Text-Image Synergy for Multimodal Retrieval and Annotation

A dissertation submitted towards the degree
Doctor of Engineering (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

by

Sreyasi Nag Chowdhury

Saarbrücken, 2021

Defense Colloquium

Date:	June 28 2021
Dean of the Faculty:	Prof. Dr. Thomas Schuster

Examination Committee

Chair:	Prof. Dr. Kurt Mehlhorn
Reviewer, Advisor:	Prof. Dr. Gerhard Weikum
Reviewer:	Prof. Dr. Gerard de Melo
Reviewer:	Prof. Dr. Klaus Berberich
Academic Assistant:	Dr. Koninika Pal

Acknowledgments

Before trying to be a grown-up and churning out an array of facts, figures, and expert jargon, let me clarify a few matters of the heart, because -

"It is only with the heart that one can see rightly;
what is essential is invisible to the eye."

Not everyone gets a chance to make acquaintances of eminence, leave alone working closely with one for a major stint. I consider myself lucky to have worked under the guidance of my "Doktorvater" Dr. Gerhard Weikum, who quite literally was a father figure of my life in academia. Not only did his professional ideas percolate in my work, I will also be forever influenced by many of his ways as a leader. I am grateful to him for giving me the opportunity to work with him.

The twists and turns in one's journey are often decided by the people they meet on the way. My professional journey till now was not an exception. One of the reasons I came to Germany for my higher studies was Dr. Subrata Dasgupta, my Computer Science professor at college, for who's guidance I'll be forever indebted. Anyone who has gone through the PhD journey is aware of the ups and downs. At a low point during my PhD, Dr. Hakan Ferhatosmanoglu's encouragement was crucial for me to revive. In Dr. Gerard de Melo I found a mentor who nudged me in the right directions and instilled confidence by acknowledging my research contributions.

I had often underestimated the importance of friendships, but looking back at my PhD journey I realise how my friends were a balm to my soul. I am thankful for the extended weekend getaways with my closest friends in Europe and in the USA, the late night musings during stay-overs, and the interesting coffee-table conversations with my friends from MPI. The memories we created helped cheer me up on-demand.

And lastly, I cannot thank enough those who I tend to take for granted. . . those who have tamed me the most. . . my constants – my parents, my sister, and my partner – for contributing in ways known and unknown towards the person I have become.

"All people have stars, but they are not the same things for different people. For some, who are travelers, the stars are guides. For others they are no more than little lights in the sky. For others, who are scholars, they are problems... But all these stars are silent. You – You alone will have stars as no one else has them."

Sreyasi Nag Chowdhury
Saarbrücken, June 28 2021

Quotes from The Little Prince by Antoine de Saint-Exupéry.

Abstract

Text and images are the two most common data modalities found on the Internet. Understanding the synergy between text and images, that is, seamlessly analyzing information from these modalities may be trivial for humans, but is challenging for software systems.

In this dissertation we study problems where deciphering text-image synergy is crucial for finding solutions. We propose methods and ideas that establish semantic connections between text and images in multimodal contents, and empirically show their effectiveness.

We present four interconnected text-and-image problems in this dissertation:

- **Image Retrieval.** Retrieving images by textual queries heavily relies on matching query keywords with words surrounding images. Images without surrounding text, or with text which is not matching but thematically related to query keywords cannot be retrieved. We propose leveraging three modalities in combination to improve image retrieval: visual (automatically detected image tags), textual (query keywords), and commonsense knowledge.
- **Image Tag Refinement.** Objects detected in images by Computer Vision tools often exhibit noise and incoherence. Correctly identifying image concepts is paramount for tasks like image retrieval with user queries. We propose to leverage commonsense knowledge to eliminate noisy and incoherent detections, and enrich image annotations by adding thematic tags which cannot be explicitly visually grounded.
- **Image-Text Alignment.** All image-text contents on the Internet (news articles, blog posts, social media stories) require images to be inserted at semantically meaningful positions in a textual write-up. We propose a framework that selects relevant images and aligns them with meaningful paragraphs of a story.
- **Image Captioning.** Images in multimodal (text-image) stories usually have captions which adhere to the context of surrounding text and elevate the reading experience. We propose the problem of contextual image captioning: generating an image caption based on surrounding text, in contrast to mainstream image captioning which only considers images alone.

Our promising results and observations open up interesting scopes for future research involving text-image data understanding.

Kurzfassung

Text und Bild sind die beiden häufigsten Arten von Inhalten im Internet. Während es für Menschen einfach ist, gerade aus dem Zusammenspiel von Text- und Bildinhalten Informationen zu erfassen, stellt diese kombinierte Darstellung von Inhalten Softwaresysteme vor große Herausforderungen.

In dieser Dissertation werden Probleme studiert, für deren Lösung das Verständnis des Zusammenspiels von Text- und Bildinhalten wesentlich ist. Es werden Methoden und Vorschläge präsentiert und empirisch bewertet, die semantische Verbindungen zwischen Text und Bild in multimodalen Daten herstellen.

Wir stellen in dieser Dissertation vier miteinander verbundene Text- und Bildprobleme vor:

- **Bildersuche.** Ob Bilder anhand von textbasierten Suchanfragen gefunden werden, hängt stark davon ab, ob der Text in der Nähe des Bildes mit dem der Anfrage übereinstimmt. Bilder ohne textuellen Kontext, oder sogar mit thematisch passendem Kontext, aber ohne direkte Übereinstimmungen der vorhandenen Schlagworte zur Suchanfrage, können häufig nicht gefunden werden. Zur Abhilfe schlagen wir vor, drei Arten von Informationen in Kombination zu nutzen: visuelle Informationen (in Form von automatisch generierten Bildbeschreibungen), textuelle Informationen (Stichworte aus vorangegangenen Suchanfragen), und Alltagswissen.

- **Verbesserte Bildbeschreibungen.** Bei der Objekterkennung durch Computer Vision kommt es des Öfteren zu Fehldetektionen und Inkohärenzen. Die korrekte Identifikation von Bildinhalten ist jedoch eine wichtige Voraussetzung für die Suche nach Bildern mittels textueller Suchanfragen. Um die Fehleranfälligkeit bei der Objekterkennung zu minimieren, schlagen wir vor Alltagswissen einzubeziehen. Durch zusätzliche Bild-Annotationen, welche sich durch den gesunden Menschenverstand als thematisch passend erweisen, können viele fehlerhafte und zusammenhanglose Erkennungen vermieden werden.

- **Bild-Text Platzierung.** Auf Internetseiten mit Text- und Bildinhalten (wie Nachrichtenseiten, Blogbeiträge, Artikel in sozialen Medien) werden Bilder in der Regel an semantisch sinnvollen Positionen im Textfluss platziert. Wir nutzen dies um ein Framework vorzuschlagen, in dem relevante Bilder ausgesucht werden und mit den passenden Abschnitten eines Textes assoziiert werden.

- **Bildunterschriften.** Bilder, die als Teil von multimodalen Inhalten zur Verbesserung der Lesbarkeit von Texten dienen, haben typischerweise Bildunterschriften, die zum Kontext des umgebenden Texts passen. Wir schlagen vor, den Kontext beim automatischen Generieren von Bildunterschriften ebenfalls einzubeziehen. Üblicherweise werden hierfür die Bilder allein analysiert. Wir stellen die kontextbezogene Bildunterschriftengenerierung vor.

Unsere vielversprechenden Beobachtungen und Ergebnisse eröffnen interessante Möglichkeiten für weitergehende Forschung zur computergestützten Erfassung des Zusammenspiels von Text- und Bildinhalten.

Contents

CHAPTER 1

Introduction

Contents

1.1 Motivation

VISION and speech are the most basic forms of human perception and communication, corroborating the importance of research in the intersection of Natural Language Processing and Computer Vision. Data on the Internet has become predominantly multi-modal, consisting of text punctuated with images. However, "understanding" of the available content by a software system is still far-fetched. Tapping into visuals along with text can go a long way in modeling human intelligence in machines, but it needs another important component – commonsense knowledge – knowledge that for humans is a product of everyday lived experiences.

In this dissertation we have studied applications at the intersection of Natural Language Processing and Computer Vision. The efforts and results are directed towards bridging the semantic gap between written expressions and high-level interpretation of visuals, leading to a better understanding of multimodal documents. We also observe that background knowledge – such as those from Commonsense Knowledge bases – aid in capturing high-level text-image synergy.

1.1.1 Multimodality in Big Data

Our experience of the real world is predominantly multimodal – we perceive the environment with our visual, auditory, tactile, olfactory and gustatory senses. These channels of information (or modalities) offer knowledge which are semantically correlated as well as complementary to each other, thus enabling the construction of patterns and connections which may not be apparent from a single source. The exponentially growing repository of unstructured data on the web reflects the multimodality of the real world. Artificial Intelligence needs to interpret such multimodal data in order to understand the real world and offer efficient assistant systems to humans.

Multimodal data is used in various everyday applications such as e-commerce (textual product details, product image/video), video games, real-time subtitle generation for the hearing-impaired and so on. Due to the COVID-19 pandemic, digital education and healthcare has become crucial and renewed the impetus for research involving multimodal data.

Various established research directions in Computer Science involve multimodality. Social Signal Processing is a cross-disciplinary research area that aims at modelling and understanding social interactions [161] such as those on social networks. Affective Computing or Affective Analysis [15] involves automatic recognition of emotions or sentiments from data produced by human actions such as written material, speech, facial expressions, brain signals etc. Sentiment Analysis and Opinion Mining are sub fields which study problems such as modelling socio-political sentiments/opinions from user reviews on public forums. A typical Google search with textual keywords returns multimodal results in the form of text, images, and videos. Research at the intersection of Natural Language Processing and Computer Vision (for example automatic generation of image captions) is inherently multimodal, and offers important building blocks for AI applications.

However, most commonly, information from each of the modalities are processed separately, independent of each other. The synergy between them is not well understood and opens up vast scope for further research.

1.1.2 Multimodality in Information Retrieval

Modern search engines not only return results from various modalities (text, image, video etc.) against a textual search query, they can also search by different input modalities, e.g., search by image and audio, search by voice query and so on. Modalities apart from the user query are also used for retrieval, e.g., the user's location and search history.

The primary challenge in multimodal Information Retrieval is tackling the semantic gap between high-level search keywords and low-level features that represent a modality (e.g., color and shape in case of images). Our work on image retrieval presented in Chapter 3 [29] mitigates this challenge by using commonsense knowledge as a semantic bridge between textual query and detected objects in images.

1.1.3 Multimodality in Computer Vision

Most Computer Vision problems deal with data that is multimodal in nature. Interesting examples include facial expression recognition [31], and eye and gesture tracking [140] where the temporal and the visual modalities are jointly interpreted.

One of the most extensively studied and elementary problems in Computer Vision is to correctly detect regions containing objects in images and map them to semantic labels which may be object categories, or words within image captions. Since object detection algorithms utilize low-level image features to classify objects, detections are often noisy and incoherent: for example, a blue wall may be tagged 'ocean'. Moreover, an image conveys more meaning to a human than merely the objects it contains. Such high-level thematic concepts (e.g., depicted emotions such as 'happy occasion' or generalizations such as 'sport' and 'entertainment') which cannot be grounded in the image are not captured by automatic object detection algorithms. We study and propose solutions to these limitations in our work on image tag refinement presented in Chapter 4 [28].

1.1.4 Multimodality in Natural Language Processing

Natural language is inherently multimodal – humans communicate through various modalities including linguistic, visual (gestures, facial expressions), voice intonation and so on. This phenomenon has transcended to material found on the Internet, most commonly a combination of text and images, in blog posts, news articles, e-commerce, marketing, digital education, and social media. Identifying and analysing gesture and expressions fall in the purview of Computer Vision, and the role of voice tones is studied through research in Speech Processing. Natural Language Processing (NLP) deals mainly with language in the form of text, increasingly accompanied with visuals.

The NLP research community has grappled to understand the inter-modality synergy through the years. This has led to important contributions in multimodal representation learning [59], which in turn benefit downstream vision-and-language applications. However, one artefact of learned representations is that they are not explainable, but rather function as a black box [1]. To initiate research in finding explainable text-image relations, we briefly study a multimodal dataset and present annotations towards text-image discourse relations [4]. We also propose two novel multimodal NLP tasks: text-image alignment (Chapter 5 [27]) and contextual image captioning (Chapter 6 [25]).

1.2 Thesis Contributions

The aim of this dissertation is to better understand the role of text and images in multimodal contents and thereby study their synergy. We propose and solve novel problems in the areas of Information Retrieval, Computer Vision, and Natural Language Processing with the observation that leveraging both modalities jointly lead to better results in vision-and-language tasks. The contributions in this dissertation can be outlined with the following scenario:

A content creator (journalist/blogger/columnist) would like to illustrate her article with appropriate images. She may search for images on the web using textual queries, for which we propose an efficient image retrieval model in Chapter 3. Alternatively, she may look for images in her personal image repository. The latter is a laborious task when the images in the repository are not associated with meaningful tags to facilitate search. To mitigate this, in Chapter 4 we propose a model which generates a coherent set of image tags comprising visual as well as thematic concepts. Next, the tool proposed in Chapter 5 helps the author to automatically select relevant images and align them to meaningful paragraphs within her write-up. Finally, having aligned images to relevant textual contexts, our model proposed in Chapter 6 can be used to automatically generate novel image captions conditioned on surrounding text.

To this end, the contributions of this thesis can be summarized as follows, along with the respective publications. As the main author in all the publications, I was the principal investigator, as well as the primary developer, with developing assistance from co-authors for the publications in WSDM 2020 and LANTERN 2021.

- Sreyasi Nag Chowdhury, Niket Tandon, and Gerhard Weikum. **Know2Look: Commonsense Knowledge for Visual Search.** *Proceedings of the 5th Workshop on Automated Knowledge Base Construction* (AKBC), pages 57–62. San Diego, California, USA, June 17 2016.

  This work aims at harnessing the multi-modality of data on the web using Commonsense Knowledge (CSK). The case we study is that of web images, especially significant due to the social media boom. Traditional search and retrieval of images is largely dependent on solely textual cues which may be ambiguous and incomplete. Inefficiencies in computer vision technologies also do not lead to perfect visual detections. We hypothesize that the use of background CSK on query terms along with the textual and visual content of images can significantly improve image retrieval performance. To this end we deploy three different modalities - text, visual cues, and CSK pertaining to the query - as a recipe for efficient search and retrieval.

- Sreyasi Nag Chowdhury, Niket Tandon, Hakan Ferhatosmanoglu, Gerhard Weikum. **VISIR: Visual and Semantic Image Label Refinement.** *Proceedings of the 11th ACM International Conference on Web Search and Data Mining* (WSDM), pages 117-125. Los Angeles, California, USA, Feb 5–9 2018.

  Advances in computer vision research has made automatic detection of visual contents in images and videos possible. Although prone to huge error margins, recent results [54] look promising with detection accuracy as high as 70% on certain benchmarks. However, all existing visual recognition architectures ignore the context of the detections. For example, a *tennis racket* and a *lemon* can be detected in an image showing a tennis game. The obvious flaw here is that because of the lack of context, a *tennis ball* has been wrongly detected as a visually similar object - *lemon*. We aim to solve this problem, thereby cleaning the detection labels, by establishing semantic relatedness relationships between different bounding box detections in the same image. We further enrich the tag space with generalizations and abstractions from Commonsense Knowledge.

- Sreyasi Nag Chowdhury, William Cheng, Gerard de Melo, Simon Razniewski, and Gerhard Weikum. **Illustrate Your Story: Enriching Text with Images.** *Proceedings of the 13th International Conference on Web Search and Data Mining* (WSDM), pages 849–852. Houston, Texas, USA, Feb 3–7 2020.

  Sreyasi Nag Chowdhury, Simon Razniewski, and Gerhard Weikum. **Story-oriented Image Selection and Placement.** *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (EACL). Kyiv, Ukraine (virtual conference), April 19–23 2021.

  The most effective way to deliver a message to humans happens to be a combination of textual and visual cues. Naturally, multimodal documents have become common on the Internet in the form of news articles, blog posts, personal travel accounts and social media stories. The generation of such multimodal content requires considerable human judgment and reasoning. Selecting relevant images from a bigger pool and placing them in the correct context within a body of text can be time-consuming and labor intensive. Examples of such tasks may be seen in small-scale generation of personal stories, or industry-scale generation of promotional content for advertising. We present a framework called SANDI, that automatizes the process of selection of images from a thematically related or unrelated pool, and aligns those images in suitable paragraphs in a given body of text.

- Sreyasi Nag Chowdhury, Rajarshi Bhowmick, Hareesh Ravi, Gerard de Melo, Simon Razniewski, and Gerhard Weikum. **Exploiting Image–Text Synergy for Contextual Image Captioning.**

Proceedings of the 3rd Workshop Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN) 2021. EACL 2021 Workshop. Kyiv, Ukraine (virtual conference), April 19–23 2021.

Automatic caption generation is a well-researched field in the intersection of Computer Vision and NLP. Existing caption generation frameworks generate captions which heavily adhere to explicit visual content, ignoring high-level thematic and sentimental narratives. However, more often than not, images are part of a story. We present the problem of generating image captions conditioned on surrounding texts. Such captions capture the thematic context of the image, and abstract away from explicit visual content.

1.3 Organization

The remainder of the thesis is organized as follows. Chapter 2 gives a background on prior research and methodology relevant for this dissertation. Chapter 3, 4, 5, and 6 describe our contributions related to image retrieval, image tag refinement, text-image alignment, and contextual image captioning respectively. We conclude with an outlook on future research directions in Chapter 7.

CHAPTER 2

Background

Contents

THE synergy between text and images have long intrigued the Computer Vision and Natural Language Processing research communities. The purpose of this chapter is to position this dissertation in the broad research spectrum.

2.1 Related Works

Prior research that motivates this dissertation can be categorized under *Language and Vision* (research in the intersection of Natural Language Processing and Computer Vision) and *Commonsense Knowledge*. Some of the work discussed in this section only remotely relates to the contributions of this dissertation. More closely related work corresponding to specific contributions is discussed in the respective chapters (3, 4, 5, 6).

2.1.1 Vision and Language

Very early work in this domain involved using textual meta-data from images (timestamps, locations, user tags) to draw associations with text. These attempts were followed by learning representation of images from low-level image features (color, texture, shape) and associating them with textual keywords. With the success of newly-introduced deep learning frameworks, more sophisticated and fine-grained representations for both images and text can now be learned. Common practice in language-and-vision tasks is to learn image and text representations from a joint embeddings space. Research in this field are summarized under the following non-exhaustive areas.

2.1.1.1 Image Retrieval

Early approaches to image retrieval relied on textual keywords associated with images such as image name, textual tags, captions, and descriptions [134]. Such Tag Based Image Retrieval (TBIR) [100, 93] rely on manual annotations which are often incomplete (not capturing all visual information featured in an image), diverse (semantically varied tags for visually similar images), and language-dependent. These drawbacks motivated the research field of Content Based Image Retrieval (CBIR) which search and retrieve images similar to a given query image based on low-level image features such as color, shape, texture etc. [159, 24, 97, 6]. Due to heavy computational requirements in CBIR, most search engines often use textual tags from a small set of visually similar images to expand the search to bigger image repositories. In text-only, vision-only, and text-cum-vision image retrieval systems, there exists a semantic gap between the users' queries and the visual content of the images in the image repository. We address this issue in Chapter 3, and attempt to bridge the semantic gap through the use of commonsense knowledge. Another similar work [68] shows how commonsense can improve image retrieval.

2.1.1.2 Image Attribute Recognition

High level concepts in images lead to better results in Vision-to-Language problems [171]. Identifying image attributes is the starting point toward understanding text-image synergy. To this end, several deep-learning based modern architectures have been built to detect visual concepts in images through object recognition [64, 129, 131], scene recognition [185], and activity recognition [55, 178, 184]. We leverage some frameworks from this category in our proposed models to detect visual concepts in images. Since all these frameworks work with low level image features like color, texture, gradient etc., noise creep in often leading to incoherent or incorrect detections. For example, a blue wall could be detected as 'ocean'. While some of the incoherence can be refined using background knowledge (as we observe in our contribution [28] discussed in Chapter 4), considerable inaccuracy still exists.

2.1.1.3 Multimodal Embeddings

A popular method of semantically comparing images and text has been to map textual and visual features into a common space of multimodal embeddings [51, 160, 47]. Semantically similar concepts across modalities can then be made to occur nearby in the embedding space. Visual-Semantic Embedding (VSE) has been used for generating captions for the whole image [47], or to associate textual cues to small image regions [71] thus aligning text and visuals. Visual features of image content (for example color, geometry, aspect-ratio) have also been used to align image regions to nouns (e.g. "chair"), attributes (e.g. "big"), and pronouns (e.g. "it") in corresponding explicit textual descriptions [73]. However, alignment of small image regions to text snippets play little role in jointly interpreting the correlation between the whole image and a larger body of text. We focus on the latter in our contribution on text-image alignment discussed in Chapter 5.

2.1.1.4 Generating Image Captions and Descriptions

Generation of natural language image descriptions is a popular problem at the intersection of computer vision, natural language processing, and artificial intelligence [12]. Image captioning was first approached as a retrieval problem [49] – retrieve a sentence that best describes an image from a sentence database. This naturally yielded inaccurate captions, especially for novel combinations of visual concepts. More modern frameworks use a combination of Convolutional Neural Network (CNN) to encode image features and Recurrent Neural Network (RNN) to generate natural language captions [43, 162, 48, 102, 71], recently also using attention-based neural networks [176]. Methodologically they are similar to VSE in that visual and textual features are mapped to the same multimodal embedding space to find semantic correlations. Prior work also explore leveraging external knowledge for image captioning [172, 187].

While most existing frameworks generate descriptive captions [176, 150, 101], some of the more recent architectures venture into generating stylized captions by the addition of linguistic flavors such as humor [52], usage of puns [17] and sentiments [166, 22, 141]. These stylized captions still largely describe the visual contents of the image, with the use of occasional abstract concepts such as “look good”, “adorable cat” etc. [116]. Recent work on description generation [76] use dense captioning (captions for different regions in the image) to generate natural language descriptive paragraphs for images. Although they claim to capture the ‘story’ of the image, the descriptions only contain visual cues ignoring high-level abstract and sentimental narratives.

The nature of captions/descriptions that accompany images on the web (on social media channels and blogs) often capture the abstract theme instead of an insipid account of the visual contents of the image. However, generation of abstract image captions is a subjective problem making evaluations challenging. One of the few research works that look beyond descriptive captions generates poetry from images [91]. Personalized image captions [119] takes into account user’s recent vocabulary to generate descriptive or abstract captions. In Chapter 6 we study the problem of generating image captions conditioned on surrounding contextual text. These captions incorporate abstract thematic concepts from the context while loosely adhering to the visual contents of the image.

2.1.2 Commonsense Knowledge

Research on Commonsense Knowledge has gained traction in the recent past in an attempt to better equip software systems to assist humans. We can define Commonsense Knowledge as knowledge about generic relationships between common concepts – for example *(flower, hasProperty, fragrant)* – that humans inherently use in daily interactions and reasoning. Interesting examples of human commonsense can be seen in natural language and visual understanding. For example, the sentence “She took the cake out of the fridge and ate it.” is easy for humans to parse – the edible object out of the two object mentions (*cake* and *fridge*) is clearly the *cake*. However, in absence of such inherent knowledge, the resolution of the pronoun *it* is tricky for a software system. To the average human, *commonsense* refers to *good judgement*, while for AI research, *commonsense* refers to a large set of basic knowledge possessed by humans. The formal way to represent the knowledge required to parse the example sentence is through subject-relation-object triples – *(cake, hasProperty, edible)*, *(fridge, hasProperty, inedible)*. Use of such latent background knowledge in many contemporary applications has the potential to bridge the gap between any software service and the end user (humans).

2.1.2.1 Commonsense Knowledge from Text

Traditionally, commonsense knowledge bases were curated manually through experts [81] or through crowd-sourcing [143]. The latter, the 1999 Open Mind Commonsense (OMCS) Project of the MIT Media Labs, is the most elaborate attempt to collect human commonsense – a World Wide Web based collaboration with over 14,000 authors. This led to the collection of ~700,000 sentences which would be used later to curate a commonsense knowledge base. OMCS differs from similar previous attempts like Cyc [81] in that it collects natural language English sentences, as opposed to using formal logic structure.

The most popular commonsense knowledge base in use today, ConceptNet [96], was initially constructed from sentence collected through the OMCS project. Section 2.3.2 elaborates on ConceptNet. More recent methods of commonsense knowledge acquisition tap into the vast text resources on the web [152] and query logs [133]. The major shortcoming of all available general purpose commonsense knowledge bases is incompleteness – there is no exhaustive list of commonsense knowledge available to date. This can also be attributed to the unavailability of a concrete definition of commonsense owing to its subjectivity. Commonsense often depends on culture (for example, what kind of present is appropriate for a friend) and life experiences (for example, what kind of clothing is appropriate for sub-zero temperatures). The other shortcoming is the lack of informative confidence scores for commonsense knowledge assertions. There has been a recent attempt to consolidate the various sources of commonsense knowledge into one comprehensive commonsense knowledge graph [69]. The authors also propose a probabilistic scoring mechanism.

2.1.2.2 Commonsense Knowledge from Visual Data

Acquiring commonsense knowledge exclusively from text has its limitations – common human wisdom is often not explicitly penned down, leaving the reader to "read between the lines". Visual sources however, are a rich source of such 'hidden' knowledge. Prior work explore ways to learn commonsense knowledge from real images [23] as well as from non-photo-realistic abstractions [157]. Recent work have leveraged commonsense knowledge for visual verification of relational phrases [136] and for non-visual tasks like fill-in-the-blanks by intelligent agents [90]. Learning commonsense from visual cues continue to be a challenge in itself.

In the era of deep learning and large language models trained on web-scale text like OpenAI's[1] GPT3 [14], commonsense knowledge bases that have been developed from external sources (like images, games etc.) seem to be losing relevance. However, since commonsense knowledge is most often not captured in writing, it is somewhat safe to assume that language models learned from web-scale text data do not necessarily posses fine-grained commonsense. Therefore, although these language models perform well on tasks like generation of congruent texts and factual question answering, they fail in conceptual tasks and decision making which require commonsense reasoning. Hence research on commonsense knowledge acquisition from non-textual sources is paramount for creating efficient intelligent systems.

Throughout this dissertation, we leverage existing commonsense knowledge bases as additional components in our models – to semantically connect user's textual query and detected image objects for image retrieval (Chapter 3), and to expand the image tag space with thematic concepts (Chapters 4 and 5). In most cases we observe an improvement in model performance.

[1] https://openai.com

2.2 Methodology

The methodology discussed in this section lay the technical groundwork for the research work presented in this dissertation.

2.2.1 Statistical Language Models

Statistical Language Models have been successfully used in many Information Retrieval (IR) and Natural Language Processing(NLP) problems like ranking, machine translation, speech recognition, and so on. In IR systems, language models are used to score documents in order to rank them according to their relevance to a given query. In NLP applications, they are used to predict typical word sequences according to the usage of a particular language. In this section, we briefly discuss the basics of statistical language modelling.

A Statistical Language Model (LM) is a probability distribution over a vocabulary of strings (for simplicity, we consider a string as a word). Given a sequence of words $(w_1, w_2, ...w_n)$, a language model generates a probability for each word, and hence can be used to predict the next word in a given sequence. The probability of a sequence of words can be decomposed into the probability of each successive word conditioned on previous words (by the chain rule of probability).

$$P(w_1, w_2, w_3, w_4) = P(w_1)P(w_2|w_1)P(w_3|w_2w_1)P(w_4|w_3w_2w_1) \tag{2.1}$$

An *unigram language model* ignores conditional dependencies and calculates the probability of each word independently. The order of words in an unigram language model is irrelevant, and is often referred to as a "bag of words" model.

$$P_{unigram}(w_1, w_2, w_3, w_4) = P(w_1)P(w_2)P(w_3)P(w_4) \tag{2.2}$$

A *bigram language model* considers the dependency between two consecutive words.

$$P_{bigram}(w_1, w_2, w_3, w_4) = P(w_1)P(w_2|w_1)P(w_3|w_2)P(w_4|w_3) \tag{2.3}$$

An *n-gram language model* is similarly defined where the dependency is only on the last $(n-1)$ words. The unigram and bigram language models are among the simplest and commonly used statistical language models.

In contrast to the continuous bag of words models (the previously discussed n-gram language models) where consecutive words are considered together, a *skip-gram language model* considers words while skipping context words in between. For the sentence "This is a sentence about skip-grams.", word-pairs ('This', 'sentence'), ('a', 'about'), ('is', 'about') are considered which would otherwise have been ignored since they do not occur consecutively.

Query-likelihood Language Model. The query-likelihood model is the most basic approach of using language model in IR [122]. It models the likelihood of a document d being relevant to a given query q. Using Bayes rule,

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \tag{2.4}$$

The prior probability of the query $P(q)$ is constant for all documents. For simplicity, the prior probability of documents $P(d)$ can be considered uniform for all documents. Hence, the likelihood of a document to be relevant to a given query is approximated by $P(q|d)$, or the probability with which the query q is generated from the language model of document d. In a document retrieval task, language models (LM) are inferred for each document d in the collection, the probability $P(q|LM_d)$ of generating the given query from the document is calculated, and the documents are ranked according to these probabilities.

Using maximum likelihood estimate, the probability of generating query q from language model LM_d is:

$$P(q|LM_d) = \prod_{w \in q} \hat{P}_{mle}(w|LM_d) = \prod_{w \in q} \frac{tf_{w,d}}{L_d} \tag{2.5}$$

where $tf_{w,d}$ is the frequency of word w in document d, and L_d is the total number of words in document d. The hat on $\hat{P}_{mle}(w|LM_d)$ indicates that the probability is estimated. If a query word does not appear in the document, $\hat{P}_{mle}(w|LM_d) = 0$. This leads to the retrieval system failing to retrieve documents which do not contain the exact query words, but are otherwise relevant. To mitigate this issue – that is, to assign some non-zero probability to unseen words and avoid over-fitting – the probability distribution is *smoothed*.

Language model *smoothing* can be done in various ways, the most basic being adding a multinomial distribution from the entire document collection with the document-specific multinomial distribution.

$$\hat{P}(w|d) = \lambda \hat{P}_{mle}(w|LM_d) + (1 - \lambda)\hat{P}_{mle}(w|LM_c) \tag{2.6}$$

where LM_c is a language model built from the entire document collection, and $\lambda \in [0, 1]$ is a controlling hyper-parameter.

Thus, the likelihood of a document d to be relevant to a given query q, can be defined as:

$$P(d|q) = P(d) \prod_{w \in q} (\lambda P(w|LM_d) + (1 - \lambda)P(w|LM_c)) \tag{2.7}$$

We will revisit query-likelihood language models in Chapter 3.

Neural Language Model. While query-likelihood language models are count-based methods, neural language models (language models based on neural networks) are continuous-space models. In neural language modelling (NLM) [11], words are first represented as vectors: for example, each word may be initialised as a 26-dimensional vector (for 26 letters of the English alphabet) where the vector indices corresponding to the letters in the word contain the value 1 while the rest of the indices contain the value 0. Through the training process sophisticated word representations in the continuous vector space (vectors with fractional values instead of only two values 1 and 0) are then learned which have properties such as representations (or embeddings) of semantically similar words are close-by in the learned vector space.

The first NLMs [11, 113] propose feed-forward neural networks that use fixed context lengths. Later, Recurrent Neural Network (RNN) based models have been proposed [106] which deal with unlimited context lengths that represent short-term memory. More recently, transformer-based language models [40] have been proposed which process sequences without recurrence, using a method called self-attention [156]. Memory is ensured by learning contextual information in the form of embeddings that represent the position of a token from the start of the sequence. This class of language models are called masked language models after their training objective: a portion of tokens in the training corpus are 'masked' or hidden from the model during training, and parameters are learned through the loss incurred in the model's prediction of these masked tokens. Although this method constitutes supervised learning, training only requires running text and is not dependent on large scale annotation efforts. We leverage learned representations from a state-of-the-art masked language model, BERT [40], for the image captioning problem presented in Chapter 6.

2.2.2 Open Information Extraction

Open Information Extraction (OpenIE) is a paradigm that deals with the extraction of relation tuples from unstructured text. In the simplest form, OpenIE extracts information in the form of subject-relation-object triples. Unlike other forms of information extraction, OpenIE does not rely on pre-defined relations and manually crafted domain-specific pattern-matching rules. Additionally, OpenIE smoothly scales to handle Web-scale corpora. The relations is OpenIE are typically just the text that links two arguments. For example, from the sentence “Albert Einstein was born in Ulm.”, OpenIE would create a triple (Albert Einstein; was born in; Ulm) corresponding to the relation was-born-in.

The predecessors of OpenIE include supervised [145] and self-supervised methods like the KnowItAll WebIE system [44]. While supervised methods relied on labelled examples to learn extraction rules, semi-supervised methods first created labelled examples automatically and then learned rules from them. In both cases the set of relations had to be manually pre-defined. In contrast, OpenIE can extract an unlimited number of relations in addition to being completely unsupervised.

Some of the well-known OpenIE systems are as follows:

TextRunner. TextRunner [180] was the first proposed OpenIE system. It identifies relations using a conditional random field (CRF) and solves a sequence labelling problem to assign labels (‘ENT’ for entity and ‘REL’ for relation) to each word in a sentence.

ReVerb. Instead of learning relations, ReVerb [45] uses syntactic constraints in the form of Part of Speech (POS) based regular expressions (capturing verb phrases) to identify relations. This eliminates incoherent and uninformative relations which appeared in TextRunner. Additionally, ReVerb considers only those relations which appear multiple times (over a defined threshold) in the corpus to avoid over-specific relations. ReVerb finds triples in two phases – first identifying verb phrases as relations, followed by identifying noun phrases connected by the relations.

OLLIE. In order to accommodate complex assertions, OLLIE [104] extracts a context for each assertion that attributes the claim in the assertion to additional entities. For example, given the sentence “Early scientists believed that the earth is the center of the universe.”, OLLIE extracts the triple *(the earth; be the center of; the university)* `Attributed To` *(believe; early scientists)*.

ClausIE. While the previous OpenIE systems all extract binary relationships, ClausIE [32] is a clause-based system that extracts relations with higher arity. For example, from the sentence “The doorman showed Albert Einstein to his office.”, ClausIE extracts the tuple *(The doorman, showed, Albert Einstein, to his office)*.

The major advantage of OpenIE is its ability to quickly gather large volumes of information from the web which can be utilized for various downstream applications. Three such applications are question-answering, fact-checking, and opinion mining. In question-answering, millions of triples can be collected based on an information need or question. For fact-checking, agreement or conflict of a given assertion can be established against a big corpus of domain-specific triples extracted through OpenIE. Similarly, public opinions can be harnessed through OpenIE triples extracted from product reviews or political discussion forums.

We use the OpenIE tool ReVerb for commonsense knowledge acquisition from the web, discussed more in Chapter 3.

2.2.3 Integer Linear Programming

Linear programming is a mathematical optimization paradigm (to achieve the best outcome such as maximum profit or lowest cost) where the model is defined by linear relationships, i.e., the objective function and the constraints are linear. An Integer Linear Program(ing) (ILP) is a linear programming problem where all the variables are restricted to be integers. ILPs are useful where the variables represent quantities which can only be integers (for example, number of cars), or when they represent decisions (binary variables 0 and 1 representing yes or no). ILPs belong to the class of NP-complete problems.

An ILP takes the following mathematical form.

Objective: $maximize\ c^T x$
Constraints: $Ax = b$ (linear constraints)
$l \leq x \leq u$ (bound constraints)

We use the Gurobi ILP solver[2]. The solver first runs the Presolve algorithm on the initial problem, followed by Branch-and-Cut, which is a Branch-and-Bound algorithm using Cutting Planes to tighten the (Linear Program) LP formulations. The LPs within the ILP are solved using the Simplex algorithm. We briefly introduce each of these algorithms here.

Presolve. Solving an ILP in time intensive. Hence reducing the size of the problem prior to finding solutions is paramount. Presolve refers to problem reductions that are applied before the branch-and-cut algorithm in order to tighten the problem formulation.

Cutting Planes. In an ILP formulation, the variables are restricted to be integers. Cutting Planes refer to the algorithm that removes fractional solutions and tightens the problem formulation without dividing the problem into sub-problems. Cutting Planes work by finding new inequalities that cut off the current solution space.

Branch and Cut. A Branch-and-bound algorithm which employs Cutting Planes during its solution is referred to as Branch-and-Cut. To begin with, all the integrality restrictions of the ILP is removed. The resultant problem, called a LP relaxation, is then solved. When a fractional solution is encountered, and Cutting Planes have tightened the problem formulation, the branch-and-bound algorithm is started by splitting the problem into two (or more) sub-problems. Each node in the branch-and-bound search-tree is a new LP. The non-integral solutions in the LP relaxations serve as upper bounds while the integral solutions serve as lower bounds. When an upper bound is lower than a lower bound, the node is pruned. If all the integrality restrictions of the original ILP are satisfied at a particular node, the solution at the node is one of the feasible solutions of the ILP. Based on the objective function of the ILP (either maximization or minimization), the optimal solution is then calculated from all feasible solutions.

Simplex Algorithm. The SimpleX Algorithm [34] solves a LP problem by constructing a feasible solution at a vertex of the convex polytope (a geometric object with flat sides) and then walking along a path on the edges of the polytope (such that the next vertex has a higher value of the objective function) until an optimal is reached.

We formulate ILP-based models for the problems presented in Chapter 4 (refinement of automatically detected image tags) and Chapter 5 (semantic alignment of images to textual paragraphs).

[2] https://www.gurobi.com/products/gurobi-optimizer/

2.2.4 Deep Learning

Deep Learning is a paradigm for automatically learning useful representations from data. In the last few years, Deep Learning has become indispensable in learning from and analysing large quantities of data, especially in the Computer Vision and Natural Language Processing communities. Here we provide a brief overview of the basic concepts of deep neural networks – networks of nodes and edges that facilitate Deep Learning. 'Deep' refers to the structure of the most efficient neural networks, which can be imagined as several networks stacked on top of each other, output of one network influencing computations in the next.

Layer. A 'layer' in a neural network is a collection of nodes that compute data representations. Neural Networks always contain an input layer (which receives the raw data, e.g. pixels in an image) and an output layer (which produces the output of the given task, e.g. classifying the image into classes 'cat' or 'dog'). Sandwiched in between are zero or more 'hidden layers' where most of the computations take place. Each node in a hidden layer receives input from one or more nodes of the previous layer. A 'dense layer' is a layer which receives input from all nodes of the previous layer. The complexity of a neural network, or how 'deep' it is depends on the number of hidden layers. Each node in a hidden layer is typically responsible for computations for one example input (a n-dimensional vector representing the data point). The computation is a weighted sum of the values of the input vector and addition of a bias:

$$\hat{y} = \sum_{i}^{n} w_i^T x_i + b \tag{2.8}$$

where x_i is a value in the n-dimensional input vector, the weights w_i and bias b are parameters which are learned during training, and $\hat{y}$ is the predicted output. The formulation in Equation 2.8 is referred to as a linear feed-forward layer.

Activation Function. Not all mathematical functions can be modeled with the linear relationship shown in Equation 2.8. A function that adds non-linearity to the model is called an 'Activation Function'. The computation at each node then takes the following form:

$$\hat{y} = f(\sum_{i}^{n} w_i^T x_i + b) \tag{2.9}$$

where f is the activation function. An example of an activation function is *sigmoid*, denoted as σ:

$$\hat{y} = \sigma(y) = \frac{1}{1 + e^{-y}} \tag{2.10}$$

The estimation of the output value $\hat{y}$ is then propagated to the next layer.

The activation function often seen in the last layer of deep neural networks is called a *Softmax* activation. The number of nodes in the last layer correspond to the number of possible outcome (for example, two in case of a binary classification problem). The Softmax activation maps the values of the nodes into probabilities.

Loss Function. Once the predicted values are propagated to the last layer of the neural network, they are compared with the original output values. The 'loss function' measures the amount of deviation (or 'loss' in accuracy) from the original output value. The commonly used loss functions are *Mean*

Squared Error for regressions problems and *Cross Entropy* for classification problems.

$$Mean\ Squared\ Error = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2 \tag{2.11}$$

where y_i is the actual output for sample i and $\hat{y_i}$ is the estimated output. m represents the number of data samples. For a binary classification problem (with two classes denoted by 0 and 1), the Cross Entropy loss is calculated as follows:

$$Cross\ Entropy = -(ylog(p) + (1-y)log(1-p)) \tag{2.12}$$

where y is the true label for the data point, and p is the probability with which the label is predicted. This can be simply extrapolated to multi-class classification.

Backpropagation. The objective of the training process is to minimise the loss by finding the optimal values for the parameters (weights w_i and bias b from Equation 2.8). Backpropagation is the algorithm that computes the gradient (rate of change) of the loss with respect to the parameters. This involves propagating the loss backward through the layers and computing partial derivatives of the cost function. The values of the weights and bias at every node are then updated as follows:

$$w_i = w_i - \left(\alpha \times \frac{\partial C}{\partial w_i}\right) \tag{2.13}$$

$$b = b - \left(\alpha \times \frac{\partial C}{\partial b}\right) \tag{2.14}$$

where C is the loss function and α is a hyperparameter called the ‘learning rate’ which controls how much the weights and bias are changed at each iteration. A high learning rate leads to faster training, but generates sub-optimal solutions. Backpropagation and gradient descent continues till convergence.

Encoder-decoder Architecture. An encoder refers to an architecture that finds patterns in raw data to construct useful representations. A decoder uses the learned representations to generate new data. An example of an encoder-decoder architecture is an image captioning system where the encoder learns representations of raw images and the decoder then generates textual captions from the image representations. The encoder and decoder architectures may be composed of several CNN [79] or RNN (for example, Long Short Term Memory Network (LSTM) [63]) units. Encoder-decoder architectures are most commonly used to build sequence-to-sequence models for tasks likes machine translation, question-answering etc.

Attention. In sequence-to-sequence (commonly called seq2seq) architectures, ‘attention mechanism’ allows the model to focus on parts of the input sequence that are important to generate the output sequence. A sophisticated deep learning architecture called ‘Transformer’ shows that attention mechanism lifts the reliance on recurrent units to capture dependencies in sequences [156]. A Transformer-based language model called BERT [40] currently offers the most efficient language representations.

We propose a deep learning based model for the image captioning problem in Chapter 6.

2.3 Resources

The models introduced in this dissertation have been integrated with background knowledge from various resources. Some of the general purpose resources which will frequently appear in the subsequent chapters are being discussed here. Other specialized resources will be discussed in respective chapters where applicable.

2.3.1 WordNet

WordNet [109] is a large lexical database of English words and their semantic relationships. It has also been constructed for ~200 other languages. WordNet has been a valuable resource for a number of NLP problems like word-sense disambiguation, text summarization, text classification, machine translation, as well as problems in IR.

The English WordNet contains nouns, adjectives, verbs, and adverbs, grouped into ~117,000 *synsets* containing synonymous word forms. The synsets are accompanied with definitions and usage examples, presenting a combined dictionary-thesaurus functionality. The relations that connect synsets are:

- Hypernymy – A is a hypernym of B if every instance of B belongs to the class A. E.g., *feline* is a hypernym of *cat*.
- Hyponymy – A is a hyponym of B if every instance of A belong to the class B. E.g. *cat* is a hyponym of *feline*.
- Holonymy – A is a holonym of B is B is a part of A. E.g., *car* is a holonym of *wheel*.
- Meronymy – A is a meronym of B if A is a part of B. E.g. *wheel* is a meronym of *car*.
- Troponymy – A is a troponym of B if B is some manner of A. E.g. *running* is a troponym of *jogging*.
- Entailment – A is entailed by B if B involves A. E.g. *eating* is entailed by *swallowing*.
- Coordinate term – Words sharing a common hypernym. E.g. *car* and *truck*; hypernym *vehicle*.

Most word-word relations are restricted to the same part of speech, forming in essence four sub-nets for nouns, adjectives, verbs, and adverbs. A few links exist between different parts of speech. These are called *morphosemantic* relations, linking word senses containing the same stem word. For e.g. the noun *teacher* and the verb *teach* are connected with a morphosemantic link.

WordNet word senses have been considered as an unifying standard in various other resources. The large image database ImageNet [38] maps images to the WordNet senses of the corresponding objects for sense disambiguation (for e.g. **bank** (institution/building) as opposed to river *bank*). However, WordNet sense distinctions are too fine-grained, which led to annotation errors in the construction of the database. For e.g., *sunglass* is defined as a convex lens, while *sunglasses* is defined as spectacles. This distinction is often irrelevant to human annotators, leading to incorrectly classified images in ImageNet. This results in errors in applications like object recognition which train on ImageNet.

We have utilised the synsets and hypernymy-hyponymy relations in WordNet to capture generalisations of word senses. Such background knowledge can be considered as *commonsense knowledge* – for e.g., *dog* is an *animal*, *woman* is a *person*, *bed* is a *furniture*, and so on. The details will be discussed in Chapters 3 and 4.

2.3.2 ConceptNet

ConceptNet [96] is a large commonsense knowledge (CSK) base automatically constructed by applying natural language processing and extraction rules to 700,000 crowd-sourced statements from the Open Mind Common Sense Project [143]. Later versions of ConceptNet include knowledge from various other resources like DBPedia, Wiktionary, and "games with purposes" like Verbosity. The version of ConceptNet at the time of writing this dissertation – ConceptNet 5.7 – contains 34,074,917 assertions regarding physical, spatial, social, temporal, and psychological aspects of daily life. The assertions are represented as *(subject, relation, object)* triples. There are 50 unique relations. Some relations with larger number of assertions are – /r/Causes, /r/LocatedNear, /r/UsedFor, /r/HasProperty, /r/PartOf, /r/RelatedTo, /r/HasA, /r/IsA, /r/CapableOf, /r/AtLocation. The English slice of ConceptNet (where both the subject and the object are English concepts) contains 3,410,732 assertions. Figure 2.1 show a depiction of the ConceptNet commonsense knowledge graph.

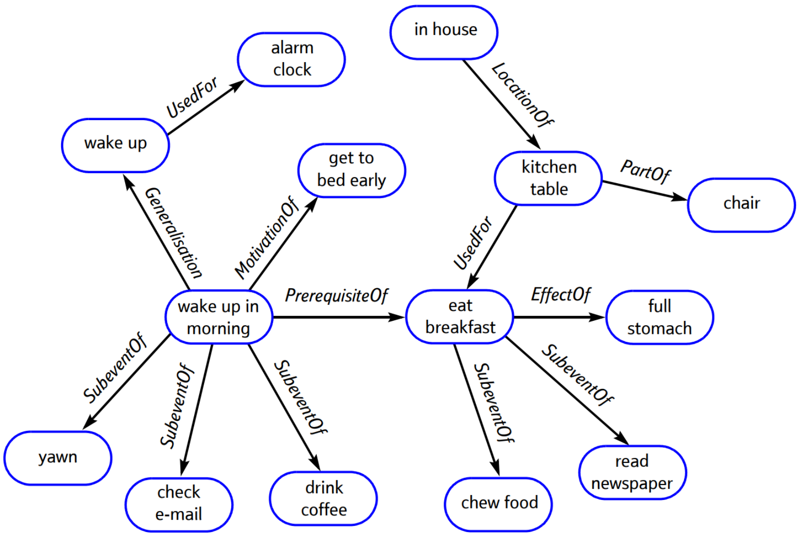


Figure 2.1: A fragment of ConceptNet. Image taken from [96].

ConceptNet is available as a web tool for exploration, as well as a natural language processing toolkit for research purposes. We use ConceptNet as a background knowledge resource for most of the research presented in this dissertation, but we pick and choose the appropriate relations since only some of them are ancillary for our purposes. The details of the relations selected for a particular task will be discussed in respective chapters.

2.3.3 Word Embeddings

Word embeddings are vector representations of words. An embedding captures the semantic representation of a word in a numeric form, thus facilitating mathematical operations on it. They can be generated by various methods. In the simplest form, a word can be represented as a one-hot vector. A one-hot vector is a vector of zeros except for the element at the index representing the corresponding word in the vocabulary. For example, let's consider a vocabulary of 10 words – (a, and, because, dog, is, let, live, place, the, world). The one-hot vector representation for the word "live" is [0 0 0 0 0 0 1

0 0 0] – all elements except that at the seventh index (which is the position of the word “live” in the vocabulary) contains 0. Such a simple word representation is ineffective for real applications since the embeddings do not capture semantic relationships between words. More sophisticated methods are required such that embeddings of words that are closer in meaning are similar. Word2Vec [107], GloVe [120], and more recently BERT [40] are the most common frameworks for generating word embeddings. We use all of these three methods at various points throughout this dissertation.

Word2Vec. Word2Vec embeddings can be constructed using two methods – Continuous Bag of Words (CBOW) and Skip Gram. The CBOW method takes a context (words surrounding a target word) as input and predicts the corresponding target word. It follows the bag-of-words assumption, that is, the order of words in the context are irrelevant for prediction. The vector representations of words are learned in the process of predicting the target word. The Skip Gram method solves the inverse problem of predicting the context words from a given word. Higher weights are assigned to words closer to the given word than those that are further away. While CBOW is faster to train, Skip Gram learns better representation for infrequent words.

GloVe. Global Vectors, or GloVe in short, are word representations learned from aggregated global word-word occurrence statistics of a large text corpus. As a prerequisite, a one-time computationally expensive pass over the entire corpus is required to populate the co-occurrence matrix. During training, only the non-zero entries are used, making the iterations less time consuming. The training objective is then to learn word representations such that their dot product equals the logarithm of the words’ frequency of co-occurrence.

BERT. Bidirectional Encoder Representations from Transformers (BERT), as the name suggests, is a bidirectional Transformer [156] based language model. BERT learns word representations by two objectives – masked language model and next sentence prediction. In the masked language model objective, about 15% of the words in a sequence are masked or hidden. The model then predicts these masked words using only their positional information (for e.g., the index of the word in the sequence). For the next sentence prediction training objective, the input to the model is a pair of sentences. The model then classifies whether the second sentence is the contextual next sentence to the first. BERT uses three special tokens for these tasks – (1) [CLS], the first token of every input sequence, (2) [SEP], a token separating two sentences in an input sequence, and (3) [MASK], to represent masked words in a sequence. In order to cover a wide range of out-of-vocabulary words, BERT uses word-piece tokenization – for example, diving the word “playing” into two tokens “play” and “##ing”. The authors publish two pretrained models of BERT – BERT-BASE (12 transformer blocks, 768 hidden layers, 12 attention heads) and BERT-LARGE (24 transformer blocks, 1024 hidden layers, 16 attention heads). The output from any of the layers may be used as word embeddings. However, as identified by the authors, summing the outputs of the last 4 layers results in the most efficient word representations.

2.3.4 Multimodal Data Sets

Research presented in this dissertation primarily involve two data modalities: text and image. The publicly available multimodal data sets that were leveraged are being discussed here. Novel data sets that were created during the course of the dissertation will be discussed in respective chapters.

2.3.4.1 Object Detection

ImageNet. ImageNet [39] is a popular data set of manually annotated images that paved the way for computer vision research, especially those involving deep neural networks trained on large amounts of data. ImageNet contains 14 million images, out of which 1 million images have annotated boxes around detected objects (popularly known as “bounding boxes”). The data set is mapped to the WordNet noun hierarchy with thousands of images per node. Following the WordNet ontology, ImageNet images fall into 21 thousand synsets (conceptually related words or synonyms).

The ImageNet Large Scale Visual recognition Challenge (ILSVRC) [135] played a pivotal role in encouraging research in computer vision. This annual competition consisted of three tasks – image classification, single-object localization, and object detection. We have leveraged the data from the object detection challenge (popularly referred to as “ILSVRC DET”), which consists of 1 million images and 1000 object classes. Our utilization of this dataset is detailed in Section 4.4.

MSCOCO. While ImageNet images are most often isolated images with closeups of single objects, the Microsoft Common Objects in Context $MSCOCO$ [89] data set contains 330 thousand images of everyday scenes with multiple objects within a single image. MSCOCO images provide contextual information required to better understand images. The drawback of MSCOCO as an object detection data set is that it only covers 80 object classes.

2.3.4.2 Image Captioning

Pascal Sentence Data Set. The Pascal Sentence Data Set [126] is on one the earliest image captioning data sets. It consists of 1000 images with 5 crowd-sourced captions per image. The small number of images in the dataset is a deterrant for modern deep learning based algorithms which rely on huge amounts of training data.

MSCOCO. Although MSCOCO is somewhat deficient for object detection, it is a valuable resource for image captioning. Each image in MSCOCO is accompanied with 5 crowd-sourced captions.

SBU Captioned Photo Data Set. The SBU (Stony Brook University) captioning data set contains 1 million images from Flickr with visually relevant captions. SBU offers original Flickr captions posted by users, whereas MSCOCO and Pascal offer crowd-sourced captions. The nature of the latter captions are hence strictly descriptive with mentions of objects seen in the images. SBU captions on the other hand also capture abstractions and sentiments related to the images.

2.3.4.3 Social Media

The prominent social media platforms that offer multimodal data are Flickr, Reddit, and Instagram. Various data sets have been curated from these sources, such as Flick30K Entities [121] (a data set of Flickr images with textual phrases mapped to image regions), InstaPIC [119] (a data set Instagram images, captions, and hashtags for personalized image captioning). In our research we leverage multimodal data from Reddit which offers a rich source of user comments along with each image-caption pair. This novel Reddit data set will be discussed in Chapter 6.

2.4 Evaluation Metrics

There are several established evaluation metrics for evaluating IR and NLP tasks . A non-exhaustive list of these metrics are being discussed here. We have used some of these metrics to evaluate (parts of) research presented in this dissertation. New metrics that were introduced in the course of the dissertation have been discussed in respective chapters.

Precision, Recall, F-score. Precision, Recall, and F1 are rudimentary evaluation measures used in IR, and machine learning.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{2.15}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2.16}$$

F-score, which is a measure of accuracy, is a harmonic mean between precision and recall.

$$F_b = (1 + b^2)\frac{Precision * Recall}{(b^2 * Precision) + Recall} \tag{2.17}$$

where b is a weight for recall. The most common variant of F-score is F1 ($b = 1$).

These metrics are insufficient when it comes to evaluating results of NLP tasks which involve long sequence of text outputs like machine translation, text summarization, question-answering, image caption generation, etc. Hence newer metrics like the ones discussed below were defined.

BLEU. The Bilingual Evaluation Understudy (BLEU) [118], originally developed to evaluate machine translation, measures the n-gram precision between reference (human) and candidate (machine-generated) translations. BLEU employs a modified n-gram precision – how many of the words (or n-grams) in the candidate sentence appears in the reference sentence, considering candidate words only as many times as they appear in the reference. All n-gram precision scores (p_n) are then combines as follows:

$$n\text{-}gram\ Precision = exp(\sum_{n=1}^{N} w_n log p_n), \text{ where } w_n = 1/n \tag{2.18}$$

The other component of BLEU is the brevity penalty (BP) – to discount the score of very short translations. Typically, there are multiple reference translations against which the accuracy of a candidate translation is measured. The average length of all references are considered into the calculation of BP.

$$BP = \begin{cases} 1, & \text{if } c > r \\ exp(1 - \frac{r}{c}) & \text{otherwise} \end{cases} \tag{2.19}$$

Here c is the length of candidate, and r is the average length of references. With short candidates, the ratio r/c is high, heading to a smaller BP. The final BLEU score is obtained by multiplying $n\text{-}gram\ Precision$ with BP.

BLEU scores may consider unigram, bigram, trigram and 4-gram precision, called BLEU-1, BLEU-2, BLEU-3, and BLEU-4 respectively. Although BLUE score is relatively simple, there are major shortcomings. Firstly, it does not account for semantic meaning (for e.g. synonyms of reference words) and syntactic structure (grammatical order of words) of the generated text [149]. It has also been found that BLEU does not correlate well to human judgements [130, 115].

ROUGE. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [87] reverses the BLEU method and measures how many n-grams in the reference sequence are present in the candidate sequence. It has been popularly used for evaluating text summarization, and mitigates some of the issues with BLEU. ROGUE has a few variants depending on the text feature used for calculation – (1) ROUGE-N is based on n-gram matches, (2) ROUGE-L depends on the longest common sub-sequence (LCS) match, (3) ROUGE-W takes into account the lengths of consecutive sub-sequence matches as weights, (4) ROUGE-S considers skip bigrams (words in the original order, separated by an arbitrary gaps). ROUGE-L/W/S use F-score, instead of only recall. ROUGE-1 (calculation based on matching unigrams) and ROUGE-L are most commonly used.

For ROUGE-L, the precision and recall are calculated as follows:

$$Precision_{LCS} = \frac{LCS(c,r)}{m} \quad \text{and} \quad Recall_{LCS} = \frac{LCS(c,r)}{n} \tag{2.20}$$

where c and r are candidate and reference sentences of length m and n respectively. ROGUE-L is then the corresponding F-score.

METEOR. While BLEU an ROUGE are measures of precision and recall respectively, Metric for Evaluation of Translation with Explicit Ordering (METEOR) [9] calculates an weighed F-score, with a penalty for incorrect word order. Unlike the previous metrics, METEOR also considers synonyms from WordNet for better matching. METEOR has been claimed to correlated better with human judgements.

To calculate the METEOR score, the largest subset of matches is gathered between candidate and reference sentences. This is done by finding exact matches, followed by matches after stemming (stripping the last few letters of words), and matches with WordNet synonyms. If M is the total number of unigram matches, precision $P = M/c$ and recall $R = M/r$ for candidate and reference sentence of length c and r respectively. The F-score if then calculated as a harmonic mean between P and R.

$$F_{mean} = \frac{10PR}{P + 9R} \tag{2.21}$$

with R weighted 9 times more than P.

METEOR accounts for the order of words by introducing a penalty score. The candidate sentence is divided into fewest possible chunks (a set of unigrams which are adjacent both in the candidate and the reference). Longer correctly generated sub-sequences lead to fewer chunks and is therefore less penalized.

$$Penalty = 0.5 * \frac{C}{M}^{3} \tag{2.22}$$

where C is the number of chunks, and M is the total number of unigram matches.

Finally, the METEOR score is calculated as $F_{mean} * (1 - Penalty)$.

CIDEr. The Consensus-based Image Description Evaluation (CIDEr) [158] was proposed to evaluate novel machine generated image captions. CIDEr calculates the consensus between candidate and reference sentences, weighted by the TF-IDF scores of each word against the entire corpus.

Each candidate and reference sentence is first stemmed, and represented with a set of n-grams ($n = 1, 2, 3, 4$). This is followed by calculation of co-occurrences of n-grams in candidate and reference sentences. Finally, cosine similarities of the candidate and reference n-grams are calculated, weighted by their TF-IDF scores over the entire corpus. The intuition behind introducing TF-IDF based weighting is – n-grams that appear in image captions frequently in the corpus are not informative for a particular image.

The CIDEr scores for n-grams between candidate and reference are calculated as follows. Note that CIDEr score is calculated between a set of candidate and reference sentences. For simplicity, we show the equations for one candidate sentence (c) and one reference sentence(r).

$$CIDEr_n(c, r) = \frac{1}{M} \sum_i \frac{\mathbf{g^n(c).g^n(r_i)}}{\|\mathbf{g^n(c)}\|.\|\mathbf{g^n(r_i)}\|} \tag{2.23}$$

where $\mathbf{g^n(c)}$ is a vector of TF-IDF scores for the n-grams in c with magnitude $\|\mathbf{g^n(r_i)}\|$. Similarly for $\mathbf{g^n(r_i)}$. The CIDEr score between c and r is then calculated as:

$$CIDEr(c, r) = \sum_{n=1}^{N} w_n CIDEr_n(c, r) \tag{2.24}$$

where $w_n = \frac{1}{N}$ with $N = 4$ has been found to be the best setting empirically.

SPICE. Semantic Propositional Image Caption Evaluation (SPICE) [7], also developed to evaluate image captions, operates on scene graphs generated from reference and candidate captions. A scene graph is a semantic representation or a parse tree of an image caption with nodes for object classes C (e.g., ball, dog, Frisbee), relations R (e.g., through, in, with), and attributes A (e.g., red, wet, green). Formally, a candidate caption c is represented as the following scene graph $G(c)$:

$$G(c) = \langle O(c), E(c), K(c) \rangle \tag{2.25}$$

where $O(c) \subseteq C$ is a set of object mentions, $E(c) \subseteq O(c) \times R \times O(c)$ is a set of edges representing object-object relations, and $K(c) \subseteq O(c) \times A$ is a set of object attributes.

SPICE score is calculated as the F1-score based on matched tuples between scene graphs of candidate and reference captions. Like METEOR, SPICE also considers synonyms from WordNet. It has been shown [7] that SPICE correlate more with human judgements than METEOR and CIDEr.

Although metrics introduced in this section, especially BLEU, ROUGE, METEOR, CIDEr, and SPICE, have been developed as yardsticks of machine generation problems, they can be used for comparing any two blocks of text, with varying efficiency. In addition to these metrics, cosine similarity between respective embedding vectors is often used as a measure of semantic relatedness between two text snippets.

For some of the problems presented in this dissertation, the metrics discussed in this section were found to be unsuitable. In such cases we have defined newer evaluation metrics. These have been discussed in respective chapters where applicable.

2.5 Crowd-sourcing

The research questions addressed in this dissertation are often of a subjective nature. Quantitative evaluation with metrics discussed in Section 2.4 are often inefficient in judging the quality of our proposed models. In such scenarios, we conduct meticulous user evaluations on crowd-sourcing platforms. In this section, we introduce the basic working concepts for crowd-sourcing.

Crowd-sourcing is the process where a task, decomposed into 'micro-tasks' (simpler tasks which are parallelizable and results combined), is solved by a network or 'crowd' of people. This 'crowd' of people, often referred to as annotators, may be geographically dispersed and may have varied levels of expertise on different topics. Inter-annotator agreement is often used as an estimate to determine the subjectivity or difficulty level of a task. The collective contribution of the annotators lead to efficient solving of time consuming and labor intensive problems such as collection of labelled data in NLP [144] and Computer Vision [39]. Popular NLP usages of crowd-sourcing include translation, summarization, word-sense disambiguation, and sentiment recognition. In Computer Vision, crowd-sourcing has been primarily used for labelling images and image segmentation.

There are different types of crowd-sourcing tasks based on the way in which the annotators are motivated to complete their micro-tasks. The motivation may be in the form of economic incentives (paid crowd-sourcing), enjoyment (popularly called 'Games with a Purpose'), or altruism (crowd-sourcing for common good such as disaster response). Crowd-sourcing has revolutionized research in Artificial Intelligence (both in the industry and academia) by making huge amounts of data available in a short time and relatively low costs. In our research work, we have deployed paid crowd-sourcing to evaluate the quality of results produced by our proposed models.

Crowd-sourcing Platforms. The most commonly used platforms for paid crowd-sourcing is Amazon Mechanical Turk and Appen (formerly called Figure8 and Crowdflower). Through these platforms, task owners (or requesters) post Human Intelligence Tasks (HITs) which are made available to a big population of crowd-workers or annotators. The annotators complete their assigned micro-tasks (for example answering 10 questions from a pool of 1000 questions) in return for a financial compensation.

Quality Controls. One of the challenges of crowd-sourcing is to control the quality of the annotations being collected. In practice, quality control is done through various checks. The requester often curates a set of questions with correct answers, referred to as gold-standard questions or 'honeypot' questions which the annotators must answer correctly in order to be eligible to complete the task. Over time, annotators on a certain platform receive reputation scores. Requesters may limit their 'crowd' to only 'highly reputed' annotators in order to ensure quality annotations.

Compensation. Ethical compensation practice is to offer annotators the standard US/EU minimum wage, which is around 10-15$ per hour. In order to estimate the time required for a micro-task, the requester completes (a portion of) the task themselves and divides the time taken with the number of constituent micro-tasks.

Result Aggregation. Based on the task, the contributions by the annotators are combined in various ways to solve the global problem. Popular methods for result aggregation are – taking the average of annotators' scores, or considering the majority score from odd number of annotations.

CHAPTER 3

Commonsense Knowledge for Visual Search

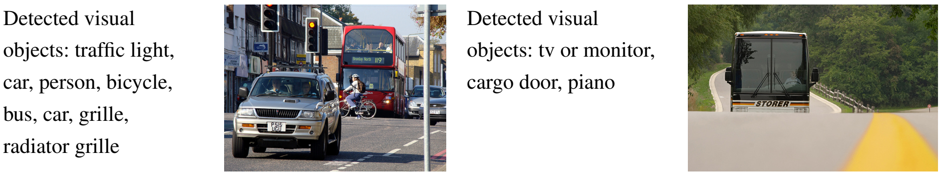
Contents

WITH the rise in popularity of social media, images accompanied by contextual text form a huge section of the web. However, search and retrieval of documents are still largely dependent on solely textual cues. Although visual cues have started to gain focus, the imperfection in object/scene detection do not lead to significantly improved results. We hypothesize that the use of commonsense knowledge on query terms can significantly aid in retrieval of documents with images. To this end we deploy three different modalities - text, visual cues, and commonsense knowledge pertaining to the query - as a recipe for efficient search and retrieval.

3.1 Introduction

Motivation: Image retrieval by querying visual contents has been on the agenda of the database, information retrieval, multimedia, and computer vision communities for decades [98, 35]. Search engines like Baidu, Bing or Google perform reasonably well on this task, but crucially rely on textual cues that accompany an image: tags, caption, URL string, adjacent text etc.

In recent years, deep learning has led to a boost in the quality of visual object recognition in images with fine-grained object labels [142, 80, 112]. Methods like $LSDA$ [65] are trained on more than 7,000 classes of $ImageNet$ [38] (which are mostly leaf-level synsets of WordNet [109]), and annotate newly seen images with class labels for bounding boxes of objects. For the image in Figure 3.1a, for example, object labels *traffic light, car, person, bicycle* and *bus* have been recognized making it easily retrievable for queries with these concepts. However, these labels come with uncertainty. For the image in Figure 3.1b, there is much higher noise in its visual object labels; so querying by visual labels would not work here.



Detected visual objects: traffic light, car, person, bicycle, bus, car, grille, radiator grille

Detected visual objects: tv or monitor, cargo door, piano

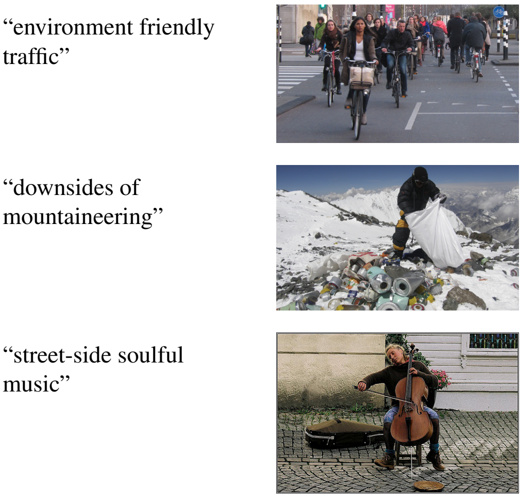
(a) Good object detection

(b) Poor object detection

Figure 3.1: Example cases where visual object detection may (a) and may not (b) aid in search and retrieval.

Opportunity and Challenge: These limitations of text-based search, on one hand, and visual-object search, on the other hand, suggest combining the cues from text and vision for more effective retrieval. Although each side of this combined feature space is incomplete and noisy, the hope is that the combination can improve retrieval quality.

Unfortunately, images that show more sophisticated scenes, or emotions evoked on the viewer are still out of reach. Figure 3.2 shows three examples, along with query formulations that would likely consider these sample images as relevant results. These answers would best be retrieved by queries with abstract words (e.g. "environment friendly") or activity words (e.g. "traffic") rather than words that directly correspond to visual objects (e.g. "car" or "bike"). So there is a vocabulary gap, or even concept mismatch, between what users want and express in queries and the visual and textual cues that come directly with an image. This is the key problem addressed in this paper.



"environment friendly traffic"

"downsides of mountaineering"

"street-side soulful music"

Figure 3.2: Sample queries containing abstract concepts and expected results of image retrieval.

Approach and Contribution: To bridge the concepts and vocabulary between user queries and image features, we propose an approach that harnesses commonsense knowledge (CSK). Recent advances in automatic knowledge acquisition have produced large collections of CSK: physical (e.g. color or shape) as well as abstract (e.g. abilities) properties of everyday objects (e.g. bike, bird, sofa, etc.) [152], subclass and part-whole relations between objects [153], activities and their participants [151], and more. This kind of knowledge allows us to establish relationships between our example queries and observable objects or activities in the image. For example, the following CSK triples establish relationships between *'backpack'*, *'tourist'* and *'travel map'*: `(backpacks,` `are carried by, tourists)`, `(tourists, use, travel maps)`. This allows for retrieval of images with generic queries like *"travel with backpack"*.

This idea is worked out into a *query expansion model* where we leverage a CSK knowledge base for automatically generating additional query words. Our model unifies three kinds of features: *textual features* from the page context of an image, *visual features* obtained from recognizing fine-grained object classes in an image, and *CSK features* in the form of additional properties of the concepts referred to by query words. The weighing of the different features is crucial for query-result ranking. To this end, we have devised a method based on statistical language models [182].

The paper's contribution can be characterized as follows. We present the first model for incorporating CSK into image retrieval. We develop a full-fledged system architecture for this purpose, along with a query processor and an answer-ranking component. Our system *Know2Look*, uses commonsense *know*ledge to *look for* images relevant to a query by *looking at* the components of the images in greater detail. We further discuss experiments that compare our approach to state-of-the-art image search in various configurations. Our approach substantially improves the query result quality.

3.2 Multimodal document retrieval

Adjoining text of images may or may not explicitly annotate their visual contents. Search engines relying on only textual matches ignore information which may be solely available in the visual cues. Moreover, the intuition behind using CSK is that humans innately interpolate visual or textual information with associated latent knowledge for analysis and understanding. Hence we believe that leveraging CSK in addition to textual and visual information would take results closer to human preferences. In order to use such background knowledge, curating a CSK knowledge base is of primary importance. Since automatic acquisition of canonicalized CSK from the web can be costly, we conjecture that noisy subject-predicate-object (SPO) triples extracted through Open Information Extraction [10] may be used as CSK. We hypothesize that the combination of the noisy ingredients – CSK, object-classes, and textual descriptions – would create an ensemble effect facilitating efficient search and retrieval. We describe the components of our architecture in the following sections.

3.2.1 Data, Knowledge and Features

We consider a document x from a collection X with two kinds of features:

- **Visual features** xv_j**:** labels of object classes recognized in the image, including their hypernyms (e.g., king cobra, cobra, snake).
- **Textual features** xx_j**:** words that occur in the text that accompanies the image, for example image caption.

We assume that the two kinds of features can be combined into a single feature vector $x = \langle x_1 \dots x_M \rangle$ with hyper-parameters α_v and α_x to weigh visual vs. textual features.

CSK is denoted by a set Y of triples $y_k (k = 1..j)$ with components ys_k, yp_k, yo_k (s - subject, p - predicate, o - object). Each component consists of one or more words. This yields a feature vector $y_k j (j = 1..M)$ for the triple y_k.

3.2.2 Language Models for Ranking

We study a variety of query-likelihood language models (LM) for ranking documents x with regard to a given query q. We assume that a query is simply a set of keywords $q_i (i = 1..L)$. In the following we formulate equations for unigram LMs, which can be simply extended to bigram LMs by using word pairs instead of single ones.

Basic LM:

$$P_{basic}[q|x] = \prod_i P[q_i|x] \tag{3.1}$$

where we set the weight of word q_i in x as follows:

$$P[q_i|x] = \alpha_x P[q_i|xx_j] P[xx_j|x] + \alpha_v P[q_i|xv_j] P[xv_j|x] \tag{3.2}$$

Here, xx_j and xv_j are unigrams in the textual or visual components of a document; α_x and α_v are hyper-parameters to weigh the textual and visual features respectively.

Smoothed LM:

$$P_{smoothed}[q|x] = \alpha P_{basic}[q|x] + (1 - \alpha) P[q|B] \tag{3.3}$$

where B is a background corpus model and $P[q|B] = \prod_i P[q_i|B]$. We use Flickr tags from the YFCC100M dataset [154] along with their frequency of occurrences as a background corpus.

Commonsense-aware LM (a translation LM):

$$P_{CS}[q|x] = \prod_i \left[\frac{\sum_k P[q_i|y_k] P[y_k|x]}{|k|} \right] \tag{3.4}$$

The summation ranges over all y_k that can bridge the query vocabulary with the image-feature vocabulary; so both of the probabilities $P[q_i|y_k]$ and $P[y_k|x]$ must be non-zero. For example, when the query asks for "electric car" and an image has features "vehicle" (visual) and "energy saving" (textual), triples such as `(car, is a type of, vehicle)` and `(electric engine, saves, energy)` would have this property. That is, we consider only commonsense triples that overlap with both the query and the image features.

The probabilities $P[q_i|y_k]$ and $P[y_k|x]$ are estimated based on the word-wise overlap between q_i and y_k and y_k and x, respectively. They also consider the confidence of the words in y_k and x.

Mixture LM (the final ranking LM):
Since a document x can capture a query term or its commonsense expansion, we formulate a mixture model for the ranking of a document with respect to a query:

$$P[q|x] = \beta_{CS} P_{CS}[q|x] + (1 - \beta_{CS}) P_{smoothed}[q|x] \tag{3.5}$$

where β_{CS} is a hyper-parameter weighing the commonsense features of the expanded query.

The formulas and descriptions of these query-likelihood language models can be seen at a glance in Table 3.1.

	Formula	Description
Basic LM	$P_{basic}[q\mid x] = \prod_i P[q_i\mid x];$	A unigram/bigram LM described by the probability of generation of a query q from a document x. The weight of the i^{th} word in q is given by $P[q_i\mid x]$. The product over all words of the query ensures a conjunctive query.
	$P[q_i\mid x] = \frac{\alpha_x}{\lvert j\rvert} \sum_j sim(q_i, xx_j)P[xx_j\mid x] + \frac{\alpha_v}{\lvert l\rvert} \sum_l sim(q_i, xv_l)P[xv_l\mid x];$	A word in the query may match with the textual or visual features of a document weighted by α_x and α_v, and normalised with number of matches $\lvert j\rvert$ and $\lvert l\rvert$ respectively.
Smoothed LM	$P_{smoothed}[q\mid x] = \alpha P_{basic}[q\mid x] + (1-\alpha)P[q\mid B];$ $P[q\mid B] = \prod_i P[q_i\mid B]$	The Basic LM after smoothing on background corpus B. The relative frequency of q_i in B $(P[q_i\mid B])$ is used for smoothing the LM.
Commonsense-aware LM	$P_{CS}[q\mid x] = \prod_i \left[\frac{\sum_k P[q_i\mid y_k]P[y_k\mid x]}{\lvert k\rvert}\right];$	A translation LM describing the probability of generation of a query from the k commonsense knowledge triples y_k. The summation over k includes all triples bridging the gap between the query vocabulary and the document vocabulary; it is normalized by the total number of such triples.
	$P[q_i\mid y_k] = \sum_j sim(q_i, y_{kj})$	The probability that the query word q_i has been generated from the CSK triple y_k is the sum of similarity scores between the two words/phrases, normalised by the number of words/phrases $(\lvert j\rvert)$ in the CSK triples.
Mixture LM	$P[q\mid x] = \beta_{CS}P_{CS}[q\mid x] + (1-\beta)P_{smoothed}[q\mid x]$	Combination of the weighted Commonsense-aware LM and Smoothed LM for ranking a document x for a query q.

Table 3.1: Mathematical formulations of Language Models for Ranking

3.2.3 Feature Weights

By casting all features into word-level unigrams, we have a unified feature space with hyper-parameters (α_x, α_v, and β_{CS}). The hyper-parameters are manually chosen. They may optionally be tuned by withheld data and using cross-validation with some performance measure (e.g. NDCG) to optimize. For weights of visual object class xv_j of document x, we consider the *confidence score* from $LSDA$ [65]. We extend these object classes with their hypernyms from WordNet which are set to the same confidence as their detected hyponyms[1]. Although not in common parlance this kind of expansion can also be considered as CSK. We define the weight for a textual unigram xx_j as its informativeness – the inverse document frequency with respect to a background corpus (Flickr tags with frequencies).

The words in a CSK triple y_k have non-uniform weights proportional to their similarity with the query words, their *idf* with respect to a background corpus, and the salience of their position – boosting the weight of words in s and o components of y. The function computing similarity between two unigrams favors exact matches to partial matches.

Tables 3.2, 3.3, 3.4 show details of the feature weights, hyper-parameter definitions, and function definitions respectively.

	Formula	Description
Textual feature weight	$P[xx_j\|x] = \frac{idf(xx_j)}{\sum_\nu idf(xx_\nu)}$	The informativeness or weight of a word/phrase xx_j in a document is captured by calculating it's idf in a large background corpus ν.
Visual feature weight	$P[xv_j\|x] = \frac{conf(xv_j)}{\sum_\nu conf(xv_\nu)} \times \frac{idf(xv_j)}{\sum_\nu idf(xv_\nu)}$	The weight of a object class xv_j in a document is calculated by the product of it's confidence (from LSDA) and it's informativeness.
CSK feature weight	$P[y_k\|x] = \frac{\sum_i \sum_j sim(y_{kj}, x_i) sal(y_{kj}) inf(y_{kj})}{\|i\|\|j\|}$	The relevance of a commonsense triple y to a document is decided by the similarity of its words/phrases y_k to the features of the document, the salience (or importance) of the match, and the informativeness of the word/phrase.

Table 3.2: Mathematical formulations of Feature Weights

Hyper-parameter	Description
α	Weight of the basic document features; $(1 - \alpha)$ being the weight for smoothing.
α_x	Weight associated with the textual features of a document.
α_v	Weight associated with the visual features of a document.
β_{CS}	Weight pertaining to the commonsense knowledge features of an expanded document.

Table 3.3: Definition of Hyper-parameters

[1] a hyponym is a word or phrase whose semantic field is included within that of another word, its hypernym

	Function	Description
Confidence	$conf(w)$	A score output by the LSDA to depict the confidence of detection of an object class. The hypernyms of the detected visual object classes are assigned the same confidence score.
Informative-ness	$inf(w) = idf_B(w)$	We measure informative-ness of a word by its idf value in a larger corpus, such that common terms are penalised.
Similarity	$sim(w_1, w_2) = \dfrac{\lvert substring(w_1, w_2)\rvert}{max[length(w_1), length(w_2)]}$	This function calculates the amount of string overlap between w_1 and w_2.
Salience	$sal(w) = \lambda_s \quad$ if $w \in subject$ $= \lambda_p \quad$ if $w \in predicate$ $= \lambda_o \quad$ if $w \in object$ where $t_{csk} = \langle subject, predicate, object\rangle$	The importance of the string match position in a commonsense knowledge triple t_{csk} is captured by this function. Intuitively, the textual features in the subject and the object are more important that those in the predicate. Therefor we assign $\lambda_s = \lambda_o > \lambda_p$ and $\lambda_s + \lambda_p + \lambda_o = 1$

Table 3.4: Function definitions

3.2.4 Example

Query string: *travel with backpack*

Commonsense triples to expand query:

$t1$:`(tourists, use, travel maps)`

$t2$:`(tourists, carry, backpacks)`

$t3$:`(backpack, is a type of, bag)`

Say we have a document x with features:

Textual - “A tourist reading a map by the road.”

Visual - person, bag, bottle, bus

The query will now successfully retrieve the above document, whereas it would have been missed by text-only systems.

3.3 Datasets

For the purpose of demonstration we choose a topical domain – *Tourism*. Our CSK knowledge base and image dataset obey this constraint.

CSK acquisition through OpenIE:

Section 2.2.2 provides a brief primer on OpenIE. OpenIE extracts *(subject, predicate, object)* triples from a sizable text corpus. We consider a slice of Wikipedia pertaining to the domain *tourism* as the text corpus to extract CSK from. Nouns from the Wikipedia article titled ‘Tourism’(seed document) constitute our basic language model. We collect articles by traversing the Wiki Category hierarchy tree while pruning out those with substantial topic drift. The Jaccard Distance (Equation 3.6) of a document from the seed document is used as a metric for pruning.

$$JaccardDistance = 1 - WeightedJaccardSimilarity \tag{3.6}$$

where,

$$WeightedJaccardSimilarity = \frac{\Sigma_n min[f(d_i, w_n), f(D, w_n)]}{\Sigma_n max[f(d_i, w_n), f(D, w_n)]} \tag{3.7}$$

In Equation 3.7, acquired Wikipedia articles d_i are compared to the seed document D; $f(d', w)$ is the frequency of occurrence of word w in document d'. For simplicity only articles with Jaccard distance of 1 from the seed document are pruned out. The corpus of domain-specific pages thus collected constitute ~5000 Wikipedia articles.

The OpenIE tool ReVerb [46] run against our corpus produces around 1 million noisy SPO triples. After filtering with our basic language model we have ~22,000 moderately clean assertions.

Image Dataset: For the purpose of experiments we construct our own image dataset. ~50,000 images with descriptions are collected from the following datasets: Flickr30k [181], *Pascal Sentence Dataset* [126], *SBU Captions Dataset* Captioned Photo Dataset [117], and *MSCOCO*O [89]. The images are collected by comparing their textual descriptions with our basic language model for *Tourism*. An existing object detection algorithm – *LSDA* [65] – is used for object detection in the images. The detected object classes are based on the 7000 leaf nodes of ImageNet [38]. We also expand these classes by adding their super-classes or hypernyms with the same confidence score.

Query Benchmark: We construct a benchmark of 20 queries from co-occurring Flickr tags from the YFCC100M dataset [154]. This benchmark is shown in Table 3.5. Each query consists of two keywords that have appeared together with high frequency as user tags in Flickr images.

aircraft international	diesel transport
airport vehicle	dog park
backpack travel	fish market
ball park	housing town
bench high	lamp home
bicycle road	old clock
bicycle trip	road signal
bird park	table home
boat tour	tourist bus
bridge road	van road

Table 3.5: Query Benchmark for evaluation

3.4 Experiments

Baseline Google search results on our image dataset form the baseline for the evaluation of *Know2Look*. We consider the results in two settings – search only on original image caption (Vanilla Google), and on image captions along with detected object classes (Extended Google). The later is done to aid Google in its search by providing additional visual cues. We exploit the domain restriction facility of Google search (*query string site:domain name*) to get Google search results explicitly on our dataset.

Know2Look In addition to the setup for Extended Google, *Know2Look* also performs query expansion with CSK. In most cases we win over the baseline since CSK captures additional concepts related to query terms enhancing latent information that may be present in the images. We consider the top 10 retrieval results of the two baselines and *Know2Look* for the 20 queries in our query benchmark[2]. We compare the three systems by Precision@10. Table 3.6 shows the values of Precision@10 averaged over 20 queries for each of the three systems – *Know2Look* performs better than the baselines.

	Average Precision@10
Vanilla Google	0.47
Extended Google	0.64
Know2Look	0.85

Table 3.6: Comparison of *Know2Look* with baselines

In this work we proposed the incorporation of commonsense knowledge for image retrieval. Our architecture, *Know2Look*, expands queries by related commonsense knowledge and retrieves images based on their visual and textual contents. By utilizing the visual and commonsense modalities we make search results more appealing to the humans than traditional text-only approaches. We support our claim by comparing *Know2Look* to Google search on our image data set. The proposed concept can be easily extrapolated to document retrieval. Moreover, in addition to using noisy OpenIE triples as commonsense knowledge, existing commonsense knowledge bases can also be leveraged.

In the next chapter we will discuss how to refine automatically detected image tags in order to eliminate some of the noise in the visual modality that negatively affects image retrieval.

[2] http://mpi-inf.mpg.de/~sreyasi/queries/evaluation.html

CHAPTER 4

Visual and Semantic Image Label Refinement

Contents

THE social media explosion has populated the Internet with a wealth of images. There are two existing paradigms for image retrieval: 1) Content Based Image Retreival (CBIR), which has traditionally used visual features for similarity search (e.g., SIFT features), and 2) Tag Based Image Retrieval (TBIR), which has relied on user tagging (e.g., Flickr tags). CBIR now gains semantic expressiveness by advances in deep-learning-based detection of visual labels. TBIR benefits from query-and-click logs to automatically infer more informative labels. However, learning-based tagging still yields noisy labels and is restricted to concrete objects, missing out on generalizations and abstractions. Click-based tagging is limited to terms that appear in the textual context of an image or in queries that lead to a click. This paper addresses the above limitations by semantically refining and expanding the labels suggested by learning-based object detection. We consider the semantic coherence between the labels for different objects, leverage lexical and commonsense knowledge, and cast the label assignment into a constrained optimization problem solved by an integer linear program. Experiments show that our method, called VIsual and Semantic Image-label Refinement (VISIR), improves the quality of the state-of-the-art visual labeling tools like $LSDA$ and $YOLO$.

(a)

(b)

Figure 4.1: Noisy and Incomplete Labels: a) from $LSDA$ [65] - *dog, Browning machine gun, greater kudu, bird* b) from flickr.com - *happiness*

4.1 Introduction

Motivation and Problem: The enormous growth of social media has populated the Internet with a wealth of images. On one hand, this makes image search easier, as there is redundancy for many keywords with informative text surrounding the images. On the other hand, it makes search harder, as there is a huge amount of visual contents that is hardly understood by the search engine. There are two paradigms for searching images: content-based image retrieval (CBIR) and tag-based image retrieval (TBIR).

CBIR finds images similar to a query image based on visual features that are used to represent an image. These features include color, shape, texture, SIFT descriptors etc. (e.g., [83, 36, 8]). Recent advances in deep-learning-based object detection have lifted this approach to a higher level, by assigning object labels to bounding boxes (e.g., [65, 128, 132, 163]). However, these labels are limited to concrete object classes (e.g., *truck, SUV, Toyota Yaris Hybrid 2016*, etc.), often trained (only) on (subsets of) the ca. 20,000 classes of ImageNet [39]. Thus, they miss out on generalizations (e.g., *vehicle*) and abstractions (e.g., *transportation, traffic jam, rush hour*). Fig 4.1 (a) shows the top-confidence visual labels by LSDA [65] for an example case of incorrect labels.

TBIR retrieves images by textual matches between user query and manually assigned image tags (e.g., from collaborative communities such as Flickr). While some of the semantic gap in CBIR is reduced in TBIR, the performance of TBIR often suffers from incomplete and ambiguous tags [61]. Figure 4.1 (b) illustrates this point: there is only a single tag *happiness* and none for the concrete objects in the image. For the big search engines, one way of overcoming this bottleneck is to exploit query-and-click logs (e.g., [33, 67, 169]). The query keyword(s) associated with a click can be treated as label(s) for the clicked image. However, this method crucially relies on the labels to appear in (sufficiently many) queries (or, traditionally, salient text surrounding the image). [85] gives a survey on TBIR and tag assignment and refinement.

Recently, the gap between the two image search paradigms is narrowing. TBIR-style tags inferred from query-and-click logs can be used to train a deep-learning network for more informative labels towards better CBIR. Also, crowdsourcing could be a way towards more semantic labels (e.g., [75]), for example, to capture human activities or emotions (e.g., [41, 55, 125, 74]). Nevertheless, there are still major shortcomings in the state-of-the-art.

This paper addresses the outlined limitations. The goal is to automatically annotate images with semantically informative tags, including generalizations and abstractions and also cleaning out noisy labels for concrete objects.

Approach and Contribution: We leverage state-of-the-art CBIR by considering the visual tags of an existing object detection tool ($LSDA$ [65] in our experiments) as a starting point. Note that there are multiple labels for each bounding box with varying confidence scores, and our goal is to compute the most informative labels for the entire image. We impose a constrained optimization on these initial labels, in order to enforce their semantic coherence. We also consider labels that are visually similar to the detected ones, to compensate for omissions. In addition, we utilize lexical and commonsense knowledge to generate candidate labels for generalizations (hypernyms from WordNet [110]) and abstractions (concepts from ConceptNet [147]). So we both refine and expand the initial labels. The joint inference on the entire label space is modeled as an optimization problem, solved by an integer linear program (viz. using the Gurobi ILP solver). Figure 4.2 shows examples for the input labels from the deep-learning-based visual object detection (left column) and the output labels that VISIR computes (right column). The labels from LSDA illustrate a clear semantic incoherence for these specific examples. VISIR labels are coherent, adds generalizations (in blue) and abstractions (in green). Incorrect labels are marked red. Although our work aligns with the existing TBIR research on social tagging and tag refinement, there are key differences.

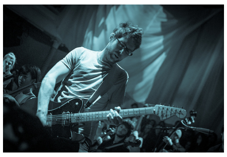
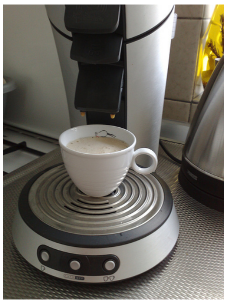
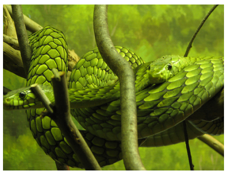
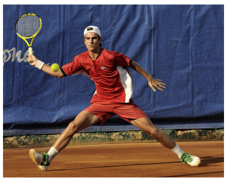
Input Image	LSDA Labels	VISIR Labels
	allosaurus loggerhead turtle person bird	person guitar *stringed instrument* *self-expression*
	bone china stove WC, loo cup or mug	food processor bowl cup or mug *utensil*
	cucumber snake green mamba	snake *reptile* *slithery* *poisonous*
	racket person bathing cap tennis ball head cabbage	tennis bat *individual* *play tennis* tennis ball

Figure 4.2: Images with labels from LSDA and VISIR

Granularity: Our starting point is labels for bounding boxes, whereas user-provided tags refer to an entire image.
Cardinality: The number of bounding boxes in one image can be quite large. Moreover, object detectors usually produce a long list of varying-confidence labels for each bounding box.
Noise: As a result, many of the visual candidate labels in our approach are of mixed quality, whereas traditional social tagging typically has few but trusted annotations per image.
For these factors, our notion of tag refinement is unlike the one in prior work. Therefore, we refer to our task as *Visual Tag Refinement*.

Visual Tag Refinement can be broken down into three sub-tasks, for which this paper provides effective solutions:

- elimination of incoherent tags[1] among the initial visual labels
- expansion of the tag space by adding visually similar tags missed by object detectors, and adding candidate tags for generalization and abstraction
- joint inference on the enriched tag space, by integer linear program

4.2 Related Work

Social Tagging. TBIR has its origin in community-based social tagging of images (e.g., Flickr), web pages or publications (e.g., Bibsonomy). Crowdsourcing to compile large training collections can be seen as a variant of this kind of user-provided tagging. There is ample research in this area [84, 103, 53], especially on learning tag recommendations. Our task of visual tag refinement differs from social tagging substantially. The CBIR-based tags that we start with, label individual objects instead of assigning tags to the overall image. Multiple candidate tags per bounding box also lead to a dense tag space in contrast to sparse user tags. Finally, the large number of varying-granularity and varying-confidence tags per image entails a much higher degree of noise in the label space, whereas social tags are usually considered trusted.

Automatic Image Tagging. Early work on this problem generated tags only for an entire image (or a single image region), but did so one class at a time (e.g., [3]). More recent methods support labeling multiple objects in the same image. One such approach, WSABIE [167], performs k-nearest-neighbor classification on embeddings of words and features to scale to many classes.

State-of-the-art work on object detection addresses both recognizing object bounding boxes and tagging them with their class labels. Such work makes heavy use of deep learning (especially CNNs). Prominent representatives are LSDA [65], and Faster R-CNN [132]. The latter improves the speed of object detection by incorporating a Region Proposal Network (RPN). Major emphasis in this line of Computer Vision work has been on coping with small, partly occluded and poorly illuminated objects. In contrast, the emphasis of VISIR is on the semantic coherence between objects and jointly modeling the uncertainty of candidate labels. Instead of speed, we optimize for higher

[1] "tag" and "label" are used interchangeably in the paper with the same meaning.

labeling quality. VISIR is agnostic of the underlying object detector; it is straightforward to plug in a different tool.

Context plays an important role in computer vision [42], and context-based object detectors were popular before the success of CNNs. These methods consider local context [16], global context [124] or their combination [179]. With the advent of CNNs, the focus shifted to representation learning and improving detection speed. However, contextual information is gaining renewed attention. The state-of-the-art method [165] is based on a CNN-RNN combination, where the recurrent neural network (RNN) captures label dependencies. $YOLO$ [128] unifies learning with global context into a single neural network for the entire image. It exploits a word tree derived from the WordNet whose leaf nodes appear in ImageNet. In our experiments, we use YOLO as a baseline for context-aware object detection. All learning-based methods crucially rely on extensive training data.

Tag Refinement. The problem of tag refinement aims at removing noisy tags from images while adding more relevant ones [85]. This line of work appears in the literature also as tag completion [170, 50] or image re-tagging [92, 94]. Most of this work uses only nouns as tags and disregards word ambiguity. Background knowledge, such as WordNet synonymy sets and other lexical relations, is rarely used. Word categories that are vital for denoting abstractions, namely, adjectives, verbs and verbal phrases, are out of scope. Moreover a common assumption is that visually similar images are semantically similar, meaning that they should have similar tags. This assumption is often invalid. This body of work employs a variety of methods, including metric learning [58], matrix completion [170, 50, 188], latent topic models [175], and more.

Commonsense Knowledge for Image Retrieval. The first work on image retrieval with commonsense knowledge (CSK) [95] exploited the Open Mind Commonsense Knowledgebase [143], a small knowledge base with simple properties of concepts for concept expansion and for activation spreading. Since then, much more comprehensive CSK knowledge bases have been constructed, most notably, ConceptNet [147] and WebChild [152]. However, such background knowledge has not been used by modern object detectors. A notable exception addresses emotions invoked by images, and tags objects with sentiment-bearing adjectives [21]. However, this work is limited to a small label space. A recent framework for image search [29] uses CSK extracted through OpenIE for query expansion.

4.3 Model and Methodology

We define the problem of *Visual Tag Refinement* as the tasks of:

- cleaning noisy object tags from low-level image features
- enriching existing detections by adding additional relevant tags
- abstracting from concrete objects towards a more conceptual space.

We first present a framework for the proposed problem, followed by the description of its individual components. We then present the optimization model to solve the problem.

4.3.1 Framework for Visual Tag Refinement

We consider an image x with multiple bounding boxes $x_1, x_2, ...x_k$. Each bounding box x_i has labels for detected physical objects along with detection confidence scores. The values of these labels and scores are outputs of an off-the-shelf object detection tool, e.g., LSDA [65]. We define three different label spaces, candidates from which would be associated either with bounding boxes of an image or would globally add semantics to it:

- A space of all possible object labels detectable by an underlying object detection tool is denoted by CL (for Concrete-object Labels). For example, the LSDA tool [65] uses ImageNet [39] object classes, which are also leaf-level synsets of WordNet [110]. There can be two sets of such classes for a bounding box x_i of image x: cl_i constitute those labels originally detected from low-level image features, cl'_i constitute undetected labels visually similar to those in cl_i.
- A space of extended labels is denoted by XL. For each image x and bounding box x_i, a subset from this space, xl_i, contains additional label candidates that generalize the classes in cl_i and cl'_i. For example, "ant" $\in cl_i \rightarrow$ "insect" $\in xl_i$. Adding generalized terms to the label space serves a dual purpose - overcoming the training bias of the object detection tools, and broadening the label space for greater web visibility. We discuss more on the issue of training bias in section 4.6.

  CL and XL contain labels signifying visual objects in an image. Hence we call the super set of such labels "visual labels" VL; $VL = CL \cup XL$.
- A space of abstract labels (utilities, emotions, themes) is denoted by AL. This constitutes abstract concepts associated with visual objects derived from commonsense knowledge bases. For example, *fragrant* $\in AL$ from the ConceptNet [96] clause `hasProperty(flower, fragrant)`.

An image x can hence be described by three sets of labels - a set of deep-learning based class labels $cl \cup cl' \in CL$, a set of extended labels $xl \in XL$, and a set of abstract labels $al \in AL$. Further, we define three different scores that act as edges between nodes of the above spaces:

- Visual Similarity $vsim(l_j, l_k)$ for $l_j, l_k \in CL$
- Semantic Relatedness $srel(l_j, l_k)$ for $l_j, l_k \in VL$
- Abstraction Confidence $aconf(l_j, al_n)$ for $al_n \in AL$ and $l_j \in VL$

  We present the visual tag refinement problem in terms of three sub-problems:

- The noisy tag problem - for each image x and bounding box x_i infer which of the labels in $cl_i \in CL$ should be accepted. We eliminate those labels which are not coherent with the other bounding box detections in the image. For example, in Figure 4.2 image 5 we eliminate the detection *cucumber* since it is not semantically related to the other labels *snake* and *green mamba*.
- The incomplete tag problem - for each image x and bounding box x_i infer which of the labels in $cl'_i \in CL$ and $xl_i \in XL$ should be additionally associated with the bounding box.
- The abstraction tag problem - infer which of $al \in AL$ should be globally associated with image x.

We solve these problems jointly and retain the most confident hypothesis for each bounding box relative to the others as well as a global hypothesis toward tag abstraction in an image. Hence, we predict a set of plausible labels $L_x \in CL \cup XL \cup AL$ for an image x.

4.3.2 Visual Similarity (or "Confusability")

Deep-learning based tools using low-level image features to predict the object classes can confuse one object to be another. We consider two labels to be visually similar if they occur as candidates in cl_i for the same bounding box x_i. We collect evidence of such visual similarity from low-level image features, in particular, from object detection results of LSDA [65]. We define the visual similarity between two labels l_j and l_k by a Jaccard-style similarity measure as shown in Equation 4.1. In this similarity measure, if labels l_j and l_k always appear together as candidates for the same bounding box, and never with any other labels, then they are considered highly visually similar, $vsim(l_j, l_k) = 1$. If labels l_j and l_k never appear together, one label is never confused by the tool to be another; in this case $vsim(l_j, l_k) = 0$, meaning l_j and l_k are not visually similar. Given that the initial object detections from low level image features are noisy in itself, this evidence would also contain noise. However, it is expected that the evidence will hold when it is computed over a large dataset.

$$vsim(l_j, l_k) = \frac{\sum\limits_{i:l_j,l_k \in cl_i} (conf_{BB}(x_i, l_j) + conf_{BB}(x_i, l_k))}{\sum\limits_{i:l_j \in cl_i} (conf_{BB}(x_i, l_j)) + \sum\limits_{i:l_k \in cl_i} (conf_{BB}(x_i, l_k))} \tag{4.1}$$

We can refer to this measure also as "confusability" since the object detection tool confuses one object to be another based on similar low-level visual features.

4.3.3 Semantic Relatedness

Semantic Relatedness between two concepts signifies their conceptual similarity. Our model uses this measure to establish the contextual coherence between labels of different bounding boxes. The relatedness between two labels l_j and l_k is defined as a weighted linear combination of their cosine similarity from word embeddings and their spatial co-location confidence.

$$srel(l_j, l_k) = \delta cosine(l_j, l_k) + (1 - \delta) coloc(l_j, l_k) \qquad for \quad l_j, l_k \in VL \tag{4.2}$$

Word Embeddings: To improve the contextual coherence between object labels in images, the context of words needs to be captured. We utilize vector space word embeddings for this purpose. A $Word2Vec$ [107] model is trained from manually annotated image descriptions from a large set of image captions, as described later in more details. The cosine similarity between two labels – $cosine(.,.)$ in Equation 4.2 – is calculated from their respective word vectors.

Spatial Co-location: Spatial relationships between concepts carry an important evidence of relatedness. For example, an "apple" and a "table" are related concepts since they occur in close spatial proximity. Similarly, a "tennis racket" and a "lemon" are unrelated. $coloc(.,.)$ in Equation 4.2 is a frequency-based co-location score mined from manual annotations of image labels.

4.3.4 Concept Generalization

A hypernym is a superordinate of a concept. In other words, a concept is an instantiation of its hypernym. For example, *fruit* is a hypernym for *apple*, i.e., *apple* IsA *fruit*. WordNet [110] provides a hierarchy of concepts and their hypernyms which we leverage to generalize our object classes. WordNet also reports different meanings (senses) of a concept; for example a *punching bag* is (*a person on whom another person vents their anger*) or (*an inflated ball or bag that is suspended and punched for training in boxing*), leading to very different hypernymy trees. For this reason, we map our object classes from ImageNet into their correct WordNet sense number, followed by traversing their hypernymy tree up to a certain level. This yields a cleaner generalization. Further more, to avoid exotic words among the hypernyms, we use their approximate Google result counts and prune out those below a threshold. Hence for the concept *ant* we retain the hypernym *insect* and prune the hypernym *hymenopteran*. Following this heuristics, we assign 1 to 3 hypernyms per object class.

4.3.5 Concept Abstraction

To introduce human factors like commonsense and subjective perception, we incorporate abstract words and phrases associated with visual concepts of an image. For example an *accordion* is “used to” *make music*. We consider two relations from ConceptNet 5 [96] for assigning the abstract labels - *usedFor*, and *hasProperty*. Some example of assigned abstract labels/phrases (in green) can be found in Figure 4.2. Abstract concepts which are assigned to images have high abstraction confidence. Abstraction confidence of a concept/phrase is defined as the joint semantic relatedness of the phrase and the refined visual labels of the image.

4.3.6 Tag Refinement Modeled as an ILP

We cast the multi-label visual tag refinement problem into an Integer Linear Program (ILP) optimization with the following definitions. We choose ILP as it is a very expressive framework for modeling constrained optimization (more powerful than probabilistic graphical models), and at the same time comes with very mature and efficient solvers like $Gurobi$ (`http://gurobi.com`). Some tools for probabilistic graphical models even use ILP for efficient MAP inference.

Given an image x, with bounding boxes $x_1, x_2, ...$, it has three sets of visual labels: cl_i (initial bounding box labels), cl'_i (labels visually similar to the original detections), and xl_i (hypernyms of labels in $cl_i \cup cl'_i$). The set $vl_i = cl_i \cup cl'_i \cup xl_i$ constitutes all visual labels which are candidates for bounding box x_i. The image would also be assigned abstract labels $al_1, al_2, ...$ globally. We thus introduce 0-1 decision variables:

$X_{ij} = 1$ if x_i should indeed have visual label vl_j, 0 otherwise
$Y_j = 1$ if x should indeed have abstract label al_j, 0 otherwise
$Z_{ijmk} = 1$ if $X_{ij} = 1$ and $X_{mk} = 1$, 0 otherwise
$W_{ijk} = 1$ if $X_{ij} = 1$ and $Y_k = 1$, 0 otherwise

Decision variables Z_{ijmk} and W_{ijk} emphasise pair-wise coherence between two visual labels and between a visual and an abstract label respectively.

Objective: Select labels for x and its bounding boxes which maximizes a weighted sum of evidence and coherence:

$$max\left[\alpha\sum_{i,j}\big(vconf(x_i,l_j)+\kappa gconf(x_i,l_j)\big)X_{ij}+ \right.$$
$$\beta\sum_{i,m}\sum_{\substack{l_j\in vl_i\\ l_k\in vl_m}} srel(l_j,l_k)Z_{ijmk}+$$
$$\left.\gamma\sum_{l_j\in VL}\sum_{k}aconf(l_j,al_k)\sum_{i}W_{ijk}\right] \tag{4.3}$$

with hyper-parameters α, β, γ, κ.

For each $l \in CL$, we define set $S(l) \subseteq CL$ of labels visually similar to l. $vsim(l, l') = 0$ if $l' \notin S(l)$. Recall the definition of $vsim(.,,)$ from Equation 4.1.

Visual Confidence, the confidence with which a visual label should be associated with an image is defined as:

$$vconf(x_i, l_j) = conf_{BB}(x_i, l_j) \text{ if } l_j \in cl_i \tag{4.4}$$
$$= \sum_{l\in cl_i} conf_{BB}(x_i, l)vsim(l, l_j) \text{ if } l_j \in cl'_i/cl_i \tag{4.5}$$

Here, a high confident original detection adds significant weight to the objective function, hence increasing the chances of its retention. Similarly, the weight of a label visually similar to multiple original labels is boosted. Also, labels visually similar to only one low confident original label is assigned less importance.

For $l \in CL$ we define a set $H(l) \in XL$ of hypernyms of l. The *Generalization Confidence* of a label l_j in bounding box x_i is defined in terms of the semantic relatedness between the label and its hypernym.

$$gconf(x_i, l_j) = \sum_{l:l_j\in H(l)} srel(l_j, l)\ if\ l_j \in xl_i \tag{4.6}$$
$$= 0\ if\ l_j \in \{cl_i \cup cl'_i\} \tag{4.7}$$

Abstraction Confidence $aconf(.,.)$ of a label l_j and an abstract concept al_k is defined as their semantic relatedness, weighted by the score of the assertion containing the abstract concept in ConceptNet [96]. For example, `hasProperty(baby, newborn)` has a score of 10.17 in ConceptNet. We name this score $CNet(al_k)$.

$$aconf(l_j, al_k) = CNet(al_k) * srel(l_j, al_k) \tag{4.8}$$

Constraints:
$\sum_j X_{ij} <= 1$: for each bounding box x_i there can be at most one visual label ($\in VL$)
$\sum_j Y_j <= 5$: one image x can have at most five abstract labels ($\in AL$)

$$\left.\begin{aligned}&(1 - Z_{ijmk}) <= (1 - X_{ij}) + (1 - X_{mk})\\&Z_{ijmk} <= X_{ij}\\&Z_{ijmk} <= X_{mk}\end{aligned}\right\}\begin{array}{l}\text{Pair-wise mutual}\\\text{coherence between}\\\text{visual labels}\end{array}$$

$$\left.\begin{aligned}&(1 - W_{ijk}) <= (1 - X_{ij}) + (1 - Y_{k})\\&W_{ijk} <= X_{ij}\\&W_{ijk} <= Y_{k}\end{aligned}\right\}\begin{array}{l}\text{Pair-wise mutual}\\\text{coherence between}\\\text{visual and abstract}\\\text{labels}\end{array}$$

The final set of visual and abstract labels per image are expected to be highly coherent. This is validated in Section 4.5.

4.4 Data Sets and Tools

In this section, we present the image data sets as well as the criteria and heuristics we follow to mine the various background knowledge utilized in our optimization model.

ImageNet Object Classes. $LSDA$ [65] is used to get the initial visual object labels from low-level image features. The LSDA tool has been trained on 7604 leaf-level nodes of $ImageNet$ [39]. Most of these object classes are exotic concepts which rarely occur in everyday images. Examples include scientific names of flora and fauna – *interior live oak, Quercus wislizenii, American white oak, Quercus alba*, and obscure terms – *pannikin, reliquary, lacrosse*. We prune those exotic classes by thresholding on their Google and Flickr search result counts. Some object class names are ambiguous where two senses of the same word from WordNet have been included. We consider only the most common sense. We work with the most frequent 1000 object classes obtained after pruning[2].

WordNet Hypernyms. For the ImageNet object classes described above, we traverse the WordNet hypernymy tree of the associated sense up to level three. We restrict the traversal level to avoid too much generalization – for example, *person* generalizing to *organism*. We prune out hypernyms with Google and Flickr result counts below a threshold. By considering the hypernyms of the 1000 ImageNet object classes mentioned above, we add 800 new visual labels to the model.

The ImageNet object classes and the WordNet hypernyms together constitute the **Visual Labels** of VISIR.

Abstract Labels. Commonsense knowledge (CSK) assertions from ConceptNet [147] contribute to concept abstraction in VISIR. For example, in Figure 4.2, the abstract concept *poisonous* is added to the labels of the fifth image. ConceptNet is a crowd-sourced knowledge base where most assertions have the default confidence score of 1.0 (as they were stated only by one person). Only popular statements like `hasProperty(apple, red fruit)` are stated by multiple people, hence raising the confidence score significantly. Certain assertions have contradictory scores – for example, `usedFor(acne medicine, clear skin)` appears twice, with scores 1.0 and -1.0.

[2] The full list is available at `http://people.mpi-inf.mpg.de/~sreyasi/visTagRef/1000classes_names.txt`

This happens when someone down-votes a statement. Using such indistinctive scores in VISIR would be uninformative. We therefore use the joint semantic relatedness of the assertion and visual labels of an image, weighted by the ConceptNet score (only positive scores), as the abstraction confidence.

Visual Similarity. The visual similarity or "confusability" scores (Equation 4.1) are mined from object detection results (from low-level image features) over 1 million images from the following data sets that are popularly employed in the computer vision community: $Flickr\ 30K$ [181], $Pascal\ Sentence\ Dataset$ [126], $SBU\ Captions\ Dataset$ [117], $MSCOCO$ [89]. All these data sets have collections of Flickr images not pertaining to any particular domain. For each detected bounding box, $LSDA$ provides a confidence score distribution over 7604 object classes (leaf nodes in ImageNet). Only predictions with a positive confidence score are considered as candidates for a bounding box. An object class pair appearing as candidates for the same bounding box are considered as visually similar. Table 4.1 shows few examples of visually similar object class pairs – *mail train* and *commuter train* are confused 91% times whereas *diaper* and *plaster cast* are confused 18% times.

object1	object2	visual similarity
mail train	commuter train	0.91
cattle	horse	0.76
soccer ball	kite baloon	0.26
Red Delicious	bowling ball	0.21
diaper	plaster cast	0.18
bicycle pump	mascara	0.17

Table 4.1: Object class pairs and visual similarity scores

Spatial Co-location. Spatial co-location scores between different object classes are mined from ground truth annotations of the detection challenge (DET) of ImageNet ILSVRC 2015 [135]. We consider two objects to be spatially co-located only if they are tagged in the same image. For simplicity, we do not consider the physical distance between the bounding boxes of the tagged object classes. A frequency-based co-location score is assigned to pairs of object classes based on evidence over the train set of ILSVRC DET. We find spatial co-location data for 200 object classes (since the detection challenge only considers 200 object classes). The top few frequently co-located objects are: *(person, microphone), (table, chair), (person, sunglasses), (person, table), (person, chair)*. A general observation would be that the image collection in ILSVRC DET has a high occurrence of *person*.

4.5 Experiments and Results

We analyze and compare the results that VISIR produces with that of two baselines: $LSDA$[3] and $YOLO$[4]. The performances of LSDA, YOLO, and VISIR are compared on the basis of precision, recall and F1-score measures.

[3] `http://lsda.berkeleyvision.org/`
[4] `https://pjreddie.com/darknet/yolo/`

4.5.1 Setup

As discussed in Section 4.3.1, we operate with three kinds of labels: visual class labels from ImageNet (CL), their generalizations (XL) which consist of WordNet hypernyms of labels $\in CL$, and abstract labels (AL) from commonsense knowledge. We evaluate the methods with respect to three different label spaces (as the combination of three types of labels): CL, $CL + XL$, and $CL + XL + AL$. LSDA and YOLO operate only on CL, while VISIR has three variants (configuring it for the above combinations of label spaces). Each system is given a label budget of 5 tags per image. For VISIR, this is enforced by an ILP constraint; for the two baselines, we use their confidence scores to pick the top-5.

Hyper-parameter Tuning: To tune the hyper-parameters for Equation 4.3 we use the annotations of the training image set of ILSVRC DET. We also extend this set by adding the hypernyms of the ground-truth labels as correct labels. A randomized search is used to tune the hyperparameters.

User Evaluation: Besides establishing semantic coherence among concrete object labels, VISIR applies concept generalization and abstraction. For modern benchmark datasets like ILSVRC 2015 DET, such enriched labeling does not exist so far. Therefore, in order to evaluate VISIR and compare to baselines, we construct a labeled image dataset by collecting human judgments about correctness of labels as discussed below.

For each label space, CL, $CL + XL$, and $CL + XL + AL$, the union of the labels produced by each method forms the set of result labels for an image. This result pool is evaluated by human annotators. Judges determined whether each label is appropriate and informative for an image. Instead of a binary assessments, annotators are asked to grade each label in the pool with 0, 1, or 2 – 0 corresponding to incorrect labels, 2 corresponding to highly relevant labels. We gather user judgments for the three label pools (corresponding to the label spaces) separately. This produces three different sets of graded labels per image. Users are not informed about the nature of the label pools. For each label pool we collect responses from at least 5 judges. The final assessment is determined by the majority of the judges (e.g., at least 3 out of 5 need to assert that a label is good).

Selection of Test Images: A major goal of this work is to make the refined labels more coherent or semantically related. Hence, we focus on the case where the deep-learning-based object detection tools produce contextually incoherent results. For the user evaluation, we collect a set of images with a reasonable context – those that have 3-7 detected bounding boxes and with LSDA labels having a semantic relatedness score less than 0.1. Such 100 images are collected from the ILSVRC 2015 DET Val image set.

4.5.2 Model Performance

Precision is estimated as the fraction of "good labels" detected, where a "good label" is one considered relevant by the majority of the human judges. We assess the recall per method as the number of labels picked from the good labels in the pool of labels generated by all three methods. The recall is artificially restricted because the label pool may contain more good labels than the label budget of the method. For example, if the label budget per method is set to 5, even if all 5 labels of a method are good, the recall for a pool with 8 good labels would only be 5/8. However, it is a fair notion across the different methods.

Relaxed vs Conservative Assessments: According to the evaluation design, labels graded 1 are either inconspicuous, or less relevant to the image than labels graded 2. In order to identify the “good labels” in a label pool, we define two methods of assessment: *Relaxed Assessment* considers all labels graded 1 or 2 as correct. *Conservative Assessment* considers only those labels graded 2 as correct, resulting in a stricter setup. The three graded label pools from the user evaluations naturally have labels in common.

Performance Results: Tables 4.2 through 4.4 compare the three methods for the three different label pools – CL, $CL + XL$, $CL + XL + AL$ – with conservative assessment. For CL, there is no real improvement over LSDA, but we see that for $CL + XL$ and $CL + AL + XL$ VISIR adds a good number of semantically informative labels and improves on the two baselines in terms of both precision and recall.

We also test VISIR’s performance with a tighter constraint on choosing the number of bounding boxes per image, by setting the label budget to 80% of all bounding boxes received as input. This variant, which we refer to as VISIR*, aims to filter out more noise in the output of the deep-learning-based object detections. Naturally, VISIR*-CL would have higher precision than VISIR-CL while sacrificing on recall. VISIR*-CL improves further on precision and F1-score because it is able to eliminate some of the initial noise the LSDA detections bring in. For pools $CL + XL$ and $CL + XL + AL$, VISIR* has higher precision than VISIR, but slighly loses in recall.

System	Precision	Recall	F1-score
LSDA	0.51	0.86	0.64
YOLO	0.49	0.56	0.52
VISIR-CL	0.51	0.86	0.64
VISIR*-CL	**0.57**	0.81	**0.67**

Table 4.2: Pool CL: Conservative Assessment

System	Precision	Recall	F1-score
LSDA	0.52	0.81	0.63
YOLO	0.49	0.51	0.50
VISIR-CL+XL	**0.54**	**0.82**	**0.65**
VISIR*-CL+XL	**0.60**	0.76	**0.67**

Table 4.3: Pool CL+XL: Conservative Assessment

System	Precision	Recall	F1-score
LSDA	0.49	0.35	0.41
YOLO	0.52	0.23	0.32
VISIR-CL+XL+AL	**0.54**	**0.91**	**0.68**
VISIR*-CL+XL+AL	**0.56**	**0.89**	**0.69**

Table 4.4: Pool CL+XL+AL: Conservative Assessment

Table 4.5 and Table 4.6 show the relaxed and conservative assessments with respect to the combined pool (i.e., for all three label spaces together) of good labels per image. It is natural that all methods perform better for the relaxed setting compared to that of the conservative assessment. However, the observation that VISIR's performance does not degrade much for the conservative assessment demonstrates its high output quality and robustness. Figure 4.3 illustrates this by anecdotal examples with the labels assigned by each of the competitors (with good labels in black and bad ones in red). In image 4, LSDA produces typically unrelated labels – a *monkey* and a *tennis ball*. This contextual incoherence likely arises due to low level color features. In contrast to LSDA, YOLO addresses the semantic coherence of the labels, however likely in expense of recall (for example in image 6). By necessitating semantic coherence among detected labels VISIR eliminates incoherent labels - for example, VISIR removes *motorcycle* from image 1, *tennis ball* from image 4, *hat with a wide brim* from image 5 and so on.

System	Precision	Recall	F1-score
LSDA	0.55	0.30	0.39
YOLO	0.57	0.19	0.29
VISIR-CL	**0.57**	0.28	0.38
VISIR-CL+XL	**0.62**	0.30	**0.40**
VISIR-CL+XL+AL	**0.71**	**0.90**	**0.79**

Table 4.5: Aggregate Pool: Relaxed Assessment

System	Precision	Recall	F1-score
LSDA	0.49	0.35	0.41
YOLO	0.52	0.23	0.32
VISIR-CL	**0.52**	0.34	0.41
VISIR-CL+XL	**0.55**	**0.35**	**0.43**
VISIR-CL+XL+AL	**0.54**	**0.91**	**0.68**

Table 4.6: Aggregate Pool: Conservative Assessment

Table 4.7 lists the new labels introduced by VISIR, each for at least 10 images. These labels are generated via generalization (from WordNet hypernyms) and abstraction (from commonsense knowledge). As none of the baselines can produce these labels, VISIR naturally achieves a recall of 1. The precision values for the labels illustrate how VISIR addresses the problem of label incompleteness. In most cases, these labels were assessed as correct by the judges.

Label	Label frequency	Precision
individual	46	0.59
man or woman	44	0.64
animal	31	0.94
human	20	0.95
canine	18	0.94
furniture	12	0.83
barking animal	11	1.00

Table 4.7: New labels suggested by VISIR

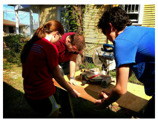
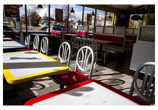
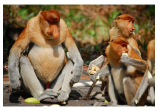
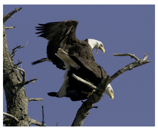
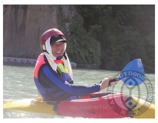
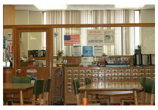
	LSDA	YOLO	VISIR-CL	VISIR-CL+XL	VISIR-CL+XL+AL
	person table motorcycle	bench person bowl	person table	person table	person table man or woman furniture
	table car	tow truck bench car chair	table chair	table chair	seat furniture chair flat dining furniture table
	monkey tennis ball	bird dog	monkey	monkey	primate monkey orangutan ape simian furry
	bird hat with a wide brim	airplane bird	bird	bird	bird avian flying animal
	helmet person watercraft smelling bottle bathing cap record sleeve impeller	person	bathing cap bib watercraft helmet person	bathing cap fabric watercraft helmet person	bathing cap cloth, fabric watercraft helmet protective hat individual person
	table baby bed swim trunks	chair	chair table	chair table	seat furniture chair table flat

Figure 4.3: Images with labels from LSDA, YOLO, and different configurations of VISIR

4.6 Discussion of Limitations

Training Bias in LSDA: The LSDA tool predicts only leaf-level object classes of ImageNet. The same limitation holds for most other state-of-the-art object detectors. Because of this incomplete tag space many important objects cannot be detected. For example, *giraffe* is not a leaf-level object class of ImageNet. Since LSDA did not see any training images of a giraffe, it mis-labels objects in Figure 4.1a according to its training. This noise propagates to our model, sometimes making it impossible to find the correct labels.

Incorrect Sense Mapping in ImageNet: LSDA trains on ImageNet images. Hence improper word sense mappings in ImageNet propagate to incorrect labels from LSDA. For example, ImageNet contains similar images for two separate synsets *Sunglass (a convex lens used to start a fire)* and *Sunglasses (shades, dark glasses)*. Naturally, these two synsets have completely different WordNet hypernyms which VISIR uses, hence introducing noise. The direct hypernym of *Sunglass* is *lens*, while that of *Sunglasses* is *glasses*.

Incomplete Spatial Co-location data: Spatial co-location patterns mined from text contain noise due to linguistic variations in the form of proverbs. For example, the commonsense knowledge base WebChild [152] assigns significant confidence to the spatial co-location of *elephant* and *room* (most likely from the idiom “the elephant in the room”). To counter such linguistic bias, we have mined spatial co-location information from the manually annotated ground truth of ILSVRC [135]. Unfortunately, annotations are available for only 200 object classes, leaving us with only a small fraction of annotated visual-label pairs. If more cues of this kind were available, we would have been able to establish stronger contextual coherence.

Incomplete and Noisy Commonsense Knowledge: ConceptNet and WebChild are quite incomplete; so we cannot assign an abstract concept to every detected visual label. Also, assertions in these knowledge bases are often contradictory and noisy. We manage to reduce the noise by considering semantic relatedness with the visual labels, but this only alleviates part of the problem.

In this chapter we presented VISIR, a new method for refining and expanding visual labels for images. Its key strengths are cleaning out noisy labels from predictions by object detection tools and adding informative labels that capture generalizations and abstractions. Our model makes this feasible by considering the visual similarity of labels, the semantic coherence across concepts, and various kinds of background knowledge. The joint inference on an enriched label candidate space is performed by means of a judiciously designed Integer Linear Program. Our experiments show the viability of the approach, and also demonstrate significant improvements over two state-of-the-art object detection and tagging tools.

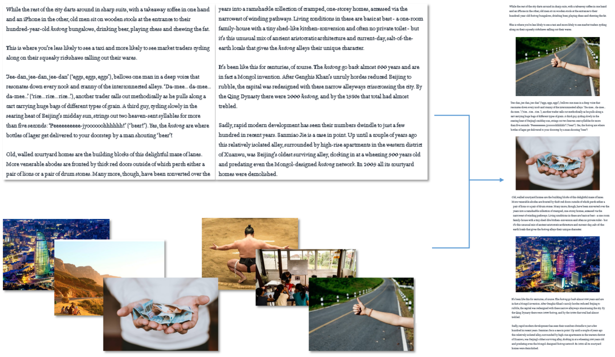
CHAPTER 5

Story-oriented Image Selection and Placement

Contents

MULTIMODAL contents have become commonplace on the Internet today, manifested as news articles, social media posts, and personal or business blog posts. Among the various kinds of media (images, videos, graphics, icons, audio) used in such multimodal stories, images are the most popular. The selection of images from a collection – either author's personal photo album, or web repositories – and their meticulous placement within a text, builds a succinct multimodal commentary for digital consumption. In this paper we present a system that automates the process of selecting relevant images for a story and placing them at contextual paragraphs within the story for a multimodal narration. We leverage automatic object recognition, user-provided tags, and commonsense knowledge, and use an unsupervised combinatorial optimization to solve the selection and placement problems seamlessly as a single unit. To this end. we present a framework called Story-AND-Images Alignment (SANDI) (which stands for Story-AND-Images).

5.1 Introduction



While the rest of the city darts around in sharp suits, with a takeaway coffee in one hand and an iPhone in the other, old men sit on wooden stools at the entrance to their hundred-year-old *hutong* bungalows, drinking beer, playing chess and chewing the fat.

This is where you're less likely to see a taxi and more likely to see market traders cycling along on their squeaky rickshaws calling out their wares.

'Jee-dan, jee-dan, jee-dan' ('eggs, eggs, eggs'), bellows one man in a deep voice that resonates down every nook and cranny of the interconnected alleys. 'Da-mee... da-mee... da-mee...' ('rice... rice... rice...'), another trader calls out methodically as he pulls along a cart carrying huge bags of different types of grain. A third guy, cycling slowly in the searing heat of Beijing's midday sun, strings out two heaven-sent syllables for more than five seconds: 'Peeeeeeeeee-jyoooooohhhhhh!' ('beer!'). Yes, the *hutong* are where bottles of lager get delivered to your doorstep by a man shouting 'beer'!

Old, walled courtyard homes are the building blocks of this delightful maze of lanes. More venerable abodes are fronted by thick red doors outside of which perch either a pair of lions or a pair of drum stones. Many more, though, have been converted over the years into a ramshackle collection of cramped, one-storey homes, accessed via the narrowest of winding pathways. Living conditions in these are basic at best - a one-room family-house with a tiny shed-like kitchen-conversion and often no private toilet - but it's this unusual mix of ancient aristocratic architecture and current-day, salt-of-the-earth locals that gives the *hutong* alleys their unique character.

It's been like this for centuries, of course. The *hutong* go back almost 800 years and are in fact a Mongol invention. After Genghis Khan's unruly hordes reduced Beijing to rubble, the capital was redesigned with these narrow alleyways crisscrossing the city. By the Qing Dynasty there were 2000 *hutong*, and by the 1950s that total had almost trebled.

Sadly, rapid modern development has seen their numbers dwindle to just a few hundred in recent years. Sanmiao Jie is a case in point. Up until a couple of years ago this relatively isolated alley, surrounded by high-rise apartments in the western district of Xuanwu, was Beijing's oldest surviving alley, clocking in at a wheezing 900 years old and predating even the Mongol-designed *hutong* network. In 2009 all its courtyard homes were demolished.

Figure 5.1: The story-and-images alignment problem: given an album of images and a textual narrative, the task it to select relevant images and place them at coherent locations in the story.

It is well-known (and supported by studies [82, 105]) that the most powerful messages are delivered with a combination of words and pictures. On the Internet, such multimodal content is abundant in the form of news articles, social media posts, and personal blog posts where authors enrich their stories with carefully chosen and placed images. As an example, consider a vacation trip report, to be posted on a blog site or online community. The backbone of the travel report is a textual narration, but the user typically places illustrative images in appropriate spots, carefully selected from her photo collection from this trip. These images can either show specific highlights such as waterfalls, mountain hikes or animal encounters, or may serve to depict feelings and the general mood of the trip, e.g., by showing nice sunsets or bar scenes. Another example is brochures for research institutes or other organizations. Here, the text describes the mission, achievements and ongoing projects, and it is accompanied with judiciously selected and placed photos of buildings, people, products and other images depicting the subjects and phenomena of interest, e.g., galaxies or telescopes for research in astrophysics.

The generation of such multimodal stories requires substantial human judgement and reasoning, and is thus time-consuming and labor-intensive. In particular, the effort on the human side includes selecting the right images from a pool of story-specific photos (e.g., the traveler's own photos) and possibly also from a broader pool for visual illustration (e.g., images licensed from a PR company's catalog or a big provider such as Pinterest). Even if the set of photos were exactly given, there is still considerable effort to place them within or next to appropriate paragraphs, paying attention to the semantic coherence between surrounding text and image. In this paper, we set out to automate this human task, formalizing it as a *story-images alignment* problem.

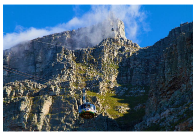
Problem Statement. Given a story-like text document and a set of images, the problem is to automatically decide where individual images are placed in the text. Figure 5.1 depicts this task. The problem comes in different variants: either all images in the given set need to be placed, or a subset of given cardinality must be selected and aligned with text paragraphs. Formally, given n paragraphs and $m \leq n$ images, assign these images to a subset of the paragraphs, such that each paragraph has at most one image. The variation with image selection assumes that $m > n$ and requires a budget $b \leq n$ for the number of images to be aligned with the paragraphs.

Prior Work and its Inadequacy. There is ample literature on computer support for multimodal content creation, most notably, on generating image tags and captions. Closest to our problem is prior work on story illustration [70, 137], where the task is to select illustrative images from a large pool. However, the task is quite different from ours, making prior approaches inadequate for the setting of this paper. First, unlike in general story illustration, we need to consider the text-image alignments jointly for all pieces of a story, rather than making context-free choices one piece at a time. Second, we typically start with a pool of story-specific photos and expect high semantic coherence between each narrative paragraph and the respective image, whereas general story illustration operates with a broad pool of unspecific images that serve many topics. Third, prior work assumes that each image in the pool has an informative caption or set of tags, by which the selection algorithm computes its choices. Our model does not depend on pre-defined set of tags, but detects image concepts on the fly.

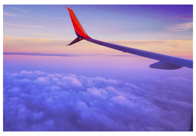
Research on Image Tagging may be based on community input, leading to so-called “social tagging” [60], or based on computer-vision methods, called “visual tagging”. In the latter case, bounding boxes are automatically annotated with image labels, and relationships between objects may also be generated [129, 99]. Recent works have investigated how to leverage commonsense knowledge as a background asset to further enhance such automatically computed tags [28]. Also, deep-learning methods have led to expressive forms of multimodal embeddings, where textual descriptions and images are projected into a joint latent space [51, 47] in order to compute multimodal similarities.

In this paper, in addition to manual image tags where available, we harness visual tags from deep neural network based object-detection frameworks and incorporate background commonsense knowledge, as automatic steps to enrich the semantic interpretation of images. This, by itself, does not address the alignment problem, though. The alignment problem is solved by combinatorial optimization. Our method is experimentally compared to baselines that makes use of multimodal embeddings.

Our Approach – SANDI. We present a framework that casts the story-images alignment task into a combinatorial optimization problem. The objective function, to be maximized, captures the semantic coherence between each paragraph and the image that is placed there. To this end, we consider a suite of features, most notably, the visual tags associated with an image (user-defined tags as well as tags from automatic computer-vision tools), text embeddings, and also background knowledge in the form of commonsense assertions. The optimization is constrained by the number of images that the story should be enriched with. As a solution algorithm, we devise an integer linear program (ILP)

Image	Ground Truth Paragraph
	. . . Table Mountain Cableway. The revolving car provides 360 degree views as you ascend this mesmerising 60-million-year-old mountain. From the cableway station. . .
	. . . On the east flank of the hill is the old Muslim quarter of the Bo-Kaap; have your camera ready to capture images of the photogenic pastel-painted pre-colonial homes. . .

(a) Sample image and corresponding paragraph from Lonely Planet

Image	Ground Truth Paragraph
	If you are just looking for some peace and quiet or hanging out with other students...library on campus, a student hangout space in the International College building. . . .
	. . . I was scared to travel alone. But I quickly realized that there's no need to be afraid. Leaving home and getting out of your comfort zone is an important part of growing up.. . .

(b) Sample image and corresponding paragraph from Asia Exchange

Figure 5.2: Image-text semantic coherence in datasets.

and employ the Gurobi ILP solver for computing the exact optimum. Experiments show that SANDI produces semantically coherent alignments.

Contributions. To the best of our knowledge, this is the first work to address the story-images alignment problem. Our salient contributions are:

1. We introduce and define the problem of story-images alignment.
2. We analyze two real-world datasets of stories with rich visual illustrations, and derive insights on alignment decisions and quality measures.
3. We devise relevant features, formalize the alignment task as a combinatorial optimization problem, and develop an exact-solution algorithm using integer linear programming.
4. We present experiments that compare our method against baselines that use multimodal embeddings.

5.2 Related Work

Story Illustration. Existing research finds suitable images from a big image collection to illustrate personal stories [70] or news posts [137, 37]. Traditionally, images are searched based on textual tags associated with image collections. Occasionally they use visual similarity measures to prune out images very similar to each other. More recent frameworks use deep neural networks to find suitable representative images for a story [127]. Story Illustration only addresses the problem of image selection, whereas we solve two problems simultaneously: image selection and image placement – making a joint decision on all pieces of the story. [127] operates on small stories (5 sentences) with simple content, and retrieves 1 image per sentence. Our stories are much longer texts, the sentences are more complex, and the stories refer to both general concepts and named entities. This makes our problem distinct. We cannot systematically compare our full-blown model with prior works on story illustration alone.

Image-text Comparison. Visual similarity of images has been leveraged to associate single words [183] or commonly occurring phrases [186] to a cluster of images. While this is an effective solution for better indexing and retrieval of images, it can hardly be used for contextual text-image alignment. For example, an image with a beach scene may be aligned with either "relaxed weekend" or "this is where I work best" depending on the context of the full text.

Yet another framework combines visual features from images with verbose image descriptions to find semantically closest paragraphs in the corresponding novels [190], looking at images and paragraphs in isolation. In a similar vein, [30] align images with one semantically closest sentence in the corresponding article for viewing on mobile devices. In contrast, we aim to generate a complete longer multimodal content to be read as a single unit. This calls for distinction between paragraphs and images, and continuity of the story-line. [2] temporally aligns images and their corresponding captions into a story sequence. Their task is much simpler since image-caption pairs are already aligned.

Commonsense Knowledge for Story Understanding. One of the earliest applications of Commonsense Knowledge to interpret the connection between images and text is a photo agent which automatically annotated images from user's multi-modal (text and image) emails or web pages, while also inferring additional commonsense concepts [86]. Subsequent works used commonsense reasoning to infer causality in stories [168], especially applicable to question answering. The most commonly used database of commonsense concepts is ConceptNet [146]. We enhance automatically detected concepts in an image with relevant commonsense assertions. This often helps to capture more context about the image.

5.3 Dataset and Problem Analysis

5.3.1 Datasets

To the best of our knowledge, there is no experimental dataset for text-image alignment, and existing datasets on image tagging or image caption generation are not suitable in our setting. We therefore compile and analyze two datasets of blogs from Lonely Planet[1] and Asia Exchange[2].

- Lonely Planet: 2178 multimodal articles containing on average 20 paragraphs and 4.5 images per article. Most images are accompanied by captions. Figure 5.2a shows two image-paragraph pairs from this dataset. Most of the images and come from the author's personal archives and adhere strictly to the content of the article.

[1] www.lonelyplanet.com/blog
[2] www.asiaexchange.org

	Criterion	% of images
Relevance	Placement specific to surrounding paragraphs	91%
	Relevant text after image	86%
	Avg. #relevant paragraphs	1.65
Main reason	Natural named objects	9%
	Human activities	12%
	Generic objects	15%
	General nature scenes	20%
	Man-made named objects	21%
	Geographic locations	29%

Table 5.1: Analysis of image placement for 50 images from Lonely Planet travel blogs.

- Asia Exchange: 200 articles about education opportunities in Asia, with an average of 13.5 paragraphs and 4 images per article. The images may be strongly adhering to the content of the article (top image in Figure 5.2b), or they may be generic stock images complying with the abstract theme as seen in the bottom image in Figure 5.2b). Most images have captions.

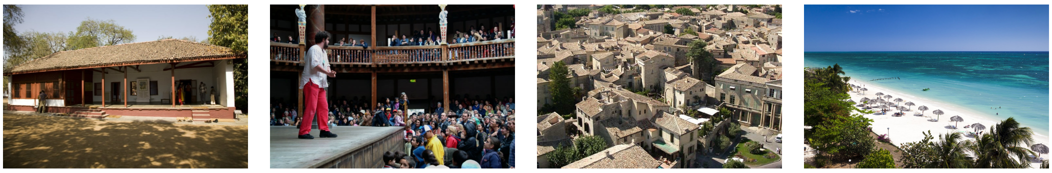
Text-Image Semantic Coherence. To understand the specific nature of this data, we had two annotators analyze the given placement of 50 randomly chosen images in articles from the Lonely Planet dataset. The annotators assessed whether the images were specific to the surrounding paragraphs as opposed to merely being relevant for entire articles. The annotators also determined to how many paragraphs an image was specifically fitting, and indicated the main reason for the given alignments. For this purpose, we defined 6 possibly overlapping meta-classes: (i) specific man-made entities such as monuments, buildings or paintings, (ii) natural objects such as lakes and mountains, (iii) general nature scenes such as fields or forest, (iv) human activities such as biking or drinking, (v) generic objects such as animals or cars, and (vi) geographic locations such as San Francisco or Rome.

The outcome of the annotation is shown in Table 5.1. As one can see, 91% of the images were indeed more specifically relevant to surrounding text than to the article in general, and 86% of these were placed before the relevant text. We therefore assume the paragraph following the image as ground truth. As to the main reasons for this relevance, we observe quite a mix of reasons, with geographic locations being most important at 29%, followed by man-made objects at 21% and general nature scenes at 20% and so on.

5.3.2 Image Descriptors

Based on the analysis in Table 5.1, we consider the following kinds of tags for describing images:

Visual Tags (CV). State-of-the-art computer-vision methods for object and scene detection yield visual tags from low-level image features. We use three frameworks for this purpose. First, deep convolutional neural networks based architectures like LSDA [64] and YOLO [129], are used to detect objects like *person*, *frisbee* or *bench*. These models have been trained on ImageNet object classes and denote "Generic objects" from Table 5.1. For stories, general scene descriptors like *restaurant* or *beach* play a major role, too. Therefore, our second asset is scene detection, specifically from the MIT Scenes Database [185]. Their publicly released pre-trained model "Places365-CNN", trained on 365 scene categories with 5000 images per category, predicts scenes in images with corresponding



CV: country store, person, bench, lodge outdoor
MAN: unassuming ashram, Mahatma Ghandi
BD: Sabarmati Ashram

CV: person, sunglasses, stage
MAN: Globe Theatre, performance, Shakespeare, Spectators
BD: Shakespeare's Globe
CSK: show talent, attend concert, entertain audience

CV: adobe brick, terra cotta, vehicle, table, village
MAN: tiled rooftops
BD: uzes languedoc, languedoc roussillon
CSK: colony, small town

CV: umbrella, beach
MAN: white sands, Playa Ancon
BD: ancon cuba, playa ancon
CSK: sandy shore, vacation

Figure 5.3: Characterization of image descriptors: CV adds visual objects/scenes, MAN and BD add location details, CSK adds high-level concepts.

confidence scores. We pick the most confident scene for each image. These constitute "General nature scenes" from Table 5.1. Thirdly, since stories often abstract away from explicit visual concepts, a framework that incorporates generalizations and abstractions into visual detections [28] is also leveraged. For e.g., the concept "hiking" is supplemented with the concepts "walking" (Hypernym of "hiking" from WordNet) and "fun" (from ConceptNet [146] assertion "hiking, HasProperty, fun").

User Tags (MAN). Owners of images often have additional knowledge about content and context – for e.g., activities or geographical information ("hiking near Lake Placid"), which, from Table 5.1 play a major role in text-image alignment. In a down-stream application, users would have the provision to specify tags for their images. For experimental purposes, we use the nouns and adjectives from image captions from our datasets as proxy for user tags.

Big-data Tags (BD). Big data and crowd knowledge allow to infer additional context that may not be visually apparent. We utilize the Google reverse image search API[3] to incorporate such tags. This API allows to search by image, and suggests tags based on those accompanying visually similar images in the vast web image repository. These tags often depict popular places and entities, such as "Savarmati Ashram", or "Mexico City insect market", and thus constitute "Natural names objects", "Man-made named objects", as well as "Geographic locations" from Table 5.1.

Commonsense Knowledge (CSK). CSK can bridge the gap between visual and textual concepts [29]. We use the following ConceptNet relations to enrich the image tag space: *used for, has property, causes, at location, located near, conceptually related to*. As ConceptNet is somewhat noisy, subjective, and diverse, we additionally filter its concepts by *informativeness* for a given image following [173]. If the top-10 web search results of a CSK concept are semantically similar to the image context (detected image tags), the CSK concept is considered to be *informative* for the image. For example, consider the image context "*hike, Saturday, waterproof boots*". CSK derived from "hike" are *outdoor activity*, and *fun*. The top-10 Bing search results for the concept *outdoor activity* are semantically similar to the image context. However, those for the term *fun* are semantically varied. Hence, *outdoor activity* is more informative than *fun* for this image. Cosine similarity between the mean vectors of the image context and the search results is used as a measure of semantic similarity.

Figure 5.3 shows examples for the different kinds of image tags.

[3] www.google.com/searchbyimage

5.4 Model for Story-Images Alignment

Without substantial amounts of labeled training data, there is no point in considering machine-learning methods. Instead, we tackle the task as a Combinatorial Optimization problem in an unsupervised way.

Our *story-images alignment* model constitutes an Integer Linear Program (ILP) which jointly optimizes the placement of selected images within an article. The main ingredient for this alignment is the pairwise similarity between images and units of text. We consider a paragraph as a text unit.

Text-Image Pairwise Similarity. Given an image, each of the four kinds of descriptors of Section 5.3.2 gives rise to a bag of features. We use these features to compute *text-image semantic relatedness scores srel*(i, t) for an image i and a paragraph t.

$$srel(i, t) = cosine(\mathbf{i}, \mathbf{t}) \tag{5.1}$$

where $\mathbf{i}$ and $\mathbf{t}$ are the mean word embeddings for the image tags and the paragraph respectively. For images, we use all detected tags. For paragraphs, we consider only the top 50% of concepts w.r.t. their Term Frequence – Inverse Document Frequency (TF-IDF) ranking over the entire dataset. Both paragraph concepts and image tags capture unigrams as well as bigrams. We use word embeddings from word2vec trained on Google News Corpus.

$srel(i, t)$ scores serve as weights for variables in the ILP. Note that model for text-image similarity is orthogonal to the combinatorial problem solved by the ILP. Cosine distance between concepts (as in Eq. 5.1) could be easily replaced by other similarity measures over the multimodal embedding space.

Tasks. Our problem can be divided into two distinct tasks:

- Image Selection – to select relevant images from an image pool.
- Image Placement – to place selected images in the story.

These two components are modelled into one ILP where Image Placement is achieved by maximizing an objective function, while the constraints dictate Image Selection. In the following subsections we discuss two flavors of our model consisting of one or both of the above tasks.

5.4.1 Complete Alignment

Complete Alignment constitutes the problem of aligning *all images* in a given image pool with relevant text units of a story. Hence, only Image Placement is applicable. For a story with $|T|$ text units and an associated image pool with $|I|$ images, the alignment of images $i \in I$ to text units $t \in T$ can be modeled as an ILP with the following definitions:

Decision Variables: The following binary decision variables are introduced:
$X_{it} = 1$ if image i should be aligned with text unit t, 0 otherwise.

Objective: Select image i to be aligned with text unit t such that the semantic relatedness over all text-image pairs is maximized:

$$max \left[\sum_{i \in I} \sum_{t \in T} srel(i, t) X_{it} \right] \tag{5.2}$$

where $srel(i, t)$ is the text-image semantic relatedness from Eq. 5.1.

Constraints:

$$\sum_i X_{it} \leq 1 \forall t \tag{5.3}$$

$$\sum_t X_{it} = 1 \forall i \tag{5.4}$$

We make two assumptions for text-image alignments: no paragraph may be aligned with multiple images (5.3), and each image is used exactly once in the story (5.4). The former is an observation from multimodal presentations on the web such as in blog posts or brochures. The latter assumption is made based on the nature of our datasets, which are fairly typical for web contents. Both are designed as hard constraints that a solution must satisfy. In principle, we could relax them into soft constraints by incorporating violations as a loss-function penalty into the objective function. However, we do not pursue this further, as typical web contents would indeed mandate hard constraints. Note also that the ILP has no hyper-parameters; so it is completely unsupervised.

5.4.2 Selective Alignment

Selective Alignment is the flavor of the model which *selects a subset* of thematically relevant images from a big image pool, and places them within the story. Hence, it constitutes both tasks – Image Selection and Image Placement. Along with the constraint in (5.3), Image Selection entails the following additional constraints:

$$\sum_t X_{it} \leq 1 \forall i \tag{5.5}$$

$$\sum_i \sum_t X_{it} = b \tag{5.6}$$

where b is the budget for the number of images for the story. b may be trivially defined as the number of paragraphs in the story, following our assumption that each paragraph may be associated with a maximum of one image. (5.5) is an adjustment to (5.4) which implies that not all images from the image pool need to be aligned with the story. The objective function from (5.2) rewards the selection of best fitting images from the image pool.

5.5 Quality Measures

In this section we define metrics for automatic evaluation of text-image alignment models. The two tasks involved – Image Selection and Image Placement – call for separate evaluation metrics as discussed below.

5.5.1 Image Selection

Representative images for a story are selected from a big pool of images. There are multiple conceptually similar images in our image pool since they have been gathered from blogs of the domain “travel”. Hence evaluating the results on strict precision (based on exact matches between selected and ground-truth images) does not necessarily assess true quality. We therefore define a relaxed precision metric (based on semantic similarity) in addition to the strict metric. Given a set of selected images I and the set of ground truth images J, where $|I| = |J|$, the precision metrics are:

$$RelaxedPrecision = \frac{\sum\limits_{i \in I} \max\limits_{j \in J}(cosine(\mathbf{i}, \mathbf{j}))}{|I|} \tag{5.7}$$

$$StrictPrecision = \frac{|I \cap J|}{|I|} \tag{5.8}$$

5.5.2 Image Placement

For each image in a multimodal story, the ground truth (Ground Truth (GT)) paragraph is assumed to be the one following the image in our datasets. To evaluate the quality of SANDI’s text-image alignments, we compare the GT paragraph and the paragraph assigned to the image by SANDI (henceforth referred to as “aligned paragraph”). We propose the following metrics for evaluating the quality of alignments:

BLEU and ROUGE. BLEU and ROUGE are classic n-gram-overlap-based metrics for evaluating machine translation and text summarization. Although known to be limited insofar as they do not recognize synonyms and semantically equivalent formulations, they are in widespread use. We consider them as basic measures of concept overlap between GT and aligned paragraphs.

Semantic Similarity. To alleviate the shortcoming of requiring exact matches, we consider a metric based on embedding similarity. We compute the similarity between two text units t_i and t_j by the average similarity of their word embeddings, considering all unigrams and bigrams as words.

$$SemSim(t_i, t_j) = cosine(\mathbf{t_i}, \mathbf{t_j}) \tag{5.9}$$

where $\mathbf{x}$ is the mean vector of words in x. For this calculation, we drop uninformative words by keeping only the top 50% with regard to their TF-IDF weights over the whole dataset.

Average Rank of Aligned Paragraph. We associate each paragraph in the story with a ranked list of all the paragraphs on the basis of semantic similarity (Eq. 5.9), where rank 1 is the paragraph itself. Our goal is to produce alignments ranked higher with the GT paragraph. The average rank of alignments produced by a method is computed as follows:

$$ParaRank = 1 - \left[\left(\frac{\sum\limits_{t \in T'} rank(t)}{|I|} - 1 \right) \Big/ \left(|T| - 1 \right) \right] \tag{5.10}$$

where $|I|$ is the number of images and $|T|$ is the number of paragraphs in the article. $T' \subset T$ is the set of paragraphs aligned to images. Scores are normalized between 0 and 1; 1 being the perfect alignment and 0 being the worst alignment.

Order Preservation. Most stories follow a storyline. Images placed at meaningful spots within the

	BLEU	ROUGE	SemSim	ParaRank	Order Preserve
Random	3.1	6.9	75.1	50.0	50.0
VSE++ [47]	11.0	9.5	84.6	59.1	55.2
VSE++ ILP	12.56	11.23	83.98	58.08	47.93
SANDI-CV	18.2	17.6	86.3	63.7	54.5
SANDI-MAN	**45.6**	**44.5**	**89.8**	72.5	**77.4**
SANDI-BD	26.6	25.1	84.7	61.3	61.2
SANDI∗	44.3	42.9	89.7	**73.2**	76.3

Table 5.2: Complete Alignment on the Lonely Planet dataset.

	BLEU	ROUGE	SemSim	ParaRank	Order Preserve
Random	6.8	8.9	70.8	50.0	50.0
VSE++ [47]	19.4	17.7	85.7	51.9	48.0
VSE++ ILP	23.5	20.11	85.98	52.55	46.13
SANDI-CV	21.5	20.6	87.8	58.4	52.0
SANDI-MAN	**35.2**	**32.2**	89.2	61.5	61.5
SANDI-BD	24.1	22.3	86.7	56.0	53.6
SANDI∗	33.4	31.5	**89.7**	**62.4**	**62.5**

Table 5.3: Complete Alignment on the Asia Exchange dataset.

story would ideally adhere to this sequence. Hence the measure of pairwise ordering provides a sense of respecting the storyline. Lets define order preserving image pairs as: $P = \{(i, i') : i, i' \in I, i \neq i', i'$ follows i in both GT and SANDI alignments$\}$, where I is the set of images in the story. The measure can be defined as number of order preserving image pairs normalized by the total number of GT ordered image pairs.

$$OrderPreserve = \frac{|P|}{(|I|(|I|-1)/2)} \tag{5.11}$$

5.6 Experiments and Results

We evaluate the two flavors of SANDI – Complete Alignment and Selective Alignment – based on the quality measures described in Section 5.5.

5.6.1 Setup

Tools. Deep convolutional neural network based architectures similar to $LSDA$ [64], $YOLO$ [129], VISIR [28] and $PlacesCNN$ [185] are used as sources of *Visual tags*. Google reverse image search tag suggestions are used as *Big-data tags*. We use the $Gurobi$ Optimizer for solving the ILP. A Word2Vec [108] model trained on the Google News Corpus encompasses a large cross-section of domains, and hence is a well-suited source of word embeddings for our purposes.

SANDI Variants. The variants of our text-image alignment model are based on the use of image descriptors from Section 5.3.2.

- SANDI-CV, SANDI-MAN, and SANDI-BD use CV, MAN, and BD tags as image descriptors respectively.
- SANDI$\ast$ combines tags from all sources.
- +CSK: With this setup we study the role of commonsense knowledge as a bridge between visual features and textual features.

Alignment sensitivity. The degree to which alignments are specific to certain paragraphs varies from article to article. For some articles, alignments have little specificity, for instance, when the whole article talks about a hiking trip, and images generically show forests and mountains. We measure alignment sensitivity of articles by comparing the semantic relatedness of an image to its ground-truth paragraph against all other paragraphs in the same article. We use the cosine similarity between the image's vector of MAN tags and the text vectors, for this purpose. The alignment sensitivity of an article then is the average of these similarity scores over all its images. We restrict our experiments to the top-100 most alignment-sensitive articles in each dataset.

5.6.2 Complete Alignment

In this section we evaluate our Complete Alignment model (defined in Section 5.4.1), which places *all* images from a given image pool within a story.

Baselines. To the best of our knowledge, there is no existing work on story-image alignment in the literature. Hence we modify methods on joint visual-semantic-embeddings (VSE) [72, 47] to serve as baselines. Our implementation of VSE is similar to [47], henceforth referred to as VSE++. We compare SANDI with the following baselines:

- Random: a simple baseline with random image-text alignments.
- VSE++ Greedy or simply VSE++: for a given image, VSE++ is adapted to produce a ranked list of paragraphs from the corresponding story. The best ranked paragraph is considered as an alignment, with a greedy constraint that one paragraph can be aligned to at most one image.
- VSE++ ILP: using cosine similarity scores between image and paragraph from the joint embedding space, we solve an ILP for the alignment with the same constraints as that of SANDI.

Since there are no existing story-image alignment datasets, VSE++ has been trained on the $MSCOCO$ captions dataset [89], which contains 330K images with 5 captions per image.

Evaluation. Tables 5.2 and 5.3 show the performance of SANDI variants across the different evaluation metrics (from Section 5.5.2) on the Lonely Planet and Asia Exchange datasets respectively. On both datasets, SANDI outperforms VSE++, especially in terms of paragraph rank (+14.1%/+10.5%) and order preservation (+11.1%/+14.5%). While VSE++ looks at each image in isolation, SANDI captures context better by considering all text units of the article and all images from the corresponding album at once in a constrained optimization problem. VSE++ ILP, although closer to SANDI in methodology, does not outperform SANDI. The success of SANDI can also be attributed to the fact that it is less tied to a particular type of images and text, relying only on word2vec embeddings that are trained on a much larger corpus than MSCOCO.

		Standard	+CSK
SANDI-CV	*SemSim*	86.2	**86.3**
	ParaRank	59.9	59.7
SANDI-MAN	*SemSim*	85.1	**85.5**
	ParaRank	53.8	**55.0**

Table 5.4: Role of Commonsense Knowledge.

Tag Space	Precision	Random	NN	VSE++	SANDI
CV	*Strict*	0.4	2.0	1.14	**4.18**
	Relaxed	42.16	52.68	29.83	**53.54**
MAN	*Strict*	0.4	3.95	-	**14.57**
	Relaxed	37.14	42.73		**49.65**
BD	*Strict*	0.4	1.75	-	**2.71**
	Relaxed	32.59	37.94		**38.86**
∗	*Strict*	0.4	4.8	-	**11.28**
	relaxed	43.84	50.06		**54.34**

Table 5.5: Image Selection on the Lonely Planet dataset.

On both datasets, SANDI-MAN is the single best configuration, while the combination, SANDI∗ marginally outperforms it on the Asia Exchange dataset. The similarity of scores across both datasets highlights the robustness of the SANDI approach.

Role of Commonsense Knowledge. While in alignment-sensitive articles the connections between paragraphs and images are often immediate, this is less the case for articles with low alignment sensitivity. Table 5.4 shows the impact of adding common sense knowledge on the 100 least alignment sensitive articles from the Lonely Planet dataset. As one can see, adding CSK tags leads to a minor improvement in terms of semantic similarity (+0.1/+0.4%), although the improvement is too small to argue that CSK is an important ingredient in text-image alignments.

Tag Space	Precision	Random	NN	VSE++	SANDI
CV	*Strict*	0.45	0.65	0.44	**0.79**
	Relaxed	55.0	**57.64**	30.05	57.2
MAN	*Strict*	0.45	0.78	-	**3.42**
	Relaxed	40.24	52.0		**52.87**
BD	*Strict*	0.45	0.82	-	**0.87**
	Relaxed	31.12	**33.27**		33.25
∗	*Strict*	0.45	1.04	-	**1.7**
	relaxed	55.68	58.1		**58.2**

Table 5.6: Image Selection on the Asia Exchange dataset.

5.6.3 Selective Alignment

This variation of our model, as defined in Section 5.4.2, solves two problems simultaneously – selection of representative images for the story from a big pool of images, and placement of the selected images within the story. The former sub-problem relates to the topic of "Story Illustration" [70, 137, 37, 127], but work along these lines has focused on very short texts with simple content (in contrast to the long and content-rich stories in our datasets).

6.3.1 Image Selection

Setup. In addition to the setup described in Section 5.6.1, following are the requirements for this task:

- Image pool – We pool images from all stories in the slice of the dataset we use in our experiments. Stories from a particular domain – for e.g. travel blogs from Lonely Planet – are largely quite similar. This entails that images in the pool may also be very similar in content – for e.g., stories on *hiking* contain images with similar tags like *mountain, person, backpack*.
- Image budget – For each story, the number of images in the ground truth is considered as the image budget b for Image Selection (Equation 5.4.2).

Baselines. We compare SANDI with the following baselines:

- Random: a baseline of randomly selected images from the pool.
- NN: a selection of nearest neighbors from a common embedding space of images and paragraphs. Images are represented as centroid vectors of their tags, and paragraphs are represented as centroid vectors of their distinctive concepts. The basic vectors are obtained from Word2Vec trained on Google News Corpus.
- VSE++: state-of-the-art on joint visual-textual embeddings; the method presented in [47] is adapted to retrieve the top-b images for a story.

Evaluation. We evaluate Image Selection by the measures in Section 5.5.1. Table 5.5 and Table 5.6 show the results for Story Illustration, that is, image selection, for SANDI and the baselines. For the Lonely Planet dataset (Table 5.5), a pool of 500 images was used. We study the effects of a bigger images pool (1000 images) in our experiments with the Asia Exchange dataset (Table 5.6). As expected, average strict precision (exact matches with ground truth) drops. Recall from Section 5.3.1 that the Asia Exchange dataset often has stock images for general illustration rather than only story-specific images. Hence the average relaxed precision on image selection is higher. The nearest-neighbor baseline (NN) and SANDI, both use Word2Vec embeddings for text-image similarity. SANDI's better scores are attributed to the joint optimization over the entire story, as opposed to

	BLEU	ROUGE	SemSim	ParaRank
Random	0.31	0.26	69.18	48.16
VSE++ [47]	1.04	0.8	79.18	53.09
VSE++ ILP	1.23	1.03	79.04	53.96
SANDI-CV	1.70	1.60	83.76	61.69
SANDI-MAN	**8.82**	**7.40**	82.95	66.83
SANDI-BD	1.77	1.69	**84.66**	**76.18**
SANDI*	6.82	6.57	84.50	75.84

Table 5.7: Selective Alignment on the Lonely Planet dataset.

Story

Clatter into Lisbon's steep, tight-packed Alfama aboard a classic yellow tram...England. Ride a regular bus for a squeezed-in-with-the-natives view of the metropolis...Venice, Italy...opting for a public vaporetto (water taxi) instead of a private punt...Hungary. Trundle alongside the Danube, with views up to the spires and turrets of Castle Hill...Istanbul, Turkey...Travel between Europe and Asia...Ferries crossing the Bosphorus strait...Sail at sunset...Monte Carlo's electric-powered ferry boats...The 'Coast Tram' skirts Belgium's North Sea shoreline...Pretty but pricey Geneva...travel on buses, trams and taxi-boats...Liverpool, England...Hop aboard Europe's oldest ferry service...just try to stop yourself bursting into song.

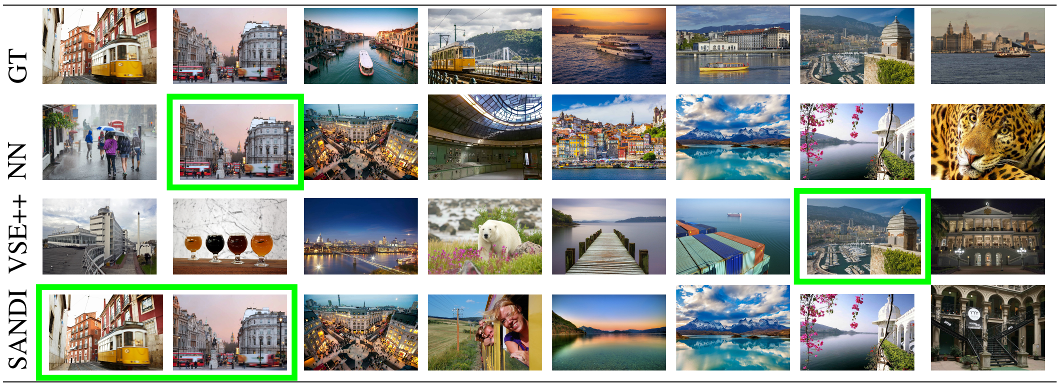


Figure 5.4: Image Selection. Images within green boxes are exact matches with ground truth (GT). SANDI retrieves more exact matches than the baselines (NN, VSE++). SANDI's non-exact matches are also much more thematically similar.

greedy selection in case of NN. VSE++ uses a joint text-image embeddings space for similarity scores. The results in the tables clearly show SANDI's advantages over the baselines.

Our evaluation metric $RelaxedPrecision$ (Eq. 5.7) factors in the semantic similarity between images which in turn depends on the image descriptors (Section 5.3.2). Hence we compute results on the different image tag spaces, where '$*$' refers to the combination of CV, MAN, and BD. The baseline VSE++ however, operates only on visual features; hence we report its performance only for CV tags.

Figure 5.4 shows anecdotal image selection results for one story. The original story contains 17 paragraphs; only the main concepts from the story have been retained in the figure for readability. SANDI is able to retrieve 2 ground-truth images out of 8, while the baselines retrieve 1 each. Note that the remaining non-exact matches retrieved by SANDI are also thematically similar. This can be attributed to the wider space of concepts that SANDI explores through the different types of image descriptors described in Section 5.3.2.

6.3.2 Image Placement

Having selected thematically related images from a big image pool, SANDI places them within contextual paragraphs of the story. Note that SANDI actually integrates the Image Selection and Image Placement stages into joint inference on selective alignment seamlessly, whereas the baselines operate in two sequential steps.

We evaluate the alignments by the measures from Section 5.5.2. Note that the measure *OrderPreserve* does not apply to Selective Alignment since the images are selected from a pool of mixed images which cannot be ordered. Tables 5.7 and 5.8 show results for the Lonely Planet and Asia

Exchange datasets respectively. We observe that SANDI outperforms the baselines by a clear margin, harnessing its more expressive pool of tags. This holds for all the different metrics (to various degrees). We show anecdotal evidence of the diversity of our image descriptors in Figure 5.3 and Table 5.10.

	BLEU	ROUGE	SemSim	ParaRank
Random	2.06	1.37	53.14	58.28
VSE++ [47]	2.66	1.39	58.00	64.34
VSE++ ILP	2.78	1.47	57.65	64.29
SANDI-CV	1.04	1.51	60.28	75.42
SANDI-MAN	**3.49**	**2.98**	61.11	**82.00**
SANDI-BD	1.68	1.52	**76.86**	70.41
SANDI∗	1.53	1.84	64.76	80.57

Table 5.8: Selective Alignment on the Asia Exchange dataset.

5.6.4 Role of Model Components

Image Descriptors. Table 5.10 shows alignments for a section of a single story from three SANDI variants. Each of the variants capture special characteristics of the images, hence aligning to different paragraphs. The paragraphs across variants are quite semantically similar. The highlighted key concepts bring out the similarities and justification of alignment. The wide variety of image descriptors that SANDI leverages (CV, BD, MAN, CSK) is unavailable to VSE++, attributing to the latter's poor performance.

Embeddings. The nature of embeddings is decisive towards alignment quality. Joint visual-semantic-embeddings trained on MSCOCO (used by VSE++) fall short in capturing high-level semantics between images and story. Word2Vec embeddings trained on a much larger and domain-independent Google News corpus better represents high-level image-story interpretations.

ILP. Combinatorial optimization (Integer Linear Programming) wins in performance over greedy optimization approaches. In Tables 5.5 and 5.6 this phenomenon can be observed between NN (greedy) and SANDI (ILP). This pair of approaches make use of the same embedding space, with SANDI outperforming NN.

5.7 *Illustrate Your Story*: A web application for SANDI

We build an end-to-end system for automatically selecting relevant images from an album and placing them in suitable contexts within a body of text. The application solves a global optimization problem that maximizes the coherence of text paragraphs and image tags, and allows for exploring explanations for the alignments. In addition to the Complete and Selective alignments discussed in Section , here we propose a functionality that ensures greater visual appeal of the generated multimodal content. This functionality, named *Spacing-aware Alignment*, aims to distribute images as uniformly as possible through the story while ensuring semantic coherence with the alignment

	BLEU	ROUGE	SemSim	ParaRank	Order Preserve	Spacing
Random	3.1	6.9	75.1	50.0	50.0	78.6
VSE++[47] ILP	12.6	11.2	84.0	58.1	47.9	79.1
Complete Alignment	45.6	44.5	89.8	72.5	77.4	84.0
Spacing-aware Alignment	43.3	41.5	89.5	71.5	75.5	89.1

Table 5.9: Influence of enforced image-spacing on the Lonely Planet dataset: slightly reduced image-text semantic coherence.

paragraphs. An automated system that can successfully select and align images to text will be useful to a multitude of end users like journalists, bloggers, authors, and commercial enterprises. Experiments show that our method can align images with texts with high semantic fit, and to user satisfaction.

Contributions. To the best of our knowledge, this is the first implementation of the story–image alignment problem. Additionally, we argue that spacing of images within a story is important for the visual appeal of the generated multimodal content, and propose a new model to that effect. We integrate a image captioning system which retrieves, for each image, the most fitting quote from a dataset of famous quotes. We also study the inherent subjectivity of the task with a suitable user study.

5.7.1 Spacing-aware Alignment

We observe from the Lonely Planet and Asia Exchange datasets that images are usually uniformly distributed throughout the story. For more visual appeal, we add placement constraints that spread images as evenly as possible across the story.

For a story with T paragraphs and I images, the number of paragraphs between successive images, m, is bounded as follows:

$$\left\lfloor \frac{T+1}{I+1} \right\rfloor \leq m \leq \left\lceil \frac{T-1}{I-1} \right\rceil \tag{5.12}$$

Additional constraints for the ILP restrict the distances between neighboring images:

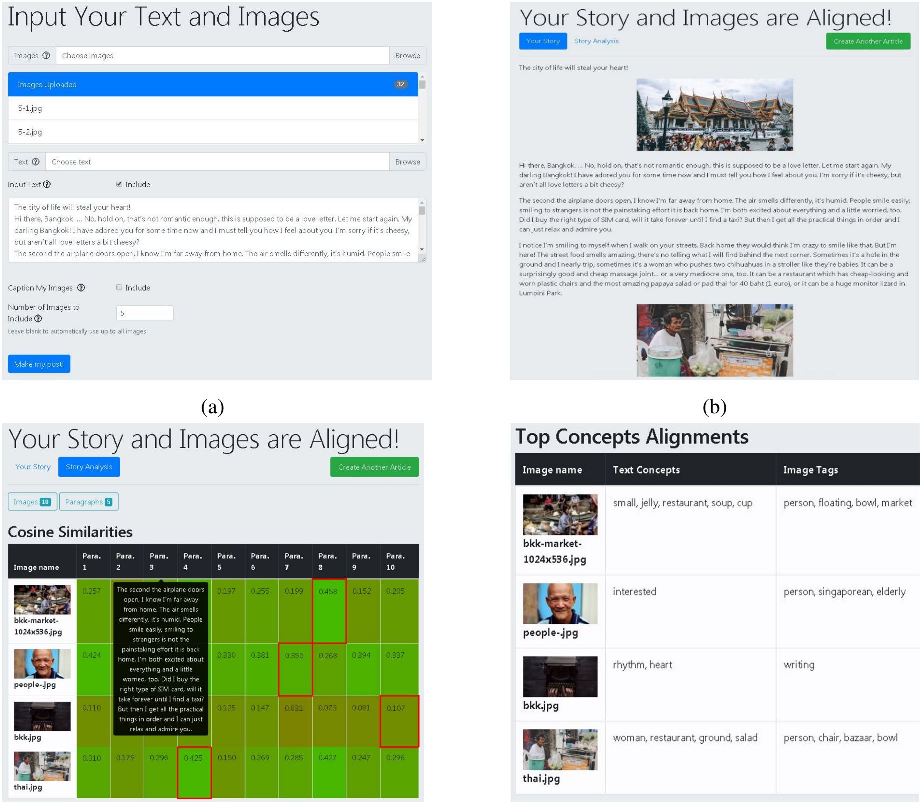
$$\sum_i \sum_{s=0}^{u-1} X_{i,t+s} \geq 1 \forall t \tag{5.13}$$

$$\sum_i \sum_{s=0}^{l-1} X_{i,t+s} \leq 1 \forall t \tag{5.14}$$

where $u = \left\lceil \frac{T-1}{I-1} \right\rceil$ and $l = \left\lfloor \frac{T+1}{I+1} \right\rfloor$.

The constraint in (5.13) ensures that there are at most u paragraphs between two images, while the constraint in (5.14) ensures that there are at least l paragraphs between two images.

Evaluating Spacing-aware Alignments. In addition to the evaluation metrics discussed in Section 5.5, we define another automatic evaluation metric in order to evaluate our Spacing-aware Alignment model.



Input Your Text and Images

Images ⓘ Choose images Browse

Images Uploaded 32

5-1.jpg

5-2.jpg

Text ⓘ Choose text Browse

Input Text ⓘ ☑ Include

The city of life will steal your heart!
Hi there, Bangkok. ... No, hold on, that's not romantic enough, this is supposed to be a love letter. Let me start again. My darling Bangkok! I have adored you for some time now and I must tell you how I feel about you. I'm sorry if it's cheesy, but aren't all love letters a bit cheesy?
The second the airplane doors open, I know I'm far away from home. The air smells differently, it's humid. People smile

Caption My Images! ⓘ □ Include

Number of Images to Include ⓘ 5

Leave blank to automatically use up to all images

Make my post!

(a)

Your Story and Images are Aligned!

Your Story Story Analysis Create Another Article

The city of life will steal your heart!

Hi there, Bangkok. ... No, hold on, that's not romantic enough, this is supposed to be a love letter. Let me start again. My darling Bangkok! I have adored you for some time now and I must tell you how I feel about you. I'm sorry if it's cheesy, but aren't all love letters a bit cheesy?

The second the airplane doors open, I know I'm far away from home. The air smells differently, it's humid. People smile easily; smiling to strangers is not the painstaking effort it is back home. I'm both excited about everything and a little worried, too. Did I buy the right type of SIM card, will it take forever until I find a taxi? But then I get all the practical things in order and I can just relax and admire you.

I notice I'm smiling to myself when I walk on your streets. Back home they would think I'm crazy to smile like that. But I'm here! The street food smells amazing, there's no telling what I will find behind the next corner. Sometimes it's a hole in the ground and I nearly trip, sometimes it's a woman who pushes two chihuahuas in a stroller like they're babies. It can be a surprisingly good and cheap massage joint... or a very mediocre one, too. It can be a restaurant which has cheap-looking and worn plastic chairs and the most amazing papaya salad or pad thai for 40 baht (1 euro), or it can be a huge monitor lizard in Lumpini Park.

(b)

Your Story and Images are Aligned!

Your Story Story Analysis Create Another Article

Images 10 Paragraphs 5

Cosine Similarities

Image name	Para. 1	Para. 2	Para. 3	Para. 4	Para. 5	Para. 6	Para. 7	Para. 8	Para. 9	Para. 10
bkk-market-1024x536.jpg	0.257				0.197	0.255	0.199	0.458	0.152	0.205
people-.jpg	0.424				0.330	0.381	0.350	0.268	0.394	0.337
bkk.jpg	0.110				0.125	0.147	0.031	0.073	0.081	0.107
thai.jpg	0.310	0.179	0.296	0.425	0.150	0.269	0.285	0.427	0.247	0.296

The second the airplane doors open, I know I'm far away from home. The air smells differently, it's humid. People smile easily; smiling to strangers is not the painstaking effort it is back home. I'm both excited about everything and a little worried, too. Did I buy the right type of SIM card, will it take forever until I find a taxi? But then I get all the practical things in order and I can just relax and admire you.

(c)

Top Concepts Alignments

Image name	Text Concepts	Image Tags
bkk-market-1024x536.jpg	small, jelly, restaurant, soup, cup	person, floating, bowl, market
people-.jpg	interested	person, singaporean, elderly
bkk.jpg	rhythm, heart	writing
thai.jpg	woman, restaurant, ground, salad	person, chair, bazaar, bowl

(d)

Figure 5.5: Our system takes an image collection and a body of text as input (a), and generates a multimodal story (b). Story analysis: the matrix in (c) shows cosine similarities between images and paragraphs, highlighting the alignments with red boxes, hovering over the column headers show the corresponding paragraphs. Similar concepts from aligned images and paragraphs can be seen in (d).

Ideal alignments would space out images evenly within a story. We propose to measure the deviation from such an ideal alignment as an indication of visual appeal. For a story with T paragraphs and I (selected) images, the ideal relative distance between two images would be $1/(|I|-1)$. $\mathbb{T} = 1, 2, ..., t$ is the set of ordered paragraphs, where $|\mathbb{T}| = t$. We define the allocation of images to paragraphs via a function $f : I \to \mathbb{T}$ defined as $f(i) = j$ if and only if $X_{ij} = 1$. X_{ij} is a binary random variable from (5.2). We now look at the co-domain of the function $\mathbb{T}' \in \mathbb{T}$, and sort its elements $1 \leq j_1 \leq j_2 \leq ...j_{|I|} \leq t$. The average image spacing variance of a story is then defined as:

$$Spacing(X) = 1 - \left[\frac{1}{|I|-1} \sum_{k=1}^{|I|-1} \left| \frac{j_{k+1} - j_k}{t-1} - \frac{1}{|I|-1} \right| \right] \tag{5.15}$$

We compare Spacing-aware Alignment with Complete Alignment and baselines (VSE++[47] ILP and a random alignment) on all the defined evaluation metrics. The results can be seen in Table 5.9; all values are scaled to [0,100]. The Spacing-aware Alignment ensures a more uniform image spacing in the story, while sacrificing slightly on semantic coherence of image–paragraph pairs. This can be justified as an acceptable trade-off since images are often just used for general illustration of the overall theme of the story, without maintaining a tight semantic fit with surrounding paragraphs. The

$Spacing$ of the ground truth alignments from the Lonely Planet dataset is 81.03.

5.7.2 System Overview

The computational model of the system (SANDI) consists of multiple image tags, word embeddings to represent concepts from images and paragraphs, and a combinatorial optimization problem solver.

Input. SANDI takes the following mandatory user inputs – an image collection and a body of text. **Output.** After computation of the image–paragraph alignments, SANDI displays the generated multimodal story.

Interactive Exploration. Let us consider a blogger writing about their last vacation trip. They took over 100 pictures during the trip, but would only like to include 5 representative ones in their trip report. With our system, they can save the time and effort of going through all the images and selecting and placing them within appropriate paragraphs of the textual narrative. Our system automatically performs Image Selection and Image Placement in one seamless optimization step.

In the home page (Figure 5.5 (a)), users are able to upload a few images or a big pool of images, enter/upload text, specify the number of images to be selected for the story, and specify their choice to automatically caption the images. Once the user uploads their files and their selections, a session is created in the back-end with a session-ID in case any further user interaction is required. This will arise when our system fails to automatically detect visual concepts from the images, and asks for user-specified image tags. Figure 5.5 (b) shows the output page, where the generated multi-modal story is displayed to the user. Additionally, a Story Analysis page provides the justifications for individual image–paragraph alignments. This is in the form of image–paragraph cosine similarities (Figure 5.5 c), and similar concepts from the two modalities (Figure 5.5 d).

The initial requests from the front-end, session handling, as well as automatic image tagging are all performed in a Flask server written in Python. The optimization is handled in a Java Servlets server to take advantage of its speed. We use the Bootstrap CSS framework to create an uncluttered and modern looking user interface.

Caption Prediction. Most images from the Lonely Planet and Asia Exchange datasets are captioned. Following this characteristic of multimodal narratives, we add an image captioning component in our application. While automatic caption generation is a well-studied problem [12], we resort to quote-based caption prediction to obtain more inspiring results. Using a visual-semantic embedding framework [47] to infer text–image similarities, we predict the top-10 related quotes for an image from the Quotes-500K dataset [56]. Among these, the quote with the highest cosine similarity to the aligned paragraph is then displayed as the image caption. Hence, our image captions are not only attractive, but also contextually meaningful. Figure 5.6 shows some examples.

5.7.3 User Study

The story-specific selection of images from an image repository and their placement within the story is to a certain extent a matter of subjective preferences. While the problem has been formally characterized and evaluated using automated means in Section 5.6, we study the inherent subjectivity that it entails through a user evaluation of image–paragraph alignments. We use a random selection of images from both datasets – Lonely Planet and Asia Exchange.

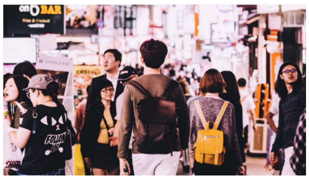
Image and Caption	Aligned Paragraph
 “What a rebellious act it is to love yourself naturally in a world of fake appearences.”	Watch out for the Korean wave! The catchy beats, colorful soap operas and gripping dramas are invading countries around the world with a massive force, and are here to stay. And while before it targeted the youger crowds, it’s now getting more and more popular among grown-up folks too!
 “All the world’s a stage and all men and women are merely players.”	No genre of media is excluded: Film, literature, graphic novels, language, food, fashion. . . But arguably, the genre with the biggest global impact is the new wave of Korean pop music, commonly referred to as K-pop, with its addicting melodies and innovative choreography.

Figure 5.6: SANDI predicts contextually meaningful captions.

The user study is designed to collect feedback, on a per-image basis, of which aligned paragraph – from the ground truth (GT) or from our system (SANDI) – is a better semantic fit for an image. We conduct the user study via Figure Eight (formerly, CrowdFlower).

Design. Each question consists of an image and two paragraphs (A and B). Contributors are asked to choose one of the following options – A More Relevant Than B, A Equally Relevant to B, A Less Relevant Than B – as answer to the question “Which text is better fitting with the given image?”. We collect 5 judgments per question for a total of 250 questions.

Avoiding Bias. The source of the paragraph (GT or SANDI) is not revealed to the contributor. The assignment of the paragraphs (GT to A, SANDI to B etc.) is randomized to eliminate bias. Moreover, data points are chosen such that GT and SANDI paragraphs are of similar length – a difference of maximum 20 words is allowed. This is done to avoid possible bias towards longer or shorter paragraphs.

Quality Assurance. Test questions are modeled such that one of the paragraphs belong to an unrelated story. This allows us to eliminate responses from inattentive contributors. The confidence score (pink horizontal bars in Figure 5.7) for each aggregated result depicts the level of agreement between multiple contributors.

Result Aggregation. For each question, the option selected by the majority of the contributors is reported. As shown in Figure 5.7, SANDI alignments were chosen to be more relevant 46.4% of times, whereas GT was chosen 40.6% of times. This shows that while GT alignments are generally chosen with care, SANDI alignments are even more semantically relevant. Both paragraphs were deemed equally relevant in 13.4% of the questions.

The observations from the user study support our hypothesis that the problem has a subjective component – the alignments in the ground truth are not absolute, and there exist other suitable alignments – which our application enables exploring.

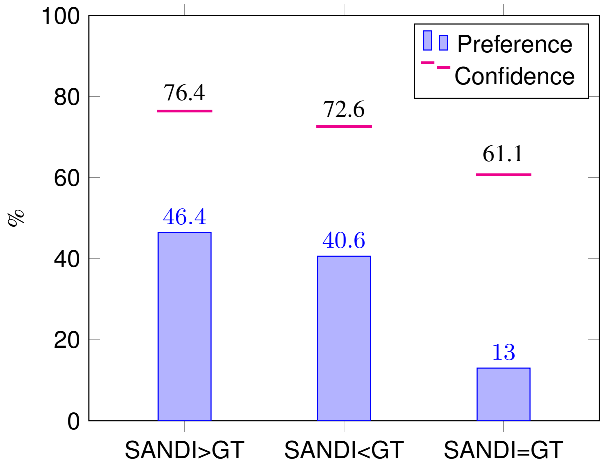


Figure 5.7: User choice of ground truth (GT) and SANDI alignments.

5.8 Discussion

5.8.1 Text-Image Discourse Relations

In unimodal contents like text documents, two sentences have explicit or implicit connections between them as studied in the Penn Discourse Treebank [111, 123]. Similarly, in multimodal contents, each of the modalities is expected to have certain contribution to the overall cognition of the content. Moreover, adjoining snippets from different modalities are anticipated to be coherent in context, one often leading to a better understanding of the other.

Naturally, for text-image alignment, it is crucial to understand the synergy between the modalities. We briefly study the coherence in text-image contexts and make observations about the role of the individual modalities. Coherence in text is characterized by distinct structures in grammar. Images lack such structures and contain largely incidental details that connect them to surrounding textual narratives. Therefore, identifying the distinctive image content that the author intended to depict is challenging [148]. Our study of multimodal instructional content (a dataset of cooking recipes) [177] provide some insights into inferential relationships between images and surrounding text [5]. Instead of proposing a taxonomy of discourse relations like the Penn Discourse Treebank, we conduct a thorough user evaluation to understand various relations between image and text – for e.g., whether the image content complementary or supplementary to the text, whether there is a temporal relationship between the events being narrated in the text and that in the supporting image, and so on. We believe that this is an important first step towards understanding the intent of aligning an image to a paragraph in a story or narrative. Follow-up research may benefit from the annotated resource that this work generated. Prior work on multimodal discourse relations studied the coherence between gesture and speech [78].

5.8.2 Future Research

Here are a few other questions that arise from this work that could encourage further research:

Additional Features. While our feature space covers most natural aspects, in downstream applications additional image metadata such as GPS location or timestamps may be available. GPS location may provide cues for geographic named entities, lifting the reliance on user-provided tags. Timestamps might prove to be useful for temporal aspects of a story-line.

Abstract and Metaphoric Relations. Our text-image alignments were focused largely on visual and contextual features. We do not address stylistic elements like metaphors and sarcasm in text, which would entail more challenging alignments. For example, the text "the news was a dagger to his heart" should not be paired with a picture of a dagger. Although user knowledge may provide some cues towards such abstract relationships, a deeper understanding of semantic coherence is desired.

Subjectivity of the Ground Truth. Based on our analysis, the articles from Lonely Planet and Asia Exchange provide only a pseudo ground truth, as other meaningful alignments exists (as seen in our user study). An interesting direction might be to explore datasets which only allow a single truth.

Image Caption Generation. From our observation of the real-world datasets, most images are accompanied by captions. Having aligned individual images to text units, a natural extension would be to generate captions for the images which are contextually meaningful to the surrounding paragraphs.

In this work we have introduced the problem of story-images alignment – selecting and placing a set of representative images for a story within contextual paragraphs. We analyzed features towards meaningful alignments from two real-world multimodal datasets – Lonely Planet and Asia Exchange blogs – and defined measures to evaluate text-image alignments. We presented SANDI, a methodology for automating such alignments by a constrained optimization problem maximizing semantic coherence between text-image pairs jointly for the entire story. In addition to ensuring a high level of semantic fit of images and aligned paragraphs, our web application also guarantees uniform spacing of images for better visual appeal. Quantitative evaluations show that SANDI produces alignments which are semantically meaningful. Our user evaluation corroborates that the quality of the obtained image–paragraph alignments are comparable to human judgments. We believe that such an application will be of assistance to online content creators such as bloggers, journalists, commercial content writers, as well as for creation of personal social media posts.

SANDI web application is available at `https://sandi.mpi-inf.mpg.de`, while a video demonstration can be viewed at `https://youtu.be/k5gu2pNxdNU`.

Image and detected concepts	SANDI-CV	SANDI-MAN	SANDI-BD
CV: cottage, flower-pot, carrier, MAN: Bilbo Baggins, Shire, BD: New Zealand, hobbit house	Take advantage of your stay here and visit the memorable scenes shown in the movies. Visit The Shire and experience firsthand the 44 Hobbit Holes from which Bilbo Baggins emerged to commence his grand adventure. Tongariro National Park is home to the feared Mount Doom in The Lord of the Rings. Other famous locations that you can visit are Christchurch, Nelson and Cromwell.	Home to hobbits, warriors, orcs and dragons. If you're a fan of the famous trilogies, Lord of the Rings and The Hobbit, then choosing New Zealand should be a no-brainer.	Take advantage of your stay here and visit the memorable scenes shown in the movies. Visit The Shire and experience firsthand the 44 Hobbit Holes from which Bilbo Baggins emerged to commence his grand adventure. Tongariro National Park is home to the feared Mount Doom in The Lord of the Rings. Other famous locations that you can visit are Christchurch, Nelson and Cromwell.
CV: snowy mountains, massif, alpine glacier, mountain range MAN: outdoor lover, New Zealand, study destination, BD: New Zealand	New Zealand produced the first man to ever climb Mount Everest and also the creator of the bungee-jump. Thus, it comes as no surprise that this country is filled with adventures and adrenaline junkies.	Moreover, the wildlife in New Zealand is something to behold. Try and find a Kiwi! (the bird, not the fruit). They are nocturnal creatures so it is quite a challenge. New Zealand is also home to the smallest dolphin species, Hector's Dolphin. Lastly, take the opportunity to search for the beautiful yellow-eyed penguin.	Home to hobbits, warriors, orcs and dragons. If you're a fan of the famous trilogies, Lord of the Rings and The Hobbit, then choosing New Zealand should be a no-brainer.
CV: cup, book, knife, art gallery, cigarette holder, MAN: New Zealand, educational rankings, BD: books	Study in New Zealand and your CV will gain an instant boost when it is time to go job hunting.	The land of the Kiwis consistently tops educational rankings and has many top ranking universities, with 5 of them in the top 300 in the world. Furthermore, teachers are highly educated and more often than not researchers themselves. Active participation, creativity and understanding of different perspectives are some of the many qualities you can pick up by studying in New Zealand.	New Zealand has countless student-friendly cities. Therefore, it should hardly come as a surprise that New Zealand is constantly ranked as a top study abroad destination. To name but a few cities: Auckland, North Palmerston, Wellington and Christchurch all have fantastic services and great universities where you can study and live to your heart's content.

Table 5.10: Example alignments. Highlighted texts show similar concepts between image and aligned paragraphs.

CHAPTER 6

Contextual Image Captioning

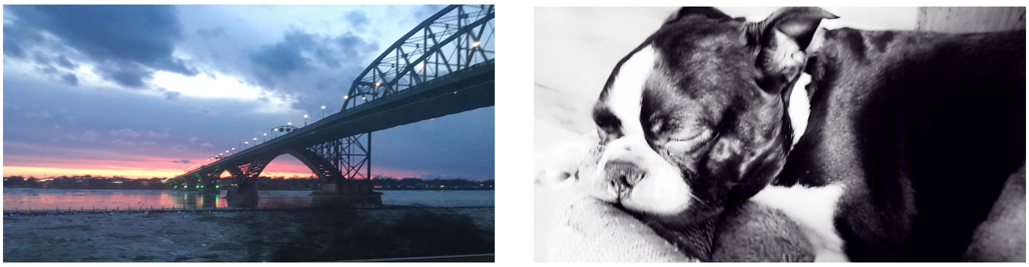
Contents

MODERN web content – news articles, blog posts, educational resources, marketing brochures – is predominantly multimodal. A notable trait is the inclusion of media such as images placed at meaningful locations within a textual narrative. Most often, such images are accompanied by captions – either factual or stylistic (humorous, metaphorical, etc.) – making the narrative more engaging to the reader. While standalone image captioning has been extensively studied, captioning an image based on external knowledge such as its surrounding text remains under-explored. In this paper, we study this new task: given an image and an associated unstructured knowledge snippet, the goal is to generate a *contextual caption* for the image.

6.1 Introduction

In multimodal (text–image) documents, images are typically accompanied by captions. These may, for instance, provide specific details about the narrative – location, names of persons etc. – or may be thematic comments grounding the sentimental value of the image in the narrative. The image captions explicitly or implicitly refer to the image and its surrounding text, and play a major role in engaging the reader. We call this type of captions *contextual captions*. To study the nature and automatic generation of *contextual captions*, we create a dataset from Reddit posts with images, titles and associated comments. Figure 6.1 illustrates the *Contextual Image Captioning* problem.

Generating captions for standalone images [66, 164] or summarizing a detailed piece of text [139, 88] are well-studied problems. However, generating image captions accounting for contextual information is a largely unexplored task and poses many challenges. Related tasks include multimodal summarization [18, 19] and title generation [114]. Multimodal summarization usually involves segmentation and sorting of both the modalities or has specific templates along which the summary



I recently moved to Buffalo, NY and every day I am discovering how beautiful this town is. I took this pic...and I was thrilled about it! I wanted to share the pallet of colors the sunset had that evening.

Generated Contextual Captions:

- *A beautiful sunset path to heaven.*
- *A sunset...unknown artist.*

RIP Echo. There will be sunny days and endless treats for him now. I still have the dog tag of my Boston Terrier on my key chain. Echo will always be with you. My condolences. So sorry for your loss. RIP.

Generated Contextual Captions:

- *Had to say goodbye to my best friend today . He passed away shaking.*
- *My best friend passed away yesterday. I think he was a fantastic.*

Figure 6.1: The *Contextual Image Captioning* problem. Novel captions generated by (variants of) our model for the shown image and its associated paragraph.

is generated [139]. In contrast, captions in our dataset are part of the story told by the comments. Generating the caption requires reasoning on the flow of the story and conditionally deciding to follow, lead or negate the context.

Inadequacy of Prior Work. Image captioning and text summarization are unimodal, and ignore important information in the dormant modality. Multimodal summarization and news image captioning [13] usually entail captions with explicit references to the context, and may be achieved with a copy mechanism [57] that can selectively copy information from paragraph to caption, e.g., named entities such as names of people and organizations, geographical locations, etc. However, most social media driven content is affective and requires implicit reasoning about the context. For example, for an image of the Grand Canyon, we might encounter captions such as “perfect for a lovely hike” or “too tired to walk?”, due to the subjectivity of the task, which requires *inference* based on the context.

Approach and Contribution.

- We formulate the novel task of *Contextual Image Captioning*.
- We create a new dataset for *contextual captions* from Reddit posts and associated comments.
- We propose an end-to-end trained neural model for *Contextual Image Captioning* and comprehensively evaluate its performance using quantitative and qualitative measures.
- We study how various factors affect the generation of novel *contextual captions*.

6.2 Related Work

Image Caption and Description Generation. Previous research on captioning conditioned only on images [49, 162, 71, 176, 76, 43, 48, 102] has been successful for descriptive captions with explicit grounding to image objects and scenes (e.g., “a boy is playing with a doll.”). Only recently, captions with sentimental and abstract concepts have been explored [52, 17, 119, 91, 141]. However, these ignore relevant text data surrounding images in social media and other domains. We aim to study a multimodal approach – given an image and adjacent text, generate a contextual caption.

Multimodal Summarization. Research on multimodal embeddings [77, 174, 138] has facilitated studying image–text data. Summarization of multimodal documents (e.g., blog/news articles) [30, 189] proceeds by aligning a subset of images with extracted [19], or generated [18] text segments from the original document. In contrast, image captions in our data do not explicitly summarize the associated text and rather act as a short commentary that connects the two modalities.

News Image Captioning. A task similar to our problem is captioning images within news articles [155, 20]. A key challenge here is to correctly identify and generate named entities [155]. It is similar to our problem in that generated concepts in the captions cannot be inferred from images alone and require contextual information from the associated text. However, news image captions tend to be descriptive compared to the subjective nature of captions in our dataset representing common social media content.

6.3 Datasets

To the best of our knowledge, the only existing image–text caption datasets are from the news domain (e.g., Daily Mail Corpus). Our task differs from news image captioning in two major ways. Firstly, we aim to generate image captions conditioned on its immediate context – a text paragraph. Most news datasets (e.g., the Daily Mail Corpus) do not contain such paragraph structure. Also, there is no clear image–text alignment. Moreover, news image captions are descriptive (containing specific information) and not particularly engaging (containing sentimental expressions). Thus, news datasets are not suitable for our task. Data from multimodal blog articles such as Lonely Planet offer a clean structure – one image aligned to a single paragraph. The paragraph containing the image provides meaningful context. Most images are accompanied by engaging captions. Such a dataset has been recently proposed [26]. However, it contains only 10,000 samples – too little to train a deep text generation model. Instead, we consider Reddit, which offers a rich source of multimodal data. Out of the image-related subreddits, /r/pics is particularly suitable for our problem because of the nature of posts. Firstly, the posts do not contain expert jargon, unlike other subreddits like /r/photographs. Secondly, the image captions are mostly affective and not drab descriptions. Lastly, post frequency is high, presenting a big dataset.

Data Scraping. We scrape the subreddit /r/pics to collect a sizable number of posts over the span of a year. For each post, we grab the image, the post title, and 1-10 comments. We choose to consider the post title as ground truth caption since it is written by the image poster, ensuring a consistent and coherent intent. The comments are concatenated, preserving their tree structure, to serve as the associated paragraph. Inappropriate posts that do not adhere to community standards and were flagged by moderators are removed.

Data Characteristics. The collected images do not adhere to any particular subject or theme. The paragraphs are ~59.2 words long, and the captions are ~10.6 words long on an average.

In some posts, captions and paragraphs may contain different named entities (NE), making prediction of the ground truth NE difficult. For example, the caption "My friend and I are en route to the Grand Canyon" may be accompanied with the paragraph "Try to hike down to the Colorado. Also visit Zion National Park!" The NEs in the paragraph (Colorado, Zion) do not match that in the caption (Grand Canyon). Owing to this characteristic, we study two distinct variants of the dataset – one containing NEs and the other without NEs. We denote these variants as +NE and ~~NE~~, respectively.

The paragraphs sometimes exhibit topic drift, e.g., comments on the post on the Grand Canyon may include "I remember my last trip to India...we had spicy food!". Hence, we also study variants with ensured context overlap – one common word (ignoring stop words) between caption and paragraph. These variants are suffixed overlap – e.g. +NE-overlap.

We report experimental results on all these variants for 30,000/8000/8000 train/val/test splits. A small sample of the dataset can be found at `https://bit.ly/3eHLnMo`. Further details are as follows:

- Total number of samples: 242,767
- Samples with named entities (NE) in caption: 137,732 (56.82%)
- Samples with no NE in caption: 104,653 (43.18%)

We ensure a context overlap between paragraph and caption with the following splits:

- *+NE* samples with one common word between paragraph and caption: 50,730 (20.93%). These are named *+NE-overlap* in Table 6.1.
- *+NE* samples with two common words between paragraph and caption: 23,283 (9.61%).
- *~~NE~~* samples with one common word between paragraph and caption: 38,301 (15.80%). These are named *~~NE~~-overlap* in Table 6.1.
- *~~NE~~* samples with two common words between paragraph and caption: 15,070 (6.22%)

We use SpaCy to detect named entities in captions. SpaCy detects 18 kinds of named entities (`https://spacy.io/api/annotation#named-entities`). `TIME, MONEY, PERCENT,` and `LANGUAGE` were ignored since they include common conversational phrases like "day before yesterday", "my two cents", "an English breakfast" etc. Examples of captions with NEs: *"Just the (Earth/LOC) letting off some steam ((Iceland/GPE))"*, *"The (first/CARDINAL) Chipotle , opened in (Denver/GPE) in (1993/DATE)."*. Examples of captions without NEs: *"Texture of the paint on a skull I painted."*, *"My girlfriend and I handle social situations differently."*. In future work, the NE types could be leveraged to learn positional relationships in sentences.

6.4 Contextual Captioning Model

Figure 6.2 shows our Contextual Captioning model architecture.

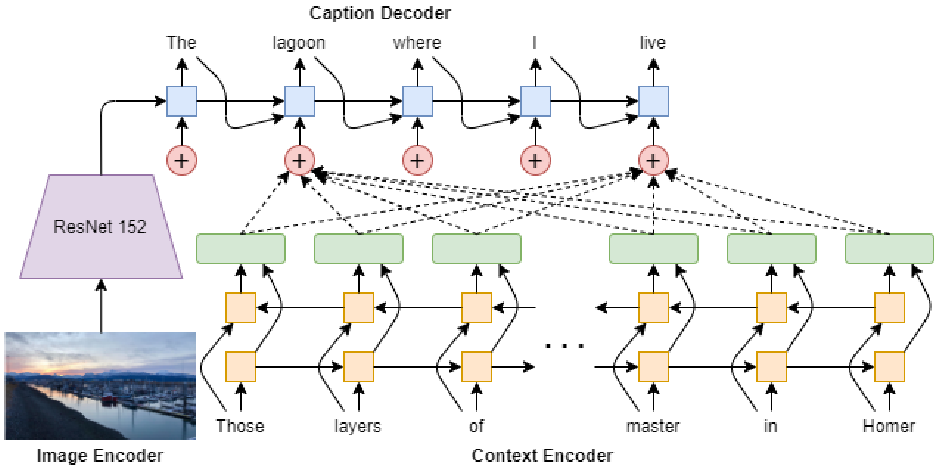


Figure 6.2: A schematic diagram of our model

Given an input image I and an associated input paragraph $P = \{w_1^p, \dots, w_M^p\}$ of length M, our model generates a caption $C = \{w_1^c, \dots, w_N^c\}$ of length N. Our model adopts a standard encoder–decoder architecture. Following prior work, we use features extracted from a pre-trained ResNet152 [62] model for image encoding. These extracted features are projected to a d-dimensional embedding space using a dense layer.

To encode the input paragraph, we deploy a bidirectional LSTM (BiLSTM). The outputs of the BiLSTM, denoted as $\mathbf{G} = \{\mathbf{g}_1, \dots, \mathbf{g}_M\}$, where $\mathbf{g}_i = \text{BiLSTM}(\mathbf{x}_i, \mathbf{g}_{i-1}) \forall i \in i \in \{1, \dots, M\}$, is the encoded representation of the input paragraph. $\mathbf{x}_i$ is the vector embedding of the word w_i^p.

We deploy a unidirectional LSTM for sequential decoding of the caption C that leverages both the encoded image and paragraph representations. The image embedding is provided as an input to the decoder LSTM at timestep $t = 1$. In all subsequent timesteps, the decoder input is the embedding $\mathbf{y}_{t-1}$ of the previous token w_{t-1}^c. The decoder state at each timestep t is obtained as $\mathbf{h}_t = \text{LSTM}(\mathbf{y}_{t-1}, \mathbf{h}_{t-1})$. To incorporate contextual information from the input paragraph, we concatenate an attention-weighted sum of the encoder states, denoted as $\tilde{\mathbf{G}}_t$, to the current state $\mathbf{h}_t$.

At each decoder time step $t \in \{2, \dots N\}$, the attention weights α^t over the encoder states depend on the current decoder state $\mathbf{h_t}$ and the encoder states $\mathbf{G}$. Formally,

$$\tilde{\mathbf{G}}_\mathbf{t} = \sum_{i=i}^{M} \alpha_i^t \mathbf{g_i} \tag{6.1}$$

$$\alpha_i^t = \frac{\mathbf{v}^\intercal(\mathbf{W_g g_i} + \mathbf{W_h h_t} + \mathbf{b})}{\sum_{i'=1}^{M} \mathbf{v}^\intercal(\mathbf{W_g g_{i'}} + \mathbf{W_h h_t} + \mathbf{b})} \tag{6.2}$$

Finally, we pass the concatenated output through two dense layers with a non-linear activation layer (e.g, ReLU) placed in between. The output logits are then passed through a Softmax function to obtain the output distribution $p(.)$ over the vocabulary. We train our model end-to-end by minimizing the negative log-likelihood, i.e., $\theta^* = \arg\min_\theta - \log p(C \mid I, P; \theta)$. Note that we obtain the input embeddings, $\mathbf{x_i}$, and $\mathbf{y_t}$, of the encoder and decoder respectively from the embedding layer of a pretrained $BERT_{\text{BASE}}$ model.

The model's objective is to learn the optimal parameters θ^* to maximize the log-likelihood $\log p(C|I, P; \theta)$. Therefore, we train our model end-to-end by minimizing the negative log-likelihood defined as:

$$\mathcal{L}(\theta) = \sum_{t=1}^{N} - \log p(w_t^c \mid w_1^c, \ldots, w_{t-1}^c, I, P; \theta) \tag{6.3}$$

6.5 Experiments and Results

6.5.1 Experimental Setup

Our architecture is developed in PyTorch. The number of samples in all train/val/test splits is 30,000/8000/8000. Each model is trained for 20 epochs, with a batch size of 16. On a Tesla V100-PCIE-16 GB GPU, training 1 epoch taken 8 min. For each model variant, the best validation model is used for testing. We experiment with models using pre-trained BERT token embeddings, as well as learning token embeddings from scratch (with a vocabulary size of 100,000). We observe that BERT token embeddings have a positive effect on the quality of captions (Figure 6.4), and hence consider this configuration as default.

6.5.2 Quantitative Evaluation

Metrics. We use MSCOCO [89] automatic caption evaluation tool[1] to quantitatively evaluate our proposed model variants. We report scores for the BLEU-1, ROUGE-L, CIDEr, and SPICE metrics. In addition, we also report scores for semantic similarity between ground truth (c_{gt}) and generated (c_{gen}) captions: $\text{SemSim}(c_{\text{gt}}, c_{\text{gen}}) = \text{cosine}(\mathbf{v}_{c_{\text{gt}}}, \mathbf{v}_{c_{\text{gen}}})$, where $\mathbf{v}_{c_{\text{gt}}}$ and $\mathbf{v}_{c_{\text{gen}}}$ are the mean vectors of constituent words in the respective captions from 300-dimensional GloVe embeddings.

Baselines. To the best of our knowledge, there is no existing work that studies contextual image captioning. Therefore, we present two baselines that can also be regarded as ablated versions of our model: Image-only and Text-only captioning.

Results. In Table 6.1, we report scores[2] for the baselines and our model variants. Recall from Section 6.3 that our models are based on various data splits: +NE, ~~NE~~, and their respective overlap variants. We observe that for the *+NE* split, contextual captions are not better than the unimodal baselines on n-gram overlap based scores. This can be attributed to the nature of the dataset: NEs in the paragraph differ from those in ground truth captions. Since contextual captions draw inference from the paragraphs, the predicted NEs differ from ground truth captions as well, leading to lower scores for n-gram overlap based metrics. For the ~~NE~~ splits as well as both the *overlap* splits, contextual captions fare better than the baselines.

The observed low scores for BLEU-1, ROUGE-L, CIDEr, and SPICE hint towards the subjectivity of the task. N-gram overlap based metrics do not accommodate varied interpretations and linguistic diversity. Figure 6.3 exemplifies how image-only captions for different images are often similar, while contextual captions (which utilized both textual and visual modalities) are linguistically richer.

[1] https://github.com/tylin/coco-caption

[2] The BLEU-1 and ROUGE-L scores are multiplied by 100, and CIDEr and SPICE scores are multiplied by 10 following the standard practice.

	BLEU-1	*ROUGE-L*	*CIDEr*	*SPICE*	*SemSim*
			+NE		
Image-only	**9.72**	**8.42**	0.42	0.18	**0.72**
Text-only	8.71	7.85	**0.68**	**0.29**	0.73
Contextual	7.94	7.82	0.50	0.17	0.71
			+NE-overlap		
Image-only	8.64	7.84	0.50	0.19	0.73
Text-only	8.34	7.48	0.53	0.20	0.73
Contextual	**10.13**	**9.57**	**0.84**	**0.31**	**0.75**
			~~NE~~		
Image-only	5.96	6.42	0.37	0.14	0.71
Text-only	5.36	5.29	0.30	0.16	0.68
Contextual	**6.37**	**6.93**	**0.45**	**0.19**	**0.72**
			~~NE~~-overlap		
Image-only	7.80	7.50	0.38	0.16	0.76
Text-only	6.87	6.54	0.61	0.36	0.72
Contextual	**9.30**	**9.68**	**0.78**	**0.50**	**0.77**

Table 6.1: Quantitative Evaluation of baselines and Contextual Captioning on different data splits

High average $SemSim$ scores of contextual captions are indicative of their thematic similarity with the ground truth. Note that the splits with enforced similarity (*-overlap*) between paragraph and caption fare better on $SemSim$, leading to the conjecture that with a cleaner dataset, it would be possible to generate very relevant contextual captions.

Validation Performances of Test Models. We train each model for 20 epochs and chose the best validation model for testing. Here, we report the validation losses of our reported test models.

Models	+NE	+NE- overlap	~~NE~~	~~NE~~- overlap
Image-only	0.89	0.70	1.41	1.05
Text-only	1.39	1.46	1.38	1.27
Contextual	1.29	1.28	1.14	1.13

Table 6.2: Validation loss of the reported test models.

6.5.3 Qualitative Evaluation

Setup. A user study was set up on the crowd-sourcing platform Appen[3] (formerly Figure8). 250 test samples were studied. Samples were chosen such that the paragraphs were of similar lengths (40-80 words). For each sample, users were shown the image, its associated paragraph, and were asked to rate 6 captions (4 contextual caption variants and 2 baselines from Table **??**) on a scale from 1 (irrelevant) to 5 (highly relevant). For analysis we consider captions rated ≥ 3 as superior, and those rated < 3 as inferior.

Observations. We observe that for 80.4% of samples (201 out of 250), at least one of the 4 contextual

[3] https://appen.com

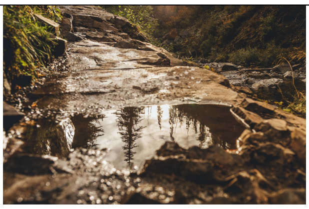
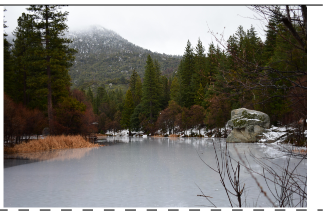
Image		
Paragraph	Made the hike to Franklin Falls and while waiting for some other people to clear my shot, I noticed how good the light looked hitting the rocks.	I was driving down the mountain. . . popped out my camera to snag this shot. It's beautiful right now. . . there wasn't nearly as much snow as last year.
Contextual	*Rush hour on the nature coast.*	*I love the snow mountains. Come in the countryside often.*
Image-only	*The view from the top of the cosmopolitan.*	*The view from the top of the moon.*
Text-only	*Pretty cool sunset.*	*Rain ready for a local bar.*

Figure 6.3: Linguistic richness of *Contextual Captions* in contrast to those generated from only image or only text.

captioning models is rated strictly higher than both baselines, and for 94.8% of samples they are at least as good as both baselines. A variant-wise analysis of this is shown in Table 6.3.

	Image-only		Text-only	
	$\geq$	$>$	$\geq$	$>$
+NE	71.6	42.4	74.4	38.4
+NE-overlap	69.6	42.8	74.0	44.4
~~NE~~	70.0	45.2	73.6	41.2
~~NE~~-overlap	**76.0**	**48.4**	**81.2**	**49.6**

Table 6.3: Percentage of samples where contextual captions are rated as good as or better than baselines.

In 75.22% of samples, contextual captions were rated highest among all 6 captions. The variant-wise analysis of the same is shown in Table 6.4.

+NE	+NE-overlap	~~NE~~	~~NE~~-overlap	Image-only	Text-only
17.79	15.83	15.86	**25.74**	14.98	9.87

Table 6.4: Percentage of samples rated highest per model.

We identify three categories of sample:

- *Significant*: samples where at least one of the 6 methods generate a caption with rating ≥ 3. These constitute 46% of samples (115/250).
- *Insignificant*: samples on which all 6 methods obtain a rating < 3. Here, paragraphs show substantial randomness and offer little context for the image. It appears impossible to generate

good contextual captions for such samples.

- *Bad-base*: samples which are insignificant (rating $<$ 3) with respect to both baselines. These constitute 80.4% of samples (201/250).

For 86.09% of *Significant* samples (99/115), contextual Captions were rated higher than the baselines. A detailed analysis is given in Table 6.5.

	Image-only		Text-only	
	$\geq$	$>$	$\geq$	$>$
+NE	66.1	55.7	67.8	47.0
+NE-overlap	64.4	53.9	69.6	54.0
~~NE~~	60.9	48.7	64.4	43.5
~~NE~~-overlap	**72.17**	**59.1**	**78.3**	**58.3**

Table 6.5: Percentage of *Significant* samples where contextual captions are rated as good as or better than baselines.

The ratings of 32.8% (67 of 201) of *Bad-base* samples were made significant, i.e., improved to strictly $\geq$ 3, by the best contextual captioning model. In other words, contextual captioning generates superior captions for samples with inferior baseline captions.

~~NE~~-overlap emerged as the best contextual captioning split, closely followed by *+NE-overlap*, in both the quantitative and qualitative evaluation.

Factorial Experiment. We conduct another study taking the form of a $2 \times 2 \times 2$ full factorial experiment based on three factors – presence of NEs, caption-paragraph overlap, and use of pre-trained BERT token embeddings. We study the effect of these factors with a user study (with a similar setup as the one described in Section 6.5.3) with all factor combinations. The effect of each of the factors can be seen in Figure 6.4. Using BERT token embeddings is by far the most effective in enhancing caption quality. It is interesting to note that presence of NEs (including its interaction with other factors) has a negative effect – captions without NEs are rated higher. Caption-paragraph overlap splits are also rated higher, which indicates that high inter-modality content overlap is necessary for generating good contextual captions.

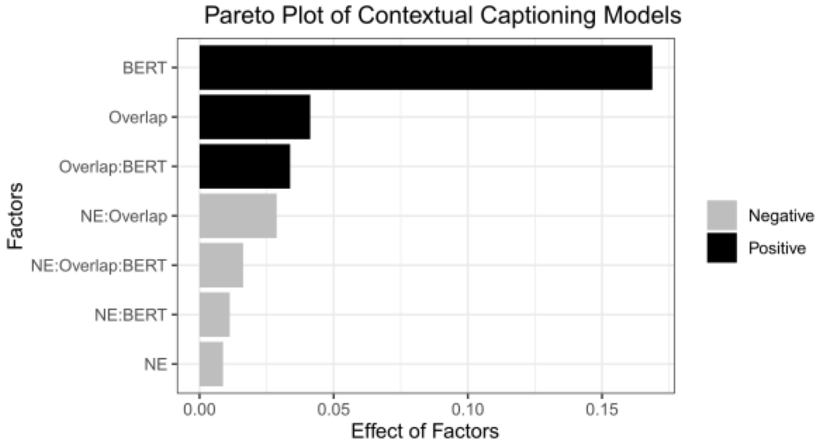


Figure 6.4: Effect of various factors of our Contextual Captioning models.

6.5.4 Further Discussion

Named Entities in Captions. The user study shows that the presence of named entities (NE) has a deteriorating effect on the caption quality. We conjecture that a lack of strong cues from the paragraphs lead to incorrectly generated NEs. Future work should also explore copy mechanisms to copy NEs from paragraphs to captions.

Presence and absence of Named Entities. Both from the quantitative and qualitative evaluations, it becomes apparent that captions without NEs are thematically closer to ground truth (*SemSim* in Table 6.1) as well as preferred by humans. This is an interesting observation on the nature of captions – a generic abstract commentary seems to be of greater value than those including specifics, as in captions with NEs.

Caption Quality. We observe that the captions generated by the baselines do not show linguistic diversity (Figure 6.3). "The view from my hotel window.", "My friend and I..." etc. are common templates learned by the models. We conjecture that training the model on samples containing cleaner paragraphs that have a high content overlap with the image would yield nicer captions. We partially emulate this in our *-overlap* splits, which indeed show better model performance.

Nature of paragraphs. Recall that the text paragraphs in our dataset are constructed by concatenating user comments on a Reddit post. Most often these paragraphs are very noisy and offer very little context overlap with the image. While this increases the linguistic diversity of generated captions, it often leads to meaningless captions which are eventually rated low by human judges. It is our conjecture that training the model on samples containing cleaner paragraphs (like the Lonely Planet dataset proposed by [26]), which have a high content overlap with the image, would lead to better quality captions. We partially emulate this in our *-overlap* splits, which indeed lead to nicer model performances.

In this chapter we proposed the novel task of *Contextual Image Captioning* that exploits image-text synergy of multimodal documents. Table 6.5 shows a few good and bad examples of contextual captions generated by our model. To facilitate a thorough study of this problem, we curate a new dataset comprising 250,000 multimodal Reddit posts. We provide a detailed analysis of the dataset along with experimental results to identify interesting factors that determine the quality of generated captions. We hope that our work will kindle and support follow-up research on this under-explored task, with downstream applications like content authoring tools and multimodal dialogue.

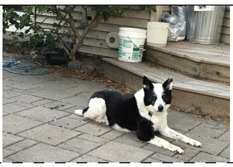
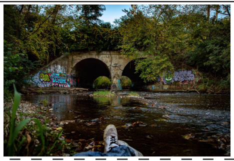
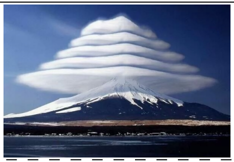
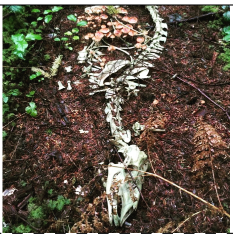
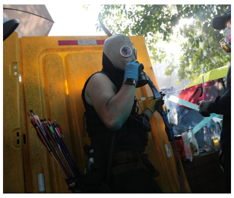
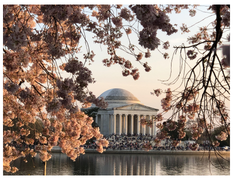
Image			
Paragraph	Shes pretty. Sorry for your loss. If only dogs had a longer lifespan than humans! If only we picked the one that would bury us. . . . That aside, I see you made each others life wonderful.	Thats a dope shot, nice. Thats cool Looks like the Rouge River. Arent those tunnels better known as bridges ? Its not really a bridge its not used for people to get across.	Lenticular clouds are cool. . . Super cool! Looks like a giant tree. . . These are called lenticular clouds. This must be the most enticing pic of a mountain I ever saw.
Contextual Captions	*My only friend passed away last year. He passed away from cancer.*	*A beautiful stream I encountered on holiday.*	*Some of these clouds are having fun as we get out of water.*
Image-only Captions	*My friend is a new friend. My son's dog.*	*A picture I took at a morning in my hometown.*	*A picture I took in my morning. The view from my hotel.*
Text-Only Captions	*Rest easy, hear surgery. Cancer, and essa. He died last year.*	*Milky way over the clouds worth it however.*	*A collage from the top of a cliff bench at midnight.*
Image			
Paragraph	You absolute legend you. Thanks so much! I saw it and thought the same thing! I tried r/skeletons just gave me spooky bone memes. I'll check fossils though! One of those students are headed to krypton.	there is nothing NSFW about this Beijing agent says what. . . US says HK police can handle this. . . HK ppl says please stop violence. . . I say i am just a HK citizen. . . Cop used to be an adventurer.	The Jefferson Memorial will always be my favorite DC monument. I've lived in the area for fifteen years and *finally* saw this beauty for myself in person. It was absolutely worth the hellish traffic!
Contextual Captions	*I drew this and I thought it looked cool. I want to be a good artist.*	*This is a real unedited picture ever taken.*	*A beautiful sunset ship swallowed sea lion waterfalls.*
Image-only Captions	*My first time on a year ago, and I just got a photo of this photo.*	*My first ever attempt at a photo of a year ago.*	*The view from my house in the morning.*
Text-Only Captions	*Danny Devito ink drawing lights up optical.*	*Hong Kong protesters fired a letter to the protest police threatening by police. Pepper spray*	*Tragedy in Pittsburgh, pa 18 years final night the highway.*

Figure 6.5: Good examples (top row) and bad examples (bottom row) of *Contextual Captions*.

CHAPTER 7

Conclusion and Outlook

This dissertation broadly addresses problems involving the interplay of images and text in web-content. The efforts are directed to encourage research in the domain of multimodal document understanding. Here we summarize our contributions and discuss the scopes for future research.

7.1 Summary

Each of the research areas discussed below, although well-studied, are rife with various shortcomings. We have addressed some of the limitations and introduced new facets through our contributions in this dissertation.

Image Retrieval. Search and retrieval of images predominantly function by matching query keywords with textual tags associated with images, like user-provided tags and automatically detected tags (objects, scenes etc.). The key shortcomings of this approach are twofold: (1) image tags are often incomplete (not encompassing all visible objects), (2) intrinsic thematic associations between text and visual content that humans recognize (for example, “environment friendly vehicle” referring to transport without carbon emissions, like bicycles) are lacking in image search systems. In Chapter 3 we propose an image retrieval system to mitigate these limitations. Our experimental results reveal that using information from three modalities – visual data from images, textual data from users’ query, and commonsense knowledge to bridge the semantic gap between them – leads to effective image retrieval.

Automatic Image Tagging. Automatic image tagging typically begins with identifying visual objects within an image by surrounding the regions with ‘bounding boxes’, followed by proposing the most confident prediction out of a limited number of object classes that the detector was trained with. This often leads to noisy object detection: for example, a ‘person’, ‘tennis racket’, ‘tennis court’, and ‘lemon’ could be detected in the same image, where the lemon is clearly thematically out of place. In Chapter 4 we propose to clean up such inter-object incoherence, and enrich the set of image tags with thematic commonsense knowledge concepts. For instance, in the above illustrative example, we replace the tag ‘lemon’ with ‘tennis ball’ based on visual similarities between an array of objects and probability of its spatial co-location with the other identified objects (‘person, ‘tennis racket, ‘tennis court’). Additionally, we add the thematic tags ‘sports’ and ‘entertainment’. A user study shows that the enriched tag space proposed by our model is more expressive than that proposed by traditional image tagging systems.

Text-Image Alignment. In Chapter 5 we study the problem of automatically selecting relevant images for a story and aligning them to textual paragraphs based on text-image semantic similarity.

While prior works propose systems to align small image regions to textual phrases, or whole images to single sentences, our contribution considers the overall context of a big story by solving a global optimization problem for a more efficient thematic alignment.

Image Captioning. Traditional image captioning tools generate straightforward descriptions of the visual content of images. Much of the prior works in this field rely on big crowd-sourced data sets of descriptive image captions. The need for crowd-sourcing can be understood in the context of available image-caption pairs on social media sites like Flickr, Pinterest etc.: these capture sentimental and thematic concepts rather than explicitly describing the visual image content. With our contribution in Chapter 6 we venture away from descriptive captions, and use contextual information (text surrounding an image, for example the associated paragraph in a multimodal story) to generate thematic captions which capture knowledge from both the textual and visual modalities. The captions generated this way are closer to the real-world image captions (on social media sites).

Our experimental results show improvements over state-of-the-art methods (where available).

7.2 Future Research Directions

We anticipate that the research presented in this dissertation will encourage the Natural Language Processing and Artificial Intelligence research communities to pursue the open challenges that have not been addressed so far. Following are some scopes that we think are worth exploring.

User Intent in Multimodal Contexts. The user's information need or intent behind a query is an important aspect in any search scenario. In multimodal contexts, this becomes even more challenging. For example, with the query "environment friendly transport" a user might want to see images of electric vehicles, or manual means of short-distance transport like bicycles, skateboards etc. While some of the intent can be deciphered from the user's search history, it is still a subjective and challenging issue. Similar challenges exist for text-image semantic alignment. In certain cases like news articles, images and adjoining paragraphs need to be explicitly depicting certain objects and events, while in other cases like personal blog posts, images are often used for broad thematic illustration. Although a high-level personalization can tackle the problem to some extent, a deeper understanding of topic, context, and user intent is required.

Understanding Linguistic Styles. Language often assumes interesting narrative structures beyond literal retelling of events and thoughts. The use of figures of speech like metaphors ("The curtain of night fell upon us."), hyperboles ("The news was a dagger to his heart.") or other linguistic devices like sarcasm ("I work 40 hours a week to be this poor."), and allegory ("All animals are equal but some animals are more equal than others.") is common practice to make narration interesting. When dealing with multimodal applications like image retrieval or text-image alignment, this warrants an additional challenge because of the lack of direct mappings from text to visual objects. Understanding high-level implications become essential in tackling such linguistic styles. For example, the above hyperbole should not be paired with images of daggers, and "animals" may refer to human beings in a bureaucratic setup for the cited allegory. A deeper understanding of such linguistic styles is necessary for sophisticated machine understanding of human language.

Multimodal Chat-bots. Person-to-person digital communication has evolved in the very recent past and has taken a multimodal direction. From SMS-interfaces using predominantly text and simple emoticons, we have expanded chats with visually appealing (often animated) emoticons, memes, and GIFs. Following the trend, chat-bots should ramp up their abilities to better connect with and assist humans. To this end, research learning associations between visual data (GIFs, memes, emoticons) and textual chats will be a valuable contribution.

We hope that the research presented in this dissertation, the new resources introduced, and the discussion that followed will foster and enrich future research in multimodal data understanding and other related topics.

Bibliography

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052 3

[2] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. 2016. Sort Story: Sorting Jumbled Images and Captions into Stories. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 925–931. https://doi.org/10.18653/v1/d16-1091 55

[3] Emre Akbas and Fatos T. Yarman-Vural. 2007. Automatic Image Annotation by Ensemble of Visual Descriptors. 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA. IEEE Computer Society. https://doi.org/10.1109/CVPR.2007.383484 38

[4] Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A Corpus of Image-Text Discourse Relations. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 570–575. https://doi.org/10.18653/v1/n19-1056 3

[5] Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A Corpus of Image-Text Discourse Relations. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 570–575. https://doi.org/10.18653/v1/n19-1056 71

[6] Ahmad AlZu'bi, Abbes Amira, and Naeem Ramzan. 2015. Semantic content-based image retrieval: A comprehensive study. J. Vis. Commun. Image Represent. 32 (2015), 20–54. https://doi.org/10.1016/j.jvcir.2015.07.012 8

[7] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V (Lecture Notes in Computer Science), Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.), Vol. 9909. Springer, 382–398. https://doi.org/10.1007/978-3-319-46454-1_24 23

[8] Relja Arandjelovic and Andrew Zisserman. 2012. Three things everyone should know to improve object retrieval. 2012 IEEE Conference on Computer Vision and Pattern Recognition,

Providence, RI, USA, June 16-21, 2012. IEEE Computer Society, 2911–2918. https://doi.org/10.1109/CVPR.2012.6248018 36

[9] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss (Eds.). Association for Computational Linguistics, 65–72. https://www.aclweb.org/anthology/W05-0909/ 22

[10] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction for the web. International Joint Conferences on Artificial Intelligence, Vol. 7. 2670–2676. 27

[11] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A Neural Probabilistic Language Model. Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA, Todd K. Leen, Thomas G. Dietterich, and Volker Tresp (Eds.). MIT Press, 932–938. http://papers.nips.cc/paper/1839-a-neural-probabilistic-language-model 12

[12] Raffaella Bernardi, Ruket Çakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. J. Artif. Intell. Res. 55 (2016), 409–442. https://doi.org/10.1613/jair.4900 9, 69

[13] Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good News, Everyone! Context Driven Entity-Aware Captioning for News Images. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 12466–12475. https://doi.org/10.1109/CVPR.2019.01275 76

[14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html 10

[15] Rafael A. Calvo and Sidney K. D'Mello. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. IEEE Trans. Affect. Comput. 1, 1 (2010), 18–37. https://doi.org/10.1109/T-AFFC.2010.1 2

[16] João Carreira, Fuxin Li, and Cristian Sminchisescu. 2012. Object Recognition by Sequential Figure-Ground Ranking. Int. J. Comput. Vis. 98, 3 (2012), 243–262. https://doi.org/10.1007/s11263-011-0507-2 39

[17] Arjun Chandrasekaran, Devi Parikh, and Mohit Bansal. 2018. Punny Captions: Witty Wordplay in Image Descriptions. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 770–775. https://doi.org/10.18653/v1/n18-2121 9, 77

[18] Jingqiang Chen and Hai Zhuge. 2018. Abstractive Text-Image Summarization Using Multi-Modal Attentional Hierarchical RNN. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 4046–4056. https://doi.org/10.18653/v1/d18-1438 75, 77

[19] Jingqiang Chen and Hai Zhuge. 2019. Extractive summarization of documents with images based on multi-modal RNN. Future Gener. Comput. Syst. 99 (2019), 186–196. https://doi.org/10.1016/j.future.2019.04.045 75, 77

[20] Jingqiang Chen and Hai Zhuge. 2019. News Image Captioning Based on Text Summarization Using Image as Query. 15th International Conference on Semantics, Knowledge and Grids, SKG 2019, Guangzhou, China, September 17-18, 2019. IEEE, 123–126. https://doi.org/10.1109/SKG49510.2019.00029 77

[21] Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. 2014. Object-Based Visual Sentiment Concept Analysis and Application. Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014, Kien A. Hua, Yong Rui, Ralf Steinmetz, Alan Hanjalic, Apostol Natsev, and Wenwu Zhu (Eds.). ACM, 367–376. https://doi.org/10.1145/2647868.2654935 39

[22] Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. 2018. "Factual" or "Emotional": Stylized Image Captioning with Adaptive Learning and Attention. Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X (Lecture Notes in Computer Science), Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11214. Springer, 527–543. https://doi.org/10.1007/978-3-030-01249-6_32 9

[23] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. NEIL: Extracting Visual Knowledge from Web Data. IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013. IEEE Computer Society, 1409–1416. https://doi.org/10.1109/ICCV.2013.178 10

[24] Ryszard S. Choras. 2006. Content-Based Image Retrieval - A Survey. Biometrics, Computer Security Systems and Artificial Intelligence Applications, Khalid Saeed, Jerzy Pejas, and

Romuald Mosdorf (Eds.). Springer, 31–44. https://doi.org/10.1007/978-0-387-36503-9_4 8

[25] Sreyasi Nag Chowdhury, Rajarshi Bhowmik, Hareesh Ravi, Gerard de Melo, Simon Razniewski, and Gerhard Weikum. 2021. Exploiting Image–Text Synergy for Contextual Image Captioning. Proceedings of the Third Workshop Beyond Vision and LANguage: inTEgrating Real-world kNowledge, EACL Workshop LANTERN 2021, Kyiv, Ukraine, April 19-23, 2021. ACL. 3

[26] Sreyasi Nag Chowdhury, William Cheng, Gerard de Melo, Simon Razniewski, and Gerhard Weikum. 2020. Illustrate Your Story: Enriching Text with Images. WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 849–852. https://doi.org/10.1145/3336191.3371866 77, 84

[27] Sreyasi Nag Chowdhury, Simon Razniewski, and Gerhard Weikum. 2021. SANDI: Story-and-Images Alignment. Proceedings of the Sixteenth Conference of the European Chapter of the Association for Computational Linguistics, EACL 2021, Kyiv, Ukraine, April 19-23, 2021. ACL. 3

[28] Sreyasi Nag Chowdhury, Niket Tandon, Hakan Ferhatosmanoglu, and Gerhard Weikum. 2018. VISIR: Visual and Semantic Image Label Refinement. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018, Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek (Eds.). ACM, 117–125. https://doi.org/10.1145/3159652.3159693 2, 8, 53, 57, 61

[29] Sreyasi Nag Chowdhury, Niket Tandon, and Gerhard Weikum. 2016. Know2Look: Commonsense Knowledge for Visual Search. Proceedings of the 5th Workshop on Automated Knowledge Base Construction, AKBC@NAACL-HLT 2016, San Diego, CA, USA, June 17, 2016, Jay Pujara, Tim Rocktäschel, Danqi Chen, and Sameer Singh (Eds.). The Association for Computer Linguistics, 57–62. https://doi.org/10.18653/v1/w16-1311 2, 39, 57

[30] Wei-Ta Chu and Ming-Chih Kao. 2017. Blog Article Summarization with Image-Text Alignment Techniques. 19th IEEE International Symposium on Multimedia, ISM 2017, Taichung, Taiwan, December 11-13, 2017. IEEE Computer Society, 244–247. https://doi.org/10.1109/ISM.2017.40 55, 77

[31] Ciprian Adrian Corneanu, Marc Oliu Simon, Jeffrey F. Cohn, and Sergio Escalera Guerrero. 2016. Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications. IEEE Trans. Pattern Anal. Mach. Intell. 38, 8 (2016), 1548–1568. https://doi.org/10.1109/TPAMI.2016.2515606 2

[32] Luciano Del Corro and Rainer Gemulla. 2013. ClausIE: clause-based open information extraction. WWW. International World Wide Web Conferences Steering Committee / ACM, 355–366. 13

[33] Nick Craswell and Martin Szummer. 2007. Random walks on the click graph. SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research

and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007, Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando (Eds.). ACM, 239–246. https://doi.org/10.1145/1277741.1277784 36

[34] George B. Dantzig. 1960. Inductive Proof of the Simplex Method. IBM J. Res. Dev. 4, 5 (1960), 505–506. 14

[35] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Ze Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40, 2 (2008), 5:1–5:60. https://doi.org/10.1145/1348246.1348248 25

[36] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Ze Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40, 2 (2008), 5:1–5:60. https://doi.org/10.1145/1348246.1348248 36

[37] Diogo Delgado, João Magalhães, and Nuno Correia. 2010. Automated Illustration of News Stories. Proceedings of the 4th IEEE International Conference on Semantic Computing (ICSC 2010), September 22-24, 2010, Carnegie Mellon University, Pittsburgh, PA, USA. IEEE Computer Society, 73–78. https://doi.org/10.1109/ICSC.2010.68 54, 64

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. IEEE Computer Society, 248–255. https://doi.org/10.1109/CVPR.2009.5206848 17, 25, 32

[39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. IEEE Computer Society, 248–255. https://doi.org/10.1109/CVPR.2009.5206848 20, 24, 36, 40, 44

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423 12, 16, 19

[41] Santosh Kumar Divvala, Ali Farhadi, and Carlos Guestrin. 2014. Learning Everything about Anything: Webly-Supervised Visual Concept Learning. 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. IEEE Computer Society, 3270–3277. https://doi.org/10.1109/CVPR.2014.412 36

[42] Santosh Kumar Divvala, Derek Hoiem, James Hays, Alexei A. Efros, and Martial Hebert. 2009. An empirical study of context in object detection. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009,

Miami, Florida, USA. IEEE Computer Society, 1271–1278. https://doi.org/10.1109/CVPR.2009.5206532 39

[43] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. 2015. Long-term recurrent convolutional networks for visual recognition and description. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 2625–2634. https://doi.org/10.1109/CVPR.2015.7298878 9, 77

[44] Oren Etzioni, Michael J. Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the Web: An experimental study. Artif. Intell. 165, 1 (2005), 91–134. 13

[45] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying Relations for Open Information Extraction. EMNLP. ACL, 1535–1545. 13

[46] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying Relations for Open Information Extraction. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, 1535–1545. https://www.aclweb.org/anthology/D11-1142/ 32

[47] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018. BMVA Press, 12. http://bmvc2018.org/contents/papers/0344.pdf 8, 53, 61, 62, 64, 66, 67, 68, 69

[48] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 1473–1482. https://doi.org/10.1109/CVPR.2015.7298754 9, 77

[49] Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. 2010. Every Picture Tells a Story: Generating Sentences from Images. Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV (Lecture Notes in Computer Science), Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.), Vol. 6314. Springer, 15–29. https://doi.org/10.1007/978-3-642-15561-1_2 9, 77

[50] Zheyun Feng, Songhe Feng, Rong Jin, and Anil K. Jain. 2014. Image Tag Completion by Noisy Matrix Recovery. Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII (Lecture Notes in Computer Science), David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.), Vol. 8695. Springer, 424–438. https://doi.org/10.1007/978-3-319-10584-0_28 39

[51] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 2121–2129. http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model 8, 53

[52] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. StyleNet: Generating Attractive Visual Captions with Styles. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 955–964. https://doi.org/10.1109/CVPR.2017.108 9, 77

[53] Yue Gao, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu. 2013. Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search. IEEE Trans. Image Process. 22, 1 (2013), 363–376. https://doi.org/10.1109/TIP.2012.2202676 38

[54] Ross B. Girshick. 2015. Fast R-CNN. 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. IEEE Computer Society, 1440–1448. https://doi.org/10.1109/ICCV.2015.169 4

[55] Georgia Gkioxari, Ross B. Girshick, and Jitendra Malik. 2015. Contextual Action Recognition with R*CNN. 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. IEEE Computer Society, 1080–1088. https://doi.org/10.1109/ICCV.2015.129 8, 36

[56] Shivali Goel, Rishi Madhok, and Shweta Garg. 2018. Proposing Contextually Relevant Quotes for Images. Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings (Lecture Notes in Computer Science), Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.), Vol. 10772. Springer, 591–597. https://doi.org/10.1007/978-3-319-76941-7_49 69

[57] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics. https://doi.org/10.18653/v1/p16-1154 76

[58] Matthieu Guillaumin, Thomas Mensink, Jakob J. Verbeek, and Cordelia Schmid. 2009. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009. IEEE Computer Society, 309–316. https://doi.org/10.1109/ICCV.2009.5459266 39

[59] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep Multimodal Representation Learning: A Survey. IEEE Access 7 (2019), 63373–63394. https://doi.org/10.1109/ACCESS.2019.2916887 3

[60] Manish Gupta, Rui Li, Zhijun Yin, and Jiawei Han. 2010. Survey on social tagging techniques. SIGKDD Explorations 12, 1 (2010), 58–72. https://doi.org/10.1145/1882471.1882480 53

[61] Harry Halpin, Valentin Robu, and Hana Shepherd. 2007. The complex dynamics of collaborative tagging. Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007, Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy (Eds.). ACM, 211–220. https://doi.org/10.1145/1242572.1242602 36

[62] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. CVPR. 770–778. 79

[63] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 16

[64] Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Ronghang Hu, Jeff Donahue, Ross B. Girshick, Trevor Darrell, and Kate Saenko. 2014. LSDA: Large Scale Detection through Adaptation. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 3536–3544. http://papers.nips.cc/paper/5418-lsda-large-scale-detection-through-adaptation 8, 56, 61

[65] Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Ronghang Hu, Jeff Donahue, Ross B. Girshick, Trevor Darrell, and Kate Saenko. 2014. LSDA: Large Scale Detection through Adaptation. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 3536–3544. http://papers.nips.cc/paper/5418-lsda-large-scale-detection-through-adaptation 25, 29, 32, 36, 37, 38, 40, 41, 44, 113

[66] MD Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. ACM Computing Surveys (CSUR) 51, 6 (2019), 118. 75

[67] Xian-Sheng Hua and Jin Li. 2014. Tell me what. 2013 IEEE International Conference on Multimedia and Expo Workshops, Chengdu, China, July 14-18, 2014. IEEE Computer Society, 1–2. https://doi.org/10.1109/ICMEW.2014.6890616 36

[68] Rodrigo Toro Icarte, Jorge A. Baier, Cristian Ruz, and Alvaro Soto. 2017. How a General-Purpose Commonsense Ontology can Improve Performance of Learning-Based Image Retrieval. Proceedings of the Twenty-Sixth International Joint Conference on Artificial

Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, Carles Sierra (Ed.). ijcai.org, 1283–1289. https://doi.org/10.24963/ijcai.2017/178 8

[69] Filip Ilievski, Pedro A. Szekely, Jingwei Cheng, Fu Zhang, and Ehsan Qasemi. 2020. Consolidating Commonsense Knowledge. CoRR abs/2006.06114 (2020). arXiv:2006.06114 https://arxiv.org/abs/2006.06114 10

[70] Dhiraj Joshi, James Ze Wang, and Jia Li. 2006. The Story Picturing Engine - a system for automatic text illustration. ACM Trans. Multim. Comput. Commun. Appl. 2, 1 (2006), 68–89. https://doi.org/10.1145/1126004.1126008 53, 54, 64

[71] Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 3128–3137. https://doi.org/10.1109/CVPR.2015.7298932 8, 9, 77

[72] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. CoRR abs/1411.2539 (2014). arXiv:1411.2539 http://arxiv.org/abs/1411.2539 62

[73] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What Are You Talking About? Text-to-Image Coreference. 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. IEEE Computer Society, 3558–3565. https://doi.org/10.1109/CVPR.2014.455 8

[74] Ronak Kosti, Jose M. Alvarez, Adrià Recasens, and Àgata Lapedriza. 2017. Emotion Recognition in Context. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 1960–1968. https://doi.org/10.1109/CVPR.2017.212 36

[75] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. 2016. Crowdsourcing in Computer Vision. Found. Trends Comput. Graph. Vis. 10, 3 (2016), 177–243. https://doi.org/10.1561/0600000071 36

[76] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A Hierarchical Approach for Generating Descriptive Image Paragraphs. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 3337–3345. https://doi.org/10.1109/CVPR.2017.356 9, 77

[77] Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards Unsupervised Image Captioning With Shared Multimodal Embeddings. 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. IEEE, 7413–7423. https://doi.org/10.1109/ICCV.2019.00751 77

[78] Alex Lascarides and Matthew Stone. 2009. Discourse coherence and gesture interpretation. Gesture 9, 2 (2009), 147–180. 71

[79] Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361, 10 (1995), 1995. 16

[80] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. 2015. Deep learning. Nat. 521, 7553 (2015), 436–444. https://doi.org/10.1038/nature14539 25

[81] Douglas B. Lenat, Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd. 1990. CYC: Toward Programs With Common Sense. Commun. ACM 33, 8 (1990), 30–49. https://doi.org/10.1145/79173.79176 10

[82] Paul Martin Lester. 2013. Visual communication: Images with messages. Cengage Learning. 52

[83] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh C. Jain. 2006. Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multim. Comput. Commun. Appl. 2, 1 (2006), 1–19. https://doi.org/10.1145/1126004.1126005 36

[84] Xirong Li, Cees G. M. Snoek, and Marcel Worring. 2009. Learning Social Tag Relevance by Neighbor Voting. IEEE Trans. Multimedia 11, 7 (2009), 1310–1322. https://doi.org/10.1109/TMM.2009.2030598 38

[85] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees G. M. Snoek, and Alberto Del Bimbo. 2016. Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement, and Retrieval. ACM Comput. Surv. 49, 1 (2016), 14:1–14:39. https://doi.org/10.1145/2906152 36, 39

[86] Henry Lieberman and Hugo Liu. 2002. Adaptive Linking between Text and Photos Using Common Sense Reasoning. Adaptive Hypermedia and Adaptive Web-Based Systems, Second International Conference, AH 2002, Malaga, Spain, May 29-31, 2002, Proceedings (Lecture Notes in Computer Science), Paul De Bra, Peter Brusilovsky, and Ricardo Conejo (Eds.), Vol. 2347. Springer, 2–11. https://doi.org/10.1007/3-540-47952-X_2 55

[87] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. Text summarization branches out. 74–81. 22

[88] Hui Lin and Vincent Ng. 2019. Abstractive Summarization: A Survey of the State of the Art. The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, 9815–9822. https://doi.org/10.1609/aaai.v33i01.33019815 75

[89] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V (Lecture Notes in Computer Science), David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.), Vol. 8693. Springer, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48 20, 32, 45, 62, 80

[90] Xiao Lin and Devi Parikh. 2015. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. IEEE Conference on Computer Vision and Pattern

Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 2984–2993. https://doi.org/10.1109/CVPR.2015.7298917 10

[91] Bei Liu, Jianlong Fu, Makoto P. Kato, and Masatoshi Yoshikawa. 2018. Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training. 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018, Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei (Eds.). ACM, 783–791. https://doi.org/10.1145/3240508.3240587 9, 77

[92] Dong Liu, Xian-Sheng Hua, Meng Wang, and Hong-Jiang Zhang. 2010. Retagging social images based on visual and semantic consistency. Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010, Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti (Eds.). ACM, 1149–1150. https://doi.org/10.1145/1772690.1772848 39

[93] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. 2009. Tag ranking. Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009, Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl (Eds.). ACM, 351–360. https://doi.org/10.1145/1526709.1526757 8

[94] Dong Liu, Shuicheng Yan, Xian-Sheng Hua, and Hong-Jiang Zhang. 2011. Image Retagging Using Collaborative Tag Propagation. IEEE Trans. Multimedia 13, 4 (2011), 702–712. https://doi.org/10.1109/TMM.2011.2134078 39

[95] Hugo Liu and Henry Lieberman. 2002. Robust photo retrieval using world semantics. Proceedings of LREC2002 Workshop: Using Semantics for IR. 15–20. 39

[96] Hugo Liu and Push Singh. 2004. ConceptNet—a practical commonsense reasoning tool-kit. BT technology journal 22, 4 (2004), 211–226. 10, 18, 40, 42, 43, 113

[97] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. 2007. A survey of content-based image retrieval with high-level semantics. Pattern Recognit. 40, 1 (2007), 262–282. https://doi.org/10.1016/j.patcog.2006.04.045 8

[98] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. 2007. A survey of content-based image retrieval with high-level semantics. Pattern Recognit. 40, 1 (2007), 262–282. https://doi.org/10.1016/j.patcog.2006.04.045 25

[99] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual Relationship Detection with Language Priors. Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I (Lecture Notes in Computer Science), Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.), Vol. 9905. Springer, 852–869. https://doi.org/10.1007/978-3-319-46448-0_51 53

[100] Dan Lu, Xiaoxiao Liu, and Xueming Qian. 2016. Tag-Based Image Search by Social Re-ranking. IEEE Trans. Multim. 18, 8 (2016), 1628–1639. https://doi.org/10.1109/TMM.2016.2568099 8

[101] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 3242–3250. https://doi.org/10.1109/CVPR.2017.345 9

[102] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2015. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6632 9, 77

[103] Leandro Balby Marinho, Alexandros Nanopoulos, Lars Schmidt-Thieme, Robert Jäschke, Andreas Hotho, Gerd Stumme, and Panagiotis Symeonidis. 2011. Social Tagging Recommender Systems. Recommender Systems Handbook. Springer, 615–644. 38

[104] Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. EMNLP-CoNLL. ACL, 523–534. 13

[105] Paul Messaris and Linus Abraham. 2001. The role of images in framing news stories. Framing public life. Routledge, 231–242. 52

[106] Tomás Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010, Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura (Eds.). ISCA, 1045–1048. http://www.isca-speech.org/archive/interspeech_2010/i10_1045.html 12

[107] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality 19, 41

[108] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality 61

[109] George A. Miller. 1995. WordNet: A Lexical Database for English. Commun. ACM 38, 11 (1995), 39–41. https://doi.org/10.1145/219717.219748 17, 25

[110] George A. Miller. 1995. WordNet: A Lexical Database for English. Commun. ACM 38, 11 (1995), 39–41. https://doi.org/10.1145/219717.219748 37, 40, 41

[111] Eleni Miltsakaki, Rashmi Prasad, Aravind K. Joshi, and Bonnie L. Webber. 2004. The Penn Discourse Treebank. Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2004/summaries/618.htm 71

[112] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. 2015. Inceptionism: Going deeper into neural networks. Google Research Blog. Retrieved June (2015). 25

[113] Frederic Morin and Yoshua Bengio. 2005. Hierarchical Probabilistic Neural Network Language Model. Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS 2005, Bridgetown, Barbados, January 6-8, 2005, Robert G. Cowell and Zoubin Ghahramani (Eds.). Society for Artificial Intelligence and Statistics. http://www.gatsby.ucl.ac.uk/aistats/fullpapers/208.pdf 12

[114] Kazuma Murao, Ken Kobayashi, Hayato Kobayashi, Taichi Yatsuka, Takeshi Masuyama, Tatsuru Higurashi, and Yoshimune Tabuchi. 2019. A Case Study on Neural Headline Generation for Editing Support. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 2 (Industry Papers), Anastassia Loukina, Michelle Morales, and Rohit Kumar (Eds.). Association for Computational Linguistics, 73–82. https://doi.org/10.18653/v1/n19-2010 75

[115] Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why We Need New Evaluation Metrics for NLG. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 2241–2252. https://doi.org/10.18653/v1/d17-1238 22

[116] Vicente Ordonez, Xufeng Han, Polina Kuznetsova, Girish Kulkarni, Margaret Mitchell, Kota Yamaguchi, Karl Stratos, Amit Goyal, Jesse Dodge, Alyssa C. Mensch, Hal Daumé III, Alexander C. Berg, Yejin Choi, and Tamara L. Berg. 2016. Large Scale Retrieval and Generation of Image Descriptions. Int. J. Comput. Vis. 119, 1 (2016), 46–59. https://doi.org/10.1007/s11263-015-0840-y 9

[117] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain, John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (Eds.). 1143–1151. http://papers.nips.cc/paper/4470-im2text-describing-images-using-1-million-captioned-photographs 32, 45

[118] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. ACL, 311–318. https://doi.org/10.3115/1073083.1073135 21

[119] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 6432–6440. https://doi.org/10.1109/CVPR.2017.681 9, 20, 77

[120] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543. https://doi.org/10.3115/v1/d14-1162 19

[121] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. ICCV. IEEE Computer Society, 2641–2649. 20

[122] Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia, W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel (Eds.). ACM, 275–281. https://doi.org/10.1145/290941.291008 11

[123] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The Penn Discourse TreeBank 2.0. Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2008/summaries/754.html 71

[124] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge J. Belongie. 2007. Objects in Context. IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007. IEEE Computer Society, 1–8. https://doi.org/10.1109/ICCV.2007.4408986 39

[125] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Chuck Rosenberg, and Fei-Fei Li. 2015. Learning semantic relationships for better action retrieval in images. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 1100–1109. https://doi.org/10.1109/CVPR.2015.7298713 36

[126] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. Proceedings of the 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, USA,

June 6, 2010, Chris Callison-Burch and Mark Dredze (Eds.). Association for Computational Linguistics, 139–147. https://www.aclweb.org/anthology/W10-0721/ 20, 32, 45

[127] Hareesh Ravi, Lezi Wang, Carlos Muñiz, Leonid Sigal, Dimitris N. Metaxas, and Mubbasir Kapadia. 2018. Show Me a Story: Towards Coherent Neural Story Illustration. 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. IEEE Computer Society, 7613–7621. https://doi.org/10.1109/CVPR.2018.00794 55, 64

[128] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 779–788. https://doi.org/10.1109/CVPR.2016.91 36, 39

[129] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 6517–6525. https://doi.org/10.1109/CVPR.2017.690 8, 53, 56, 61

[130] Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. Comput. Linguistics 44, 3 (2018). https://doi.org/10.1162/coli_a_00322 22

[131] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 91–99. http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks 8

[132] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 6 (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031 36, 38

[133] Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhadeo, and Gerhard Weikum. 2019. Commonsense Properties from Query Logs and Question Answering Forums. Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 1411–1420. https://doi.org/10.1145/3357384.3357955 10

[134] Yong Rui, Thomas S. Huang, and Shih-Fu Chang. 1999. Image Retrieval: Current Techniques, Promising Directions, and Open Issues. J. Vis. Commun. Image Represent. 10, 1 (1999), 39–62. https://doi.org/10.1006/jvci.1999.0413 8

[135] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and

Fei-Fei Li. 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y 20, 45, 50

[136] Fereshteh Sadeghi, Santosh Kumar Divvala, and Ali Farhadi. 2015. VisKE: Visual knowledge extraction and question answering by visual verification of relation phrases. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 1456–1464. https://doi.org/10.1109/CVPR.2015.7298752 10

[137] Katharina Schwarz, Pavel Rojtberg, Joachim Caspar, Iryna Gurevych, Michael Goesele, and Hendrik P. A. Lensch. 2010. Text-to-Video: Story Illustration from Online Photo Collections. Knowledge-Based and Intelligent Information and Engineering Systems - 14th International Conference, KES 2010, Cardiff, UK, September 8-10, 2010, Proceedings, Part IV (Lecture Notes in Computer Science), Rossitza Setchi, Ivan Jordanov, Robert J. Howlett, and Lakhmi C. Jain (Eds.), Vol. 6279. Springer, 402–409. https://doi.org/10.1007/978-3-642-15384-6_43 53, 54, 64

[138] Thomas Scialom, Patrick Bordes, Paul-Alexis Dray, Jacopo Staiano, and Patrick Gallinari. 2020. BERT Can See Out of the Box: On the Cross-modal Transferability of Text Representations. CoRR abs/2002.10832 (2020). arXiv:2002.10832 https://arxiv.org/abs/2002.10832 77

[139] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 1073–1083. https://doi.org/10.18653/v1/P17-1099 75, 76

[140] Gihan Shin and Junchu Chun. [n. d.]. 2

[141] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging Image Captioning via Personality. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 12516–12526. https://doi.org/10.1109/CVPR.2019.01280 9, 77

[142] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. (2015). http://arxiv.org/abs/1409.1556 25

[143] Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open Mind Common Sense: Knowledge Acquisition from the General Public. On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002 Irvine, California, USA, October 30 - November 1, 2002, Proceedings, Robert Meersman and Zahir Tari (Eds.). Lecture Notes in Computer Science, Vol. 2519. Springer, 1223–1237. https://doi.org/10.1007/3-540-36124-3_77 10, 18, 39

[144] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. EMNLP. ACL, 254–263. 24

[145] Stephen Soderland. 1999. Learning Information Extraction Rules for Semi-Structured and Free Text. Mach. Learn. 34, 1-3 (1999), 233–272. 13

[146] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, Satinder P. Singh and Shaul Markovitch (Eds.). AAAI Press, 4444–4451. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972 55, 57

[147] R. Speer and Catherine Havasi. 2012. Representing General Relational Knowledge in ConceptNet 5. Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), 3679–3686. http://www.lrec-conf.org/proceedings/lrec2012/summaries/1072.html 37, 39, 44

[148] Matthew Stone and Una Stojnic. 2015. Meaning and demonstration. Review of Philosophy and Psychology 6, 1 (2015), 69–97. 71

[149] Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is Not Suitable for the Evaluation of Text Simplification. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 738–744. https://doi.org/10.18653/v1/d18-1081 22

[150] Ying Hua Tan and Chee Seng Chan. 2016. phi-LSTM: A Phrase-Based Hierarchical LSTM Model for Image Captioning. Computer Vision - ACCV 2016 - 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V (Lecture Notes in Computer Science), Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato (Eds.), Vol. 10115. Springer, 101–117. https://doi.org/10.1007/978-3-319-54193-8_7 9

[151] Niket Tandon, Gerard de Melo, Abir De, and Gerhard Weikum. 2015. Knowlywood: Mining Activity Knowledge From Hollywood Narratives. Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015, James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu (Eds.). ACM, 223–232. https://doi.org/10.1145/2806416.2806583 26

[152] Niket Tandon, Gerard de Melo, Fabian M. Suchanek, and Gerhard Weikum. 2014. WebChild: harvesting and organizing commonsense knowledge from the web. Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY,

USA, February 24-28, 2014, Ben Carterette, Fernando Diaz, Carlos Castillo, and Donald Metzler (Eds.). ACM, 523–532. https://doi.org/10.1145/2556195.2556245 10, 26, 39, 50

[153] Niket Tandon, Charles Hariman, Jacopo Urbani, Anna Rohrbach, Marcus Rohrbach, and Gerhard Weikum. 2016. Commonsense in Parts: Mining Part-Whole Relations from the Web and Image Tags. (2016), 243–250. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12337 26

[154] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. The New Data and New Challenges in Multimedia Research. CoRR abs/1503.01817 (2015). arXiv:1503.01817 http://arxiv.org/abs/1503.01817 28, 32

[155] Alasdair Tran, Alexander Patrick Mathews, and Lexing Xie. 2020. Transform and Tell: Entity-Aware News Image Captioning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. IEEE, 13032–13042. https://doi.org/10.1109/CVPR42600.2020.01305 77

[156] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. NIPS. 5998–6008. 12, 16, 19

[157] Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Learning Common Sense through Visual Abstraction. 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. IEEE Computer Society, 2542–2550. https://doi.org/10.1109/ICCV.2015.292 10

[158] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087 23

[159] Remco C. Veltkamp, Mirela Tanase, and Danielle Sent. 1999. Features in Content-based Image Retrieval Systems: a Survey. State-of-the-Art in Content-Based Image and Video Retrieval [Dagstuhl Seminar, 5-10 December 1999] (Computational Imaging and Vision), Remco C. Veltkamp, Hans Burkhardt, and Hans-Peter Kriegel (Eds.), Vol. 22. Kluwer / Springer, 97–124. https://doi.org/10.1007/978-94-015-9664-0_5 8

[160] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-Embeddings of Images and Language. 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1511.06361 8

[161] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. Image Vis. Comput. 27, 12 (2009), 1743–1759. https://doi.org/10.1016/j.imavis.2008.11.007 2

[162] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935 9, 77

[163] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. IEEE Trans. Pattern Anal. Mach. Intell. 39, 4 (2017), 652–663. https://doi.org/10.1109/TPAMI.2016.2587640 36

[164] Haoran Wang, Yue Zhang, and Xiaosheng Yu. 2020. An Overview of Image Caption Generation Methods. Comp. Int. and Neurosc. 2020 (2020), 3062706:1–3062706:13. 75

[165] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. CNN-RNN: A Unified Framework for Multi-label Image Classification. 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2285–2294. https://doi.org/10.1109/CVPR.2016.251 39

[166] Ke Wang and Xiaojun Wan. 2018. SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, Jérôme Lang (Ed.). ijcai.org, 4446–4452. https://doi.org/10.24963/ijcai.2018/618 9

[167] Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. WSABIE: Scaling Up to Large Vocabulary Image Annotation. IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011, Toby Walsh (Ed.). IJCAI/AAAI, 2764–2770. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-460 38

[168] Bryan Williams, Henry Lieberman, and Patrick H. Winston. 2017. Understanding Stories with Large-Scale Common Sense. Proceedings of the Thirteenth International Symposium on Commonsense Reasoning, COMMONSENSE 2017, London, UK, November 6-8, 2017 (CEUR Workshop Proceedings), Andrew S. Gordon, Rob Miller, and György Turán (Eds.), Vol. 2052. CEUR-WS.org. http://ceur-ws.org/Vol-2052/paper21.pdf 55

[169] Fei Wu, Xinyan Lu, Jun Song, Shuicheng Yan, Zhongfei (Mark) Zhang, Yong Rui, and Yueting Zhuang. 2016. Learning of Multimodal Representations With Random Walks on the Click Graph. IEEE Trans. Image Process. 25, 2 (2016), 630–642. https://doi.org/10.1109/TIP.2015.2507401 36

[170] Lei Wu, Rong Jin, and Anil K. Jain. 2013. Tag Completion for Image Retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 35, 3 (2013), 716–727. https://doi.org/10.1109/TPAMI.2012.124 39

[171] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R. Dick, and Anton van den Hengel. 2016. What Value Do Explicit High Level Concepts Have in Vision to Language Problems?. 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV,

USA, June 27-30, 2016. IEEE Computer Society, 203–212. https://doi.org/10.1109/CVPR.2016.29 8

[172] Qi Wu, Chunhua Shen, Peng Wang, Anthony R. Dick, and Anton van den Hengel. 2018. Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. IEEE Trans. Pattern Anal. Mach. Intell. 40, 6 (2018), 1367–1381. https://doi.org/10.1109/TPAMI.2017.2708709 9

[173] Zhaohui Wu and C. Lee Giles. 2013. Measuring Term Informativeness in Context. Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (Eds.). The Association for Computational Linguistics, 259–269. https://www.aclweb.org/anthology/N13-1026/ 57

[174] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, Xin Liu, and Ming Zhou. 2020. XGPT: Cross-modal Generative Pre-Training for Image Captioning. CoRR abs/2003.01473 (2020). arXiv:2003.01473 https://arxiv.org/abs/2003.01473 77

[175] Hao Xu, Jingdong Wang, Xian-Sheng Hua, and Shipeng Li. 2009. Tag refinement by regularized LDA. Proceedings of the 17th International Conference on Multimedia 2009, Vancouver, British Columbia, Canada, October 19-24, 2009, Wen Gao, Yong Rui, Alan Hanjalic, Changsheng Xu, Eckehard G. Steinbach, Abdulmotaleb El-Saddik, and Michelle X. Zhou (Eds.). ACM, 573–576. https://doi.org/10.1145/1631272.1631359 39

[176] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (JMLR Workshop and Conference Proceedings), Francis R. Bach and David M. Blei (Eds.), Vol. 37. JMLR.org, 2048–2057. http://proceedings.mlr.press/v37/xuc15.html 9, 77

[177] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 1358–1368. https://doi.org/10.18653/v1/d18-1166 71

[178] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas J. Guibas, and Fei-Fei Li. 2011. Human action recognition by learning bases of action attributes and parts. IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011, Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool (Eds.). IEEE Computer Society, 1331–1338. https://doi.org/10.1109/ICCV.2011.6126386 8

[179] Jian Yao, Sanja Fidler, and Raquel Urtasun. 2012. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012. IEEE Computer Society, 702–709. https://doi.org/10.1109/CVPR.2012.6247739 39

[180] Alexander Yates, Michele Banko, Matthew Broadhead, Michael J. Cafarella, Oren Etzioni, and Stephen Soderland. 2007. TextRunner: Open Information Extraction on the Web. HLT-NAACL (Demonstrations). The Association for Computational Linguistics, 25–26. 13

[181] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguistics 2 (2014), 67–78. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/229 32, 45

[182] ChengXiang Zhai. 2008. Statistical Language Models for Information Retrieval. (2008). https://doi.org/10.2200/S00158ED1V01Y200811HLT001 27

[183] Baopeng Zhang, Yanyun Qu, Jinye Peng, and Jianping Fan. 2017. An automatic image-text alignment method for large-scale web image retrieval. Multim. Tools Appl. 76, 20 (2017), 21401–21421. https://doi.org/10.1007/s11042-016-4059-x 55

[184] Zhichen Zhao, Huimin Ma, and Shaodi You. 2017. Single Image Action Recognition Using Semantic Body Part Actions. IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. IEEE Computer Society, 3411–3419. https://doi.org/10.1109/ICCV.2017.367 8

[185] Bolei Zhou, Àgata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 487–495. http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database 8, 56, 61

[186] Ning Zhou and Jianping Fan. 2015. Automatic image-text alignment for large-scale web image indexing and retrieval. Pattern Recognit. 48, 1 (2015), 205–219. https://doi.org/10.1016/j.patcog.2014.07.001 55

[187] Yimin Zhou, Yiwei Sun, and Vasant G. Honavar. 2019. Improving Image Captioning by Leveraging Knowledge Graphs. IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019. IEEE, 283–293. https://doi.org/10.1109/WACV.2019.00036 9

[188] Guangyu Zhu, Shuicheng Yan, and Yi Ma. 2010. Image tag refinement towards low-rank, content-tag prior and error sparsity. Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010, Alberto Del Bimbo, Shih-Fu Chang, and Arnold W. M. Smeulders (Eds.). ACM, 461–470. https://doi.org/10.1145/1873951.1874028 39

[189] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal Summarization with Multimodal Output. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 4154–4164. https://doi.org/10.18653/v1/d18-1448 77

[190] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. IEEE Computer Society, 19–27. https://doi.org/10.1109/ICCV.2015.11 55

List of Figures

List of Tables

Abbreviations

BERT Bidirectional Encoder Representations from Transformers. 19

BLEU Bilingual Evaluation Understudy. 21, 60, 80

CBIR Content Based Image Retrieval. 8, 35, 36, 38

CIDEr Consensus-based Image Description Evaluation. 23, 80

CNN Convolutional Neural Network. 9, 16, 38, 39

CSK Commonsense Knowledge. 3, 26, 27, 29, 30, 31, 32, 33

GT Ground Truth. 60

ILP Integer Linear Program(ing). 14, 37, 42, 61, 67

IR Information Retrieval. 11, 17, 21

LM Language Model. 11, 28

LSTM Long Short Term Memory Network. 16, 79

METEOR Metric for Evaluation of Translation with Explicit Ordering. 22

NLP Natural Language Processing. 3, 11, 17, 21, 24

OpenIE Open Information Extraction. 13, 31

POS Part of Speech. 13

RNN Recurrent Neural Network. 9, 12, 16, 39

ROUGE Recall-Oriented Understudy for Gisting Evaluation. 22, 60, 80

SANDI Story-AND-Images Alignment. 52

SPICE Semantic Propositional Image Caption Evaluation. 23, 80

TBIR Tag Based Image Retrieval. 8, 35, 36, 38

TF-IDF Term Frequence – Inverse Document Frequency. 58

VISIR VIsual and Semantic Image-label Refinement. 35, 61

VSE Visual-Semantic Embedding. 8, 9, 61

External Datasets and Tools

BERT Bidirectional Word Embeddings. `https://github.com/google-research/bert`. 79

Flickr 30K Image Captioning Dataset. `https://www.kaggle.com/hsankesara/flickr-image-dataset`. 45

Gurobi Integer Linear Program Optimizer `https://www.gurobi.com/`. 42, 61

ImageNet Image dataset following the WordNet hierarchy. `http://www.image-net.org/`. 25, 44

LSDA Object detection tool. `https://github.com/jhoffman/lsda`. 25, 29, 32, 35, 36, 37, 44, 45, 61, 113

MSCOCO Microsoft Common Objects in Context. `https://cocodataset.org/#home`. 20, 32, 45, 62

Pascal Sentence Dataset Image captioning dataset. `https://vision.cs.uiuc.edu/pascal-sentences/`. 32, 45

PlacesCNN MIT Scenes Dataset. `http://places.csail.mit.edu/downloadCNN.html`. 61

SBU Captions Dataset Image captioning dataset. `http://www.cs.virginia.edu/~vicente/sbucaptions/`. 32, 45

Word2Vec Word Embeddings. `https://code.google.com/archive/p/word2vec/`. 41

YOLO Object detection tool. `https://pjreddie.com/darknet/yolo/`. 35, 39, 45, 61