Abstract of "Data-driven Image Captioning" by Rebecca Mason, Ph.D., Brown University, May 2016.

Human communication is naturally multimodal. On the web, images frequently appear alongside text: for example, product images and descriptions on shopping websites, or social media users commenting on an image or a video. Image captions can serve many purposes: describing the salient content of an image, giving background information that is relevant to understanding the image, and allowing for images to be indexed and retrieved on search engines. Automatic image captioning is a challenging task involving several open problems in the fields of Natural Language Processing (NLP) and Computer Vision (CV).

This thesis presents work toward image captioning methods that learn from weakly-supervised examples of previously captioned images. These approaches employ text-to-text natural language generation techniques, which generate image captions by adapting text from captions of visually similar images. Using automatic and human evaluations, we demonstrate that our models can produce coherent and informative captions of images.

The work in this thesis will help enable the development of data-oriented image captioning systems which can be used to generate captions that describe the same relevant features that are described by humans, even in specific domains with few CV resources.

Data-driven Image Captioning

by
Rebecca Mason
A. B., Mount Holyoke College, 2009
Sc. M., Brown University, 2011

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Computer Science at Brown University

Providence, Rhode Island
May 2016

This dissertation by Rebecca Mason is accepted in its present form by
the Department of Computer Science as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date _____ _____

Eugene Charniak, Director

Recommended to the Graduate Council

Date _____ _____

James Hays, Reader

Date _____ _____

Yejin Choi, Reader
(University of Washington)

Approved by the Graduate Council

Date _____ _____

Dean Weber
Dean of the Graduate School

A picture is worth one thousand words.

Chinese Proverb

As the Chinese say, one thousand and one words is worth more than a picture.

John McCarthy

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is concerned with the task of automatically generating image captions. In general, image captioning refers to the following problem: given an image, generate text that describes the image. Automatic captioning methods for images (as well as video and other multimedia) are intended to reduce the amount of human labor needed for organizing, retrieving, and analyzing digital media.

1.1 Applications of Image Captioning

In recent years, digital photography has become cheap and ubiquitous, and the cost to store and retrieve large amounts of image data has also decreased. For social media, image sharing is a popular mode of communication, and is used not just for communicating with friends and family, but also for citizen journalism and activism. In scientific and medical fields, images are used for things like cataloging and identifying types of species and diseases, or capturing the output of experiments. Images are also collected and stored by professionals in many other fields, such as history and the arts.

There is a wide variety of potential applications for image captioning technology:

Background and Context Text that is associated with an image can provide information to help a user understand what they are seeing.

Search and Retrieval Humans often use natural language to describe images that they wish to search for. Image captions provide text to compare against search queries.

Accessibility Captions provide an alternate way to access information that is shared in an image, in cases where the user is unable to access or view the image directly. This is important not just for blind and visually impaired users, but also for users who access the internet using mobile devices or limited bandwidth connections.

Decision Making Image captions can help users to make decisions about what they are seeing in an image. For example, online shopping websites use captions alongside images of products to

help users understand situations where they might use the product. In a scenario where the computer is making a decision, image captioning could be used to ask a human for help or feedback.

1.2 Contributions of this Thesis

This thesis presents work toward data-driven approaches to image captioning. In this work, vison and language features are learned jointly in a statistical model that is trained on images and human-authored captions found on the web.

Specifically, this thesis introduces the use of *text-to-text natural language processing* for generation of image captions. The text-to-text generation process begin with some source text as input, and then adapts that text in order to meet the objective for some output. By collecting examples of images and human-written captions that are similar to those that we wish to produce in our system, we can create models that generate captions for new images that are taken in a similar context.

This thesis presents two new techniques using text-to-text generation for image captioning. First, we propose a nonparametric model for estimation of the content of a query image by examining human-written captions of images with similar spatial features. *Our model produces captions that are 48% more relevant than the previous best approach using spatial feature matching, and 34% more relevant than the previous best approach using a more complex visual model*, according to human judgments of relevance.

Second, we describe a new task setting for *domain-specific image captioning* in which many relevant visual details cannot be captured by off-the-shelf visual entity extractors. *We develop a joint model of visual and textual bag-of-word features, and use this model to adapt existing human-authored captions to new query images. We implement our model using a large, unlabeled dataset of women's shoes images and natural language descriptions* (Berg et al., 2010) *and use human and automatic evaluation to demonstrate the effectiveness of our proposed method.*

An additional contribution of this thesis is that we present an examination of datasets and evaluation techniques for previous work on annotating online images. While image caption generation is a very new task, there is a longer history of previous work generating image keyword annotations by selecting words from text that is associated with the query image – such as an image that appears on a website or next to a news article. Examining this related work gives us insight into some of the challenges in designing and evaluating applications that combine Natural Language Processing and Computer Vision. *We reimplement previous image annotation models and present a series of simple baselines that outperform previously published systems for this task.* We describe how these results motivate the approaches used in our image captioning research.

1.3 Outline

Chapter 2 provides background on relevant work in Computer Vision and Natural Language Generation, a brief summary of existing image captioning research, and conventional methods of evaluation.

Chapter 3 describes our work examining datasets and evaluation for online image annotation. Work in this chapter is based on Mason and Charniak (2012).

Chapter 4 introduces our nonparametric content estimation model for image captioning. We describe a dataset and prior work in image caption retrieval, present our model, and describe the experimental tasks used for evaluation. This work is previously described in Mason and Charniak (2014b).

Chapter 5 describes our setup for domain-specific image captioning, the dataset we use, and our models for content estimation and for adapting existing captions. This work was introduced in Mason and Charniak (2013) and Mason (2013) and the final model and experiments were presented in Mason and Charniak (2014a).

Finally, Chapter 6 gives the conclusions of this study, and suggestions for future work.

Chapter 2

Background

The main components of the automatic captioning process are image understanding and language generation. Each are very difficult problems, motivating entire communities of researchers from the fields of Natural Language Processing (NLP) and Computer Vision (CV). This chapter provides a brief background of work that is relevant for understanding existing approaches to image captioning. This is followed by a summary of existing image captioning research, including approaches to evaluation and what are the attributes of a good image caption.

2.1 Image Understanding

Image understanding refers to the process of determining the content and meaning of a source image. This process is typically associated with Computer Vision, a field of research concerned with automatically reconstructing properties of the real world according to visual input (Szeliski, 2010). However there are some applications where this information can be recovered from alternate sources, such as meta-information, or related text. This section briefly reviews image understanding approaches which are relevant to automatic image captioning.

2.1.1 Computer Vision

Visual Object Recognition

Visual object recognition is one of the central problems of CV research, and is an important component of many automatic image captioning approaches. The core problem is to learn generic categories of visual objects, and to locate and identify new instances of these categories (Grauman and Leibe, 2010). The human vision system is able to perform this task with very little effort, considering the difficulty of the task. Consider, for example, the task of recognizing a table. There are many different kinds of tables, such as a dining table, a workbench, an operating table, or a ping-pong table. Yet humans are able to recognize that these all belong to the same conceptual category, and infer the identity of objects that have not been seen before. Even the exact same object can vary in

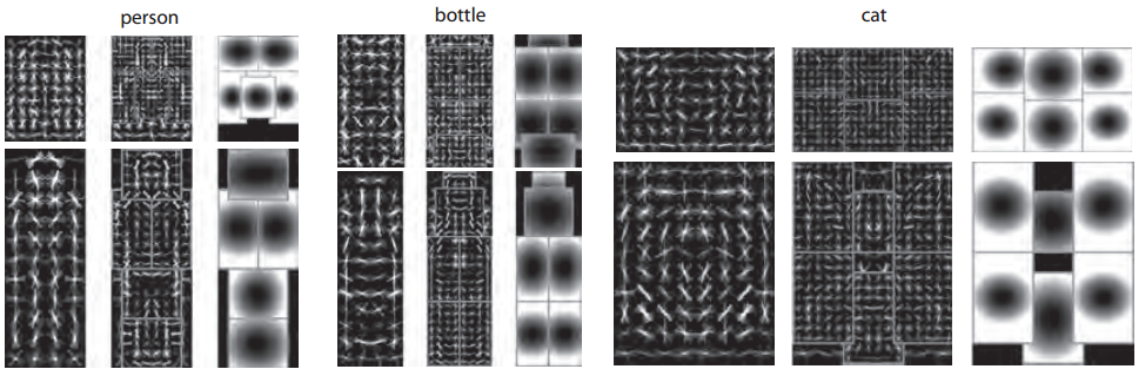


Figure 2.1: Representations from the Deformable Part-Based Model (Felzenszwalb et al., 2010, 2008) for categories, learned on the PASCAL (Everingham et al., 2008) dataset. Figure from Felzenszwalb et al. (2010).

appearance in different images. This may be due to differences in viewpoints or illumination, or the view of the object may be partially occluded by another object or cut off by the edge of the image frame.

The automatic object recognition system which has most prominently been used for image captioning is the Deformable Part-Based Model (Felzenszwalb et al., 2010, 2008). Part-based models are particularly helpful for recognizing object categories such as humans which appear in different poses. It represents images using low-level HOG features (Dalal and Triggs, 2005), which measure the direction of the change of intensity at different parts of the image. To train their object detector, they match the movable parts of the object in the training image, such as wheels on a bicycle, or limbs on a person. They then use a latent SVM to discriminatively learn the different objects. Figure 2.1 shows examples of learned representations. Supervised models such as the Deformable Part-based Model require images with labeled instances of objects for training. Typically each label corresponds to a "bounding box" that indicates the location of the object in the image.

Scene Recognition

Another fundamental problem in Computer Vision is scene recognition. Many scenes can be characterized by their global spatial properties. The well-known GIST (Oliva and Torralba, 2001) feature is a global image descriptor related to perceptual dimensions such as "naturalness", "roughness", and "ruggedness". These features are coarsely localized in order to describe the structure of the image. Another well-known global image descriptor is the TinyImage descriptor (Torralba et al., 2008), which resizes the image to a 32x32 thumbnail, so that the structure of the scene can be described using the overall layout of the colors in the thumbnail image.

Both GIST and TinyImage descriptors can be used for classifying types of scenes, such as of different kinds of scenes: beach, forest, city street, and so on. They can also help in recognizing different attributes of scenes (Patterson and Hays, 2012), such as man-made vs natural environments,

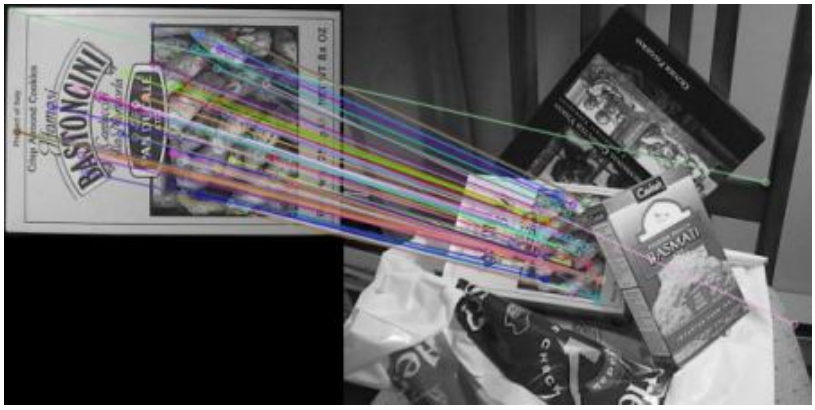


Figure 2.2: Interest point matching using SIFT descriptors. Image from Lowe (1999). SIFT features generated by OpenCV (http://opencv.org/).

or indoor lighting vs outdoor lighting. Finally, scene-level image descriptors can also be used as a measure for comparing images in many data-driven Computer Vision applications. These methods reduce an inference problem for an unknown image to finding an existing labeled image that is similar. For example, the Im2Text system (Ordonez et al., 2011) finds a caption for an image using GIST and TinyImage descriptors to find the most similar image from a database of captioned images.

Feature Recognition

Finally, there are also "bag-of-image-word" visual features which are computed at various points on the image, and contain information such as the color, shape, texture, or lighting at that point. Like a bag-of-words model for text, the bag-of-image-words model does not consider the position of the features in the image. The features are quantized in to discrete words using the k-means algorithm. A single "image word" does not carry semantic meaning like a single word of English text.

Some standard bag-of-image-word features that are often used in CV are SIFT (Lowe, 1999), HOG (Dalal and Triggs, 2005), and Textons (Leung and Malik, 2001). A SIFT descriptor describes which way edges are oriented at a certain point in an image (Lowe, 1999). Bag-of-HOG (histogram of gradients) features describe gradients and curvature at a point (Dalal and Triggs, 2005). For texton features, images are convolved with Gabor filters at multiple orientations and scales, sampled at the locations where the image words will be (Leung and Malik, 2001).

Figure 2.2 shows an example of SIFT features being used on the interest points of an object. This allows the same object to be recognized in a different image despite differences in scale and rotation.

2.1.2 Non-visual Approaches

There are also non-CV approaches to image understanding, which are used for image captioning, retrieval, and annotation. Image search engines on the web, such as `images.google.com` typically

only use text that is related to the image in order to decide which images to retrieve for a query. Previous work has used related text and meta-information such as an article related to a news image (Deschacht et al., 2007; Feng and Lapata, 2010b,a), the webpage where the image comes from (Leong et al., 2010), or the GPS coordinates where the image was taken (Fan et al., 2010).

2.2 Natural Language Generation

Natural language generation is an area of NLP that deals with the automatic production of text or speech according to a certain input (Jurafsky and James, 2009). Generation methods are often categorized as either concept-to-text methods, which produce textual output from non-linguistic input; or as text-to-text methods which produce textual output using input text from human-authored sources (Reiter et al., 2000). However, much previous work in image captioning uses a hybrid of these approaches.

2.2.1 Concept-to-text Generation

The most basic steps of a traditional concept-to-text generation pipeline are selection of content to be in the output text, and realization of the natural language output. Content selection is determined by the input data – such as the output of a visual detection system – as well as the communication objective for the output, and a set of constraints capturing linguistic or other knowledge. This objective may be reached using various AI planning algorithms (Hovy, 1991; Koller and Stone, 2007).

Surface realization is a linguistic process of constructing a sentence using the choices of words and syntactic structures found in the content selection stage (Prevost and Steedman, 1993). It involves applying morphological and syntactic rules so that the output text sounds natural and correct. The rules governing this process are relatively well-understood, and there are several software systems available for realization (Bateman, 1997; Gatt and Reiter, 2009). However, understanding which concepts are important, and selecting words and phrases to describe that content, is still an open research question.

2.2.2 Text-to-text Generation

In text-to-text generation, content is typically specified by some textual input source. The objective is to preserve the meaning of the input text, while transforming it to better meet the communication objective. Some examples of text-to-text generation are:

Summarization : Generating a summary that contains only the most important information in a document or group of documents. Extractive summarization methods select relevant sentences from the original document or documents and using that text as the summary (Nenkova and Vanderwende, 2005; Haghighi and Vanderwende, 2009). Abstractive summarization methods

generate novel sentences to describe relevant content from the source documents (Murray et al., 2010; Cheung and Penn, 2013).

Compression : Decreasing the length of an input sentence by deleting words that are not relevant, without making the sentence ungrammatical (Clarke and Lapata, 2008; Martins and Smith, 2009).

Paraphrasing : Rewording and rearranging phrases or sentences in a different way from the original (Barzilay and Lee, 2003).

Simplification : Rewriting a sentence to make it easier to understand (Vanderwende et al., 2007; Zhu et al., 2010).

Fusion : Combining the relevant content of two sentences into one single sentence (Barzilay and McKeown, 2005; Elsner and Santhanam, 2011).

Text-to-text generation methods are typically guided by some notion of relevance. In some cases, relevance is determined using intrinsic qualities of the input text, such as the frequency of a word in a document, or the positions of noun phrases in the grammatical structure of the text. Outside sources, including non-linguistic information, can also be used to guide selection of relevant content.

2.3 Image Description

There is a variety of interesting research at the intersection between Computer Vision and Natural Language Processing. In this section, we provide background on two main tasks in describing images: generating keyword annotations, and generating full-sentence image captions.

2.3.1 Image Annotation

Image annotation is the task of taking in an image and generating relevant descriptive keywords that describe the visual content of the image. Image annotation is an important area of research with applications such as tagging, indexing, and retrieval. Keyword annotations can be used to approximate the content of the query image, and as a source of content words for generating an output caption (Feng and Lapata, 2010a).

The Computer Vision literature contains countless approaches to this task, using image understanding techniques (Section 2.1) to select annotation keywords for a query image. A survey of methods for image annotation be found in Hanbury (2008).

In addition to Computer Vision approaches, there is research using Natural Language Processing to discover visually descriptive keywords. Text processing is computationally less expensive than image processing and may provide information that is difficult to learn visually. Instead of selecting descriptive keywords according to a visual-to-textual representation dictionary, descriptive words can be mined from natural language text that is associated with the image. Most commercial image

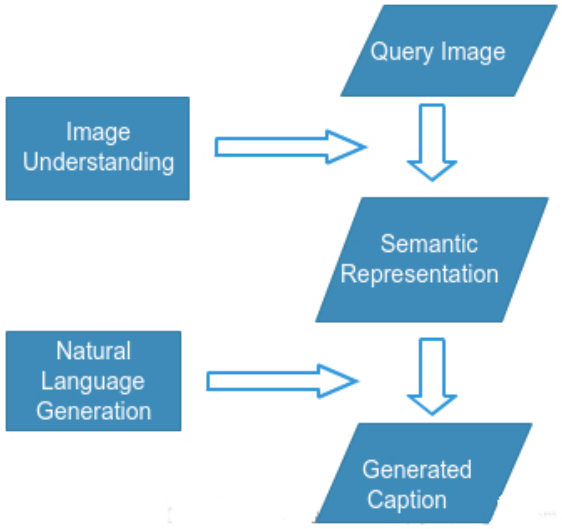


Figure 2.3: A general pipeline for image captioning.

search websites use surrounding text as a source of information for understanding the content of an image (Frankel et al., 1996).

Some examples of text that may be used are text or captions on the webpage that contains the image, the title of the webpage, or the URL or filename of the image itself. For example, Deschacht et al. (2007) use named-entity recognition on news articles to identify people in images that are associated with the article. Feng and Lapata (2010b) learn topic models to learn descriptive keywords for news images given both the image and the associated news article. Boiy et al. (2008) and Leong et al. (2010) use term association to estimate the "visualness" of words in order to select words that are likely to describe visual content of images. The Wikipedia Retrieval Task at ImageCLEF 2011 (Tsikrika et al., 2011) compared several approaches to image retrieval on Wikipedia, including both systems which rely solely on text-based approaches and systems which incorporate Computer Vision approaches.

2.3.2 Image Captioning

In recent years, there has been an increasing interest in systems that describe images using natural language – phrases or captions, rather than keyword-length descriptions. The objective is to generate image captions which describe the relevant content in the image. Natural language captions are helpful for describing the relationships between objects in images, or for describing images to humans.

An image caption is the output of a complex process which involves understanding the query image, grounding the visual representation to a semantic representation of what is relevant in the image, and then natural language generation of the output caption. Figure 2.3 shows an example pipeline of an image captioning system. However, exact formulations of the image captioning task

Query Image		
Caption (automatic)	Last year, thousands of Tongans took part in unprecedented demonstrations to demand greater democracy and public ownership of key national assets.	Contaminated Cadbury's chocolate was the most likely cause of an outbreak of salmonella poisoning, the Health Protection Agency has said.
Caption (human)	King Tupou, who was 88, died a week ago.	Cadbury will increase its contamination testing levels.

Table 2.1: Example query images from Feng and Lapata (2010a), with automatic captions generated by extracting text from a related text document, and human-authored captions for comparison.

vary across previously published image captioning methods.

This thesis will focus on models that can learn from images and captions that co-occur naturally on the web. Therefore, we primarily focus on more recent image captioning work that seeks to integrate both image understanding and natural language generation components.

Generating Captions for Images with Associated Text Documents

A handful of approaches have been proposed in the literature to incorporate knowledge in the form of text documents into the image captioning pipeline. For each query image, we assume that we are given or are able to retrieve a related text document. The output caption should contain a summary of information in the document that is relevant to the query image. This task formulation is similar to query-focused automatic summarization, but with an image serving as the focus for the output summary.

For example, Feng and Lapata (2010a) generate captions of news images using both summarization methods on the news articles that appear with each image. They use a joint model of visual and textual information to select relevant content. They integrate these models using a topic model based on Latent Dirichlet Allocation (Blei et al., 2003) which incorporates a bag-of-words model of the article text, and a bag-of-image-words model of SIFT features computed from the query image. Once the relevant content is identified, they present methods for caption generation using both extractive and abstractive summarization. Table 2.1 shows examples of query images, human-authored captions, and captions automatically generated using the extractive method.

Aker and Gaizauskas (2010) and Fan et al. (2010) present image captioning systems that model image content using GPS coordinates of where the image was taken – information which is often recorded by cameras in mobile phones. These systems retrieve text documents related to landmarks near where a query photo was taken, then generate concise summaries of those text documents.

A related task which has seen some interest in the natural language processing community is generation of descriptions for information graphics (Mittal et al., 1995; Greenbacker et al., 2011; Demir et al., 2012). Information graphics such as line graphs and plots exist in many documents, but the information contained in them is often not described in the document, and inaccessible to users such as the visually impaired. In these systems, the content of the image is determined by directly accessing the data used to generate the information graphic, as well as analyzing the accompanying text document.

One shortcoming of these approaches to image captioning is that not all of the text in the related document will be related to the visual content in the image. For example in Table 2.1, the automatic caption on the left gives information that is not relevant to what is shown in the image. Another concern is that these methods rely on extra data outside of the content of the image, since they assume that the related text document will be available. These approaches are most applicable for specific domains (e.g., news, travel, financial reports) for which it can be assumed that these documents exist and can be retrieved in a structured way.

Generating Captions for Images by Caption Transfer

Data-driven matching methods have shown to be effective for a variety of complex problems in Computer Vision. These methods reduce an inference problem for an unknown image to finding an existing labeled image which is semantically similar. When generating captions for query images which do not have a corresponding text document available, one can reduce the captioning problem to finding a semantically similar captioned image, and transferring the existing caption to the query image.

The IM2TEXT model by Ordonez et al. (2011) presents the first web-scale approach to image caption generation. IM2TEXT retrieves an image which is the closest visual match to the query image, and transfers its description to the query image. Visual matches are computed using a combination of visual object detectors and scene based descriptors such as TinyImage and GIST. Ordonez et al. (2011) also present a new corpus, the SBU-Flickr dataset[1], which is made of 1 million images and human-authored captions uploaded by users of the website `flickr.com`.

Kuznetsova et al. (2012) present a related approach also using the SBU-Flickr corpus which uses trained CV recognition systems to detect a variety of visual entities in the query image. A separate description is retrieved for each visual entity, which are then fused into a single output caption. Like IM2TEXT, their approach uses visual similarity as a proxy for textual relevance, but their sentence fusion approach allows greater flexibility for generating output that matches the visual content.

Other related work models the text more directly, but is more restrictive about the source and quality of the human-written training data. Farhadi et al. (2010) and Hodosh et al. (2013) learn joint representations for images and captions, but can only be trained on data with very strong alignment between images and descriptions (i.e. captions written by Mechanical Turkers).

[1] `http://vision.cs.stonybrook.edu/~vicente/sbucaptions/`

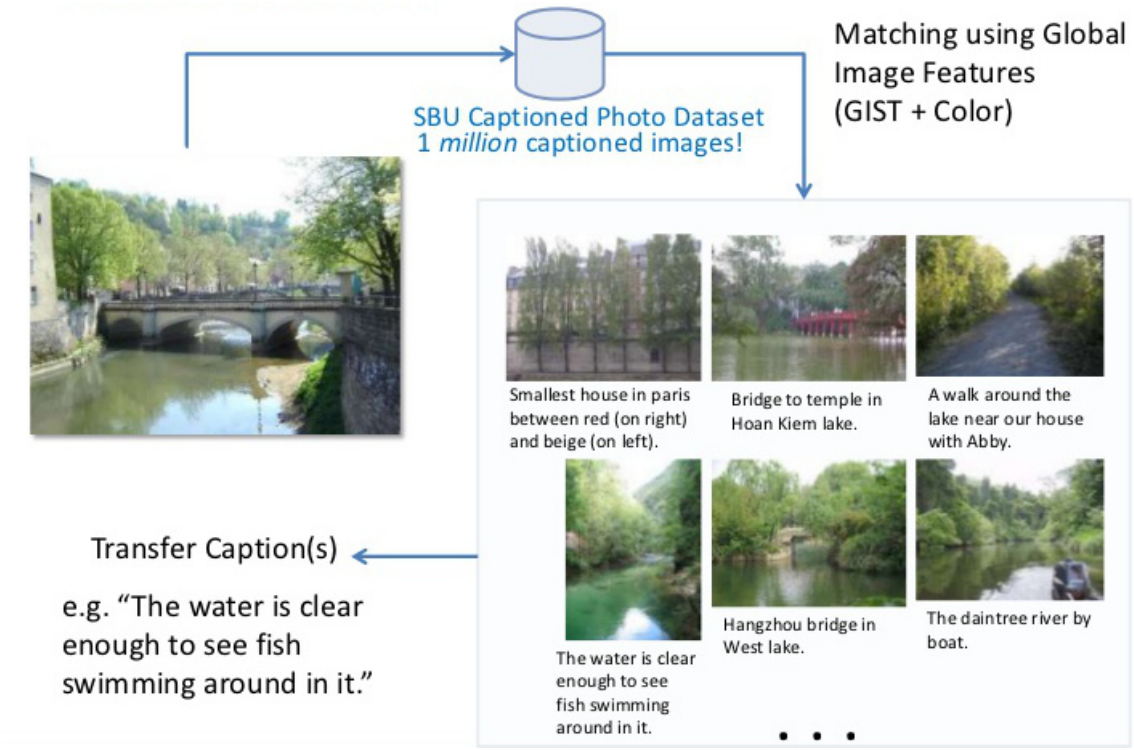


Figure 2.4: An illustration of the Im2Text captioning system from Ordonez et al. (2011). For the query image, the system selects a captioned image that is the closest match according to visual similarity. The caption from the matched image is selected to caption the query image as well.

Integrating Visual and Linguistic Models

The most recent approaches to image captioning integrate image understanding with linguistic models. This allows for smoothing of the noisy visual detection scores using knowledge learned from large linguistic corpora.

For example, Kulkarni et al. (2011) uses a conditional random field to predict the most likely labeling of objects in a scene, incorporating both the detection scores with text-based potentials computed from large text corpora. Predicted labels are used to complete sentence templates which provide form for the generated captions. Template-based generation is also used by Yang et al. (2011), who use an HMM-based approach. In addition to correcting noisy initial detections, the linguistic model can also be used to predict verbs and preposition words which are difficult to determine visually.

Later work such as Li et al. (2011) and Mitchell et al. (2012) generate more natural-sounding captions using more flexibile models that learn from examples of n-grams and syntactic structures in larger text corpora. Yu and Siskind (2013) present a model that learns the alignment between individual words and object detection classes.

Finally, the past few months have seen a large amount of interest in exploiting deep neural networks for the task of image captioning (Vinyals et al., 2014; Karpathy and Fei-Fei, 2014; Kiros

et al., 2014; Donahue et al., 2014; Mao et al., 2014; Fang et al., 2014). Further progress in this area of research will likely lead to models which provide richer representations of visual scenes.

2.4 Evaluation

Image captions are evaluated using the same basic techniques used for evaluation other kinds of automatically generated language: intrinsic evaluations in which humans rate or compare the quality of generated text; extrinsic evaluations in which peformance is measured by how much it helps humans can perform some task; and automatic metrics which compare generated text to a human-authored gold standard.

Intrinsic evaluations have humans rate image captions based on the quality of their language (grammaticality) and content (relevance to the image). Likert scales (or rating scales) are one approach that has been used for caption evaluation (Feng and Lapata, 2010a; Kuznetsova et al., 2012; Mitchell et al., 2012). The advantage of Likert scales is that they can be used to measure the degree to which one method is better than another. However, they can provide noisy measurements and require careful calibration, especially for human studies that are conducted over Mechanical Turk. Forced choice evaluations, where users must choose between two different captions that are shown, have also been used to evaluate image captioning systems (Ordonez et al., 2011; Kuznetsova et al., 2013). Forced choice evaluations are easier to perform over the internet, and show how often a proposed method improves over the baseline.

Extrinsic evaluations are rare in previous captioning work, since they are more difficult to measure, and because caption generation is such a novel topic that it is not yet clear which kinds of tasks will provide interesting research questions and practical applications in the long term. As an example, Table 2.2 shows some images found on the web, and some examples of captions that could be used to describe the images in different contexts. If we were to perform the discrimination task from Ordonez et al. (2011) using these two images, all of the example captions would perform equally well, because none of them could be applied to the other image. The three captions under each image all choose to include different information: a thorough description of the entire scene, labeling only the relevant nouns in the scene, or providing information about the context where the image was taken.

Automatic metrics compare generated text to a human-authored gold standard. Automatic evaluation metrics are useful for system development, where they can provide feedback much quicker than the time it would take to run a human evaluation. They are also useful for situations where human evaluations are very expensive, due to requiring human evaluators with expensive expertise. However, it is difficult to provide a fair gold-standard for human-authored texts, as human authors generally have a variety of different ways to express the same idea. Automatic metrics also have difficulty measuring the grammaticality or coherence of a generated text. However, for a large enough test set, automatic metrics often capture significant differences in how well different generation systems capture content words that are in the human-authored texts.

	
A red bird is standing on a glass table against a green background.	A man wearing a blue suit and a woman wearing black clothes are standing behind a microphone. A curtain is in the background.
Northern Cardinal **	The president of the United States, Barack Obama and the vice chairwoman of the Federal Reserve, Janet Yellen.
Photo taken while trying out my new camera in the backyard.	Janet L. Yellen, 67, would be the first woman to lead the Federal Reserve if the Senate confirms her nomination for a four-year term. **

Table 2.2: Images from the web, along with example captions that could serve different purposes. Captions followed by ** are the original captions.

In NLP, there are a variety of specialized metrics for evaluating different kinds of generated language, but there is no metric that is specifically developed for the task of evaluating generated image captions. In previous image captioning work, the BLEU (Papineni et al., 2002) metric is most frequently used (Farhadi et al., 2010; Kulkarni et al., 2011; Ordonez et al., 2011; Kuznetsova et al., 2012). The ROUGE (Lin, 2004) metric has been also used for evaluating generated image captions (Yang et al., 2011), and recent work shows that ROUGE scores correlate better with human judgments of image caption quality (Elliott and Keller, 2014).

Here, we describe both the BLEU and ROUGE metrics in greater detail, as they are used (in addition to human evaluations) in our experiments in later chapers.

2.4.1 BLEU

BLEU (Papineni et al., 2002) is a modified unigram precision metric, originally developed for evaluating automatic translations. The quality of a machine translation is considered to be how well it resembles a translation generated by a professional human translator. BLEU scores range from 0 to 1, where scores closer to 1 are assumed to be closer to that of a human. Even a human translator will rarely achieve a perfect BLEU score, because there may be many correct translations for the same source text. BLEU scores are highly correlated with professional judgments of translation quality (Papineni et al., 2002; Coughlin, 2003).

In order to compute BLEU scores, it is necessary to have a reference translation to compare against. When available, BLEU may consider multiple reference translations, in order to cover more of the possible correct translations. Using an unmodified precision metric, it is possible to achieve very high scores by repeatedly guessing frequent words in the target language. BLEU modifies the precision metric to discourage this kind of guessing.

2.4.2 ROUGE

ROUGE (Lin, 2004) is a recall-oriented metric developed for summarization evaluation. Similar to BLEU, ROUGE is a measure of how well a candidate text resembles human-generated text. There are different ROUGE measures which count the number of overlapping units such as ngrams, word sequences, and word pairs between the candidate text and the ideal text created by humans. ROUGE is a widely-used metric for summarization evalution, and is used to compare systems in shared summarization tasks such as DUC[2] and TAC[3].

[2] http://duc.nist.gov/

[3] http://www.nist.gov/tac/

Chapter 3

Evaluating Image Annotations

This chapter presents our work examining the baselines used to evaluate image annotations that are generated by Natural Language Processing systems. In these systems, descriptive keywords are generated for a query image by selecting words from some text that is associated with the image to be annotated. In this chapter, we develop a series of baseline measures for this task, inspired by baselines used in Information Retrieval, Computer Vision, and extractive summarization. We compare the results on two recent datasets (Section 3.1) developed for the task of image annotation using Natural Language Processing techniques.

The main contributions of this chapter are that we present image annotation methods that match or exceed the best published scores for image annotation on these datasets (Section 3.3.3 and 3.4.3), and provide an explanation of experimental practices that can lead to misleading results in research combining language and vision processing techniques (Section 3.5).

In the previous chapter, we introduced the task of image annotation. More specifically, we introduced image annotation approaches which use Natural Language Processing to examine text that is associated to a query image, and then select keywords from that text that are descriptive of the query image. In Section 3.1 we describe in detail two representative datasets used for this image annotation task, and the typical steps used to generate keyword annotations from natural language text.

In Section 3.2 we describe a series of baseline methods for image annotation which we then evaluate (in Sections 3.3 and 3.4) on the two datasets respectively. Many previously published annotation models are based on modeling association between visual and textual features, or modeling term relationships between different words in order to identify which words are more likely to describe visual content. In contrast, we develop image annotation models similar to frequency baselines used in automatic summarization. The performance of our approach demonstrates that image annotation with associated text is perhaps better thought of a summarization task. This finding motivates the approaches we use for image captioning in later chapters of this document.

3.1 Data and Preprocessing

Table 3.1 provides an overview of the datasets; while remainder of this section covers the source of the datasets and their gold annotations in more detail.

3.1.1 BBC Dataset

The BBC Dataset Feng and Lapata (2008)[1] contains news articles, image captions, and images taken from the BBC News website. Each training instance consists of a news article, and the associated image and caption from the same news story. The annotation keywords for the training set are generated by selecting "descriptive" words from the image captions. These words are defined as the nouns, adjectives, and certain kinds of verbs that are found in these captions.

To address the problem of converting natural language into annotations, a large amount of preprocessing is performed. In Feng and Lapata (2008), Feng and Lapata (2010b), and Feng and Lapata (2010a), the established preprocessing procedure for this dataset is to lemmatize and POS-tag using TreeTagger (Schmid, 1994). This leaves a total text vocabulary of about 32K words, which is further reduced by removing words that appear fewer than five times in the training set articles. Table 3.1 shows the number of word tokens and types after performing these steps.

3.1.2 UNT Dataset

The UNT Dataset (Leong et al., 2010)[2] consists of images and co-occurring text from webpages. The webpages are found by querying Google Image Search with frequent English words, and randomly selecting from the results.

Each image in UNT is annotated by five people via Mechanical Turk. In order to make human and system results comparable, human annotators are required to only select words and collocations that are directly extracted from the text, and the gold annotations are the count of how many times each keyword or collocation is selected. The human annotators write keywords into a text box; while the collocations are presented as a list of candidates and annotators mark which ones are relevant. Human annotators tend to select subsets of collocations in addition to the entire collocation. For example, the gold annotation for one image has "university of texas", "university of texas at dallas", "the university of texas", and "the university of texas at dallas", each selected by at least four of the five annotators. Additionally, annotators can select multiple forms of the same word (such as "tank" and "tanks"). Gold annotations are stemmed after they are collected, and keywords with the same stem have their counts merged. For this reason, many keywords have a higher count than the number of annotators.

[1] Downloaded from `http://homepages.inuf.ed.ac.uk/s0677528/data.html`

[2] Downloaded from `http://lit.csci.unt.edu/index.php?P=research/downloads`

Dataset:	**BBC**	**UNT**
data instances	article, image, and caption from a news story	image and text from a webpage
source of data	scraped from BBC News website	Google Image Search results
candidate keywords or collocations for annotation	descriptive unigram words from training data	$n \leq 7$-grams extracted from co-occurring text; collocations must appear as article names on Wikipedia
gold annotations	descriptive words from held-out image captions	multiple human-authored annotations for each image
evaluation metric	precision and recall against gold annotations	metrics adapted from evaluation of lexical substitutions (SemEval)
number of instances in training set	3121 articles, images, and image captions	299 websites, images, and human-generated keyword annotations (train using cross-validation)
number of instances in test set	240 articles and images, held-out image captions	300 websites, images, held out annotations
preprocessing procedure	lemmatize tokens, remove from dataset all words that are not descriptive or that appear fewer than five times in training articles	stem all tokens
average number of text tokens after preprocessing	169 tokens per article, 4.5 tokens per caption	278 tokens per webpage
average document title length	4 tokens	6 tokens
total vocabulary after preprocessing	10479 types	8409 types

Table 3.1: Comparison of the BBC (Feng and Lapata, 2008) and UNT (Leong et al., 2010) image annotation datasets.

3.2 Baselines

For this study, we retrieved the datasets and performed preprocessing on the text as described in the previous section. We then implement five baselines to compare against the captioning models in the previous work.

First, we implement the following two baselines, both of which are previously used for comparison in Feng and Lapata (2008), Feng and Lapata (2010b) and Leong et al. (2010):

Document Title The title of a document gives important information about the text of the full document. Document titles frequently contain keywords that are important to the body of the text, or even describe ideas at a more conceptual level than is contained in the full text (Montes-y Gómez et al., 2000).

In the BBC dataset, the headline for the news article is used as the document title. For the UNT dataset, the title of the webpage is used as the document title.

tf*idf Short for "term frequency–inverse document frequency", tf*idf is a statistic that gives weight to a term according to how important the term is in a particular document. It is a standard baseline used for information retrieval tasks, based on the intuition that a word that appears in a smaller number of documents is more likely to be meaningful than a word that appears in many documents.

tf*idf is the product of term frequency and inverse document frequency, where N is the number of documents, and n_i is the number of documents that contain the term t_i:

$$idf(t_i) = \log \frac{N}{n_i}$$

To run the tf*idf baseline on the BBC dataset, we base the idf weights on the document frequency of the training articles. On the UNT dataset, we follow Leong et al. (2010) who uses the British National Corpus to calculate the idf scores.[3]

Next, we implement our own baselines, inspired by previous work in automatic summarization and image annotation in Computer Vision.

Corpus Frequency Image annotation keywords tend to be distributed with with a relatively small number of frequently occurring keywords, and a long tail of keywords that only appear a few times. Related work for image annotation in the field of Computer Vision (Section 2.3.1) shows that the background frequency of a keyword in the corpus is a very powerful feature on its own. For UNT, we use the total keyword frequency of all the gold annotations, except for the one document that we are currently scoring. For BBC, we only measure the frequency of keywords in the training set captions, since we are specifically interested in the frequency of terms in captions.

[3] We also implemented a cross-validation tf*idf baseline where for each document we recalculate idf using the other 299 documents. But we did not get any meaningful change in the output annotations.

Term Frequency Term frequency has been shown to be a powerful feature in summarization (Nenkova and Vanderwende, 2005). Words that appear frequently in a document are considered more meaningful than infrequent words. Term frequency is the number of times a term (excluding function words) appears in a document, divided by the total number of terms in that document. For our image annotation baseline, the document is the text that appears alongside the image – the text of the website where the image appears, or the news article that corresponds with the image. On the UNT dataset we use the stopword list included with the MALLET toolkit (McCallum, 2002). For the BBC dataset, we compute term frequency of the "descriptive words" that were pulled out during preprocessing.

Sentence Extraction This baseline extracts the most central sentence from the co-occurring text, and uses descriptive words from that sentence as the image annotation. Unlike sentence extraction techniques from Feng and Lapata (2010a), our baseline selects a sentence based only on the text of the document, not the estimated content of the image. We extract the sentence with the minimum KL-divergence to the entire document, using the KLSum algorithm (Haghighi and Vanderwende, 2009).[4]

3.3 BBC Dataset Experiments

3.3.1 System Comparison

In addition to the baselines, we compare against the Mix LDA and Text LDA systems from Feng and Lapata (2010b). In Mix LDA, each instance is represented as a bag of textual features (unigrams) and visual features (SIFT features quantized to discrete "image words" using k-means). A Latent Dirichlet Allocation topic model (Blei et al., 2003) is trained on articles, images, and captions from the training set. Keywords are generated for an unseen image and article pair by estimating the distribution of topics that generates the test instance, then multiplying them with the word distributions in each topic to find the probability of textual keywords for the image. The Text LDA model is similar to Mix LDA but excludes the SIFT image words from both train and test instances.

3.3.2 Evaluation

In previous work using the BBC dataset Feng and Lapata (2008, 2010b); Leong et al. (2010), systems are evaluated by measuring precision and recall against the keywords found in the human-authored captions. For term frequency, tf*idf, corpus frequency, and the Mix LDA system, the generated annotation for each test image is its ten most likely keywords. We also run all baselines and the Mix LDA system on an unpruned version of the dataset, where infrequent terms are not removed from training data, test data, or the gold annotations. The purpose of this evaluation is to see if candidate keywords which are "unlearnable" for Mix LDA system can be learned by our systems.

[4]The KLSum algorithm is described in futher detail in the next chapter, Section 4.4.

	Standard			Include-infrequent		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Term Frequency	13.13	27.84	17.84	13.62	25.71	17.81
tf * idf	9.21	19.97	12.61	7.25	13.52	9.44
Doc Title	17.23	13.70	15.26	15.91	11.86	13.59
Corpus Frequency	3.17	6.52	4.26	3.17	6.02	4.15
Sentence Extraction	16.67	15.61	16.13	18.62	16.83	17.68
Mix LDA	7.30	16.16	10.06	7.50	13.98	9.76
Text LDA	8.38	17.46	11.32	7.79	14.52	10.14

Table 3.2: Image annotation results for previous systems and our proposed baselines on the BBC Dataset.

3.3.3 Results

The evaluation results for the BBC Dataset are shown in Table 3.2.[5] Term frequency is a stronger baseline than tf*idf. Since nearly all of BBC's function words are removed during preprocessing, the words that are downweighted by the idf score are common – but meaningful – words such as *police* or *government* which appear frequently in the BBC news articles.

Recall that Feng and Lapata (2008, 2010b) remove keywords that appear fewer than five times in the training section of the BBC corpus. They remove these keywords are removed because their experiments compare against Computer Vision models such as CorrLDA (Blei and Jordan, 2003) which need to see a keyword multiple times in the training set in order to learn a visual model. However, our experiments show that text-only baselines, such as term frequency, are able to predict some of these keywords. As shown in Table 3.3, many of the keywords which are removed from the corpus in preprocessing are meaningful to the content of the images.

While sentence extraction has a lower recall than term frequency, it is the only baseline or system that has improved recall when including infrequent words. This is unexpected because our baseline selects a sentence based on the term frequency of the document, and the recall for term frequency fell. One possible explanation is that extraction implicitly uses correlations between keywords. Probabilities of objects appearing together in an image are not independent; and the accuracy of annotations can be improved by generating annotation keywords as a set Moran and Lavrenko (2011). Recent works in image captioning also use these correlations: explicitly, using graphical models Kulkarni et al. (2011); Yang et al. (2011); and implicitly, using language models Feng and Lapata (2010a). In comparison, sentence extraction is very implicit.

[5] We are unable to reproduce the scores reported in Feng & Lapata Feng and Lapata (2008, 2010a,b). We have contacted the authors, who were unable to identify the reason for this discrepancy. All system and baseline scores presented on the BBC corpus are of our own implementation, and may not match those reported in previous publications.

	
Cadbury increase **contamination testing level**	malaria parasite **spread** mosquito

Table 3.3: Examples of gold annotations from the test section of the BBC Dataset. The bolded words are the ones that appear five or more times in the training set; the unbolded words appear fewer than five times and would be removed from both the candidate and gold keywords in the standard BBC evaluation.

3.4 UNT Dataset Experiments

3.4.1 System Comparison

We evaluate against the text mining system from Leong et al. (2010). Their system generates image keywords by extracting text from the co-occurring text of an image. It uses three features for selecting keywords.

Flickr Picturability queries the Flickr API with words from the text in order to find related image tags. Retrieved tags that appear as surface forms in the text are rewarded proportionally to their frequency in the text.

Wikipedia Salience assigns scores to words based on a graph-based measure of importance that considers each term's document frequency in Wikipedia.

Pachinko Allocation Model is a topic model that captures correlations between topics (Li and McCallum, 2006). PAM infers subtopics and supertopics for the text, then retrieves top words from the top topics as annotations.

There is also a combined model of these features. The feature weights are trained using an SVM with 10-fold cross-validation.

3.4.2 Evaluation

Evaluation on UNT uses a framework originally developed for the SemEval lexical substitution task (McCarthy and Navigli, 2007). This framework accounts for disagreement between annotators by weighting each generated keyword by the number of human annotators who also selected that keyword. The scoring framework consists of four evaluation measures: *best normal*, *best mode*, *oot* (out-of-ten) *normal*, and *oot mode*. Both the original framework and its adaptation by Leong et al.

	Best		Out-of-ten (oot)	
	Normal	**Mode**	**Normal**	**Mode**
Term Frequency	5.67	14.29	33.40	89.29
tf * idf	5.94	14.29	38.40	78.57
Doc Title	6.40	7.14	35.19	92.86
Corpus Frequency	2.54	75.00	8.22	82.14
Flickr Picturability	6.32	78.57	35.61	92.86
Wikipedia Salience	6.40	7.14	35.19	92.86
Topic Model (PAM)	5.99	42.86	37.13	85.71
Combined (SVM)	6.87	67.49	37.85	100.00

Table 3.4: Image annotation results for our proposed baselines, and the text mining systems from Leong et al. (2010)

(2010) give precision and recall for each of the evaluation measures. However, in practice, precision and recall are identical for all baselines and systems.

The two *best* evaluations find the accuracy of a single "best" keyword generated by the system. *Best normal* measures the accuracy for each system annotation a_j as the number of times a_j appears in the R_j, the multi-set union of human tags, and averages over all the test images.

$$Bestnormal = \frac{\sum_{i_j \in I} \frac{|a_j \in R_j|}{|R_j|}}{|I|}$$

In *oot normal*, up to ten unordered guesses can be made without penalty.

$$ootnormal = \frac{\sum_{i_j \in I} \frac{\sum_{a_j \in A_j} |a_j \in R_j|}{|R_j|}}{|I|}$$

where A_j is the set of ten system annotations for image i_j.

The *best mode* and *oot mode* metrics are the same as the *normal* metrics except they only evaluate system annotations for images where R_j contains a single most frequent tag. We use the scoring software provided by SemEval[6] with the gold annotation file provided in the UNT Dataset.

3.4.3 Results

The results of the lexical substitution evaluation on the UNT Dataset are shown in Table 3.4. For this corpus, the tf*idf baseline outperforms term frequency. This is because UNT is a web-based corpus with noisier text. Even though a stopword filter is used, many documents contain text such as copyright disclaimers which are not relevant to the content of the image.

Recall that the *mode* evaluation is only measured on data instances where the gold annotations have a single most frequent keyword. While running the evaluation script on the gold annotation file that came with the UNT dataset, we discover that SemEval only identifies 28 of the 300 instances as having a single mode annotation, and that for 21 of those 28 instances, the mode keyword

[6] `http://nlp.cs.swarthmore.edu/semeval/tasks/task10/data.shtml`

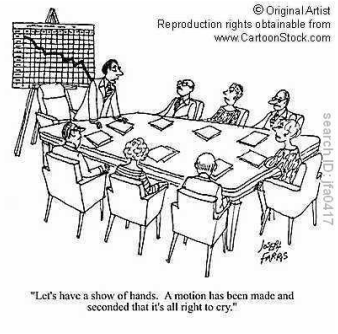
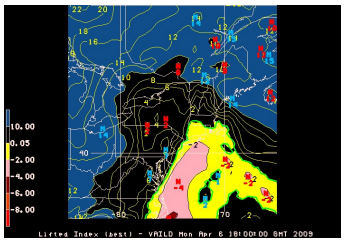
		
cartoon(6), market(5), market share(5), declin(3), imag(3), share(3), pictur(1), illustr(1), cartoonstock(1), origin(1), artist(1), meet(1), jfa0417(1), meetingcopyright(1)	cartoon(6), bill gate(5), gate(4), monopoli(4), pearli gate(4), bill(3), imag(3), caricatur(2), pictur(2), illustr(1), copyright(1), artist(1), own(1), pearli(1)	lift index(5), gener(3), index(3), condit(2), comput(2), comput gener(2), unstabl(2), zone(2), area(1), field(1), between(1), stabl(1), encyclopedia(1), thunderstorm(1), lift(1), free encyclopedia(1), wikipedia(1)

Table 3.5: Examples of gold annotations from the UNT Dataset.

is "cartoon". Those 21/28 images correspond to the 75% *best mode* score obtained by Corpus Frequency baseline.

When we investigated the reason why the word "cartoon" appears so frequently in the UNT corpus, we discovered that 45 of the 300 images were collected from a single online cartoon library. Although Leong et al. (2010) describe some steps that they take to encourage diversity in the sources of images collected for this corpus, apparently those steps were not sufficient.

Additionally, we discovered that the co-occurring text to many of these images contains a long list of keywords, with little full-sentence text that is relevant to the image. We looked at a small sample of the rest of the dataset and found that many of the other text documents in UNT also contain keyword lists. This makes it difficult to measure the benefits of using complex techniques like topic modeling and graph similarity to find and extract keyword annotations. This is shown in the *normal* evaluation results, where the combined system from Leong et al. (2010) is only slightly better at selecting the single best keyword, and no better than tf*idf for the *out-of-ten* measure.

3.5 Conclusion

This chapter presents our work examining evaluation of image annotation systems which use co-occurring natural language text as a source of annotation keywords. We proposed image annotation baselines and tested them on two different datasets.

In Section 3.3.3, we showed that making prior assumptions about which keywords are possible are possible to learn may discount . We especially found good performance from baselines inspired

by those used in document summarization, such as term frequency.

In Section 3.4.3 we found that datasets constructed using search queries are vulnerable to over-representing certain sources or certain styles of text that tend to perform well in search engine results. We are not aware of previous work that has brought this issue to attention, although many datasets which are also used for image captioning, such as SBU-Flickr (Ordonez et al., 2011) are also constructed using search results.

Chapter 4

Nonparametric Image Captioning

This chapter presents our work on image captioning via a caption transfer model. The major contribution of this chapter is a state-of-the-art result for image captioning on the SBU-Flickr dataset, obtained in Section 4.5.

In the previous chapter, we examined the task of annotating images with keywords, by extracting relevant words from natural language text that is related to the query image. We demonstrate that for certain tasks, simple term frequency-based approaches can match or exceed the performance of more complex visual or linguistic models. This is because more complex models often impose constraints and assumptions on the form of the data being modeled. This restricts the effectiveness of these models in situations where the data does not match the form that is specified *a priori*. An example can be seen in Figure 3.3, where restrictions cause the image annotation systems to ignore many relevant terms as potential keywords.

More generally, we are interested in further exploration of nonparametric methods for image description. Nonparametric models are commonly used in statistics to describe and analyze data when the underlying distribution for the data is unknown.

In Computer Vision, the availability of larger datasets has enabled development of nonparametric matching methods which reduce the inference problem for an unknown image to finding an existing labeled image which is semantically similar. Data-driven matching methods have shown to be very effective for a variety of challenging vision problems (Hays and Efros, 2008; Makadia et al., 2008; Tighe and Lazebnik, 2010; Liu et al., 2011).

Visual matching approaches for image captioning (Section 2.3.2) take as input a query image with no associated text, and generate captions by transferring captions from visually similar images retrieved from a large database of captioned images. Unlike the image annotation methods in the previous chapters, these methods do not rely on the availability of additional information besides the query image itself. However, there are some challenges with this approach. Consider the example in Table 4.1. Both images are portraits of human subjects, both facing directly at the camera, and are photographed in black and white against a white background. However, the caption of the retrieved image would not be relevant or descriptive for the query image.

Query Image	Retrieved Image
	 3 month old baby girl with blue eyes in her crib

Table 4.1: Example of a query image from the SBU-Flickr dataset, with the most visually similar captioned image in the database. Visual similarity computed using scene attributes (Patterson et al., 2014).

This example illustrates two main challenges. First, visual similarity measures do not capture all of the relevant visual details which humans might describe in a caption. Second, the text of retrieved caption may be poorly aligned with the visual features used for matching. This second problem is of particular concern because many image captions on the web contain contextual or background information which is not related to the visual content of the image.

In this chapter, we present a nonparametric density estimation technique for estimating the content of the generated caption. This content is modeled using a term frequency distribution which is found by smoothing the term counts over multiple retrieved images. For example, words that appear in multiple captions in Figure 4.2 are likely to be visually relevant to the query image.

The output caption is generated by extracting the caption which best represents the mutually shared content. This task is cast as extractive multi-document summarization, a well-studied problem in Natural Language Processing. The objective of extractive multi-document summarization is to generate a summary of a document collection by extracting sentences with content that is relevant to the entire document collection, which is typically represented using unigram word frequency models (Nenkova and Vanderwende, 2005; Haghighi and Vanderwende, 2009).

Before we apply this approach, we formally define the task (Section 4.1) and describe the dataset to be used (Section 4.2). In Section 4.3 we propose our model, first defining the feature space for visual similarity, then formulating a density estimation problem with the aim of modeling the words which are used to describe the images that are visually similar to the query image. In Section 4.4 we explore methods for selecting which caption to transfer to the query image. Experimental setup and results are presented in Section 4.5, where we show that our model strongly outperforms two state-of-the-art caption extraction systems according to human judgments of caption relevance. Section 4.6 is the conclusion.

	3 month old baby girl with blue eyes in her crib
	A photo from the Ismail's portrait shoot
	A portrait of a man, in black and white
	Portrait in black and white with the red rose
	I apparently had this saved in black and white as well
	Portrait in black and white

Table 4.2: Captions from images retrieved using a k nearest-neighbor approach. Our proposed approach identifies words that appear in multiple captions.

4.1 Formulation

The image caption transfer task is as follows. Given a query image I_q, the goal is to generate a relevant description. The description is generated by selecting a single caption from a database, $\mathcal{C}$. This database consists of a very large number of captioned images, collected from the web.

Previous approaches to this task such as Ordonez et al. (2011) and Patterson et al. (2014) select a match image in $\mathcal{C}$ which is the most visually similar to I_q according to some computed measure. Then simply transfer the caption of the matched image to the query image.

Our approach has an intermediate step of obtaining a probability density estimate of the caption text, given the query image. $p(w)$ is the prior probability for content words for all the captions in $\mathcal{C}$. $p(w|I_q)$ is a word distribution conditioned on the query image.

4.2 SBU-Flickr Dataset

For the experiments in this chapter, we use the SBU-Flickr dataset (Ordonez et al., 2011)[1]. The SBU-Flickr dataset contains one million images from `Flickr.com`, along with the corresponding captions which were uploaded by the users. This dataset was constructed by querying for words from a term list of common visual entities. To encourage visual descriptiveness in the collection, they enforce that all descriptions must contain at least two words belonging to their term list, and at least one prepositional word that indicates spatial relationships (Ordonez et al., 2011).

Due to its size, SBU-Flickr corpus has enabled notable research in both Computer Vision and Natural Language Processing. In particular, it is used in a variety of tasks in work stemming from the 2011 Johns Hopkins-CLSP Summer Workshop, such as image captioning (Mitchell et al., 2012; Kuznetsova et al., 2012), identifying visual descriptions in text (Dodge et al., 2012), and identifying semantically relevant regions of images (Berg et al., 2012).

However, the SBU-Flickr corpus is known to have many misalignments between images and caption content, because Flickr users often use captions to describe background information about the image. Further analysis by Hodosh et al. (2013) shows that the majority of the captions in (∼67%) in SBU-Flickr describe information that cannot be obtained from the image itself, while a substantial fraction (∼23%) contain almost no description of visual content.

4.3 Approach

4.3.1 Measuring Visual Similarity

Many Computer Vision matching methods compute global (scene-based) descriptors rather than object and entity detections. Scene-based techniques in Computer Vision are generally more robust, and can be computed more efficiently on large datasets.

The basic Im2Text model from Ordonez et al. (2011) uses an equally weighted average of GIST Oliva and Torralba (2001) and TinyImage Torralba et al. (2008) features, which coarsely localize low-level features in scenes. The output is a multi-dimensional image space where semantically similar scenes (e.g. streets, beaches, highways) are projected near each other. However, recent work by Patterson and Hays (2012) and Patterson et al. (2014) shows that "scene attribute" representations can provide improved matching for image captioning over the basic Im2Text model. Scene attribute representation are characterized using low-level perceptual attributes as used by GIST (e.g. openness, ruggedness, naturalness), as well as high-level attributes informed by open-ended crowd-sourced image descriptions (e.g., indoor lighting, running water, places for learning).

For our experiments, we use the publicly available[2] scene attributes from Patterson and Hays (2012) to compute representations for all the query images and database images in the SBU-Flickr

[1] `http://tamaraberg.com/CLSP11/`

[2] `https://github.com/genp/sun_attributes`

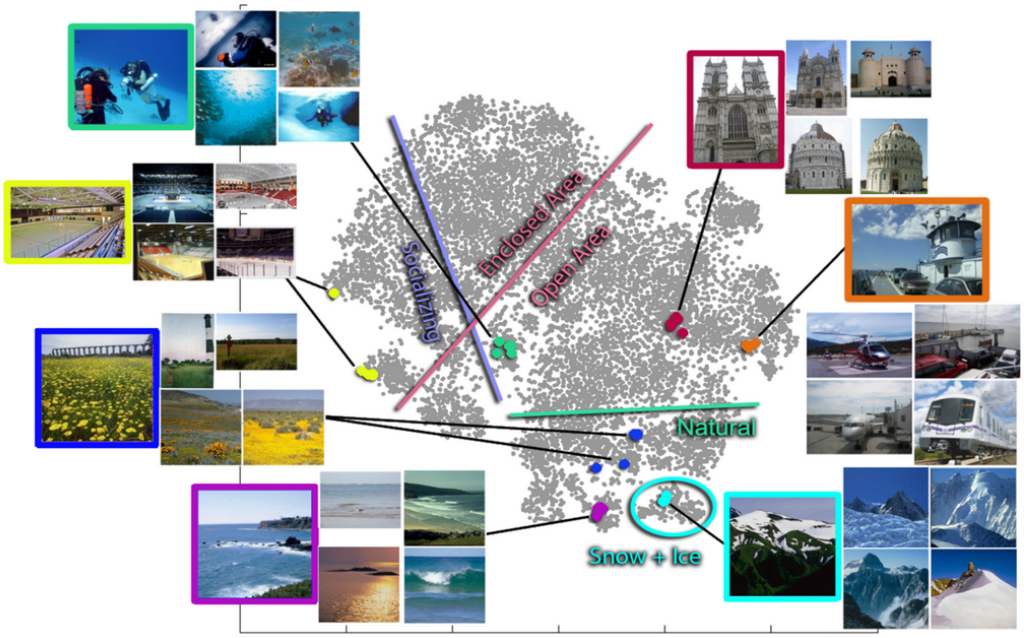


Figure 4.1: A 2d visualization of the multi-dimensional scene attribute space for images. Semantically similar images are projected near to each other. Image provided by Patterson and Hays (2012).

corpus. Images are represented using 102-dimensional real-valued vectors, and similarity between images is measured using the Euclidean distance.

4.3.2 Density Estimation

As shown in Bishop (2006), probability density estimates at a particular point can be obtained by considering points in the training data within some local neighborhood. In our case, we define some region $\mathcal{R}$ in the image space which contains I_q. The probability mass of that space is

$$P = \int_{\mathcal{R}} p(I_q) dI_q \tag{4.1}$$

and if we assume that $\mathcal{R}$ is small enough such that $p(I_q)$ is roughly constant in $\mathcal{R}$, we can approximate

$$p(I_q) \approx \frac{k^{img}}{n^{img} V^{img}} \tag{4.2}$$

where k^{img} is the number of images within $\mathcal{R}$ in the training data, n^{img} is the total number of images in the training data, and V^{img} is the volume of $\mathcal{R}$. In this paper, we fix k^{img} to a constant value, so that V^{img} is determined by the training data around the query image.[3]

[3] Alternately, instead of using k nearest-neighbors, one could use a kernel density approach by fixing the value of V^{img} and determining k^{img} from the number of points in $\mathcal{R}$. This technique is called Parzen-Window Density Estimation, and is useful to ensure that samples far away from the query are not selected in cases where the query is in a sparse area of the data. However, it can lead to over-smoothing in more dense areas. Due to the large number of samples in the SBU-Flickr dataset, the k nearest-neighbor approach is more appropriate.

At this point, we extend the density estimation technique in order to estimate a smoothed model of descriptive text. Let us begin by considering $p(w|I_q)$, the conditional probability of the word[4] w given I_q. This can be described using a Bayesian model:

$$p(w|I_q) = \frac{p(I_q|w)p(w)}{p(I_q)} \tag{4.3}$$

The prior for w is simply its unigram frequency in $\mathcal{C}$, where n_w^{txt} and n^{txt} are word token counts:

$$p(w) = \frac{n_w^{txt}}{n^{txt}} \tag{4.4}$$

Note that n^{txt} is not the same as n^{img} because a single captioned image can have multiple words in its caption. Likewise, the conditional density

$$p(I_q|w) \approx \frac{k_w^{txt}}{n_w^{txt} V^{img}} \tag{4.5}$$

considers instances of observed words within $\mathcal{R}$, although the volume of $\mathcal{R}$ is still defined by the image space. k_w^{txt} is the number of times w is used within $\mathcal{R}$ while n_w^{txt} is the total number of times w is observed in $\mathcal{C}$.

Combining Equations 2, 4, and 5 and canceling out terms gives us the posterior probability:

$$p(w|I_q) = \frac{k_w^{txt}}{k^{img}} \cdot \frac{n^{img}}{n^{txt}} \tag{4.6}$$

If the number of words in each caption is independent of its image's location in the image space, then $p(w|I_q)$ is approximately the observed unigram frequency for the captions inside $\mathcal{R}$.

4.4 Output Caption Selection

We compare two selection methods for extractive caption generation:

1. SumBasic SumBasic (Nenkova and Vanderwende, 2005) is a sentence selection algorithm for extractive multi-document summarization which exclusively maximizes the appearance of words which have high frequency in the original documents. Here, we adapt SumBasic to maximize the average value of $p(w|I_q)$ in a single extracted caption:

$$output = \max_{c^{txt} \in \mathcal{R}} \sum_{w \in c^{txt}} \frac{1}{|c^{txt}|} p(w|I_q) \tag{4.7}$$

The candidate captions c^{txt} do not necessarily have to be observed in $\mathcal{R}$, but in practice we did not find increasing the number of candidate captions to be more effective than increasing the size of $\mathcal{R}$ directly.

2. KL Divergence We also consider a KL Divergence selection method. This method outperforms the SumBasic selection method for extractive multi-document summarization (Haghighi and

[4]Here, we use word to refer to non-function words, and assume all function words have been removed from the captions.

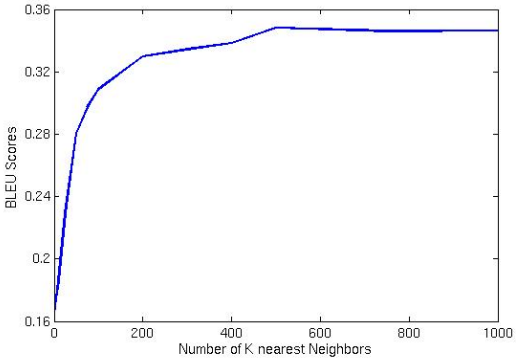


Figure 4.2: BLEU scores vs k images retrieved for our nonparametric model using SumBasic caption selection.

Vanderwende, 2009). It also generates the best extractive captions for Feng and Lapata (2010a), who caption images by extracting text from a related news article. The KL Divergence method is

$$output = \min_{c^{txt} \in \mathcal{R}} \sum_{w} p(w|I_q) \log \frac{p(w|I_q)}{p(w|c^{txt})} \tag{4.8}$$

4.5 Evaluation

4.5.1 Automatic Evaluation

BLEU scores (Papineni et al. (2002), Section 2.4.1) are widely used for image caption evaluation. We compute BLEU scores using the scoring software from NIST, and comparing against the original captions for the query images which were held out during the captioning process.

However, we find BLEU scores to be poor indicators of the quality of our model. As shown in Figure 4.2, the BLEU scores increase as we increase the number of k nearest-neighbor in the model, even as the density estimations seem to get washed out by oversmoothing. BLEU scores continue to improve until $k = 500$ but only because the generated captions become increasingly shorter, and use more general words like "picture" which could be accurate for almost any photo. Furthermore, although we observe that our system captions selected using SumBasic obtain consistently higher BLEU scores, our personal observations find that captions selected using the KLSum method are more relevant, as the SumBasic captions tend to be very short. These findings are consistent with work by Elliott and Keller (2014) which recently shows that BLEU scores tend to reward brevity rather than relevance for image caption evaluation.

Nevertheless, BLEU scores are the accepted metric for recent work, and as shown in Table 4.3, our KLSum captions with $k = 25$ still outperform all other previously published systems and baselines. We have made our full BLEU setup, as well as the captions for all systems and baselines, available

System	BLEU@1
Scene Attributes	.1640
Collective	.1654
System (SumBasic)	.2294
System (KLSum)	.1886

Table 4.3: BLEU scores for our system, Scene Attributes nearest-neighbor baseline (Patterson et al., 2014), and Collective caption generation (Kuznetsova et al., 2012). See Section 4.5.1.

in the ACL Anthology[5], in order to allow our work to be evaluated using future automatic metrics.

4.5.2 Human Evaluation

We generate captions using our system with KL Divergence sentence selection and $k = 25$. We also evaluate the original Human captions for the query image, as well as generated captions from two recently published caption transfer systems. First, we consider the Scene Attributes system (Patterson et al., 2014), which represents both the best scene-based transfer model and a $k = 1$ nearest-neighbor baseline for our system. We also compare against the Collective system (Kuznetsova et al., 2012), which is the best object-based transfer model.

We perform our human evaluation of caption relevance using a similar setup to that of Kuznetsova et al. (2012), who have humans rate the image captions on a 1-5 scale (5: perfect, 4: almost perfect, 3: 70-80% good, 2: 50-70% good, 1: totally bad). Evaluation is performed using Amazon Mechanical Turk. Evaluators are shown both the caption and the query image. Evaluators are specifically instructed to ignore errors in grammaticality and coherence, so to not unfairly advantage systems such as ours and Scene Attributes which select entire human-written captions, against systems such as Collective which construct new captions out of transferred phrases. An example of the human evaluation task is shown in Figure 4.3.

In order to facilitate comparison, we use the same test/train split that is used in the publicly available system output for the Collective system[6]. However, we remove some query images which have contamination between the train and test set (this occurs when a photographer takes multiple shots of the same scene and gives all the images the exact same caption). We also note that their test set is selected based on images where their object detection systems had good performance, and may not be indicative of their performance on other query images.

Table 4.4 shows the results of our human study. Captions generated by our system have 48% improvement in relevance over the Scene Attributes system captions, and 34% improvement over the Collective system captions. Although our system captions score lower than the human captions on average, there are some instances of our system captions being judged as more relevant than the human-written captions.

[5] http://www.aclweb.org/anthology/attachments/P/P14/P14-2097.Datasets.zip

[6] http://www.cs.sunysb.edu/~pkuznetsova/generation/cogn/captions.html

System	Relevance
Collective	2.38 ($\sigma = 1.45$)
Scene Attributes	2.15 ($\sigma = 1.45$)
System	3.19 ($\sigma = 1.50$)
Human	4.09 ($\sigma = 1.14$)

Table 4.4: Human evaluations of relevance: mean ratings and standard deviations. See Section 4.5.2.

4.6 Discussion and Examples

Example captions are shown in Table 4.5. In many instances, scene-based image descriptors provide enough information to generate a complete description of the image, or at least a sufficiently good one. However, there are some kinds of images for which scene-based features alone are insufficient. For example, the last example describes the small pink flowers in the background, but misses the bear.

Image captioning is a relatively novel task for which the most compelling applications are probably not yet known. Much previous work in image captioning focuses on generating captions that concretely describe detected objects and entities (Kulkarni et al., 2011; Yang et al., 2011; Mitchell et al., 2012; Yu and Siskind, 2013). However, human-generated captions and annotations also describe perceptual features, contextual information, and other types of content. Additionally, our system is robust to instances where entity detection systems fail to perform. However, one could consider combined approaches which incorporate more regional content structures. For example, previous work in nonparametric hierarchical topic modeling (Blei et al., 2010) and scene labeling (Liu et al., 2011) may provide avenues for further improvement of this model. We leave these ideas for future work.

Please rate the image caption.

Good image captions:

- Describe what is important in the image
- Might give background information about the image, like names of people or locations
- Do not have to be long (a few words is fine)
- Are allowed to have spelling or grammar mistakes

Bad image captions:

- Describe things that are not in the image
- Describe a scene that doesn't match the image

5 (perfect): Good description for the image. All details in the caption describe something in the image. Background information such as names of people or places seem plausible.

4 (mostly relevant): One or two moderate mistakes in the caption. For example, mistaking a cat for a dog, or a man for a woman.

3 (somewhat relevant): About half of the details described in the caption are wrong. For example, the caption describes the setting of the image correctly, but the main objects in the image are completely wrong.

2 (mostly irrelevant): Only one or two words in the caption are related to the image at all.

1 (completely irrelevant): The image and the caption describe completely different scenes. Background information is clearly not at all related to this image.

pussy cat sitting in the sun

5 (perfect) 4 (mostly relevant) 3 (about half relevant) 2 (mostly irrelevant) 1 (completely irrelevant)

Optional: If you have additional comments, observations, or feedback, please share them here.)

Submit

Figure 4.3: Human evaluation task.

				
COLLECTIVE:	One of the birds seen in company of female and juvenile.	View of this woman sitting on the sidewalk in Mumbai by the stained glass. The boy walking by next to matching color walls in gov t building.	Found this mother bird feeding her babies in our maple tree on the phone.	Found in floating grass spotted alongside the scenic North Cascades Hwy near Ruby arm a black bear.
SCENE ATTRIBUTES:	This small bird is pretty much only found in the ancient Caledonian pine forests of the Scottish Highlands.	me and allison in front of the white house	The sand in this beach was black...I repeat BLACK SAND	Not the green one, but the almost ghost-like white one in front of it.
SYSTEM:	White bird found in park standing on brick wall	by the white house	pine tree covered in ice :)	Pink flower in garden w/ moth
HUMAN:	Some black head bird taken in bray head.	Us girls in front of the white house	Male cardinal in snowy tree knots	Black bear by the road between Ucluelet and Port Alberni, B.C., Canada

Table 4.5: Example query images and generated captions.

Chapter 5

Domain-Specific Image Captioning

Consider the following example:

Query Image	Retrieved Image
	 This sporty sneaker clog keeps foot cool and comfortable and fully supported.

Table 5.1: Example of a query image and nearest-neighbor match from the Attribute Discovery dataset. Visual similarity computed using GIST (Oliva and Torralba, 2001).

The image on the left is the query image, while the image on the right is its nearest-neighbor in a database, using GIST (Oliva and Torralba, 2001) nearest neighbors [1]. In this example, the caption of the retrieved image is somewhat accurate for the query image – both images show clog-style shoes, both shoes appear to be comfortable. However, some words in the retrieved caption such as "sporty" and "sneaker" do not accurately describe the query image.

In the previous chapter, we presented an approach for image captioning using multi-document summarization to select an existing caption which best fits the query image. However, there are some limitations to this approach. The output would be limited to existing captions in the database. Additionally, it is difficult to determine how many nearest-neighbor matches the nonparametric density estimator should smooth over, and to find the best balance between brevity and detail. This is particularly true for the task of *domain-specific image captioning*, where fine-grained details

[1] We do not use the scene attributes from Patterson and Hays (2012) because the images are not of natural scenes.

become more relevant. For example, if the query image above were to appear on an online shopping website, simply captioning it as "A shoe" would be completely inadequate.

Domain-specific image captioning also presents additional challenges. As we mentioned in previous chapters, many current approaches for image captioning rely on the use of general-domain entity detectors. These detectors typically require accurate hand-labeled training data, which are not available for many specific domains where automatic image captioning would be useful. Ideally, a domain-specific captioning system would learn in a less supervised fashion, using captioned images found on the web. Less supervised captioning methods could be used to generate detailed and accurate descriptions for a variety of long-tail domains of captioned image data, such as in nature and medicine.

In this chapter, we present a data-driven framework for domain-specific image caption generation, which adapts existing captions in the manner shown. Our framework has three main components. We *extract* an existing description from a database of human-captions, by projecting query images into a multi-dimensional space where structurally similar images are near each other. We also train a *joint topic model* to discover the latent topics which generate both captions and images. We combine these two approaches using *sentence compression* to delete modifying details in the extracted caption which are not relevant to the query image. An example is shown in Table 5.

	Query Image	Retrieved Image
Extract		 This sporty sneaker clog keeps foot cool and comfortable and fully supported.
Topic Model		This sporty sneaker clog keeps foot cool and comfortable and fully supported.
Compress	This clog keeps foot comfortable and supported.	

Our domain-specific captioning framework is inspired by several recent approaches at the intersection of Natural Language Processing and Computer Vision, including our own work described in earlier chapters of this document. These include:

Caption Transfer Our method extends previous work such as Farhadi et al. (2010), Ordonez et al. (2011), and the nonparametric method presented in Chapter 4 of this thesis.

Image Annotation While recent improvements in state-of-the-art visual object class detections

(Felzenszwalb et al., 2010) have enabled much recent work in image caption generation (Farhadi et al., 2010; Ordonez et al., 2011; Kulkarni et al., 2011; Yang et al., 2011; Mitchell et al., 2012; Yu and Siskind, 2013), trained detectors are often not available for domain-specific entities. Instead, we develop a joint topic model to learn the latent topics that generate both images and captions. Previous work by Berg et al. (2010) and Feng and Lapata (2010b) has also explored using natural language caption text as a source of image annotations.

Sentence Compression Typical models for sentence compression (Knight and Marcu, 2002; Furui et al., 2004; Turner and Charniak, 2005; Clarke and Lapata, 2008) have a summarization objective: reduce the length of a source sentence without changing its meaning. In contrast, our objective is to change the meaning of the source sentence, letting its overall correctness relative to the query image determine the length of the output. Our work can also be constrasted to that of Kuznetsova et al. (2013), who compress image caption sentences with the objective of creating a corpus of generally transferrable image captions, while our objective is to compress captions to generate a more accurate caption for a specific query image.

In Section 5.1 we describe the Attribute Discovery Dataset (Berg et al., 2010), which will be used in experiments for the rest of this chapter. Section 5.2 describes the approach used for transferring captions, and Section 5.3 describes the topic model. Section 5.4 describes how transferred captions are adapted for a specific query image using sentence compression. Section 5.5 describes automatic and human evaluations that we use to show that our captioning method effectively deletes inaccurate words from extracted captions while maintaining a high level of detail in the generated output.

5.1 Dataset

The dataset we use is the women's shoes section of the publicly available Attribute Discovery Dataset[2] from Berg et al. (2010), which consists of product images and captions scraped from the shopping website `Like.com`. We use the women's shoes section of the dataset which has 14764 captioned images. Product descriptions describe many different attributes such as styles, colors, fabrics, patterns, decorations, and affordances (activities that can be performed while wearing the shoe). Some captions also include non-visual information such as sizing or whether the item is on sale. Some examples are shown in Table 5.2.

For our experiments, we first determine an 80/20% train test split. We define a textual vocabulary of "descriptive words", which are non-function words – adjectives, adverbs, nouns (except proper nouns), and verbs. This gives us a total of 9578 descriptive words in the training set, with an average of 16.33 descriptive words per caption.

[2] `http://tamaraberg.com/attributesDataset/index.html`

	Two adjustable buckle straps top a classic rubber rain boot grounded by a thick lug sole for excellent wet-weather traction.
	Available in Plus Size. Faux snake skin flats with a large crossover buckle at the toe. Padded insole for a comfortable all day fit.
	Glitter-covered elastic upper in a two-piece dress sandal style with round open toe. Single vamp strap with contrasting trim matching elasticized heel strap crisscrosses at instep.
	Explosive! These white leather joggers are sure to make a big impression. Details count, including a toe overlay, millennium trim and lightweight raised sole.

Table 5.2: Example data from the Attribute Discovery Dataset (Berg et al., 2010). See Section 5.1.

5.2 Caption Transfer

Our overall process is to first find a caption sentence from our database to use as a template, and then adapt the template sentences using sentence compression. We compress by removing details that are probably not correct for the5sec:eval query image. For example, if the sentence describes "a red slipper" but the shoe in the query image is yellow, we want to remove "red" and keep the rest.

As in this simple example, the basic paradigm for compression is to keep the head words of phrases ("slipper") and remove modifiers. Thus we want the extraction stage of our scheme to be more likely to find a candidate sentence with correct head words, figuring that the compression stage can edit the mistakes. Our hypothesis is that headwords tend to describe more spatially structured visual concepts, while modifier words describe those that are more easily represented using local or unstructured features.[3] Table 5.3 contains additional example captions with parses.

GIST (Oliva and Torralba, 2001) is a commonly used feature in Computer Vision which coarsely localizes perceptual attributes (e.g. rough vs smooth, natural vs manmade). By computing the GIST of the images, we project them into a multi-dimensional Euclidean space where images with semantically similar structures are located near each other. Thus the extraction stage of our caption generation process selects a sentence from the GIST nearest-neighbor to the query image.

5.3 Topic Model

The second component of our framework incorporates visual and textual features using a less structured model. We use a multi-modal topic model to learn the latent topics which generate bag-of-words features for an image and its caption.

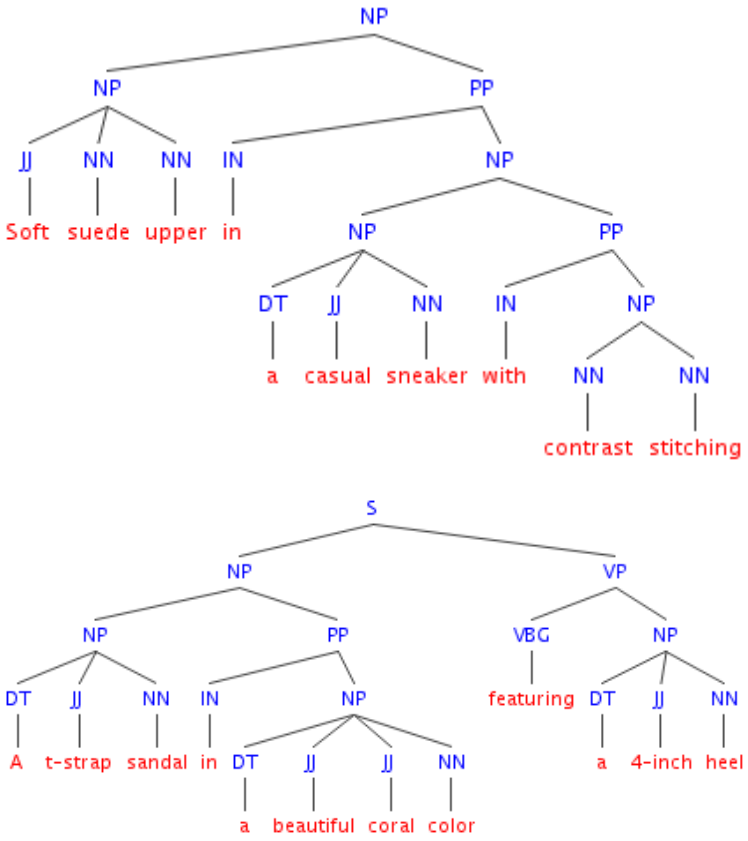
The bag-of-words model for Computer Vision represents images as a mixture of topics. Measures of shape, color, texture, and intensity are computed at various points on the image and clustered into discrete "codewords" using the k-means algorithm. Unlike text words, an individual codeword has little meaning on its own, but distributions of codewords can provide a meaningful, though unstructured, representation of an image.

An image and its caption do not express exactly the same information, but they are topically related. We employ the Polylingual Topic Model (Mimno et al., 2009), which is originally used to model corresponding documents in different languages that are topically comparable, but not parallel translations. A plate diagram for the polylingual topic model is shown in Figure 5.1. We extend this work to model shopping images and captions.

The generative process is defined for a captioned image: the pair $w =< w^{img}, w^{txt} >$. It starts with a single topic distribution drawn from concentration parameter α and base measure m:

$$\theta \sim Dir(\theta, \alpha m) \tag{5.1}$$

[3] For example, the color "red" can be described using a bag of random pixels, while a "slipper" is a spatial configuration of parts in relationship to each other.



5sec:topicmodel

Table 5.3: Example parses of women's shoes descriptions. Our hypothesis is that the headwords in phrases are more likely to describe visual concepts which rely on spatial locations or relationships, while modifiers words can be represented using less-structured visual bag-of-words features.

Modality-specific latent topic assignments z^{img} and z^{txt} are drawn for each of the text words and codewords:

$$\mathbf{z}^{img} \sim P(\mathbf{z}^{img}|\theta) = \prod_n \theta_{z_n^{img}} \tag{5.2}$$

$$\mathbf{z}^{txt} \sim P(\mathbf{z}^{txt}|\theta) = \prod_n \theta_{z_n^{txt}} \tag{5.3}$$

Observed words are generated according to their probabilities in the modality-specific topics:

$$\mathbf{w}^{img} \sim P(\mathbf{w}^{img}|\mathbf{z}^{img}, \Phi^{img}) = \phi^{img}_{w_n^{img}|z_n^{img}} \tag{5.4}$$

$$\mathbf{w}^{txt} \sim P(\mathbf{w}^{txt}|\mathbf{z}^{txt}, \Phi^{txt}) = \phi^{txt}_{w_n^{txt}|z_n^{txt}} \tag{5.5}$$

Given the uncaptioned query image q^{img} and the trained multi-modal topic model, it is now possible to infer the shared topic proportion for q^{img} using Gibbs sampling:

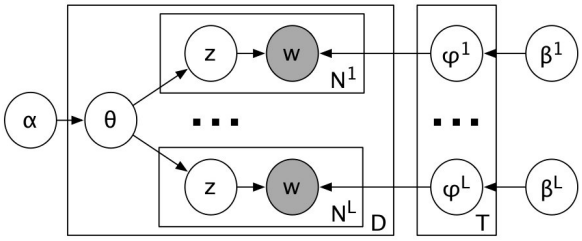


Figure 5.1: Polylingual topic model (Mimno et al., 2009)

$$P(z_n = t|q^{img}, z_{\backslash n}, \Phi^{img}, \alpha m) \propto \phi^{img}_{q_n^{img}|t} \frac{(N_t)_{\backslash n} + \alpha m_t}{\sum_t N_t - 1 + \alpha} \tag{5.6}$$

5.4 Compression

Let $\mathbf{w} = w_1, w_2, ..., w_n$ be the words in the extracted caption for q^{img}. For each word, we define a binary decision variable δ, such that $\delta_i = 1$ if w_i is included in the output compression, and $\delta_i = 0$ otherwise. Our objective is to find values of δ which generate a caption for q^{img} which is both semantically and grammatically correct.

We cast this problem as an Integer Linear Program (ILP), which has previously been used for the standard sentence compression task (Clarke and Lapata, 2008; Martins and Smith, 2009). ILP is a mathematical optimization method for determining the optimal values of integer variables in order to maximize an objective given a set of constraints. Sentence compression is modeled as an integer optimization problem because the decision to include each word is a binary integer decision.

5.4.1 Compression Objective

The objective for the sentence compression is to maximize a weighted linear combination of two measures which represent the correctness and fluency of the output compression. For correctness, recall in Section 5.1 we defined words as either descriptive words or function words. For each descriptive word, we estimate the probability of the word given the query image, $P(w_i|q^{img})$, using topic proportions estimated using Equation 5.6:

$$P(w_i|q^{img}) = \sum_t P(w_i|z_t^{txt}) P(z_t|q^{img}) \tag{5.7}$$

This estimate is used to find $I(w_i)$, which is a function of the likelihood of each word in the extracted caption. This function considers the prior probability of w_i because frequent words often have a high posterior probability even when they are inaccurate. Function words (such as articles and prepositions) get a score of zero, because they do not contribute to the accuracy of a generated caption.

$$I(w_i) = \begin{cases} P(w_i|q^{img}) - P(w_i), & \text{if descriptive} \\ 0, & \text{function word} \end{cases} \tag{5.8}$$

Thus the sum $\sum_{i=1}^{n} \delta_i \cdot I(w_i)$ is the overall measure of the correctness of a proposed caption conditioned on q^{img}.

Next, we measure the fluency of the output caption using a trigram language model. This requires additional binary decision variables. These binary decision variables are α_i which equals 1 if w_i begins the output compression, β_{ij} which equals 1 if the bigram sequence w_i, w_j ends the compression, and γ_{ijk} which equals 1 if the trigram sequence w_i, w_j, w_k is in the compression. We also add a special variable $\delta_0 = 1$, to represent the "start token" for the output compression.

This language model favors shorter sentences, which is not necessarily the objective for image captioning, so we introduce a weighting factor, λ, to lessen the effect.

Here is the combined objective, using P to represent $\log P$:

$$\begin{aligned} \max z \;=\; & \Bigg(\sum_{i=1}^{n} \alpha_i \cdot P(w_i|\text{start}) \\ & + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} \gamma_{ijk} \cdot P(w_k|w_i, w_j) \\ & + \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} \beta_{ij} \cdot P(\text{end}|w_i, w_j) \Bigg) \cdot \lambda \\ & + \sum_{i=1}^{n} \delta_i \cdot I(w_i) \end{aligned} \tag{5.9}$$

5.4.2 Compression Constraints

The compression constraints for the ILP ensure the mathematical validity of the model, as well as the grammatical correctness of its output.

Sequential Constraints As defined in Clarke and Lapata (2008), these constraints ensure that the ordering of the trigrams is valid, and that the mathematical validity of the model holds. These constraints are:

1. Only one word can be the first word in the output compression.

$$\sum_{i} \alpha_i = 1 \tag{5.10}$$

2. If a word is included in the compression, it is either the first word in the compression, or it follows another word in the compression.[4]

[4]The second word in the compression is the last word of the trigram starting with the special start token.

$$\delta_k - \alpha_k - \sum_{i=0}^{k-2} \sum_{j=1}^{k-1} \gamma_{ijk} = 0$$

$$\forall k : k \in 1...n \tag{5.11}$$

3. If a word appears in the compression, it either is followed by another word, or it ends the compression.

$$\sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} \gamma_{ijk} - \sum_{j=i+1}^{n} \beta_{ij} - \sum_{h=0}^{i-1} \beta_{hi} - \delta_i = 0$$

$$\forall i : i \in 1...n \tag{5.12}$$

$$\tag{5.13}$$

4. Only one bigram can end the compression.

$$\sum_{i=0}^{n-1} \sum_{j=i+1}^{n} \beta_{ij} = 1 \tag{5.14}$$

Modifier Constraints

Modifier constraints ensure that the sentence is coherent. Using the "semantic head" variation of the headfinder from Collins (1999), these constraints are:

1. The head word of the sentence and the head words of noun phrases must be included.
2. If $headof(w_i) = w_j$, then $\delta_i \leq \delta_j$ for words that are not punctuation or coordinating conjunctions.

Other Constraints

Finally, there are other constraints that ensure a minimum length for the compressed output ($\sum_i \delta_i \geq 3$) and define valid use of punctuation and coordinating conjunctions.

5.5 Evaluation

We evaluate using both automatic metrics and a human study. Automatic metrics provide a simple and objective method to evaluate nearly 3000 captions generated from the images in our test set, and allow us to explore the trade-off between recall and precision in our sentence compression model. A human study more accurately measures the quality of generated captions, since automatic metrics fail to capture variance in human descriptions. Humans are also better at measuring the grammaticality of a compressed caption.

ROUGE-2	Average	95% Confidence int.
KL (Extraction)		
P	.06114	(.05690 - .06554)
R	.02499	(.02325 - .02686)
F	.03360	(.03133 - .03600)
GIST (Extraction)		
P	.10894	(.09934 - .11921)
R	.05474	(.04926 - .06045)
F	.06863	(.06207 - .07534)
LM-Only (Compression)		
P	.13782	(.12602 - .14864)
R	.02437	(.02193 - .02700)
F	.03864	(.03512 - .04229)
System (Compression)		
P	**.16752**	(.15679 -.17882)
R	.05060	(.04675 - .05524)
F	.07204	(.06685 - .07802)

Table 5.4: ROUGE-2 (bigram) scores. The precision of our system compression (bolded) significantly improves over the caption that it compresses (GIST), without a significant decrease in recall.

5.5.1 Setup

We compare the following systems and baselines:

KL (Extraction): The top performing extractive model from Feng and Lapata (2010a), and the second-best captioning model overall. Using estimated topic distributions from our joint model, we extract the source with minimum KL Divergence from q^{img}.

GIST (Extraction): The sentence extracted using GIST nearest-neighbors, and the uncompressed source for the compression systems.

LM-Only (Compression): This baseline changes the function $I(w)$ (Equation 5.8) to simply give the prior $P(w)$ in the case a word is descriptive. This causes the ILP to ignore the content objective and only maximize the trigram language model (still subject to the constraints). We include this baseline to demonstrate that our model is effectively conditioning output compressions on the query image, as opposed to generating a more generally transferrable caption as does Kuznetsova et al. (2013).

System (Compression): Our full system.

Unfortunately, we cannot compare our system against prior work in general-domain image captioning, because those models use visual detection systems which train on labeled data that is not available in our domain.

	BLEU@1
KL (Extraction)	.2098
GIST (Extraction)	.4259
LM-Only (Compression)	.4780
System (Compression)	**.4841**

Table 5.5: BLEU@1 scores of generated captions against human authored captions. Our model (bolded) has the highest BLEU@1 score with significance.

	System		LM-Only	
	Yes	No	Yes	No
Compression improves accuracy	63.2%	36.8%	42.6%	57.4%
Compression is grammatical	73.1%	26.9%	82.2%	17.8%

Table 5.6: Human evaluation results.

5.5.2 Automatic Evaluation

We perform automatic evaluation using similarity measures between automatically generated and human-authored captions. Note that currently our system and baselines only generate single-sentence captions, but we compare against entire held-out captions in order to increase the amount of text we have to compare against.

ROUGE (Lin, 2004) is a summarization evaluation metric which has also been used to evaluate image captions (Yang et al., 2011). It is usually a recall-oriented measure, but we also report precision and f-measure because our sentence compressions do not improve recall. Table 5.4 shows ROUGE-2 (bigram) scores computed without stopwords.

We observe that our system very significantly improves ROUGE-2 precision of the GIST extracted caption, without significantly reducing recall. While LM-Only also improves precision against GIST extraction, it indiscriminately removes some words which are relevant to the query image. We also observe that GIST extraction strongly outperforms the KL model, which demonstrates the importance of visual structure.

We also report BLEU (Papineni et al., 2002) scores in Table 5.5, which are the most popularly accepted automatic metric for captioning evaluation (Farhadi et al., 2010; Kulkarni et al., 2011; Ordonez et al., 2011; Kuznetsova et al., 2012, 2013). Results are very similar to the ROUGE-2 precision scores, except the difference between our system and LM-Only is less pronounced because BLEU counts function words, while ROUGE does not.

5.5.3 Human Evaluation

We perform human evaluation of compressions generated by our system and LM-Only.[5] Users are shown the query image, the original uncompressed caption, and a compressed caption, and are asked two questions: does the compression improve the accuracy of the caption, and is the compression grammatical.

We collect 553 judgments from six women who are native English-speakers and knowledgeable about fashion. Users were recruited via email and did the study over the internet.

Table 5.6 reports the results of the human evaluation. Users report 63.2% of SYSTEM compressions improve accuracy over the original, while the other 36.8% did not improve accuracy. (Keep in mind that a bad compression does not make the caption less accurate, just less descriptive.) LM-ONLY improves accuracy for less than half of the captions, which is significantly worse than SYSTEM captions (Fisher exact test, two-tailed p less than 0.01).

Users find LM-Only compressions to be slightly more grammatical than System compressions, but the difference is not significant. ($p > 0.05$)

[5] About 15% of output compressions are the same for both systems, and about 10% have no deleted words in the output compression. We include the former in the human evaluation, but not the latter.

Query Image	Nearest Neighbor
	
Extraction: Shimmering snake-embossed leather upper in a slingback evening dress sandal style with a round open toe. **Compression:** Shimmering upper in a slingback evening dress sandal style with a round open toe.	
Query Image	Nearest Neighbor
	
Extraction: This sporty sneaker clog keeps foot cool and comfortable and fully supported. **Compression:** This clog keeps foot comfortable and supported.	
Query Image	Nearest Neighbor
	
Extraction: Italian patent leather peep-toe ballet flat with a signature tailored grosgrain bow. **Compression:** leather ballet flat with a signature tailored grosgrain bow.	
Query Image	Nearest Neighbor
	
Extraction: Platform high heel open toe pump with horsebit available in silver guccissima leather with nickel hardware with leather sole. **Compression:** Platform high heel open toe pump with horsebit available in leather with nickel hardware with leather sole.	

Table 5.7: Example output from our full system. Red underlined words indicate the words which are deleted by our compression model.

Query Image	Nearest Neighbor
	
Extraction: Classic ballet flats with decorative canvas strap and patent leather covered buckle. **Compression:** Classic ballet flats covered.	
Query Image	Nearest Neighbor
	
Extraction: This shoe is the perfect shoe for you , featuring an open toe and a lace up upper with a high heel , and a two tone color . **Compression:** This shoe is the shoe , featuring an open toe and upper with a high heel .	

Table 5.8: Examples of bad performance. The top example is a parse error, while the bottom example deletes modifiers that are not part of the image description.

Chapter 6

Conclusion

Bibliography

Aker, A. and Gaizauskas, R. (2010). Generating image descriptions using dependency relational patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1250–1258, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barzilay, R. and Lee, L. (2003). Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 16–23. Association for Computational Linguistics.

Barzilay, R. and McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Bateman, J. A. (1997). Enabling technology for multilingual natural language generation: the kpml development environment. *Natural Language Engineering*, 3(01):15–55.

Berg, A. C., Berg, T. L., Daume, H., Dodge, J., Goyal, A., Han, X., Mensch, A., Mitchell, M., Sood, A., Stratos, K., et al. (2012). Understanding and predicting importance in images. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3562–3569. IEEE.

Berg, T. L., Berg, A. C., and Shih, J. (2010). Automatic attribute discovery and characterization from noisy web data. In *Proceedings of the 11th European conference on Computer vision: Part I*, ECCV'10, pages 663–676, Berlin, Heidelberg. Springer-Verlag.

Bishop, C. M. (2006). *Pattern recognition and machine learning*, volume 1. Springer New York.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30.

Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Boiy, E., Deschacht, K., and Moens, M. (2008). Learning visual entities and their visual attributes from text corpora. In *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*, pages 48–53. IEEE.

Cheung, J. C. K. and Penn, G. (2013). Towards robust abstractive multi-document summarization: A caseframe analysis of centrality and domain.

Clarke, J. and Lapata, M. (2008). Global inference for sentence compression an integer linear programming approach. *J. Artif. Int. Res.*, 31(1):399–429.

Collins, M. J. (1999). *Head-driven statistical models for natural language parsing*. PhD thesis, Philadelphia, PA, USA.

Coughlin, D. (2003). Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*, pages 63–70.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886 –893 vol. 1.

Demir, S., Carberry, S., and McCoy, K. F. (2012). Summarizing information graphics textually. *Computational Linguistics*, 38(3):527–574.

Deschacht, K., Moens, M.-F., et al. (2007). Text analysis for automatic image annotation. In *ACL*, volume 7, pages 1000–1007.

Dodge, J., Goyal, A., Han, X., Mensch, A., Mitchell, M., Stratos, K., Yamaguchi, K., Choi, Y., Daumé III, H., Berg, A. C., et al. (2012). Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 762–772. Association for Computational Linguistics.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2014). Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*.

Elliott, D. and Keller, F. (2014). Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 452–457.

Elsner, M. and Santhanam, D. (2011). Learning to fuse disparate sentences. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 54–63. Association for Computational Linguistics.

Everingham, M., Van Gool, L., Williams, C., Winn, J., and Zisserman, A. (2008). The pascal visual object classes challenge 2008 (voc2008) results. `http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html`.

Fan, X., Aker, A., Tomko, M., Smart, P., Sanderson, M., and Gaizauskas, R. (2010). Automatic image captioning from the web for gps photographs. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 445–448, New York, NY, USA. ACM.

Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., et al. (2014). From captions to visual concepts and back. *arXiv preprint arXiv:1411.4952*.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: generating sentences from images. In *Proceedings of the 11th European conference on Computer vision: Part IV*, ECCV'10, pages 15–29, Berlin, Heidelberg. Springer-Verlag.

Felzenszwalb, P. F., Girshick, R. B., and McAllester, D. (2008). Discriminatively trained deformable part models, release 4. `http://people.cs.uchicago.edu/~pff/latent-release4/`.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Feng, Y. and Lapata, M. (2008). Automatic image annotation using auxiliary text information. In *ACL*, pages 272–280.

Feng, Y. and Lapata, M. (2010a). How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1239–1249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Feng, Y. and Lapata, M. (2010b). Topic models for image annotation and text illustration. In *HLT-NAACL*, pages 831–839.

Frankel, C., Swain, M. J., and Athitsos, V. (1996). Webseer: An image search engine for the world wide web.

Furui, S., Kikuchi, T., Shinnaka, Y., and Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE TRANS. ON SPEECH AND AUDIO PROCESSING*, 12(4):401–408.

Gatt, A. and Reiter, E. (2009). Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93. Association for Computational Linguistics.

Grauman, K. and Leibe, B. (2010). *Visual object recognition*. Morgan & Claypool Publishers.

Greenbacker, C. F., Carberry, S., and McCoy, K. F. (2011). A corpus of human-written summaries of line graphs. In *Proceedings of the UCNLG+ Eval: Language Generation and Evaluation Workshop*, pages 23–27. Association for Computational Linguistics.

Haghighi, A. and Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics.

Hanbury, A. (2008). A survey of methods for image annotation. *J. Vis. Lang. Comput.*, 19:617–627.

Hays, J. and Efros, A. A. (2008). Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: data, models and evaluation metrics. *JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH*, 47:853–899.

Hovy, E. H. (1991). *Approaches to the planning of coherent text*. Springer.

Jurafsky, D. and James, H. (2009). Speech and language processing an introduction to natural language processing, computational linguistics, and speech.

Karpathy, A. and Fei-Fei, L. (2014). Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artif. Intell.*, 139(1):91–107.

Koller, A. and Stone, M. (2007). Sentence generation as a planning problem.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. In *CVPR*, pages 1601–1608.

Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., and Choi, Y. (2013). Generalizing image captions for image-text parallel corpus. In *ACL*.

Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., and Choi, Y. (2012). Collective generation of natural image descriptions. In *ACL*.

Leong, C. W., Mihalcea, R., and Hassan, S. (2010). Text mining for automatic image tagging. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 647–655. Association for Computational Linguistics.

Leung, T. and Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., and Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 220–228, Stroudsburg, PA, USA. Association for Computational Linguistics.

Li, W. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 577–584, New York, NY, USA. ACM.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In Marie-Francine Moens, S. S., editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, C., Yuen, J., and Torralba, A. (2011). Nonparametric scene parsing via label transfer. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(12):2368–2382.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150 –1157 vol.2.

Makadia, A., Pavlovic, V., and Kumar, S. (2008). A new baseline for image annotation. In *Computer Vision–ECCV 2008*, pages 316–329. Springer.

Mao, J., Xu, W., Yang, Y., Wang, J., and Yuille, A. L. (2014). Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

Martins, A. F. T. and Smith, N. A. (2009). Summarization with a joint model for sentence extraction and compression. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, ILP '09, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mason, R. (2013). Domain-independent captioning of domain-specific images. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 69–76, Atlanta, Georgia. Association for Computational Linguistics.

Mason, R. and Charniak, E. (2012). Apples to oranges: Evaluating image annotations from natural language processing systems. In *NAACL-2012: Main Proceedings*, Montreal, Canada. Association for Computational Linguistics.

Mason, R. and Charniak, E. (2013). Annotation of online shopping images without labeled training examples. In *Proceedings of Workshop on Vision and Language*, Atlanta, Georgia. Association for Computational Linguistics.

Mason, R. and Charniak, E. (2014a). Domain-specific image captioning. In *CoNLL-2014*, page 11.

Mason, R. and Charniak, E. (2014b). Nonparametric method for data-driven image captioning. In *ACL-2014: Main Proceedings*, Baltimore, Maryland. Association for Computational Linguistics.

McCallum, A. K. (2002). {MALLET: A Machine Learning for Language Toolkit}.

McCarthy, D. and Navigli, R. (2007). Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 880–889, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A. C., Berg, T. L., and Daumé III, H. (2012). Midge: Generating image descriptions from computer vision detections. In *European Chapter of the Association for Computational Linguistics (EACL)*.

Mittal, V. O., Roth, S., Moore, J. D., Mattis, J., and Carenini, G. (1995). Generating explanatory captions for information graphics. In *IJCAI*, pages 1276–1283.

Montes-y Gómez, M., López-López, A., and Gelbukh, A. (2000). Information retrieval with conceptual graph matching. In *Database and Expert Systems Applications*, pages 312–321. Springer.

Moran, S. and Lavrenko, V. (2011). Optimal tag sets for automatic image.

Murray, G., Carenini, G., and Ng, R. (2010). Interpretation and transformation for abstracting conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 894–902. Association for Computational Linguistics.

Nenkova, A. and Vanderwende, L. (2005). The impact of frequency on summarization.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175.

Ordonez, V., Kulkarni, G., and Berg, T. (2011). Im2text: Describing images using 1 million captioned photographs. In *NIPS*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Patterson, G. and Hays, J. (2012). Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE.

Patterson, G., Xu, C., Su, H., and Hays, J. (2014). The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*.

Prevost, S. and Steedman, M. (1993). Generating contextually appropriate intonation. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, pages 332–340. Association for Computational Linguistics.

Reiter, E., Dale, R., and Feng, Z. (2000). *Building natural language generation systems*, volume 33. MIT Press.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer.

Tighe, J. and Lazebnik, S. (2010). Superparsing: scalable nonparametric image parsing with superpixels. In *Computer Vision–ECCV 2010*, pages 352–365. Springer.

Torralba, A., Fergus, R., and Freeman, W. T. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11):1958–1970.

Tsikrika, T., Popescu, A., and Kludas, J. (2011). Overview of the wikipedia image retrieval task at imageclef 2011. In *CLEF (Notebook Papers/Labs/Workshop)*.

Turner, J. and Charniak, E. (2005). Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 290–297, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. (2007). Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*.

Yang, Y., Teo, C. L., Daumé III, H., and Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, Scotland.

Yu, H. and Siskind, J. M. (2013). Grounded language learning from video described with sentences. In *ACL (1)*, pages 53–63.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1353–1361. Association for Computational Linguistics.