

Context-specific Approaches for Generation of Image Captions

Rebecca Mason

October 28, 2013

Abstract

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). I propose work towards image captioning methods that learn from weakly-supervised examples of previously captioned images in a same context. This approach employs data-oriented text-to-text generation techniques, such as extractive multi-document summarization and sentence compression, which generate novel image captions by adapting text from captions of visually similar images.

In this proposal, I will present my ongoing image captioning research which demonstrates how text-to-text generation techniques can produce coherent and informative captions for images which are difficult or impossible to caption using previously published methods. I propose to extend this work to broader domains of images and text, to further develop models for analyzing and understanding caption text, to construct joint models for transferring and editing previously written captions, and to investigate the efficacy of various evaluation practices for image captioning.

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. On the web, some motivations for image captioning are providing background and context to images, improving accessibility of websites for visually-impaired users, and improving image retrieval by providing text to search user queries against.

Caption generation has potential impact for the wide variety of users who interact with large image databases. On social media, image sharing is a popular mode of interaction. **Facebook.com** has over 250 billion images uploaded as of June 2013, with its 1.15 billion registered users uploading 350 million images a day on average.¹ In addition to sharing with friends and family, social media images are valuable for other applications such as tracking consumer trends, or helping journalists and law enforcement officials understand an event. 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.

Here are the intended contributions for this thesis work:

1. **Generally-applicable methods for context-specific captioning.** This work introduces the task of context-specific image captioning. This thesis proposal describes preliminary work that learns to generate captions for online shopping images, using models that are trained on similar images and their captions. It also proposes future work for identifying specific contexts that exist within larger general-domain datasets.
2. **Image captioning without labeled training data or auxiliary information.** This research is novel in that it addresses the problem of image understanding without using either supervised visual detection algorithms, or auxiliary information such as natural language text that is directly related to the image. Preliminary work shows that this approach is able to generate captions for images that could not be captioned using previously existing techniques.
3. **Text-to-text methods for image caption generation.** Text-to-text is a natural language generation technique where some source text

¹https://fbcdn-dragon-a.akamaihd.net/hphotos-ak-ash3/851590_229753833859617_1129962605_n.pdf

is used as input to the language generation process. These techniques are useful for the task of context-specific image captioning, because a human-authored text that is relevant to one image, is often also relevant to similar images. This thesis explores methods for identifying text that is relevant to a query image, and for adapting visual text from the input source caption to an output caption for the query image.

This thesis proposal is structured as follows:

- The second section provides background information on image captioning and natural language generation, a brief survey of previous image captioning research, and methods of evaluation for automatically generated image captions.
- The third section describes my preliminary research. This includes previous related work I have done at Brown, and work on caption generation for online shopping images.
- The fourth section outlines research that I propose to do for this thesis.
- The fifth section contains a list of papers I have written.

2 Background

The main components of the automatic captioning process are image understanding and language generation. Each are very difficult problems, with large communities of researchers from the fields of Natural Language Processing (NLP) and Computer Vision (CV). In this section, I will briefly cover approaches to each of these components as well as previous work in image caption generation. I will then discuss evaluation methods and what are the attributes of a good image caption.

2.1 Image Understanding

Image understanding refers to determining the content and meaning of a source image. This usually implies that techniques from Computer Vision are used, but this does not always have to be the case.

Recent advances in general-domain image captioning have been enabled by improvements in visual object recognition, especially the deformable part model (Felzenszwalb et al., 2010, 2008). It represents images using low-level HOG features (Dalal and Triggs, 2005), which measure the direction of the change in light 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.

Because the deformable part model is supervised, it requires many labeled images to train on. Typically each label corresponds to a “bounding box” in the image that indicates where in the image the object begins and ends. However, there are some examples of image understanding where the label does not correspond to a specific part of the object. For example, Barnard et al. (2003) use a less-supervised topic model to learn both the labels for an image, as well as what part of the image they correspond to.

There are also visual classifiers that are based on the entire image. These classifiers can employ scene features such as Tiny Image (Torralba et al., 2008) and GIST (Oliva and Torralba, 2001), which characterize the spatial layout of the entire image. GIST coarsely localizes low-level features for what direction the light is oriented, but does not capture color or fine details. Tiny Image features are basically computed by scaling down the size of an image until all that is left is the basic structure of the colors in the scene. GIST and Tiny Image features can help classify different types 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, or indoor lighting vs outdoor lighting.

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. 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).

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 in the NLP community, 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* (*shallow generation*) methods which produce textual output using input text from human-authored sources. However, most previous image captioning work uses a hybrid of these approaches.

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), and the communication objective for the output. Realization of the selected content can be carried out in a variety of ways. In the simplest cases, a system might select the best output from a collection of previously written text. Other systems use

templates or grammars which support greater flexibility, but have a greater likelihood of producing ungrammatical or incoherent output. To reduce these errors, these systems may include more sophisticated components for generating *referring expressions*, which are phrases that identify particular nouns in text.

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. Summaries can be *extractively*, by selecting relevant sentences from the original document(s); or *abstractively*, by generating new sentences using realization methods from concept-to-text generation.

Compression : Decreasing the length of an input sentence by deleting words that are not relevant, without making the sentence ungrammatical.

Paraphrasing : Rewording and rearranging phrases or sentences in a different way from the original.

Simplification : Rewriting a sentence to make it easier to understand.

Fusion : Combining the relevant content of two sentences into one single sentence.

As shown by these examples, text-to-text generation methods are usually 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. However, outside sources, including non-linguistic information, can also be used to guide selection of relevant content.

2.3 Related Image Captioning Work

Image captioning technology has benefitted from a recent surge of interest in the Natural Language Processing (NLP) and Computer Vision (CV) research communities.

Some earlier NLP frameworks for image captioning rely on non-visual information such as the GPS coordinates of an image, to retrieve documents that are related to the query image (Aker and Gaizauskas, 2010; Fan

et al., 2010). In this work, image captions are generated by summarizing the related documents. Deschacht et al. (2007) introduce visual information in their work to a limited degree, using a face detection algorithm to find the number of people in a news image, and a named-entity recognition algorithm to select possible names from the related news article. Feng and Lapata (2010a) generate captions for news images using both extractive and abstractive summarization methods on the related articles, but they use a joint model of visual and textual information to select relevant content. Because all of these systems rely on textual documents that are directly related to the query image, they tend to be most applicable for specific domains such as news or travel images.

Other image captioning frameworks do not rely on related textual documents, and are applicable for general-domain image captioning tasks. The presence of visual object classes, scenes, or attributes are detected using trained visual detection systems. These visual detections are manually aligned with textual words and descriptions, and placed into templates or grammar structures which provide form for the caption (Kulkarni et al., 2011; Li et al., 2011; Yang et al., 2011; Mitchell et al., 2012). Linguistic models also have an important role in these systems. They are used to correct noisy initial detections, predict verbs and preposition words which are difficult to determine visually, and help generate more natural-sounding output.

Farhadi et al. (2010) and Ordonez et al. (2011) generate descriptions using *caption transfer*. They retrieve text from other captioned images, and transfer the text to the query image. Transferred captions are selected using some measure of content similarity, either between the query image and the candidate text, or between two images directly. Kuznetsova et al. (2012) transfer phrases from images based on detected objects and attributes, and combine multiple phrases into a single output sentence.

There is also research in related tasks such as image annotation, visual attribute discovery, and caption generalization. This work will be discussed at various points during later sections of this document.

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 performance is measured by how much it helps humans can perform some task; and *automatic metrics* which compare generated



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

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 common approach for caption evaluation. 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, are easier to perform over the internet. They also 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 1 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. In NLP, there are a variety of specialized metrics for different language generation tasks, but there is no metric that is specifically developed for the task of evaluating 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), while ROUGE (Lin, 2004) is a similar metric that has been used for summarization as well. Both measure similarity between automatically generated and human-authored captions. BLEU is a modified unigram precision metric, originally developed for evaluating automatic translations. ROUGE is a recall-oriented metric developed for summarization evaluation.

Neither BLEU or ROUGE capture the variety of different ways that humans can describe an image, but for a large enough dataset, they can capture significant differences between how well different image captioning methods can resemble the content of the human-authored captions. However, neither measure is very good at measuring the grammaticality or coherence of generated image captions.

3 Preliminary Work

This section describes preliminary work for this thesis in the areas of image annotation, image annotation evaluation, and image caption generation.

3.1 Evaluating Image Annotations

This work examines evaluation methods for systems that automatically annotate images using related text. We obtain datasets that are used in previously published work for this task, and implement a series of baseline measures inspired by those used in information retrieval, computer vision, and extractive summarization. The objective is to learn what are the most promising methods for applications that combine vision and language, and to develop good practices for understanding them. Although this work was focused on the task of annotating images in the case where related documents are available, many of the ideas are also applicable to image caption generation.

Sometimes, different conventions in NLP and CV can conflict, causing misleading experimental results. For example, a good way to collect data in CV is to query terms into an image search engine, and collect the results (Berg and Forsyth, 2006). Since search engines rely on related textual information, this provides a visually diverse set of images that are related to the term. However, if both the image and the related document are to be collected, it can be a problem because the textual information is less diverse. All of the text in the dataset share common features that allowed it to have a high rank with the search engine. In our work, we examine the dataset from Leong et al. (2010) where many of the queries end up at image repository websites where the related text document is a list of keywords. Their best performing system (which includes complex topic models, measures of visualness, and a salience measure relying on syntax), has similar performance to our simple term-frequency baseline. This problem is relevant to some of the work proposed in Section 4 of this document, which will use the SBU dataset (Ordonez et al., 2011), which is also collected using query terms.

Another dataset we examined in this project was the BBC News Images Dataset² which is also used as a dataset for image caption generation in Feng and Lapata (2010a). For their captioning system, the input is a set of image annotations that are generated using both the related text and bag-of-SIFT features computed on the query image (Feng and Lapata, 2010b). However, we found that on the BBC dataset, their annotation model does

²<http://homepages.inuf.ed.ac.uk/s0677528/data.html>

not effectively use learn from the visual features, because the size of the training set is too small (3121 images). This problem is relevant to work in Section 3.2 of this thesis proposal.

The reason why we find so much previous work on this task to be misleading, is because it is incredibly difficult to decide which “gold” annotations should be tested against, and which baselines should be compared against. In the traditional CV annotation task, there is typically a fixed vocabulary of labels that a system can generate, and the ground-truth annotations are also limited to those labels. It is also typical to remove infrequent labels from the ground-truth data because it is not reasonable for a CV system to learn a label from only a few examples. However, if the image annotation system has any textual input, such as a document that is related to the query image, then it should be evaluated like a text-to-text generation system, in which that information is considered intrinsic to the input. Furthermore, when different image annotation systems are developed for a particular kinds of data and input, cross-system evaluations can be misleading. The systems compared against do not use all the features that are available, or the performance is degraded in the process of adapting the model. This problem is relevant to work in Sections 3.3 and 4.2 of this proposal, which introduce approaches to image captioning that do not require labeled training data or auxiliary information.

3.2 Annotation of Online Shopping Images without Labeled Training Examples

This work is concerned with the task of generating image annotations for online shopping images, using models that are trained on images and natural language captions. As mentioned in Section 3.1, the traditional image annotation task assumes that there is a fixed vocabulary of labels. One challenge of training annotation models on image captions, is that some visual content can be described using several different words, while other visual content might not be considered relevant to describe at all. Additionally, some words describe background information that is not shown visually, or contextual information that is interpreted by the user. Rather than modeling images and text such that one generates the other, we train a topic model based on LDA Blei et al. (2003) where both an image and its caption are generated by a shared latent distribution of topics.

Previous work by Feng and Lapata (2010b) shows that topic models where image features or regions generate text features (such as (Barnard et al., 2003)) are not appropriate for modeling images with captions or other

 <p>Two adjustable buckle straps top a classic rubber rain boot grounded by a thick lug sole for excellent wet-weather traction.</p>	 <p>Size(s) Available: 6, 11.5. Brand & Style - VANS Kvd Width - Medium (B, M) Heel Height - Shoe Size is Womens Size 11.5 = Mens Size 10 1 Inch Heel Material - Canvas Upper and Man Made Sole</p>	 <p>Carlo Fellini - Evening clutch beaded on a wave pattern</p>
---	--	---

Table 2: Example data from Attribute Discovery Dataset Berg et al. (2010).

collocated text. Our work in Section 3.1 shows it is also difficult to learn from visual features on small datasets that have very complex images, such as the BBC News Images. However, Berg et al. (2010) show that it is possible to learn relationships between descriptive words and visual features, using a larger dataset where the images are more similar to each other. In Berg et al. (2010), their primary interest is to characterize attributes according to how they are visually represented: global or local; color, texture, or shape. Here, our concern is the task of predicting attributes for query images, which is not addressed in their work.

Berg et al. (2010) collect images and captions from the shopping website aggregator *Like.com*, and make it publically available as the Attribute Discovery Dataset³. We use the women’s shoes and handbags sections of their dataset. The womens shoes section has 14764 captioned images, and the handbags section has 9145 captioned images. We are interested in this dataset because it provides a focused domain for both the images and the captions. Objects are typically oriented the same way in the images, and are set against a plain white background. The text is from a real-life situation where stores provide descriptions of products to encourage people to buy them. However, the products also have a lot of diversity. For example, there are many types of shoes such as hiking boots, stilettos, flip flops, and running shoes; and many different attributes such as materials, colors, and patterns. The captions are varied in quality, as shown in Table 2. While some captions have full-length grammatical sentences, others are just a va-

³<http://tamaraberg.com/attributesDataset/index.html>

riety of keywords. The captions describe visual features, but also non-visual qualities such as sizing and shipping, or what is the proper occasion to wear the item.

Since the Attribute Discovery Dataset does not have a manually-crafted set of labels, we consider all “descriptive” words⁴ in the training image captions as possible labels, as is done in Feng and Lapata (2008). Unlike Berg et al. (2010), we do not attempt to discover multi-word attributes or cluster synonymous terms. This leaves a vocabulary of 9578 descriptive words for shoes, and 6309 descriptive words for handbags. On average, there are about 16 descriptive words per caption. For images, we compute SIFT descriptors at points of interest, and use k-means to cluster them into 750 visual terms.

We use a topic model designed for multi-lingual data, specifically the Polylingual Topic Model Mimno et al. (2009). This model was developed for correlated documents in different languages that are topically similar, but are not direct translations, such as Wikipedia or news articles in different languages. We train the topic model with images and text as two languages. For query images, we estimate the topic distribution that generated just the image. Finally, we find the probability for each word in the vocabulary being generated by that topic distribution.

To evaluate the generated annotations, we perform an annotation task similarly to Feng and Lapata (2010b): we compare precision and recall of the annotations against descriptive words in the original caption. Our method has a 30-35% improvement in finding words from the held-out image caption, compared to previous methods and baselines. We also compute additional visual bag-of-word features, to capture color, texture, shading, and curvature, and include them in our model, which yields a modest, but significant gain in performance.

3.3 Captioning Online Shopping Images using Sentence Compression

Using the model and the dataset that are introduced in Section 3.2, we introduce a *domain-specific* image captioning task, and present a framework for caption transfer and compression. This framework first transfers a caption from the training set image to the query image, using a nearest-neighbor algorithm based on the spatial layout of the images. The multi-modal topic model (trained on bag-of-word image features) is used to estimate how ac-

⁴Adjectives, adverbs, nouns, and verbs (except for stopwords and proper nouns)

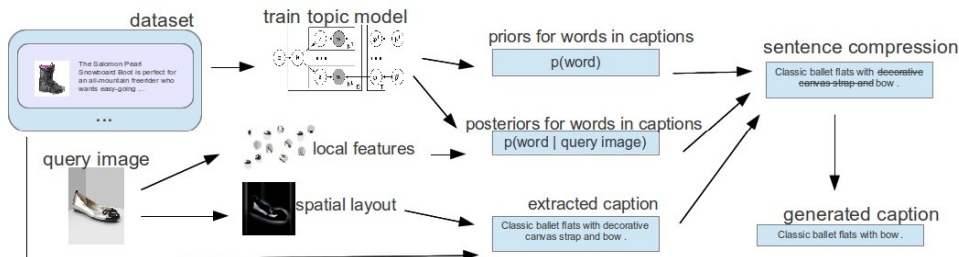


Figure 1: An overview of our image captioning system.

curate are the individual words in the transferred caption. Then, a sentence compression model is used to optimize the accuracy of the transferred caption, while retaining its grammaticality. Both automatic and human evaluations are performed, and show that our model improves the accuracy of the transferred captions.

This framework treats the overall shape of the object as the “scene” and extracts a caption sentence using GIST Oliva and Torralba (2001) nearest neighbors between the query image and the images in the training set. Because similar objects and attributes tend to appear in similar scenes, we expect that at least some of the extracted caption will describe local attributes that are also in the query image. GIST has previously been used as a baseline for caption transfer in Ordonez et al. (2011). We also compare against the best performing transfer method from Feng and Lapata (2010a) (using the ROUGE and BLEU automatic metrics), but find that GIST has better performance.

The rest of the framework finds and removes the parts of the transferred caption that are not accurate to the query image. Our model finds the optimal set of words to delete from the transferred caption using an ILP (Integer Linear Program) to optimize the content of the caption without deleting words that would make the caption become ungrammatical. While we are recently aware that Kuznetsova et al. (2013) published first on applying sentence compression to image captions, our work is still first to compress image captions with the goal of adaptation, rather than generalization. In other words, our objective is to find the optimal compression of the source text with respect to a specific query image. In contrast, Kuznetsova et al. (2013) compress with the objective of generally *generally* transferrable image captions.

Our ILP model draws inspiration from sentence compression work by

Clarke and Lapata (2008). The ILP has three inputs: the extracted caption; the prior probabilities words appearing in captions, $p(w)$; and the posterior probabilities of words appearing in captions given the query image, $p(w|query)$. The latter is estimated using the topic model from Section 3.2. The output of the ILP is a compressed image caption where the inaccurate words have been deleted.

The formal ILP objective⁵ is to maximize a weighted linear combination of two measures. The first measure in the objective is a trigram language model. The second is $\sum_{i=1}^n \delta_i \cdot I(w_i)$, where w_i, \dots, w_n are words in the extracted caption, δ_i is a binary decision variable which is true if we include w_i in the compressed output, and $I(w_i)$ is a score for the accuracy of each word. For non-function words, $I(w_i) = p(w|query) - p(w)$, which can have a positive or negative value. We do not use $p(w_i|query)$ directly in order to distinguish between cases where $p(w_i|query)$ is low because w_i is inaccurate, and cases where $p(w_i|query)$ is low because $p(w_i)$ is low generally. Function words do not affect the accuracy of the generated caption, so $I(w_i) = 0$.

We evaluate this work using ROUGE and BLEU scores against extractive baselines and a compression baseline that is based on maximizing the language model and does not consider the bag-of-words image features. Our model generates the most accurate captions (best precision) over all the baselines, and does not significantly decrease recall from the original transferred caption. Additionally, we run a human study evaluation, where we ask participants whether the compressed captions are more accurate or less grammatical than the original transferred caption. Both our compression model and the baseline compression model preserve the accuracy of the caption the majority of the time (73.1% for the system, and 82.2% for the baseline; not a significant difference). However, our model improves the accuracy of the transferred caption 63.2% percent of the time, versus only 42.6% of the time for the baseline compression method.⁶ In the majority of the time, our model effectively improves the accuracy of transferred captions to a query image, without making the generated caption less grammatical.

⁵To formulate this problem as a linear program, the probabilities are actually log probabilities, but the logs are omitted here to save space.

⁶Note that for the other 36.8% of transferred captions, even if the compression model deletes an accurate word, this does not mean the caption is less accurate, only less descriptive.



Figure 2: Images from the SBU-Flickr dataset that have been clustered into k-means using scene attributes (Patterson and Hays, 2012).

4 Proposed Work

4.1 Context Discovery in General Domain Data

One clear direction for future work is to extend our image captioning framework to natural images. By “natural images” we refer to images of everyday scenes seen by people, unlike the shopping images, where objects tend to be posed in similar positions against plain backgrounds. Instead of previously-defined domains such as handbags and shoes, we propose to cluster general domain images based on visual scene domains. We are interested in finding new domains of images and captions where we could try to learn local image features the way we did for the shoes images. Although visual recognition is generally much more difficult in natural scenes than in posed images, our hope is to find domains where types and attributes of objects, and the ways that they appear visually, will be much more limited.

As an example, we use the scene attributes and classifiers by Patterson and Hays (2012), which build an attribute-based taxonomy of scene types using crowd-sourcing, in order to k-means cluster the 1 million captioned images in the SBU-Flickr dataset from Ordonez et al. (2011). Figure 2 shows a few examples of images and text from one of the clusters. Although the clusters were generated using only visual information, there are words such as “building”, “blue sky”, and “outside” that are shared between the captions as well. Although these words do not appear in every caption that has these features, we could use them as a feature in order to learn words and scene attributes that are related to each other from noisy images and captions that appear together naturally. This information would be helpful for caption transfer, because we could give more weight to visual features

that are more highly correlated with text features, and vice versa.

So far, we have made a very small preliminary effort to explore this topic, but we are interested in using canonical correlation analysis (CCA) which is a way of finding the linear relationships between two sets of multidimensional variables. This method has recently been used to discover the shared basis of images and natural language captions for image caption search and retrieval (Hodosh et al., 2013). Spectral methods are used for many other NLP tasks, and have the advantage that they can be trained very quickly on large amounts of data.

4.2 Caption Transfer by Consensus

Even scene clusters that only use visual scene data can still be used to improve transferred captions. We are interested in applying ideas from multi-document summarization to nearest-neighbor image captions. In many instances, transferring the first nearest-neighbor caption using GIST or some other method does not result in a good caption. This can be because there are relevant objects or features that the image descriptor does not capture, or because the nearest-neighbor image has a caption which is unlikely to be relevant to any other visually similar image.

In such cases, we propose the following: instead of retrieving the caption of the nearest-neighbor image to the query image, we retrieve the k nearest-neighbor images and their captions. We then approach the image transfer problem as an extractive multi-document summarization task, where the objective is to retrieve text that contains the key shared relevant information between the documents (Goldstein et al., 2000).

Although there is much work to be done in figuring out the optimal summarization techniques to be used for image captions, we run a small toy example retrieving the 25 nearest-neighbor images of query images from the SBU-Flickr dataset according to scene attributes from Patterson and Hays (2012). We use the TOPICSUM model from Haghighi and Vanderwende (2009) to determine the relevant content words from each set of captions.

Figure 3 shows some examples of the top shared relevant words that were found for different query images. Some of the words are very accurate, but there are also some limitations to this approach because it only captures words that describe scene attributes. Also, it can inaccurately suggest words if those words appear frequently in the captions of the nearest-neighbor images. However, it is possible that some of these issues may be mitigated by the choice of summarization objective, which we have not yet explored.

We are interested in multi-document summarization for caption transfer,

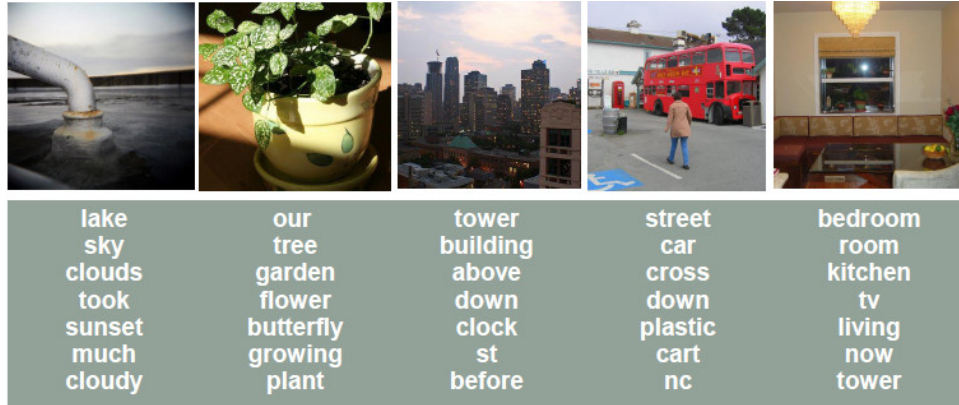


Figure 3: Top shared relevant words from the 25 nearest-neighbors captions for different query images.

because it provides a clear benefit for query images where object detections are not available, or are too difficult to obtain. However, there are other problems that we propose to apply this to as well. We could use the summarization model to identify words in captions that are likely to describe scene attributes, as opposed to words that are background words (stopwords) in the language, or words that describe non-visual or non-scene-based content. This model could also provide an alternative approach to the image caption generalization task that is introduced by Kuznetsova et al. (2013). Instead of using sentence compression to delete words that are not general enough, we could use the summarization model to select alternate captions that are more general.

4.3 Joint Model for Caption Transfer and Compression

Our caption generation framework from Section 3.3 currently extracts only a single caption sentence to compress, while recent work in summarization has focused on the problem of learning how to jointly extract and compress Martins and Smith (2009); Berg-Kirkpatrick et al. (2011). Developing a joint model for caption transfer and compression to generate image captions for the shoes images would connect the compression model in the previous work to the new techniques that we wish to develop in our proposed work.

A considerable challenge in this proposed work is to be able to efficiently search for the optimal compression, when multiple transfer candidates are

being considered. In a traditional summarization task, this is somewhat easier because both the extraction and compression have similar objectives of describing relevant content without taking up too much space. However, in our caption generation framework the transfer and compression objectives are less similar, because they separately consider spatial and bag-of-word visual features. A joint model would require developing a new joint objective and learning how to balance content that describes the two.

We would also need to develop a new framework for evaluation. Our work in Section 3.3 was concerned with evaluating the performance of our compression model, so both the human and automatic studies mostly measured improvement over the GIST transferred caption. We would need to employ different methods of evaluation in order to measure the performance of a joint model.

Since a poor extraction choice can make finding an accurate compression impossible, we should also study different methods of extraction to learn about what kinds of features are most likely to help us find good sentences. In Section 3.3, we found that a global feature descriptor was better than a bag of image word descriptors for caption transfer in the shopping dataset. As we extend our framework to other domains of images, we are interested in finding whether scene-based descriptors and classifiers in general are better at finding good sentences than local descriptors, and whether there is a connection between region and phrase-based detectors correlating better with sentence and phrase-length text, while local image descriptors are more related to single words. Finding patterns like this in visual text in general would be helpful for many other tasks besides image captioning.

5 List of Papers

Mason, R. and Charniak, E. (2011b). Extractive multi-document summaries should explicitly not contain document-specific content. In *Proceedings of the ACL 2011 Workshop on Automatic Summarization for Different Genres, Media, and Languages*, Portland, Oregon. Association for Computational Linguistics

Mason, R. and Charniak, E. (2011a). Bllip at tac 2011: A general summarization system for a guided summarization task. In *Proceedings of TAC 2011*

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

(Note: The work from Section 3.3 is currently under submission to a conference.)

References

- 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.
- Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D., and Jordan, M. (2003). Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135.
- 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.
- Berg, T. L. and Forsyth, D. A. (2006). Animals on the web. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1463–1470. IEEE.
- Berg-Kirkpatrick, T., Gillick, D., and Klein, D. (2011). Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 481–490, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.
- Clarke, J. and Lapata, M. (2008). Global inference for sentence compression an integer linear programming approach. *J. Artif. Int. Res.*, 31(1):399–429.
- 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.
- Deschacht, K., Moens, M.-F., et al. (2007). Text analysis for automatic image annotation. In *ACL*, volume 7, pages 1000–1007.
- 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.

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.

Goldstein, J., Mittal, V., Carbonell, J., and Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization-Volume 4*, pages 40–48. 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.

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.

- Jurafsky, D. and James, H. (2009). Speech and language processing an introduction to natural language processing, computational linguistics, and speech.
- 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.
- 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.
- 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.
- 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 Language 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. (2011a). Bllip at tac 2011: A general summarization system for a guided summarization task. In *Proceedings of TAC 2011*.
- Mason, R. and Charniak, E. (2011b). Extractive multi-document summaries should explicitly not contain document-specific content. In *Proceedings of the ACL 2011 Workshop on Automatic Summarization for Different Genres, Media, and Languages*, Portland, Oregon. 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.
- 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)*.
- 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.
- 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.
- 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.