Domain-Independent Captioning of Domain-Specific Images

Rebecca Mason
Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University, Providence, RI 02912
rebecca@cs.brown.edu

Abstract

Automatically describing visual content is an extremely difficult task, with hard AI problems in Computer Vision (CV) and Natural Language Processing (NLP) at its core. Previous work relies on supervised visual recognition systems to determine the content of images. These systems require massive amounts of hand-labeled data for training, so the number of visual classes that can be recognized is typically very small. We argue that these approaches place unrealistic limits on the kinds of images that can be captioned, and are unlikely to produce captions which reflect human interpretations.

We present a framework for image caption generation that does not rely on visual recognition systems, which we have implemented on a dataset of online shopping images and product descriptions. We propose future work to improve this method, and extensions for other domains of images and natural text.

1 Introduction

As the number of images on the web continues to increase, the task of automatically describing images becomes especially important. Image captions can provide background information about what is seen in the image, can improve accessibility of websites for visually-impaired users, and can improve image retrieval by providing text to search user queries against. Typically, online search engines rely on collocated textual information to resolve queries, rather than analyzing visual content directly. Likewise, earlier image captioning research from the Natural Language Processing (NLP) community use collocated information such as news articles or GPS coordinates, to decide what information to include in the generated caption (Deschacht and Moens, 2007; Aker and Gaizauskas, 2010; Fan et al., 2010; Feng and Lapata, 2010a).

However, in some instances visual recognition is necessary because collocated information is missing, irrelevant, or unreliable. Recognition is a classic Computer Vision (CV) problem including tasks such as recognizing instances of object classes in images (such as `car`, `cat`, or `sofa`); classifying images by scene (such as `beach` or `forest`); or detecting attributes in an image (such as `wooden` or `feathered`). Recent works in image caption generation represent visual content via the output of trained recognition systems for a pre-defined set of visual classes. They then use linguistic models to correct noisy initial detections (Kulkarni et al., 2011; Yang et al., 2011), and generate more natural-sounding text (Li et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012).

A key problem with this approach is that it assumes that image captioning is a grounding problem, with language acting only as labels for visual meaning. One good reason to challenge this assumption is that it imposes unrealistic constraints on the kinds of images that can be automatically described. Previous work only recognizes a limited number of visual classes – typically no more than a few dozen in total – because training CV systems requires a huge amount of hand-annotated data. For example, the PASCAL VOC dataset[1] has 11,530 training im-

[1] http://pascallin.ecs.soton.ac.uk/

Proceedings of the NAACL HLT 2013 Student Research Workshop, pages 69–76,
Atlanta, Georgia, 13 June 2013. ©2013 Association for Computational Linguistics

ages with 27,450 labeled objects, in order to learn only 20 object classes. Since training visual recognition systems is such a burden, "general-domain" image captioning datasets are limited by the current technology. For example, the SBU-Flickr dataset (Ordonez et al., 2011), which contains 1 million images and captions, is built by first querying Flickr using a pre-defined set of queries, then further filtering to remove instances where the caption does not contain at least two words belonging to their term list. Furthermore, detections are too noisy to generate a good caption for the majority of images. For example, Kuznetsova et al. (2012) select their test set according to which images receive the most confident visual object detection scores.

We instead direct our attention to the *domain-specific* image captioning task, assuming that we know a general object or scene category for the query image, and that we have access to a dataset of images and captions from the same domain. While some techniques may be unrealistic in assuming that high-quality collocated text is always available, assuming that there is no collocated information at all is equally unrealistic. Data sources such as file names, website text, Facebook likes, and web searches all provide clues to the content of an image. Even an image file by itself carries metadata on where and when it was taken, and the camera settings used to take it. Since visual recognition is much easier for domain-specific tasks, there is more potential for natural language researchers to do research that will impact the greater community.

Finally, labeling visual content is often not enough to provide an adequate caption. The meaning of an image to a user is more than just listing the objects in the image, and can even change for different users. This problem is commonly known as "bridging the semantic gap":

> "The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation. A linguistic description is almost always contextual, whereas an image may live by itself." (Smeulders et al., 2000)

General-domain models of caption generation fail to capture context because they assume that all the relevant information has been provided in the image. However, training models on data from the same domain gives *implicit* context about what information should be provided in the generated text.

This thesis proposes a framework for image captioning that does not require supervision in the form of hand-labeled examples. We train a topic model on a corpus of images and captions in the same domain, in order to jointly learn image features and natural language descriptions. The trained topic model is used to estimate the likelihood of words appearing in a caption, given an unseen query image. We then use these likelihoods to rewrite an extracted human-written caption to accurately describe the query image. We have implemented our framework using a dataset of online shopping images and captions, and propose to extend this model to other domains, including natural images.

2 Framework

In this section, we provide an overview of our image captioning framework, as it is currently implemented. As shown in Figure 1, the data that we use are a set of images and captions in a specific domain, and a query image that is from the same domain, but is not included in the training data. The training data is used in two ways: for **sentence extraction** from the captions of training images that are visually similar to the query image overall; and for training a **topic model** of individual words and local image features, in order to capture fine-grained details. Finally, a **sentence compression** algorithm is used to remove details from the extracted captions that do not fit the query image.

The work that we have done so far has been implemented using the Attribute Discovery Dataset (Berg et al., 2010), a publicly available dataset of shopping images and product descriptions.[2] Here, we run our framework on the women's shoes section, which has over 14000 images and captions, representing a wide variety of attributes for texture, shapes, materials, colors, and other visual qualities. The women's shoes section is formally split

challenges/VOC/

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

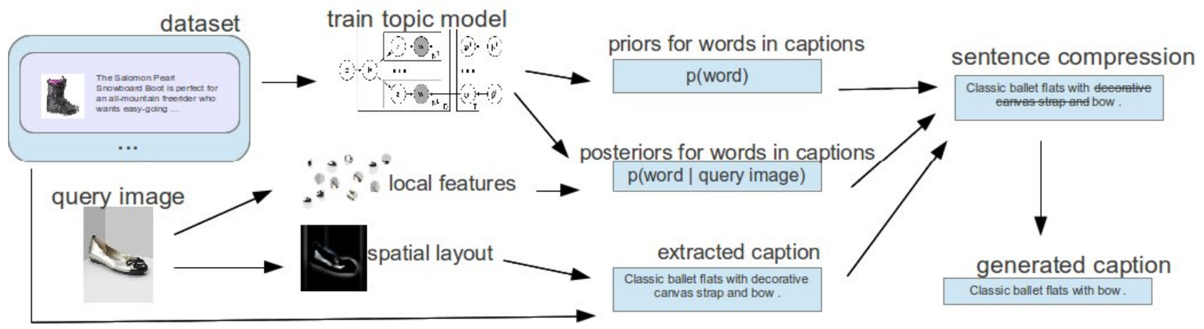

dataset
The Salomon Pearl Snowboard Boot is perfect for an all-mountain freerider who wants easy-going ...
...
train topic model
priors for words in captions
p(word)
sentence compression
Classic ballet flats with ~~decorative canvas strap and~~ bow .
posteriors for words in captions
p(word | query image)
query image
local features
spatial layout
extracted caption
Classic ballet flats with decorative canvas strap and bow .
generated caption
Classic ballet flats with bow .

Figure 1: Overview of our framework for image caption generation.

into ten subcategories, such as `wedding shoes`, `sneakers`, and `rainboots`. However, many of the subcategories contain multiple visually distinct kinds of shoes. We do not make use of the subcategories, instead we group all of the categories of shoe images together. The shoes in the images are mostly posed against solid color backgrounds, while the captions have much more variability in length and linguistic quality.

For our thesis work, we intend to extend our current framework to different domains of data, including natural images. However, it is important to point out that no part of the framework as it is currently implemented is specific to describing shoes or shopping images. This will be described in Section 4.

2.1 Sentence Extraction

GIST (Oliva and Torralba, 2001) is a *global* image descriptor which describes how gradients are oriented in different regions of an image. It is commonly used for classifying background scenes in images, however images in the Attribute Discovery Dataset do not have "backgrounds" per se. Instead, we treat the overall shape of the object as the "scene" and extract a caption sentence using GIST 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. The rest of our framework finds and removes the parts of the extracted caption that are not accurate to the query image.

2.2 Topic Model

Image captions often act as more than labels of visual content. Some visual ideas can be described using several different words, while others are typically not described at all. Likewise, 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 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 Blei and Jordan (2003)) are not appropriate for modeling images with captions or other collocated text. 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, and then In the model, images and their captions are represented using *bag-of-words*, a commonly-used technique for document representation in both CV and NLP research. The textual features are non-function words in the model, including words that describe specific objects or attributes (such as `boot`, `snake-skin`, `buckle`, and `metallic`) in addition to words that describe more abstract attributes and affordances (such as `professional`, `flirty`, `support`,

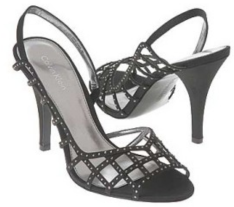
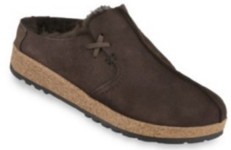
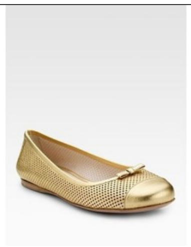
 **Original:** Go all-out glam in the shimmering Dyeables Roxie sandals. Metallic faux leather upper in a dress thong sandal style with a round open toe. ...	 **Original:** Find the softness of shearling combined with support in this clog slipper. The cork footbed mimics the foot's natural shape, offering arch support, while a flexible outsole flexes with your steps and resists slips. ...	 **Original:** Perforated leather with cap toe and bow detail.
Extracted: Shimmering snake-embossed leather upper in a slingback evening dress sandal style with a round open toe .	**Extracted:** This sporty sneaker clog keeps foot cool and comfortable and fully supported.	**Extracted:** Italian patent leather peep-toe ballet flat with a signature tailored grosgrain bow .
System: Shimmering upper in a slingback evening dress sandal style with a round open toe .	**System:** This clog keeps foot comfortable and supported.	**System:** leather ballet flat with a signature tailored grosgrain bow .

Table 1: Some examples of shoes images from the Attribute Discovery Dataset and performance with our image captioning model. Left: Correctly removes explicitly visual feature "snake-embossed leather" from extraction; leaves in correct visual attributes "shimmering", "slingback", and "round open toe". Center: Extracted sentence with some contextually visual attributes; the model correctly infers that "sporty" and "cool" are not likely given an image of a wool bedroom slipper, but "comfortable" and "supported" are likely because of the visible cork soles. Right: Extracted sentence with some non-visual attributes; model removes "Italian" but keeps "signature tailored".

and `waterproof`). For "image words", we compute features at several points in the image such as the color values of pixels, the angles of edges or corners, and response to various filters, and cluster them into discrete image words. However, the information that an image word conveys is very different than the information conveyed in a text word, so models which require direct correspondence between features in the two modalities would not be appropriate here.

We train the topic model with images and text as two languages. We estimate the probabilities of textual words given a query image by first estimating the topic distribution that generated the image, and then using the same distribution to find the probabilities of textual words given the query image. However, we also perform an annotation task similarly to Feng and Lapata (2010b), in order to evaluate the topic model on its own. Our method has a 30-35% improvement in finding words from the held-out image caption, compared to previous methods and baselines.

2.3 Sentence Compression via Caption Generation

We describe an ILP for caption generation, drawing 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 their posterior probabilities of words appearing in captions given the query image, $p(w|query)$. The latter is estimated using the topic model we have just described. The output of the ILP is a compressed image caption where the inaccurate words have been deleted.

Objective: The formal ILP objective[3] is to maximize a weighted linear combination of two measures. The first we define as $\sum_{i=1}^{n} \delta_i \cdot I(w_i)$, where $w_i, ..., w_n$ 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,

[3] To formulate this problem as a linear program, the probabilities are actually log probabilities, but we omit the logs in this paper to save space.

$I(w_i) = p(w|query) - p(w)$, 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 $I(w_i) = 0$.

The second measure in the objective is a trigram language model, described in detail in Clarke (2008). In the original sentence compression task, the language model is a component as it naturally prefers shorter output sentences. However, our objective is not to generate a shorter caption, but to generate a more accurate caption. However, we still include the language model in the objective, with a weighting factor ϵ, as it helps remove unnecessary function words and help reduce the search space of possible sentence compressions.

Constraints: The ILP constraints include sequential constraints to ensure the mathematical validity of the model, and syntactic constraints that ensure the grammatical correctness of the compressed sentence. We do not have space here to describe all of the constraints, but basically, using the "semantic head" version of the headfinder from Collins (1999), we constrain that the head word of the sentence and the head word of the sentence's object cannot be deleted, and for any word that we include in the output sentence, we must include its head word as well. We also have constraints that define valid use of coordinating conjunctions and punctuation.

We evaluate generated captions using automatic metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). These metrics are commonly used in summarization and translation research and have been previously used in image captioning research to compare automatically generated captions to human-written captions for each image (Ordonez et al., 2011; Yang et al., 2011; Kuznetsova et al., 2012). Although human-written captions may use synonyms to describe a visual object or attribute, or even describe entirely different attributes than what is described in the generated captions, computing the automatic metrics over a large test set finds statistically significant improvements in the accuracy of the extracted and compressed captions over extraction alone.

For our proposed work (Section 4), we also plan to perform manual evaluations of our captions based on their content and language quality. However, cross-system comparisons would be more difficult because our method uses an entirely different kind of data. In order to compare our work to related methods (Section 3), we would have to train for visual recognition systems for hundreds of visual attributes, which would mean having to hand-label the entire dataset.

3 Related Work in Image Captioning

In addition to visual recognition, caption generation is a very challenging problem. In some approaches, sentences are constructed using templates or grammar rules, where content words are selected according to the output of visual recognition systems (Kulkarni et al., 2011; Yang et al., 2011; Mitchell et al., 2012). Function words, as well as words like verbs and prepositions which are difficult to recognize visually, may be selected using a language model trained on non-visual text. There is also similar work that uses large-scale ngram models to make the generated output sound more natural (Li et al., 2011).

In other approaches, captions are extracted in whole or in part from similar images in a database. For example, Farhadi et al. (2010) and Ordonez et al. (2011) build semantic representations for visual content of query images, and extract captions from database images with similar content. Kuznetsova et al. (2012) extract phrases corresponding to classes of objects and scenes detected in the query image, and combine extracted phrases into a single sentence. Our work is different than these approaches, because we directly measure how visually relevant individual words are, rather than only using visual similarity to extract sentences or phrases.

Our method is most similar to that of Feng and Lapata (2010a), who generate captions for news images. Like them, we train an LDA-like model on both images and text to find latent topics that generate both. However, their model requires both an image and collocated text (a news article) to estimate the topic distribution for an unseen image, while our topic model only needs related text for the training data. They also use the news article to help generate captions, which means that optimizing their generated output for content and grammaticality is a much easier problem. Although their model combines phrases and n-grams from different sentences to form an image caption, they only consider the text from a single news article for extraction, and they can assume that the text is mostly accurate and relevant to the content of the image.

In this sense, our method is more like Kuznetsova et al. (2012), which also uses an Integer Linear Program (ILP) to rapidly optimize how well their generated caption fits the content of the image model. However, it is easier to get coherent image captions from our model since we are not combining parts of sentences from multiple sources. Since we build our output from extracted sentences, not phrases, our ILP requires fewer grammaticality and coherence constraints than it would for building new sentences from scratch. We also model how relevant each individual word is to the query image, while they extract phrases based on visual similarity of detected objects in the images.

4 Proposed Work

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 domains such as handbags and shoes, we propose to cluster the training data based on visual scene domains such as mountains, beaches, and living rooms. We are particularly interested in the scene attributes and classifiers by Patterson and Hays (2012) which builds an attribute-based taxonomy of scene types using crowd-sourcing, rather than categorical scene types which are typically used.

Visual recognition is generally much more difficult in natural scenes than in posed images, since lighting and viewpoints are not consistent, and objects may be occluded by other objects or truncated by the edge of the image. However, we are optimistic because we do not need to solve the *general* visual recognition task, since our model only learns how visual objects and attributes appear in specific domains of scenes, a much easier problem. Additionally, the space of likely objects and attributes to detect is limited by what typically appears in that type of scene. Finally, we can use the fact that our image captioning method is not grounded in our favor, and assume that if an object is partially occluded or truncated in an image, than it is less likely that the photographer considered that object to be interesting, so it is not as important whether that object is described in the caption or not.

Finally, there is also much that could be done to improve the text generation component on its own. Our framework 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). 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. As mentioned in Section 2.1, we have already found that global feature descriptors are better than bag of image word descriptors for extracting sentences to use in image caption compressions 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.

References

Ahmet Aker and Robert Gaizauskas. 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.

Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. 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.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 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.

David M. Blei and Michael I. Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 127–134, New York, NY, USA. ACM.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

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

James Clarke. 2008. *Global Inference for Sentence Compression: An Integer Linear Programming Approach*. Dissertation, University of Edinburgh.

Michael John Collins. 1999. *Head-driven statistical models for natural language parsing*. Ph.D. thesis, Philadelphia, PA, USA. AAI9926110.

Koen Deschacht and Marie-Francine Moens. 2007. Text analysis for automatic image annotation. In *ACL*, volume 45, page 1000.

Xin Fan, Ahmet Aker, Martin Tomko, Philip Smart, Mark Sanderson, and Robert Gaizauskas. 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.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 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.

Yansong Feng and Mirella Lapata. 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.

Yansong Feng and Mirella Lapata. 2010b. Topic models for image annotation and text illustration. In *HLT-NAACL*, pages 831–839.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, pages 1601–1608.

Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *ACL*.

Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

André F. T. Martins and Noah A. Smith. 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.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 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.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alexander C. Berg, Tamara L. Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *European Chapter of the Association for Computational Linguistics (EACL)*.

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

G. Patterson and J. Hays. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 0:2751–2758.

Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, December.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, Scotland.