

Domain-Specific Image Captioning

Rebecca Mason and Eugene Charniak

Brown Laboratory for Linguistic Information Processing (BLLIP)

Brown University, Providence, RI 02912

{rebecca, ec}@cs.brown.edu

Abstract

We present a data-driven framework for image caption generation which incorporates visual and textual features with varying degrees of spatial structure. We propose the task of domain-specific image captioning, where many relevant visual details cannot be captured by off-the-shelf general-domain entity detectors. We extract previously-written descriptions from a database and adapt them to new query images, using a joint visual and textual bag-of-words model to determine the correctness of individual words. We implement our model using a large, unlabeled dataset of women’s shoes images and natural language descriptions (Berg et al., 2010). Using both automatic and human evaluations, we show that our captioning method effectively deletes inaccurate words from extracted captions while maintaining a high level of detail in the generated output.

1 Introduction

Broadly, the task of image captioning is: given a query image, generate a natural language description of the image’s visual content. Both the image understanding and language generation components of this task are challenging open problems in their respective fields. A wide variety of approaches have been proposed in the literature, for both the specific task of caption generation as well as related problems in understanding images and text.

Typically, image understanding systems use supervised algorithms to detect visual entities and concepts in images. However, these typically require accurate hand-labeled training data, which is not available in most specific domains. Ideally,

-
1. Extract existing human-authored caption according to similarity of coarse visual features.



Query Image



Nearest-Neighbor

Nearest-neighbor caption: *This sporty sneaker clog keeps foot cool and comfortable and fully supported.*

-
2. Estimate correctness of extracted words using domain-specific joint model of text and visual bag-of-word features.

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

-
3. Compress extracted caption to adapt its content while maintaining grammatical correctness.

Output: *This clog keeps foot comfortable and supported.*

a domain-specific image captioning system would learn in a less supervised fashion, using captioned images found on the web.

This paper focuses on image caption generation for a specific domain – images of women’s shoes, collected from online shopping websites. 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.

Our captioning framework is inspired by several recent approaches at the intersection of Natural Language Processing and Computer Vision. Previous work such as Farhadi et al. (2010) and Ordonez et al. (2011) explore extractive methods for image captioning, but these rely on general-domain visual detection systems, and only gener-

ate extractive captions. Other models learn correspondences between domain-specific images and natural language captions (Berg et al., 2010; Feng and Lapata, 2010b) but cannot generate descriptions for new images without the use of auxiliary text. Kuznetsova et al. (2013) propose a sentence compression model for editing image captions, but their compression objective is not conditioned on a query image, and their system also requires general-domain visual detections. This paper proposes an image captioning framework which extends these ideas and culminates in the first domain-specific image caption generation system.

More broadly, our goal for image caption generation is to work toward less supervised captioning methods which 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.

2 Related Work

Our framework for domain-specific image captioning consists of three main components: extractive caption generation, image understanding through topic modeling, and sentence compression.¹ These methods have previously been applied individually to related tasks such as general domain image captioning and annotation. We briefly describe some of the related work:

2.1 Extractive Caption Generation

In previous work on image caption extraction, captions are generated by retrieving human-authored descriptions from visually similar images. Farhadi et al. (2010) and Ordonez et al. (2011) retrieve whole captions to apply to a query image, while Kuznetsova et al. (2012) generate captions using text retrieved from multiple sources. The descriptions are related to visual concepts in the query image, but these models use visual similarity to approximate textual relevance; they do not model image and textual features jointly.

2.2 Image Understanding

Recent improvements in state-of-the-art visual object class detections (Felzenszwalb et al., 2010)

¹A research proposal for this framework and other image captioning ideas was previously presented at NAACL Student Research Workshop in 2013 (Mason, 2013). This paper presents a completed project including implementation details and experimental results.

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). However, these systems typically rely on a small number of detection types, e.g. the twenty object categories from the PASCAL VOC challenge.² These object categories include entities which are commonly described in general domain images (people, cars, cats, etc) but these require labeled training data which is not typically available for the visually relevant entities in specific domains.

Our caption generation system employs a multi-modal topic model from our previous work (Mason and Charniak, 2013) which generates descriptive words, but lacks the spatial structure needed to generate a full sentence caption. Other previous work uses topic models to learn the semantic correspondence between images and labels (e.g. Blei and Jordan (2003)), but learning from natural language descriptions is considerably more difficult because of polysemy, hypernymy, and misalignment between the visual content of an image and the content humans choose to describe. The MixLDA model (Feng and Lapata, 2010b; Feng and Lapata, 2010a) learns from news images and natural language descriptions, but to generate words for a new image it requires both a query image and query text in the form of a news article. Berg et al. (2010) use discriminative models to discover visual attributes from online shopping images and captions, but their models do not generate descriptive words for unseen images.

2.3 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 objective differs from that of Kuznetsova et al. (2013), who compress image caption sentences with the objective of creating a corpus of generally transferrable image captions. Their compression objective is to maximize the probability of a caption conditioned on the source

²<http://pascallin.ecs.soton.ac.uk/challenges/VOC/>



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 1: Example data from the Attribute Discovery Dataset (Berg et al., 2010). See Section 3.

image, while our objective is conditioned on the query image that we are generating a caption for. Additionally, their model also relies on general-domain trained visual detections.

3 Dataset and Preprocessing

The dataset we use is the women’s shoes section of the publicly available Attribute Discovery Dataset³ 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 examples are shown in Table 1.

For preprocessing in our framework, 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.

4 Image Captioning Framework

4.1 Extraction

To repeat, our overall process is to first find a caption sentence from our database to use as a template, and then correct the template sentences using sentence compression. We compress by remov-

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

ing details that are probably not correct for the test 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 to 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.⁴ Table 2 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

4.2 Joint 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

⁴For 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.

⁵See Section 5.1 for additional implementation details.

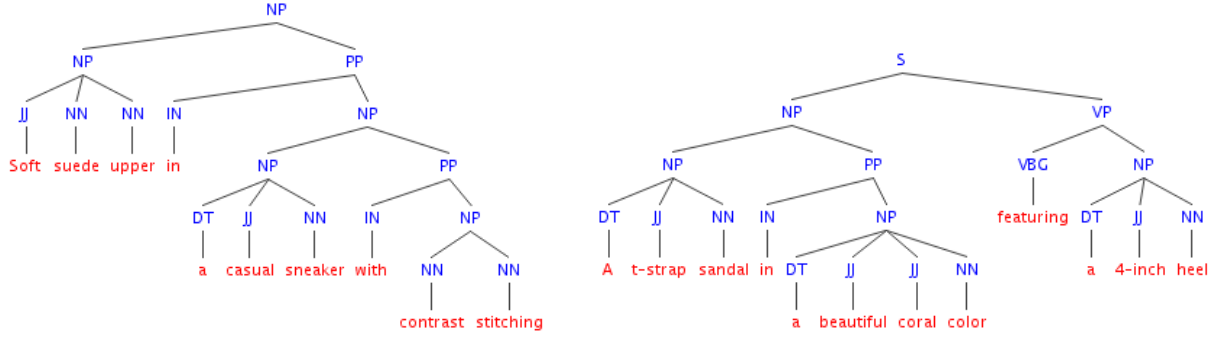


Table 2: Example parses of women’s shoes descriptions. Our hypothesis is that the headwords in phrases are more likely to describe visual concepts which rely on spatial locations or relationships, while modifiers words can be represented using less-structured visual bag-of-words features.

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.⁶ 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. In particular, we employ our previous work (Mason and Charniak, 2013) which extends this model to topically similar images and natural language captions. The generative process for a captioned image starts with a single topic distribution drawn from concentration parameter α and base measure m :

$$\theta \sim \text{Dir}(\theta, \alpha m) \quad (1)$$

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

$$\mathbf{z}^{img} \sim P(\mathbf{z}^{img}|\theta) = \prod_n \theta_{z_n^{img}} \quad (2)$$

⁶While space limits a more detailed explanation of visual bag-of-word features, Section 5.2 provides a brief overview of the specific visual attributes used in this model.

$$\mathbf{z}^{txt} \sim P(\mathbf{z}^{txt}|\theta) = \prod_n \theta_{z_n^{txt}} \quad (3)$$

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

$$\mathbf{w}^{img} \sim P(\mathbf{w}^{img}|\mathbf{z}^{img}, \Phi^{img}) = \phi_{w_n^{img}|z_n^{img}}^{img} \quad (4)$$

$$\mathbf{w}^{txt} \sim P(\mathbf{w}^{txt}|\mathbf{z}^{txt}, \Phi^{txt}) = \phi_{w_n^{txt}|z_n^{txt}}^{txt} \quad (5)$$

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:

$$P(z_n = t|q^{img}, z_{\setminus n}, \Phi^{img}, \alpha m) \propto \phi_{q_n^{img}|t}^{img} \frac{(N_t)_{\setminus n} + \alpha m_t}{\sum_t N_t - 1 + \alpha} \quad (6)$$

4.3 Sentence Compression

Let $\mathbf{w} = w_1, w_2, \dots, w_n$ 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.

4.3.1 Objective

The ILP objective is a weighted linear combination of two measures which represent the correctness and fluency of the output compression:

Correctness: Recall in Section 3 we defined words as either descriptive words or function words. For each descriptive word, we estimate $P(w_i|q^{img})$, using topic proportions estimated using Equation 6:

$$P(w_i|q^{img}) = \sum_t P(w_i|z_t^{txt})P(z_t|q^{img}) \quad (7)$$

This is used to find $I(w_i)$, a function of the likelihood of each word in the extracted caption:

$$I(w_i) = \begin{cases} P(w_i|q^{img}) - P(w_i), & \text{if descriptive} \\ 0, & \text{function word} \end{cases} \quad (8)$$

This function considers the prior probability of w_i because frequent words often have a high posterior probability even when they are inaccurate. 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} .

Fluency: We formulate a trigram language model as an ILP, which requires additional binary decision variables: $\alpha_i = 1$ if w_i begins the output compression, $\beta_{ij} = 1$ if the bigram sequence w_i, w_j ends the compression, $\gamma_{ijk} = 1$ if the trigram sequence w_i, w_j, w_k is in the compression, and a special ‘‘start token’’ $\delta_0 = 1$. 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 $\log P$:

$$\begin{aligned} \max z = & \left(\sum_{i=1}^n \alpha_i \cdot P(w_i|\text{start}) \right. \\ & + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^n \gamma_{ijk} \cdot P(w_k|w_i, w_j) \\ & \left. + \sum_{i=0}^{n-1} \sum_{j=i+1}^n \beta_{ij} \cdot P(\text{end}|w_i, w_j) \right) \cdot \lambda \\ & + \sum_{i=1}^n \delta_i \cdot I(w_i) \end{aligned} \quad (9)$$

Sequential	1.) $\sum_i \alpha_i = 1$ 2.) $\delta_k - \alpha_k - \sum_{i=0}^{k-2} \sum_{j=1}^{k-1} \gamma_{ijk} = 0$ $\forall k : k \in 1 \dots n$ 3.) $\delta_j - \sum_{i=0}^{j-1} \sum_{k=j+1}^n \gamma_{ijk} - \sum_{i=0}^{j-1} \beta_{ij} = 0$ $\forall j : j \in 1 \dots n$ 4.) $\sum_{j=i+1}^{n-1} \sum_{k=j+1}^n \gamma_{ijk} - \sum_{j=i+1}^n \beta_{ij} - \sum_{h=0}^{i-1} \beta_{hi} - \delta_i = 0$ $\forall i : i \in 1 \dots n$ 5.) $\sum_{i=0}^{n-1} \sum_{j=i+1}^n \beta_{ij} = 1$
Modifier	1. If head of the extracted sentence = w_i , then $\delta_i = 1$ 2. If w_i is head of a noun phrase, then $\delta_i = 1$ 3. Punctuation and coordinating conjunctions follow special rules (below). Otherwise, if $\text{headof}(w_i) = w_j$, then $\delta_i \leq \delta_j$
Other	1. $\sum_i \delta_i \geq 3$ 2. Define valid use of punctuation and coordinating conjunctions.

Table 3: Summary of ILP constraints.

4.3.2 ILP Constraints

The ILP constraints ensure both the mathematical validity of the model, and the grammatical correctness of its output. Table 3 summarizes the list of constraints. Sequential constraints are defined as in Clarke (2008) ensure that the ordering of the trigrams is valid, and that the mathematical validity of the model holds.

5 Implementation Details

5.1 Extraction

GIST features are computed using code by Oliva and Torralba (2001)⁷. GIST is computed with images converted to grayscale; since color features tend to act as modifiers in this domain. Nearest-neighbors are selected according to minimum distance from q^{img} to both a regularly-oriented and a horizontally-flipped training image.

Only one sentence from the first nearest-neighbor caption is extracted. In the case of multi-sentence captions, we select the first suitable sentence according to the following criteria 1.) has at least five tokens, 2.) does not contain NNP or NNPS (brand names), 3.) does not fail to parse using Stanford Parser (Klein and Manning, 2003). If the nearest-neighbor caption does not have any sentences meeting these criteria, caption sentences from the next nearest-neighbor(s) are considered.

⁷<http://people.csail.mit.edu/torralba/code/spatialenvelope/>

5.2 Joint Topic Model

We use the Joint Topic Model that we implemented in our previous work; please see Mason and Charniak (2013) for the full model and implementation details. The topic model is trained with 200 topics using the polylingual topic model implementation from MALLET⁸. Briefly, the code-words represent the following attributes:

SHAPE: SIFT (Lowe, 1999) describes the shapes of detected edges in the image, using descriptors which are invariant to changes in rotation and scale.

COLOR: RGB (red, green, blue) and HSV (hue, saturation, value) pixel values are sampled from a central area of the image to represent colors.

TEXTURE: Textons (Leung and Malik, 2001) are computed by convolving images with Gabor filters at multiple orientations and scales, then sampling the outputs at random locations.

INTENSITY: HOG (histogram of gradients) (Dalal and Triggs, 2005) describes the direction and intensity of changes in light. These features are computed on the image over a densely sampled grid.

5.3 Compression

The sentence compression ILP is implemented using the CPLEX optimization toolkit⁹. The language model weighting factor in the objective is $\lambda = 10^{-3}$, which was hand-tuned according to observed output. The trigram language model is trained on training set captions using BerkeleyLM (Pauls and Klein, 2011) with Kneser-Ney smoothing. For the constraints, we use parses from Stanford Parser (Klein and Manning, 2003) and the “semantic head” variation of the Collins headfinder Collins (1999).

6 Evaluation

6.1 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} .

⁸<http://mallet.cs.umass.edu/>

⁹<http://www-01.ibm.com/software/integration/optimization/cplex-optimization-studio/>

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 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.

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

LM-ONLY (COMPRESSION): We include this baseline to demonstrate that our model is effectively conditioning output compressions on q^{img} , as opposed to simply generalizing captions as in Kuznetsova et al. (2013)¹⁰. We modify the compression ILP to ignore the content objective and only maximize the trigram language model (still subject to the constraints).

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.

6.2 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

¹⁰Technically their model is conditioned on the source image, in order to address alignment issues which are not applicable in our setup.

	BLEU@1
KL (EXTRACTION)	.2098
GIST (EXTRACTION)	.4259
LM-ONLY (COMPRESSION)	.4780
SYSTEM (COMPRESSION)	.4841

Table 5: BLEU@1 scores of generated captions against human authored captions. Our model (bolded) has the highest BLEU@1 score with significance.

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 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, 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; Kuznetsova et al., 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.

6.3 Human Evaluation

We perform human evaluation of compressions generated by our system and LM-Only. 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



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

Table 6: Example output from our full system. Red underlined words indicate the words which are deleted by our compression model.

	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 7: Human evaluation results.

about fashion.¹¹ Users were recruited via email and did the study over the internet.

Table 7 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. ($p > 0.05$)

7 Conclusion

We introduce the task of domain-specific image captioning and propose a captioning system which is trained on online shopping images and natural language descriptions. We learn a joint topic model of vision and text to estimate the correctness of extracted captions, and use a sentence compression model to propose a more accurate output caption. Our model exploits the connection between image and sentence structure, and can be used to improve the accuracy of extracted image captions.

The task of domain-specific image caption generation has been overlooked in favor of the general-domain case, but we believe the domain-specific case deserves more attention. While image captioning can be viewed as a complex grounding problem, a good image caption should do more than label the objects in the image. When an expert looks at images in a specific domain, he or she makes inferences that would not be made by a non-expert. Providing this information to non-

¹¹About 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	GIST Nearest-Neighbor
	
Extraction: Classic ballet flats <u>with decorative canvas strap and patent leather</u> covered <u>buckle</u> .	Extraction: Classic ballet flats covered.
Compression: Classic ballet flats covered.	
Query Image	GIST Nearest-Neighbor
	
Extraction: This shoe is the <u>perfect shoe for you</u> , featuring an open toe and <u>a lace up</u> upper with a high heel, <u>and a two tone color</u> .	Extraction: This shoe is the shoe, featuring an open toe and upper with a high heel.
Compression: This shoe is the shoe, featuring an open toe and upper with a high heel.	

Table 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.

expert users in the form of an image caption will greatly expand the utility for automatic image captioning.

References

- 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.
- 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.
- 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.

- N. Dalal and B. Triggs. 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, June.
- 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.
- P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. 2010. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.
- 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.
- Sadaoki Furui, Tomonori Kikuchi, Yousuke Shinnaka, and Chiori Hori. 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE TRANS. ON SPEECH AND AUDIO PROCESSING*, 12(4):401–408.
- Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03*, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artif. Intell.*, 139(1):91–107, July.
- 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*.
- Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2013. Generalizing image captions for image-text parallel corpus. In *ACL*.
- T. Leung and J. Malik. 2001. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44.
- 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.
- D.G. Lowe. 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.
- 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 Language Processing, ILP '09*, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.
- R. Mason and E. Charniak. 2013. Annotation of online shopping images without labeled training examples. Workshop on Vision and Language (WVL).
- Rebecca Mason. 2013. Domain-independent captioning of domain-specific images. NAACL Student Research Workshop.
- 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.

Adam Pauls and Dan Klein. 2011. Faster and smaller n-gram language models. In *Proceedings of ACL*, Portland, Oregon, June. Association for Computational Linguistics.

Jenine Turner and Eugene Charniak. 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.

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.

Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 53–63, Sofia, Bulgaria. Association for Computational Linguistics.