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, 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. 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) 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. 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 multimodal 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

---

1 A 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.

2 http://pascalin.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 criss-crosses 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.
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. 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 $\alpha$ and base measure $m$:

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

(1)

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

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

(2)

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

(3)

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

$$w^{img} \sim P(w^{img}|z^{img}, \Phi^{img}) = \phi^{img}_w$$

(4)

$$w^{txt} \sim P(w^{txt}|z^{txt}, \Phi^{txt}) = \phi^{txt}_w$$

(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:

$$P(z_n = t|q^{img}, z_{\neq n}, \Phi^{img}, \alpha m) \propto \phi^{img}_t \frac{(N_t)_n + \alpha m_t}{\sum_t (N_t)_n + \alpha}$$

(6)

4.3 Sentence Compression

Let $w = w_1, w_2, ..., w_n$ be the words in the extracted caption for $q^{img}$. For each word, we define a binary decision variable $\delta$, 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 $\delta$ 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:

\[
P(w_i|q^{img}) = \sum_l P(w_i|z_l^{txt})P(z_l|q^{img}) \quad (7)
\]

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

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

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, \( \lambda \), to lessen the effect.

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

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

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)\(^2\). 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.

---

\(^2\)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 MALLET. 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 toolkit. 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 \( q^{img} \).

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). 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

<table>
<thead>
<tr>
<th>ROUGE-2</th>
<th>Average</th>
<th>95% Confidence int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL (EXTRACTION)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>.06114</td>
<td>(.05690 - .06554)</td>
</tr>
<tr>
<td>R</td>
<td>.02499</td>
<td>(.02325 - .02686)</td>
</tr>
<tr>
<td>F</td>
<td>.03360</td>
<td>(.03133 - .03600)</td>
</tr>
<tr>
<td>GIST (EXTRACTION)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>.10894</td>
<td>(.09934 - .11921)</td>
</tr>
<tr>
<td>R</td>
<td>.05474</td>
<td>(.04926 - .06045)</td>
</tr>
<tr>
<td>F</td>
<td>.06863</td>
<td>(.06207 - .07534)</td>
</tr>
<tr>
<td>LM-ONLY (COMPRESSION)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>.13782</td>
<td>(.12602 - .14864)</td>
</tr>
<tr>
<td>R</td>
<td>.02437</td>
<td>(.02193 - .02700)</td>
</tr>
<tr>
<td>F</td>
<td>.03864</td>
<td>(.03512 - .04229)</td>
</tr>
<tr>
<td>SYSTEM (COMPRESSION)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>.16752</td>
<td>(.15679 - .17882)</td>
</tr>
<tr>
<td>R</td>
<td>.05060</td>
<td>(.04675 - .05524)</td>
</tr>
<tr>
<td>F</td>
<td>.07204</td>
<td>(.06685 - .07802)</td>
</tr>
</tbody>
</table>

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.

Technically their model is conditioned on the source image, in order to address alignment issues which are not applicable in our setup.
Table 5: BLEU@1 scores of generated captions against human authored captions. Our model (bolded) has the highest BLEU@1 score with significance.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL (Extraction)</td>
<td>0.2098</td>
</tr>
<tr>
<td>GIST (Extraction)</td>
<td>0.4259</td>
</tr>
<tr>
<td>LM-ONLY (Compression)</td>
<td>0.4780</td>
</tr>
<tr>
<td>SYSTEM (Compression)</td>
<td>0.4841</td>
</tr>
</tbody>
</table>

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.

Table 6: Example output from our full system. Red underlined words indicate the words which are deleted by our compression model.
Table 7: Human evaluation results.

<table>
<thead>
<tr>
<th>Compression improves accuracy</th>
<th>Yes</th>
<th>No</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>63.2%</td>
<td>36.8%</td>
<td>42.6%</td>
<td>57.4%</td>
</tr>
<tr>
<td>LM-Only</td>
<td>73.1%</td>
<td>26.9%</td>
<td>82.2%</td>
<td>17.8%</td>
</tr>
</tbody>
</table>

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

### References


