Annotation of Online Shopping Images without Labeled Training Examples

Rebecca Mason and **Eugene Charniak**
Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University, Providence, RI 02912
{rebecca,ec}@cs.brown.edu

Abstract

We are interested in the task of image annotation using noisy natural text as training data. An image and its caption convey different information, but are generated by the same underlying concepts. In this paper, we learn latent mixtures of topics that generate image and product descriptions on shopping websites by adapting a topic model for multilingual data (Mimno et al., 2009). We use the trained model to annotate test images without corresponding text. We capture visual properties such as color, texture, shape, and orientation by computing low-level image features, and measure the contribution of each type of visual feature towards the accuracy of the model. Our model significantly outperforms both a competitive baseline and a previous topic model-based system.

1 Introduction

Image annotation is a classic problem in Computer Vision. Given a query image, the task is to generate a set of textual labels that describe the visual content. The typical approach to these problems is to use supervised models, which require large numbers of hand-annotated examples for each of the labels. However, the amount of information available on the web continues to grow, the task of organizing and describing visual data becomes increasingly complex. For example, a shopping website might arrange products into broad categories such as "shoes" and "handbags" with each category containing tens of thousands of products that are difficult for users to search and navigate. It is often infeasible to discover all of the attributes within those categories that are relevant to users and create labeled training examples for each of them.

Instead, we approach this problem by discovering visual attributes from noisy natural language captions. That is, given a collection of images and captions found on the web, we learn a model of visual and textual features. Then given a query image with no text, we can generate likely descriptive words. This is a difficult task because image captions on the web are often noisy and incomplete: some captions might not describe a particular visual feature, might use a synonym for that feature, or might describe information that is not visual in the image at all.

A secondary motivation for this work is to use the image annotations as a component in language generation systems such as for automatic image captioning. We point to examples of previous work such as Feng and Lapata (2010a) where image annotations generated from a topic model are used to help generate full sentences to describe images. Much of the current research in image captioning is limited by the current technology for object recognition in Computer Vision. For example, SBU-Flickr dataset (Ordonez et al., 2011) with 1 million images and captions, is considered to be general-domain but is actually built by querying Flickr using a pre-defined term list related to visual attributes that there are trained recognition systems for. While these systems can accurately generate descriptions for common visual objects and attributes, they are not as well-suited for describing the "long-tail" of visual attributes which appear in many domain-specific

Proceedings of the NAACL HLT Workshop on Vision and Language (WVL '13), pages 1–9,
Atlanta, Georgia, 14 June 2013. ©2013 Association for Computational Linguistics

		
Two adjustable buckle straps top a classic rubber rain boot grounded by a thick lug sole for excellent wet-weather traction.	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	Carlo Fellini - Evening clutch beaded on a wave pattern

Table 1: Examples of data from the Attribute Discovery Dataset (Berg et al., 2010). The images are fairly clean and uniform, while captions have more noise and variation.

datasets.

In this paper, we model image and text features from the training data using a generative model. We adapt the Polylingual topic model from Mimno et al. (2009) to train on multi-modal data, and then use the trained model to generate annotations for test images. We evaluate our model on two categories of shopping images using a variety of types of computed image features. For image annotation we outperform both a difficult baseline and previous work.

2 Related Work

We use the polylingual topic model from Mimno et al. (2009), which was developed to model multilingual corpora that are *topically comparable* between languages – the documents are not direct translations, but they cover the same ideas. For example, English and Finnish Wikipedia pages about skiing are roughly similar, but the subject is covered more thoroughly in Finnish. Therefore, the number of tokens assigned to the Finnish topic for skiing is much higher than it is in the English. While Mimno et al. (2009) show that the model is effective in tasks such as modeling topically comparable documents across languages, our work is the first to show that this model can be used to model data of different *modalities*. Another quality of the polylingual topic model is that words in different languages do not directly correspond with each other. This is a feature of other multi-lingual topic models but would not work for multi-modal data because a textual word can carry more meaning by itself than an image feature can.

Countless approaches have been proposed for the use of topic models in image annotation, but the vast majority of these approaches consider the text modality merely as labels for the image modality. The most highly cited of these is the Correspondence LDA (corr-LDA) model of Blei et al. (2003), where topics are learned using the image modality alone, and each textual word must be generated by a specific region in the image. However, more recent work has started to recognize the textual modality as a source of information in its own right. Jia et al. (2011) present a model that allows different information to be emphasized in each modality, but it requires very clean text; they do not use documents with captions that cannot be easily parsed or processed. Then, they stem all words, and disregard sparse word tokens. This works when working with sources such as Wikipedia, where text captions are highly edited and consistently formatted. In comparision, our work can be trained on corpra where the text has poor or inconsistant quality. Additionally, their work was for the task of image retrieval from a text query, while we are generating text annotations for a query image.

Our work is most similar to the MixLDA model of Feng and Lapata (2010b), except MixLDA mod-

els images and their related text as a single bag-of-features, with visual and textual features coming from the same vocabulary. This means that some topics should have a greater proportion of features from one of modalities, if there is an idea that is better expressed in one over the other. Their model was developed for finding descriptive words given both an image and a news article, and can also be used on large and noisy amounts of data, so we compare MixLDA against our model in the experiments.

Although we use the Attribute Discovery Dataset of Berg et al. (2010), their work is different from ours in both problem formulation and the types of attributes discovered. Their primary interest is to characterize attributes according to how they are visually represented: global or local; color, texture, or shape. Their work does not address the task of predicting attributes for unseen images. Additionally, they do not work with individual descriptive words, but cluster them using mutual information of visual attributes, creating a smaller number of "visual synsets". For example, one of their visual synsets for images and descriptions of womens handbags is {*mesh, interior, metal*} and another is {*silver, metallic*}. In comparison, in the topic model the same word can be generated by more than one topic.

Liu et al. (2010) examine the use of a variety of image features in a Bayesian model in order to measure which are the best for classifying diverse materials such as stone, glass, and plastic. They found that the image features they used for shape and color were better indicators of the material of an object than texture features, and their best combined model did not include texture as a feature at all. We are also interested in finding out whether our performance on generating descriptive words is affected by different types of image features.

3 Dataset

We use the Attribute Discovery Dataset from Berg et al. (2010).[1] The dataset consists of pairs of images and captions taken from the shopping website `like.com`. The data has four categories: women's shoes, handbags, earrings, and neckties. We run our model on two categories, shoes and handbags, due to their larger sizes – 14764 and 9145 image-caption pairs respectively – and diversity of features. This is a reasonable amount of data in the shopping images domain; more than half of the number of comparable products sold on large retail websites such as `Zappos.com` or `Amazon.com`.

Compared to general datasets such as Pascal Sentences, the images in the Attribute Discovery Dataset are more uniform. All image files are 280x280 pixel JPEGs, and images of products are typically taken from similar angles against a white or a light-colored solid background. Only rarely do the images have noisy backgrounds, such as a person wearing the item, or the same item displayed in multiple colors in one image. However, this does not necessarily make our task much easier, since the visual attributes we wish to learn are not pre-defined as they are in a general-domain dataset. And the lack of hand-annotated data means no *negative examples* of when an attribute is not present, which are typically used to train visual classifiers.

Furthermore, the captions are extremely noisy in this dataset. Compared to the 20 object types in the Pascal Sentences dataset, or about one hundred in COREL, here there are thousands of words that can be used to describe features in the images, including synonyms, multiple stems of words, and misspellings. In addition to explicit visual descriptions of the products, the captions describe "less visual" features such as details about the construction of the item, during which season or activity it would be appropriate to wear, or feelings that could be evoked by looking at the item. These features are difficult to represent as specific visual attributes, but can be identified visually by domain-experts. Captions can also include information that is non-visual such as sizing and shipping information, or whether the item is on sale.

The captions can be either full English sentences, a list of features, or sometimes just a few words. Longer captions in the dataset are truncated to 250 characters in length.

From our own obervations, we estimate about 10% of the captions in the shoes dataset contain few or no descriptive words. At least 3.7% of the shoes captions are entirely Javascript code, have significant portions of code, or very long URLs. Another 5-6% either contain no information besides

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

sizing or shipping information, only the brand name or model number of the shoe, or the caption is so short that there are only one or two descriptive words that could be used in our model. In the womens' shoes category, we take some simple steps to remove URLs and code to avoid learning accidental correlation with legitimate features.[2] However, we still use all image and caption pairs in the training set, including those which end up having empty captions, since they are still useful for learning topics for visual features. For the handbags captions, we did not try to remove code or long URLs since it seemed to be less of a problem in that category.

4 Feature Representation

4.1 Text Features

The bag-of-words model is used for text. We use Mxterminator (Reynar and Ratnaparkhi, 1997) to split sentences in the captions (in many instances, nothing is done in this step becuase there are no full sentences in the caption), Stanford POS Tagger(Toutanova et al., 2003) to tag words, then include adjectives, adverbs, verbs, and nouns in the topic model (except for proper nouns and common background English words from a stoplist). However, these tags are really more a rough estimate of parts of speech due to the number of incomplete sentences and phrases, and the fact that many of the words used to describe styles or attributes of clothing have different meanings in colloqueal English.[3]

All tokens are converted to lower case, but there is no stemming or lemmatization. After preprocessing, the size of the shoes text vocabulary is 9578 words, with an average of 16.33 descriptive words per image, while the bags have a text vocabulary of 6309 word types with 15.41 descriptive words per image on average.

4.2 Visual Features

The bag-of-features model is used for visual features as well. Most of these features are standard in computer vision research, and are also used in work we cited in Section 2.

[2]Tokens removed: URLs, all tokens that end in ".sh", and a few tokens obviously related to Javascript eg *script, src, typeof, var*.

[3]Some examples of domain-specific words used in shopping image descriptions: `www.zappos.com/glossary`

Shape: A SIFT descriptor describes which way edges are oriented at a certain point in an image (Lowe, 1999). It was develped to recognize the same object under different scales and rotations. However, it is also commonly used for recognizing more generalized types or features of objects. We use the VLFeat open source library (Vedaldi and Fulkerson, 2008) to compute SIFT features at points of interest and to cluster the SIFT features into discrete "visual terms" using the k-means algorithm. There are 750 visual terms for SIFT features.

Color: We use two representations for color, RGB (red, green, blue) and HSV (hue, saturation, value). 25 pixels are sampled from the center 100x100 pixels of the image (to avoid sampling from the background of the image). Those pixel values are also clustered to visual terms using k-means, with 100 visual terms each.

Texture: Images are convolved with Gabor filters at multiple orientations and scales, sampled at random locations, then clustered to form *texton* features for texture (Leung and Malik, 2001). We convert all images to grayscale, then sample 25 locations from the center of the image, and cluster to 100 visual terms. We also have a color texton feature, where we sample and cluster textons separately for the red, green, and blue color channels.

Reflectance/Curvature:[4] We use three types of related features for gradients and curvature. The first is a bag-of-HOG (histogram of gradients) feature set (Dalal and Triggs, 2005) computed over a regular grid on the image to measure changes in intensity.[5] The most significant of those features (as determined by L^2 norm) are selected for each image, and like previous features are clustered into visual terms using k-means. The second two types are derivatives of HOG which include information about the amount of curvature at each orientation of the HOG descriptor.[6]

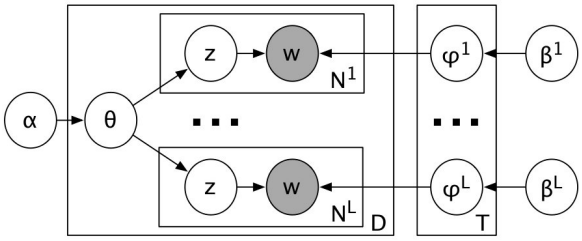

Figure 1: Polylingual topic model (Mimno et al., 2009)

5 Model

We model textual and visual features using the polylingual topic model by Mimno et al. (2009). In this section, we describe how the generative process and inference of this model is adapted to topically comparable multi-modal data.

Figure 1 shows the original polylingual topic model. We model multi-modal data using two "languages": *txt* for the bag-of-words captions, and *img* for the combined visual terms. The generative process is defined for an image and caption pair, $w =< w^{img}, w^{txt} >$:

$$\theta \sim Dir(\theta, \alpha m)$$

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

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

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

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

First, a topic distribution for w is drawn from an asymmetric Dirichlet prior with concentration parameter α and base measure m. Then a latent topic assignment is drawn for each word token in w^{txt}, and each discrete image feature in w^{img}. Once the topic assignments are sampled, the observed tokens are sampled according to their probability in the modality-specific topics $\Phi^{img} = \{\phi_1^{img}, ..., \phi_T^{img}\}$ and $\Phi^{txt} = \{\phi_1^{txt}, ..., \phi_T^{txt}\}$.

[4]Note: These features are implemented using code from (Felzenszwalb et al.,).

[5]There is significant overlap between these features, although the benefits of overlap are lost due to the bag-of-features model.

[6]Personal correspondance, work in progress.

To find the most probable descriptive words for an unseen image, the first step is to estimate the topic distribution that generated the image. Gibbs sampling is used to sample topic assignments for visual terms in the test image d^{img}:

$$P(z_n = t|d^{img}, z_{\backslash n}, \Phi^{img}, \alpha m)$$

$$\propto \phi^{img}_{d_n^{img}|t} \frac{(N_t)_{\backslash n} + \alpha m_t}{\sum_t N_t - 1 + \alpha}$$

Assuming that the descriptive words are independent, the probability of text word w_i given d^{img} is:

$$P(w_i|d^{img}) = \sum_t P(w_i|z_t^{txt}) P(z_t|d^{img})$$

summing over all topics $t \in T$.

For training the model, we used the Polylingual topic model implementation from the Mallet toolkit (McCallum, 2002) (with some small modifcations to use it for generation). We use 1000 iterations for inference, with hyperparameter optimization every 10 iterations. In both shoes and bags categories, the number of topics is 200, which was minimally tuned by hand on the shoes data.

6 Experimental Setup and Evaluation

We first run our model on the larger category, shoes. For both systems and baselines, we find the 10, 15, and 20 most likely words for the test images. We evaluate by computing precision and recall against descriptive words from the held-out captions for those images.[7] We compute macro-averages of these scores because there is a lot of variation between the sizes of the captions in the dataset. The split between training and test instances is 80/20%.

We also evaluate the contributions of different types of image features. We evaluate the model for each image feature individually (along with the text features), as well as combinations of image features.

We compare against the MixLDA system and a strong baseline. We choose MixLDA because it is relatively easy to re-implement and because it

[7]We find descriptive words for test instances in the exact same way we did for training instances in Section 4.1. Instances where we did not find any useable descriptive words did not count towards the evaluation.

	10 words			15 words			20 words		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baselines									
MixLDA	21.02	13.80	16.66	17.41	17.15	17.13	14.88	19.53	16.89
Corpus frequency	21.03	13.73	16.61	17.51	17.14	17.32	15.41	20.12	17.45
Single Attribute									
SIFT	27.00	16.30	20.34	22.84	20.65	21.69	20.09	24.22	21.96
Grayscale Texture	21.26	13.88	16.80	18.25	17.87	18.06	15.71	20.52	17.80
RGB Texture	24.77	14.93	18.63	21.01	18.99	19.95	18.49	22.29	20.21
HSV Color	22.17	13.35	16.67	18.59	16.79	17.65	16.48	19.85	18.01
RGB Color	23.21	13.98	17.45	19.78	17.88	18.78	17.53	21.12	19.15
HOG	26.33	15.87	19.80	22.36	20.21	21.23	19.60	23.62	21.42
TriHOG	24.60	14.82	18.50	20.64	18.66	19.60	18.14	21.87	19.83
TriHOG-Polar	26.03	15.69	19.58	22.06	19.94	20.95	19.32	23.29	21.12
Combined Models									
All-Color	24.22	14.60	18.22	20.62	18.65	19.59	18.11	21.83	19.80
All-Texture	25.50	15.41	19.24	21.63	19.55	20.53	18.88	22.75	20.64
All-HOG	27.36	16.50	20.58	23.31	21.07	22.14	20.40	24.58	22.30
Combine All	29.31	17.70	22.04	24.88	22.49	23.63	21.71	26.16	23.73
SIFT+RGB Texture+HOG	28.62	17.25	21.52	24.35	22.01	23.12	21.20	25.55	23.17

Table 2: Results of evaluation in the women's shoes cateogory (top 10-20 words).

has previously outperformed other image annotation systems when trained on natural language captions. Because the MixLDA model originally only used SIFT features, we compare it against the SIFT-only version of our model, with each system using the same computed image and text features. We re-implement the MixLDA system mostly as it is described in Feng and Lapata (2010b), with a few changes to make it more comparable to our model: Obviously in our version of MixLDA the test instances are only the unseen image as there is no other surrounding text. The number of topics is 200 (the original MixLDA had more but that did not seem to help here), and the α and β hyperparameters are optimized every 10 iterations.[8]

We also compare our model against corpus frequency of words in the training set. Although this may seem like a trivial baseline, previous work on image annotation from both computer vision (Müller et al., 2002; Monay and Gatica-Perez, 2003; Barnard et al., 2003) and natural language processing (Mason and Charniak, 2012) has shown that a large portion of the keyword probability mass can often be accounted for by a very small number of words, allowing systems to game better-looking results by simply guessing the frequency distribution of the text vocabulary. We find this to be especially true in the domain-specific case, where common terms (eg *shoe, sole, heel, upper*) are used in almost every caption, and in some captions account for most words used (such as the second example in Table 1). While domain-frequent words are also needed for generating new captions, we don't want them to account for all of the words our system generates. Of course, a human evaluation would be another possible way of addressing this issue, but it would be difficult and expensive to find enough people who have sufficient knowledge of womens' clothing and would be able to accurately say whether the generated words are appropriate or not (words such as *hobo, PU, stacked, upper,* and *vamp*). Also, although the gold image captions are noisy, the num-

[8]We used the Mallet toolkit's Parallel LDA sampler for inference, while a variational approach is used in the original. However, we do not believe this would change the outcome of this experiment. We also tried MixLDA without hyperparameter optimization but we do not show those results as they are significantly worse.

		
sole upper detail heel print fun fabric patent uppers soles high shoe rounded leather rubber lining elastic animal toe feet	style upper heel leather strap sandal lining toe dress satin shoe comfort ankle sole adjustable outsole platform stiletto rhinestone sandals	bag, leather, zip, pocket, hardware, features, shoulder, flap, main, cell, perfect, length, drop, zipper, closure, bold, phone, evening, holds, hobo
This high heel platform shoe has a patent leather upper with an ornamenting bow at the toe, a leather lining, a rounded toe, and a rubber bottom.Available Colors: Black Patent, Cheetah Print PU.	Create a timeless look with these Andie dress sandals from Coloriffics. Dyeable white satin matte satin or metallic satin upper in a two-piece dress-sandal style with an open round toe crossing pleated vamp straps with a dazzling rhinestone clasp and a wraparound heel strap with an adjustable buckle closure.	Treesje Dakota Shoulder Bag Black Shine - Designer Handbags

Table 3: Example results for unseen images. Both the top words generated by our model and the original held-out captions for the images are shown. (Note: In the third example, "hobo" is actually the term that is used to describe that shape of handbag.)

ber of test documents is very large so we can find significance on precision and recall using bootstrap resampling.

We also ran the baseline system and our system on the handbags category of the dataset. We did not modify the system in any way when using the bags dataset, just gave it different file for input.

7 Results and Discussion

The results of our evaluations are in Table 2. As we expected, the corpus frequency baseline does very well. It is comparable to MixLDA for 10 and 15 words, and significantly better than MixLDA for over 20 words. However, the Polylingual topic model using only SIFT features and text is much better than both. The trained MixLDA model has topics with both image and text features, so when estimating topics given only an image, it estimates that it was generated by topics that have a high proportion of image features. Though it also estimates some topics that have a mix of visual and text features, being able to generate good text descriptions from those topics, the topics that have less text features will be mainly determined by the smoothing parameter – the uniform distribution, worse than guessing the corpus distribution.

Out of the single attribute models, all except three of the single feaure models were significantly higher than corpus frequency on both precision and recall at 10, 15, and 20 words. The exceptions are the two color features and grayscale texture. For grayscale texture, we had expected it would correlate well with the material of the shoe; but either the low resolution of the images makes it difficult to distinguish materials by their texture, or materials don't correlate with the "less visual" features as much as we expected. Interestingly, since the material of an item tends to correlate strongly with other attributes such as shape and color, so our model still generates correct descriptive words for material in many cases.

While neither color nor texture were useful fea-

	10 words			15 words			20 words		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Corpus frequency	17.58	13.19	11.76	13.19	12.70	12.94	11.76	15.10	13.22
Combined Model	24.41	15.67	19.09	21.01	20.22	20.61	18.76	24.09	21.10

Table 4: Results of evaluation in the handbags cateogory (top 10-20 words).

tures on their own, RGB Texture did very well as a single attribute, and was within signficance of both the combined color and combined texture models. This may be related to the fact that RGB Textons have a larger number of visual terms than those other features, 100 for each of the three color channels. Unlike material, we observed that the color of an object is often not mentioned in the human-written caption (as seen in the examples in Table 1), or several colors are described in the caption where only one is seen in the image (seen in some of the examples in Table 3). We also observed that our system generates very few color words.

The gradient and shape-based features have the best single-attribute performance by far. Both SIFT and HOG capture shape at local points, but while SIFT features are invariant to differences in position or scale, HOG features are more sensitive to the way the item is oriented in the image. Although the curvature features TriHOG and TriHog-Polar are nearly as good as HOG on their own, combining the three HOG features does not significantly improve performance of the model over HOG alone.

Not all of the single-attribute models performed as well as others, but there was no case where removing one of the features improved the performance of the combined model. The fewest number of image attributes that our model could use and still get within significance to the full combined model is three – SIFT, RGB Texture, and HOG. However, we found that each image attribute does slightly improve, the model, even if not by a significant amount.

Our results on the handbags category of the dataset are shown in Table 4. Although our scores are not as high as they were in the shoes category, the scores of the corpus frequency baseline are not as high either, and our model does about as well over the baseline in each category. But is worth reiterating that we were able to run our system on both the bags and shoes shopping categories with absolutely no modifications or tuning of parameters.

8 Conclusion and Future Work

In conclusion, we have shown that the polylingual topic model works well for modeling topically comparable images and related text, and obtain competitive results for the image annotation task. Our model is trained on noisy image captions from the web, rather than hand-labeled data.

For future work, we would like to further adapt the polylingual topic model for multi-modal data by allowing some topics to be generated only by one modality or the other. We are also interested in characterizing the image annotations in order to generate a single most likely annotation for different types of features such as texture or color. Finally, we are interested in extending this model to use with other domains of data. For natural images, we could use image segmentation algorithms to separate the object of interest from the background of the image, or we could use scene classifcation to cluster the training images by their background scene and train seprate models for each.

References

K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D.M. Blei, and M.I. Jordan. 2003. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135.

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

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.

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

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.

Yangqing Jia, M. Salzmann, and T. Darrell. 2011. Learning cross-modality similarity for multinomial data. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2407 –2414, nov.

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.

C. Liu, L. Sharan, E.H. Adelson, and R. Rosenholtz. 2010. Exploring features in a bayesian framework for material recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 239–246. IEEE.

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.

R. Mason and E. Charniak. 2012. Apples to oranges: Evaluating image annotations from natural language processing systems. NAACL.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

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.

Florent Monay and Daniel Gatica-Perez. 2003. On image auto-annotation with latent space models. In *Proceedings of the eleventh ACM international conference on Multimedia*, Multimedia '03, pages 275–278, New York, NY, USA. ACM.

Henning Müller, Stéphane Marchand-Maillet, and Thierry Pun. 2002. The truth about corel - evaluation in image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, CIVR '02, pages 38–49, London, UK, UK. Springer-Verlag.

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

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *In Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 16–19.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *In Proceedings of HLT-NAACL 2003*, pages 252–259.

A. Vedaldi and B. Fulkerson. 2008. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`.