

Nonparametric Method for Data-driven 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 nonparametric density estimation technique for image caption generation. Data-driven matching methods have shown to be effective for a variety of complex problems in Computer Vision. These methods reduce an inference problem for an unknown image to finding an existing labeled image which is semantically similar. However, related approaches for image caption generation (Ordonez et al., 2011; Kuznetsova et al., 2012) are hampered by noisy estimations of visual content and poor alignment between images and human-written captions. Our work addresses this challenge by estimating a word frequency representation of the visual content of a query image. This allows us to cast caption generation as an extractive summarization problem. Our model strongly outperforms two state-of-the-art caption extraction systems according to human judgments of caption relevance.

1 Introduction

Automatic image captioning is a much studied topic in both the Natural Language Processing (NLP) and Computer Vision (CV) areas of research. The task is to identify the visual content of the input image, and to output a relevant natural language caption.

Much prior work treats image captioning as a retrieval problem (see Section 2). These approaches use CV algorithms to retrieve similar images from a large database of captioned images, and then transfer text from the captions of those images to the query image. This is a challenging problem for two main reasons. First, visual similarity measures do not perform reliably and do not

Query Image: Captioned Images:



- 1.) 3 month old baby girl with blue eyes in her crib
- 2.) A photo from the Ismail's **portrait** shoot
- 3.) A **portrait** of a man, in **black** and **white**
- 4.) **Portrait** in **black** and **white** with the red rose
- 5.) I apparently had this saved in **black** and **white** as well
- 6.) **Portrait** in **black** and **white**

Table 1: Example of a query image from the SBU-Flickr dataset (Ordonez et al., 2011), along with scene-based estimates of visually similar images. Our system models visual content using words that are frequent in these captions (highlighted) and extracts a single output caption.

capture all of the relevant details which humans might describe. Second, image captions collected from the web often contain contextual or background information which is not visually relevant to the image being described.

In this paper, we propose a system for transfer-based image captioning which is designed to address these challenges. Instead of selecting an output caption according to a single noisy estimate of visual similarity, our system uses a word frequency model to find a smoothed estimate of visual content across multiple captions, as Table 1 illustrates. It then generates a description of the query image by extracting the caption which best represents the mutually shared content.

The contributions of this paper are as follows:

1. Our caption generation system effectively leverages information from the massive amounts of human-written image captions on the internet. In particular, it exhibits strong performance on the SBU-Flickr dataset (Ordonez et al., 2011), a noisy corpus of one million captioned images collected from the web. We achieve a remarkable 34% improvement in human relevance scores over a recent state-of-the-art image captioning system (Kuznetsova et al., 2012), and 48% improvement over a scene-based retrieval system (Patterson et al., 2014) using the same computed image features.

2. Our approach uses simple models which can be easily reproduced by both CV and NLP researchers. We provide resources to enable comparison against future systems.¹

2 Image Captioning by Transfer

The IM2TEXT model by Ordonez et al. (2011) presents the first web-scale approach to image caption generation. IM2TEXT retrieves the image which is the closest visual match to the query image, and transfers its description to the query image. The COLLECTIVE model by Kuznetsova et al. (2012) is a related approach which uses trained CV recognition systems to detect a variety of visual entities in the query image. A separate description is retrieved for each visual entity, which are then fused into a single output caption. Like IM2TEXT, their approach uses visual similarity as a proxy for textual relevance.

Other related work models the text more directly, but is more restrictive about the source and quality of the human-written training data. Farhadi et al. (2010) and Hodosh et al. (2013) learn joint representations for images and captions, but can only be trained on data with very strong alignment between images and descriptions (i.e. captions written by Mechanical Turkers). Another line of related work (Fan et al., 2010; Aker and Gaizauskas, 2010; Feng and Lapata, 2010) generates captions by extracting sentences from documents which are related to the query image. These approaches are tailored toward specific domains, such as travel and news, where images tend to appear with corresponding text.

¹See http://bllip.cs.brown.edu/download/captioning_resources.zip or ACL Anthology.

3 Dataset

In this paper, we use the SBU-Flickr dataset². Ordonez et al. (2011) query Flickr.com using a huge number of words which describe visual entities, in order to build a corpus of one million images with captions which refer to image content. However, further analysis by Hodosh et al. (2013) shows that many captions in SBU-Flickr (~67%) describe information that cannot be obtained from the image itself, while a substantial fraction (~23%) contain almost no visually relevant information. Nevertheless, this dataset is the only web-scale collection of captioned images, and has enabled notable research in both CV and NLP.³

4 Our Approach

4.1 Overview

For a query image I_q , our task is to generate a relevant description by selecting a single caption from \mathcal{C} , a large dataset of images with human-written captions. In this section, we first define the feature space for visual similarity, then formulate a density estimation problem with the aim of modeling the words which are used to describe visually similar images to I_q . We also explore methods for extractive caption generation.

4.2 Measuring Visual Similarity

Data-driven matching methods have shown to be very effective for a variety of challenging problems (Hays and Efros, 2008; Makadia et al., 2008; Tighe and Lazebnik, 2010). Typically these methods compute global (scene-based) descriptors rather than object and entity detections. Scene-based techniques in CV are generally more robust, and can be computed more efficiently on large datasets.

The basic IM2TEXT model uses an equally weighted average of GIST (Oliva and Torralba, 2001) and TinyImage (Torralba et al., 2008) features, which coarsely localize low-level features in scenes. The output is a multi-dimensional image space where semantically similar scenes (e.g. streets, beaches, highways) are projected near each other.

²<http://tamaraberg.com/CLSP11/>

³In particular, papers stemming from the 2011 JHU-CLSP Summer Workshop (Berg et al., 2012; Dodge et al., 2012; Mitchell et al., 2012) and more recently, the best paper award winner at ICCV (Ordonez et al., 2013).

Patterson and Hays (2012) present “scene attribute” representations which are characterized using low-level perceptual attributes as used by GIST (e.g. openness, ruggedness, naturalness), as well as high-level attributes informed by open-ended crowd-sourced image descriptions (e.g., indoor lighting, running water, places for learning). Follow-up work (Patterson et al., 2014) shows that their attributes provide improved matching for image captioning over IM2TEXT baseline. We use their publicly available⁴ scene attributes for our experiments. Training set and query images are represented using 102-dimensional real-valued vectors, and similarity between images is measured using the Euclidean distance.

4.3 Density Estimation

As shown in Bishop (2006), probability density estimates at a particular point can be obtained by considering points in the training data within some local neighborhood. In our case, we define some region \mathcal{R} in the image space which contains I_q . The probability mass of that space is

$$P = \int_{\mathcal{R}} p(I_q) dI_q \quad (1)$$

and if we assume that \mathcal{R} is small enough such that $p(I_q)$ is roughly constant in \mathcal{R} , we can approximate

$$p(I_q) \approx \frac{k^{img}}{n^{img} V^{img}} \quad (2)$$

where k^{img} is the number of images within \mathcal{R} in the training data, n^{img} is the total number of images in the training data, and V^{img} is the volume of \mathcal{R} . In this paper, we fix k^{img} to a constant value, so that V^{img} is determined by the training data around the query image.⁵

At this point, we extend the density estimation technique in order to estimate a smoothed model of descriptive text. Let us begin by considering $p(w|I_q)$, the conditional probability of the word⁶ w given I_q . This can be described using a

⁴https://github.com/genp/sun_attributes

⁵As an alternate approach, one could fix the value of V^{img} and determine k^{img} from the number of points in \mathcal{R} , giving rise to the kernel density approach (a.k.a. *Parzen windows*). However we believe the KNN approach is more appropriate here, because the number of samples is nearly 10000 times greater than the number of dimensions in the image representation.

⁶Here, we use word to refer to non-function words, and assume all function words have been removed from the captions.

Bayesian model:

$$p(w|I_q) = \frac{p(I_q|w)p(w)}{p(I_q)} \quad (3)$$

The prior for w is simply its unigram frequency in \mathcal{C} , where n_w^{txt} and n^{txt} are word token counts:

$$p(w) = \frac{n_w^{txt}}{n^{txt}} \quad (4)$$

Note that n^{txt} is not the same as n^{img} because a single captioned image can have multiple words in its caption. Likewise, the conditional density

$$p(I_q|w) \approx \frac{k_w^{txt}}{n_w^{txt} V^{img}} \quad (5)$$

considers instances of observed words within \mathcal{R} , although the volume of \mathcal{R} is still defined by the image space. k_w^{txt} is the number of times w is used within \mathcal{R} while n_w^{txt} is the total number of times w is observed in \mathcal{C} .

Combining Equations 2, 4, and 5 and canceling out terms gives us the posterior probability:

$$p(w|I_q) = \frac{k_w^{txt}}{k^{img}} \cdot \frac{n^{img}}{n^{txt}} \quad (6)$$

If the number of words in each caption is independent of its image’s location in the image space, then $p(w|I_q)$ is approximately the observed unigram frequency for the captions inside \mathcal{R} .

4.4 Extractive Caption Generation

We compare two selection methods for extractive caption generation:

1. SumBasic SumBasic (Nenkova and Vanderwende, 2005) is a sentence selection algorithm for extractive multi-document summarization which exclusively maximizes the appearance of words which have high frequency in the original documents. Here, we adapt SumBasic to maximize the average value of $p(w|I_q)$ in a single extracted caption:

$$output = \arg \max_{c^{txt} \in \mathcal{R}} \sum_{w \in c^{txt}} \frac{1}{|c^{txt}|} p(w|I_q) \quad (7)$$

The candidate captions c^{txt} do not necessarily have to be observed in \mathcal{R} , but in practice we did not find increasing the number of candidate captions to be more effective than increasing the size of \mathcal{R} directly.

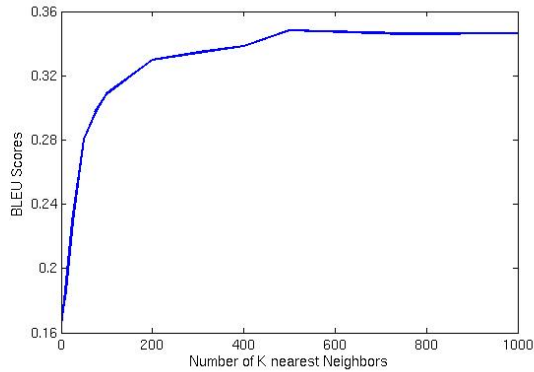


Figure 1: BLEU scores vs k for SumBasic extraction.

2. KL Divergence We also consider a KL Divergence selection method. This method outperforms the SumBasic selection method for extractive multi-document summarization (Haghighi and Vanderwende, 2009). It also generates the best extractive captions for Feng and Lapata (2010), who caption images by extracting text from a related news article. The KL Divergence method is

$$output = \arg \min_{c^{txt} \in \mathcal{R}} \sum_w p(w|I_q) \log \frac{p(w|I_q)}{p(w|c^{txt})} \quad (8)$$

5 Evaluation

5.1 Automatic Evaluation

Although BLEU (Papineni et al., 2002) scores are widely used for image caption evaluation, we find them to be poor indicators of the quality of our model. As shown in Figure 1, our system’s BLEU scores increase rapidly until about $k = 25$. Past this point we observe the density estimation seems to get washed out by oversmoothing, but the BLEU scores continue to improve until $k = 500$ but only because the generated captions become increasingly shorter. Furthermore, although we observe that our SumBasic extracted captions obtain consistently higher BLEU scores, our personal observations find KL Divergence captions to be better at balancing recall and precision. Nevertheless, BLEU scores are the accepted metric for recent work, and our KL Divergence captions with $k = 25$ still outperform all other previously published systems and baselines. We omit full results here due to space, but make our BLEU setup with captions for all systems and baselines available for documentary purposes.

System	Relevance
COLLECTIVE	2.38 ($\sigma = 1.45$)
SCENE ATTRIBUTES	2.15 ($\sigma = 1.45$)
SYSTEM	3.19 ($\sigma = 1.50$)
HUMAN	4.09 ($\sigma = 1.14$)

Table 2: Human evaluations of relevance: mean ratings and standard deviations. See Section 5.2.

5.2 Human Evaluation

We perform our human evaluation of caption relevance using a similar setup to that of Kuznetsova et al. (2012), who have humans rate the image captions on a 1-5 scale (5: perfect, 4: almost perfect, 3: 70-80% good, 2: 50-70% good, 1: totally bad). Evaluation is performed using Amazon Mechanical Turk. Evaluators are shown both the caption and the query image, and are specifically instructed to ignore errors in grammaticality and coherence.

We generate captions using our system with KL Divergence sentence selection and $k = 25$. We also evaluate the original HUMAN captions for the query image, as well as generated captions from two recently published caption transfer systems. First, we consider the SCENE ATTRIBUTES system (Patterson et al., 2014), which represents both the best scene-based transfer model and a $k = 1$ nearest-neighbor baseline for our system. We also compare against the COLLECTIVE system (Kuznetsova et al., 2012), which is the best object-based transfer model.

In order to facilitate comparison, we use the same test/train split that is used in the publicly available system output for the COLLECTIVE system⁷. However, we remove some query images which have contamination between the train and test set (this occurs when a photographer takes multiple shots of the same scene and gives all the images the exact same caption). We also note that their test set is selected based on images where their object detection systems had good performance, and may not be indicative of their performance on other query images.

Table 2 shows the results of our human study. Captions generated by our system have 48% improvement in relevance over the SCENE ATTRIBUTES system captions, and 34% improve-

⁷<http://www.cs.sunysb.edu/~pkuznetsova/generation/cogn/captions.html>


				
COLLECTIVE:	One of the birds seen in company of female and juvenile.	View of this woman sitting on the sidewalk in Mumbai by the stained glass. The boy walking by next to matching color walls in gov t building.	Found this mother bird feeding her babies in our maple tree on the phone.	Found in floating grass spotted alongside the scenic North Cascades Hwy near Ruby arm a black bear.
SCENE ATTRIBUTES:	This small bird is pretty much only found in the ancient Caledonian pine forests of the Scottish Highlands.	me and allison in front of the white house	The sand in this beach was black...I repeat BLACK SAND	Not the green one, but the almost ghost-like white one in front of it.
SYSTEM:	White bird found in park standing on brick wall	by the white house	pine tree covered in ice :)	Pink flower in garden w/ moth
HUMAN:	Some black head bird taken in bray head.	Us girls in front of the white house	Male cardinal in snowy tree knots	Black bear by the road between Ucluelet and Port Alberni, B.C., Canada

Table 3: Example query images and generated captions.

ment over the COLLECTIVE system captions. Although our system captions score lower than the human captions on average, there are some instances of our system captions being judged as more relevant than the human-written captions.

6 Discussion and Examples

Example captions are shown in Table 3. In many instances, scene-based image descriptors provide enough information to generate a complete description of the image, or at least a sufficiently good one. However, there are some kinds of images for which scene-based features alone are insufficient. For example, the last example describes the small pink flowers in the background, but misses the bear.

Image captioning is a relatively novel task for which the most compelling applications are probably not yet known. Much previous work in image captioning focuses on generating captions that concretely describe detected objects and entities (Kulkarni et al., 2011; Yang et al., 2011; Mitchell et al., 2012; Yu and Siskind, 2013). However, human-generated captions and annotations also describe perceptual features, contextual information, and other types of content. Additionally, our system is robust to instances where entity detection systems fail to perform. However, one could

consider combined approaches which incorporate more regional content structures. For example, previous work in nonparametric hierarchical topic modeling (Blei et al., 2010) and scene labeling (Liu et al., 2011) may provide avenues for further improvement of this model. Compression methods for removing visually irrelevant information (Kuznetsova et al., 2013) may also help increase the relevance of extracted captions. We leave these ideas for future work.

References

- Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, pages 1250–1258, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Alexander C Berg, Tamara L Berg, Hal Daume, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. 2012. Understanding and predicting importance in images. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3562–3569. IEEE.
- Christopher M Bishop. 2006. *Pattern recognition and machine learning*, volume 1. Springer New York.
- David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested chinese restaurant process

- and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30, February.
- Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C. Berg, and Tamara L. Berg. 2012. Detecting visual text. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Xin Fan, Ahmet Aker, Martin Tomko, Philip Smart, Mark Sanderson, and Robert Gaizauskas. 2010. Automatic image captioning from the web for gps photographs. In *Proceedings of the international conference on Multimedia information retrieval, MIR '10*, pages 445–448, New York, NY, USA. ACM.
- Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: generating sentences from images. In *Proceedings of the 11th European conference on Computer vision: Part IV, ECCV'10*, pages 15–29, Berlin, Heidelberg. Springer-Verlag.
- Yansong Feng and Mirella Lapata. 2010. 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.
- Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics.
- James Hays and Alexei A Efros. 2008. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.
- 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*.
- Ce Liu, Jenny Yuen, and Antonio Torralba. 2011. Nonparametric scene parsing via label transfer. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(12):2368–2382.
- Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. 2008. A new baseline for image annotation. In *Computer Vision–ECCV 2008*, pages 316–329. Springer.
- 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)*.
- Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization.
- 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*.
- Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. From large scale image categorization to entry-level categories. In *International Conference on Computer Vision*.
- 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.
- Genevieve Patterson and James Hays. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE.
- Genevieve Patterson, Chen Xu, Hang Su, and James Hays. 2014. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*.
- Joseph Tighe and Svetlana Lazebnik. 2010. Superparsing: scalable nonparametric image parsing with superpixels. In *Computer Vision–ECCV 2010*, pages 352–365. Springer.
- Antonio Torralba, Robert Fergus, and William T Freeman. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11):1958–1970.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yian-nis 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.