Im2Text: Describing Images Using 1 Million Captioned Photographs

Vicente Ordonez **Girish Kulkarni** **Tamara L Berg**

Stony Brook University
Stony Brook, NY 11794
{vordonezroma or tlberg}@cs.stonybrook.edu

Abstract

We develop and demonstrate automatic image description methods using a large captioned photo collection. One contribution is our technique for the automatic collection of this new dataset – performing a huge number of Flickr queries and then filtering the noisy results down to 1 million images with associated visually relevant captions. Such a collection allows us to approach the extremely challenging problem of description generation using relatively simple non-parametric methods and produces surprisingly effective results. We also develop methods incorporating many state of the art, but fairly noisy, estimates of image content to produce even more pleasing results. Finally we introduce a new objective performance measure for image captioning.

1 Introduction

Producing a relevant and accurate caption for an arbitrary image is an extremely challenging problem, perhaps nearly as difficult as the underlying general image understanding task. However, there are already many images with relevant associated descriptive text available in the noisy vastness of the web. The key is to find the right images and make use of them in the right way! In this paper, we present a method to effectively skim the top of the image understanding problem to caption photographs by collecting and utilizing the large body of images on the internet with associated visually descriptive text. We follow in the footsteps of past work on internet vision that has demonstrated that big data can often make big problems – e.g. image localization [13], retrieving photos with specific content [27], or image parsing [26] – much more bite size and amenable to very simple non-parametric matching methods. In our case, with a large captioned photo collection we can create an image description surprisingly well even with basic global image representations for retrieval and caption transfer. In addition, we show that it is possible to make use of large numbers of state of the art, but fairly noisy estimates of image content to produce more pleasing and relevant results.

People communicate through language, whether written or spoken. They often use this language to describe the visual world around them. Studying collections of existing natural image descriptions and how to compose descriptions for novel queries will help advance progress toward more complex human recognition goals, such as how to *tell the story behind an image*. These goals include determining what content people judge to be most important in images and what factors they use to construct natural language to describe imagery. For example, when given a picture like that on the top row, middle column of figure 1, the user describes the girl, the dog, and their location, but selectively chooses not to describe the surrounding foliage and hut.

This link between visual importance and descriptions leads naturally to the problem of text summarization in natural language processing (NLP). In text summarization, the goal is to select or generate a summary for a document. Some of the most common and effective methods proposed for summarization rely on extractive summarization [25, 22, 28, 19, 23]. where the most important or

Figure 1: **SBU Captioned Photo Dataset:** Photographs with user-associated captions from our web-scale captioned photo collection. We collect a large number of photos from Flickr and filter them to produce a data collection containing over 1 million well captioned pictures.

relevant sentence (or sentences) is selected from a document to serve as the document's summary. Often a variety of features related to document content [23], surface [25], events [19] or feature combinations [28] are used in the selection process to produce sentences that reflect the most significant concepts in the document.

In our photo captioning problem, we would like to generate a caption for a query picture that summarizes the salient image content. We do this by considering a large relevant document set constructed from related image captions and then use extractive methods to select the best caption(s) for the image. In this way we implicitly make use of human judgments of content importance during description generation, by directly transferring human made annotations from one image to another.

This paper presents two extractive approaches for image description generation. The first uses global image representations to select relevant captions (Sec 3). The second incorporates features derived from noisy estimates of image content (Sec 5). Of course, the first requirement for any extractive method is a document from which to extract. Therefore, to enable our approach we build a web-scale collection of images with associated descriptions (ie captions) to serve as our document for relevant caption extraction. A key factor to making such a collection effective is to filter it so that descriptions are likely to refer to visual content. Some small collections of captioned images have been created by hand in the past. The UIUC Pascal Sentence data set[1] contains 1k images each of which is associated with 5 human generated descriptions. The ImageClef[2] image retrieval challenge contains 10k images with associated human descriptions. However neither of these collections is large enough to facilitate reasonable image based matching necessary for our goals, as demonstrated by our experiments on captioning with varying collection size (Sec 3). In addition this is the first – to our knowledge – attempt to mine the internet for general captioned images on a web scale!

In summary, our contributions are:

- A large novel data set containing images from the web with associated captions written by people, filtered so that the descriptions are likely to refer to visual content.
- A description generation method that utilizes global image representations to retrieve and transfer captions from our data set to a query image.
- A description generation method that utilizes both global representations and direct estimates of image content (objects, actions, stuff, attributes, and scenes) to produce relevant image descriptions.

1.1 Related Work

Studying the association between words with pictures has been explored in a variety of tasks, including: labeling faces in news photographs with associated captions [2], finding a correspondence between keywords and image regions [1, 6], or for moving beyond objects to mid-level recognition elements such as attribute [16, 8, 17, 12].

Image description generation in particular has been studied in a few recent papers [9, 11, 15, 30]. Kulkarni et al [15] generate descriptions from scratch based on detected object, attribute, and prepositional relationships. This results in descriptions for images that are usually closely related to image content, but that are also often quite verbose and non-humanlike. Yao et al [30] look at the problem

[1] http://vision.cs.uiuc.edu/pascal-sentences/
[2] http://www.imageclef.org/2011

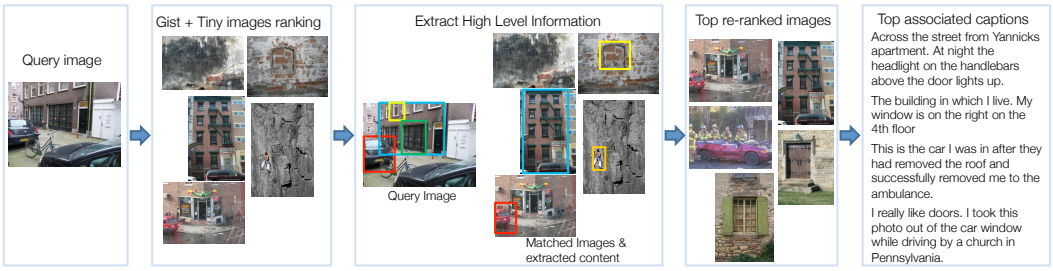

Figure 2: **System flow:** 1) Input query image, 2) Candidate matched images retrieved from our web-scale captioned collection using global image representations, 3) High level information is extracted about image content including objects, attributes, actions, people, stuff, scenes, and tfidf weighting, 4) Images are re-ranked by combining all content estimates, 5) Top 4 resulting captions.

of generating text using various hierarchical knowledge ontologies and with a human in the loop for image parsing (except in specialized circumstances). Feng and Lapata [11] generate captions for images using extractive and abstractive generation methods, but assume relevant documents are provided as input, whereas our generation method requires only an image as input.

A recent approach from Farhadi et al [9] is the most relevant to ours. In this work the authors produce image descriptions via a retrieval method, by translating both images and text descriptions to a shared meaning space represented by a single $< object, action, scene >$ tuple. A description for a query image is produced by retrieving whole image descriptions via this meaning space from a set of image descriptions (the UIUC Pascal Sentence data set). This results in descriptions that are very human – since they were written by humans – but which may not be relevant to the specific image content. This limited relevancy often occurs because of problems of sparsity, both in the data collection – 1000 images is too few to guarantee similar image matches – and in the representation – only a few categories for 3 types of image content are considered.

In contrast, we attack the caption generation problem for much more general images (images found via thousands of Flickr queries compared to 1000 images from Pascal) and a larger set of object categories (89 vs 20). In addition to extending the object category list considered, we also include a wider variety of image content aspects, including: non-part based stuff categories, attributes of objects, person specific action models, and a larger number of common scene classes. We also generate our descriptions via an extractive method with access to much larger and more general set of captioned photographs from the web (1 million vs 1 thousand).

2 Overview & Data Collection

Our captioning system proceeds as follows (see fig 2 for illustration): 1) a query image is input to the captioning system, 2) Candidate match images are retrieved from our web-scale collection of captioned photographs using global image descriptors, 3) High level information related to image content, e.g. objects, scenes, etc, is extracted, 4) Images in the match set are re-ranked based on image content, 5) The best caption(s) is returned for the query. Captions can also be generated after step 2 from descriptions associated with top globally matched images.

In the rest of the paper, we describe collecting a web-scale data set of captioned images from the internet (Sec 2.1), caption generation using a global representation (Sec 3), content estimation for various content types (Sec 4), and finally present an extension to our generation method that incorporates content estimates (Sec 5).

2.1 Building a Web-Scale Captioned Collection

One key contribution of our paper is a novel web-scale database of photographs with associated descriptive text. To enable effective captioning of novel images, this database must be good in two ways: 1) It must be large so that image based matches to a query are reasonably similar, 2) The captions associated with the data base photographs must be visually relevant so that transferring captions between pictures is useful. To achieve the first requirement we query Flickr using a huge number of pairs of query terms (objects, attributes, actions, stuff, and scenes). This produces a very large, but noisy initial set of photographs with associated text. To achieve our second requirement

Figure 3: **Size Matters:** Example matches to a query image for varying data set sizes.

we filter this set of photos so that the descriptions attached to a picture are relevant and visually descriptive. To encourage visual descriptiveness in our collection, we select only those images with descriptions of satisfactory length based on observed lengths in visual descriptions. We also enforce that retained descriptions contain at least 2 words belonging to our term lists and at least one prepositional word, e.g. "on", "under" which often indicate visible spatial relationships.

This results in a final collection of over 1 million images with associated text descriptions – the *SBU Captioned Photo Dataset*. These text descriptions generally function in a similar manner to image captions, and usually directly refer to some aspects of the visual image content (see fig 1 for examples). Hereafter, we will refer to this web based collection of captioned images as C.

Query Set: We randomly sample 500 images from our collection for evaluation of our generation methods (exs are shown in fig 1). As is usually the case with web photos, the photos in this set display a wide range of difficulty for visual recognition algorithms and captioning, from images that depict scenes (e.g. beaches), to images with a relatively simple depictions (e.g. a horse in a field), to images with much more complex depictions (e.g. a boy handing out food to a group of people).

3 Global Description Generation

Internet vision papers have demonstrated that if your data set is large enough, some very challenging problems can be attacked with very simple matching methods [13, 27, 26]. In this spirit, we harness the power of web photo collections in a non-parametric approach. Given a query image, I_q, our goal is to generate a relevant description. We achieve this by computing the global similarity of a query image to our large web-collection of captioned images, C. We find the closest matching image (or images) and simply transfer over the description from the matching image to the query image. We also collect the 100 most similar images to a query – our matched set of images $I_m \in M$ – for use in our our content based description generation method (Sec 5).

For image comparison we utilize two image descriptors. The first descriptor is the well known gist feature, a global image descriptor related to perceptual dimensions – naturalness, roughness, ruggedness etc – of scenes. The second descriptor is also a global image descriptor, computed by resizing the image into a "tiny image", essentially a thumbnail of size 32x32. This helps us match not only scene structure, but also the overall color of images. To find visually relevant images we compute the similarity of the query image to images in C using a sum of gist similarity and tiny image color similarity (equally weighted).

Results – Size Matters! Our global caption generation method is illustrated in the first 2 panes and the first 2 resulting captions of Fig 2. This simple method often performs surprisingly well. As reflected in past work [13, 27] image retrieval from small collections often produces spurious matches. This can be seen in Fig 3 where increasing data set size has a significant effect on the quality of retrieved global matches. Quantitative results also reflect this (see Table 1).

4 Image Content Estimation

Given an initial matched set of images $I_m \in M$ based on global descriptor similarity, we would like to re-rank the selected captions by incorporating estimates of image content. For a query image, I_q and images in its matched set we extract and compare 5 kinds of image content:

- Objects (e.g. cats or hats), with shape, attributes, and actions – sec 4.1
- Stuff (e.g. grass or water) – sec 4.2

- People (e.g. man), with actions – sec 4.3
- Scenes (e.g. pasture or kitchen) – sec 4.4
- TFIDF weights (text or detector based) – sec 4.5

Each type of content is used to compute the similarity between matched images (and captions) and the query image. We then rank the matched images (and captions) according to each content measure and combine their results into an overall relevancy ranking (Sec 5).

4.1 Objects

Detection & Actions: Object detection methods have improved significantly in the last few years, demonstrating reasonable performance for a small number of object categories [7], or as a mid-level representation for scene recognition [20]. Running detectors on general web images however, still produces quite noisy results, usually in the form of a large number of false positive detections. As the number of object detectors increases this becomes even more of an obstacle to content prediction. However, we propose that if we have some prior knowledge about the content of an image, then we can utilize even these imperfect detectors. In our web collection, C, there are strong indicators of content in the form of caption words – if an object is described in the text associated with an image then it is likely to be depicted. Therefore, for the images, $I_m \in M$, in our matched set we run only those detectors for objects (or stuff) that are mentioned in the associated caption. In addition, we also include synonyms and hyponyms for better content coverage, e.g. "dalmatian" triggers "dog" detector. This produces pleasingly accurate detection results. For a query image we can essentially perform detection verification against the relatively clean matched image detections.

Specifically, we use mixture of multi-scale deformable part detectors [10] to detect a wide variety of objects – 89 object categories selected to cover a reasonable range of common objects. These categories include the 20 Pascal categories, 49 of the most common object categories with reasonably effective detectors from Object Bank [20], and 20 additional common object categories.

For the 8 animate object categories in our list (e.g. cat, cow, duck) we find that detection performance can be improved significantly by training *action specific detectors*, for example "dog sitting" vs "dog running". This also aids similarity computation between a query and a matched image because objects can be matched at an action level. Our object action detectors are trained using the standard object detector with pose specific training data.

Representation: We represent and compare object detections using 2 kinds of features, shape and appearance. To represent *object shape* we use a histogram of HoG [4] visual words, computed at intervals of 8 pixels and quantized into 1000 visual words. These are accumulated into a spatial pyramid histogram [18]. We also use an *attribute representation* to characterize object appearance. We use the attribute list from our previous work [15] which cover 21 visual aspects describing color (e.g. blue), texture (e.g. striped), material (e.g. wooden), general appearance (e.g. rusty), and shape (e.g. rectangular). Training images for the attribute classifiers come from Flickr, Google, the attribute dataset provided by Farhadi et al [8], and ImageNet [5]. An RBF kernel SVM is used to learn a classifier for each attribute term. Then appearance characteristics are represented as a vector of attribute responses to allow for generalization.

If we have detected an object category, c, in a query image window, O_q and a matched image window, O_m, then we compute the probability of an object match as:

$$P(O_q, O_m) = e^{-D_o(O_q, O_m)}$$

where $D_o(O_q, O_m)$ is the Euclidean distance between the object (shape or attribute) vector in the query detection window and the matched detection window.

4.2 Stuff

In addition to objects, people often describe the stuff present in images, e.g. "grass". Because these categories are more amorphous and do not display defined parts, we use a region based classification method for detection. We train linear SVMs on the low level region features of [8] and histograms of Geometric Context output probability maps [14] to recognize: sky, road, building, tree, water, and grass stuff categories. While the low level features are useful for discriminating stuff by their appearance, the scene layout maps introduce a soft preference for certain spatial locations dependent on stuff type. Training images and bounding boxes are taken from ImageNet and evaluated at test time on a coarsely sampled grid of overlapping square regions over whole images. Pixels in any

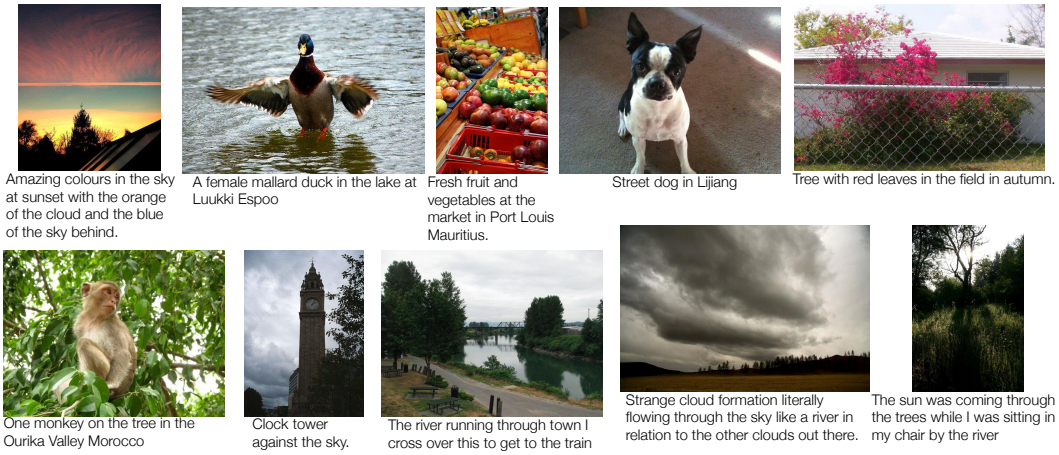

Figure 4: **Results:** Some good captions selected by our system for query images.

region with a classification probability above a fixed threshold are treated as detections, and the max probability for a region is used as the potential value.

If we have detected a stuff category, s, in a query image region, S_q and a matched image region, S_m, then we compute the probability of a stuff match as:

$$P(S_q, S_m) = P(S_q = s) * P(S_m = s)$$

where $P(S_q = s)$ is the SVM probability of the stuff region detection in the query image and $P(S_m = s)$ is the SVM probability of the stuff region detection in the matched image.

4.3 People & Actions

People often take pictures of people, making "person" the most commonly depicted object category in captioned images. We utilize effective recent work on pedestrian detectors to detect and represent people in our images. In particular, we make use of detectors from Bourdev et al [3] which learn poselets – parts that are tightly clustered in configuration and appearance space – from a large number of 2d annotated regions on person images in a max-margin framework. To represent activities, we use follow on work from Maji et al [21] which classifies actions using a the poselet activation vector. This has been shown to produce accurate activity classifiers for the 9 actions in the PASCAL VOC 2010 static image action classification challenge [7]. We use the outputs of these 9 classifiers as our action representation vector, to allow for generalization to other similar activities.

If we have detected a person, P_q, in a query image, and a person P_m in a matched image, we compute the probability that the people share the same action (pose) as:

$$P(P_q, P_m) = e^{-D_p(P_q, P_m)}$$

where $D_p(P_q, P_m)$ is the Euclidean distance between the person action vector in the query detection and the person action vector in the matched detection.

4.4 Scenes

The last commonly described kind of image content relates to the general scene where an image was captured. This often occurs when examining captioned photographs of vacation snapshots or general outdoor settings, e.g. "my dog at the beach". To recognize scene types we train discriminative multi-kernel classifiers using the large-scale SUN scene recognition data base and code [29]. We select 23 common scene categories for our representation, including indoor (e.g. kitchen) outdoor (e.g. beach), manmade (e.g. highway), and natural (pasture) settings. Again here we represent the scene descriptor as a vector of scene responses for generalization.

If a scene location, L_m, is mentioned in a matched image, then we compare the scene representation between our matched image and our query image, L_q as:

$$P(L_q, L_m) = e^{-D_l(L_q, L_m)}$$

where $D_l(L_q, L_m)$ is the Euclidean distance between the scene vector computed on the query image and the scene vector computed on the matched image.

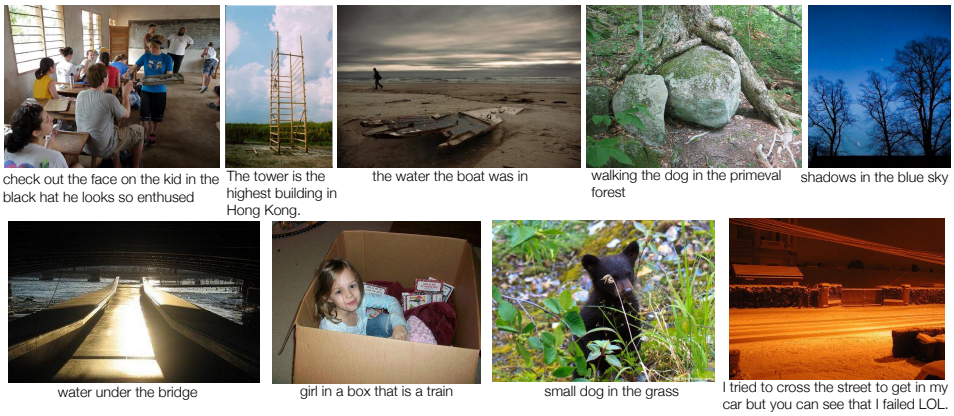

Figure 5: **Funny Results:** Some particularly funny or poetic results.

4.5 TFIDF Measures

For a query image, I_q, we wish to select the best caption from the matched set, $I_m \in M$. For all of the content measures described so far, we have computed the similarity of the query image content to the content of each matched image independently. We would also like to use information from the entire matched set of images and associated captions to predict importance. To reflect this, we calculate TFIDF on our matched sets. This is computed as usual as a product of term frequency (tf) and inverse document frequency (idf). We calculate this weighting both in the standard sense for matched caption document words and for detection category frequencies (to compensate for more prolific object detectors).

$$tfidf = \frac{n_{i,j}}{\sum_k n_{k,j}} * log \frac{|D|}{|j : t_i \in d_j|}$$

We define our matched set of captions (images for detector based tfidf) to be our document, j and compute the tfidf score where $n_{i,j}$ represents the frequency of term i in the matched set of captions (number of detections for detector based tfidf). The inverse document frequency is computed as the log of the number of documents $|D|$ divided by the number of documents containing the term i (documents with detections of type i for detector based tfidf).

5 Content Based Description Generation

For a query image, I_q, with global descriptor based matched images, $I_m \in M$, we want to re-rank the matched images according to the similarity of their content to the query. We perform this re-ranking individually for each of our content measures: object shape, object attributes, people actions, stuff classification, and scene type (Sec 4). We then combine these individual rankings into a final combined ranking in two ways. The first method trains a linear regression model of feature ranks against BLEU scores. The second method divides our training set into two classes, positive images consisting of the top 50% of the training set by BLEU score, and negative images from the bottom 50%. A linear SVM is trained on this data with feature ranks as input. For both methods we perform 5 fold cross validation with a split of 400 training images and 100 test images to get average performance and standard deviation. For a novel query image, we return the captions from the top ranked image(s) as our result.

For an example matched caption like "The little boy sat in the grass with a ball", several types of content will be used to score the goodness of the caption. This will be computed based on words in the caption for which we have trained content models. For example, for the word "ball" both the object shape and attributes will be used to compute the best similarity between a ball detection in the query image and a ball detection in the matched image. For the word "boy" an action descriptor will be used to compare the activity in which the boy is occupied between the query and the matched image. For the word "grass" stuff classifications will be used to compare detections between the query and the matched image. For each word in the caption tfidf overlap (sum of tfidf scores for the caption) is also used as well as detector based tfidf for those words referring to objects. In the event that multiple objects (or stuff, people or scenes) are mentioned in a matched image caption the

object (or stuff, people, or scene) based similarity measures will be a sum over the set of described terms. For the case where a matched image caption contains a word, but there is no corresponding detection in the query image, the similarity is not incorporated.

Results & Evaluation: Our content based captioning method often produces reasonable results (exs are shown in Fig 4). Usually results describe the main subject of the photograph (e.g. "Street dog in Lijiang", "One monkey on the tree in the Ourika Valley Morocco"). Sometimes they describe the depiction extremely well (e.g. "Strange cloud formation literally flowing through the sky like a river...", "Clock tower against the sky"). Sometimes we even produce good descriptions of attributes (e.g. "Tree with red leaves in the field in autumn"). Other captions can be quite poetic (Fig 5) – a picture of a derelict boat captioned "The water the boat was in", a picture of monstrous tree roots captioned "Walking the dog in the primeval forest". Other times the results are quite funny. A picture of a flimsy wooden structure says, "The tower is the highest building in Hong Kong". Once in awhile they are spookily apropos. A picture of a boy in a black bandana is described as "Check out the face on the kid in the black hat. He looks so enthused." – and he doesn't.

We also perform two quantitative evaluations. Several methods have been proposed to evaluate captioning [15, 9], including direct user ratings of relevance and BLEU score [24]. User rating tends to suffer from user variance as ratings are inherently subjective. The BLEU score on the other hand provides a simple objective measure based on n-gram precision. As noted in past work [15], BLEU is perhaps not an ideal measure due to large variance in human descriptions (human-human BLEU scores hover around 0.5 [15]). Nevertheless, we report it for comparison.

Method	BLEU
Global Matching (1k)	0.0774 +- 0.0059
Global Matching (10k)	0.0909 +- 0.0070
Global Matching (100k)	0.0917 +- 0.0101
Global Matching (1million)	0.1177 +- 0.0099
Global + Content Matching (linear regression)	0.1215 +- 0.0071
Global + Content Matching (linear SVM)	0.1259 +- 0.0060

Table 1: Automatic Evaluation: BLEU score measured at 1

As can be seen in Table 1 data set size has a significant effect on BLEU score; more data provides more similar and relevant matched images (and captions). Local content matching also improves BLEU score somewhat over purely global matching.

In addition, we propose a new evaluation task where a user is presented with two photographs and one caption. The user must assign the caption to the most relevant image (care is taken to remove biases due to placement). For evaluation we use a query image and caption generated by our method. The other image in the evaluation task is selected at random from the web-collection. This provides an objective and useful measure to predict caption relevance. As a sanity check of our evaluation measure we also evaluate how well a user can discriminate between the original ground truth image that a caption was written about and a random image. We perform this evaluation on 100 images from our web-collection using Amazon's mechanical turk service, and find that users are able to select the ground truth image *96%* of the time. This demonstrates that the task is reasonable and that descriptions from our collection tend to be fairly visually specific and relevant. Considering the top retrieved caption produced by our final method – global plus local content matching with a linear SVM classifier – we find that users are able to select the correct image *66.7%* of the time. Because the top caption is not always visually relevant to the query image even when the method is capturing some information, we also perform an evaluation considering the top 4 captions produced by our method. In this case, the best caption out of the top 4 is correctly selected *92.7%* of the time. This demonstrates the strength of our content based method to produce relevant captions for images.

6 Conclusion

We have described an effective caption generation method for general web images. This method relies on collecting and filtering a large data set of images from the internet to produce a novel web-scale captioned photo collection. We present two variations on our approach, one that uses only global image descriptors to compose captions, and one that incorporates estimates of image content for caption generation.

References

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

[2] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, E. Learned-Miller, Y. Teh, and D. Forsyth. Names and faces. In *CVPR*, 2004.

[3] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[6] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation. In *ECCV*, 2002.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

[8] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[9] A. Farhadi, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth. Every picture tells a story: generating sentences for images. In *ECCV*, 2010.

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

[11] Y. Feng and M. Lapata. How many words is a picture worth? automatic caption generation for news images. In *Proc. of the Assoc. for Computational Linguistics*, ACL '10, pages 1239–1249, 2010.

[12] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007.

[13] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[14] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *Int. J. Comput. Vision*, 75:151–172, October 2007.

[15] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. In *CVPR*, 2011.

[16] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[17] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching. In *CVPR*, June 2006.

[19] W. Li, W. Xu, M. Wu, C. Yuan, and Q. Lu. Extractive summarization using inter- and intra- event relevance. In *Int Conf on Computational Linguistics*, 2006.

[20] E. P. X. Li-Jia Li, Hao Su and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2010.

[21] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

[22] R. Mihalcea. Language independent extractive summarization. In *National Conference on Artificial Intelligence*, pages 1688–1689, 2005.

[23] A. Nenkova, L. Vanderwende, and K. McKeown. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *SIGIR*, 2006.

[24] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu. Bleu: a method for automatic evaluation of machine translation. pages 311–318, 2002.

[25] D. R. Radev and T. Allison. Mead - a platform for multidocument multilingual text summarization. In *Int Conf on Language Resources and Evaluation*, 2004.

[26] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*, 2010.

[27] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *PAMI*, 30, 2008.

[28] K.-F. Wong, M. Wu, and W. Li. Extractive summarization using supervised and semi-supervised learning. In *International Conference on Computational Linguistics*, pages 985–992, 2008.

[29] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[30] B. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2t: Image parsing to text description. *Proc. IEEE*, 98(8), 2010.