Robust image retrieval using topic modeling on captioned image data

Jia-Shen Boon, Akshay Sood, Meenakshi Syamkumar
Department of Computer Sciences
University of Wisconsin-Madison
[boon,sood,ms]@cs.wisc.edu

Abstract

Image retrieval based only on image features emphasizes only visual similarity and does not capture semantic similarity between images. In order to capture the semantic similarity textual data associated with images can be very useful. We demonstrate that some semantics of an image, while poorly captured by the image alone, can be captured by text that accompanies the image. These semantics include artistic feel and sociocultural events. We capture these semantics by modeling the topics generated by the accompanying text, referred to as captions, whie visual features are extracted with a deep convolutional network. A joint model of text and images is applied to the Flickr8K dataset. We also collect a custom dataset of over 32K images and 110K captions, crawled from Imgur. This model has applications in general to image retrieval, as well as generating links to similar images within popular photo-sharing websites such as Imgur. In the latter application, such links would allow users to 'account-hop', which would increase visitor stay duration.

1. Introduction

The availability of a large and growing corpus of image data, associated with accompanying text in the form of crowd-sourced commentaries, presents an opportunity to augment the task of retrieval of images similar to a query image. By leveraging this text, significant improvements can be made in the identification and retrieval of images similar in semantic content.

The semantic content in images can be captured at multiple levels, as shown in Figure 1. The top image depicts Rowan Atkinson playing Mr. Bean at the 2012 Summer Olympics opening ceremony, and we consider the task of identifying images similar to this image from a pool of images. The bottom-left image captures low-level visual features by depicting a man on a beach. The bottom-center image captures high-level visual features by depicting Mr. Bean in another scene. The bottom-right image captures non-visual similarity by showing another scene from the opening ceremony, depicting the Queen of England jumping off a helicopter. Methods to identify similar images based only on image data can perform the first task with relative ease, whereas the second task requires more sophistication. The third task requires leveraging non-visual similarity, so cannot be directly performed using image data alone.

Figure 1: Capturing image similarity at different levels. The query image depicts Rowan Atkinson playing Mr. Bean at the 2012 Summer Olympics opening ceremony. (a) captures low-level visual features by depicting a man on a beach. (b) captures high-level visual features by depicting Mr. Bean in another scene. (c) captures non-visual similarity by showing another scene from the opening ceremony depicting the Queen of England jumping off a helicopter.

We propose the use of text data associated with images in order to capture image similarity at multiple levels. In particular, we use a topic modeling approach for the text data. We hypothesize that it would allow us to address two challenges for image retrieval based on image data alone: identifying visually noisy or absent but semantically related ideas, and distinguishing visually similar but semantically distinct images. Figure 2 gives an example of how topic modeling over text may be able to address these issues.

Image-sharing sites like Imgur, Flickr, Instagram and Facebook provide excellent platforms to associate multitudes of images with useful descriptions via comments and

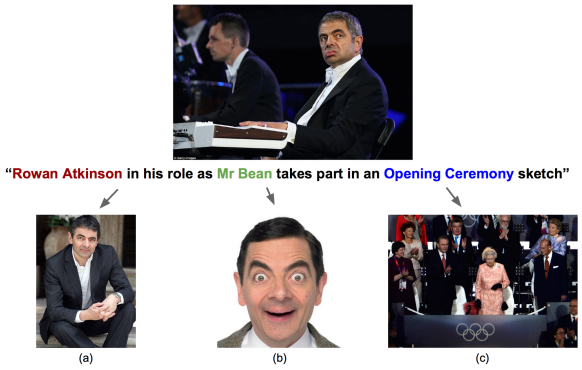


Figure 2: Leveraging topic modeling over text associated with images can allow us to identify visually noisy or absent but semantically related ideas: for instance (c) is representative of the topic associated with the 2012 Summer Olympics opening ceremony. It also allows to distinguish visually similar but semantically distinct ideas: for instance (a) captures the actor Rowan Atkinson while (b) captures Mr. Bean, the character he plays. Image/Caption source: http://www.dailymail.co.uk/news/article-2179920/Olympics-Opening-Ceremony-London-gets-2012-Games-way-Greatest-Show-On-Earth-rounded-Macca-course.html

tags. Leveraging the associated textual aids in addition to image data in performing image retrieval would allow better capture of associated semantics in the image matching process, thereby facilitating better retrieval of similar images. For this task, we propose using topic modeling for the text data using Latent Dirichlet Allocation and convolutional neural networks for the image data. These may then be combined to form a better reverse image search engine, as well as to retrieve captioned images given a query captioned image, i.e., captioned image retrieval.

2. Related Work

Researchers in image retrieval have long known that image features are complementary to, but not a substitute for text features [1]. The seminal work in [2] defined two types of gaps in content-based image retrieval (CBIR): the *sensory gap* between a real world object and what a sensor can capture from the object, and the *semantic gap* between what is captured by the sensor and the user's interpretation of the sensor data. In our work, we bridge this semantic gap with image captions.

[3] has jointly modeled images and captions with deep Boltzmann machines. [4] has used topic modeling using probabilistic latent semantic analysis (pLSA) for multimodal image retrieval (using images and image tags). [5] has used neural networks to jointly model images and captions by fragmenting images and caption and embedding them in a common space, using these subsequently for bidirectional image-sentence retrieval. In contrast, we train images and captions separately, which should reduce computational training time considerably. [6] has done image-only retrieval using deep autoencoders. For efficient look-up, images are mapped to binary codes. In contrast, our proposed model infers codes in continuous space, allowing us to represent images in a larger space for the same number of dimensions.

[7] proposes determining a correspondance LDA model which finds conditional relationship between latent variables that represent the set of interesting image regions and the set of words from their captions and applications of their model to automatic image annotation and image retrieval from textual queries. We intend to use an LDA model for topic modeling over the image captions only, instead of computing correspondences between captions and image regions; we leverage two separate models on captions and images to provide a combined text-image based retrieval application.

3. Datasets

3.1. Available datasets

Several datasets exist for the purpose of general CBIR. Captioned image datasets include Pascal 1K, Flickr8K and Flickr30K [8, 9]. Each image in these datasets is annotated by 5 people to produce 5 captions describing the image. We apply our models to the Flickr8k dataset in this paper. The COREL dataset consists of 10.8K images manually clustered into 80 concept groups, for the purpose of CBIR [10]. However, the images have no associated text. ImageCLEF is an annual image retrieval competition [11]. Its Wikipedia image retrieval task includes a dataset of over 200K images and their accompanying text, but the task was terminated in 2011. [12] has crawled Flickr for over a million images and their respective captions. The purpose is to automatically generate image captions, and as such these captions generally only describe the visual features of the image rather than the broader semantics that we want to capture with text.

3.2. Custom dataset

In order to establish the significance of extracting public data available on social media platforms like imgur, Facebook, Instagram, Pinterest and so on, we created our custom dataset, *DataM*, which is generated by crawling imgur. A public application was registered for the crawling in order to obtain *client_id* and *client_secret*, which provide access to *imgurAPI*. In order to perform the crawling, we used *gallery_search* operation by using specific search keywords. For maintaining the standard for the search keywords, we obtained a list of 1000 class synsets from [13]. These standard synset classes are from the WordNet database [14]. We

performed crawling with each of these search keywords and downloaded the resultant images. Along with the images, we also crawl the title and top 4 comments to form the captions associated with the images. Apart from the WordNet [14] synset list, we also created custom generated keywords to represent human sports related activities like *snowboarding, running, swimming, etc.*. In order to prevent the application from running into "Ratelimit per application" issue, we got our application white-listed, as ours is a cost-free application. In addition, to overcome the "Ratelimit per IP address" issue, we provisioned multiple Ubuntu EC2 instancess in parallel on Amazon AWS [15] and executed our application on those instances. We were able to generate a corpus containing 32840 images and 110683 captions.

Much of the text from captions is uninformative, but some contain artistic or sociocultural references that would be extremely difficult for a computational model (e.g., a neural network) to deduce from the image data alone, as we have hypothesized (see Figure 3). These references include regality, mythical creatures such as dragons and kaiju (Japanese monsters), and pop culture icons like tardis (a spaceship in the Dr Who science fiction series) and Spider-Man.

4. Methodology

Our overarching approach to image retrieval is simple. We extract features from training images using a deep convolutional neural network; we extract features from training captions using LDA; to find similar captioned images given a query captioned image, we perform the same feature extraction procedure on the query and return the training examples that are nearest to the query in the feature space, i.e., using K-Nearest Neighbors and some distance metric. Details are provided in the following subsections. All of our implementations can be found at [16].

4.1. Baseline

For a baseline image retrieval system, we perform (a) image-only retrieval using SIFT features and spatial pyramid matching, and (b) text-only retrieval using scores based on TF-IDF and kurtosis. Since our evaluation is qualitative, we do not provide the baseline results here.

4.2. Image feature extraction

Deep neural networks have the ability to learn a hierarchy of features of image data. This may be done in a supervised [17] or unsupervised manner [18]. Lower layers of the hierarchy learn low level visual features like Gabor filters, while higher layers learn higher level visual features like patterns and parts of objects, as alluded to in Figure 1.

Our deep convolutional neural network is based on CaffeNet, a variation of AlexNet [17]. AlexNet achieved

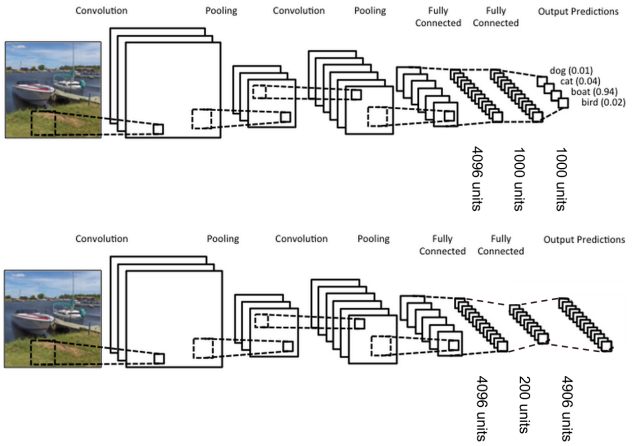


Figure 4: Simplified representation of CaffeNet without (top) and with (bottom) finetuning.
Top - the original CaffeNet ends with several fully-connected layers followed by a 1000-unit softmax layer. We remove this softmax layer because we do not want its "squashing" nature during KNN search.
Bottom - in the finetuned CaffeNet, we remove both the softmax layer and the fully connected layer below that, replacing them with a shallow autoencoder. The new architecture reduces computational resources, and may also improve the quality of retrieval results.

state-of-the-art results on the ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2010 classification task. A pre-trained model of CaffeNet is provided by the Caffe deep learning framework [19]. This model has been trained on 1.2 million (labeled) ImageNet images over 310K iterations. We test two variations of CaffeNet on the Flickr8k dataset - the pre-trained model as-is, and the pre-trained model with finetuning, as shown in Figure 4.

To finetune CaffeNet, we first replace its softmax layer and the fully connected layer below with a shallow autoencoder. Adding the autoencoder serves two purposes. Firstly, the extracted image feature vector has fewer dimensions because of the bottleneck of the autoencoder. This dimensionality reduction not only reduces the data storage requirement but also speeds up image retrieval during KNN. Secondly, by finetuning the model with our own dataset, the model fits the regularity of this dataset and therefore the quality of the retrieved images should improve as well.

4.3. Topic modeling over captions

We use a topic modeling approach for the captioned text associated with the images. This is motivated by the assumption that the text is generated using underlying latent topics, and that similar images should have captions with similar distributions over topics. Thus, by identifying the

topic distributions for each caption, we may be able to retrieve similar images by retrieving similar captions. To this end, we use the widely used Latent Dirichlet Allocation [20] (LDA) for topic modeling over the caption data.

We first process text data by lemmatization and filter out stop words to generate our vocabulary. We then apply a TF-IDF weighting scheme to the data before topic modeling.

We use LDA with Gibbs Sampling to estimate distributions over topics for each caption and for each word in the vocabulary. Given a query caption, we estimate its topic distribution using the topic distributions of the words that it comprises. We use K-Nearest Neighbors to identify the captions most similar to the query caption, using symmetrized KullbackLeibler divergence as a distance metric. We use Laplace smoothing to mitigate the effects of zero probabilities in computing KL divergence.

4.4. Combined Model

In order to create a combined model, we perform a merge of the feature vectors which are used as input for k-NN's that are used in the two models (LDA and AlexNet). We generate the feature vectors for the training and the test sets by executing the above two models for image feature and topic model extraction. To make the visual and textual features comparable in feature space, we scale image features. Textual features are left unscaled since they represent a probability distribution over topics and should sum to 1. The image-based features for the training set are normalized by dividing the feature vector by the standard deviation of image features and multiply by the standard deviation of the LDA features, from the training set. The image-based test set is also normalized by the same factor from the training set. Once these modifications are completed, the training and the test feature vectors are concatenated. We experiment with different types of distance metrics to generate the result for k-NN search (with k = 3): L2 distance on the whole feature set, L2 distance for image-based features and symmetrised KL divergence for LDA features.

5. Results

5.1. Image retrieval with image data only

Both CaffeNet models, with and without finetuning, take as input a 227x227 color image. Before feeding an image to a network, we resize it to 227x227 and subtract it by a pre-computed mean image.

To perform feature extraction on the non-finetuned CaffeNet, we perform a forward pass on the network, stopping just before the (last) softmax layer. The softmax layer allows the network to output a class probability, as required in the network's original purpose, but it also tends to cause all features but one to approach zero. We remove this layer since we want all features - not just one - to be informative during the KNN search.

As aforementioned, we replace the last two layers of CaffeNet with an autoencoder before finetuning (see Figure 5). Concretely, the layers after the 4096-unit full connected layer are originally: relu, dropout, 1000-unit fully connected layer and softmax. We replace these with: sigmoid, 200-unit fully connected layer, 4096-unit fully connected layer and sigmoid. *sigmoid* refers to a sigmoid layer. Rectified linear unit (relu) and dropout are general neural network tricks introduced in the original AlexNet paper [17]. We stick with sigmoid units and no dropout in our autoencoder to keep the architecture simple.

The weights of the modified CaffeNet are finetuned as follows. The weights of the new layers are initialized randomly. Training images are fed into the network in minibatches of 50 images each. Training minimizes the L2 error between the output of final sigmoid layer and the first 4096-unit layer, as expected of an autoencoder. Error derivatives of the previous layers are calculated by backpropagation. We **do not** backpropagate error to layers earlier than the autoencoder. That is, weights of the layers outside of the autoencoder are not finetuned. The base learning rate is 0.001 and drops by a factor of 10 every 3000 iterations. As the weights are updated according to the training dataset, we track the network's L2 error on the validation set and we stop training when the validation error plateaus. With this procedure, training took about 20 hours and 9000 iterations on an Intel Core i5-4260U CPU with 4 1.40GHz cores and 4 GB of RAM.

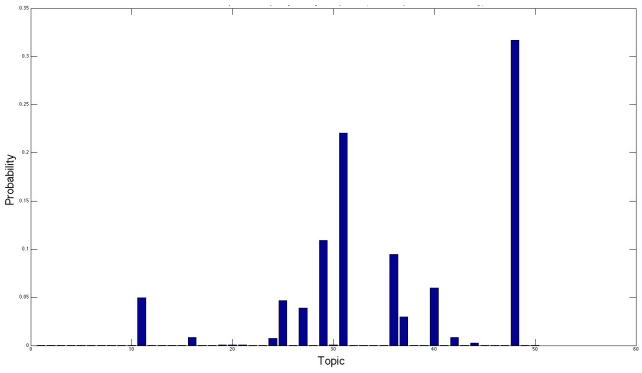
A sample of the retrieval results of CaffeNet are shown in Figure 5. Both pre- and post-finetuned networks seem to have captured object-level concepts such as *dog*, *person*, *street*. This is expected since the network was pretrained on classification-labeled data. Also because ImageNet, the pre-training data, consists of image with a clear subject, the pre-tuned network performs poorly when the query does not have a single, prominent subject, e.g., in a cluttered scene or when there's multiple subjects. Finetuning does not appear to improve the quality of retrieval results, but it is important to emphasize that the finetuned model still has faster retrieval times and a smaller data footprint than the model without finetuning.

5.2. Text-only image retrieval

Using LDA with Gibbs sampling, we estimate distributions over topics for each word and for each caption. We set LDA hyperparameters as $\alpha = \frac{50}{T}$ as suggested in [21] and $\beta = \frac{200}{W}$, where T is the number of topics and W is the vocabulary size.

For the Flickr8k dataset, we fix $T = 50$. We use lemmatized versions of captions, and combine multiple captions for an image into a single caption associated with that image. Figure 6 shows a query image and captions associated

(a)



(b)

Figure 6: (a) Query image from Flickr8k with associated lemmatized captions, concatenated to form a single caption: A man leap into the air with a mountain vista behind him; A man pose as he jump from rock to rock in a forest; A man step off of a boulder; A young man leap through the air from a rocky mountaintop (b) Distribution over topics for that caption (with Laplace smoothing)

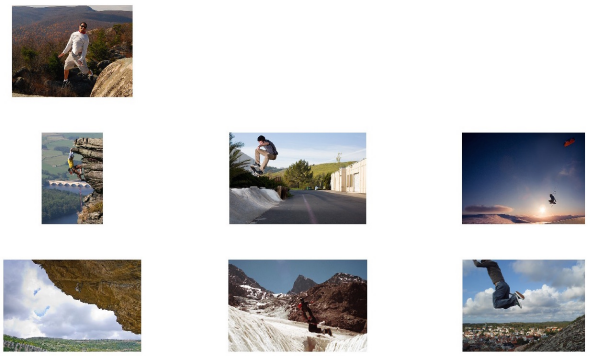


Figure 7: Retrieval results using text only. *Top left* - query image. *Bottom two rows* - retrieval results in order of decreasing similarity (row-major order)

with that image, concatenated into a single caption, along with the distribution over topics for that caption. Figure 7 shows the images corresponding to the captions that are the nearest neighbors for the query caption, in the sense of minimizing symmetrized KL divergence.

Using the distribution over topics for the query caption, we can also compute the most likely topics represented in

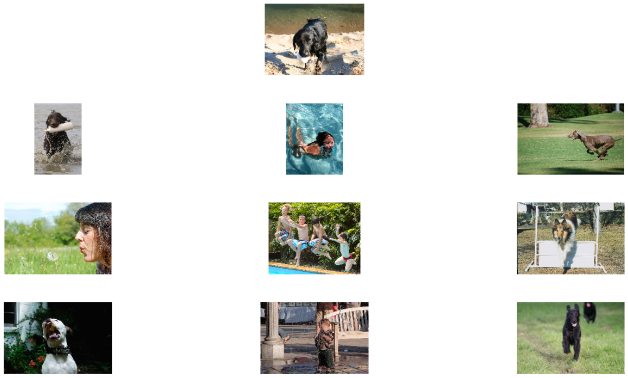


Figure 8: Given a query image, we can retrieve the images representative of the most likely topics in the caption of the query image. Each column here corresponds to images representative of such a topic. Most likely words for 1st topic: MOUTH CARRY STICK; most likely words for 2nd topic: WATER SWIM POOL; most likely words for 3rd topic: FIELD GRASS GRASSY

that caption. We can then determine which images are most representative of each of these topics and return these topic-wise similar images. Figure 8 shows this for an example query image.

Figure 9 shows some image retrieval results using LDA over the caption data for our custom Imgur dataset DataM. Due to the diversity of topics in our dataset, we use LDA with $T = 500$ topics to model the data. The results show that topic modeling can capture semantic ideas in the image, for both visually present and absent ideas.

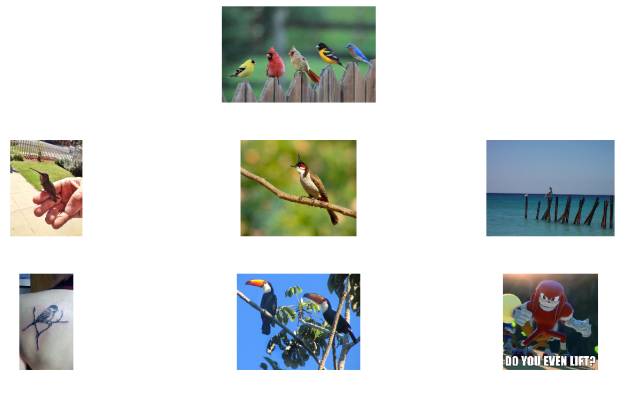
5.3. Text-image combined image retrieval

We executed the combined model described in section 4.3, by altering the weightage associated with the feature set from the two different models. Given the features of two captioned images $< p_1, c_1 >$ and $< p_2, c_2 >$, where p stands for picture and c stands for caption, the distance between these captioned images is given by

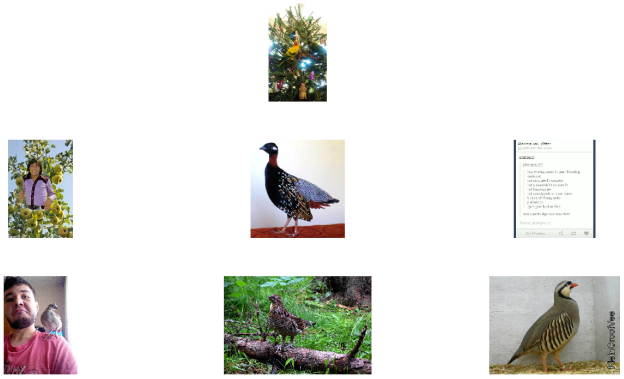
$$dist = f_p * ||p_1 - p_2||_2 + f_c * KL(c_1, c_2)$$

.

A higher factor f is a downweighting of the corresponding features. KL stands for symmetric KL divergence. We tried three combinations of weights: image and text features being equally weighted, image based features being weighted more and text based features being weighted more. The top results, i.e., 1-NN result, of each weightage combiation is shown in Figure 10. When image features are more heavily weighted, the results tend to be more visually similar; when textual features are more heavily weighted, the similarites are more abstract, which is to be expected



(a)



(b)

Figure 9: Image retrieval results using LDA over custom Imgur dataset DataM. The top image in each sub-figure is the query image. (a) demonstrates how our approach is able to identify the latent topic 'bird' underlying the query image - which consists of multiple bird species - and retrieve a diverse set of images depicting different bird species. (b) shows retrieved images capturing the idea of 'partridge in a pear tree' (from the famous Christmas carol 'The Twelve Days of Christmas') regardless of visual similarity.

since captions carry only coarse visual information.

6. Conclusion

We hypothesized that captions are a useful additional source of information to measure image similarity. We crawled Imgur to create a dataset *DataM* of over 32K images and their accompanying text. By manually inspecting DataM, we found some evidence to validate our hypothesis, although it is a challenge to find the signal over the noisy textual data. Nevertheless, we obtained meaningful results by performing image retrieval over DataM using a text-only topic modeling approach, validating the usefulness of this approach for enhancing image retrieval in real applications. When the text data is fairly clean, as in the case of Flickr8k, we demonstrated that image retrieval using captions alone is feasible using LDA. Topic modeling over the text also allowed us to retrieve images representative of the salient topics captured in an image caption - investigation into the value that this may add for data gathered from social media should be fertile ground for future enquiry.

We showed that using a deep pre-trained neural network, even one that is trained for classification rather than CBIR, is a quick-and-dirty way to perform CBIR with image data alone, that yields very decent results. Adding an autoencoder to the network and finetuning the weights reduces computational requirements, and can also improve the quality of retrieval results.

We combined the image-only and text-only models to obtain a joint model that extracts similarity captured in either the image or the text data or both. We demonstrated that the joint model can improve image retrieval performance over either model for the Flickr8k dataset, based on a qualitative comparison.

Acknowledgements

We would like to extend our heartfelt thanks to Prof. Vikas Singh for his invaluable guidance and suggestions for this project. We would also like to thank Jia Xu for guiding us through all assignments.

References

[1] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *Journal of visual communication and image representation*, vol. 10, no. 1, pp. 39–62, 1999. 2

[2] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349–1380, 2000. 2

[3] N. Srivastava, R. R. Salakhutdinov, and G. E. Hinton, "Modeling documents with deep boltzmann machines," *arXiv preprint arXiv:1309.6865*, 2013. 2

[4] S. Romberg, R. Lienhart, and E. Hörster, "Multimodal image retrieval," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 1, pp. 31–44, 2012. 2

[5] A. Karpathy, A. Joulin, and F. F. F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in Neural Information Processing Systems*, 2014, pp. 1889–1897. 2

[6] A. Krizhevsky and G. E. Hinton, "Using very deep autoencoders for content-based image retrieval." in *ESANN*. Citeseer, 2011. 2

[7] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 127–134. 2

[8] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 139–147. 2

[9] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014. 2

[10] J. Z. Wang, J. Li, and G. Wiederhold, "Simplicity: Semantics-sensitive integrated matching for picture libraries," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 9, pp. 947–963, 2001. 2

[11] I. Eggel and H. Müller, "The imageclef management system," in *Multilingual Information Access Evaluation II. Multimedia Experiments*. Springer, 2010, pp. 332–339. 2

[12] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Advances in Neural Information Processing Systems*, 2011, pp. 1143–1151. 2

[13] "1000 class Synset words from WordNet." https://github.com/sh1r0/caffe-android-demo/blob/master/app/src/main/assets/synset_words.txt. 2

[14] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995. 2, 3

[15] "Amazon AWS." http://aws.amazon.com/. 3

[16] "Github link for our implementation." https://github.com/msyamkumar/captioned_image_retrieval_CS766_final. 3

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 3, 4

[18] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8595–8598. 3

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014. 3

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003. 4

[21] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004. 4

(a) Regal Ring-necked Snake.
- **Regal** as f***.
- Its like a gummy... nope.
- It's so pretty. o_o
- Y'all motherf****** need to learn about generally non-venomous critters.

(b) Blue Dragon Sea Slug
- That sh*t looks majestic.
- This is a baby **dragon** and nothing you say will make me think otherwise.
- Infant **kaiju**.
- must have

(c) Saw one of you driving down from Snowbird ski area, Salt Lake Cityish area. P.S you were speeding and almost hit you when taking photo.Mobile Upload
- I'm so glad you're a safe and logical driver.
- You idiot?
- It's a **TARDIS** in disguise. There is no speed limit.
- so if the **whovian** was speeding where you speeding too to keep up to take the photo?

(d) **Spiderman** colored Agama Mwanzae lizard
- Mas: http://i.imgur.com/DrFX6Gc.jpg http://i.imgur.com/FQ4nByC.jpg http://i.imgur.com/fKdBdbp.jpg
- New villain in the next Spider-Man movie
- Does whatever a lizard can.
- Ah the Red Headed Agama, not just for Mondays I see

Figure 3: A sample of the images, captions and comments crawled from imgur. Below each image is a caption by the uploader followed by the top four comments, as shown in bullets. Emphasis is ours.

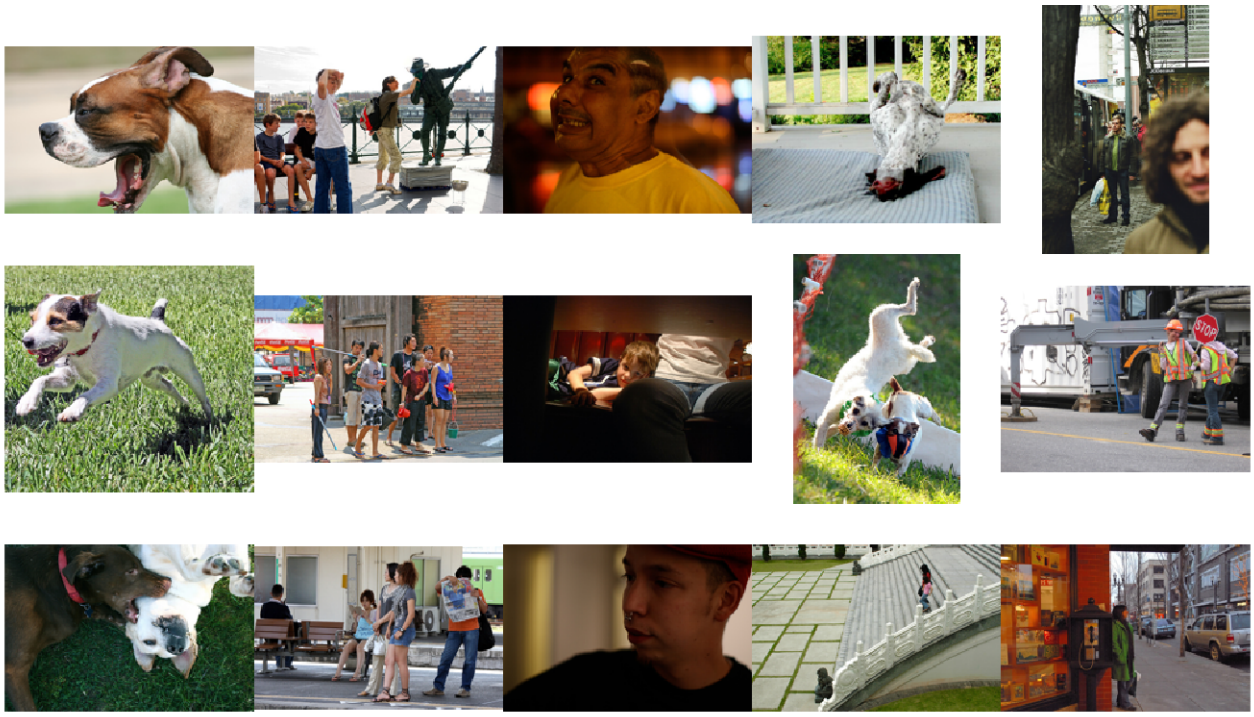


Figure 5: Retrieval results before and after finetuning CaffeNet (without using captions). *Top* - query images from Flickr8k's test set; *middle* - images retrieved before finetuning; *bottom* - images retrieved after finetuning.

Figure 10: Retrieval results from a joint model of text and images. *First column* - query images from Flickr8k's test set; *Second column* - retrieval results when text and image are equally weighted; *Third column* - retrieval results when the image is weighted more heavily; *Fourth column* - retrieval results when the text is weighted more heavily