Semantic image search using queries

Shabaz Basheer Patel, Anand Sampat
Department of Electrical Engineering
Stanford University
CA 94305
shabaz@stanford.edu,asampat@stanford.edu

Abstract

Previous work, on image search looks into RNNs and DT-RNNs and applying it over the queries. In this work, we model to generate natural language descriptions for images and this is utilized in order to represent the image. It is achieved by forming a model based on a combination of Convolutional Neural Networks over image regions, along with a Recurrent Neural Networks over sentences. This work uses LSTMs and the experiments are done over Flickr8K datasets.

1 Introduction

When a human looks at an image, he makes many inferences from that image. However, it is very difficult for a computer to infer such an information. Previously most works focused on labeling images with tags with the constituents in the image such as rCNN[1], OverFeat[2]. However, it doesn't capture all the descriptions which a human can understand. This work looks into achieving a closed vocabularies of visual constituents to describe the image. We use this information in order to form a representation for the image. Hence, for a particular query we provide images with the learnt representations.

Most of the work have used RNNs[3] and also using DT-RNNs[4] over the queries, and had provided good results. In this work, we plan to use LSTMs over the CNN features and understand the outputs observed from using this new RNN.

In this method, we approach it by extracting image features using VGG 16-layer CNN. Once, we generate the image context vector, we provide it to the LSTM only at the first iteration. By this, we get words for every iteration through the LSTM, which represents the image. Over this words we then use GloVe vectors to represent words and map this sentence in the same embedded vector space along with mapping the query in the same vector space.

2 Related Work

Work done by Tracy et. al. uses distributional similarity of words in semantic vector space[5]. Usually tf-idf is used in order to describe each word. Most of such compositional algorithms uses a maximum of two-words in their query phrase and then analyze similarities computed by the cosine distance or any other similar metric.

Socher et. al.[6] projects words and image regions into the same multimodal embedded space using kernelized canonical correlation anaysis. Socher et. al. projects a single word vector embedding in order to perform zero shot learning[7]. Such a mapping enables to classify unseen images. Thus,

it explains that such a multimodal embedding helps in extracting the semantic information from the image.

Generating contextual information to improve recognition has been a recent research where Duygulu et. al.[8] performs such an analysis.Farhadi et al[9] performs an automatic method to parse images. It uses a triple of objects estimated for an image to retrieve sentences from a query. Our work approaches the problem in a similar manner where we generate sentences for our images and embed this sentence in the vector space where we also map our query. Our work initially focuses on generating natural language descriptions for an image. We use data from the MSCOCO datasets to train a model to generate sentences from image features using a CNN and RNN alike. We build upon the publicly available code of Andrej et. al.[10] which uses VGG CNN[11] to get image features and then followed by using LSTM to generate the descriptions.

We expect annotated images along with a comparison of sentences generated by our algorithm vs. the sentence per image in the training set. Our evaluation will calculate an image-sentence score by mapping regions of the image to words in the phrase - probability for each region/word pair. BLEU scores will be generated for each image in the test set and averaged over the entire set. From these generated sentences for the image, we map into word vector space either by averaging both the sentence and query. We also do this using a RNN to embed both these into a multimodal embedded space.

3 Technical Approach

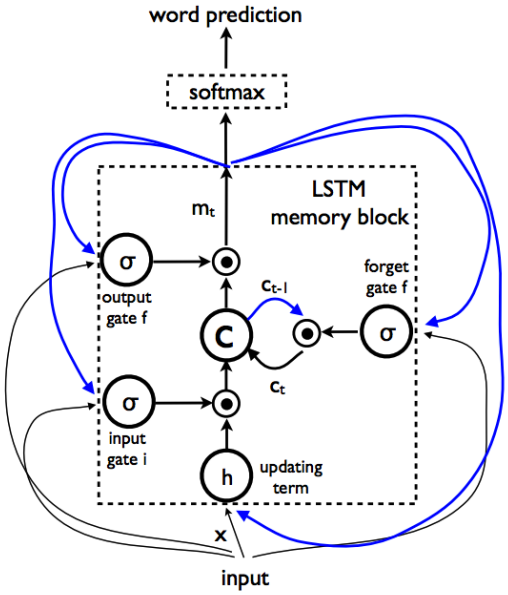

Figure 1: LSTM: the memory block contains a cell c which is controlled by three gates

We evaluate the performance of those recently proposed recurrent units (LSTMs) on sequence modeling for generating descriptions. Before the evaluation, we first describe the LSTMs in this section. Figure 1 shows the mostly used gated neural network, Long Short-Term Memory[12].

LSTMs adaptively captures dependencies of different time scales. It has gating units that modulate the flow of information inside the unit. The definition of the gates, cell update and output are as follows:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1})$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1})$$
$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1})$$
$$c_t = f_t \circ c_{t-1} + i_t \circ h(W_{cx}x_t + W_{cm}m_{t-1})$$
$$m_t = o_t \circ c_t$$
$$p_{t+1} = Softmax(m_t)$$

$\circ$ represents the hadamard product.W represents the trained parameters, σ and $tanh$ are non-linearity functions used in the above equations.

3.1 Pipeline for the Image Retreival System

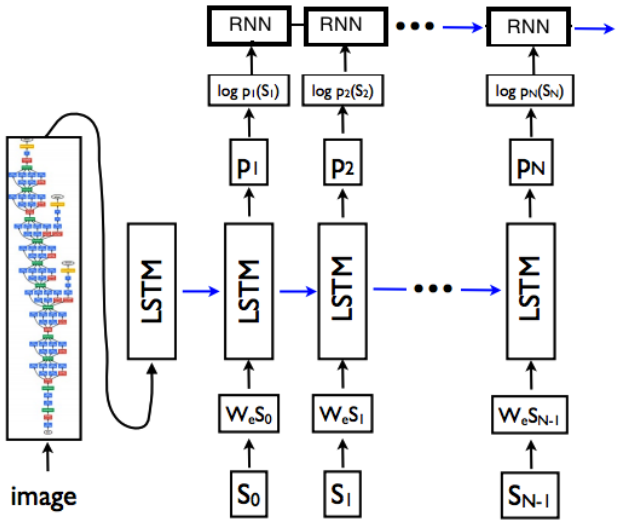

Figure 2: the pipeline for the mapping the image into a multimodal embedded space

We assume an input set of images and their textual descriptions for the training set. The main challenge is to design a model that can predict a variable-sized sequence of words given an image. The developed language models is based on LSTMs, it is achieved by defining a probability distribution of the next word in a sequence given the previous and current words. In this model, Andrej et. al.[10] uses a simple but a effective extension that additionally conditions the generative process on the content of an input image. During the training, our neural network takes the image and a sequence of input vectors $(x_1, ..., x_t)$. It then computes a sequence of hidden states $(h_1, ..., h_t)$ and a sequence of outputs $(y_1, ..., y_t)$ by iterating the following recurrence relation for $t = 1$ to T :

$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + 1(\text{t=1}) \circ b_v)$$
$$y_t = softmax(W_{oh}h_t + b_0)$$

In the equations, $W_{hi}, W_{hx}, W_{hh}, W_{oh}, x_i$ and b_h, b_o are the parameters for which we train the the system. We extract features using the last layer of the CNN when using $CNN_{\theta_c}(I)$. The output vector y_t holds the log probabilities of words in the dictionary and one additional dimension for a special END token. We provide context vector b_v to the RNN only at the first iteration.

From the generated description for the image using the above described model. We map this sentence into a multimodal embedded space. The input query is also mapped onto the same space. This is done by the following methods,

a) By utilizing 300-D GloVe vectors for very word for the generated sentence and for the query. It is mapped by averaging these vectors.

b) We utilize a RNN, which is trained over the captions in the COCO training dataset. The generated sentence is passed into RNN and this sentence is mapped in the final 100-D hidden vector from the RNN. In the same manner mapping for the query is done using the same RNN. From the above techniques we get vectors representing all images and the query in the same vector space. Then, in order to get closest image to our query, we perform cosine distance over all images and return those images to the user which very closely represents the query.

4 Experiments Results

4.1 Generating Sentences

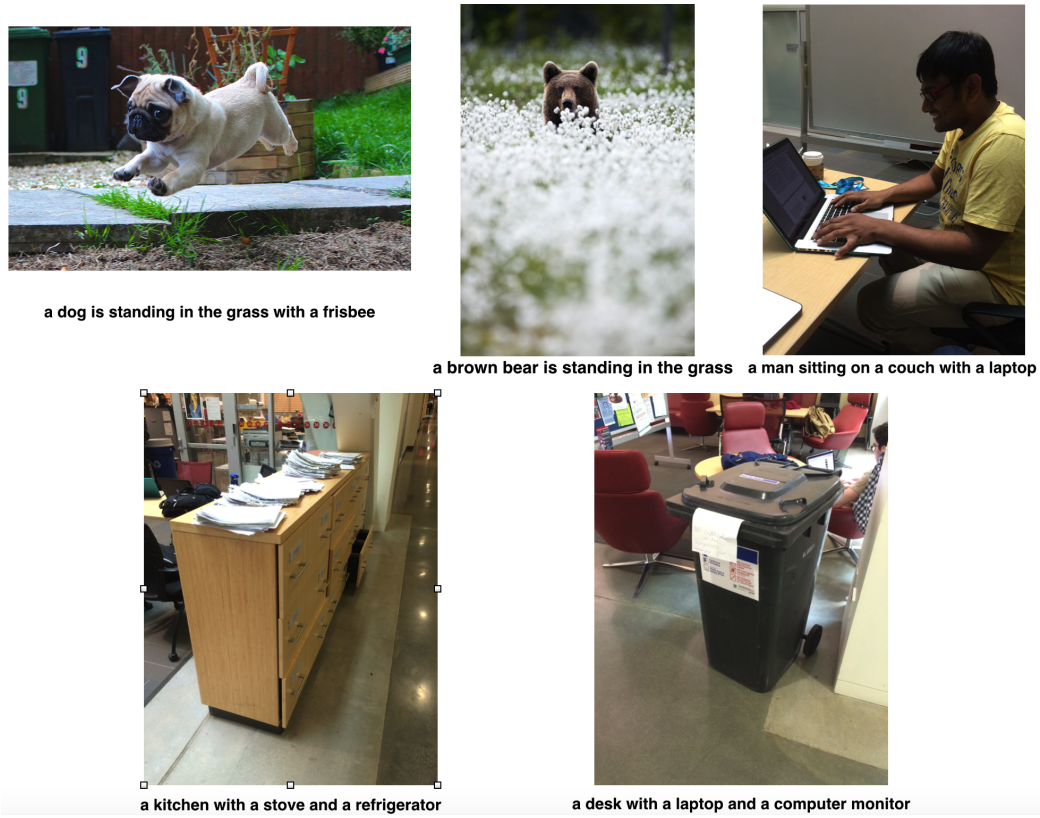

Figure 2: Test Images and their descriptions

Our first results consist of running the CNN model and model parameters from [7]. This results in the 4096-D vector that is fed into the LSTM network trained on MS-COCO dataset. The results are shown below for 5 test images (3 of which we took ourselves and 2 provide in the dataset). The evaluation metric used is a BLEU score (for unigram, bigram, trigram, and 4-gram) for each image and its result is averaged over the test set.

B-1: 59.1, **B-2**: 38.9, **B-3**: 16.5, **B-4**: 0.0

As expected the 4-gram score averages 0 as it is tougher to find matches between the candidate prediction and the reference sentences.

4.2 Semantic Search

Below we employed two methods to relate search queries to indexed images. We assessed both with a mean rank score over 1000 test images in the Flickr8k dataset. Also, since both the average GloVe vector approach and RNN approach have embedded non-linearities in the high-dimensional space, our visualizations employ t-SNE to better maintain this in 2D space (PCA is too lossy).

4.2.1 Average GloVe Vectors

Query	Mean Rank
A boy in his blue swim shorts at the beach .	107.3
A blond woman in a blue shirt appears to wait for a ride .	-
A lady and a man with no shirt sit on a dock .	35.7
A snowboarder takes a rest on the mountainside .	234.3
This man is smiling very big at the camera .	313.2

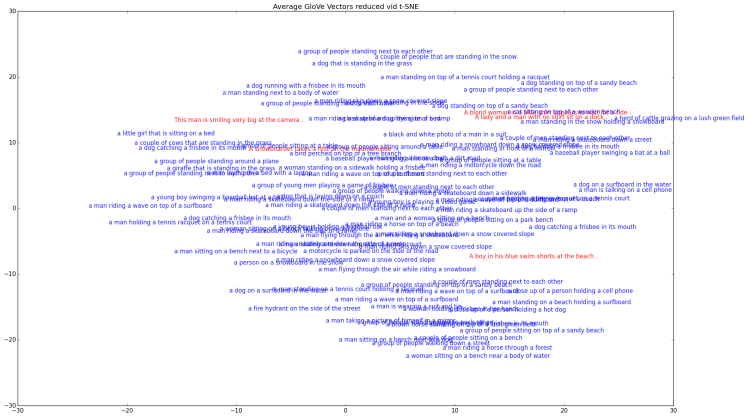

Figure 3: Semantic vector space via 300D GloVe vector average reduced to 2D via TSNE. red = queries, blue = generated captions

Since this representation equally weights all words in the sentence, the clustering tends to get confused very easily by smaller words and related suffixes (e.g. gerunds like sitting/smiling and 'is', 'the', etc). The result is a poor grouping of elements. The final mean rank scores are highly varied - for those queries with all specific and important words like 'A lady and a man with no shirt sit on a dock' the score is very good, but for generic queries like 'The man is smiling very big at the camera' the ambiguity of the words results in a very poor rank.

4.2.2 Recurrent Neural Net Vector Outputs

Query	Mean Rank
A boy in his blue swim shorts at the beach .	170.4
A blond woman in a blue shirt appears to wait for a ride .	92.8
A lady and a man with no shirt sit on a dock .	-
A snowboarder takes a rest on the mountainside .	74.3
This man is smiling very big at the camera .	203.5

The Recurrent Neural Net better captures local semantic relations between queries. While this is clear in the visualization, the mean rank scores do not match as our queries ended up in a similar

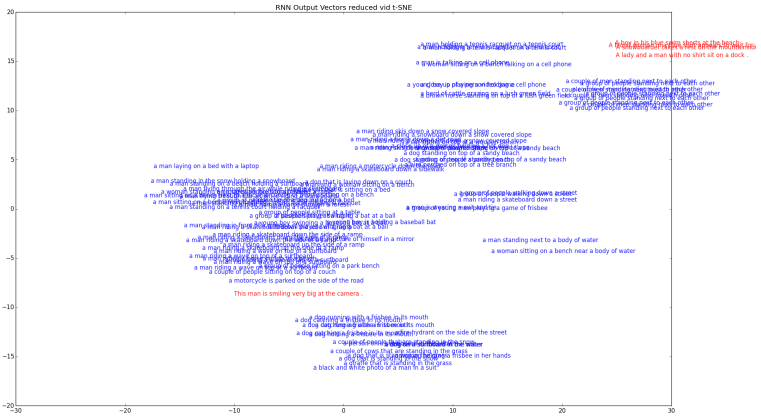

Figure 4: Semantic vector space via 100D RNN output vectors average reduced to 2D via TSNE. red = queries, blue = generated captions

semantic space. This resulted in the algorithm having issues distinguishing differences between the 4 of the queries. Nevertheless, the average mean rank is lower than the average GloVe vector results.

4.3 Error Analysis

Below we address systematic errors in both methodologies.

Figure 5: Close up on GloVe vector graph.

As seen above 'A boy in his blue swim shorts at the beach' is a classic query that incorporates many different semantic concepts. While the overall meaning is tied to the concept of 'beach', the GloVe vector method incorrectly embeds it far from any cluster since 'his', 'blue' and 'shorts' all are part of other concepts. This results in ambiguity in the final ranking since the embedding is roughly equidistant from multiple concept clusters (e.g. action-related queries above and beach-related once below).

The RNN captures many concepts much better than the previous method. Some high level concepts are well clustered like 'snow', 'tennis', 'next to each other', etc. However, the query sentences

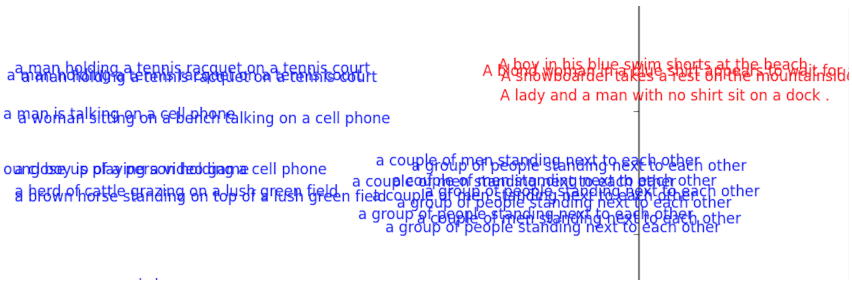

Figure 6: Close up on RNN Vector graph.

seem to cluster as well despite having no clear conceptual similarity. Since the training set has both lowercase and capital letters and capital letters are much less prevalent, the queries end up far away from many of the generated captions. Despite this error, we still see improvement over the previous average GloVe vector approach.

5 Conclusion

In this work, we learnt and built upon the model for generating natural language descriptions about the image. Having captured the semantics in the image, we address the image retrieval problem. For future ideas, we can approach this problem by applying other neural networks such as Recursive Neural Networks in order to embed the generated sentence and the query into same space.

References

[1] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014.

[2] Sermanet, Pierre, et al. "Overfeat: Integrated recognition, localization and detection using convolutional networks." arXiv preprint arXiv:1312.6229 (2013).

[3] D. Turney and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37:141188.

[4]Grounded Compositional Semantics for Finding and Describing Images with Sentences

[5]Turney, Peter D., and Patrick Pantel. "From frequency to meaning: Vector space models of semantics." Journal of artificial intelligence research 37.1 (2010): 141-188.

[6]Socher, Richard, and Li Fei-Fei. "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[7]Socher, Richard, et al. "Zero-shot learning through cross-modal transfer." Advances in neural information processing systems. 2013.

[8] Duygulu, Pinar, et al. "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary." Computer VisionECCV 2002. Springer Berlin Heidelberg, 2002. 97-112.

[9] Farhadi, Ali, et al. "Every picture tells a story: Generating sentences from images." Computer VisionECCV 2010. Springer Berlin Heidelberg, 2010. 15-29.

[10] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." arXiv preprint arXiv:1412.2306 (2014).

[11] Simonyan, Karen, et al. "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv preprint arXiv:1409.1556 (2015).

[12] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." arXiv preprint arXiv:1411.4555 (2014).