Image Captioning using Visual Attention

Anadi Chaman(12105) and K.V.Sameer Raja(12332)

October 4, 2015

1 Objective

This project aims at generating captions for images using neural language models. There has been a substantial increase in number of proposed models for image captioning task since neural language models and convolutional neural networks(CNN) became popular. Our project has its base on one of such works, which uses a variant of Recurrent neural network coupled with a CNN. We intend to enhance this model by making subtle changes to the architecture and using phrases as elementary units instead of words, which may lead to better semantic and syntactical captions.

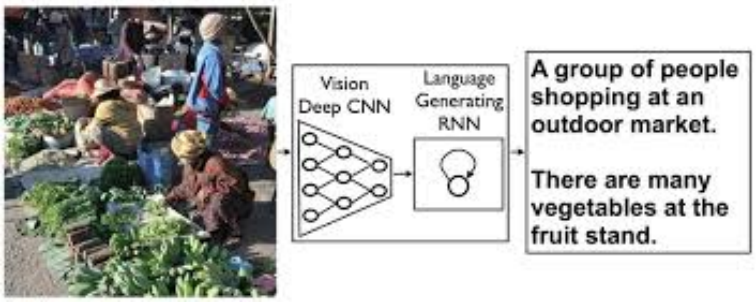

Figure 1: Neural language model
src : img.techxplore.com/newman/gfx/news/hires/2014/1-imagedescrip.png

2 Motivation

Generating captions for images is a vital task relevant to the area of both Computer Vision and Natural Language Processing. Mimicking the human ability of providing descriptions for images by a machine is itself a remarkable step along the line of Artificial Intelligence. The main challenge of this task is to capture how objects relate to each other in the image and to express them in a natural language (like English).Traditionally, computer systems have been using pre-defined templates for generating text descriptions for images. However,

this approach does not provide sufficient variety required for generating lexically rich text descriptions. This shortcoming has been suppressed with the increased efficiency of neural networks. Many state of art models use neural networks for generating captions by taking image as input and predicting next lexical unit in the output sentence.

3 Previous work

Karpathy et.al [1] developed a model that generated text descriptions for images based on labels in the form of a set of sentences and images. They use multi-modal embeddings to align images and text based on a ranking model they proposed. Their model was evaluated on both full frame and region level experiments and it was found that their Multimodal Recurrent Neural Net architecture outperformed retrieval baselines.

Kiros et.al [2] proposed a neural network based approach for generating text descriptions from image and for image retrieval from text. They used a Multimodal log bilinear model that was biased by the features of the input image.

Vinyals et.al [4] have proposed a model which uses CNN to generate image features which are passed to an LSTM network. Using word embeddings and feature vectors as intermediaries in determining gate values, they used beam search on obtain multinomial distribution to generate image captions.

4 Dataset

For training and validation purposes we would be using either Microsoft-COCO data set, which has 82,783 images with 5 captions per image , or IAPR TC-12 dataset, having 20,000 multi object images and captions in English, German and another Spanish languages. In additional to these, we would be using Flickr8k and Flickr30k data sets which contain 8000 and 30000 images respectively from flickr website and each image has 5 different captions associated with it.

5 Methodology

In this project we try to enhance the work of Xu et.al [5] by using phrases instead of words to generate captions.

- The first task would be to obtain phrases for a given input sentence using SENNA software. We are banking on the statistics reported by Remi Lebret in his paper [3] which says that identifying Noun phrases (NP) , Verb phrases (VP) and Prepositional phrases (PP) suffices to capture the whole caption in flickr8k, flickr30k and MS COCO datasets.

- Embeddings for above obtained phrases are calculated by taking element wise summation of vectors for word in corresponding phrases. The corpus

for training word vectors is generated in such a way that every word in caption dataset is seen in training corpus.

- CNN is used to obtain feature vectors from images. We are trying to use pre-trained CNN for generating feature vectors since the number of model parameters in our architecture are large due to LSTM RNN.
- Above obtained image feature vectors and phrase embeddings are passed as input to LSTM network. We use attention models for extracting context vector which is passed as an additional input to LSTM network.
- Finally beam search is applied on the output of LSTM (which gives a multinomial distribution over vocabulary) to generate sentences in image captions.

Evalutaion : We intend to report accuracies using METEOR(Metric for Evaluation of Translation with Explicit ORdering) which scores machine translation hypotheses by aligning them to one or more reference translations and is designed to overcome some of the shortcomings of BLUE Metric.

References

[1] Karpathy, A., and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306* (2014).

[2] Kiros, R., Salakhutdinov, R., and Zemel, R. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (2014), pp. 595–603.

[3] Lebret, R., Pinheiro, P. O., and Collobert, R. Phrase-based image captioning. *arXiv preprint arXiv:1502.03671* (2015).

[4] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555* (2014).

[5] Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044* (2015).