Visual Attention based Image Captioning

Anadi Chaman(12105), K.V.Sameer Raja(12332)
Dr. Amitabha Mukherjee
Indian Institute of Technology, Kanpur

Introduction

In this project, we worked to generate descriptive captions for images using neural language models. Our work is a variant of the CNN-LSTM architecture based on visual attention models proposed by Kelvin Xu et al. We have incorporated the use of phrase embeddings for generating captions, and compared the performance obtained here, with that from word embeddings.

Previous Work

- Ryan Kiros[3] proposed a neural network based caption generating model. It used Multi-modal log bilinear model that was biased by the features obtained from input image.
- Andrej Karpathy [1] developed a model that uses multi-modal embeddings to align images features and text based on a ranking model. Their Multimodal neural network architecture was found to outperform retrieval baselines.
- Oriol Vinyals[5] proposed a CNN-LSTM architecture, where they used feature vectors obtained from CNN, and word embeddings to determine LSTM gate values. Beam search was finally used at the output to generate captions.

Architecture

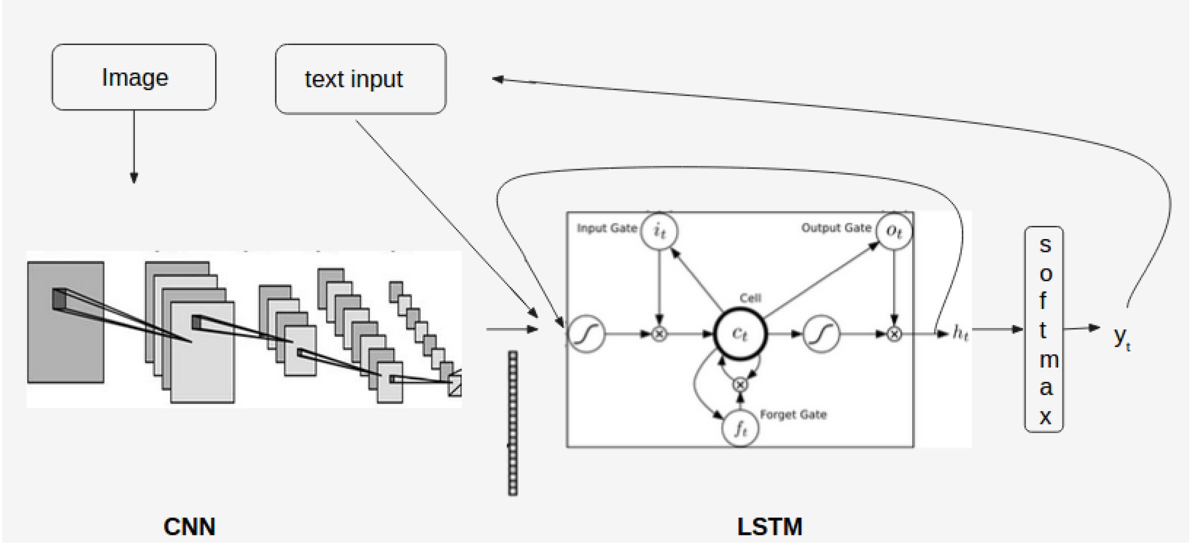

Figure 1: System flow Diagram

Convolution Neural Network

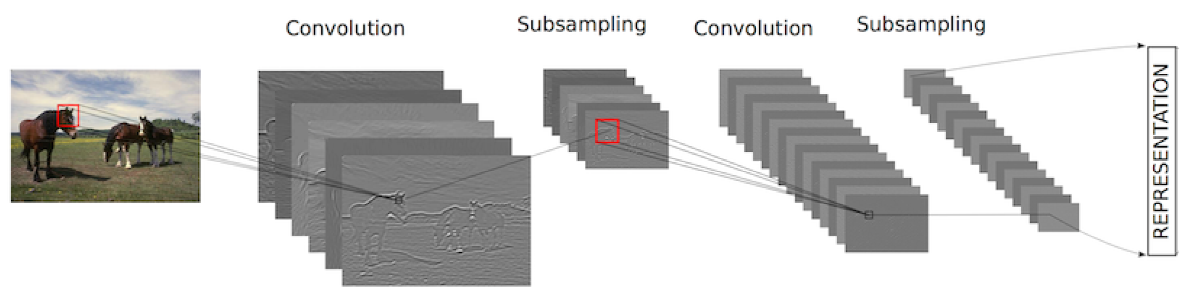

Figure 2: Feature map extraction using CNN [4]

- CNNs are feedforward type of neural networks that convolute and sub-sample an image at successive stages to yield feature maps.
- We pass images of size 24×24 as an input to a pre-trained CNN , where they are convolved with 4 different filters to yield 4 sub-images. This process was continued till 512 feature maps of size 14×14 each were yielded.
- Looking across the images, we got 196 different annotation vectors each of dimensionality 512.

Attention Model

- Given the annotation vectors, a context vector is generated which points to different portions of the given image. Mathematically:

$$\hat{z}_t = \sum_i s_{t,i}\boldsymbol{a_i}$$

- We have employed a 'Hard attention' model which is a stochastic mechanism. The weights $s_{t,i}$ are sampled from a multinuolli(α_i) distribution
- These α_i's are learned using a network with previous hidden state (h_{t-1}) and annotation vectors as input.
- New objective function accounting for sampling is given by :

$$L_s = \sum_s p(s|\boldsymbol{a}) log(p(y|s,\boldsymbol{a}))$$
$$\leq log \sum_s p(s|\boldsymbol{a})p(\boldsymbol{y}|s,\boldsymbol{a}) = log(p(\boldsymbol{y}|\boldsymbol{a}))$$

LSTM RNN

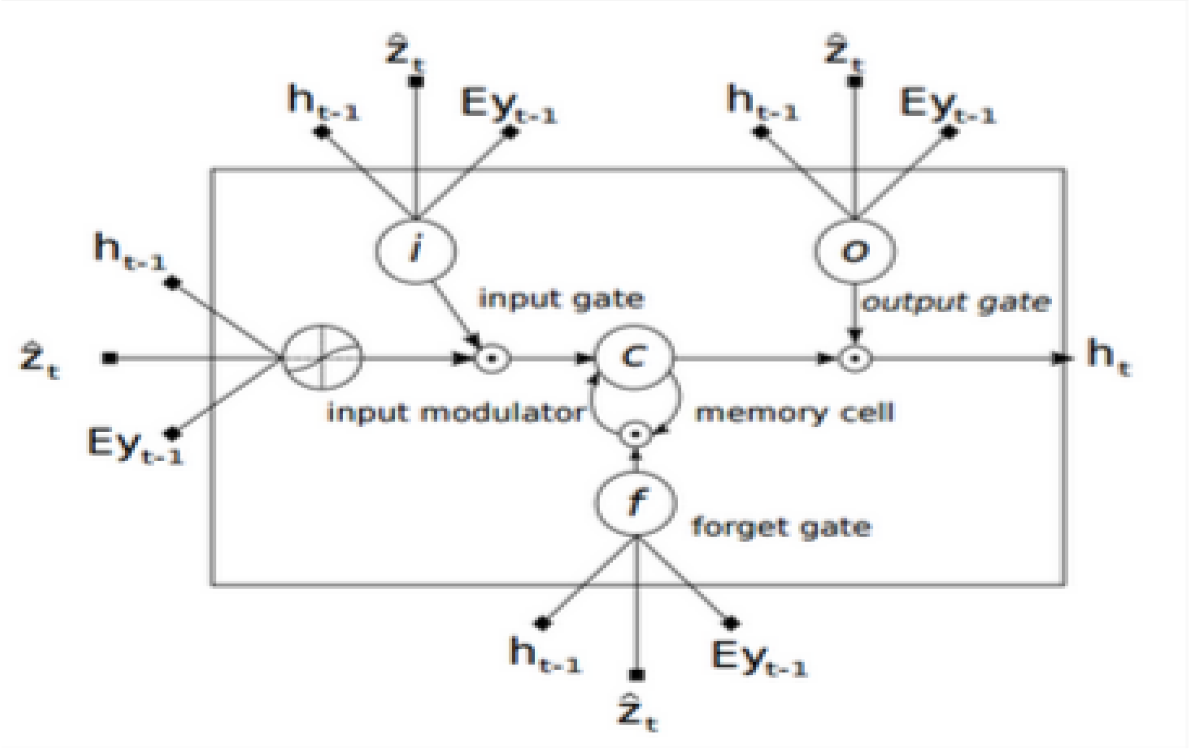

Figure 3: An LSTM cell [6]

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ g_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{pmatrix} T_{D+m+n,n} \begin{pmatrix} Ey_{t-1} \\ h_{t-1} \\ \hat{z}_t \end{pmatrix}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ g_t$$
$$h_t = o_t \circ tanh(c_t)$$
$$p(y_t|a, y_{t-1}) \propto exp(L_o(Ey_{t-1} + L_h h_t + L_z \hat{z}_t))$$

Approach

- Senna software has been used to obtain phrases from captions, available as a part of training data.
- Embedding of a phrase is obtained by taking the sum of embeddings of words belonging to that phrase.
- For generating annotation vectors, we used a pre-trained model of CNN namely Oxford VGGnet trained on Imagenet Dataset, and are using an LSTM architecture
- Due to large vocabulary size of phrases, we found the ones with the highest frequency, and replaced the rest with UNK symbol). This reduced our vocabulary to 10000 phrases.

Dataset and Resources

- Flickr8k dataset contains 8000 images and each image has 5 captions describing it, summing to 40000 caption and image pairs.
- We used 30000 captions for training, 5000 captions for validation and 5000 captions for testing purposes respectively.
- Word embeddings used for obtaining phrase embeddings are derived from pre-trained word2vec model trained on google news corpus.

Results

Input	vocabulary	METEOR
phrases	10000	0.062
phrases(pre embeddings)	10000	0.06
phrases	36220	0.041
words[2]	9630	0.067
words(beam search)	9630	0.089

Conclusions

- The marginal decrease in accuracy when using phrases is due to replacement of large number of phrases in the training data with UNK symbol.
- If an efficient phrase vocabulary reduction technique is employed, we hope that phrase input will have better accuracy compared to word input.

References

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

[2] kelvinxu.
arctic-captions.
https://github.com/kelvinxu/arctic-captions.

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

[4] A. Sironi.
Bigger Faster Convolutional Neural Networks.
cvlabwww.epfl.ch/projects/bigger_faster_convolutional_neural_networks.html.

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

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