IE643: Deep Learning : Theory and Practice **Jul-Nov 2018**

Project Report : Automated Image Captioning

Student Name: Gaurav Didwania, Aman Jain, Sushil Khyalia Roll No: 16005020,160050034,160050035

Abstract

This project report contains the details on our project which aims to create a deep neural network model which will be able automatically generate captions for a given input image. We describe the deep learning tools involved in our project. In particular, we explain about the deep neural network architectures, the loss functions and activation functions used and the training algorithm used.

1 Introduction

In the past few years, the problem of generating descriptive sentences automatically for images has led to a rising interest in natural language processing and computer vision research. Image captioning is a fundamental task which requires semantic understanding of images and the ability of generating description sentences with proper and correct structure. In this project, we propose a hybrid system employing the use of multilayer Convolutional Neural Network (CNN) to generate vocabulary describing the images and a Long Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. The Bleu metric is an algorithm for evaluating the performance of a machine translation system by grading the quality of text translated from one natural language to another. The performance of the proposed model is evaluated using standard evaluation matrices.

2 Literature Survey

Our project draws inspiration from a closely related work by Vinyals et al. [4]. In their paper [4], the authors tried to tackle the problem of captioning by combining the then state-of-art sub-networks for vision and language models. It had a LSTM model which combined with a CNN image embedder and word embeddings (as shown in figure 5).

Also, in a very similar work by Xu et al. [5] had a very similar structure of a CNN encoder followed by a LSTM decoder network but it used something called as attention based caption generation. Instead of using a fully connected layer as it's final layer the CNN network had a lower convolutional layer at the end of the network. This allowed the decoder to selectively focus on certain parts of an image by selecting a subset of all the feature vectors. The inclusion of attention gave the advantage of ability to visualize what the model actually "saw".

In another work by Karpathy and Fei-Fei [2] they tried to generate a model which was able to describe different regions of the input image. They used a CNN to represent regions of image and used a Bidirectional Recurrent Neural Network (BRNN) for representing sentences and finally used a Multimodal Recurrent Neural Network for generating descriptions.

3 Proposed project

We planned to develop an end to end model that takes as input an image and returns a sentence describing the image which we have completed. The sentence generation is inspired from various advances in machine translations that take as input a sentence and outputs another describing sentence in another language. In our case the input sentence is replaced by the image feature vector. This image feature vector is obtained as an output to a CNN which is then fed into the LSTM to generate output sentences.
There are certain key points in this implementation according to the implementation of show and tell [1] paper:

1) Our model is trained to maximize the likelihood p(S|I) of producing a target sequence of words $S = \{S_1, S_2, ...\}$ where each word S_t comes from a given dictionary, that describes the image adequately.

2) Beam Search : Iteratively consider the set of the k best sentences up to time t as candidates to generate sentences of size $t+1$, and keep only the resulting best k of them. As the number of nodes to expand from is fixed, this algorithm is space-efficient and allows more potential candidates than a best-first search.

3) Instead of VGGNet object identifier of the keras, we will use ResNet (ResidualNet). Our motivation to replace VGGNet with ResNet are the results of the image classification results as given in figure 1. These are the results for the imagenet classification task over the years.

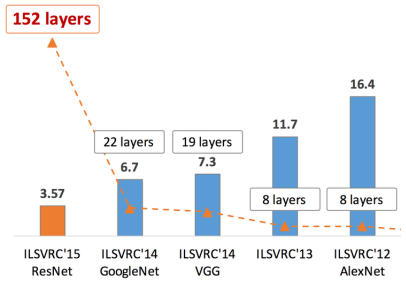

Figure 1: Error rates on ImageNet classification task

Also it has been empirically observed from these results and numerous others, that ResNet can encode better image features [6].
A brief overlook of unrolled version (unrolling with respect to time) of our model is summarized in the following image:

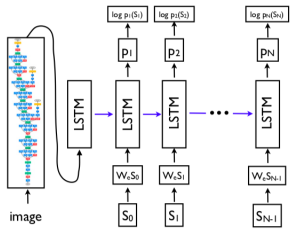

Figure 2: LSTM Network without attention

We modified the above model to add an attention layer before our Decoder (LSTM) which is based on principle of giving weights to the feature vectors of the image provided by the encoder before passing them

as input to decoder. This enhances the network performance.

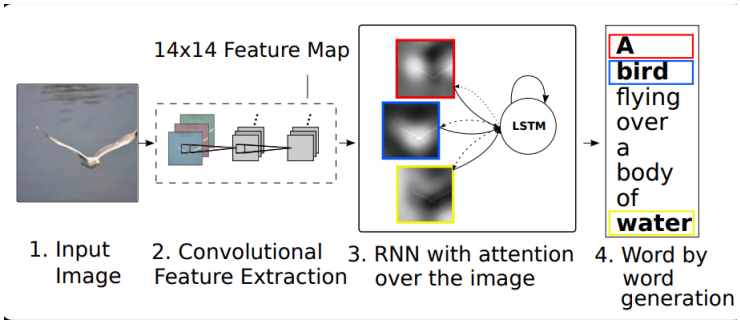

Figure 3: Complete Encoder-Decoder model with attention

4 Results

After running 5 epochs we were able to get decent captions on majority of the images. The image given below shows example of one bad and one good caption generated by our neural network. We have used a MS-COCO validation set (2014) to check robustness of our network.

```
Epoch: 4 - Step: 0 - Mini Eval Loss: 3.40189528465271
Sample: a man standing next to a red fire hydrant .
Target: a plastic yellow cup with two tooth brushes and some toothpaste
Epoch: 4 - Step: 100 - Mini Eval Loss: 2.7256839275360107
Sample: a man standing on a skateboard in the middle of a street .
Target: male skateboarder at the peak of a stunt in a public skate park .
```

Figure 4: Sample caption generated vs an actual caption of an image in Validation set

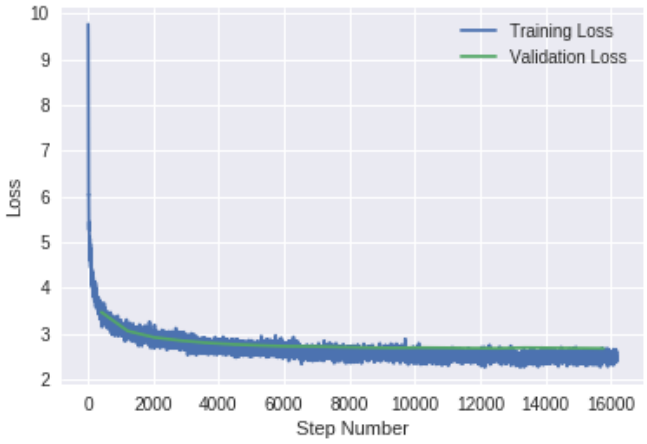

Figure 5: Loss vs Time steps

5 Possible improvements over our work

Taking inspiration from [6],one can modify the algorithm according to [4], i.e. modifying the standard long short-term memory (LSTM)based decoder by introducing a gate function to reduce the search scope of the vocabulary for any given image, which is termed the word gate decoder. The reason for this modification is to reduce the large action space which otherwise makes it difficult for the model to accurately predict the current word.

One better and different approach to image captioning problem is using reinforcement learning as explored in [6]

6 Conclusion

Accurate image captioning with the use of multimodal neural networks has been a hot topic in the field of Deep Learning. It has wide number of applications such as for automating the work of a person who interprets the image manually and for visually impaired people as well as to extract insights from the images or videos.

References

[1] kth-sml project, `https://github.com/amundv/kth-sml-project`, 2017.

[2] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, 2017.

[3] Vishal Motwani Vikram Mullachery. Image captioning. 2015.

[4] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015.

[6] Xinxin Zhu, Lixiang Li, Jing Liu, Longteng Guo, Zhiwei Fang, Haipeng Peng, and Xinxin Niu. Image captioning with word gate and adaptive self-critical learning. *Applied Sciences*, 8(6):909, Jan 2018.