Image captioning using attention

Aadishesh Sharma
CSE
aadishes@buffalo.edu

Sreeram Kashyap
Industrial Engineering
kashyapc@buffalo.edu

Nitesh Shantha Kumar
Robotics
nshantha@buffalo.edu

Leena Patil
CSE
leenaman@buffalo.edu

Abstract

Generating descriptions for images is among the most interesting applications of DNN it is aimed at bridging the gap between computer vision and natural language processing. There has been considerable research on this topic which explored different models for the same. All of them utilize an encoder and a decoder to read the image and generate captions for it. An improvement of interest in the research is the application of attention mechanism, originally proposed for neural translation. Substantial increase in accuracy of output sentences has been observed across many datasets with application of this mechanism. In this study we further explore this idea and make use of different models in the literature to improve the accuracy of this model. Our work involves amalgamation of Convolutional Neural Networks and Recurrent Neural Networks. The CNN is used as an encoder and RNN as a decoder with implementation of attention mechanism. For the evaluation of results of our model, BLEU score has been used.

1 Introduction

There have been recent advancements in the field of generating automatic captions for images. An image captioning model generally finds a relationship between the objects contained in an image with its suitable sentence structure in a natural language such as English. The current models[1] [2] in this domain take leverage of attention mechanism to maxim ize the output which enables to store and report the information and relationship between noticeable or important features in an image. Generally, attention mechanism incorporate two types: 'hard' and 'soft' attention. However, in our work, soft attention mechanism is applied. Attention models can be understood by drawing parallels to the biological phenomenon of focusing on a smaller part rather than the whole scene, thus attention.

Our project focuses on image-to-sentence generation which bridges the gap between two different domains i.e. computer vision and natural language process. We intend to leverage Natural Language Processing techniques to understand images. Convolutional Neural Networks are used to extract features from an image and used as an encoder and then Recurrent Neural Networks used as a decoder which does language modelling upto the word level. Additionally, attention mechanism has also been applied which recognizes what a word is referring to in an image and draws the relationship between clusters in an image. Attention mechanism helps in selecting relevant information from a large input of data coming from CNN encoder. Usually, RNN models are computationally expensive to train and by using attention which focuses on relevant parts only it addresses this problem as well.

There has been a lot of work done[3] [4] in the domain of attention models and image captioning. However, most of the models use LSTM for decoding but in our work, we have implemented GRU. The main reason behind using GRU is that it is considered to have performance at par with LSTM and is less computationally expensive.

2 Experiments

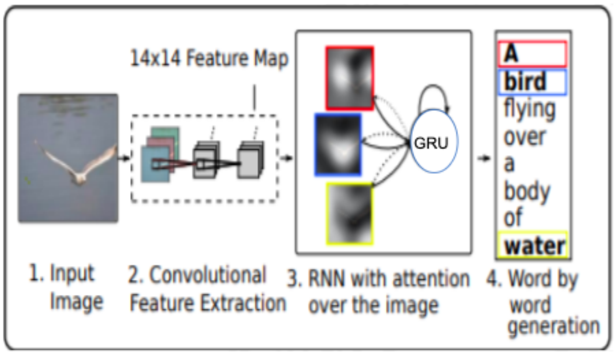

14x14 Feature Map
GRU
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Figure 1: Image Captioning Model

2.1 Datasets

There are many datasets available but for this project, we chose Common Objects in Context dataset provided by Microsoft commonly known as MSCOCO dataset which is also widely used in this domain. The dataset contains more than 55,000 images splitted into training, validation and testing sets with each image having corresponding annotations. Each annotation has around

5-20 words and there are multiple captions for each image.

3 Model

In this section we describe the model employed for the different tasks in generating captions. Firstly we use the encoder-decoder model used in [5]. Then we explain the attention mechanism employs for learning the semantic interactions.

3.1 Architecture

The encoder Reads the image and reads the different features present in the image. It uses convolutional neural networks for the purpose. In this study we mainly explored different models of encoder models. The decoder is used to learn the semantic interactions among the features extracted by the encoder.
The most important work done in this paper is the implementation of bahdenue attention for generating captions first proposed in [6]. The main issue with encoder-decoder models is that the decoder takes the entire vector provided by the encoder at once and processes. By using attention the model will takes in each word and tries to relate it to one of the visual features found by the encoder. This way it works similar to attention in humans. Instead of encoding the input sequence into a single fixed context vector, the attention model develops a context vector that is filtered specifically for each output time step. In the work done in [1] they proposed two attention models namely soft attention and hard attention. In soft attention the context vector from the encoder is used to calculate the probability od each word being related to a particular feature in the image. In hard or global attention crisp decisions are made about elements in the context vector for each word. An illustration of attention mechanism is shown in the Figure 2

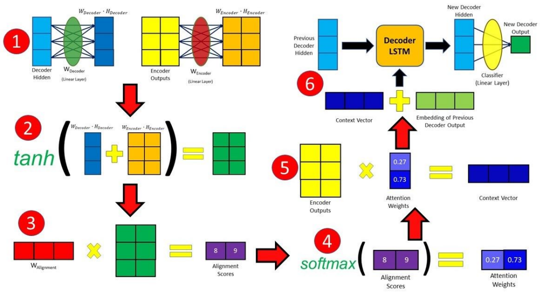


Figure 2: Attention mechanism shows considerable increase in accuracy of image captioning over the earlier models

3.2 Models used

We explored previous research for models employed for capturing the features in an image. The largest work on images was on imagenet dataset. Since its release in 2014 there has been considerable research on this dataset. Many customized architectures have been proposed which vary in accuracy and time complexity. We considered two main factors for selecting the models. Firstly, generalizable i.e. how good the model can be generalized for other cases than imagenet dataset. Secondly, time complexity i.e. how long it takes to train for accpetable accuracy.In the current implementation[5] inceptionv3 has been used because of the advantages it presents over many other models along with acceptably high accuracy.But we wanted to compare the performance of other models and evaluate results. From a spectrum of models available under imagenet research we selected three models described below Firstly we wanted to test a relatively simple and fast architecture that improvised over the first accurate model prsented under imagenet i.e. alexnet[7]. So we chose VGG16 proposed by oxford research group. The VGG16[8] shows a simple architecture comprising of conv2d layers with reducing sizes and intermediate pooling layers. It was one of the simplest models which depicted considerable accuracy in predictions.
The second model selected is Deep residual learning model also known as ResNet[9] proposed by microsoft research group. It uses a network in network architecture to show a relatively high accuracy. One of the biggest advantages of resnet is that even though the network is very deep, training takes very less time and it avoids negative outcomes, and shows an increase in accuracy. The resnet architecture is shown in fig 3.

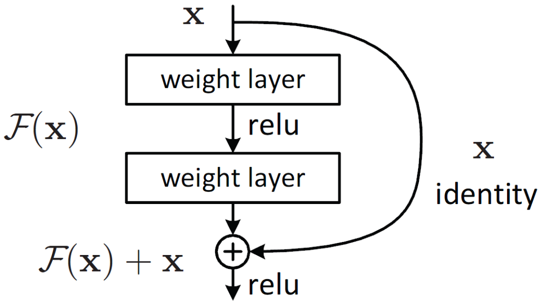


Figure 3: Residual learning mechanism

The last model used is NASNet[10].NASNet is the latest architecture provided by google brain research group. It gives a very good generalization error compared to any other model employed on imagenet. The main proposition of this research is the use of scheduled droppath mechanism. DropPath mechanism is a regularization method which drop random cells in the path with a fixed probability throughout the training. This idea is enhanced in scheduled droppath mechanism by gradually increasing the probability of dropping the cells in the path. This helps reduce overfitting and increases generalizing ability. NasNet architecture consists of two main modules namely normal cell and and reduction cell. A normal cell is a group of convolution operations that gives feature map of the same size as input. A reduction cell reduces the size of the feature maps by a factor of 2 along each dimension compared to the input.NASNet uses Normal and reduction cells stacked in a custom combination to achieve high accuracy. The main disadvantage observed with this model is that it takes around 4-5 times to train the model compared to other models.
Two main models were considered for the decoders-GRU and LSTM. We explored both the models and found out that GRU was a better application, because

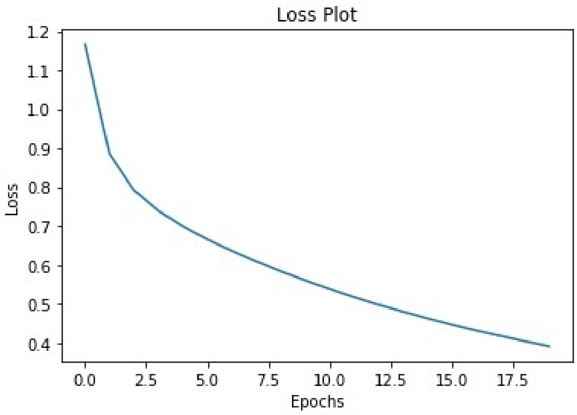
of its simple design and faster processing. GRUs are simpler and thus easier to modify, for example adding new gates in case of additional input to the network.

3.3 Metrics

The evaluation of image to sentence models is generally performed using several metrics such as BLEU, METEOR, ROUGE or CIDER, all of which mainly measure the word overlap between generated and reference captions[11]. However, we chose to use Bilingual Evaluation Understudy also known as BLEU scores as it is mostly used in academia to evaluate models and it's easy for us to compare our model with the work of others. In BLEU, the score is calculated between 0 to 1. The higher the score, the better it is. In simple terms, BLEU metric uses precision i.e. how many words are appeared in the original annotations generated by the model. However, BLEU uses a modified precision i.e. for each occurence of a n-gram in ground truth y can account for only one occurence in x.

4 Results

Since, the dataset is huge and for evaluation we had 6000 images and it's not possible to iterate through each image showing BLEU score, we averaged out sentence BLEU score for the total 6000 images. The results we got are promising. In most of the architectures, LSTM is used as a decoder but we used GRU considering it to be less computationally expensive and performance at par with LSTMs. We implemented three different models namely, VGG16, RESNET-50 and NASNET for encoding with GRU used as a decoder. Image features were extracted using pre trained CNN models. For the optimization, Adam optimizer was used. A total of 5000 most occurring words were used as vocabulary and the rest of the words were replaced with "unknown" token. All tokens less than the maximum length of caption were padded with 0 to make the length consistent. Each model was trained for different epochs and generated different loss values. For VGG16 model, 30 epochs were used, 40 for RESNET-50 and 175 epochs for NASNET. The loss plots are displayed below:



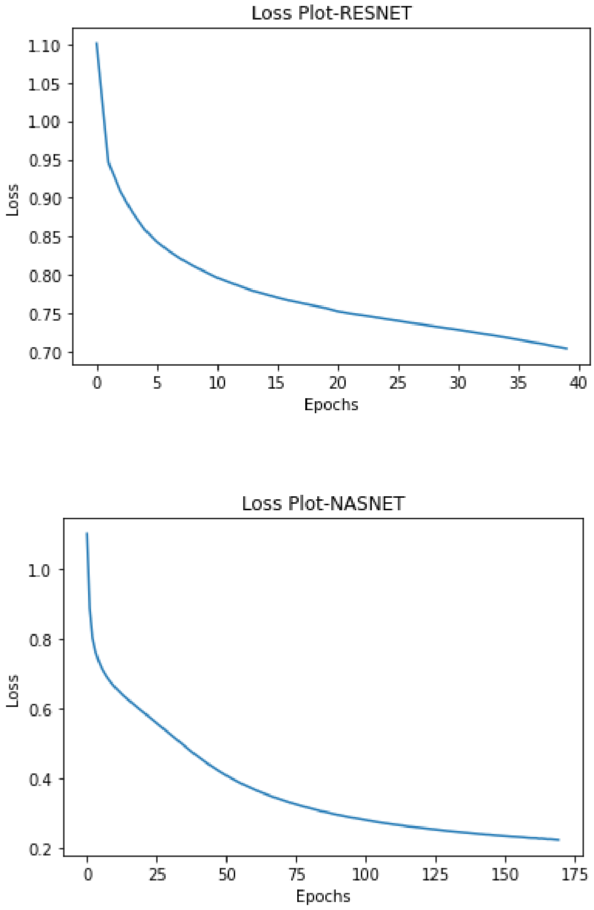


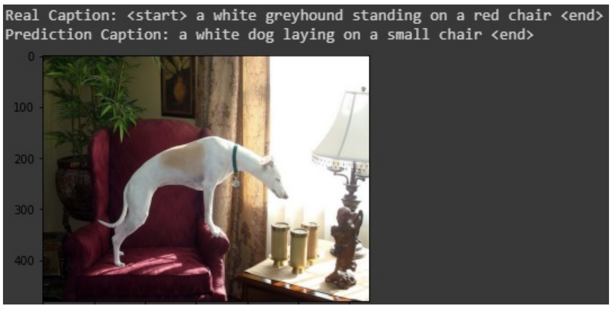
Figure 4: Loss Plots of our models

For evaluation, we used BLEU score. We can observe that for BLEU-4, all of our models certainly did better than other models.

Architecture	*BLEU-4*
Google NIC	24.6
Soft Attention	24.3
Hard Attention	25.0
Our Model 1-VGG	**.31**
Our Model 2-ResNet	**.34**
Our Model 3-NasNet	**.351**

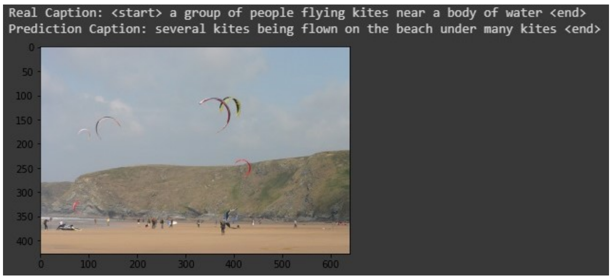
Table 1: Comparison of existing and our models

Following below are the outputs of some of the captions generated by our model:



```
Real Caption: <start> a white greyhound standing on a red chair <end>
Prediction Caption: a white dog laying on a small chair <end>
```

Figure 5: Output of our model



```
Real Caption: <start> a group of people flying kites near a body of water <end>
Prediction Caption: several kites being flown on the beach under many kites <end>
```

Figure 6: Output of our model

5 Conclusion

Our work gives a useful insight into the working of image captioning problem. Firstly, we mainly focused on studying the encoder part. We kept the decoder architecture undisturbed for further studies. We found that selecting a suitable architecture for the encoder plays a crucial role in determining the accuracy of the final output. In this study the three models that we selected broadly covered the entire spectrum of the models available in literature. We would like to point out an anomaly that we noted with the use of NASNet. With this model we recorded a comparatively lower net-loss during training. So, we expected to see a considerable increase in accuracy as well as bleu score. But contradictory to our hypothesis we did not observe any measurably high increase in accuracy. It was not very high compared to the other models. Considering the fact that it took 4 times longer to train this model, we propose that it is not a very good contender in this scenario of image captioning. ResNet model showed less BLEU score than NASNet but took far less time to train, almost equal to VGG16. This is a notable result of this study.

The decoder part needs to be studied further. Particularly, application of LSTM and bi-directional LSTM. Further studies are required to compare the combination of LSTM with the above encoders, which could not be covered in this project. LSTM and Bi directional LSTM have shown a better performance in image captioning without attention mechanism in [11] [12]. Further studies need to explore more combinations to find better architectures suitable for generalizing, so as to be used in other scenarios that are yet unexplored.

References

[1] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[2] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge J. Belongie. Learning to evaluate image captioning. *CoRR*, abs/1806.06422, 2018.

[3] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[4] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[10] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

[11] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997, 2016.

[12] Minghai Chen, Guiguang Ding, Sicheng Zhao, Hui Chen, Qiang Liu, and Jungong Han. Reference based lstm for image captioning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.