Visual Attention model for Image Captioning using Deep learning

Dinesh Naik
Information Technology
National Institute of Technology Surathkal
Mangalore, India
din_nknitk.edu.in

Devaguptam Sreegeethi
Information Technology
National Institute of Technology Surathkal
Mangalore, India
sreegeethidev@gmail.com

Lakshmi S Raj
Information Technology
National Institute of Technology Surathkal
Mangalore, India
lakshmi.shiburaj@gmail.com

Kogatam Thanmai
Information Technology
National Institute of Technology Surathkal
Mangalore, India
kthanmai.4@gmail.com

***Abstract*—Caption creation is a challenging artificial intelligence problem where a text description has to be made for a given image. It requires both methods from a computer perspective to understand the content of the image and the language model from the field of natural language processing in order to convert the image into words in the correct order. Recently, deep learning methods have found extensive results in examples of this problem. What is most impressive about these methods is that one end-to-end model can be defined to predict captions, given a picture, instead of requiring complex data processing.In this paper, it discusses about a system that generates contextual description about images using deep learning model and also incorporate an attention model to compare their performance accuracy of image captioning. These models are run on Flicker 8k dataset and uses BLEU score metrics to evaluate their performance. According to the BLEU scores calculated, we conclude that the attention model works better than the merge model for image captioning, this is likely because attention model gives priority to relevant parts of the image while merge model gives equal priority to all parts.**

***Index Terms*—Deep Learning, Merge Model, Attention Model, Image Captioning, Flicker 8k**

I. INTRODUCTION

A large amount of information is stored in the image. Every day large image data is generated on social media and in viewing sites. Deep learning can be used to automatically interpret these images, thus replacing the manual annotations made. This will greatly reduce human error and effort by eliminating the need for human intervention. Production of captions from images has a number of practical advantages, ranging from helping the visually impaired, allowing automatic labeling, saving millions of images uploaded daily to online, app editing recommendations, beneficial to visual assistants of photography, visually impaired, social media, and a few other natural language processing applications[1]. This field brings together high-end models in Natural Language Processing and Computer Vision, two major components in Artificial Intelligence.What is most impressive about these methods is that a end-to-end model can be defined to predict captions, give a picture, instead of requiring complex data editing or a pipeline of specially designed models. One of the challenges is finding a large number of images with ever-increasing related text[2]. However, most of this data is audio and therefore cannot be used directly in the image caption model. In order to train the modeling of image captioning, a large data set with well-defined annotations is required. In this paper, it deploys a system that generates a context description of objects in images.

Encouraged by neural machine translation, many common image captions systems use encoder-decoder frame, in which the embedded image is encoded into a vectors according to the information contained within the image, and then converted to a descriptive text sequence[3]. This coding may contain a single output element CNN vector, or many visual features found in different regions within the image. In the latter case, the regions can be taken with the uniform samples, or guided by an object detector shown to produce improved performance. Although these are the state of the art models, they currently do not apply information about the geographical relationship between the findings, such as the relative area size[4]. This information is usually not relevant to understanding the content within the image, however,and is used by humans in consultation with the physical world. Related position, for example, can help distinguish between a “girl on a horse” and a “girl standing next to a horse”. Assuming that an attention model will help tackle this problem better[5].

In this paper, it is deployed with a Deep Learning model and compare its performance with an attention model to identify the more powerful model. The Deep Learning Model uses CNN as an encoder to get information from the images and RNN is used as a decoder which converts the image description into natural language processing. The attention model we are considering assigns a weight to each feature and these weighted features are sent into an encoder. It also

TABLE I
SUMMARY OF LITERATURE SURVEY

Authors	Methodology	Merits	Limitations	Additional Details
[6]Quanzeng You et al.	Semantic attention architecture which aggregates top-down and bottom-up approaches	Outperforms state-of-art algorithm	Uses global features only	Microsoft COCO and Flickr30K
[7]Ali Farhadi et al.	Architecture that assumes meanings between images and sentences	Finds images from given description	Oversimplified Sentence model	Own dataset (PASCAL 2008)
[8]Girish Kulkarni et al.	Conditional Random Field architecture	More accurate captions compared to previous works	Does not deal with more than 20 object categories	UIUC PASCAL dataset
[9]Desmond Elliott et al.	PROXIMITY, CORPUS, STRUCTURE and PARALLEL models	Visual dependency representations and automatic evaluations	Only parallel model is effective	PASCAL dataset
[10]P. Kuznetsov et al.	Holistic data driven approach	More complex and human-like captions	Didn't evaluate using standard scores	Custom Dataset

uses an RNN model as a decoder.

Key contributions of the project :

- Used two of the most popular image captioning algorithms.
- Comparison of these algorithms based on BLEU score.

The report is divided into different sections as stated here. Section II contains description of studies related to image captioning. Section III comprises the methodology for implementing the said model. Section IV has the discussions and conclusions of the project. Section V contains the comparison of implemented models with previous works. Section VI contains limitations of the project. Section VII contains conclusions of the project.

A. Motivation

Caption creation is a challenging artificial intelligence problem where a text description has to be made for a given image. It requires both methods from a computer perspective to understand the content of the image and the language model from the field of natural language processing in order to convert the image into words in the correct order. Recently, deep learning methods have found extensive results in examples of this problem.Although these are the state of the art models, they currently do not apply information about the geographical relationship between the findings, such as the relative area size.We assuming that an attention model will help tackle this problem better and want to compare its performance to a traditional Deep Learning Model.

B. Problem Statement

We aim to create an Attention based Image Captioning Model and compare its performance to an deep learning-merge model on the Flicker 8k dataset. For the comparison and analysis we use BLEU scores generated on each model.

II. Related Work

In paper[6], they have combined both top-down and bottom-up approaches with the help of semantic attention. This algorithm tries to particularly attend to semantic postulates, approaches and club them into recurrent neural networks(RNN)'s hidden states and outputs. A feedback connects top-down and bottom-up computation using selection and fusion. They have evaluated the model using Microsoft COCO and Flickr30k data-sets.

Bottom-up approaches are classical ones. They generally start with visual objects, words, phrases, concepts and attributes, and club these into sentences using semantic models. [7] and [8] have discerned concepts and used templates to get sentences, while [9] pieces together detected concepts. Whereas, in papers [10] and [11], they have used more powerful language models. Papers [12] and [13] are the newest attempts in this field and they have got results close to the state-of-the-art on different image captioning benchmarks data-sets.

Whereas, top-down approaches are the modern ones, they solve this problem as machine translation problem [14, 15, 16]. Rather than translating into different languages this approach translates visual representation to a language equivalent. They used pre-trained models to get visual representation which have been trained for image classification using large data-sets[17]. Language counterpart is attained through Recurrent Neural Networks(RNN) models. One of the main advantages of this approach is that the entire model can be trained end-to-end. Differences in the approaches lie in the Recurrent Neural Network models used. Top-down approach is considered state-of-this-art in image captioning problem. In [18], they have used encoder/decoder system with CNN and RNN in addition with attention mechanism. They claimed that results were promising.

The Approaches to solve image captioning problem can be classified into 3 categories, namely

1. Retrieval-based image captioning - these models describe images using pre-existing captions from a repository

2. Template-based image captioning - In this method, set of objects and attributes are defined as templates. These models can create grammatically correct captions but not variable length captions

3. Deep Learning based image captioning - In this process, it uses CNN as an encoder to get information from the images and RNN is used as a decoder which converts the image description into natural language processing.

III. Methodology

A. Merge model

In this project, performance of two models for image caption generation is compared. First model is a merge model and second is attention model. For the merge model, the methodology is split into 5 parts. The five parts are as follows Photo caption dataset, prepare photo data, Prepare Text Data, develop Deep Learning Model, evaluate model.

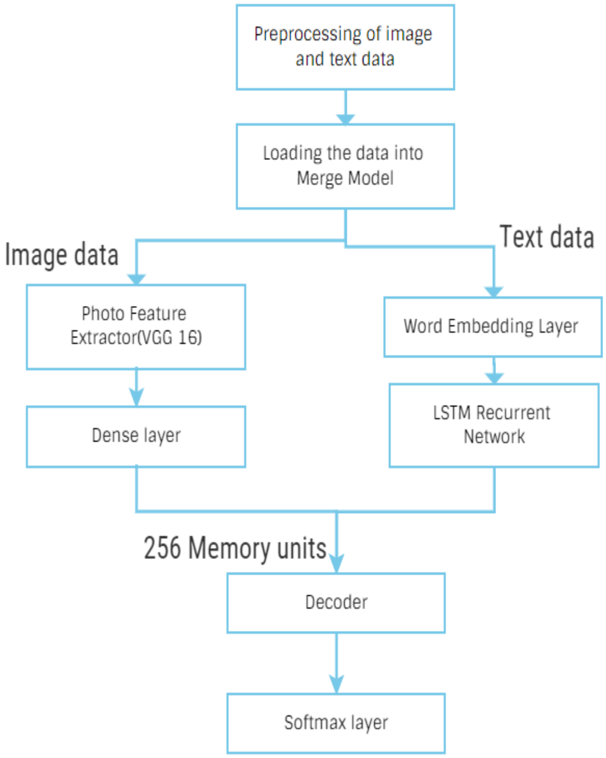


Fig. 1. Proposed Architecture for the Image captioning using Merge Model.

Figure 1 depicts the proposed architecture of Merge model used for image captioning. Here, VGG 16 is used as a feature extractor and Encoder-Decoder architecture for training and generation of the captions.

1) Prepare photo data: A pre-trained model, Visual Geometry Group (VGG) model is used to extract features from the images. This model is can classify images but it takes a lot of time. So instead of classifying the images using VGG, last layer of the model (which is used for classification of images) has been removed to extract the features of images as output of the said model. Extracted features of each image are stored in files and then fed into the merge model as interpretation of given image in the dataset.

2) Prepare Text Data: The Flickr8k dataset contains multiple descriptions for each image. To get better output, text descriptions require minimal cleaning. Each photo is assigned a unique identifier which is used to identify the image's respective description in text file. We cleaned the text in the below given ways to reduce the length of the vocabulary of words :

- Convert all words to lowercase.
- Remove all punctuation, words with length 1,and words that include numbers.

3) Developing a deep learning model: The development of the merge model is split into three parts

- Photo Feature Extractor - Extraction of features from the given images of a dataset is done in this part using 16 layer VGG model (by removing the output layer). The extracted features are fed as input to merge model.
- Sequence Processor - This part deals with handling text input. The sequence processor takes text with fixed length (34 words) as input for word embedding layer which uses a mask to not consider padded values. The output of this layer are used as input for LSTM model which has memory of size 256 units.
- Decoder - In this part of the merge model final prediction of the caption is done by merging the outputs of photo feature extractor and sequence processor which are 256 element vectors and then processing the merged output by dense layer.

VGG takes input photo features in a vector of 4,096 elements. The output of this model are processed by a dense layer to produce a 256 element representation of the photo in photo feature extractor part. The Decoder model feeds the merged output to a dense 256 neuron layer and this layer is connected to output dense layer that makes softmax prediction over entire output text for next word of the sequence.

4) Evaluate model: The model is evaluated by generating descriptions for all photos in the test dataset and by calculating the BLEU scores. Evaluation of mode3l is done by comparing each generated description against given descriptions for the image in test dataset. BLEU scores for 1, 2, 3 and 4 cumulative n-grams have been calculated to assess the performance of the model.

B. Attention model

In attention model, priority is given to most relevant part of the image while generating next word for caption generation. Image caption generation using attention model process is split into 3 parts which are feature extraction of images, prepare text data, defining attention model.

.

1) Feature extraction: A pre-trained Inception model is used for extraction of features for images in the dataset. In this model also , last layer has been removed to get features of the images instead of the classification of images. Inception model contains a series of CNN layers which are used for extraction of relevant features from the images and output a encoded feature map representation.

2) Prepare Text Data: In this part, minimal text cleaning is done for the captions of the images in the dataset. Start sequence and end sequence tokens are at the beginning and end of the sentence to know when the sentence starts and ends while generating captions. Tokenization of the sentences is done by mapping each word to an integer word ID.

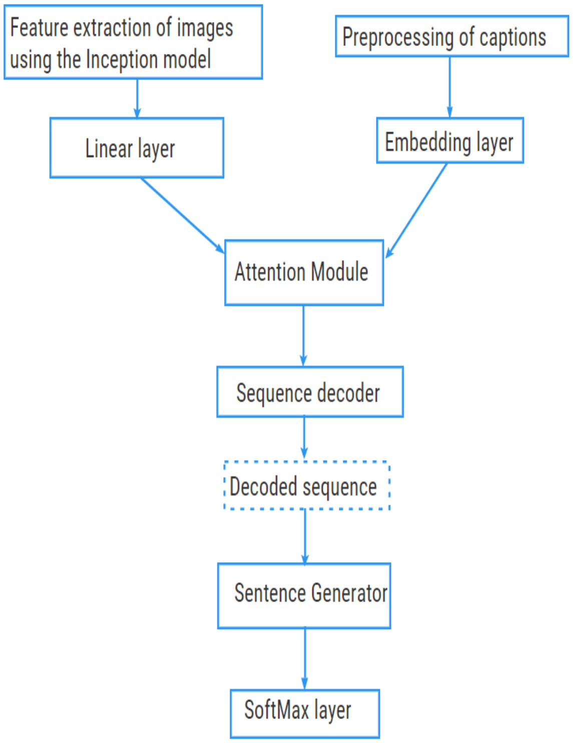


Fig. 2. Proposed Architecture for the Image captioning using Attention Model.

Figure 2 depicts the proposed architecture of Merge model used for image captioning. Here, Inception is used as a feature extractor and Encoder-Decoder architecture with attention module is used for training and generation of the captions.

Each sentence is extended to same length by padding tokens since the model takes input of sentences in fixed length.

3) Defining attention model: The attention model is split into 4 components

- Encoder - In this part of the model, the extracted encoded feature map from Inception model is passed into decoder. It also encodes the image feature vectors.
- Sequence Decoder - This part deals with handling text input. Sequence decoder has a recurrent network built and the captions of the images are taken as inputs after going through an embedding layer. Sequence decoder also initializes hidden state of the image.
- Attention - This part takes the encoded image feature vector from encoder and hidden state from the sequence decoder and calculates the weighted attention score of the features. This attention weighted score along with input sequence is passed through an embedding layer and gives combined input sequence as output. The combined input sequence is sent to Sequence decoder to give a new hidden state and an output sequence.
- Sentence generator - The output from the sequence decoder which is a output sequence is processed by the sentence generator and generates its predicted word probabilities.

IV. Discussion and Conclusions

A. *Dataset description*

Flickr8K dataset is the dataset that we used for image captioning. This dataset has 8000 images and these images are split into training, test and validation sets with 6000, 1000, 1000 images respectively. Each image has 5 different captions. This dataset is used for our project because it is relatively small compared to MS COCO and can be used to easily build mage captioning models.

In this paper, we deployed two most popular algorithms for image captioning i.e., merge and attention model. The models have been trained using Flickr8k dataset using progressive uploading. Both the models use Encoder-Decoder architecture. In both merge and attention model, both the encoded image and text description generated so far is sent to the decoder every time however in attention model weighted attention scores are assigned to encoded images.

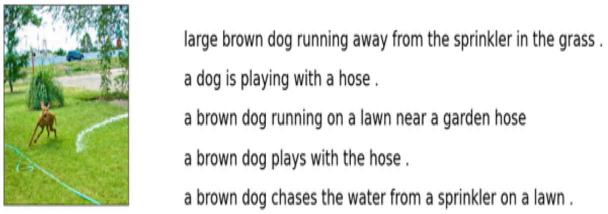

large brown dog running away from the sprinkler in the grass .
a dog is playing with a hose .
a brown dog running on a lawn near a garden hose
a brown dog plays with the hose .
a brown dog chases the water from a sprinkler on a lawn .

Fig. 3. Sample Image and its 5 captions.

Figure 3 depicts a data sample of the dataset Flickr8k i.e., each image of the dataset has 5 captions.

B. *Experimental Procedures*

- Collection of dataset (captions and photos), with regards to this paper it's Flickr8k dataset.
- Preparation of images and captions as expected by respective models.
- Build-out of the deep learning models.
- Training the models using Flickr8k dataset with 6,000 images and 1000 images as validation set.
- Evaluating the model using 1000 images as testing set from Flick8k dataset.
- Generating captions for the queried image.

C. *Experiment Environment*

TABLE II
Required Entities

Information Table
Python 3 Python SciPy Keras + TensorFlow Numpy and NLTK GPU runtime

Table II informs us the required entities that must be present in the system to run these particular models.

D. Graphs required for training and testing loss.

In this paper, Crossentropy loss function and adam optimizer is used. Formula for calculating crossentropy loss function is

$$L_{CE} = -\sum_{i=1}^{n} t_i log(p_i)$$

, where n stands for number of classes, t_i is the actual label and p_i is the confidence value of the predicted label.

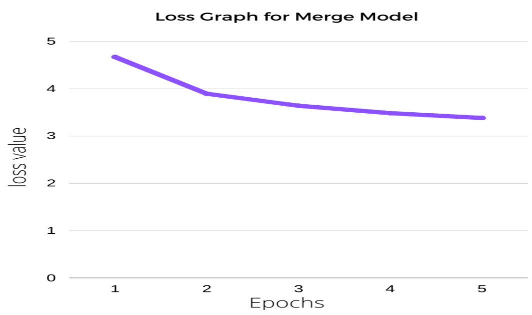


Fig. 4. Loss graph for Merge model used for image captioning.

Fig 4 shows the loss graph of merge model. 5 epochs are used to train the model and their respective loss values are plotted.

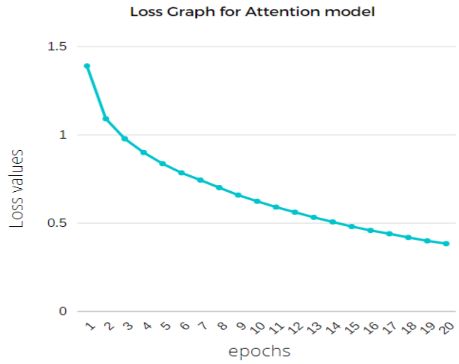


Fig. 5. Loss graph for Attention model used for image captioning.

Fig 5 shows the loss graph of attention model. 20 epochs are used to train the model and their respective loss values are plotted.

E. Metric of performance

In this paper, BLEU(Bilingual Evaluation Understudy) score is used to evaluate the models. BLEU score is a metric used to calculate how similar is the candidate text with the reference text. It's value varies from 0 to 1, where perfect match gets the score of 1 and perfect mismatch score of 0.

Table III contains the BLEU score values of Merge model for image captioning when used with Flickr8k dataset. BLEU-1 with maximum score of 0.528440.

Table IV has the BLEU scores of Attention model for image captioning when used with Flickr8k dataset. BLEU-4 shows highest score of 0.610474 compared to other BLEU scores.

TABLE III
BLUE SCORE TABLE FOR MERGE MODEL FOR FLICKR8K DATASET

BLUE scores	Value
BLEU-1	0.528440
BLEU-2	0.277373
BLEU-3	0.183888
BLEU-4	0.080878

TABLE IV
BLUE SCORE TABLE FOR ATTENTION MODEL FOR FLICKR8K DATASET

BLUE scores	Value
BLEU-1	0.555556
BLEU-2	0.372678
BLEU-3	0.553096
BLEU-4	0.610474

F. Information table for several Epochs

TABLE V
EPOCHS TABLE FOR MERGE MODEL

Epoch	Loss value	Time taken (in sec)
1	4.677	1648
2	3.892	1601
3	3.630	1597
4	3.481	1606
5	3.375	1612

Table V depicts Epochs results for Merge Model with average loss value 3.8117 and 1613 sec time taken.

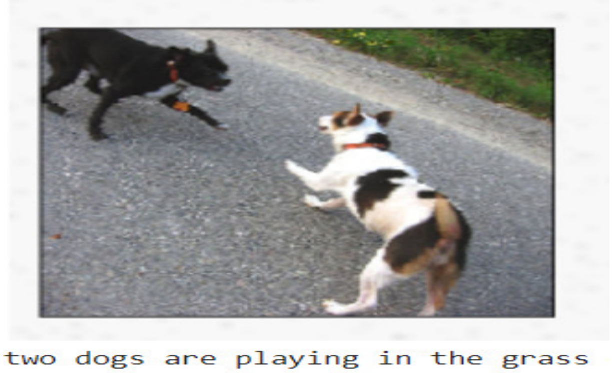
TABLE VI
EPOCHS TABLE FOR ATTENTION MODEL

Epoch	Loss value	Time taken (in sec)
1	1.390	257.80
2	1.091	165.28
3	0.977	165.51
4	0.900	165.16
5	0.837	165.39
6	0.785	165.21
7	0.744	165.28
8	0.701	165.04
9	0.659	165.04
10	0.624	165.36
11	0.591	165.41
12	0.562	165.12
13	0.533	165.41
14	0.506	165.19
15	0.481	164.73
16	0.459	165.06
17	0.440	165.98
18	0.419	164.88
19	0.400	164.78
20	0.384	164.54

Table VI shows Epochs details of Attention model for image captioning. Average loss value observed in this model is 0.6741 and time taken in seconds is 169.8 .

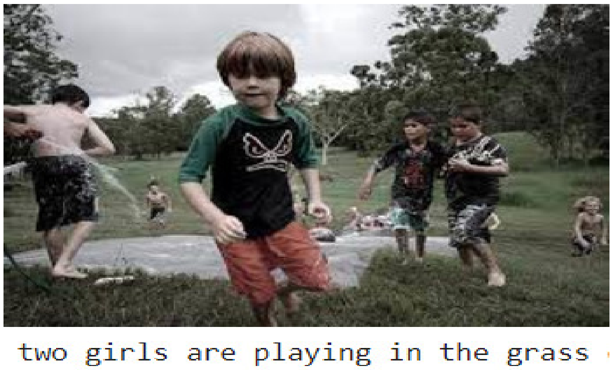
G. Sample inputs and outputs

Here are the sample outputs of image captioning :



two dogs are playing in the grass

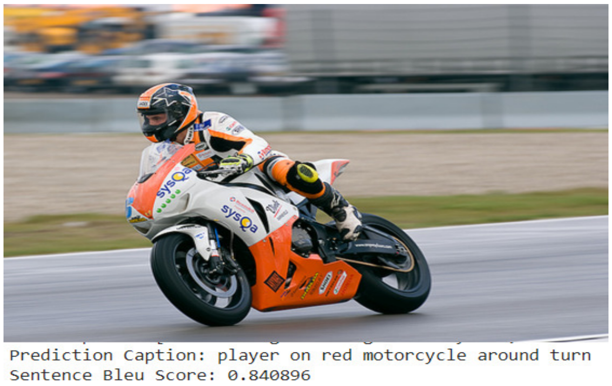
Fig. 6. Caption generated using merge model



two girls are playing in the grass

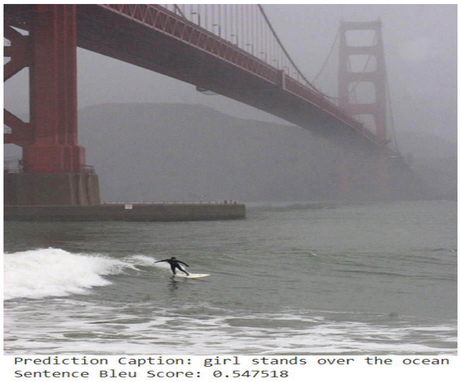
Fig. 7. Caption generated using merge model

Figures 6 and 7 show the captions generated for the respective pictures using Merge model.



```
Prediction Caption: player on red motorcycle around turn
Sentence Bleu Score: 0.840896
```

Fig. 8. Caption generated using attention model



```
Prediction Caption: girl stands over the ocean
Sentence Bleu Score: 0.547518
```

Fig. 9. Caption generated using attention model

Figures 8 and 9 depict the caption generated for the image using Attention model.

V. Comparison with previous works

In paper [21], they have used VGG19 for feature extraction of the images of the Flickr8k dataset. CNN Encoder- RNN decoder architecture with attention mechanism using 20 epochs has got the following results and also tried fine-tuning the encoder network.

TABLE VII
Comparsion of the results with previous works

Models	BLEU-4 score
VGG19 + Soft Attention Model	0.196
ResNet101	0.205
DenseNet101	0.195
Merge Model	0.080
VGG16 + Attention Model	0.610

Table VII tells us that VGG16 + attention model which has encoder and decoder mechanism in it gives the optimal results.

VI. Limitations

The major limitation of the paper is that it can be further improved by training merge and attention models with large datasets like Flickr30k and MSCOCO. Improvement can also be done by increasing number of epochs to get better BLUE scores and accuracy for merge model. Luong attention module can be used for better results since in this module both local attention and global attention are included.

VII. Conclusions

According to the BLEU scores calculated above, it can be concluded that attention model works well for image captioning. This is because in attention model priority is given to most relevant part of the image where as in merge model equal priority is to all parts of the image while generating next word of the caption. In the future, we would like to try getting more natural captions and also try for Flickr 30k dataset and MSCOCO dataset.

The problem with merge model approach is that, tends to focus on all parts of the image and generate words which might not be relevant to the context of the image. Simple CNN models also cannot decode the location and position of an object ,which leads to data loss and more training data is needed. Attention models are better in this context as they decode image in a way human natural tend to do ,by focusing on certain points of interest.

References

[1] Huang, Lun, et al. ”Attention on attention for image captioning.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[2] He, Sen, et al. ”Human attention in image captioning: Dataset and analysis.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[3] Zhou, Luowei, et al. ”Watch what you just said: Image captioning with text-conditional attention.” Proceedings of the on Thematic Workshops of ACM Multimedia 2017. 2017.

[4] Chen, Shi, and Qi Zhao. ”Boosted attention: Leveraging human attention for image captioning.” Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[5] Gao, Lianli, et al. ”Deliberate attention networks for image captioning.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

[6] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, “Image captioning with semantic attention,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[7] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In Computer Vision–ECCV 2010, pages 15–29. Springer, 2010.

[8] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In Proceedings of the 24th CVPR. Citeseer, 2011.

[9] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale ngrams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 220–228. Association for Computational Linguistics, 2011.

[10] D. Elliott and F. Keller. Image description using visual dependency representations. In EMNLP, pages 1292–1302, 2013.

[11] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pages 359–368. Association for Computational Linguistics, 2012.

[12] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, et al. From ´ captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1473–1482, 2015

[13] R. Lebret, P. O. Pinheiro, and R. Collobert. Simple image description generator via a linear phrase-based approach. Proceedings of the International Conference on Learning Representations (ICLR), 2015

[14] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112, 2014

[15] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[16] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, ¨ F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 2014.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[18] Ansar Hani; Najiba Tagougui; Monji Kherallah. Image Caption Generation Using A Deep Architecture. Publisher : IEEE

[19] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 1171–1179. Curran Associates, Inc., 2015.

[20] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, “Spice: Semantic propositional image caption evaluation,” in Computer Vision – ECCV 2016, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds., Cham, 2016, pp. 382–398, Springer International Publishing.

[21] Aditya Bhattacharya, Eshwar Shamanna Girishekar, Padmakar Anil Deshpande : ”Empirical Analysis of Image Caption Generation using Deep Learning”.