Image Captioning

Bhavya Narang
2019462
bhavya19462@iiitd.ac.in

Yash Aggarwal
2019480
yash19480@iiitd.ac.in

1 Introduction

Image captioning is the task of generating descriptions for an image automatically and reduce human efforts in the task of captioning. It plays an important role in multiple tasks such as generating multiple captions for unlabelled textbook images, providing descriptive summary to blind people, etc.

Not only the summary should tell about the objects present in the image and the relation between them but also the description should be fluent and coherent in a human understandable context and that is what makes the task of image captioning difficult.

This problem lies at the intersection of both Computer Vision and Natural language Processing. The advances in deep learning in both the fields of CV and NLP have led to the problem being more widely pursued. On the CV side of the solution we want to extract the objects and predicates in the most significant manner so that we can get the output of this system into the NLP pipeline along with the required caption. Now on the NLP side of the problem given the object and predicates we want to interpret the relations between them and finally pass them onto models which output a natural language form of the statement.

2 Problem Statement

We will try to analyze the problem of image captioning from the vision and natural language aspect. For the vision part of the problem, we will focus on majorly two problems: feature extraction and object recognition. This will form the basis of the our analysis further. We will try to use different architectures for this problem. Some of them include CNN+RNN, CNN+Transformers and other techniques. This forms the basis of the natural language part of the problem. Using the results generated, we will perform caption generation task using various methods such as LSTM, encoder-decoders and other techniques.

3 Dataset

We will be using the Flickr8k dataset, which comprises of 2 parts that is, images and corresponding captions. The dataset comprises of 2 parts, images and captions. There are a total of 8092 images with each image having a unique id. Corresponding to each image, we have 5 captions, with a total of 40460 captions.

The average length of the captions comes out to be 55.13, with the max length being 199 and the minimum length being 1. The median length of the captions comes out to be 53.

To obtain a better representation of the textual data, I created a word cloud of the dataset shown in the figure below:

4 Literature Survey

Vinyals et al. (2015) is motivated by the progress and recent development (for 2015) of the state of the art machine translation systems which use recurrent neural networks. These RNNs take a sentence as input and try to produce another sentence in some other translation. The paper takes an RNN based encoder-decoder and in the generation task replaces the encoder RNN with a deep convolutional neural network architecture (such as image net) by pre-training the network for image classification initially and then removing the last layer and uses the embedding thus obtained to feed into the decoder RNN for caption generation. The paper uses LSTM architecture for RNNs to tackle the vanishing and exploding gradient problems thus faced. Finally, the paper uses BLUE score for testing and analysis and claims to have

achieved a state of the art results at that time.

Xu et al. (2015) is motivated by the paper "Show and Tell" as the name suggests itself. Instead of only using the encoder-decoder architecture as suggested by "Show and Tell" which takes a compressed static representation of the image, this work uses attention to compute which objects are related to the context of the current word which is going to be predicted. Also, instead of taking just the second last layer which contains high-level features, the model tries to advocate the use of lower-level features which are obtained by previous layers which help in preserving more information. The paper uses both hard and soft attention mechanisms and also provides a visualization mechanism that brightens the part of the image which has a higher probability score in the case of soft attention mechanism and the considered object in hard attention respectively. Finally, the paper uses BLUE score for testing and analysis which is done on Flickr8k, Flickr30k and COCO datasets.

Moving away from the traditional convolutional approach, Liu et al. (2021) use sequential patches for generating embeddings, which are later used in self-attention mechanisms. The CPTR model suggested in this paper uses a cross attention layer in the decoder for word-to–patches attention. For the encoder, they use identical N layers of multi-head self attention sublayer which allows the model to encode multiple subspaces. In a similar fashion, the decoder layer consists of stacked layers containing multi-head self attention sub layer. For evaluation of the said model, they have used BLEU-1,2,3,4, METEOR, ROUGE and CIDEr scores. These metrics calculate the fluency score for the output generated. Higher the value, more fluent the translation. This model performs better than convolutional models, such as CNN + RNN and CNN + Transformer. This paper clearly demonstrates that rather than depending on a CNN based approach, using a sequential patch encoder utilizes longer dependencies and the decoder is also much more precise in assigning "words" to the "patches".

Contrary to the traditional encoder-decoder structures which use image features, Herdade et al. (2019) uses object relational transformers, which use the information about the spatial relationships between input objects using geometric attention. This is an important feature to encode as it allows to capture reasoning in the physical world. The difference between "a man standing beside the chair" and "a man sitting on the chair" can be captured usingh the spatial information about the man and the chair itself. This paper uses object relational transformers to show the usefulness of geometric attention and improved captions. At each of the encoders, they use a multi head self attention layer where geometric features between bounding boxes for different objects are concatenated together. This approach is a possible replacement for the original positional encoding. A possible improvement for this model would be to include geometric attention in decoder cross attention layers.

Proposing a variation of the usual problem statement, Anderson et al. (2018) suggests a semi supervised task for image captioning where they propose learning from a partially specified sequence data. This approach lifts the restriction of using image-caption corpora, and provides novel state of the art results on existing neural captioning models. A partial sequence data only provides parts of text data, while the information is hidden. They use a finite state automata in these cases and using beam search find appropriate results for undefined states. They propose a novel algorithm, PS3(Partially-specified sequence supervision) for training neural networks on partially sequenced data using FSA. Their proposed model achieves state of the art results on COCO dataset and is able to describe new visual concepts on the existing Open Images dataset.

Shi et al. (2020) lays emphasis on high-level semantic contexts present in the image which have not been much used till now to perform image captioning. The use of semantics is made by preparing a visual relationship detection mechanism in a weakly supervised setting from both the image and textual data by extracting the objects and the predicates from them respectively. Also, the work considers the multi-task learning-based approach to extract the objects in the image and then a visual relationship graph is made by passing the objects and predicates thus obtained into a graph convolutional network-based model. The work uses an implementation of faster R-CNN and other feature extraction models such as Resnet-101 for multi

task object detection.This Finally, this model also uses an attention-based mechanism to output the final obtained caption and claims this to be the state of the art technique (2021). The work is done on MSCOCO dataset and uses metrics such as BLEU-1, BLEU-4, SPICE, etc.

5 Methodology

We have followed the approach of an encoder decoder mechanism for our task. The encoder consists of pre trained resnet on Image Net dataset and we have removed the last 2 layers that are linear and softmax to extract the features.

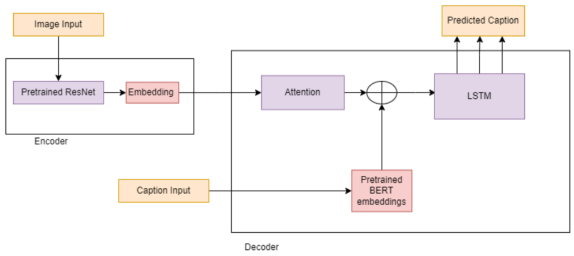

Figure 1: overall encoder-decoder structure of our model

These image embeddings are then fed into the decoder along with the caption of the image and use of attention is also made for this. Caption text is passed into BERT model and the text encoding is passed and then concatenated with the embeddings obtained after attention with the image.

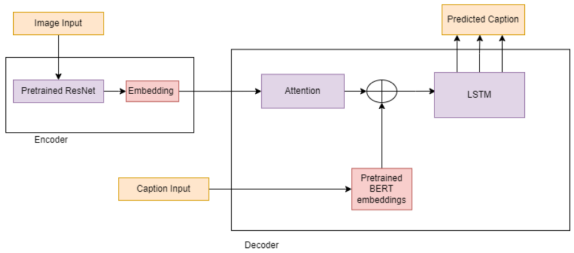

Figure 2: model inference structure using previous word as inference to next.

Finally the concatenated embeddings have been passed through RNN that is LSTM to obtain the word predictions one by one. We have modelled the prediction of the words as a multi class classification problem and hence we have used cross entropy loss for the same.

6 Experiments

For experiments we have used unfreezing of layers of both the encoder and the decoder and have observed the impact of this after training the models on the loss function.

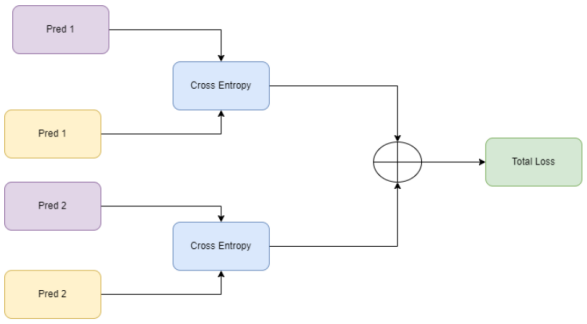

Figure 3: our loss function for the image captioning model.

7 Results and Analysis

Using the model defined above, we generated the captions for the test set of the Flickr Dataset. Some of the results are shown in the figures 1, 2, and 3.

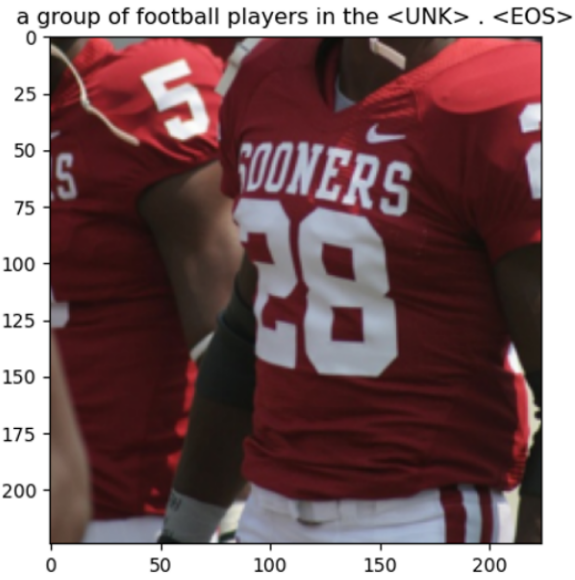

Figure 4: Caption: A group of football players

These were some of the results that we found to be much more coherent and comprehensible as compared to the other results. Some of the results which were not as coherent as the others are shown in Figures 4, 5 and 6.

7.1 Bias in Results

We also see that the model didn't learn properly on small no. of epochs, and found that certain biases got introduced in it. For instance, with the image of a man or a boy wearing a shirt, we saw that the model would associate 'red' color with the shirt.

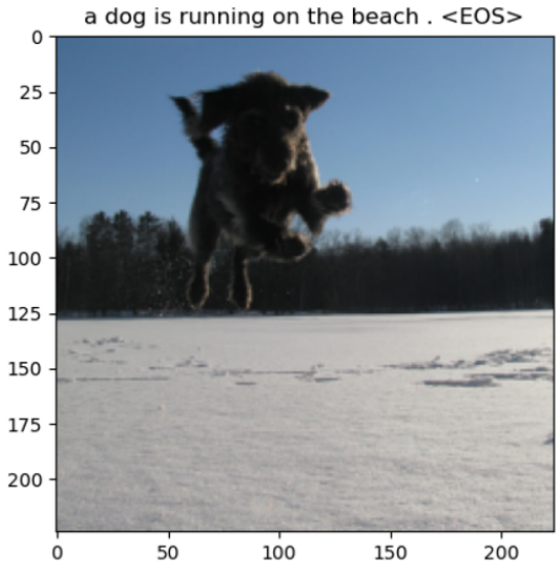

Figure 5: Caption: A dog is running on the beach

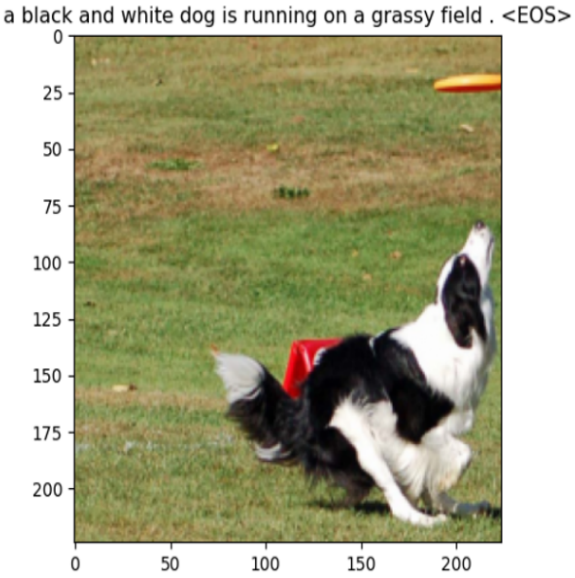

Figure 6: Caption: A black and white dog is playing hoop

One of the instances is given in Figure 7.

For the experiments defined in the experiments section, we found the following results for different layers:

Loss(100 batches)	ResNet	DistilBert
224.94	0	0
218.67	0	10
219.77	10	0
215.86	10	10

Table 1: Loss with different no. of layers freezed in ResNet and DistilBert.

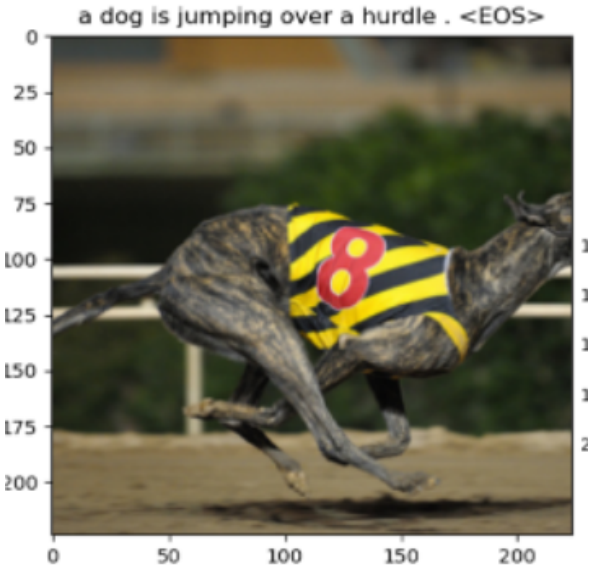

Figure 7: Caption: A dog is running in a race. We see that the image is not able to distinguish between the fences in a background and a hurdle.

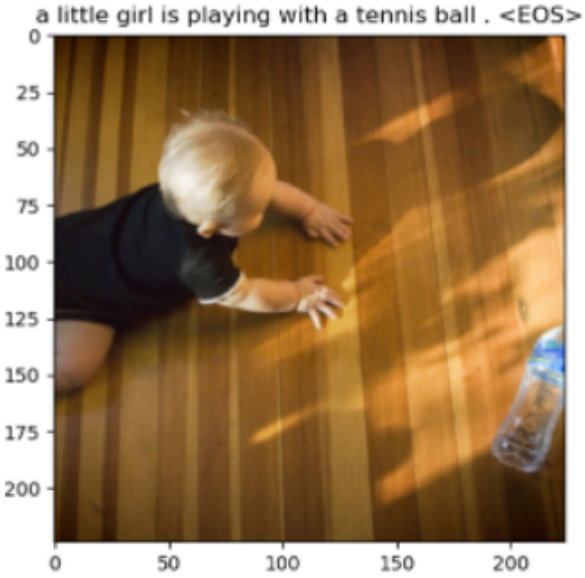

Figure 8: Caption: A little girl is crawling on the floor. We see that the shadow of girl's head has been labeled as a ball in the image.

8 Individual Contributions

The individual contributions were as follows: The initial dataset collection, cleaning and EDA was done by Yash. The models were created by both Bhavya and Yash. The experiments with the model were planned and performed by Bhavya.

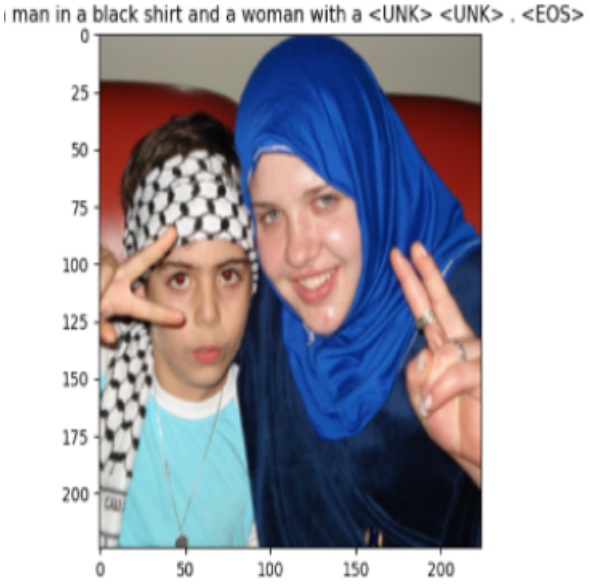

Figure 9: Caption: A boy and a smiling woman. We see that because of the black and white head band, the color of the boy's shirt has been marked as black. Also, woman's photo is not defined clearly as well.

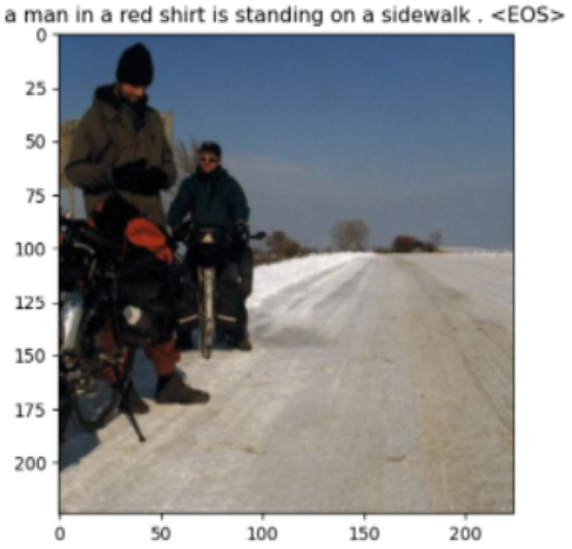

Figure 10: Biased Output: We can clearly see that the man is not wearing a red shirt. However because of a large no. of images with a man wearing a red shirt, we see that the model learns this bias.

References

Peter Anderson, Stephen Gould, and Mark Johnson. 2018. Partially-supervised image captioning. *Advances in Neural Information Processing Systems*, 31.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, 32.

Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. 2021. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. 2020. Improving image captioning with better use of caption. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7454–7464, Online. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

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