Image To Text Matching Captioning For News Images

Report submitted in fulfillment of the requirements for the B.Tech. Project of

Third Year B.Tech.

by

Deepesh Tank (18075017)
and
Dishant Chourasia (18075018)

Under the guidance of

Dr. Tanima Dutta

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY (BHU) VARANASI
Varanasi 221005, India
May 2021

Declaration

We certify that

1. The work contained in this report is original and has been done by ourselves and the general supervision of our supervisor.
2. The work has not been submitted for any project.
3. Whenever we have used materials (data, theoretical analysis, results) from other sources, we have given due credit to them by giving their details in the references.

Place: IIT (BHU) Varanasi
Date: May 2021

Deepesh Tank (18075017)
Dishant Chourasia (18075018)
B.Tech. Students
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Certificate

This is to certify that the work contained in this report entitled "Image to text matching captioning for news images" being submitted by Deepesh Tank (Roll No. 18075017) and Dishant Chourasia (Roll No. 18075018), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bona fide work of our supervision.

Dr. Tanima Dutta
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Place: IIT (BHU) Varanasi
Date: May 2021

Acknowledgments

We want to express our sincere gratitude to the people who have helped us the most throughout our project. We are grateful to our project supervisor (Dr. Tanima Dutta) for providing us an opportunity to implement the paper entitled: “Image To Text Matching Captioning For News Image” and her constant support for the project. At last, but not the least we would like to thank our tutors who encouraged us and helped us out in finalizing the project.

Place: IIT (BHU) Varanasi

Date: May 2021

Deepesh Tank (18075017)
Dishant Chourasia (18075018)

Abstract

News image consists of an image, an article and a caption.News Image Captioning is the task of automatically generating a caption for a news image.The motive is to automatically generate captions for news images which if needed can then be used as reference captions for manually creating news image captions.

We have proposed a deep learning model for news-image captioning. We are using ROI pooling to extract image features and Order embedding to embed the article and caption. Then, we are using LSTM and the Transformer network for developing the model architecture.

Content

1. Introduction
2. Related Work
3. Proposed Methodology
4. Experimental Details

    a. Dataset

    b. Training Details
5. Qualitative Results
6. Ablation Study

    a. Transformer + VGG-19

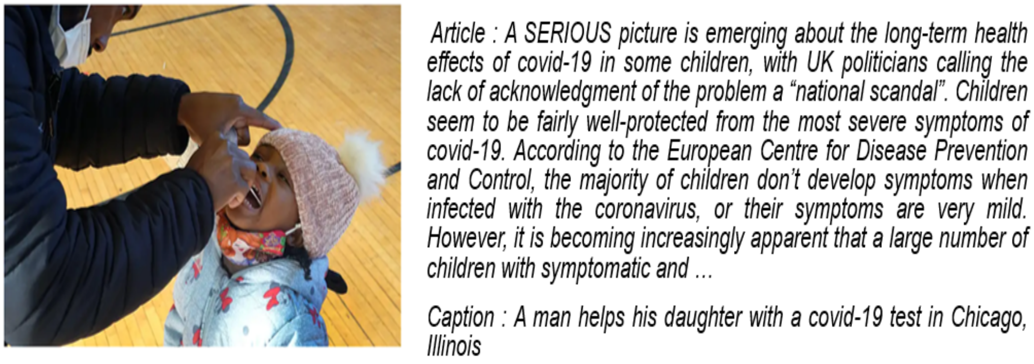
    b. Transformer + Article Only
7. Conclusion
8. References

1. Introduction

Image Captioning is the task of generating a description for an image. It is a significant research field in computer vision. Image captioning has several applications. It is important in image understanding like recognition of objects, scenes and object to object relationships.

News Image Captioning is an advanced level form of Image captioning. The motive is to automatically generate captions for news images which can then be used as reference captions for manually creating news image captions. Two problems faced by Traditional Image Captioning are : i) unawareness of names, places, and named entities, ii) linguistic expressiveness i.e. generated captions are usually shorter than human-generated captions. News image captioning tackles both of these two problems. News image captions include specific person's name, organization names, places and also provide good contextual information.

A news image is shown in the figure below. It might be hard to detect the main object in the image. For example: person, cap, mask, basketball court. The caption holds much information (e.g. covid-19 test, Chicago, Illinois) that would be quite hard to detect if we only use the image. In news, the article is the major medium of information, the caption and image provide extra explanations to keep up the article.



Article : A SERIOUS picture is emerging about the long-term health effects of covid-19 in some children, with UK politicians calling the lack of acknowledgment of the problem a "national scandal". Children seem to be fairly well-protected from the most severe symptoms of covid-19. According to the European Centre for Disease Prevention and Control, the majority of children don't develop symptoms when infected with the coronavirus, or their symptoms are very mild. However, it is becoming increasingly apparent that a large number of children with symptomatic and ...

Caption : A man helps his daughter with a covid-19 test in Chicago, Illinois

Figure 1: An example of a news image along with its article and caption.

In this project, we present a model for news-image captioning which is based on ROI pooling and the Transformer network (a successful architecture for different NLP tasks). It takes image and text features into consideration to generate a caption that contains both text and image features of the image-article pairs.

The results of the experiment show that the predicted captions can give relevant information related to the news article which can be used to produce captions that are written by a person. They can also help in finding images that appropriately support the article/text.

2. Related Work

Our work is related to image captioning, object detection and encoder-decoder architecture.

Object detection: Modern machines face a challenge in recognizing, classifying and detecting numerous objects present in the images. However, in recent years, a tremendous amount of effort has been carried out for detecting and identifying objects using convolutional deep learning neural networks. With the rise in the improvement of advanced hardware devices like TPU, Super Computer etc. and new training techniques for larger dataset neural networks have become more prominent in the object detection and recognition fields. In recent studies, it is found that using deep learning, features that are extracted directly from image pixels are giving more promising results than manually selected features in the object detection field. Newly developed, deep learning-based algorithms replaced the traditional manual feature extraction methods by methods that directly extract features from the original images.

Image captioning: Automatic caption generation of an image is the most elemental problem that links Computer Vision and Natural Language Processing. There has been a lot of work in object detection, image annotation and image classification, but less attention has been paid in generating sentence descriptions.

A solution is to use the results of these methods where an image is annotated with a set of keywords and then these keywords are put through some other stage that puts these keywords in a sentence.

Encode-Decoder Architecture: In deep learning, an encoder takes a sentence as input and then transforms it into an embedded vector of fixed length. This embedded vector is fed into a decoder which generates sentences in a particular target language. This architecture based on encoder and decoder has been used in captioning images. Here the captioning problem is viewed as a translation problem where the image is of the source language, the caption is of the target. For encoding purpose, CNN is taken and for decoding purpose, RNN is taken.

3. Proposed Methodology

In this part, we propose a Deep Neural Network to do this task.

The problem statement is devised as : Given a pair of news image “I” and its related article “A” , select a sentence “B” from the sentences present in the article “A”, such that the selected sentence “B” is best for expressing the given image “I” along with its article “A”. The training data comprises tuples of article-caption-image. During testing, we are given a news image and an article as input and we have to output a caption.

For the image, we first feed it into the VGG-19 model (pre-trained) to extract the feature map for the image. This feature map of dimension (1,28,28,512) is received from the “block4_conv4” layer of the VGG-19 model. Then, we feed the images into Faster-RCNN to get bounding boxes and scores for the image. In descending order of scores, we take the first 3 corresponding bounding boxes (ROI) for the image. The ROIs give us the features of the specific objects present in the image and thus the caption generation will be highly impacted by those objects. Then, using the feature map of the images and ROIs of the images we perform ROI pooling with pooled_height and pooled_width set to 7, and finally, reshape it. After this process, we get the Image processed vector of dimension (1,75264) for an image.

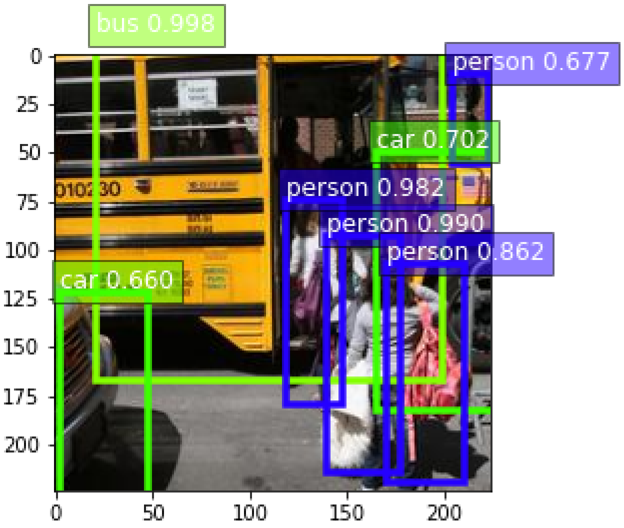


Figure 2: Bounding Boxes for a test image.

For the article, we have capped the maximum number of sentences to 40. Then we make a list of sentences and then tokenize each sentence. Tokenization lowercases the sentences, removes useless symbols, and also removes stop words. Stop words are words that provide no meaning to the sentence. Removing stop words will not affect the processing of text. The aim is to reduce the noise and dimension of the feature set.

After this, using a pre-trained Order-Embedding model we convert these sentences into a sequence of vectors which embeds each sentence into a 1024 length vector. Therefore, for every article, we have a sequence of 40 vectors each having a length of 1024. This step is needed as the neural network has significance for numbers only, not words.

For the caption, we first tokenize it and then convert it into a Caption vector using an Order-Embedding model(pre-trained) to get a 1024 length Caption vector. Now, the model takes one input as the embedded sentence vector sequence and feeds it to a LSTM network. It has been used to learn the sequence of sentence vectors of the article. It outputs a 1024 length Intermediate output vector.

Also, the model takes the image feature vector as a second input and by virtue of a dense layer outputs a 1024 length vector. Then both of these 1024 length vectors received from the image and article are concatenated to get a Concatenated vector. Concatenation is done to merge both image and article features in a single vector. This Concatenated vector is embedded using Position-Embedding and then fed into the Transformer network. The vector obtained is then passed to the Global average pooling layer.

Global Average Pooling layer can be used instead of fully connected layers that are generally used in Convolutional Neural Networks. We simply take the mean of each feature map, and then the output vector is propagated further.

The advantage of using this layer is that there are zero parameters that need optimization and therefore in this layer, overfitting is avoided.

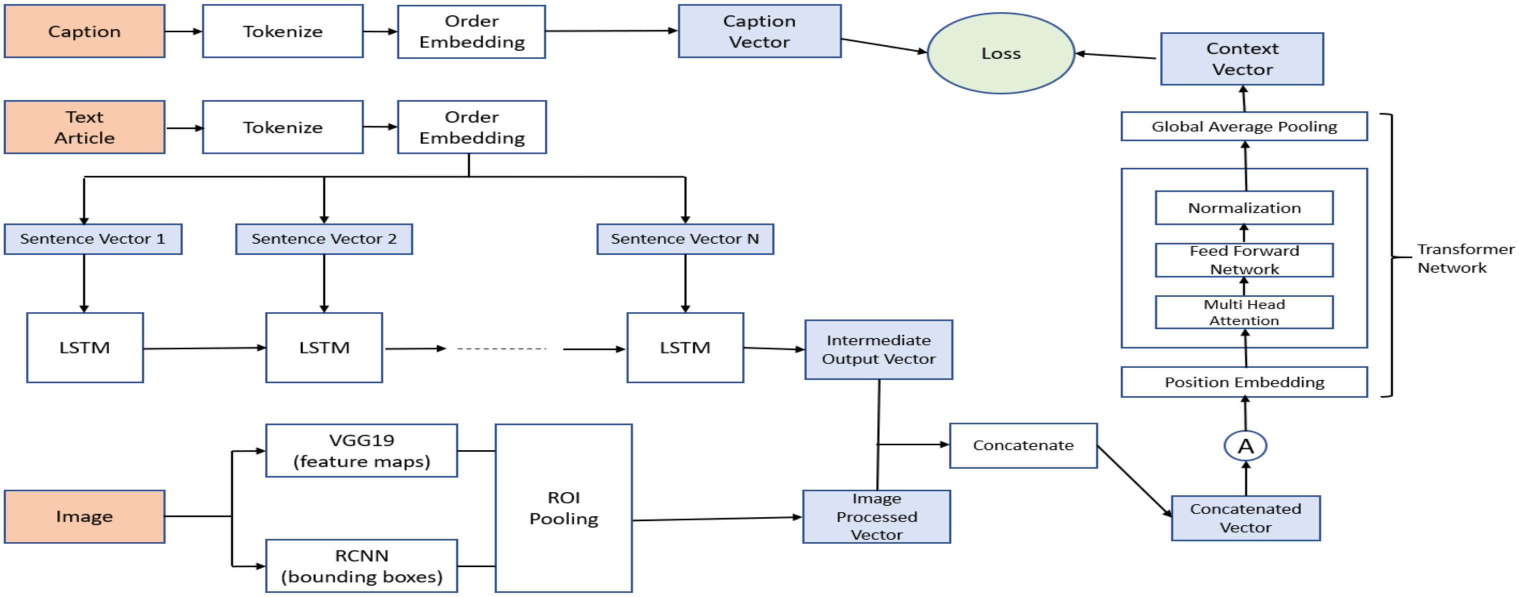


Figure 3: Block diagram of the model architecture(Transformer + RCNN).

This vector and the Concatenated vector are then passed into an Attention layer. Recurrent Neural Networks is unable to remember longer sequences due to the problem of vanishing gradient. So, the attention layer helps in memorizing long sequences. The attention mechanism helps the model to put more emphasis /attention on relevant parts of the input sequence
The output of the attention layer and the output of the transformer network (after global average pooling) are concatenated and squeezed to a 1024 length vector (same dimension as caption vector to calculate loss). Hence, we receive a context vector of length 1024. This context vector is used to calculate loss against the caption vector using categorical cross-entropy and then backpropagate this loss to train the model.

For testing, we put an image and its article as input and receive a context vector. Then we calculate cosine similarity between this context vector and each embedded sentence vector of the article. The sentence with the highest cosine similarity is chosen as the caption for the news image.

4. Experimental Details

4. a) Dataset

We used GoodNews Dataset, which is the largest dataset for news-image captions. We considered using this dataset but found some issues. The biggest issue in the GoodNews Dataset is that many instances contain incomplete text. Therefore, we try to solve this problem by taking only those articles which have both images and captions. We have only taken the article in which the number of sentences is less than or equal to 40. If the number of sentences is less than 40, we append empty sentences to make all articles have equal lengths of sentences. The images contained in the dataset are of random sizes, for the purpose of gaining uniformity we have reduced the size of all images to a fixed size of 224 x 224 pixels.

Another problem is that many instances of the dataset lack several leading sentences, which usually convey essential information on news articles. This problem can only be solved manually, which requires a lot of manual work therefore we have not taken this into consideration.

4. b) Training Details

After refining the dataset, the total number of news-image captions are 1609 . After the split, the training dataset contains 1500 and the testing dataset contains 109 news-image captions. Parameters and Hyperparameters for the models are:

1. Optimizer : SGD

2. Learning rate : 0.1

3. Momentum : 0.7

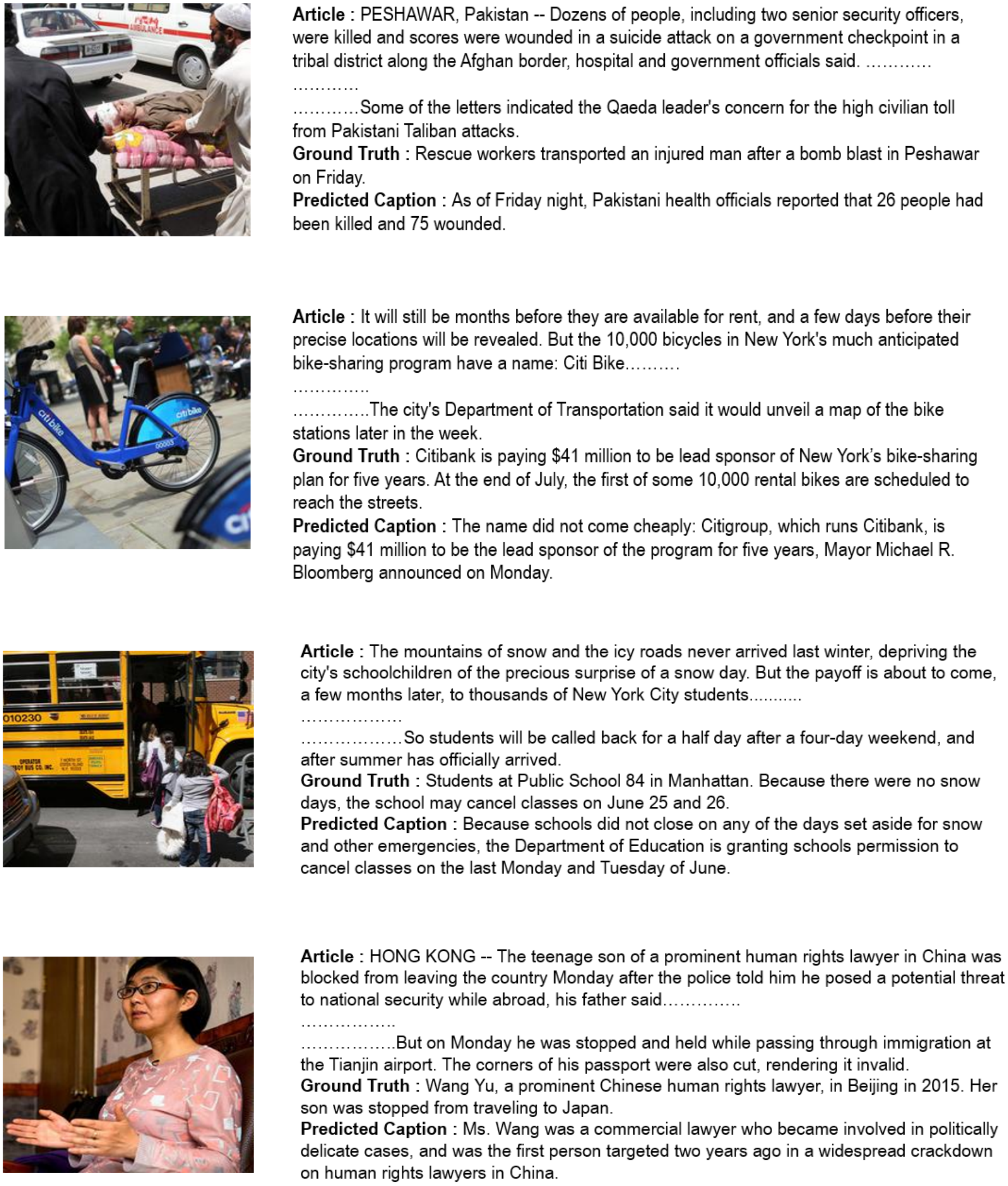
4. Loss function Used : categorical_crossentropy

5. Number of epochs : 20

6. Activation function: Relu

7. Number of hidden layers in feed-forward network inside transformer: 32

5. Qualitative Results



Article : PESHAWAR, Pakistan -- Dozens of people, including two senior security officers, were killed and scores were wounded in a suicide attack on a government checkpoint in a tribal district along the Afghan border, hospital and government officials said. …………

…………

…………Some of the letters indicated the Qaeda leader's concern for the high civilian toll from Pakistani Taliban attacks.

Ground Truth : Rescue workers transported an injured man after a bomb blast in Peshawar on Friday.

Predicted Caption : As of Friday night, Pakistani health officials reported that 26 people had been killed and 75 wounded.

Article : It will still be months before they are available for rent, and a few days before their precise locations will be revealed. But the 10,000 bicycles in New York's much anticipated bike-sharing program have a name: Citi Bike………

…………..

…………..The city's Department of Transportation said it would unveil a map of the bike stations later in the week.

Ground Truth : Citibank is paying $41 million to be lead sponsor of New York's bike-sharing plan for five years. At the end of July, the first of some 10,000 rental bikes are scheduled to reach the streets.

Predicted Caption : The name did not come cheaply: Citigroup, which runs Citibank, is paying $41 million to be the lead sponsor of the program for five years, Mayor Michael R. Bloomberg announced on Monday.

Article : The mountains of snow and the icy roads never arrived last winter, depriving the city's schoolchildren of the precious surprise of a snow day. But the payoff is about to come, a few months later, to thousands of New York City students..........

………………

………………So students will be called back for a half day after a four-day weekend, and after summer has officially arrived.

Ground Truth : Students at Public School 84 in Manhattan. Because there were no snow days, the school may cancel classes on June 25 and 26.

Predicted Caption : Because schools did not close on any of the days set aside for snow and other emergencies, the Department of Education is granting schools permission to cancel classes on the last Monday and Tuesday of June.

Article : HONG KONG -- The teenage son of a prominent human rights lawyer in China was blocked from leaving the country Monday after the police told him he posed a potential threat to national security while abroad, his father said…………..

……………..

……………..But on Monday he was stopped and held while passing through immigration at the Tianjin airport. The corners of his passport were also cut, rendering it invalid.

Ground Truth : Wang Yu, a prominent Chinese human rights lawyer, in Beijing in 2015. Her son was stopped from traveling to Japan.

Predicted Caption : Ms. Wang was a commercial lawyer who became involved in politically delicate cases, and was the first person targeted two years ago in a widespread crackdown on human rights lawyers in China.

Figure 4: Results obtained on Testing data

We have implemented the main model along with its two other variants for the comparison of results. The metric BLEU score, Meteor score, and Cider Scores are used as evaluation parameters.

Bleu score measures word-to-word similarity directly and also word clusters in two sentences are the same to what extent. It is also possible that good outputs that are using different words score badly just because they don't match the human-written sentence.

Both news articles and news-image captions have a certain number of named entities that carry relevant contextual information. Therefore, we are using CoverageNE to measure the coverage of named entities in predicted captions. The coverage for a pair of predicted and true caption is defined as recall of named entities in the caption:

$$\text{CoverageNE} = [\, E\ (\text{predicted}) \cap E\ (\text{gold})\,] \,/\, [\, E\ (\text{gold})\,]$$

Here, E (predicted) and E (gold) are sets of named entities in predicted and true captions respectively. We have used SpaCy (industrial-strength natural language processing tool) to identify named entities in the ground truth captions, and then find the number of exact matches in predicted captions.

Model	Trainable Parameters	BLEU	Meteor	Cider	CoverageNE
Transformer + RCNN (object based feature extraction)	123 M	**0.2532**	**0.054**	**0.101**	**0.2235**
Transformer + VGG-19 (full image-based feature extraction)	46 M	0.2170	0.044	0.062	0.1692
Transformer + Article Only (image features not considered)	19 M	0.2088	0.043	0.061	0.1679

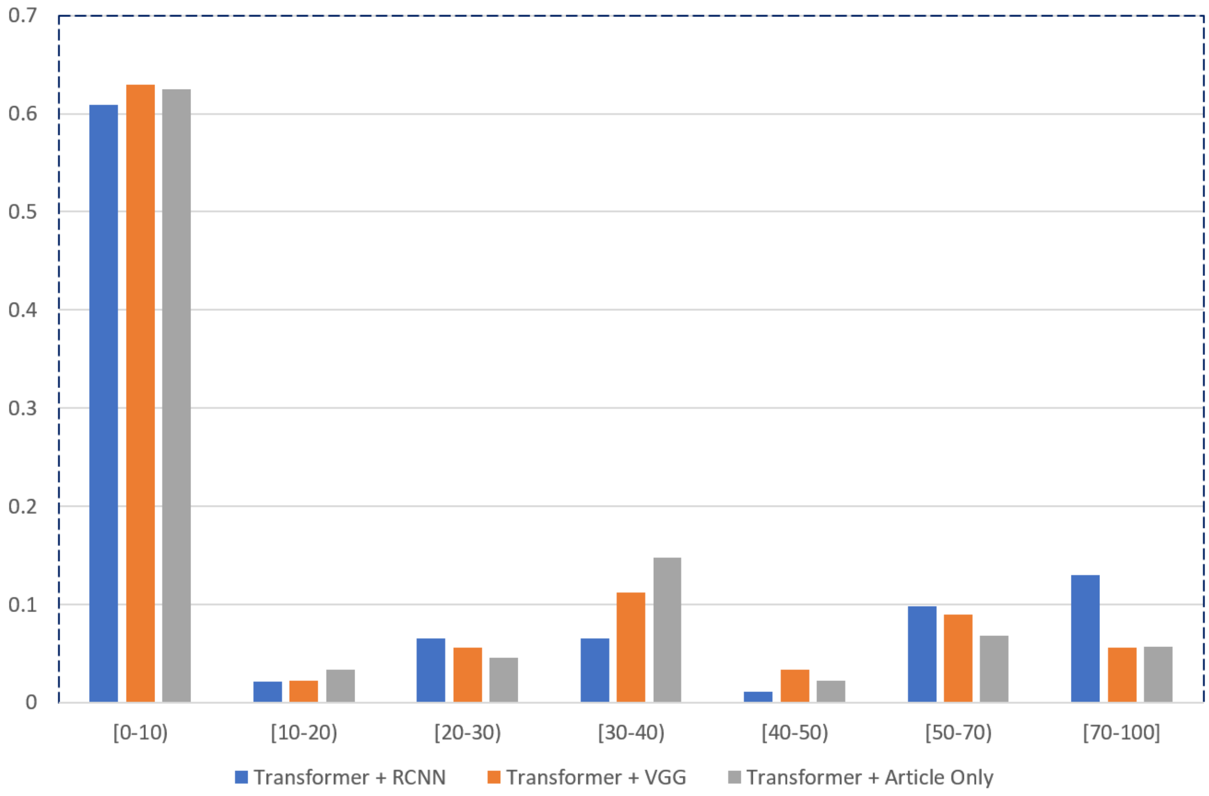


Figure 5: Distributions of CoverageNE scores for the main model and its variants.

We multiplied the received CoverageNE score by 100 and plotted the distribution of the score range to the fraction of testing data for each model. We can observe that for higher CoverageNE score range i.e, (50-70) and (70-100), the Transformer + RCNN model performs best as compared to other variations.

6. Ablation Study

6.a) Transformer + VGG-19: For this variation, we use VGG-19 instead of Faster-RCNN and ROI pooling to get full image features. We use the "fc2" layer having a vector of size 4096 to get an image feature. Both image features and text features are embedded by a pre-trained order embedding model such that both of them are represented in a 1,024-dimensional semantic space. The rest of the network remains the same. A few changes to the model were done to implement this variation and the model architecture is shown in the figure.

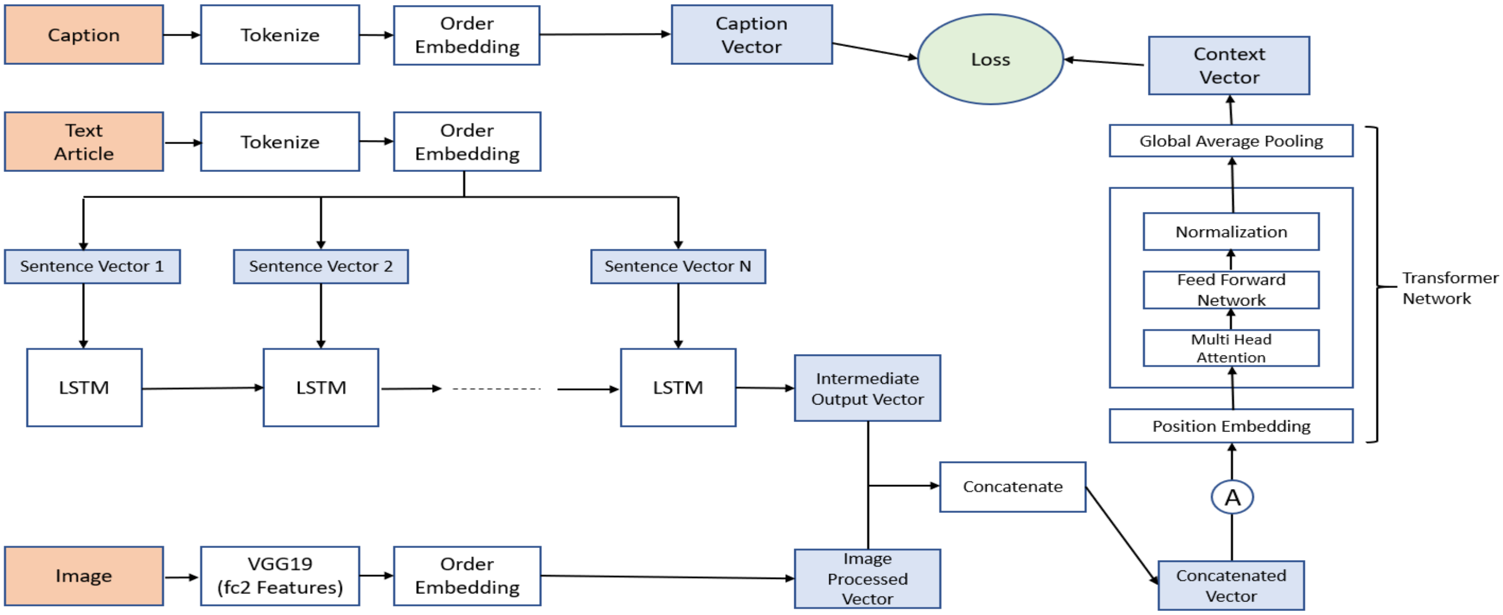


Figure 6: Block diagram of the model architecture(Transformer + VGG)

6.b) Transformer + Article Only: For this variation, we don't consider image features and train the network only on the basis of articles and captions. For testing, we input the article only to get the caption. Several changes were made in the model to implement this variation and the model architecture is shown in the figure.

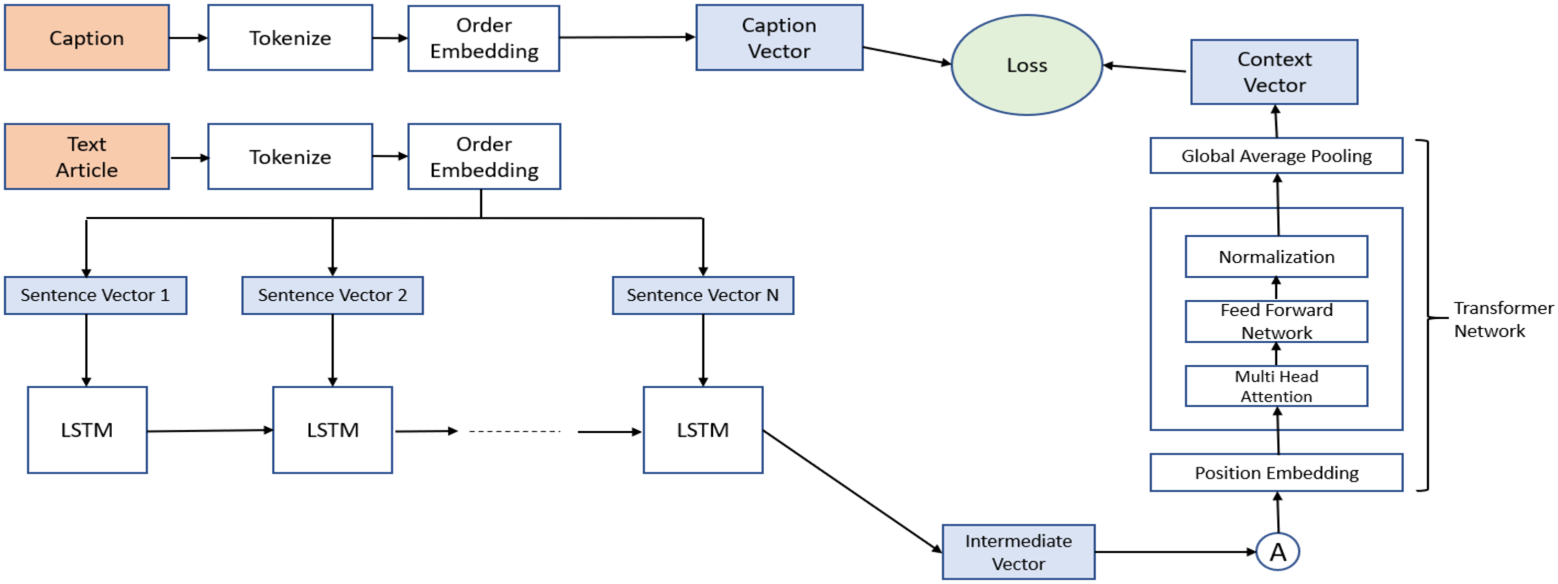


Figure 7: Block diagram of the model architecture(Transformer + Article only)

7. Conclusion

In this paper, we presented a method for news-image captioning based on the Transformer model using RCNN and ROI pooling for image feature extraction. This method can integrate image and text features in predicting a caption. The experimental results demonstrated our proposed model overperforms other variants by significant margins, therefore it can integrate visual and textual information in generating captions effectively. In the future, our model can improve its results on a more refined dataset specifically designed for new image captioning tasks. Our model is flexible with the choice of method for extraction of text and image features, thus can be used in comparing the performance of various feature extraction models.

8. References

1. https://www.newscientist.com/article/mg24933233-600-children-are-getting-long-covid-and-being-left-with-lasting-problems/
2. https://paperswithcode.com/method/global-average-pooling#:~:text=Global%20Average%20Pooling%20is%20a,in%20the%20last%20mlpconv%20layer.
3. https://www.analyticsvidhya.com/blog/2019/11/comprehensive-guide-attention-mechanism-deep-learning/
4. https://www.quora.com/What-is-the-benefit-of-using-attention-mechanism-in-OCR
5. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.
6. Ivan Vendrov, Ryan Kiros, Sanja Fidler, Raquel Urtasun. "Order-Embeddings of Images and Language." arXiv preprint arXiv:1511.06361 (2015).
7. Yang, Zhishen and Okazaki, Naoaki "Image Caption Generation for News Articles" Image Caption Generation for News Articles , Barcelona, Spain (Online) ,Dec 2020.
8. A. F. Biten, L. Gomez, M. Rusiñol and D. Karatzas, "Good News, Everyone! Context Driven Entity-Aware Captioning for News Images," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12458-12467, doi: 10.1109/CVPR.2019.01275.