IMAGE TO TEXT MATCHING CAPTIONING FOR NEWS IMAGES

BTech Project

Group Members:

Deepesh Tank (18075017)

Dishant Chourasia (18075018)

INTRODUCTION: PROBLEM STATEMENT

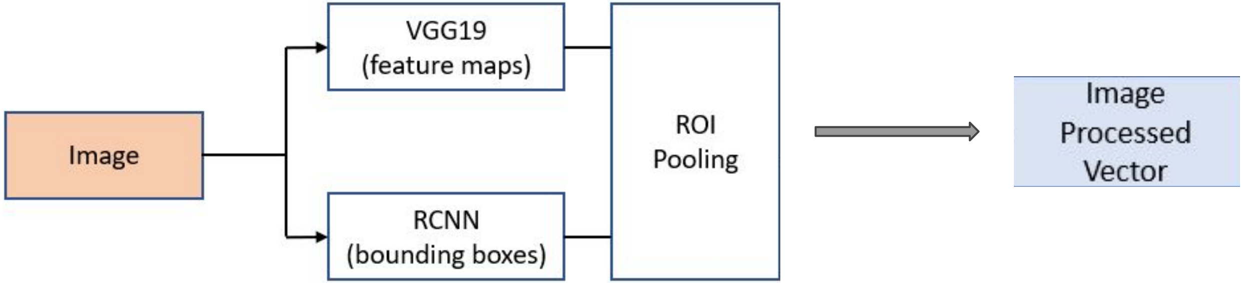
Our problem is formulated as follows:

- Given a news image I and its associated article D, choose a sentence S from the article that best describes the image given D. The training data thus consists of article-image-caption tuples.
- During testing, we are given a document and an associated image for which we need to output a caption.

Used Transformer network and ROI Pooling based Architecture to develop the model.

PROPOSED METHODOLOGY

FOR IMAGE



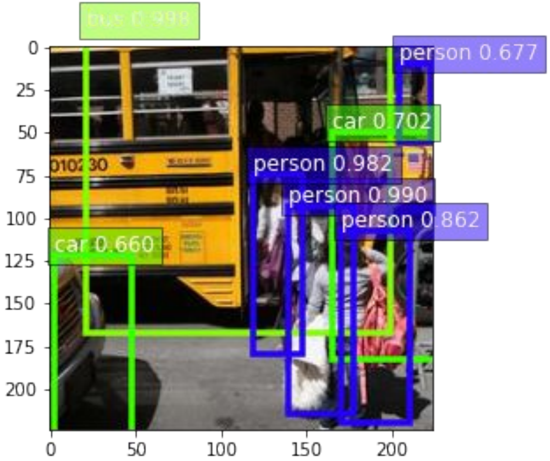
Image

VGG19 (feature maps)

RCNN (bounding boxes)

ROI Pooling

Image Processed Vector



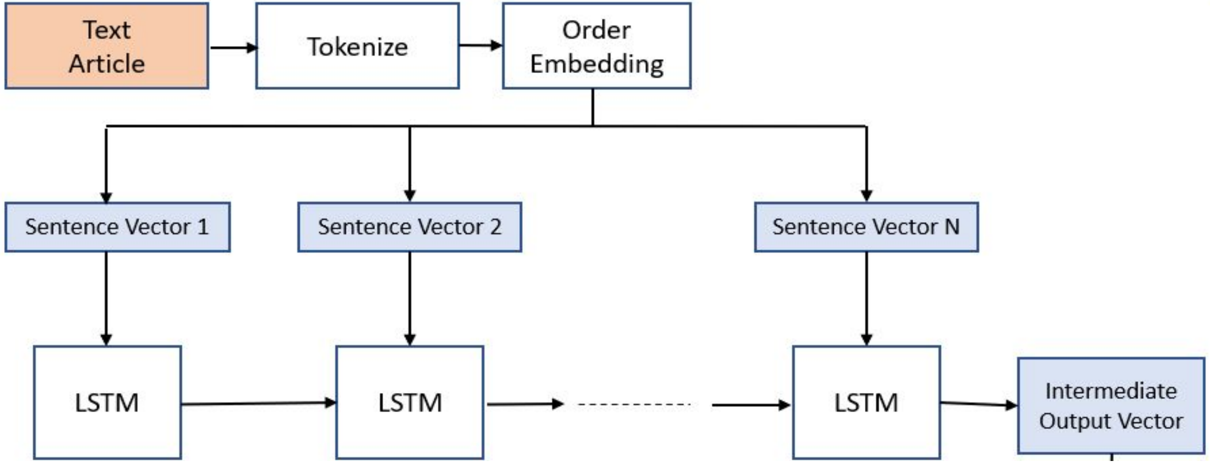
- We first feed it into the VGG-19 model pre-trained on imagenet, to extract the feature map (“block4_conv4” layer) for the image.
- Then, we feed the images into Faster-RCNN as well to get bounding boxes and scores for the image.

- 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.

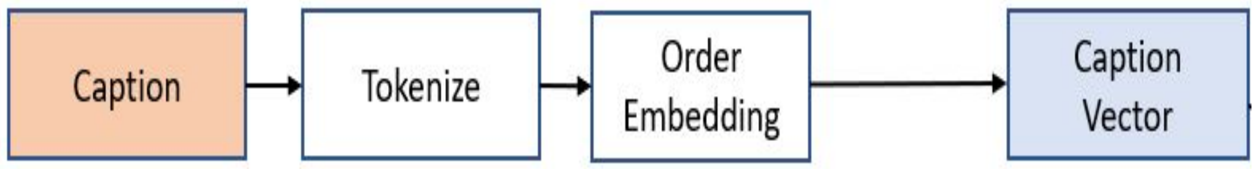
- Then, using the feature maps of the image and ROIs of the image we perform ROI pooling to get an Image Processed Vector.

FOR ARTICLE



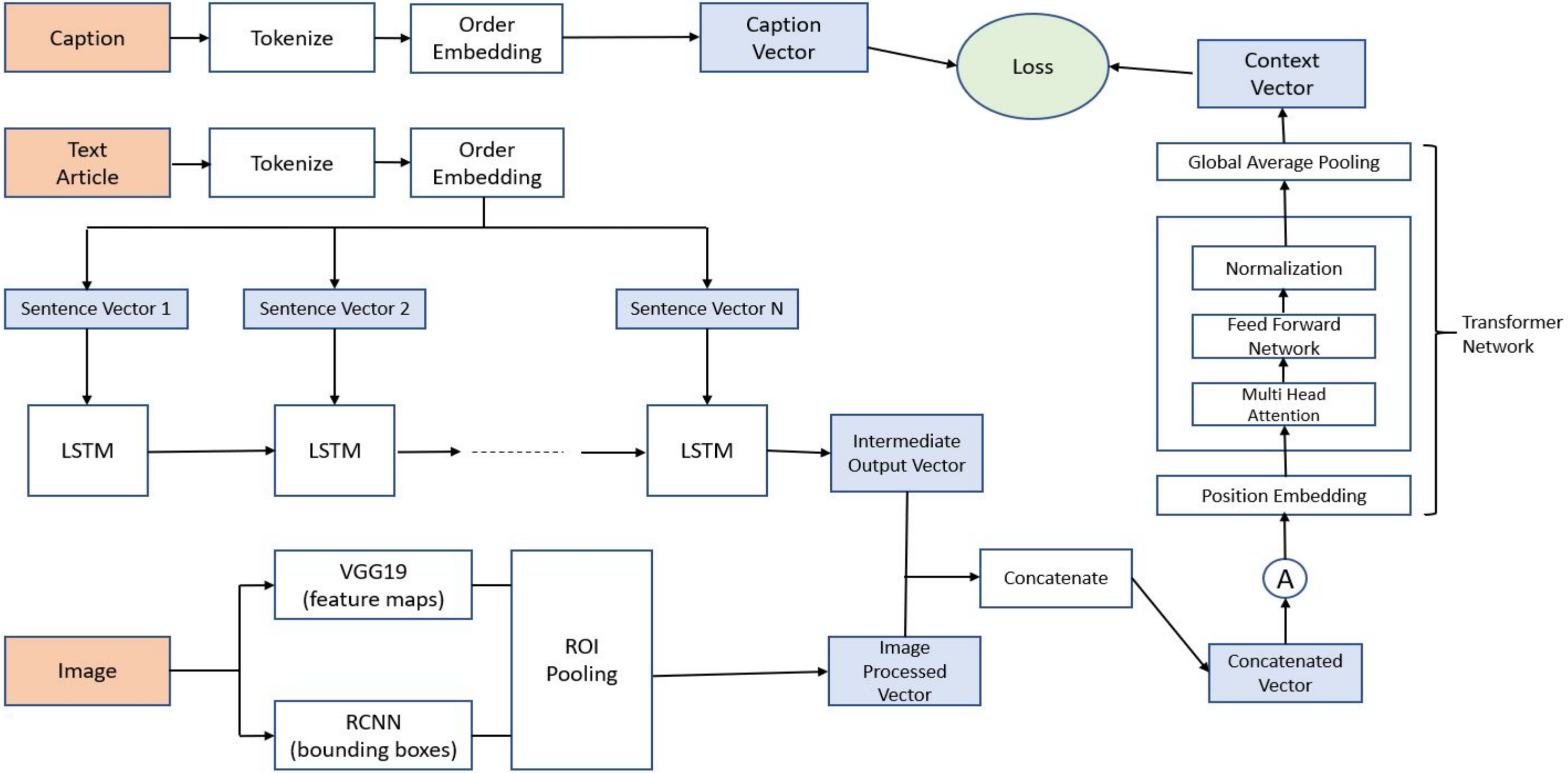
- We tokenize each sentence. After this, we convert these sentences into a sequence of vectors using a pre-trained Order-Embedding model which embeds each sentence.
- Sequence of sentence vectors is fed into to a LSTM network gives us the Intermediate Output vector.

FOR CAPTION



- We first tokenize it and then convert it into a Caption vector using a pre-trained Order-Embedding model .
- It gives us a 1024 length Caption vector.

MODEL ARCHITECTURE

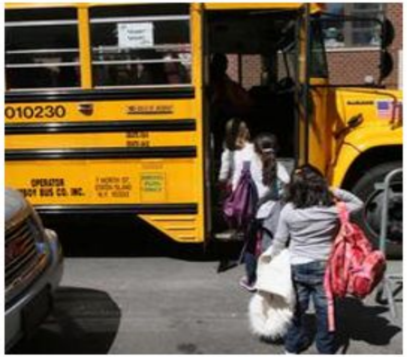


- The Intermediate output vector (received from the article) and the Image processed vector(received from the image) are concatenated to get a Concatenated vector.
- Then, this concatenated vector is fed into transformer network.
- Global average pooling(average of each feature map) is applied on the output of the transformer network.
- Attention layer is used on the output of transformer network and the concatenated vector(consisting of image and article features) to get a Context vector.

TRAINING

- Good News dataset is used (1500 training data points and 109 testing data points)
- The Context vector(output) of the model is used to calculate loss against the Caption Vector(received from caption) and back-propagated to train the model.
- Training details:
  - Loss function : categorical_crossentropy
  - Optimizer : SGD (learning rate = 0.1, momentum = 0.7)
  - Epochs : 20

Results



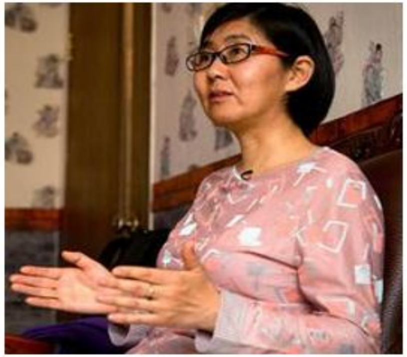
Article : The mountains of snow and the icy roads never arrived last winter, depriving the city's schoolchildren of the precious surprise of a snow day. But the payoff is about to come, a few months later, to thousands of New York City students...........

....

.................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.



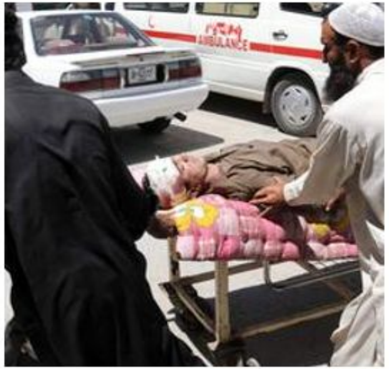
Article : HONG KONG -- The teenage son of a prominent human rights lawyer in China was blocked from leaving the country Monday after the police told him he posed a potential threat to national security while abroad, his father said.............

....

................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.



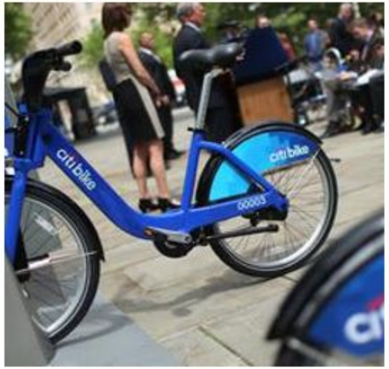
Article : PESHAWAR, Pakistan -- Dozens of people, including two senior security officers, were killed and scores were wounded in a suicide attack on a government checkpoint in a tribal district along the Afghan border, hospital and government officials said. …………

…………

…………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.………

…………..

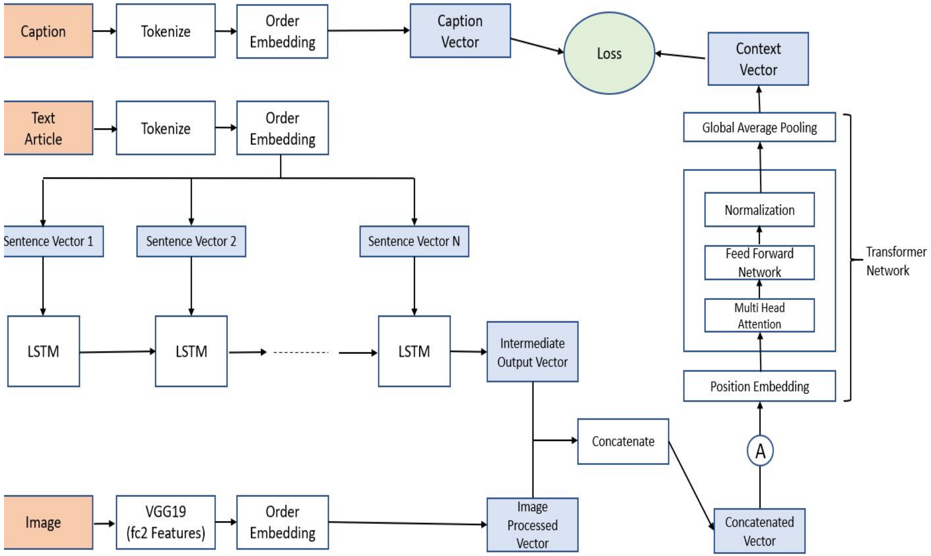
…………..The city's Department of Transportation said it would unveil a map of the bike stations later in the week.

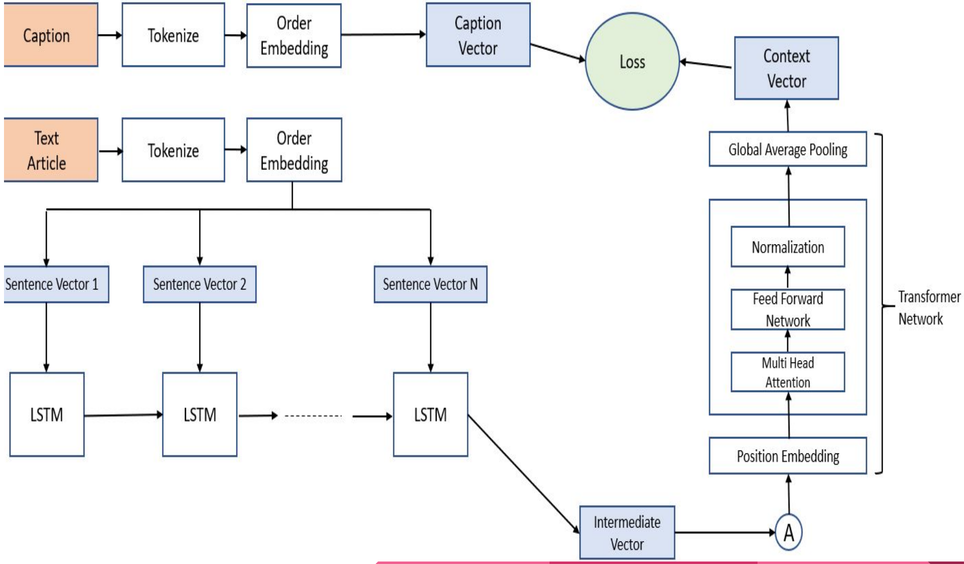
Ground Truth : Citibank is paying $41 million to be lead sponsor of New York’s bike-sharing plan for five years. At the end of July, the first of some 10,000 rental bikes are scheduled to reach the streets.

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.

ABLATION STUDY

- **Transformer + VGG:**
  - For this variation , we use VGG-19(fc2 layer) as an image feature extractor instead of Faster-RCNN and ROI pooling to get full image features.
- **Transformer + Article:**
  - For this variation, we don’t take image features into consideration take articles only as input.



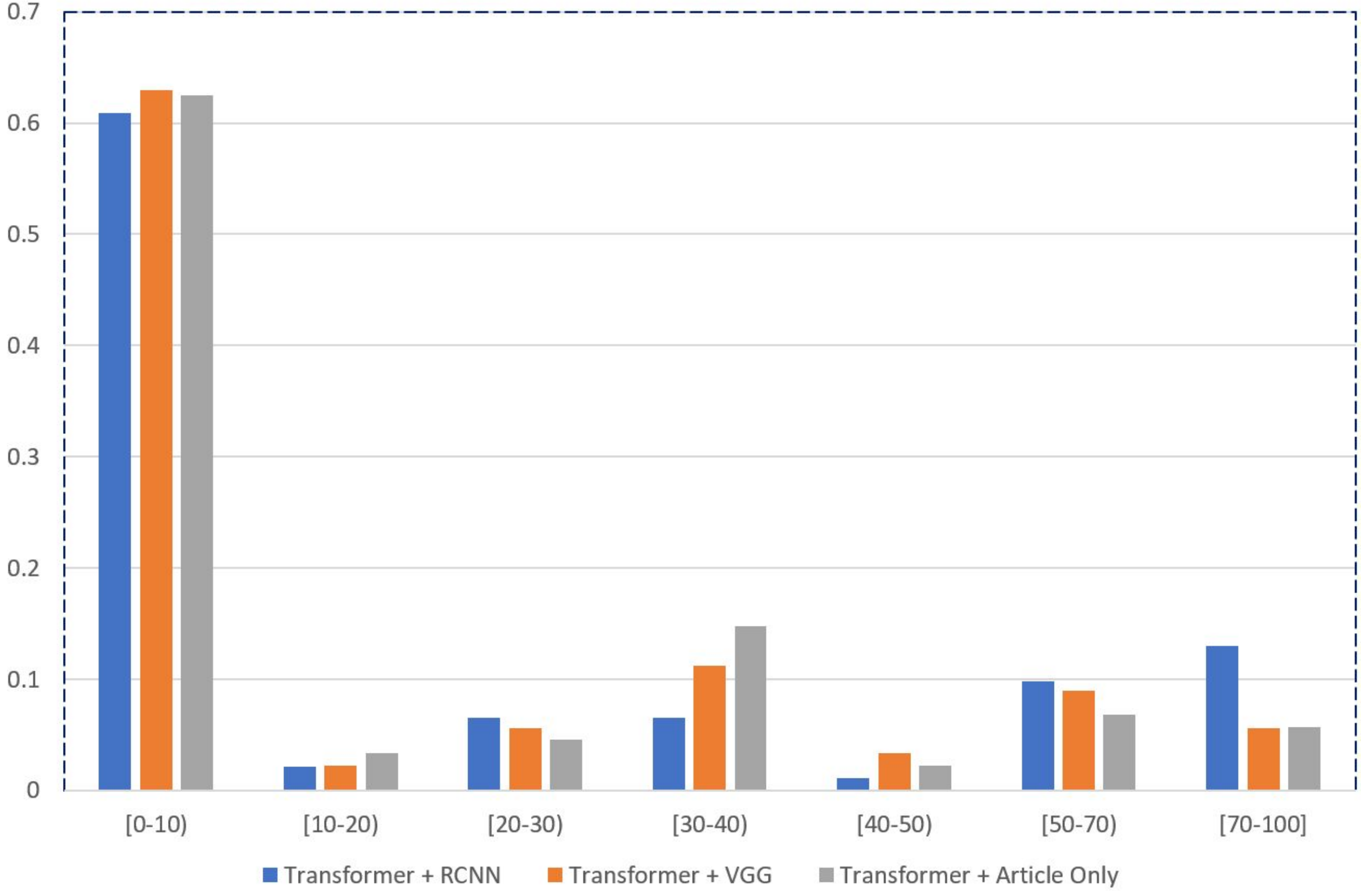


COMPARISON RESULTS

Model	**Parameters**	**BLEU**	**Meteor**	**Cider**	**CoverageNE**
Transformer (RCNN)	123 M	**0.2532**	**0.054**	**0.101**	**0.2235**
Transformer (VGG)	46 M	0.2170	0.044	0.062	0.1692
Transformer (Article Only)	19 M	0.2088	0.043	0.061	0.1679

CoverageNE = [E (predicted) ∩ E (gold)] / [E (gold)]
measure the coverage of named entities in predicted caption.

Distributions of CoverageNE scores for model and its variants.



CONCLUSION

- The experimental results demonstrated our proposed model overperforms the other variants by significant margins, therefore it could integrate visual and textual information in generating captions effectively.
- The experimental results show that the predicted captions can provide primary information to reproduce news-image captions written by journalists
- Our model is flexible with choice of method for extraction of text and image features, thus can be used in comparing the performance of various feature extraction models.