CPSC 8810: Deep Learning

Homework 2 Report

Namita Vagdevi Cherukuru – C32671672 (ncheruk@clemson.edu)

Video Caption Generation

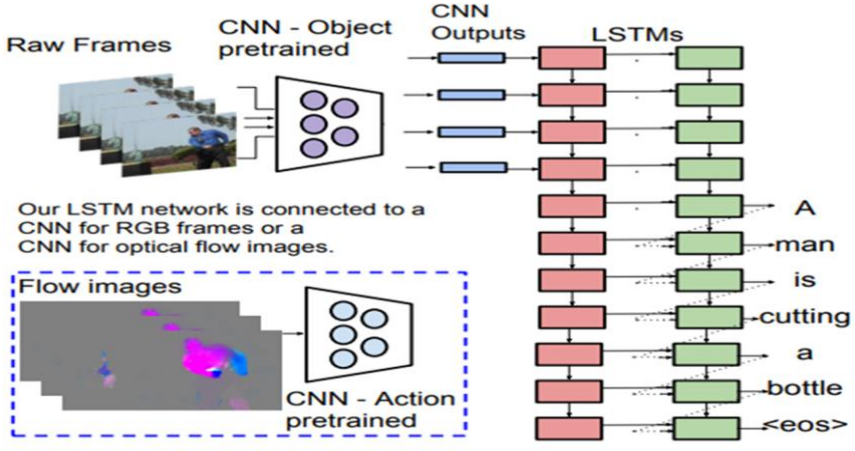
The basic idea of Video Caption Generation is to input a video and get a stream of captions for the actions happening in the video. This is achieved by following Deep learning techniques: Sequence to sequence model and training of the dataset provided.

Requirements:

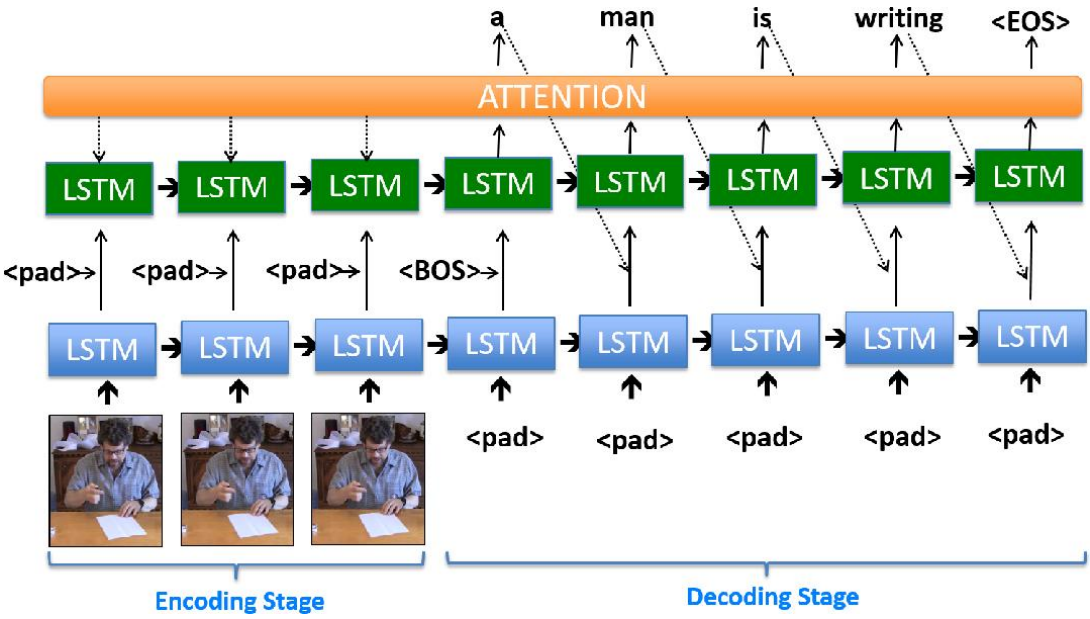
- Python 3
- Tensorflow 1.15
- MSVD Dataset (1450 videos for training, 100 for testing)
- Sequence to sequence model

Processing of Video to Captions:

- The model uses 2 layers of LSTM (RNNs), in the first layer of the RNN, the Video is encoded, and an output is generated with the help of the decoder.
- In the decoding process, tokens are created to segment the captions based on the beginning and ending verse and perform processing over the video to produce the actual words based on the pictorial representation using a library.
- As part of sequence to sequence model, use of S2VT is to map the video frames to the words.
- The following picture depicts the same,



- While mapping the words and creating a caption, it is made sure that there is no anomaly in generation with the help of beam search which helps lean towards one path while generating.
- To reduce the exposure issues, it is important to give groundtruth as feed while processing at different input odds.
- We can reach the baseline of the model by computing the BLEU value which is based on the length and occurrences of the captions. BLEU score will unlock the efficiency at which the captions are generated.
- Attention mechanism can help enable sequence to sequence model to function effectively. Without the attention mechanism, BLEU score is 0.9854, with the attention mechanism, the BLEU score is 0.7854.



Experiment and Results:

- The experimental parameters are the following for one of the scenarios:
  rnn_size = 1089
  num_of_layers=2 (LSTM)
  dim_video_feat = 2963
  embed_size = 5698
  lr_rate = 0.001 (learning rate)
  batch_size = 100
  max_gradnorm = 6 (gradient norm)
  max_encoder_steps = 125
  max_decoder_steps = 26
  sample_size = 1090

dim_video_frame = 100

Method	BLEU Score
Without Attention	0.9855
Attention with scale	0.7584
Attention with gradient norm	0.7100

- One of the results were following:

<BOS> a man

<BOS> a man is

<BOS> a man is

<BOS> a man is feeding

<BOS> a man is feeding a

<BOS> a man is feeding a dog

<BOS> a man is feeding a dog.

Second data testing:

rnn_size = 1024
num_of_layers=2 (LSTM)
dim_video_feat = 1859
embed_size = 2564
lr_rate = 0.0001 (learning rate)
batch_size = 85
max_gradnorm = 4 (gradient norm)
max_encoder_steps = 98
max_decoder_steps = 20
sample_size = 985
dim_video_frame = 85

Method	BLEU Score
Without Attention	0.7026
Attention with scale	0.6127
Attention with gradient norm	0.6917

Third data testing:

rnn_size = 985
num_of_layers=2 (LSTM)
dim_video_feat = 1045
embed_size = 2012
lr_rate = 0.0001 (learning rate)
batch_size = 50
max_gradnorm = 5 (gradient norm)
max_encoder_steps = 70
max_decoder_steps = 18
sample_size = 1852
dim_video_frame = 62

Method	BLEU Score
Without Attention	0.6924
Attention with scale	0.5988
Attention with gradient norm	0.7154