Deep Learning

Name: Ponugoti vinay

CUID: C13375304
Email: vponugo@g.clemson.edu
GitHub Link: https://github.com/Vinay-ponugoti/DeepLearning/tree/main/HW2/HW2_1

Introduction:

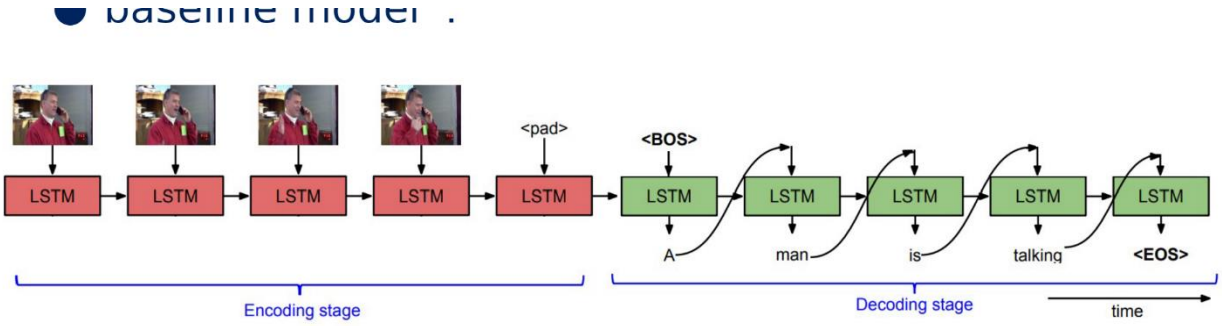
The generation of video captions in a naturalistic manner has consistently posed a challenge for both natural language processing and computer vision. RNNs have shown effectiveness in interpreting visual sequences, as they can capture the dynamics of sequences over time. However, creating video descriptions requires handling varying lengths of input frames and word sequences. Moving away from template-based methods, this model uses a Subject-Object-Verb format to generate captions based on the likelihood of correct word matches. While this approach is useful, its fixed structure can lead to irrelevant captions and restrict real-time adaptability. To address these issues, a dynamic sequence-to-sequence model utilizing Long Short-Term Memory (LSTM) will be implemented. This model will process video frames and produce flexible, variable-length text while continuously learning from new visual data.

Recent advancements in deep learning, particularly with RNNs and CNNs, have greatly propelled the field of natural language processing. LSTM networks have shown remarkable success in sequence-to-sequence tasks such as speech recognition and machine translation. A common method involves using a stacked LSTM to encode input frames in sequence. However, employing mean-pooling to create a single feature vector from CNN outputs has a significant limitation—it does not capture temporal information or maintain the order of video frames. To overcome this issue, researchers have suggested using Conditional Random Fields (CRFs) to produce semantic tuples (like activity, object, and location) that LSTM models can then convert into phrases. Another approach introduces a hierarchical recurrent neural encoder with an added temporal filter layer made up of LSTM cells. Additionally, attention mechanisms have gained popularity, enhancing the performance of neural networks in various tasks such as machine translation, image captioning, and video captioning.

Method Applied:

Deep Learning

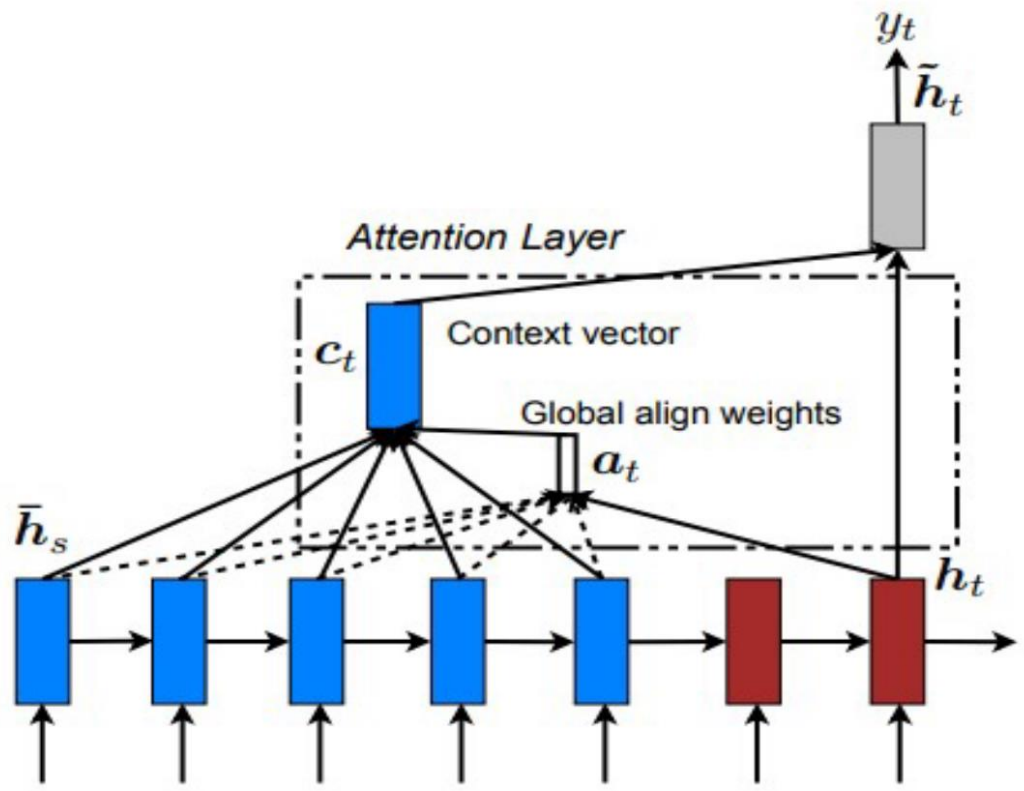
The model analyzes a sequence of inputs to create a series of outputs. By utilizing the pre-trained CNN VGG19, it produces an 80x4096 spatial feature map that captures detailed nuances from video frames. The LSTM architecture is designed to manage the variability in input and output sizes. An attention mechanism is incorporated, allowing the decoder to concentrate on the most pertinent parts of the input sequence, which improves the interaction between the encoder and decoder. This combined method enables the model to produce natural language descriptions of video content effectively.



Attention Layer:

Attention is a machine learning mechanism that enables the decoder to focus on relevant and significant segments of the input sequence by utilizing information from each hidden state of the encoder. When dealing with lengthy input sequences or scenarios where the lengths of the input and output sequences differ, this approach is very helpful. It is now widely used in many projects related to natural language processing to improve the accuracy and fluency of generated text, and it has grown in popularity.

Deep Learning



Data Preprocess:

Data preparation uses the tokens listed below.

- Dictionary - most frequently word or min count
- other tokens : <PAD>, <BOS>, <EOS>, <UNK>
  - <PAD> : Pad the sentencen to the same length
  - <BOS> : Begin of sentence, a sign to generate the output sentence.
  - <EOS> : End of sentence, a sign of the end of the output sentence.
  - <UNK> : Use this token when the word isn't in the dictionary or just ignore the unknown word.

Deep Learning

Model evaluation & parameters:

HyperParameter	Value
Learning rate	0.001
Use_attention	True
Max_encoder_steps	64
Max_decoder_steps	15
Embedding_size	1024
Batch_size	50
Beam_size	5(if beam search is True)

Result:

Before implemented our model, the mean BLEU score was 0.27. But it has now risen to 0.55. After 200 epochs, the BLEU score saw further improvement, reaching around 0.58.