CPSC 8430: Deep Learning
HW2
Submitted by: *Abyad Enan*

**Git Hub Link: https://github.com/AbyadEnan/CPSC-8430-Deep-Learning/tree/main/hw2/hw2_1**

**Objective:**

The objective of this project is to generate captions for videos based on the actions of the persons or objects in the videos.

**What is video captioning?**

A sequence-to-sequence model is created to create model, train, and test for generating captions for videos, where the captions represent the action, done by persons/objects. This report briefly explains how to achieve the goal.

**Dataset:**

MSVD Dataset (Number of videos for; training: 1450, testing:100)

**Tools:**

- Python 3.9.7
- Pytorch 1.10.2
- Pandas 1.4.1
- Scipy 1.8.0

**Dictionary:**

A dictionary is created for this work from the labeled file provided in the dataset. The dictionary is created using few tokens in the process as follows:

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.

**Model:**

The baseline model is as follows:

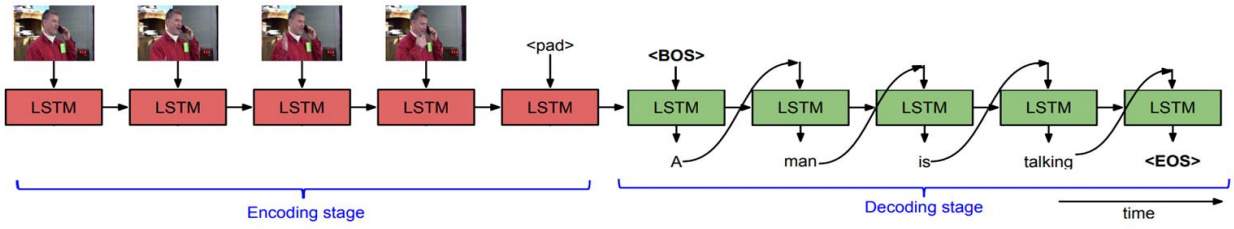


Figure1: Sequence-to-sequence model

To get better result, an attention layer is added. The attention layer is implemented on the encoder's hidden state. The decoder's hidden state and the encoder's output are utilized as a matching function to generate a scalar that travels through the softmax layer before being passed to the next time step of the decoder. The attention layer is as follows:

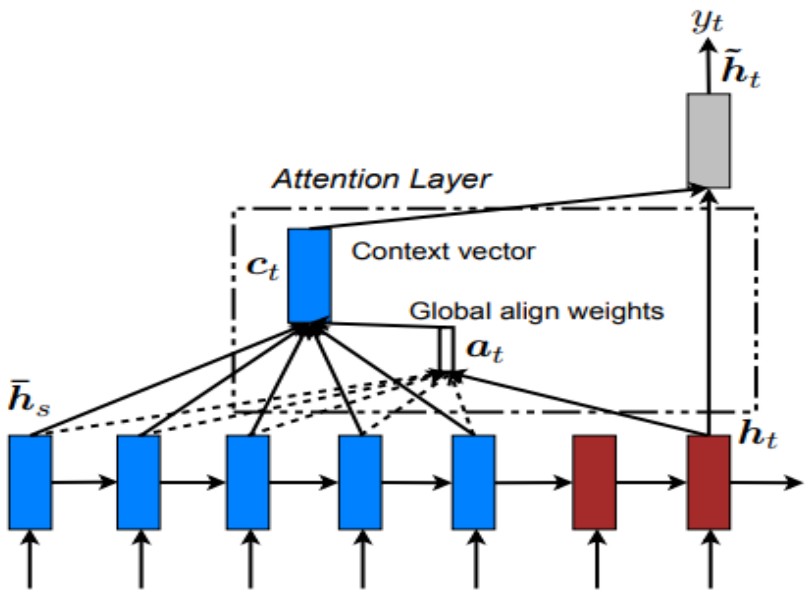


Figure 2: Attention layer

A problem, called exposure bias, might arise while training. This problem is resolved by using ground truth as input. The process is depicted in the figure as follows:

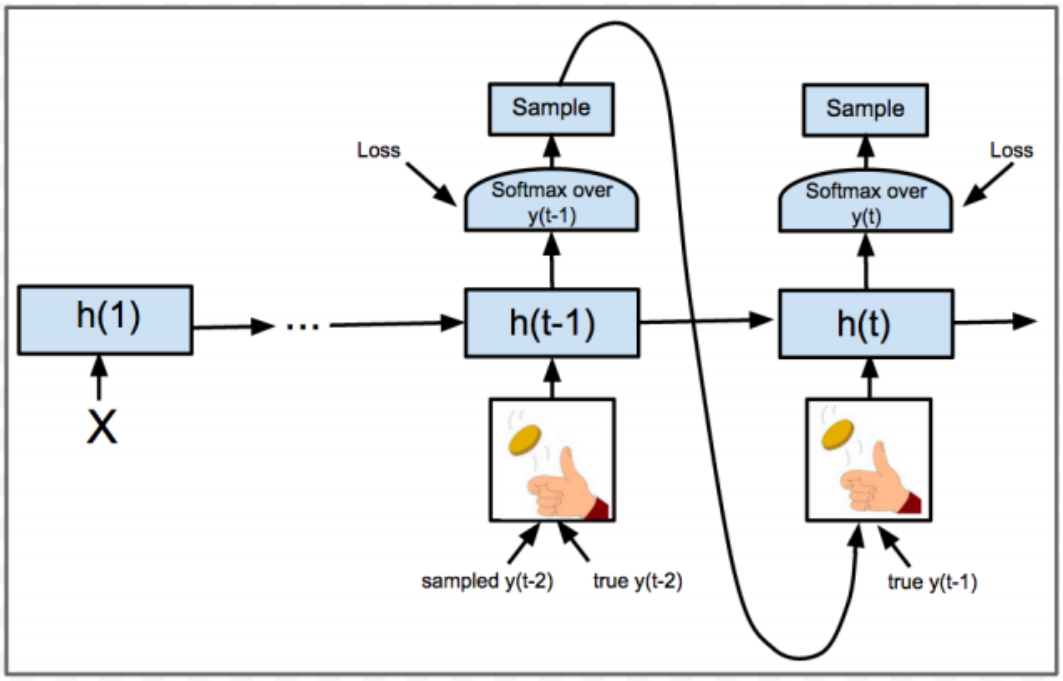


Figure 3: Schedule Sampling

The rest of the process is summarized as follows:

Loss function: CrossEntropyLoss

BLEU score of the model: bleu_eval

Models: 10 models are created by varying the parameters- hidden layer sizes, batch sizes, dropout percentages and word dimensions. Among them, 4 models are trained by 10 epochs, depicted in the table below. For training, the model wit highest BLEU score is used for 30 epochs to make the final model.

| Models | Learning rate | Size of batch | Size of hidden layer | Dropout % | BLEU score |
|---|---|---|---|---|---|
| **model1** | 0.001 | 16 | 128 | 0.4 | 0.695791746 |
| **model2** | 0.001 | 16 | 128 | 0.2 | 0.690616515 |
| **model3** | 0.001 | 32 | 128 | 0.3 | 0.678419199 |
| **model4** | 0.001 | 16 | 128 | 0.3 | 0.623407207 |