CPSC 8430: Deep Learning

Homework 2

Video Caption Generation using S2VT

Reek Majumder

Email : rmajumd@clemson.edu

GitHub:- https://github.com/reek129/CPSC8430_HW2_V2

GitHub Based on Hw2 requirements: - https://github.com/reek129/CPSC_8430--Deep_Learning_HW1

Introduction

In this project, we present, train, and test a sequential-to-sequential model to generate a caption for videos. The basic concept behind Video Caption Generation is that we can upload a video and get a stream of captions for the actions that are taking place in the video. This is accomplished by using deep learning techniques and training on the provided dataset.

Requirements:

Python -3.9.7
torch 1.10
SciPy 1.7
MSVD Dataset (1450 videos for training and 100 for testing)

Note: - For this project we are using the feat folder which is the extracted features from Video data and used it for training and testing purposes.

Dictionary:

The first step in our project is to load the label file and create a dictionary. All of the critical words from captions are stored in the dictionary. We test the number of times a word is repeated in a video caption and encode a word to a unique index and vice versa for all caption data to determine the importance of the word for a specific video.
The following are some of the tokens used to store data in the dictionary:

- <PAD>:- Pad the sentences to the same length to maintain uniformity.
- <BOS>:- Beginning of the sentence, an identifier to generate the output sentence.
- <EOS>:- End of Sentence, an identifier to signal the system end of output sentence.
- <UNK>:- Use this token when the word isn't in the dictionary or just ignore the unknown word.

Base Model

The seq2seq (S2VT) model's baseline consists of two Recurrent Neural Networks (RNNs) layers. The "encoderRNN" class in the python script is responsible for processing and encoding the videos in the first layer. The "decoderRNN" class is responsible for decoding and generating the output in the second layer. "decoderRNN" is written to segment captions using tokens based on the beginning and ending verses of a sentence, then process the video to generate the actual words as the output(Figure 1). The process of encoding and decoding using the "encoderRNN" and "decoderRNN" classes is depicted in the diagram (Figure2).

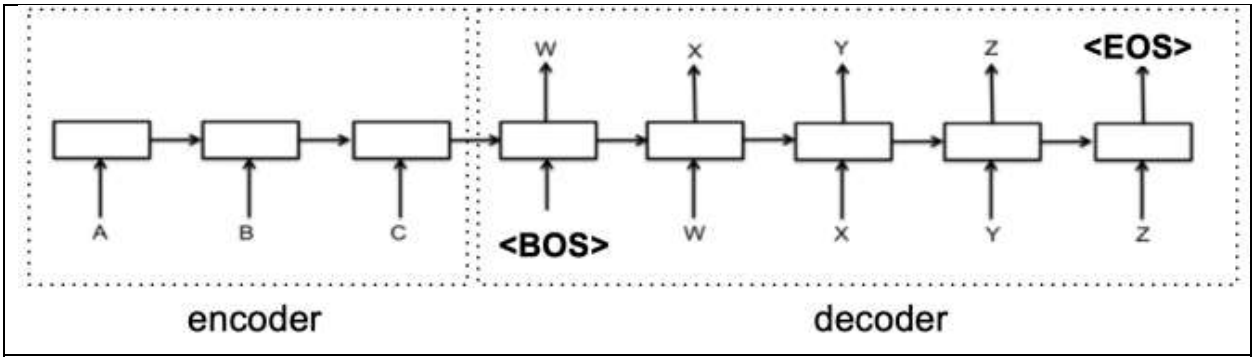

Figure 1

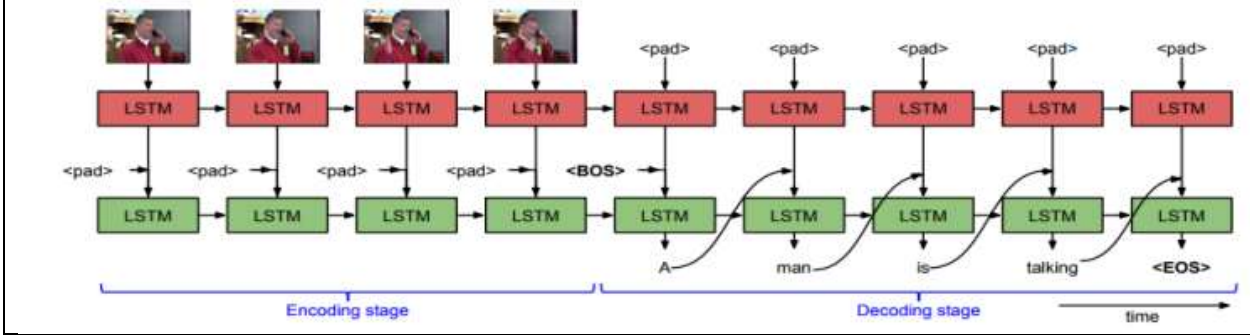

Figure 2

Attention Layer

- An attention layer was implemented on encoder hidden states to improve the base model's performance. At each decoding time step, the model can peek at different sections of the inputs.
- The decoder's hidden state and encoder's output are used as a matching function to generate a scalar, which is then passed through softmax, and the decoder's last hidden state is sent to the next time step.

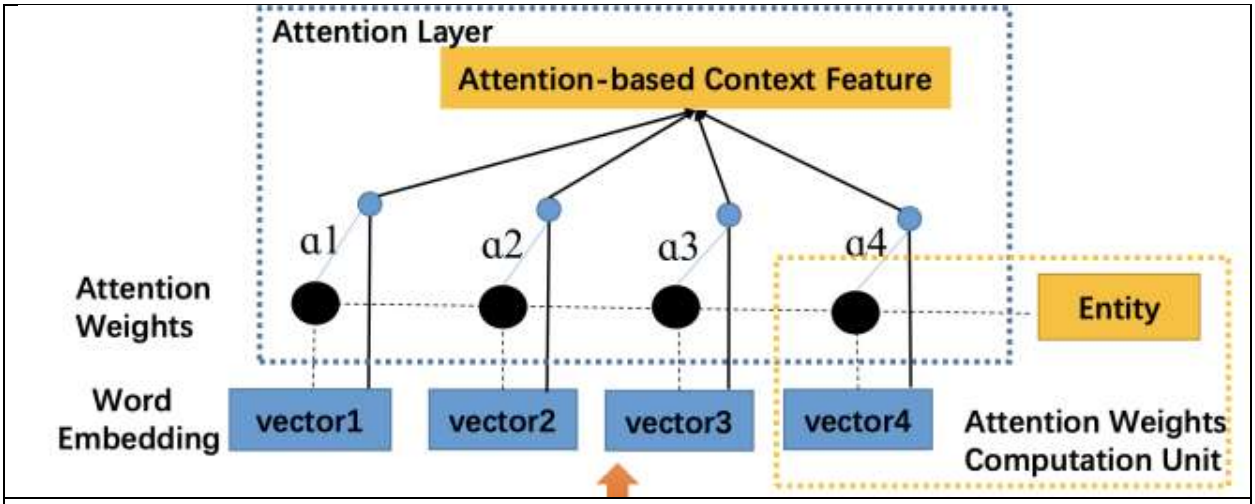

Figure 3

Schedule Sampling

For inference, previous unknown tokens get replaced by the tokens generated by the model. This can cause errors to be accumulated over the sequences.

Results

In SavedModel_reek_v2 we have 42 models with size less than 100 MB.

Model Time for different batch size [16, 32] dropout rate [0.1,0.2,0.3,0.4], hidden size [128,256,512] and word dimension [1024,2048], learning rate 0.001 and minimum vocab size 3

Model Name	time
model_v2_batchSize_16_hidSize_128_dropPer_0.1_wordDim_1024	181.7233012
model_v2_batchSize_16_hidSize_128_dropPer_0.1_wordDim_2048	182.661222
model_v2_batchSize_16_hidSize_128_dropPer_0.2_wordDim_1024	180.6251535
model_v2_batchSize_16_hidSize_128_dropPer_0.2_wordDim_2048	183.2000308
model_v2_batchSize_16_hidSize_128_dropPer_0.3_wordDim_1024	181.8157299
model_v2_batchSize_16_hidSize_128_dropPer_0.3_wordDim_2048	182.9103966
model_v2_batchSize_16_hidSize_128_dropPer_0.4_wordDim_1024	182.5373514
model_v2_batchSize_16_hidSize_128_dropPer_0.4_wordDim_2048	180.0731709
model_v2_batchSize_16_hidSize_256_dropPer_0.1_wordDim_1024	181.3336871
model_v2_batchSize_16_hidSize_256_dropPer_0.1_wordDim_2048	182.4557693
model_v2_batchSize_16_hidSize_256_dropPer_0.2_wordDim_1024	183.1876993
model_v2_batchSize_16_hidSize_256_dropPer_0.2_wordDim_2048	183.0412009
model_v2_batchSize_16_hidSize_256_dropPer_0.3_wordDim_1024	183.0597632
model_v2_batchSize_16_hidSize_256_dropPer_0.3_wordDim_2048	183.3421581
model_v2_batchSize_16_hidSize_256_dropPer_0.4_wordDim_1024	182.8700683
model_v2_batchSize_16_hidSize_256_dropPer_0.4_wordDim_2048	181.9862154
model_v2_batchSize_16_hidSize_512_dropPer_0.1_wordDim_1024	185.3970366
model_v2_batchSize_16_hidSize_512_dropPer_0.1_wordDim_2048	188.8678737
model_v2_batchSize_16_hidSize_512_dropPer_0.2_wordDim_1024	188.3107727
model_v2_batchSize_16_hidSize_512_dropPer_0.2_wordDim_2048	189.1402884
model_v2_batchSize_16_hidSize_512_dropPer_0.3_wordDim_1024	189.354532
model_v2_batchSize_16_hidSize_512_dropPer_0.3_wordDim_2048	187.4530709
model_v2_batchSize_16_hidSize_512_dropPer_0.4_wordDim_1024	188.2411404
model_v2_batchSize_16_hidSize_512_dropPer_0.4_wordDim_2048	188.0374002
model_v2_batchSize_32_hidSize_128_dropPer_0.1_wordDim_1024	126.7612121
model_v2_batchSize_32_hidSize_128_dropPer_0.1_wordDim_2048	126.6435373
model_v2_batchSize_32_hidSize_128_dropPer_0.2_wordDim_1024	127.3719897
model_v2_batchSize_32_hidSize_128_dropPer_0.2_wordDim_2048	125.8590796
model_v2_batchSize_32_hidSize_128_dropPer_0.3_wordDim_1024	125.9560549
model_v2_batchSize_32_hidSize_128_dropPer_0.3_wordDim_2048	126.4872873
model_v2_batchSize_32_hidSize_128_dropPer_0.4_wordDim_1024	127.2450922
model_v2_batchSize_32_hidSize_128_dropPer_0.4_wordDim_2048	126.6008744
model_v2_batchSize_32_hidSize_256_dropPer_0.1_wordDim_1024	127.7535567
model_v2_batchSize_32_hidSize_256_dropPer_0.1_wordDim_2048	127.3069122
model_v2_batchSize_32_hidSize_256_dropPer_0.2_wordDim_1024	128.5894663

model_v2_batchSize_32_hidSize_256_dropPer_0.2_wordDim_2048	127.8471286
model_v2_batchSize_32_hidSize_256_dropPer_0.3_wordDim_1024	127.7121215
model_v2_batchSize_32_hidSize_256_dropPer_0.3_wordDim_2048	128.2831359
model_v2_batchSize_32_hidSize_256_dropPer_0.4_wordDim_1024	127.1641772
model_v2_batchSize_32_hidSize_256_dropPer_0.4_wordDim_2048	127.8938625
model_v2_batchSize_32_hidSize_512_dropPer_0.1_wordDim_1024	129.754009
model_v2_batchSize_32_hidSize_512_dropPer_0.1_wordDim_2048	129.7946036
model_v2_batchSize_32_hidSize_512_dropPer_0.2_wordDim_1024	131.0556335

Model Vs Bleu Score for 5 epochs

Bleu Score	**Model Location**
0.704488	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_256_dropPer_0.2_wordDim_2048.h5
0.699342	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_128_dropPer_0.2_wordDim_2048.h5
0.695792	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_128_dropPer_0.4_wordDim_1024.h5
0.693332	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_128_dropPer_0.4_wordDim_2048.h5
0.690617	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_256_dropPer_0.4_wordDim_1024.h5
0.690617	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_512_dropPer_0.2_wordDim_1024.h5
0.690617	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_512_dropPer_0.3_wordDim_1024.h5
0.690617	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_128_dropPer_0.2_wordDim_2048.h5
0.690617	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_512_dropPer_0.4_wordDim_1024.h5
0.690617	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_256_dropPer_0.1_wordDim_2048.h5
0.690617	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_256_dropPer_0.1_wordDim_1024.h5
0.690617	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_128_dropPer_0.1_wordDim_2048.h5
0.690617	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_512_dropPer_0.1_wordDim_2048.h5
0.690617	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_256_dropPer_0.3_wordDim_2048.h5
0.690617	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_128_dropPer_0.2_wordDim_1024.h5
0.680368	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_256_dropPer_0.3_wordDim_1024.h5

0.678969	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_256_dropPer_0.4_wordDim_2048.h5
0.678419	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_128_dropPer_0.3_wordDim_2048.h5
0.677991	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_128_dropPer_0.2_wordDim_1024.h5
0.67757	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_128_dropPer_0.3_wordDim_1024.h5
0.674488	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_128_dropPer_0.1_wordDim_2048.h5
0.674419	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_256_dropPer_0.4_wordDim_1024.h5
0.673517	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_256_dropPer_0.3_wordDim_2048.h5
0.672665	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_128_dropPer_0.1_wordDim_1024.h5
0.670595	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_256_dropPer_0.3_wordDim_1024.h5
0.666792	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_128_dropPer_0.4_wordDim_2048.h5
0.661986	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_256_dropPer_0.1_wordDim_2048.h5
0.660553	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_128_dropPer_0.1_wordDim_1024.h5
0.659637	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_256_dropPer_0.2_wordDim_1024.h5
0.658776	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_256_dropPer_0.1_wordDim_1024.h5
0.656419	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_512_dropPer_0.4_wordDim_2048.h5
0.656419	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_256_dropPer_0.2_wordDim_2048.h5
0.656419	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_512_dropPer_0.1_wordDim_1024.h5
0.652983	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_128_dropPer_0.4_wordDim_1024.h5
0.623407	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_128_dropPer_0.3_wordDim_1024.h5
0.615637	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_512_dropPer_0.2_wordDim_1024.h5
0.600094	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_256_dropPer_0.2_wordDim_1024.h5
0.59977	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_512_dropPer_0.1_wordDim_1024.h5
0.59977	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_512_dropPer_0.2_wordDim_2048.h5

0.591742	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_128_dropPer_0.3_wordDim_2048.h5
0.357386	SavedModel_reek_v2/model_v2_batchSize_32_hidSize_256_dropPer_0.4_wordDim_2048.h5
0.138889	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_512_dropPer_0.3_wordDim_2048.h5
0.138889	SavedModel_reek_v2/model_v2_batchSize_16_hidSize_512_dropPer_0.4_wordDim_1024.h5

Best Model Comparison

Top two models with higher Bleu Score was iterated for 50 epochs

Model Name	Time
model_v2_batchSize_32_hidSize_128_dropPer_0.2_wordDim_2048	2324.841
model_v2_batchSize_32_hidSize_256_dropPer_0.2_wordDim_2048	2261.451

Bleu Score	Model Name
model_v2_batchSize_32_hidSize_128_dropPer_0.2_wordDim_2048	0.569901
model_v2_batchSize_32_hidSize_256_dropPer_0.2_wordDim_2048	0.663685 (**Best Model**)