Will Sherrer
Dr Feng Lou
March 25, 2022

This implementation is an S2VT model for video captioning for the MSVD dataset. The main files I created are:

Dictionary.py: this file contains the vocabulary that the model will use. It parses through the training data to build a vocabulary.

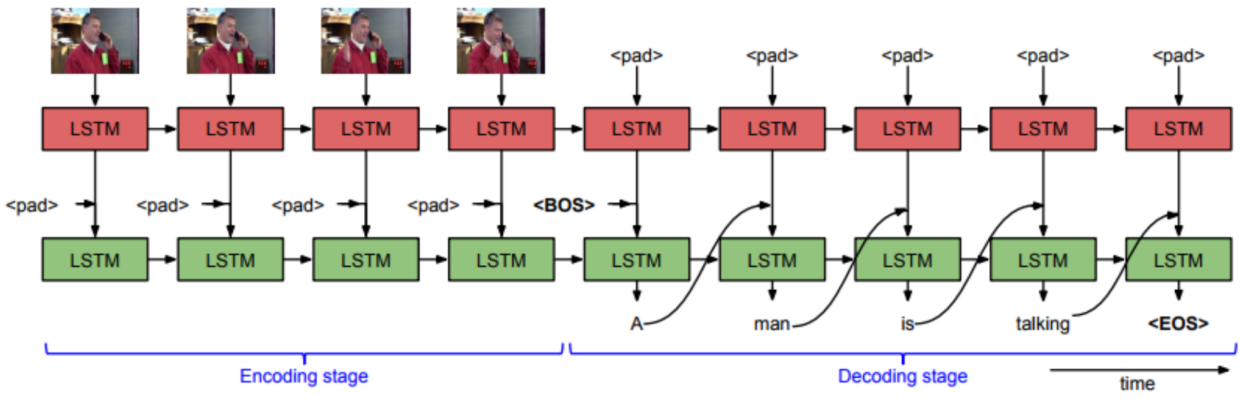
S2VT_model.py: this file contains the S2VT model.

Model_seq2seq.py: contains the testing and training code. This contains hyperparameters and also prints the output of the model to an output file. When submitting the num_epochs will be set to 0 to skip testing and the model will load parameters already trained.

Utils.py: this file contains functions for other tasks like fetching data for example

When training the model for evaluation I have decided to only train for 00 epochs. The output is not as good as it should be however it does exceed the baseline. The training dictionary is shuffled every epoch in a small attempt to counteract overfitting, as well as the caption that will be used as the ground truth for that video. This is to feed the model with differing ground truths in an effort to expose it to more of the vocabulary while training.

I trained with 2 different batch sizes, several epochs at 50 and several at 32, this is to try and help the model specialize better as well as generalize. For Final training I trained with a batch size of 10.



Issues: When training for only several epochs my relative loss is average however the model outputs the same sentence for each caption. This can be fixed with more training however with owning a Macbook Pro which does not have access to cuda and uses CPU training times are very slow. I had several issues while building the model which I thankfully was able to resolve. I

believe the most difficult part of this was taking the general knowledge of how this model should work and putting it into code and figuring out little details such as padding. I had many issues when implementing the batch size which required me to change how my training, testing, data fetching, dictionary, and model all worked. However, I was able to successfully solve that. I also ran into the problem of long training times. Since my mac doesn't use cuda, and my laptop is a few years old, It takes a long time to train, so while I might have finished my model earlier I spent many days at the end training and tweaking parameters and retraining.

Credit: https://vsubhashini.github.io/s2vt.html - model architecture