Homework 2 CPSC 8430

Andrew Wright

GitHub Link: https://github.com/andrewright17/Homework-2-Deep-Learning

Model

This seq2seq model was developed to generate captions for video input from the MSVD dataset. The model architecture is an encoder-decoder LSTM model with an attention mechanism that had a hidden_size of 512.

Training

The model was trained for 100 epochs on 1450 videos with a learning rate of 0.0001. A build_dictionary function was used to construct mappings for words to indices (and vice versa) for words used more than 4 times among the training captions. The labels were also preprocessed using regex to clean up punctuation that may have complicated the training.

Testing

Testing the model on 100 videos showed an average BLEU evaluation score of 0.684, which is above the baseline of 0.6. This shows that the model is accurately generating captions, but there is still room for improvement. Likely, I would make some changes to the architecture by adjusting the dimensions of the LSTM and attention layers as a first step. Below is an example of a generated caption from the model.

Example frame from a test video:

True captions:

[['A woman cuts an oion.', 'A girl is cutting an onion.', 'A girl is cutting an onion.', 'A girl is cutting onion into small pieces.', 'A girl is dicing up an onion.', 'A woman chops a white onion.', 'A woman cuts and dices an onion.', 'A woman is chopping an onion into small pieces.', 'A woman is chopping an onion slice into fine pieces.', 'A woman is cutting onions.', 'A woman is slicing a onion.', 'A Woman is slicing some vegetables.', 'A woman slices an onion.', 'An onion is chopped up.', 'The cook is dicing onions.', 'The girl is slicing an onion.']]

Model caption:

"a young woman is cutting a something of an something"