Nathaniel Todd and Colton Current

CS 497 - Neural Networks

Final Report

May 3rd, 2024

A Project Working With English Speech To French Text

**Overview:**

For our project we wanted to try and tackle the issue of language barriers. We wanted to try and take English speech and translate it into French text. Originally, we wanted to try and translate sentences of any length, but after some testing we swapped to a max length of 20 words. Our project has two main sides: the speech to text and the translator. Both sides were vital to our vision for the final product.

**Data Description:**

For our speech to text side of our project, we needed to find a dataset that has a lot of recordings of different words. We already had access to one dataset that had about 10 different commands and decided to try and expand on that idea. Quickly we found a few different sets including another, larger command set and a set of common words. We choose the larger command dataset as the common words set would have needed a lot more time in order to just obtain the set. The larger dataset we used was called the “Speech Command Dataset” and can be found on the hugging face website (<https://huggingface.co/datasets/speech_commands>) and contains 31 (for version 0.01) or 36 (for version 0.02) different word classes. Each version was made up of thousands of recordings in a .WAV format. We decided to work with version 0.02 for our project and had to do a decent amount of preprocessing before we used it.

To preprocess our data, we first had to read it using the dataset library from hugging face. We then had to sub-section our data to just the files in the train section and grab only 1500 from each class. After that, we had to transform these files into spectrograms so we could begin training. This is where we ran into issues, as for some reason we could not determine, once the .WAV files were transformed into spectrograms there were some that were different sizes than the majority. To account for this, we only took the spectrograms of a certain size to use for our training.

For our translation side of our project, we wanted to find a dataset that included a large vocabulary for both english and french translations. The idea behind this was to find a dataset that could work with any sentences we threw at our machine translator. For our dataset we chose the main dataset of the Workshop of Statistical Machine Translation which can be found on Kaggle(<https://www.kaggle.com/datasets/dhruvildave/en-fr-translation-dataset>) and contains 22.5 million english sentences and their translations. For the training of our actual model, we only used 500,000 to 5,000,000 of these sentences before preprocessing our dataset of all special/unusable characters and filtering out sentences smaller than our max length.

**Testing:**

To test our project, we used hyperparameter tuning similar to gridsearch, saved those models, and used the BLEU translation testing metric to see how our different models held up against a pretrained translator from the NeMo collections.

**Limitations:**

When training our different models and testing out different max lengths with different batch sizes, we found that we would always run out of memory eventually. This would cause the program to crash but we would save the results and models that were trained before the crash. Depending on the model our project only has the capacity to translate sentences from 1 to 20 or 1 to 48 words. Going past that would result in only part of the sentence being translated.

**Future Work:**

Continuing this project we want to try and increase the validity of our translation model and its outputs. To do this we would want to start with using a different dataset with only the words in our speech recognizer dataset and their translations. Once this is done, I believe we would need less data to train the models, and we could successfully test more hyperparameters without having memory errors. We also want to make use of the Silence class for our speech to text model. To do this we need to sample some background noise from the environment in which the would be run, such as classrooms or lecture halls. This would decrease the chance of our speech to text to detect the wrong word and break our sentence structure. We would also like to find a dataset that is not so limiting in terms of the words it contains. To accomplish this we would need to either source another dataset or create one of our own.