# Homework 3 CPSC 8430

Andrew Wright

GitHub Link: https://github.com/andrewright17/Homework-3-Deep-Learning

## Task

For homework 3, we are tasked with fine-tuning a BERT model for extractive question answering on the Spoken-SQuAD dataset. Our goal is to train a model that can accurately extract an answer to a question from a given context. A testing accuracy above 50% would be considered a success.

## Datasets

The Spoken-SQUAD dataset is similar to the famousd SQuAD dataset, but it adds a layer of complexity to the problem. The Spoken-SQuAD dataset contains the same documents as the SQUAD dataset, except that the document is fed into Google's text-to-speech system to generate an audio version, and then ASR transcription is used to convert the audio back into text. In an ideal world, the documents in both datasets would be the same. However, these systems are not perfect, and thus we get a noisy dataset that contains some errors once the documents reach their final transcripted form.

In this project, I have fine-tuned BERT on both the SQuAD and Spoken-SQuAD datasets. By fine-tuning on SQuAD, I was able to learn the general structure of the dataset and how to fine-tune BERT for extractive question answering. This allowed me to extend the task to the Spoken-SQuAD dataset.

## BERT Model and Training

The BERT model that I used from HuggingFace was 'bert-base-uncased'. On HuggingFace, there are a few models that have already been fine-tuned on SQuAD or Spoken-SQuAD. Almost every one uses 'bert-base-uncased' as its beginning model. This was my rationale for choosing the model I chose.

In order to apply learning rate decay, I used get_scheduler from the transformers library. By applying a linear learning rate, we improve the model's training process allowing it to learn much more at the beginning of training than towards the end. I also added a doc stride to reduce the compute needed. A doc stride essentially splits the document into windows, or smaller sequences, and then each window is scored to find the highest probable section of the document with the answer.

The model was trained on 4 A6000 GPUs using an internal server at MUSC. Total training time was around 80 minutes for fine-tuning on SQuAD and roughly 30 minutes for fine-tuning on Spoken-SQuAD.

# Evaluation

The evaluation of the BERT model was as follows:

- SQuAD dataset training, SQuAD dataset testing, Exact Match: 5.85, F1: 44.94
- Spoken SQuAD dataset training, SQuAD dataset testing, Exact Match: 4.87, F1: 40.90

These metrics indicate to me that there is an error somewhere in the training process or in the validation. I would expect the models to perform with an F1 at least above 70%. At the time of this submission, I am retraining the models to see if the weights of the model were incorrectly stored. I will continue to troubleshoot this in order to improve these metrics.