# Machine Learning Final Project Report

Erik Rhodes, Dylan Sain, Seth Silva

March 2021

## 1 Project Topic

As students, our primary way of showing our knowledge of a certain topic is test taking. Of course, to be good at test taking requires a long time spent studying, as well as time spend improving study methods. We created a machine learning algorithm that will be able to do the same things that students do every day: to search through a series of notes or a textbook and answer a question about the topic. Can a machine learning algorithm beat a human at learning?

For this project, we explored using Transformer models to do natural language processing, in particular, extractive question answering.

## 2 Data

For this project, we primarily used data from two particular sources. The first is the Stanford Question Answering Data set, or SQuAD. This is a collection of English text passages primarily from Wikipedia containing information on a variety of subjects, and an associated set of questions and answers. We used a model that was trained on SQuAD for comparison to models that were trained on a different data set: an online history textbook.

The second component of our data is made up of multiple choice questions from the end of selected chapters of an online world history textbook (see appendix). This is the data that we worked to collect and clean ourselves, and what we used to train and test our selected models.

## 3 Data Cleaning

The first component of our data, SQuAD, is effectively delivered to us in a clean, formatted .json file. Furthermore, the model we were using for comparison was conveniently pre-trained on this data set. Naturally, not much cleaning needed to take place here.

However, the multiple choice questions sourced from the aforementioned online history textbook needed proper formatting. Below you will see an example of the textbook pages containing the questions.

# STANDARDIZED TEST PRACTICE

**TEST-TAKING TIP**

Make sure to read the entire question and each possible answer before deciding on the correct answer.

## Reviewing Vocabulary

*Directions: Choose the word or words that best complete the sentence.*

**1.** Archeologists study _____, or objects made by humans.

**A** rivers
**B** animals
**C** artifacts
**D** oceans

**2.** Donald Johanson and his team found an example of a/an _____ in Ethiopia.

**A** *Homo habilis*
**B** *Australopithecus*
**C** *Homo erectus*
**D** Neanderthal

**3.** The keeping of animals and the growing of food on a regular basis is known as _____

**A** systematic agriculture
**B** domesticated agriculture
**C** Neolithic agriculture
**D** Paleolithic agriculture

**4.** _____ were skilled workers who made products such as weapons and jewelry.

**A** Farmers
**B** Anthropologists
**C** Artisans
**D** Priests

## Reviewing Main Ideas

*Directions: Choose the best answers to the following questions.*

**Section 1** *(pp. 4–11)*

**5.** Which type of scientist uses fossils and artifacts to study early humans?

**A** Chemists
**B** Physicists
**C** Anthropologists
**D** Geologists

**6.** Which hominids do scientists believe were probably the first to leave Africa?

**A** *Homo erectus*
**B** *Australopithecus*
**C** *Homo sapiens sapiens*
**D** Neaderthals

**7.** How did early humans adapt in order to survive?

**A** Following animal migrations
**B** Painting cave art
**C** Living in isolation
**D** Keeping records

**8.** Which invention made hunting easier for early humans?

**A** Cave art
**B** Fire
**C** The spear
**D** Farming

**Need Extra Help?**

| If You Missed Questions . . . | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Go to Page . . . | 4 | 7 | 14 | 16 | 4 | 7 | 7 | 8 |

GO ON

As one can see, it would be difficult to sift through all of the formatted text and graphics to get the information we need. Furthermore, we wanted to eliminate the multiple choice aspect of the questions so our models could be tested solely on their ability to extract the pertinent information (as seems to be the standard in the field of ML question answering). Our first step was to eliminate or slightly change questions that would be ambiguous to the models relative to the contexts provided.

As a little background, SQuAD is separated into "questions" and "contexts" where the former is exactly what it sounds like and the latter is the paragraph of human-readable information that the model is supposed to answer the questions based off of. Our goal was to roughly model our textbook data off of SQuAD and as you can see, at the bottom of the textbook pages it tells us the page number of where the pertinent information for each question lies. So for each of the chosen questions, we went to the page where the information lies and manually located a paragraph or "context" that contains the answer. These "contexts" are what we trained and tested the models off of.

So, an example of a question that we know would be ambiguous to the model is as follows...

**context:** 'For decades, scientists assumed these earliest of upright creatures must also have used tools. In 1974, Donald Johanson challenged this theory when his team found a new skeleton in Ethiopia. Johanson nicknamed the female skeleton "Lucy" and suggested that she was the common ancestor for several types of early human life. Scientists called this type of hominid Australopithecus (aw•STRAY•loh•PIH• thuh•CUS), or "southern ape." It flourished in eastern and southern Africa.'
**question:** Donald Johanson and his team found an example of a/an _____ in Ethiopia.
**answer:** 'Australopithecus'

From a strictly NLP point of view, this question would be somewhat difficult and potentially defective to our model. So we changed the question to get rid of the fill in the blank and include the word "hominid".

# 4 Exploratory Data Analysis (EDA)

As a natural language processing task, our data doesn't lend itself to most common data processing. Below you will see a graph of the most common words in our data. You'll notice that a large portion of these are typically treated as stop words and removed. If we had removed these from our data, we would certainly have much less data to run through, which would likely lead to an increase in our performance. However, in a natural language understanding task like question answering, these stop words are often important to the meaning of a sentence that the model is trying to understand. For that reason, we decided to leave these in.

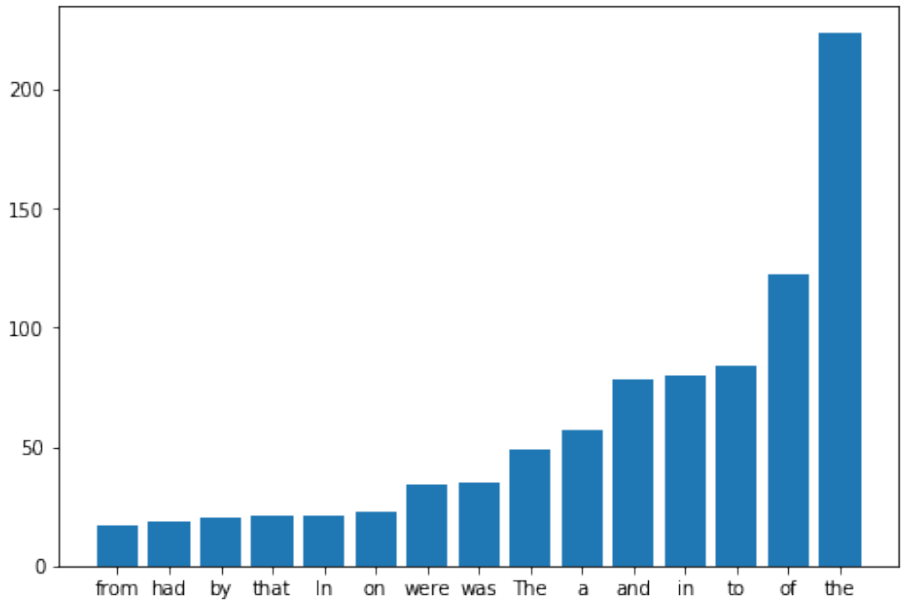



Figure 1: Simple count of words in our data.

# 5 Models

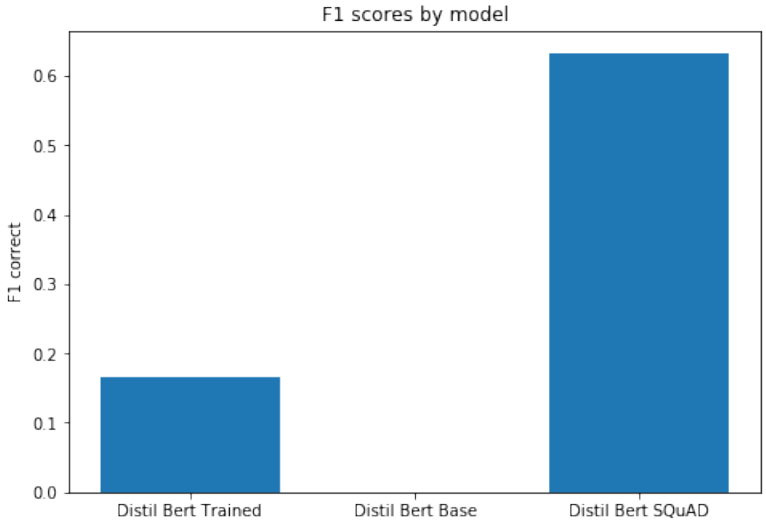
Until recently, most NLP tasks involved the use of RNNs and LSTMs, like the ones we investigated in this course. However, after the introduction of the Transformer model in 2017, it quickly became the industry standard for many NLP applications, including question answering. Most applications of the Transformer model today involve BERT, a language model developed at Google for use in processing search queries. We used the Huggingface implementation of the Transformer, which uses models pretrained on large English datasets fine tuned for specific tasks. In particular, we compared the results from three particular Transformer models. Each of these was a variation of the DistilBERT language model, which is a smaller, faster version of Google's BERT, that is by default trained on a large corpus of unlabeled English text. It is intended to be fine-tuned by training on examples relevant to a particular task, such as masked language modeling, named entity recognition, or question answering. The three models we investigated were:

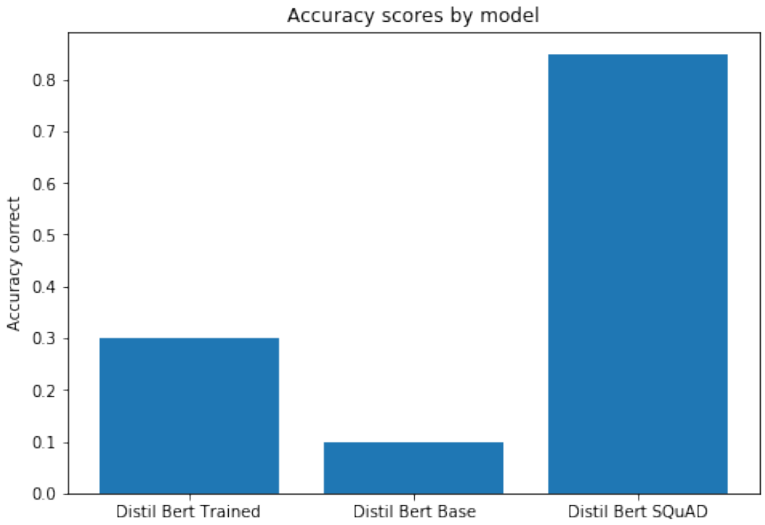
1. The default model for the question answering task, fine tuned on the SQuAD dataset.
2. A base DistilBERT given no further training.
3. DistilBERT trained on the textbook data we made ourselves.

In our explorations with different Transformer-based models, we also investigated BERT and several other derivatives. We decided to focus on DistilBERT in particular as it was a good all-round model for a large variety of tasks that isn't as large or slow as BERT.

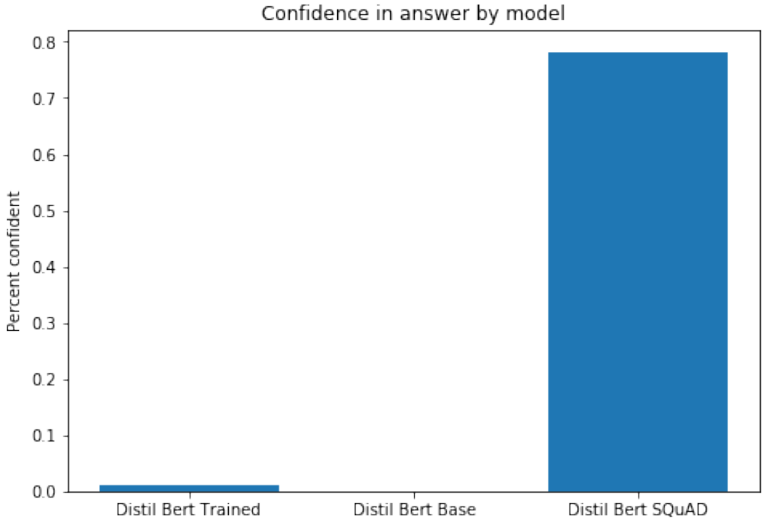
# 6 Results and Analysis

J After an initial run through of the data with all three models we decided to randomly choose training and testing sets for multiple iterations. As some of the questions were more obviously harder than others this ensured that all of the 'harder' questions weren't subject to be exclusively training or testing. This allowed for different models to be tested. In the end we took three metrics from the models: F1 score, accuracy score, and confidence score. The F1 score was the exactly correct answer being given. This is a decent metric, but as some answers were correct but in the wrong form, i.e. 'The Taj Mahal' vs 'Taj Mahal', accuracy score was the better metric. This allowed those particular answers to be counted as correct. The final metric we took was an average confidence score of the answer. Each answer submitted a probability that it was the correct answer.






As you can see above there is a significant difference between the three models, with the SQuAD trained data set preforming much better than the other two models. However it is clear that we were able to improve the base model by fine tuning and training it on our data set. This is a significant increase for both F1 and accuracy scores.




The final metric, confidence, has a huge difference in percentages. The SQuAD data set was reporting a confidence of around 78%. Despite getting almost 80% accuracy score. In contrast getting a healthy 30% correct for how small our data was, the trained distil bert model was averaging 0.945%. It got a lot right but was never confident it was right. The untrained distil bert was even worse brining an average confidence of 0.0525%.

# 7 Discussion and Conclusion

With the difficulties presented in creating out own model to analyze the data in a timely manner we were able to adapt and eventually prove that transfer learning (or fine tuning) for these huge data sets does work. We saw a steady increase of accuracy and confidence from the trained model and the untrained model. However the small data set we were able to collect wasn't able to compete with the 86,000 entries from SQuAD, despite not being trained on our specific data. In the future we would like to have more data and more time to train these base models. We would like to see how efficient transfer learning can be and if we can improve on the SQuAD model by specializing it for a certain data set.

# Appendix

- `https://github.com/chillva/MLTestTaker`

  Our GitHub repository containing our code and data.

- `http://glhssocialstudies.weebly.com/world-history-textbook---pdf-copy.html`

  World History textbook from which we collected our data.

- `https://rajpurkar.github.io/SQuAD-explorer/`

  The Stanford Question Answering Dataset.

- `https://huggingface.co/transformers/model_doc/distilbert.html#overview`

  The documentation for the DistilBERT model we used.