AAI 520 Natural Language Processing

Team 6

Date: 10/8/2023

Team Project Report

Team Members: Eric Barnes, Jon Agustin, Massimiliano Repupilli

**Running the DistillBERT Transformer on the SQuAD Dataset**

In this final project report, we outline the implementation of a question and answering system using the DistillBERT transformer. We utilized the Stanford Question Answering Dataset (SQuAD) v2.0, a substantial dataset with training and validation splits. By loading this dataset using the Hugging Face Datasets library, we ensured access to a diverse set of questions and contexts for training and evaluation purposes.

Question answering comes in many forms. In this work, we examine a particular form of extractive QA that involves answering a question about a passage by highlighting the segment of the passage that answers the question. This involves fine-tuning a model which predicts a start position and an end position in the passage. The goal of Question Answering is to find the answer to a question given a question and an accompanying context. The predicted answer will be either a span of text from the context or an empty string (indicating the question cannot be answered from the context).

**Data Quality of the SQuAD Dataset**

The Stanford Question Answering Dataset (SQuAD) is a popular dataset for training and evaluating question answering (QA) models. It is a collection of over 100,000 question-answer pairs derived from Wikipedia articles. The questions are open-ended and can be answered by any sequence of tokens in the given text.

The quality of SQuAD is generally high. The questions are well-written, cover a wide range of topics and the answers are also accurate and comprehensive. There are however some potential data quality issues with this particular dataset. SQuAD is crowdsourced, so it is possible that some of the questions and answers contain errors. For example, some questions may be ambiguous or have multiple possible answers. Some answers may also be incomplete or inaccurate. Also, because SQuAD is based on Wikipedia articles (which are known to be biased towards certain topics and perspectives) our model may ultimately be biased itself. Finally, SQuAD is relatively small and in some contexts may not contain a diverse enough range of questions and answers. This may limit the ability of question-answering models trained on the dataset to perform well on niche data.

Despite these potential issues, SQuAD is still a valuable resource for training and evaluating QA models. It is one of the largest and most well-known QA datasets available, and it has been used to train some of the most successful QA models (Bard, 2023)

**Key Challenges**

The question answering process consists of five major steps:

1. Preprocessing: Cleaning and formatting the input data.
2. Model Initialization: Loading a pre-trained model or training a new one.
3. Model Training: Feeding the model preprocessed data to learn the relationships between questions and answers.
4. Model Evaluation: Assessing the model's performance.
5. Prediction: Feeding the model a new question to predict the answer.

As is the case with many machine learning workflows, finding an effective method of model training on resource-limited hardware was a challenge in development of the QA system. To address this challenge we utilized the HFTrainer class from the txtai library (NeuML LLC, 2023) and we were able to perform training on the entire dataset in approximately 1 hour.

DistilBERT was selected as our model architecture for its computational efficiency and language understanding capabilities. With 67 million parameters, it is a more lightweight version of BERT, making it well-suited for our resource-constrained environment. Furthermore, DistilBERT's pre-training on extensive English text from BookCorpus and Wikipedia enables it to capture a wide array of language patterns and knowledge, making it an ideal choice for our question-answering task.

**Evaluation of Performance**

Our QA system achieved the following evaluation scores:

|  |  |  |  |
| --- | --- | --- | --- |
|  | DistilBERT | ROBERTA | BERT |
| F1-score |  |  |  |
| Recall |  |  |  |
| Accuracy |  |  |  |
| Precision |  |  |  |
| ~~BLEU~~ |  |  |  |
| ~~EM~~ |  |  |  |

* F1-score: The F1-score is a weighted average of the precision and recall. It is a good measure of the overall performance of a model, as it takes into account both the model's ability to identify the correct answers and its ability to avoid false positives.
* Recall: Recall is the fraction of true positives that are correctly identified by the model. It is a measure of the model's ability to find all of the relevant answers.
* Accuracy: Accuracy is the fraction of all predictions that are correct. It is a good measure of the overall performance of a model, but it can be misleading if the dataset is imbalanced.
* Precision: Precision is the fraction of predicted positives that are actually true positives. It is a measure of the model's ability to avoid false positives (OpenAI, 2023)

In the above example, the DistilBERT model has a high F1-score, recall, and precision. This means that our model was able to both identify the correct answers and avoid false positives. The model also has a high accuracy, but this is likely due to the fact that SQuAD is relatively balanced.

References

Google AI. (2023). Bard language model [Large language model]. Retrieved from <https://bard.google.com/>

OpenAI. (2023). ChatGPT [Large language model]. Retrieved from <https://chat.openai.com/>

Hugging Face. (2023). Transformers: State-of-the-art Natural Language Processing for PyTorch and TensorFlow. Retrieved from <https://huggingface.co/transformers/>

NeuML LLC. (2023). txtai: All-in-one embeddings database. Retrieved from https://neuml.github.io/txtai