CSE576 Project Phase 1- Baseline Testing

Members : Ujun Jeong, Nickolas Dodd, Jordan Miller, Raha Moraffah, John Kevin Cava

Baselines

This phase of the project included having each member of the team run different models from HuggingFace in order to establish a baseline model which could later be fine-tuned. For our experiment, we've used almost the same parameters from the original paper as much as possible. We've conducted three experiments and made three baselines with these parameter settings:the learning rate is 1e-5, batch size is 1, and we also used accumulation steps 25 to help with the low batch size.

Three baselines can be found here: [Colab Link] (This link is available for only ASU accounts)

1) BERT model on the Quail Dataset: This is exactly the same baseline model used in QuAIL paper. We've also trained this model on QuAIL dataset. This baseline test has been completed. The performance of this model was:
- Loss = 1.04721
- Accuracy = 0.56054
- Epoch = 1.99

2) Longformer model on the Quail Dataset: The longformer was chosen because it works well with longer documents and the Quail Dataset is our test set. This baseline test has been completed. The performance of this model was:
- Loss = 0.8903
- Accuracy = 0.6682
- Epoch = 1.99

3) Roberta on the Quail Dataset: In order to process the Quail Dataset with the existing training script, we've tried two methods: changing the format into Swag and training directly, and creating a preprocessor which is committed to the API of the huggingface/datasets. As a result, we found that 2nd method trains faster, and we decided to use this method from now on. This baseline test has been completed. The performance of this model was:
- Loss = 0.8762
- Accuracy = 0.6598
- Epoch = 1.99

It was also meaningful to try the 1st method because we learned that training speed can change a little depending on how data is handled. In particular, the training speed was so slow in the early stages that we looked into the GPU usage, and learned how to properly connect the GPU to process the converted dataset.

Extra experiment

We've also trained Robert model on the Swag Dataset. The Swag baseline was ran on the Swag dataset because part of our synthetically created datasets have a shorter context than the others. The team wanted to have an idea on how well this dataset would perform on its own so we could understand how it affects the overall performance.

We've modified batch size as 8 in this extra experiment as contexts in this dataset(Swag) is so short that we could train it with bigger size of batch.

This baseline test has been completed. The performance of this model was:
- Loss = 1.018
- Accuracy = 0.7937
- Epochs = 3

The model can be found here: [Colab Link]

Reference

- Github(hugging face): [Repository Link]