L06 AWS MLU Lab Reflection Journal

Ryan Yauch

ITAI 2376

Lab Module 2

2/25/2025

Lab 05: Fine-Tuning BERT

Learning Insights

Lab 5 went over Bidirectional Encoder Representations from Transformers, or BERT for short. Although I've used BERT before for assignments, and researched it for projects I haven't gotten the chance to work and mess with the values in a model.

The main new thing I learned is how much computational power BERT takes, running it through the lab occasionally gave me errors and it was still only restricted to 2000 data points. When using BERT on an entire dataset, it may be better to do it in chunks over time rather than all at once.

Challenges and Struggles

As outlined at the start of the lab, BERT kept using up the notebook's memory, or giving an error after attempting to run for a while. To fix this problem I had to reduce the batch sizes to increase performance and decrease the memory usage. But on a second run of the lab, I no longer had the same memory usage issue I had before for reasons I haven't figured out. This was the only main thing I struggled with when working on this lab, but it resolved itself. These issues did start my train of thought on performance which I mention later on.

Personal Growth

I found one of my more surprising parts of the learning experience was when it took a singular example from the dataset to get a closer look at how they were tokenized/encoded. I like seeing both sides of these processes working on smaller and larger sets of data.

I think I also gained an understanding of how powerful systems need to be to run on a large scale, an assignment I recently worked on went over the benefits of a multi-GPU system compared to a base CPU system. I think this can also be applied to an actual job field, as you have to weigh the costs of the system and how powerful it has to be, versus how much money having these models could save.

Critical Reflection

I think the main concept I took from this lab is surprisingly just the new thought process of cost/performance of models. Although I always thought about model performance through how fast they finished and accurate they were, I didn't consider the system usage and potential for these not to function at a higher level. For something like BERT to work properly and efficiently on a large dataset, you'd need a somewhat decent system for it to operate on. I think companies also have to take this into consideration on how much AI is worth implementing, since it would take a large amount of infrastructure to build upon.