ANNETTE BAZAN

ITAI 2372:

DEEP LEARNING

FEBUARY 25, 2025

LAB 06: AWS ML

UNIVERSITY MODULE 2

LAB 05-FINE TUNING BERT

Module 2, Lab 5 of the AWS Machine Learning University introduced me to the process of fine-tuning BERT, also known as Bidirectional Encoder Representations from Transformers. Specifically using the lighter DistilBERT variant, to classify sentiment in Amazon product reviews. The lab offered a hands-on opportunity to work with a pre-trained transformer model, adapt it to a specific task, and evaluate its performance. Through this experience, I gained valuable insights into NLP (natural language processing), encountering some troubleshooting moments and reflecting on my growth during this lab. This journal captures the key takeaways, troubleshooting, and critical perspectives on the process.

The lab provided a structured introduction to BERT and its application in sentiment analysis, deepening my understanding of transformer-based models. One of the most significant insights was the power of transfer learning leveraging pre-trained models like DistilBERT, which has learned general language patterns, and fine-tuning it for a specific task in a smaller dataset. This approach not only saves computational resources but also enables high performance with limited data as evidenced by the lab's use of just 2,000 data points from much larger Amazon review dataset. This lab helped highlight the importance of data preprocessing in NLP. Tokenizing text using the DistilBERT tokenizer and creating a custom Review Dataset class. Highlighting how raw text must be transformed into a format compatible with transformer models, such as input IDs and attention mask. The lab's emphasis on freezing all model weights except the final classification layer was particularly enlightening, as it demonstrated a practical strategy to manage computational complexity while still adapting the model to the task at hand. By the end of the lab, the model achieved a validation accuracy of 0.88 after the 20 epochs reinforcing the effectiveness of this approach.

Despite the lab's clear guidance, I encountered several challenges that tested my technical skills and patience. One primary struggle was managing computational resources. BERT models are notoriously resource-intensive, and even with DistilBERT, I faced memory constraints on my instance. When this occurred, I followed the lab's advice to restart the kernel and reduce the batch size, which resolved the issue but highlighted the importance of hardware considerations in machine learning workflows. This lab prompted me to think more about scalability and optimization in real-world applications. Another challenge was interpreting the training and validation loop outputs. While the code executed successfully, understanding the fluctuations in validation accuracy. Peaking at 0.89 in epoch 18 before dropping to 0.88 in epoch 20, requiring careful analysis. Initially, assuming more epochs would improve the performance, but the results suggested potential for overfitting or data variability. Debugging required revisiting the lab's instructions to see where my code was failing me, but I did use Gemini to help after exhausting what I could understand and reflect on.

This lab helps my progress in developing a stronger, confident in myself as a machine learning practitioner. Handling a model like BERT and complex NLP tasks that deepened my appreciation for the balance between theory and hands-on. I grew more comfortable with PyTorch and the Hugging face transformers library tools which I have only stuck my toe in so to speak. Writing and modifying code such as adjusting the number of epochs from 10 to 20 and observing the impact on validation loss. I did struggle finding this part of the code to change I will admit. Beyond technical skills I developed greater resilience in troubleshooting. When faced with memory errors or unexpected time run errors in my outputs, I learned to approach problems more methodically. Using the lab's instructions and my own critical reasoning. This shift helped me work through my frustration to problem-solving reflects a maturing mindset that I believe will serve me well as we progress through more modules.

Looking back through lab 05 was a well-structured hands-on exercise that was prompting me to question certain aspects of the process. The decision to use only 2,000 data points, while practical for training time, left me wondering about the model's generalizability. How could or would the performance change with the full 56,000 entry data set? It was hinted at, but resource constraints prevented me from testing it fully. This is where the limitation highlighted the tension in machine learning, the trade-off between simplicity and real-world complexity. Additionally, freezing all but the classifier layer while it was efficient it seemed like a simplification of BERT's potential. Fine turning the entire model might yield better accuracy at a higher computational cost. The trade-off raises questions about when such compromises are justified in practice but perhaps in resource limited settings, less so in industry environments with access to powerful GPUs. The dataset itself has a binary "isPositive" label oversimplifies sentiment while ignoring nuances like neutrality. Any future iterations could possibly benefit from a more granular classification scheme to capture real-world complexity maybe.

Lab 5 was another rich hands-on learning experience that helped expand my technical knowledge and critical thinking skills. I gained a practical understanding of fine-tuning BERT, navigating resource-related challenges that helped grow my confidence in this course and my ability to adapt to specific tasks that I may not know much about. The lab inspired me to consider how I can use this lab, to approach similar problems in a new light. I wonder if larger datasets, unfrozen layers, or more sophisticated evaluation metrics. This reflection has motivated me to continue exploring NLP and transformer models, with an eye toward applying what I have retained and learned between this lab and real-world applications.