Chaklam Silpasuwanchai, Todsavad Tangtortan Updated 18 Mar 2025
AT82.05 Artificial Intelligence: Natural Language Understanding (NLU)

A7: Training Distillation vs LoRA

In this assignment, we will explore the comparison between Odd Layer and Even Layer Student Training Models and LoRA (Low-Rank Adaptation) on a distillation task using BERT from Huggingface.

Note: You are ENCOURAGED to work with your friends, but DISCOURAGED to blindly copy other's work. Both parties will be given 0.

Note: Comments should be provided sufficiently so we know you understand. Failure to do so can raise suspicion of possible copying/plagiarism.

Note: You will be graded upon (1) documentation, (2) experiment, (3) implementation.

Note: This is a one-week assignment, but start early.

Deliverables: The GitHub link containing the Jupyter notebook, a README.md of the GitHub repository, and the folder of your web application called 'app'.

Task 1. Hate Speech/Toxic Comment Dataset - Find and load a dataset that includes toxic comments or hate speech. This dataset will be used for training and evaluating the models. (1 point)

Task 2. Odd Layer vs Even Layer Training - Based on the case-studies/distilBERT.ipynb, modify as follows:

1) Train the student model using the odd layers {1, 3, 5, 7, 9, 11} from the 12-layer teacher to the 6-layer student. (1 point)
2) Train the student model using the even layers {2, 4, 6, 8, 10, 12} from the 12-layer teacher to the 6-layer student. (1 point)

Task 3. LoRA (Low-Rank Adaptation) - Implement LoRA to train the 12-layer student model. (1 point)

Task 4. Evaluation and Analysis

1) Evaluate the models on the test set, and analyze the performance of the models trained with Odd Layers, Even Layers, and LoRA. Discuss the differences in performance across the three methods. (0.5 point)
2) Discuss the challenges encountered during the implementation, specifically comparing distillation fine-tuning models (Odd and Even Layer) with LoRA fine-tuning. Propose improvements or modifications to address the challenges. (0.5 point)

Model Type	Training Loss	Test Set Performance
Odd Layer		
Even Layer		
LoRA		

Task 5. Web Application - Develop a simple web application that classifies whether a given text input is toxic or hate speech. (1 point) The web application should:

1) Include an input box where users can type in a text prompt.
2) Based on the input, the model should classify and display whether the text is toxic or not. For example, if the input is "I hate you", the model might classify it as toxic.

Good luck :-)