Generative AI Technical and Social: Homework 2

September 18, 2025

In this exercise, we will explore transformer models and text generation. Please review the code provided in the Colab notebook and solve the following problems.

Problem 1

1. **[1 point]** How many trainable parameters does the GPT transformer model in Colab have?
2. **[2 points]** If we modify the GPT transformer model in Colab to have 96 layers, an embedding dimension of 12,288, 96 attention heads, and a maximum sequence length of 2048, how many trainable parameters would the model have? (This configuration is in fact the GPT-3!)
3. **[Bonus: 2 points]** For training the current model with 32-bit precision and a batch size of 1, how much GPU memory is required roughly?

Problem 2

1. **[2 points]** Train the model from scratch by removing the weight loading from GPT-2 and report the training accuracy and loss curve.
2. **[1 point]** Test the model on some examples. Are you surprised by the generalization capabilities of models like ChatGPT, which shares the same architecture but on a larger scale?

Problem 3

1. **[2 points]** Experiment with the ChatGPT playground using different temperature values, Top-p, and max-token values. Explain their meanings.
2. **[Bonus: 2 points]** Add comments for each line to this notebook, an implementation of MultiHeadAttention from scratch. Then answer the question: how many tunable parameters are in this model?

Problem 4 [5 points]

In class we have discussed benchmark-driven research. Use OpenAI Deep Research to select a benchmark in a field that interests you. Read the paper and write a one-page report on this benchmark. Your report should include

- A description of what the benchmark is and what it measures.
- The values, goals, or merits that this benchmark promotes.
- The criticisms, limitations, or trade-offs associated with this benchmark. How could the benchmark be improved, or where might it fall short in your own research.

Examples of these benchmarks include: AccountingBench, FinanceBench, τ-bench, VisualWebArena, OSWorld, Bright (Reasoning benchmark), Last Human Exam, ToolBench / Gorilla, AlfWorld, AgentBench, etc.