CIS 620, Advanced Topics in Deep Learning, Spring 2024

Homework 1
Due: Monday, January 29, 11:59pm
Submit to Gradescope

Learning Objectives

After completing this assignment, you will:

- Use BERT model to inference on a long input sequence
- Experiment with chunking method and see its limitations
- Learn about Longformer and how it address context length problems

Deliverables

This is an **individual** assignment for both the written and coding portions.

1. **A PDF with your name in the agreement**

   Copy and edit the google doc to enter your name in the Student Agreement and answer the questions mentioned below.
2. **hw1.ipynb file with the functions implemented**

   Complete the coding assignment in the Jupyter Notebook and upload the file to Gradescope.

Note that there is a separate assignment for the papers, which will be handed in separately.

Homework Submission Instructions

Written Homeworks

All written homework must be submitted as a PDF to Gradescope. **Handwritten assignments (scanned or otherwise) will not be accepted.**

Coding Homeworks

All coding assignments will be done in Jupyter Notebooks. We will provide a .ipynb template for each assignment as well as function stubs for you to implement. You are free to use your own installation of Jupyter for the assignments, or Google Colab, which provides a Jupyter Environment connected to Google Drive along with a hosted runtime containing a CPU and GPU.

Questions

1. **Long Context Sequence**
    a. Complete the **Context Chunking** section of the notebook by filling out the code. Experiment with different hyperparameters and answer the question at the end of the section in PDF.
        i. Explain Chunking in a few sentences.
        ii. Why does chunking fail in the later example? Can you give another example on which chunking also fails?
    b. Answer the questions about **LongFormer** and complete the second section. Report your results and compare with results from chunking.
        i. Read the [Longformer paper] (https://arxiv.org/abs/2004.05150).
            1. What is the Sliding Window Attention and Global Attention for Longformer?
            2. How do they capture information in a long context?
        ii. **(Optional)** Read about the ["Lost in the Middle"] (https://arxiv.org/abs/2307.03172) phenomena of LLMs.
            1. What is the phenomena describing? What might be a reason for it to happen?
            2. Based on your understanding of Longformer, will it also suffer the same problem?

ANSWERS:-

1.a.i) Chunking involves breaking down large texts into smaller and more manageable segments. This is particularly useful when a text is so large that it exceeds the token limit of a model. Each chunk is processed independently and the results from these chunks are then aggregated and analyzed to form a comprehensive understanding or answer to a query.

1.a.ii) When a large text is chunked, each chunk is processed independently. This can lead to loss of context if the answer to a question lies at the boundary of chunks or requires information from multiple chunks. Chunking can introduce artificial discontinuity in the text. If a sentence or idea is split across multiple chunks, the model may not be able to understand the full context, leading to incorrect or incomplete answers. In this case, the example has a higher chance to fail because there are incorrect locations in its surrounding key words and Philadelphia is located in the chunk where University of Pennsylvania is present. There are many more incorrect locations which are likely to be associated with the Penn Museum.

Another example where chunking fails would be to ask a question like “What caused X civilization to fall?” based on a detailed historical text, chunking might fail to provide a comprehensive answer as the reasons could be scattered throughout the text.

1.b.i.1. Sliding Window is an attention mechanism where the attention pattern provides each token a fixed-size attention window surrounding it. It scales linearly with the input sequence length, making it efficient for long sequences. This windowed attention allows each token to attend to a set of neighboring tokens, ensuring local context is captured and processed effectively.

Global attention is a mechanism that allows certain designated tokens to attend to all tokens in the sequence and vice versa. This is useful for tasks where specific tokens need a global view of the entire input. Global Attention tokens are pre-selected based on the task and provides a way to incorporate task-specific information into the model.

1.b.i.2. When we employ the sliding window attention mechanism along with the global attention mechanism, we are effectively trying to combat the quadratic nature of self attention mechanisms present in transformer models. The sliding window attention is used to capture local context where every token is given a window to accommodate the tokens surrounding it.

In order to compensate for tokens that require a broader view of the input like essential entities, question tokens or classification tokens, we employ the global attention mechanism as described in the previous answer. This mechanism allows the model to maintain an understanding of the entire document, enabling it to handle tasks that require a global understanding of the text.

For both the examples given in the colab notebook, the answer was Philadelphia, Pennsylvania for both chunking and the longformer.

I tried other questions like “Who is the university’s founder?” and “Who was the university’s first professor?” and it gave the correct answers to both the questions.

The example I created that failed the chunking model was “Where is Memphis?” and the answer I got was “Philadelphia.”
This example failed the longformer model as well and the answer I got was “Philadelphia, Pennsylvania.”
The correct answer is Egypt.

When I use the question “Which country is Memphis located in?”, the chunking model gave the correct answer- “Egypt” but the longformer model failed.