Reading Assignment 7: Longformers

Q1. Why does the memory complexity of a transformer expand quadratically when the input sequence only expands linearly? How does this limit our ability to build larger and larger models?

In a standard Transformer, eachtoken in a sequence must attend to every other token through the self-attention mechanism. This means that for a sequence of n tokens, the model computes an $n \times n$attention matrix, producing both computational and memory costs that scale as $O(n^2)$. Even though the number of tokens grows linearly, the number of pairwise interactions grows quadratically, leading to exponential memory demand as the input length increases. This quickly becomes infeasible on modern GPUs, which have limited memory bandwidth, forcing practitioners to truncate or split long documents into smaller chunks. As a result, long-range dependencies and global context are often lost, constraining both model size and the kinds of problems Transformers can handle efficiently.

Q2. How does the Longformer try to improve on this complexity?

The Longformer replaces the full quadratic self-attention with a sparse, hybrid attention pattern that scales linearly with sequence length. It introduces a sliding window attention, where each token only attends to a fixed number of nearby tokens, and global attention, where a few special tokens (like [CLS] or question words) can attend to all tokens and vice versa. This design preserves local context for most tokens while still allowing a small number of tokens to aggregate global information, achieving $O(n)$complexity. By doing so, Longformer can efficiently process sequences of thousands of tokens without partitioning or truncation. The result is a model that captures both fine-grained local patterns and document-level understanding while remaining computationally feasible for long-context NLP tasks. More recent architectures, like Transformer-XL, RETRO, and Gemini 1.5 Pro, extend context through recurrent caching or retrieval-augmented memory, which allows them to “remember” past segments without recomputing attention over the entire sequence.