Integration of Retrieval-Augmented Generation in StreamingLLM for Enhanced Recall

Andrew Jenkins, Brian Bailey
{awj@mit.edu, bbailey@mit.edu}

Summary

The evolution of language processing technologies in the Han Lab at MIT has introduced StreamingLLM, a sophisticated system designed to handle continuous data streams. This model represents a significant advancement from traditional fixed-context window approaches, primarily through the introduction of attention sinks—specific tokens within the model's memory that anchor the attention mechanism and provide a semblance of stability over long sequences. Despite this innovation, StreamingLLM is still bound by a finite attention window, leading to an inevitable loss of context as dialogues extend beyond this window, and earlier tokens are evicted to make room for new ones.

Our study aims to expand the capabilities of StreamingLLM by incorporating a Retrieval-Augmented Generation (RAG) mechanism, which enables the model to access a broader context beyond its immediate attention window. We have engineered a RAG-enhanced StreamingLLM that exhibits a marked improvement in maintaining narrative coherence and accuracy, showcasing its potential to effectively manage prolonged and detailed interactions. Additionally, we have curated a task-specific dataset for testing purposes.

Additional Details

StreamingLLM utilizes attention sinks to ensure a degree of reliability that surpasses conventional sliding-window language models. These attention sinks are a set of initial tokens that the model consistently references, which helps to stabilize the attention distribution throughout the dialogue. However, the fixed size of the attention window, which we have adapted to 250 tokens for our study, remains a bottleneck when it comes to processing lengthy text where context outside of this window is discarded.

In our pursuit to mitigate the context retention challenges of StreamingLLM, our initial approach harnessed the KV values directly within the model. This method aimed to directly incorporate the distilled essence of the most relevant context into the current attention window. However, this direct utilization of KV values for RAG yielded unsatisfactory results.

Confronted with these limitations, we pivoted to a more conventional RAG implementation. We adopted a sentence transformer to embed entire sentences within the narrative, including those previously evicted as newer text streamed in. Through a cosine similarity comparison of these sentence embeddings with the query embeddings, we devised a mechanism that could select and reintegrate the most contextually significant segments. This refined RAG approach yielded substantially better outcomes, evidenced by an average accuracy increase of 25.4% across the board and a remarkable 76% improvement for queries addressing the beginning of the stories. Furthermore, we altered the questions in the testing data to be more similar to the sentences in which the true answers lie, so as to make the RAG mechanism more accurate. On this data, the RAG integrated model (k=2) showed performance increases of 38% on average and 76% on questions about the beginning of stories. The switch to embedding sentences rather than relying on KV vectors enabled the model to effectively reconstruct lost context, thereby addressing StreamingLLM's limited retrospective capacity.

Our dataset, derived from TinyStories, was carefully curated to challenge the model's recall capabilities. We created questions that targeted unique details within the stories—details strategically chosen or inserted to evaluate the model's ability to recall and use specific information. This curation ensured a robust assessment of the model's enhanced recall abilities, reflecting the RAG mechanism's success in bridging the context gap inherent in the base StreamingLLM.

The integration of the RAG mechanism into StreamingLLM signifies a leap forward in natural language processing, particularly for applications requiring deep and sustained interactions. Our findings underscore the enhanced capacity of language models to engage in more meaningful dialogues by retaining a comprehensive understanding of the context, thereby paving the way for AI systems to participate in complex conversations with human-like continuity.

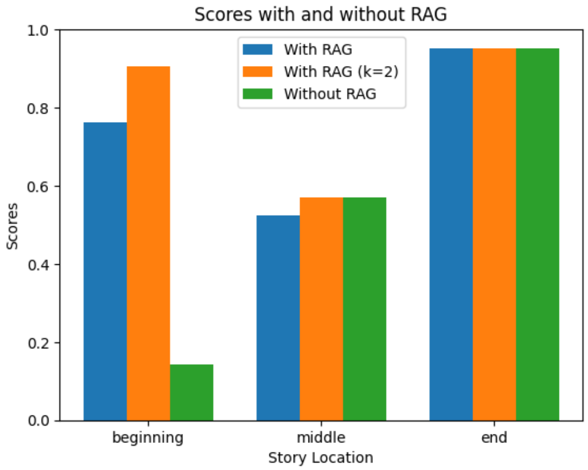

Figure 1: This figure shows the difference in accuracy between the Top 2 RAG-integrated model, Top 1 RAG-integrated model, and the base StreamingLLM.