Preliminary Research Idea (5): HDP-Attention: Mimicking Human Hierarchical Memory for Efficient Long-Context Large Language Models

Professor Richard Yi Da Xu
HKBU and TadReamk Limited

To address the computational inefficiency and the "Lost in the Middle" phenomenon faced by large language models when processing ultra-long context windows, this proposal introduces HDP-Attention (Hierarchical Dirichlet Process Attention), inspired by human hierarchical memory. Current models typically treat memory as a flat buffer, applying equal attention to every historical token, which leads to quadratically growing computational costs. In contrast, humans maintain detailed working memory for immediate conversation, but compress past events into semantic summaries, only recalling specific details when triggered. By applying Bayesian nonparametric techniques—specifically, the Hierarchical Dirichlet Process (HDP)—this method aims to mimic this efficiency by dynamically organizing context into a structured tree-like hierarchy rather than a linear sequence.

The technical mechanism utilizes the "Chinese Restaurant Process" to automatically cluster tokens into topics without requiring a preset limit on the number of topics. For recent or currently active topics, the model computes attention at the individual token level to ensure high fidelity. However, as topics become inactive, they are compressed into single "Topic Vectors" (parent nodes), significantly reducing memory footprint. If a new query matches a compressed topic vector, the system "expands" that node to restore access to specific tokens within it. This approach effectively achieves infinite context length by maintaining low-resolution summaries of the past while preserving high-resolution access to the current context.