Milestone Deliverable Review Report

Deep Funding Round: 3

Project code: DFR3-RFP4

Project title: Memory-augmented LLMs: Retrieving Information using a Gated Encoder LLM (RIGEL)

Milestone number: 4

Milestone deliverable: The current version of the code to produce the compressed hierarchy can be found here: https://github.com/mlabs-haskell/rigel/blob/staging/generate_hier_cv_db.py

The PR that constituted the most important contributions to this milestone can be found here: mlabs-haskell/rigel#10

The output produced by this script can be found here: https://drive.google.com/drive/folders/1TlTZmaezqUWiTNUWvQOfw4j9PpqGhbqH?usp=sharing

Date: 11/11/2024

Status: Accepted

Feedback (Why accepted, why rejected?):
on this challenging milestone #4 of the RIGEL project, the MLabs Team focused on optimizing context vector retrieval by compressing context vectors into a hierarchical structure. Starting from an uncompressed size of 4096, they reduced vectors incrementally through levels of 1024, 256, 64, and finally 16, allowing for efficient storage and retrieval.

Initially, the team considered clustering algorithms and locality-sensitive hashing but found better results using a custom hierarchical search algorithm. The search begins with the most compressed vector representation, progressively moving up levels of resolution, narrowing down to vectors with the highest similarity. The final selection is made from the uncompressed level, providing the most accurate match.

Despite concerns over processing load, the team discovered that compressed vectors, paired with GPU parallelism, mitigated memory constraints and enhanced processing efficiency. The accompanying code outlines the process for generating, storing, and verifying these hierarchical context vectors. This approach facilitates rapid retrieval, supporting the

RIGEL model's goal of efficient and scalable memory augmentation. Future updates may focus on refining this hierarchical system for further speed and accuracy.

Then we can find the PR that constituted the most important contributions to this milestone, the project enhances Llama2 to extract context vectors and store them efficiently with cv_storage, leveraging indexed_binary_db for quick retrieval. Compressed layers, like mipmaps, speed up searches, guided by generated prompts and hierarchical structures.

Finally the autoput produced by this script can be found in a drive folder with the total of index , .db and jason files, 11 in total.

The repository shows constant activity by two team members responsible for the development.

Incredible work carried out by MLabs Team.

If rejected, suggested changes: