Yibo Peng

☎ 1-412-508-0360 ✉ yibop@andrew.cmu.edu linkedin.com/in/yibo-peng-cs github.com/pppyb

Education

Carnegie Mellon University **Pittsburgh, PA**
Master of Science in Artificial Intelligence Engineering GPA:3.74/4.0 *(Expected) Aug 2023 - Dec 2024*

Beijing Jiaotong University & Lancaster University **Beijing, CN & Lancaster, UK**
*Bachelor of Science in Computer Science (**Honours**) Rank: 10/98* *Aug 2018 - July 2022*

Research Experience

Language Technologies Institute, Carnegie Mellon University **Pittsburgh, PA**
Evaluating LCMs vs. RAG for Code Generation, Advisor: Graham Neubig & Daniel Fried *Aug 2024 – Present*

- Implemented an experimental framework to compare long context models with **Retrieval-Augmented Generation (RAG)** for code generation, using Unlimiformer to **extend model's input context lengths**.
- Conducted experiments adjusting context lengths to explore their impact on code generation quality, identifying **optimal ranges** and revealing **trade-offs between context length and noise**.
- Demonstrated that RAG maintained **superior performance** over long context models, highlighting its effectiveness in organizing and utilizing large-scale information even when extensive context is available.

Language Technologies Institute, Carnegie Mellon University **Pittsburgh, PA**
Unlimiformer: Long-Range Transformers with Unlimited Length Input *June 2024 – Aug 2024*

- Reproduced **Unlimiformer** to extend input length in code generation tasks, using **k-nearest neighbors (kNN)** retrieval to handle long-distance inputs without increasing computational complexity.
- Introduced **Repocoder RAG method** for retrieval and tailored to the code snippets of code generation.
- Improved model performance with Unlimiformer, achieving **significant gains** in **EM (Exact Match) and ES (Evaluation Score)** when handling input sequences beyond the original context length limit.
- Integrated Unlimiformer into various **encoder-decoder models**, including **LLaMA** series models.

Language Technologies Institute, Carnegie Mellon University **Pittsburgh, PA**
RepoCoder: Repository-Level Code Completion Through Iterative Retrieval and Generation *April 2024 – June 2024*

- Set up an experimental environment for testing, including line, API, and function level **code completion tasks**.
- Enhanced **retrieval strategy** to analyze code generation performance on **state-of-the-art LLM**.
- Optimized the retrieval-generation pipeline through **adding the prompt length** which improved retrieval quality and increased the EM (Exact Match) score by over **5 %** compared to the original **baseline**.

ECE Department, Carnegie Mellon University **Pittsburgh, PA**
Speculative Decoding with LLM *May 2024 – Jul 2024*

- Implemented the Speculative Decoding algorithm, improving inference speed for large Transformer models through parallel computation, achieving **2-3x** acceleration in practical tests.
- Designed **KV Cache optimization** to reduce memory bandwidth bottlenecks and enhance inference efficiency.
- Applied the acceleration technique to code generation tasks, conducting experimental validation using Salesforce Codegen model series (ranging from 350M to 6B parameters).

Work Experience

PricewaterhouseCoopers LLP (PwC) **Beijing, CN**
Development Engineer Intern – Quantitative Model Expert Team *Nov 2021 – Apr 2022*

- Developed an large VBA application to assess and calculate Expected Credit Loss(ECL) of accounts receivable.
- Reduced calculation time from **15 minutes to 10 seconds** by transitioning calculations to the database.
- Improved code efficiency by simplifying loops, reducing global variable usage, and optimizing function calls.
- Collaborated with cross-functional teams to integrate the model and over **230 listed companies** used it.

Publication & Patent

- B. Hu, ”Intelligent Home Standards and Technologies.” Human-Computer Interaction Application & Entertainment Type Equipment. Ed. **Y. Peng**, Beijing: Tsinghua University Press, 2022. 144-155 & 212-247.
- Y. Tian, Z. Li, **Y. Peng**, 2021. Automatic control system and network of circulating water degassing devices. CN Patent Application 202120168813.7, filed January 2021.