REST: RETRIEVAL-BASED SPECULATIVE DECODING

Đỗ Minh Khôi - 240101051

Tóm tắt

- Lớp: CS2205.FEB2025
- Link Github của nhóm:
- Link YouTube video:

Đỗ Minh Khôi - 240101051

- Tổng số slides không vượt quá 10

Giới thiệu

- Các LLMs (GPT, Llama, Vicuna, ...) có khả năng suy luận rất mạnh, nhưng để chạy lại rất tốn tài nguyên.
- Phương pháp thông thường khi chạy LLMs:

  Sinh từng token -> Push forward LLMs -> Choose probability -> Choose token -> Loop

- Phương pháp Speculative Decoding:

  Dùng các LLMs nhỏ hơn nhiều lần dự đoán trước k token -> Verify sử dụng LLMs giống phương pháp trước -> Loop

Giới thiệu

- Phương pháp Speculative Decoding tuy đã tăng tốc được cho LLMs nhưng vẫn cần phải train lại các LLMs nhỏ và kiểm định chất lượng tạo sinh của các LLMs này
- REST đã ra đời phương pháp mới để tăng tốc LLMs vẫn sử dụng cơ chế Speculative Decoding nhưng không sử dụng các LLMs nhỏ làm draft model thay vào đó sẽ xây dựng một cơ sở dữ liệu (datastore) để xây dựng các draft token và verify lại các token này sử dụng LLMs

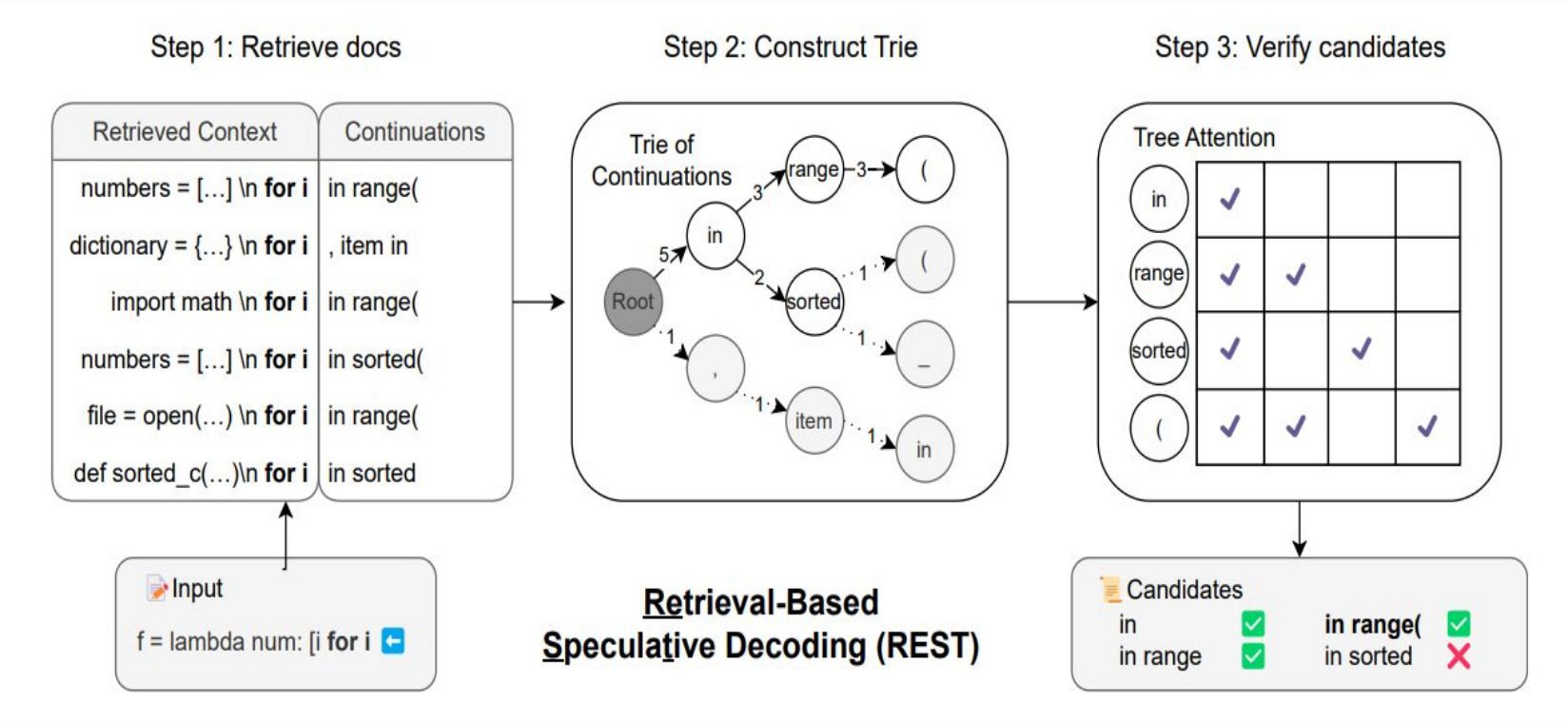
Mục tiêu

- Tăng tốc các LLMs (GPT, Llama, Vicuna, …) giảm thiểu số lần push forward qua các LLMs này khi sinh văn bản
- Xây dựng được một datastore đủ lớn, đủ độ phủ để tăng độ chính xác cũng như tốc độ khi sinh văn bản của các LLMs

Nội dung và Phương pháp

- REST gồm 3 bước chính, được thực hiện tuần tự tại mỗi bước sinh token:
  - Truy xuất (Retrieval)
  - Xây dựng cây Trie (Draft Construction)
  - Xác minh bằng Tree Attention (Verification)

Nội dung và Phương pháp



Kết quả dự kiến

- Bộ dữ liệu đánh giá:
  - HumanEval (sinh mã Python)
  - MT-Bench (hội thoại nhiều lượt)
- Mô hình thử nghiệm:
  - CodeLlama 7B/13B
  - Vicuna 7B/13B
- Tăng tốc đạt được:
  - REST đạt tốc độ nhanh hơn so với phương pháp speculative decoding truyền thống
  - Datastore càng lớn chất lượng truy vấn và tốc độ càng cao

Kết quả dự kiến

Benchmark	Model	Method	*Mean Token Time*(↓)	Speedup(↑)
HumanEval (1 shot)	CodeLlama 7B	Autoregressive (Greedy)	27.89 ms/token	1×
	CodeLlama 7B	Speculative (Greedy)	15.90 ms/token	1.75×
	CodeLlama 7B	REST (Greedy)	11.82 ms/token	2.36×
	CodeLlama 13B	Autoregressive (Greedy)	44.32 ms/token	1×
	CodeLlama 13B	Speculative (Greedy)	19.39 ms/token	2.29×
	CodeLlama 13B	REST (Greedy)	19.53 ms/token	2.27×
HumanEval (10 shot)	CodeLlama 7B	Autoregressive (Nucleus)	27.99 ms/token	1×
	CodeLlama 7B	Speculative (Nucleus)	18.83 ms/token	1.49×
	CodeLlama 7B	REST (Nucleus)	13.18 ms/token	2.12×
	CodeLlama 13B	Autoregressive (Nucleus)	44.46 ms/token	1×
	CodeLlama 13B	Speculative (Nucleus)	22.68 ms/token	1.96×
	CodeLlama 13B	REST (Nucleus)	20.47 ms/token	2.17×
MT-Bench	Vicuna 7B	Autoregressive (Greedy)	25.48 ms/token	1×
	Vicuna 7B	Speculative (Greedy)	19.44 ms/token	1.31×
	Vicuna 7B	REST (Greedy)	15.12 ms/token	1.69×
	Vicuna 13B	Autoregressive (Greedy)	44.30 ms/token	1×
	Vicuna 13B	Speculative (Greedy)	29.80 ms/token	1.49×
	Vicuna 13B	REST (Greedy)	25.08 ms/token	1.77×
MT-Bench	Vicuna 7B	Autoregressive (Nucleus)	25.93 ms/token	1×
	Vicuna 7B	Speculative(Nucleus)	20.65 ms/token	1.26×
	Vicuna 7B	REST(Nucleus)	16.02 ms/token	1.62×
	Vicuna 13B	Autoregressive (Nucleus)	44.32 ms/token	1×
	Vicuna 13B	Speculative (Nucleus)	31.78 ms/token	1.39×
	Vicuna 13B	REST (Nucleus)	25.92 ms/token	1.71×

Tài liệu tham khảo

- REST: Retrieval-Based Speculative Decoding
- Language models are few-shot learners. In Advances in Neural Information Processing Systems (NeurIPS).
- Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality
- Fast Inference from Transformers via Speculative Decoding
- LLM-Pruner: On the Structural Pruning of Large Language Models
- Geralization through memorization: Nearest neighbor language model