Quantifying Latency-Accuracy of LLMs

Zhuohao Li, Ying Li

Motivation and Objectives

Objective: How fast (latency) and how well (accuracy) LLMs that understands and generates language works.

Importance: If we get this right, people will have a better, smoother experience and the LLMs will be more useful in various deployment settings.

Goals and Deliverables:

measure and improve the balance between speed (how fast you get an answer) and quality (how good or accurate that answer is).

A clear report on how our LLMs performs now

Technical Approach and Novelty

Current Practice:

Current Evaluation mainly focus on how accurate LLMs is, with less emphasis on how fast it responds, also less care about cost-efficiency esp on computation-intensive devices.

Our Approach:

Measure both speed and accuracy together on different platforms

Novelty:

The integrated evaluation framework that equally weighs speed and accuracy, providing a more holistic view of LLM performance.

Methods

Algorithms & models:

Deploy LLMs on different platforms, and use a combination of LLMs to assess both response time and answer quality.

Llama-2 7B/13B, LoRA, GPT-2, etc.

Datasets & Task definition:

Document summarization, code generation (debugging, continuous writing), Emotional analysis, Machine translation, logic calculation

Deployment env:

1) Server: 3 Nvidia A6000s
2) Edge: Mac M2
3) Cloud: Hugging face

Evaluation and Metrics

Latency Measurement:

We assess latency by timing the response duration from query input to model output

Accuracy Measurement:

Correctness checks

Balanced Scored:

Equal weight to speed and quality, a holistic view of model performance

Current Status and Next Steps

- Current:
  - Deployed different LLMs (Llama-2 7B, Llama-2 13B) on different platforms(Nvidia A6000, hugging face, Mac M2)
  - Successfully tested and executed different categories of tasks including document summarization, code generation, emotional analysis in practice
- Next:
  - Deploy more LLMs (fine-tuned, open-sourced: * some of them might not be available on HF)
  - Design the benchmark carefully
  - Insert measuring part in our test
  - Report / Repository