Mixture-of-Experts - KL-Divergence Threshold

Composing Small LLMS for Efficient Inference with a Reference Foundation Model

Alex Towell
lex@metafunctor.com

2023-09-30

Abstract

We investigate using KL divergence to dynamically switch between a large and small LLM based on ongoing quality assessments.

Contents

0.1 Introduction

This paper introduces a method to dynamically manage computational demands when utilizing Large Language Models (LLMs). The approach employs Kullback-Leibler (KL) divergence to facilitate the switching between a computationally intensive large LLM and a smaller, more efficient counterpart. This strategy is designed for scenarios where rapid and cost-effective content generation is as important as maintaining a reasonable quality of outputs, especially in real-time and multimodal AI interactions. It presents a potential path towards enhancing the practicality and scalability of LLM applications.

Addressing the growing need for computational efficiency, this paper proposes a method mindful of the current GPU scarcity and the high expense associated with LLM operations. Efficiency not only conserves resources but expands the potential for creative and analytical tasks within LLMs. This approach also promises advancements in real-time, multimodal AI interactions, suggesting a new paradigm for AI assistants to engage more naturally with users.

0.2 Literature Review

This section reviews the current state of research on computational efficiency in LLMs, the impact of hardware limitations, and the potential of dynamic computation strategies.

0.2.1 GPU Scarcity and Expense

- Briefly discuss the high demand and low supply of GPUs.
- Mention how this affects the accessibility of LLMs.

0.2.2 Efficiency in Model Exploration

- Cover the importance of computational efficiency for exploring model spaces.
- Introduce the concept of “tree of thoughts” and how efficiency can expand these explorations.

0.2.3 Sampling Procedures and Output Quality

- Explore how different sampling methods affect LLM outputs.
- Discuss exploitation versus exploration tradeoffs and how efficiency can facilitate more exploration.
- Related to this, show that one way to explore is by incresing the temperature of the sampling procedure.
  - However, since it flattens the distribution, a smaller model with a similar tempoerature will have a more similiar output to the larger model than a smaller model with a lower temperature.
  - Hypothesis: smaller models are therefore particularly relevant for exploration. See: tree-of-thoughts.

0.2.4 Multimodal AI and Real-Time Interactions

- Review the emerging trend of multimodal AI applications.
- Delve into the challenges of real-time AI interactions and how computational efficiency can improve user experiences.

0.2.5 Hybrid Local-Cloud Systems

- Discuss the potential of combining local and cloud computing for LLMs.
- Highlight how this hybrid approach can address both computational and cost-efficiency.

0.2.6 Embedding Consistency in LLMs

- Discuss the significance of consistent embeddings across different LLMs, facilitating a shared understanding of the token space.
- Examine how using the same embeddings helps maintain semantic continuity when generating content, especially when switching between models of different sizes.

The literature review will conclude by linking these areas to the proposed dynamic LLM inference method and suggesting directions for future research.

0.3 Theoretical Framework

KL divergence, or Kullback-Leibler divergence, serves as a measure of how one probability distribution diverges from a second, reference probability distribution. In the domain of LLMs, which generate outputs by producing probability distributions over a vocabulary, KL divergence quantifies the difference in the predicted probabilities between a large, more accurate model and a smaller, faster model:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right)$$

Here, P and Q represent the probability distributions of the large and small LLMs, respectively. This measure is crucial in our approach as it provides a metric for assessing the ‘closeness’ of the small model’s output to that of the large model, which can be interpreted as a proxy for output quality. The suitability of KL divergence for this task lies in its sensitivity to differences in output probabilities, making it an appropriate tool for decision-making in dynamically selecting the appropriate model for inference tasks.

0.4 Methodology

Our methodology employs an iterative feedback mechanism to dynamically switch between a large and a small LLM based on ongoing quality assessments.

1. **Initialization**: We initiate content generation with the large LLM to establish a context with potentially higher coherence and relevance.

2. **Transition to Small LLM**: For subsequent content generation, we transition to the small LLM to benefit from its relative computational efficiency.

3. **Periodic Divergence Checks**:

   - At defined intervals, we generate text with the large LLM.
   - We then compute D_{KL} to evaluate the quality consistency between the models.
   - A decision to continue with the small LLM is made if D_{KL} is below a certain threshold, suggesting comparable quality.

4. **Iterative Process**: The cycle of generating content and assessing divergence constitutes a feedback loop, aiming to balance efficiency with content quality.

The threshold for D_{KL} is a hyperparameter informed by preliminary observations and could be subject to tuning as experimentation progresses.

0.5 Implementation

The practical deployment of this system includes:

1. **Model Selection**: Select large and small LLMs with a shared vocabulary, appropriate for the target application.

2. **Infrastructure Setup**: Configure computational resources to support both models and the dynamic switching mechanism.

3. **Divergence Threshold Setting**: Define an initial D_{KL} threshold based on exploratory tests, considering task-specific requirements for quality and computational efficiency.

4. **Monitoring System**: Implement a real-time monitoring system for periodic D_{KL} checks to determine when to switch between models.

The focus is on establishing a reliable and responsive system that can be tested and iterated upon to assess the viability of the approach in practical scenarios.

0.6 Experimentation

The system's performance will be empirically evaluated through a series of controlled tests:

- **Problem-Solving Tasks**: The large and small LLMs, along with the hybrid system, will be tasked with solving problems that have known solutions.

- **Correctness Assessment**: The correctness of the answers will be automatically checked, establishing a baseline success rate for each model.

- **Performance Metric**: We will calculate the relative performance improvement of the hybrid system over the small LLM, using the large LLM as a benchmark.

This approach allows us to quantify the effectiveness of the dynamic threshold mechanism in enhancing the problem-solving capabilities of the small LLM, providing a clear measure of the system's added value.

0.7 Experimentation

The empirical evaluation will focus on a set of reasoning tasks. These tasks will be automatically generated and will have objective measures of correctness. The large and small LLMs, alongside the hybrid system, will be compared based on their performance on these tasks.

- **Task Generation**: A suite of reasoning problems (e.g., integration, word problems, algorithmic problems like sorting) will be synthetically generated for testing.
- **Automated Evaluation**: A system will be put in place to automatically assess the correctness of the LLMs' solutions.
- **Relative Performance Calculation**: The relative performance improvement (RPI) of the hybrid system is defined as:

$$RPI = \frac{Correct_{hybrid} - Correct_{small}}{Correct_{large} - Correct_{small}}$$

  where $Correct_{hybrid}$, $Correct_{small}$, and $Correct_{large}$ are the numbers of correctly solved problems by the hybrid system, the small LLM, and the large LLM, respectively.

- **Experiment Details**:
  - *[Placeholder for specifics on problem types, generation methods, and solution evaluation criteria.]*
  - *[Placeholder for the selection criteria for problems included in the test suite.]*
  - *[Placeholder for the description of how the large and small models will be accessed and used in the experiment.]*
  - *[Placeholder for any preprocessing or postprocessing steps in the problem-solving pipeline.]*

This structured approach will allow us to quantitatively assess the value added by implementing the KL divergence-based switching mechanism in LLMs.

0.8 Experimental Results

The results of our experiments provide insight into the effectiveness of the dynamic threshold mechanism for LLM inference. We present detailed analyses, both qualitative and quantitative, to illustrate the impact of our approach.

0.8.1 Qualitative Analysis

- *[Placeholder for individual problem outputs and discussions on specific cases.]*

0.8.2 Quantitative Analysis

- *[Placeholder for visualizations showing RPI versus KL divergence threshold.]*
- *[Placeholder for statistical analysis and interpretation of results.]*

0.8.3 Troublesome Case Studies

- *[Placeholder for discussion on specific problems where the hybrid system's dynamic switching proved critical.]*

These results demonstrate the practical implications of our method and its potential to enhance the efficiency of LLMs in problem-solving tasks.

0.9 Future Work

0.9.1 Adaptive Strategies for Model Selection

Exploring adaptive strategies for model selection, including but not limited to Reinforcement Learning (RL), opens up promising research directions. These methods could dynamically adjust key parameters like the KL divergence threshold based on the context and task demands. For instance, using RL, the system could learn optimal adjustments over time, enhancing decision-making in a mixture of models framework. Each "expert" model in this mixture could be characterized by unique strengths, such as speed or analytical depth, offering a tailored approach to multimodal AI interactions and problem-solving tasks.

0.9.2 Mixture-of-Experts To Approximate Foundation Reference Model

Future research could investigate a system that employs a foundation reference LLM as a benchmark for quality. Smaller, specialized LLMs could be dynamically combined or selected based on their KL divergence from this reference model. When the divergence indicates that the smaller LLMs are significantly deviating from the expected output of the reference model, the system would default to the reference LLM. This approach could also explore weighted combinations of smaller LLM outputs, adjusted by their divergence, to optimize the balance between efficiency and fidelity to the reference model.

0.9.3 Divergence Over a Higher-Level Concept Space

The current approach uses KL divergence to compare the token probabilities of LLMs. It may be useful to map the token space to higher-level concepts, like `king` and `queen` to `ruler`, and then compare the distributions over these concepts. This would allow us to compare the models at a more abstract level, which would generally reduce the divergence between models, as the higher-level concepts would be more similar than the specific tokens, but it may be more relevant for certain applications.

0.10 Conclusion

This adaptive method aims to balance computational or cost efficiency with the quality of generated text, presenting a significant step towards more versatile text generation paradigms.