Uncovering the Mysteries of MoE: Developing Better Understanding and Performance of Mixture-of-Experts Transformers

General Public Abstract

Recent advancements in artificial intelligence have transformed our digital world, with large language models (LLMs) like ChatGPT becoming powerful tools capable of complex tasks such as text generation, translation, and problem-solving. These models, however, require immense computational resources due to their large number of parameters and extensive training datasets.

A promising innovation called Mixture-of-Experts (MoE) significantly enhances the efficiency of these models. Imagine MoE like a team of specialists (experts) working together. Each expert knows how to perform specific tasks exceptionally well. When a new task arises, the model selects only those experts needed for that particular task instead of involving everyone. This selective participation reduces computational effort and improves efficiency. One of the techniques that we have investigated in the past is called granularity. Granular MoE consists of more experts of smaller size. As we have previously shown, such (fine-grained) MoE models are even more efficient than regular (coarse-grained) models.

Despite their widespread use, our understanding of MoE models remains limited. This project aims to answer several key research questions:

- Can existing techniques for understanding traditional neural networks be effectively applied to MoE models?
- Do MoE models behave similarly to traditional models with very wide layers, or do they offer unique advantages?
- How does the granularity of the MoE model (fine vs. coarse-grained) change the representations (thought patterns) of MoE models?
- What role does having many possible paths for data processing within an MoE model play in its effectiveness?
- Under what circumstances do fine-grained MoE models outperform coarse-grained ones in practical applications?

By addressing these questions, we aim to better understand MoE models, allowing us to develop more efficient, powerful, and interpretable artificial intelligence. Improved interpretability is crucial for ensuring these technologies are transparent and more trustworthy

Our findings will guide future AI research, benefiting a wide range of applications and contributing to the development of smarter, more accessible technologies.