Week 11 Thursday Paper 1: A REVIEW OF SPARSE EXPERT MODELS IN DEEP LEARNING

Explained by

Sparse Expert Models (Switch Transformers, GLAM, and mo…

Supplemental: Sparse Expert Models: Past and Future

1. What forms of sparsity are used in this paper?
Ans: The paper discusses sparse expert models, including mixture-of-experts, Switch Transformers, Routing Networks, and BASE layers. These models operate under the principle that each input example is acted upon by a subset of the model's parameters, creating a form of sparsity. This approach decouples the parameter count from the compute per example, allowing for large yet computationally efficient models

2. What are the scaling properties of sparse expert models? Why are these properties desirable?
Ans: Sparse expert models excel in large data scenarios due to their unique scaling properties. These models can scale in terms of parameter count without a proportional increase in computational cost per example, which is desirable because it allows for significantly larger model sizes while maintaining computational efficiency. This scalability has been shown to yield significant improvements across domains like natural language processing, computer vision, and speech recognition. Moreover, the models adapt well to distributed computing environments, fitting naturally with parallelism schemes and benefiting from dynamic routing to efficiently use computational resources

3. Briefly explain the routing algorithms discussed in the paper.
Ans: The paper discusses several routing algorithms, including the original top-k routing mechanism and new advances. Top-k routing is foundational, where input tokens are routed to the top-k experts based on scores computed from token embeddings and expert embeddings. New advances in routing algorithms aim to improve efficiency and load balancing, such as reinforcement learning-based routing, which selects the top-1 expert with a reward mechanism, and other algorithms that address token overflow and

prioritize inputs with higher routing scores or adaptively compute based on token demands

4. Why is NLP a natural fit for sparse expert and MoE architectures? Do the potentially make sense for vision?
Ans: Sparse expert and mixture-of-experts architectures are well-suited for NLP due to the large, easily available datasets and the effectiveness of self-supervised learning techniques like next-word prediction. These models excel in efficiently handling the immense diversity and complexity of natural language. While most work on sparse expert models has been in NLP, their use is expanding into computer vision and speech recognition, indicating a potential for broader applicability. The adaptability of these models to different domains suggests they could also be effectively utilized in vision tasks, especially with the universal applicability of Transformer architectures

5. How does the paper address domain adaptation?

6. How can probing help to understand sparse expert models?
Ans: Probing, or the analysis of what information is captured by different parts of a model, could help understand sparse expert models by examining the specialization of experts. Identifying which types of inputs are routed to which experts can reveal how the model organizes knowledge and processes information, potentially offering insights into model behavior, efficiency, and areas for improvement.

7. What is “adaptive computation” and how would you address it given the tools and techniques of today?
Ans: Adaptive computation involves models that adjust the amount or type of computation based on the input, which is conceptually aligned with the idea behind sparse expert models where each input is processed by a different subset of parameters. Addressing adaptive computation with current tools and techniques could involve developing more advanced routing algorithms that dynamically allocate computational resources based on input characteristics or leveraging systems like Pathways, which facilitate efficient implementation of heterogeneous architectures on modern hardware