DSelect-k: Differentiable Selection in the Mixture of Experts with Applications to Multi-Task Learning

Main Idea: sparse gates commonly used for MoE, like top-k, are not smooth (continuously differentiable), which can lead to performance issues in gradient-based methods. Dselect-k presents a fully differentiable sparse gate for MoE.

Differentiability – a function that is differentiable is a function which has a defined derivative at every point. This is a requirement for gradient-based methods.

Continuous Differentiability – a function is continuously differentiable if it is fully differentiable AND the corresponding derivative is continuous. Continuous meaning smooth, with no abrupt changes and bumps. Continuous differentiability is not a requirement but optimizes performance of gradient-based methods.

- Top-k routing is not continuously differentiable due to the router's hard selection of experts. This hard routing leads to it being possible for small changes in the input score to have large changes in the expert weights, which is not ideal.
- Dselect-k achieves continuous differentiability through smoothing techniques.

My takeaways:

- Although Dselect-k in theory should perform better than top-k, this technique has not been applied much in practice. This can be attributed to the increased computational complexity it brings, as well as to the simplicity and proven practical use of top-k.
    - Some other recently proposed continuous differentiability methods for MoE routing, check soft MoE (optimized for vision) and Mixture-of-Tokens (optimized for text generation).