Soft Merging of Experts with Adaptive Routing

Main Idea: develop a technique called SMEAR (Soft Merging of Experts with Adaptive Routing) – single merged expert constructed via a weighted average of all the experts' parameters - to address the non-differentiability issue of discrete routing in MoE, hypothesizing that this lack of differentiability is what causes instabilities and underperformance in MoE.

Past research points that stable task/domain-level learned experts are possible (like in the DEMix line of work), but this is harder to achieve at the token-level. A few works showing the challenges of learned MoE at the token-level:

- Hash layer (random routing based on a fixed heuristic) achieves comparable results through a fixed random strategy.
- Switch and the Scaling Laws paper find that increasing the active parameters and the number of experts provides a predictable performance improvement, but this is not the same when just scaling the total number of parameters (this shows limited returns).
  - This can perhaps be explained by suboptimal routing.

With SMEAR, the authors hypothesize that these inefficiencies in MoE are caused by gradient estimation issues. First, they explore if fixed heuristic routing can overperform learned routing, and then compare that to SMEAR (which is fully differentiable).

SMEAR

- In traditional MoE routing, the router training needs to resort to gradient estimation techniques. The goal of SMEAR is to develop an architecture that enables end-to-end gradient-based training (fully differentiable, no gradient estimation) without an increase in computational costs.
- Ensemble routing
    - Would allow for an end-to-end gradient-based training but with a significant increase in computational costs.
- Merging of Experts
    - Recent work has shown that averaging the parameters of models that share a common architecture can often produce an aggregate model that shares the capabilities of the individual models.
- SMEAR
    - Constructs a single merged expert whose parameters are computed as the weighted average of the experts within a routing block.
        - Each expert's set of weights is set by the corresponding routing probability generated by the router.
    - Instead of only taking the top-k experts selected by the router, which is the discrete step in the strategy, SMEAR weighs each expert's parameters according to the weight given by the router and merges them into a single expert.
        - Allows updating each expert in each forward pass in a fully differentiable manner.

        - Almost equivalent (slightly higher due to the cost of merging) cost of top-1 routing at inference but more expensive training costs (due to having to backpropagate through each expert after each forward pass).

Experimental Setup

- Main question to be answered is if SMEAR can outperform heuristic routing strategies.
- Use T5 fine-tuned on GLUE for NLP tasks, while also conducting computer vision experiments based on ResNet.
- Used a “tag routing” strategy as one of the baselines, which is a routing strategy based on metadata (oracle routing).
- Add experts to existing pre-trained network (models are not trained from scratch and are based off pre-trained dense models).
    - Similarly to adding adapters for fine-tuning (all pre-trained parameters are kept frozen).
- Router is a simple linear classifier.
- Each layer has 8 experts.
- No balance loss was used.

Results

- Models using learned routing strategies learned through gradient estimation (thus not fully differentiable) often underperform heuristic routing strategies.

- SMEAR outperforms every routing strategy (heuristic or learned) in both NLP and Vision settings, including tag routing (determined by metadata) and a parameter-matched (in terms of total parameters) dense baseline.
    - Consistent with DEMix line of research, which says that a good learned routing strategy should be better than routing determined by metadata.
- SMEAR performs comparably to a fully active MoE ensemble (especially in T5-GLUE), which is seen as the upper bound of this approach.
- In terms of inference, SMEAR performs comparably to the top-1 routing strategy.
- Doubling the number of experts (from 8 to 16) in SMEAR led to a slight performance boost in Vision but no notable difference in T5-GLUE.
- Significant sparsity observed when visualizing the router's distribution, suggesting expert specialization.

My takeaways:

- SMEAR offers a novel training framework that might set a precedent for future MoE models by mitigating the non-differentiability issue common in discrete routing decisions, thereby leading to more stable and efficient learning.
- The gradual diversification from a single expert to a full MoE configuration in SMEAR could inspire new initialization techniques for complex neural networks, ensuring a smoother transition to specialized expert utilization.

- Given SMEAR’s performance improvements and computational efficiency, it would be worthwhile to investigate how it could be adapted to real-world tasks requiring modularity and efficiency, such as personalized recommendation systems or multi-domain language models.