Learning Factored Representations in a Deep Mixture-of-Experts

Main Idea:

- To apply stacked layers of mixture-of-experts, so to have multiple sets of (gating, experts). This allows multiple combinations of experts to be called while keeping a modest model size.

The problem they are trying to solve for is that deep neural networks are expensive to compute at inference time since all the neurons are used.

The solution proposed is to implement stacked MoE layers, where multiple expert combinations are possible, and the gating mechanism ensures only useful neurons for that input are used (experts on the specific input space). This gives better computational efficiency at inference, allowing for a model that is both large and efficient.

Approach:

- The input is first passed through the first MoE layer (represented by z1):
    - $z_1 = \sum_{i=1}^{N} g_{i1}(x) \cdot f_{i1}(x)$, where $g_{i1}(x)$ and $f_{i1}(x)$represent the gating probability and expert output for expert *i* at layer 1, respectively.
        - both the gating mechanism and the expert function use a non-linearity (ReLU)

- The outputs of the first layer (z1) are then passed as an input to the next MoE layer z2, which replaces x with z1.
- z2 is then passed through a final layer (f3) and a softmax is applied (in the context of classification)
    - $F(x) = z_3 = Softmax(f_3(z_2))$

The network is trained with SGD with a caveat to help balance the training through the experts:

- The mean of all experts' total assignment is compared to each expert's running total assignment. If an expert is found to have a running total assignment significantly higher than the mean, its training is paused temporarily to allow for the training of other experts.
- This strategy is found to mostly be useful in early stages of training, where the experts have not yet specialized significantly on a part of the input space. After some training, the experts are expected to have some specialization, and thus this constraint can be lifted.

This paper makes use of conditional computation, although the details about this are not shown in-depth.

Results:

- The stacked MoE layer showed promising results, as it came close to fully dense networks in terms of performance while having significant inference pros due to conditional computation.
- Experiments in specific tasks also showed that different experts indeed did specialize in different clusters of the data.

My takeaways:

- This paper is revolutionary in terms of the idea presented in terms of stacking MoE layers in a deep neural network and trying to find a way to balance the load between experts.
- Introduces the idea that MoE can have improved performance when stacked, paving the way for adding this as a modular component that can be added to other architectures.
- This strategy is still not sparse (top-k), but it opens the field to the idea that a top-k strategy is possible as a future line of research.