BASE Layers: Simplifying Training of Large, Sparse Models

Main Idea: introduces a new routing approach that approaches the problem as a linear assignment. This ensures load balancing without the need for auxiliary losses or adjusting CF. BASE also shows that a single expert/MoE layer can be effective.

- Makes use of top-1 routing like Switch.
- The linear assignment problem is designed to maximize token-expert affinities and has the constraint of balanced loads.

BASE Algorithm

1. Compute token-expert score for all experts.
2. Solve the linear assignment problem.
    a. Goal - Maximize token-expert affinity.
    b. Constraint – ensure balanced loads to experts at a batch-level.
3. Route tokens to experts.
4. Compute the expert scores as a weighted sum based on the routing weights.
    a. Top-2 routing is used at training.
5. Return the output to the original worker.

This approach is only used during training, as during test time the strategy of top-1 routing without load balancing is taken.

Results

- Having a single BASE layer in the network can be effective.
- Expert layers are robust to changes in the expert-shared parameters ratio and the position(s) of the layer in the network.
- Exploration of which inputs are assigned to each expert shows the same specialization patterns of other works: experts specialize on simple input patterns related to semantics and syntax.