1. Motivation

Mixture of Experts (MoE) are rising in popularity as a means to train extremely large models yet allowing for a reasonable computational cost at inference time. However, state-of-the-art approaches either:

- (large-scale MoE) utilize many experts and routing decisions that have to be trained jointly, which leads to training instabilities and can make it hard to implement the routing in practice
- (hierarchical classifiers) define rigid per-class routing that might not be optimal subsets of the data to train on

We propose to revisit the single-gate MoE and improve its accuracy-efficiency trade-off, as well as training practicality. Key to our work are:

- A full base model branch acting both as an early-exit (efficiency) and an ensembling regularization scheme (accuracy)
- A simple and efficient asynchronous training pipeline without router collapse issues
- An automatic per-sample clustering-based initialization.

2. Our Improved Single-gated MoE Design

Asynchronous Training algorithm

Step 1: Train the base model $\phi$ (or use off-the-shelf) then freeze

Step 2 (init routing): Cluster the base model embeddings using K-means, obtaining cluster centers $c_k$.

Define target gate $g^* = argmax_k$ then freeze

For $k = 1$ to $K$ (asynchronous) do

- Initialize $k$-th expert from the base model's weights
- Sample training example set $D_k$ by following the distribution given by $g^*$, where $g$ is a regularization noise
- Train the $k$-th expert on $D_k$

3. Training Scheme

The components of our model are:

- The base model $\phi$ is network trained on the whole dataset, and is executed for every input. It captures shared generic knowledge.
- Experts $e_k$ take as input an intermediate feature map of the base model. At inference, the most probable expert is executed. They capture specialized knowledge.
- Ensembles $d_k$ combines outputs of the base model and selected expert. We experimented with several ensembling designs and use bagging in practice: $d_k (x) = \phi (x) + e_k (x)$

4. Any-time Performance for Maximized Efficiency

Default Behavior (static): Select the top-1 expert chosen by the gate

Early-exiting (dynamic): Exit after the base model forward pass

Top-k experts (dynamic): Select more than one expert and combines their output via ensembling

We find that we can implement both dynamic behavior with a simple thresholding rule and achieve good performance. More complex (e.g., learned) early-exiting strategies did not help.

$\alpha_k = g(k | x) \cdot (1 - max \phi (y | x))$

Combined gate and base model confidence:

$e(x) = 1$ if $\forall k \alpha_k (x) < \tau$

Early exit if no expert is confident enough

5. Results on Image Classification

Comparison to baseline models:

- # gates
- Acc
- #MACs
- # train.

6. Per-sample Assignment

The per-sample routing uncovers meaningful intra-class variations. This shows the limits of per-class routing (e.g., hierarchical classification) as it can sometimes be too rigid to capture data diversity

The class king-penguin (left) co-occurs with other animals (right) for full-view images.

but is grouped with e.g., bell pepper when the image is a close-up of its orange beak

Conclusions

- We augment MoE with a novel ensembling scheme and a simple asynchronous and stable training pipeline leveraging a per-sample clustering-based initialization.
- Our model consistently reaches higher accuracy than hierarchical classifiers and a 1-expert ensembling baseline, revealing the benefits of training specialized experts with per-sample routing.
- Finally, maintaining the base model as an independent branch allows us to further save computations at inference time using a simple threshold-based conditional rule to adapt the computational budget without retraining.


Qualcomm Technologies Netherlands, B.V.

Mikel Urrutia

Piotr Uszynski

Tijmen Blankevoort

Amélie Royer

Amir Salleh

Babak Ashrafipour

Tijmen Blankevoort

Qualcomm Technologies Netherlands, B.V.

Amélie Royer, Amir Salleh, Babak Ashrafipour, Tijmen Blankevoort

Tijmen Blankevoort