BLOG ›

Mixture-of-Experts with Expert Choice Routing

WEDNESDAY, NOVEMBER 16, 2022

Posted by Yanqi Zhou, Research Scientist, Google Research, Brain Team

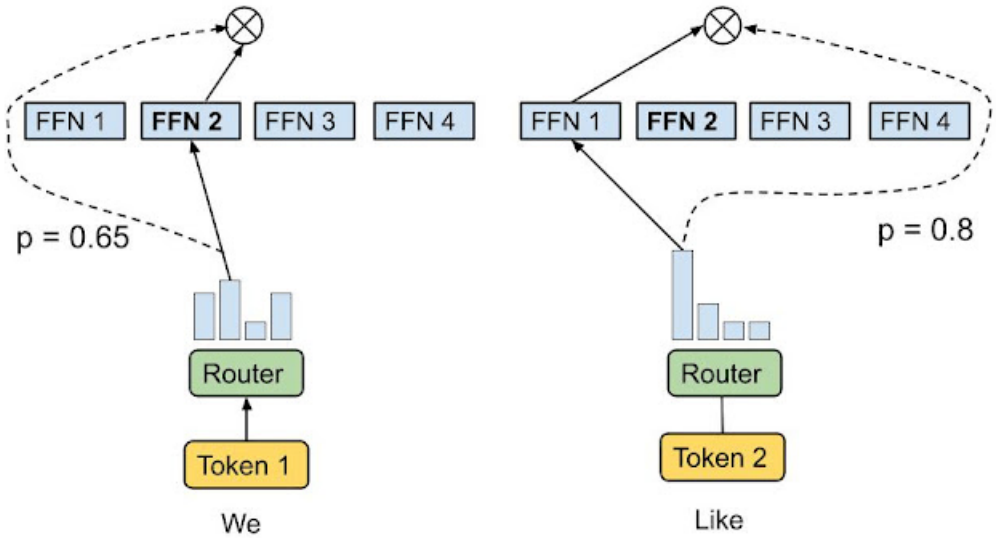
The capacity of a neural network to absorb information is limited by the number of its parameters, and as a consequence, finding more effective ways to increase model parameters has become a trend in deep learning research. **Mixture-of-experts** (MoE), a type of conditional computation where parts of the network are activated on a per-example basis, has been proposed as a way of dramatically increasing model capacity without a proportional increase in computation. In sparsely-activated variants of MoE models (e.g., **Switch Transformer**, **GLaM**, **V-MoE**), a subset of experts is selected on a per-token or per-example basis, thus creating sparsity in the network. Such models have demonstrated better scaling in multiple domains and better retention capability in a continual learning setting (e.g., **Expert Gate**). However, a poor expert routing strategy can cause certain experts to be under-trained, leading to an expert being under or over-specialized.

In "**Mixture-of-Experts with Expert Choice Routing**", presented at **NeurIPS 2022**, we introduce a novel MoE routing algorithm called Expert Choice (EC). We discuss how this novel approach can achieve optimal load balancing in an MoE system while allowing heterogeneity in token-to-expert mapping. Compared to token-based routing and other routing methods in traditional MoE networks, EC demonstrates very strong training efficiency and downstream task scores. Our method resonates with one of the vision for **Pathways**, which is to enable heterogeneous mixture-of-experts via Pathways MPMD (multi program, multi data) support.

Overview of MoE Routing

MoE operates by adopting a number of experts, each as a sub-network, and activating only one or a few experts for each input token. A gating network must be chosen and optimized in order to route each token to the most suited expert(s). Depending on how tokens are mapped to experts, MoE can be sparse

work has implemented sparse routing via *k-means clustering*, linear assignment to maximize token-expert affinities, or hashing. Google also recently announced GLaM and V-MoE, both of which advance the state of the art in natural language processing and computer vision via sparsely gated MoE with top-*k* token routing, demonstrating better performance scaling with sparsely activated MoE layers. Many of these prior works used a *token choice* routing strategy in which the routing algorithm picks the best one or two experts for each token.

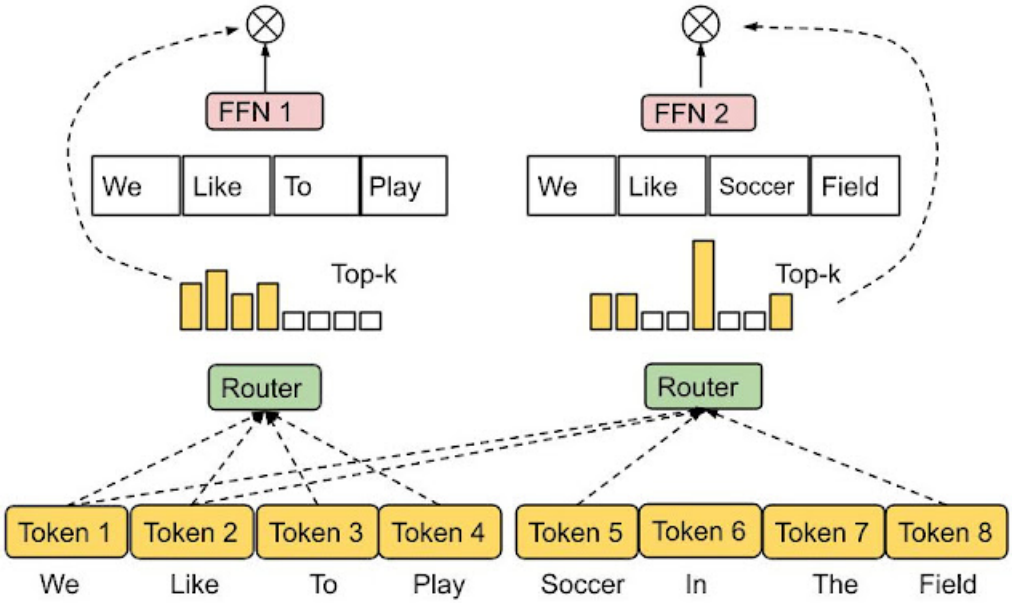


Token Choice Routing. The routing algorithm picks the top-1 or top-2 experts with highest affinity scores for each token. The affinity scores can be trained together with model parameters.

The independent token choice approach often leads to an imbalanced load of experts and under-utilization. In order to mitigate this, previous sparsely gated networks introduced additional auxiliary losses as regularization to prevent too many tokens being routed to a single expert, but the effectiveness was limited. As a result, token choice routings need to overprovision expert capacity by a significant margin (2x–8x of the calculated capacity) to avoid dropping tokens when there is a buffer overflow.

In addition to load imbalance, most prior works allocate a fixed number of experts to each token using a top-*k* function, regardless of the relative importance of different tokens. We argue that different tokens should be received by a variable number of experts, conditioned on token importance or difficulty.

the **expert choice** routing method illustrated below. Instead of having tokens select the top-*k* experts, the experts with predetermined buffer capacity are assigned to the top-*k* tokens. This method guarantees even load balancing, allows a variable number of experts for each token, and achieves substantial gains in training efficiency and downstream performance. EC routing speeds up training convergence by over 2x in an 8B/64E (8 billion activated parameters, 64 experts) model, compared to the top-1 and top-2 gating counterparts in **Switch Transformer**, **GShard**, and **GLaM**.



Expert Choice Routing. Experts with predetermined buffer capacity are assigned top-k tokens, thus guaranteeing even load balancing. Each token can be received by a variable number of experts.

In EC routing, we set expert capacity *k* as the average tokens per expert in a batch of input sequences multiplied by a *capacity factor*, which determines the average number of experts that can be received by each token. To learn the token-to-expert affinity, our method produces a token-to-expert score matrix that is used to make routing decisions. The score matrix indicates the likelihood of a given token in a batch of input sequences being routed to a given expert.

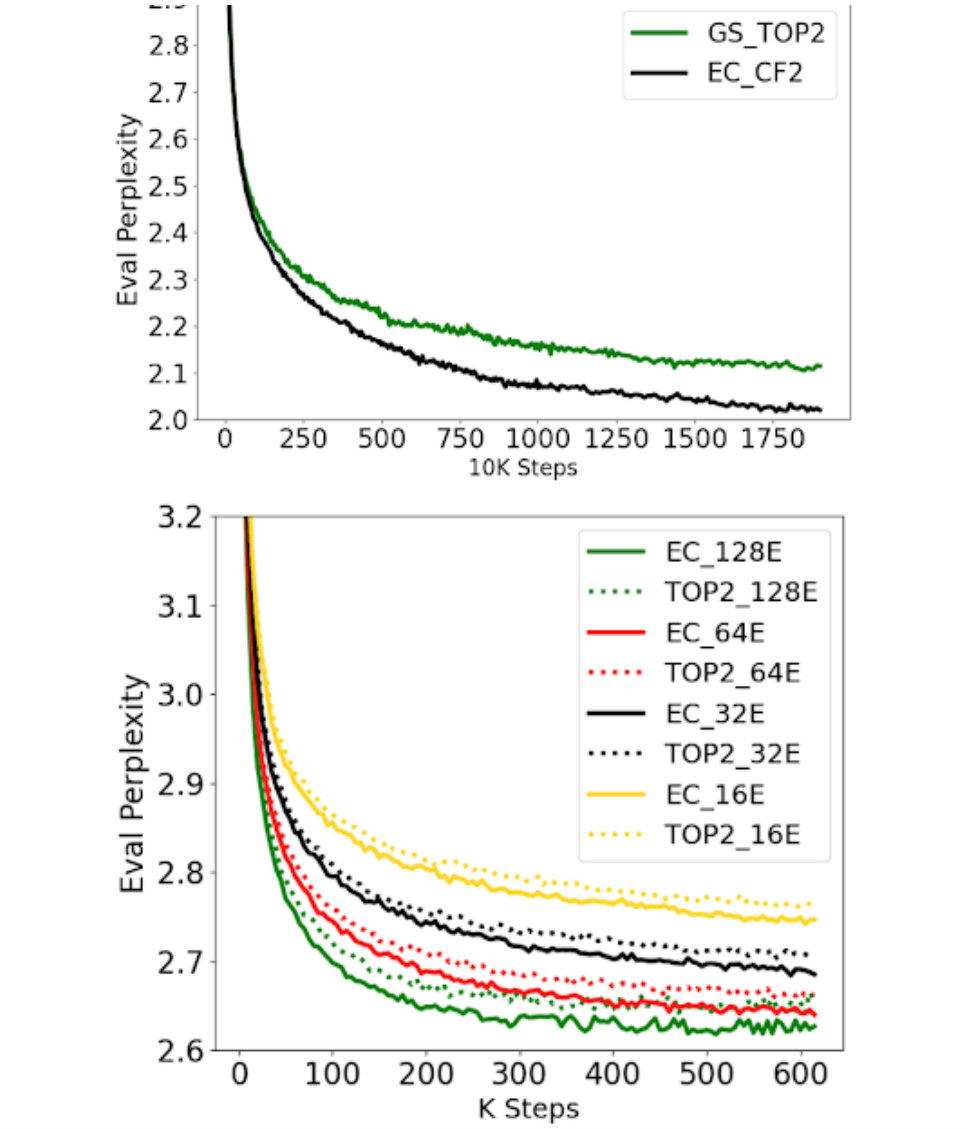
Similar to Switch Transformer and GShard, we apply an MoE and gating function in the dense **feedforward** (FFN) layer, as it is the most computationally expensive part of a Transformer-based network. After producing the token-to-expert score matrix, a top-*k* function is applied along the token dimension for each expert to pick the most relevant tokens. A permutation function is then

such that all experts can execute the same computational kernel concurrently on a subset of tokens. Because a fixed expert capacity can be determined, we no longer overprovision expert capacity due to load imbalancing, thus significantly reducing training and inference step time by around 20% compared to GLaM.

Evaluation

To illustrate the effectiveness of Expert Choice routing, we first look at training efficiency and convergence. We use EC with a capacity factor of 2 (EC-CF2) to match the activated parameter size and computational cost on a per-token basis to GShard top-2 gating and run both for a fixed number of steps. EC-CF2 reaches the same perplexity as GShard top-2 in less than half the steps and, in addition, we find that each GShard top-2 step is 20% slower than our method.

We also scale the number of experts while fixing the expert size to 100M parameters for both EC and GShard top-2 methods. We find that both work well in terms of perplexity on the evaluation dataset during pre-training — having more experts consistently improves training perplexity.



*Evaluation results on training convergence: EC routing yields 2x faster convergence at 8B/64E scale compared to top-2 gating used in GShard and GLaM (**top**). EC training perplexity scales better with the scaling of number of experts (**bottom**).*

To validate whether improved perplexity directly translates to better performance in downstream tasks, we perform fine-tuning on 11 selected tasks from GLUE and SuperGLUE. We compare three MoE methods including Switch Transformer top-1 gating (ST Top-1), GShard top-2 gating (GS Top-2) and a version of our method (EC-CF2) that matches the activated parameters and computational cost of GS Top-2. The EC-CF2 method consistently outperforms the related methods and yields an average accuracy increase of more than 2% in a large 8B/64E setting. Comparing our 8B/64E model against its dense counterpart, our method achieves better fine-tuning results, increasing the average score by 3.4 points.

allowing a variable number of experts per token is indeed helpful. On the other hand, we compute statistics on token-to-expert routing, particularly on the ratio of tokens that have been routed to a certain number of experts. We find that a majority of tokens have been routed to one or two experts while 23% have been routed to three or four experts and only about 3% tokens have been routed to more than four experts, thus verifying our hypothesis that expert choice routing learns to allocate a variable number of experts to tokens.

Final Thoughts

We propose a new routing method for sparsely activated mixture-of-experts models. This method addresses load imbalance and under-utilization of experts in conventional MoE methods, and enables the selection of different numbers of experts for each token. Our model demonstrates more than 2x training efficiency improvement when compared to the state-of-the-art GShard and Switch Transformer models, and achieves strong gains when fine-tuning on 11 datasets in the GLUE and SuperGLUE benchmark.

Our approach for expert choice routing enables heterogeneous MoE with straightforward algorithmic innovations. We hope that this may lead to more advances in this space at both the application and system levels.

Acknowledgements

Many collaborators across google research supported this work. We particularly thank Nan Du, Andrew Dai, Yanping Huang, and Zhifeng Chen for the initial ground work on MoE infrastructure and Tarzan datasets. We greatly appreciate Hanxiao Liu and Quoc Le for contributing the initial ideas and discussions. Tao Lei, Vincent Zhao, Da Huang, Chang Lan, Daiyi Peng, and Yifeng Lu contributed significantly on implementations and evaluations. Claire Cui, James Laudon, Martin Abadi, and Jeff Dean provided invaluable feedback and resource support.

Labels: Machine Learning NeurIPS

Previous posts

NOV 11, 2022

Characterizing Emergent Phenomena in

NOV 10, 2022

Multi-layered Mapping of Brain Tissue via

NOV 9, 2022

ReAct: Synergizing Reasoning and Acting in Language