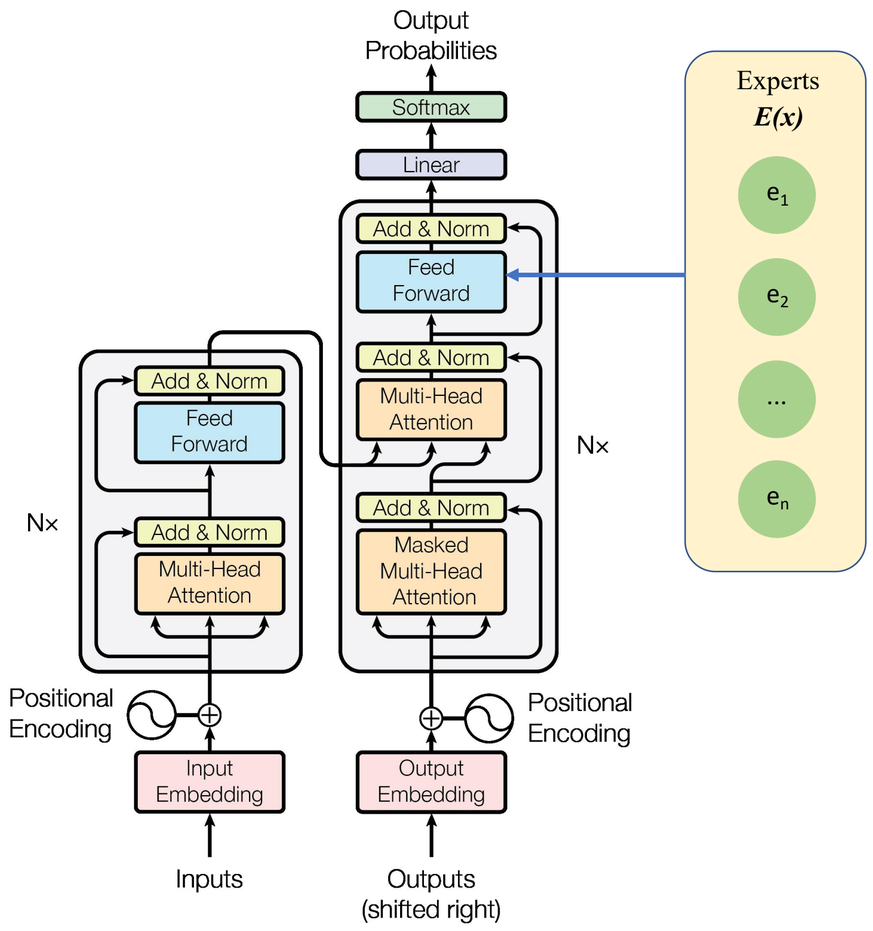
Design and Implement of Model Inference Acceleration via Mixture-of-Experts

Sen Yang[1], Xiaolong Liu[1], Ye Ji[1], Aobo Guo[1]

NLP Lab, School of artificial intelligence, Northeastern University[1]

Background

Structure of Transformer

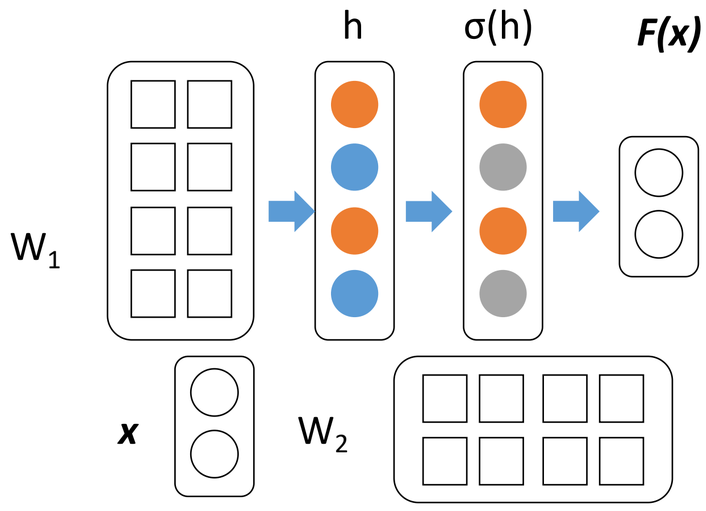


We divide the FFN layer in Decoder of Transformer models into experts that focus on different neurons. This improves the computation and performance of our model.

Sparsity Verification

Sparsity in FFN layer

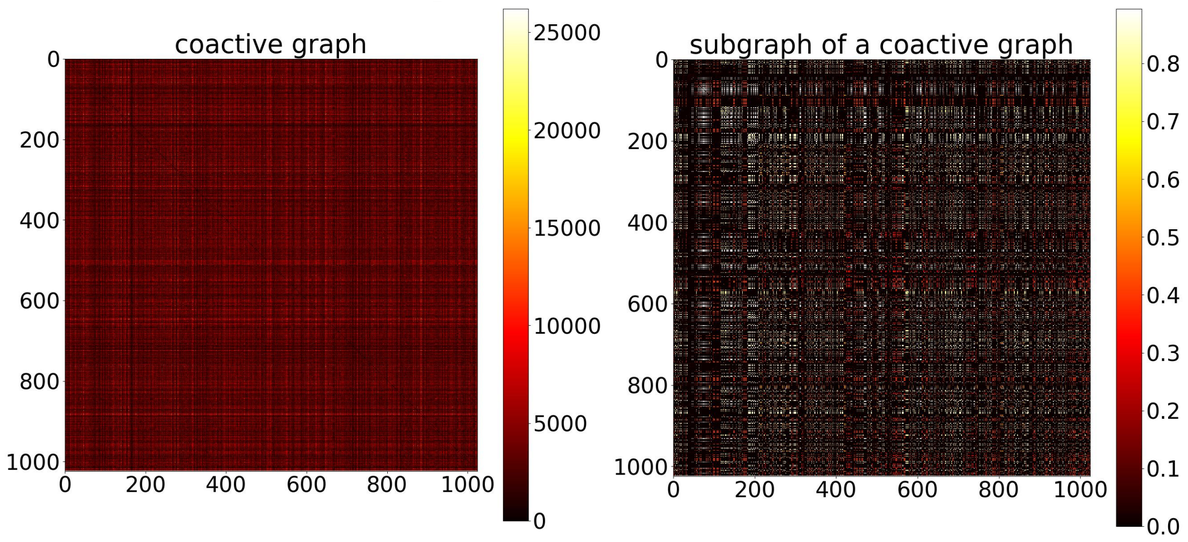
Calculating the ratio of neurons recorded as negative in FFN, it was found that about 90% of neurons in Encoder and about 80% to 90% in Decoder were not activated, which means only a small fraction of the neurons are actually involved in work. We found that it can conditionally use only 10% ~ 20% of the FFN parameters while maintaining more than 95% of the original performance, consistents with the MoEfication[1] which is 10%~30%.



MoE Network

Co-activation diagram

We use a co-activation graph to represent each FFN. The nodes in the graph are neurons and the edge weight is the number of times two neurons are activated at the same time. We divide the graph into several internally connected subgraphs (Karypis and Kumar,1998)[2], each of which is an expert.



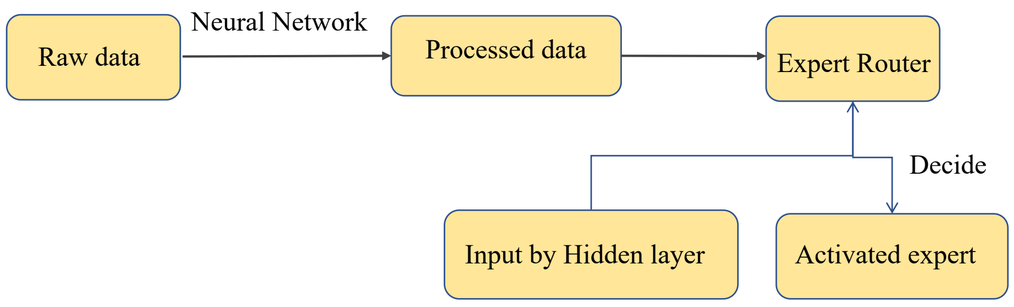
Cluster splitting

We assume that similar vector operations will activate similar neurons. We regard the column of W_1 matrix in FFN as the vector set of d_{model} dimension, where d_{model} is the size of the encoder/decoder input. We use the balanced k-means of k clusters to divide these vectors into experts.

Gate Network

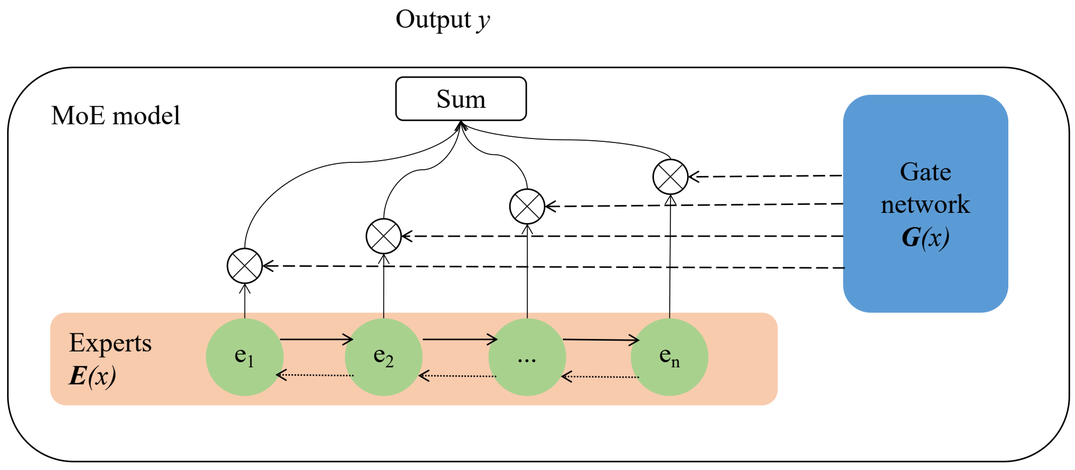
Expert Router

Provide as many activated neurons as possible with limited experts.



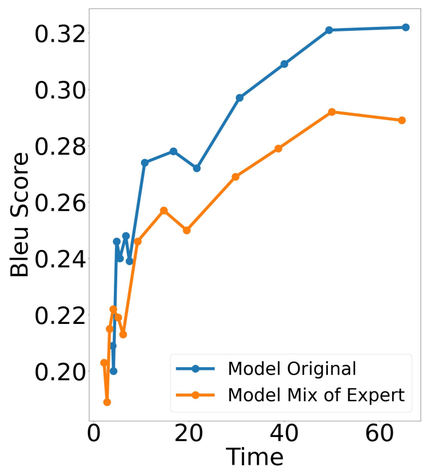
Network Frame

Flow chart

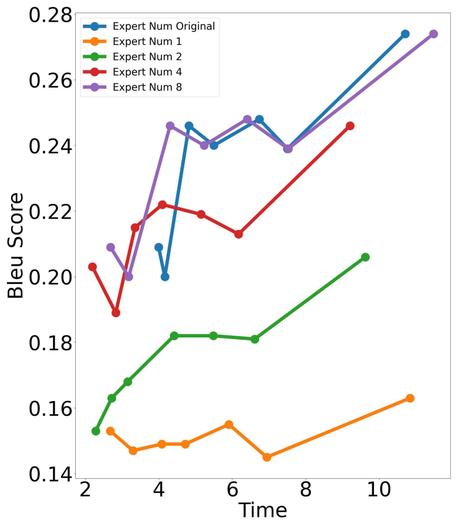


Results

Whether to use MoE



How many experts to use



Sample	1	2	3	4	5	6	7	8	9	10
Origin-Time	4.00	4.16	4.82	5.50	6.73	7.50	10.71	16.74	21.63	30.62
Origin-Bleu	0.21	0.20	0.25	0.24	0.25	0.24	0.27	0.28	0.27	0.30
MoE-Time	2.19	2.82	3.35	4.09	5.15	6.17	9.21	14.71	19.51	29.78
MoE-Bleu	0.20	0.19	0.22	0.22	0.21	0.25	0.26	0.25	0.27	0.28

Conclusion

- ✓ When using only half of the experts in Decoder layer, the speed is improved by about 20% compared with the original model, and the accuracy loss is about 10%.

Reference

[1] Zhang, Zhengyan, et al. "MoEfication: Transformer Feed-forward Layers are Mixtures of Experts." arXiv preprint arXiv:2110.01786 (2021).

[2] Karypis, George, and Vipin Kumar. "A fast and high-quality multilevel scheme for partitioning irregular graphs." SIAM Journal on scientific Computing 20.1 (1998): 359-392.