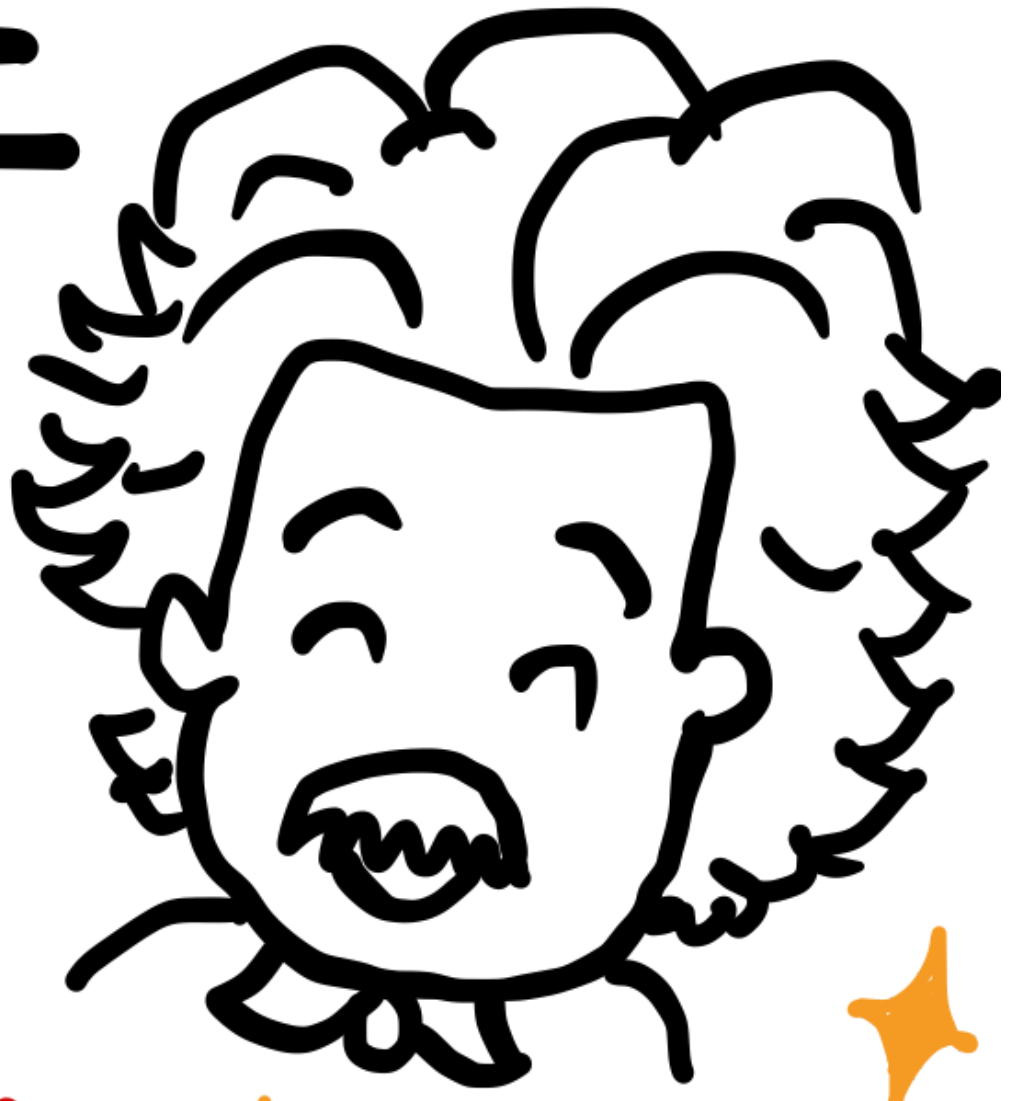
MIXTURE Of EXPERTS Explained



- by @ydnysh

Scaling is one of the important parameters to achieve better model quality. But scaling comes with its own set of challenges like increased model size and compute requirements.

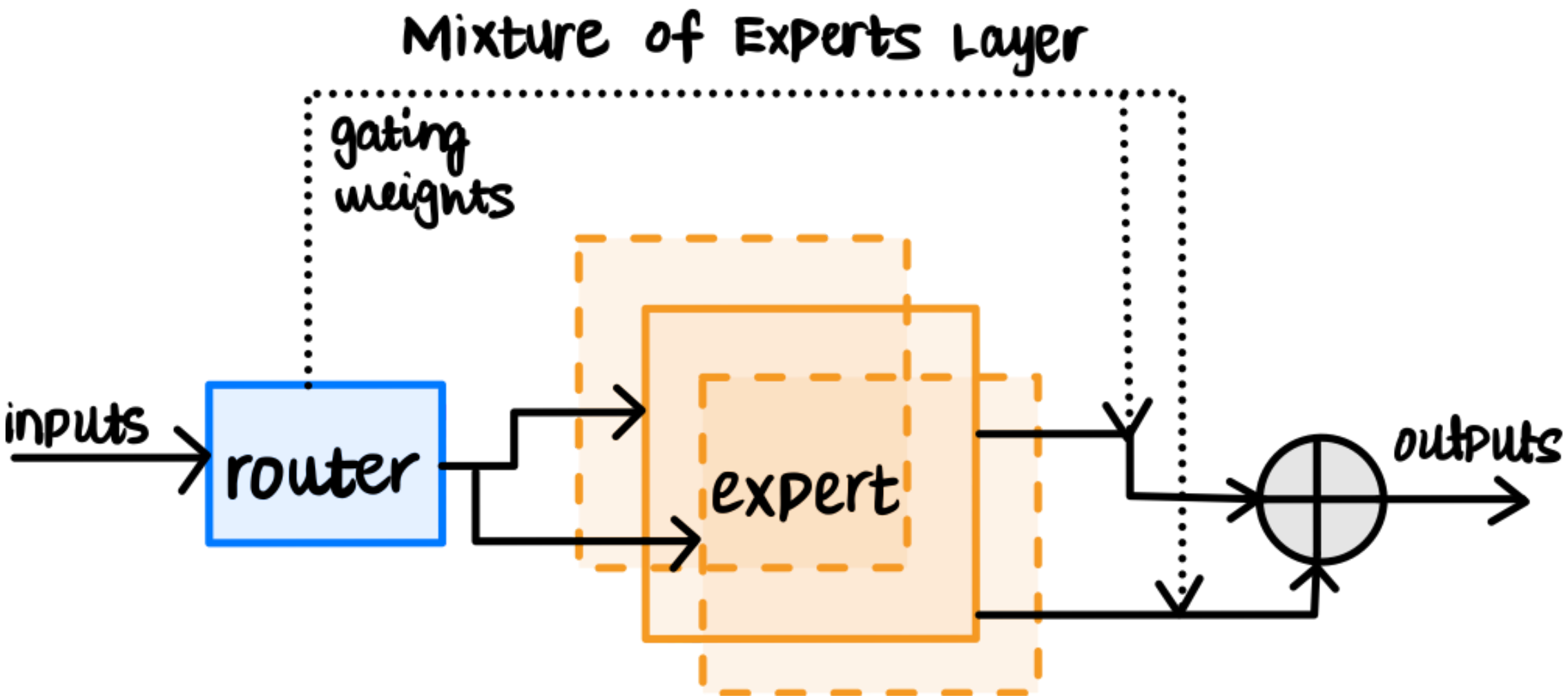
To address these challenges the Mixture of Experts (MoE) model was introduced.

MoE enable models to be pretrained with far less compute and helps us to scale up the model size with same compute budget as the dense model.

MoE has the following two main components :

1) Sparse MoE Layers

2) Gate Network / Router



1) Sparse MoE Layers :

↳ replace traditional layers in a model with MoE layers

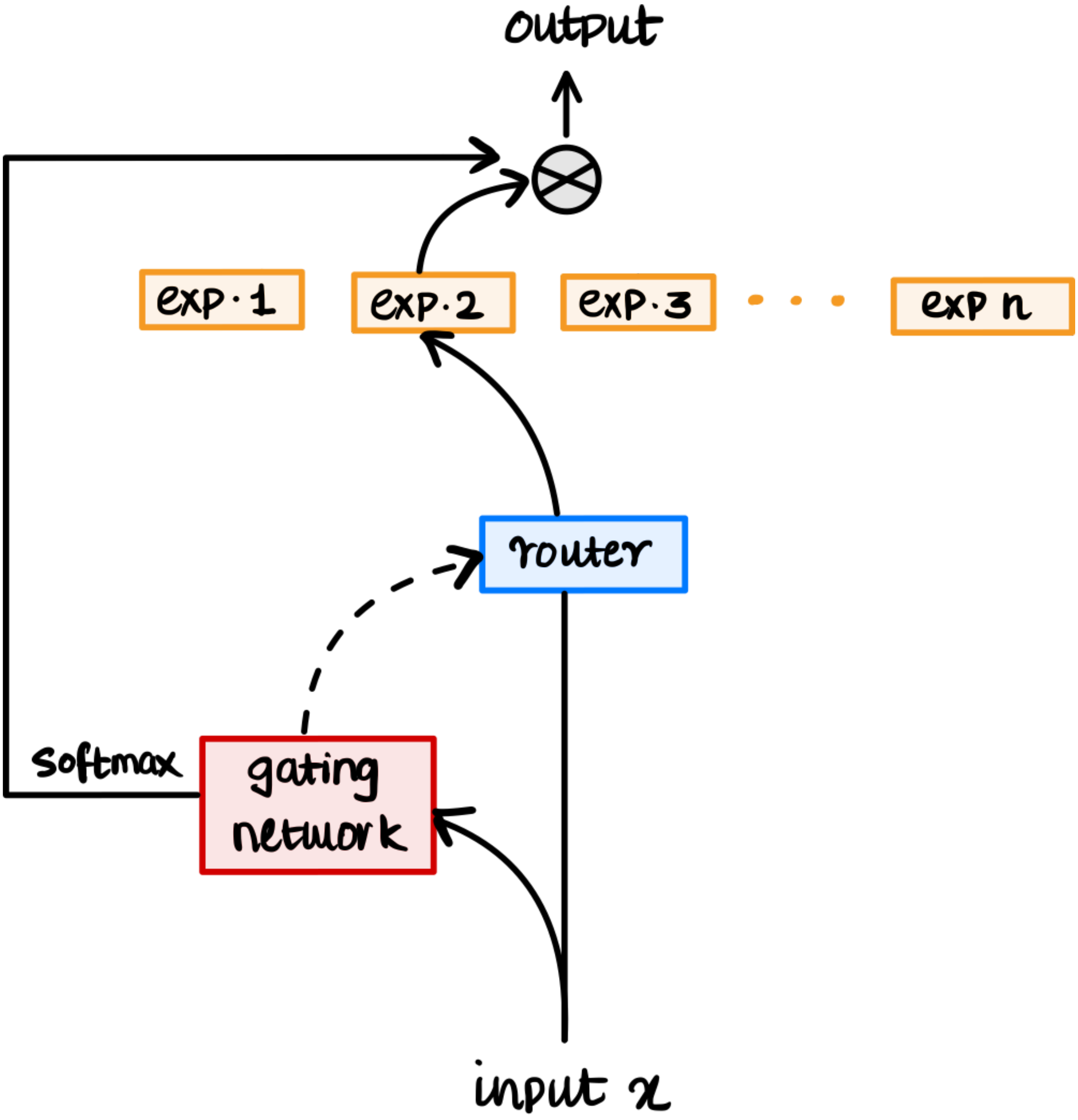
↳ these layers consists of different experts, each of them functioning as a small neural network

↳ for e.g : 5 experts → 5 specialized neural networks.

2) Gate Network / Router :

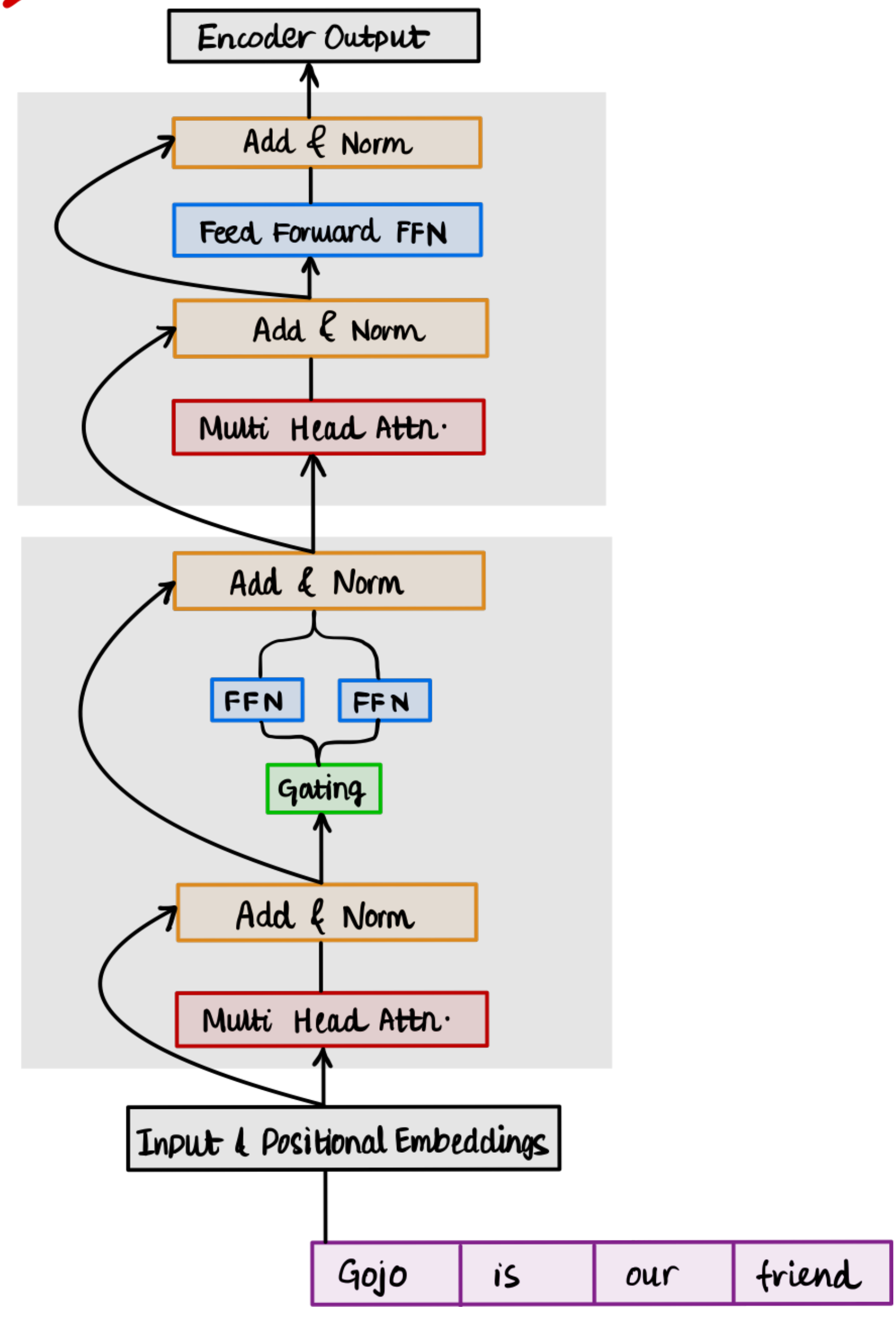
↳ it's a routing mechanism deciding which expert handles a specific input.

↳ it's like a traffic police directing cars to different lanes.


output
exp·1
exp·2
exp·3
exp n
router
Softmax
gating
network
input x

Routing a token/input to an expert is a crucial task for the MoEs. The router is composed of learned parameters and is pretrained at the same time as the rest of the network.

MOE in Transformers



Benefits of MoEs :

↳ Effecient Pretraining -
Models can be pretrained faster with less computational efforts.

↳ training a larger model for fewer steps yields better results than training a small model for more steps.

↳ Faster Inference -

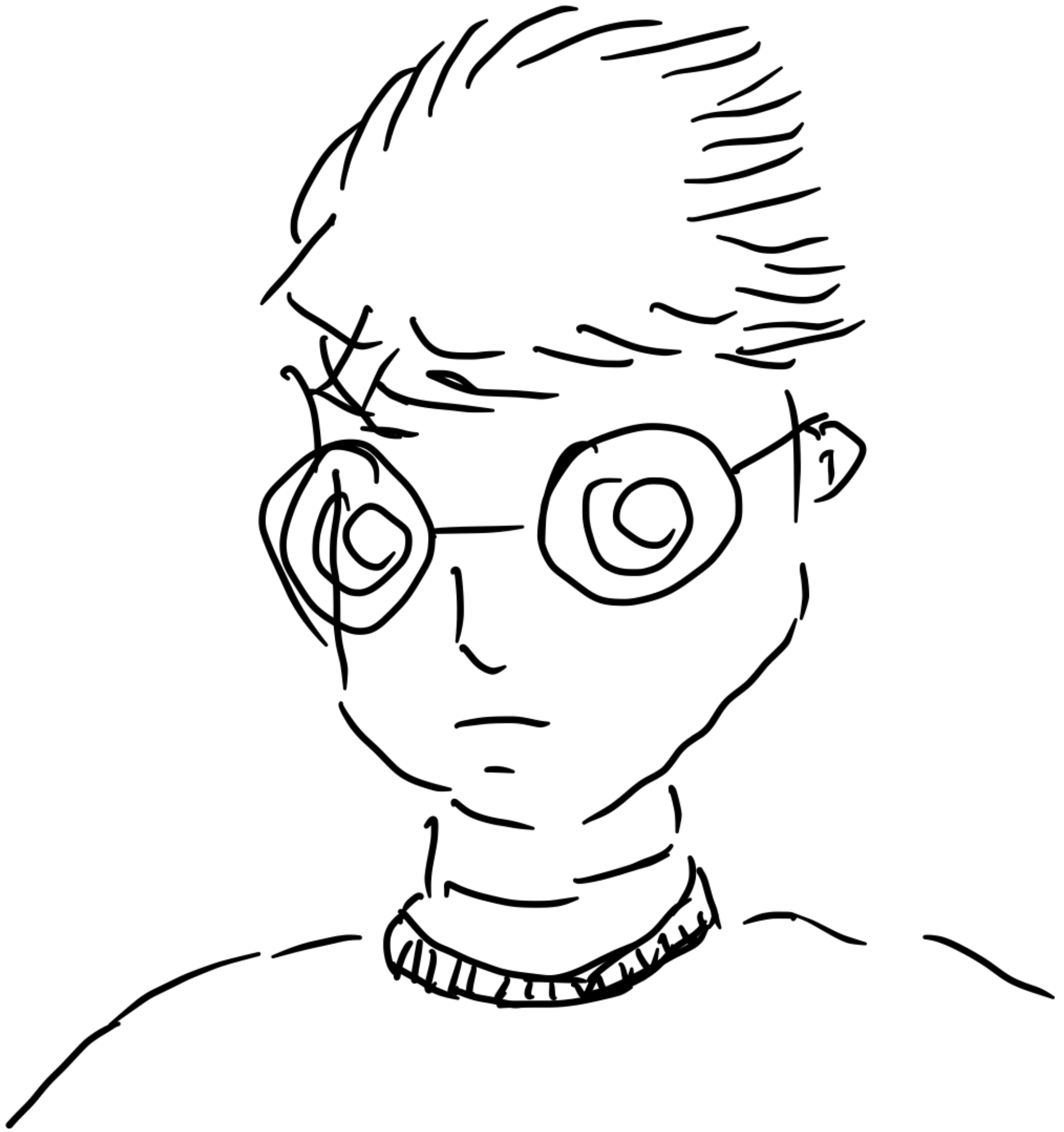
During inference MOE models exhibit faster speed than the dense ones w/ same no. of parameters.

↳ Despite having many parameters, only a subset are used for faster predictions.

Training Challenges :

↳ MOEs have faced issues with generalization during fine tuning, resulting in overfitting.

↳ MOE instruction-tuning has shown promising results addressing the fine-tuning challenges.



@ yadnyesh