

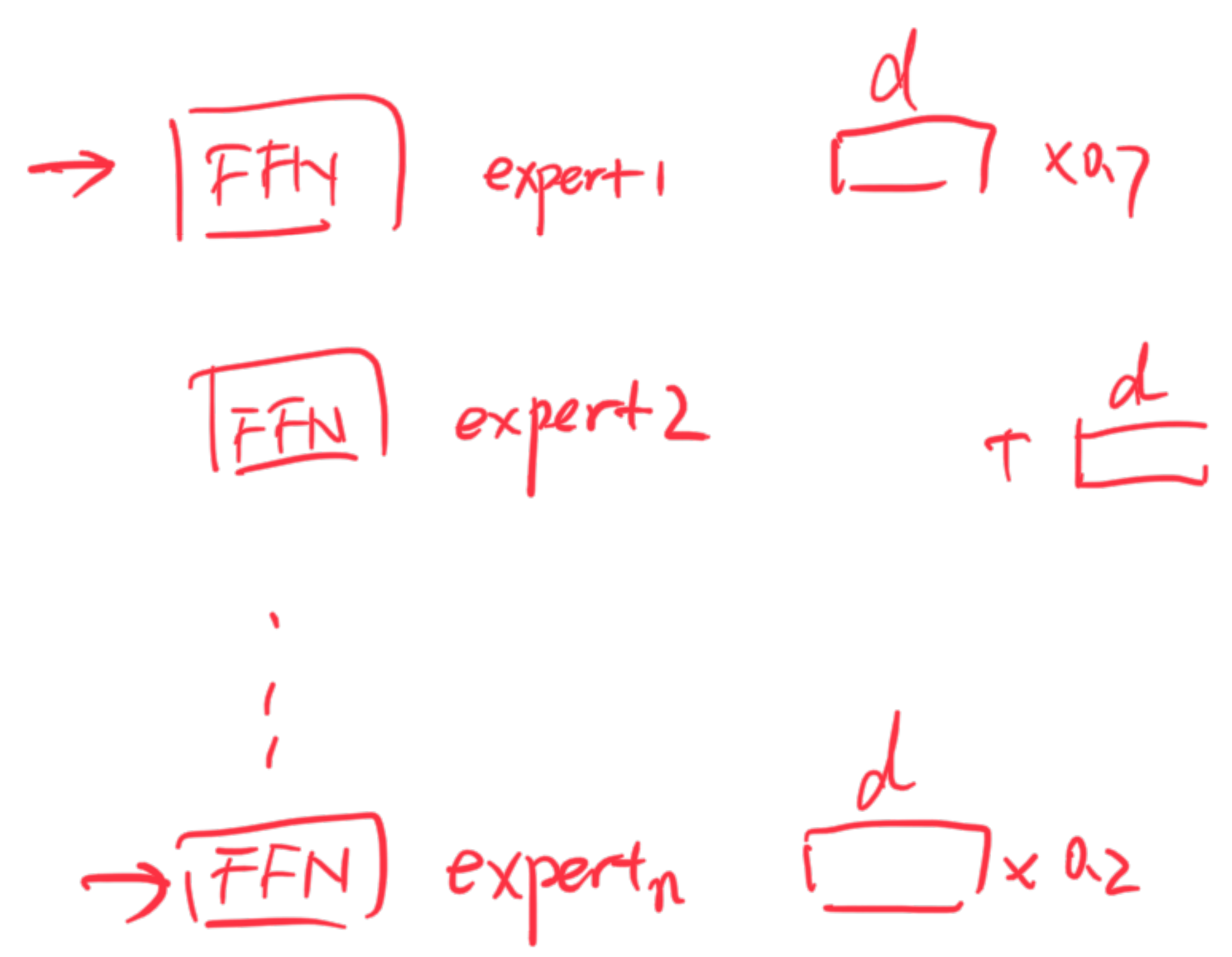
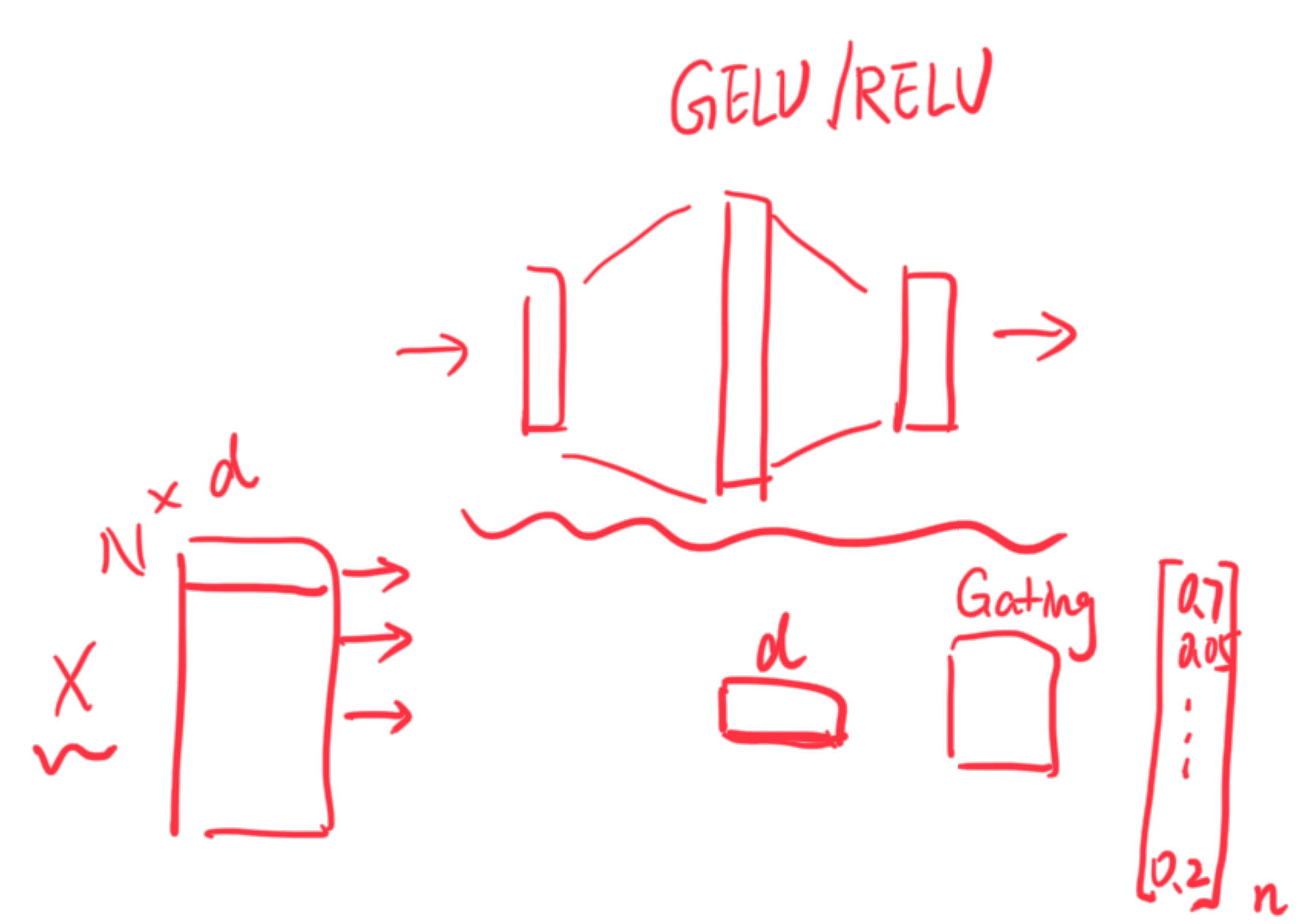
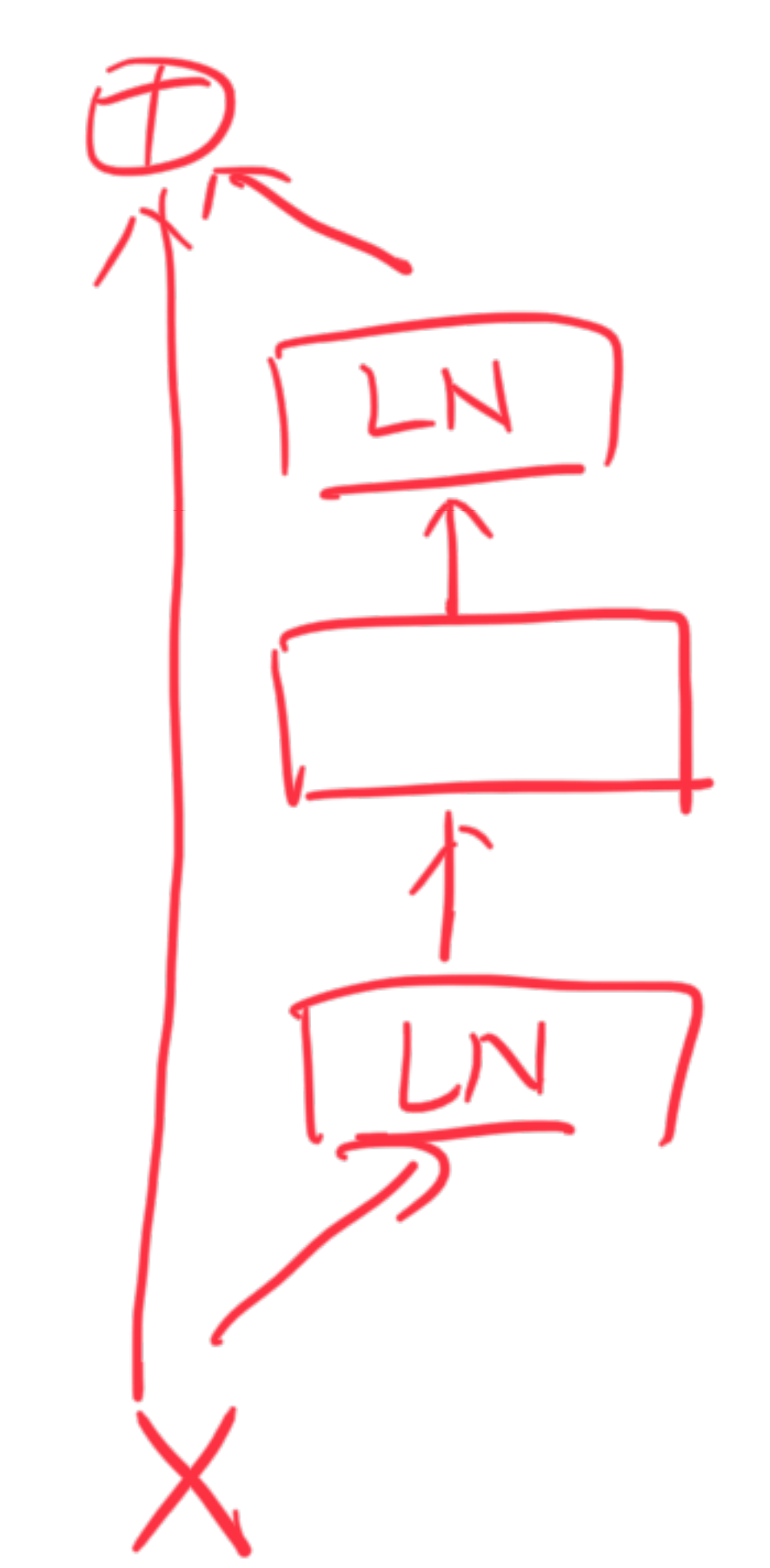
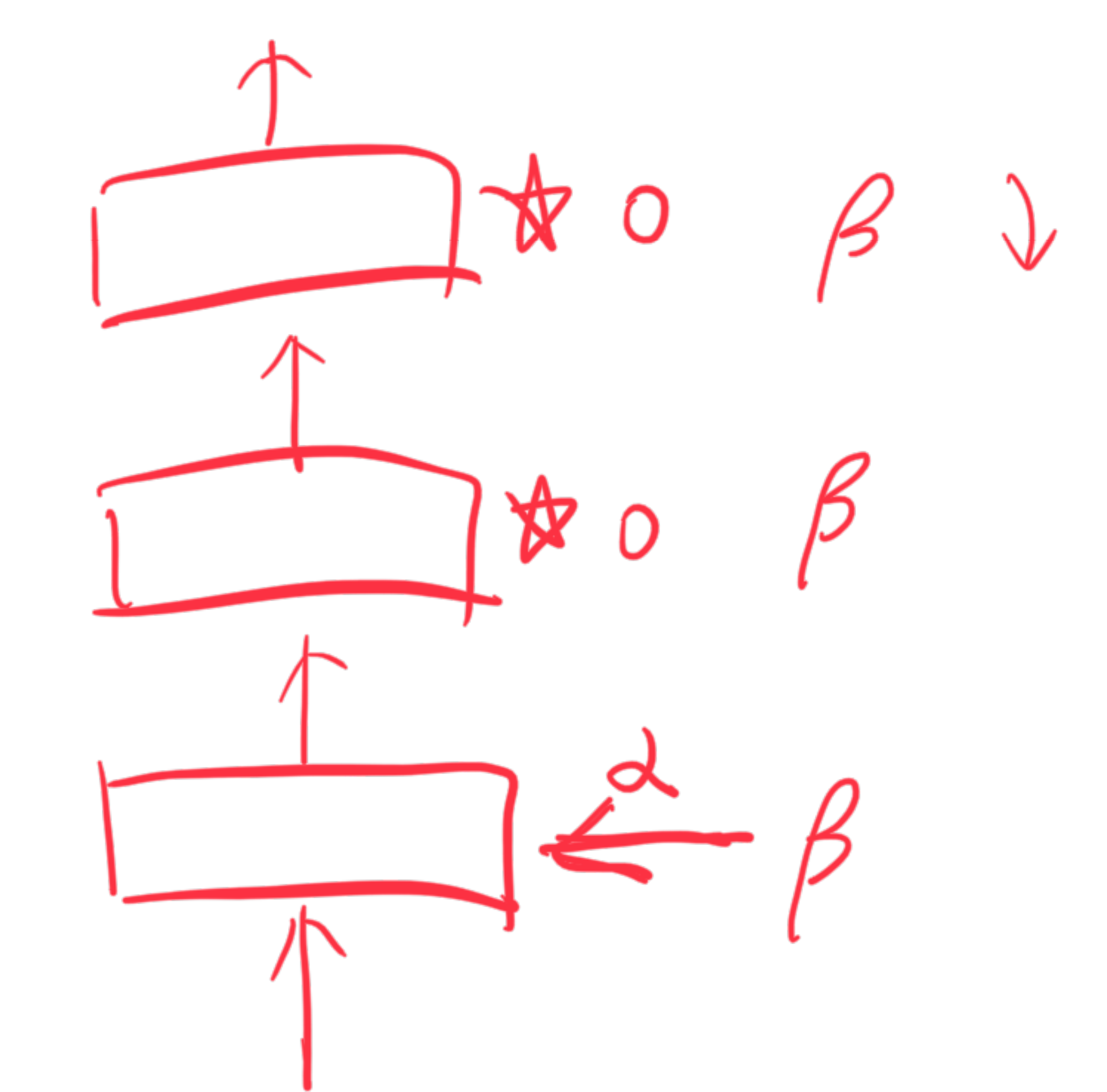
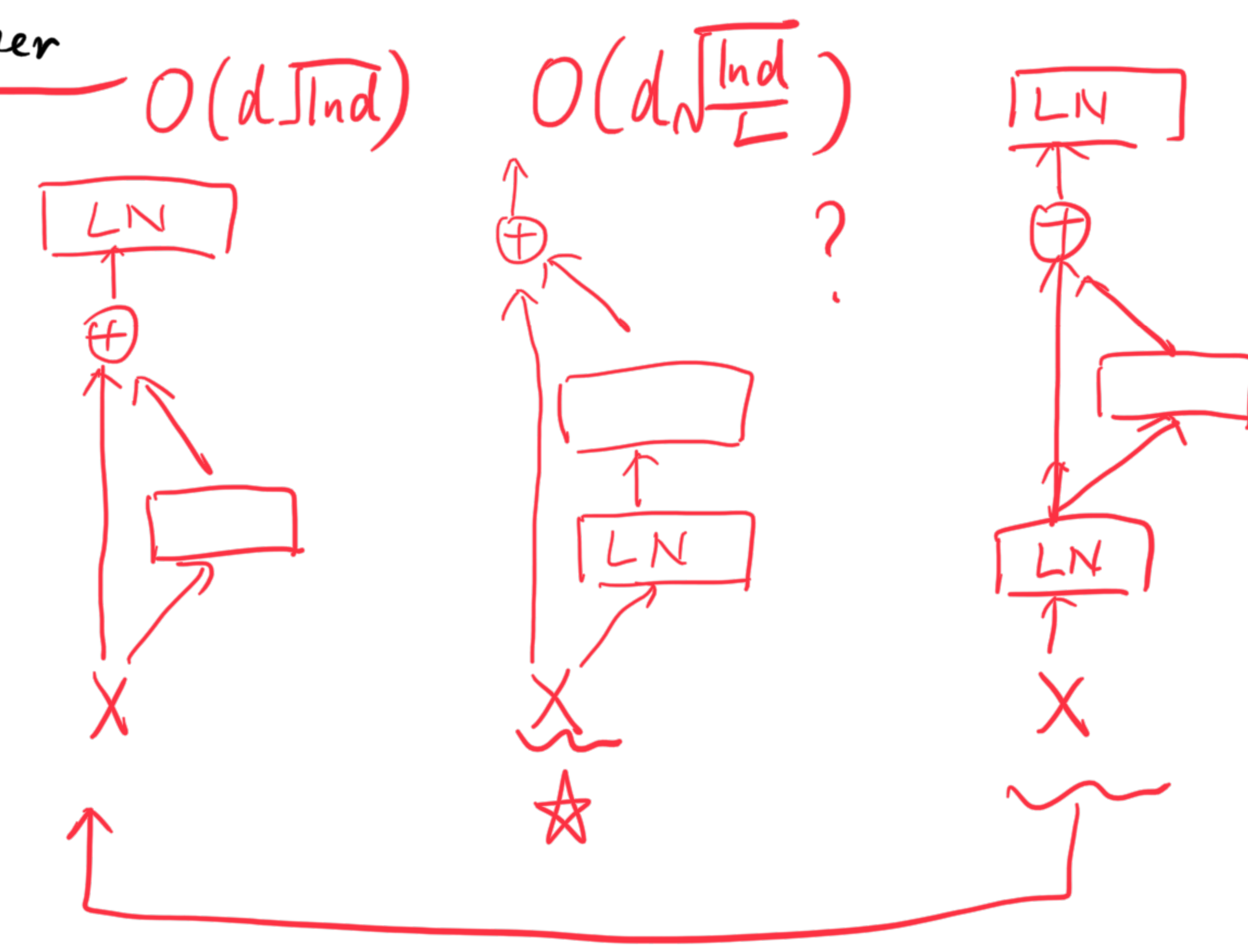
Transformer Evolution

X { Position Encoding
Image Patch

Attention → Efficient Transformer

LayerNorm { Post-LN
Pre-LN
Sandwich-LN

FFN { MoE
MoE-Transformer / Switch-Transformer
Fast MoE



标题	年
Switch Transformers: Scaling to Trillion Parameter Models...	2021
Outrageously Large Neural Networks: The Sparsely-Gated M...	2017
GShard: Scaling Giant Models with Conditional Computati...	2020
On Layer Normalization in the Transformer Architecture	2020
FastMoE: A Fast Mixture-of-Expert Training System	2021
CogView: Mastering Text-to-Image Generation via Tran...	2021