Mixture of Depths

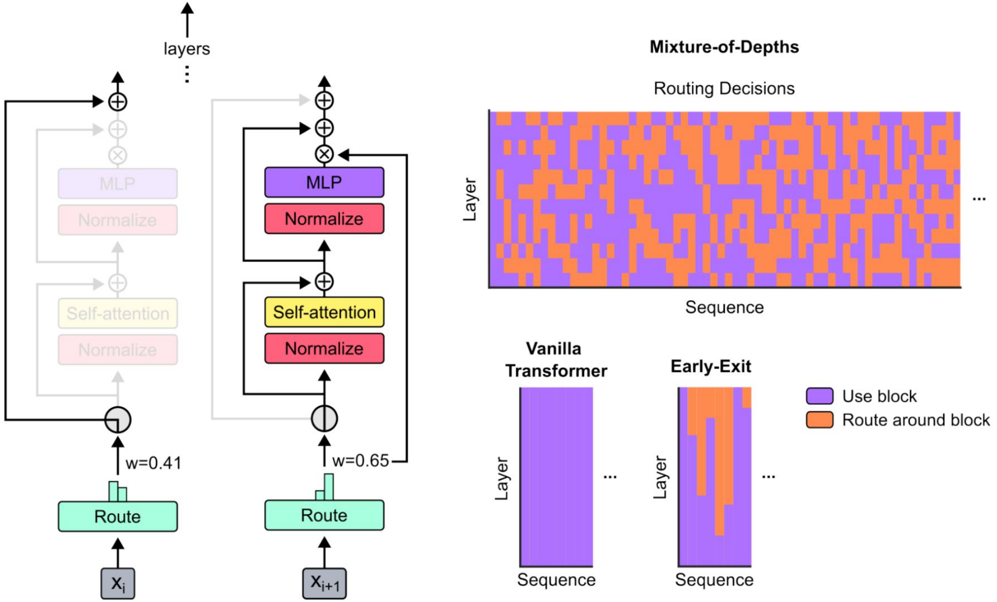
SDML
Ryan Chesler

Problem

- A lot of compute is wasted on transformers attending to all tokens all the time even for simple things
- Would be powerful to find some scheme that can variably allocate more or less capacity to tokens

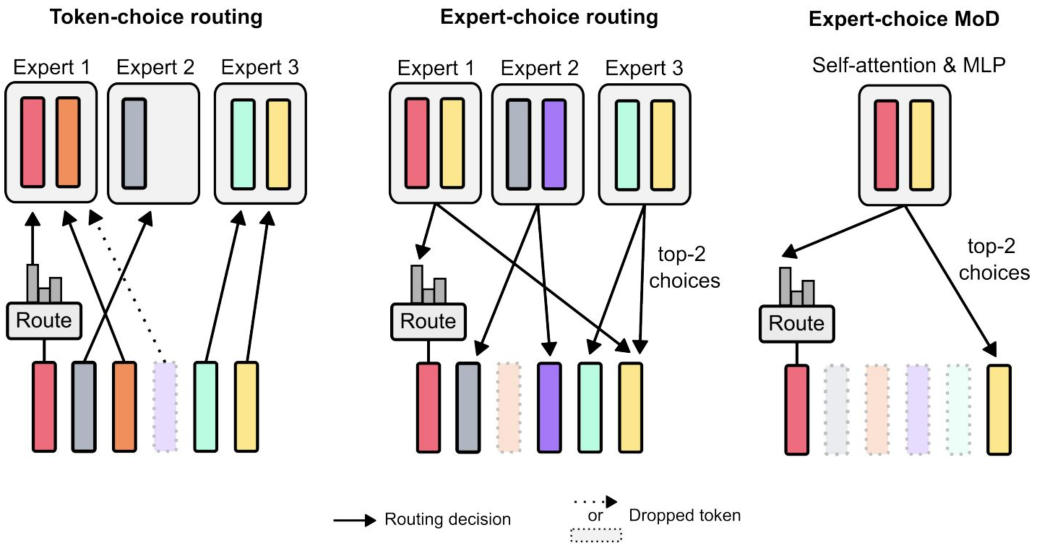
Mixture of Depths

- Taking inspiration from mixture of experts, we can route tokens so not all of them are attended to all the time
- This can potentially save a ton of compute



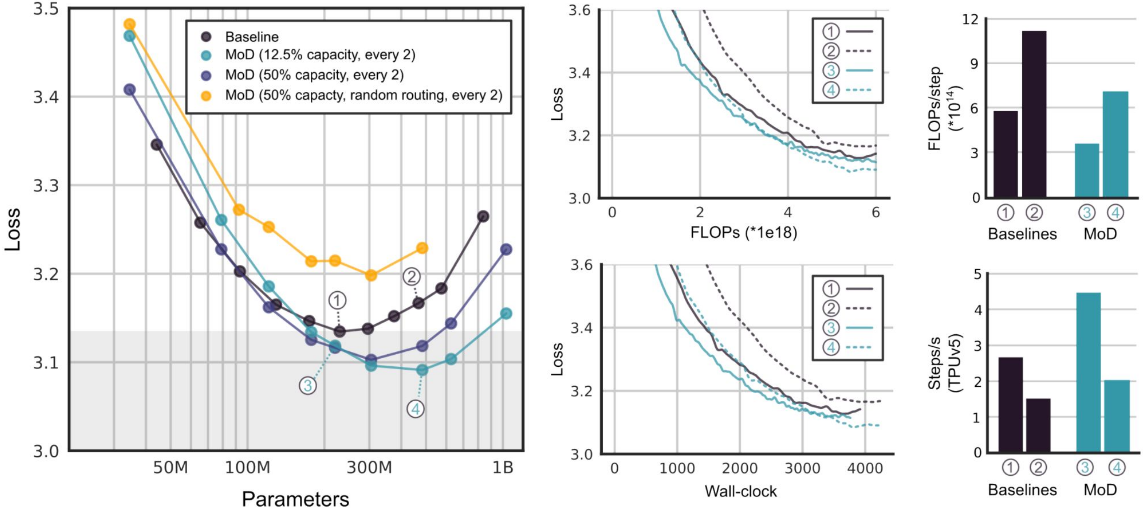
Routing Schemes

- Token-choice routing
  - Each token picks its own path
- Expert-choice routing
  - Each expert picks top-k tokens to attend to
- Expert-choice MoD
  - A single top-k choice is made, eliminating a fixed proportion of tokens



Results

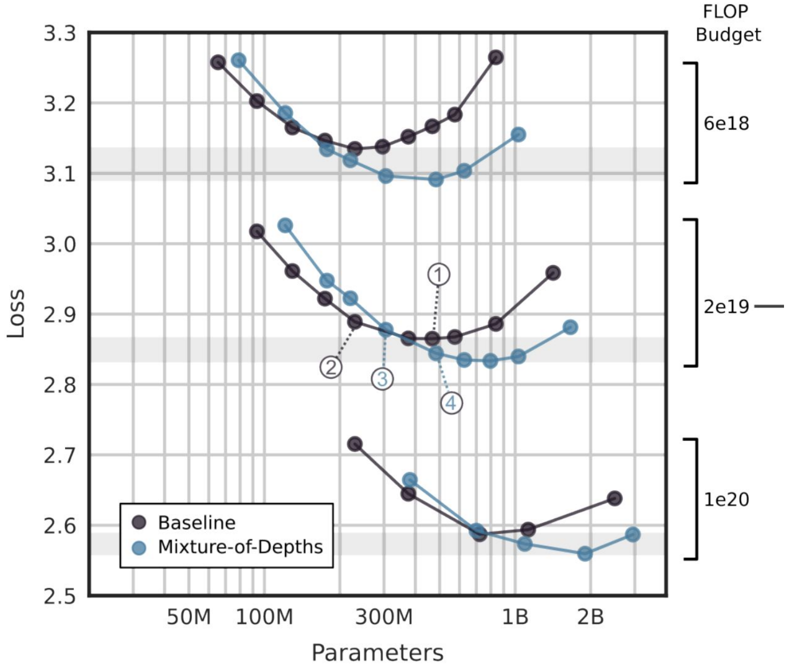
- Better performance per FLOPs
- Has regions that beat compute optimal vanilla transformers



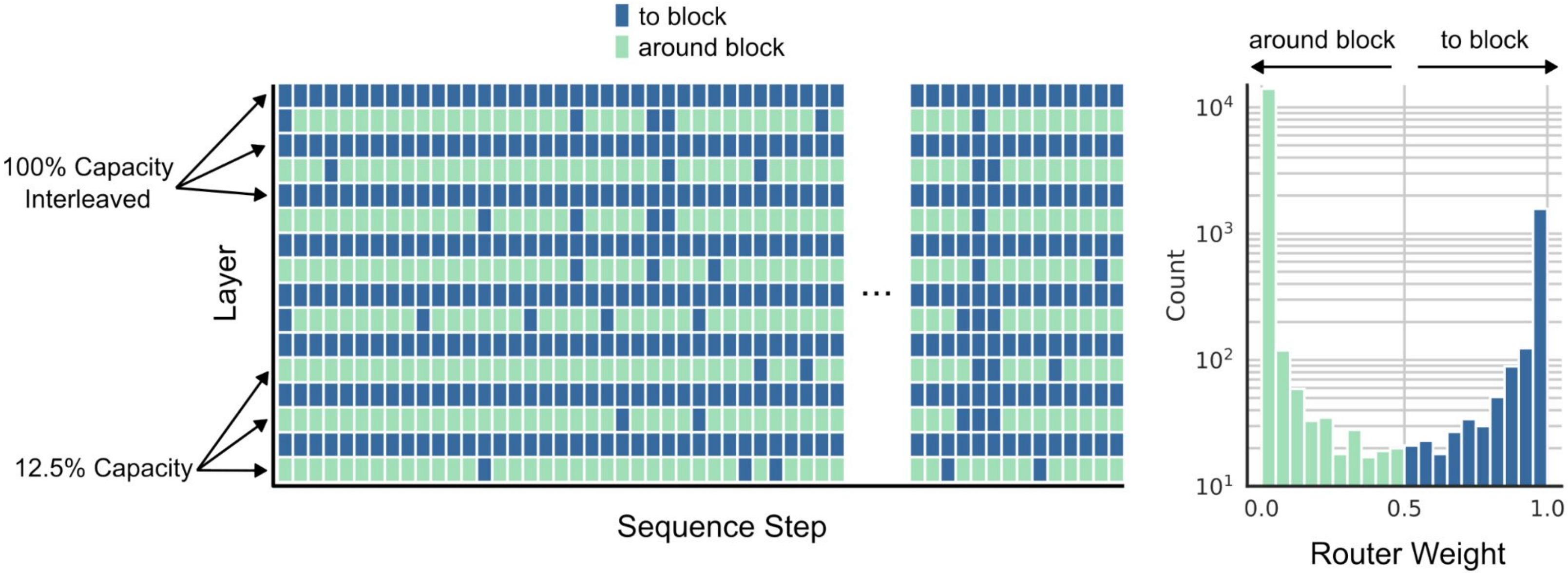
Results

Compute optimal comparison

Allows more training steps to be taken given a certain compute budget and scale to larger models

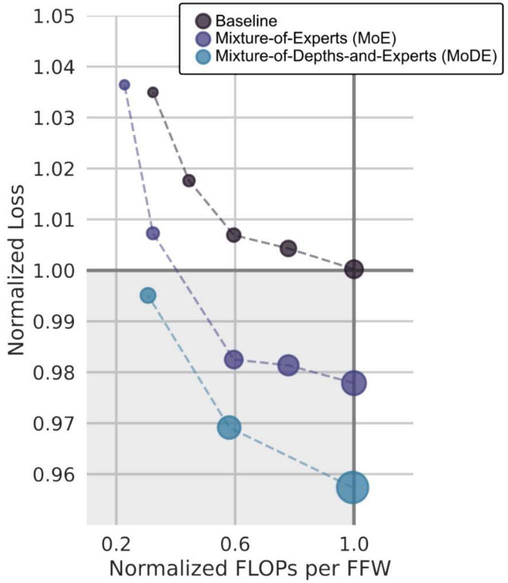


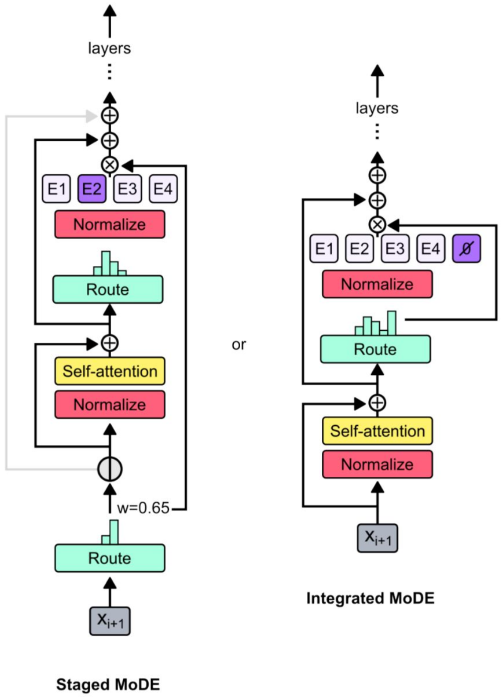
Routing pattern



Combining with Mixture of Experts(MoDE)

Able to combine both mixture of experts and mixture of depth together to get even better results





Staged MoDE

Integrated MoDE