CMSC828G: Special Topics in Systems for Machine Learning

Mixture-of-Experts Inference

Instructor: Yafan Huang

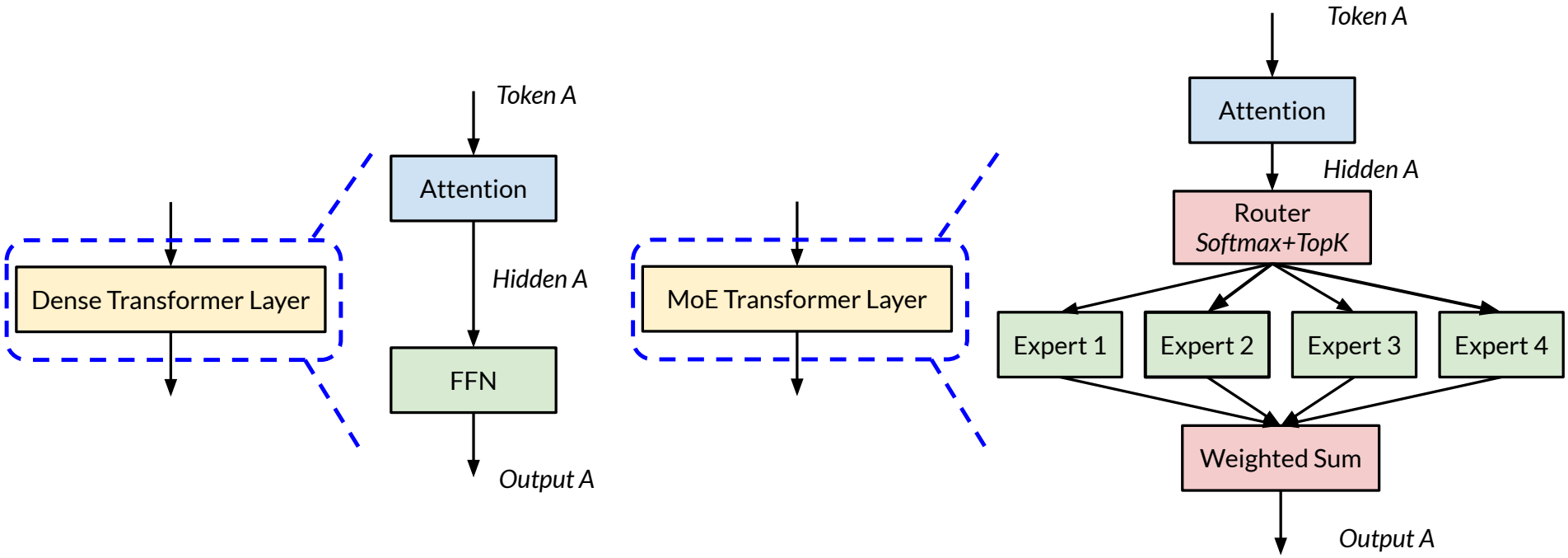


DEPARTMENT OF
COMPUTER SCIENCE

Mixture-of-Experts (MoE) Transformer

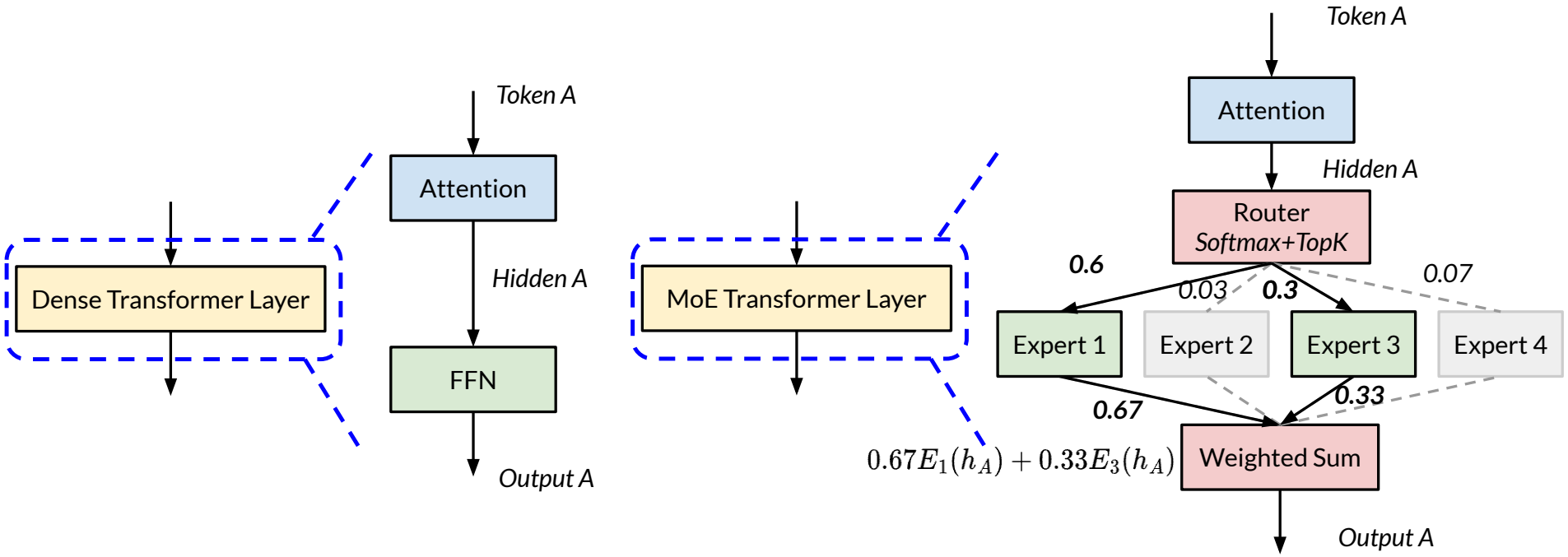
- MoE layer consists of multiple feed-forward networks (FFNs), called experts.
    - Multiple FFNs within a transformer layer share the same attention.
- A learned router selects a subset of experts for each input token (top-k routing).
- The selected expert outputs are combined into a weighted sum.
- Sparse computation: only a subset of the experts is activated per token.

MoE Forward Pass: Single Token



**Omit residual connection and layer normalization for simplicity.*

MoE Forward Pass: Single Token



**Omit residual connection and layer normalization for simplicity.*

MoE Forward Pass: Multiple Tokens

- Each token independently selects its top-k experts.
- Tokens assigned to the same expert are grouped for computation.
- Different experts may receive different numbers of tokens, resulting in imbalanced workloads.

Token Routing

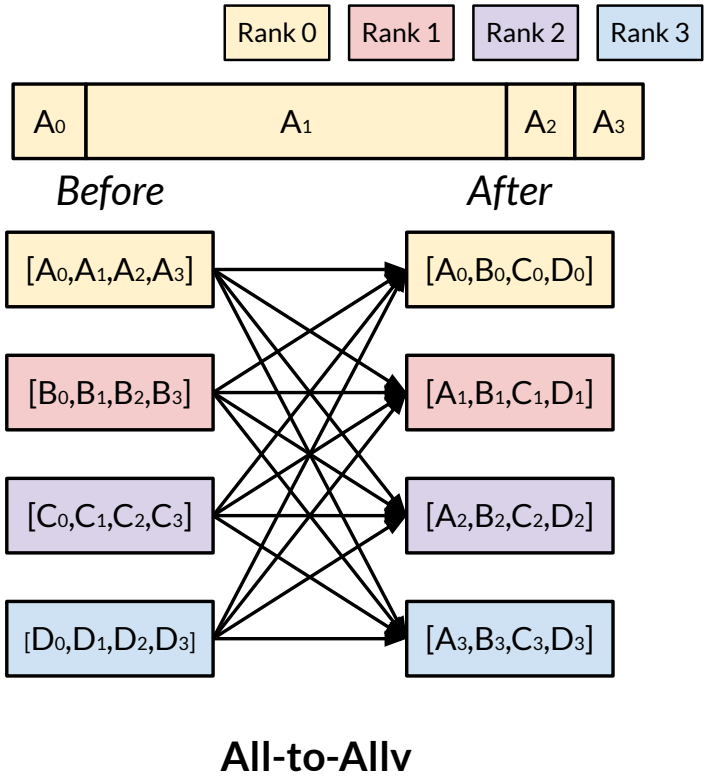
A	-> E1, E3
B	-> E2, E3
C	-> E1, E4
D	-> E3, E4

Grouped by Expert

E1	A, C
E2	B
E3	A, B, D
E4	C, D

Expert Parallelism: MoE across GPUs

- Expert Parallelism (EP) distributes expert weights across multiple GPUs.
- Token activations are dispatched to their selected experts and returned for aggregation.
- **All-to-Allv**-style communication supports variable-sized token transfers between GPUs.
- EP improves model scalability but introduces communication overhead and load imbalance.



Elastic Expert Parallelism

- Traditional EP: Fixed GPU allocation and expert placement.
- **Elastic EP**: Dynamically scales the number of GPUs. Example: vLLM and SGLang.
- Reconfiguration: Redistributes experts and updates communication groups.
    - Multiple experts can be placed on a single GPU.
    - Popular experts can be replicated across multiple GPUs for load balancing.
- Challenge is to minimize schedule overhead while maintaining serving availability.

System Research Directions in MoE

- GPU Efficiency: Exploit GPU computation capability for small expert workloads.
- Load Balancing: How to dynamically place and replicate experts?
- Communication: Minimize/overlap transfer and deal with load imbalance.
- Elasticity: How to scale GPU resources with changing workloads?
- Disaggregation: How to optimize attention and expert deployment?
- Fault Tolerance: Recover from GPU failures without service interruption.

Student Presentations

- [SC’22] DeepSpeed Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale
- [SIGCOMM’25] MegaScale-Infer: Efficient Mixture-of-Experts Model Serving with Disaggregated Expert Parallelism