Gating Dropout: Communication-efficient Regularization for Sparsely Activated Transformers

Rui Liu[*1], Young Jin Kim[2], Alexandre Muzio[2], Hany Hassan Awadalla[2]

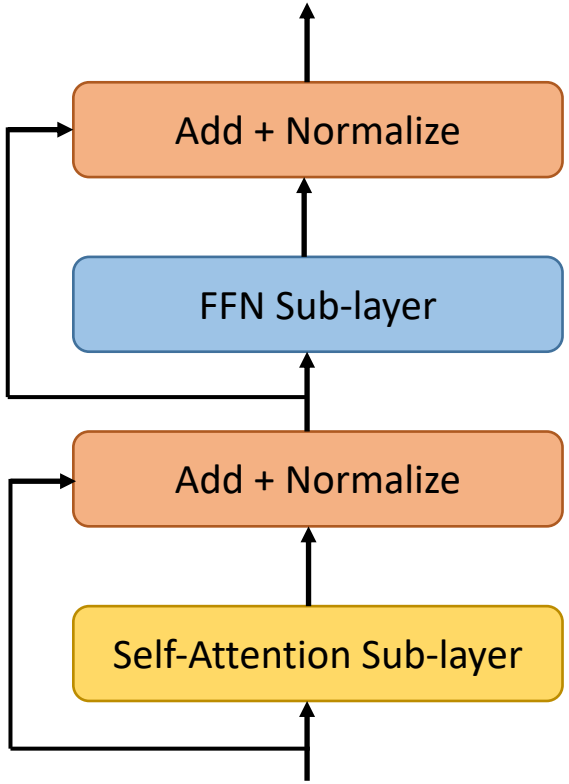
[1] University of Michigan, Ann Arbor

[2] Microsoft

[*]Work done while a research intern at Microsoft

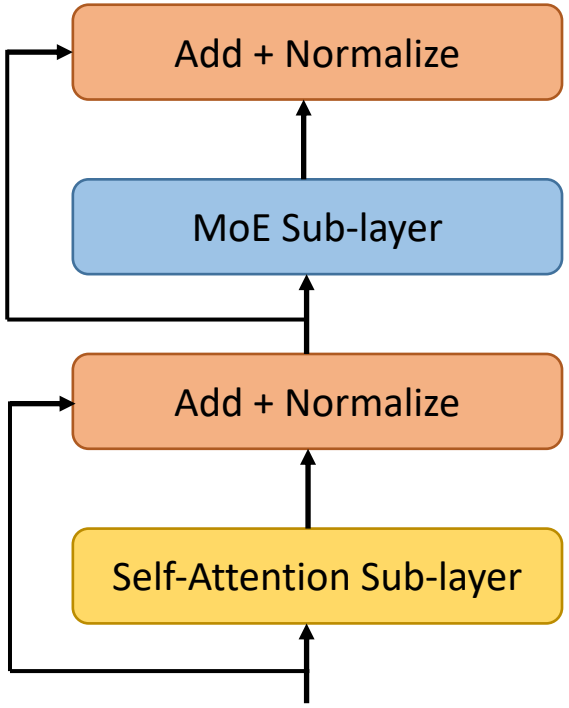
ICML 2022

Background: MoE Transformer



Dense transformer [1,2]

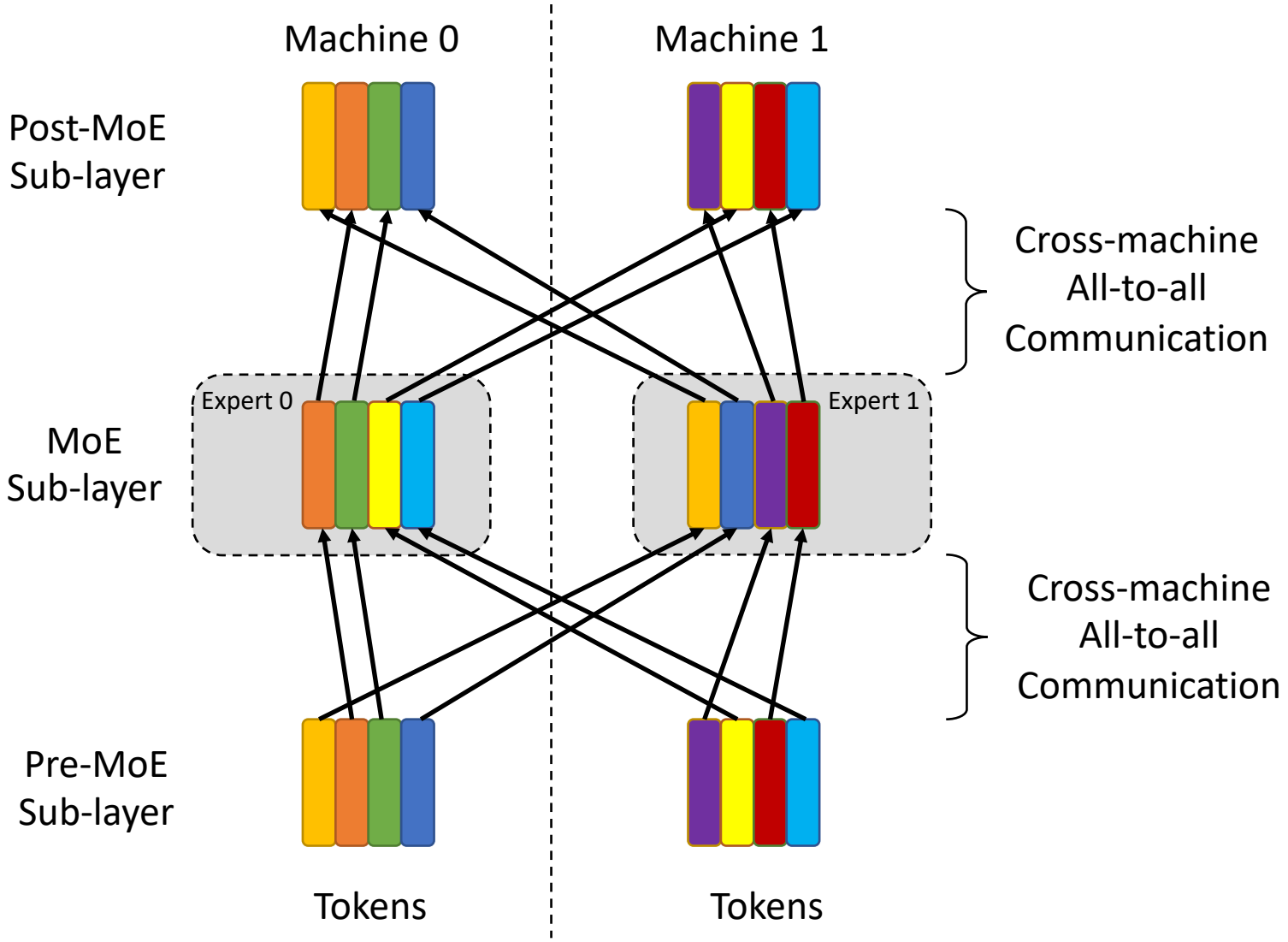
replace the FFN sublayer with MoE sub-layer (*Mixture of Experts*, i.e., a set of FFN sublayers residing at different machines)

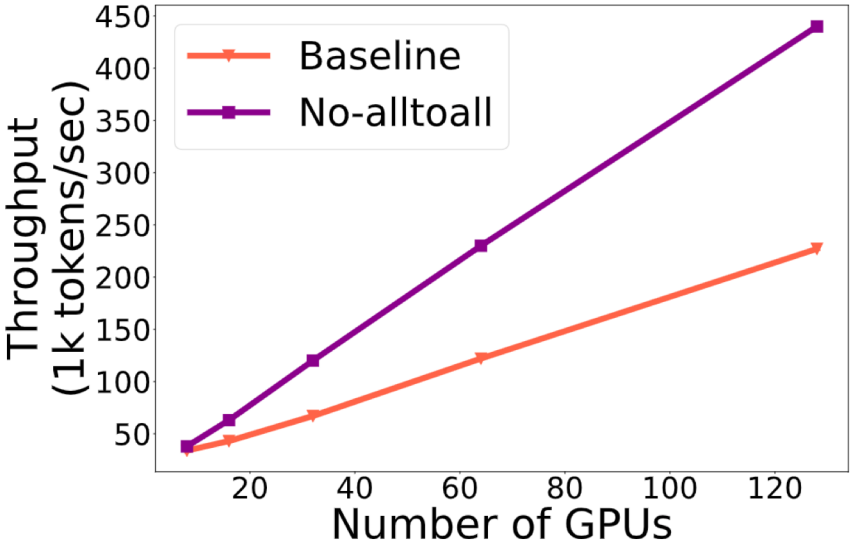


MoE Transformer [3,4]

[1] Vaswani, Ashish, et al. "Attention is all you need." NeurIPS (2017).
[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv (2018).
[3] Fedus, William, Barret Zoph, and Noam Shazeer. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity." *arXiv* (2021).
[4] Kim, Young Jin, et al. "Scalable and efficient moe training for multitask multilingual models." arXiv (2021).

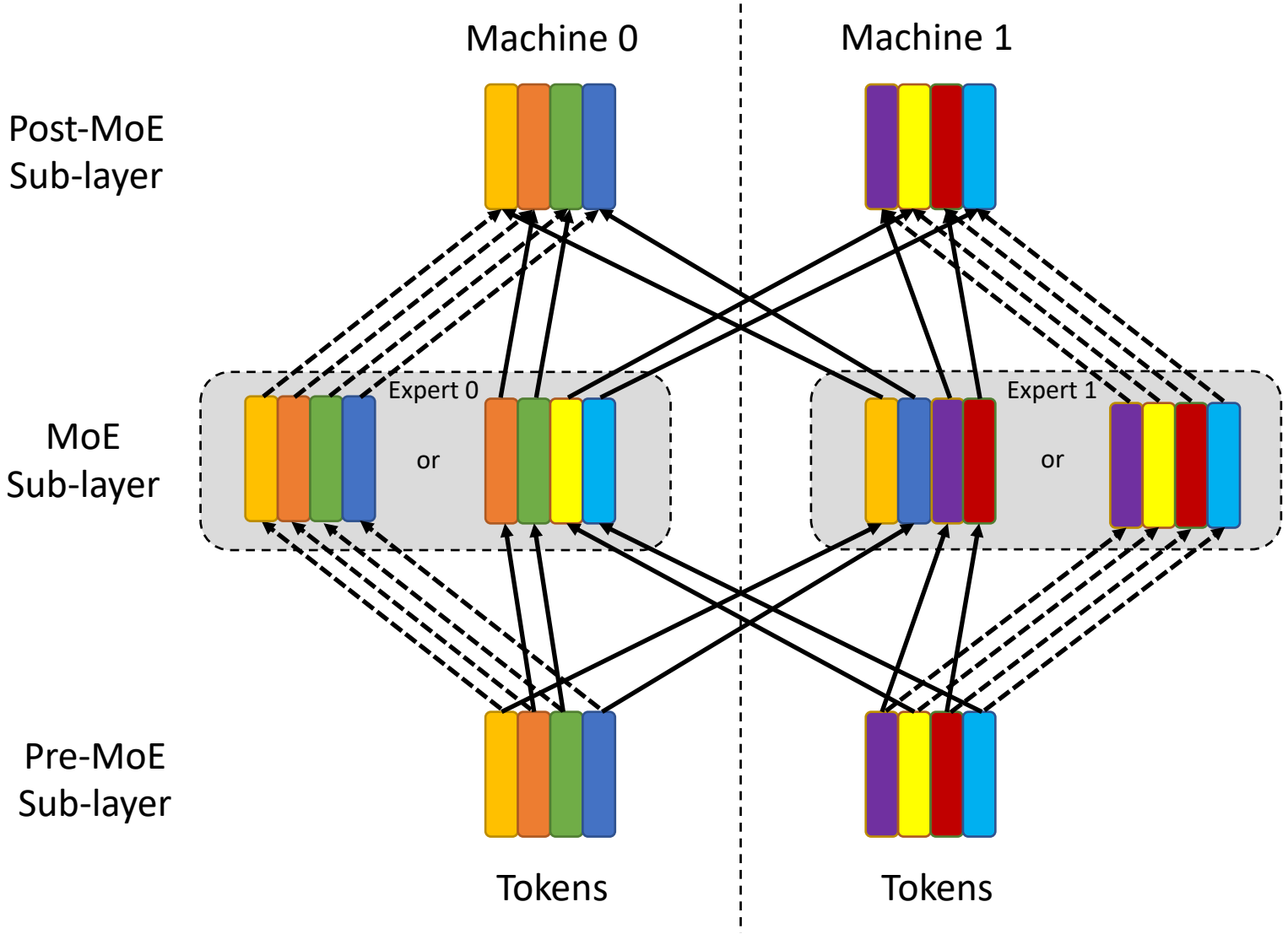
Communication Cost in MoE Transformer





Significant throughput improvement when removing all-to-all operations

Gating Dropout: Communication-efficient Regularization



Happen with probability p (Without all-to-all Communication)

Happen with probability $1-p$ (With all-to-all Communication)

Communication Efficiency:
reduce communication cost by skipping all-to-all operations

Regularization Effect:
encourage learning a general ability as well as exploring other experts without relying on the gating network

Experimental Setup

Datasets (multilingual translation)

- WMT-10: contains 32.5 million parallel sentences in 10 languages
- Web-50: contains 700 million parallel sentences in 50 languages

Models

- WMT-10: Transformer-base architecture **[1]** (5.6 billion parameters)
- Web-50: Transformer-big architecture **[1]** (10 billion parameters)

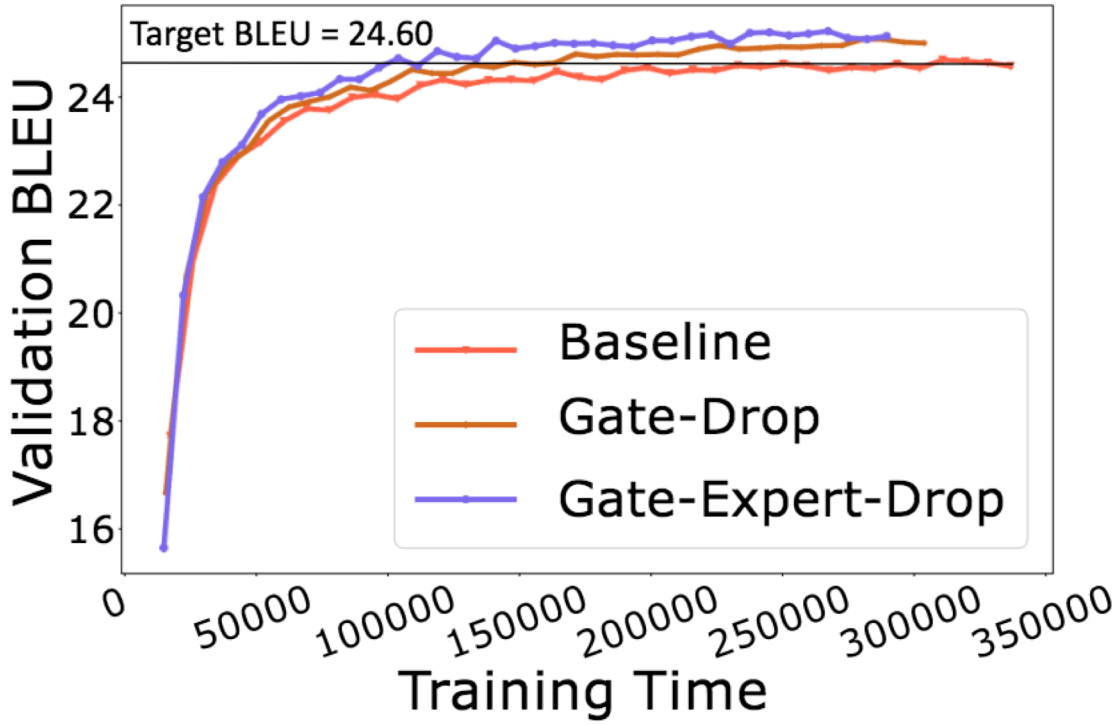
Methods

- Baseline: Z-code M3 **[2]**
- Our methods: Gate-Drop and Gate-Expert-Drop

[1] Vaswani, Ashish, et al. "Attention is all you need." NeurIPS (2017).
[2] Kim, Young Jin, et al. "Scalable and efficient moe training for multitask multilingual models." arXiv (2021).

Experimental Results

WMT-10



Note:
E -> X means translation from English to other languages.
Low means only low-resource languages are considered.

Web-50

Method	BLEU (avg)	E→X	E→X (low)	X→E	X→E (low)
Baseline	28.63	23.01	22.15	34.26	33.89
Gate-Drop	**29.22**	**23.86**	**22.87**	**34.59**	**34.34**
Gate-Expert-Drop	28.85	23.22	22.61	34.49	34.22

Conclusion

- Propose a new variant of dropout Gating Dropout for training MoE transformers
- Benefits: Communication efficiency and Regularization effect
- Demonstrate faster wall-clock convergence speed and higher converged BLEU scores on multilingual translation datasets

For complete details on this work, please refer to our paper

Rui Liu, Young Jin Kim, Alexandre Muzio and Hany Hassan Awadalla. **Gating Dropout: Communication-efficient Regularization for Sparsely Activated Transformers**, *ICML* 2022

University of Michigan

Microsoft