µParametrization for Mixture of Experts

Jan Małaśnicki*, Kamil Ciebiera*, Mateusz Boruń,
Maciej Pióro, Jan Ludziejewski, Maciej Pióro, Maciej Stefaniak, Michał Krutul,
Sebastian Jaszczur, Marek Cygan, Kamil Adamczewski, Jakub Krajewski

IDEAS NCBR
UNIVE OF WA
Wroclaw University of Science and Technology

TL;DR

This work explores MoE parametrizations which allow for LR transfer across model widths. We derive µP for MoE, parameterization with theoretical guarantees on feature learning in the width limit. We also test the more straightforward SimpleP parametrization, which also enables LR transfer.

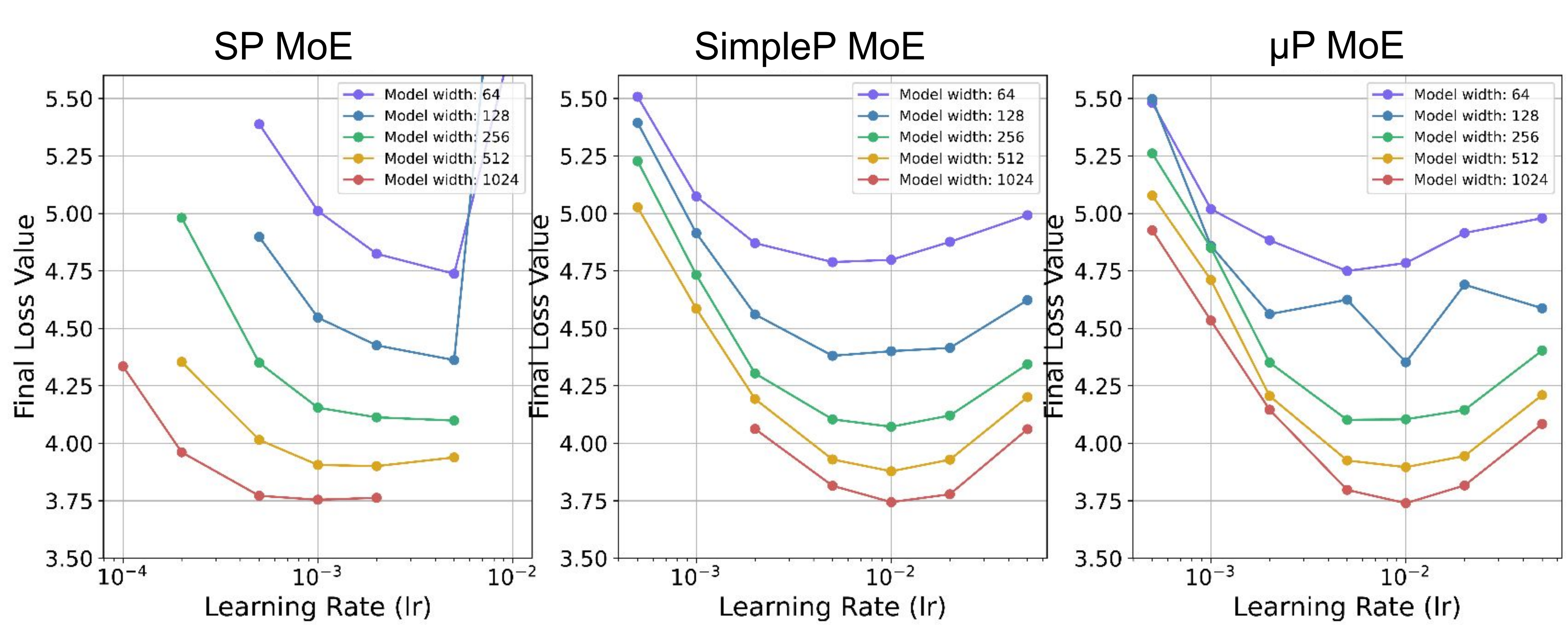
How to parametrize the model?

	Embedding	Unembedding	Attention (Q, K, V, O)	Feed-forward (dense)	Experts (MoE)	Router (MoE)
Init. Var.	1.0	1.0	1/fan_in	1/fan_in	1/fan_in	1/fan_in | 1.0
Multiplier	1.0	1/fan_in	1.0	1.0	1.0	1.0 | 1/fan_in
LR (Adam)	1.0	1.0	1/fan_in	1/fan_in	1/fan_in | 1/fan_in	1.0

µP (dense, TP5) | SimpleP (MoE) | µP MoE

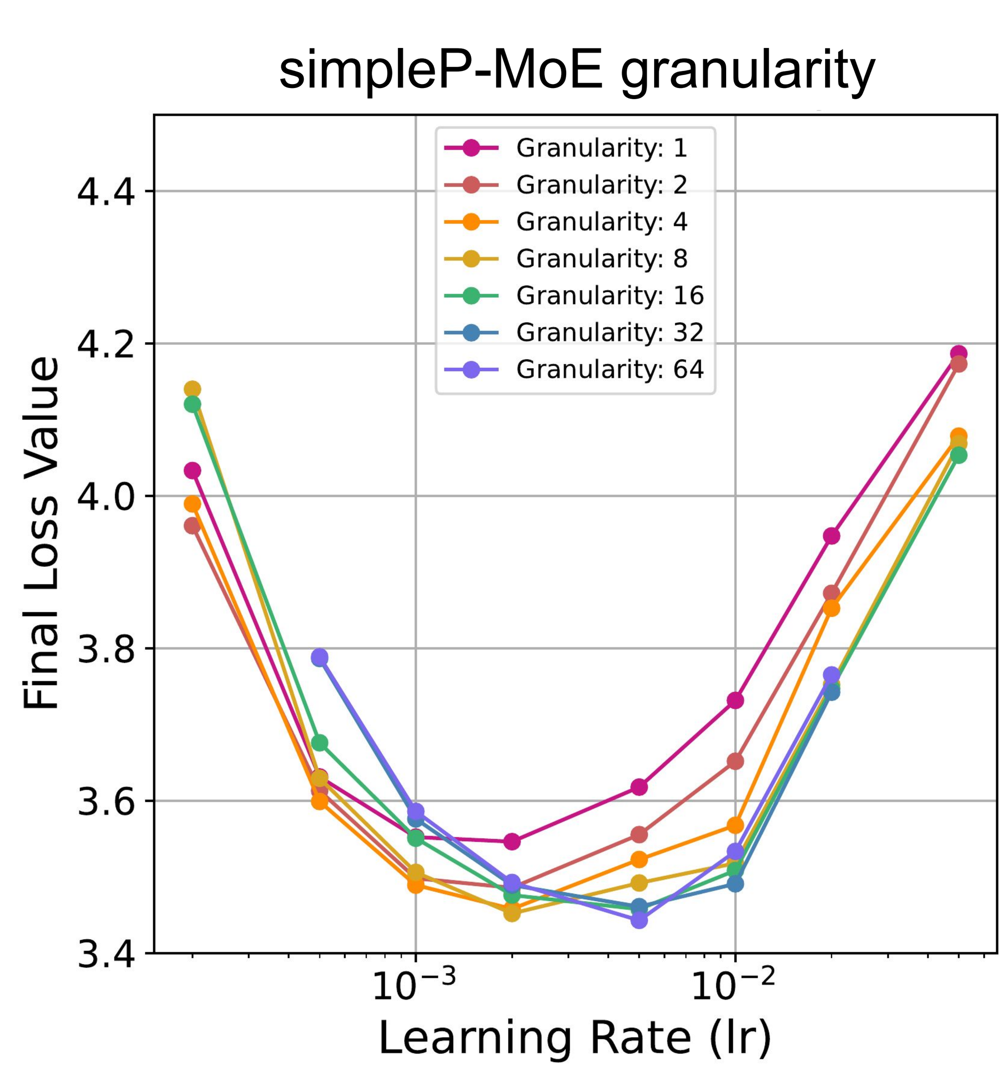
- Setting: model width scaling. Constant number of experts, layers, *top-k*, etc.
- SimpleP MoE: The most straightforward application of dense µP to MoE (no theory)
- µP MoE:
  - Theoretically derived
  - Intuition: Treat router weight like output weight

LR transfer in different MoE parametrizations

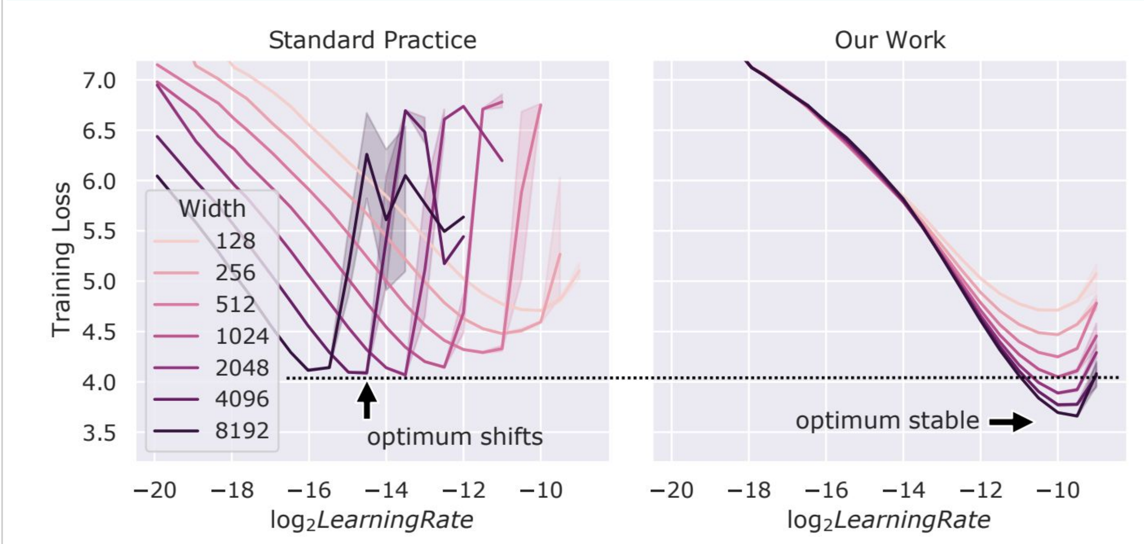


SimpleP MoE ablations

- LR transfer is achieved by both methods when scaling model width and number of experts
- Let's try to find where simpleP MoE fails
- Granularity: increases the number of experts and *top-k* while keeping the computations and memory constant (more, smaller experts + higher *top-k*)
- Number of experts: over varying number of experts the optimal LR is almost stable.



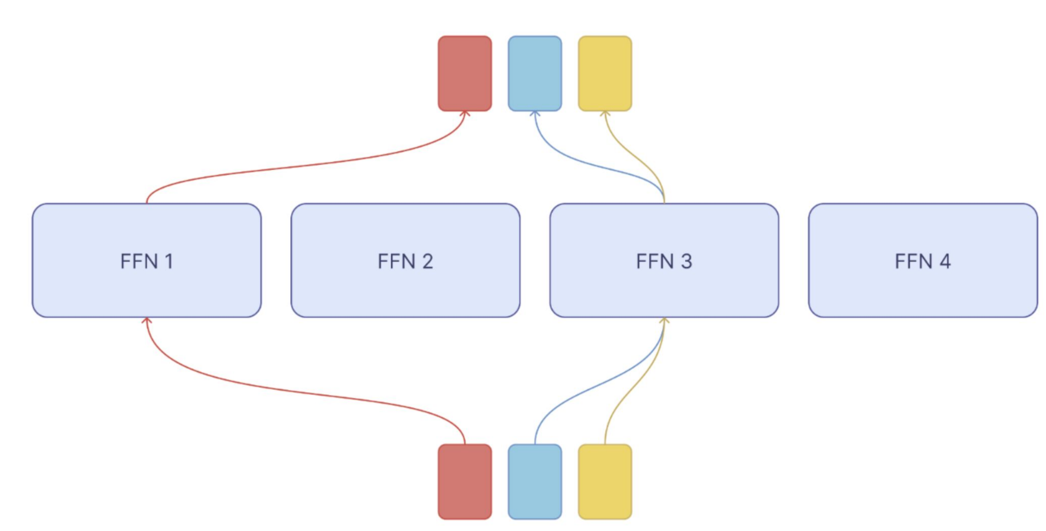
What is µP?



Tensor Programs 5 (Yang et. al. 2022)

µP MoE math intuition

- TP5 has 3 distinct weight types:
  - input weight - fixed to scalable
  - hidden weight - scalable to scalable
  - output weight - scalable to fixed
- Intuition:
  - Experts are hidden weights
  - Router is output weight
- This is just a guiding heuristic, but we do formal analysis following TP5 and get these results



Conclusions

- MoE is µ-parametrizable
- Both simpleP and muP for MoE achieve LR transfer
- In the paper we have the derivation of µP for MoE
- Future work on scaling *top-k* number of experts and expert size

github.com/llm-random/llm-random

Acknowledgements: This work was funded by IDEAS NCBR. NCN grant 2022/45/N/ST6/02222, NCBR grant POIR.01.01.01-00-0392/17-00, ERC Starting Grant TOTAL.
Computational resources were provided by: PL-Grid (grant PLG/2024/017060), University of Warsaw, IDEAS NCBR, and Writer.