香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

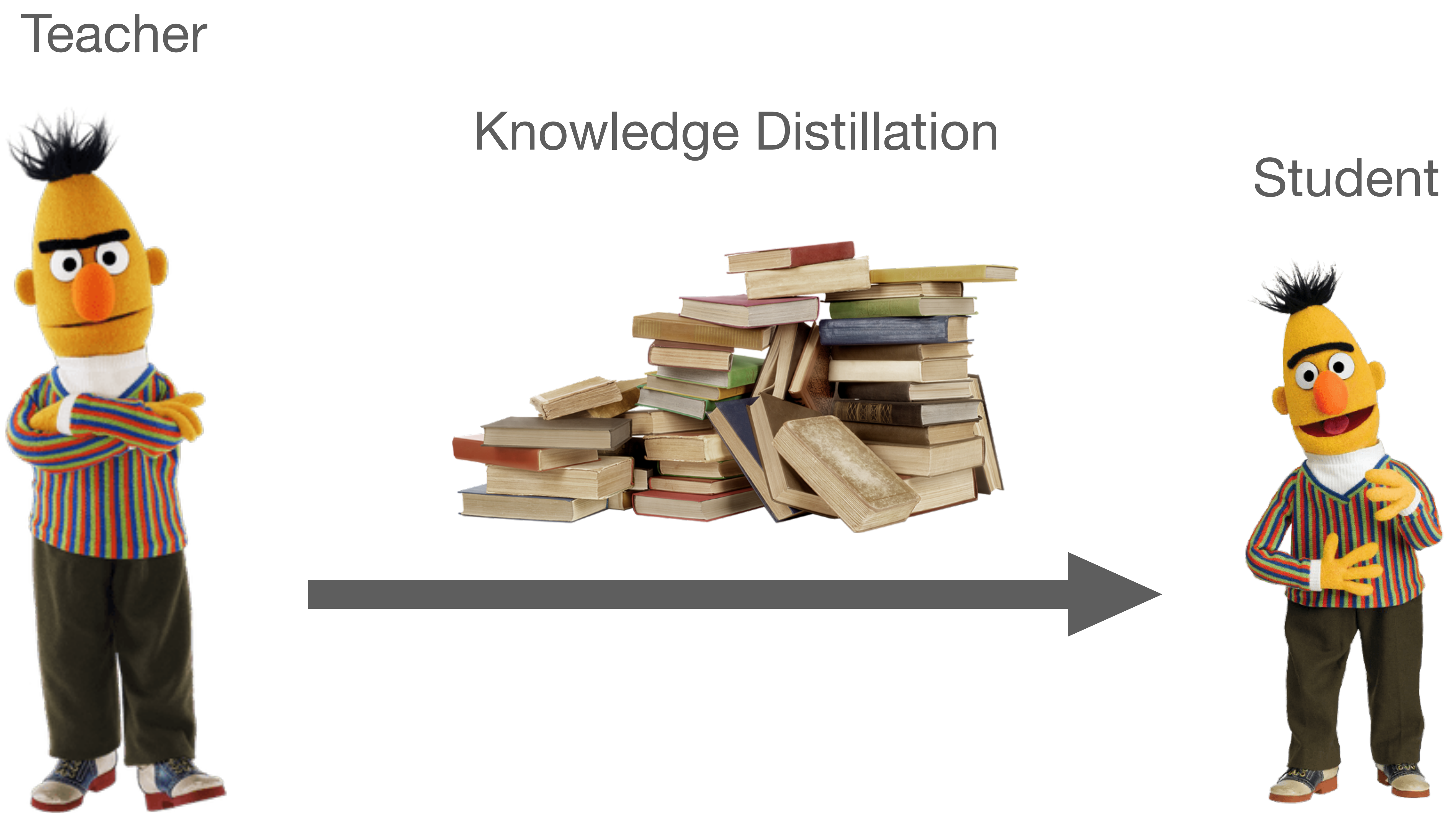
Lifting the Curse of Capacity Gap in Distilling Language Models

Chen Zhang, Yang Yang, Jiahao Liu, Jingang Wang, Yunsen Xian, Benyou Wang, Dawei Song

Language Model Distillation

Teacher-student Paradigm

- Language model (LM) distillation aims at reducing inference compute by distilling the large LM into a small LM under a teacher-student paradigm.



Language Model Distillation

Existing Methods

- Task-specific distillation with finetuning data (e.g., MRPC).
  - KD (Hinton, et al.)
  - MiniDisc (Zhang, et al.)
- Task-agnostic distillation with pretraining data (e.g., Wikipedia).
  - MiniLM (Wang, et al.)
  - TinyBERT (Jiao, et al.)
- Task-agnostic distillation is commonly viewed as a better choice.

Curse of Capacity Gap

Theoretical Intuition

- The curse of capacity gap is not new, but has already been recognized in vision community as \`large teacher, poor student\` for *task-specific* vision model distillation. We leave a minor theoretical justification says increasing teacher capacity introduces a tradeoff between errors of the teacher ($\epsilon_{\mathscr{T}}$) and the capacity gap ($\epsilon_{\mathscr{G}}$).

Proposition 1 (VC dimension theory, Vapnik, 1998)**.** *Assuming that the teacher function is $f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}$, the labeling function is $f \in \mathcal{F}$, and the data is $\mathcal{D}$, we have:*

$$r(f_{\mathcal{T}}) - r(f) \leq \epsilon_{\mathcal{T}} + o(\frac{|\mathcal{F}_{\mathcal{T}}|_c}{|\mathcal{D}|}),$$

where $r(\cdot)$ is the risk function, $|\cdot|_c$ is the function class capacity measure, and $|\cdot|$ is the data scale measure. It should be highlighted that the approximation error $\epsilon_{\mathcal{T}}$ is negatively correlated with the capacity of the teacher model while the estimation error $o(\cdot)$ is correlated with the learning optimization.

Proposition 2 (Generalized distillation theory, Lopez-Paz et al., 2016)**.** *Additionally providing that the student function is $f_{\mathcal{S}} \in \mathcal{F}_{\mathcal{S}}$, we have:*

$$r(f_{\mathcal{S}}) - r(f_{\mathcal{T}}) \leq \epsilon_{\mathcal{G}} + o(\frac{|\mathcal{F}_{\mathcal{S}}|_c}{|\mathcal{D}|^{\alpha}}),$$

where the approximation error $\epsilon_{\mathcal{G}}$ is positively correlated with the capacity gap between the teacher and the student models, and $1/2 \leq \alpha \leq 1$ is a factor correlated to the learning rate.

Theorem 1. *The bound for the student function at a learning rate can be written as:*

$$\begin{aligned} r(f_{\mathcal{S}}) - r(f) &\leq \epsilon_{\mathcal{T}} + \epsilon_{\mathcal{G}} + o(\frac{|\mathcal{F}_{\mathcal{T}}|_c}{|\mathcal{D}|}) + o(\frac{|\mathcal{F}_{\mathcal{S}}|_c}{|\mathcal{D}|^{\alpha}}) \\ &\leq \epsilon_{\mathcal{T}} + \epsilon_{\mathcal{G}} + o(\frac{|\mathcal{F}_{\mathcal{T}}|_c + |\mathcal{F}_{\mathcal{S}}|_c}{|\mathcal{D}|^{\alpha}}), \end{aligned}$$

Proof. The proof is rather straightforward by combining Proposition 1 and 2. $\square$

Curse of Capacity Gap

Empirical Investigation

- Little work has systematically verified that the curse for both *task-specific and task-agnostic* LM distillation. We mainly focus on task-agnostic one. The curse indeed exists in LM distillation and existing methods cannot simply tackle the curse.

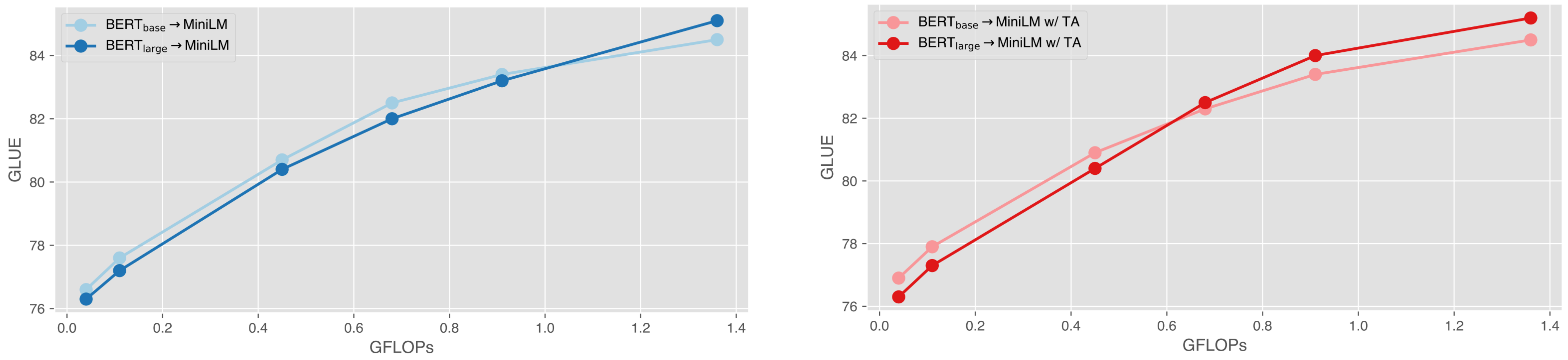
GLUE (Wang et al., 2019). The △ denotes the performance difference of preceding two numbers. To ensure students at similar scales, the student/teacher scale ratios are properly reduced for some methods.

Method	$\text{BERT}_{\text{base}}$	$\text{BERT}_{\text{large}}$	△
Teacher	86.7	88.3	+1.6
$\text{KD}_{10\%/5\%}$ (2015)	81.3	80.8	−0.5
$\text{DynaBERT}_{15\%/5\%}$ (2020)	81.1	79.2	−1.9
$\text{MiniDisc}_{10\%/5\%}$ (2022a)	82.4	82.1	−0.3
$\text{TinyBERT}_{4L;312H}$ (2020)	82.7	82.5	−0.2
$\text{MiniLM}_{3L;384H}$ (2021b)	82.5	82.0	−0.5
$\text{MiniMoE}_{3L;384H}$ (**ours**)	82.6	83.1	+0.5

Curse of Capacity Gap

Empirical Investigation

- Little work has systematically verified that the curse for both *task-specific and task-agnostic* LM distillation. We mainly focus on task-agnostic one. The curse indeed exists in LM distillation and existing methods cannot simply tackle the curse.



Curse Lifting

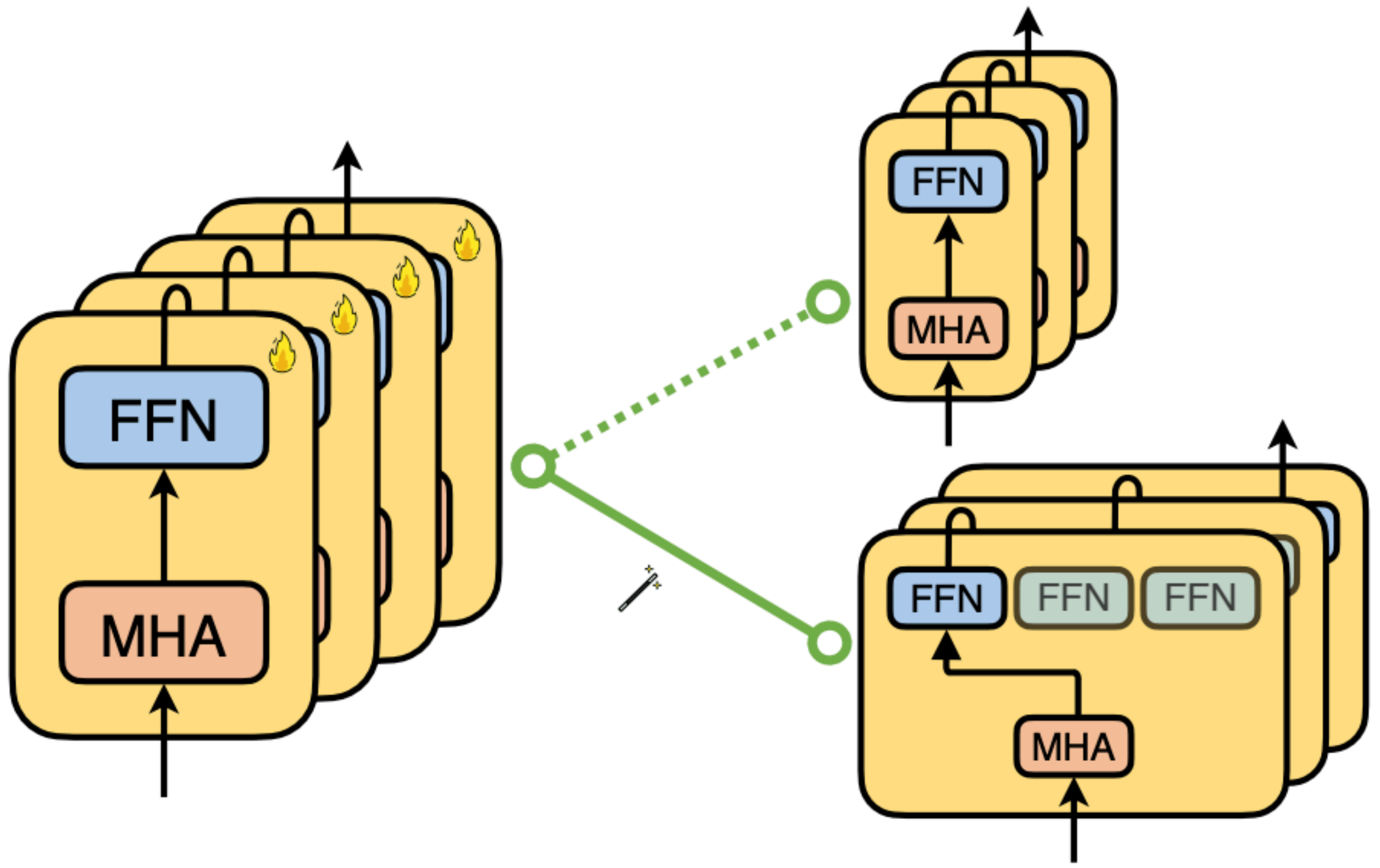
Potential Solutions

- An intuitive guideline is enlarging student capacity without increasing inference compute.
- Quantized student, enlarging student capacity with lower precision, yet not distillation-friendly.
- Depth-adaptive student, enlarging student capacity with adaptive depths (e.g., early exiting), yet not in constant compute.
- *Mixture of experts* (MoE) student, enlarging student capacity with sparse experts.

Curse Lifting

MiniMoE

- We incorporate the merits of MoE in the design of the distillation.
- We start from a task-agnostic distillation baseline MiniLM and propose to replace the student in it with a MoE one (thus named *MiniMoE* for mixture of minimal experts).



Experiments

Setup

- Distillation on Wikipedia.
- Finetuning on GLUE (sequence and sequence-pair classification) and CoNLL (named entity recognition).
- BERT-base and BERT-large as teachers of different scales.
- All students default to 4 experts.

Experiments

Lifted Curse

- The curse is not lifted by MiniLM and MiniLM w/ TA until MiniMoE.

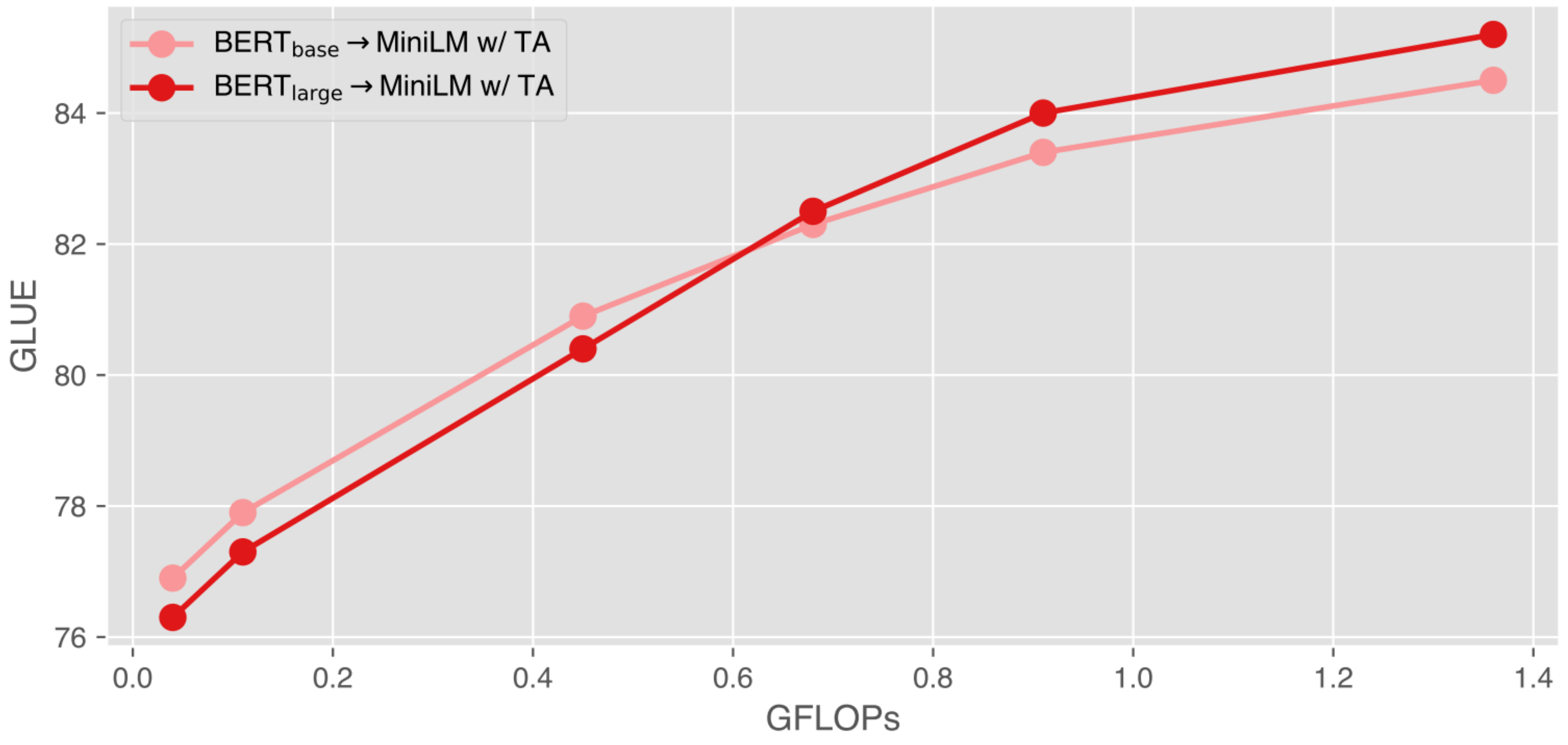
Method	Teacher	SST-2 Acc	MRPC F1	STS-B SpCorr	QQP F1	MNLI-m/mm Acc	QNLI Acc	RTE Acc	GLUE Score	CoNLL F1
MiniLM$_{6L;384H}$	BERT$_{base}$	91.1	90.1	88.1	86.7	81.5/81.8	89.2	67.9	84.5	93.2
	BERT$_{large}$ $\Uparrow$	90.9	90.6	89.0	86.9	81.8/82.4	88.8	70.0	85.1	93.2
w/ TA	BERT$_{base}$	91.3	90.3	88.2	86.8	81.4/81.6	89.7	66.8	84.5	93.2
	BERT$_{large}$ $\Uparrow$	91.4	89.8	88.5	87.0	81.9/81.6	89.5	71.5	85.2	93.2
MiniMoE$_{6L;384H}$	BERT$_{base}$	91.3	90.2	88.6	86.5	81.6/81.5	89.5	68.6	84.7	93.3
	BERT$_{large}$ $\Uparrow$[1]	90.5	90.0	88.8	86.8	81.8/82.2	90.8	70.4	85.2	93.3
MiniLM$_{4L;384H}$	BERT$_{base}$	90.0	88.6	87.2	86.1	80.0/80.3	87.9	67.2	83.4	91.5
	BERT$_{large}$ $\Downarrow$	89.3	87.5	88.1	85.9	79.9/80.2	87.6	67.2	83.2	91.2
w/ TA	BERT$_{base}$	90.0	88.5	87.3	86.3	80.1/80.7	88.0	66.4	83.4	91.8
	BERT$_{large}$ $\Uparrow$	90.6	88.7	88.1	86.3	80.5/80.7	87.9	69.0	84.0	92.2
MiniMoE$_{4L;384H}$	BERT$_{base}$	90.8	88.1	88.2	85.9	79.8/80.4	88.6	69.3	83.9	92.3
	BERT$_{large}$ $\Uparrow$	90.5	88.0	88.7	86.7	80.9/80.9	89.2	69.0	84.2	92.4
MiniLM$_{3L;384H}$	BERT$_{base}$	89.1	89.1	86.6	85.4	77.8/78.4	87.2	66.1	82.5	90.1
	BERT$_{large}$ $\Downarrow$	89.1	86.1	87.1	85.1	78.6/78.5	86.0	65.7	82.0	87.3
w/ TA	BERT$_{base}$	89.8	87.8	86.0	85.5	77.6/78.5	86.8	66.1	82.3	90.4
	BERT$_{large}$ $\Downarrow$	89.7	84.9	87.2	85.2	78.5/79.1	86.6	66.4	82.2	90.2
MiniMoE$_{3L;384H}$	BERT$_{base}$	89.3	87.4	87.8	85.6	78.2/78.7	87.2	67.0	82.6	90.7
	BERT$_{large}$ $\Uparrow$	89.1	88.4	87.6	86.2	78.8/79.5	87.5	67.9	83.1	91.6

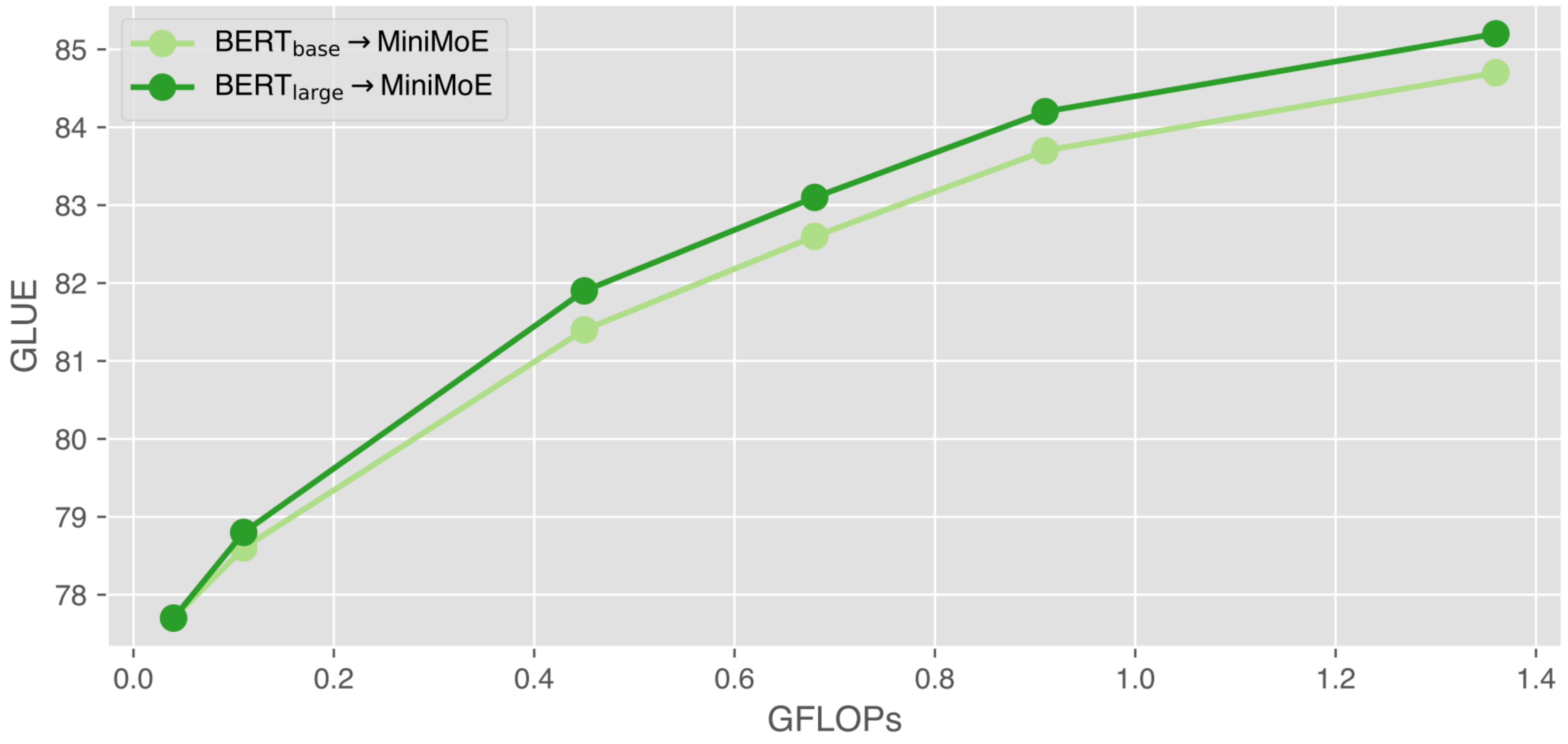
[1] $\Uparrow$ is used to indicate the deficiency is tackled on both GLUE and CoNLL, otherwise $\Downarrow$ is used.

Experiments

Lifted Curse

- The curse is not lifted by MiniLM and MiniLM w/ TA until MiniMoE.





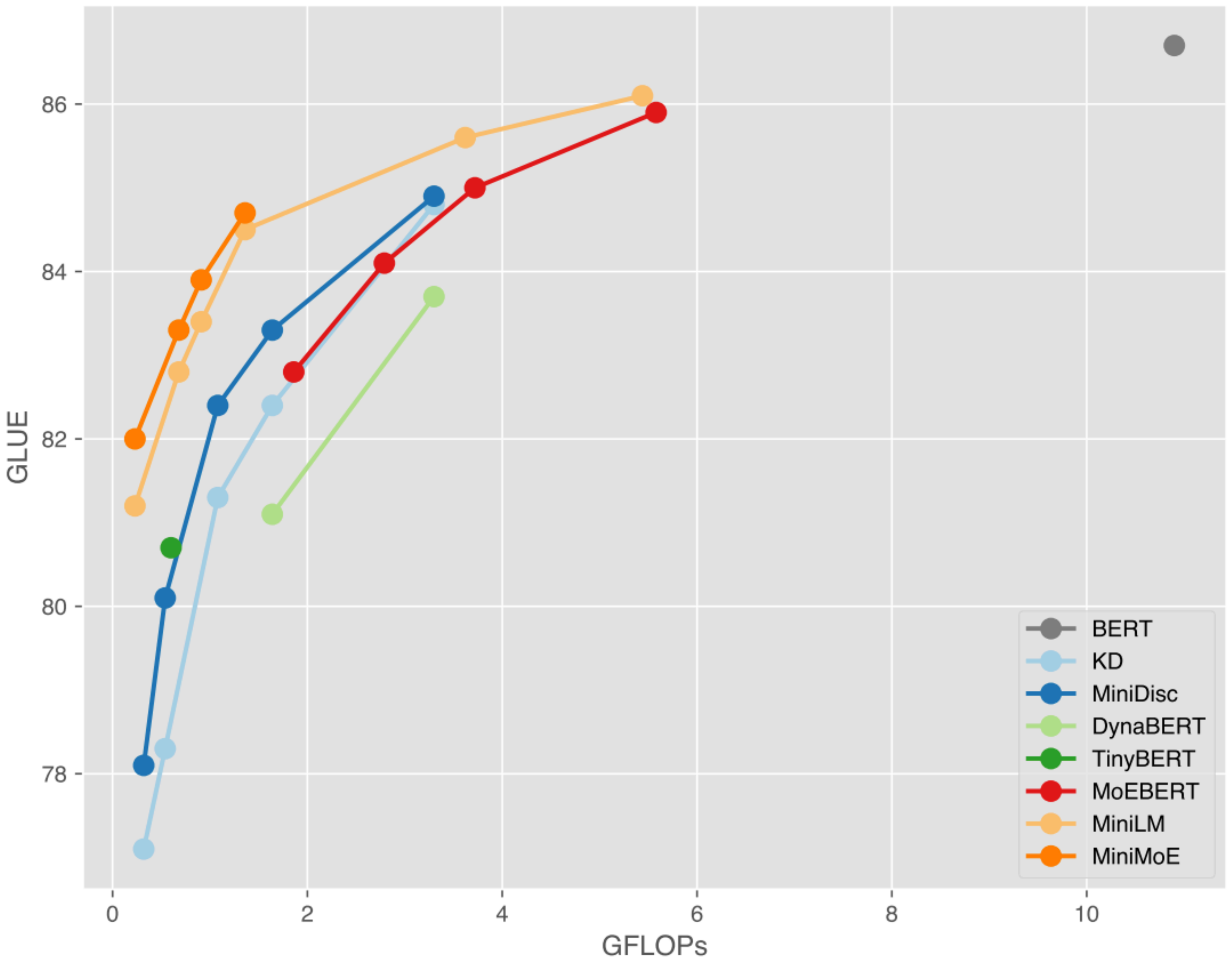
Experiments

Superior Performance

- MiniMoE also achieves new SOTA results.

Method	GFLOPs		SST-2 Acc	MRPC F1	STS-B SpCorr	QQP F1	MNLI-m/mm Acc	QNLI Acc	RTE Acc	GLUE Score	CoNLL F1
$\text{BERT}_{\text{base}}$	10.9		93.8	91.5	87.1	88.4	84.9/84.9	91.9	71.5	86.7	94.8
$\text{KD}_{15\%}$	1.64	6~8×	89.9	88.6	85.1	86.2	79.8/80.2	85.6	63.9	82.4	92.8
$\text{PKD}_{15\%}$	1.64		90.0	88.2	85.5	86.4	80.4/79.6	85.9	63.9	82.5	92.9
$\text{MoEBERT}_{17\%}$[1]	1.86		89.6	88.4	85.1	86.8	80.4/80.5	86.6	65.0	82.8	92.7
$\text{DynaBERT}_{15\%}$[2]	1.64		89.1	85.1	84.7	84.3	78.3/79.0	86.6	61.4	81.1	-
$\text{MiniDisc}_{15\%}$[3]	1.64		89.8	88.2	85.8	86.6	80.3/79.9	87.3	68.2	83.3	93.0
$\text{MiniLM}_{6L;384H}$	1.36		91.1	90.1	88.1	86.7	81.5/**81.8**	89.2	67.9	84.5	93.2
w/ TA	1.36		91.3	**90.3**	88.2	**86.8**	81.4/81.6	**89.7**	66.8	84.5	93.2
$\text{MINIMOE}_{6L;384H}$	1.36		**91.3**	90.2	**88.6**	86.5	**81.6**/81.5	89.5	**68.6**	**84.7**	**93.3**
$\text{KD}_{10\%}$	1.08	10~12×	88.2	87.6	84.0	84.4	77.6/77.4	84.3	67.2	81.3	91.2
$\text{MiniDisc}_{10\%}$	1.08		89.1	88.4	85.4	84.9	78.2/78.6	86.3	68.2	82.4	91.9
$\text{MiniLM}_{4L;384H}$	0.91		90.0	**88.6**	87.2	86.1	80.0/80.3	87.9	67.2	83.4	91.5
w/ TA	0.91		90.0	88.5	87.3	**86.3**	**80.1/80.7**	88.0	66.4	83.4	91.8
$\text{MINIMOE}_{4L;384H}$	0.91		**90.8**	88.1	**88.2**	85.9	79.8/80.4	**88.6**	**69.3**	**83.9**	**92.3**
$\text{KD}_{5\%}$	0.54	16~20×	85.6	84.0	83.8	82.5	72.6/73.2	81.6	63.2	78.3	83.1
$\text{MiniDisc}_{5\%}$	0.54		86.9	87.6	84.8	83.5	72.7/74.5	84.0	66.8	80.1	85.6
$\text{TinyBERT}_{4L;312H}$[4]	0.60		88.5	87.9	86.6	85.6	**78.9/79.2**	**87.3**	**67.2**	**82.7**	-
$\text{MiniLM}_{3L;384H}$	0.68		89.1	**89.1**	86.6	85.4	77.8/78.4	87.2	66.1	82.5	90.1
w/ TA	0.68		**89.8**	87.8	86.0	85.5	77.6/78.5	86.8	66.1	82.3	90.4
$\text{MINIMOE}_{3L;384H}$	0.68		89.3	87.4	**87.8**	**85.6**	78.2/78.7	87.2	67.0	82.6	**90.7**
$\text{KD}_{3\%}$	0.32	34~47×	85.2	83.6	81.9	82.1	71.9/72.7	81.9	57.4	77.1	74.3
$\text{MiniDisc}_{3\%}$	0.32		85.9	85.7	83.6	83.1	72.9/73.6	81.9	58.1	78.1	80.5
$\text{MiniLM}_{4L;192H}$	0.23		86.9	**86.4**	85.4	84.3	77.5/77.5	85.9	65.3	81.2	90.0
w/ TA	0.23		87.2	85.6	86.2	84.6	77.3/**78.0**	86.6	64.6	81.3	89.9
$\text{MINIMOE}_{4L;192H}$	0.23		**88.1**	86.1	**86.2**	**84.8**	**77.7**/77.8	**86.6**	**68.6**	**82.0**	**91.3**

[1] Each FFN is split to 8 experts and each MHA to 4 to reach the sparsity.
[2] The results are produced from the released code.
[3] The results are mainly taken from the original papers.
[4] The results are produced without additional task-specific distillation.



Experiments

Practical Compute

- We also offer the practical compute consumed by MiniMoE. The imposed MoE student would not add to the inference compute that much.

Method	GFLOPs	Throughput	Params
$\text{BERT}_{\text{base}}$	10.9	80.8 tokens/ms	109.5 M
$\text{KD}_{5\%}$	0.54	544.7 tokens/ms	28.7 M
$\text{MiniLM}_{\text{3L;384H}}$	0.68	485.3 tokens/ms	17.2 M
$\text{MINIMOE}_{\text{3L;384H}}$	0.68	433.1 tokens/ms	28.3 M

Experiments

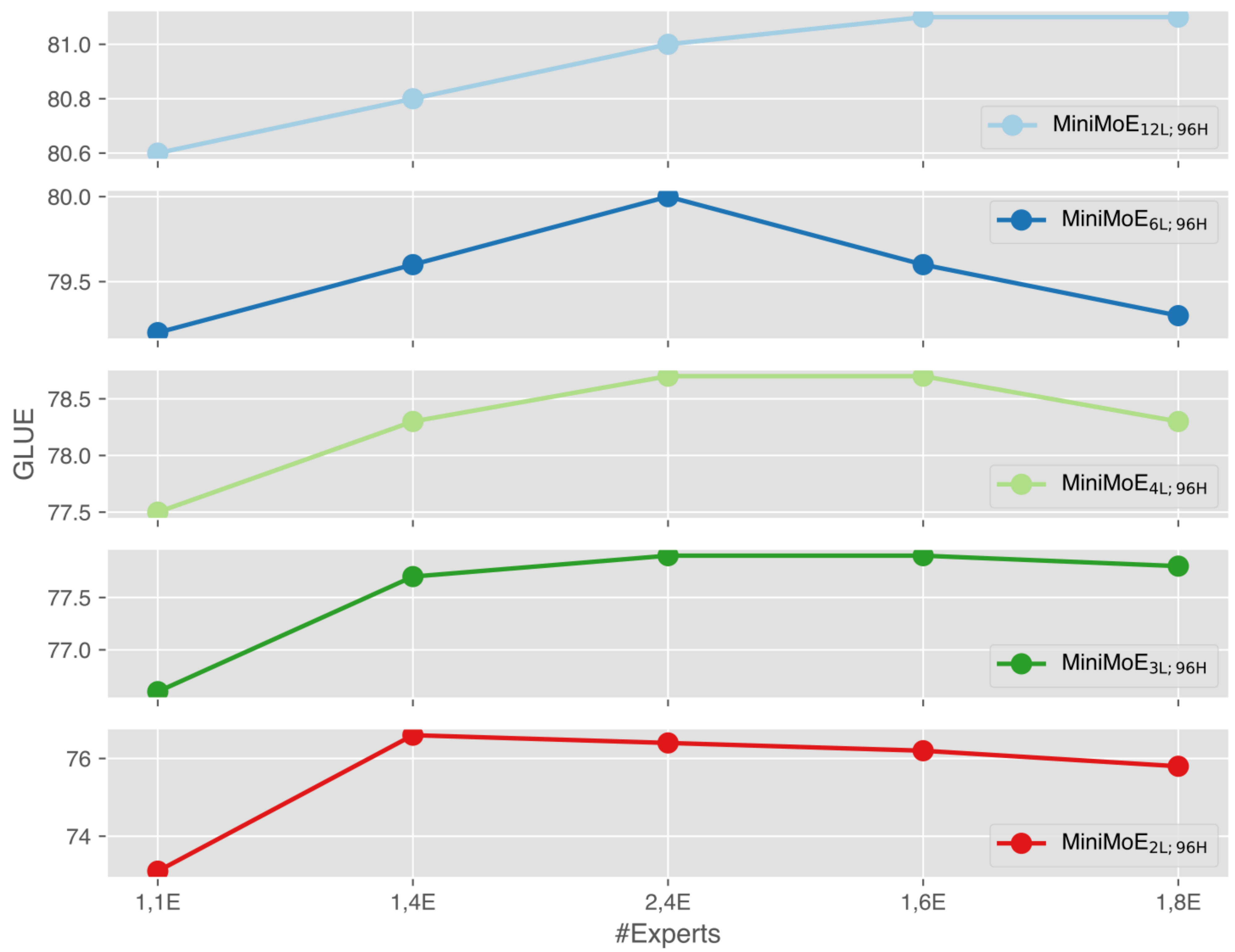
Memory Efficiency

- Nonetheless, would it be possible to reduce the parameter amount? Yes, we find that every expert can be more or less pruned from a SVD (singular value) perspective. We leave this to future.

Method	**%Value>0.2**	**%Value>0.1**	**%Value>0.05**	**Trm Params (Value>0.1)**
$\text{MiniLM}_{3L;384H}$ dense	315/384=82%	356/384=93%	373/384=97%	5.3M→5.1M
$\text{MiniMoE}_{3L;384H}$ expert #1	6/384=2%	82/384=21%	275/384=72%	-
$\text{MiniMoE}_{3L;384H}$ expert #2	34/384=9%	220/384=57%	361/384=94%	-
$\text{MiniMoE}_{3L;384H}$ expert #3	15/384=4%	175/384=46%	338/384=88%	-
$\text{MiniMoE}_{3L;384H}$ expert #4	24/384=6%	200/384=52%	357/384=93%	-
$\text{MiniMoE}_{3L;384H}$ all experts	79/384/4=5%	677/384/4=44%	1331/384/4=87%	16.4M→8.2M

Experiments
Expert Number

- We also explore how would the expert number impact the performance. Adding experts can increase the variance thus lead to degraded performance for students that are already small (with high variance).



Conclusion

- MiniMoE can largely lift the curse, but still leaves the room for improvement.
- However, given that MiniMoE yet cannot fully lift the curse, we are still wondering whether the issue of capacity gap is really a curse that should be lifted or just a law that should be adopted?
- arXiv: https://arxiv.org/abs/2305.12129
- GitHub: https://github.com/GeneZC/MiniMoE
- HuggingFace: https://huggingface.co/GeneZC/bert-base-minimoe-6L-384H and more
- Slides: https://genezc.github.io/assets/files/ACL2023_MiniMoE.pdf
- Thank you all!