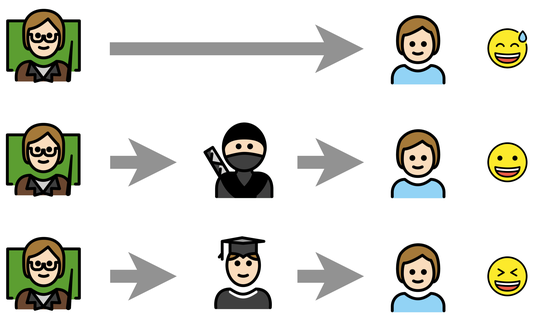
Minimal Distillation Schedule for Extreme Language Model Distillation

Chen Zhang[1], Yang Yang[2], Qifan Wang[3], Jiahao Liu[2], Jingang Wang[2], Wei Wu[2], Dawei Song[1*]
[1]Beijing Institute of Technology, [2]Meituan NLP, [3]Meta AI, *Corresponding author

Motivation



- distillation directly from the teacher to the student may suffer from teacher-student capacity gap.
- distillation with the teacher assistant is way better yet may suffer from assistant sub-optimality.
- distillation with the optimal teacher assistant is urgent.

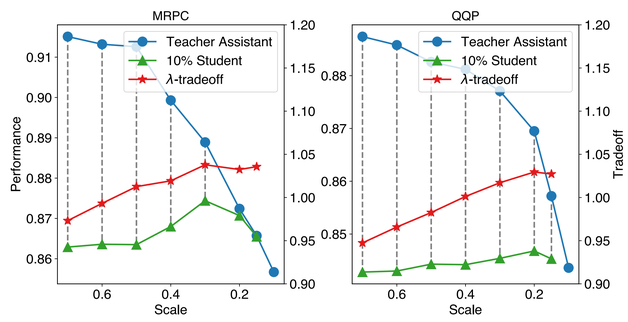
Results

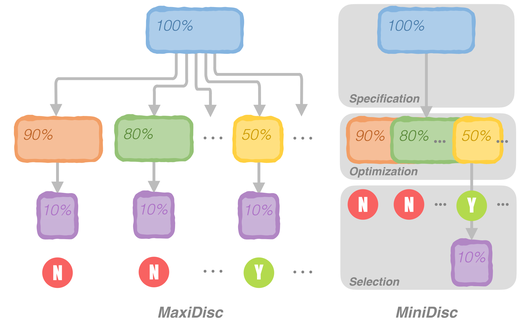
Method	FLOPs	SST-2	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	Average	GPUs
$\text{BERT}_{\text{base}}$	10.9G	93.8	91.5	87.1	88.4	84.9/84.9	91.9	71.5	86.7	–
Teacher Assistant-based Distillation										
$\text{TA}_{15\%}$ (2020)	1.6G	89.3	87.7	85.3	85.7	80.0/80.3	88.1	68.4	83.1	2×
$\text{MAXIDISC}_{15\%}$	1.6G	89.8	87.7	85.4	86.9	81.0/80.1	86.1	68.2	83.2	40×
$\text{MINIDISC}_{15\%}$	1.6G	89.8	88.2	85.8	86.6	80.3/79.9	87.3	68.2	**83.3**	4×
$\text{TA}_{10\%}$ (2020)	1.1G	89.1	87.9	83.1	84.7	77.8/77.9	85.7	68.6	81.8	2×
$\text{MAXIDISC}_{10\%}$	1.1G	89.0	88.2	84.8	84.8	78.3/77.8	85.3	66.8	81.9	40×
$\text{MINIDISC}_{10\%}$	1.1G	89.1	88.4	85.4	84.9	78.2/78.6	86.3	68.2	**82.4**	4×
$\text{TA}_{5\%}$ (2020)	0.5G	86.5	86.5	82.2	83.2	73.3/73.7	82.6	65.3	79.2	2×
$\text{MAXIDISC}_{5\%}$	0.5G	86.9	88.3	84.8	83.7	74.4/76.3	83.5	65.0	**80.4**	40×
$\text{MINIDISC}_{5\%}$	0.5G	86.9	87.6	84.8	83.5	72.7/74.5	84.0	66.8	80.1	4×

- comprehensive results demonstrate the improved efficiency, see more results in our paper.

Method

- scale-performance tradeoff is a good indicator of assistant optimality, which is quantitatively defined by a lambda-tradeoff without an individual run to the student:
  - $t = m + lambda * (1 - s)$, where m is the performance and $(1 - s)$ is the sparsity w.r.t. the teacher scale.





- the sparsity is a trivial measure while the performance is a measure that requires training, which is efficiently estimated by a sandwich framework:
  - *parameter sharing* among these candidate assistants admits one-run and more efficient training of all these assistants.

Conclusions

- Having observed that the scale-performance tradeoff of the teacher assistant is of great importance to the student, we introduce a lambda-tradeoff. To efficiently compute the measures for teacher assistant candidates, we design a sandwich optimization for these candidates.
- This work is funded in part by the Natural Science Foundation of China (grant no: 62376027) and Beijing Municipal Natural Science Foundation (grant no: 4222036 and IS23061).