Rethinking Network Pruning under the Pre-train and Fine-tune Paradigm

Dongkuan Xu[1], Ian En-Hsu Yen[2], Jinxi Zhao[2], Zhibin Xiao[2]
[1]The Pennsylvania State University, [2]Moffett AI

Welcome! Email: *dux19@psu.edu* Twitter: *DongkuanXu* WeChat: *xudongkuan220019*

Background

☑ Model Compression

- **Compression is desirable**

Fig. Resources are constrained [1]

- **Pruning is a popular approach**

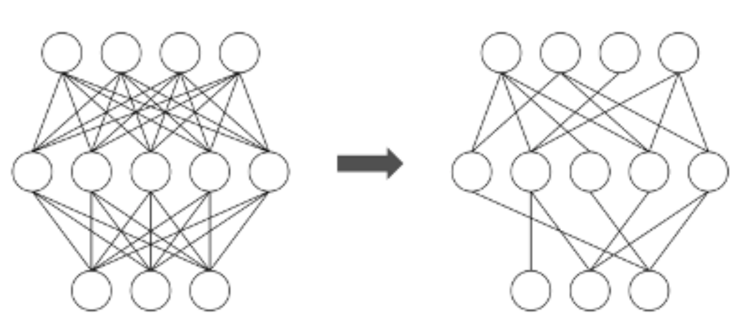


Fig. Illustration of network pruning

☑ Structural vs. Sparse

- **Structural pruning: a layer**
- **Sparse pruning: a neuron**

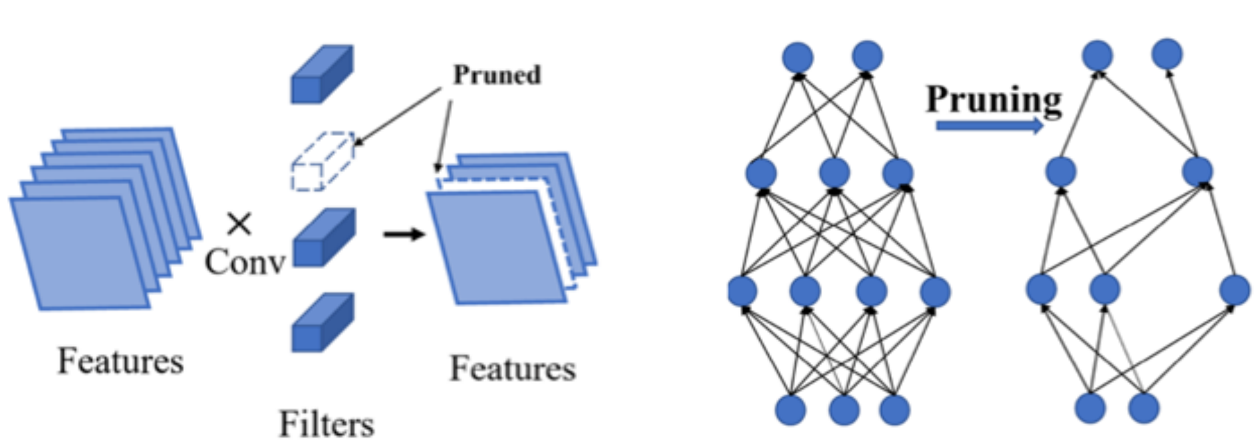


Fig. Structural pruning vs. sparse pruning [2]

Motivation

- **Sparse pruning is more impressive than structural pruning in CNN community**

Width	Sparsity	NNZ params	Top-1 acc.	Top-5 acc.
0.25	0%	0.46M	50.6%	75.0%
0.5	0%	1.32M	63.7%	85.4%
0.75	0%	2.57M	68.4%	88.2%
1.0	0%	4.21M	70.6%	89.5%
	50%	2.13M	69.5%	89.5%
	75%	1.09M	67.7%	88.5%
	90%	0.46M	61.8%	84.7%
	95%	0.25M	53.6%	78.9%

Structural result

Sparse result

Fig. MobileNets sparse vs structural results [3]

- **However, existing sparse pruning of BERT yields inferior results than its small-dense counterparts**

Proposed: Knowledge-Aware Sparse Pruning

☑ Proposed: Pruning At Distilling

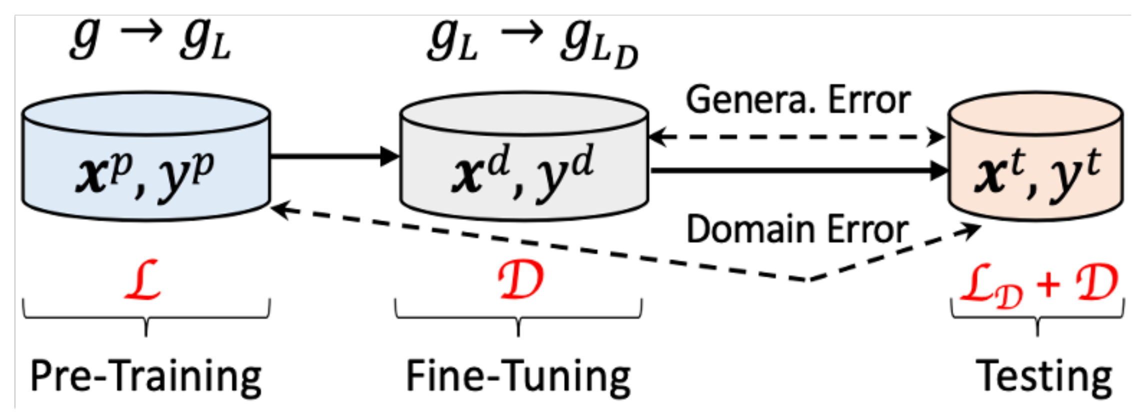


Fig. General pre-training & fine-tuning

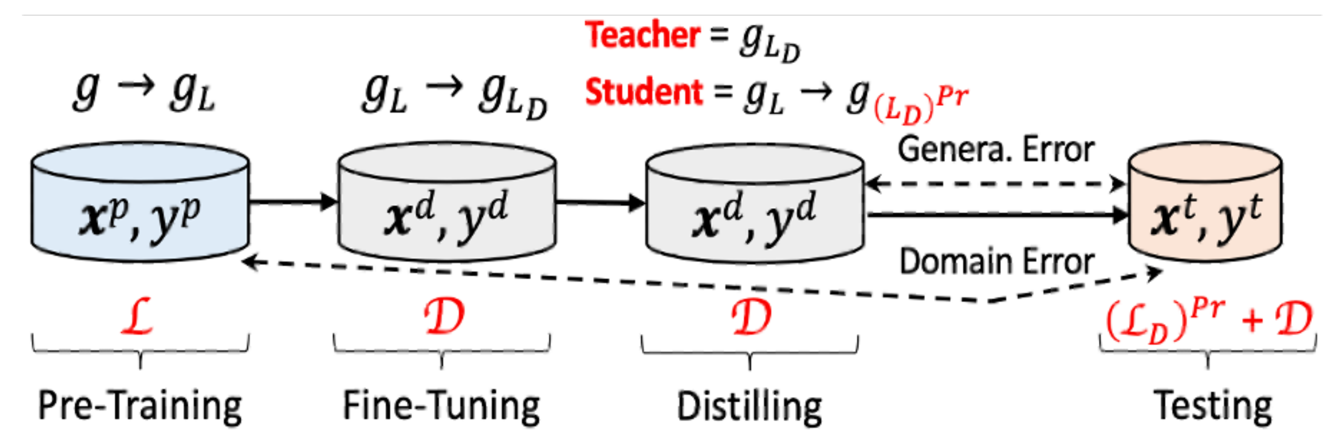


Fig. Pruning at distilling (proposed)

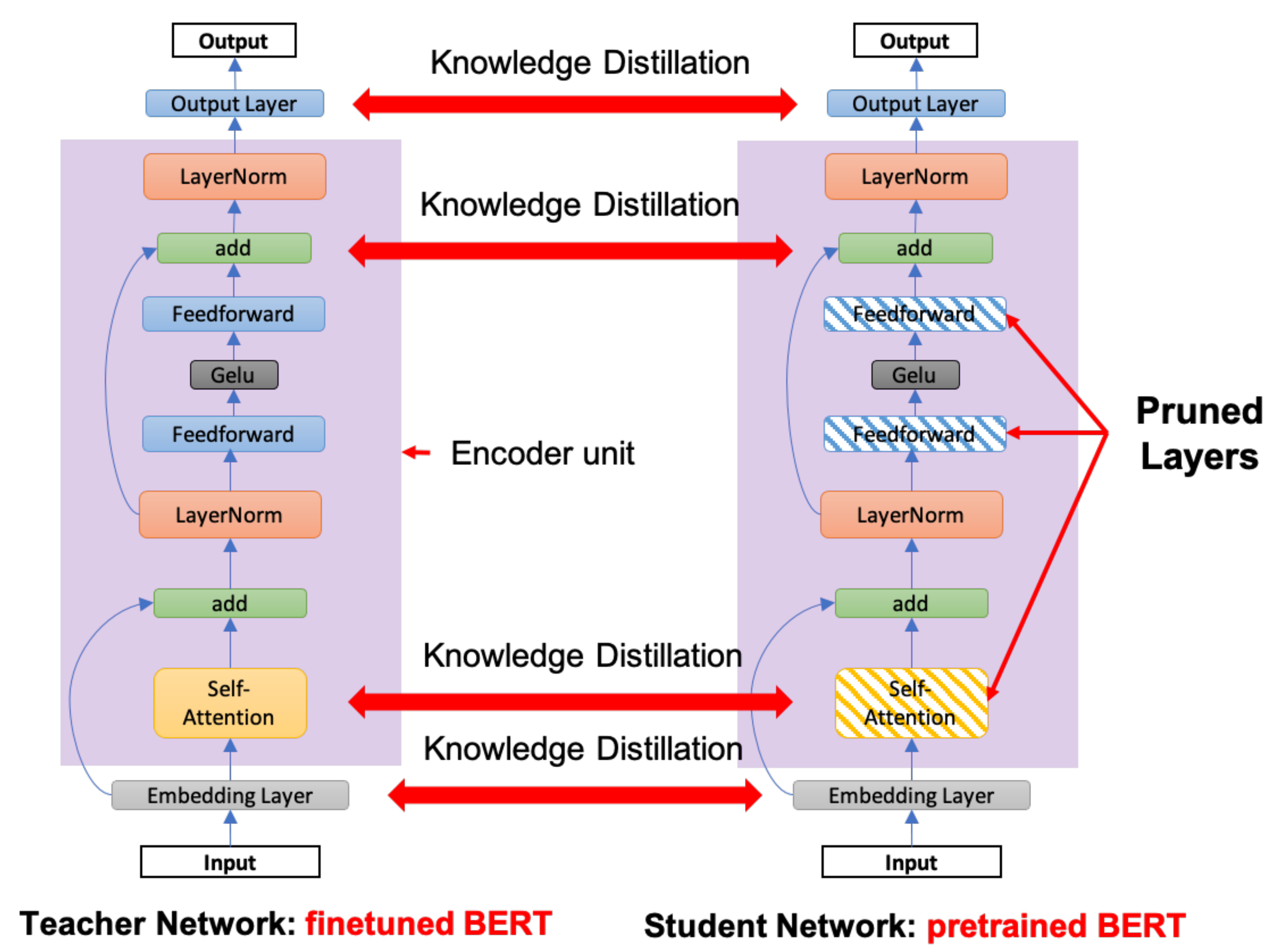


Fig. Distillation details

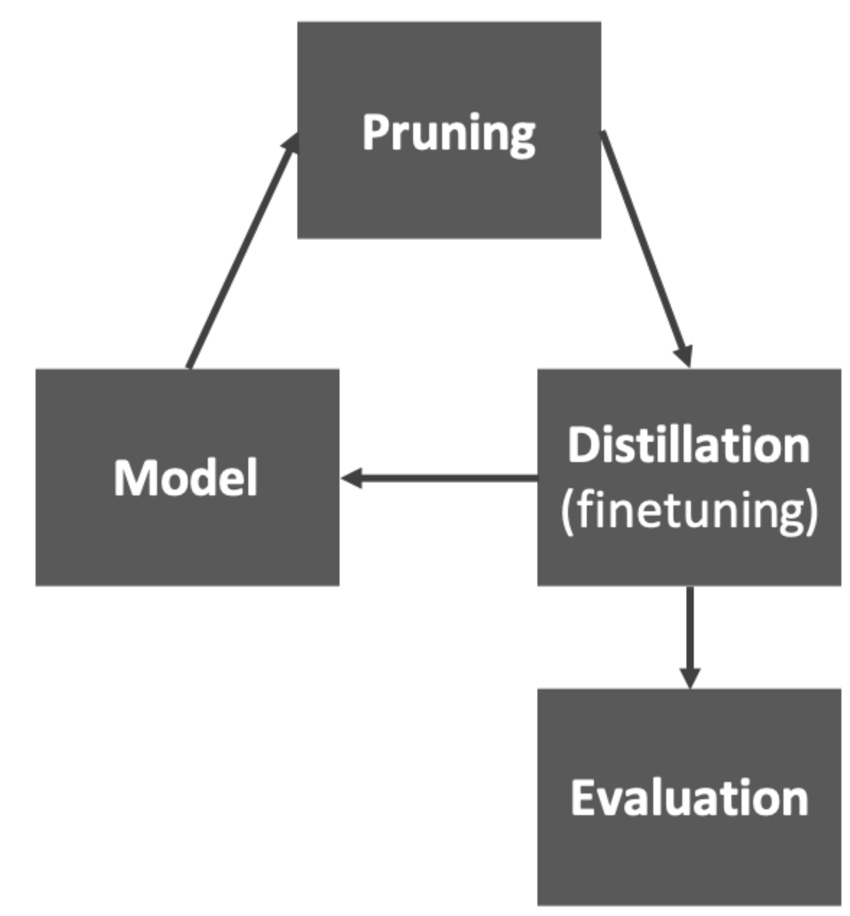


Fig. Pruning details

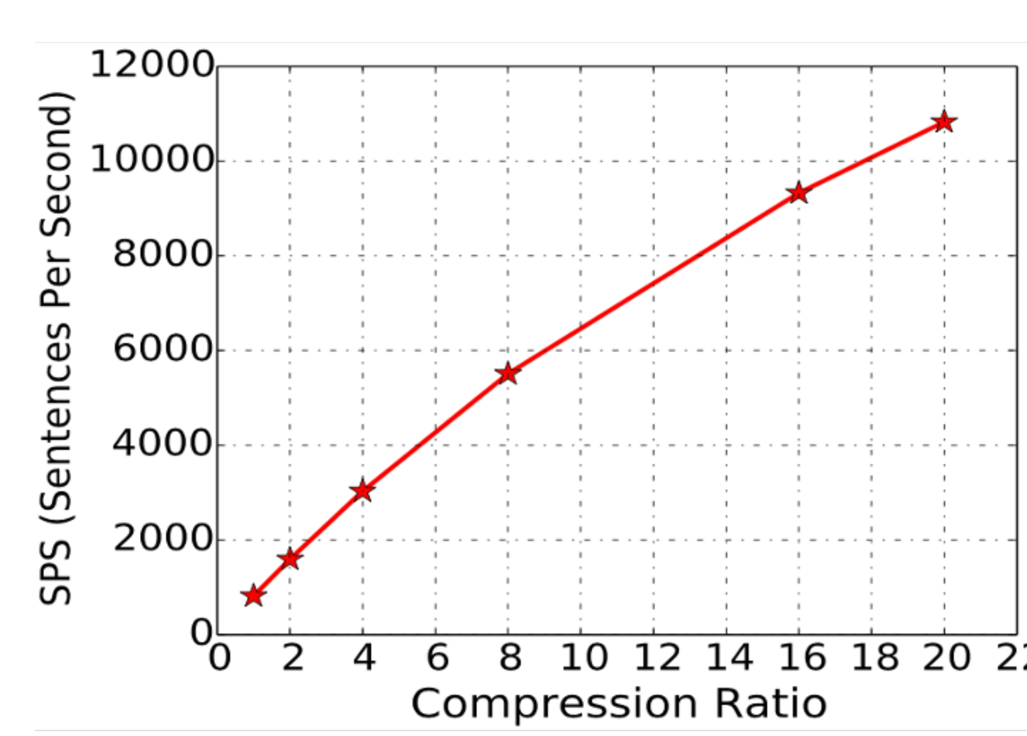
Experiments

☑ GLUE (dev set)

Method	Remain. Weights	QNLI (Acc)	MRPC (F1)	RTE (Acc)	CoLA (Mcc)	Avg.
Without Pruning						
BERT-base	-	91.8	88.6	69.3	56.3	76.5
ELMo	-	71.1	76.6	53.4	44.1	61.3
Structural Pruning						
$BERT_6$-PKD	50%	89.0	85.0	65.5	45.5	71.3
BERT-of-Theseus	50%	89.5	89.0	68.2	51.1	74.5
DistilBERT	50%	89.2	87.5	59.9	51.3	72.0
$MiniLM_6$	50%	91.0	88.4	71.5	49.2	75.0
$TinyBERT_6$	50%	90.4	87.3	66.0	54.0	74.4
$TinyBERT_4$	18%	88.7	86.8	66.5	49.7	72.9
Sparse Pruning						
BERT-Tickets	30-50%	88.9	84.9	66.0	53.8	73.2
CompressBERT	10%	76.8	-	-	-	-
RPP	11.6%	88.0	81.9	67.5	-	-
SparseBERT	5%	90.6	88.5	69.1	52.1	75.1

Compression ratio = x20,
but only 1.4% performance drop

☑ Hardware Performance



Performance under different compression ratios on MRPC (Moffett AI's latest hardware platform ANTOM)

[1] https://www.robots.ox.ac.uk/~namhoon/doc/slides-snip-msr.pdf
[2] Chen, L., Chen, Y., Xi, J., & Le, X. (2021). Knowledge from the original network: restore a better pruned network with knowledge distillation. Complex & Intelligent Systems, 1-10.
[3] Zhu, M., & Gupta, S. (2017). To prune, or not to prune: exploring the efficacy of pruning for model compression. arXiv preprint arXiv:1710.01878.