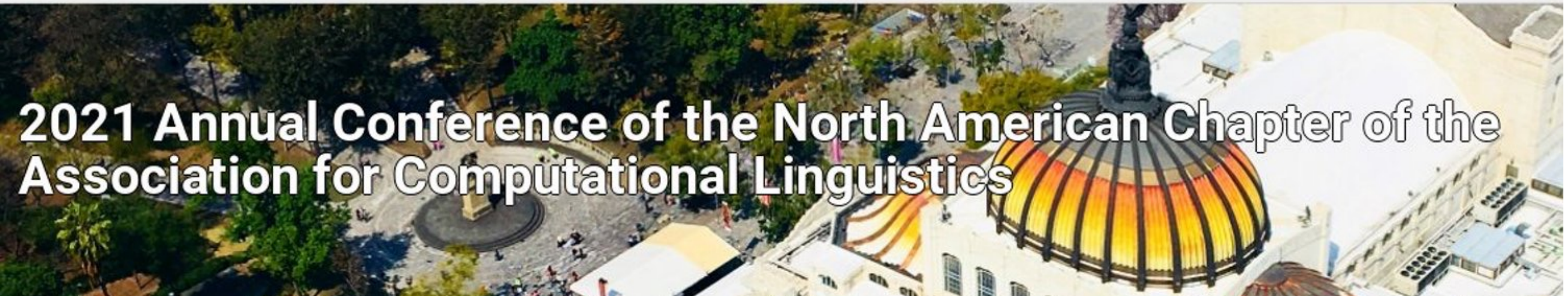


Rethinking Network Pruning under the Pre-train and Fine-tune Paradigm

Dongkuan Xu[1], Ian En-Hsu Yen[2], Jinxi Zhao[2], Zhibin Xiao[2]

[1]The Pennsylvania State University

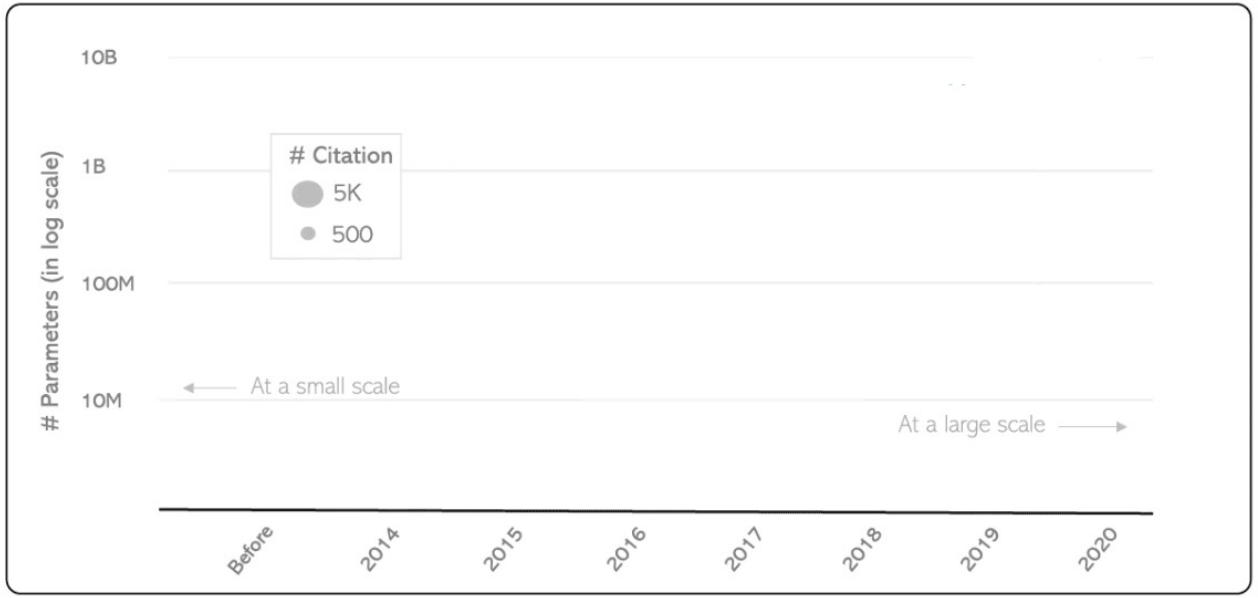
[2]Moffett AI

PennState

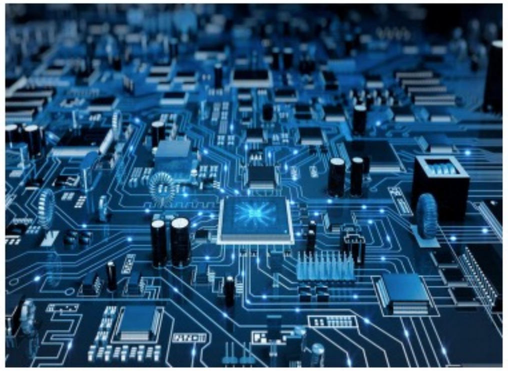
MOFFETT AI

Background

- Neural networks become larger and larger [1]
- But resources (memory, computation, power) are constrained [2]



Evolution of deep learning models over time, in terms of **model size**



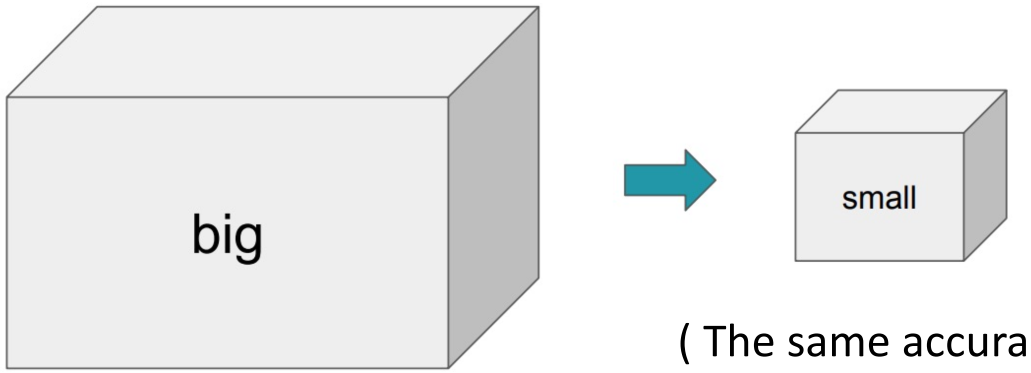
Embedded Systems e.g., Mobile Devices

Real-Time Tasks e.g., Autonomous Car

[1] https://www.microsoft.com/en-us/research/blog/a-deep-generative-model-trifecta-three-advances-that-work-towards-harnessing-large-scale-power/
[2] https://www.robots.ox.ac.uk/~namhoon/doc/slides-snip-msr.pdf

Background

- Compression Is Desirable
- Pruning is a popular compression approach



(The same accuracy)

Illustration of model compression [1]

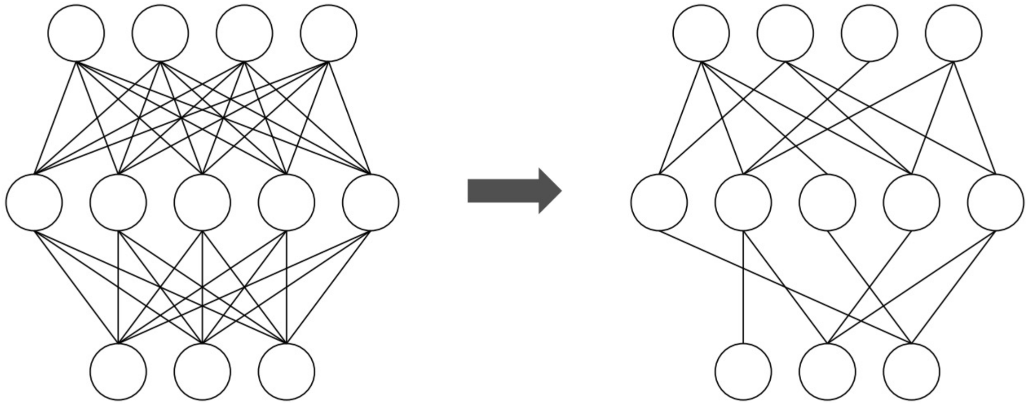
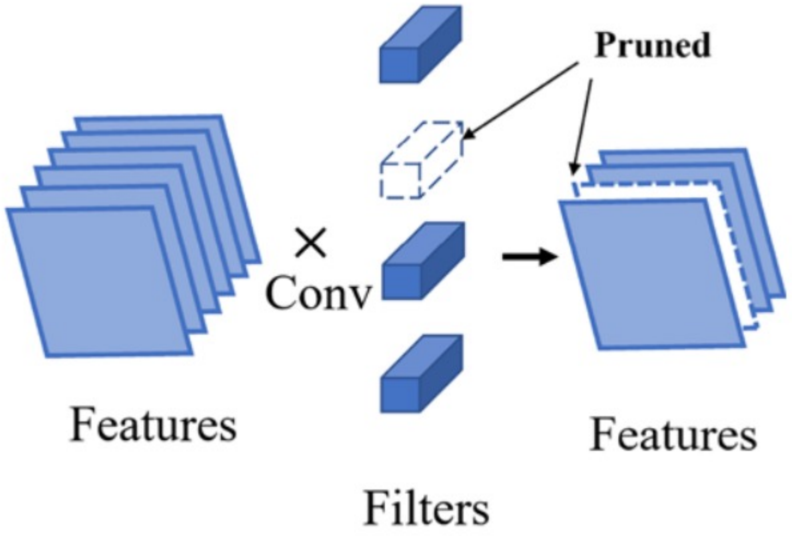


Illustration of network pruning [2]

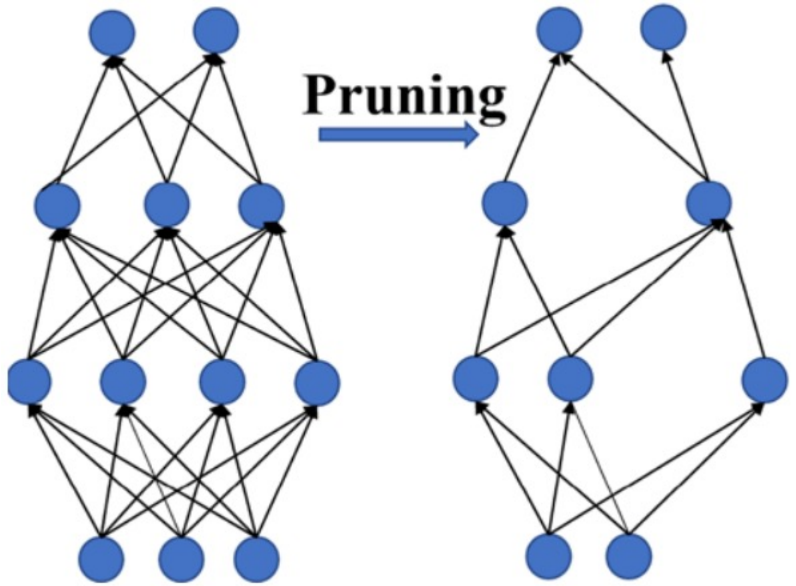
[1] https://www.robots.ox.ac.uk/~namhoon/doc/slides-snip-msr.pdf
[2] https://mlsys.org/media/Slides/mlsys/2020/balla(02-14-30)-02-14-30-1413-what_is_the.pdf

Background: Structural Pruning vs. Sparse Pruning

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



Structural pruning for CNN [1]



Sparse pruning for fully connected networks [1]

[1] 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.

Motivation: Sparse Pruning

- 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

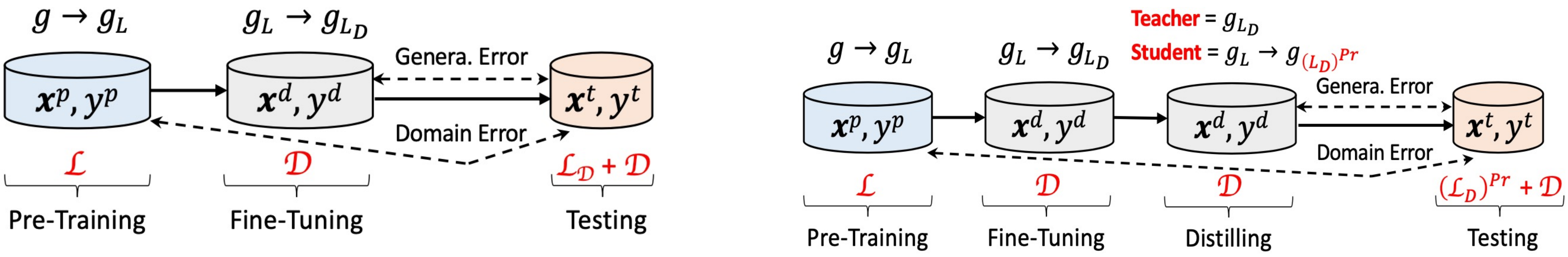
MobileNets sparse vs structural results [1]

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

[1] Zhu, M., & Gupta, S. (2017). To prune, or not to prune: exploring the efficacy of pruning for model compression. arXiv preprint arXiv:1710.01878.

Proposed: Knowledge-Aware Sparse Pruning

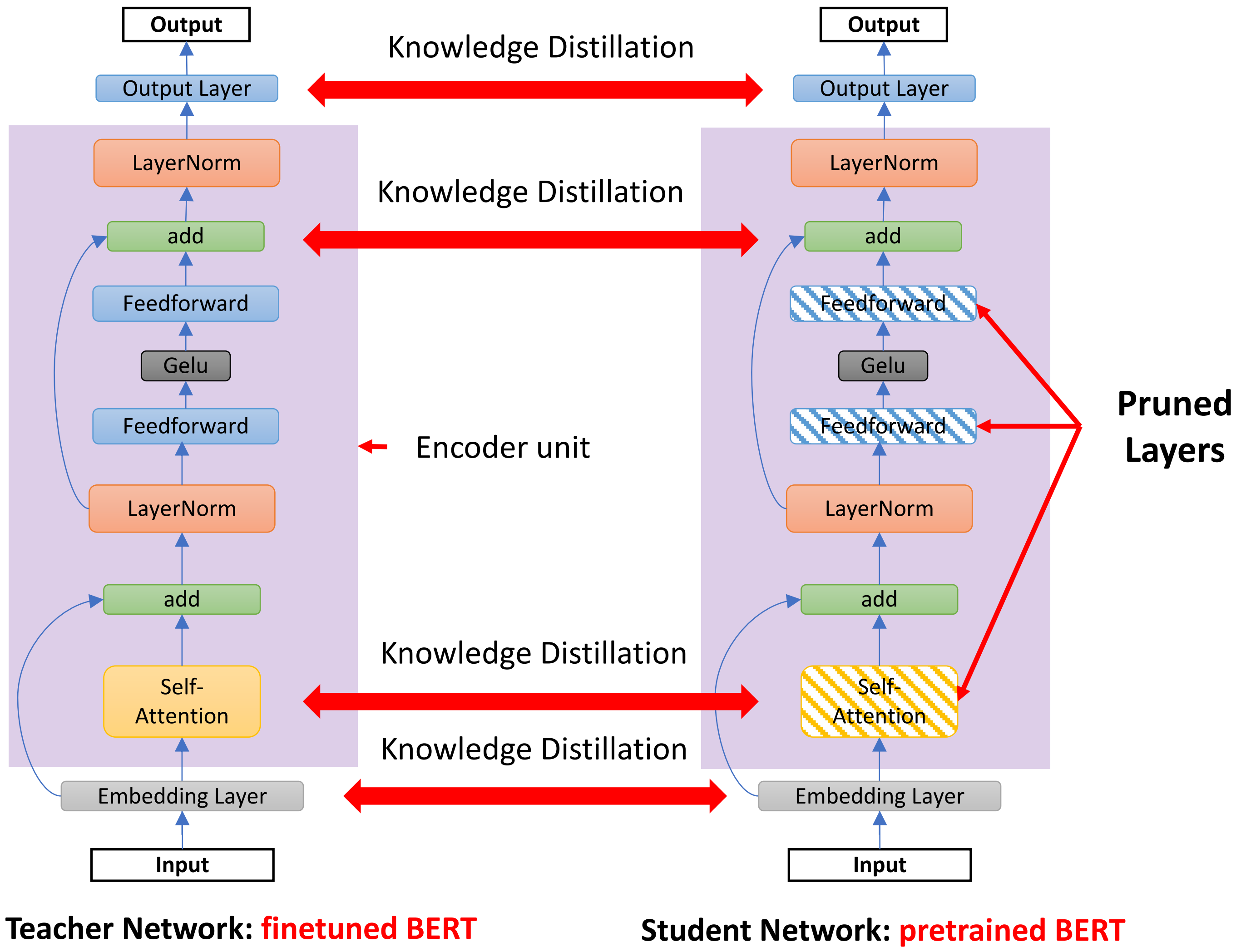
- Two gaps in pre-training & fine-tuning procedure
- Proposed: pruning at distilling



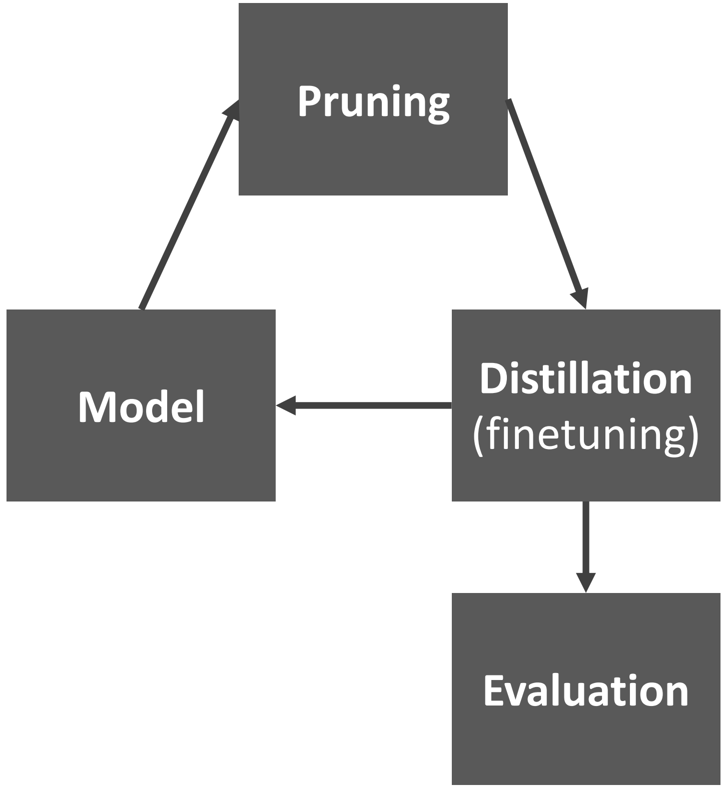
General pre-training & fine-tuning

Pruning at Distilling (Proposed)

Proposed: Knowledge-Aware Sparse Pruning



Distillation details



Cyclic pruning

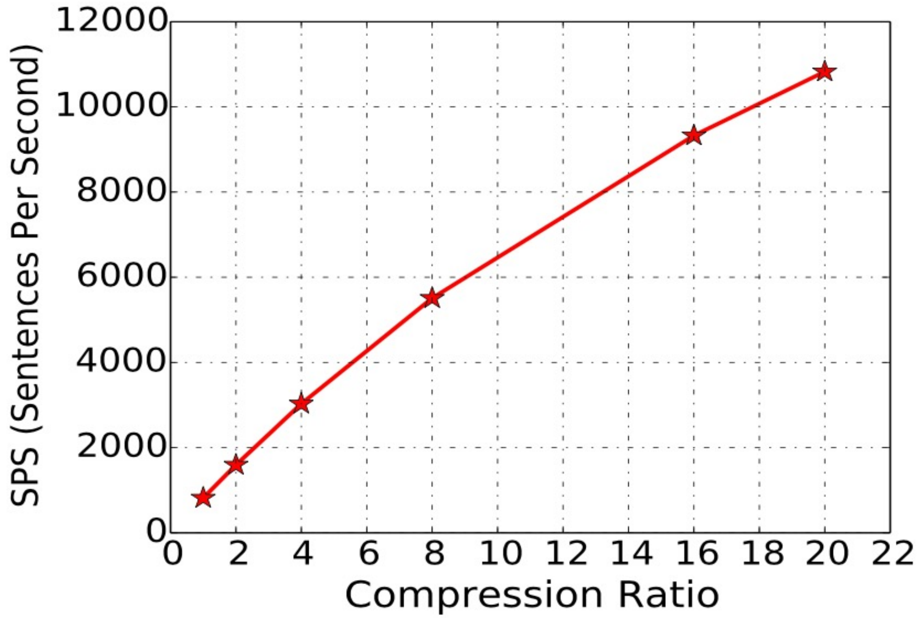
Experimental Results

- GLUE

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

Comparison on the dev sets:
Compression ratio = x20, but only 1.4% performance drop

- Hardware performance



Performance under different compression ratios on the MRPC dataset (sentences per second)

Discussion

- Why is sparse pruning an important topic?
  - Sparse pruning leads to significantly higher compression ratio than structural pruning
  - Commercial hardware platforms have been starting to support sparse tensor operation

- Sparse pruning is trending and with promising future

Accelerating inference speed from 1.2 to 2.4 times

Google (March 9, 2021): comparison of the processing time for the dense (left) and sparse (right) models of the same quality [1]

[1] https://ai.googleblog.com/2021/03/accelerating-neural-networks-on-mobile.html

Q & A

Web: www.personal.psu.edu/dux19/

Email: dux19@psu.edu