Compressing Large-Scale Transformer-Based Models: A Case Study on BERT

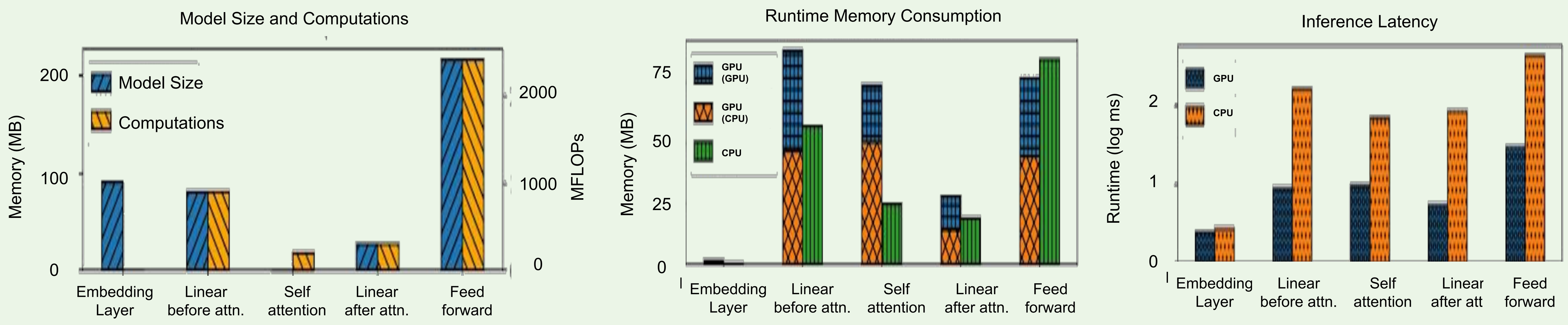
Prakhar Ganesh[1*] Yao Chen[1*] Xin Lou[1] Mohammad Ali Khan[1]
Yin Yang[2] Hassan Sajjad[3] Preslav Nakov[3] Deming Chen[1,4] Marianne Winslett[1,4]

[1]Advanced Digital Sciences Center, Illinois at Singapore [2]College of Science and Engineering, Hamad Bin Khalifa University, Qatar
[3]Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar [4]University of Illinois at Urbana-Champaign, USA

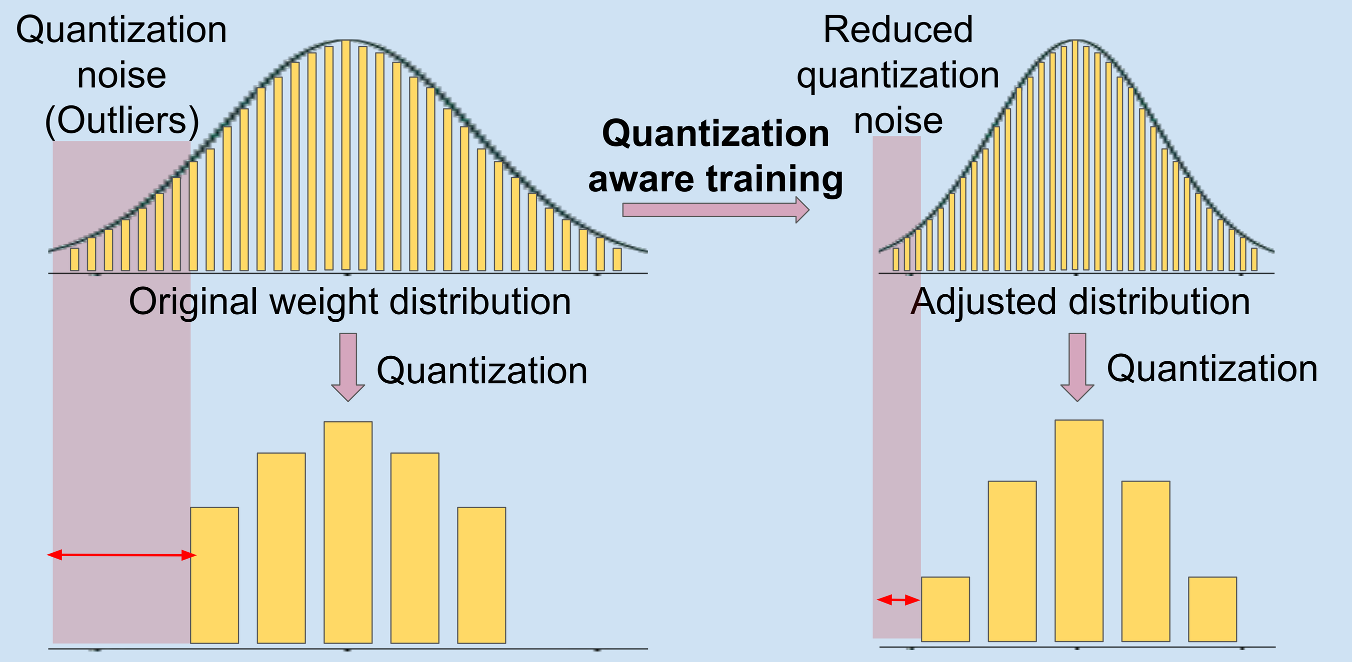
Motivation

- ➢ Growing size of pre-trained models, approaching trillions of parameters
- ➢ Deployment requires access to cloud computing or high-performance clusters
- ➢ Solution: **Model Compression!**
- ➢ **We offer a comprehensive systematic study of model compression for Transformer-based large-scale NLP models, with focus on BERT**

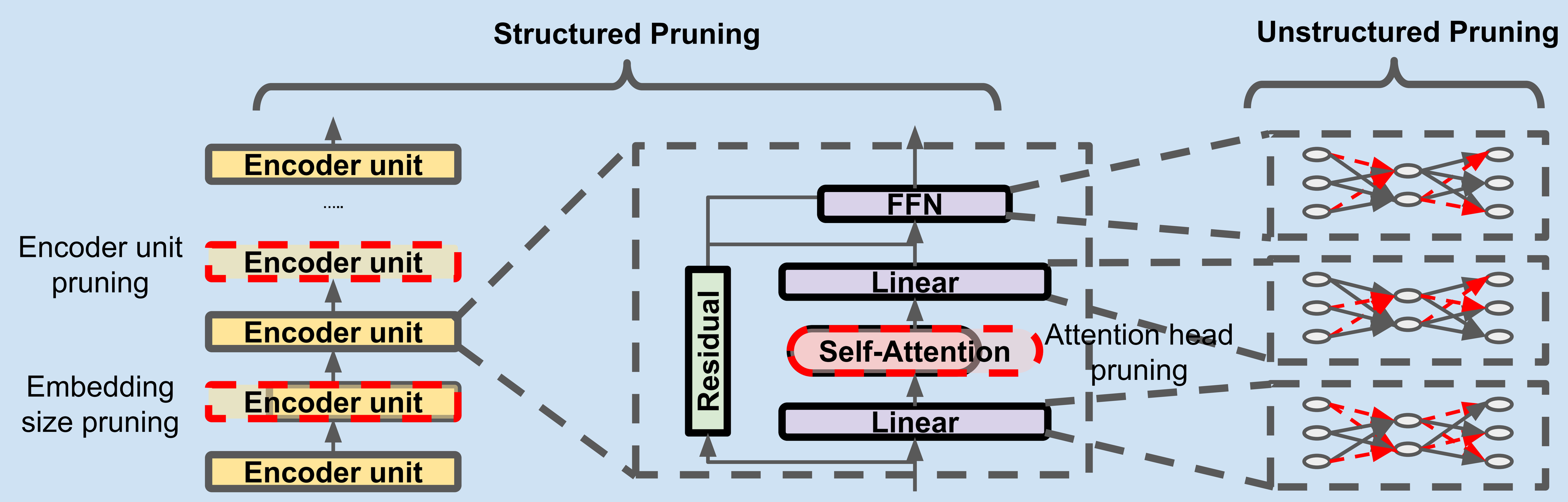
BERT Breakdown Analysis



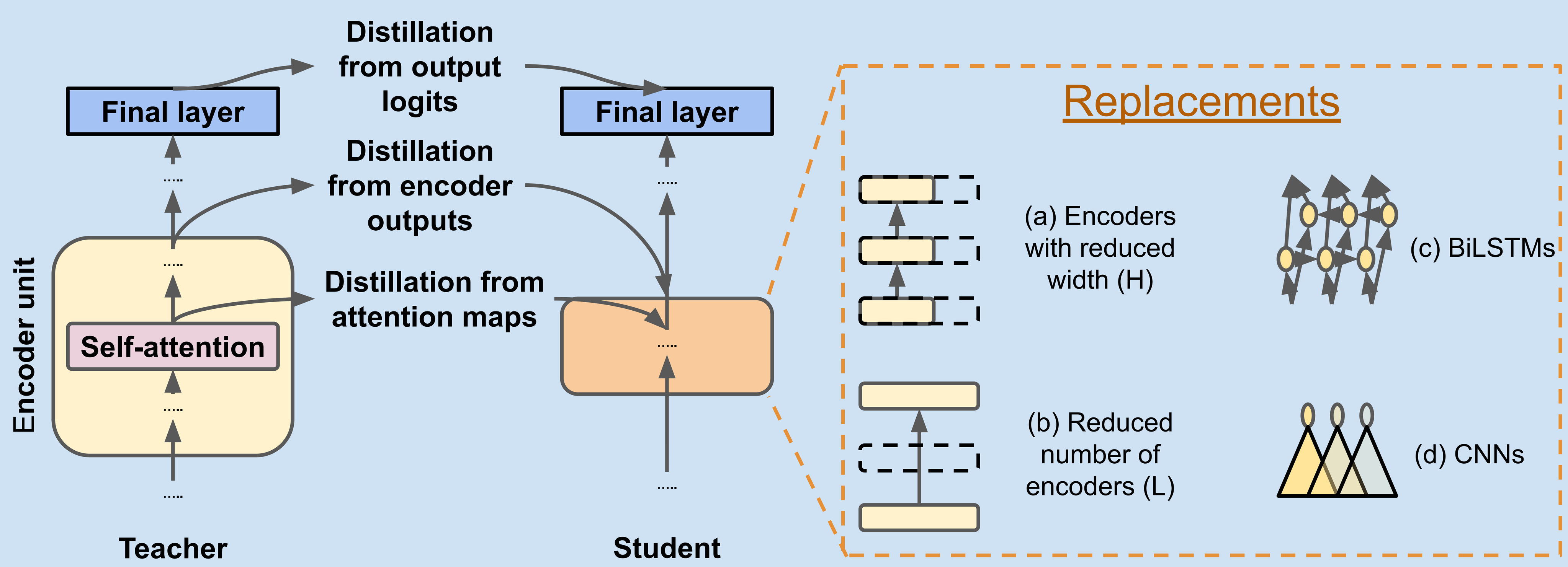
Quantization



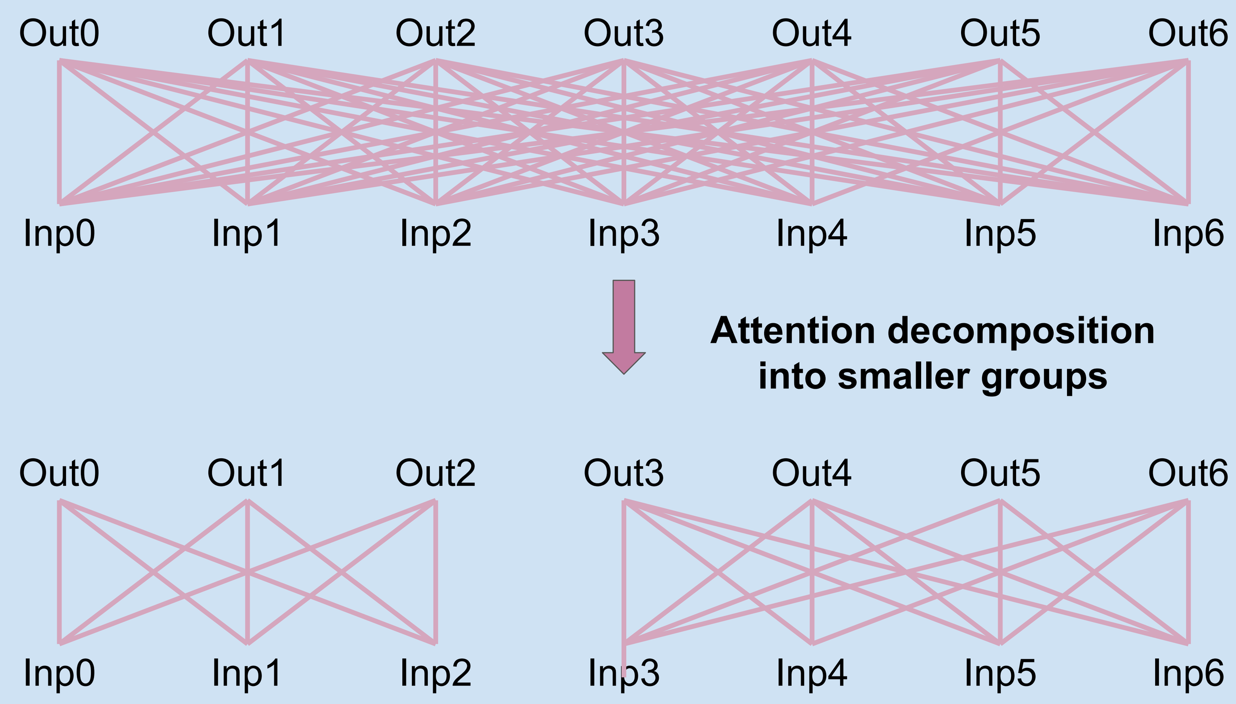
Pruning



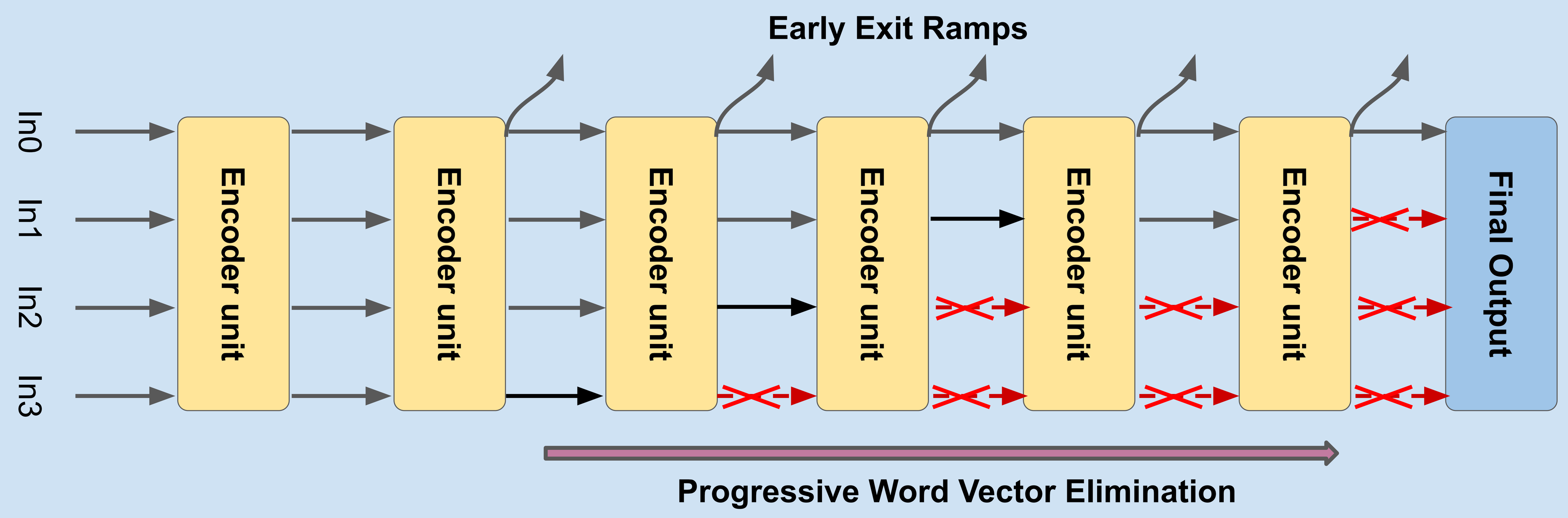
Knowledge Distillation



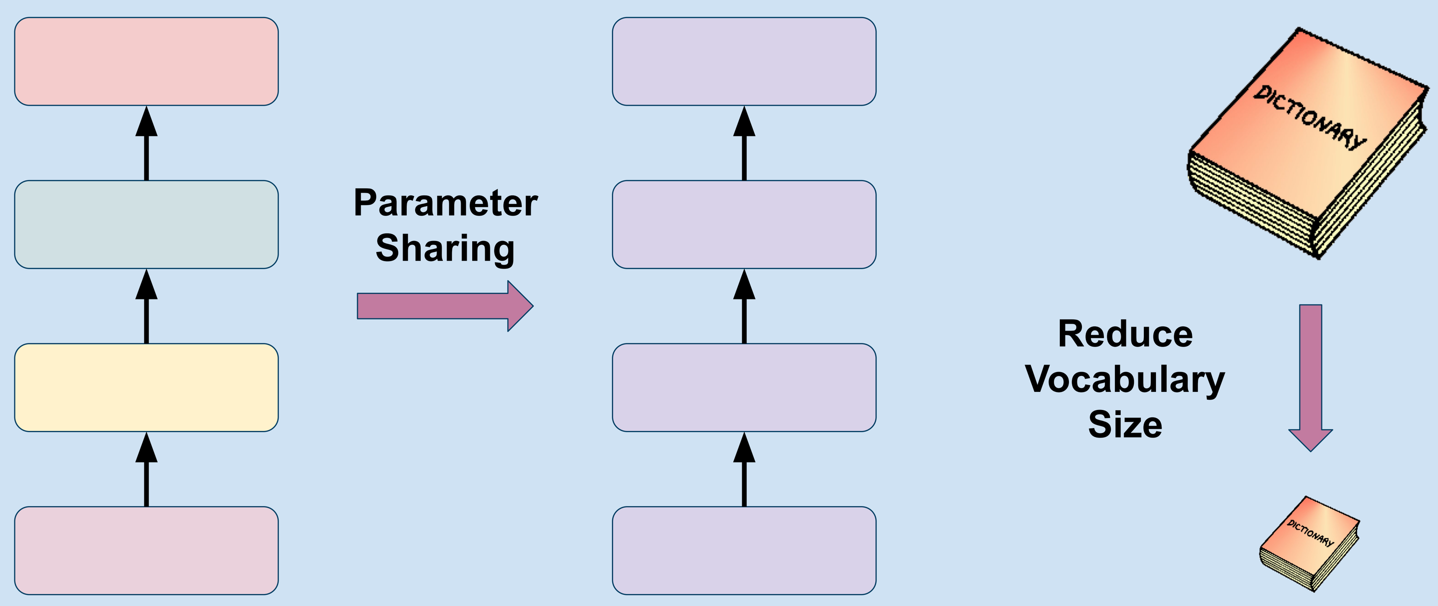
Matrix Decomposition



Dynamic Inference Acceleration



Other Methods



Effectiveness of Compression

Methods	Model Size w/ emb	Model Size w/o emb	Speedup GPU	Speedup CPU	Accuracy/F1 MNLI	Accuracy/F1 QQP	Accuracy/F1 SST-2	Accuracy/F1 SQD	Avr. Drop
$BERT_{BASE}$	100%	100%	1x	1x	84.6	89.2	93.5	88.5	0.0
Quantization	15%	12.5%	1x	1x	83.9	–	92.6	88.3	-0.6
	10.2%	5.5%	1x	1x	83.7	–	–	–	-0.9
Unstructured Pruning	67.6%	58.7%	1x	1x	–	–	–	88.5	0.0
	48.9%	35.1%	1x	1x	83.1	89.5	92.9	87.8	-0.63
	23.8%	3%	1x	1x	79.0	89.3	–	79.9	-4.73
Structured Pruning	60.7%	50%	–	–	–	88.9	91.8	–	-1.0
	39.1%	38.8%	2.93x	2.76x	83.4	–	90.9	86.7	-1.86
KD from Output Logits	22.8%	10.9%	6.25x	7.09x	–	88.6	92.9	–	-0.6
	24.1%	3.3%	10.7x	8.6x	78.6	88.6	91.0	–	-3.03
	7.4%	4.8%	19.5x	–	81.6	88.7	91.8	–	-2.06
KD from Attn.	60.7%	50%	1.94x	1.73x	84.0	91.0	92.0	–	-0.1
Multiple KD combined	60.7%	50%	1.94x	1.73x	82.2	88.5	91.3	86.9	-1.73
	23.1%	24.8%	3.9x	4.7x	83.3	–	92.8	90.0	-0.16
	13.3%	6.4%	9.4x	9.3x	82.5	89.2	92.6	–	-1.0
	1.6%	1.8%	25.5x	22.7x	71.3	–	82.2	–	-12.3
Matrix Decomposition	60.6%	49.1%	0.92x	1.05x	84.8	89.7	92.4	–	-0.13
	100%	100%	3.14x	3.55x	82.6	90.3	–	87.1	-0.76
Dynamic Inference	100%	100%	1.25x	1.28x	83.9	89.2	93.4	–	-0.26
	100%	100%	2.5x	3.1x	83.8	–	92.1	–	-1.1
Param. Sharing	10.7%	8.8%	1.2x	1.2x	84.3	89.6	90.3	89.3	-0.58
Pruning with KD	40.0%	37.3%	1x	1x	83.5	88.9	92.8	–	-0.7
	31.2%	12.4%	5.9x	8.7x	82.0	90.4	92.0	–	-0.96
Quantization with KD	7.6%	3.9%	1.94x	1.73x	82.0	–	–	–	-2.6
	5.7%	6.1%	3.9x	4.7x	83.3	–	92.6	90.0	-0.23
Compound	1.3%	0.9%	1.83x	–	84.4	89.8	88.5	–	-1.53

- ➢ Quantization is the best single compression method for accuracy-size trade-off
- ➢ Unstructured pruning can reduce size without accuracy drop, but fails for extreme compression
- ➢ Model agnostic distillation allows training BiLSTM and CNN student models for tremendous speedup
- ➢ Combining multiple distillation methods, specially attention distillation, can improve performance
- ➢ Pruning and Quantization can be guided using distillation for better accuracy
- ➢ Compounding multiple compression methods together can help with extreme compression at minimal drop in accuracy

Practical Suggestions

- ➢ Choose an appropriate baseline based on the downstream task requirement
- ➢ Use specialised hardware and accelerators
- ➢ Investigate the target setup
  - Choose a compression method based on the acceleration requirement, i.e., accelerator characteristics
  - Choose an appropriate student model
- ➢ Compound different compression methods
  - Combine various BERT-specific methods
  - Use knowledge distillation as a guide for other forms of compression