E.T.: Re-Thinking Self-Attention for Transformer Models on GPUs

Shiyang Chen[1], Shaoyi Huang[2], Santosh Pandey[1], Guang R. Gao[3], Long Zheng[3], Caiwen Ding[2], Hang Liu[1]
Stevens Institute of Technology[1]
University of Connecticut[2]
University of Delaware[3]

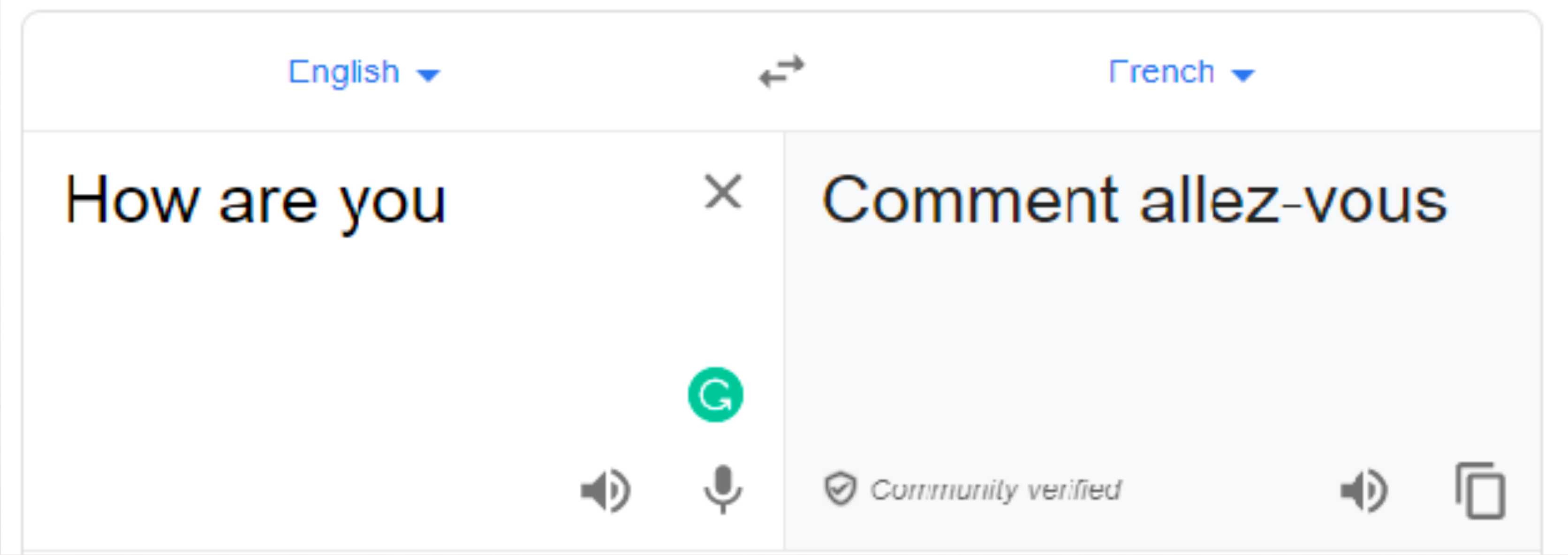
December 9, 2021

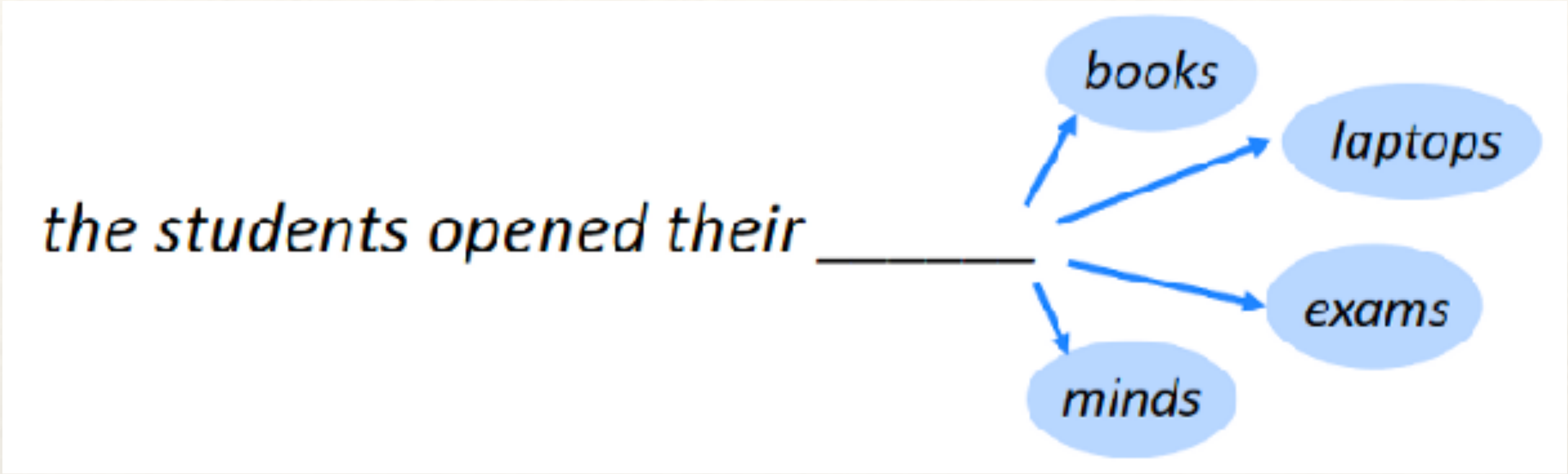
Outline

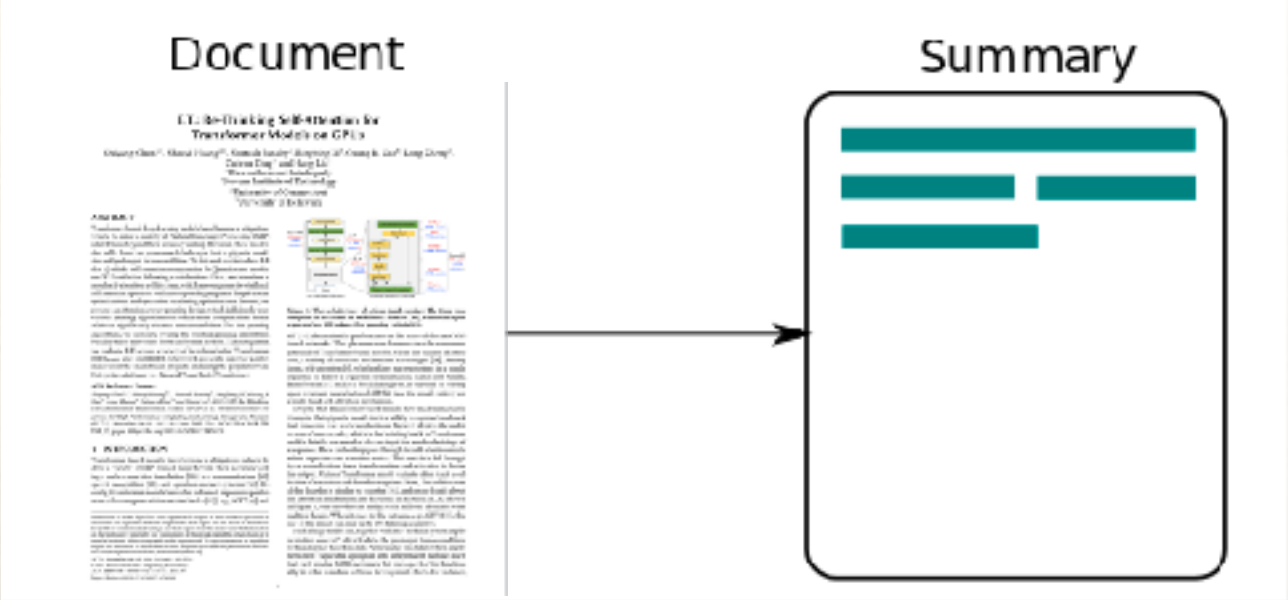
- Motivation
- Challenge #1: Long turnaround time
- Challenge #2: Gigantic model size
- Technique #1: Self-Attention Primitives
- Technique #2: Attention-Aware, Tensor-Core Friendly Pruning
- Evaluation
- Conclusion

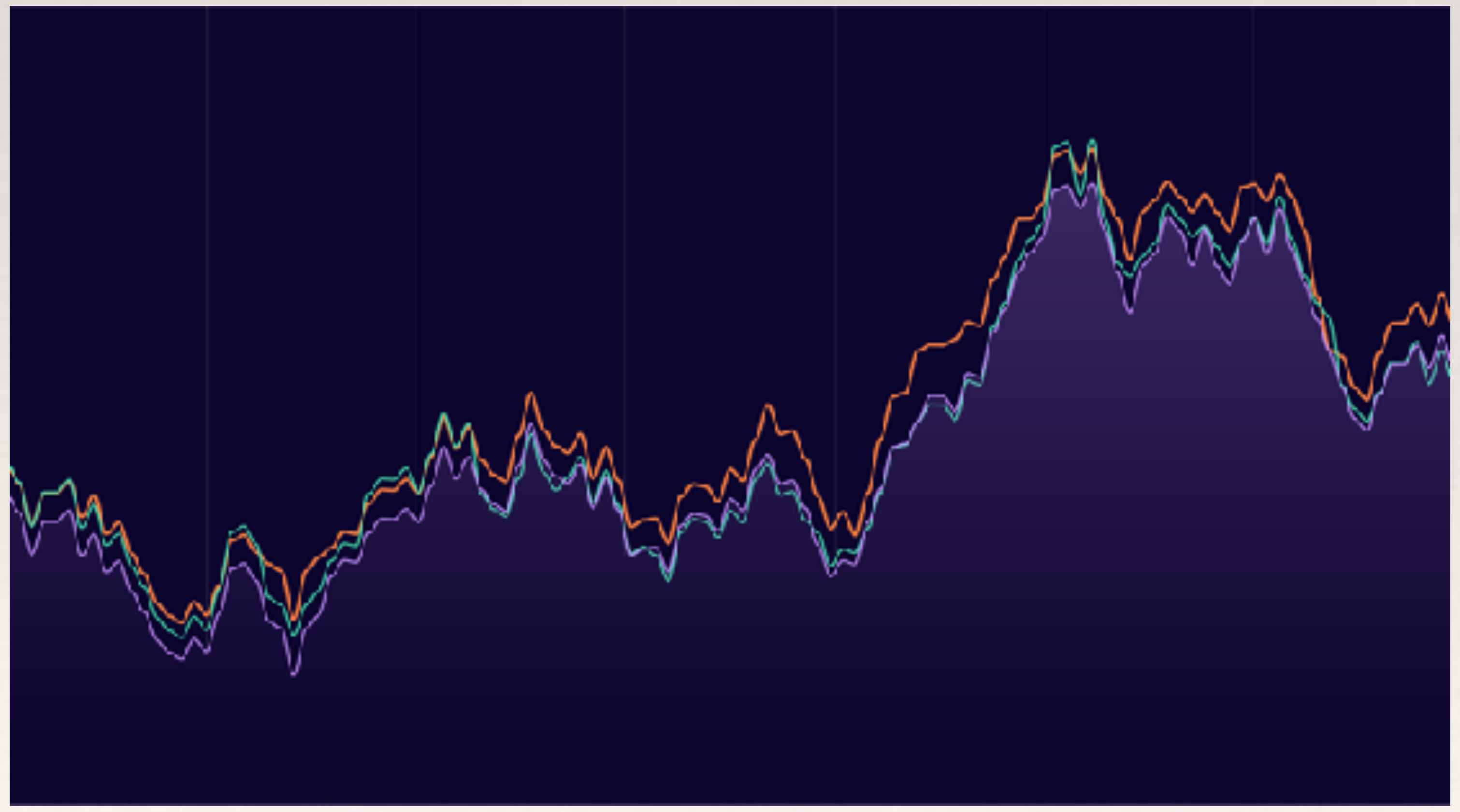
STEVENS INSTITUTE of TECHNOLOGY THE INNOVATION UNIVERSITY®

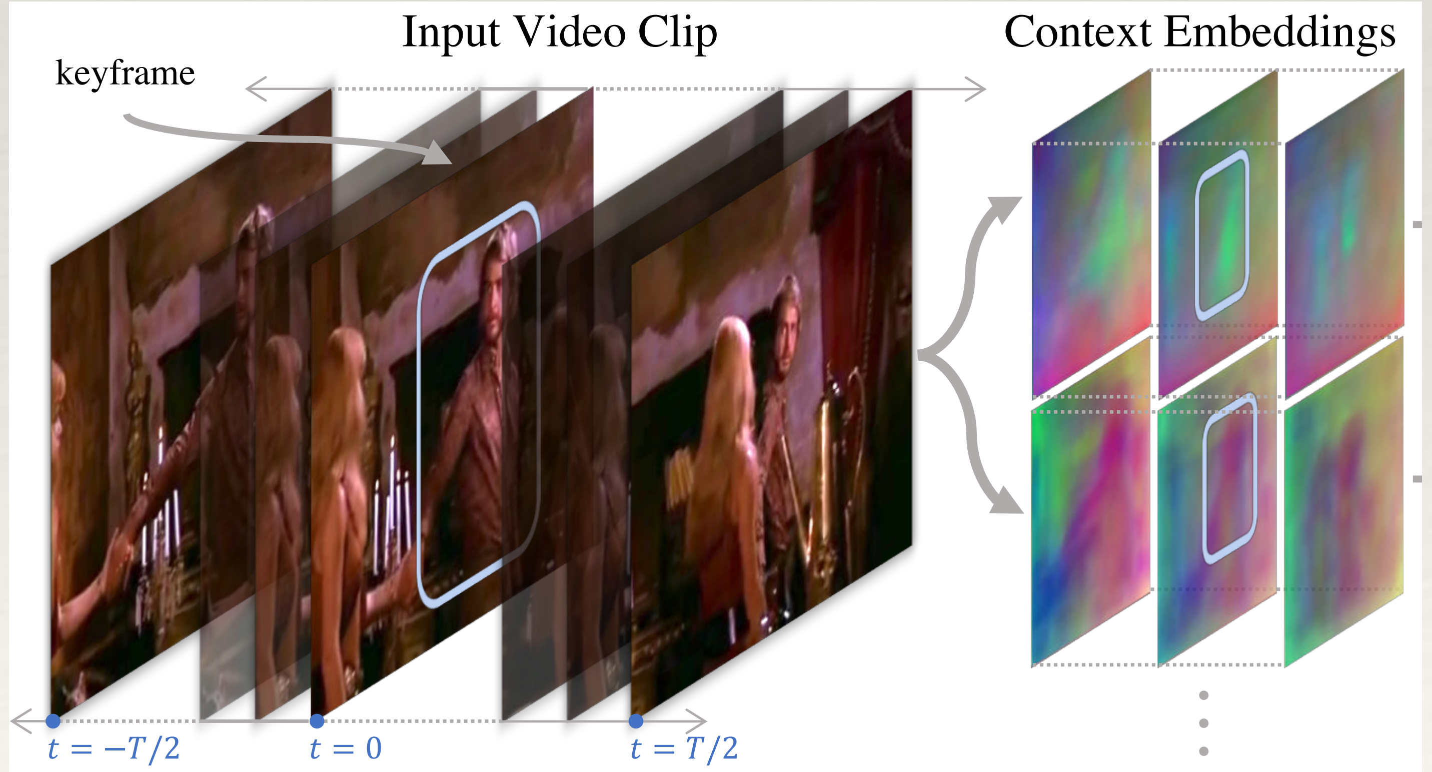
Motivation: Sequence-based Problems is Everywhere …







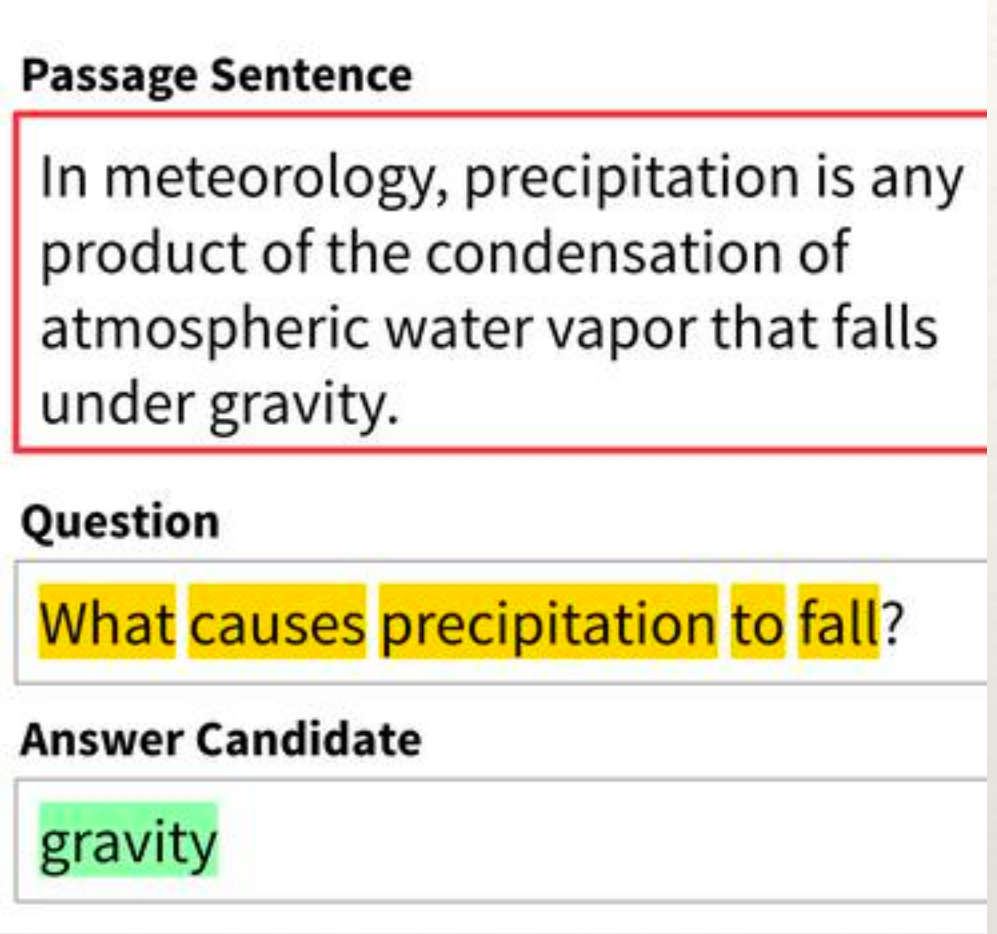




[1]. Girdhar, Rohit, et al. "Video action transformer network." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

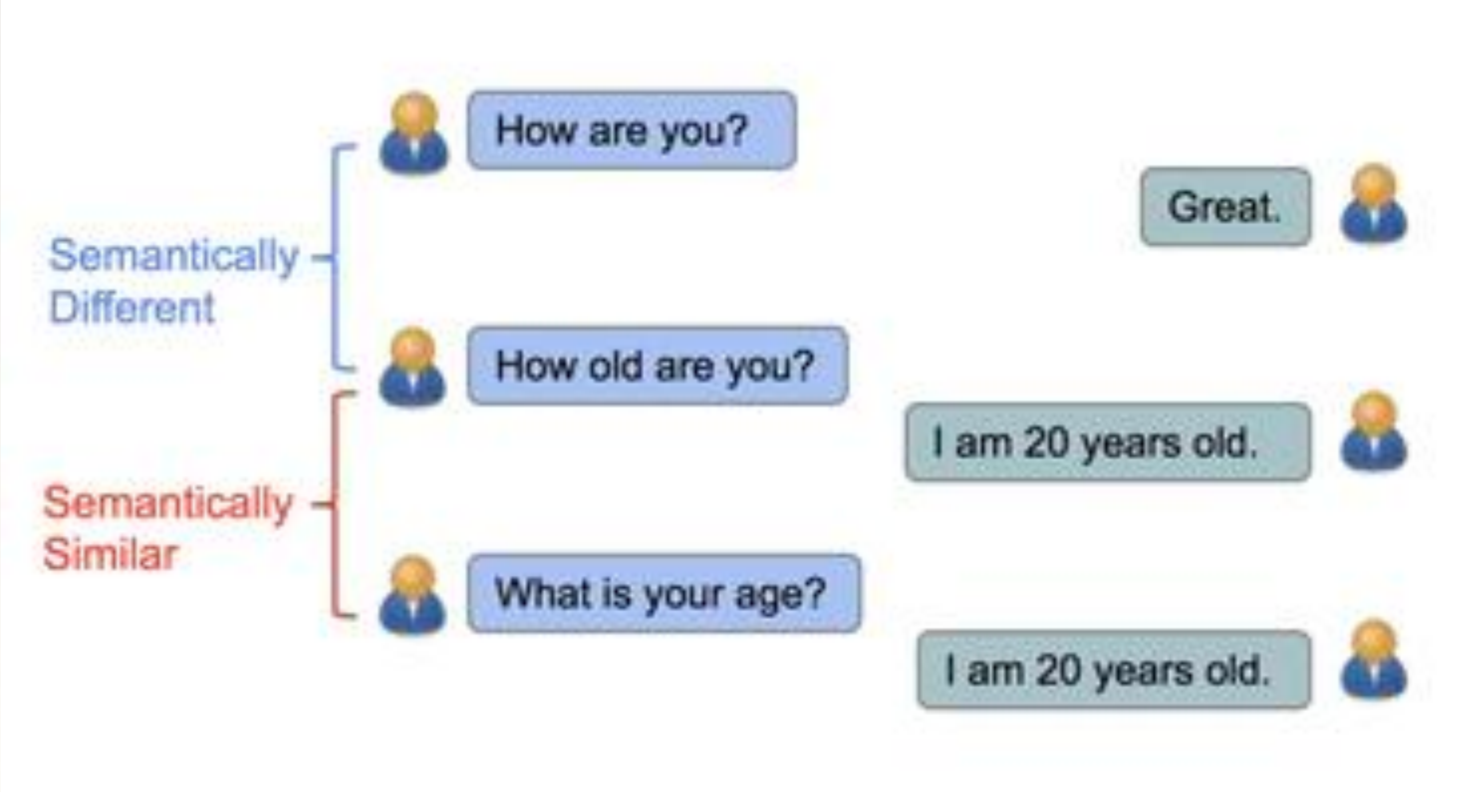
STEVENS INSTITUTE of TECHNOLOGY THE INNOVATION UNIVERSITY®

Attention is All You Need!



Question Answering on SQuAD2.0 dev

Rank	Model	F1	EM	Extra Training Data	Paper	Code	Result	Year
1	XLNet (single model)	90.6	87.9	×	XLNet: Generalized Autoregressive Pretraining for Language Understanding			2019
2	XLNet+DSC	89.51	87.65	×	Dice Loss for Data-imbalanced NLP Tasks			2019
3	RoBERTa (no data aug)	89.4	86.5	✓	RoBERTa: A Robustly Optimized BERT Pretraining Approach			2019

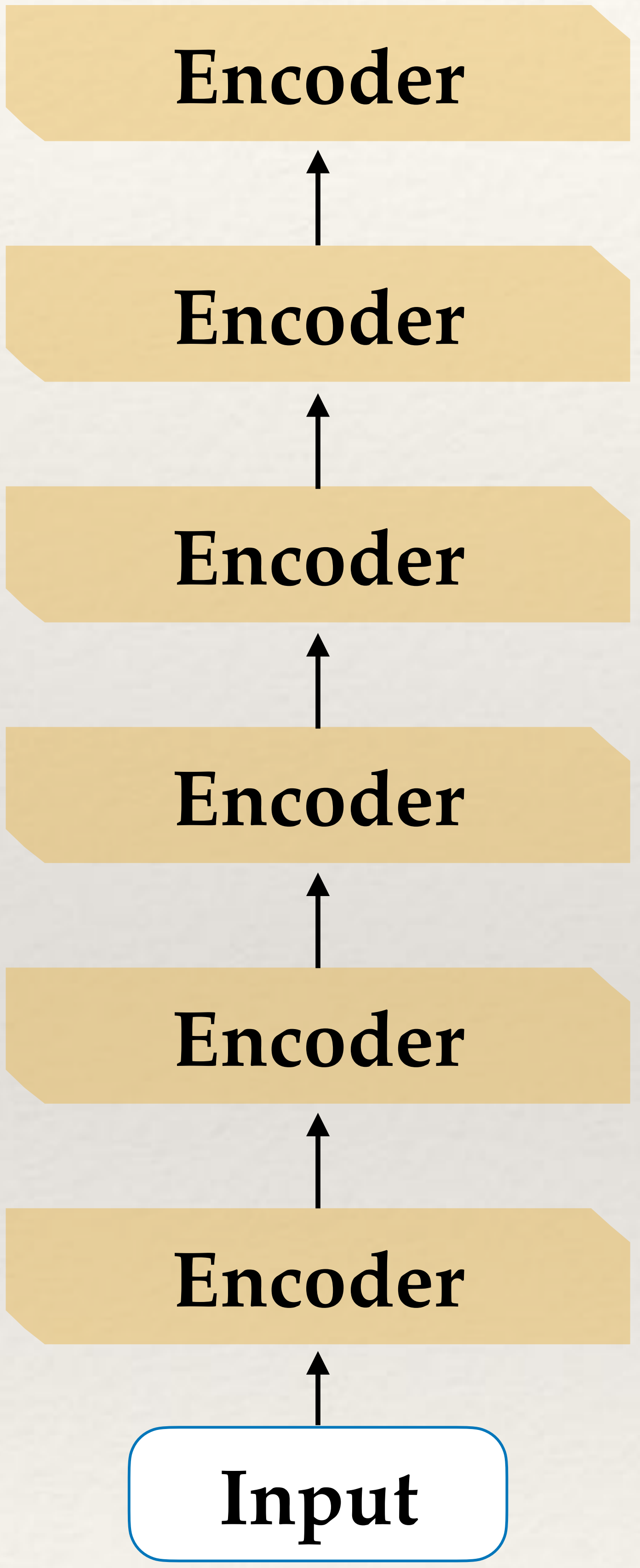


Semantic Textual Similarity on MRPC

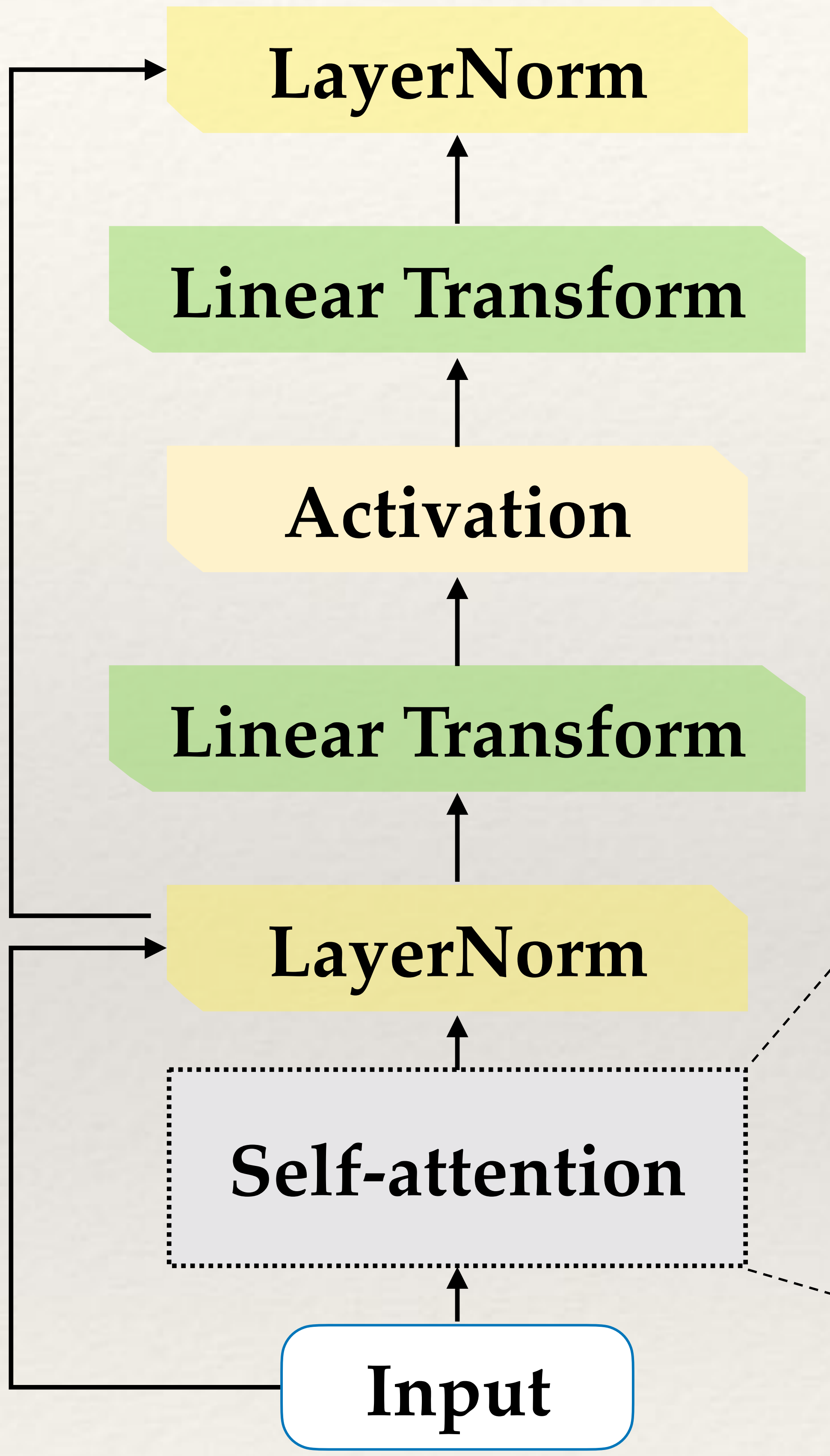
Rank	Model	Accuracy	F1	Paper	Code	Result	Year
1	SMART-RoBERTa Large	93.7%		SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization			2019
2	ALBERT	93.4%		ALBERT: A Lite BERT for Self-supervised Learning of Language Representations			2019
3	RoBERTa	92.3%		RoBERTa: A Robustly Optimized BERT Pretraining Approach			2019

[1]. https://paperswithcode.com/sota/question-answering-on-squad20-dev
[2]. https://paperswithcode.com/sota/semantic-textual-similarity-on-mrpc

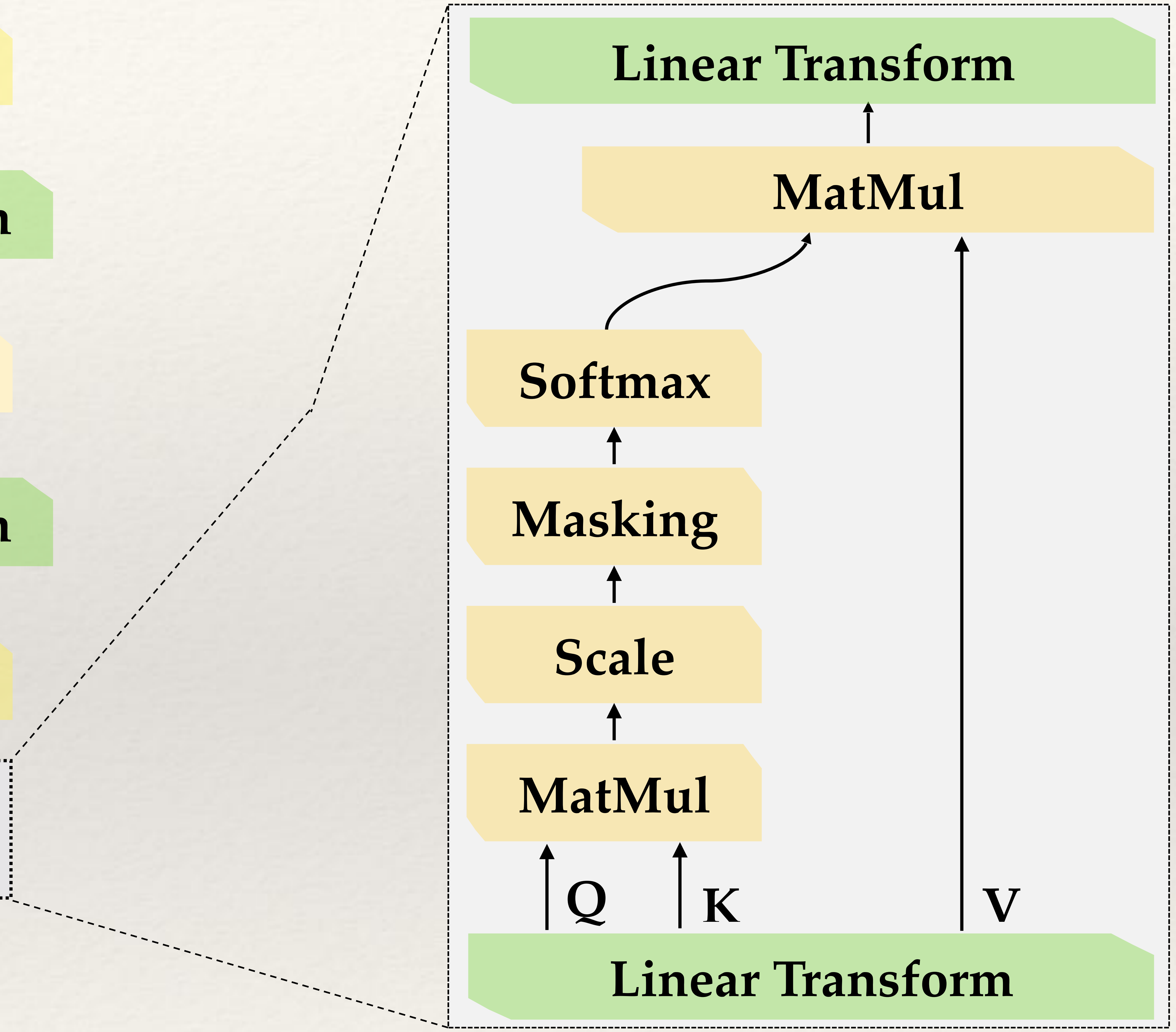
Challenge #1. Long Turnaround Time



Model workflow

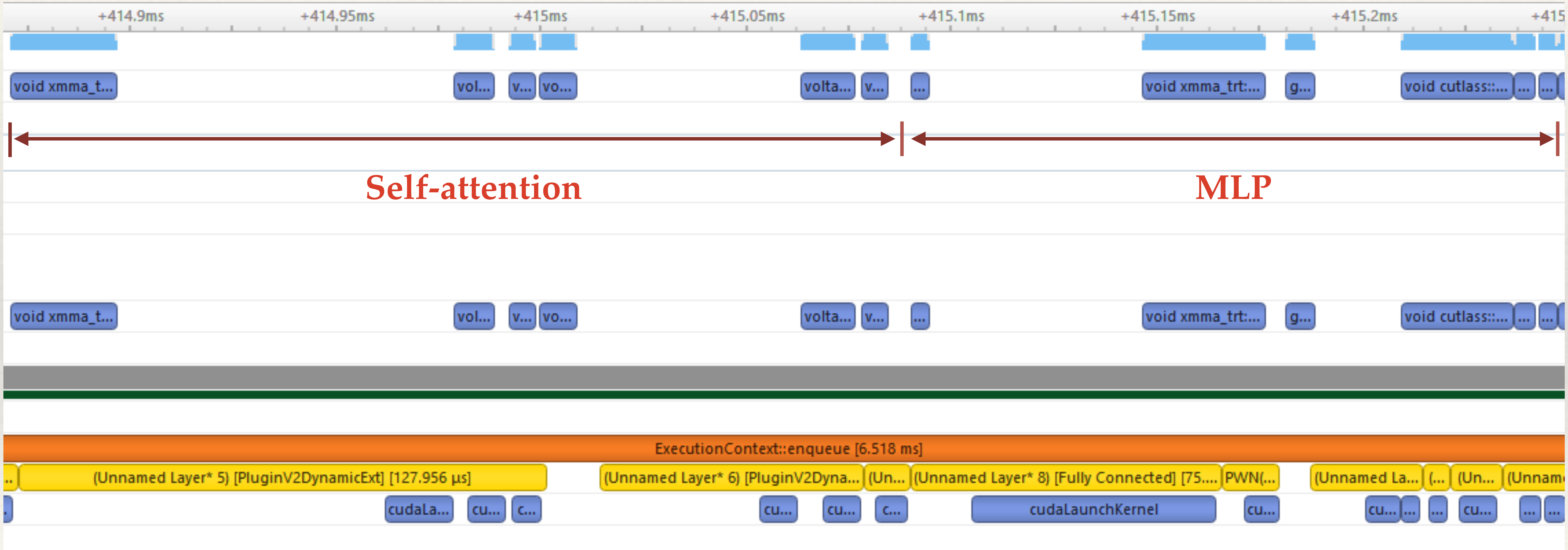


Encoder workflow



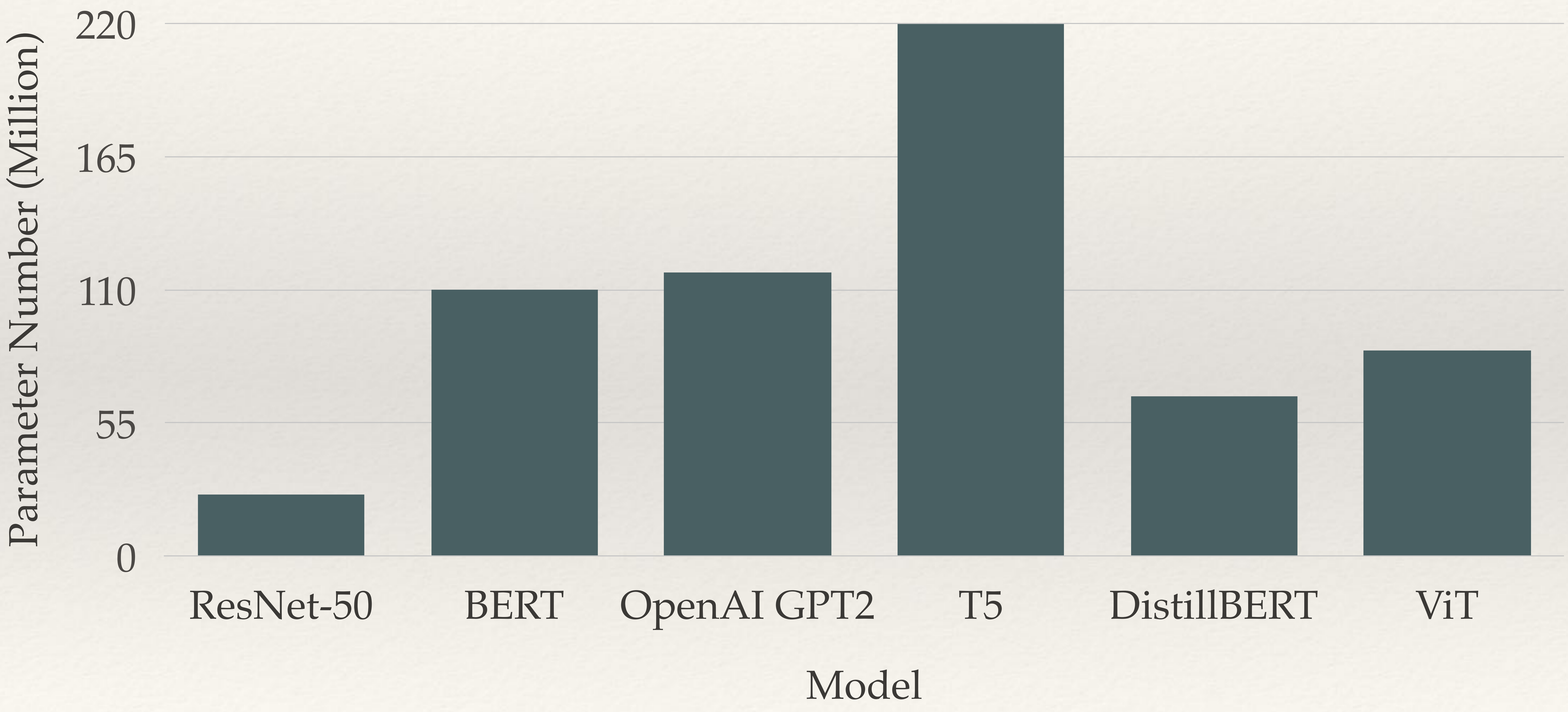
Self-attention workflow

Challenge #1. Long Turnaround Time

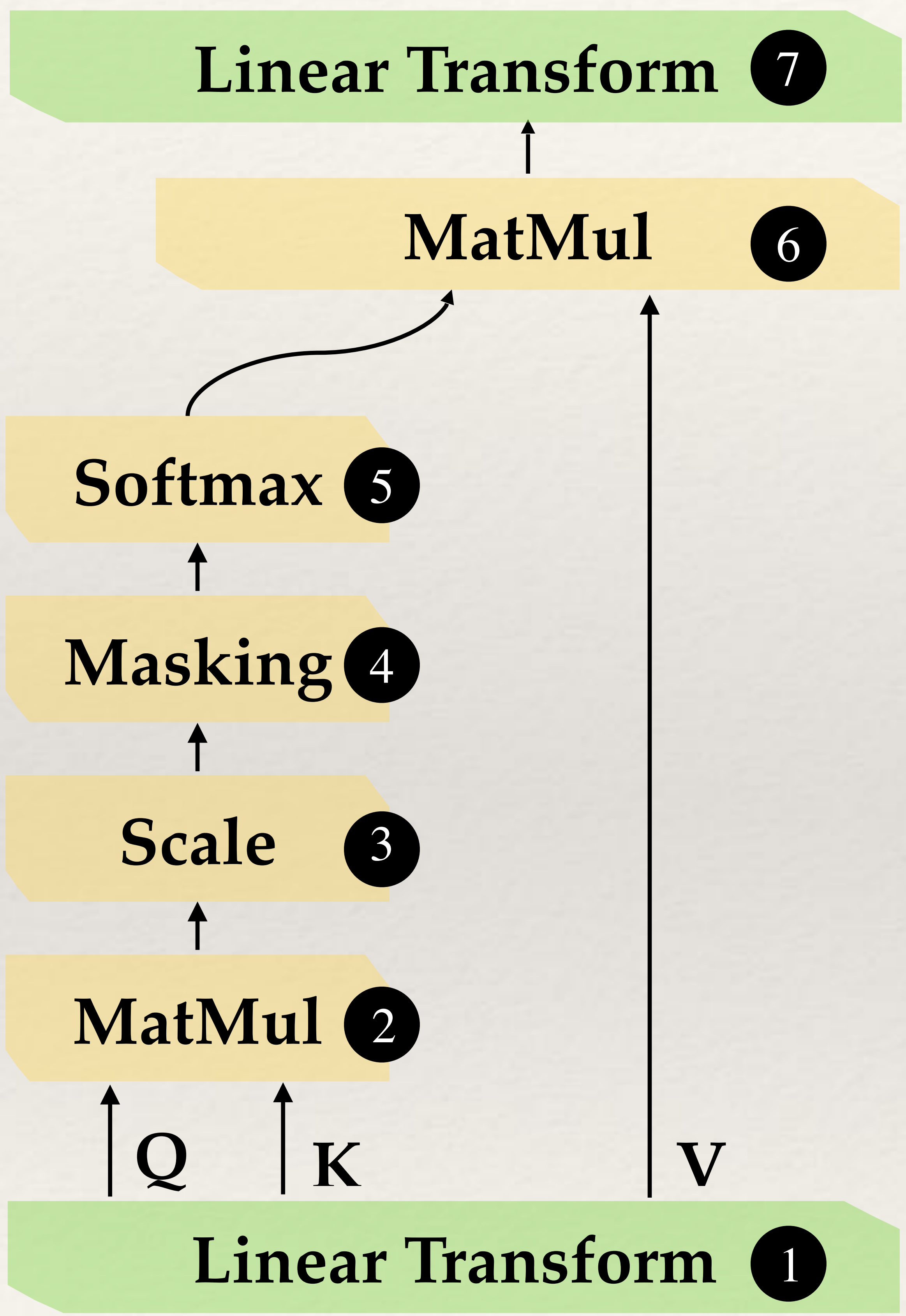


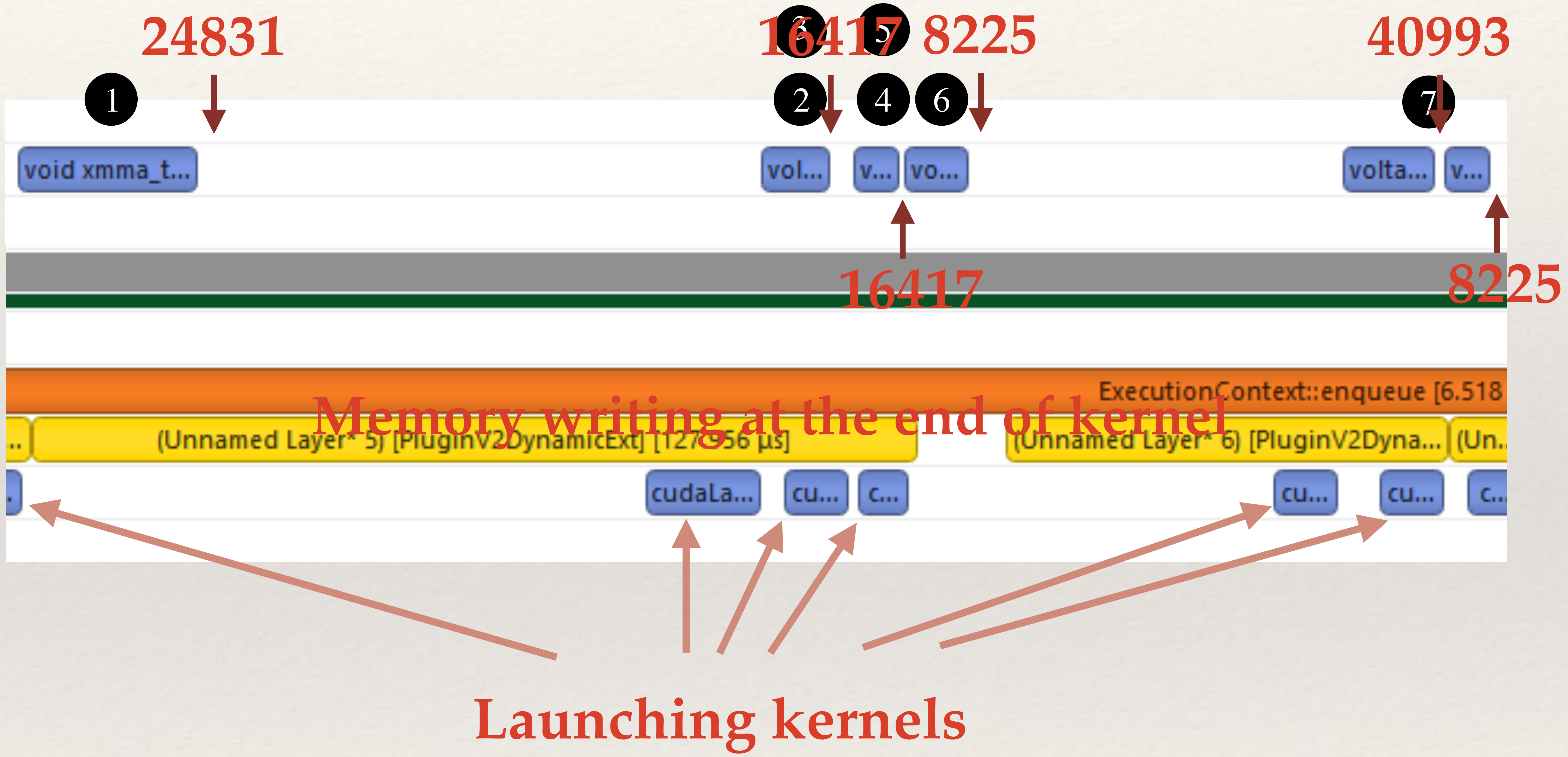
[1]. https://github.com/NVIDIA/TensorRT/tree/master/demo/BERT

Challenge #2. Gigantic model size

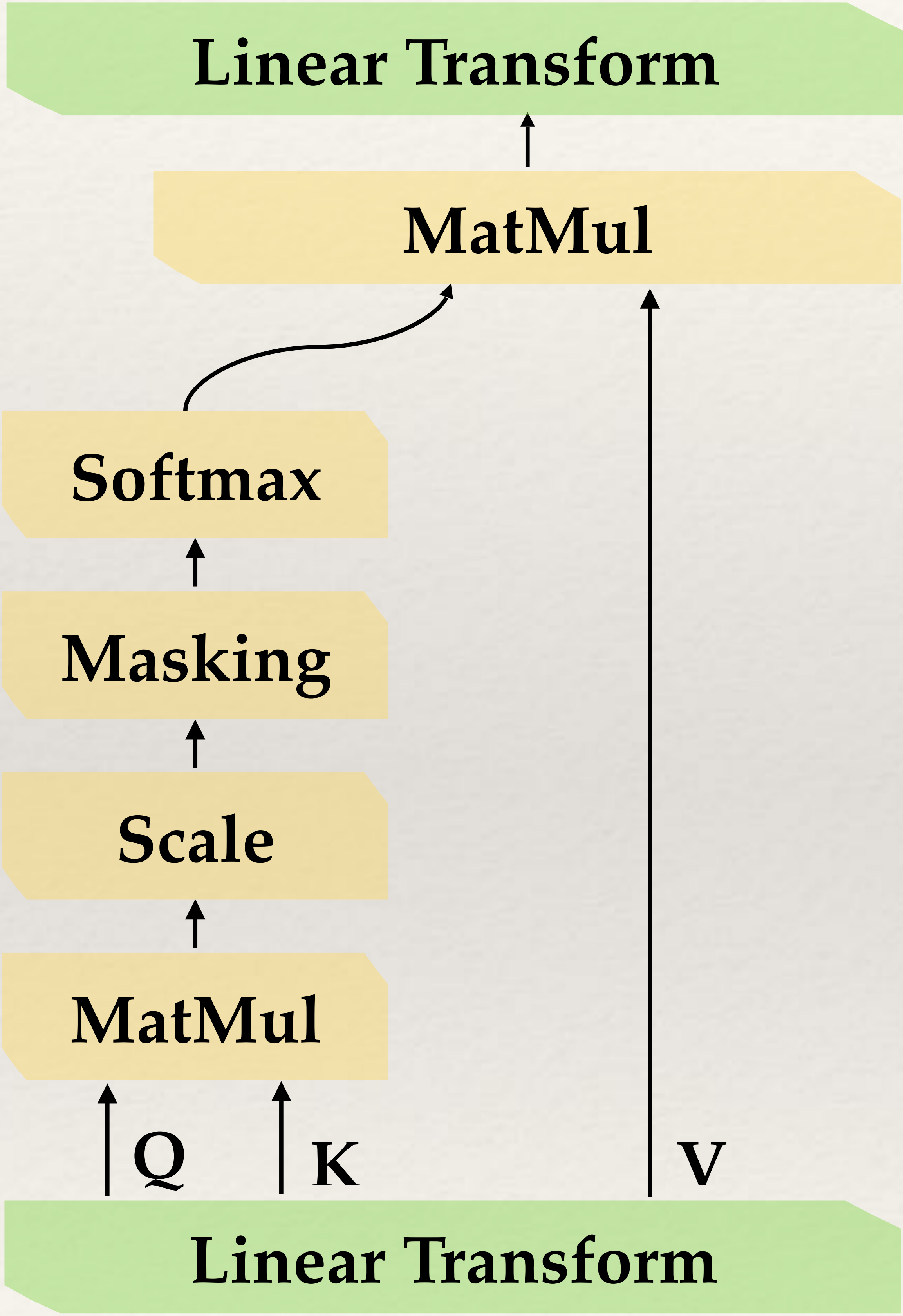


Kernels are not free





Think self-attention as a primitive

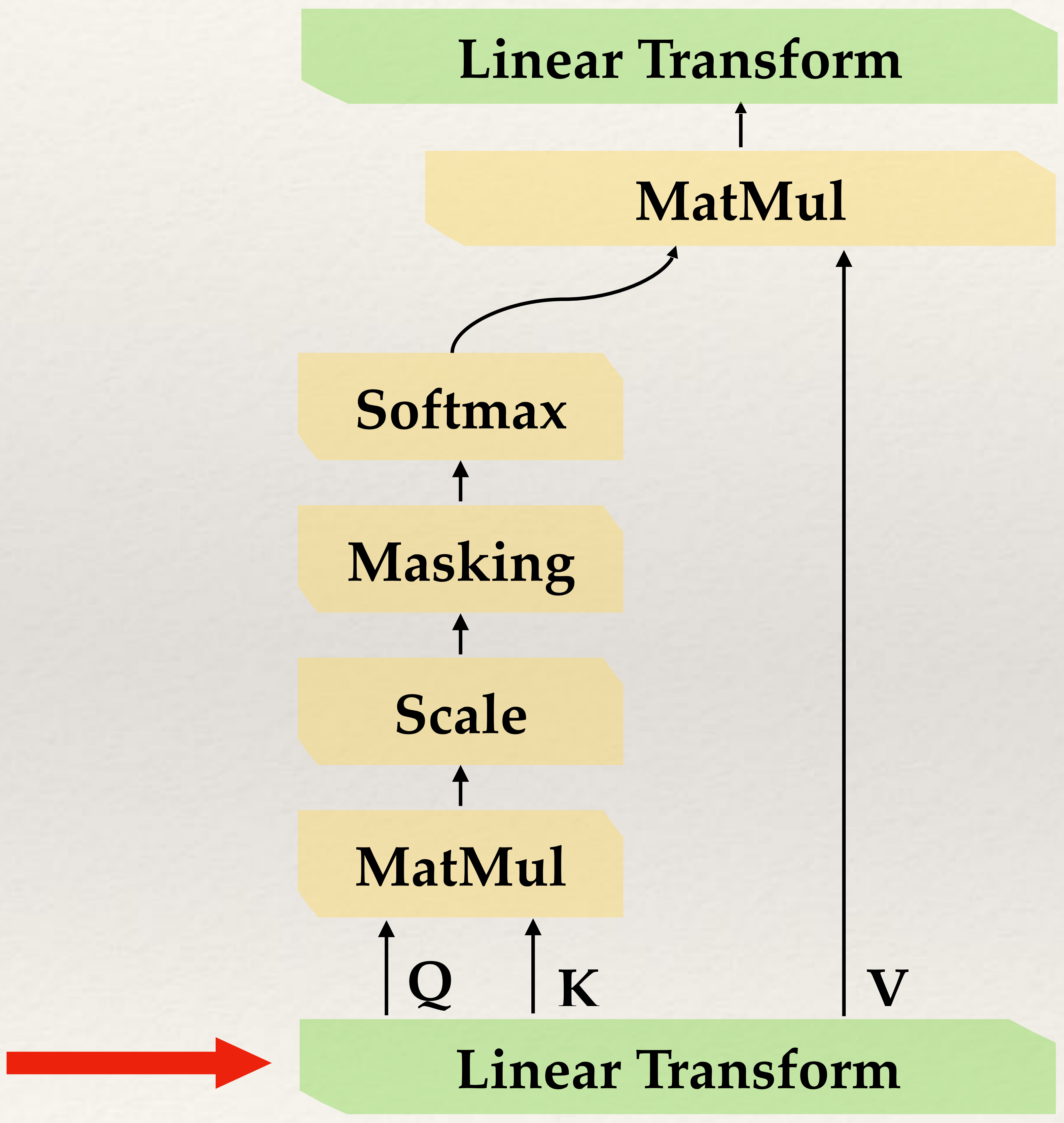


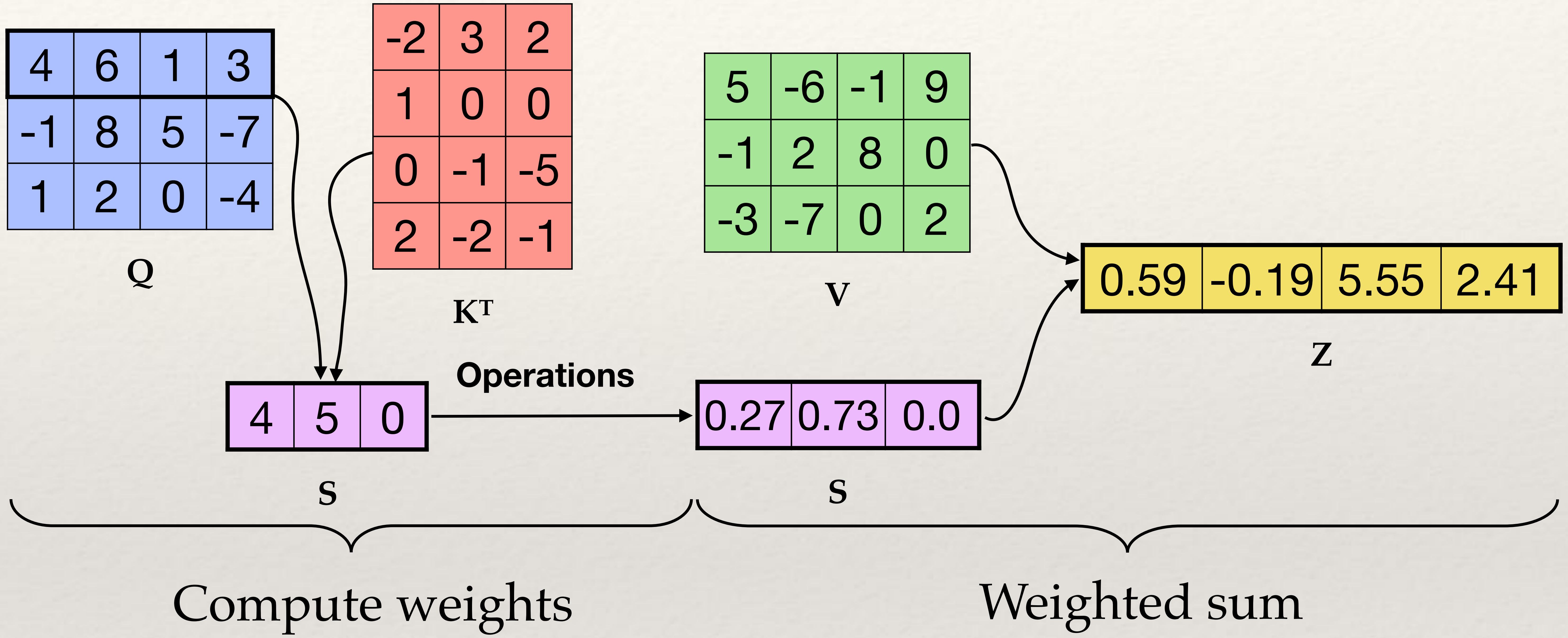
How are you →

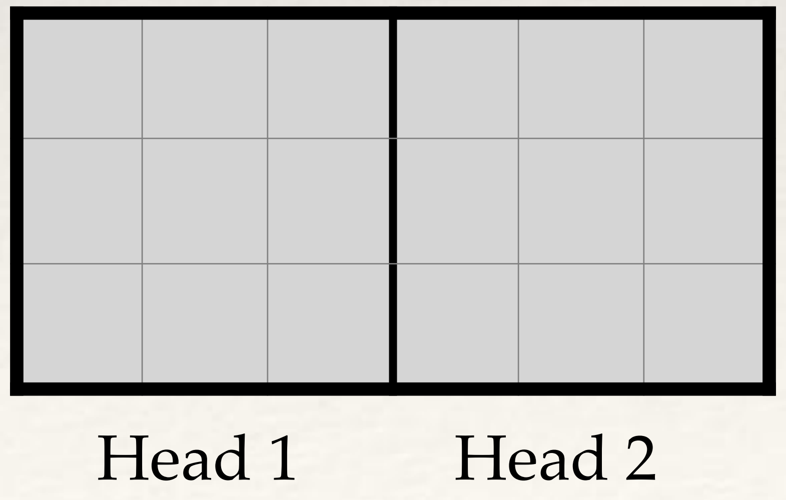
how	5	2	1	4
are	-3	-1	0	0
you	1	-2	-1	-1

Input

Think self-attention as a primitive

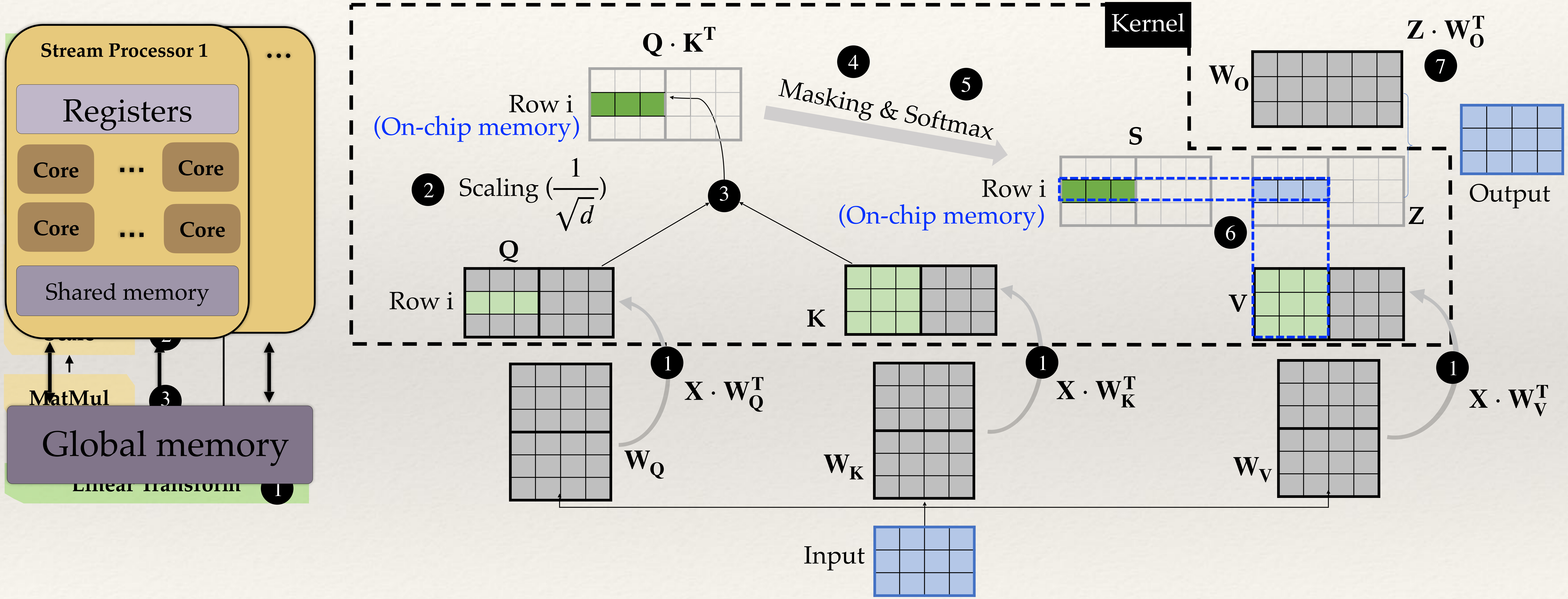




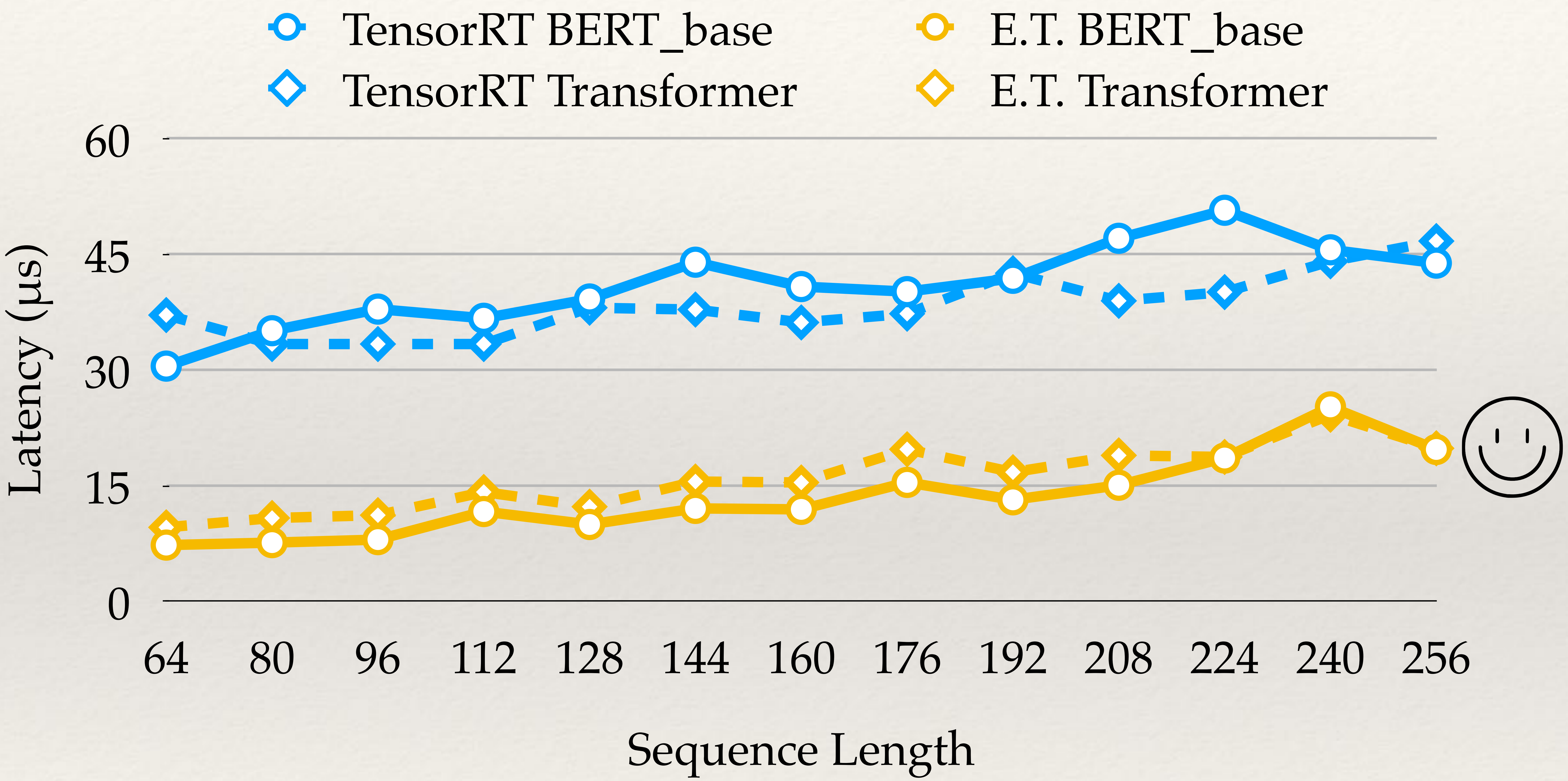


STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Compute the self-attention on-the-fly



Evaluate on-the-fly-attention



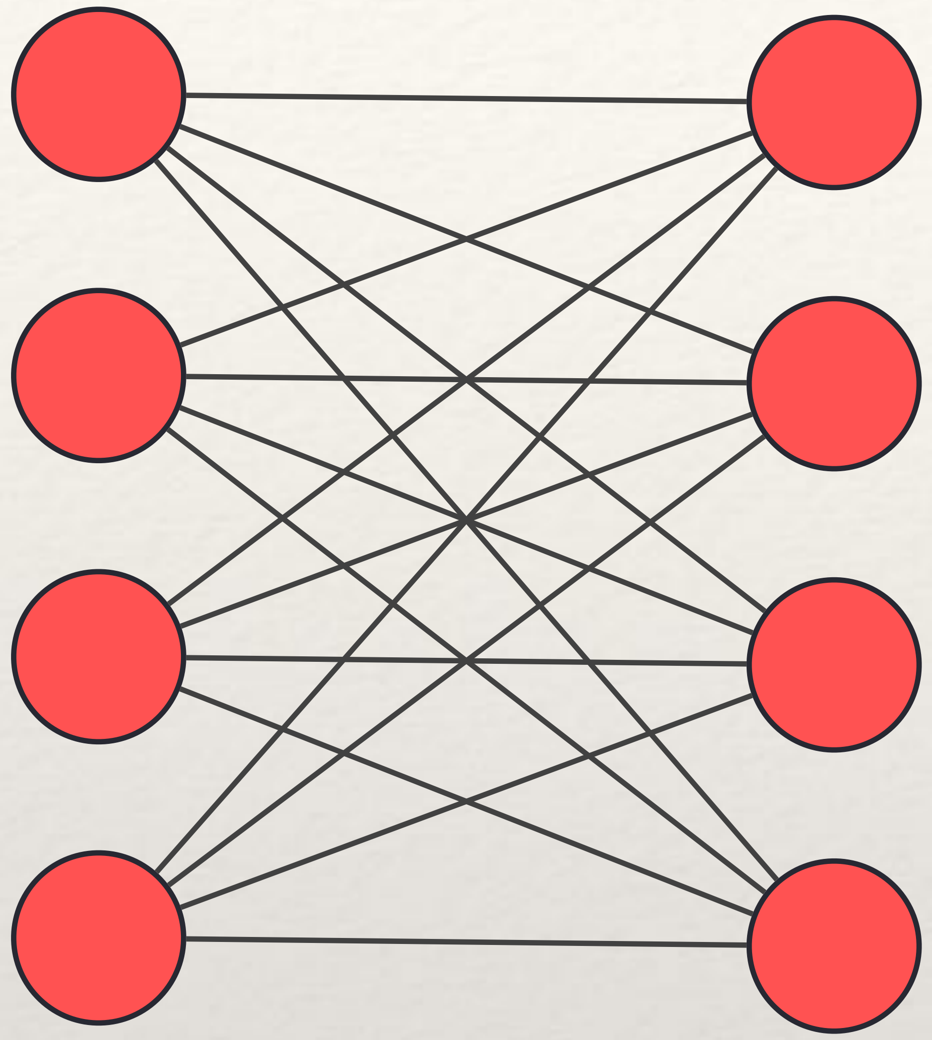
BERT_base:

- Model Size: 768
- Number of heads: 12

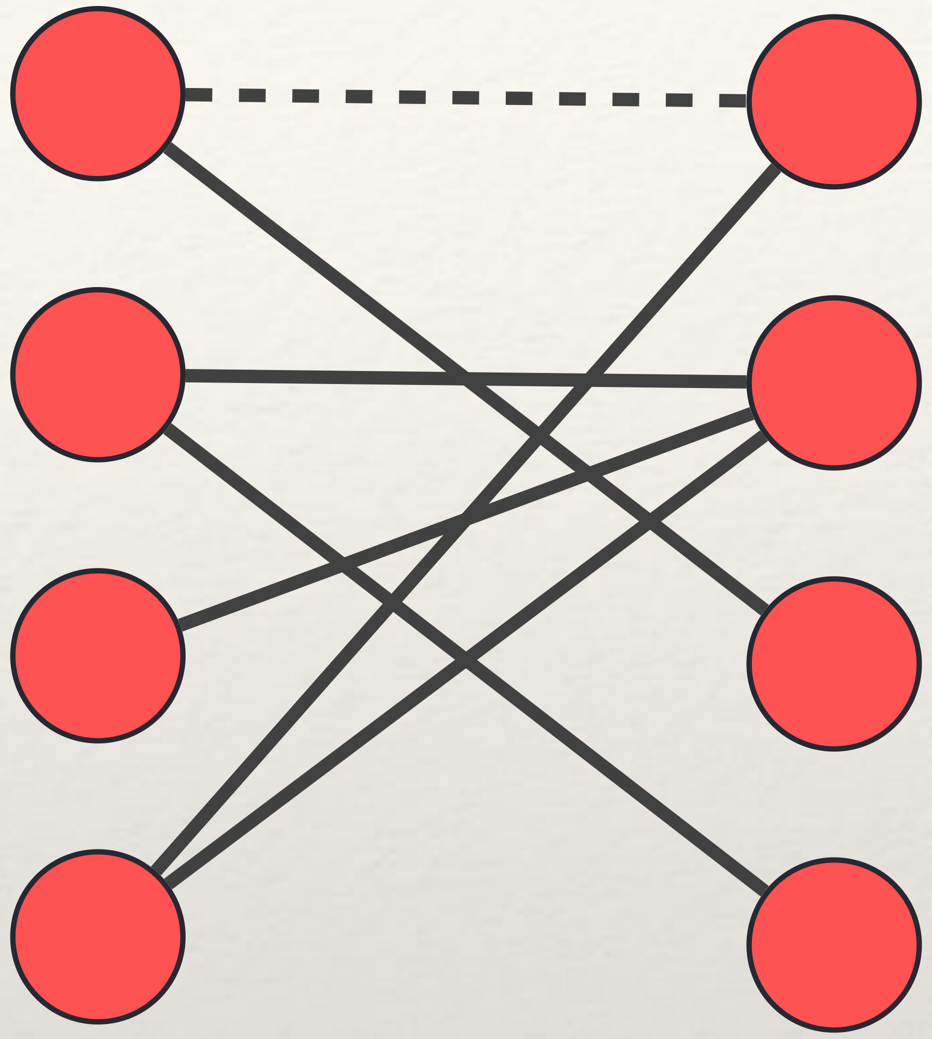
Transformer:

- Model Size: 800
- Number of heads: 4

Pruning makes the model small



Pruning



Dense

4	3	6	9
8	7	6	2
4	8	3	2
5	2	4	9

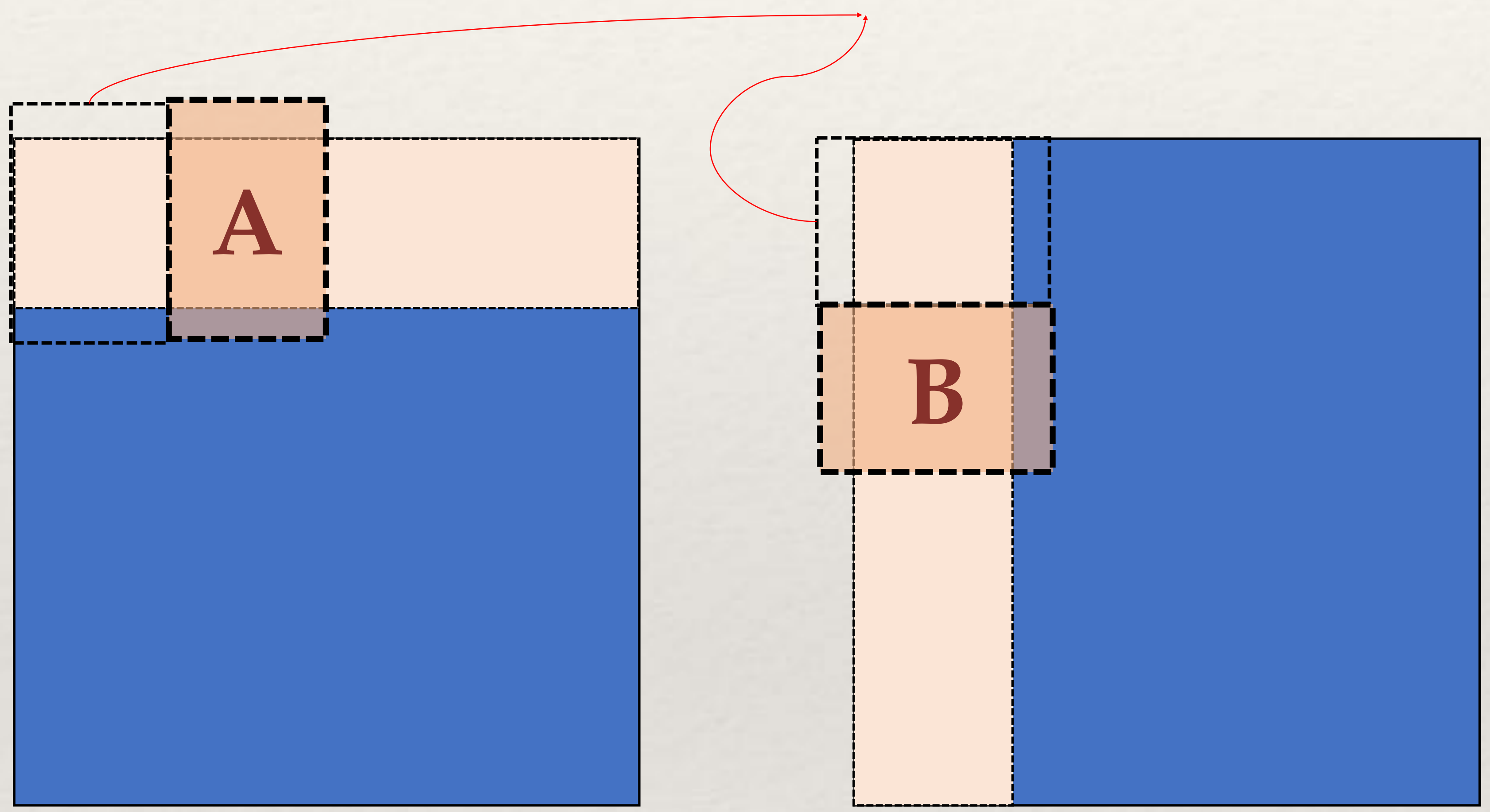
Pruning

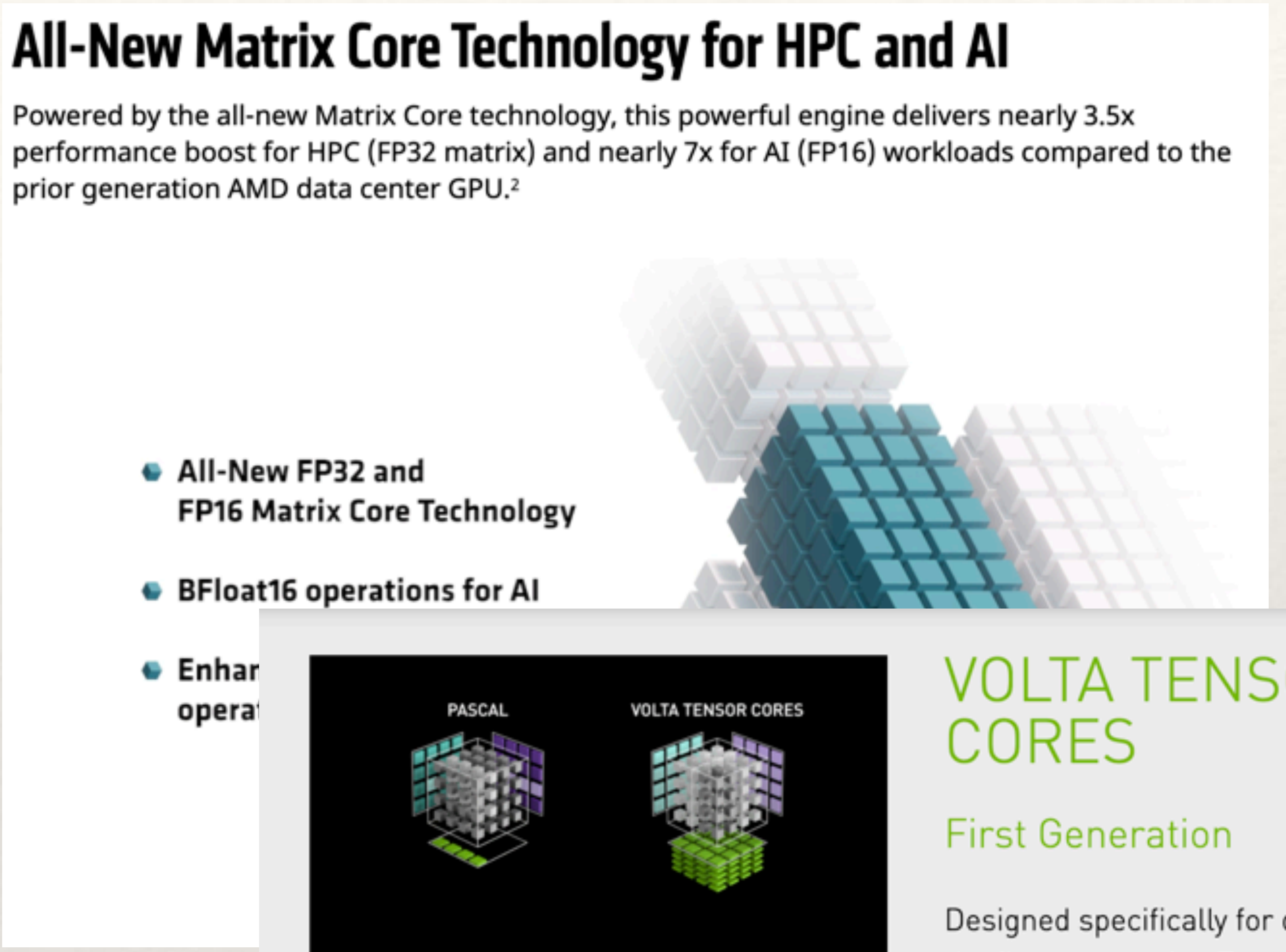
Sparse

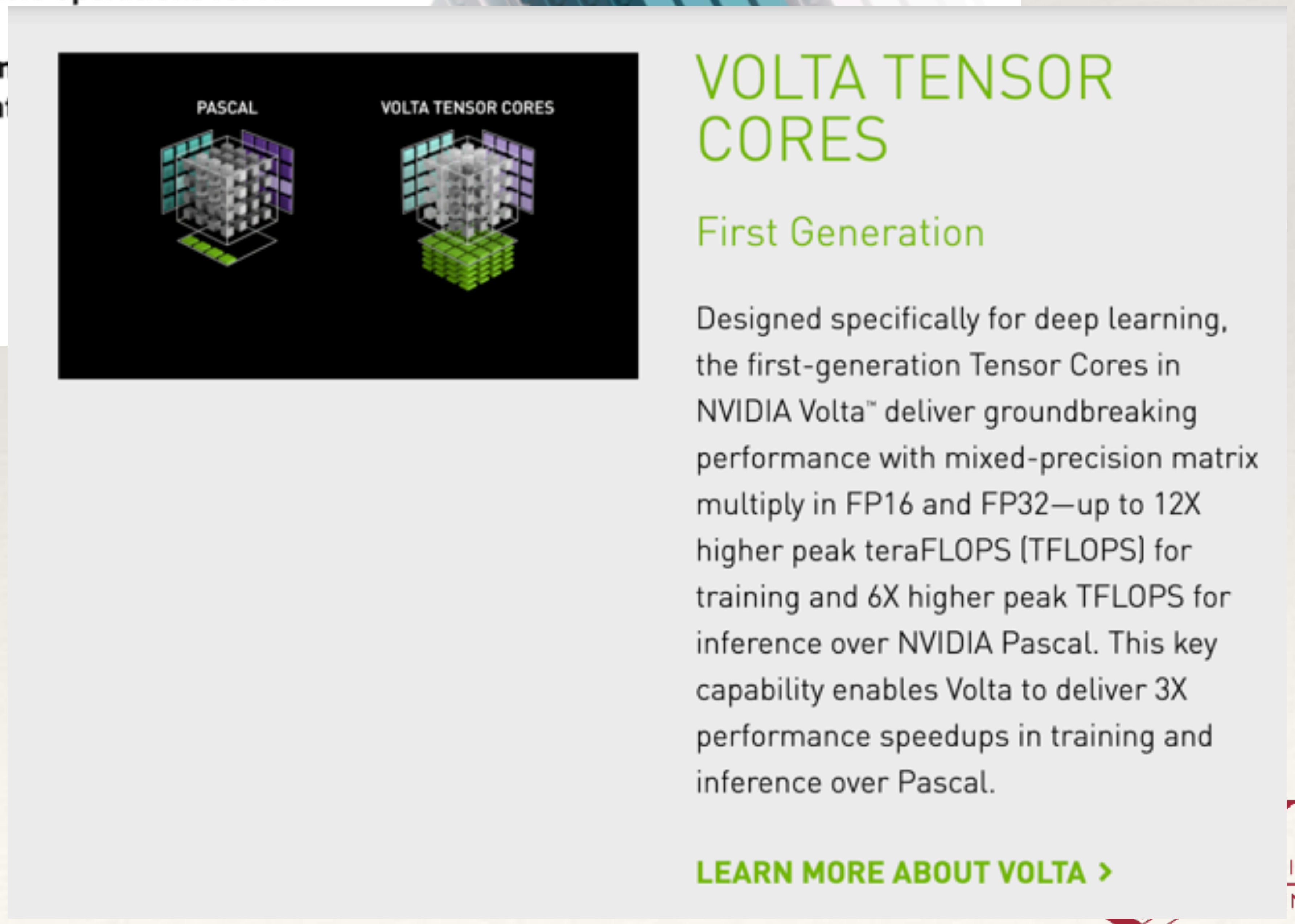
0	0	6	0
0	7	0	2
0	8	0	0
5	2	0	0

Using emerging hardware

$$C = A \cdot B + C$$







[1]. https://www.amd.com/en/technologies/cdna
[2]. https://www.nvidia.com/en-us/data-center/tensor-cores

Efficient computing on sparse models

Irregular

0	0	6	0
0	7	0	2
0	8	0	0
5	2	0	0

Row

4	3	6	9
0	0	0	0
4	8	3	2
0	0	0	0

Column

4	0	6	0
8	0	6	0
4	0	3	0
5	0	4	0

Tensor-tile

0	0	6	9
0	0	6	2
4	8	0	0
5	2	0	0

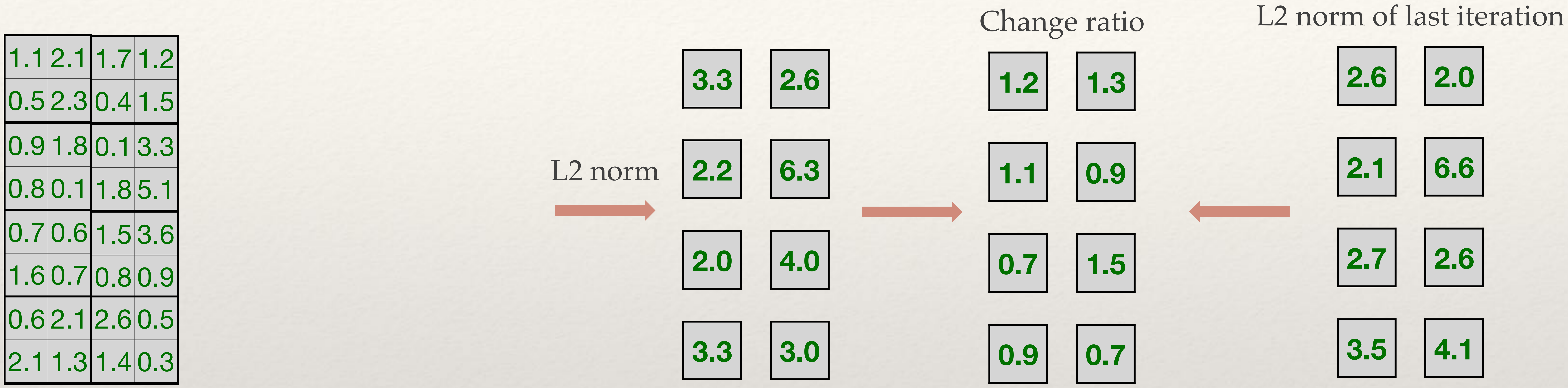
$$\mathbf{W} = \begin{bmatrix} 4 & 3 & 6 & 9 \\ 0 & 0 & 0 & 0 \\ 4 & 8 & 3 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \mathbf{W}_{\text{pruned}} = \begin{bmatrix} 4 & 3 & 6 & 9 \\ 4 & 8 & 3 & 2 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & -2 & 0 \\ -3 & 1 & 1 & 1 \\ 0 & -1 & 1 & 0 \\ 2 & 0 & -3 & 1 \end{bmatrix} \times \mathbf{W}^{T}_{\text{pruned}} = \begin{bmatrix} 4 & 4 \\ 3 & 8 \\ 6 & 3 \\ 9 & 2 \end{bmatrix} \rightarrow \begin{bmatrix} -8 & 0 & -2 & 0 \\ 6 & 0 & 1 & 0 \\ 3 & 0 & -5 & 0 \\ -1 & 0 & 1 & 0 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & -2 & 0 \\ -3 & 1 & 1 & 1 \\ 0 & -1 & 1 & 0 \\ 2 & 0 & -3 & 1 \end{bmatrix} \qquad \mathbf{W} = \begin{bmatrix} 4 & 0 & 6 & 0 \\ 8 & 0 & 6 & 0 \\ 4 & 0 & 3 & 0 \\ 5 & 0 & 4 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 4 & 6 \\ 8 & 6 \\ 4 & 3 \\ 5 & 4 \end{bmatrix}$$

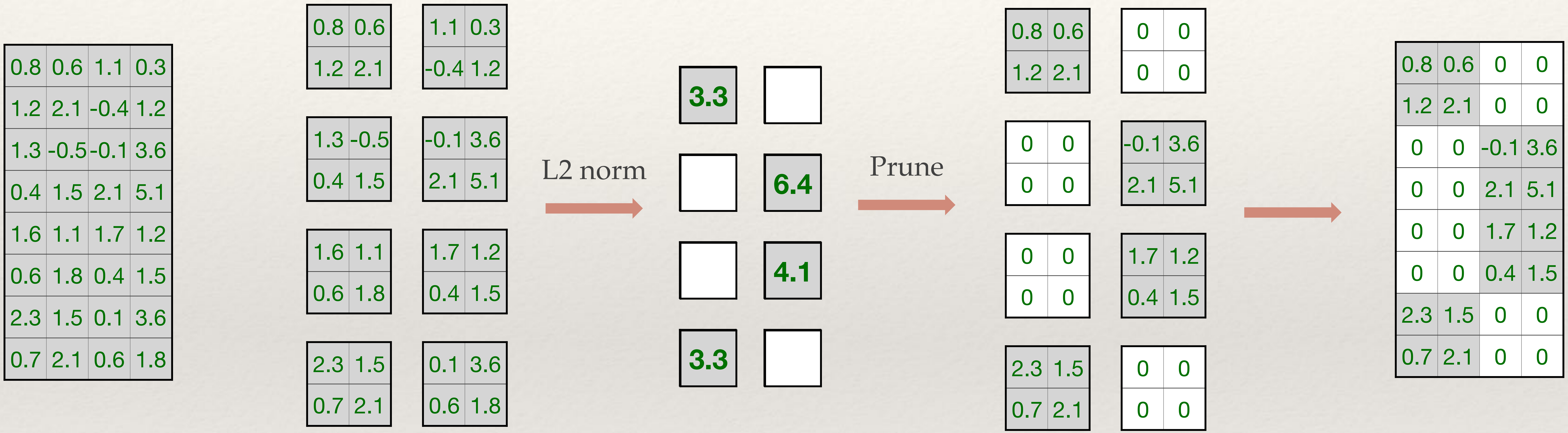
$$\mathbf{X}\text{ (adjusted)} = \begin{bmatrix} 1 & -2 \\ -3 & 1 \\ 0 & 1 \\ 2 & -3 \end{bmatrix} \times \mathbf{W}^{T}_{\text{pruned}} = \begin{bmatrix} 4 & 8 & 4 & 5 \\ 6 & 6 & 3 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -2 & 0 \\ -3 & 1 & 1 & 1 \\ 0 & -1 & 1 & 0 \\ 2 & 0 & -3 & 1 \end{bmatrix}$$

Prune the model as fine-tuning

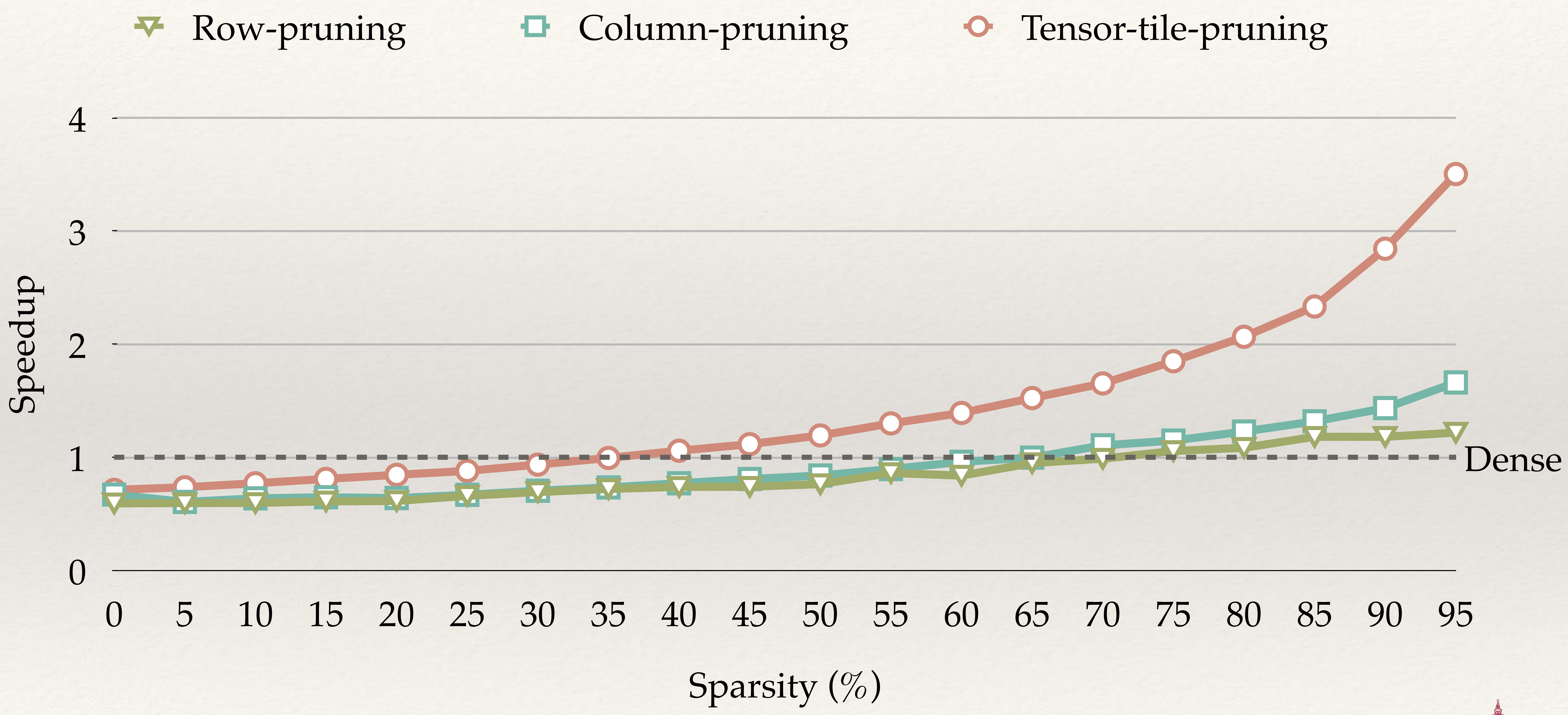


$$\min \quad \underbrace{f\left(\{W^k\}_{k=1}^N, \{b^k\}_{k=1}^N\right)}_{\text{Original loss}} + \lambda \underbrace{\sum_{k=1}^{N}\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\|W_{ij}^k\|_2}{\|W_{ij}^{k-1}\|_2 + \epsilon}}_{\text{Regularizer}}$$

Prune the model as fine-tuning



Performance gain from pruning



STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Efficient computing on sparse models

Irregular

0	0	6	0
0	7	0	2
0	8	0	0
5	2	0	0

Row

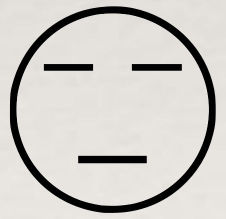
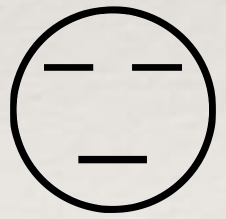
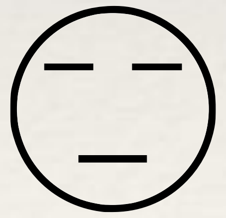
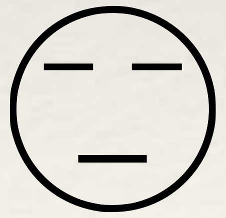
4	3	6	9
0	0	0	0
4	8	3	2
0	0	0	0

Column

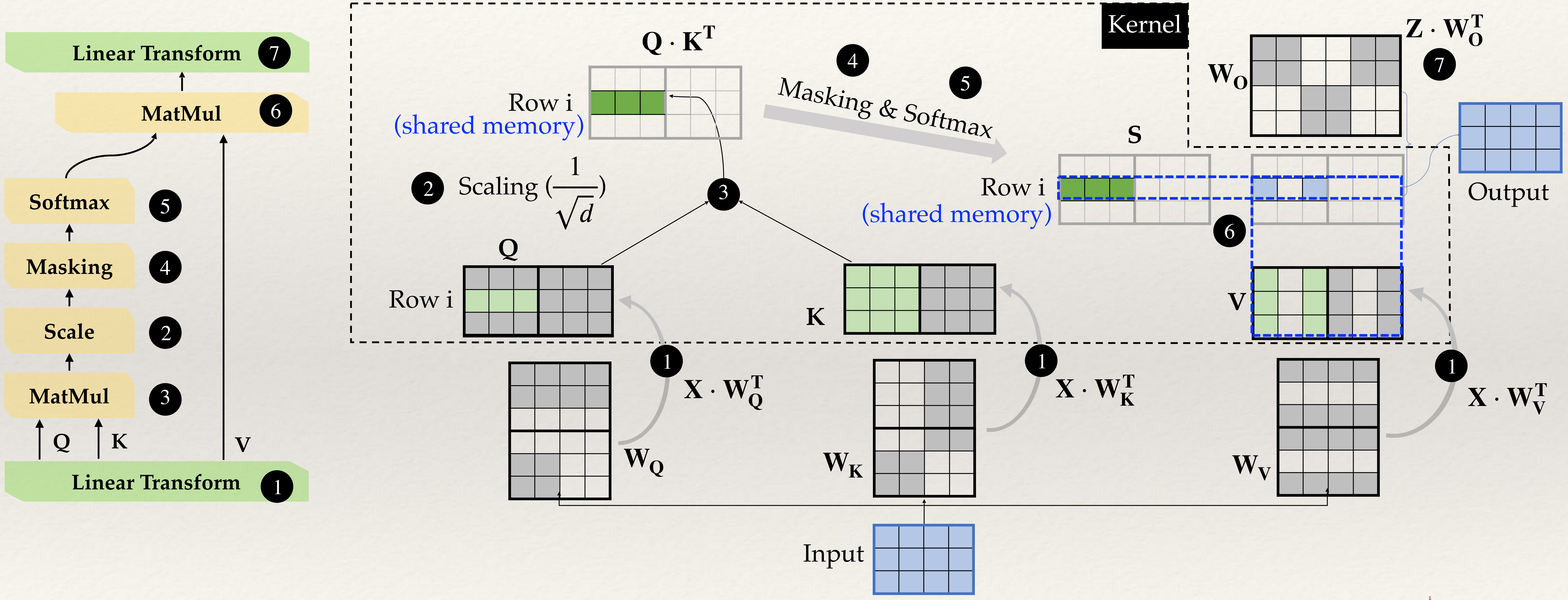
4	0	6	0
8	0	6	0
4	0	3	0
5	0	4	0

Tensor-tile

0	0	6	9
0	0	6	2
4	8	0	0
5	2	0	0

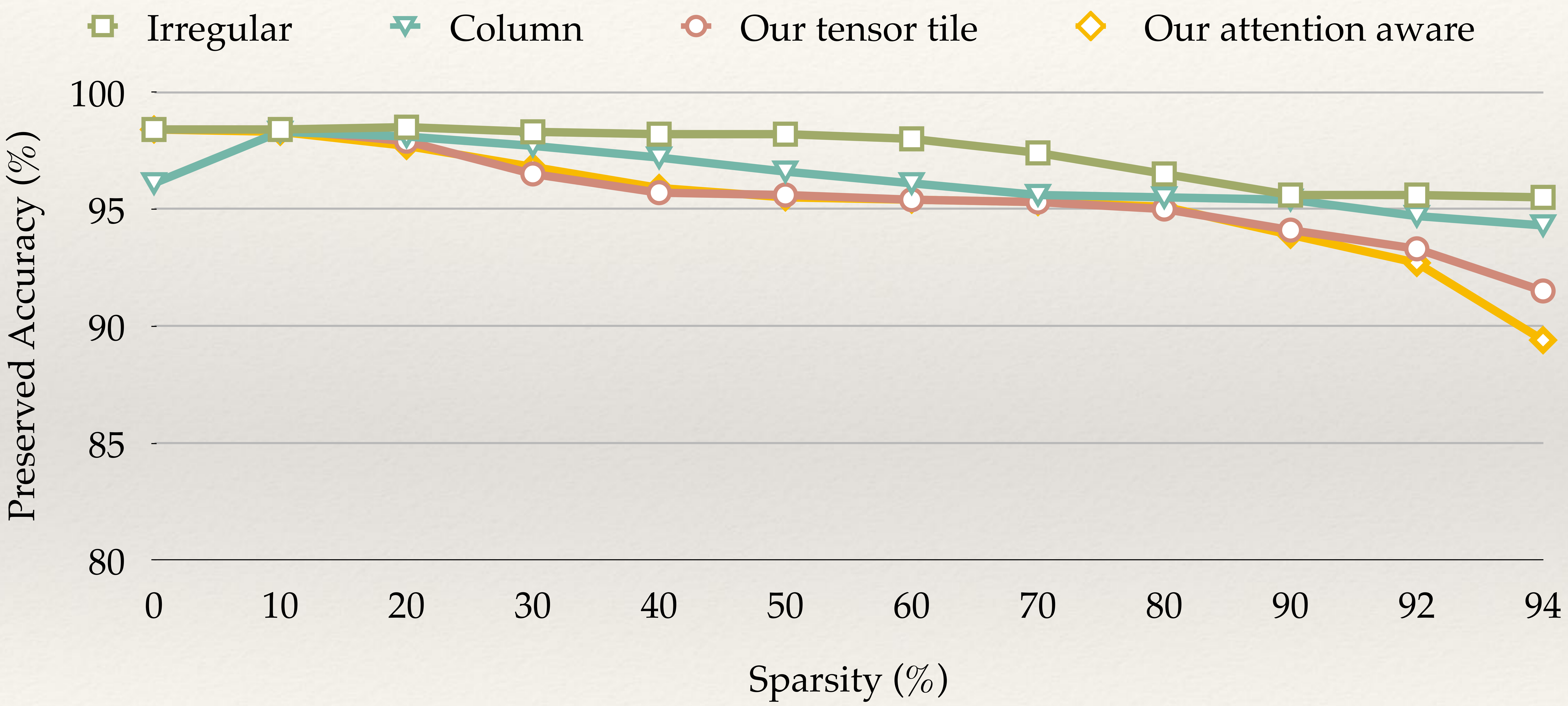
	Irregular	Row	Column	Tensor-tile
Accuracy				
Latency				

Attention-aware pruning



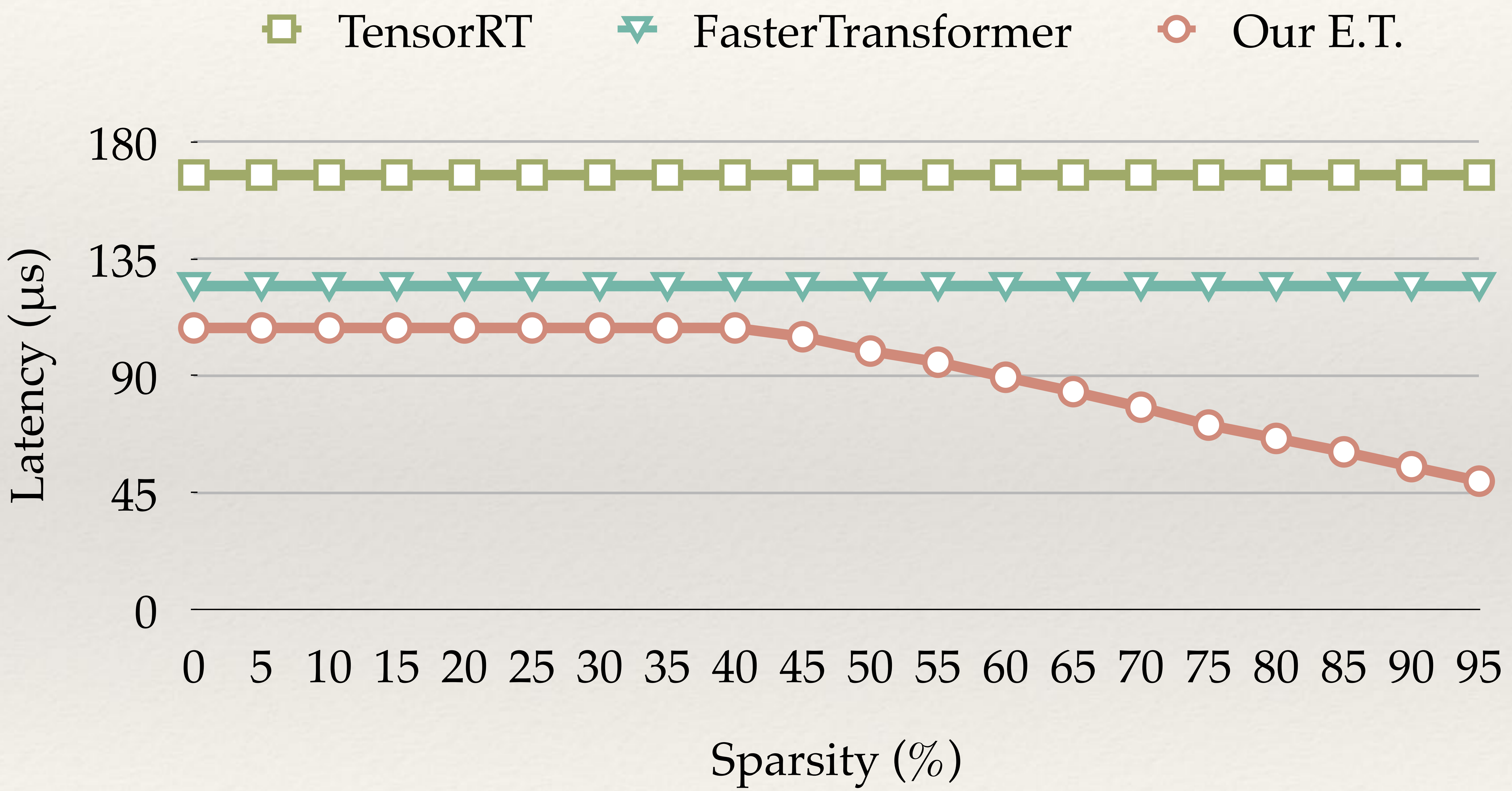
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Evaluate pruning algorithms



STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Compare with state-of-the-art



STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Conclusion

- We design a novel self-attention architecture with 2.5x speedup compared with TensorRT
- We introducing tensor-tile pruning algorithms and model-aware pruning.
- E.T. is avaliable at:
    -

Thank You & Questions?