LINFORMER

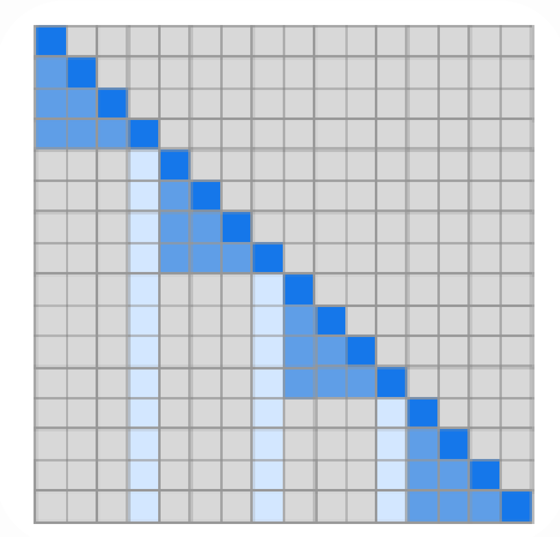
SELF-ATTENTION WITH LINEAR COMPLEXITY

STANISLAS BOITON LEDENKO, JACQUES MAUBERT, MAXENCE LASBORDES

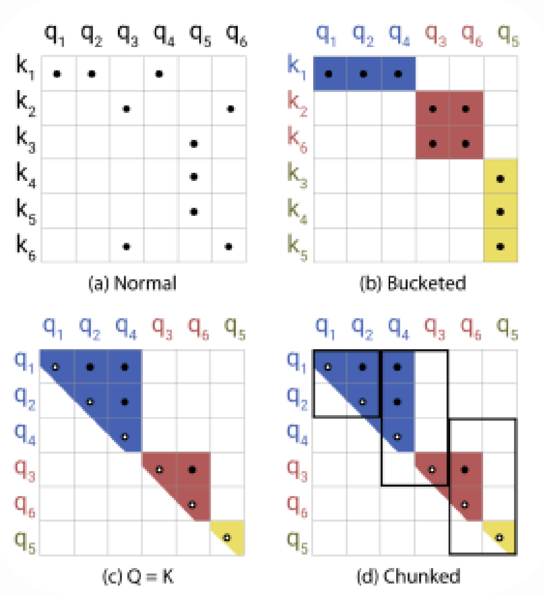
Introduction

- Transformers are $O(n^2)$ in time and memory complexity.
- Self-attention is the bottleneck of Transformers.
- BERT, GPT, T5, RoBERTa

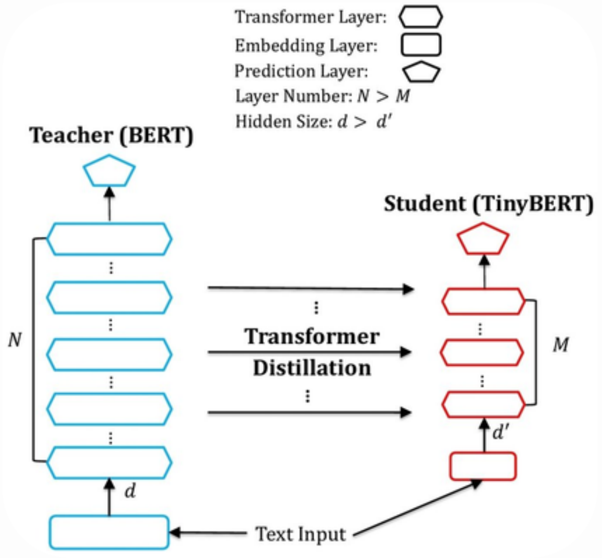
Related Works



Sparse Attention



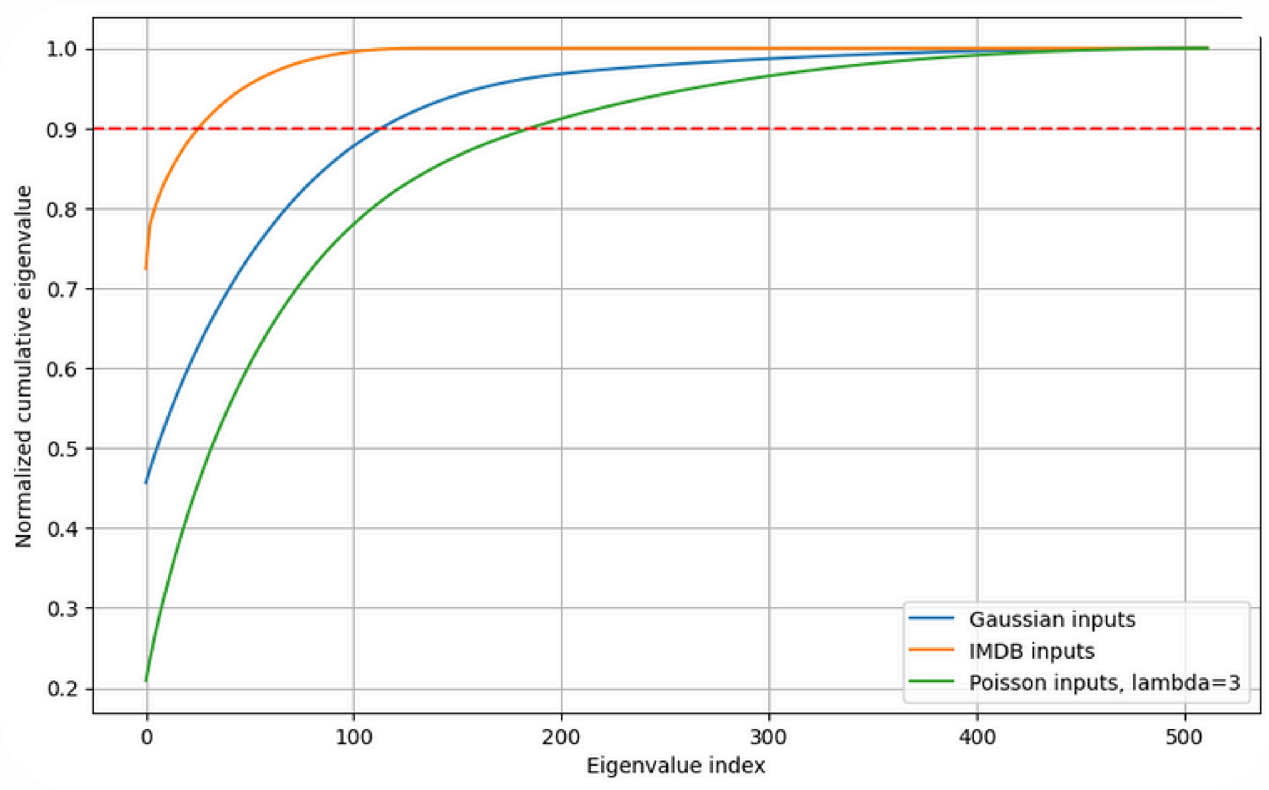
LSH Attention

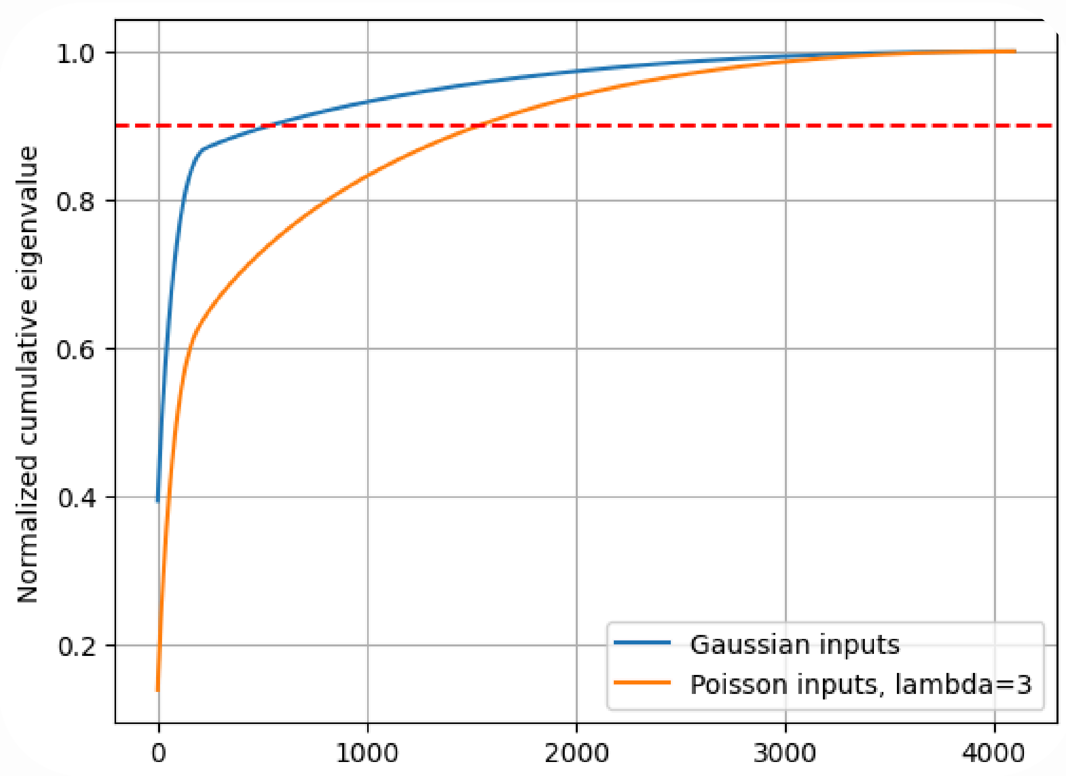


Knowledge Distillation

Idea and key principle

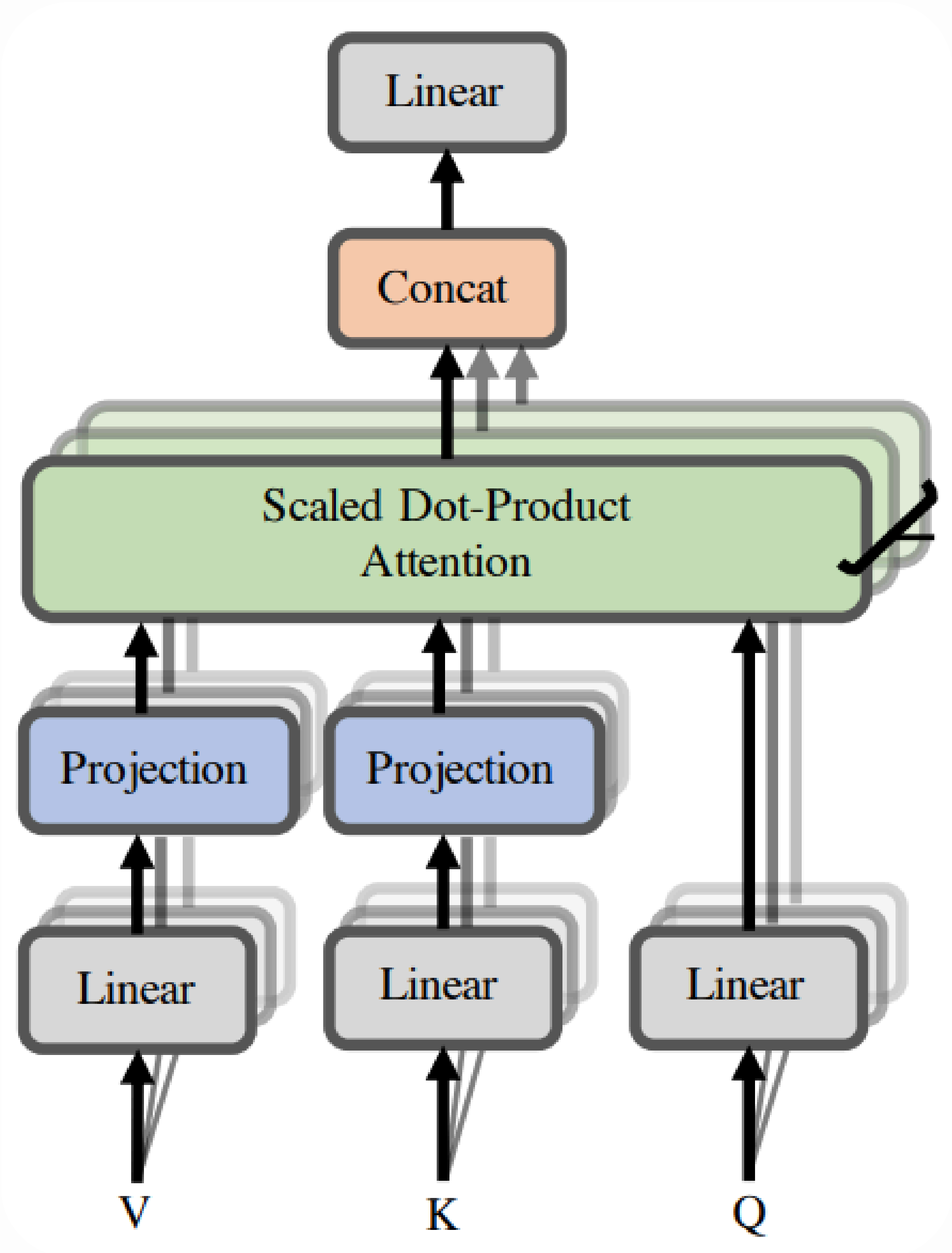
The context mapping matrix is approximaly low rank





Model

- Self-Attention is O(n) in time and memory complexity.



Results/Limits

n	Model	SST-2	IMDB	QNLI	QQP	Average
	Liu et al. (2019), RoBERTa-base	93.1	94.1	90.9	**90.9**	92.25
	Linformer, 128	92.4	94.0	90.4	90.2	91.75
	Linformer, 128, shared kv	**93.4**	93.4	90.3	90.3	91.85
	Linformer, 128, shared kv, layer	93.2	93.8	90.1	90.2	91.83
512	Linformer, 256	93.2	94.0	90.6	90.5	92.08
	Linformer, 256, shared kv	93.3	93.6	90.6	90.6	92.03
	Linformer, 256, shared kv, layer	93.1	94.1	**91.2**	90.8	**92.30**
512	Devlin et al. (2019), BERT-base	92.7	93.5	91.8	89.6	91.90
	Sanh et al. (2019), Distilled BERT	91.3	92.8	89.2	88.5	90.45
	Linformer, 256	93.0	93.8	90.4	90.4	91.90
1024	Linformer, 256, shared kv	93.0	93.6	90.3	90.4	91.83
	Linformer, 256, shared kv, layer	93.2	**94.2**	90.8	90.5	92.18

Comparing the pretraining perplexities of various models.

Inference-time Efficiency Results

length n	projected dimensions k				
	128	256	512	1024	2048
512	1.5x	1.3x	-	-	-
1024	1.7x	1.6x	1.3x	-	-
2048	2.6x	2.4x	2.1x	1.3x	-
4096	3.4x	3.2x	2.8x	2.2x	1.3x
8192	5.5x	5.0x	4.4x	3.5x	2.1x
16384	8.6x	7.8x	7.0x	5.6x	3.3x
32768	13x	12x	11x	8.8x	5.0x
65536	20x	18x	16x	14x	7.9x

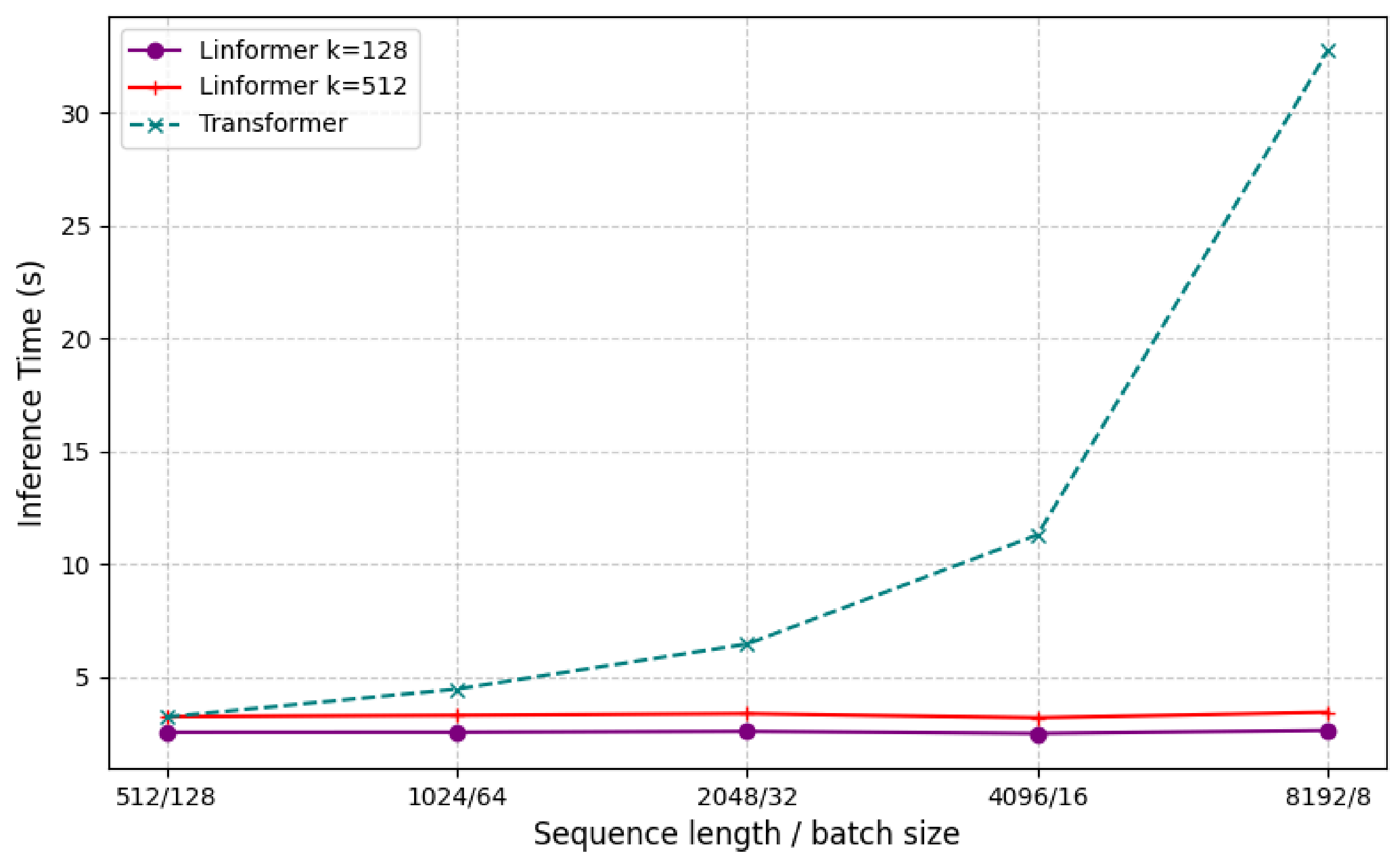
Time

length n	projected dimensions k				
	128	256	512	1024	2048
512	1.7x	1.5x	-	-	-
1024	3.0x	2.9x	1.8x	-	-
2048	6.1x	5.6x	3.6x	2.0x	-
4096	14x	13x	8.3x	4.3x	2.3x
8192	28x	26x	17x	8.5x	4.5x
16384	56x	48x	32x	16x	8x
32768	56x	48x	36x	18x	16x
65536	60x	52x	40x	20x	18x

Memory

Experiment

- Implementation of the Linformer self-attention & Transformer self-attention



Limits

- Efficient for long sequences tasks only
- Quality loss
- Not used nowadays

Conclusion

- Transformer models are slow to train and deploy.
- Theoretical and empirical demonstration.
- Efficient self-attention mechanism, O(n) with respect to sequence length.
- Other solutions: Flash attention, quantization techniques

Thanks!