VCC

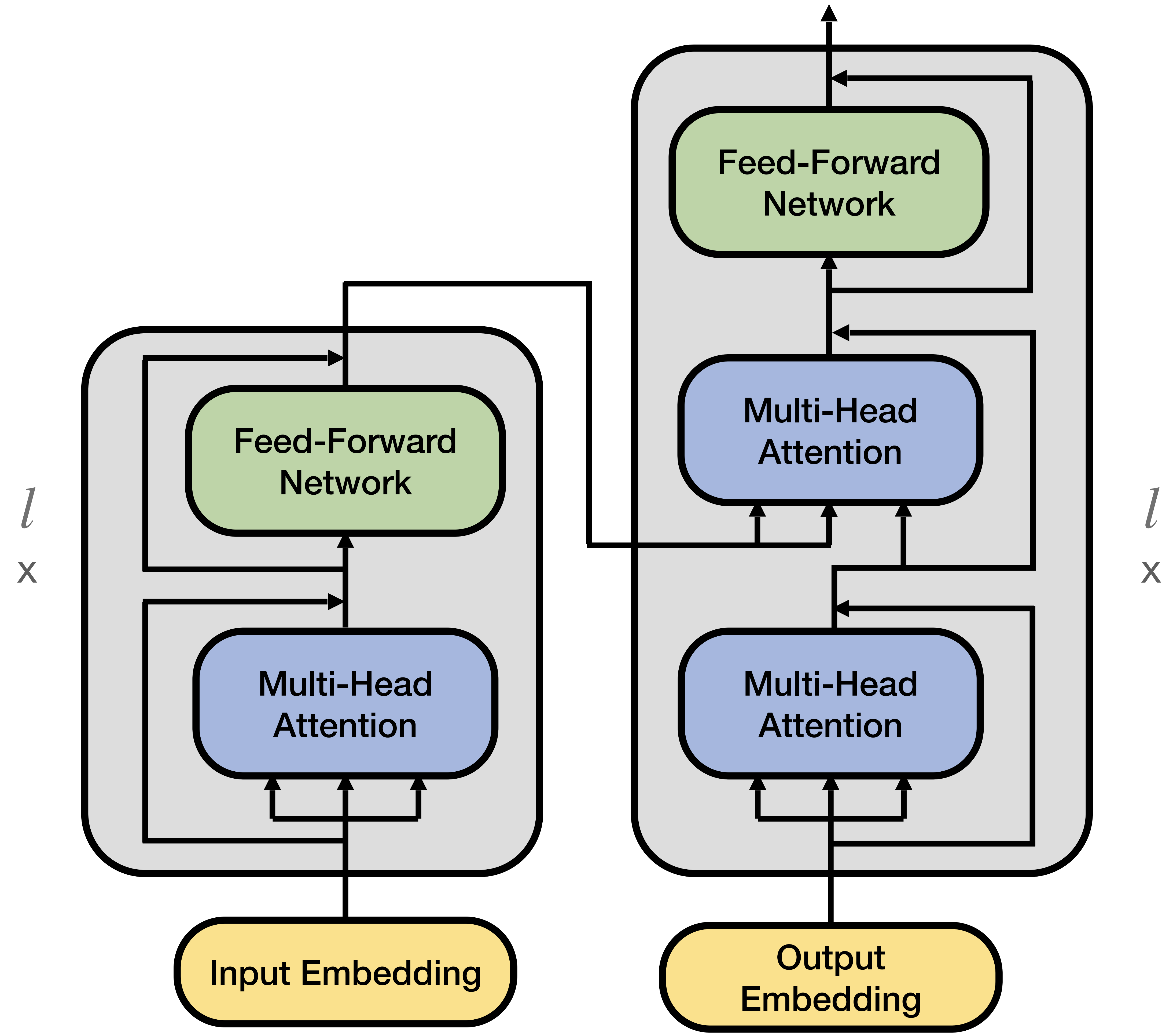
Scaling Transformers to 128K Tokens or More by Prioritizing Important Tokens

Zhanpeng Zeng, Cole Hawkins, Mingyi Hong, Aston Zhang,
Nikolaos Pappas, Vikas Singh, Shuai Zheng

Motivation

Transformer

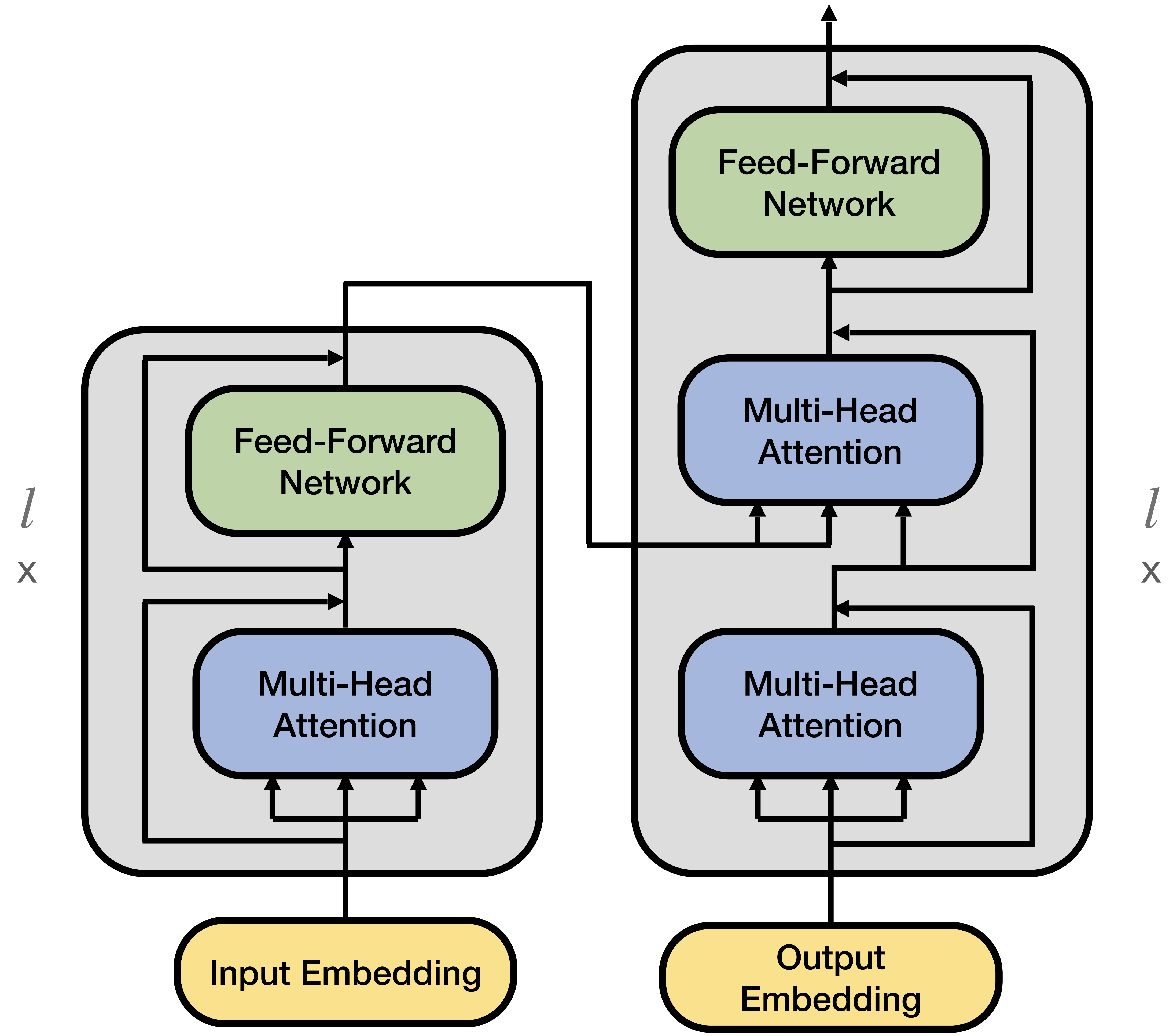
- Transformers are extremely efficient on LLM and visions



Motivation

Transformer

- Transformers are extremely efficient on LLM and visions
- But Transformers are extremely compute intensive when processing long sequences



Motivation

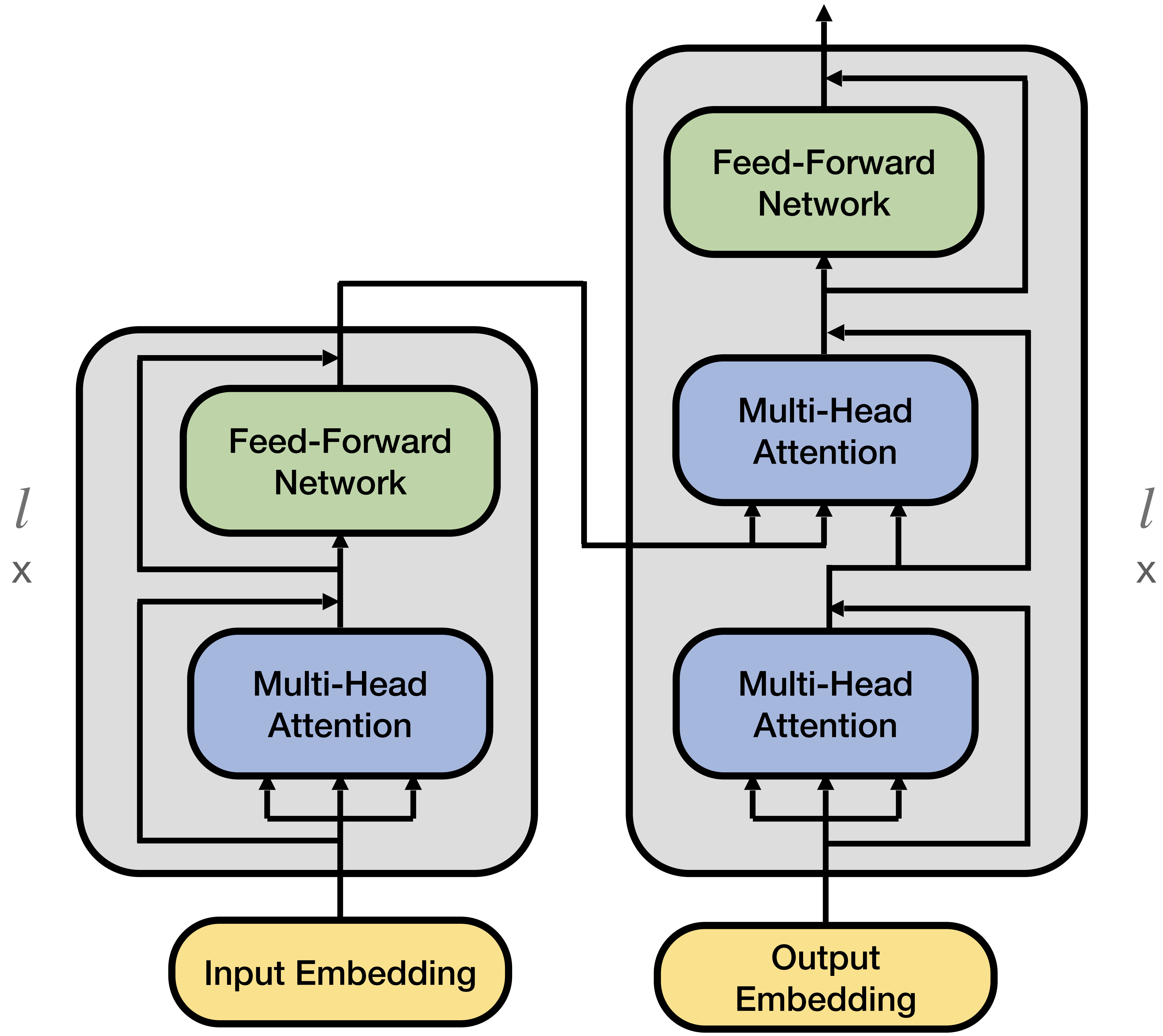
Transformer

- Transformers are extremely efficient on LLM and visions

- But Transformers are extremely compute intensive when processing long sequences

$$O(ln^2d + lnd^2)$$

- l: number of layers
- n: sequence lengths
- d: model dimension

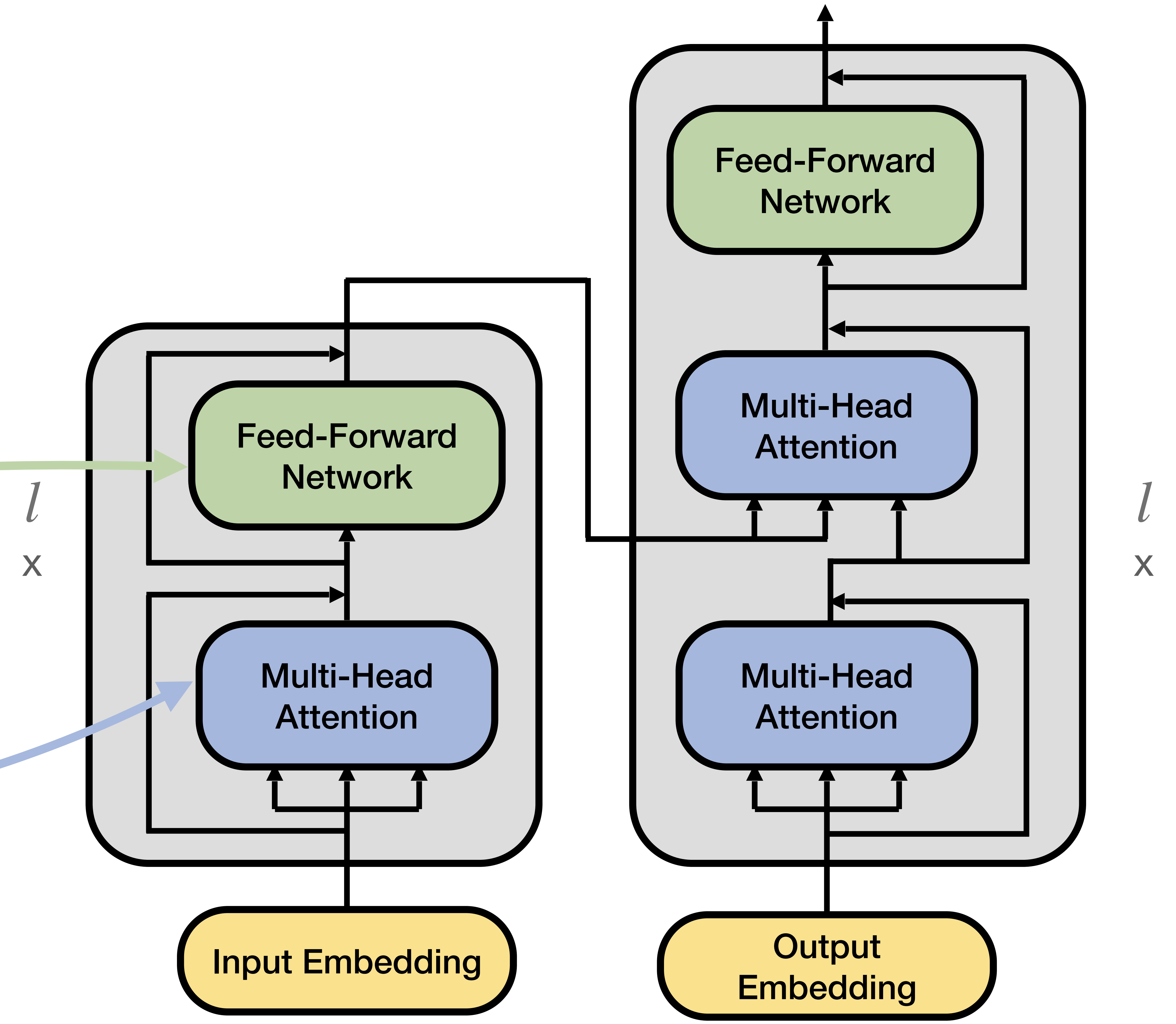


Motivation

Transformer

- Transformers are extremely efficient on LLM and visions
- But Transformers are extremely compute intensive when processing long sequences

$$O(ln^2d + lnd^2)$$



- l: number of layers
- n: sequence lengths
- d: model dimension

Motivation

Efficient Transformers

- Performer, Random Feature Attention, Nystromformer, Longformer, Big Bird, Reformer, YOSO, MRA Attention, Memorizing Transformers, RMT…

$$O(ln^2d + lnd^2)$$

$$\downarrow$$

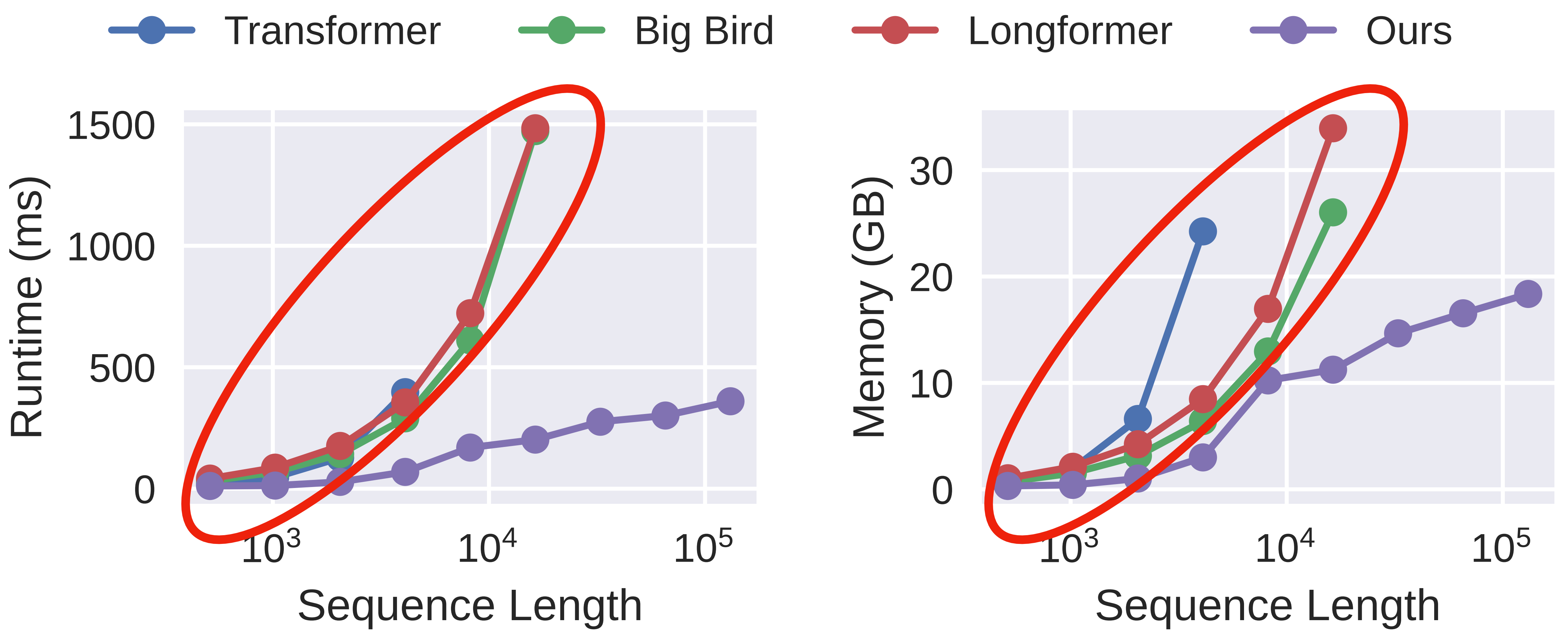
$$O(lnmd + lnd^2)$$

Motivation

Efficient Transformers

- Performer, Random Feature Attention, Nystromformer, Longformer, Big Bird, Reformer, YOSO, MRA Attention, Memorizing Transformers, RMT…

$$O(lnmd + lnd^2)$$

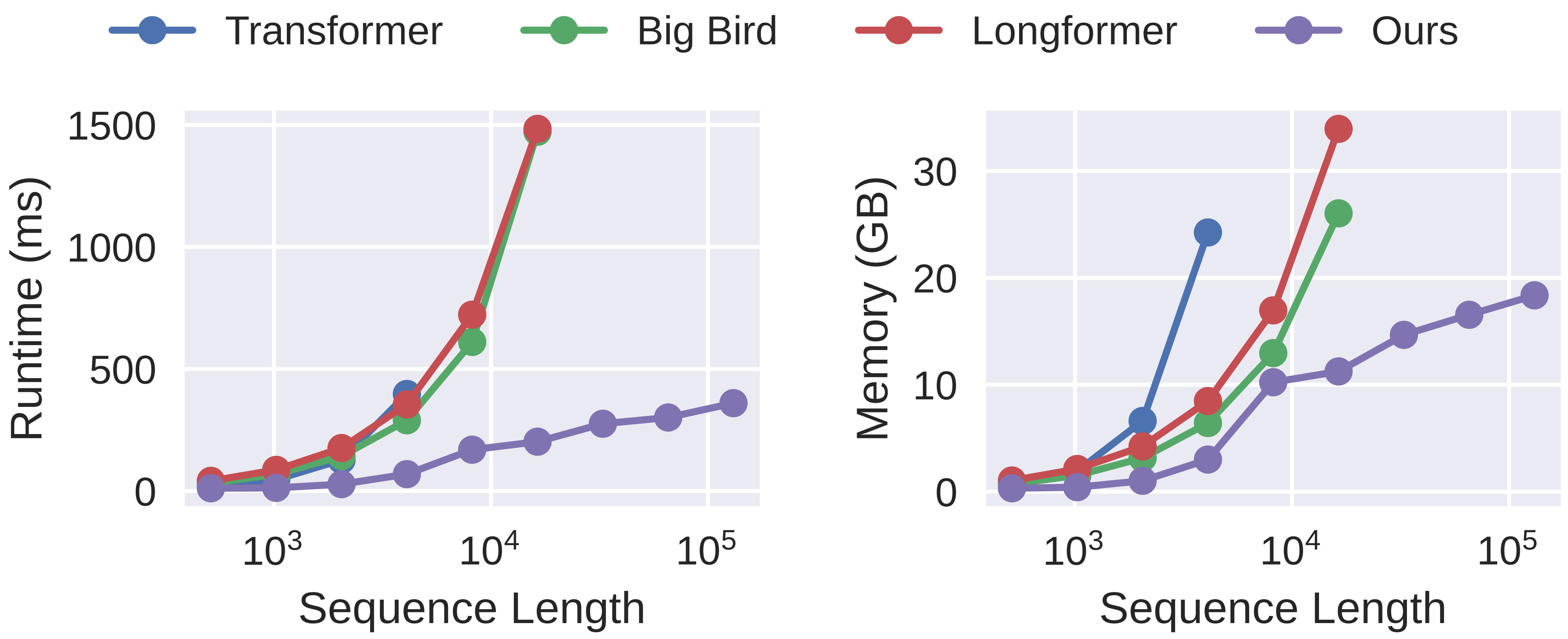


Motivation

Efficient Transformers

- Performer, Random Feature Attention, Nystromformer, Longformer, Big Bird, Reformer, YOSO, MRA Attention, Memorizing Transformers, RMT…

$$O(l\textcolor{red}{n}md + l\textcolor{red}{n}d^2)$$



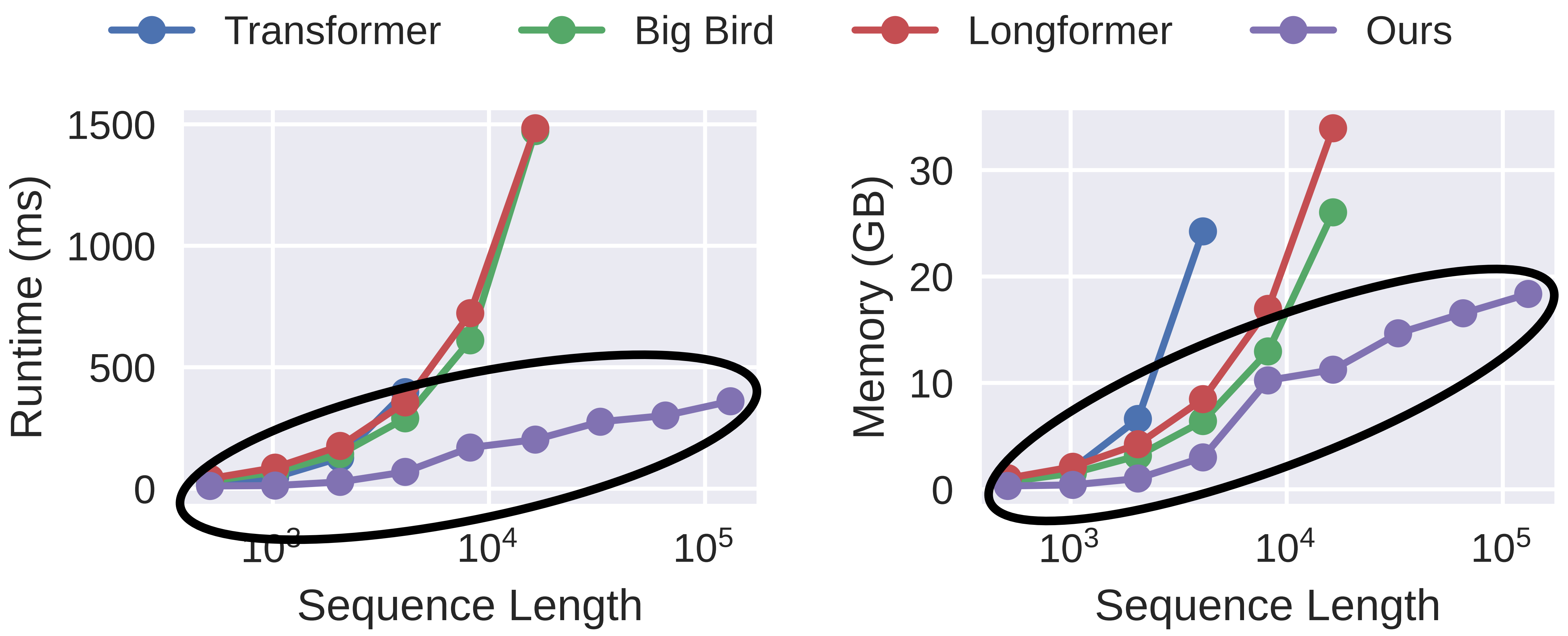
Can we do better?

Motivation

Efficient Transformers

- Performer, Random Feature Attention, Nystromformer, Longformer, Big Bird, Reformer, YOSO, MRA Attention, Memorizing Transformers, RMT…

$$O(l\textcolor{red}{n}md + l\textcolor{red}{n}d^2)$$



The goal of this work

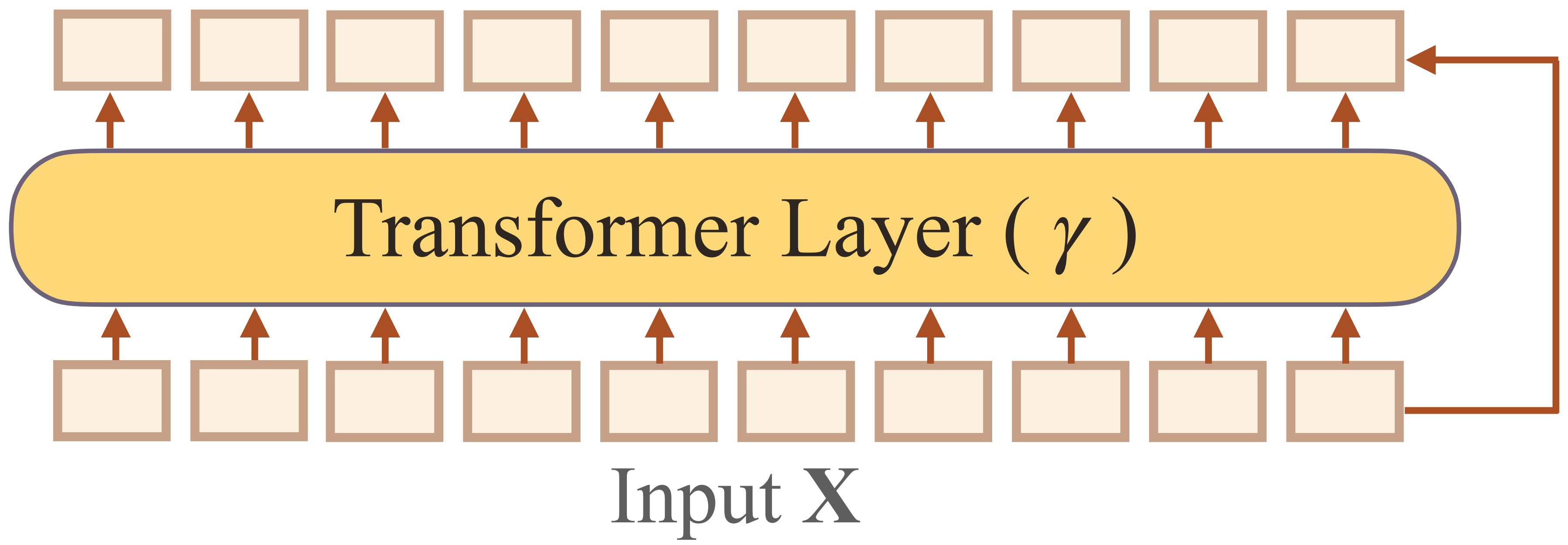
VIP-Token Centric Compression (VCC)

Overview

Vanilla Transformer

$$X_{new} = \beta(\alpha(X) + X) + \alpha(X) + X$$

$$= \gamma(X) + X$$



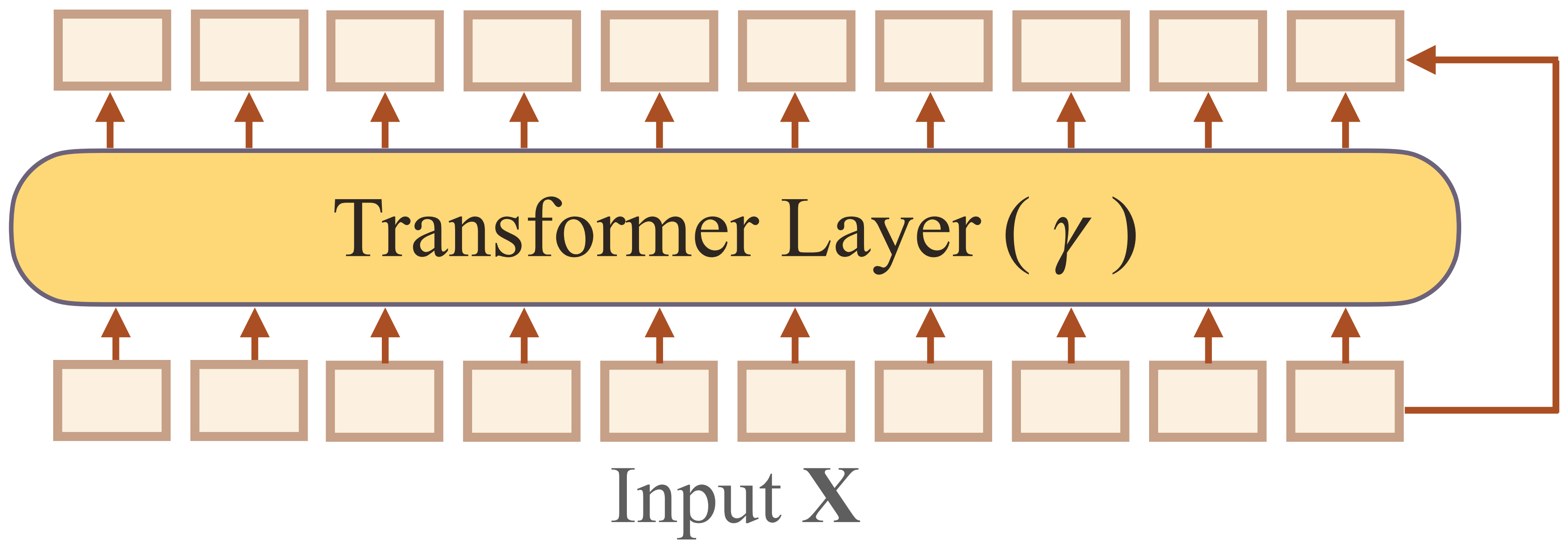
VIP-Token Centric Compression (VCC)

Overview

Vanilla Transformer

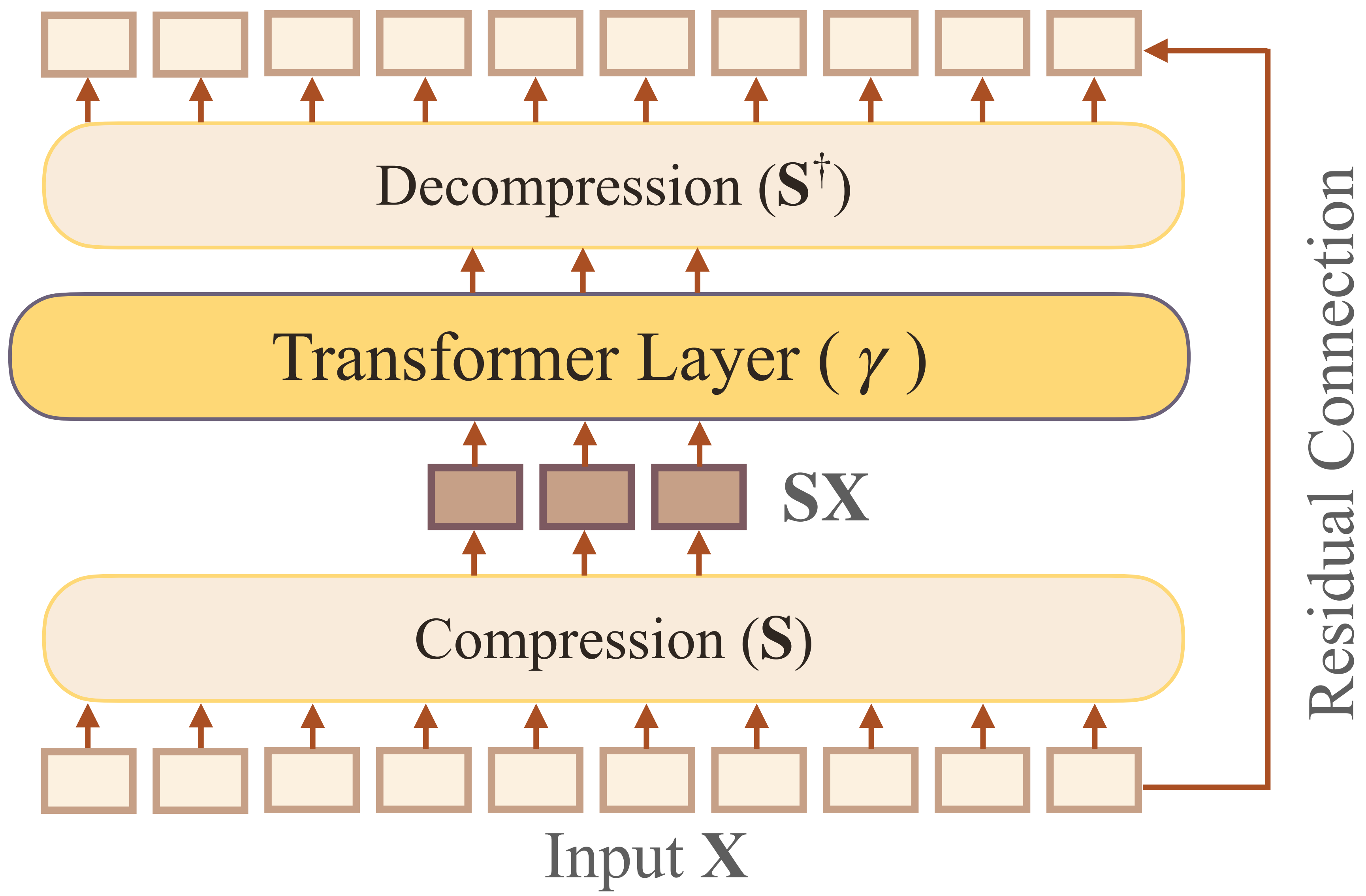
$$X_{new} = \beta(\alpha(X) + X) + \alpha(X) + X$$

$$= \gamma(X) + X$$



Compression

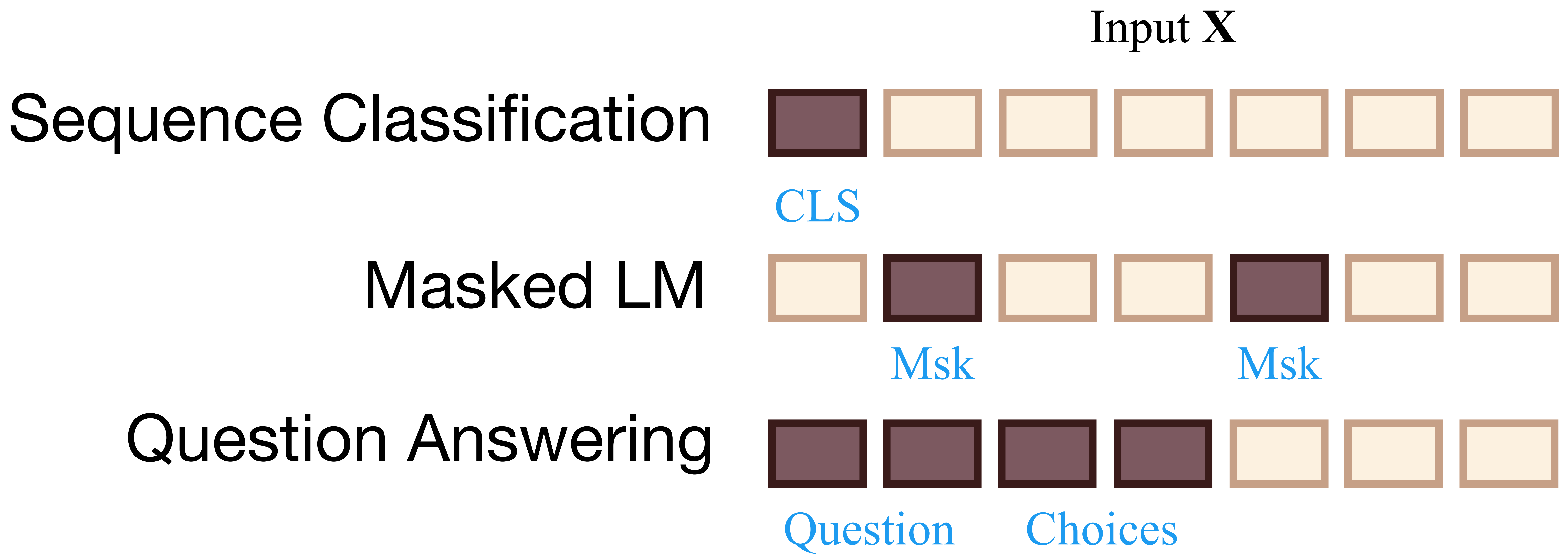
$$X_{new} = S^{\dagger}\gamma(SX) + X$$



VIP-Token Centric Compression (VCC)

VIP-Tokens

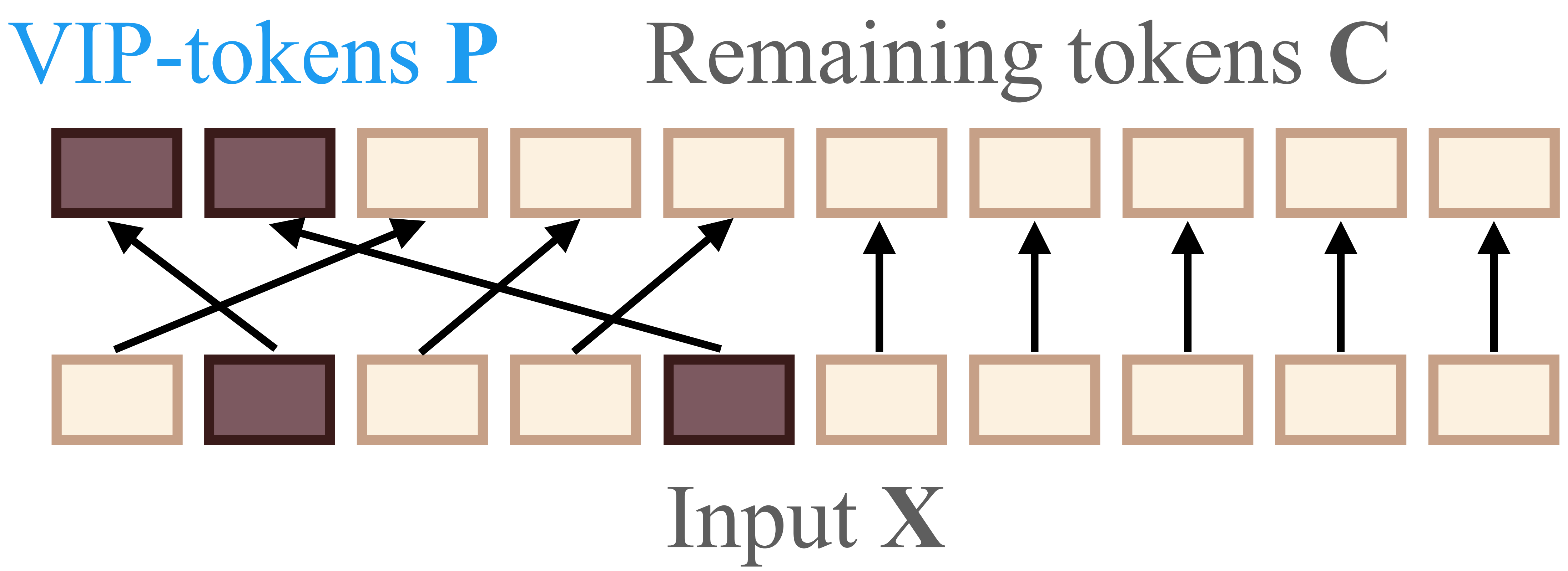
- Elevating the importance of a few tokens: VIP-Tokens



VIP-Token Centric Compression (VCC)

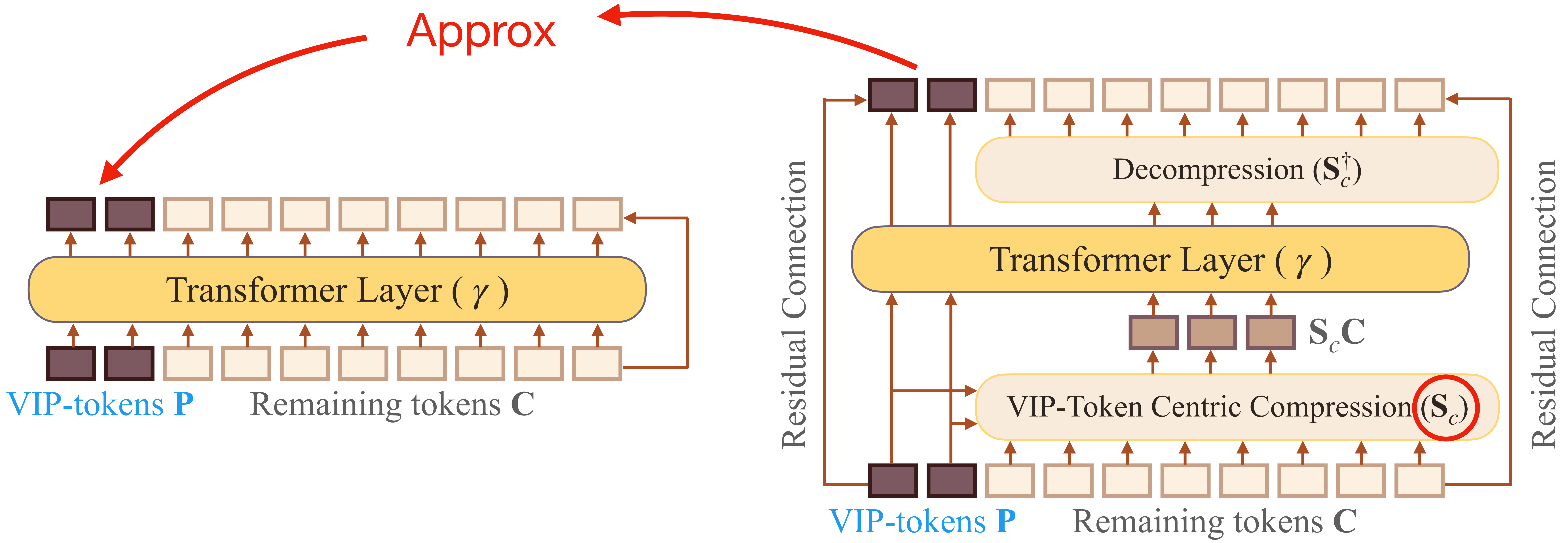
VIP-Tokens

- VIP-tokens occupy the front seats



VIP-Token Centric Compression (VCC)

Goal

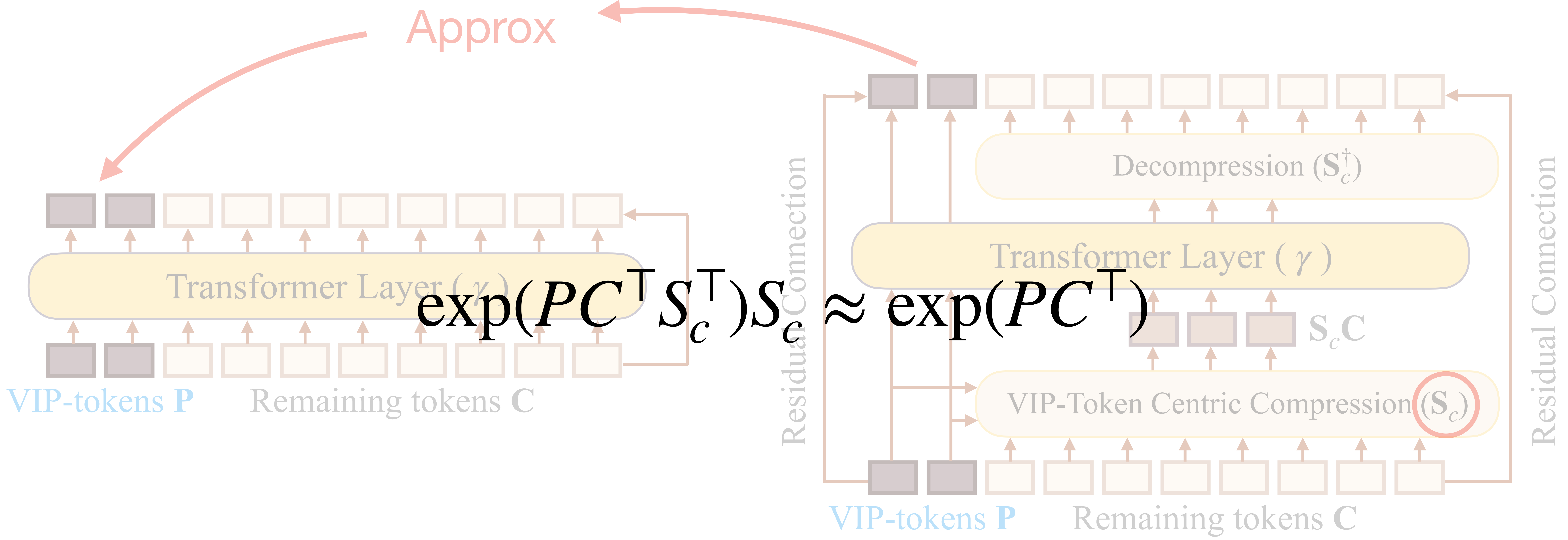


Ground Truth

Approx when applying VCC

VIP-Token Centric Compression (VCC)

Observation



$$\exp(PC^\top S_c^\top)S_c \approx \exp(PC^\top)$$

Ground Truth

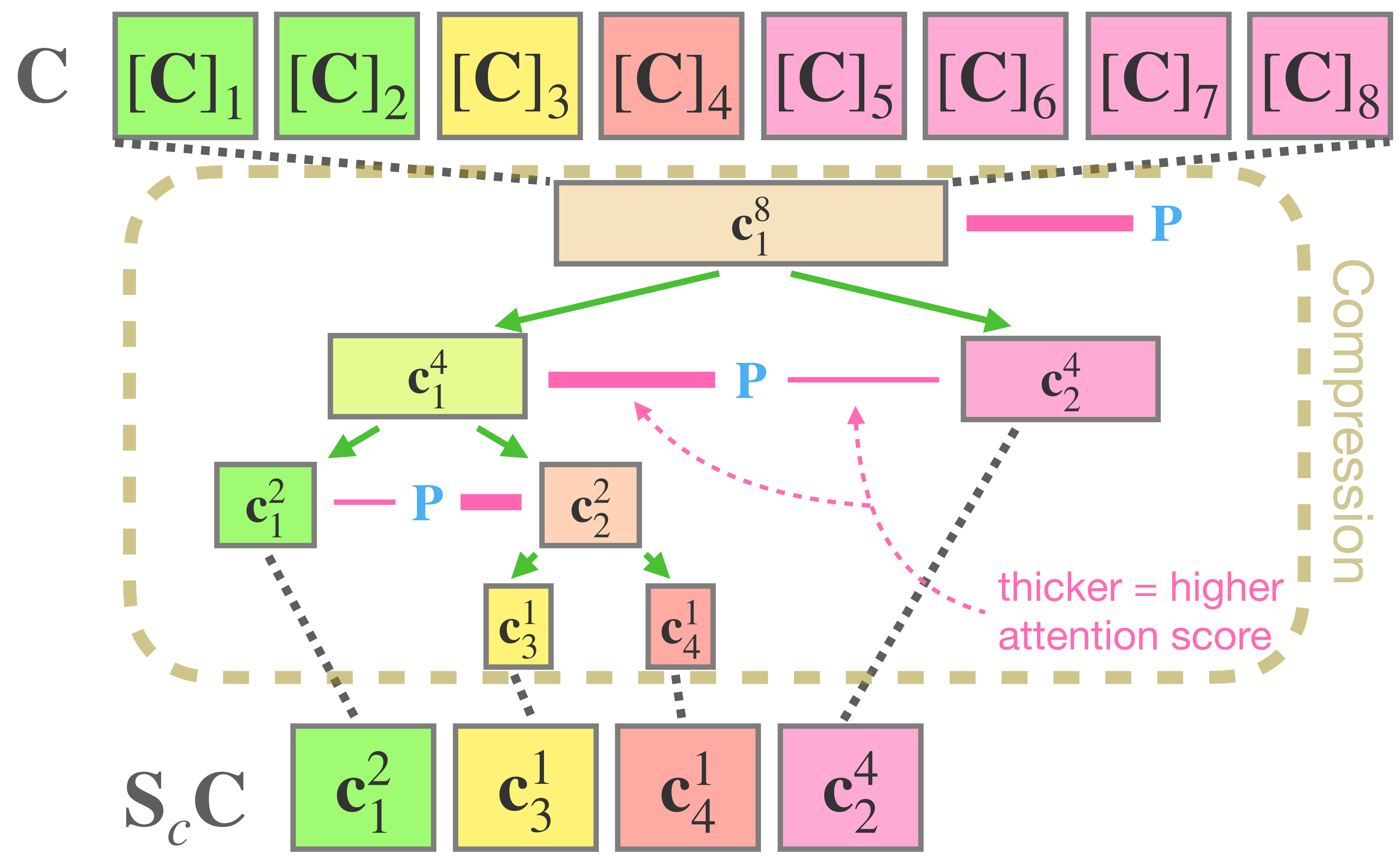
Approx when applying VCC

VIP-Token Centric Compression (VCC)

Instantiation of VCC

$$\exp(PC^\top S_c^\top)S_c \approx \exp(PC^\top)$$

Multi-Resolution Compression

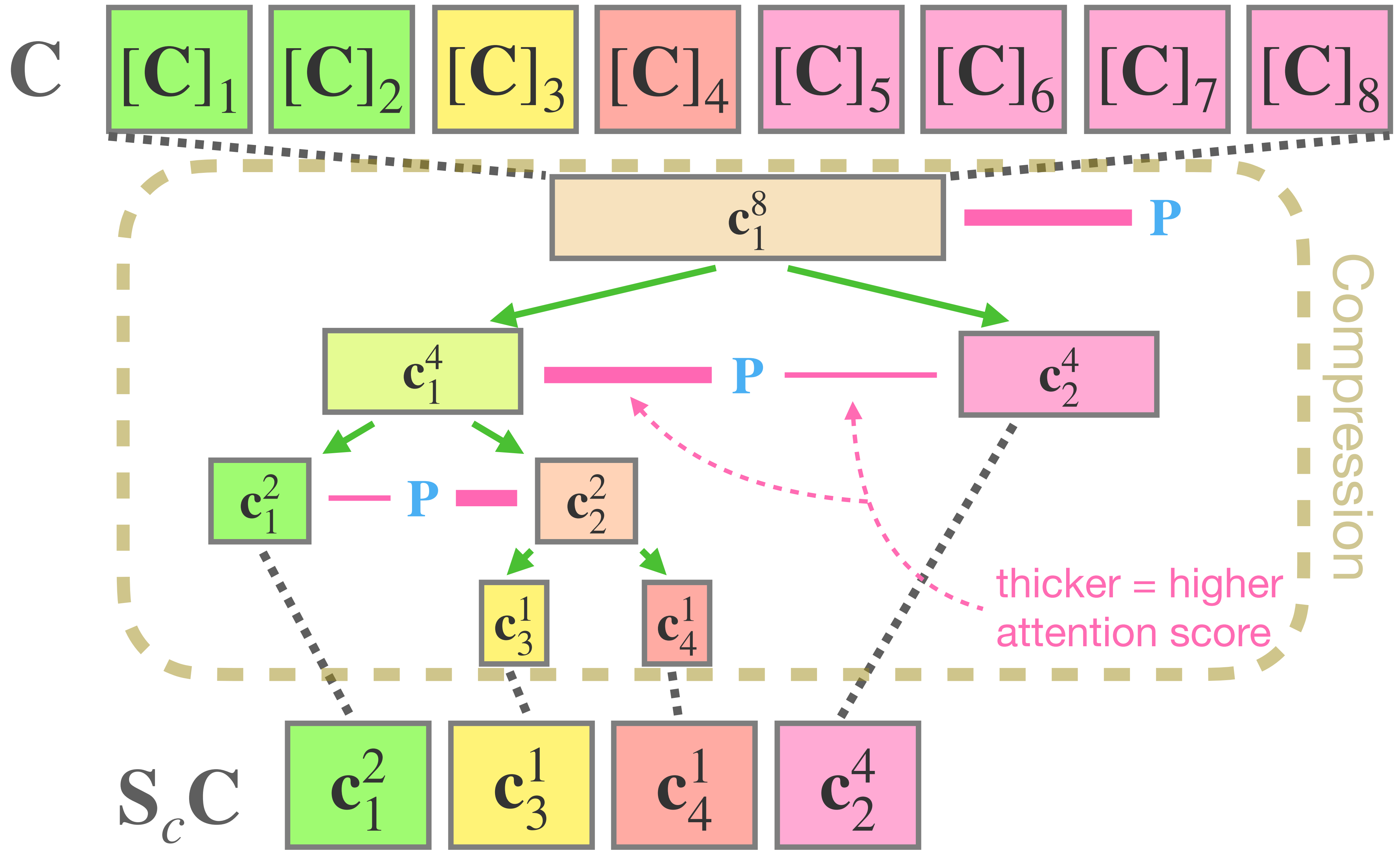


VIP-Token Centric Compression (VCC)

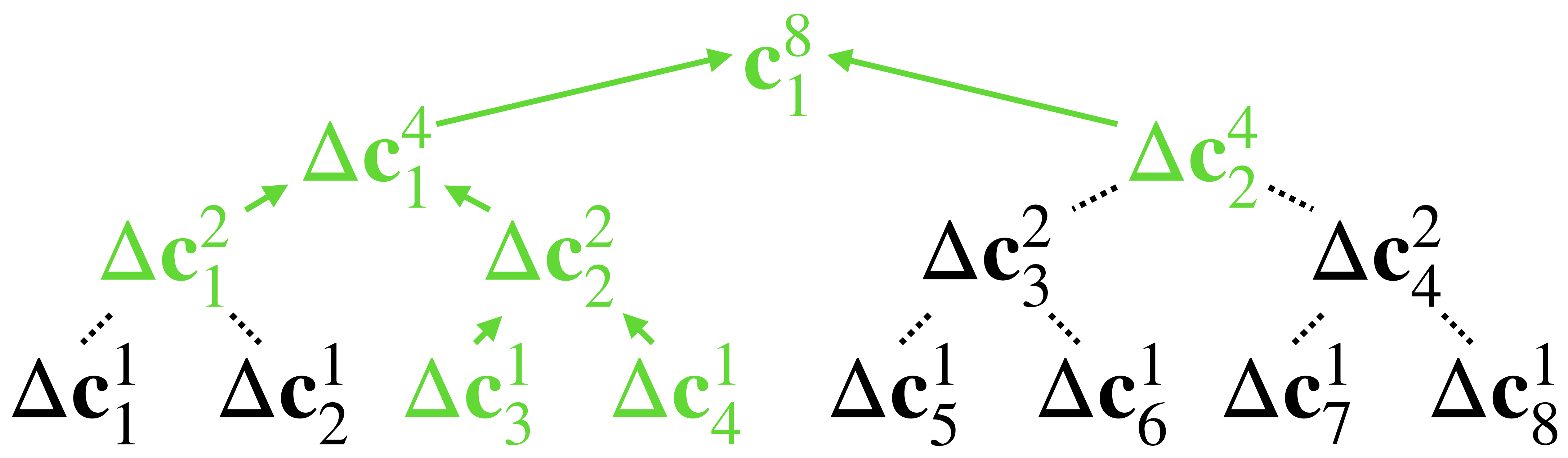
Instantiation of VCC

$$\exp(PC^\top S_c^\top)S_c \approx \exp(PC^\top)$$

Multi-Resolution Compression



Efficient Data Structure for Efficient Compression and Decompression



VIP-Token Centric Compression (VCC)

Efficiency Gain

$$O(ln^2d + lnd^2)$$

$$\downarrow$$

$$O(lr^2d + lrd^2) + O(lr\log(n_c)d + lrn_pd) + O(nd)$$

r: length of compressed sequence, n_p: number of VIP-tokens, n_c: number of remaining tokens

VIP-Token Centric Compression (VCC)

Efficiency Gain

$$O(ln^2d + lnd^2)$$

$$O(lr^2d + lrd^2) + O(lr\log(n_c)d + lrn_pd) + O(nd)$$

Cost of
Transformer

r: length of compressed sequence, n_p: number of VIP-tokens, n_c: number of remaining tokens

VIP-Token Centric Compression (VCC)

Efficiency Gain

$$O(ln^2d + lnd^2)$$

$$\downarrow$$

$$O(lr^2d + lrd^2) + O(lr\log(n_c)d + lrn_pd) + O(nd)$$

$\downarrow$ Cost of Transformer

$\downarrow$ Overhead for de/compression

r: length of compressed sequence, n_p: number of VIP-tokens, n_c: number of remaining tokens

VIP-Token Centric Compression (VCC)

Efficiency Gain

$$O(ln^2d + lnd^2)$$

$$\downarrow$$

$$O(lr^2d + lrd^2) + O(lr\log(n_c)d + lrn_pd) + O(nd)$$

$$\downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow$$

Cost of Transformer

Overhead for de/compression

Overhead for efficient data structure

r: length of compressed sequence, n_p: number of VIP-tokens, n_c: number of remaining tokens

Evaluation

Evaluation
Encoder-Only Models

Table 2: Dev set results for encoder-only models.

Method	Size	Length	HotpotQA			QuALITY		WikiHop	
			Time	EM	F1	Time	Accuracy	Time	Accuracy
RoBERTa	base	512	19.9	35.1	44.9	21.2	39.0	19.6	67.6
RoBERTa	base	4K	422.3	62.2	76.1	403.2	39.5	414.1	75.2
Big Bird	base	4K	297.9	59.5	73.2	307.0	38.5	293.3	74.5
Longformer	base	4K	371.0	59.9	73.6	368.0	27.9	369.7	74.3
MRA Attention	base	4K	203.5	63.4	77.0	200.5	38.7	199.2	76.1
Ours	base	4K	114.6	60.9	74.6	126.4	39.6	108.0	75.9
Ours*	base	4K	114.6	61.4	75.0	125.7	39.5	108.0	76.1
Ours*	large	4K	285.8	66.7	80.0	390.8	41.8	394.3	79.6

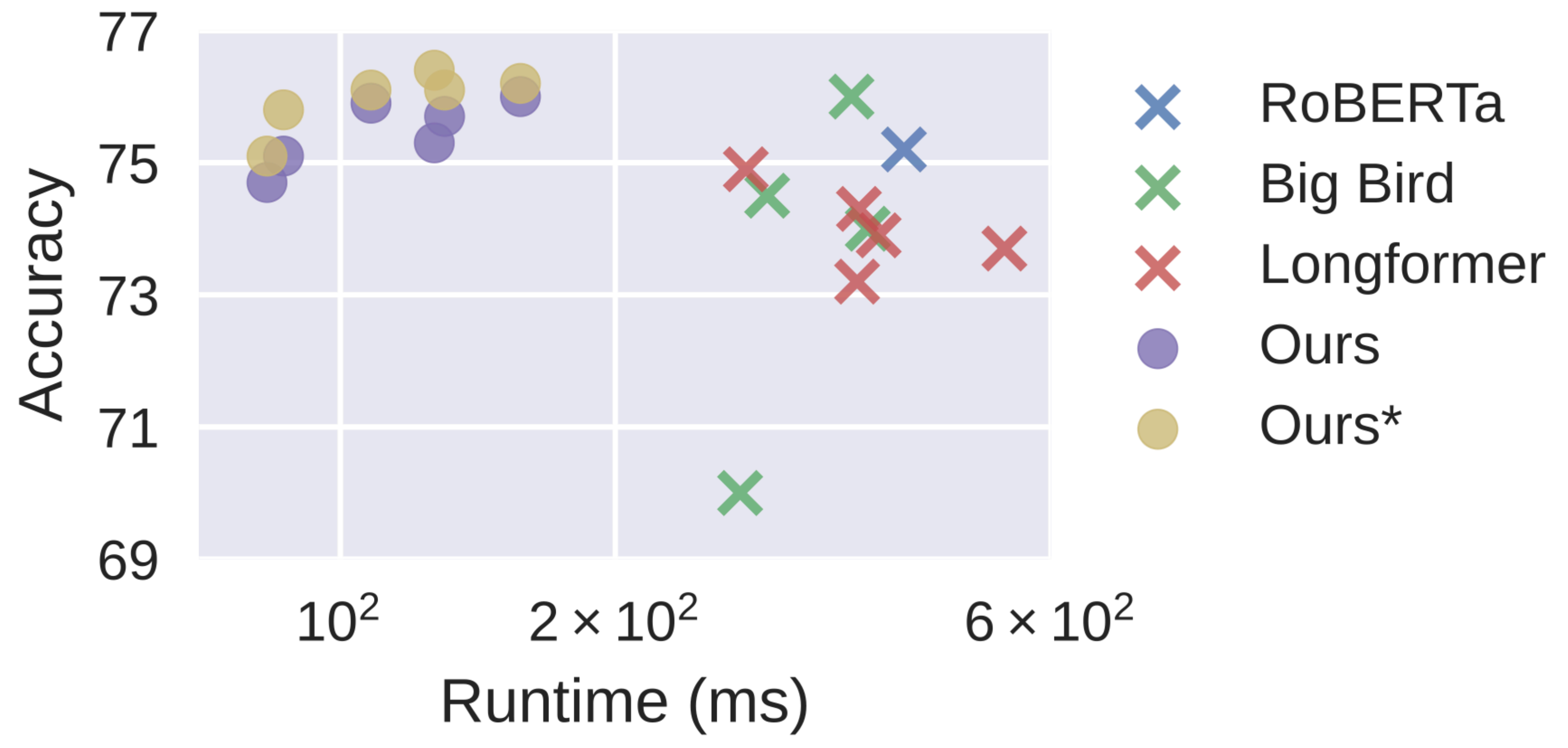


Figure 5: Model runtime vs WikiHop dev accuracy when using different model specific hyperparameters

Evaluation

Encoder-Decoder Models

Table 3: Dev set results for encoder-decoder models. The left / right values of runtime columns are the runtime for the entire model / the encoder.

Method	Size	# Param	Length	WikiHop			HotpotQA			CNN/Dailymail				MediaSum			
				Runtime	EM	F1	Runtime	EM	F1	Runtime	R-1	R-2	R-L	Runtime	R-1	R-2	R-L
T5	base	223M	512	25.7 / 20.5	66.7	69.1	26.3 / 20.5	34.1	44.4	40.0 / 20.5	43.3	20.5	40.4	39.9 / 20.5	30.7	14.5	28.1
T5	base	223M	4K	594.3 / 553.7	76.2	78.1	594.3 / 550.6	64.2	77.5	614.4 / 549.4	43.8	20.9	41.0	613.5 / 552.9	34.9	17.2	31.9
LongT5	base	248M	4K	270.7 / 233.9	72.7	74.8	271.3 / 233.7	62.3	75.7	291.6 / 234.9	43.3	20.6	40.5	287.3 / 229.5	34.9	17.3	32.0
LED	base	162M	4K	236.6 / 222.9	70.0	72.4	237.4 / 222.9	55.1	67.9	249.4 / 221.8	43.3	20.0	40.5	- / -	-	-	-
Ours	base	223M	4K	181.7 / 148.1	76.7	78.4	155.4 / 127.4	64.5	77.7	195.8 / 139.9	43.6	20.7	40.7	196.7 / 140.2	34.8	17.3	31.9
T5	large	738M	512	83.5 / 67.0	69.1	71.4	84.1 / 67.0	36.9	47.8	124.6 / 67.0	43.8	20.7	40.9	124.5 / 67.0	31.9	15.5	29.1
T5	large	738M	4K	1738.7 / 1601.0	79.1	80.7	1598.1 / 1598.1	68.0	81.3	1824.8 / 1600.4	44.3	21.0	41.4	- / -	-	-	-
Ours	large	738M	4K	561.4 / 460.6	79.0	80.6	485.3 / 382.8	67.8	81.0	608.1 / 433.8	44.4	21.4	41.5	609.7 / 434.4	35.8	18.2	32.8
Ours	3b	3B	4K	1821.5 / 1441.2	80.8	82.3	1547.7 / 1197.1	70.2	83.2	1930.7 / 1364.8	44.8	21.5	41.9	1930.7 / 1364.8	36.3	18.5	33.3

Evaluation
Encoder-Decoder Models

Table 3: Dev set results for encoder-decoder models. The left / right values of runtime columns are the runtime for the entire model / the encoder.

Method	Size	# Param	Length	Qasper			QuALITY			Arxiv				SummScreenFD			
				Runtime	EM	F1	Runtime	EM	F1	Runtime	R-1	R-2	R-L	Runtime	R-1	R-2	R-L
T5	base	223M	512	31.8 / 20.5	10.8	16.4	29.3 / 20.5	33.6	47.3	59.0 / 20.5	28.9	8.6	25.6	59.1 / 20.5	27.0	4.8	23.5
T5	base	223M	4K	608.2 / 551.7	13.2	29.1	596.3 / 551.2	34.7	47.4	645.4 / 549.1	44.4	18.4	39.9	647.9 / 551.1	31.6	6.8	27.6
LongT5	base	248M	16K	1628.5 / 1421.3	16.2	33.4	1633.1 / 1439.7	35.8	48.5	1699.7 / 1370.4	48.5	21.7	43.7	1763.4 / 1427.8	33.1	7.3	28.5
LED	base	162M	16K	- / -	-	-	- / -	-	-	1055.8 / 923.6	47.8	20.6	43.2	- / -	-	-	-
Ours	base	223M	16K	538.3 / 391.6	16.0	30.8	557.1 / 419.2	36.5	48.7	672.8 / 392.1	48.5	21.4	43.9	670.5 / 390.9	33.1	7.3	28.6
T5	large	738M	512	101.9 / 66.4	11.3	17.0	95.8 / 67.1	35.3	49.0	182.2 / 67.1	30.5	9.1	27.1	180.9 / 66.5	28.3	4.9	24.9
T5	large	738M	4K	- / -	-	-	1760.5 / 1596.4	37.8	50.5	1901.5 / 1598.8	46.0	19.4	41.4	- / -	-	-	-
Ours	large	738M	16K	1679.6 / 1120.2	16.3	33.7	1753.6 / 1210.7	40.3	52.5	1959.1 / 1111.0	49.5	22.2	44.7	1957.1 / 1109.2	34.3	7.6	29.6
Ours	3b	3B	16K	6165.4 / 4637.3	19.0	38.2	6398.8 / 4962.7	45.2	56.0	7676.3 / 4642.2	49.8	22.4	45.0	7641.5 / 4631.3	34.7	7.8	30.1

Evaluation
Encoder-Decoder Models

Table 3: Dev set results for encoder-decoder models. The left / right values of runtime columns are the runtime for the entire model / the encoder.

Method	Size	# Param	Length	ContractNLI			NarrativeQA			GovReport				QMSum			
				Runtime	EM	F1	Runtime	EM	F1	Runtime	R-1	R-2	R-L	Runtime	R-1	R-2	R-L
T5	base	223M	512	24.0 / 20.5	73.5	73.5	26.8 / 20.5	2.0	11.3	59.1 / 20.5	40.5	14.8	38.2	43.5 / 20.5	30.2	8.0	26.5
T5	base	223M	4K	579.0 / 551.6	86.8	86.8	593.4 / 547.6	3.8	13.3	648.3 / 551.5	54.0	25.2	51.4	620.2 / 551.5	31.1	8.2	27.4
LongT5	base	248M	16K	1564.2 / 1462.5	85.1	85.1	1541.7 / 1370.2	5.2	15.6	1726.4 / 1387.7	55.8	27.9	53.2	1721.4 / 1450.7	35.7	11.7	31.4
Ours	base	223M	16K	484.2 / 393.1	87.0	87.0	518.2 / 394.4	5.0	15.8	674.0 / 391.6	55.2	27.1	52.6	623.1 / 396.5	31.8	8.8	27.9
T5	large	738M	512	78.1 / 67.1	74.3	74.3	- / -	-	-	180.9 / 67.0	43.3	16.2	41.1	136.4 / 67.1	31.7	8.1	27.6
T5	large	738M	4K	1702.4 / 1601.2	87.2	87.2	- / -	-	-	- / -	-	-	-	- / -	-	-	-
Ours	large	738M	16K	1440.6 / 1122.6	87.8	87.8	1551.7 / 1133.9	6.6	18.7	1955.5 / 1113.8	56.3	28.0	53.8	1816.4 / 1134.6	34.8	10.4	30.7
Ours	3b	3B	16K	5850.2 / 4665.9	88.5	88.5	6055.4 / 4659.4	8.2	21.2	7668.2 / 4642.7	56.9	28.5	54.3	7146.7 / 4655.6	35.7	10.9	31.1

Evaluation

Scaling Length to 128K

Table 4: Dev results of NarrativeQA on base model when scaling sequence length from 16K to 128K.

Length	Runtime (ms)	k	h	EM	F1
16K	518.2 / 394.4 / 162.4	16	90	5.9	16.6
32K	946.8 / 671.6 / 212.6	32	55	6.6	17.5
32K	1027.9 / 751.0 / 298.0	16	90	6.4	17.5
64K	1848.7 / 1177.2 / 254.8	64	30	7.2	18.4
64K	2244.8 / 1574.2 / 659.4	16	90	7.5	19.3
128K	6267.8 / 5125.9 / 1902.2	16	90	8.0	19.6

Takeaway

VCC reduces the overall complexity dependency on the sequence length without sacrificing the model accuracy.

VCC uses the standard Transformer blocks (with standard feed-forward network and self-attention) while achieving efficiency gain.

VCC can be directly incorporated into existing pertained models with some additional training.

GitHub mlpen/VCC

zzeng38@wisc.edu

End