

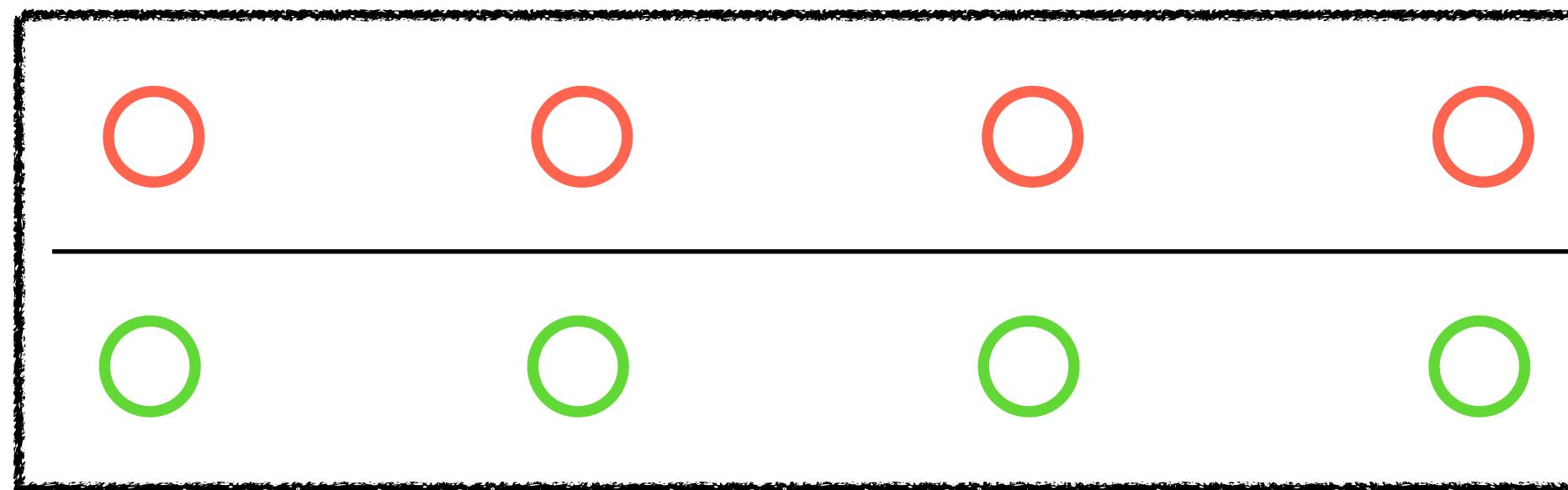
Transformers (2)

COMP3361 – Week 5

Lingpeng Kong

Department of Computer Science, The University of Hong Kong

Dot-Product-Softmax Attention



○
Query

Memory (key-value pairs)

$$\text{○ ○ } q \cdot k_1$$

$$\text{○ ○ } q \cdot k_2$$

$$\text{○ ○ } q \cdot k_3$$

$$\text{○ ○ } q \cdot k_4$$

$$q \cdot k_1$$

$$q \cdot k_2$$

$$q \cdot k_3$$

$$q \cdot k_4$$

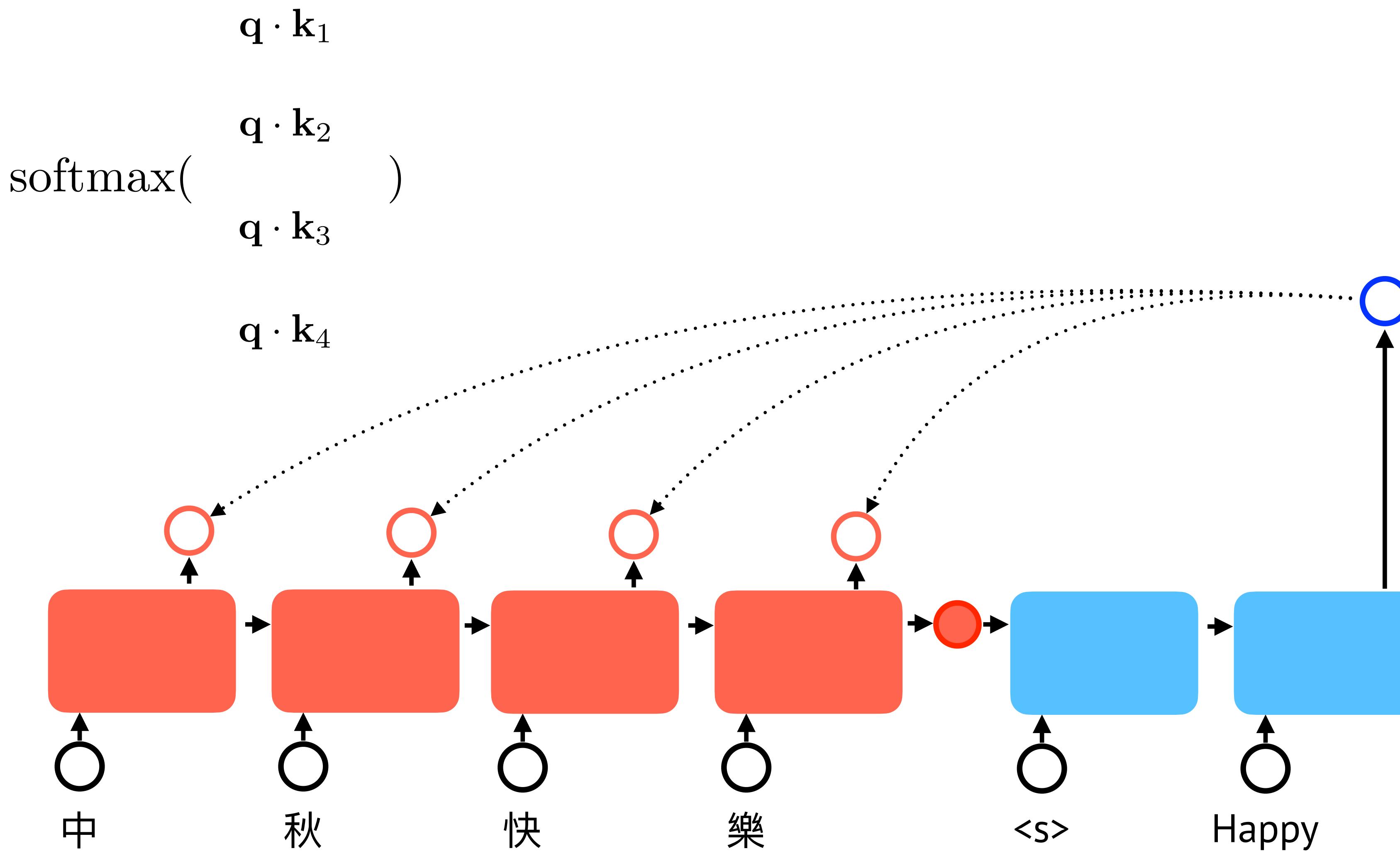
$$\text{softmax}(\quad \quad \quad) \rightarrow$$

$$\begin{bmatrix} 0.6 \\ 0.1 \\ 0.2 \\ 0.1 \end{bmatrix}$$

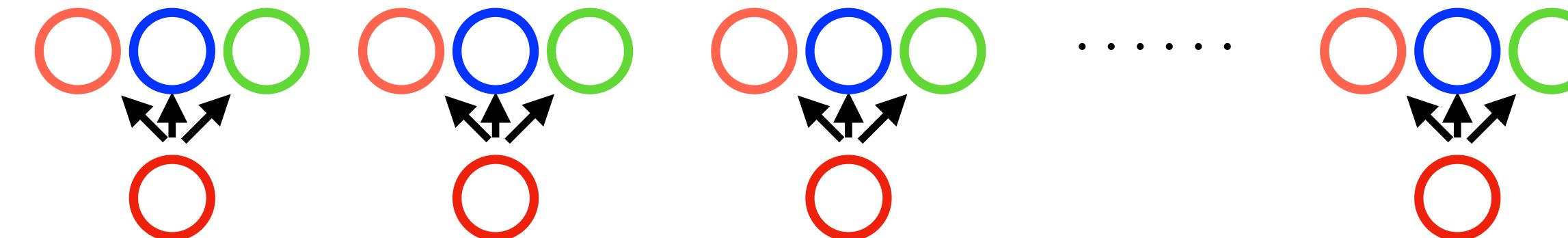
$$\begin{bmatrix} \text{○} \\ \text{○} \\ \text{○} \\ \text{○} \end{bmatrix}$$

$$\rightarrow 0.6 \text{○} + 0.1 \text{○} + 0.2 \text{○} + 0.1 \text{○} = \text{○} \text{ context vector } \mathbf{c}$$

Attention Mechanism

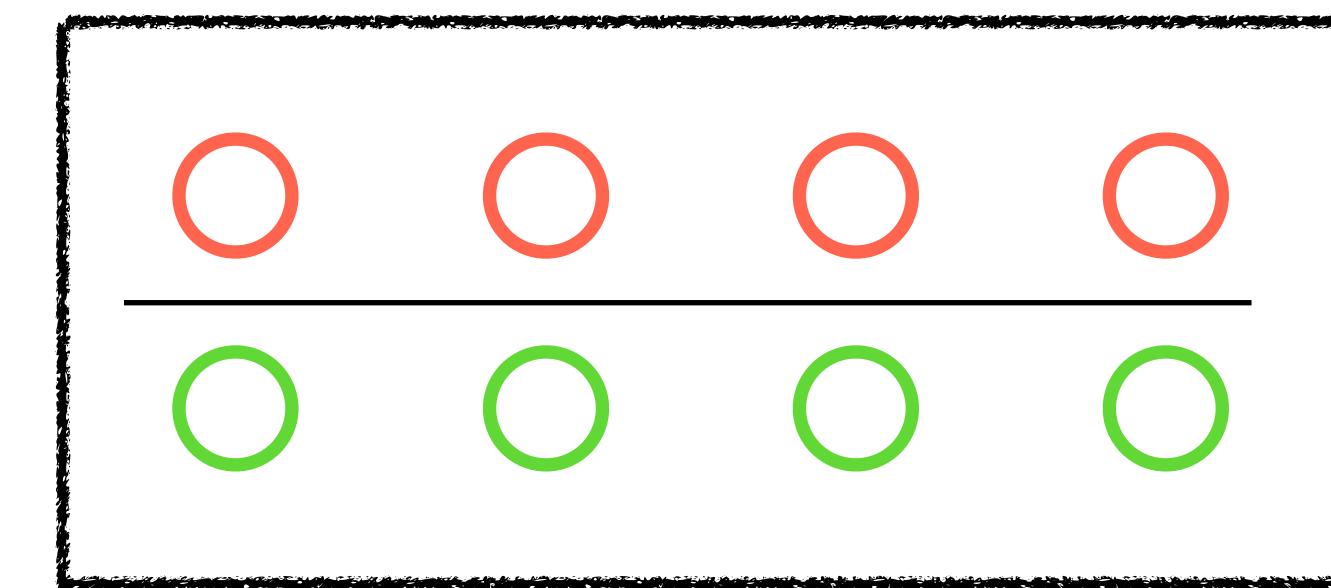


Considering the full sequence as context



Attention Mechanism

○
Query



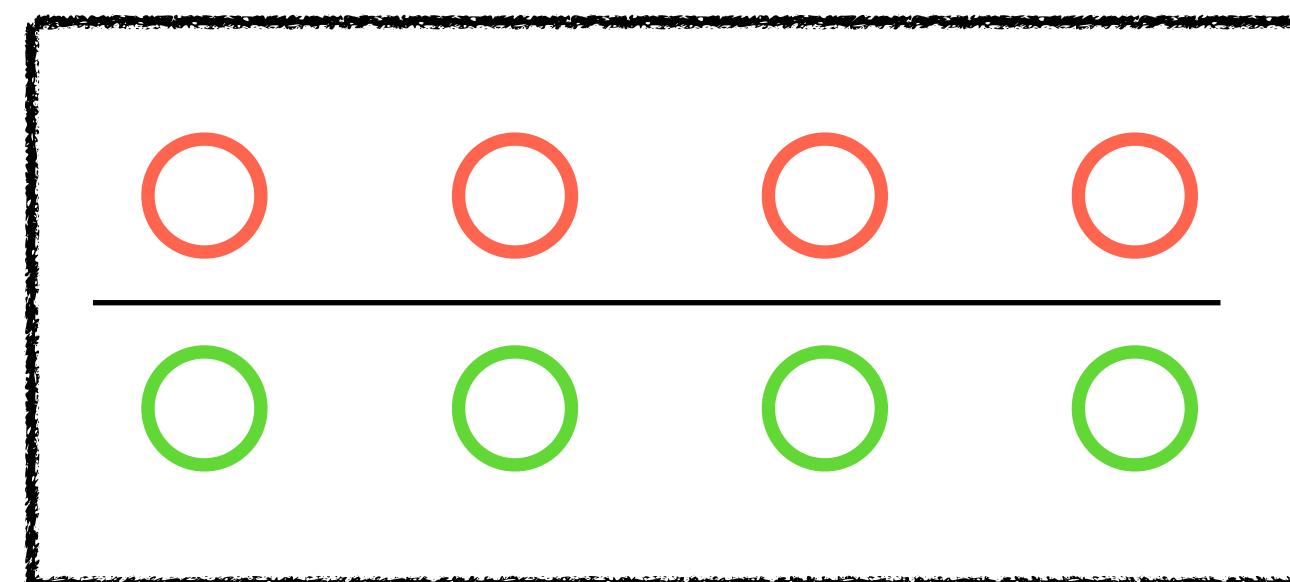
Memory (key-value pairs)

○ ○ ○ ○

Attention Mechanism

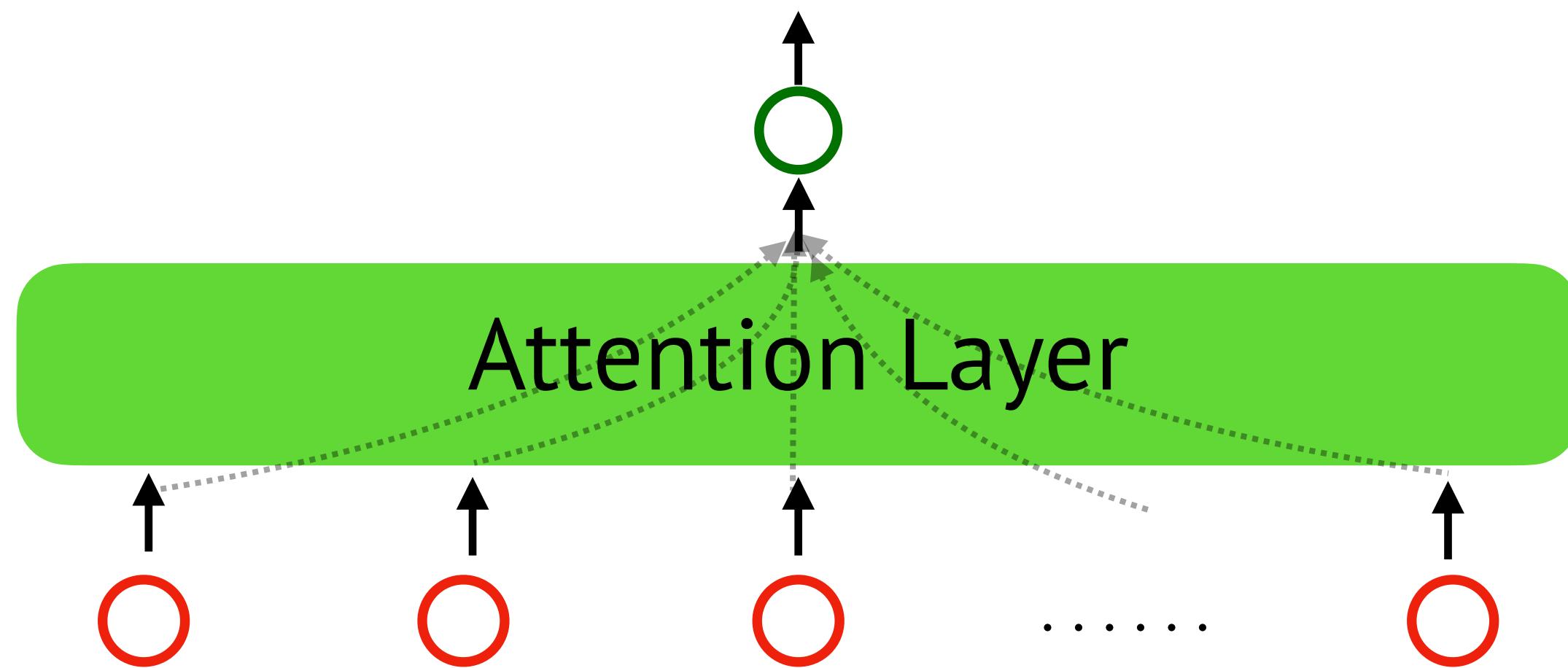
○
Query

$$0.6 \text{ ○} + 0.1 \text{ ○} + 0.2 \text{ ○} + 0.1 \text{ ○} \\ = \text{○} \text{ context vector } \mathbf{c}$$



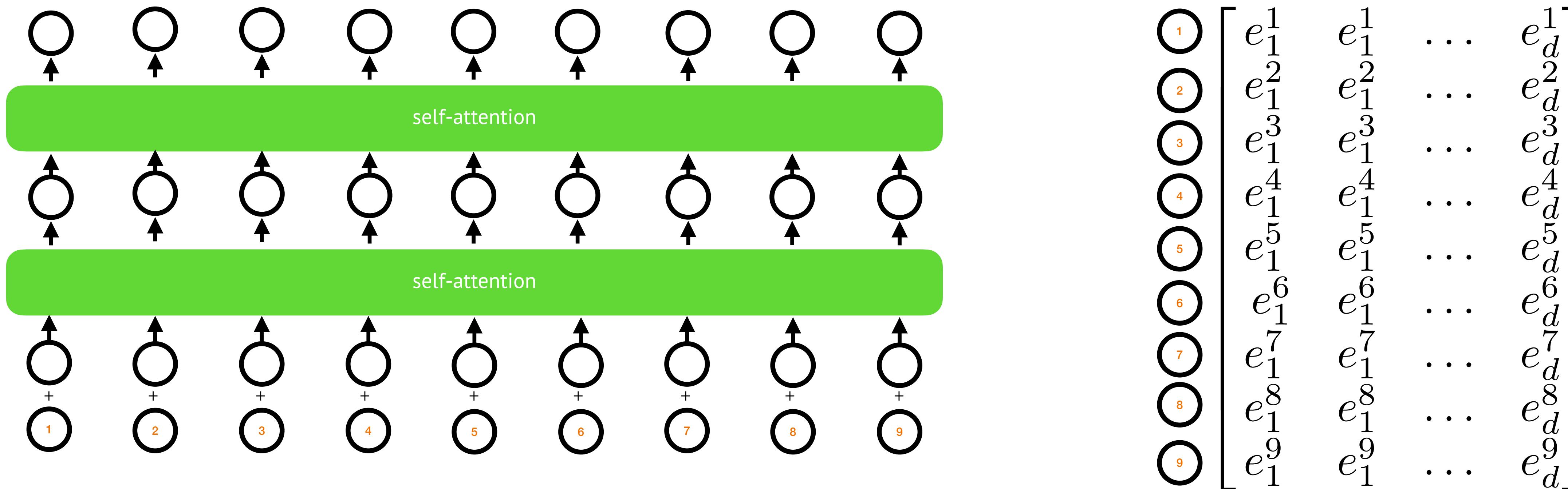
○ ○ ○ ○

Self-attention



This is almost transformer – except a few things.

Transformer (positional embedding)

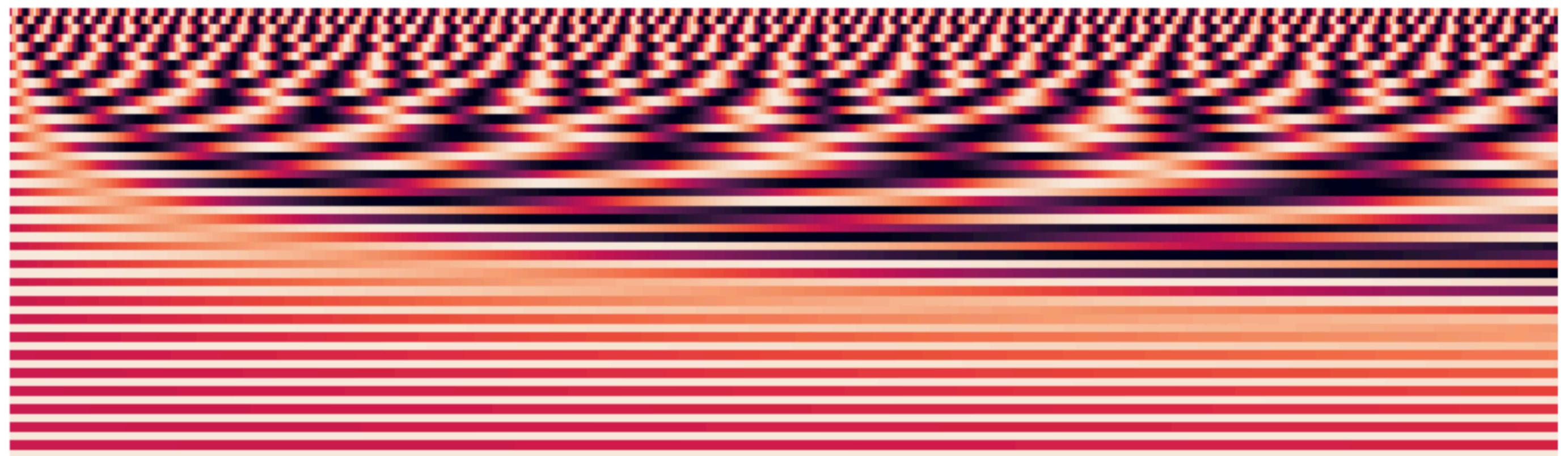


Positional Encoding

$$\begin{bmatrix} \sin\left(\frac{i}{10000^{2 \times \frac{1}{d}}}\right) \\ \cos\left(\frac{i}{10000^{2 \times \frac{1}{d}}}\right) \\ \vdots \\ \vdots \\ \sin\left(\frac{i}{10000^{2 \times \frac{d/2}{d}}}\right) \\ \cos\left(\frac{i}{10000^{2 \times \frac{d/2}{d}}}\right) \end{bmatrix}$$



Dimension



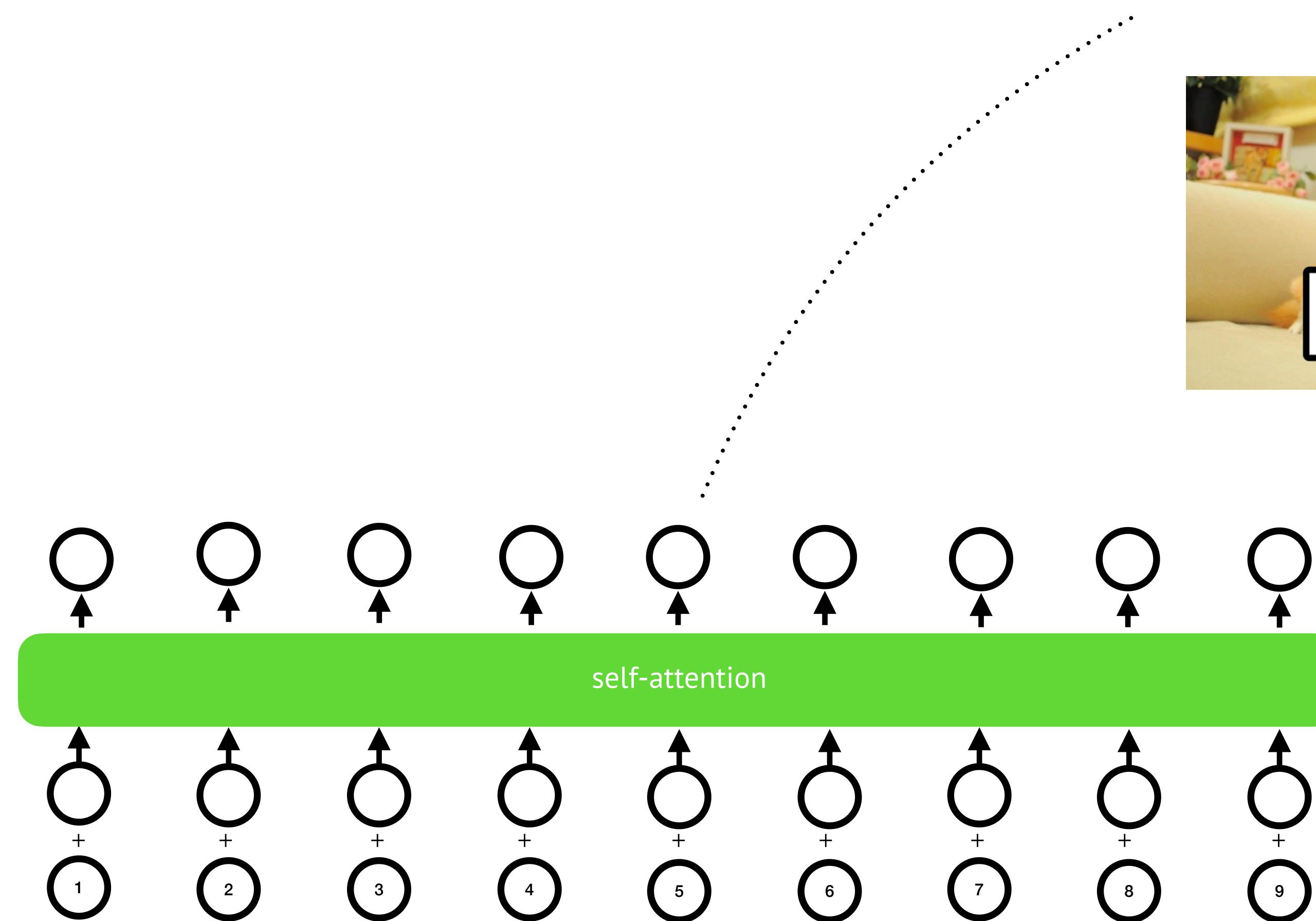
Index in the sequence

The idea of relative position

Positional Encoding

$$\begin{bmatrix} \sin\left(\frac{i}{10000^{2 \times \frac{1}{d}}}\right) \\ \cos\left(\frac{i}{10000^{2 \times \frac{1}{d}}}\right) \\ \vdots \\ \vdots \\ \sin\left(\frac{i}{10000^{2 \times \frac{d/2}{d}}}\right) \\ \cos\left(\frac{i}{10000^{2 \times \frac{d/2}{d}}}\right) \end{bmatrix}$$

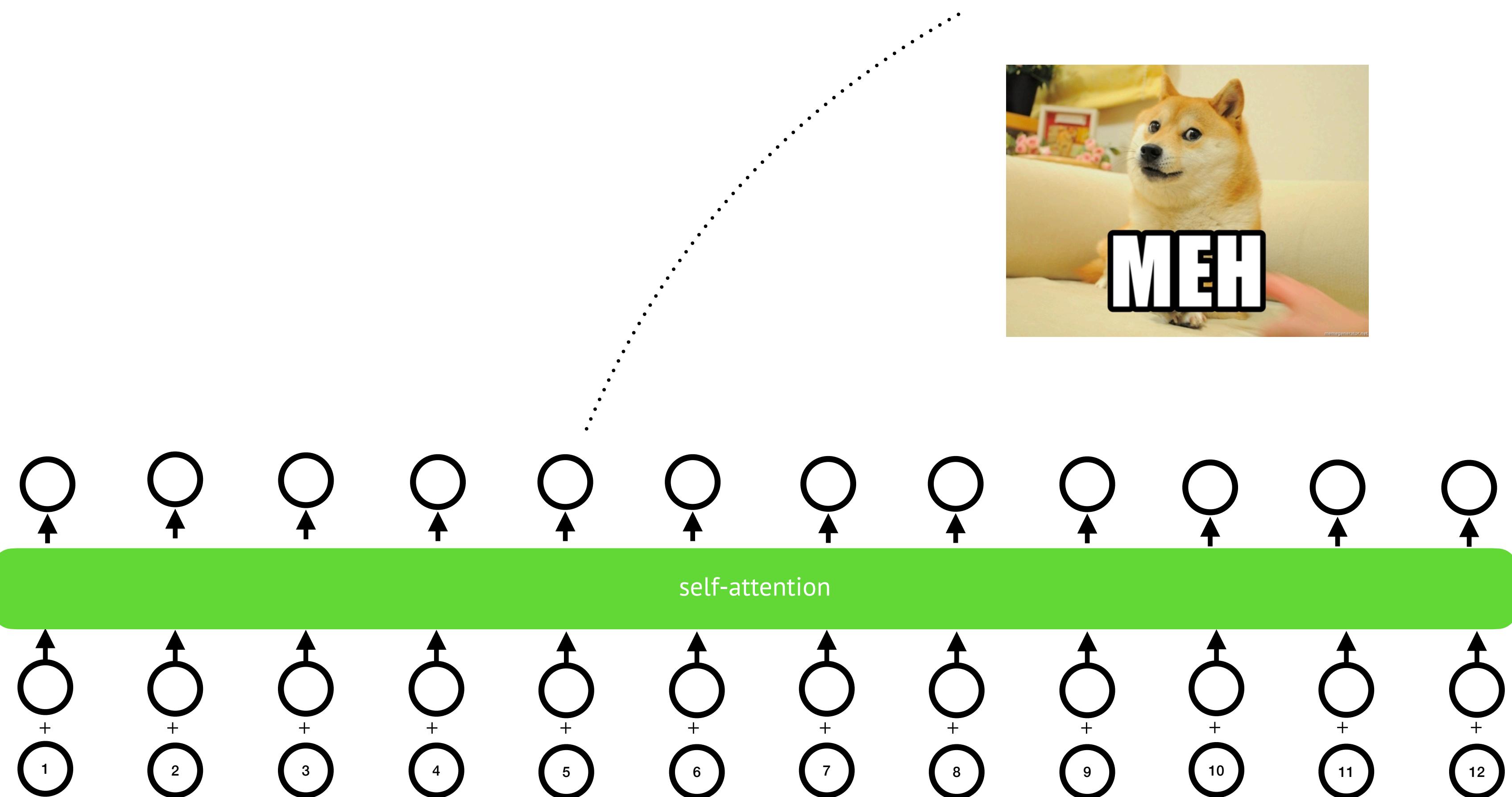
Periodic: Hope this will work in extrapolation. (No)



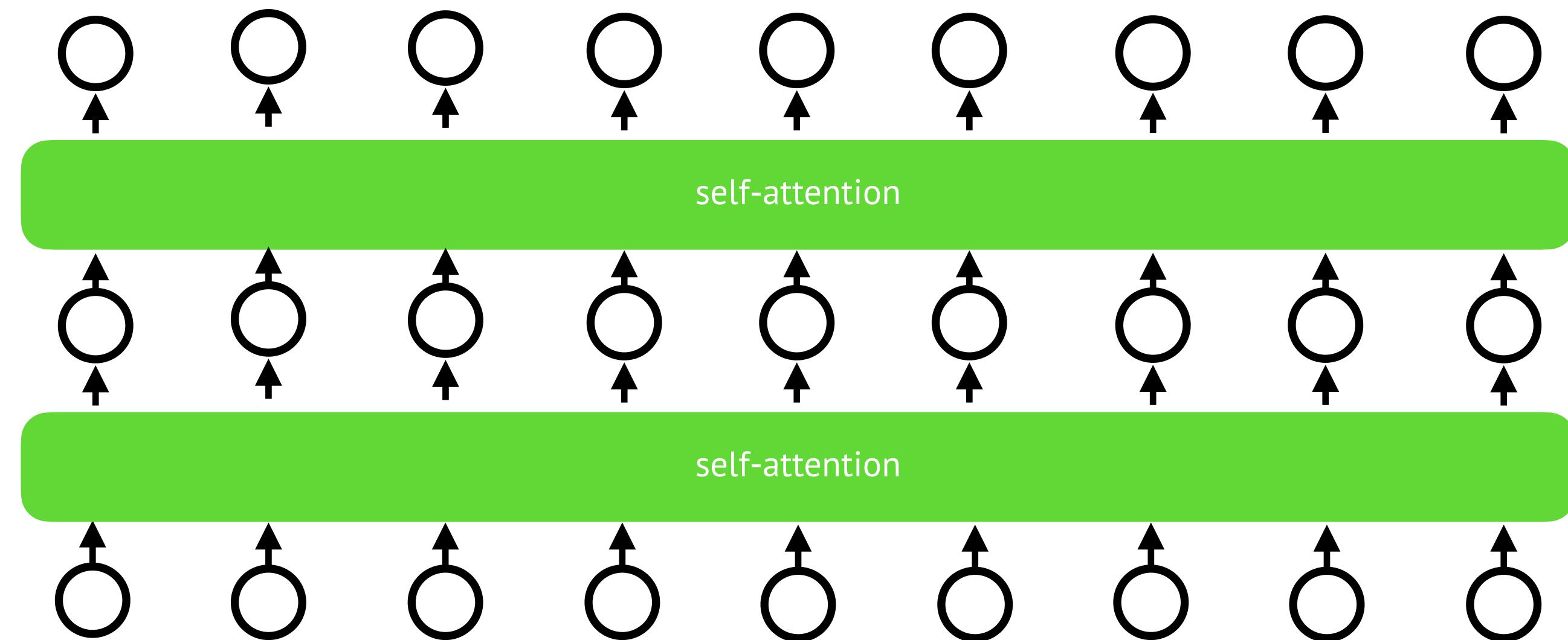
Positional Encoding

$$\begin{bmatrix} \sin\left(\frac{i}{10000^{2 \times \frac{1}{d}}}\right) \\ \cos\left(\frac{i}{10000^{2 \times \frac{1}{d}}}\right) \\ \vdots \\ \vdots \\ \sin\left(\frac{i}{10000^{2 \times \frac{d/2}{d}}}\right) \\ \cos\left(\frac{i}{10000^{2 \times \frac{d/2}{d}}}\right) \end{bmatrix}$$

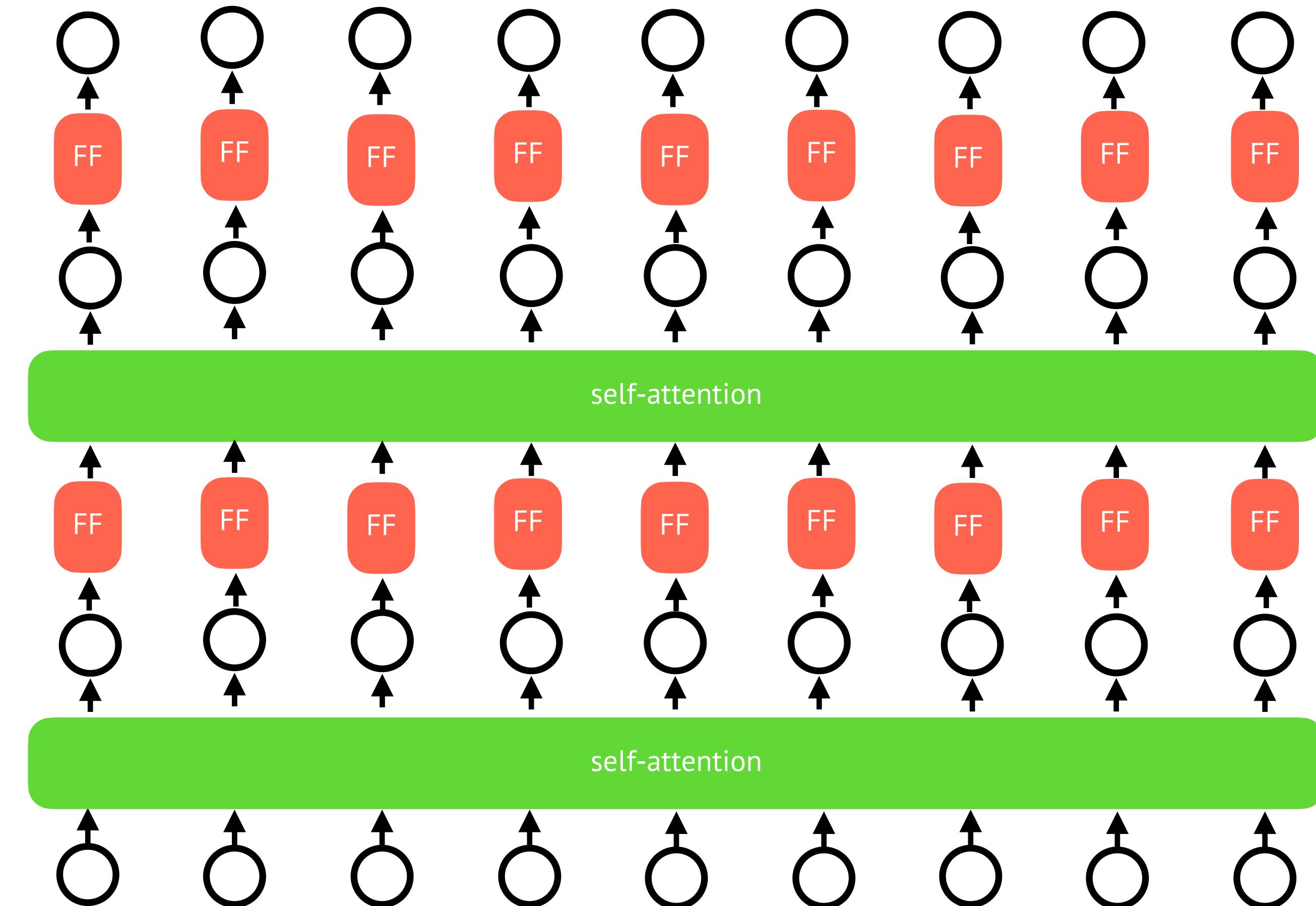
Periodic: Hope this will work in extrapolation. (No)



Feed Forward Layer



Feed Forward Layer

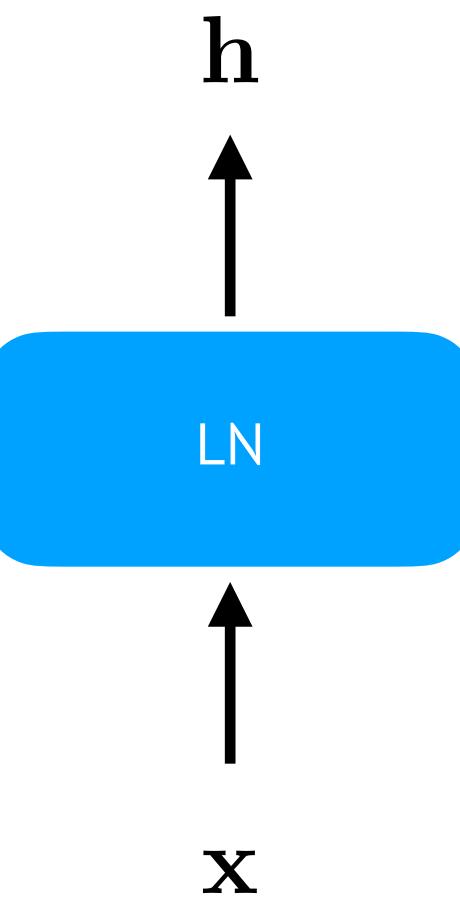


Layer Normalization (Ba et al, 2016)

$$\mathbf{h} = \mathbf{g} \odot N(\mathbf{x}) + \mathbf{b}$$

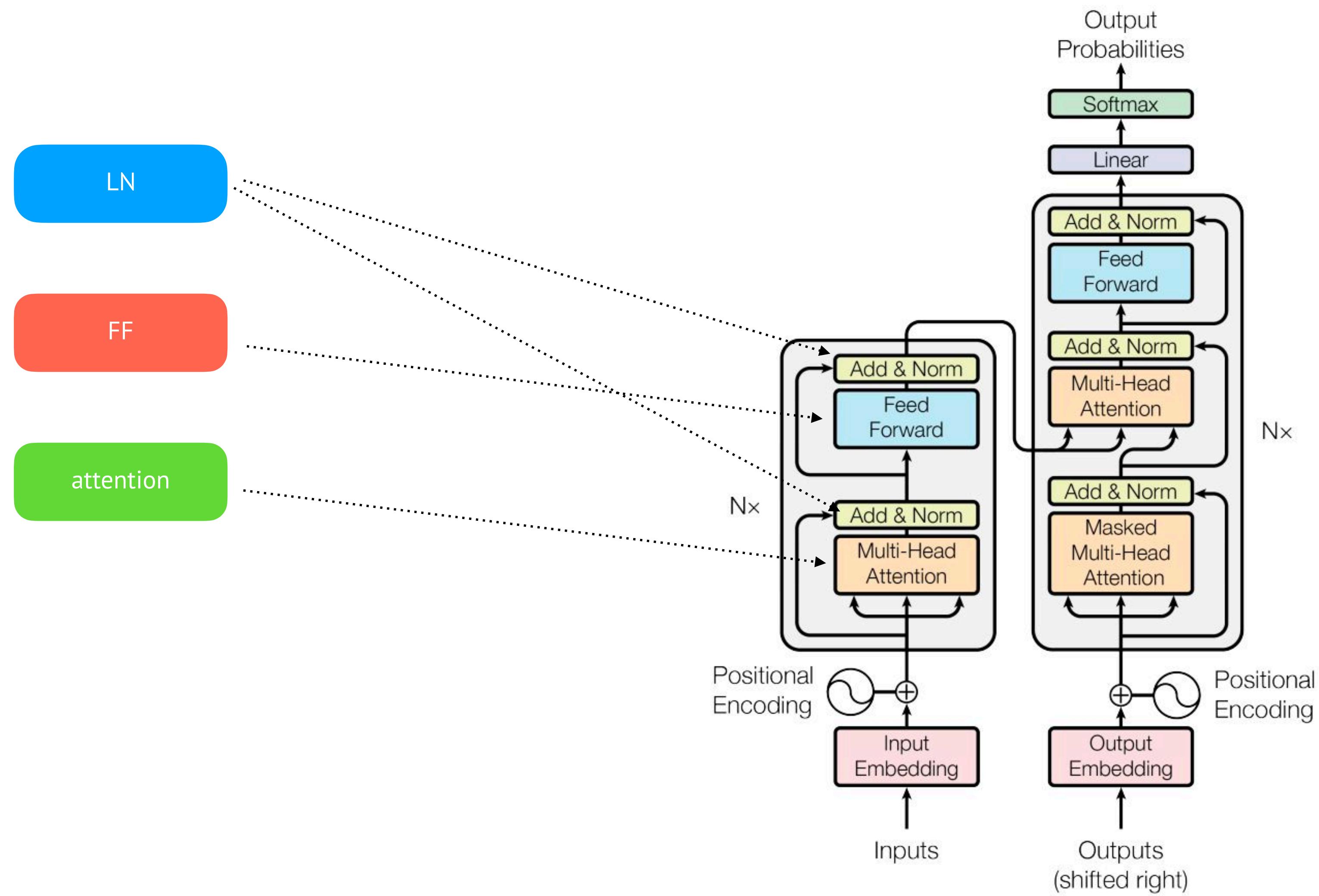
$$N(\mathbf{x}) = \frac{\mathbf{x} - \mu}{\sigma}$$

$$\mu = \frac{1}{H} \sum_{i=1}^H x_i \quad \sigma = \sqrt{\frac{1}{H} \sum_{i=1}^H (x_i - \mu)^2}$$



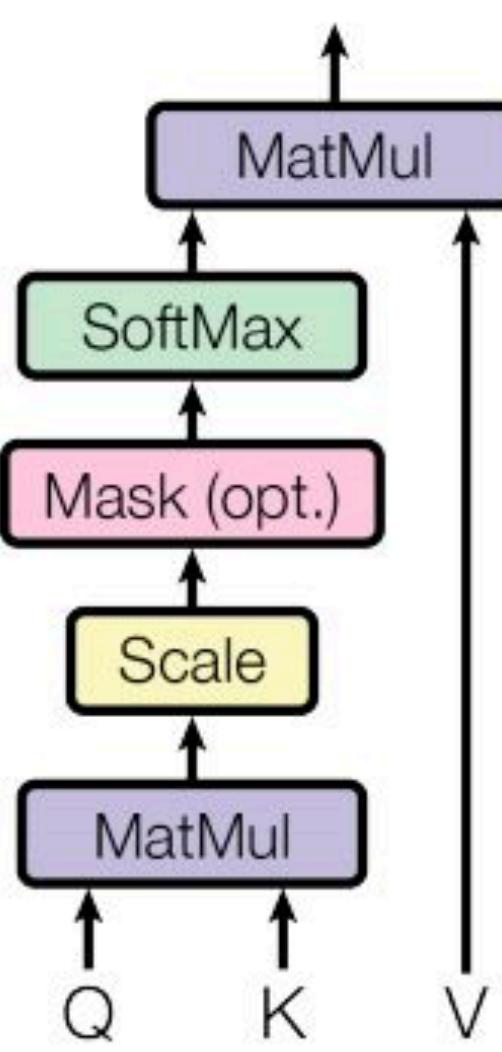
Smoother gradients, faster training and better generalization accuracy. ([Xu et al, Neurips 2019](#))

Layer Normalization

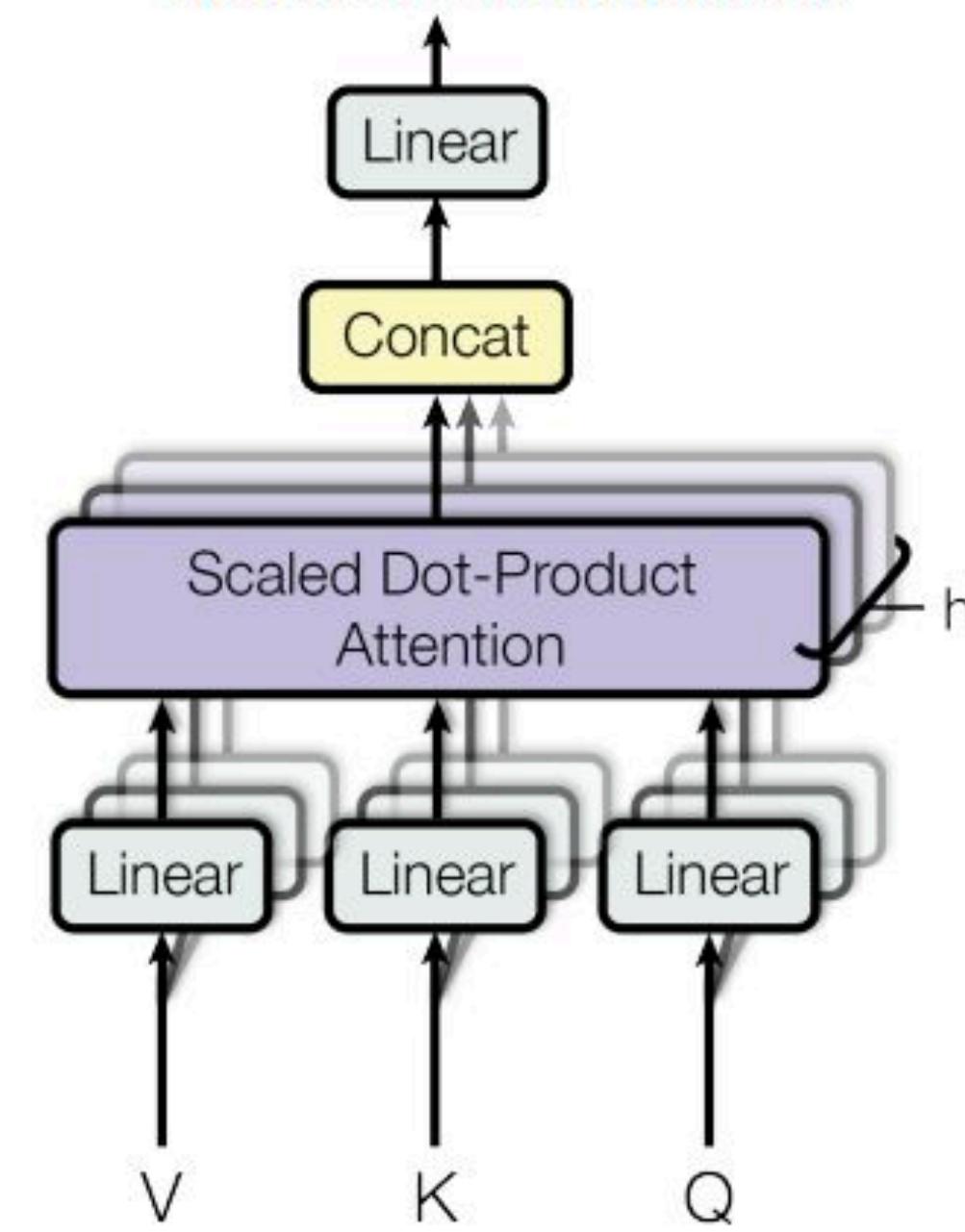


Multi-head Attention

Scaled Dot-Product Attention



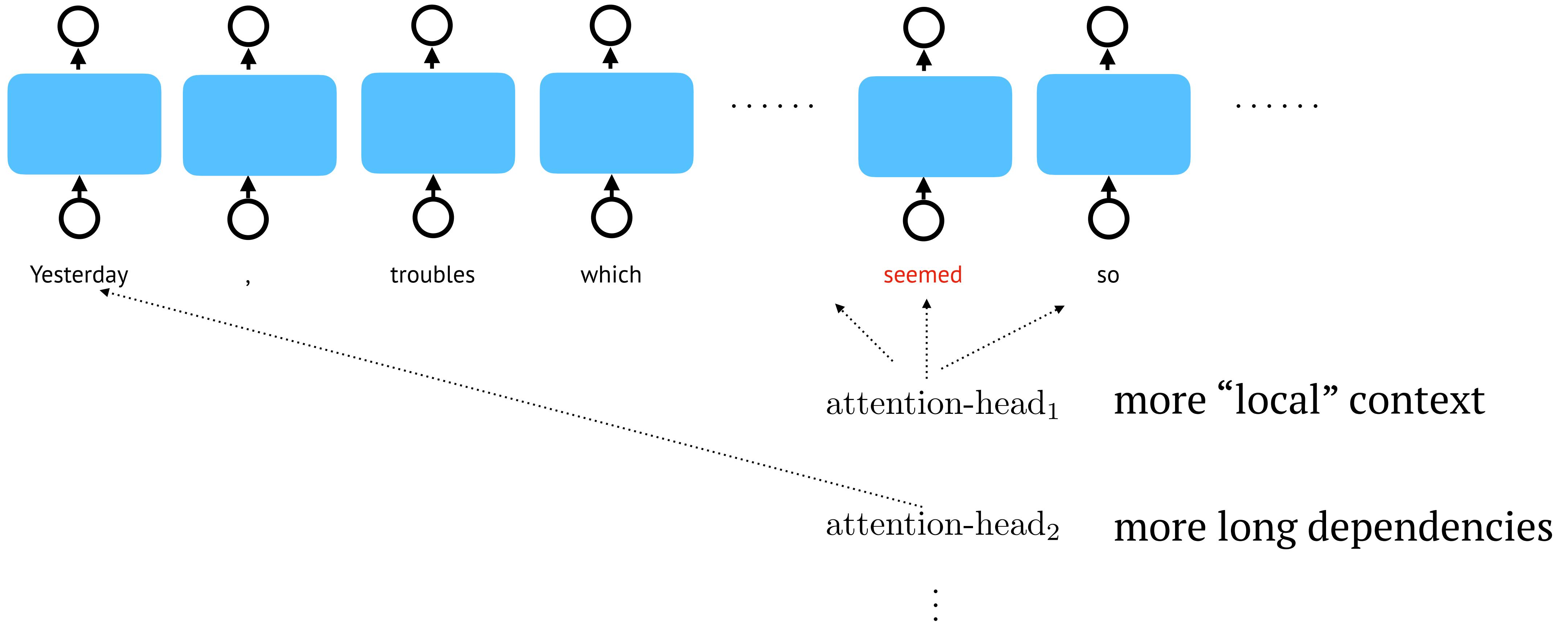
Multi-Head Attention



$$\text{score}(q, k) = \frac{q^T k}{\sqrt{d_k}}$$

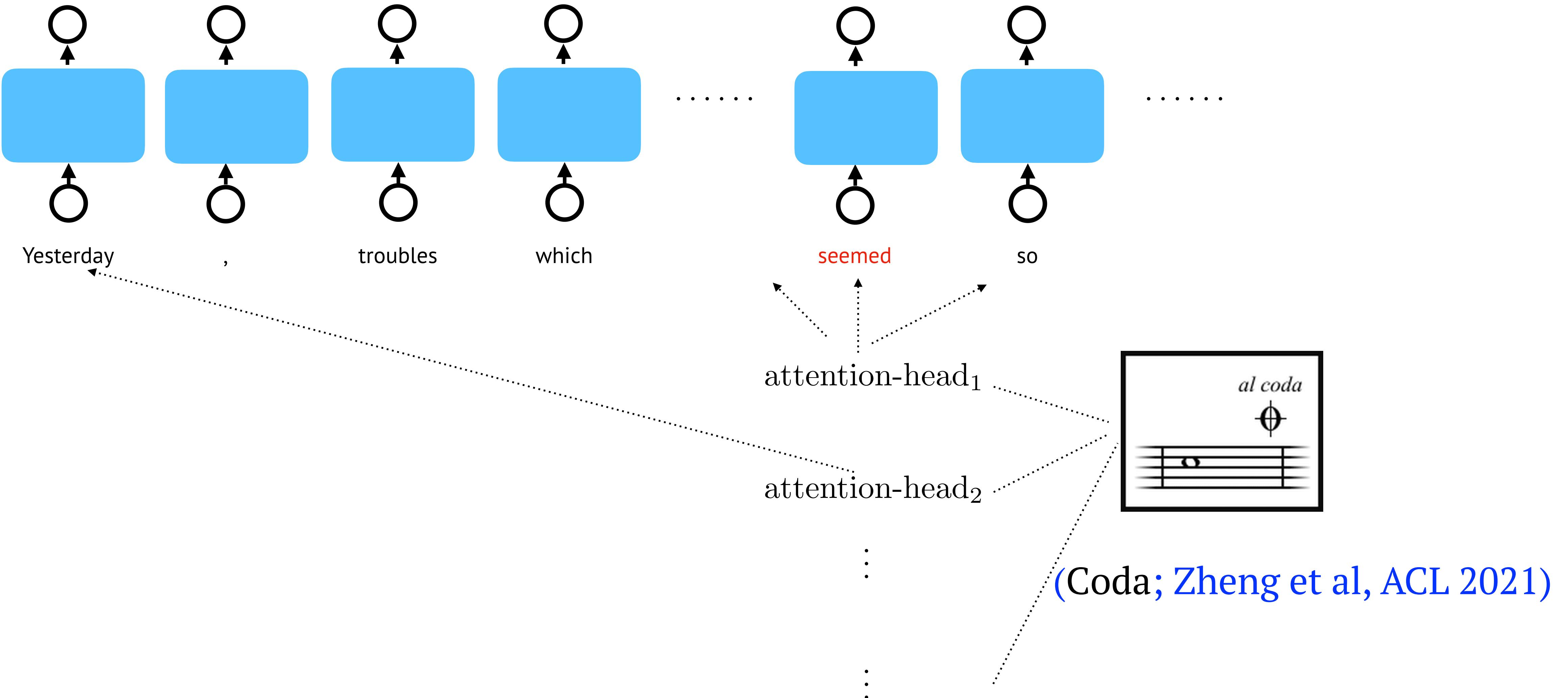
multiple copies

Multi-head Attention

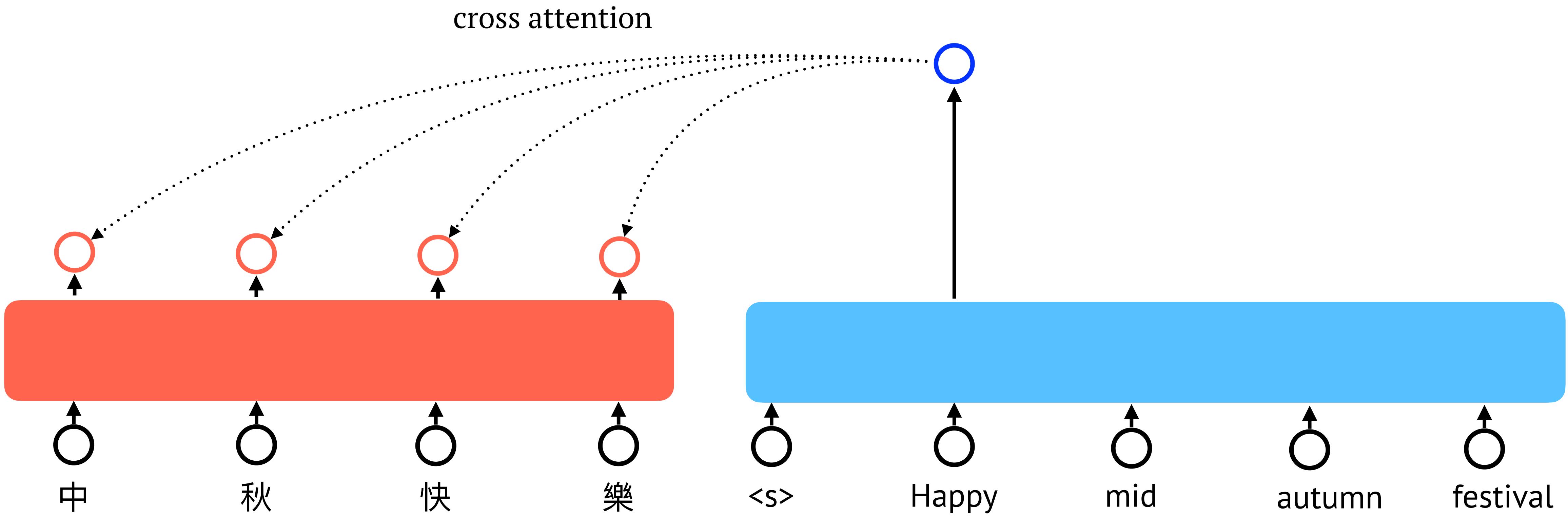


Improve the “resolution” of the attention mechanism.

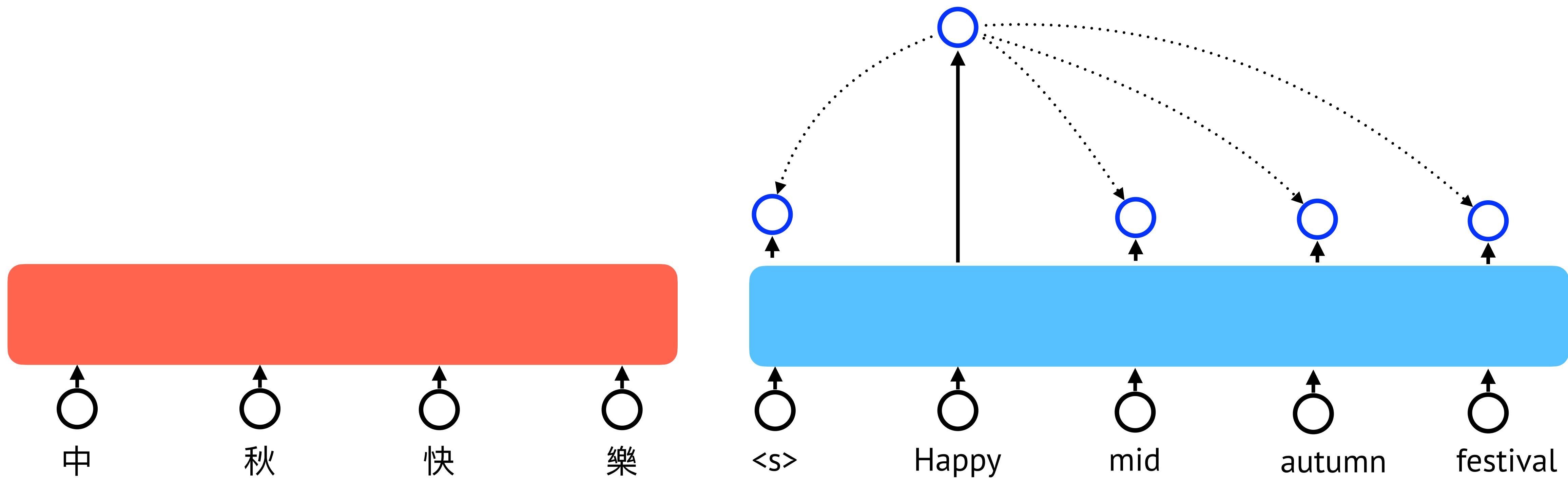
Multi-head Attention



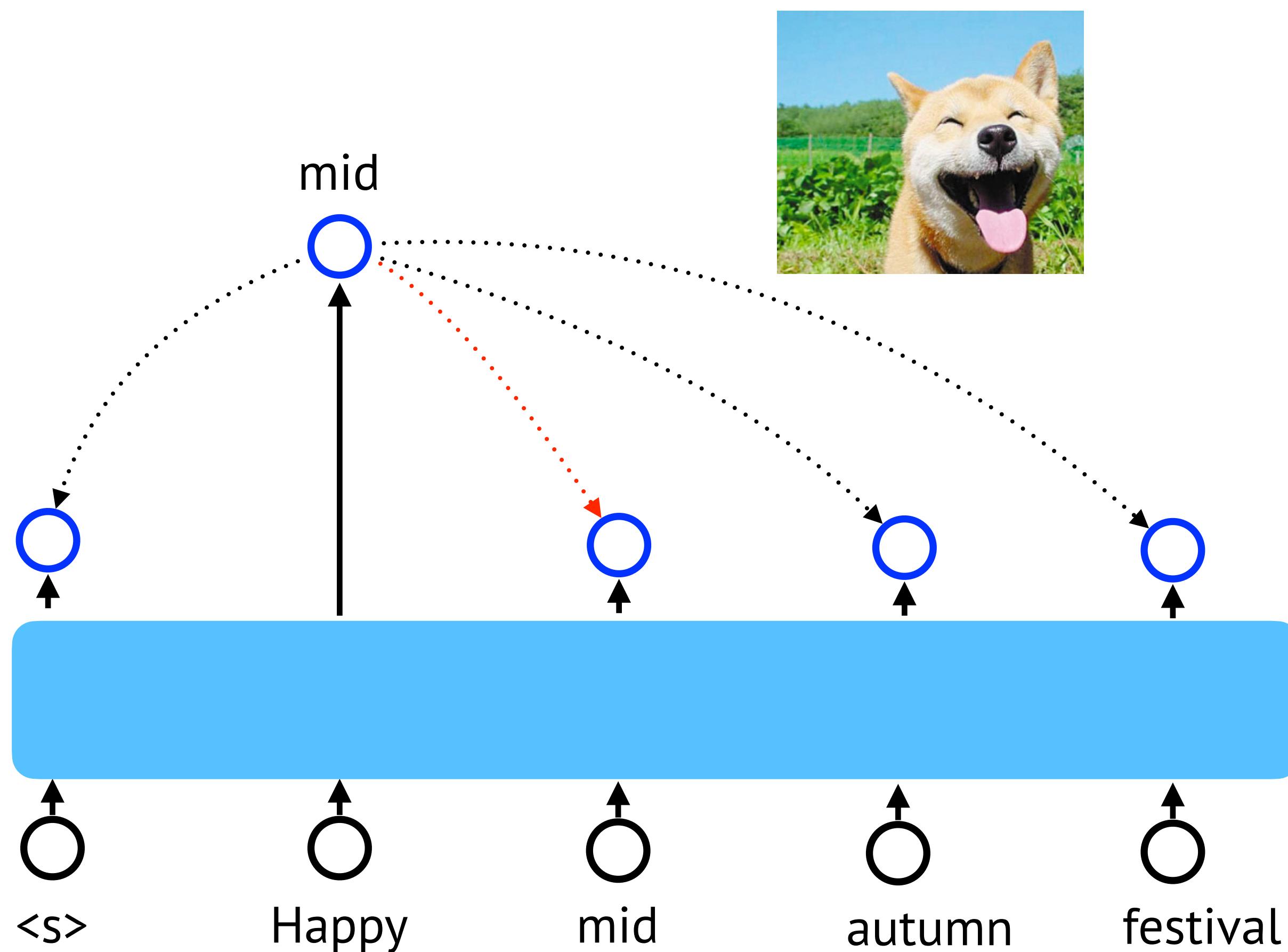
Transformer as Decoder



Transformer as Decoder



Transformer as Decoder



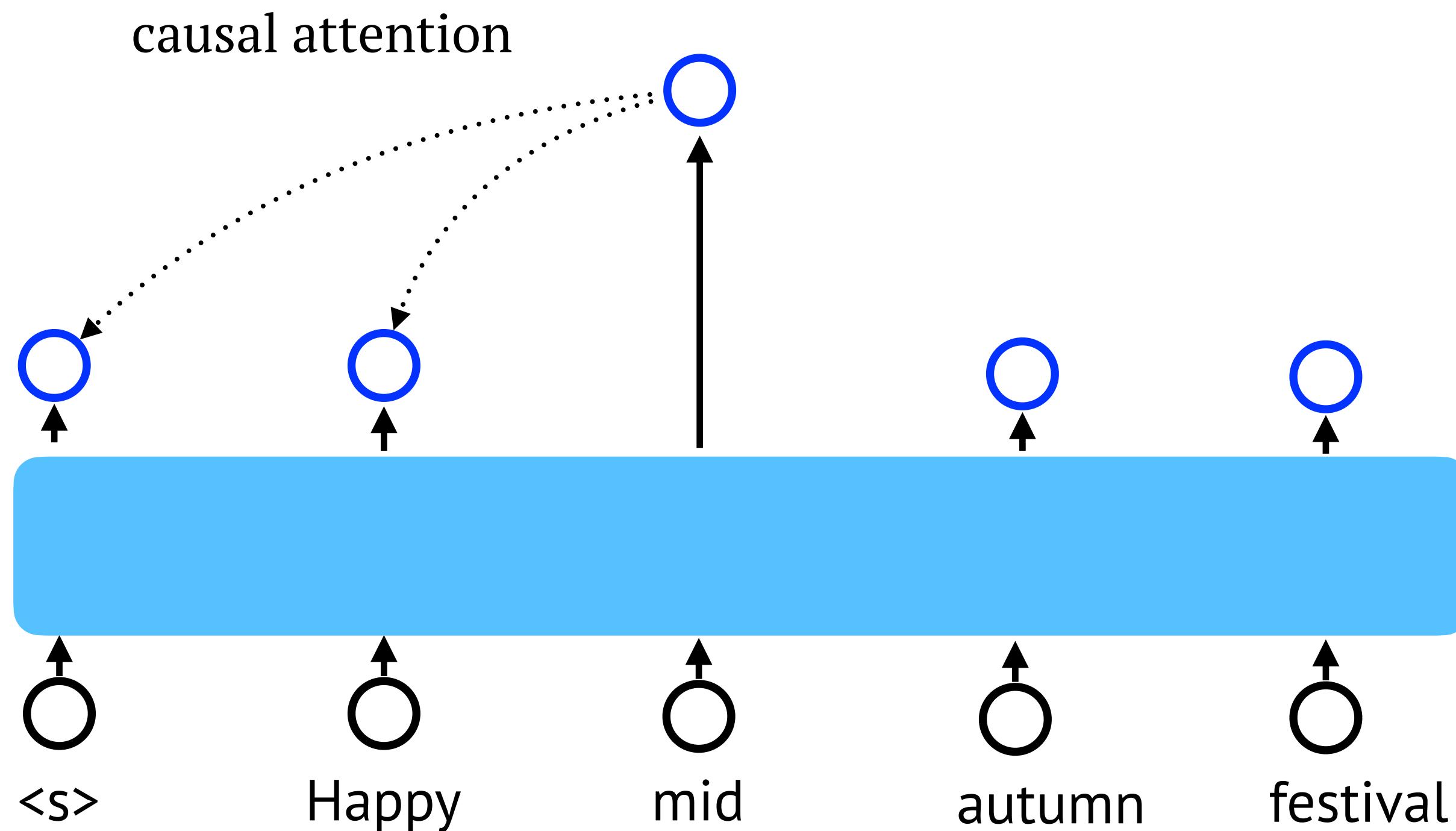
Need to prevent the attention the future words.

	Happy	mid	autumn	festival
Happy	$-\infty$	$-\infty$	$-\infty$	$-\infty$
mid		$-\infty$	$-\infty$	$-\infty$
autumn			$-\infty$	$-\infty$
festival				$-\infty$

$$e_{ij} = \begin{cases} q_i^\top k_j, & j < i \\ -\infty, & j \geq i \end{cases}$$

Transformer as Decoder

Need to prevent the attention the future words.

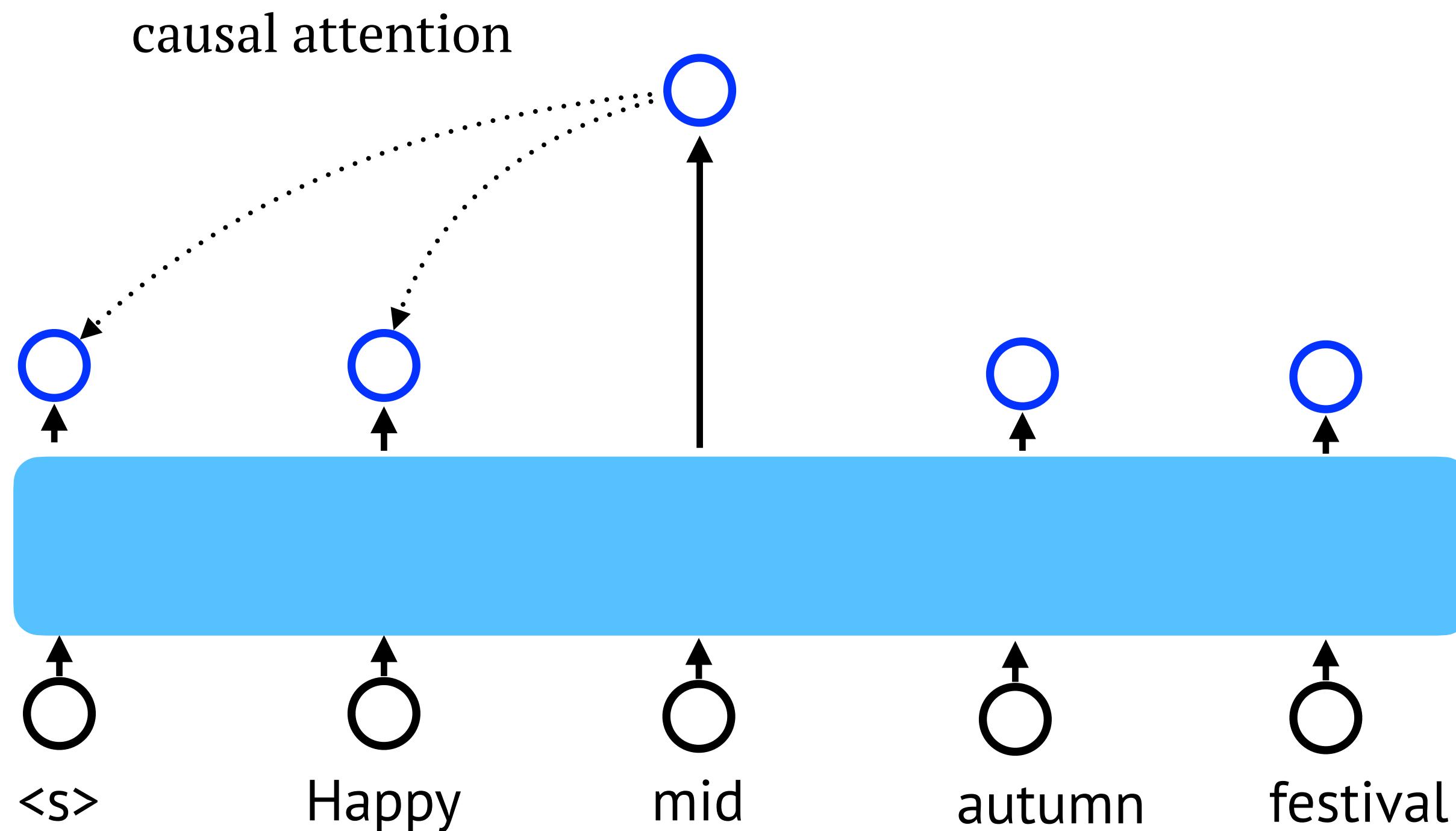


	Happy	mid	autumn	festival
Happy	−∞	−∞	−∞	−∞
mid		−∞	−∞	−∞
autumn			−∞	−∞
festival				−∞

$$e_{ij} = \begin{cases} q_i^\top k_j, & j < i \\ -\infty, & j \geq i \end{cases}$$

Transformer as Decoder

Need to prevent the attention the future words.

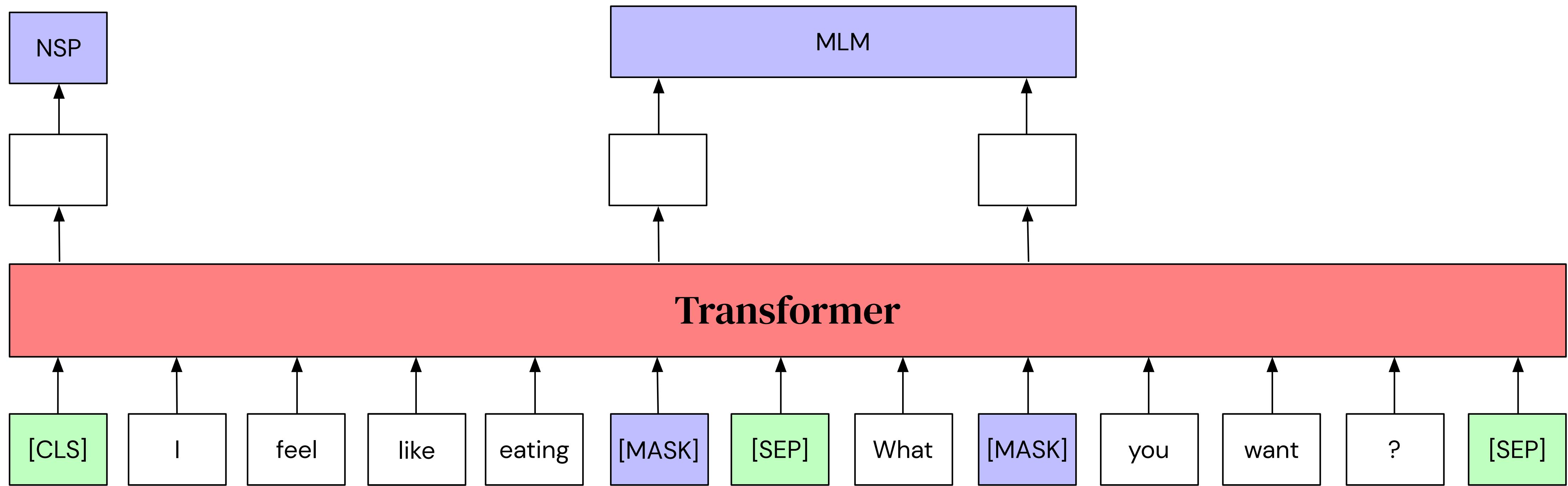


	Happy	mid	autumn	festival
Happy	−∞	−∞	−∞	−∞
mid		−∞	−∞	−∞
autumn			−∞	−∞
festival				−∞

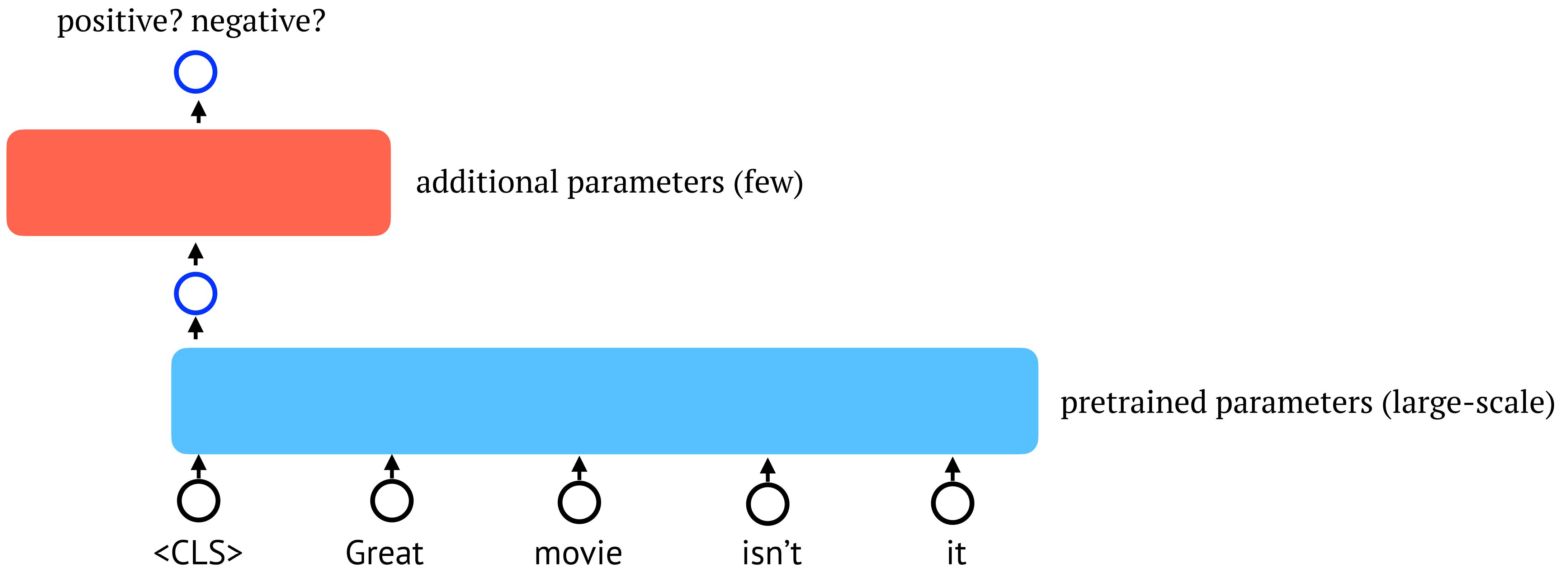
$$e_{ij} = \begin{cases} q_i^\top k_j, & j < i \\ -\infty, & j \geq i \end{cases}$$

Transformer for Pretraining

$$\mathbb{E}_{p(x_i, \hat{x}_i)}[p(x_i \mid \hat{x}_i)]$$



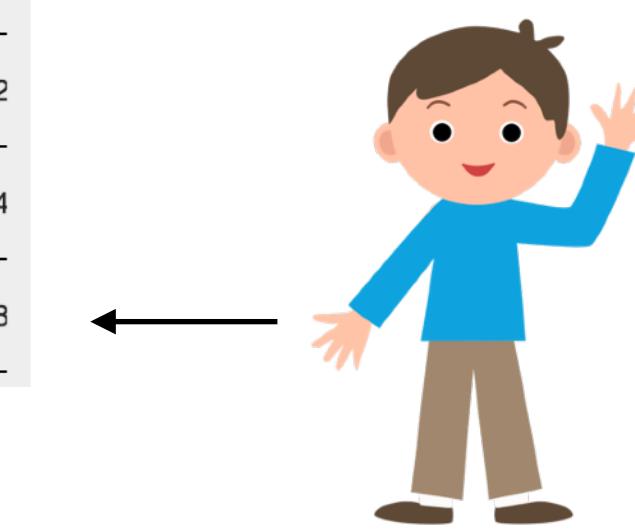
Transformer for Finetuning



GLUE Benchmark

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm
1	ERNIE Team - Baidu	ERNIE	🔗	91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7
2	AliceMind & DIRL	StructBERT + CLEVER	🔗	91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5
3	DeBERTa Team - Microsoft	DeBERTa / TuringNLv4	🔗	90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6
4	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1
+ 5	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3
6	Liangzhu Ge	Deberta + CLEVER		90.5	72.7	97.5	92.7/90.3	93.2/92.9	76.3/90.8	92.1	91.7
7	T5 Team - Google	T5	🔗	90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9
8	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART	🔗	89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8
+ 9	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	91.5	91.3
+ 10	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)	🔗	89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1
+ 11	ELECTRA Team	ELECTRA-Large + Standard Tricks	🔗	89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8
+ 12	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	🔗	88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7
13	Junjie Yang	HIRE-RoBERTa	🔗	88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4
14	Facebook AI	RoBERTa	🔗	88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2
+ 15	Microsoft D365 AI & MSR AI	MT-DNN-ensemble	🔗	87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4
16	GLUE Human Baselines	GLUE Human Baselines	🔗	87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8

Sept 27, 2021



GLUE Benchmark

QQP: Quora Question Pairs (detect paraphrase questions)

SST-2: Sentiment analysis

.....

Problem solve?

Context: Aaron is an editor. Mark is an actor.

Question: Who is not an actor?

Correct Answer: **Aaron**

BERT Prediction: **Mark**

Context: Jose hates Lisa. Kevin is hated by Lisa.

Question: Who hates Kevin?

Correct Answer: **Lisa**

BERT Prediction: **Jose**

(Ribeiro et al., 2020)

Adversarial Attacks

Dataset				Label
MNLI	Ori	Some rooms have balconies .	Hypothesis	All of the rooms have balconies off of them . Contradiction
	Adv	Many rooms have balconies .	Hypothesis	All of the rooms have balconies off of them . Neutral
IMDB	Ori	it is hard for a lover of the novel northanger abbey to sit through this bbc adaptation and to keep from throwing objects at the tv screen... why are so many facts concerning the tilney family and mrs . tilney ' s death altered unnecessarily ? to make the story more ' horrible ? '		Negative
	Adv	it is hard for a lover of the novel northanger abbey to sit through this bbc adaptation and to keep from throwing objects at the tv screen... why are so many facts concerning the tilney family and mrs . tilney ' s death altered unnecessarily ? to make the plot more ' horrible ? '		Positive
IMDB	Ori	i first seen this movie in the early 80s .. it really had nice picture quality too . anyways , i 'm glad i found this movie again ... the part i loved best was when he hijacked the car from this poor guy... this is a movie i could watch over and over again . i highly recommend it .		Positive
	Adv	i first seen this movie in the early 80s .. it really had nice picture quality too . anyways , i 'm glad i found this movie again ... the part i loved best was when he hijacked the car from this poor guy... this is a movie i could watch over and over again . i inordinately recommend it .		Negative

(Li et al., 2020)