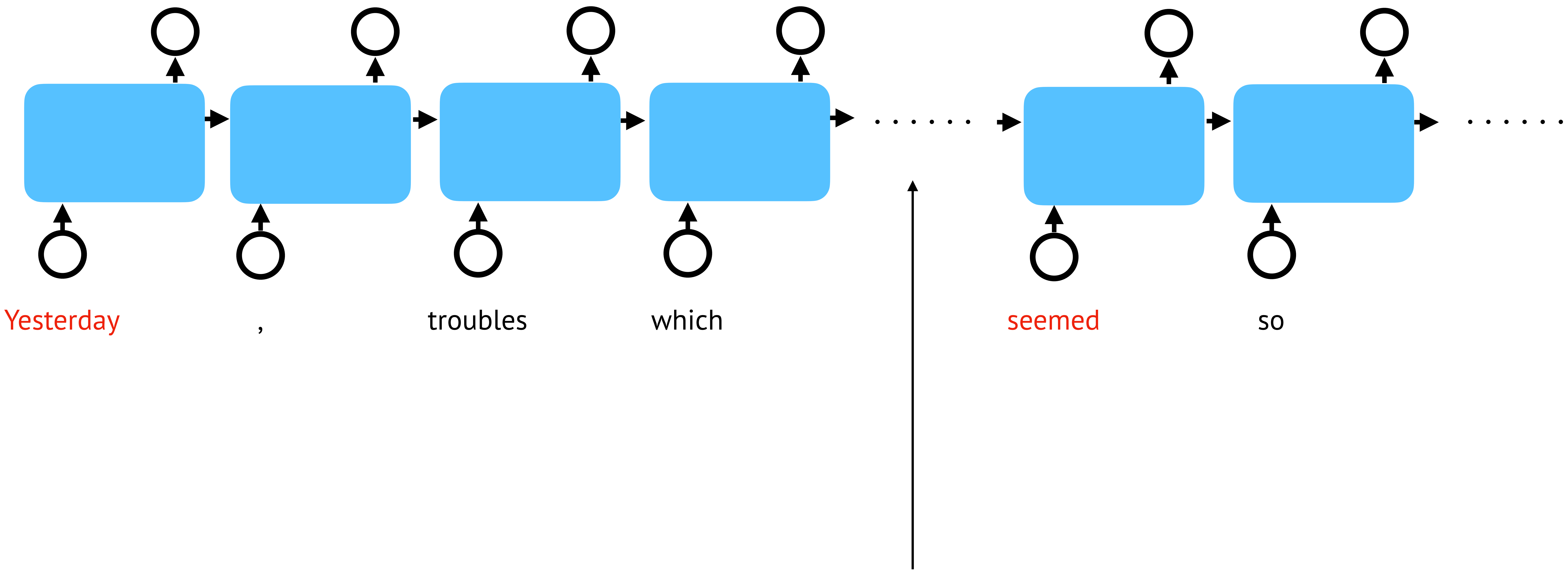
Transformers

COMP7607 — Lecture 4

Lingpeng Kong

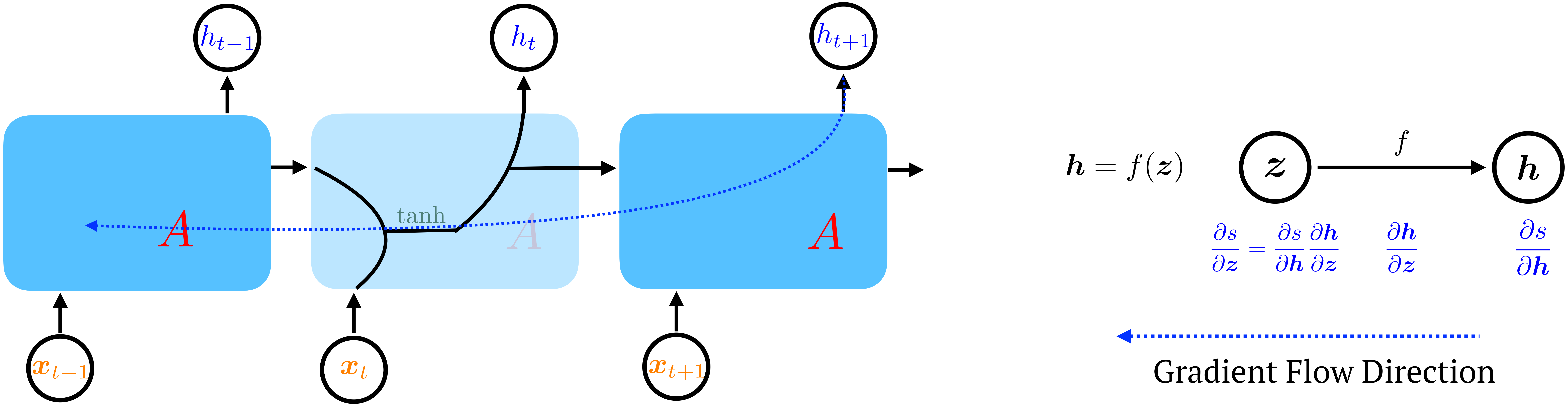
Department of Computer Science, The University of Hong Kong

Recurrent Neural Network



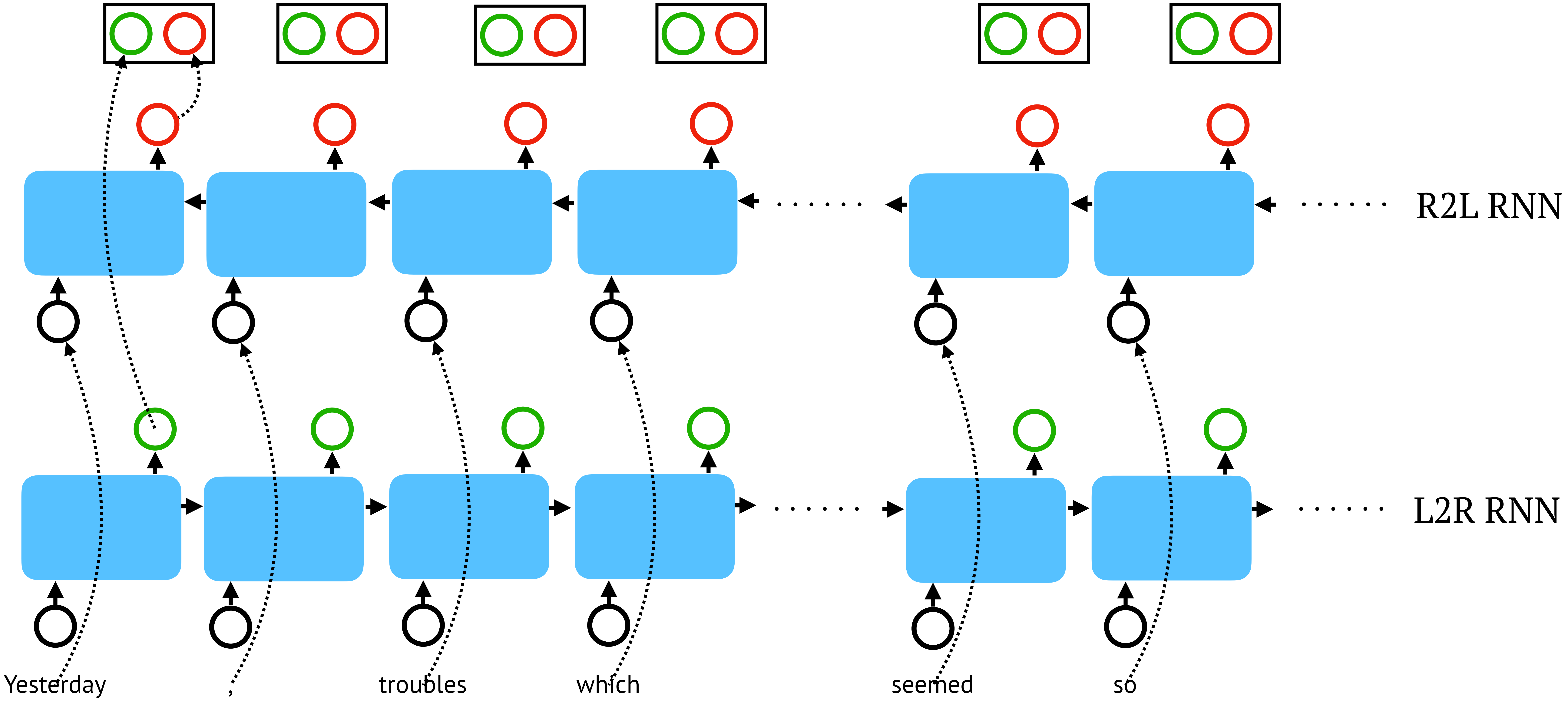
Possibly many steps [O(N)] steps before “yesterday” and “seemed” interact.

Vanishing Gradient in RNNs

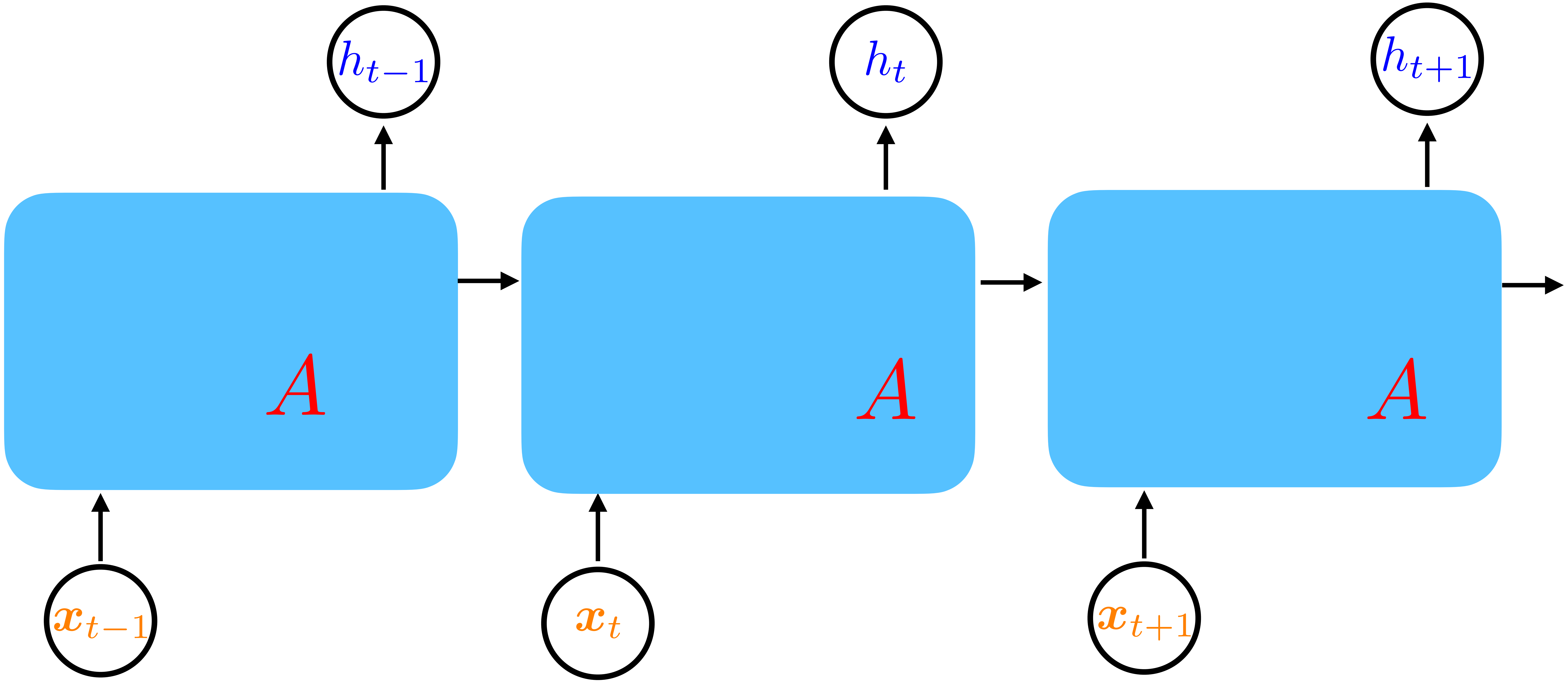


In general, the longer the path, the smaller the gradient signal.

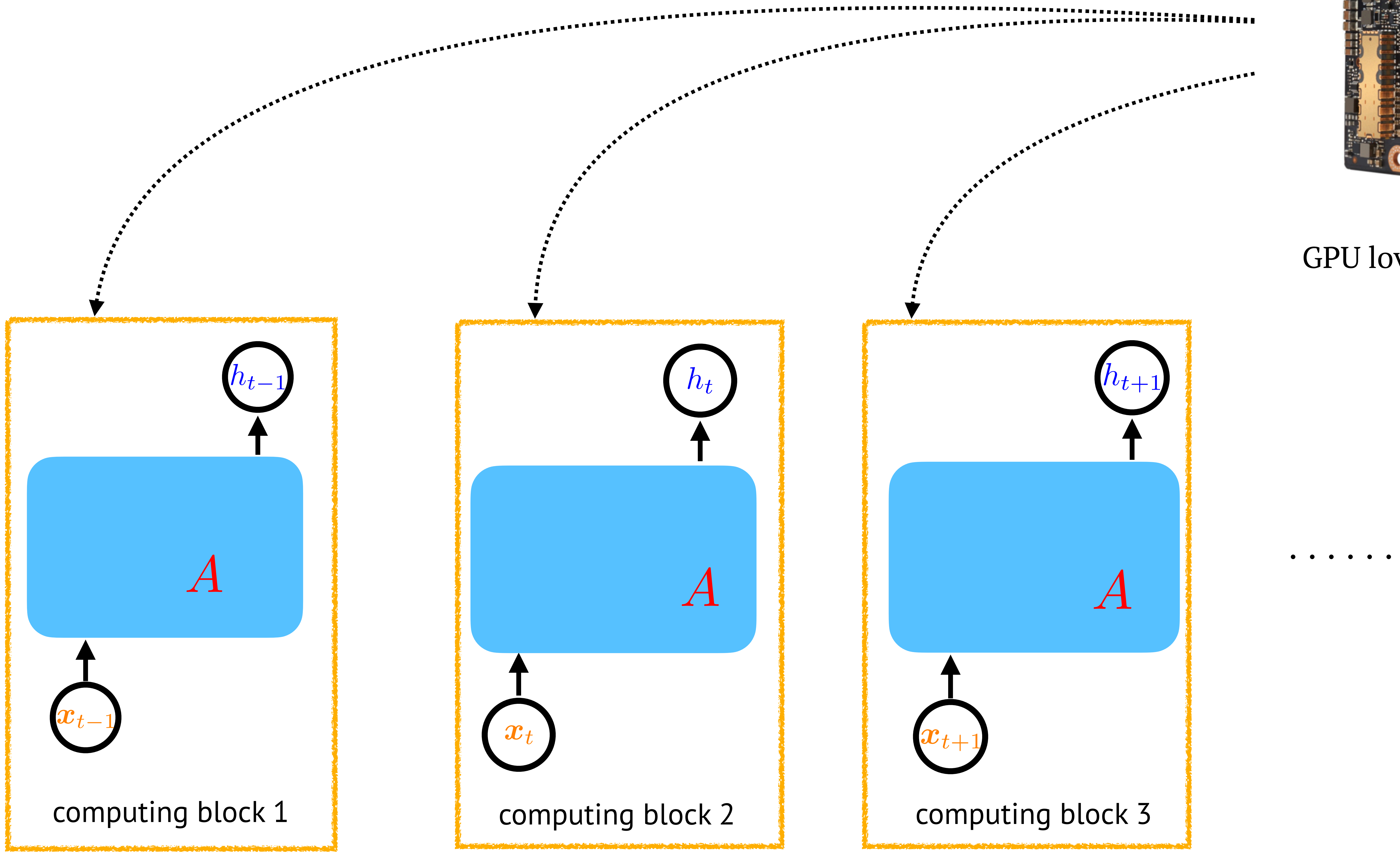
Bidirectional Recurrent Neural Network

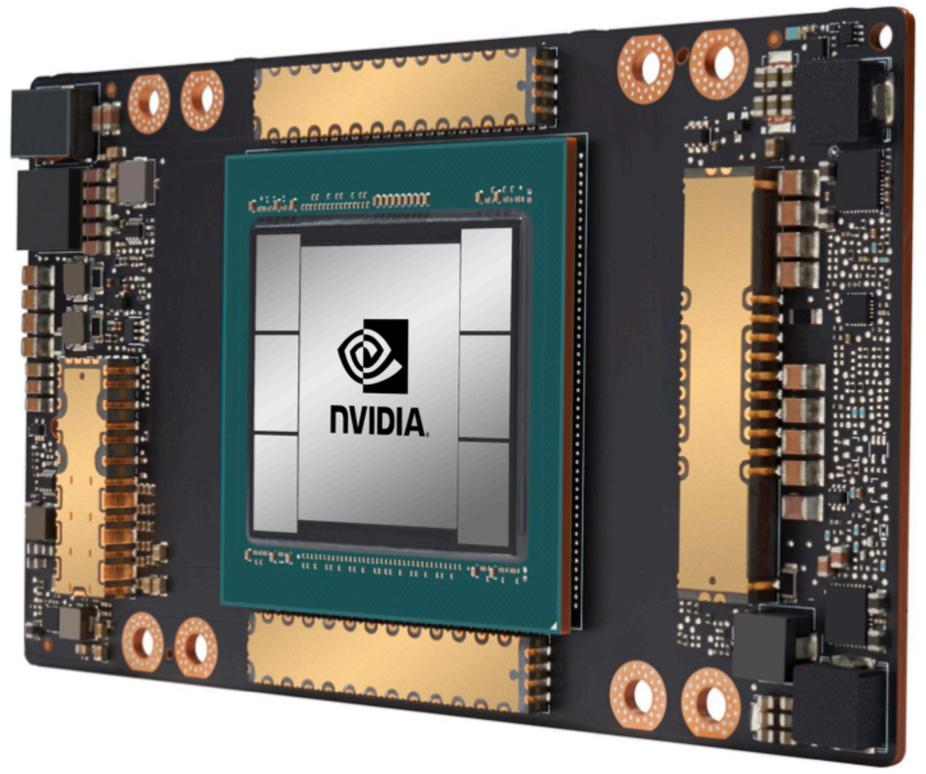


Sequential Computation



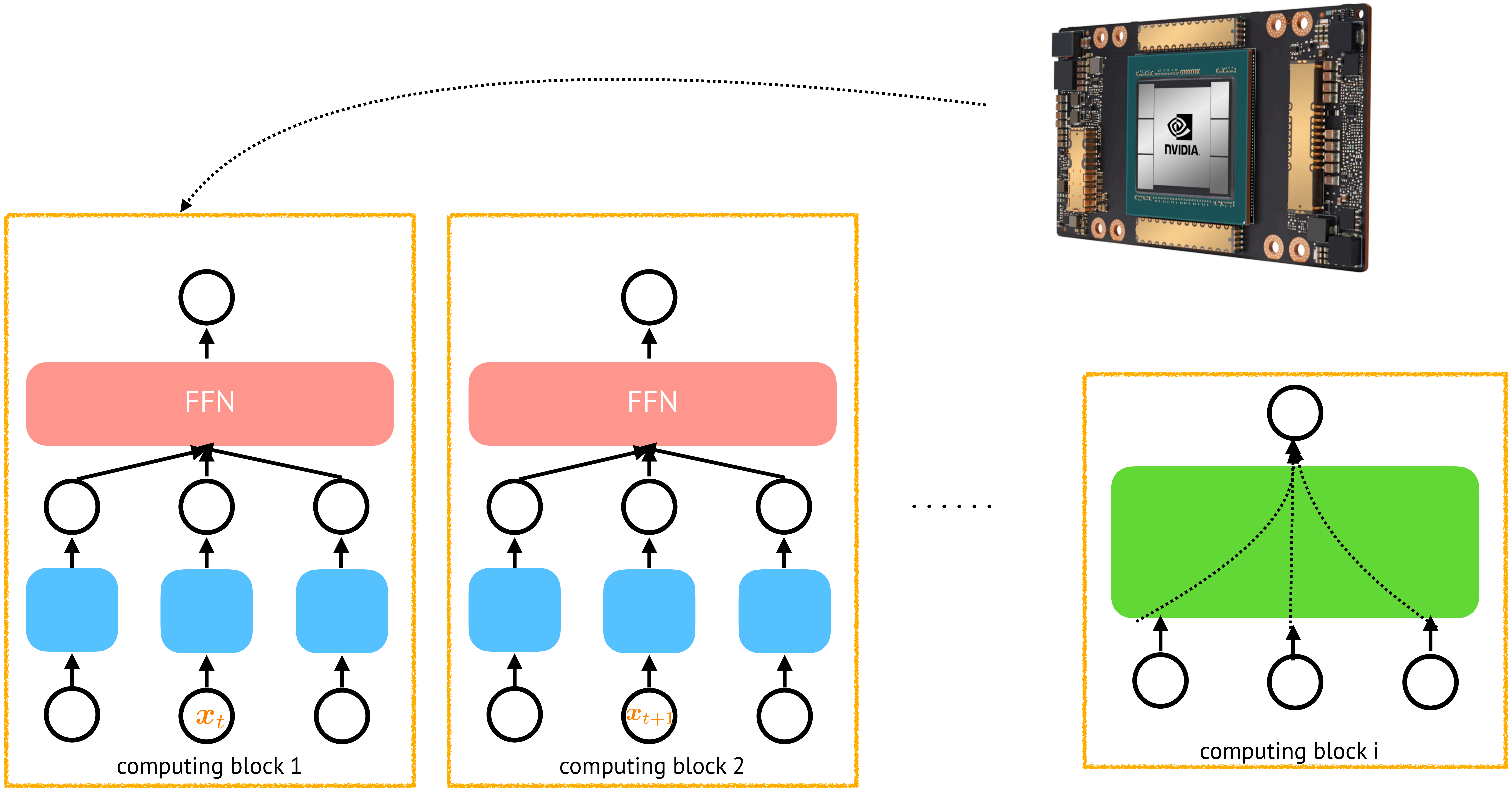
Parallel Computing?



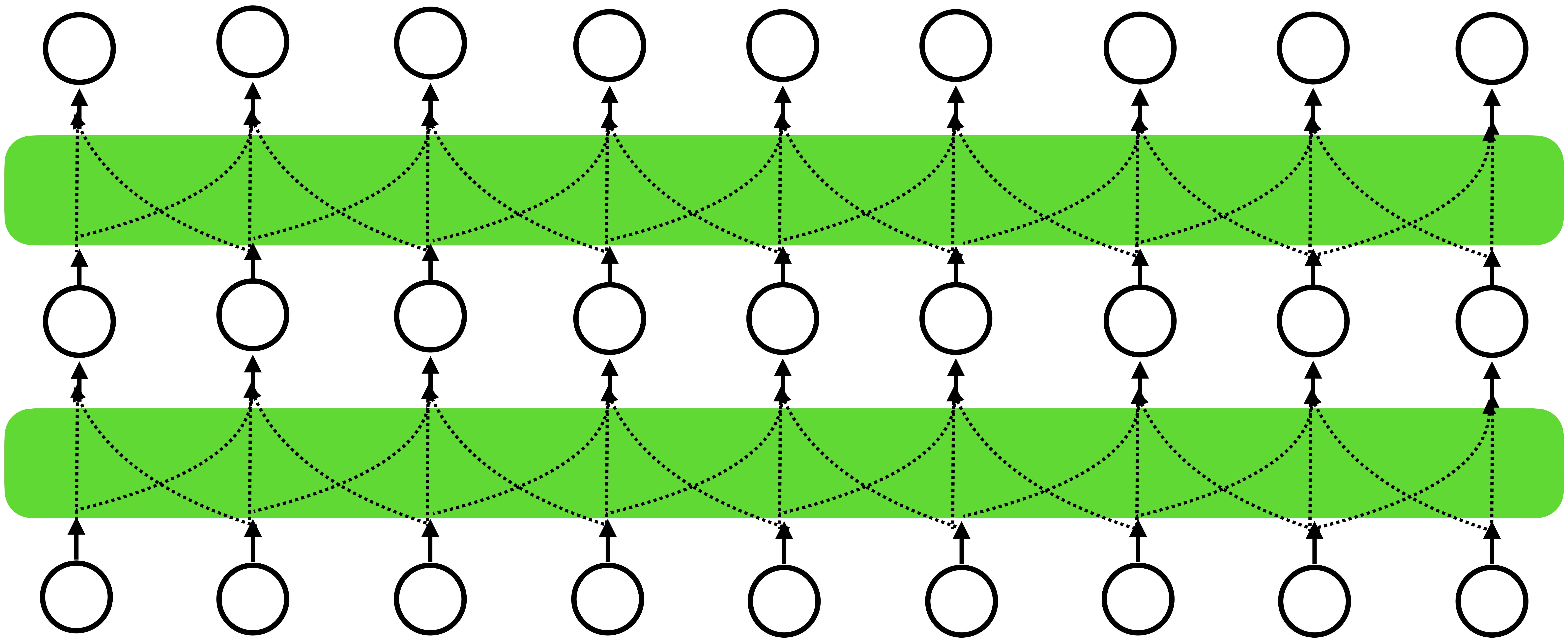


GPU loves parallel computing blocks!

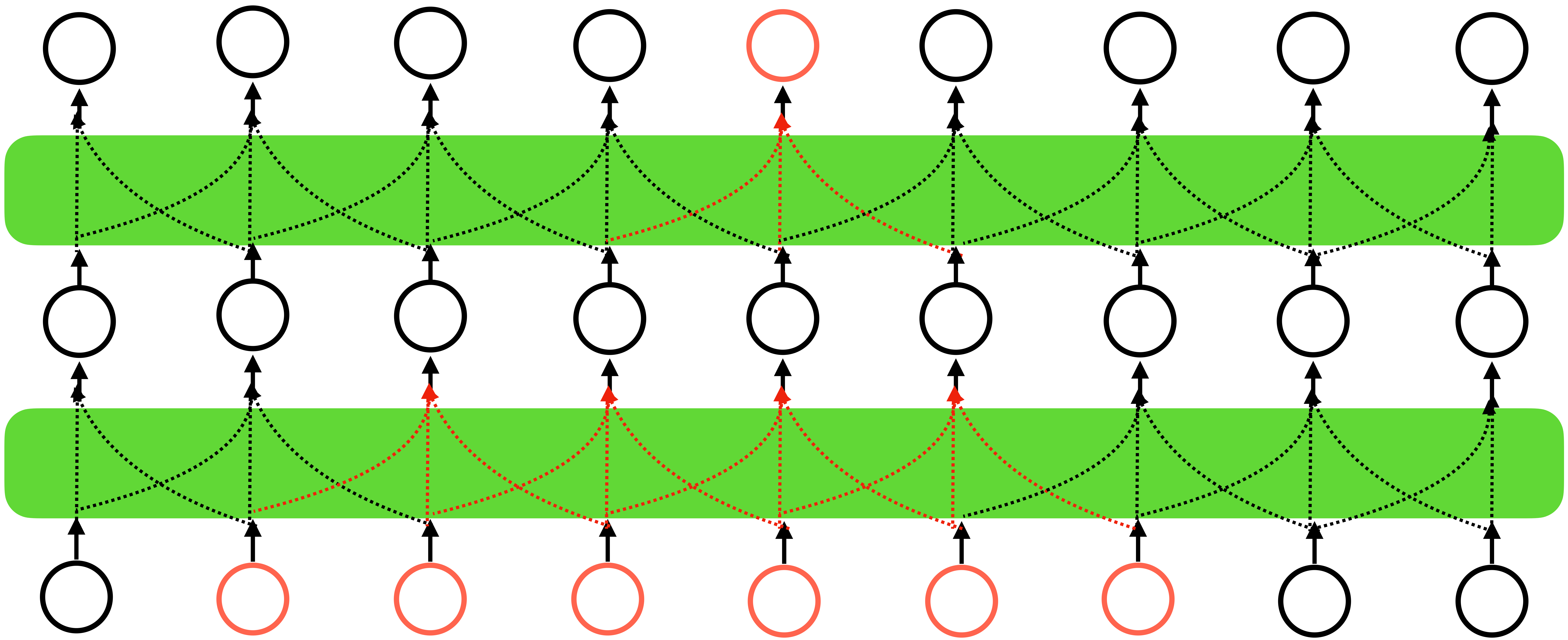
Parallel Computing?



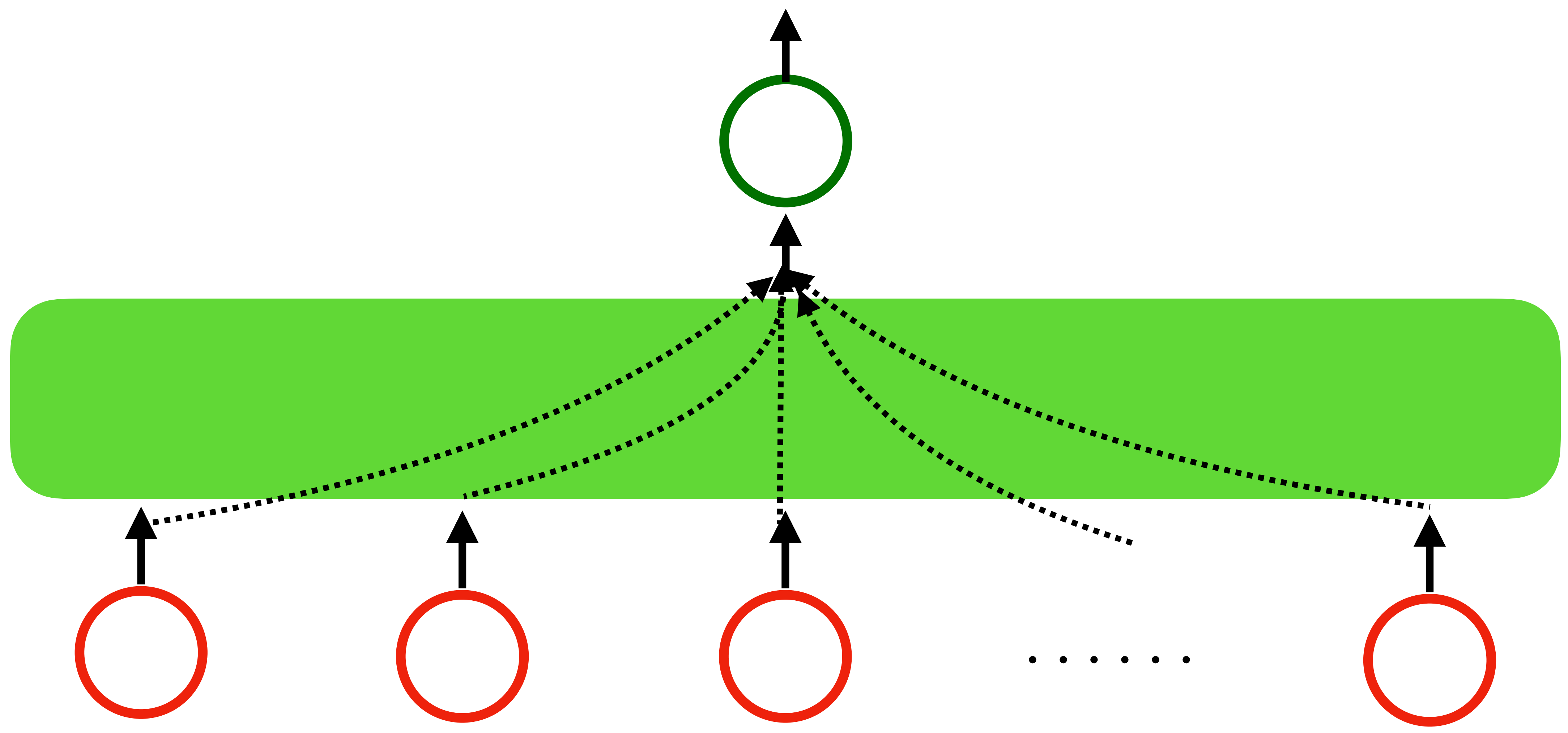
Convolution Style Models



Convolution Style Models

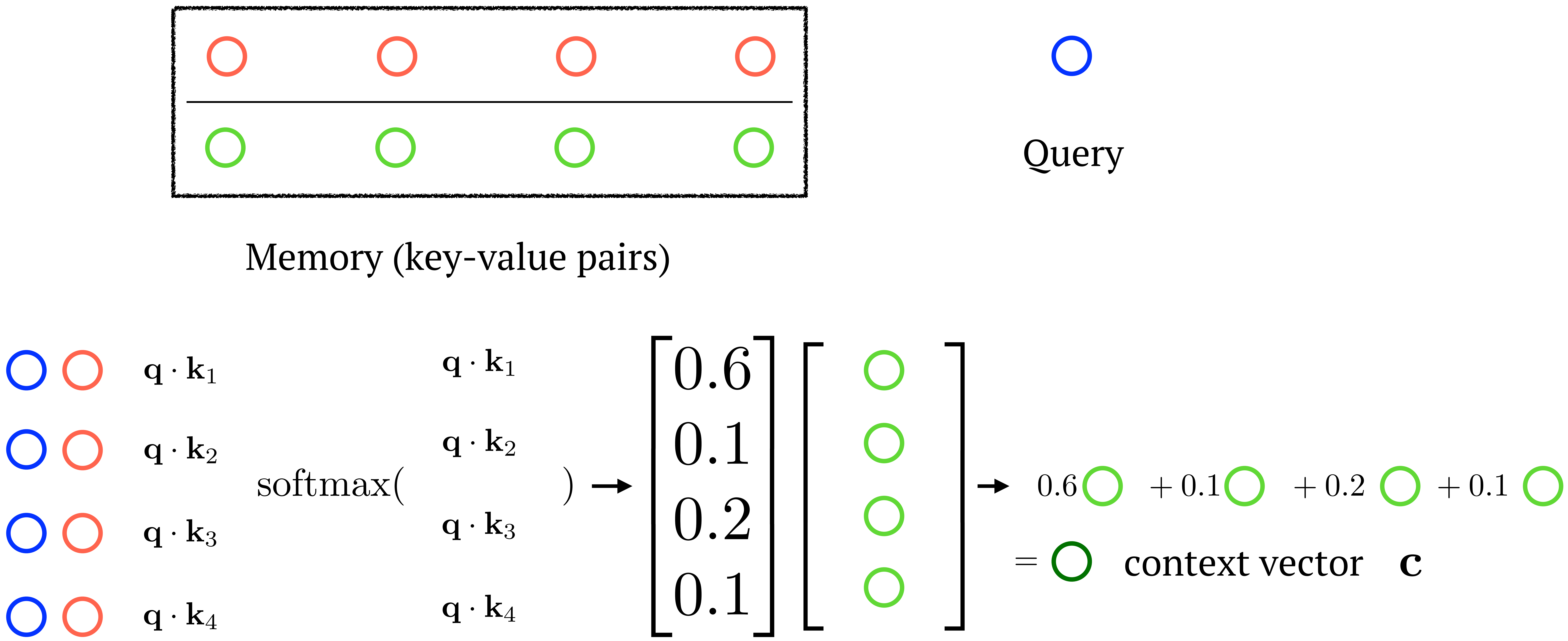


Considering the full sequence as context

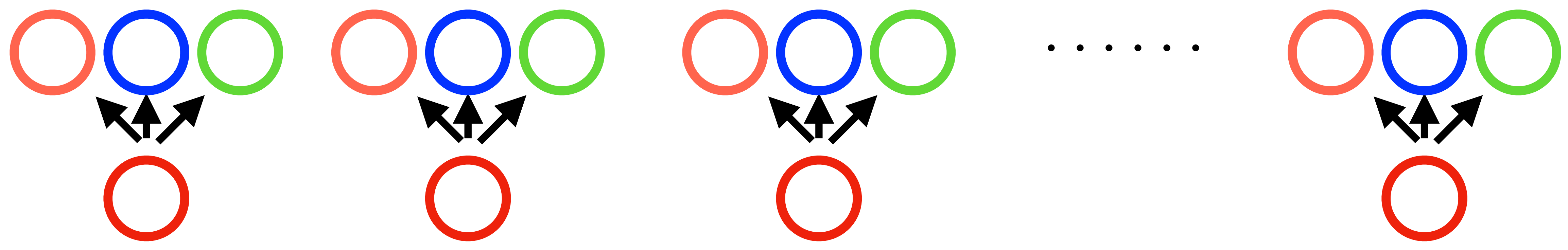


How can we achieve this?

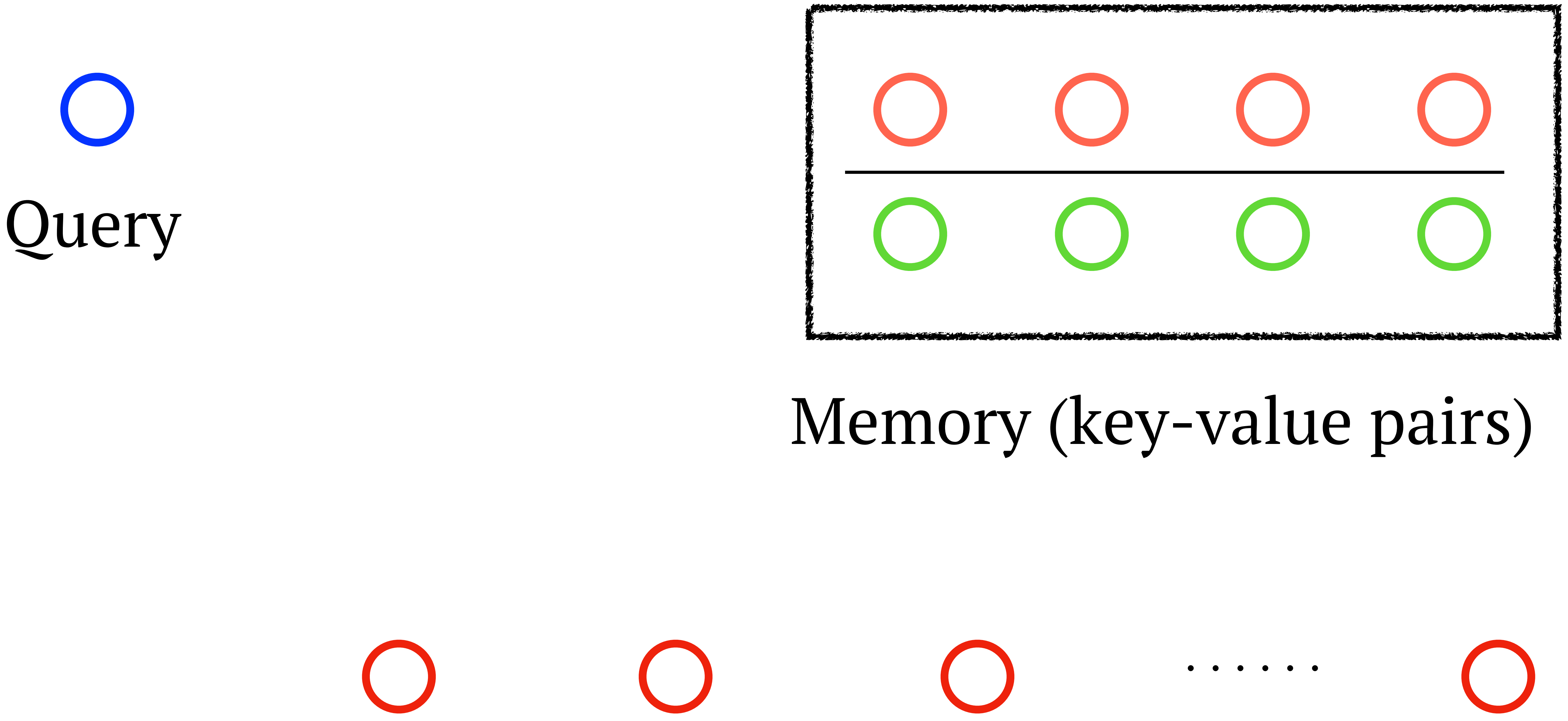
Dot-Product-Softmax Attention



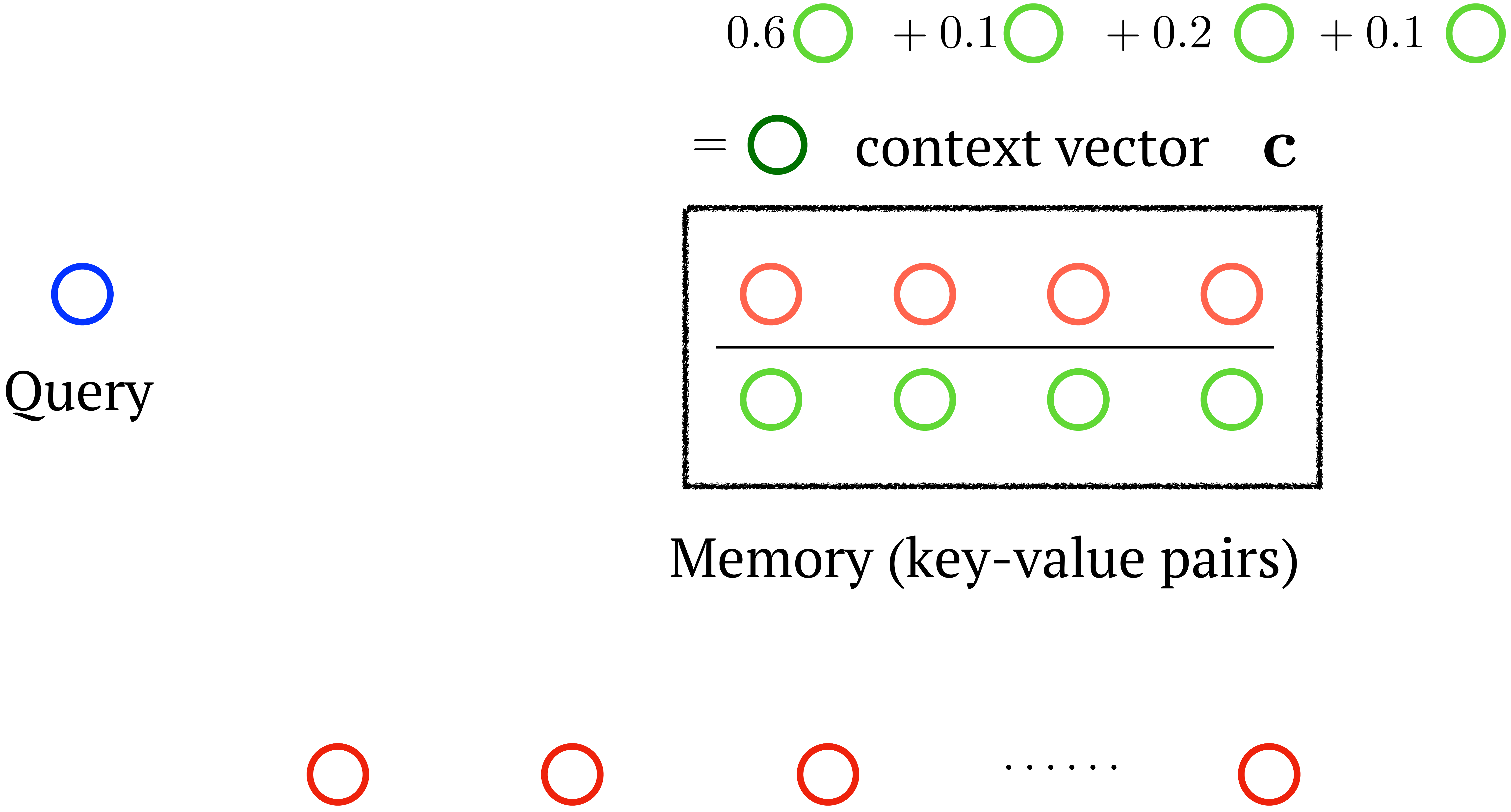
Considering the full sequence as context



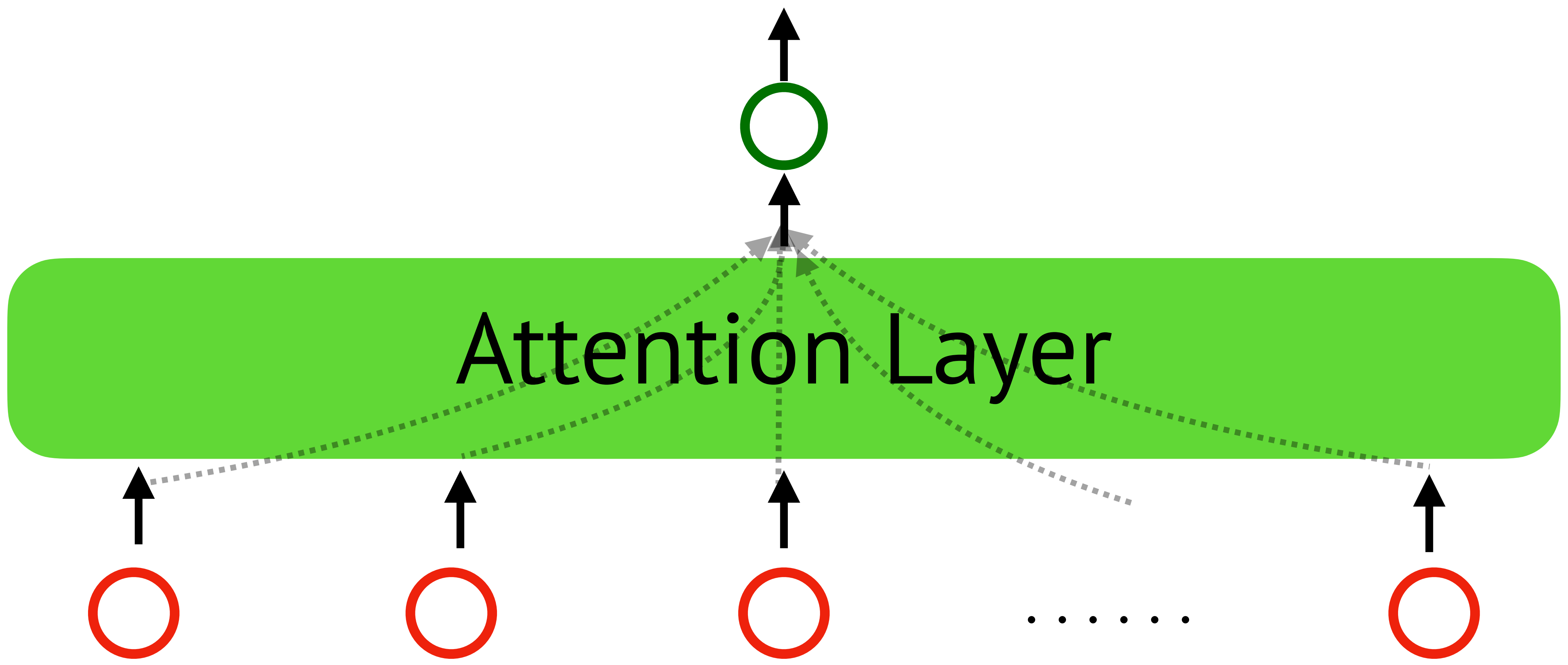
Attention Mechanism



Attention Mechanism

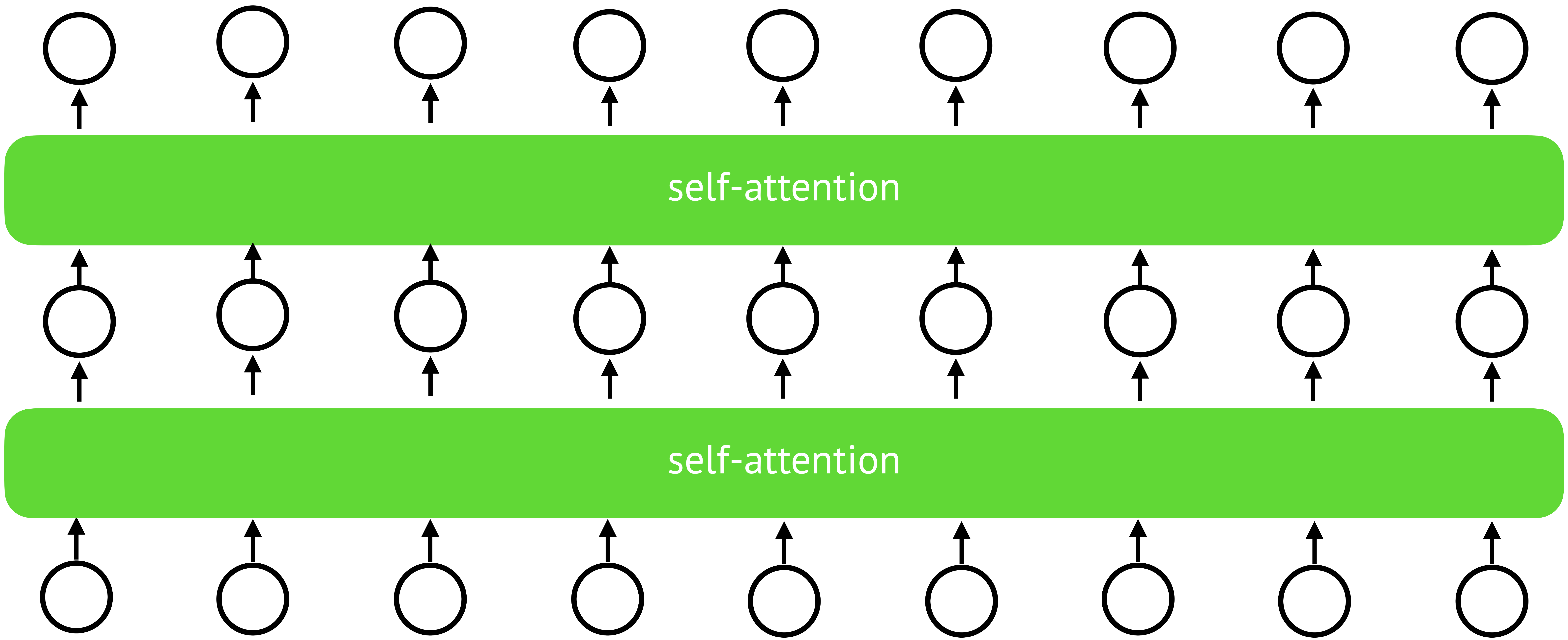


Self-attention

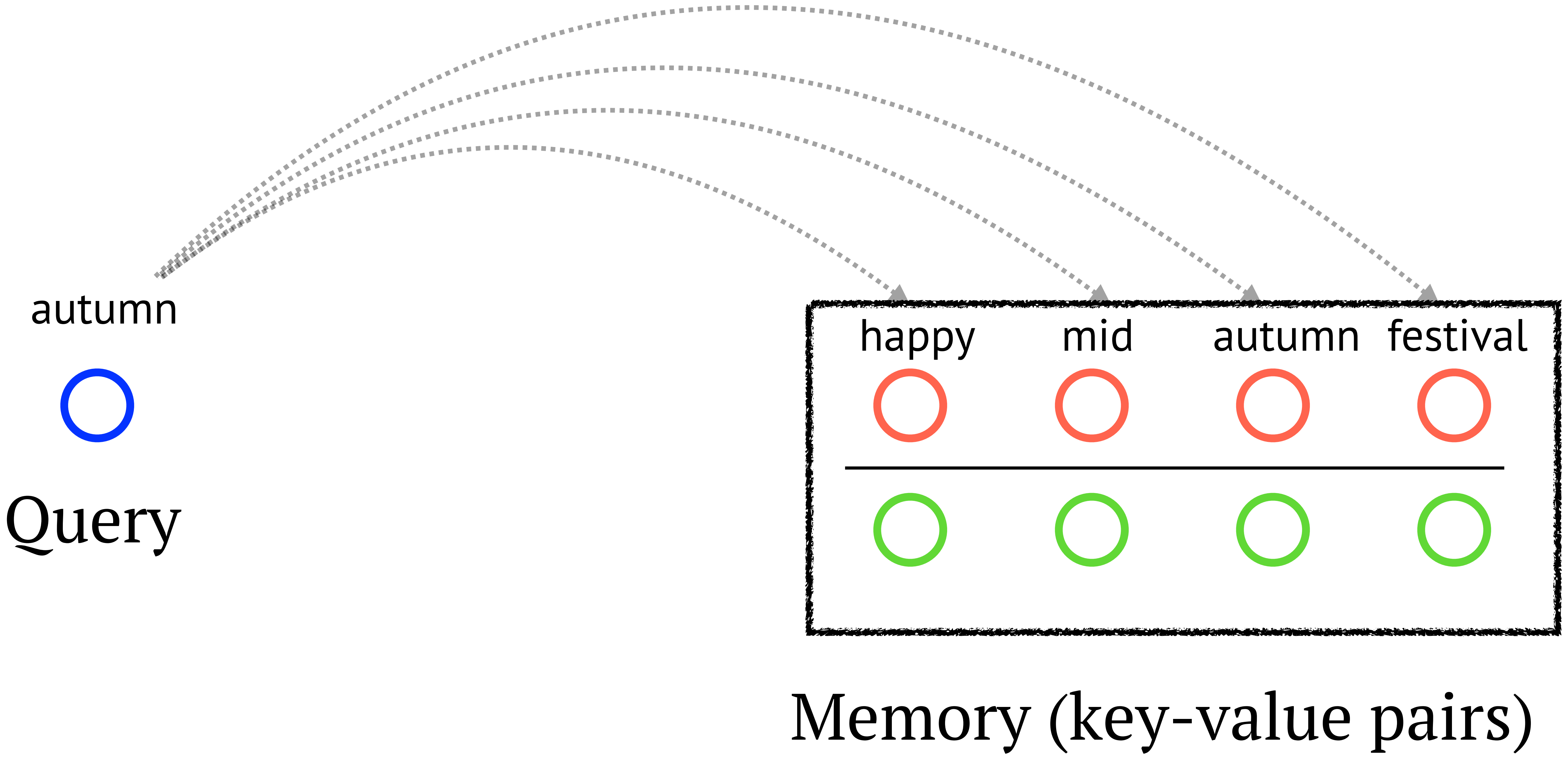


This is almost transformer — except a few things.

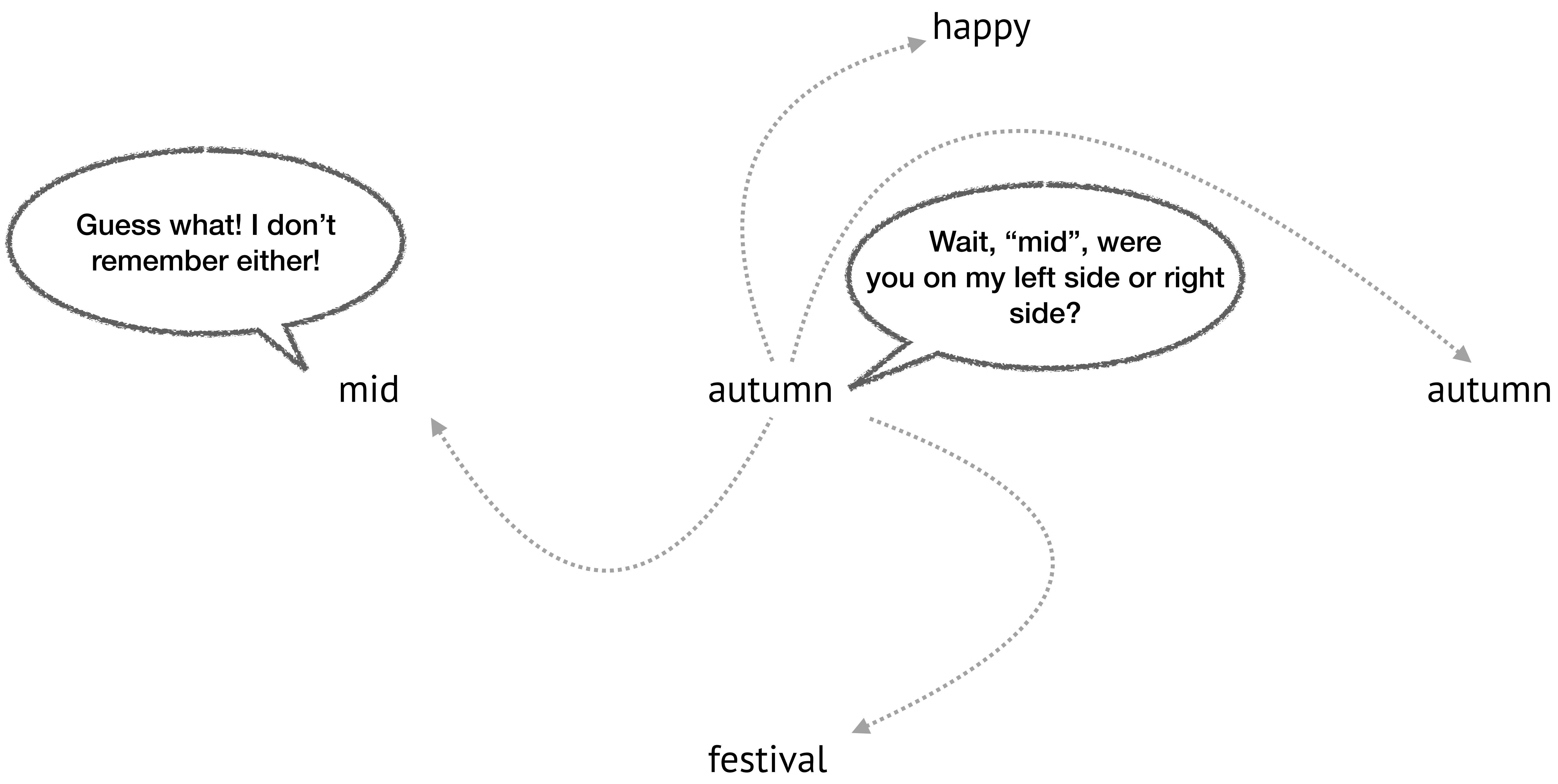
Transformer (almost)



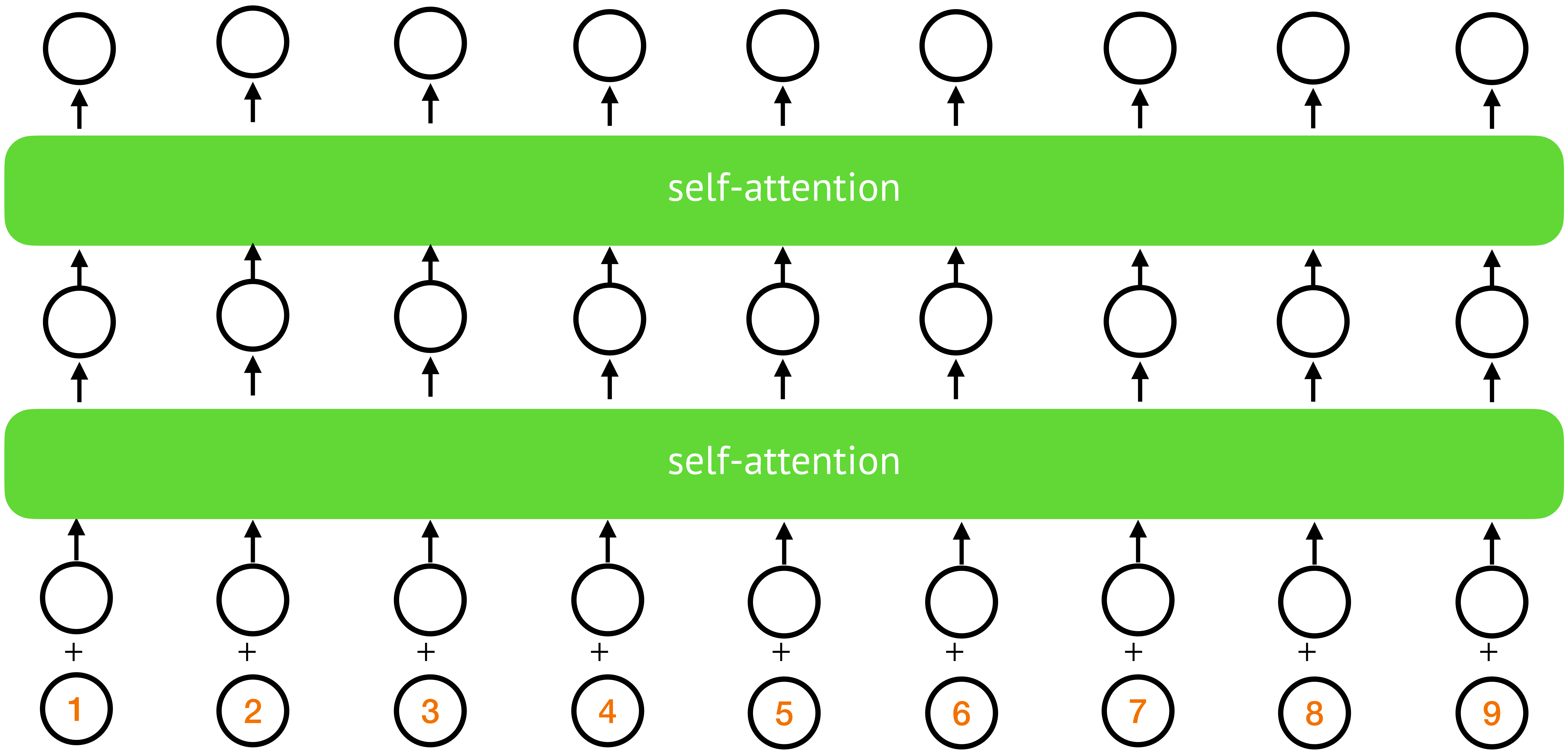
Self-attention in Transformer



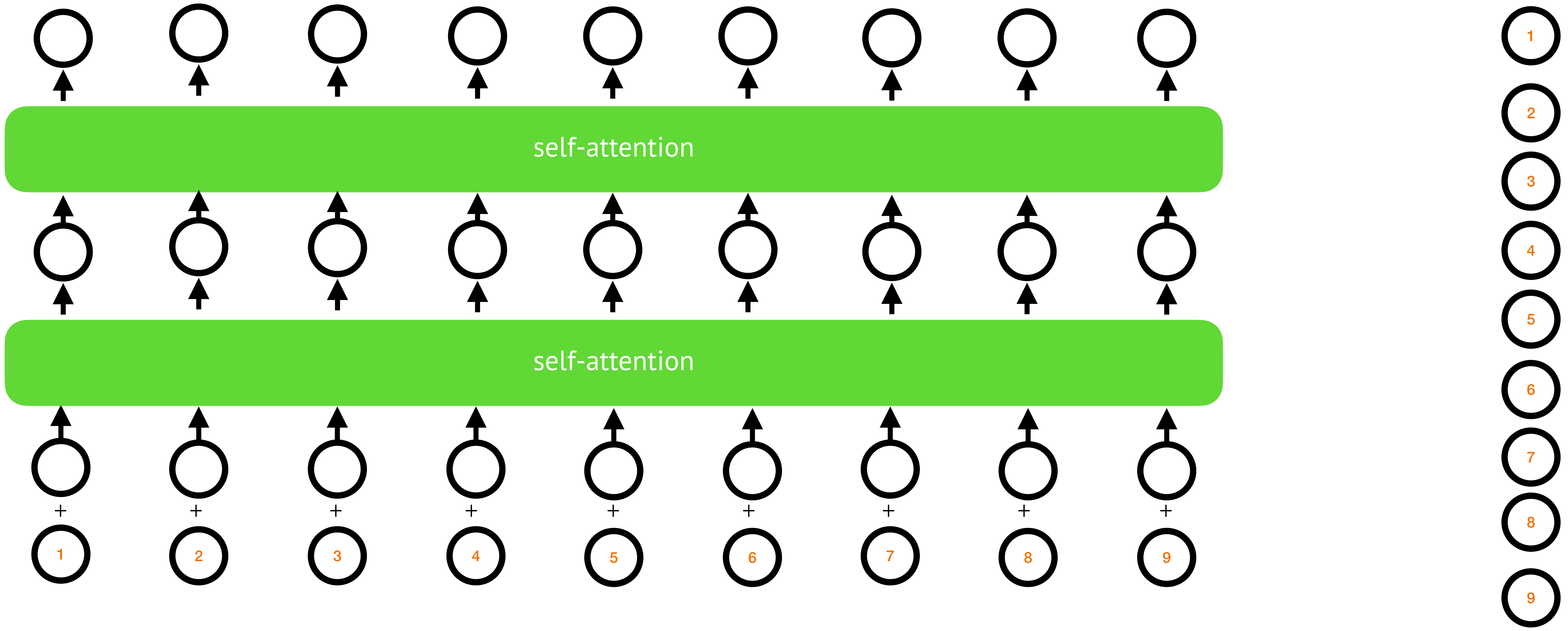
Self-attention in Transformer



Positional Embeddings

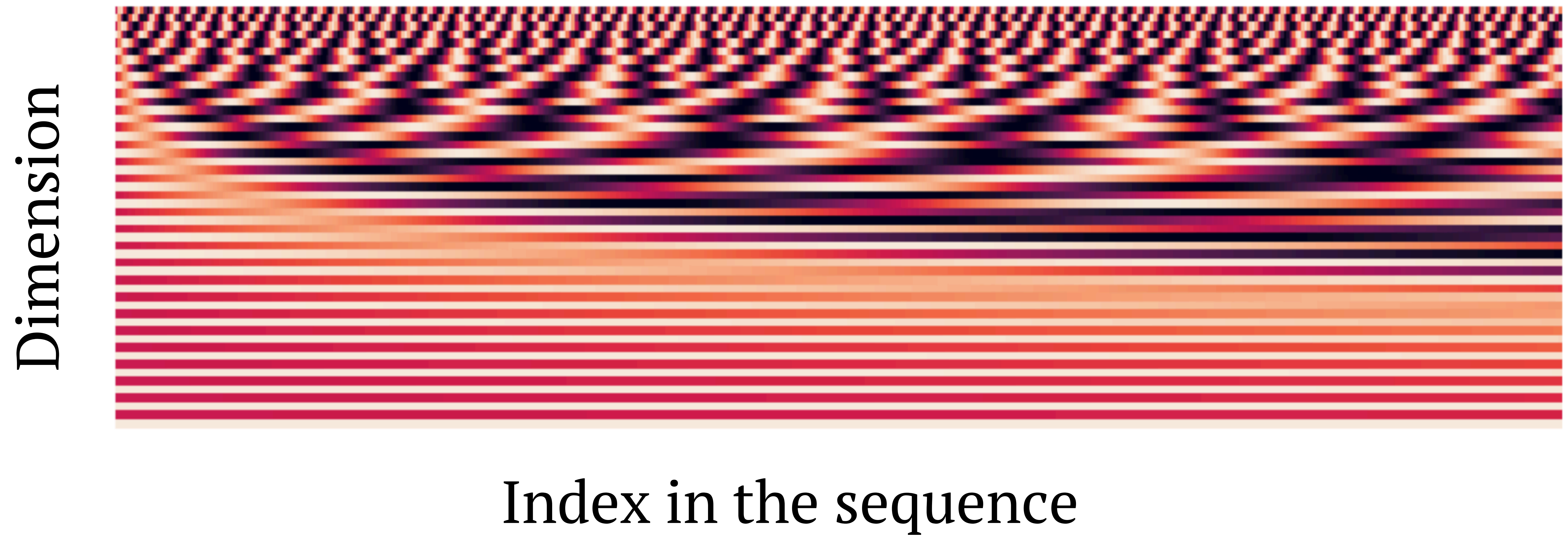


Transformer (positional embedding)



Positional Encoding

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

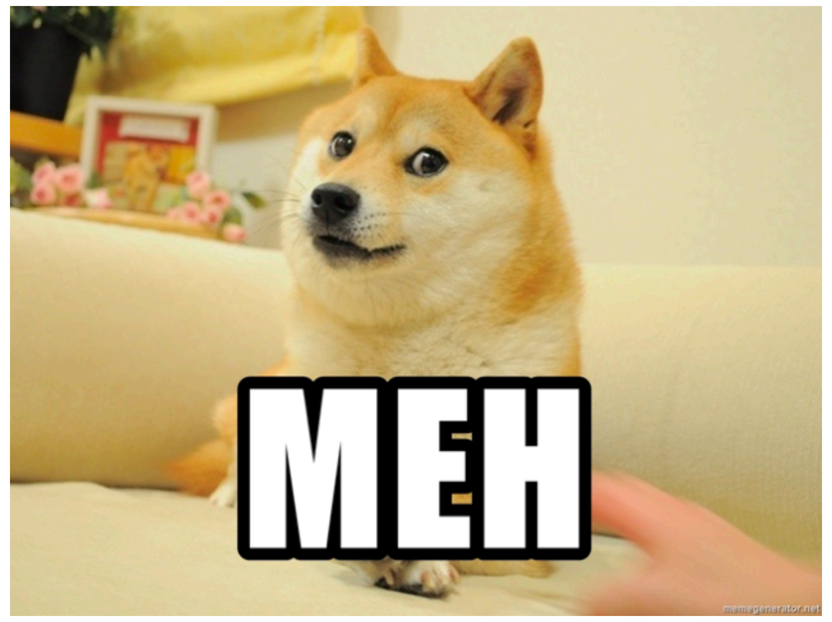


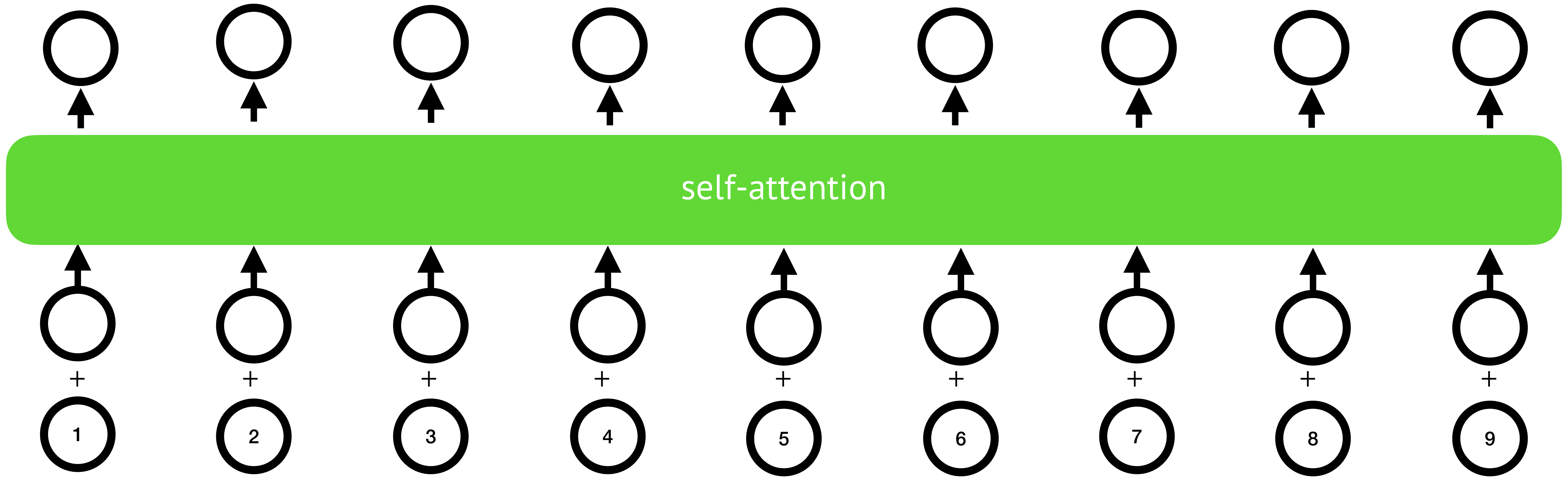
The idea of relative position

Positional Encoding

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

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

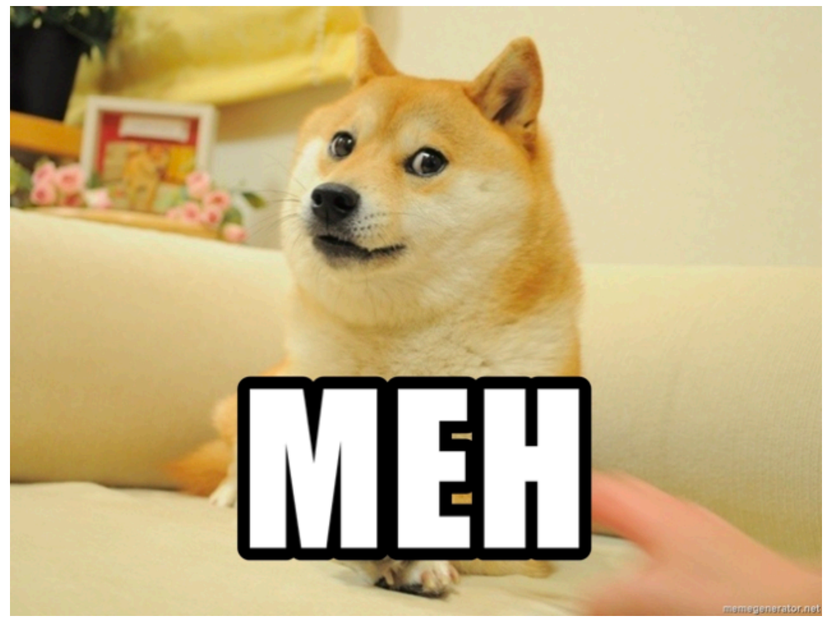


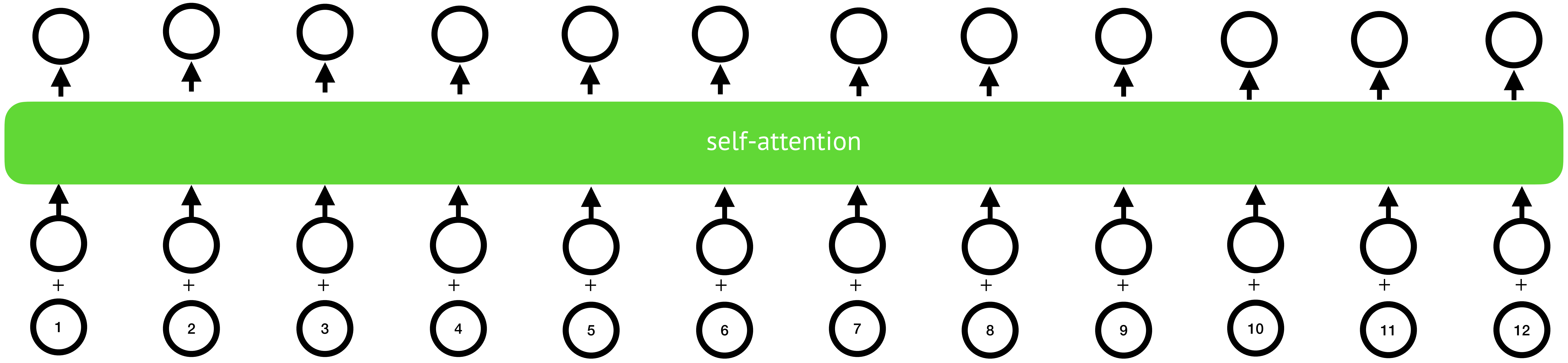


Positional Encoding

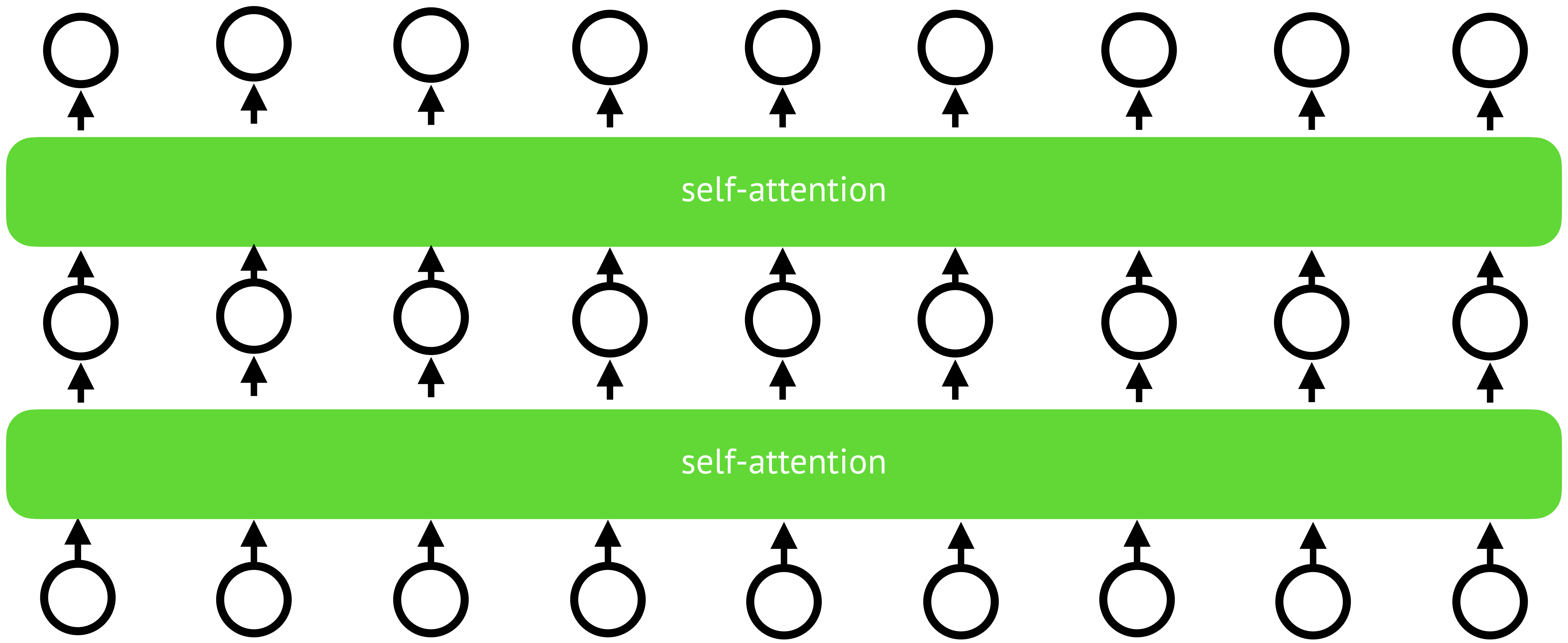
$$\begin{bmatrix} \sin\left(\frac{i}{10000^{2\times\frac{1}{d}}}\right) \\ \cos\left(\frac{i}{10000^{2\times\frac{1}{d}}}\right) \\ \vdots \\ \sin\left(\frac{i}{10000^{2\times\frac{d/2}{d}}}\right) \\ \cos\left(\frac{i}{10000^{2\times\frac{d/2}{d}}}\right) \end{bmatrix}$$

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

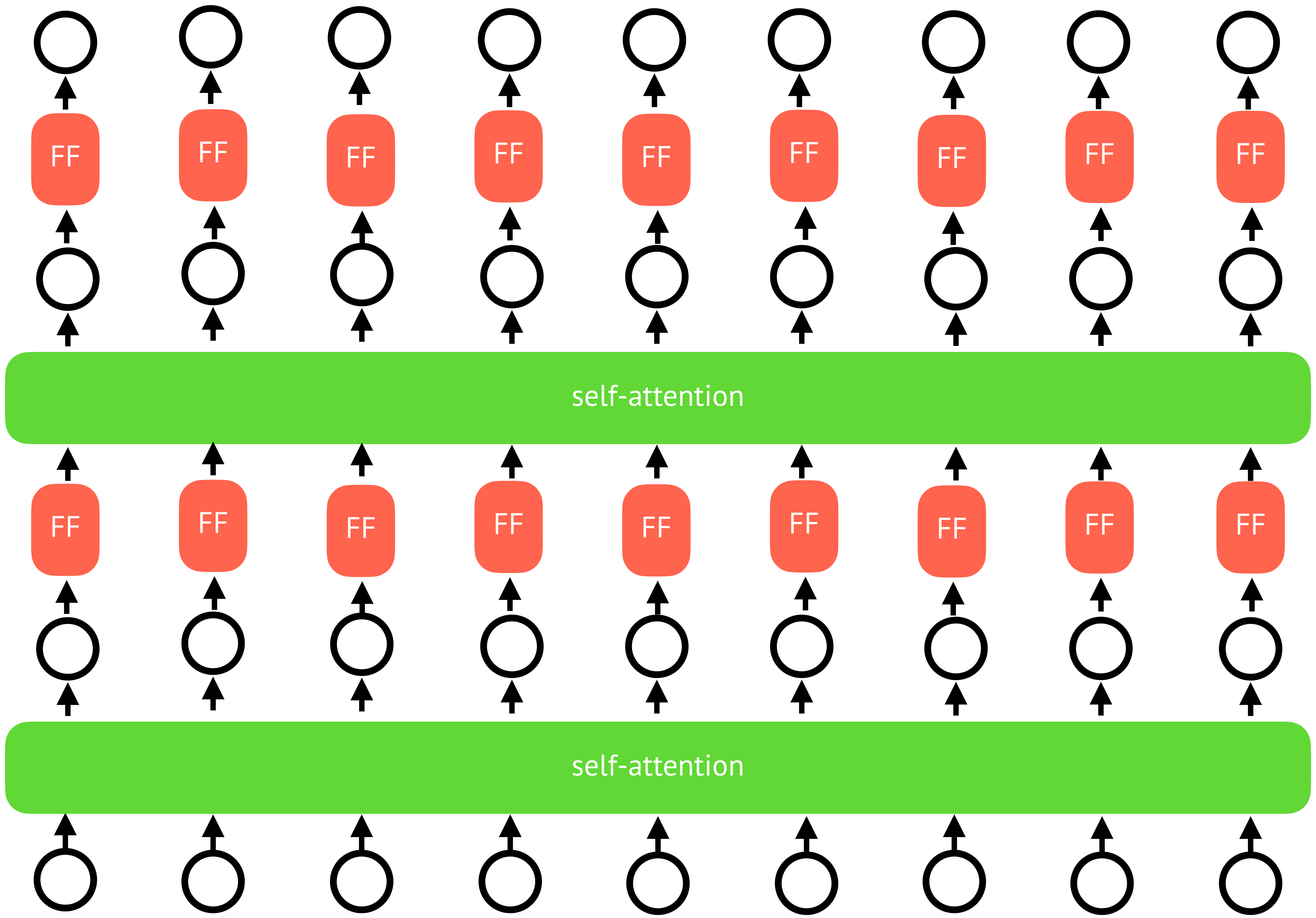




Feed Forward Layer



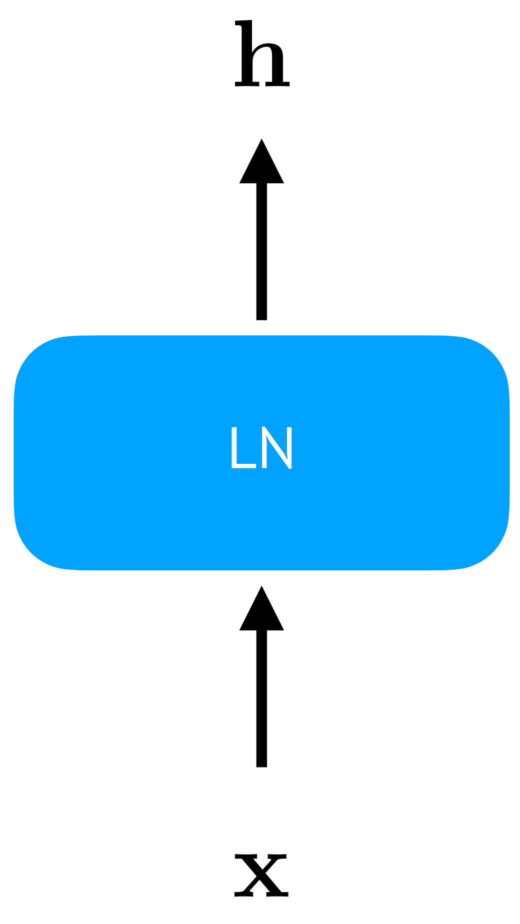
Feed Forward Layer



Layer Normalization (Ba et al, 2016)

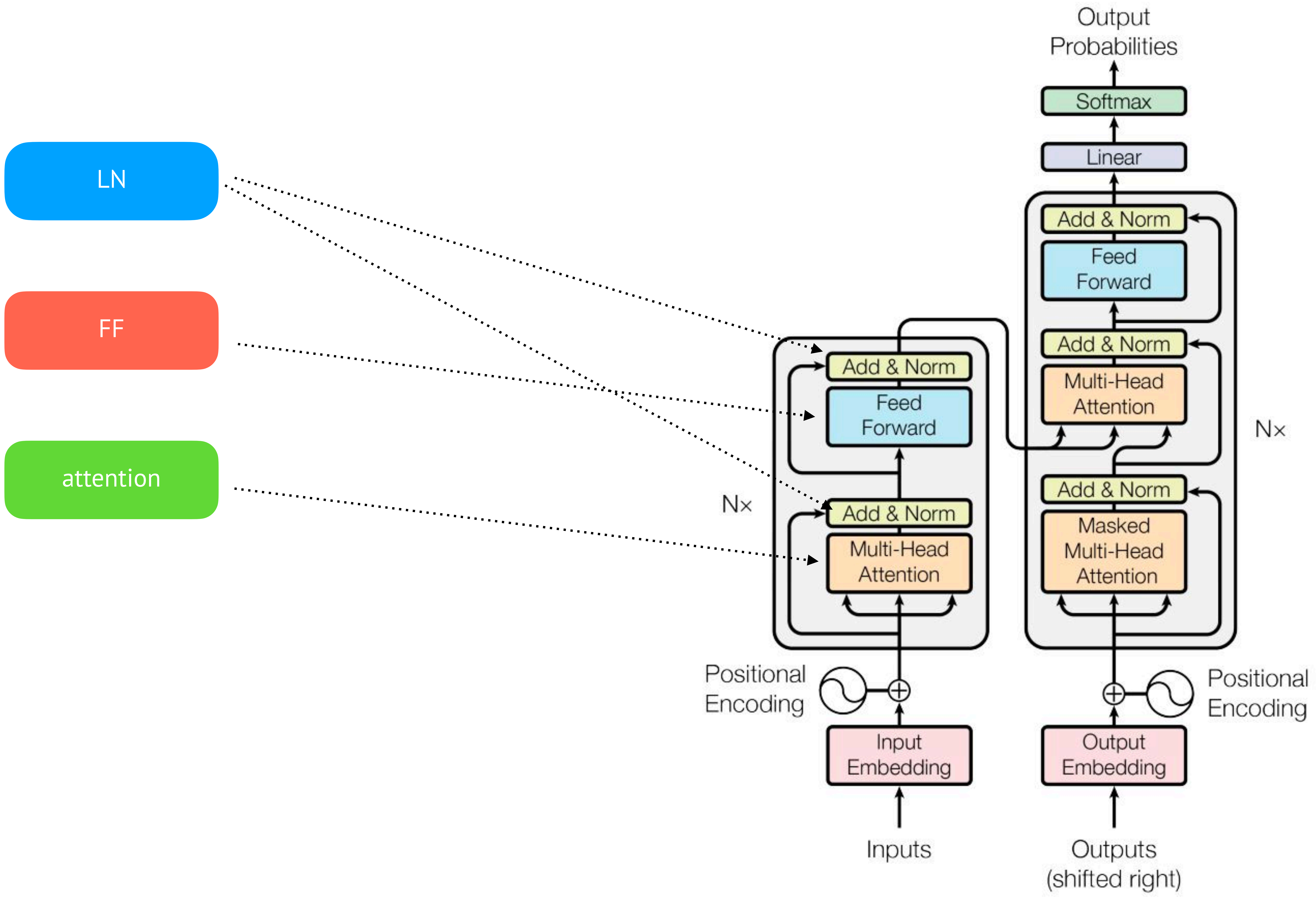
$$\mathbf{h} = \mathbf{g} \odot N(\mathbf{x}) + \mathbf{b}$$

$$N(\mathbf{x}) = \frac{\mathbf{x} - \mu}{\sigma} \qquad \mu = \frac{1}{H}\sum_{i=1}^{H} x_i \qquad \sigma = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(x_i - \mu)^2}$$

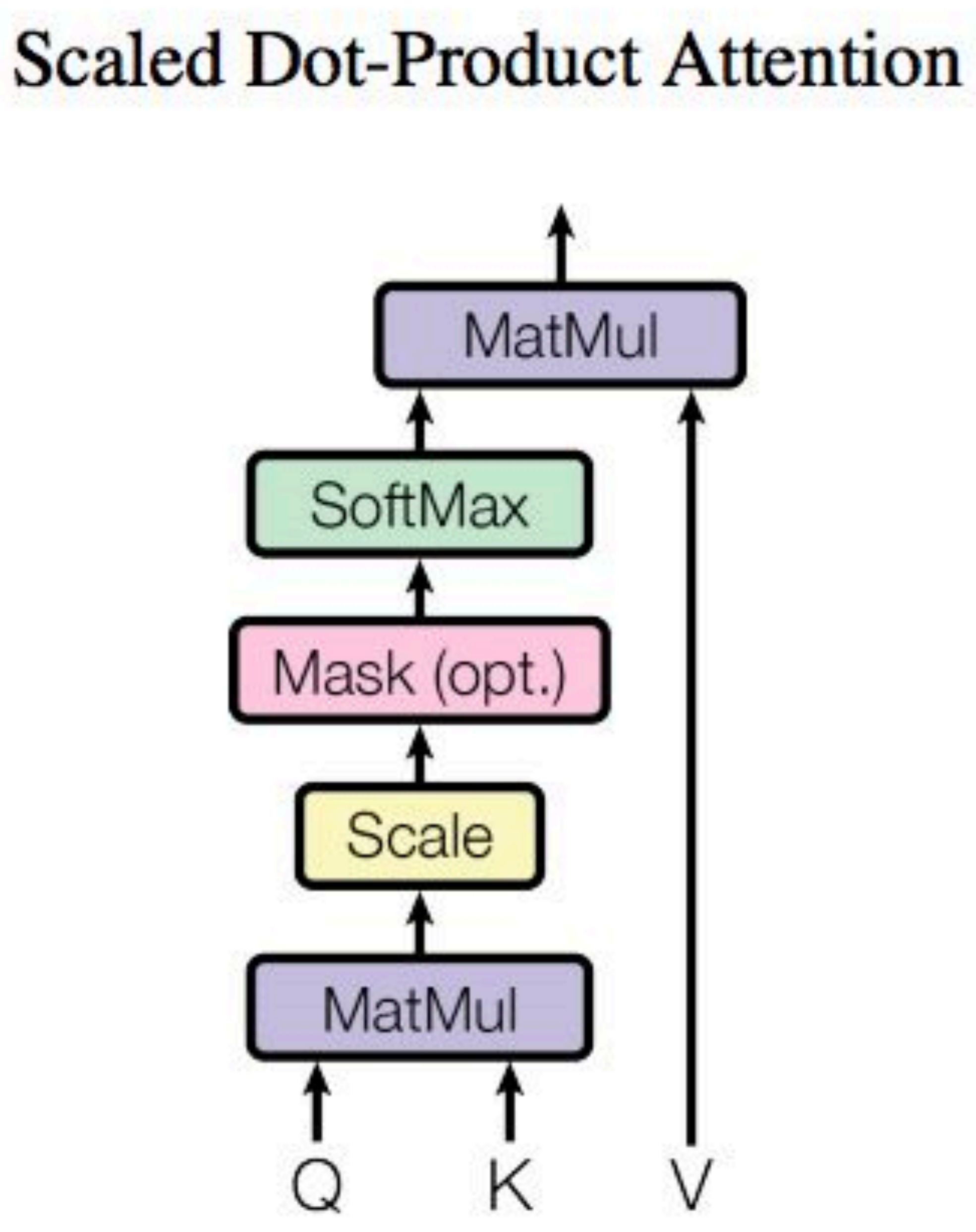


Smoother gradients, faster training and better generalization accuracy. (Xu et al, Neurips 2019)

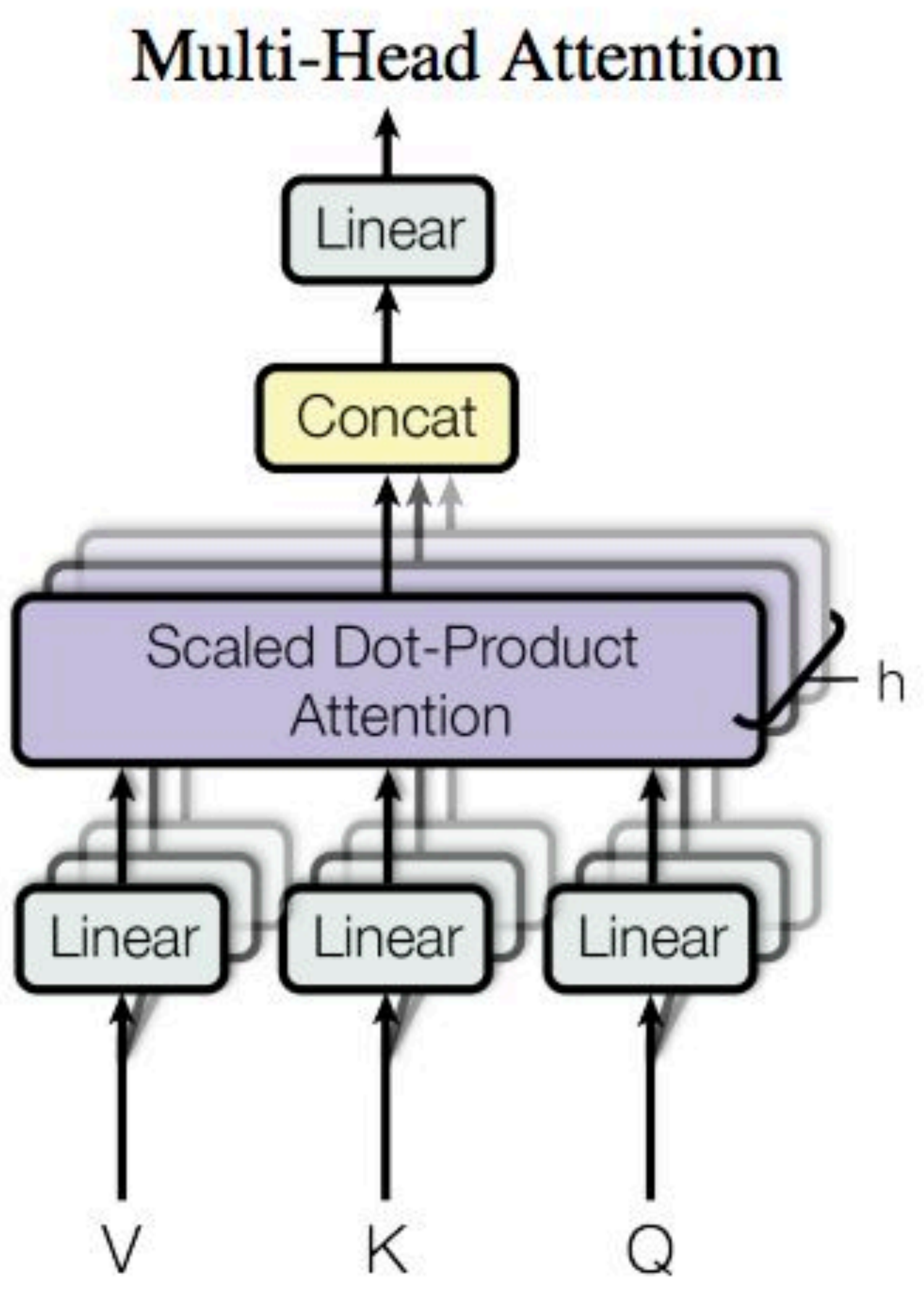
Layer Normalization



Multi-head Attention

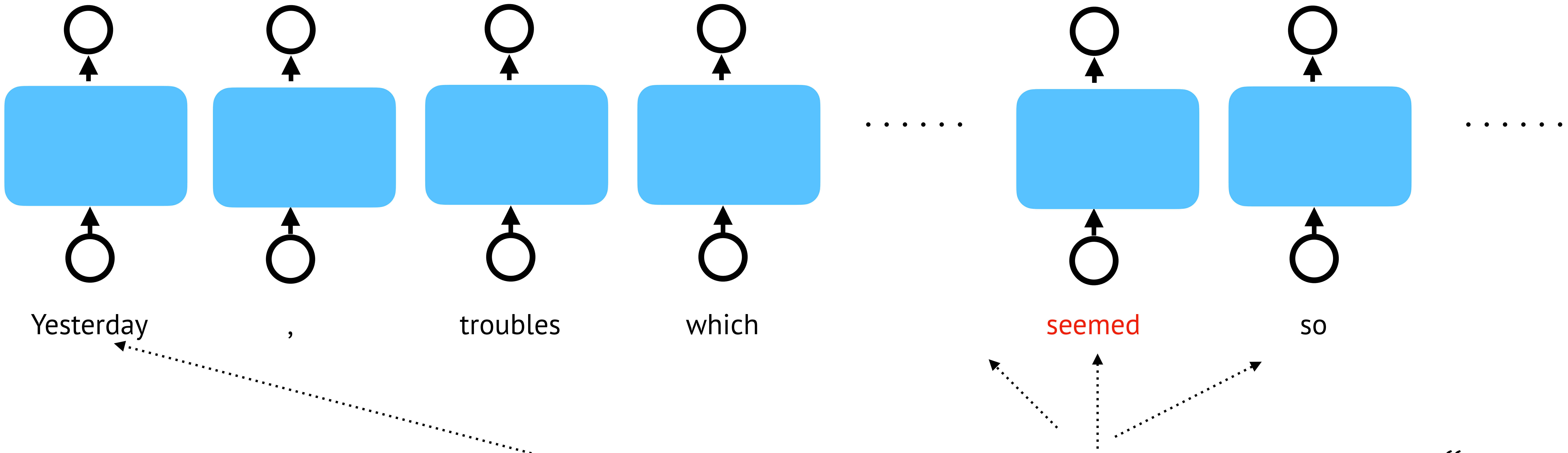


$$\mathrm{score}(q, k) = \frac{q^T k}{\sqrt{d_k}}$$



multiple copies

Multi-head Attention



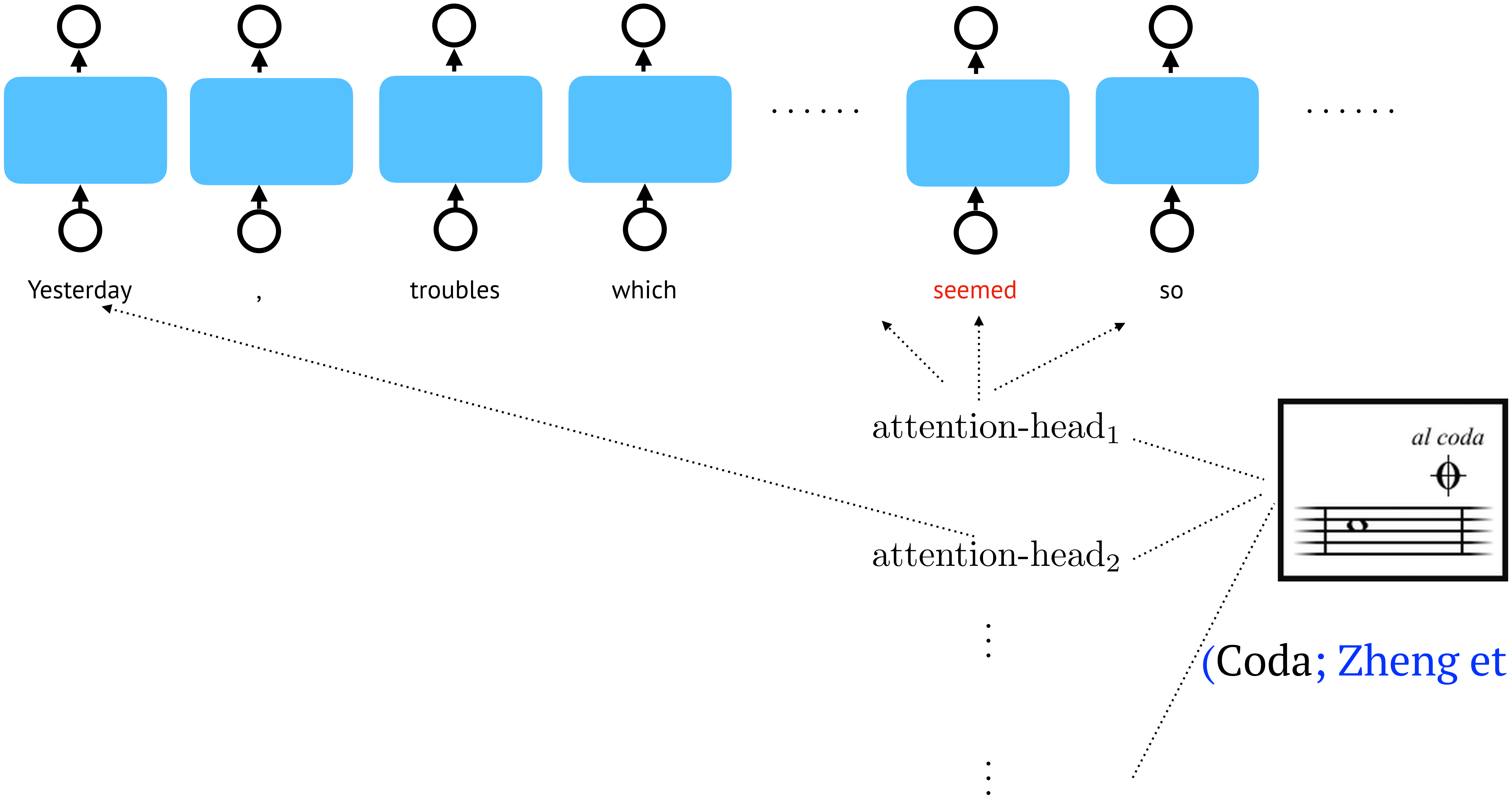
attention-head_1 more "local" context

attention-head_2 more long dependencies

$\vdots$

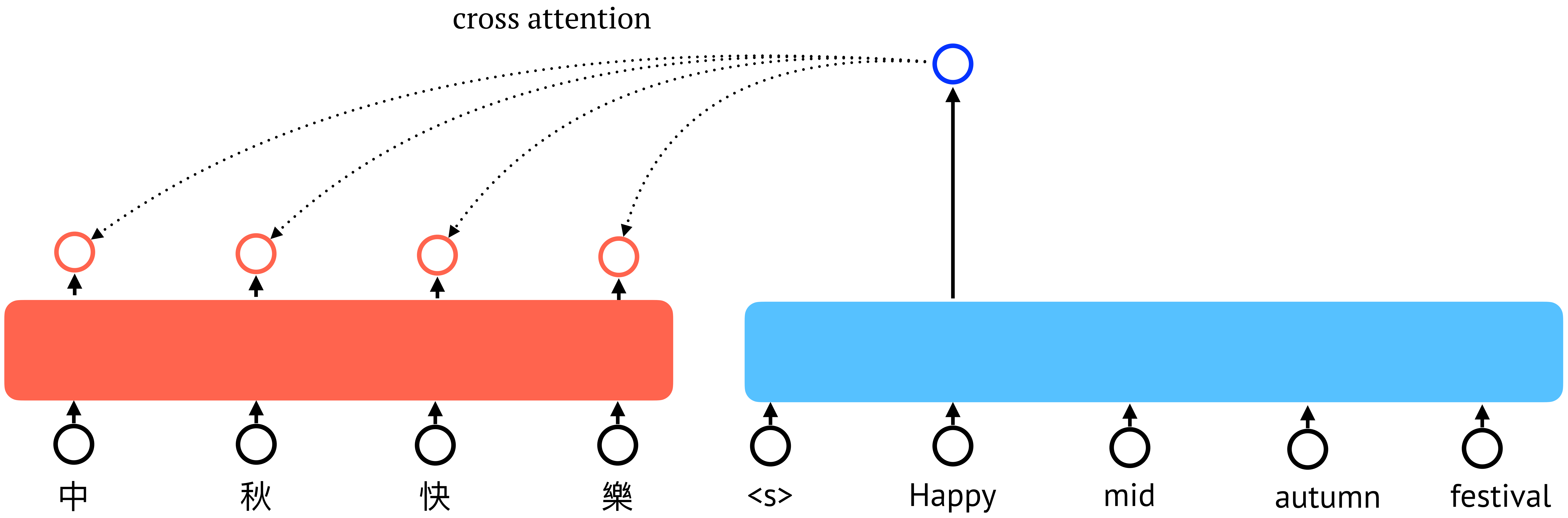
Improve the "resolution" of the attention mechanism.

Multi-head Attention

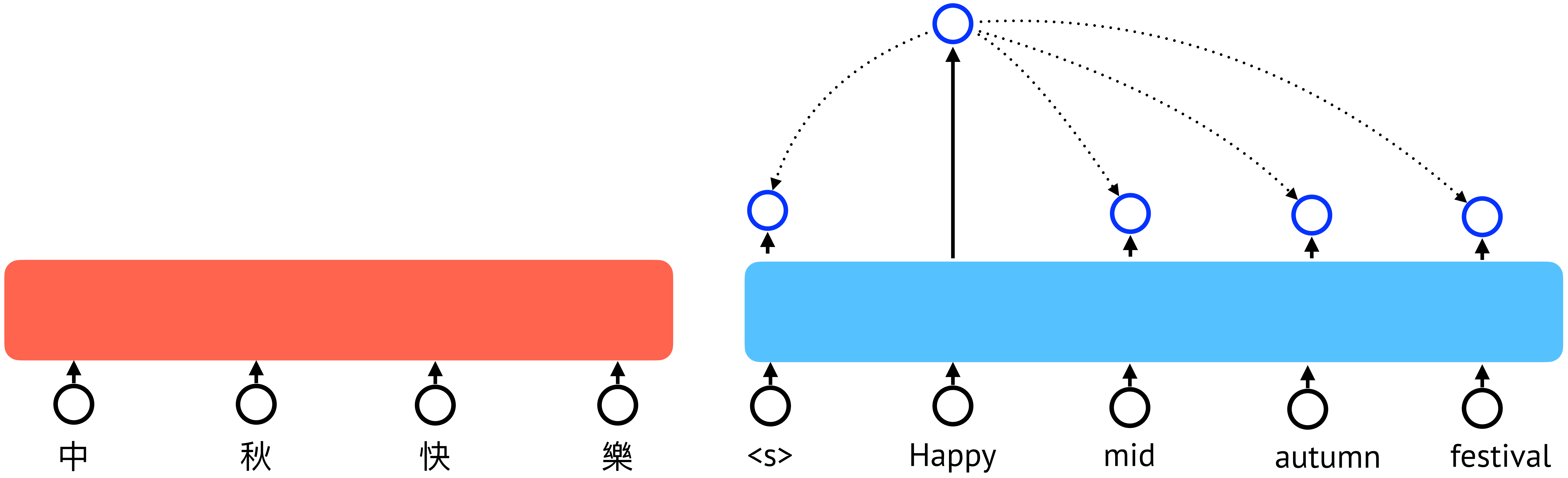


(Coda; Zheng et al, ACL 2021)

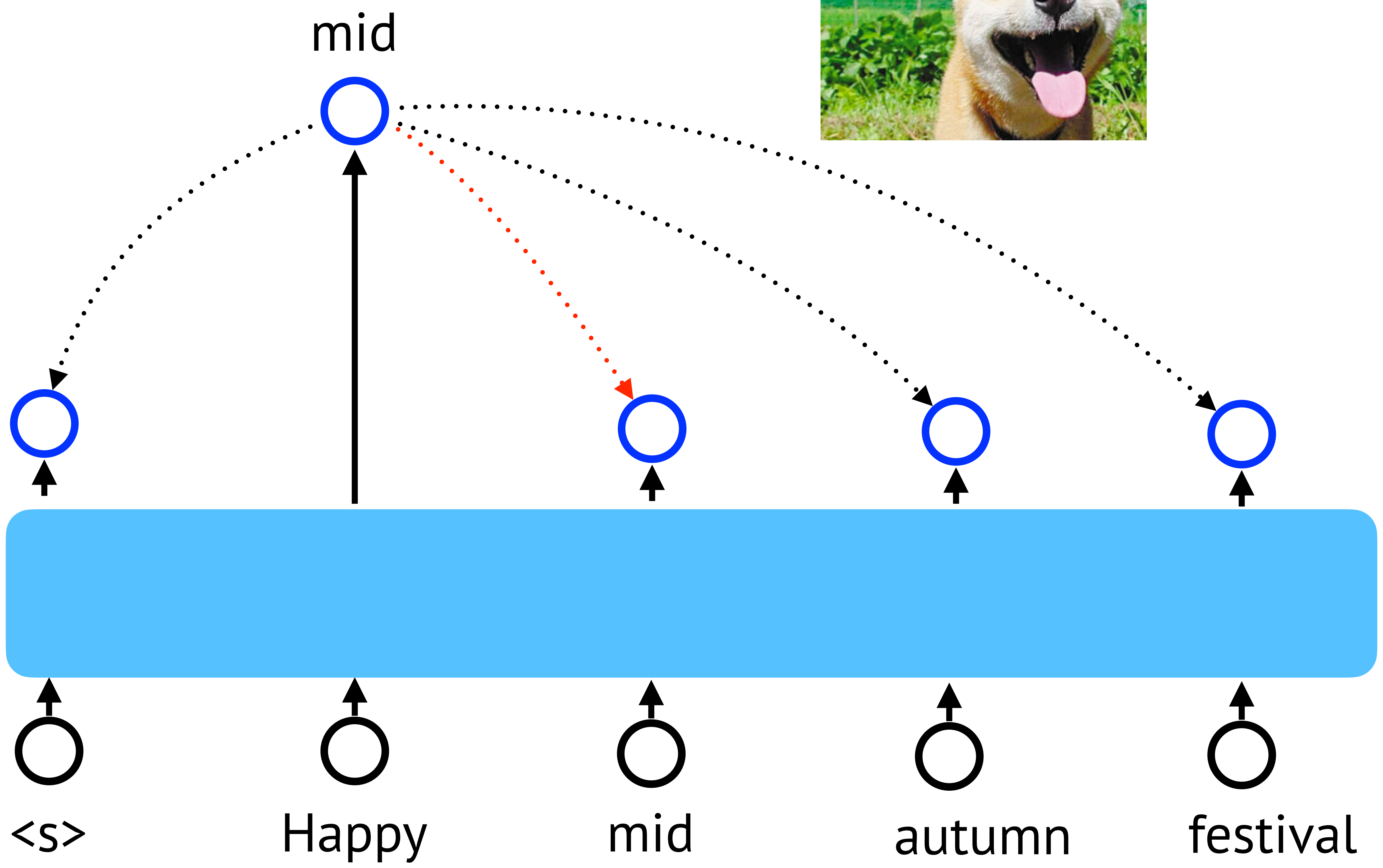
Transformer as Decoder

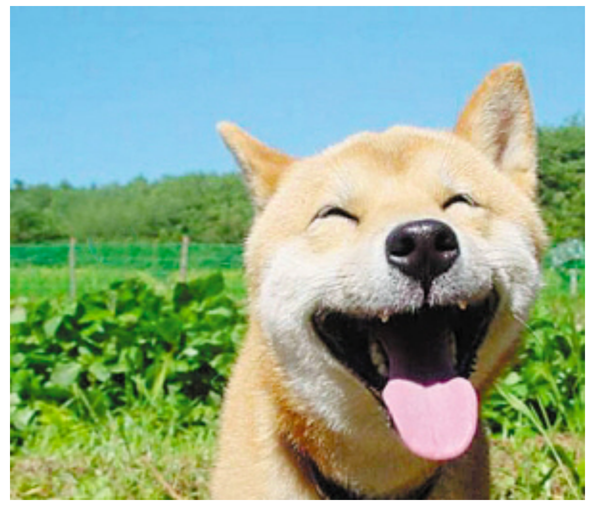


Transformer as Decoder



Transformer as Decoder



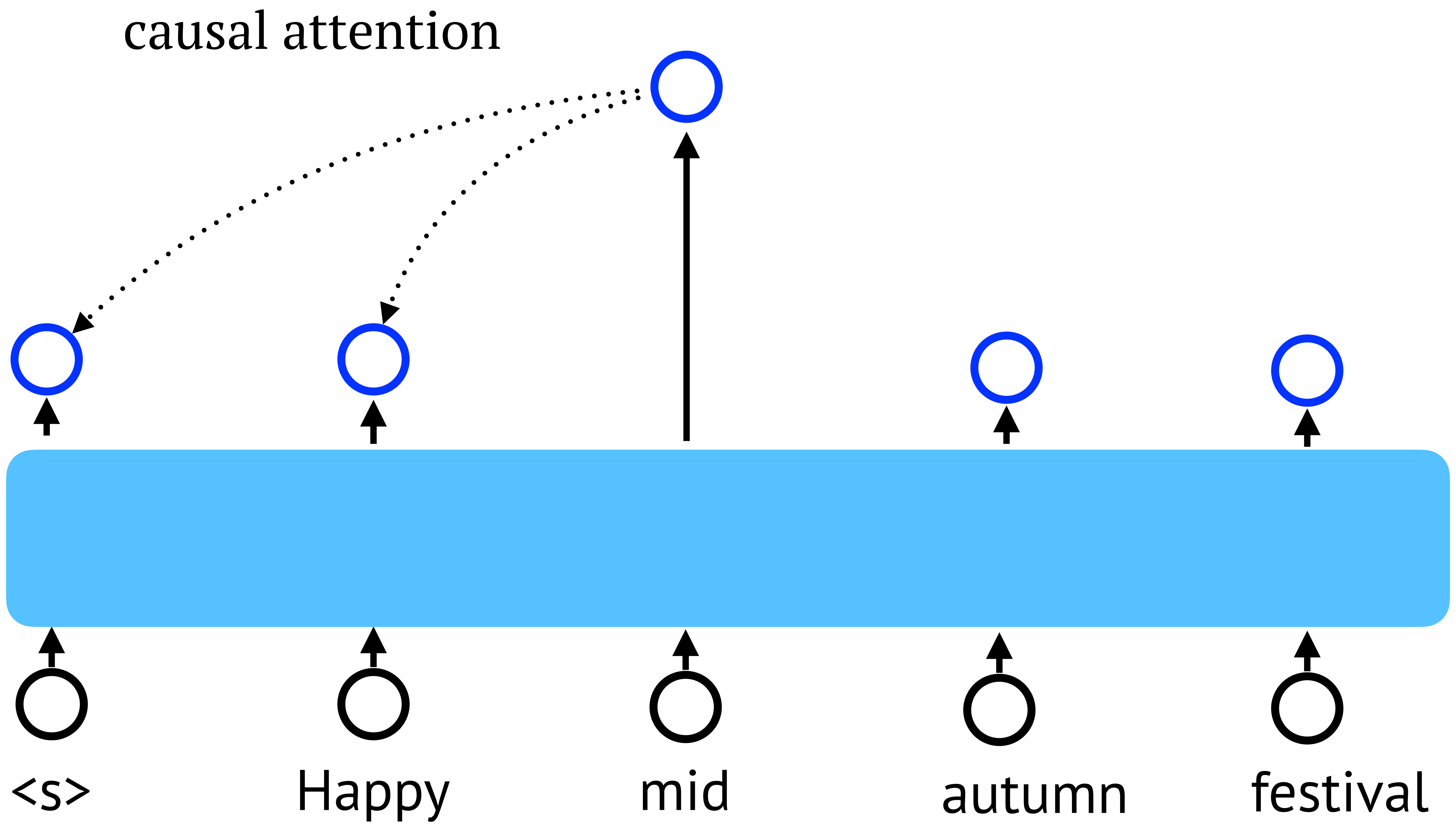


Need to prevent the attention the future words.

	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

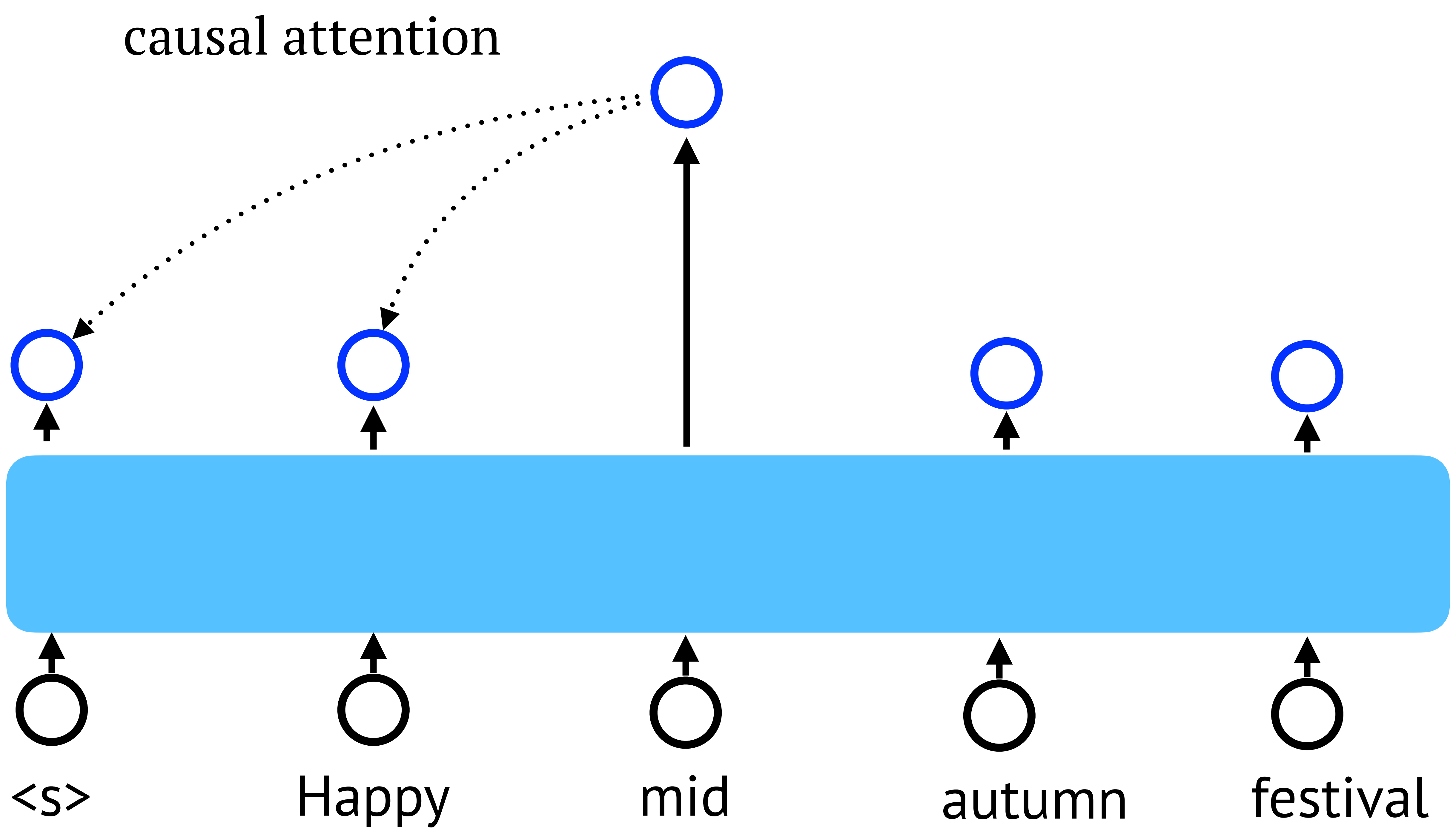


Need to prevent the attention the future words.

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

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