

Transformers, LLMs, pretraining, instruction-tuning

Reasoning in AI, week 3

Brief overview of LLM training

Next Token Prediction

Note: All equations in this course at best true up to transposes of matrices or permuting axes of tensors

Input: ~~Tokens $x_1, x_2, \dots, x_n \in T$ (e.g. $T = \{1, \dots, 128,000\}$)~~

Vectors $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n \in \mathbb{R}^d$

Output: ~~Token x_{n+1}~~ Vector $\vec{x}_{n+1} \in \mathbb{R}^d$

$$t \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

t -th standard basis vector

$d \times |T|$ embedding matrix E – embedding of token t is $E e_t$

$|T| \times d$ unembedding matrix U If $U \vec{y} = e_t$ then \vec{y} is prediction that $\Pr[x_{n+1} = t] = 1$

A vector $\vec{y} \in \mathbb{R}^d$ defines distribution $D_{\vec{y}}$ over T with $\Pr[D_{\vec{y}} = t] \propto \exp(\tau^{-1}(U \vec{y})_t)$

τ = temperature

$\tau = \infty$: distribution is uniform

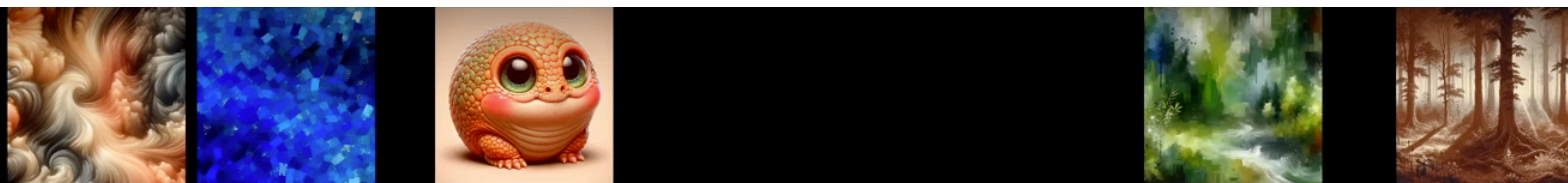
$\tau = 0$: all weight on $\arg \max_t (U \vec{y})_t$

Architecture for Next Token Prediction: Map from \mathbb{R}^{dn} to \mathbb{R}^d

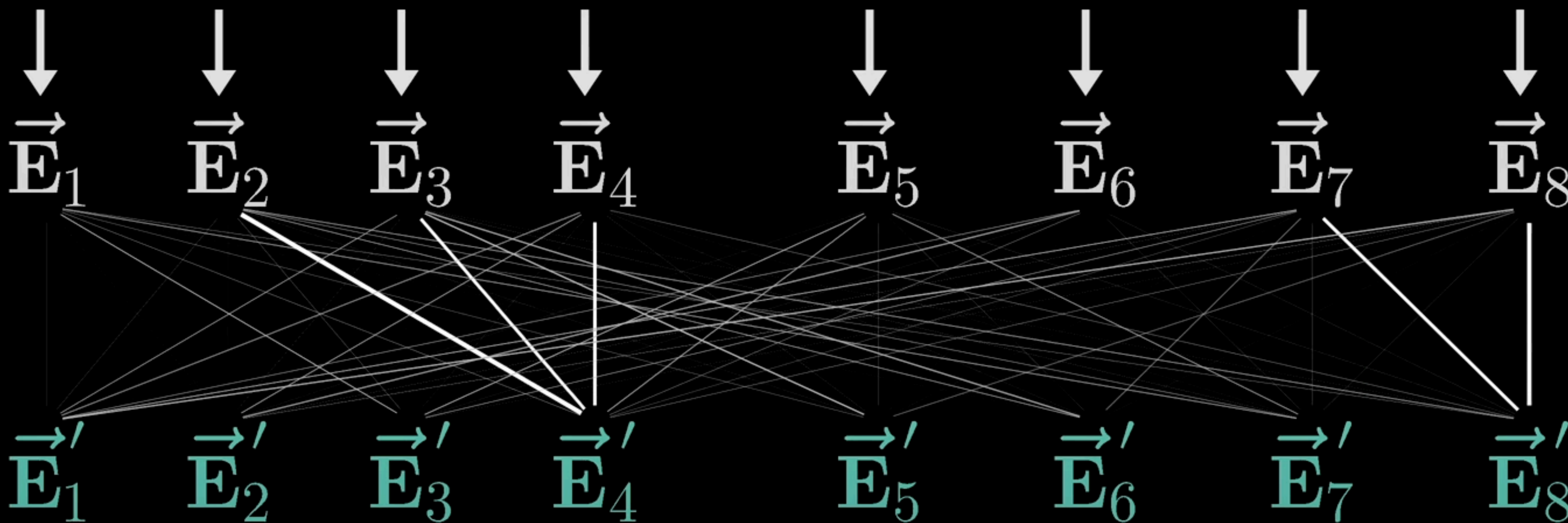
History

- Previous models for text generation: RNN, LSTM, GRU
- Common issues:
 - - Temporal Bottleneck: Vanishing gradients stop many RNN architectures from learning **long-range dependencies**
 - - Parallelisation Bottleneck: RNN states depend on previous time step hidden state, so must be **computed in series**

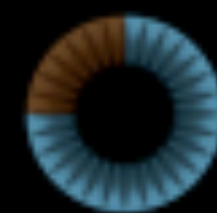
Transformer



a fluffy blue creature roamed the verdant forest

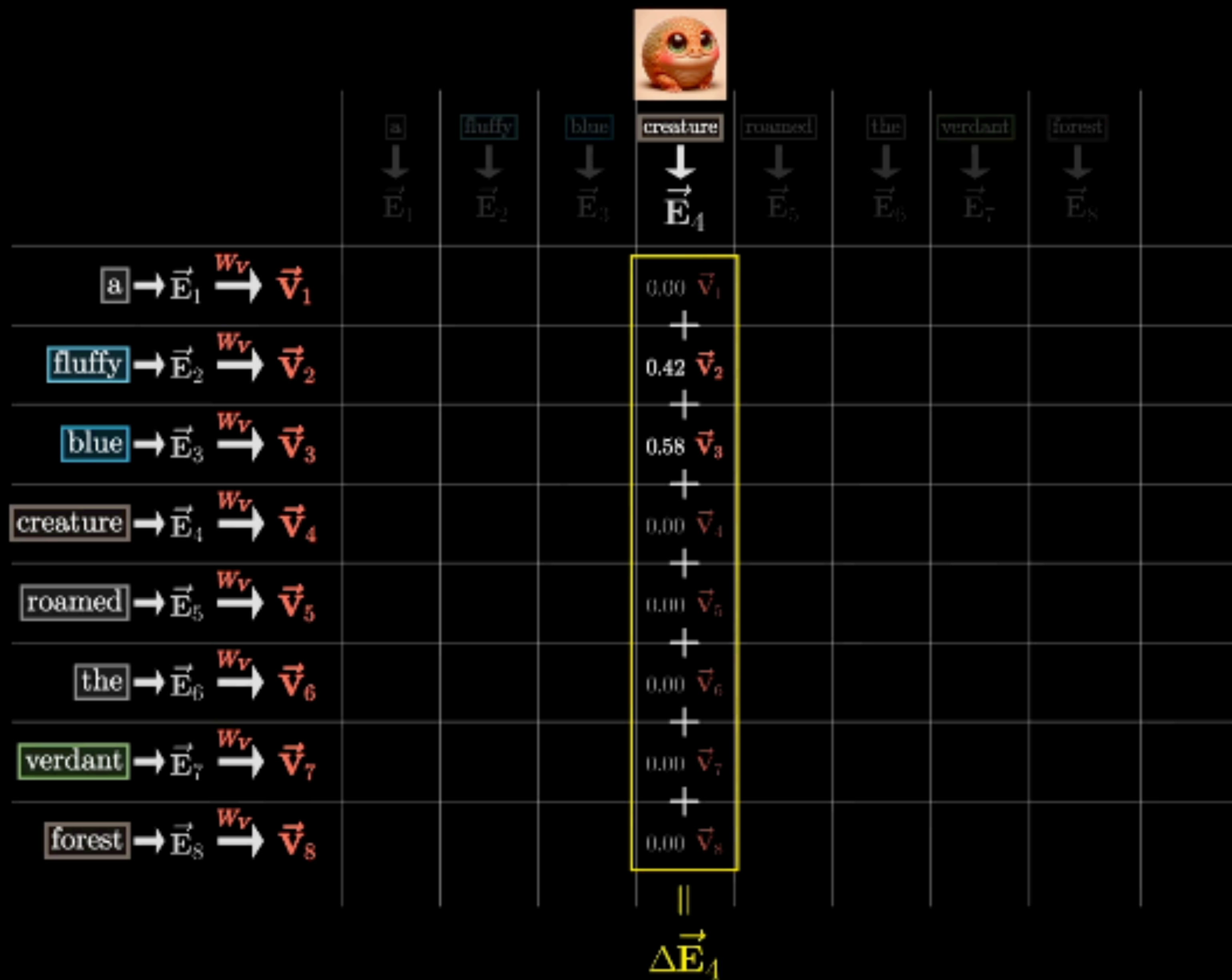


	a	fluffy	blue	creature	roamed	the	verdant	forest	
	\downarrow \vec{E}_1	\downarrow \vec{E}_2	\downarrow \vec{E}_3	\downarrow \vec{E}_4	\downarrow \vec{E}_5	\downarrow \vec{E}_6	\downarrow \vec{E}_7	\downarrow \vec{E}_8	
	\downarrow^{W_Q} \vec{Q}_1	\downarrow^{W_Q} \vec{Q}_2	\downarrow^{W_Q} \vec{Q}_3	\downarrow^{W_Q} \vec{Q}_4	\downarrow^{W_Q} \vec{Q}_5	\downarrow^{W_Q} \vec{Q}_6	\downarrow^{W_Q} \vec{Q}_7	\downarrow^{W_Q} \vec{Q}_8	
$\boxed{\text{a}} \rightarrow \vec{E}_1 \xrightarrow{W_k} \vec{K}_1$	$\vec{K}_1 \cdot \vec{Q}_1$	$\vec{K}_1 \cdot \vec{Q}_2$	$\vec{K}_1 \cdot \vec{Q}_3$	$\vec{K}_1 \cdot \vec{Q}_4$	$\vec{K}_1 \cdot \vec{Q}_5$	$\vec{K}_1 \cdot \vec{Q}_6$	$\vec{K}_1 \cdot \vec{Q}_7$	$\vec{K}_1 \cdot \vec{Q}_8$	
$\boxed{\text{fluffy}} \rightarrow \vec{E}_2 \xrightarrow{W_k} \vec{K}_2$	$\vec{K}_2 \cdot \vec{Q}_1$	$\vec{K}_2 \cdot \vec{Q}_2$	$\vec{K}_2 \cdot \vec{Q}_3$	$\vec{K}_2 \cdot \vec{Q}_4$	$\vec{K}_2 \cdot \vec{Q}_5$	$\vec{K}_2 \cdot \vec{Q}_6$	$\vec{K}_2 \cdot \vec{Q}_7$	$\vec{K}_2 \cdot \vec{Q}_8$	
$\boxed{\text{blue}} \rightarrow \vec{E}_3 \xrightarrow{W_k} \vec{K}_3$	$\vec{K}_3 \cdot \vec{Q}_1$	$\vec{K}_3 \cdot \vec{Q}_2$	$\vec{K}_3 \cdot \vec{Q}_3$	$\vec{K}_3 \cdot \vec{Q}_4$	$\vec{K}_3 \cdot \vec{Q}_5$	$\vec{K}_3 \cdot \vec{Q}_6$	$\vec{K}_3 \cdot \vec{Q}_7$	$\vec{K}_3 \cdot \vec{Q}_8$	
$\boxed{\text{creature}} \rightarrow \vec{E}_4 \xrightarrow{W_k} \vec{K}_4$	$\vec{K}_4 \cdot \vec{Q}_1$	$\vec{K}_4 \cdot \vec{Q}_2$	$\vec{K}_4 \cdot \vec{Q}_3$	$\vec{K}_4 \cdot \vec{Q}_4$	$\vec{K}_4 \cdot \vec{Q}_5$	$\vec{K}_4 \cdot \vec{Q}_6$	$\vec{K}_4 \cdot \vec{Q}_7$	$\vec{K}_4 \cdot \vec{Q}_8$	
$\boxed{\text{roamed}} \rightarrow \vec{E}_5 \xrightarrow{W_k} \vec{K}_5$	$\vec{K}_5 \cdot \vec{Q}_1$	$\vec{K}_5 \cdot \vec{Q}_2$	$\vec{K}_5 \cdot \vec{Q}_3$	$\vec{K}_5 \cdot \vec{Q}_4$	$\vec{K}_5 \cdot \vec{Q}_5$	$\vec{K}_5 \cdot \vec{Q}_6$	$\vec{K}_5 \cdot \vec{Q}_7$	$\vec{K}_5 \cdot \vec{Q}_8$	
$\boxed{\text{the}} \rightarrow \vec{E}_6 \xrightarrow{W_k} \vec{K}_6$	$\vec{K}_6 \cdot \vec{Q}_1$	$\vec{K}_6 \cdot \vec{Q}_2$	$\vec{K}_6 \cdot \vec{Q}_3$	$\vec{K}_6 \cdot \vec{Q}_4$	$\vec{K}_6 \cdot \vec{Q}_5$	$\vec{K}_6 \cdot \vec{Q}_6$	$\vec{K}_6 \cdot \vec{Q}_7$	$\vec{K}_6 \cdot \vec{Q}_8$	
$\boxed{\text{verdant}} \rightarrow \vec{E}_7 \xrightarrow{W_k} \vec{K}_7$	$\vec{K}_7 \cdot \vec{Q}_1$	$\vec{K}_7 \cdot \vec{Q}_2$	$\vec{K}_7 \cdot \vec{Q}_3$	$\vec{K}_7 \cdot \vec{Q}_4$	$\vec{K}_7 \cdot \vec{Q}_5$	$\vec{K}_7 \cdot \vec{Q}_6$	$\vec{K}_7 \cdot \vec{Q}_7$	$\vec{K}_7 \cdot \vec{Q}_8$	
$\boxed{\text{forest}} \rightarrow \vec{E}_8 \xrightarrow{W_k} \vec{K}_8$	$\vec{K}_8 \cdot \vec{Q}_1$	$\vec{K}_8 \cdot \vec{Q}_2$	$\vec{K}_8 \cdot \vec{Q}_3$	$\vec{K}_8 \cdot \vec{Q}_4$	$\vec{K}_8 \cdot \vec{Q}_5$	$\vec{K}_8 \cdot \vec{Q}_6$	$\vec{K}_8 \cdot \vec{Q}_7$	$\vec{K}_8 \cdot \vec{Q}_8$	



	a	fluffy	blue	creature	roamed	the	verdant	forest	
	\vec{E}_1	\vec{E}_2	\vec{E}_3	\vec{E}_4	\vec{E}_5	\vec{E}_6	\vec{E}_7	\vec{E}_8	
	\vec{Q}_1	\vec{Q}_2	\vec{Q}_3	\vec{Q}_4	\vec{Q}_5	\vec{Q}_6	\vec{Q}_7	\vec{Q}_8	
$\text{a} \rightarrow \vec{E}_1 \xrightarrow{W_k} \vec{K}_1$	+0.7	-83.7	-24.7	-27.8	-5.2	-89.3	-45.2	-36.1	
$\text{fluffy} \rightarrow \vec{E}_2 \xrightarrow{W_k} \vec{K}_2$	-73.4	+2.9	-5.4	+93.0	-48.2	-87.3	-49.7	+7.8	
$\text{blue} \rightarrow \vec{E}_3 \xrightarrow{W_k} \vec{K}_3$	-53.4	-5.7	+1.8	+93.4	-55.6	-56.0	-26.1	-62.1	
$\text{creature} \rightarrow \vec{E}_4 \xrightarrow{W_k} \vec{K}_4$	-21.5	-29.7	-56.1	+4.9	-32.4	-92.3	-9.5	-28.1	
$\text{roamed} \rightarrow \vec{E}_5 \xrightarrow{W_k} \vec{K}_5$	-20.1	-40.9	-87.8	-55.4	+0.6	-64.7	-96.7	-18.9	
$\text{the} \rightarrow \vec{E}_6 \xrightarrow{W_k} \vec{K}_6$	-87.9	-33.3	-22.6	-31.4	+5.5	+0.6	-4.6	-96.8	
$\text{verdant} \rightarrow \vec{E}_7 \xrightarrow{W_k} \vec{K}_7$	-41.2	-55.5	-42.3	-59.8	-79.0	-97.9	+3.7	+93.8	
$\text{forest} \rightarrow \vec{E}_8 \xrightarrow{W_k} \vec{K}_8$	-58.9	-75.5	-91.1	-90.6	-75.6	-89.0	-70.8	+4.7	





creature



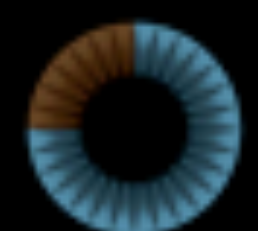
\vec{E}_4



$\Delta \vec{E}_4$



\vec{E}'_4



One head of attention

	▣	Daffy	the	creature	roamed	the	verlax	forest
	\vec{E}_1	\vec{E}_2	\vec{E}_3	\vec{E}_4	\vec{E}_5	\vec{E}_6	\vec{E}_7	\vec{E}_8
$\vec{E}_1 \xrightarrow{W_V} \vec{v}_1$	1.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1
$\vec{E}_2 \xrightarrow{W_V} \vec{v}_2$	0.00 \vec{v}_2	1.00 \vec{v}_2	0.00 \vec{v}_2	0.42 \vec{v}_2	0.00 \vec{v}_2	0.00 \vec{v}_2	0.00 \vec{v}_2	0.00 \vec{v}_2
$\vec{E}_3 \xrightarrow{W_V} \vec{v}_3$	0.00 \vec{v}_3	0.00 \vec{v}_3	1.00 \vec{v}_3	0.58 \vec{v}_3	0.00 \vec{v}_3	0.00 \vec{v}_3	0.00 \vec{v}_3	0.00 \vec{v}_3
$\vec{E}_4 \xrightarrow{W_V} \vec{v}_4$	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4
$\vec{E}_5 \xrightarrow{W_V} \vec{v}_5$	0.00 \vec{v}_5	0.00 \vec{v}_5	0.00 \vec{v}_5	0.00 \vec{v}_5	0.01 \vec{v}_5	0.00 \vec{v}_5	0.00 \vec{v}_5	0.00 \vec{v}_5
$\vec{E}_6 \xrightarrow{W_V} \vec{v}_6$	0.00 \vec{v}_6	0.00 \vec{v}_6	0.00 \vec{v}_6	0.00 \vec{v}_6	0.99 \vec{v}_6	1.00 \vec{v}_6	0.00 \vec{v}_6	0.00 \vec{v}_6
$\vec{E}_7 \xrightarrow{W_V} \vec{v}_7$	0.00 \vec{v}_7	0.00 \vec{v}_7	0.00 \vec{v}_7	0.00 \vec{v}_7	0.00 \vec{v}_7	0.00 \vec{v}_7	1.00 \vec{v}_7	1.00 \vec{v}_7
$\vec{E}_8 \xrightarrow{W_V} \vec{v}_8$	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8
	$\Delta \vec{E}_1$	$\Delta \vec{E}_2$	$\Delta \vec{E}_3$	$\Delta \vec{E}_4$	$\Delta \vec{E}_5$	$\Delta \vec{E}_6$	$\Delta \vec{E}_7$	$\Delta \vec{E}_8$

\vec{E}_1	\vec{E}_2	\vec{E}_3	\vec{E}_4	\vec{E}_5	\vec{E}_6	\vec{E}_7	\vec{E}_8
+	+	+	+	+	+	+	+
$\Delta \vec{E}_1$	$\Delta \vec{E}_2$	$\Delta \vec{E}_3$	$\Delta \vec{E}_4$	$\Delta \vec{E}_5$	$\Delta \vec{E}_6$	$\Delta \vec{E}_7$	$\Delta \vec{E}_8$
\vec{E}'_1	\vec{E}'_2	\vec{E}'_3	\vec{E}'_4	\vec{E}'_5	\vec{E}'_6	\vec{E}'_7	\vec{E}'_8



One head attention

$x_1, x_2, \dots, x_n \in R^{d_{in}}$ - token embeddings

embedding matrix: $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in R^{n \times d_{in}}$

d_k - dim of key & query vectors $d_v = d_{in}$ - dim of value vectors

$W_Q, W_K \in R^{d_{in} \times d_k}, W_V \in R^{d_{in} \times d_v}$ - query, key and value transform matrices

Query, key, and value matrices:

$$Q = XW_Q \in R^{n \times d_k} \quad K = XW_K \in R^{n \times d_k} \quad V = XW_V \in R^{n \times d_v}$$

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

One head attention

$x_1, x_2, \dots, x_n \in R^{d_{in}}$ - token embeddings

embedding matrix: $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in R^{n \times d_{in}}$

d_k - dim of key & query vectors $d_v = d_{in}$ - dim of value vectors

$W_Q, W_K \in R^{d_{in} \times d_k}, W_V \in R^{d_{in} \times d_v}$ - query, key and value transform matrices

Query, key, and value matrices:

$$Q = XW_Q \in R^{n \times d_k}$$

$$K = XW_K \in R^{n \times d_k}$$

$$V = XW_V \in R^{n \times d_v}$$

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

Time complexity: $O(nd_{in}^2 + n^2d_{in})$

Multihead attention

$$W_i^Q, W_i^K \in R^{d_{in} \times d_k}, W_i^V \in R^{d_{in} \times d_v}, i \in [n] \quad W^O \in R^{hd_{in} \times d_v}$$

$$head_i = Attention(XW_i^Q, XW_i^K, XW_i^V)$$

$$MultiHead(\{W_i^Q, W_i^K, W_i^V\}_{i=1}^N) = Concat(head_1, \dots, head_n)W^O$$

Multihead attention

$$W_i^Q, W_i^K \in R^{d_{in} \times d_k}, W_i^V \in R^{d_{in} \times d_v}, i \in [h] \quad W^O \in R^{hd_{in} \times d_v}$$

$$head_i = Attention(XW_i^Q, XW_i^K, XW_i^V)$$

$$MultiHead(\{W_i^Q, W_i^K, W_i^V\}_{i=1}^N) = Concat(head_1, \dots, head_h)W^O$$

Good practice: $d_k = d_v = \frac{d_{in}}{h} \Rightarrow$ **Time complexity:** $O(nd_{in}^2 + n^2d_{in})$

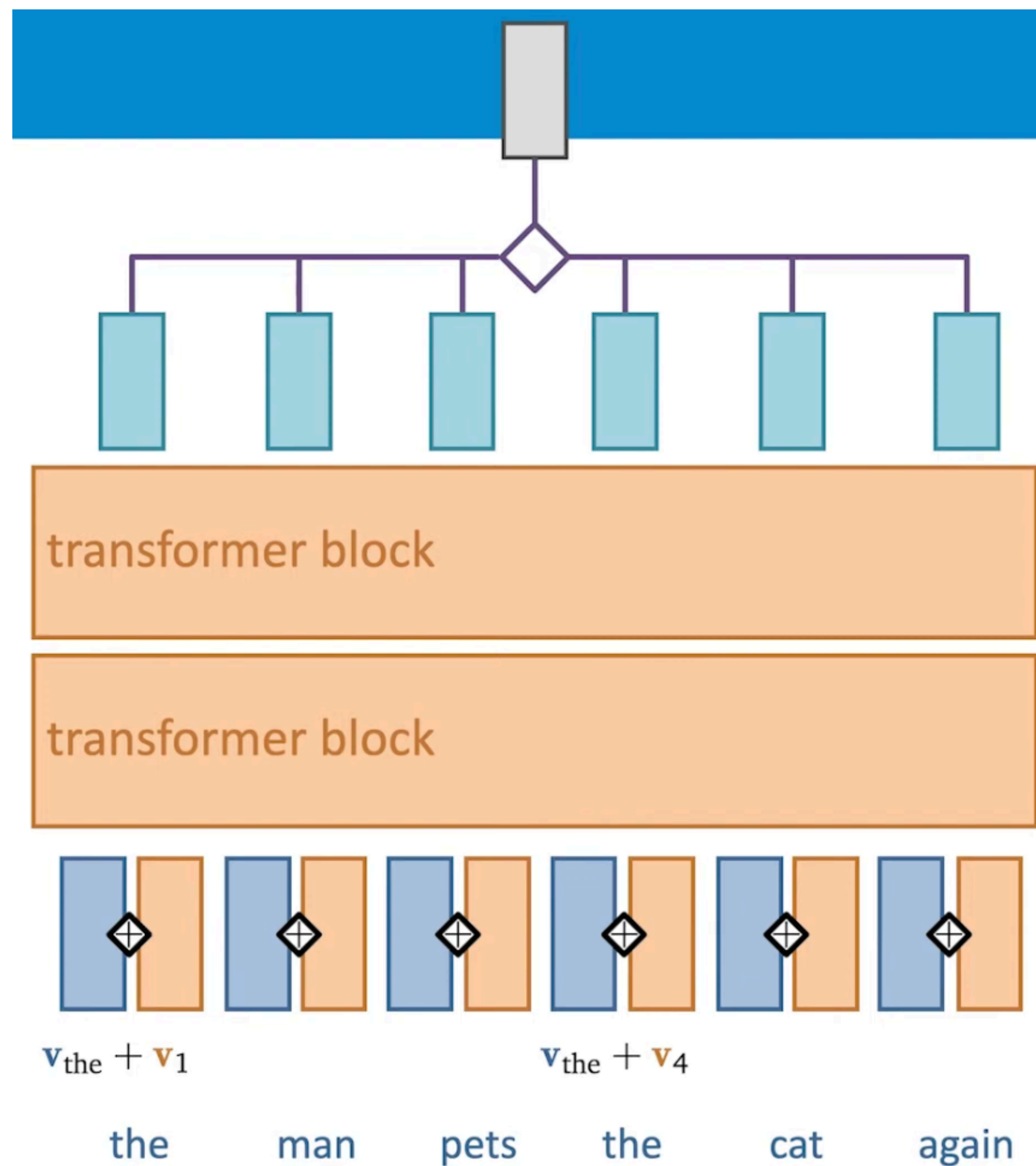
Position embeddings

word embeddings:

\mathbf{v}_{the} , \mathbf{v}_{man} , \mathbf{v}_{pets} , \mathbf{v}_{cat} , $\mathbf{v}_{\text{again}}$

position embeddings:

\mathbf{v}_1 , \mathbf{v}_2 , \mathbf{v}_3 , \mathbf{v}_4 , \mathbf{v}_5 , ...



Transformer architecture

- **Layer Normalisation**

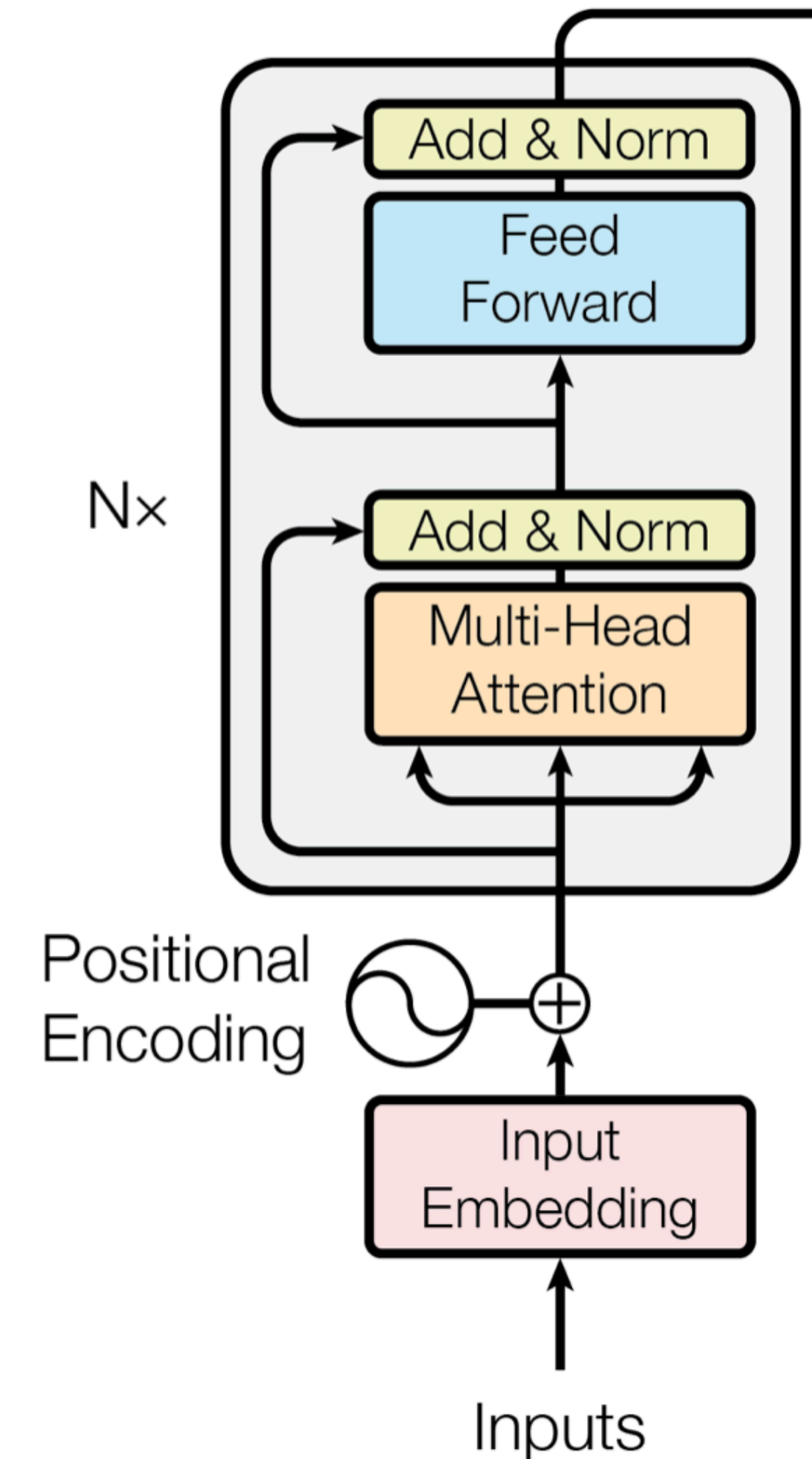
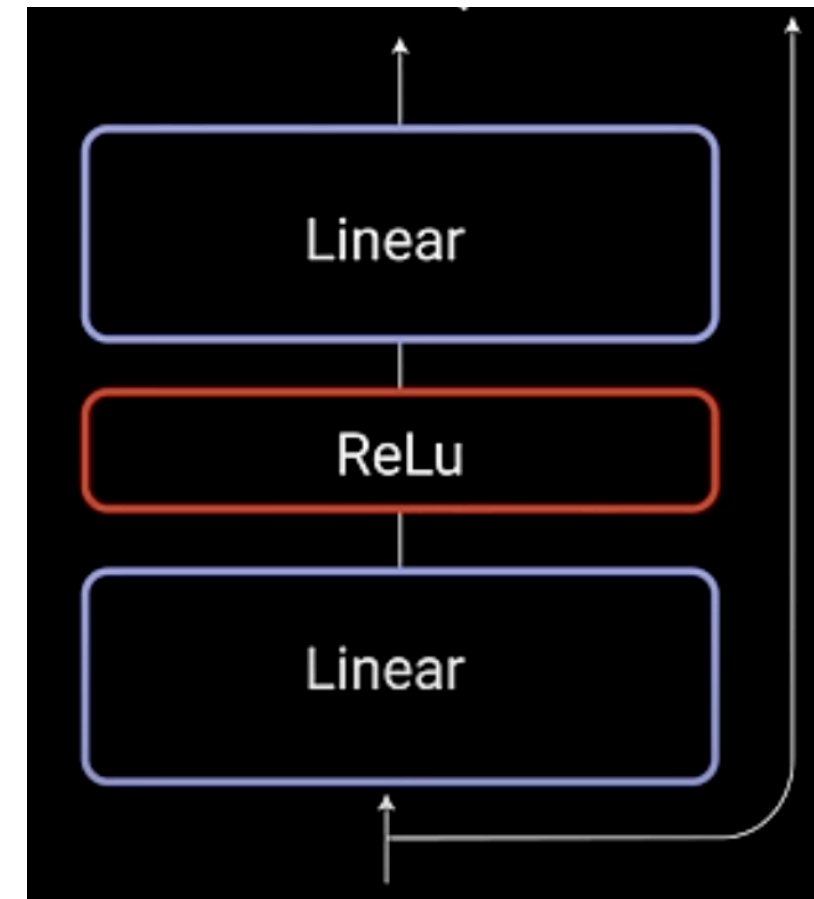
- Normalize the outputs of different modules

$$y = \frac{x - E[x]}{\sqrt{\text{Var}[x] + \epsilon}} * \gamma + \beta$$

- **Residual Connections**

- Add the input of a module to its output
- $\text{LayerNorm}(x + \text{Sublayer}(x))$

Feed-Forward



Transformer architecture

- **Layer Normalisation**

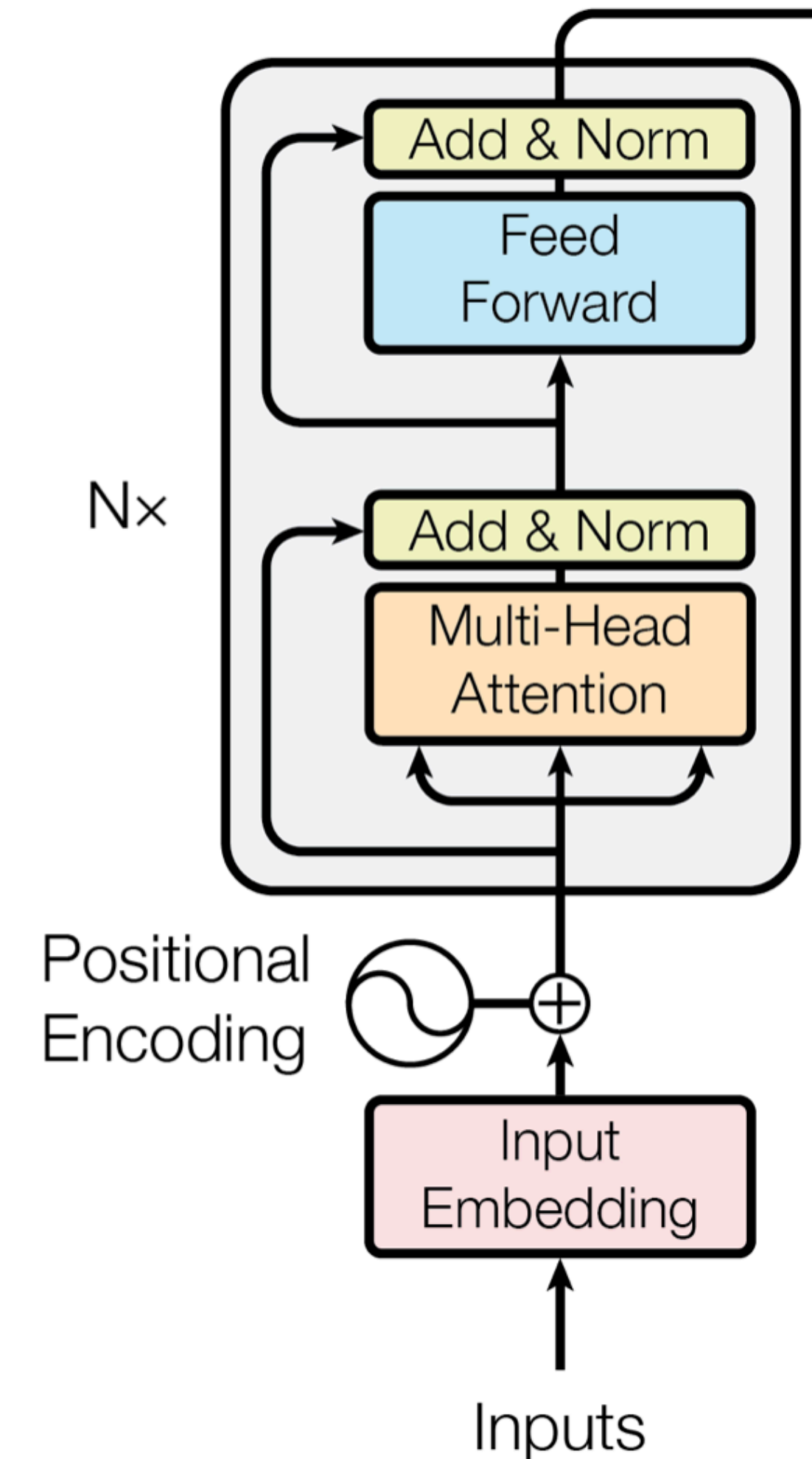
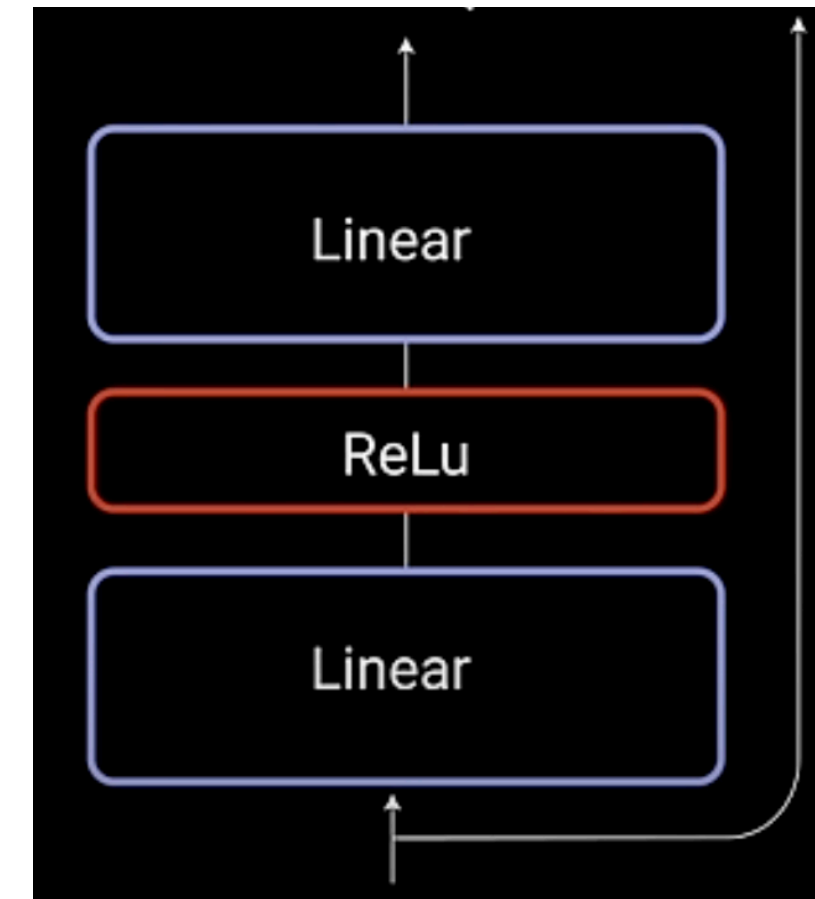
- Normalize the outputs of different modules

$$y = \frac{x - E[x]}{\sqrt{\text{Var}[x] + \epsilon}} * \gamma + \beta$$

- **Residual Connections**

- Add the input of a module to its output
- $\text{LayerNorm}(x + \text{Sublayer}(x))$

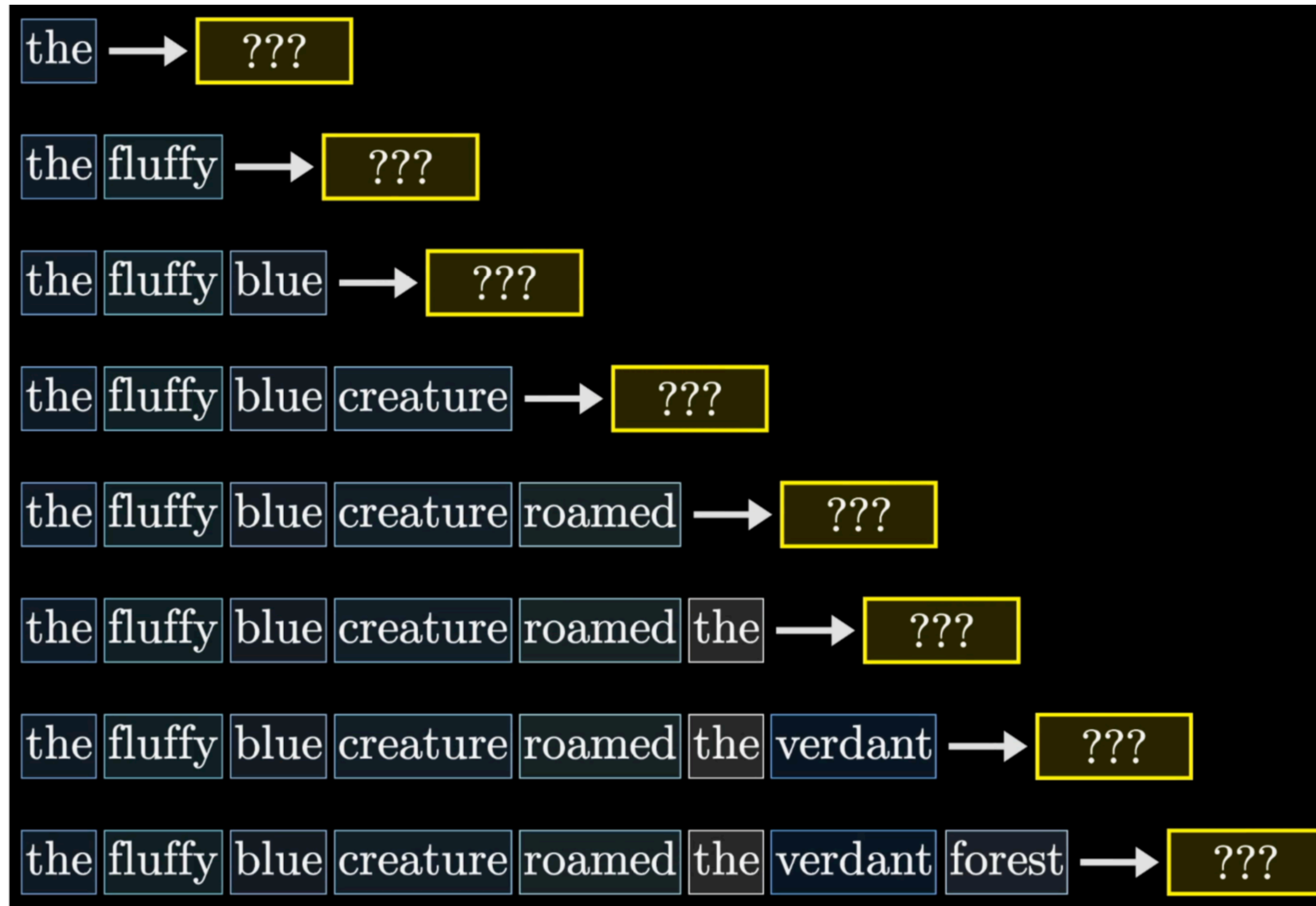
Feed-Forward



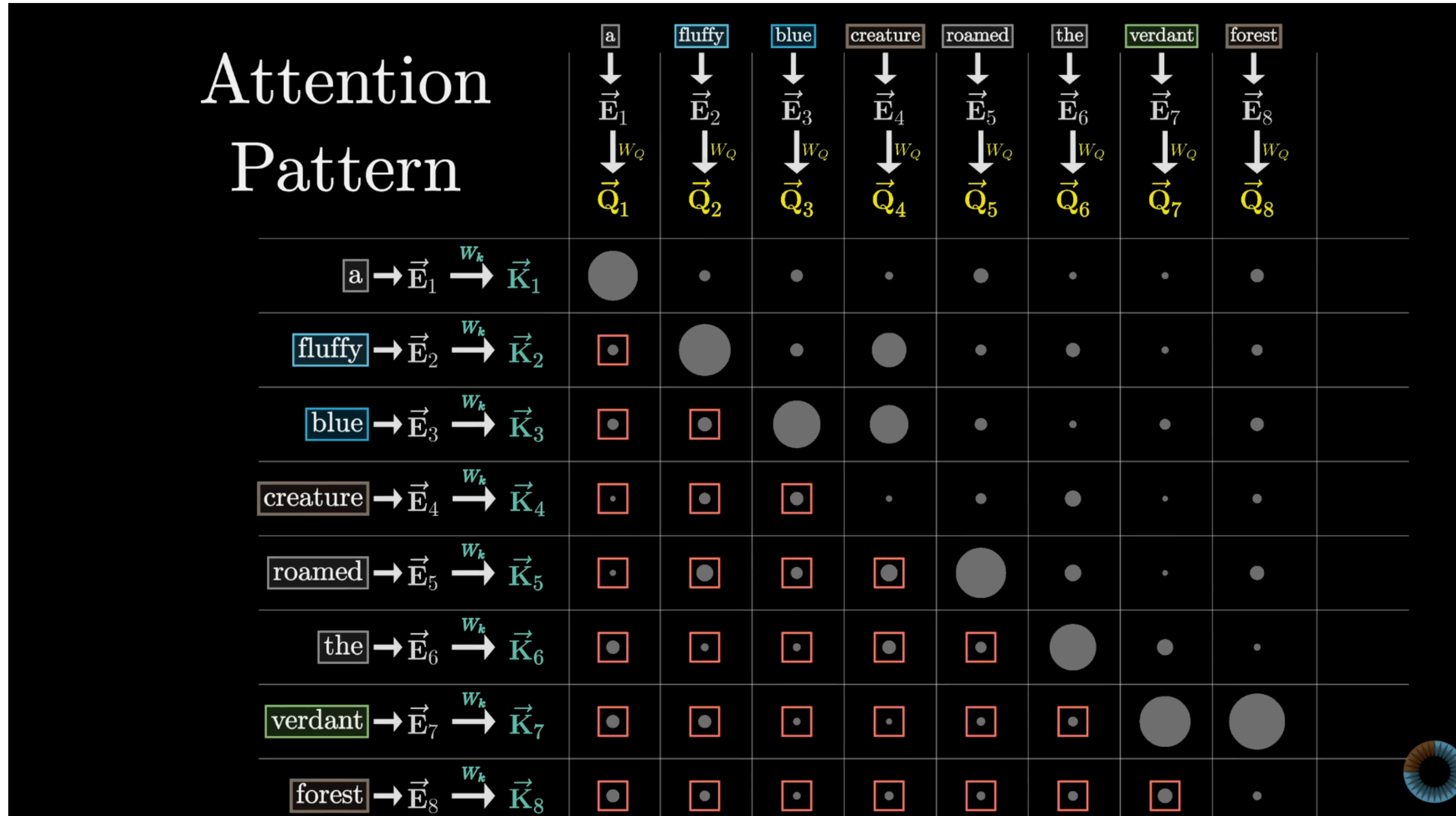
Feed-Forward layer is applied to tokens pointwise.

Tokens only interact in Attention layers!

Transformer for next token prediction



Attention Masking



Attention Masking

Unnormalized
Attention Pattern

+3.53	+0.80	+1.96	+4.48	+3.74	-1.95
$-\infty$	-0.30	-0.21	+0.82	+0.29	+2.91
$-\infty$	$-\infty$	+0.89	+0.67	+2.99	-0.41
$-\infty$	$-\infty$	$-\infty$	+1.31	+1.73	-1.48
$-\infty$	$-\infty$	$-\infty$	$-\infty$	+3.07	+2.94
$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	+0.31

softmax
→

Normalized
Attention Pattern

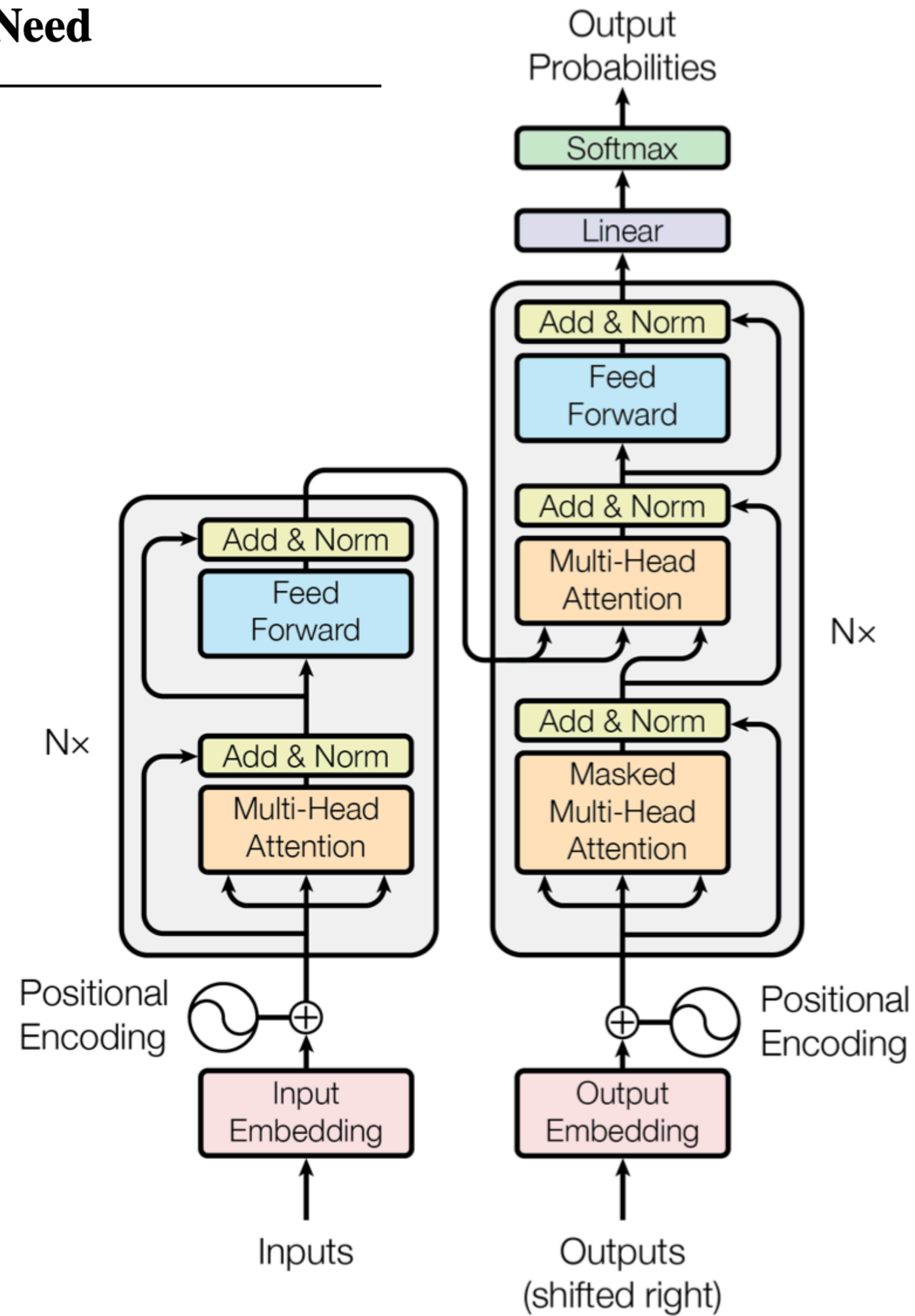
1.00	0.75	0.69	0.92	0.46	0.00
0.00	0.25	0.08	0.02	0.01	0.46
0.00	0.00	0.24	0.02	0.22	0.02
0.00	0.00	0.00	0.04	0.06	0.01
0.00	0.00	0.00	0.00	0.24	0.48
0.00	0.00	0.00	0.00	0.00	0.03

Terminology

- Attention without masking = bidirectional self-attention
- Attention with masking = autoregressive self-attention

Transformer: original encoder-decoder architecture (obsolete)

Attention Is All You Need



Transformers: advantages

- Attention mechanism: any token can attend to any other => good at long-term dependancies
- Architecture is designed to be highly parallelizable
- Universality: no inductive bias in architecture

Transformer: downsides

- Bottleneck: $O(n^2)$ computation and memory complexity
- More efficient attention mechanisms were proposed (active area of research)
- Requires a lot of data for training