自然語言處理與應用

Natural Language Processing and Applications

Transformers

Instructor: 林英嘉 (Ying-Jia Lin)

2025/03/24

Course GitHub

Slido # NLP_0324

Outline

- Weakness of RNNs and CNNs
- Transformers
- HW2

Tokens vs. words

- token 是 (語言) 模型在每個時間點處理的單位
- word 是語言本身的單位
  - token 可以是 word，也可以是 sub-word
  - 一個 word 可以是一個 token，但單純講 token 不一定指的是 word

Traditional word tokenization

Sub-word Tokenization

Issues with RNNs: Linear Interaction Distance

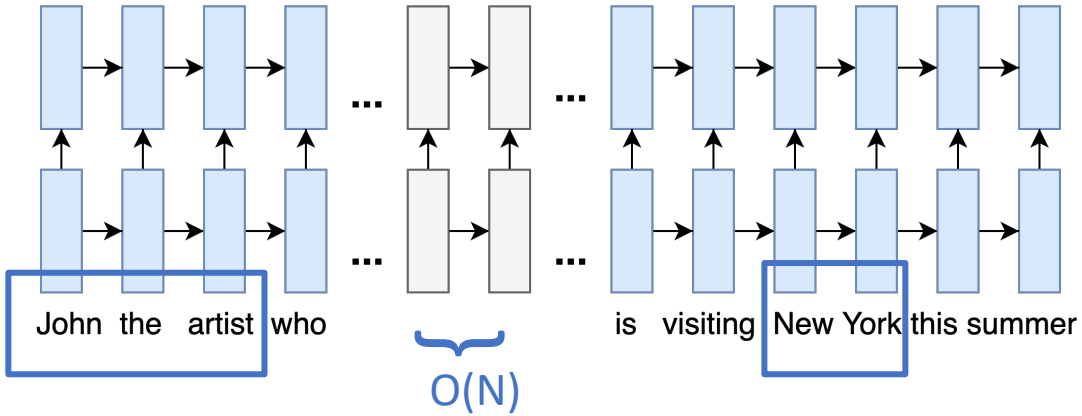
- In RNNs, the degree of words interact with each other is decided by their distance, but we already know that linear importance is not the right way to understand a sentence...

A Long Relative Clause

John, the artist who has painted many Murals in Vienna, is visiting New York this summer.

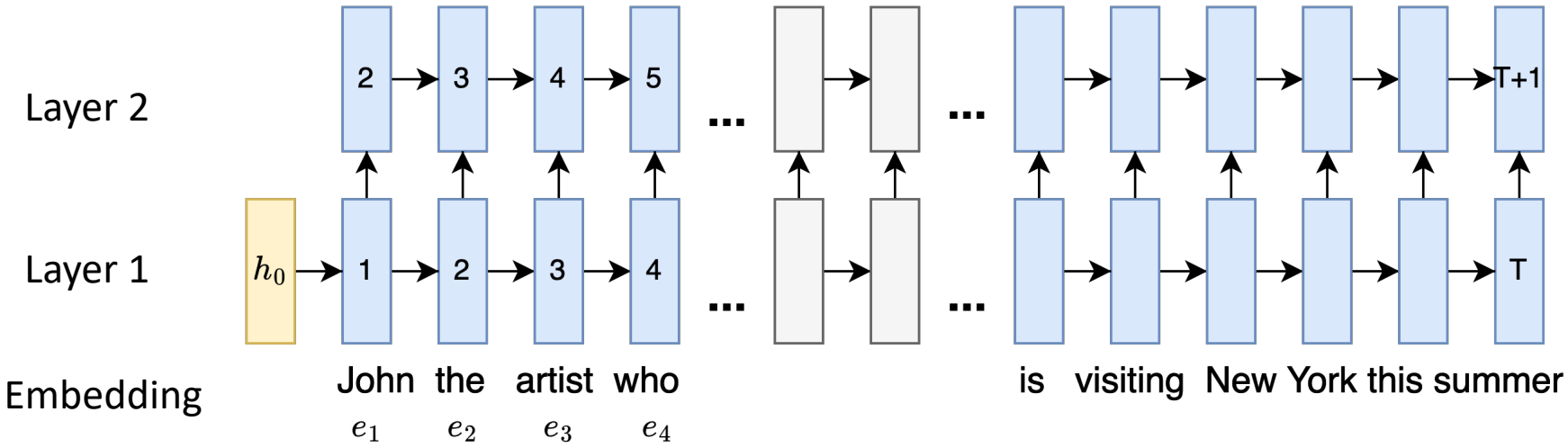
Where is John the artist visiting this summer?
A) Vienna
B) New York

▷ But the RNN may fail to identify New York as the correct answer since the info of "John" has been propagated for almost O(N) (N: sequence length).



Issues with RNNs: No Parallelizability

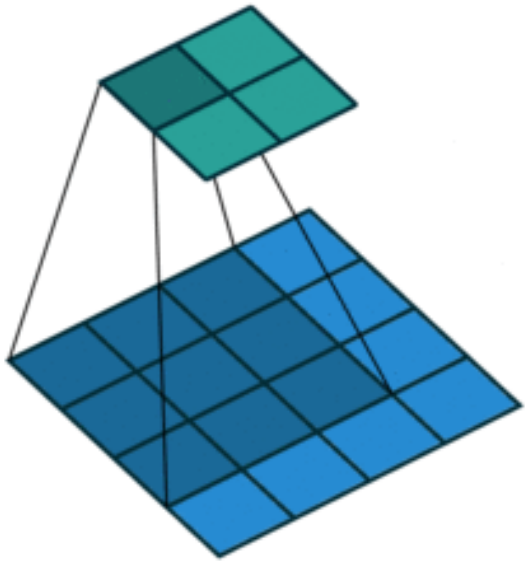
- Past hidden states need to be fully computed before the next hidden state to be ready for computation, which means it is unparallelizable between each timestep. This prevents RNN's use on large corpus or long text.



△ This is a 2-layer, uni-directional RNN. Each blue rectangle is a hidden state at the time-step of the layer. The number indicates the min # of steps required for the hidden state to be computed.

Issues with CNNs: long-term dependencies

Padding = 0, stride = 1



Padding = 0, stride = 2

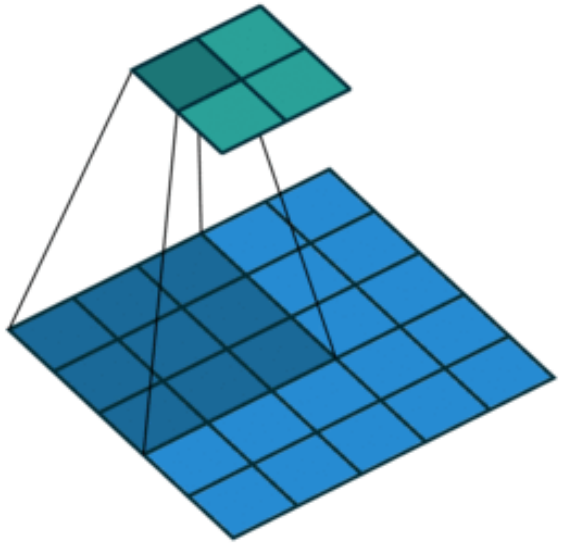
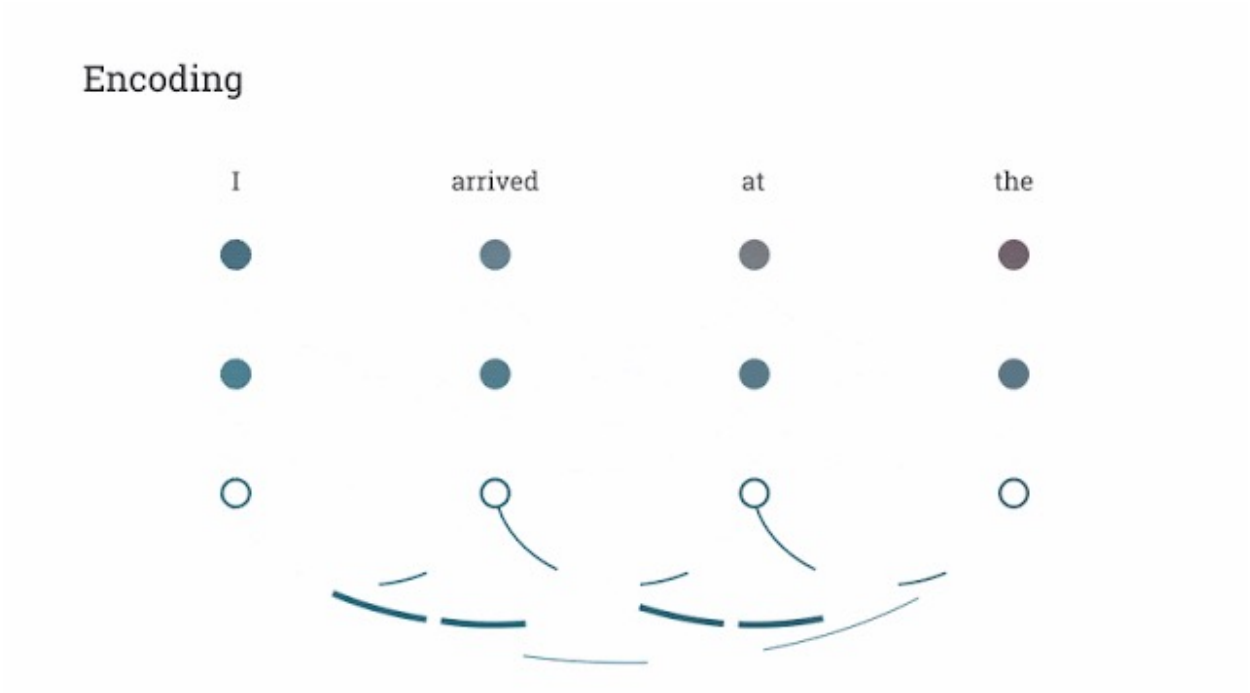


Figure source: https://hannibunny.github.io/mlbook/neuralnetworks/convolutionDemos.html

Transformers

- Attention Is All You Need (Vaswani et al., NeurIPS 2017)

Self-attention : 句子內的每個 token 自己跟自己做attention



Source: An example of the self-attention mechanism following long-distance... | Download Scientific Diagram (researchgate.net)

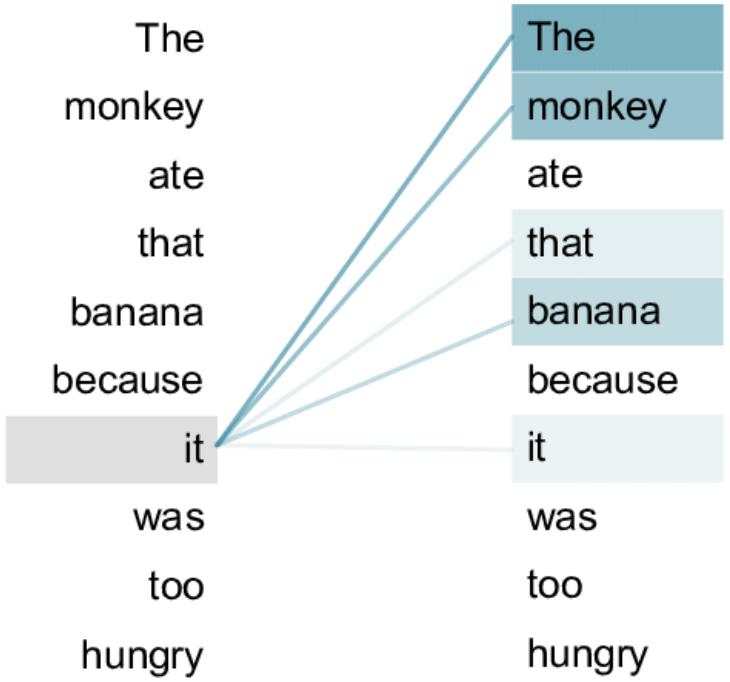


Figure source: https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

Attention Is All You Need

Essential AI

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Anthropic

Niki Parmar* Google Research nikip@google.com

Inceptive

Jakob Uszkoreit* Google Research usz@google.com

Cohere

Sakana AI

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

OpenAI

NEAR

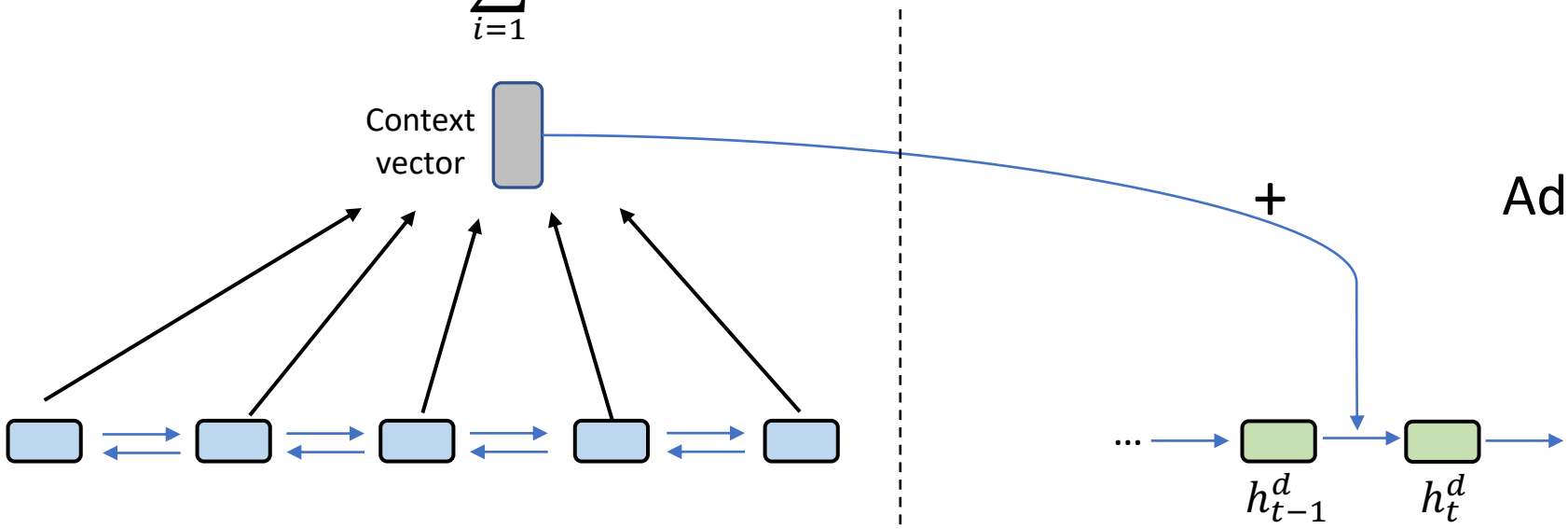
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

https://arxiv.org/abs/1706.03762v6

[Recap] Attention mechanism

i: encoder timestep
t: decoder timestep

$$c_t^e = \sum_{i=1}^{n} \alpha_{ti}^e h_i^e$$



Additive attention

$$h_i^e = [h_i^{e\_fwd} \parallel h_i^{e\_bwd}]$$

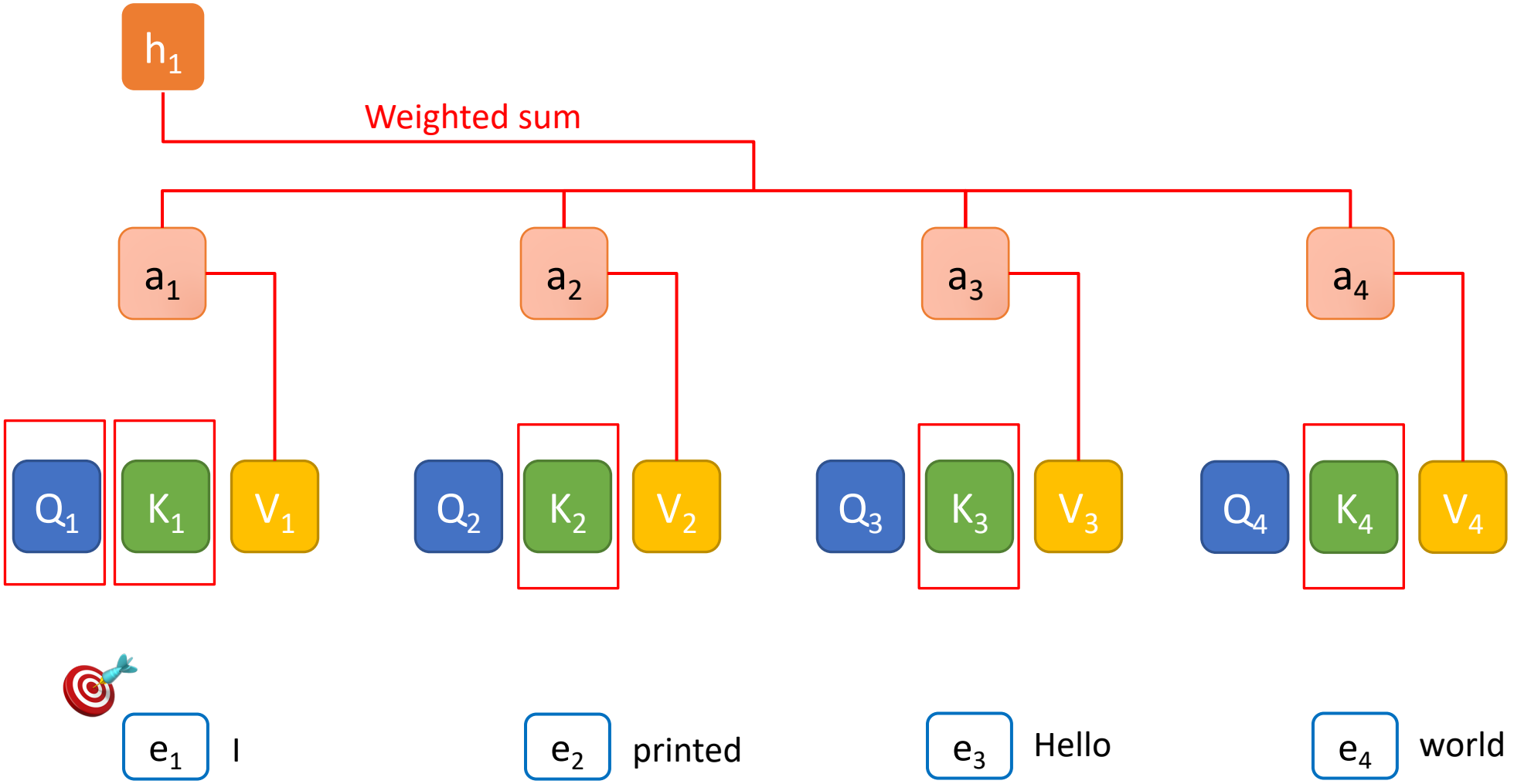
$$e_{ti} = \text{align}(h_{t-1}^d, h_i^e)$$

$$\alpha_{ti}^e = \frac{\exp(e_{ti})}{\sum_{j=1}^{n} \exp(e_{tj})}$$

Vanilla Transformers

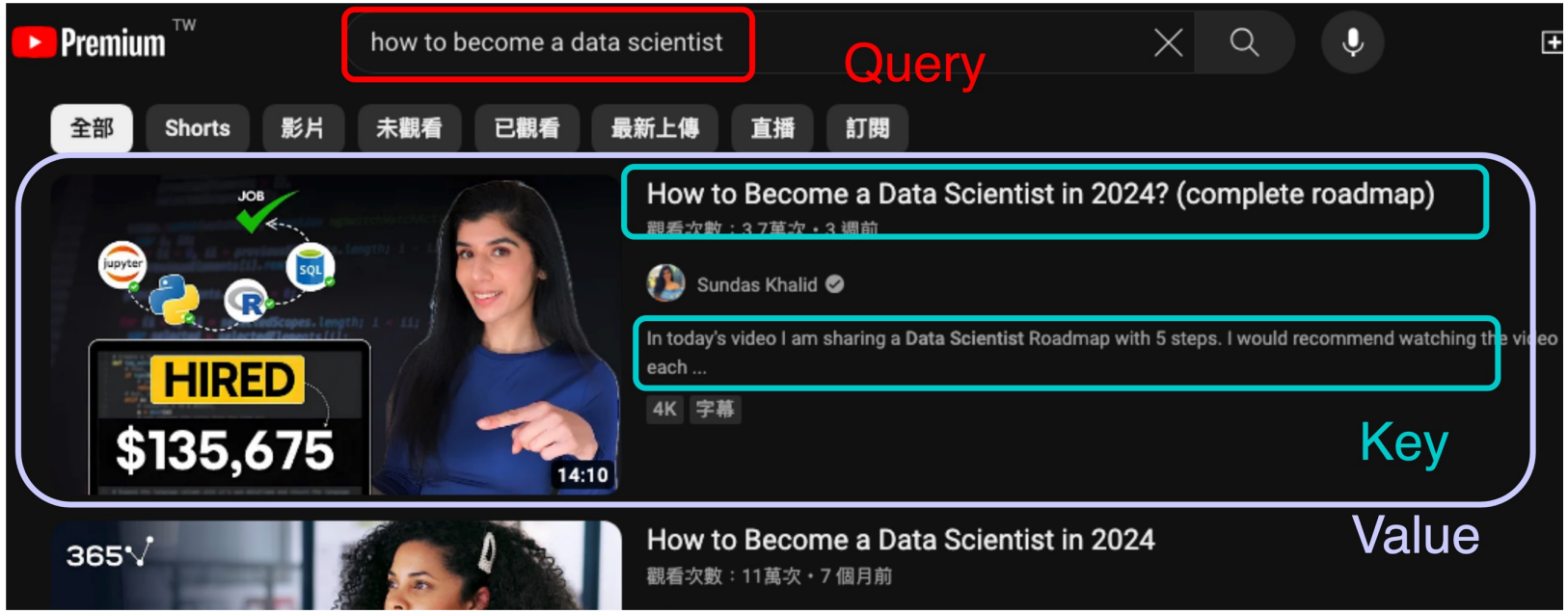
- The vanilla transformers (Vaswani et al., 2017) are also an encoder-decoder model!
- Self-attention exists in both the encoder and decoder parts.

Query (Q), Key (K), and Value (V)



The Concept of Query (Q), Key (K), and Value (V)

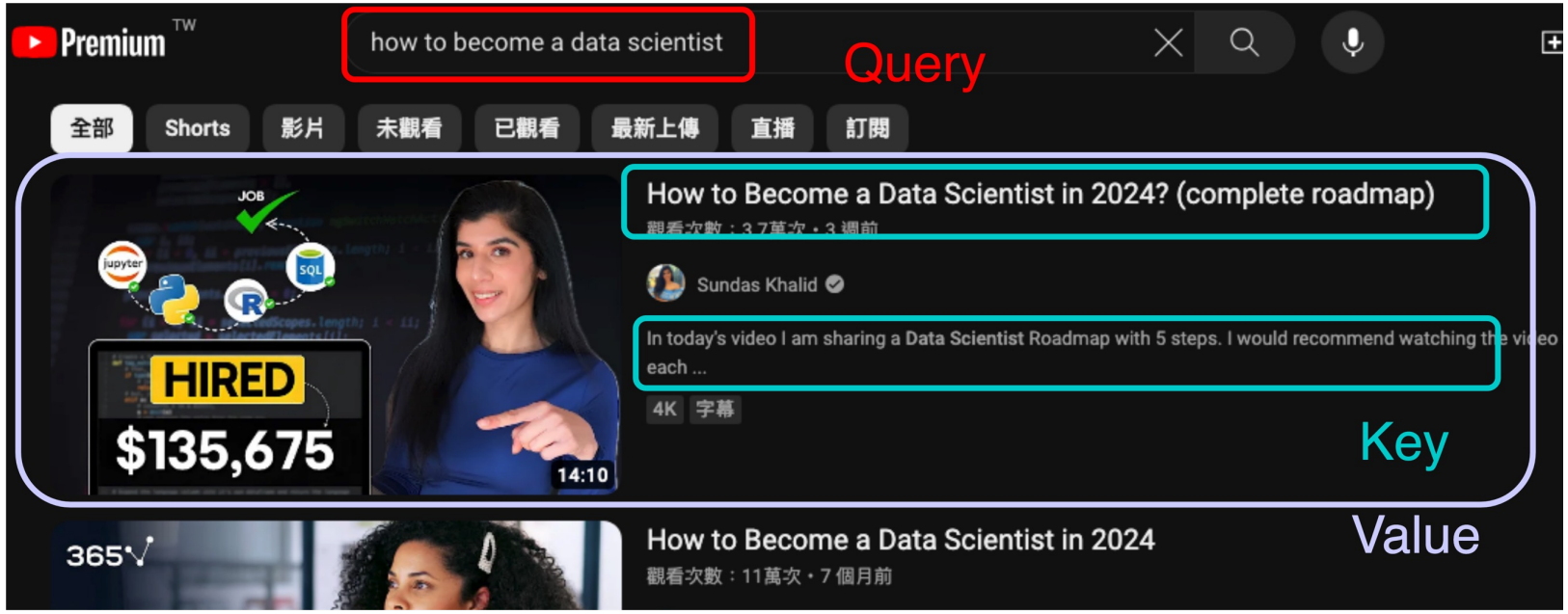
Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).



- The self-attention proposed in Vaswani et al. 2017 is query-key-value attention (QKV attention).
- What is query-key-value attention?

The Concept of Query (Q), Key (K), and Value (V)

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).



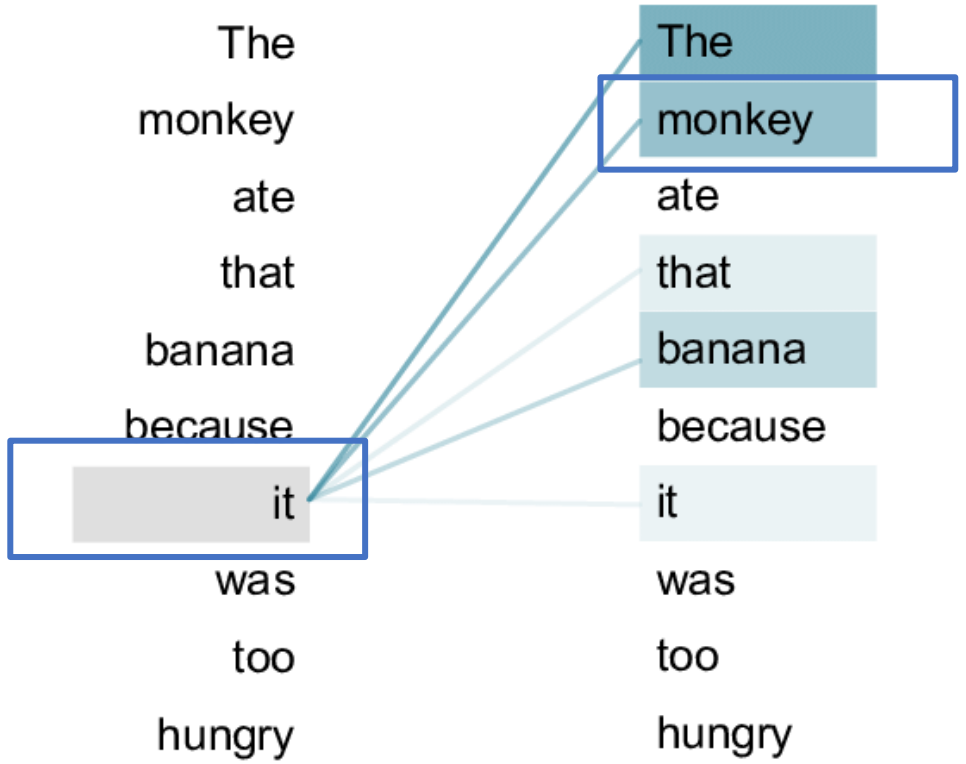
- The self-attention proposed in Vaswani et al. 2017 is query-key-value attention (QKV attention).
- What is query-key-value attention?
  - The whole process is analogous to a retrieval system.
  - We query the search engine (eg. YouTube), which tries to map our query to the keys such as video title and video descriptions in its database, and then return some best matched videos (values).

QKV Attention

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

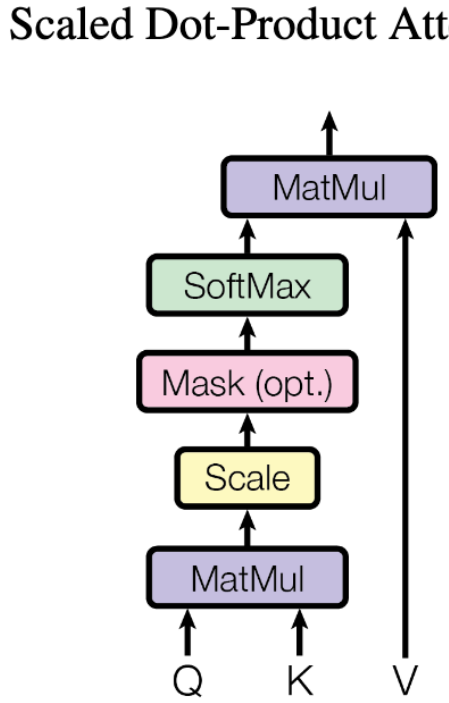
Source: An example of the self-attention mechanism following long-distance... | Download Scientific Diagram (researchgate.net)

- Back to the Natural Language example
  - The query is the information being looked for
    - Coreference Resolution object of “it”
  - The key is the context or reference
    - This is abstract ...
    - Each word w in the sentence gives hints of what “it” could be; eg. the **relative distance between w and it, `because`, `was` (neighboring) part-of-speech tags**, etc.
  - The value is the content being searched
    - One of the word in the sentence is the answer (monkey)

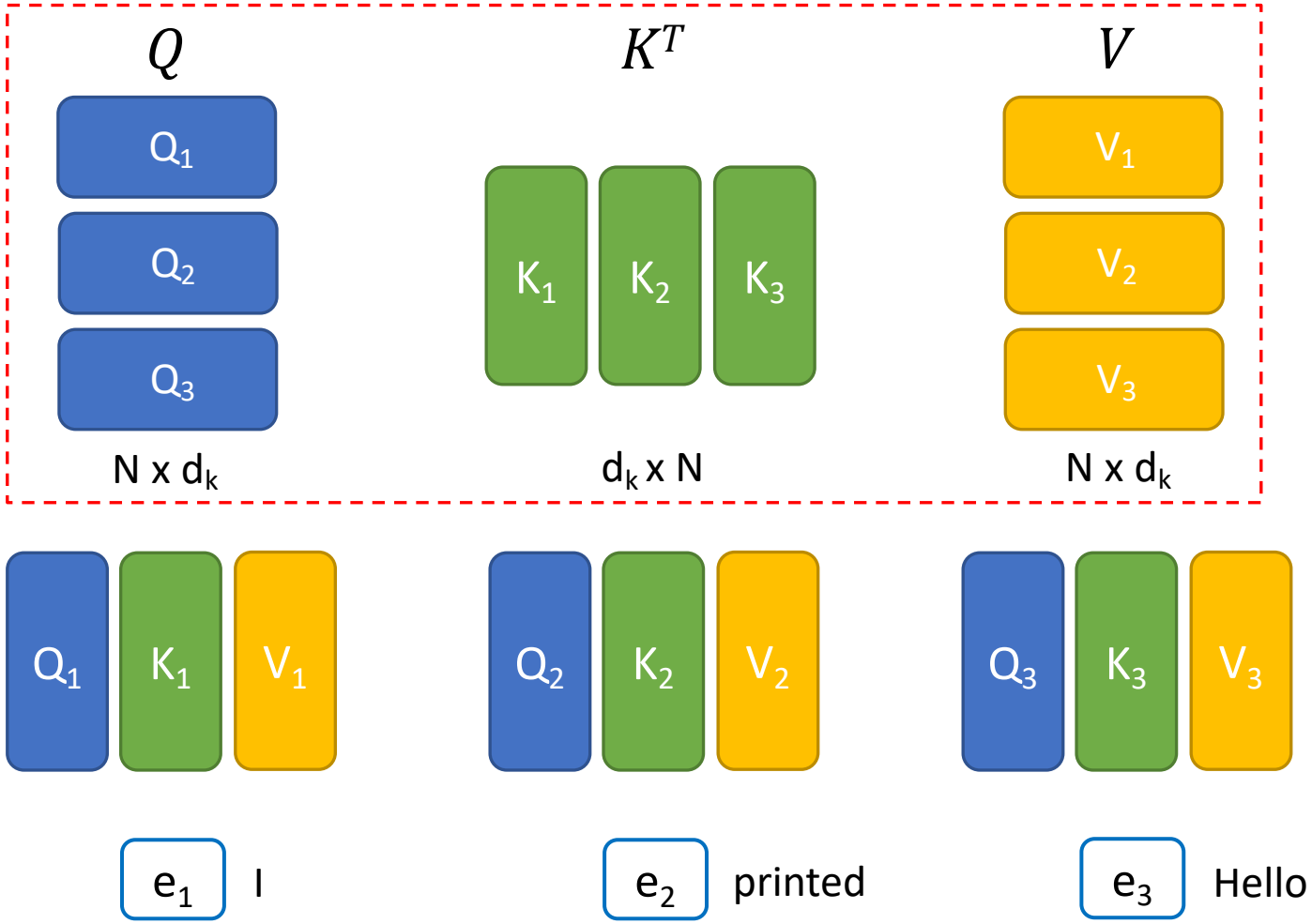


為什麼 Transformers 可以平行化？

(自注意力機制可以利用矩陣乘積來進行平行化計算)

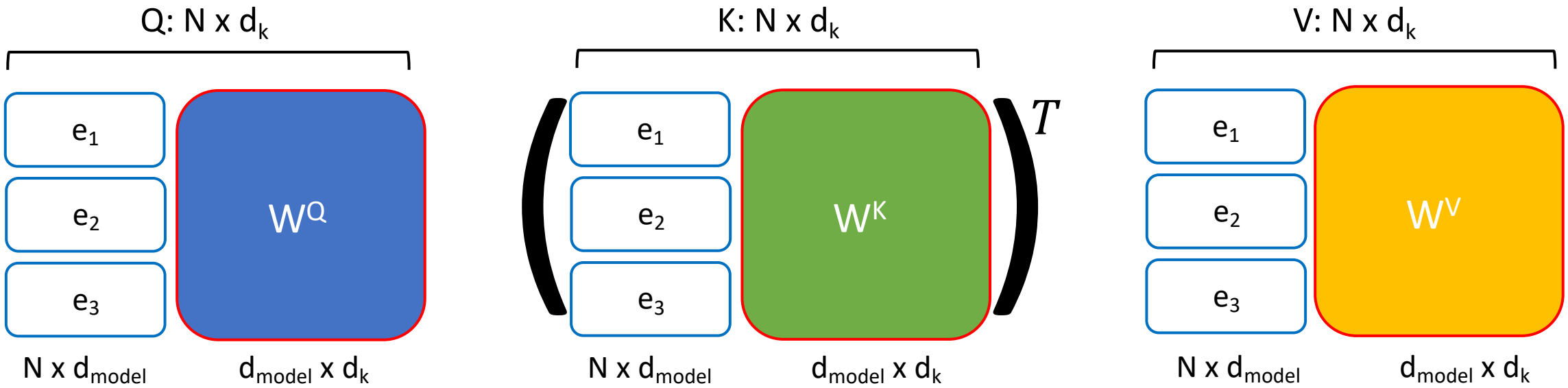


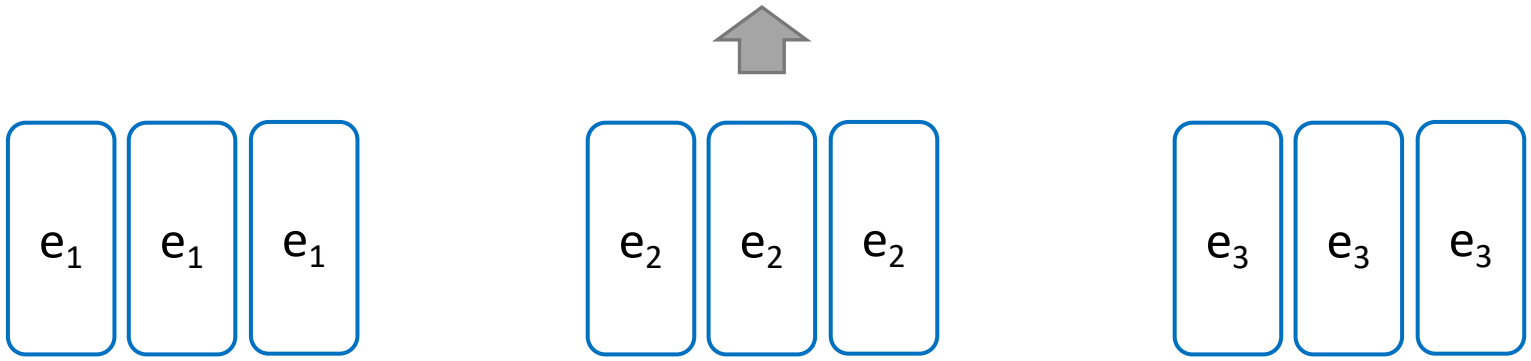
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



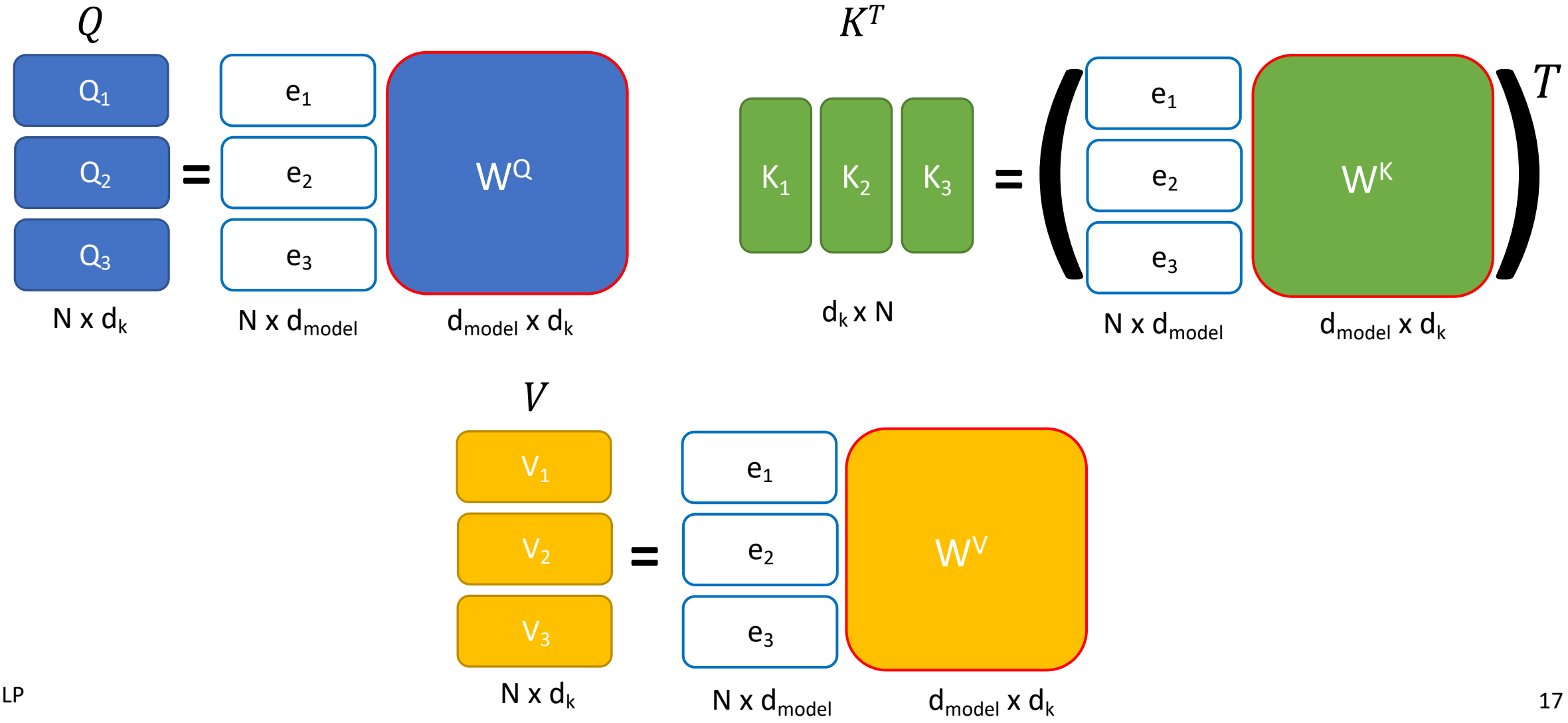
Transformers 的權重值 (1/2)

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





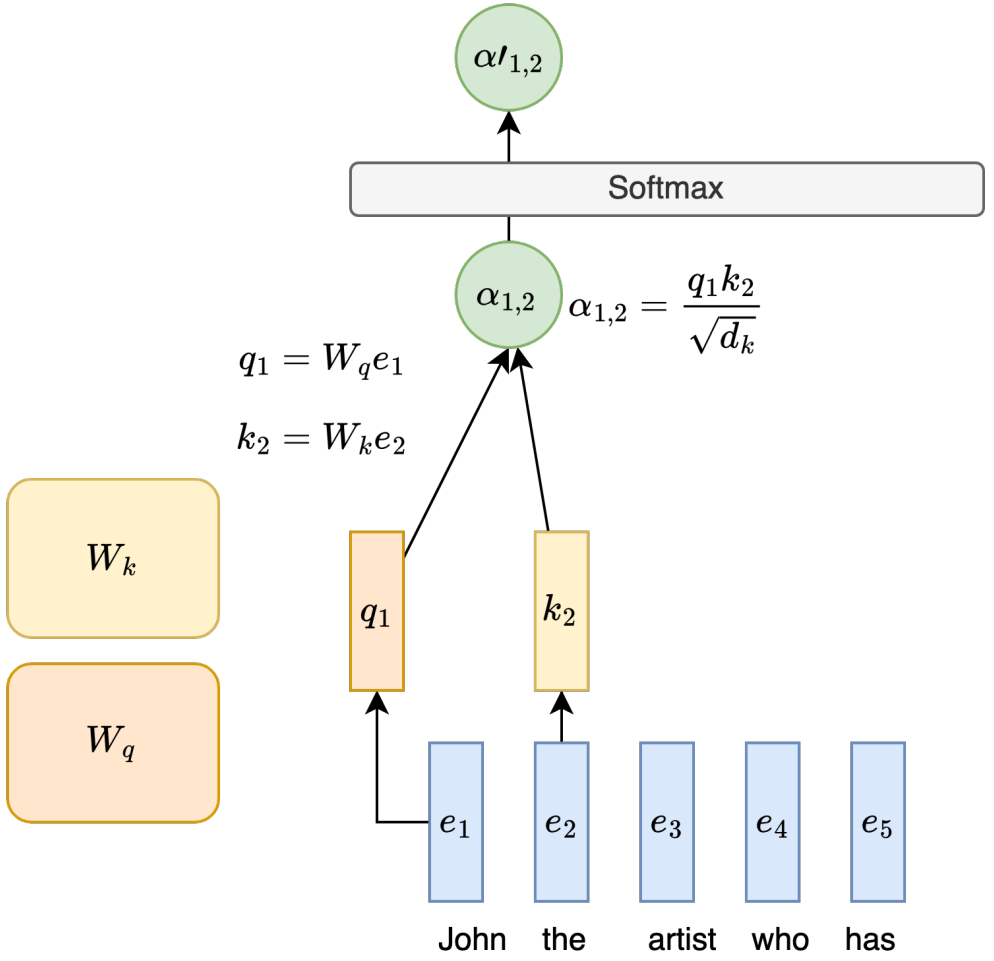
Transformers 的權重值 (2/2)



Self-Attention

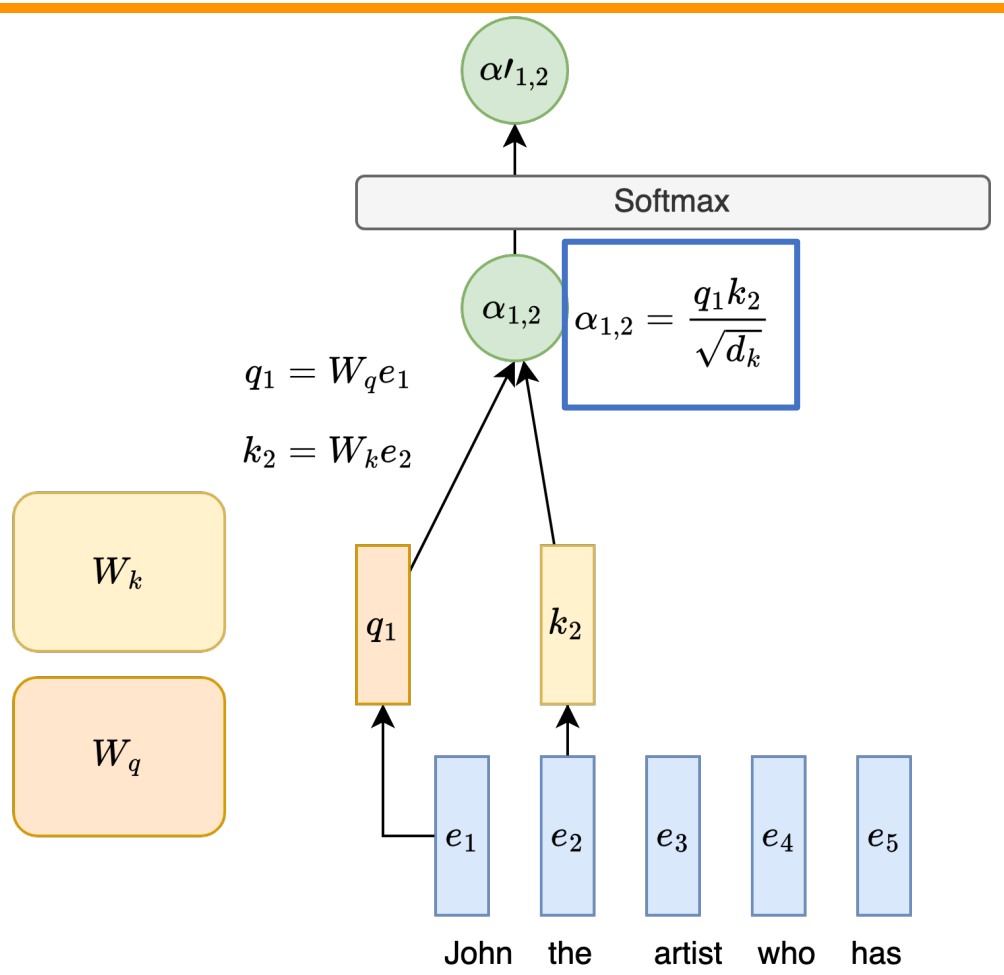
Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

- Let $e_1, .., e_N; e_i \in \mathbb{R}^d$ be input embeddings of the text sequence.
- Initialize 3 matrices W_q, W_k, W_v.
  - The matrices are used to project the input from $e_1, .., e_N$ to $q_1, .., q_N$, $k_1, .., k_N$ and $v_1, .., v_N$ respectively.
- Do a scaled dot product to get a scalar $\alpha_{1,2}$, meaning "attention score from word 1 to 2."



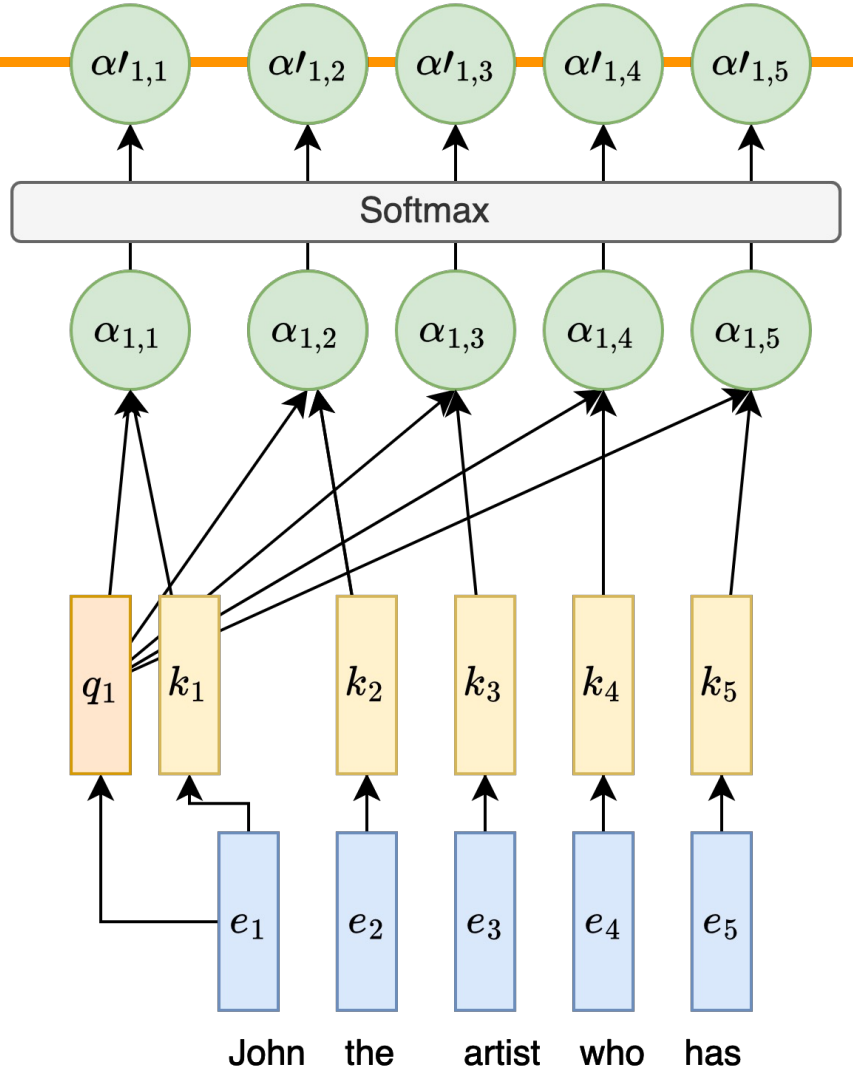
Attention Score

- The original Transformer uses scaled dot product (there are many others).
  - $score(e_i, e_j) = \frac{q_i k_j}{\sqrt{d_k}}$
  - d_k (= dim of keys = dim of queries)
  - Why Scaling?
    - Because greater d_k with Softmax() results in vanishing gradients.
    - By initializing all queries and keys to have $\mu = 0, \sigma = 1$, we can ensure $score(e_i, e_j)$ to be numerically stable, as long as we scaled by $\sqrt{d_k}$.



Attention Score

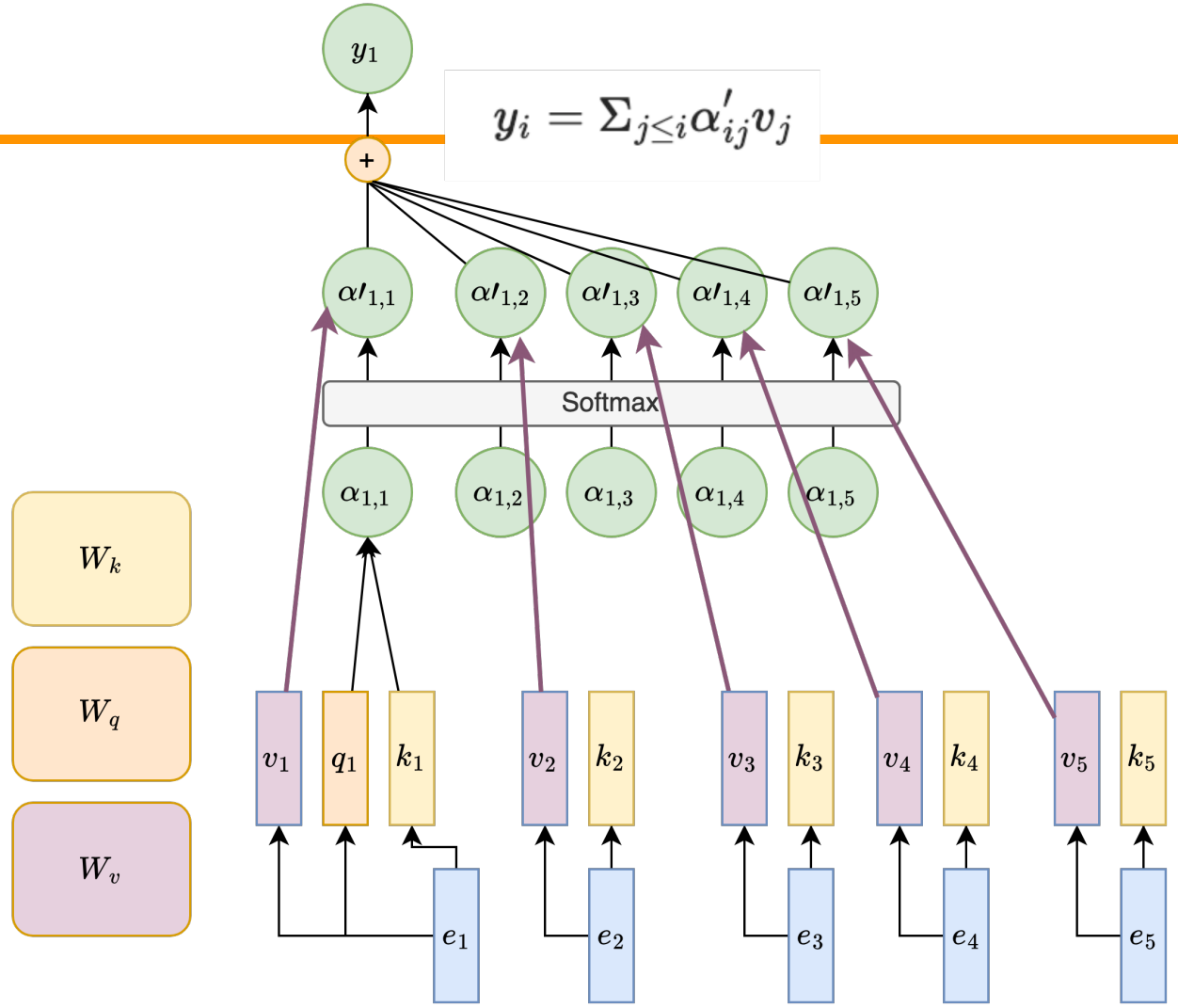
- For every pair of i, j, We calculate the scores. The figure illustrates q = 1.



$$\alpha'_{1,i} = \frac{\exp(\alpha_{1,i})}{\sum_j \exp(\alpha_{1,j})}$$

Self-Attention

- The figure illustrates when q = 1.
- The → indicates scalar vector multiplication.



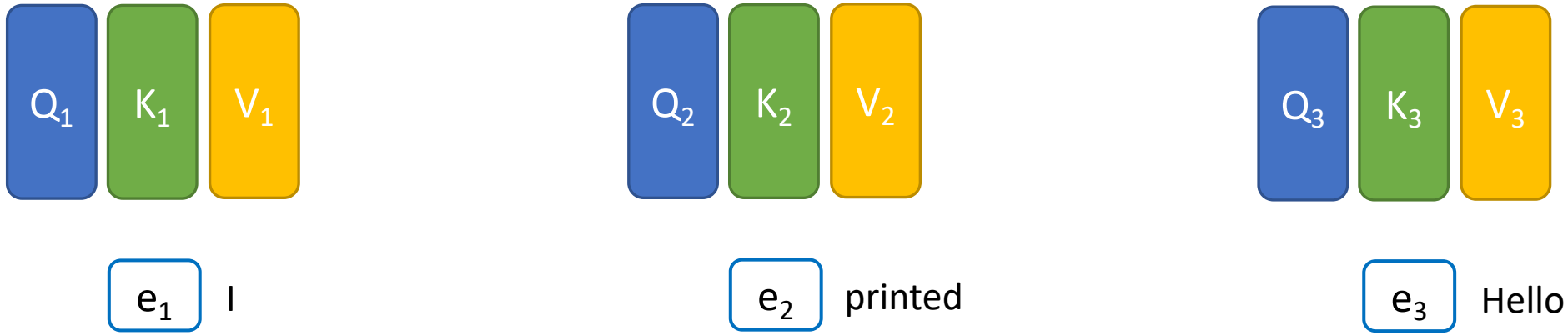
Multi-Head Attention (MHA)

- In p.9, we describe self-attention as QKV attention, where
  - we query the search engine
  - search engine tries to map our query to the keys
  - output some best matched videos (values)
- What if we want to have queries focusing on different aspects?
  - For example, one set of queries focusing on semantic similarity, another set focusing on passive/active voice.
  - **We need multiple attention mechanisms, i.e. multiple (attention) heads!**

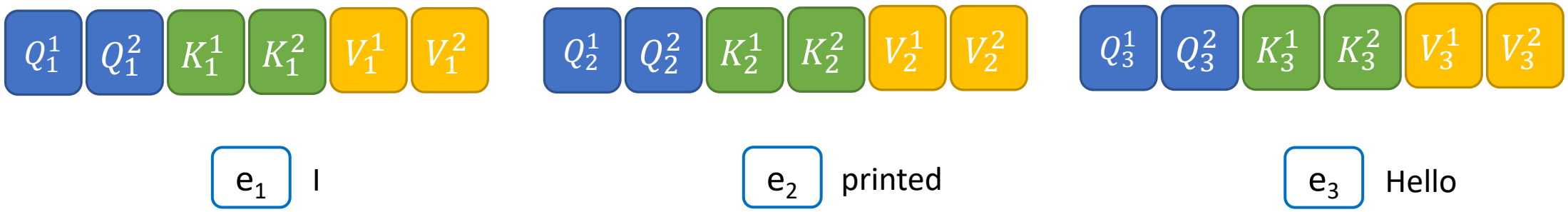
Multi-Head Attention (MHA)

上標: head
下標: token index

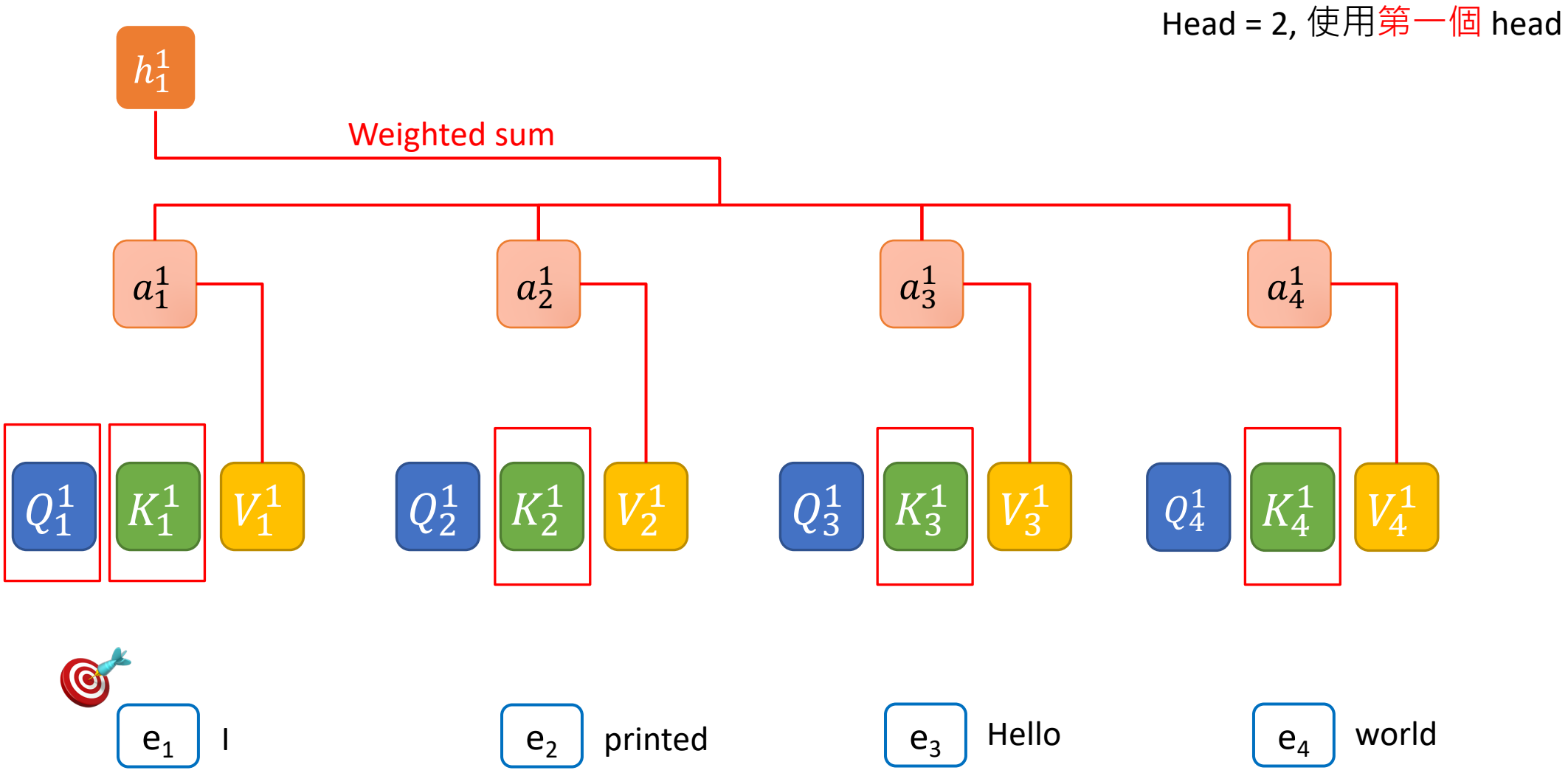
Head = 1 (without MHA)



Head = 2 (with MHA)

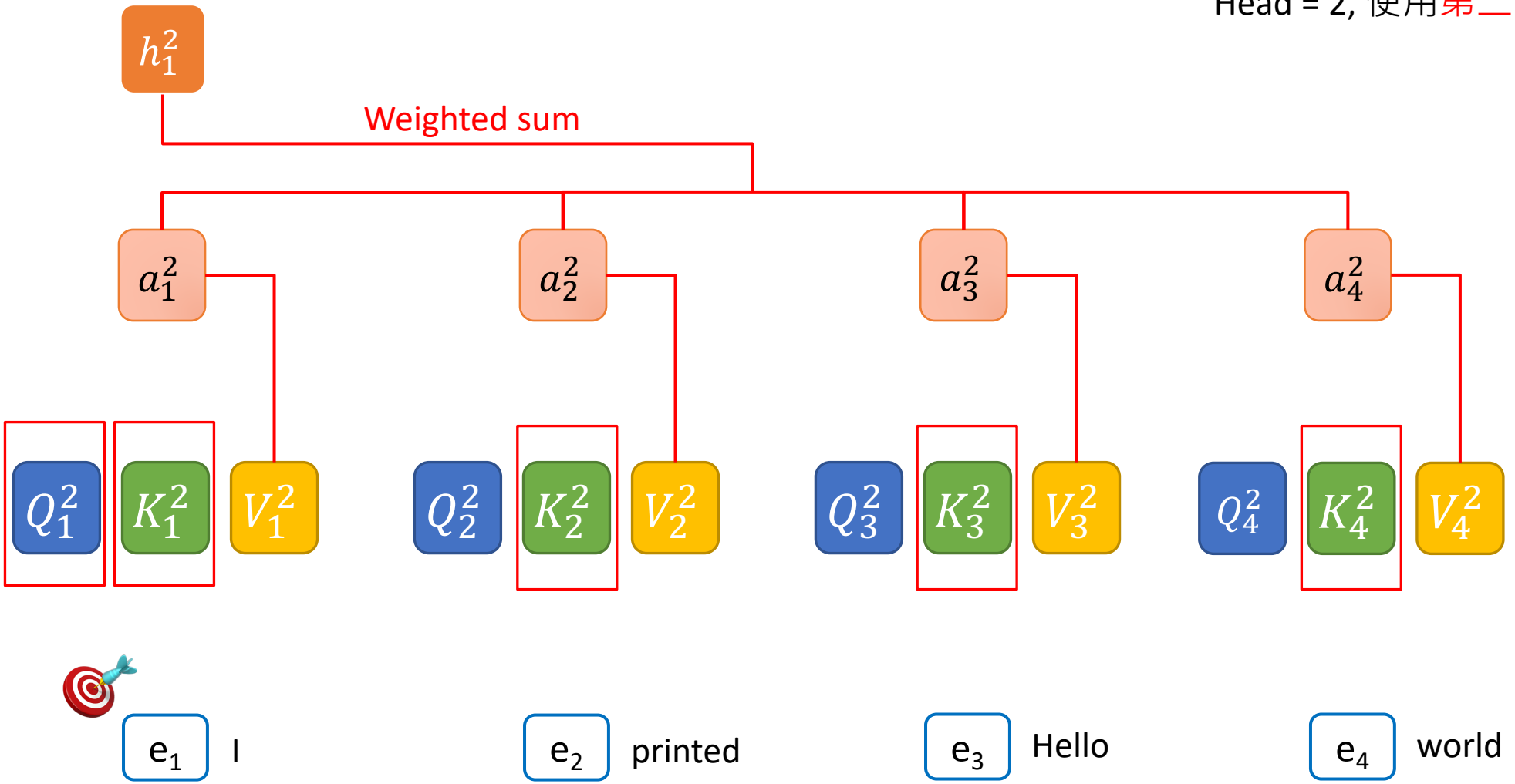


Query (Q), Key (K), and Value (V) with MHA



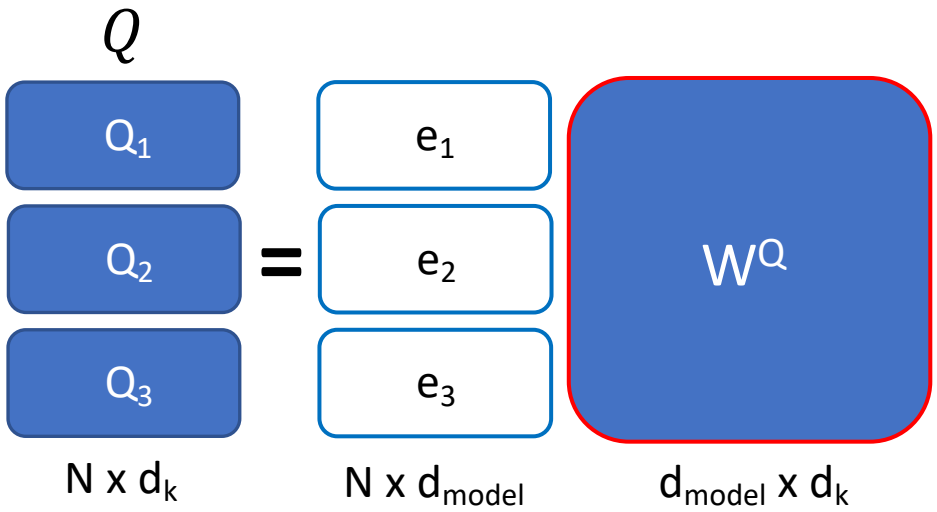
Query (Q), Key (K), and Value (V) with MHA

Head = 2, 使用第二個 head

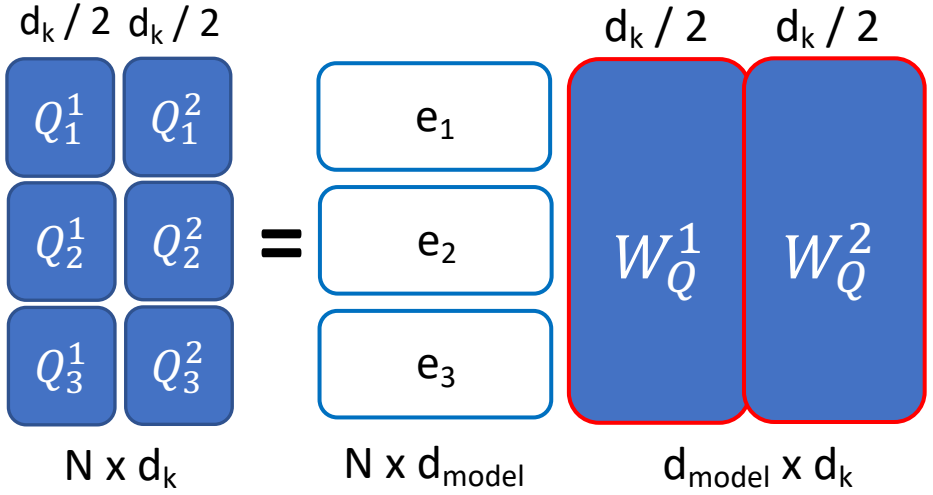


Input Dimensions of MHA

Head = 1 (without MHA)



Head = 2 (with MHA)

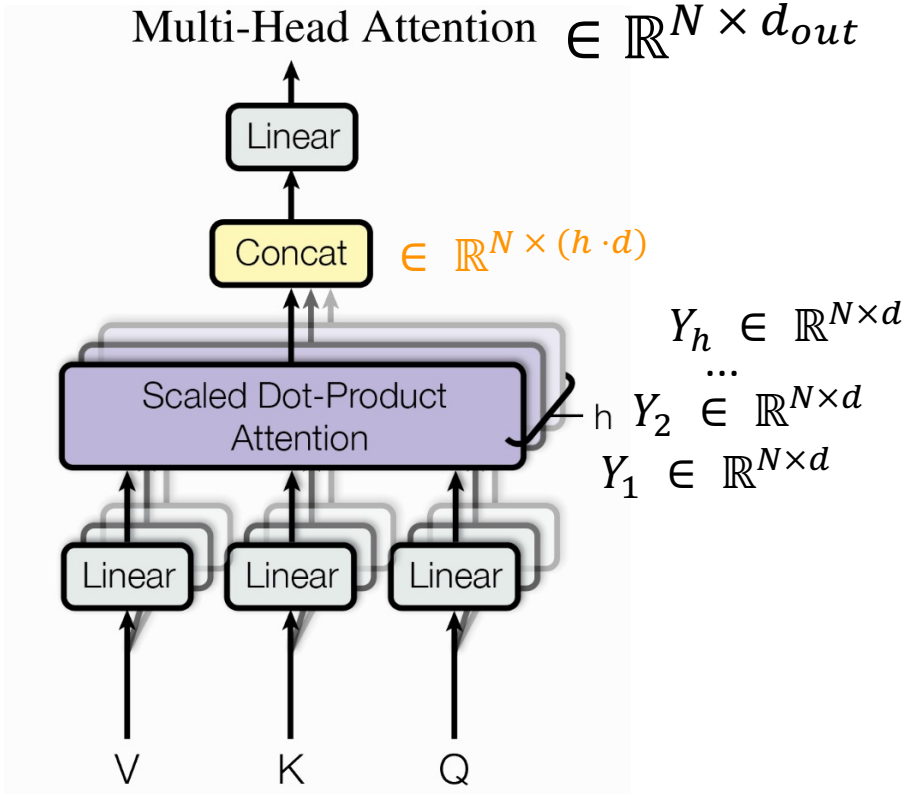


Output Dimensions

$$\text{Multihead}(Q, K, V) = \underbrace{\text{Concat}(\text{head}_1, \ldots, \text{head}_h)}_{\in \mathbb{R}^{N \times (h \cdot d)}} \underbrace{W^O}_{\in \mathbb{R}^{(h \cdot d) \times d_{out}}}$$

Head = 1 (without MHA)

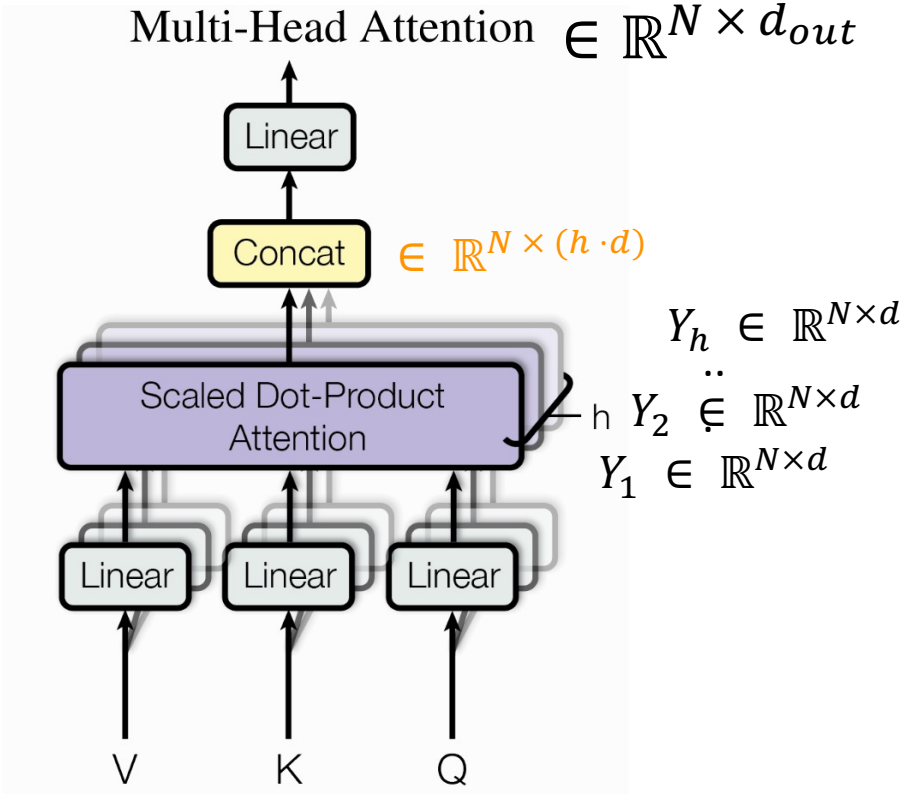
Head = 2 (with MHA)



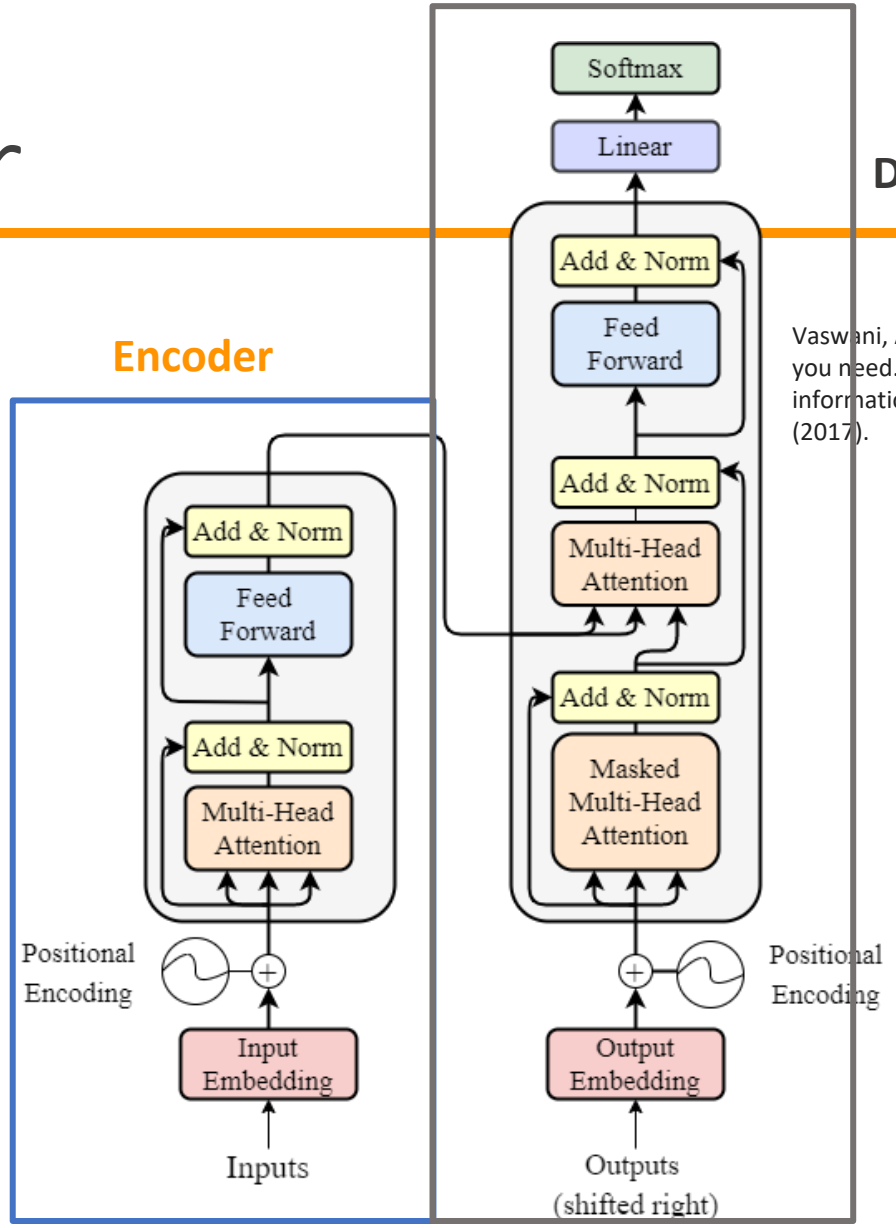
https://aclanthology.org/P19-1580/

Summary of MHA

- With W^O, the concatenated head outputs can be projected to a preferred dimension (hyperparameter: d_{out}).
- No worries on how #head changes dimension of outputs!



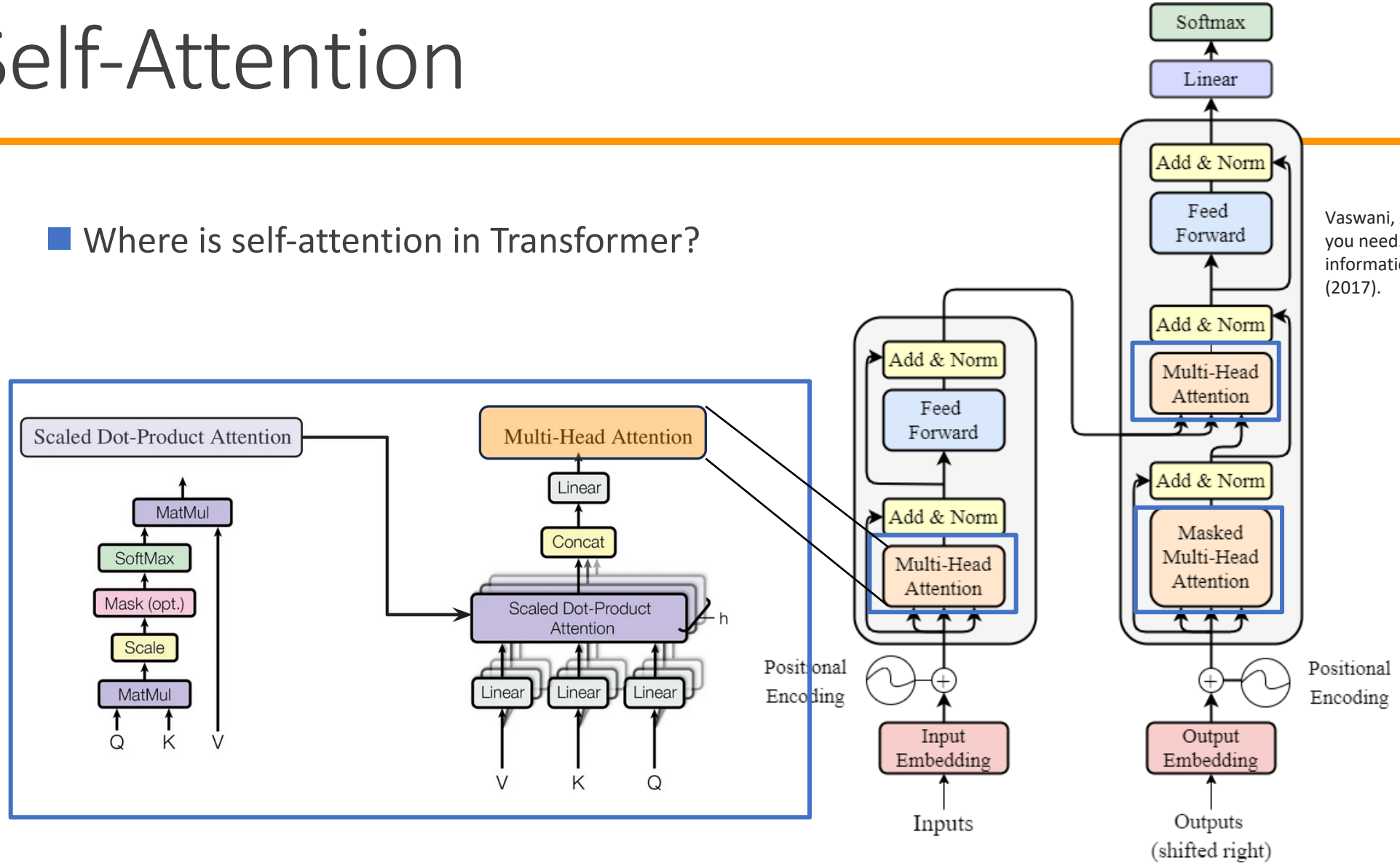
Transformer Encoder-Decoder



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Self-Attention

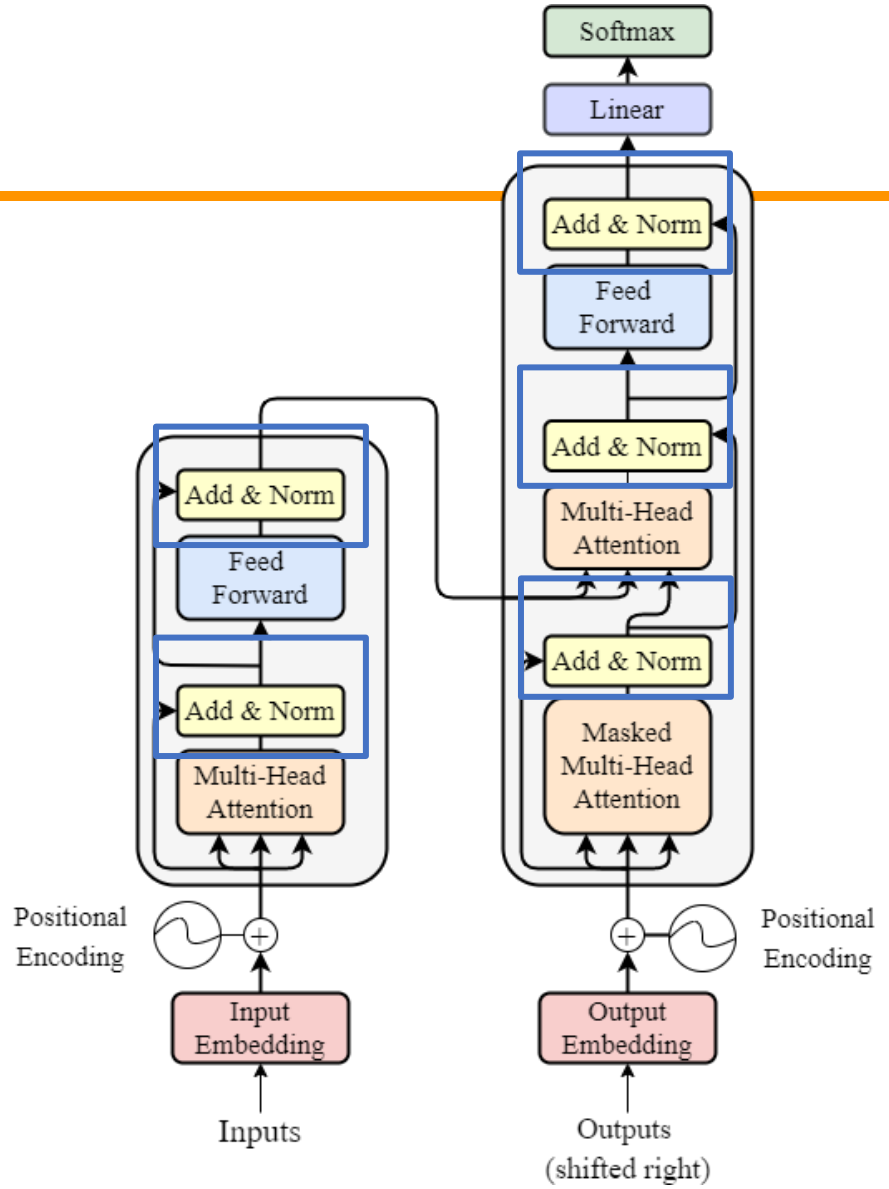
- Where is self-attention in Transformer?



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Add & Norm

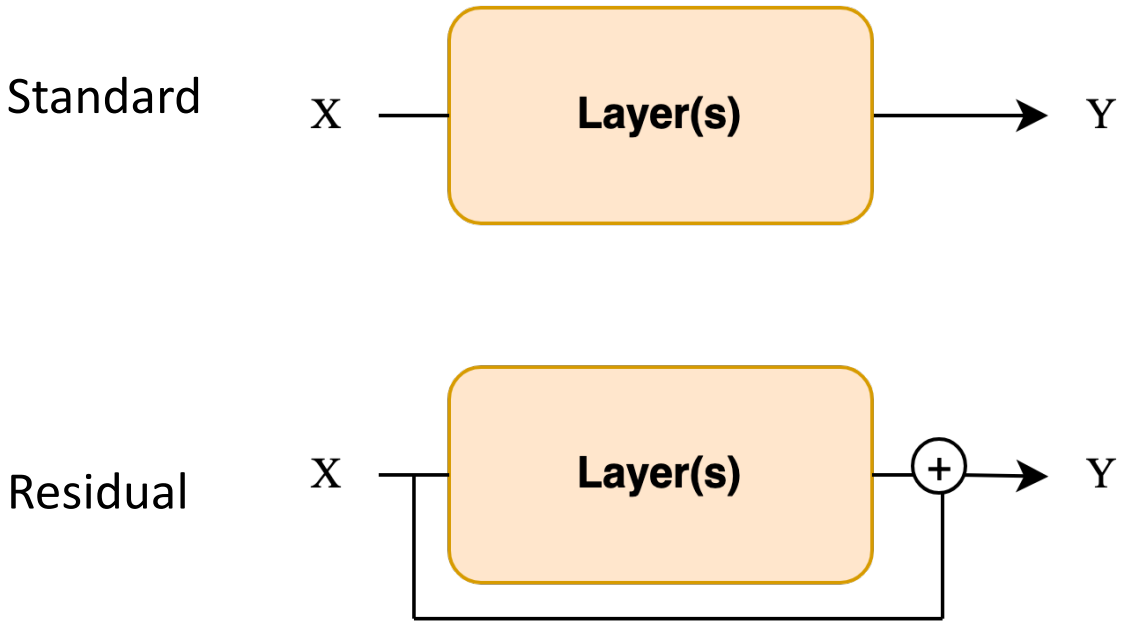
- Add: Residual Connection
- Norm: Layer Normalization



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Add & Norm

- Research shows that Residual Connection (He, Kaiming, et al., 2016) stabilizes training.
- Let the layers in between learn the residual (i.e. $Y - X$).



He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition (2016).

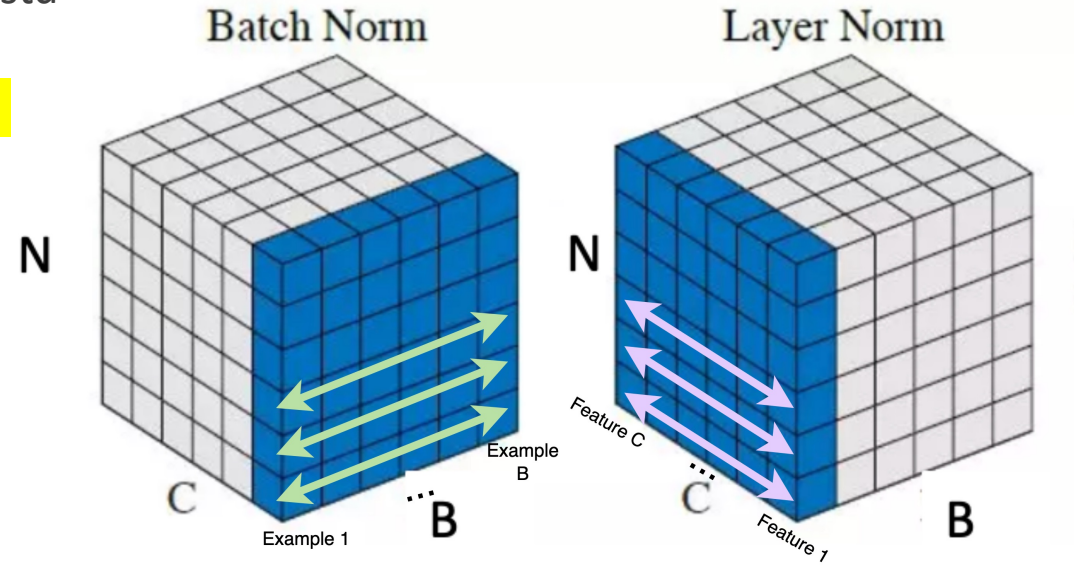
Add & Norm

B: Batch Size
C: Channels / Embedding Feature Dimension
N: Sequence length

- Normalization is also a way to stabilize training.
  - Batch Norm: For the same embedding dimension in the same sequence index, calculate its mean μ and std σ across the samples within the same batch.
  - Layer Norm: For an input vector, calculate its mean and std across embedding dimension.
  - $\mathbf{norm}(\boldsymbol{X}) = \frac{\boldsymbol{X}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}$
- Vaswani et al. (2017) used Layer Normalization.

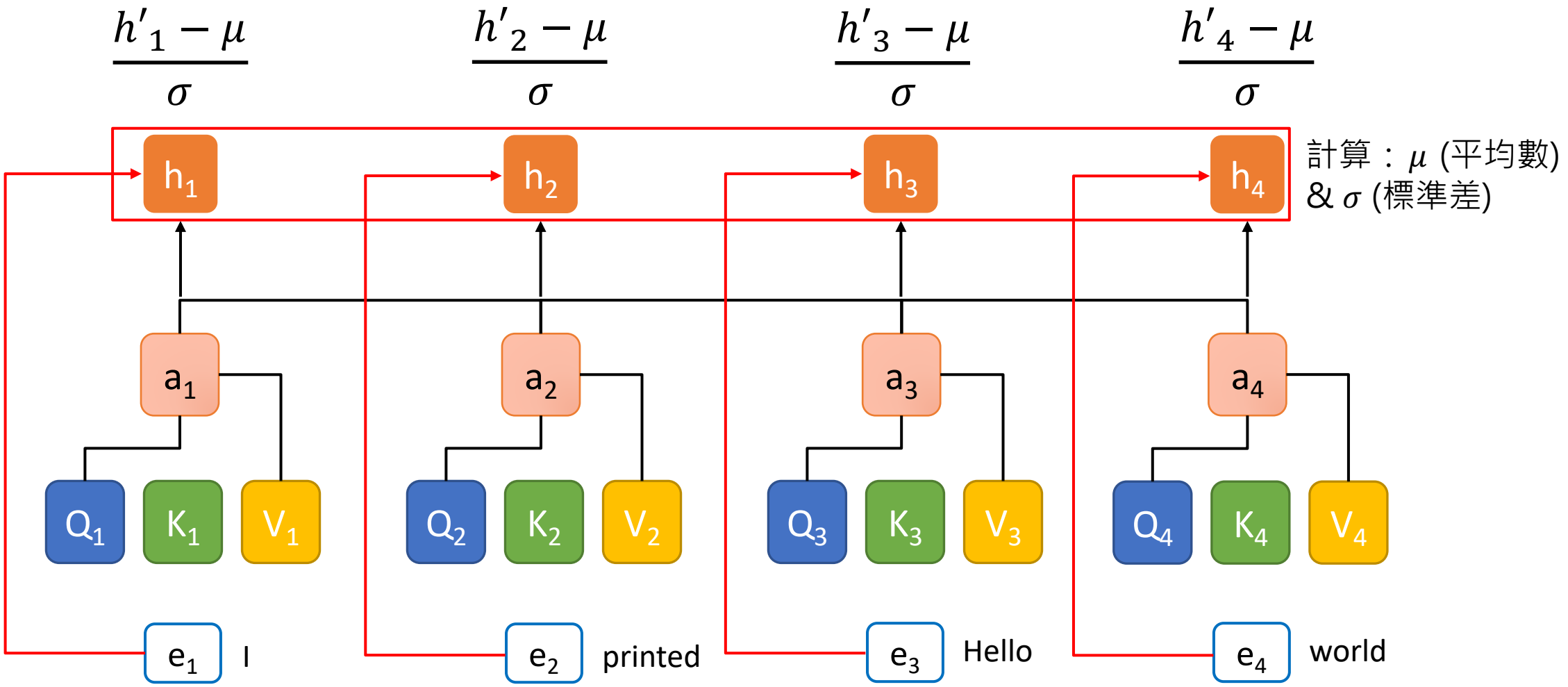
Why LN?

- No specific exp. from Vaswani et al., 2017.
- BN needs batch mean and std during inference, unsuitable for varying seq length in NLP.
- LN yields good results.



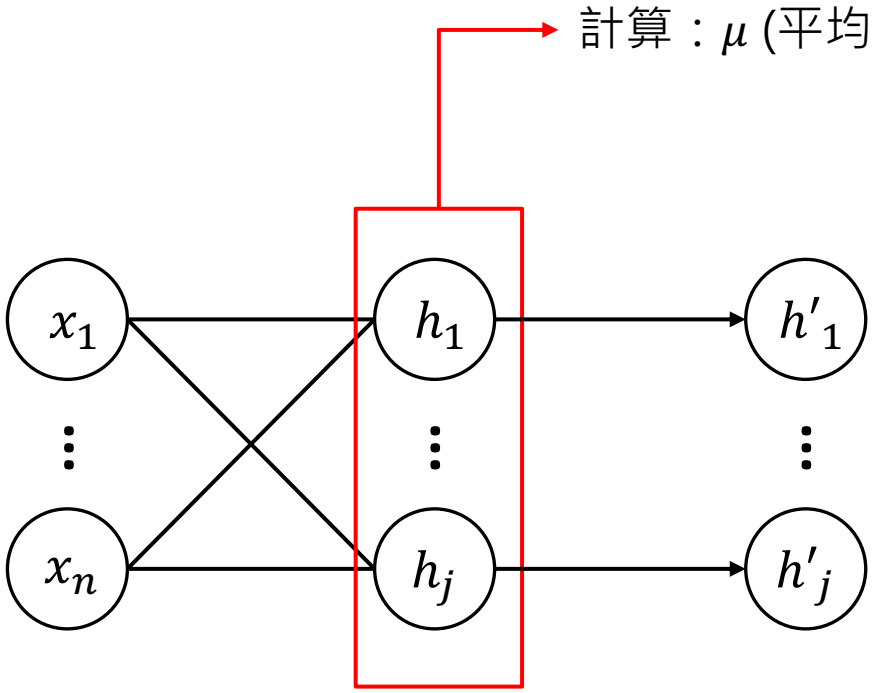
Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton. "Layer normalization." arXiv preprint arXiv:1607.06450 (2016).

Add & Norm



Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton. "Layer normalization." arXiv preprint arXiv:1607.06450 (2016).

Batch Normalization



$$h'_1 = \frac{h_1 - \mu}{\sigma}$$

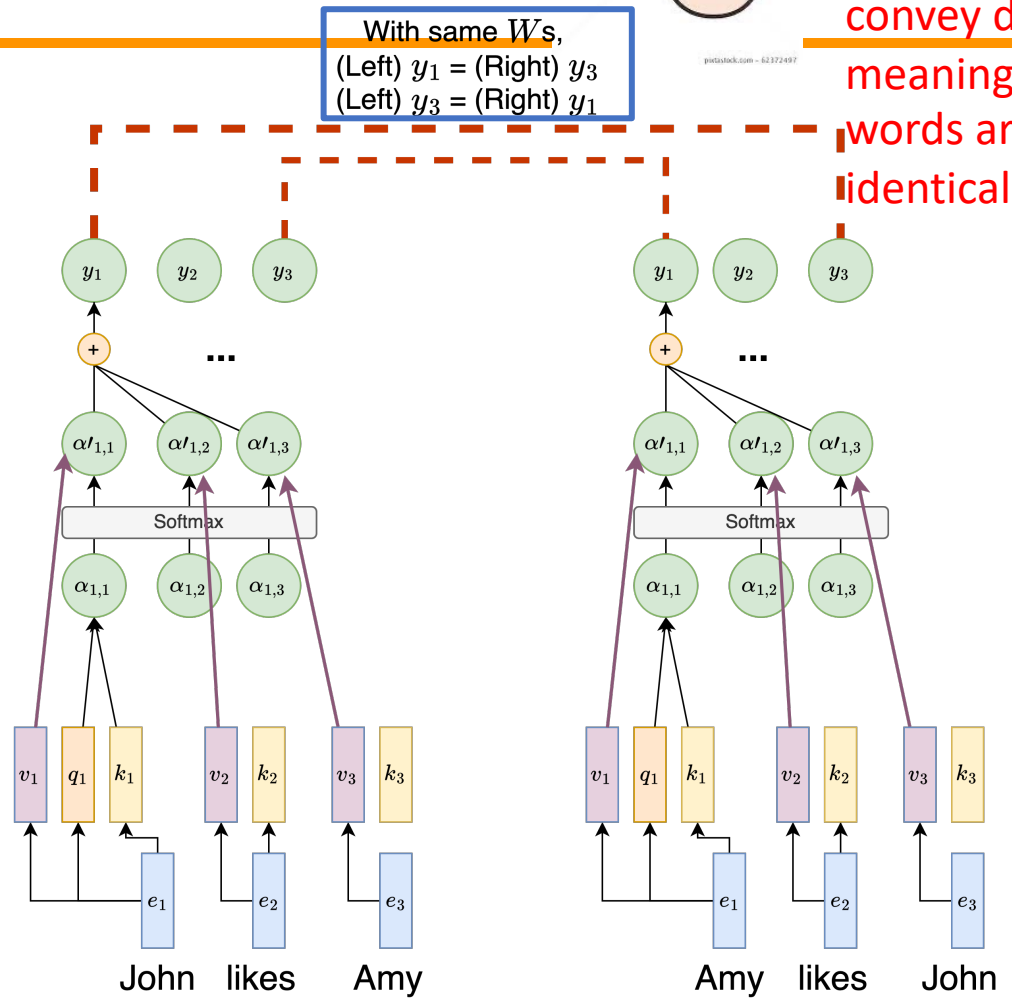
$$h'_j = \frac{h_j - \mu}{\sigma}$$

Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International conference on machine learning. 2015.

Lack of Position Modeling

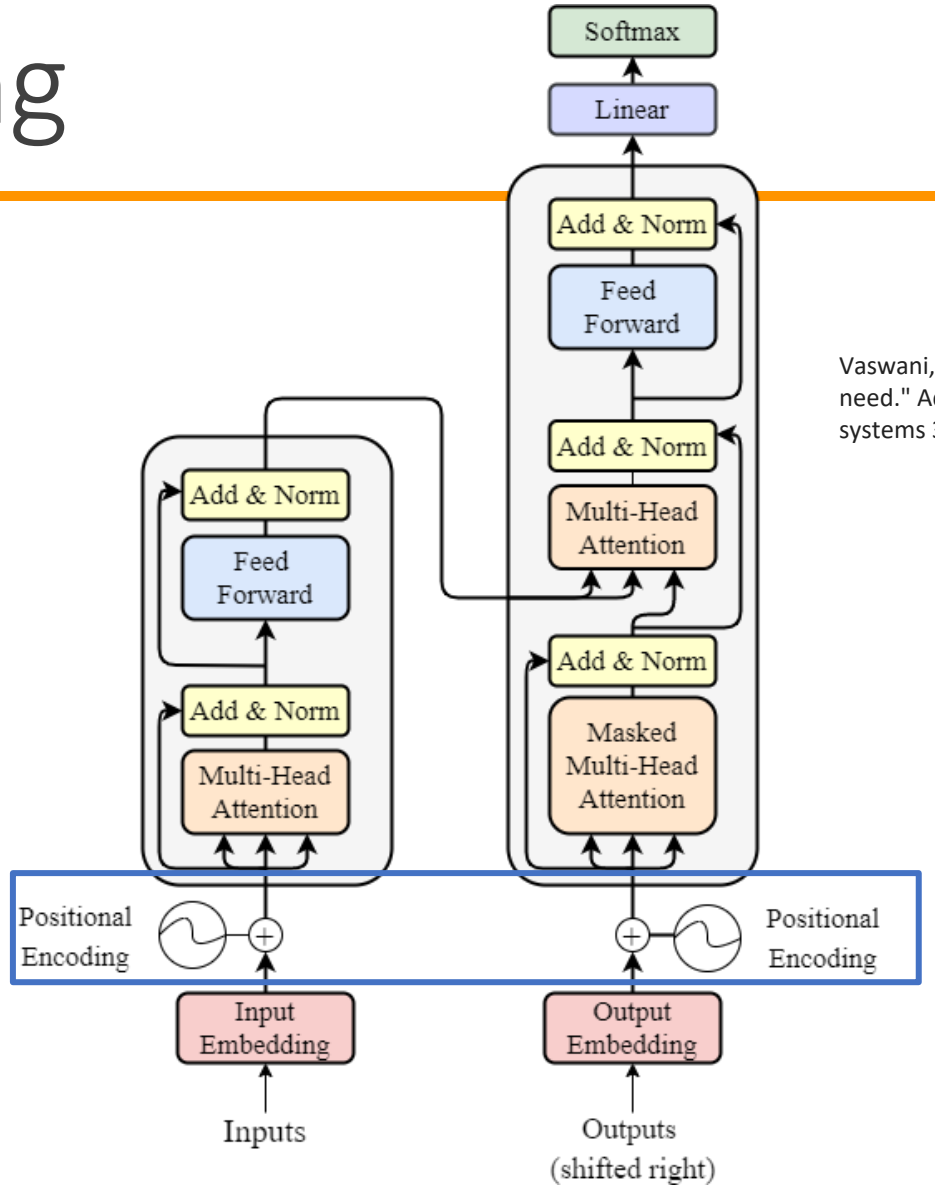
Problematic because these 2 sentences convey different meanings, but same words are encoded identically!

- Recall the 2 issues of RNNs:
  - Linear interaction distance
    - [solved] by directly learning attention weight instead of decaying the degree by distance.
  - Lack of parallelizability
    - [solved] by using matrix multiplication.
- But ... We have no way to know the relative distance of one word to another!
  - The same word, eg. “Amy”, in the 2 sentences are encoded to the same y (with the same weight matrices).



Positional Encoding

- Where is the Positional Encoding inside the transformer?
- In brief, it is a way to tell the model the position of each token in a sequence.



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Positional Encoding Equations

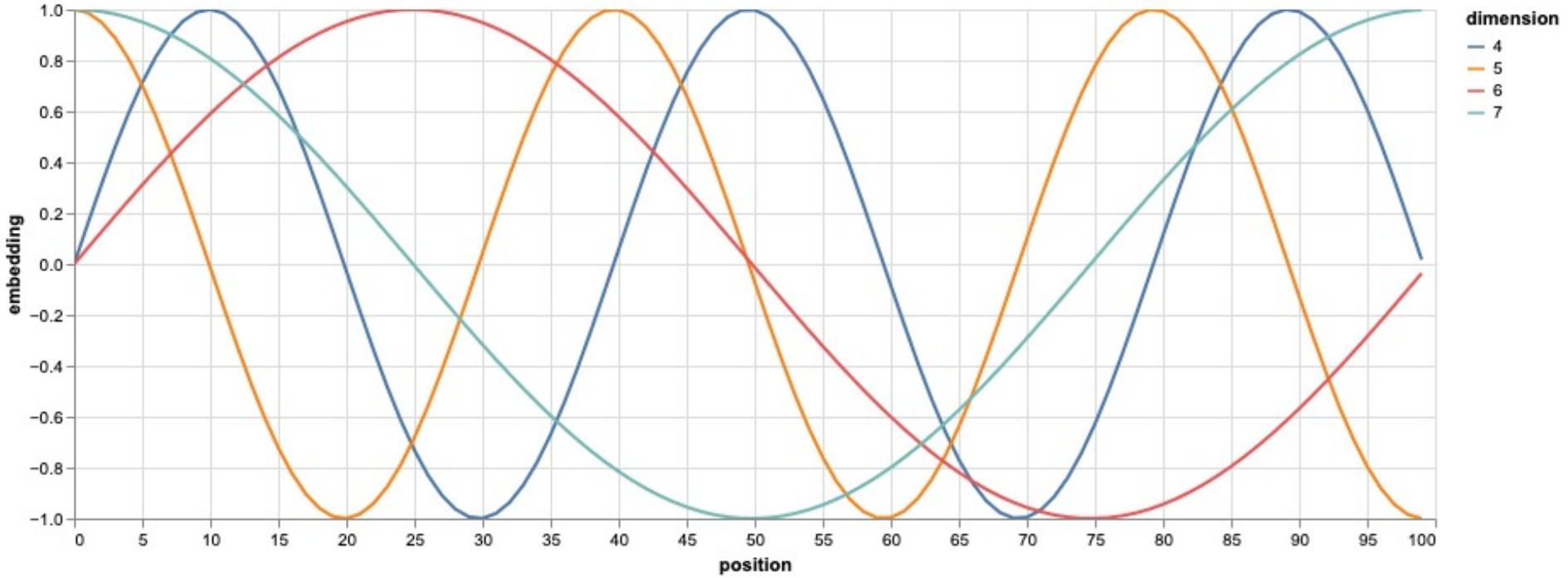
- 創造出相對位置的 embeddings
- 核心精神：利用三角函數的週期性來建構序列資訊
  - 偶數的 embedding 單元 (2i , i = 0, 1, …, d_{model})：Sin 函數
  - 奇數的 embedding 單元 (2i+1, i = 0, 1, …, d_{model})： Cosine 函數

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

輸入序列中任一token的位置

Figure source: https://nlp.seas.harvard.edu/annotated-transformer/#positional-encoding

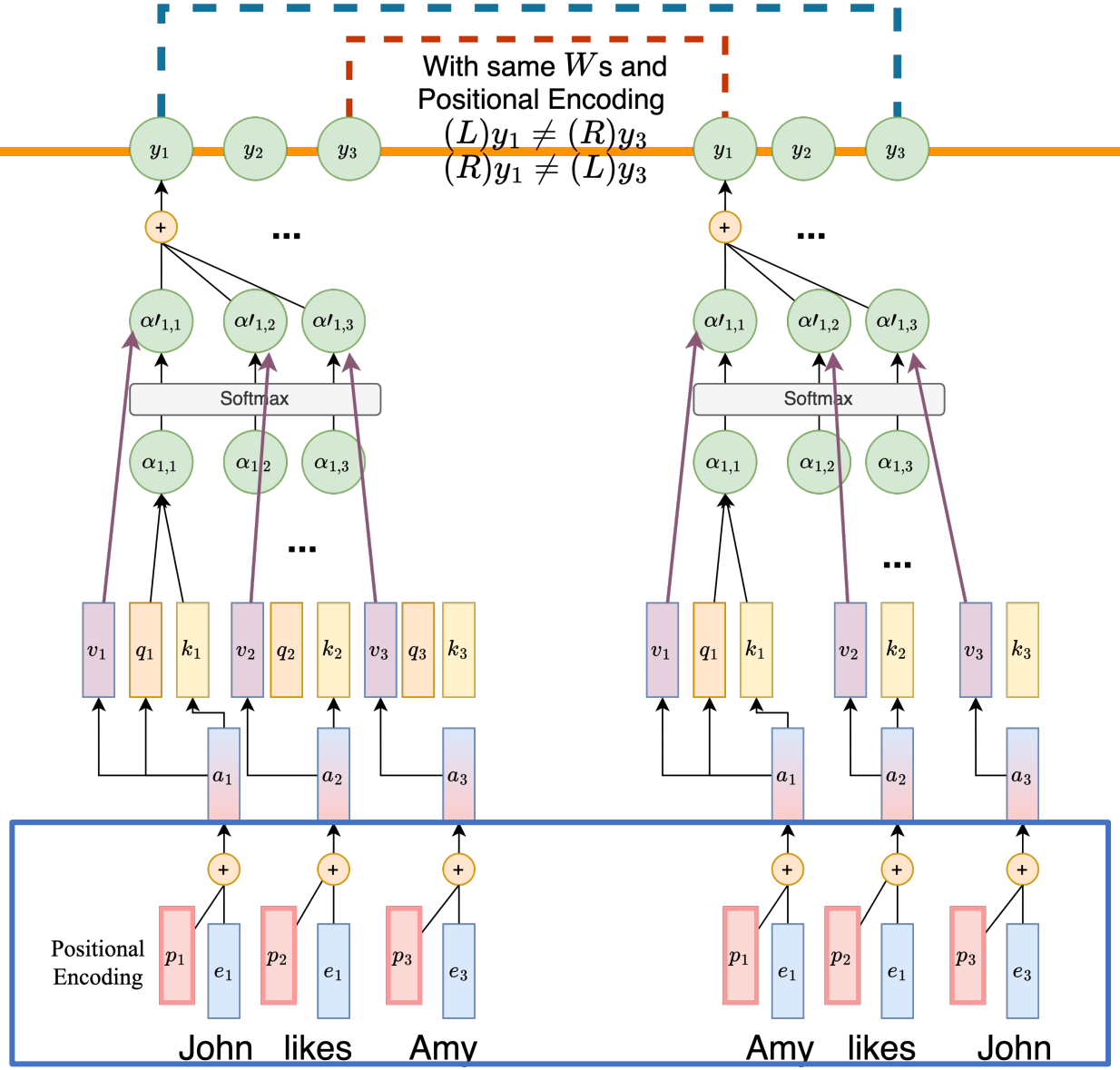
Positional Encoding Visualization



- Question: 為什麼需要同時有 Sin 和 Cosine 的函數？

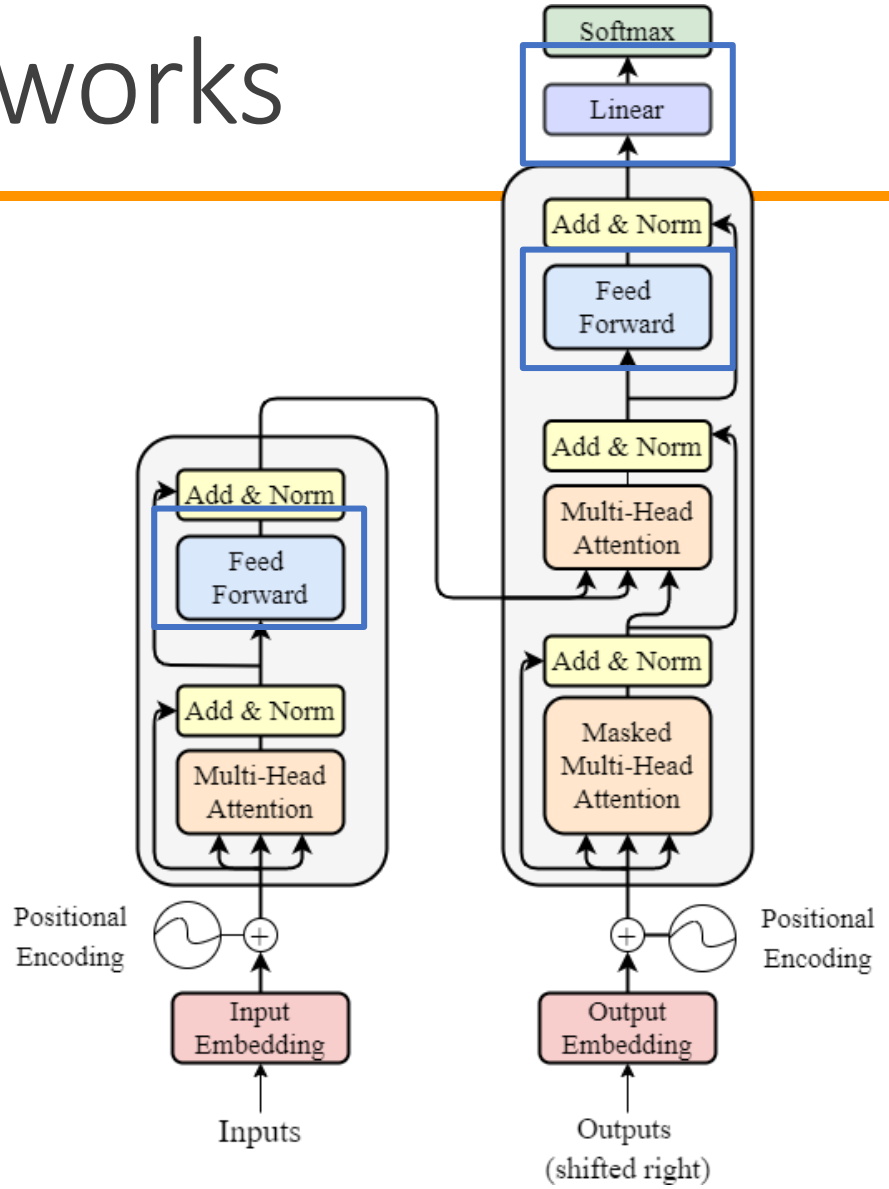
Figure source: https://nlp.seas.harvard.edu/annotated-transformer/#positional-encoding

Positional Encoding



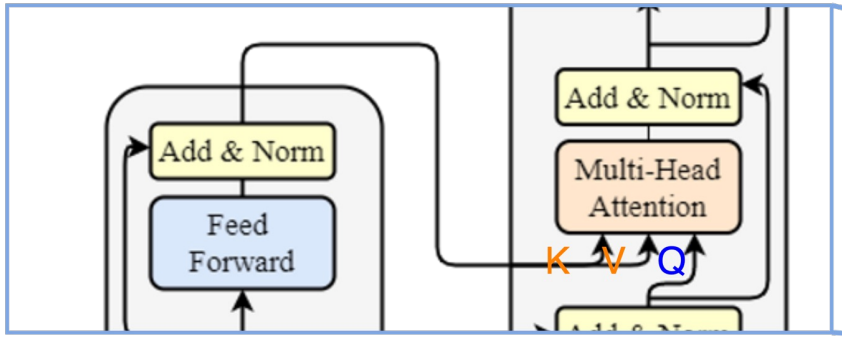
Feed Forward Networks

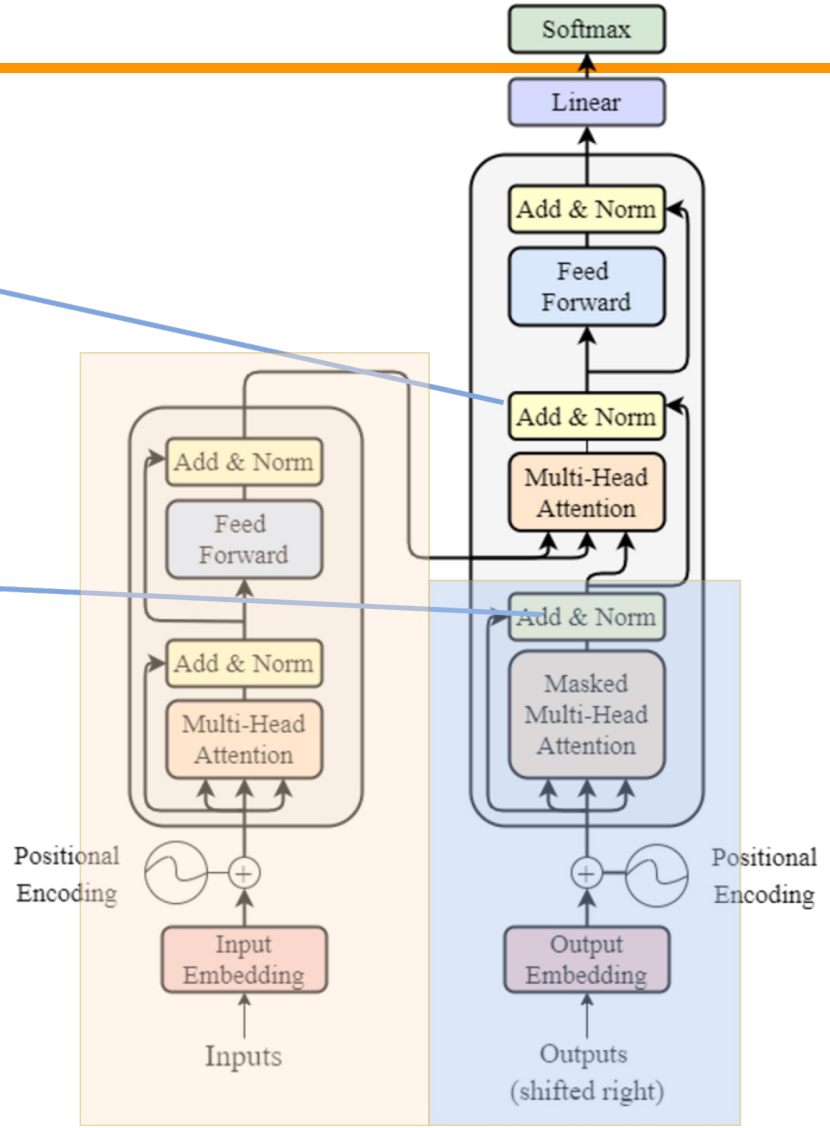
- Simply multiply by a weight matrix.
- Used to project to a specific dimension.



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Decoder: Cross-Attention





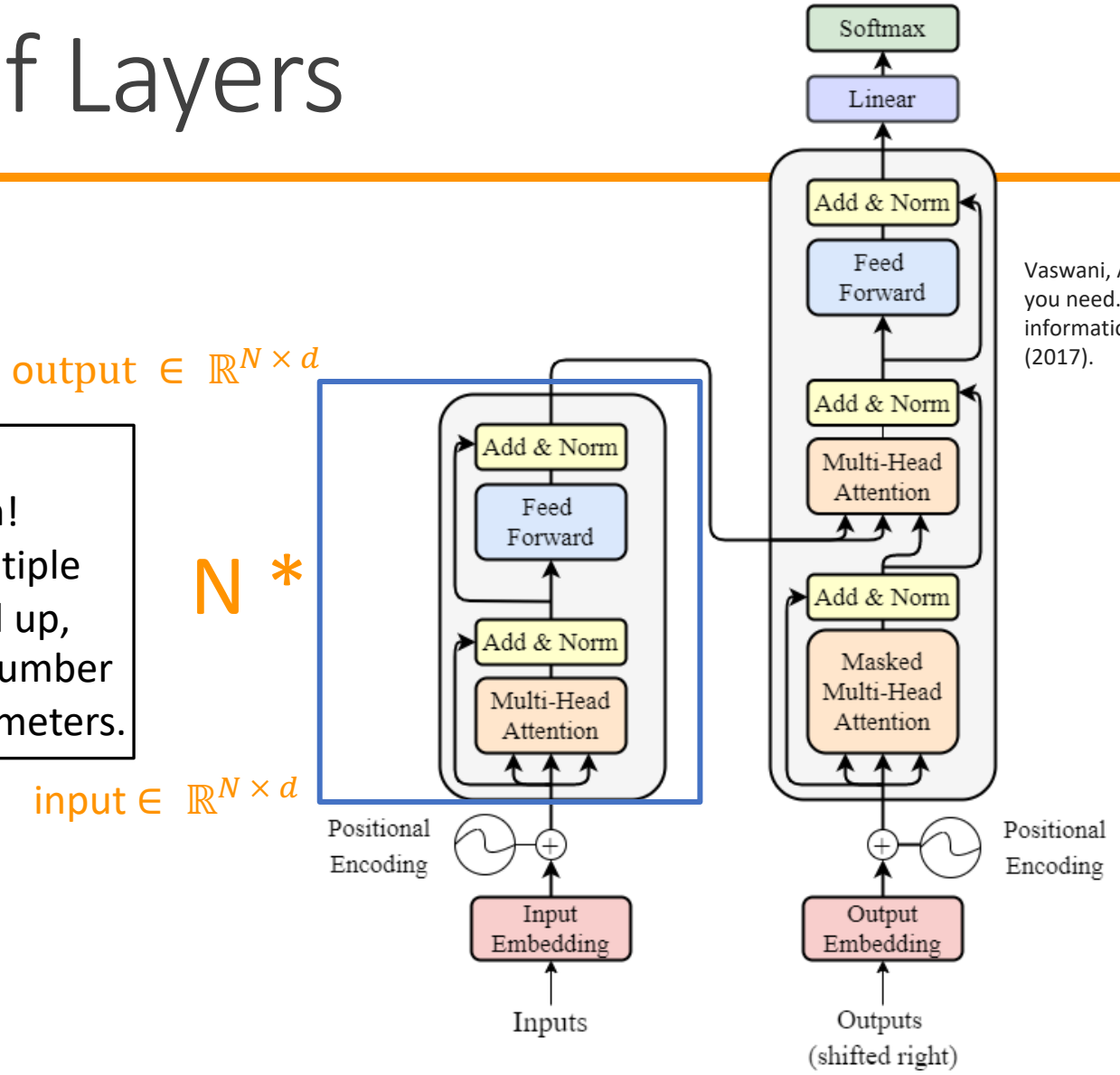
K, V uses the output from encoder as their X;

Q uses the decoder's information as X.

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

Number of Layers

Input, output dimension match!
We can have multiple encoders stacked up, to increase the number of trainable parameters.



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformers' Achievements

1. In original paper, State-of-The-Art (SoTA) of machine translation.

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

6 encoders,
6 decoders,
h = 512, 100K train steps.

6 encoders,
6 decoders,
h = 1024, 300K train steps.

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
https://arxiv.org/abs/1706.03762

Transformers' Achievements

- (Chat)GPT is just a pretrained transformer (it is just n layers of transformer decoders).
- Pretraining is similar to let the model read through a lot of books, and get a sense of the context word distribution so that it can infer the meaning.
  - For more on linguistic theory, check "distributional hypothesis."
  - For more on pretraining, wait till next week (W7) 's topic of BERT and its Family!

Source: https://stock.adobe.com/tw/search?k=robot+reading&asset_id=570899814

Transformer Variants (there are many more)

- Universal Transformer
  - Adds Adaptive Computation Time
  - Dehghani, Mostafa, et al. "Universal transformers." arXiv preprint arXiv:1807.03819 (2018).
- Longformer
  - Tackles long documents
  - Iz Beltagy, et al. (2020). Longformer: The Long-Document Transformer. arXiv:2004.05150.
- Roformer
  - Changes the design of positional encoding
  - Su, Jianlin, et al. "RoFormer: Enhanced Transformer with Rotary Position Embedding. CoRR abs/2104.09864 (2021)." arXiv preprint arXiv:2104.09864 (2021).
- Vision Transformer
  - Tackles vision tasks
  - Alexey Dosovitskiy, et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” International Conference on Learning Representations. 2021.

Additional Links

- https://nlp.seas.harvard.edu/annotated-transformer/
- https://datascience.stackexchange.com/questions/82451/why-is-10000-used-as-the-denominator-in-positional-encodings-in-the-transformer

Thank you!

Instructor: 林英嘉

✉ yjlin@cgu.edu.tw

TA: 吳宣毅

✉ m1161007@cgu.edu.tw