Bert for Gendered Pronoun Resolution

Gong Qiong
2019.06.05

Outline

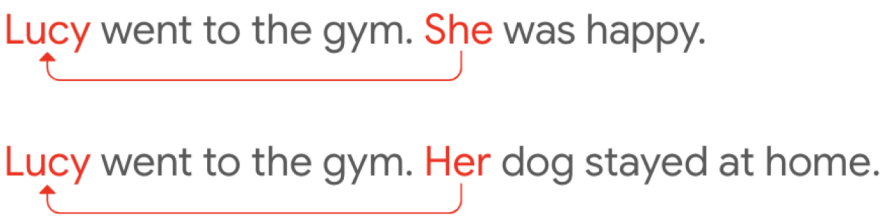
- Problem
- Self-Attention
- Transformer
- BERT
- Fine-tune BERT

Outline

- Problem
- Self-Attention
- Transformer
- BERT
- Fine-tune BERT

Problem

- Coreference Resolution
  - Coreference Resolution is the task of identifying clusters of mentions referring to the same real-world entity
- Pronoun Resolution
  - task of identifying for a specified pronoun in a passage, which named entity antecedent it refers to.



Problem

- Example of Gendered Pronoun Resolution

  Kathleen Nott was born in Camberwell, London. Her father, Philip, was a lithographic printer, and her mother, Ellen, ran a boarding house in Brixton; Kathleen was their third daughter. ***She*** *was educated at Mary Datchelor Girls' School (now closed), London, before attending King's College, London.*

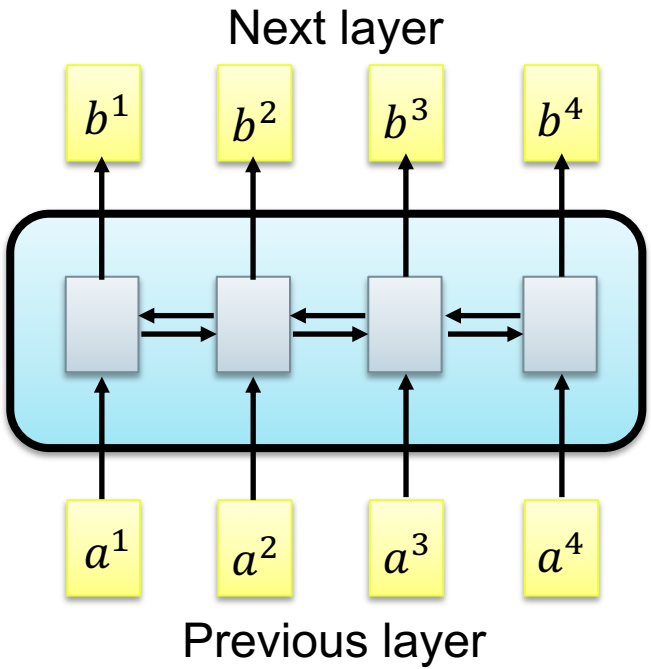
  Target pronoun: She,Candidate A: Ellen, Candidate B: Kathleen

  Ground Truth Class (one of A, B or NEITHER): Kathleen

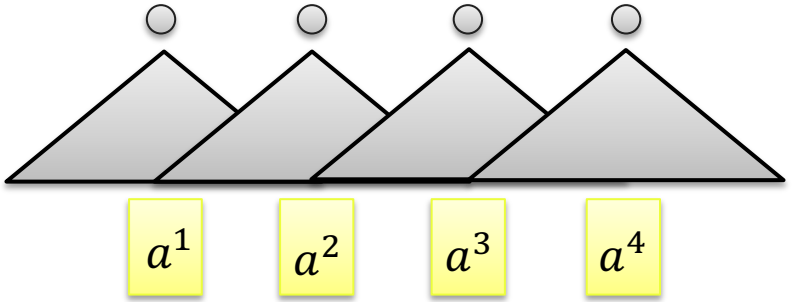
Outline

- Problem
- Self-attention
- Transformer
- BERT
- Fine-tune BERT

RNN for Sequence

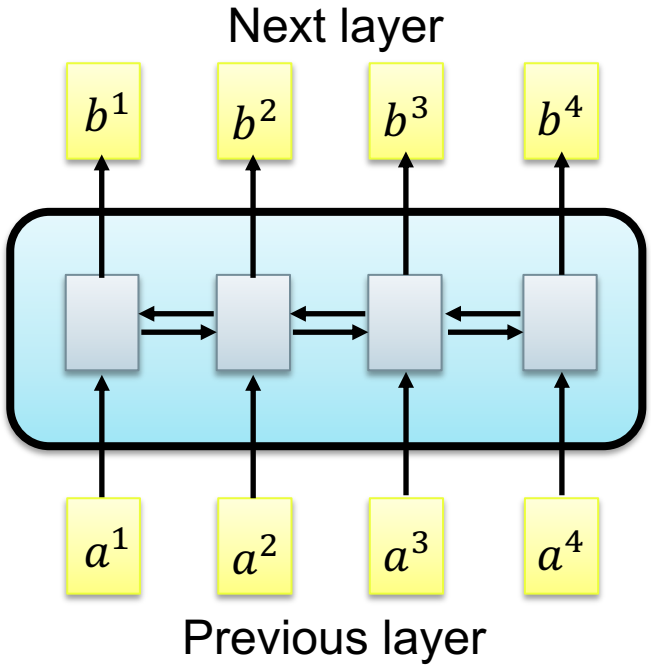


Hard to parallel !



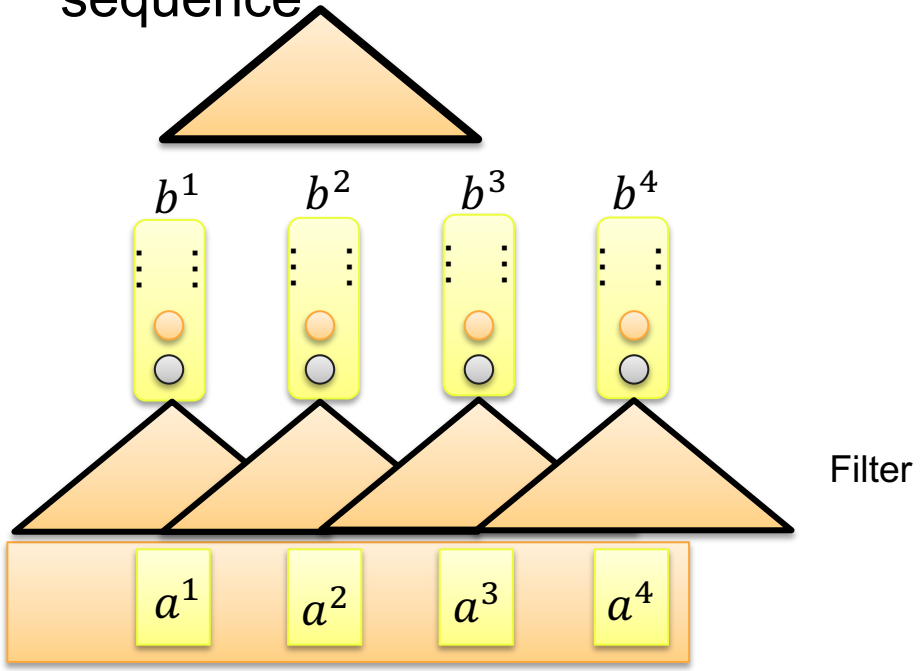
Using CNN to replace RNN

Sequence



Hard to parallel

Filters in higher layer can consider longer sequence

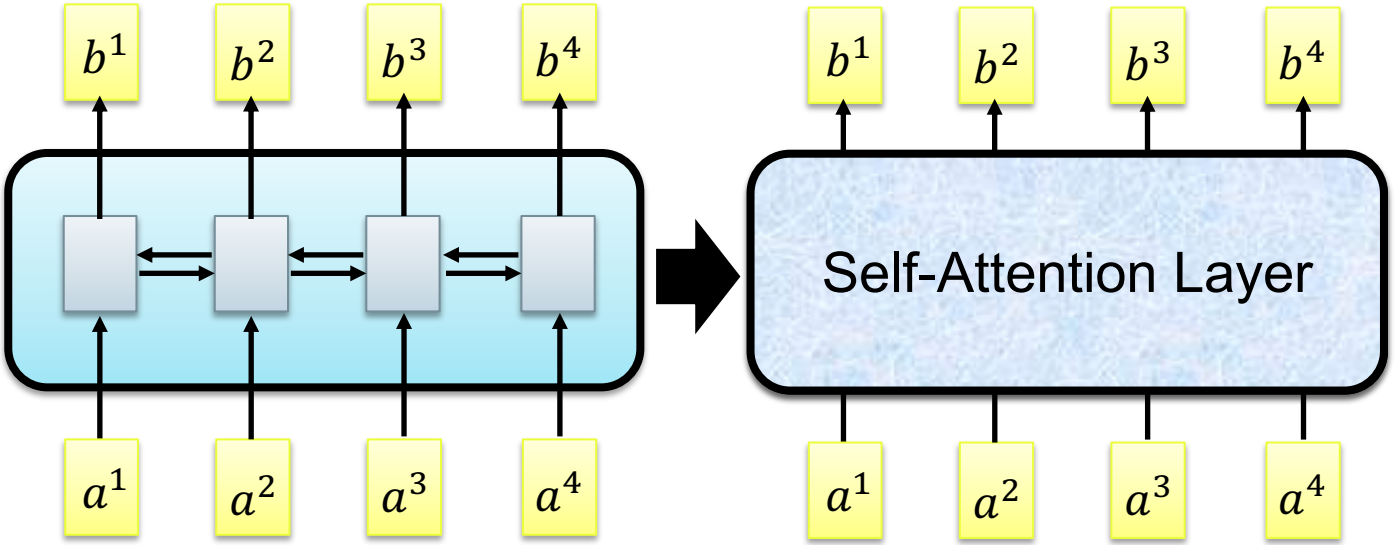


Using CNN to replace RNN (CNN can parallel)

Self-Attention

b^i is obtained based on the whole input sequence.

b^1, b^2, b^3, b^4 can be parallelly computed.



You can try to replace any thing that has been done by RNN with self-attention.

Self-Attention

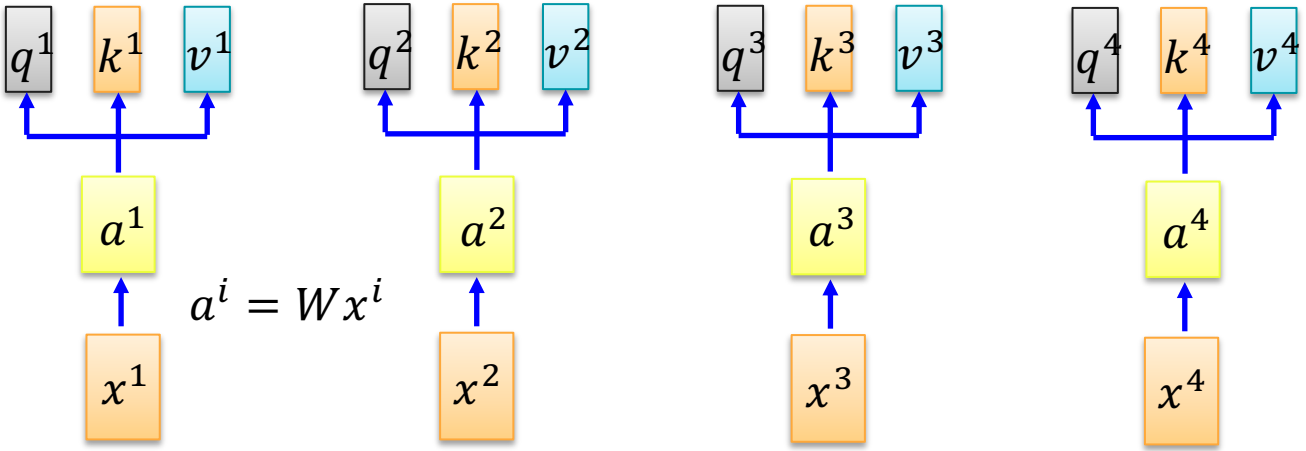
Attention is all you need.

https://arxiv.org/abs/1706.03762

q: query (to match others) $q^i = W^q a^i$

k: key (to be matched) $k^i = W^k a^i$

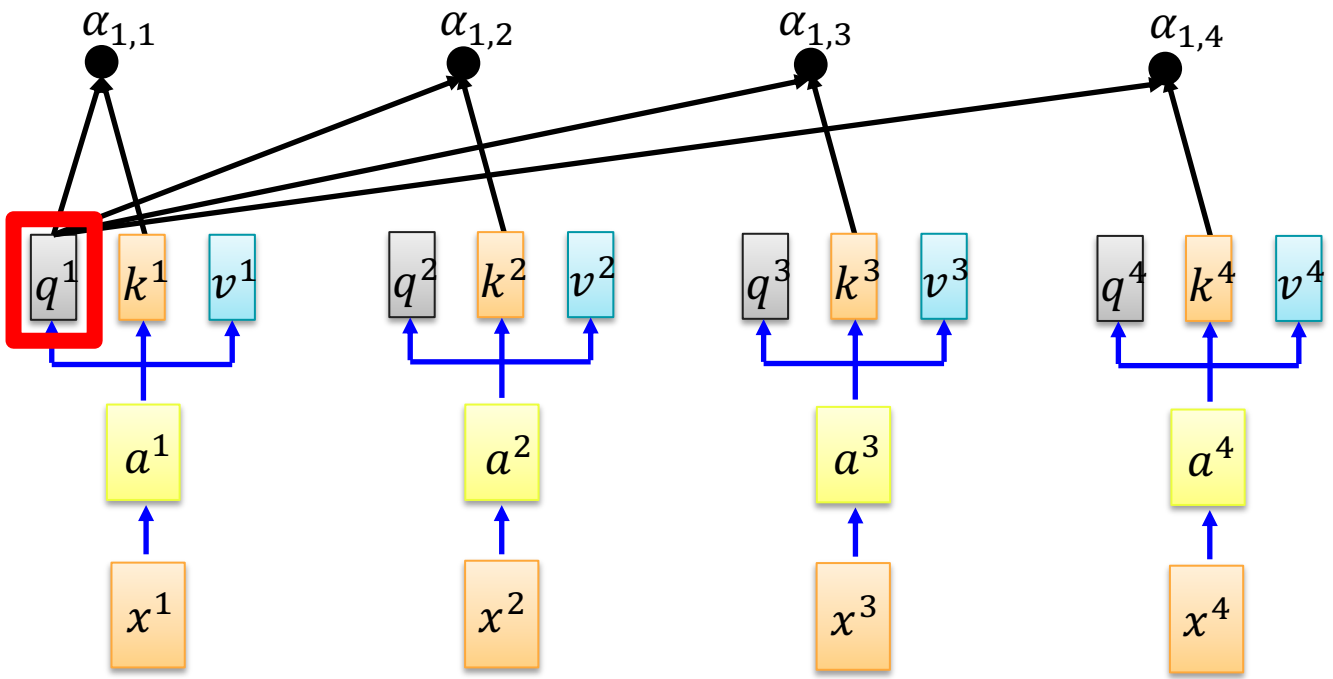
v: information to be extracted $v^i = W^v a^i$



Self-Attention

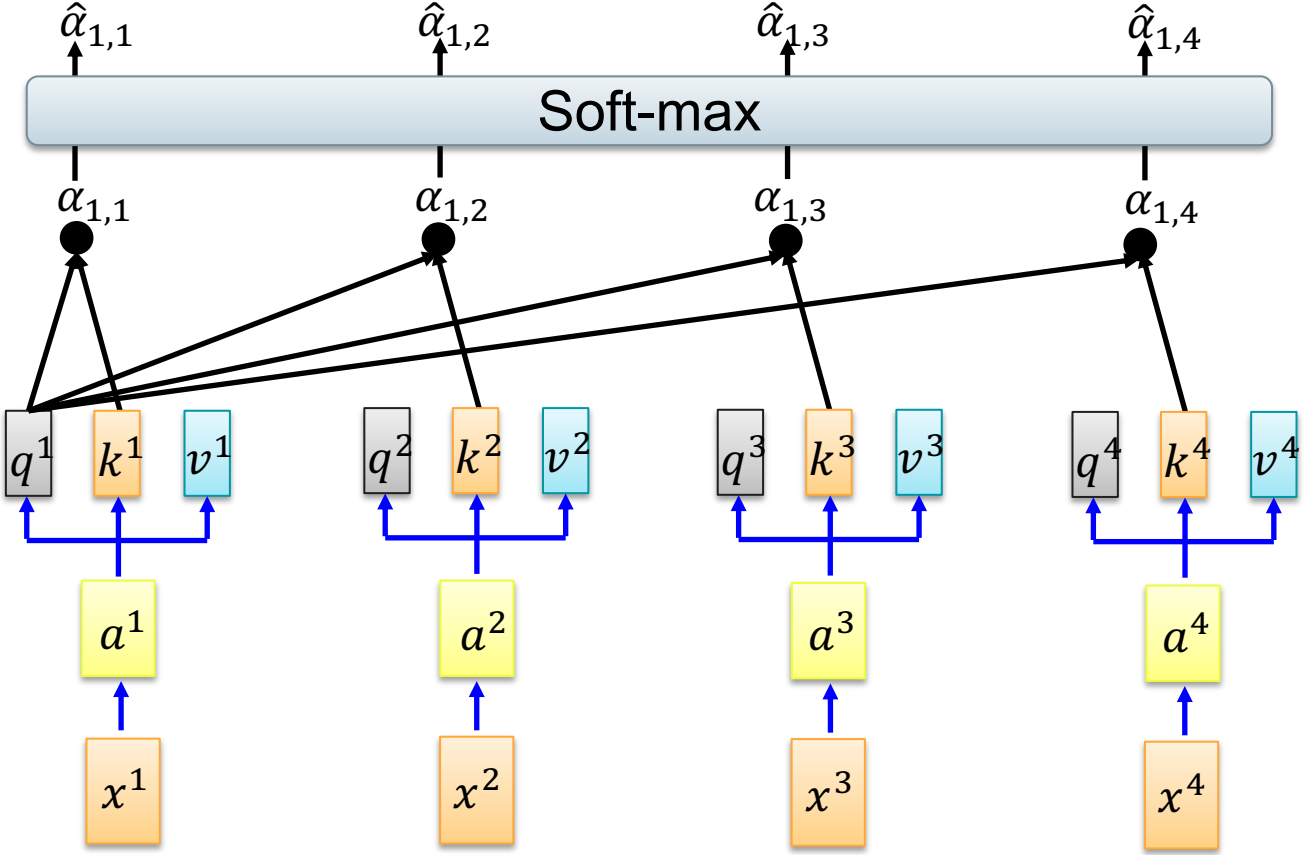
d is the dim of q and k

Scaled Dot-Product Attention: $\alpha_{1,i} = \underbrace{q^1 \cdot k^i}_{\text{dot product}} / \sqrt{d}$



Self-Attention

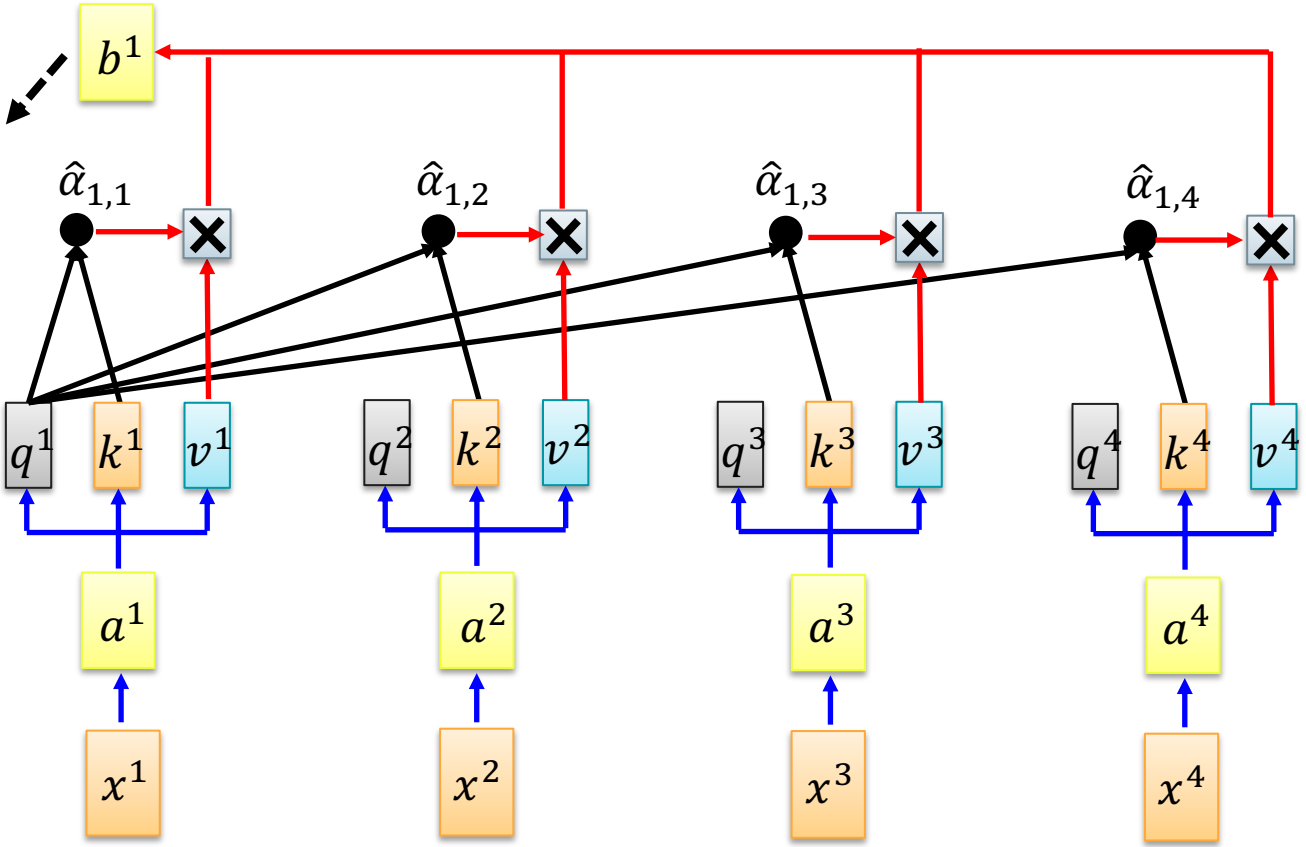
$$\hat{\alpha}_{1,i} = exp(\alpha_{1,i}) / \sum_j exp(\alpha_{1,j})$$



Self-Attention

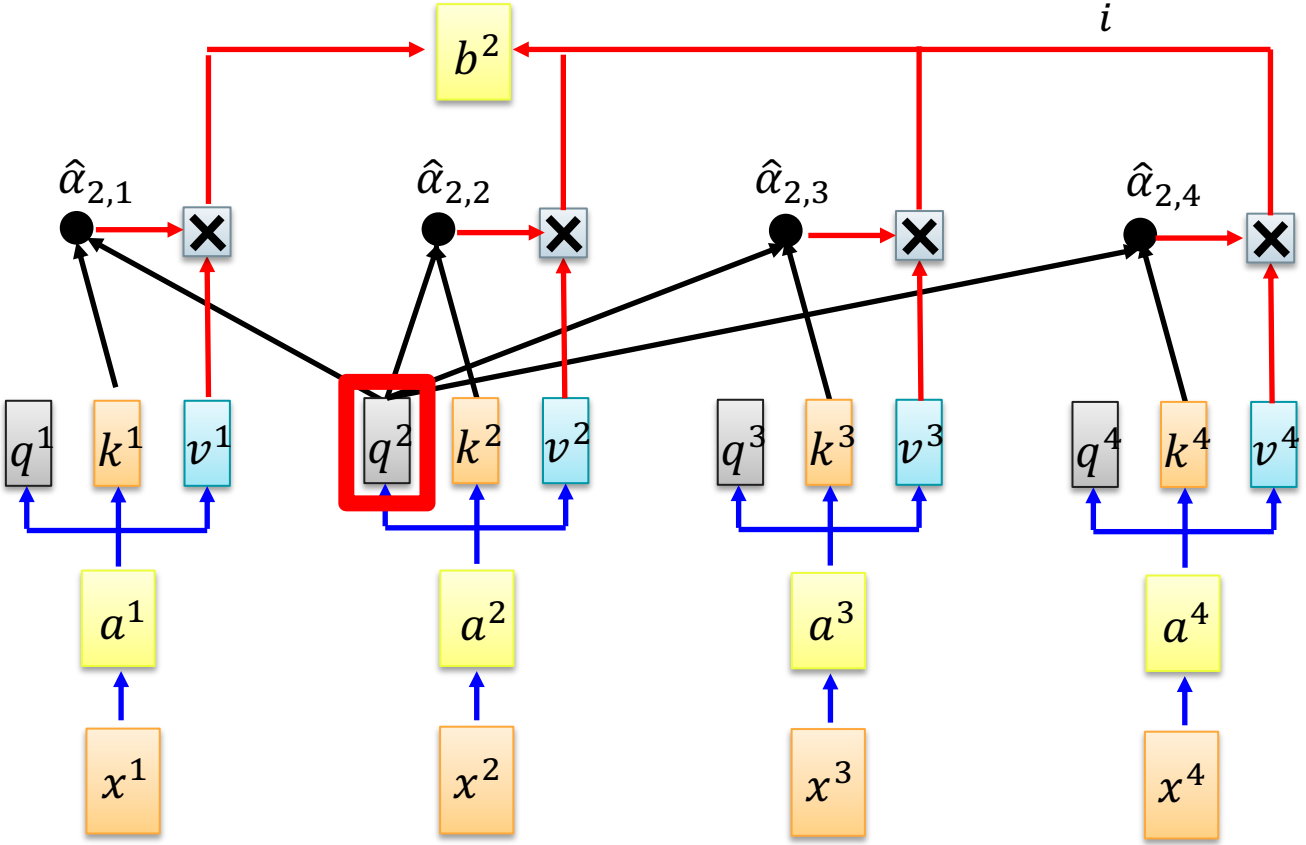
$$b^1 = \sum_i \hat{\alpha}_{1,i} v^i$$

Considering the whole sequence



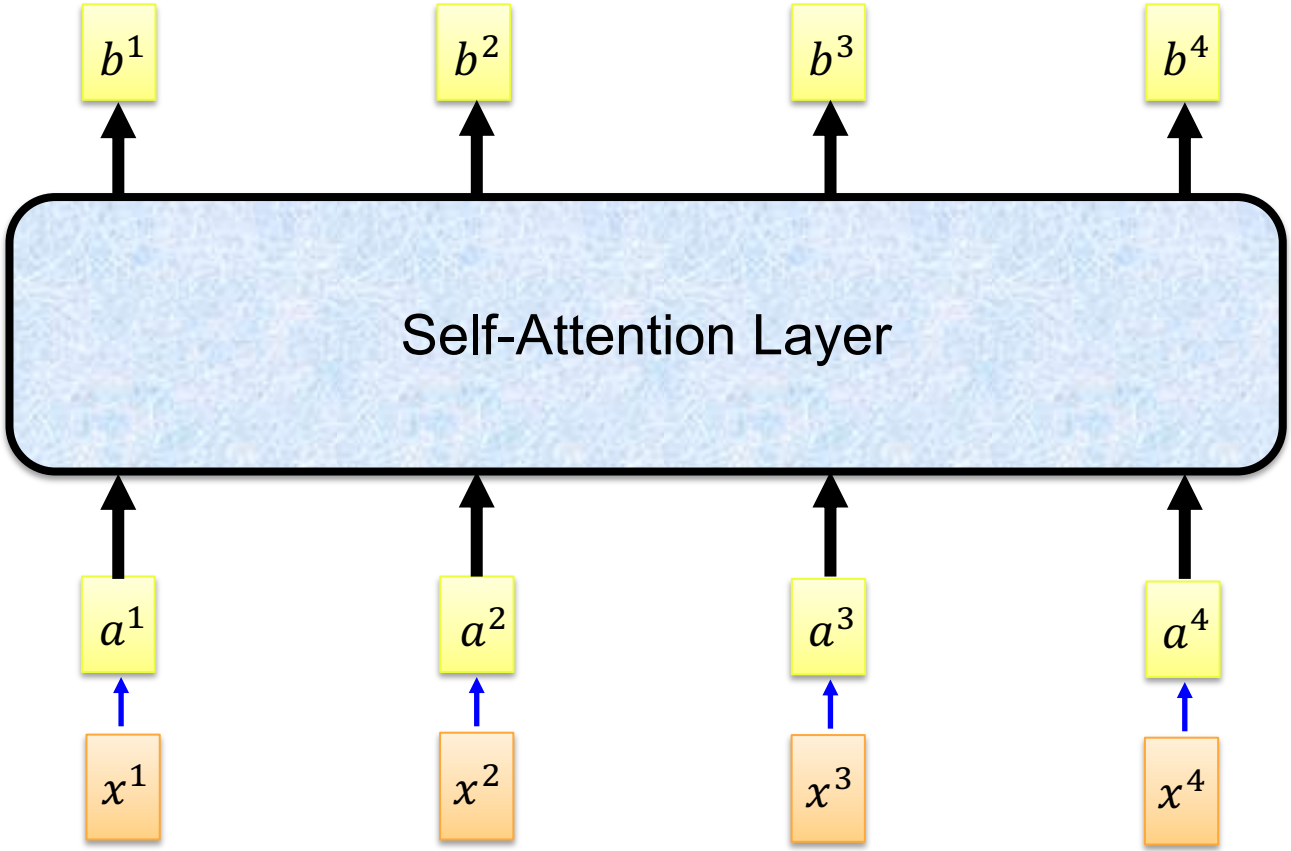
Self-Attention

$$b^2 = \sum_i \hat{\alpha}_{2,i} v^i$$



Self-Attention

b^1, b^2, b^3, b^4 can be parallelly computed.



Self-Attention

$$q^i = W^q a^i$$

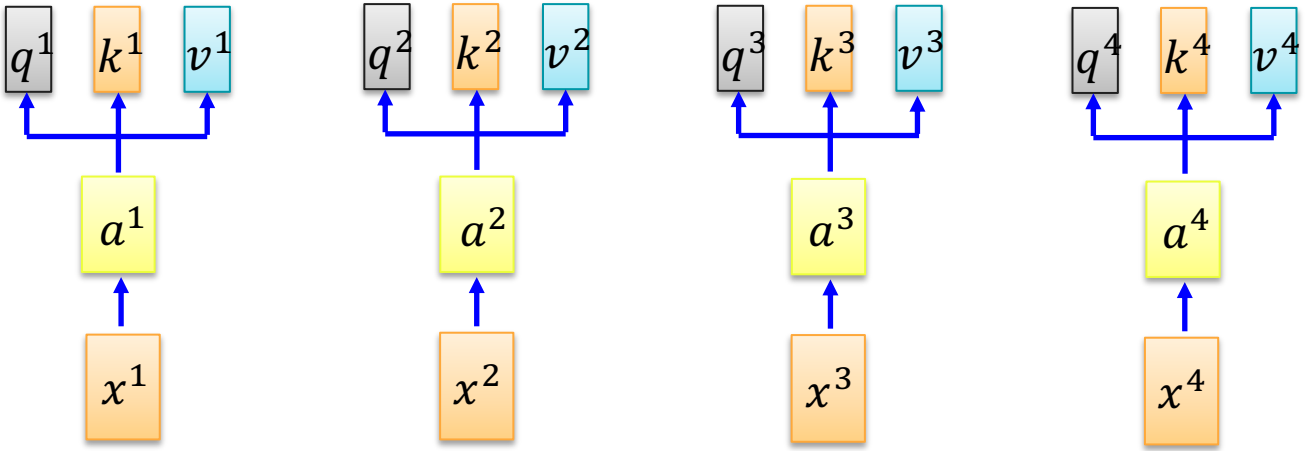
$$k^i = W^k a^i$$

$$v^i = W^v a^i$$

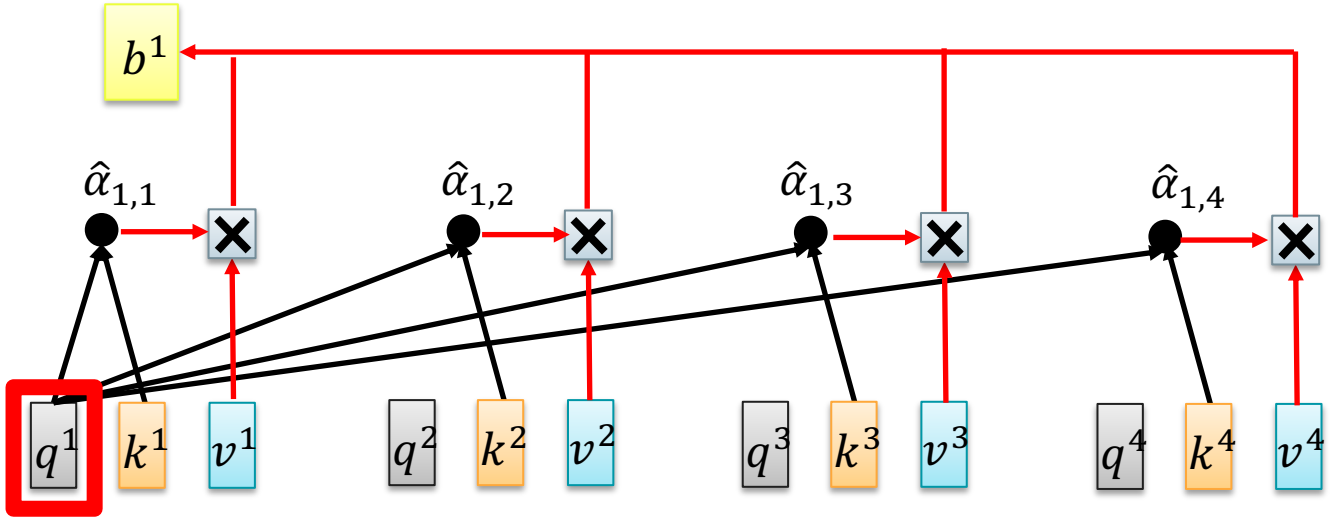
$$\underbrace{q^1\, q^2\, q^3\, q^4}_{Q} = W^q \underbrace{a^1\, a^2\, a^3\, a^4}_{I}$$

$$\underbrace{k^1\, k^2\, k^3\, k^4}_{K} = W^k \underbrace{a^1\, a^2\, a^3\, a^4}_{I}$$

$$\underbrace{v^1\, v^2\, v^3\, v^4}_{V} = W^v \underbrace{a^1\, a^2\, a^3\, a^4}_{I}$$



Self-Attention



$\alpha_{1,1} = k^1 q^1$ $\alpha_{1,2} = k^2 q^1$

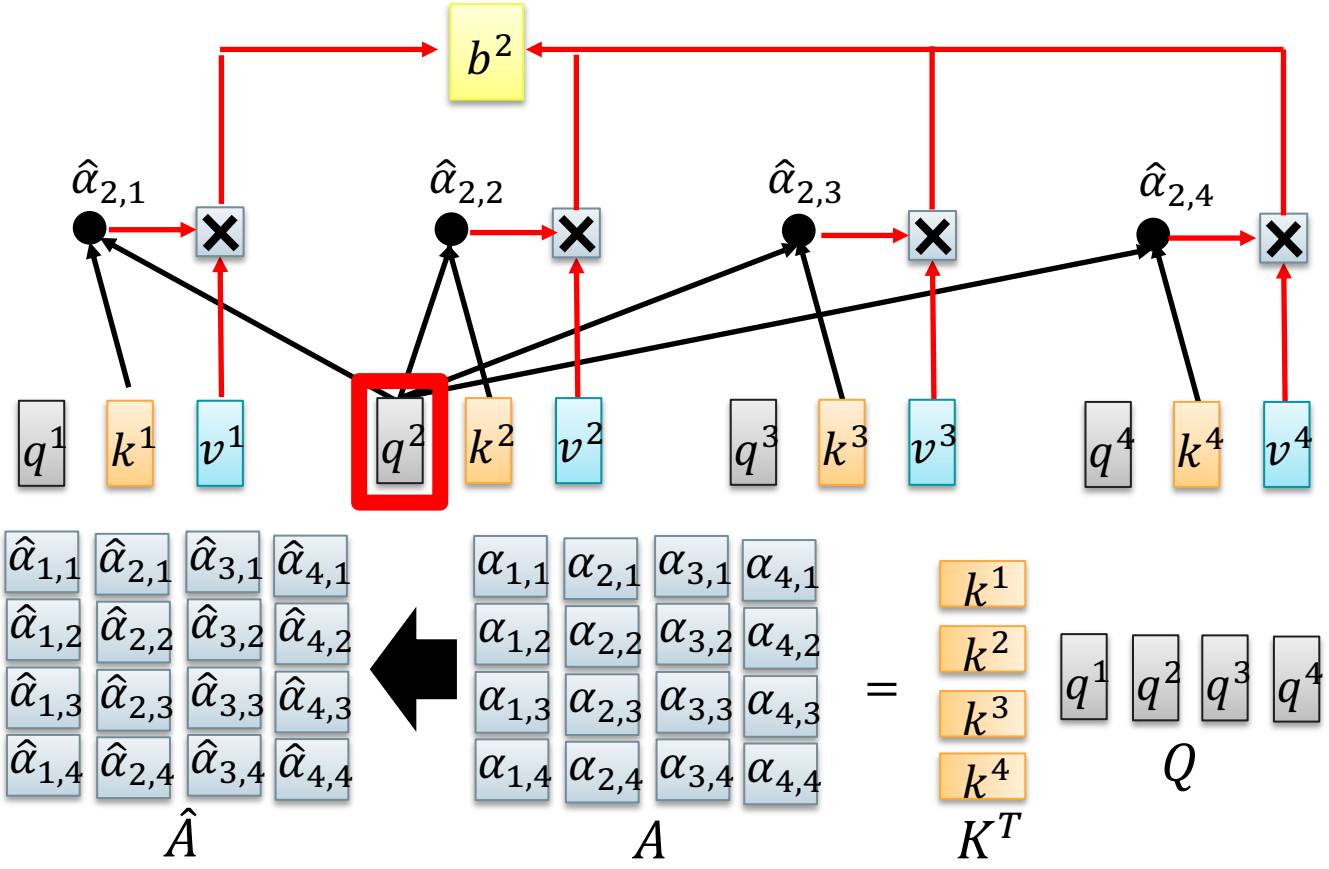
$\alpha_{1,3} = k^3 q^1$ $\alpha_{1,4} = k^4 q^1$

(ignore $\sqrt{d}$ for simplicity)

$$\begin{bmatrix}\alpha_{1,1}\\ \alpha_{1,2}\\ \alpha_{1,3}\\ \alpha_{1,4}\end{bmatrix} = \begin{bmatrix}k^1\\ k^2\\ k^3\\ k^4\end{bmatrix} q^1$$

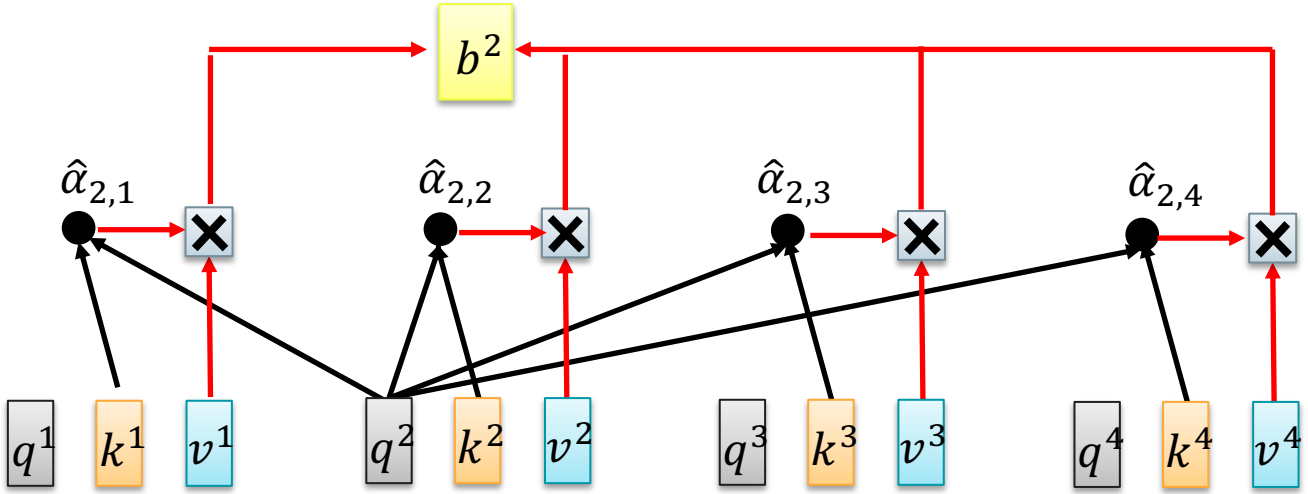
Self-Attention

$$b^2 = \sum_i \hat{\alpha}_{2,i} v^i$$



Self-Attention

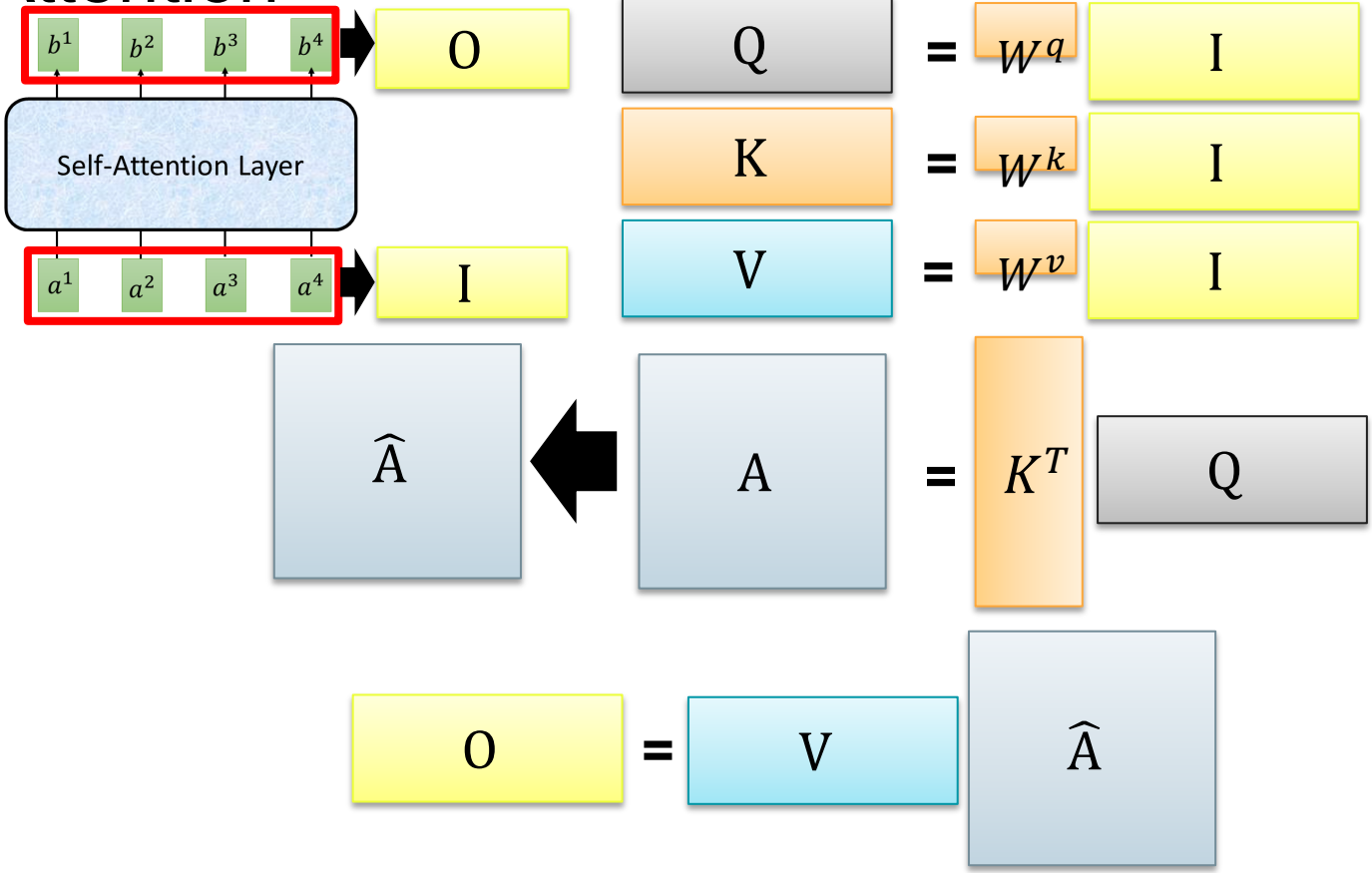
$$b^2 = \sum_i \hat{\alpha}_{2,i} v^i$$



$$\begin{matrix} b^1 & b^2 & b^3 & b^4 \end{matrix} = \begin{matrix} v^1 & v^2 & v^3 & v^4 \end{matrix} \begin{bmatrix} \hat{\alpha}_{1,1} & \hat{\alpha}_{2,1} & \hat{\alpha}_{3,1} & \hat{\alpha}_{4,1} \\ \hat{\alpha}_{1,2} & \hat{\alpha}_{2,2} & \hat{\alpha}_{3,2} & \hat{\alpha}_{4,2} \\ \hat{\alpha}_{1,3} & \hat{\alpha}_{2,3} & \hat{\alpha}_{3,3} & \hat{\alpha}_{4,3} \\ \hat{\alpha}_{1,4} & \hat{\alpha}_{2,4} & \hat{\alpha}_{3,4} & \hat{\alpha}_{4,4} \end{bmatrix}$$

O V $\hat{A}$

Self-Attention

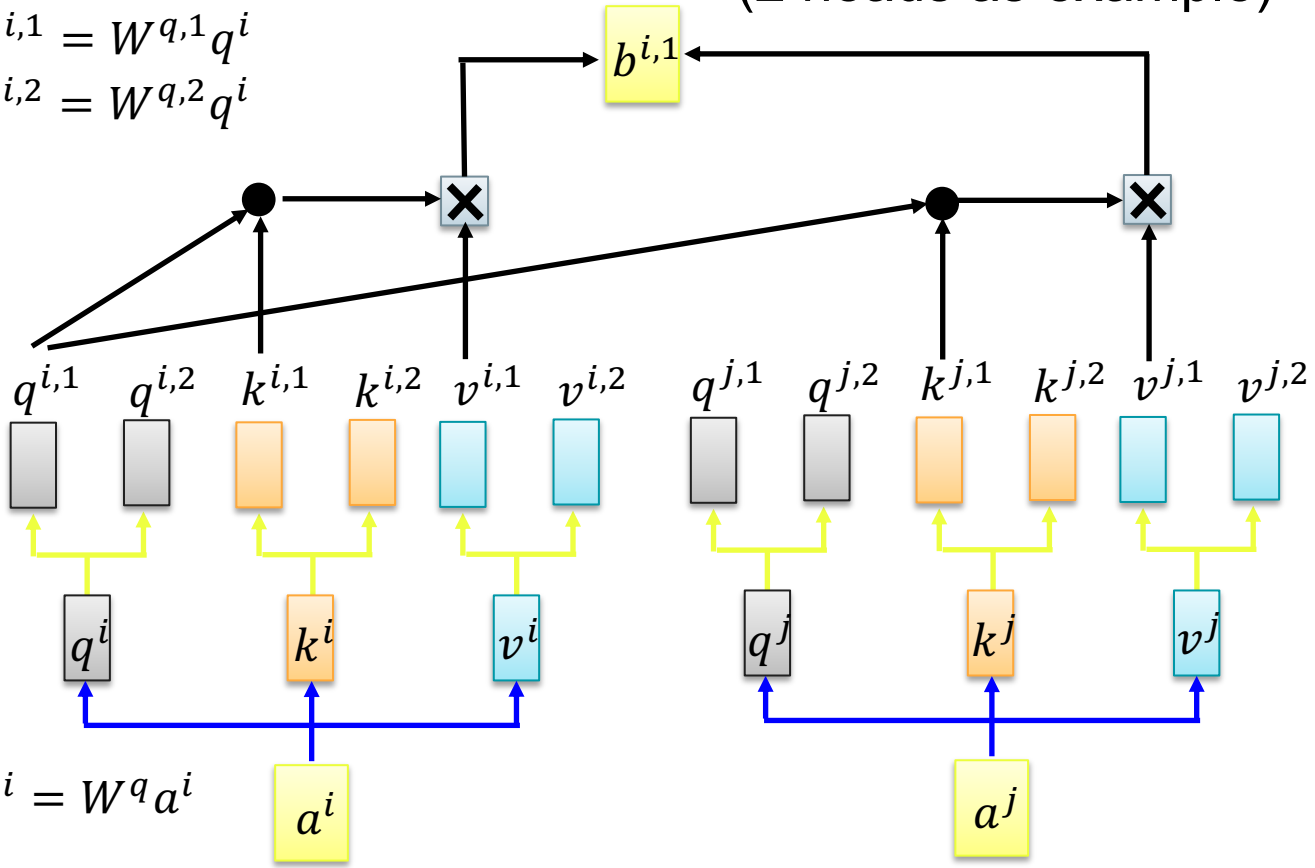


Multi-head Self-Attention

(2 heads as example)

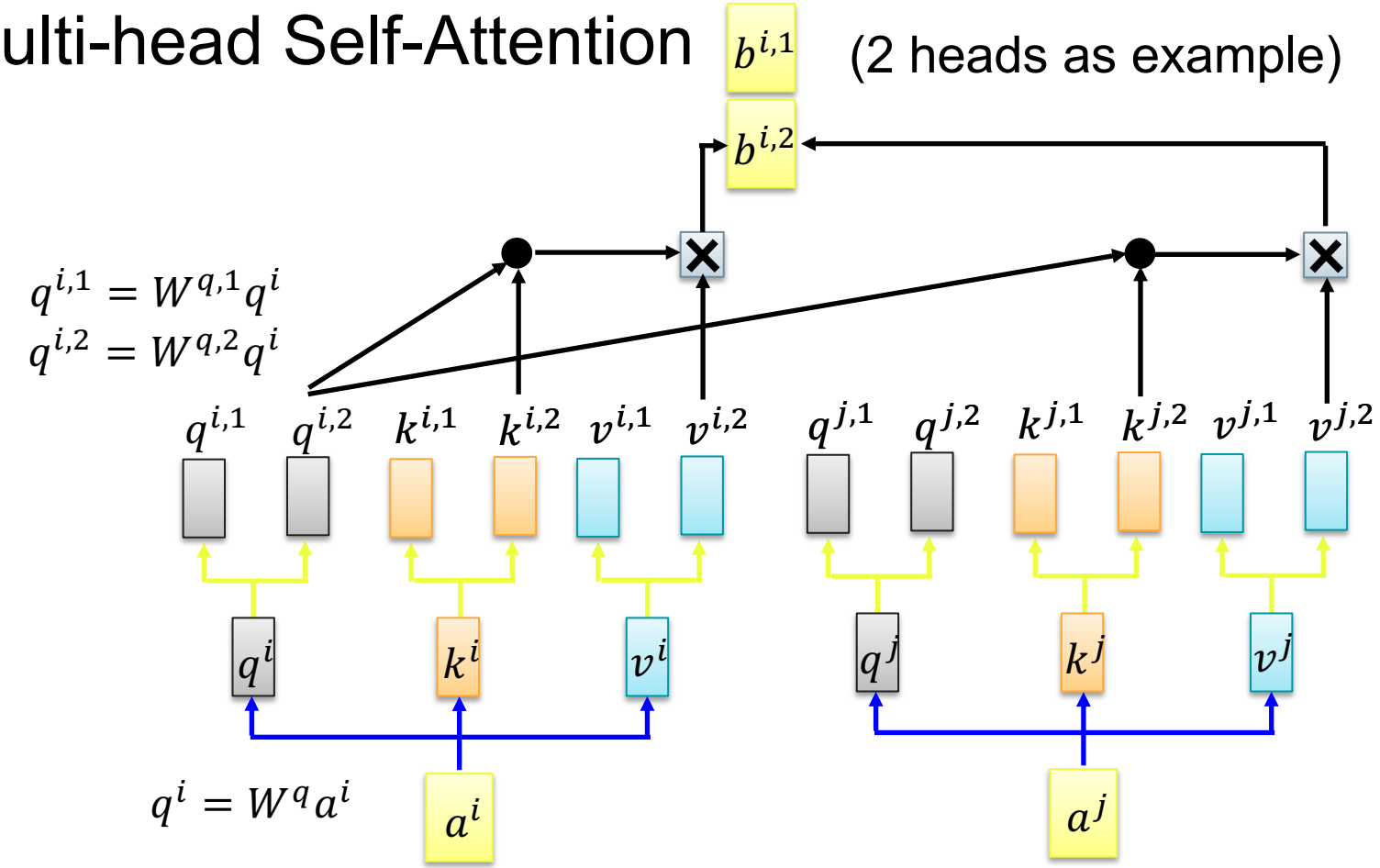
$q^{i,1} = W^{q,1} q^i$

$q^{i,2} = W^{q,2} q^i$



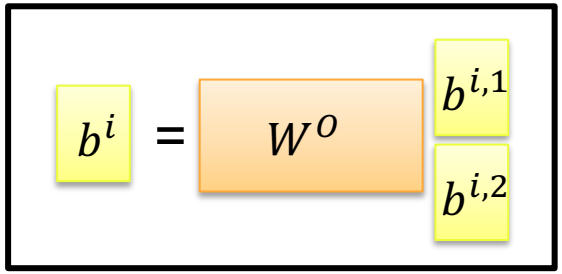
Multi-head Self-Attention

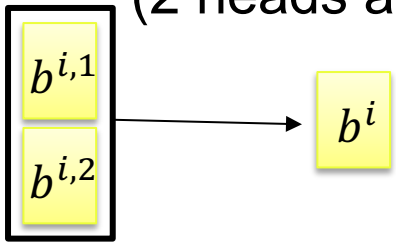
(2 heads as example)

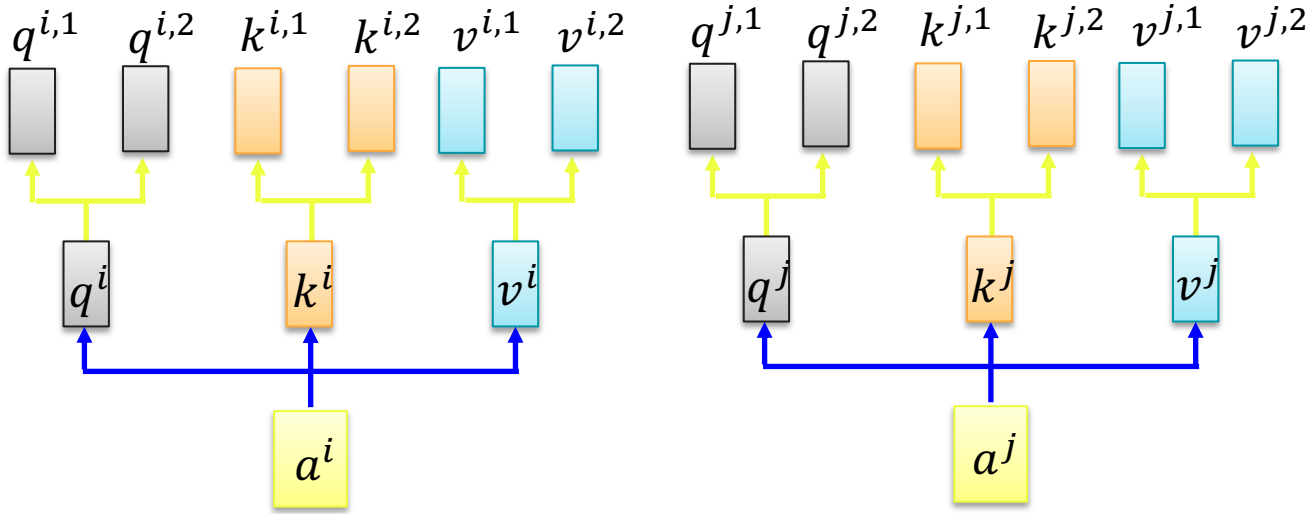


Multi-head Self-Attention

(2 heads as example)





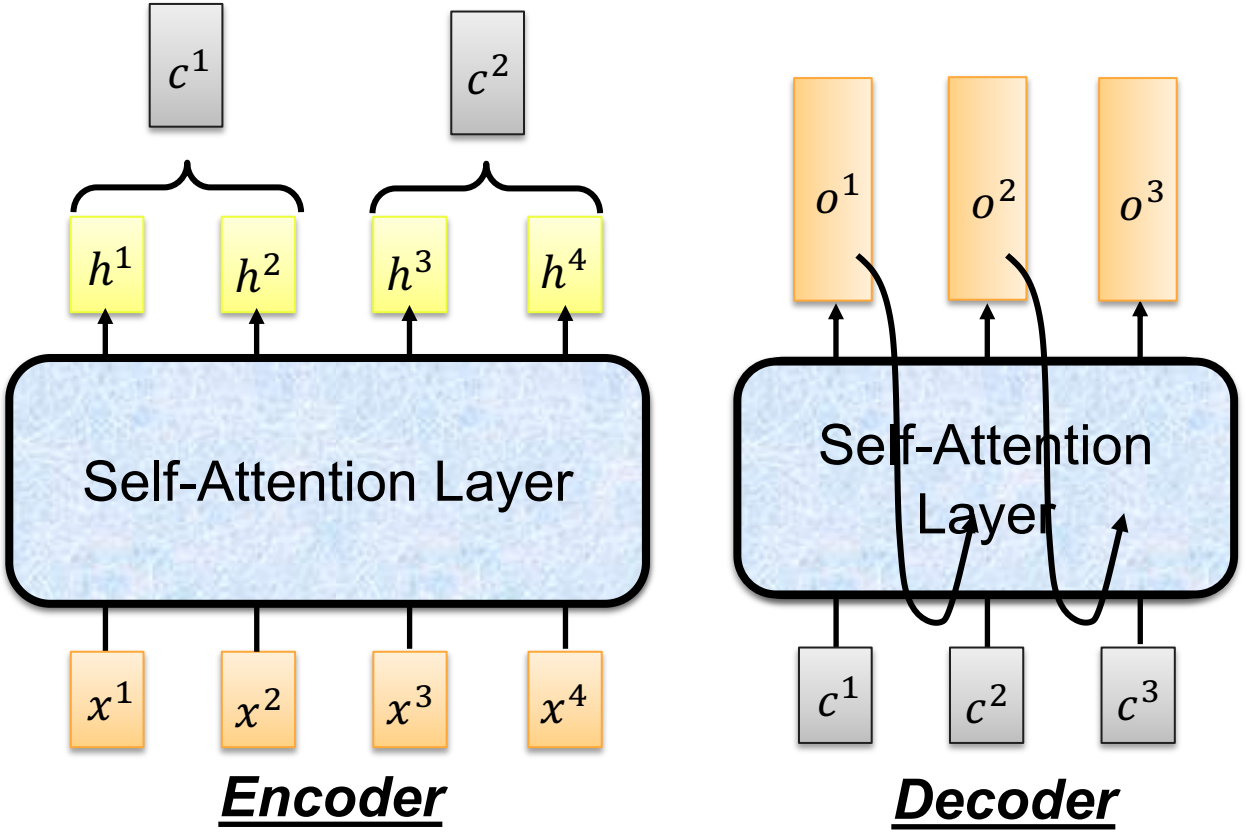


Outline

- Problem
- Self-Attention
- Transformer
- BERT
- Evaluation

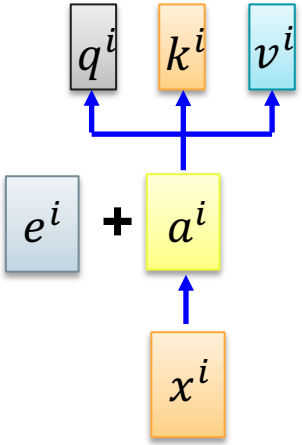
Review: https://www.youtube.com/watch?v=ZjfjPzXw6og&feature=youtu.be

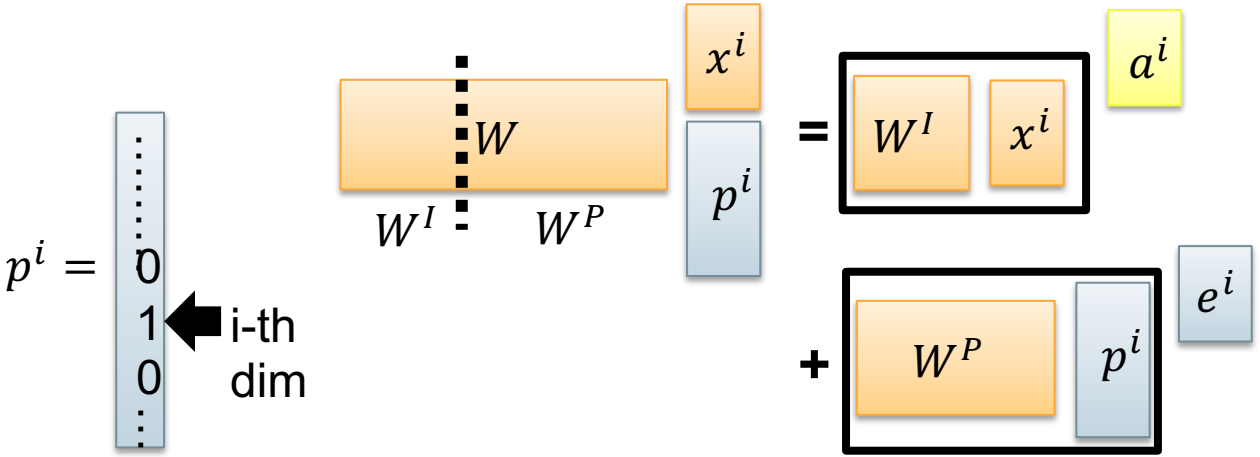
Seq2seq with Attention

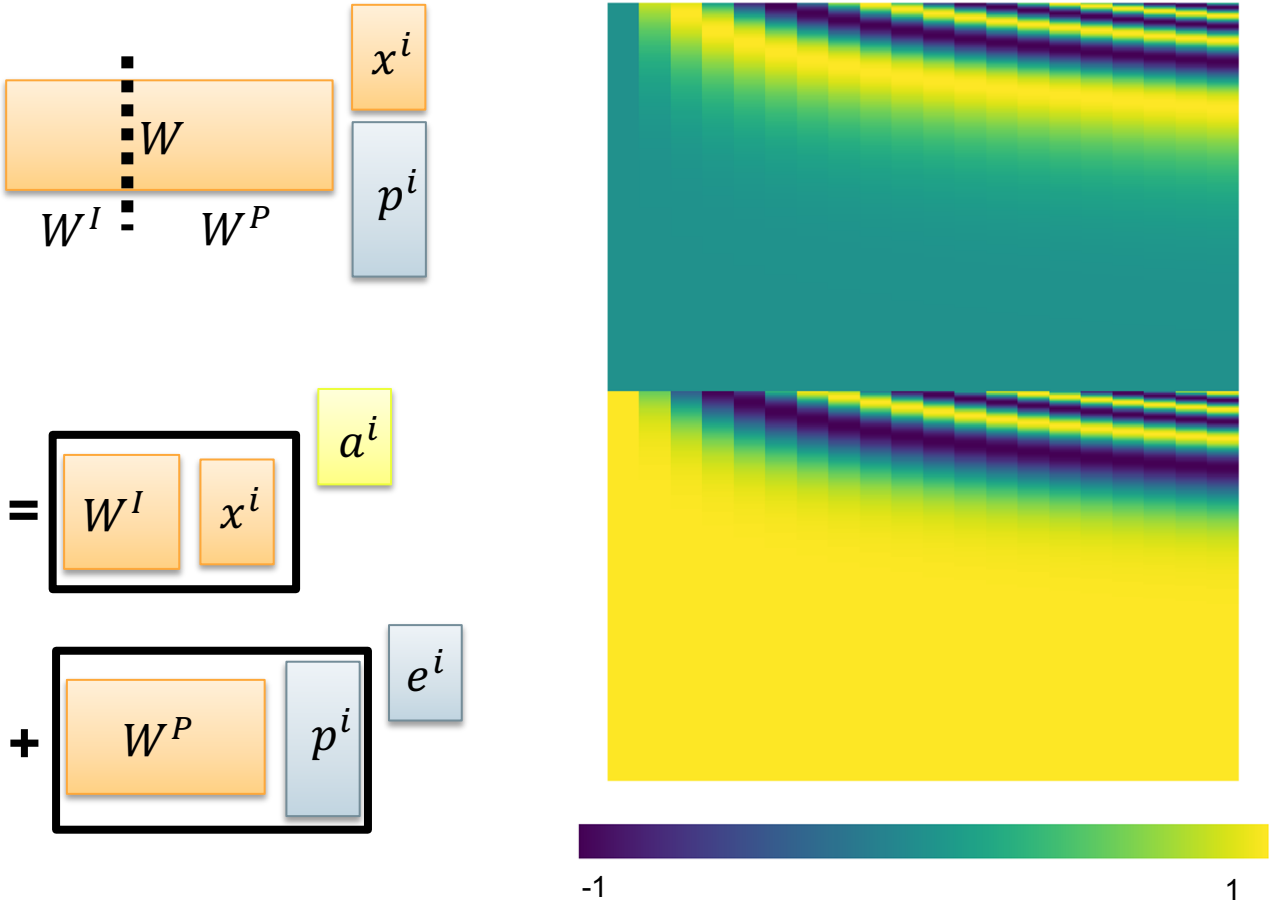


Positional Encoding

- No position information in self-attention.
- Original paper: each position has a unique positional vector e^i (not learned from data)
- In other words: each x^i appends a one-hot vector p^i







source of image: http://jalammar.github.io/illustrated-transformer/

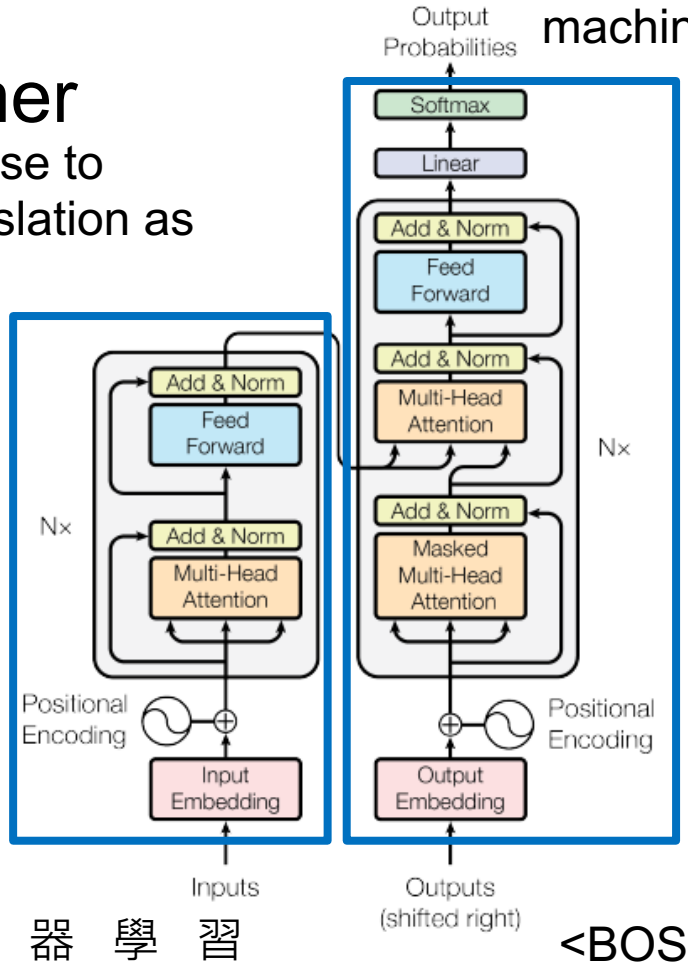
Outline

- Problem
- Gender Bias
- Transformer
- BERT
- Evaluation

https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

BERT-Transformer

Using Chinese to English translation as example



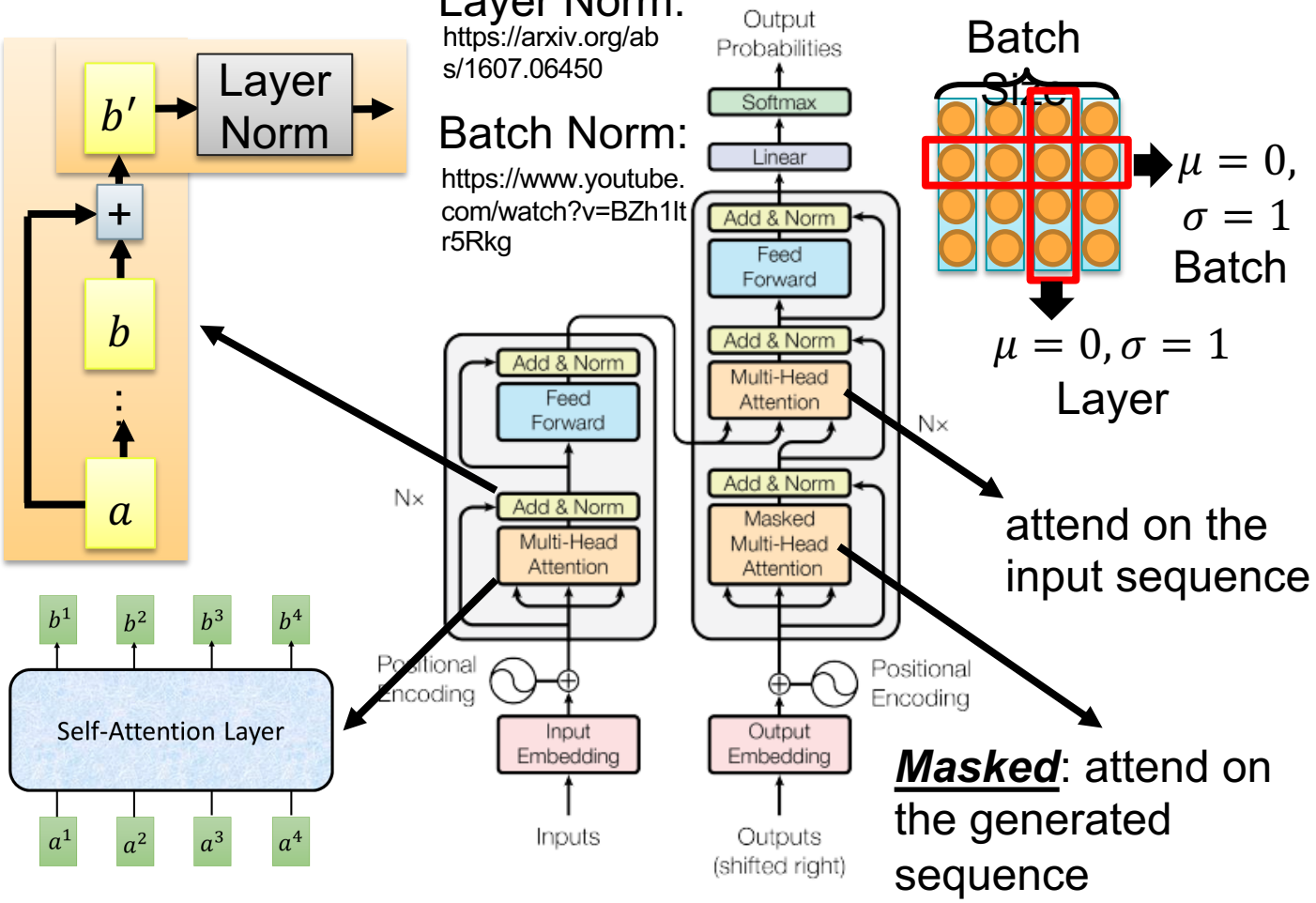
Encoder

Decoder

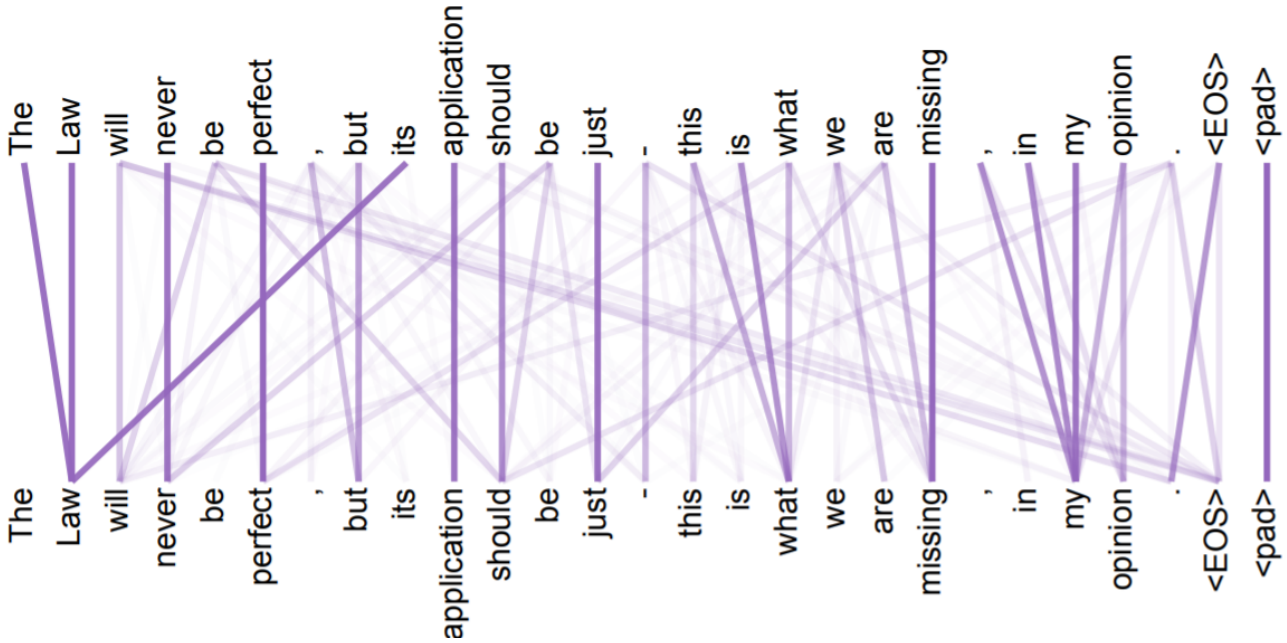
機 器 學 習

<BOS> machine

machine learning

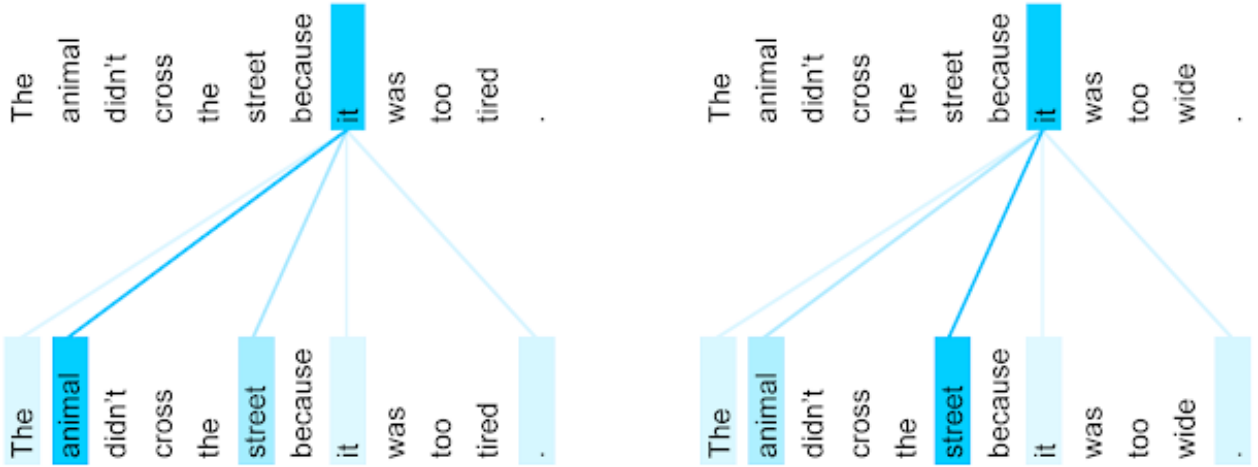

Layer Norm:
https://arxiv.org/abs/1607.06450
Batch Norm:
https://www.youtube.com/watch?v=BZh1Ir5Rkg
Layer Norm
b'
$+$
b
a
Batch Size
$\mu = 0, \sigma = 1$
Batch
$\mu = 0, \sigma = 1$
Layer
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Nx
attend on the input sequence
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
b^1 b^2 b^3 b^4
Self-Attention Layer
a^1 a^2 a^3 a^4
Masked: attend on the generated sequence

Attention Visualization



https://arxiv.org/abs/1706.03762

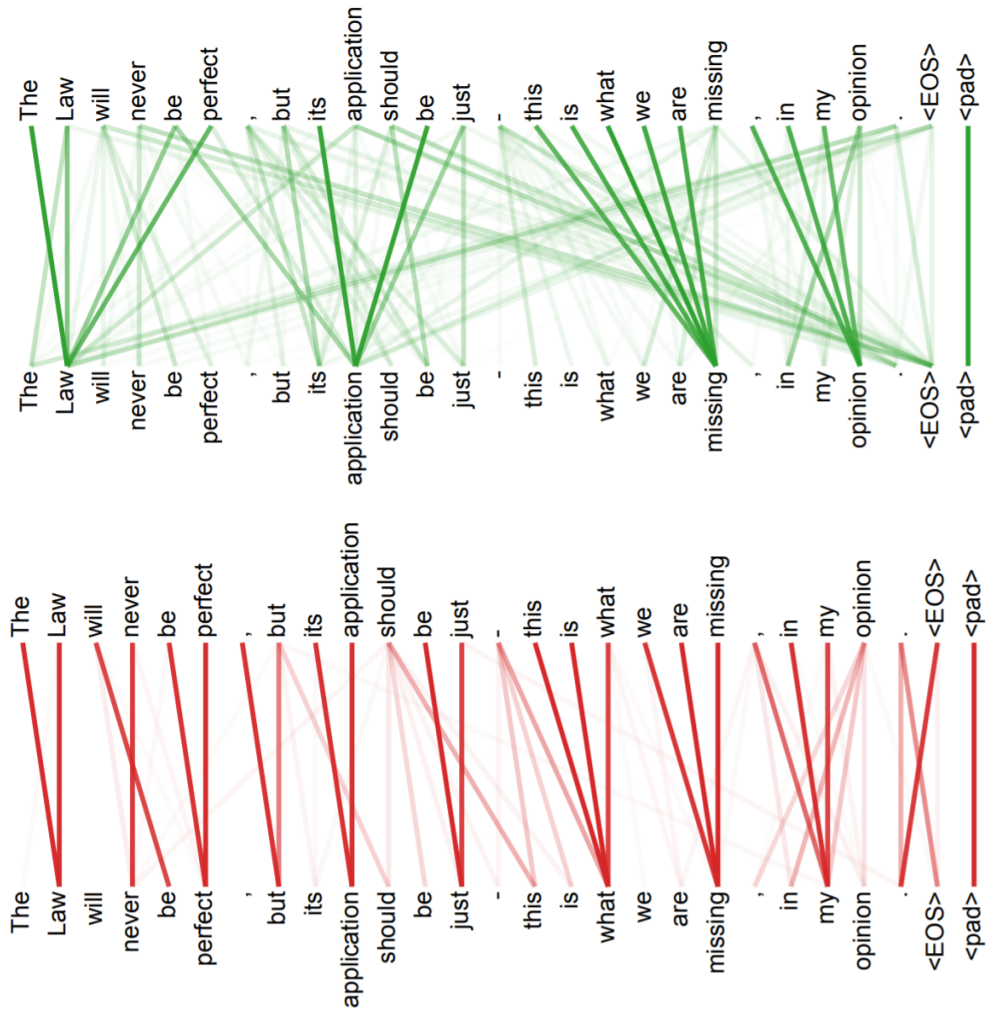
Attention Visualization



The encoder self-attention distribution for the word “it” from the 5th to the 6th layer of a Transformer trained on English to French translation (one of eight attention heads).

https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

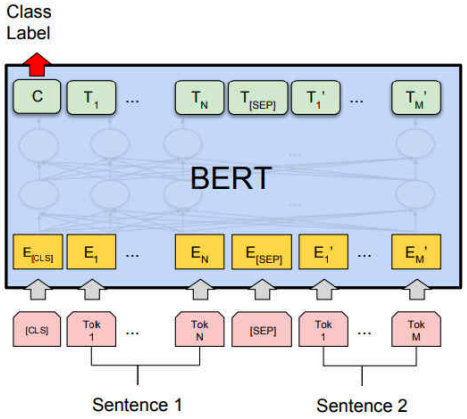
Multi-head Attention Visualization



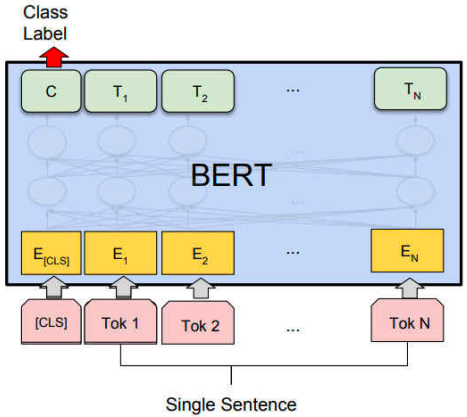
Outline

- Problem
- Self-Attention
- Transformer
- BERT
- Fine-tune BERT

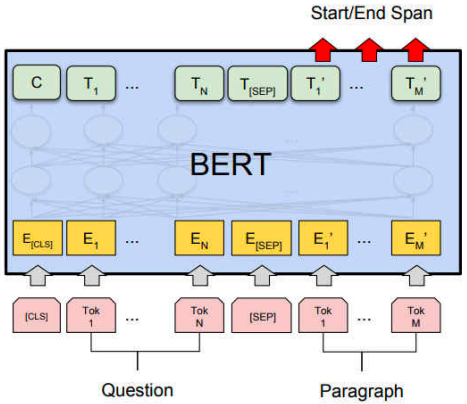
Fine-tun



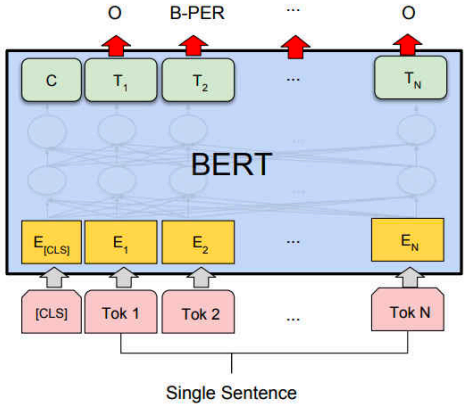
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Fine-tune BERT

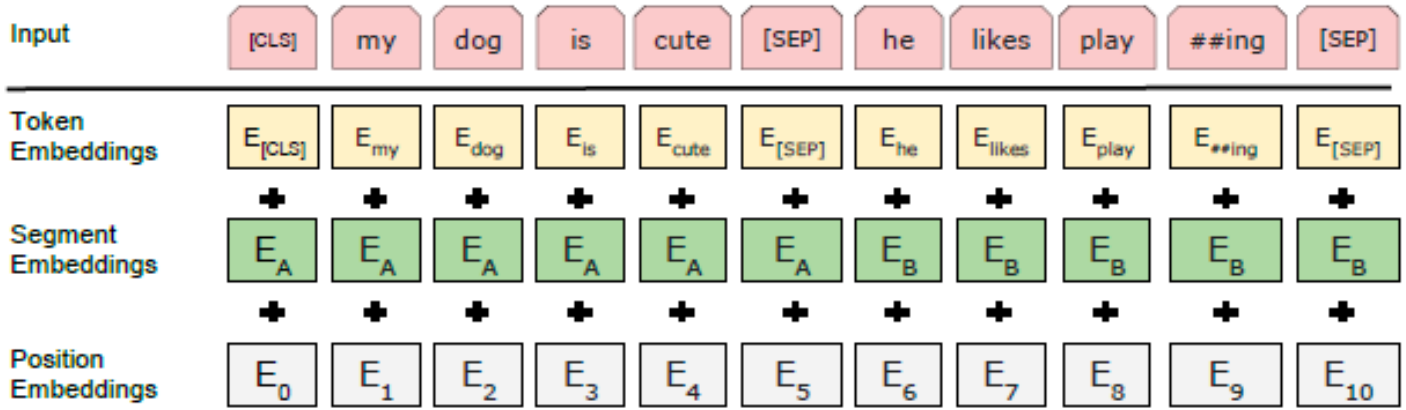


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Fine-tune BERT

References

- Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim. "A machine learning approach to coreference resolution of noun phrases." *Computational linguistics* 27.4 (2001): 521-544.
- Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.
- Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).