深度學習基礎概論

0408

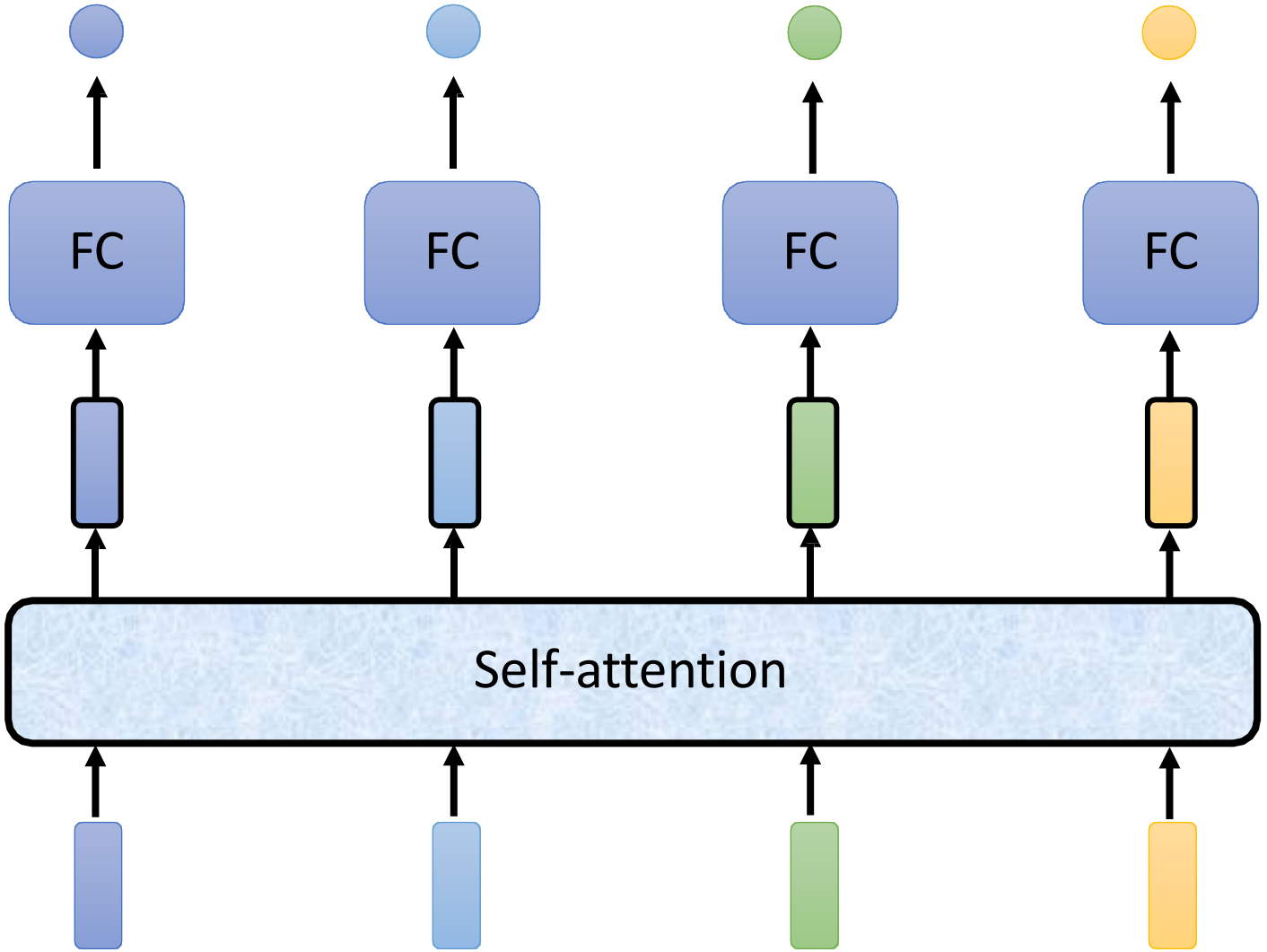
目錄

- Self attention
- Postional Coding
- BERT
- Code

Self attention

- 在一段句子中，相同單字在不同位置中會有不同意義
- E.g : I saw the saw. 第一個 saw 是 see 的過去式，而第二個是鋸子
- 若要準確翻譯或讀懂句子便要考慮單字和前後文之間的關係，而 self-attention 便是在做這件事

Self-attention

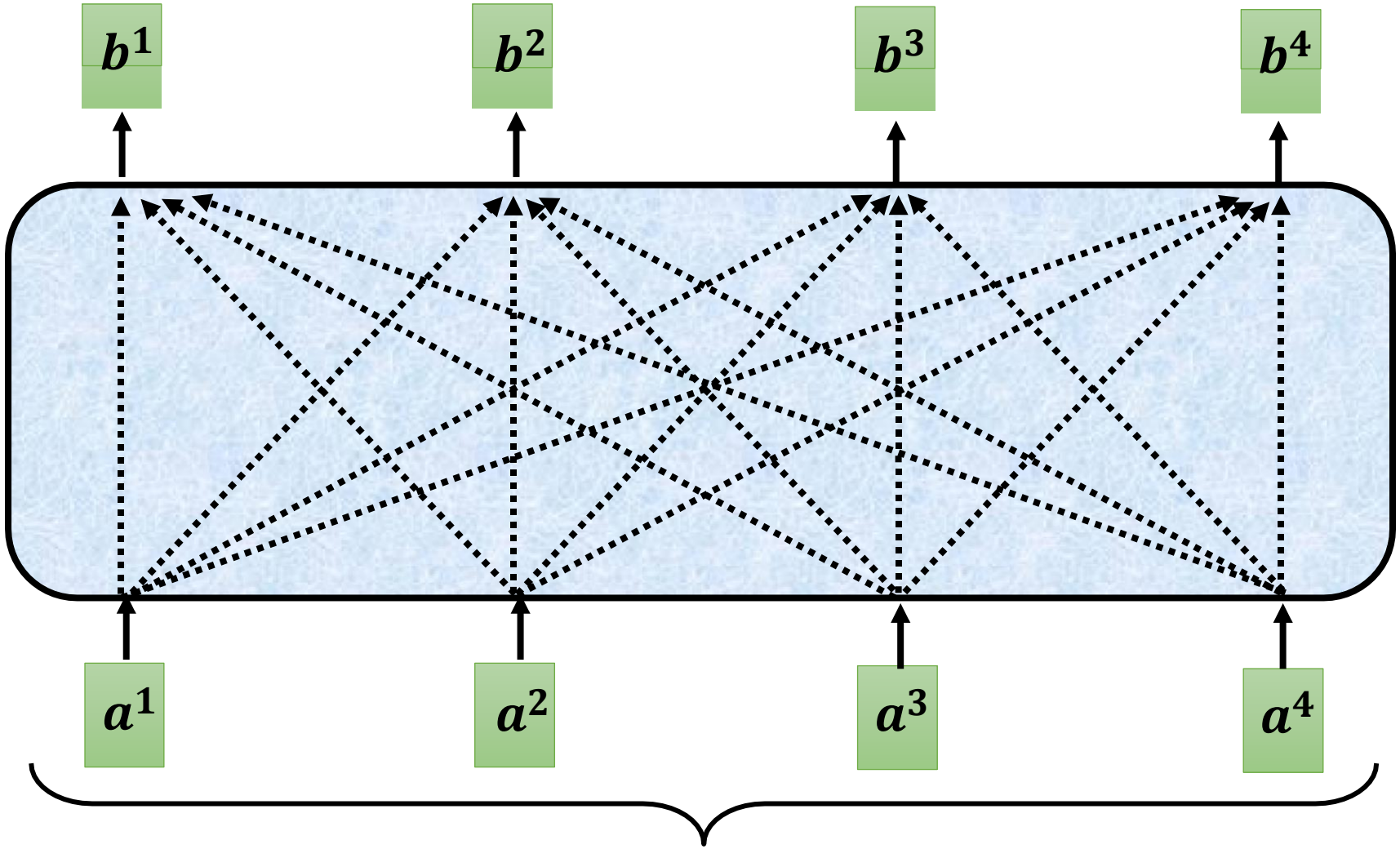


Ouput : vector，一段句子

Input : vector，一段句子

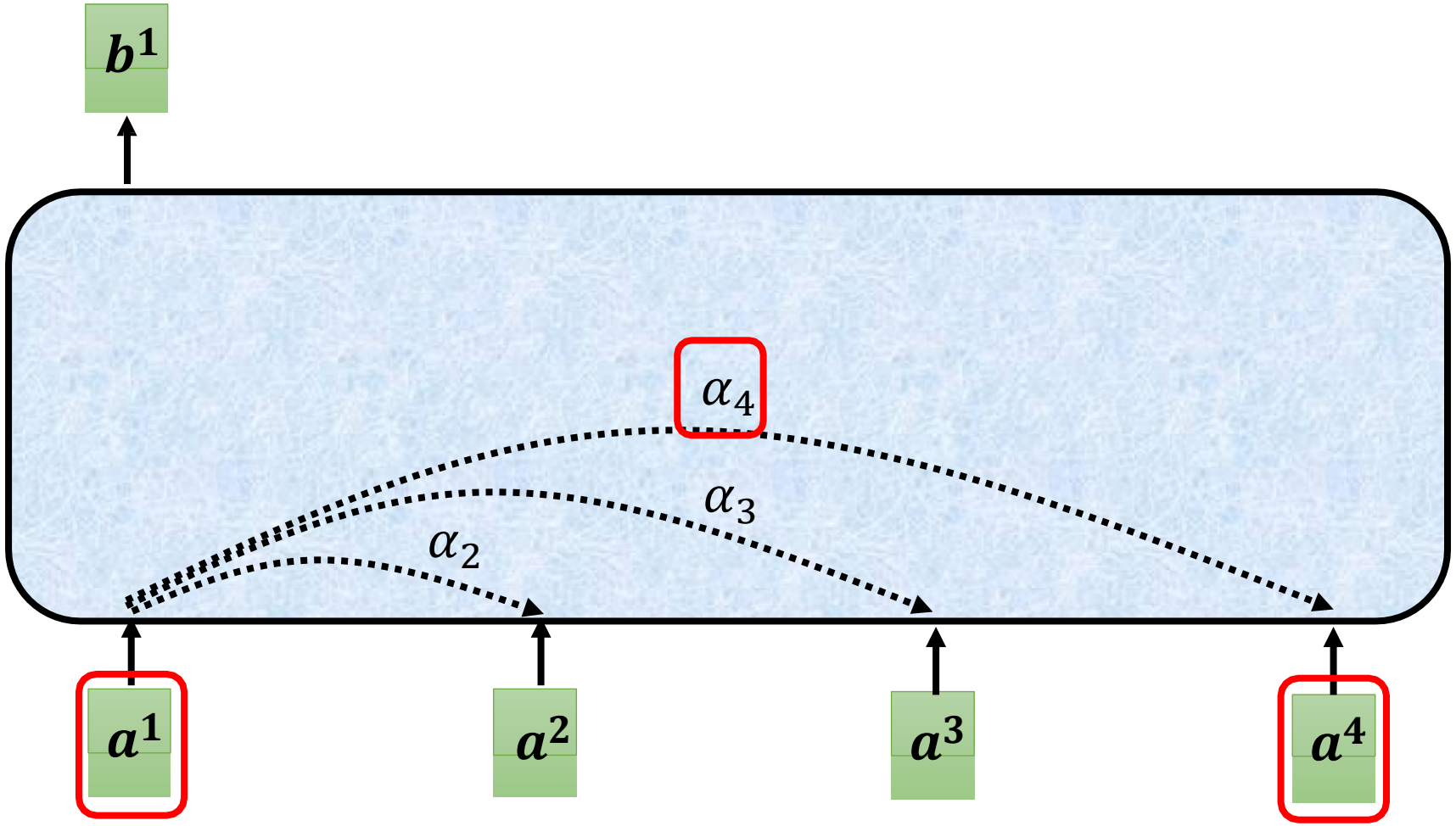
Self-attention

考慮字跟字之間的關係



假設有一個由四個字組成的句子

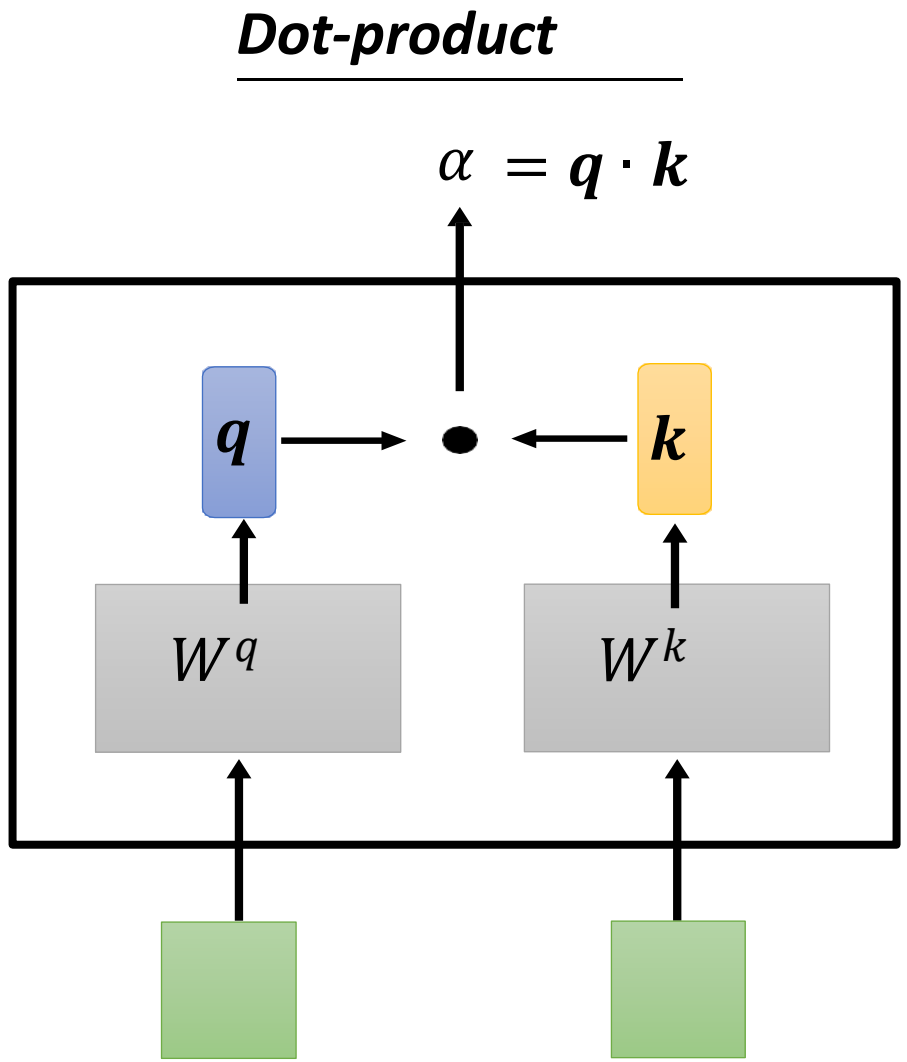
Self-attention



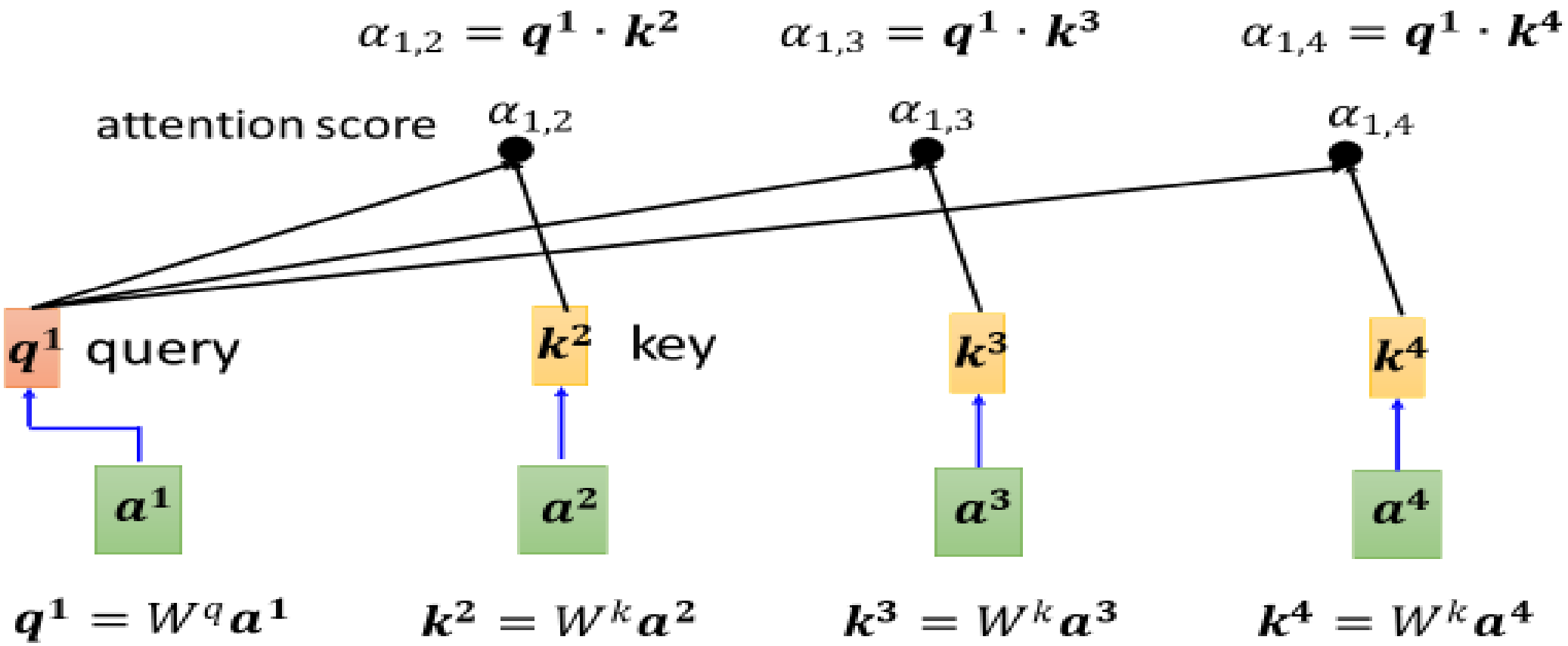
α 是字跟字之間的相關性

Self-attention

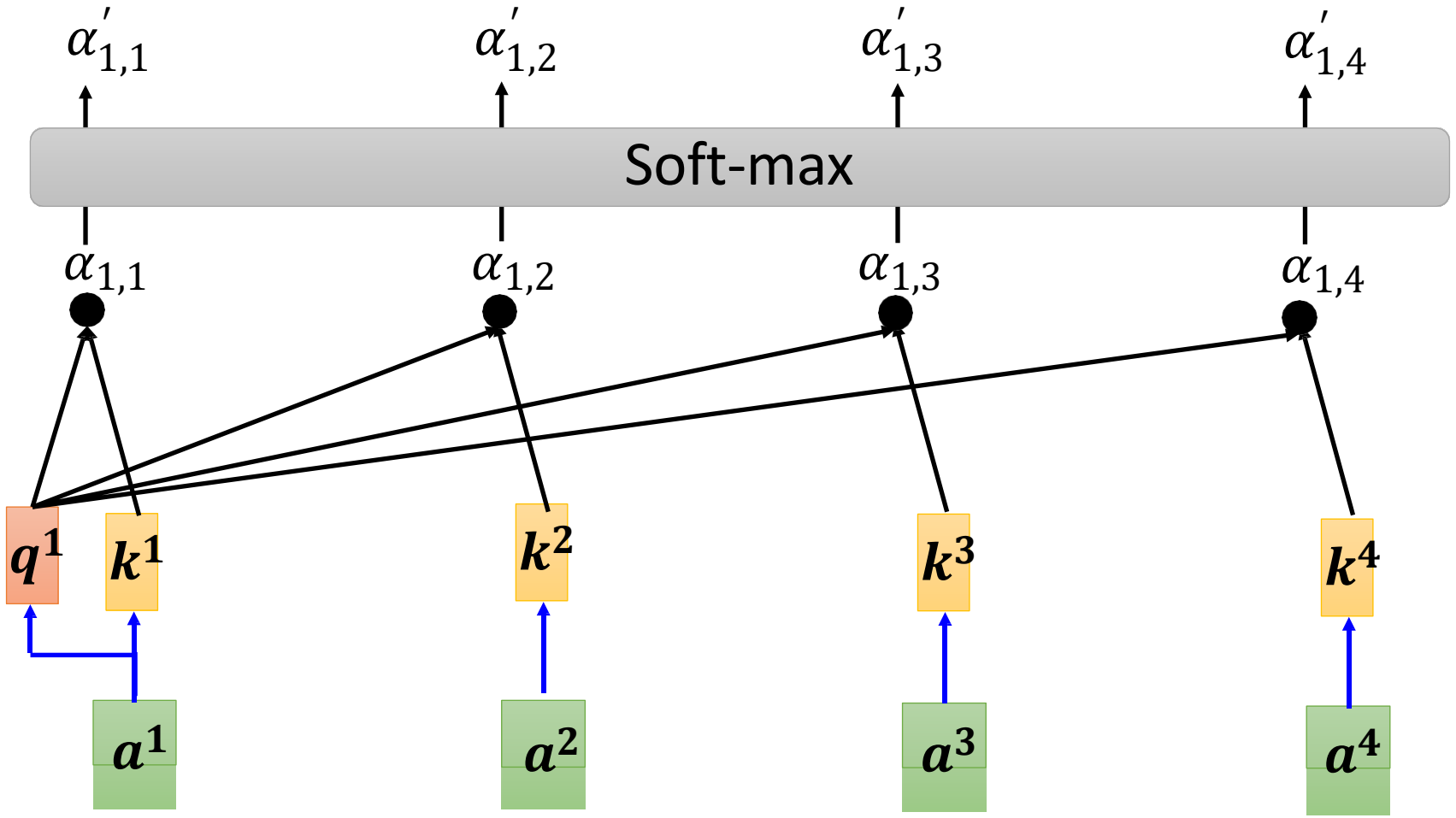
w^q、w^k 是一個 matrix ，裡面係數由 model train 出來的



Self-attention



Self-attention



$q^1 = W^q a^1$ $\quad k^2 = W^k a^2$ $\quad k^3 = W^k a^3$ $\quad k^4 = W^k a^4$

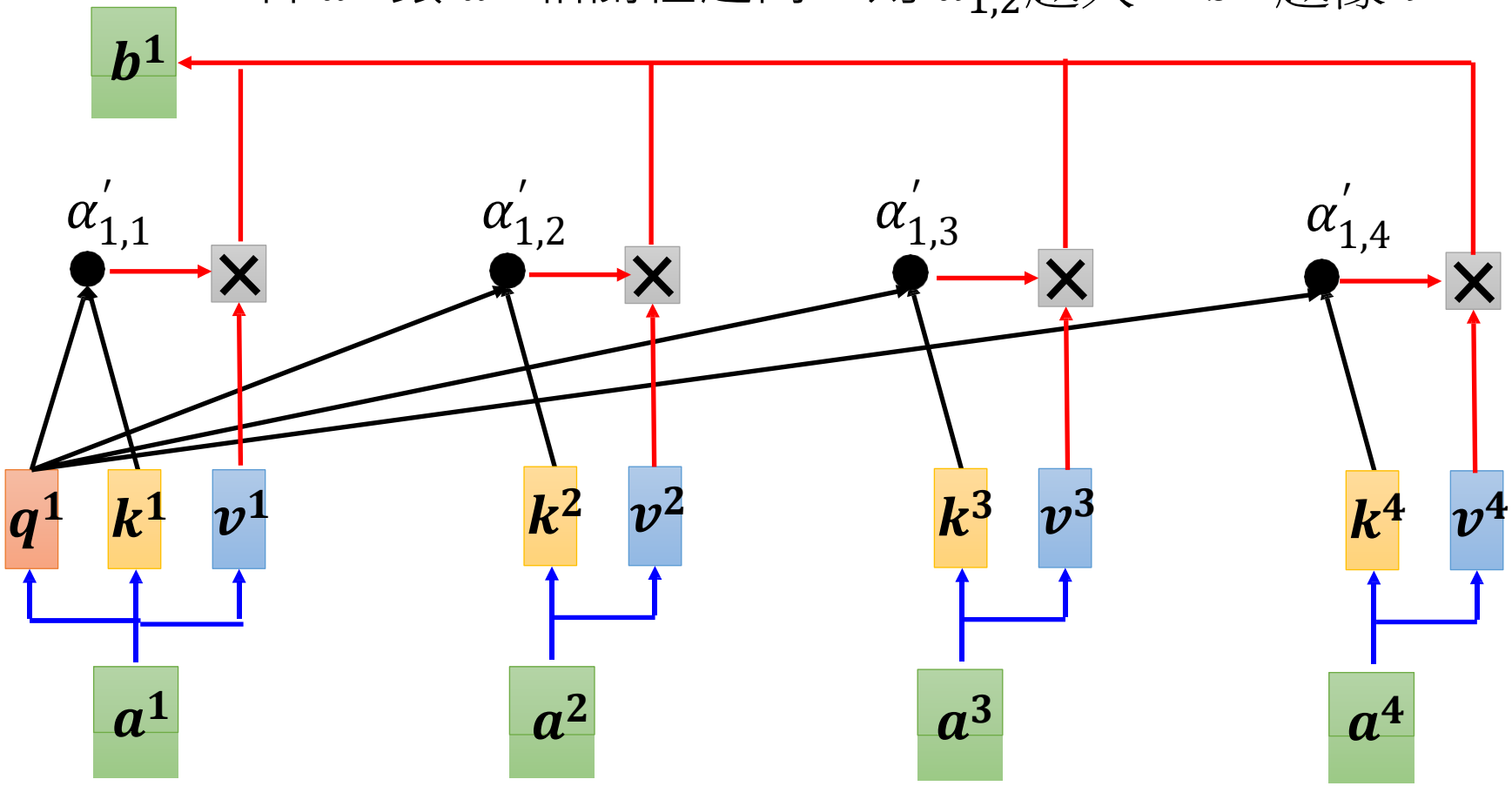
$k^1 = W^k a^1$

Self-attention

Extract information based on attention scores

v^i 也是個 matrix， $b^1 = \sum a^2_{1,i} * v^i$

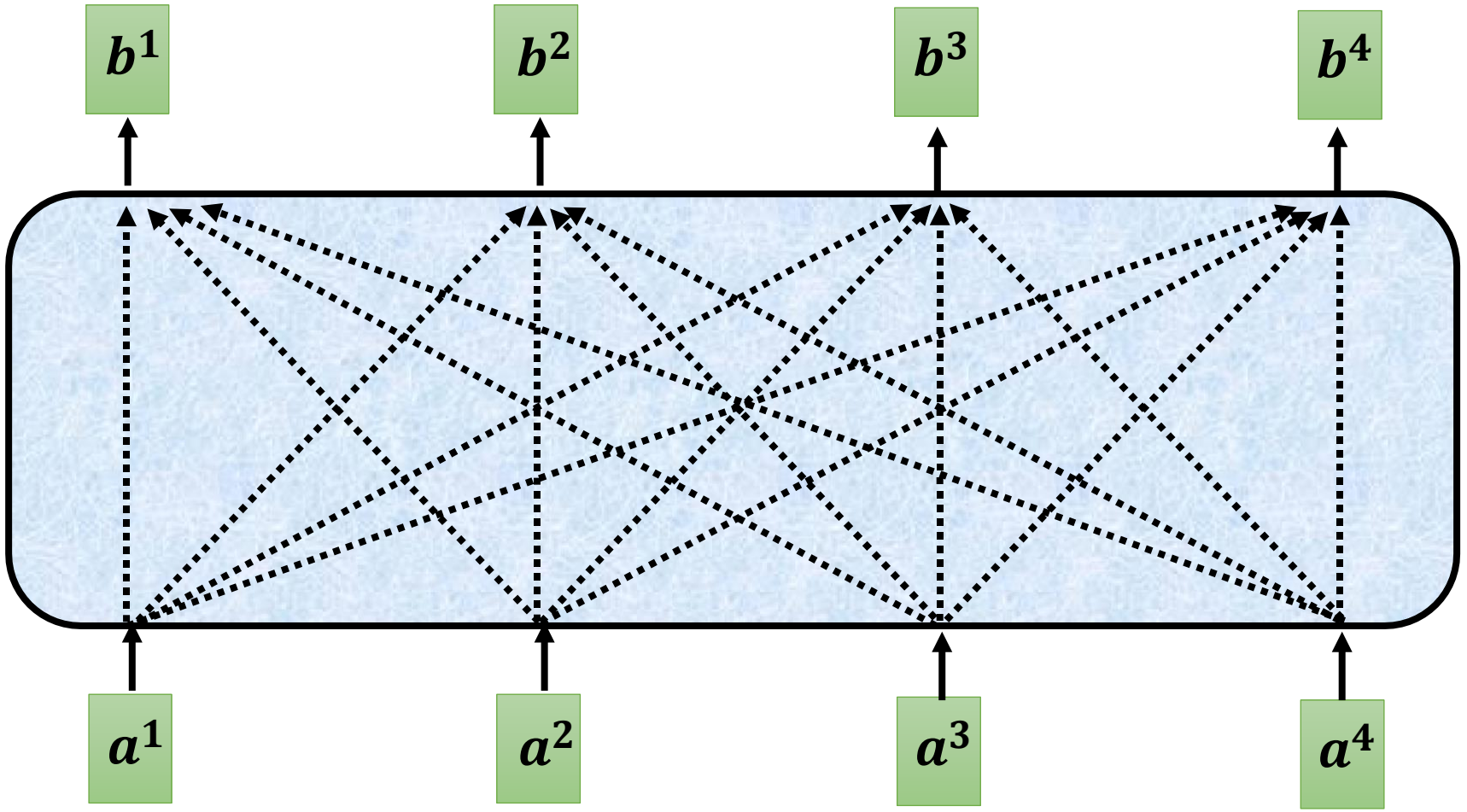
若 a^1 跟 a^2 相關性越高，則 $\alpha'_{1,2}$越大， b^1 越像 v^2



$v^1 = W^v a^1$ $v^2 = W^v a^2$ $v^3 = W^v a^3$ $v^4 = W^v a^4$

Self-attention

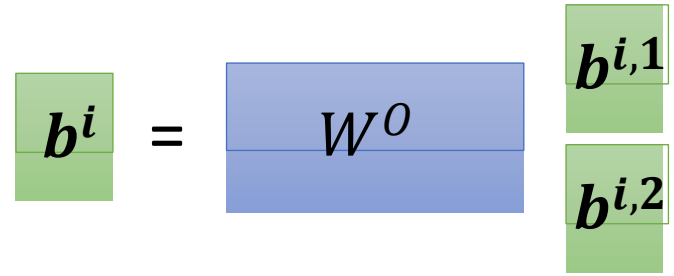
$b^1 \sim b^4$ 不是依序產生，而是可以一起被計算出來的

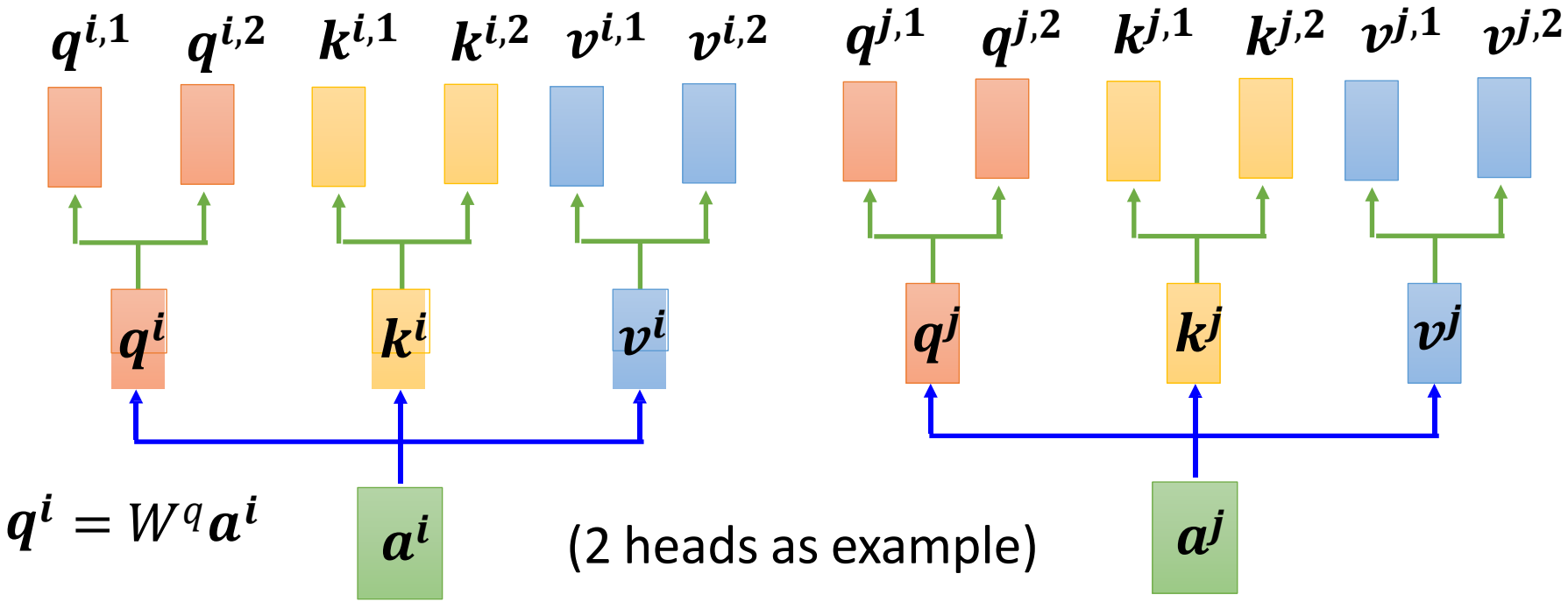


Multi-head Self-attention

可能存在不只一種相關性，做多次 self-attention

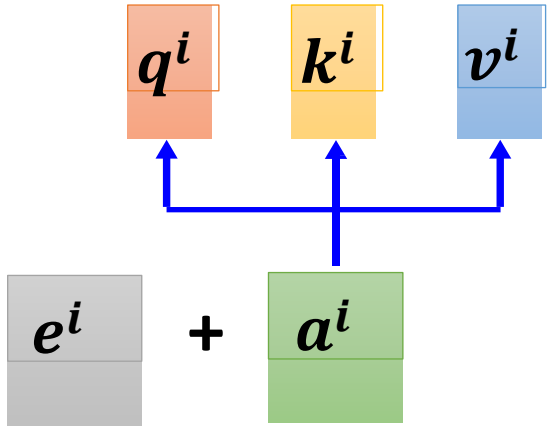
w^0 是一個 matrix ，也是由 model 去 learn 出來的





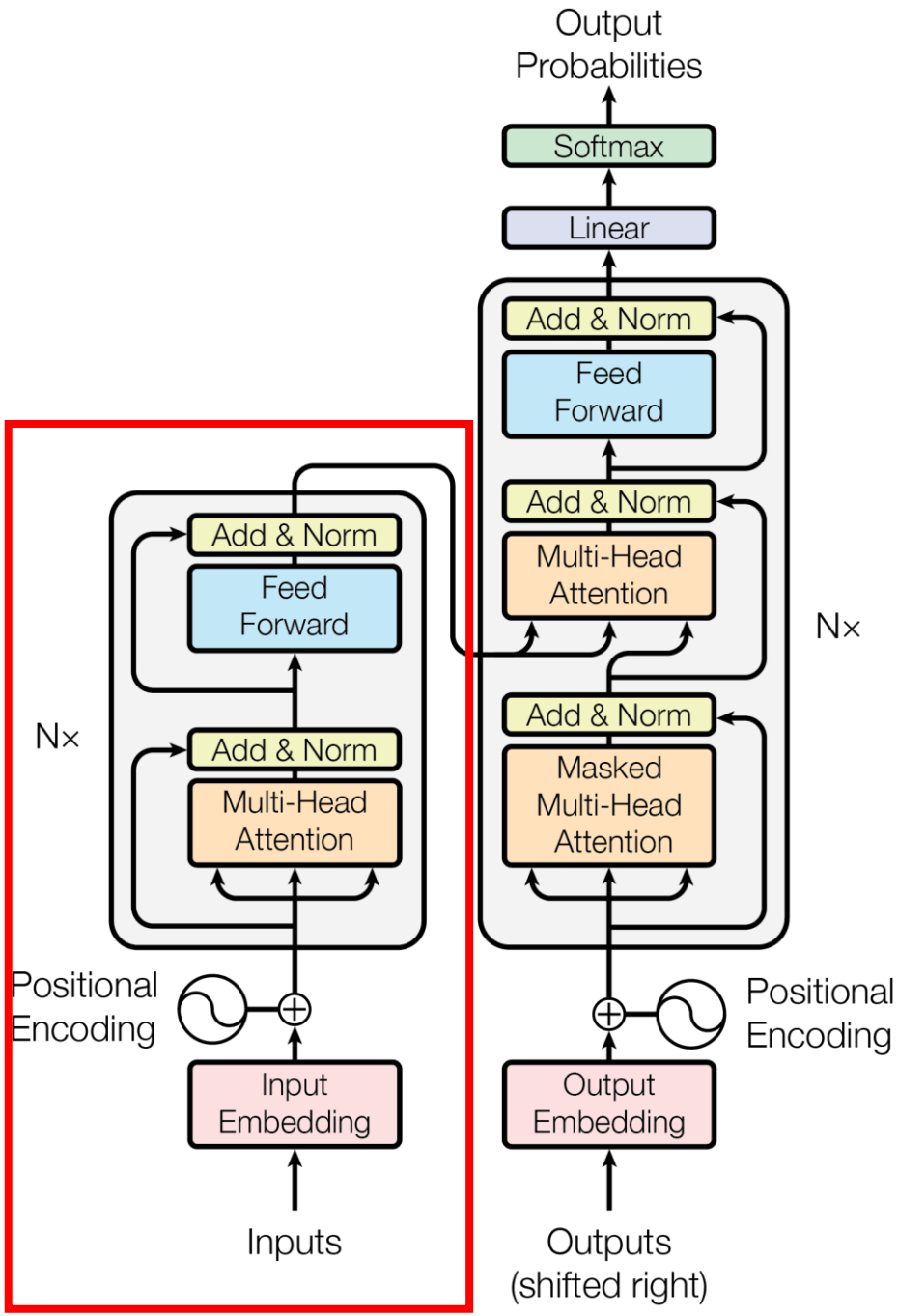
Positional Encoding

- 因為在做 self-attention 中 $a^1 \sim a^4$ 實際上是同時做的，並沒有先後順序，但先後順序對一個句子是重要的資訊
- positional encoding 就是將位置資訊放進去，不同的位置會有不同的 positional vector e^i
- e^i 是手工設定的，最早的 transformer 是用 sin cos 函數
- 但函數的設定尚未有明確的答案



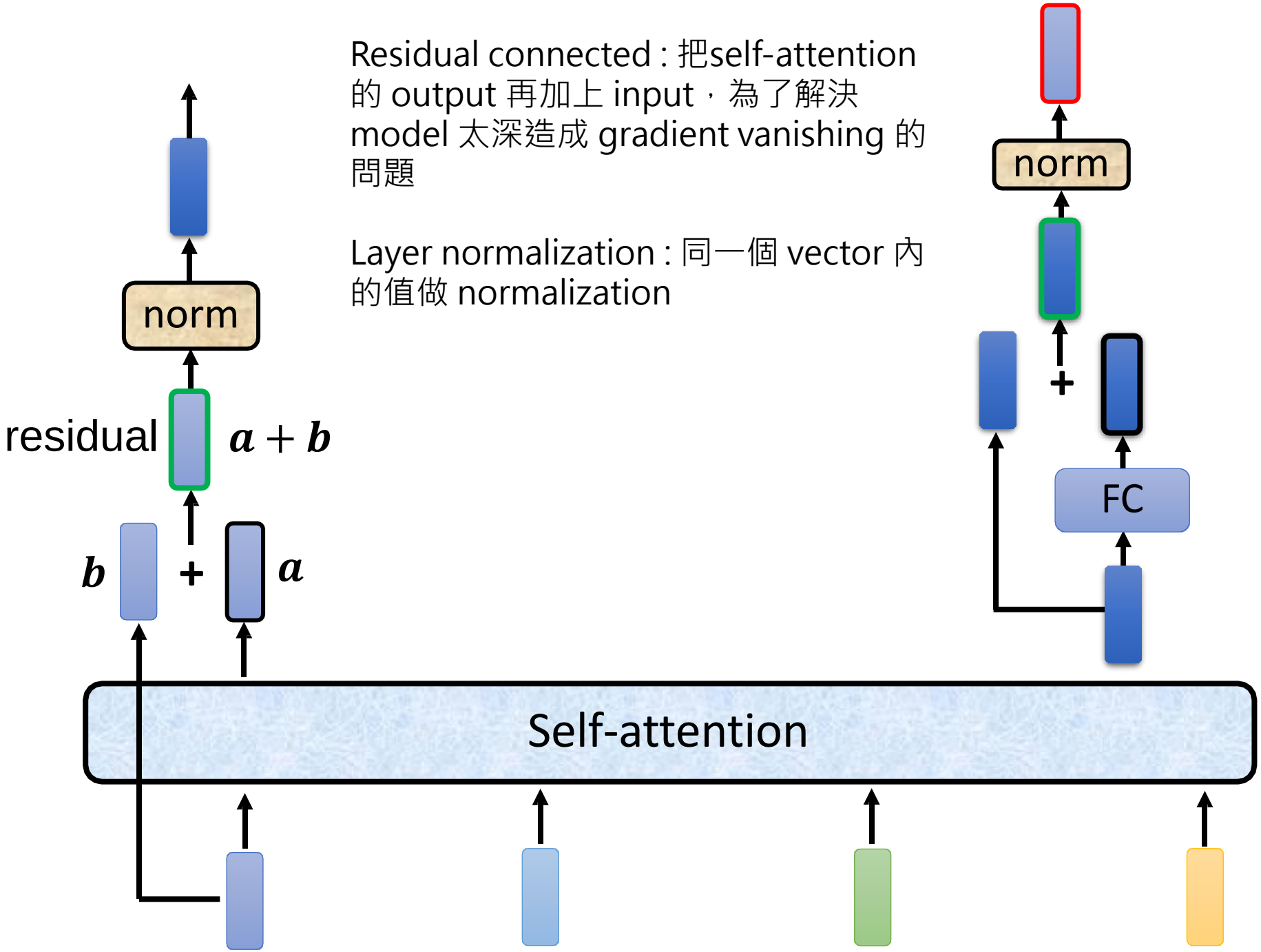
BERT

- 一個完整的 Transformer 是由 encoder 和 decoder 組合
- 而BERT 只有使用 encoder 的部分
- Feed forward : fully connected layer
- Add & Norm : residual connected + layer normalization



Residual connected : 把self-attention 的 output 再加上 input，為了解決 model 太深造成 gradient vanishing 的問題

Layer normalization : 同一個 vector 內的值做 normalization



Code

Precision、recall、F1-score

一個二分類 model 的檢測結果會有下列四種可能性:

True Positive(TP) : 預測值為1，實際值也為1，檢測正確

False Positive(FP) : 預測值為1，但實際值為0，檢測錯誤 type one error

True Negative(TN) : 預測值為0，實際值也為0，檢測正確

False Negative(FN) : 預測值為0，實際值為1，檢測錯誤 type two error

$Precision = \frac{TP}{TP+FP}$ model 猜1的情況下，實際為 1 的比例

$Recall = \frac{TP}{TP+FN}$ 所有值為 1 樣本下，model 猜 1的比例

Precision 和 Recall 是 trade off ，當 precision 高時 recall 低，所以 F1 score 就是為了總和考慮這兩個指標

$$F1\ score = \frac{2 * precision * recall}{precision + recall}$$