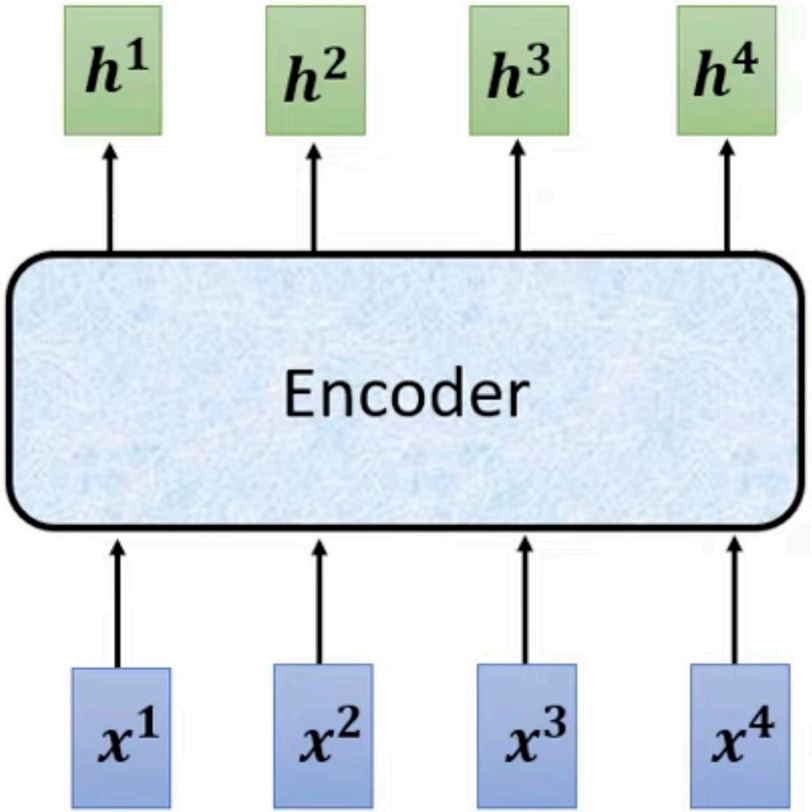
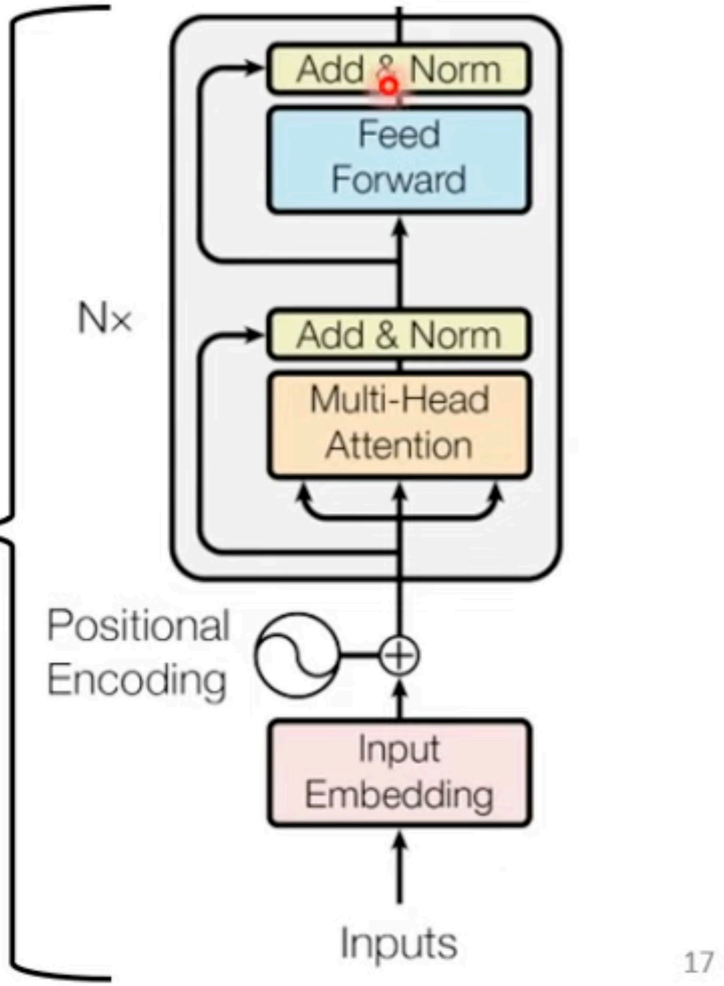
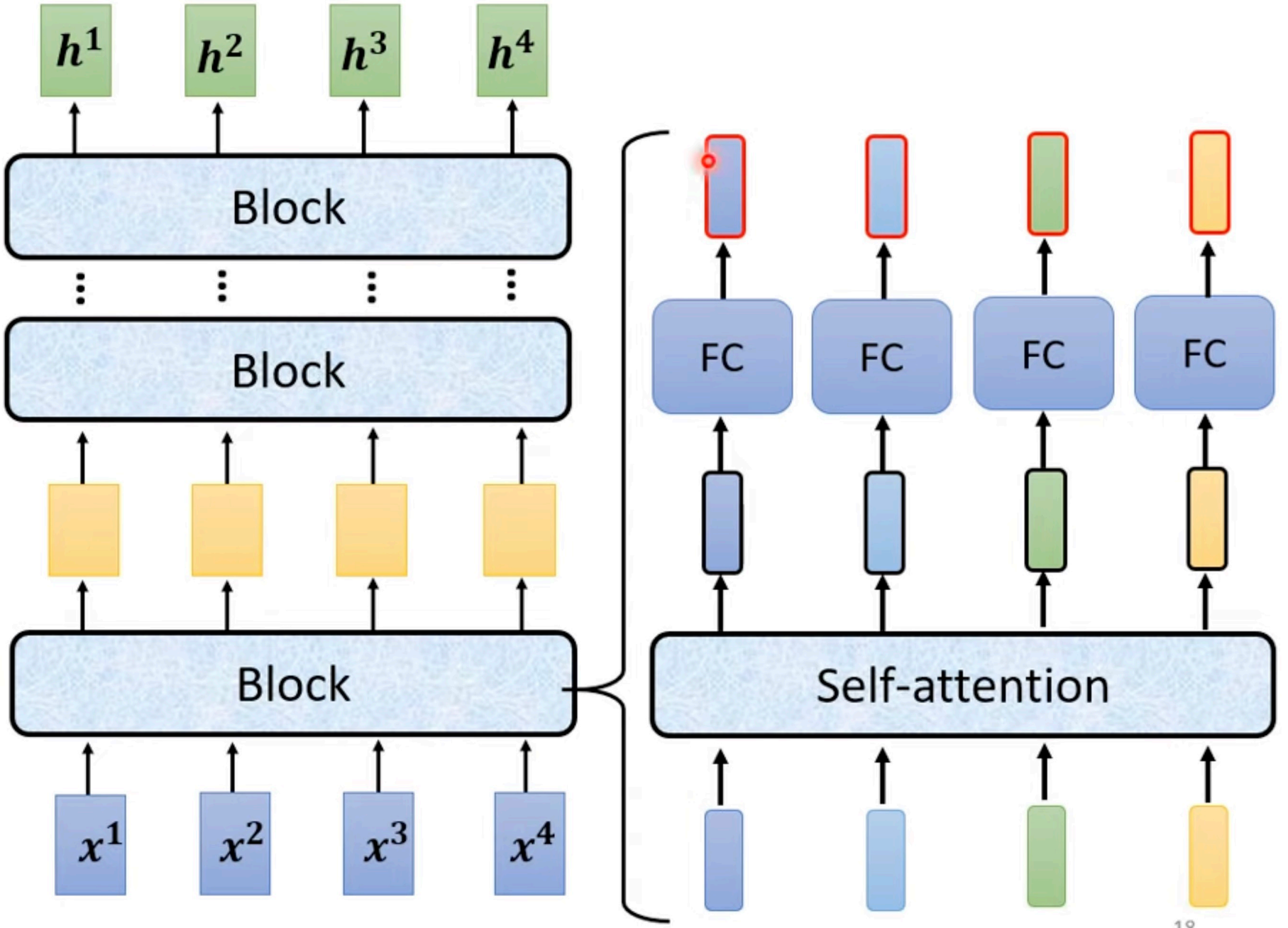
Encoder

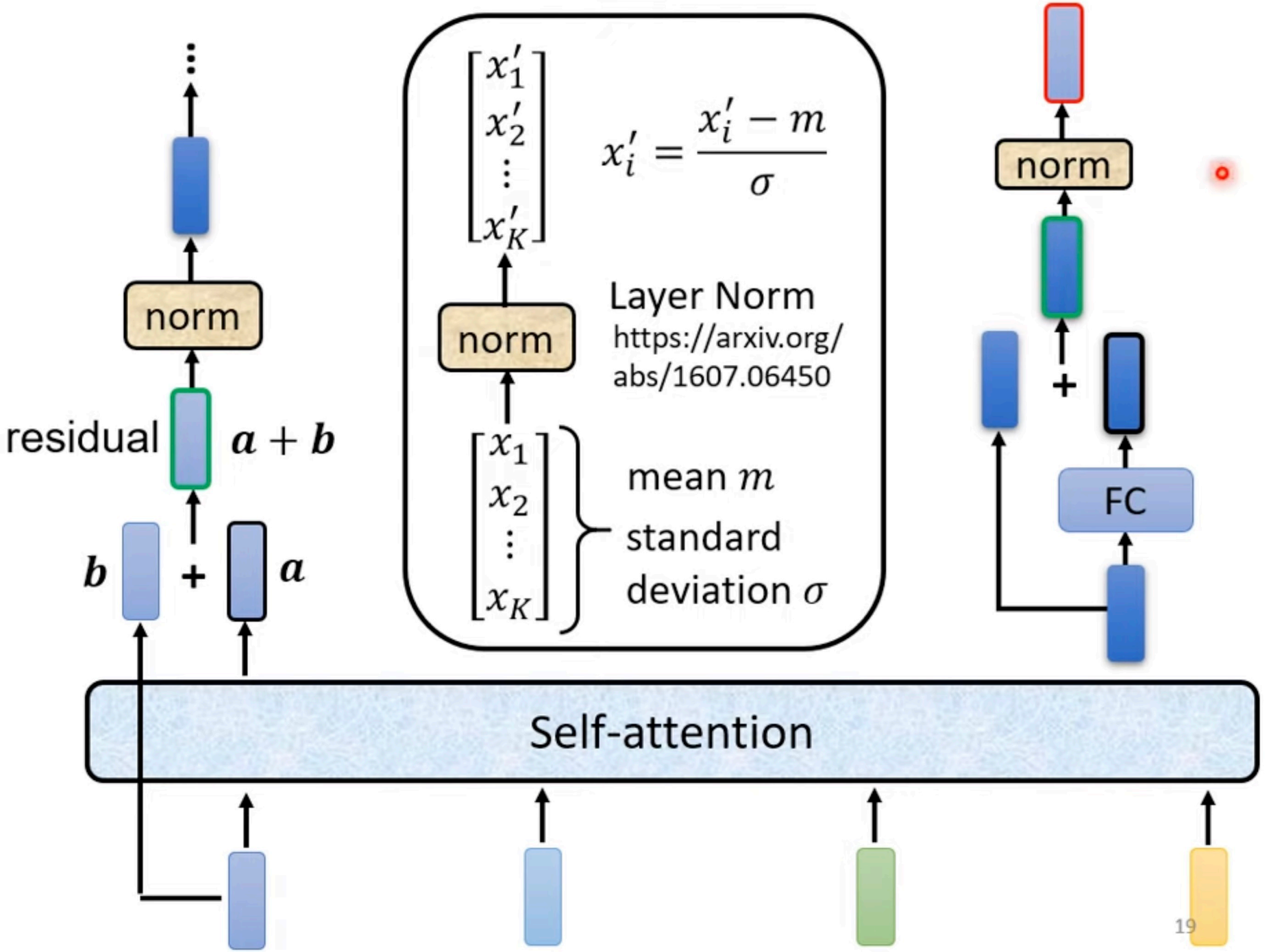
You can use **RNN** or **CNN**.



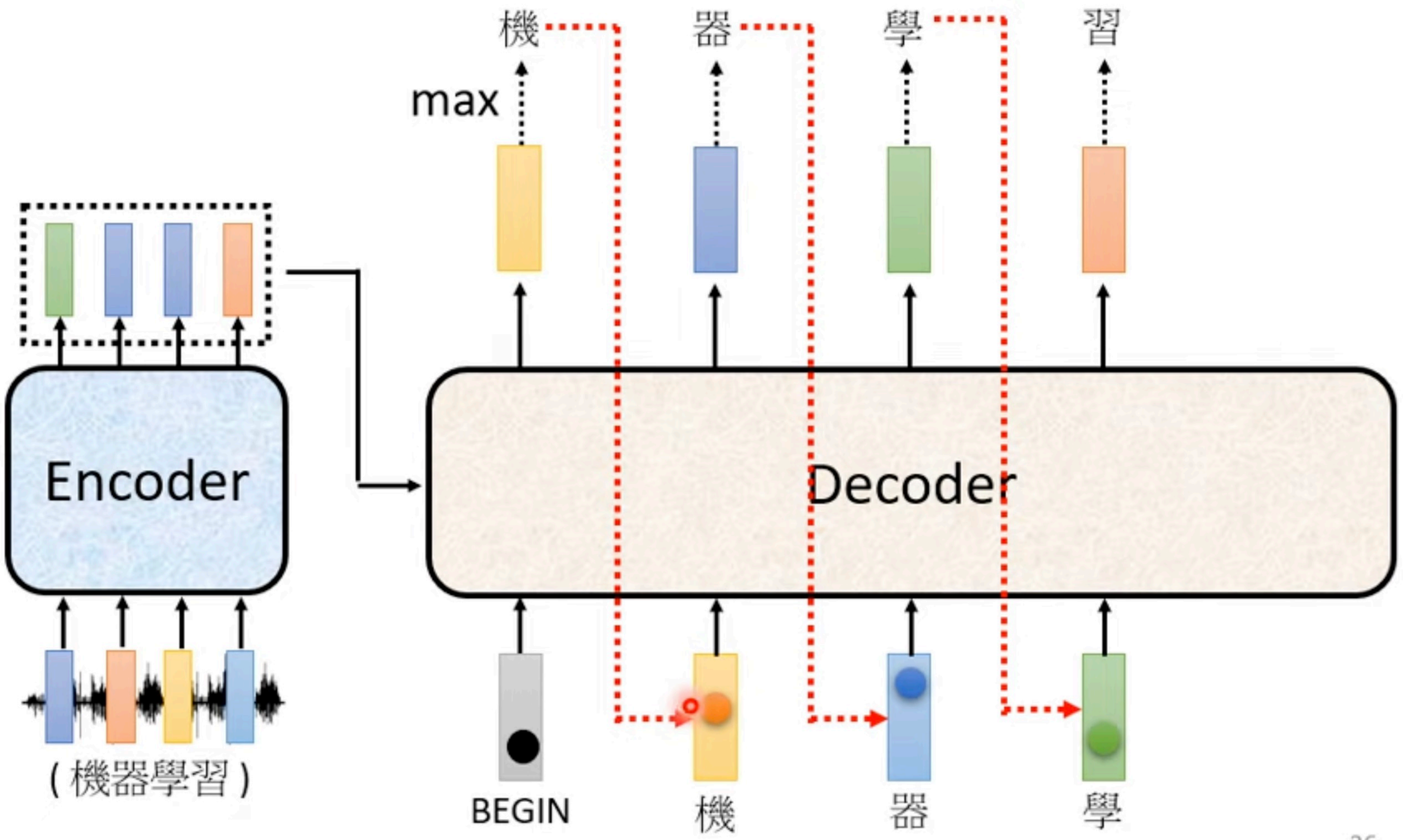
Transformer's Encoder

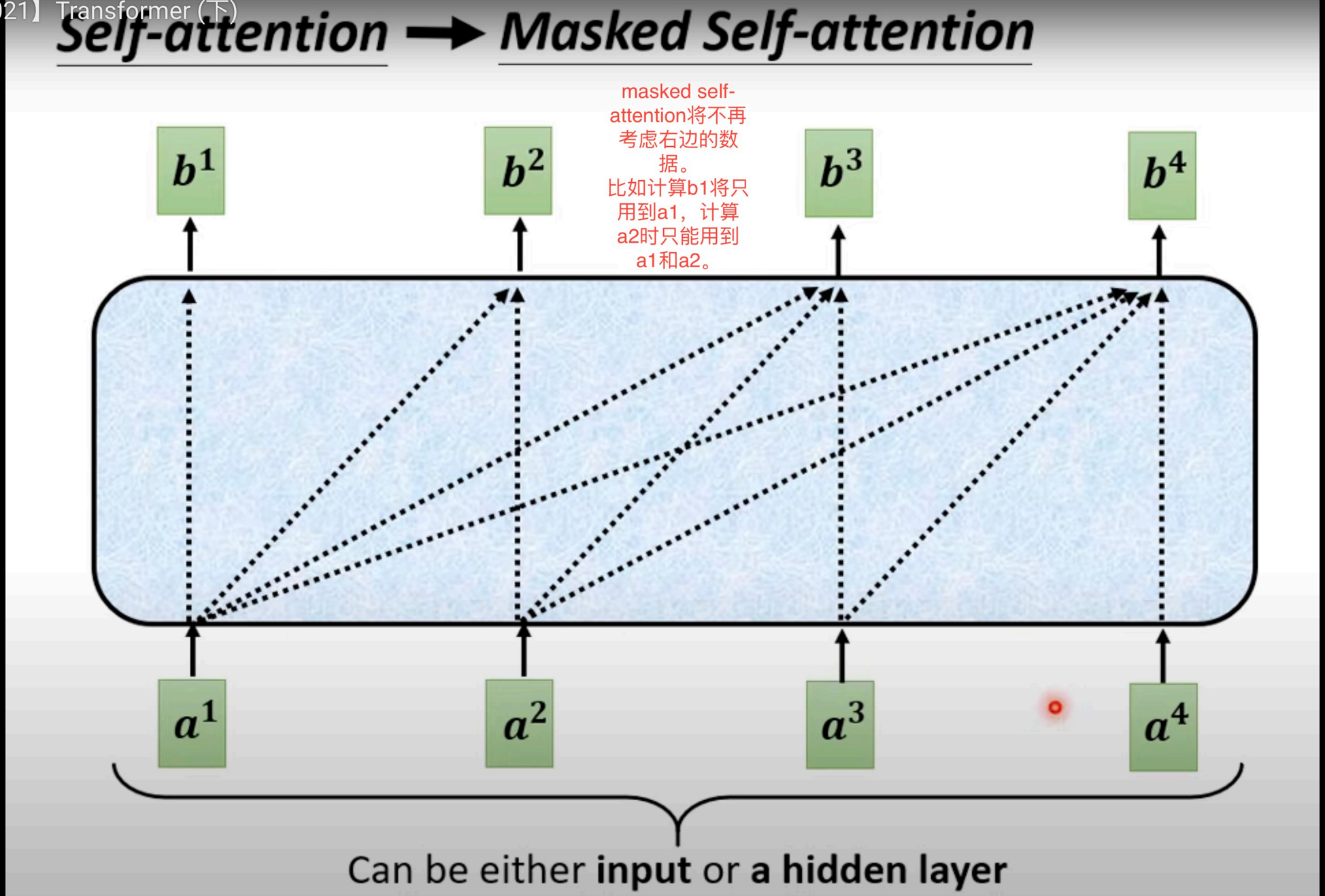



h^1
h^2
h^3
h^4
Block
Block
Block
x^1
x^2
x^3
x^4
FC
FC
FC
FC
Self-attention

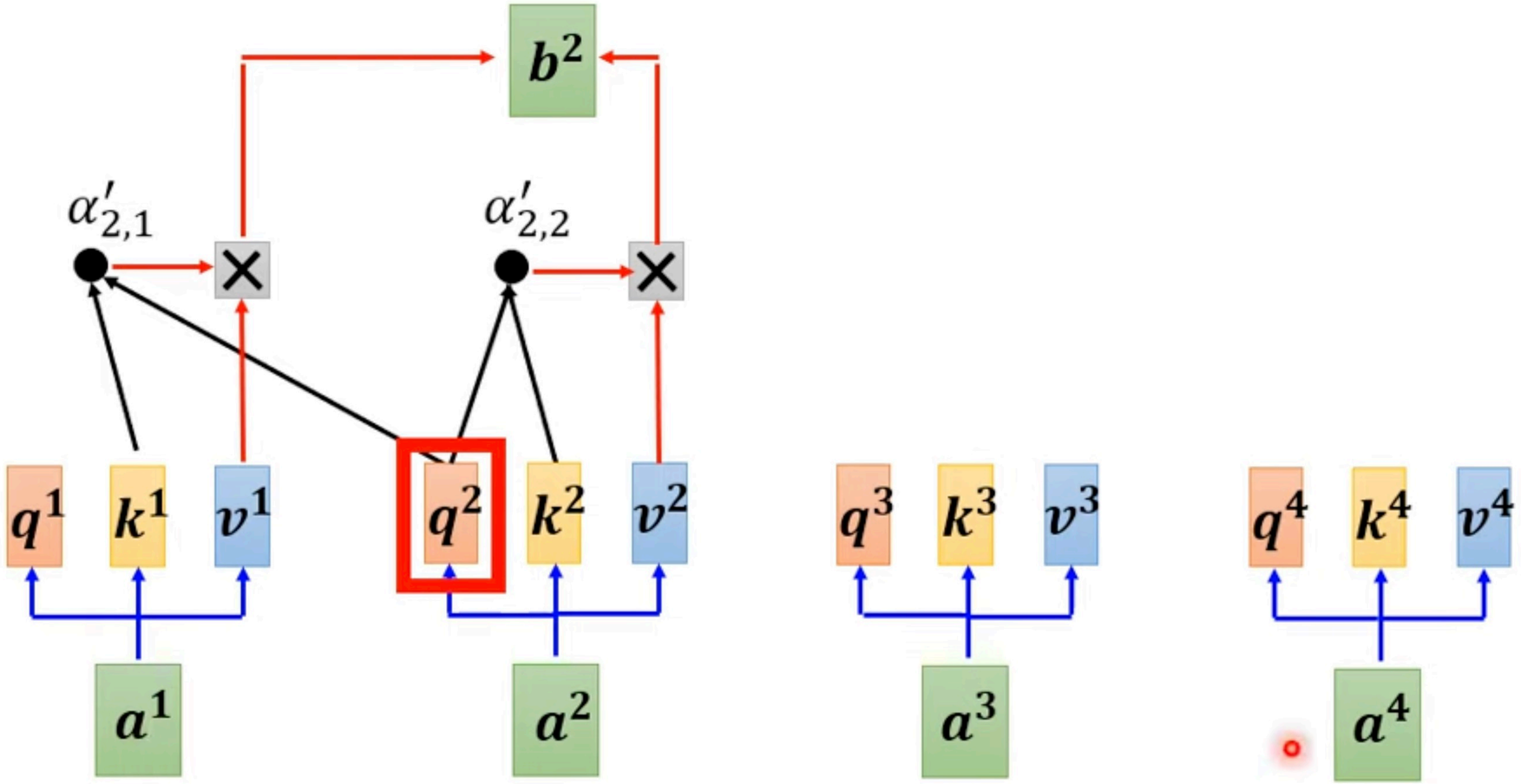

norm
residual
$a + b$
b
+
a
$\begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_K \end{bmatrix}$
$x'_i = \frac{x'_i - m}{\sigma}$
norm
Layer Norm
https://arxiv.org/abs/1607.06450
$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix}$
mean m
standard deviation σ
norm
+
FC
Self-attention

Autoregressive



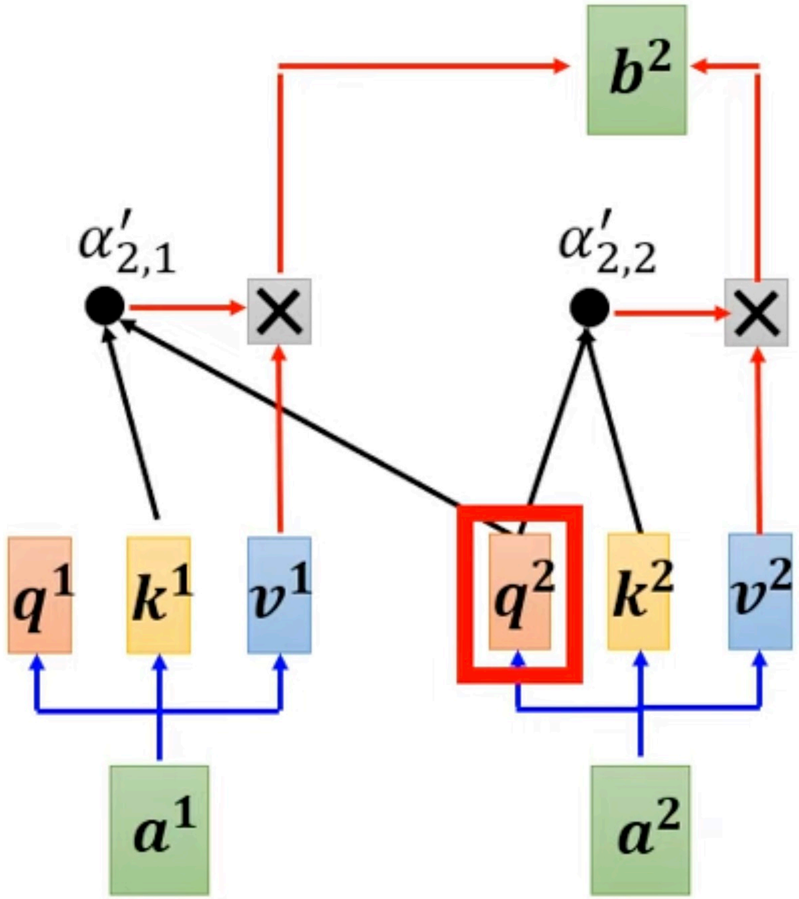

Self-attention → Masked Self-attention
masked self-attention将不再考虑右边的数据。
比如计算b1将只用到a1，计算a2时只能用到a1和a2。
b^1
b^2
b^3
b^4
a^1
a^2
a^3
a^4
Can be either input or a hidden layer
29

Self-attention → Masked Self-attention



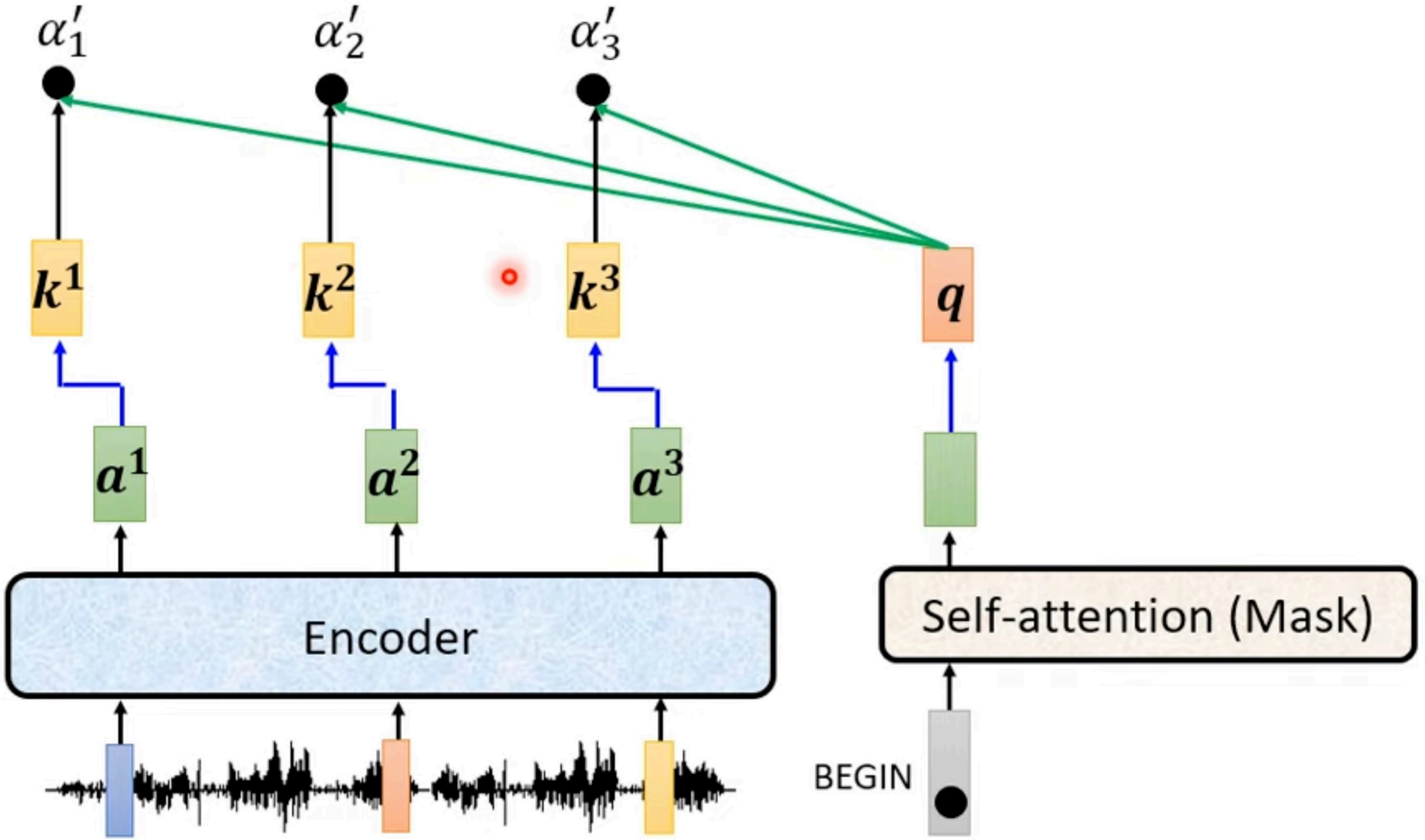
Why masked?

Self-attention → Masked Self-attention

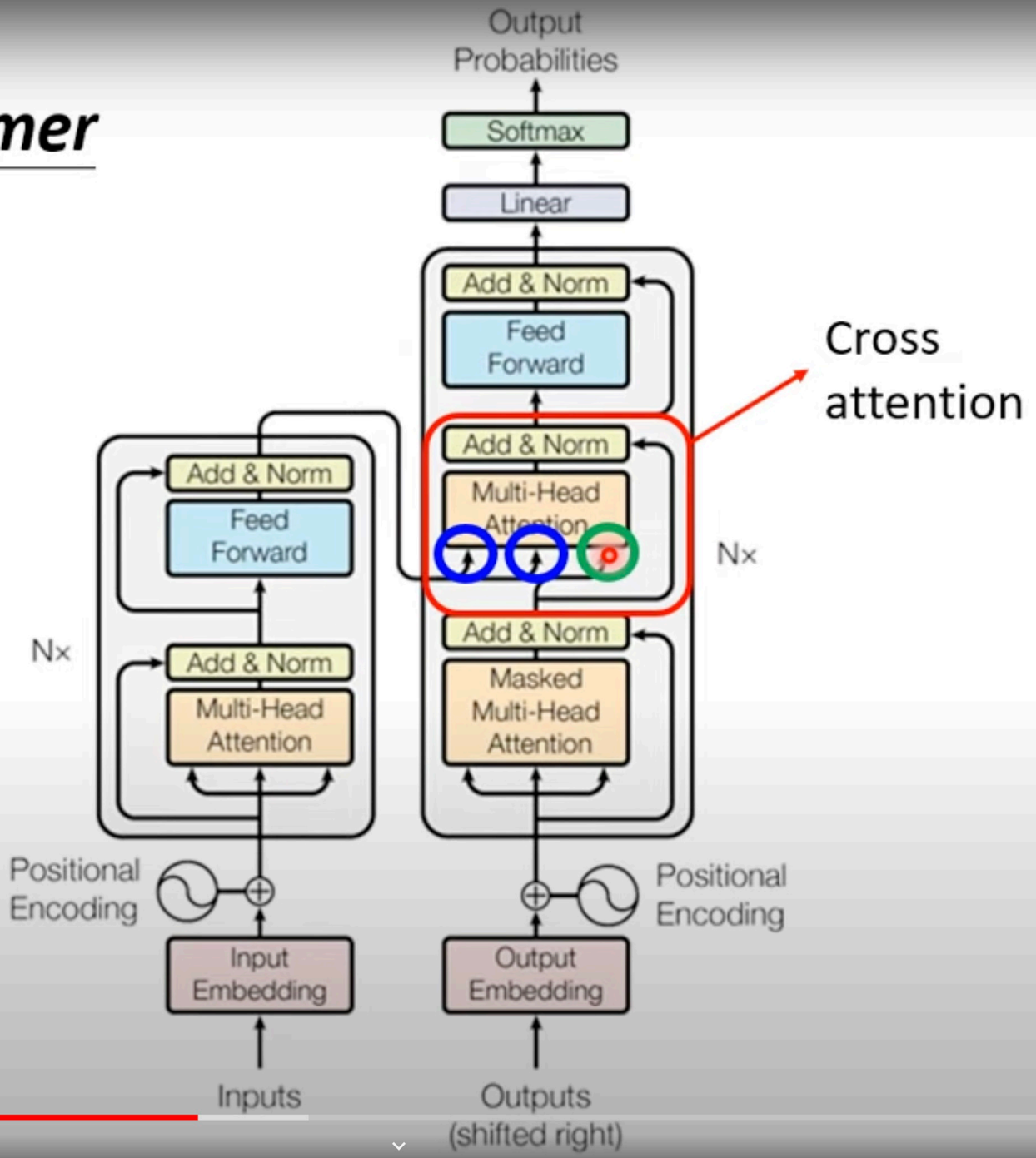


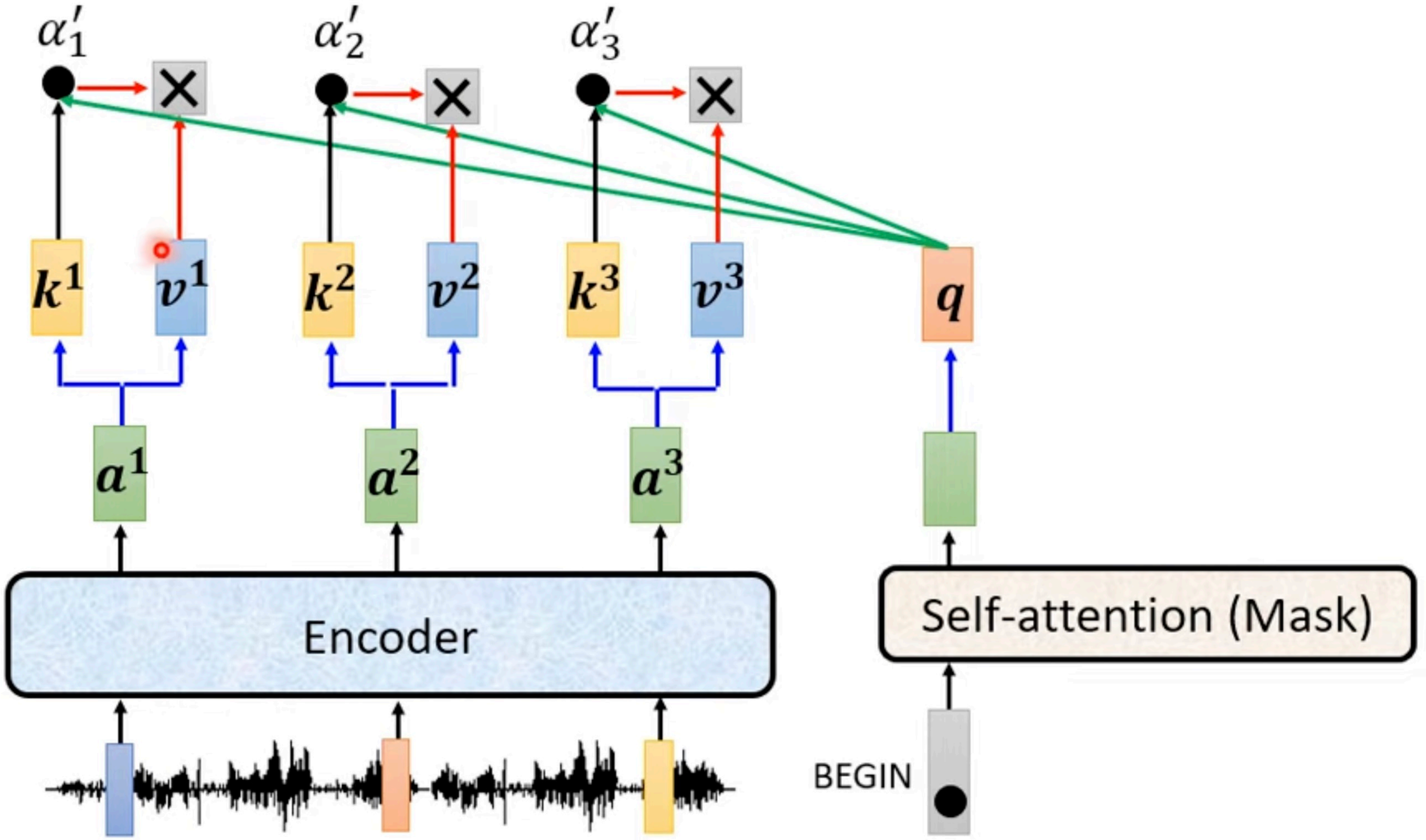
Why masked? :
由于decoder的输入是一个
一个产生，因此在计算b2时
也没有办法得到a3、a4。

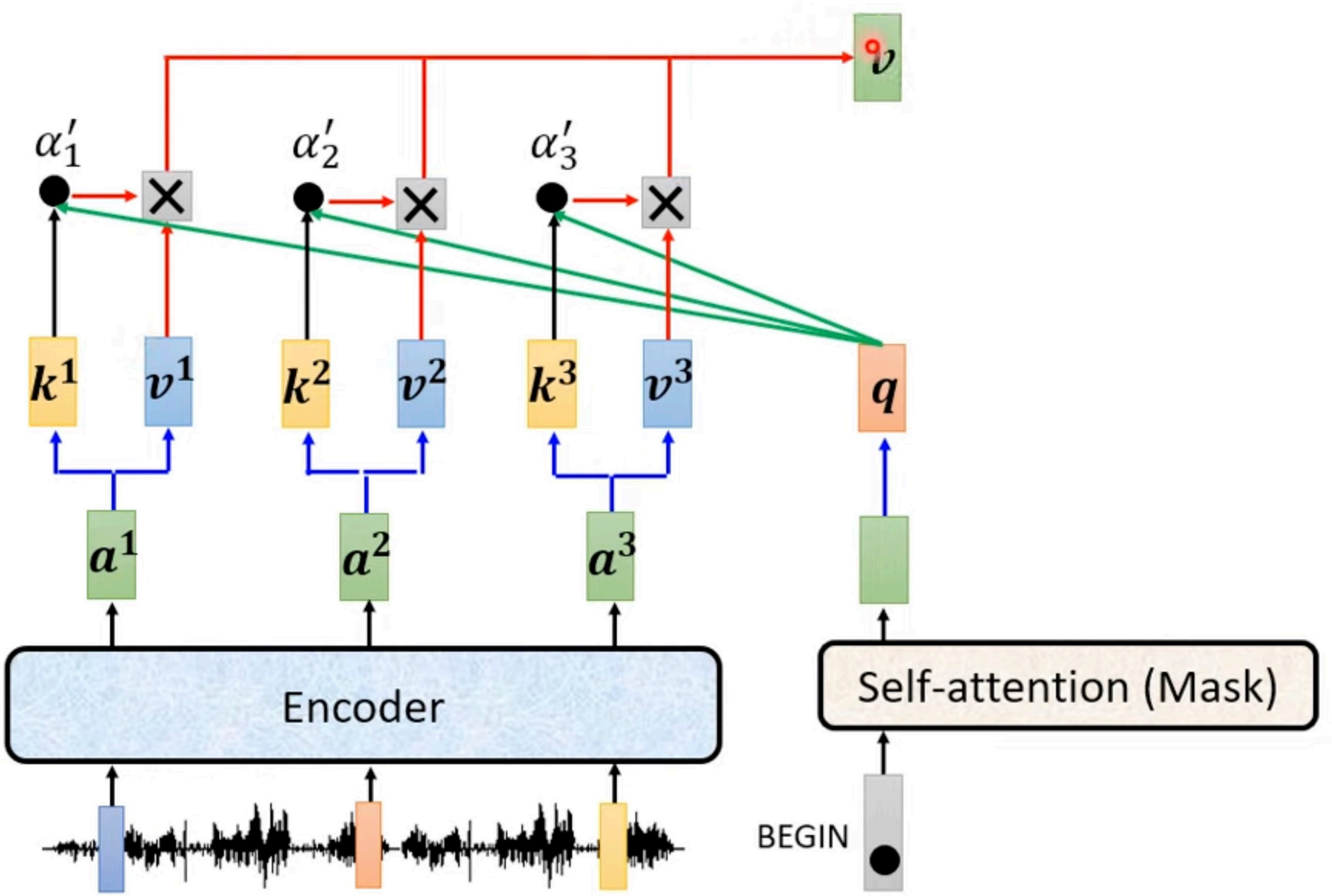
Why masked? Consider how does decoder work

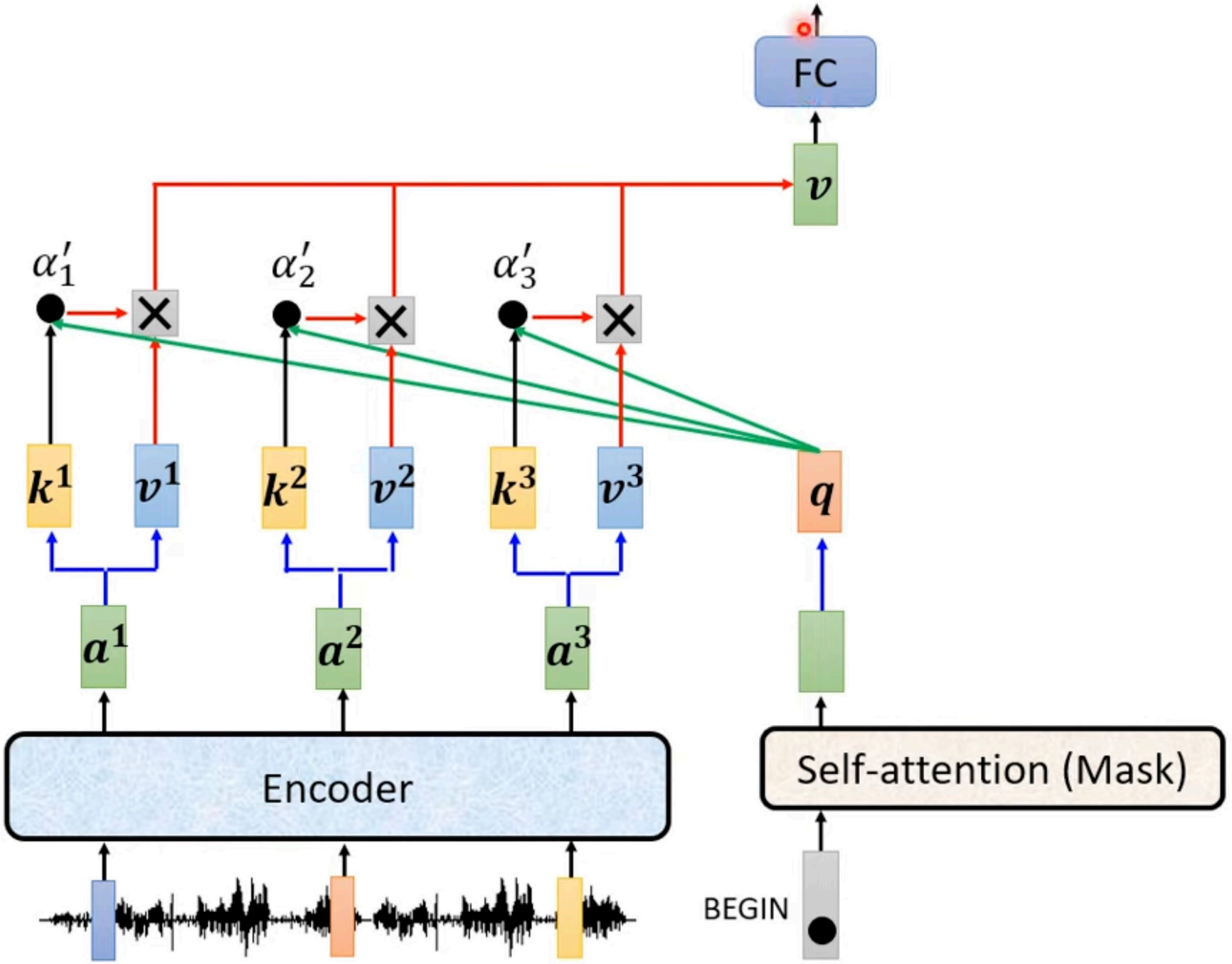

α'_1
α'_2
α'_3
k^1
k^2
k^3
q
a^1
a^2
a^3
Encoder
Self-attention (Mask)
BEGIN

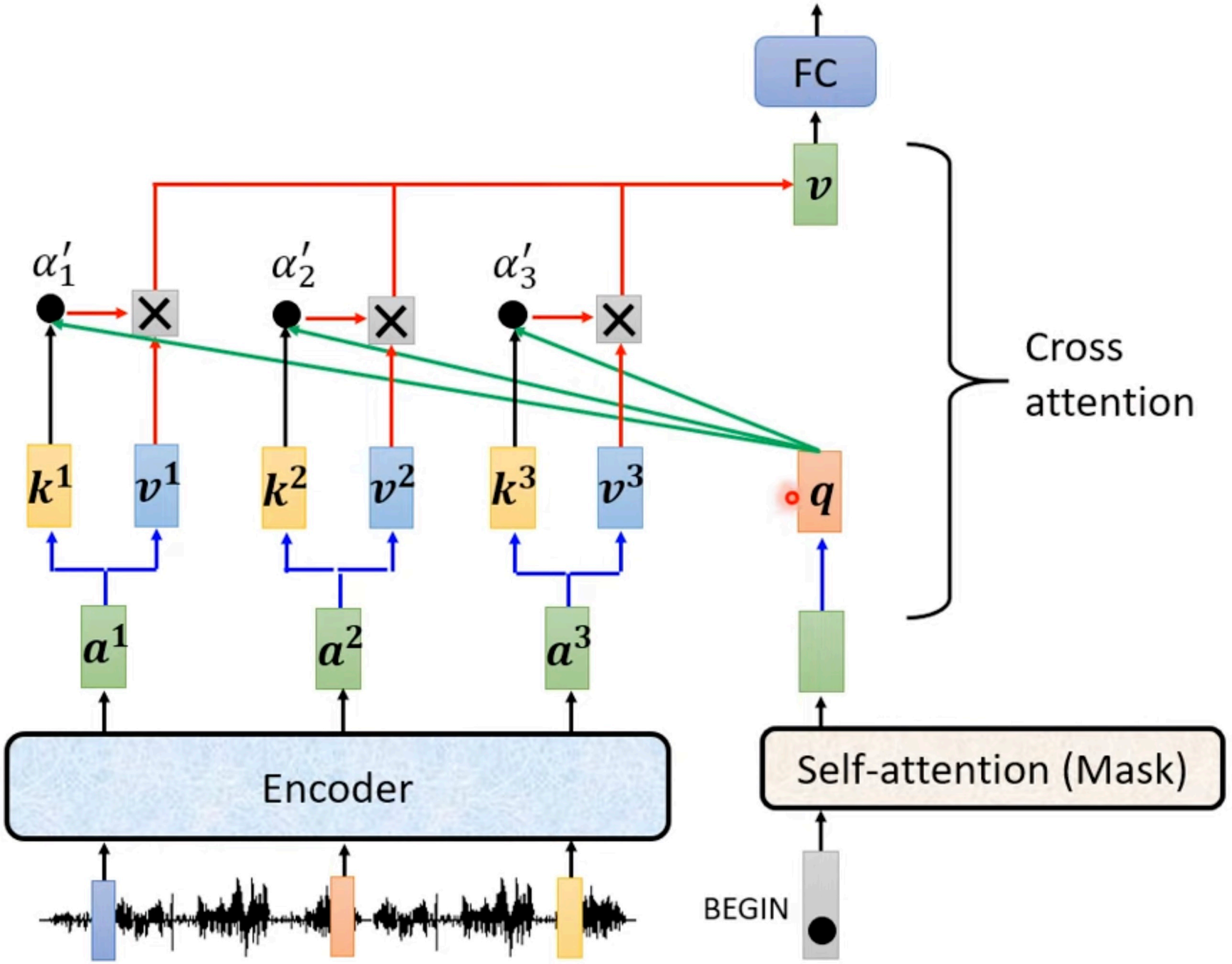
Transformer

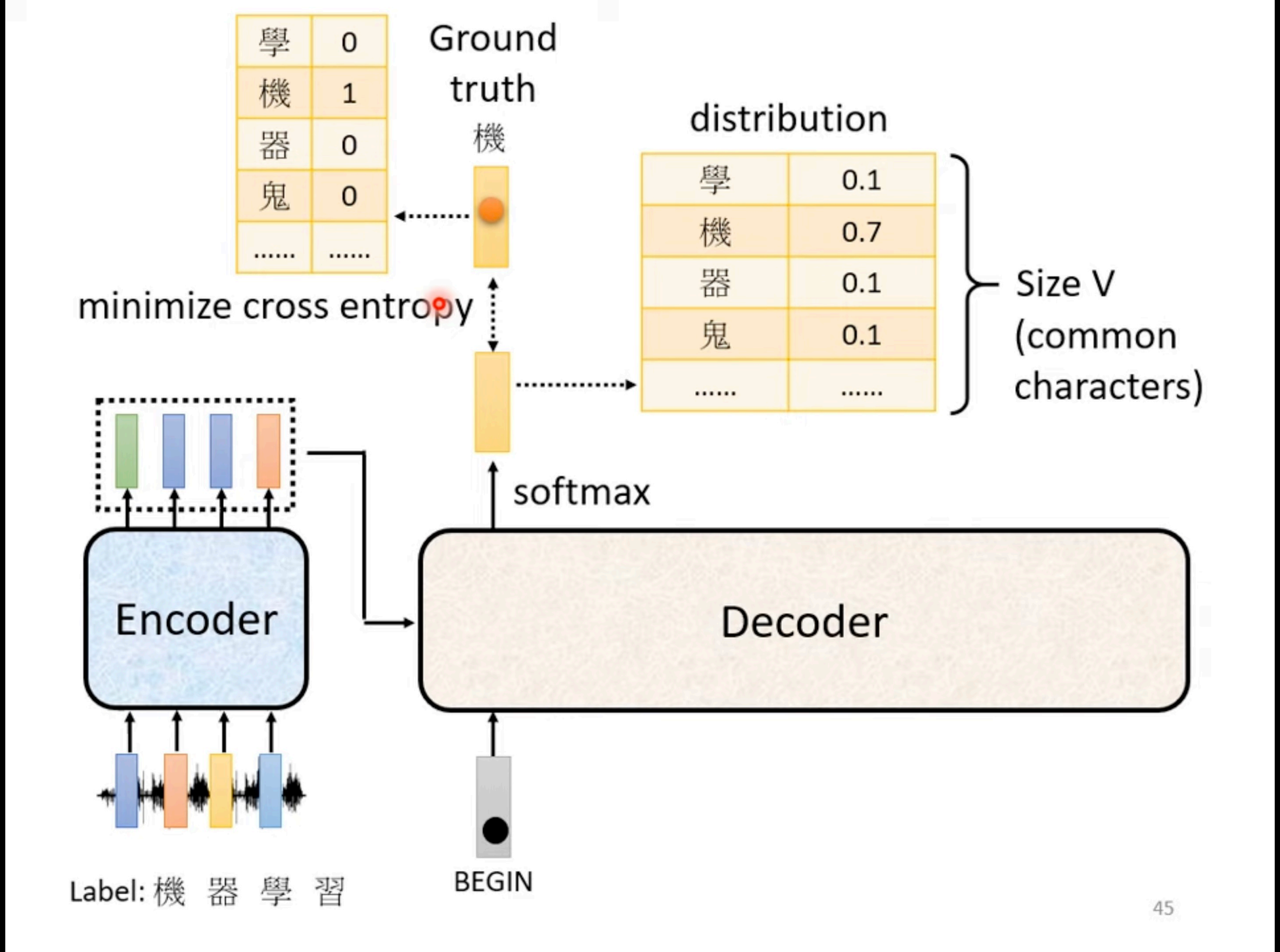


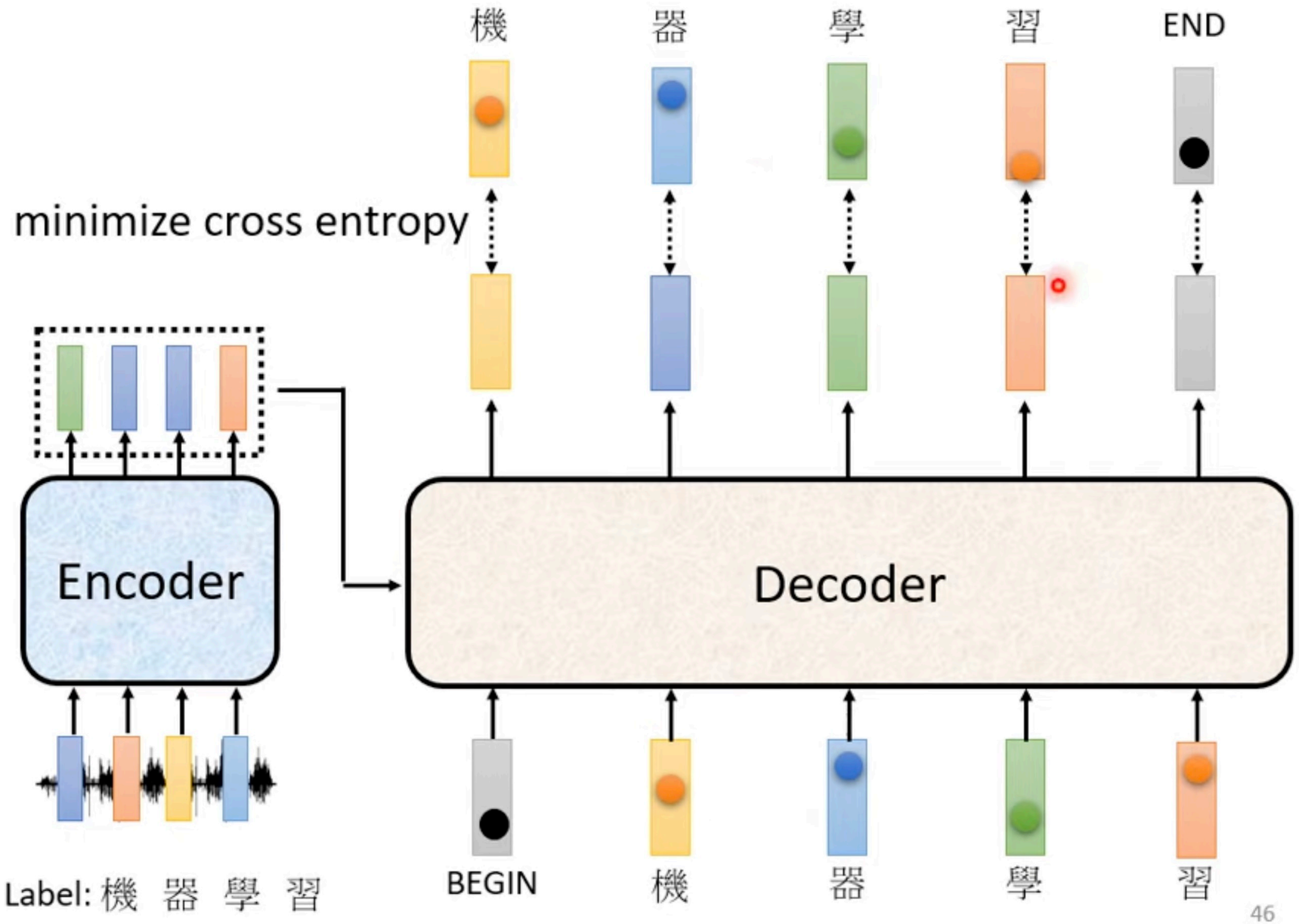

α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q
a^1
a^2
a^3
Encoder
Self-attention (Mask)
BEGIN


v
α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q
a^1
a^2
a^3
Encoder
Self-attention (Mask)
BEGIN

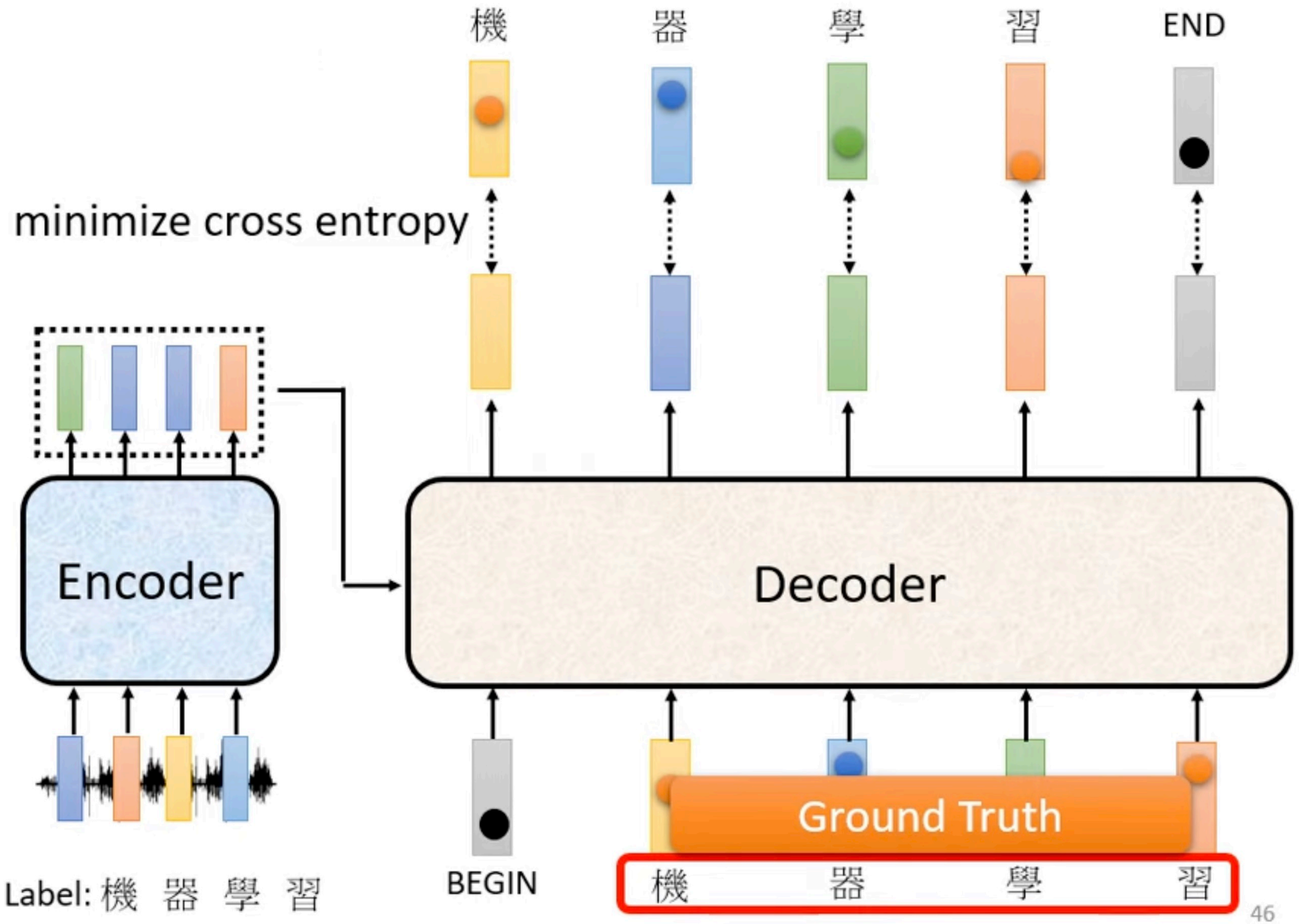

FC
v
α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q
a^1
a^2
a^3
Encoder
Self-attention (Mask)
BEGIN


FC
v
α'_1
α'_2
α'_3
Cross attention
k^1
v^1
k^2
v^2
k^3
v^3
q
a^1
a^2
a^3
Encoder
Self-attention (Mask)
BEGIN


學 0
機 1
器 0
鬼 0
....
Ground truth
機
distribution
學 0.1
機 0.7
器 0.1
鬼 0.1
....
Size V (common characters)
minimize cross entropy
softmax
Encoder
Decoder
BEGIN
Label: 機 器 學 習


機
器
學
習
END
minimize cross entropy
Encoder
Decoder
BEGIN
機
器
學
習
Label: 機 器 學 習

Teacher Forcing: using the ground truth as input.



Optimizing Evaluation Metrics?