# Quiz 9: Self Attention Layer

Please enter your name: *

HM

## The Self Attention Layer: Description

We would like to learn contextual embeddings for the words in "Tom a été entarté cet été".

In order to use the attention mechanism, we define the projections of the embeddings (X^t) onto the d_q-dimensional query space, d_k-dimensional key space and d_v-dimensional value space:

$$\mathbb{R}^{d_q} \ni q^t = W_Q^T X^t$$
$$\mathbb{R}^{d_k} \ni k^t = W_K^T X^t$$
$$\mathbb{R}^{d_v} \ni v^t = W_V^T X^t$$

What is the shape of W_Q

1 point

$W_Q \in \mathbb{R}^{D \times d_q}$

(a)

$W_Q \in \mathbb{R}^{d_q \times D}$

(b)

What condition should be satisfied to calculate the scaled dot product alignment function used in Section 1 1 point

$d_q = d_k$

(a)

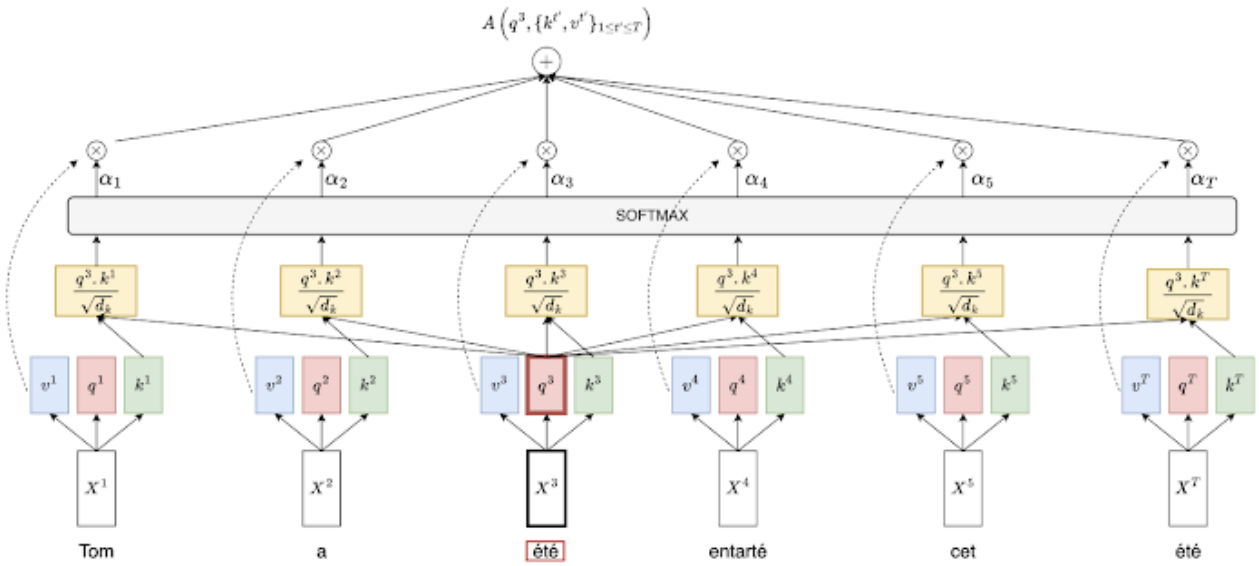
$d_v = d_k$

(b)

$d_q = d_v$

(c)

Let us focus on the first "été" in "Tom a été entarté cet été".



Which expression is correct if we use the scaled dot product as an alignment function and the softmax as the distribution function ? 1 point

$$A\left(q^3, \{k^{t'}, v^{t'}\}_{1\leq t'\leq T}\right) = \sum_{t=1}^{T} \frac{\exp(\frac{q^3.k^t}{\sqrt{d_k}})}{\sum_{t'=1}^{T} \exp(\frac{q^3.k^{t'}}{\sqrt{d_k}})} v^t$$

(a)

$$A\left(q^3, \{k^{t'}, v^{t'}\}_{1\leq t'\leq T}\right) = \sum_{t=1}^{T} \frac{\exp(\frac{q^t.k^t}{\sqrt{d_k}})}{\sum_{t'=1}^{T} \exp(\frac{q^t.k^{t'}}{\sqrt{d_k}})} v^t$$

(b)

$$A\left(q^3, \{k^{t'}, v^{t'}\}_{1\leq t'\leq T}\right) = \sum_{t=1}^{T} \frac{\exp(q^t.k^t)}{\sum_{t'=1}^{T} \exp(q^t.k^{t'})} v^t$$

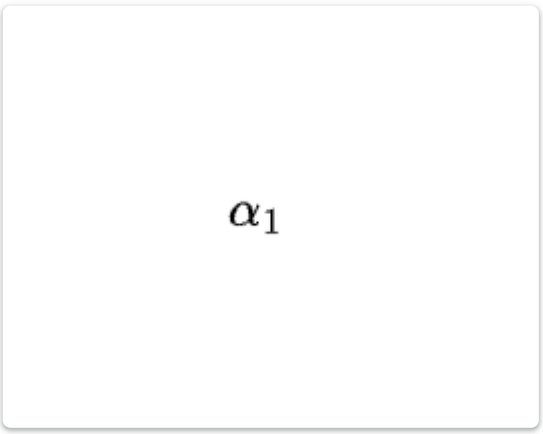
(c)

What is the interpretation of the attention vector: 1 point

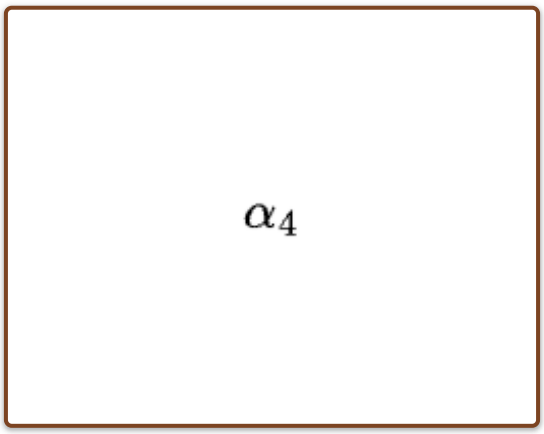
$$A\left(q^3, \{k^{t'}, v^{t'}\}_{1\leq t'\leq T}\right)$$

- It represents the contextual embedding of the word "été" in the first position
- It represents the contextual embedding of the word "été" in the second position
- It represents the contextual embedding of the word "été" in both positions

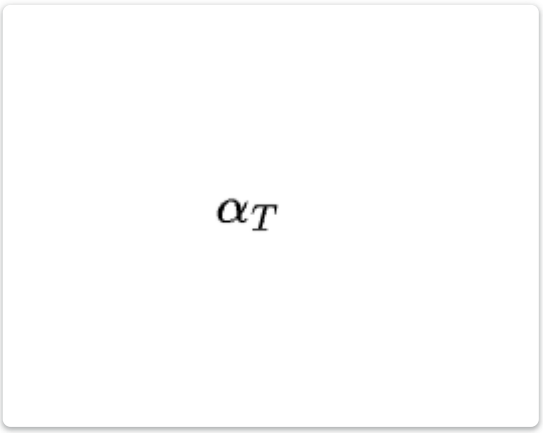
Suppose the query q^3 represents the question "What happened to Tom ?". Which attention weight will have the highest value ? 1 point




(a)




(b)




(c)

Suppose the query q^3 represents the question "When does that happen to Tom ?". Which attention weight will be the highest ? 1 point

$\alpha_1$

(a)

$\alpha_4$

(b)

$\alpha_T$

(c)

Let N be the batch size. After applying the Self Attention Layer to the whole batch, what is the change in the tensor shape ? 1 point

$(N, T, D) \rightarrow (N, T, d_q)$

(a)

$(N, T, D) \rightarrow (N, T, d_v)$

(b)

$(N, T, D) \rightarrow (N, d_k)$

(c)

The Learning Process

What are the parameters of the previous layer ? 1 point

$W_Q, W_K, W_V$

(a)

$\{q^t, k^t, v^t\}_{1 \leq t \leq T}$

(b)

$\{q_i, k_i, v_i\}_{1 \leq i \leq n}$

(c)

What is the total number of parameters ? 1 point

$D(d_q + d_k + d_v)$

(a)

$d_q + d_k + d_v$

(b)

Does the Self Attention Layer take into account the sequentiality of the data ? 1 point

Yes

No