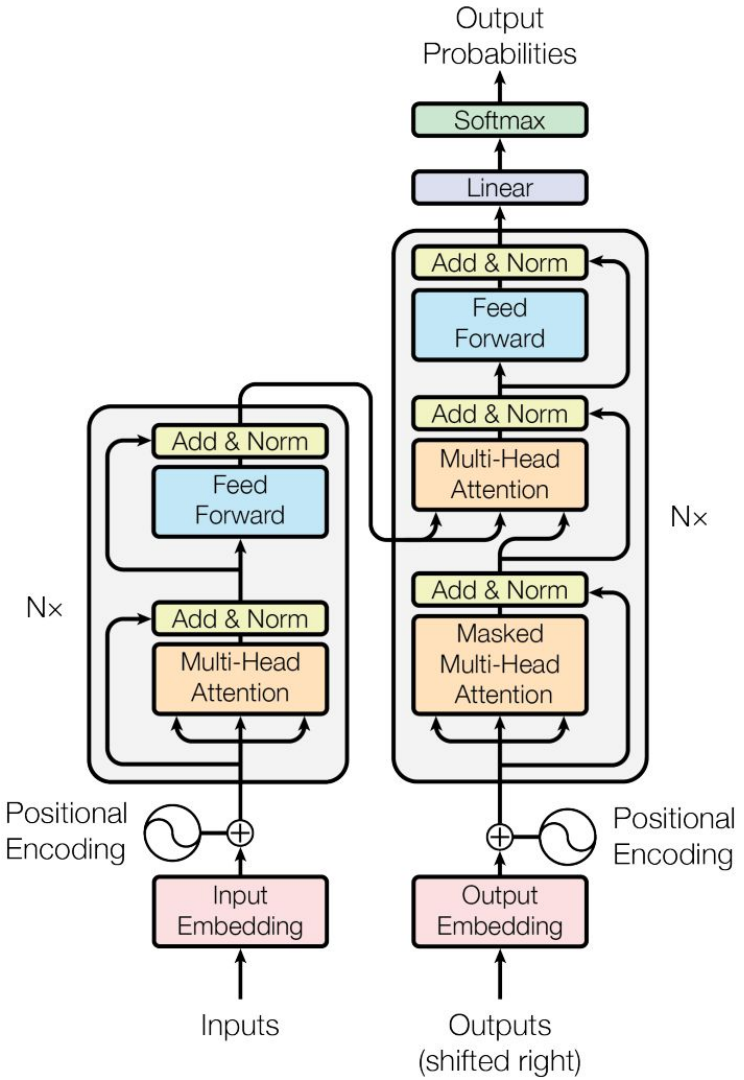
Transformers positional encoding

Orlando Ramos Flores

Transformer

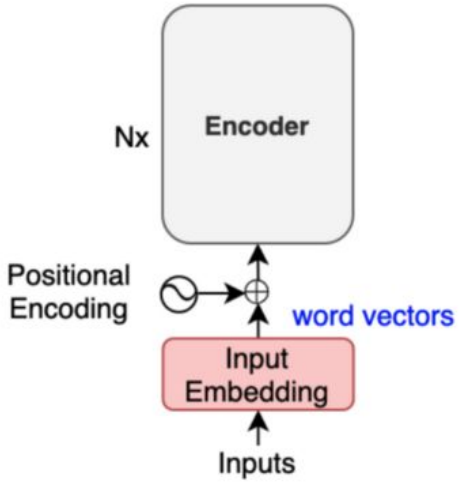
- Arquitectura del Tranformer del paper “Attention is all you need”



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Positional encoding (codificaciones posicionales)

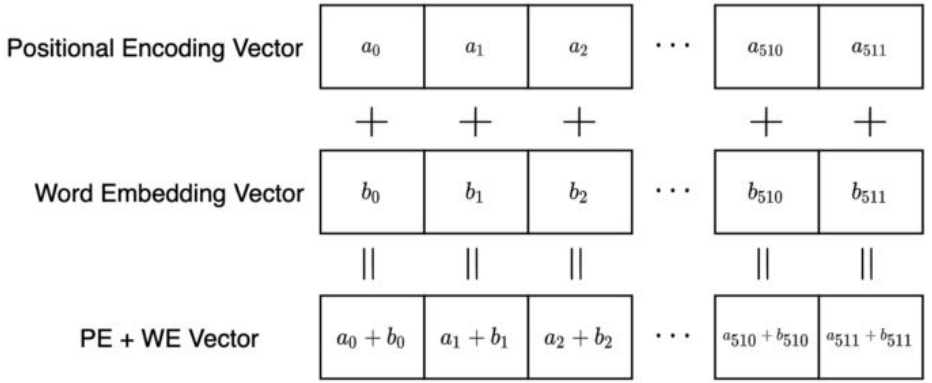
- "*Dado que nuestro modelo no contiene recurrencia ni convolución, para que el modelo haga uso del orden de la secuencia, debemos inyectar cierta información sobre la posición relativa o absoluta de los tokens en la secuencia.*"
- *Agregamos "**positional encoding**" a los embeddings de entrada en la parte inferior de las pilas de **encoder** y **decoder**.*



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Positional encoding

- El Transformer base utiliza word embeddings de 512 dimensiones (elementos).
- Por lo tanto, el **positional encoding** también tiene 512 elementos, para que se pueda sumar un vector de word embeddings y un vector de positional encoding mediante la suma de elementos (element-wise).
- El modelo puede aprender a usar la información posicional sin confundirse con la información de los embeddings (semántica).



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Positional encoding

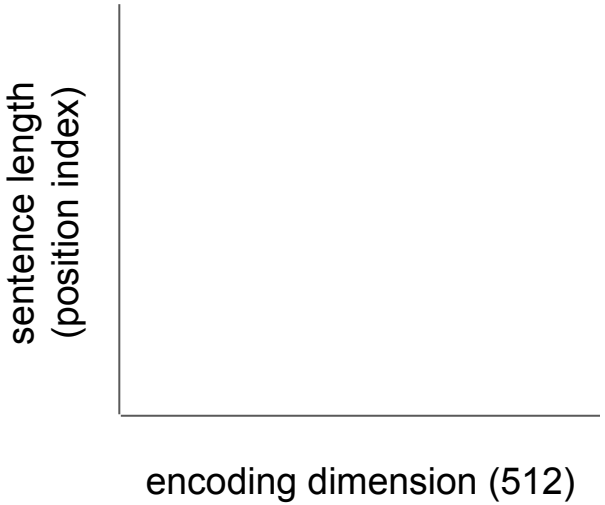
- Para calcular positional encodings se usaron funciones de seno y coseno de diferentes frecuencias.
- Donde pos es la posición e i es la dimensión.
- Es decir, cada dimensión del positional encoding corresponde a una sinusoide.
- Las longitudes de onda forman una progresión geométrica de 2π a 10000 · 2π.
- d_{model}=512 cada token es representado como un vector de 512 dimensiones (del elemento 0 al elemento 511).

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

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Positional encoding



$$PE(pos, 0) = \sin\left(\frac{pos}{10000^{\frac{0}{512}}}\right)$$
$$PE(pos, 1) = \cos\left(\frac{pos}{10000^{\frac{0}{512}}}\right)$$
$$PE(pos, 2) = \sin\left(\frac{pos}{10000^{\frac{2}{512}}}\right)$$
$$PE(pos, 3) = \cos\left(\frac{pos}{10000^{\frac{2}{512}}}\right)$$
$$PE(pos, 4) = \sin\left(\frac{pos}{10000^{\frac{4}{512}}}\right)$$
$$PE(pos, 5) = \cos\left(\frac{pos}{10000^{\frac{4}{512}}}\right)$$
$$\vdots$$
$$PE(pos, 510) = \sin\left(\frac{pos}{10000^{\frac{510}{512}}}\right)$$
$$PE(pos, 511) = \cos\left(\frac{pos}{10000^{\frac{510}{512}}}\right)$$

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Positional encoding

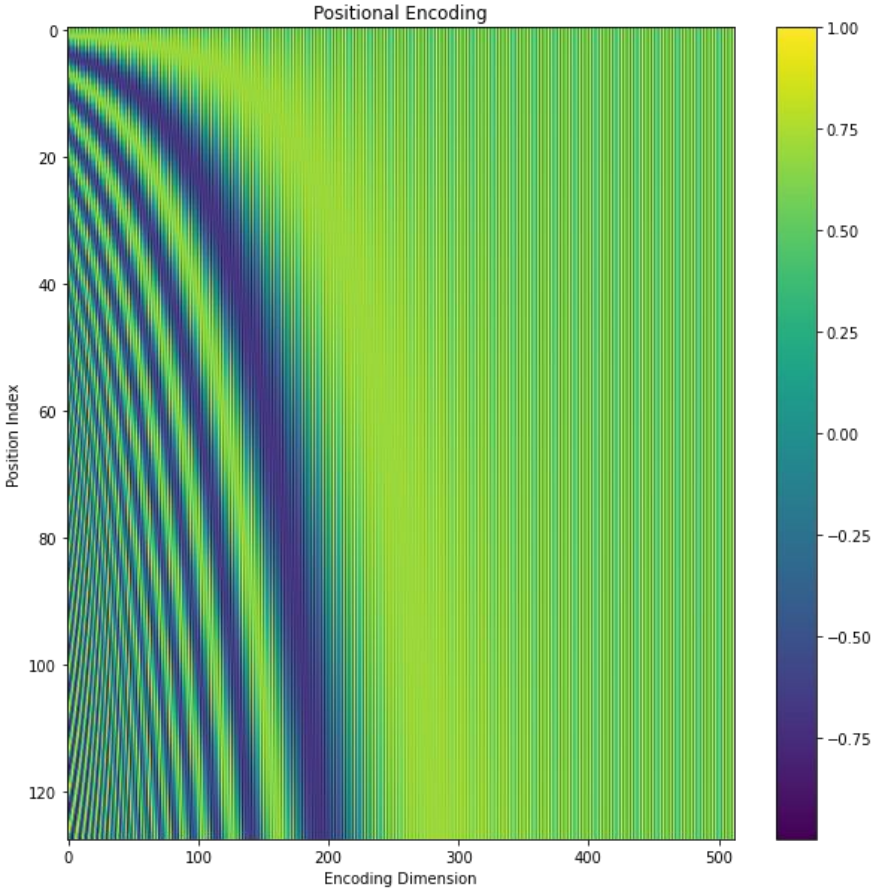
$$
\begin{aligned}
PE(0) &= \left(\sin\left(\frac{0}{10000^{\frac{0}{512}}}\right), \cos\left(\frac{0}{10000^{\frac{0}{512}}}\right), \sin\left(\frac{0}{10000^{\frac{2}{512}}}\right), \cos\left(\frac{0}{10000^{\frac{2}{512}}}\right), \ldots, \sin\left(\frac{0}{10000^{\frac{510}{512}}}\right), \cos\left(\frac{0}{10000^{\frac{510}{512}}}\right)\right) \\
&= \left(\sin(0), \quad \cos(0), \quad \sin(0), \quad \cos(0), \quad \ldots, \quad \sin(0), \quad \cos(0)\right) \\
&= (0, \quad 1, \quad 0, \quad 1, \quad \ldots, \quad 0, \quad 1)
\end{aligned}
$$

$$
\begin{aligned}
PE(1) &= \left(\sin\left(\frac{1}{10000^{\frac{0}{512}}}\right), \cos\left(\frac{1}{10000^{\frac{0}{512}}}\right), \sin\left(\frac{1}{10000^{\frac{2}{512}}}\right), \cos\left(\frac{1}{10000^{\frac{2}{512}}}\right), \ldots, \sin\left(\frac{1}{10000^{\frac{510}{512}}}\right), \cos\left(\frac{1}{10000^{\frac{510}{512}}}\right)\right) \\
&= \left(0.8414, \quad 0.5403, \quad 0.8218, \quad 0.5696, \quad \ldots, \quad 0.0001, \quad 0.9999\right)
\end{aligned}
$$

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Positional encoding

- longitud máxima de las secuencias = 128, d_model = 512
- En la Figura. Los valores están entre -1 y 1 ya que son de funciones seno y coseno.
- Los colores más oscuros indican valores más cercanos a -1, y los colores más brillantes (amarillos) están más cerca de 1.
- Los colores verdes indican valores intermedios. Por ejemplo, los colores verde oscuro indican valores alrededor de 0.



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Positional encoding

- Implementar una función en Python para calcular el Positional Encoding.