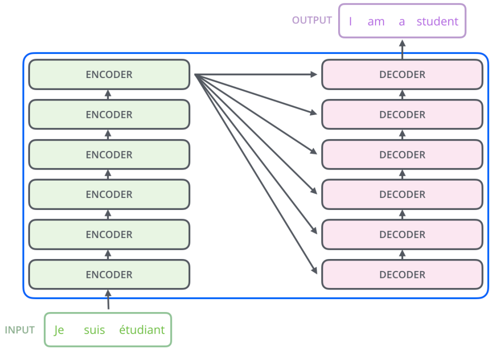
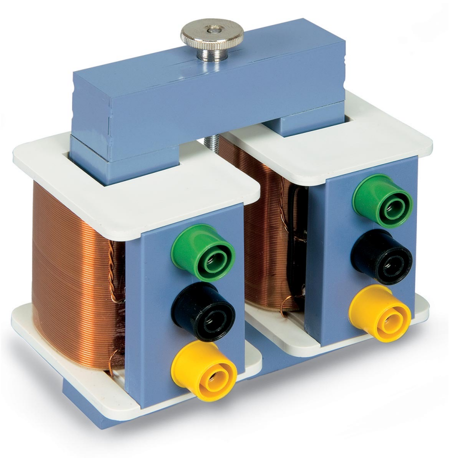
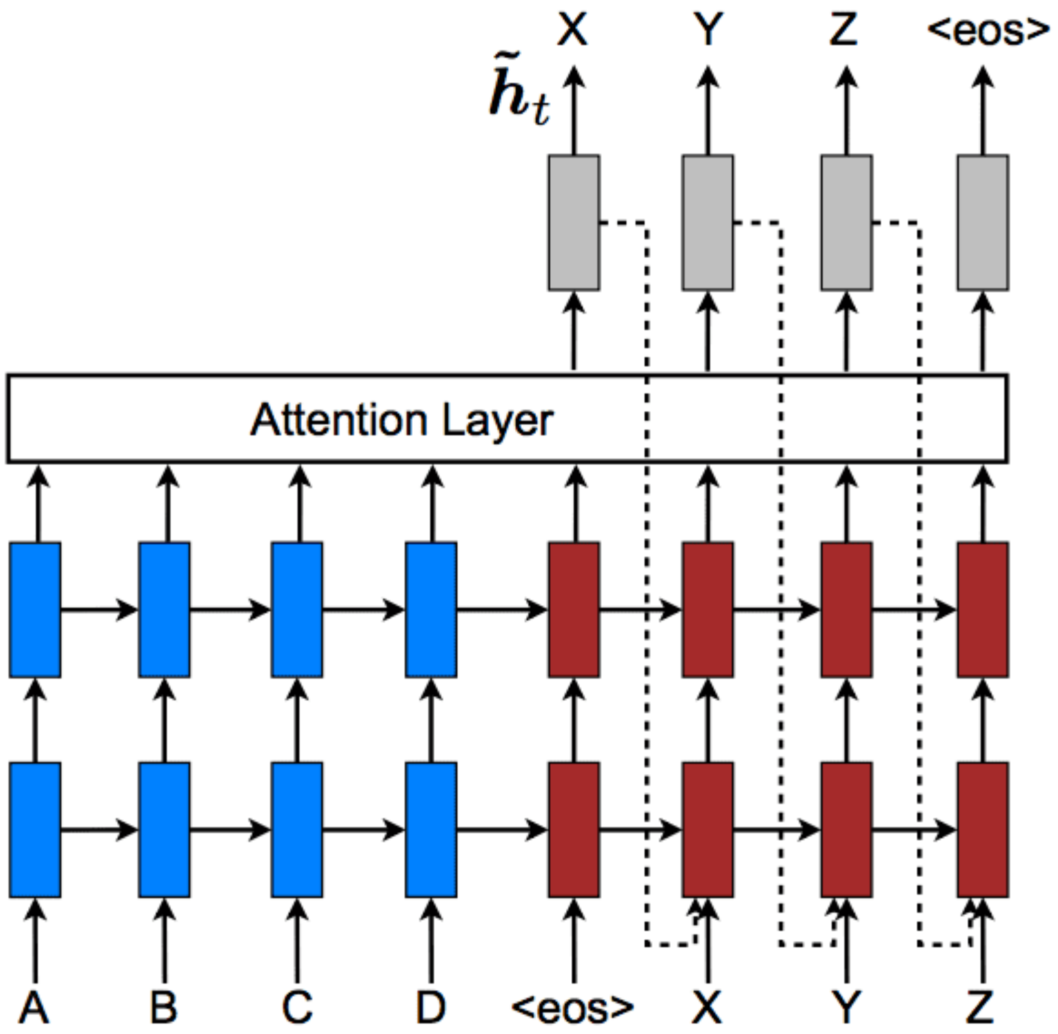
Transformers

Julio Waissman





Antes de los transformers



El artículo que cambió todo

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

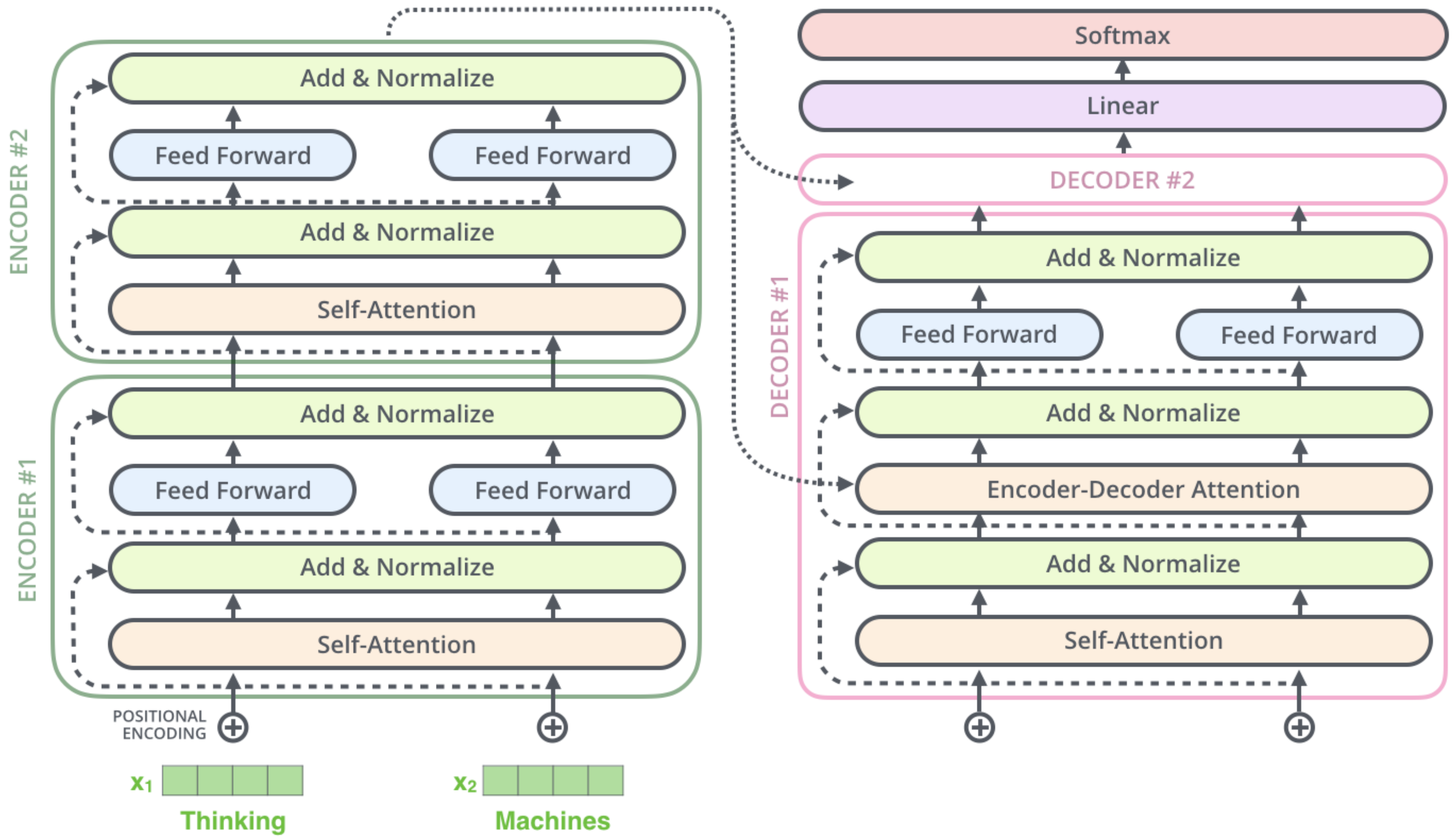
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

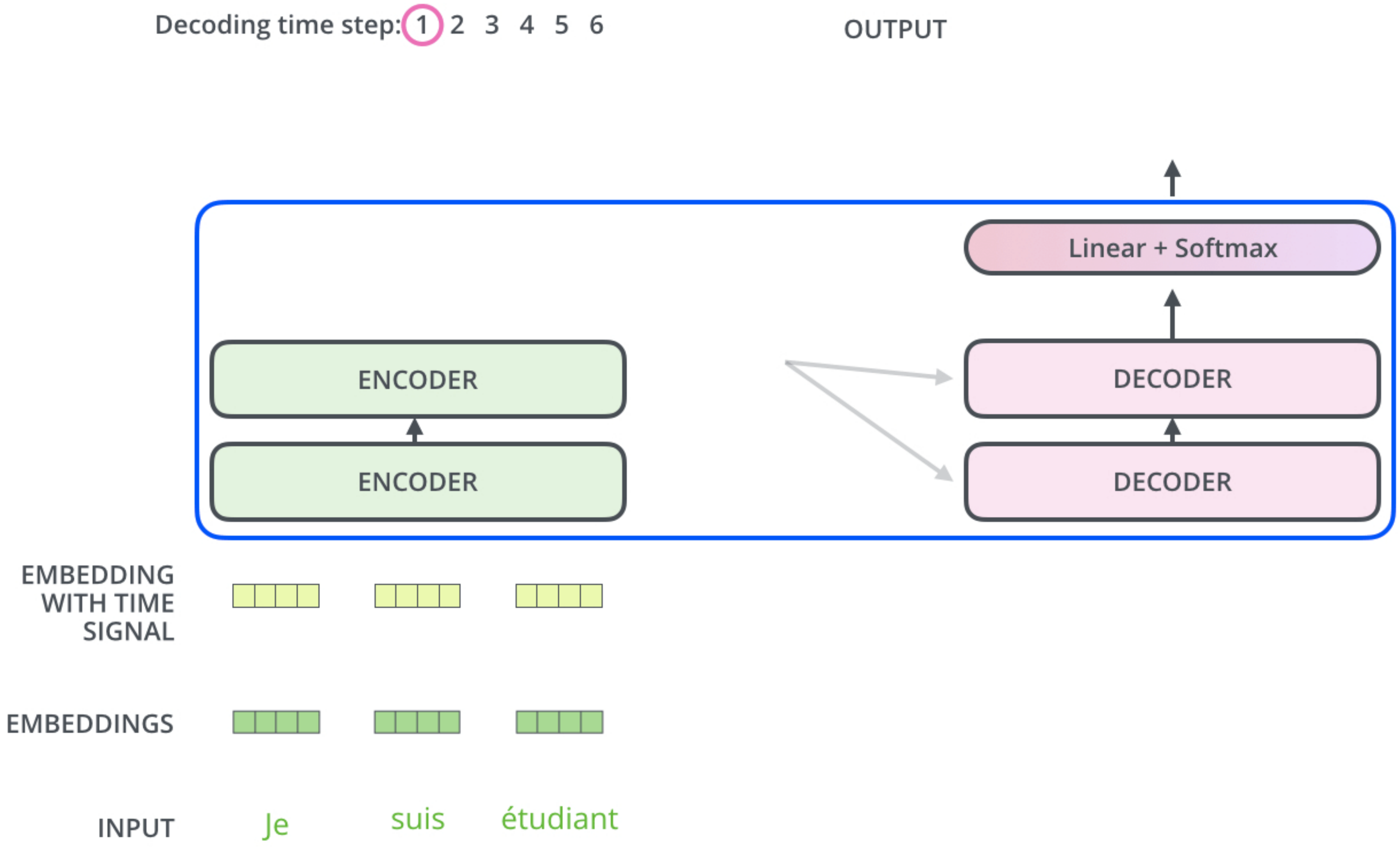
Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Modelo Propuesto (Transformer)

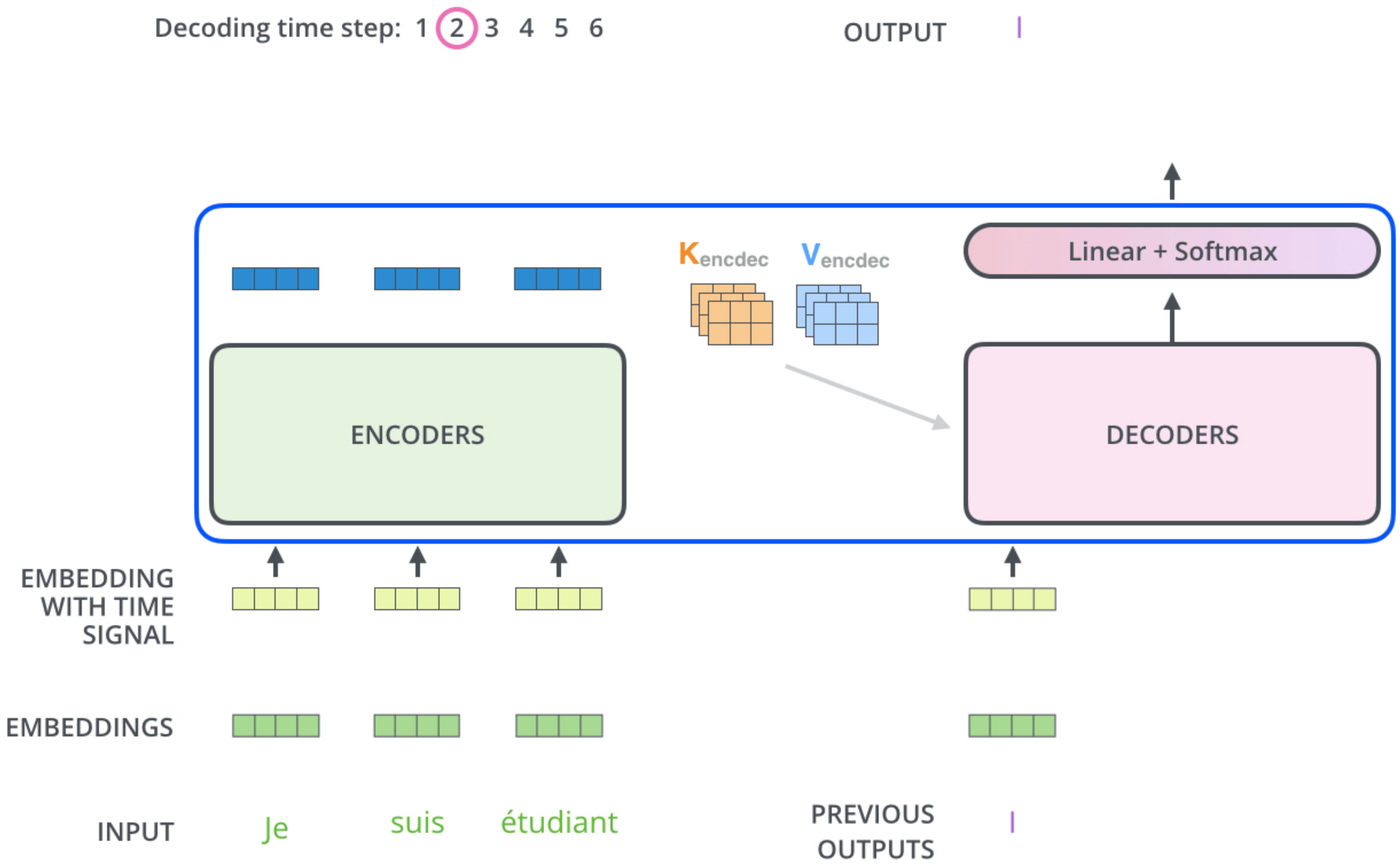
¿Cómo funciona todo?



¿Cómo funciona todo?

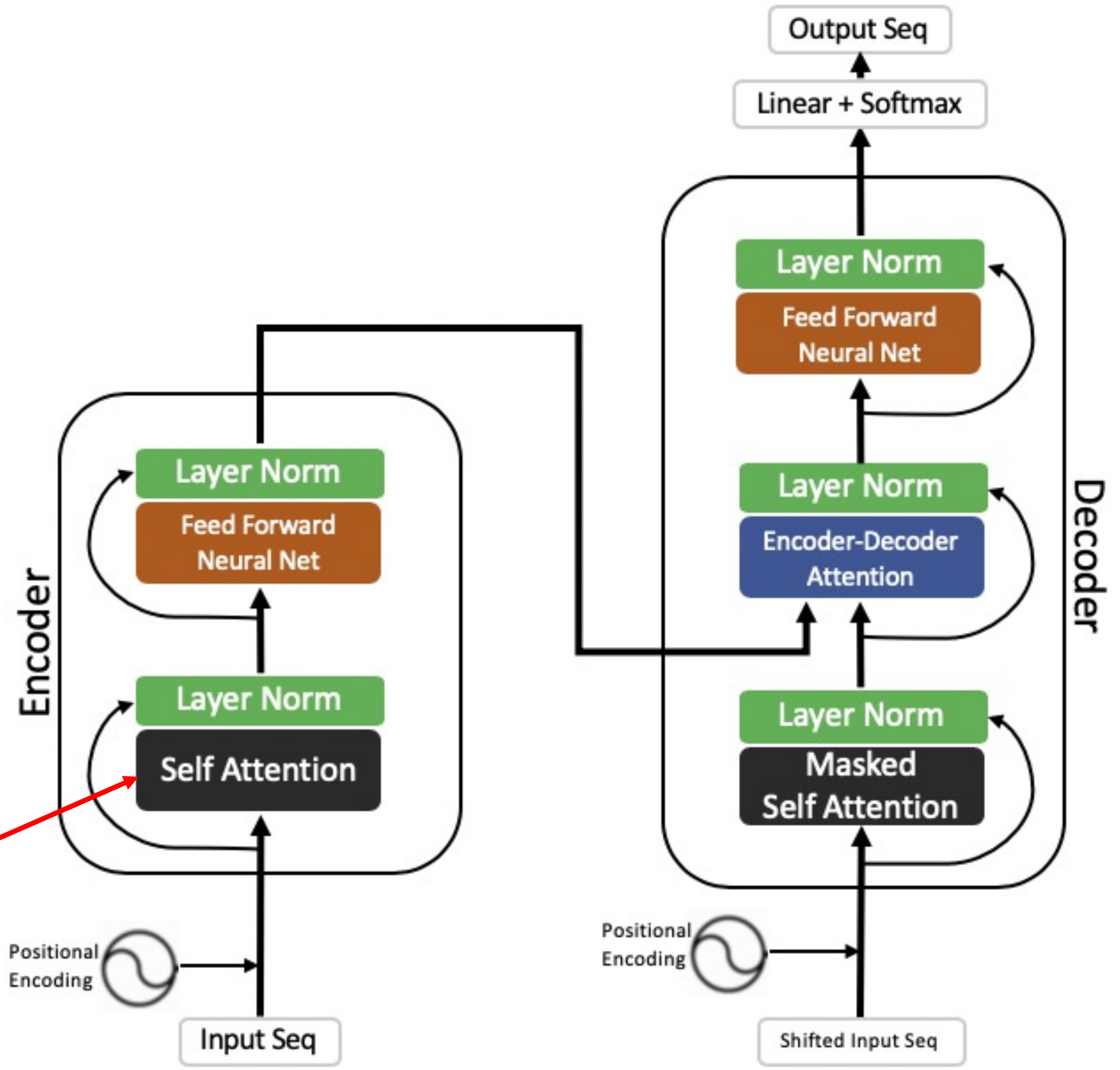


¿Cómo funciona todo?



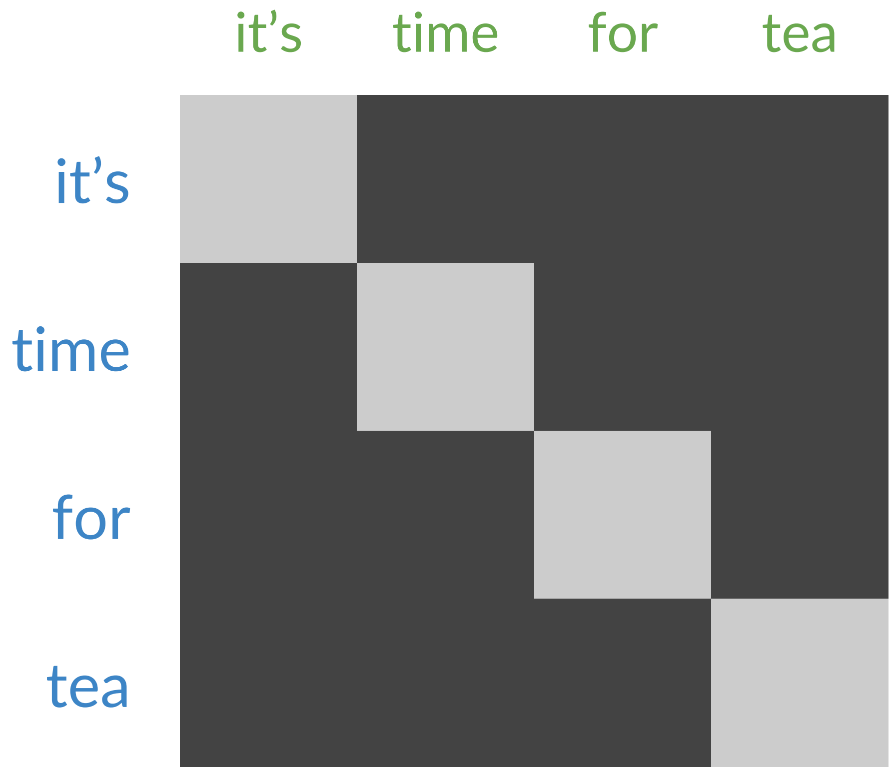
Modelo Propuesto (Transformer)

La unidad básica



Autoatención

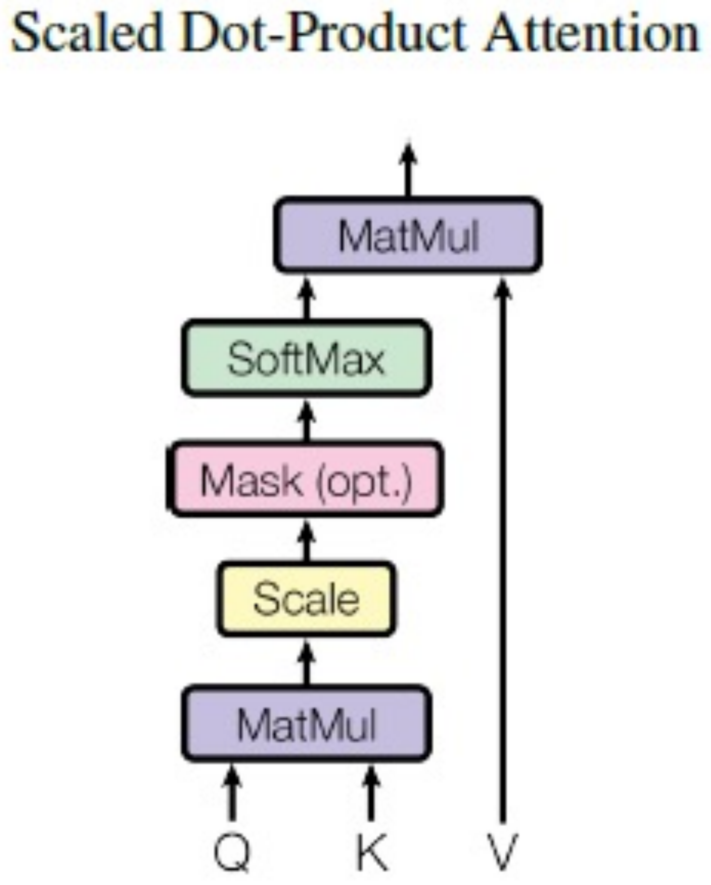
Queries, keys and values come from the **same sentence**

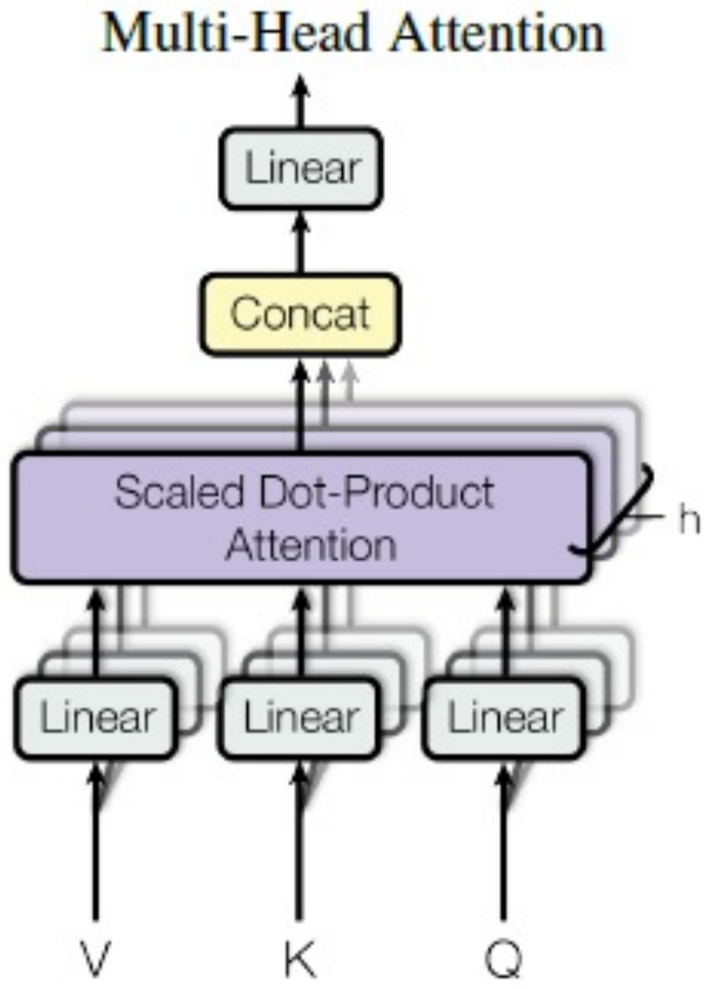


Weight matrix

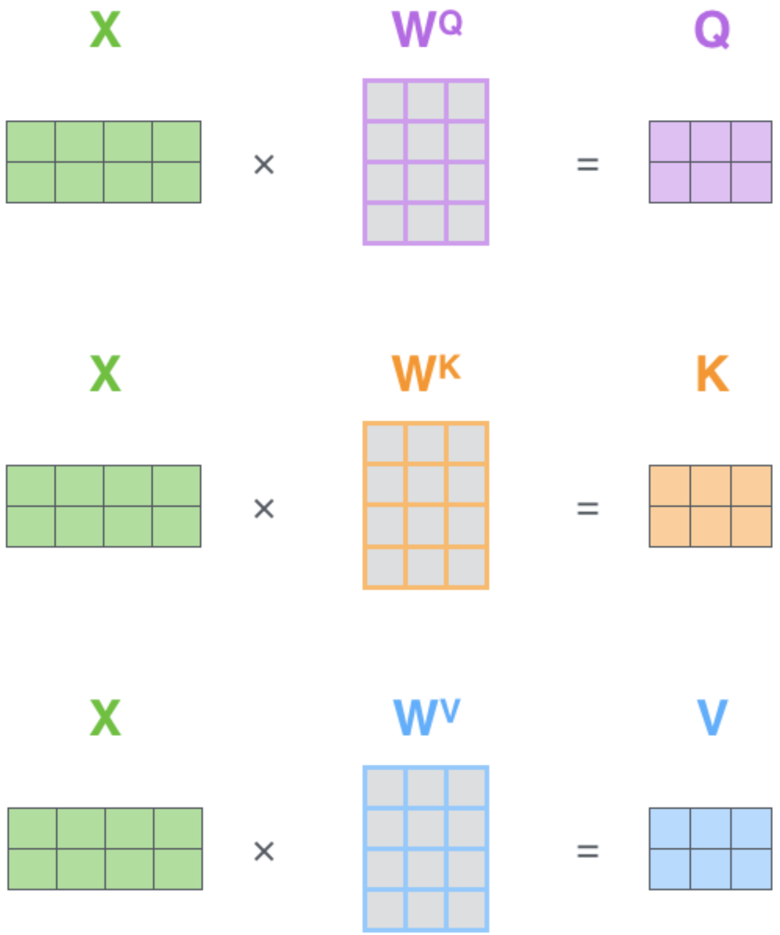
Meaning of each word **within** the sentence

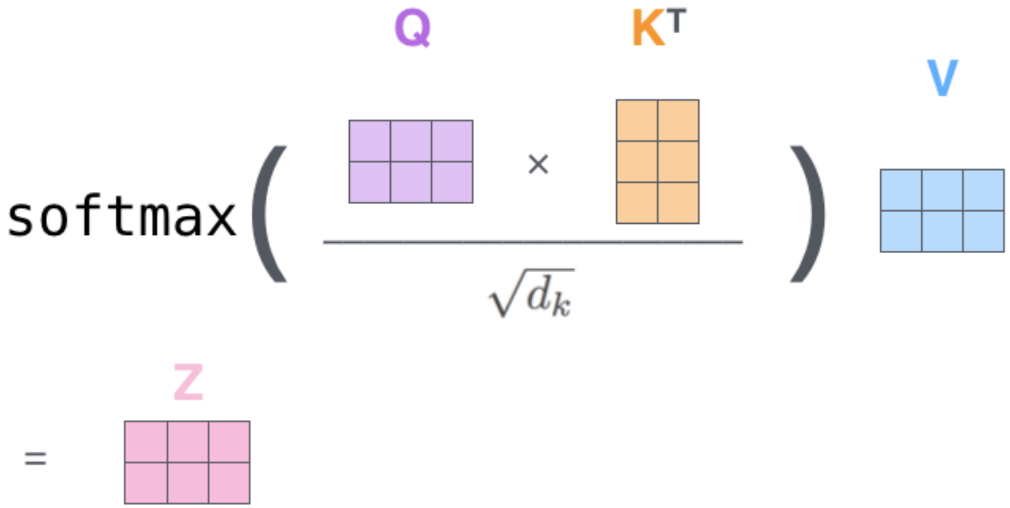
Attention vs MultiHead Attention



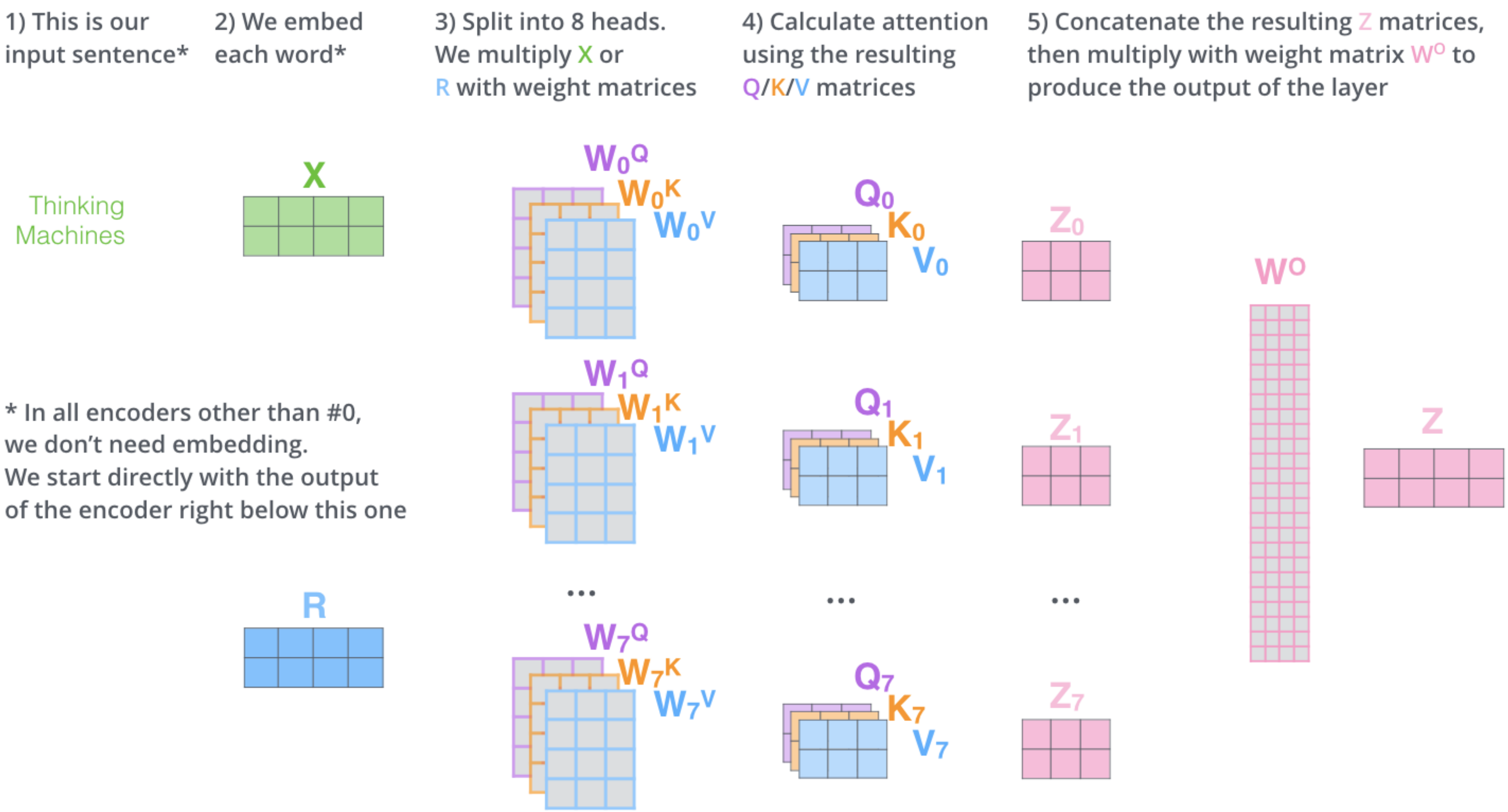


Self Attention



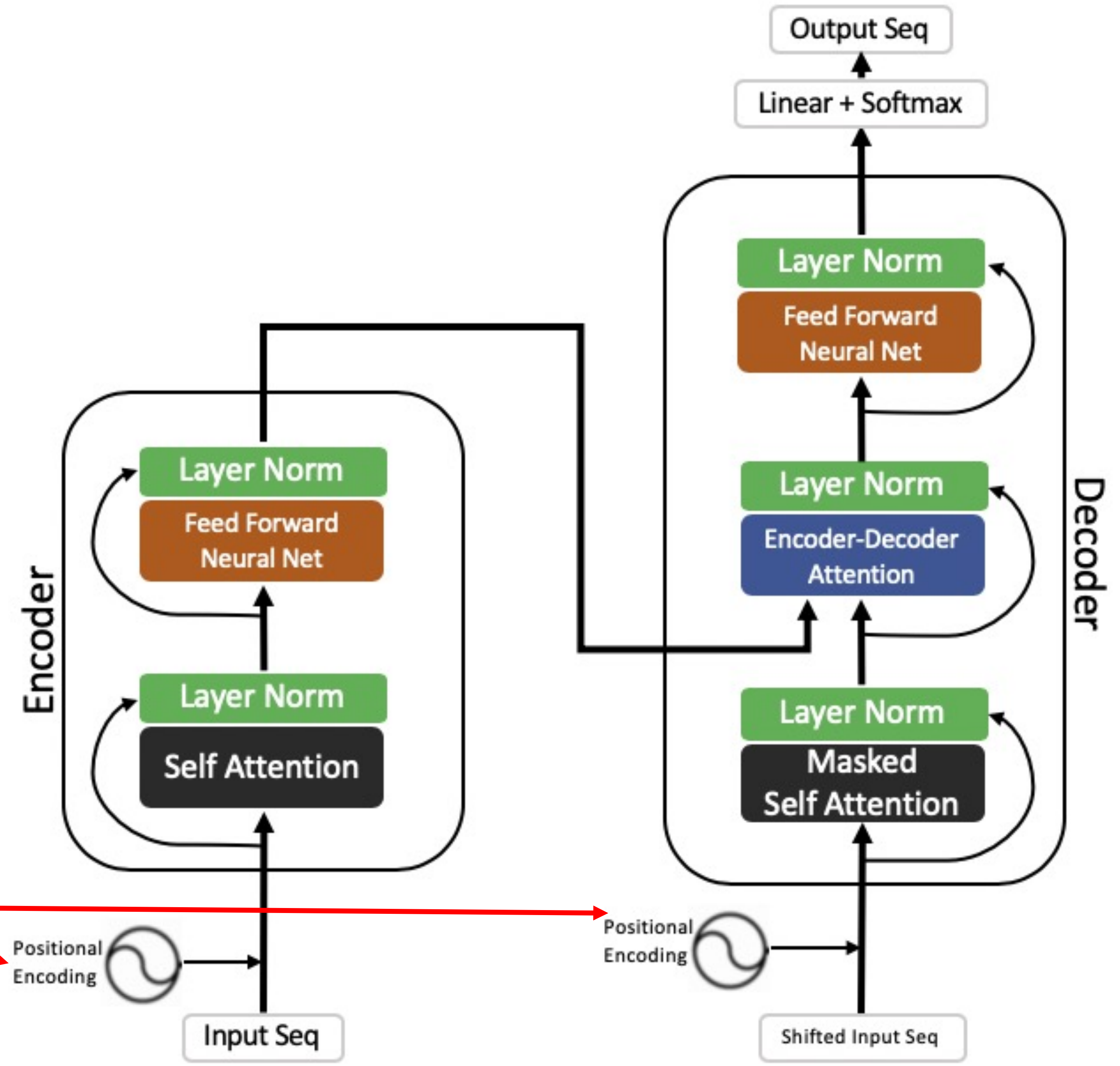


Multi-head Self Attention

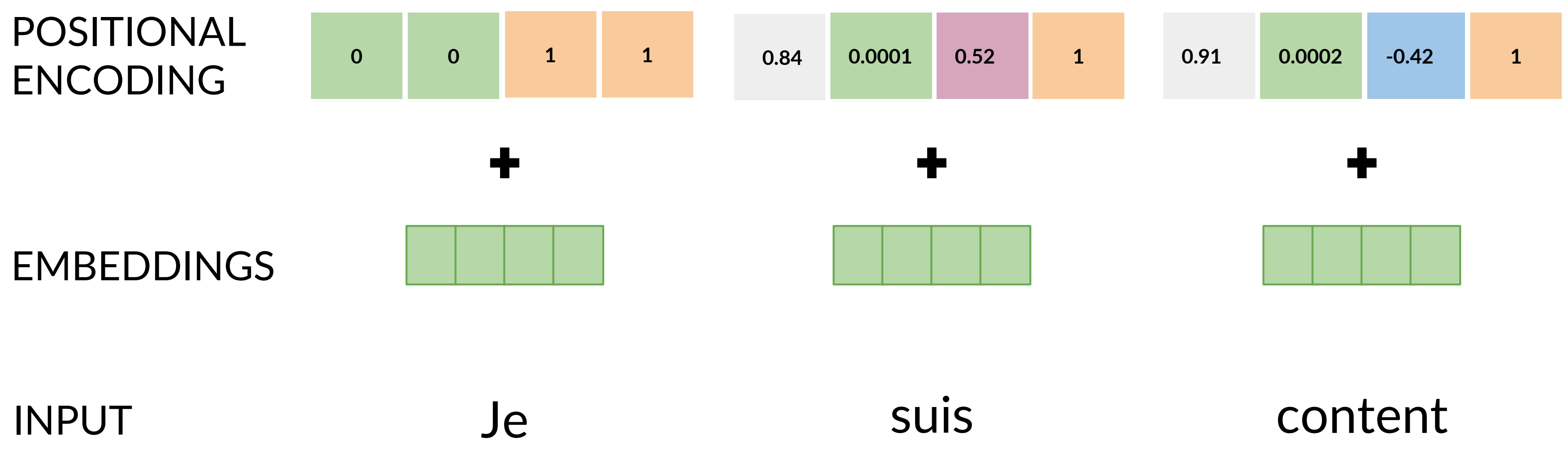


Modelo Propuesto (Transformer)

Mantener el orden de entrada



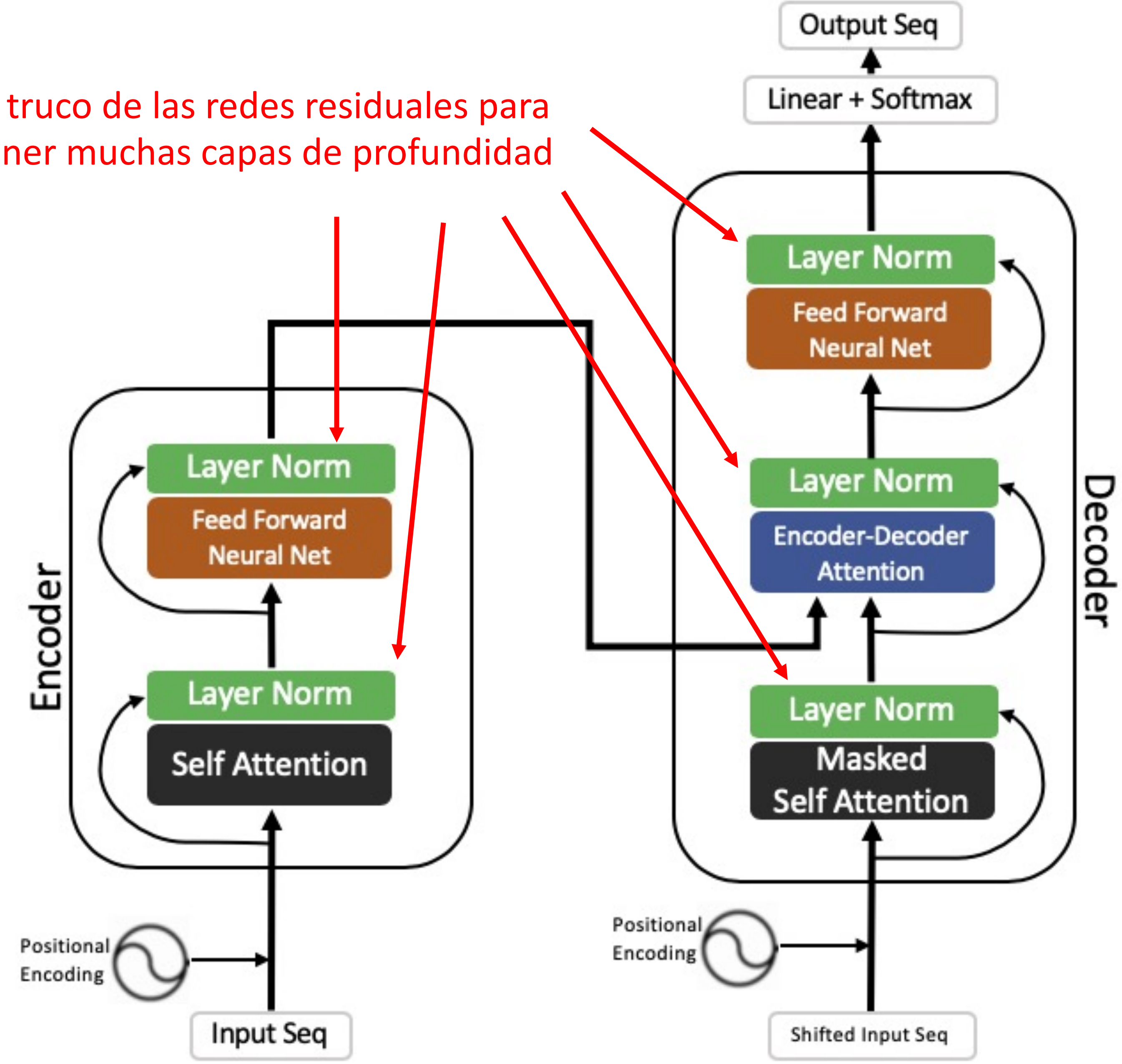
Positional encoding



RNNs vs Transformer: Positional Encoding

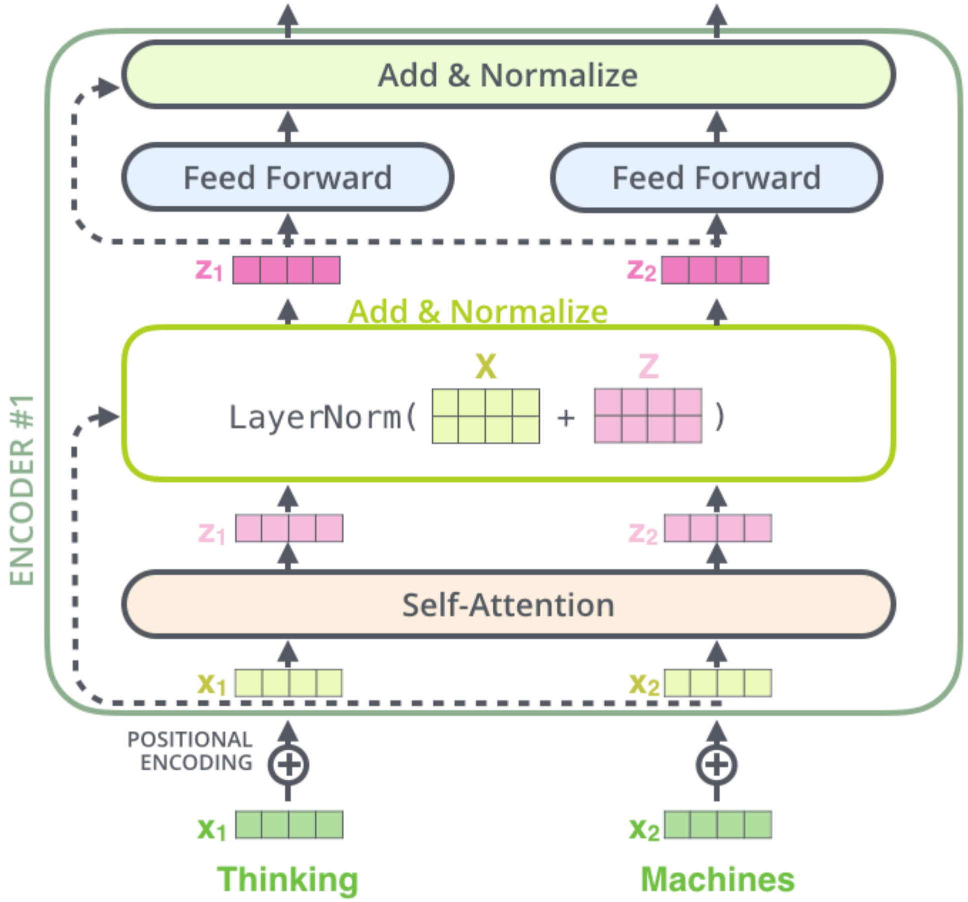
Modelo Propuesto (Transformer)

El truco de las redes residuales para
tener muchas capas de profundidad

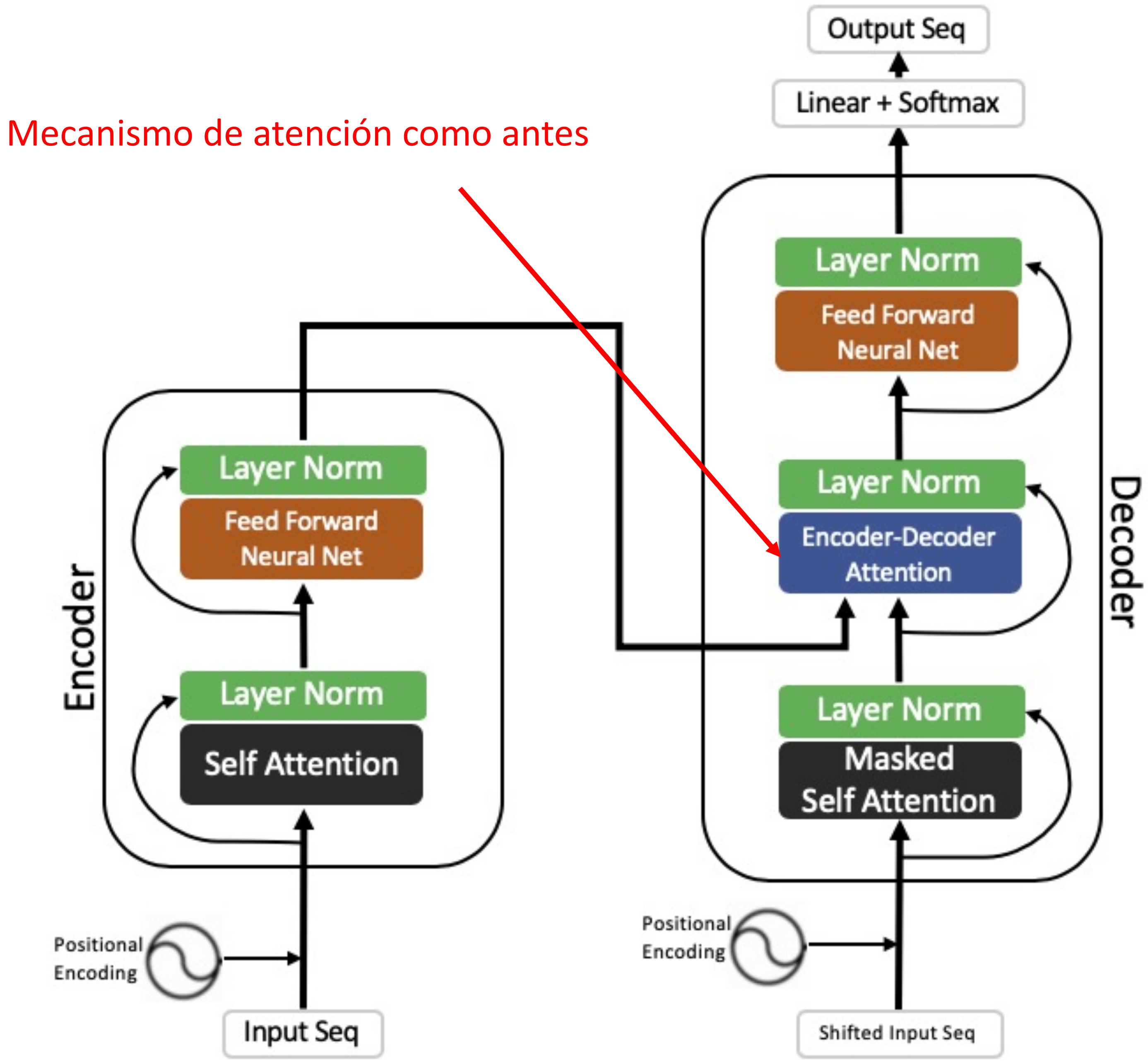


Deep Residual Learning for
Image Recognition

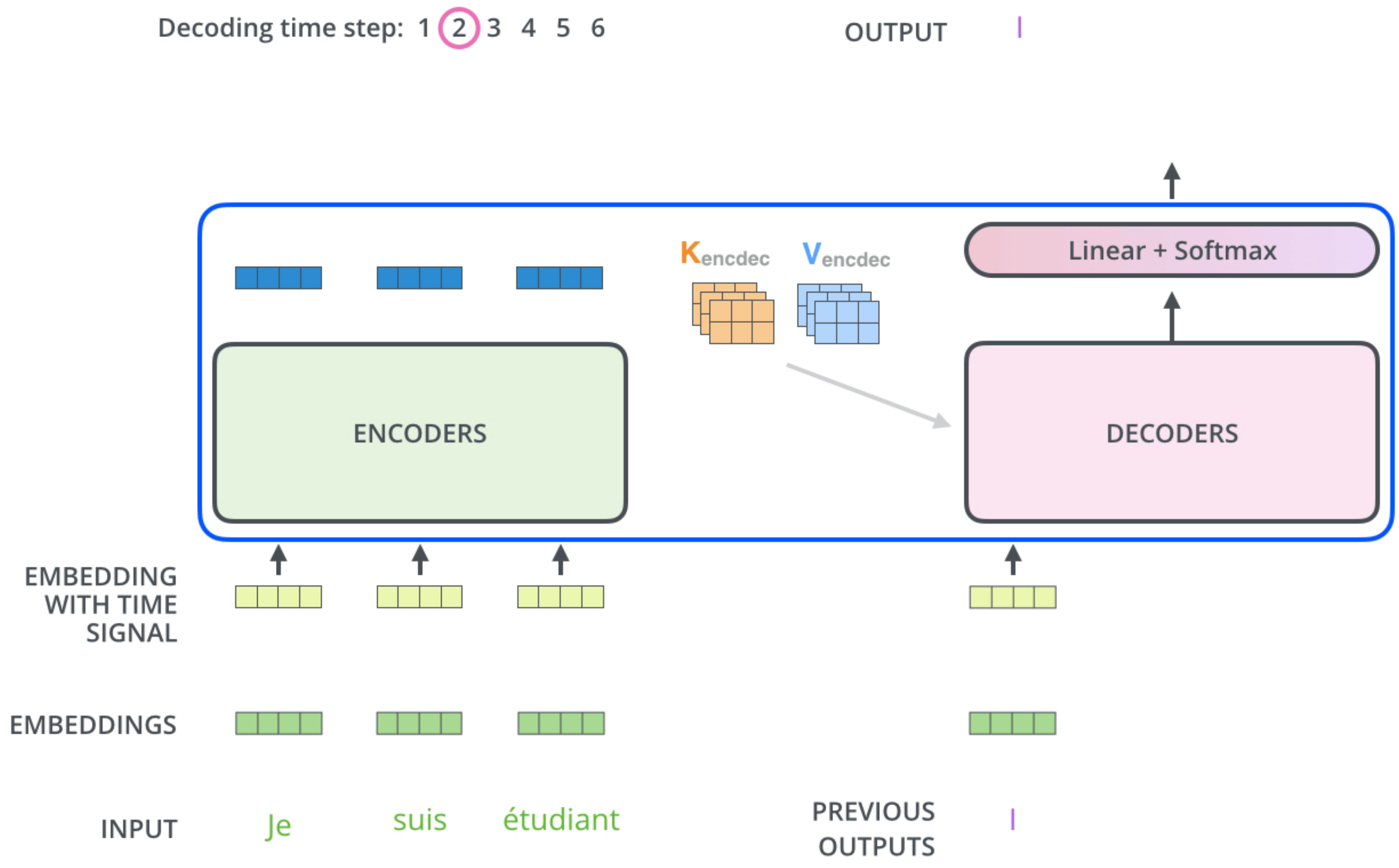
Add & Normalize



Modelo Propuesto (Transformer)

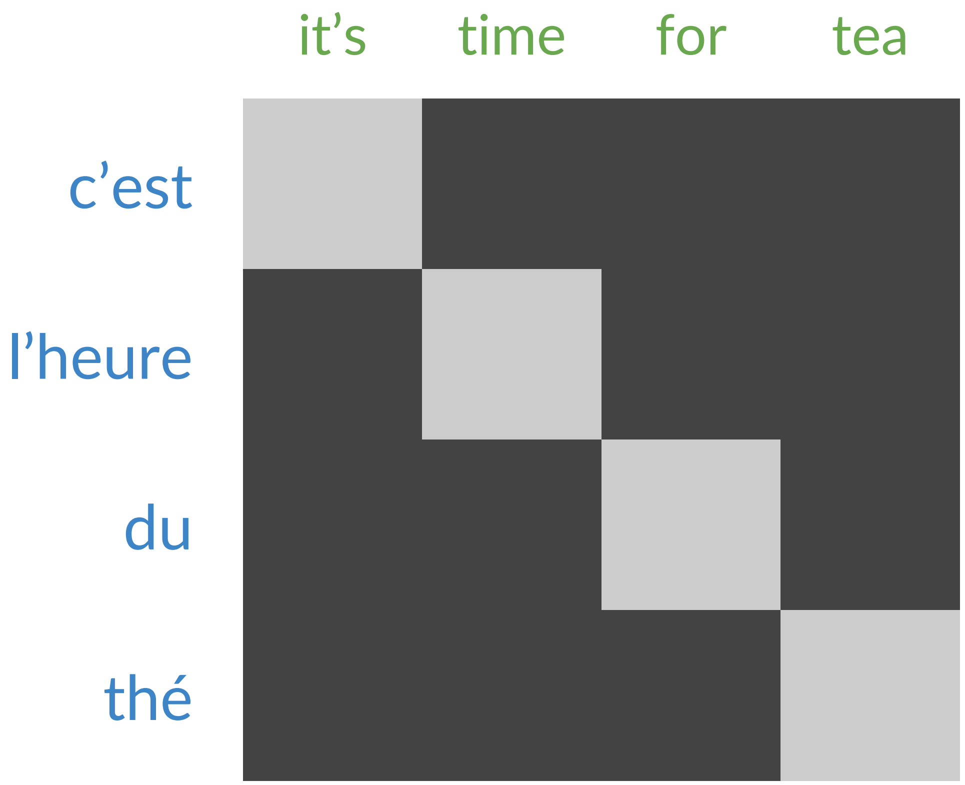


Encode-Decoder Self-Attention



Encoder-Decoder Attention

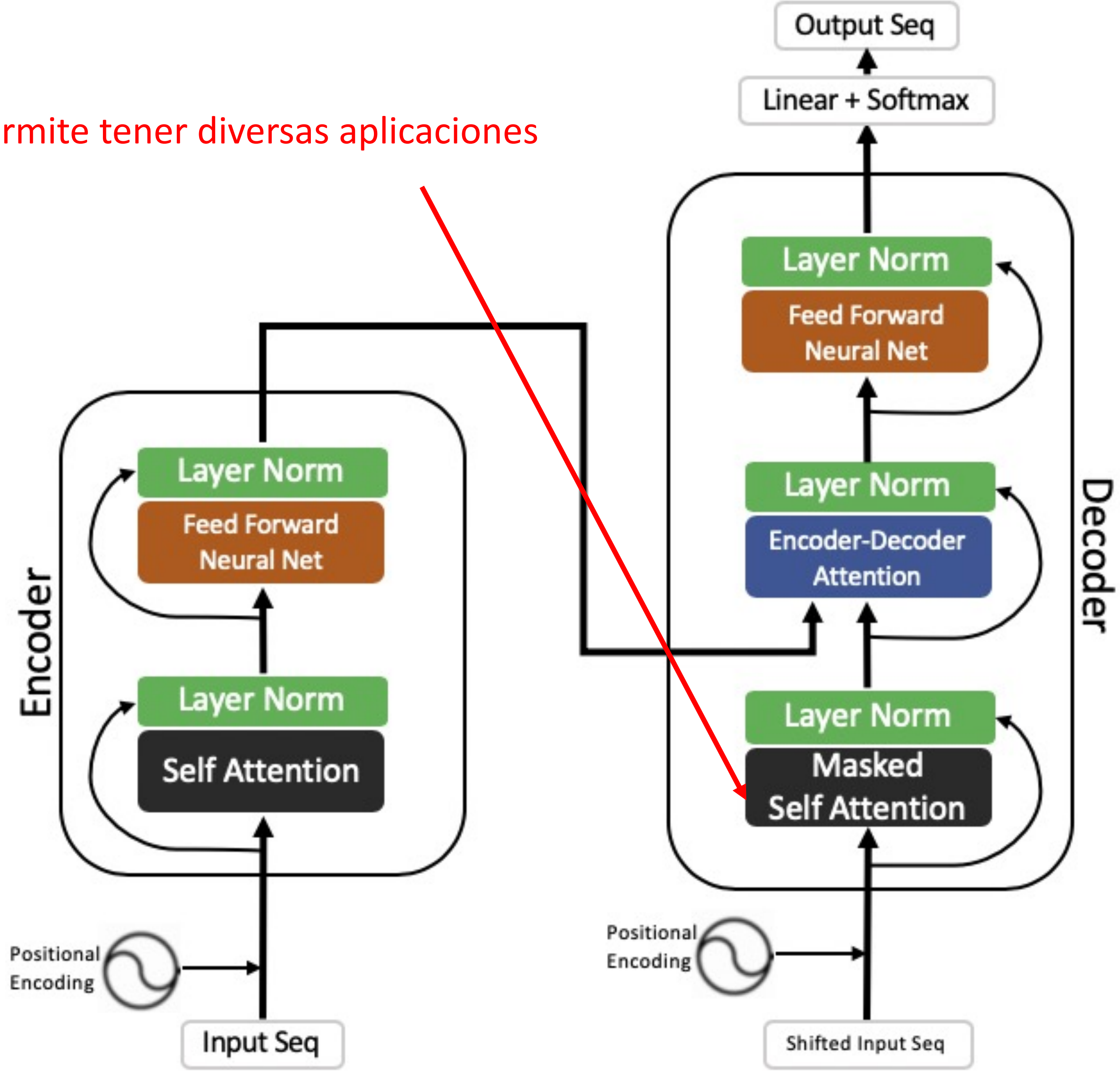
Queries from one sentence, keys and values from another



Weight matrix

Modelo Propuesto (Transformer)

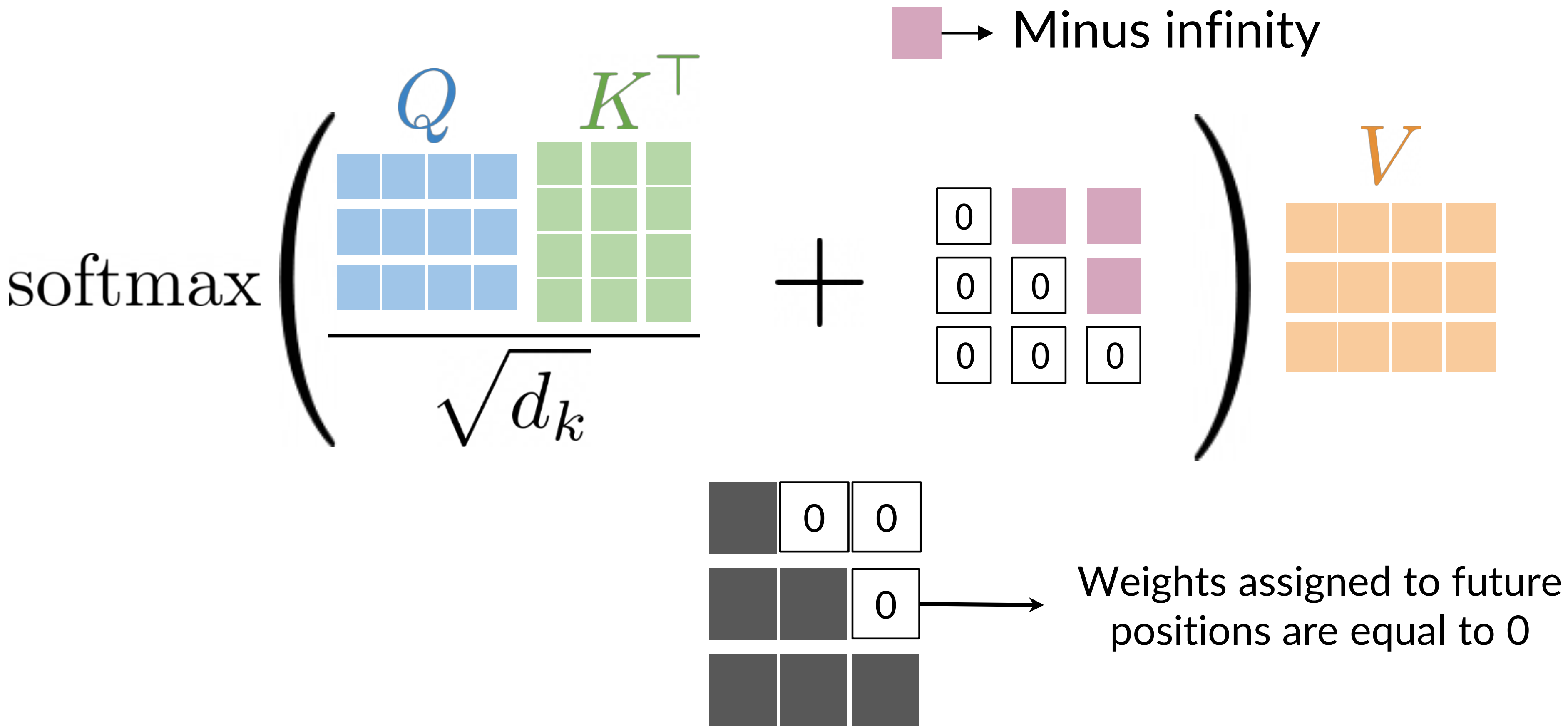
Permite tener diversas aplicaciones



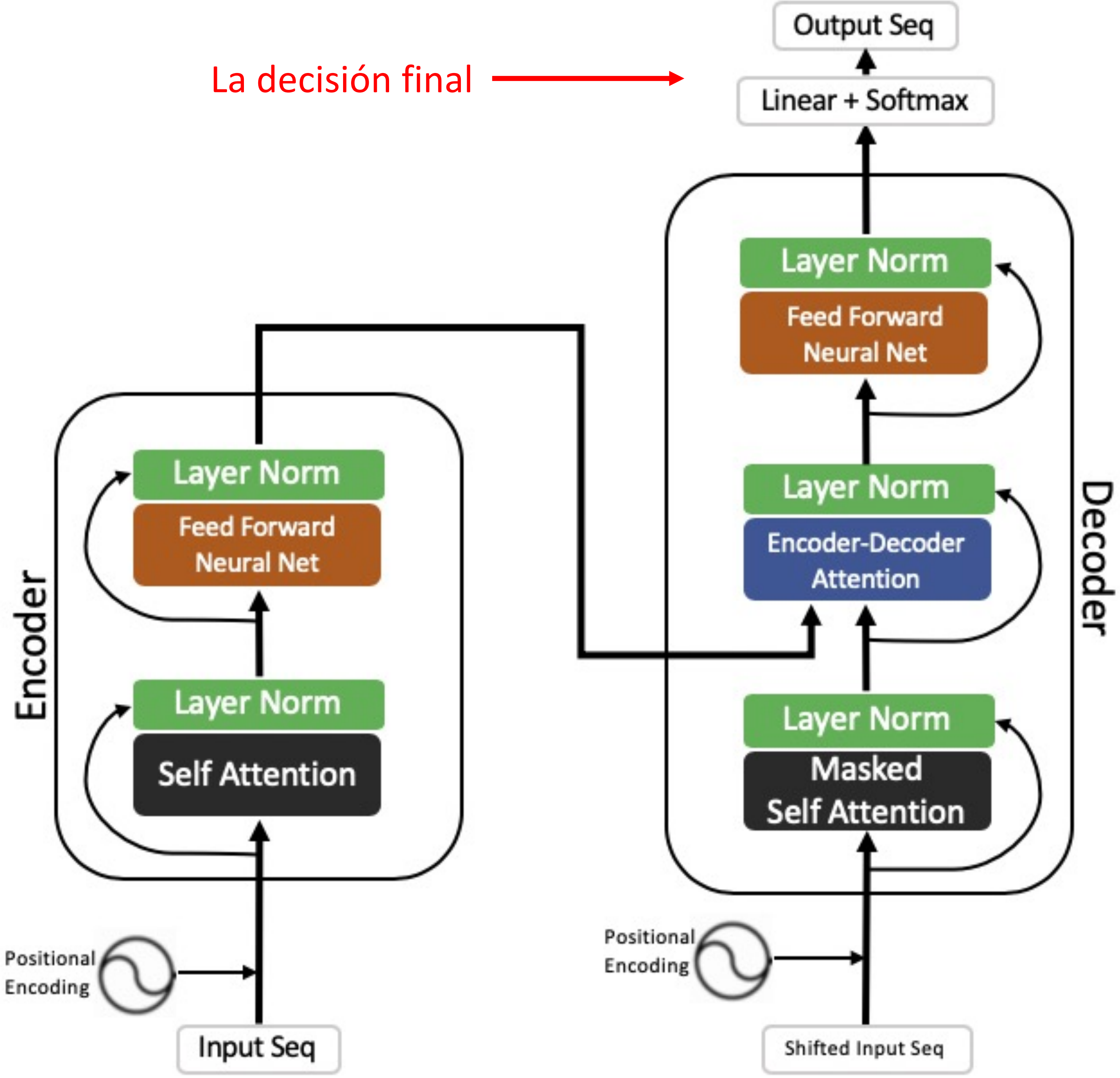
Masked Self-Attention

Queries, keys and values come from the **same sentence**. Queries don't attend to future positions.

¿Y cómo se hace el Masked Self-Attention?



Modelo Propuesto (Transformer)



Capa final de los decoders