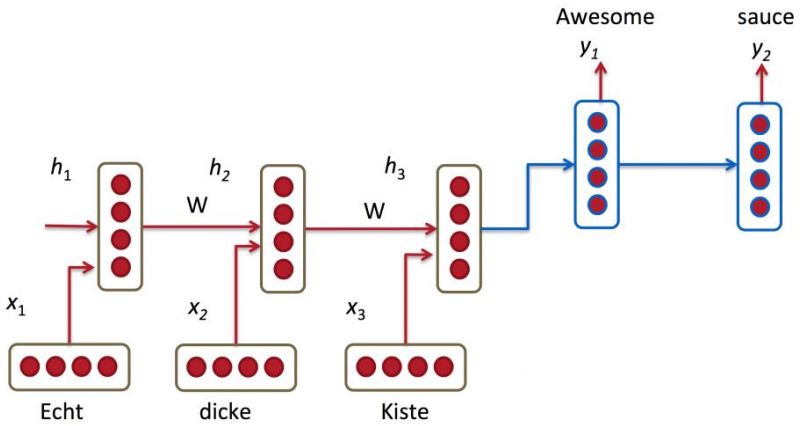
Mecanismos de atención en redes neuronales

Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones

Atención

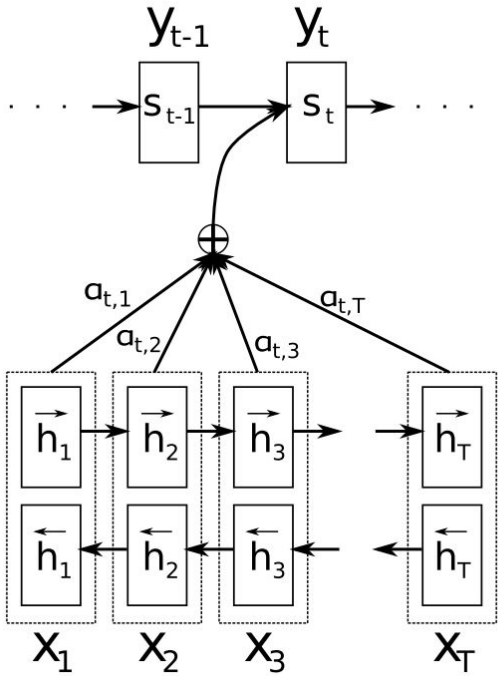
Un mecanismo de atención permite al modelo aprender qué input es más relevante para cada predicción

Atención en AMT



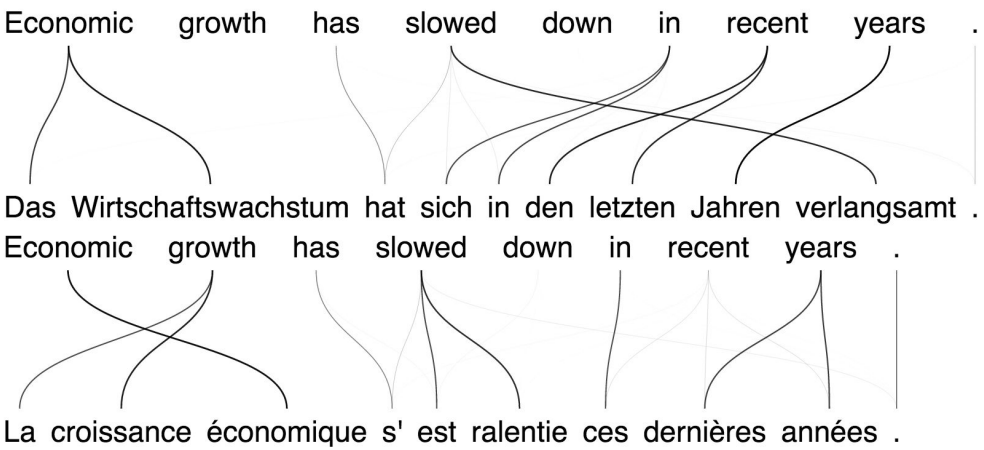
En los problemas de Automatic Machine Translation tenemos que leer y generar una secuencia

Atención en AMT



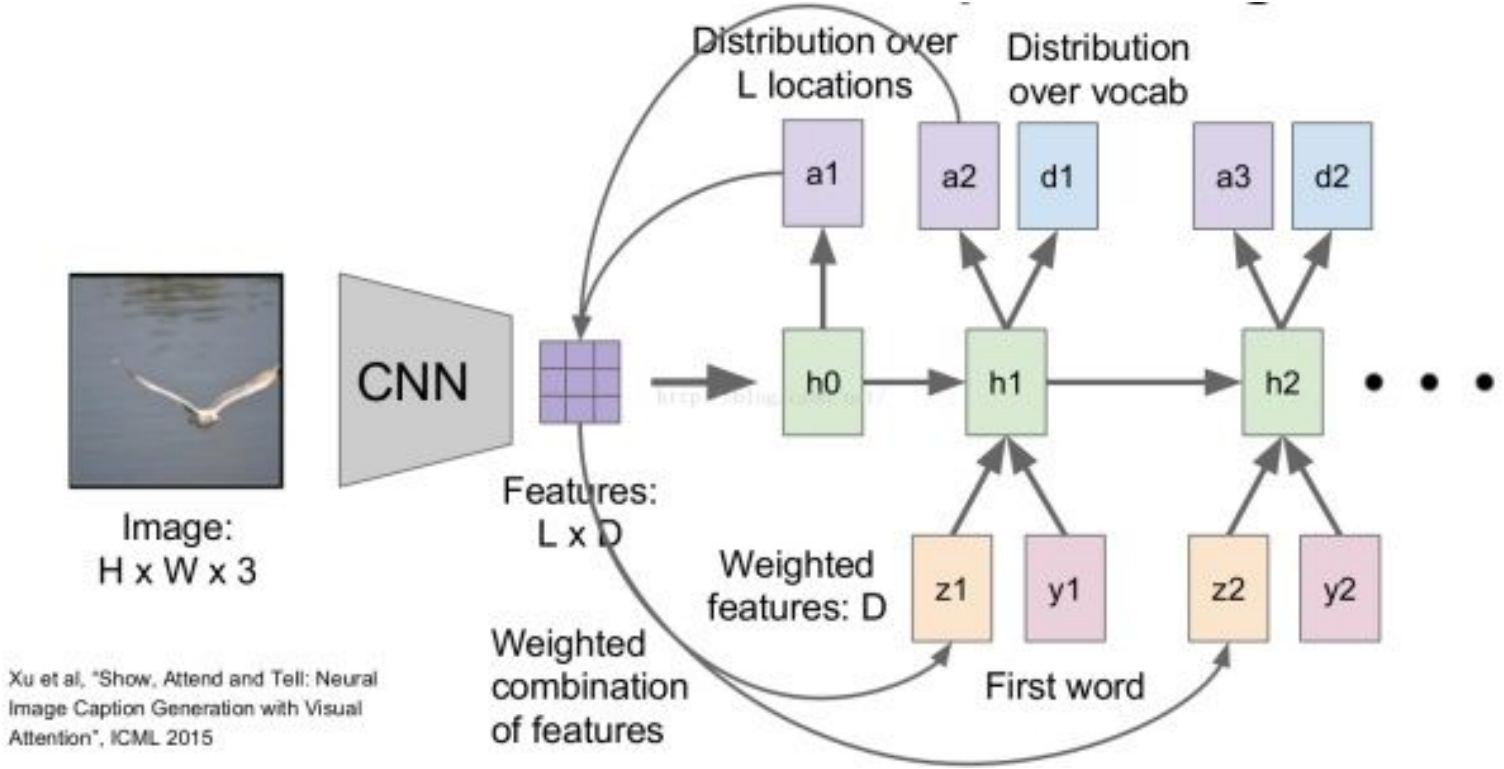
Para generar cada palabra en el idioma target, pesamos todos los output de la capa recurrente de la red.

Atención en AMT



Podemos visualizar qué tan importante era cada palabra en la oración original para generar las palabras de la oración objetivo.

Atención en Image captioning [2]



Xu et al, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

a

giraffe

and

two

zebras

standing

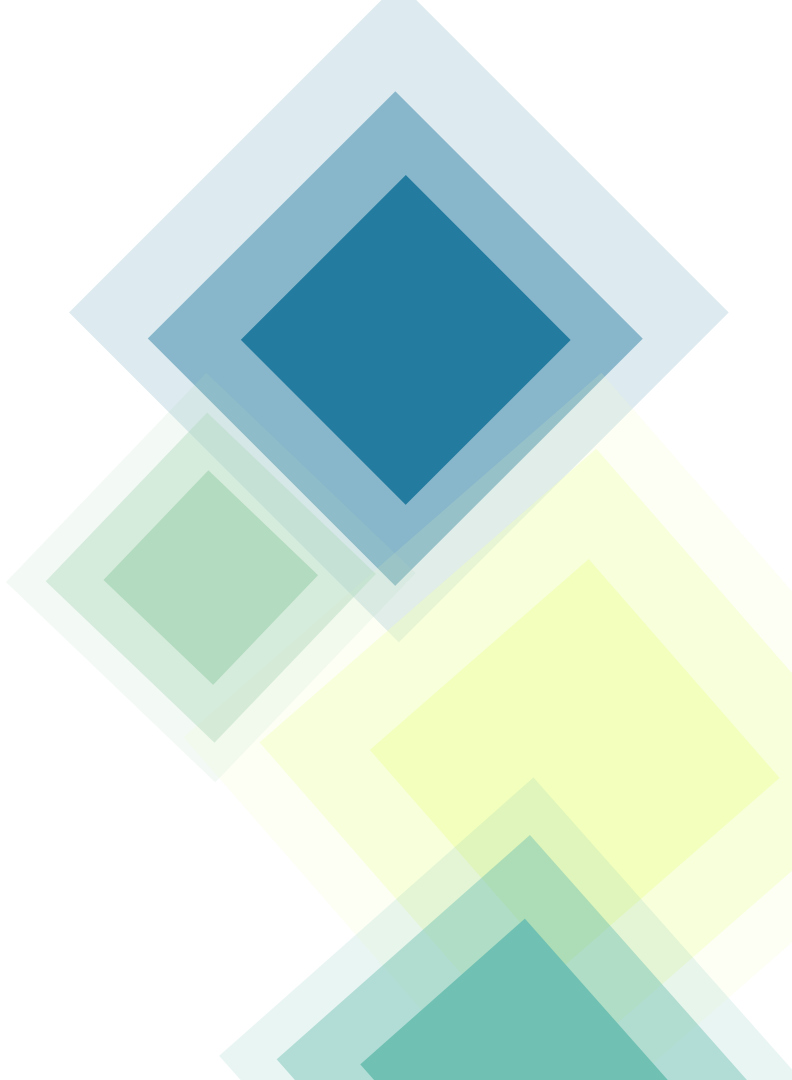
in

a

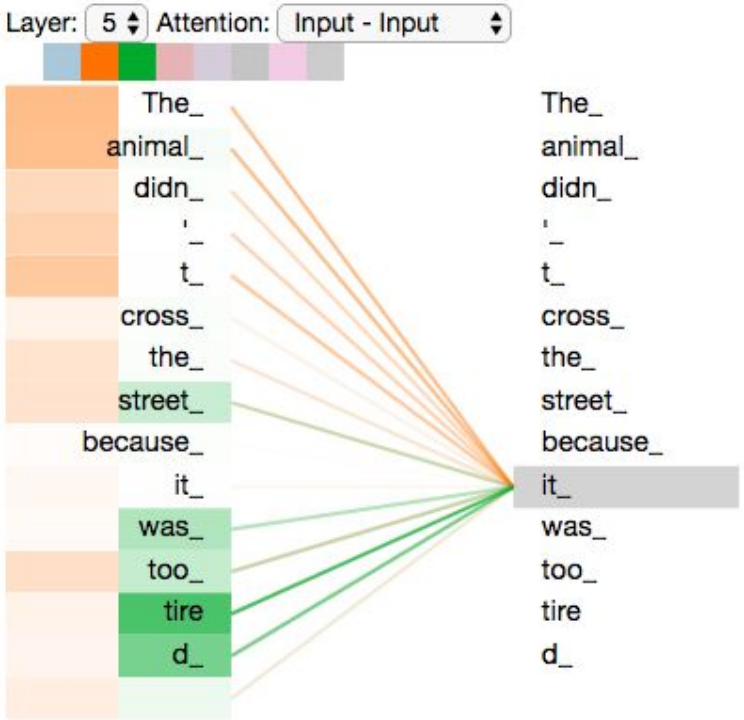
field

.

Self-attention

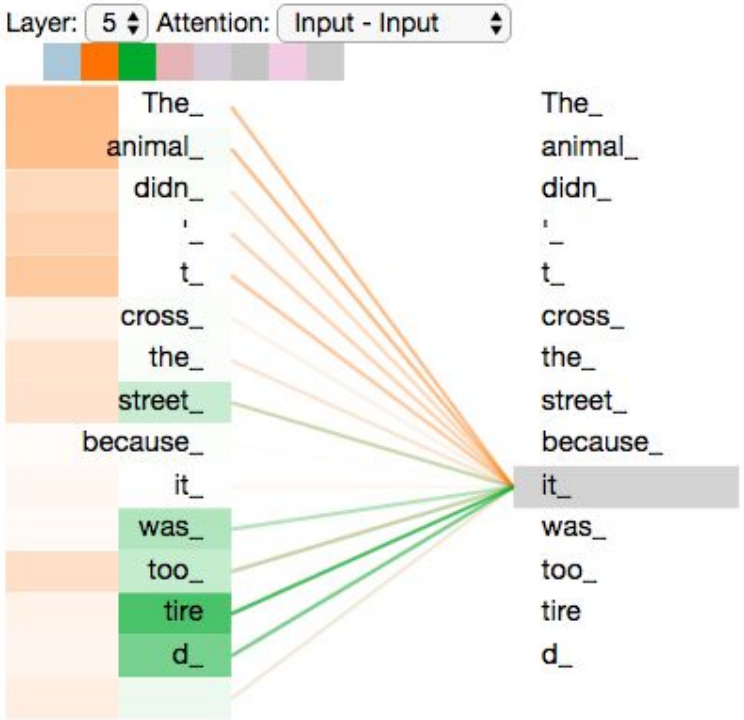


Mecanismo de Self-attention [3]



El self-attention calcula las relaciones entre todas las palabras de dos oraciones.

Mecanismo de Self-attention [3]



La atención se calcula de una oración a sí misma!

Mecanismo de Self-attention [3]

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Mecanismo de Self-attention [3]

Representación de cada una de las palabras de la oración

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Mecanismo de Self-attention [3]

Representación de cada una de las palabras de la oración

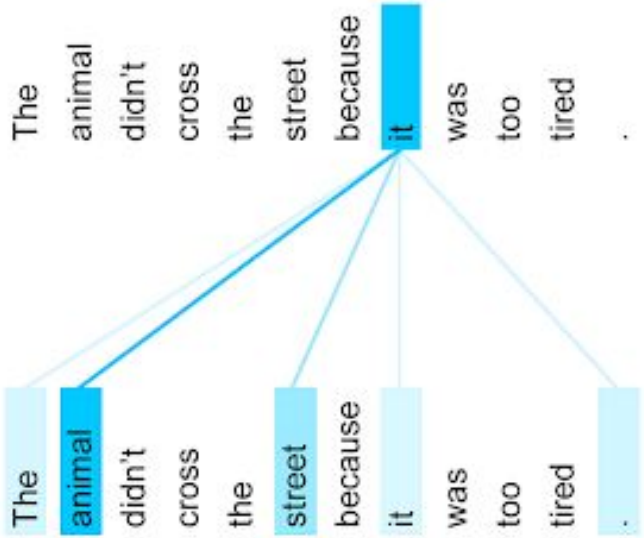
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

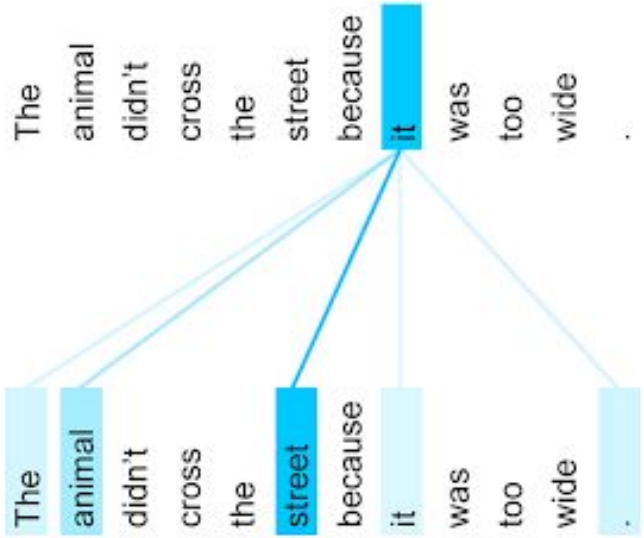
Mecanismo de Self-attention [3]

Representación de cada una de las palabras de la oración

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

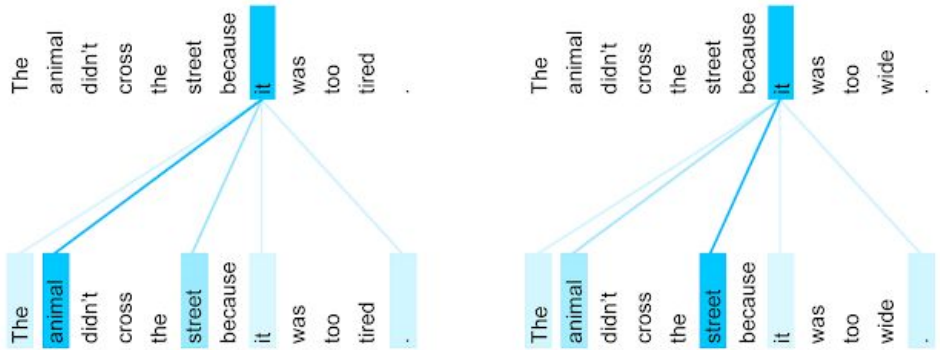
Multiples heads [3]



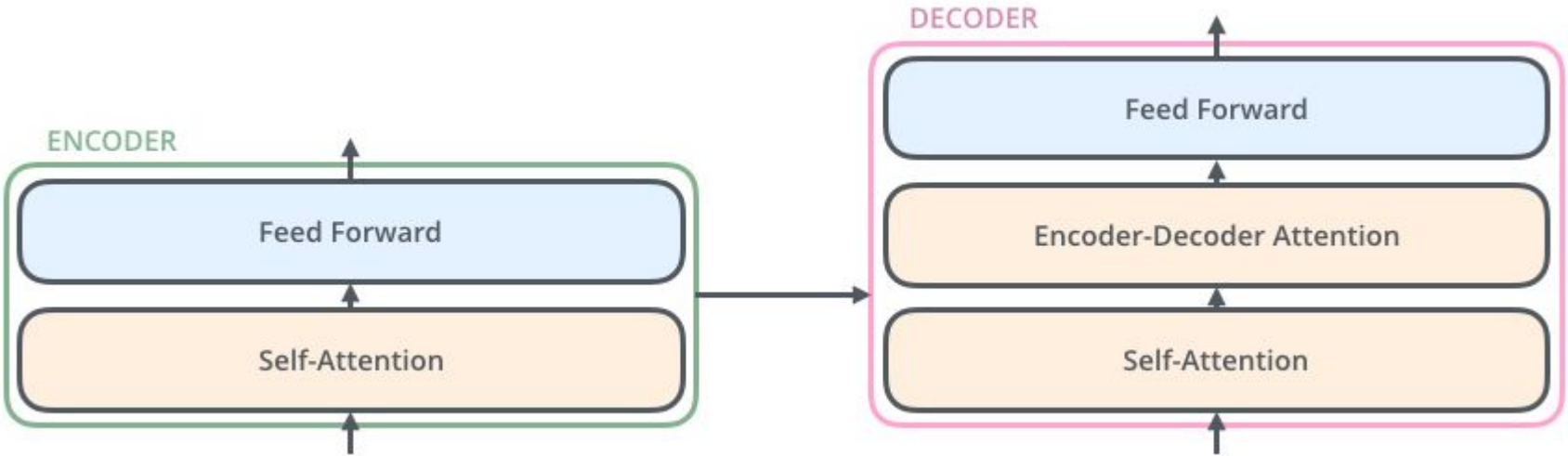


Multiples heads [3]

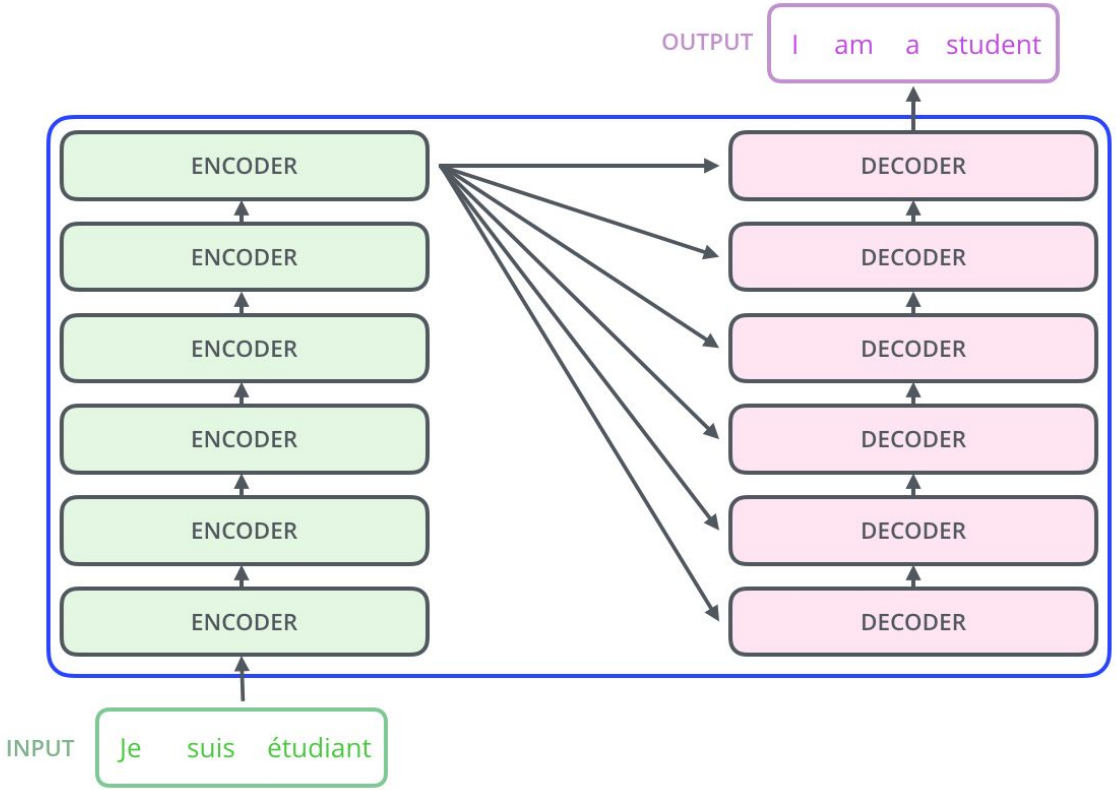
Cada head de atención aprende (potencialmente) un tipo de relación distinta entre las palabras.



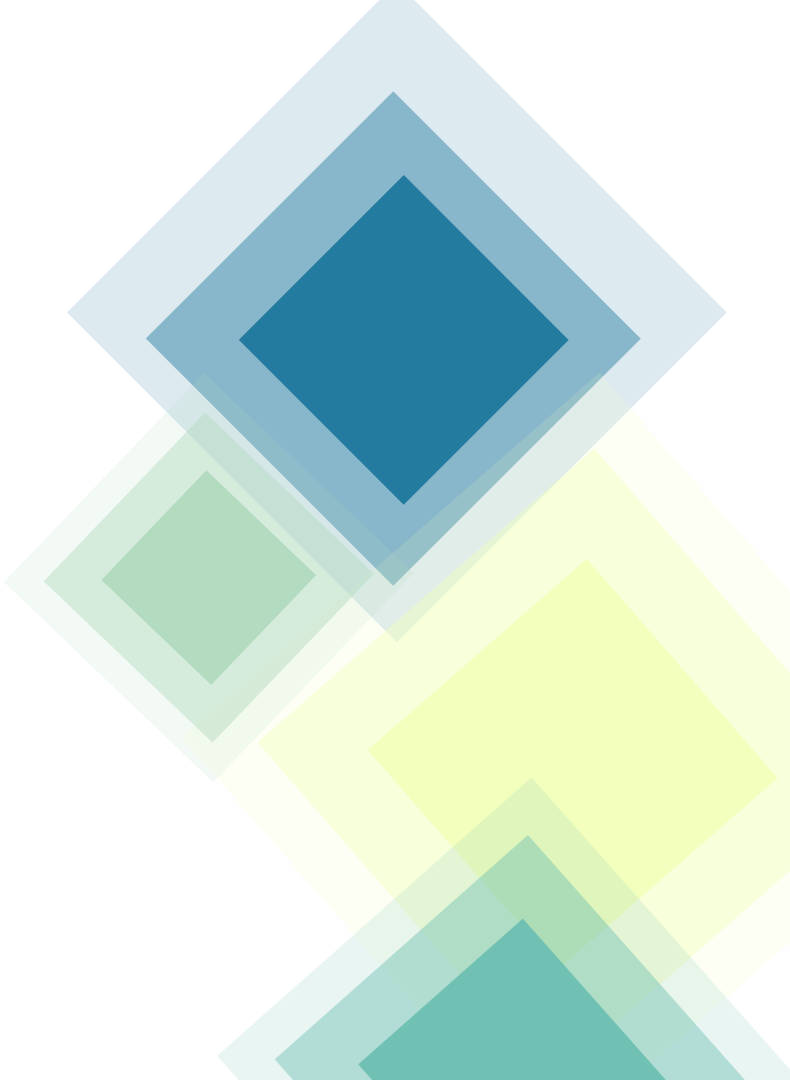
Transformer [3]



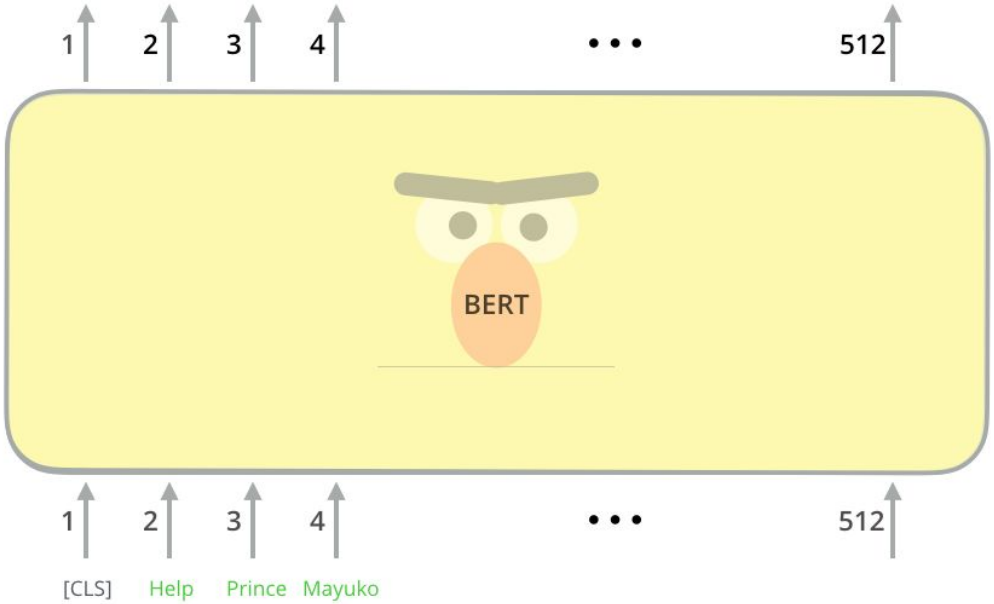
Transformer [3]



Bert

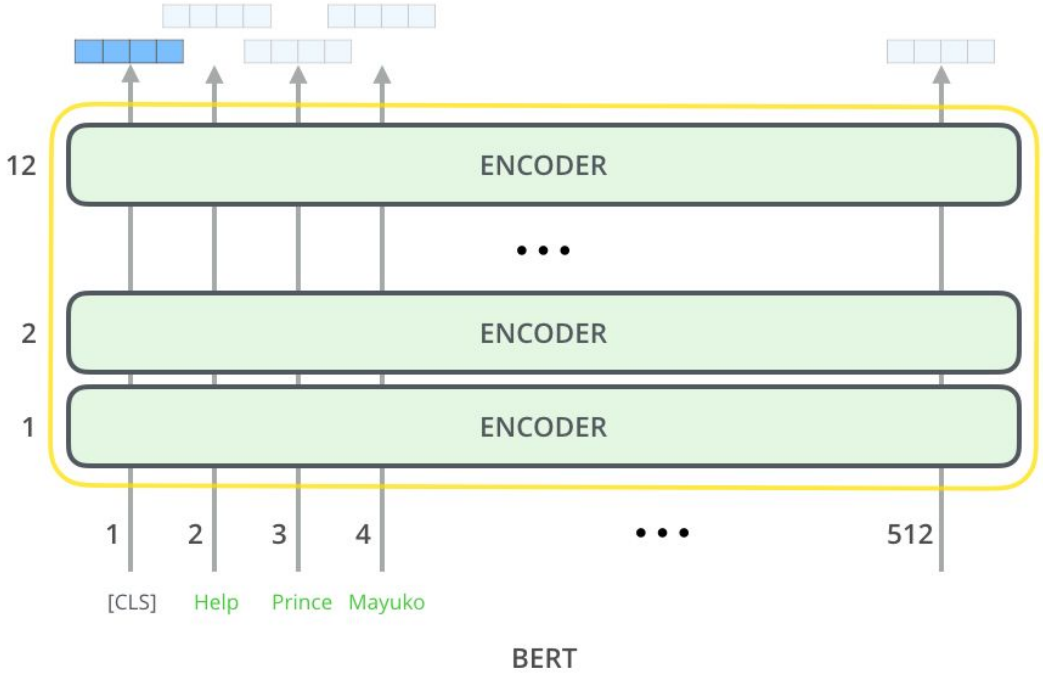


Bert [4]



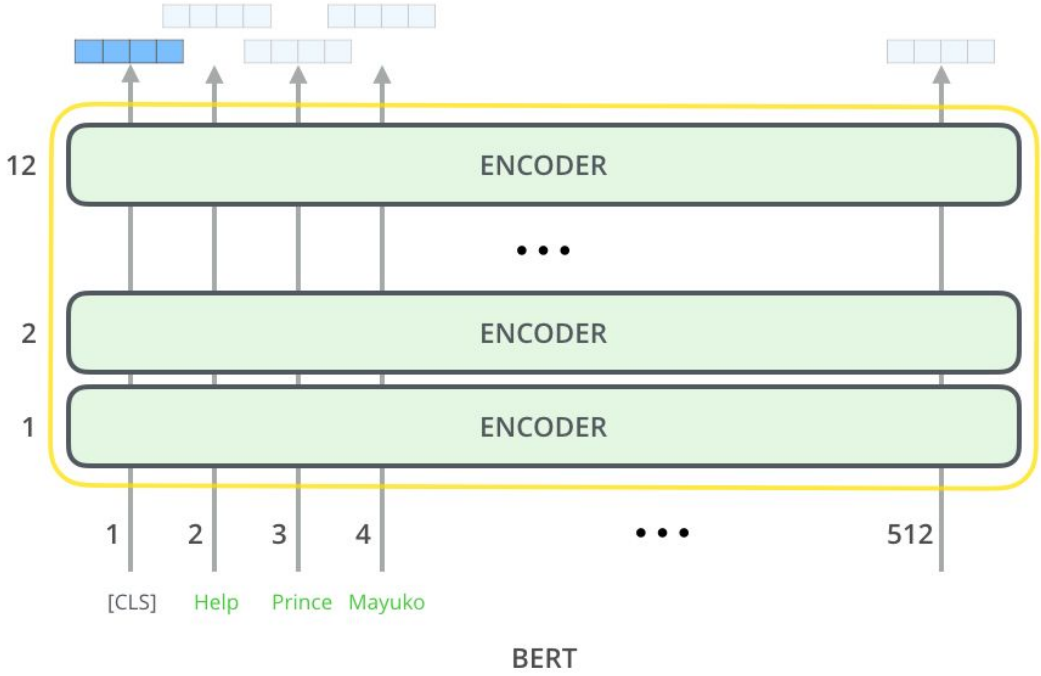
Bert es un modelo para procesado de secuencias que puede utilizarse para múltiples tareas

Bert [4]



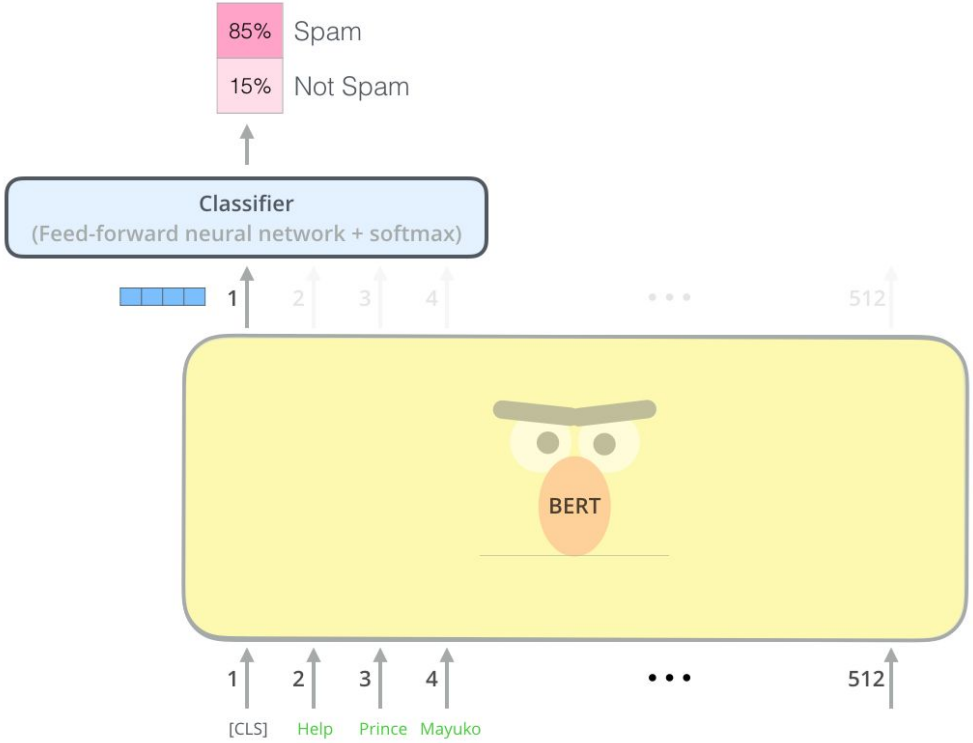
Por dentro, es un apilado de encoders que toman la concatenación de los embeddings de cada palabra.

Bert [4]



En este caso, se puede predecir un vector por cada entrada.

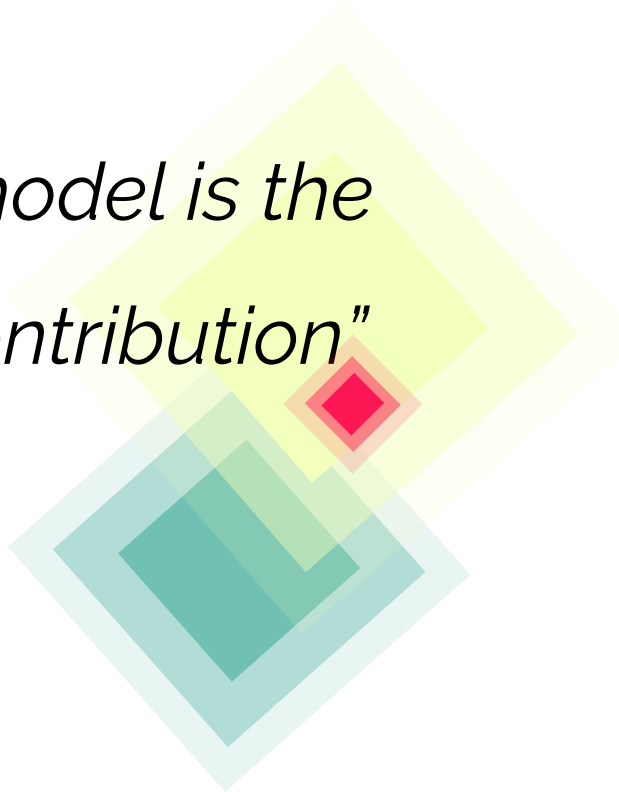
Bert [4]



O se puede utilizar sólo la salida del primer momento de tiempo para resolver un problema de sequence classification

Bert y otros modelos más modernos **no utilizan** recurrencias ni convoluciones para procesado de secuencias

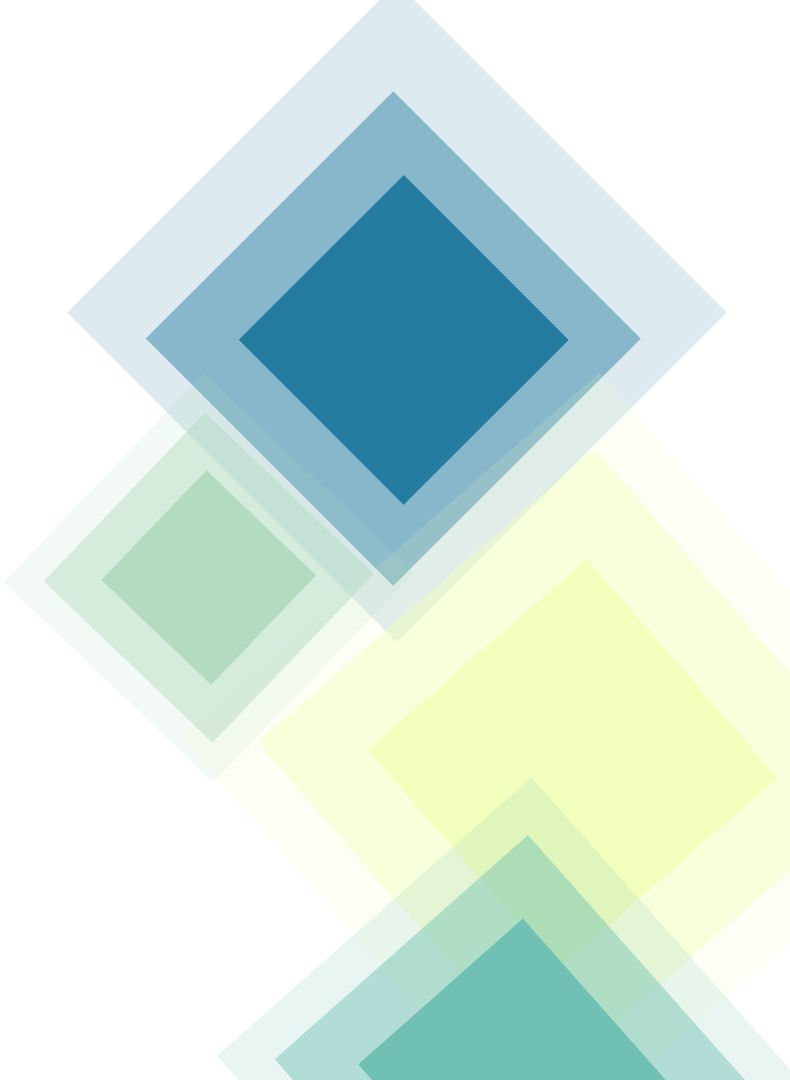
*“**bidirectional** nature of our model is the single most important new contribution”*



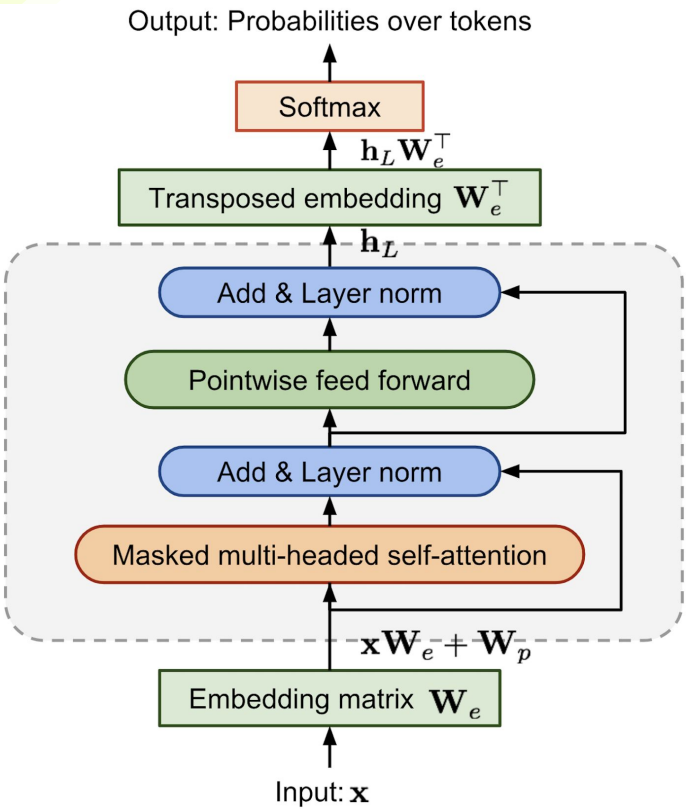
omnidireccional tal vez?

*"**bidirectional** nature of our model is the single most important new contribution"*

GPT y GPT 2



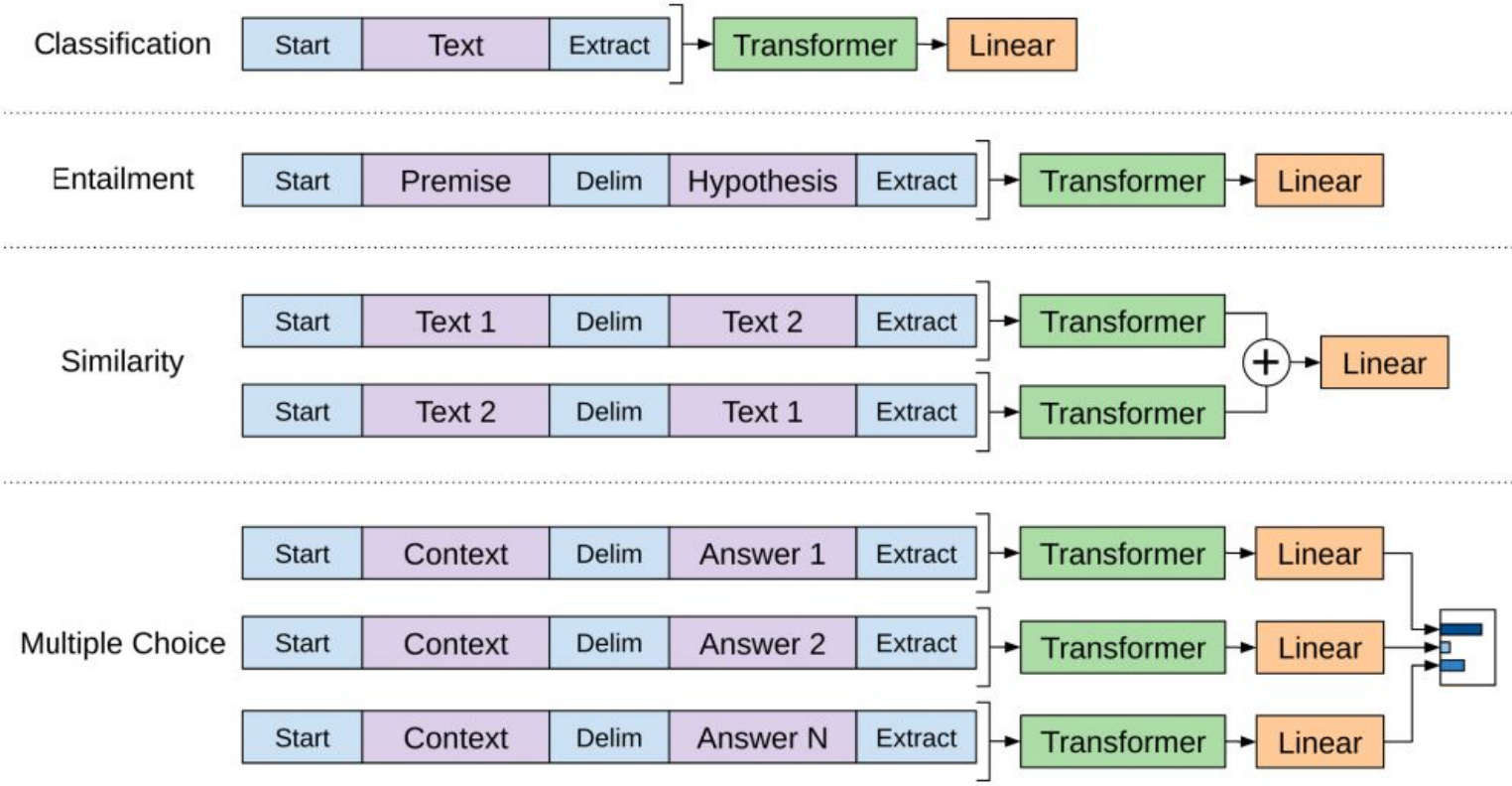
GPT [5]



Generative Pre-training Transformer utiliza un apilado de encoders (12).

La hipótesis es que un único modelo debería servir para mútiples tareas

GPT [5]


Classification
Start
Text
Extract
Transformer
Linear
Entailment
Start
Premise
Delim
Hypothesis
Extract
Transformer
Linear
Similarity
Start
Text 1
Delim
Text 2
Extract
Transformer
Start
Text 2
Delim
Text 1
Extract
Transformer
Linear
Multiple Choice
Start
Context
Delim
Answer 1
Extract
Transformer
Linear
Start
Context
Delim
Answer 2
Extract
Transformer
Linear
Start
Context
Delim
Answer N
Extract
Transformer
Linear

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Referencias

[1] Attention and memory in deep learning and nlp. Wildml

[2] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

[3] Attention Is All You Need

The illustrated transformer

The illustrated bert

[4] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Generalized Language Models

[5] Improving Language Understanding by Generative Pre-Training

[6] Language Models are Unsupervised Multitask Learners

Better language models and their implications