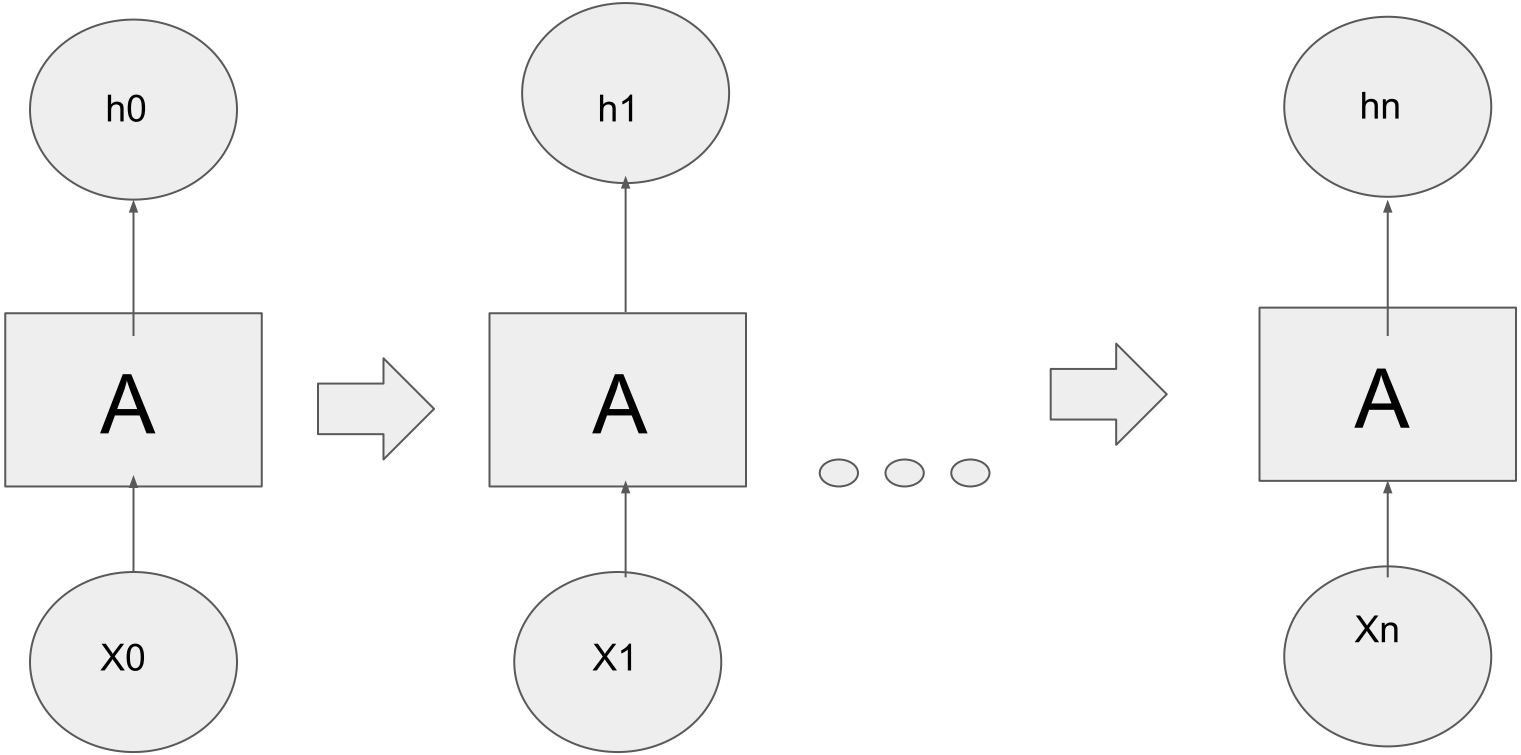
Machine Translation

mediante Redes Transformer

NLP, modelos anteriores:

- Bag of Words
- RNN (seq2seq)
- LSTM (seq2seq)
- Transformer

RNN



Vanishing & Exploding Gradients

$$H_{i+1} = A(H_i, x_i)$$

$$H_3 = A(A(A(H_0, x_0), x_1), x_2)$$

$$A(H, x) := \mathbf{W}x + \mathbf{Z}H$$

$$H_N = \mathbf{W}^N x_0 + \mathbf{W}^{N-1} x_1 + \ldots$$

```
>>> 0.9 ** 100
2.6561398887587544e-05
>>> 1.1 ** 100
13780.6123398238
>>> 0.9 ** 200
7.055079108655367e-10
>>> 1.1 ** 200
189905276.4604649
```

https://www.youtube.com/watch?v=S27pHKBEp30

LSTM

- Difíciles de entrenar
- Transfer learning difícilmente funcionan en estas redes
- Necesita un dataset específico para cada tarea
- Caminos de gradiente muy largos

Transformers

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

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after

Arquitectura

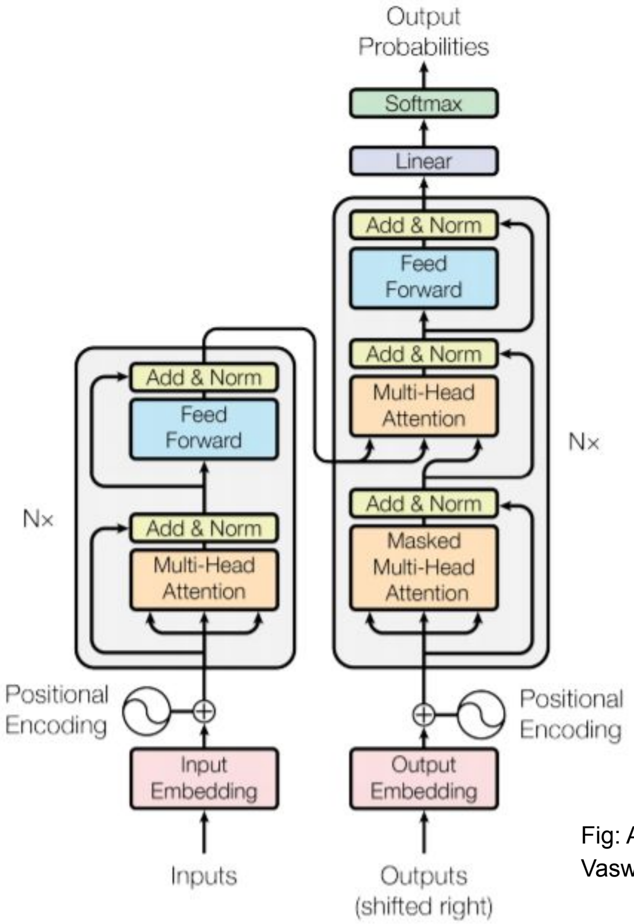


Figure 1: The Transformer - model architecture.

Fig: Arquitectura de la red Transformer, (Ashish Vaswani)

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

Figura: Attention is all you need,(Ashish Vaswani)

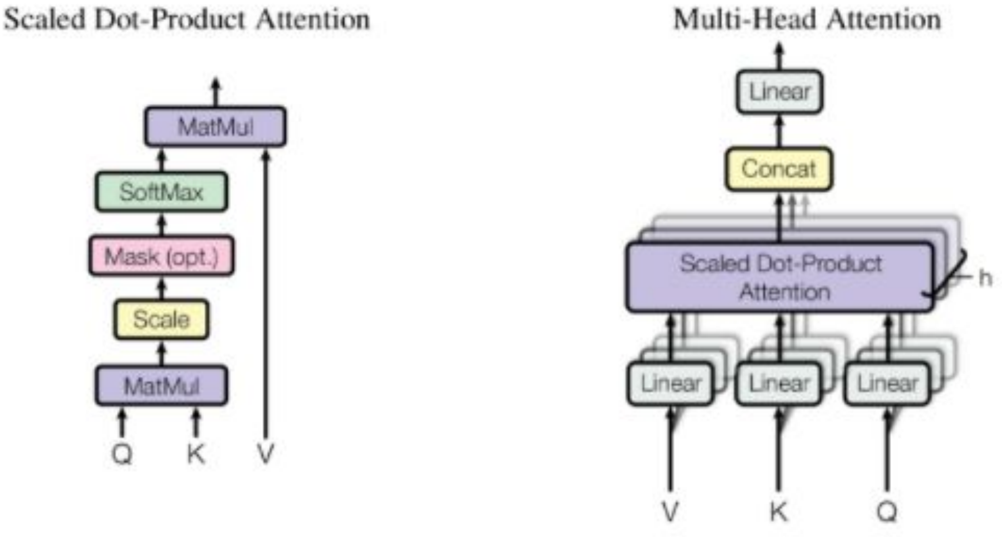
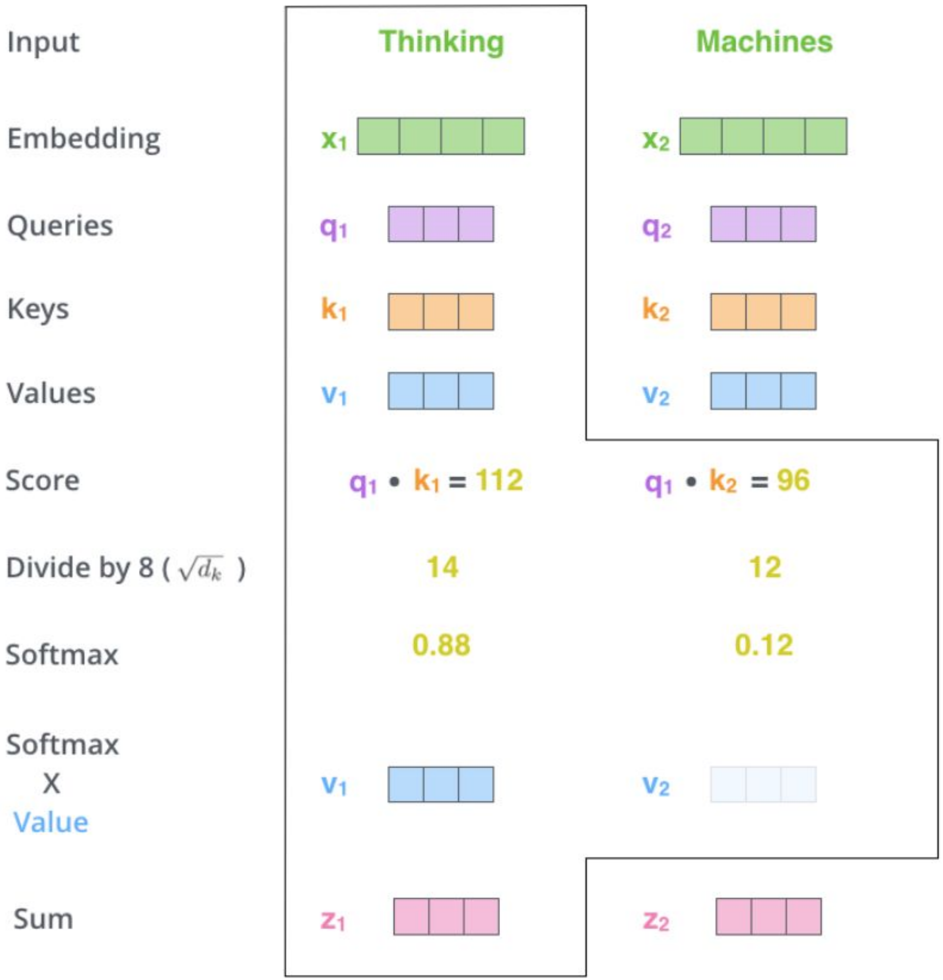
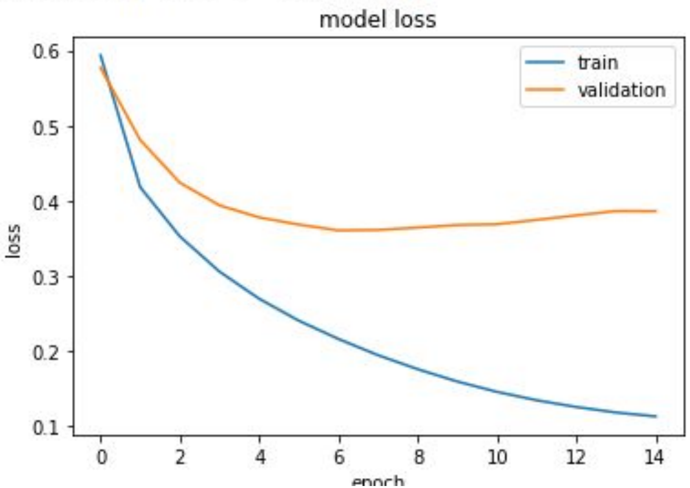


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

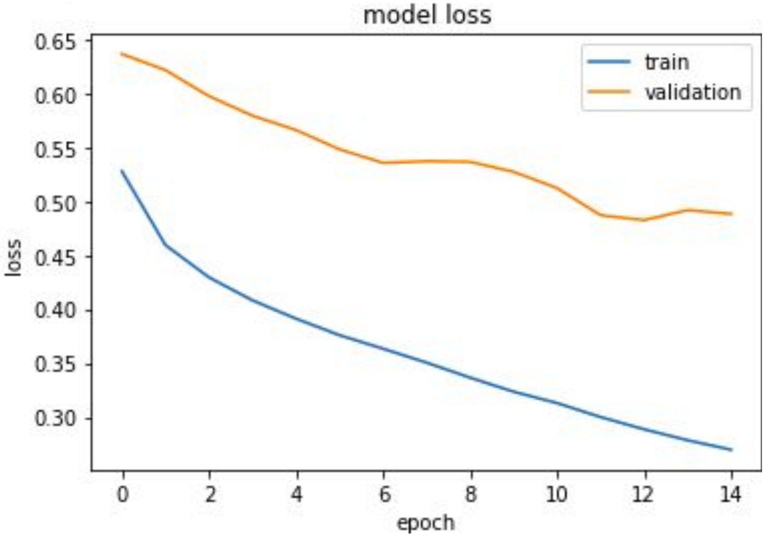
Figura: Attention is all you need,(Ashish Vaswani)



Input
Thinking
Machines
Embedding
x_1
x_2
Queries
q_1
q_2
Keys
k_1
k_2
Values
v_1
v_2
Score
$q_1 \cdot k_1 = 112$
$q_1 \cdot k_2 = 96$
Divide by 8 ($\sqrt{d_k}$)
14
12
Softmax
0.88
0.12
Softmax
X
Value
v_1
v_2
Sum
z_1
z_2



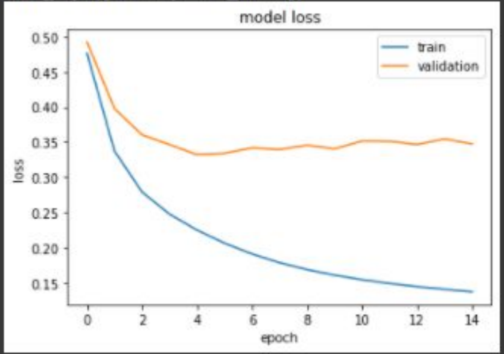
Modelo 2 heads, 256 batch, 1.9 mas frases en el dataset y shuffle True



Modelo 2 heads, 50 batch, 1.9 mas frases en el dataset y shuffle False

Resultados

```
dict_keys(['loss', 'val_loss'])
```




```
translate('the day is warm and sunny')
```

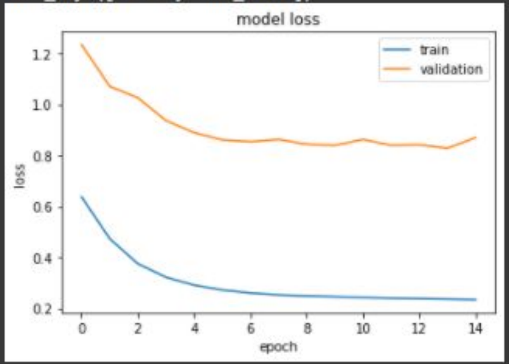
```
Frase original: the day is warm and sunny
Traducción: el dia esta caliente y crecio
```

```
translate('i am tired of college')
```

```
Frase original: i am tired of college
Traducción: estoy cansado de la universidad
```

Modelo 4 head, dataset 2 (1.9 mas)

```
dict_keys(['loss', 'val_loss'])
```



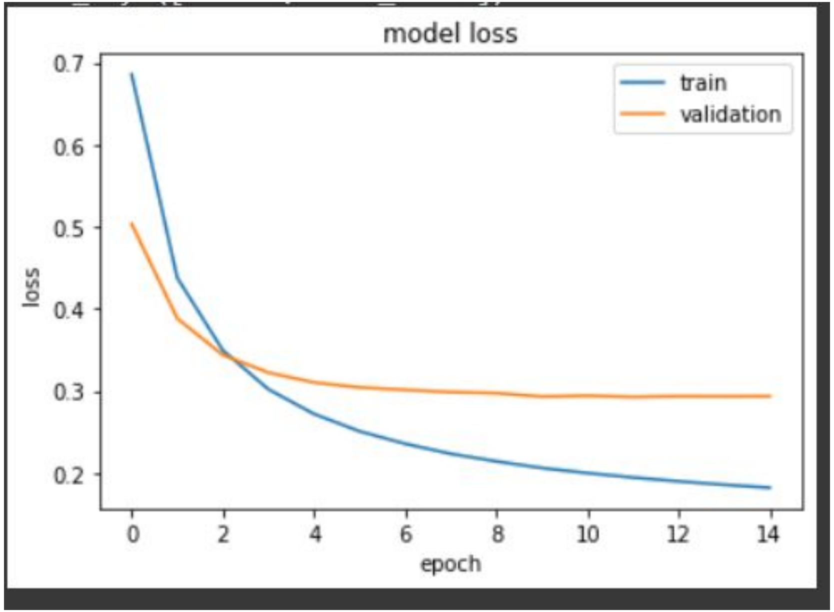

```
translate('the day is warm and sunny')
```

```
Frase original: the day is warm and sunny
Traducción: el dia tiene que tomar tiempo
```

```
[28] translate('i am tired of work even more of studying')
```

```
Frase original: i am tired of work even more of studying
Traducción: estoy cansado de trabajar incluso de mas de estudiar
```

Modelo 2 head,dataset 2 ,shuffle false



Modelo 2 head, dataset org, shuffle true