Attention Is All You Need

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., et al.

Presentador: Matías Marambio Jiménez

Transduction, Sequence modelling

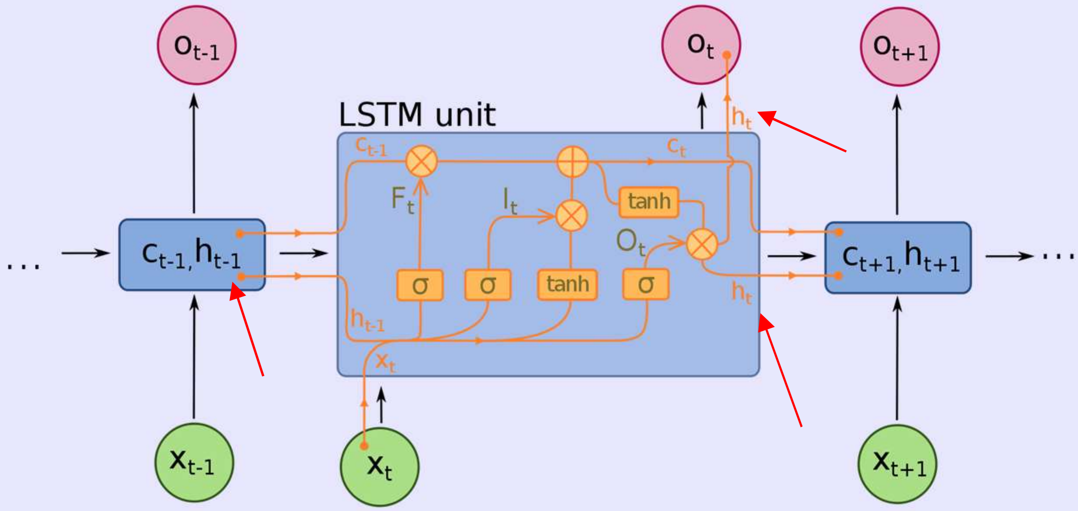
- Modelamiento de Lenguaje.
- Traducción Automática (Machine Translation).

Transduction, Sequence modelling

¿cómo se hace? (2017)

- Arquitecturas codificador-decodificador.
- Self-attention
- Modelos de lenguaje recurrentes.

$$h_t(h_{t-1})$$



¿Qué pasa con las secuencias largas?

Nuevo método: Transformer

- Arquitectura codificador-decodificador. ✓
- Sólo utiliza self-attention. ✓

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

1. Baja complejidad por capa.
2. Cantidad de computación que puede ser paralelizada.
3. Caminos cortos entre dependencias largas de la secuencia

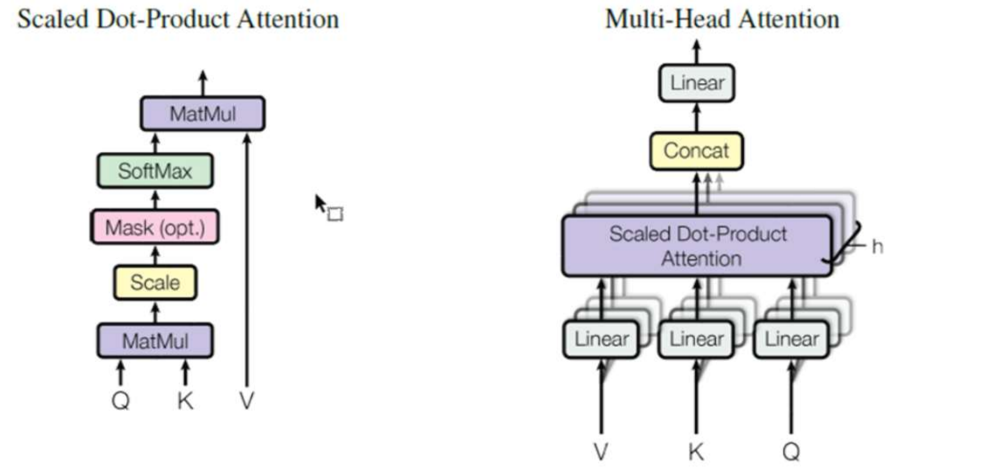
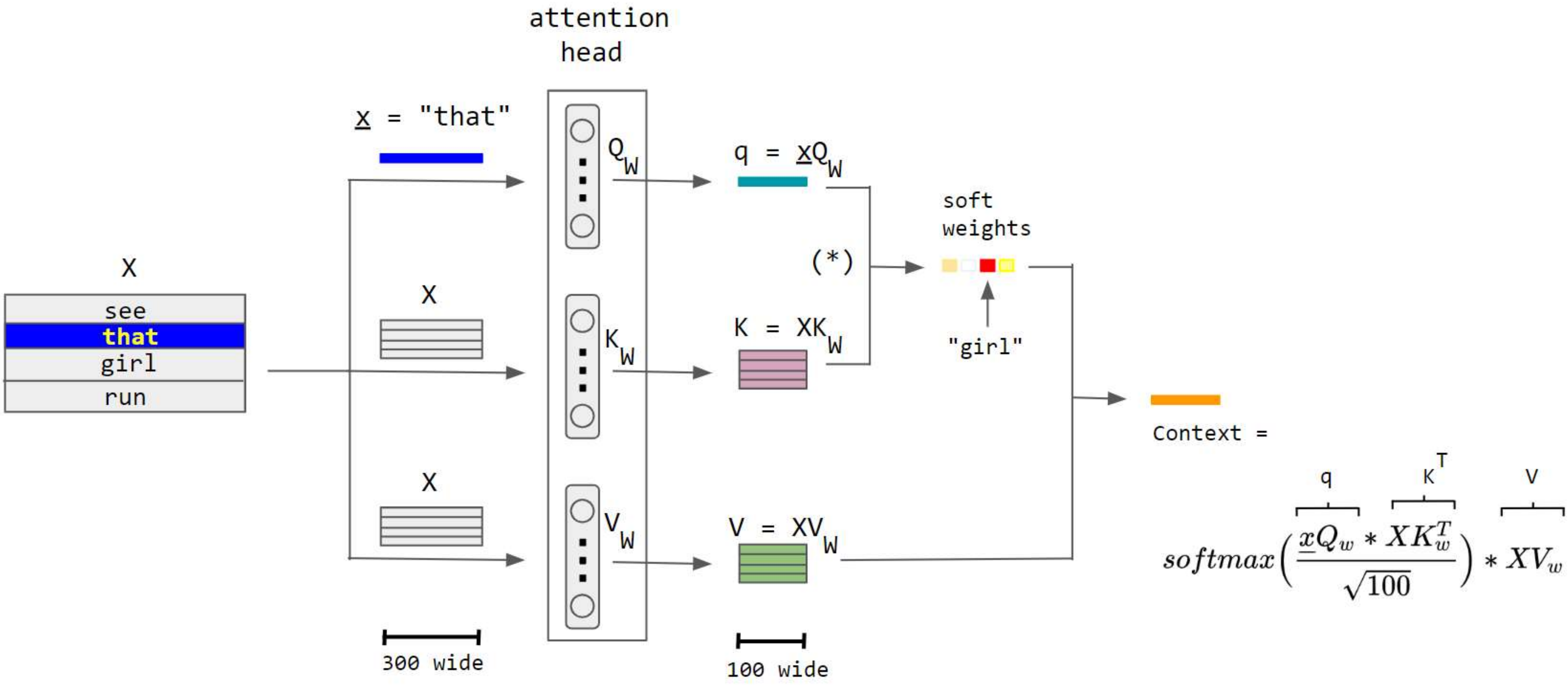


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

[redacted]



attention head

$\underline{x}$ = "that"

X

see

that

girl

run

X

X

Q_W

K_W

V_W

$q = \underline{x}Q_W$

$K = XK_W$

$V = XV_W$

(*)

soft weights

"girl"

Context =

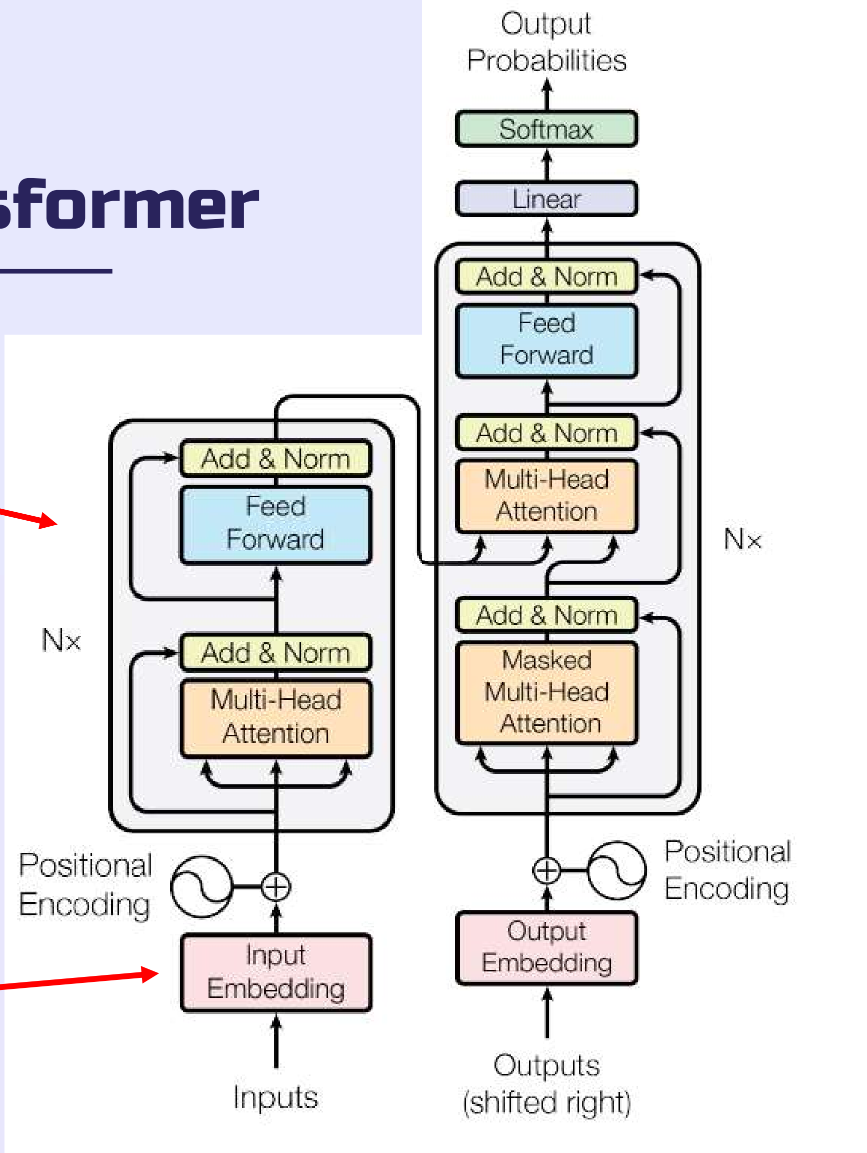
$$softmax\left(\frac{\overbrace{\underline{x}Q_w}^{q} * \overbrace{XK_w^T}^{K^T}}{\sqrt{100}}\right) * \overbrace{XV_w}^{V}$$

300 wide

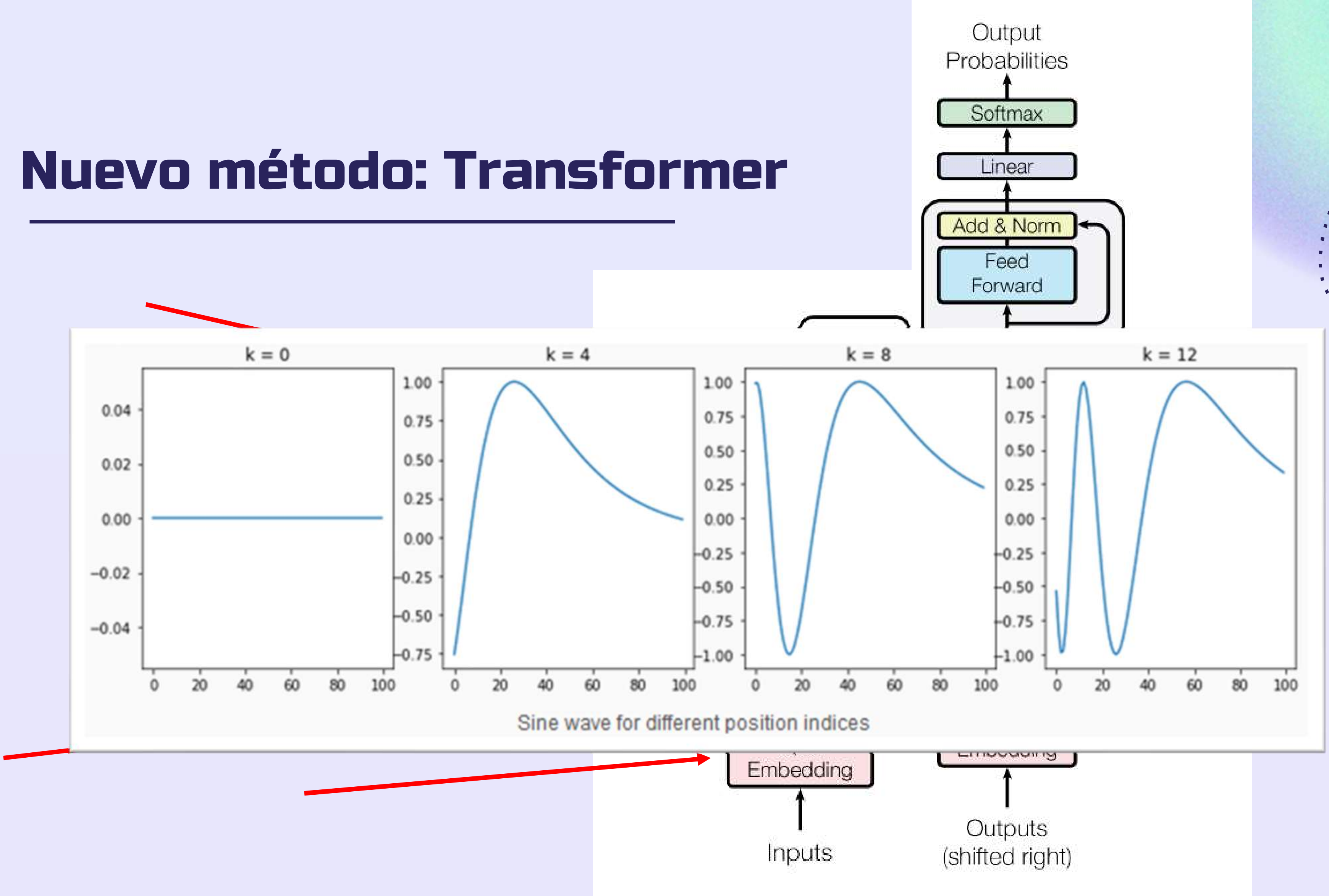
100 wide

[redacted]

Nuevo método: Transformer



Nuevo método: Transformer



Sine wave for different position indices

Resultados

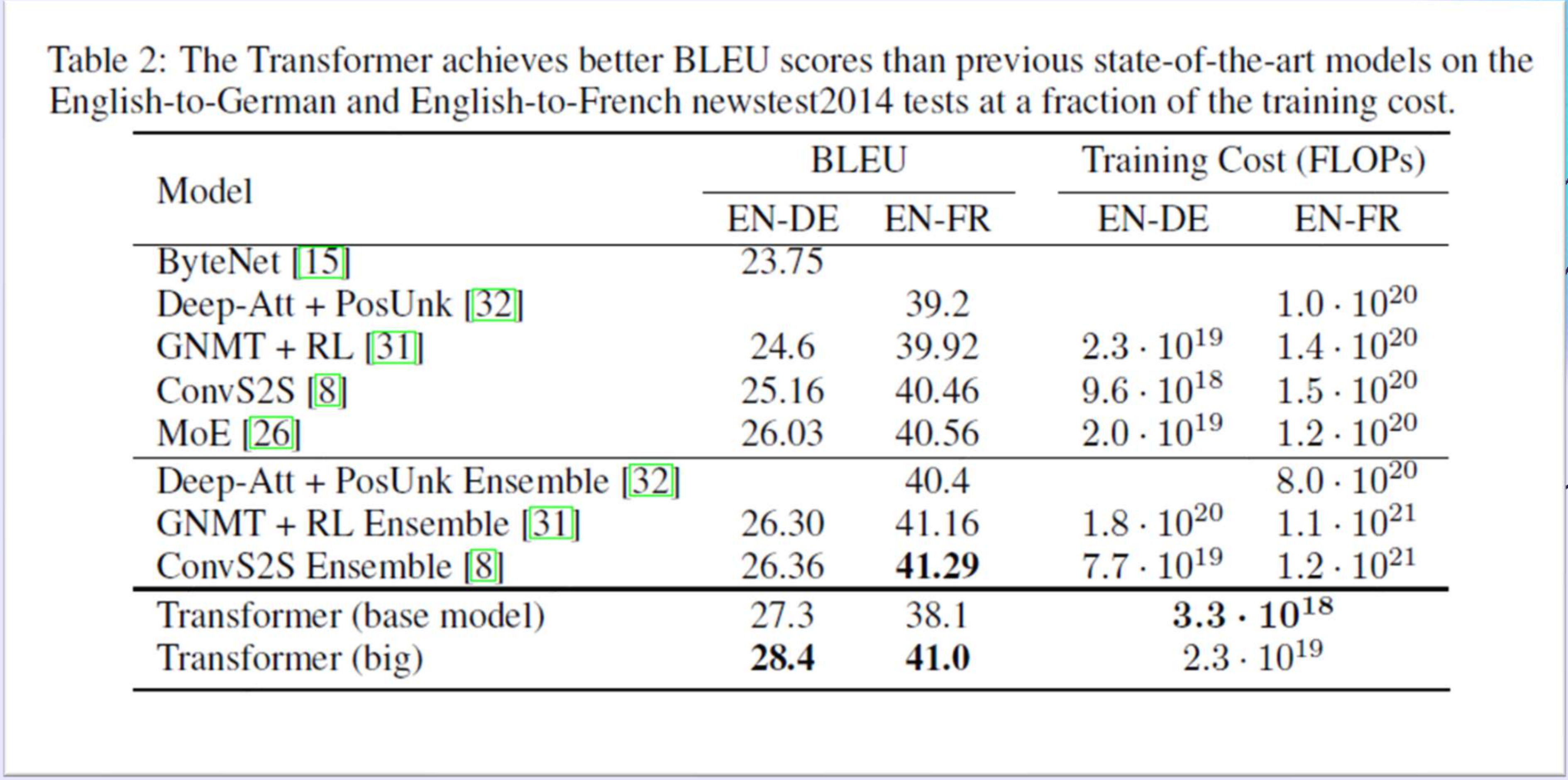


Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Conclusiones

Transformer

Basado en self-attention

Entrenan rápido

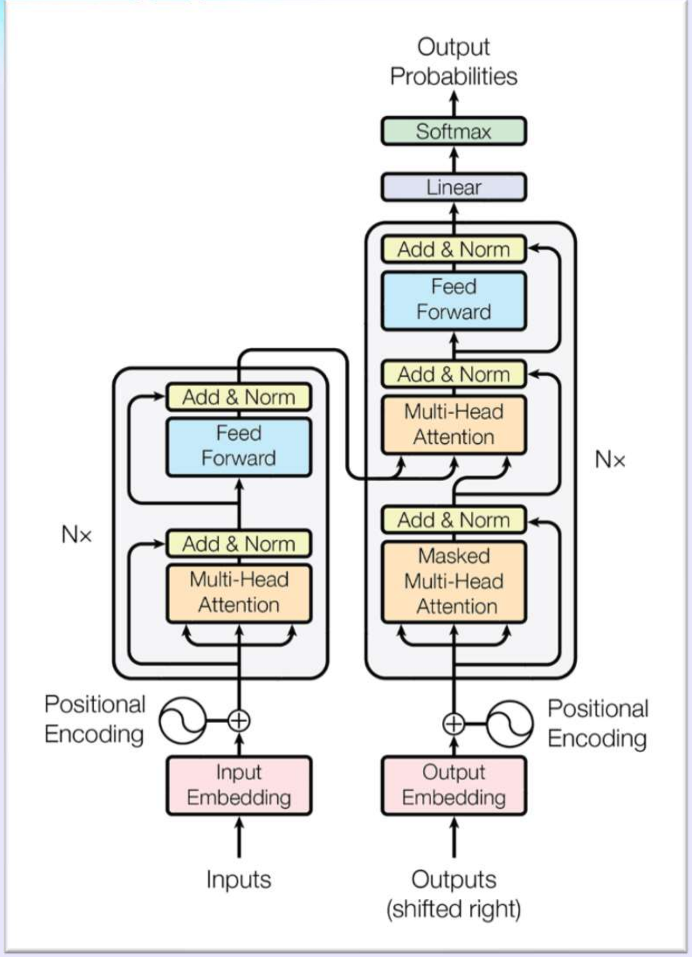
Comparado con recurrente y convolucional

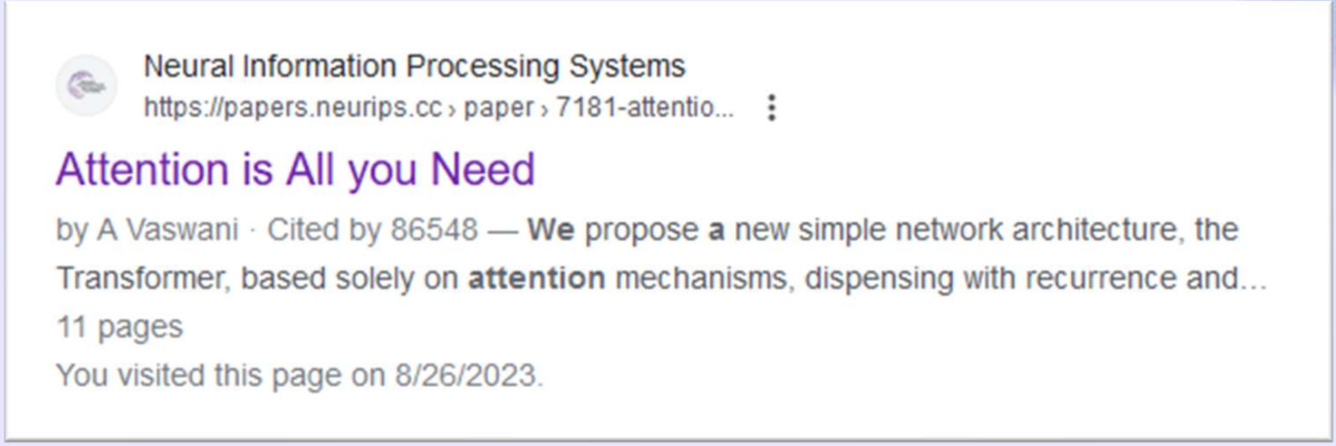
Nuevo estado del arte

WMT2014 EN-DE
WMT2014 EN-FR

Trabajo futuro

Inputs que no sean texto





Neural Information Processing Systems
https://papers.neurips.cc › paper › 7181-attentio...

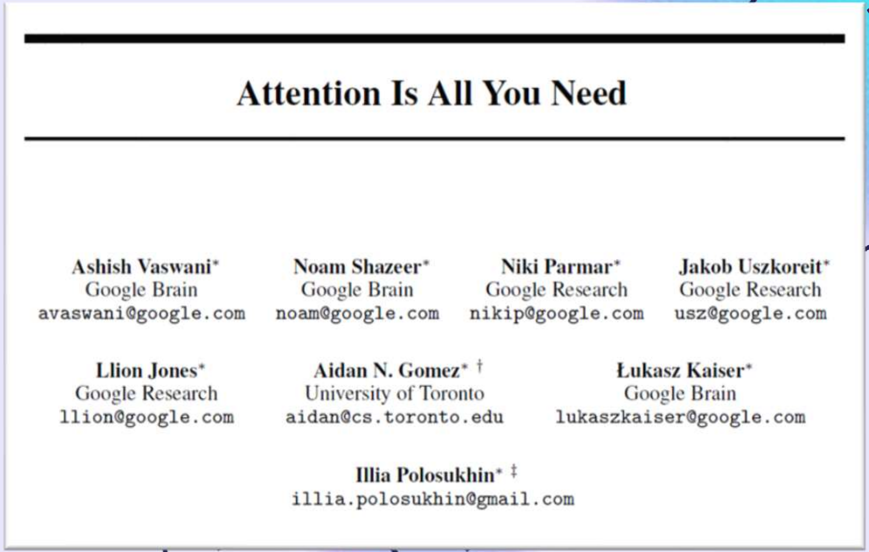
Attention is All you Need

by A Vaswani · Cited by 86548 — **We** propose **a** new simple network architecture, the Transformer, based solely on **attention** mechanisms, dispensing with recurrence and...

11 pages

You visited this page on 8/26/2023.

Gracias



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

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**