TRANSFORMERS

SISTEMAS DE INTELIGENCIA ARTIFICIAL - 2025

TABLA DE CONTENIDOS

01

INTRODUCCIÓN

¿Qué es un Transformer?

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

[Paper](#)

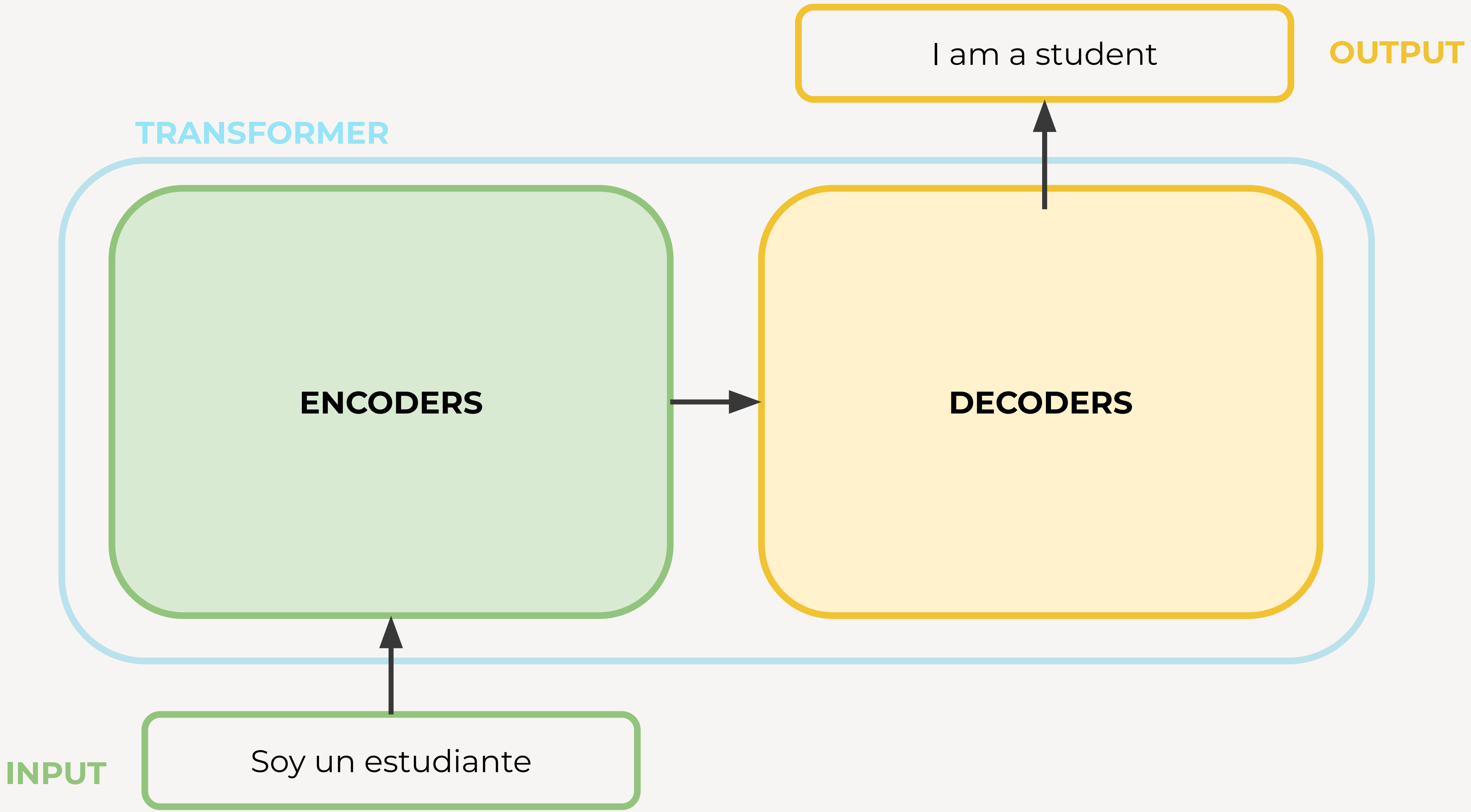
¿QUÉ ES UN TRANSFORMER?

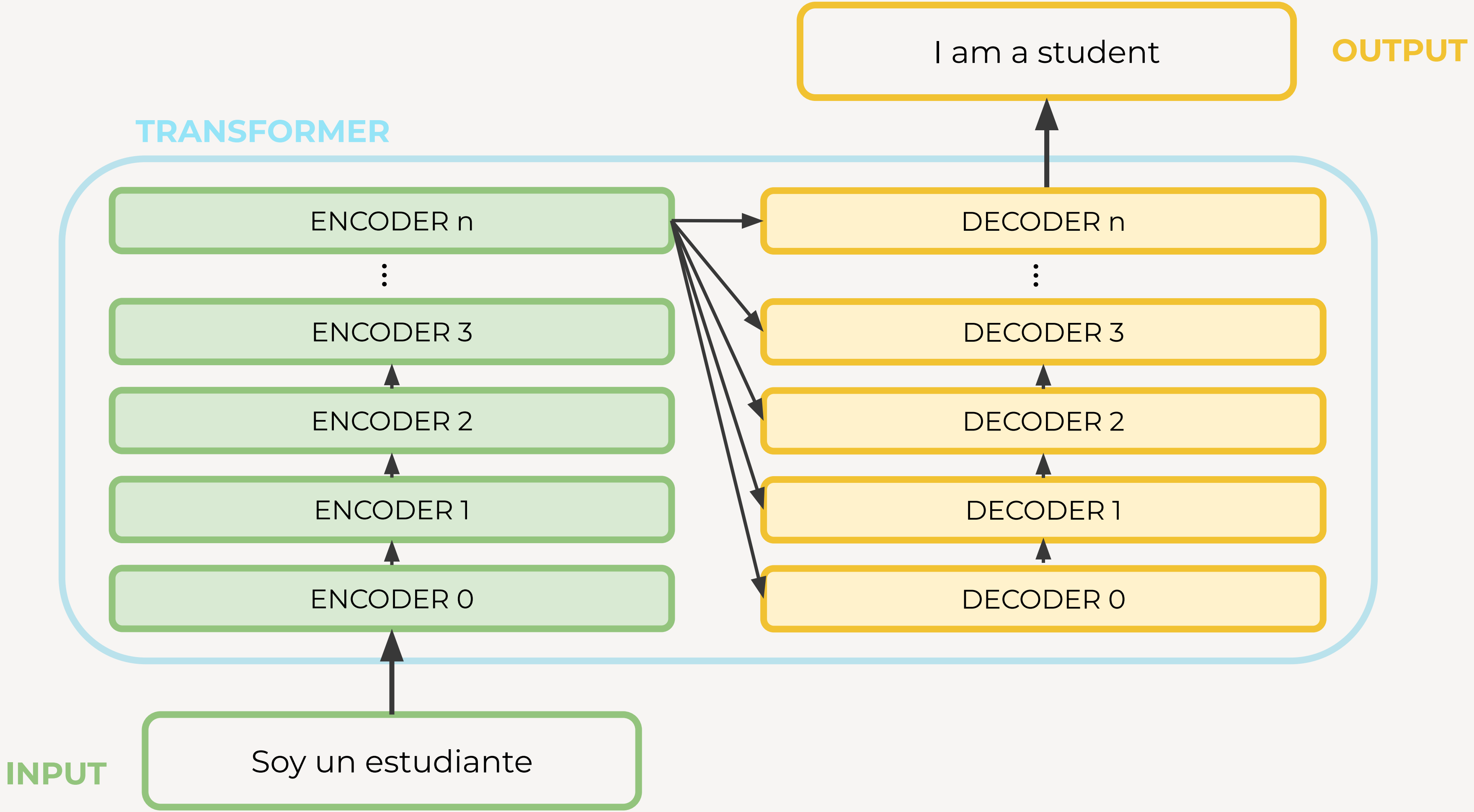
Red neuronal utilizada para transformar una secuencia en otra basada en un mecanismo de **atención**.

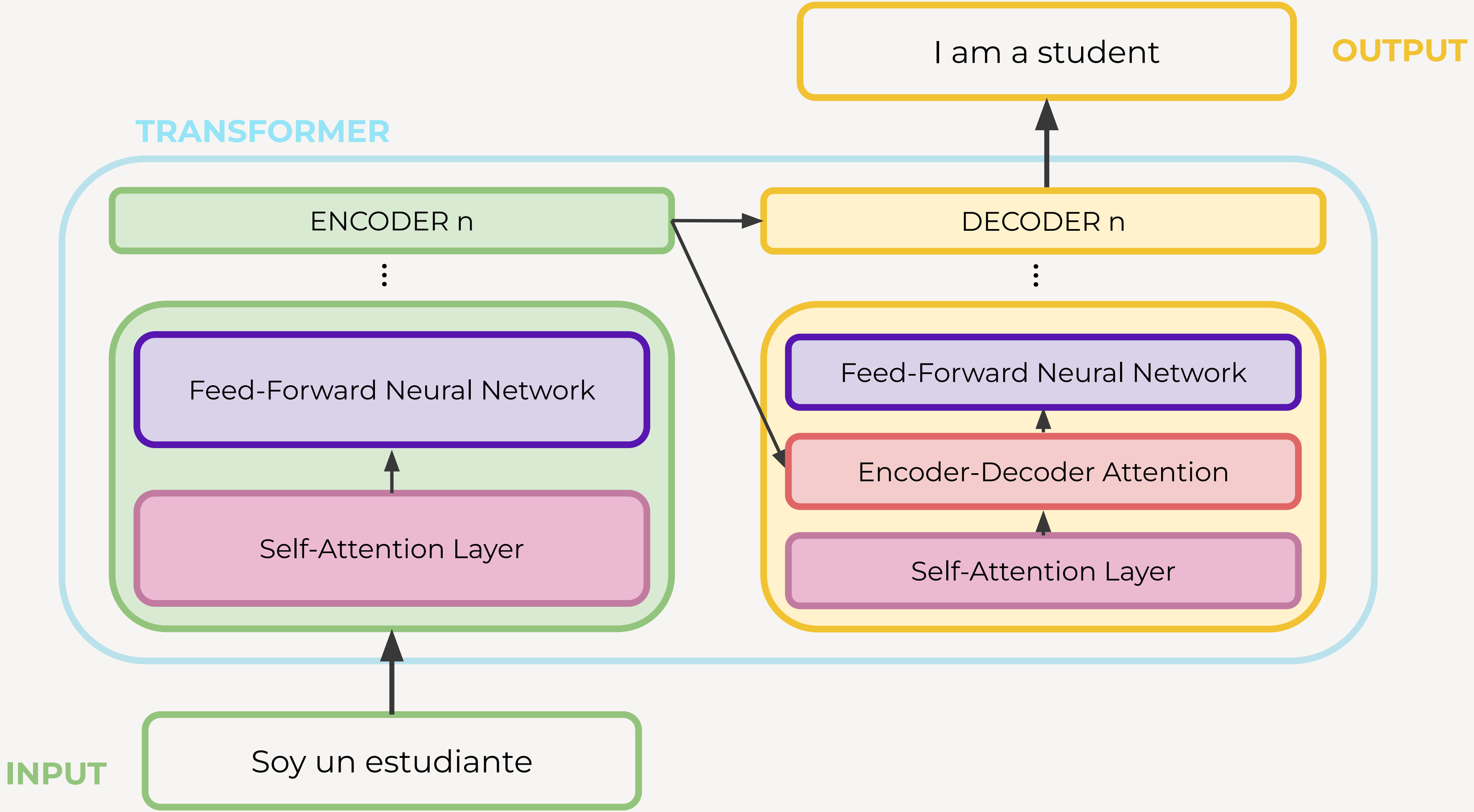
02

ARQUITECTURA

¿Cuáles son las capas ocultas de un Transformer?


OUTPUT
I am a student
TRANSFORMER
ENCODERS
DECODERS
INPUT
Soy un estudiante


I am a student
OUTPUT
TRANSFORMER
ENCODER n
ENCODER 3
ENCODER 2
ENCODER 1
ENCODER 0
DECODER n
DECODER 3
DECODER 2
DECODER 1
DECODER 0
INPUT
Soy un estudiante


I am a student
OUTPUT
TRANSFORMER
ENCODER n
DECODER n
Feed-Forward Neural Network
Self-Attention Layer
Feed-Forward Neural Network
Encoder-Decoder Attention
Self-Attention Layer
INPUT
Soy un estudiante

SELF-ATTENTION LAYER

Ayudar al Encoder a **mirar otras palabras** en la secuencia de input mientras analiza una palabra en particular.

En el Decoder solo puede mirar palabras previas en la secuencia de output.

INPUT

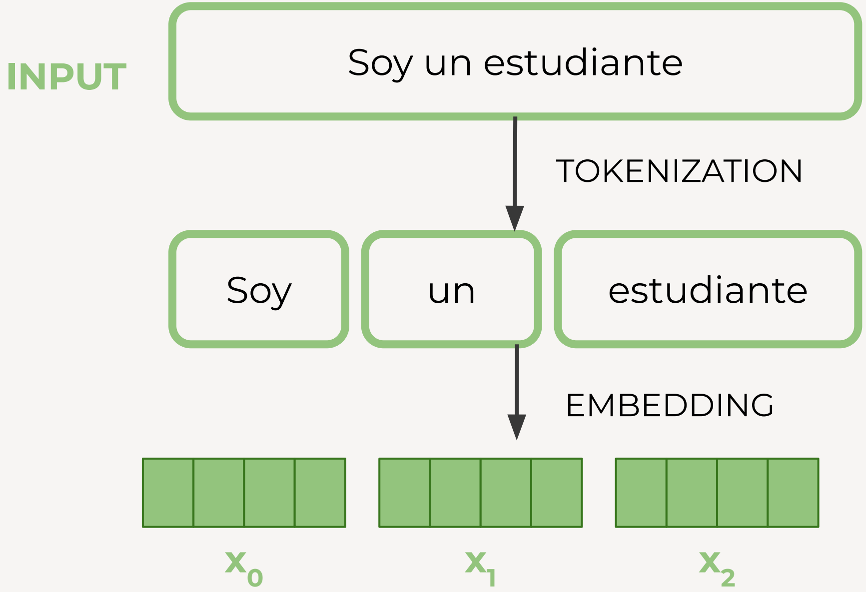
The animal didn't cross the street because **it** was tired

¿La palabra *"it"* se refiere al animal o a la calle ?

EMBEDDING

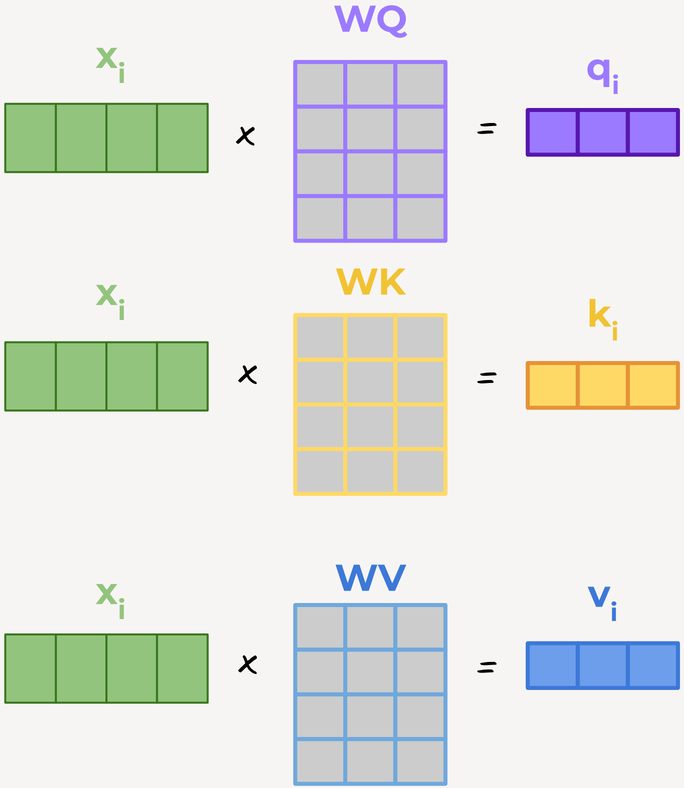
- Cada token es representado en un vector de números reales de dimensión 512.
  A esto se lo llama ***embedding*** (algoritmos: *GloVe* o *word2vec: SkipGram - CBOW*)

- El primer encoder recibe como input el *embedding*.



CÁLCULO DE SELF-ATTENTION

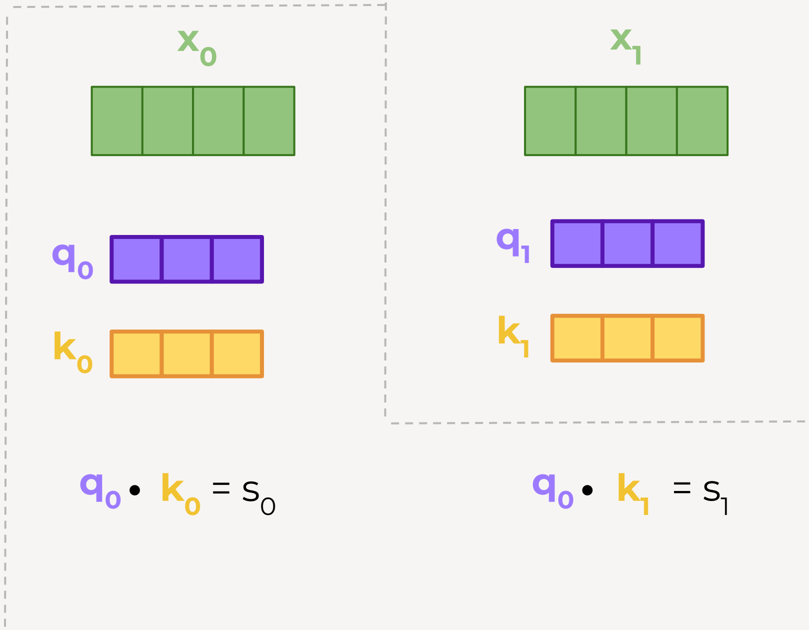
1. Por cada embedding x_i , se calculan 3 vectores llamados **query**, **key** y **value** multiplicando los embeddings por 3 matrices de pesos (*Attention Weights*) **WQ**, **WK** y **WV**



CÁLCULO DE SELF-ATTENTION

2. Se calcula un "score" por cada embedding que determina cuánta atención se le pone a ese embedding mientras se está procesando otro.

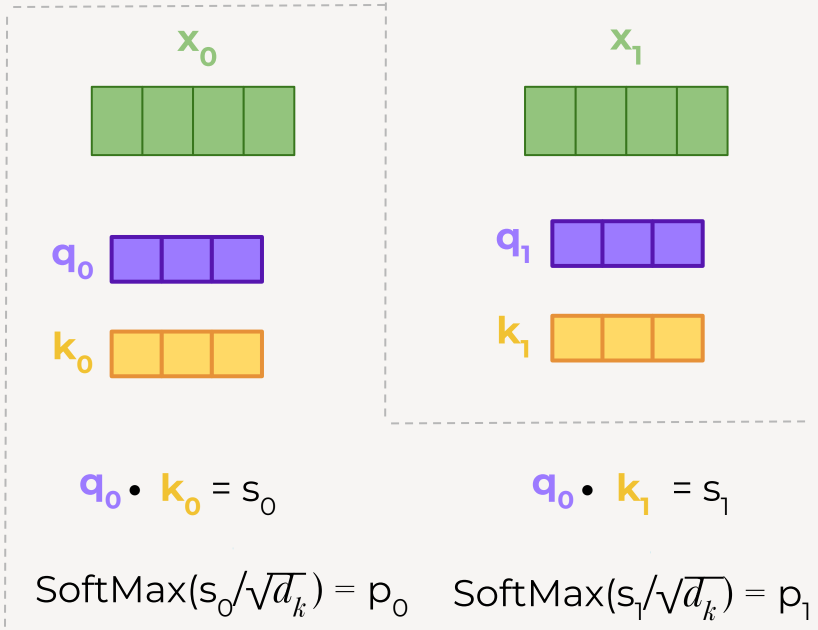
Se hace el producto escalar entre **q** y el vector **k** de cada embedding.



CÁLCULO DE SELF-ATTENTION

3. Se dividen los *scores* por la raíz de la dimensión de vector **k** (d_k)

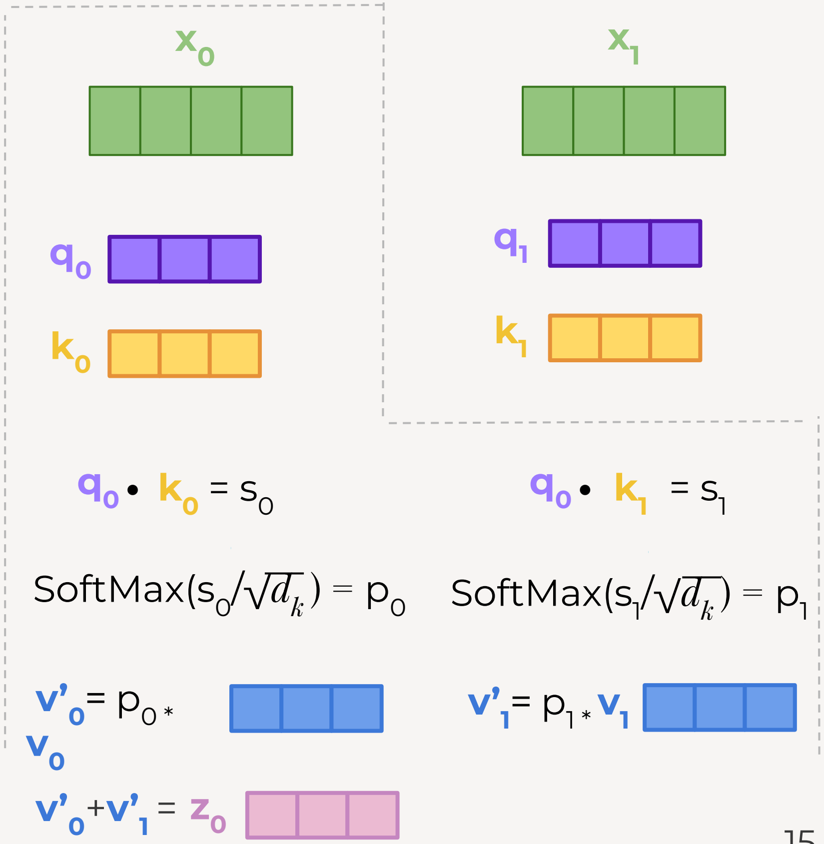
4. Usando *Softmax* para los scores *se* obtienen probabilidades (p_i).



CÁLCULO DE SELF-ATTENTION

5. Se multiplica cada vector **v** por lo obtenido en la función de Softmax.

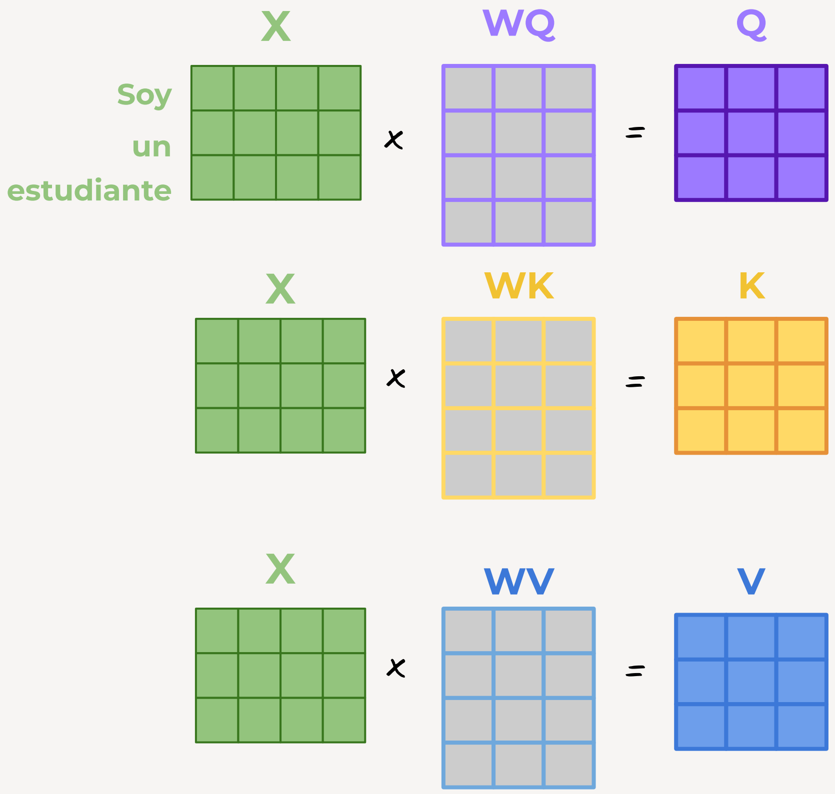
6. Se suman los vectores ponderados **v'** obteniendo el output de la capa de Self-Attention para cada token (en este ejemplo, para el primero)

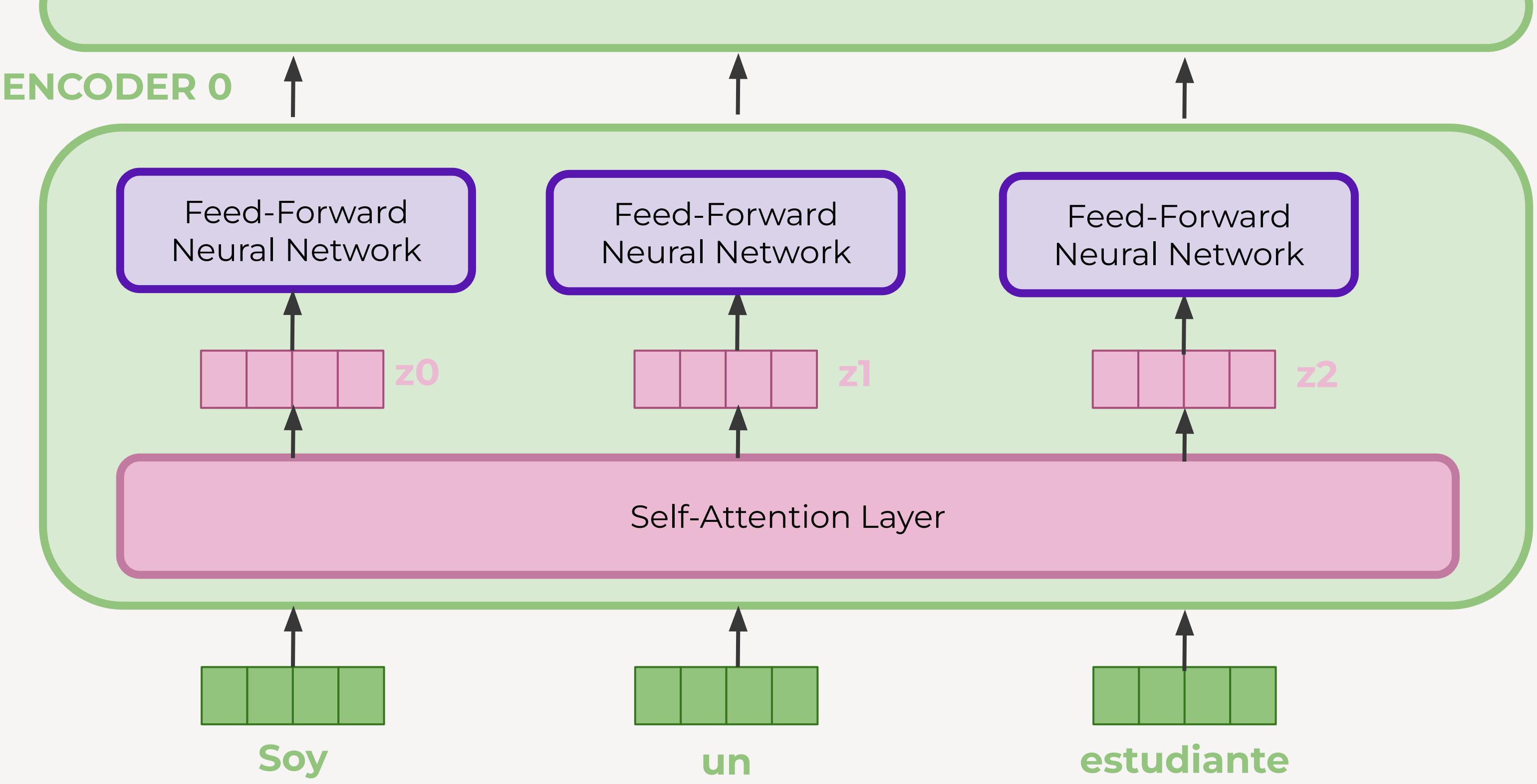


CÁLCULO DE SELF-ATTENTION

Para implementarlo se calcula de forma matricial para un procesamiento más rápido

Attention(K,Q,V) = SoftMax(**QK'** / $\sqrt{d_k}$) * **V** = **Z**



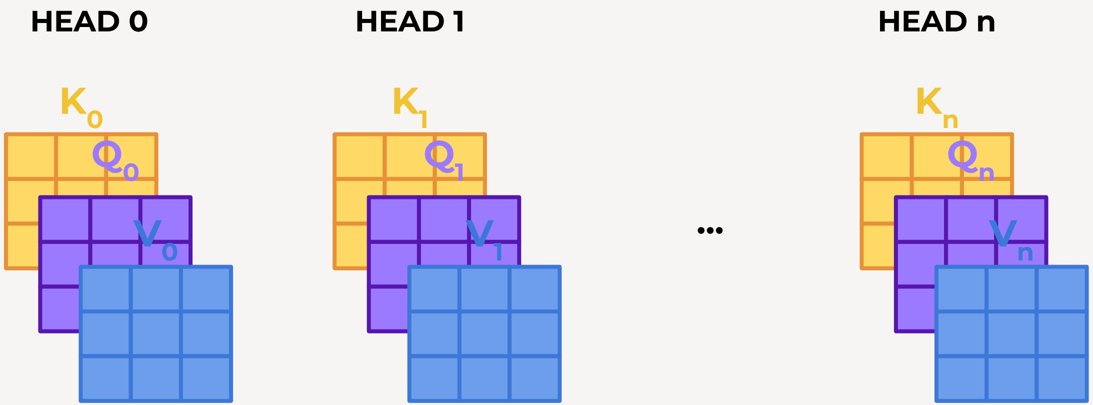


- Cada token sigue su **propio camino** en el encoder.
- Hay **dependencias** entre los caminos en la self-attention layer pero no en la red feed-forward.
- Los caminos se pueden ejecutar en **paralelo,** pasando por **redes feed-forward idénticas**

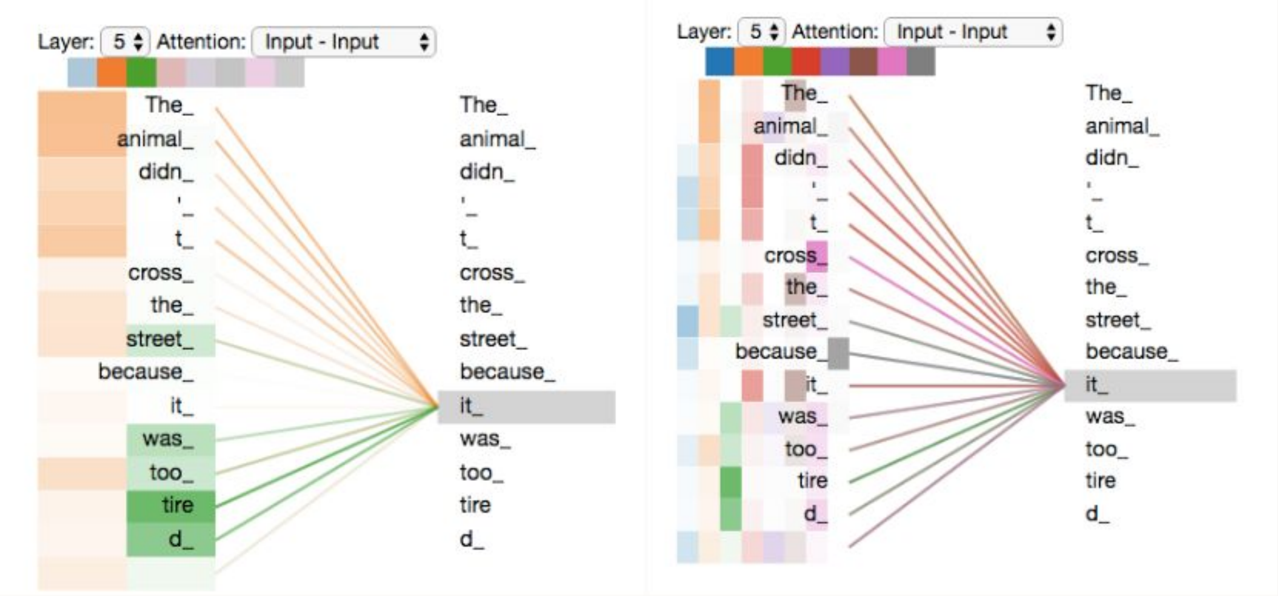
MULTI-HEADED ATTENTION

Mecanismo que mejora la performance de la capa de self-attention:

- El Encoder se puede enfocar en distintas posiciones de la secuencia **simultáneamente**
- Se tienen múltiples sets **K Q V**, llamados *“heads”*



MULTI-HEADED ATTENTION



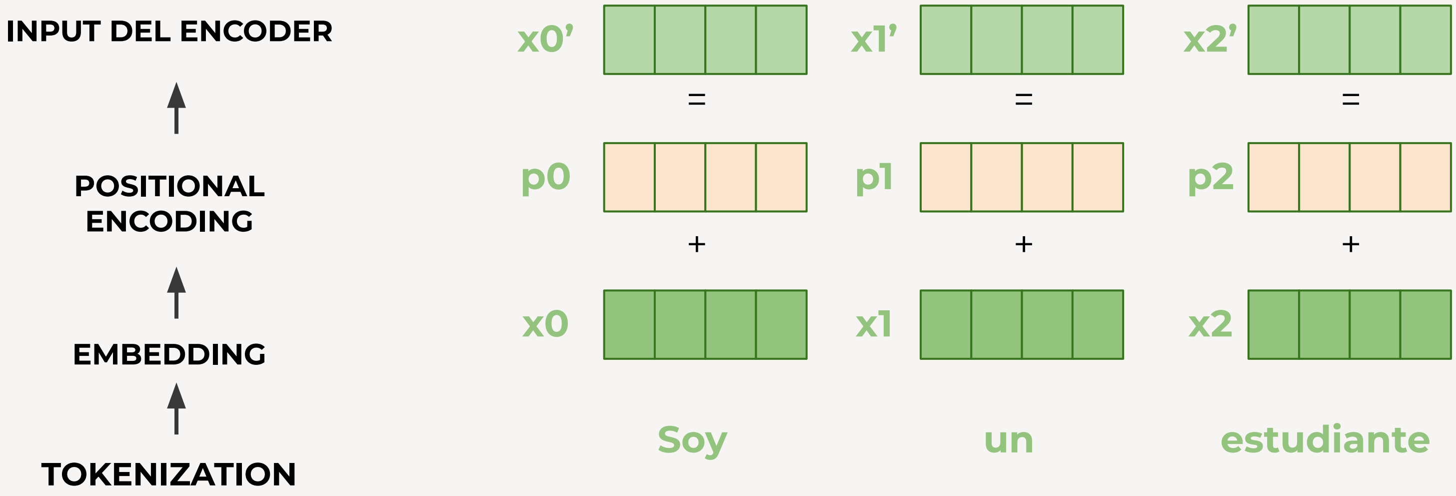
[Multi-Headed Attention]

POSITIONAL ENCODING

Representa el orden de los tokens en la secuencia tomada como input

A cada embedding se le agrega un vector **p**

Los vectores p siguen un patrón que el modelo aprende para determinar la distancia entre tokens

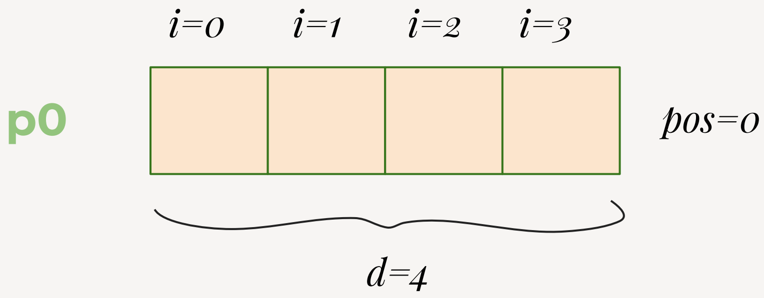


POSITIONAL ENCODING

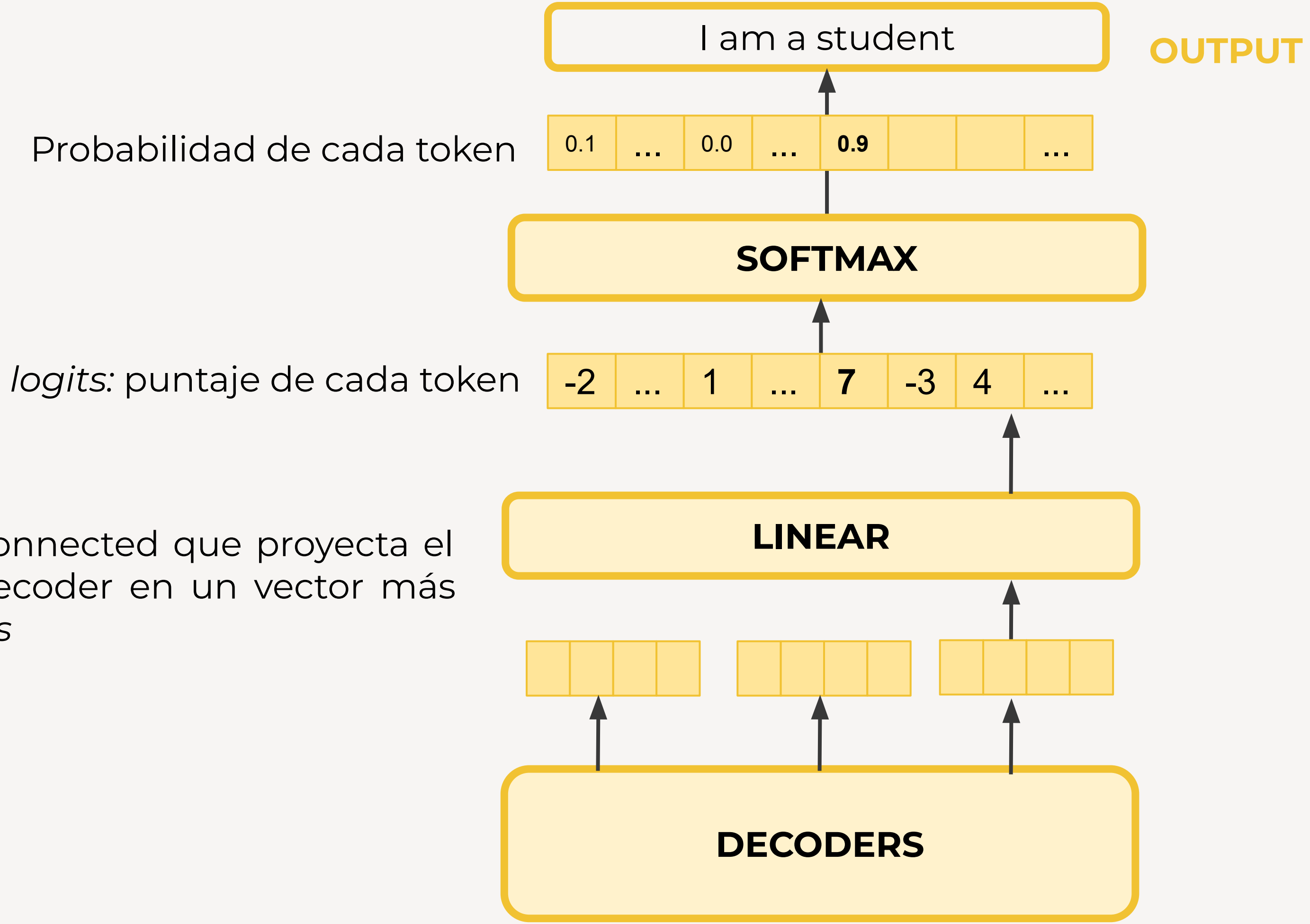
$$PE(pos, 2i) \quad = sen(pos/\ 10000^{2i/d})$$

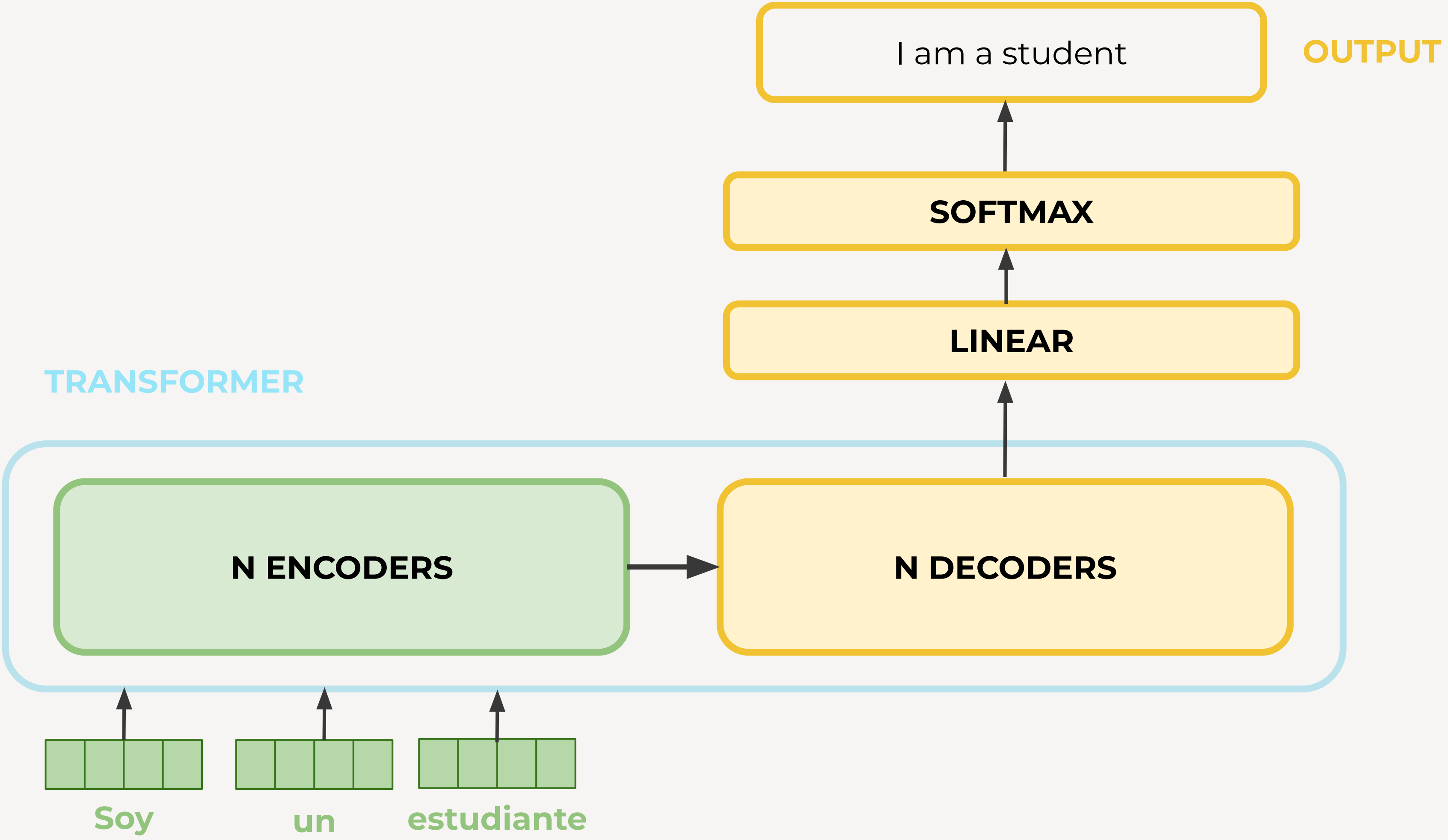
$$PE(pos, 2i+1) = cos(pos/\ 10000^{2i/d})$$

Donde:
pos es la posición del token en la secuencia
i es el índice del vector de posición
d es la dimensión de un embedding, la misma del vector de posición p_i



LINEAR
Red neuronal fully-connected que proyecta el output del último decoder en un vector más grande llamado *logits*




I am a student
OUTPUT
SOFTMAX
LINEAR
TRANSFORMER
N ENCODERS
N DECODERS
Soy
un
estudiante

03

ENTRENAMIENTO

¿Cómo aprende un Transformer?

CASO DE USO - TRADUCCIÓN

Se quiere traducir *“gracias”* a *“thanks”*.

Se tiene un vocabulario de salida *v = (a, am, i, thanks, student, <eof>).*

ENTRENAMIENTO

Los **pesos se inicializan de forma aleatoria**, entonces el modelo sin entrenar devuelve una distribución de probabilidades con valores arbitrarios para cada token

Se pueden comparar las probabilidades y por **backpropagation** nos vamos acercando al output deseado.

	a	am	i	***thanks***	student	<eof>
OUTPUT OBTENIDO	0.2	0.2	0.1	0.2	0.2	0.1
OUTPUT DESEADO	0.0	0.0	0.0	1.0	0.0	0.0

COSTO

Entrenar un Transformer requiere de gran costo computacional

6 Results

6.1 Machine Translation

On the WMT 2014 English-to-German translation task, the big transformer model (Transformer (big) in Table 2) outperforms the best previously reported models (including ensembles) by more than 2.0 BLEU, establishing a new state-of-the-art BLEU score of 28.4. The configuration of this model is listed in the bottom line of Table 3. Training took 3.5 days on 8 P100 GPUs. Even our base model surpasses all previously published models and ensembles, at a fraction of the training cost of any of the competitive models.

Vaswani, A., et al. (2017). *Attention is all you need.* Advances in neural information processing systems, 30.

VARIACIÓN DE PARÁMETROS

Table 3: Variations on the Transformer architecture. Unlisted values are identical to those of the base model. All metrics are on the English-to-German translation development set, newstest2013. Listed perplexities are per-wordpiece, according to our byte-pair encoding, and should not be compared to per-word perplexities.

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65
(A)				1	512	512				5.29	24.9	
				4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16					5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4									5.19	25.3	50
	8									4.88	25.5	80
		256			32	32				5.75	24.5	28
		1024			128	128				4.66	26.0	168
			1024							5.12	25.4	53
			4096							4.75	26.2	90
(D)							0.0			5.77	24.6	
							0.2			4.95	25.5	
								0.0		4.67	25.3	
								0.2		5.47	25.7	
(E)		positional embedding instead of sinusoids								4.92	25.7	
big	6	1024	4096	16			0.3		300K	**4.33**	**26.4**	213

Vaswani, A., et al. (2017). *Attention is all you need.* Advances in neural information processing systems, 30.

04

APLICACIONES

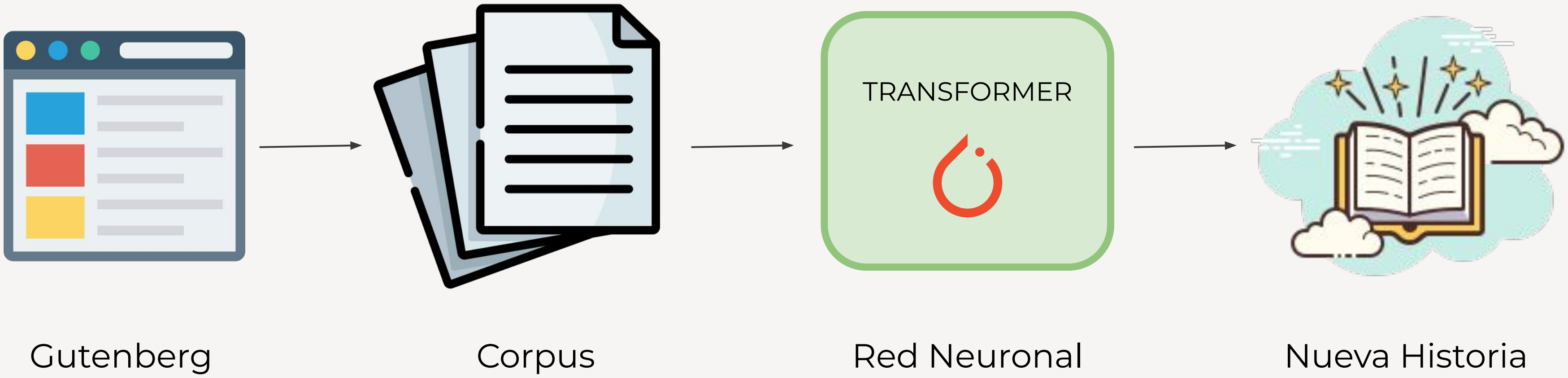
¿Para qué puedo usar un Transformer?

4.1

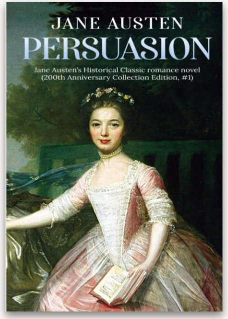
TEXTO

Aplicaciones

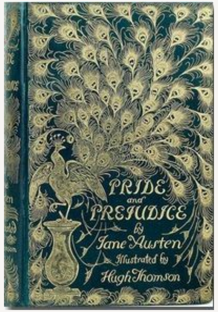
BARD TEXT GENERATOR

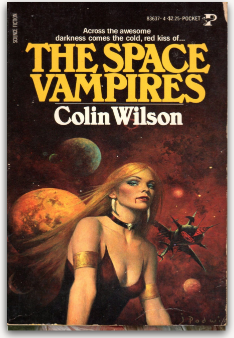


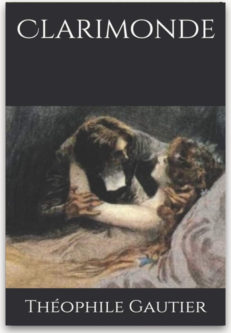
CORPUS

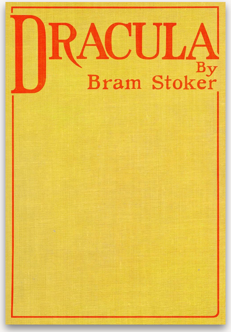

JANE AUSTEN
PERSUASION
Jane Austen's Historical Classic romance novel
(200th Anniversary Collection Edition, #1)

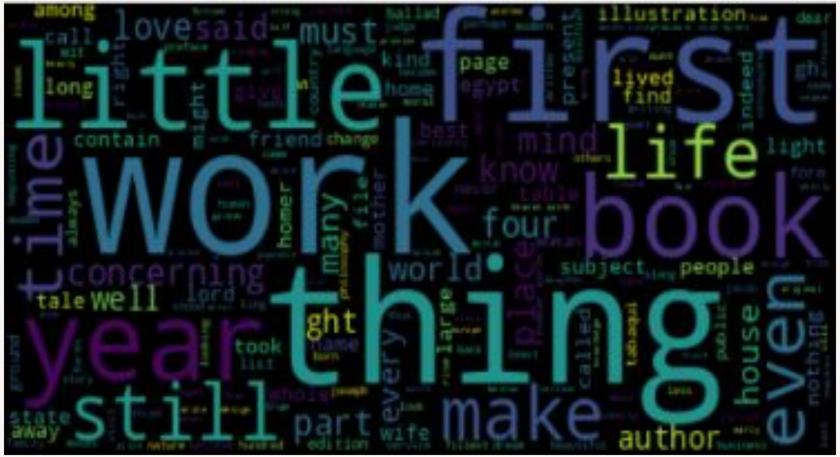

LITTLE WOMEN
MEG, JO, BETH, AND AMY
LOUISA M. ALCOTT


PRIDE and PREJUDICE
by
Jane Austen
Illustrated by
Hugh Thomson


Across the awesome
darkness comes the cold, red kiss of...
THE SPACE VAMPIRES
Colin Wilson


CLARIMONDE
THÉOPHILE GAUTIER


DRACULA
By
Bram Stoker


little
first
work
book
life
thing
year
still
make
time
four
world
author
even
house
love
said
must


something
vampyre
kennedy
admiral
francis
marchdale
anything
strange
present
bannerworth
chillingworth
feeling
looking
morning
brother
charles
believe
perhaps
together
brought
replied

GENERACIÓN

Época 1/100: “This story is about a woman”

Época 2/100: “This story is about a jungle for one , that he was the”

…

Época 20/100: “This story is about a good fortune and messages”

…

Época 60/100: “This story is about a to new truth , and be impossible for _us _ to visit”

…

Época 100/100: this story is about a man of the to names of ten years say ,

GENERACIÓN

ÉPOCAS MODELO	10 FineTuned
MYSTERY CORPUS	

A small, tightly **wound** body found in San Bernardino County that **police** say may have come from the former San Bernardino **shooter.**

Police at San Bernardino County State Police began searching for the 32-month-old San Bernardino resident, who it said was in the early morning hours of Nov. 7, 2015.

Authorities were also interviewing a man the county's first **responders found at the scene.**

The woman's death is the latest in a string of shootings in

GENERACIÓN

ÉPOCAS MODELO

10 FineTuned

VAMPYRE CORPUS

A dark shadow has been found in the sky. The **ghostly creature's** name has not been found in the sky with **no explanation,** and no one knows the cause of it.

You can see it coming out of your mouth on this little piece of paper. It's tiny, yet you need only look at it to know if it actually was the ghost of the man who had killed the Earth

She was in the throes of **madness and insanity. He had her hands clenched around her neck** and she could hear her breathing on the floor of his room.

"Ahm...**what's your name?**" I looked over at her face. "Ahm...you haven't seen me like before."

TEXT COMPLETION

ÉPOCAS
MODELO

10
FineTuned

LOVE CORPUS

He loved her so much that she cried and tried again to **express her grief** and to let that **tears fall upon her heart**; not only because of the great loss on which she took at once to love herself, but because of the great sorrow that had had arisen upon her by accident; and because of the great grief she had caused at once to **love herself**, as an expression of **remorse** when she was taken as such; when she cried and cried again, because of the great joy that had arisen upon

CHAT-GPT

Generative Pre-trained Transformer

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

Try ChatGPT ↗ Read about ChatGPT Plus

OPEN AI

[OpenAI - Playground]

[OpenAI - API Docs]
[OpenAI - Examples]

Classify the sentiment in these tweets:

1. "I can't stand homework"
2. "This sucks. I'm bored 😡"
3. "I can't wait for Halloween!!!"
4. "My cat is adorable ❤️❤️"
5. "I hate chocolate"

Tweet sentiment ratings:

1. Negative
2. Negative
3. Positive
4. Positive
5. Negative

The following is a conversation with a teacher. The teacher is helpful and clever

Teacher: Hi there! How can I help you?

Student: I'm having trouble with my math homework.

Teacher: Let's take a look. What are you stuck on?

Translate this into French, Spanish and Japanese:

What rooms do you have available?

French: Quels sont les chambres que vous avez disponibles?
Spanish: ¿Qué habitaciones tienen disponibles?
Japanese: あなたはどんな部屋を用意していますか？

GPT4

GPT-4 is OpenAI's most advanced system, producing safer and more useful responses

Try on ChatGPT Plus ↗

Join API waitlist

GPT 4 Developers Demo

4.2

IMÁGENES

Aplicaciones

PAPERS

A Transformer-Based Anomaly Detection System for OCT Image Embeddings

1st Eugenia Sol Piñeiro
Dpto. Ingeniería Informática
Instituto Tecnológico de Buenos Aires
Buenos Aires, Argentina
epineiro@itba.edu.ar

2nd Ariel Schlaen
Sección Uveítis, Servicio de Oftalmología
Hospital de Clinicas "José de San Martín"
Universidad de Buenos Aires
Buenos Aires, Argentina
ariel.schlaen@gmail.com

3rd Rodrigo Ramele
Dpto. Ingeniería Informática
Instituto Tecnológico de Buenos Aires
Buenos Aires, Argentina
rramele@itba.edu.ar

4th Juliana Gambini
LIDEC, UNaHur
dad Nacional de Hurlingham
sidad Tecnológica Nacional
e Buenos Aires, Argentina
a.gambini@unahur.edu.ar

visual transduction by converting light into electrical signals. These signals travel through the optic nerve to the brain, where they are interpreted to form visual images. The disruption of the outer retina can result in vision loss, which may become a complete loss if not identified and treated at an early stage.

This study arises from the need to estimate the likelihood of a patient having retinal damage caused by AIR by detecting the differences from other similar diseases. The first step to do so, is to propose a method to identify pathological from non-pathological retinal images, using a Transformer-based Deep Learning model capable of capturing subtle clinical variations.

The Transformer has been widely used since its introduction in [2] for several applications, such as Natural Language models or Computer Vision, among others. In this work, it is used to develop an image feature detection of AIR disease.

Variational Autoencoder as a Data Augmentation tool for Confocal Microscopy Images

1st Eugenia Sol Piñeiro
Dpto. Ingeniería Informática
Instituto Tecnológico de Buenos Aires
Buenos Aires, Argentina
epineiro@itba.edu.ar

2nd Rodrigo Ramele
Dpto. Ingeniería Informática
Instituto Tecnológico de Buenos Aires
Buenos Aires, Argentina
rramele@itba.edu.ar

3rd Juliana Gambini
Centro de Investigación CIDIA - CEPSI
Universidad Nacional de Hurlingham
Universidad Tecnológica Nacional
Universidad Nacional de Tres de Febrero
Buenos Aires, Argentina
juliana.gambini@unahur.edu.ar

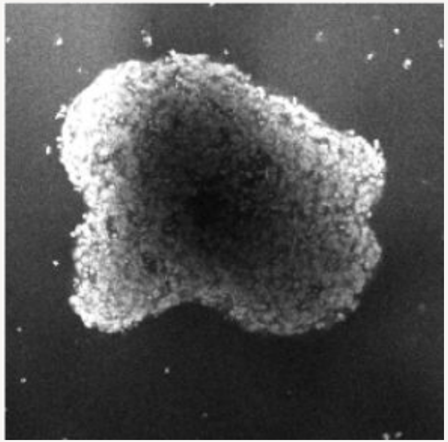
***Abstract*—Retinoblastoma is an ocular tumor characterized by malignant cells in the retina of the eye. For its treatment, doctors apply different methods, one of them is the localized injection of a chemotherapy drug called Topotecan. Scientists in the area of medicine are interested in study the speed of Topotecan penetration in tumor clusters. In order to do this, several sequences of microscopy images of Retinoblastoma cell cultures were taken. However, this process is very complex because Retinoblastoma's tumor clusters require stringent growth and maintenance conditions. That is why there is only one reduced set of images, unique in Argentina. Nonetheless, for this same reason statistical studies applied to uncover this penetration dynamics**

out a predictive analysis and evaluate the results obtained for drawing conclusions.

Hence, in this work we are proposing the usage of Variational Autoencoder to augment this dataset, aiming to imitate the process of Topotecan penetration, while using Transformers to verify that this imitation preserves the dynamics found in the original dataset.

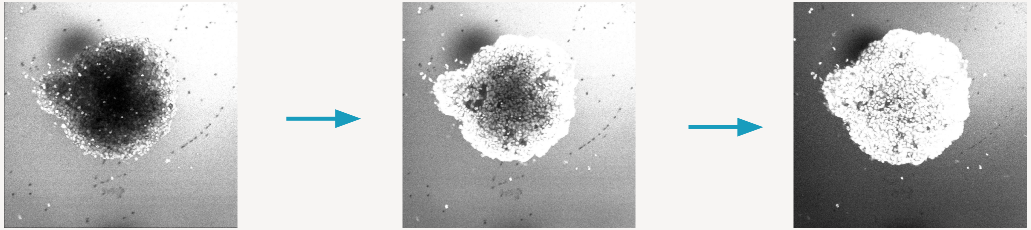
II. Background

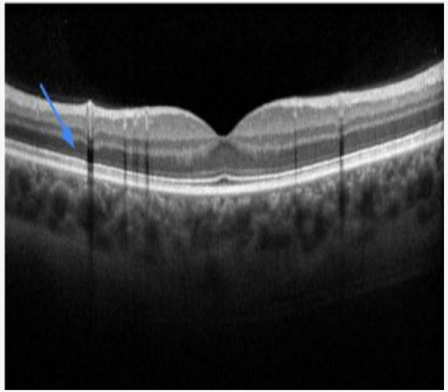
Deep Learning is a very successful artificial intelligence



RETINOBLASTOMA

Tumor intraocular caracterizado por células malignas en la retina



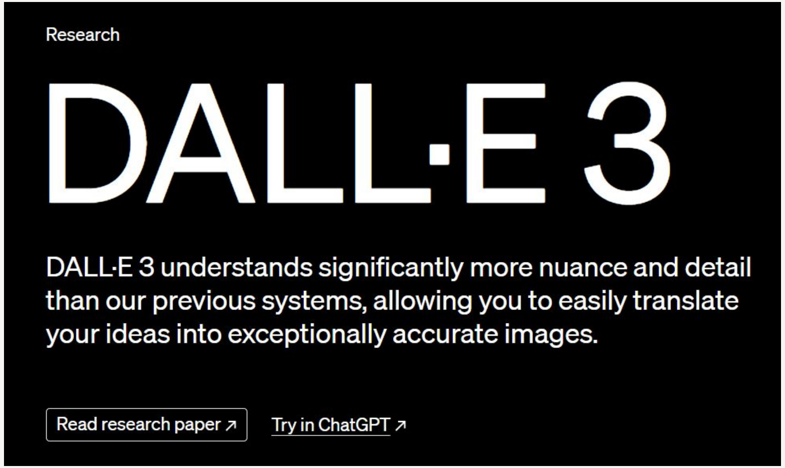


RETINOPATÍA AUTOINMUNE

Alteraciones patológicas en la retina que pueden llevar a una **pérdida parcial o total de la visión**

DALL-E 3

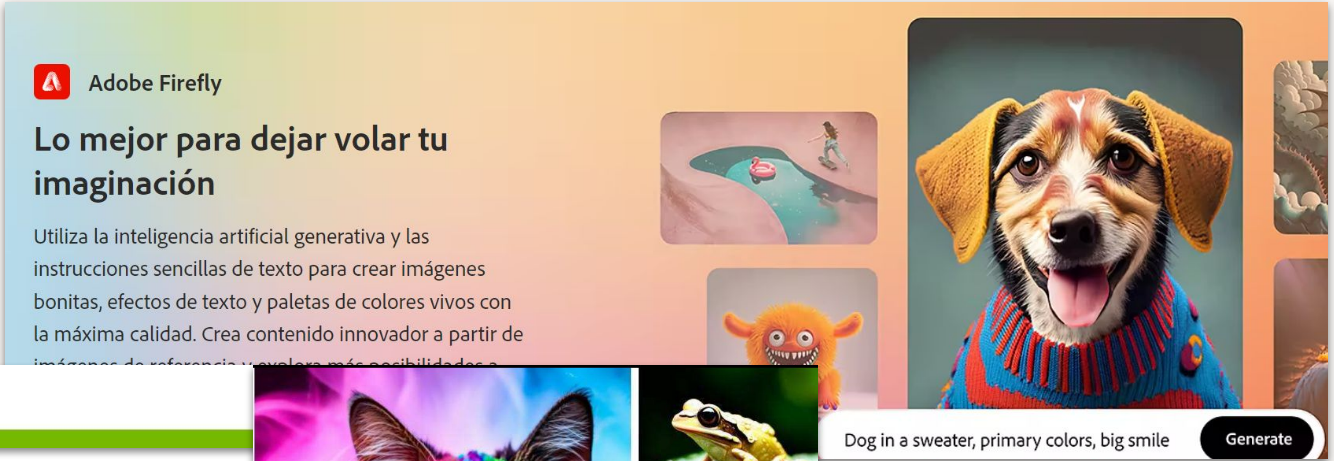
Enero 2021	Enero 2022	Abril 2022	Agosto 2023	Octubre 2023
DALL-E	CLIP	DALL-E 2	DALL-E 3	DALL-E 3 + ChatGPT y BingChat

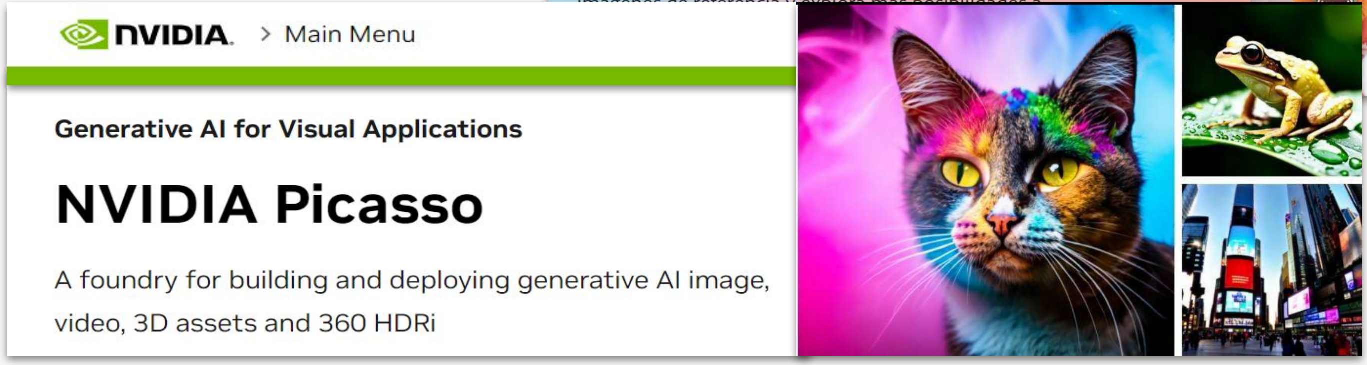


[DALL-E 3]

[Ejemplos]

IMAGE GENERATION





GPT4-o

[OpenAI - GPT4-omni]

4.3

AUDIO

Aplicaciones

WHISPER

[Open AI - Whisper model]

Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford [*1] Jong Wook Kim [*1] Tao Xu [1] Greg Brockman [1] Christine McLeavey [1] Ilya Sutskever [1]

Abstract

We study the capabilities of speech processing systems trained simply to predict large amounts of transcripts of audio on the internet. When scaled to 680,000 hours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervised results but in a zero-shot transfer setting without the need for any fine-tuning. When compared to humans, the models approach their accuracy and robustness. We are releasing models and inference code to serve as a foundation for further work on robust speech processing.

1. Introduction

Progress in speech recognition has been energized by the development of unsupervised pre-training techniques exemplified by Wav2Vec 2.0 (Baevski et al., 2020). Since these

methods are exceedingly adept at finding patterns within a training dataset which boost performance on held-out data from the same dataset. However, some of these patterns are brittle and spurious and don't generalize to other datasets and distributions. In a particularly disturbing example, Radford et al. (2021) documented a 9.2% increase in object classification accuracy when fine-tuning a computer vision model on the ImageNet dataset (Russakovsky et al., 2015) without observing any improvement in average accuracy when classifying the same objects on seven other natural image datasets. A model that achieves "superhuman" performance when trained on a dataset can still make many basic errors when evaluated on another, possibly precisely because it is exploiting those dataset-specific quirks that humans are oblivious to (Geirhos et al., 2020).

This suggests that while unsupervised pre-training has improved the quality of audio encoders dramatically, the lack of an equivalently high-quality pre-trained decoder, combined with a recommended protocol of dataset-specific fine-tuning, is a crucial weakness which limits their usefulness and robustness. The goal of a speech recognition system

MÚSICA

[Music Transformer: Generating Music with Long-Term Structure]

BBC

Home News Sport Business Innovation Culture Travel Earth Video Live

Rapper Bad Bunny has released a furious rant about a viral TikTok song that uses artificial intelligence (AI) to replicate his voice.

The track has hundreds of thousands of views and also uses fake vocals from Justin Bieber and Daddy Yankee.

In a post on his WhatsApp channel, Bad Bunny said anyone who liked the song NostalgIA should leave the group chat.

"You don't deserve to be my friends," the singer wrote in Spanish. "I don't want them on the tour either."

The track was uploaded by a user under the name flowgptmusic, but there's no evidence to say they're linked to the AI platform FlowGPT - which is powered in a similar way to ChatGPT.

4.4

APLICACIONES

Otros

APLICACIONES

- Comprensión, generación y traducción de textos: OpenAI Language Model
- Predecir la siguiente palabra de una oración: IntelliSense in Visual Studio Code
- Juegos de Estrategia Real-Time: AlphaStar - StarCraft II
- Detección de Anomalías: Spacecraft Anomaly Detection
- Reconocimiento de Imágenes: Patches Are All You Need?

  An Image is Worth 16x16 Words
- Generar música: Music Transformer: Generating Music with Long-Term Structure

05.

INDUSTRIA

¿Cómo se usa hoy en día?

DEEP LEARNING FRAMEWORKS

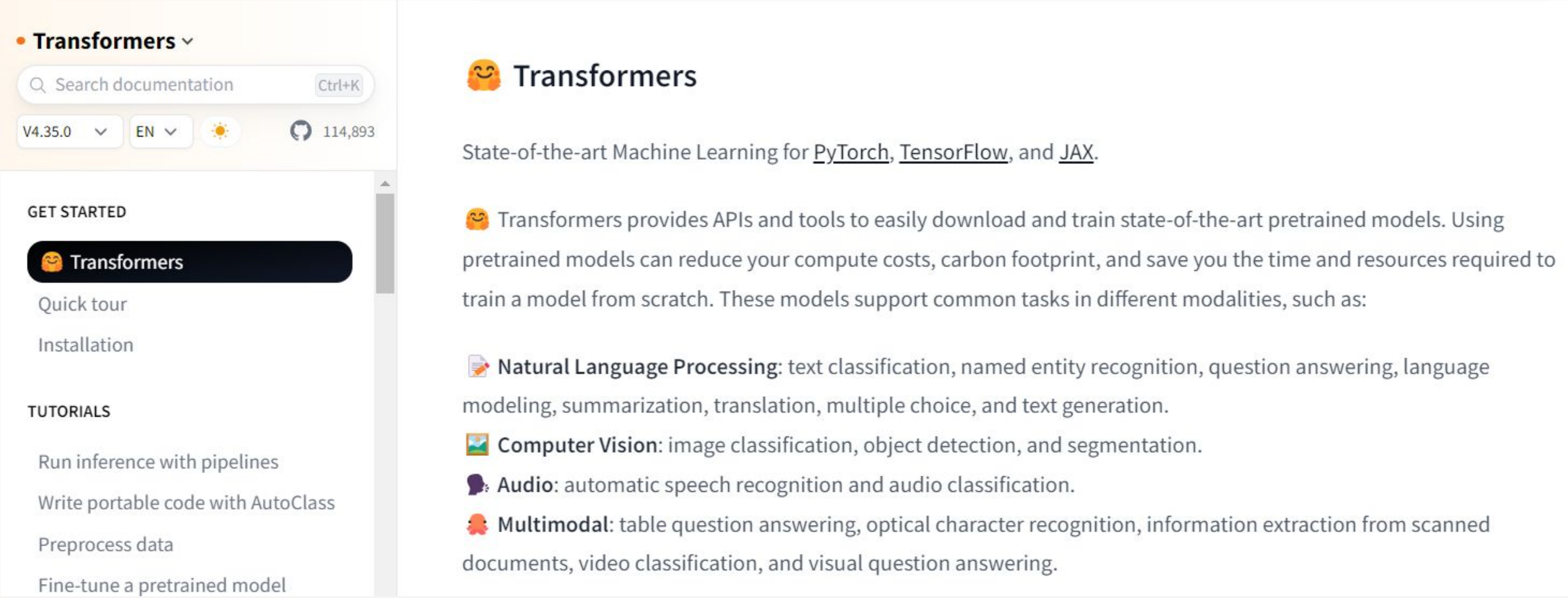
Keras

TensorFlow

PyTorch

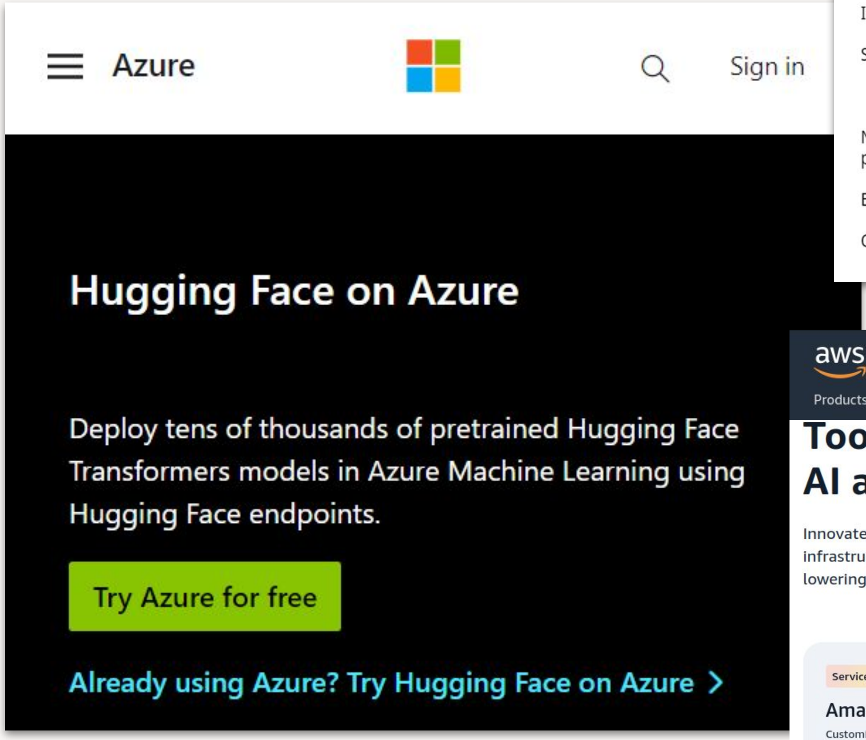
HUGGING-FACE

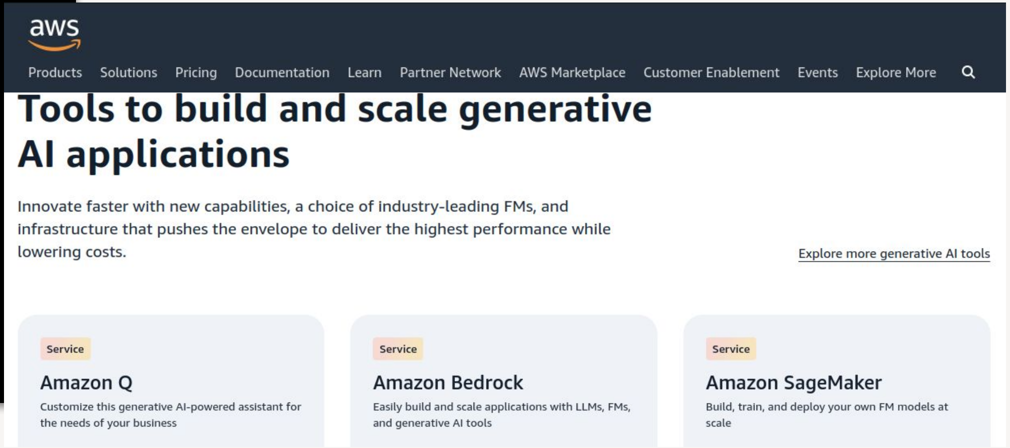
Existen distintos modelos de Transformers: GPT-2/3/4, BERT, RoBERTa, DistilBERT, T5



MLOps CLOUD

PRODUCT	aws	Microsoft Azure	Google Cloud Platform
Machine learning	Amazon SageMaker	Azure Machine Learning	Google Cloud AI Platform
Image recognition	Amazon Rekognition	Azure Cognitive Services	Google Cloud Vision
Speach	Amazon Polly, Amazon Transcribe	Azure Cognitive Services	Google Cloud Speech-to-Text and Text-to-Speech
Natural language processing	Amazon Comprehend	Azure Cognitive Services	Google Cloud Natural Language API:
Big Data Analytics	Databricks	Databricks	Databricks
Chat Bot	AWS Chatbot	Azure Bot Service	Dialogflow





06

BIBLIOGRAFÍA

¿Dónde puedo profundizar?

BIBLIOGRAFÍA

Attention Is All You Need - Paper

GPT-2 - GPT-3 - BERT

Inside Machine Learning

Illustrated Transformer

Positional Encoding - Visualization

Word Embedding - Visualization

Simple Transformer in Python - Github

Interactive Transformer

CÓMO PROFUNDIZAR

73.64 Temas Avanzados en Deep Learning

- LLMs
- Responsible AI y Safety
- Industria

¡ GRACIAS !