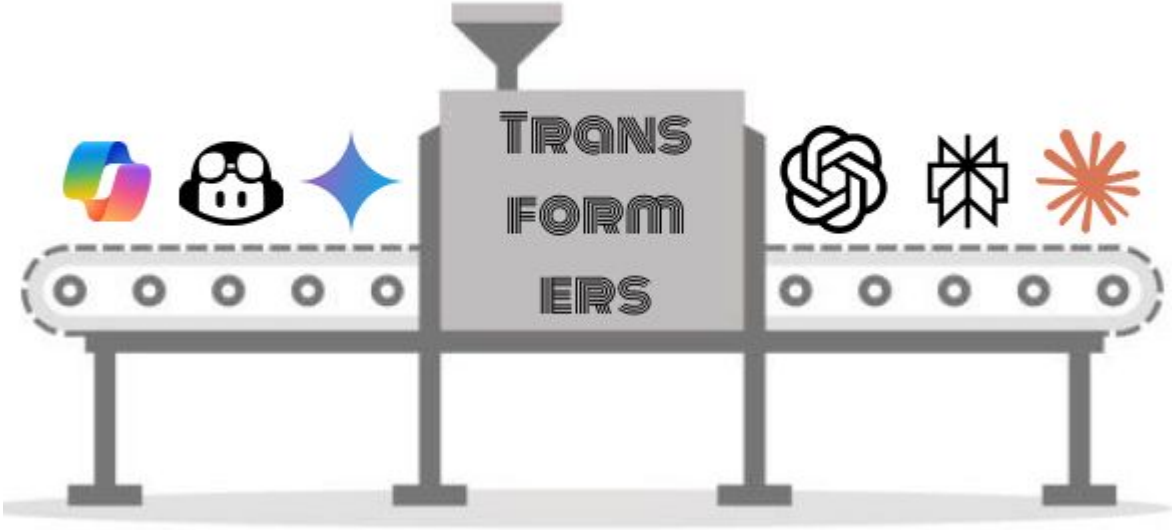
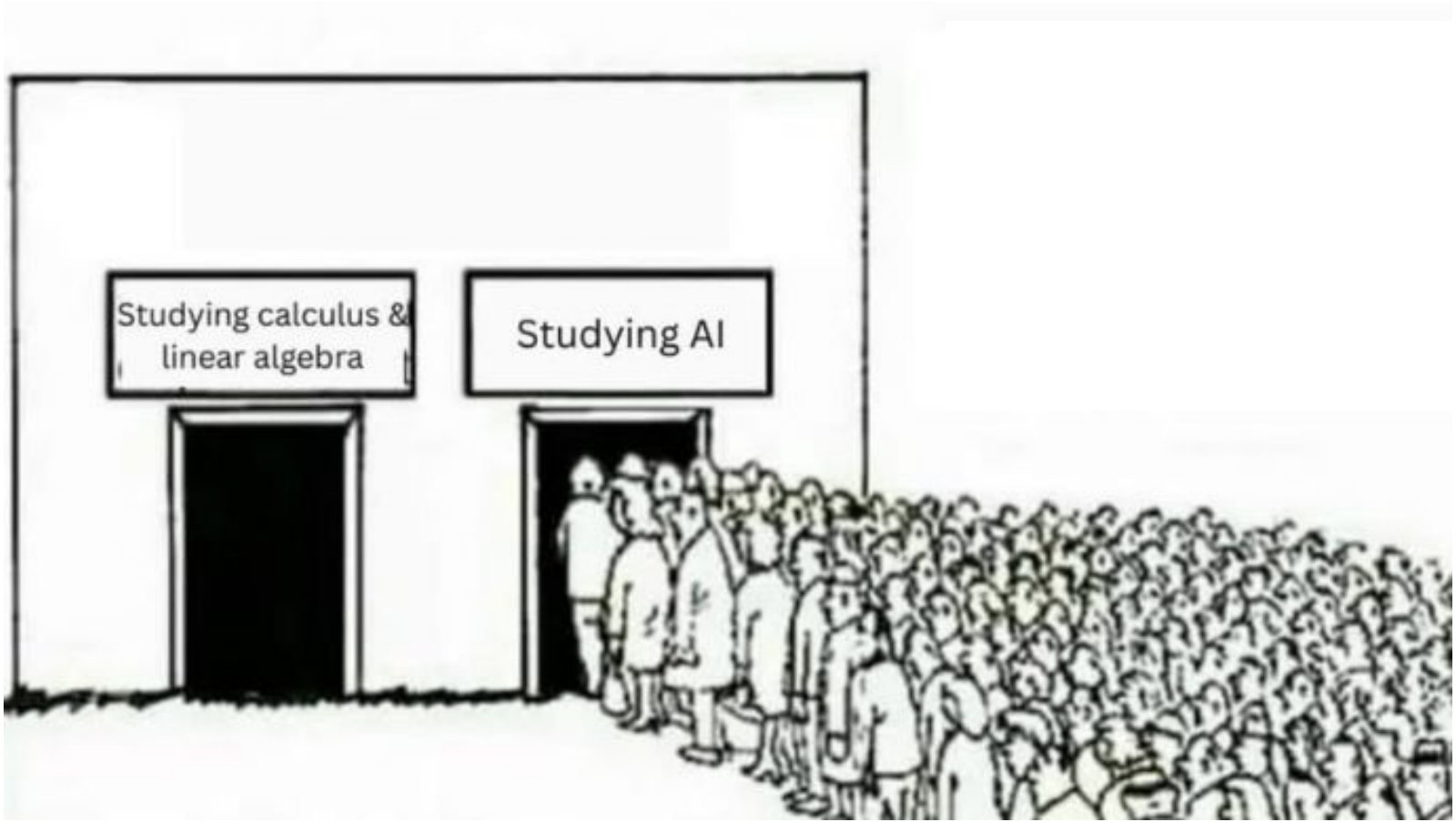
Transformers: El corazón de la IA Moderna

Transformers

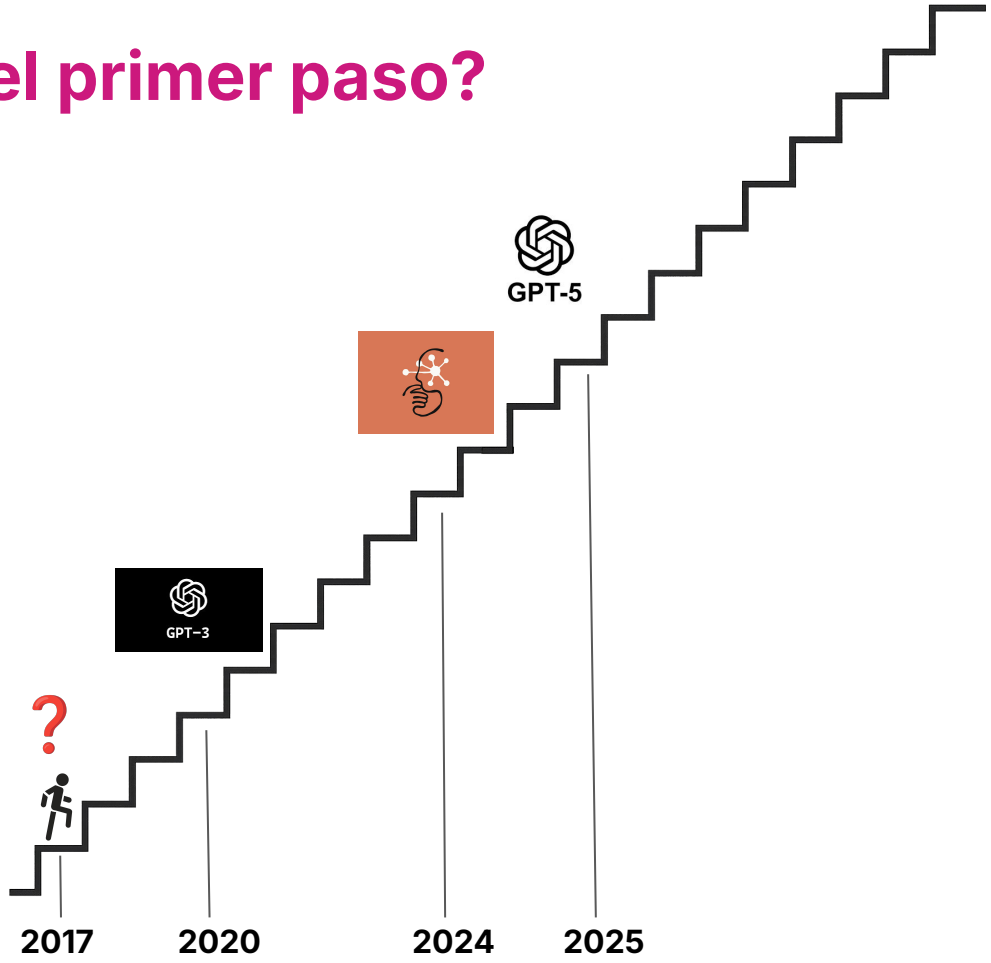
PyTorch



Lauren Gallego Ropero
4º Ciencia e Ingeniería de Datos


Studying calculus & linear algebra
Studying AI

¿Quién dio el primer paso?



“Attention is All You Need”

- Paper original de la arquitectura Transformers
- Presentada en NeurIPS 2017
- Citada en otros 250,000 papers/libros/docs...
- Pensado como modelo de traducción automática

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 training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

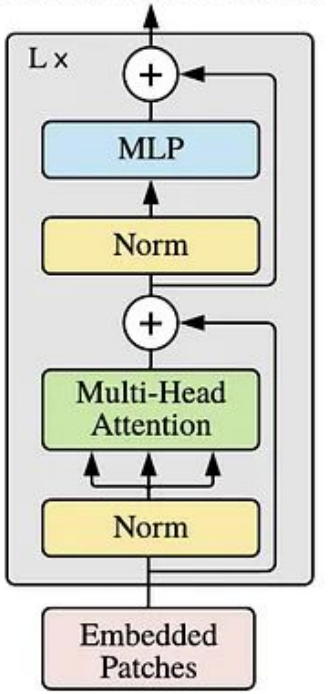
*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

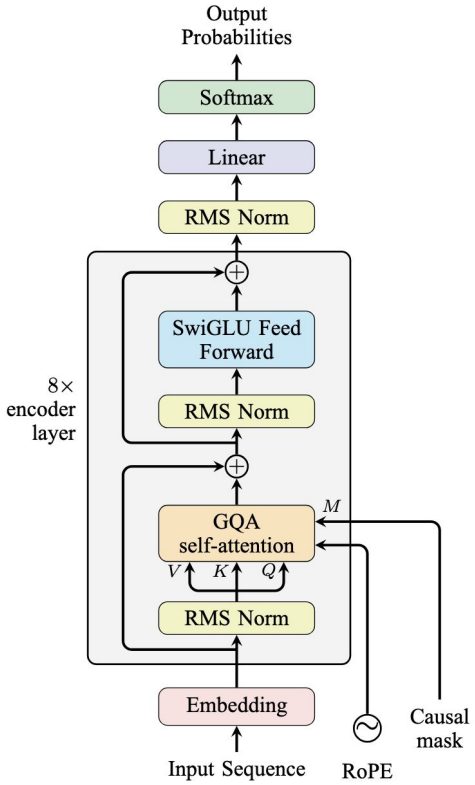
‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

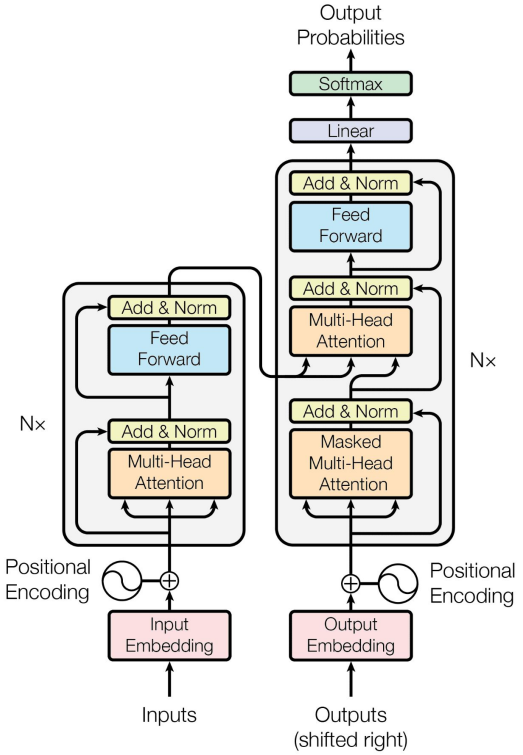
Encoder-only



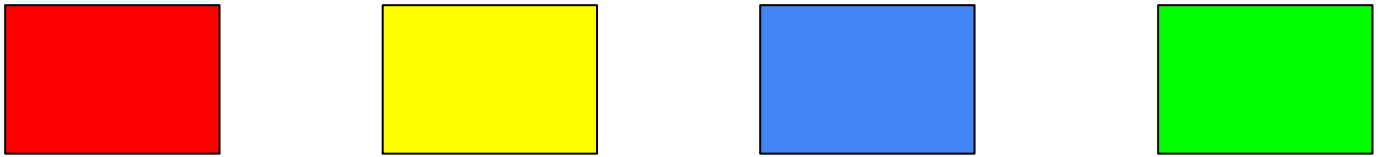
Decoder-only



Encoder-Decoder



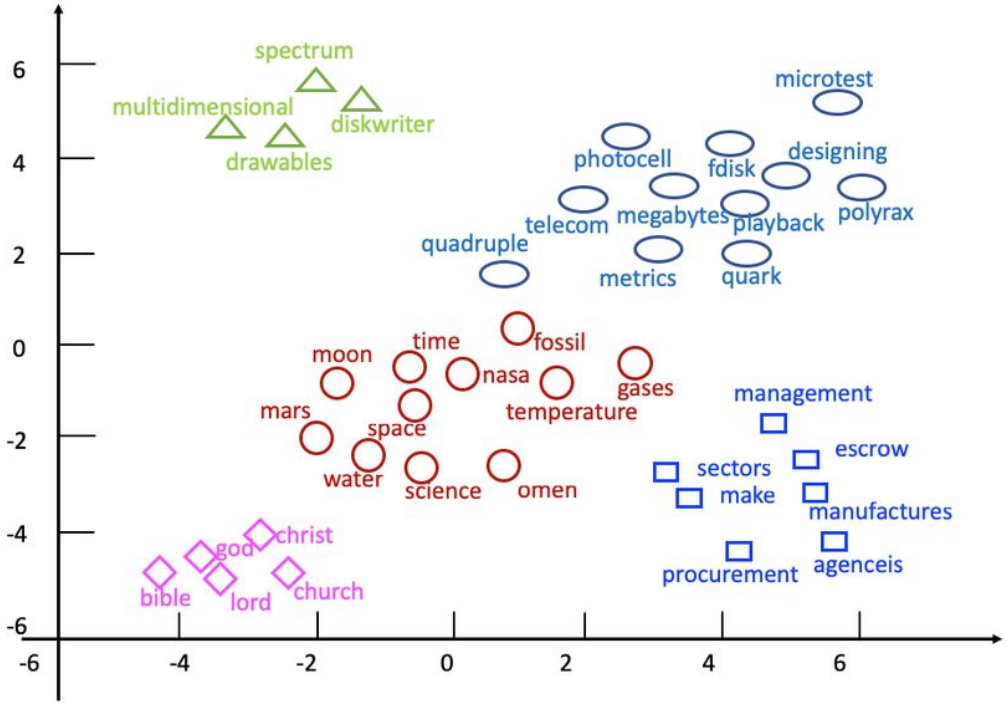
Input



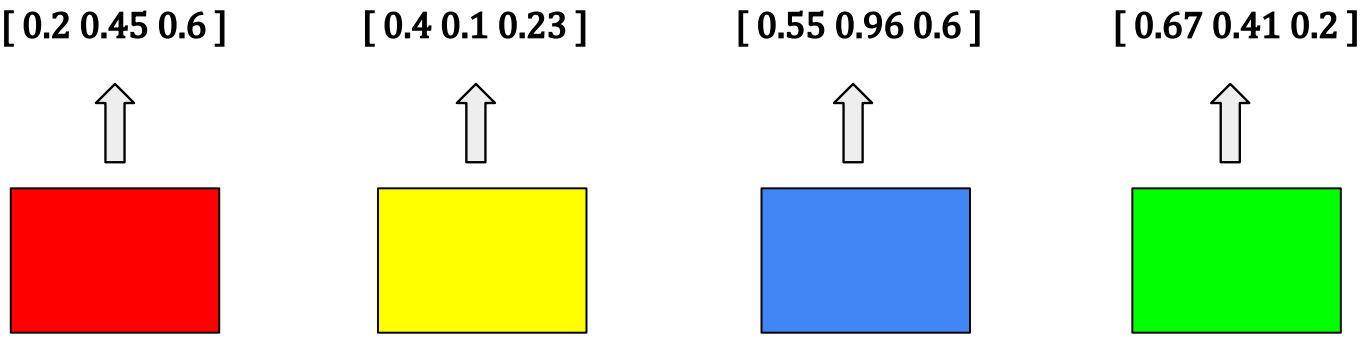
let’s go shopping <EOS>

Word Embeddings

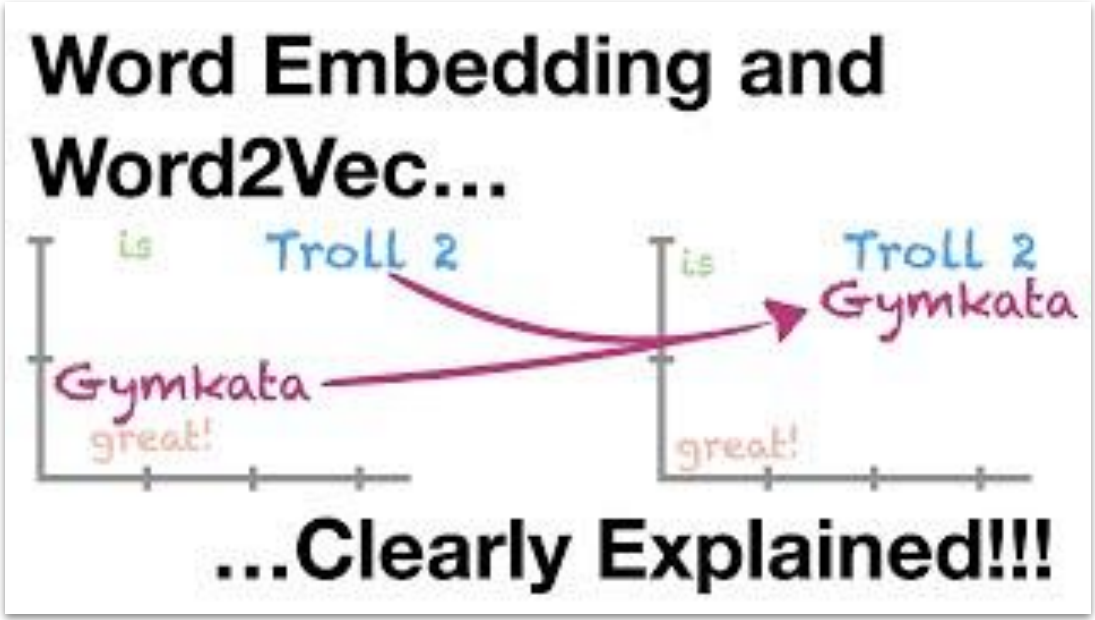
- Tokens → Vectores
- Alta dimensión
- Capturan significados
  - Rey - Hombre + Mujer = Reina
- Capturan contexto
  - París - Francia + Italia = Roma
- Token → Representación única



Word Embeddings



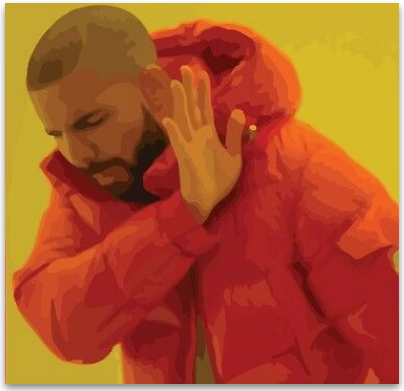
¿Cómo se obtienen los embeddings?



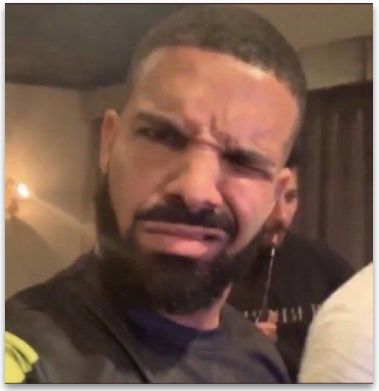
Positional Encoding

Drake dejó a Rihanna

Rihanna dejó a Drake

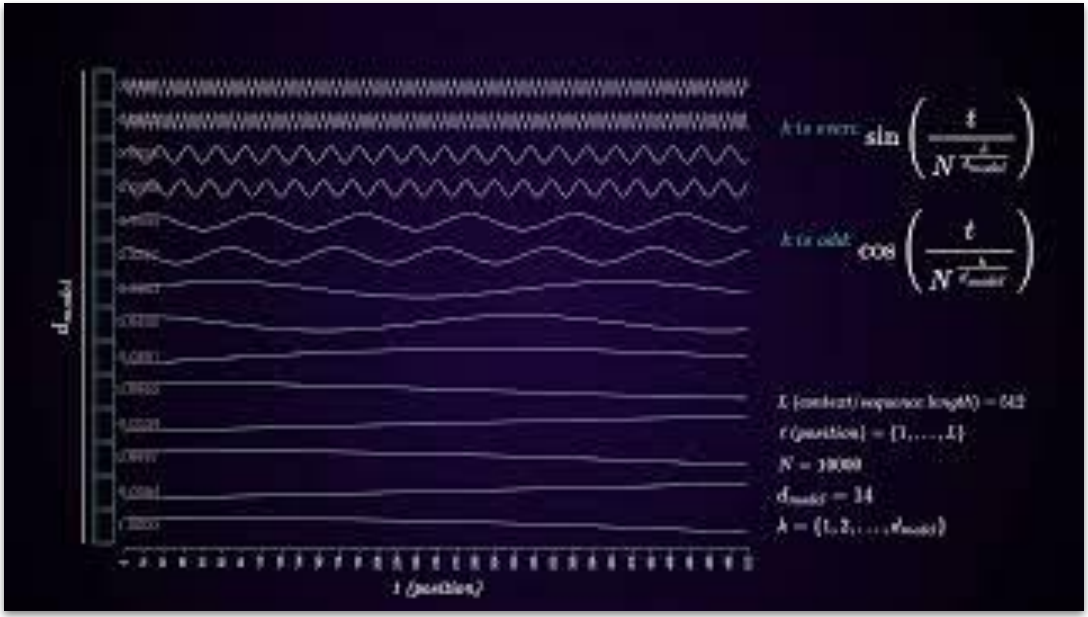


Drake Rihanna a dejó

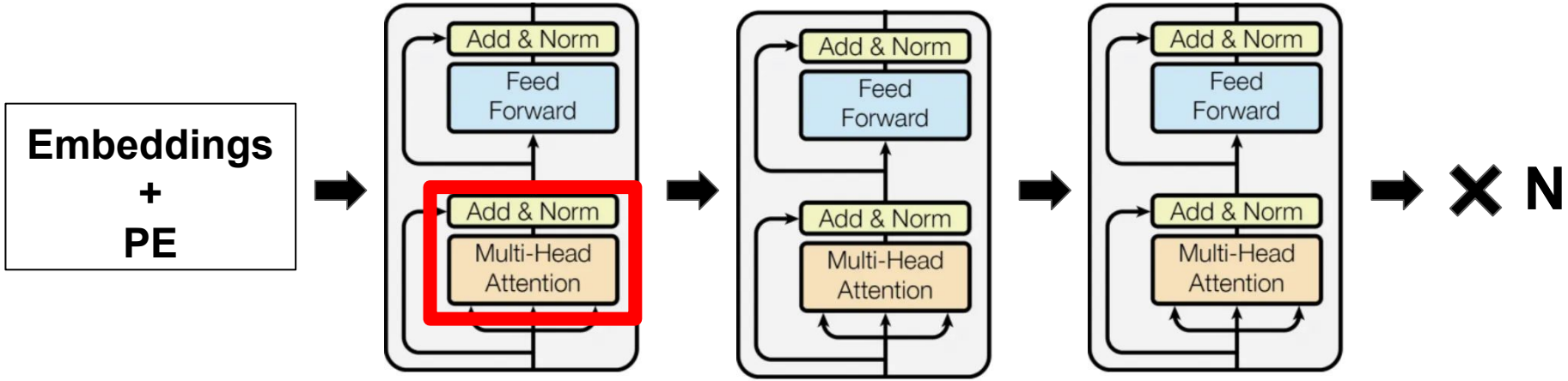


Positional Encoding

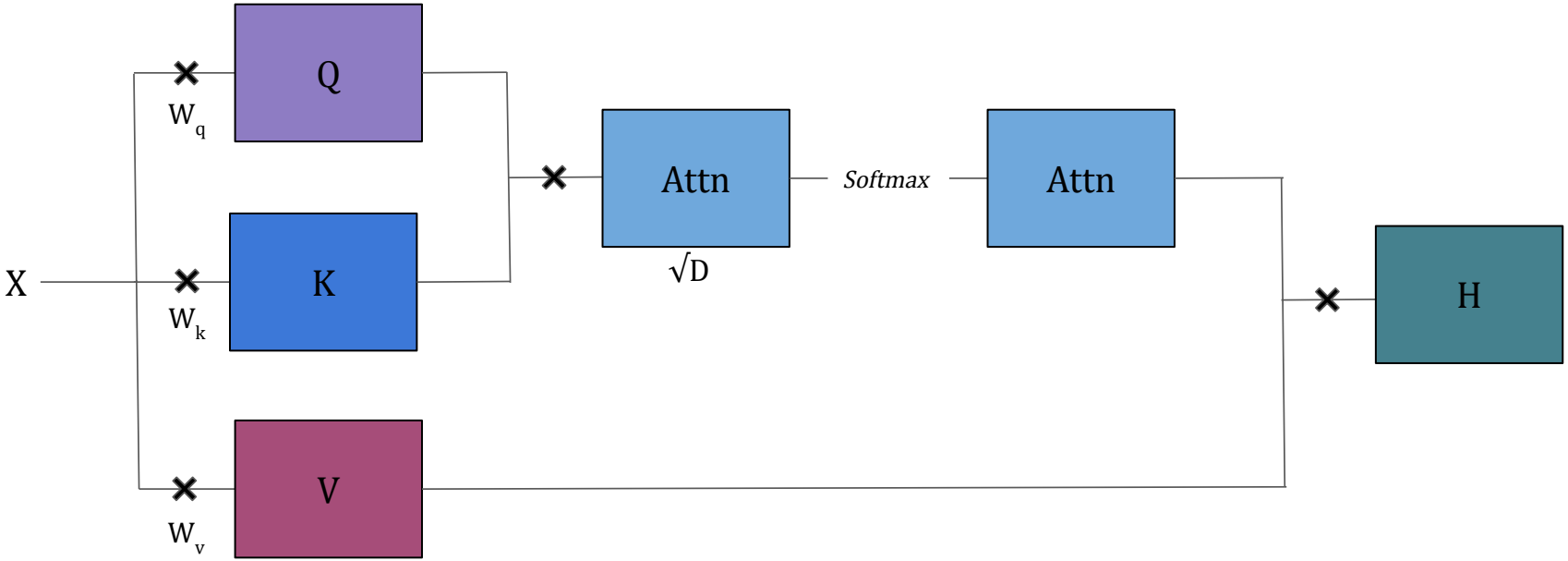
¿De dónde salen estas sumas?



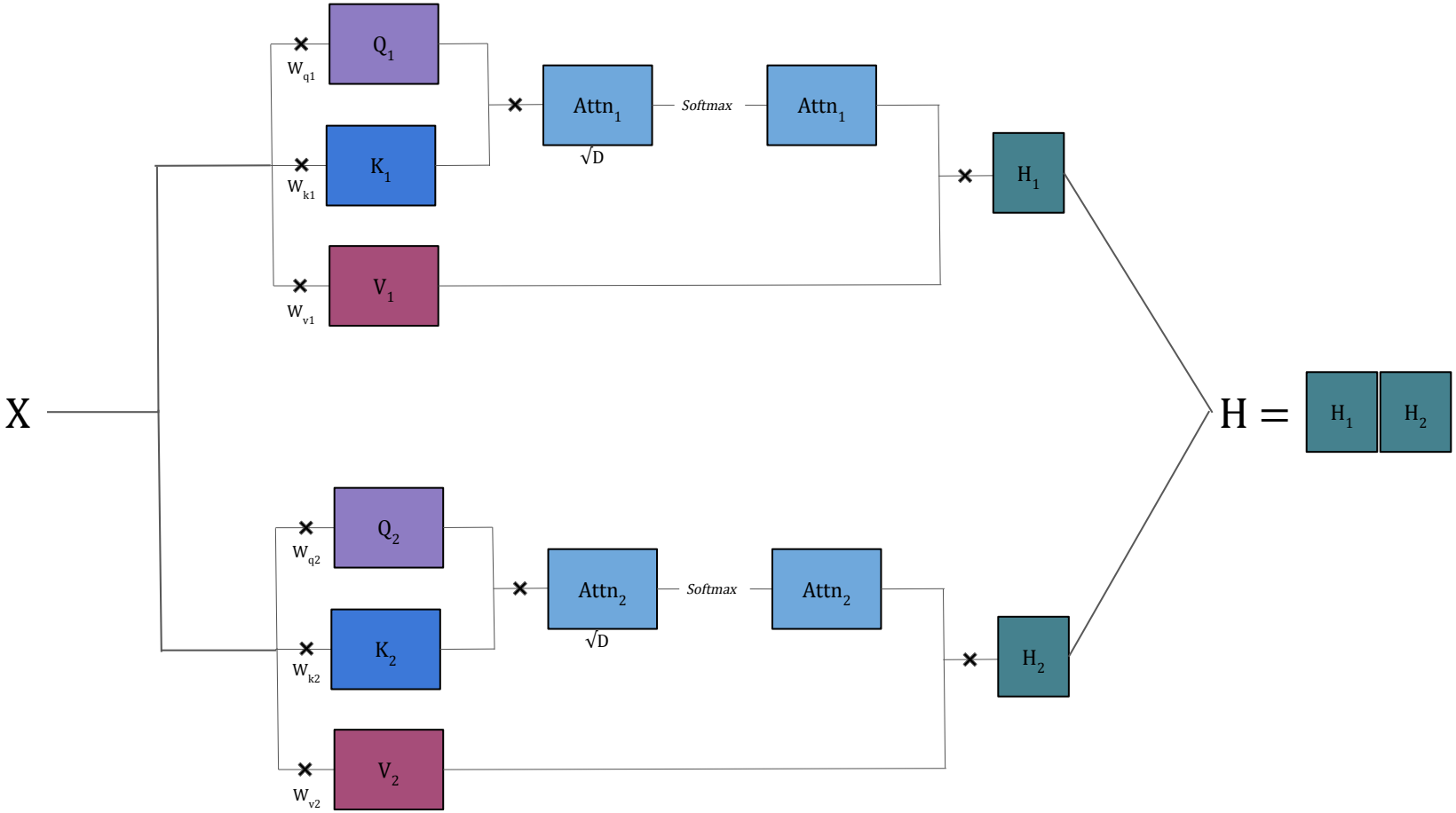
Transformer layer



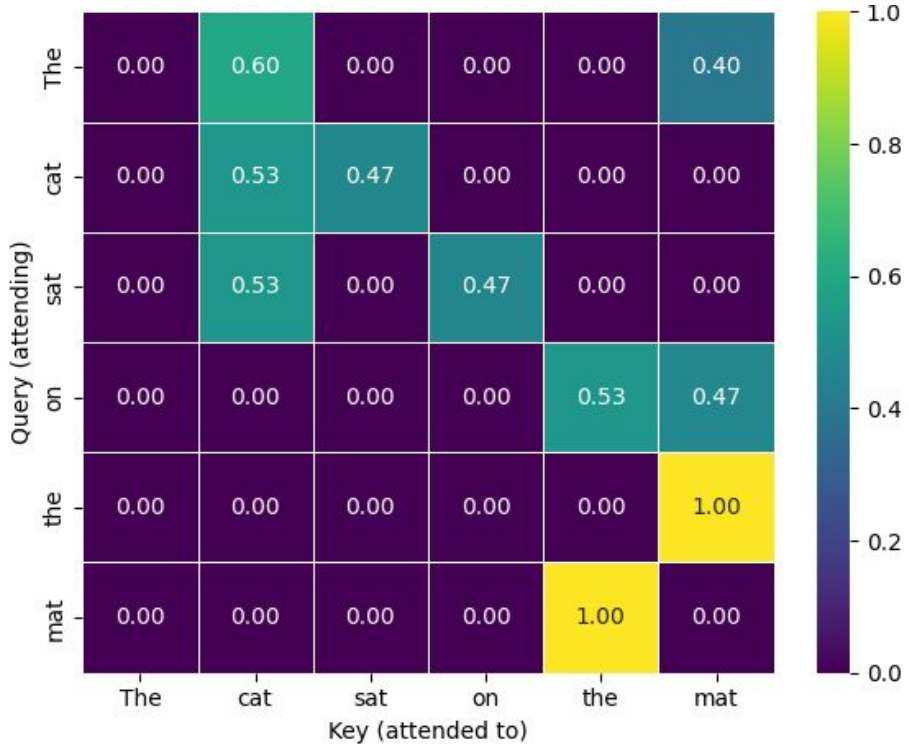
Softmax Attention



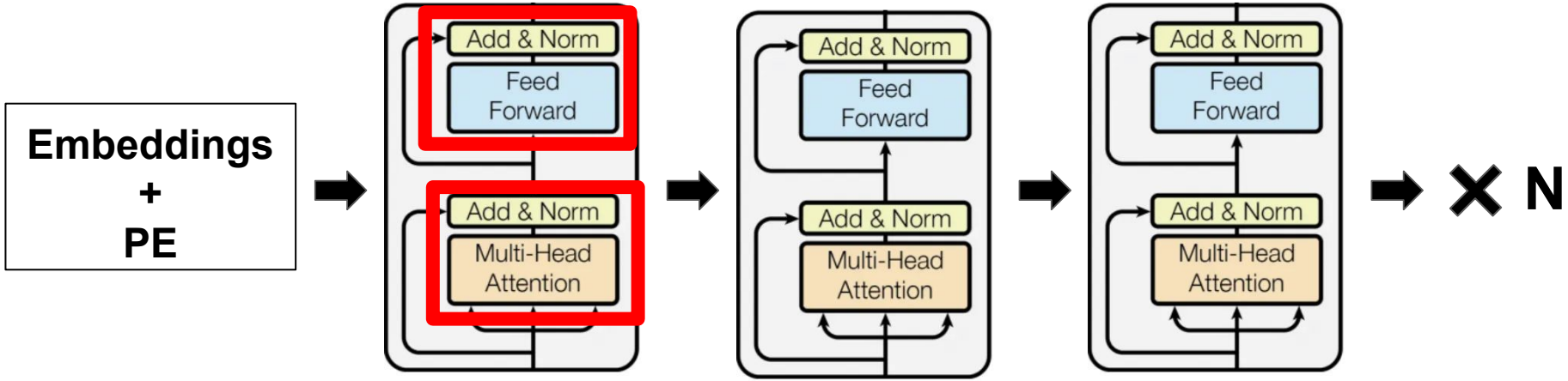
Multi-Head Attention



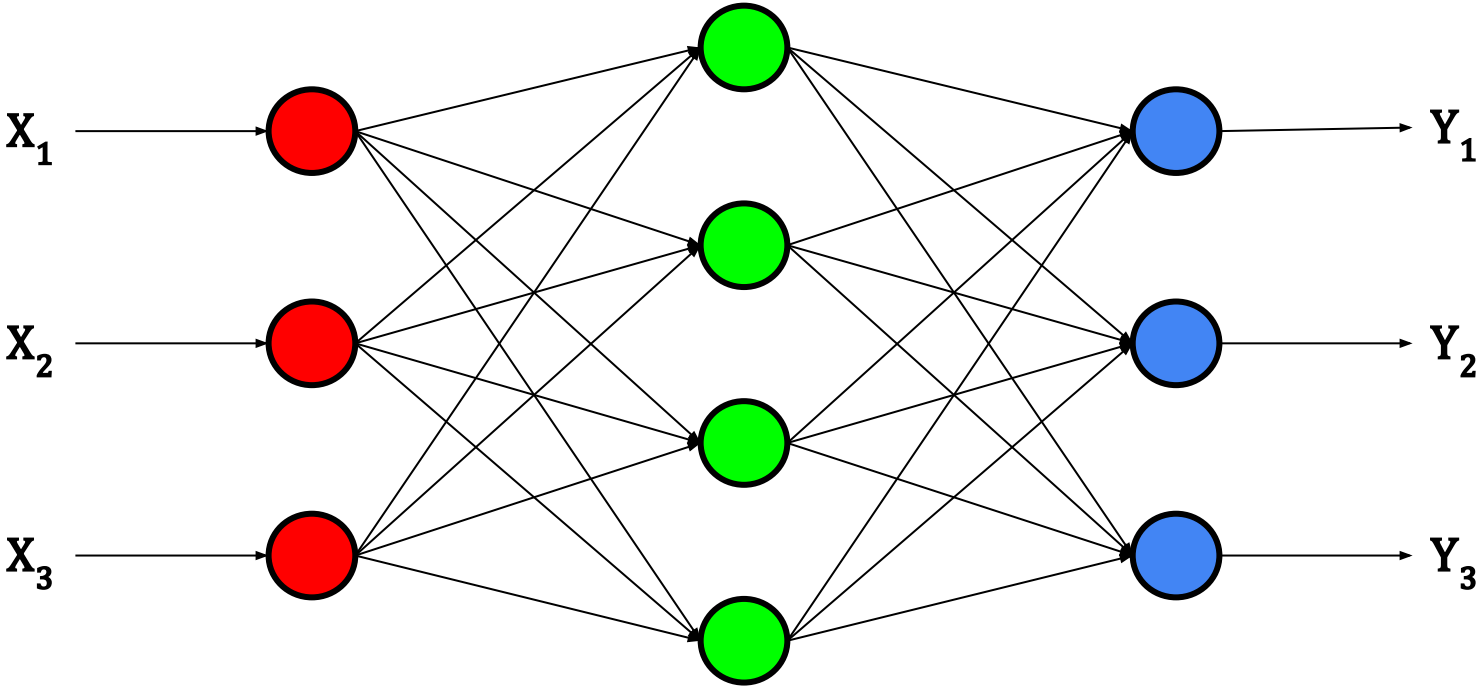
Ejemplo: “The cat sat on the mat”



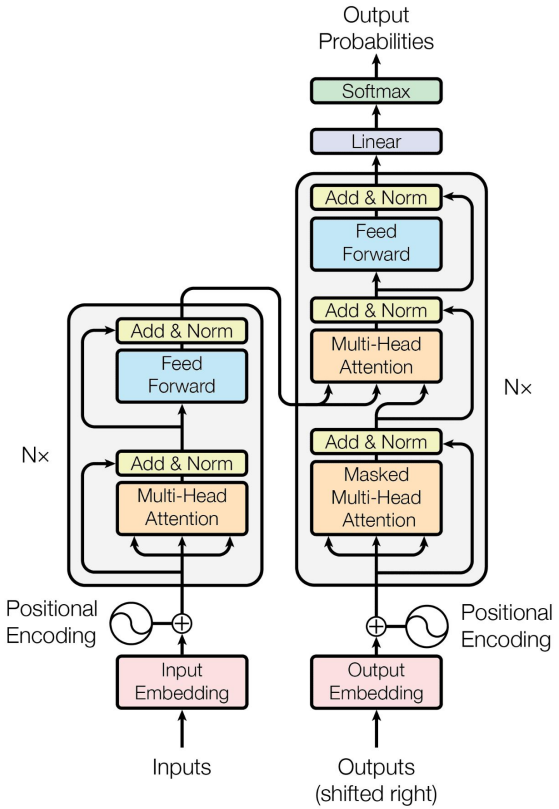
Transformer layer



Feed forward / MLP



Recap


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

¡A PRACTICAR!