An Introduction to Transformers

Jack Elliot Collier Ryder

2025-06-18

Plan of Action

- Some background
- How they work
- How to make them work

Some background to Transformers

- Neural Network

- “Transformers are models that learn to transform input data into more useful representations, helping them understand patterns, relationships, and meaning — whether in language, images, or proteins.” - ChatGPT

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.

How they work

- Encoder & Decoder
- We'll focus on just the decoder
- And we'll consider language

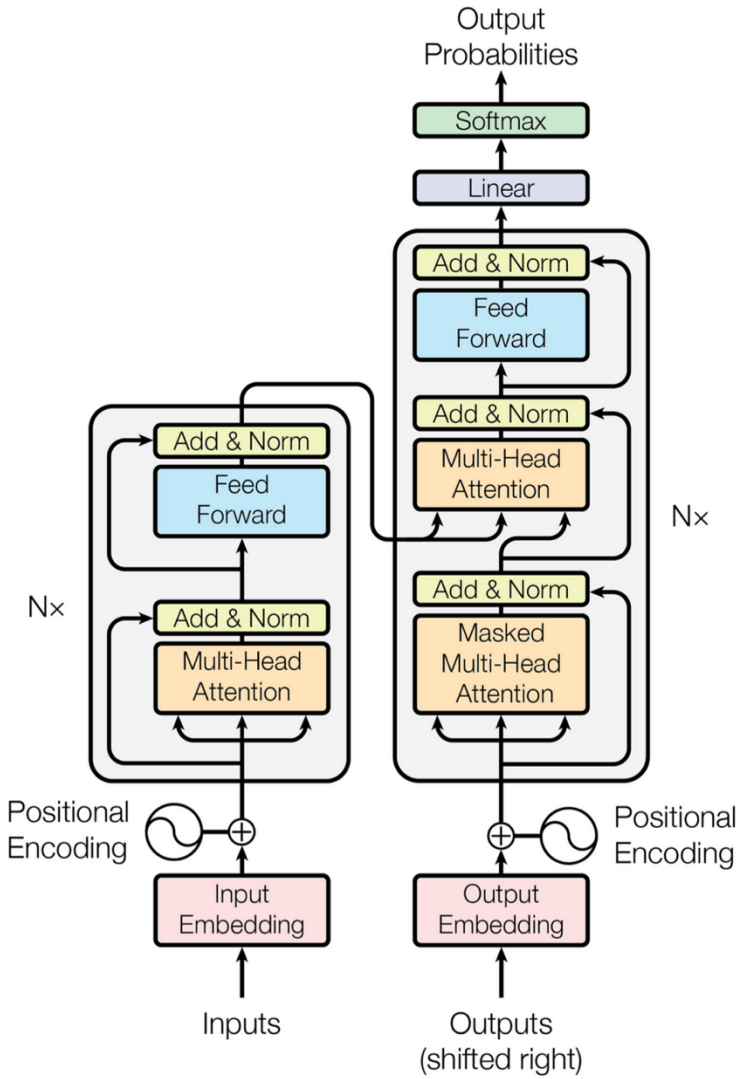


Figure 1: The Transformer - model architecture.

Input embedding

- Goal: Create an “effective” vector representation of data
- One-hot encoding
- Embedding matrix

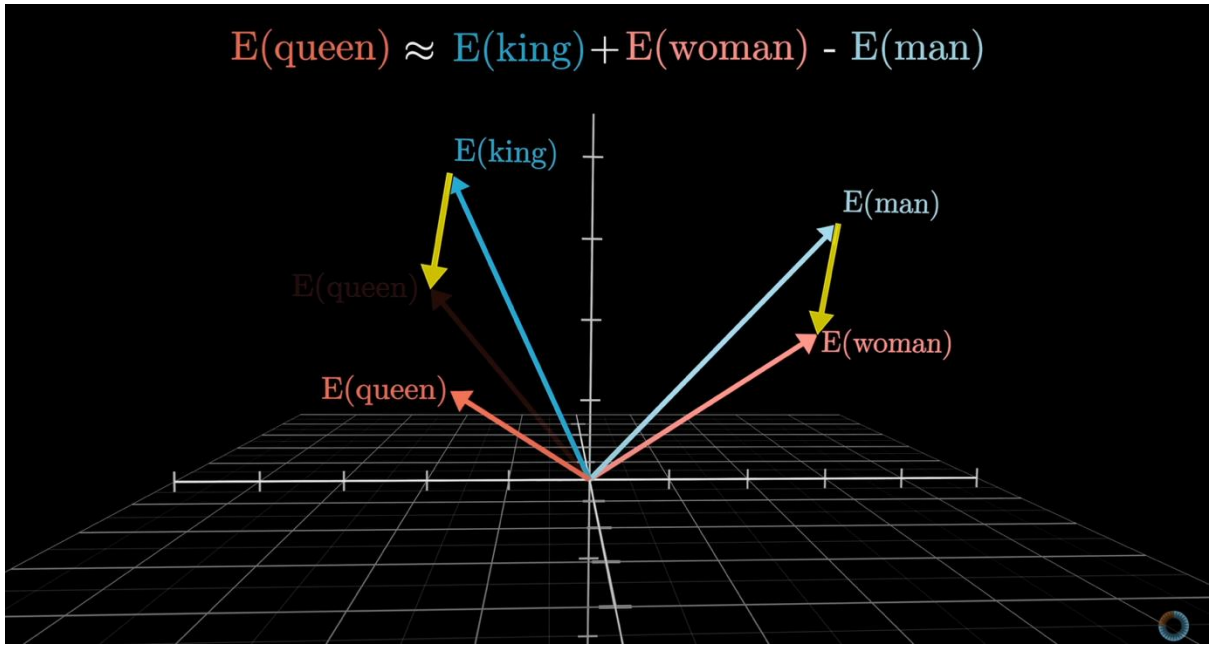
Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

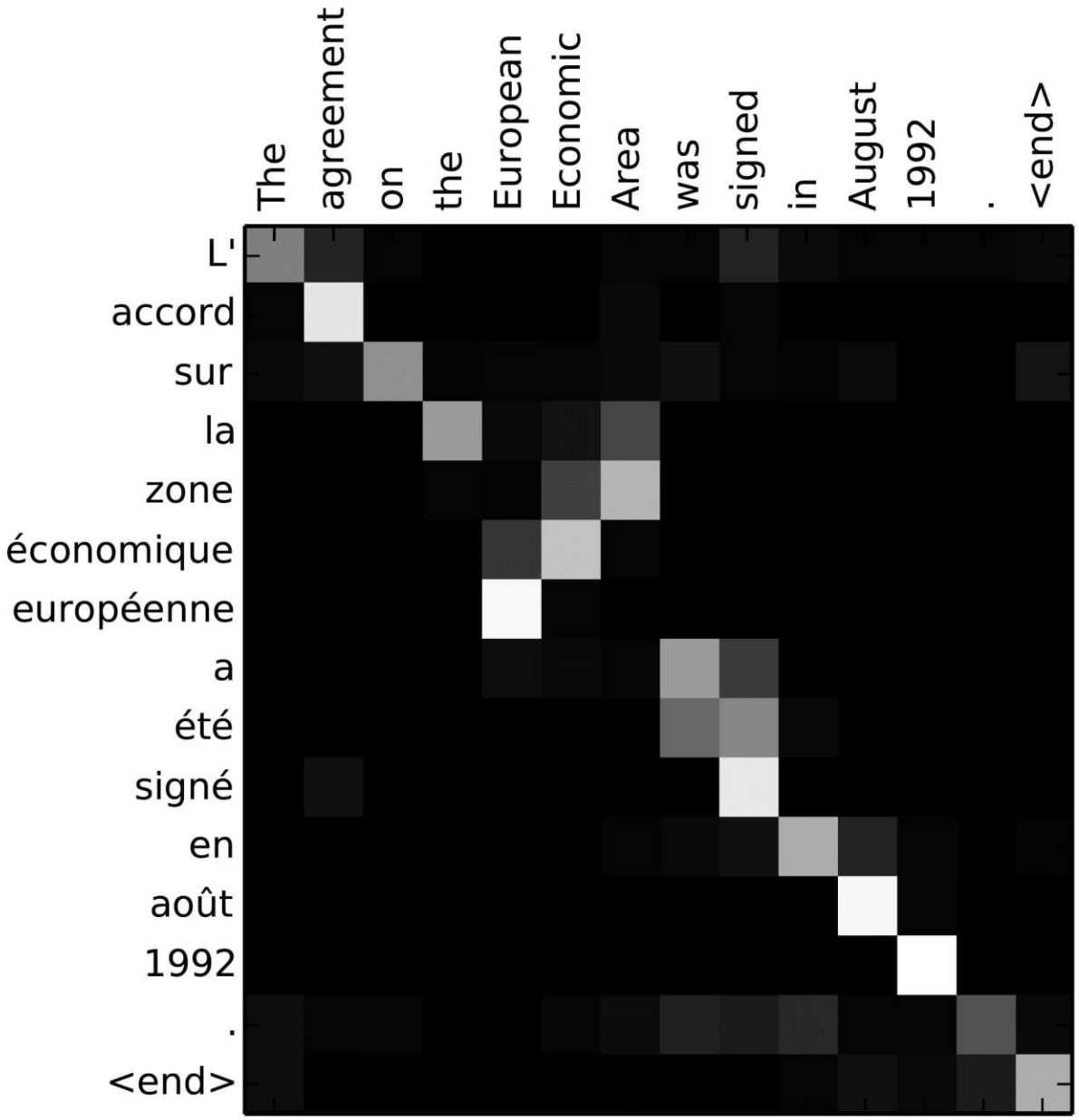
Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50



Attention

- In NLP: “Lets the words look at each other”

- *I hit the ball with a bat*
  vs.
  A bat flew by



Processing

- Feedforward networks
- Output projection
- Training…

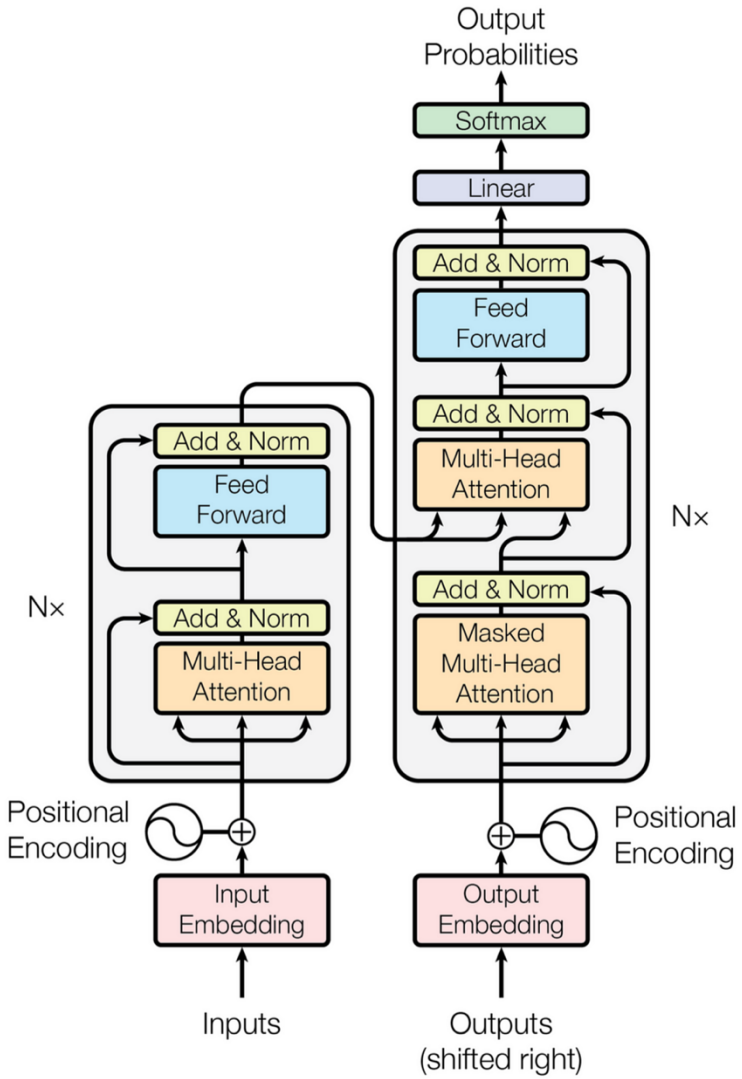


Figure 1: The Transformer - model architecture.

I skipped some things

- Positional encodings
- Dropout
- LayerNorm
- Gradient checkpointing
- Residual connections
- Etc.