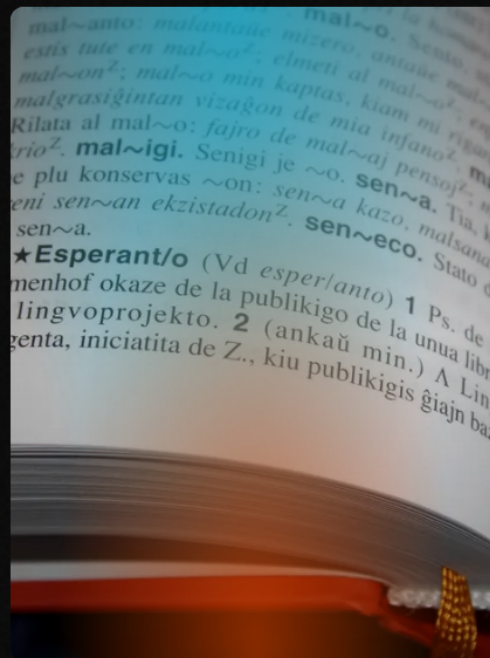
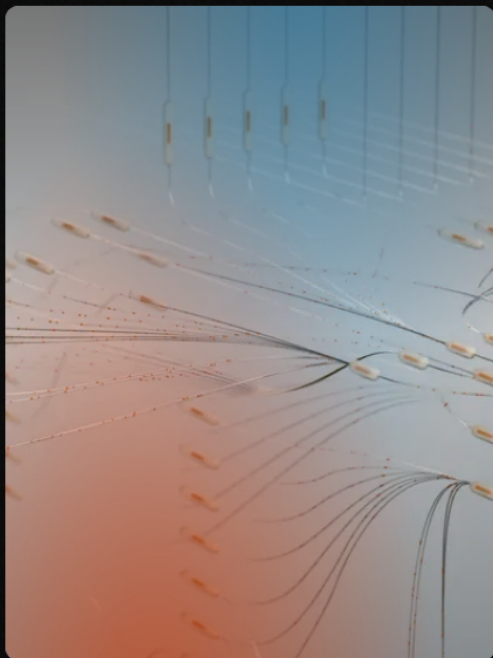


Paper Club

ARXIV 1706.03762



Attention Is All You Need, in six slides





THE CLAIM

Before this paper, the best translation models were recurrent or convolutional encoder-decoders, with attention bolted on to connect the two halves. Recurrent models process tokens one after another, which blocks parallel training.

The Transformer drops recurrence and convolutions entirely and uses attention alone. The authors report better quality, more parallelism and much less training time.

28.4

BLEU, WMT 2014 ENGLISH-GERMAN

41.8

BLEU, WMT 2014 ENGLISH-FRENCH

3.5 days

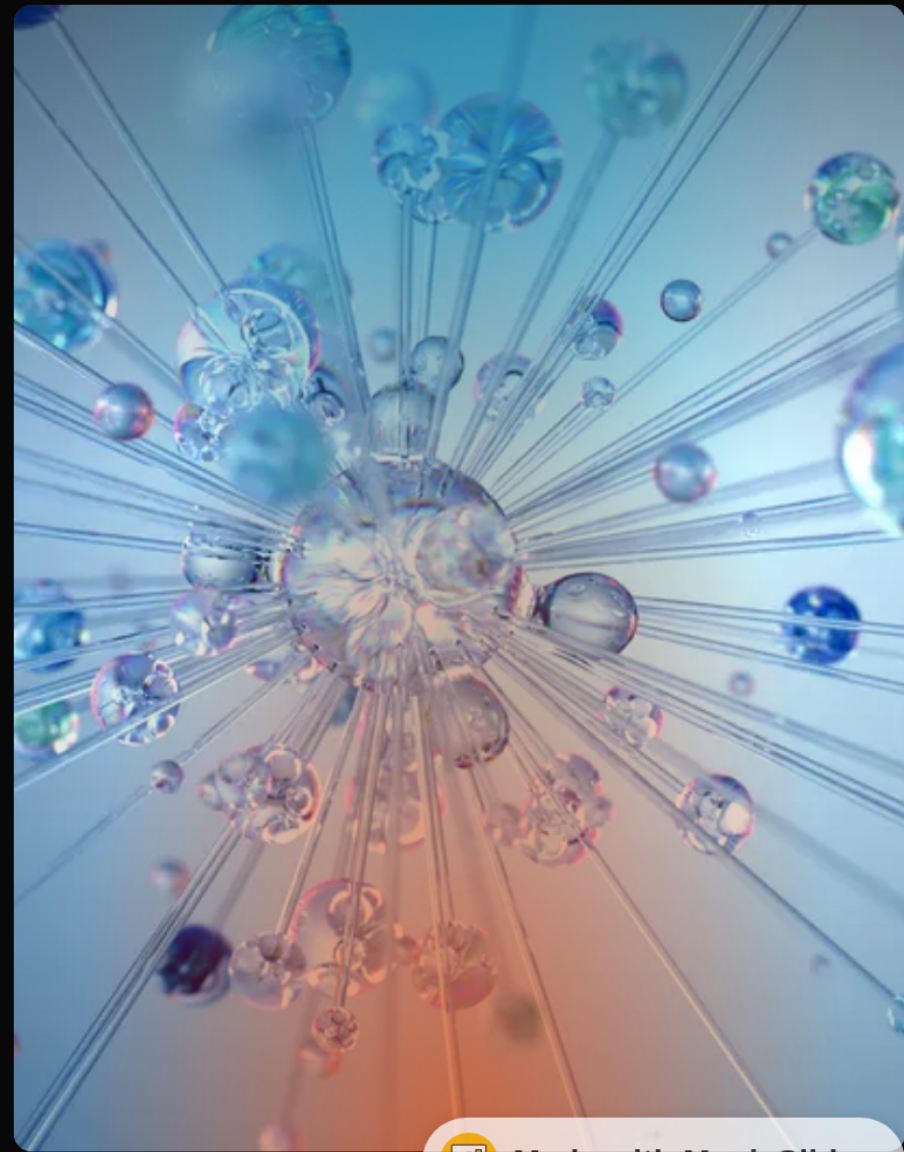
128 MODEL TRAINING ON 8 P100 GPUS



Made with MagicSlides

THE CORE IDEA

Self-attention connects every position to every other position in a constant number of sequential steps. A recurrent layer needs $O(n)$ steps. That one change makes long-range dependencies easier to learn and the whole model easier to parallelize.



Made with MagicSlides

Table 1: self-attention has $O(1)$ sequential operations and $O(n)$ maximum path length.

An encoder and a decoder, each a stack of **6 identical layers.**

Scaled dot-product attention

Dot products of queries and keys, scaled by the square root of the key dimension, then a softmax over the values.

Multi-head attention

8 parallel heads attend to different representation subspaces at once.

Positional encoding

Sine and cosine signals added to embeddings, since the model has no recurrence to track order.





How it was trained

01

Hardware

One machine with 8 NVIDIA P100 GPUs. Base model: 100,000 steps in 12 hours.

02

Optimizer

Adam with a learning rate that warms up over 4,000 steps, then decays.

03

Regularization

Residual dropout of 0.1 and label smoothing of 0.1.

512

MODEL DIMENSION

2048

FEED-FORWARD INNER SIZE

8

NUMBER OF HEADS



Made with MagicSlides

State-of-the-art translation at **a fraction of the training cost.**

Beyond translation

A 4-layer Transformer scored 91.3 F1 on English constituency parsing with WSJ data only, 92.7 semi-supervised.

Parallel by design

No sequential recurrence means training scales across GPUs.

Open question in 2017

The authors planned restricted attention for very long inputs like images, audio and video.

