The first transformer (2017)

Original paper: *Attention is all you need* (2017)

30K citations

First neural language model to process sequence without use of recurrent connections or convolutions.

Reached SOTA BLEU / ROUGE scores

NMT w/encoder-decoder architecture

Uses position encodings

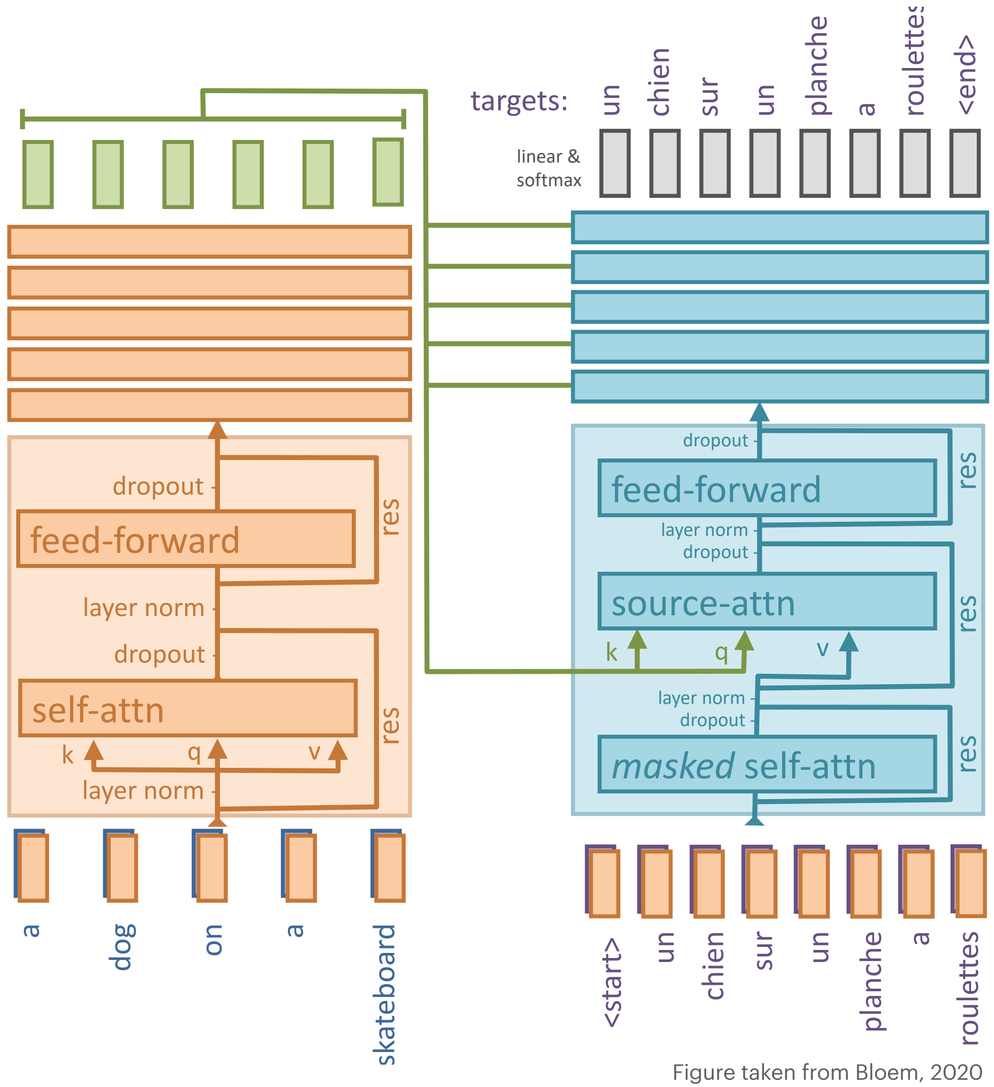


Figure taken from Bloem, 2020

Vaswani et al., *Attention is All You Need*, 2017

Attention and sequential structure

Position Embeddings

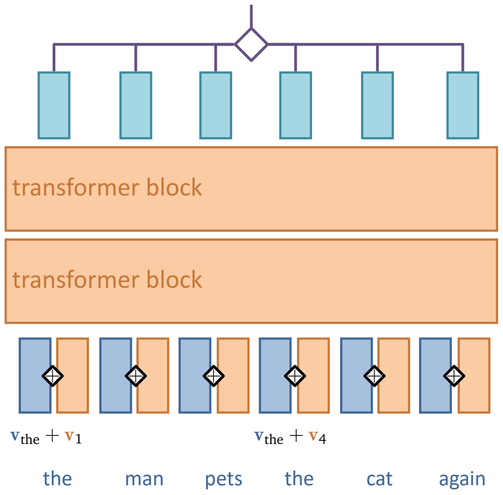
Conceptually simple

Easy to implement

Adds set of learnable parameters

Maximum context length at test time limited to max sequence length in training set

Embedding quality diminishes (in theory) with *t*

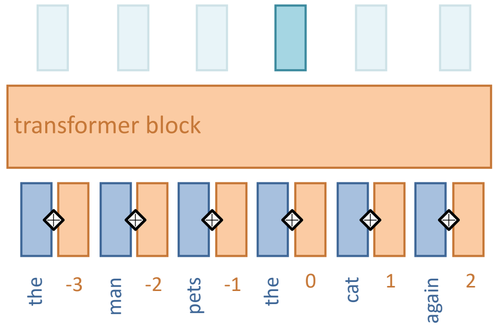


Relative position embeddings

Captures the property of absolute position invariance which is desirable

Like absolute position embeddings, limited in context length at inference

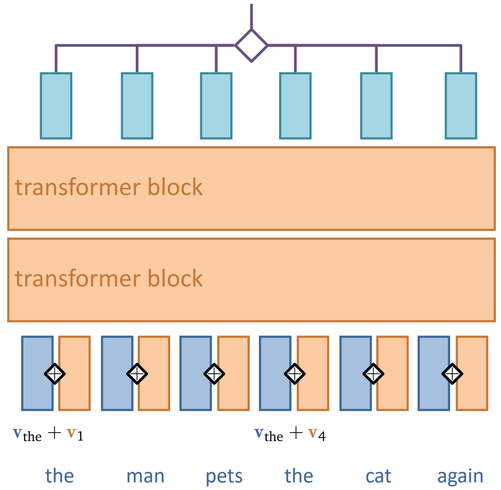
Conceptually and practically more complex; on the surface this results in (T-1)^2 different inputs because each input will have T-1 separate representations (there is a hack that turns this into 2T-1)

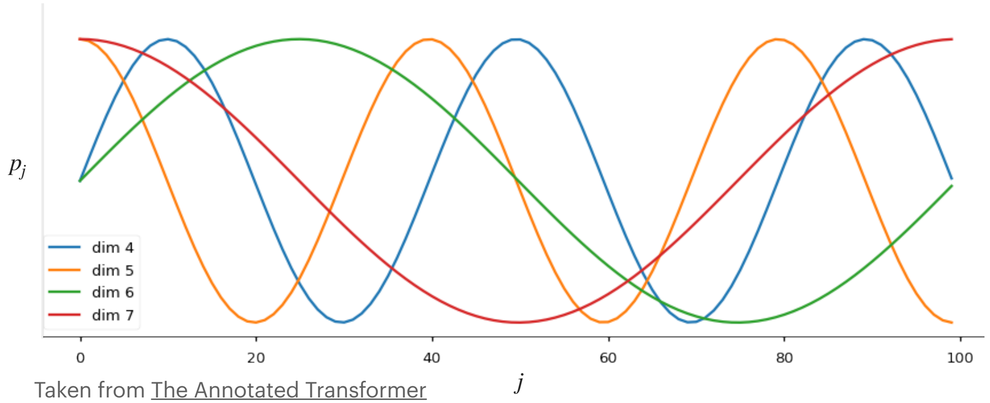


Position encodings

Captures the property of absolute position invariance which is desirable

Conceptually and practically more complex; on the surface this results in (T-1)^2 different inputs because each input will have T-1 separate representations (there is a hack that turns this into 2T-1)





Taken from The Annotated Transformer

Embedding illustrations taken from Bloem (2020)