The transformer architecture

Definition: A neural network architecture that uses self-attention (between layers) as a primary means of expressing relationships between the random variables in the sequence.

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

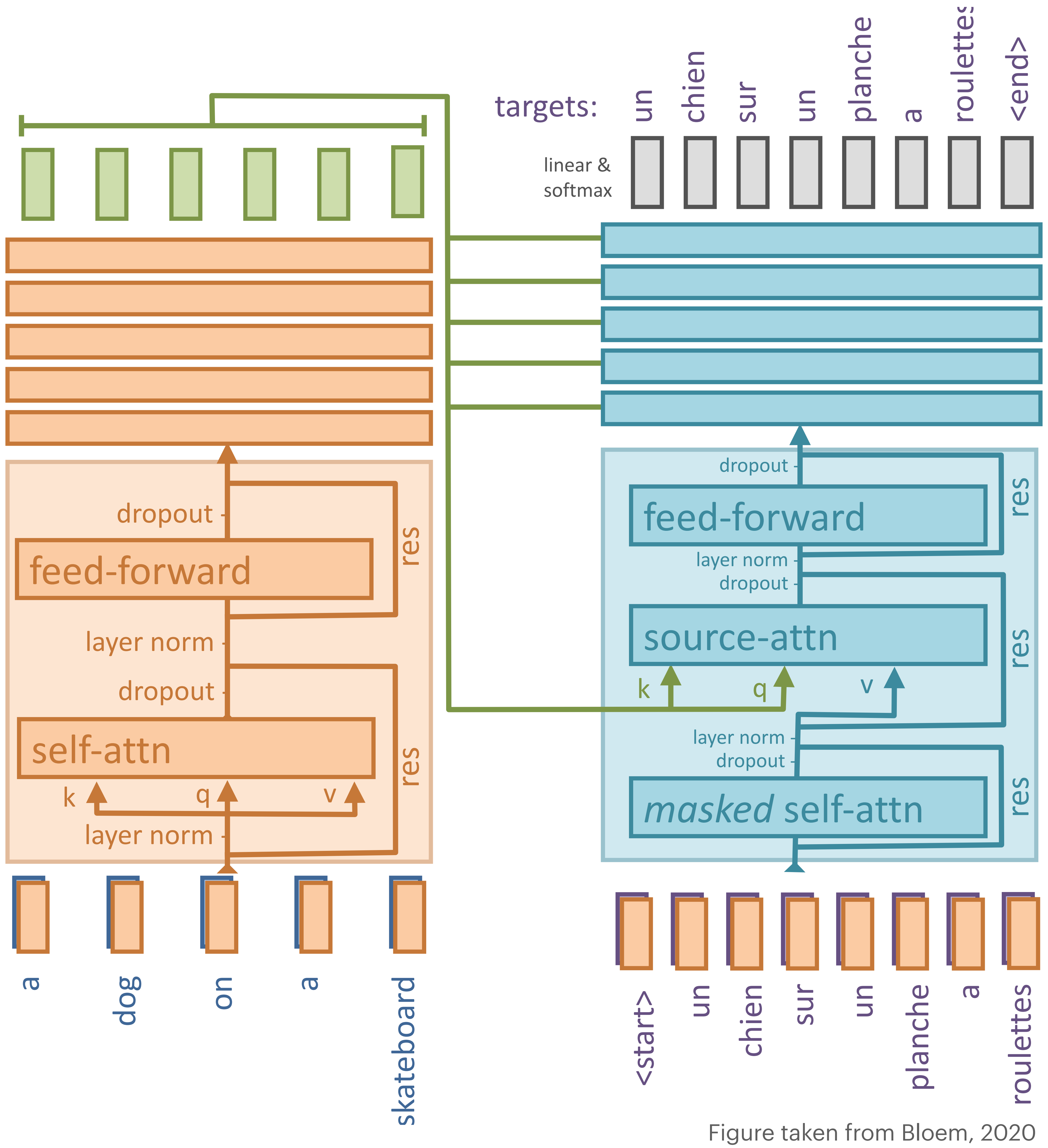

Figure taken from Bloem, 2020

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