1 | Overview

Transformers are an architecture for sequence to sequence translation which lends itself most intuitively to language translation.

They are composed of an encoder and decoder, each of which itself is actually a stack of six smaller encoders/decoders. These encoder/decoder layers (which I'll refer to as just encoder layers for now) are composed of an attention layer and a feedforward neural network.

2 | Encoders

The raw input in the context of language translation would be a vector embedding of each word. Each of these embeddings would be inputted to the layer "on its own path", separate from the rest of the sentence. The attention layer would output intermediate vectors used as input for the neural network, and the vector output of the network would be used as the input for the next encoder in the stack.

There is a layer normalization step after each sublayer (attention, ANN) inside an encoder, in which the initial input embedding matrix (which is just each word embedding stacked) is added to the outputted intermediate representation matrix (also stacked) which is then normalized.

3 | Decoders

Decoders act similarly but have encoder-decoder attention sublayers. There is also a final layer after the decoder stack that translates the vector of floats into words (by having a predetermined dictionary and using the floats as probability for each one).