Back-Background: Transformer-based architectures, primarily used in modeling language understanding tasks, eschew recurrence in neural networks, and trust entirely on self-attention mechanisms to draw global dependencies between inputs and outputs.

What is Self-Attention?

A self-attention module takes in n inputs and returns n outputs. The self-attention mechanism allows inputs to interact with each other, and find out who they should pay more attention to. The outputs are aggregates of these interactions and attention scores.

Background: Self attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence.

Self-attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment, and learning task-independent sentence representations. End-to-end memory networks are based on a recurrent attention mechanism instead of sequence aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks.

The Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output using sequence aligned RNNs or convolution.

Model Architecture: Most competitive neural sequence transduction models have an encoder-decoder structure. Here, the encoder maps an input sequence of symbol representations $(x_1, \ldots, x_n)$ to a sequence of continuous representations $z = (z_1, \ldots, z_n)$. Given z, the decoder then generates an output sequence $(y_1, \ldots, y_m)$ of symbols one element at a time. At each step the model is auto-regressive, consuming the previously generated symbols as additional input when generating the next.