# Summary (short)

The Transformer uses multi-head attention in three different ways. In a self-attention layer all of the keys, values and queries come from the same place. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0. In this work, we presented the Transformer, the ■rst sequence transduction model based entirely on self-attention. We replace the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-Attention. In row (E) we replace our sinusoidal positional encoding with learned positional embeddings. The Transformer is the ■rst transduction model relying on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution. At each step the model is auto-regressive[9], consuming the previously generated symbols as additional input when generating the next.