

Summary (extended)

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 after training for 3.5 days on eight GPUs. The Transformer allows for significantly more parallelization and can reach a new state of the art in machine translation quality. In this work, we presented the Transformer, the first 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. We measure the change in performance on English-to-German translation on the development set, newstest2013. The Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. The Transformer is the first 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. The encoder is composed of a stack of $N=6$ identical layers. In the embedding layers, we multiply those weights by different positions of a single sequence in order to compute a representation.