Attention Is All You Need

Abandonment of Recurrence and Convolution**:** The new model, Transformer, relies entirely on Attention Mechanisms and completely dispenses with the use of Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) in sequence transduction tasks.

Full Reliance on Self-Attention: The model uses the self-attention mechanism to draw global dependencies between inputs and outputs, allowing each position in the sequence to attend to all other positions.

High Parallelism: The Transformer architecture allows for significantly greater parallelism in computations compared to recurrent networks, as the operations do not depend on sequential time steps.

Superior Performance in Machine Translation: The model achieves better results in machine translation quality on the WMT 2014 English-to-German and English-to-French translation tasks, surpassing previous state-of-the-art results (including ensembles) by a significant margin in the English-to-German task.

Training Efficiency**:** The Transformer requires considerably less training time compared to competing models to achieve similar or better performance.

Key Improvements and Additions:

Multi-Head Attention: Instead of using a single attention function, the model runs several attention functions in parallel on different linear projections of the queries, keys, and values. This allows the model to jointly attend to information from different representation subspaces at different positions.

Position-wise Feed-Forward Networks: Each layer in the encoder and decoder contains a fully connected feed-forward network that is applied separately and identically to each position. It consists of two linear transformations with a ReLU activation function in between.

Weight Sharing Between Embedding Layers and the Linear Transformation Before Softmax: The same weight matrix is shared between the input and output embedding layers and the linear transformation preceding the Softmax function in the decoder.

The Transformer introduces a new architecture that relies entirely on attention mechanisms, leading to significant improvements in performance and efficiency in sequence transduction tasks such as machine translation, while incorporating innovative components like multi-head attention and positional encoding.