1) Introduction:

Summary: This paper mainly depicted the difficulty of using recurrent networks with attention mechanisms and its computational overhead(sequential computation and memory constraints) and the attempts of different works to tackle this issue using various techniques.

Learnings: In this paper, the researchers presented the Transformer, a model architecture that foregoes recurrence in favor of drawing global relationships between input and output via an attention mechanism, allowing for substantially higher parallelization and can achieve a new state of the art in translation quality after only twelve hours of training on eight P100 GPUs.

2) Background:

Summary: The Transformer model tries to minimize the problem of sequential processing faced by other solutions like e Extended Neural GPU, ByteNet and ConvS2S by lowering the number of operations to a constant number at the expense of lower effective resolution owing to averaging attention-weighted locations. Transformer is the first transduction model that calculates representations of its input and output using just self-attention rather than sequence aligned RNNs or convolution.

Learnings: Self-attention, also known as intra-attention, is an attention mechanism that connects distinct points in a single sequence to compute a representation of the sequence. Reading comprehension, abstractive summarization, textual entailment, and learning task-independent sentence representations have all been effectively utilized with self-attention.

3) Model Architecture:

Summary: The summary of this study is the transformer model architecture, which employs stacked self-attention and point-wise, completely linked layers for both the encoder and decoder, as well as the functioning processes of the encoder and decoder. Multi head self attention and residual connections are also used by the encoder and decoder. The attention output is produced as a weighted sum of the values, with the weight allocated to each value determined by the query's compatibility function with the relevant key. They also have embeddings, Softmax functions, and position-based Feed-Forward Networks. Because the model has neither recurrence or convolution, "positional encodings" are added to the input embeddings to allow the model to use the order of the sequence.

Learnings: The process through which the encoder maps the input sequences to a sequence of continuous representations, which the decoder utilizes to construct the output sequence one element at a time, is fascinating. At each phase, the previously created symbols are used as extra input for generating the next.

4) Why Self-Attention:

Summary: This section states the things the researchers took into account for using self-attention. The overall computational complexity for each layer is the first one. Another factor to consider is the amount of computing that can be parallelized, as determined by the minimal quantity of sequential operations necessary. The network's path length between long-distance dependents is the third factor.

Learnings:

Self-attention layers are faster than recurrent layers in terms of computational complexity when the sequence length n is less than the representation dimensionality d, which is most often the case with sentence representations used by advanced machine translation models, such as word-piece and byte-pair representations. Self-attention might be limited to evaluating just a neighborhood of size r in the input sequence centered around the relevant output position to increase computing performance for applications requiring very lengthy sequences. The greatest path complexity would thus be O(n/r).