**SEMINAR**

**Attention is all you need**

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

The goal of reducing sequential computation also forms the foundation of the Extended Neural GPU[20], ByteNet [ 15 ] and ConvS2S [ 8],