SOFTWARE

ATTENTION IS ALL YOU NEED

12.07.2025

Here is the link to the famous landmark paper in the recent history
https://arxiv.org/abs/1706.03762

Before this paper, most sequence modeling (e.g., for language) used recurrent neural networks (RNNs) or convolutional neural networks (CNNs). These had significant limitations, such as a difficulty with long-range dependencies and slow training due to sequential processing. The Transformer replaced recurrence with self-attention, enabling parallelization and faster training, while better capturing dependencies in data. So the transformer architecture became the foundation for nearly all state-of-the-art NLP models. This enabled training models with billions of parameters, which is key to achieving high performance in AI tasks.