Intro to Transformer

Hi everyone, in this video I will explain to you all the concepts and features of Transformer, an AI model that has been revolutionizing not only the NLP space but also the whole AI world.

The Transformer’s story starts with the famous paper: ‘Attention Is All You Need’, published by Vaswani and co-authors in 2017 at Google Brain and Google Research.

But why this publication is a game changer? The answer is that it solved the problems remaining in NLP for a long time till 2017, and then open the way to the super AI model that can understand and generate human-like language.

So first, let’s take a look which model has dominated NLP before Transformer.

Until 2017, the common architectures like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU) were the mainstream in NLP tasks.

Take machine translation from English to French using RNN as an example, It encode the input text and decode the output, sequentially one word as a time. As it processes each word, it generates a hidden state that captures the information of the sequence up to that point.

Due to this architecture, there are 3 main challenges of RNN that make it extremely difficult to train very-large dataset and learn from long context:

* Firstly, sequential computation: the fact that the sequential processing of text (one word of at a time) made it difficult to leverage modern, highly parallel hardware like GPUs effectively. Longer sequence, larger dataset, make training extremely long.
* Secondly, loss of information: When the distance between relevant information and the point where it is needed increases they tend to lose the ability to learn the connections.
* Thirdly, vanishing problem: made it hard for the model to learn correlations between distant events in a sequence