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.

Now let’s see the some limitations of RNNs.

Take language modeling task as an example. Given a sentence starts with some words, like “She stands up and opens the …” ,the model needs to predict the next word. By using RNN, the model processes one word at a time, to generate the hidden state for the next computation, meaning that it is a sequential computation. As a consequence, there are less rooms for parallel computation. So longer sequence, longer computation time.

Next, due to its nature of sequential computation, the contributions of initial states or information to the final state or the prediction are very small for long sequence context. This leads to the loss of information for long-range dependencies.

Another major limitation of RNNs is that they are prone to vanishing or exploding gradient problem. RNN use Backpropagation Through Time to updates the weights. For example, to calculate the derivative of loss function L with respect to the weights, it uses the chain rule, to calculate the gradient products from the last layer/ last instant to the initial layer or initial install. If the gradient of the hidden state t with respect to the previous state t-1 is very small than 1, then the gradient products are very very small, at certain level, we can say that it is vanished. On the other hand, if the component gradients are very large than 1, then the gradient products are very large, at certain level, we can say that it is exploded. These problems make training RNN unstable and extremely hard.