One of the most exciting developments in deep learning has been the transformer Network, or sometimes called Transformers. This is an architecture that has completely taken the NLP world by storm. And many of the most effective albums for NLP today are based on the transformer architecture. It is a relatively complex neural network architecture, but in this and the next three videos after this will go through it piece by piece. So that by the end of this next four videos, you have a good sense of how the transformer Network works and we'll be able to apply zero problems.

As the complexity of your sequence task increases, so does the complexity of your model. We have started this course with the RNN and found that it had some problems with vanishing gradients, which made it hard to capture long range dependencies and sequences. We then looked at the GRU and then the LSTM model as a way to resolve many of those problems where you may use of gates to control the flow of information. And so each of these units had a few more computations.

While these editions improved control over the flow of information, they also came with increased complexity. So as we move from our RNNs to GRU to LSTM ,the models became more complex. And all of these models are still sequential models in that they ingested the input, maybe the input sentence one word or one token at the time. And so, as as if each unit was like a bottleneck to the flow of information. Because to compute the output of this final unit, for example, you first have to compute the outputs of all of the units that come before. In this video, you learned about the transformer architecture, which allows you to run a lot more of these computations for an entire sequence in parallel. So you can ingest an entire sentence all at the same time, rather than just processing it one word at a time from left to right.

The Transformer Network was published in a seminal paper by a Vaswani ,Norm Shakespeare, Nikki Palmer, Jacob was great ,line James, Gomez, Lucas Kaiser and earlier Polish. One of the inventors of the Transformer network, Lucas Kaiser, is also co instructor of the NLP specialization with deep learning dot AI. So you can check that out as well when you're done with this deep learning specialization. The major innovation of the transformer architecture is combining the use of attention based representations and a CNN convolutional neural network style of processing. So an RNN may process one output at the time, and so maybe y(0) feeds in to them that you compute y(1) and then this is used to compute y(2).

This is a very sequential way of processing tokens, and you might contrast with a CNN that can take input a lot of pixels.

Yeah, or maybe a lot of words and can compute representations for them in parallel. So what you see in the Attention Network is a way of computing very rich, very useful representations of words. But with something more akin to this CNN style of parallel processing. To understand the attention network, there will be two key ideas will go through in the next few videos.

The first is self attention. The goal of self attention is, if you have, say, a sentence of five words will end up computing five representations for these five words, was going to write A1,A2,A3, A4 and A5. And this will be an attention based way of computing representations for all the words in your sentence in parallel. Then multi headed attention is basic A for loop over the self attention process. So you end up with multiple versions of these representations. And it turns out that these representations, which will be very rich representations, can be used for machine translation or other NLP toss to create effectiveness. So in the next video, let's jump in to learn about self attention, to compute these rich representations. The video after that, we'll talk about multi headed attention. And then the final video on transforming networks will put all of these together so that you understand how the entire transformer architecture works into end. Let's go to the next video.