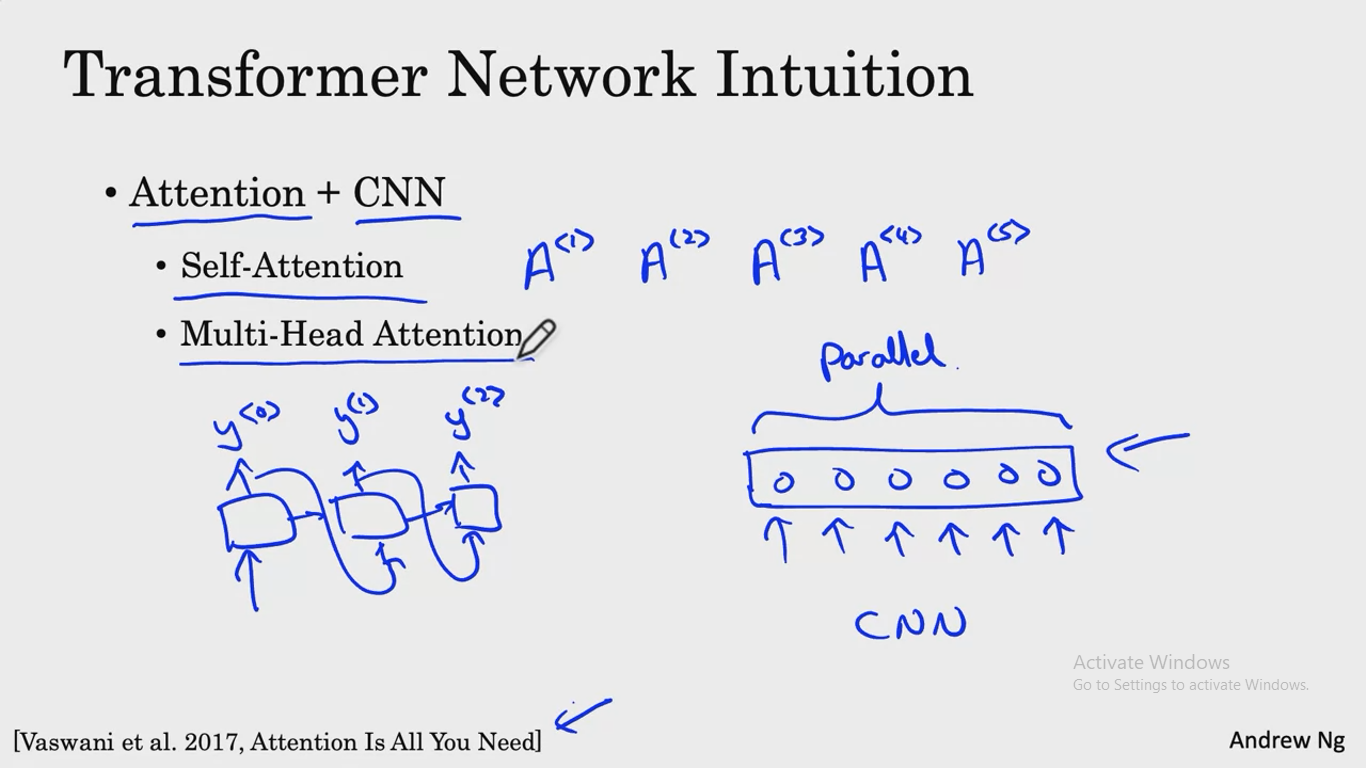
**Transformers**



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.