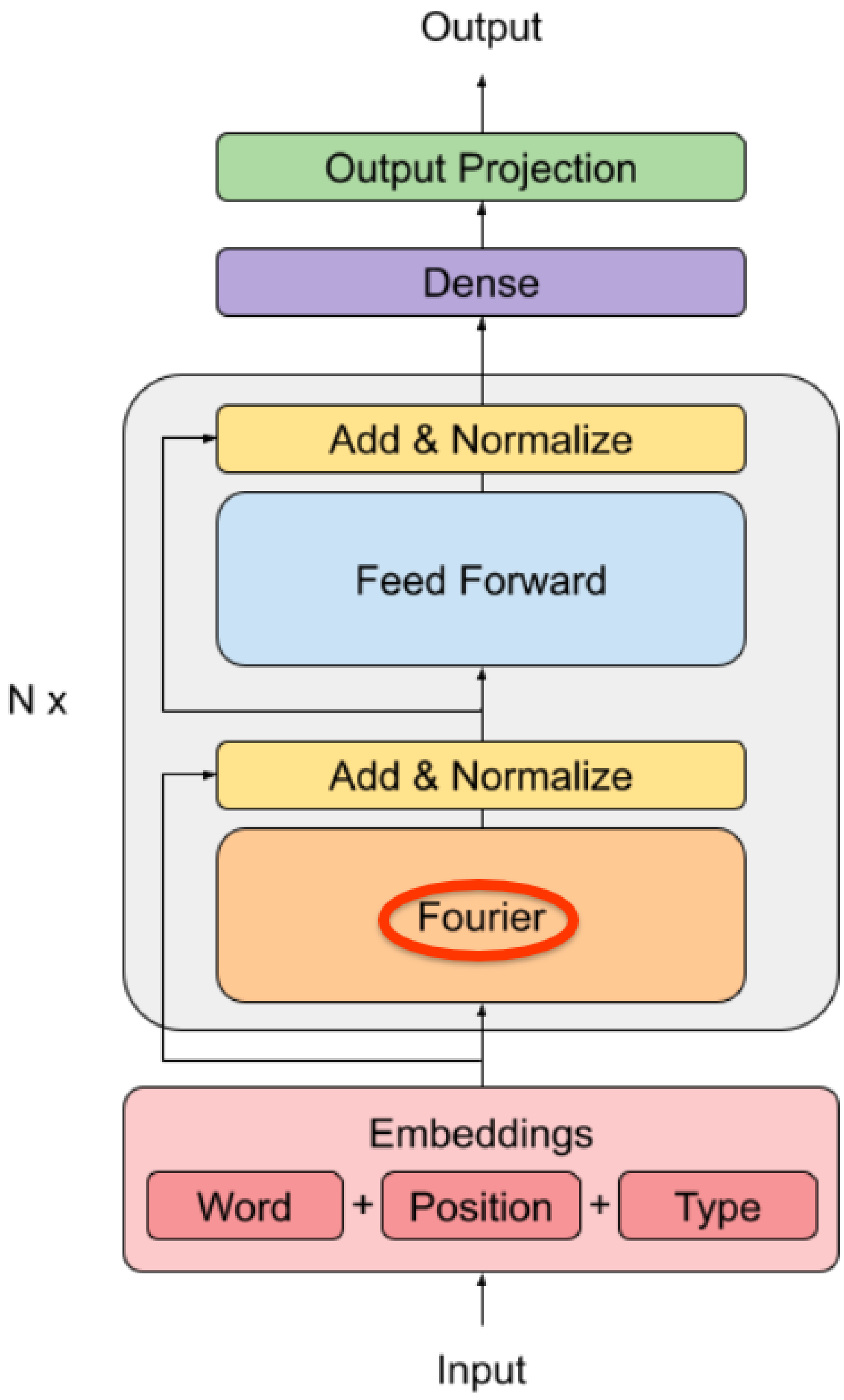
Twitter Thread by ilyaeck

ilyaeck
@ilyaeck

Attention may be all you *want*, but what you *need* is effective token mixing!
In which we replace Transformers' self-attention with FFT and it works nearly as well but faster/cheaper.
https://t.co/GiUvHkB3SK
By James Lee-Thorpe, Joshua Ainslie, @santiontanon and myself, sorta



Attention clearly works - but why? What's essential in it and what's secondary? What needs to be adaptive/learned and what can be precomputed?

The paper asks these questions, with some surprising insights.

These questions and insights echo other very recent findings like @ytay017's Pretrained CNNs for NLP https://t.co/k0jOuYMxzz and MLP-Mixer for Vision from @neilhoulsby and co. (Like them, we also found combos of MLP to be promising).