Andrew Taylor EN.705.743 10/02/25

Reading Assignment 5: Transformers

When I think about the original transformer architecture described in Vaswani et al.'s 2017 paper, it was built around an encoder–decoder structure. The encoder processed the input sequence into a representation, and the decoder generated an output sequence step by step. This design was perfect for translation tasks, where the model needed to map one language to another. However, over time, researchers realized that many tasks didn't actually need this two-part pipeline. For example, if I'm just trying to predict the next word in a sequence (language modeling), why force the model to carry around the overhead of an encoder when the decoder by itself (the autoregressive part as he called it)could do the job?

Modern transformers, like GPT, distill the architecture down to only that decoder stack. Others, like BERT, focus only on the encoder. The choice depends on the type of task: encoders are great for understanding and classification, while decoders are better for generation. To me, this shift is like moving from a Swiss Army knife to a specialized tool. The original design could do everything, but in practice, it turned out that using only the blade, or only the screwdriver was more efficient for specific tasks.

Another change is in the use of components like positional encodings, normalization, and attention masking. In the earliest models, sinusoidal positional encodings were fixed, but modern architectures often learn positional embeddings, which makes them more flexible. Similarly, researchers experimented with where to put normalization layers, and small tweaks here ended up improving training stability.

Overall, the architecture evolved because researchers wanted to simplify, specialize, and scale. By stripping away the "historical baggage," transformers became faster to train and easier to adapt. In the end, it wasn't about throwing away the original idea, but about realizing that the most efficient solutions are often the ones that use only the parts that are truly necessary for the task at hand.