Sai Priya Veerabomma
002814292

Summary: “Attention Is All You Need”

The paper *“Attention Is All You Need”* introduced the Transformer model, which completely changed the way we approach natural language processing (NLP). Before this, most models used RNNs or LSTMs, which processed text one word at a time. That made them slow to train and not great at understanding relationships between words that were far apart in a sentence. CNNs were a bit better at handling these relationships, but they still had to go deep to catch long-range dependencies.

The Transformer got rid of both recurrence and convolution altogether and instead used something entirely different **attention**. This meant it could look at all words in a sentence at once and figure out how they relate to each other, no matter how far apart they are. Because of this, it trained much faster and understood context better.

The model follows an encoder-decoder setup, with both sides made up of six repeating layers. One of the core ideas here is **scaled dot-product attention**, where the model figures out how much focus to give each word by calculating scores based on queries, keys, and values. Instead of doing this once, it uses **multi-head attention** essentially several attention processes running in parallel so it can look at different aspects of the sentence at the same time.

Since attention doesn’t keep track of word order on its own, the model adds **positional encodings** using sinusoidal functions. This lets it understand where each word is in a sequence. After attention, each word passes through a simple feed-forward network, and **residual connections** and **layer normalization** help the model learn more effectively and avoid issues during training.

What’s really impressive is how well the Transformer performed. It beat previous models in machine translation tasks, hitting 28.4 BLEU on English-to-German and 41.8 on English-to-French (WMT 2014). And it did this in just 12 hours of training on 8 GPUs—compared to the days it used to take with RNNs. It also did well on parsing tasks, showing that it could generalize to different kinds of language problems.

The attention mechanism itself became widely talked about not just in research papers but in educational videos and blog posts too. A lot of these explain how attention lets each word "look around" and decide which other words are most relevant. Visualizing what different attention heads focus on has helped people understand what the model is doing under the hood.

The Transformer didn’t just influence NLP. It laid the groundwork for models like BERT, GPT, and T5, and even extended into other areas like computer vision (with ViT), multimodal models (like CLIP and DALL·E), and scientific research (such as AlphaFold for protein folding). It showed that scaling up attention-based models can produce powerful, general-purpose systems—basically laying the foundation for what we now call **foundation models**.

That said, it's not perfect. The biggest issue is that its self-attention mechanism requires a lot of memory it scales quadratically with sequence length, which makes it hard to use with really long texts. Also, the original paper focused mostly on translation, so other applications came later. And while the model worked well from the start, researchers didn't fully understand *why* it worked so well something that's still being explored.

In short, the Transformer proved that attention mechanisms alone could handle complex sequences, opening the door to faster and more powerful models across multiple domains. Its impact has been massive, and it continues to shape the future of AI.