Attention Is All You Need - Paper Summary

# 🧠 Core Message of the Paper

We can completely replace recurrent (RNN/LSTM) or convolutional models for sequence processing with a new architecture called the Transformer, which is based only on attention mechanisms — and it works better, faster, and is more parallelizable.

# ⚙️ What Problem Does It Solve?

Before Transformers, models like RNNs and LSTMs processed data sequentially, which made:  
- Training slow (no parallel processing)  
- Long-range dependencies hard to capture  
- Gradient issues common  
  
The Transformer solves these problems by using self-attention, which:  
- Lets the model look at all positions at once  
- Captures global context more easily  
- Enables parallel computation

# 🏗️ What Is the Architecture?

The model has two parts:  
1. Encoder – processes the input sentence (e.g., English)  
2. Decoder – generates the output sentence (e.g., French)  
  
Each part is built using layers of:  
- Multi-Head Self-Attention  
- Feed-Forward Networks  
- Residual Connections + Layer Normalization

# 🔍 What Is Attention? (In Simple Terms)

Think of attention as:  
“For each word, figure out which other words in the sentence it should pay attention to.”  
  
This is done by:  
- Computing Query (Q), Key (K), and Value (V) vectors for each word  
- Using dot products between Q and K to compute attention scores  
- These scores determine how much each word contributes to the final representation of another word

# 🧩 Important Components

Component - Purpose  
Self-Attention - Each word looks at every other word to understand context  
Multi-Head Attention - Captures different kinds of relationships simultaneously  
Positional Encoding - Injects word position into the model  
Feed-Forward Layer - Adds non-linearity and depth  
Layer Normalization & Residuals - Stabilize and speed up training

# ⚡ Why Is This Important?

- Faster: No sequential processing means better use of GPUs  
- Better Performance: Beats RNN-based models on machine translation  
- Scalable: Basis for large models like GPT, BERT, etc.

# 🧪 Experimental Result

On machine translation (e.g., English → German), the Transformer outperforms state-of-the-art models while being much faster to train.

# 🚀 What Happened After the Paper?

The architecture became the standard for almost all NLP models today, including:  
- BERT – encoder-only  
- GPT – decoder-only  
- T5, BART – encoder-decoder