**What I Understood from the Paper "Attention Is All You Need"**

The paper introduces a new architecture for sequence transduction tasks (like translation) called the **Transformer**, which **completely removes recurrence and convolution** and instead relies **entirely on attention mechanisms**.

**Key Points I Understood:**

1. **Main Idea**:
   * The paper proposes that **attention mechanisms alone** can handle sequence modeling tasks effectively, **without using RNNs or CNNs**.
   * This simplifies the architecture and allows for more parallelization during training.
2. **Transformer Architecture**:
   * The model is built with an **encoder-decoder** structure:
     + **Encoder**: Maps input sequence to a continuous representation.
     + **Decoder**: Uses this representation to generate output one token at a time.
3. **Attention Mechanism**:
   * The core component is **Scaled Dot-Product Attention**, where queries (Q), keys (K), and values (V) are projected from the input.
   * The output is a weighted sum of values, where weights are calculated from the dot product of queries and keys.
4. **Multi-Head Attention**:
   * Instead of applying a single attention mechanism, the model uses **multiple attention heads** in parallel to capture different types of relationships.
5. **Positional Encoding**:
   * Since the model doesn’t use recurrence, **positional encodings** are added to the input embeddings to retain the order of the sequence.
6. **Advantages**:
   * **Faster training** thanks to parallelism.
   * **Better performance** on tasks like machine translation (e.g., English-to-German).
   * **Simpler architecture** with fewer layers and operations than RNN-based models.
7. **Results**:
   * The Transformer achieved **state-of-the-art results** on translation benchmarks at the time, with significantly **less training time**