Summary of Neural Machine Translation by Jointly Learning to Align and Translate

This paper introduced an innovative approach to Neural Machine Translation (NMT) by integrating an **attention mechanism** into the encoder-decoder architecture. Traditional NMT models encoded the entire source sentence into a single fixed-length vector, which often struggled with long or complex sentences. The authors proposed allowing the model to **dynamically focus on different parts of the source sentence** during translation, thereby improving performance and handling longer inputs more effectively

Example

Input Sentence (English):
"I am a student"

Expected Output (French):
"Je suis un étudiant"

Step 1: Traditional Encoder-Decoder (No Attention)

1. Encoder reads the entire input: "I am a student"
2. It converts the sentence into a single fixed-length vector like this:
   h = Encoder("I am a student") → [0.2, -0.5, ..., 0.1]
3. Decoder takes this vector and tries to generate: "Je suis un étudiant"

Problem: The single vector must capture all sentence information, which is hard for long or complex sentences. Translation quality drops for long sentences.

Step 2: Attention-Based NMT

Instead of one fixed vector, the attention mechanism allows the decoder to look at different parts of the input sentence for each word it generates.