Summary of Effective Approaches to Attention-based Neural Machine Translation

Global Attention: Considers all source words when generating each target word.

Local Attention: Focuses on a subset of source words, reducing computational load while maintaining performance.

Example

Translate English to French

Input (English):

"The cat sits on the mat."

Without Attention (Standard Encoder-Decoder)

The entire sentence is encoded into a **single fixed-length context vector**, which the decoder uses to generate the translation.

Problem:

- This fixed vector may not represent all parts of long or complex sentences well.
- The decoder has no direct access to individual source words while generating each target word.

Output (French):

"Le chat est sur."
(*"The cat is on."* — incomplete translation)

With Attention (Global or Local)

How it works:

- While generating each word, the decoder **looks back (attends)** to relevant parts of the source sentence.
- For example, when generating "tapis" (mat), it focuses on the representation of "mat" in the encoder output.

Output (French):

"Le chat est assis sur le tapis."
(*"The cat is sitting on the mat."* — accurate and complete)

Explanation:

- While decoding "tapis", attention weights will be highest on "mat".
- While decoding "assis", attention focuses on "sits".