Effective Approaches to Attention-based Neural Machine Translation Summary

Paper Link: https://arxiv.org/pdf/1508.04025

This paper looks at how to make Neural Machine Translation (NMT) more effective by using attention mechanisms. NMT systems usually rely on encoder-decoder architectures based on recurrent neural networks, which try to model the probability of a target sentence given a source one. But traditional models compress the entire source sentence into a single vector which leads to loss of information.

The paper introduced two main attention strategies: **Global Attention** and **Local Attention**. Global attention looks at all the words in the source sentence for every word it generates. It's similar to previous work but designed to be simpler. Local attention is more selective: it only looks at a small window of source words around a predicted position. It comes in two variations, one assumes words are aligned roughly in order (local-m), and the other predicts the alignment dynamically (local-p). This local setup is faster and more efficient, especially with long sentences, and still performs well.

Another key idea in the paper is the **input-feeding approach**. Instead of just feeding in the target word at each step, the model also feeds in the attention output from the previous step. This gives the model a sense of what it previously focused on, helping it make better alignment decisions over time.