Effective Approaches to Attention-based Neural Machine Translation
arXiv:1508.04025v5 [cs.CL] 20 Sep 2015 Authors: Minh-Thang Luong, Hieu Pham, Christopher D. Manning (Stanford University)

NMT achieved state-of-the-art (SOTA) results but traditional NMT models process the entire source sentence before translating, potentially struggling with long sentences.

The concept of "attention" in neural networks allows models to focus selectively on parts of the input when generating output.

Bahdanau et al. (2015) successfully applied attention to NMT to jointly translate and align words. This paper explores alternative attention architectures.

Global Attention:

- The decoder attends to **all** source words for each target word prediction.
- Computationally more expensive, especially for long source sentences.
- Resembles the approach in Bahdanau et al. (2015) but is presented as architecturally simpler.

Local Attention:

- The decoder attends only to a **subset** of source words within a predicted window for each target word prediction.
- Computationally less expensive than global attention.
- Can be viewed as a blend between hard (non-differentiable) and soft (differentiable) attention: it's mostly differentiable, making training easier than pure hard attention.
- Two variants explored:
    - **Monotonic Alignment (local-m):** Assumes the target word aligns monotonically to a position around the source word at the same index (plus an offset).
    - **Predictive Alignment (local-p):** The model predicts the aligned position in the source sentence for each target word.

Alignment Functions:

- Different functions are examined to score how well each source hidden state aligns with the current target decoder state (before the softmax to get attention weights):
    - Location-based
    - Content-based (dot, general, concat)

Both approaches are effective.