**SOME NOTES :**

This paper examines two simple and effective classes of attentional mechanism: a global approach which always attends to all source words and a local one that only looks at a subset of source words at a time.

Neural machine translation has shown promising results, but existing models struggle with long sentences due to the fixed-length context vector.

The encoder-decoder framework processes the input sequence into a fixed-length vector, which the decoder uses to generate the target sequence.

Attention mechanisms address this limitation by dynamically computing a context vector based on the input sequence at each decoding step.

In global attention, the model considers all source positions when computing the context vector, ensuring comprehensive coverage.

Local attention restricts the focus to a fixed window of source positions, reducing computational complexity for long sequences.

We evaluate three alignment functions: dot product, general, and concat, each offering different trade-offs in expressiveness and efficiency.

They train the models on the WMT’14 English-to-German dataset, consisting of 4.5 million sentence pairs, and the IWSLT English-to-Vietnamese dataset."

Performance is measured using the BLEU score, a standard metric for evaluating translation quality against reference translations.

Our attention-based models outperform non-attentional baselines by up to 5.9 BLEU points on the English-to-German task.

Local attention achieves comparable performance to global attention while being significantly faster, especially for long sentences.

The dot product alignment function proves effective, offering simplicity and performance competitive with more complex alternatives.

Visualizations of attention weights reveal that our models learn meaningful alignments, often corresponding to linguistic intuitions.

The input-feeding approach consistently reduces translation errors, such as repetitions, improving overall fluency.

Paper link: <https://arxiv.org/abs/1508.04025>