Neural Machine Translation by Jointly Learning to Align and Translate

Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio

Neural Machine Translation (NMT) is an emerging approach that replaces traditional phrase-based systems with end-to-end neural networks capable of directly mapping a source sentence to its translation. Early NMT models, mainly encoder–decoder architectures, compressed an entire sentence into a fixed-length vector before decoding. While effective for short inputs, this approach degraded significantly with long sentences, as crucial contextual information was lost. Prior works by Kalchbrenner & Blunsom (2013), Sutskever et al. (2014), and Cho et al. (2014) demonstrated the potential of NMT but highlighted this bottleneck, motivating the search for more flexible representations that preserve long-range dependencies.

To address this, the authors introduce **RNNsearch**, an attention-based model that jointly learns to align and translate. Instead of a single vector, the encoder produces a sequence of context-rich annotations using a bidirectional RNN, while the decoder dynamically applies soft alignment to focus on the most relevant parts of the input during each step of translation. Experiments on the large-scale English–French WMT’14 dataset show that RNNsearch not only surpasses conventional encoder–decoder models but also achieves translation performance comparable to state-of-the-art phrase-based systems. Notably, the model proves more robust for long sentences and its learned soft alignments align well with linguistic intuition, offering both improved accuracy and interpretability.