Neural Machine Translation by Jointly Learning to Align and Translate

Authors: Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio
Conference: ICLR 2015
arXiv ID: 1409.0473v7

Background

- Traditional NMT compresses the entire source sentence into one fixed vector → poor for long sentences.
- Attention-based mechanism allows dynamic focus on relevant parts of the input during translation.

Proposed Approach: RNNsearch

- Instead of a single vector, the encoder outputs a **sequence of annotations** (context-rich vectors).
- Decoder uses a **soft attention mechanism** to compute a context vector from source annotations.

Model Architecture

- **Encoder**: Bidirectional RNN (BiRNN)
  - Combines forward and backward hidden states for each source word.
- **Decoder**: RNN with attention
  - At each step, generates a word using previous word, previous hidden state, and a context vector.
- **Context Vector (c_i)**:
  - Weighted sum of encoder annotations.
  - Weights (α_{ij}) represent soft alignment between target word and source annotations.
  - Alignment model is a feedforward neural network trained jointly with the rest of the model.

Results

- **BLEU Scores**:
  - RNNsearch significantly outperformed RNNencdec on all benchmarks.
  - RNNsearch-50 achieved BLEU = 34.16 (close to Moses system: 35.63).
- **Effect of Sentence Length**:
  - RNNencdec degraded with longer sentences.

    - RNNsearch models were more stable and accurate on long sentences.
- **Qualitative Analysis**:
    - Attention weights (α_{ij}) show meaningful, mostly monotonic alignments.
    - Able to handle non-monotonic alignments (e.g., adjective-noun inversion).

Conclusions

- Fixed-length vectors limit traditional NMT performance.
- Attention mechanism improves translation, especially for long sentences.
- RNNsearch:
    - Learns useful alignments.
    - Achieves performance comparable to phrase-based systems.
    - End-to-end trainable.