NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

[1409.0473] Neural Machine Translation by Jointly Learning to Align and Translate

Key Innovation: Attention Mechanism

1. **Problem:** Traditional Seq2Seq models compressed entire source sentences into a single fixed-length vector, creating an information bottleneck.
2. **Solution:** Introduced soft attention to dynamically focus on relevant source words while generating each target word.

Architecture (RNNsearch)

1. **Bidirectional Encoder:**
    - Uses forward/backward RNNs to encode source sentence into annotations (hidden states).
2. **Decoder with Attention:**
    - At each step, computes a context vector as a weighted sum of source annotations.
    - Weights are learned via an alignment model (a feedforward network).
3. **Joint Training:**
    - Alignment and translation are learned simultaneously (end-to-end).

Advantages Over Seq2Seq

- **Handles Long Sentences:** No longer limited by fixed-length vectors.
- **Interpretability:** Attention weights visualize word alignment (e.g., verbs aligning to verbs).
- **Performance:** Achieved BLEU scores comparable to phrase-based SMT on English-French (WMT'14).

Limitations

- **Computational Cost**: Global attention scales with source length (later addressed by Luong's local attention).
- **No Input-Feeding:** Decoder doesn't explicitly track past alignment decisions (improved by Luong et al.).