Effective Approaches to Attention-based Neural Machine Translation

[1508.04025] Effective Approaches to Attention-based Neural Machine Translation

Objective:

The paper explores **attention mechanisms** in neural machine translation (NMT), proposing two novel attention-based models to improve translation quality by dynamically focusing on relevant parts of the source sentence.

Key Contributions:

1. **Attention Mechanisms**:
   - **Global Attention**: Considers all source words for each target word (computationally expensive but comprehensive).
   - **Local Attention**: Focuses on a small window of source words around a predicted position (efficiency-speed trade-off).
2. **Model Variants**:
   - **Input-feeding**: Integrates previous attention information into the current step (improves coherence).
   - **Location-based**: Uses positional features to handle alignment monotonicity (e.g., for languages like German→English).
3. **Experiments**:
   - **Datasets**: WMT English-German (4.5M sentences) and English-Czech (15M sentences).
   - **Results**:
     - **Global Attention**: Achieved +2.8 BLEU over non-attentional baselines.
     - **Local Attention**: Matched global attention quality with 50% fewer computations.
     - **Input-feeding**: Added +1.3 BLEU by maintaining attention history.
4. **Findings**:
   - **Attention is Crucial**: Both global and local attention outperform non-attentional models.

- **Hybrid Approaches**: Combining local attention with input-feeding yielded the best results.
- **Scalability**: Local attention scaled better to long sentences without quality loss.

Significance:

- Introduced **practical attention variants** balancing accuracy and efficiency.
- Demonstrated that **input-feeding** stabilizes training.
- Inspired later architectures like the Transformer .

Limitations:

- Global attention remains expensive for very long sequences.
- Handcrafted features (e.g., positional bias) were later superseded by learned attention.