Name: Suhaila Ahmed Hassan
Track: AI
Branch: Alexandria

Effective Approaches to Attention-based Neural Machine Translation

Introduction

Paper introduces two new attention approaches that significantly improve translation quality:

- Global attention, which considers all source words.
- Local attention, which focuses on a subset of source words at a time.

Attention-based Models

Global Attention:

- Focuses on all the source positions when deriving the context vector.
- A variable-length alignment vector is calculated by comparing the target hidden state with each source hidden state.
- The context vector is then calculated as a weighted average of all source states based on this alignment.
- Drawback: It has to attend to all words on the source side for each target word, which is expensive and can potentially render it impractical to translate longer sequences.

Local Attention:

- Focuses on a small subset of the source positions for each target word.
- It is less expensive and more practical for longer sequences.
- It selects a small window of context around the aligned position for the target word.
- Two variants of local attention are:
    a. Monotonic Alignment: Assumes source and target sequences are monotonically aligned.
    b. Predictive Alignment: Predicts the aligned position for each target word.

Global & Local Attention Models:

- Both use the hidden state from the top layer of a stacking LSTM at each decoding step to generate a context vector that captures relevant source-side information for predicting the current target word
- The target hidden state and the source-side context vector are combined using a concatenation layer to produce an attention hidden state.
- This attention vector is then fed into a softmax layer to produce the probability distribution for the target word.

Input-feeding Approach:

- In both models, alignment decisions are made independently, which can be suboptimal.
- The input-feeding approach addresses this by concatenating attentional vectors with the input at the next time steps.
- This approach makes the model aware of previous alignment decisions and helps in creating deeper networks.

Experiments

- Evaluates models on WMT English-German translation tasks, testing both directions.
- Development set: newstest2013 (3000 sentences) for hyperparameter selection.
- Test sets: newstest2014 (2737 sentences) and newstest2015 (2169 sentences).
- Translation quality reported using case-sensitive BLEU (tokenized and NIST BLEU), allowing comparison with existing NMT work.

Results

- Attentional models perform better than non-attentional models in terms of minimizing test cost, especially with dropout regularization.
- The attentional models are more effective when translating longer sentences, complex names and sentences.
- For choices of attentional architecture, the local attention model with predictive alignments "local-p" model performed best, achieving low perplexity and high BLEU scores.
- Local attention models achieved better alignment quality than global ones. However, translation quality and alignment quality do not always correlate well.