Neural Machine Translation By Jointly Learning To Align And Translate

Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio
Motivation

Existing encoder-decoder frameworks

Decoder

Encoder

Fixed-length representation is a potential bottleneck for long sentences

Evaluated on news-test-2014 from WMT '14 English-French parallel corpora
Method

For each generated word $y_t$ in the translation, soft-searches for a set of positions $(1, ..., T)$ in a source sentence $x=(x_1, ..., x_T)$ where the most relevant information is concentrated.

The predicted target word $y_t$ is based on the context vectors $c_t=\sum a_{t,j} h_j$ associated with these source positions and all the previous generated target words $s_{t-1}, y_{t-1}$.
Encoder - Bidirectional RNN

- Map input sentence $x$ to a sequence of annotations $h$
- Each annotation summarizes the preceding words and the following words

$$x = (x_1, \cdots, x_{T_x}) \quad h_t = f(x_t, h_{t-1}) \quad \overrightarrow{h_1}, \cdots, \overrightarrow{h_{T_x}} \quad \text{Concat} \quad h = (h_1, \cdots, h_{T_x})$$

where $h_j = [\overrightarrow{h_j}; \overleftarrow{h_j}]^\top$
Decoder - Alignment Model

- Compute alignment $\alpha_{ij}$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

How well the inputs around position $j$ and the output at position $i$ match

Use simple feedforward NN to compute $e_{ij}$ based on $s_{i-1}$ and $h_j$

$$e_{ij} = v^T \tanh(Ws_{i-1} + Vh_j)$$

where $v, W, V$ are trainable weights
Decoder

- Compute context \( c_i \)
  \[
  c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j
  \]
- Compute new decoder state \( s_i \)
  \[
  s_i = f(s_{i-1}, y_{i-1}, c_i)
  \]
- Generate new output \( y_i \)
  \[
  \text{argmax } p(y_i | y_1, \ldots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i)
  \]
Evaluation

Models
- RNNenc-max sentence length (baseline, blue)
- RNNsearch-max sentence length (red)
- Trained by minimizing mean $\log(y|x, \theta)$

Dataset
- WMT ‘14 English-French parallel corpora (348M words, 30k frequent words)
- Test split: 3003 sentences

Metrics
- BLEU score
Summary

- A fixed-length context vector is the bottleneck for translating long sentences.
- Alignment mechanism, i.e. soft-search for a set of input words or annotations, enhances translation on longer sentences.
- Alignment models trained and evaluated on English-to-French translation achieve higher accuracy than fixed-length encoder-decoder models, especially on longer sentences.
References