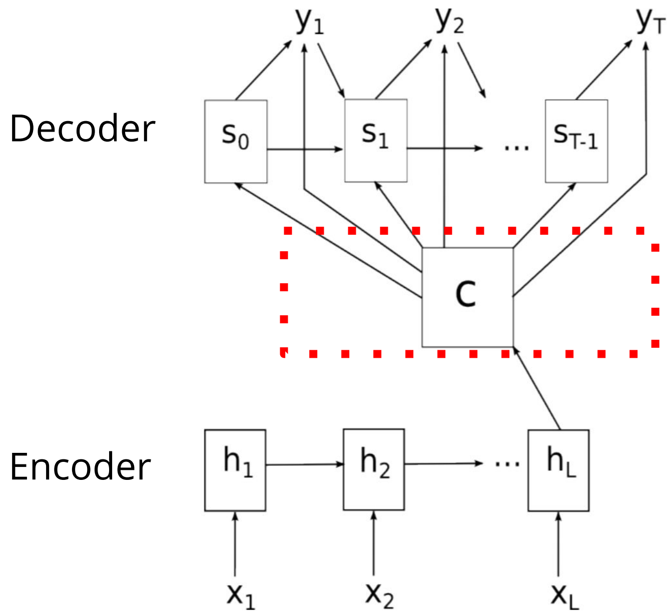
Neural Machine Translation By Jointly Learning To Align And Translate

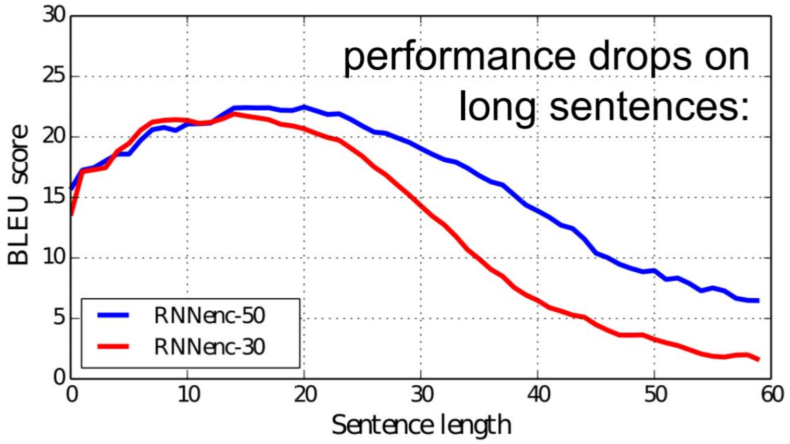
Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio

Motivation

Existing encoder-decoder frameworks



Fixed-length representation is a potential bottleneck for long sentences

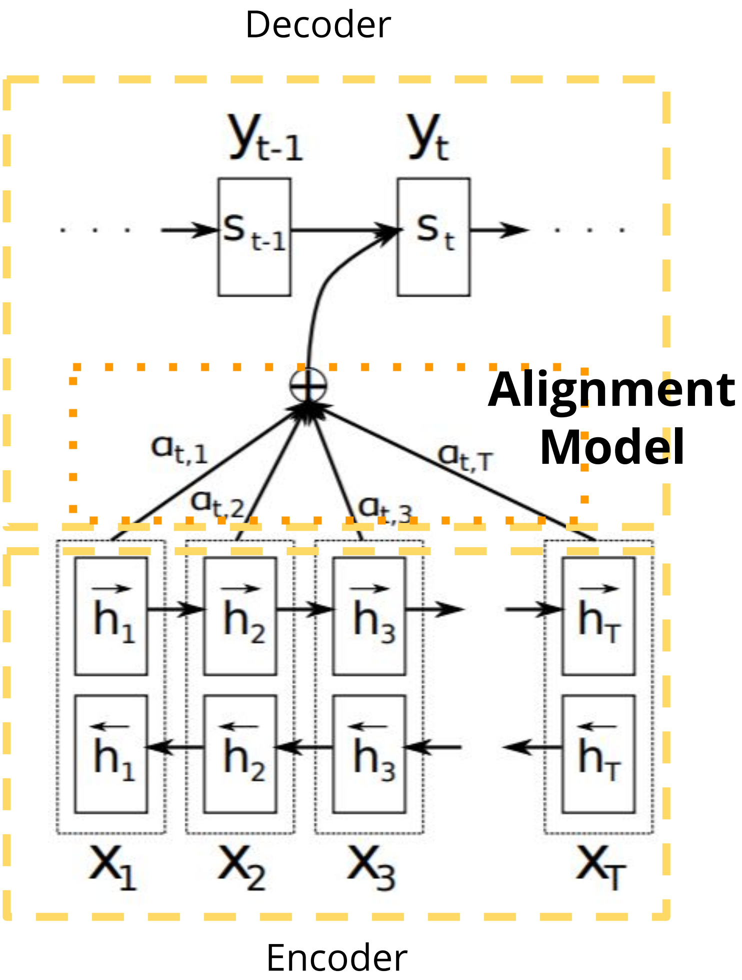


Evaluated on news-test-2014 from
WMT ’14 English-French parallel corpora

Method

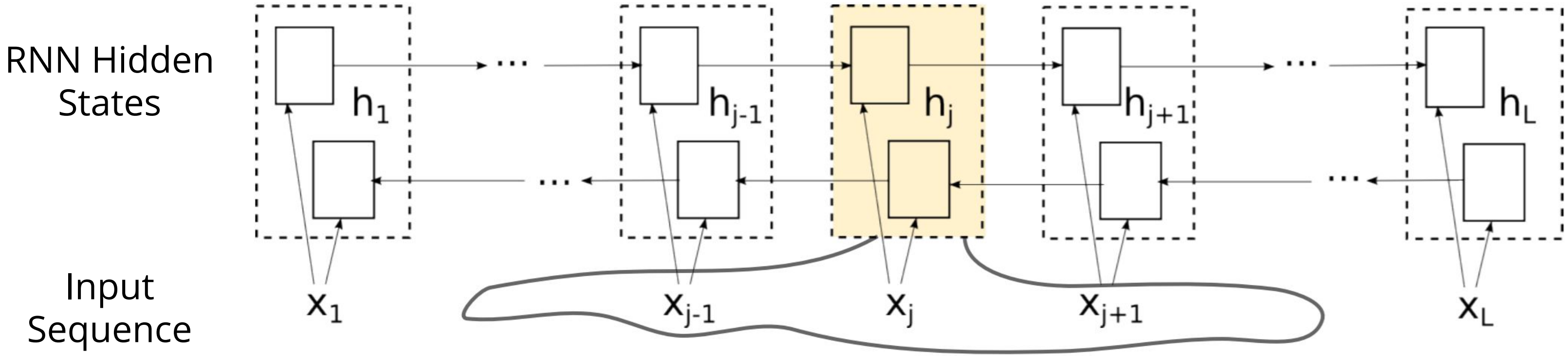
For each generated word y_t in the translation, soft-searches for a set of positions $(1, \ldots, T)$ in a source sentence $\mathbf{x}=(x_1, \ldots, x_T)$ where the most relevant information is concentrated.

The predicted target word y_t is based on the context vectors $\mathbf{c}_t=\sum\alpha_{t,j}\,\boldsymbol{h}_j$ associated with these source positions and all the previous generated target words s_{t-1}, y_{t-1}.



Encoder - Bidirectional RNN

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



$$\mathbf{x} = (x_1, \cdots, x_{T_x}) \xrightarrow{h_t = f(x_t, h_{t-1})} \begin{matrix} (\overrightarrow{h}_1, \cdots, \overrightarrow{h}_{T_x}) \\ (\overleftarrow{h}_1, \cdots, \overleftarrow{h}_{T_x}) \end{matrix} \xrightarrow{\text{Concat}} \boldsymbol{h} = (h_1, \cdots, h_{T_x})$$

where $h_j = \left[\overrightarrow{h}_j^\top ; \overleftarrow{h}_j^\top\right]^\top$

Decoder - Alignment Model

- Compute alignment α_{ij}

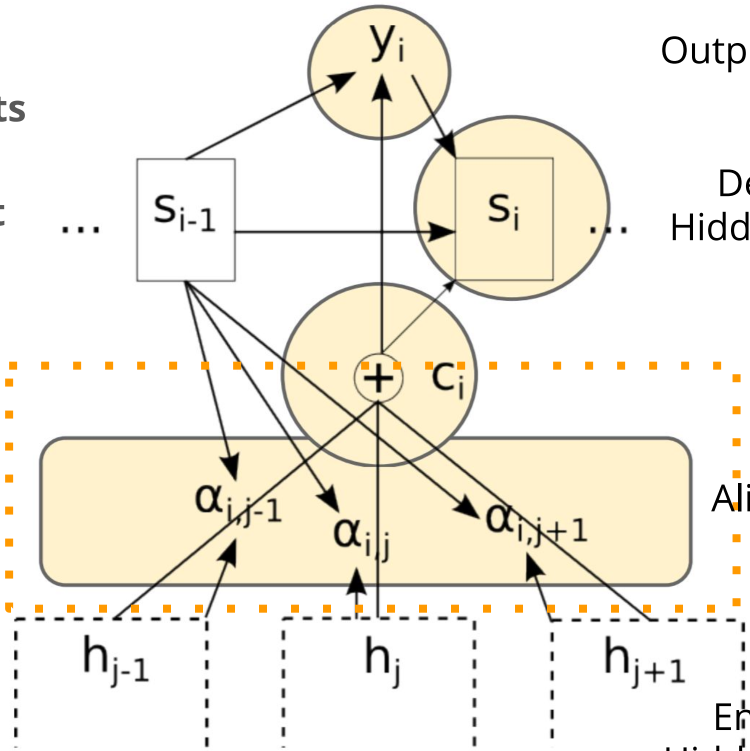
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

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

$$e_{ij} = v^T \tanh(W s_{i-1} + V h_j)$$

where v, W, V are trainable weights



Decoder

- Compute context c_i

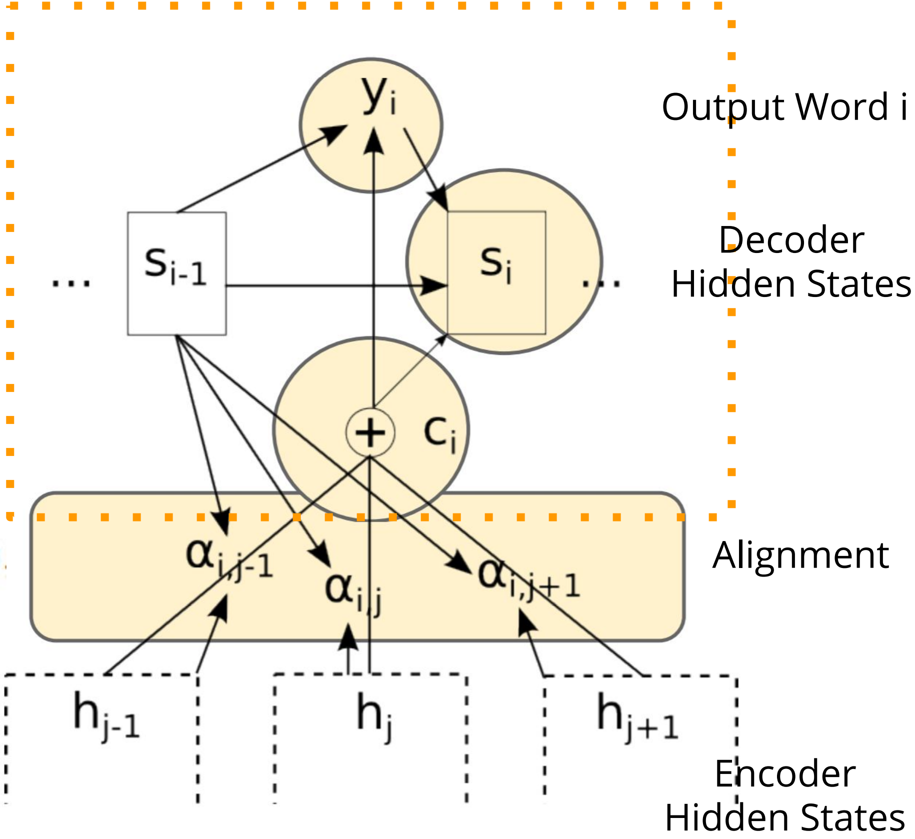
$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

- Compute new decoder state s_i

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

- Generate new output y_i

$$\text{argmax} \quad p(y_i | y_1, \ldots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i)$$



Evaluation

Models

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

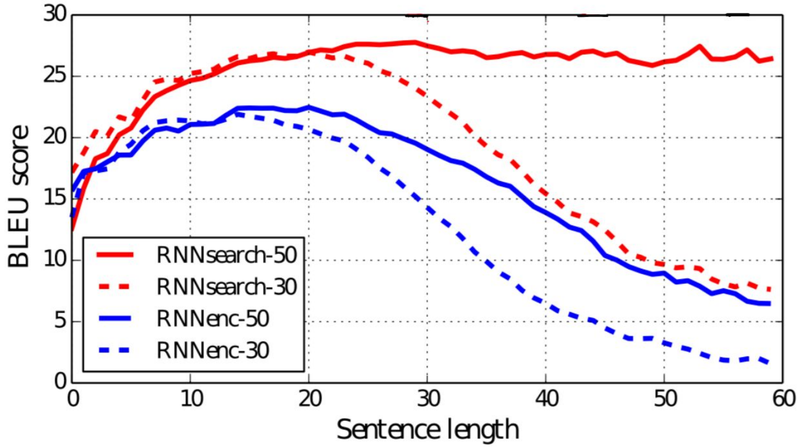
Dataset

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

Metrics

- BLEU score

Model	All	No UNK°
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50*	28.45	36.15
Moses	33.30	35.63



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

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In proceedings of the 2015 International Conference on Learning Representations, May 2015. https://arxiv.org/pdf/1409.0473.pdf. slides.
- Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder–Decoder approaches. In Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. 2014. https://arxiv.org/pdf/1409.1259.pdf.