Neural Machine Translation by Jointly Learning to Align and Translate

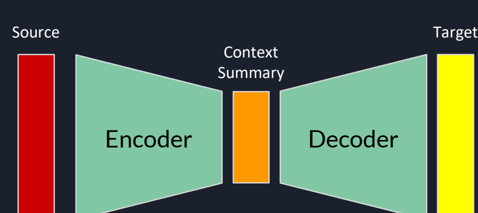
D. Bahdanau, K. Cho, Y. Bengio (**ICLR 2015**)

KAMRAN ALIPOUR

APRIL 2019

OUTLINE

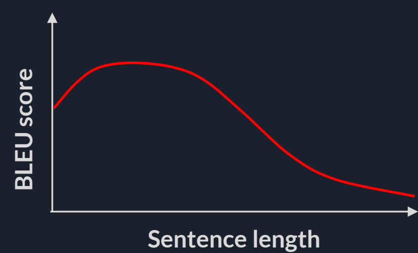
- **PROBLEM**:
  - Fixed-length vector representation is a bottleneck
  - Difficult to cope with long sentences,
    especially when longer than the sentences in the training corpus.



2

OUTLINE

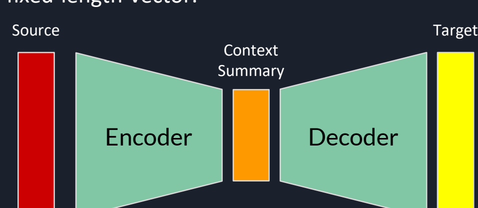
- Previous work shows performance of a basic encoder–decoder deteriorates rapidly as the length of an input sentence increases.



3

OUTLINE

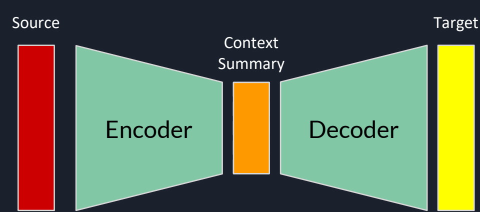
- **SOLUTION**:
  - An extension which learns to **align** and **translate** jointly.
  - Does NOT encode a whole input sentence into a single fixed-length vector.



4

OUTLINE

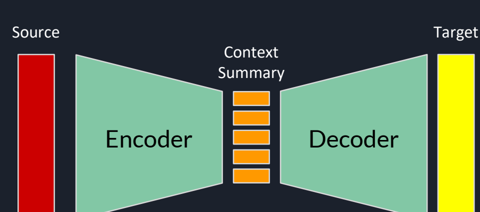
- **SOLUTION**:
  - Encode the input into a sequence of vectors and choose a subset of them adaptively while decoding.



5

OUTLINE

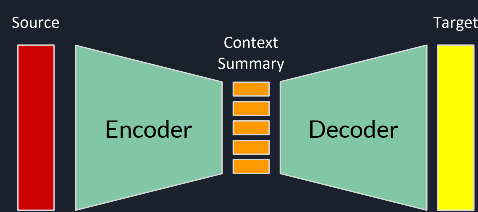
- **SOLUTION**:
  - For each translation, it (soft-)searches for a set of positions in a source sentence where the most relevant information is concentrated.



6

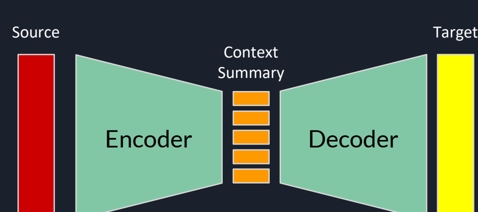
OUTLINE

- **SOLUTION**:
  - The model predicts a target word based on the context vectors associated with source positions and all the previously generated target words.



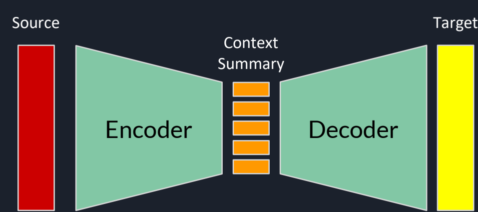
OUTLINE

- **SOLUTION**:
  - The model does not squash all the information of a source sentence, regardless of its length, into a fixed-length vector



OUTLINE

- **SOLUTION**:
  - State-of-the-art phrase-based system on the task of English-to-French translation.



9

PREVIOUS WORK

Phrase-based translation system
- Tuning sub-components separately (e.g. *Koehn et al. 2003*)

Neural machine translation
- Proposed by:
  - *Kalchbrenner and Blunsom (2013)*
  - *Cho et al. (2014)*
  - *Sutskever et al. (2014)*
- Mostly encoder-decoder architectures
- Encoders and decoders for each language
- Encoder-decoder jointly trained

10

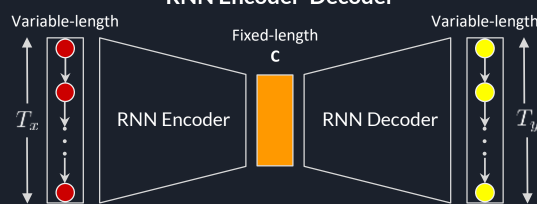
BACKGROUND

NEURAL MACHINE TRANSLATION

$$\arg\max_y p(\mathbf{y}|\mathbf{x})$$

Cho et al. (2014) & *Sutskever et al. (2014)*

RNN Encoder-Decoder



$$p(y_1, y_2, ..., y_{T_y} | x_1, x_2, ..., x_{T_x})$$

11

RNN ENCODER-DECODER

Introduced by: *Cho et al. (2014)*

$$\mathbf{x} = (x_1, ..., x_{T_x})$$

$$h_t = f(x_t, h_{t-1}) \qquad h_t \in \mathbb{R}^n$$

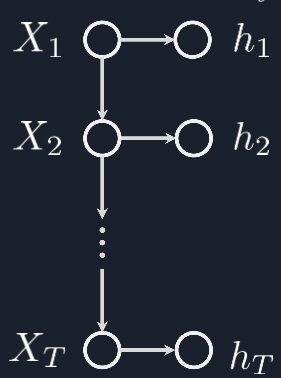
$$c = q(\{h_1, ..., h_{T_x}\})$$

12

RNN ENCODER-DECODER

Encoder:

$$h_t = f(x_t, h_{t-1})$$



$$c = q(\{h_1, \ldots, h_{T_x}\})$$

Sutskever et al. (2014):

LSTM as $f(x, h)$

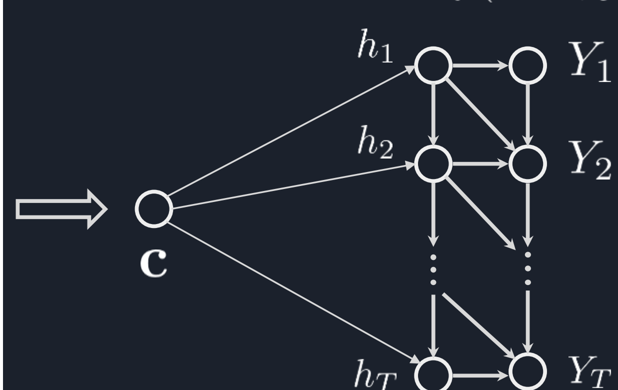
$$q(\{h_1, \ldots, h_T\}) = h_T$$

13

RNN ENCODER-DECODER

Decoder:

$$h_t = f(h_{t-1}, y_{t-1}, \mathbf{c})$$



$$\mathbf{y} = (y_1, \ldots, y_{T_y})$$

$$p(\mathbf{y}) = \prod_{t=1}^{T} p(y_t | \{y_1, \ldots, y_{t-1}\}, c)$$

$$p(y_t | \{y_1, \ldots, y_{t-1}\}, \mathbf{c}) = g(h_t, y_{t-1}, \mathbf{c})$$

14

RNN ENCODER-DECODER

Encoder-decoder jointly trained to maximize the conditional log-likelihood:

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log p_{\theta}(y_n | x_n)$$

15

RNN ENCODER-DECODER

- Better captures the linguistic regularities in the phrase table
- Indirectly explains the quantitative improvements in the overall translation performance
- Learns a continuous space representation of a phrase that preserves both the semantic and syntactic structure of the phrase.

16

ALIGN AND TRANSLATE

New Architecture

Encoder:

- Bidirectional (BiRNN, Schuster and Paliwal, 1997)
- Annotation of each word summarizes not only the preceding words, but also the following words.

17

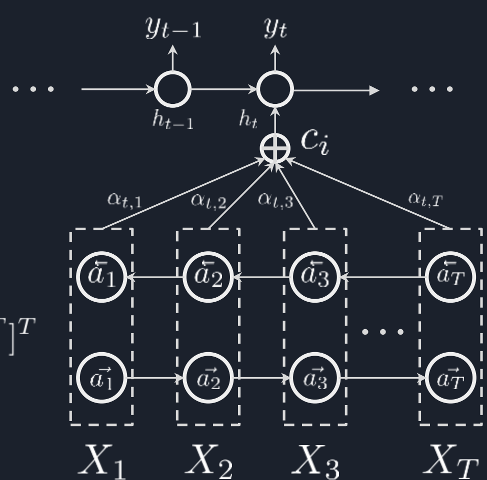
ENCODER: BIDIRECTIONAL RNN
FOR ANNOTATING SEQUENCES

BiRNN:

- Introduced by *Schuster and Paliwal, 1997*
- Also successfully used for speech recognition (*Graves et al., 2013*)
- Consists of a forward and a backward RNN

18

Align and Translate



$$a_j = [\vec{a_j}^T; \overleftarrow{a_j}^T]^T$$

$$h_i = f(h_{i-1}, y_{i-1}, c_i)$$

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

19

Align and Translate

Annotations: $(a_1, \ldots, a_{T_x})$

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

$$h_i = f(h_{i-1}, y_{i-1}, c_i)$$

$$p(y_i | y_1, \ldots, y_{i-1}, \mathbf{X}) = g(y_{i-1}, h_i, c_i)$$

20

Align and Translate

Decoder:

- Searching through source sentence while decoding a translation

Alignment Model

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

Alignment model:

$$e_{ij} = a(h_{i-1}, aj) = \nu_a^{\intercal} \tanh(W_a h_{j-1} + U_a a_j)$$

$$W_a \in \mathbb{R}^{n \times n} \qquad U_a \in \mathbb{R}^{n \times 2n} \qquad \nu_a \in \mathbb{R}^n$$

Alignment Model

Notes about alignment model :

- Parameterized a as a feedforward neural network
- Trained jointly with all the other components of the proposed model
- Alignment is NOT a latent variable
- Directly computes a soft alignment
- The gradient of the cost function can backpropagate through it

23

ALIGNMENT MODEL

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

"Expected Annotation"

α_{ij} **is the probability that the target word** y_i **is aligned to, or translated from, a source word** x_j

24

Alignment Model

Attention based decoder

- Decoder decides which parts of sentence to pay attention to.
- Encoder relieved from encoding all the information in a fixed-length vector.

25

Experiment Settings

Task: English to French translation

Bilingual, parallel corpora provided by ACL WMT '14

Results compared with original RNN Encoder-Decoder by *Cho et al. '14*

26

EXPERIMENT SETTINGS

Dataset: WMT '14

Europarl	61M words
News commentary	5.5M
UN	421M
Crawled corpora	90M + 272.5M
Total	850M

27

EXPERIMENT SETTINGS

Selected Data : 384M words

Using data selection method by *Axelrod et al. 2011*

28

EXPERIMENT SETTINGS

Training

Validation	News-test-2012 + news-test-2013
Test Set	News-test-2014 from WMT '14

29

EXPERIMENT SETTINGS

Training Models:
- RNN Encoder-Decoder
- RNNsearch

Each model trained twice:

RNNencdec-30	RNN Encoder-Decoder	30 words
RNNencdec-50	RNN Encoder-Decoder	50 words
RNNsearch-30	RNNsearch	30 words
RNNsearch-50	RNNsearch	50 words

30

Experiment Settings

- 1000 hidden units in both models
- A multilayer network with a single maxout hidden layer (Goodfellow *et al.* ‘13)
- Using a mini-batch stochastic gradient descent (SGD) algorithm together with Adadelta (Zeiler 2012)
- Training for each model: 5 days

31

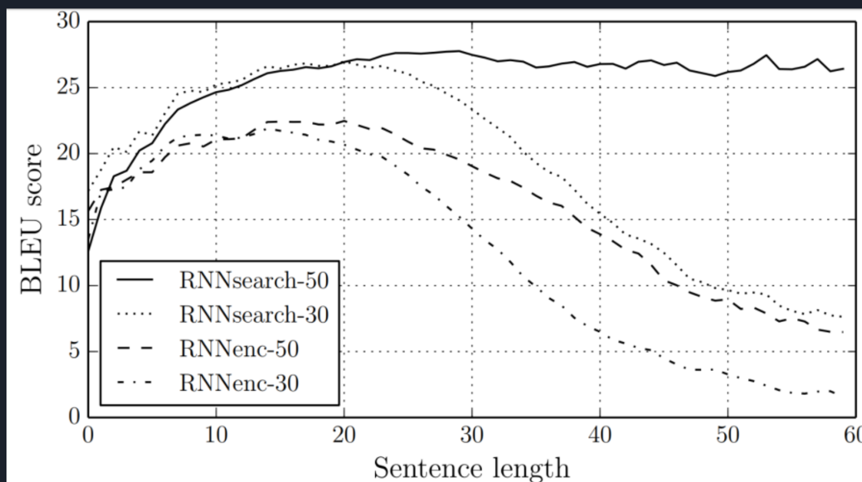
Results

BLEU Scores

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

32

RESULTS

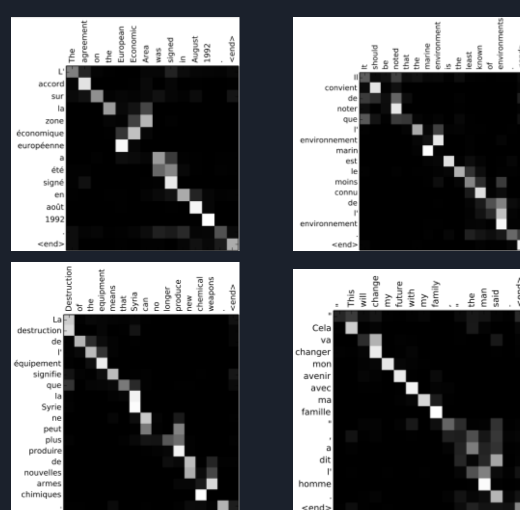


33

RESULTS

Strength of sof-alignments:

- Understands change of order
- Looks at close words for better translation
- Naturally deals with different lengths of source and target



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

34

Conclusion

- Use of a fixed-length context vector is problematic for translating long sentences
- Novel architecture that models a (soft-)search for a set of input words, or their annotations computed by an encoder, when generating each target word
- Better results on long sentences
- All components (including annotations) are jointly trained
- Performance comparable to the existing phrase-based statistical machine translation