Effective Approaches to Attention-based Neural Machine Translation - Luong et al.

M Saiful Bari

Research Assistant
School of Computer Science and Engineering
Nanyang Technological University

March 12, 2018

Outline

Intution

Attentional Mechanism

Selectively focusing on parts of the sentence.

Intution

Attentional Mechanism

Selectively focusing on parts of the sentence.

Find if a sentence is toxic or not.

Intution

Attentional Mechanism

Selectively focusing on parts of the sentence.

Find if a sentence is toxic or not.

I am going to shoot you. - toxic

Intution

Attentional Mechanism

Selectively focusing on parts of the sentence.

Find if a sentence is toxic or not.

I am going to shoot you. - toxic
I am going to love you. - not toxic

Outline

Abstract

This paper examines two simple and effective classes of attentional mechanism.

Abstract

This paper examines two simple and effective classes of attentional mechanism.

- Global attention (similar Bahdanau et al, 2015) but more simpler

Abstract

This paper examines two simple and effective classes of attentional mechanism.

- Global attention (similar Bahdanau et al, 2015) but more simpler
- Local attention (similar Xu et al, 2015) but differentiable

Abstract

This paper examines two simple and effective classes of attentional mechanism.

- Global attention (similar Bahdanau et al, 2015) but more simpler
- Local attention (similar Xu et al, 2015) but differentiable

+5.0 BLUE points compared to non-attentional system.

Abstract

This paper examines two simple and effective classes of attentional mechanism.

- Global attention (similar Bahdanau et al, 2015) but more simpler
- Local attention (similar Xu et al, 2015) but differentiable

+5.0 BLUE points compared to non-attentional system.

SOTA for WMT'15, English-German MT task. 25.9 BLEU (+1 increment).

SOTA - State Of The Art

Abstract

This paper examines two simple and effective classes of attentional mechanism.

- Global attention (similar Bahdanau et al, 2015) but more simpler
- Local attention (similar Xu et al, 2015) but differentiable

+5.0 BLUE points compared to non-attentional system.

SOTA for WMT'15, English-German MT task. 25.9 BLEU (+1 increment).

SOTA - State Of The Art

Cells are stacking LSTM.

Outline

Traditional NMT system

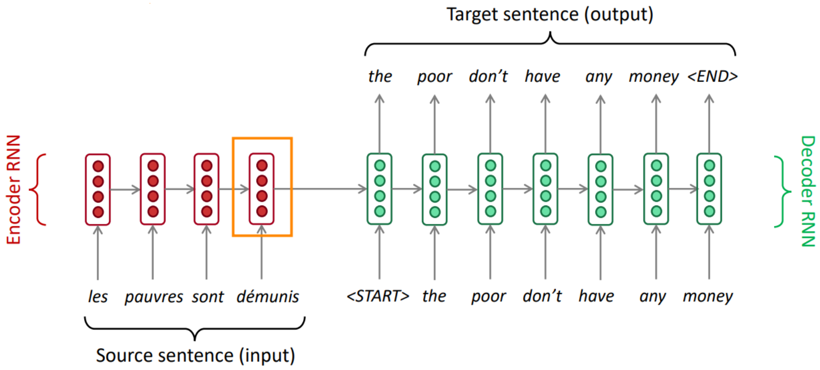


Figure: Neural machine translation a stacking recurrent architecture for translating a source sequence A B C D into a target sequence X Y Z. Here, $< eos >$ marks the end of a sentence.

Problem of Traditional NMT system

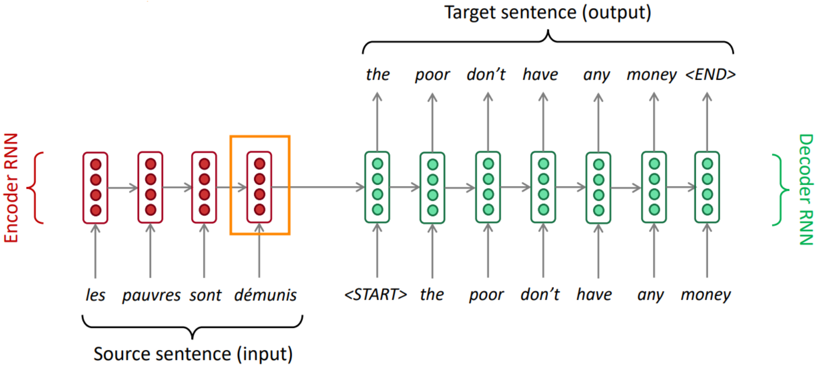


Figure: Orange rec-tangled hidden layer is the encoding of the source sentence. One single layer can't carry much info. (specially when the sentence is longer.)

Outline

Probabilistic Analysis

- **Encoder-Decoder** Model.
- a NN that directly models the **conditional probability** $p(y|x)$ of translating a source sequence, and $x_1, ..., x_n$ to a target sequence $y_1, ..., y_m$. If **s** is set of source hidden states,

$$x = les\ pauvres\ sont\ de'munis$$
$$y = the\ poor\ dont\ have\ any\ money$$

$$p(y|x) = \sum_{j=1}^{m} \log p(y_j|y_{<j}, \mathbf{s})$$

$$p(y_j|y_{<j}, \mathbf{s}) = softmax(g(h_j))$$

$$h_j = f(h_{j-1}, \mathbf{s})$$

$$J_t = \sum_{(x,y)\in\mathcal{D}} -\log p(y|x)$$

Types of Attention

$$p(y|x) = \sum_{j=1}^{m} \log p(y_j | y_{<j}, \mathbf{s})$$

- Global attention: **s** contains all the hidden layers of source.
- Local attention: **s** contains subset of hidden layers of source.

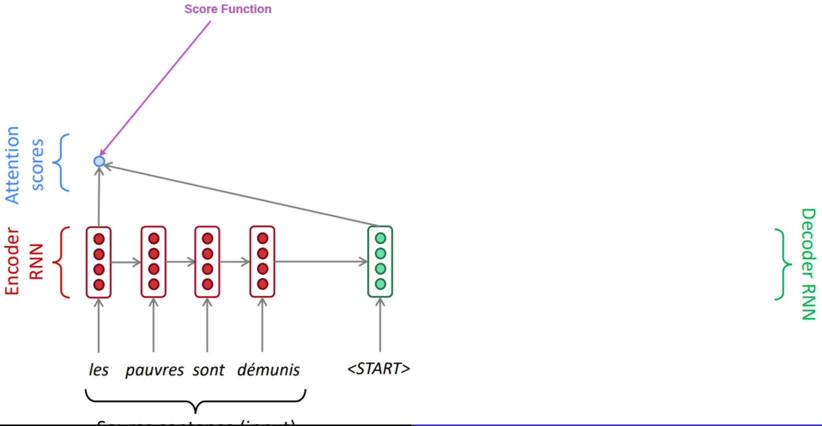
Types of Attention

$$p(y|x) = \sum_{j=1}^{m} \log p(y_i | y_{<j}, \mathbf{s})$$

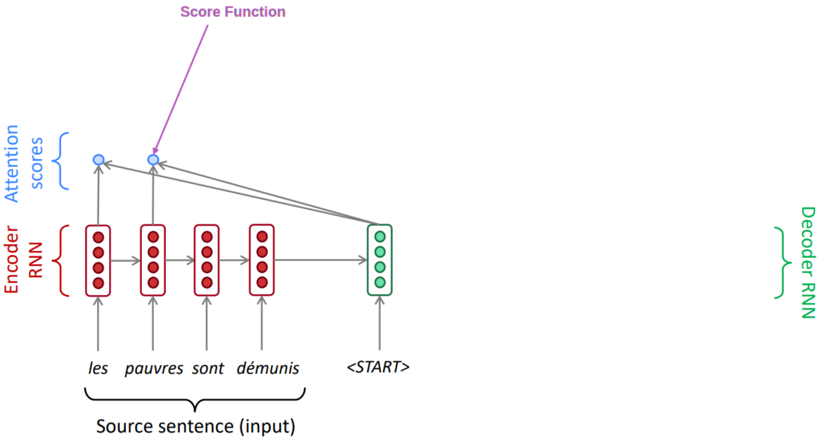
- Global attention: **s** contains all the hidden layers of source.
- Local attention: **s** contains subset of hidden layers of source.

Focus on a particular part of the source sequence.

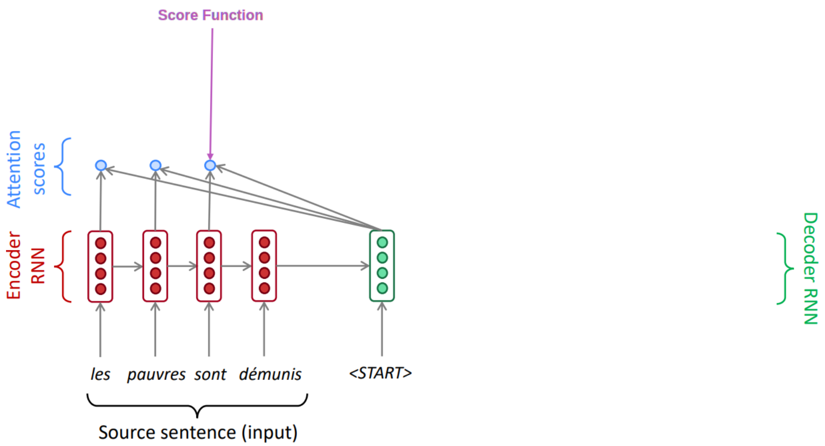
Global attention



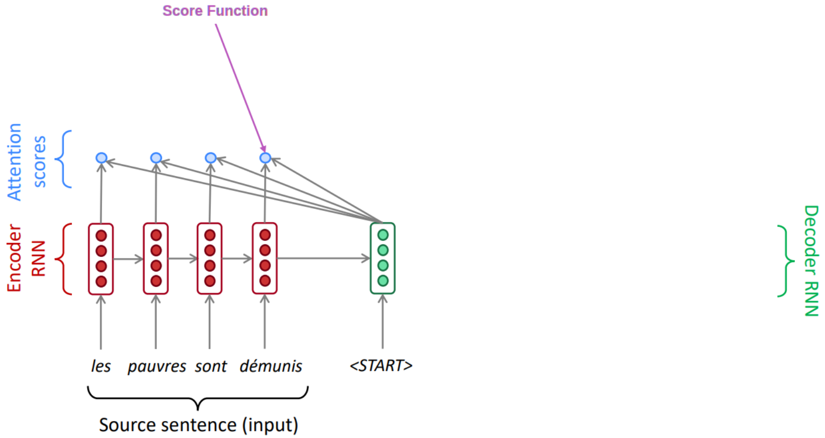
Global attention



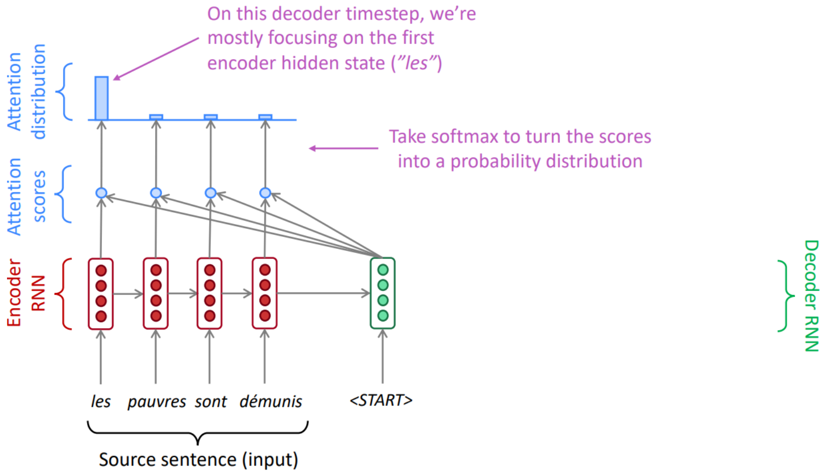
Global attention



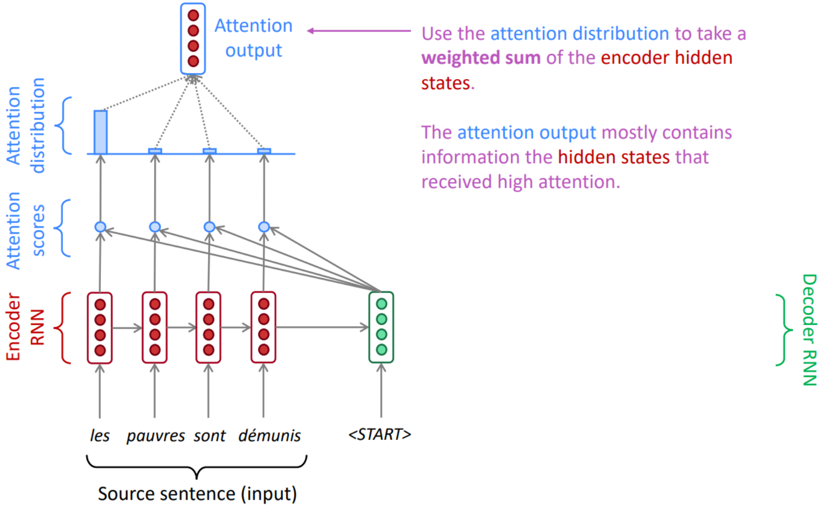
Global attention



Global attention



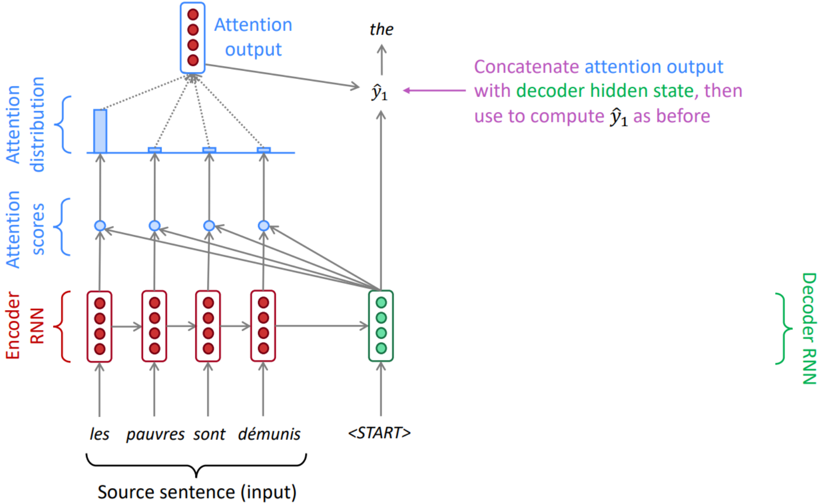
Global attention



Use the attention distribution to take a **weighted sum** of the **encoder hidden states**.

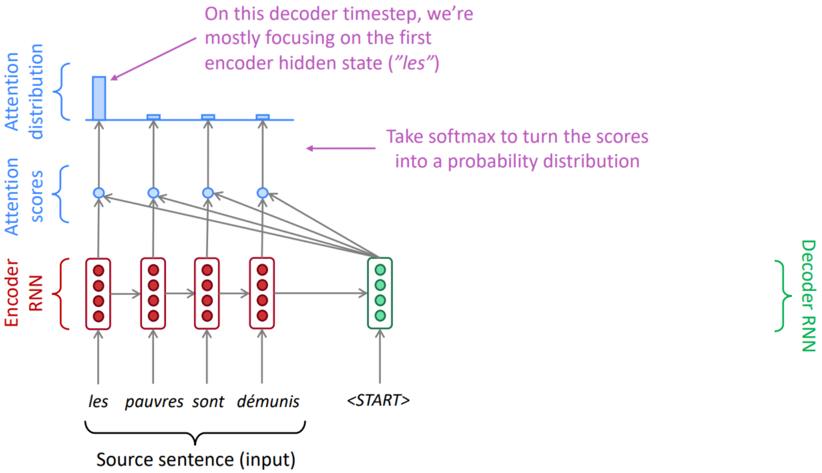
The attention output mostly contains information the **hidden states** that received high attention.

Global attention

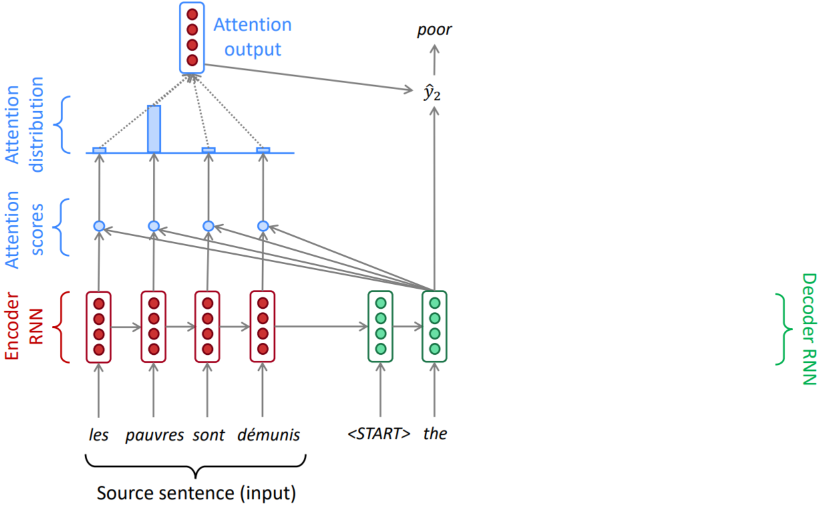


Concatenate attention output with decoder hidden state, then use to compute $\hat{y}_1$ as before

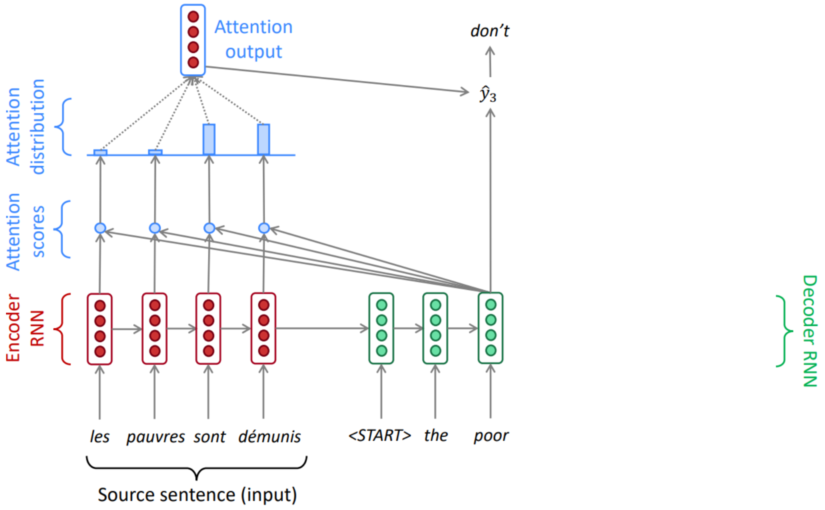
Global attention



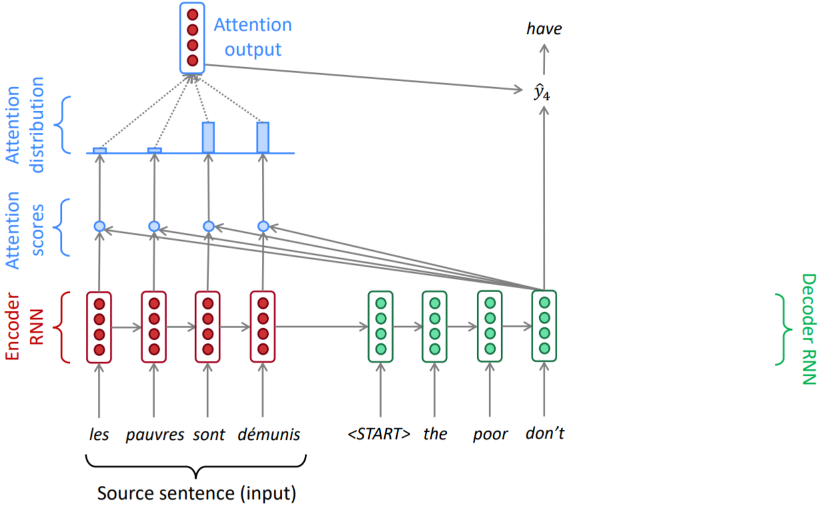
Global attention



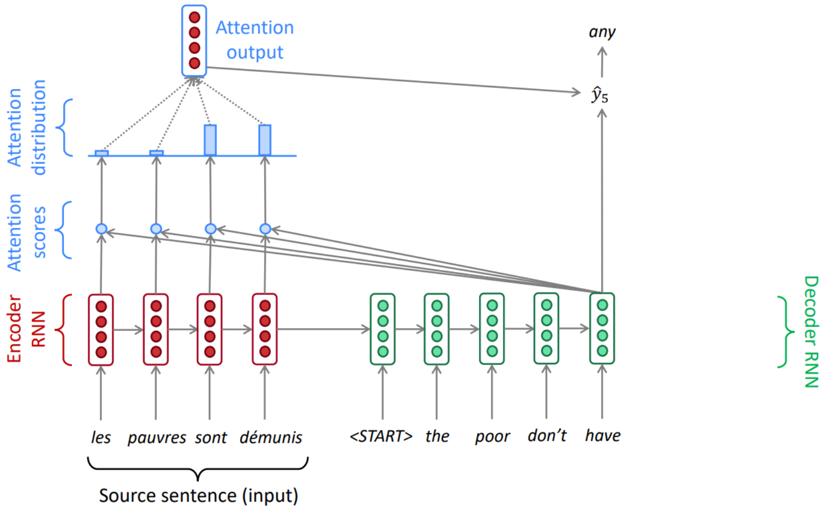
Global attention



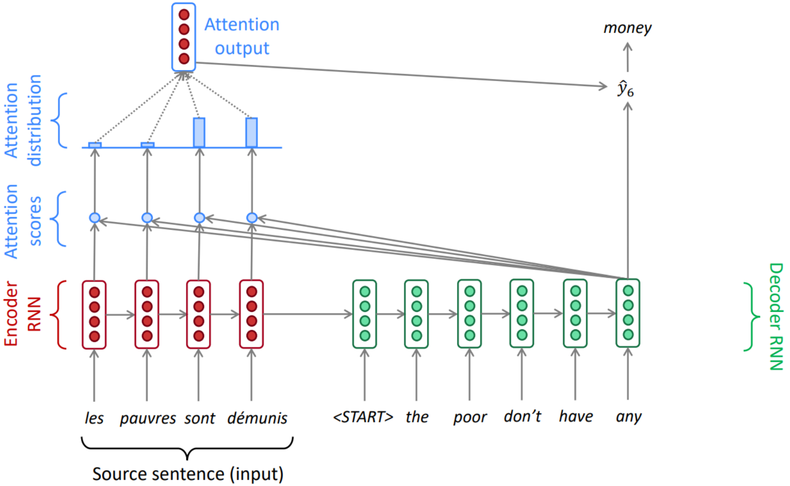
Global attention



Global attention



Global attention



Global attention: Algorithm

- Encoder (h_{et} - e for encoder, t is the time stamp)
  - Calculate the encode state, $H_e = h_{e1}, h_{e2}, ...h_{eN}$
- Decoder (h_{dt} - d for decoder, t is the time stamp)
  - for each time step t calculate score compared to H_e,

$$\begin{aligned} a_t(s) &= align(h_{dt}, h_{es}) \\ &= \frac{exp(score(h_{dt}, h_{es})}{\sum_{s'} exp(score(h_{dt}, h_{es}))} \end{aligned}$$

(s is the sequence length)

Global attention: Algorithm

- Decoder
  - for each time step t calculate score compared to H_e,
    $a_t(s) = align(h_{dt}, h_{es}) = \frac{exp(score(h_{dt},h_{es})}{\sum_{s'} exp(h_{dt},h_{es})}$ (s is the sequence length)
  - $$score(h_{dt}, h_{es}) = \begin{cases} h_{dt}^T h_{es} & dot \\ h_{dt}^T W_a h_{es} & general \\ v_a^T \tanh(W_a[h_{dt}^T; h_{es}]) & concat \end{cases}$$
  - $$c_t = \sum_{i=1}^{s} a_t(i) h_{ei}$$
  - $h_{dout} = \tanh(W_c[c_t; h_{dt}])$
  - $p(y_t|y_{<t}, x) = softmax(W_s h_{dout})$

Global attention

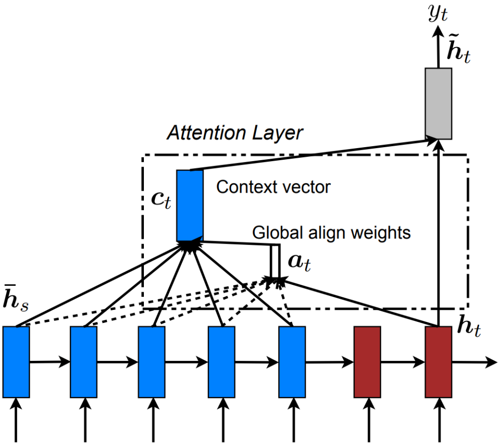


Figure: Global attention figure from the paper.

Local Attention

- For each decoder step t,
  the model first generates an aligned position p_t.
- Calculate the attention in window $[p_{t-D}, p_{t+D}]$
- D is empirically selected.
- Unlike the global approach, the local alignment vector a_t is now fixed-dimensional.

Local attention

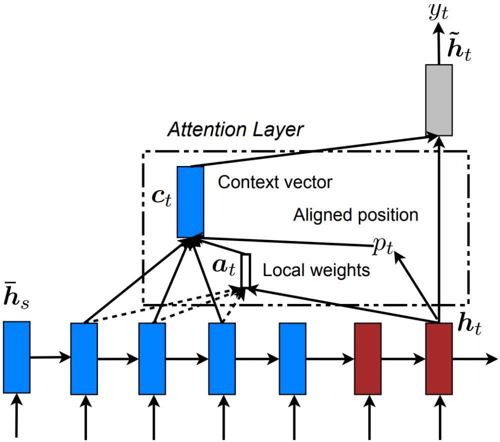


Figure: Local attention figure from the paper.

p_t selection

To calculate the value of p_t the paper consider 2 approaches,

- **Monotonic alignment (local-m) :** Simply set $p_t = t$ assuming that source and target sequences are roughly monotonically aligned
- **Predictive alignment (local-p) :** use a NN to predict the position,

$$p_t = S * sigmoid(v_p tanh(W_p h_{dt}))$$

$$a_t(s) = align(h_{et}, h_{ds}) exp(-\frac{(s - p_t)^2}{2\sigma^2})$$

where, $\sigma = D/2.p_t$ is a real number. s is an integer within the window centered at p_t

Outline

Input feeding Method

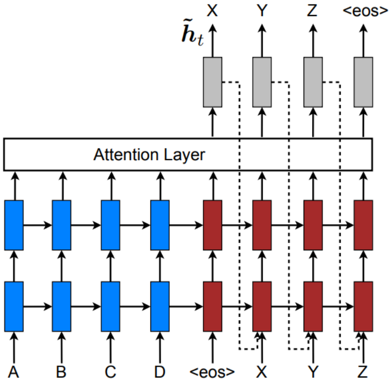


Figure: Attentional vectors $\tilde{h}_t$(decoder side) are fed as inputs to the next time steps to inform the model about past alignment decisions

Experiment

System	Ppl	BLEU
Winning WMT'14 system – *phrase-based* + *large LM* (Buck et al., 2014)		20.7
Existing NMT systems		
RNNsearch (Jean et al., 2015)		16.5
RNNsearch + unk replace (Jean et al., 2015)		19.0
RNNsearch + unk replace + large vocab + *ensemble* 8 models (Jean et al., 2015)		**21.6**
Our NMT systems		
Base	10.6	11.3
Base + reverse	9.9	12.6 (+*1.3*)
Base + reverse + dropout	8.1	14.0 (+*1.4*)
Base + reverse + dropout + global attention (*location*)	7.3	16.8 (+*2.8*)
Base + reverse + dropout + global attention (*location*) + feed input	6.4	18.1 (+*1.3*)
Base + reverse + dropout + local-p attention (*general*) + feed input	5.9	19.0 (+*0.9*)
Base + reverse + dropout + local-p attention (*general*) + feed input + unk replace		20.9 (+*1.9*)
Ensemble 8 models + unk replace		**23.0** (+*2.1*)

Figure: WMT14 English-German results

Length Analysis

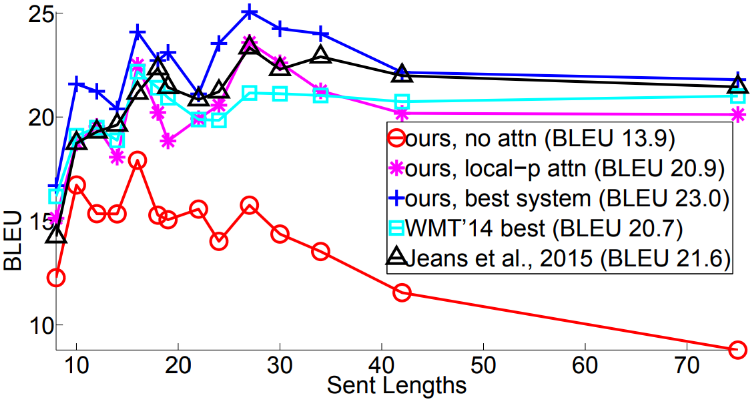


Figure: Length Analysis translation qualities of different systems as sentences become longer.

Summary

Summary

- Attention significantly improves NMT performance
  - Its very useful to allow decoder to focus on certain parts of the source. (local attention seem better than others)

Summary

- Attention significantly improves NMT performance
  - Its very useful to allow decoder to focus on certain parts of the source. (local attention seem better than others)
- Attention solves the bottleneck problem
  - Attention allows decoder to look directly at source; bypass bottleneck

Summary

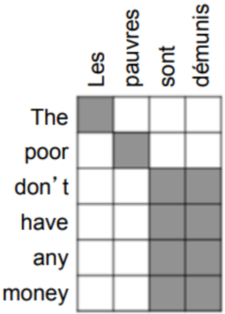
- Attention significantly improves NMT performance
  - Its very useful to allow decoder to focus on certain parts of the source. (local attention seem better than others)
- Attention solves the bottleneck problem
  - Attention allows decoder to look directly at source; bypass bottleneck

- Attention helps with vanishing gradient problem
  - Provides shortcut to faraway states

Summary

- Attention significantly improves NMT performance
  - Its very useful to allow decoder to focus on certain parts of the source. (local attention seem better than others)
- Attention solves the bottleneck problem
  - Attention allows decoder to look directly at source; bypass bottleneck

- Attention helps with vanishing gradient problem
  - Provides shortcut to faraway states
- Attention provides some interpretability.



Credits

- Effective Approaches to Attention-based Neural Machine Translation Luong et al.
- Dr. Shafiq Joty.
- Stanford CS224n Slides.
- Christopher D. Manning's Lecture.