Effective Approaches to Attention-based Neural Machine Translation

Minh-Thang Luong **Hieu Pham** **Christopher D. Manning**
Computer Science Department Stanford University, Stanford, CA, 94305
{lmthang,hyhieu,manning}@stanford.edu

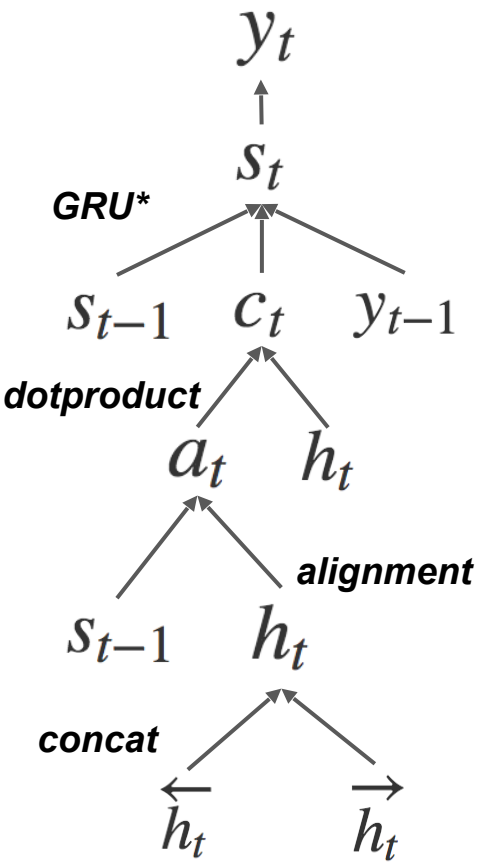
http://aclweb.org/anthology/D15-1166

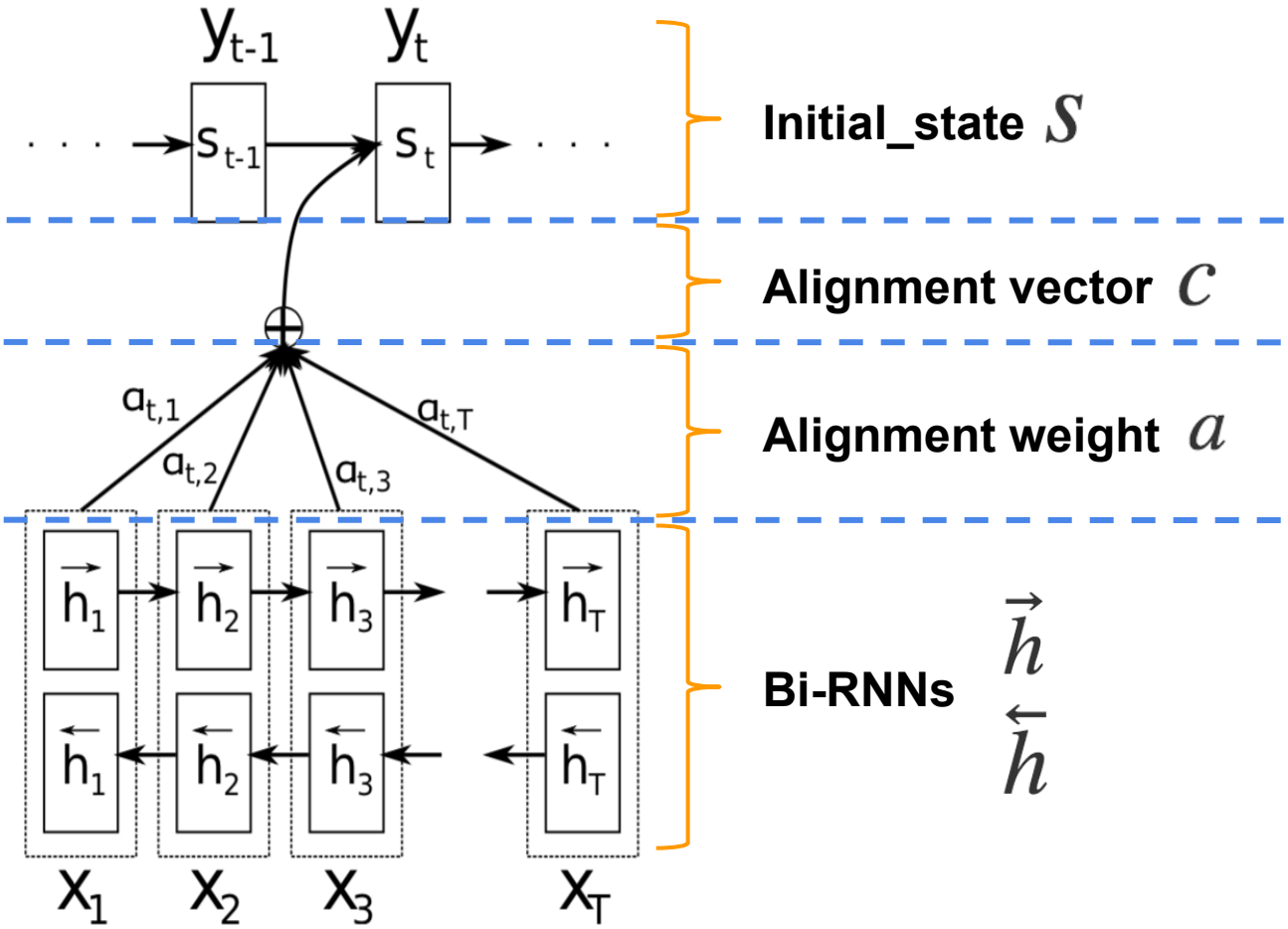
Kang Wen Wei

- **Bahdanau 2015**
- Purpose
- Attention mechanism
- Datasets
- Evaluate

Bahdanau 2015

https://arxiv.org/abs/1409.0473





Purpose

1. **Global attention**：

Consider *all the hidden-states of the encoder* when deriving the context vector c_t

- **New aligned metrics (content based function)**

1. **Local attention**：

Choose to *focus on a subset of hidden-states of the encoder* per target word

- **Local-m (Monotonic alignment)**
- **Local-p (Predictive alignment)**

1. **Input-feeding Approach**：

Choose to focus on a subset of hidden-states of the encoder per target word

Mechanism-Global attention

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_a \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{W}_a [\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s] & \textit{concat} \end{cases}$$

dot : $\vec{a} \cdot \vec{b} = |\vec{a}|\,|\vec{b}| \cos\theta$

$\because$ $\vec{a}$ and $\vec{b}$ are same direction, $\theta = 0$

$\therefore$ $\cos\theta = 1$, score will be more large

general : linear transform

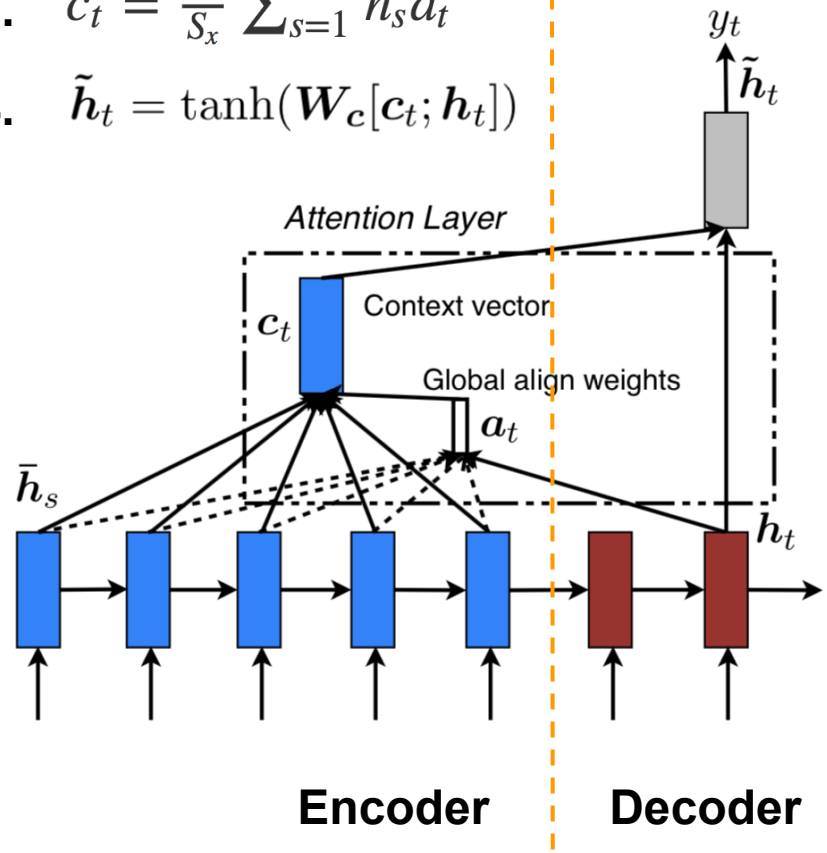
2. $\boldsymbol{a}_t(s) = \text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)$

$$= \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)}$$

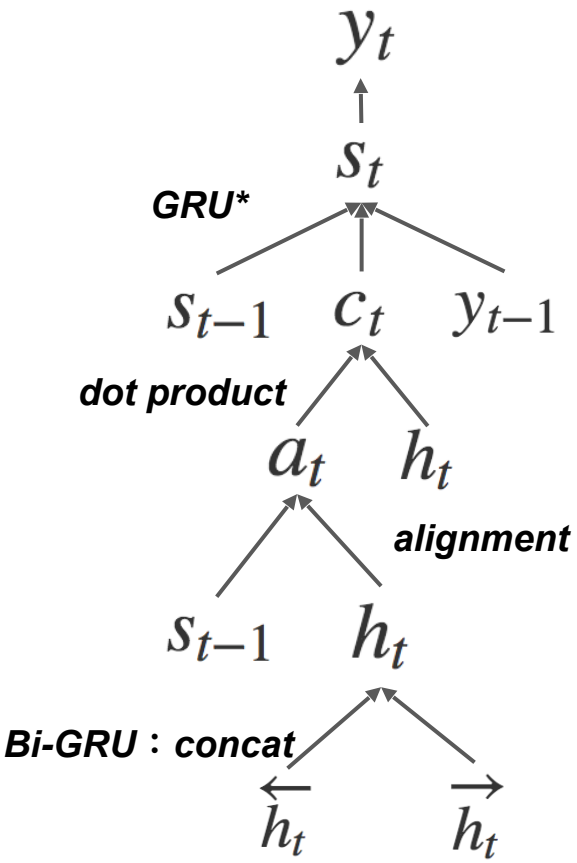
* $\boldsymbol{a}_t = \text{softmax}(\boldsymbol{W}_a \boldsymbol{h}_t)$ *location*

3. $c_t = \frac{1}{S_x} \sum_{s=1}^{S_x} h_s a_t$

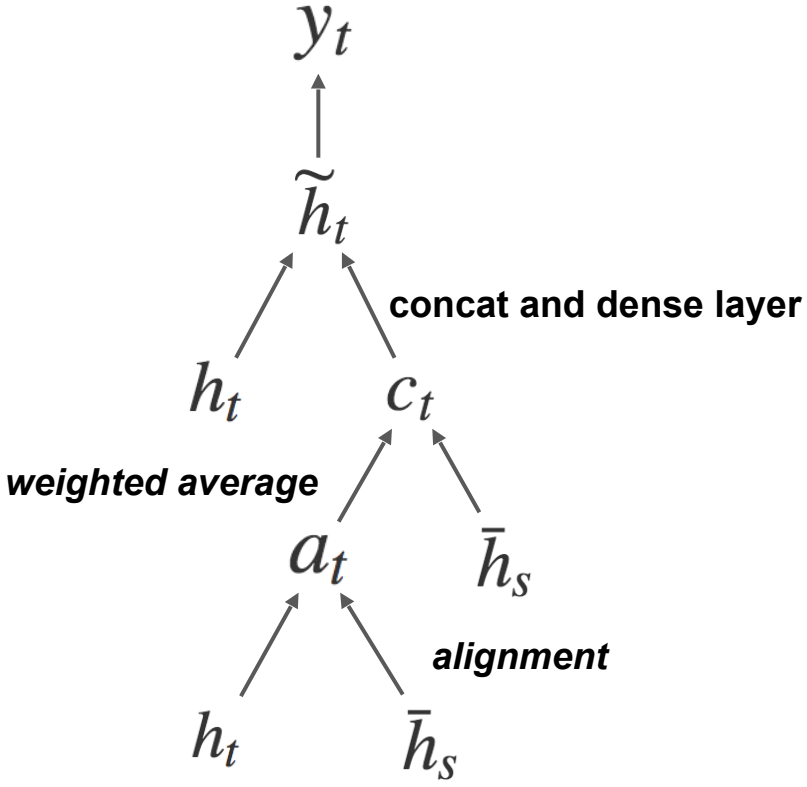
4. $\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_c [\boldsymbol{c}_t; \boldsymbol{h}_t])$



Comparison with Bahdanau 2015



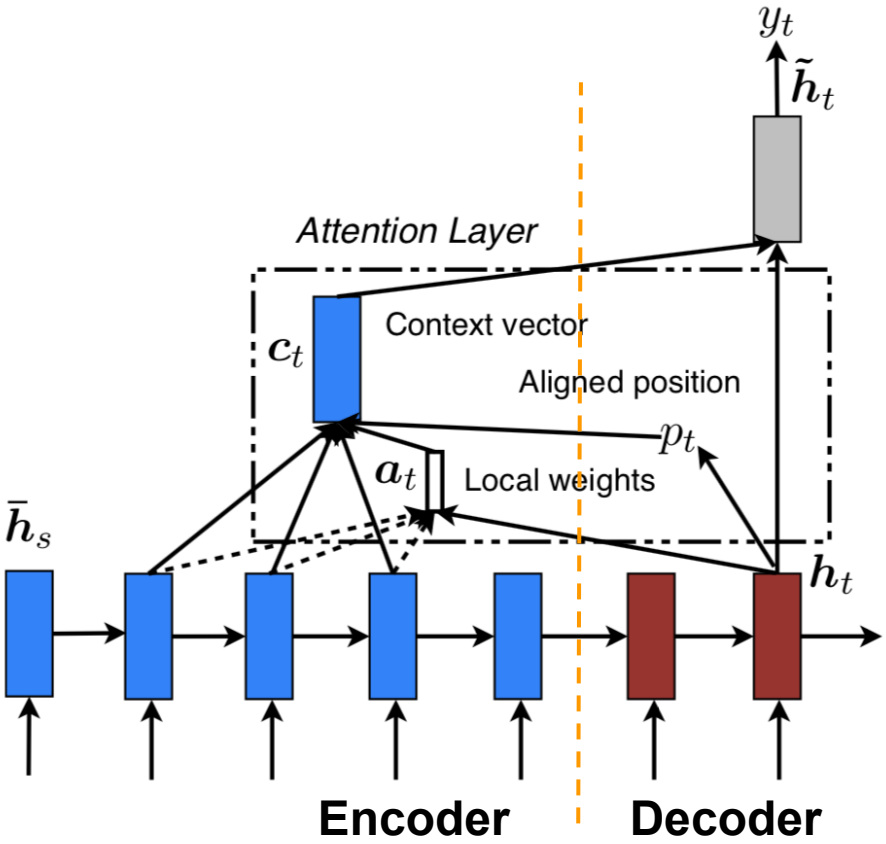
Bahdanau 2015



Global attention

Mechanism-Local attention

*Global attention attends to all words on the source side for each target word, which is expansive and can potentially render it impractical to **translate longer sequences**.*

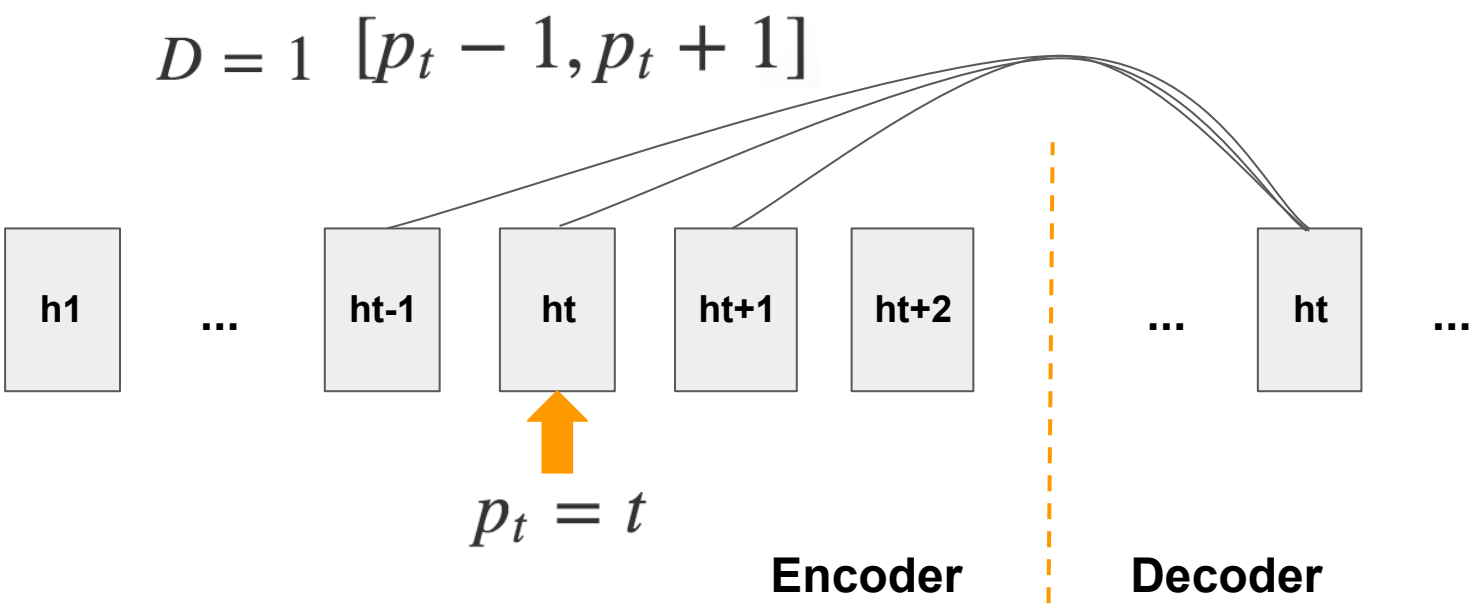


1. Generate an aligned position p_t and aligned boundary $[p_t - D, p_t + D]$
   D is empirically selected
   - **Monotonic alignment (local-m)**
   - **Predictive alignment (local-p)**
1. Compute aligned weight a_t

1. Compute context vector (weighted average) c_t

Monotonic alignment (local-m)

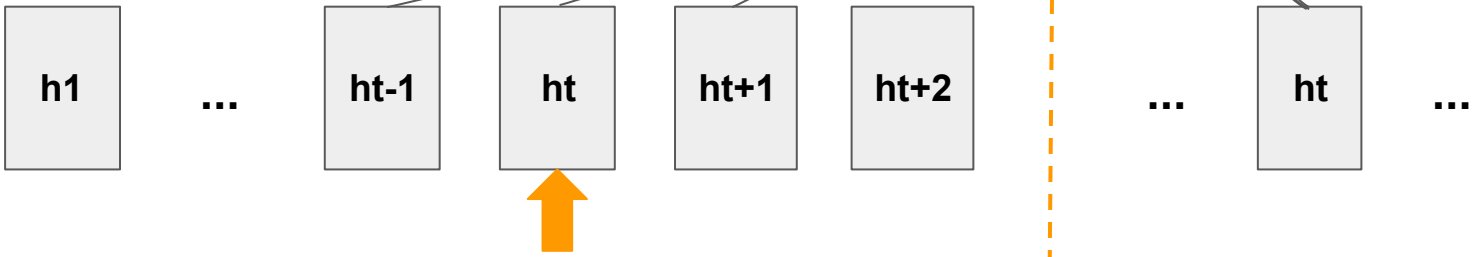
$$a_t(s) = \text{align}(h_t, \bar{h}_s) \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right)$$



Predictive alignment (local-p)

$$a_t(s) = \text{align}(h_t, \bar{h}_s) \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right)$$

$$D = 1 \quad [p_t - 1, p_t + 1]$$



$$p_t = S \cdot \text{sigmoid}(v_p^\top \tanh(W_p h_t))$$

S : *source sentence length*

v_p, W_p : weight matrix

Encoder **Decoder**

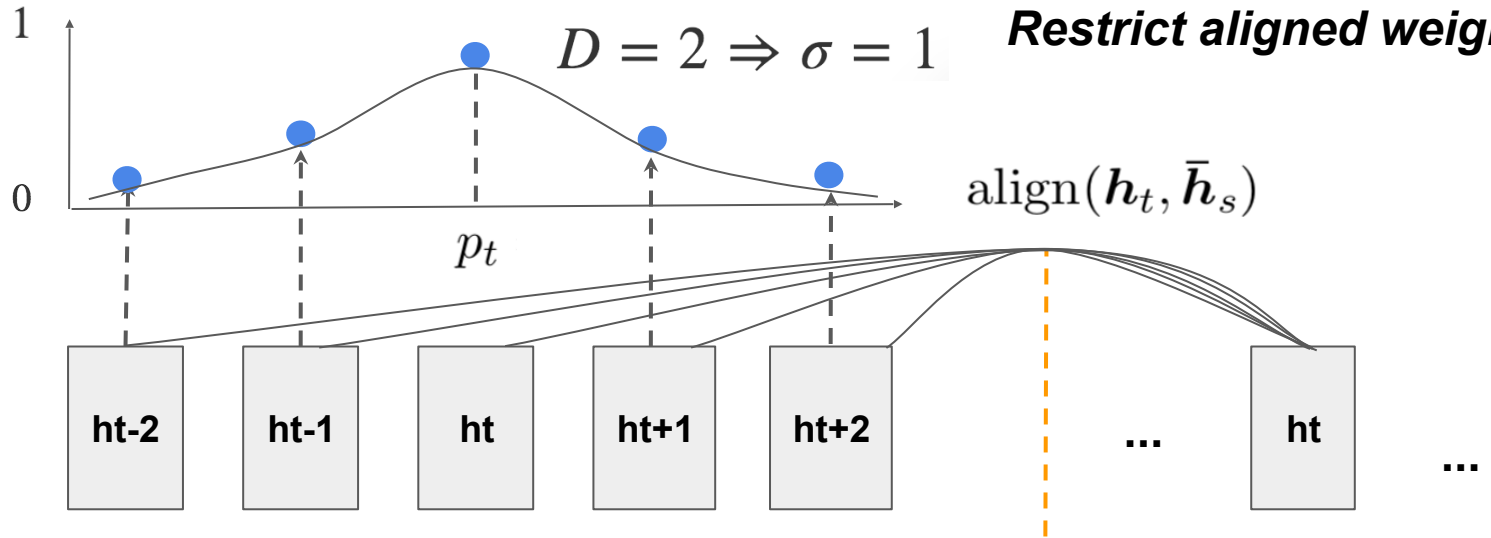
Monotonic alignment (local-p)

Normal distribution

$$a_t(s) = \text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) \exp\left(-\frac{(s-p_t)^2}{2\sigma^2}\right) \quad \Leftrightarrow \quad f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

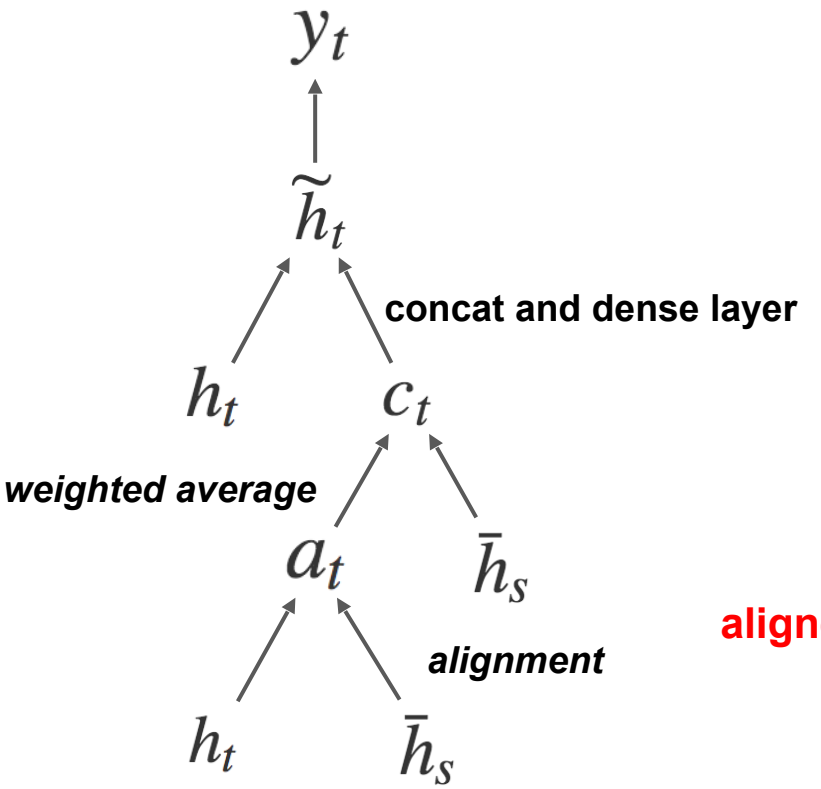
$$\frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)}$$

$$\begin{cases} s : \text{integer within the window centered at } p_t \\ \sigma = \frac{D}{2} \end{cases}$$

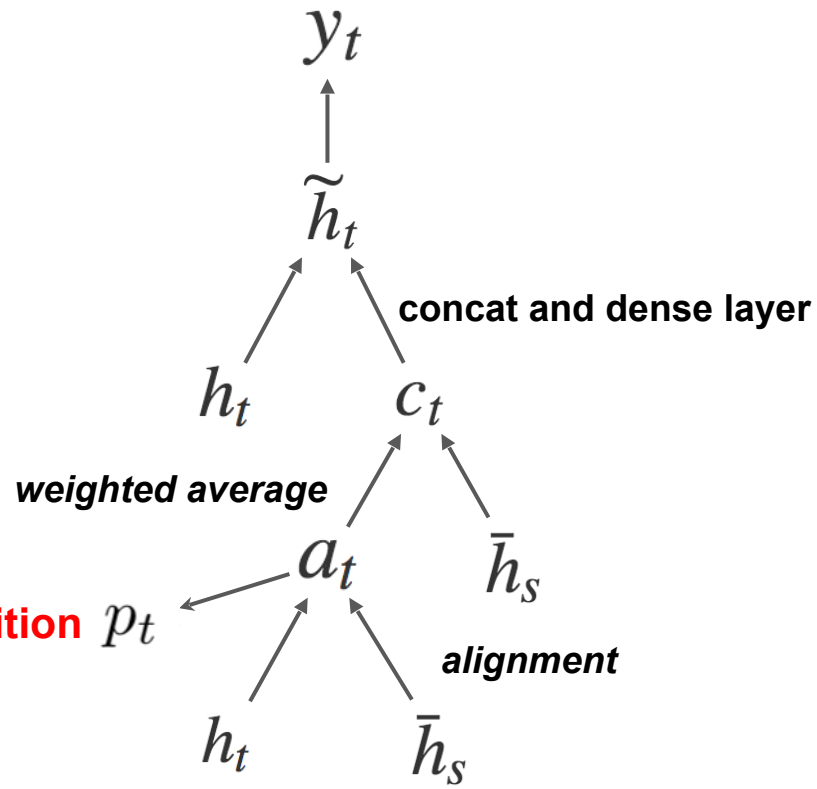


Restrict aligned weight a_t ***again !***

Monotonic alignment (local attention)



Global attention



Local attention

Attention - alignment pair

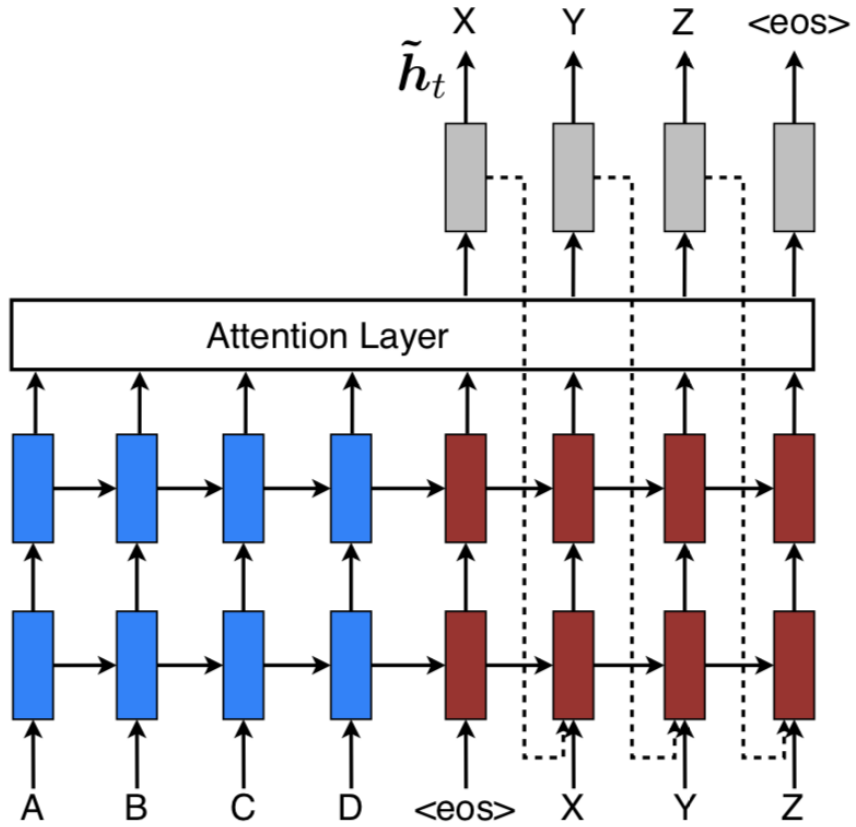
New aligned metrics

		location	dot	general	concat
Global Attention	global attention	global (location)	global (dot)	global (general)	
Local Attention	local-m		local-m (dot)	local-m (general)	
	local-p		local-p (dot)	local-p (general)	

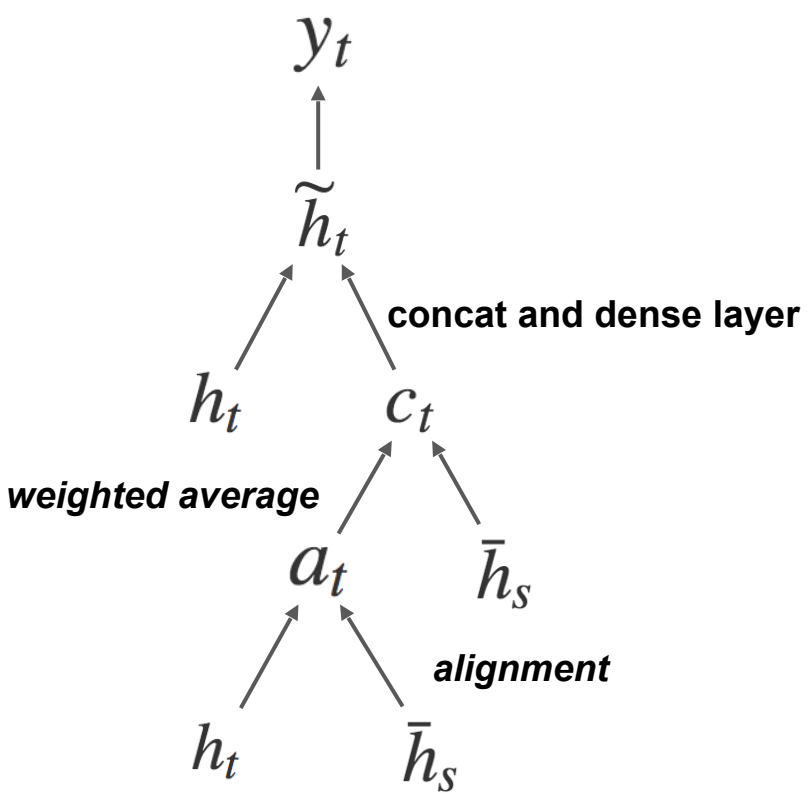
Input-feeding Approach

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

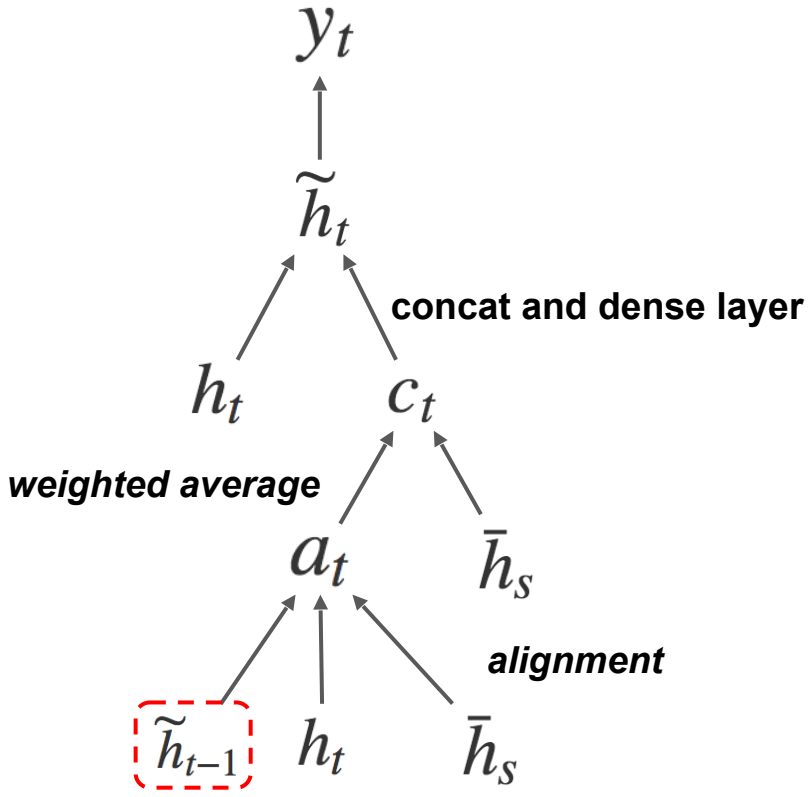
- Hope to make the model fully aware of previous alignment choices
- Create a very deep network spanning both horizontally and vertically



Input-feeding Approach



Global attention



Input-feeding

Datasets (English - German in both directions)

WMT'14

Training data：4.5M sentences pairs (116M English words, 110M German words)

Use Top 50k most frequent words for both languages, others replaced <unk>

Filter out sentence pairs whose lengths exceed 50 words

Validation：newstest-2013(3000sentences)

Test： newstest-2014(2737sentences), newstest-2015(2169sentences)

- 4 layers LSTM、1000 cells、1000 dimensional embedding
- Parameters are uniformly initialized in [-0.1, 0.1]
- Train for 10 epochs using SGD
- Start with learning rate of 1, after 5 epochs, begin halve the learning rate every epoch
- Batch_size: 128
- Normalized gradient rescaled whenever its' norm exceeds 5

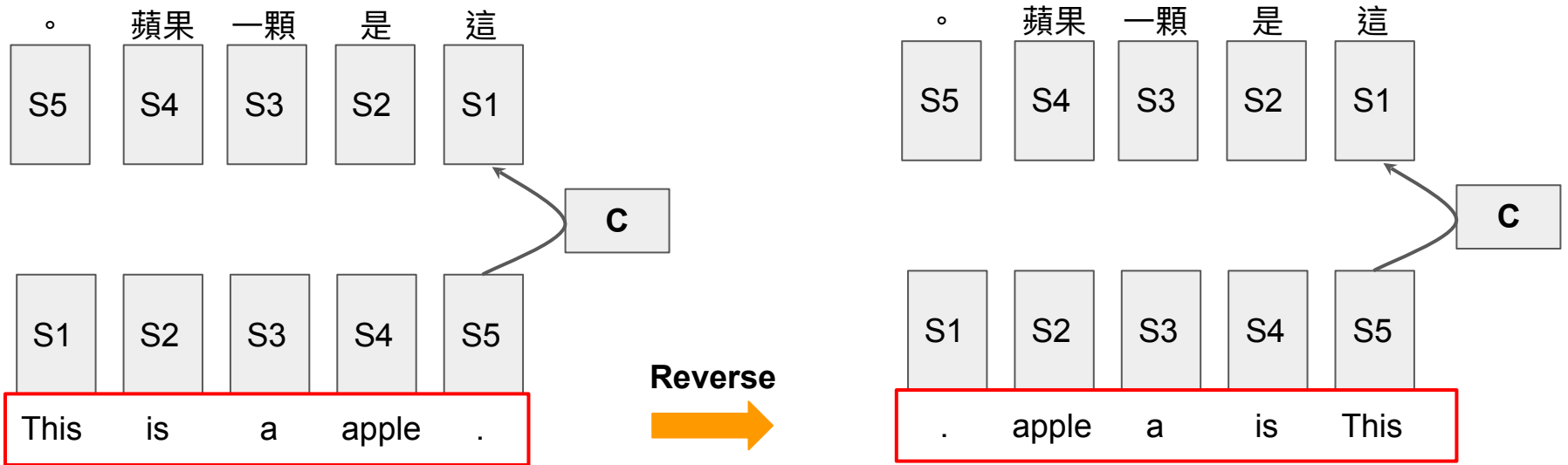
- For Dropout models, train for 12 epochs and start halving the learning rate after 8 epochs

Performance (WMT'14 English - German Result)

RNNsearch (Jean et al., 2015) : Proposed a method which solves the softmax of output layer.

unk replace (Jean et al., 2015) : Replacing each <unk> token with the aligned source word.

reverse : Reverse the words in the source sentence.



Performance (WMT'14 English - German Result)

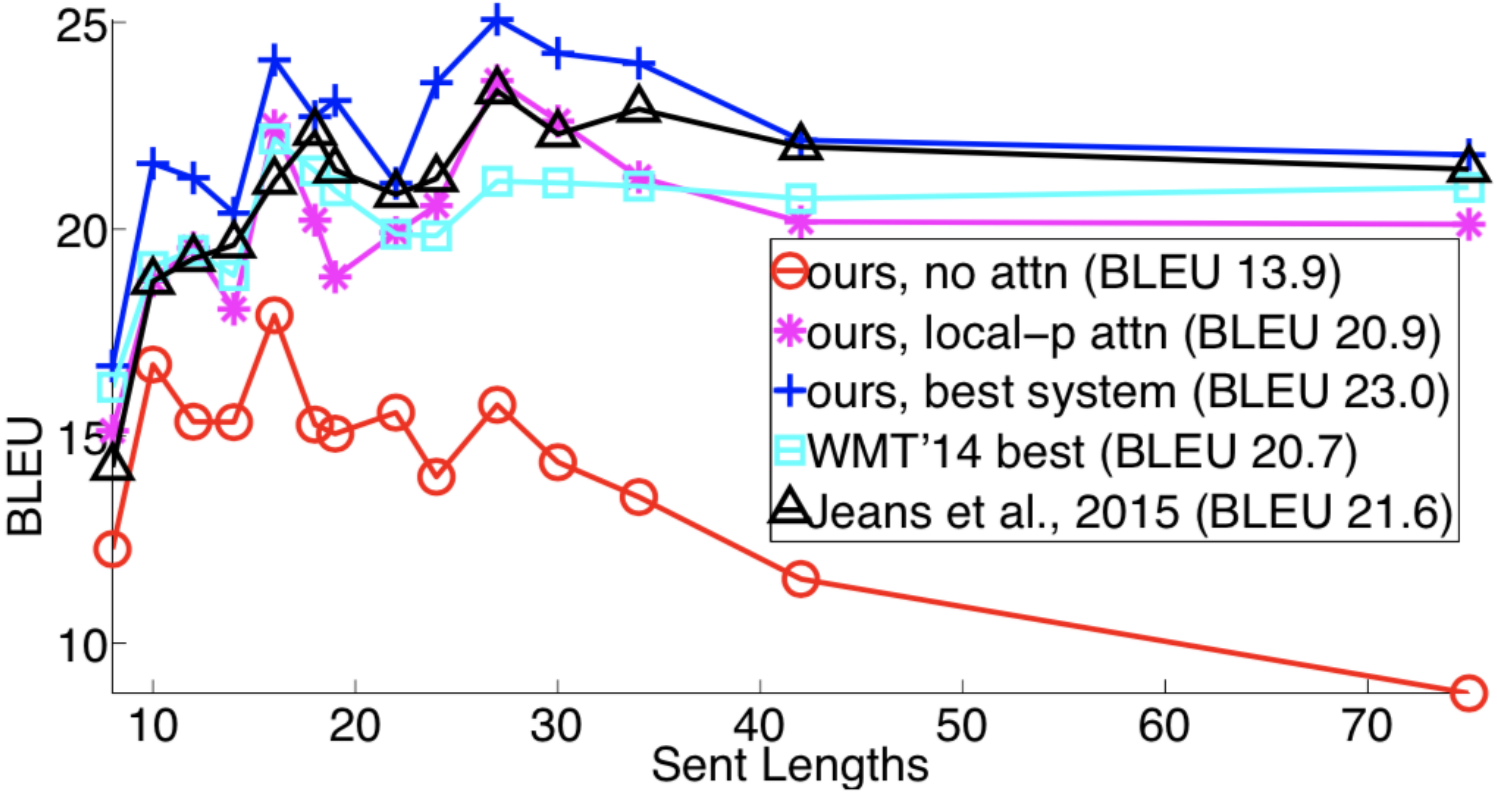
System	Ppl	BLEU
SOTA WMT'14 system – *phrase-based* + *large LM* (Buck et al., 2014)		20.7
Existing NMT systems		
RNNsearch (Jean et al., 2015)		16.5
RNNsearch + unk replace (Jean et al., 2015)		19.0
RNNsearch + unk replace + large vocab + *ensemble* 8 models (Jean et al., 2015)		**21.6**
Our NMT systems		
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*)

Using different attention approaches, with and without dropout

Performance (WMT'15 German - English Result)

System	Ppl.	BLEU
WMT'15 systems		
SOTA – *phrase-based* (Edinburg)		**29.2**
NMT + 5-gram rerank (MILA)		27.6
Our NMT systems		
Base (reverse)	14.3	16.9
+ global (*location*)	12.7	19.1 (+*2.2*)
+ global (*location*) + feed	10.9	20.1 (+*1.0*)
+ global (*dot*) + drop + feed	9.7	22.8 (+*2.7*)
+ global (*dot*) + drop + feed + unk		24.9 (+*2.1*)

Performance (Length Analysis)



Performance (Attention Architectures)

System	Ppl	BLEU	
		Before	After unk
global (location)	6.4	18.1	19.3 (+1.2)
global (dot)	6.1	18.6	20.5 (+1.9)
global (general)	6.1	17.3	19.1 (+1.8)
local-m (dot)	>7.0	x	x
local-m (general)	6.2	18.6	20.4 (+1.8)
local-p (dot)	6.6	18.0	19.6 (+1.9)
local-p (general)	**5.9**	**19**	**20.9 (+1.9)**