Effective approaches to attention-based neural machine translation

논문리뷰

이도연

목차

1 Introduction

2 NMT (Neural Machine Translation)

3 Attention-based Models

- Global Attention
- Local Attention
- Input-feeding Approach

4 Experiments & Analysis

5 Conclusion

1. Introduction

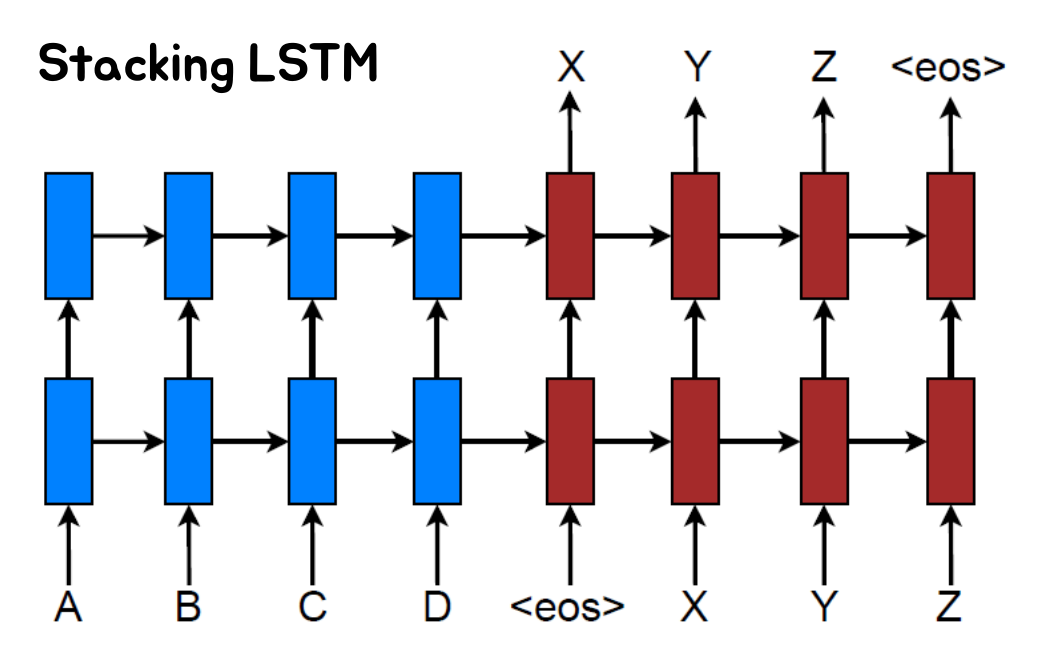
- Attentional mechanism이 NMT의 성능 향상을 위해 사용됨

- NMT에 Attention을 더 효과적으로 사용하기 위한 방법 제시

  -> Global Attention / Local Attention

- SOTA for both WMT'14 and WMT'15

2. NMT (Neural Machine Translation)



Source sentence x에서 Target sentence y가 생성될 확률

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

각 단어 y를 decoding 할 확률

$$p\left(y_j|y_{<j}, \boldsymbol{s}\right) = \mathrm{softmax}\left(g\left(\boldsymbol{h}_j\right)\right)$$

이전 hidden state로부터 현재의 hidden state를 계산

$$\boldsymbol{h}_j = f\left(\boldsymbol{h}_{j-1}, \boldsymbol{s}\right)$$

f는 RNN, LSTM, GRU

Source sentence $x_1, \ldots, x_n$ Target sentence $y_1, \ldots, y_m$ ↓

Encoder - 각 source 문장을 대표하는 s 계산

Decoder - 한 번에 하나의 target word를 생성

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

3. Attention-based Models

Common to these two types of models is the fact that at each time step t in the decoding phase, both approaches first take as input the hidden state $\boldsymbol{h}_t$ at the top layer of a stacking LSTM. The goal is then to derive a context vector $\boldsymbol{c}_t$ that captures relevant source-side information to help predict the current target word y_t. While these models differ in how the context vector $\boldsymbol{c}_t$ is derived, they share the same subsequent steps.

Specifically, given the target hidden state $\boldsymbol{h}_t$ and the source-side context vector $\boldsymbol{c}_t$, we employ a simple concatenation layer to combine the information from both vectors to produce an attentional hidden state as follows:

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W_c}[\boldsymbol{c}_t; \boldsymbol{h}_t]) \quad (5)$$

The attentional vector $\tilde{\boldsymbol{h}}_t$ is then fed through the softmax layer to produce the predictive distribution formulated as:

$$p(y_t|y_{<t}, x) = \text{softmax}(\boldsymbol{W_s}\tilde{\boldsymbol{h}}_t) \quad (6)$$

We now detail how each model type computes the source-side context vector $\boldsymbol{c}_t$.

- Stacked LSTM layer의 맨 위층의 hidden state를 사용
- 현재 target hidden state $\boldsymbol{h}_t$
- Target Word를 예측하기 위한 Source의 정보를 담고 있는 context vector $\boldsymbol{c}_t$
- attentional hidden state vector $\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W_c}[\boldsymbol{c}_t; \boldsymbol{h}_t])$

  $\boldsymbol{h}_t$, $\boldsymbol{c}_t$ concat

$$p(y_t|y_{<t}, x) = \text{softmax}(\boldsymbol{W_s}\tilde{\boldsymbol{h}}_t)$$

$\boldsymbol{c}_t$ 를 구하는 방법에 따라 Global/Local Attention

3. Attention-based Models

Global Attention

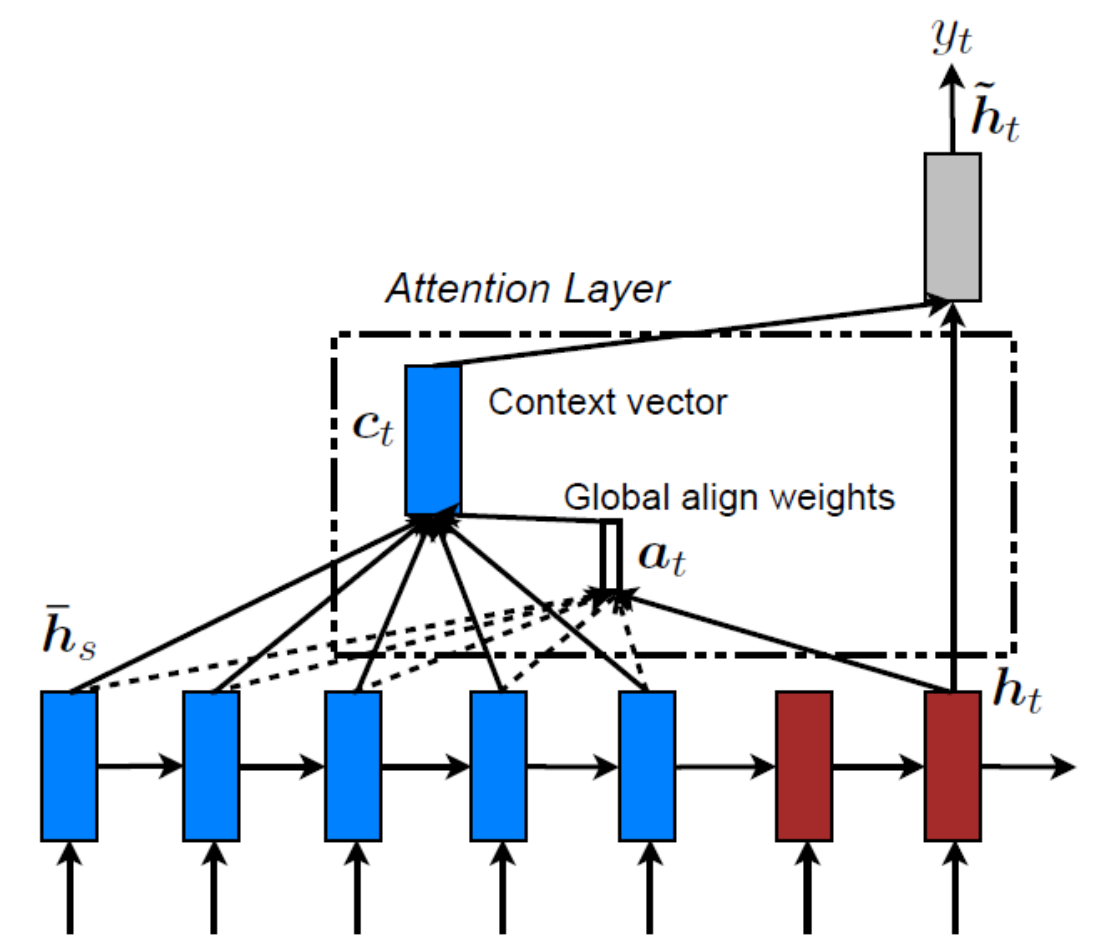
Encoder의 모든 hidden state를 고려

$$\boldsymbol{a}_t(s) = \text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) \quad \text{Score Softmax}$$

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

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

$\boldsymbol{c}_t$ 는 $\boldsymbol{a}_t$ 랑 $\bar{\boldsymbol{h}}_s$ weighted average



3. Attention-based Models

cf) 다양한 score function

content-based

$$\begin{aligned} \boldsymbol{a}_t(s) &= \text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) \\ &= \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \end{aligned}$$

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

location-based

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

Encoder의 hidden state가 사용되지 X

각각 이름과 식을 통해 생각해보면,
Encoder의 각 hidden state와 Decoder의 hidden state
각 content들을 비교한 content-based와
Decoder의 hidden state의 위치만을 고려한 location-based

3. Attention-based Models

cf) Luong Attention(Global Attention) vs Bahdanau Attention

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

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

3. Attention-based Models

cf) Luong Attention(Global Attention) vs Bahdanau Attention

Hidden State

- Encoder, Decoder 모두 stacked LSTM layer의 맨 위층의 hidden state를 사용

- bi-directional Encoder의 hidden state와 non-stacked uni-directional Decoder의 hidden state를 사용

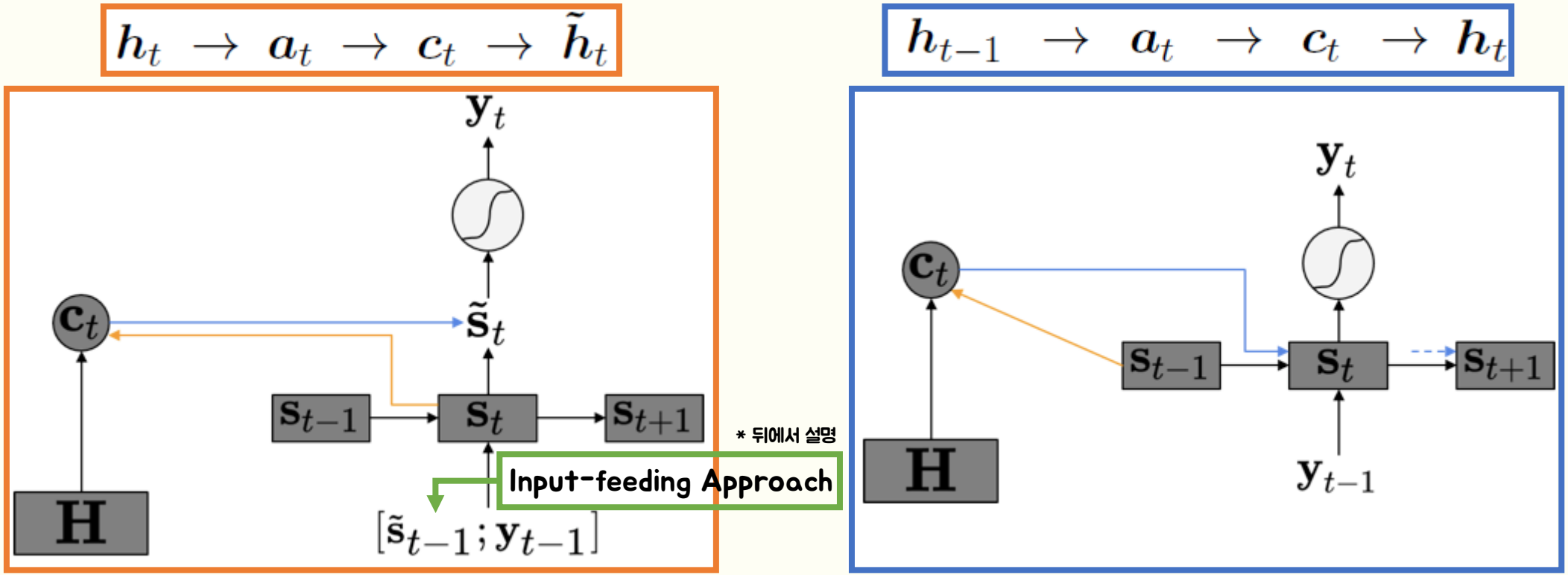
Computation path

$$h_t \rightarrow a_t \rightarrow c_t \rightarrow \tilde{h}_t$$

$$h_{t-1} \rightarrow a_t \rightarrow c_t \rightarrow h_t$$

3. Attention-based Models

cf) Luong Attention(Global Attention) vs Bahdanau Attention



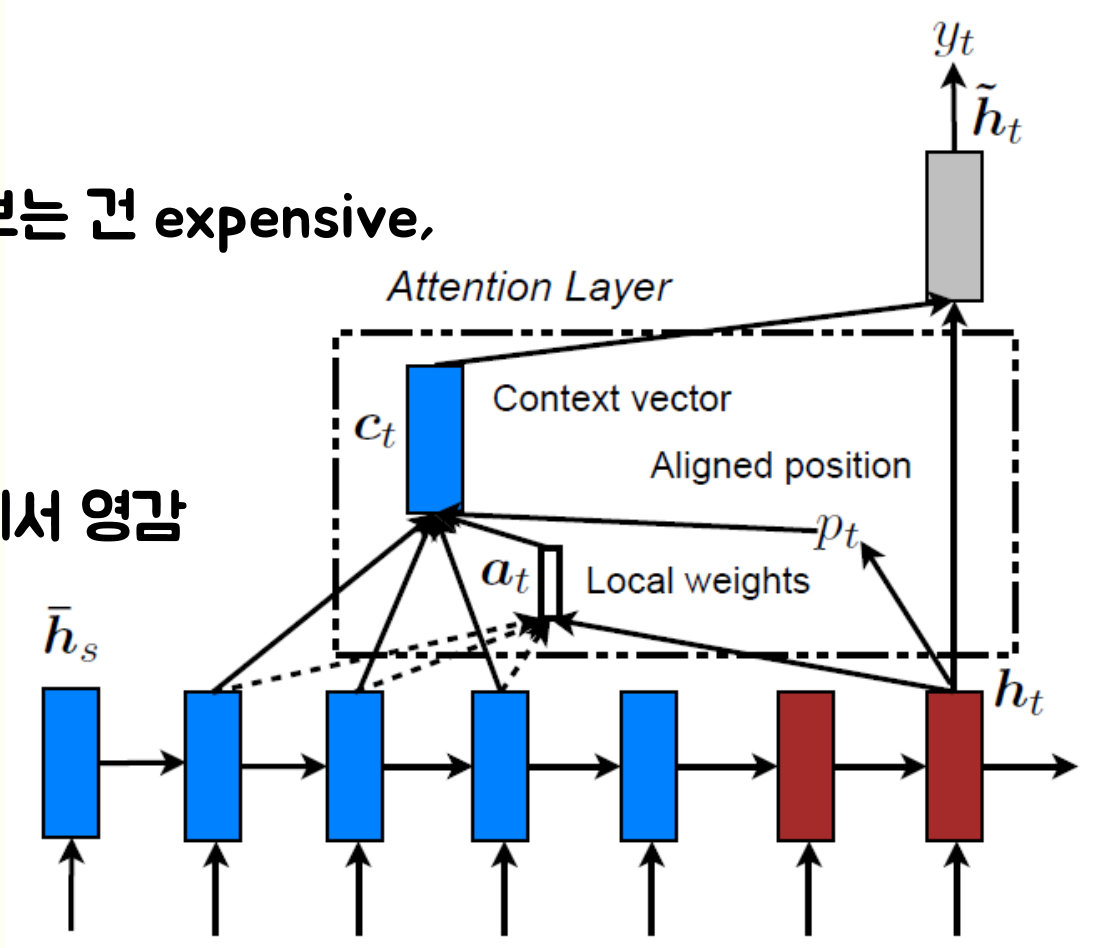
Bahdanau에서는 RNN의 재귀 연산이 수행되는 도중에 c_t 가 구해질 때까지 기다려야 하지만
Luong의 경우에는 y_t 를 구하는 부분과 RNN의 재귀 연산이 수행되는 부분을 분리할 수 있어 Computation path가 간소화!

3. Attention-based Models

Local Attention

Global Attention에서 source의 모든 단어를 보는 건 expensive,
긴 sequence에서 실용적이지 못하다
-> source position의 small subset에 집중

Soft and Hard Attention 사이의 tradeoff에서 영감
https://arxiv.org/abs/1502.03044



3. Attention-based Models

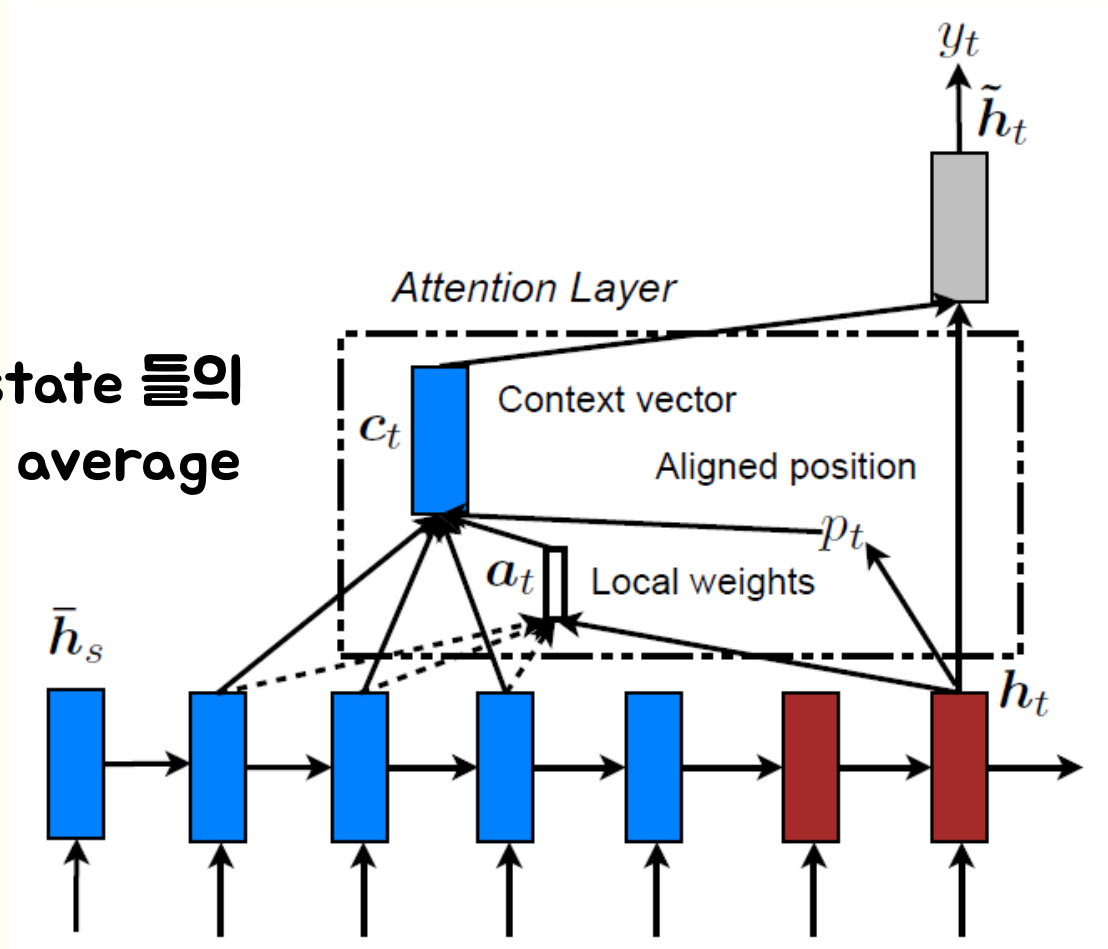
Local Attention

- Aligned Position p_t
- Context vector

  $[p_t - D, p_t + D]$ 사이에서의 source hidden state 들의 weighted average
- Local alignment vector a_t 는

  fixed-dimensional 2D+1

  +) global 에서 a_t 는 source의 timestamp 개수와 동일한 크기



3. Attention-based Models

Local Attention

$[p_t - D, p_t + D]$ Alignment

1. Monotonic Alignment (local-m)

$p_t = t$ t시점 같은 위치에 있는 단어끼리 연관이 클 것이라는 생각

2. Predictive Alignment (local-p)

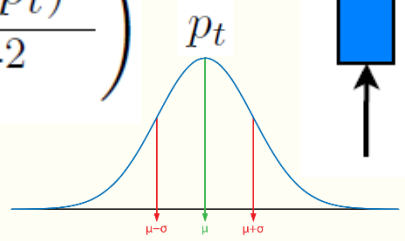
$p_t \in [0, S]$

$$p_t = S \cdot \text{sigmoid}(\boldsymbol{v}_p^\top \tanh(\boldsymbol{W}_p \boldsymbol{h}_t))$$

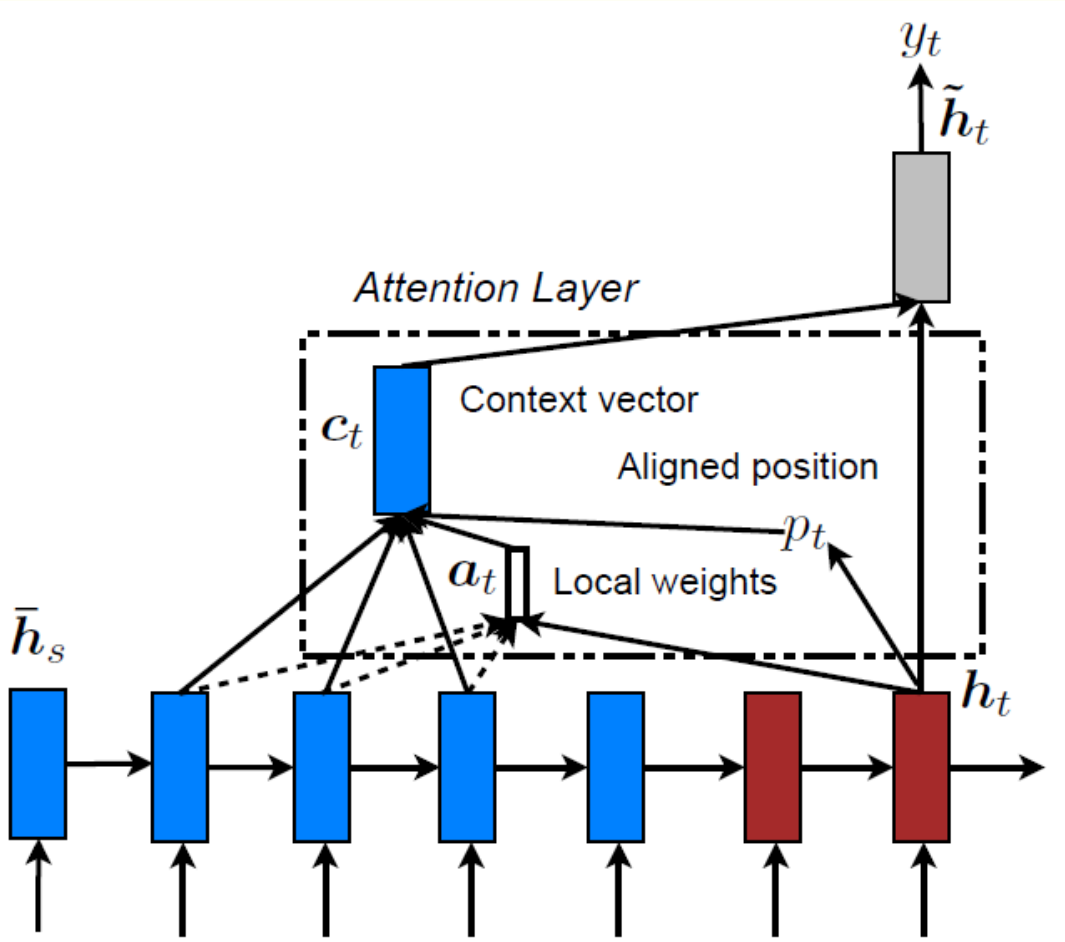
S: 문장의 길이

$\sigma = \frac{D}{2}$

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



p_t 를 기준으로 Gaussian 분포에 따라 주변 단어들이 의미를 가질 것이라는 생각



3. Attention-based Models

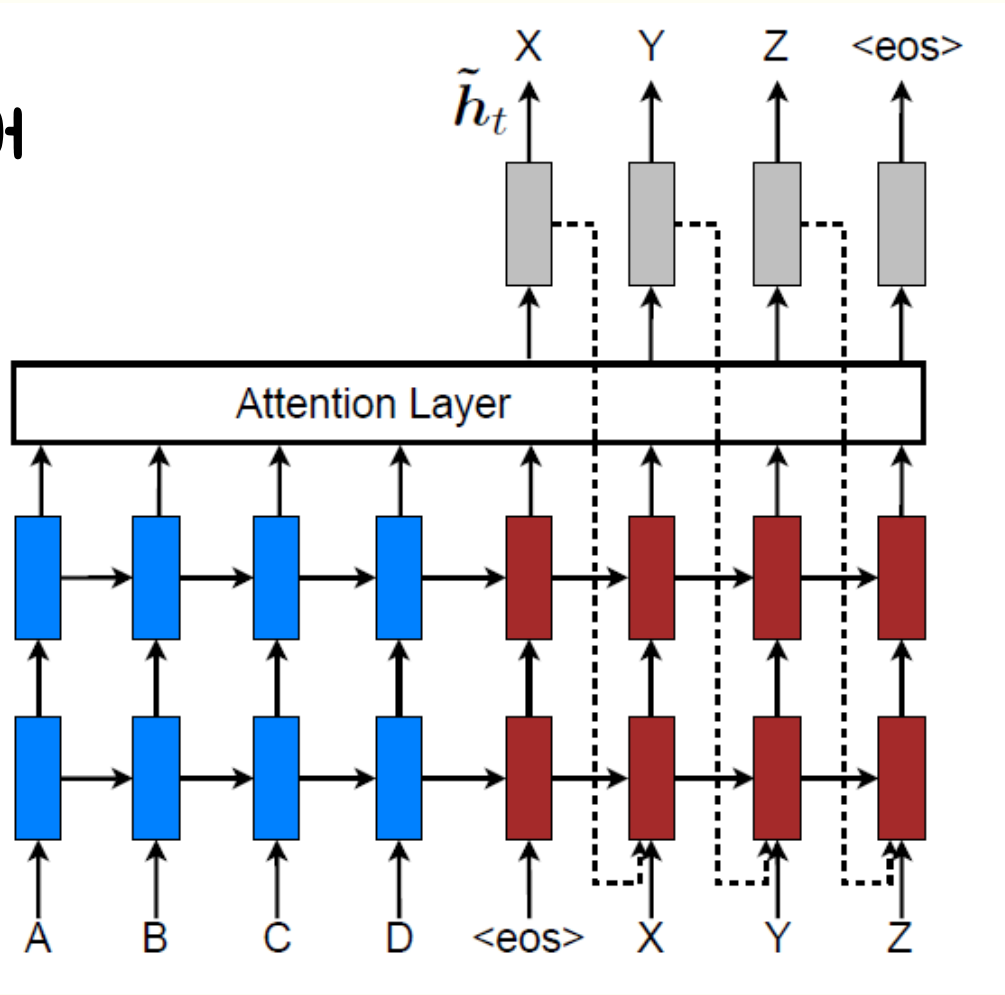
Input-feeding Approach

Standard MT에서는 coverage set이 계속 유지되어
어떤 source word가 번역되었는지 계속 따라간다

Attentional NMT에서도
이전 alignment 정보가 지금 alignment decision에
고려되도록 한다!

Attentional vector $\tilde{h}_t$ 를
다음 시점의 input과 concat

-> 이전 Alignment 정보를 잘 알 수 있도록 하고,
수평/수직적으로 매우 깊은 네트워크를 만든다



4. Experiments & Analysis

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*)

Table 1: **WMT'14 English-German results** – shown are the perplexities (ppl) and the *tokenized* BLEU scores of various systems on newstest2014. We highlight the **best** system in bold and give *progressive* improvements in italic between consecutive systems. *local-p* referes to the local attention with predictive alignments. We indicate for each attention model the alignment score function used in pararentheses.

System	BLEU
Top – *NMT + 5-gram rerank* (Montreal)	24.9
Our ensemble 8 models + unk replace	**25.9**

Table 2: **WMT'15 English-German results** *NIST* BLEU scores of the winning entry in WMT'15 and our best one on newstest2015.

System	Ppl.	BLEU
WMT'15 systems		
SOTA – *phrase-based* (Edinburgh)		**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*)

Table 3: **WMT'15 German-English results** – performances of various systems (similar to Table 1). The *base* system already includes source reversing on which we add *global* attention, *drop*out, input *feed*ing, and *unk* replacement.

SOTA for both WMT'14 and WMT'15

4. Experiments & Analysis

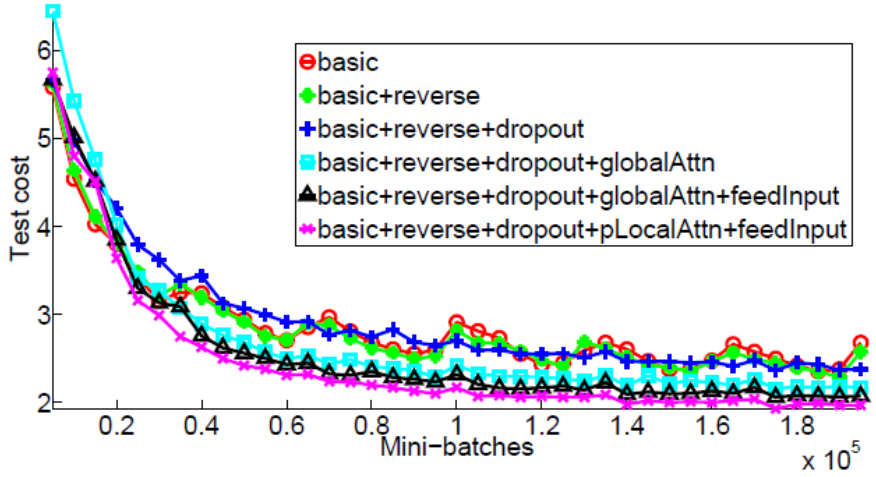


Figure 5: **Learning curves** – test cost (ln perplexity) on newstest2014 for English-German NMTs as training progresses.

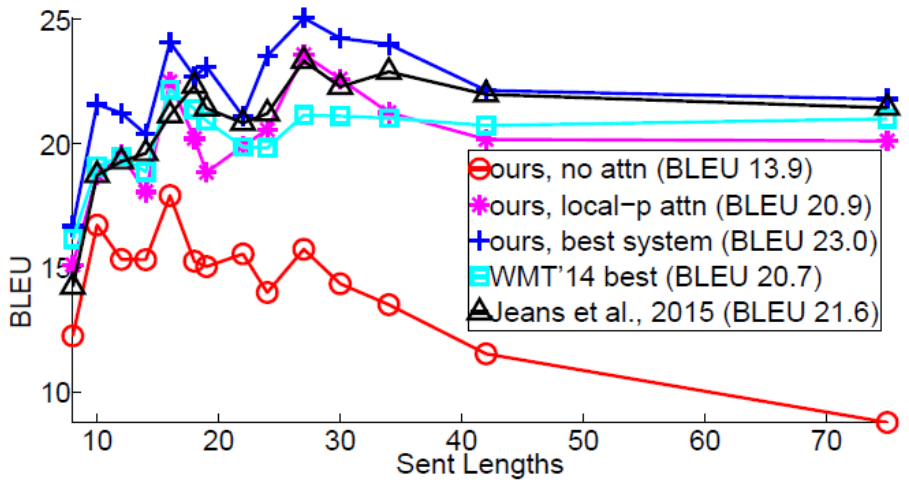


Figure 6: **Length Analysis** – translation qualities of different systems as sentences become longer.

Attention 적용했을 때 학습의 수렴 속도가 빠르다
문장의 길이가 길어질수록 번역 퀄리티 차이 보인다

4. Experiments & Analysis

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)**

Table 4: **Attentional Architectures** – performances of different attentional models. We trained two local-m (dot) models; both have ppl > 7.0.

Method	AER
global (location)	0.39
local-m (general)	0.34
local-p (general)	0.36
ensemble	0.34
Berkeley Aligner	0.32

Table 6: **AER scores** – results of various models on the RWTH English-German alignment data.

Alignment Error Rate

Attention 구조에 따른 번역 성능과 Alignment Quality

4. Experiments & Analysis

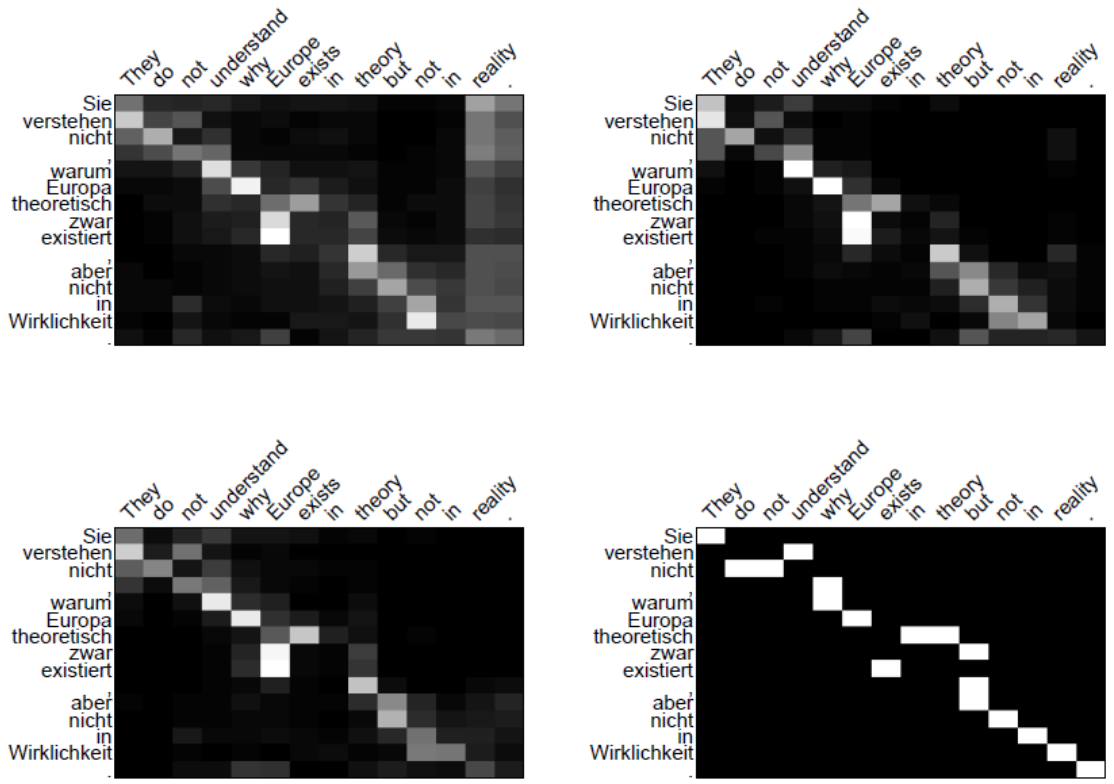


Figure 7: **Alignment visualizations** – shown are images of the attention weights learned by various models: (top left) global, (top right) local-m, and (bottom left) local-p. The *gold* alignments are displayed at the bottom right corner.

4. Experiments & Analysis

English-German translations	
src	Orlando Bloom and Miranda Kerr still love each other
ref	Orlando Bloom und *Miranda Kerr* lieben sich noch immer
best	Orlando Bloom und *Miranda Kerr* lieben einander noch immer .
base	Orlando Bloom und **Lucas Miranda** lieben einander noch immer .
src	" We ' re pleased the FAA recognizes that an enjoyable passenger experience is not incompatible with safety and security , " said Roger Dow , CEO of the U.S. Travel Association .
ref	" Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht im Widerspruch zur Sicherheit steht " , sagte *Roger Dow* , CEO der U.S. Travel Association .
best	" Wir freuen uns , dass die FAA anerkennt , dass ein angenehmes ist nicht mit Sicherheit und Sicherheit *unvereinbar* ist " , sagte *Roger Dow* , CEO der US - die .
base	" Wir freuen uns über die <unk> , dass ein <unk> <unk> mit Sicherheit nicht **vereinbar** ist mit Sicherheit und Sicherheit " , sagte *Roger* **Cameron** , CEO der US - <unk> .
German-English translations	
src	In einem Interview sagte Bloom jedoch , dass er und Kerr sich noch immer lieben .
ref	However , in an interview , Bloom has said that he and *Kerr* still love each other .
best	In an interview , however , Bloom said that he and *Kerr* still love .
base	However , in an interview , Bloom said that he and **Tina** were still <unk> .
src	Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
ref	The *austerity imposed by Berlin and the European Central Bank , coupled with the straitjacket* imposed on national economies through adherence to the common currency , has led many people to think Project Europe has gone too far .
best	Because of the strict *austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket* in which the respective national economy is forced to adhere to the common currency , many people believe that the European project has gone too far .
base	Because of the pressure **imposed by the European Central Bank and the Federal Central Bank with the strict austerity** imposed on the national economy in the face of the single currency , many people believe that the European project has gone too far .

Table 5: **Sample translations** – for each example, we show the source (*src*), the human translation (*ref*), the translation from our best model (*best*), and the translation of a non-attentional model (*base*). We italicize some *correct* translation segments and highlight a few **wrong** ones in bold.

5. Conclusion

- NMT에 Attention을 더 효과적으로 사용하기 위한 방법 제시
  -> Global Attention / Local Attention

- 다양한 Alignment Function 비교

- Attention-based NMT 짱!
  ex) Translating names and Handling long sentences

- SOTA for both WMT'14 and WMT'15

Q&A

참고

https://hcnoh.github.io/2018-12-11-bahdanau-attention Bahdanau Attention 개념 정리
https://hcnoh.github.io/2019-01-01-luong-attention Luong Attention 개념 정리