attention

강사 : 백병인

pi.paek@modulabs.co.kr

모두의연구소 Research Scientist

seq2seq vs. attention

- seq2seq

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

- attention

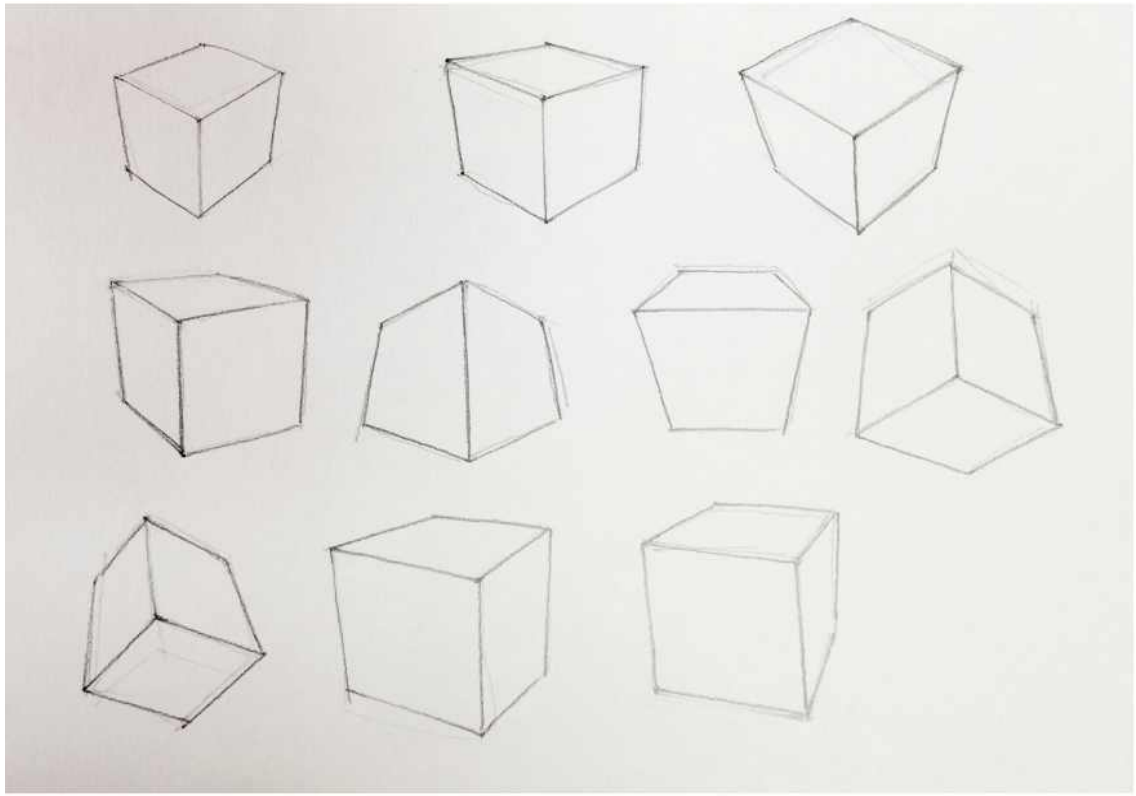
$$p(y_i|y_1,\ldots,y_{i-1},\mathbf{x}) = g(y_{i-1},s_i,c_i)$$

[틀린그림찾기] 두 수식에서 다른 부분은 어디일까요?
그 차이점이 의미하는 것은 무엇일까요?

관측자의 위치와 시선의 변화

- State Vector라는 개념의 전제
- 동일한 X를 해석(Encode)한 state c는 그 해석을 받아들이는 Decoder의 현재 위치에 무관하게 항상 **동일하게** 표현(represent)될 수 있다.

- 과연?

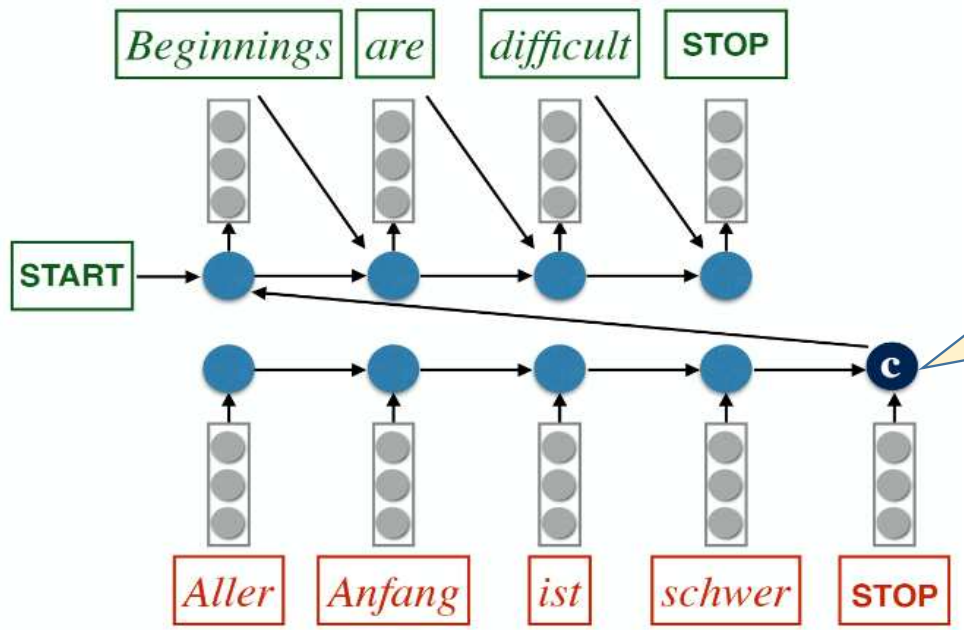


다양한 각도에서 본 육면체

(image) https://brunch.co.kr/@sooscape/4

C_i의 i가 의미하는 것

- 맥락 c는 Decoder의 포지션 i에 따라 **<u>다르게</u>** 표현(represent)되어야 한다.

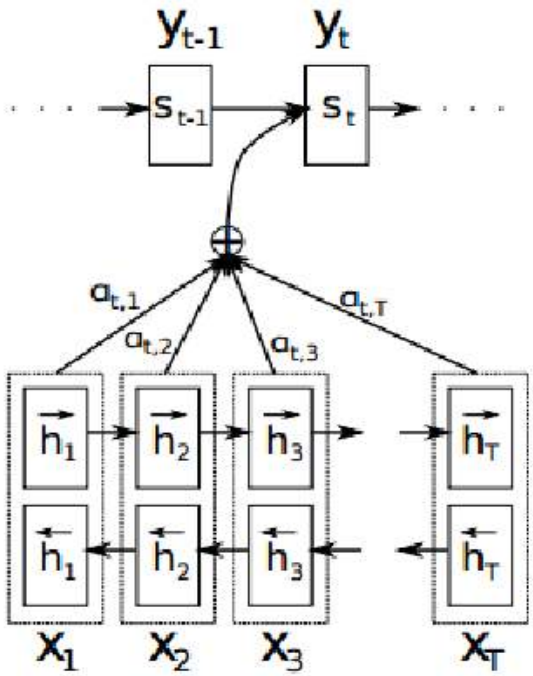


seq2seq은 Encoder가 생성한 feature representation(c)이 Decoder의 매 decoding step에 무관하게 **<u>항상 일정하다</u>**.

https://github.com/oxford-cs-deepnlp-2017/lectures/blob/master/Lecture%207%20-%20Conditional%20Language%20Modeling.pdf

NMT with attention

- Bahdanau et al. (2014) Neural Machine Translation by Jointly Learning to Align and Translate



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

Decoder Language Model

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

S_i : decoder state

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

C_i : state from attention

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

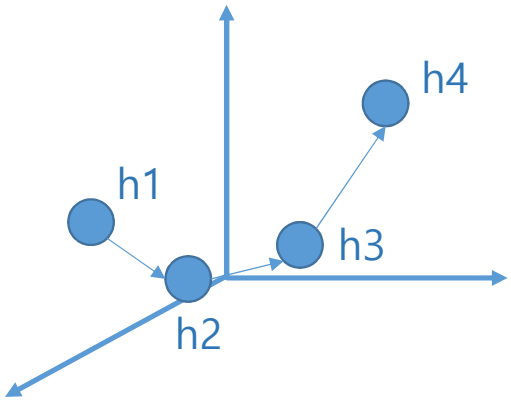
α_{ij} : attention weight w/ softmax

$$e_{ij} = a(s_{i-1},h_j)$$

e_{ij} : E-D state alignment

alignment of states (1)

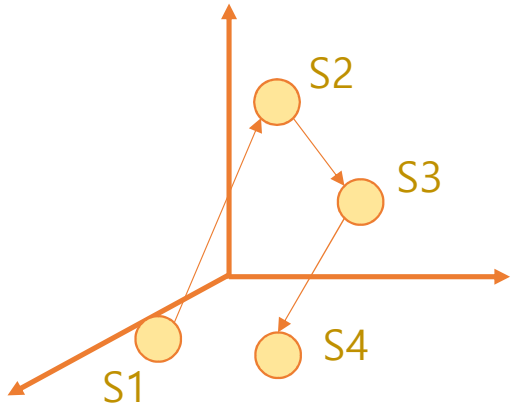
- h_j는 Encoder state space의 원소이다.
- S_i는 Decoder state space의 원소이다.
- 서로 다른 벡터공간 상의 두 벡터의 alignment를 어떻게 비교할 수 있을까?



Encoder State Space

비교 불가

$$e_{ij} = a(s_{i-1}, h_j)$$



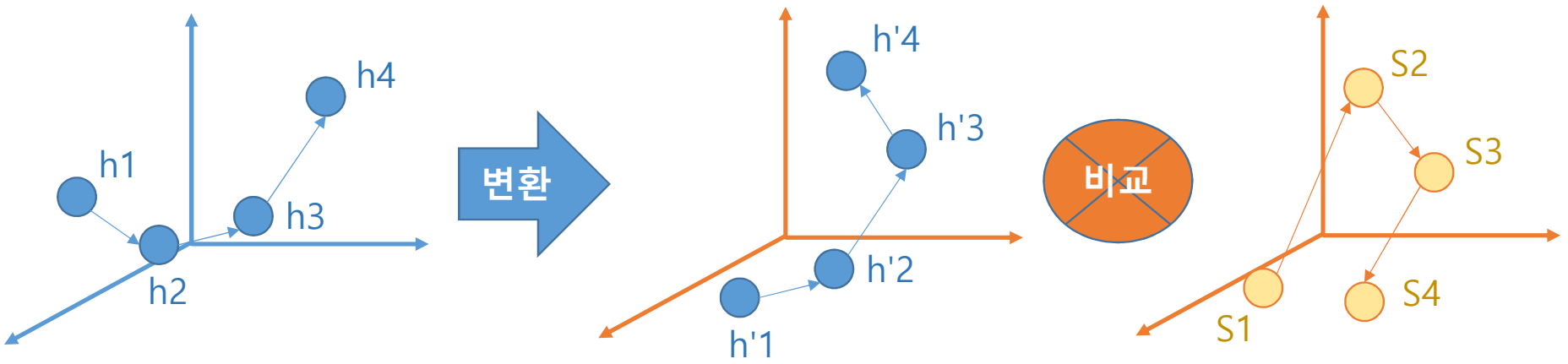
Decoder State Space

alignment of states (2)

- alignment score functions

  Luong et al. [2015] Effective Approaches to Attention-based Neural Machine Translation

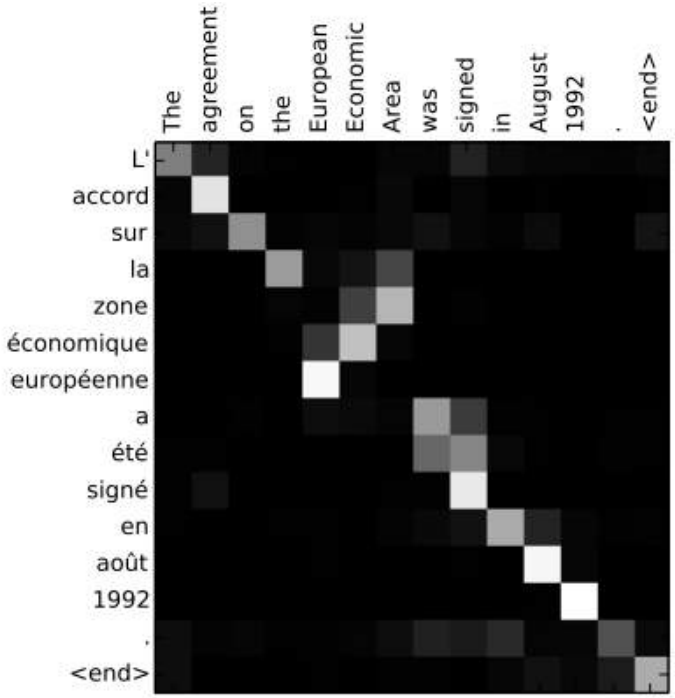
$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_a \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_a [\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right)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

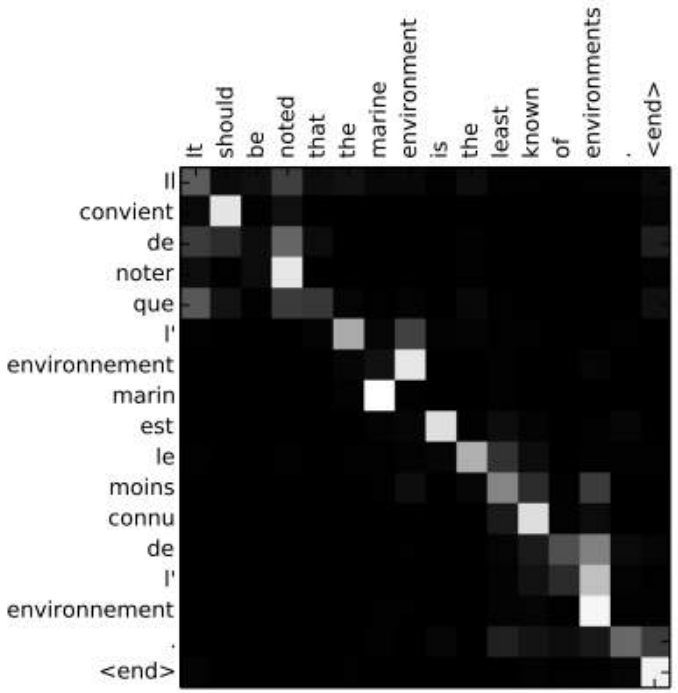


attention weight matrix

- Visualization of attention weight matrix

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





Bahdanau et al. (2014) Neural Machine Translation by Jointly Learning to Align and Translate https://arxiv.org/pdf/1409.0473

어순이 다른 언어간 번역

- attention을 통해 어순이 다른 언어간 번역 품질 향상
- 파파고 사례
  - Attention Map
    - 입력: 존과 메리는 아들과 딸이 있다.
    - 결과: John and Mary have a son and a daughter.

	<s>	존/NOUN	과/JOSA	메리/PROPERNOUN	는/JOSA	아들/NOUN	과/JOSA	딸/NOUN	이/JOSA	있/NORMALVERB	다/EOMI	./ETC	</s>
<s>	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
John/NNP	0.07	0.18	0.13	0.12	0.11	0.04	0.04	0.03	0.06	0.08	0.06	0.04	0.04
and/CC	0.07	0.03	0.44	0.06	0.06	0.01	0.06	0.02	0.05	0.07	0.05	0.04	0.04
Mary/NNP	0.02	0.04	0.03	0.52	0.08	0.02	0.01	0.12	0.05	0.03	0.03	0.02	0.02
have/VBP	0.06	0.02	0.09	0.06	0.07	0.04	0.06	0.05	0.10	0.19	0.10	0.08	0.08
a/DT	0.05	0.03	0.05	0.04	0.07	0.11	0.12	0.09	0.12	0.11	0.07	0.08	0.07
son/NN	0.04	0.03	0.04	0.05	0.06	0.15	0.11	0.14	0.09	0.09	0.06	0.08	0.06
and/CC	0.05	0.02	0.07	0.02	0.06	0.07	0.21	0.07	0.09	0.08	0.07	0.10	0.09
a/DT	0.03	0.02	0.02	0.10	0.05	0.06	0.06	0.30	0.09	0.07	0.06	0.08	0.07
daughter/NN	0.02	0.01	0.01	0.08	0.03	0.05	0.03	0.49	0.10	0.04	0.04	0.07	0.04
./.	0.10	0.01	0.03	0.02	0.06	0.04	0.08	0.07	0.11	0.10	0.10	0.14	0.12
</s>	0.10	0.03	0.03	0.04	0.07	0.04	0.05	0.05	0.11	0.10	0.09	0.11	0.20

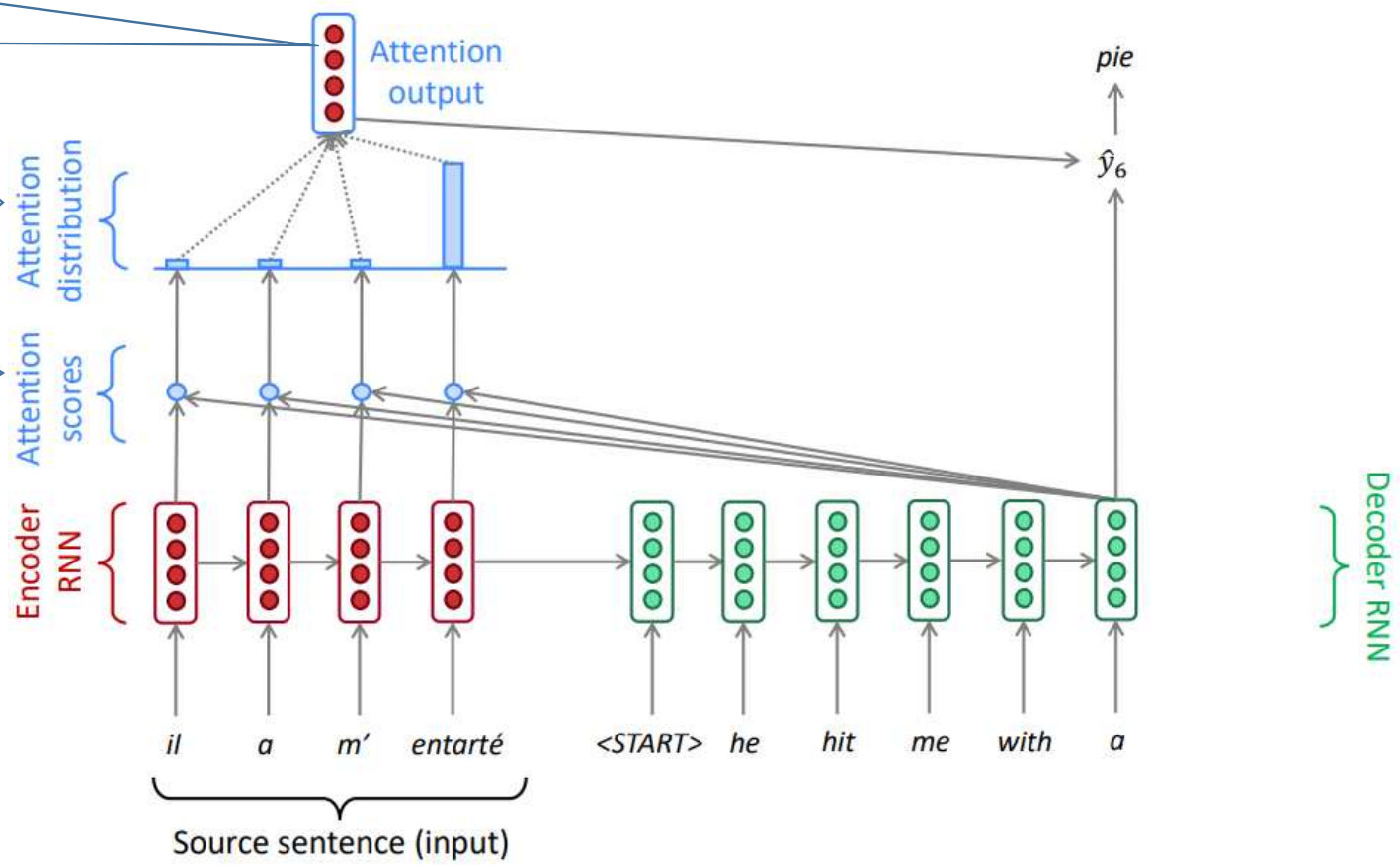
https://www.slideshare.net/deview/papagon2mt

seq2seq with attention

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

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

$$e_{ij} = a(s_{i-1}, h_j)$$



http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture08-nmt.pdf

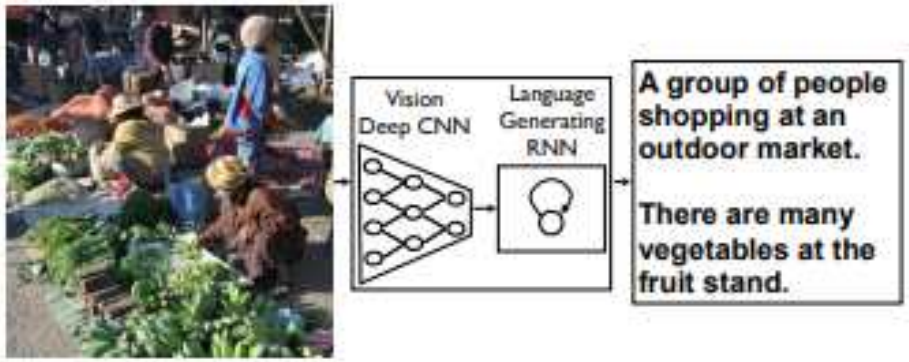
Attention의 장점

- NMT의 정확도 향상
- seq2seq의 단점 개선
  - state bottleneck의 구조적 단점 극복
  - encoder state에 decode가 직접 접근하므로 vanishing gradient 극복
- Attention Matrix가 주는 interpretability
- General Deep Learning Technique

> - More general definition of attention:
>   - Given a set of vector *values*, and a vector *query*, attention is a technique to compute a weighted sum of the values, dependent on the query.

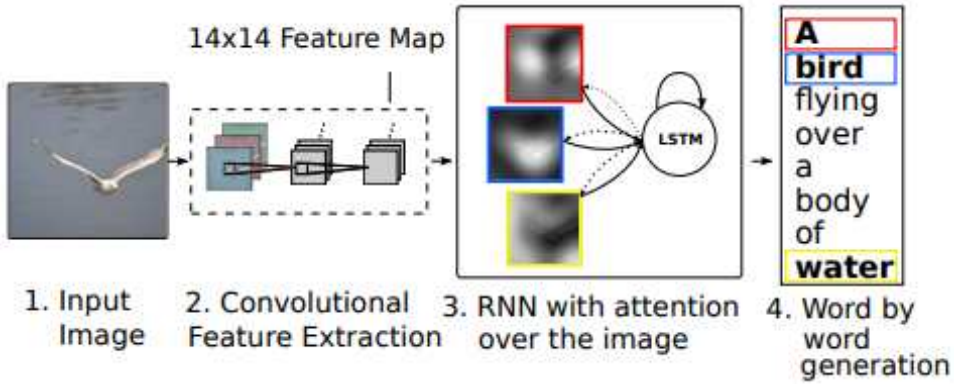
Image to Text

Attention이 적용되지 않은 Image2Text



- Vinyals et al.[2014] **Show and Tell**: A Neural Image Caption Generator

Attention이 적용된 Image2Text



- Xu et al.[2015] **Show, Attend and Tell**: Neural Image Caption Generation with Visual Attention