Show, Attend and Tell : Neural Image Caption Generation with Visual Attention

Undergraduate Researcher at CVLab

Lee Dohyeong

2025.2.20

Contents

- Introduction
- Related Work
- Method
- Experiments
- Conclusion

Introduction

- Proposal of an attention-based caption generation model

- Focus on specific image regions when generating words

- Visualization and interpretability of attention

- Visually reveal internal decision-making

- Performance improvement

- State-of-the-art captioning performance

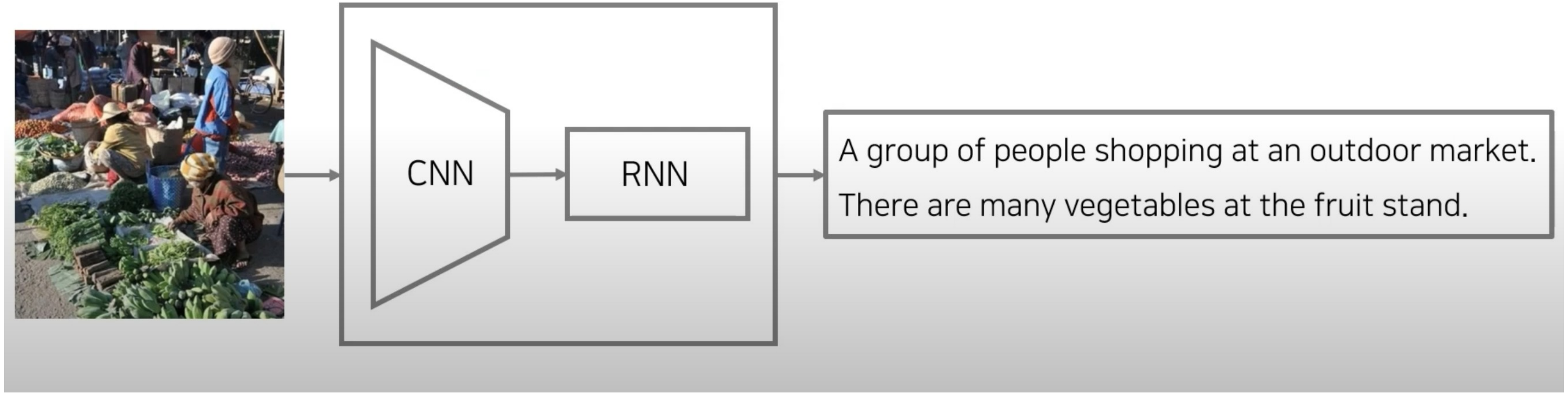
Introduction

- Image Caption

– concise sentence that describes the content and context of an image.

- NIC (Neural Image Caption)

– CNN + RNN(LSTM), End-to-End



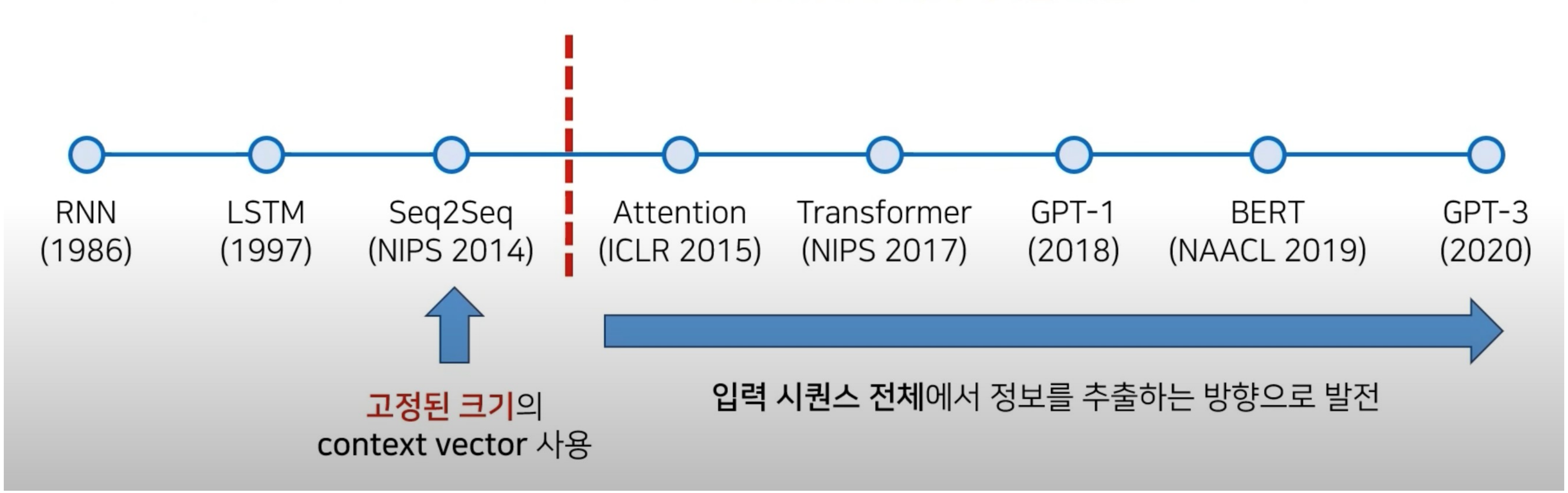
A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

Introduction

● Development process

- Utilizing an LSTM-based seq2seq-like architecture
- Recent models are based on the Transformer architecture.



RNN (1986)

LSTM (1997)

Seq2Seq (NIPS 2014)

Attention (ICLR 2015)

Transformer (NIPS 2017)

GPT-1 (2018)

BERT (NAACL 2019)

GPT-3 (2020)

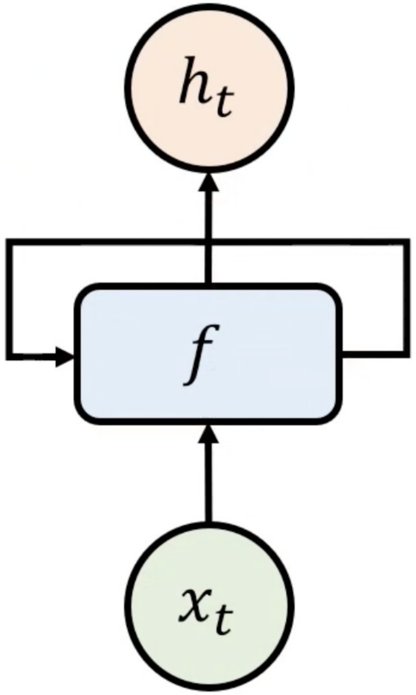
고정된 크기의 context vector 사용

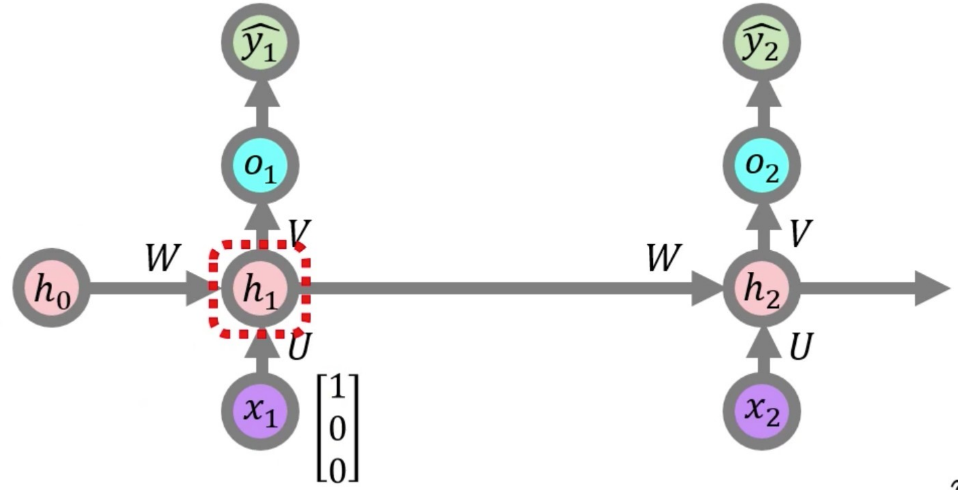
입력 시퀀스 전체에서 정보를 추출하는 방향으로 발전

Related Work

- RNN(Recurrent Neural Network)
  - Input : x
  - Hidden State : h
  - output : o

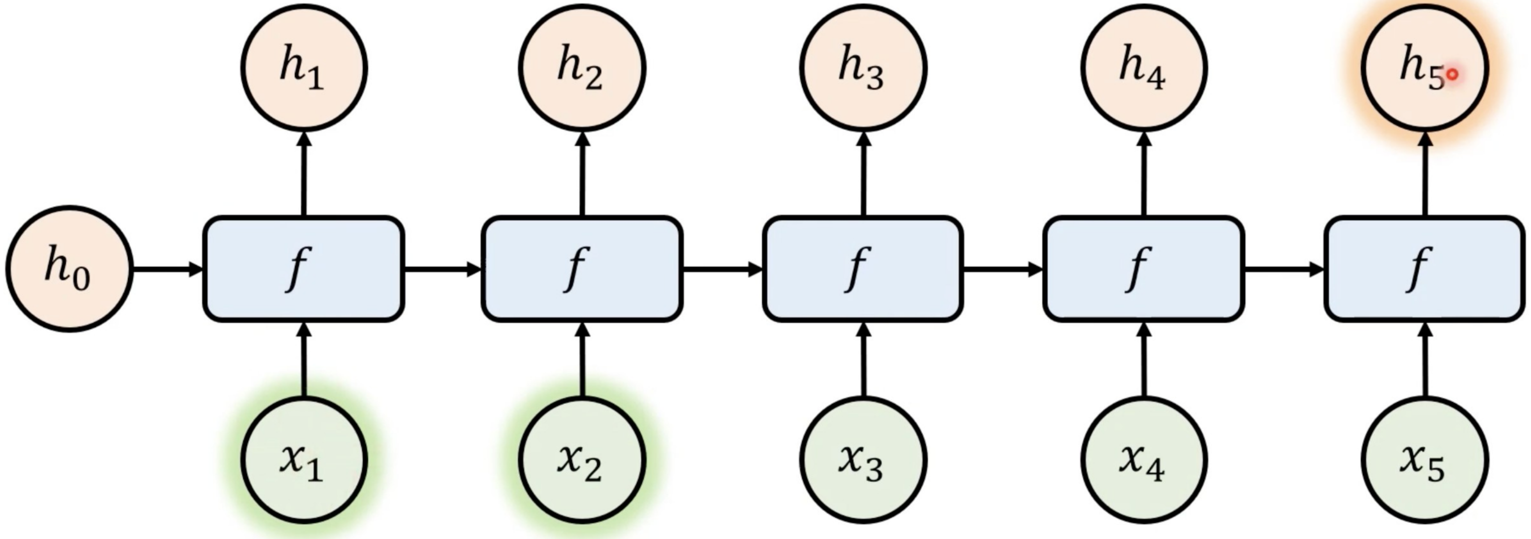
$$h_1 = \tanh(Wh_0 + Ux_1)$$





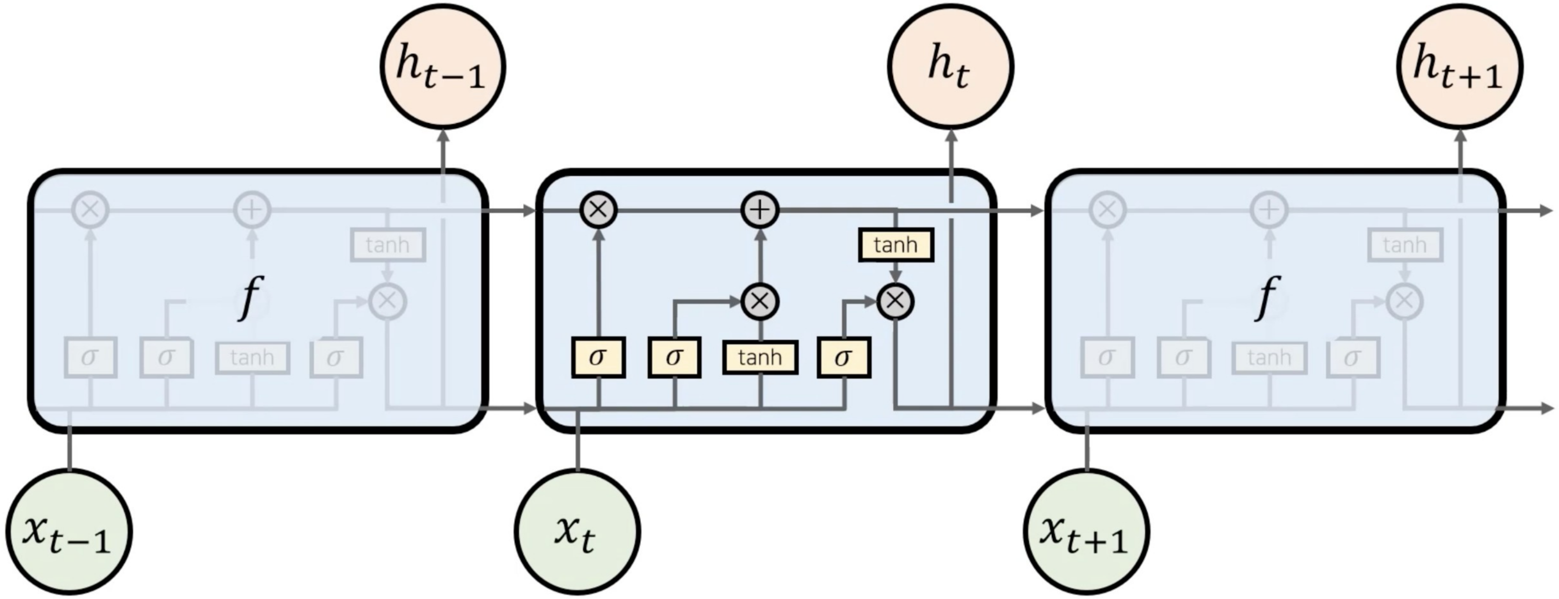
Related Work

- RNN(Recurrent Neural Network)
  - Hard to learn long-term dependencies.
    - vanishing and exploding gradients



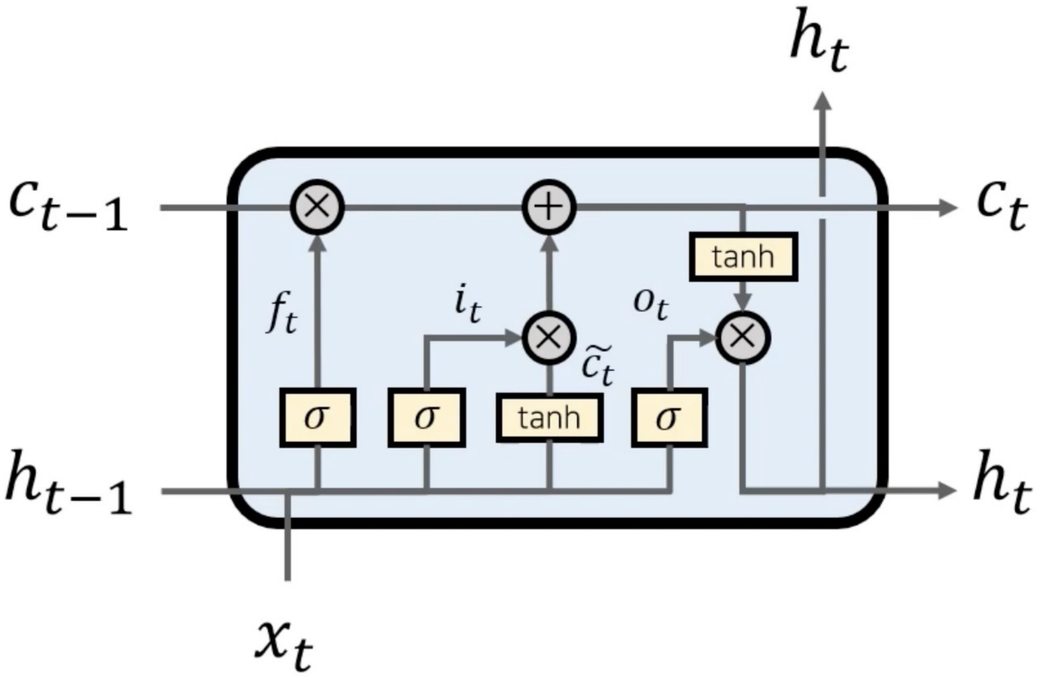
Related Work

- LSTM(Long Short-Term Memory)
  - Long Term Memory : Cell State
  - Short Term Memory : Hidden State



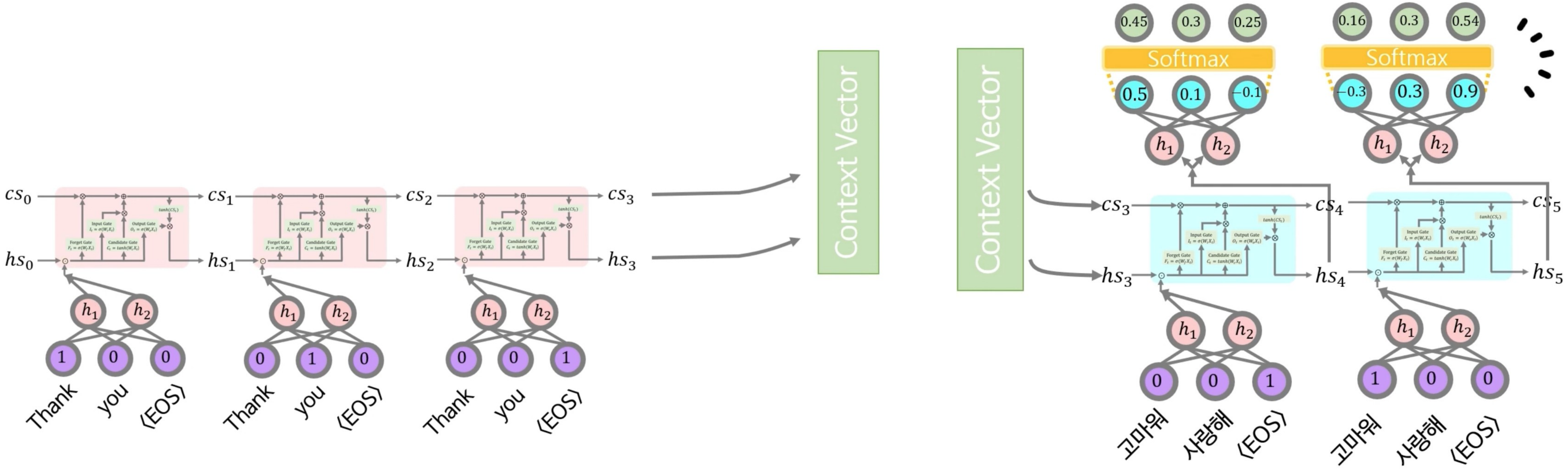
Related Work

- LSTM(Long Short-Term Memory)
  - Forget Gate : Forgets unnecessary old information.
  - Input Gate : Integrates new information into long-term memory.



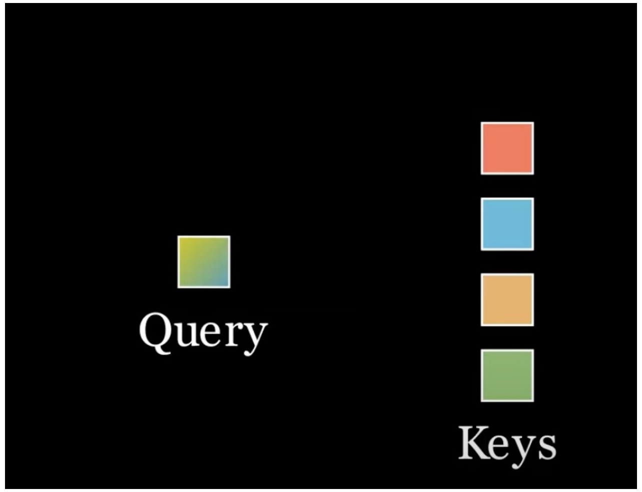
Related Work

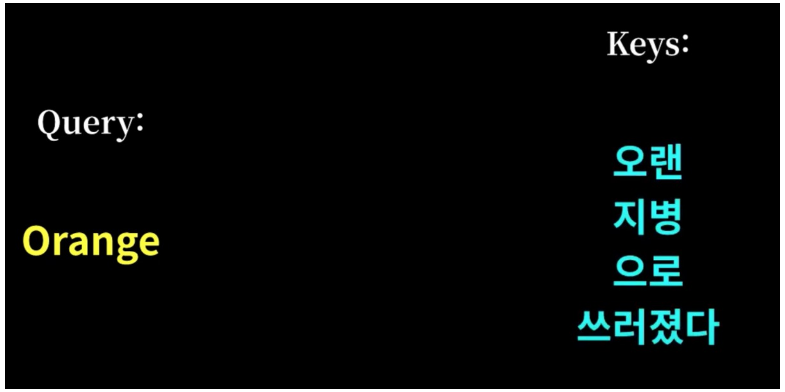
- Seq2Seq
  - Encoder - Decoder
  - End-to-End Learning
  - Variable-Length Sequences



Related Work

- Attention
  - Focus on the necessary parts
    - Dot product between Query and Key
    - Apply Softmax(Dot product result)
    - Weighted Sum = Dot product result * Value

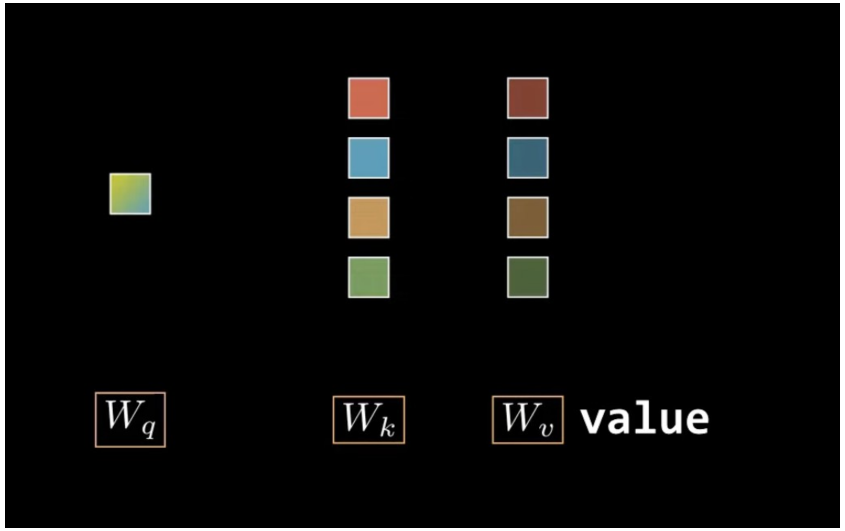



Query:

Orange

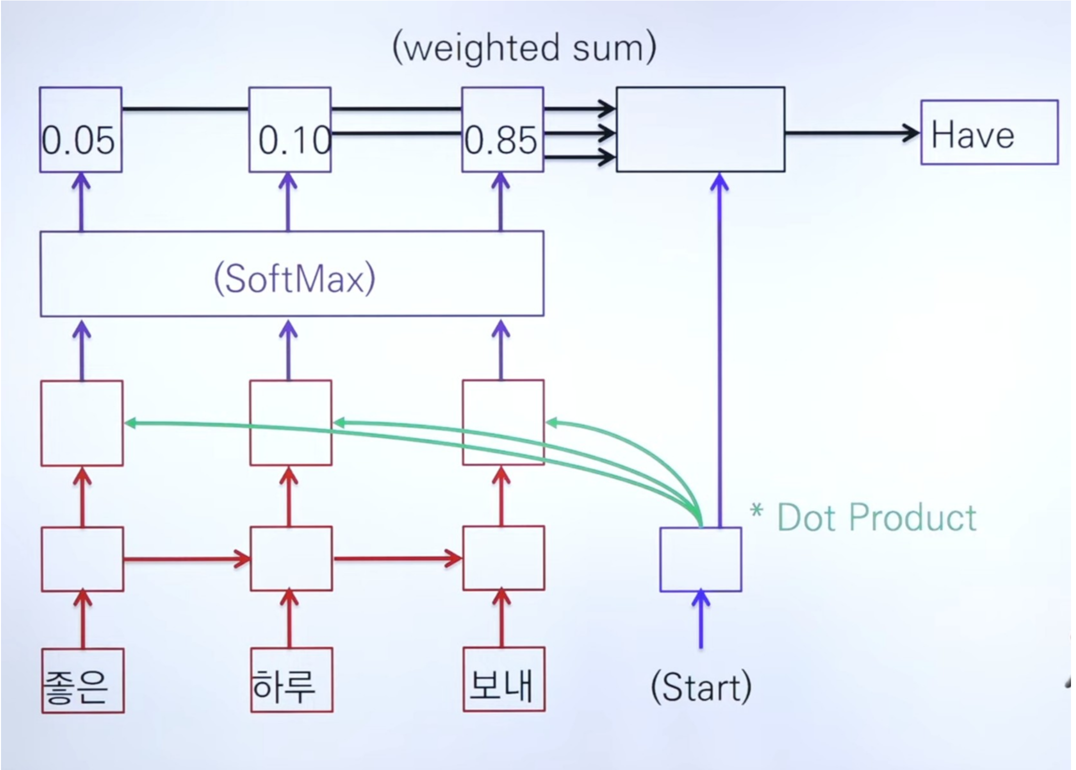
Keys:

오랜
지병
으로
쓰러졌다



Related Work

- Seq2Seq With Attention
  - Seq2Seq + Attention
  - Decode : References all of the encoder's(RNN) outputs.



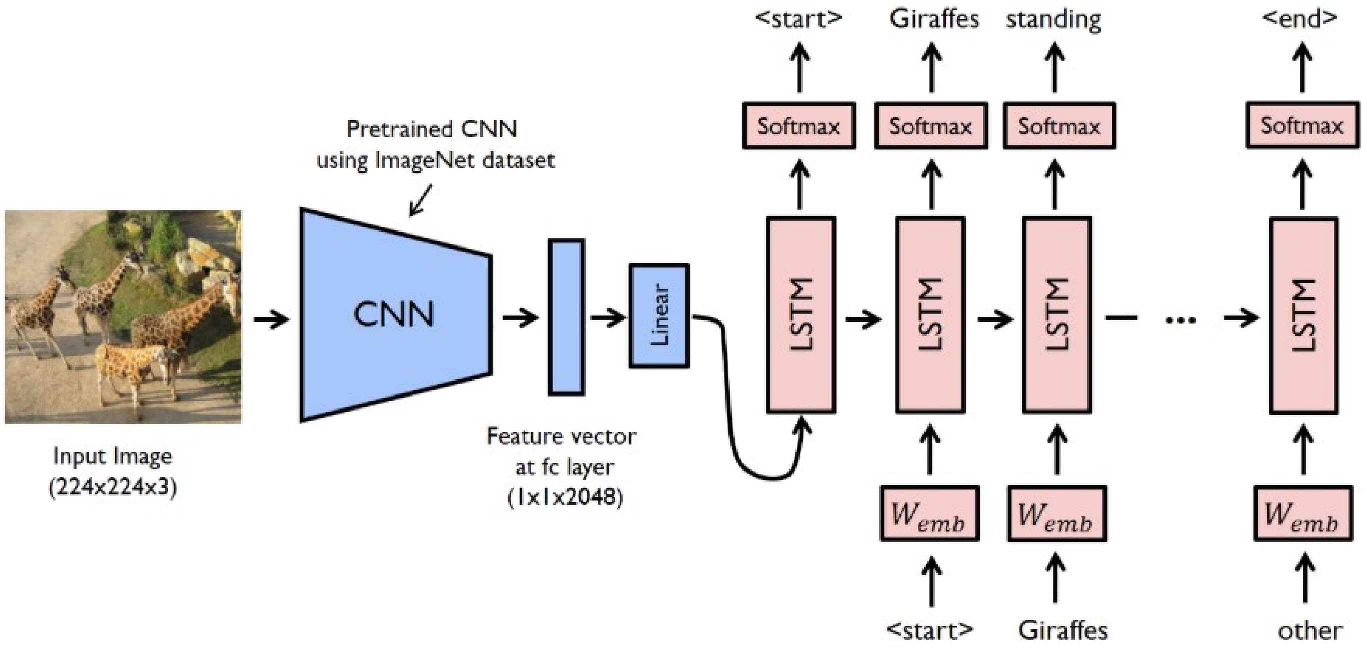
Related Work

- Show and Tell : A Neural Image Caption Generator
  - **최대우도추정**(MLE, Maximum Likelihood Estimation)

$$\theta^{\star} = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta)$$

Related Work

- Show and Tell : A Neural Image Caption Generator
  - Image embedding : Seq2Seq[RNN]>>CNN(GoogleNet) ; Encoder
  - Generating sentences : RNN(Seq2Seq) ; Decoder
  - Likelihood P(S | I) -> Maximization : S(Sentence) , I(Image)

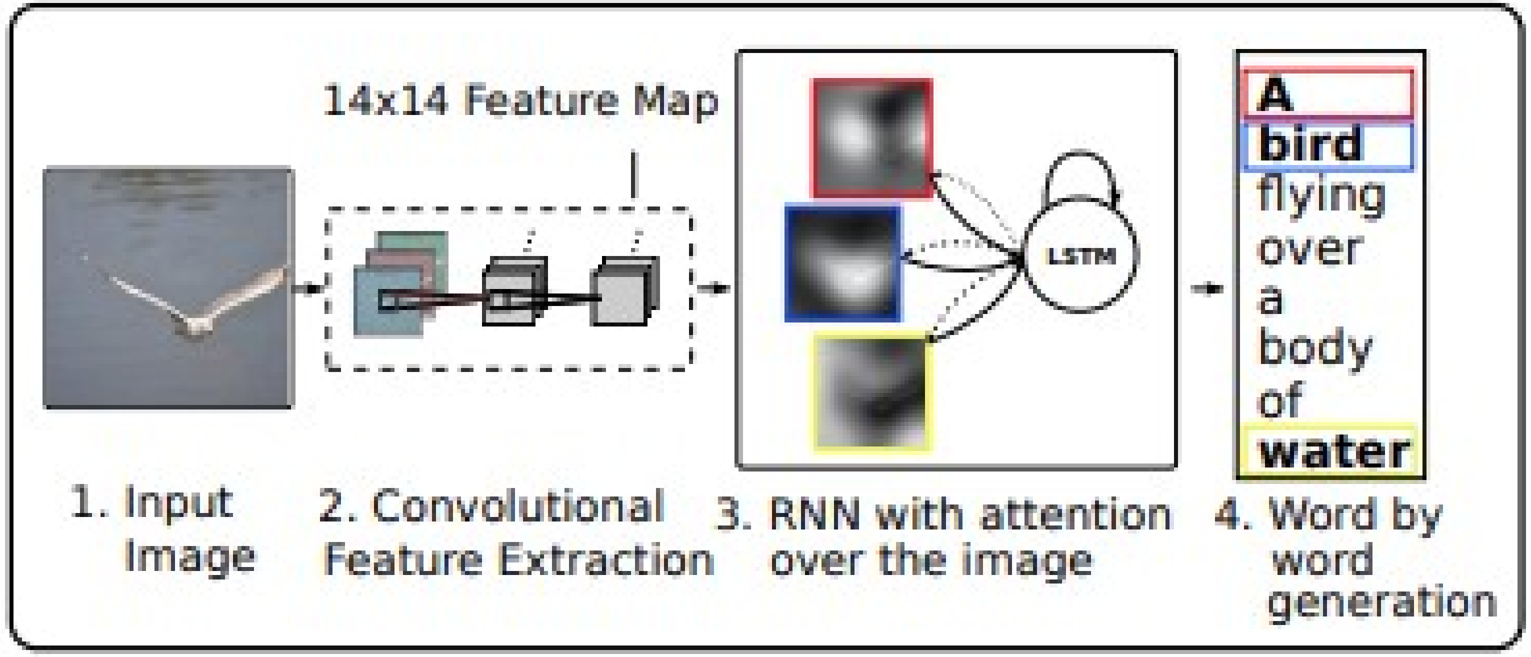


$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t) \ .$$

Method

- Show, Attend and Tell

Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in section 3.1 & 5.4

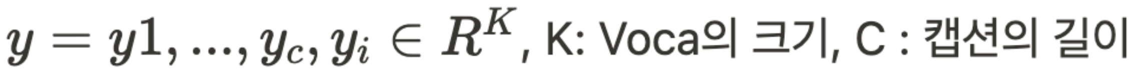

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Method

- Show, Attend and Tell
  - Background : Similar to Encoder−Decoder structure, similar to 'translating' images into sentences
  - Attention : "Extract only needed parts.",
    - Learn latent alignment between images and words from start
    - Can effectively capture spatial details, colors, and ambiance
  - Attetion Method :
    1. CNN -> (extract) 2D Feature map[Annoation Vectors]
    2. Soft / Hard Attention decides which feature map locate
    3. End to end
  - Effect :
    1. Generality : Can handle new situations beyond learned object categories.
    2. Interpretability : Visually explains which parts were attended to.

Method

- Show, Attend and Tell
  - [Encoder]
    - Raw Image -> 1 hot Encoding -> caption y
    - Use CNN -> Annotation vector(Feature Vector)



$y = y1, ..., y_c, y_i \in R^K$, K: Voca의 크기, C : 캡션의 길이

Method

- Show, Attend and Tell
  - [Decoder]
    - LSTM
    - Init,c , init,h
    - Embedding : Word -> Vector

$$c_0 = f_{\text{init,c}}\left(\frac{1}{L}\sum_i a_i\right), \quad h_0 = f_{\text{init,h}}\left(\frac{1}{L}\sum_i a_i\right)$$

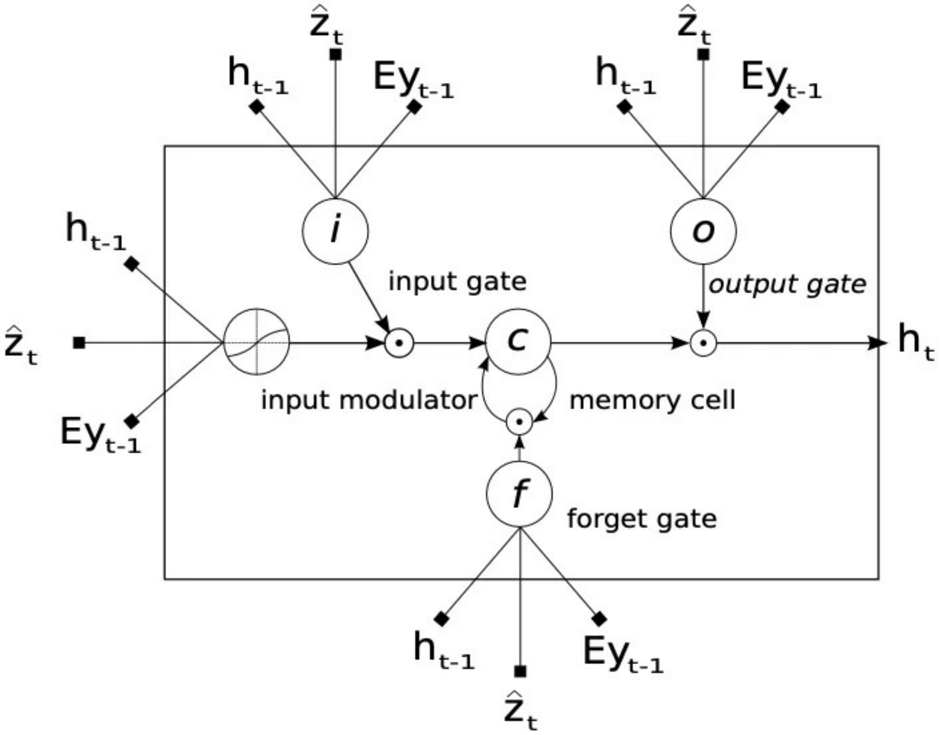
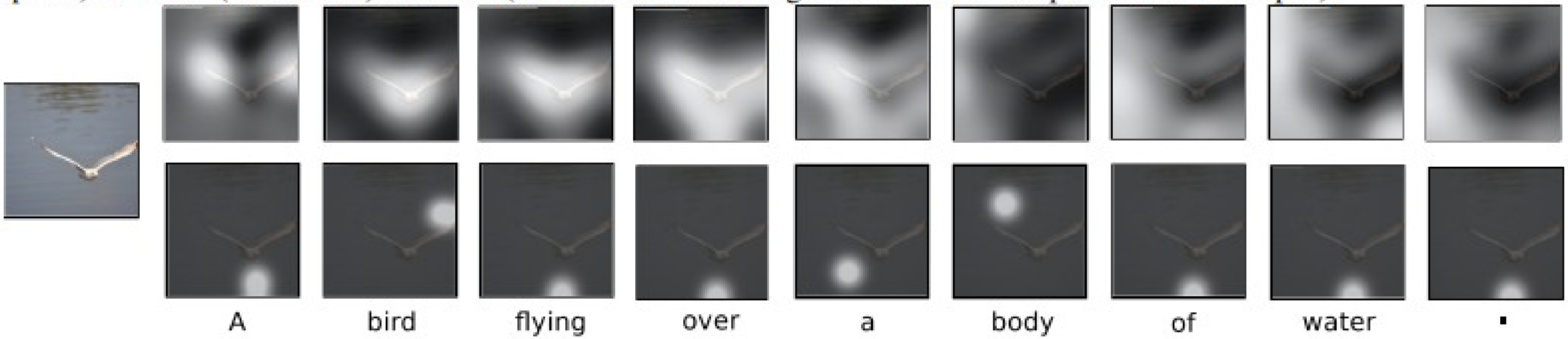


Figure 4. A LSTM cell, lines with bolded squares imply projections with a learnt weight vector. Each cell learns how to weigh its input components (input gate), while learning how to modulate that contribution to the memory (input modulator). It also learns weights which erase the memory cell (forget gate), and weights which control how this memory should be emitted (output gate).

Method

- Show, Attend and Tell
  - Hard Attention : Select only one location probabilistically / Reinforcement Learning
  - Soft Attention : Weighted average of all locations (Softmax) → Standard backprop (end-to-end)

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)



Method

- Show, Attend and Tell
  - Hard Attention :

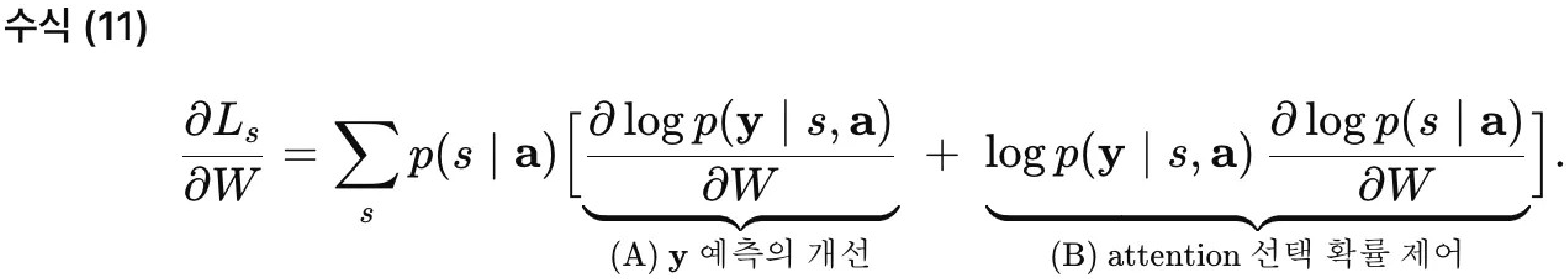
$$\alpha_{t,i} = p(s_{t,i} = 1 \mid s_{j<t}, \mathbf{a}) \ldots (8)$$

$$\hat{\mathbf{z}}_t = \sum_i s_{t,i} \, \mathbf{a}_i \ldots (9)$$

$$\begin{aligned} L_s &= \sum_s p(s \mid \mathbf{a}) \log p(\mathbf{y} \mid s, \mathbf{a}) \\ &\leq \log \sum_s p(s \mid \mathbf{a}) p(\mathbf{y} \mid s, \mathbf{a}) \\ &= \log p(\mathbf{y} \mid \mathbf{a}) \end{aligned}$$

Method

- Show, Attend and Tell
    - Hard Attention :

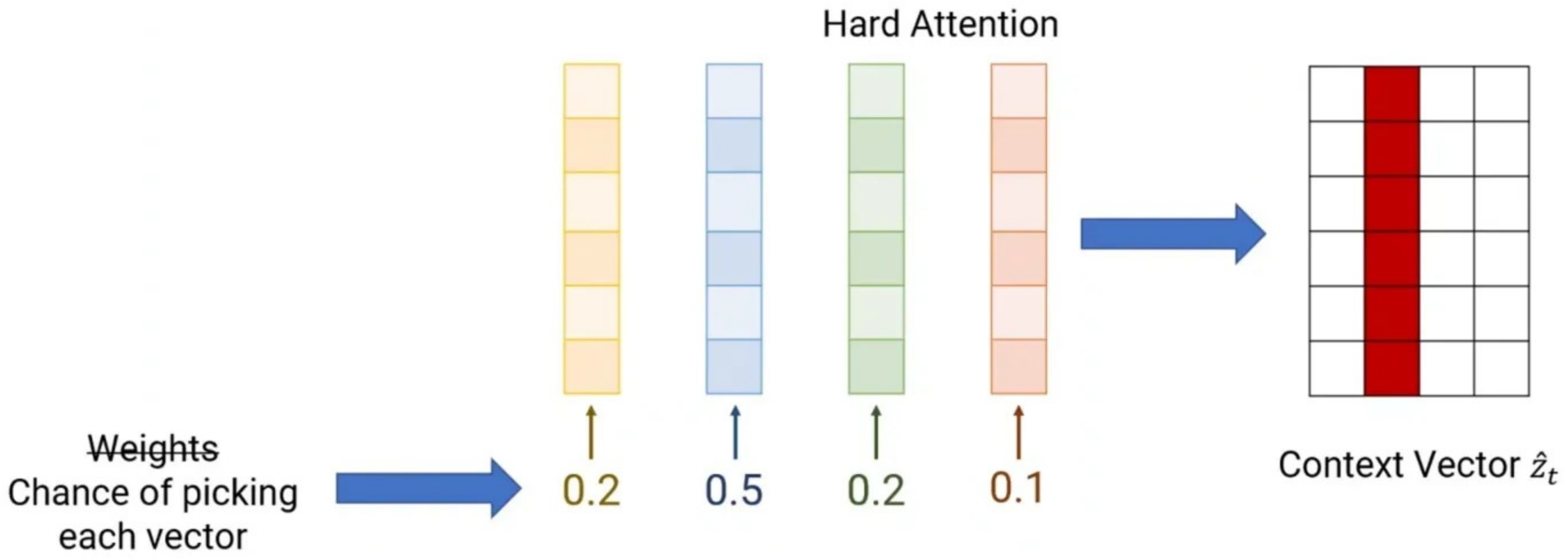


수식 (11)

$$\frac{\partial L_s}{\partial W} = \sum_s p(s \mid \mathbf{a}) \left[\underbrace{\frac{\partial \log p(\mathbf{y} \mid s, \mathbf{a})}{\partial W}}_{\text{(A) } \mathbf{y} \text{ 예측의 개선}} + \underbrace{\log p(\mathbf{y} \mid s, \mathbf{a}) \frac{\partial \log p(s \mid \mathbf{a})}{\partial W}}_{\text{(B) attention 선택 확률 제어}} \right].$$

Method

- Show, Attend and Tell
  - Stochastic Hard Attention : Monte carlo sampling, Backpropagation X



Method

- Show, Attend and Tell
  - Stochastic Hard Attention :

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{L} \exp(e_{t,k})}.$$

$$e_{ti} = f_{att}(a_i, h_{t-1}) = v_a^T \tanh(W_a a_i + W_h h_{t-1})$$

Method

- Show, Attend and Tell
  - Deterministic Soft Attention :
    - NWGM : “Normalized Weighted Geometric Mean”

$$\mathbf{NWGM}[p(y_t = k \mid a)] = \frac{\prod_i \exp(n_{t,k,i})\, p(s_{t,i}{=}1|a)}{\sum_j \prod_i \exp(n_{t,j,i})\, p(s_{t,i}{=}1|a)} = \frac{\exp\left(\mathbb{E}_{p(s_t|a)}[n_{t,k}]\right)}{\sum_j \exp\left(\mathbb{E}_{p(s_t|a)}[n_{t,j}]\right)}.$$

$$NWGM[p(y_t = k \mid \mathbf{a})] \approx \mathbb{E}[p(y_t = k \mid \mathbf{a})]$$

Method

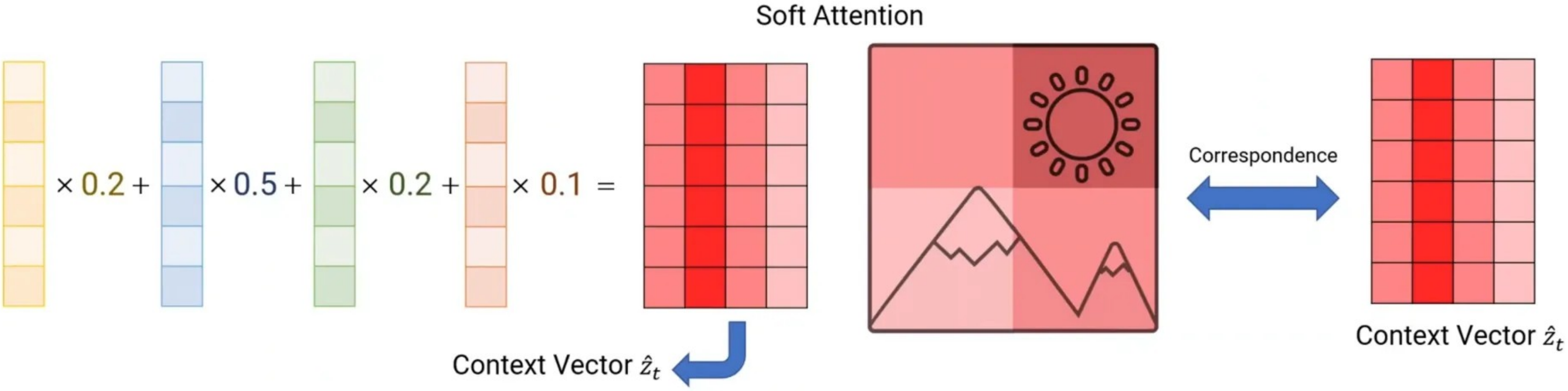
- Show, Attend and Tell
    - Deterministic Soft Attention : Use mean of a probability distribution

$$\mathbb{E}_{p(s_t|a)}\left[\hat{z}_t\right] = \sum_{i=1}^{L} \alpha_{t,i}\, a_i \qquad \ldots(13)$$

$$n_t = L_o(Ey_{t-1} + L_h h_t + L_z \hat{z}_t), n_{t,i}$$

Method

- Show, Attend and Tell
  - Deterministic Soft Attention : (=Attention), Backpropagation, end to end



$$\hat{z}_t = \sum_i \alpha_i a_i$$

Method

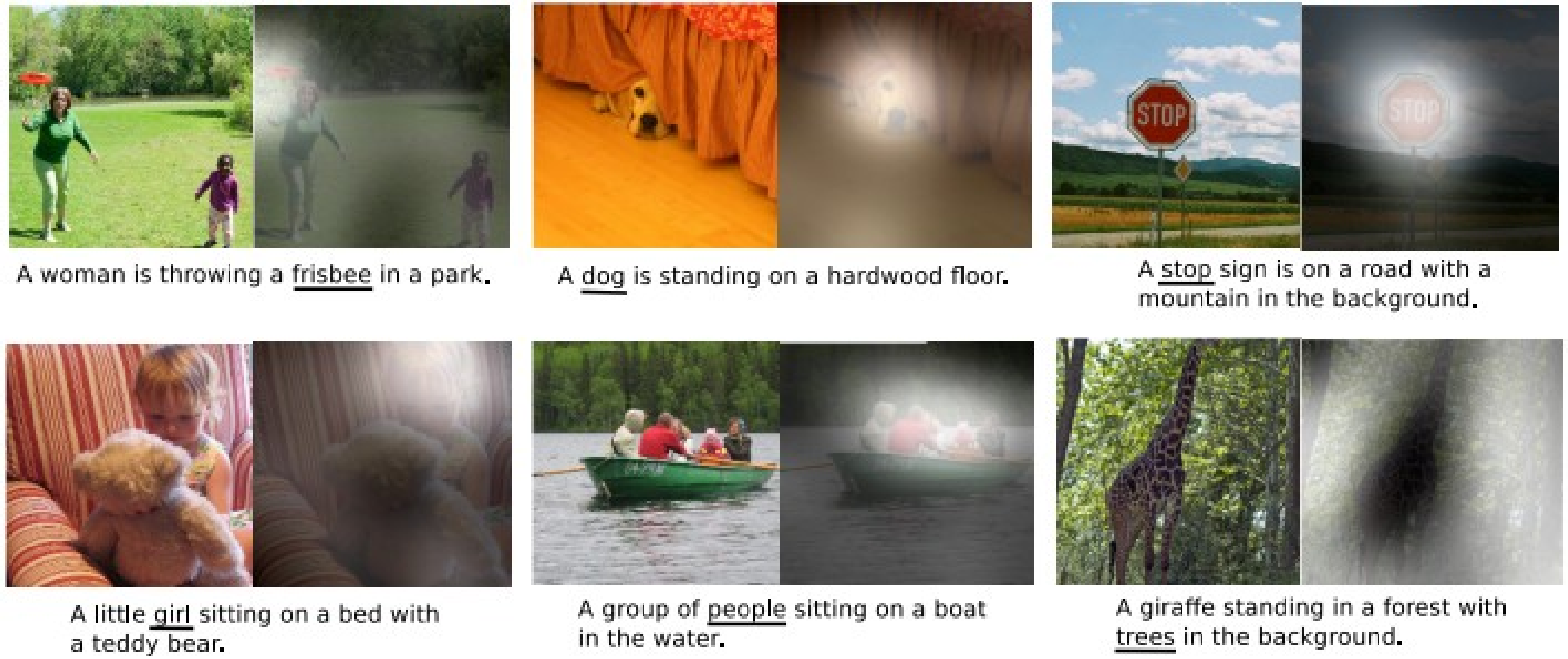
- Show, Attend and Tell
  - Doubly stochastic attention

$$L_d = -\log\left(P(\mathbf{y} \mid \mathbf{x})\right) + \lambda \textstyle\sum_i^L \left(1 - \sum_t^C \alpha_{ti}\right)^2 \ldots(14)$$

Method

- Show, Attend and Tell
  - Model focuses on when generating words

Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)



A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

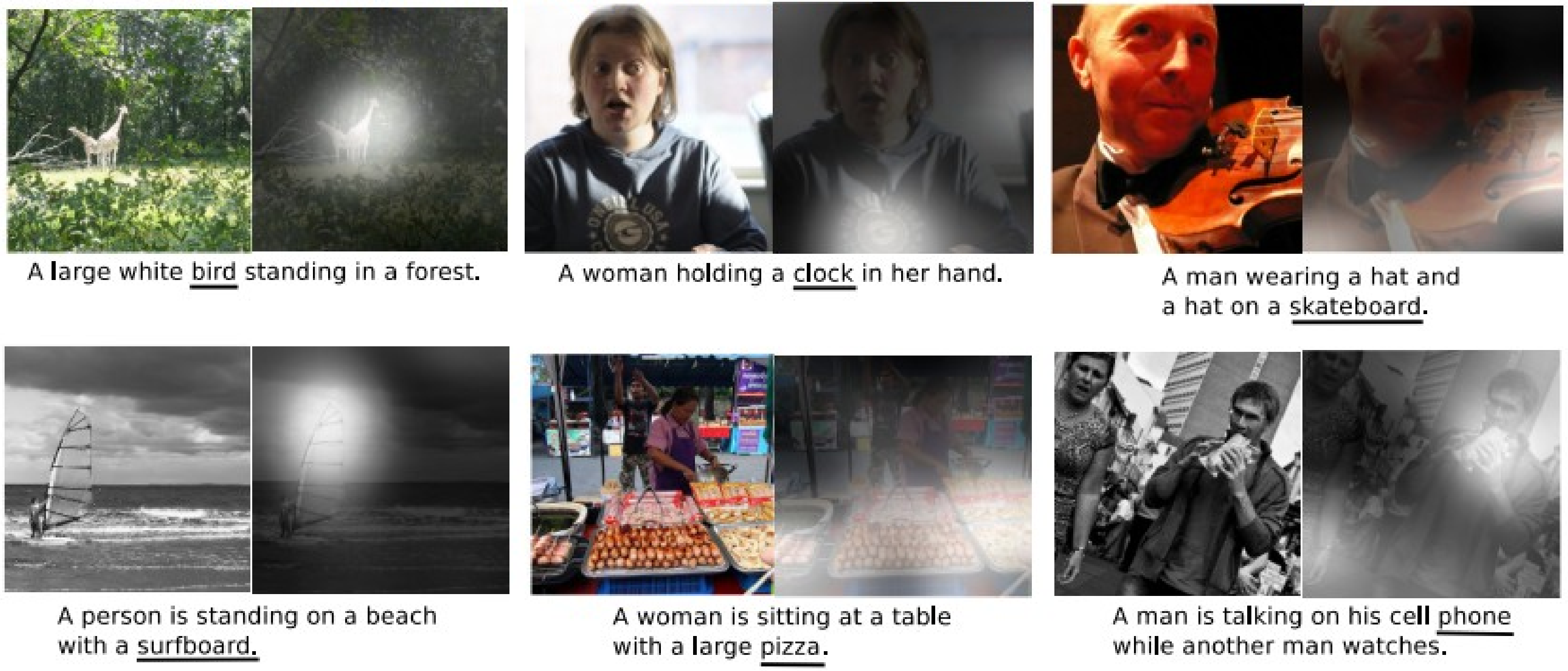
A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Method

- Show, Attend and Tell
  - Spot mistakes or illusions what it's focusing on.

Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.



A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

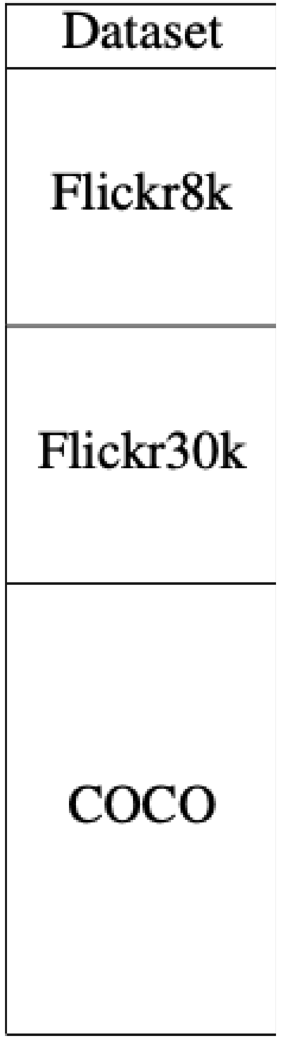
A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.

Experiments

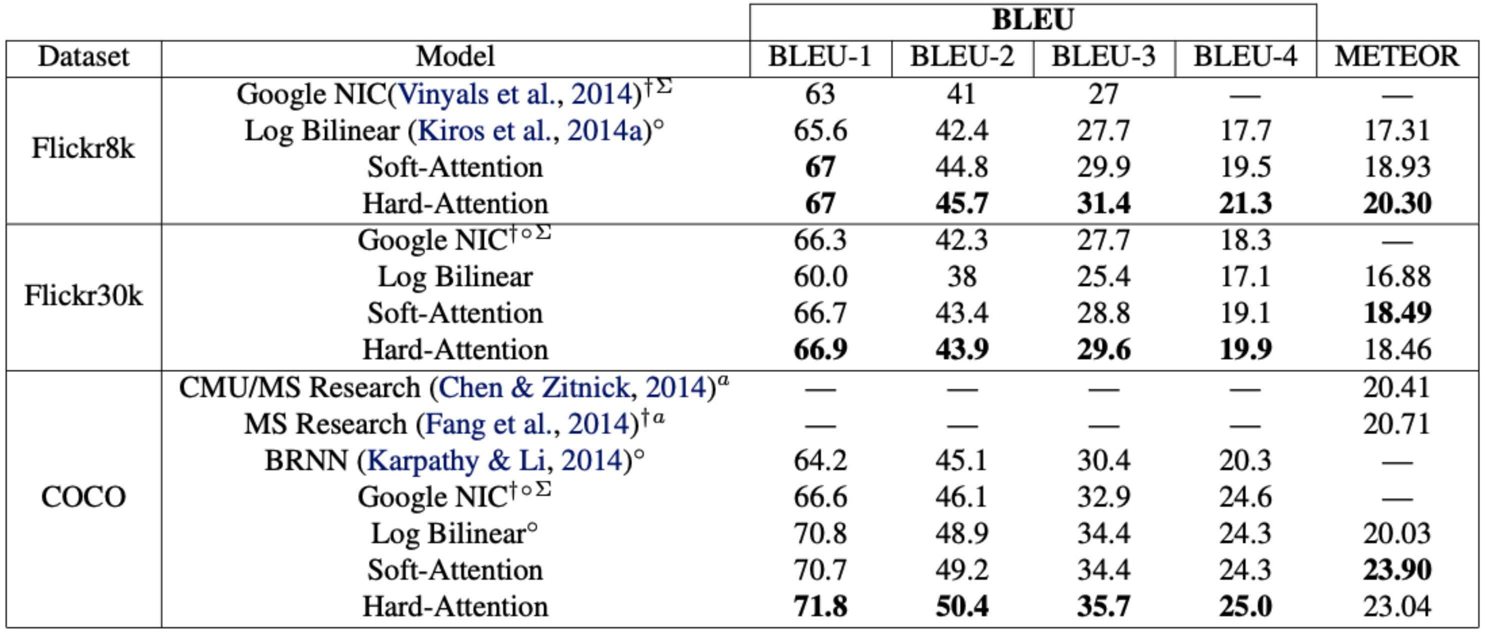
- Dataset
  - Flickr8k : Image 8k, small dataset, 5 caption
  - Flickr30k : Image 30k, big dataset, 5 caption
  - COCO : Image 80k Train set, 40k Test set



Dataset
Flickr8k
Flickr30k
COCO

Experiments

- Dataset
  - BLEU : Measures n-gram overlap
  - METEOR : Considers synonyms, morphology, and word order.



		BLEU				
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

Conclusion

- Show and Tell
  - End to End(CNN + RNN) 모델
  - 중요한 내용을 좀 더 집중해서 문장을 생성 할 수 있음
- Show, Attend and tell
  - Attention 가중치 시각화를 통해 모델이 어떤 부분에 집중했는지 확인 및 내부 결정 과정 판단 가능
  - 기존 이미지 캡션 모델보다 높은 평가지표 기록
- Transformer 기반 모델이 주류이지만 모델과 어텐션을 결합한 초기 시도 중 하나