CVPR 2015

Show and Tell: A Neural Image Caption Generator

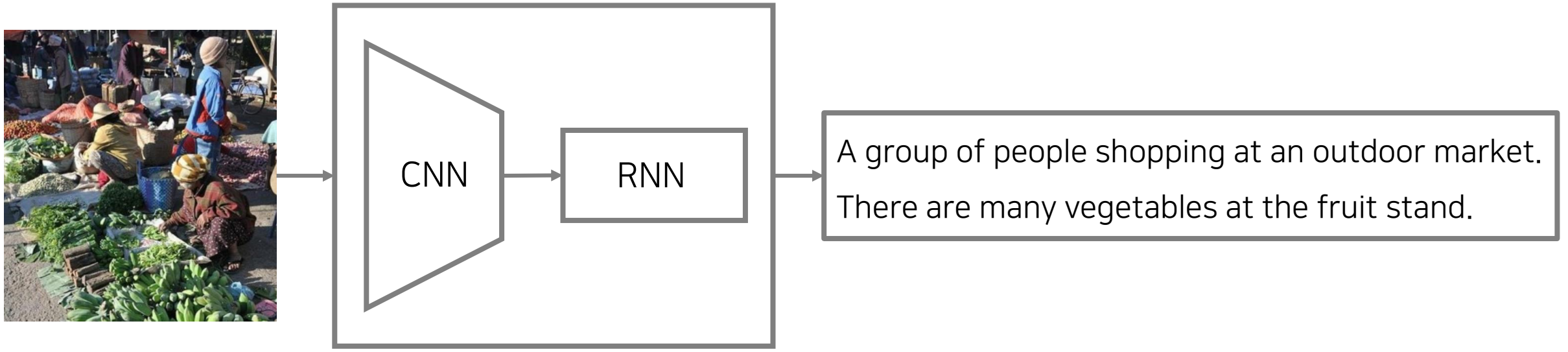
Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan

Google

나동빈

이미지 캡션 생성(Image Caption Generation)이란?

- **이미지를 설명하는 문장을 생성**하는 기술 분야를 의미합니다.
- 대표적인 모델로는 오늘 소개하는 **Neural Image Caption(NIC)**가 있습니다.
  - CNN 네트워크를 이용해 이미지의 특징을 추출한 뒤에 RNN을 거쳐 문장을 생성할 수 있습니다.



이미지 캡션 생성을 "이미지를 번역"하는 문제로 보기

- **입력**: 이미지 I
- **출력**: 목표 문장 $S = \{S_1, S_2, \ldots, S_n\}$

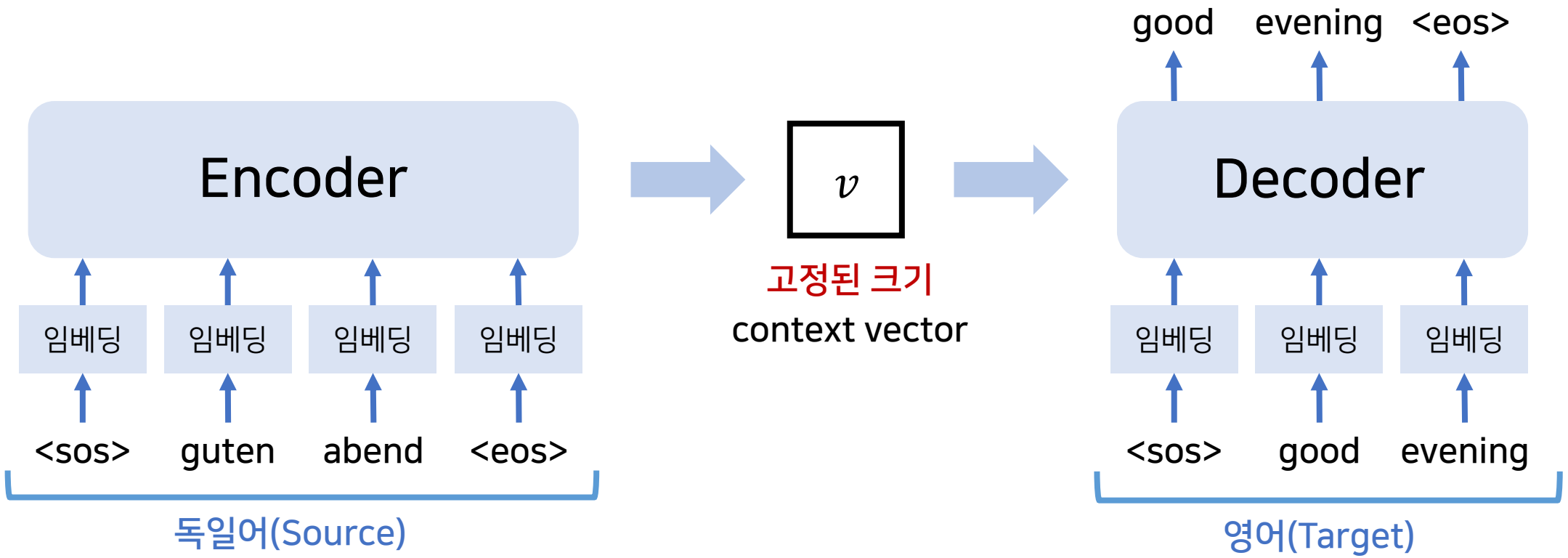
A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

- 따라서 가능도(likelihood) $P(S|I)$를 **최대화(maximization)**하는 문제로 정의할 수 있습니다.

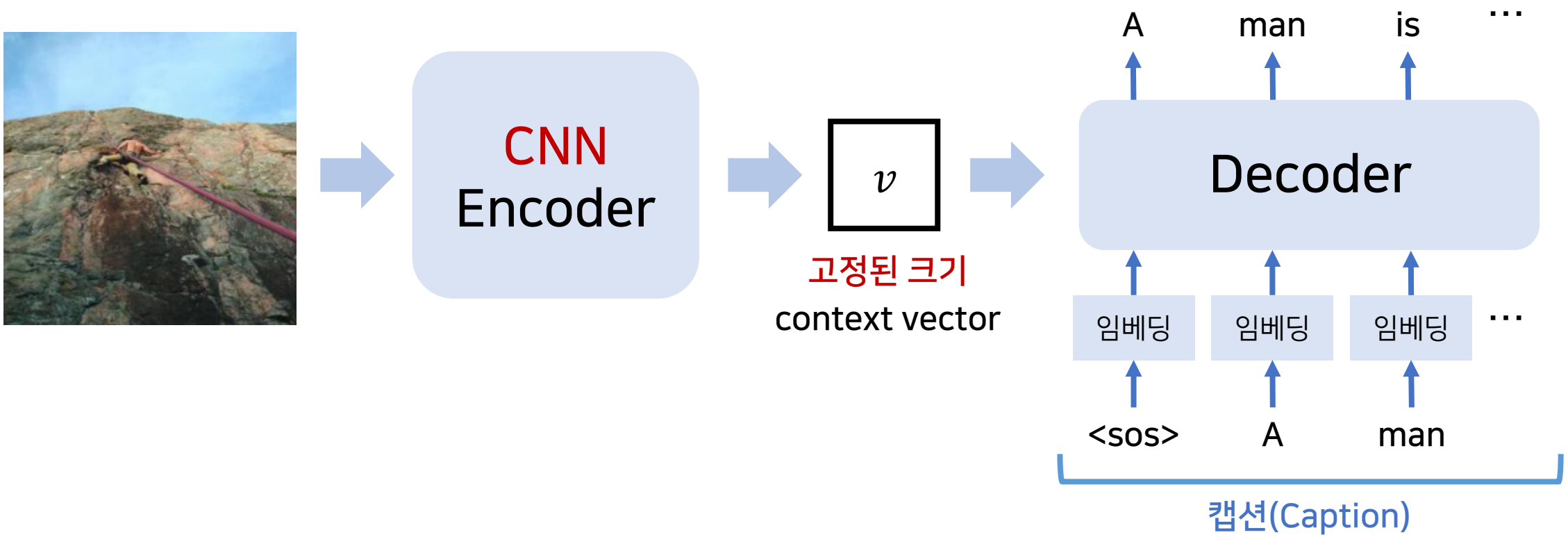
이미지 캡션 생성과 기계 번역의 공통점

- 기계 번역에서는 $P(T|S)$를 **최대화(maximizing)**합니다.
  - 소스 문장(source sentence)을 대표하는 하나의 **문맥 벡터(context vector)를 이용**합니다.



이미지 캡션 생성과 기계 번역의 공통점

- 기계 번역 작업에서의 **인코더(encoder)를 CNN으로 대체**하여 이미지 캡션을 생성할 수 있습니다.



공식(Formulation)

- 하나의 이미지를 **설명(description)**으로 변역(translation)하는 작업에 비유할 수 있습니다.

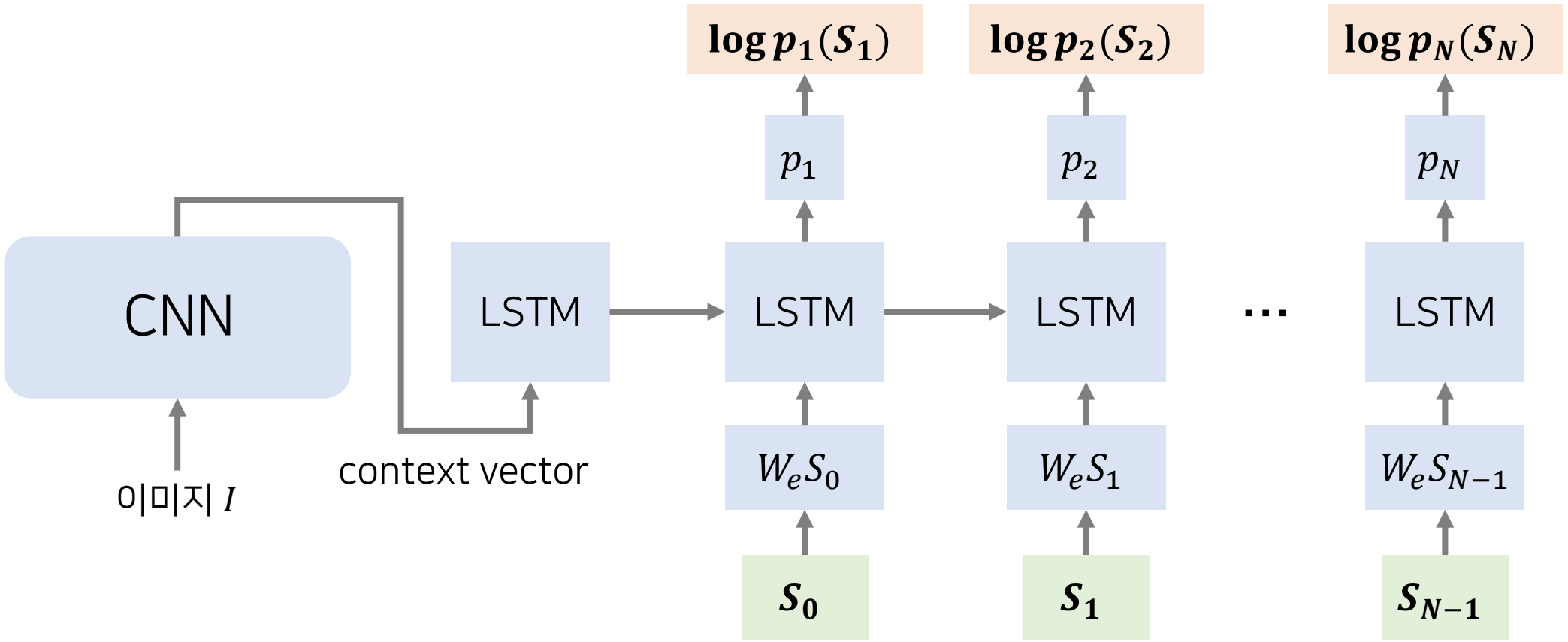
$$\theta^* = arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta)$$

- **연쇄 법칙(chain-rule)**을 이용해 다음의 식으로 전개할 수 있습니다.

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

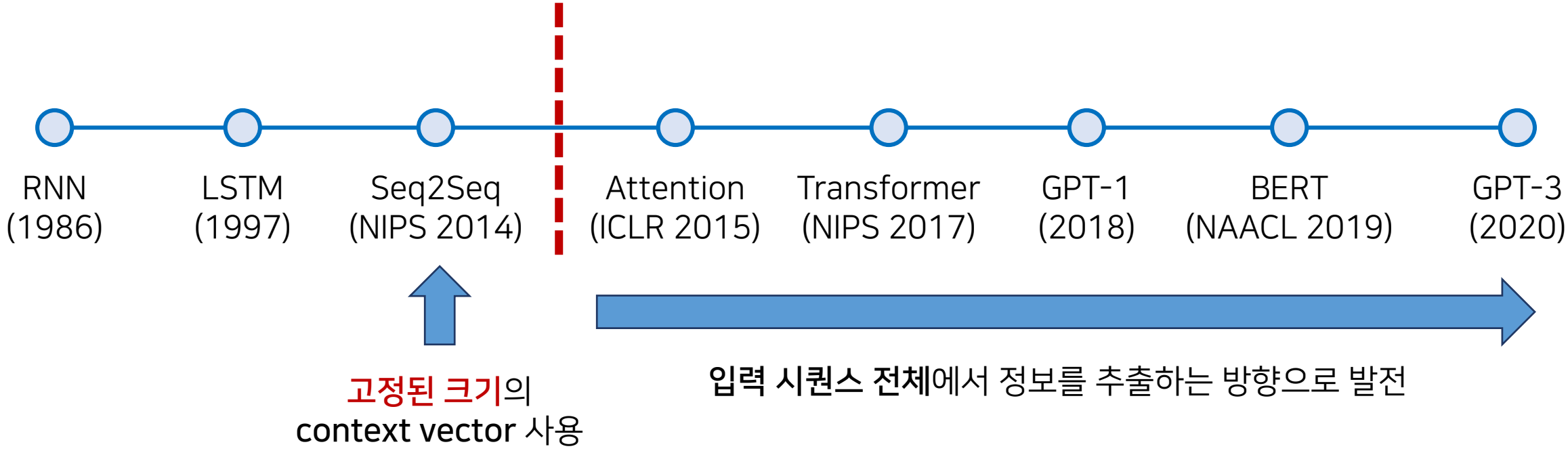
NIC(Neural Image Caption) 아키텍처

- 하나의 이미지를 **설명(description)**으로 변역(translation)하는 작업에 비유할 수 있습니다.



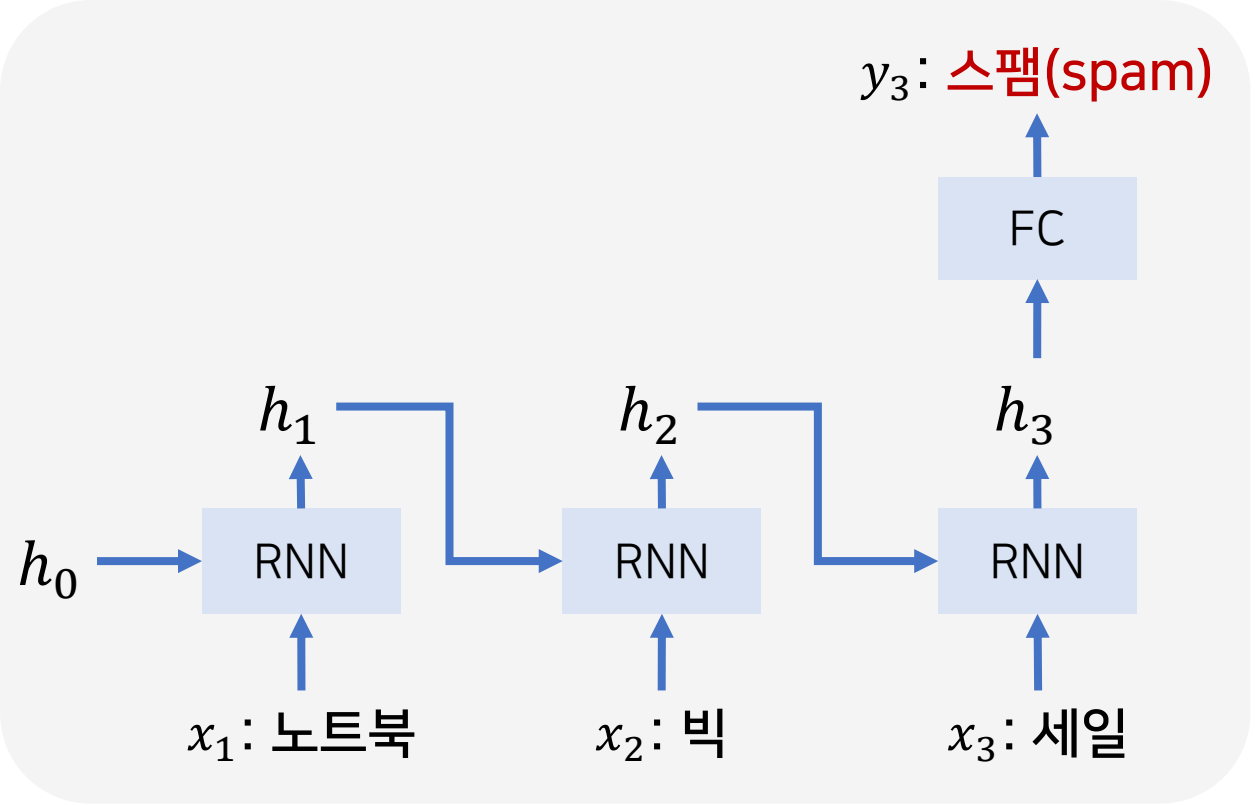
[배경지식] 딥러닝 기반의 언어 모델 발전 과정

- 딥러닝 기반의 언어 모델의 발전 과정은 다음과 같습니다.
  - 오늘 소개하는 논문에서는 LSTM을 활용한 **Seq2Seq와 유사한 아키텍처를 사용**합니다.
  - 2021년 기준으로 최신 고성능 모델들은 **Transformer 아키텍처를 기반으로** 하고 있습니다.



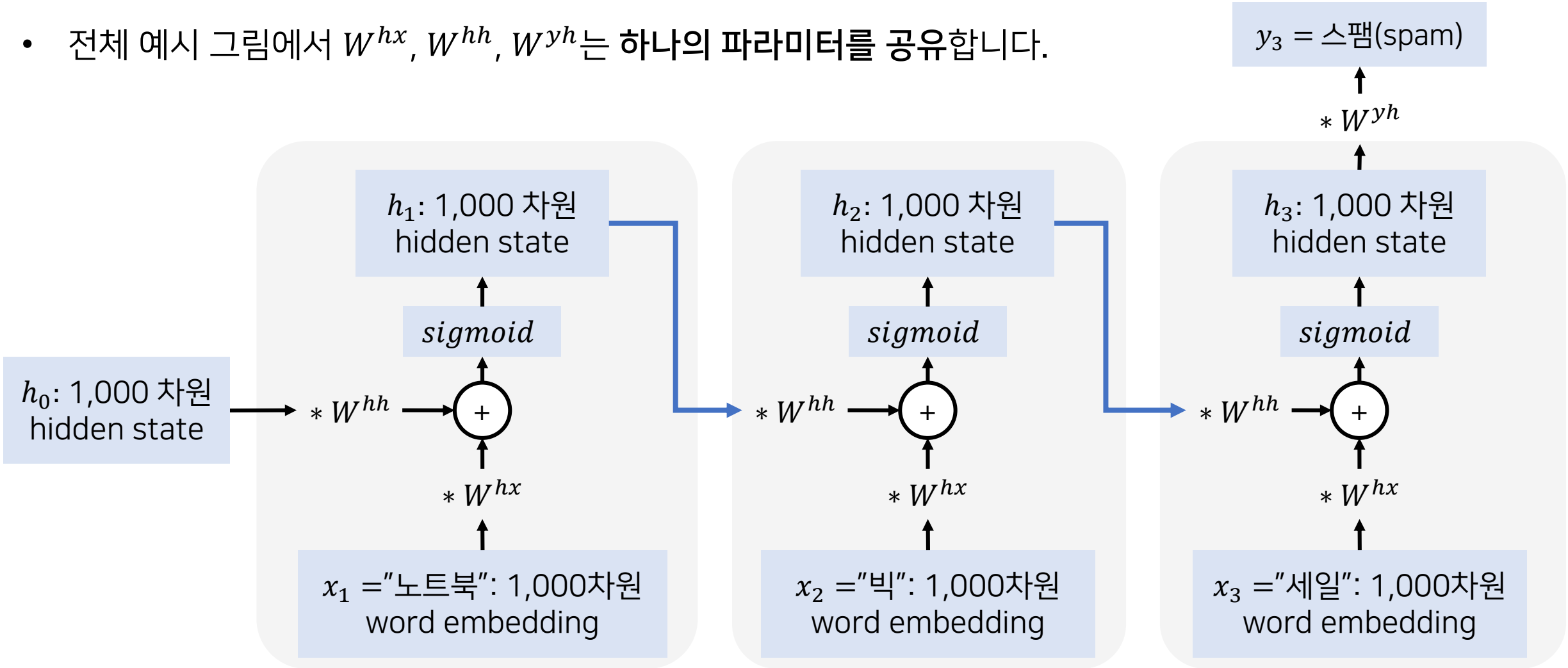
RNN 자세히 알아보기

- 입력(input)
  - x_t = 각각의 입력 단어
- 히든 상태(hidden state)
  - $h_t = sigmoid\left(W^{hx}x_t + W^{hh}h_{t-1}\right)$
- 출력(output)
  - $y_t = W^{yh}h_t$



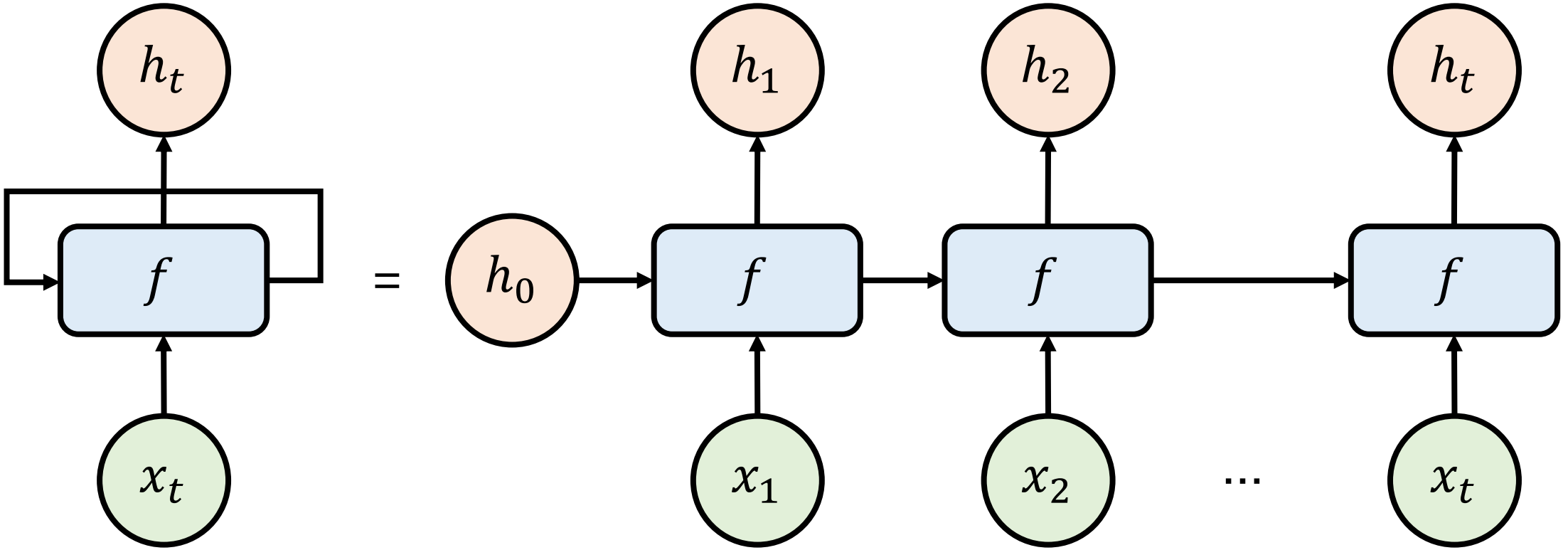
RNN 자세히 알아보기

- 전체 예시 그림에서 W^{hx}, W^{hh}, W^{yh}는 **하나의 파라미터를 공유**합니다.



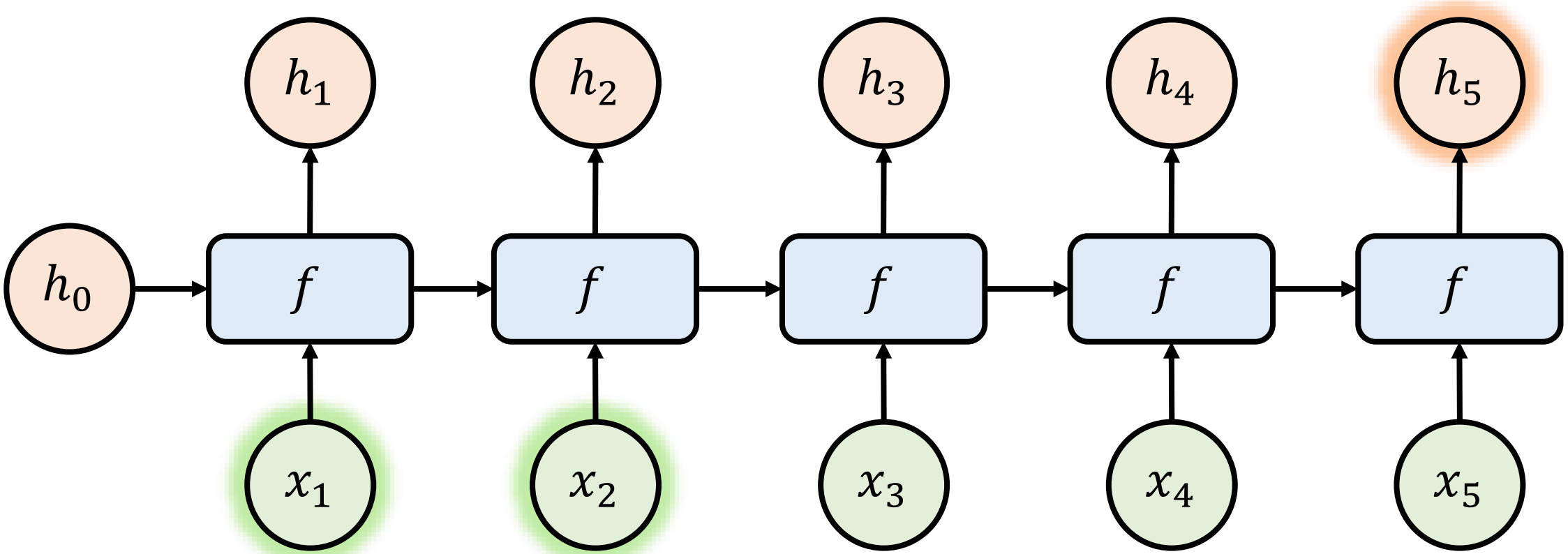
RNN의 한계점

- 이론적으로는 RNN을 이용하여 긴 길이의 순차적인 데이터를 효과적으로 처리할 수 있습니다.

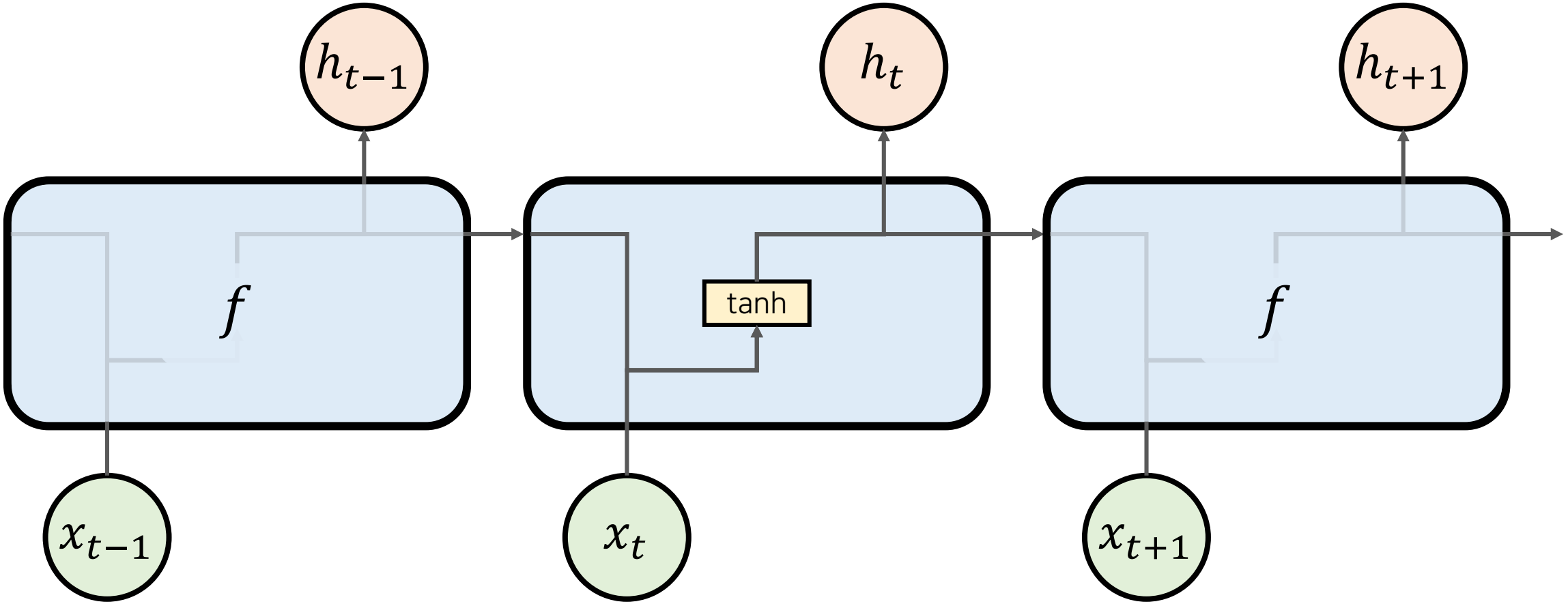


RNN의 한계점

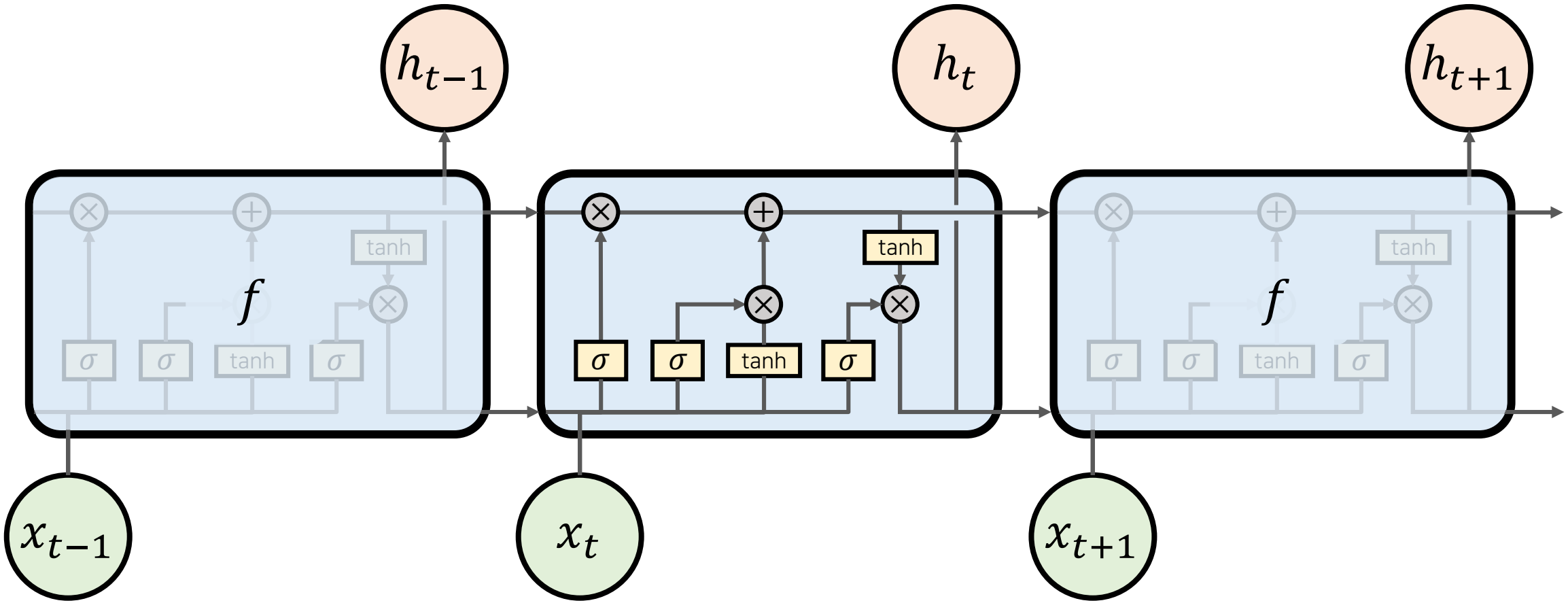
- 실제로는 토큰(token) 사이의 거리가 먼 경우 연속적인 정보가 잘 전달되지 않을 수 있습니다.



RNN(Recurrent Neural Network) 아키텍처

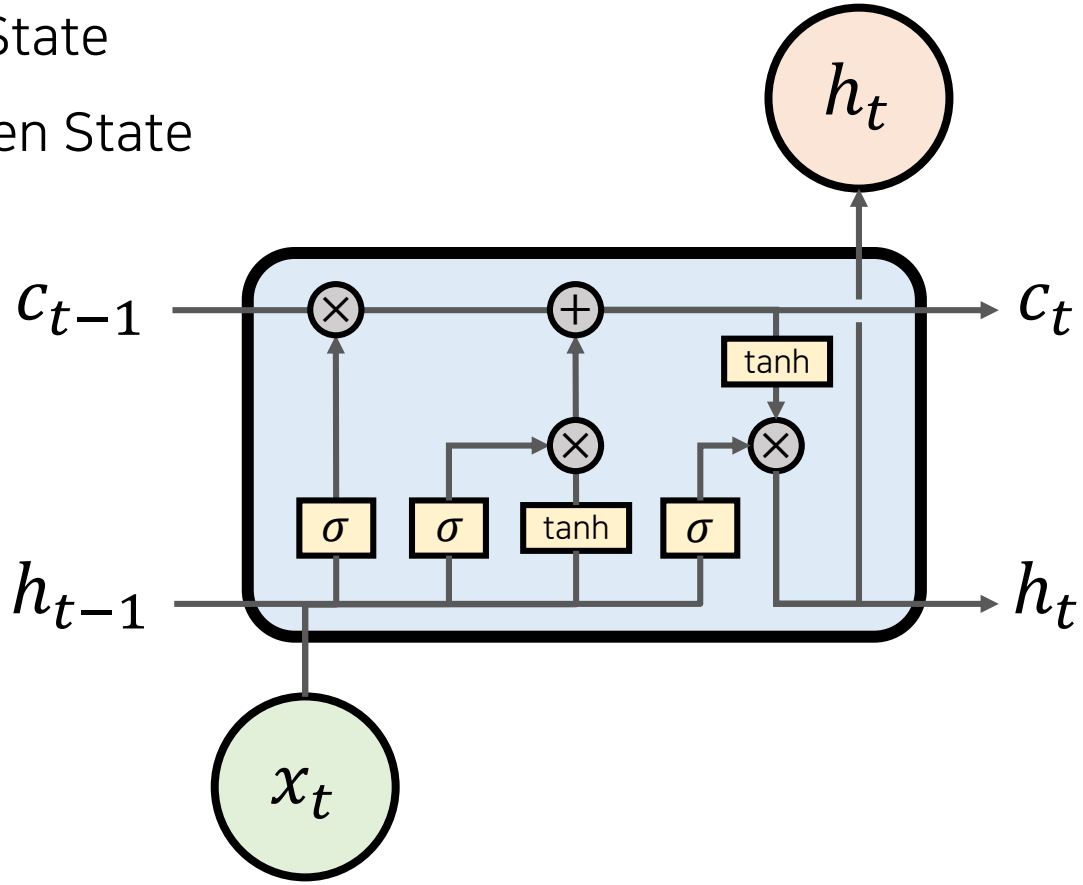


LSTM(Long Short-Term Memory) 아키텍처



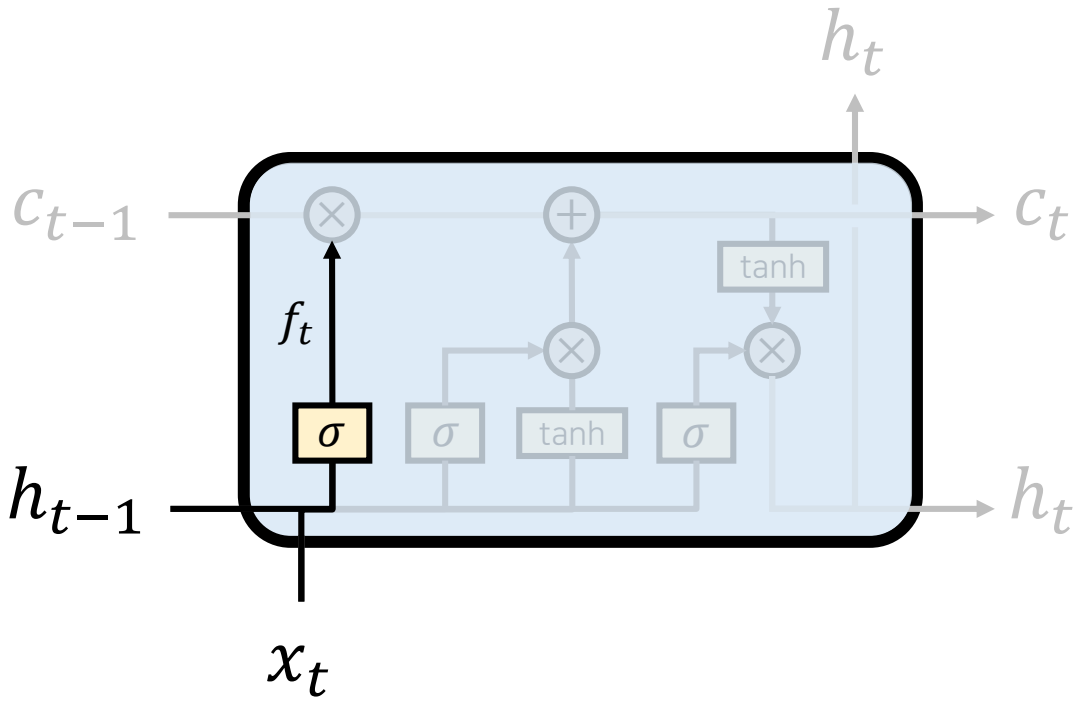
LSTM 핵심 아이디어: 두 개의 상태 정보

- LSTM은 RNN과는 다르게 **두 개의 상태 정보**를 저장하고 처리합니다.
  - **장기 기억**: Cell State
  - **단기 기억**: Hidden State



LSTM 핵심 아이디어: 게이트(Gates)

- Forget Gate는 어떠한 정보를 잊게 만들지 결정하는 레이어입니다.
  - 오래된 정보 중에서 필요 없는 정보는 잊게 됩니다.

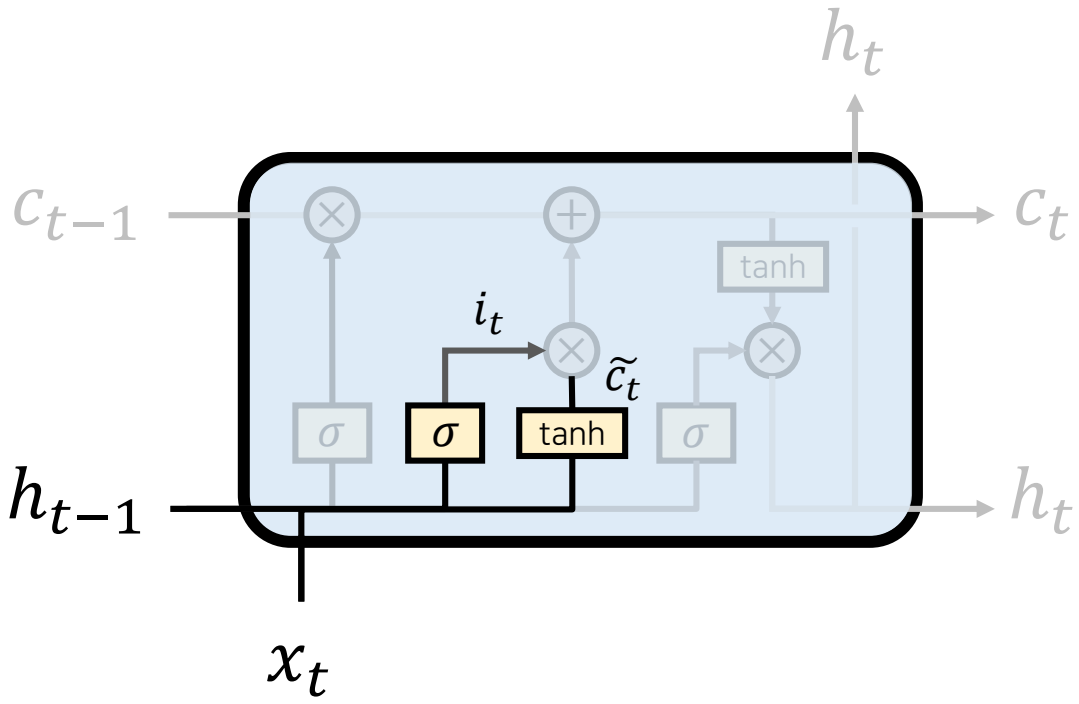


Formulation

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1})$$

LSTM 핵심 아이디어: 게이트(Gates)

- Input Gate는 새로운 정보를 장기 기억(Cell State)에 반영하는 역할을 수행합니다.
  - 새롭게 특정한 정보를 기억하도록 만듭니다.



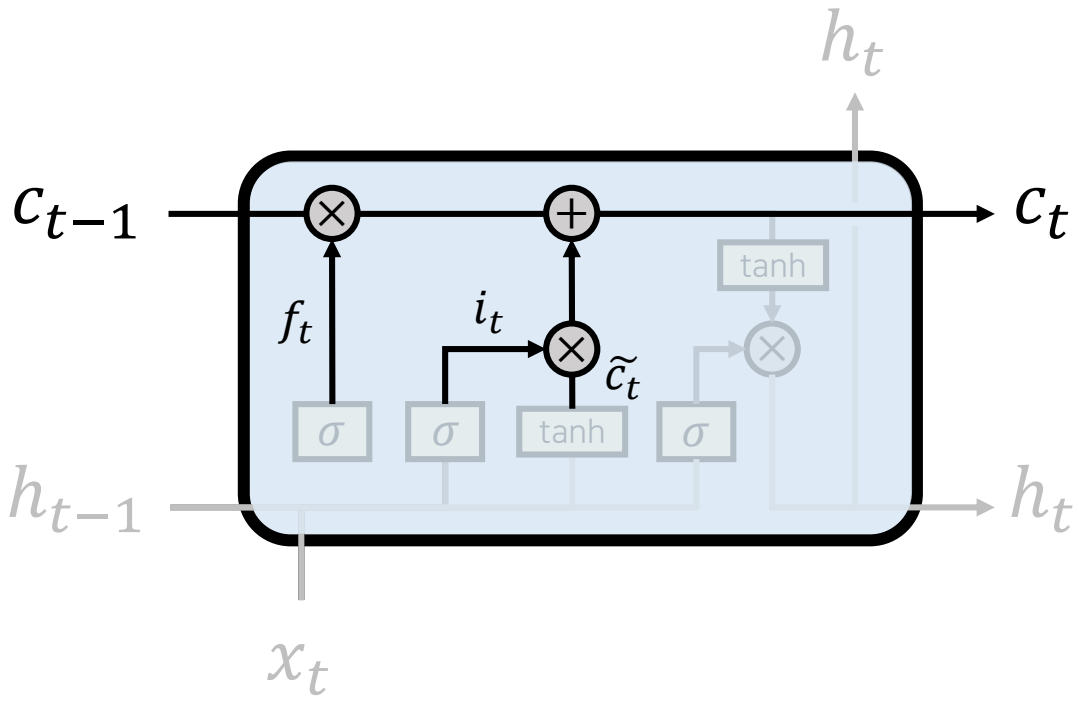
Formulation

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1})$$

$$\tilde{c}_t = tanh(W_{cx}x_t + W_{ch}h_{t-1})$$

LSTM 핵심 아이디어: 게이트(Gates)

- **장기 기억(Cell State)**은 Forget Gate와 Input Gate를 이용하여 업데이트됩니다.

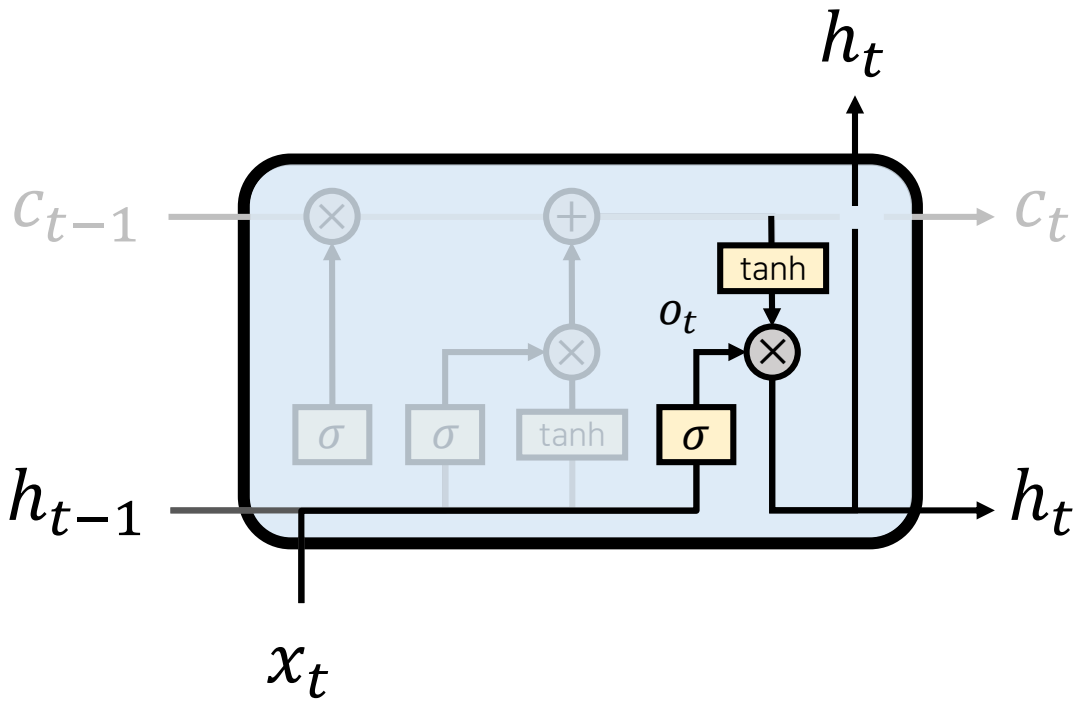


Formulation

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$$

LSTM 핵심 아이디어: 게이트(Gates)

- **Output Gate**는 장기 기억과 현재의 데이터를 이용해 단기 기억(Hidden State)을 갱신합니다.



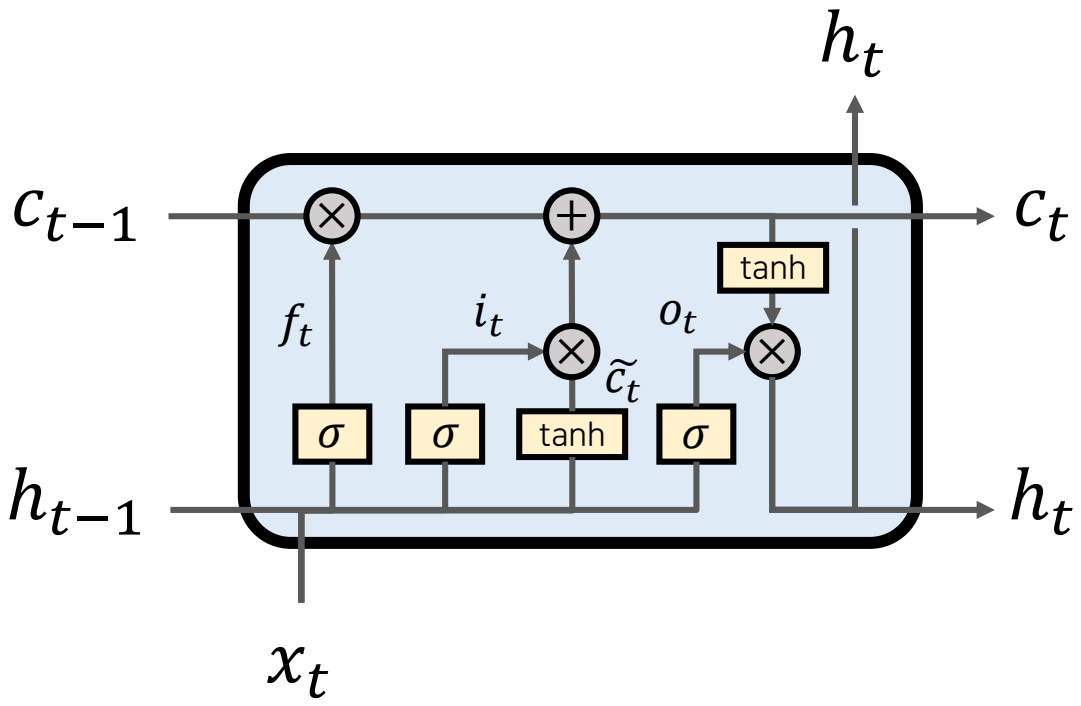
Formulation

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1})$$

$$h_t = o_t * tanh(c_t)$$

LSTM 전체 공식

- LSTM 전체 공식은 다음과 같습니다.
  - 공식에 등장하는 **모든 가중치(weight)는 공유**됩니다.



Formulation

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1})$$

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1})$$

$$\tilde{c}_t = tanh(W_{cx}x_t + W_{ch}h_{t-1})$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1})$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$$

$$h_t = o_t * tanh(c_t)$$

생성 결과 (Generation Results)

- 대표적인 평가 지표를 이용해 NIC를 이용해 생성된 결과를 평가한 것은 다음과 같습니다.

Metric	BLEU-4	METEOR	CIDER
NIC	**27.7**	**23.7**	**85.5**
Random	4.6	9.0	5.1
Nearest Neighbor	9.9	15.7	36.5
Human	21.7	25.2	85.4

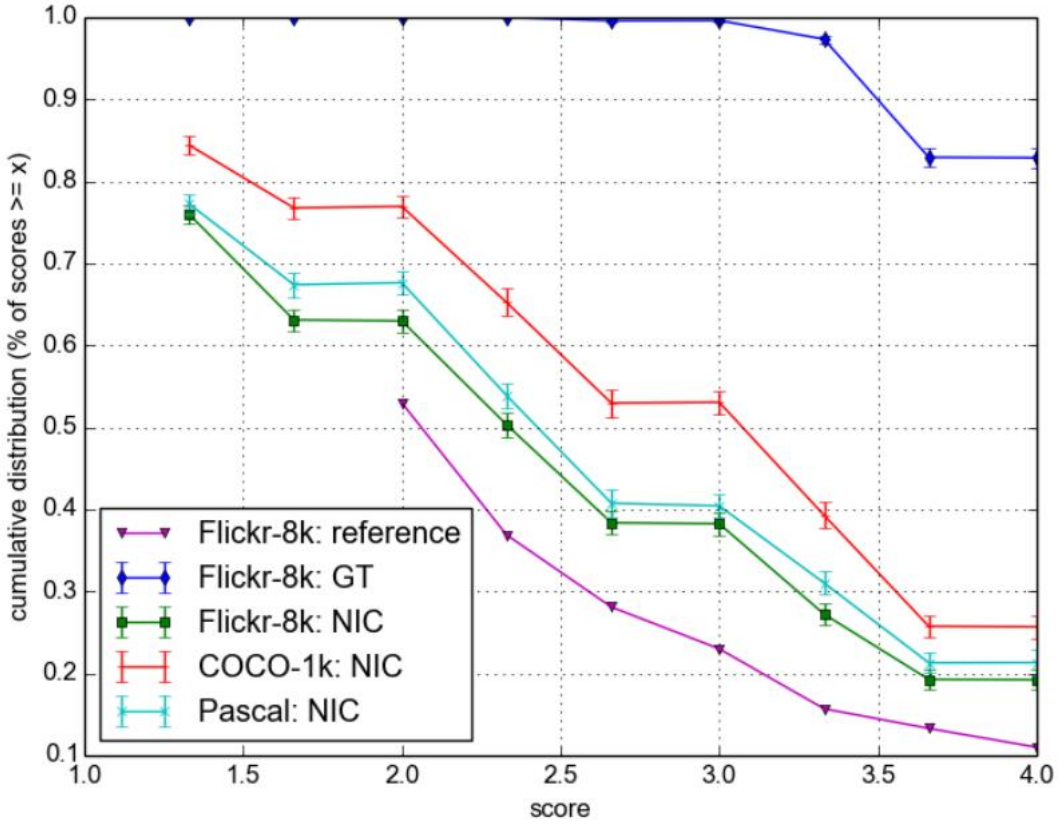
[Table] Scores on the MSCOCO development set.

Approach	PASCAL (xfer)	Flickr 30k	Flickr 8k	SBU
Im2Text [24]				11
TreeTalk [18]				19
BabyTalk [16]	25			
Tri5Sem [11]			48	
m-RNN [21]		55	58	
MNLM [14][5]		56	51	
SOTA	25	56	58	19
NIC	**59**	**66**	**63**	**28**
Human	69	68	70	

[Table] BLEU-1 scores. Authors only report previous work results when available.

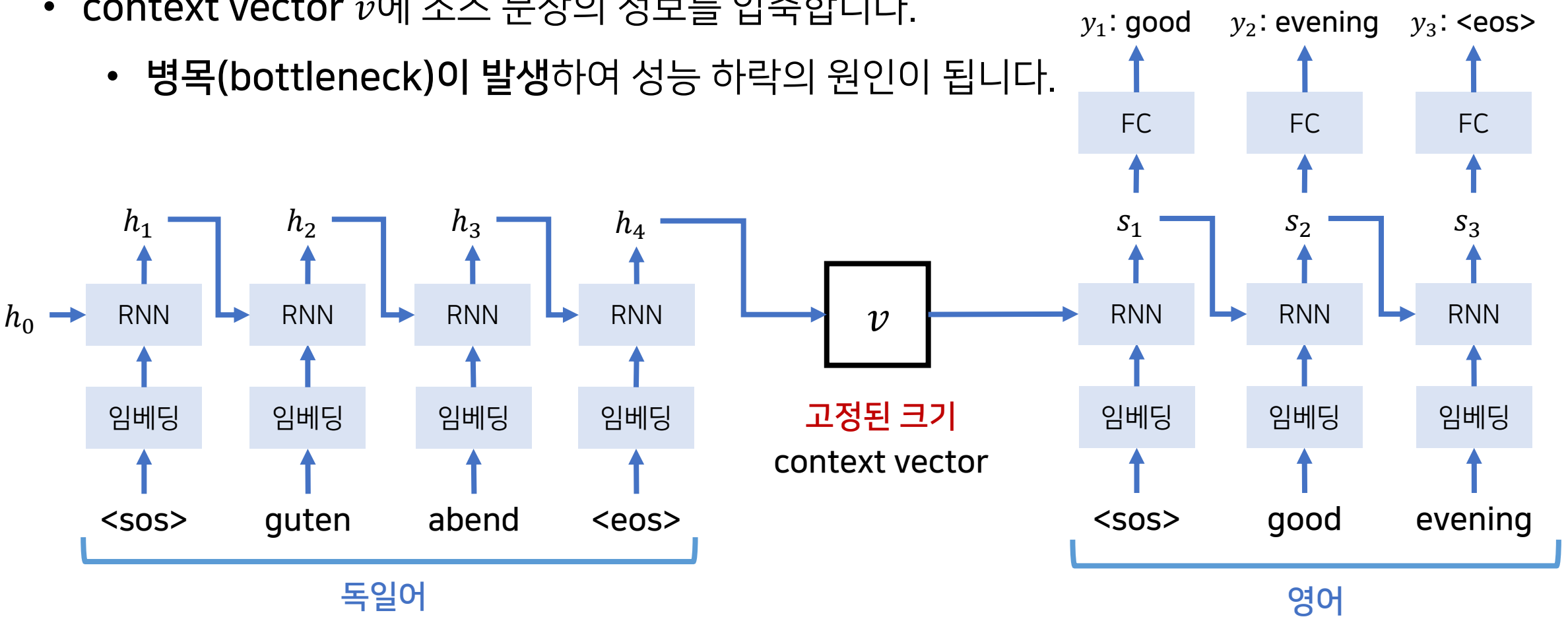
Human Evaluation

- **사람**이 1점(worst)부터 4점(best)까지의 점수로 평가한 결과는 다음과 같습니다.
  - 실제 정답(GT)에 비하면 매우 점수가 낮습니다.
  - 그래도 이전까지의 모델보다 성능이 뛰어납니다.



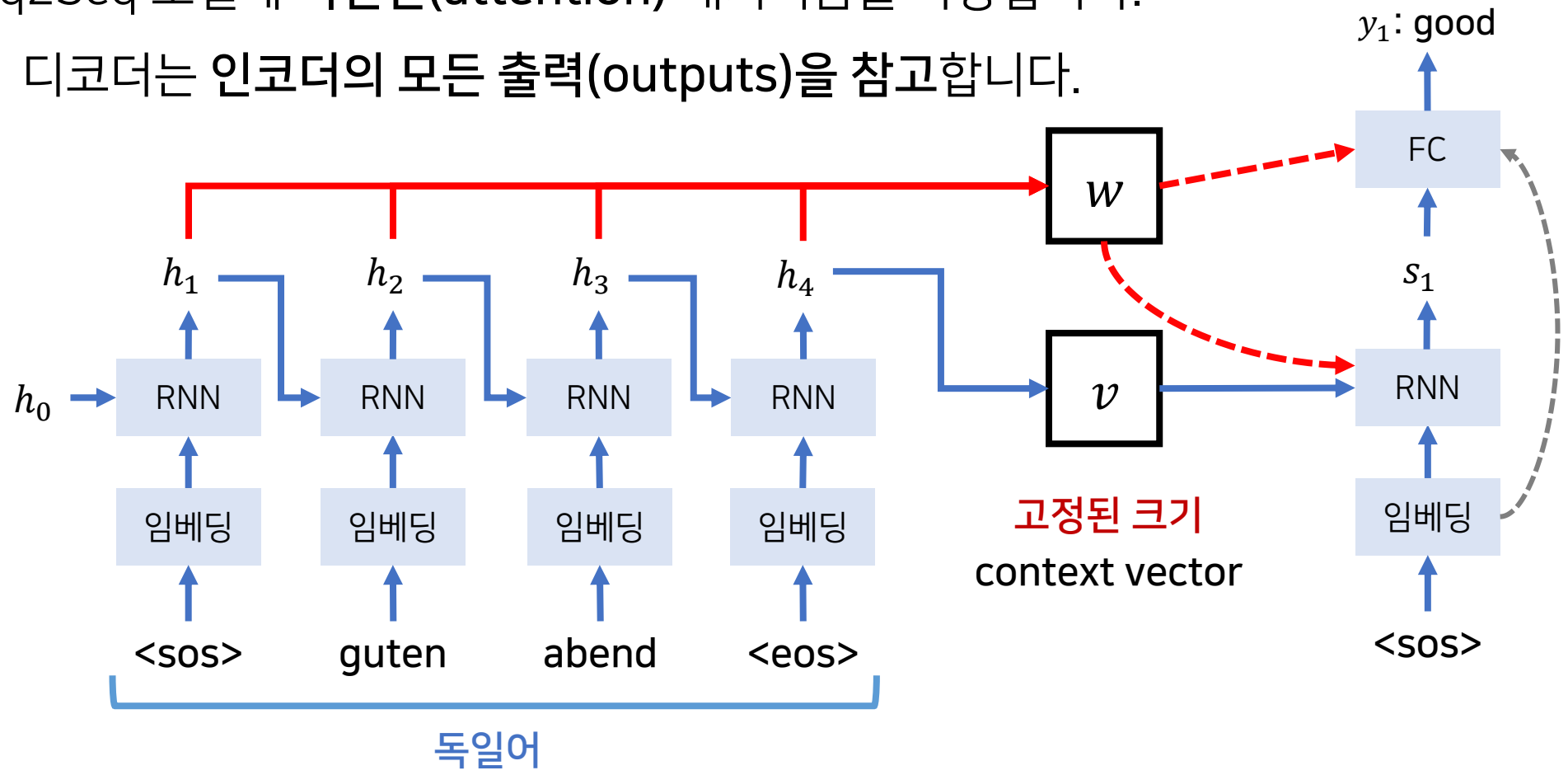
[후속 연구] 기존 Seq2Seq 모델들의 한계점

- context vector v에 소스 문장의 정보를 압축합니다.
  - 병목(bottleneck)이 발생하여 성능 하락의 원인이 됩니다.



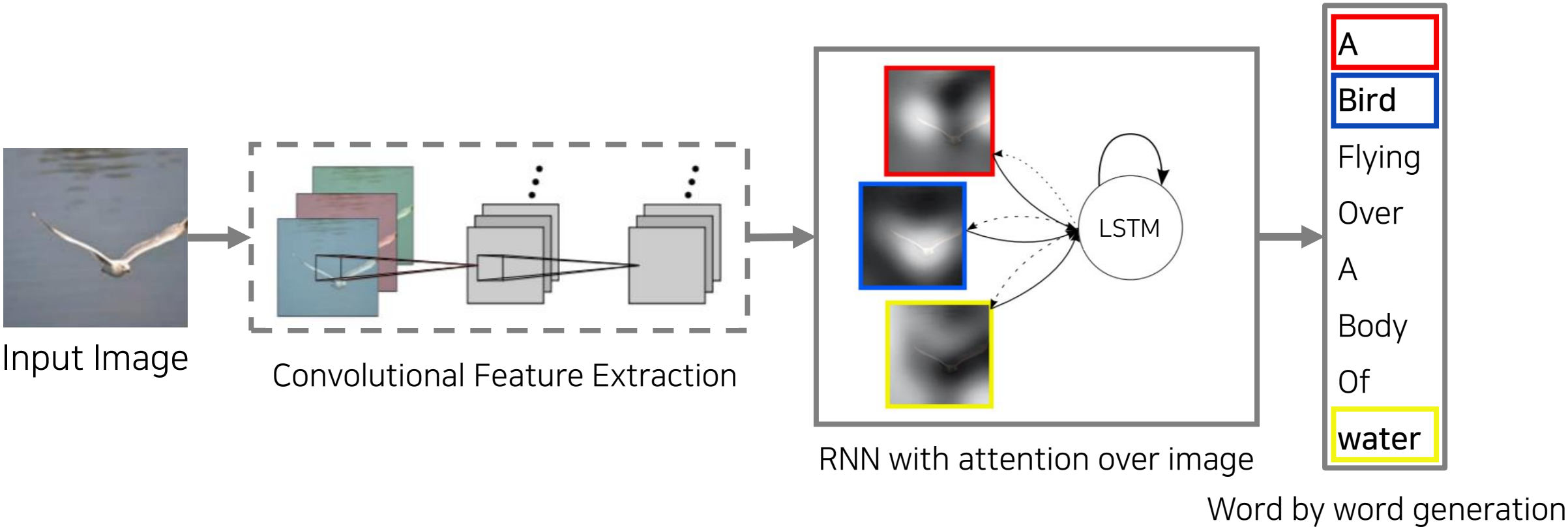
[후속 연구] Seq2Seq with Attention

- Seq2Seq 모델에 **어텐션(attention)** 매커니즘을 사용합니다.
  - 디코더는 **인코더의 모든 출력(outputs)을 참고**합니다.



[후속 연구] Neural Image Caption Generation with Visual Attention (ICML 2015)

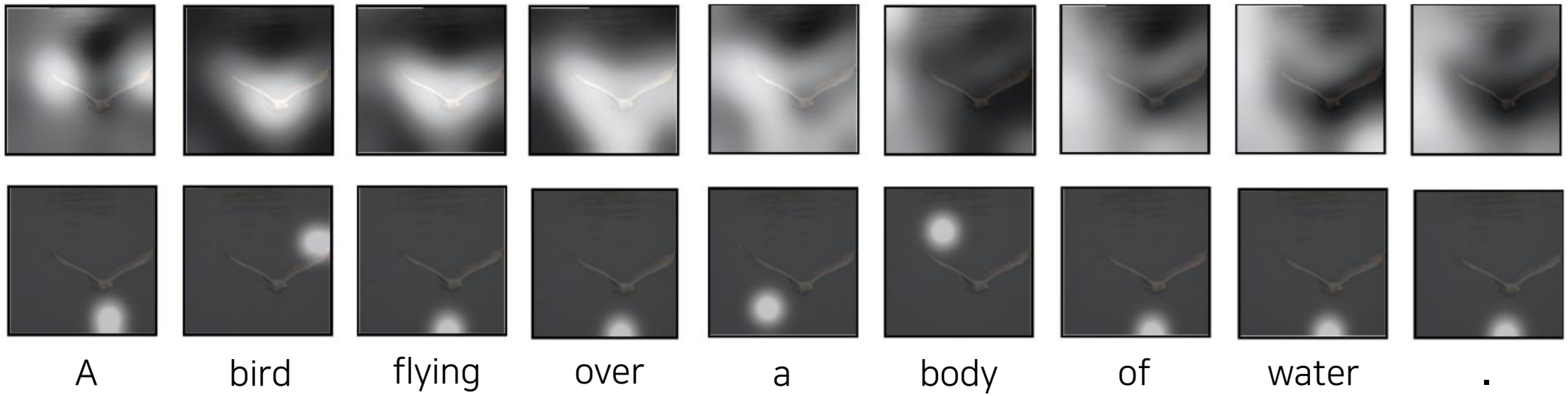
- 본 논문에서는 **Neural Image Caption(NIC)** 네트워크에 어텐션(attention) 기법을 적용합니다.
  - RNN을 사용할 때 이미지 전체에 대한 **어텐션(attention)** 정보를 활용합니다.



[후속 연구] Neural Image Caption Generation with Visual Attention (ICML 2015)

- 차례대로 단어(word)를 하나씩 생성할 때의 어텐션(attention)을 시각화한 것은 다음과 같습니다.
  - 본 논문에서는 ① **소프트(soft) 어텐션**과 ② **하드(hard) 어텐션**을 제안합니다.

: 입력 이미지(Input Image)



[후속 연구] Neural Image Caption Generation with Visual Attention (ICML 2015)

- **소프트(soft) 어텐션**의 다양한 예시를 확인해 봅시다.

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.