Language and Vision

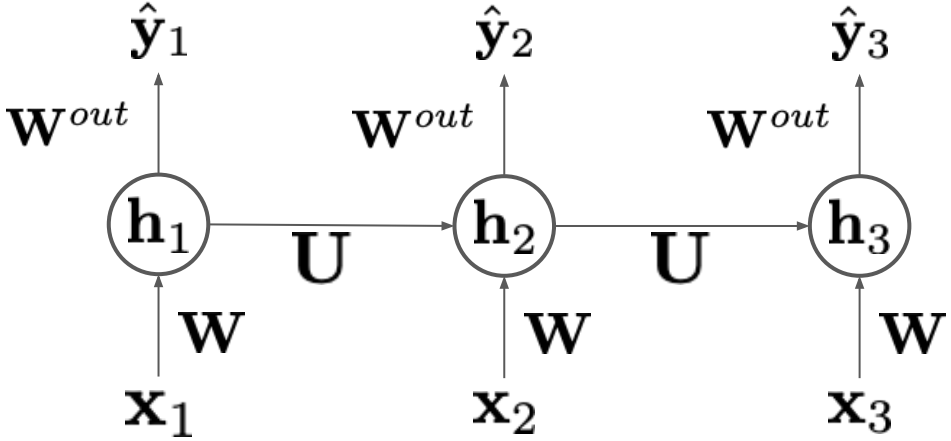
Instructor: Seunghoon Hong

Course logistics

- We will release the midterm score this week
  - The claim sessions will be announced later.
  - We do not release the correct answers.

- The third assignment will be released today
  - **Due date:** Midnight Nov. 3rd
  - **Quiz:** Nov. 6th (in class)

Recap: (Vanilla) Recurrent Neural Network



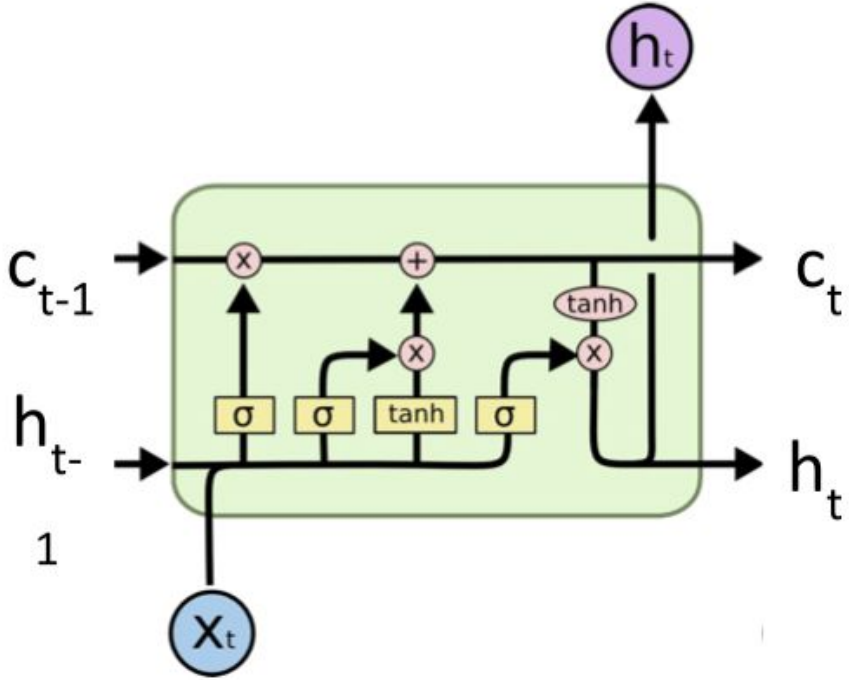
In general, for any $t \geq 1$,

$$\mathbf{h}_t = \sigma(\mathbf{U}\mathbf{h}_{t-1} + \mathbf{W}\mathbf{x}_t + \mathbf{b})$$

$$\hat{\mathbf{y}}_t = \mathbf{W}^{out}\mathbf{h}_t$$

$$\mathbf{h}_0 = \mathbf{0}$$

Recap: Long-Short Term Memory



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

Today’s agenda

- Language modeling using RNNs
- Image captioning
  - Naive image captioning, image captioning with attention
- Visual question answering
  - Naive visual question answering, memory network

Today’s agenda

- **Language modeling using RNNs**
- Image captioning
  - Naive image captioning, image captioning with attention
- Visual question answering
  - Naive visual question answering, memory network

Modeling language

seq2seq

Sentence generation

f_θ → I am very hungry at the end of every class

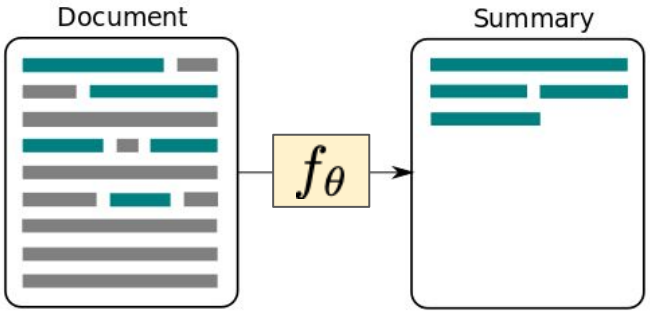
Machine translation

The agreement on the European Economic Area was signed in August 1992.

f_θ

L'accord sur l'Espace économique européen a été signé en août 1992.

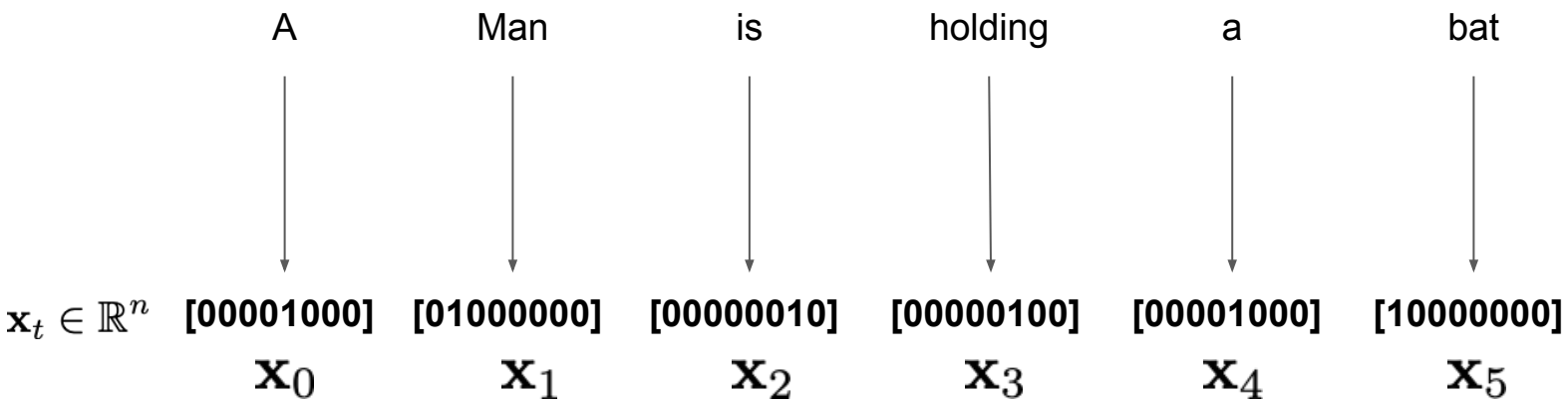
Text summarization



And so many more...

Modeling language

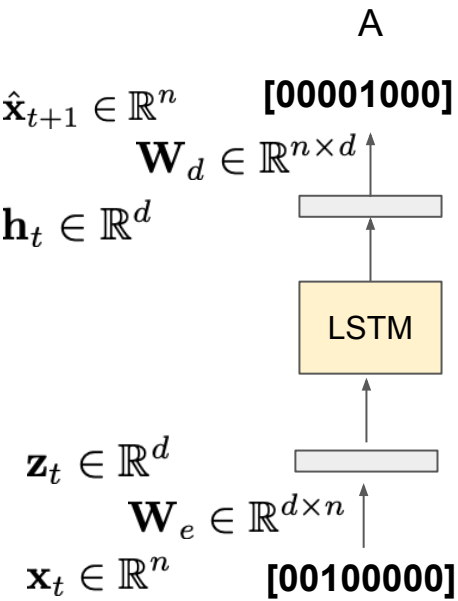
- Sentence = a sequence of discrete symbols



One-hot encoding of discrete symbols (tokenization)

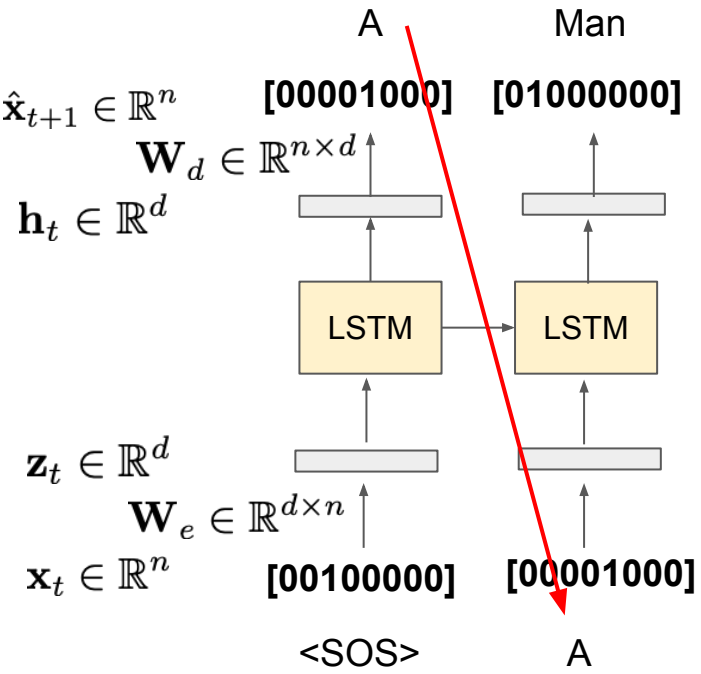
RNN as a language model

- Sentence generation = predicting a next token



RNN as a language model

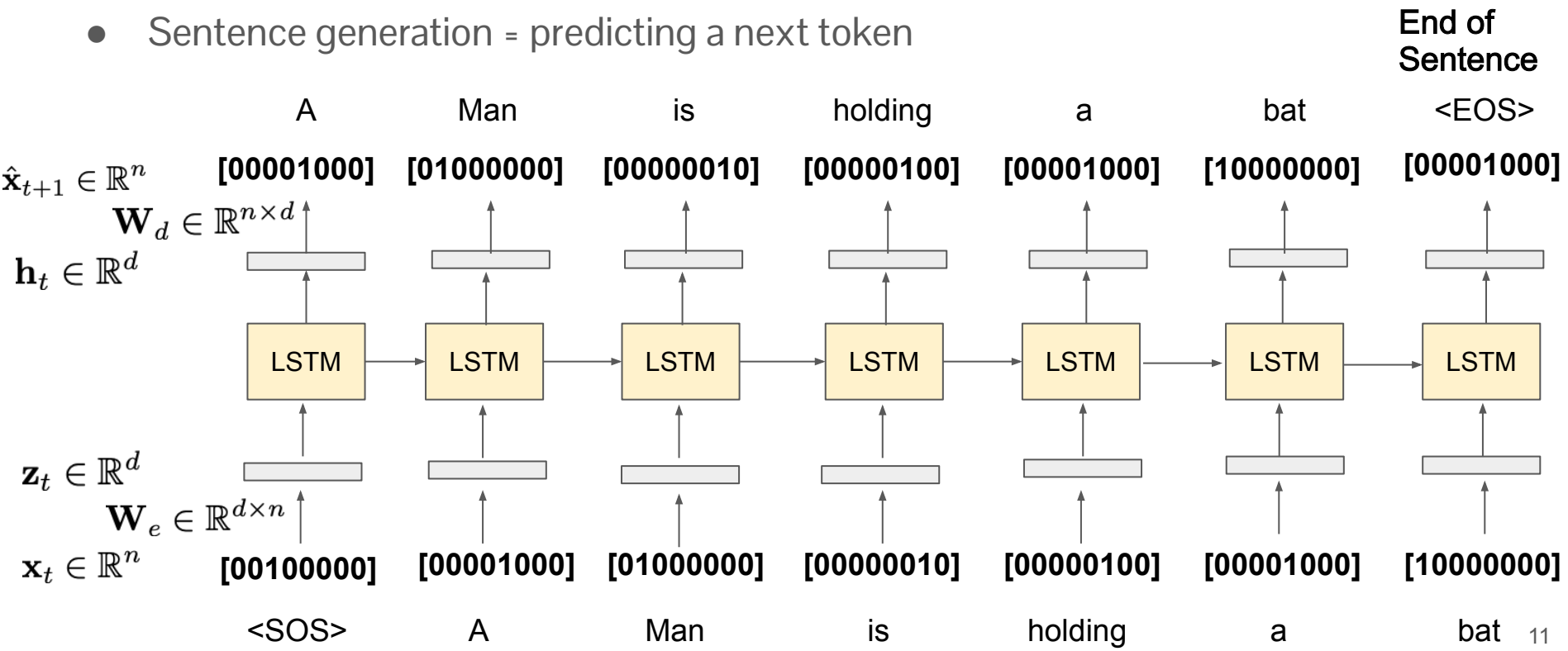
- Sentence generation = predicting a next token



Note: accumulated!! 현재 input주는 거만 영향 주는것은 아니다

RNN as a language model

- Sentence generation = predicting a next token

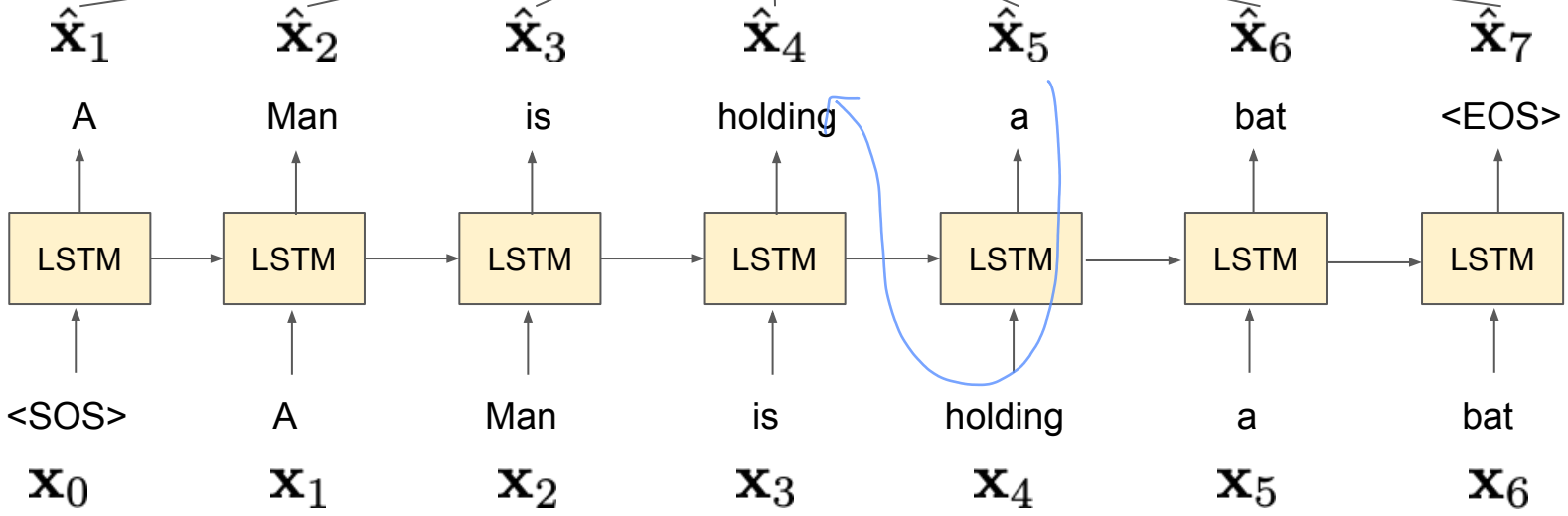


Training: RNN-based language model

P1: Error accumulated
P2: Gradient flow through input to the output. But sampling from the probability vector (making one hot encoding) is not differentiable, i.e., cannot backprop.

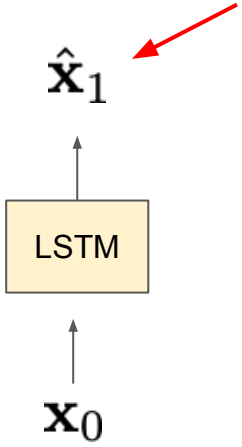
$$\sum_{t=1}^{T} \mathcal{L}(\hat{\mathbf{x}}_t, \mathbf{x}_t)$$

Cross-entropy loss (word-level classification)

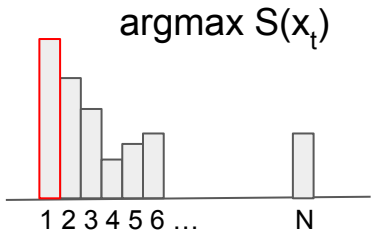


S: We feed ground-truth words as inputs (also known as **teacher forcing**)

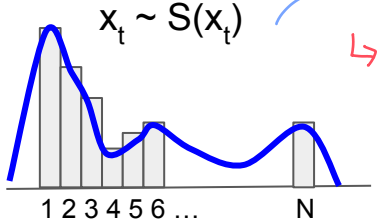
Inference: RNN-based language model



- **For each step, sample a word from the output score**
- **Sampling methods:**
  - Take the word with maximum score (greedy, deterministic)
  - Sample a word according to score probability (stochastic)



Greedy method

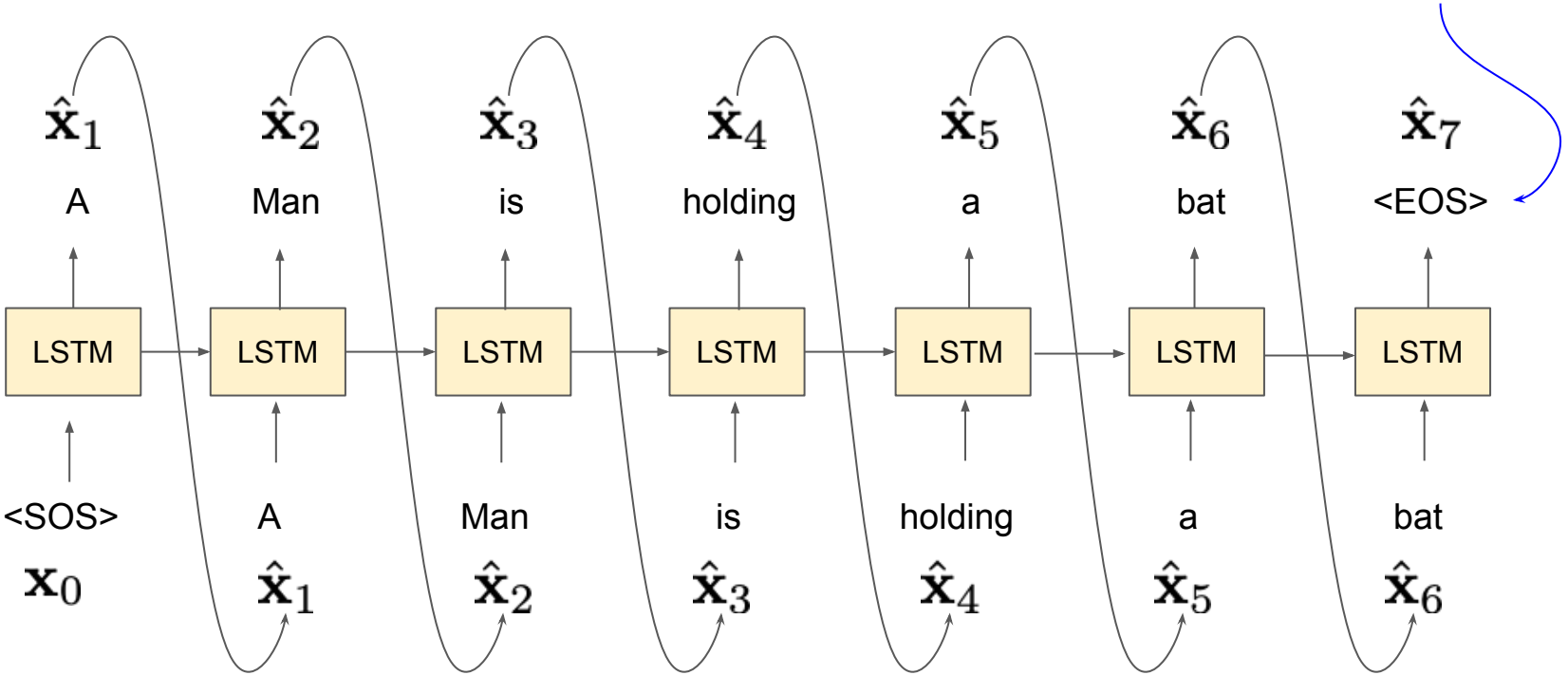


Probabilistic method

Inference time! Backprop 걱정할 필요 X

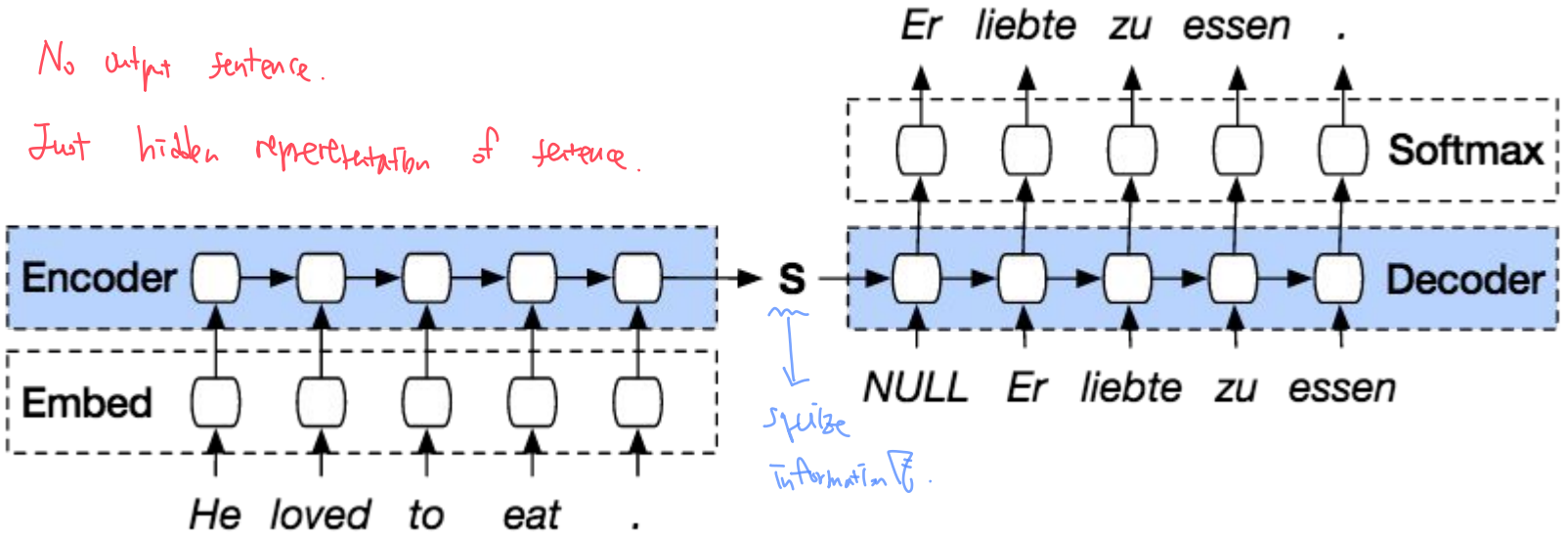
Inference: RNN-based language model

Stop sampling when it samples the end-of-sentence symbol



Machine translation

- Translate a sentence in one language to another



Summary: LSTM-based language model

- Sentence = a sequence of discrete symbols
- RNN (e.g. LSTM) for modeling a sequence of discrete symbols
  - Each word: an one-hot encoding
  - Sentence generation: prediction of the next word given the previous words
  - Training: sequential classification (classification of each word at a time)
  - Inference: sequentially predict a word and use it as an input to the next step

Today's agenda

- Language modeling using RNNs
- **Image captioning** Predicting sentence from image
  - Naive image captioning, image captioning with attention
- Visual question answering
  - Naive visual question answering, memory network

Image captioning

- Task definition: describe an image using natural language (sentence)

x: image

y: sentence

$f_\theta(y \mid g_\phi(x))$

$f_\theta(\mathbf{y} \mid f_\phi(\mathbf{x}))$

"man in black shirt is playing guitar."

1. Recognizing visual patterns in an image

2. Generate a sentence conditioned on image

CNN

RNN

Naive image captioning

$$f_\theta(\mathbf{y}|g_\phi(\mathbf{x}))$$

Pretrained CNN
using ImageNet dataset

Input image
(224x224x3)

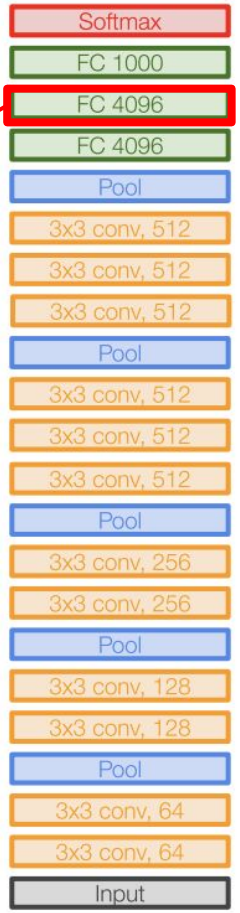
CNN

Feature vector
at fc layer
(1x1x4096)

1000 안쓰는 pre-defined class 없이고
classification (RNN output) 에 쓰일 수
없어서

RNN의 input으로 쓰기 위해 sequence로 되어 있는 것을 썼다

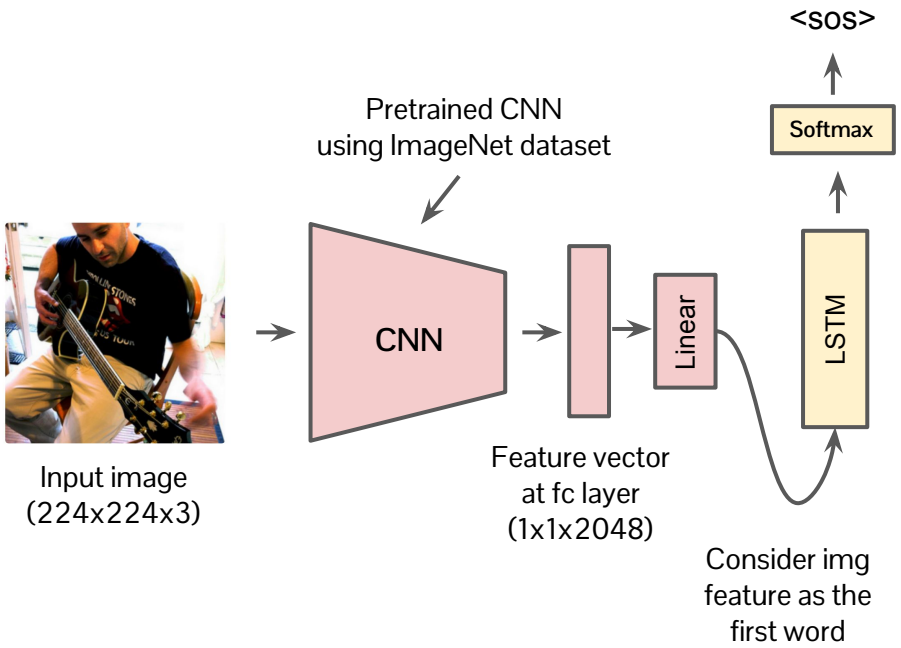
Spatial features를 고려하려면 Pooling layer를 쓸 수도.. (나중에)



VGG16

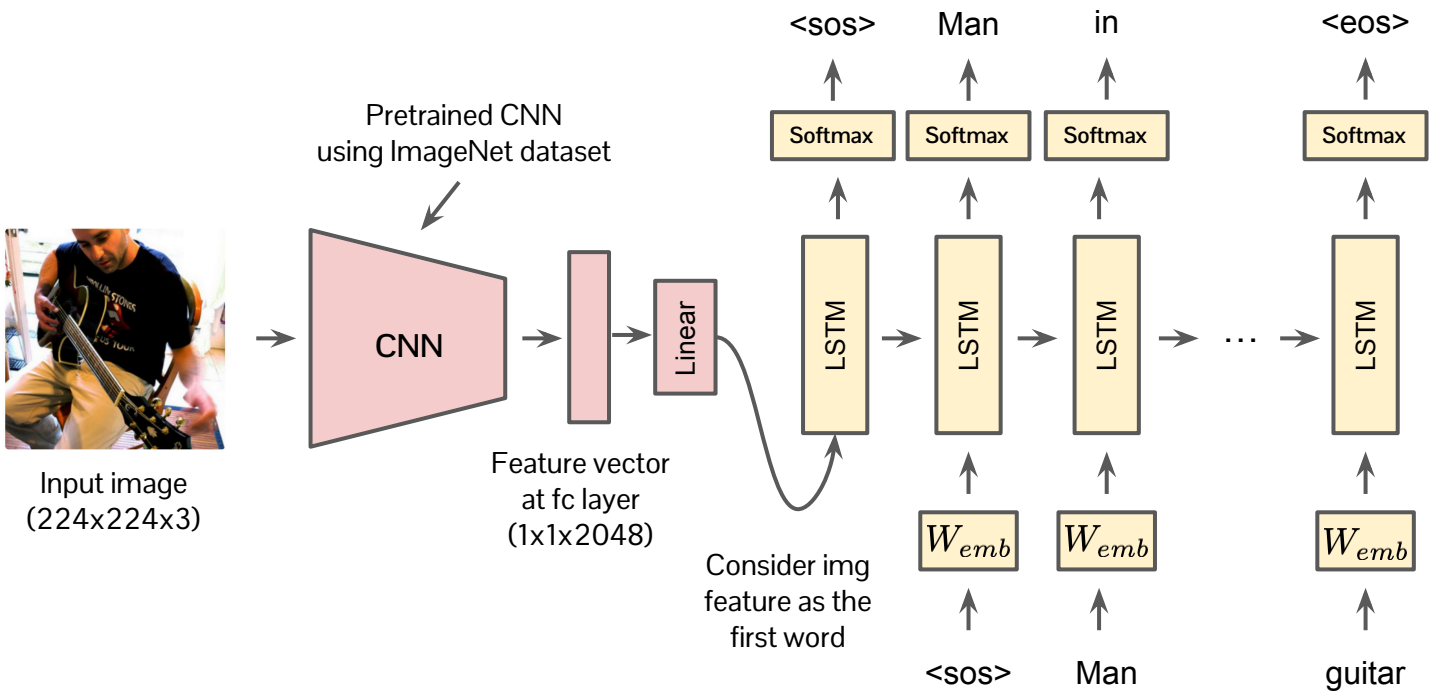
Naive image captioning

$$f_\theta(\mathbf{y}|g_\phi(\mathbf{x}))$$



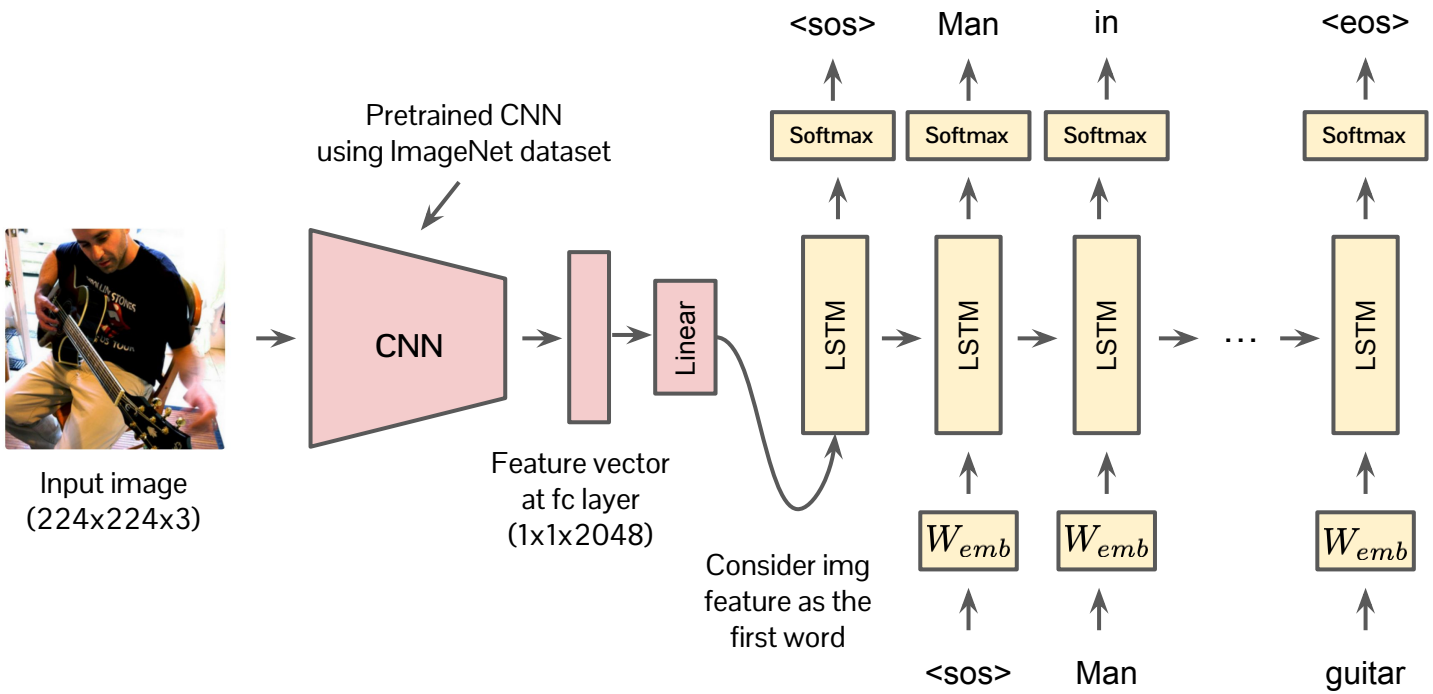
Naive image captioning

$$f_\theta(\mathbf{y}|g_\phi(\mathbf{x}))$$



Naive image captioning

$$f_\theta(\mathbf{y}|g_\phi(\mathbf{x}))$$

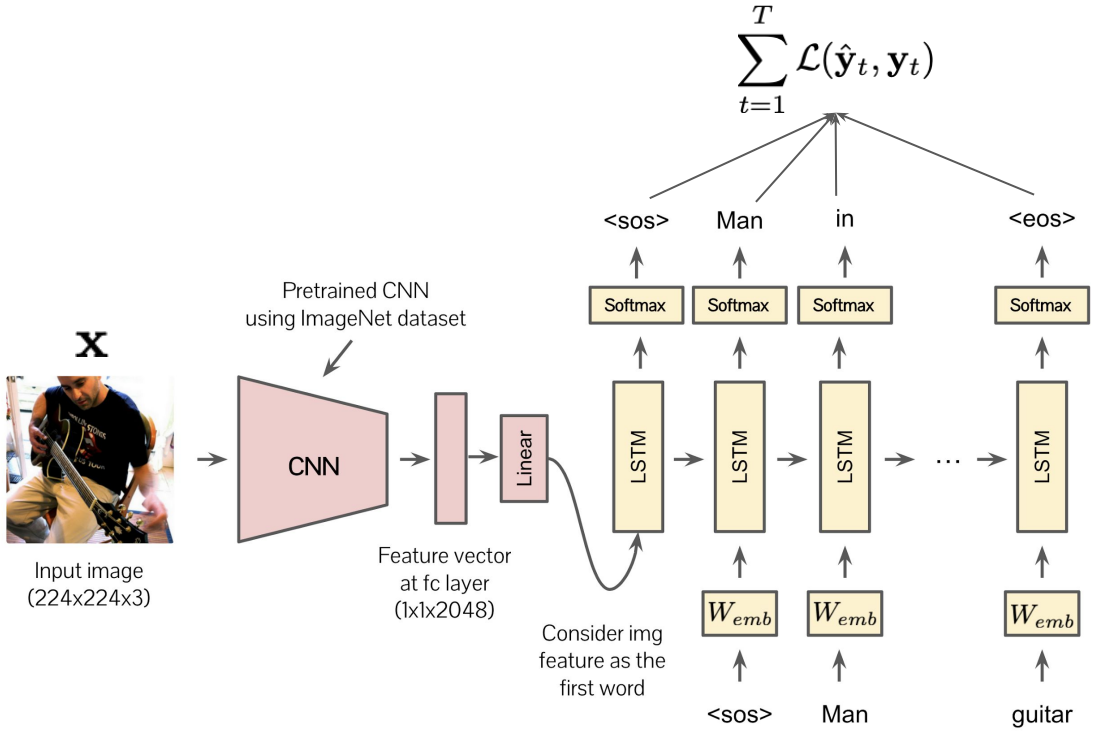


Naive image captioning: Training

Training data
(image, sentence) pairs

x	y
	A person on a beach flying a kite.
	A black and white photo of a train on a train track.
	A person skiing down a snow covered slope.
⋮	

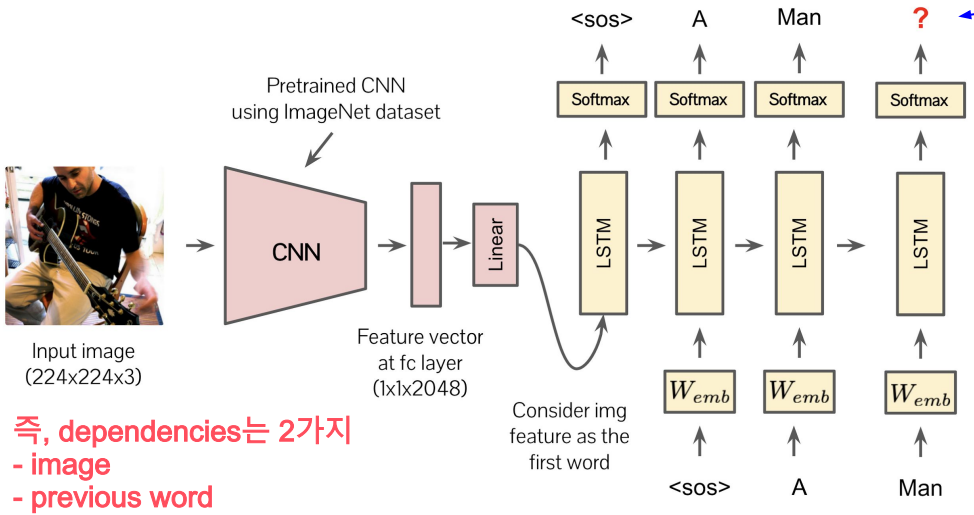
Cross-entropy loss
(same sentence generation)



Practical issue

is 넣은이유?
이전 단어 보면 문법적으로 맞아서.
play?
이전 단어 위치 보면 동사 ok. 동시에 이미지랑 관련 있다.
<eos>
역시 이전 단어와 연관이 있음

1. is
2. plays
3. <eos>



Can you guess what would be the following word?

- If RNN is strong enough, it can generate reasonable sentences **conditioned only on previous words** (not an image)
- This is especially prominent since the previous words has more direct impact on prediction of next words (shorter dependency length)
- In order to make captioning conditioned on image content, we have to **strengthen** the conditioning to image

즉, dependencies는 2가지
- image
- previous word

previous word에만 중점을 주면.. can start dirft.
즉 short term dependency에 강력하게 주는 것은 좋지 않음.
long term dependency는 backprop 쭉 해야하기 때문에 short term dependency보다 어려움..

Improving image conditioning: shortcuts

Input image
(224x224x3)

Pretrained CNN
using ImageNet dataset

CNN

Feature vector
at fc layer
(1x1x2048)

Linear

Consider img
feature as the
first word

<sos> Man in <eos>

Softmax Softmax Softmax Softmax

LSTM LSTM LSTM ... LSTM

W_{emb} W_{emb} W_{emb}

<sos> Man guitar

Add image conditioning
at every step
(short-cut connections)

Pros

Reduce the
dependency distance to
1 gradient flow directly
for ∀ time step.

Cons

Still can ignore the
image conditioning

concanate linear layer!
그런데 연결된 weights가 0이면
ignore image conditioning!

Improving image conditioning: attention

현저한, 돌출된

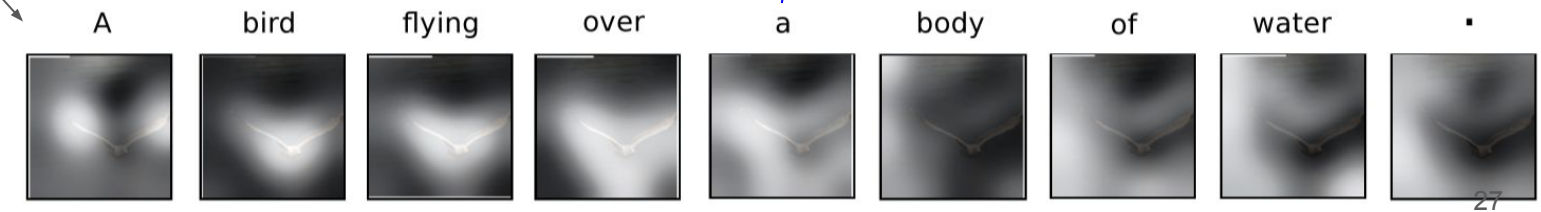
- Make the model "gaze" on salient objects for generating corresponding words

No attention (= uniform attention)

A bird flying over a body of water.

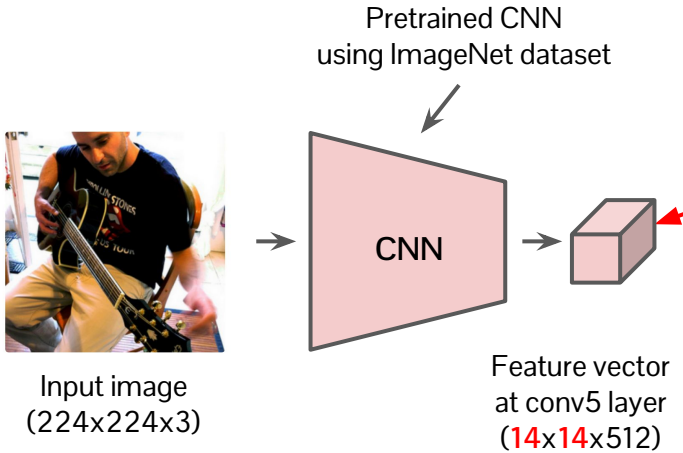
Word prediction is conditioned on parts of an image

Soft attention

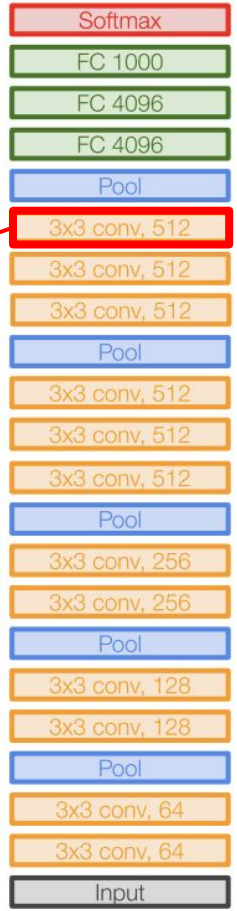


흰색이
집중하는 곳

Attention-based image captioning



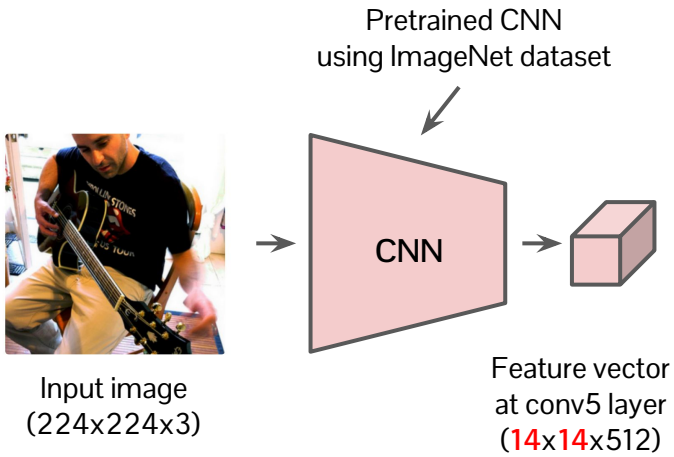
Attention 적용하려면 spatial info.필요하므로 이거 선택!



VGG16

Attention-based image captioning

직관: 사람이 모든 곳에 동시에 집중 못하듯이.. Attention is limited! 한쪽에 집중하면 다른 곳의 집중도는 떨어질 수밖에



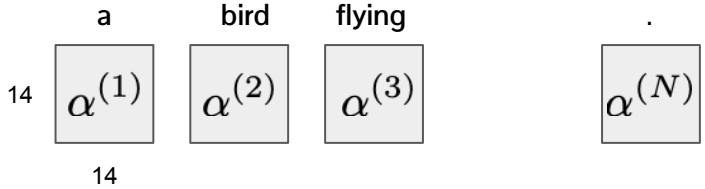
Attention:

- A positive matrix that has same spatial dimension as feature map

$$\alpha^{(t)} \in \mathbb{R}^{W \times H} \qquad \sum_{i,j} \alpha_{i,j}^{(t)} = 1, \quad \forall \alpha_{i,j}^{(t)} \geq 0$$

size of feature map

- We want to compute attention for each word



- Attention is used to abstract image feature

$$\mathbf{z}^{(t)} = \sum_{i,j} \alpha_{i,j}^{(t)} \mathbf{x}_{i,j} \in \mathbb{R}^C$$

called context feature

Attention-based image captioning

Challenges:

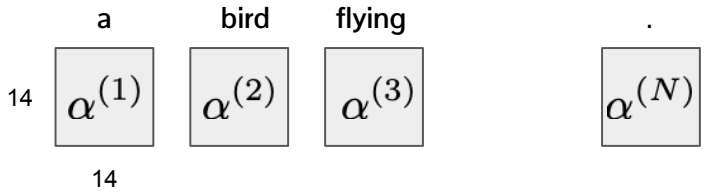
- How do we compute the attention? →
- How do we use it to predict the word? →

Attention:

- A positive matrix that has same spatial dimension as feature map

$$\alpha^{(t)} \in \mathbb{R}^{W \times H} \qquad \sum_{i,j} \alpha_{i,j}^{(t)} = 1$$

- We want to compute attention for each word



- Attention is used to abstract image feature

$$\mathbf{z}^{(t)} = \sum_{i,j} \alpha_{i,j}^{(t)} \mathbf{x}_{i,j} \in \mathbb{R}^{C}$$

Attention-based image captioning

Challenges:

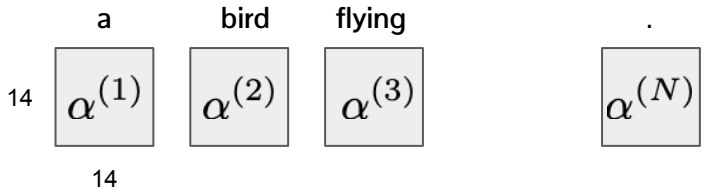
- How do we compute the attention?
- How do we use it to predict the word?

Attention:

- A positive matrix that has same spatial dimension as feature map

$$\alpha^{(t)} \in \mathbb{R}^{W \times H} \qquad \sum_{i,j} \alpha_{i,j}^{(t)} = 1$$

- We want to compute attention for each word

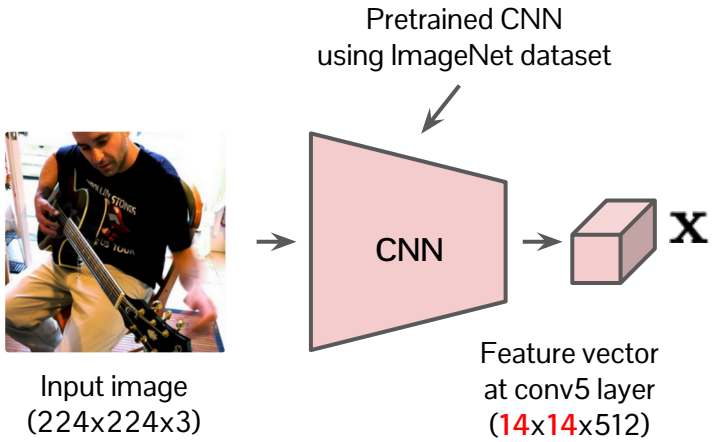


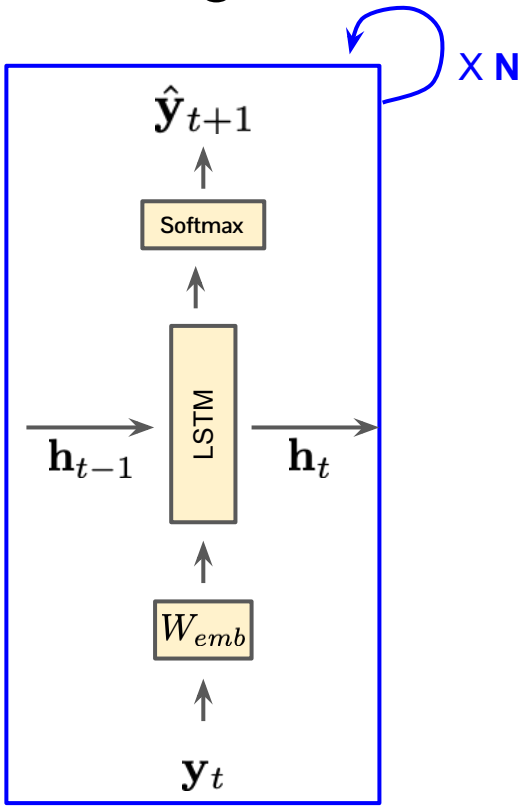
- Attention is used to abstract image feature

$$\mathbf{z}^{(t)} = \sum_{i,j} \alpha_{i,j}^{(t)} \mathbf{x}_{i,j} \in \mathbb{R}^{C}$$

Attention-based image captioning

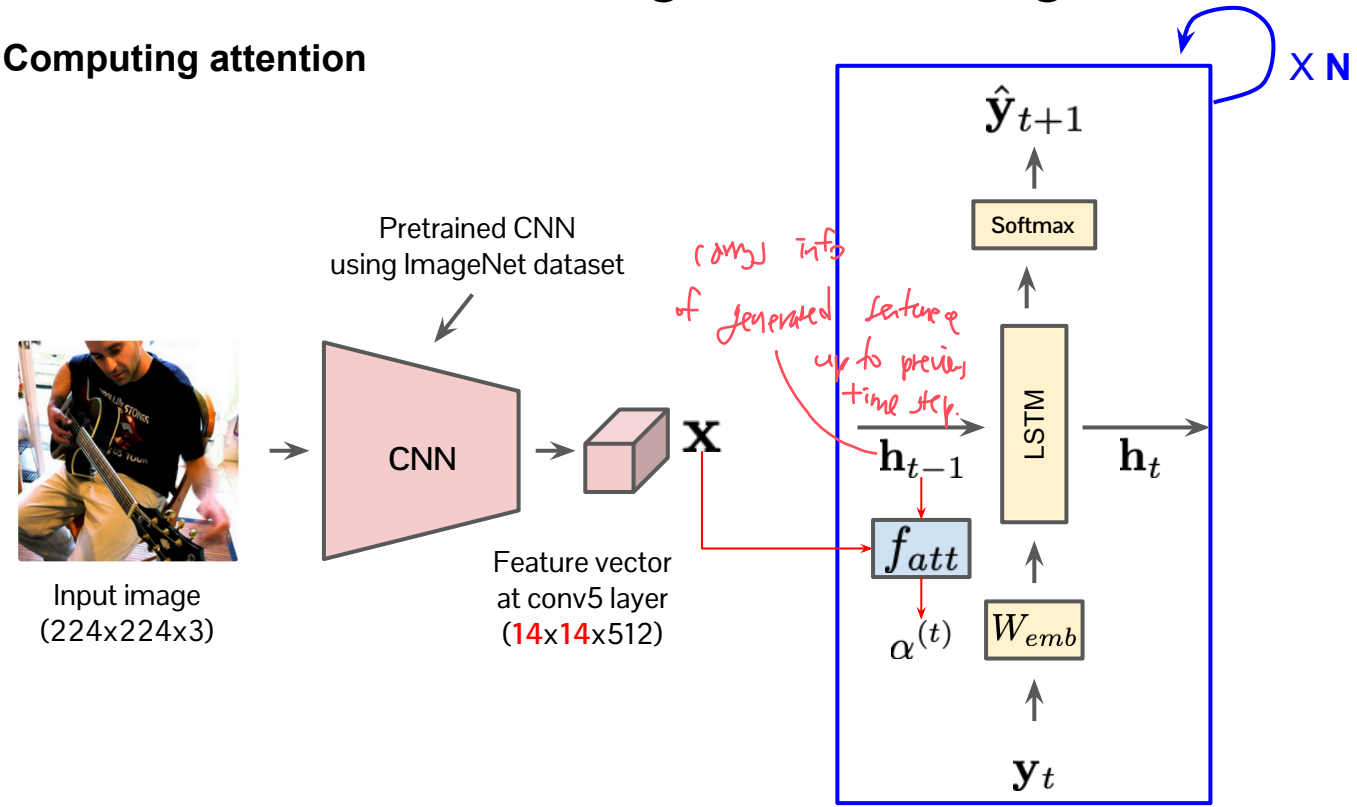
Computing attention





Attention-based image captioning

Computing attention



Attention module

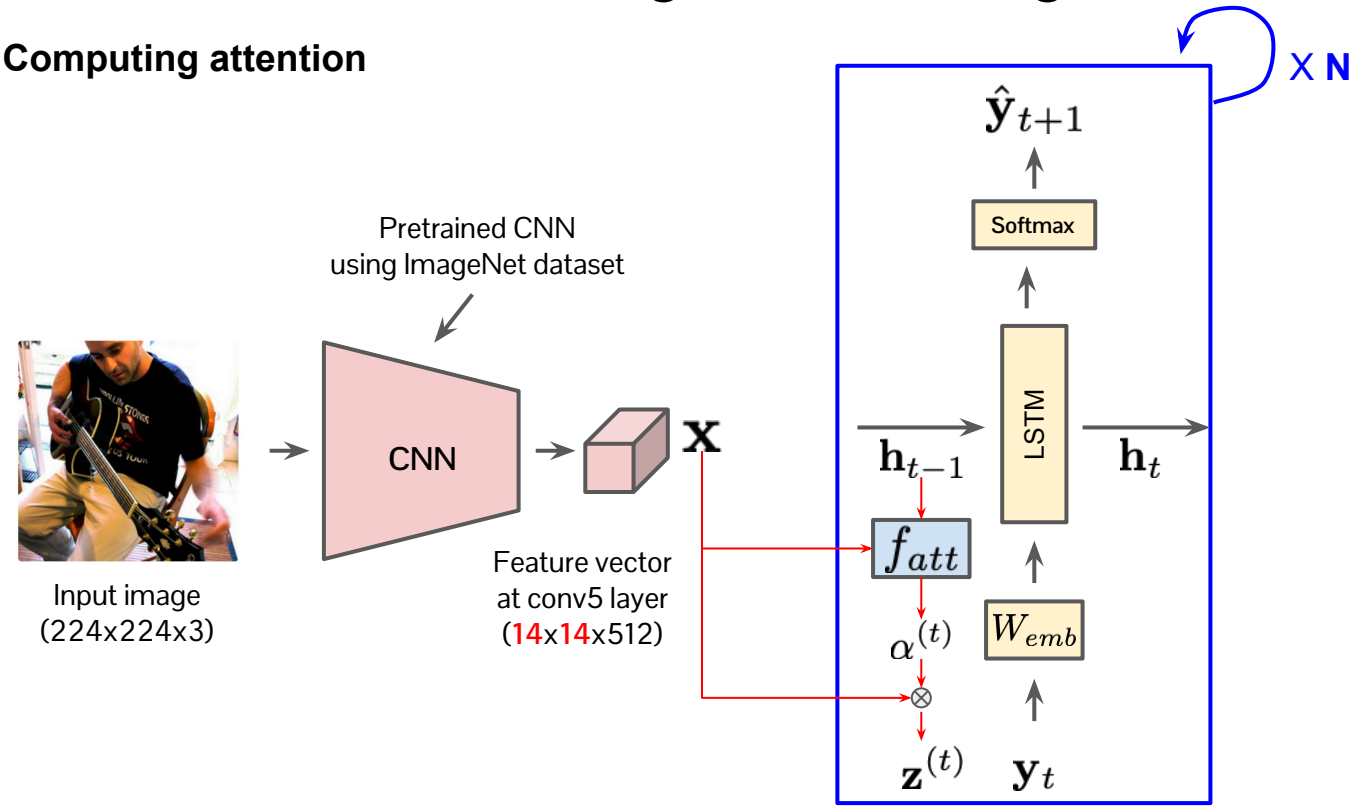
$$e^{(t)} = f_{att}(\mathbf{x}, \mathbf{h}_{t-1})$$

$$\alpha^{(t)} = \frac{\exp(e^{(t)})}{\sum_{i,j} \exp(e^{(t)}_{i,j})}$$

↳ to make sum 1.

Attention-based image captioning

Computing attention



Attention module

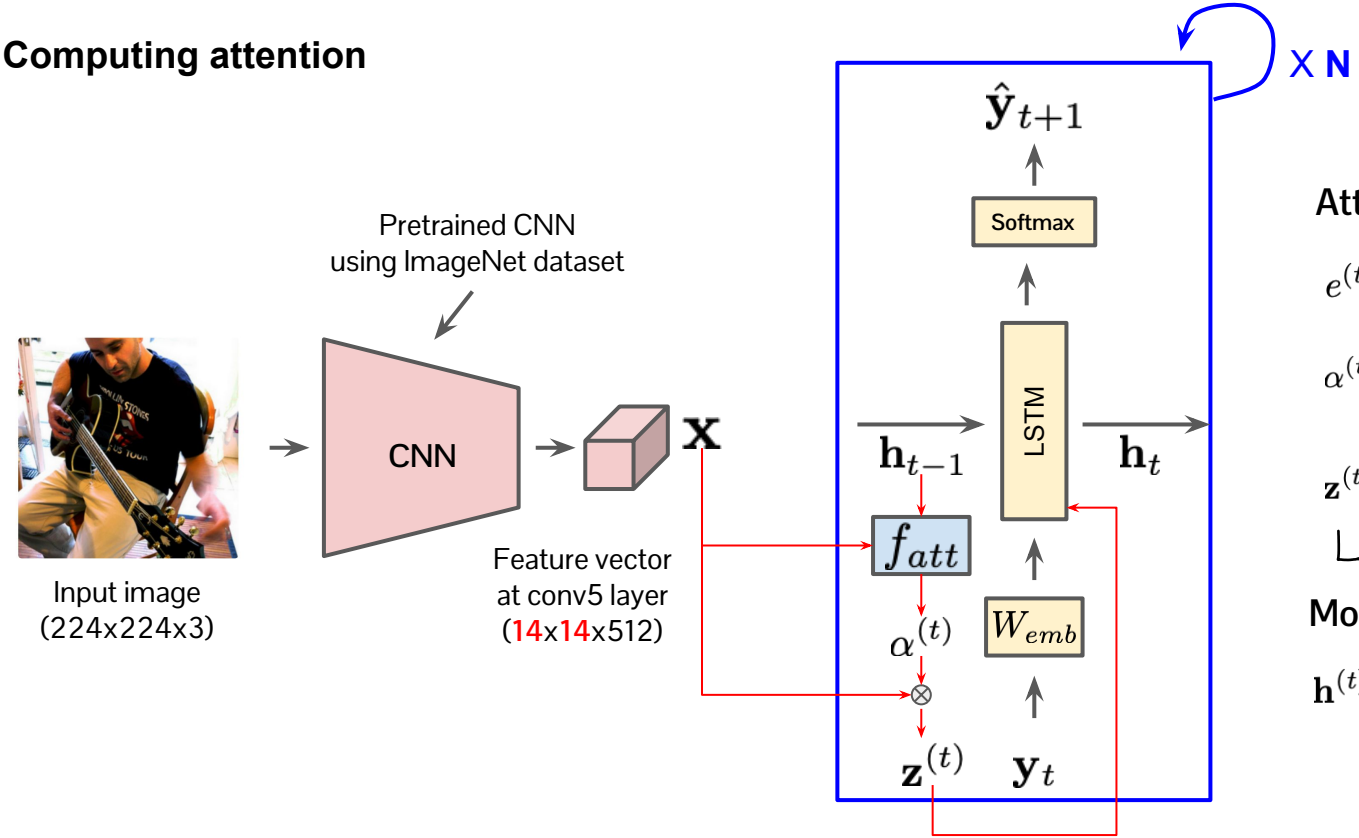
$$e^{(t)} = f_{att}(\mathbf{x}, \mathbf{h}_{t-1})$$

$$\alpha^{(t)} = \frac{\exp(e^{(t)})}{\sum_{i,j} \exp(e^{(t)}_{i,j})}$$

$$\mathbf{z}^{(t)} = \sum_{i,j} \alpha^{(t)}_{i,j} \mathbf{x}_{i,j}$$

Attention-based image captioning

Computing attention



Attention module

$$e^{(t)} = f_{att}(\mathbf{x}, \mathbf{h}_{t-1})$$

$$\alpha^{(t)} = \frac{\exp(e^{(t)})}{\sum_{i,j} \exp(e^{(t)}_{i,j})}$$

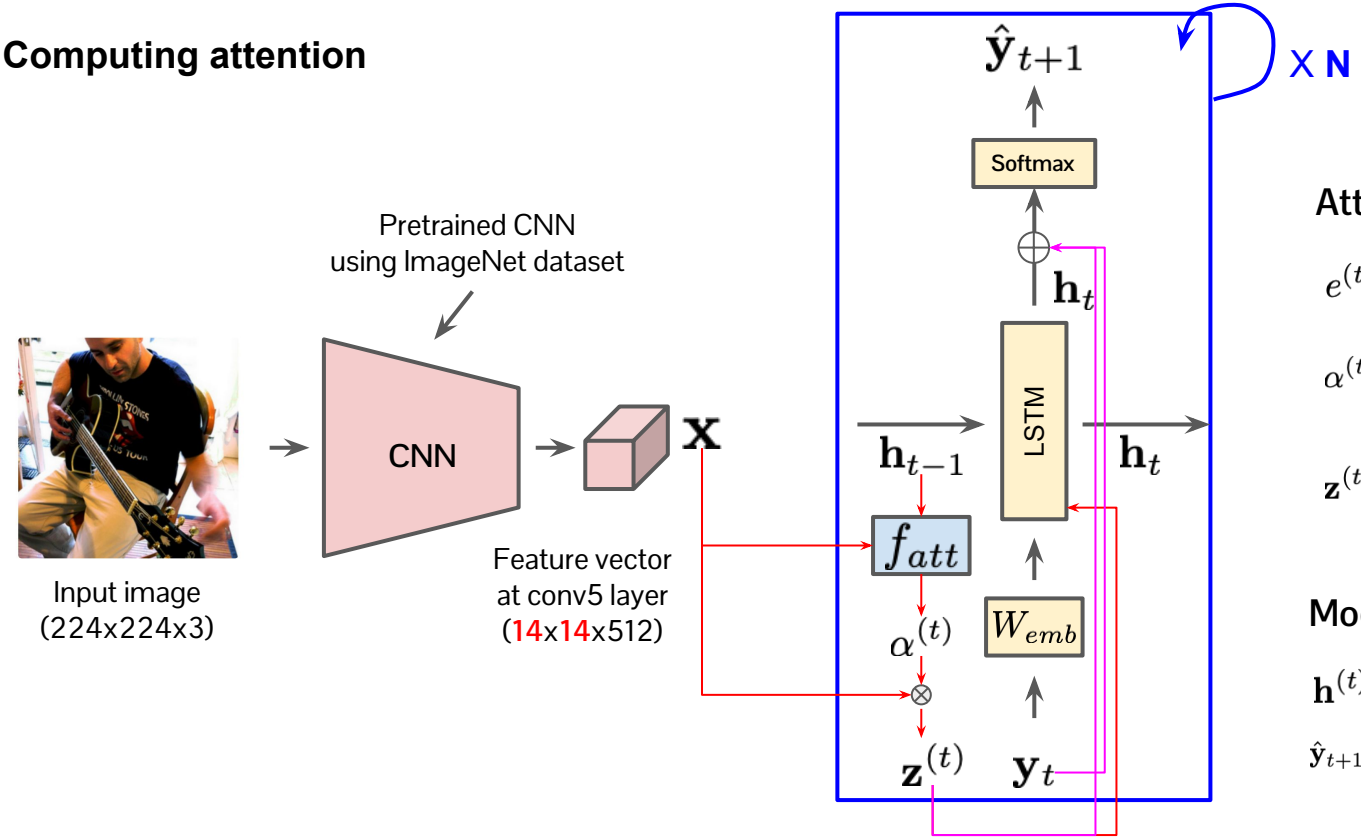
$$\mathbf{z}^{(t)} = \sum_{i,j} \alpha^{(t)}_{i,j} \mathbf{x}_{i,j}$$

Modified LSTM

$$\mathbf{h}^{(t)} = LSTM(\mathbf{h}_{t-1}, W_{emb}\mathbf{y}_t, \mathbf{z}^{(t)})$$

Attention-based image captioning

Computing attention



Attention module

$$e^{(t)} = f_{att}(\mathbf{x}, \mathbf{h}_{t-1})$$

$$\alpha^{(t)} = \frac{\exp(e^{(t)})}{\sum_{i,j} \exp(e^{(t)}_{i,j})}$$

$$\mathbf{z}^{(t)} = \sum_{i,j} \alpha^{(t)}_{i,j} \mathbf{x}_{i,j}$$

Modified LSTM

$$\mathbf{h}^{(t)} = LSTM(\mathbf{h}_{t-1}, W_{emb}\mathbf{y}_t, \mathbf{z}^{(t)})$$

$$\hat{\mathbf{y}}_{t+1} = \exp(W^o(W_{emb}\mathbf{y}_t + W^h\mathbf{h}_t + W^z\mathbf{z}_t))$$

이 예제를 보면 description은 틀렸지만 attention을 봄으로써 word의 image에의 relevancy를 확인할 수 있다.


A(0.96)


group(0.27)


of(0.27)


people(0.21)


sitting(0.28)


on(0.22)


a(0.21)


boat(0.19)


in(0.13)

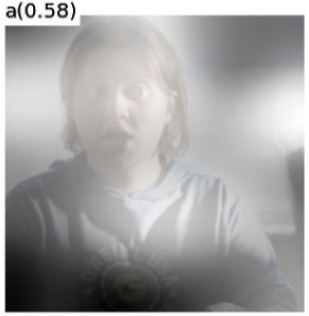

the(0.10)


water(0.30)

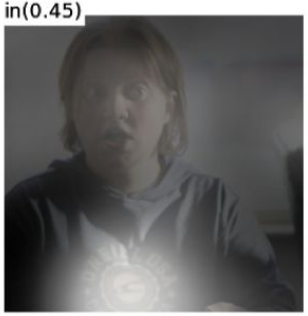

A(1.00)

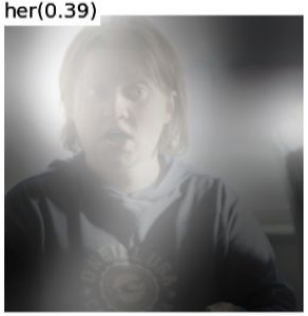

woman(0.80)

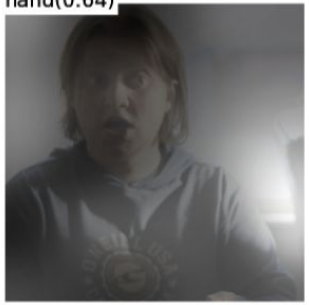

holding(0.68)

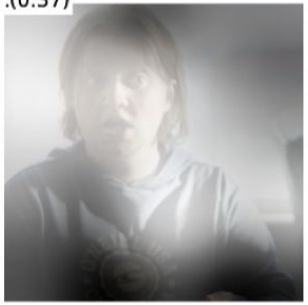

a(0.58)


clock(0.62)


in(0.45)


her(0.39)


hand(0.64)

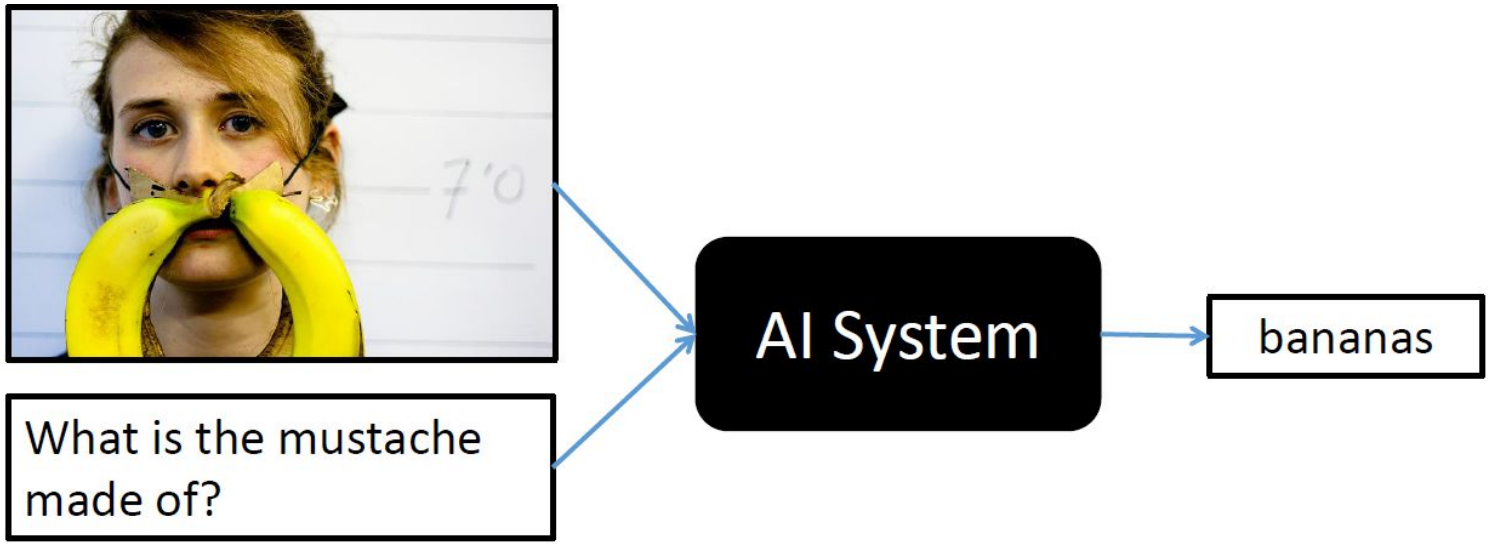

.(0.37)

Today’s agenda

- Language modeling using RNNs
- Image captioning
  - Naive image captioning, image captioning with attention
- **Visual question answering**
  - Naive visual question answering, memory network

Visual question answering

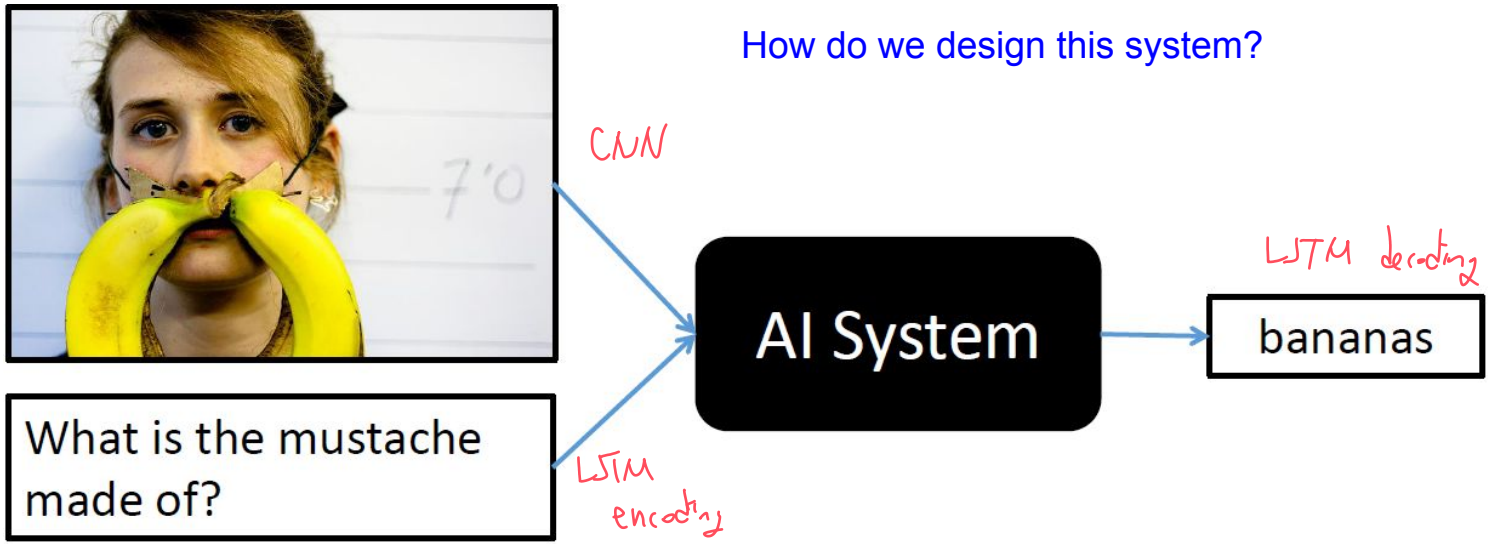
- Objective: given an image and a question about an image, predict an answer.



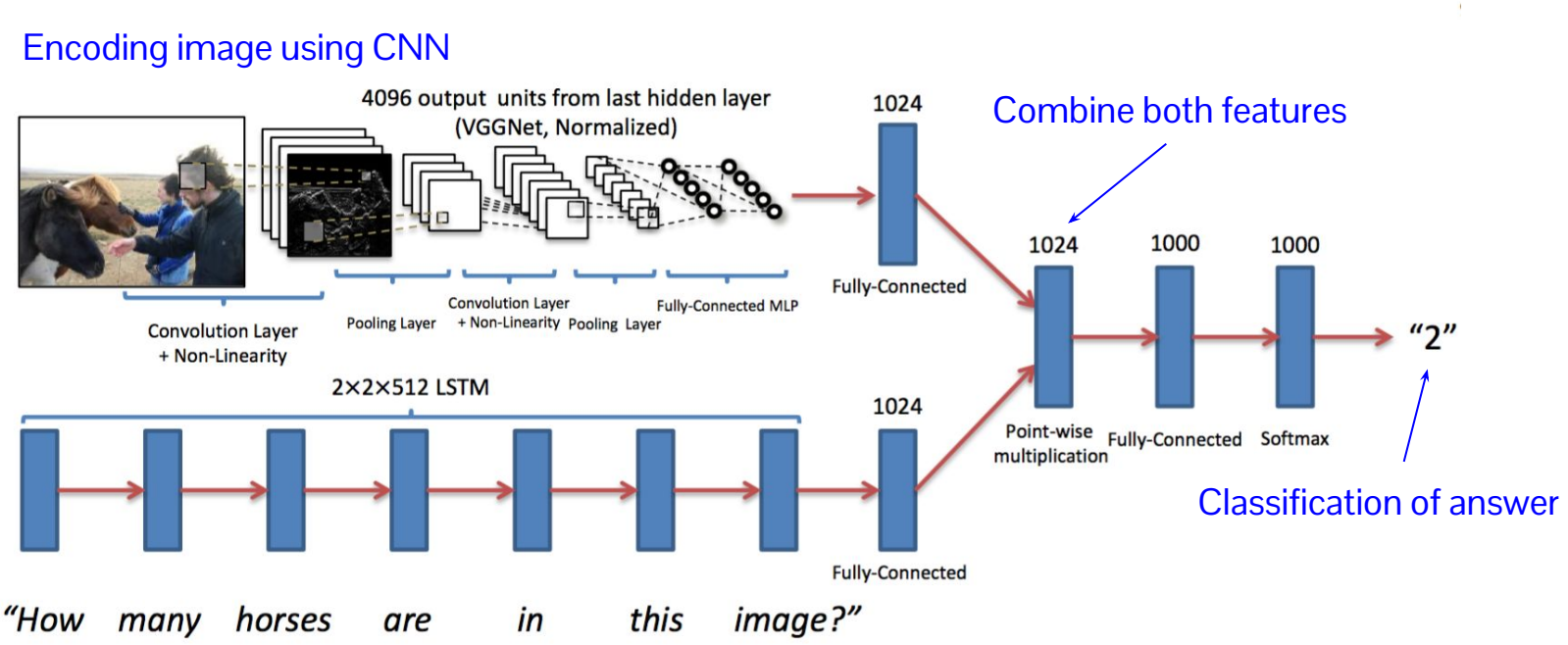
Visual question answering

Multi-step reasoning이 필요한 경우도 있다.
e.g. 자전거 핸들의 바구니에 사과를 옮기는 경우..
Q. 자전거는 몇 개의 바구니를 옮기고 있나?
-> Attention 이용하여 해결!!

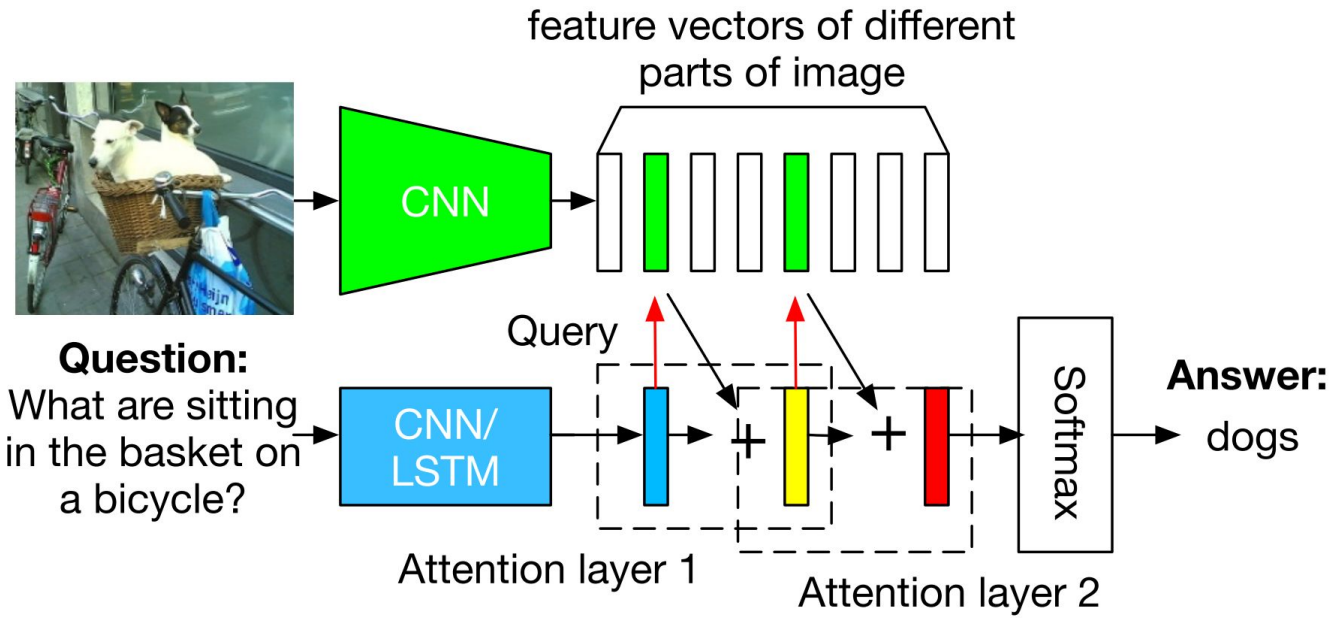
- Objective: given an image and a question about an image, predict an answer.



Visual question answering



VQA with attention



Yang et al., Stacked Attention Networks for Image Question Answering, In CVPR, 2016

VQA with attention

Question: What are sitting in the basket on a bicycle?

Original Image First Attention Layer Second Attention Layer