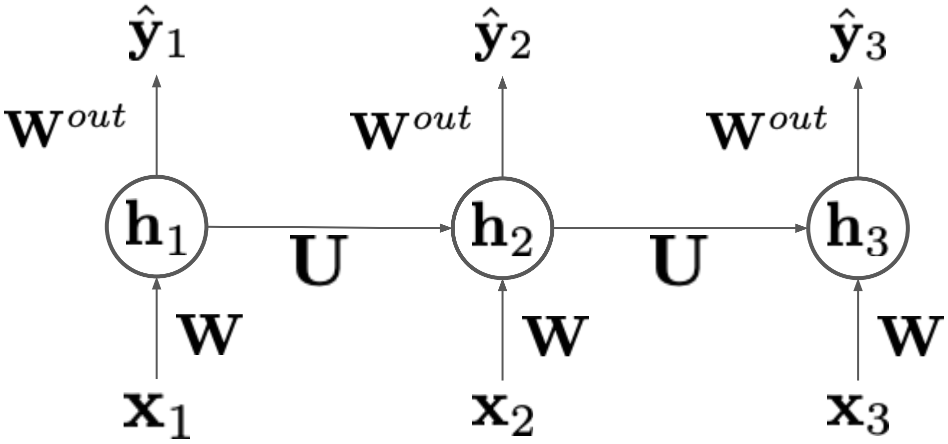
Language and Vision

Instructor: Seunghoon Hong

Course logistics

- No lecture on the next week

Recap: (Vanilla) Recurrent Neural Network



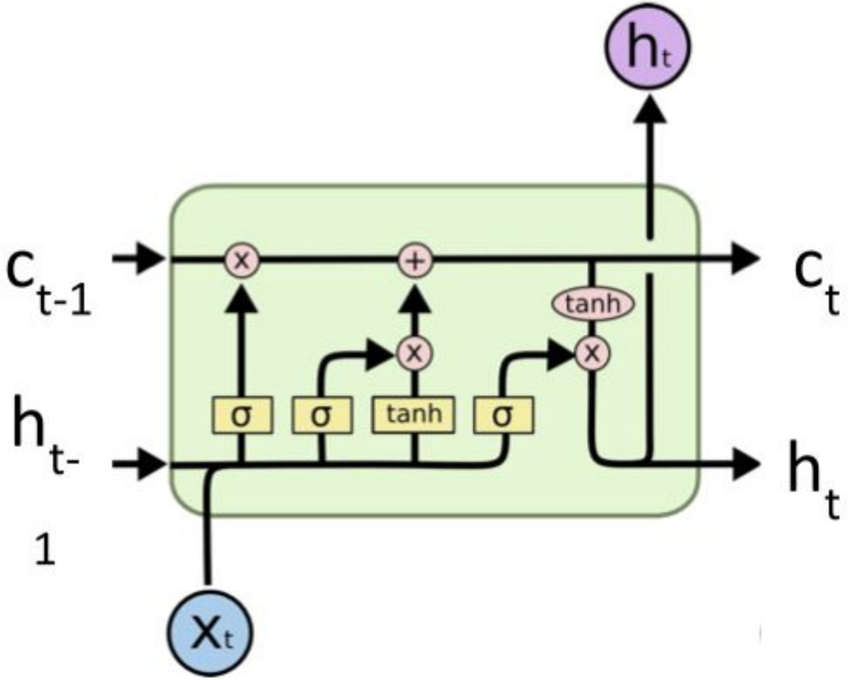
In general, for any $t \geq 1$,

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

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

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

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

Sentence generation

$f_\theta \rightarrow$ I am very hungry at the end of every class

Machine translation

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

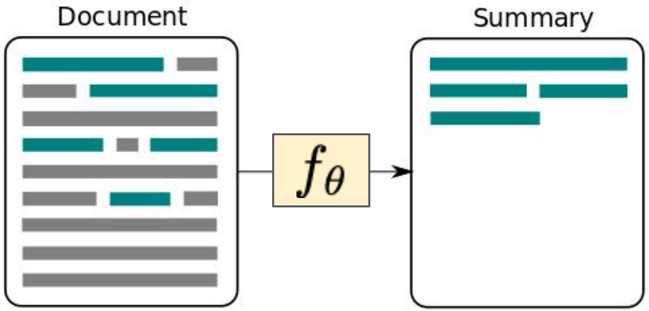
$\downarrow$

f_θ

$\downarrow$

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

Text summarization



And so many more...

Modeling language

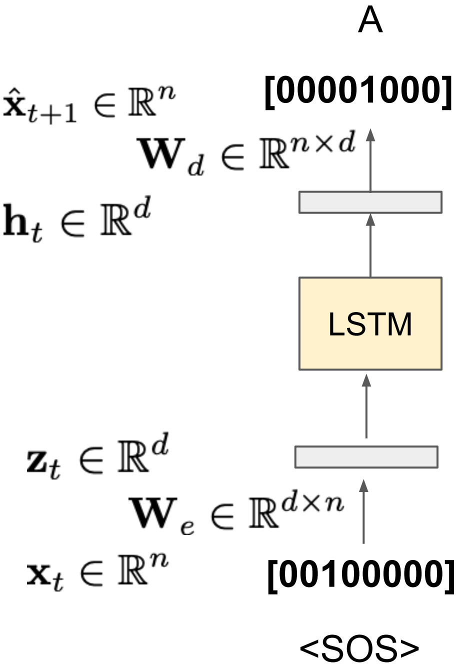
- Sentence = a sequence of discrete symbols

	A	Man	is	holding	a	bat
$\mathbf{x}_t \in \mathbb{R}^n$	**[00001000]**	**[01000000]**	**[00000010]**	**[00000100]**	**[00001000]**	**[10000000]**
	$\mathbf{x}_0$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$

One-hot encoding of discrete symbols
(tokenization)

RNN as a language model

- Sentence generation = predicting a next token



Classification layer $\hat{\mathbf{x}}_{t+1} = \mathbf{W}_d \mathbf{h}_t + \mathbf{b}_d$

Continuous output embedding $\mathbf{h}_t = \mathrm{LSTM}(\mathbf{z}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}; \theta)$

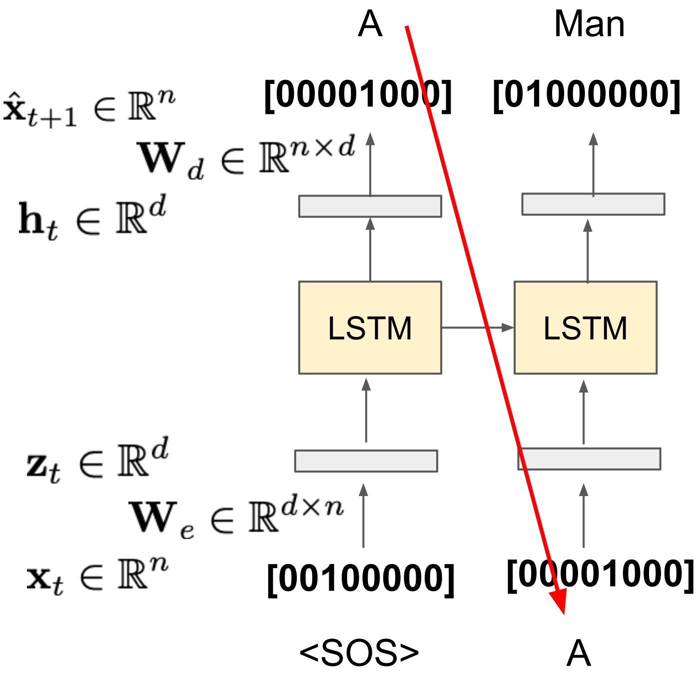
Recurrent neural network

Continuous word embedding $\mathbf{z}_t = \mathbf{W}_e \mathbf{x}_t + \mathbf{b}_e$

One-hot encoding

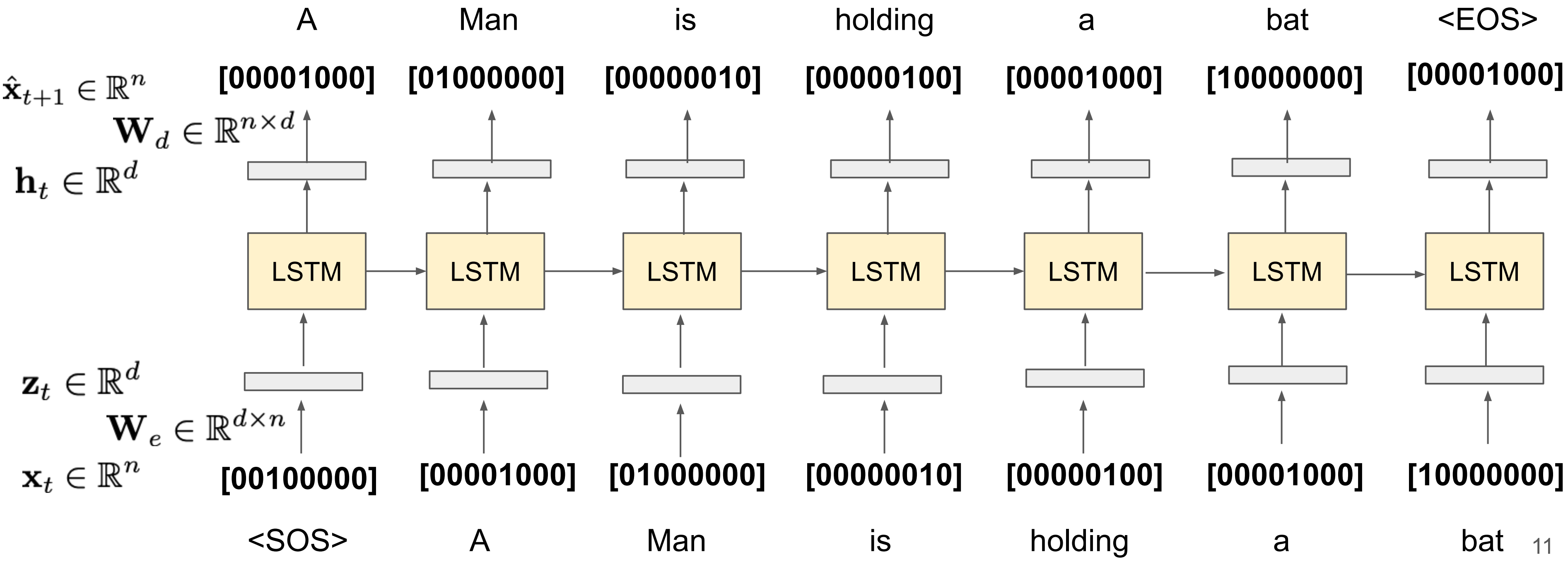
RNN as a language model

- Sentence generation = predicting a next token



RNN as a language model

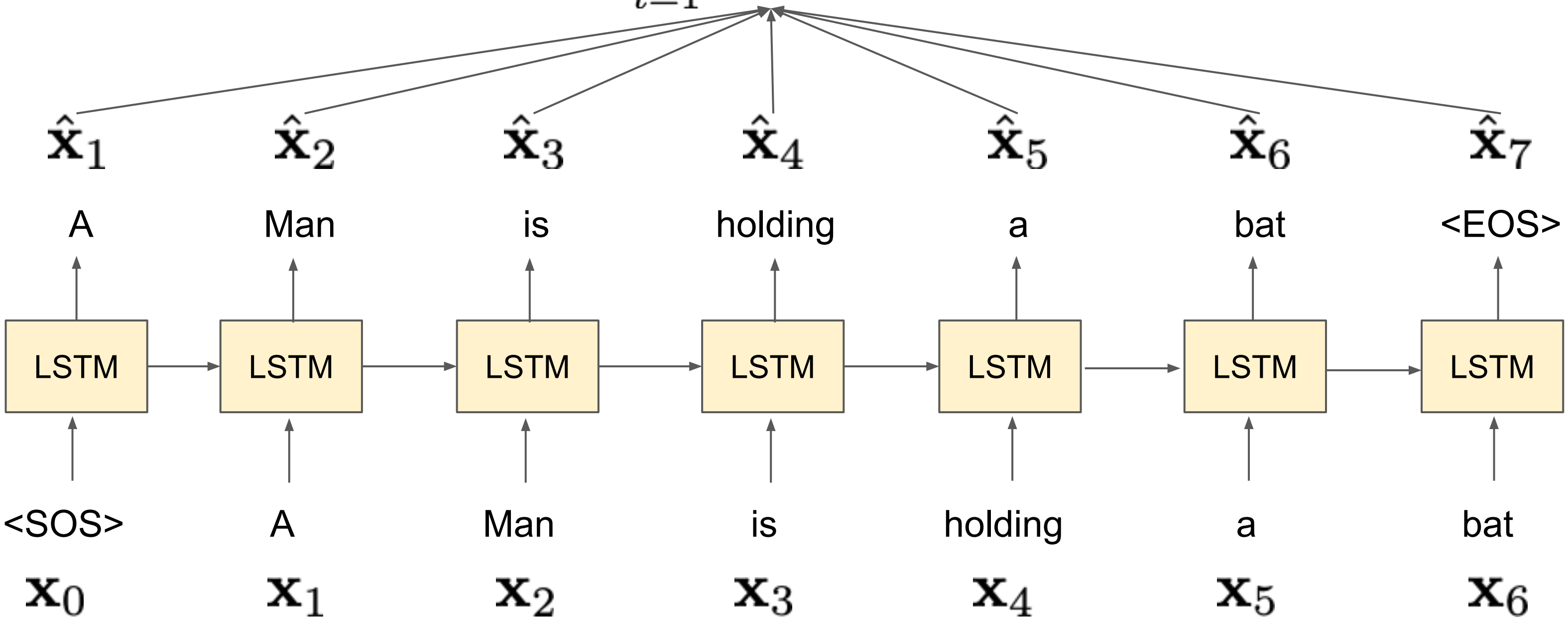
- Sentence generation = predicting a next token



Training: RNN-based language model

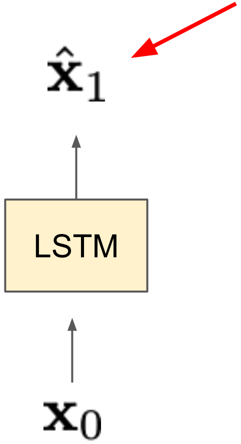
$$\sum_{t=1}^{T} \mathcal{L}(\hat{\mathbf{x}}_t, \mathbf{x}_t)$$

Cross-entropy loss (word-level classification)

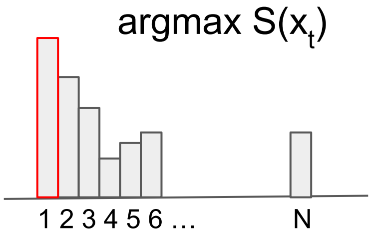


We feed ground-truth words as inputs (also known as **teacher forcing**)

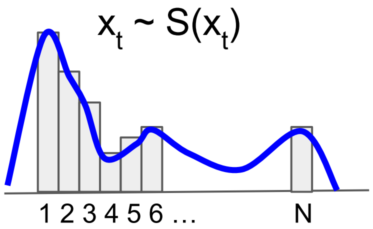
Inference: RNN-based language model



- **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)



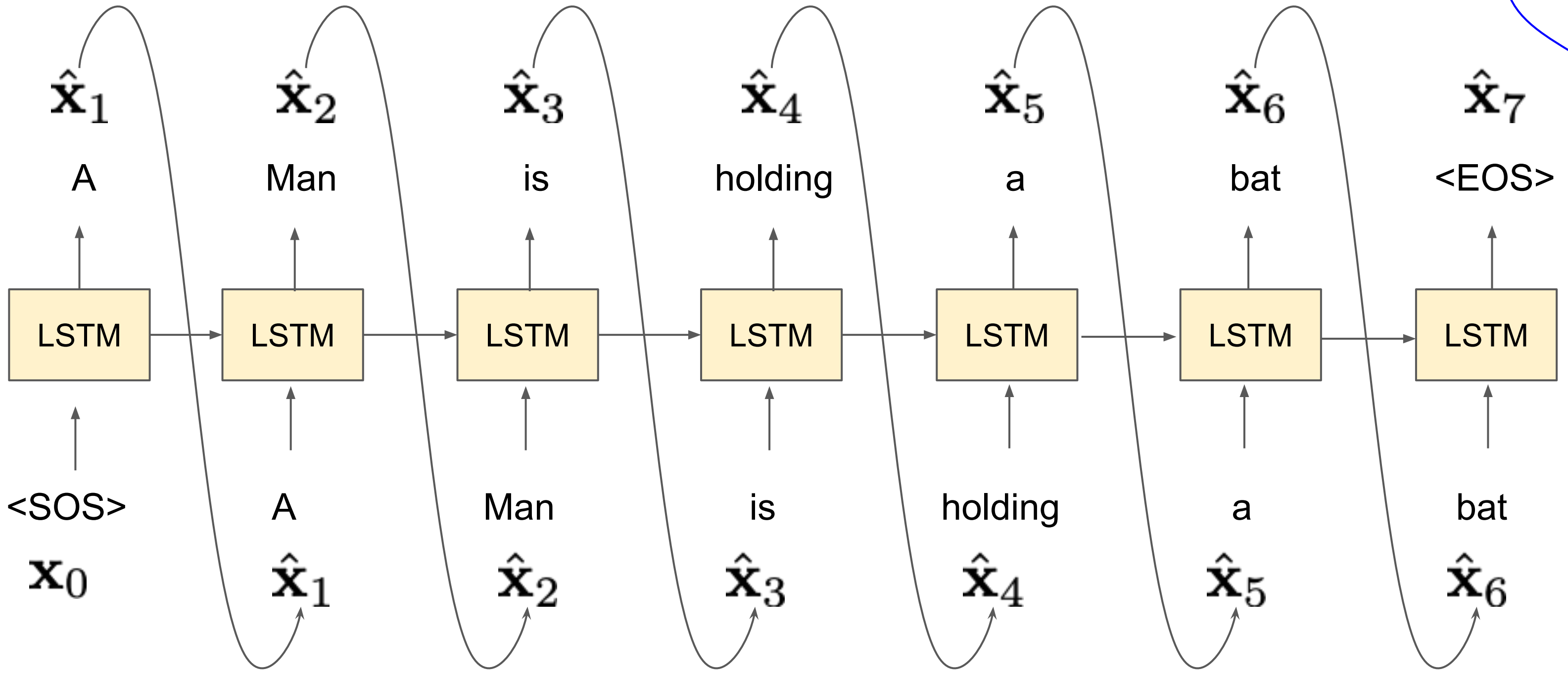
Greedy method



Probabilistic method

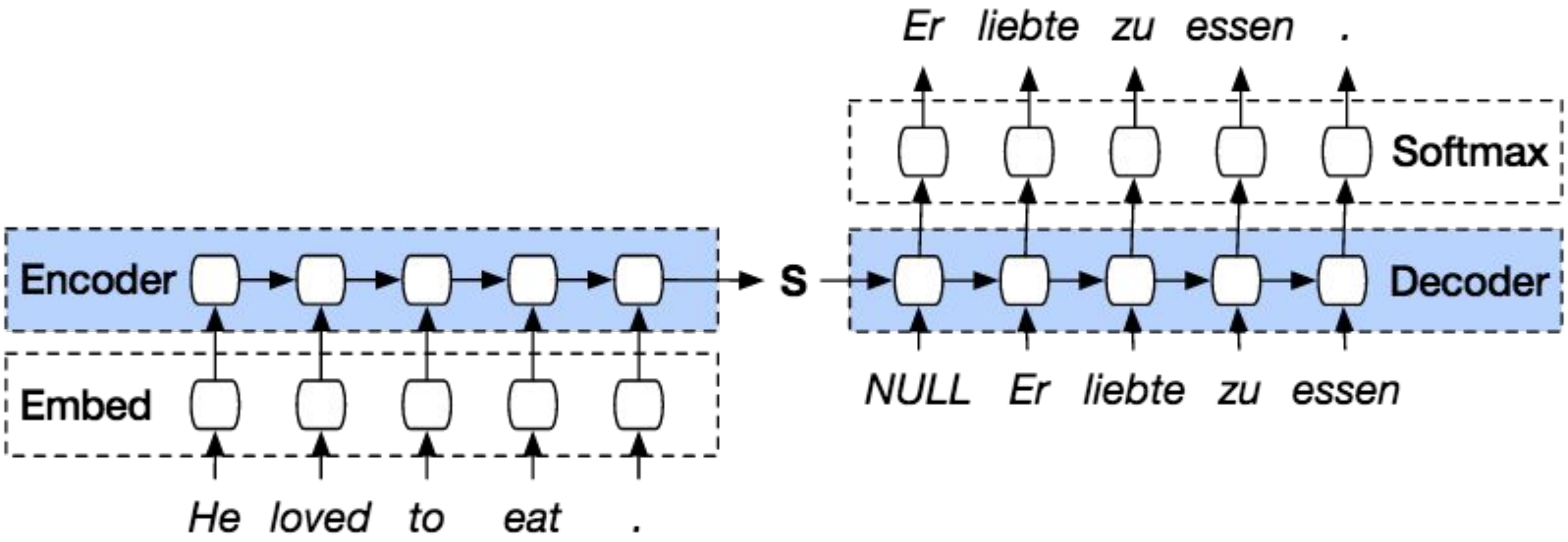
Inference: RNN-based language model

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**
    - 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

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

y: sentence

"man in black shirt is playing guitar."

1. Recognizing visual patterns in an image

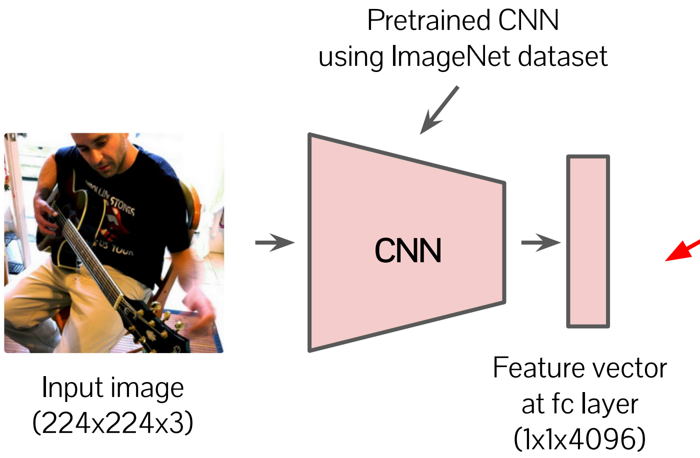
2. Generate a sentence conditioned on image

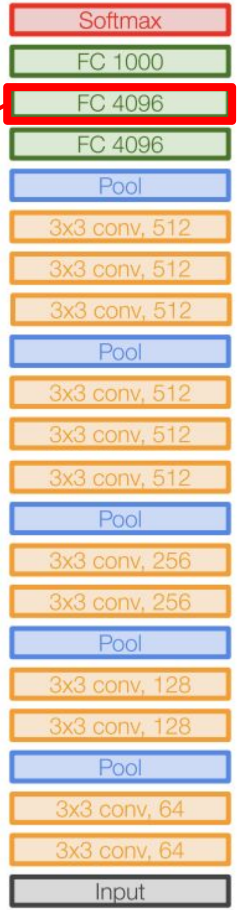
CNN

RNN

Naive image captioning

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

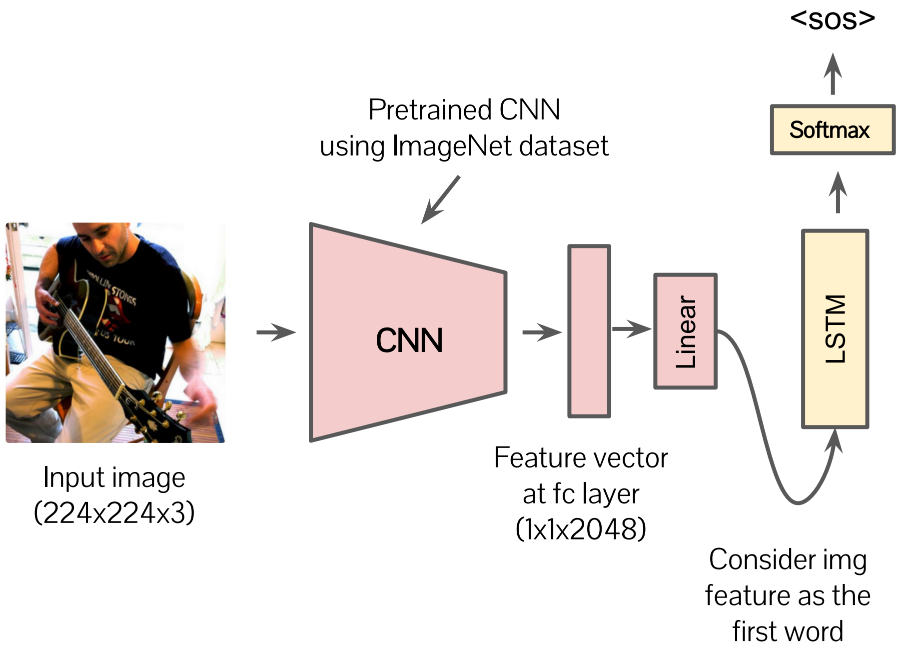




VGG16

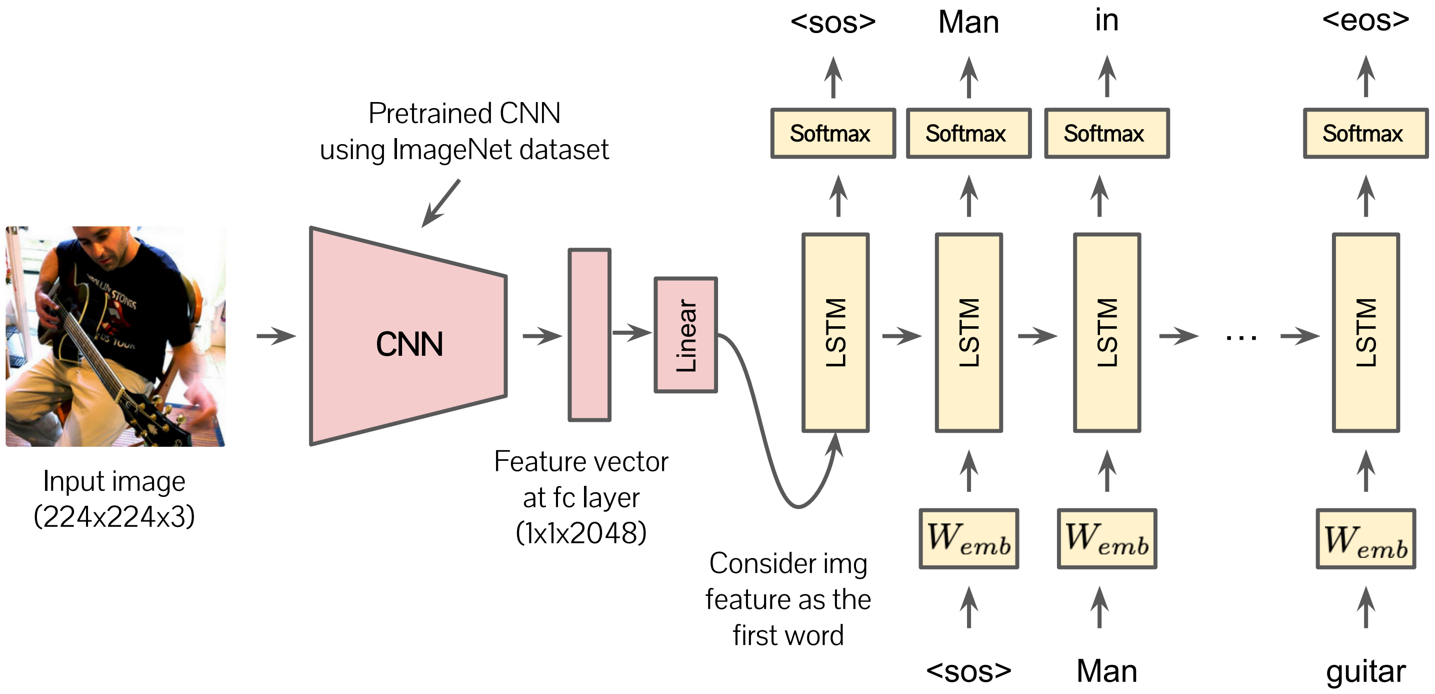
Naive image captioning

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



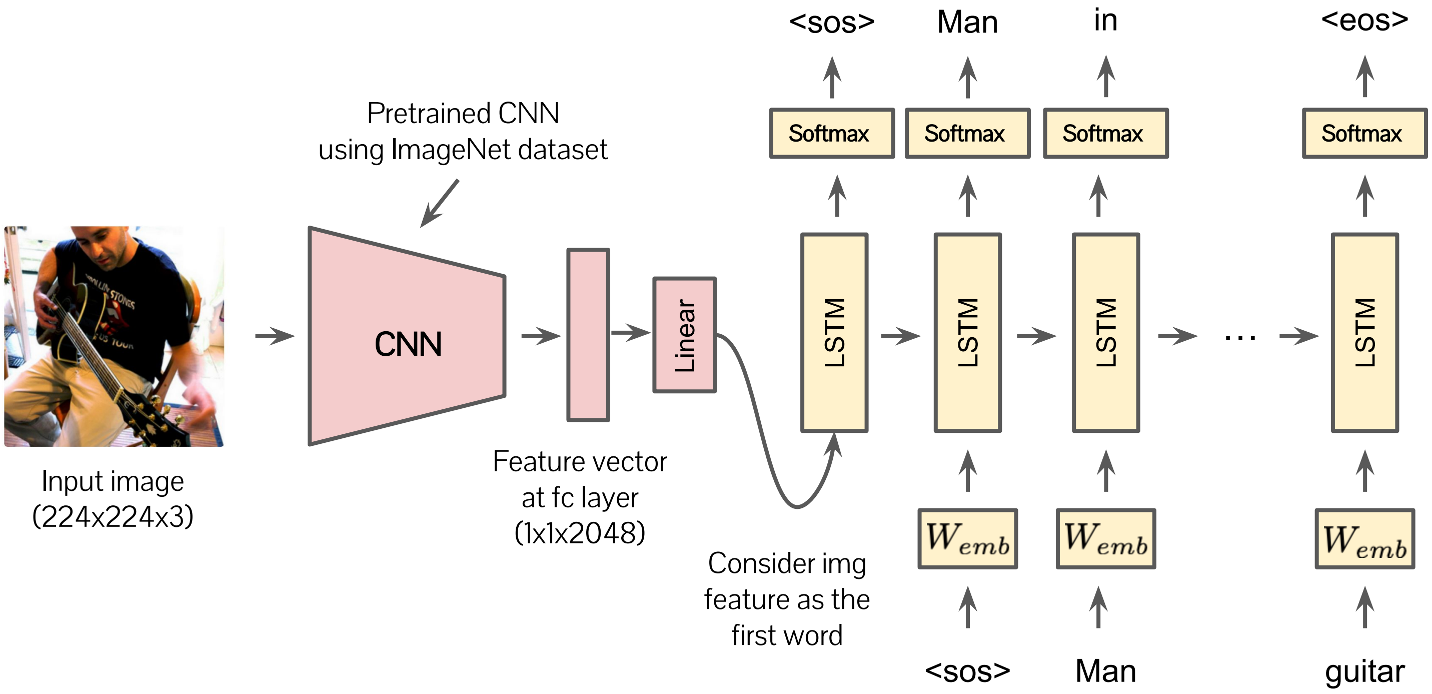
Naive image captioning

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



Naive image captioning

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

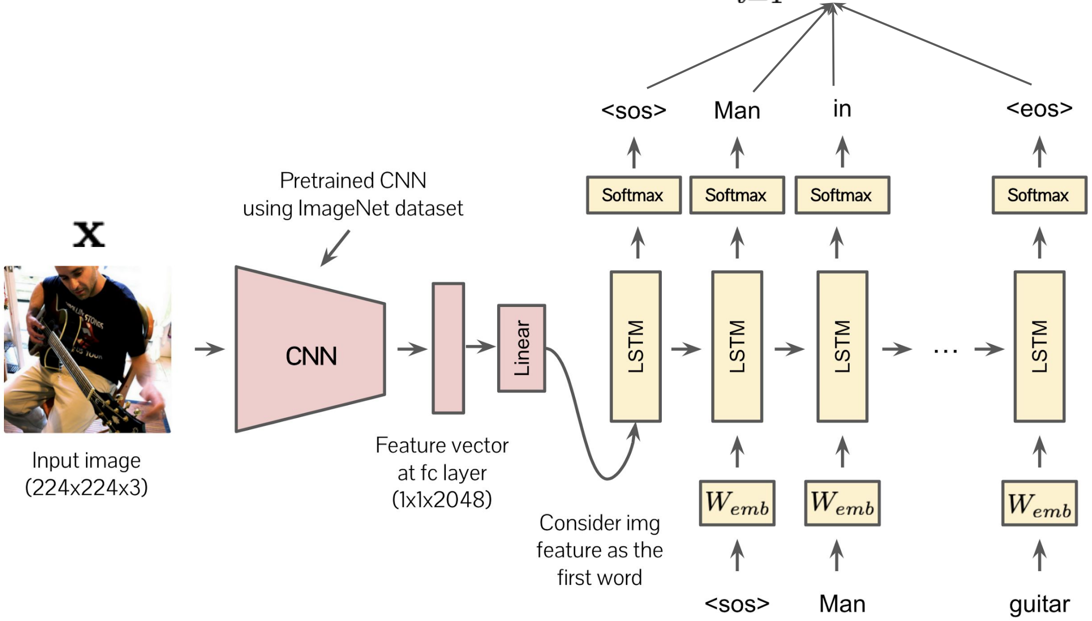


Naive image captioning: Training

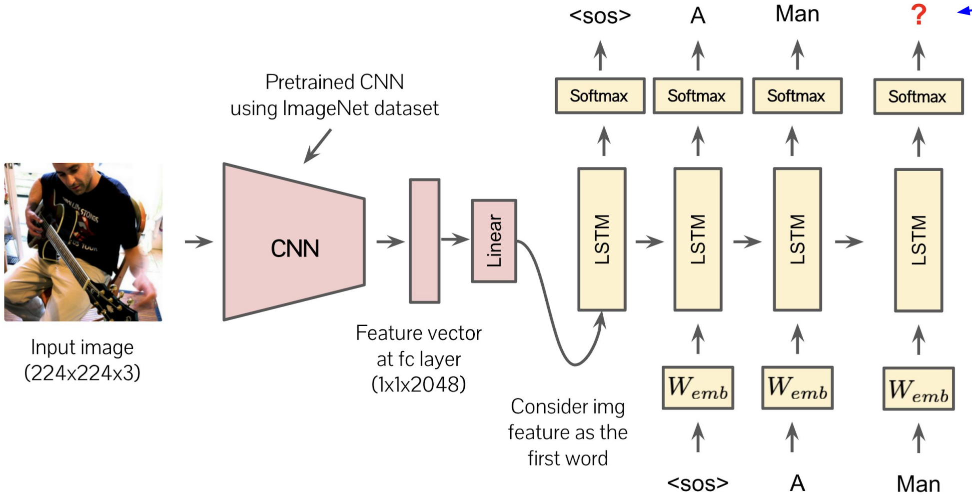
Training data
(image, sentence) pairs

Cross-entropy loss
(same as sentence generation)

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



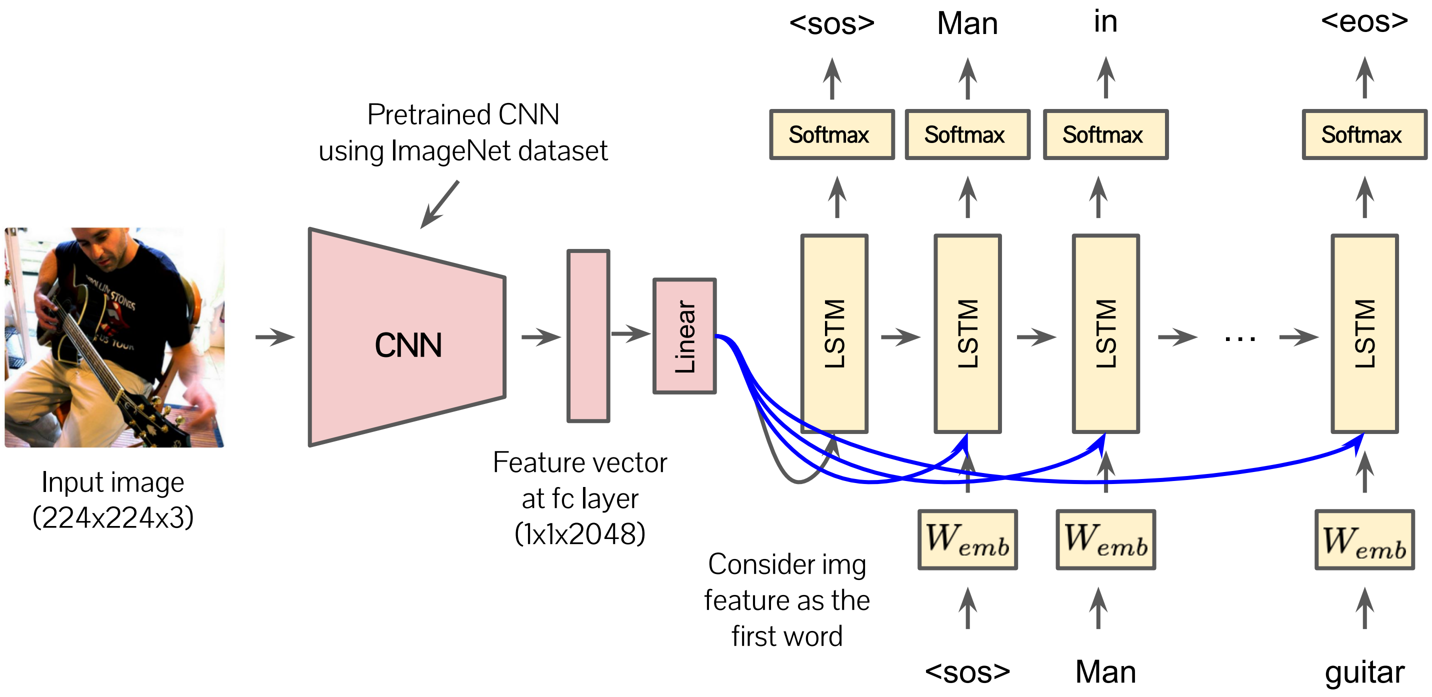
Practical issue



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

Improving image conditioning: shortcuts



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

Pros

Reduce the dependency distance to 1

Cons

Still can ignore the image conditioning

Improving image conditioning: attention

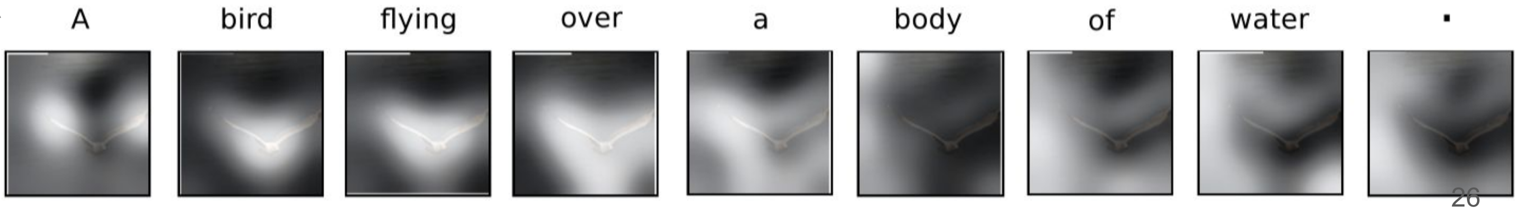
- Make the model “gaze” on salient objects for generating corresponding words

No attention (= uniform attention)

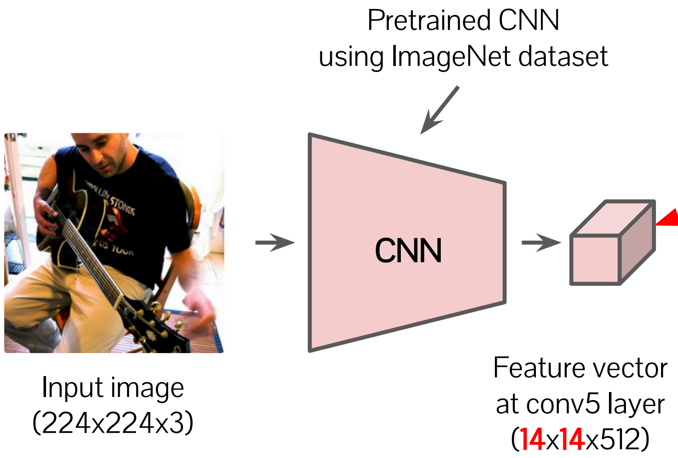
A bird flying over a body of water.

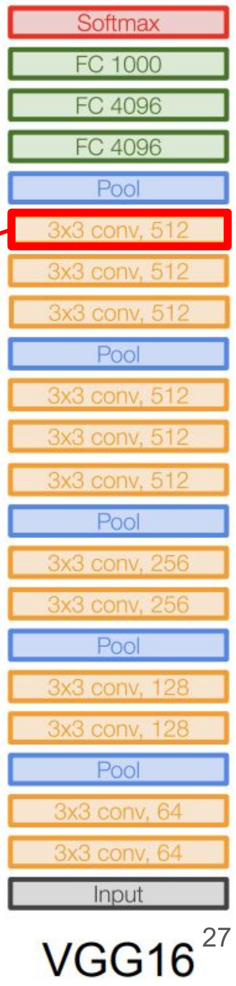
Word prediction is conditioned on parts of an image

Soft attention

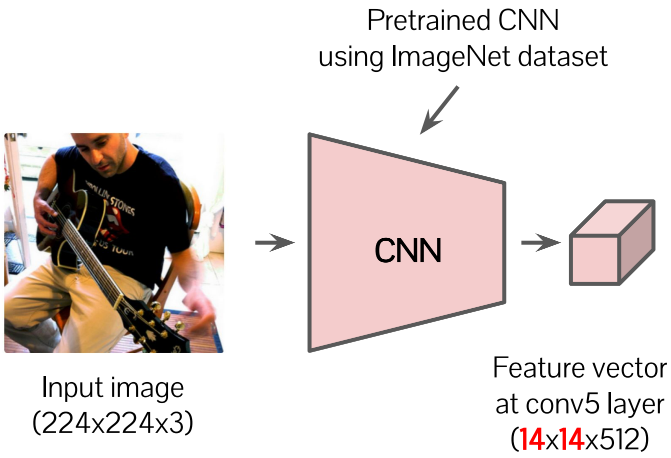


Attention-based image captioning





Attention-based image captioning

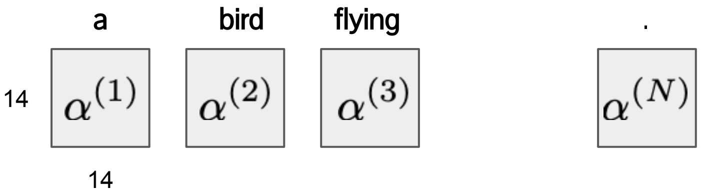


Attention:

- 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

a	bird	flying	.
$\alpha^{(1)}$	$\alpha^{(2)}$	$\alpha^{(3)}$	$\alpha^{(N)}$

14

14

- 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

a	bird	flying	.
$\alpha^{(1)}$	$\alpha^{(2)}$	$\alpha^{(3)}$	$\alpha^{(N)}$

14

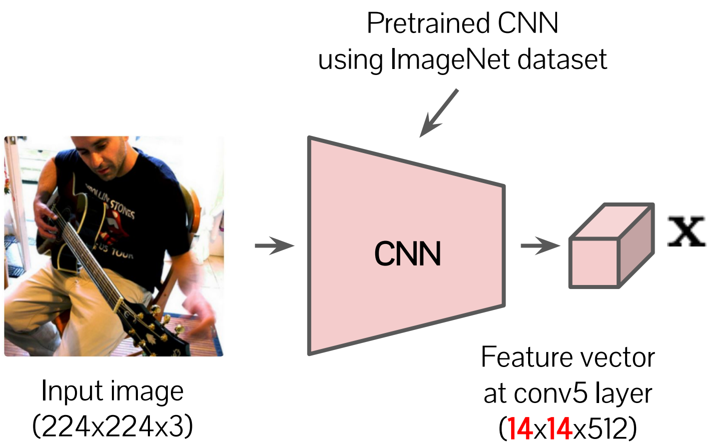
14

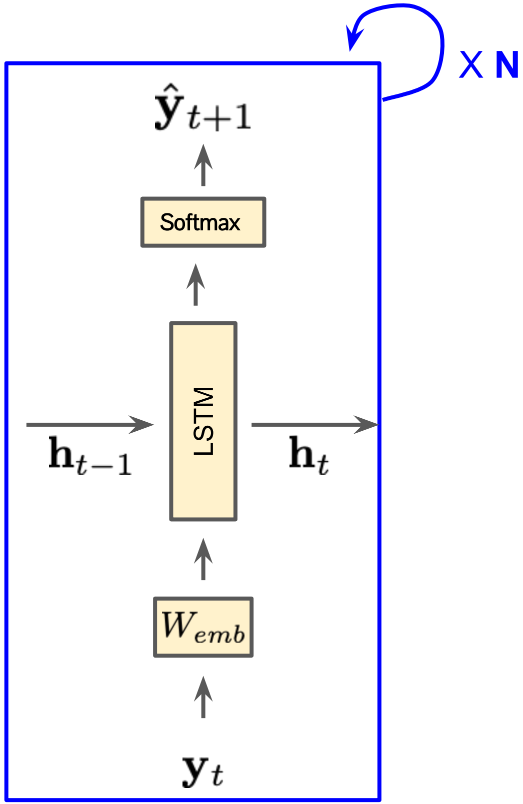
- Attention is used to abstract image feature

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

Attention-based image captioning

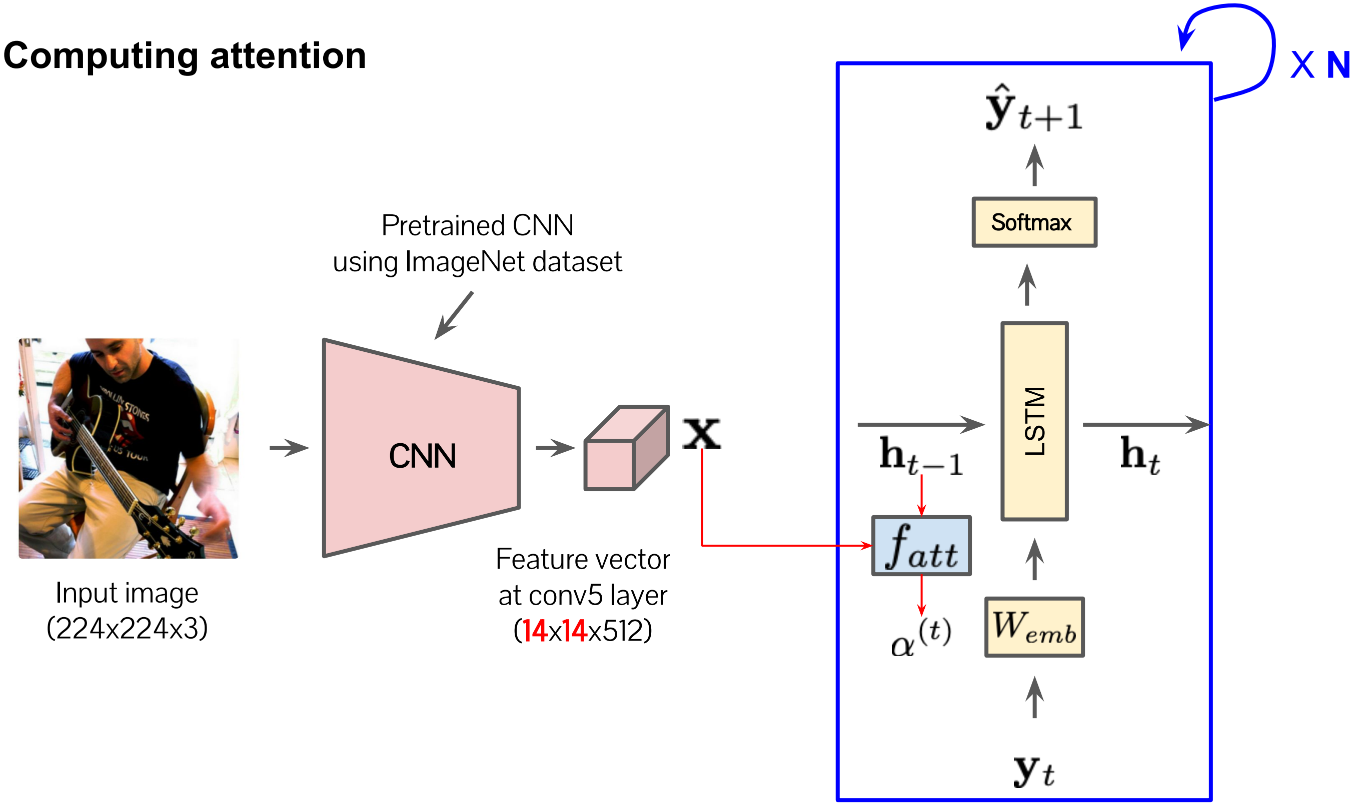
Computing attention





Attention-based image captioning

Computing attention



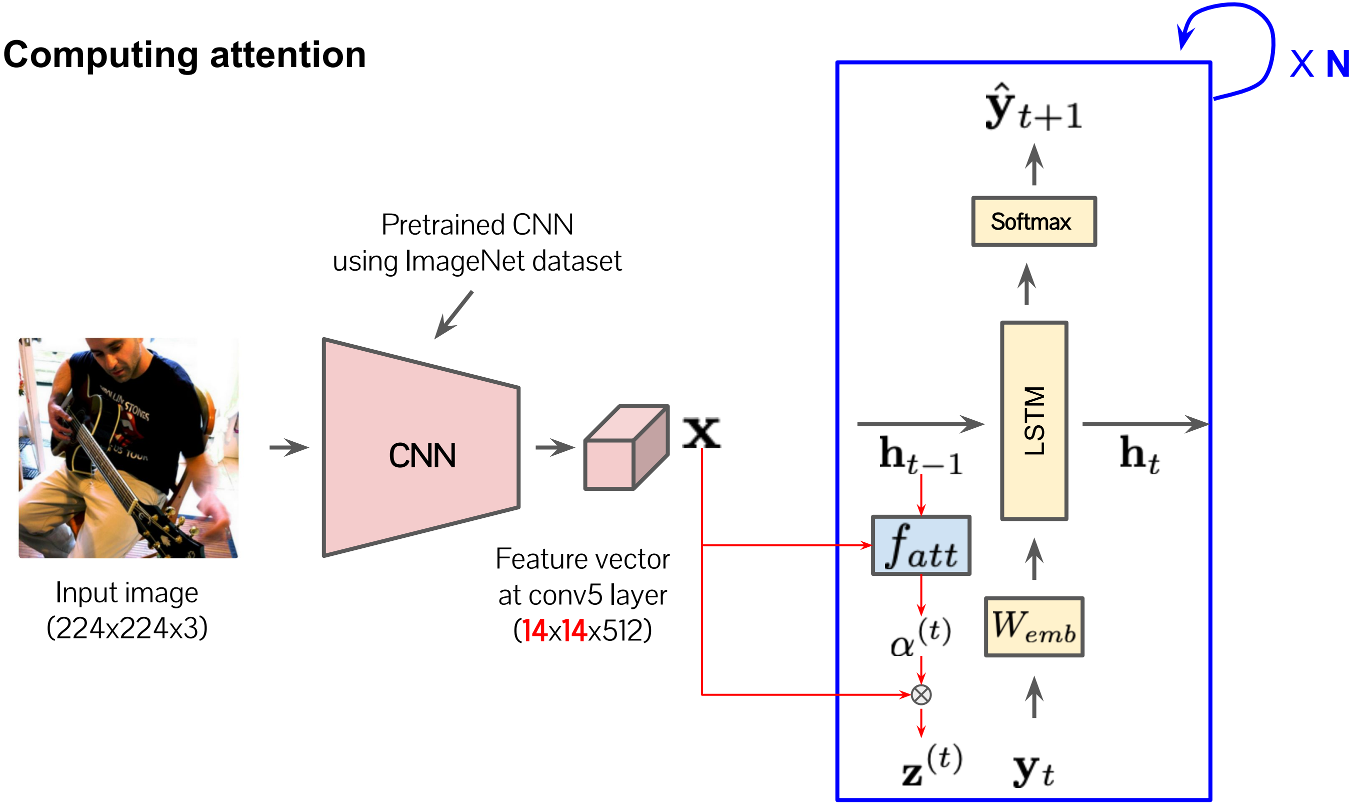
Attention module

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

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

Attention-based image captioning

Computing attention



Attention module

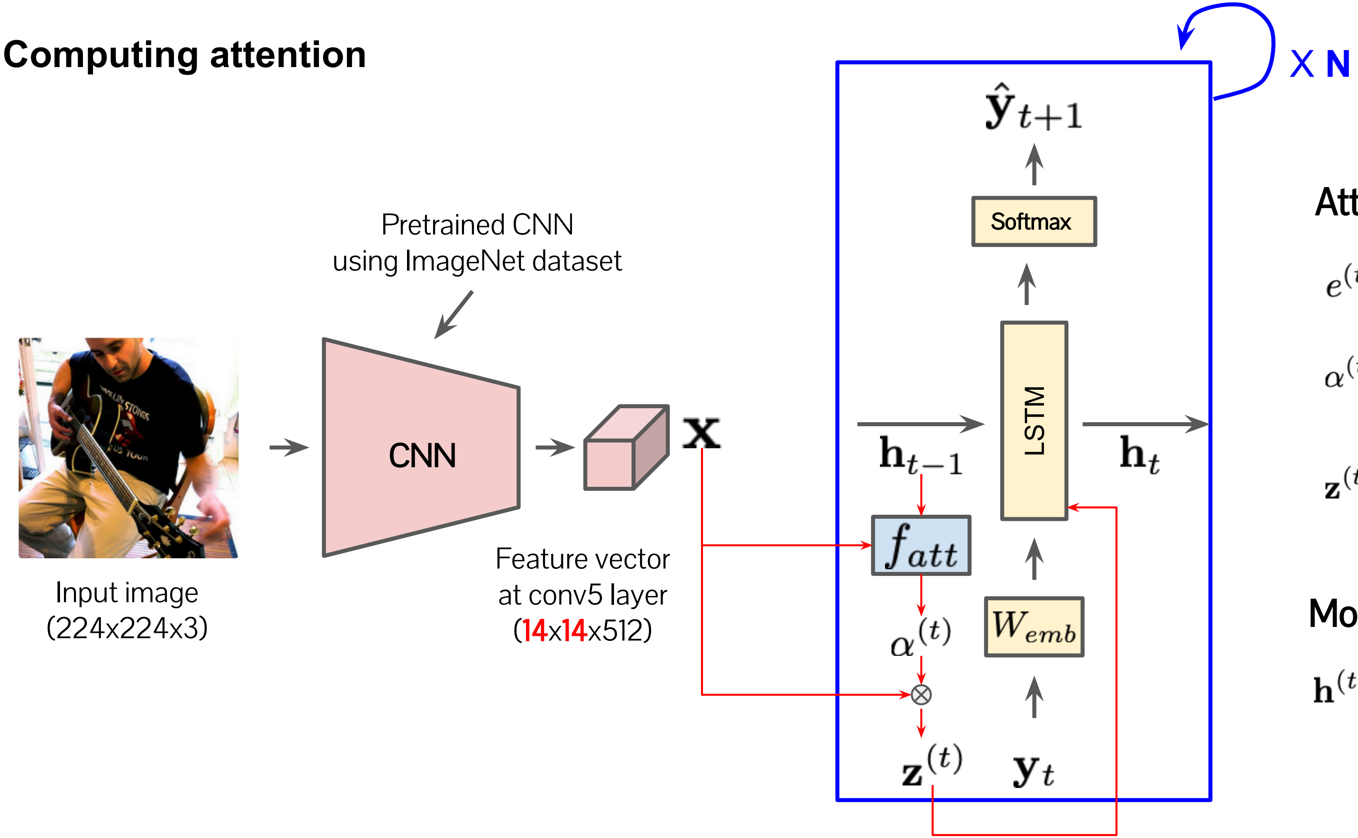
$$e^{(t)} = f_{att}(\mathbf{x}, \mathbf{h}_{t-1})$$

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

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

Attention-based image captioning

Computing attention



Attention module

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

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

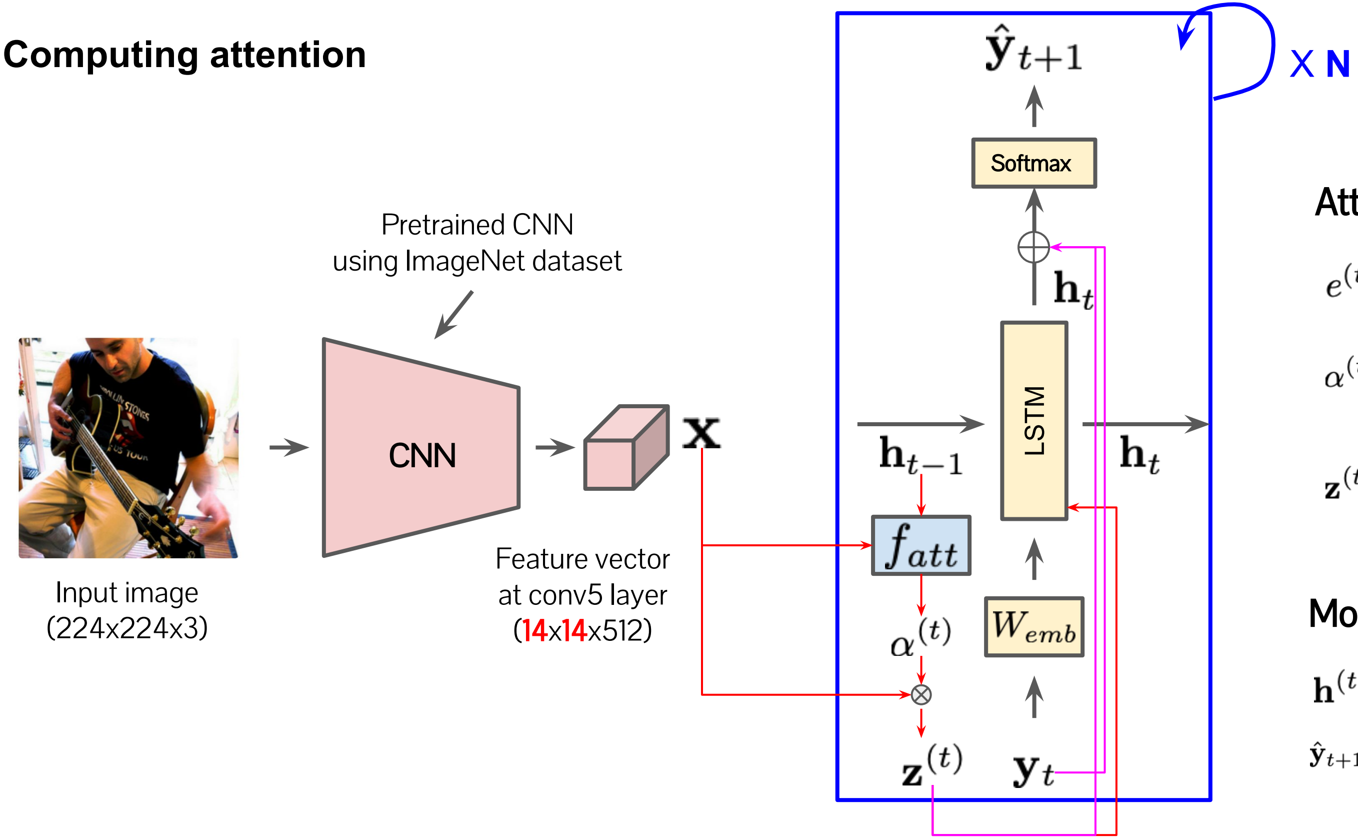
$$\mathbf{z}^{(t)} = \sum_{i,j} \alpha^{(t)}_{i,j} \mathbf{x}_{i,j}$$

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

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

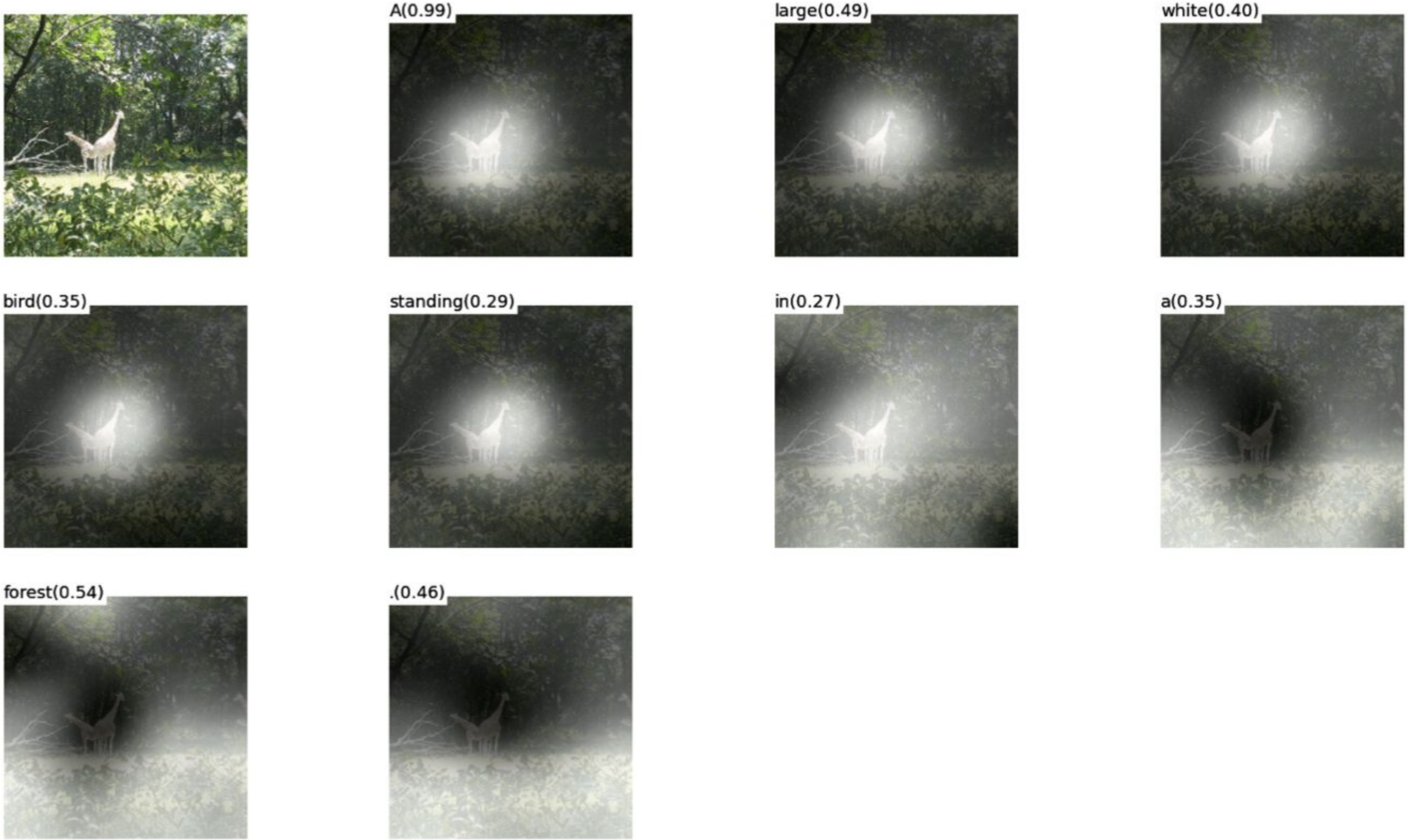
$$\alpha^{(t)} = \frac{\exp(e^{(t)})}{\sum_{i,j} \exp(e^{(t)}_{i,j})}$$

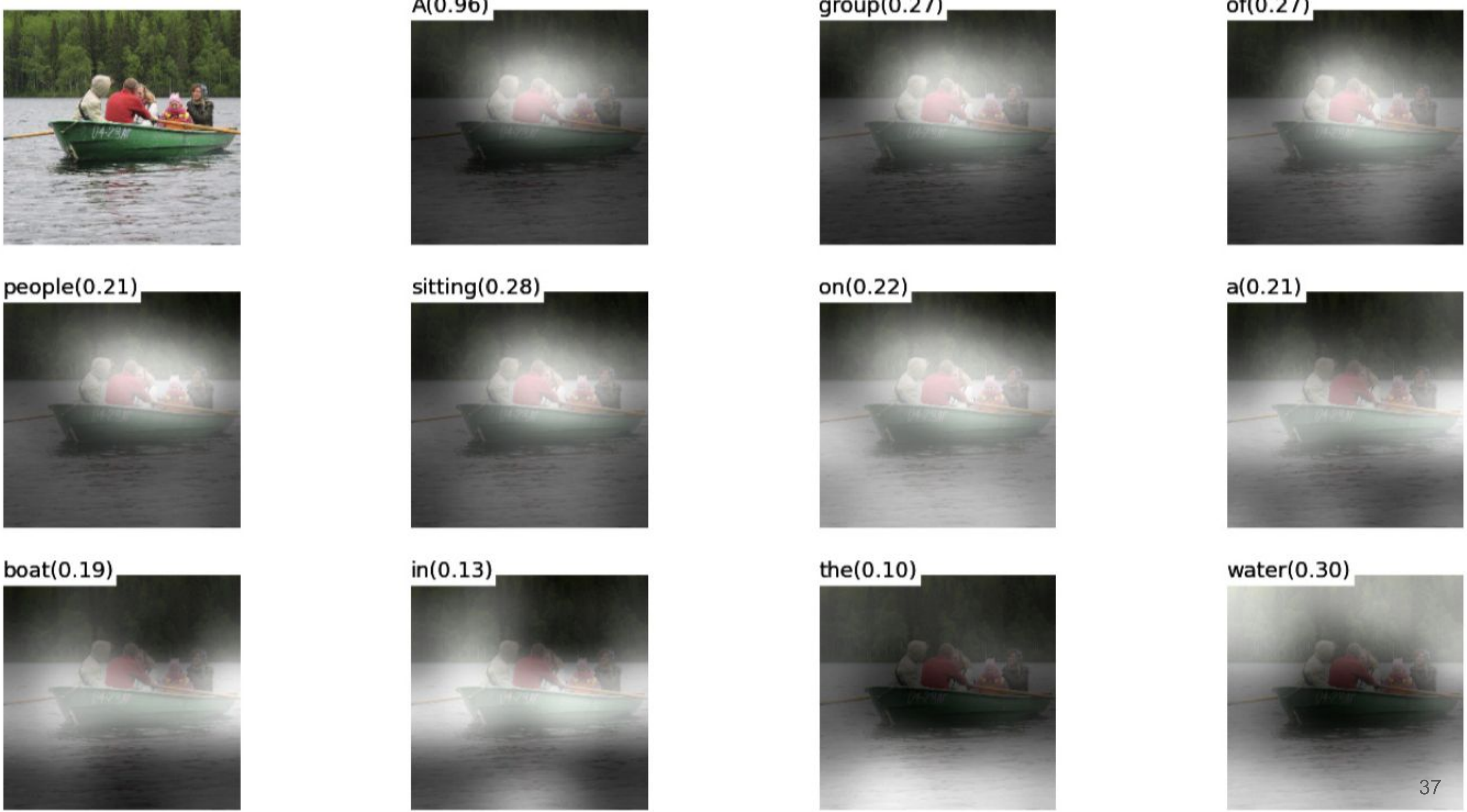
$$\mathbf{z}^{(t)} = \sum_{i,j} \alpha^{(t)}_{i,j} \mathbf{x}_{i,j}$$

Modified LSTM

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

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


A(0.99)
large(0.49)
white(0.40)
bird(0.35)
standing(0.29)
in(0.27)
a(0.35)
forest(0.54)
.(0.46)


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)

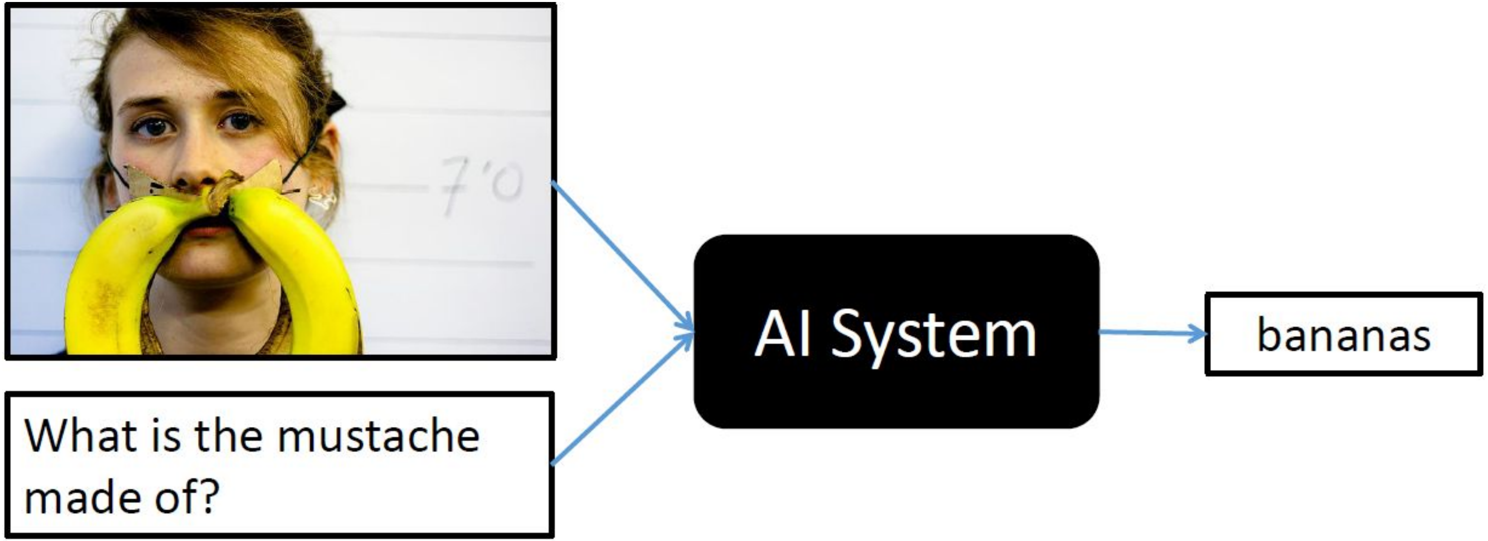

A(1.00)
woman(0.80)
holding(0.68)
a(0.58)
clock(0.62)
in(0.45)
her(0.39)
hand(0.64)
.(0.37)

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

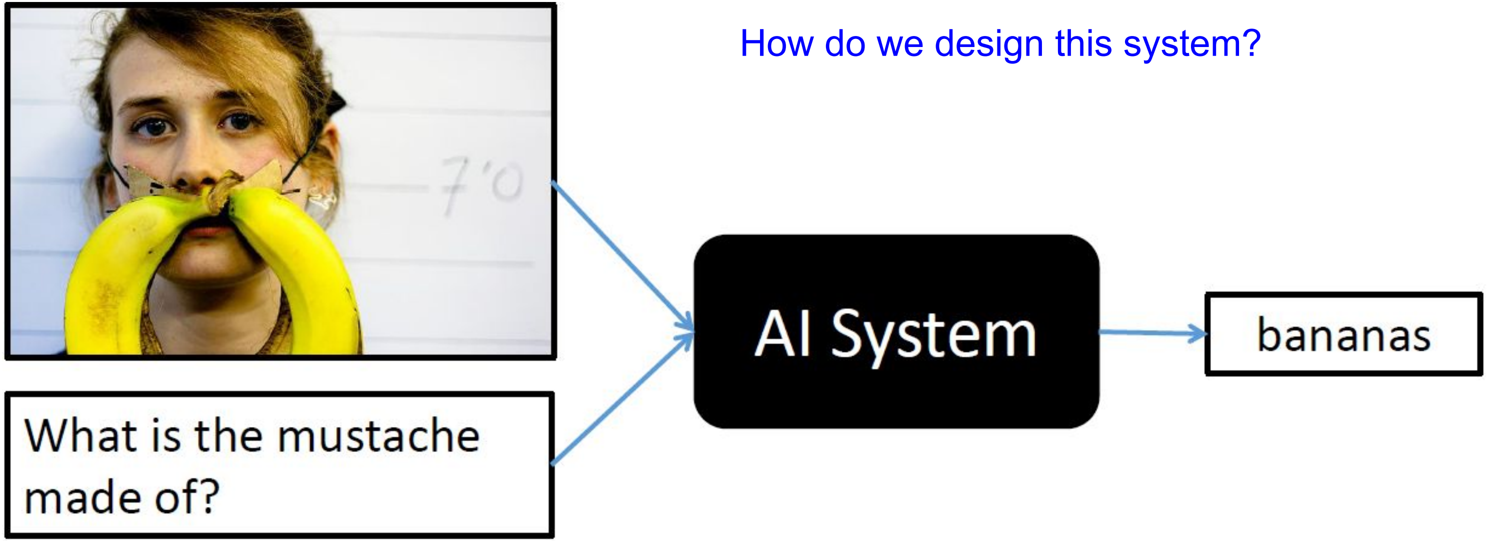
- Objective: given an image and a question about an image, predict an answer.



Visual question answering

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

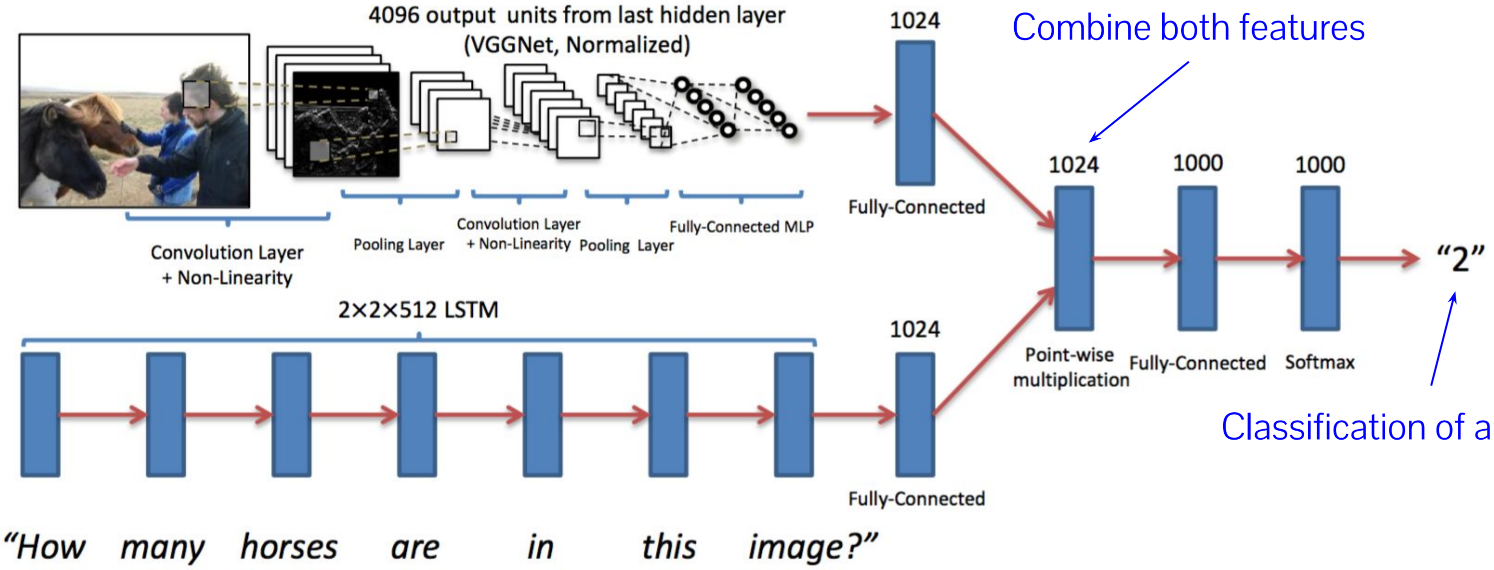
How do we design this system?



Visual question answering

Encoding image using CNN

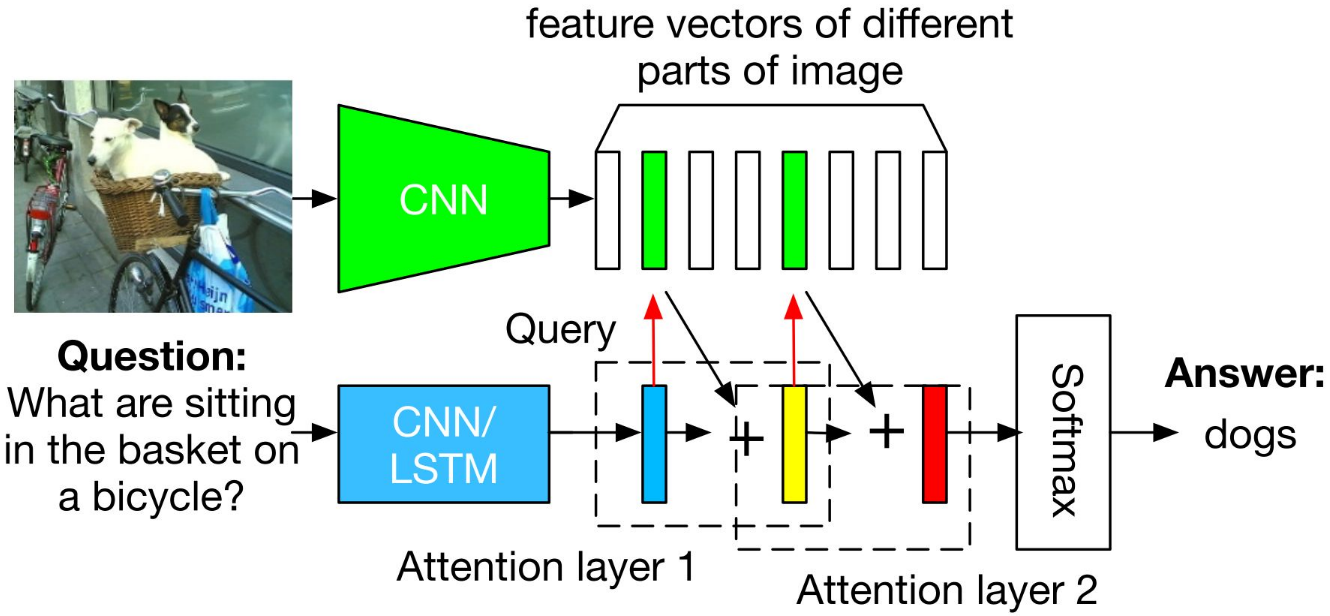
Combine both features



Classification of answer

Encoding question using LSTM

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