THE UNIVERSITY OF BRITISH COLUMBIA

Topics in AI (CPSC 532S):
Multimodal Learning with Vision, Language and Sound

Lecture 12: RNN Applications

Logistics

Assignment 4 is out … is due March 8th

— You have a choice of implementing <u>**one**</u> of two parts

— You can start on the assignment today

Logistics

Project Groups — Group formation survey will go out today/tomorrow. Groups formed by early next week. Fill out one survey per "Group". If you don't have a group fill it out as an individual. The group will be assigned to you.

Survey Option — Instructions coming today/tomorrow. Read: Deep Audio-Visual Learning: A Survey (https://arxiv.org/pdf/2001.04758)

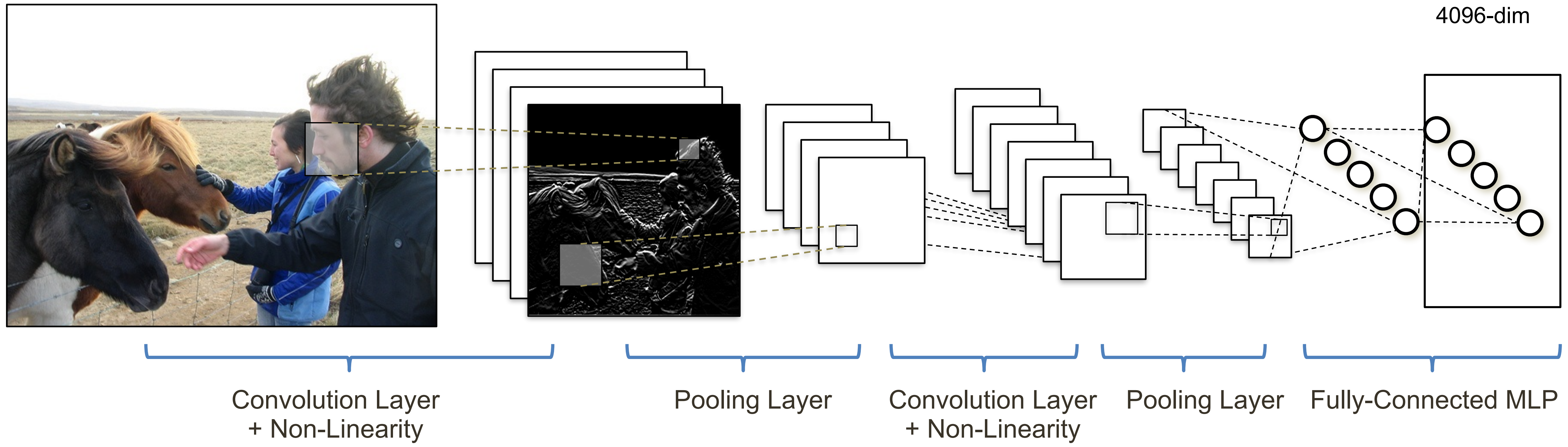
Project Proposals — due **March 12th** (2-4 pages; 4 pages is a hard max)

Applications: Neural Image Captioning

* slide from Dhruv Batra

Applications: Neural Image Captioning

Image Embedding (VGGNet)

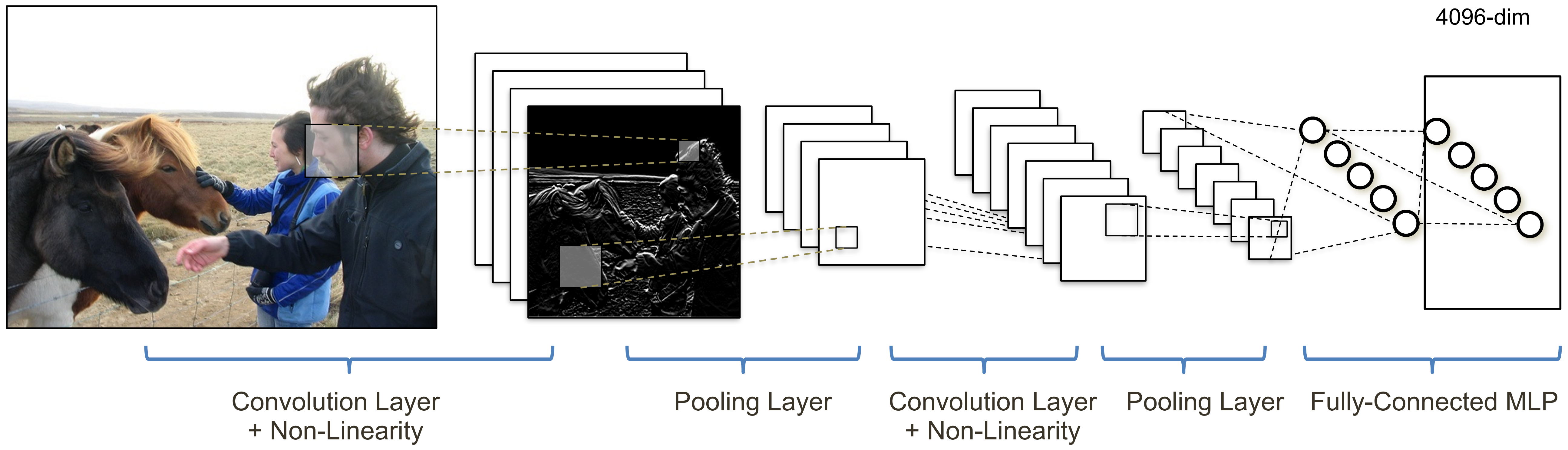


Assignment 2: Load the VGG-16 model, remove last layer

* slide from Dhruv Batra

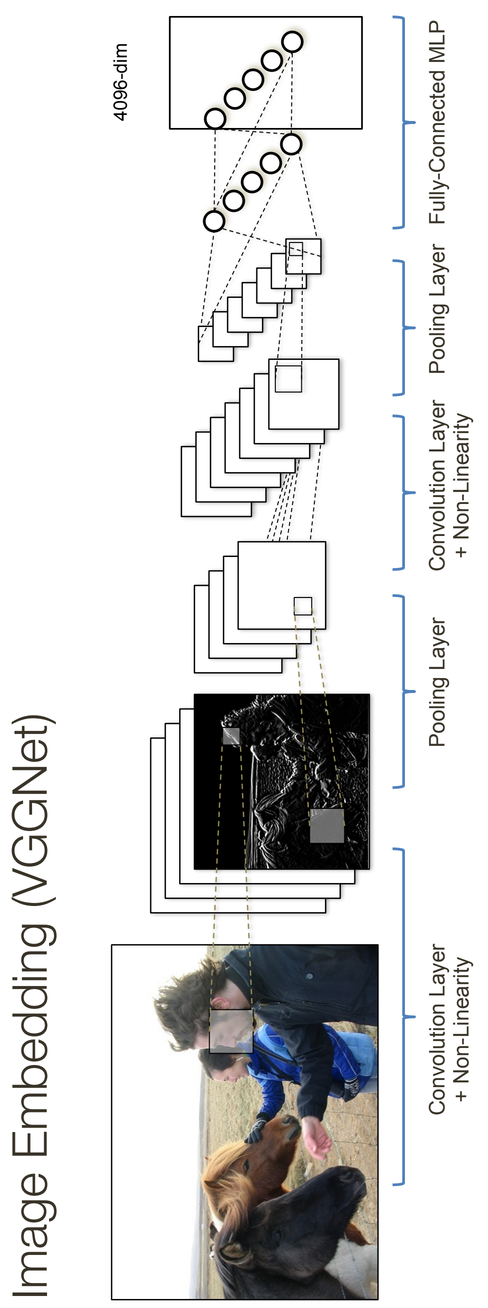
Applications: Neural Image Captioning

Image Embedding (VGGNet)



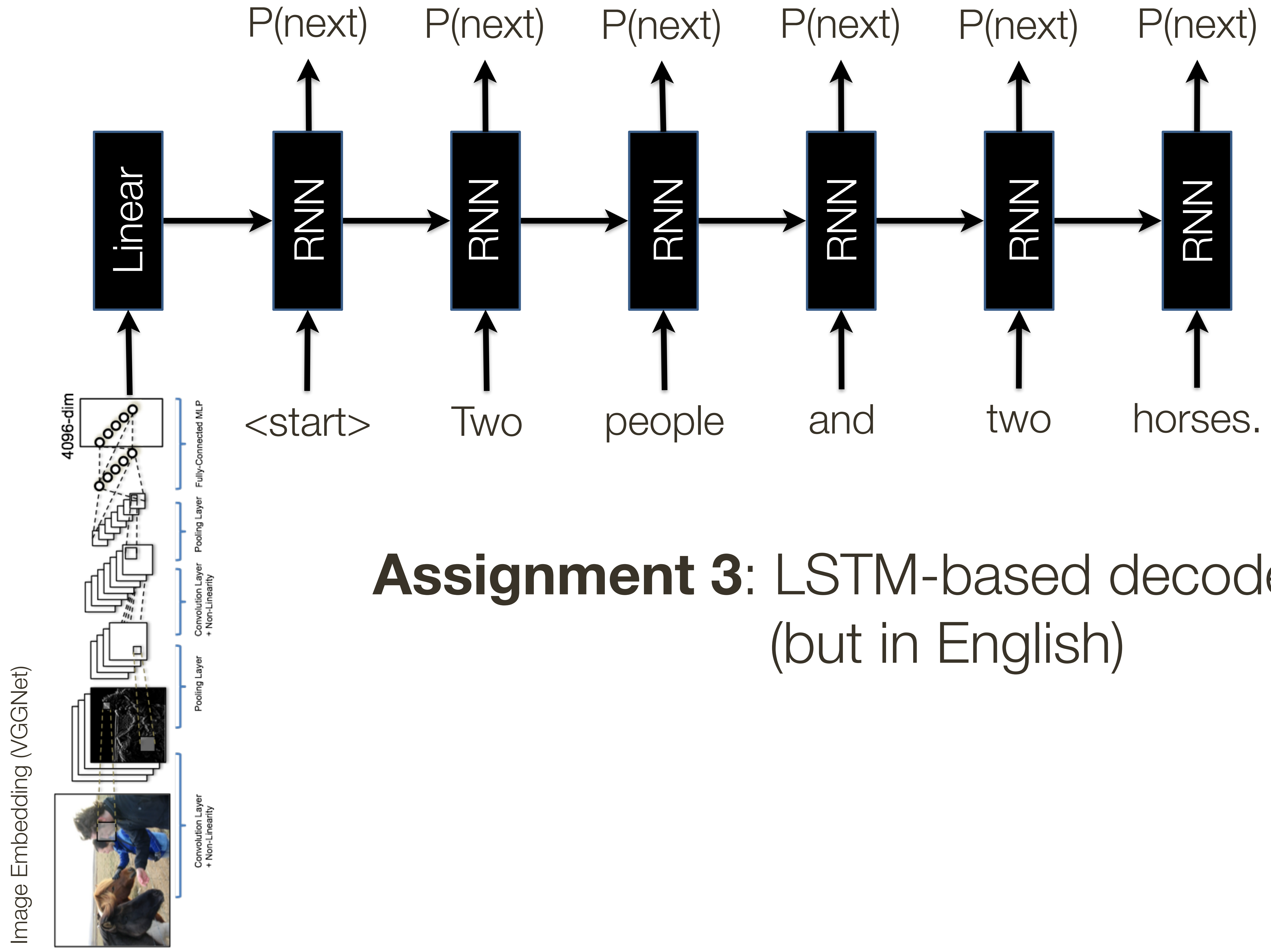
* slide from Dhruv Batra

Applications: Neural Image Captioning



* slide from Dhruv Batra

Applications: Neural Image Captioning



Assignment 3: LSTM-based decoder
(but in English)

* slide from Dhruv Batra

Applications: Neural Image Captioning

Good results

A cat sitting on a suitcase on the floor

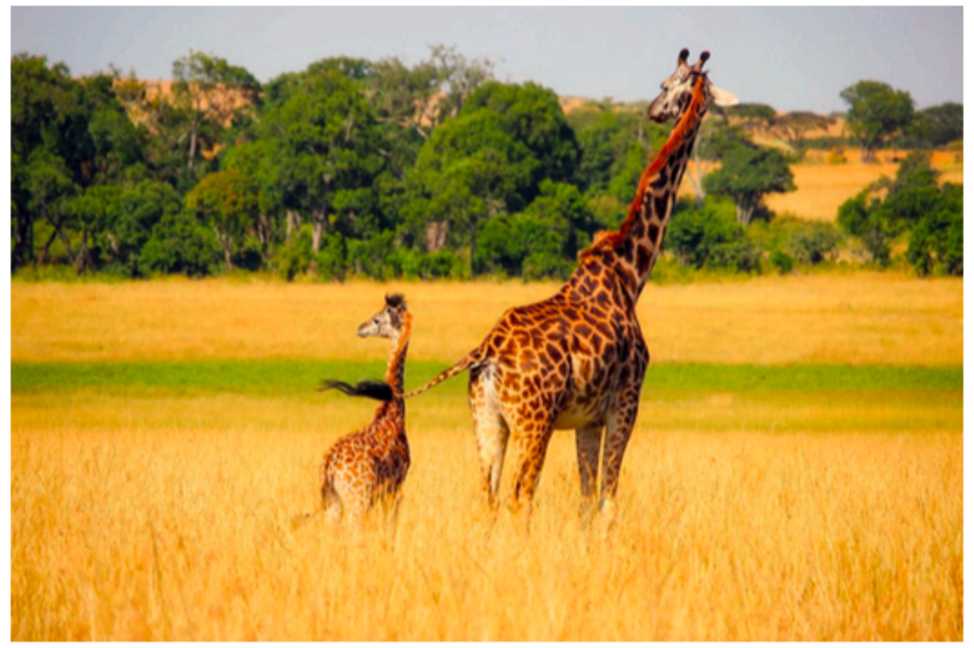
A cat is sitting on a tree branch

A dog is running in the grass with a frisbee

A white teddy bear sitting in the grass

Two people walking on the beach with surfboards

A tennis player in action on the court



Two giraffes standing in a grassy field

A man riding a dirt bike on a dirt track

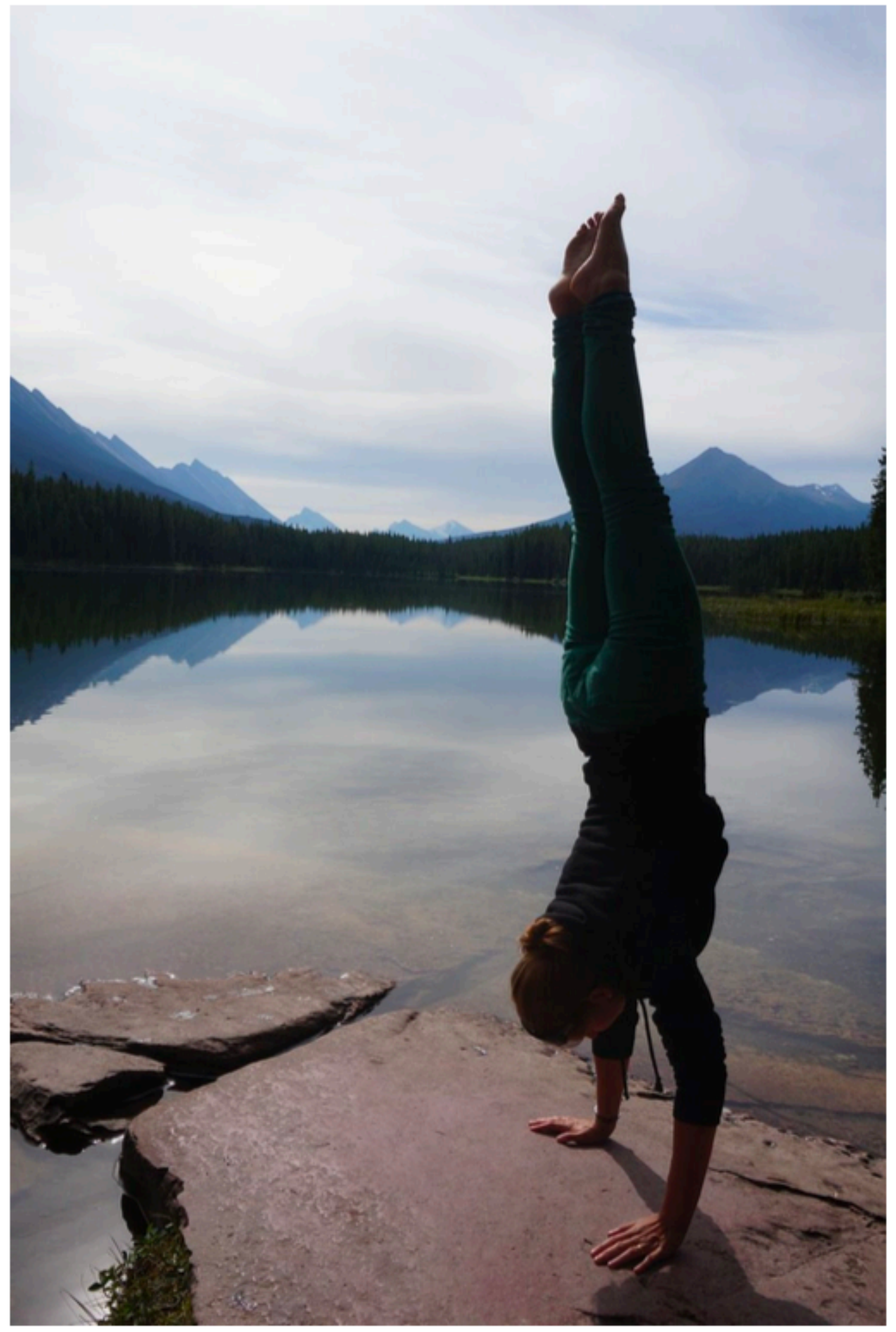
* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Applications: Neural Image Captioning

Failure cases

A woman is holding a cat in her hand

A person holding a computer mouse on a desk



A woman standing on a beach holding a surfboard

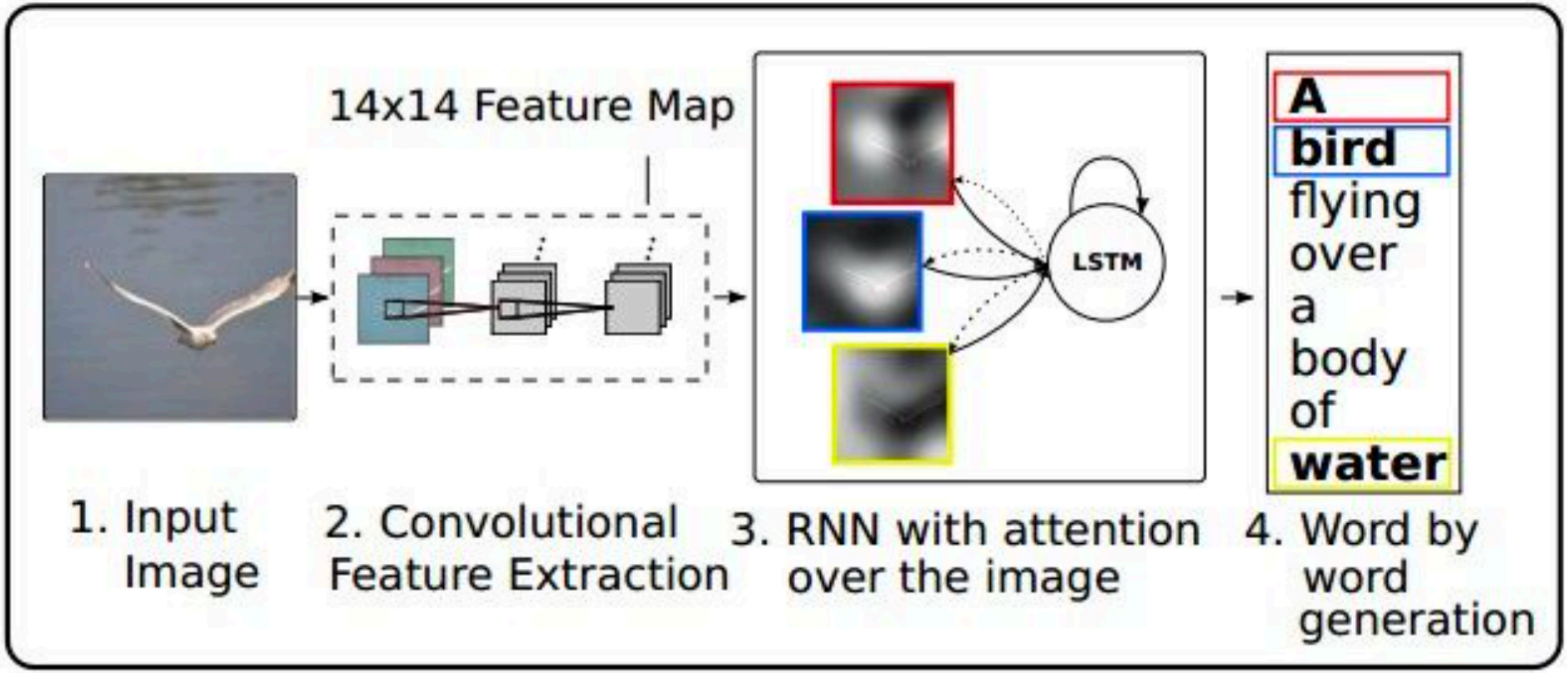
A bird is perched on a tree branch

A man in a baseball uniform throwing a ball

Applications: Image Captioning with Attention

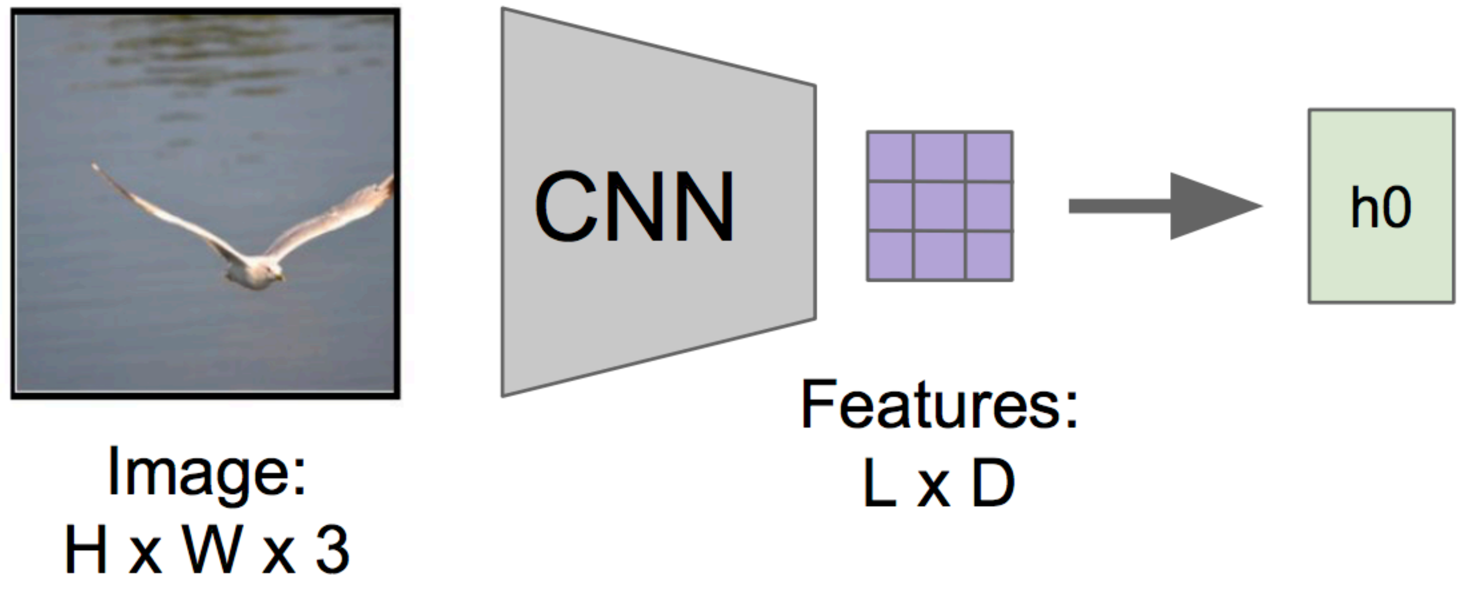
[Xu et al., ICML 2015]

RNN focuses its attention at a different spatial location when generating each word



Applications: Image Captioning with Attention

[Xu et al., ICML 2015]



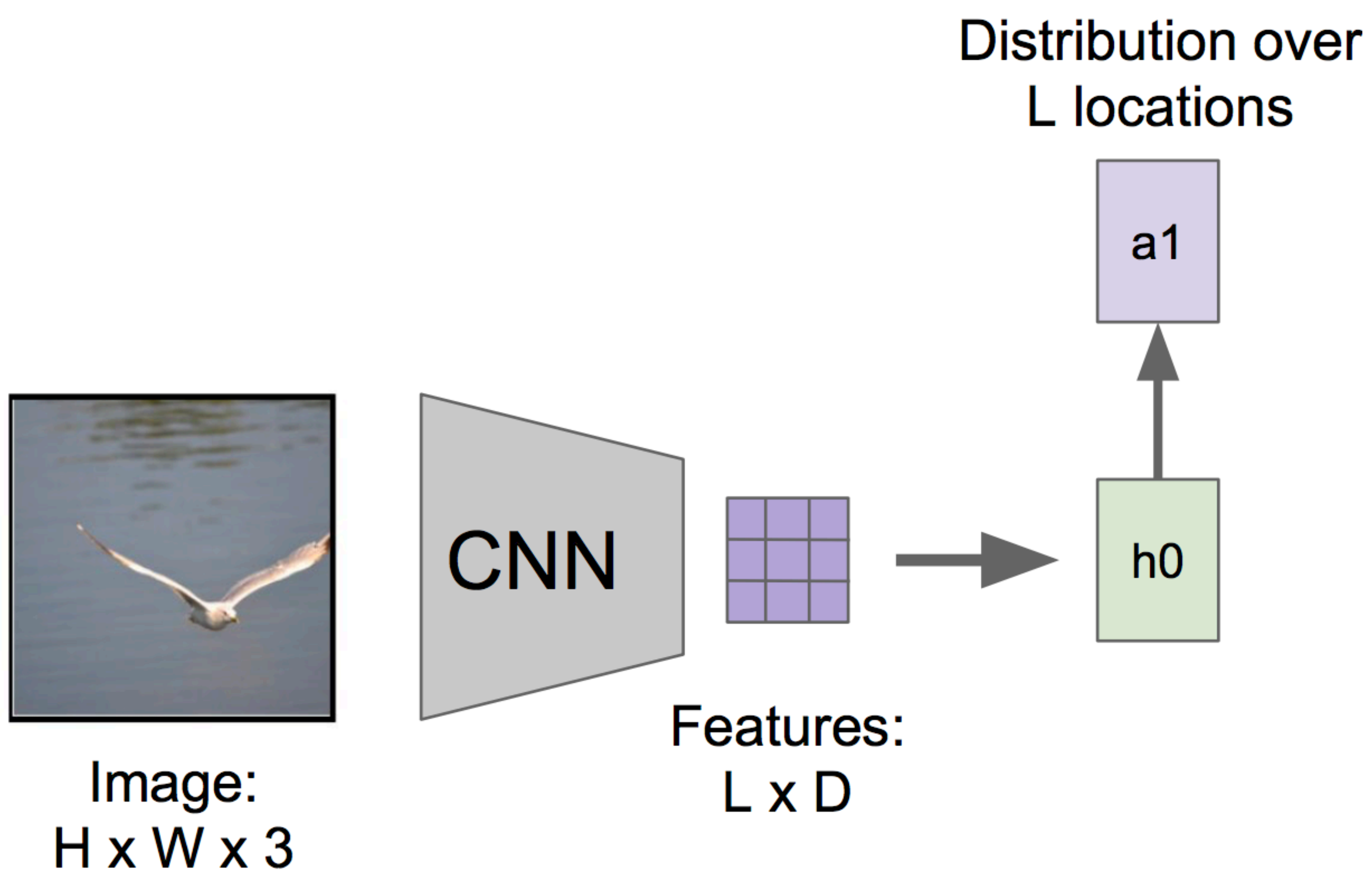
(7x7x512) = (49x512)

or

(1x7x7x512) = (1x49x512)

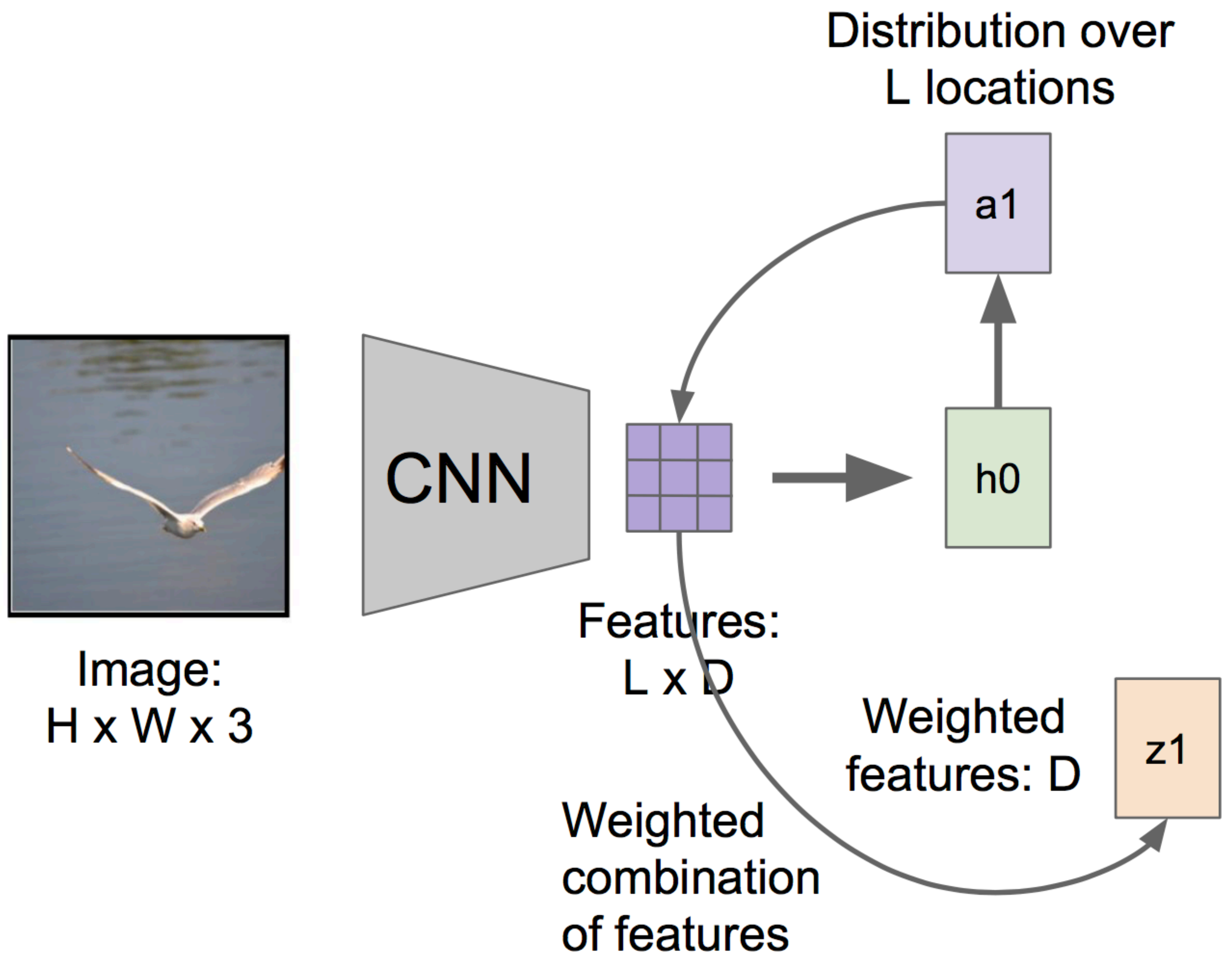
Applications: Image Captioning with Attention

[Xu et al., ICML 2015]



Applications: Image Captioning with Attention

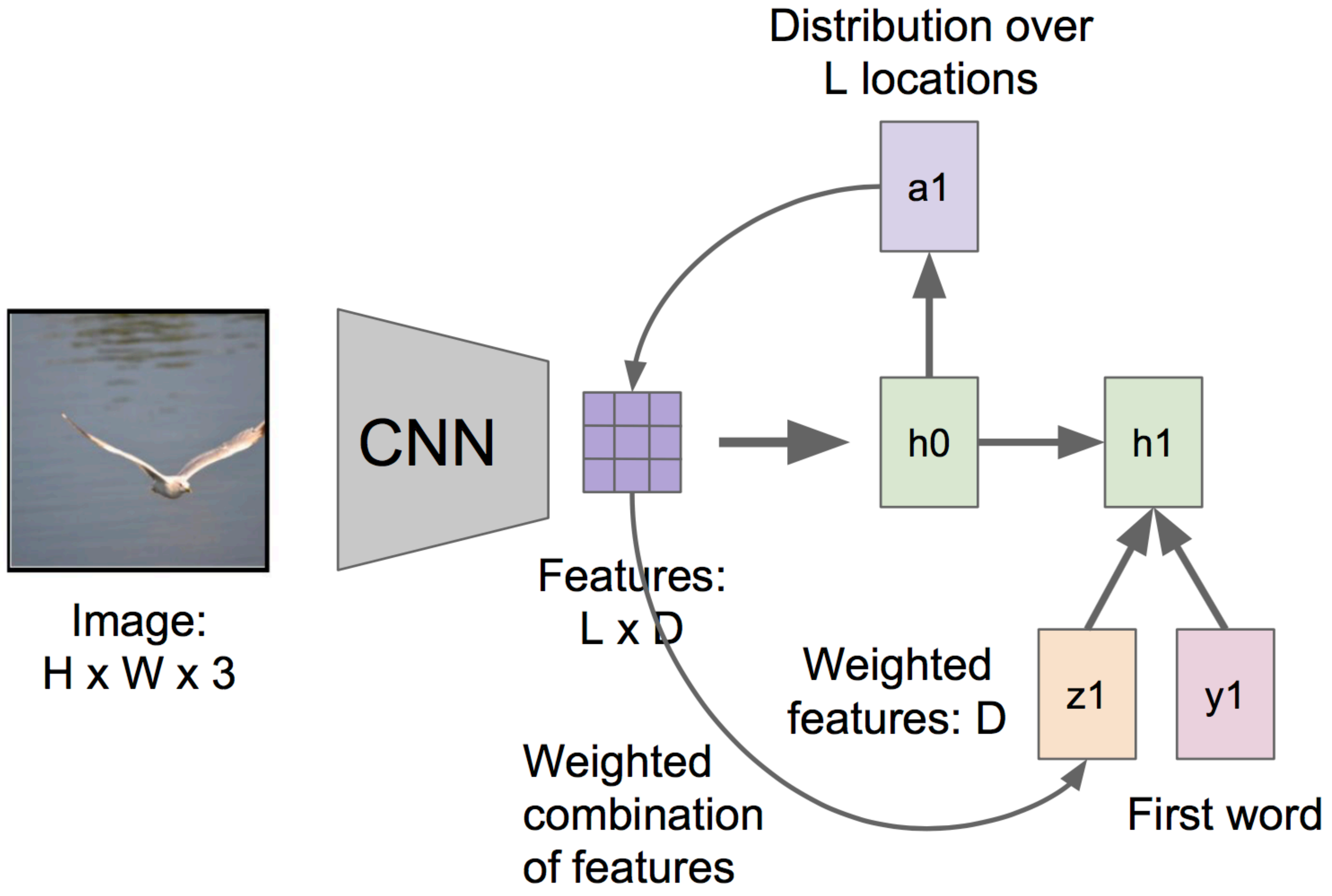
[Xu et al., ICML 2015]



$$z = \sum_{i=1}^{L} p_i v_i$$

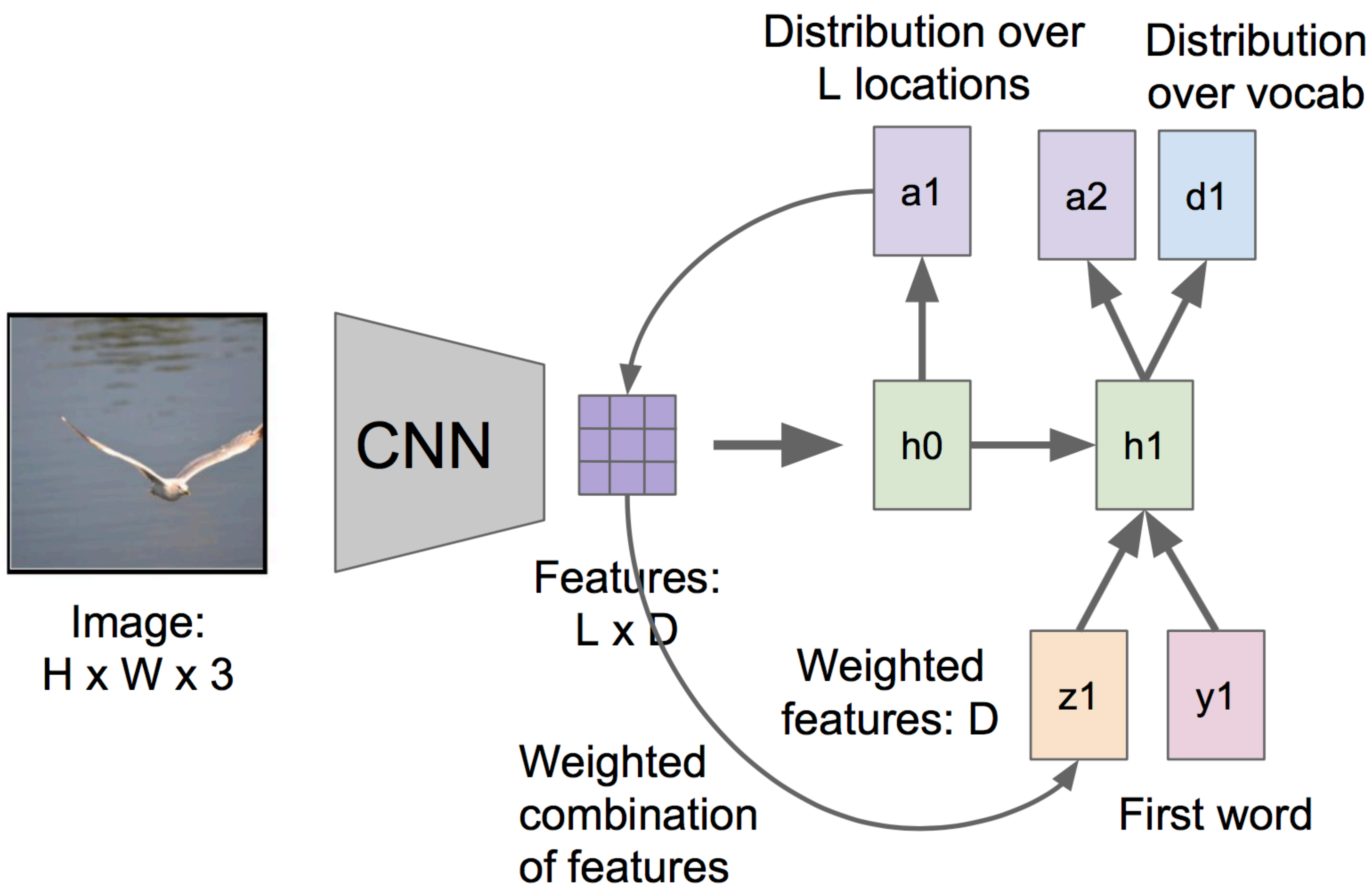
Applications: Image Captioning with Attention

[Xu et al., ICML 2015]



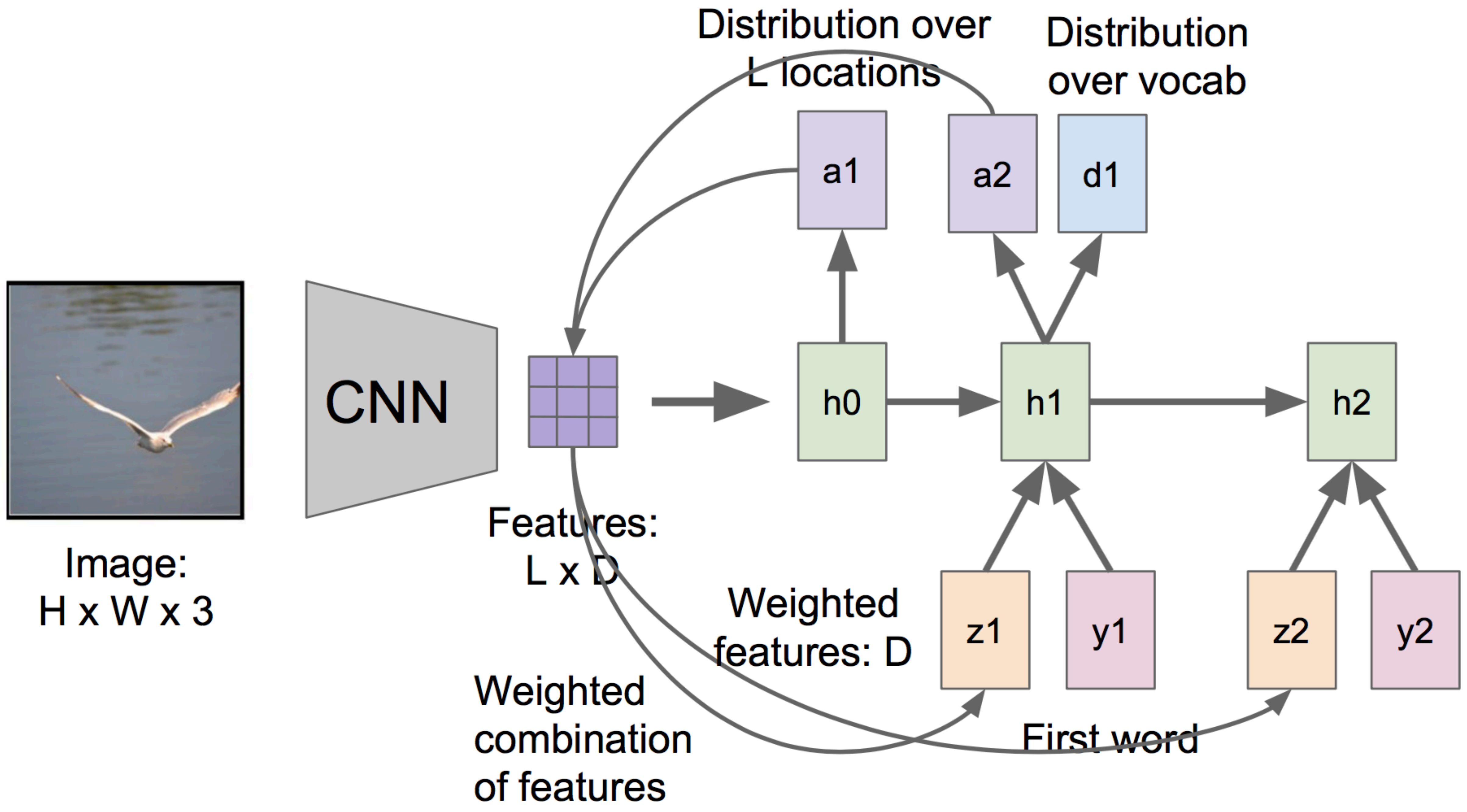
Applications: Image Captioning with Attention

[Xu et al., ICML 2015]



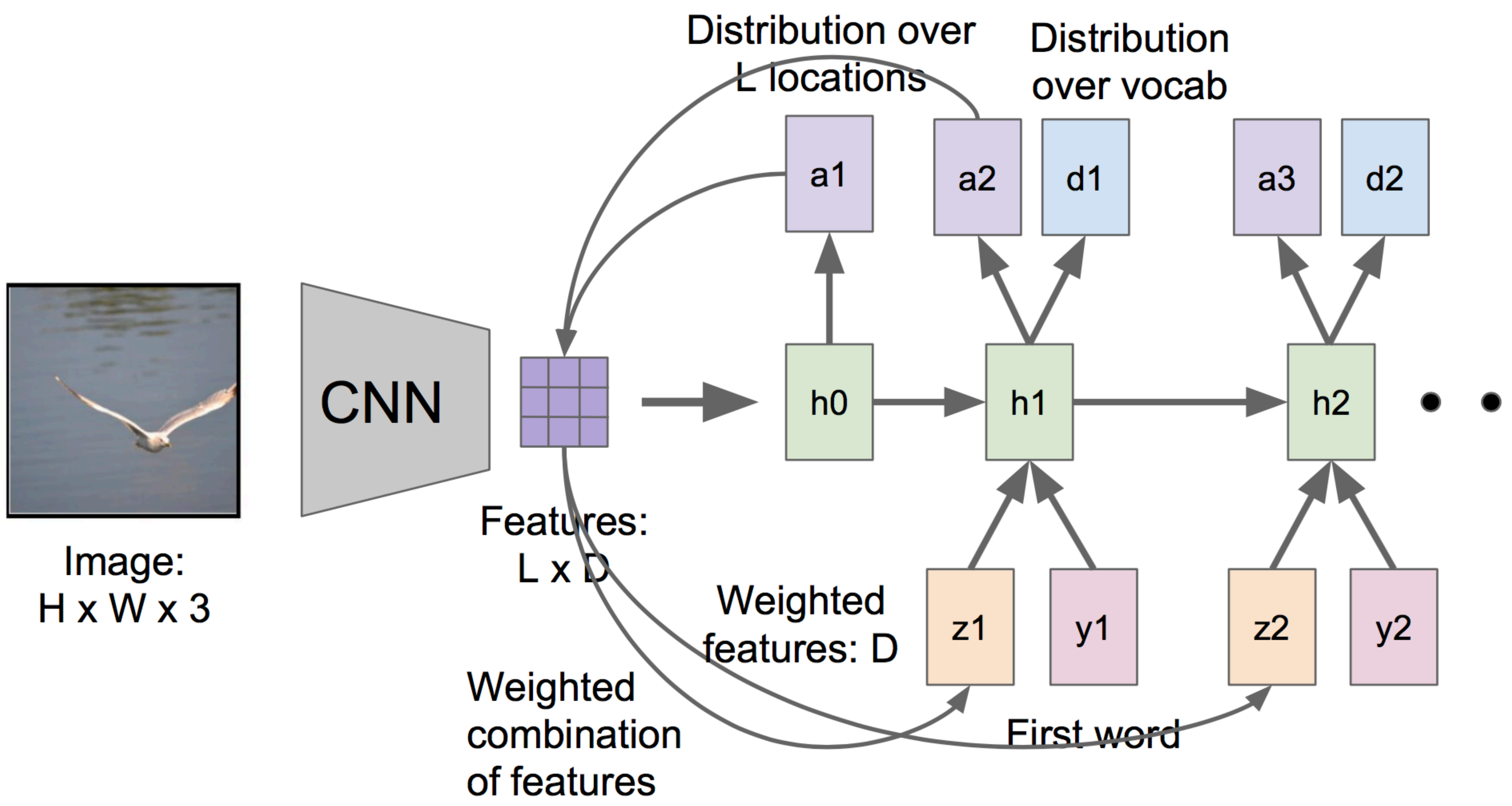
Applications: Image Captioning with Attention

[Xu et al., ICML 2015]



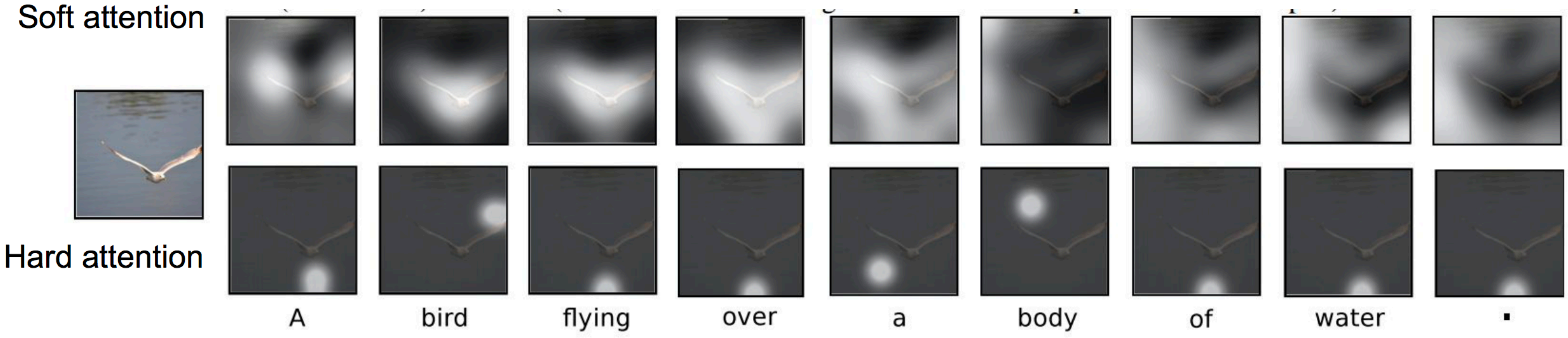
Applications: Image Captioning with Attention

[Xu et al., ICML 2015]



Applications: Image Captioning with Attention

[Xu et al., ICML 2015]



Applications: Image Captioning with Attention

[Xu et al., ICML 2015]

Good results

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.

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

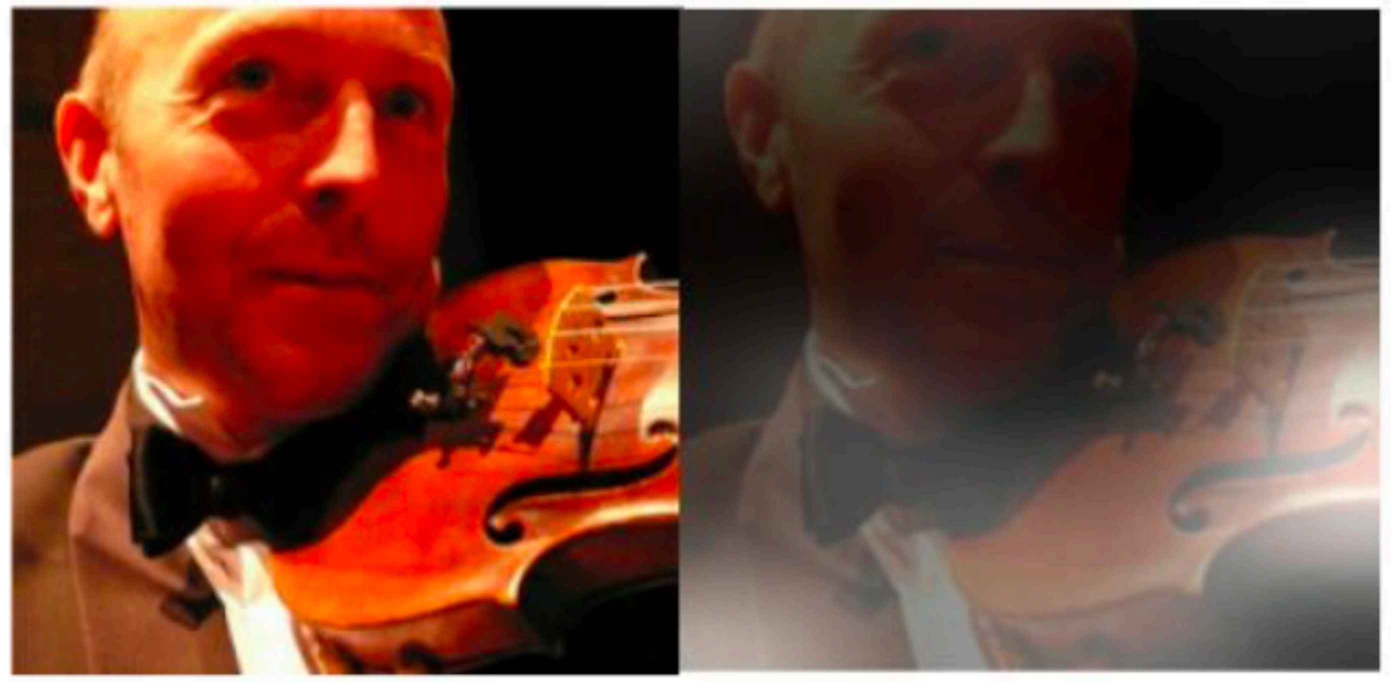
Applications: Image Captioning with Attention

[Xu et al., ICML 2015]

Failure results

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.

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Applications: Typical Visual Question Answering (VQA)

Image

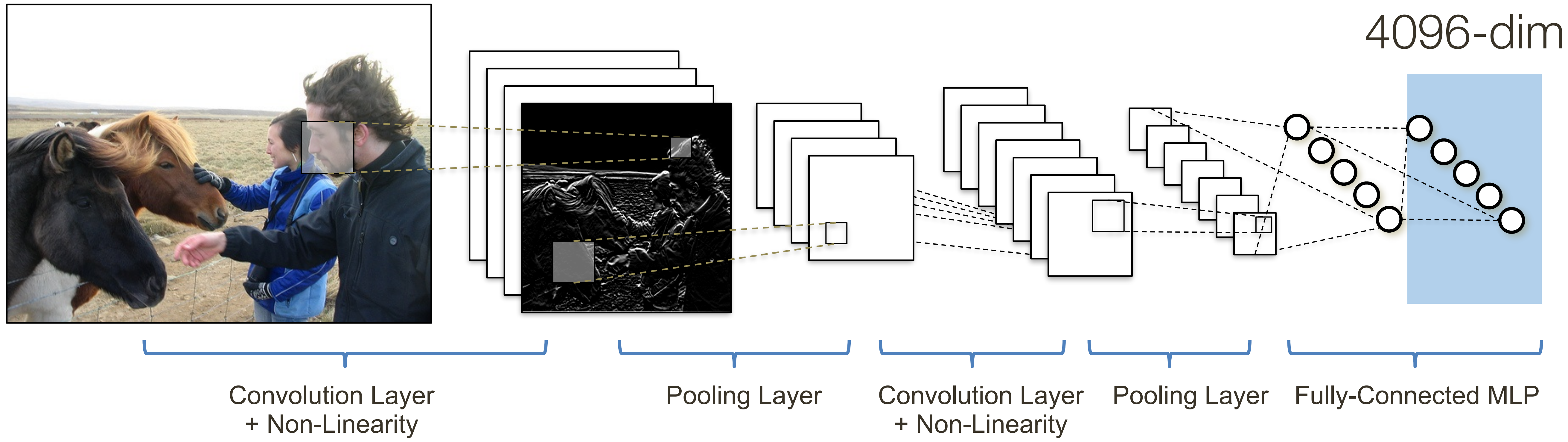
Question

"How many horses are in this image?"

* slide from Dhruv Batra

Applications: Typical Visual Question Answering (VQA)

Image Embedding (VGGNet)



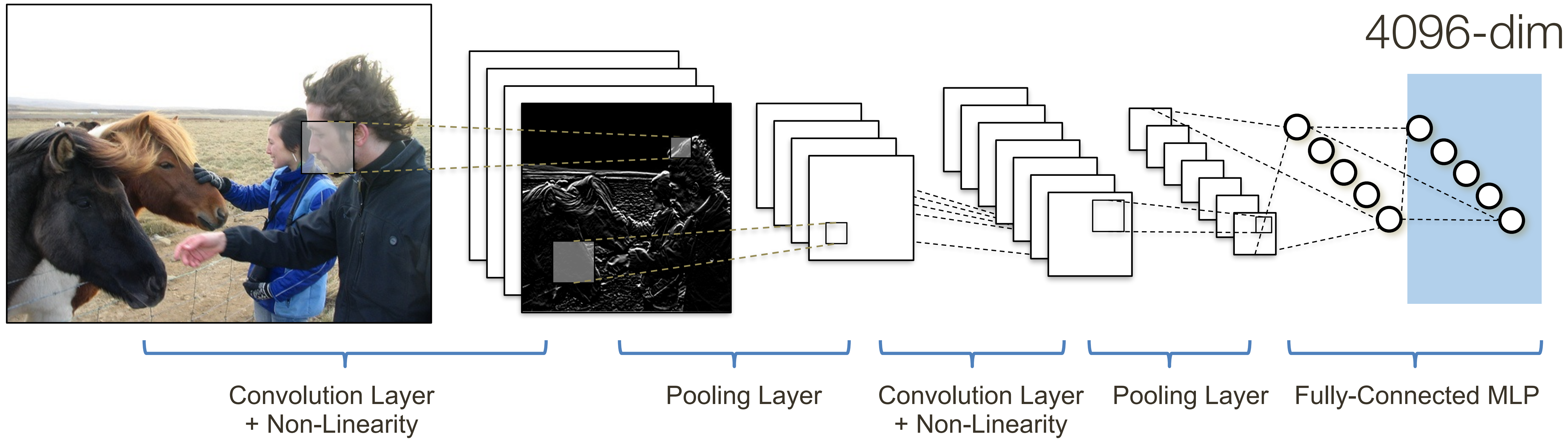
Question

"How many horses are in this image?"

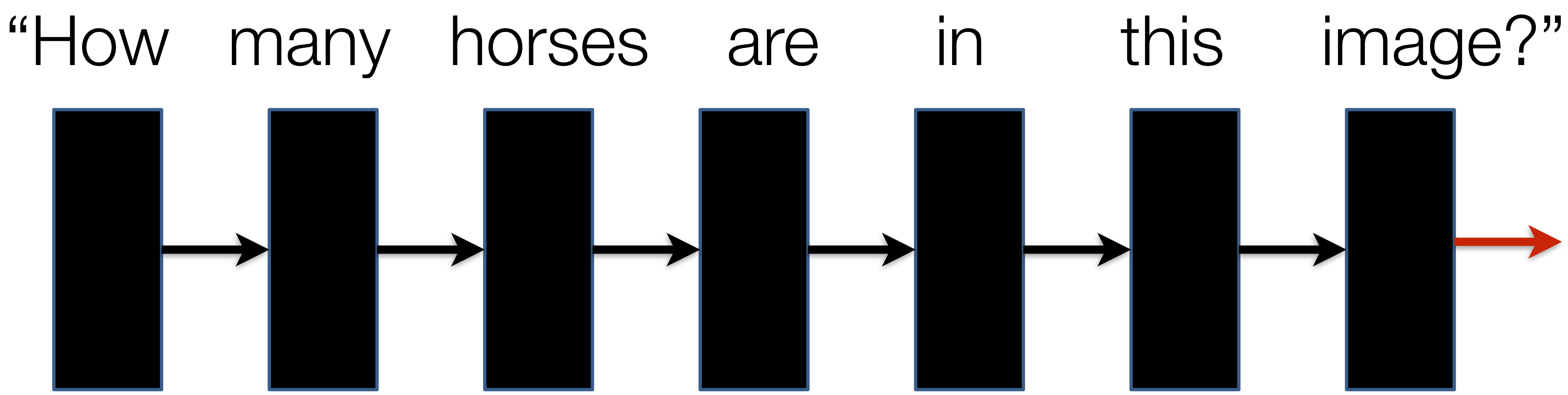
* slide from Dhruv Batra

Applications: Typical Visual Question Answering (VQA)

Image Embedding (VGGNet)

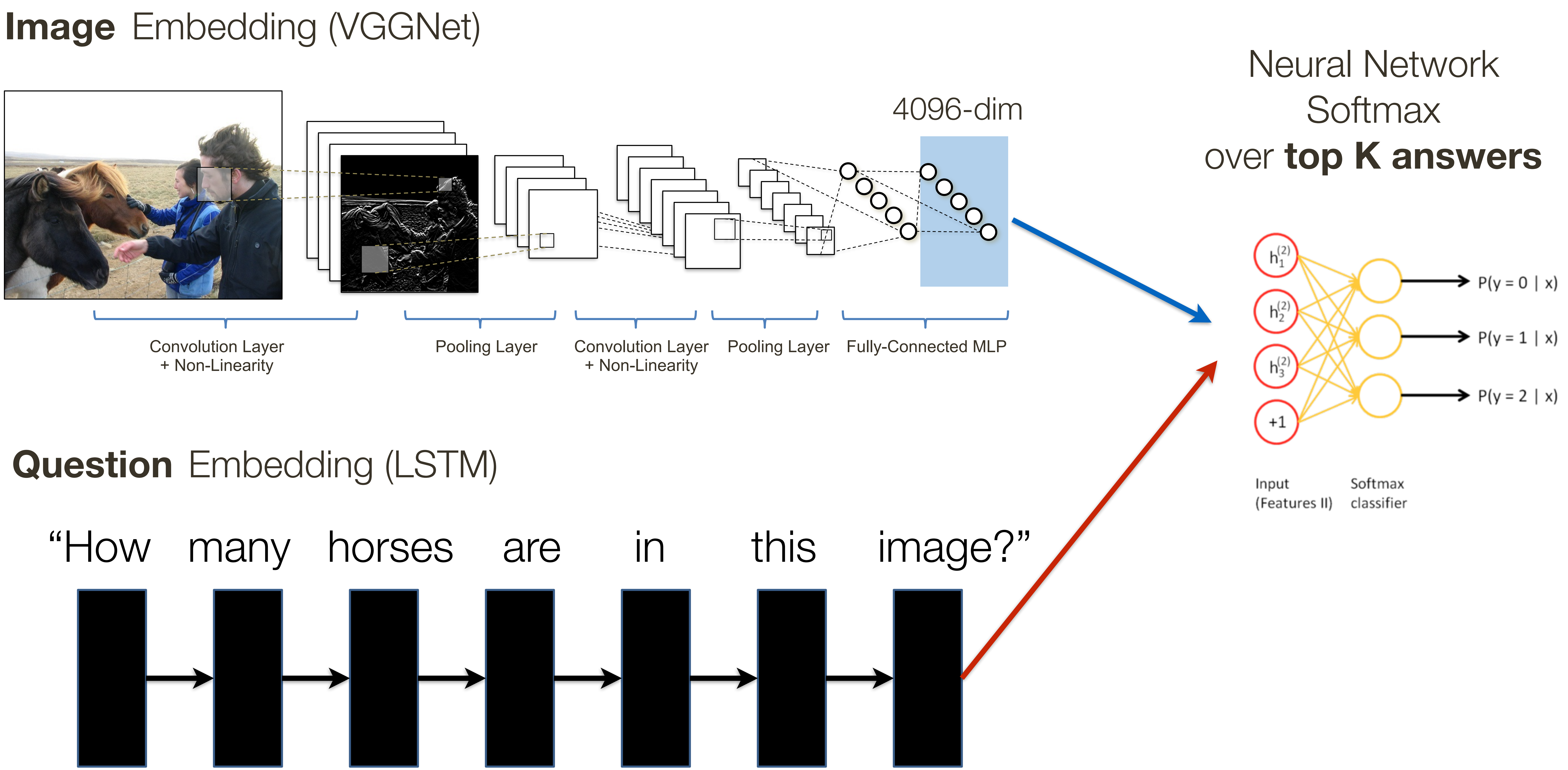


Question Embedding (LSTM)



* slide from Dhruv Batra

Applications: Typical Visual Question Answering (VQA)



* slide from Dhruv Batra

Applications: Visual Dialogs

[Seo et al., NIPS 2017]

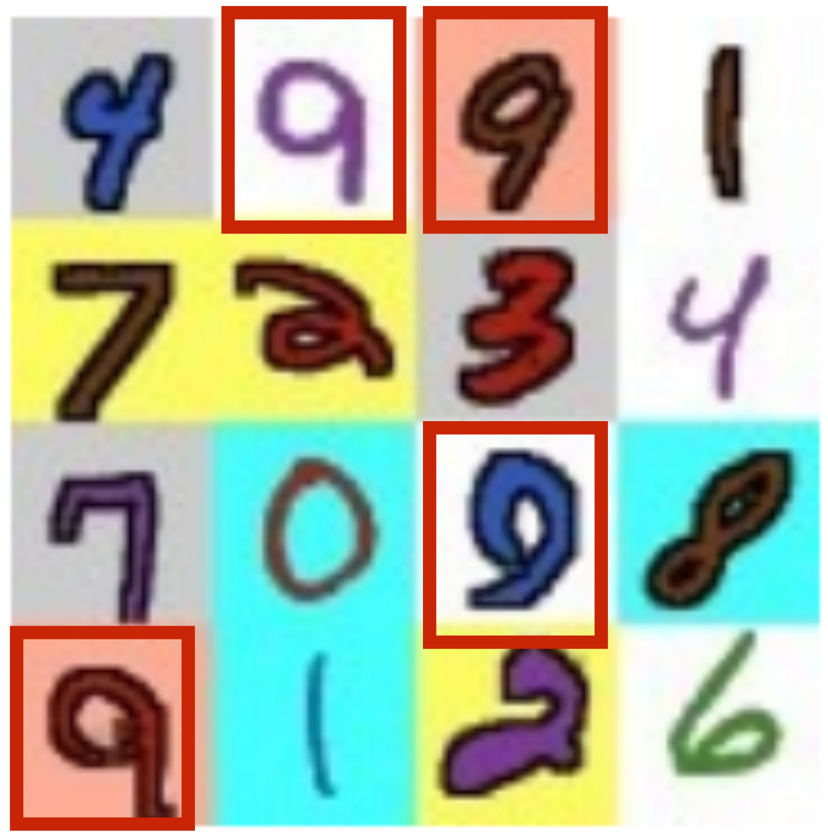
Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

#	Question	Answer
1	How many 9's are there in the image?	

Applications: Visual Dialogs

[Seo et al., NIPS 2017]

Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

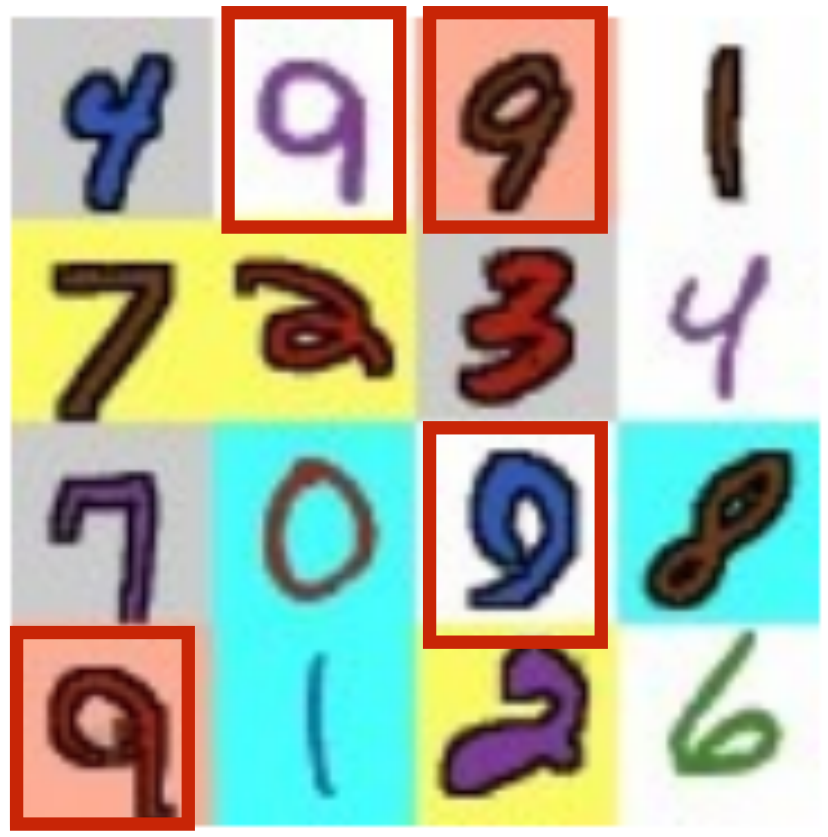


#	Question	Answer
1	How many 9's are there in the image?	-

Applications: Visual Dialogs

[Seo et al., NIPS 2017]

Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history



#	Question	Answer
1	How many 9's are there in the image?	four

Visual Dialog Task

[Seo et al., NIPS 2017]

Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	

Visual Dialog Task

[Seo et al., NIPS 2017]

Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	

Visual Dialog Task

[Seo et al., NIPS 2017]

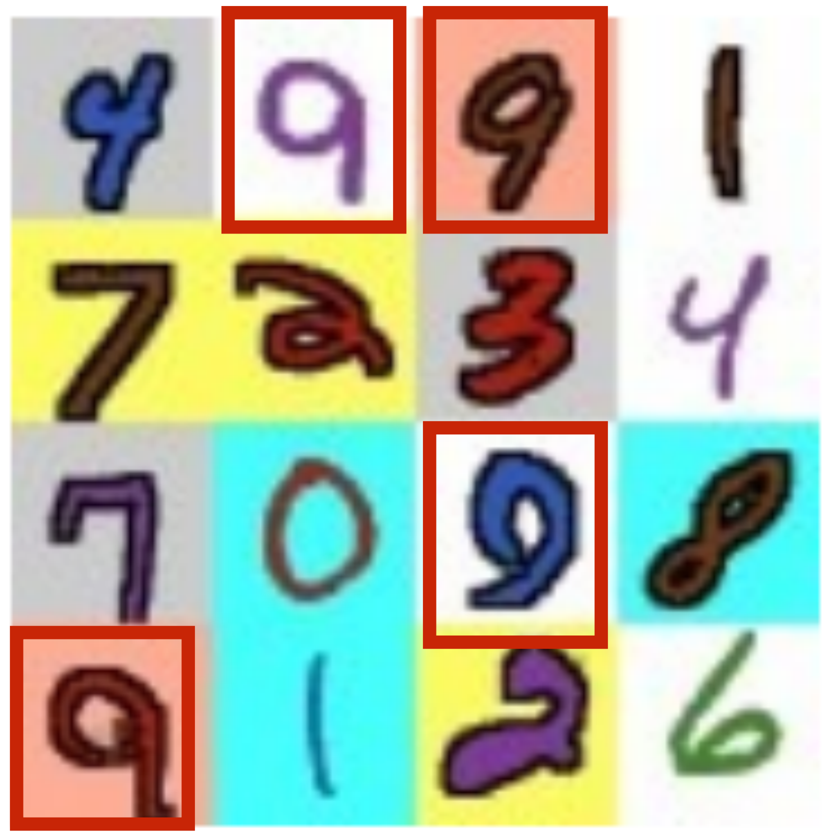
Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	

Visual Dialog Task

[Seo et al., NIPS 2017]

Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history



#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	

Visual Dialog Task

[Seo et al., NIPS 2017]

Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	one

Visual Dialog Task

[Seo et al., NIPS 2017]

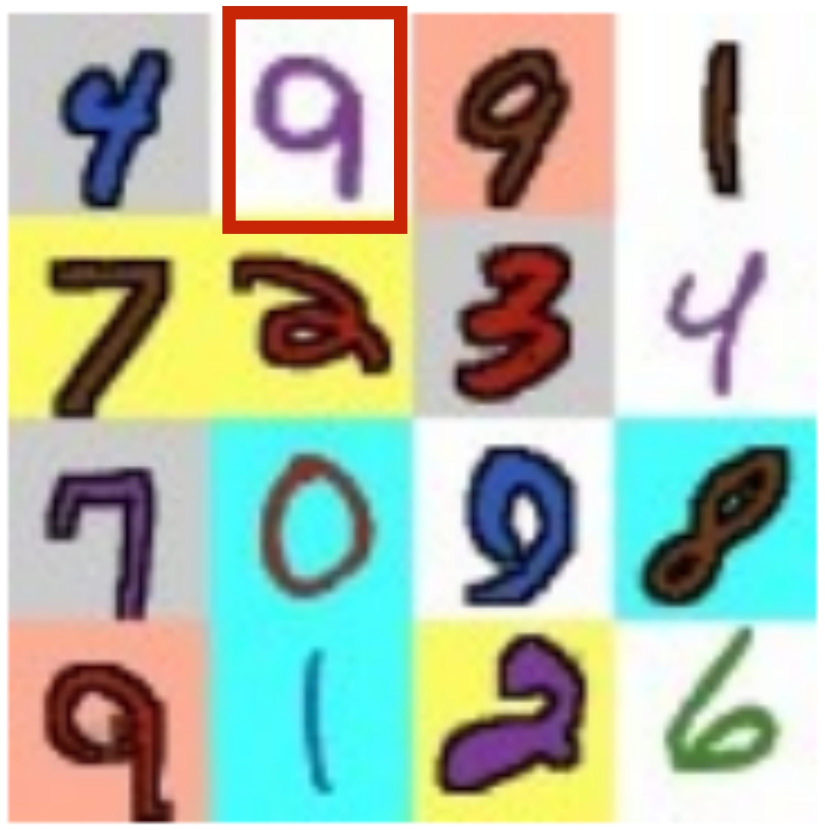
Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	one
3	What is the background color of the digit at the left of it?	white

Visual Dialog Task

[Seo et al., NIPS 2017]

Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

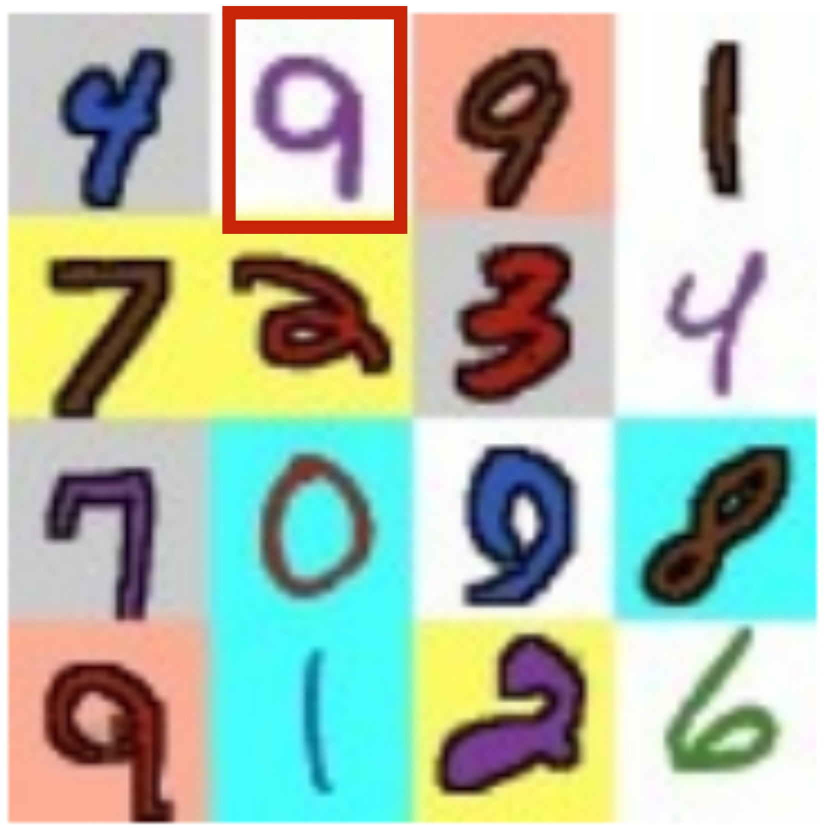


#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	one
3	What is the background color of the digit at the left of it?	white

Visual Dialog Task

[Seo et al., NIPS 2017]

Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history



#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	one
3	What is the background color of the digit at the left of it?	white
4	What is the style of the digit?	flat
5	What is the color of the digit at the left of it?	blue
6	What is the number of the blue digit?	4
7	Are there other blue digits?	two

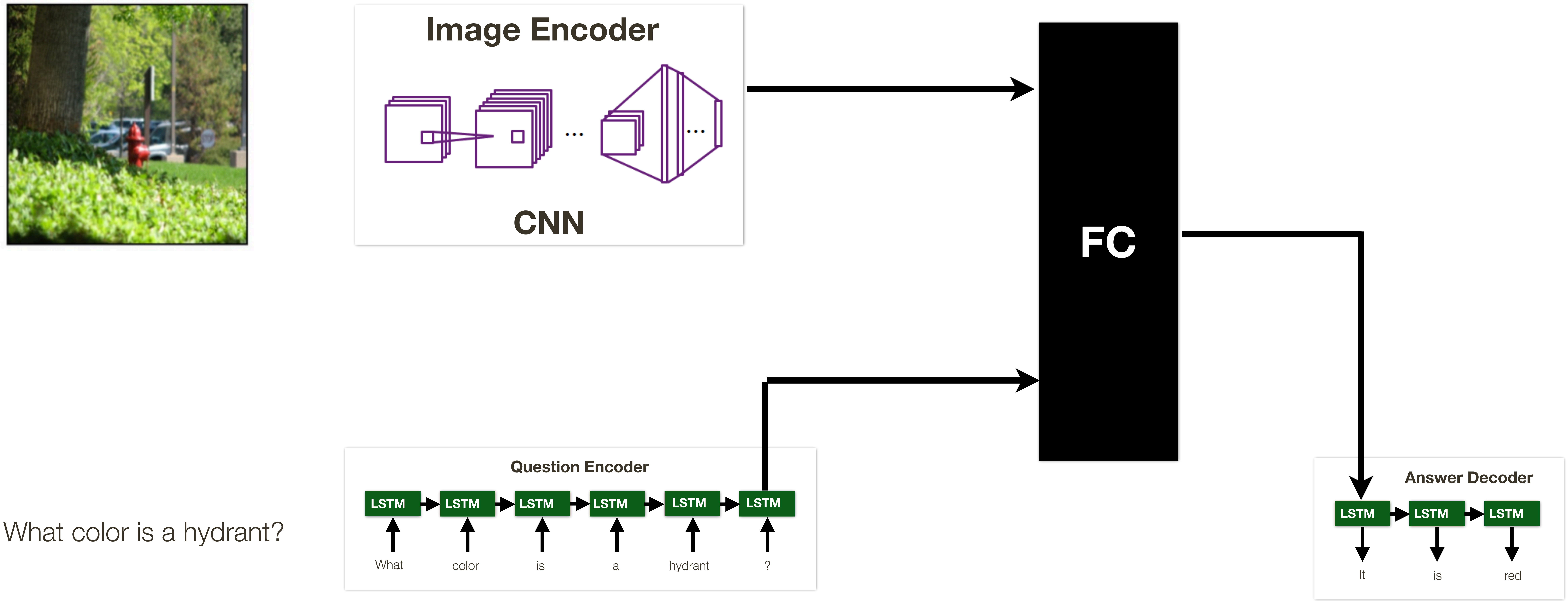
Simple Visual Question Answering

[Seo et al., NIPS 2017]

Q: What color is a hydrant?

Simple Visual Question Answering

[Seo et al., NIPS 2017]

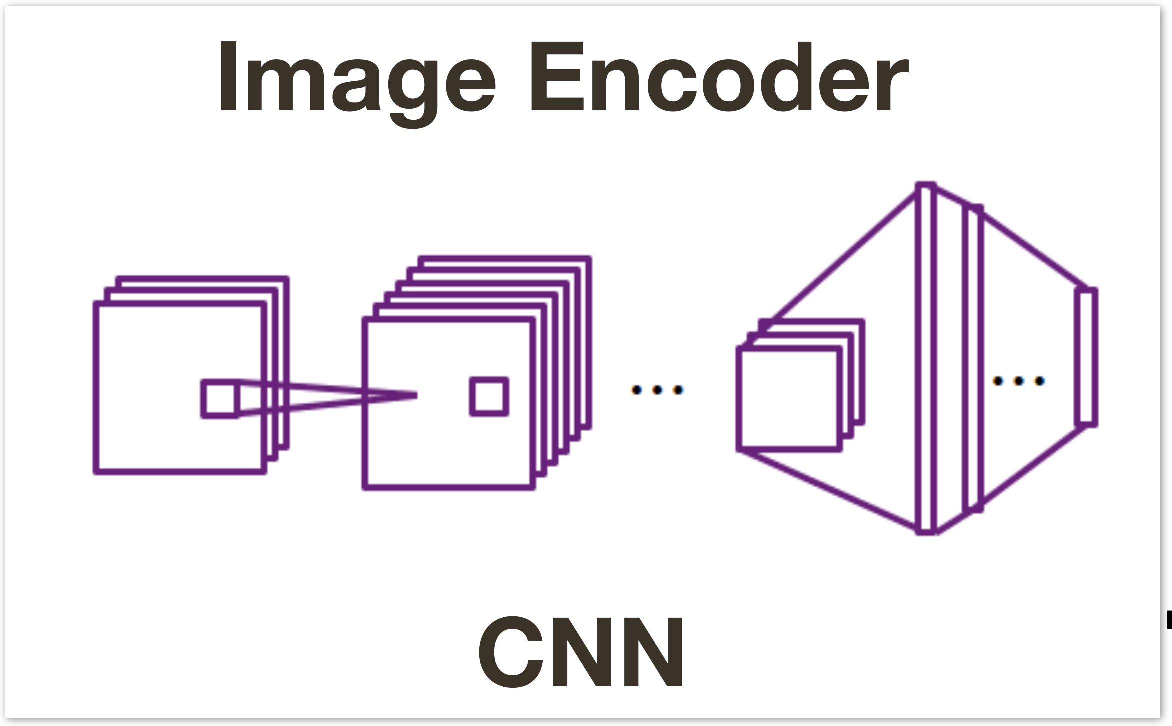


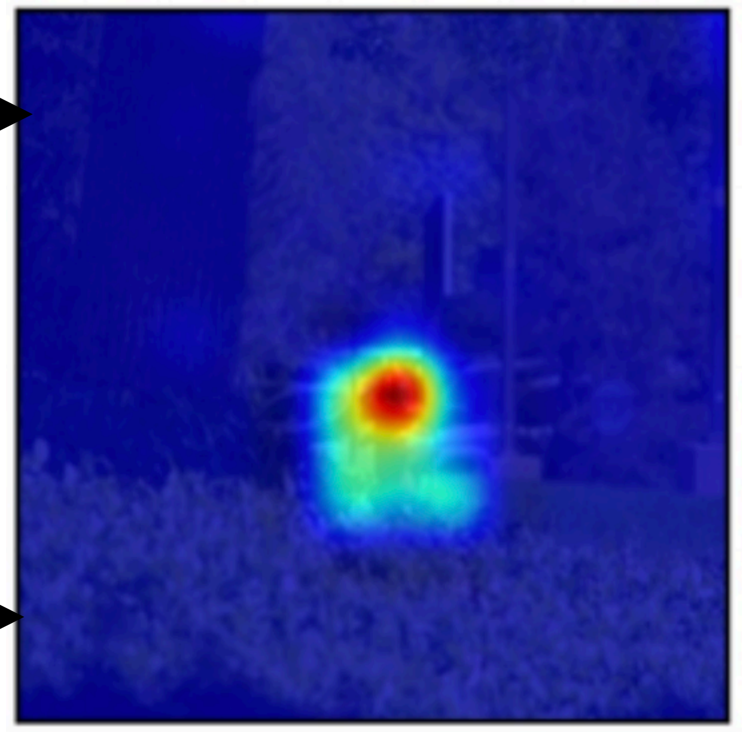
Q: What color is a hydrant?

A: It is red

Attention Networks for Visual Question Answering

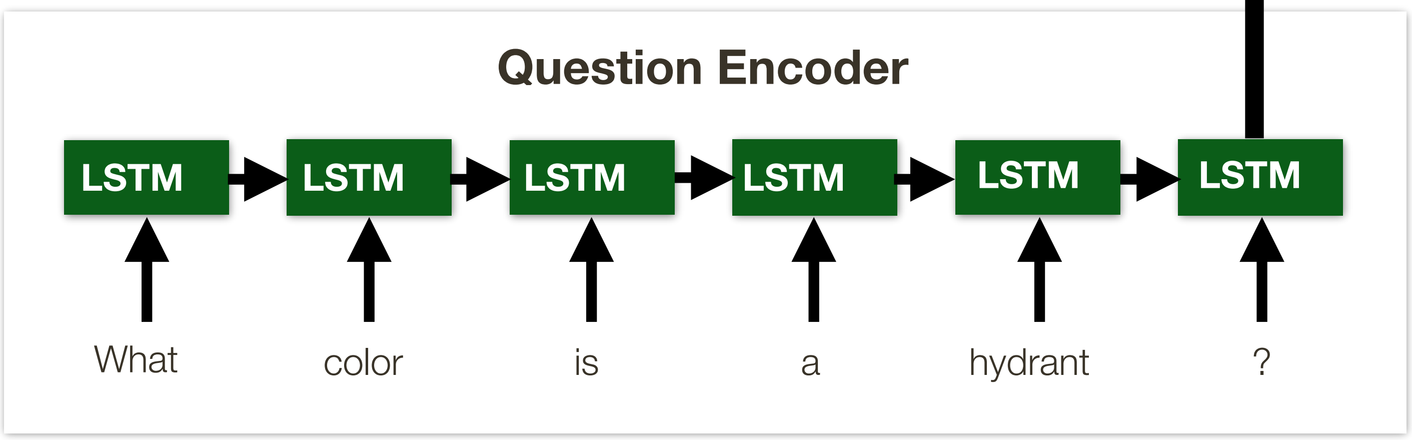
[Seo et al., NIPS 2017]





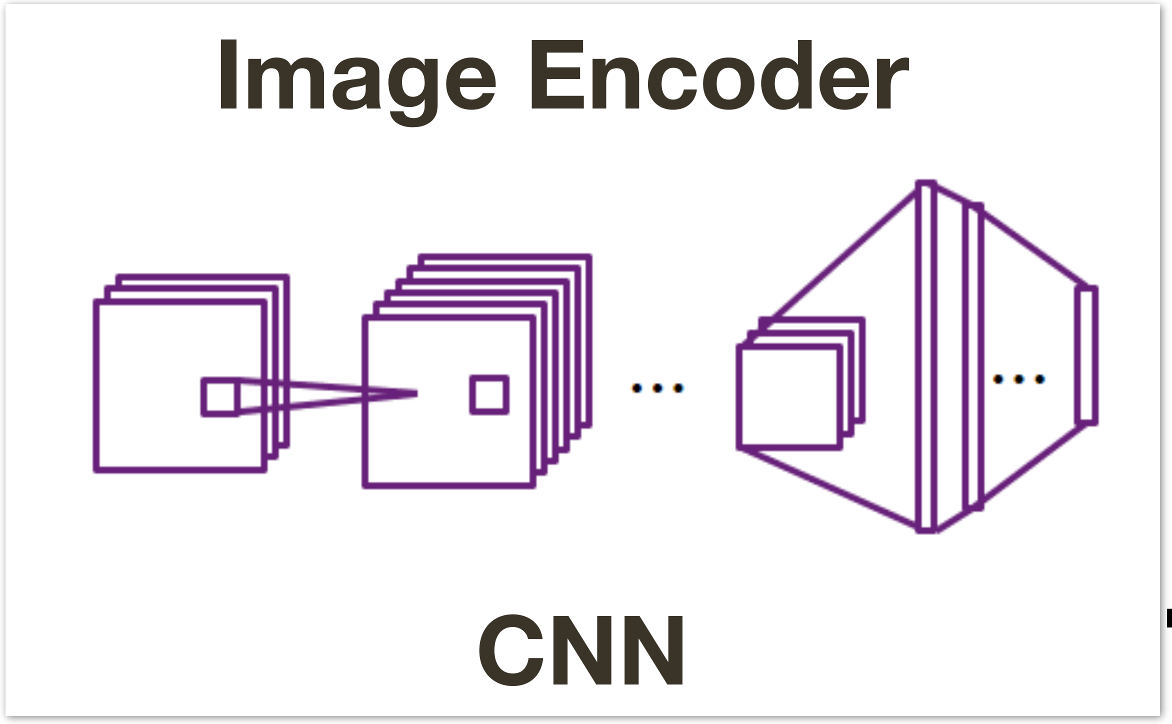
Tentative Attention

Q: What color is a hydrant?

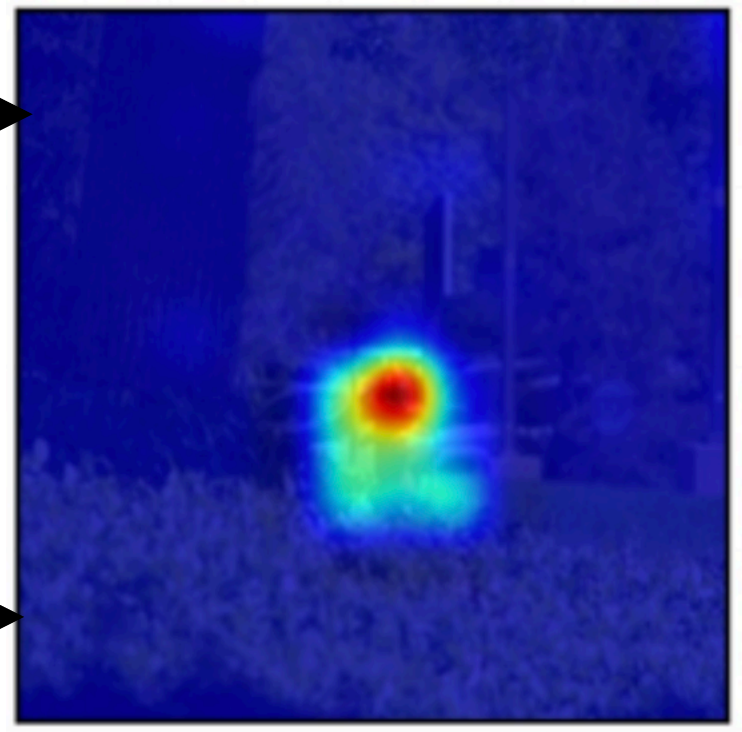


Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]

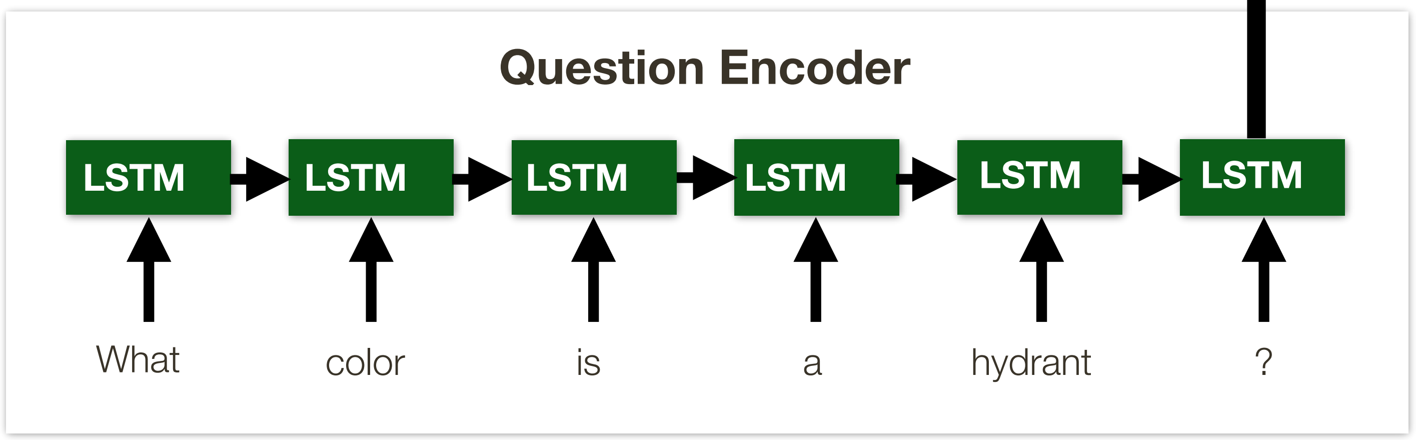


$M \, x \, M = n$ grid with values between 0 and 1, indicating which part of the image to pay attention to in order to answer the question



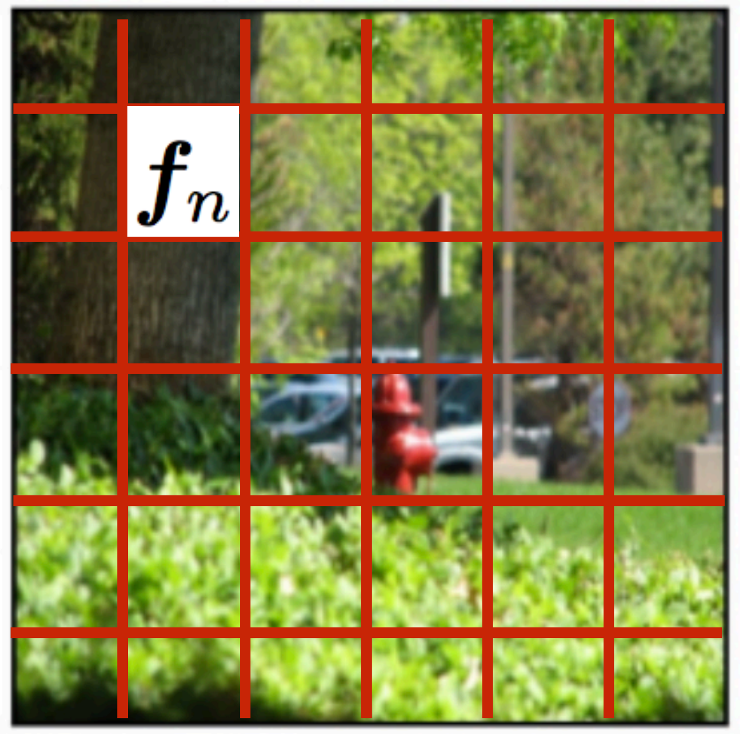
Tentative Attention

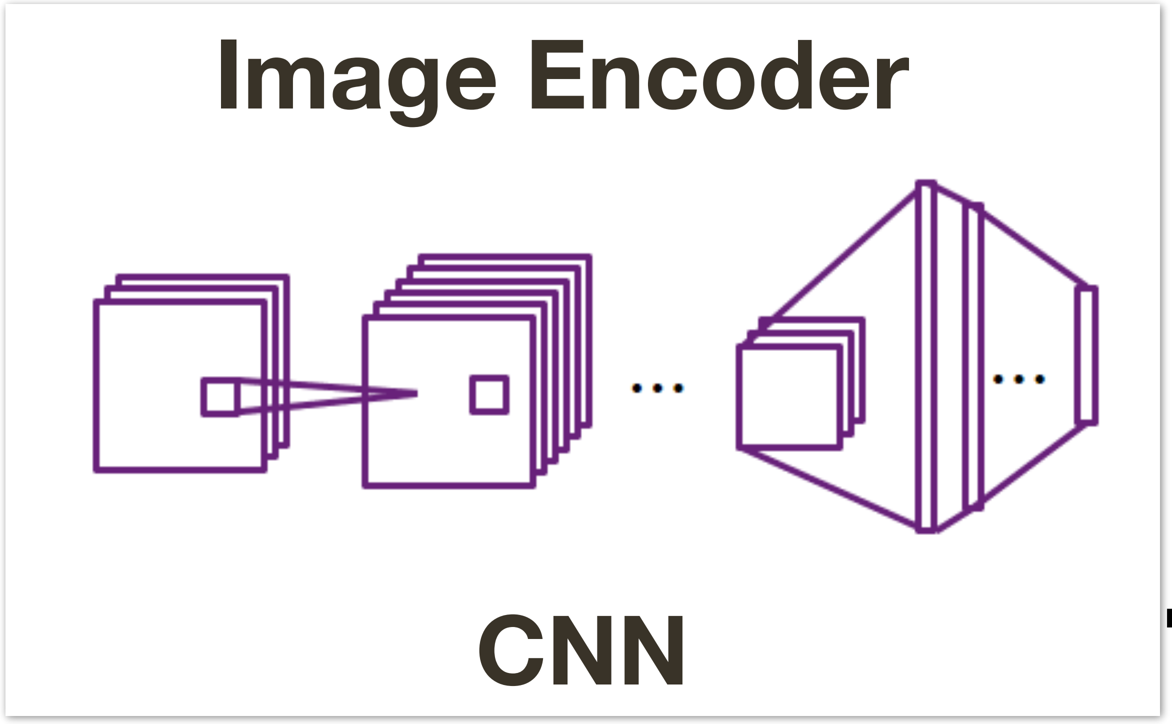
Q: What color is a hydrant?



Attention Networks for Visual Question Answering

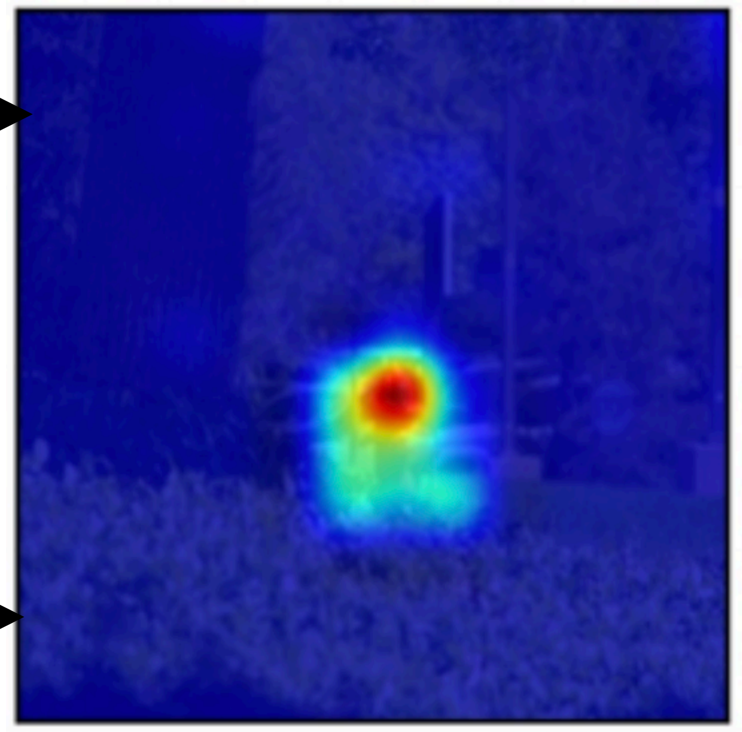
[Seo et al., NIPS 2017]





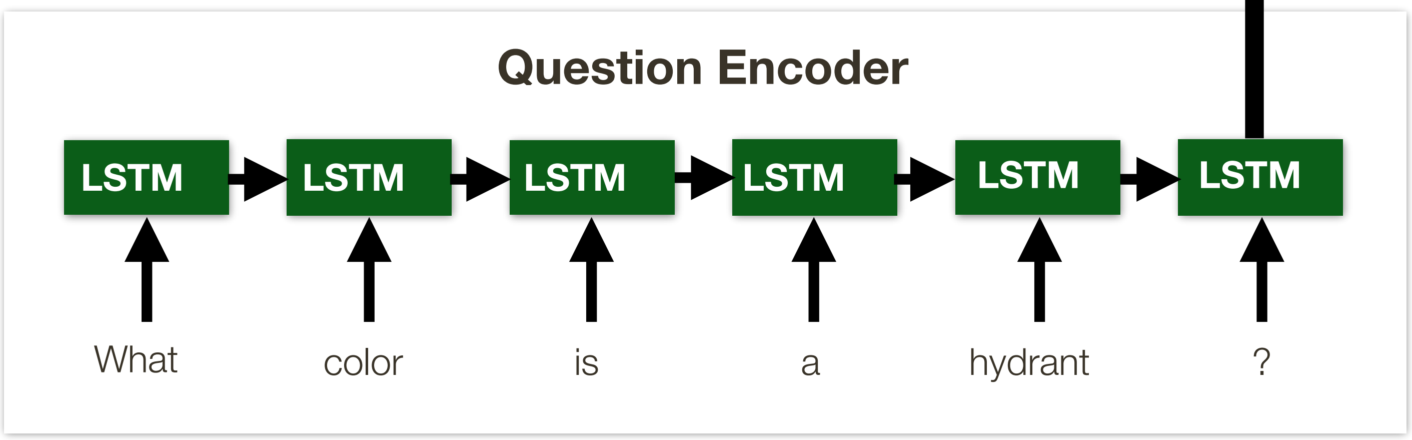
$M \, x \, M = n$ grid with values between 0 and 1, indicating which part of the image to pay attention to in order to answer the question

c_t



Tentative Attention

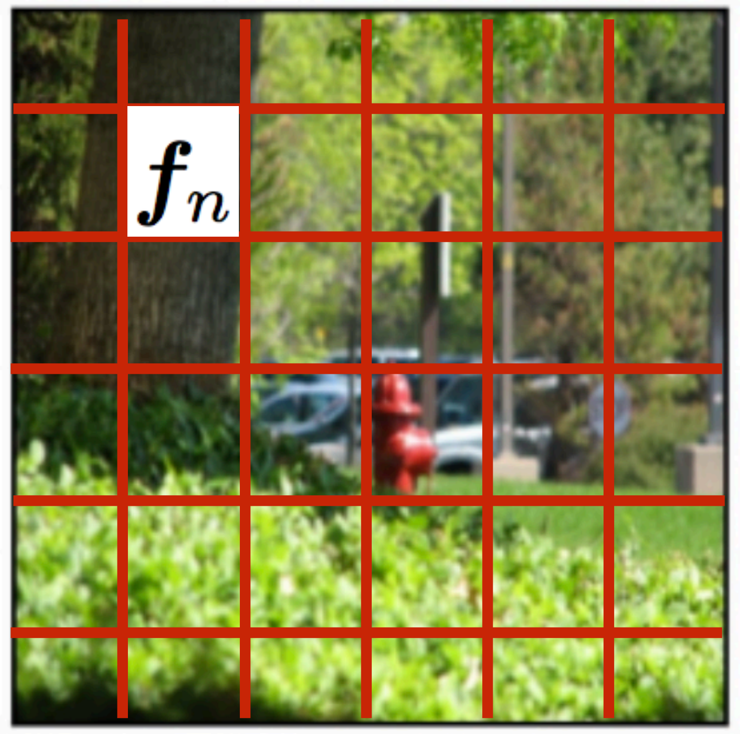
Q: What color is a hydrant?

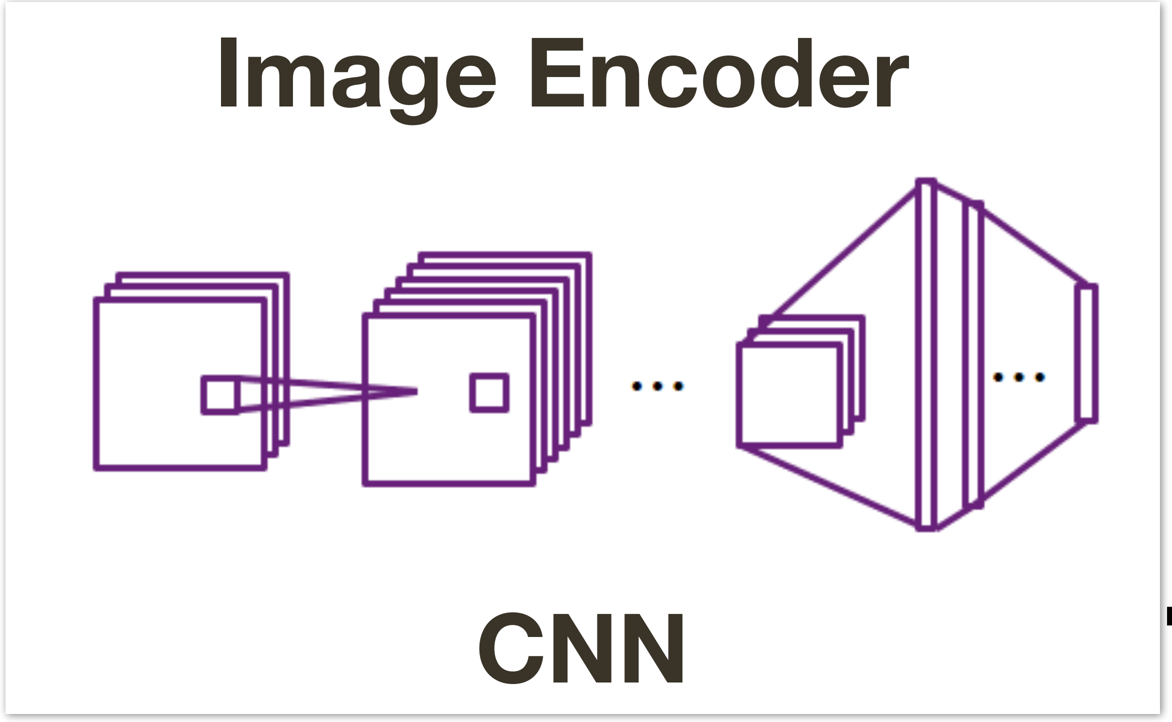


$$s_{t,n} = \left(\mathbf{W}_c^{\text{tent}} \boldsymbol{c}_t\right)^\top \left(\mathbf{W}_f^{\text{tent}} \boldsymbol{f}_n\right)$$

Attention Networks for Visual Question Answering

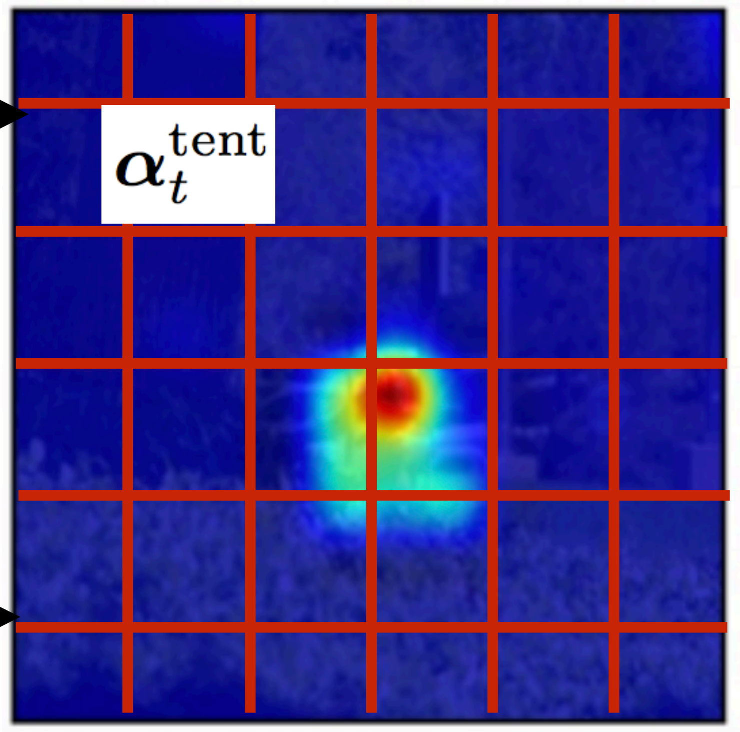
[Seo et al., NIPS 2017]





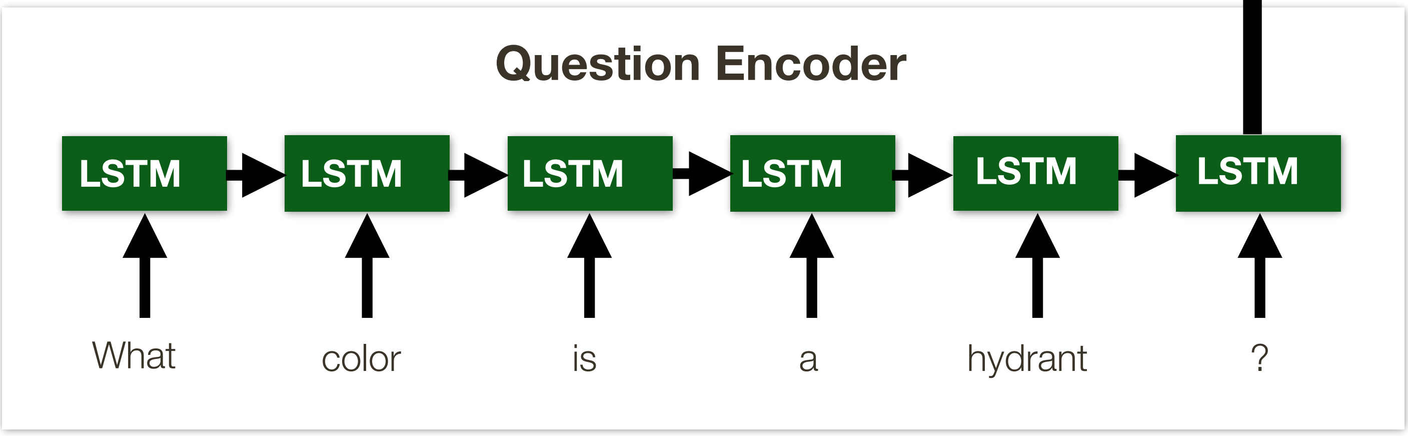
$M \, x \, M = n$ grid with values between 0 and 1, indicating which part of the image to pay attention to in order to answer the question

c_t



Tentative Attention

Q: What color is a hydrant?

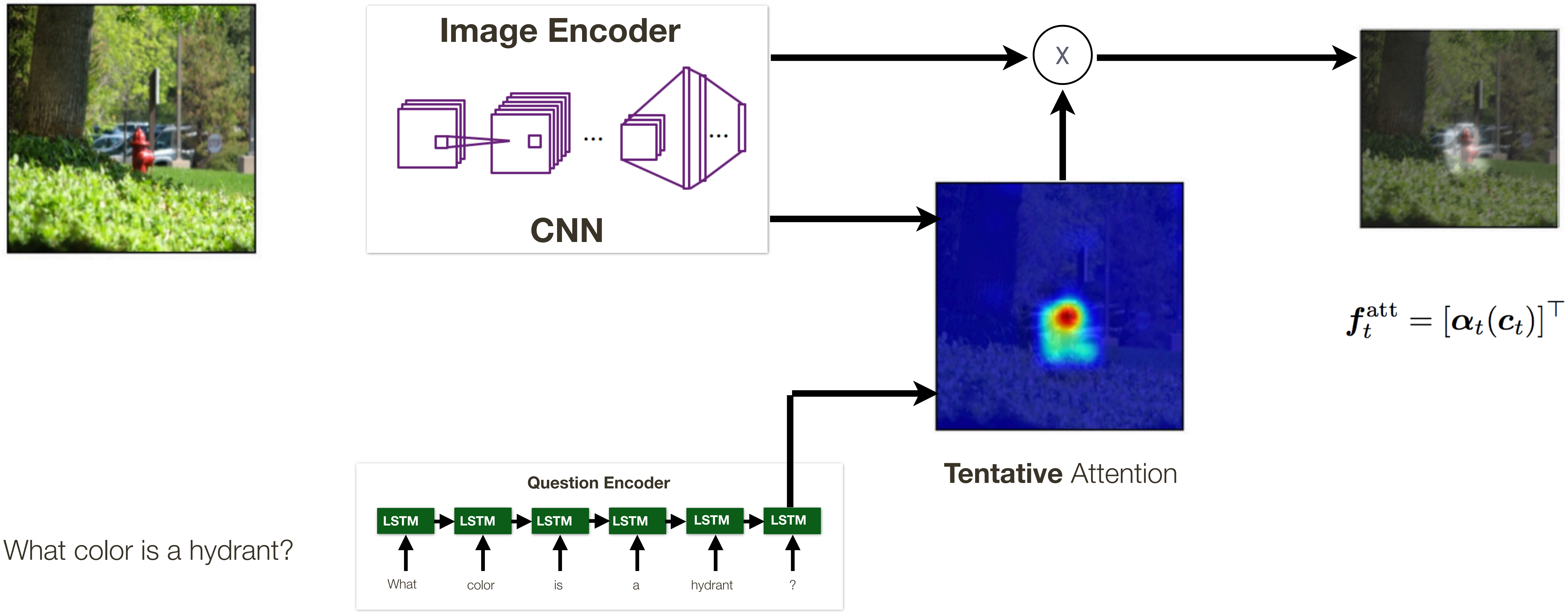


$$s_{t,n} = \left(\mathbf{W}_c^{\text{tent}} \boldsymbol{c}_t\right)^\top \left(\mathbf{W}_f^{\text{tent}} \boldsymbol{f}_n\right)$$

$$\boldsymbol{\alpha}_t^{\text{tent}} = \text{softmax}\left(\{s_{t,n}, 1 < n < N\}\right)$$

Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]



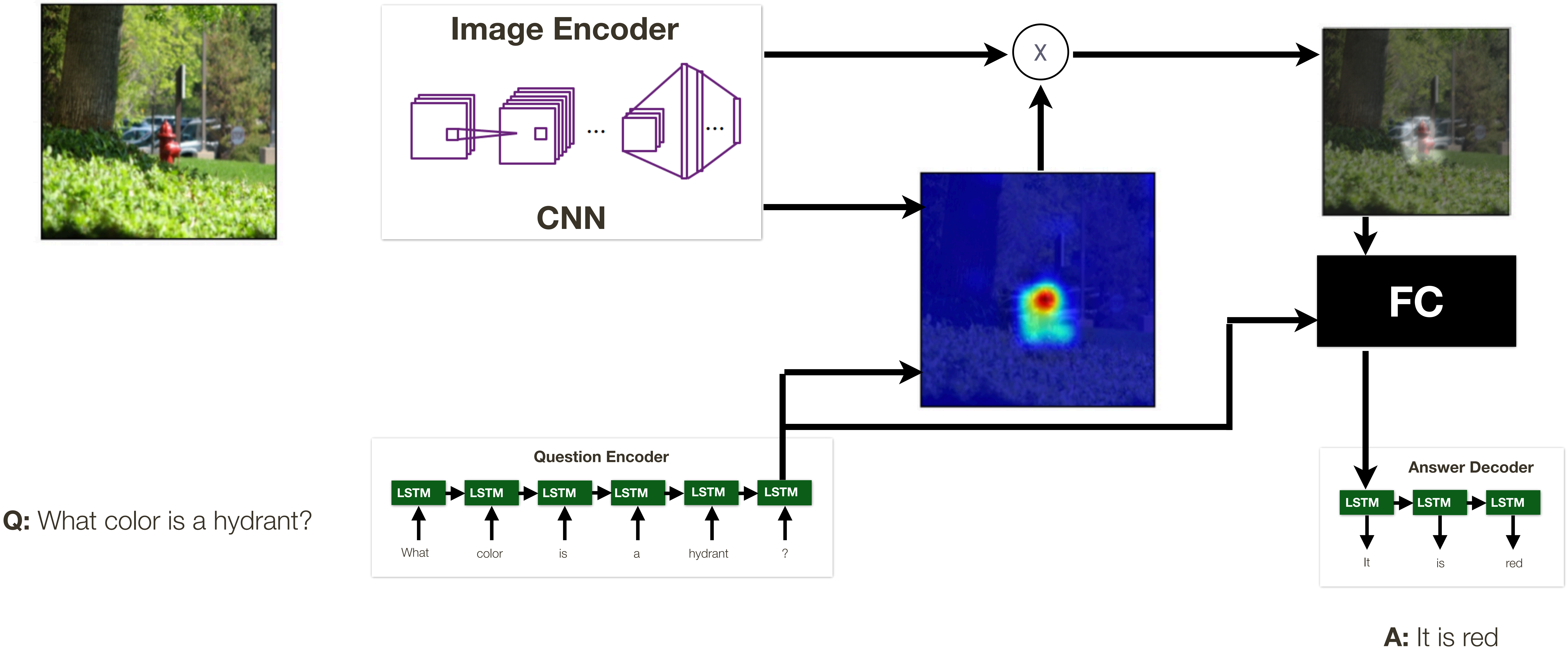
Tentative Attention

$$\boldsymbol{f}_t^{\text{att}} = [\boldsymbol{\alpha}_t(\boldsymbol{c}_t)]^\top \cdot \boldsymbol{f}$$

Q: What color is a hydrant?

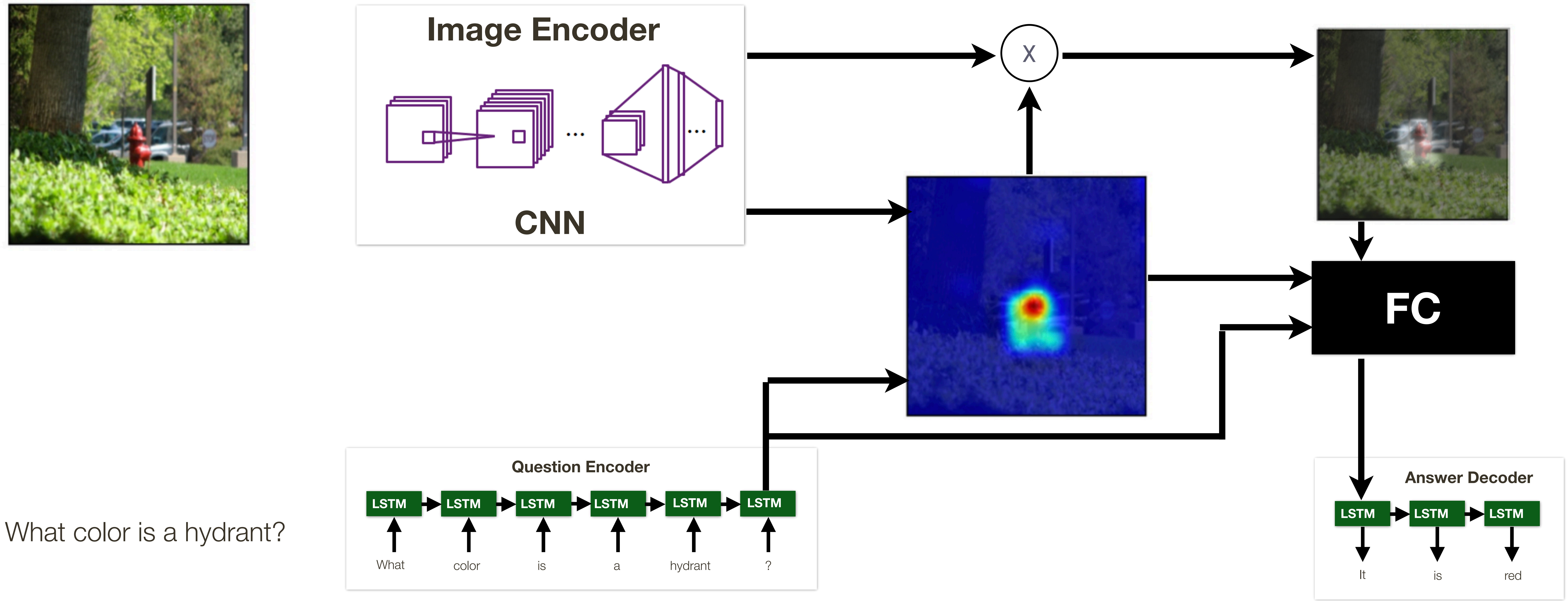
Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]



Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]

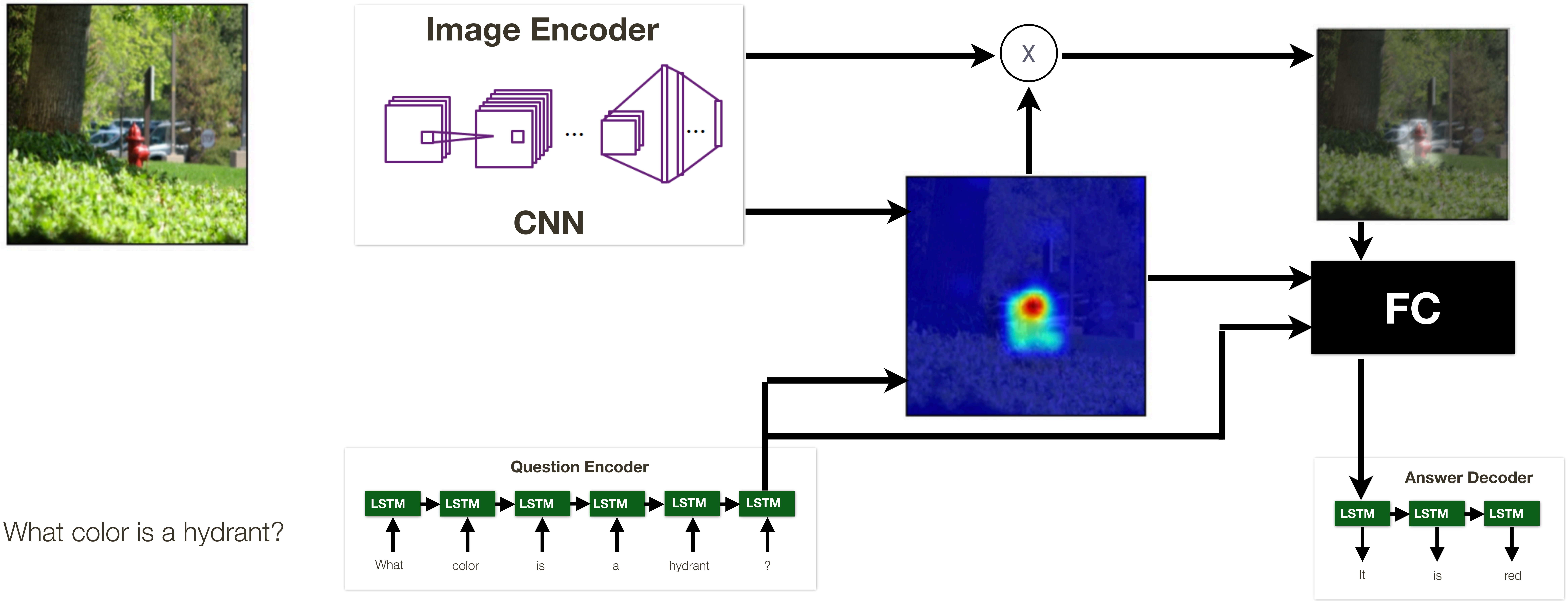


Q: What color is a hydrant?

A: It is red

Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]

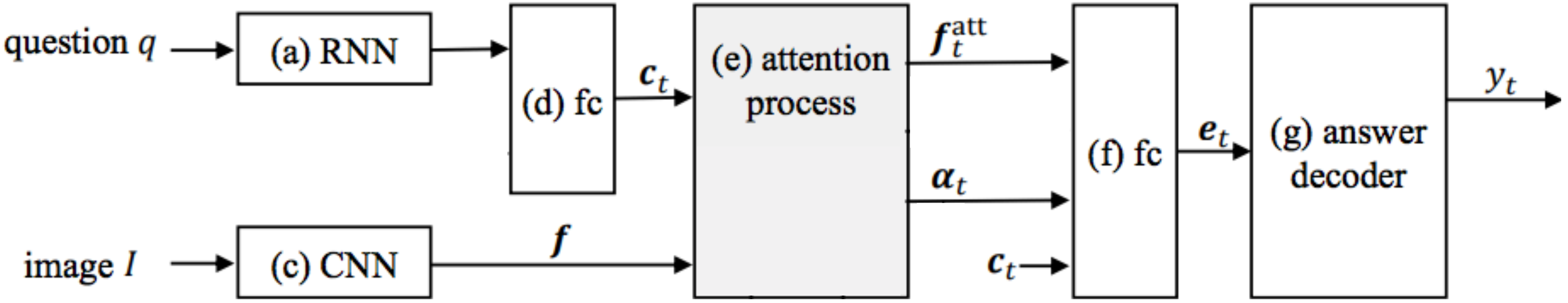


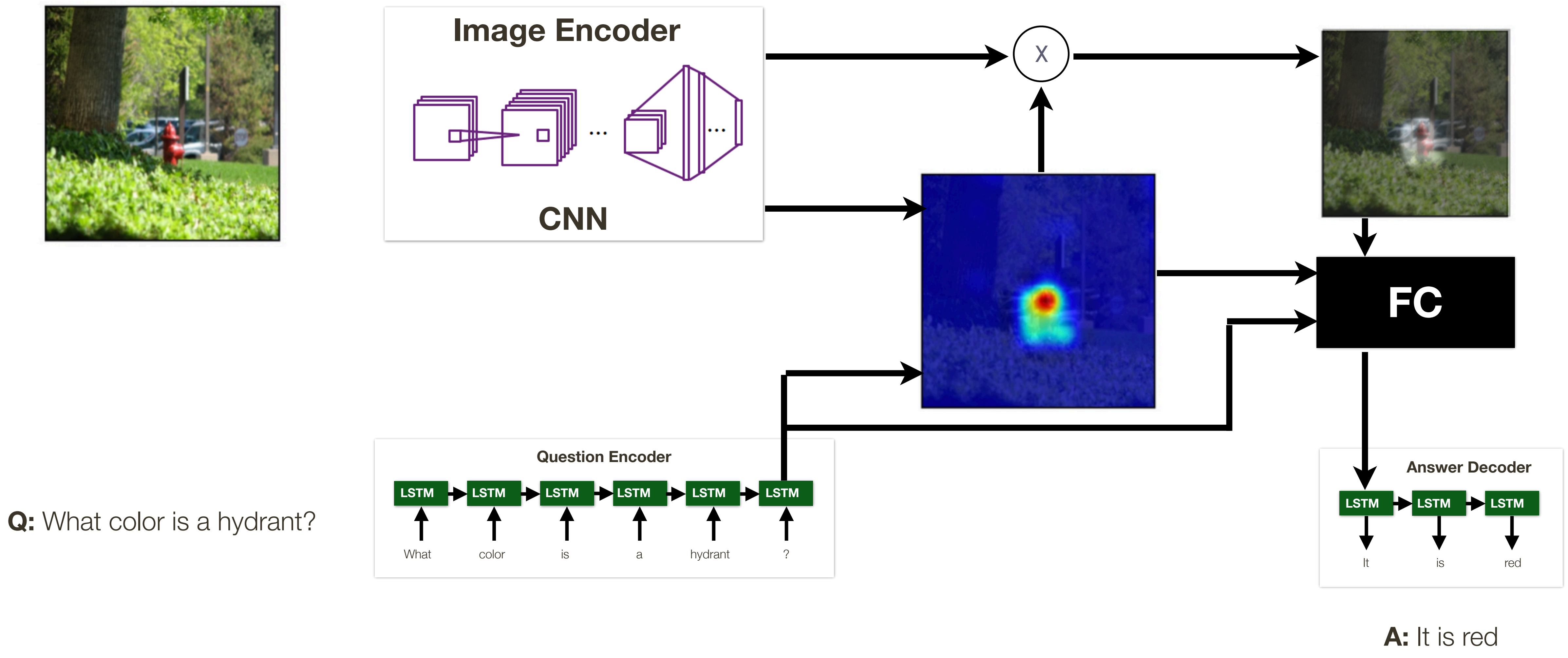
Q: What color is a hydrant?

A: It is red

Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]





Visual Dialog Task

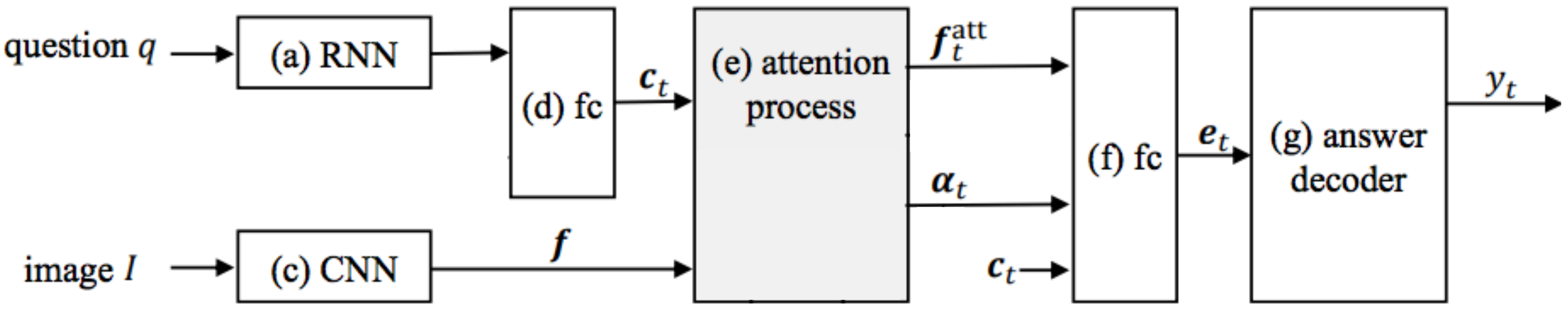
[Seo et al., NIPS 2017]

Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	one

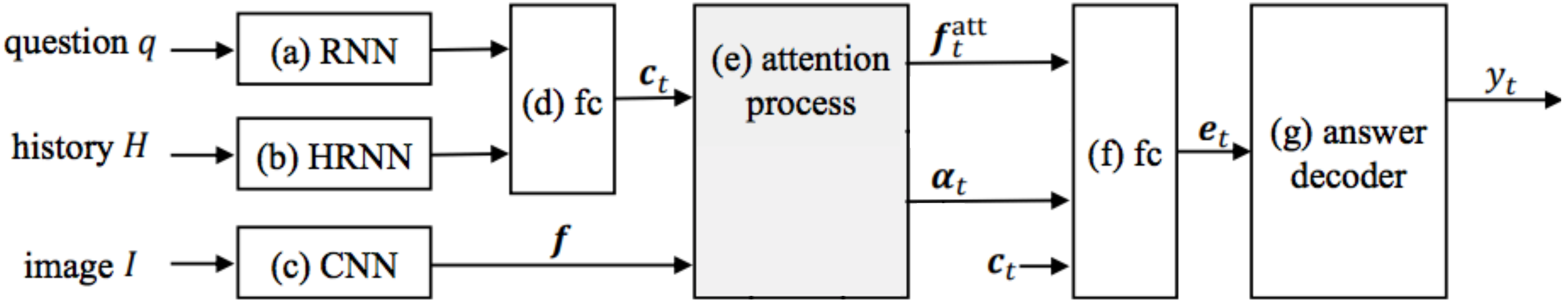
Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]



Attention Networks for Visual Dialogs

[Seo et al., NIPS 2017]



Hierarchical RNN (**HRNN**):

— Encode the question using LSTM

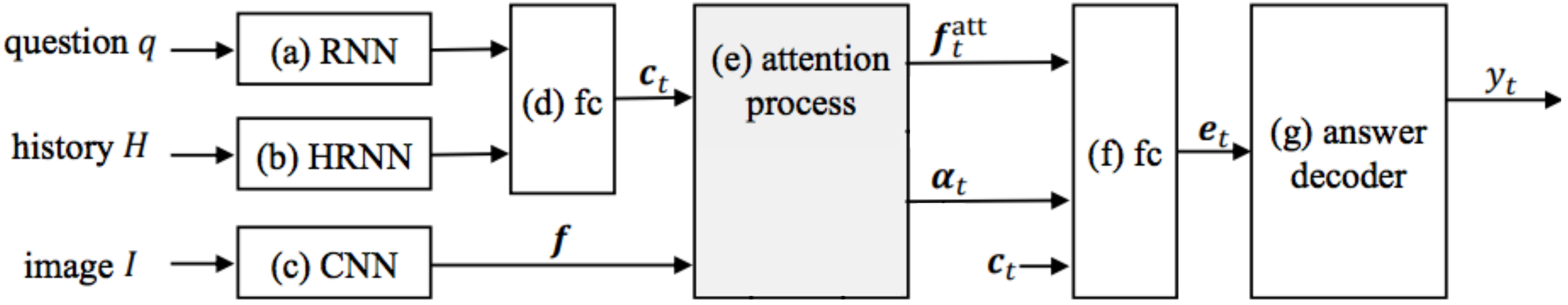
— Encode the answer using LSTM

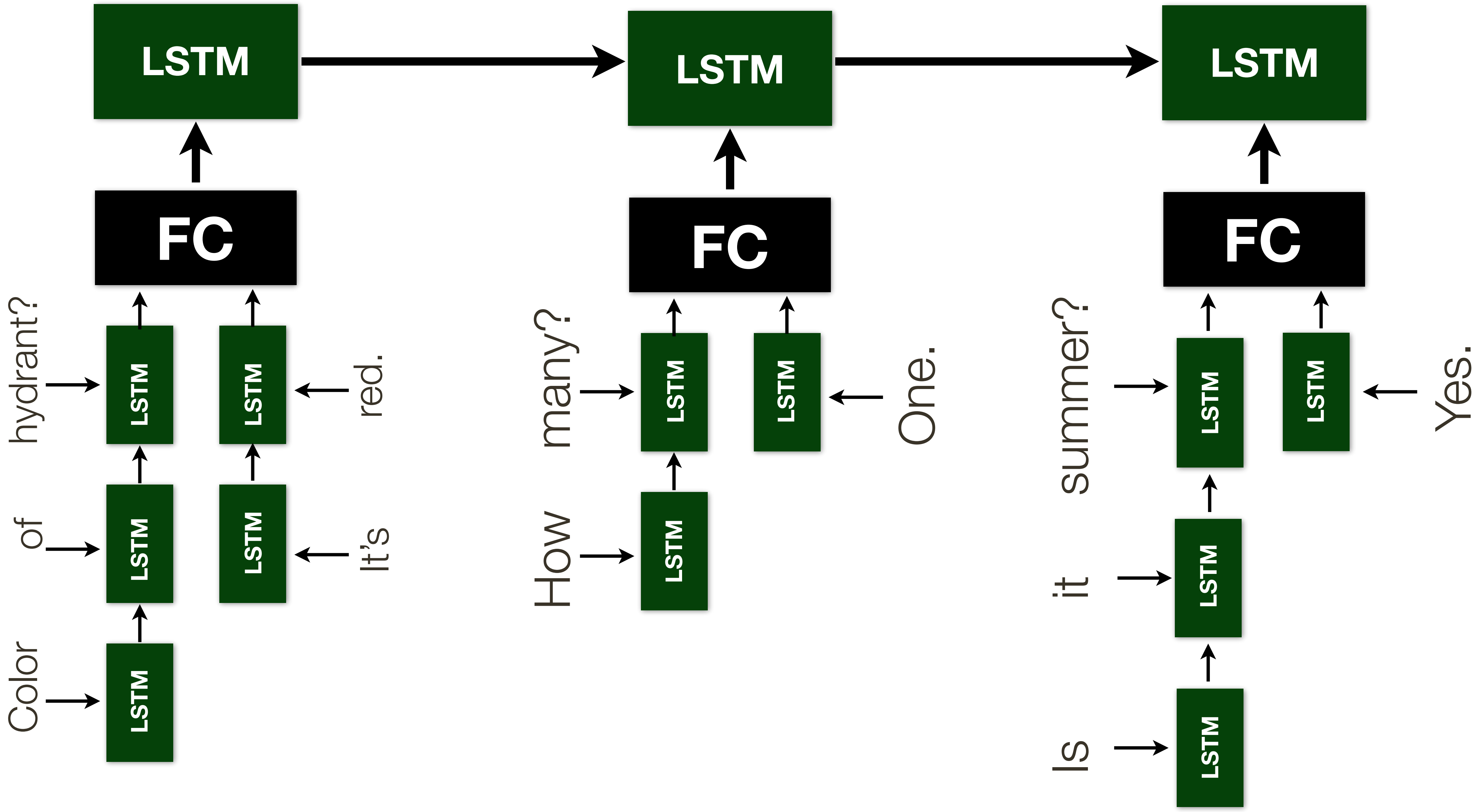
— Obtain QA embedding by fusing them using FC layer

— QA embeddings along the dialog are then encoded using higher-level LSTM

Attention Networks for Visual Dialogs

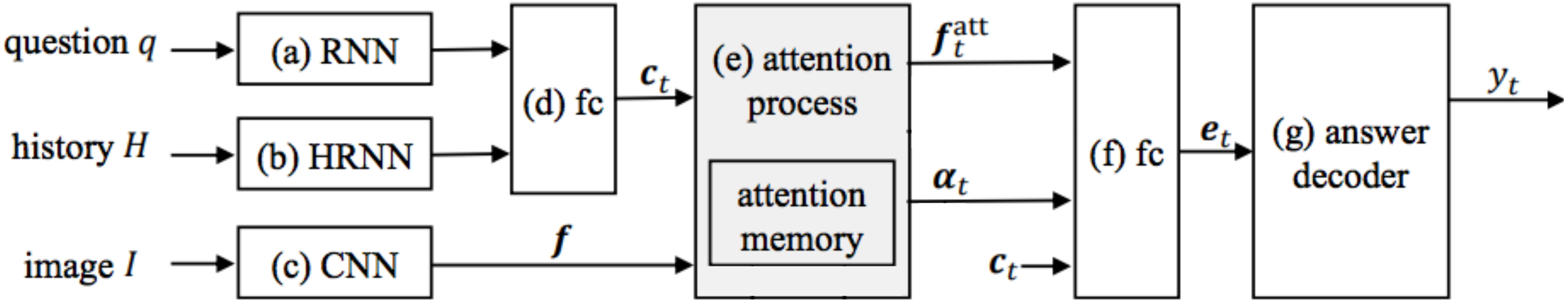
[Seo et al., NIPS 2017]



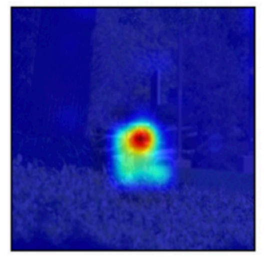
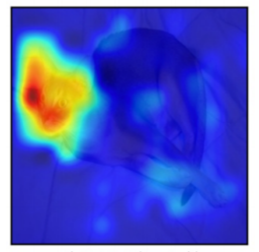


Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]



Associative Memory:

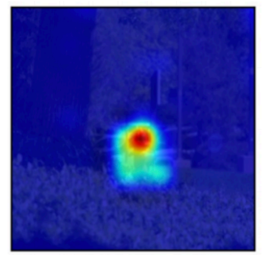
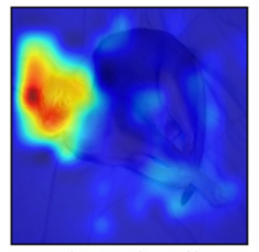
Question Turn	Key (hash)	Memory
1	f (**H:** Empty; **Q:** What color is a hydrant? **A:** It is red)	
2	f (**H:** …; **Q:** Is there a tree? **A:** Yes)	

Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]

Q3: What color is it?

Associative Memory:

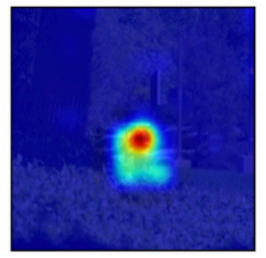
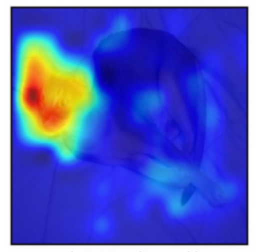
Question Turn	Key (hash)	Memory
1	f (**H:** Empty; **Q:** What color is a hydrant? **A:** It is red)	
2	f (**H:** …; **Q:** Is there a tree? **A:** Yes)	

Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]

Q3: What color is it?

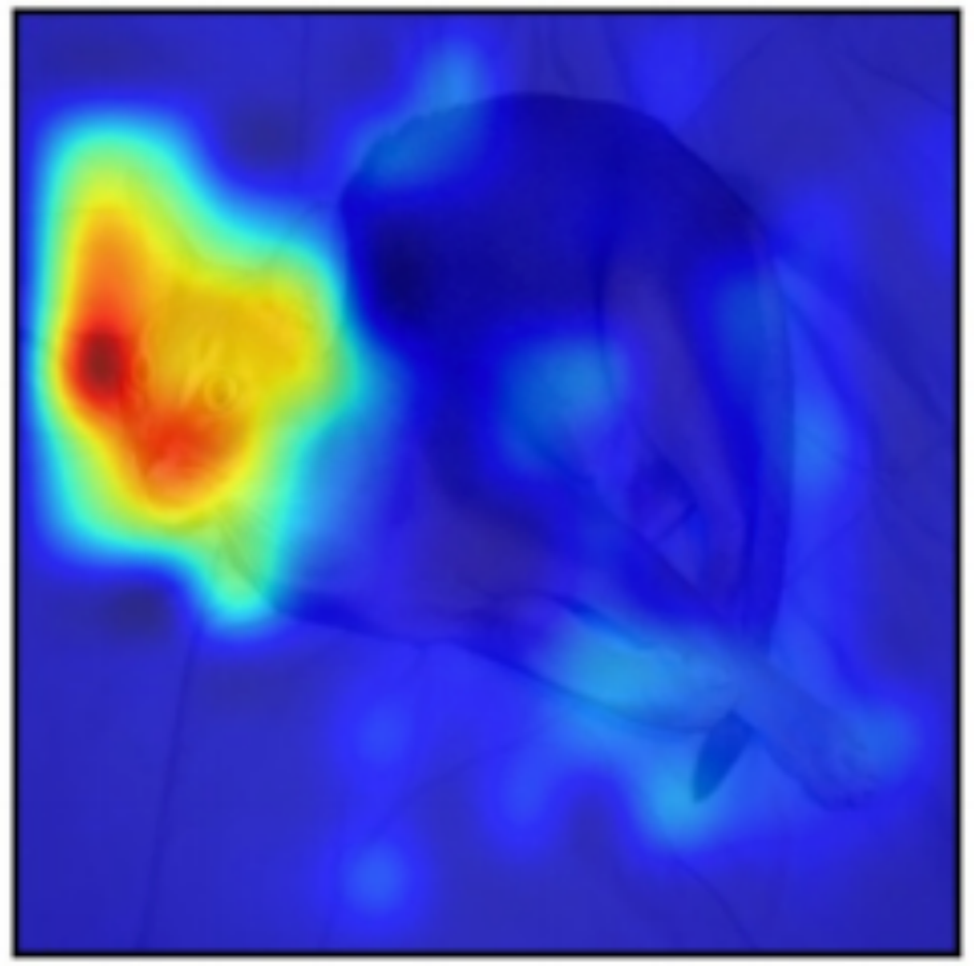
Associative Memory:

Question Turn	Key (hash)	Memory
1	f (**H:** Empty; **Q:** What color is a hydrant? **A:** It is red)	
2	f (**H:** …; **Q:** Is there a tree? **A:** Yes)	

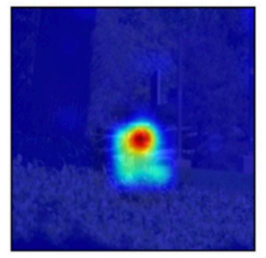
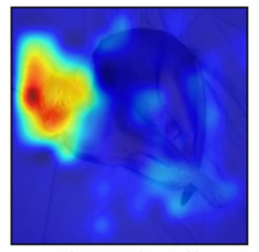
Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]

Q3: What color is it?



Associative Memory:

Question Turn	Key (hash)	Memory
1	f (**H:** Empty; **Q:** What color is a hydrant? **A:** It is red)	
2	f (**H:** …; **Q:** Is there a tree? **A:** Yes)	

Associative Memory Attention

[Seo et al., NIPS 2017]

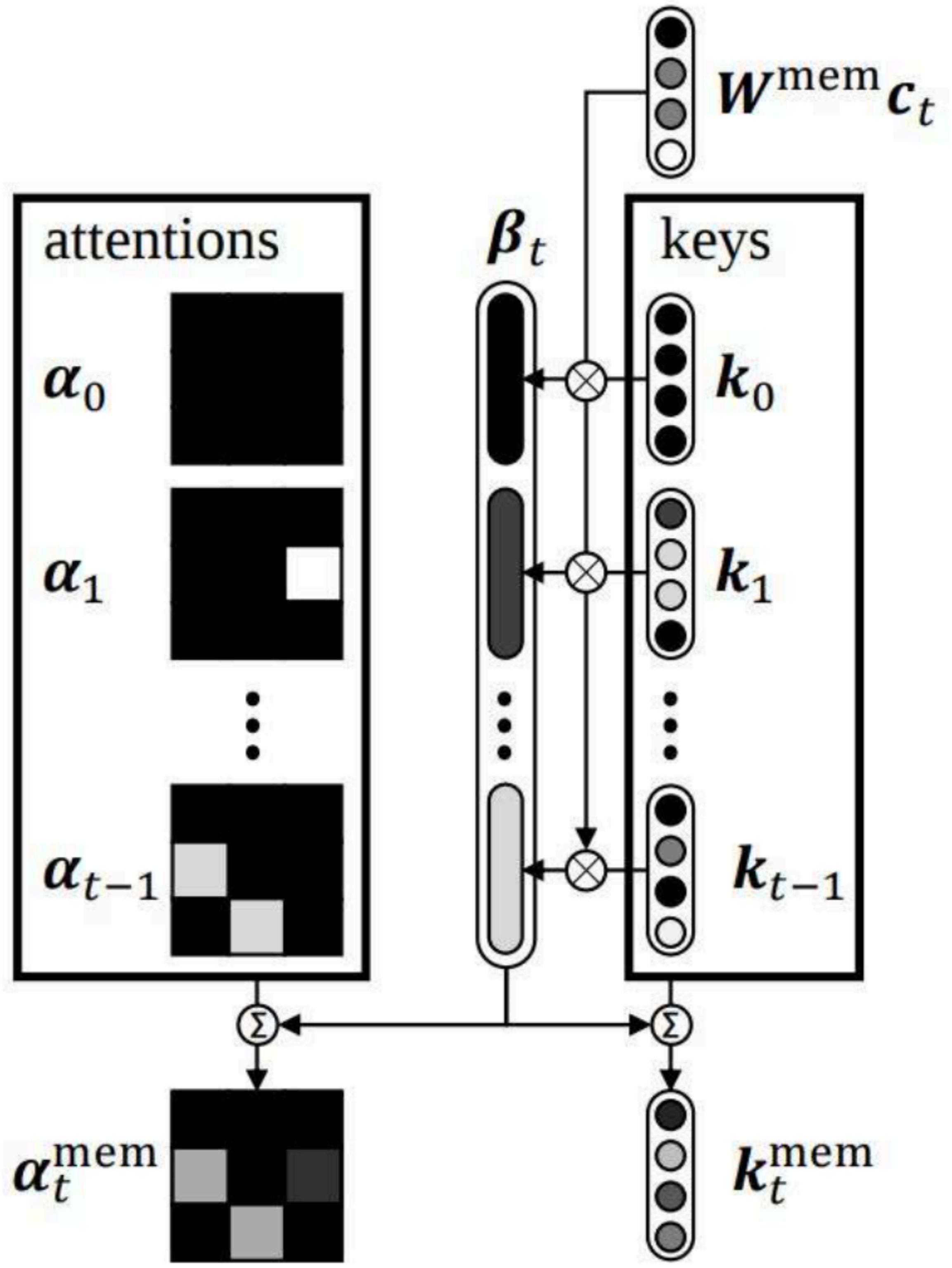
Key Idea: Every item in memory is **`(attention, key)`** pair — explicitly storing attentions used to answer previous questions

Associative Memory Attention

[Seo et al., NIPS 2017]

Key Idea: Every item in memory is **`(attention, key)`** pair — explicitly storing attentions used to answer previous questions

Intuition: How similar is the current turn's context to each of the previous response scenarios?

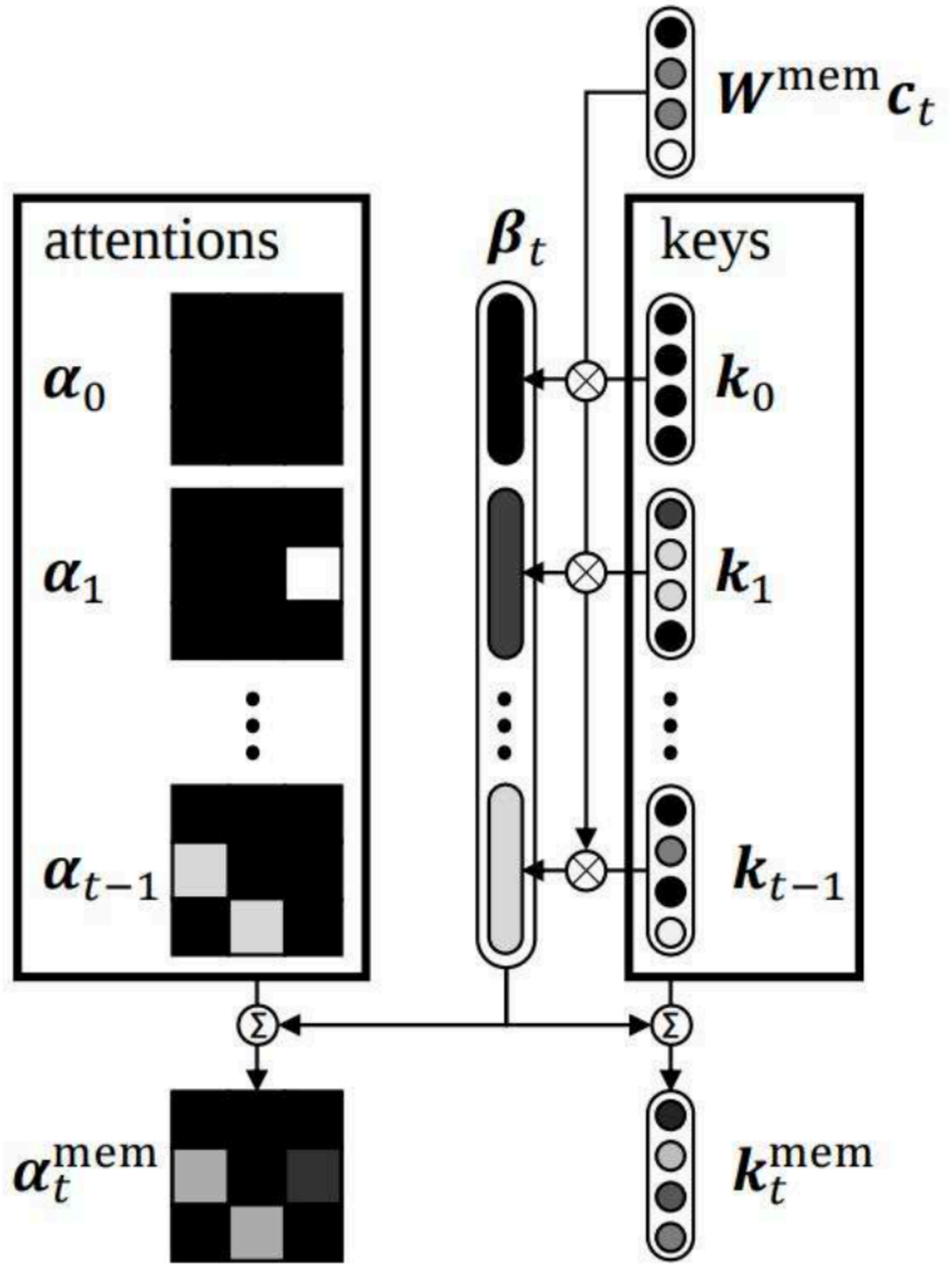


Associative Memory Attention

[Seo et al., NIPS 2017]

Key Idea: Every item in memory is **`(attention, key)`** pair — explicitly storing attentions used to answer previous questions

Intuition: How similar is the current turn's context to each of the previous response scenarios?



Observation: This formulation gives all previous turns equal weight (uniform prior)

Associative Memory Attention

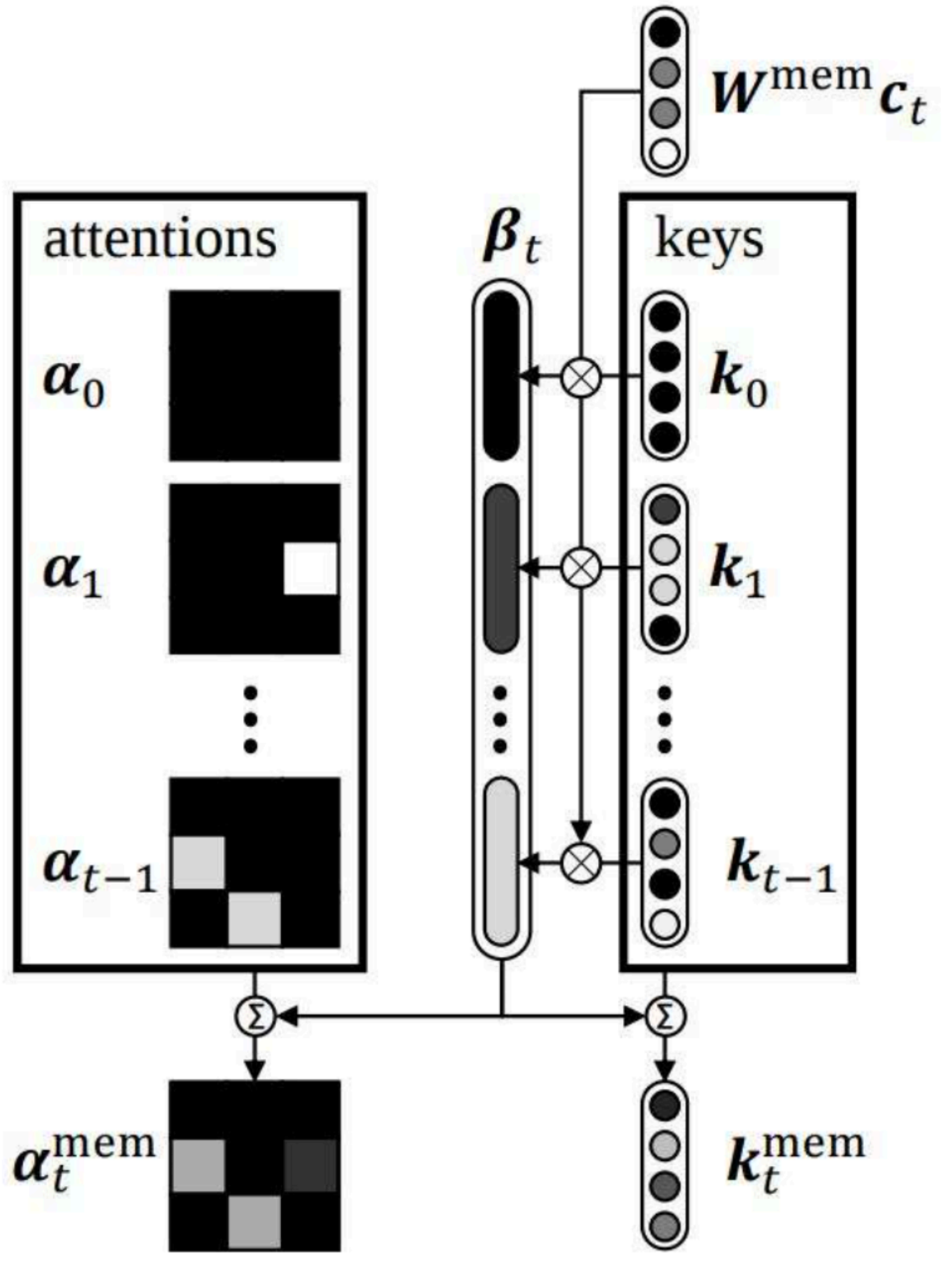
[Seo et al., NIPS 2017]

Key Idea: Every item in memory is **`(attention, key)`** pair — explicitly storing attentions used to answer previous questions

Intuition: How similar is the current turn's context to each of the previous response scenarios?

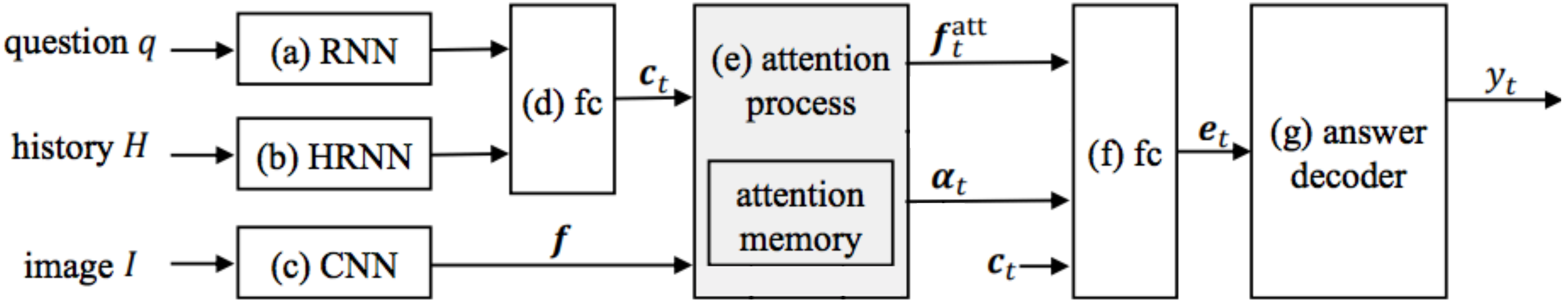
Observation: This formulation gives all previous turns equal weight (uniform prior)

Intuition: More recent questions are likely more relevant



Dynamic Attention Combination

[Seo et al., NIPS 2017]



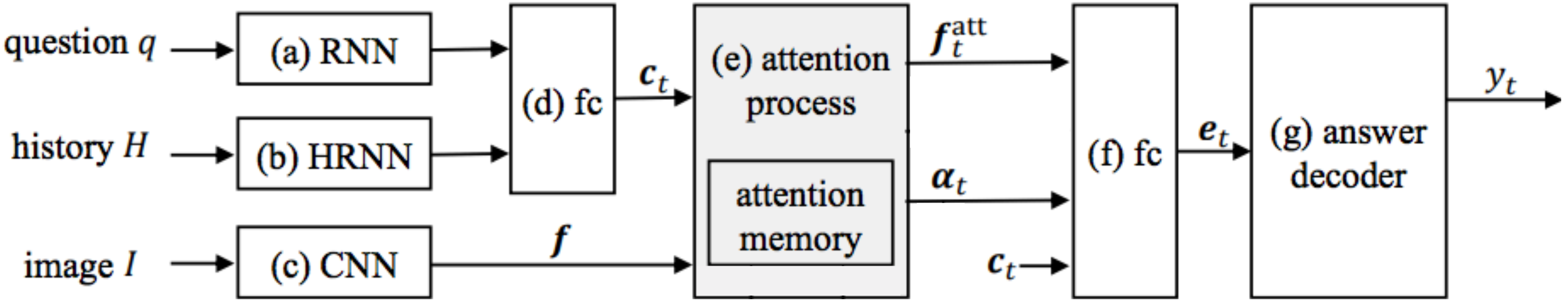
Two types of attention that focus on distinctly different aspect:

— **Tentative** Attention: What do we need to focus on given the current question

— **Associative Memory** Attention: What regions (attentions) used by previous turns are useful for the current question (a.k.a. visual reference resolution)

Dynamic Attention Combination

[Seo et al., NIPS 2017]



Two types of attention that focus on distinctly different aspect:

— **Tentative** Attention: What do we need to focus on given the current question

— **Associative Memory** Attention: What regions (attentions) used by previous turns are useful for the current question (a.k.a. visual reference resolution)

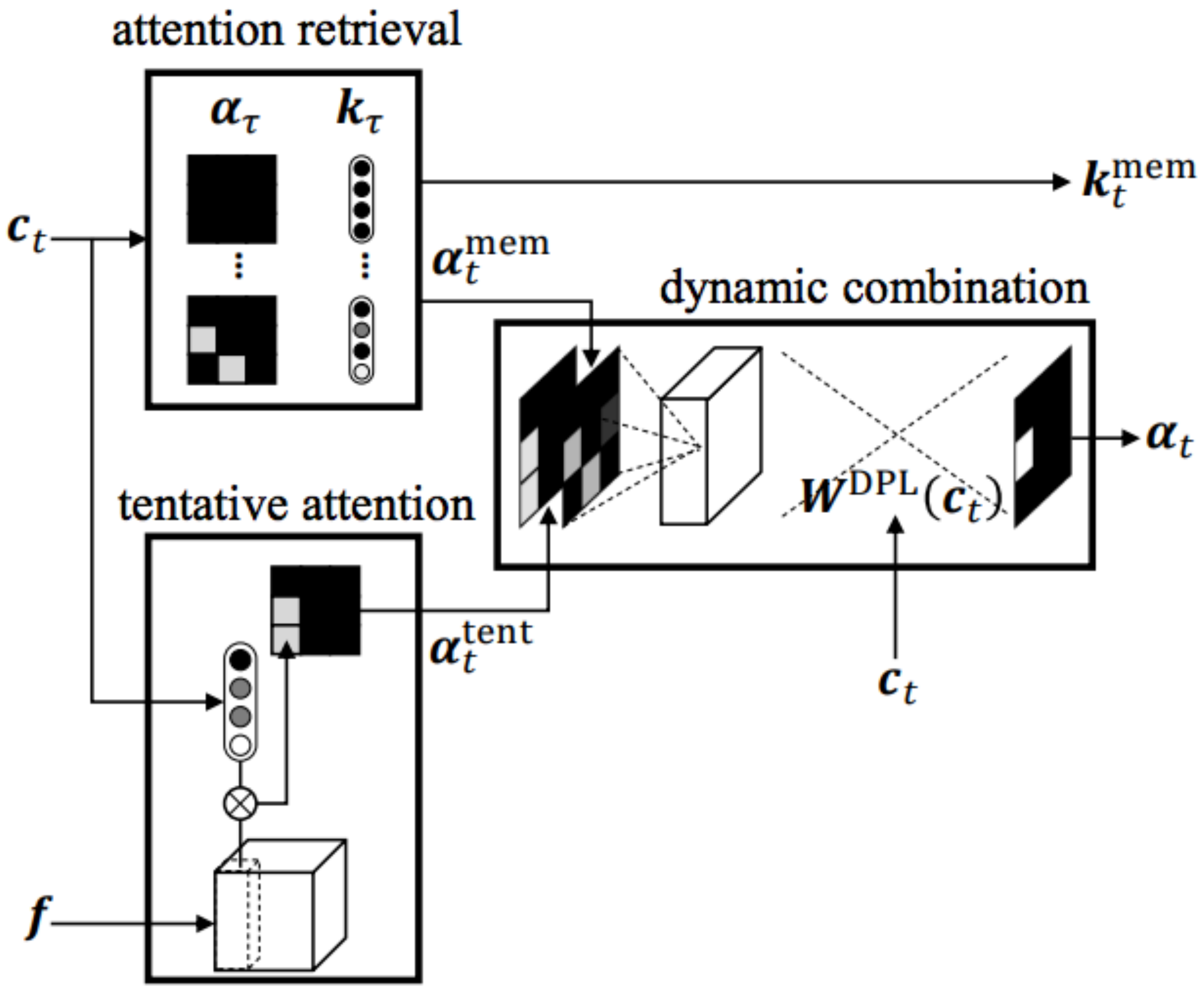
Intuition: We need a dynamic mechanism to fuse these attention models

[Noh et al., CVPR 2016]

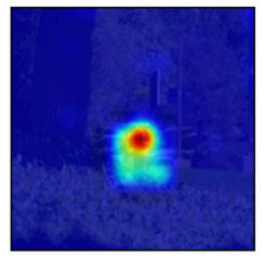
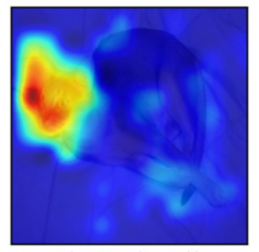
Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]

Q3: What color is it?

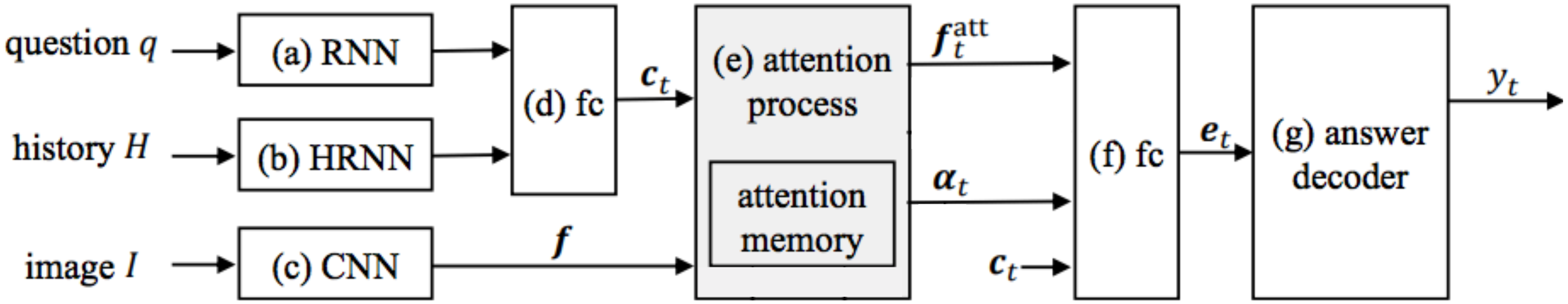


Associative Memory:

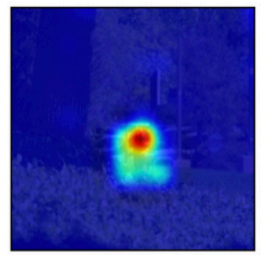
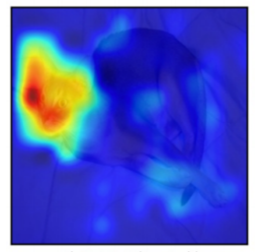
Question Turn	Key (hash)	Memory
1	f (**H:** Empty; **Q:** What color is a hydrant? **A:** It is red)	
2	f (**H:** …; **Q:** Is there a tree? **A:** Yes)	

Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]

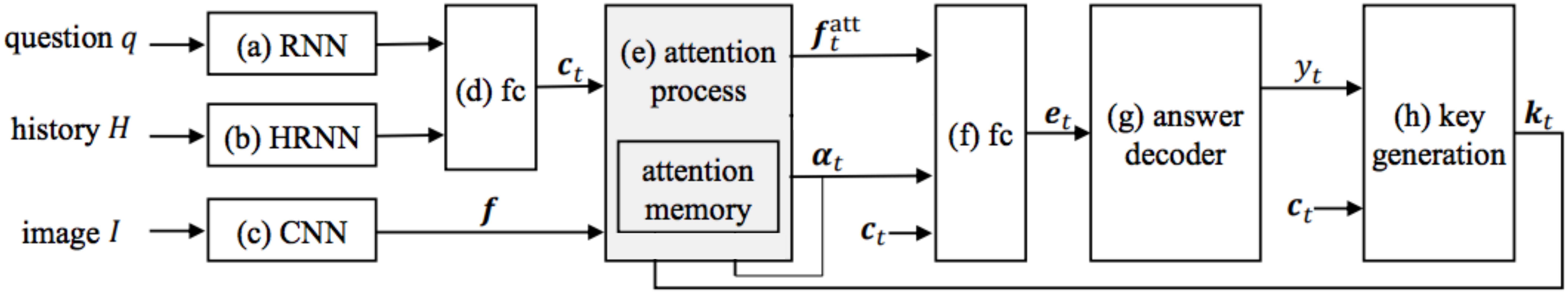


Associative Memory:

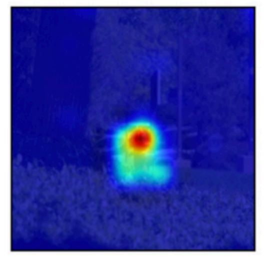
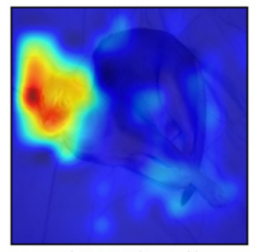
Question Turn	Key (hash)	Memory
1	f (**H:** Empty; **Q:** What color is a hydrant? **A:** It is red)	
2	f (**H:** …; **Q:** Is there a tree? **A:** Yes)	
	????	

Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]

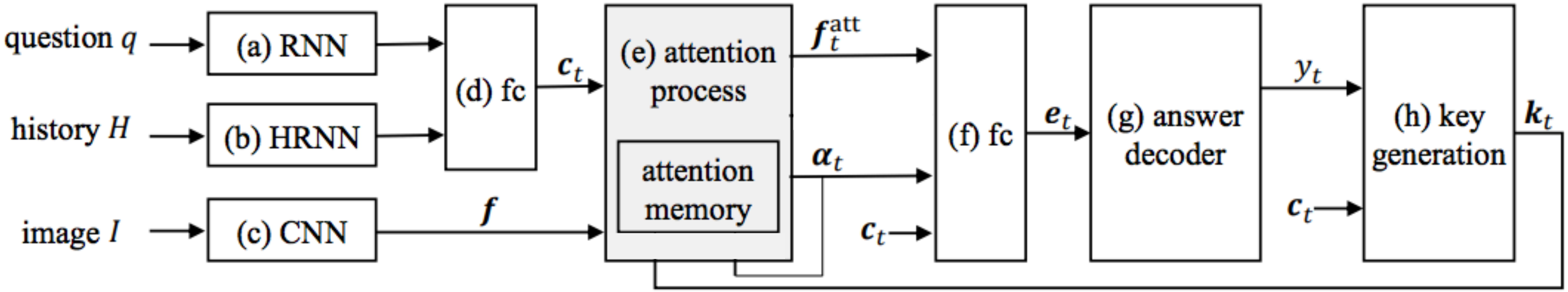


Associative Memory:

Question Turn	**Key (hash)**	**Memory**
1	f (**H:** Empty; **Q:** What color is a hydrant? **A:** It is red)	
2	f (**H:** …; **Q:** Is there a tree? **A:** Yes)	
	????	

Training

[Seo et al., NIPS 2017]



Network is **fully differentiable**, can be trained using BackProp

Experiments

[Seo et al., NIPS 2017]

MNIST Dialog Dataset (Programmatically Generated)

— 4x4 grid of MNIST digits

— Each digit has 4 **attributes** (color, background, numbers style)

— **Questions**: counting, attribute

— **Answers**: single word

Experiments

[Seo et al., NIPS 2017]

MNIST Dialog Dataset (Programmatically Generated)

— 4x4 grid of MNIST digits

— Each digit has 4 **attributes** (color, background, numbers style)

— **Questions**: counting, attribute

— **Answers**: single word

VisDial Dataset (Real images + AMT)

— MS-COCO images + Caption

— **Questions:** unconstrained

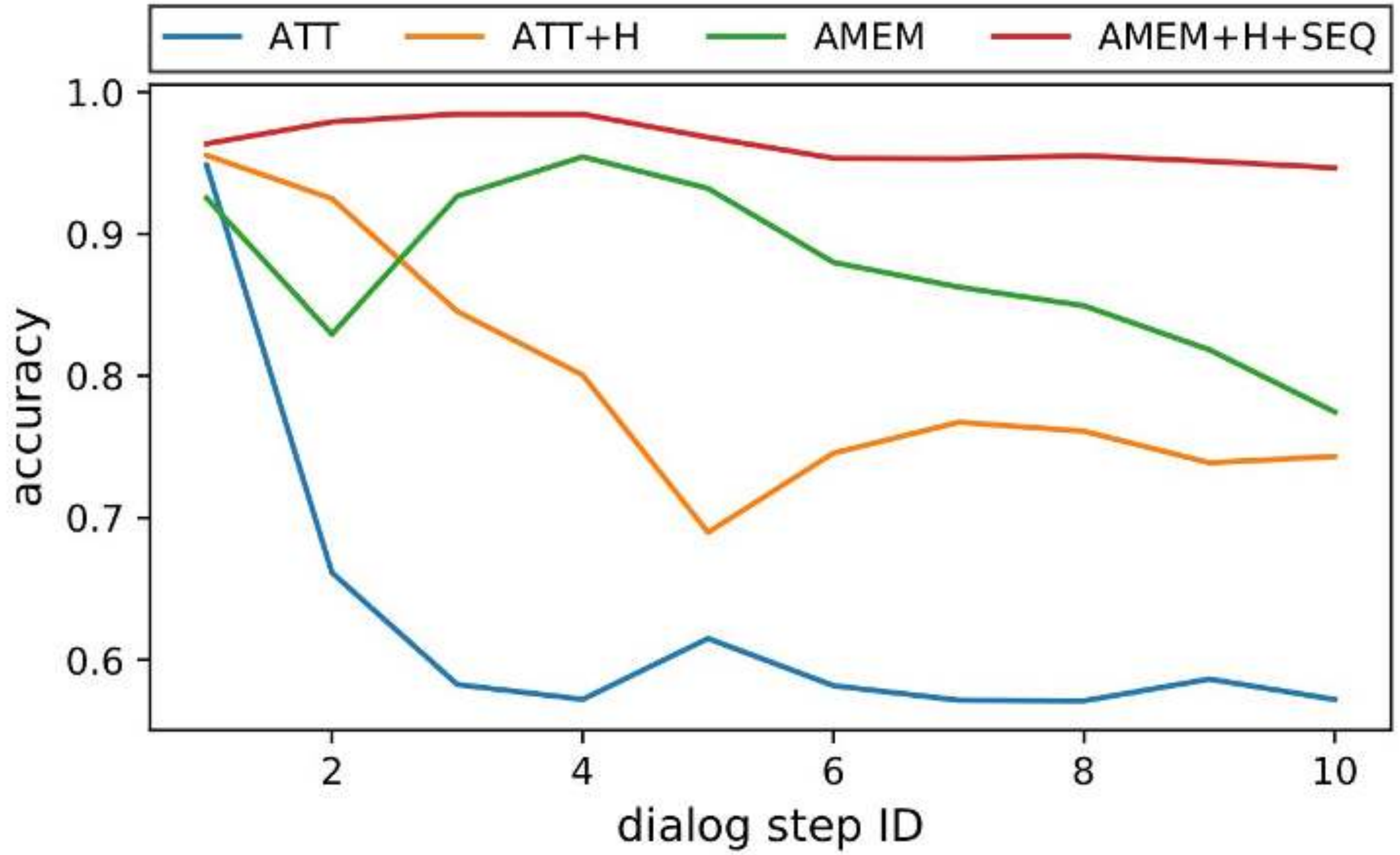
— **Answers:** free form text, 100 candidates

[Das, Kottur, Gupta, Singh, Yadav, Moura, Lee, Parikh, Batra, ICCV 2017]

Results: MNIST Dialog

[Seo et al., NIPS 2017]

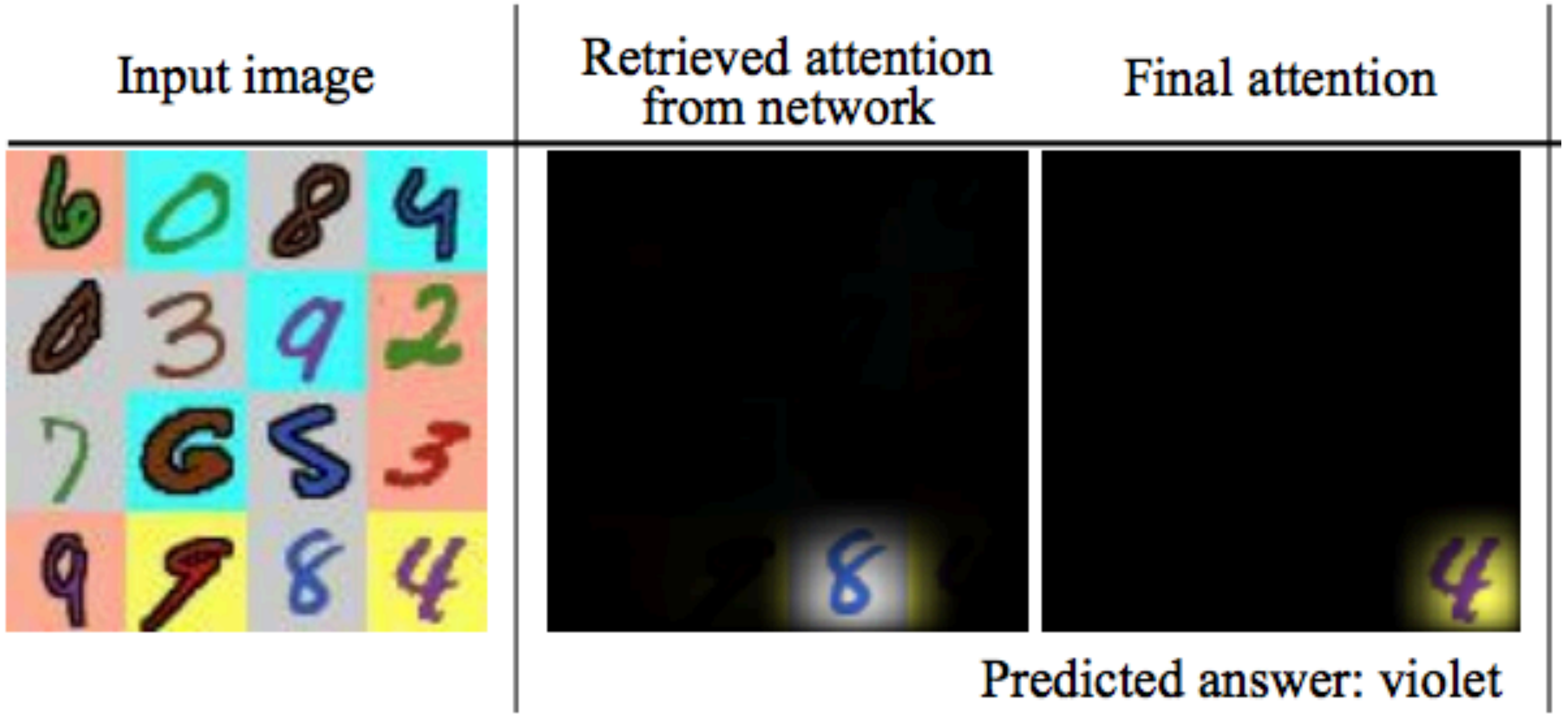
Basemodel	+H	+SEQ	Accuracy
I	–	–	20.18
Q	–	–	36.58
	✓	–	37.58
LF [1]	✓	–	45.06
HRE [1]	✓	–	49.10
MN [1]	✓	–	48.51
ATT	–	–	62.62
	✓	–	79.72
AMEM	–	–	87.53
	✓	–	89.20
	–	✓	90.05
	✓	✓	**96.39**



Results: Interpretability / Implicit Reasoning

[Seo et al., NIPS 2017]

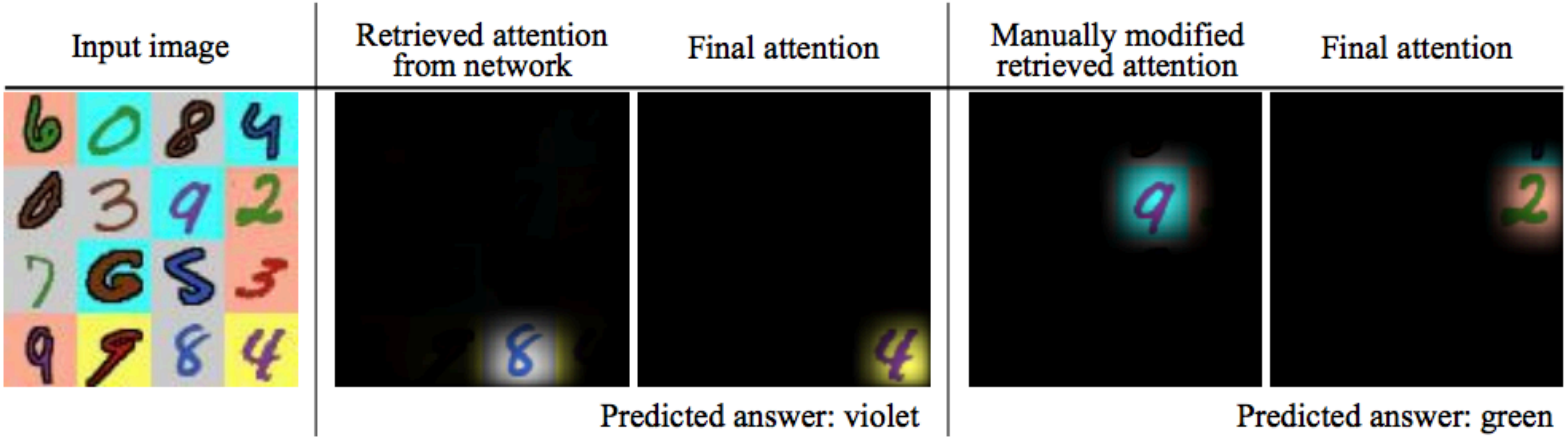
History:	Are there any 9's in the image ?	three
	How many digits in a yellow background are there among them ?	one
	What is the color of the digit ?	red
	What is the color of the digit at the right of it ?	blue
	What is the style of the blue digit ?	flat
Current QA:	What is the color of the digit at the right of it ?	violet



Results: Interpretability / Implicit Reasoning

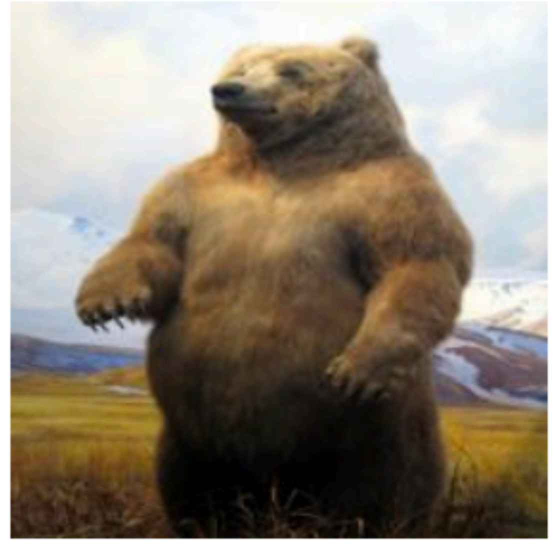
[Seo et al., NIPS 2017]

History:	Are there any 9's in the image ?	three
	How many digits in a yellow background are there among them ?	one
	What is the color of the digit ?	red
	What is the color of the digit at the right of it ?	blue
	What is the style of the blue digit ?	flat
Current QA:	What is the color of the digit at the right of it ?	violet



Results: VisDial

[Seo et al., NIPS 2017]

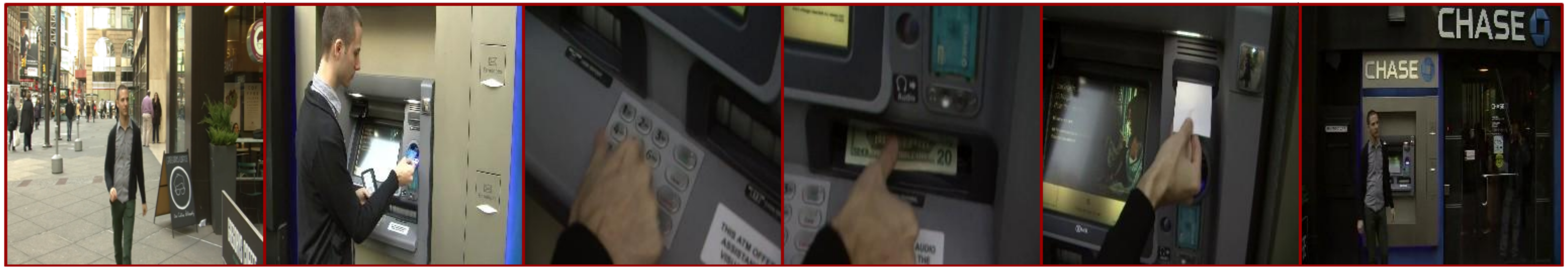
Dialog Information	Input image	Attended image
Caption: *A large bear standing upright with mountains in the background* Previous QA: *Is this the only bear here ? / yes* Current question: *What color is it's fur ?* GT answer: *Brown* Predicted answer: *Brown* Rank of GT: *1*		
Caption: *A train that is on a large rail way* Previous QA: *Is the train moving ? / No it is stopped* Current question: *What color is the train ?* GT answer: *It is white and red with some blue on it* Predicted answer: *It is white and red with some blue on it* Rank of GT: *1*		
Caption: *An airplane parked in the middle of a runway* Previous QA: *Can you see the airport ? / No* Current question: *Is it a sunny day ?* GT answer: *Yes* Predicted answer: *Yes* Rank of GT: *1*		

Results: VisDial

[Seo et al., NIPS 2017]

Model	+H	ATT	# of params	MRR	R@1	R@5	R@10	MR
Answer prior [24]	–	–	n/a	0.3735	23.55	48.52	53.23	26.50
LF-Q [24]	–	–	8.3 M (3.6x)	0.5508	41.24	70.45	79.83	7.08
LF-QH [24]	✓	–	12.4 M (5.4x)	0.5578	41.75	71.45	80.94	6.74
LF-QI [24]	–	–	10.4 M (4.6x)	0.5759	43.33	74.27	83.68	5.87
LF-QIH [24]	✓	–	14.5 M (6.3x)	0.5807	43.82	74.68	84.07	5.78
HRE-QH [24]	✓	–	15.0 M (6.5x)	0.5695	42.70	73.25	82.97	6.11
HRE-QIH [24]	✓	–	16.8 M (7.3x)	0.5846	44.67	74.50	84.22	5.72
HREA-QIH [24]	✓	–	16.8 M (7.3x)	0.5868	44.82	74.81	84.36	5.66
MN-QH [24]	✓	–	12.4 M (5.4x)	0.5849	44.03	75.26	84.49	5.68
MN-QIH [24]	✓	–	14.7 M (6.4x)	0.5965	45.55	76.22	85.37	5.46
SAN-QI [9]	–	✓	n/a	0.5764	43.44	74.26	83.72	5.88
HieCoAtt-QI [14]	–	✓	n/a	0.5788	43.51	74.49	83.96	5.84
AMEM-QI	–	✓	**1.7 M (0.7x)**	0.6196	48.24	78.33	87.11	4.92
AMEM-QIH	✓	✓	2.3 M (1.0x)	0.6192	48.05	78.39	87.12	4.88
AMEM+SEQ-QI	–	✓	**1.7 M (0.7x)**	**0.6227**	**48.53**	**78.66**	**87.43**	**4.86**
AMEM+SEQ-QIH	✓	✓	2.3 M (1.0x)	0.6210	48.40	78.39	87.12	4.92

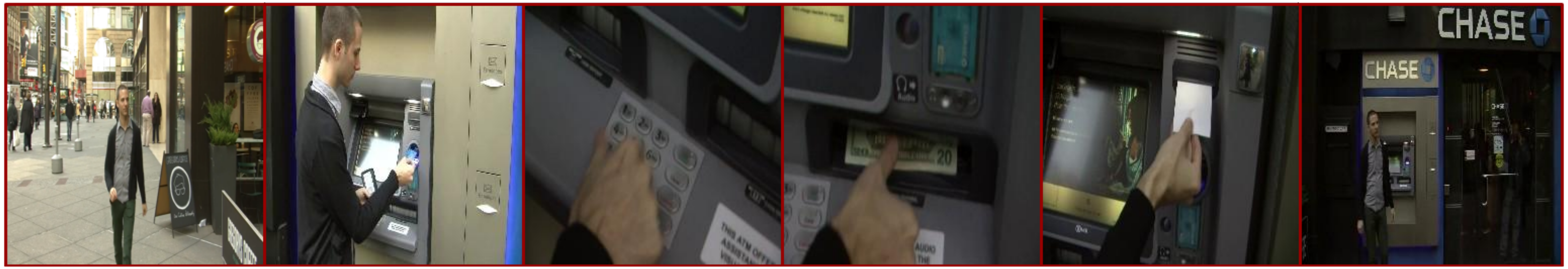
Applications: Activity Detection



[Ma et al., 2014]

Applications: Activity Detection

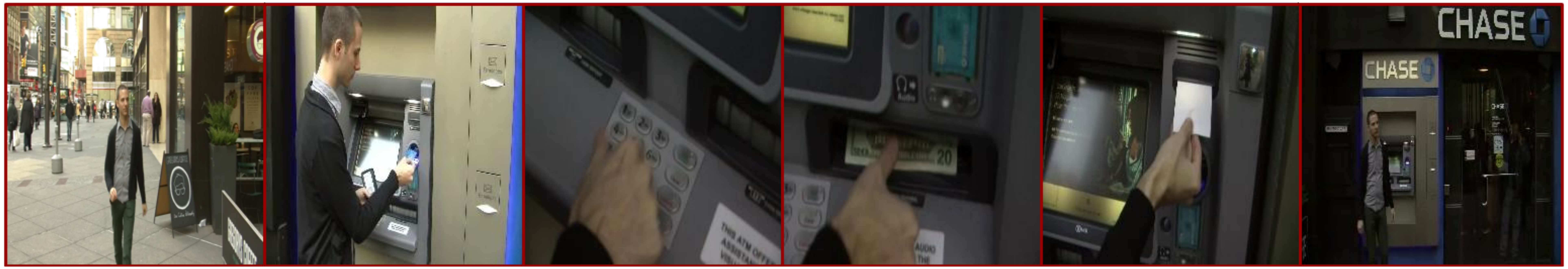
Activity: A collection of human/object movements with a particular semantic meaning



[Ma et al., 2014]

Applications: Activity Detection

Activity: A collection of human/object movements with a particular semantic meaning



Action Recognition: Finding if a video segment contains such a movement

[Ma et al., 2014]

Applications: Activity Detection

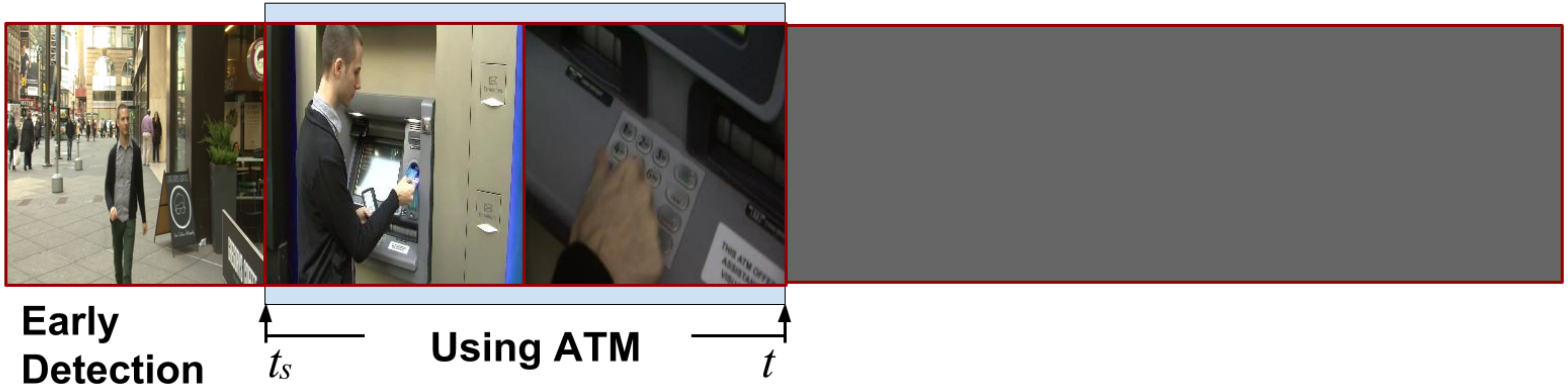
Activity: A collection of human/object movements with a particular semantic meaning

Action Recognition: Finding if a video segment contains such a movement

Action Detection: Finding a segment (beginning and start) and recognize the action in it

[Ma et al., 2014]

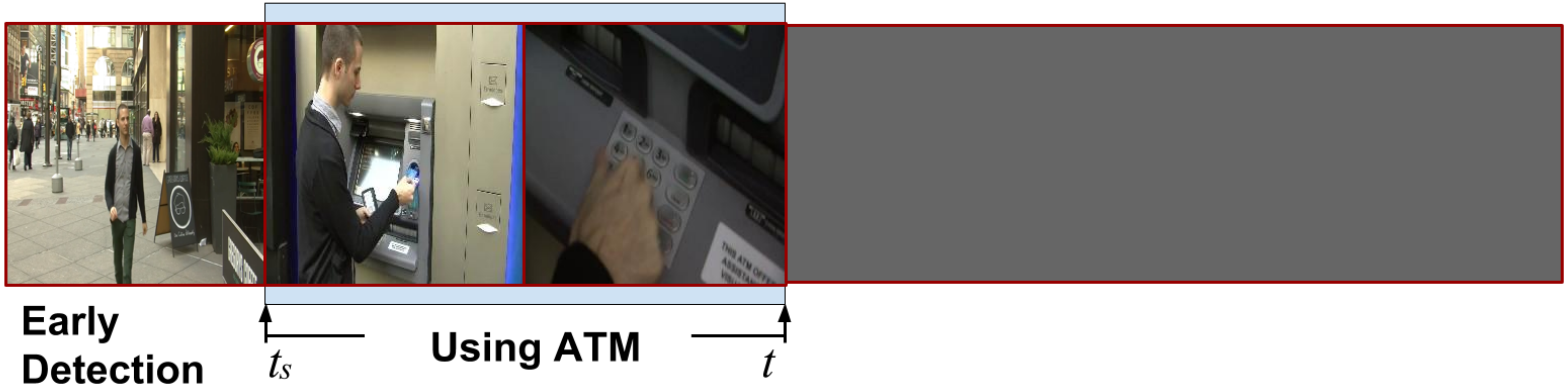
Applications: Activity Detection



[Ma et al., 2014]

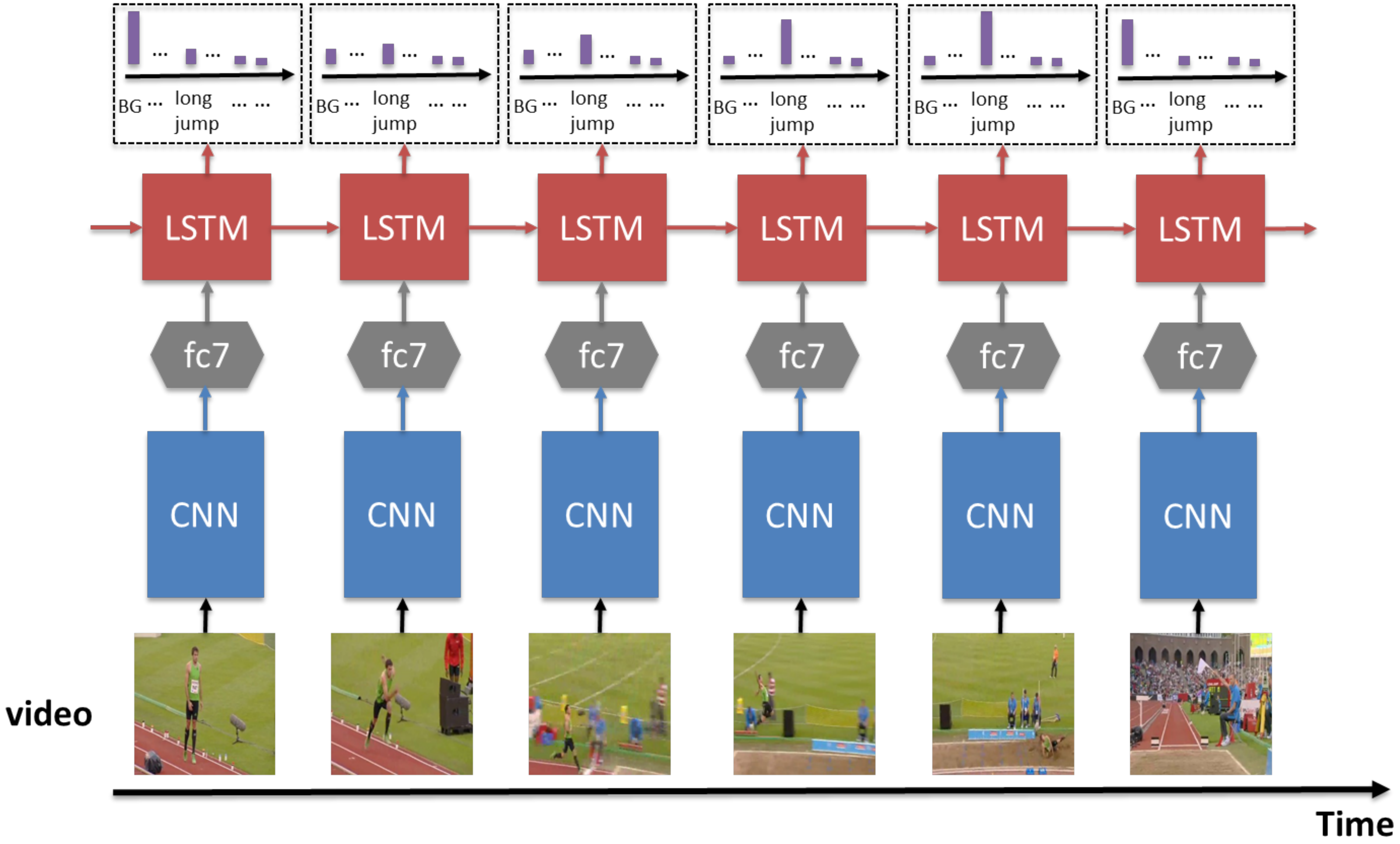
Applications: Activity Detection

Early Detection: Recognize when an action starts and try to predict which action is performed as quickly as possible.



[Ma et al., 2014]

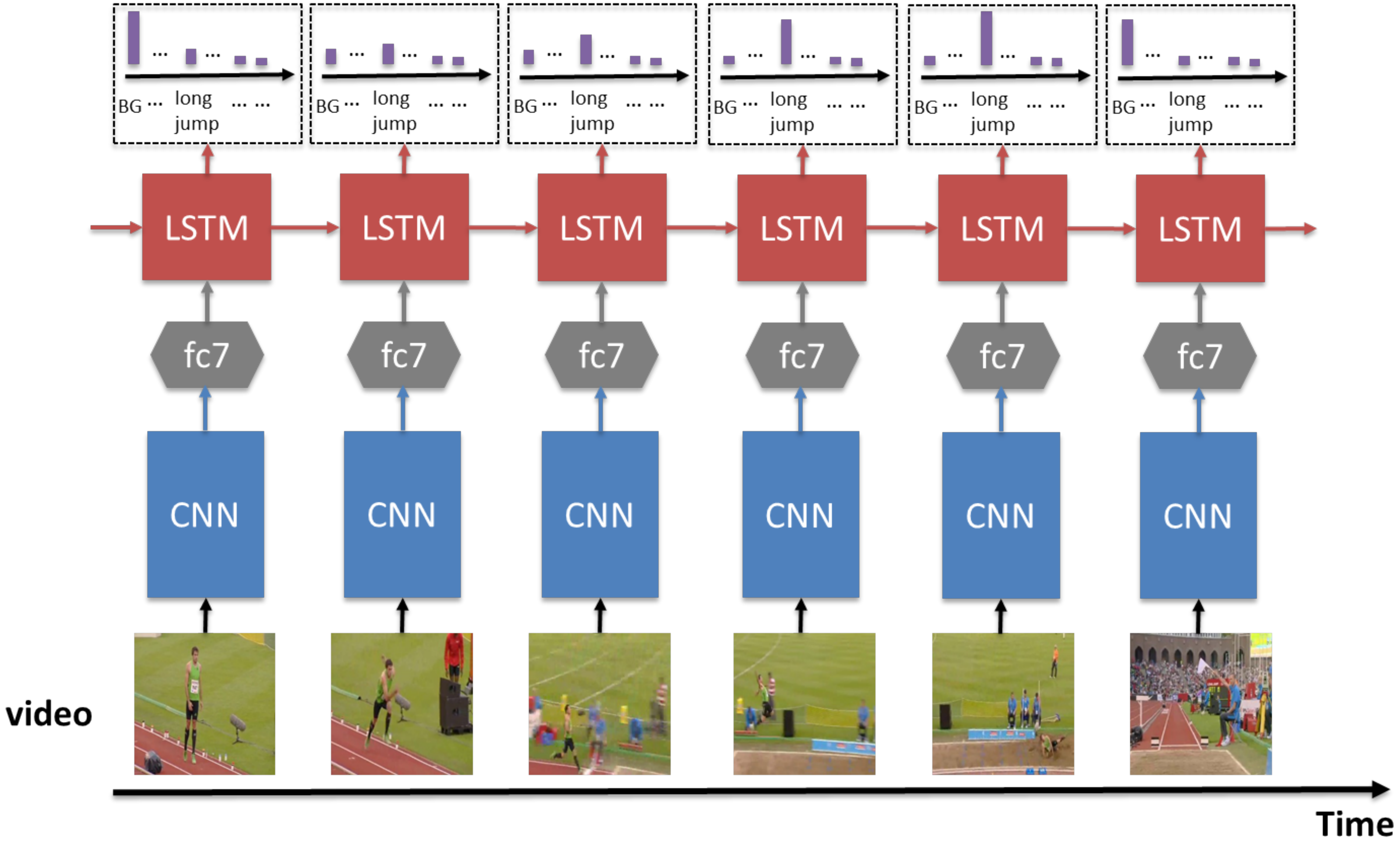
Applications: Activity Detection



[Ma et al., 2014]

Applications: Activity Detection

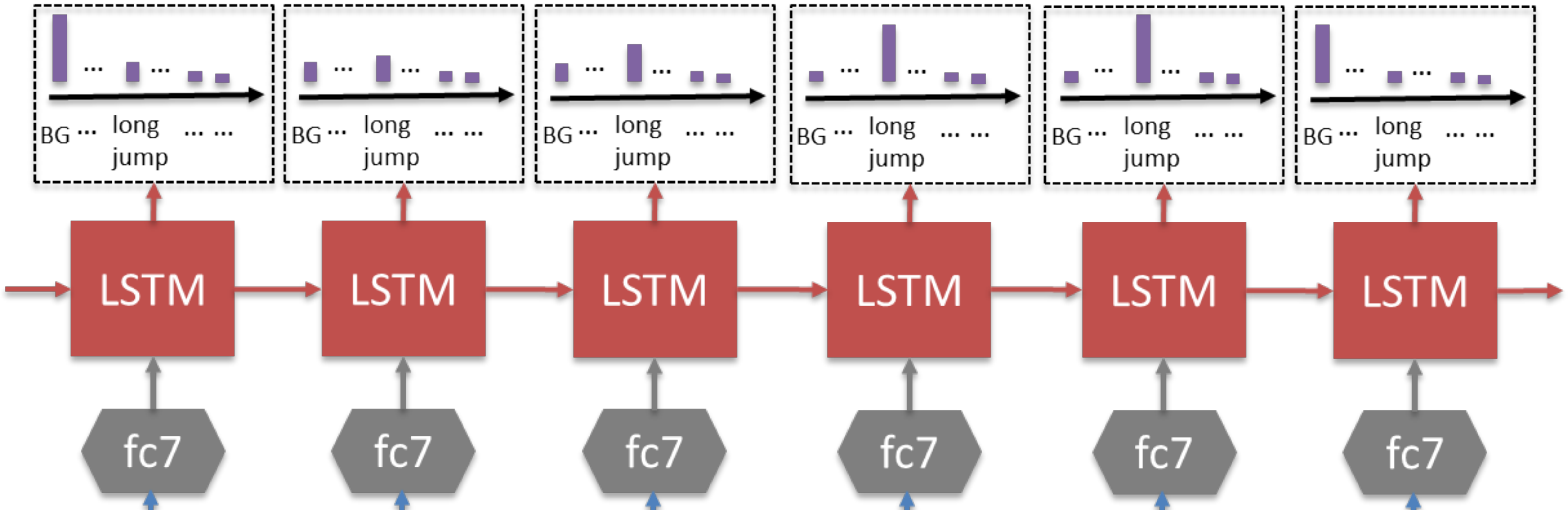
Penalty at every time step is the same

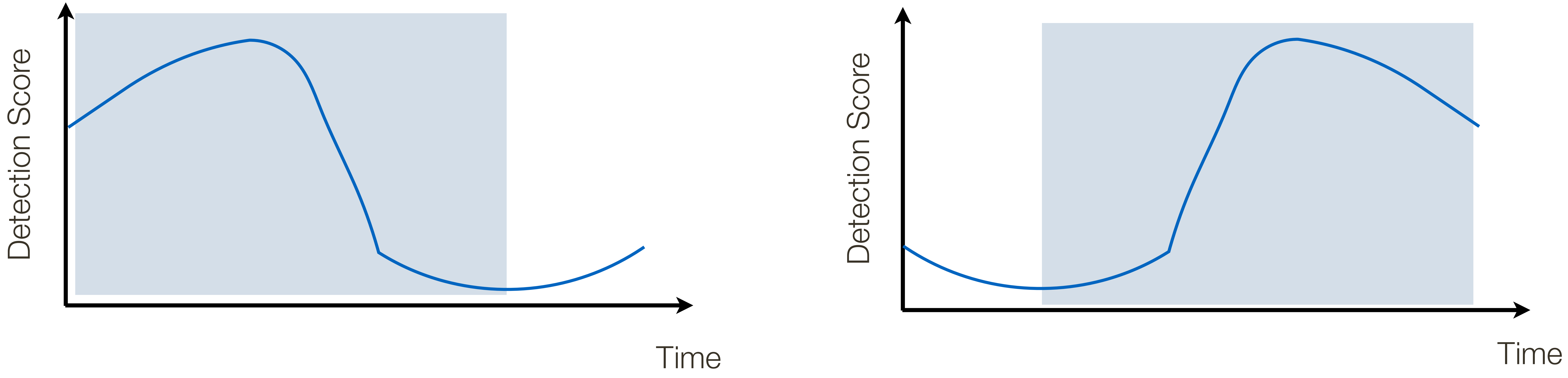


[Ma et al., 2014]

Applications: Activity Detection

Penalty at every time step is the same





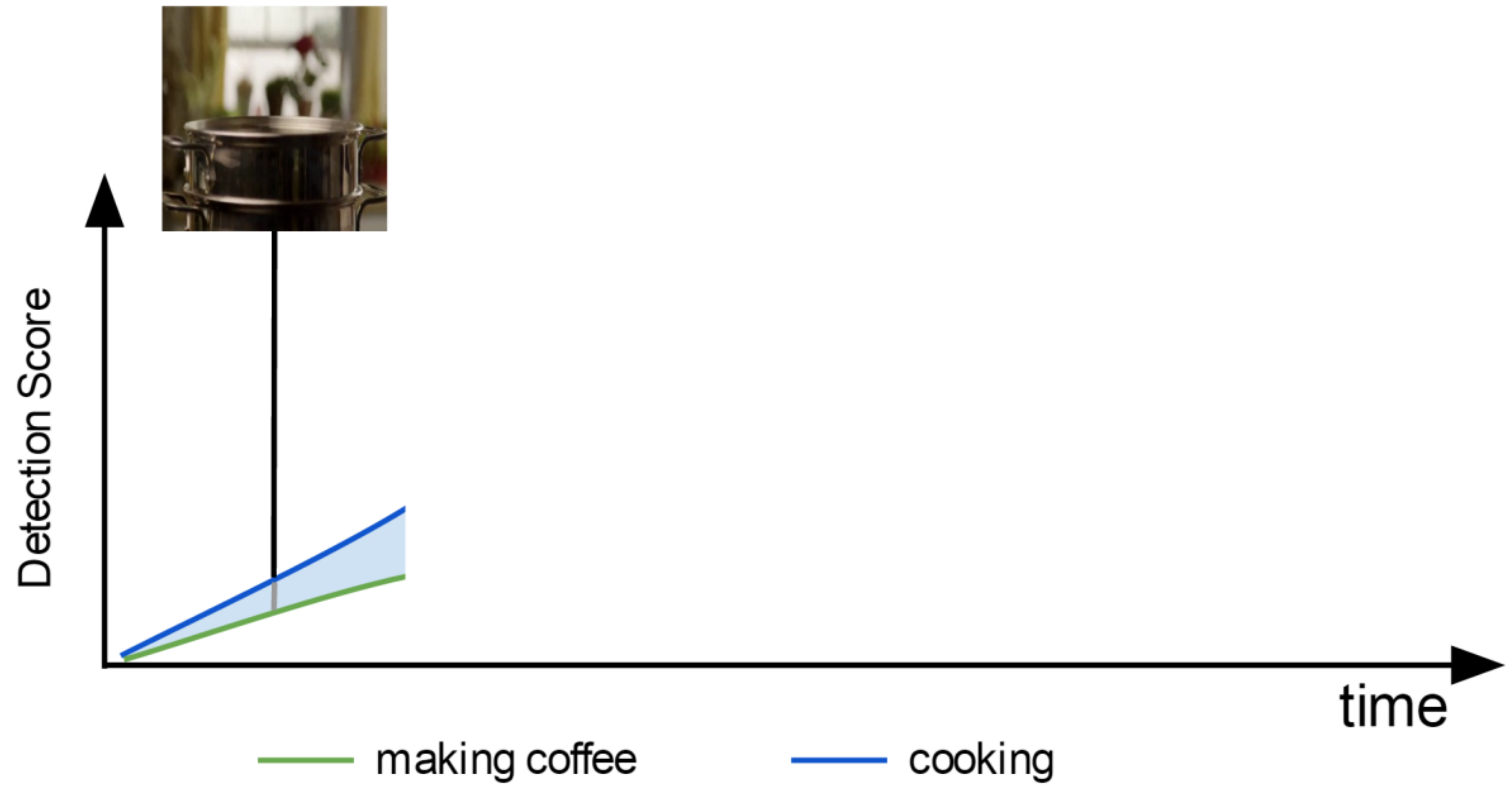
[Ma et al., 2014]

Applications: Activity Detection

As the detector sees more of an action, it should become more confident of

— Detecting the correct action class

— More confident that it is not the incorrect action class



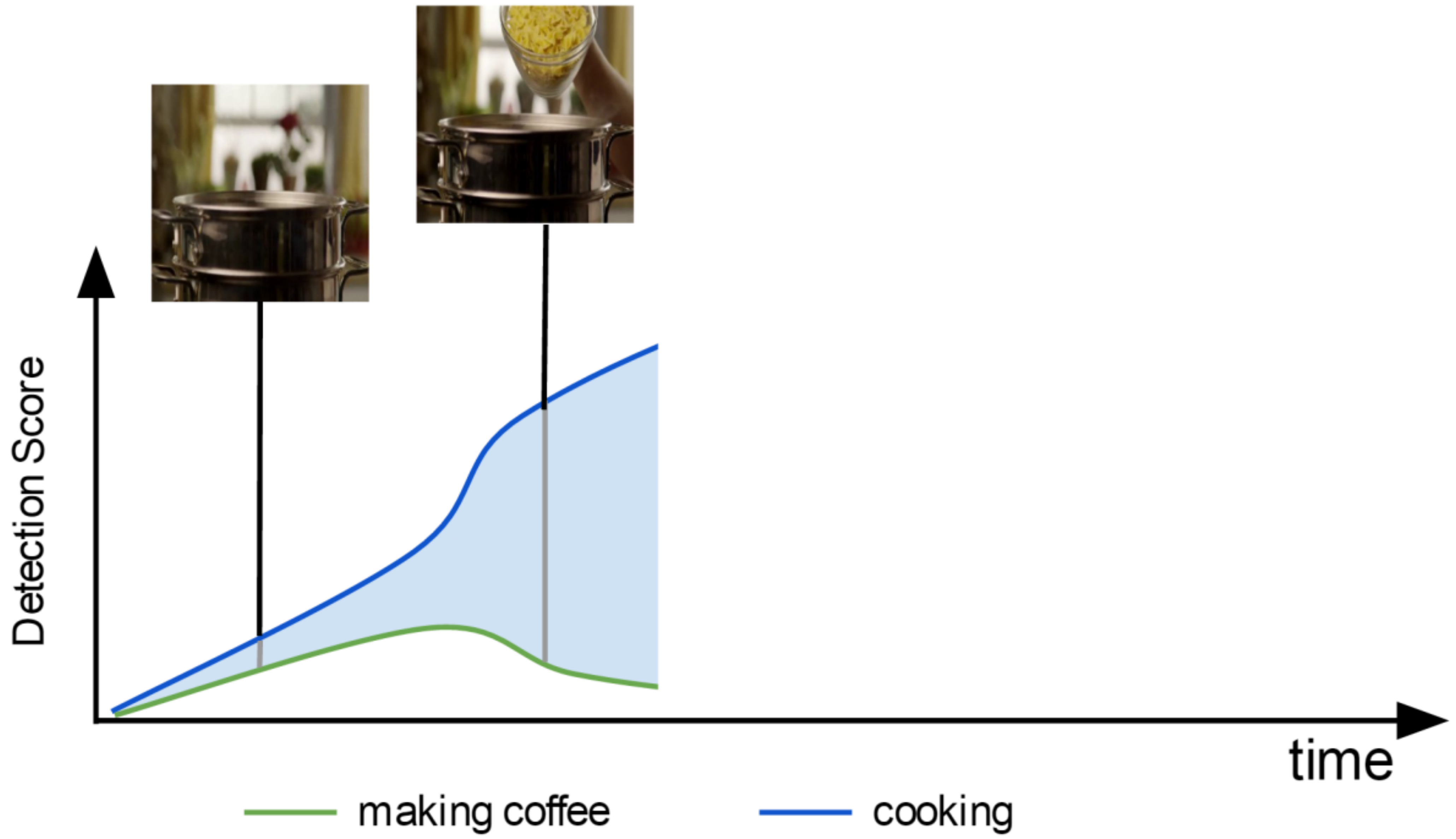
[Ma et al., 2014]

Applications: Activity Detection

As the detector sees more of an action, it should become more confident of

— Detecting the correct action class

— More confident that it is not the incorrect action class



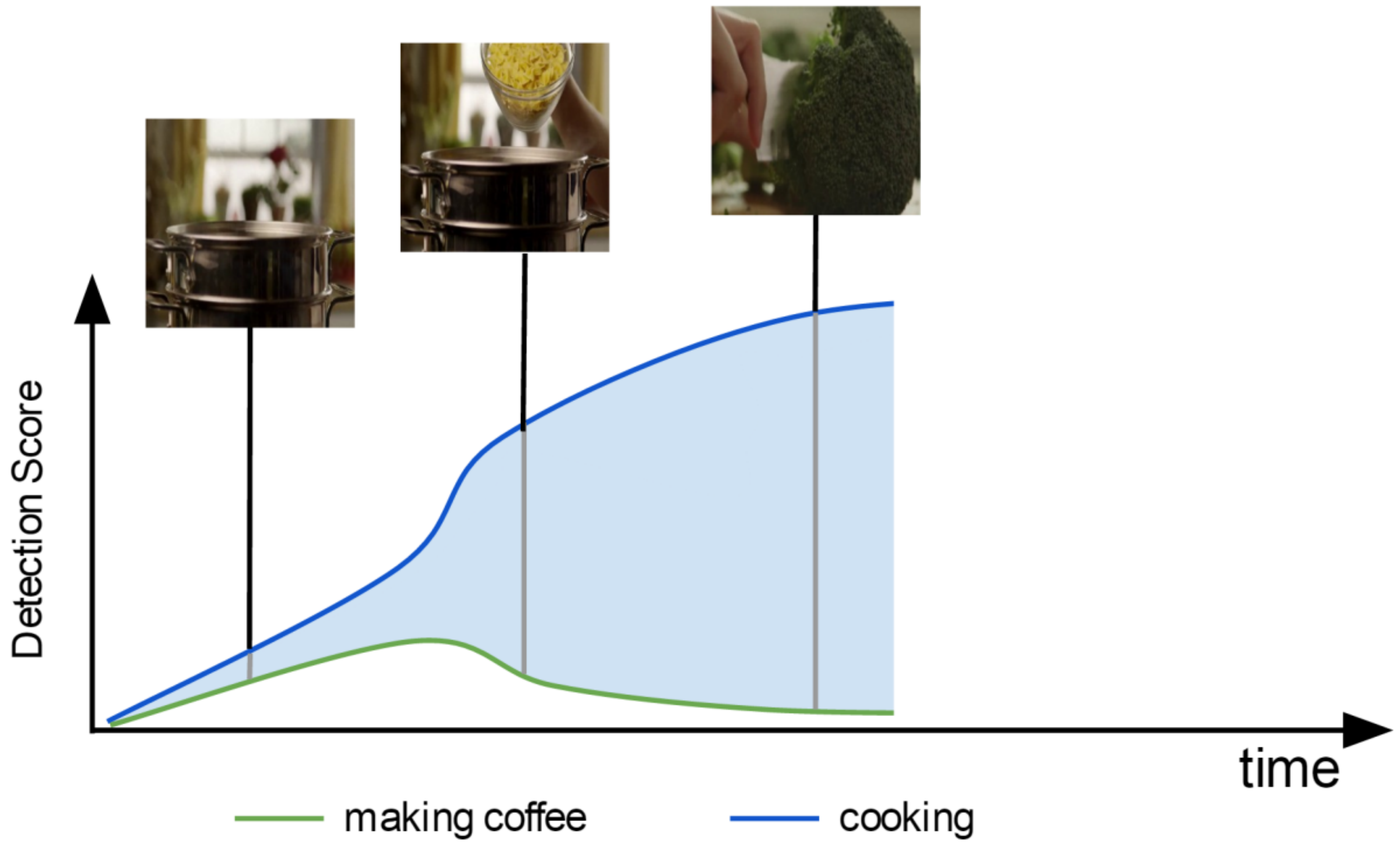
[Ma et al., 2014]

Applications: Activity Detection

As the detector sees more of an action, it should become more confident of

— Detecting the correct action class

— More confident that it is not the incorrect action class



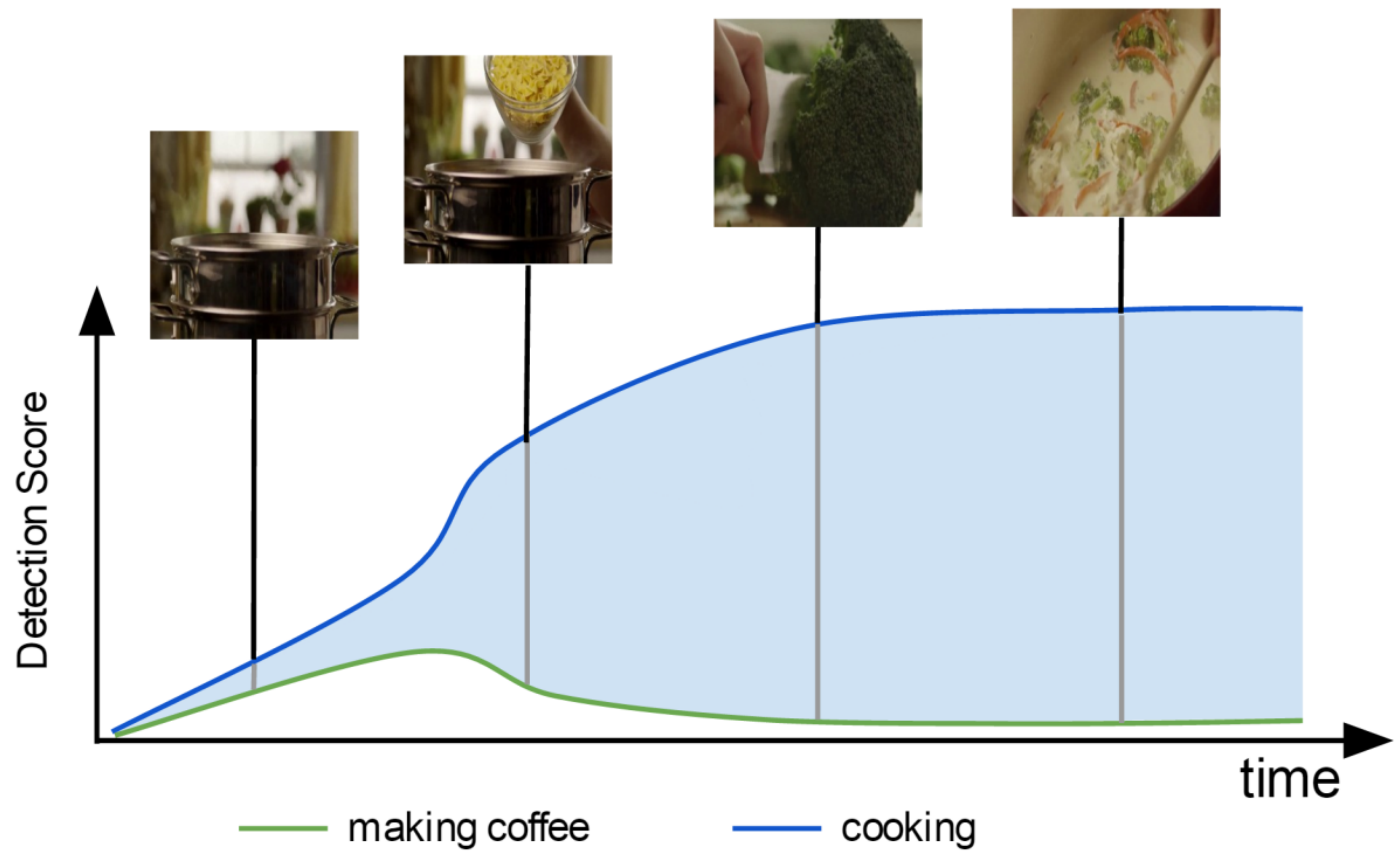
[Ma et al., 2014]

Applications: Activity Detection

As the detector sees more of an action, it should become more confident of

— Detecting the correct action class

— More confident that it is not the incorrect action class



[Ma et al., 2014]

New Class of Loss Functions

Classification loss at time t

Training loss at time t: $\mathcal{L}^t = \mathcal{L}_c^t + \lambda_r \mathcal{L}_r^t$

Ranking loss at time t

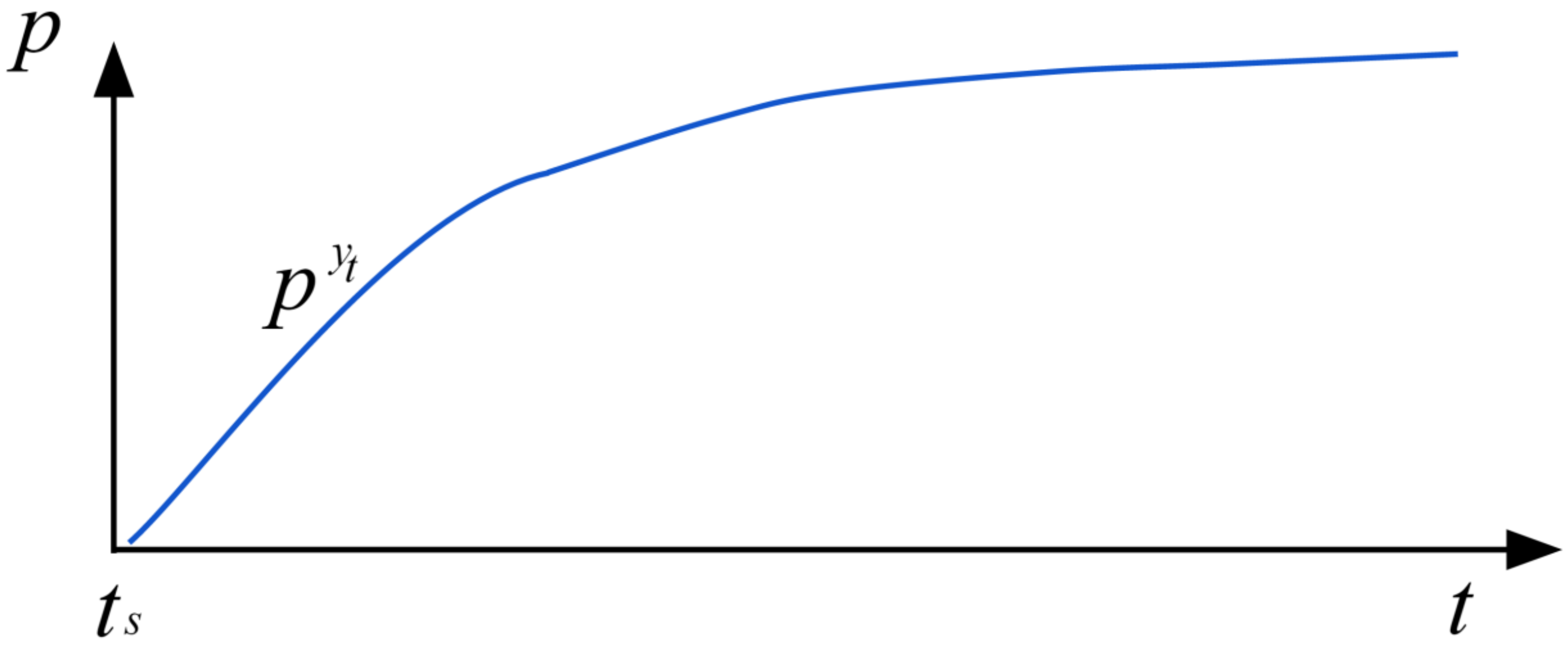
$\mathcal{L}_r^t$ is one of the following:

- $\mathcal{L}_s^t$ ranking loss on detection score
- $\mathcal{L}_m^t$ ranking loss on discriminative margin

[Ma et al., 2014]

Ranking Loss on Detection Score $\mathcal{L}_s^t$

Ideally what we want:

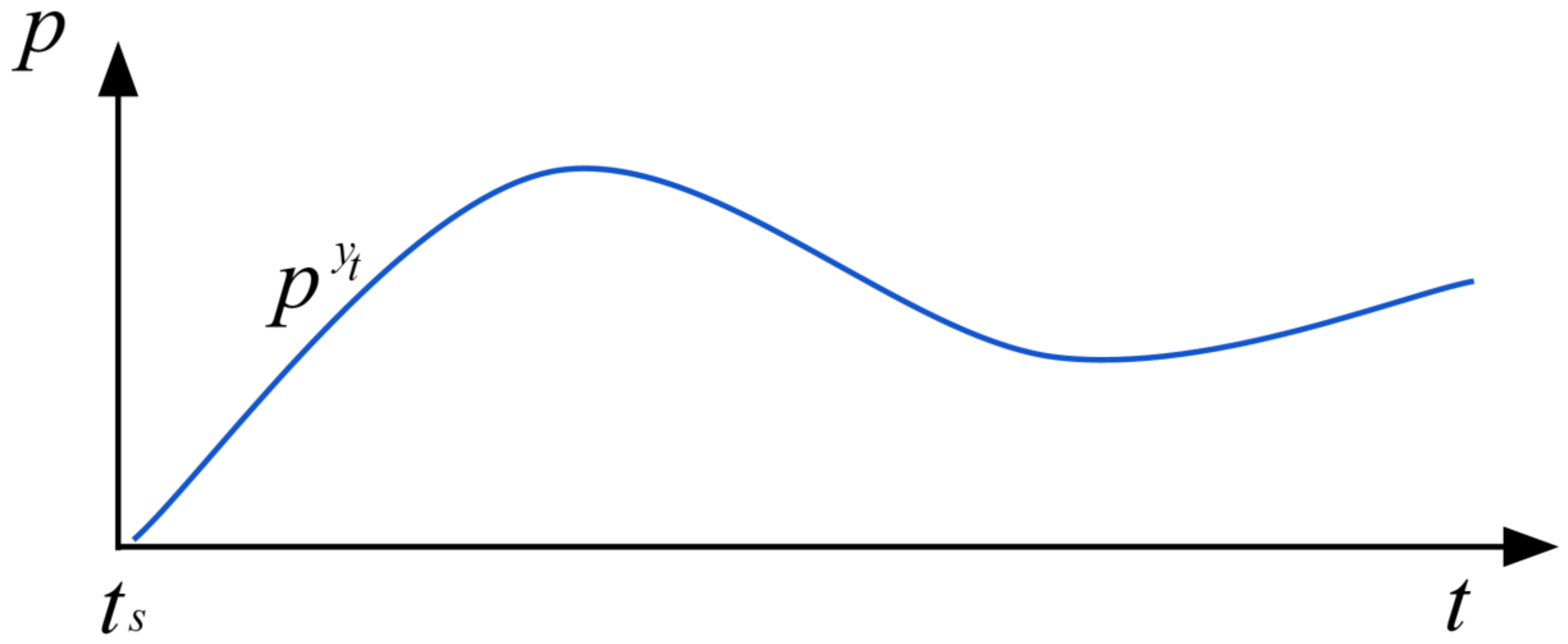


Prediction score of the ground truth action label

[Ma et al., 2014]

Ranking Loss on Detection Score $\mathcal{L}_s^t$

In **Practice:**



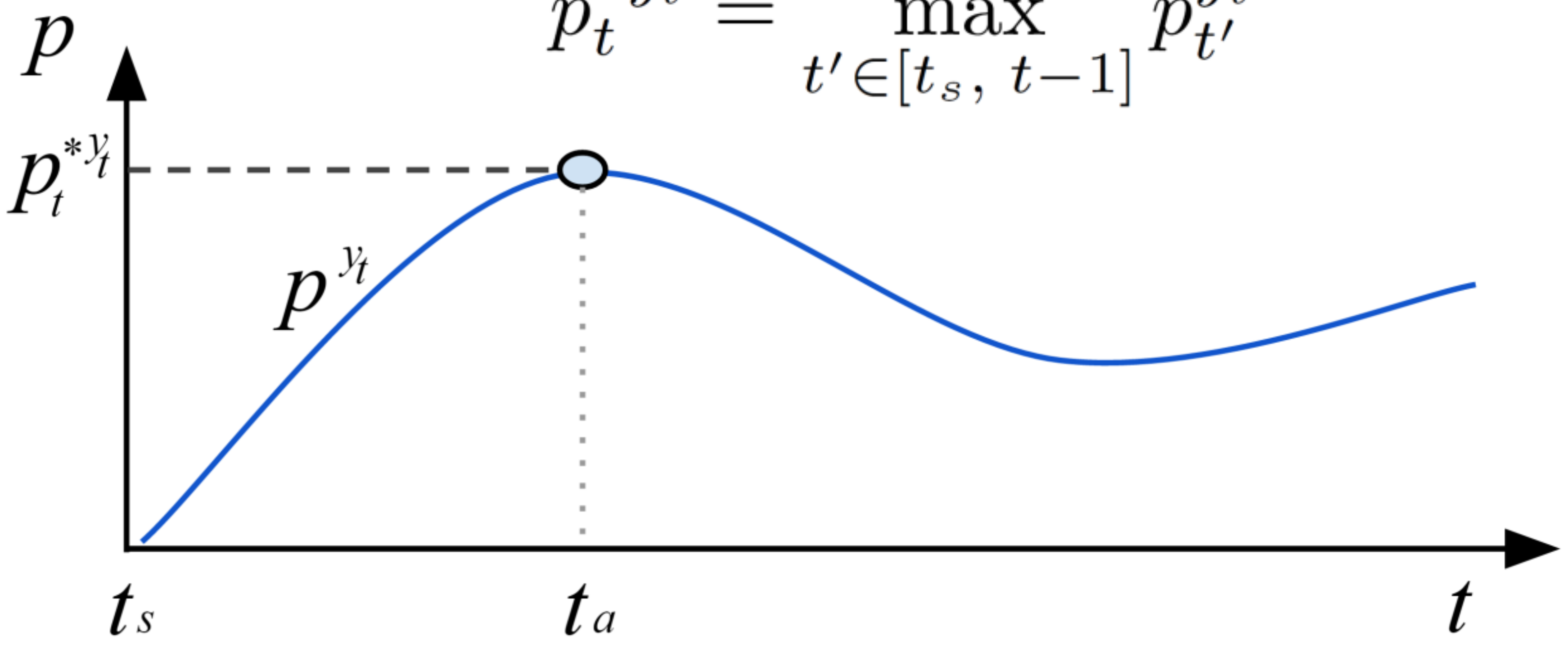
Prediction score of the ground truth action label

[Ma et al., 2014]

Ranking Loss on Detection Score $\mathcal{L}_s^t$

In **Practice:**

$$p_t^{*y_t} = \max_{t' \in [t_s,\, t-1]} p_{t'}^{y_t}$$

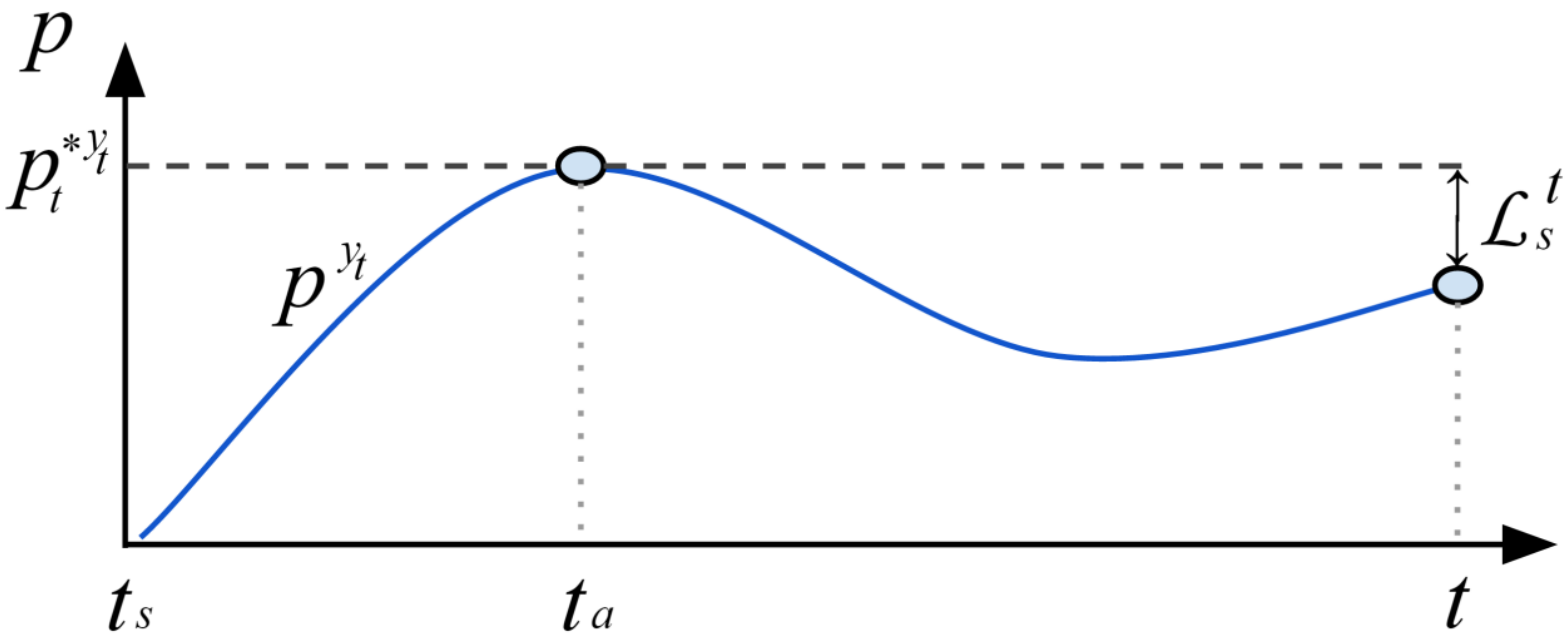


Prediction score of the ground truth action label

[Ma et al., 2014]

Ranking Loss on Detection Score $\mathcal{L}_s^t$

In **Practice:**



Prediction score of the ground truth action label

[Ma et al., 2014]

Applications: Activity Detection

Activity detection performance measured in mAP at different IOU thresholds

Model	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$
Heilbron *et al.*	12.5%	11.9%	11.1%	10.4%	9.7%	-	-	-
CNN	30.1%	26.9%	23.4%	21.2%	18.9%	17.5%	16.5%	15.8%
LSTM	48.1%	44.3%	40.6%	35.6%	31.3%	28.3%	26.0%	24.6%
LSTM-m	52.6%	48.9%	45.1%	40.1%	35.1%	31.8%	29.1%	27.2%
LSTM-s	**54.0%**	**50.1%**	**46.3%**	**41.2%**	**36.4%**	**33.0%**	**30.4%**	**28.7%**

LSTM-m LSTM trained using both classification loss and rank loss on *discriminative margin*.

LSTM-s LSTM trained using both classification loss and rank loss on *detection score*.

[Ma et al., 2014]

Applications: Early Activity Detection

Activity early detection performance measured in mAP at different IOU thresholds

Model	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$
CNN	27.0%	23.4%	20.4%	17.2%	14.6%	12.3%	11.0%	10.3%
LSTM	49.5%	44.7%	38.8%	33.9%	29.6%	25.6%	23.5%	22.4%
LSTM-m	52.6%	47.9%	41.5%	36.2%	31.4%	27.1%	24.8%	23.5%
LSTM-s	**55.1%**	**50.3%**	**44.0%**	**38.9%**	**34.1%**	**29.8%**	**27.4%**	**26.1%**

Note: first 3/10 of activity is seen by a detector

LSTM-m LSTM trained using both classification loss and rank loss on *discriminative margin*.

LSTM-s LSTM trained using both classification loss and rank loss on *detection score*.

[Ma et al., 2014]

Applications: Early Activity Detection

Activity early detection performance measured in mAP at different IOU thresholds

Model	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$
CNN	27.0%	23.4%	20.4%	17.2%	14.6%	12.3%	11.0%	10.3%
LSTM	49.5%	44.7%	38.8%	33.9%	29.6%	25.6%	23.5%	22.4%
LSTM-m	52.6%	47.9%	41.5%	36.2%	31.4%	27.1%	24.8%	23.5%
LSTM-s	**55.1%**	**50.3%**	**44.0%**	**38.9%**	**34.1%**	**29.8%**	**27.4%**	**26.1%**

Note: first 3/10 of activity is seen by a detector

LSTM-m LSTM trained using both classification loss and rank loss on *discriminative margin*.

LSTM-s LSTM trained using both classification loss and rank loss on *detection score*.

Take home: Early detection is only 1-3% worse than sewing the whole sequence

[Ma et al., 2014]

Applications: Activity Detection

[Ma et al., 2014]

Applications: Activity Detection

[Ma et al., 2014]