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

Lecture 12: RNN Applications (Part 2)

Logistics

— **Assignment 1 & 2** grading is ongoing (a bit slow)

— **Assignment 3** is now due end-of-day Sunday

— **Assignment 4** will be available tonight

— **Quiz for final project groups** is on Canvas (due Thursday)

Review: Generalized Soft Attention

$$\beta_{i,t} = score(\mathbf{h}_i^{(enc)}, \cancel{\mathbf{h}_t^{(dec)}})$$

Relevance of encoding at token i for decoding token t

Query: $\mathbf{Q}_t$

$$\beta_{i,t} = score(\mathbf{W}_k \mathbf{h}_i^{(enc)}, \mathbf{W}_q \mathbf{x}_t^{(dec)})$$

$$\beta_{i,t} = score(\mathbf{W}_k \mathbf{h}_i^{(enc)}, \mathbf{W}_q \mathbf{h}_{t-1}^{(dec)})$$

$$\beta_{i,t} = score(\mathbf{W}_k \mathbf{h}_i^{(enc)}, \mathbf{W}_q [\mathbf{x}_t^{(dec)}, \mathbf{h}_{t-1}^{(dec)}])$$

Key: $\mathbf{K}_i$

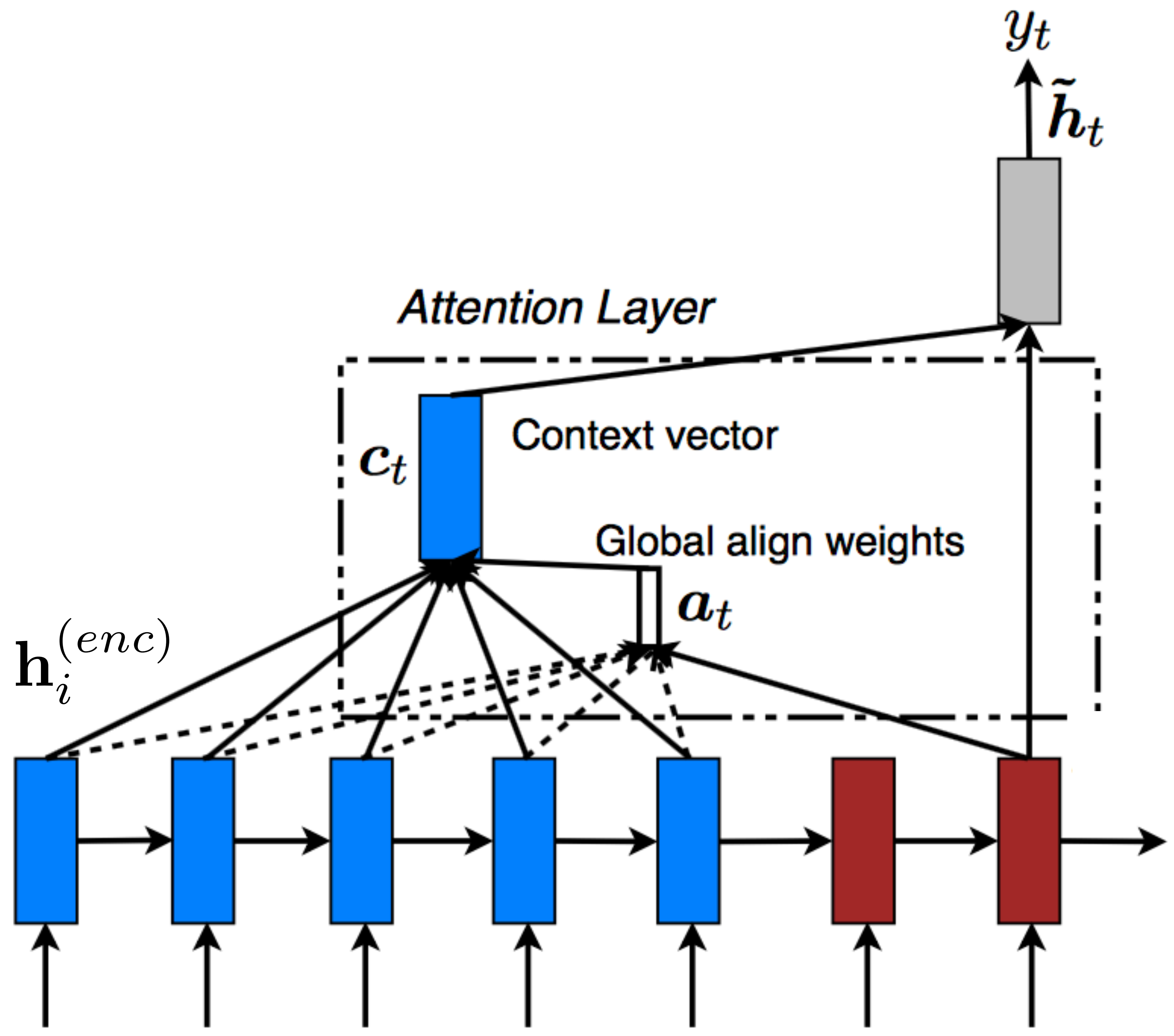
$$\alpha_{i,t} = \text{Softmax}(\beta_{i,t})$$

Normalize the weights to sum to 1

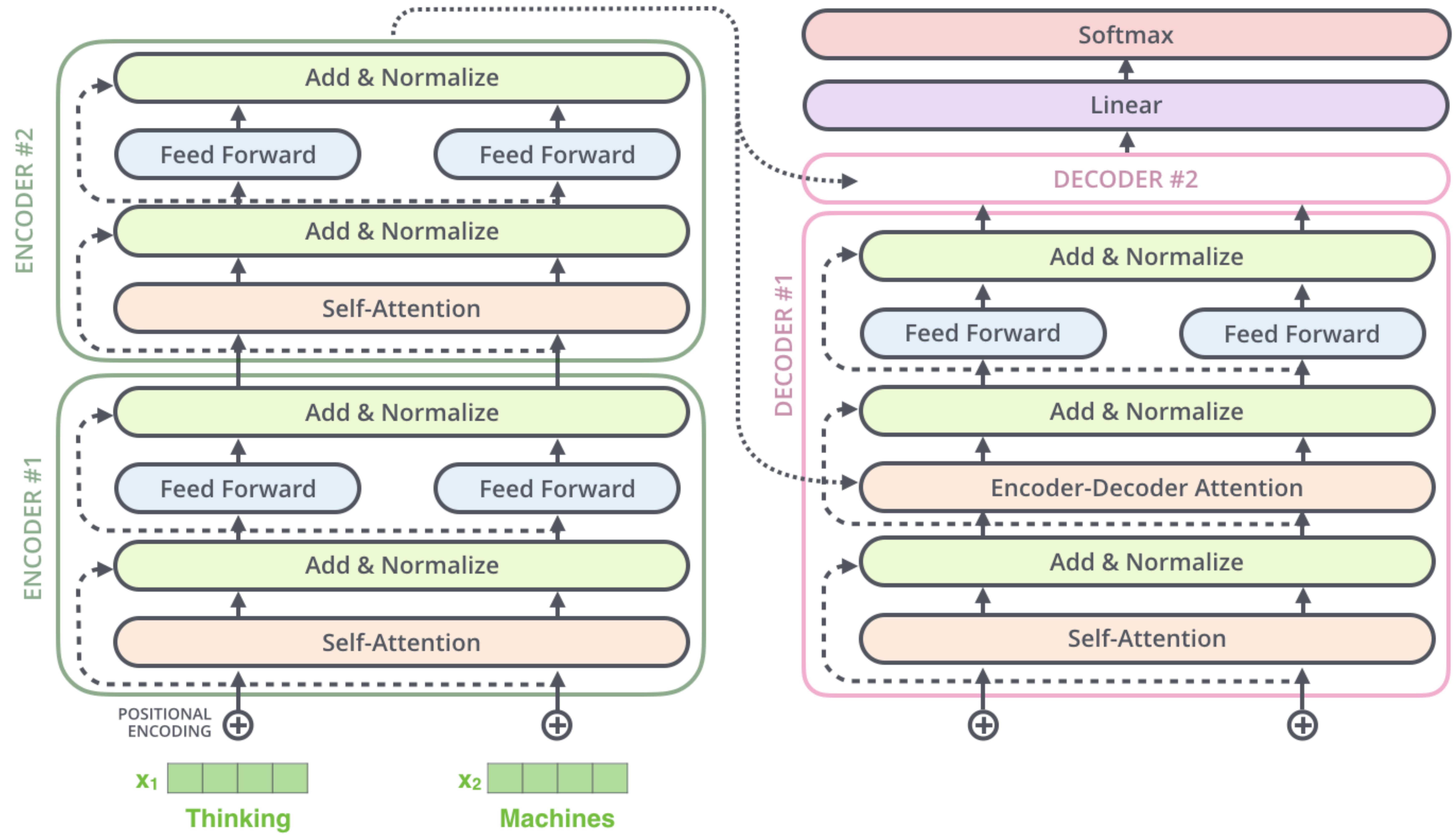
$$\mathbf{c}_t = \sum_i \alpha_{i,t} \mathbf{W}_v \mathbf{h}_i^{(enc)}$$

Value: $\mathbf{V}_i$

Form a context vector that would simply be added to the standard decoder input

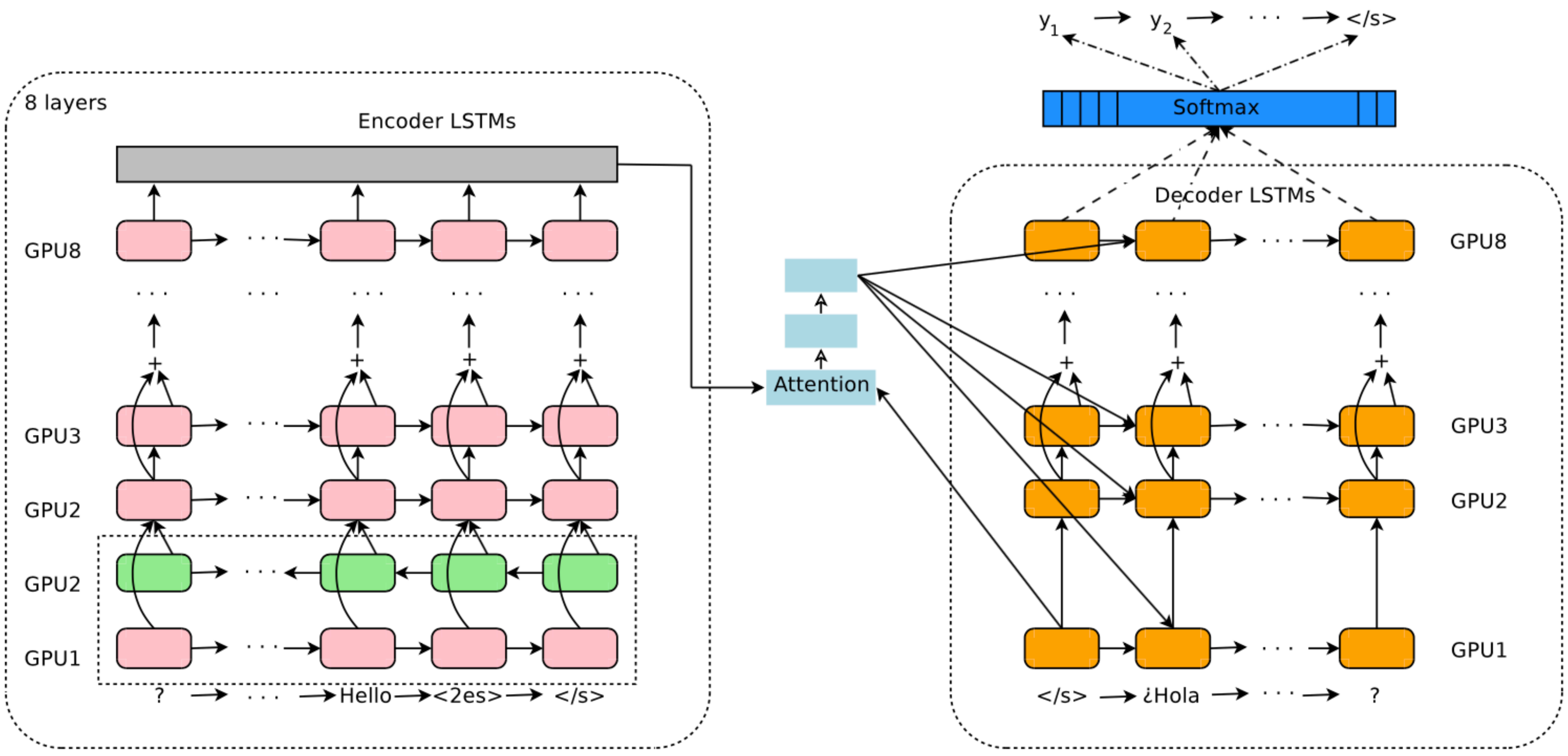


Review: Transformers



Applications: Google Language Translation

One model to translate from **any language** to any other language



[Johnson et al., 2017]

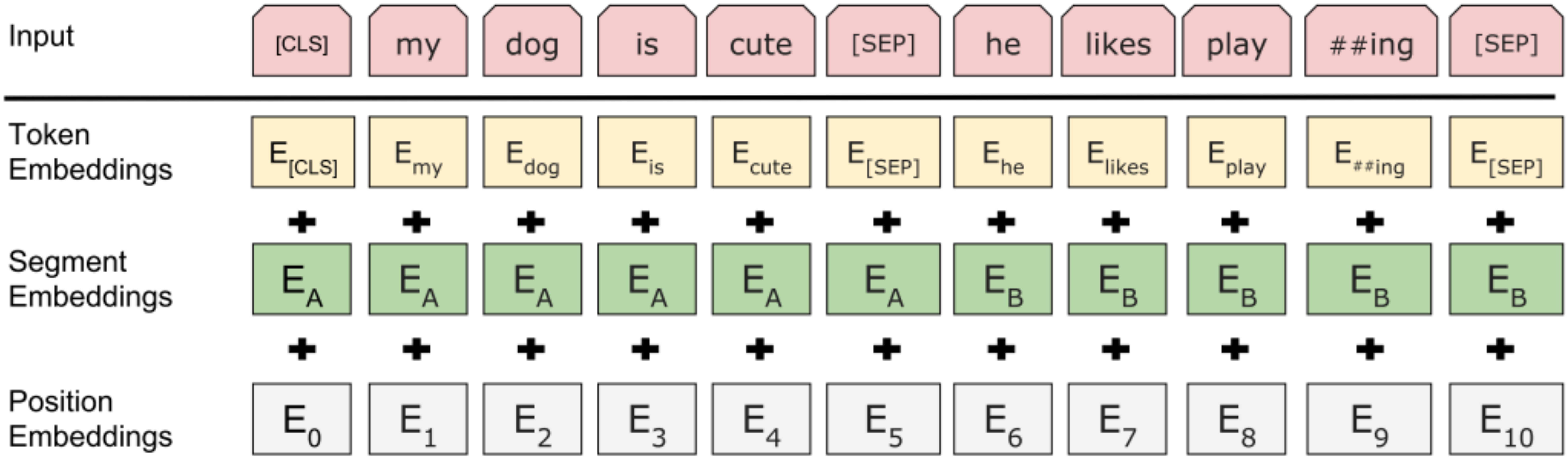
Applications: Masked Language Modeling (BERT)

To learn relationships between sentences, predict whether Sentence B is actual sentence that **proceeds** Sentence A, or a random sentence

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless
Label = NotNextSentence

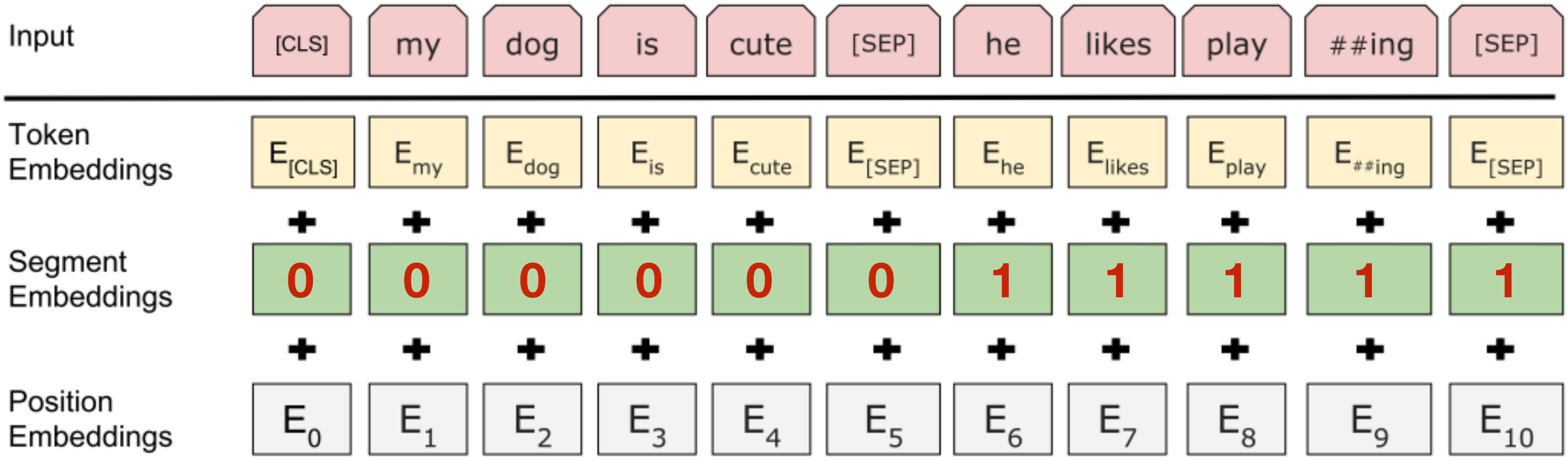
Applications: Masked Language Modeling (BERT)



Use 30,000 WordPiece **vocabulary**

Each token is a **sum of three** embeddings

Applications: Masked Language Modeling (BERT)



Use 30,000 WordPiece **vocabulary**

Each token is a **sum of three** embeddings

Applications: Masked Language Modeling (BERT)

Multi-headed self **attention**

— Models context

Feed-forward layers

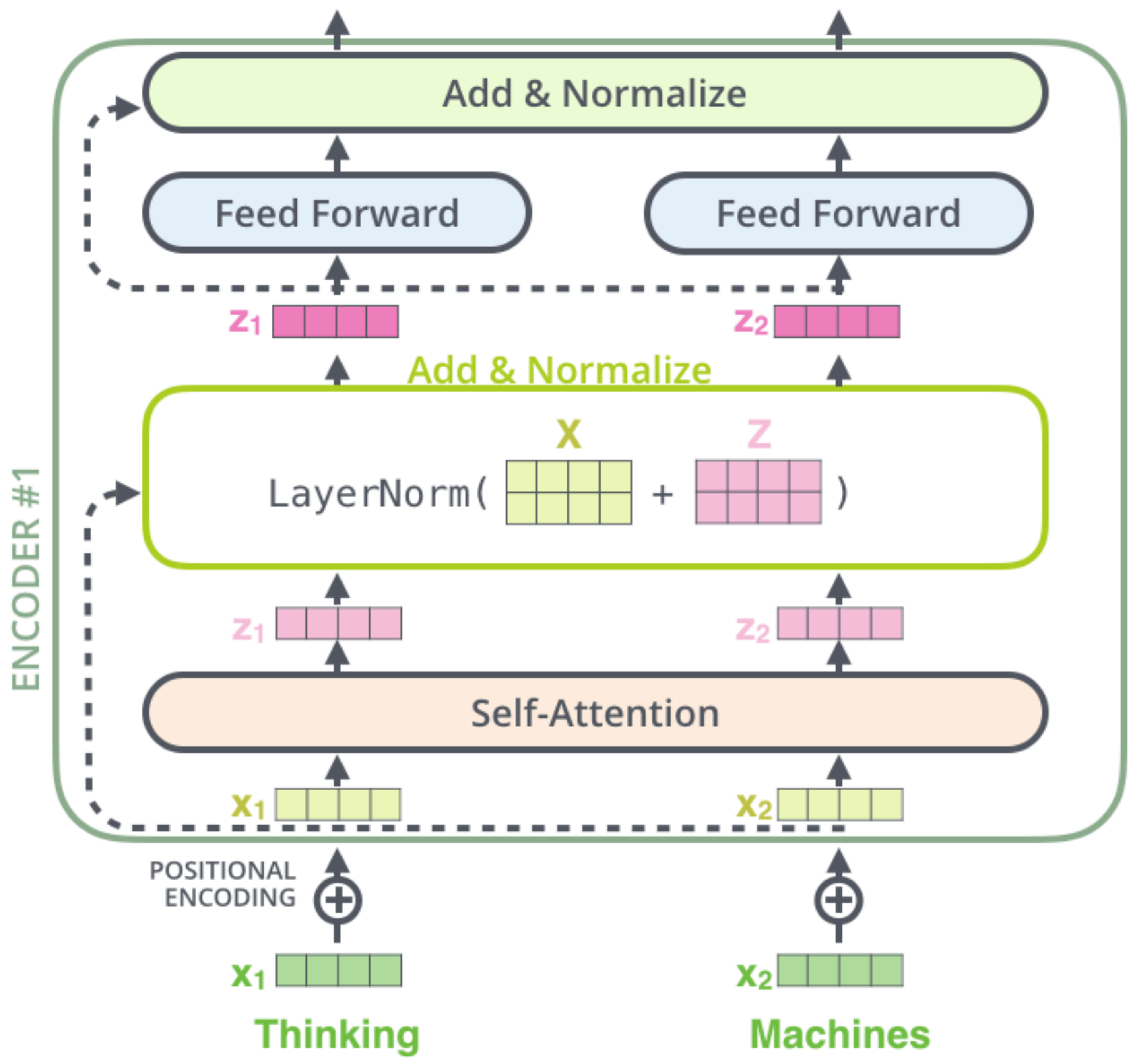
— Computes non-linear hierarchical features

Layer norm and **residuals**

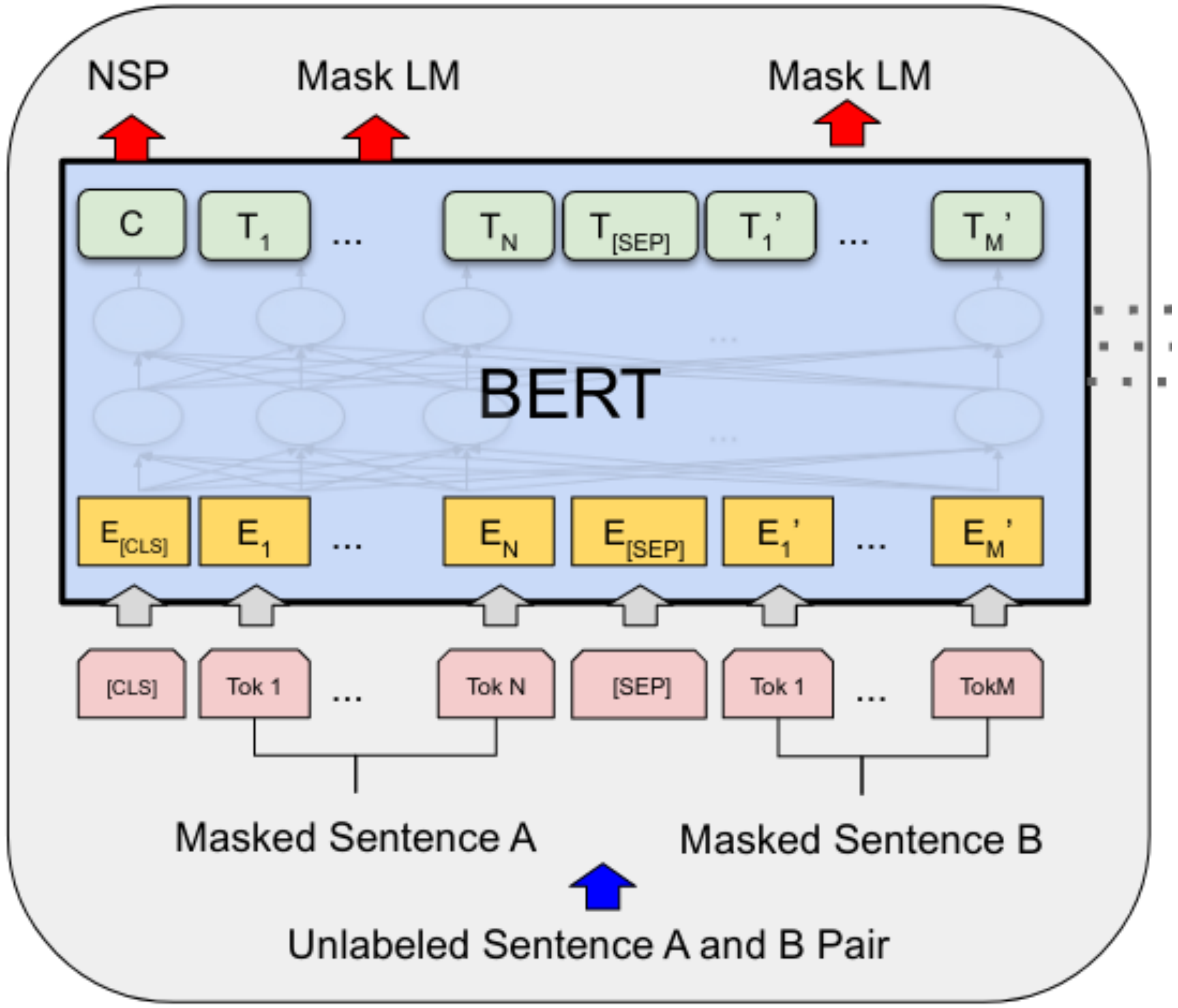
— Makes training deep neural network (e.g., 12 layers possible)

Positional Embeddings

— Allows model to learn relative positioning

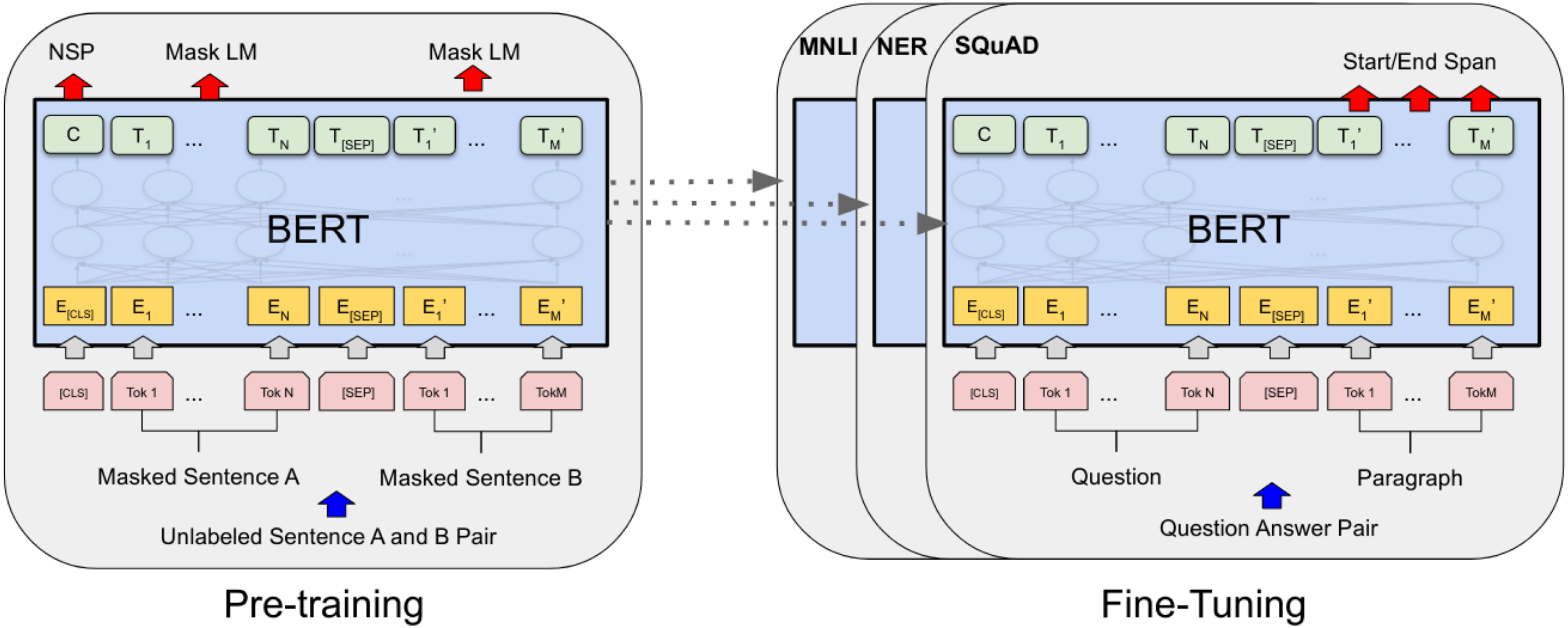


Applications: Masked Language Modeling (BERT)

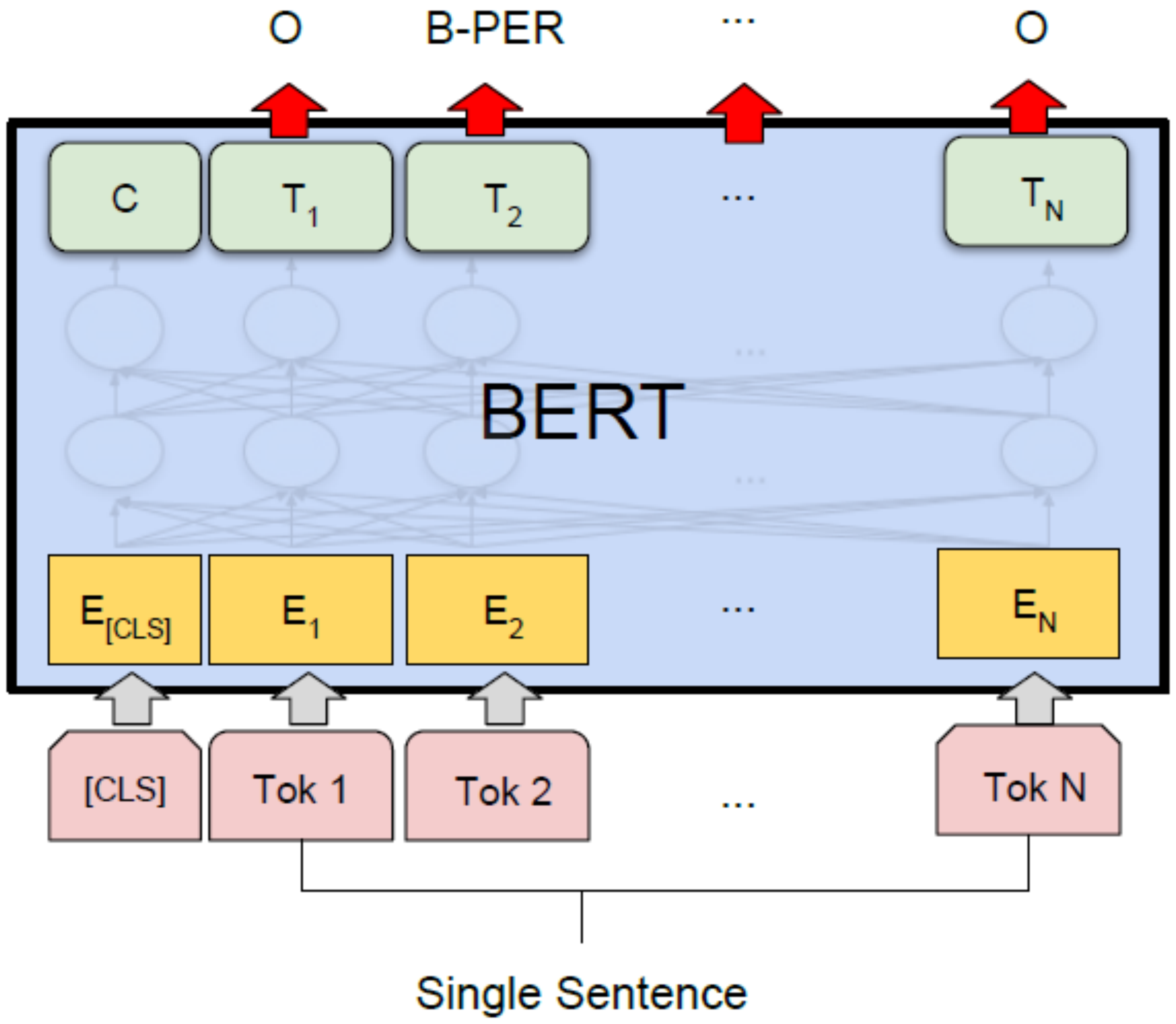


Pre-training

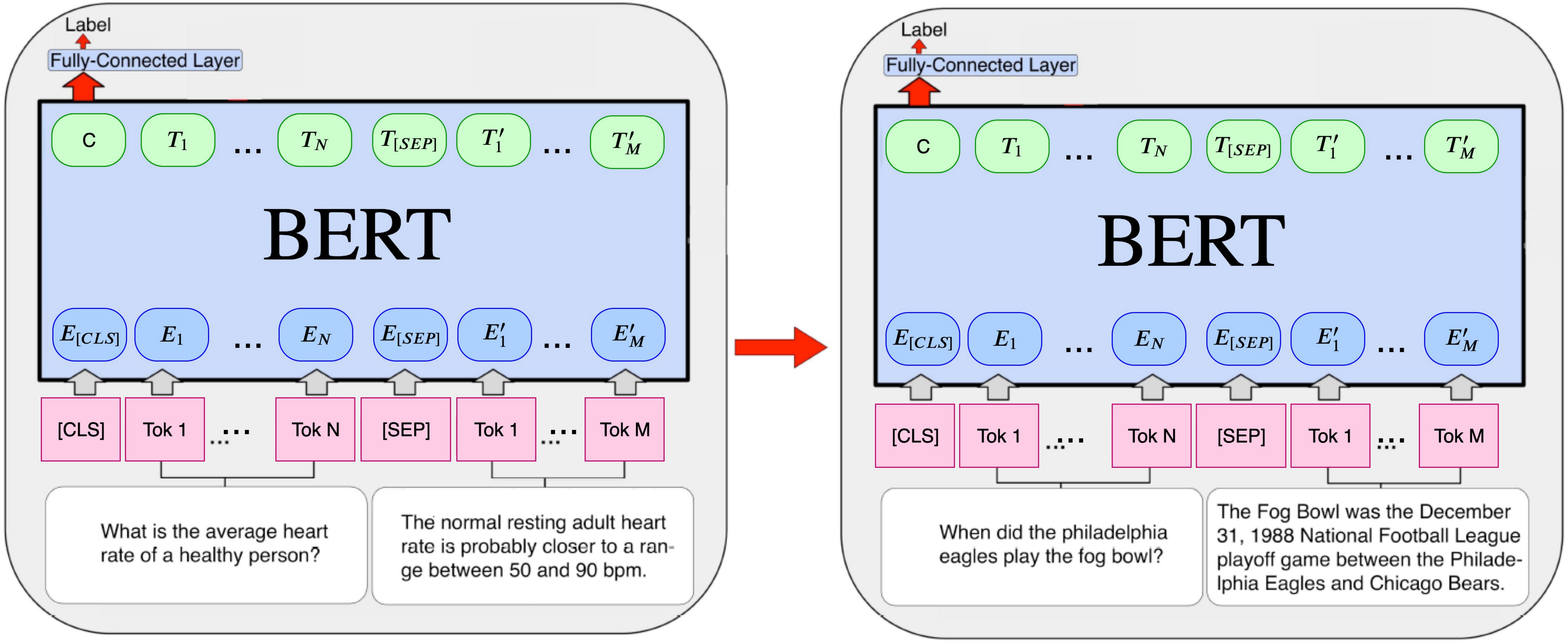
Applications: Masked Language Modeling (BERT)



Applications: Masked Language Modeling (BERT)



Applications: Masked Language Modeling (BERT)

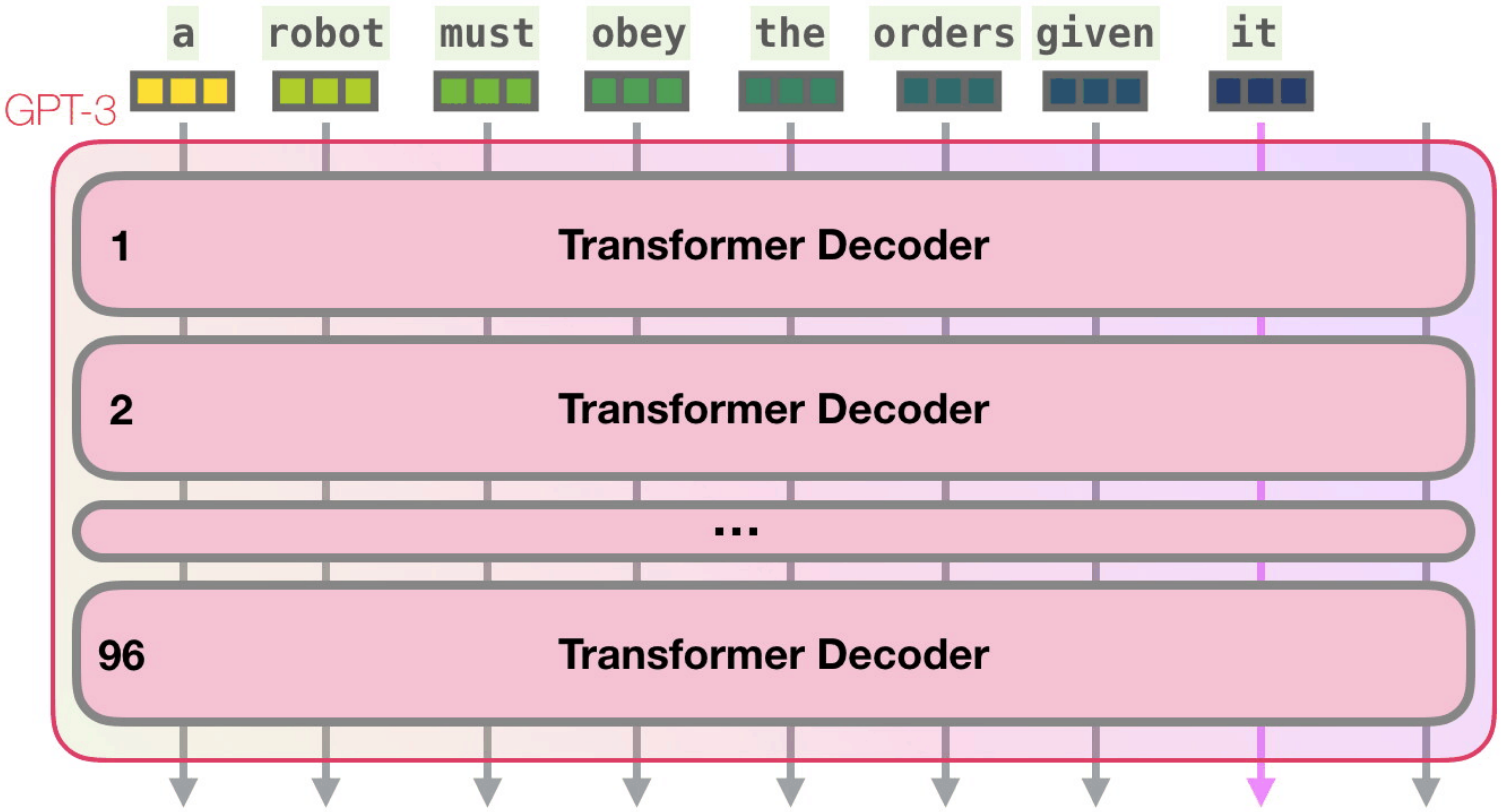


Transfer: ASNQ Dataset

Adapt: Target Dataset

Applications: Language Modeling (GPT3)

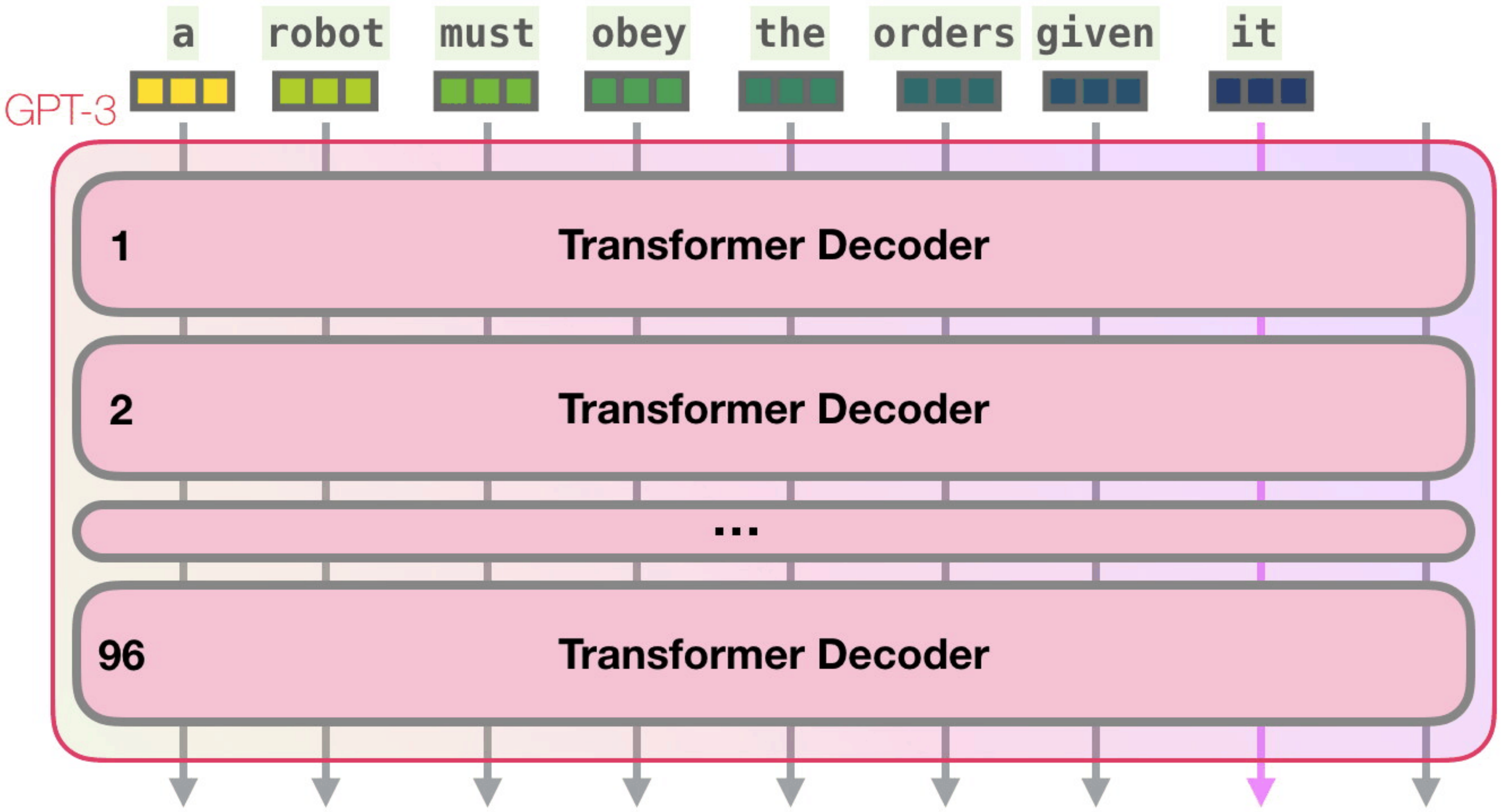
Task: Sentence completion (basically next token prediction)



Video source: https://jalammar.github.io/how-gpt3-works-visualizations-animations/

Applications: Language Modeling (GPT3)

Task: Sentence completion (basically next token prediction)



Video source: https://jalammar.github.io/how-gpt3-works-visualizations-animations/

Applications: Language Modeling (GPT3)

ELMo: 93M params, 2-layer biLSTM

BERT-base: 110M params, 12-layer Transformer

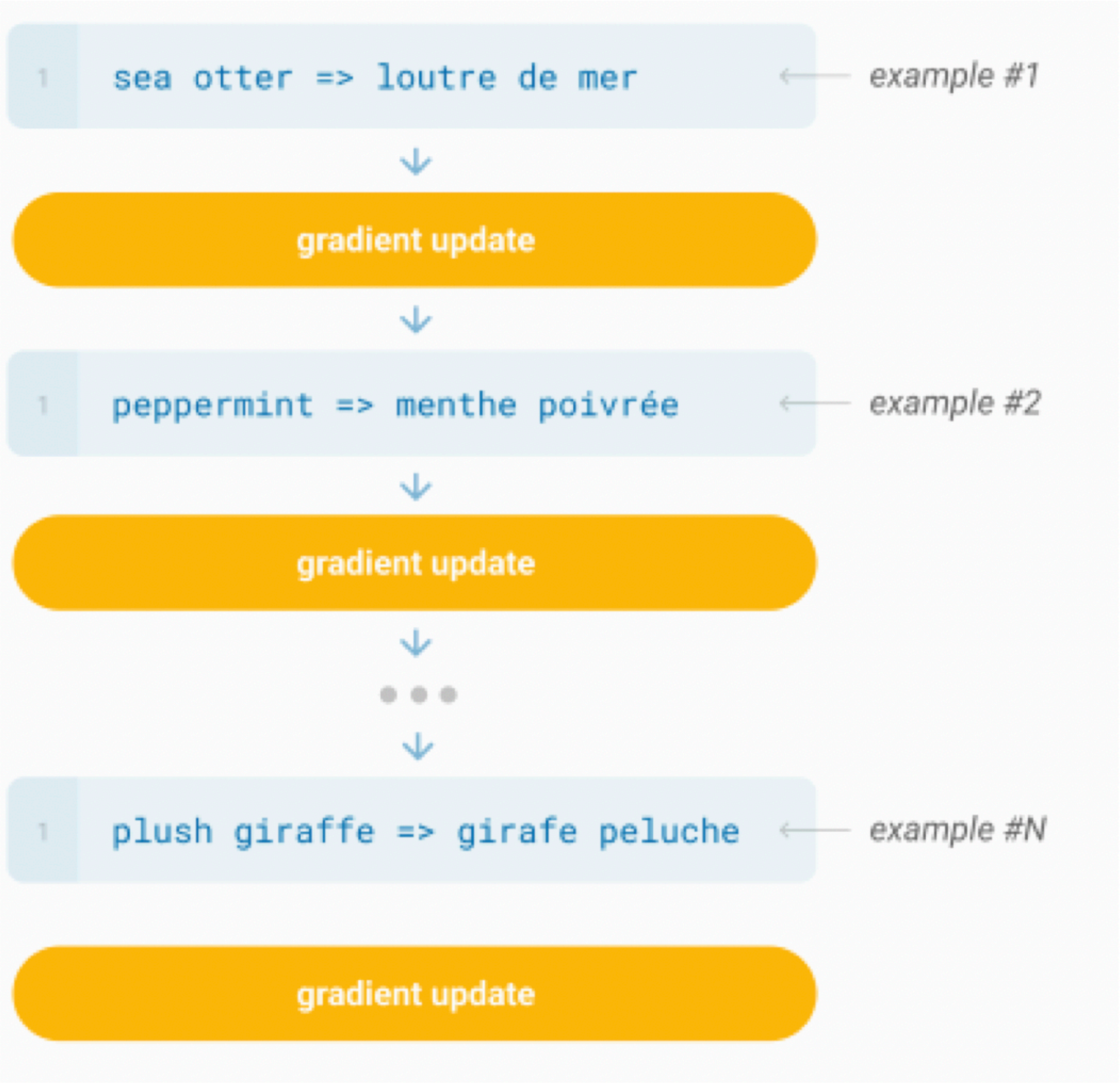
BERT-large: 340M params, 24-layer Transformer

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

BERT-like **Fine-tuning** (not used in GPT3)

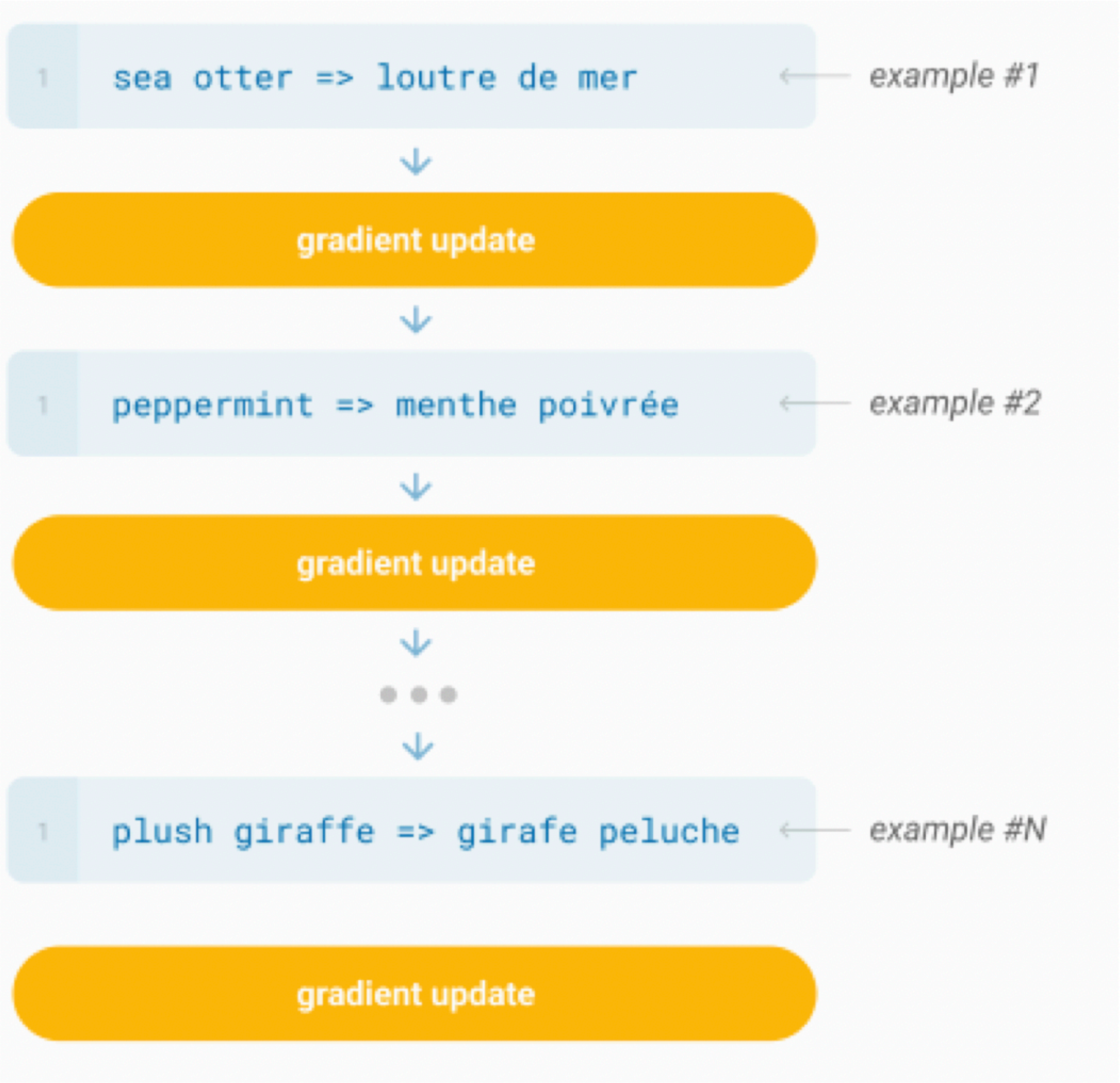
Downstream data **fine-tuning**



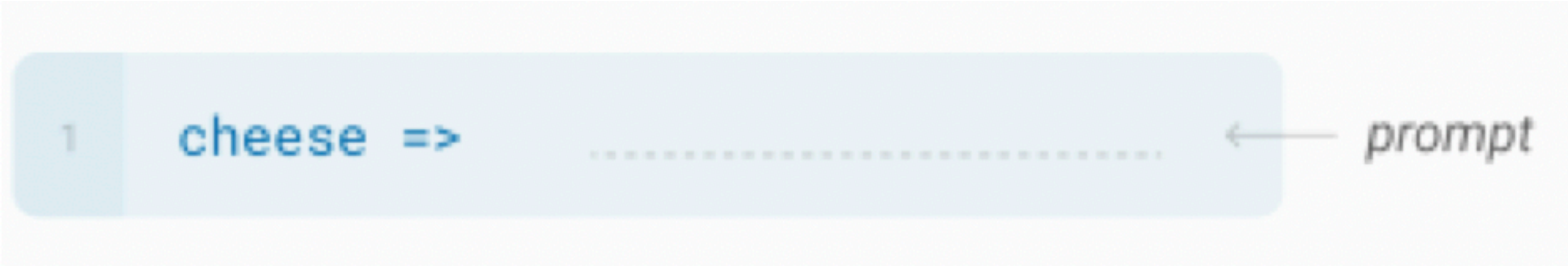
Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

BERT-like **Fine-tuning** (not used in GPT3)

Downstream data **fine-tuning**



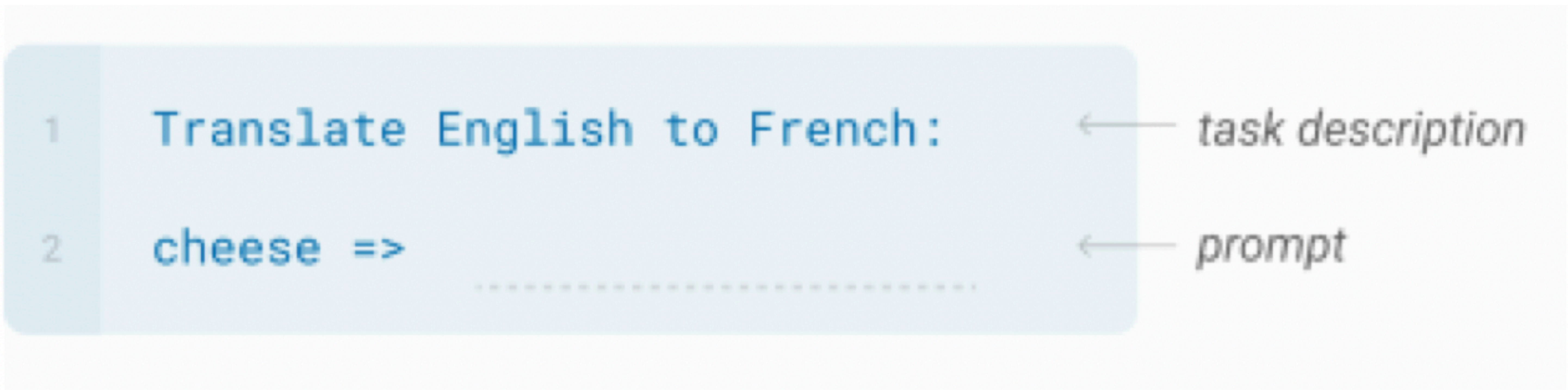
Downstream data **testing**



Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

GPT-3 Zero-shot inference

The model predicts the answer given only a natural language description of the task. **No gradient updates are performed.**



No fine-tuning! Literally just take a rpretrained GPT3 and give it prefix above

Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

GPT-3 One-shot inference

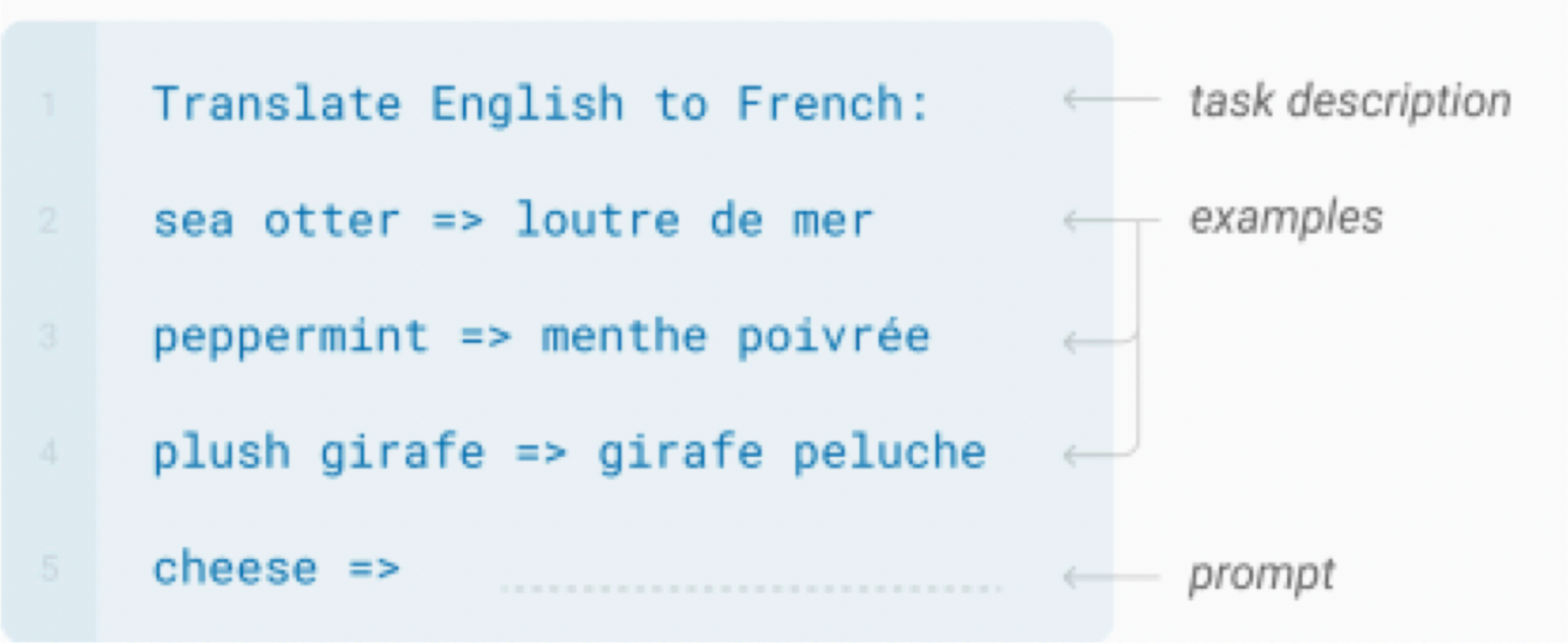
In addition to the task description, the model sees a single example of the task.
No gradient updates are performed.

No fine-tuning! Literally just take a rpretrained GPT3 and give it prefix above

Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

GPT-3 Few-shot inference

In addition to the task description, the model sees a few examples of the task.
No gradient updates are performed.



No fine-tuning! Literally just take a rpretrained GPT3 and give it prefix above

Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

How does Pre-training + Fine tuning Compare to GPT3

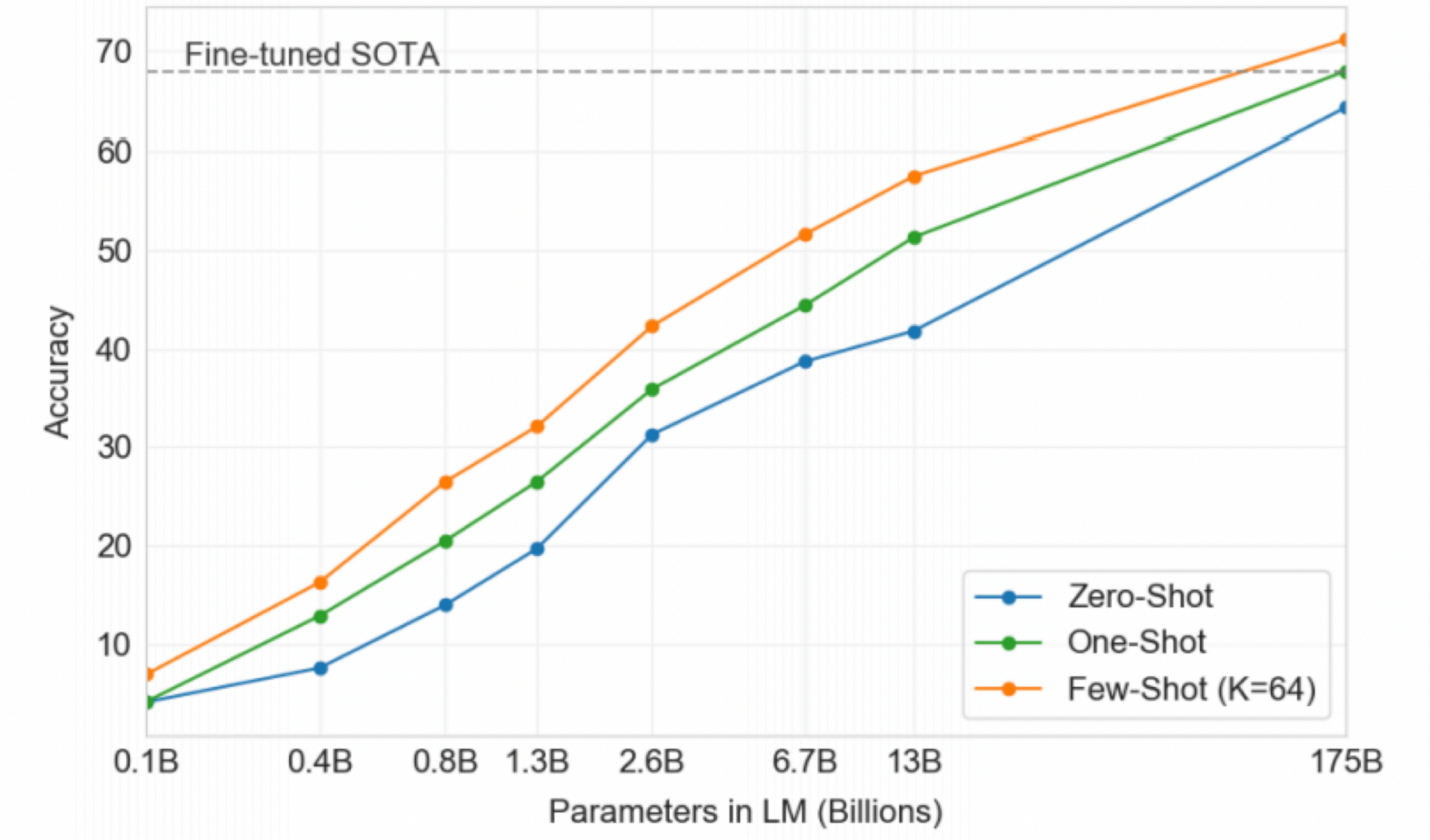
Task: Trivia QA

Question
Miami Beach in Florida borders which ocean?
What was the occupation of Lovely Rita according to the song by the Beatles
Who was Poopdeck Pappys most famous son?
The Nazi regime was Germany's Third Reich; which was the first Reich?
At which English racecourse did two horses collapse and die in the parade ring due to electrocution, in February 2011?
Which type of hat takes its name from an 1894 novel by George Du Maurier where the title character has the surname O'Ferrall ?
What was the Elephant Man's real name?

Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

How does Pre-training + Fine tuning Compare to GPT3

Task: Trivia QA



Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

How does Pre-training + Fine tuning Compare to GPT3

Task: Comprehension QA

Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**[d]	**90.0**[e]	**93.1**[e]
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

Performance is generally worse on “harder” datasets

Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

GPT3 for language translation

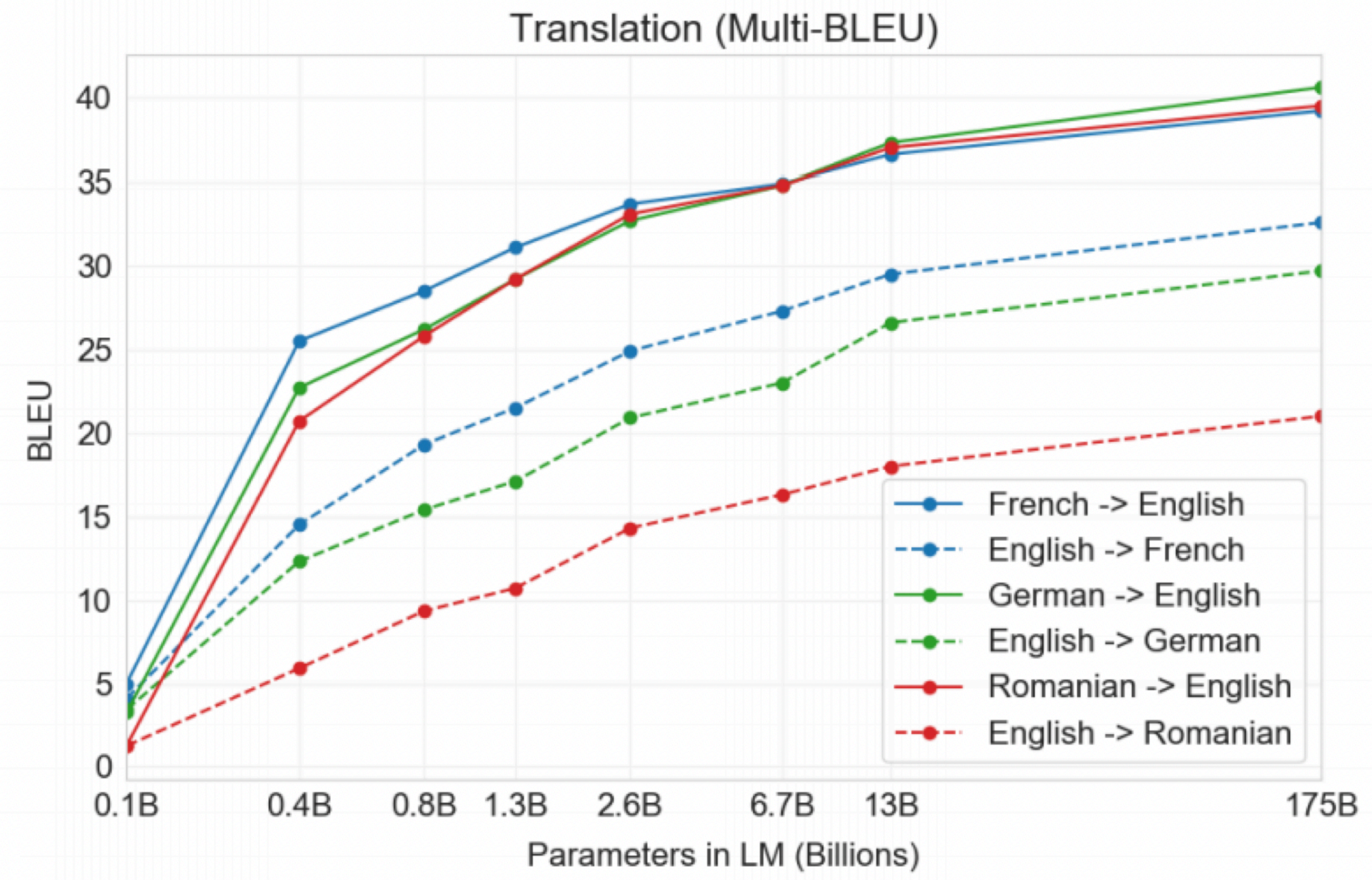
Task: Language translation (about 7% of training data is from languages other than English

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

GPT3 for language translation

Task: Language translation (about 7% of training data is from languages other than English



Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

Task: Note performance hasn't asymptoted

Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

GPT3 doing mathematics

- **2 digit addition (2D+)** – The model is asked to add two integers sampled uniformly from $[0, 100)$, phrased in the form of a question, e.g. "Q: What is 48 plus 76? A: 124."
- **2 digit subtraction (2D-)** – The model is asked to subtract two integers sampled uniformly from $[0, 100)$; the answer may be negative. Example: "Q: What is 34 minus 53? A: -19".
- **3 digit addition (3D+)** – Same as 2 digit addition, except numbers are uniformly sampled from $[0, 1000)$.
- **3 digit subtraction (3D-)** – Same as 2 digit subtraction, except numbers are uniformly sampled from $[0, 1000)$.
- **4 digit addition (4D+)** – Same as 3 digit addition, except uniformly sampled from $[0, 10000)$.
- **4 digit subtraction (4D-)** – Same as 3 digit subtraction, except uniformly sampled from $[0, 10000)$.
- **5 digit addition (5D+)** – Same as 3 digit addition, except uniformly sampled from $[0, 100000)$.
- **5 digit subtraction (5D-)** – Same as 3 digit subtraction, except uniformly sampled from $[0, 100000)$.
- **2 digit multiplication (2Dx)** – The model is asked to multiply two integers sampled uniformly from $[0, 100)$, e.g. "Q: What is 24 times 42? A: 1008".
- **One-digit composite (1DC)** – The model is asked to perform a composite operation on three 1 digit numbers, with parentheses around the last two. For example, "Q: What is 6+(4*8)? A: 38". The three 1 digit numbers are selected uniformly on $[0, 10)$ and the operations are selected uniformly from {+,-,*}.

Setting	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3 Zero-shot	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3 One-shot	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3 Few-shot	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3

Slide source: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/11-gpt3.pdf

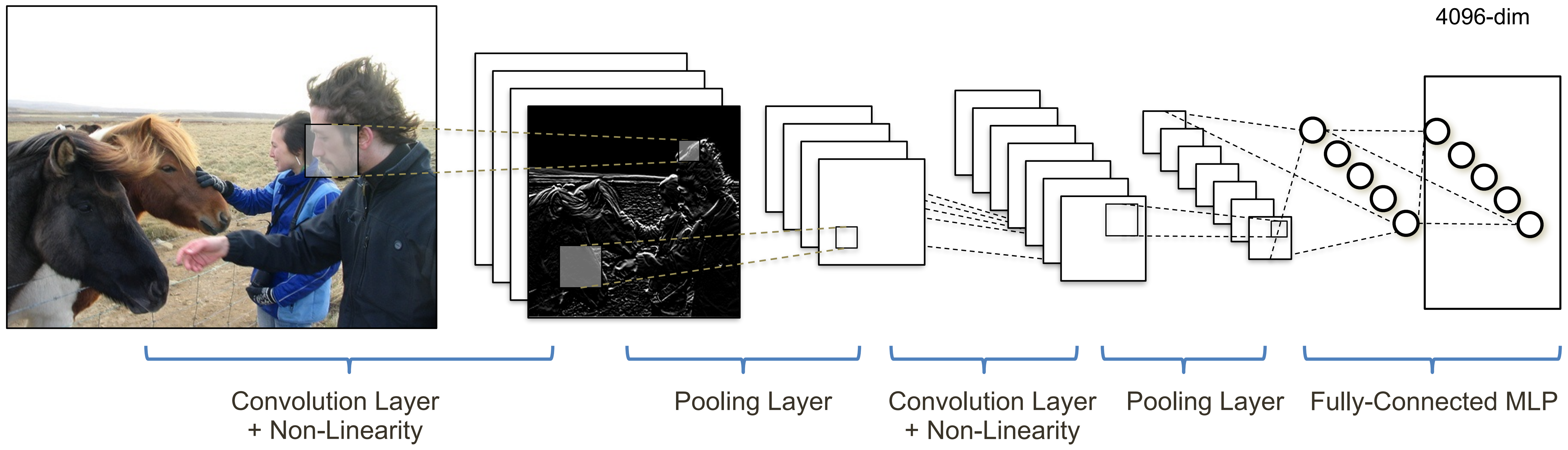
Let us look at some multi-modal architectures now that use RNNs

Applications: Neural Image Captioning

* slide from Dhruv Batra

Applications: Neural Image Captioning

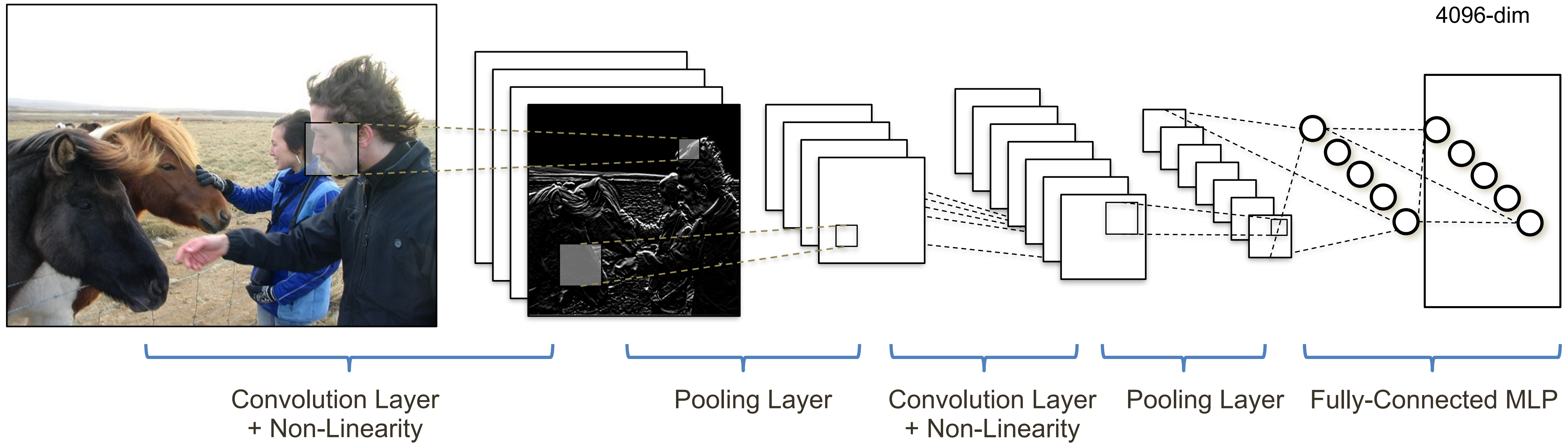
Image Embedding (VGGNet)



* slide from Dhruv Batra

Applications: Neural Image Captioning

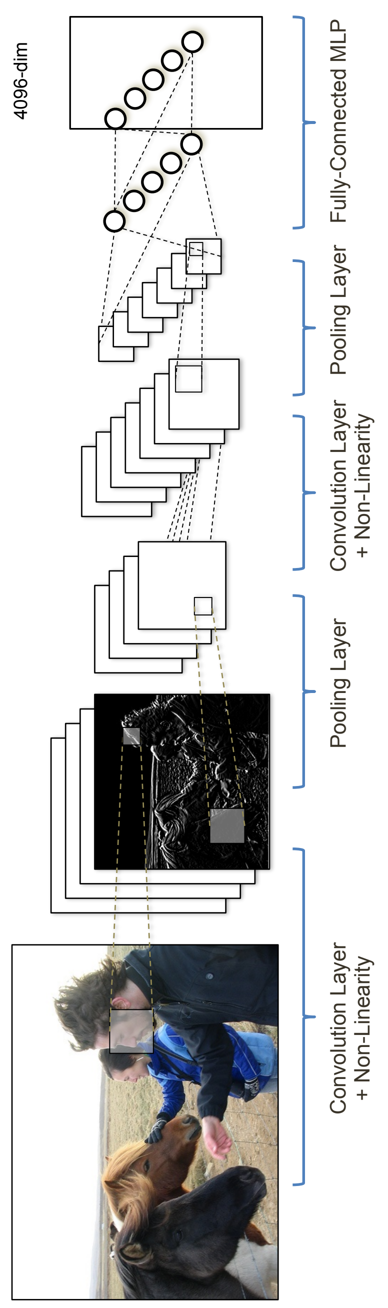
Image Embedding (VGGNet)



* slide from Dhruv Batra

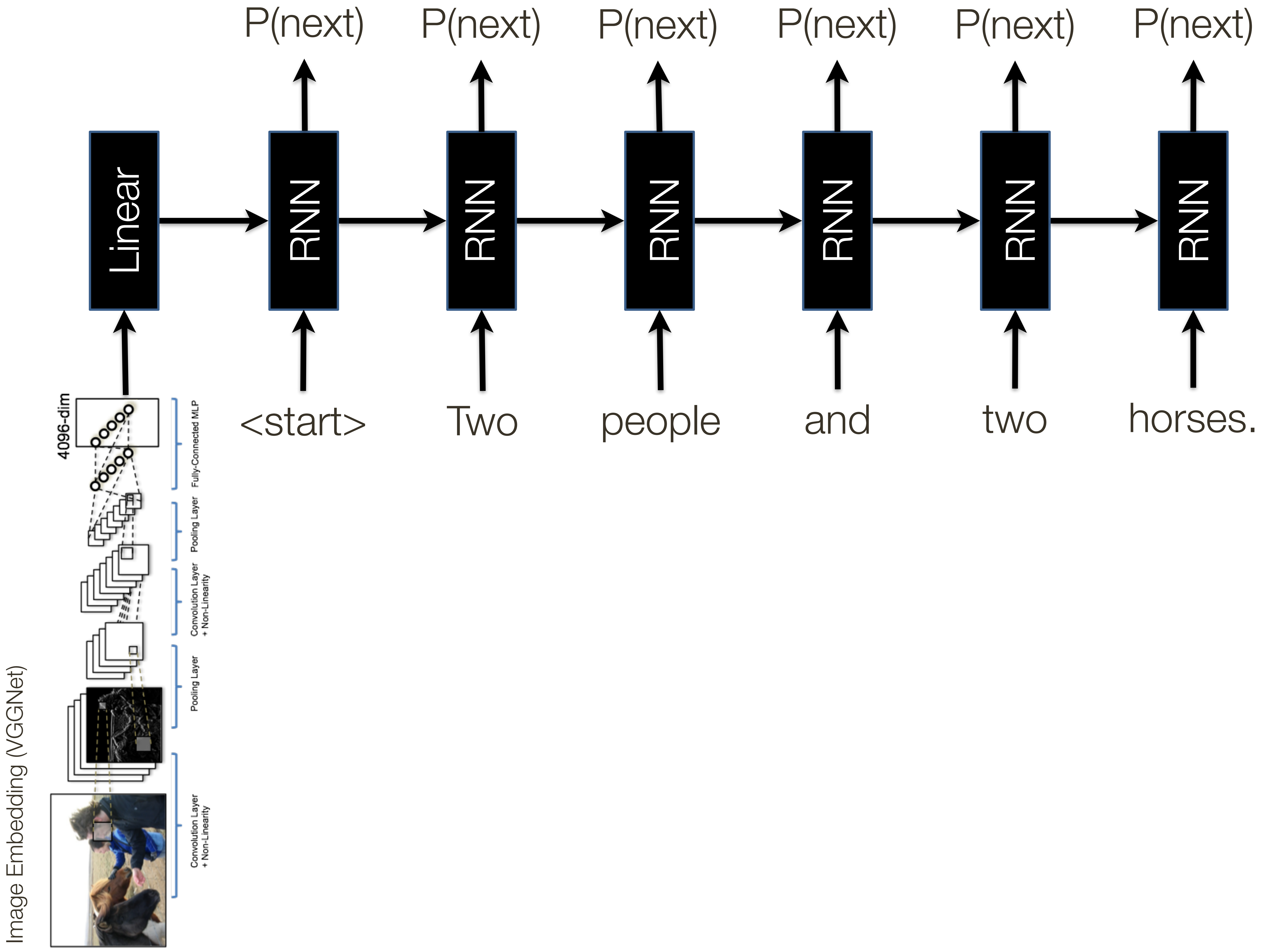
Applications: Neural Image Captioning

Image Embedding (VGGNet)



* slide from Dhruv Batra

Applications: Neural Image Captioning



* slide from Dhruv Batra

Applications: Neural Image Captioning

Good results

A cat sitting on a suitcase on the floor

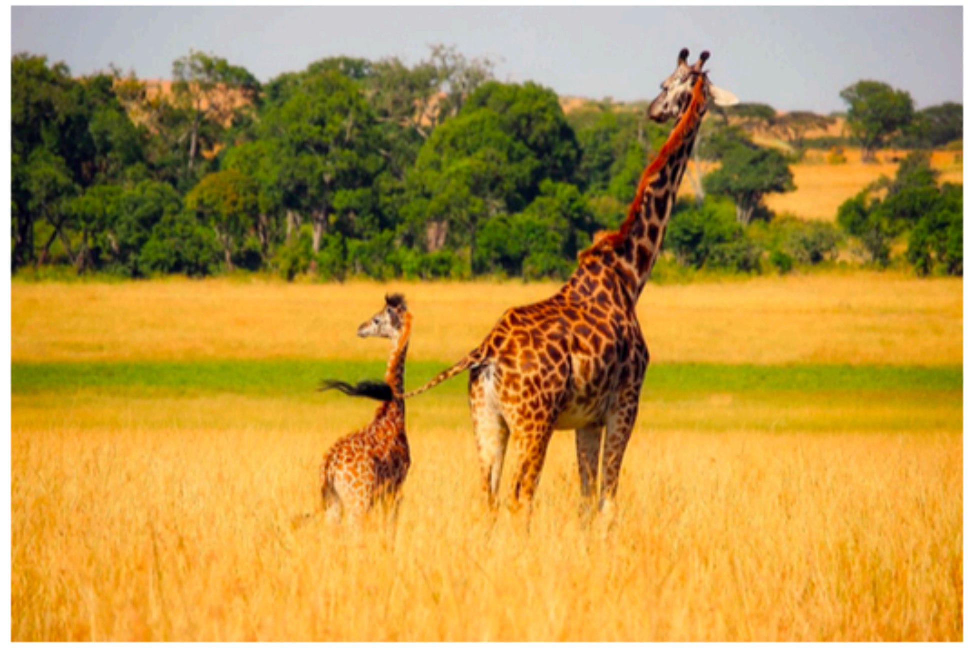
A cat is sitting on a tree branch

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

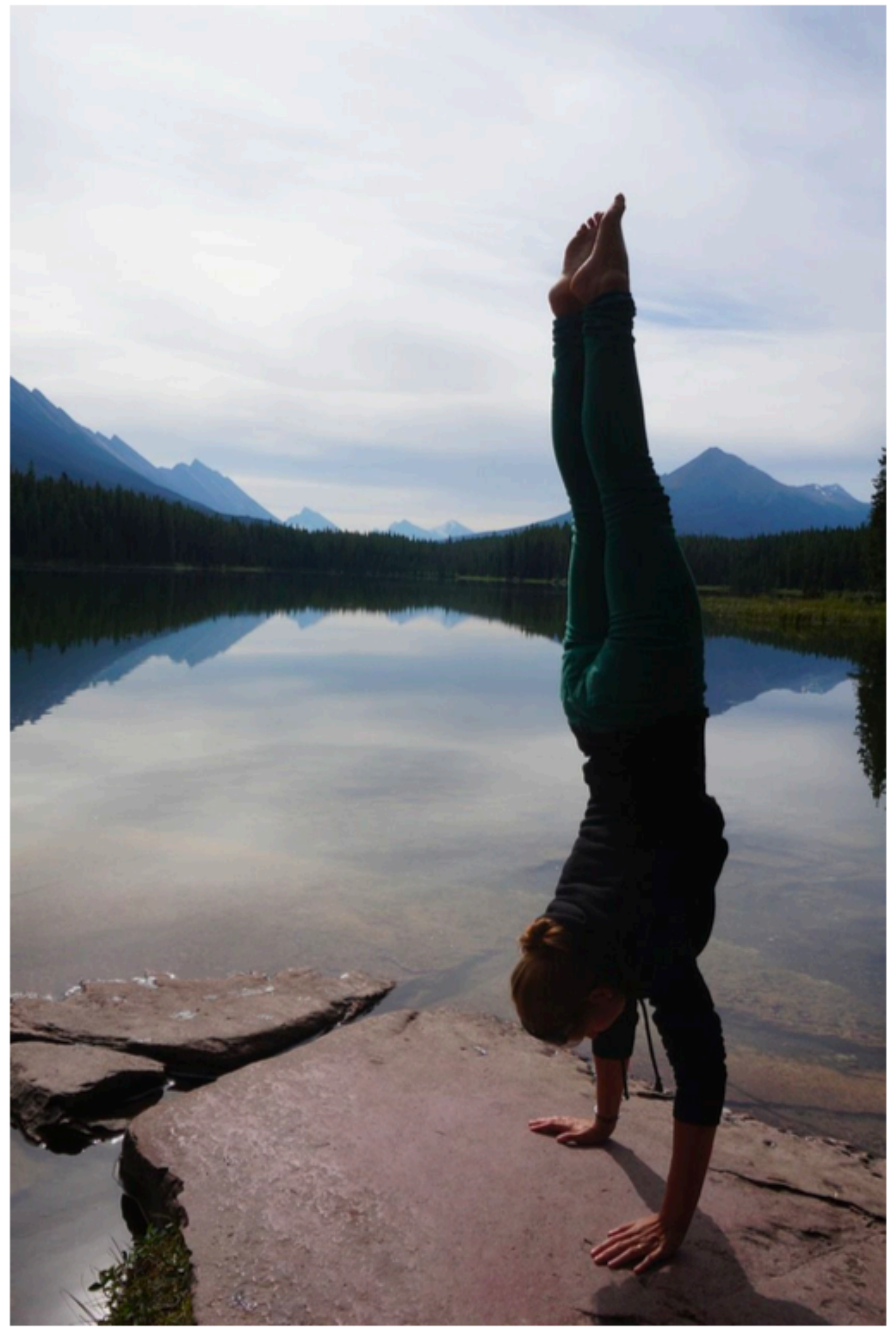
* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

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

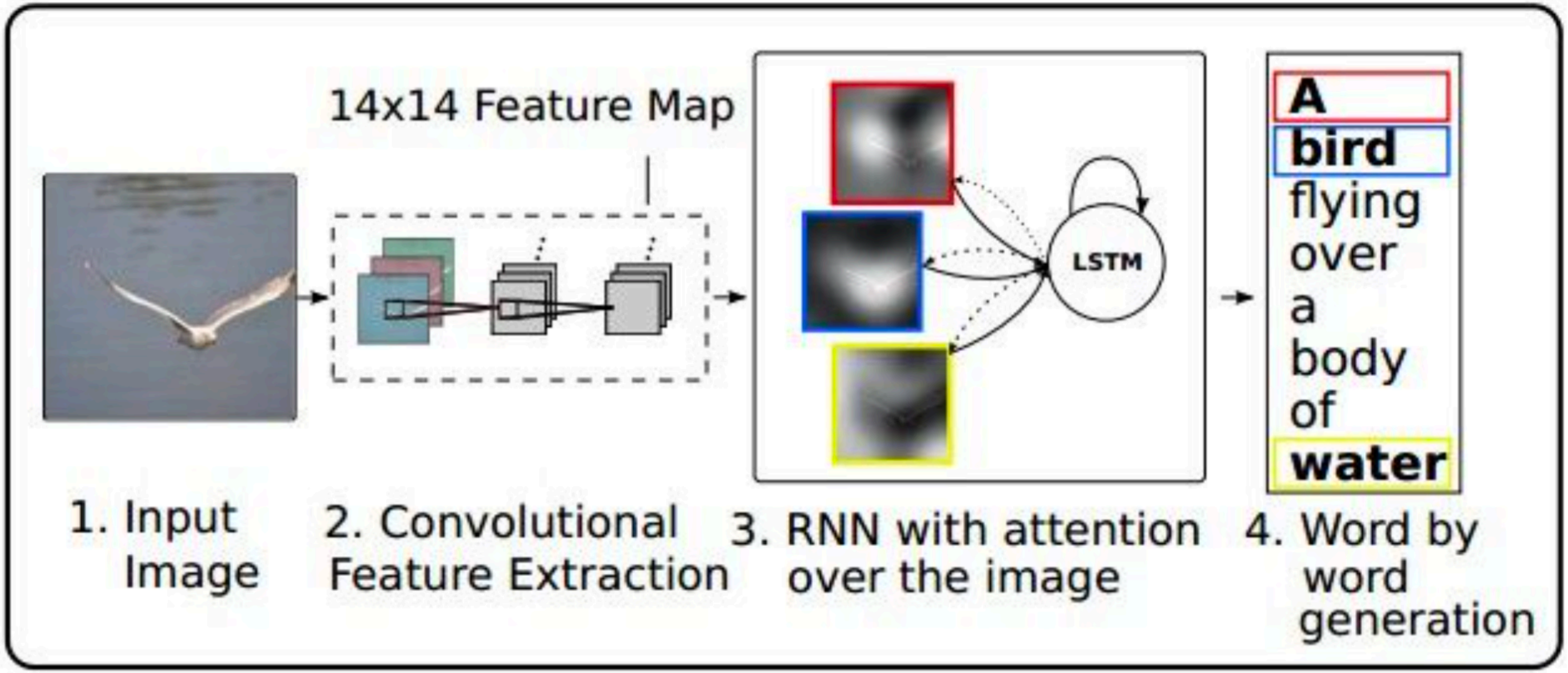
A bird is perched on a tree branch

A man in a baseball uniform throwing a ball

Applications: Image Captioning with Attention

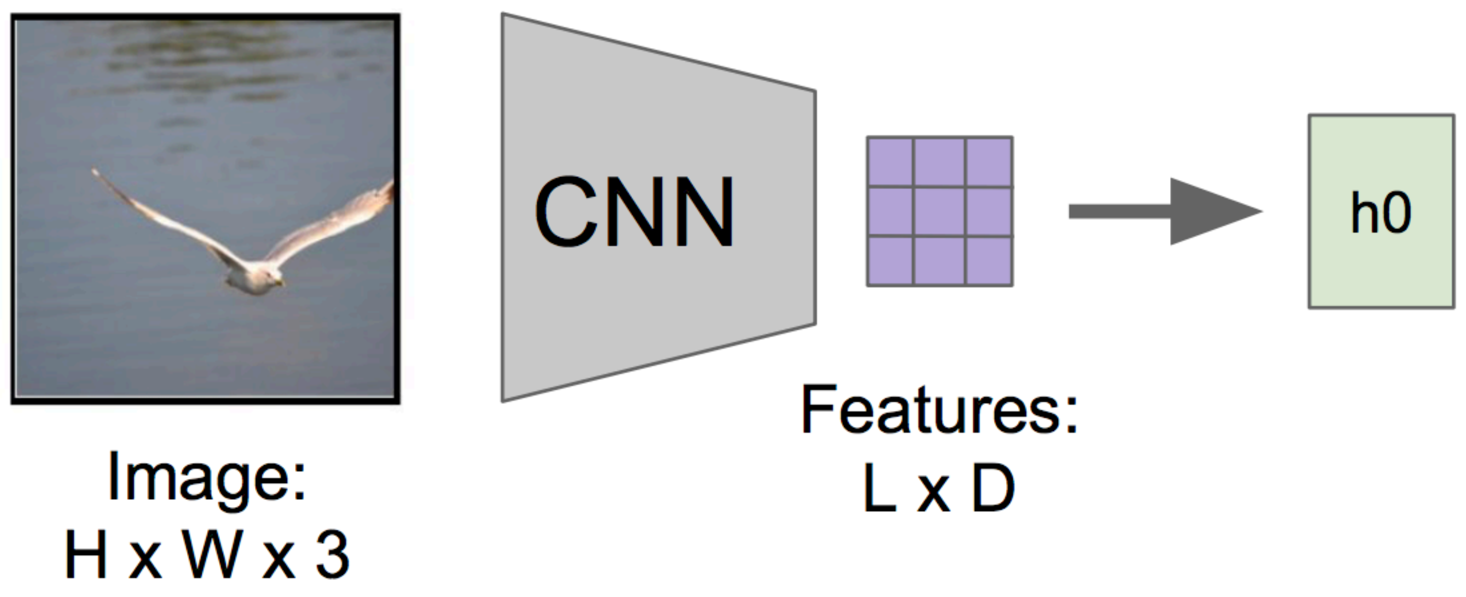
[Xu et al., ICML 2015]

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



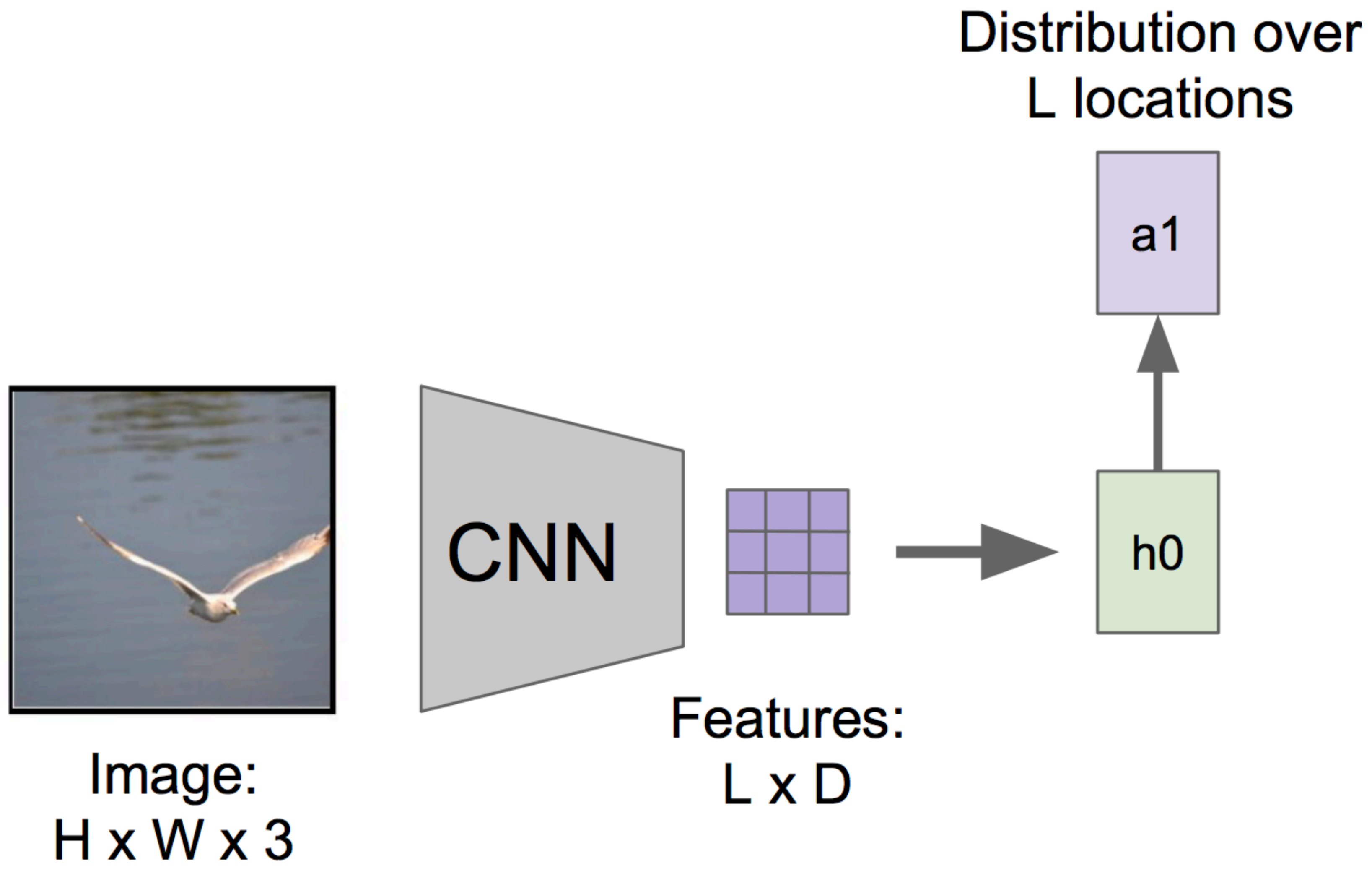
Applications: Image Captioning with Attention

[Xu et al., ICML 2015]



Applications: Image Captioning with Attention

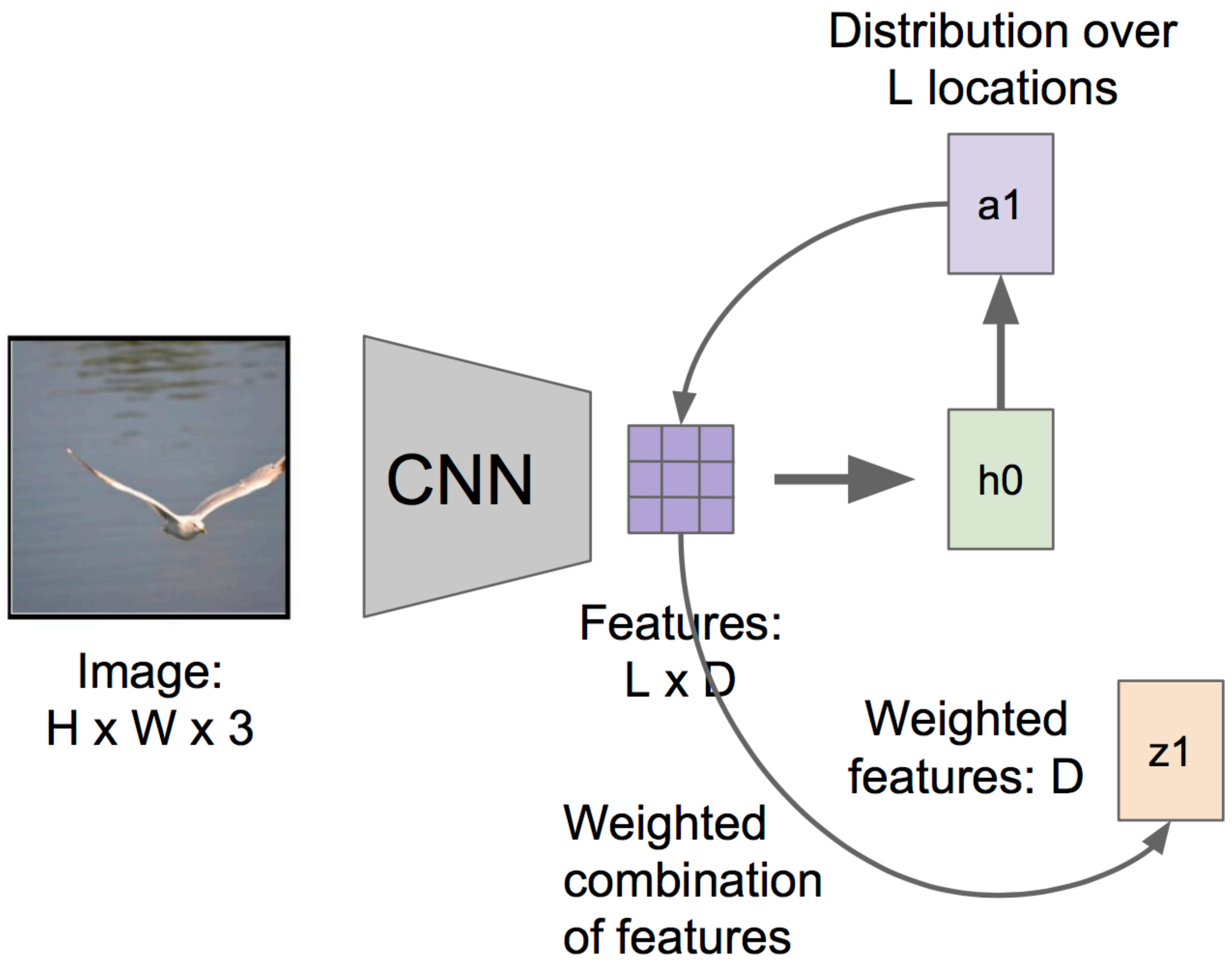
[Xu et al., ICML 2015]



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

Applications: Image Captioning with Attention

[Xu et al., ICML 2015]

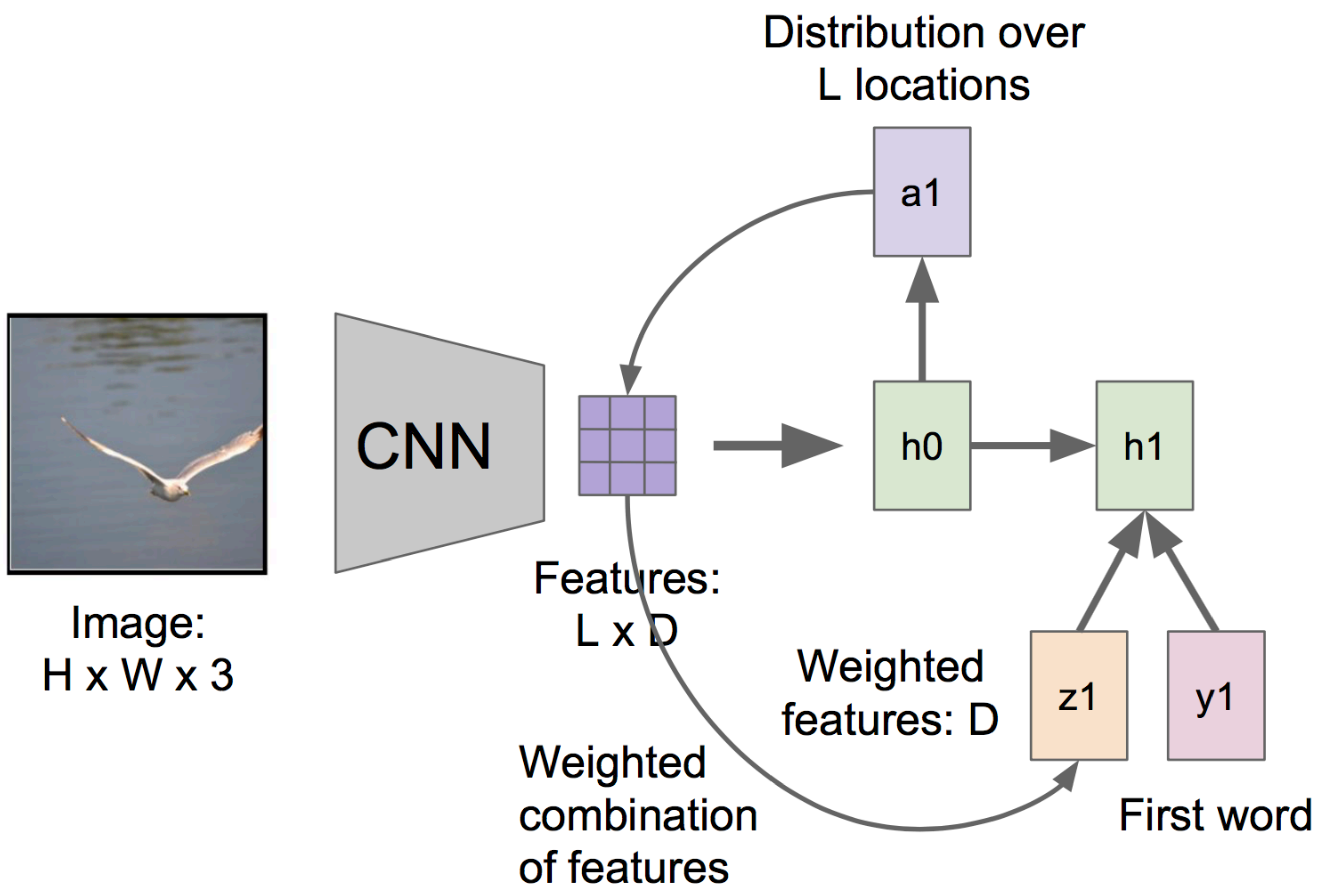


$$z = \sum_{i=1}^{L} p_i v_i$$

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

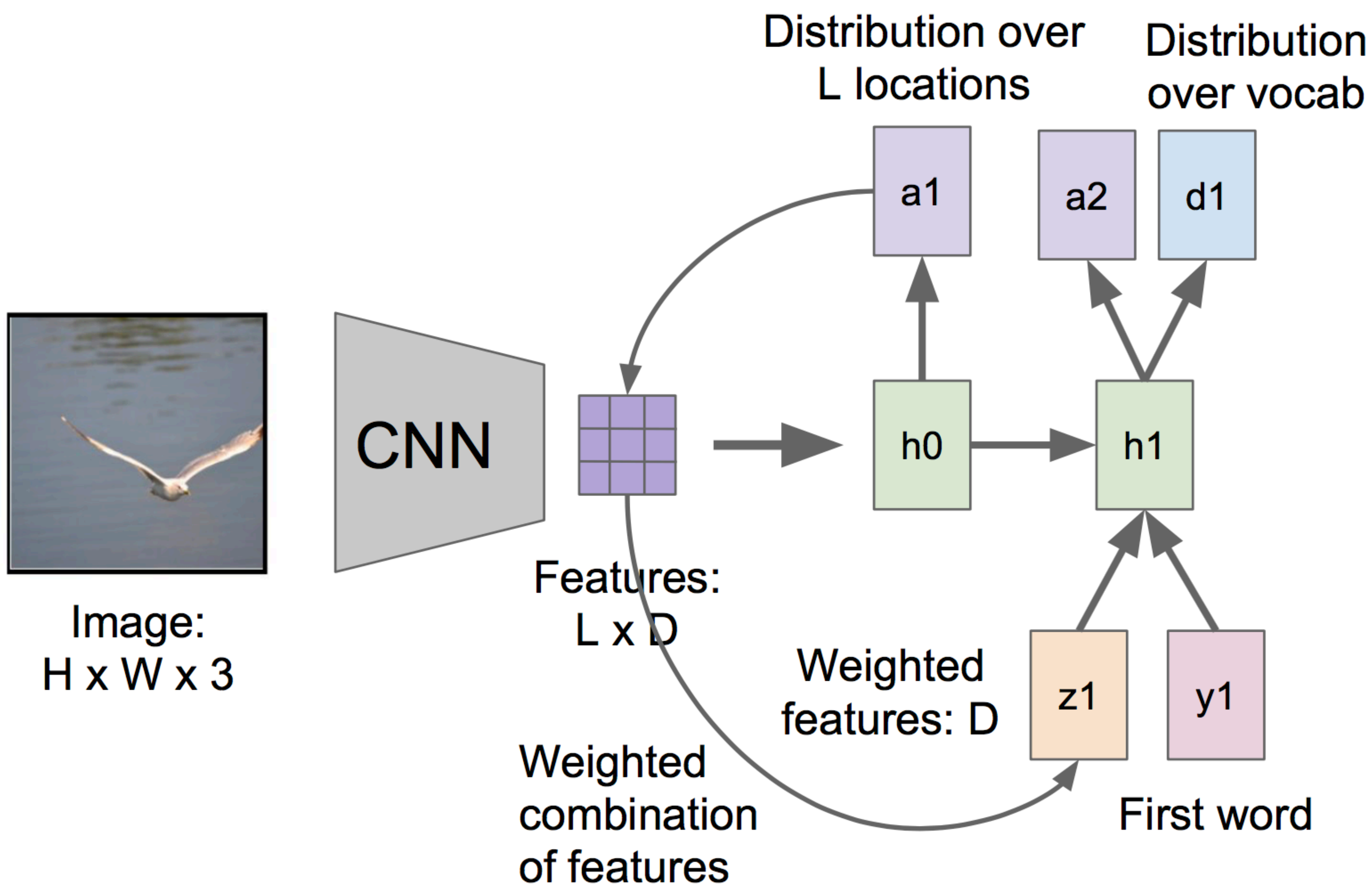
Applications: Image Captioning with Attention

[Xu et al., ICML 2015]



Applications: Image Captioning with Attention

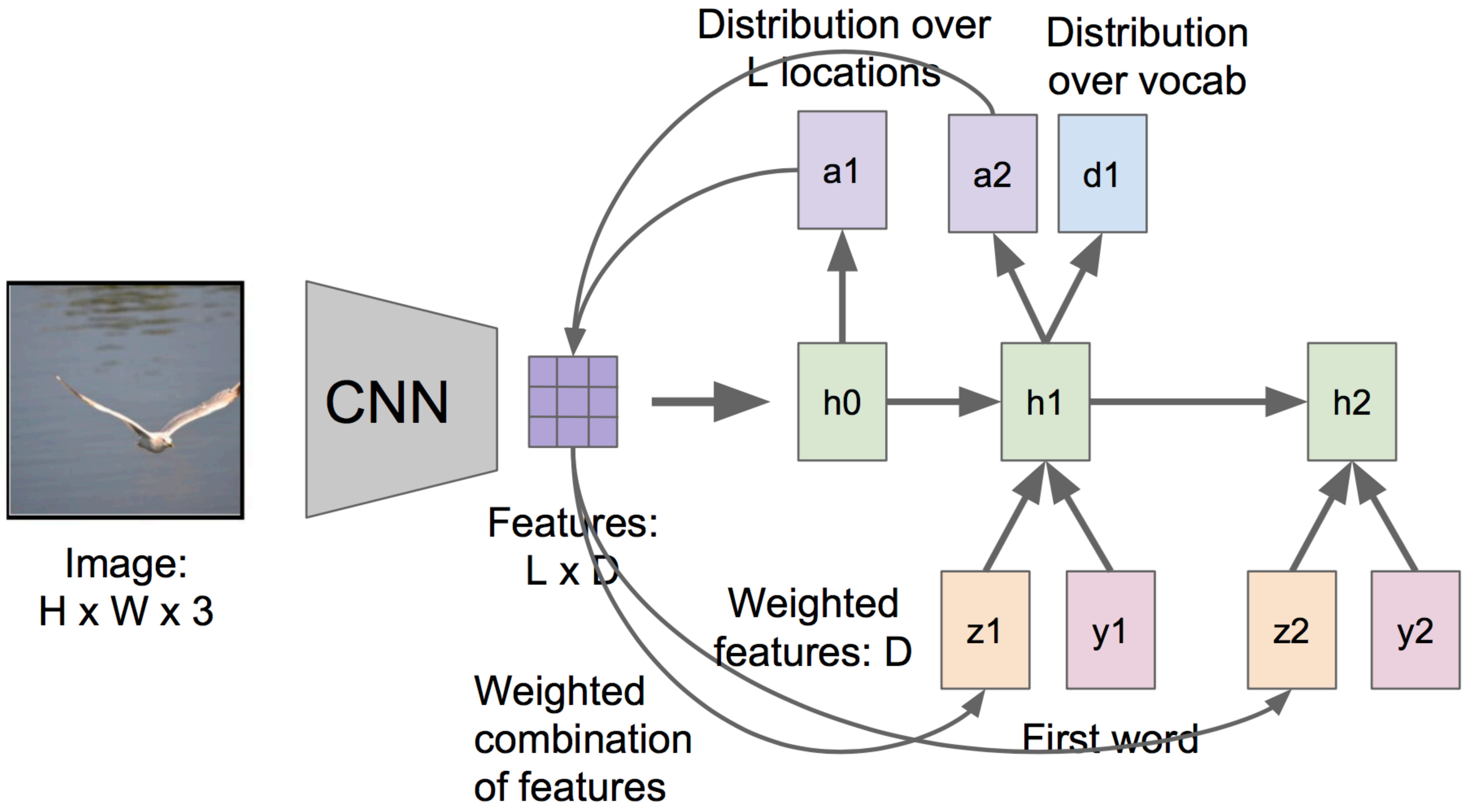
[Xu et al., ICML 2015]



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

Applications: Image Captioning with Attention

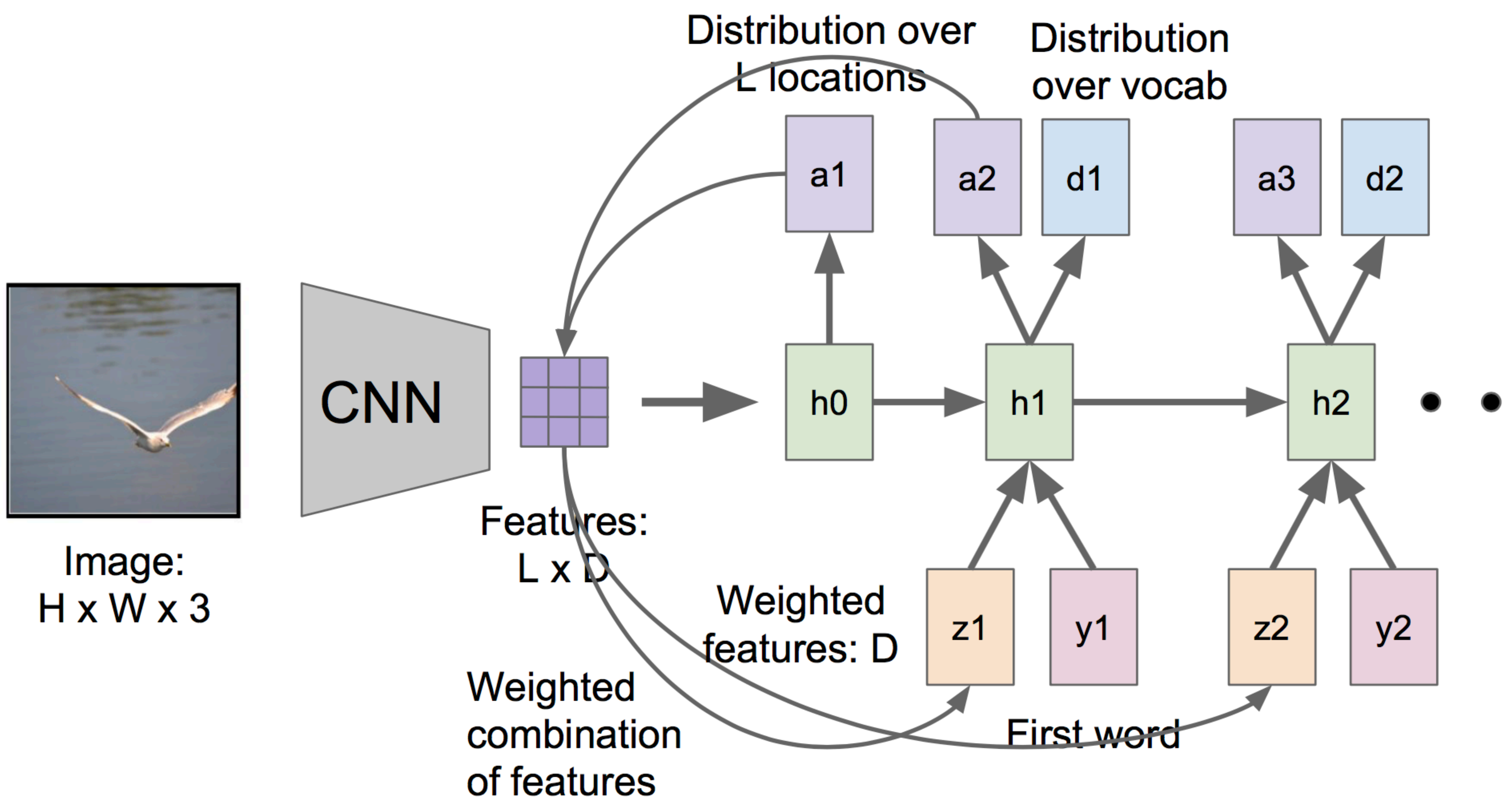
[Xu et al., ICML 2015]



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

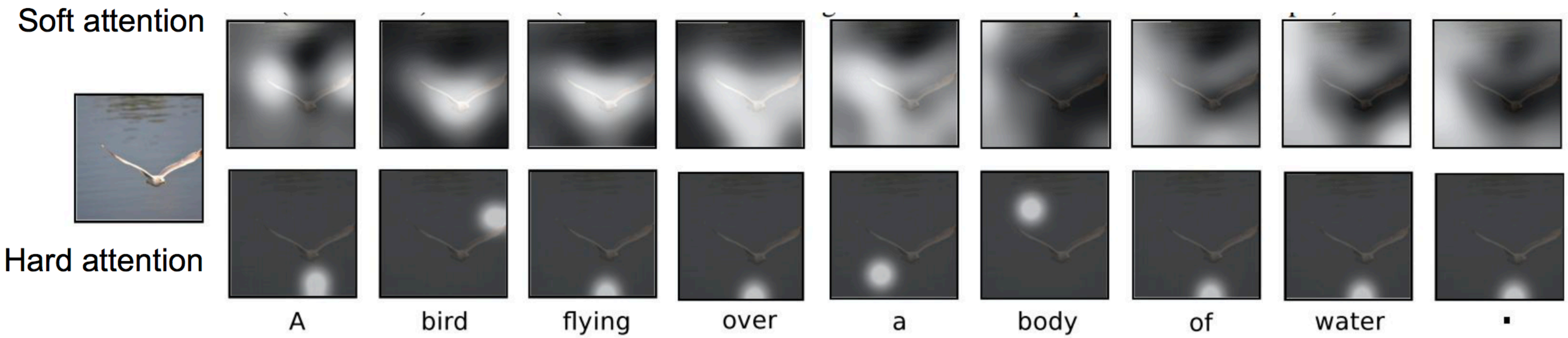
Applications: Image Captioning with Attention

[Xu et al., ICML 2015]



Applications: Image Captioning with Attention

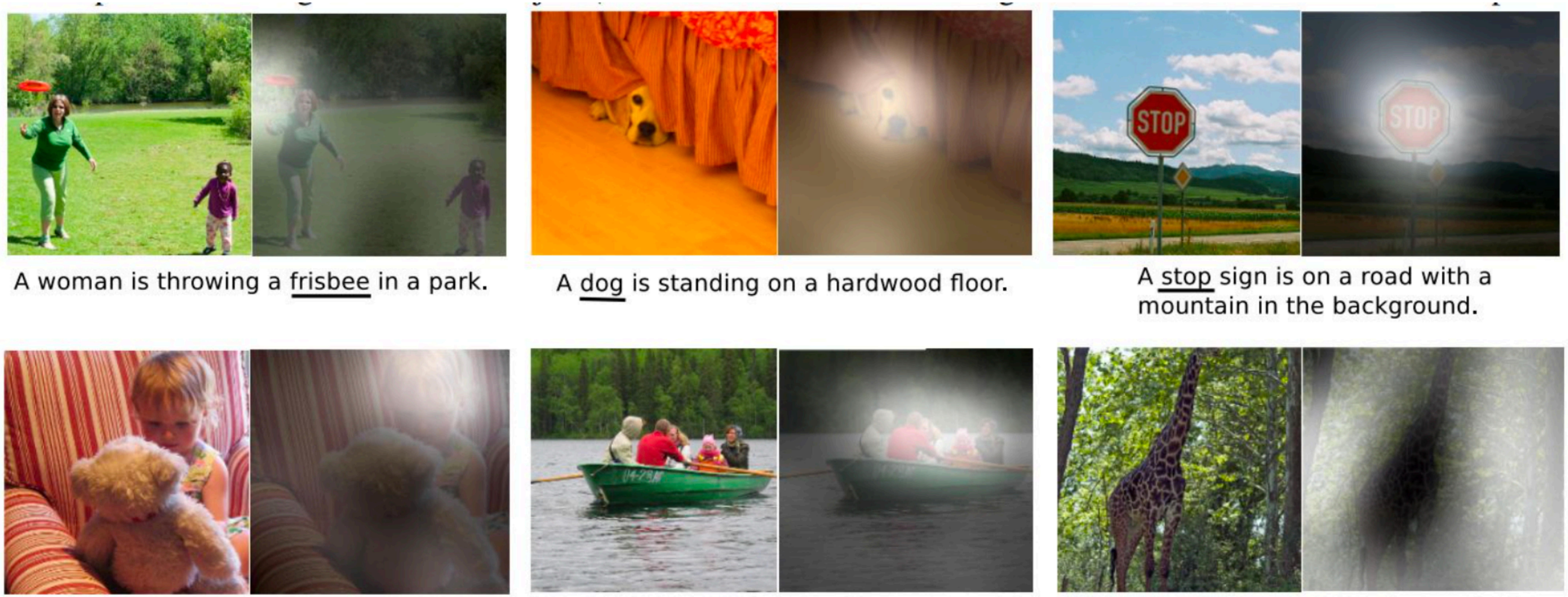
[Xu et al., ICML 2015]



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.

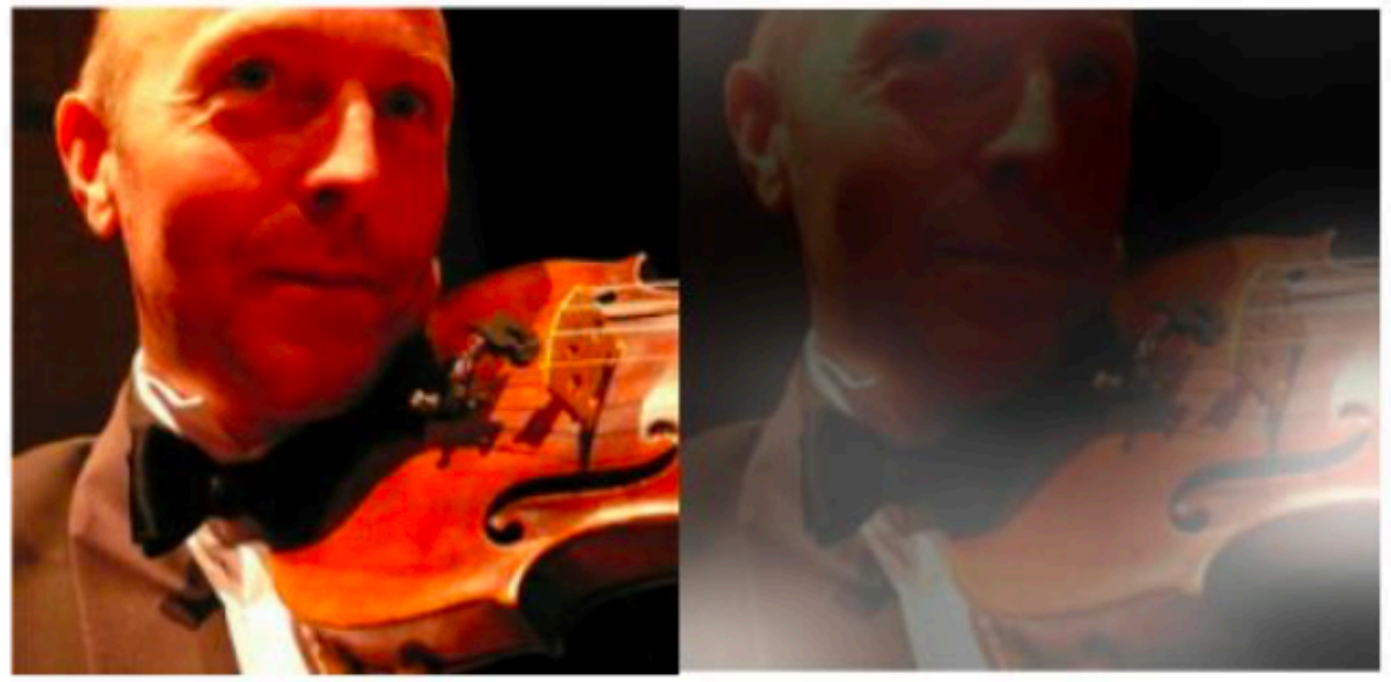
Applications: Image Captioning with Attention

[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

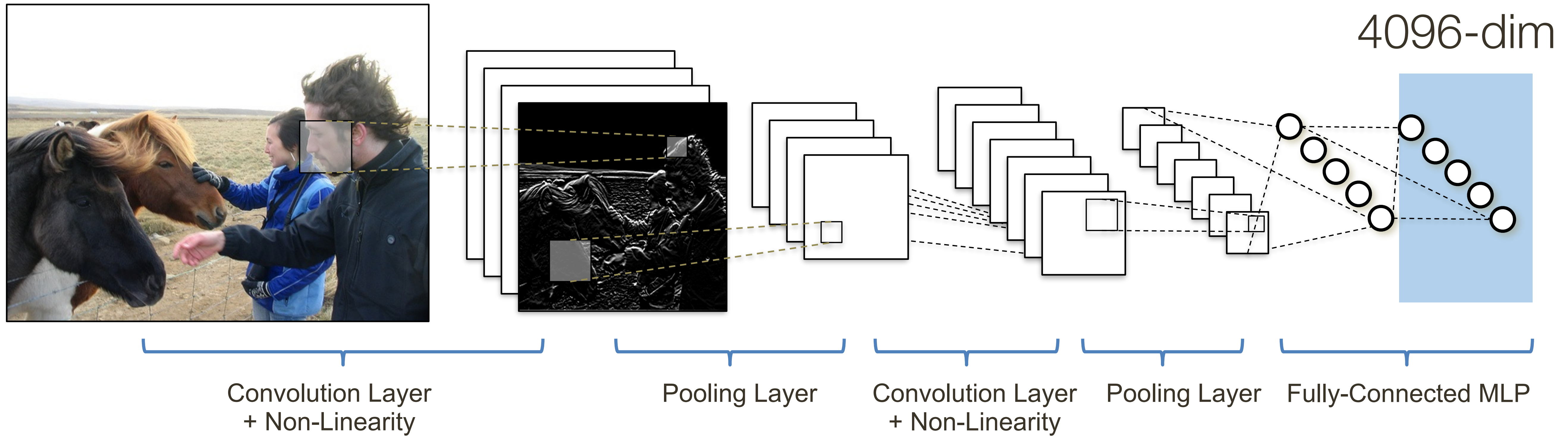
Question

"How many horses are in this image?"

* slide from Dhruv Batra

Applications: Typical Visual Question Answering (VQA)

Image Embedding (VGGNet)



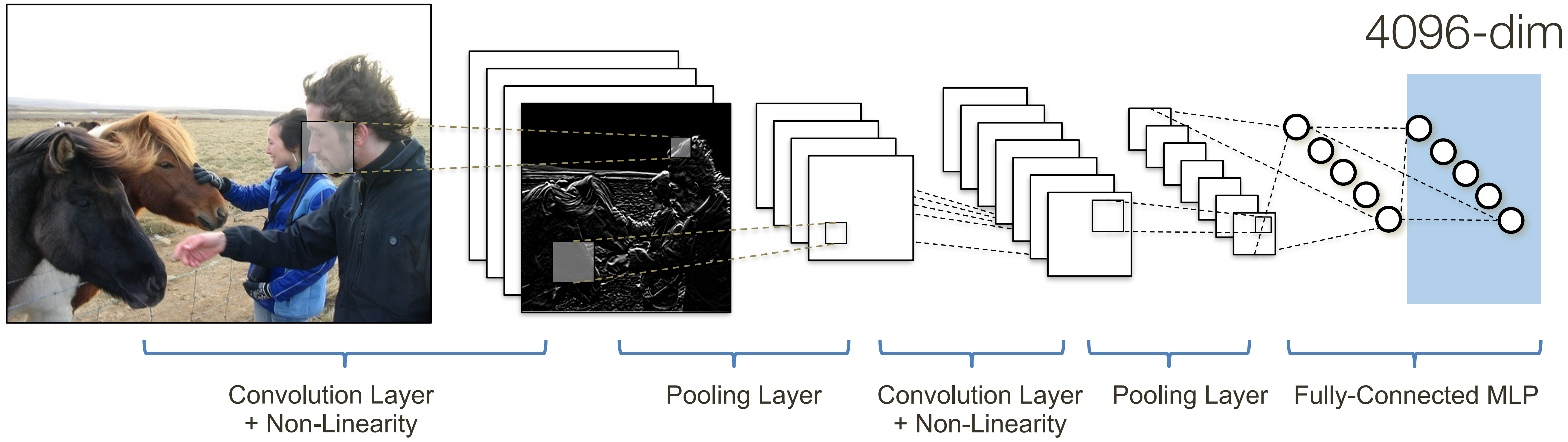
Question

"How many horses are in this image?"

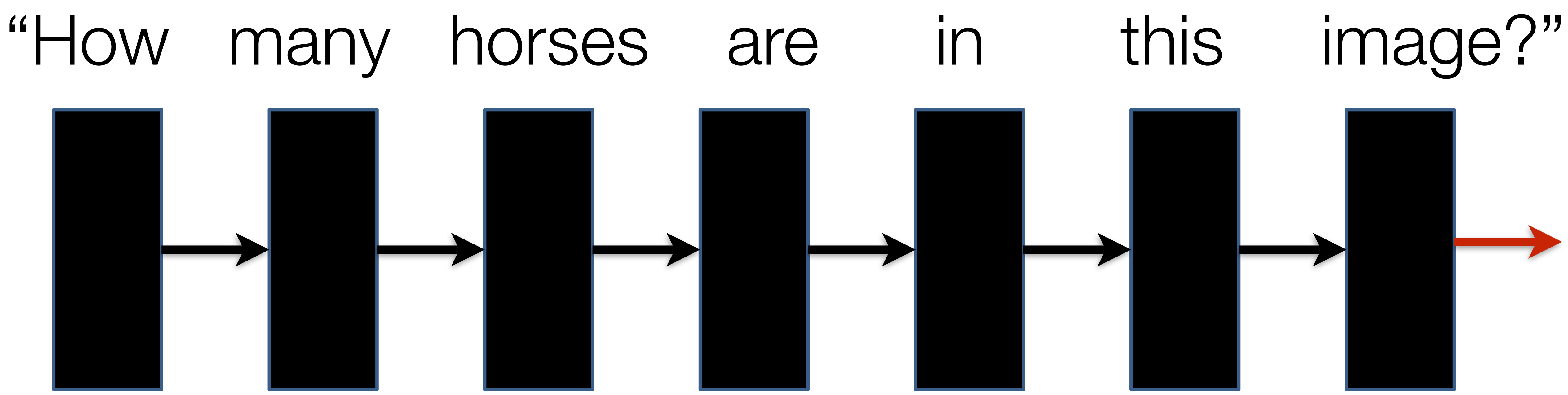
* slide from Dhruv Batra

Applications: Typical Visual Question Answering (VQA)

Image Embedding (VGGNet)

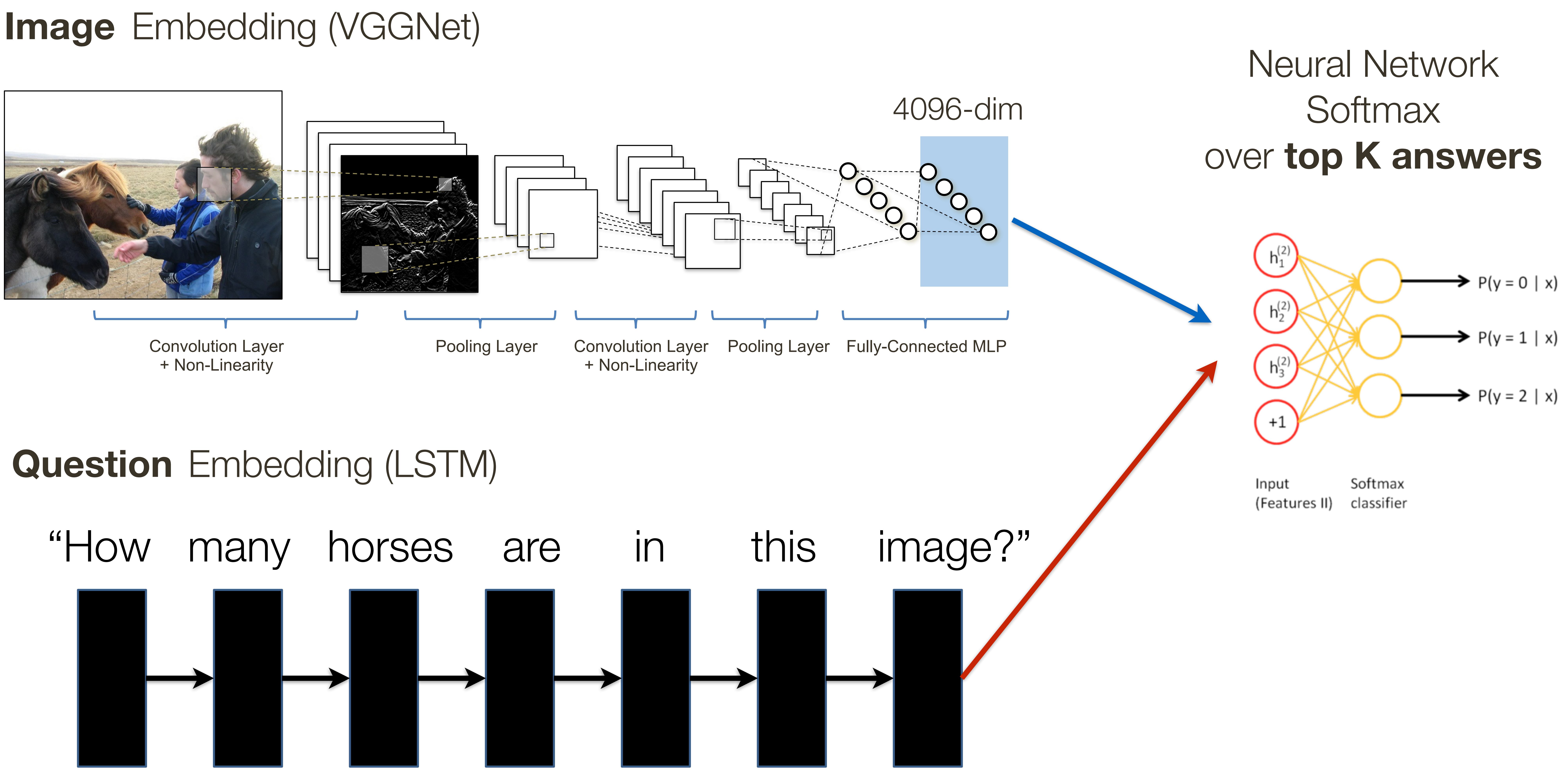


Question Embedding (LSTM)



* slide from Dhruv Batra

Applications: Typical Visual Question Answering (VQA)



* slide from Dhruv Batra

Applications: Visual Dialogs

[Seo et al., NIPS 2017]

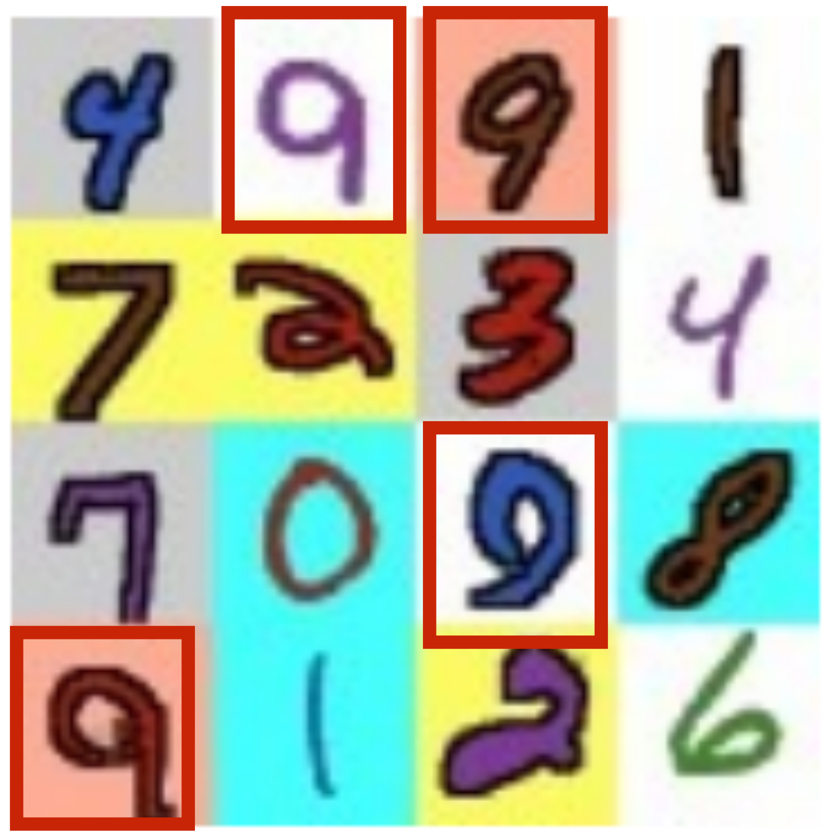
Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

#	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

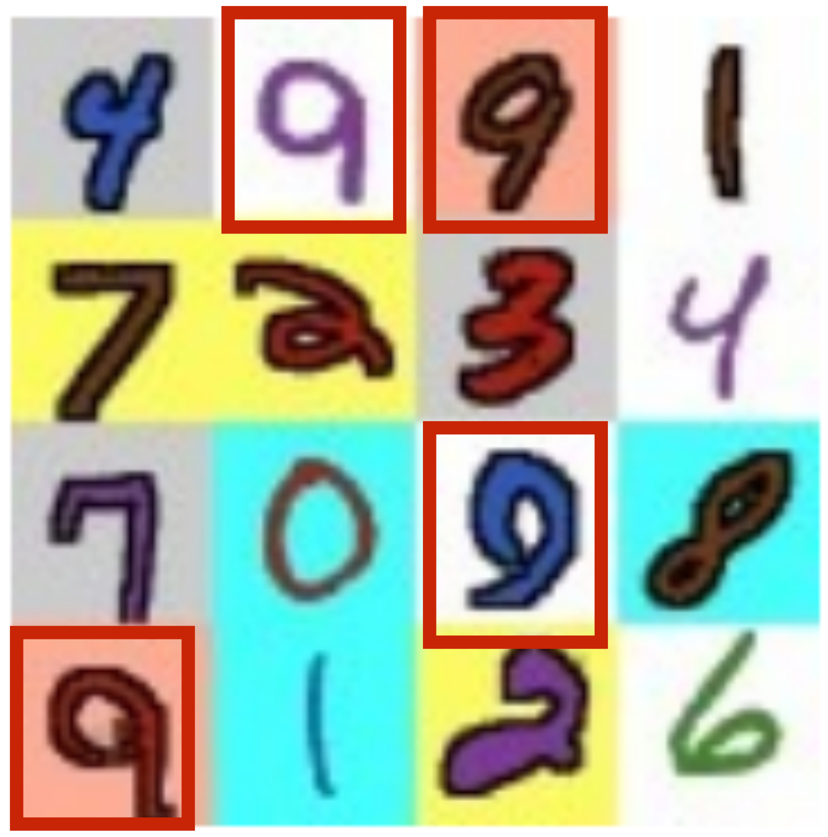


#	Question	Answer
1	How many 9's are there in the image?	-

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]

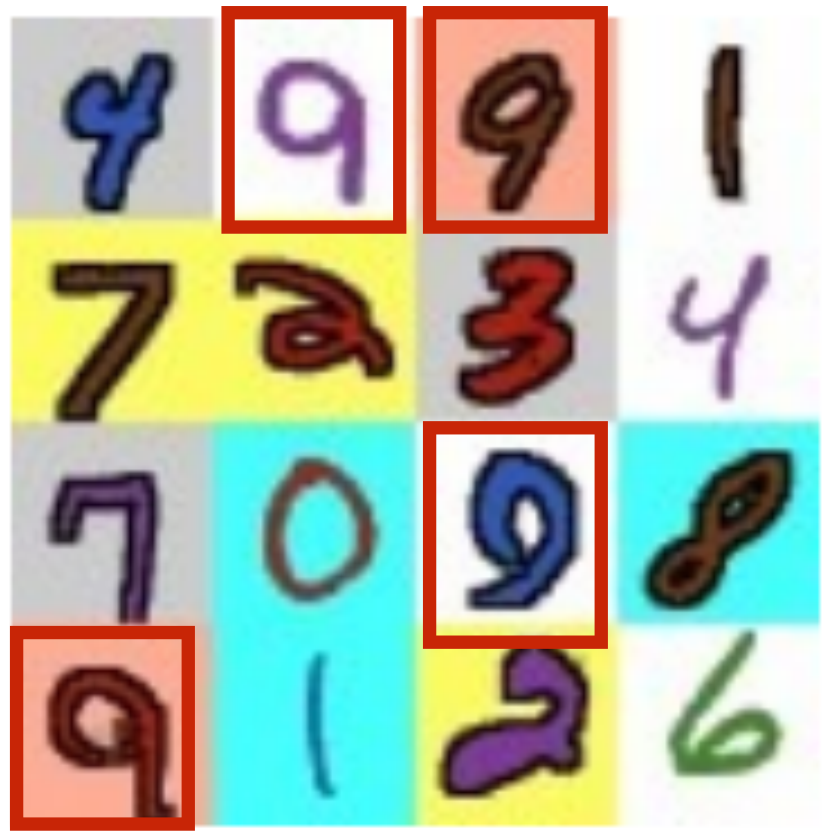
Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

#	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]

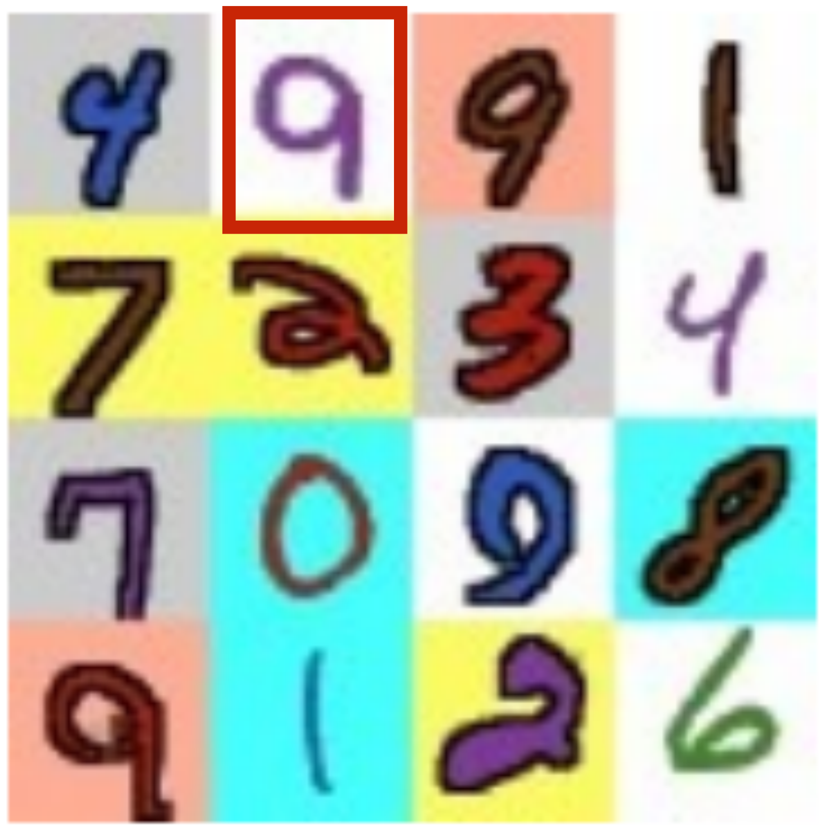
Interconnected questions in sequence: Typically questions later in the dialog make references to the earlier questions in the dialog history

#	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

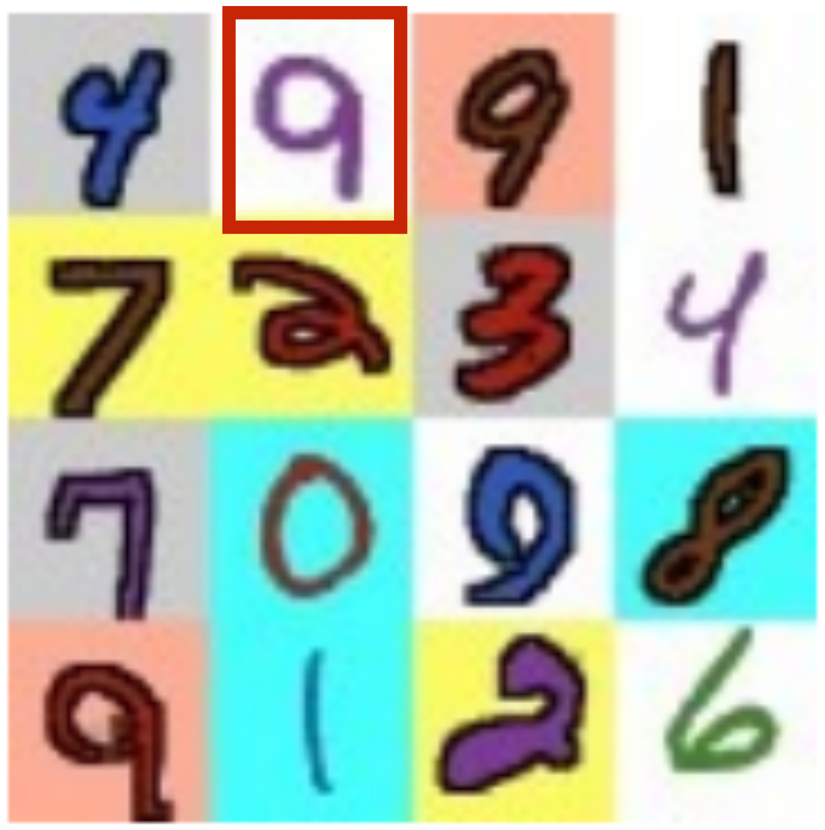


#	Question	Answer
1	How many 9's are there in the image?	four
2	How many brown digits are there among them?	one
3	What is the background color of the digit at the left of it?	white

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

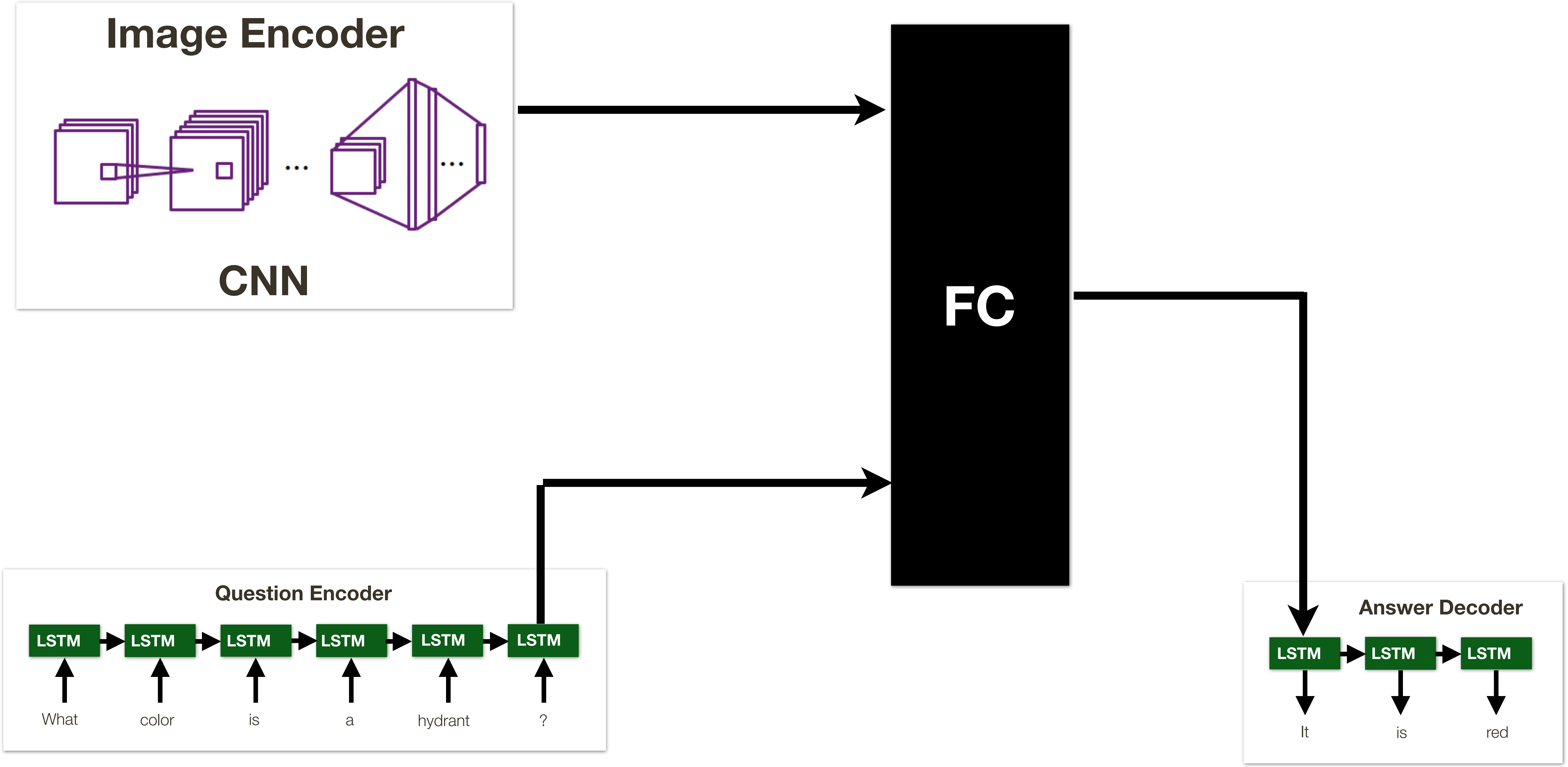
Simple Visual Question Answering

[Seo et al., NIPS 2017]

Q: What color is a hydrant?

Simple Visual Question Answering

[Seo et al., NIPS 2017]

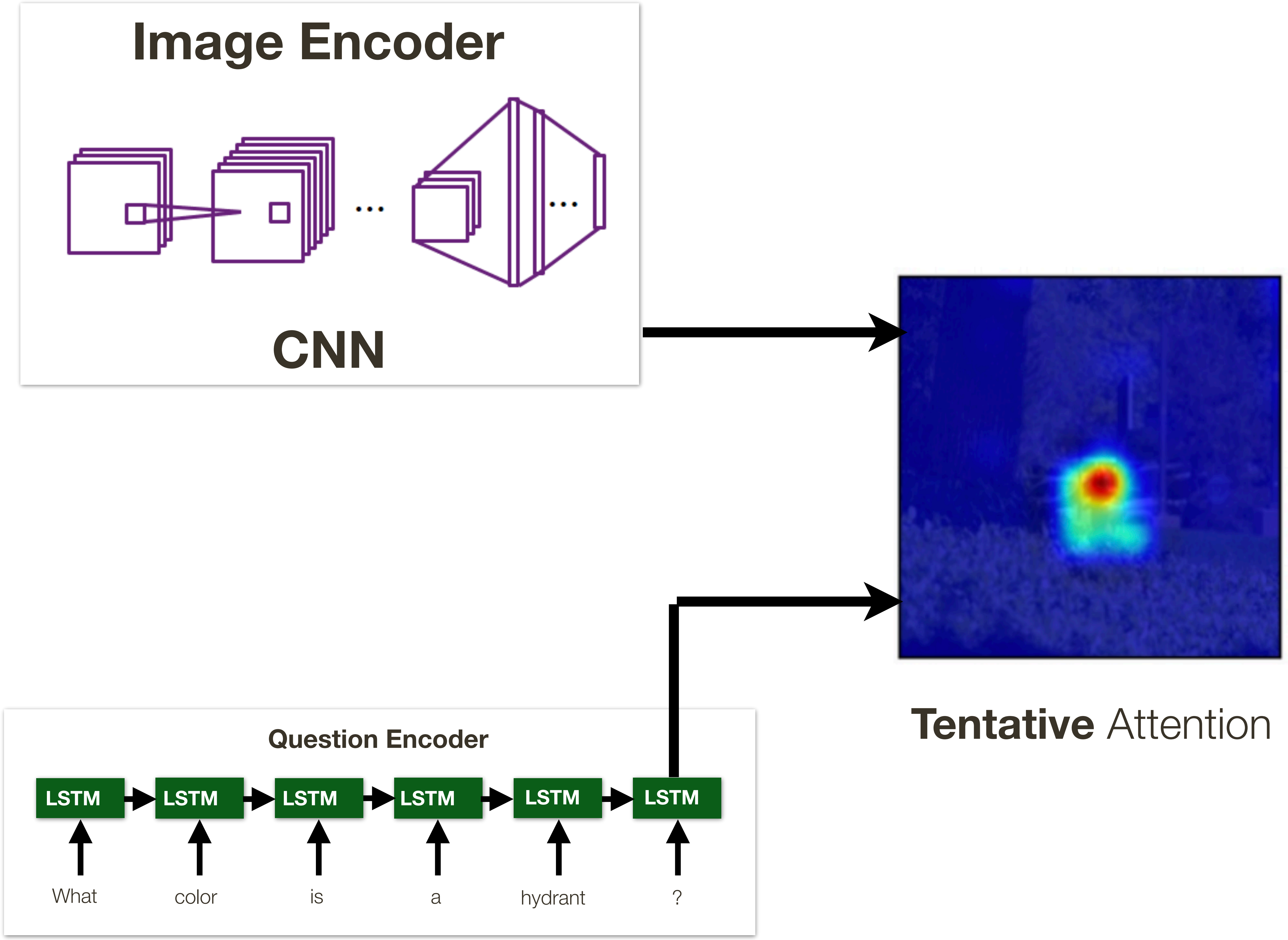


Q: What color is a hydrant?

A: It is red

Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]

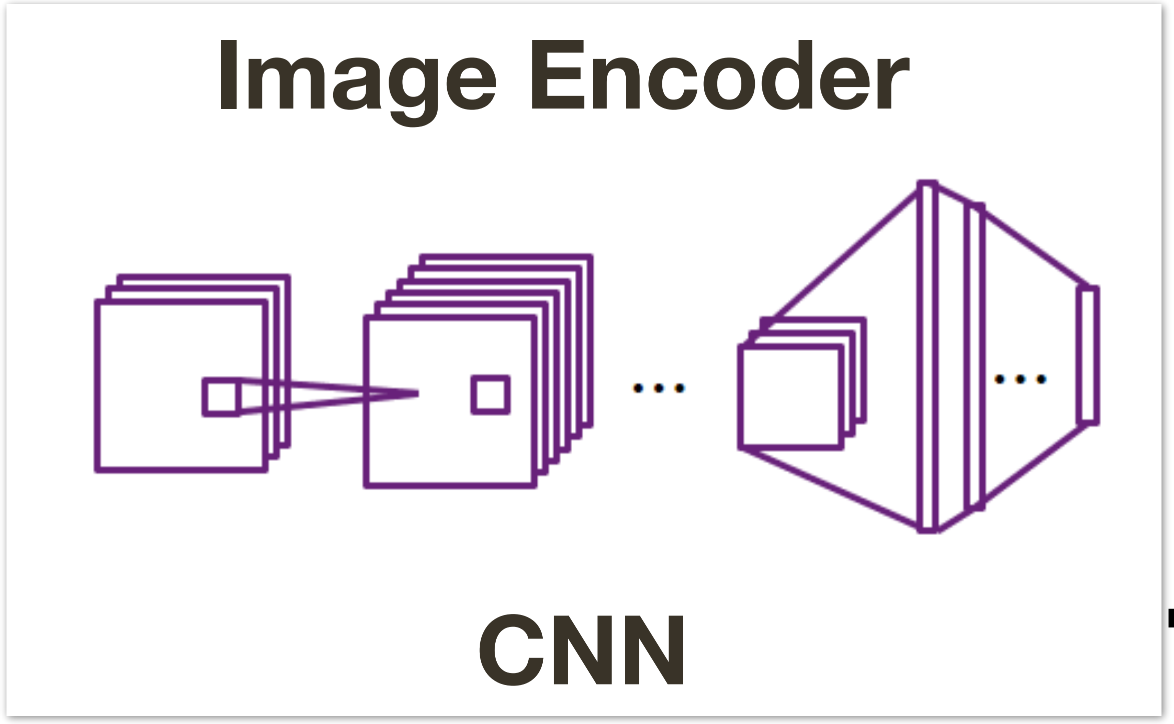


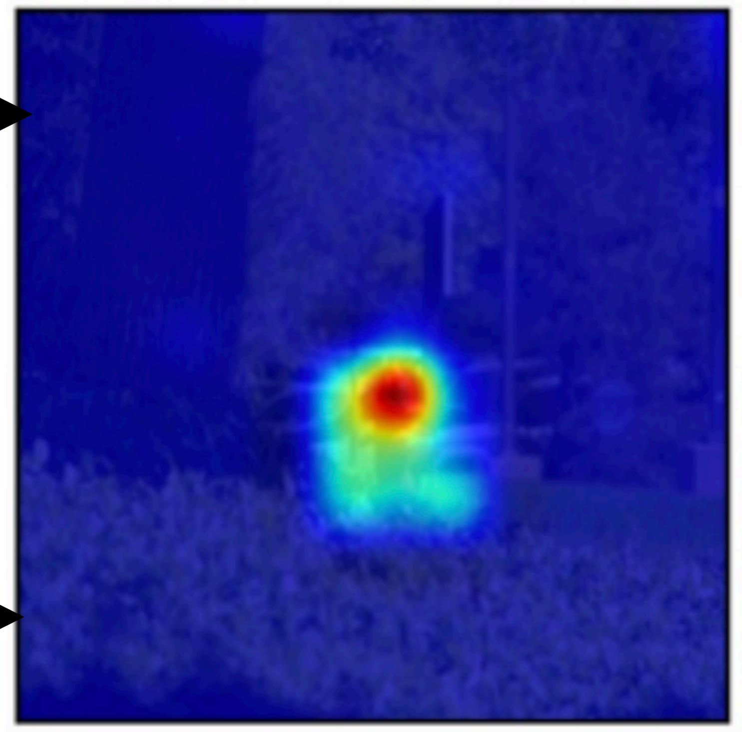
Q: What color is a hydrant?

Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]

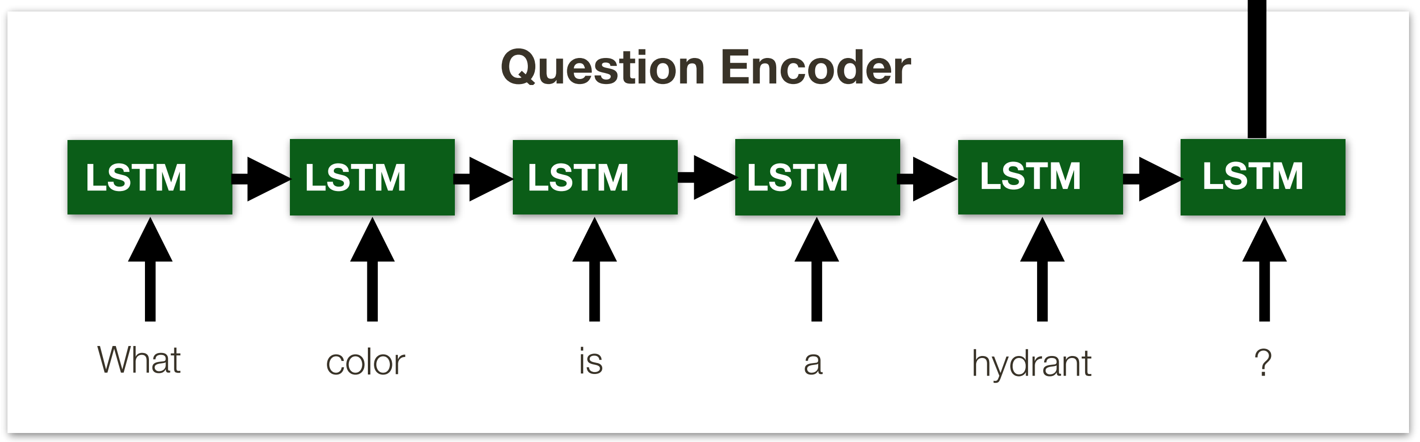
$M \times M = n$ grid with values between 0 and 1, indicating which part of the image to pay attention to in order to answer the question





Tentative Attention

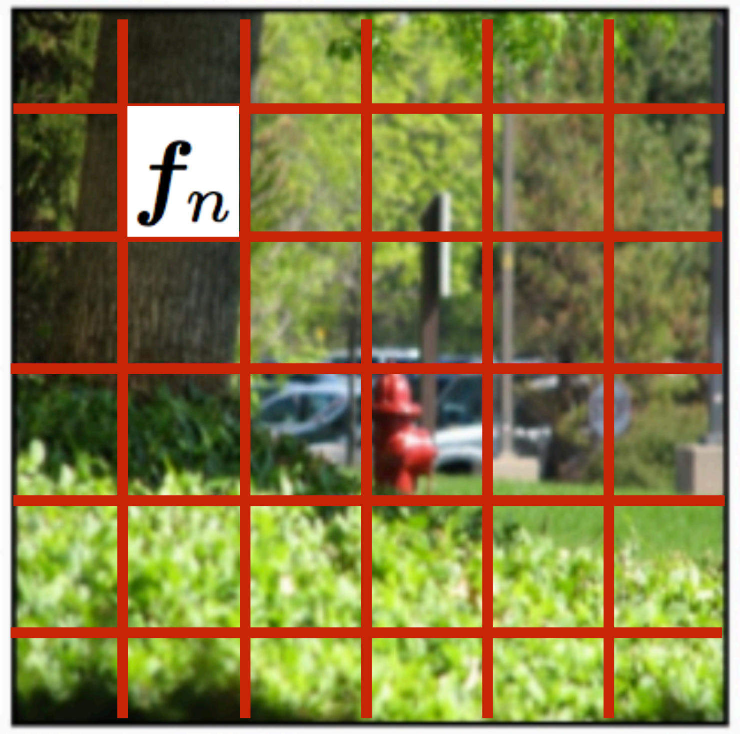
Q: What color is a hydrant?

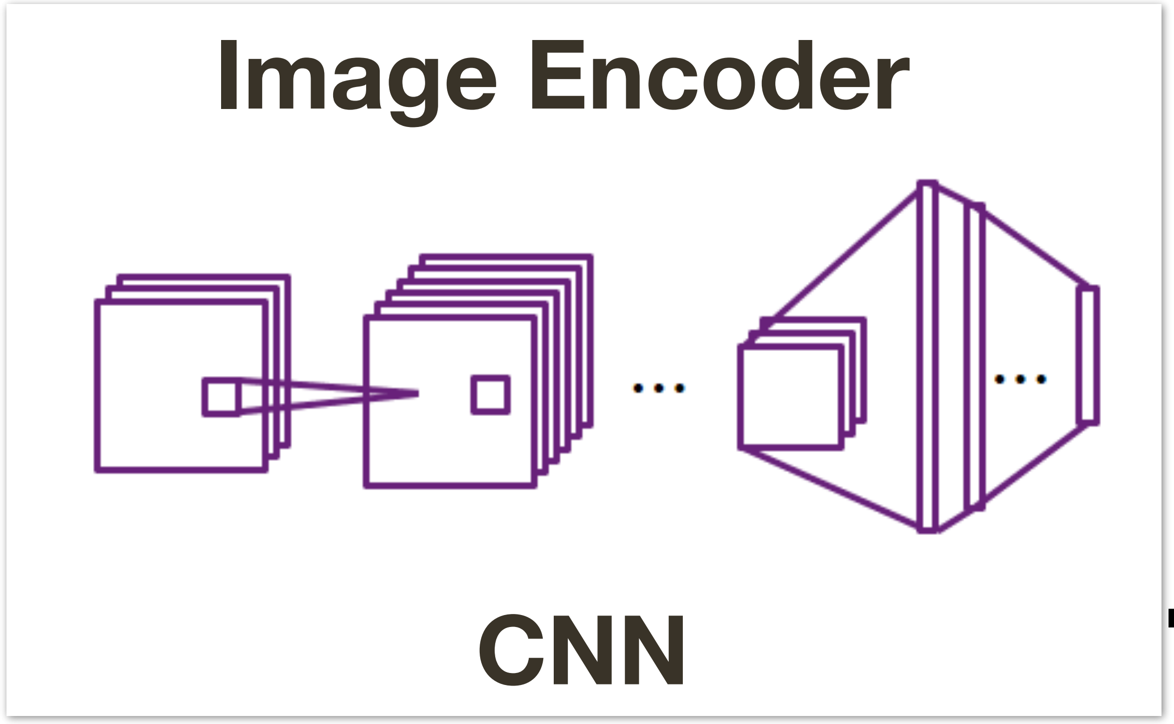


Attention Networks for Visual Question Answering

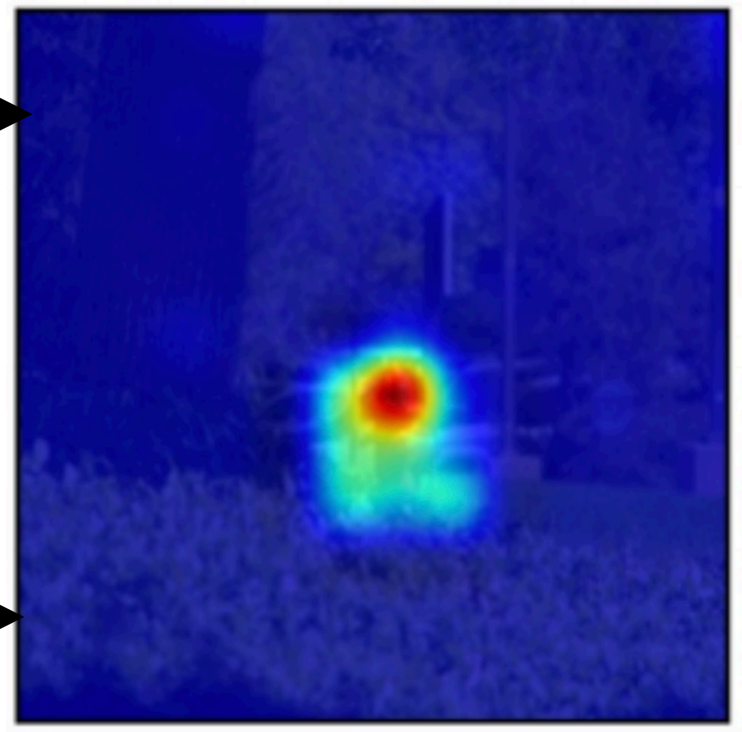
[Seo et al., NIPS 2017]

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



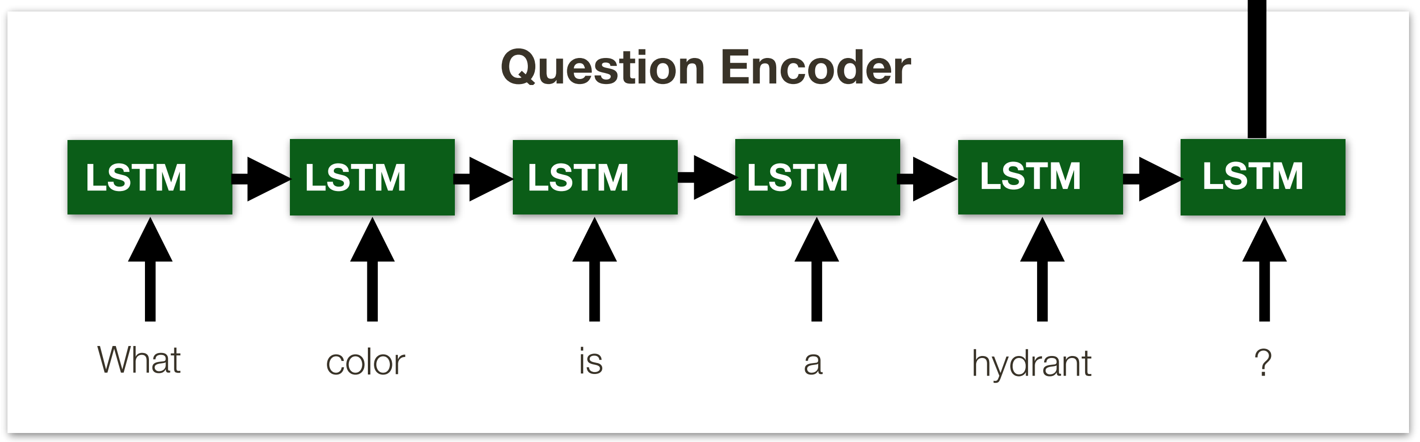


c_t



Tentative Attention

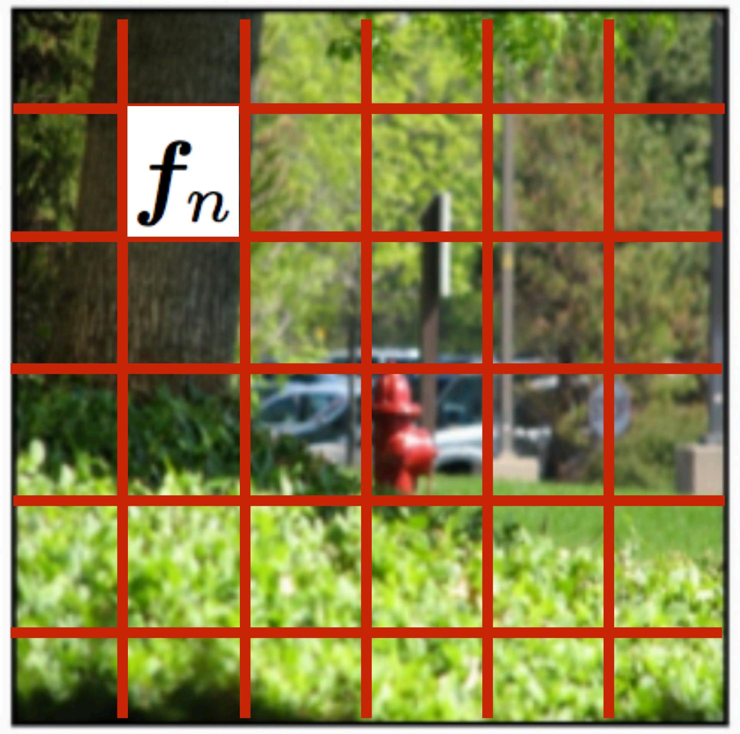
Q: What color is a hydrant?

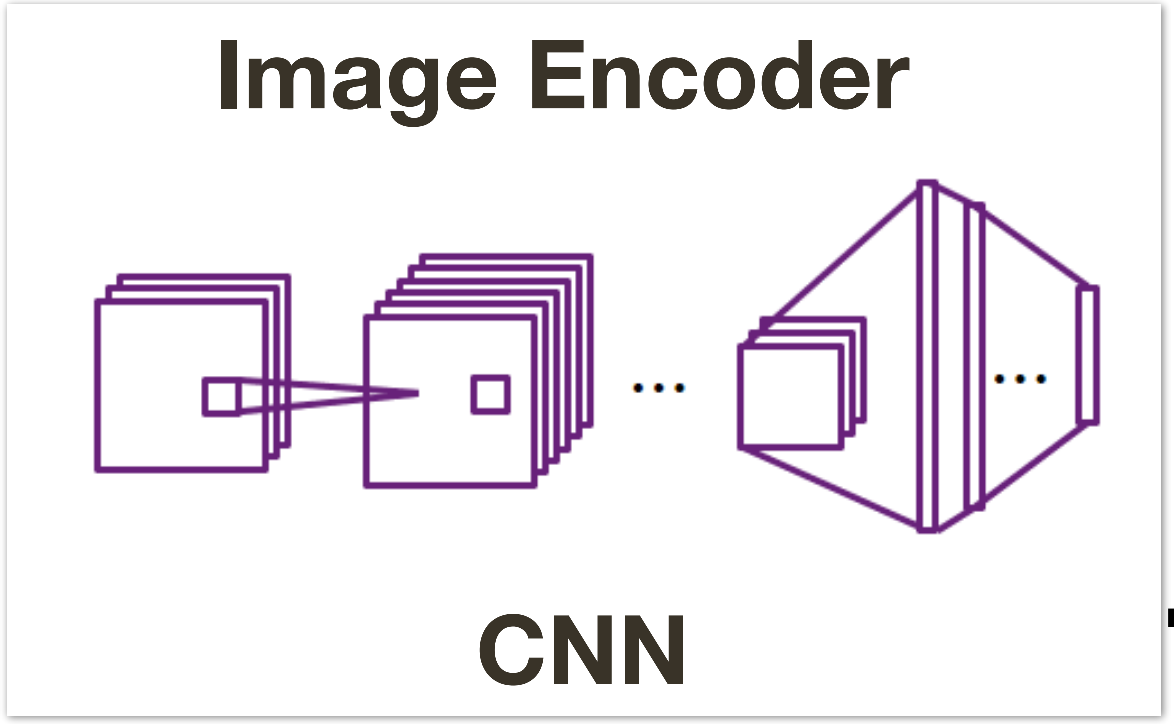


$$s_{t,n} = \left(\mathbf{W}_c^{\text{tent}} \boldsymbol{c}_t\right)^\top \left(\mathbf{W}_f^{\text{tent}} \boldsymbol{f}_n\right)$$

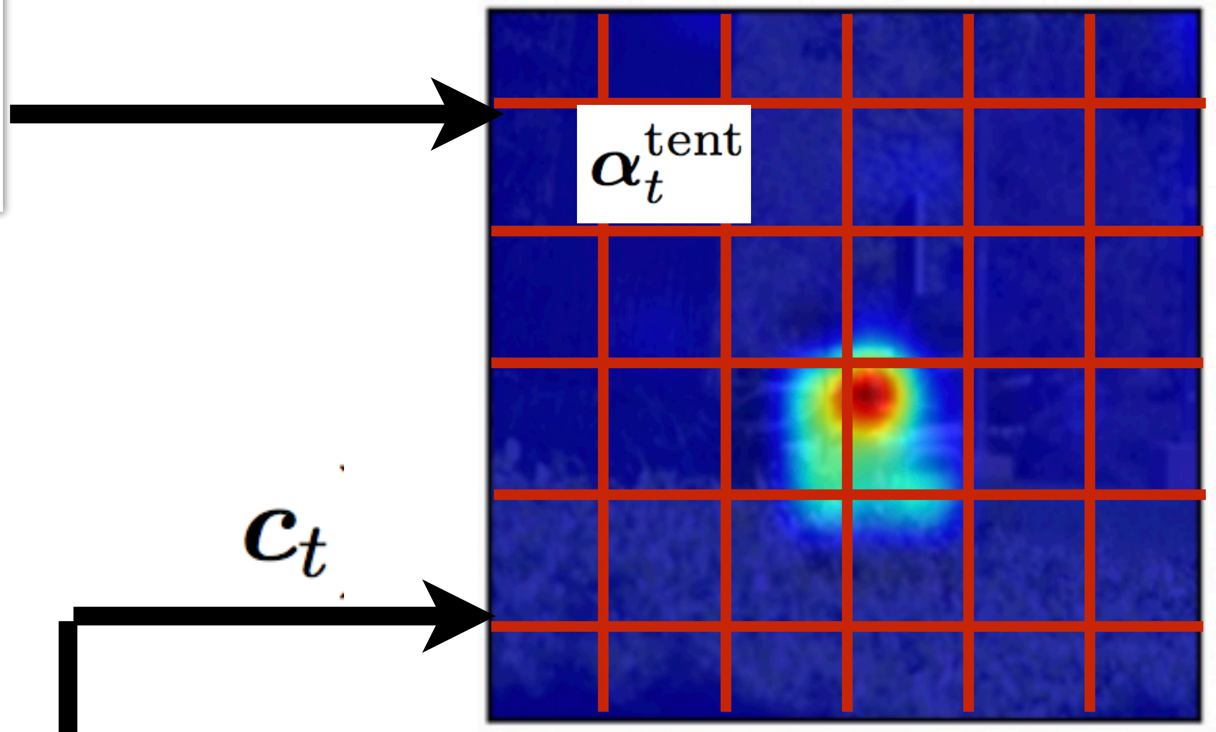
Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]



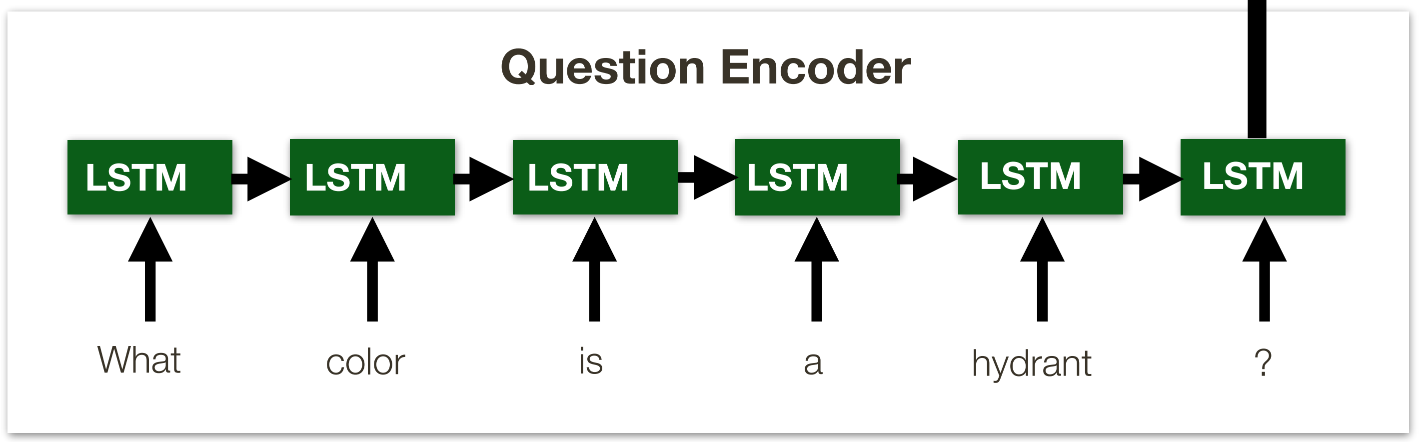


$M \; x \; M = n$ grid with values between 0 and 1, indicating which part of the image to pay attention to in order to answer the question



Tentative Attention

Q: What color is a hydrant?

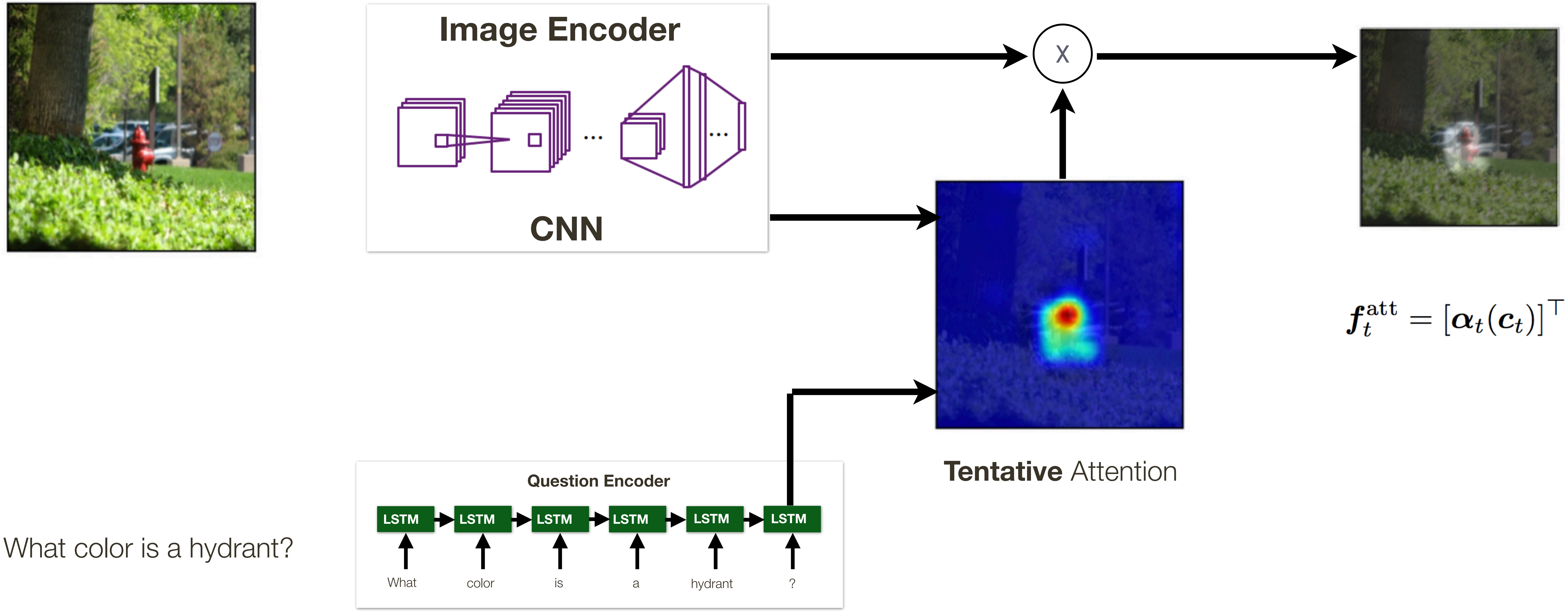


$$s_{t,n} = \left(\mathbf{W}_c^{\text{tent}} \boldsymbol{c}_t\right)^\top \left(\mathbf{W}_f^{\text{tent}} \boldsymbol{f}_n\right)$$

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

Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]

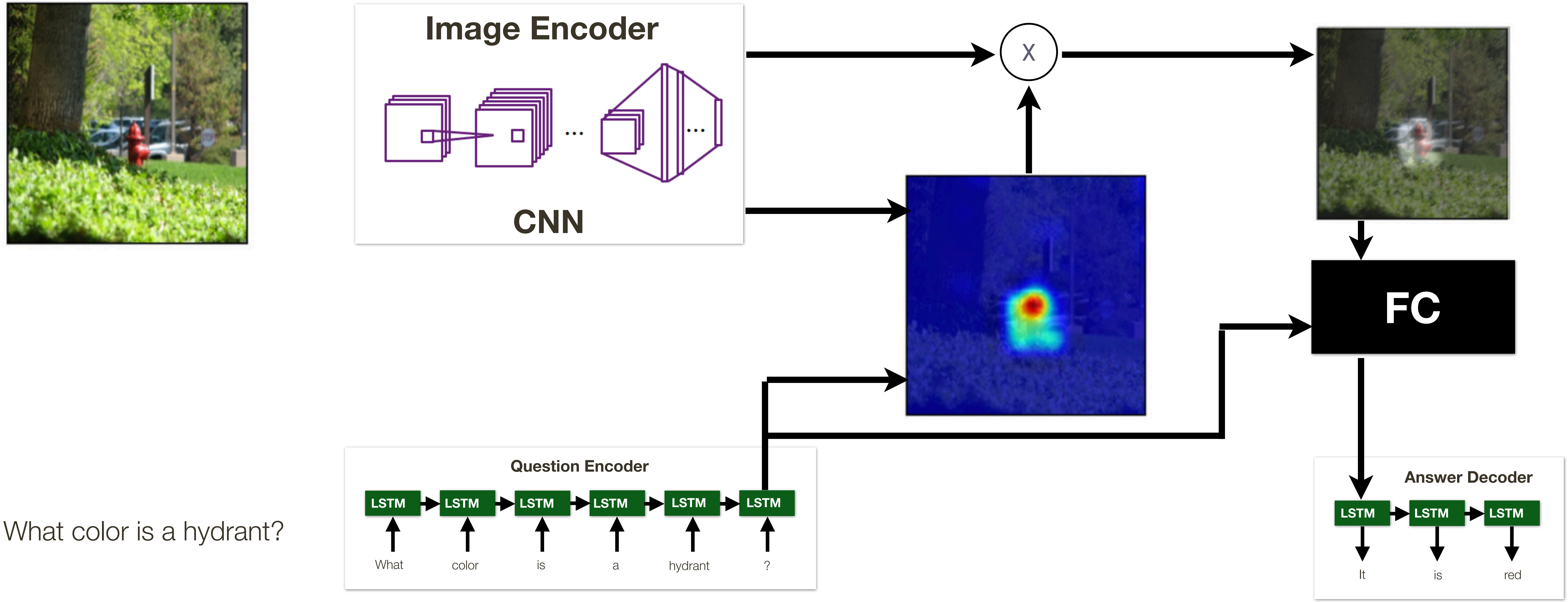


$$\boldsymbol{f}_t^{\mathrm{att}} = [\boldsymbol{\alpha}_t(\boldsymbol{c}_t)]^\top \cdot \boldsymbol{f}$$

Q: What color is a hydrant?

Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]

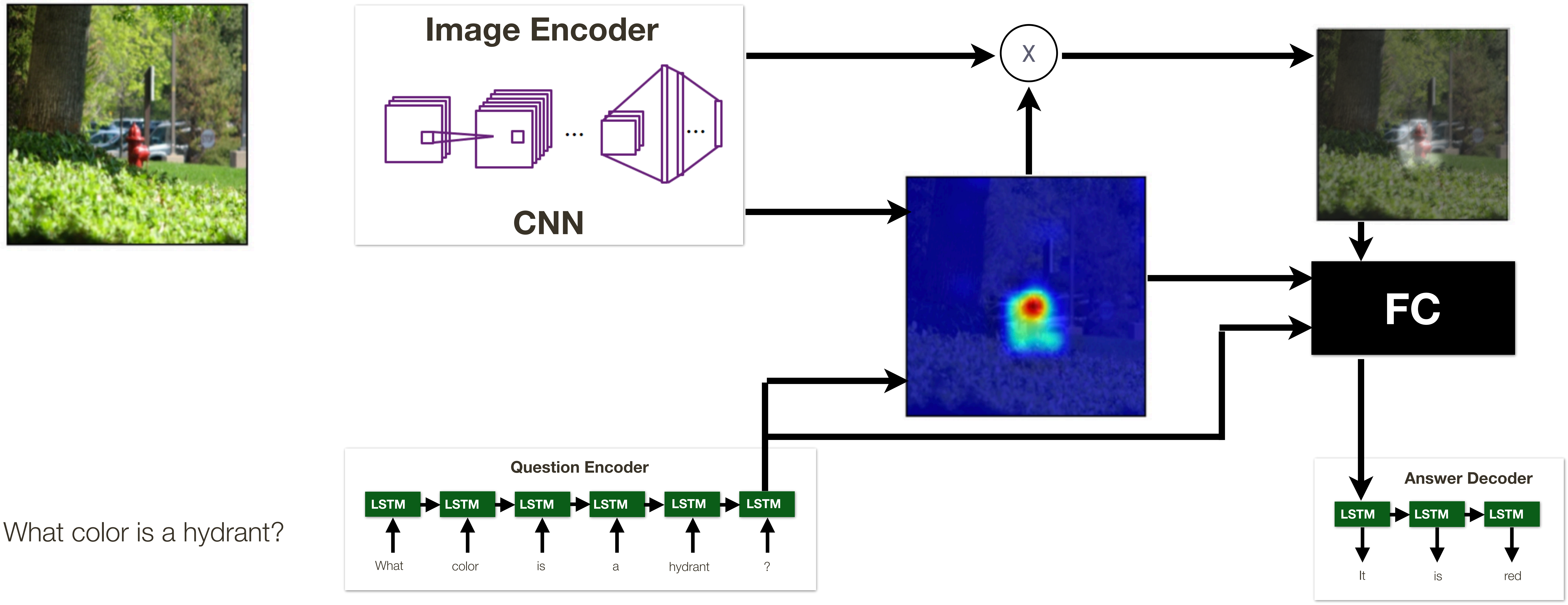


Q: What color is a hydrant?

A: It is red

Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]

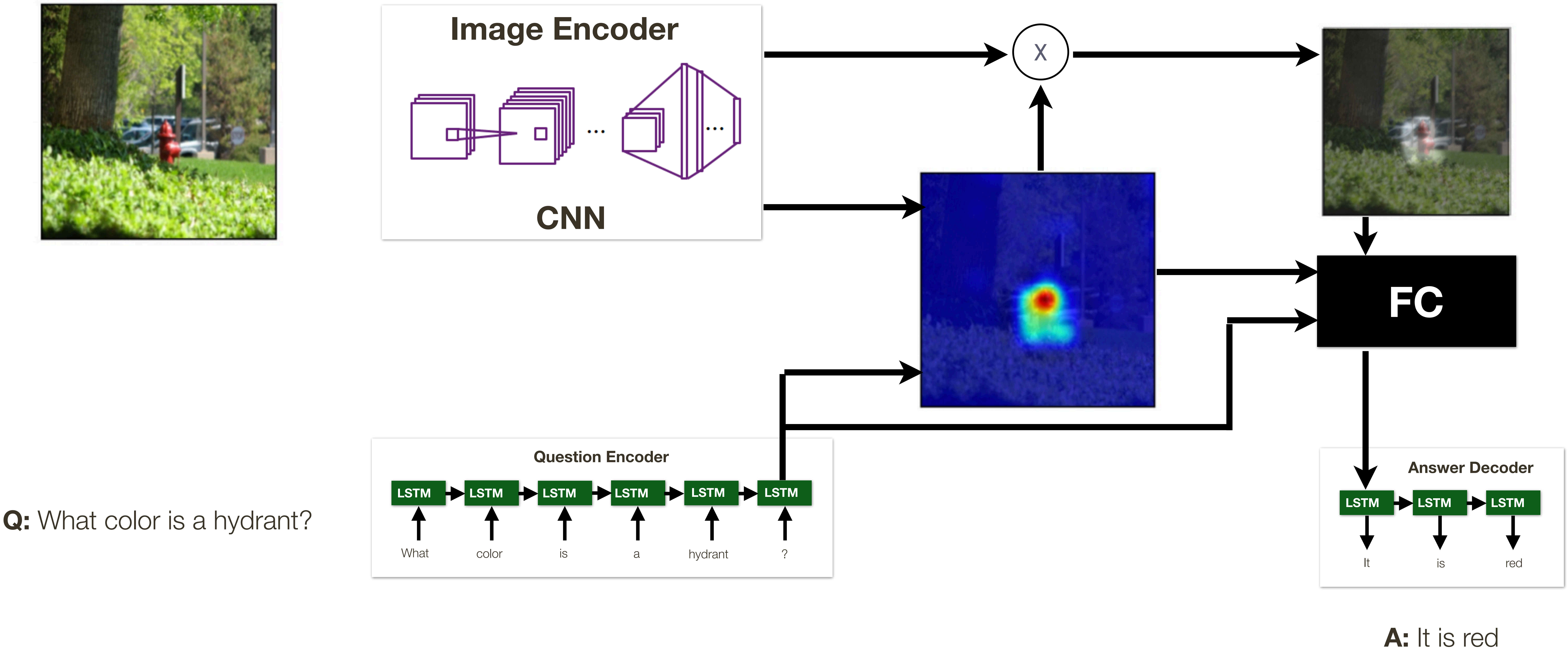


Q: What color is a hydrant?

A: It is red

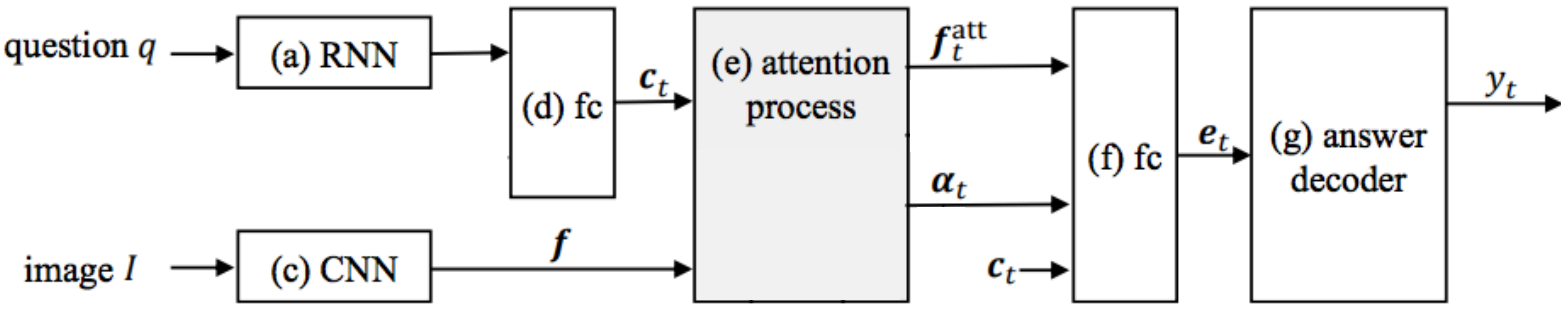
Attention Networks for Visual Question Answering

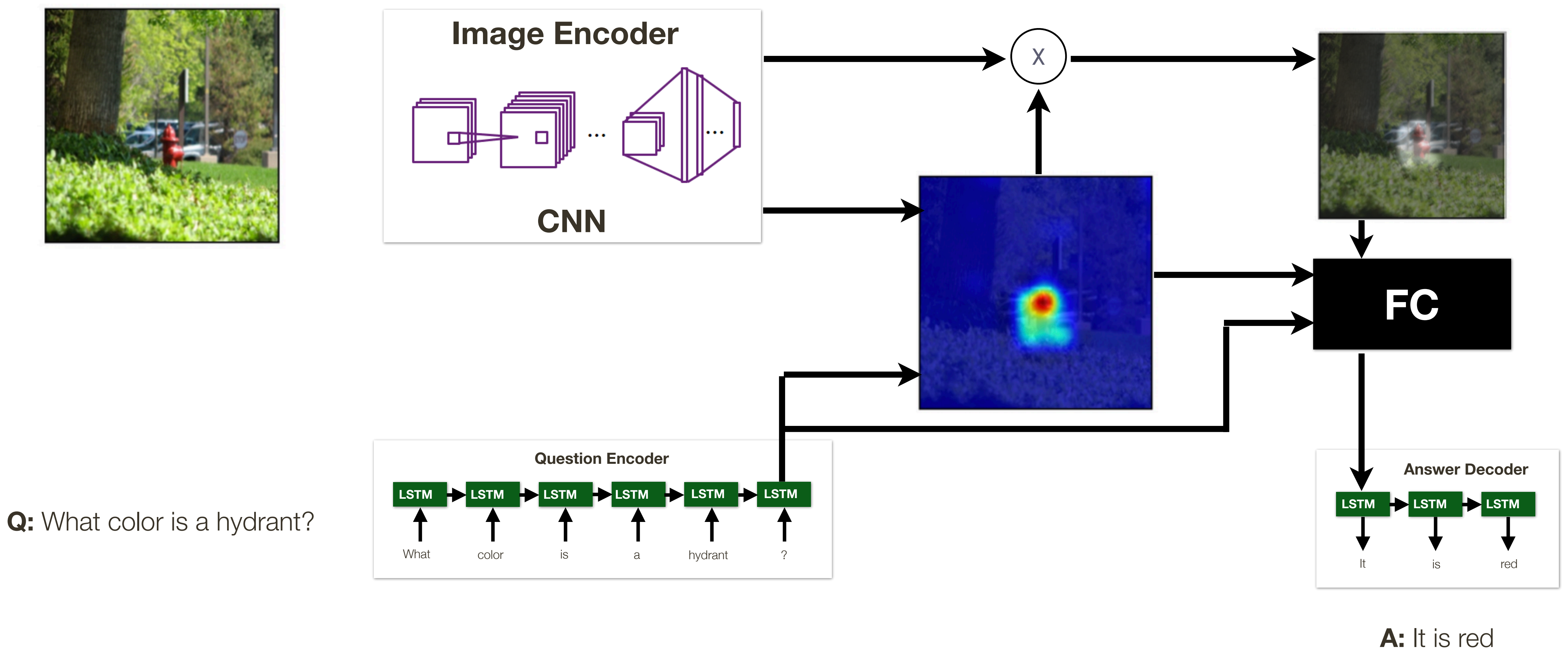
[Seo et al., NIPS 2017]



Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]





Visual Dialog Task

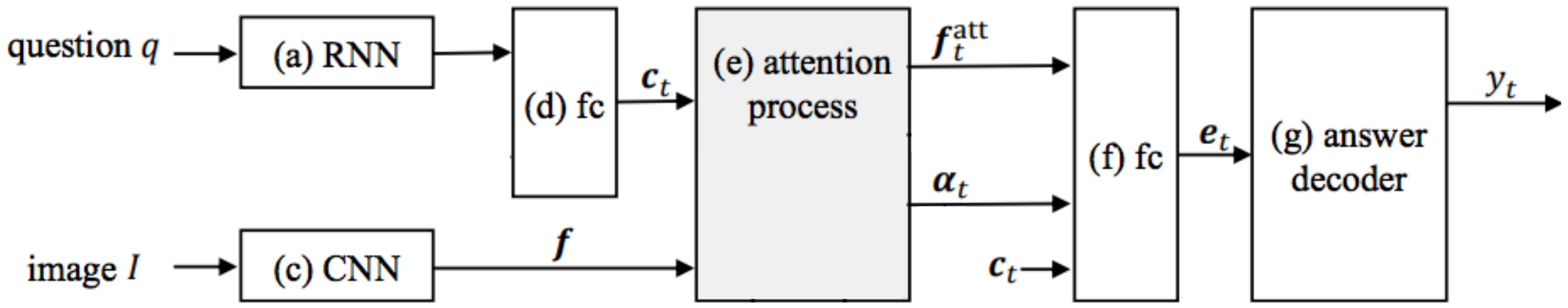
[Seo et al., NIPS 2017]

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

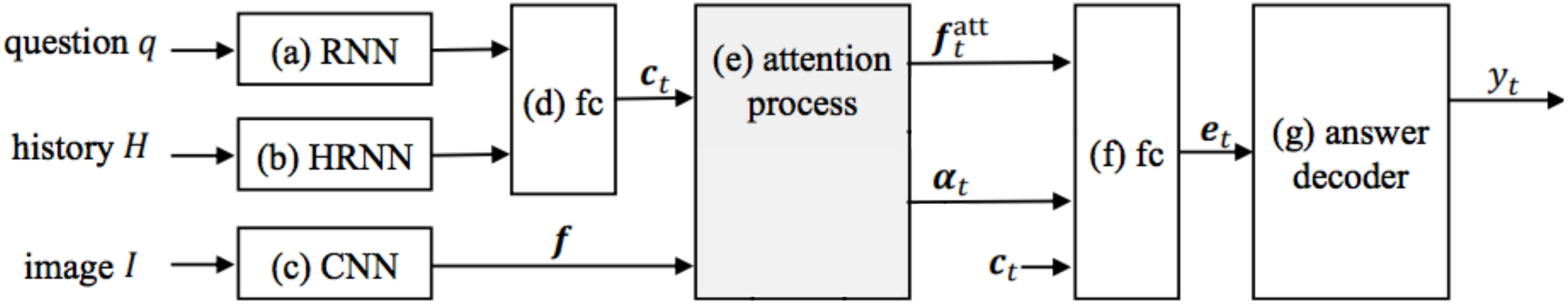
Attention Networks for Visual Question Answering

[Seo et al., NIPS 2017]



Attention Networks for Visual Dialogs

[Seo et al., NIPS 2017]



Hierarchical RNN (**HRNN**):

— Encode the question using LSTM

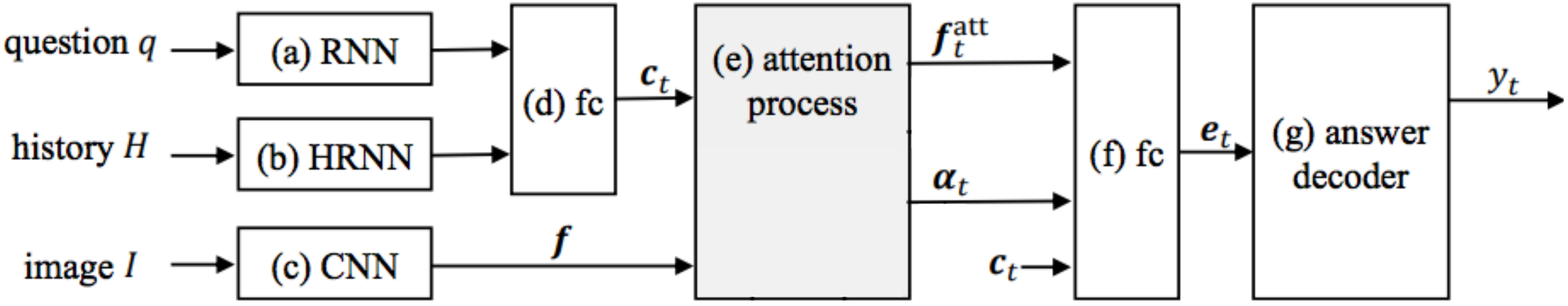
— Encode the answer using LSTM

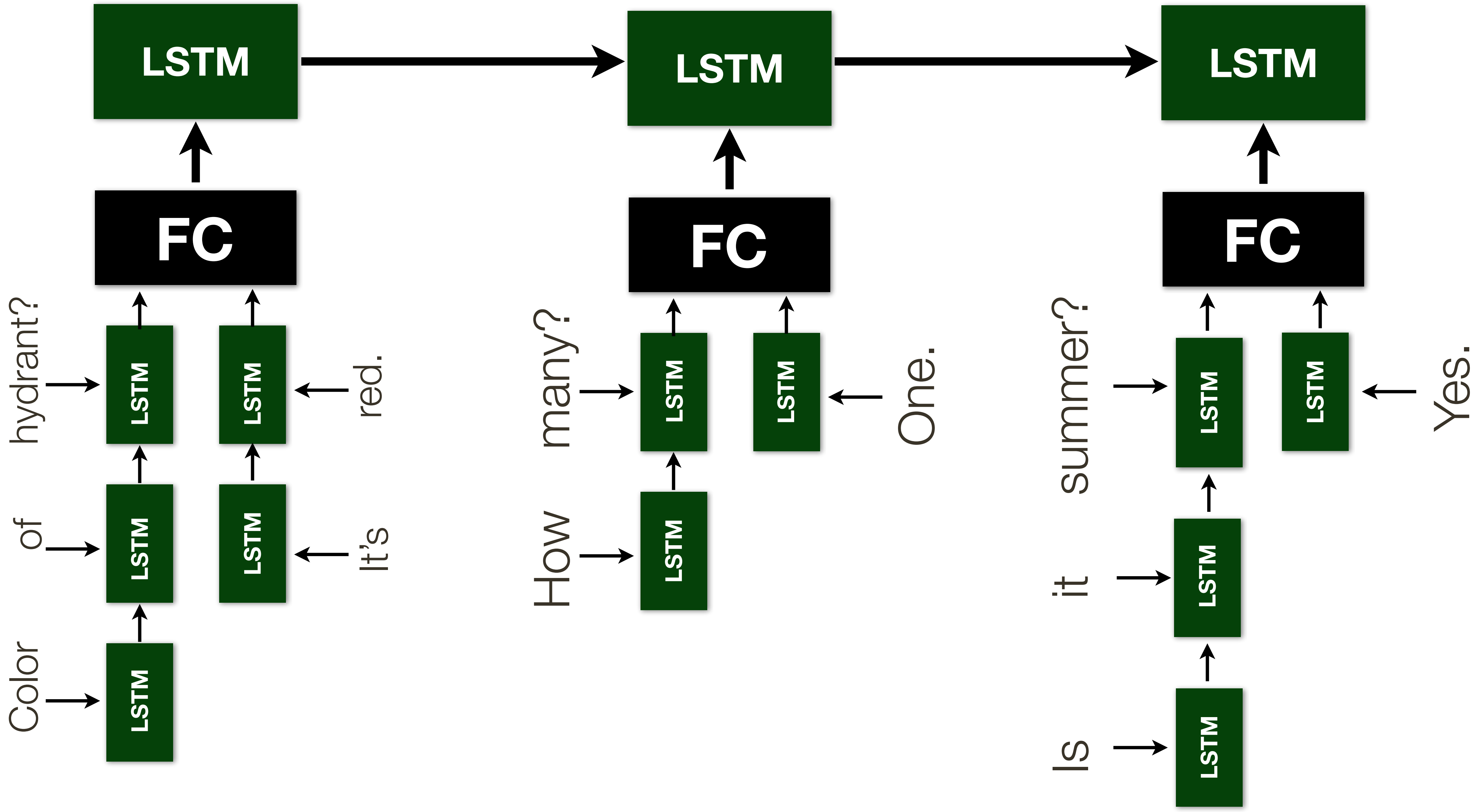
— Obtain QA embedding by fusing them using FC layer

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

Attention Networks for Visual Dialogs

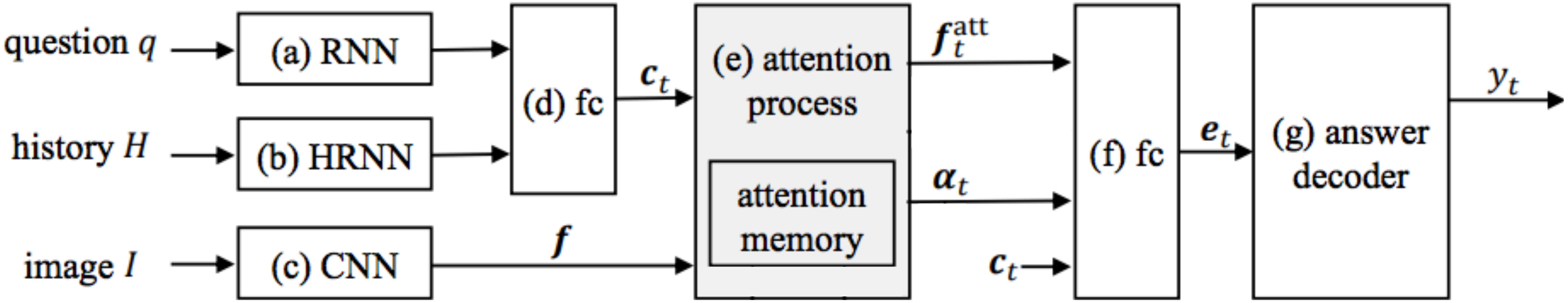
[Seo et al., NIPS 2017]



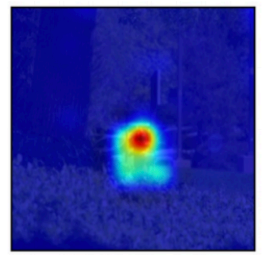
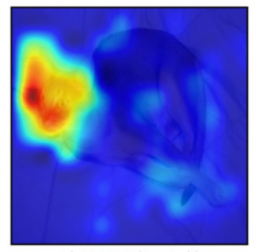


Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]



Associative Memory:

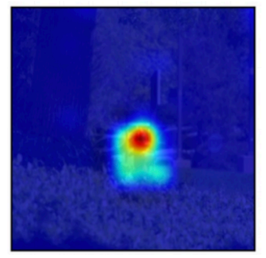
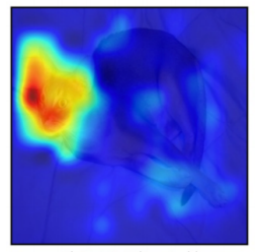
Question Turn	Key (hash)	Memory
1	f (**H:** Empty; **Q:** What color is a hydrant? **A:** It is red)	
2	f (**H:** …; **Q:** Is there a tree? **A:** Yes)	

Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]

Q3: What color is it?

Associative Memory:

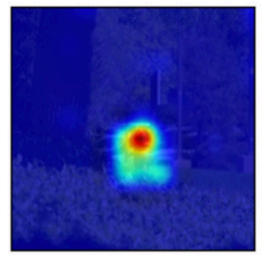
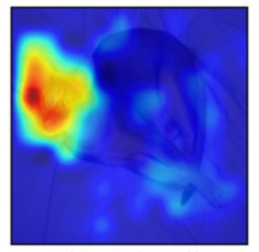
Question Turn	Key (hash)	Memory
1	f (**H:** Empty; **Q:** What color is a hydrant? **A:** It is red)	
2	f (**H:** …; **Q:** Is there a tree? **A:** Yes)	

Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]

Q3: What color is it?

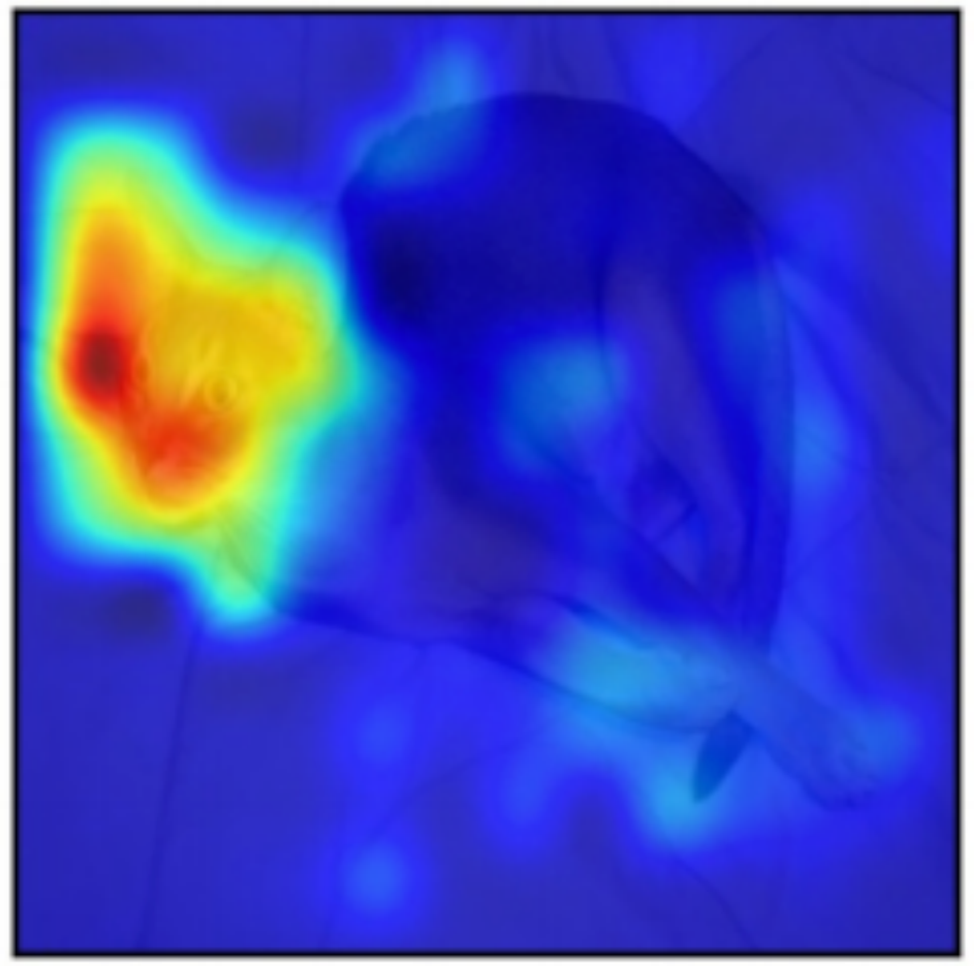
Associative Memory:

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)	

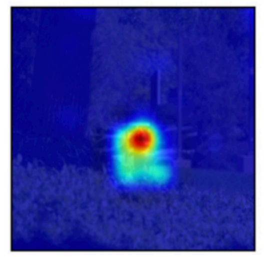
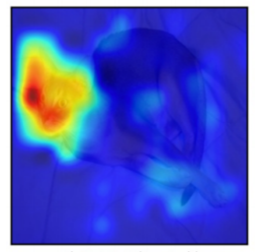
Memory Networks for Visual Dialogs

[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]

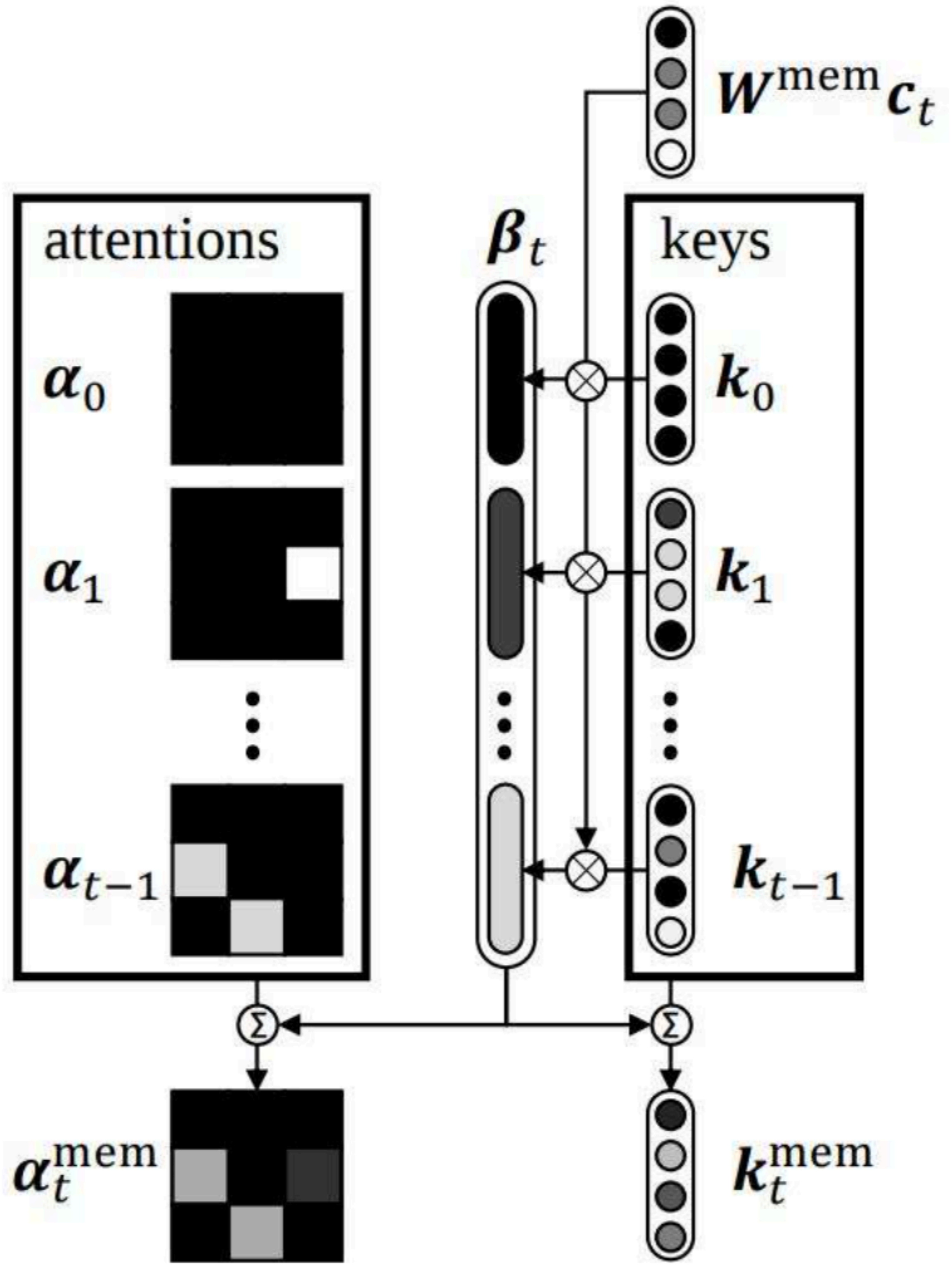
Key Idea: Every item in memory is **`(attention, key)`** pair — explicitly storing attentions used to answer previous questions

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?



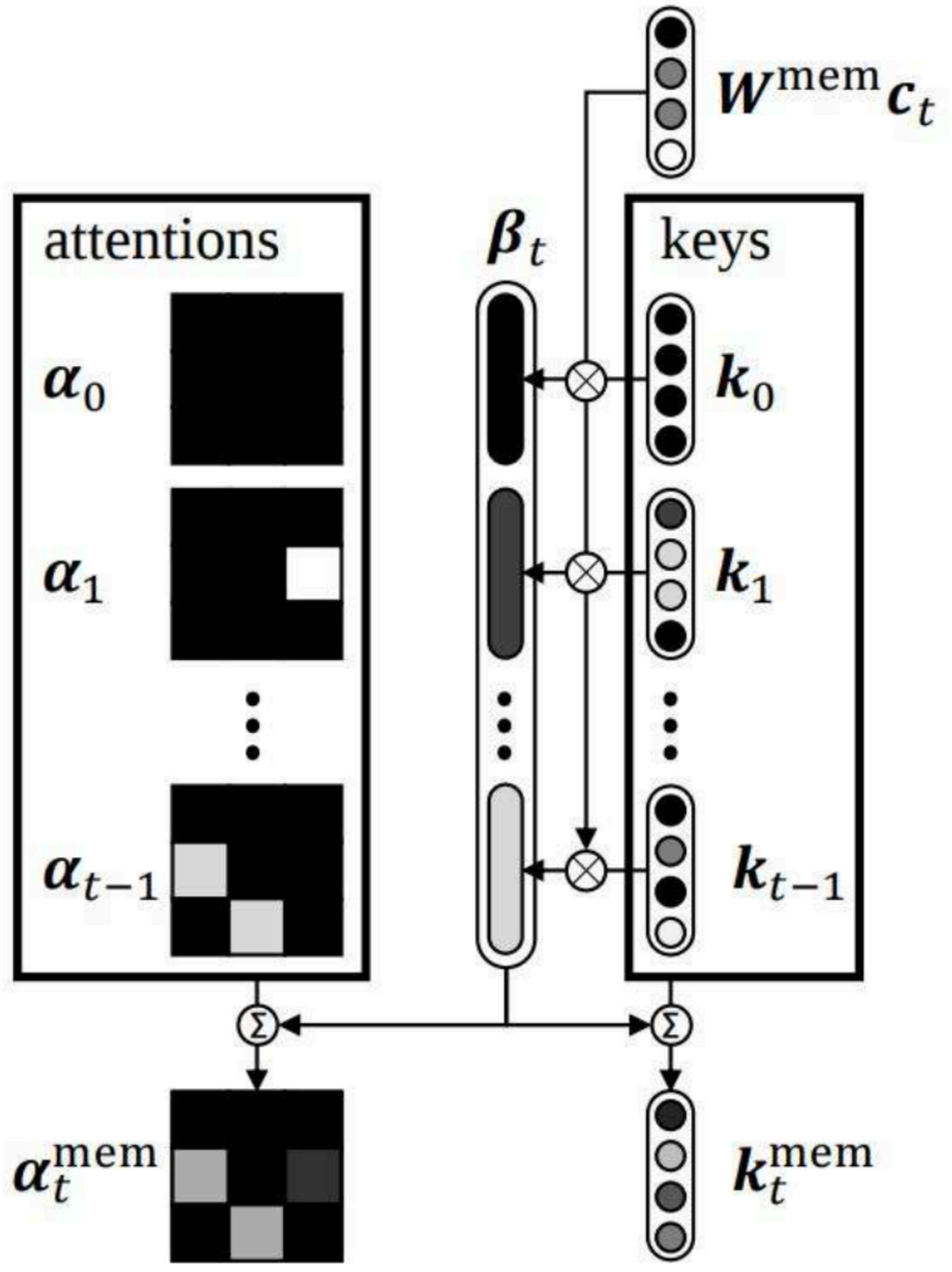
Associative Memory Attention

[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

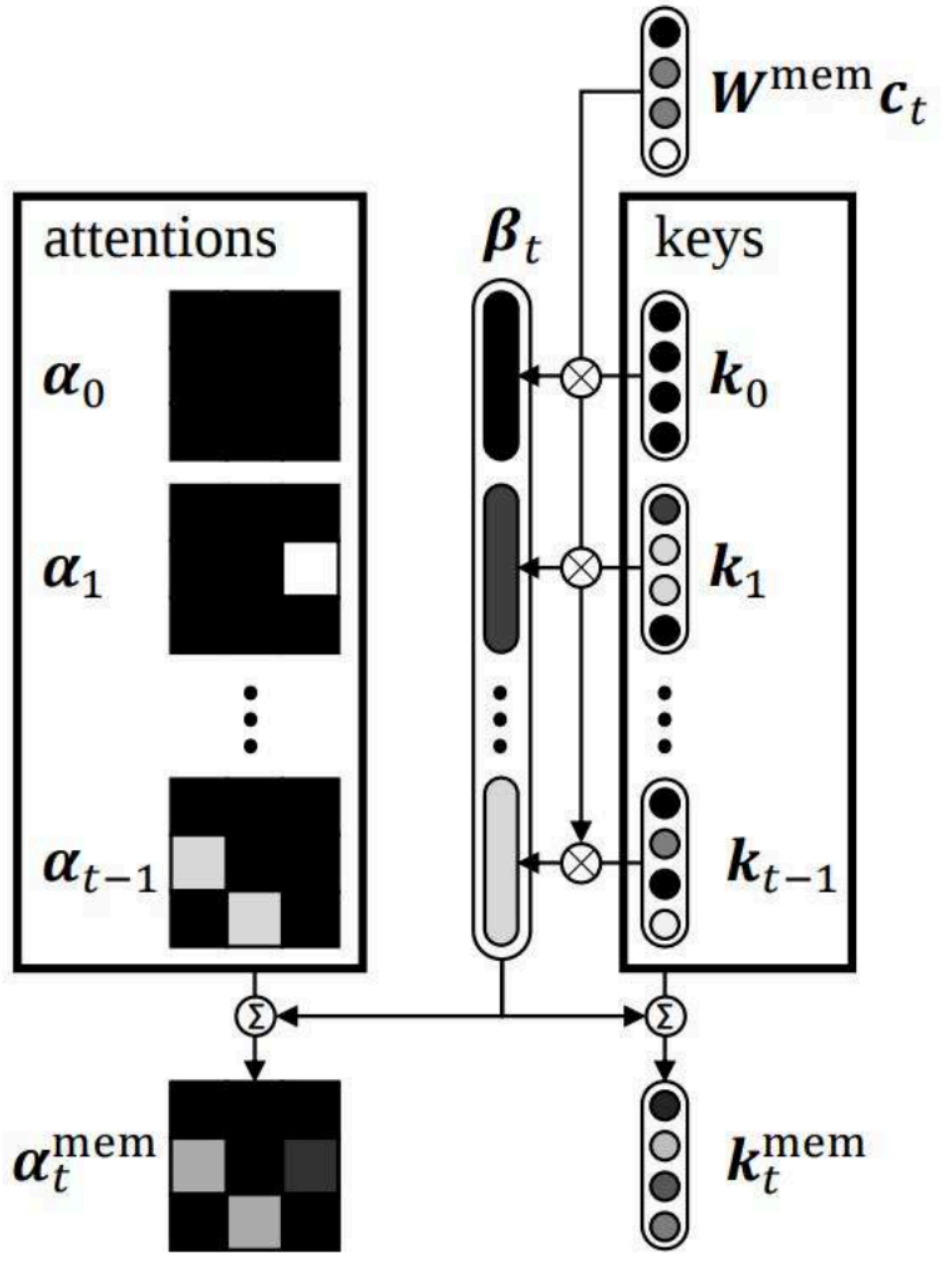
[Seo et al., NIPS 2017]

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



Associative Memory Attention

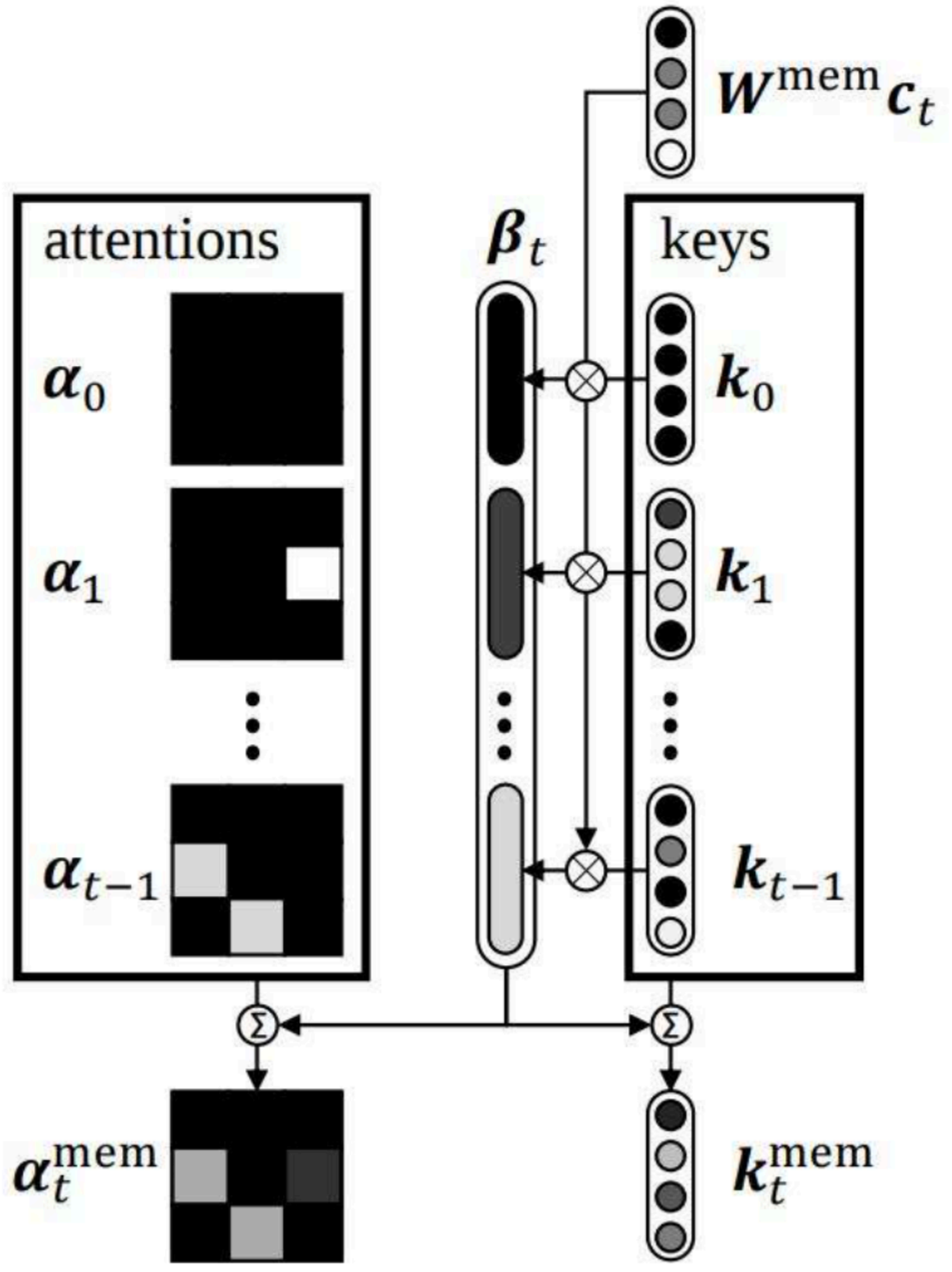
Observation: This formulation gives all previous turns equal weight (uniform prior)

$$m_{t,\tau} = (\boldsymbol{W}^{\text{mem}}\boldsymbol{c}_t)^\top \boldsymbol{k}_\tau$$

$$\boldsymbol{\beta}_t = \text{softmax}\left(\{m_{t,\tau}, 0 < \tau < t-1\}\right)$$

$$\boldsymbol{\alpha}_t^{\text{mem}} = \sum_{\tau=0}^{t-1} \boldsymbol{\beta}_{t,\tau}\boldsymbol{\alpha}_\tau$$

Weighted combination
of **attention maps**



Associative Memory Attention

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

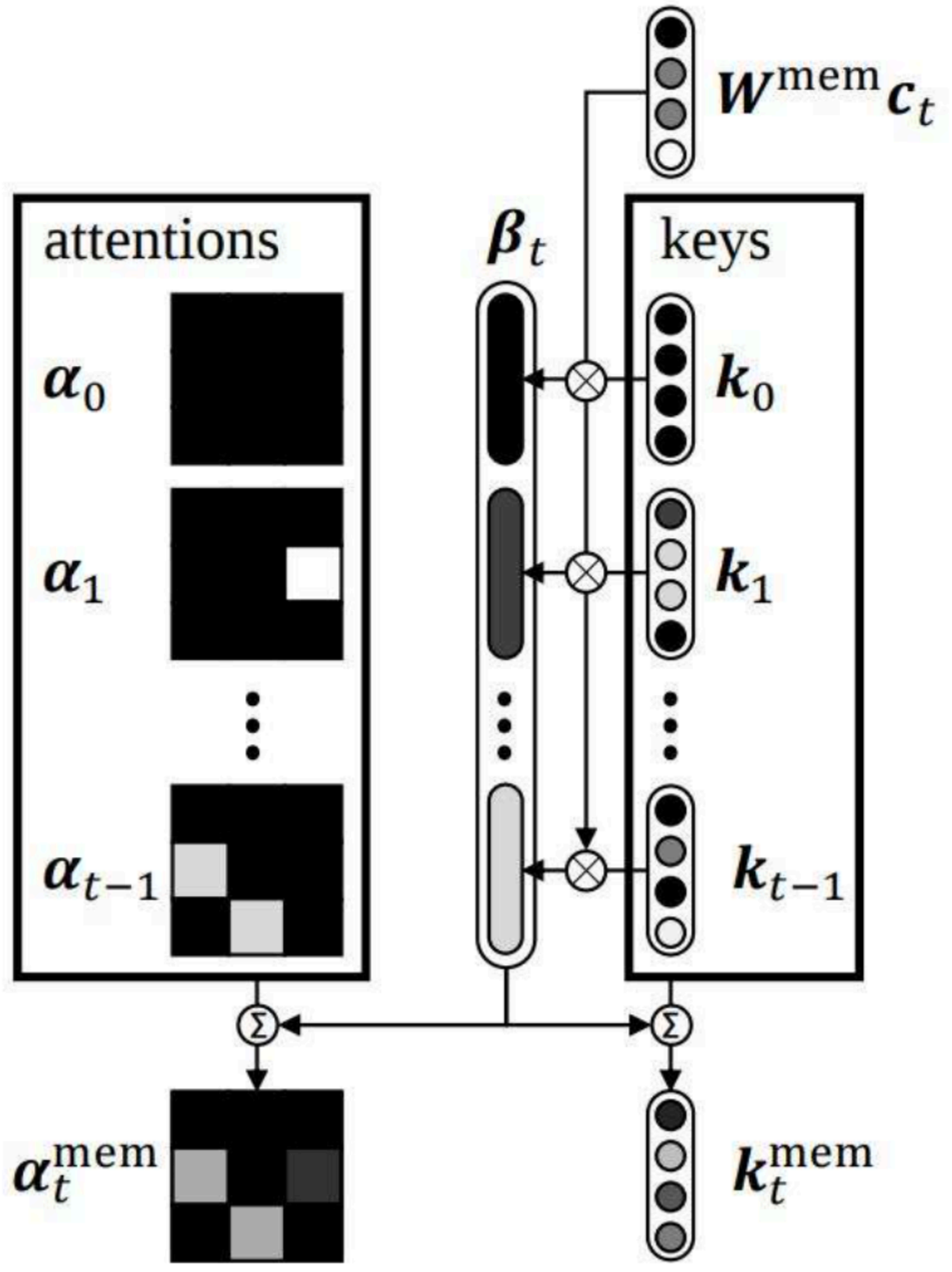
Intuition: More recent questions are likely more relevelant

$$m_{t,\tau} = \left(\boldsymbol{W}^{\text{mem}} \boldsymbol{c}_t\right)^\top \boldsymbol{k}_\tau$$

$$\boldsymbol{\beta}_t = \text{softmax}\left(\{m_{t,\tau}, 0 < \tau < t-1\}\right)$$

$$\boldsymbol{\alpha}_t^{\text{mem}} = \sum_{\tau=0}^{t-1} \beta_{t,\tau} \boldsymbol{\alpha}_\tau$$

Weighted combination
of **attention maps**



Associative Memory Attention

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

Intuition: More recent questions are likely more relevelant

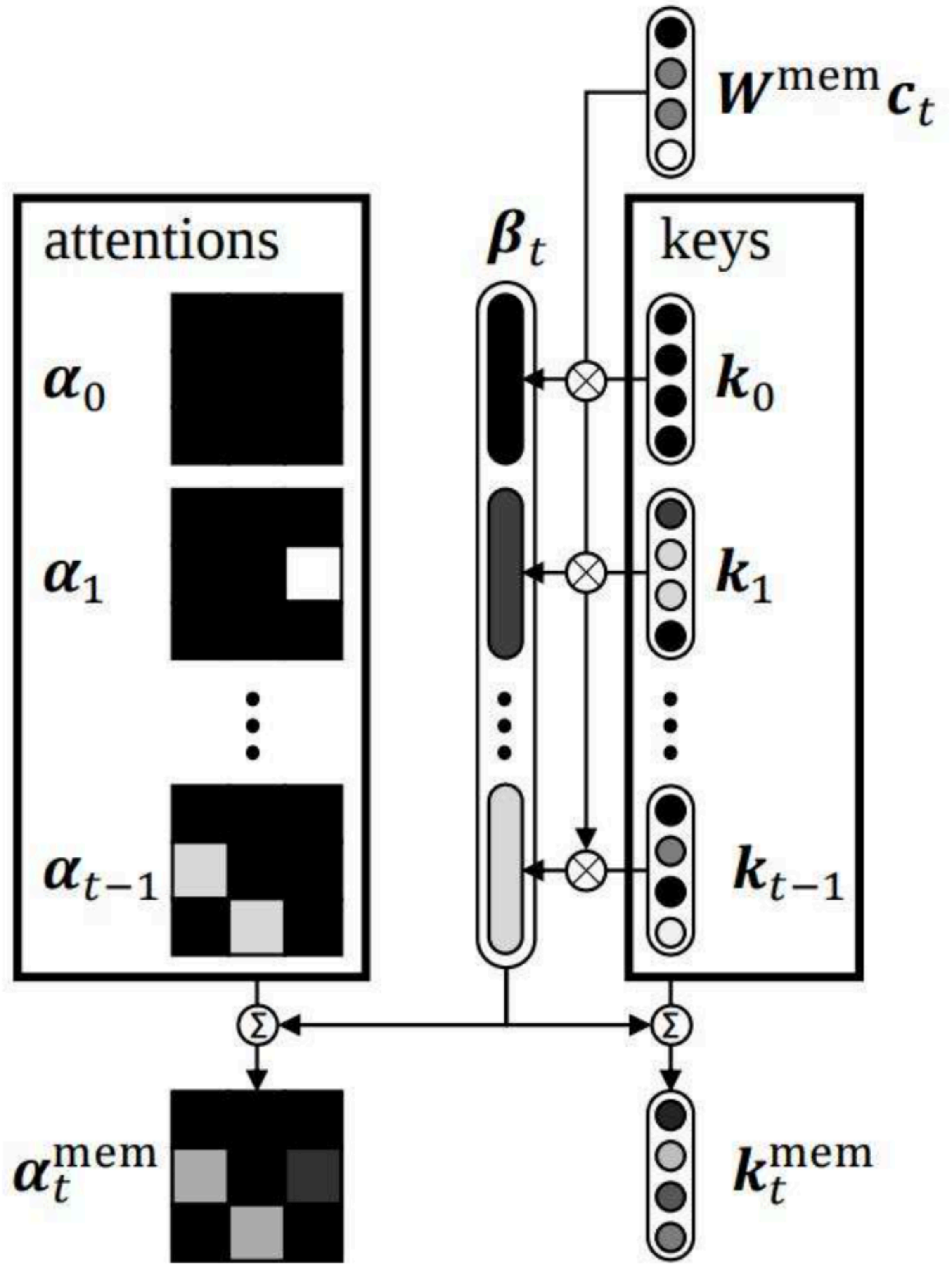
* learnable parameter

$$m_{t,\tau} = \left(\boldsymbol{W}^{\text{mem}}\boldsymbol{c}_t\right)^\top \boldsymbol{k}_\tau \boxed{+\,\theta\,(t-\tau)}$$

$$\boldsymbol{\beta}_t = \text{softmax}\left(\{m_{t,\tau}, 0 < \tau < t-1\}\right)$$

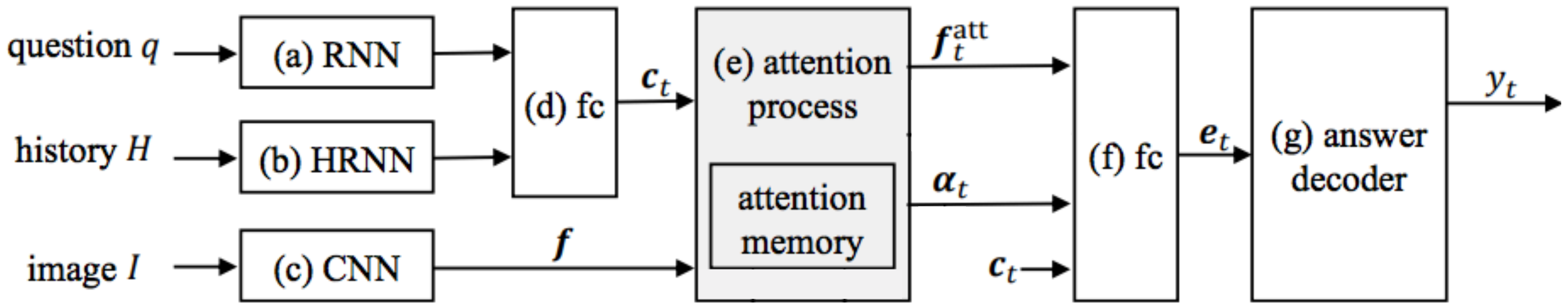
$$\boldsymbol{\alpha}_t^{\text{mem}} = \sum_{\tau=0}^{t-1} \beta_{t,\tau}\boldsymbol{\alpha}_\tau$$

Weighted combination
of **attention maps**



Dynamic Attention Combination

[Seo et al., NIPS 2017]



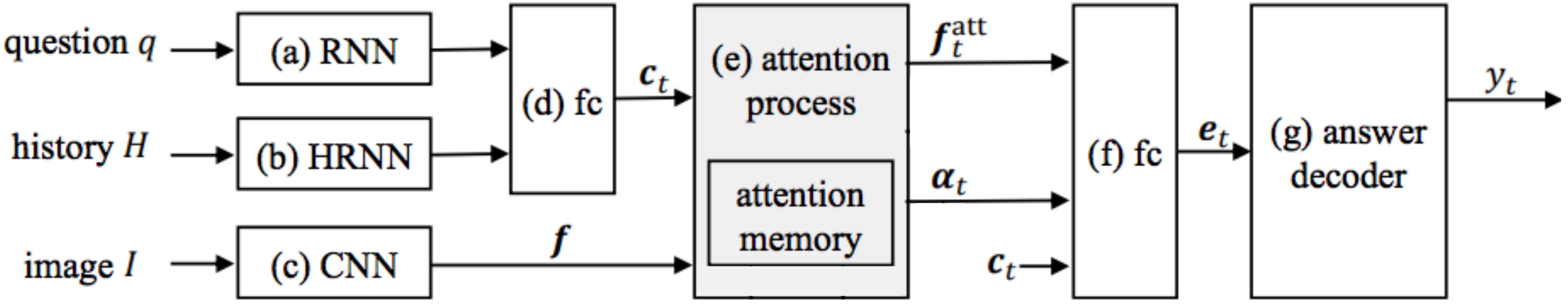
Two types of attention that focus on distinctly different aspect:

— **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)

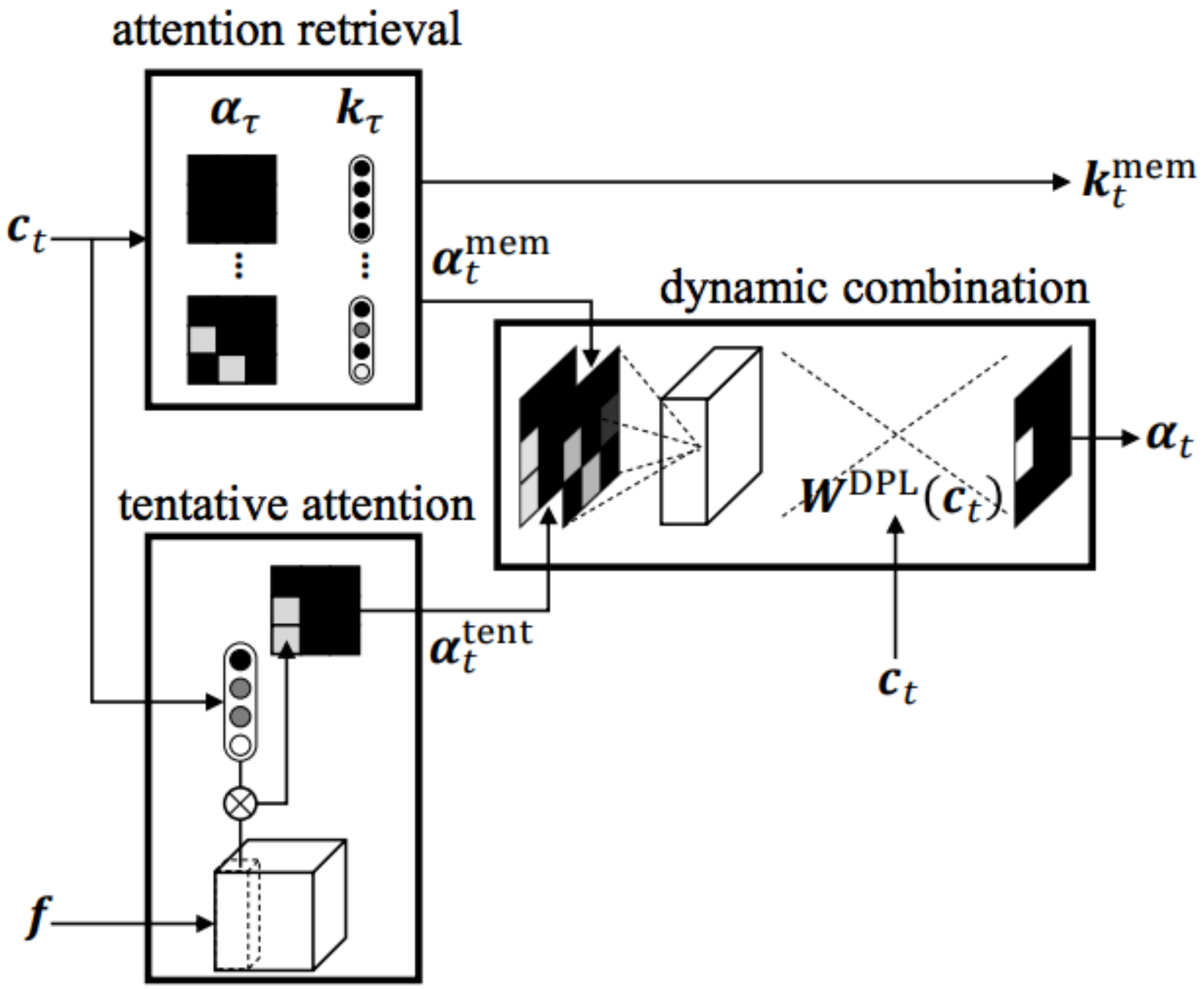
Intuition: We need a dynamic mechanism to fuse these attention models

[Noh et al., CVPR 2016]

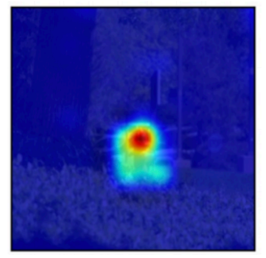
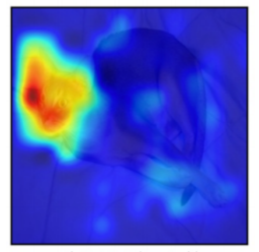
Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]

Q3: What color is it?

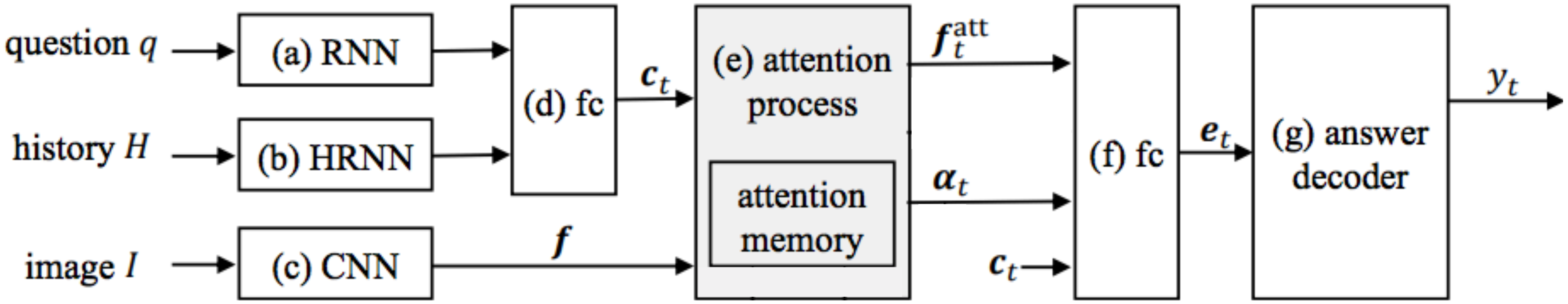


Associative Memory:

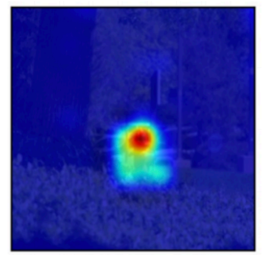
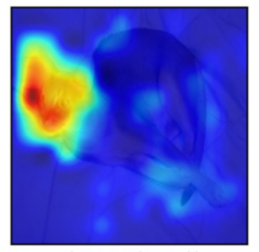
Question Turn	Key (hash)	Memory
1	f (**H:** Empty; **Q:** What color is a hydrant? **A:** It is red)	
2	f (**H:** …; **Q:** Is there a tree? **A:** Yes)	

Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]

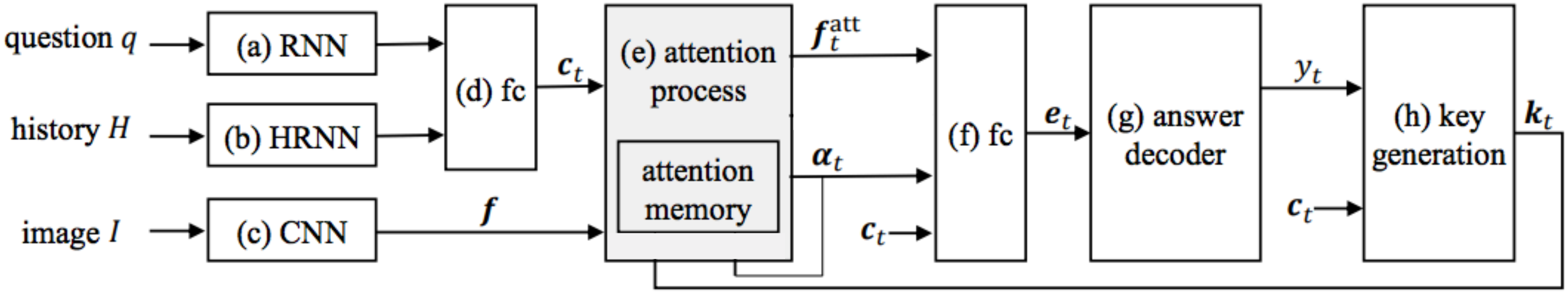


Associative Memory:

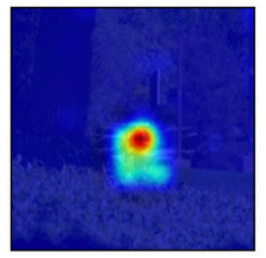
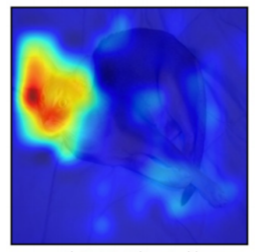
Question Turn	Key (hash)	Memory
1	f (**H:** Empty; **Q:** What color is a hydrant? **A:** It is red)	
2	f (**H:** …; **Q:** Is there a tree? **A:** Yes)	
	????	

Memory Networks for Visual Dialogs

[Seo et al., NIPS 2017]

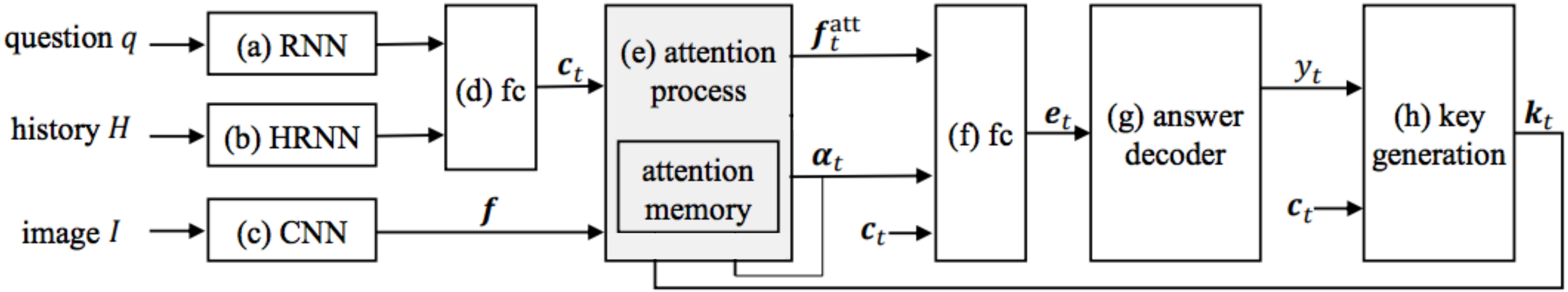


Associative Memory:

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

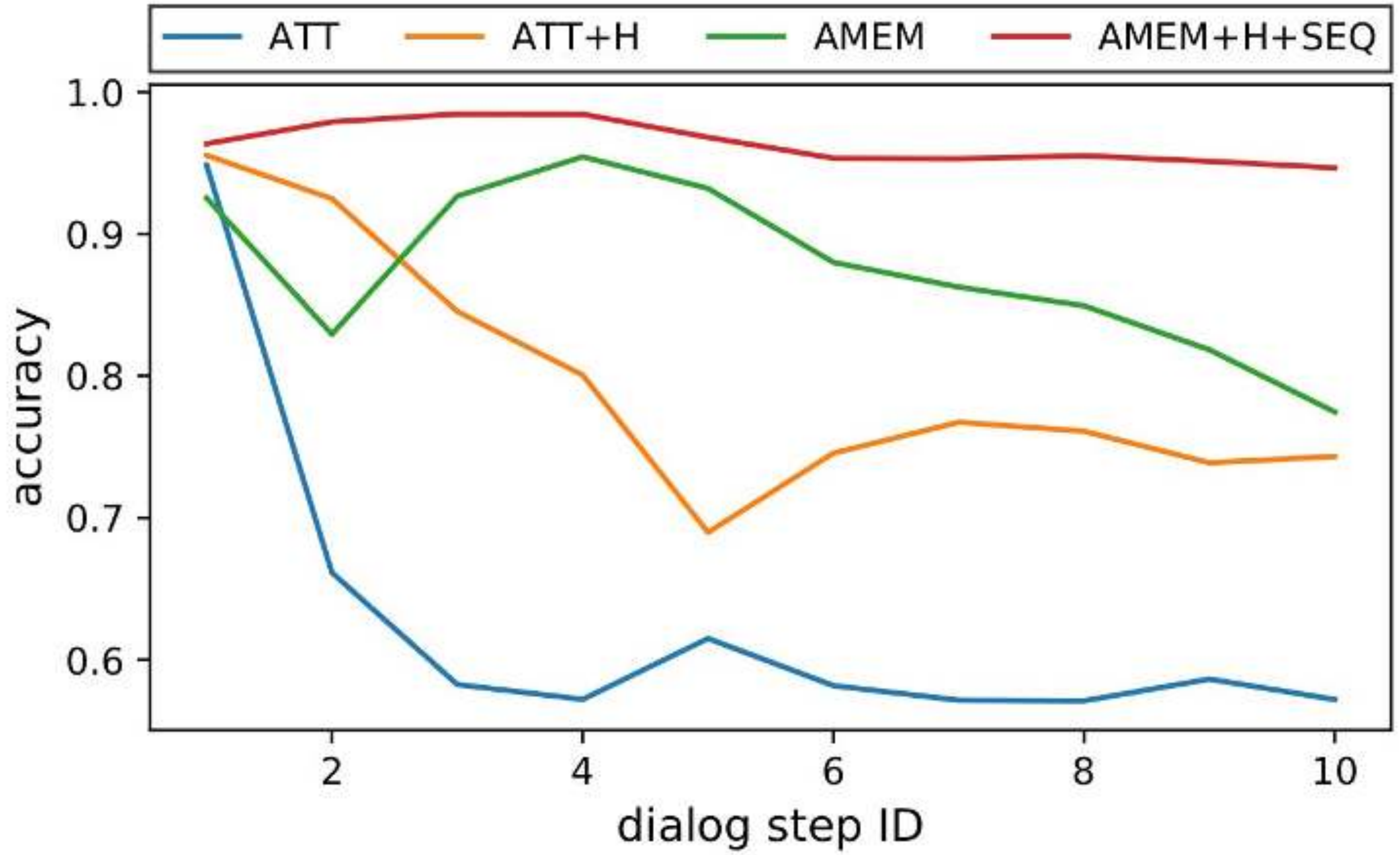
— **Answers:** free form text, 100 candidates

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

Results: MNIST Dialog

[Seo et al., NIPS 2017]

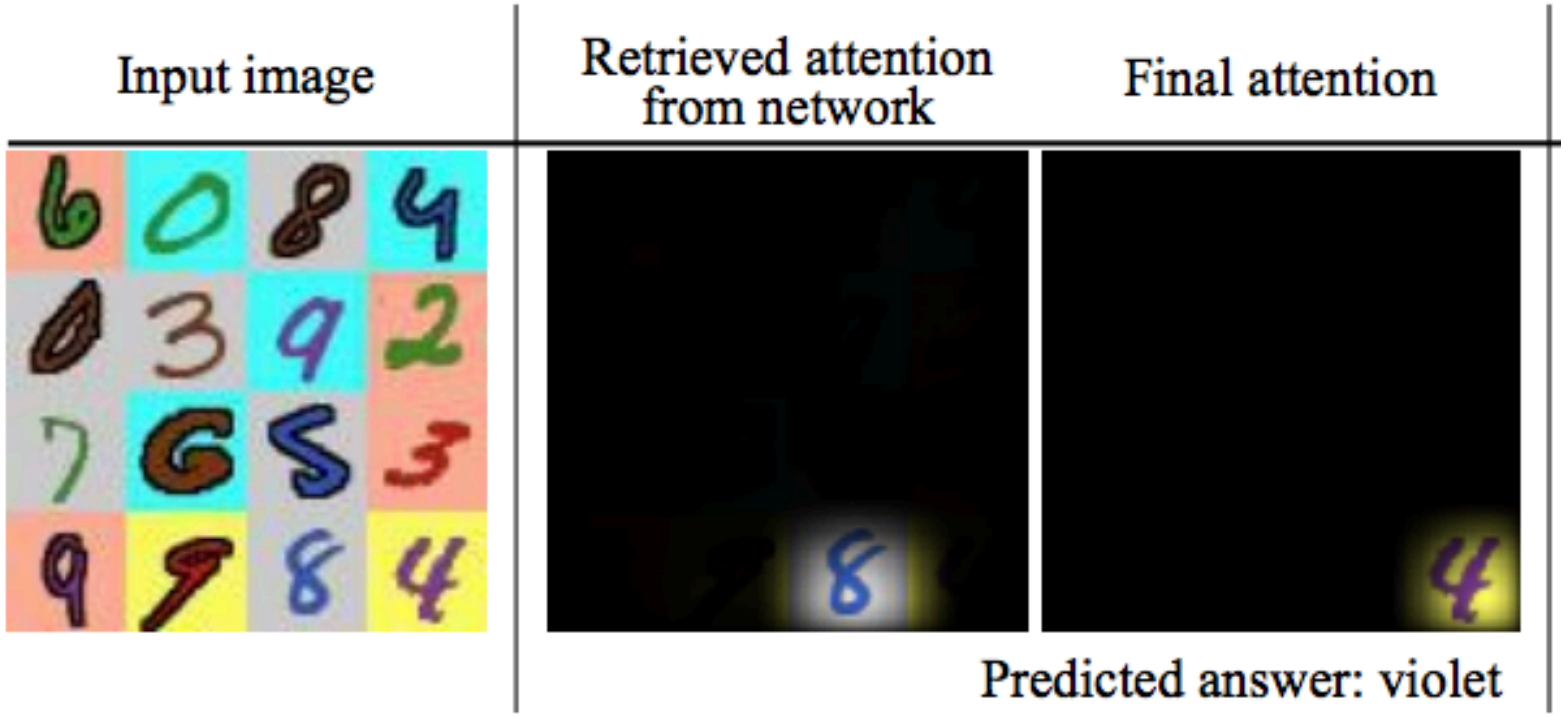
Basemodel	+H	+SEQ	Accuracy
I	–	–	20.18
Q	–	–	36.58
	✓	–	37.58
LF [1]	✓	–	45.06
HRE [1]	✓	–	49.10
MN [1]	✓	–	48.51
ATT	–	–	62.62
	✓	–	79.72
AMEM	–	–	87.53
	✓	–	89.20
	–	✓	90.05
	✓	✓	**96.39**



Results: Interpretability / Implicit Reasoning

[Seo et al., NIPS 2017]

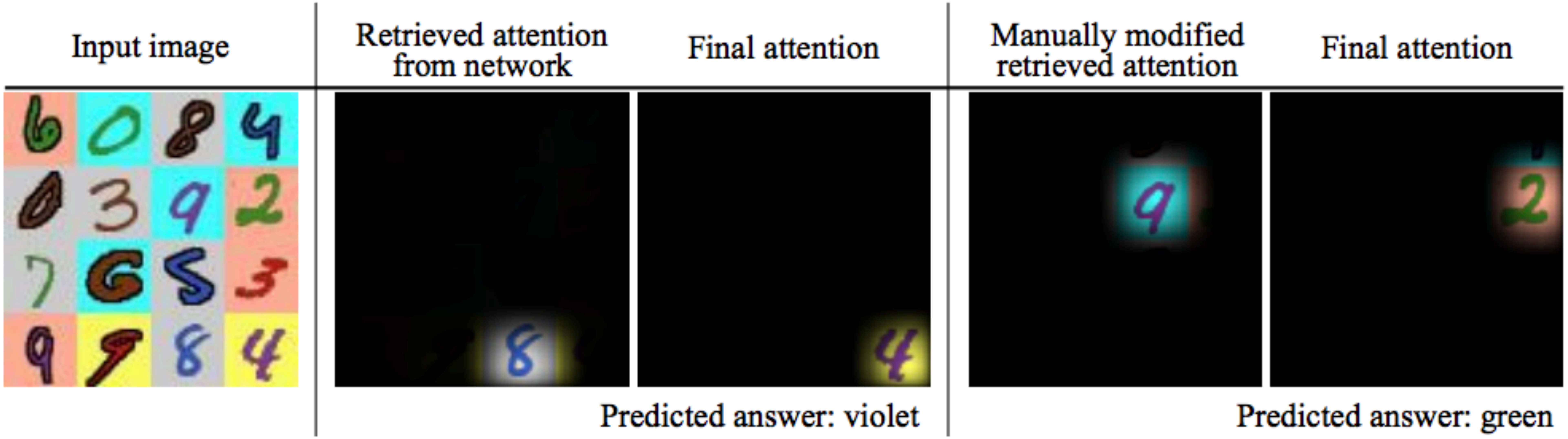
History:	Are there any 9's in the image ?	three
	How many digits in a yellow background are there among them ?	one
	What is the color of the digit ?	red
	What is the color of the digit at the right of it ?	blue
	What is the style of the blue digit ?	flat
Current QA:	What is the color of the digit at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



Results: Interpretability / Implicit Reasoning

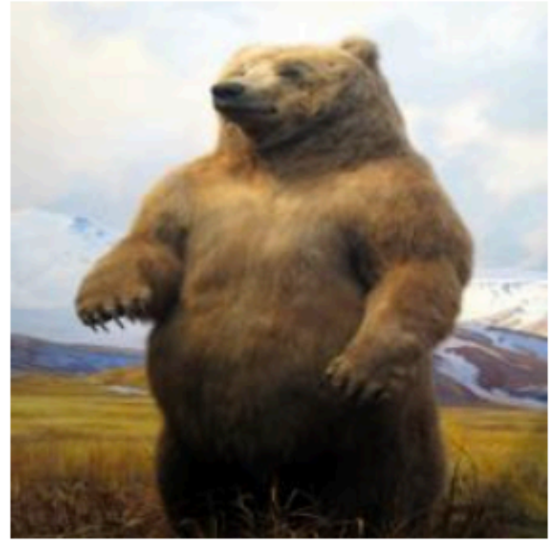
[Seo et al., NIPS 2017]

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]

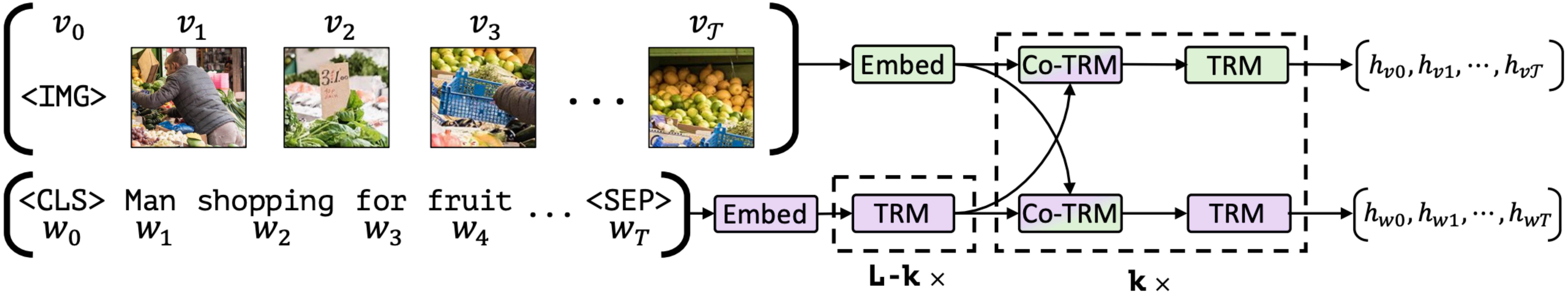
Dialog Information	Input image	Attended image
Caption: *A large bear standing upright with mountains in the background* Previous QA: *Is this the only bear here ? / yes* Current question: *What color is it's fur ?* GT answer: *Brown* Predicted answer: *Brown* Rank of GT: *1*		
Caption: *A train that 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*		

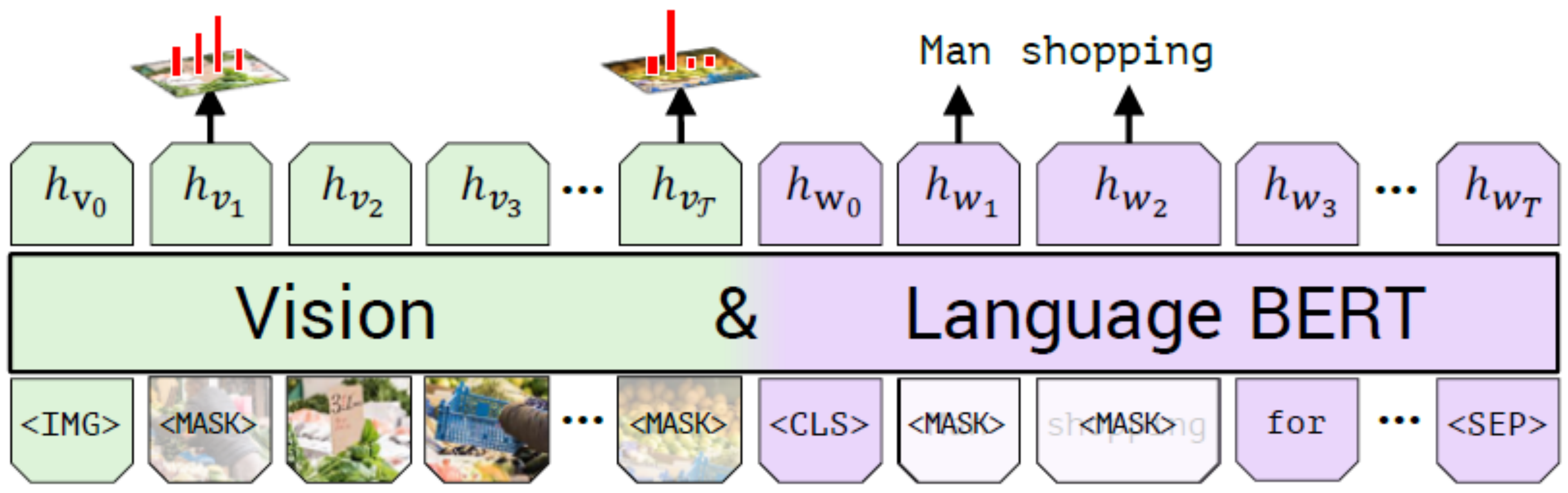
Results: VisDial

[Seo et al., NIPS 2017]

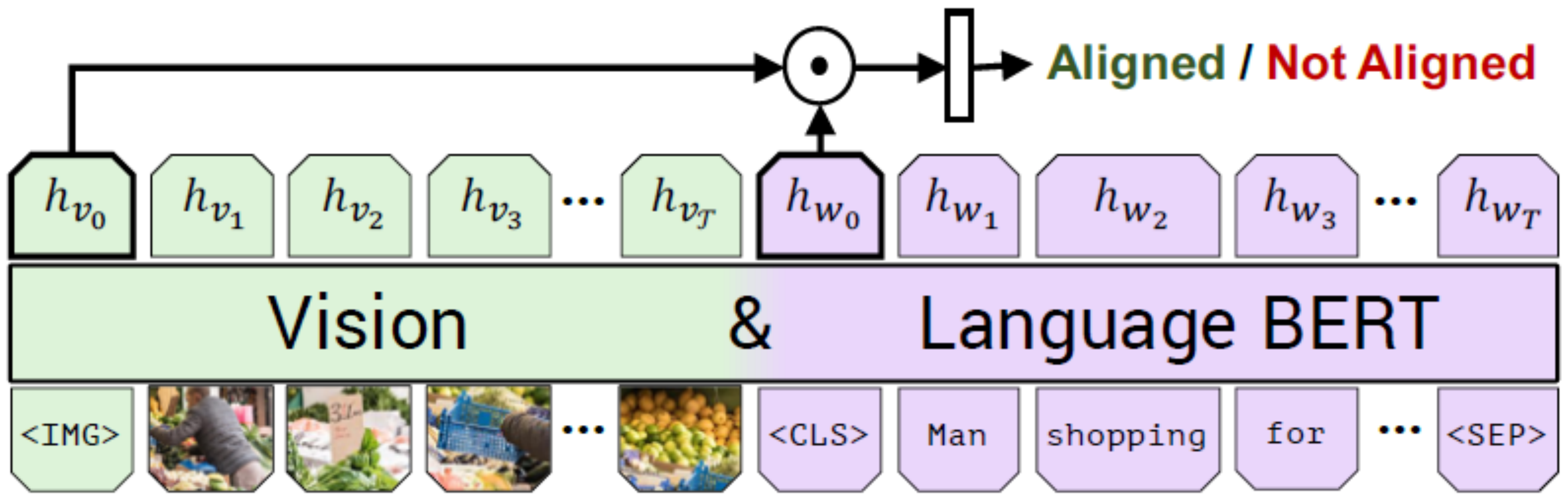
Model	+H	ATT	# of params	MRR	R@1	R@5	R@10	MR
Answer prior [24]	–	–	n/a	0.3735	23.55	48.52	53.23	26.50
LF-Q [24]	–	–	8.3 M (3.6x)	0.5508	41.24	70.45	79.83	7.08
LF-QH [24]	✓	–	12.4 M (5.4x)	0.5578	41.75	71.45	80.94	6.74
LF-QI [24]	–	–	10.4 M (4.6x)	0.5759	43.33	74.27	83.68	5.87
LF-QIH [24]	✓	–	14.5 M (6.3x)	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

Visual BERT (ViIBERT)





(a) Masked multi-modal learning



(b) Multi-modal alignment prediction