

Going beyond Captioning: Visual QA, Visual Dialog

Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad



Review: Question

Can we do the opposite (caption-to-image) of what we learned in the previous lecture? How?

Review: Question

Can we do the opposite (caption-to-image) of what we learned in the previous lecture? How?
Yes, we will see deep generative models soon that can do this

Visual Question Answering (VQA): Task Overview¹



Credit: Aishwarya Agrawal, Devi Parikh, Georgia Tech

¹Agrawal et al, VQA: Visual Question Answering, IJCV 2015

Visual Question Answering (VQA): Task Overview¹

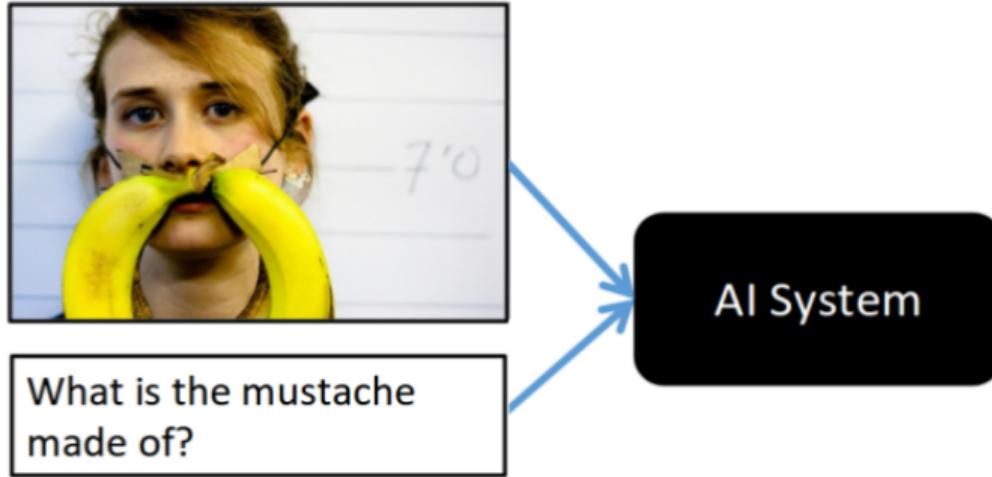


What is the mustache
made of?

Credit: Aishwarya Agrawal, Devi Parikh, Georgia Tech

¹Agrawal et al, VQA: Visual Question Answering, IJCV 2015

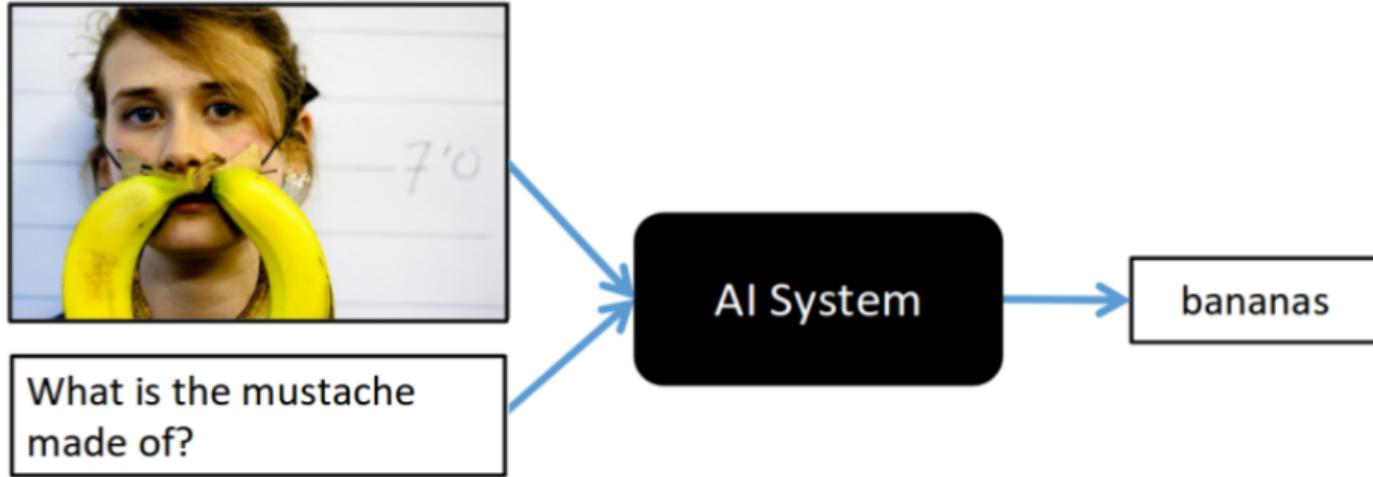
Visual Question Answering (VQA): Task Overview¹



Credit: Aishwarya Agrawal, Devi Parikh, Georgia Tech

¹Agrawal et al, VQA: Visual Question Answering, IJCV 2015

Visual Question Answering (VQA): Task Overview¹



Credit: Aishwarya Agrawal, Devi Parikh, Georgia Tech

¹Agrawal et al, VQA: Visual Question Answering, IJCV 2015

VQA CloudCV Demo



Where is the kid sitting?

Submit

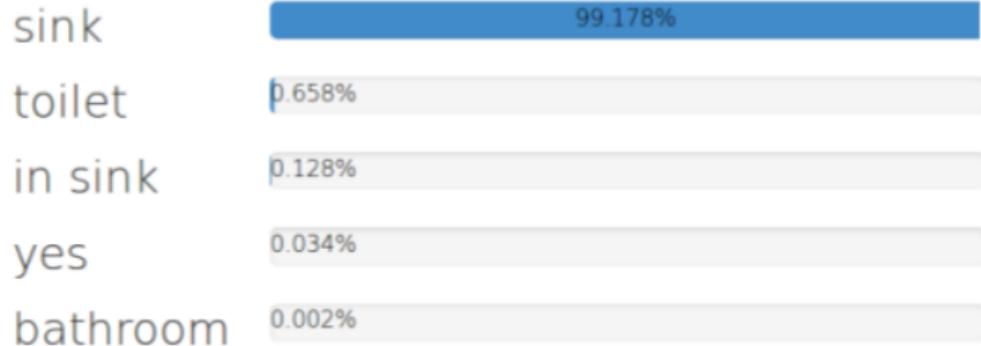
VQA CloudCV Demo



Where is the kid sitting?

Submit

Predicted top-5 answers with confidence:



Feel free to try VQA on your images on this [link](#)!

VQA Dataset²



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

²Agrawal et al, VQA: Visual Question Answering, IJCV 2015

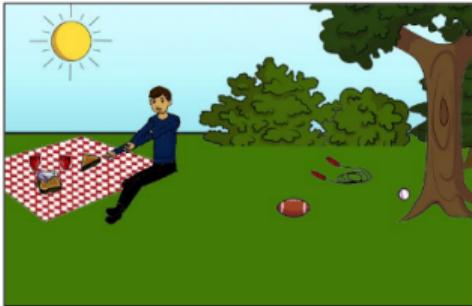
VQA Dataset²



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

- **Open-ended answers and Multiple-choice answers**

²Agrawal et al, VQA: Visual Question Answering, IJCV 2015

VQA Dataset²



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

- **Open-ended answers and Multiple-choice answers**
- 250K images (MS COCO + 50K abstract images)

²Agrawal et al, VQA: Visual Question Answering, IJCV 2015

VQA Dataset²



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

- **Open-ended answers and Multiple-choice answers**
- $250K$ images (MS COCO + $50K$ abstract images)
- $750K$ questions, $10M$ answers

²Agrawal et al, VQA: Visual Question Answering, IJCV 2015

VQA Dataset²



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

- **Open-ended answers and Multiple-choice answers**
- $250K$ images (MS COCO + $50K$ abstract images)
- $750K$ questions, $10M$ answers
- Each question is answered by 10 human annotators

²Agrawal et al, VQA: Visual Question Answering, IJCV 2015

COCO-QA³



COCOQA 5078

How many leftover donuts is the red bicycle holding?

Ground truth: three



COCOQA 1238

What is the color of the tee-shirt?

Ground truth: blue



COCOQA 26088

Where is the gray cat sitting?

Ground truth: window

³Ren et al, Exploring Models and Data for Image Question Answering, NeurIPS 2015

COCO-QA³



COCOQA 5078

How many leftover donuts is the red bicycle holding?

Ground truth: three



COCOQA 1238

What is the color of the tee-shirt?

Ground truth: blue



COCOQA 26088

Where is the gray cat sitting?

Ground truth: window

- Automatically generate QA pairs with MS COCO captions
- $118K$ QA pairs on $123K$ images
- 4 types of questions: **What object, How many, What color, Where**

³Ren et al, Exploring Models and Data for Image Question Answering, NeurIPS 2015

Visual7W⁴



Q: What endangered animal is featured on the truck?

- A: A bald eagle.
- A: A sparrow.
- A: A humming bird.
- A: A raven.



Q: Where will the driver go if turning right?

- A: Onto 24 1/2 Rd.
- A: Onto 25 1/4 Rd.
- A: Onto 23 1/4 Rd.
- A: Onto Main Street.



Q: When was the picture taken?

- A: During a wedding.
- A: During a bar mitzvah.
- A: During a funeral.
- A: During a Sunday church service.

- 7W stands for what, where, when, who, why, how and

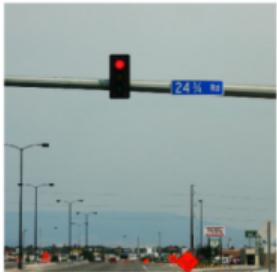
⁴Zhu et al, Visual7W: Grounded Question Answering in Image, CVPR 2016

Visual7W⁴



Q: What endangered animal is featured on the truck?

- A: A bald eagle.
- A: A sparrow.
- A: A humming bird.
- A: A raven.



Q: Where will the driver go if turning right?

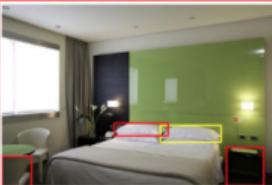
- A: Onto 24 1/2 Rd.
- A: Onto 25 1/4 Rd.
- A: Onto 23 3/4 Rd.
- A: Onto Main Street.



Q: When was the picture taken?

- A: During a wedding.
- A: During a bar mitzvah.
- A: During a funeral.
- A: During a Sunday church service.

- 7W stands for what, where, when, who, why, how and which



Q: Which pillow is farther from the window?



Q: Which step leads to the tub?



Q: Which is the small computer in the corner?

⁴Zhu et al, Visual7W: Grounded Question Answering in Image, CVPR 2016

Visual7W⁴



Q: What endangered animal is featured on the truck?

A: A bald eagle.
A: A sparrow.
A: A humming bird.
A: A raven.



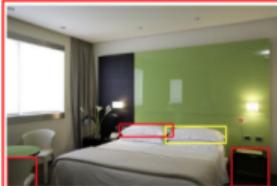
Q: Where will the driver go if turning right?

A: Onto 24 1/2 Rd.
A: Onto 25 1/4 Rd.
A: Onto 23 1/4 Rd.
A: Onto Main Street.



Q: When was the picture taken?

A: During a wedding.
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.



Q: Which pillow is farther from the window?



Q: Which step leads to the tub?



Q: Which is the small computer in the corner?

- 7W stands for **what, where, when, who, why, how and which**
- 328K QA pairs on $\sim 47K$ images

⁴Zhu et al, Visual7W: Grounded Question Answering in Image, CVPR 2016

Visual7W⁴



Q: What endangered animal is featured on the truck?

A: A bald eagle.
A: A sparrow.
A: A humming bird.
A: A raven.



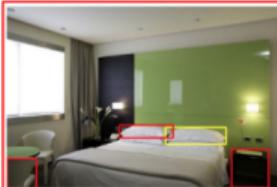
Q: Where will the driver go if turning right?

A: Onto 24 ½ Rd.
A: Onto 25 ¼ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.



Q: When was the picture taken?

A: During a wedding.
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.



Q: Which pillow is farther from the window?



Q: Which step leads to the tub?

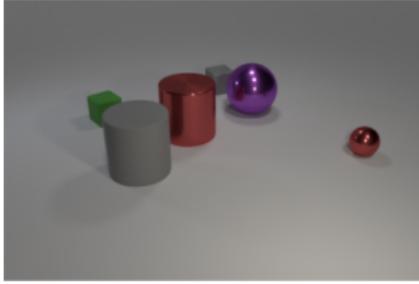


Q: Which is the small computer in the corner?

- 7W stands for **what, where, when, who, why, how** and **which**
- 328K QA pairs on $\sim 47K$ images
- Two types of tasks: **telling** and **pointing**

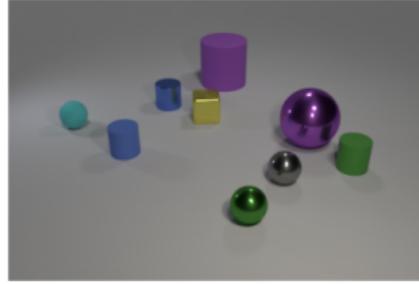
⁴Zhu et al, Visual7W: Grounded Question Answering in Image, CVPR 2016

CLEVR⁵



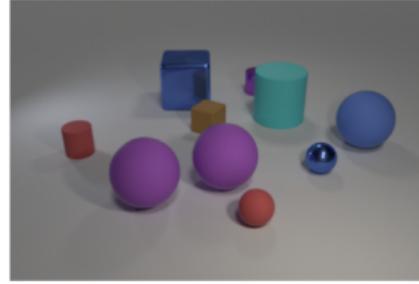
Q: How big is the gray rubber object that is behind the big shiny thing behind the big metallic thing that is on the left side of the purple ball?
A: small
Q-type: query_size
Size: 9

Q: There is a purple ball that is the same size as the red cylinder; what material is it?
A: metal
Q-type: query_material
Size: 17



Q: There is a tiny rubber thing that is the same color as the metal cylinder; what shape is it?
A: cylinder
Q-type: query_shape
Size: 9

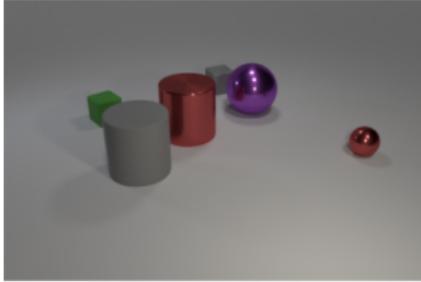
Q: What is the shape of the tiny green thing that is made of the same material as the large cylinder?
A: cylinder
Q-type: query_shape
Size: 9



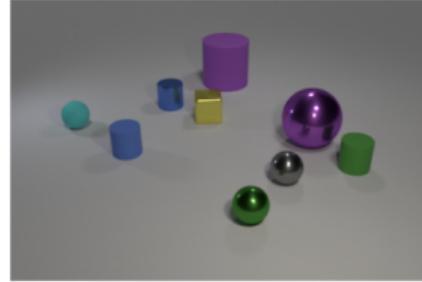
Q: Is the size of the ball that is made of red rubber the same as the purple large block; what metal thing? color is it?
A: yes
Q-type: equal_size
Q-type: query_color
Size: 12
Size: 9

⁵ Johnson et al, CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, CVPR 2017

CLEVR⁵

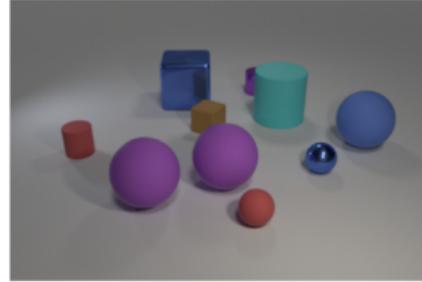


Q: How big is the gray rubber object that is behind the big shiny thing behind the big metallic thing that is on the left side of the purple ball?
A: small
Q-type: query_size
Size: 9



Q: There is a tiny rubber thing that is the same color as the metal cylinder; what shape is it?
A: cylinder
Q-type: query_shape
Size: 9

Q: What is the shape of the tiny green thing that is made of the same material as the large cylinder?
A: cylinder
Q-type: query_shape
Size: 9

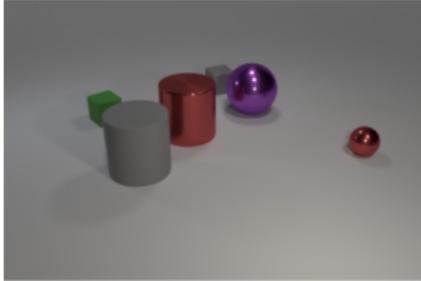


Q: Is the size of the ball that is made of red rubber the same as the purple metal thing?
A: yes
Q-type: equal_size
Q-type: query_color
Size: 9

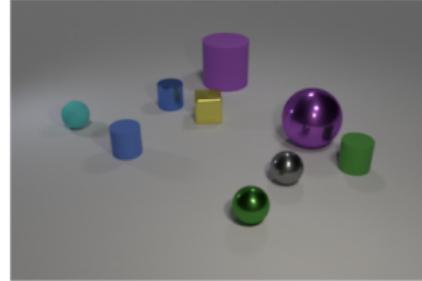
- 100K **rendered images** and 1M **automatically generated questions**

⁵Johnson et al, CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, CVPR 2017

CLEVR⁵

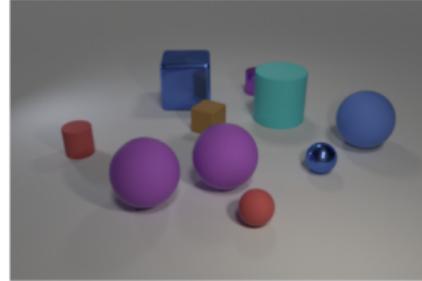


Q: How big is the gray rubber object that is behind the big shiny thing behind the big metallic thing that is on the left side of the purple ball?
A: small
Q-type: query_size
Size: 9



Q: There is a tiny rubber thing that is the same color as the metal cylinder; what shape is it?
A: cylinder
Q-type: query_shape
Size: 9

Q: What is the shape of the tiny green thing that is made of the same material as the large cylinder?
A: cylinder
Q-type: query_shape
Size: 9



Q: Is the size of the ball that is made of red rubber the same as the purple metal thing?
A: yes
Q-type: equal_size
Size: 12

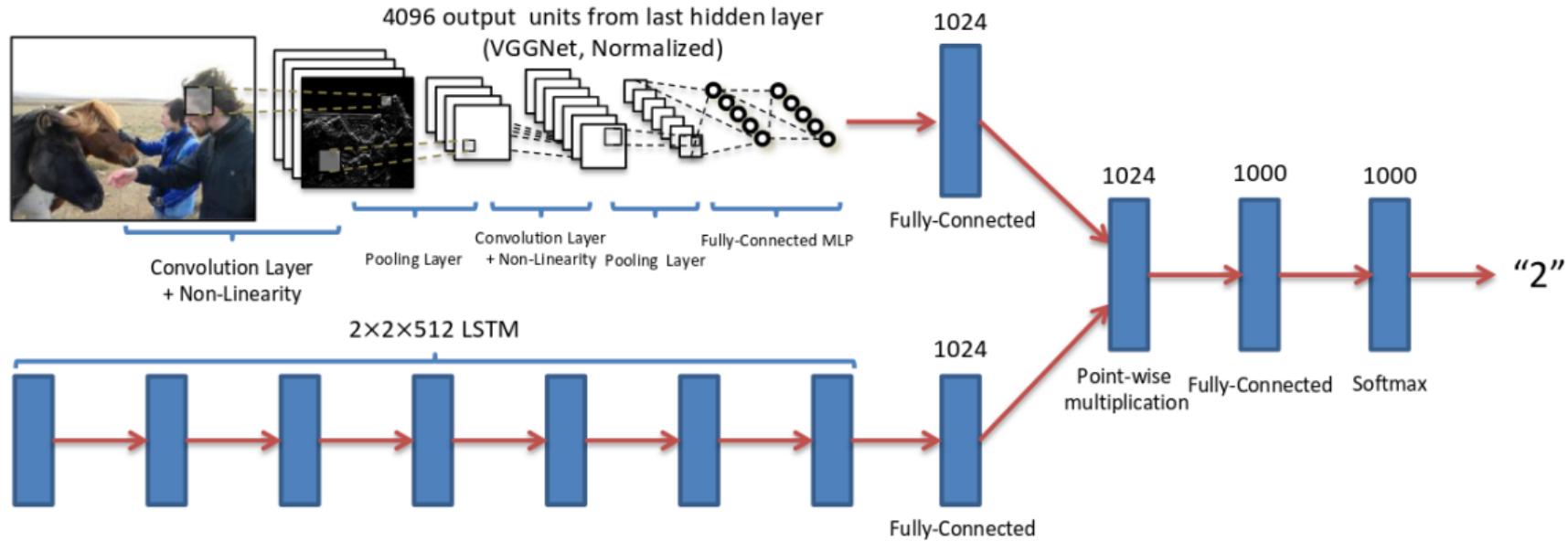
Q: What is the color of the large block; what color is it?
A: blue
Q-type: query_color
Size: 9

- 100K **rendered images** and 1M **automatically generated questions**
- Questions are complex and require reasoning skills

⁵ Johnson et al, CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, CVPR 2017

VQA Models

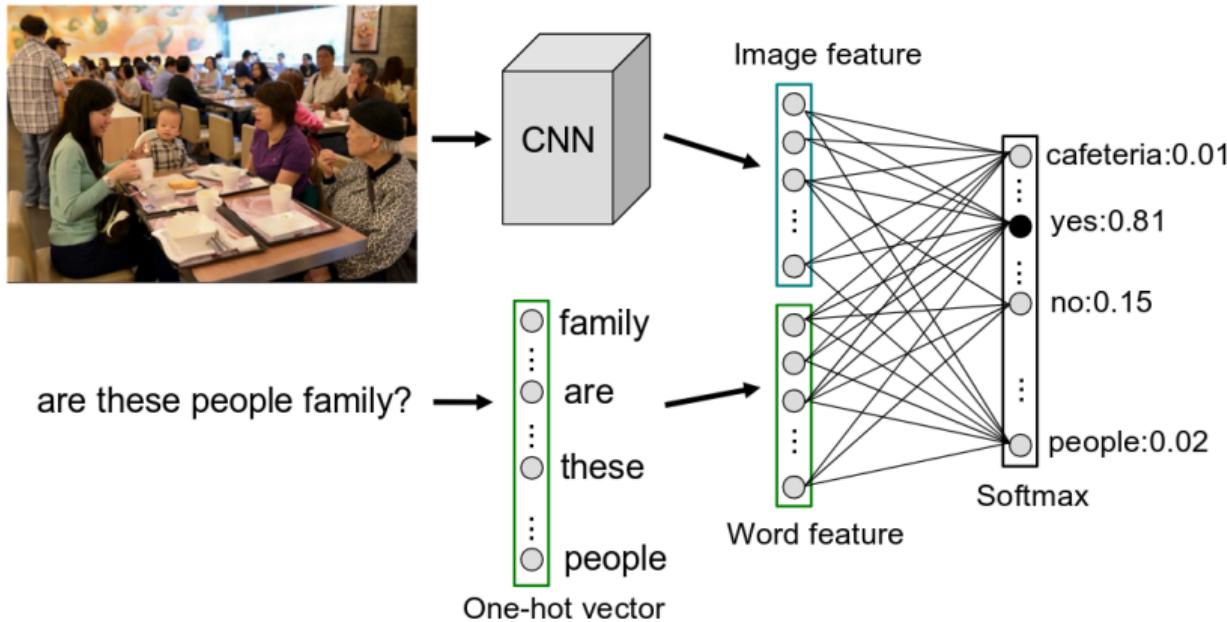
Simple Baseline Model: **LSTM + Image feature**⁶



"How many horses are in this image?"

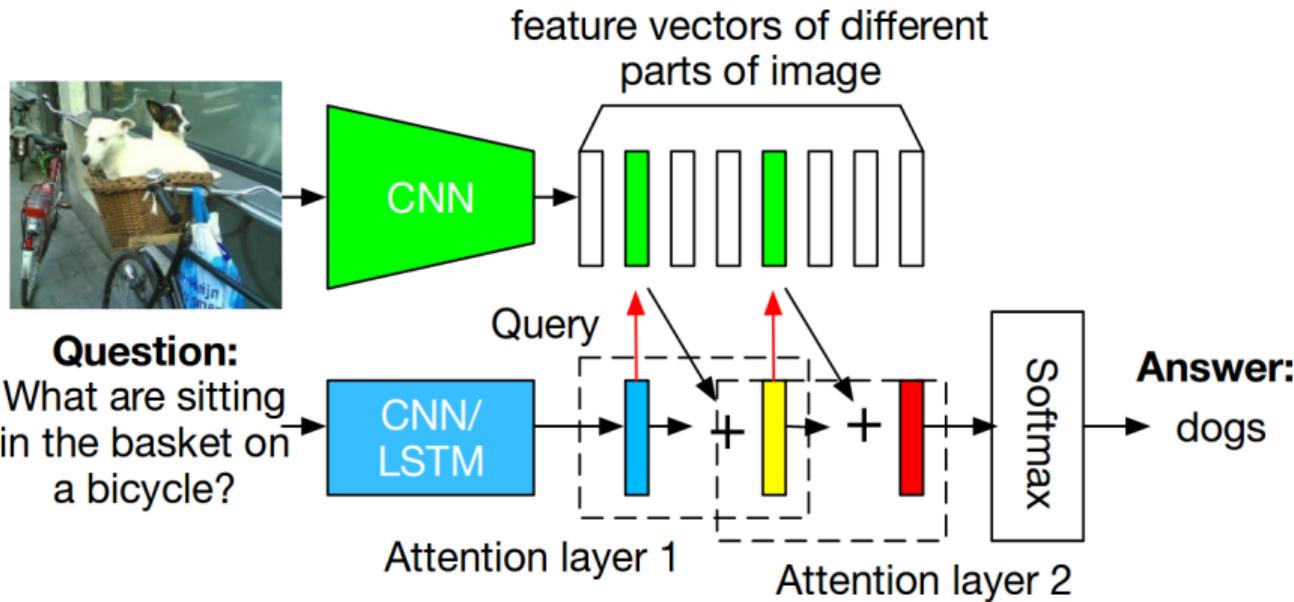
⁶Agrawal et al, VQA: Visual Question Answering, IJCV 2015

VQA Models: Bag-of-words + Image Features (iBOWIMG)⁷



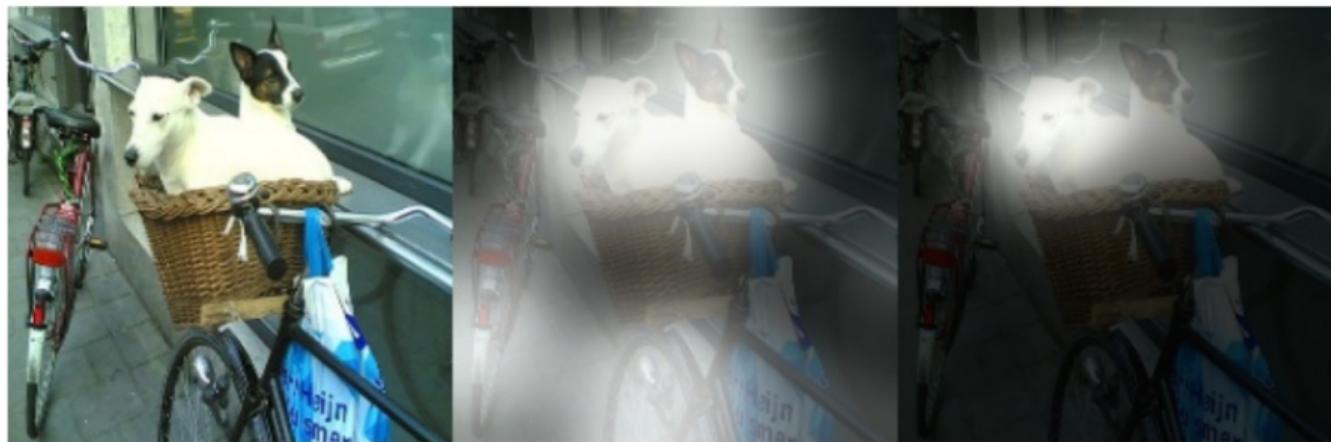
⁷Zhou et al, Simple Baseline for Visual Question Answering, arXiv 2015

VQA Models: Stacked Attention Networks for Image Question Answering⁸



⁸Yang et al, Stacked Attention Networks for Image Question Answering, CVPR 2016

VQA Models: Stacked Attention Networks for Image Question Answering⁸



Original Image First Attention Layer Second Attention Layer

The stacked attention network **first focuses on all referred concepts**, e.g., *bicycle*, *basket* and objects in the basket (*dogs*) in the first attention layer

⁸Yang et al, Stacked Attention Networks for Image Question Answering, CVPR 2016

VQA Models: Stacked Attention Networks for Image Question Answering⁸



Original Image First Attention Layer Second Attention Layer

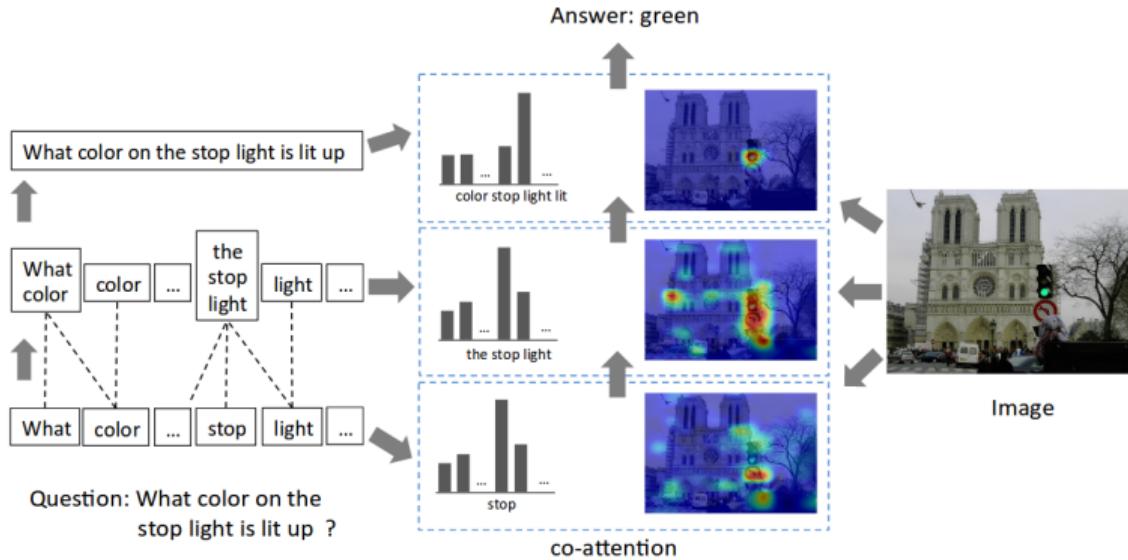
The stacked attention network **first focuses on all referred concepts**, e.g., *bicycle*, *basket* and objects in the basket (*dogs*) in the first attention layer and **then further narrows down the focus in the second layer** and finds out the answer "**dog**".

⁸Yang et al, Stacked Attention Networks for Image Question Answering, CVPR 2016

VQA Models

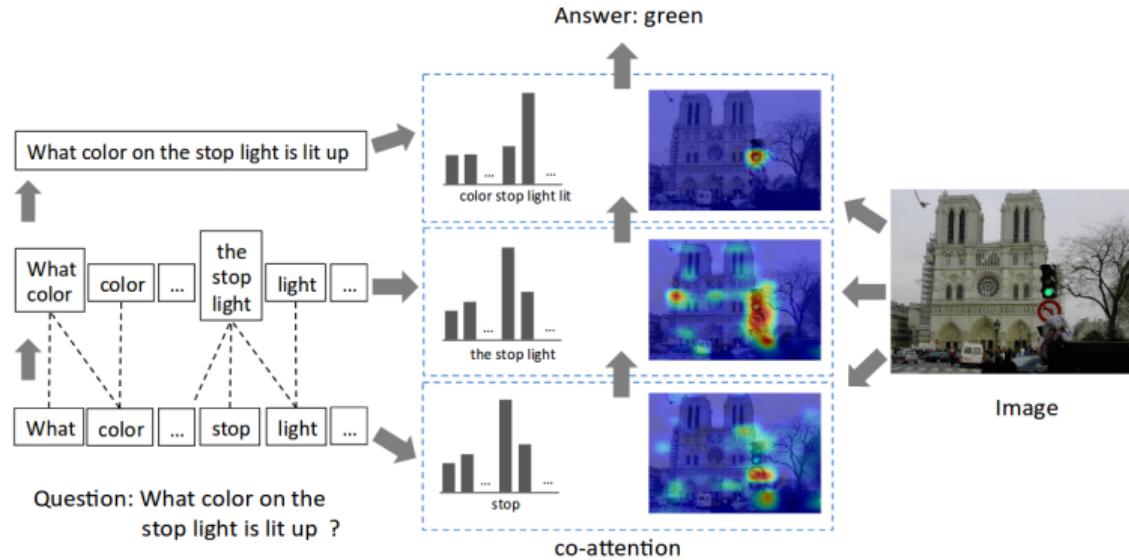
Can we do attention on question as well?

VQA Models: Hierarchical Co-Attention Model⁹



⁹Lu et al, Hierarchical Question-Image Co-Attention for Visual Question Answering, NeurIPS 2016

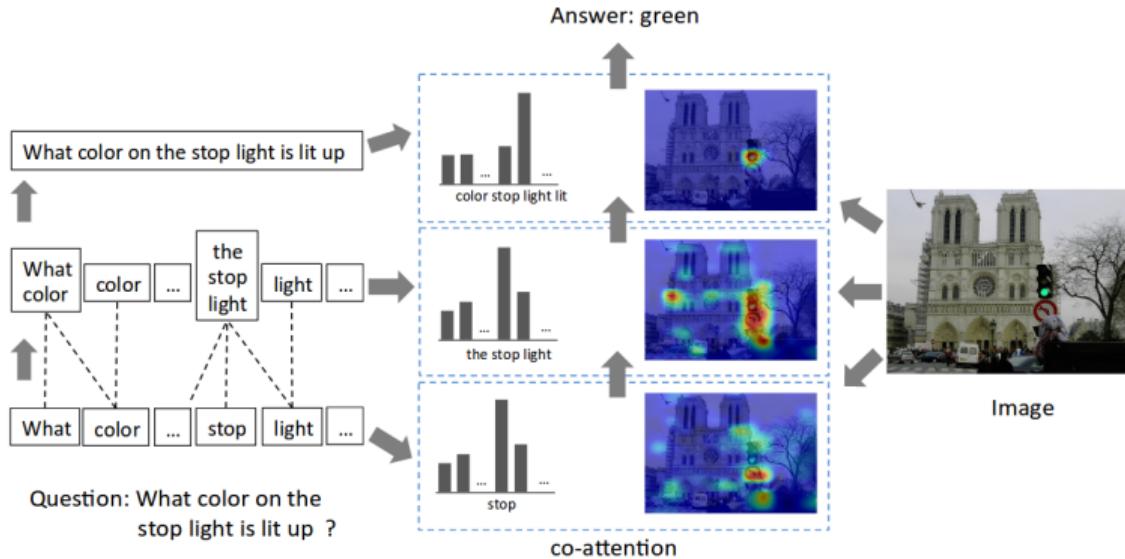
VQA Models: Hierarchical Co-Attention Model⁹



Given a question, extract its word level, phrase level and question level embeddings

⁹Lu et al, Hierarchical Question-Image Co-Attention for Visual Question Answering, NeurIPS 2016

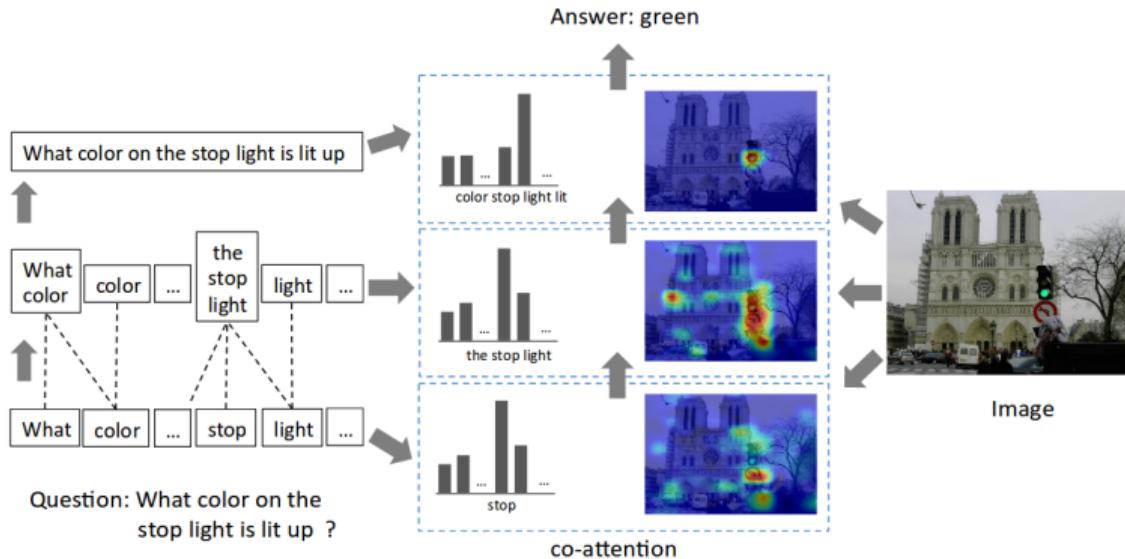
VQA Models: Hierarchical Co-Attention Model⁹



Given a question, extract its word level, phrase level and question level embeddings
At each level, apply co-attention on both image and question

⁹Lu et al, Hierarchical Question-Image Co-Attention for Visual Question Answering, NeurIPS 2016

VQA Models: Hierarchical Co-Attention Model⁹



Given a question, extract its word level, phrase level and question level embeddings

At each level, apply co-attention on both image and question

Final answer prediction is based on all co-attended image and question features

⁹Lu et al, Hierarchical Question-Image Co-Attention for Visual Question Answering, NeurIPS 2016

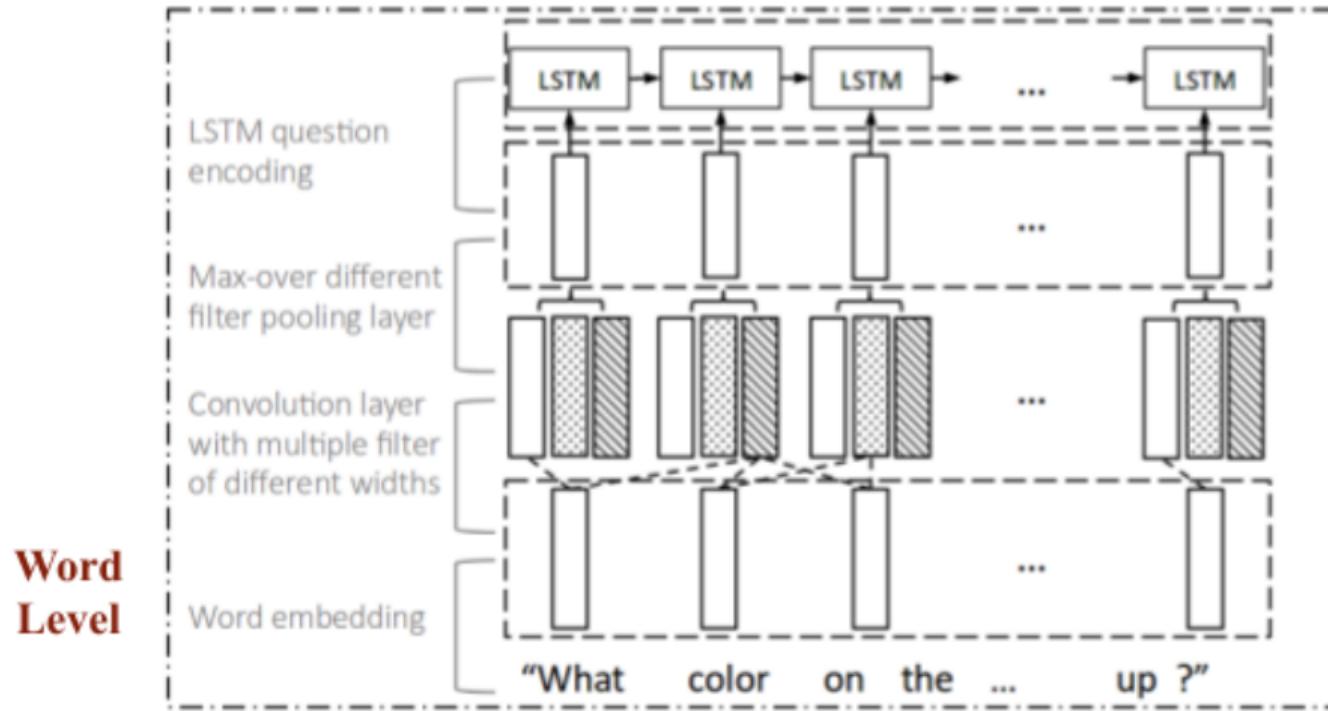
VQA Models: Hierarchical Co-Attention Model

Question hierarchy

1

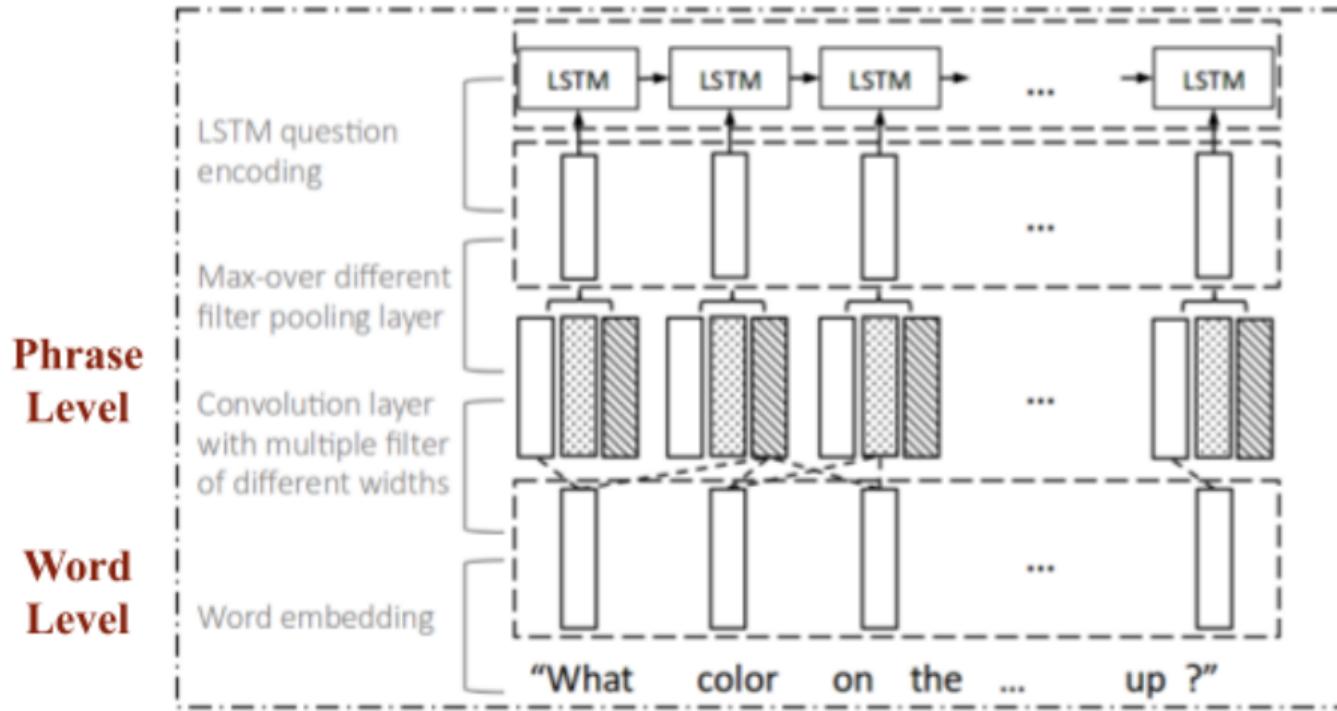
VQA Models: Hierarchical Co-Attention Model

Question hierarchy



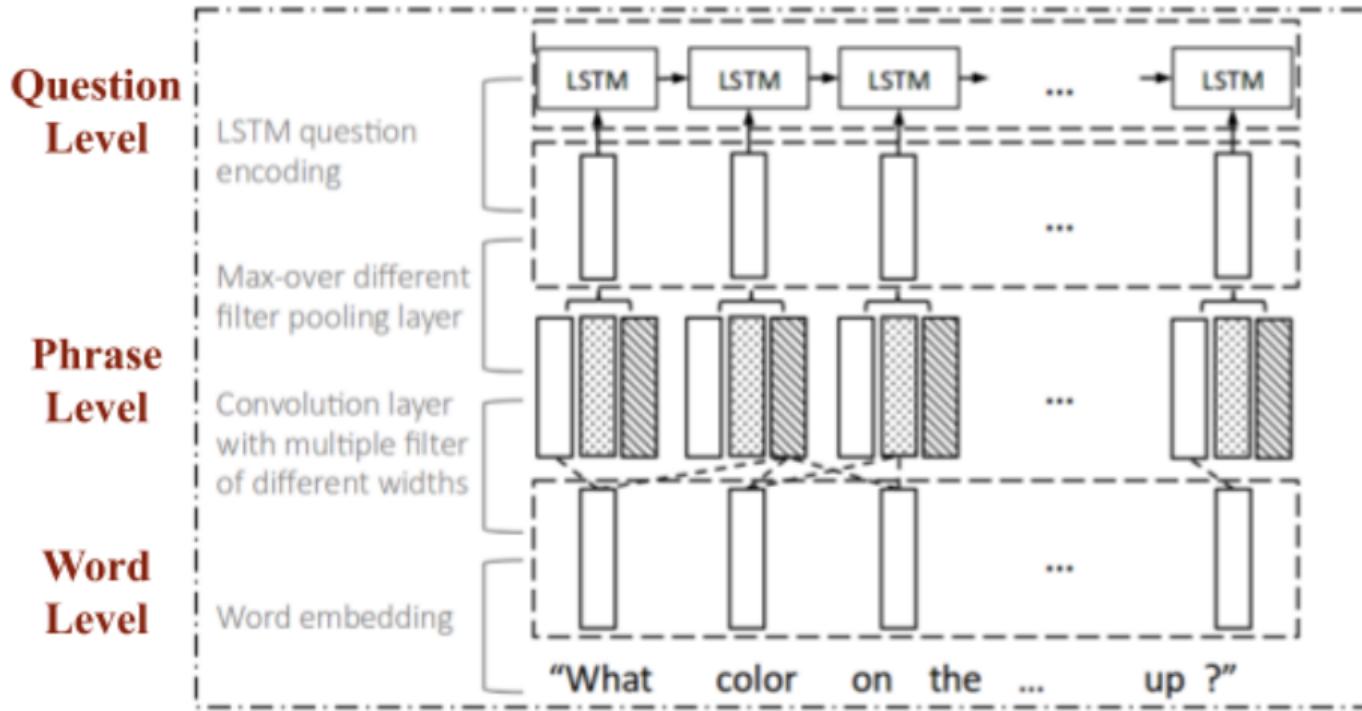
VQA Models: Hierarchical Co-Attention Model

Question hierarchy



VQA Models: Hierarchical Co-Attention Model

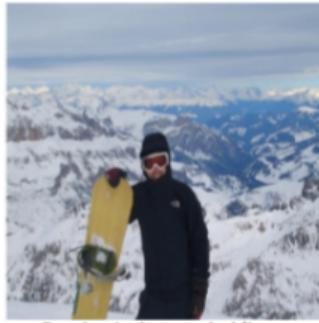
Question hierarchy



VQA Models: Hierarchical Co-Attention Model

Visualization of image and question co-attention map

Image and question
pairs



Q: what is the man holding a
snowboard on top of a snow
covered? **A:** mountain

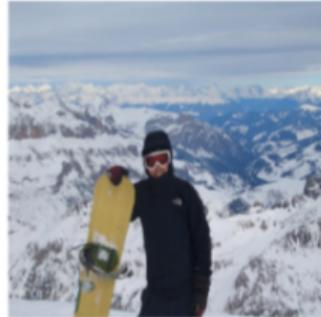


Q: what is the color of the bird? **A:**
white

VQA Models: Hierarchical Co-Attention Model

Visualization of image and question co-attention map

Image and question pairs



Q: what is the man holding a snowboard on top of a snow covered? A: mountain

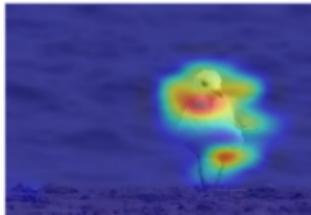
Word level co-attention maps



what is the man holding a snowboard on top of a snow covered



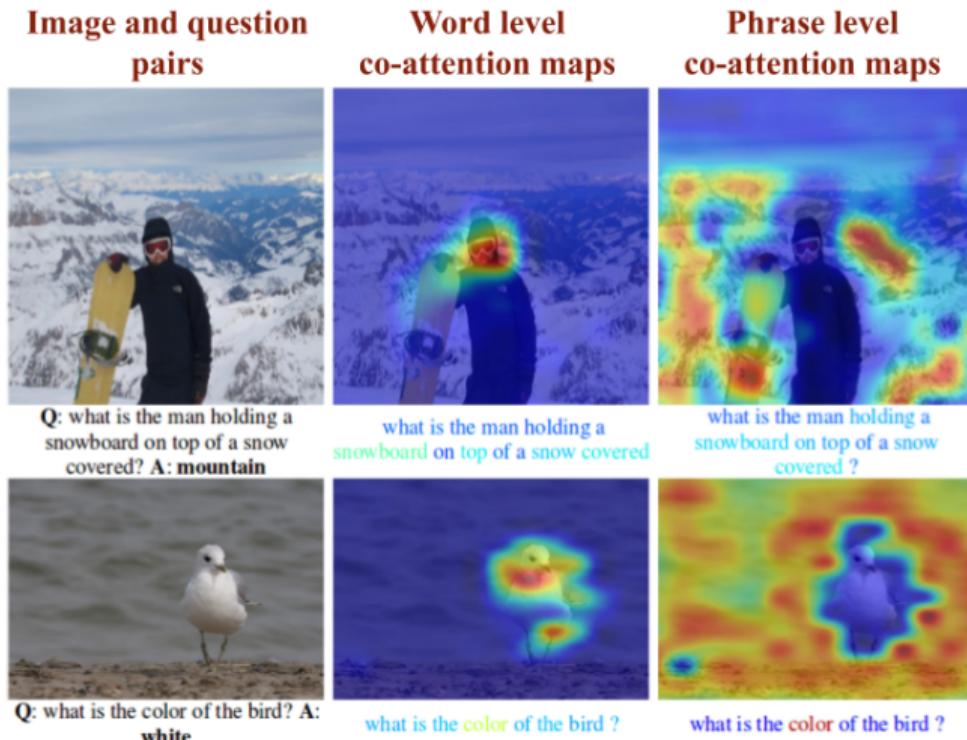
Q: what is the color of the bird? A: white



what is the color of the bird ?

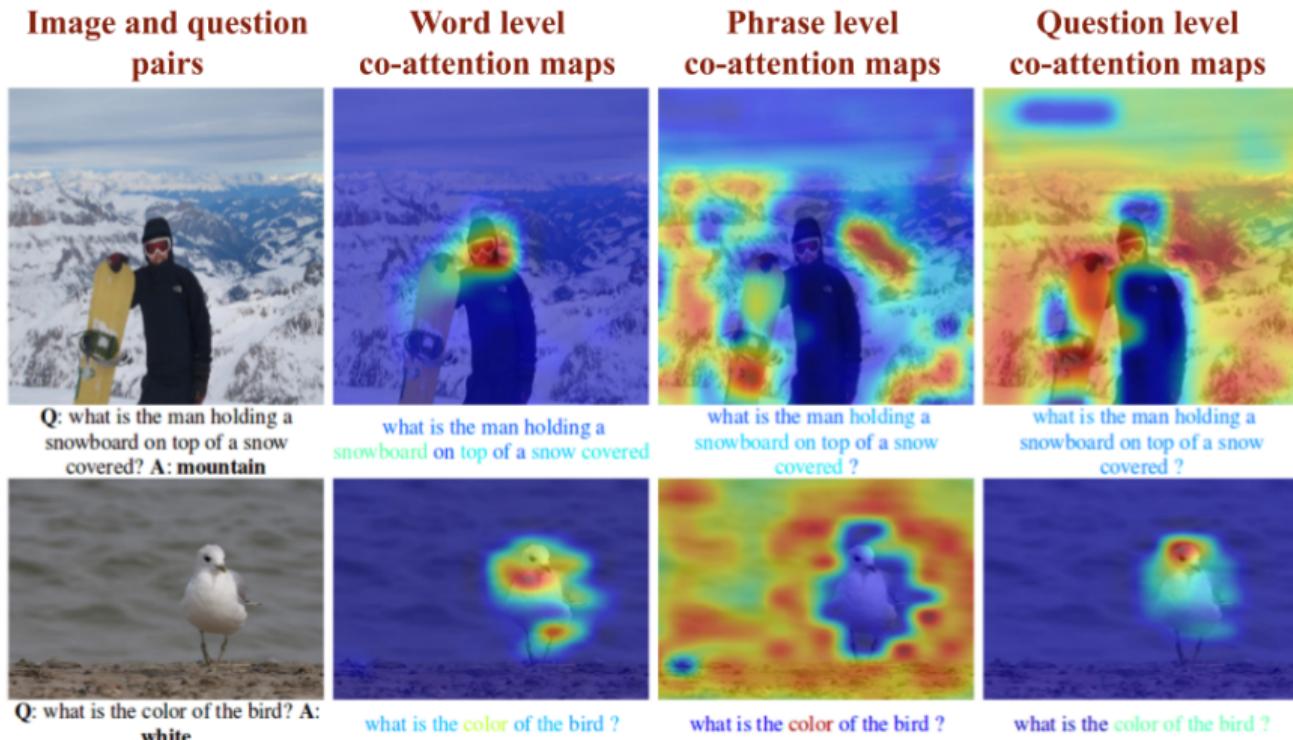
VQA Models: Hierarchical Co-Attention Model

Visualization of image and question co-attention map



VQA Models: Hierarchical Co-Attention Model

Visualization of image and question co-attention map



Visual Dialog: Task Overview¹⁰

Visual Dialog



 White and red

 No, something is there can't tell what it is

 Yes, they are

 Yes, magazines, books, toaster and basket, and a plate





Start typing question here ...



¹⁰Das et al, Visual Dialog, CVPR 2017

Visual Dialog: CloudCV Demo



Visual Dialog: CloudCV Demo



A man and a woman are holding umbrellas

Visual Dialog: CloudCV Demo



A man and a woman are holding umbrellas



What color is his umbrella?

Visual Dialog: CloudCV Demo



A man and a woman are holding umbrellas



His umbrella is black

What color is his umbrella?



Visual Dialog: CloudCV Demo



A man and a woman are holding umbrellas



His umbrella is black

What color is his umbrella?



What about hers?



Visual Dialog: CloudCV Demo



A man and a woman are holding umbrellas



His umbrella is black



Hers is multi-colored

What color is his umbrella?



What about hers?



Visual Dialog: CloudCV Demo



A man and a woman are holding umbrellas



His umbrella is black



Hers is multi-colored

What color is his umbrella?



What about hers?



How many other people are in the image?



Visual Dialog: CloudCV Demo



A man and a woman are holding umbrellas



His umbrella is black



Hers is multi-colored



I think 3. They are occluded

What color is his umbrella?



What about hers?



How many other people are in the image?



You can try Visual Dialog on this [link](#)!

Visual Dialog: Task Description

- **Given**

- *Image I*



Credit: Abhishek Das, Georgia Tech

Visual Dialog: Task Description

- **Given**

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Visual Dialog: Task Description

- **Given**

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Q: Is the other one holding anything?

Visual Dialog: Task Description

- **Given**

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t

- **Task**

- Produce free-form natural language answer A_t



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Q: Is the other one holding anything?

A: He is not

Visual Dialog: Evaluation

- A fundamental challenge in dialog systems is **evaluation**¹¹

¹¹Liu et al, How NOT To Evaluate Your Dialogue System, EMNLP 2016

Visual Dialog: Evaluation

- A fundamental challenge in dialog systems is **evaluation**¹¹
- Existing **word-overlap based metrics** such as BLEU, METEOR, ROUGE are known to correlate poorly with human judgement

¹¹Liu et al, How NOT To Evaluate Your Dialogue System, EMNLP 2016

Visual Dialog: Evaluation

- A fundamental challenge in dialog systems is **evaluation**¹¹
- Existing **word-overlap based metrics** such as BLEU, METEOR, ROUGE are known to correlate poorly with human judgement
- Human Turing test
 - expensive
 - subjective

¹¹Liu et al, How NOT To Evaluate Your Dialogue System, EMNLP 2016

Visual Dialog: Evaluation Protocol

- Given

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Q: Is the other one holding anything?

A: He is not

Credit: Dhruv Batra, Georgia Tech

Visual Dialog: Evaluation Protocol

- Given

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t
- 100 answer options



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Q: Is the other one holding anything?

A: He is not

Credit: Dhruv Batra, Georgia Tech

Visual Dialog: Evaluation Protocol

- Given

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t
- 100 answer options
 - Answers to 50 most similar questions



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Q: Is the other one holding anything?

A: He is not

Credit: Dhruv Batra, Georgia Tech

Visual Dialog: Evaluation Protocol

- Given

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t
- 100 answer options
 - Answers to 50 most similar questions
 - 30 popular answers



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Q: Is the other one holding anything?

A: He is not

Credit: Dhruv Batra, Georgia Tech

Visual Dialog: Evaluation Protocol

- Given

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t
- 100 answer options
 - Answers to 50 most similar questions
 - 30 popular answers
 - 20 random answers



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Q: Is the other one holding anything?

A: He is not

Credit: Dhruv Batra, Georgia Tech

Visual Dialog: Evaluation Protocol

- **Given**

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t
- **100 answer options**
 - Answers to 50 most similar questions
 - 30 popular answers
 - 20 random answers

- **Evaluation Task**



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Q: Is the other one holding anything?

A: He is not

Credit: Dhruv Batra, Georgia Tech

Visual Dialog: Evaluation Protocol

- **Given**

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t
- **100 answer options**
 - Answers to 50 most similar questions
 - 30 popular answers
 - 20 random answers

- **Evaluation Task**

- Rank the list of 100 options



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Q: Is the other one holding anything?

A: He is not

Credit: Dhruv Batra, Georgia Tech

Visual Dialog: Evaluation Protocol

- **Given**

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t
- **100 answer options**
 - Answers to 50 most similar questions
 - 30 popular answers
 - 20 random answers

- **Evaluation Task**

- Rank the list of 100 options

- **Accuracy/Error**



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Q: Is the other one holding anything?

A: He is not

Credit: Dhruv Batra, Georgia Tech

Visual Dialog: Evaluation Protocol

- **Given**

- Image I
- Human dialog history
 $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t
- **100 answer options**
 - Answers to 50 most similar questions
 - 30 popular answers
 - 20 random answers

- **Evaluation Task**

- Rank the list of 100 options

- **Accuracy/Error**

- Mean rank w.r.t. ground truth, Mean reciprocal rank

Credit: Dhruv Batra, Georgia Tech



Q: How many people on wheelchairs?

A: Two

Q: What gender are the people in the wheelchairs?

A: One is female, one is male

Q: Which one is holding the racket?

A: The female

Q: Is the other one holding anything?

A: He is not

Visual Dialog: Models

Encoder-Decoder frameworks

Credit: Abhishek Das, Georgia Tech

Visual Dialog: Models

Encoder-Decoder frameworks

- Encoders

Credit: Abhishek Das, Georgia Tech

Visual Dialog: Models

Encoder-Decoder frameworks

- Encoders
 - Late Fusion Encoder

Credit: Abhishek Das, Georgia Tech

Visual Dialog: Models

Encoder-Decoder frameworks

- Encoders
 - Late Fusion Encoder
 - Hierarchical Recurrent Encoder

Credit: Abhishek Das, Georgia Tech

Visual Dialog: Models

Encoder-Decoder frameworks

- Encoders

- Late Fusion Encoder
- Hierarchical Recurrent Encoder
- Memory Network Encoder

Credit: Abhishek Das, Georgia Tech

Visual Dialog: Models

Encoder-Decoder frameworks

- Encoders
 - Late Fusion Encoder
 - Hierarchical Recurrent Encoder
 - Memory Network Encoder
- Decoders

Credit: Abhishek Das, Georgia Tech

Visual Dialog: Models

Encoder-Decoder frameworks

- Encoders
 - Late Fusion Encoder
 - Hierarchical Recurrent Encoder
 - Memory Network Encoder
- Decoders
 - Generative

Credit: Abhishek Das, Georgia Tech

Visual Dialog: Models

Encoder-Decoder frameworks

- Encoders
 - Late Fusion Encoder
 - Hierarchical Recurrent Encoder
 - Memory Network Encoder
- Decoders
 - Generative
 - Discriminative

Credit: Abhishek Das, Georgia Tech

Visual Dialog: Late Fusion Encoder



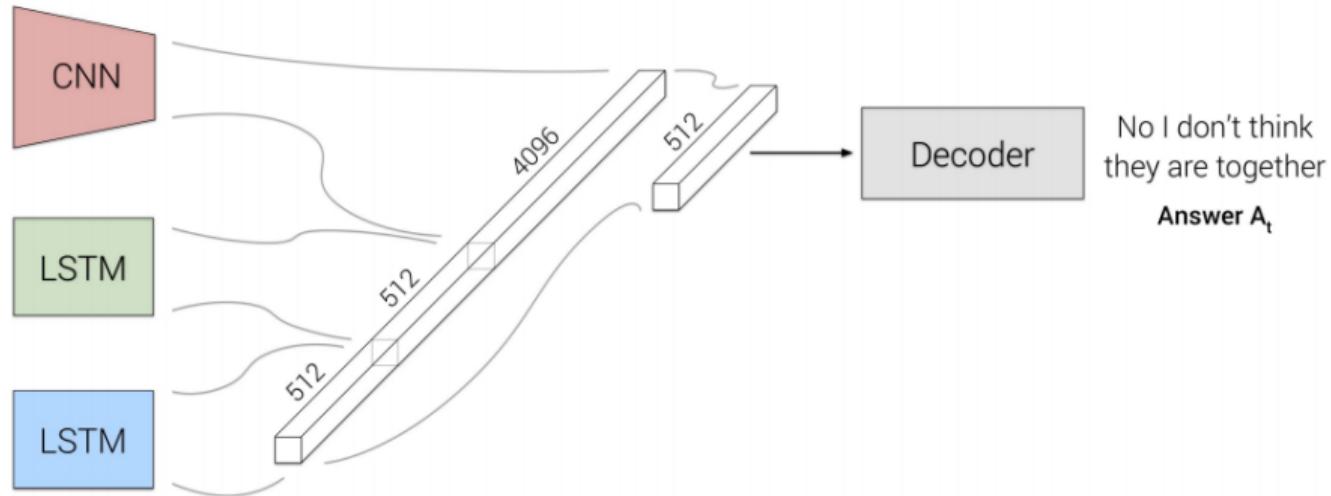
Image I

Do you think the woman is with him?

Question Q_t

The man is riding his bicycle on the sidewalk. Is the man wearing a helmet? No he does not have a helmet on. ... Are there any people nearby? Yes there's a woman walking behind him.

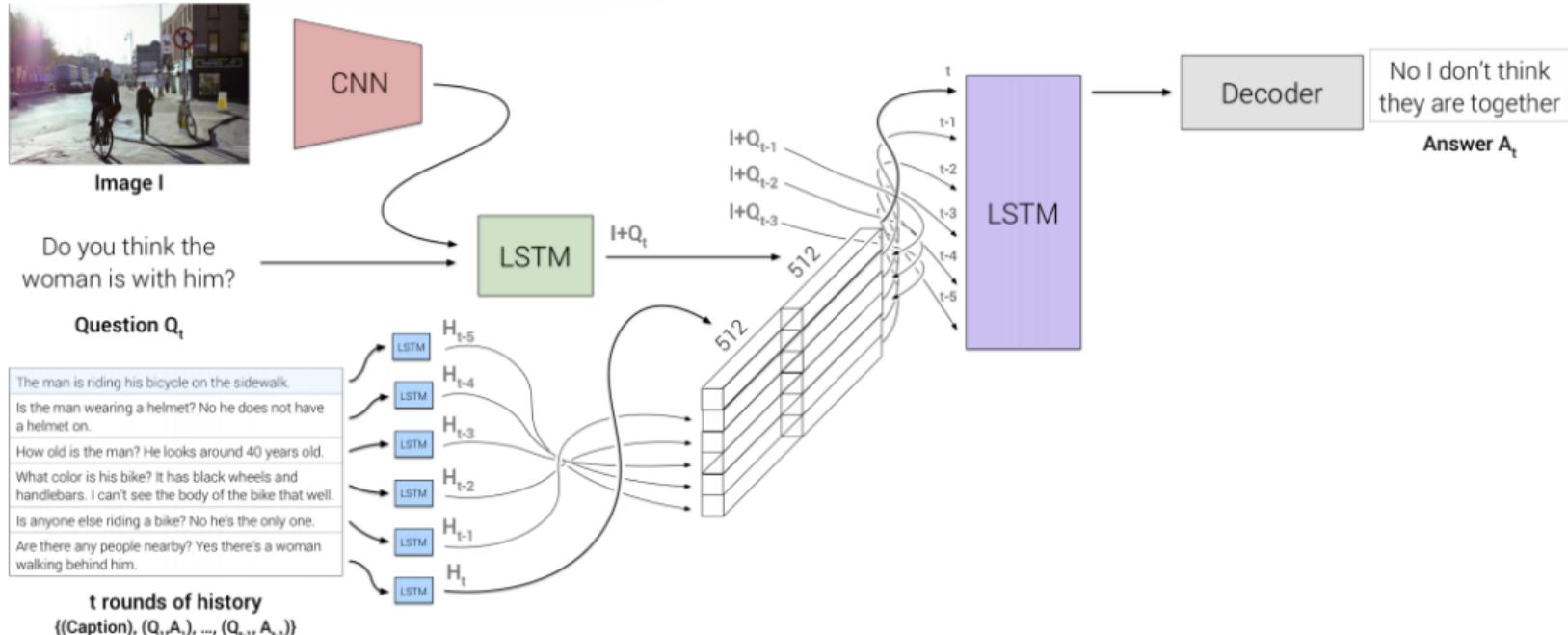
t rounds of history
(concatenated)



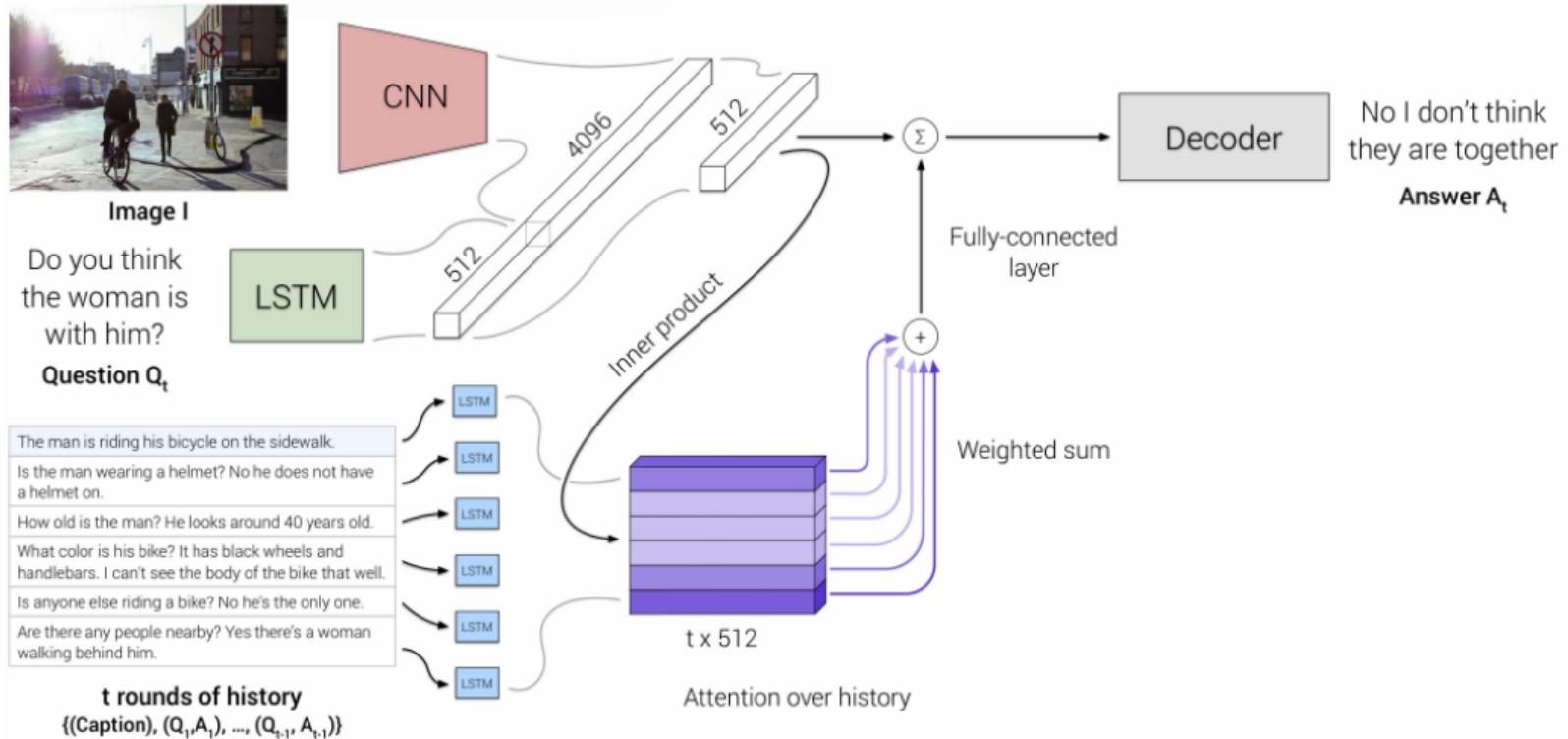
No I don't think
they are together

Answer A_t

Visual Dialog: Hierarchical Recurrent Encoder



Visual Dialog: Memory Network Encoder



Visual Dialog: Decoders

- Generative (LSTM) Decoder

Visual Dialog: Decoders

- Generative (LSTM) Decoder
 - During training, maximize log-likelihood of ground truth answer sequence

Visual Dialog: Decoders

- Generative (LSTM) Decoder
 - During training, maximize log-likelihood of ground truth answer sequence
 - During evaluation, use model's log likelihood scores and rank candidate answers

Visual Dialog: Decoders

- Generative (LSTM) Decoder
 - During training, maximize log-likelihood of ground truth answer sequence
 - During evaluation, use model's log likelihood scores and rank candidate answers
- Discriminative (softmax) Decoder

Visual Dialog: Decoders

- Generative (LSTM) Decoder
 - During training, maximize log-likelihood of ground truth answer sequence
 - During evaluation, use model's log likelihood scores and rank candidate answers
- Discriminative (softmax) Decoder
 - Computes dot product similarity between input encoding and LSTM encoding of each answer option

Visual Dialog: Decoders

- Generative (LSTM) Decoder
 - During training, maximize log-likelihood of ground truth answer sequence
 - During evaluation, use model's log likelihood scores and rank candidate answers
- Discriminative (softmax) Decoder
 - Computes dot product similarity between input encoding and LSTM encoding of each answer option
 - Dot products are fed into a softmax to compute posterior probability over the options

Visual Dialog: Decoders

- Generative (LSTM) Decoder
 - During training, maximize log-likelihood of ground truth answer sequence
 - During evaluation, use model's log likelihood scores and rank candidate answers
- Discriminative (softmax) Decoder
 - Computes dot product similarity between input encoding and LSTM encoding of each answer option
 - Dot products are fed into a softmax to compute posterior probability over the options
 - During training, maximize log-likelihood of correct option

Visual Dialog: Decoders

- Generative (LSTM) Decoder
 - During training, maximize log-likelihood of ground truth answer sequence
 - During evaluation, use model's log likelihood scores and rank candidate answers
- Discriminative (softmax) Decoder
 - Computes dot product similarity between input encoding and LSTM encoding of each answer option
 - Dot products are fed into a softmax to compute posterior probability over the options
 - During training, maximize log-likelihood of correct option
 - During evaluation, options are simply ranked based on their posterior probabilities

Visual Dialog: Results



Caption: The skiers stood on top of the mountain.

Person A (1): how many skiers are there

Person B (1): hundreds

Person A (2): are they getting ready to go downhill

Person B (2): i think so my view is at end of line

Person A (3): is it snowing

Person B (3): no, there is lot of snow though

Person A (4): can you see anybody going downhill

Person B (4): no my view shows people going up small hill on skis i can't see what's going on from there

Person A (5): do you see lift

Person B (5): no

Person A (6): can you tell if they are male or female

Person B (6): skiers closest to me are male

Person A (7): are there any children

Person B (7): i don't see any but there could be it's huge crowd

Person A (8): does anybody have hat on

Person B (8): they all have winter hat of some sort on

Person A (9): is sun shining

Person B (9): yes, all blue sky

Person A (10): do you see any clouds

Person B (10): no clouds

Visual Dialog: Results



Caption: A dog with goggles is in a motorcycle side car.

Person A (1): can you tell what kind of dog this is

Person B (1): he looks like beautiful pit bull mix

Person A (2): can you tell if motorcycle is moving or still

Person B (2): it's parked

Person A (3): is dog's tongue lolling out

Person B (3): not really

Person A (4): is picture in color

Person B (4): yes it is

Person A (5): what color is dog

Person B (5): light tan with white patch that runs up to bottom of his chin and he has white paws on 2 front feet

Person A (6): can you see motorcycle

Person B (6): from side, yes

Person A (7): what color is motorcycle

Person B (7): black with white or silver accents, sun is glaring so it's hard to tell

Person A (8): is there anybody sitting on motorcycle

Person B (8): no

Person A (9): is there anybody in picture

Person B (9): in cars on street behind motorcycle

Person A (10): does dog look like he's having fun

Person B (10): yes

Homework

Questions

- Are there better methods to evaluate Visual Dialog systems? How?

Readings (Optional)

VQA

- Deep Compositional Question Answering with Neural Module Networks, CVPR 2016
- Learning to Reason: End-to-End Module Networks for Visual Question Answering, ICCV 2017
- MUTAN: Multimodal Tucker Fusion for Visual Question Answering, ICCV 2017
- Learning Conditioned Graph Structures for Interpretable Visual Question Answering, NeurIPS 2018
- Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, CVPR 2018
- Towards VQA Models That Can Read, CVPR 2019
- Deep Modular Co-Attention Networks for Visual Question Answering, CVPR 2019

Homework

Readings (Optional)

- Visual Dialog
 - [Evaluating Visual Conversational Agents via Cooperative Human-AI Games](#), HCOMP 2017
 - [Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning](#), ICCV 2017
 - [Visual Coreference Resolution in Visual Dialog using Neural Module Networks](#), ECCV 2018
 - [Improving Generative Visual Dialog by Answering Diverse Questions](#), EMNLP 2019
 - [Large-scale Pretraining for Visual Dialog: A Simple State-of-the-Art Baseline](#), ECCV 2020
 - [Efficient Attention Mechanism for Visual Dialog that can Handle All the Interactions between Multiple Inputs](#), ECCV 2020
 - [History for Visual Dialog: Do we really need it?](#), ACL 2020
 - [Iterative Context-Aware Graph Inference for Visual Dialog](#), CVPR 2020

References I

- [1] Mateusz Malinowski and M. Fritz. "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input". In: *NIPS*. 2014.
- [2] Aishwarya Agrawal et al. "VQA: Visual Question Answering". In: *International Journal of Computer Vision* 123 (2015), pp. 4–31.
- [3] Mengye Ren, Ryan Kiros, and R. Zemel. "Exploring Models and Data for Image Question Answering". In: *NIPS*. 2015.
- [4] A. Fukui et al. "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding". In: *ArXiv* abs/1606.01847 (2016).
- [5] C. Liu et al. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation". In: *ArXiv* abs/1603.08023 (2016).
- [6] Jiasen Lu et al. "Hierarchical Question-Image Co-Attention for Visual Question Answering". In: *ArXiv* abs/1606.00061 (2016).
- [7] Zichao Yang et al. "Stacked Attention Networks for Image Question Answering". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 21–29.

References II

- [8] Yuke Zhu et al. "Visual7W: Grounded Question Answering in Images". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4995–5004.
- [9] J. Johnson et al. "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 1988–1997.
- [10] Abhishek Das et al. "Visual Dialog". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2019), pp. 1242–1256.
- [11] Parikh, Devi, CS 8803 - Vision and Language (Fall 2017). URL:
http://www.prism.gatech.edu/~arjun9/CS8803_CVL_Fall17/ (visited on 09/27/2020).