Visual Question and Answering
Lecture 8

Slides from Devi Parikh, Dhruv Bhatra, Ethan Perez, Jacob Andreas, Marcus Rorbach & others

Visual Question Answering

What is the mustache made of?

AI System

bananas

Devi Parikh

Virginia Tech

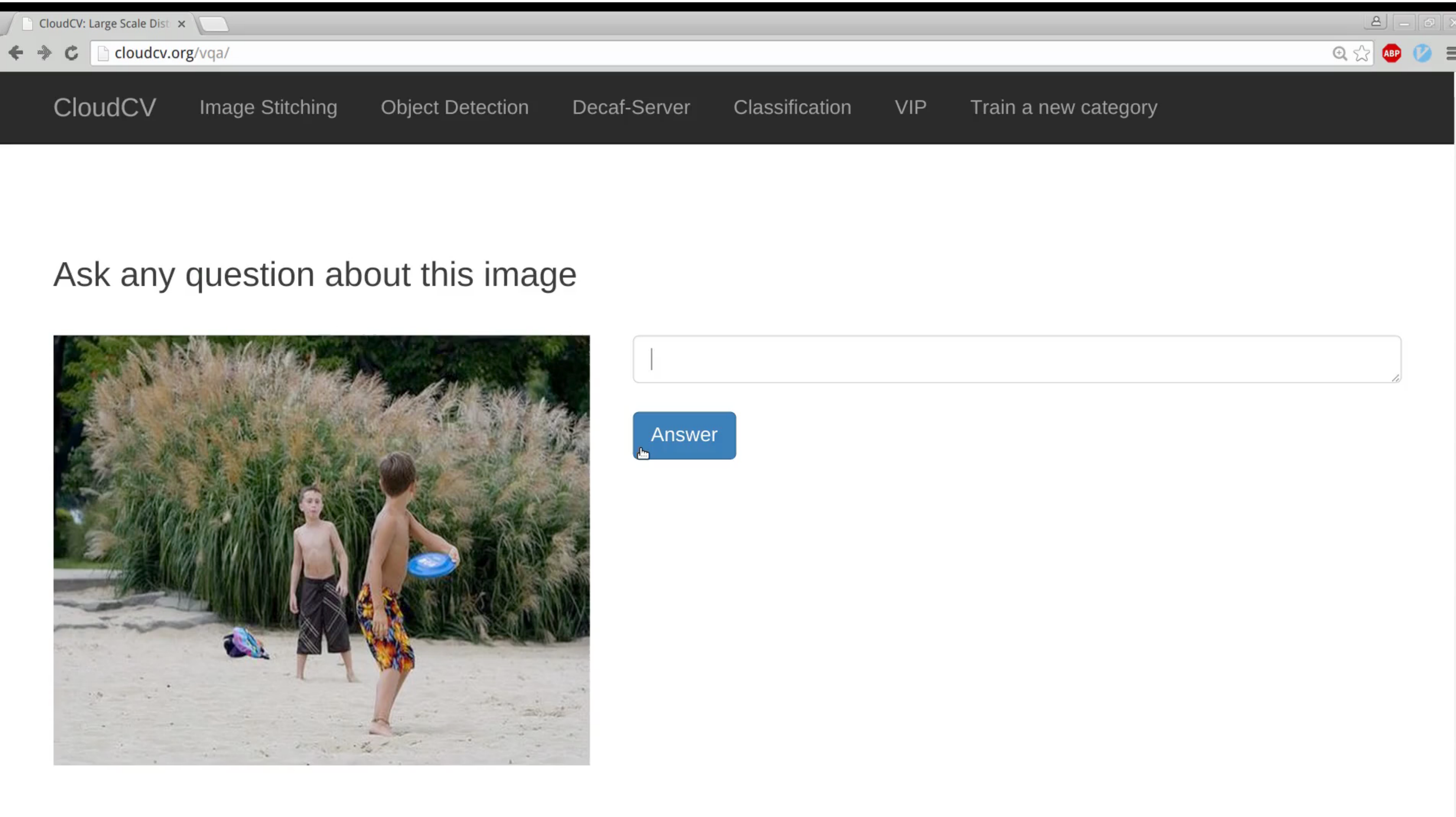
Slide credit: Devi Parikh

Visual Question Answering (VQA)

Task

- Given
  - An image
  - A natural language open-ended question
- Generate
  - A natural language answer

Visual Question Answering (VQA)



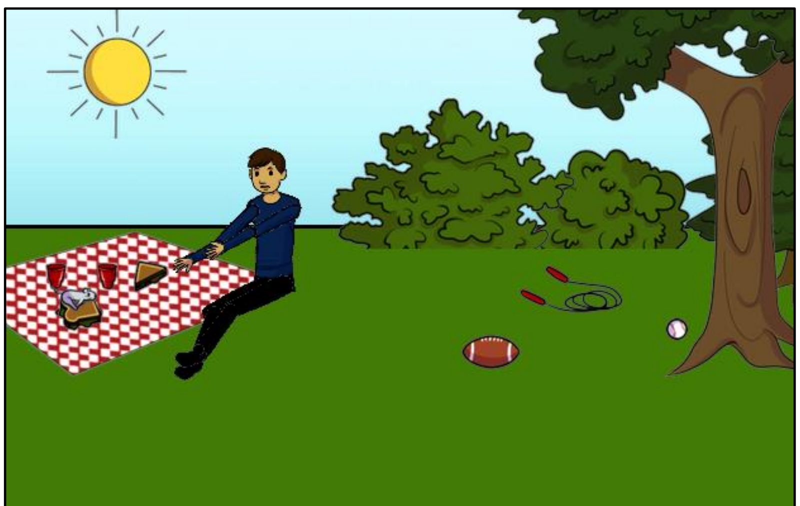
www.visualqa.org

Slide credit: Devi Parikh

Visual Question Answering (VQA)

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?

Slide credit: Devi Parikh

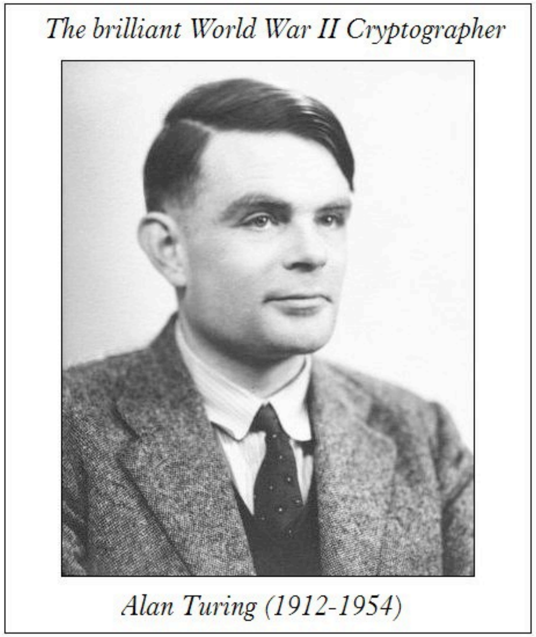
Visual Question Answering (VQA)

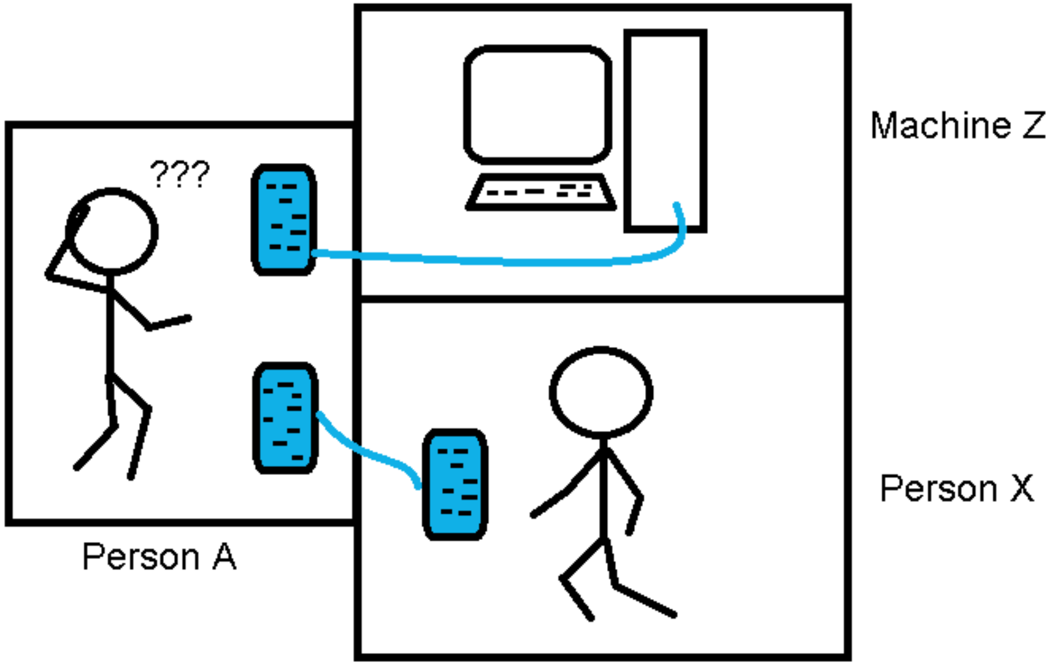
- Details of the image
- Common sense + knowledge base
- Task-driven
- Holy-grail of semantic image understanding

Slide credit: Devi Parikh

Turing Test

"Can machines think"





Q: Please write me a sonnet on the subject of the Forth Bridge.

A: Count me out on this one. I never could write poetry.

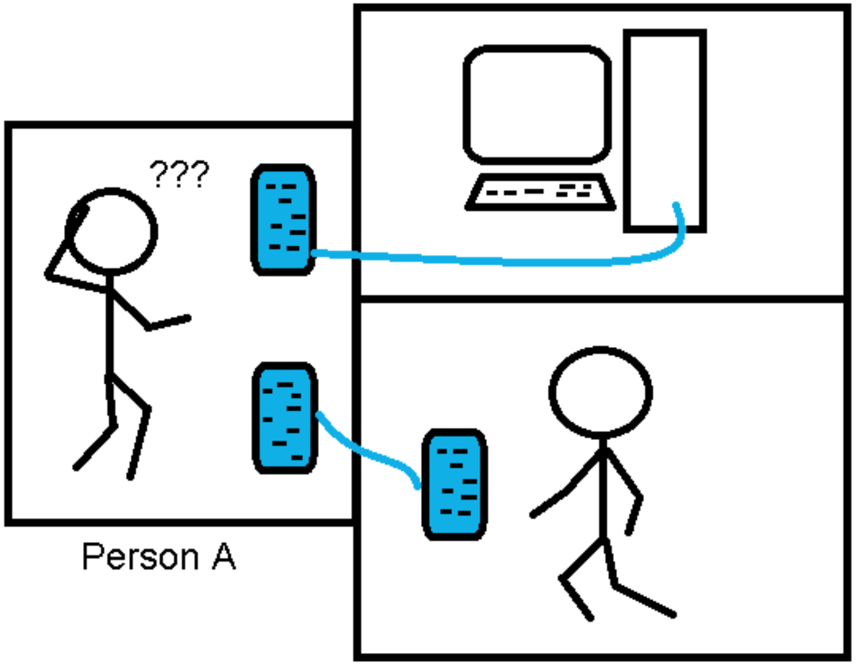
Q: Add 34957 to 70764.

A: (Pause about 30 seconds and then give as answer) 105621.

Visual Turing Test

Q: How many slices of pizza are there?

A: 6



Slide credit: Dhruv Batra

Datasets

Models

Current Status

Ongoing Efforts

Datasets

Models

Current Status

Ongoing Efforts

Slide credit: Devi Parikh

Visual Turing Test [Geman 2014]

- 2591 street city images

Slide credit: Devi Parikh

Vocabulary

- Types of objects
  - People, vehicles, building, windows, doors
- Type-dependent attributes
  - Clothing and activities of people
  - Types and colors of vehicles
- Type-dependent relationships
  - Ordered: person entering a building
  - Unordered: two people walking together

- Questions
  - Existence
  - Uniqueness
  - Attribute
  - Relationship

- Story line
- Query generator
- Human-in-the-loop

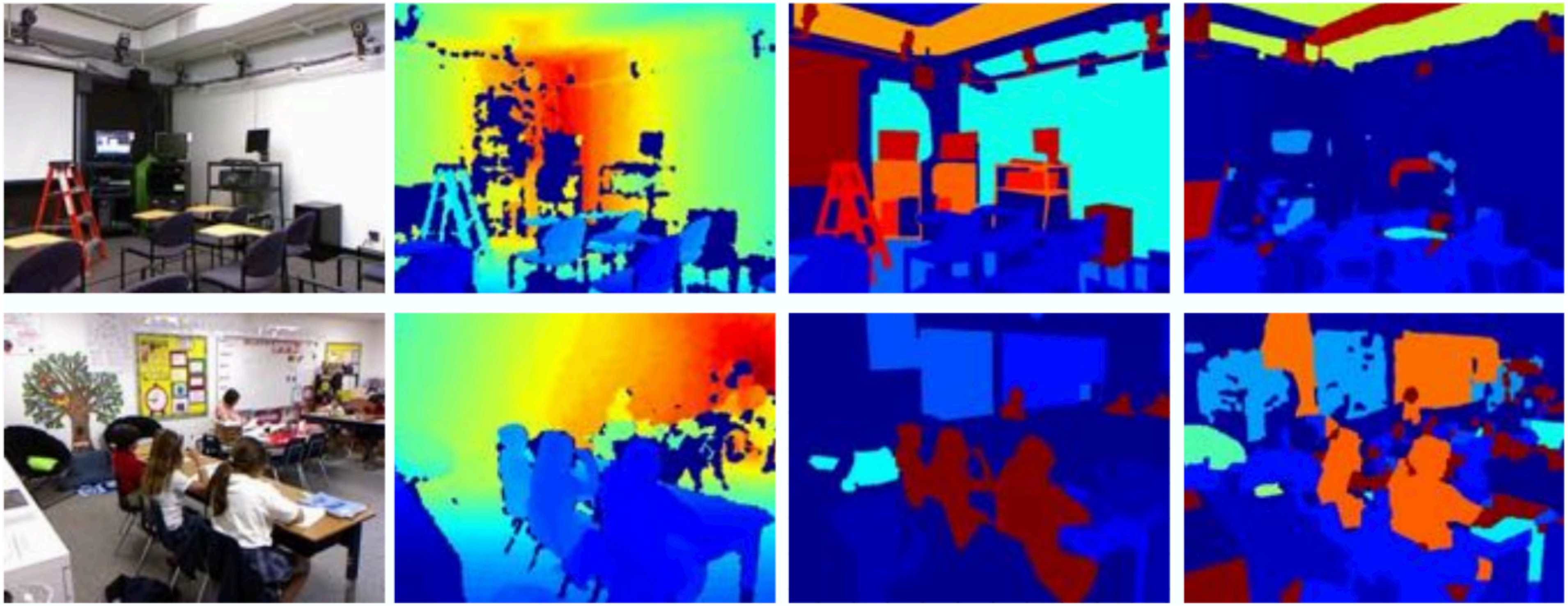
- No NLP required, vision is key

1. Q: Is there a person in the blue region? **A: yes**
2. Q: Is there a unique person in the blue region? **A: yes**
(Label this person 1)
3. Q: Is person 1 carrying something? **A: yes**
4. Q: Is person 1 female? **A: yes**
5. Q: Is person 1 walking on a sidewalk? **A: yes**
6. Q: Is person 1 interacting with any other object? **A: no**
⋮
9. Q: Is there a unique vehicle in the yellow region? **A: yes**
(Label this vehicle 1)
10. Q: Is vehicle 1 light-colored? **A: yes**
11. Q: Is vehicle 1 moving? **A: no**
12. Q: Is vehicle 1 parked and a car? **A: yes**
⋮
14. Q: Does vehicle 1 have exactly one visible tire? **A: no**
15. Q: Is vehicle 1 interacting with any other object? **A: no**
17. Q: Is there a unique person in the red region? **A: no**
18. Q: Is there a unique person that is female in the red region? **A: no**
19. Q: Is there a person that is standing still in the red region? **A: yes**
20. Q: Is there a unique person standing still in the red region? **A: yes**
(Label this person 2)
⋮
23. Q: Is person 2 interacting with any other object? **A: yes**
24. Q: Is person 1 taller than person 2? **A: amb**
25. Q: Is person 1 closer (to the camera) than person 2? **A: no**
26. Q: Is there a person in the red region? **A: yes**
27. Q: Is there a unique person in the red region? **A: yes**
(Label this person 3)
⋮
36. Q: Is there an interaction between person 2 and person 3? **A: yes**
37. Q: Are person 2 and person 3 talking? **A: yes**

Slide credit: Devi Parikh

DAQUAR [Malinowski 2014]

- DAtaset for QUestion Answering on Real-world images (DAQUAR)
- 1449 images from NYU v2



DAQUAR [Malinowski 2014]

- Synthetic QA pairs
  - 140 training
  - 280 test

	Description	Template	Example
Individual	counting	How many {object} are in {image_id}?	How many cabinets are in image1?
	counting and colors	How many {color} {object} are in {image_id}?	How many gray cabinets are in image1?
	room type	Which type of the room is depicted in {image_id}?	Which type of the room is depicted in image1?
	superlatives	What is the largest {object} in {image_id}?	What is the largest object in image1?
set	counting and colors	How many {color} {object}?	How many black bags?
	negations type 1	Which images do not have {object}?	Which images do not have sofa?
	negations type 2	Which images are not {room_type}?	Which images are not bedroom?
	negations type 3	Which images have {object} but do not have a {object}?	Which images have desk but do not have a lamp?

Slide credit: Devi Parikh

DAQUAR [Malinowski 2014]

- Human QA pairs
  - 6794 training
  - 5675 test
- Valid answers
  - Colors, numbers, objects, or sets

Slide credit: Devi Parikh

DAQUAR [Malinowski 2014]

Slide credit: Devi Parikh

DAQUAR [Malinowski 2014]

- Accuracy
- Wu-Palmer similarity (WUPS)
  - WUPS 0.0
  - WUPS 0.9

Slide credit: Devi Parikh

Toronto COCO-QA [Ren 2015]

- COCO dataset
- Caption → QA pair (automatically)
  - 123287 images
  - 78736 train questions
  - 38948 test questions
- 4 types of questions:
  - object, number, color, location
- Answers are all one-word

Slide credit: Devi Parikh

Toronto COCO-QA [Ren 2015]

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

Slide credit: Devi Parikh

Toronto COCO-QA [Ren 2015]

Q. The very old looking what is on display?

A. pot

Q. What swim in the ocean near two large ferries?

A. ducks

Q. What next to the large umbrella attached to a table?

A. trees

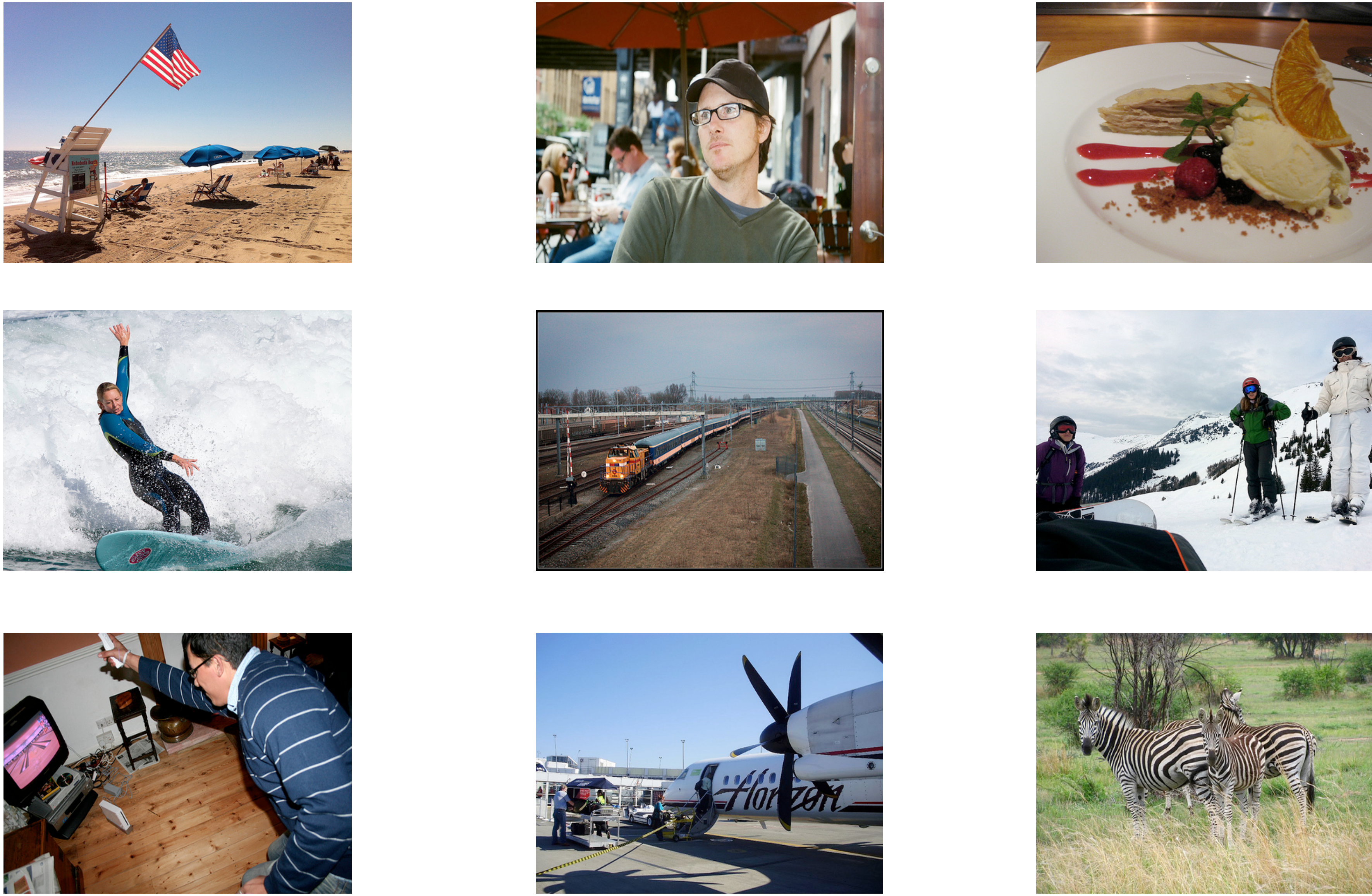
Slide credit: Devi Parikh

Toronto COCO-QA [Ren 2015]

- Accuracy
- Wu-Palmer similarity (WUPS)
  - WUPS 0.0
  - WUPS 0.9

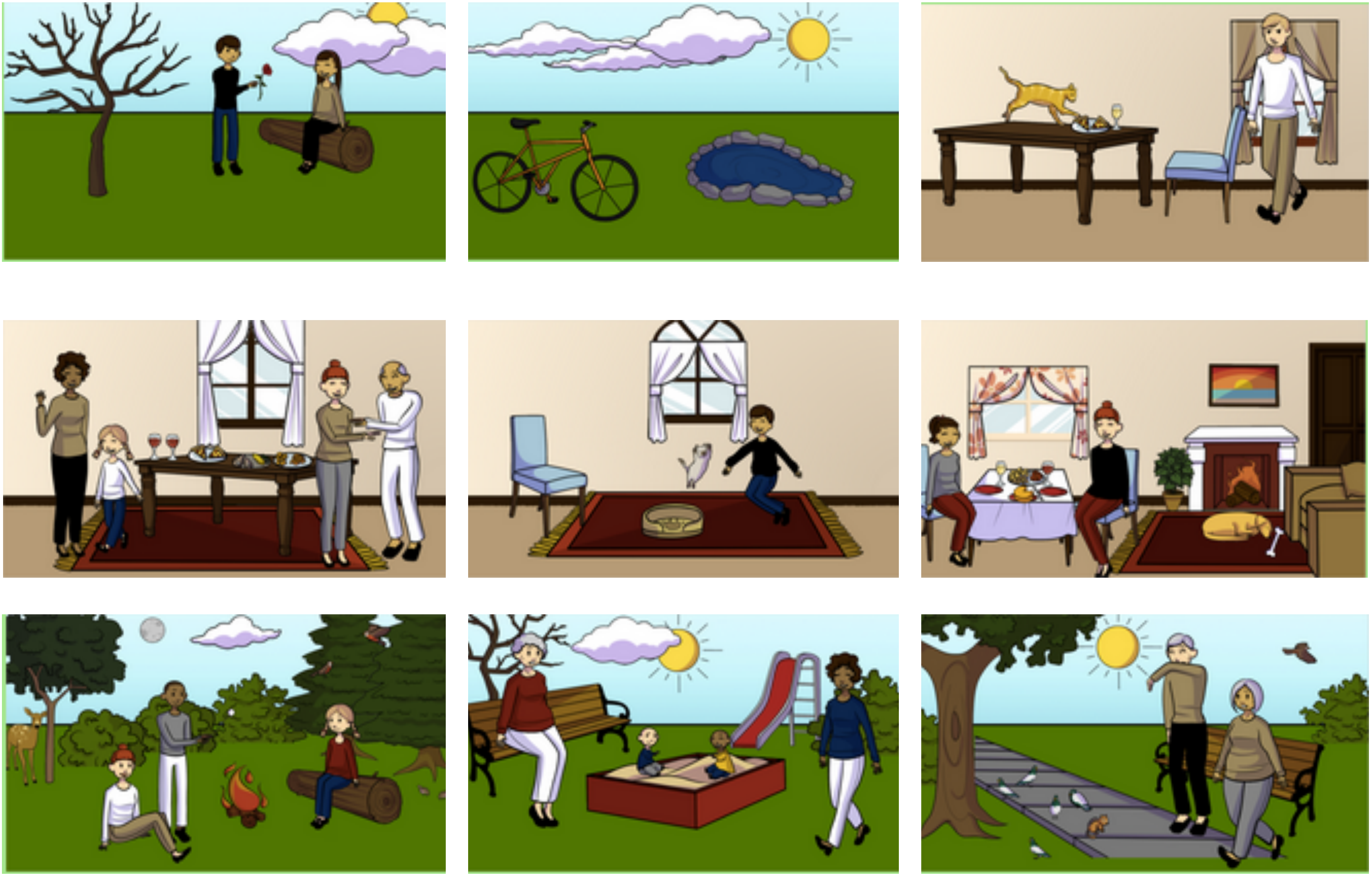
>0.25 million images

Slide credit: Devi Parikh



254,721 images (COCO)

Slide credit: Devi Parikh



50,000 scenes

Slide credit: Devi Parikh

>0.25 million images

>0.76 million questions

~10 million answers

Questions

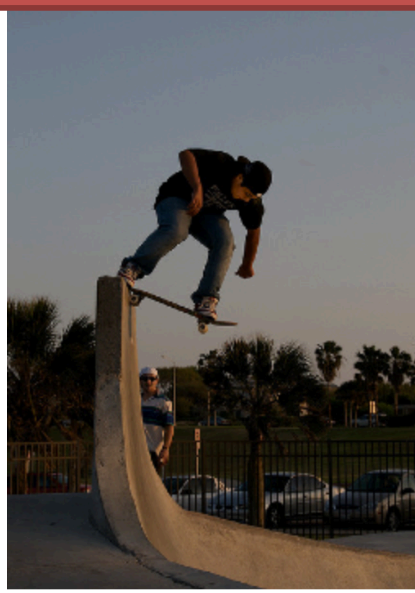
Stump a smart robot! Ask a question about this image that a human can answer, but a smart robot probably can't!

Stump a smart robot!
Ask a question that a human can answer,
but a smart robot probably can't!

We have built … **can recognize the scene (e.g, kitchen, beach)** … **smart robot!**

Ask a question …
IMPORTANT: T … **should not be able to answer the question wi** …

… **ns below:**



- Do not repeat questions. Do not ask the same questions or the same questions with minor variations over and over again across images. Think of a new question each time specific to each image.
- Each question should be a single question. Do not ask questions that have multiple parts or multiple sub-questions in them.
- Do not ask generic questions that can be asked of many other images. Ask questions specific to each image.

Please ask a question about this image that a human can answer *if* looking at the image (and not otherwise), but would stump this smart robot:

Q1: Write your question here to stump this smart robot.

Slide credit: Devi Parikh

>0.25 million images

>0.76 million questions

~10 million answers

>20 person-job-years

Slide credit: Devi Parikh

Taxing the Turkers

- *Beware also the lasting effects of doing too many --for hours after the fact you will not be able to look at any photo without automatically generating a mundane question for it.*
- *If I were in possession of state secrets they could be immediately tortured out of me with the threat of being shown images of: skateboards, trains, Indian food and [long string of expletives] giraffes.*
- *(Please...I will tell you everything...just no more giraffes...)*

Slide credit: Devi Parikh

VQA Visual Question Answering

VirginiaTech Invent the Future®

Microsoft Research

Home People Code Demo Download Evaluation Challenge Browse Visualize Workshop Sponsors Terms

Reset

Top Answers

Slide credit: Devi Parikh

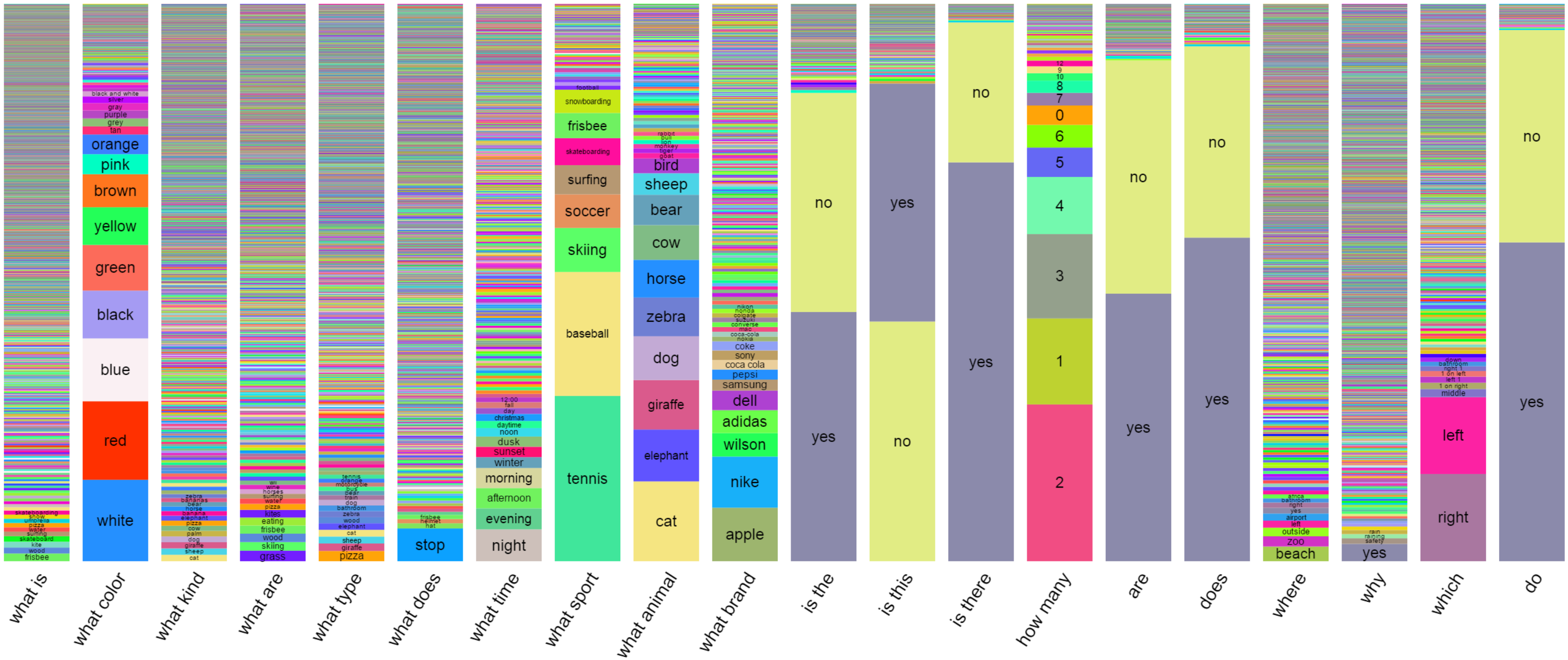
Answers

- 38% of questions are binary yes/no

- 99% questions have answers <= 3 words
  - Evaluation is feasible
  - 23k unique 1 word answers

Slide credit: Devi Parikh

Answers



Slide credit: Devi Parikh

Evaluation Formats

- Open answer
  - Input = image, question
- Multiple choice
  - Input = image, question, 18 answer options
  - Avoids language generation
  - Evaluation (even more) feasible
  - Options = {correct, plausible, popular, random} answers

Slide credit: Devi Parikh

Plausible Answers

Q. What is he playing?

(a) guitar
(b) drums
(c) baseball

Slide credit: Devi Parikh

Accuracy Metric

$$\mathrm{Acc}(\color{red}{ans}) = \min\left\{\frac{\#\text{humans that said } \color{red}{ans}}{3}, 1\right\}$$

1940. COCO_train2014_000000012015

Open-Ended/Multiple-Choice/Ground-Truth

Q: WHAT OBJECT IS THIS

Ground Truth Answers:

(1) television
(2) tv
(3) tv
(4) tv
(5) television
(6) television
(7) television
(8) tv
(9) tv
(10) television

Q: How old is this TV?

Ground Truth Answers:

(1) 20 years
(2) 35
(3) old
(4) more than thirty years old
(5) old
(6) old
(7) 80 s
(8) 30 years
(9) 15 years
(10) very old

Q: Is this TV upside-down?

Ground Truth Answers:

(1) yes
(2) yes
(3) yes
(4) yes
(5) yes
(6) yes
(7) yes
(8) yes
(9) yes
(10) yes

Slide credit: Devi Parikh

Human Accuracy, Inter-Human Agreement

Dataset	Input	All	Yes/No	Number	Other
Real	Question	40.81	67.60	25.77	21.22
	Question + Caption*	57.47	78.97	39.68	44.41
	Question + Image	83.30	95.77	83.39	72.67
Abstract	Question	43.27	66.65	28.52	23.66
	Question + Caption*	54.34	74.70	41.19	40.18
	Question + Image	87.49	95.96	95.04	75.33

Slide credit: Devi Parikh

Human Accuracy, Inter-Human Agreement

Dataset	Input	All	Yes/No	Number	Other
	Question	40.81	67.60	25.77	21.22
Real	Question + Caption*	57.47	78.97	39.68	44.41
	Question + Image	83.30	95.77	83.39	72.67
	Question	43.27	66.65	28.52	23.66
Abstract	Question + Caption*	54.34	74.70	41.19	40.18
	Question + Image	87.49	95.96	95.04	75.33

Slide credit: Devi Parikh

VQA Common Sense

Do These Questions Need Commonsense to Answer?

We will present you with a series of questions about images. For each question, please indicate whether or not you think the question **requires commonsense** in order to answer. A question requires commonsense to answer if answering the question **requires some knowledge beyond what is directly shown in the image.** Some examples are provided below.

Show Examples | Hide Examples

To answer this question, is commonsense required?

- 1. yes
- 2. no

Slide credit: Devi Parikh

VQA Common Sense

- Our best algorithm has* 17% common sense!
- Average common sense required = 31%.

* as estimated by untrained crowd-sourced workers in uncontrolled environment

VQA Age

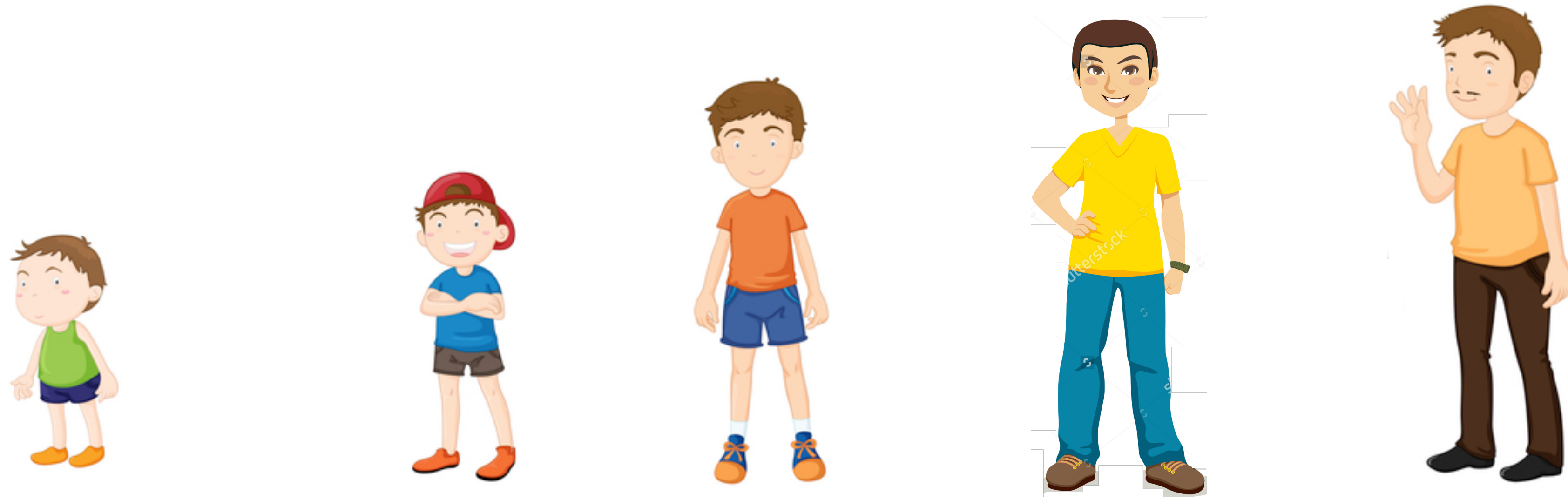
How Old Do You Think a Person Needs to be to Answer These Questions?

We will present you with a series of questions about images. For each question, please select **the youngest age group** that you think a person must be in order to be able to correctly answer the question.

To answer this question, I would expect a person to have to at least be a:

- 1. toddler (3-4)
- 2. younger child (5-8)
- 3. older child (9-12)
- 4. teenager (13-17)
- 5. adult (18+)

Slide credit: Devi Parikh



3-4 (15.3%)	**5-8** (39.7%)	**9-12** (28.4%)	**13-17** (11.2%)	**18+** (5.5%)
Is that a bird in the sky?	How many pizzas are shown?	Where was this picture taken?	Is he likely to get mugged if he walked down a dark alleyway like this?	What type of architecture is this?
What color is the shoe?	What are the sheep eating?	What ceremony does the cake commemorate?	Is this a vegetarian meal?	Is this a Flemish bricklaying pattern?
How many zebras are there?	What color is his hair?	Are these boats too tall to fit under the bridge?	What type of beverage is in the glass?	How many calories are in this pizza?
Is there food on the table?	What sport is being played?	What is the name of the white shape under the batter?	Can you name the performer in the purple costume?	What government document is needed to partake in this activity?
Is this man wearing shoes?	Name one ingredient in the skillet.	Is this at the stadium?	Besides these humans, what other animals eat here?	What is the make and model of this vehicle?

Slide credit: Devi Parikh

Question	Average Age
what brand	12.5
why	11.18
what type	11.04
what kind	10.55
is this	10.13
what does	10.06
what time	9.81
who	9.58
where	9.54
which	9.32
does	9.29
do	9.23
what is	9.11
what are	9.04
are	8.65
is the	8.52
is there	8.24
what sport	8.06
how many	7.67
what animal	6.74
what color	6.6

Slide credit: Devi Parikh

VQA Age

- Our best algorithm =* 4.84 years old!
- Average "age of questions" = 8.98 years.

* age as estimated by untrained crowd-sourced workers in uncontrolled environment

Datasets

Models

Current Status

Ongoing Efforts

Slide credit: Devi Parikh

Challenges in VQA

- Image representation
- Language representation
- Combining the modalities
- Attention
- Question-specific reasoning
- External knowledge

Slide credit: Devi Parikh

Basic Approach

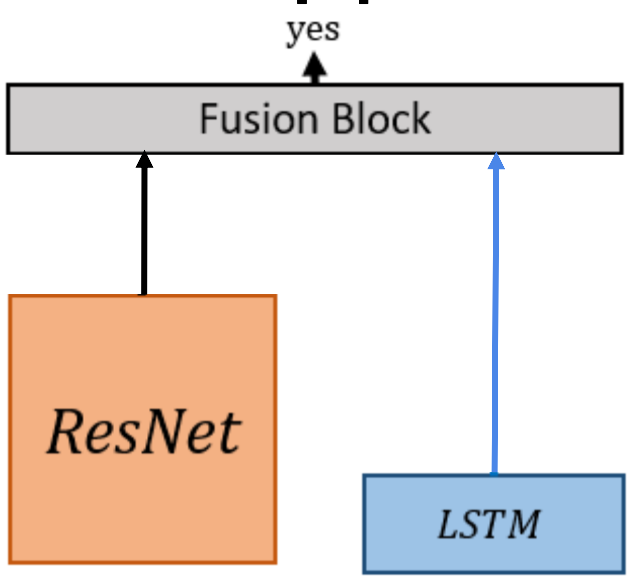


Figure from de Vries et al. "Modulating early visual processing by language." arXiv 2017.

Slide credit: Devi Parikh

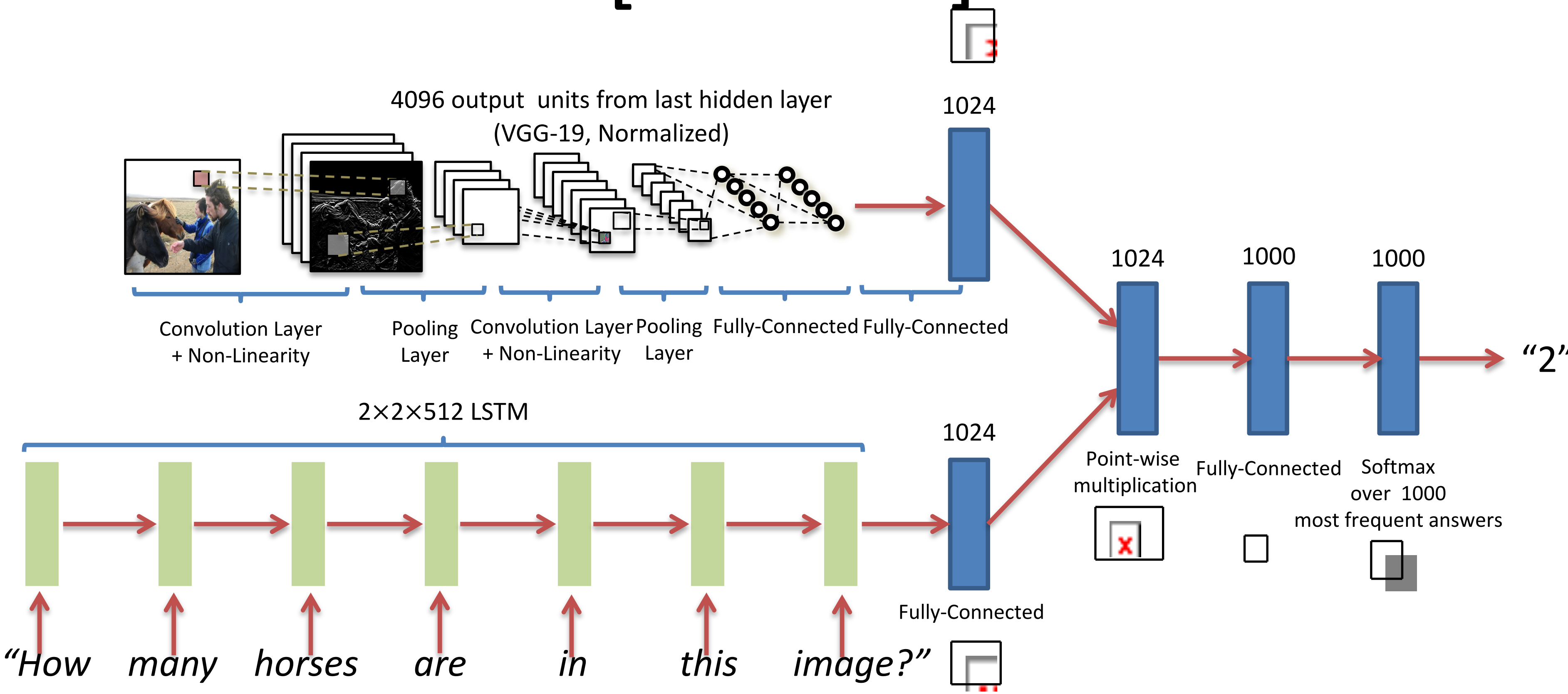
[Lu 2015]

- Input: Image, Question
- Output: Answer

- Image:
  - Convolutional Neural Network (CNN) [Fukushima 1980, LeCun et al. 1989]
- Question:
  - Recurrent Neural Network
  - Specifically, a Long Short-Term Memory (LSTM) [Hochreiter & Schmidhuber, 1997]
- Output: 1 of K most common answers

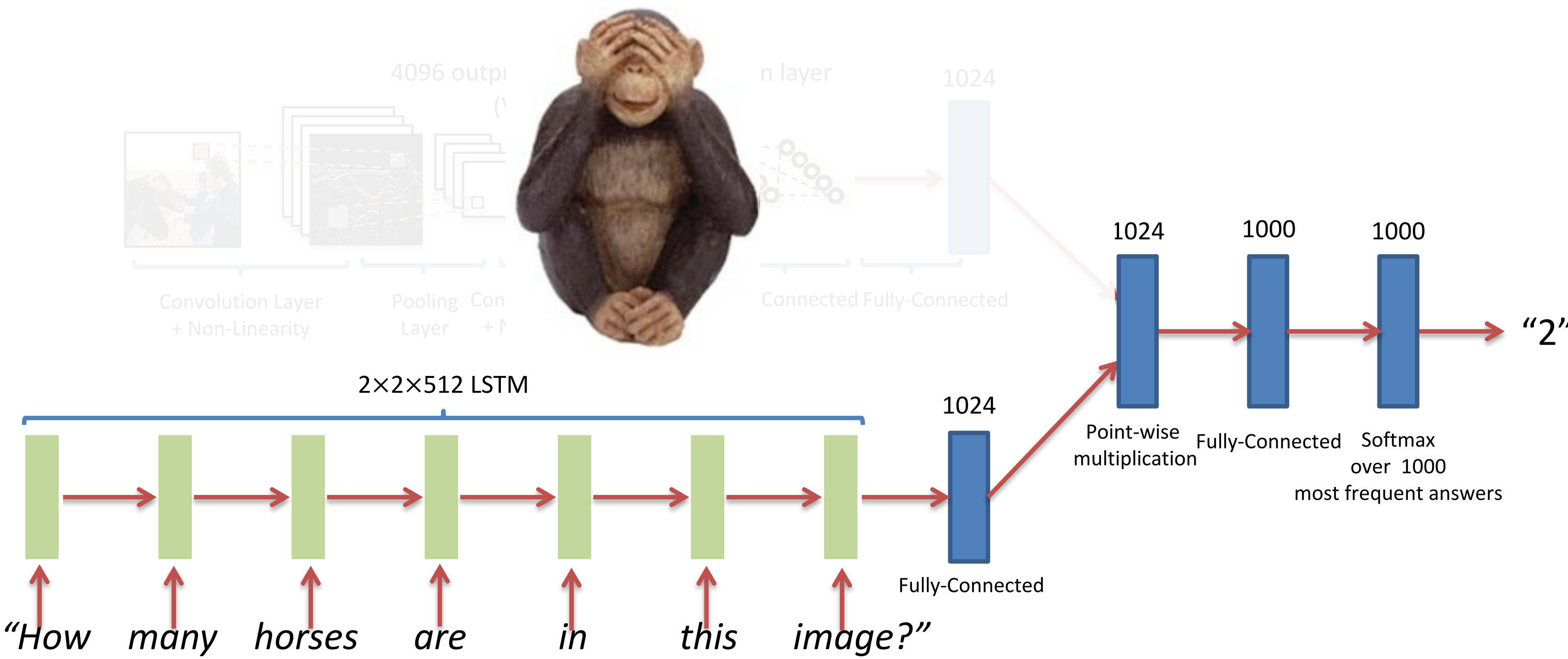
Slide credit: Devi Parikh

[Lu 2015]



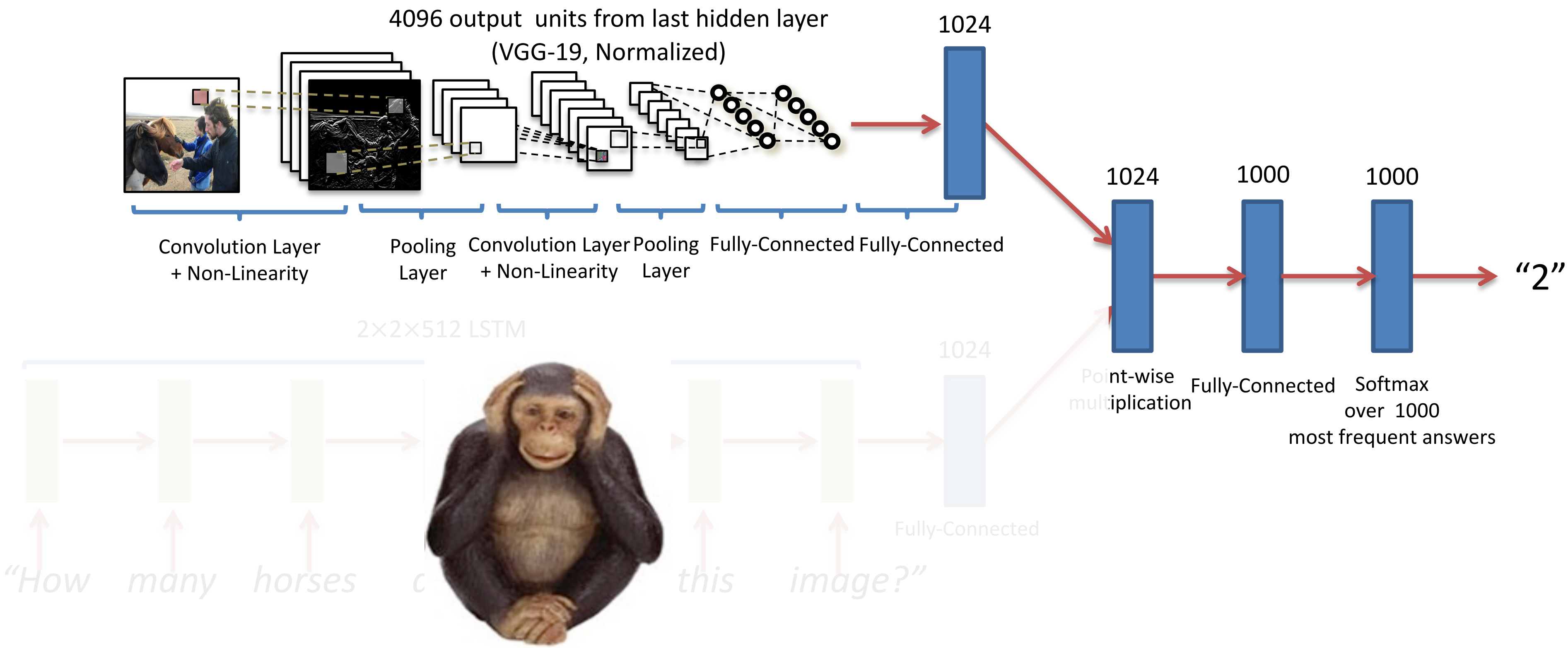
4096 output units from last hidden layer
(VGG-19, Normalized)
1024
1024
1000
1000
Convolution Layer
+ Non-Linearity
Pooling
Layer
Convolution Layer
+ Non-Linearity
Pooling
Layer
Fully-Connected
Fully-Connected
"2"
2×2×512 LSTM
1024
Point-wise
multiplication
Fully-Connected
Softmax
over 1000
most frequent answers
Fully-Connected
"How many horses are in this image?"

Ablation #1: Language-alone



Slide credit: Devi Parikh

Ablation #2: Vision-alone



Slide credit: Devi Parikh

Results

	Open-Answer				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
Question	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
Image	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
Q+I	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTM Q	50.39	78.41	34.68	30.03	55.88	78.45	35.91	41.13
LSTM Q+I	$\boxed{57.75}$	80.5	36.77	43.08	62.7	80.52	38.22	53.01
Caption	26.70	65.50	02.03	03.86	28.29	69.79	02.06	03.82
Q+C	54.70	75.82	40.12	42.56	59.85	75.89	41.16	52.53

“yes”	29.27
k-NN	40.61

Code available!

- Multiple-Choice > Open-Ended
- Question alone does quite well
  - Better than humans
- Image helps

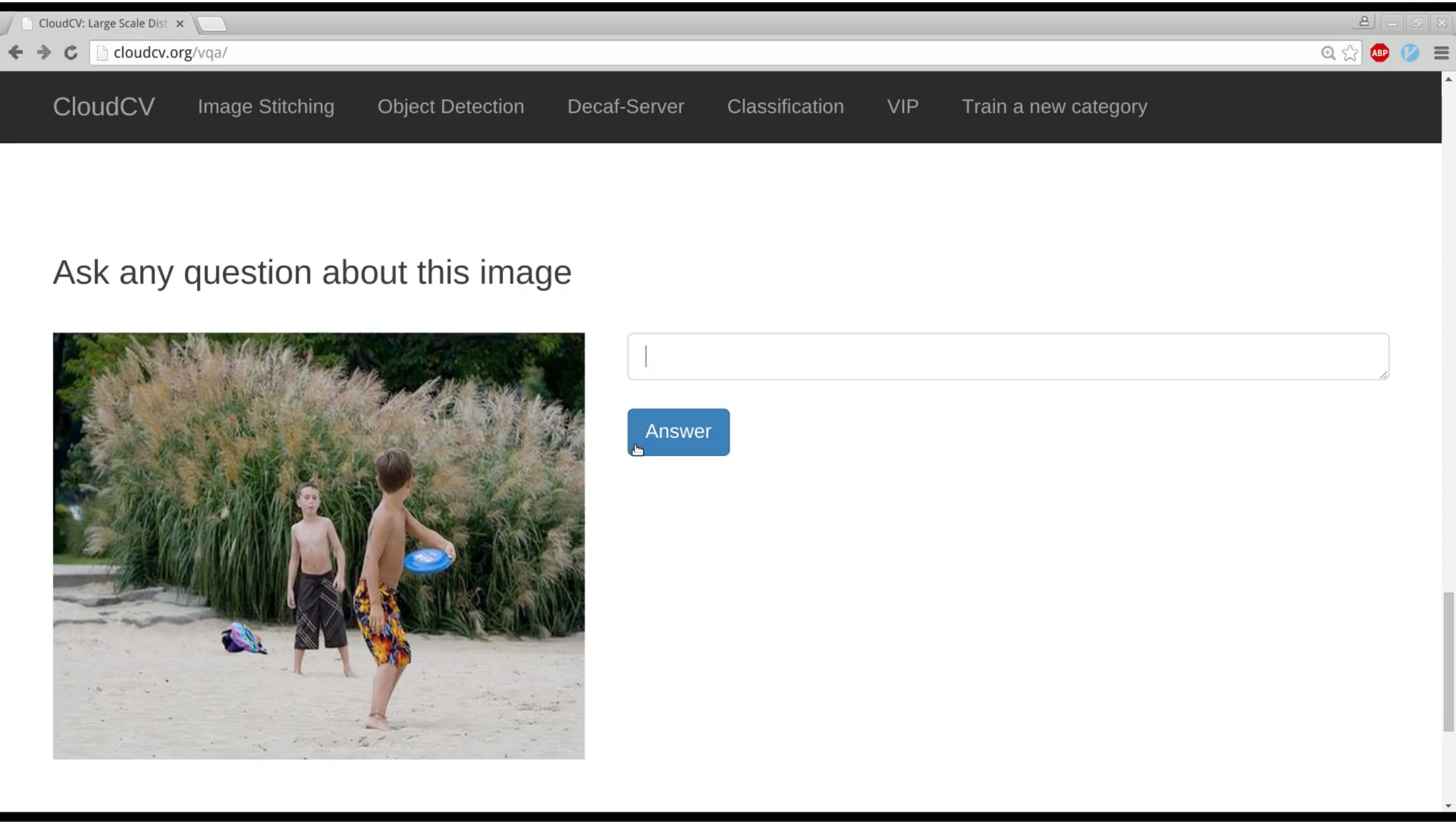
Slide credit: Devi Parikh

Results

	Open-Ended				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
prior ("yes")	29.66	70.81	00.39	01.15	29.66	70.81	00.39	01.15
per Q-type prior	37.54	71.03	35.77	09.38	39.45	71.02	35.86	13.34
nearest neighbor	42.70	71.89	24.36	21.94	48.49	71.94	26.00	33.56
BoW Q	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
I	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
BoW Q + I	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTM Q	48.76	78.20	35.68	26.59	54.75	78.22	36.82	38.78
LSTM Q + I	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
deeper LSTM Q	50.39	78.41	34.68	30.03	55.88	78.45	35.91	41.13
deeper LSTM Q + norm I	**57.75**	**80.50**	**36.77**	**43.08**	**62.70**	**80.52**	**38.22**	**53.01**
Caption	26.70	65.50	02.03	03.86	28.29	69.79	02.06	03.82
BoW Q + C	54.70	75.82	40.12	42.56	59.85	75.89	41.16	52.53

Slide credit: Devi Parikh

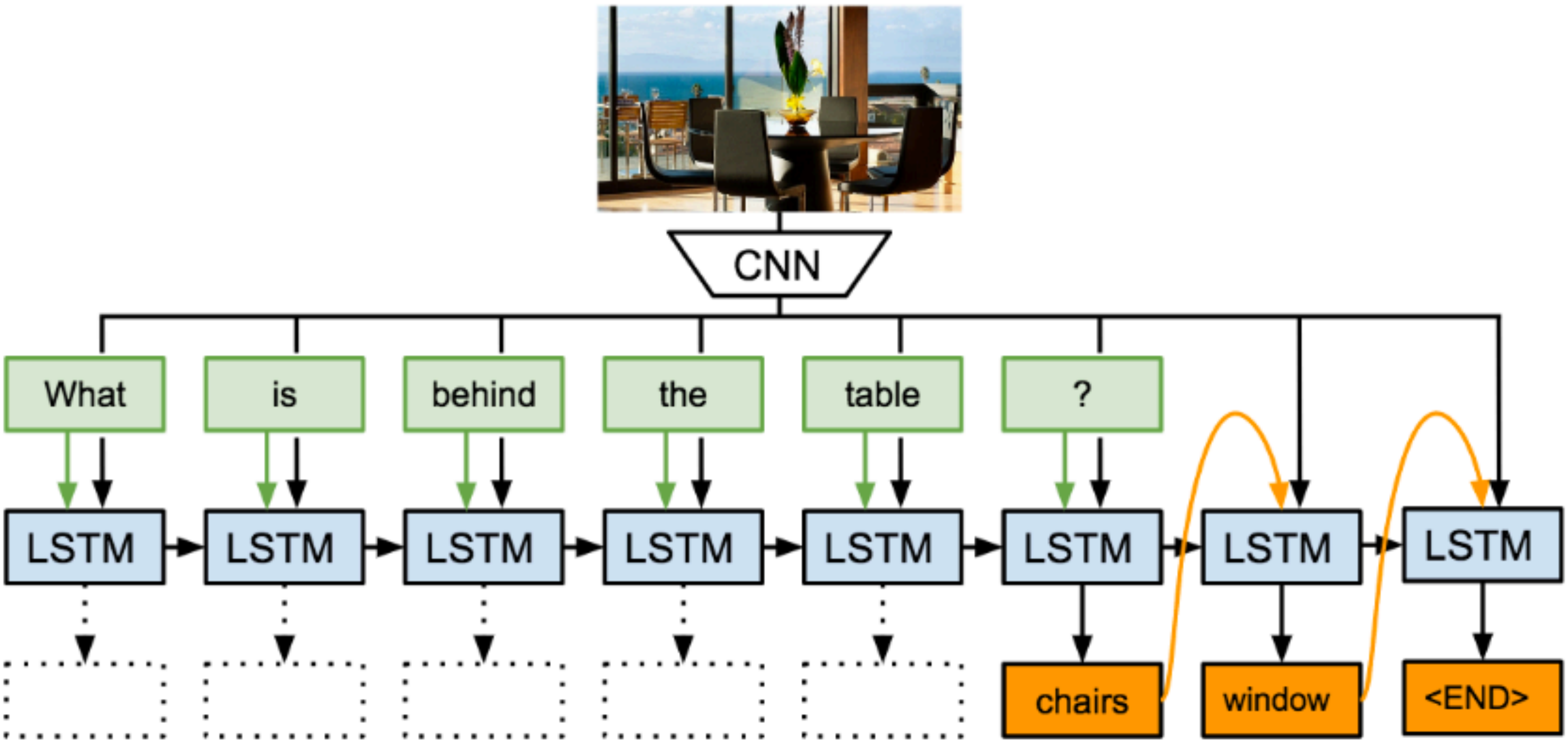
Demo



www.visualqa.org

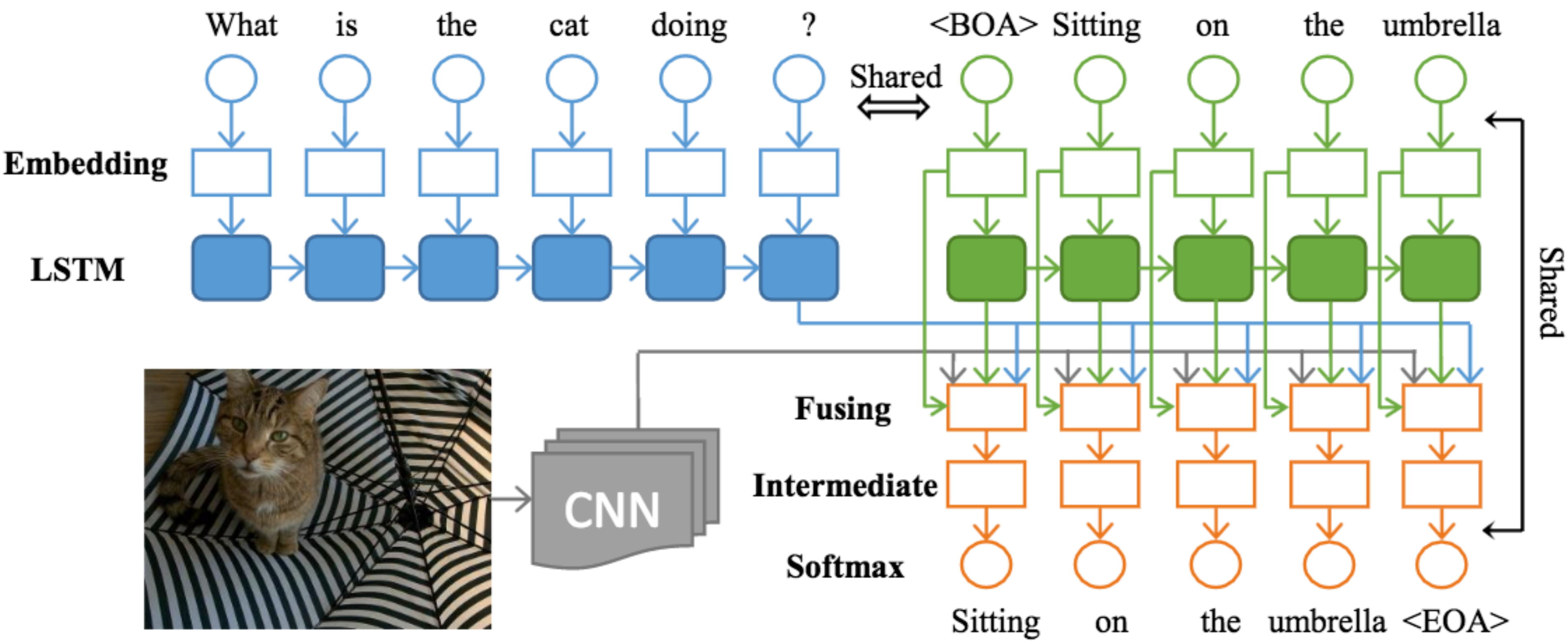
Slide credit: Devi Parikh

[Malinowski 2015]



Slide credit: Devi Parikh

[Gao 2015]



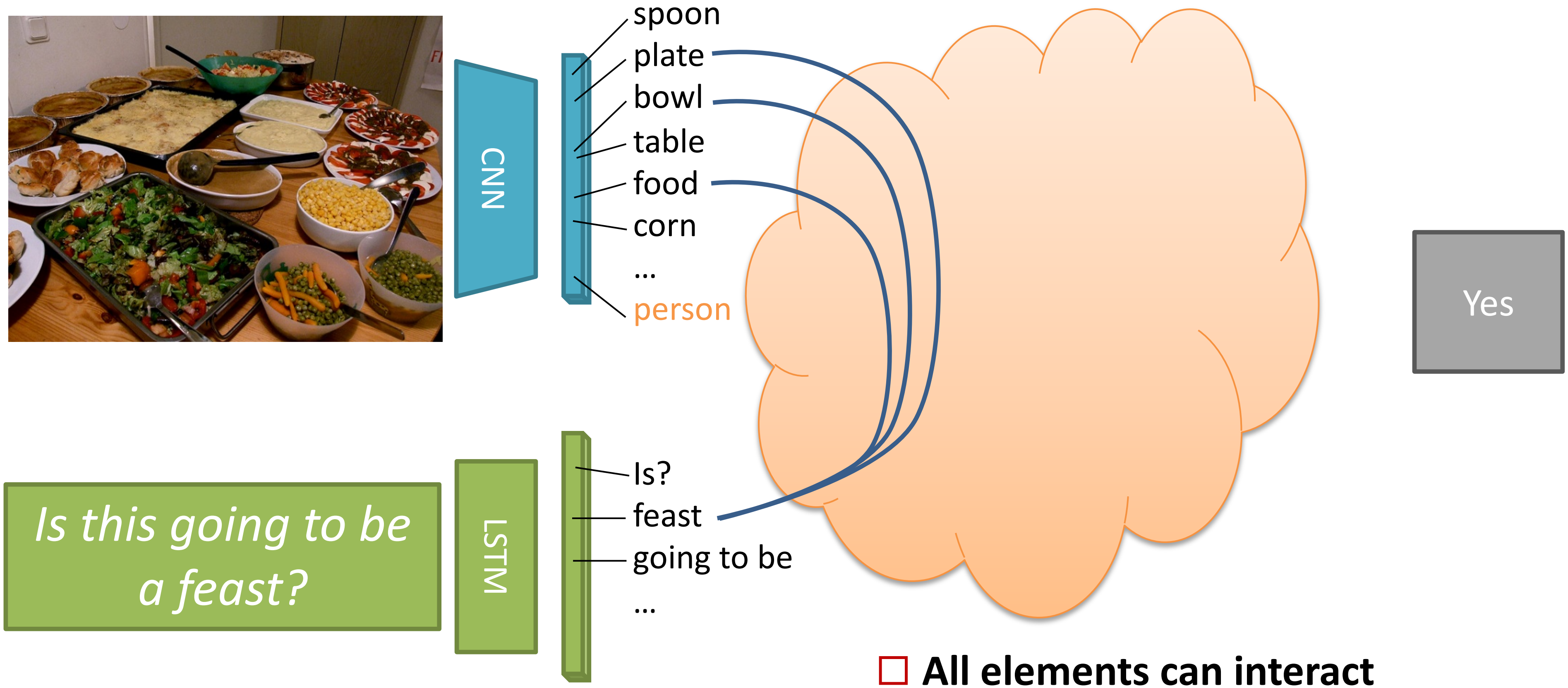
Slide credit: Devi Parikh

Challenges in VQA

- Image representation
- Language representation
- Combining the modalities
- Attention
- Question-specific reasoning

Slide credit: Devi Parikh

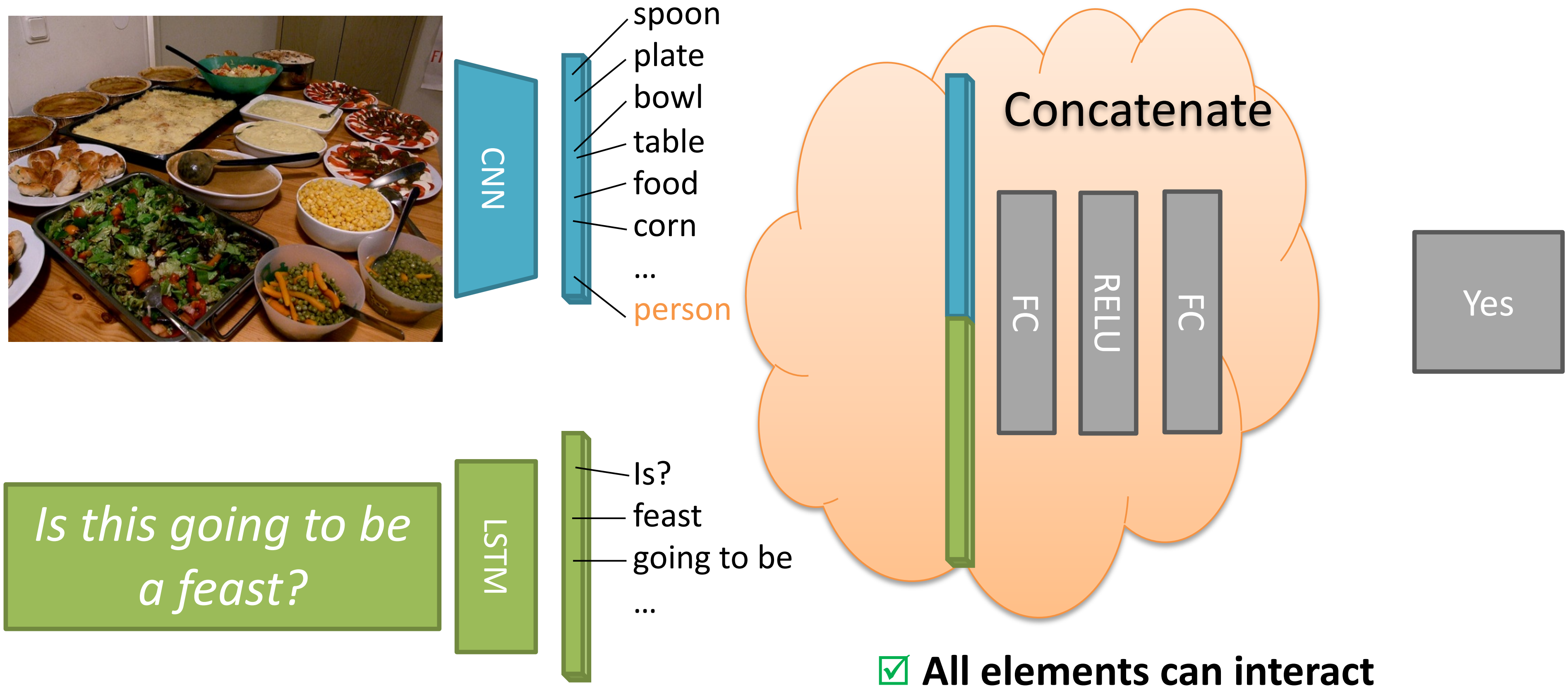
How to Combine
Image Representation and Question Representation?



- ☐ **All elements can interact**
- ☐ **Multiplicative interaction**

Slide credit: Akira Fukui and Marcus Rohrbach

How to Combine
Image Representation and Question Representation?



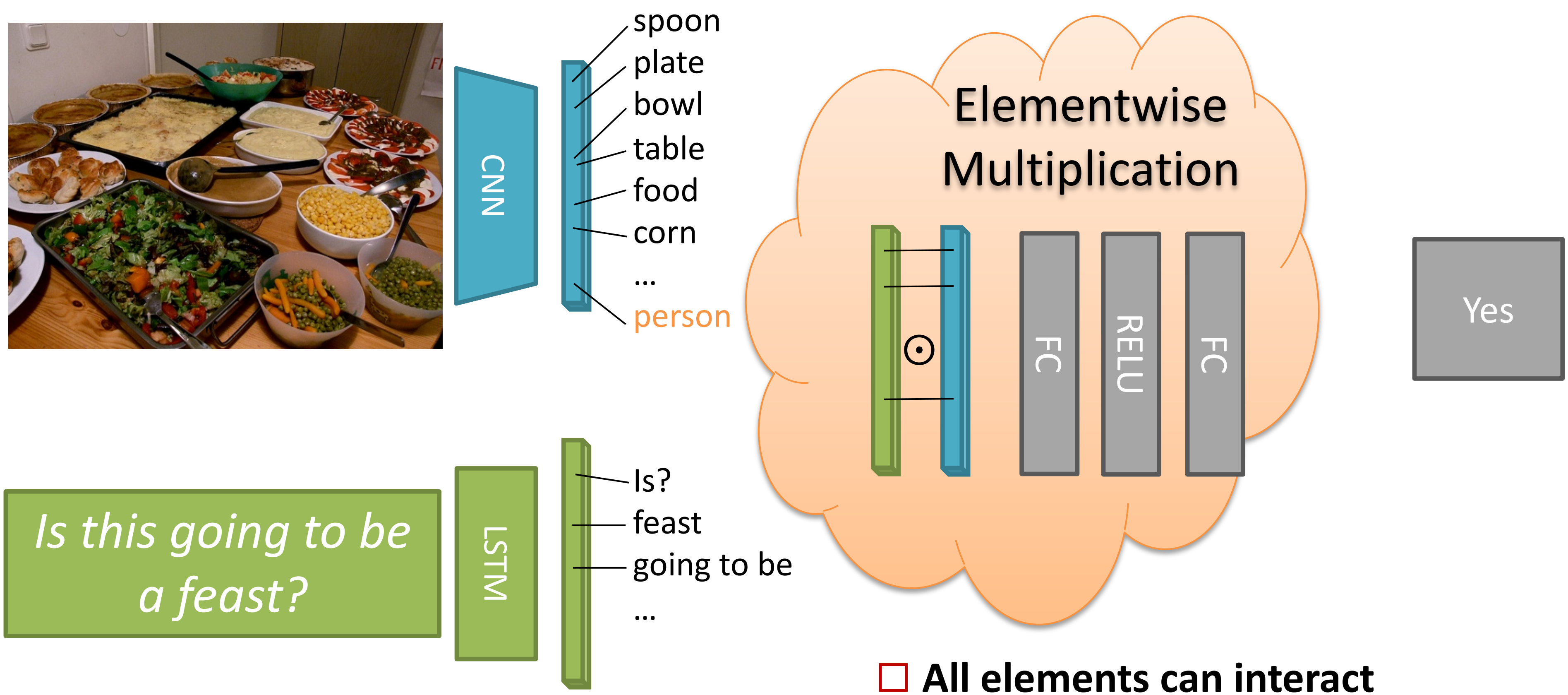
☑ **All elements can interact**

☐ **Multiplicative interaction**

- Difficult to learn output classification

Slide credit: Akira Fukui and Marcus Rohrbach

How to Combine
Image Representation and Question Representation?

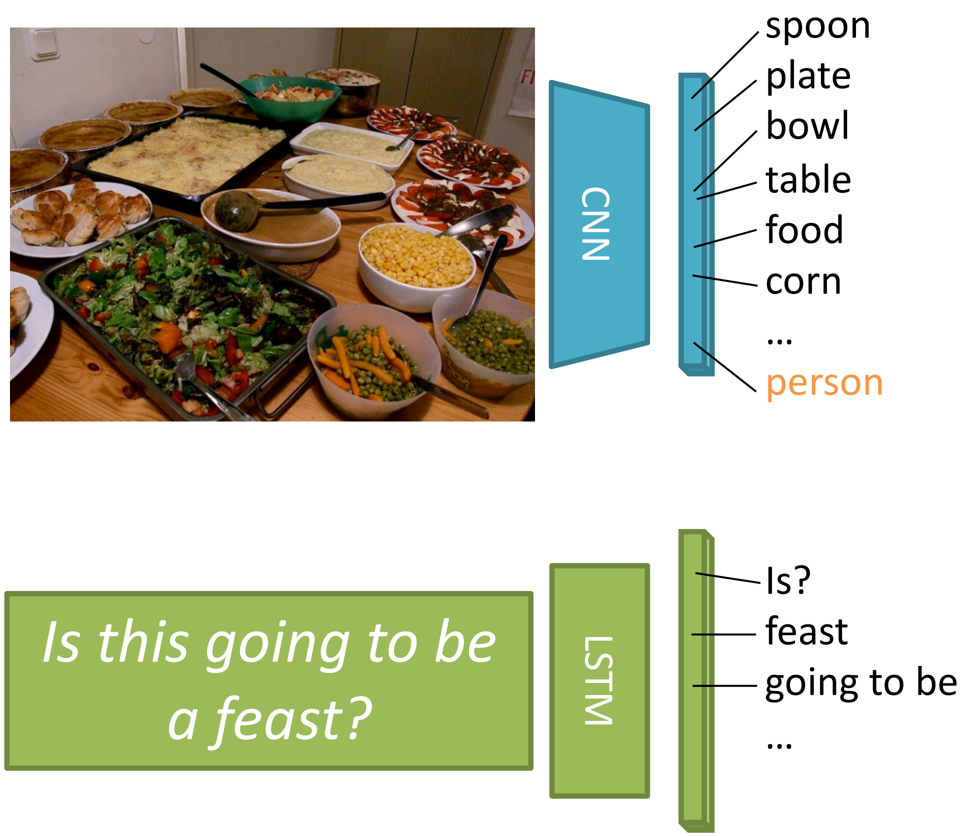


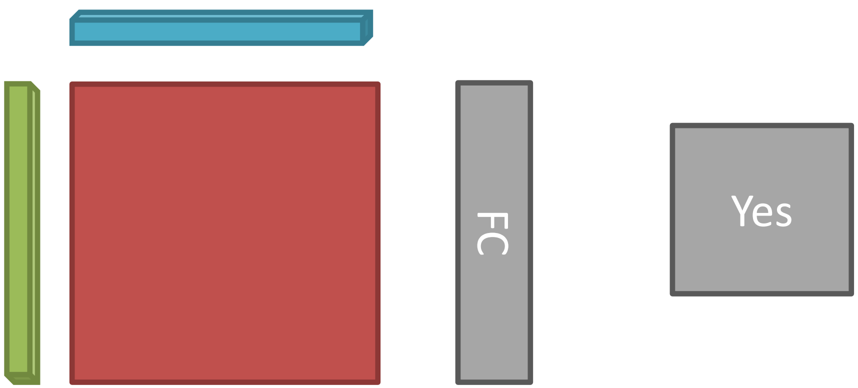
- ☐ **All elements can interact**
- ☑ **Multiplicative interaction**
- - Difficult to learn input embedding

Slide credit: Akira Fukui and Marcus Rohrbach

How to Combine
Image Representation and Question Representation?

Outer Product / Bilinear Pooling [Lin ICCV 2015]





☑ **All elements can interact**

☑ **Multiplicative interaction**

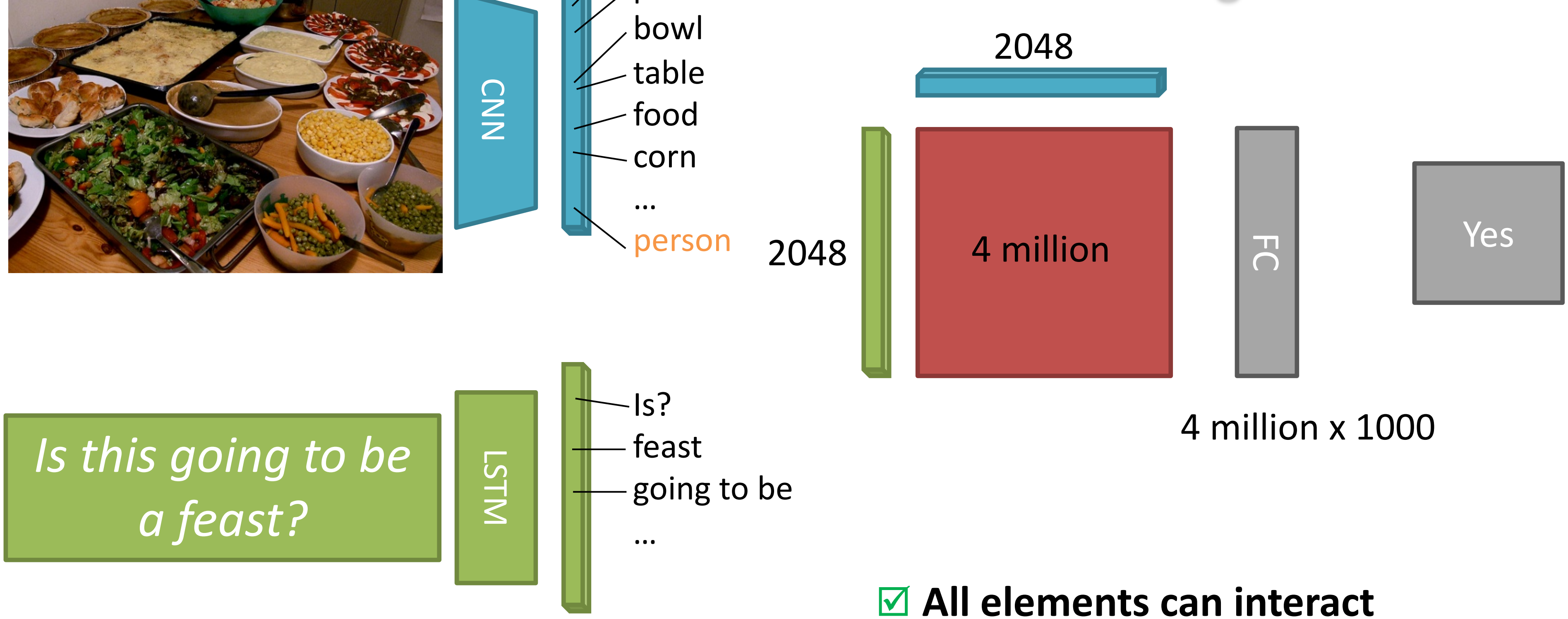
[Lin ICCV 2015] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. ICCV 2015

Slide credit: Akira Fukui and Marcus Rohrbach

How to Combine
Image Representation and Question Representation?

Outer Product / Bilinear Pooling

[Lin ICCV 2015]



- ☑ **All elements can interact**
- ☑ **Multiplicative interaction**
- ☐ **High #activations & computation**
- ☐ **High #parameters**

[Lin ICCV 2015] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. ICCV 2015

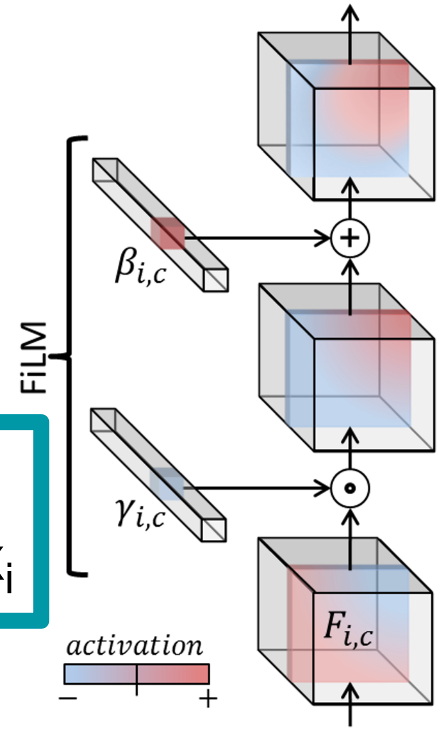
Slide credit: Akira Fukui and Marcus Rohrbach

FiLM: **F**eature-w**i**se **L**inear **M**odulation

$$FiLM(\boldsymbol{F}_{i,c}|\gamma_{i,c},\beta_{i,c}) = \gamma_{i,c}\boldsymbol{F}_{i,c} + \beta_{i,c}$$

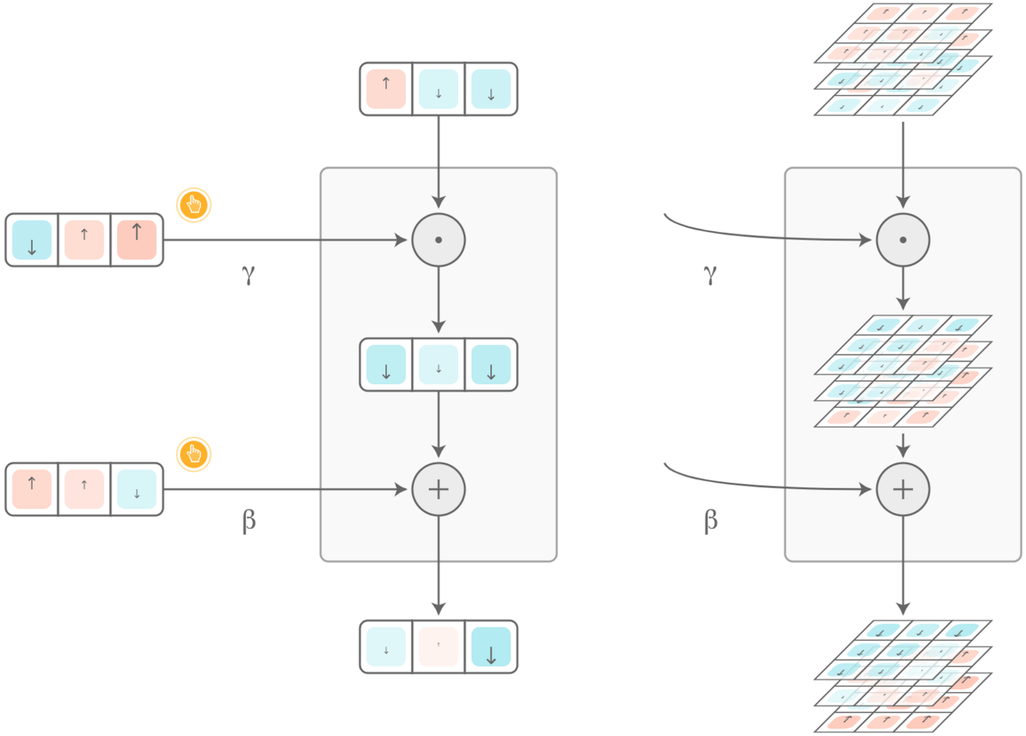
$$\gamma_{i,c} = f_c(\boldsymbol{x}_i) \qquad \beta_{i,c} = h_c(\boldsymbol{x}_i)$$

γ, β change how features are used as learned functions of conditioning input x_i



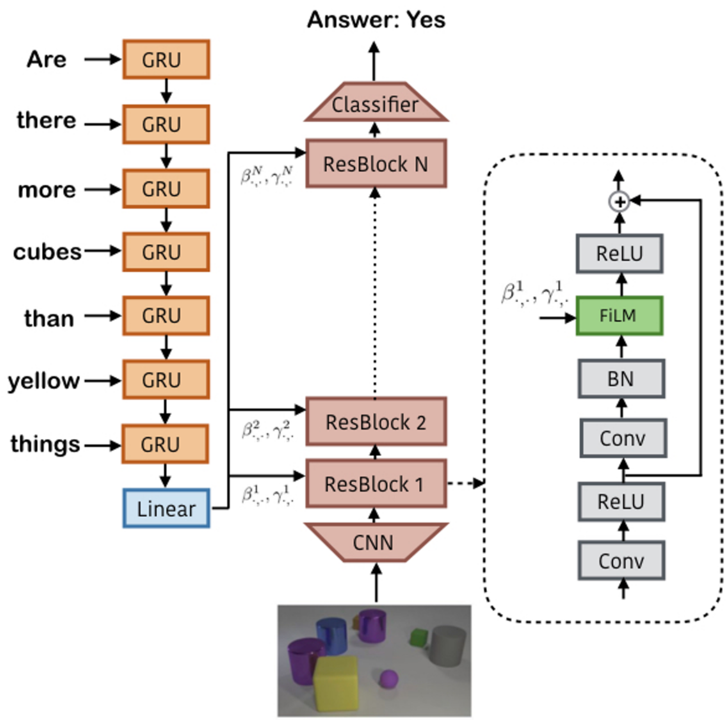
Perez et al. "FiLM: Visual Reasoning with a General Conditioning Layer". AAAI 2018.

FiLM: **F**eature-w**i**se **L**inear **M**odulation



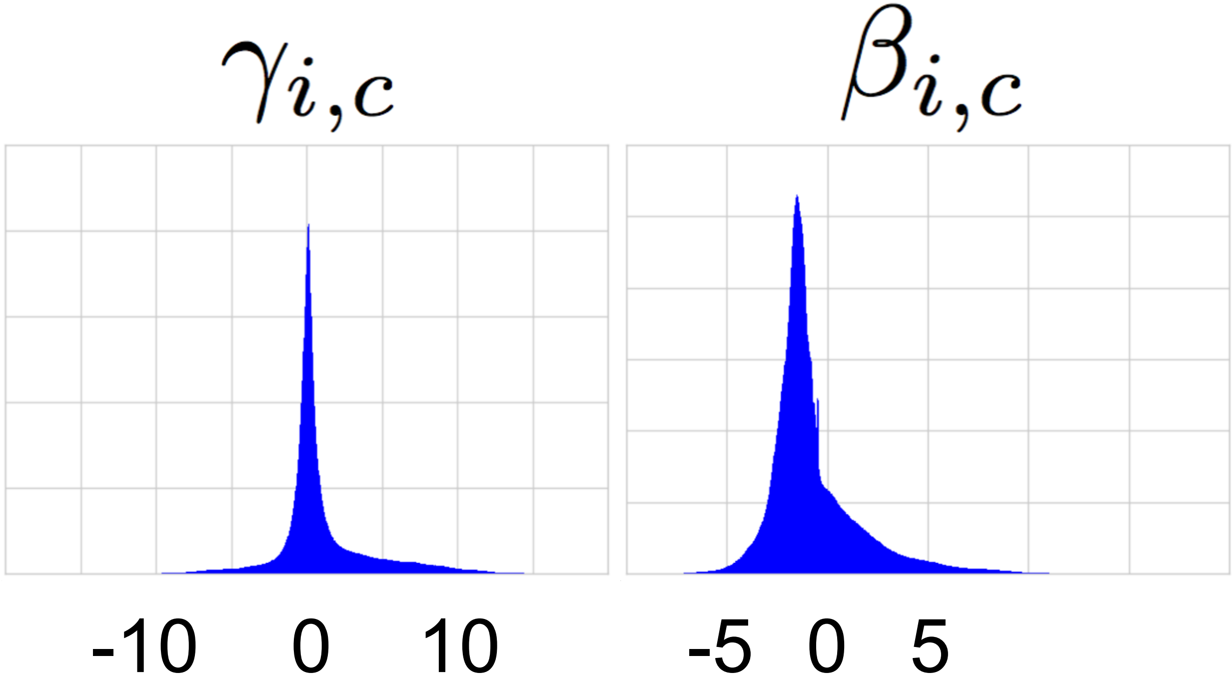
Dumoulin, Perez, Schucher et al. "Feature-wise transformations". Distill 2018.

FiL**M**: **F**eature-w**i**se **L**inear **M**odulation



Perez et al. "FiLM: Visual Reasoning with a General Conditioning Layer". AAAI 2018.

Histogram of FiLM Parameter Values



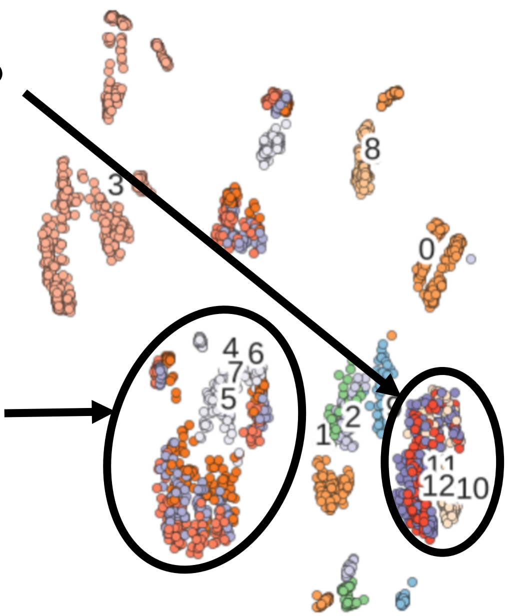
Perez et al. "FiLM: Visual Reasoning with a General Conditioning Layer". AAAI 2018.

t-SNE of FiLM Parameter Values

<u>Last FiLM Layer</u>

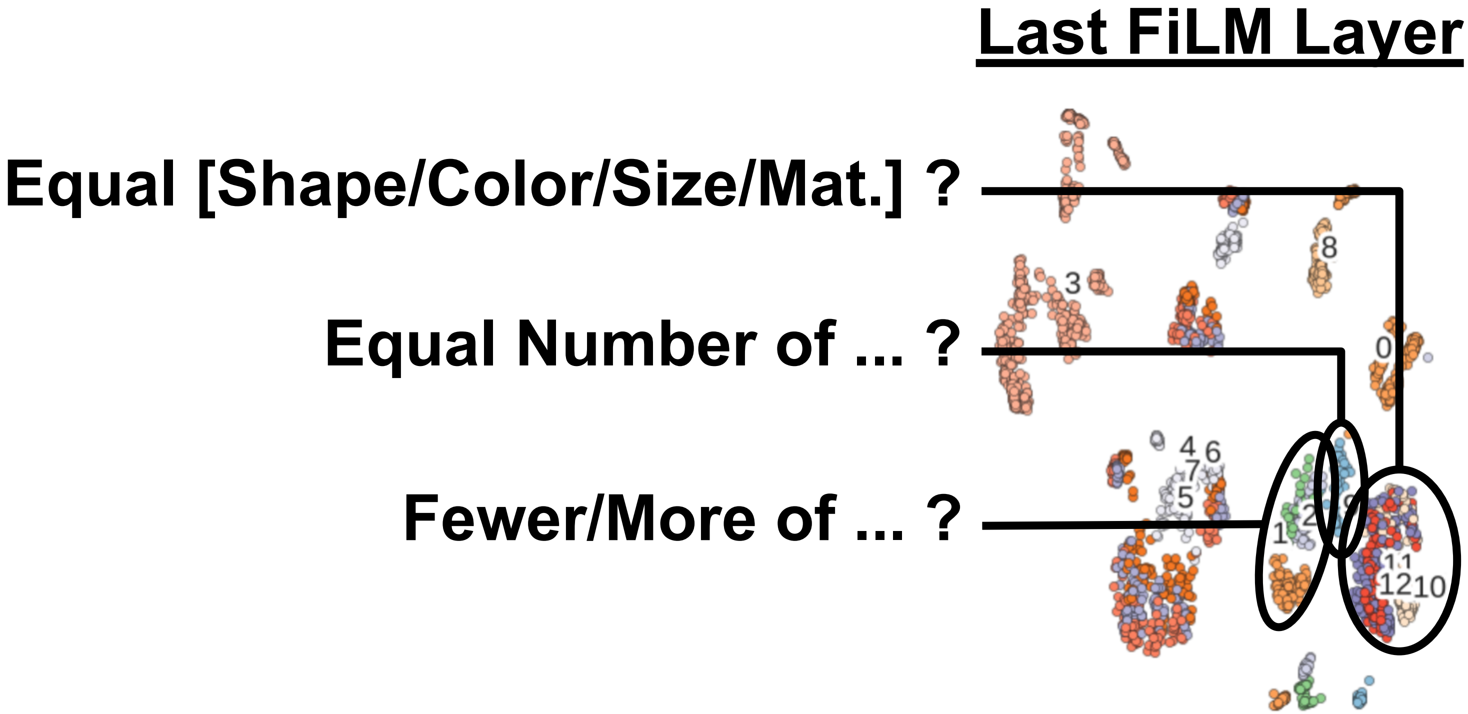
Equal [Shape/Color/Size/Mat.] ?

What [Shape/Color/Size/Mat.] ?



Perez et al. "FiLM: Visual Reasoning with a General Conditioning Layer". AAAI 2018.

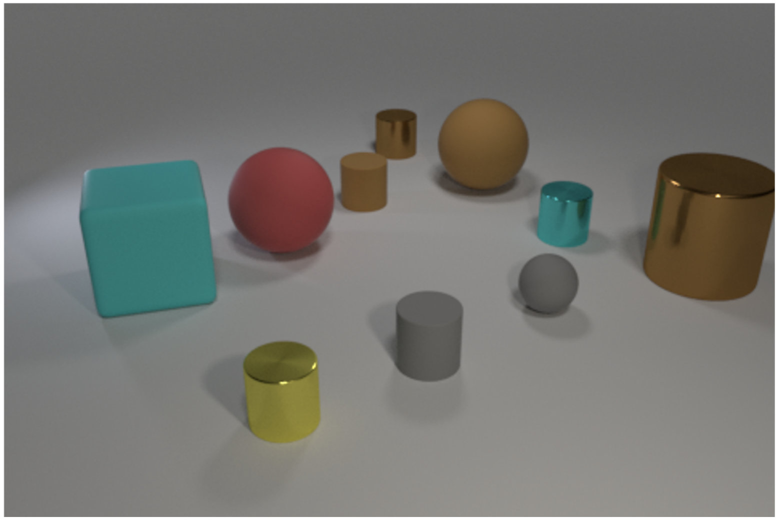
t-SNE of FiLM Parameter Values



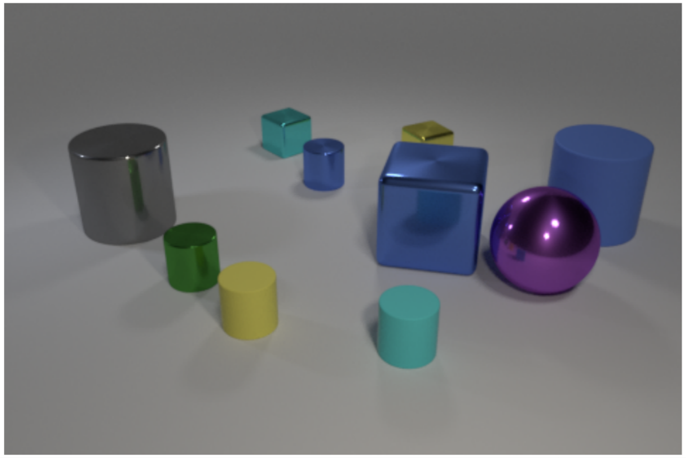
Perez et al. "FiLM: Visual Reasoning with a General Conditioning Layer". AAAI 2018.

Live Demo

- Example Questions:
  - “What is the shape of the gray matte object to the right of the large ball that is right of the yellow cylinder?”
  - “What number of things are matte objects that are behind the large cube or big purple shiny balls?”
  - “How many…”
  - “What material is…”
  - “Is there…”
  - “Are there more…”

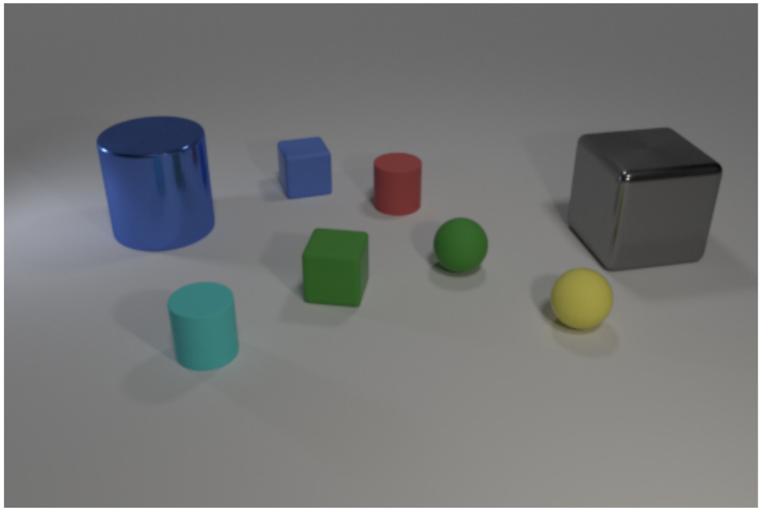


Logical Inconsistency



Question	Answer
How many gray things are there?	1
How many cyan things are there?	2
Are there as many gray things as cyan things?	**Yes**
Are there more gray things than cyan things?	No
Are there fewer gray things than cyan things?	Yes

Zero-Shot Generalization with FiLM

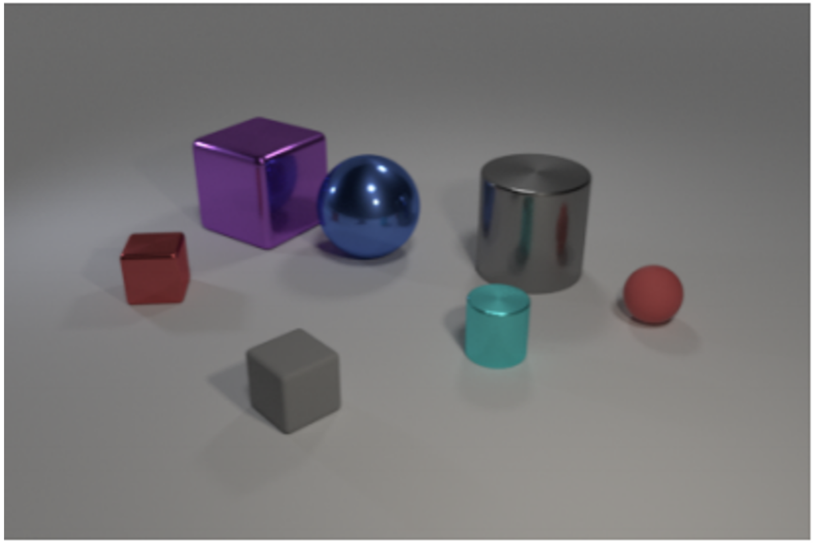


	Question	What is the blue big cylinder made of?
	(1) Swap shape	What is the blue big **sphere** made of?
+	(2) Swap color	What is the **green** big cylinder made of?
-	(3) Swap shape/color	What is the **green** big **sphere** made of?

Perez et al. "FiLM: Visual Reasoning with a General Conditioning Layer". AAAI 2018.

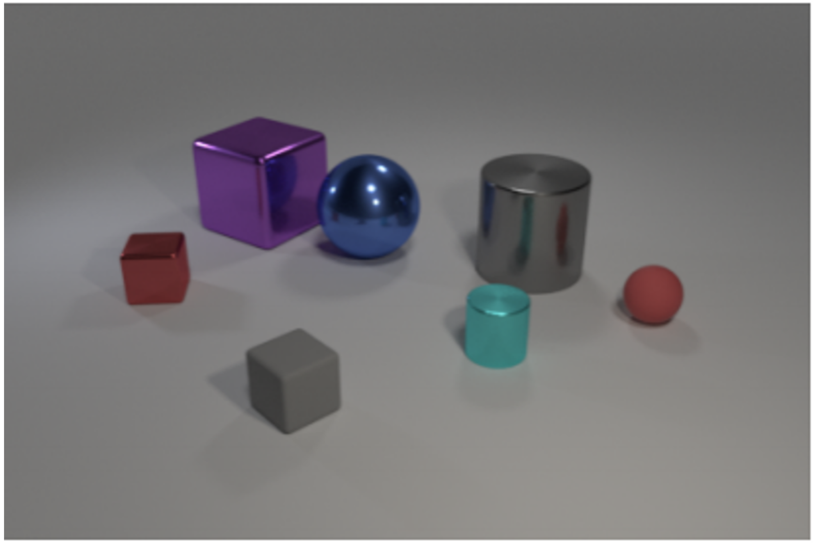
Activation Visualizations

Q: *What shape is the...* *...purple thing?* **A:** *cube*



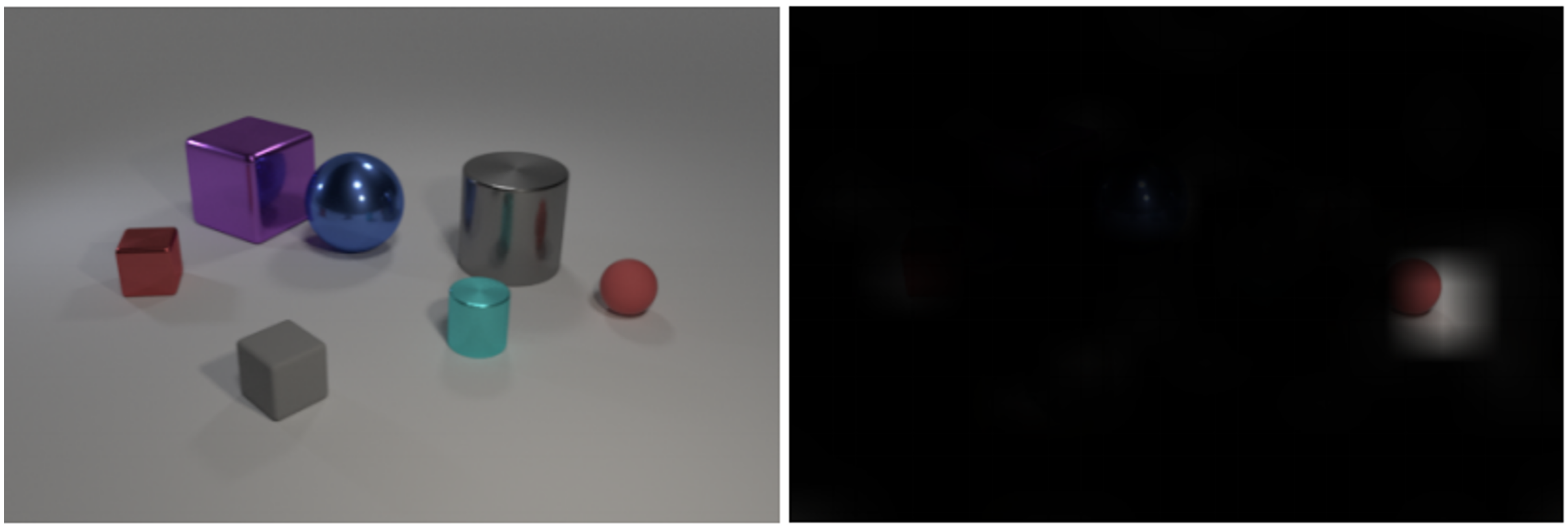
Activation Visualizations

Q: *What shape is the...* *...blue thing?* **A:** *sphere*



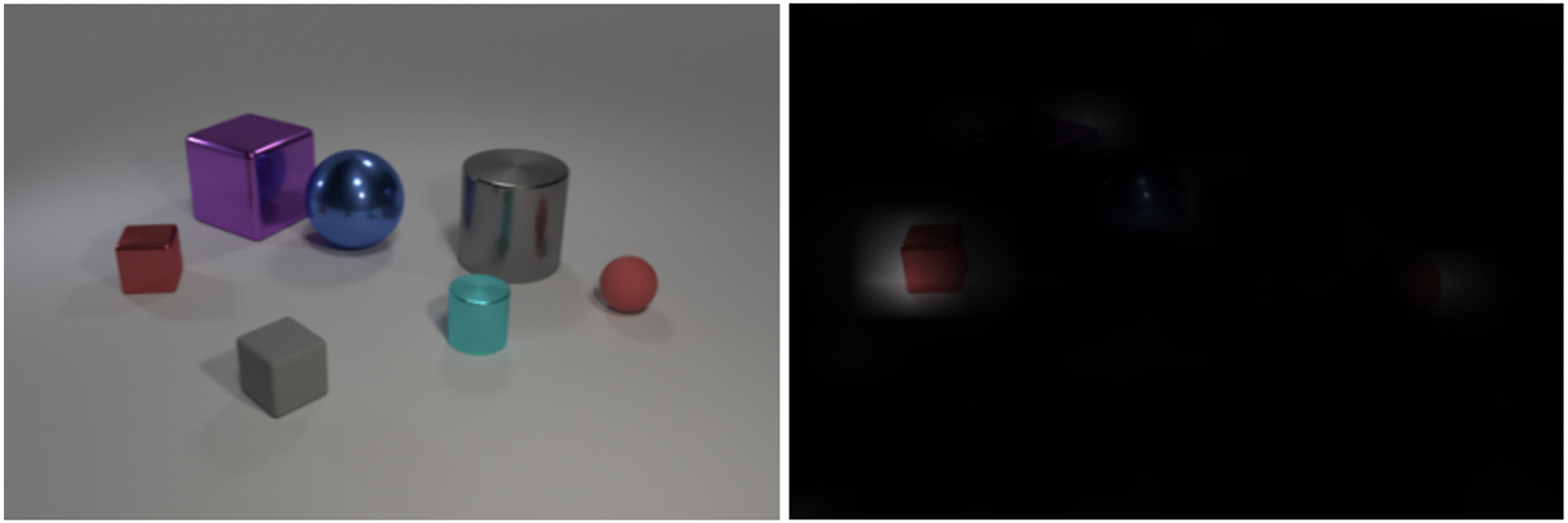
Activation Visualizations

Q: *What shape is the... ...red thing right of the blue thing?* **A:** *sphere*



Activation Visualizations

Q: *What shape is the...* *...red thing left of the blue thing?* **A:** *cube*



Challenges in VQA

- Image representation
- Language representation
- Combining the modalities
- Attention
- Question-specific reasoning

Slide credit: Devi Parikh

Standard Approach

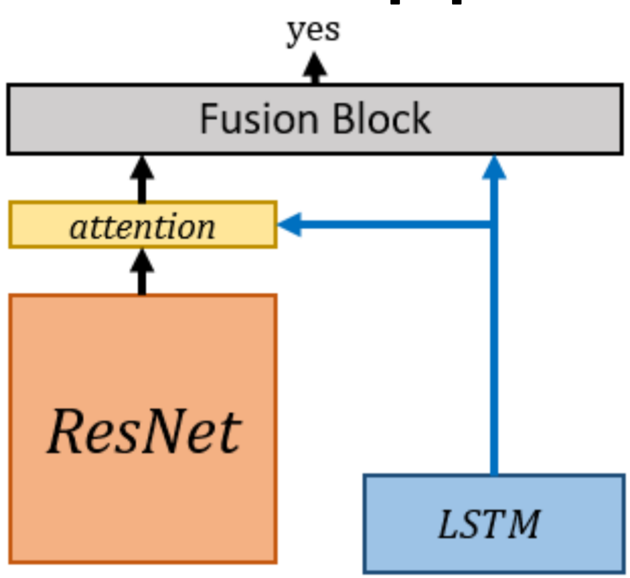


Figure from de Vries et al. "Modulating early visual processing by language." arXiv 2017.

Slide credit: Devi Parikh

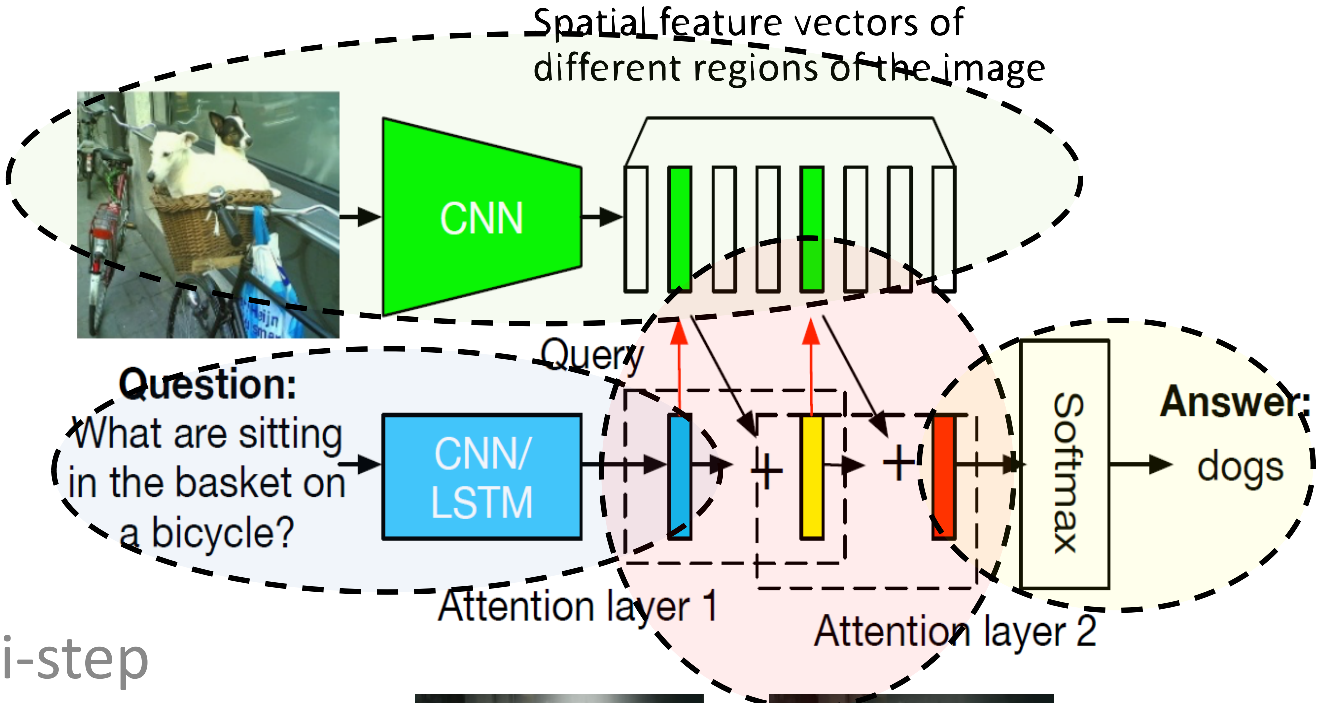
[Yang 2016] Stacked Attention Network (SAN)

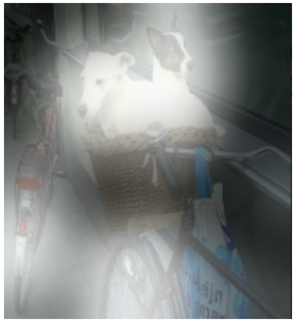
Original Image **First Attention Layer** **Second Attention Layer**

What are sitting in the basket of a bicycle?

Slide credit: Akira Fukui and Marcus Rohrbach

Stacked Attention Networks



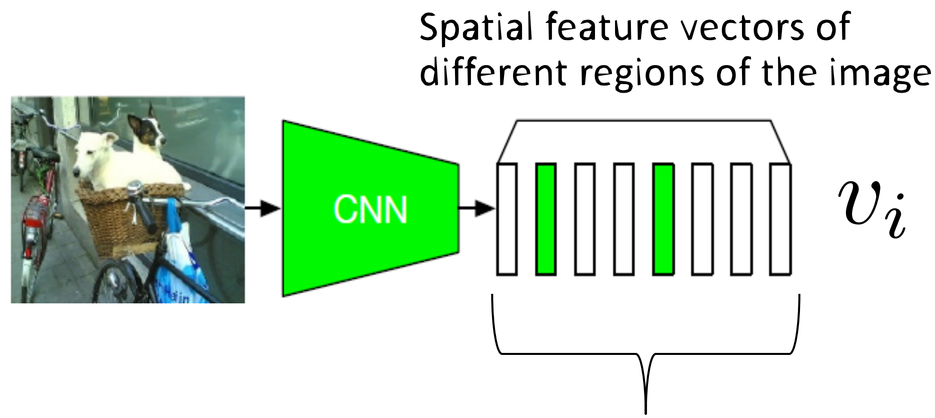


SANs perform multi-step reasoning

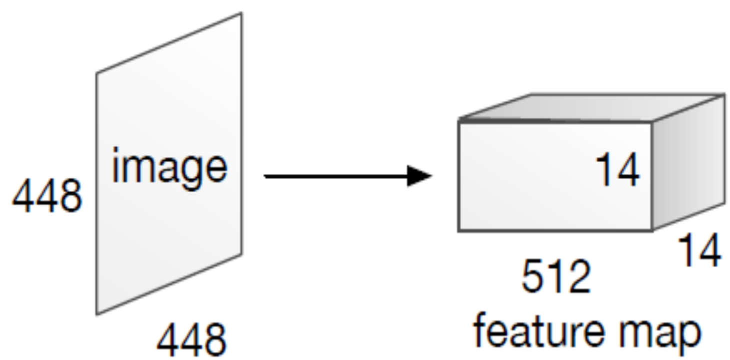
1. Question model
2. Image model
3. Multi-level attention model
4. Answer predictor

Slide credit: Xiaodong He

1. The image model in the SAN



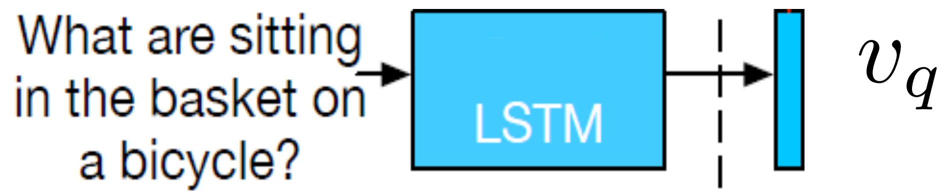
v_i 196 vectors (14 x14)

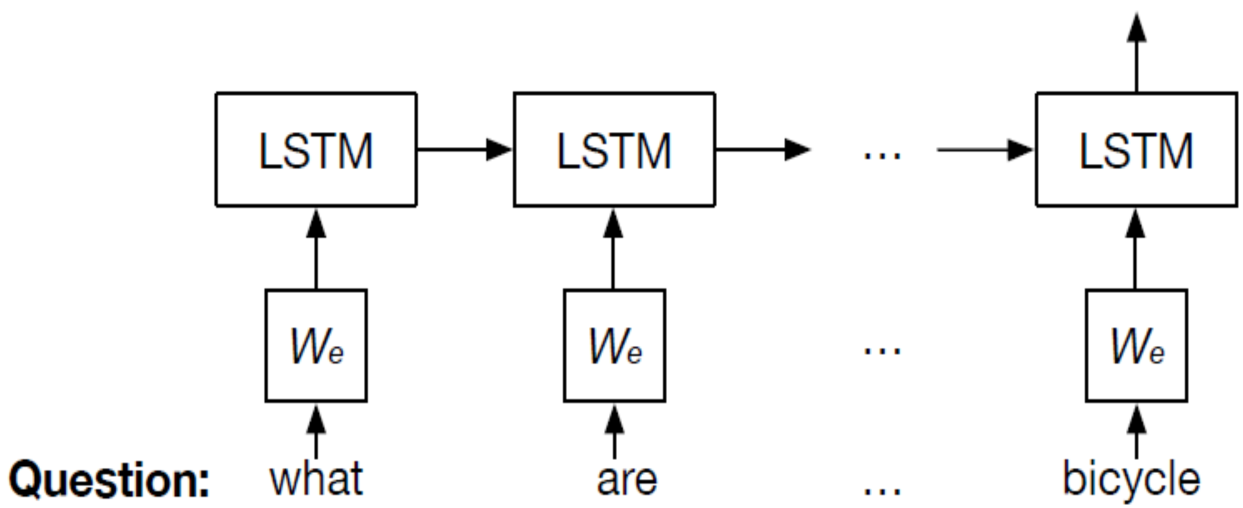


Slide credit: Adapted from Xiaodong He

2. The question model in the SAN

Code the question into a vector using an LSTM

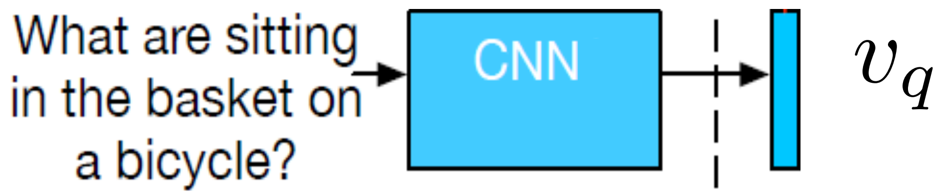


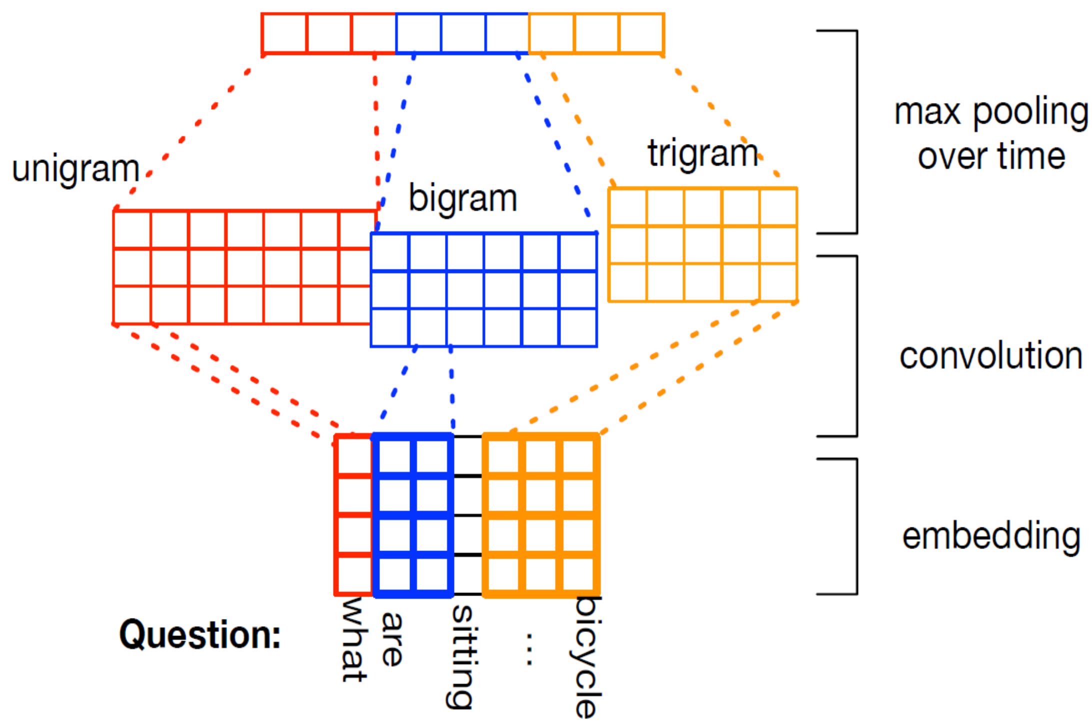


Slide credit: Adapted from Xiaodong He

2. The question model in the SAN (alternative)

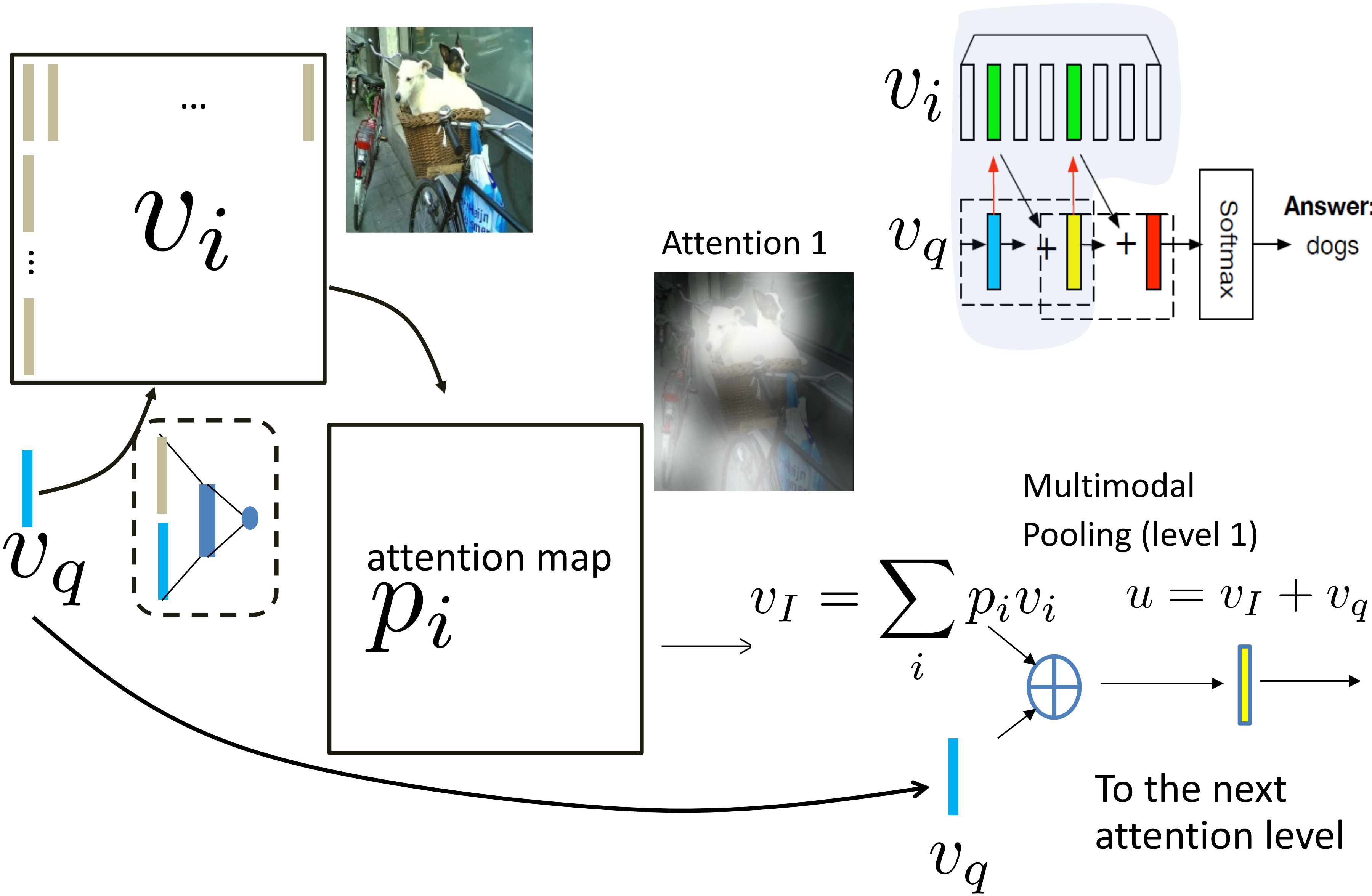
Code the question into a vector using a CNN





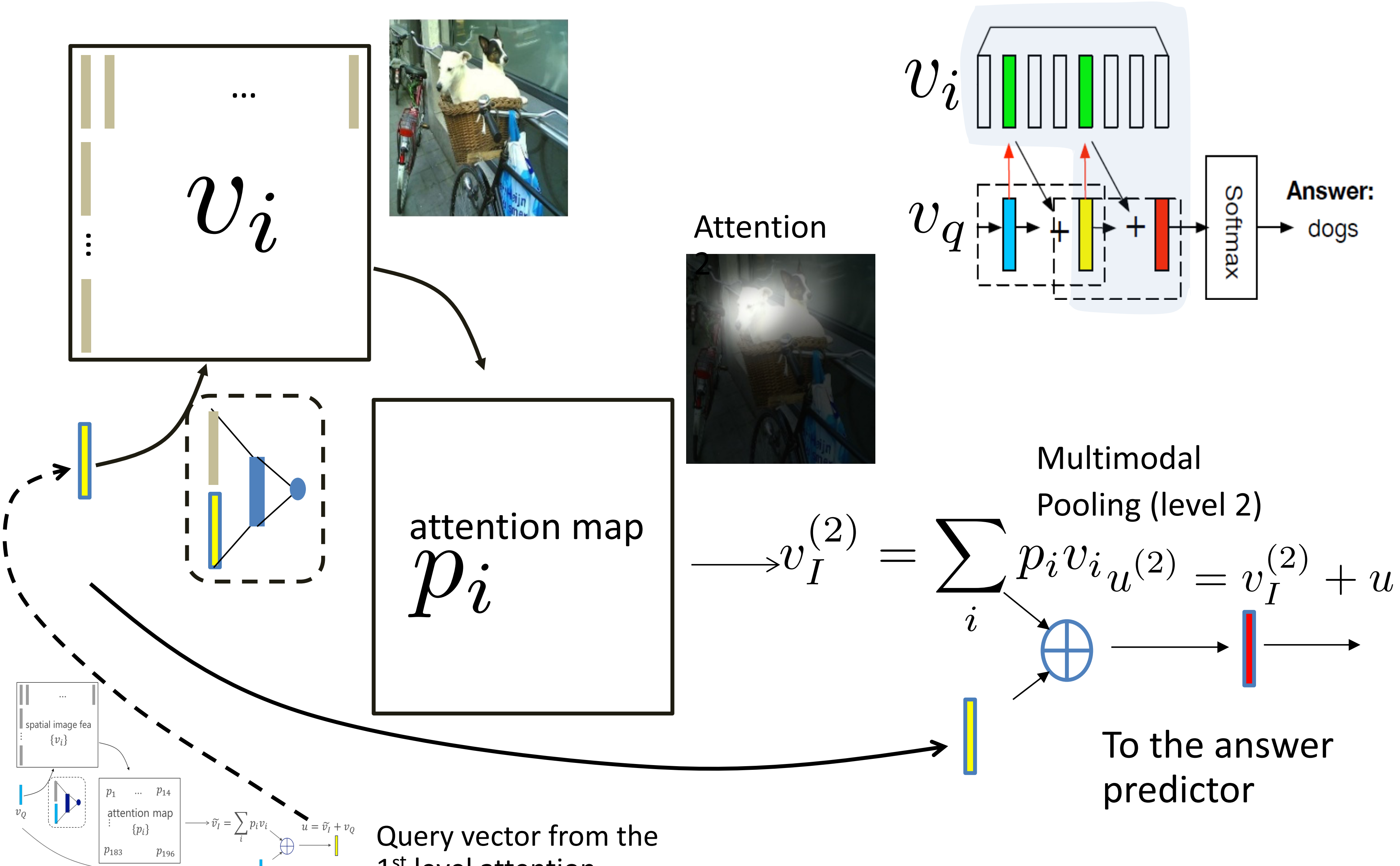
Slide credit: Xiaodong He

3. SAN: Computing the 1st level attention

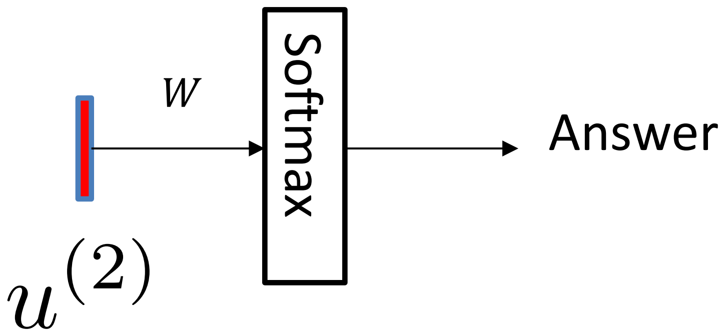


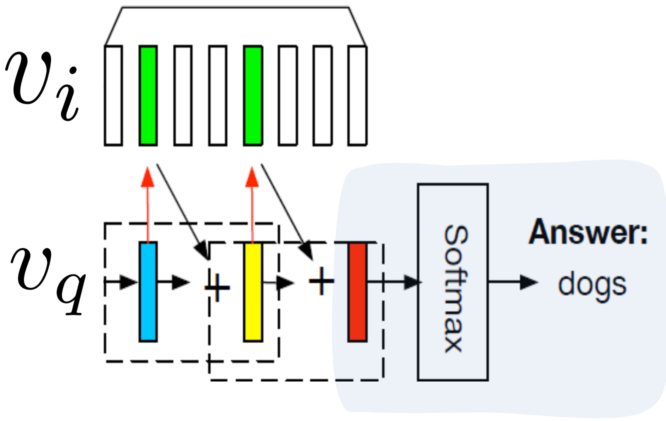
Slide credit: Adapted from Xiaodong He

3. SAN: Compute the 2nd level attention



4. Answer prediction





Results

	test-dev				test-std
Methods	All	Yes/No	Number	Other	All
VQA: [1]					
Question	48.1	75.7	36.7	27.1	-
Image	28.1	64.0	0.4	3.8	-
Q+I	52.6	75.6	33.7	37.4	-
LSTM Q	48.8	78.2	35.7	26.6	-
LSTM Q+I	53.7	78.9	35.2	36.4	54.1
SAN(2, CNN)	**58.7**	**79.3**	**36.6**	**46.1**	**58.9**

Table 5: VQA results on the official server, in percentage

Other:
Object
Color
Location
…

Big improvement on the VQA benchmark (and COCO-QA, DAQUAR)
Improvement is mainly in the *Other* category.

Slide credit: Xiaodong He

Results

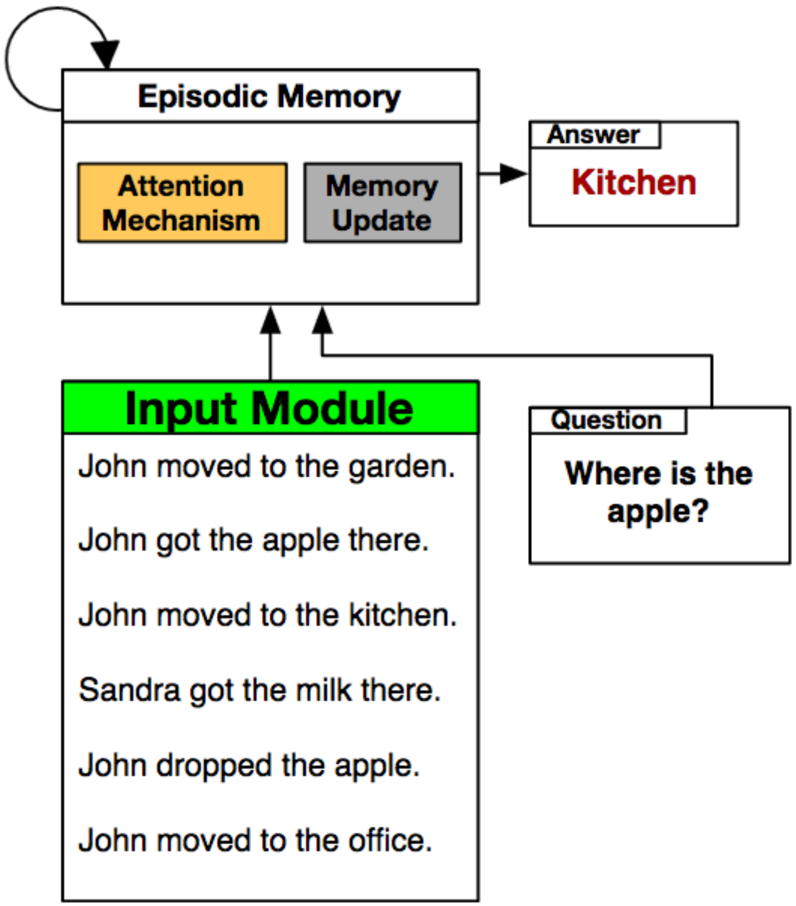
Methods	All	Yes/No 36%	Number 10%	Other 54%
SAN(1, LSTM)	56.6	78.1	41.6	44.8
SAN(1, CNN)	56.9	78.8	42.0	45.0
SAN(2, LSTM)	57.3	78.3	**42.2**	45.9
SAN(2, CNN)	**57.6**	78.6	41.8	**46.4**

Table 6: VQA results on our partition, in percentage

Using multi-level attentions improve the performance significantly (also mainly in the *Other* category)

Slide credit: Xiaodong He

[Xiong 2016]



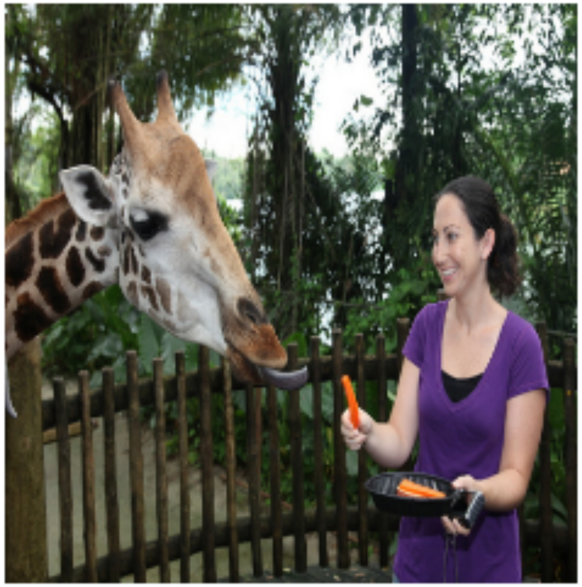
Text Question-Answering

Slide credit: Devi Parikh

MCB: Attention Visualizations

What is the woman feeding the giraffe?

Carrot

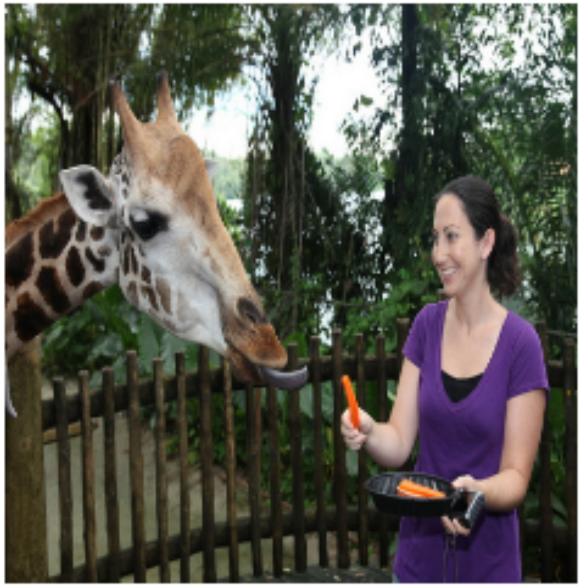


Slide credit: Akira Fukui and Marcus Rohrbach

MCB: Attention Visualizations

What is her hairstyle for the picture?

Ponytail



Slide credit: Akira Fukui and Marcus Rohrbach

MCB: Attention Visualizations

What color is the chain on the red dress?

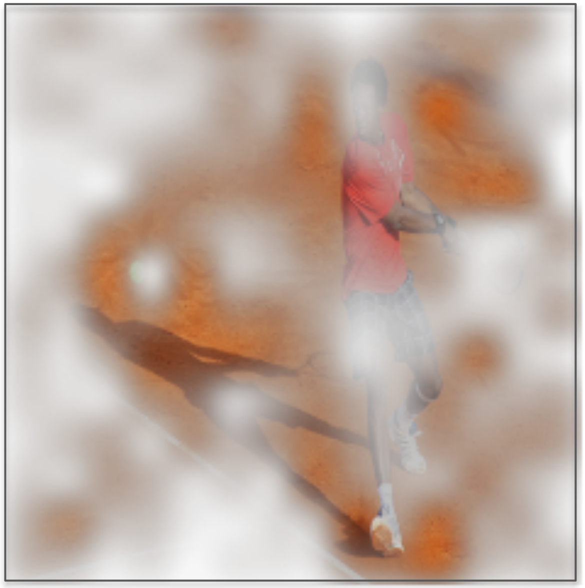
Pink

- Correct Attention, Incorrect Fine-grained Recognition

MCB: Attention Visualizations

Is the man going to fall down?

No

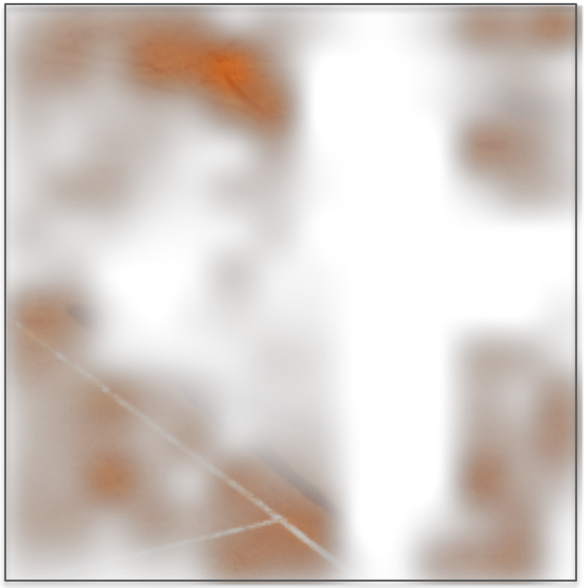


Slide credit: Akira Fukui and Marcus Rohrbach

MCB: Attention Visualizations

What is the surface of the court made of?

Clay

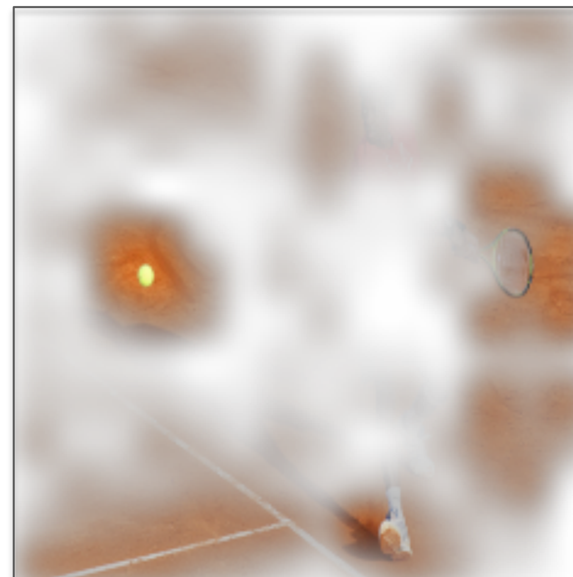


Slide credit: Akira Fukui and Marcus Rohrbach

MCB: Attention Visualizations

What sport is being played?

Tennis



Slide credit: Akira Fukui and Marcus Rohrbach

MCB: Attention Visualizations

What does the shop sell?

Clocks

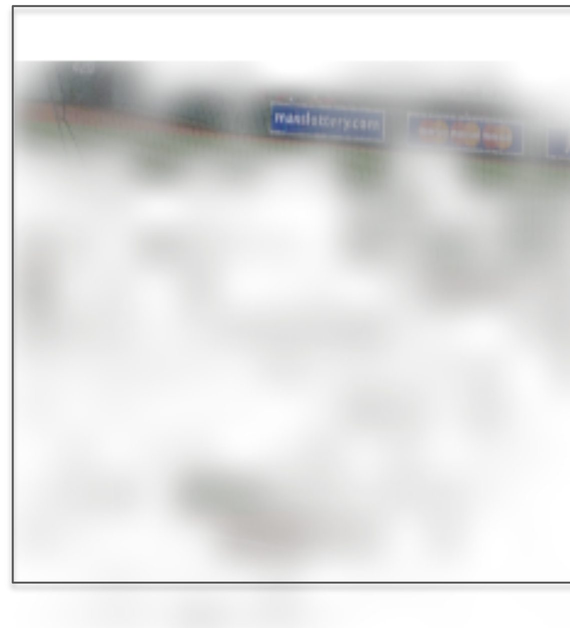
- Incorrect Attention

Slide credit: Akira Fukui and Marcus Rohrbach

MCB: Attention Visualizations

What credit card company is on the banner in the background?

Budweiser



- Correct Attention, Incorrect Concept Association

Challenges in VQA

- Image representation
- Language representation
- Combining the modalities
- Attention
- Question-specific reasoning

Slide credit: Devi Parikh

Neural Module Network (NMN) [Andreas 2016]

Slide credit: Devi Parikh

Grounded question answering

What color is the necktie?

yellow

Slide credit: Jacob Andreas

Grounded question answering

What rivers are in South Carolina?

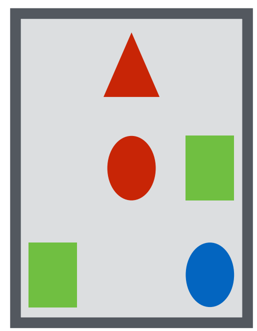
name	type	coastal
Columbia	city	no
Cooper	river	yes
Charleston	city	yes

Cooper

Slide credit: Jacob Andreas

Grounded question answering

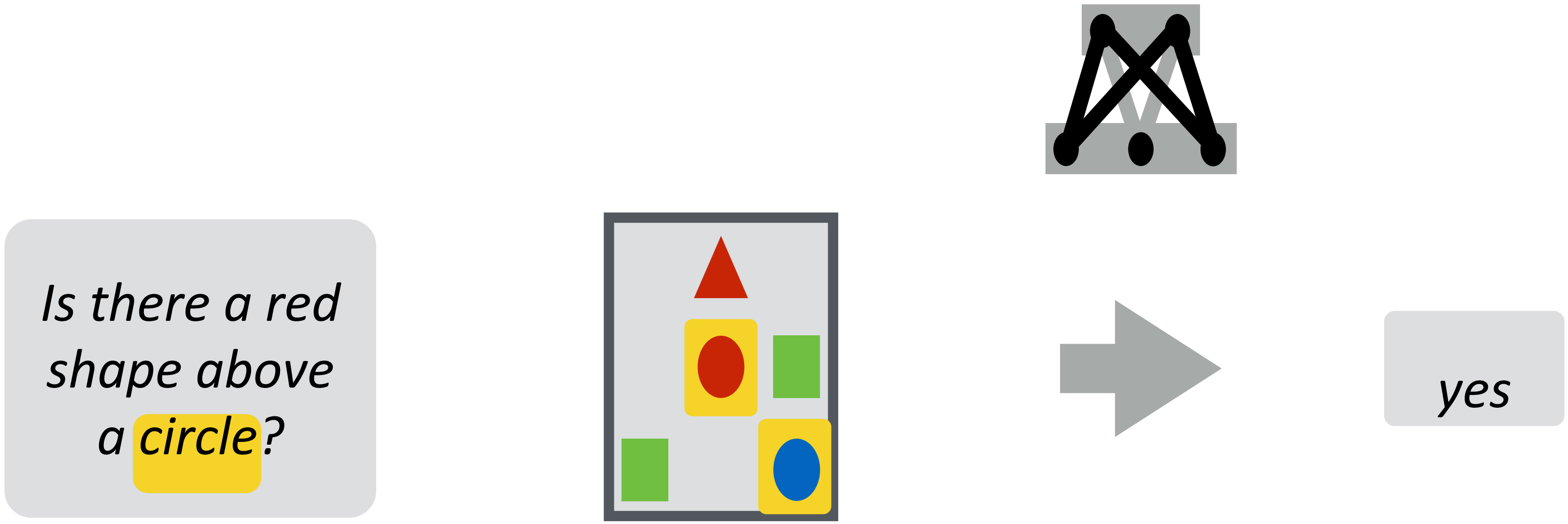
Is there a red shape above a circle?



yes

Slide credit: Jacob Andreas

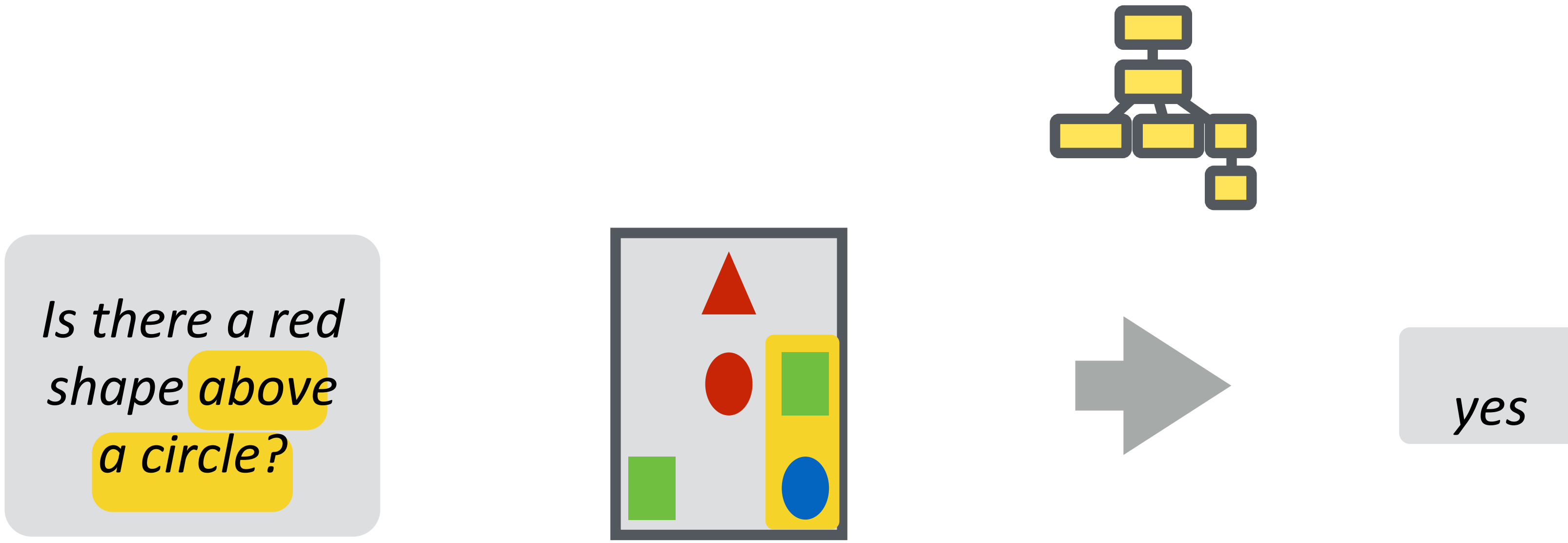
Neural nets learn lexical groundings



[Iyyer et al. 2014, Bordes et al. 2014,
Yang et al. 2015, Malinowski et al., 2015]

Slide credit: Jacob Andreas

Semantic parsers learn composition

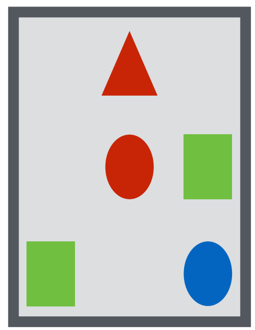


[Wong & Mooney 2007, Kwiatkowski et al. 2010, Liang et al. 2011, A et al. 2013]

Slide credit: Jacob Andreas

Neural module networks learn both!

Is there a red shape above a circle?

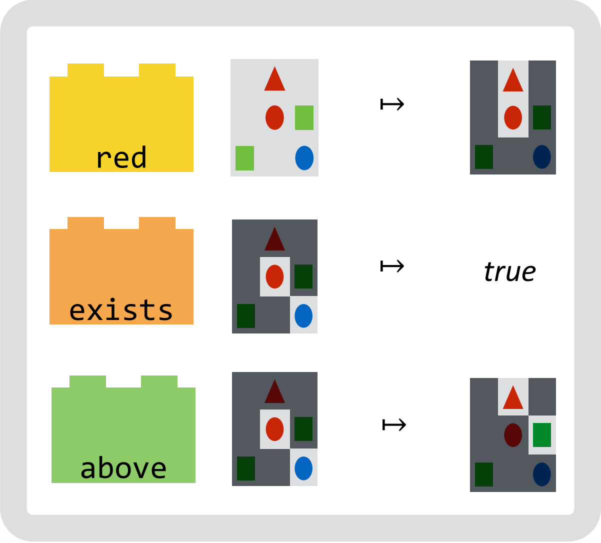


yes

Slide credit: Jacob Andreas

Neural module networks

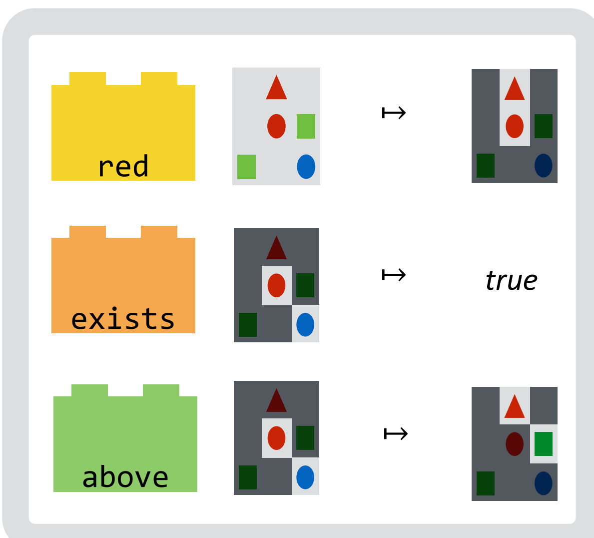
Is there a red shape above a circle?

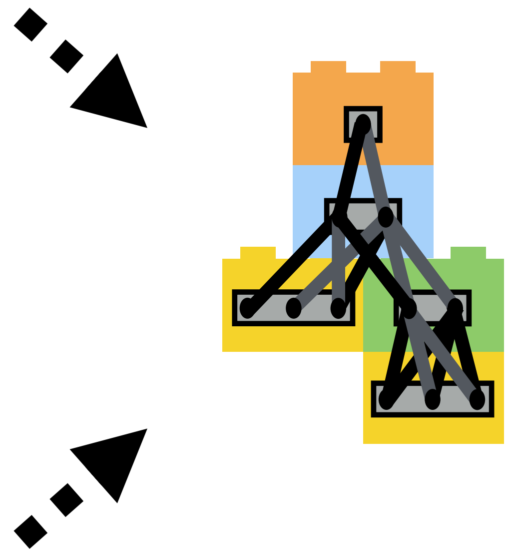


Slide credit: Jacob Andreas

Neural module networks

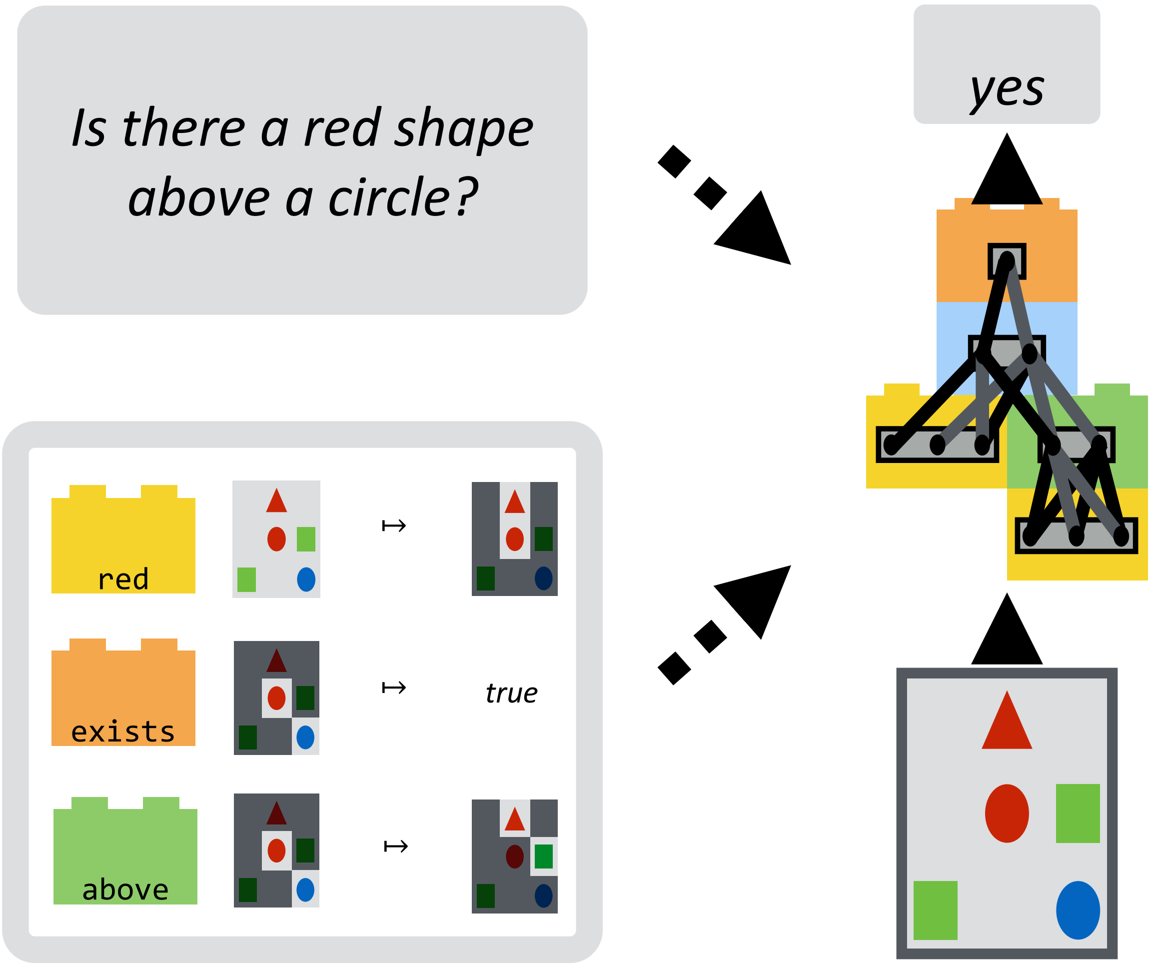
Is there a red shape above a circle?





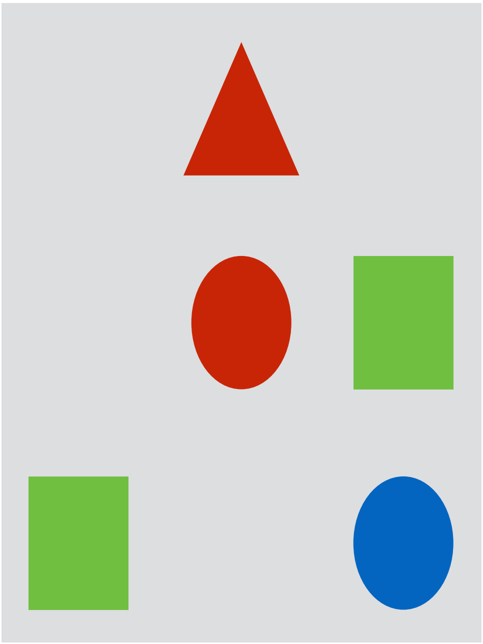
Slide credit: Jacob Andreas

Neural module networks



Slide credit: Jacob Andreas

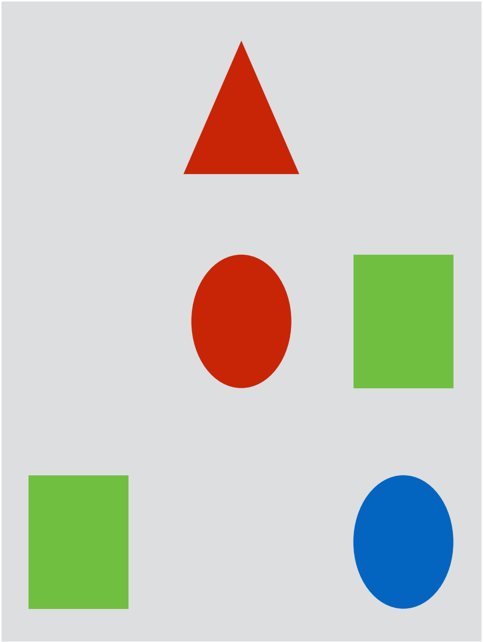
Representing meaning



Is there a red shape above a circle?

Slide credit: Jacob Andreas

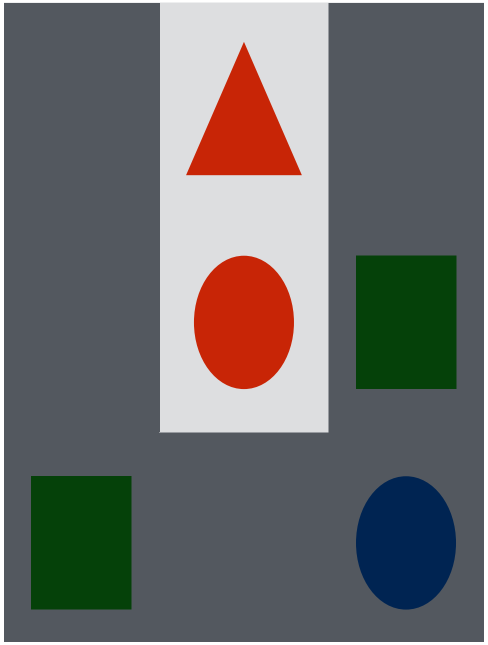
Representing meaning



Is there a red shape above a circle?

Slide credit: Jacob Andreas

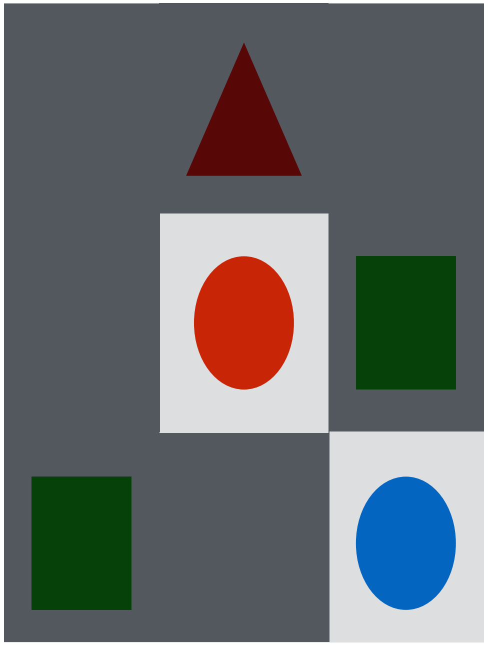
Sets encode meaning



Is there a red shape above a circle?

Slide credit: Jacob Andreas

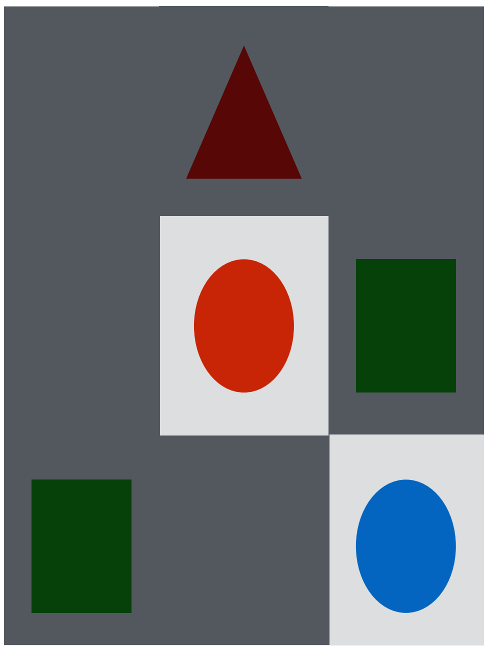
Sets encode meaning



Is there a red shape above a circle?

Slide credit: Jacob Andreas

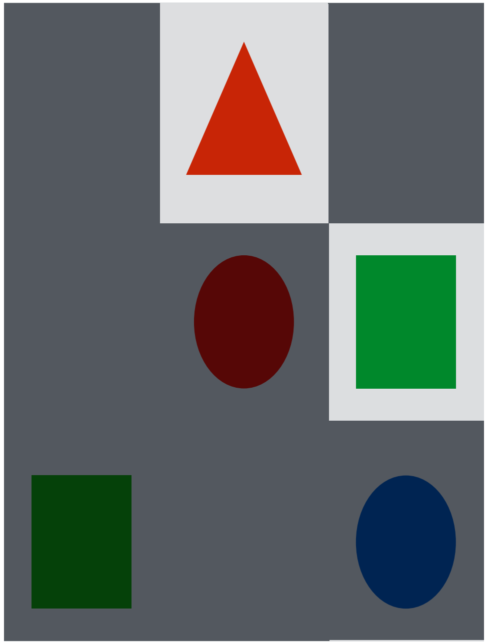
Set transformations encode meaning



Is there a red shape above a circle?

Slide credit: Jacob Andreas

Set transformations encode meaning

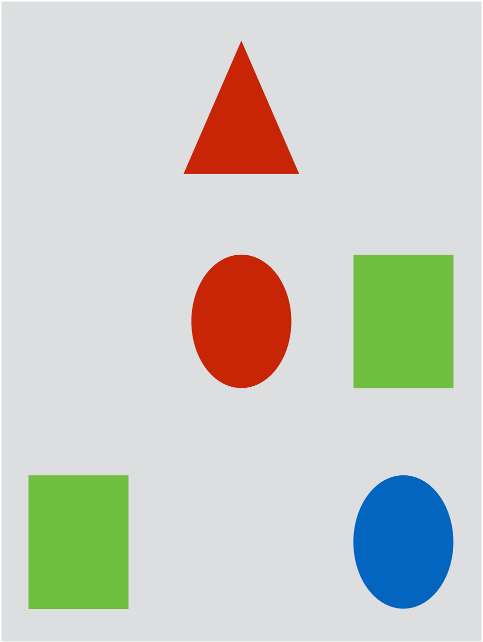


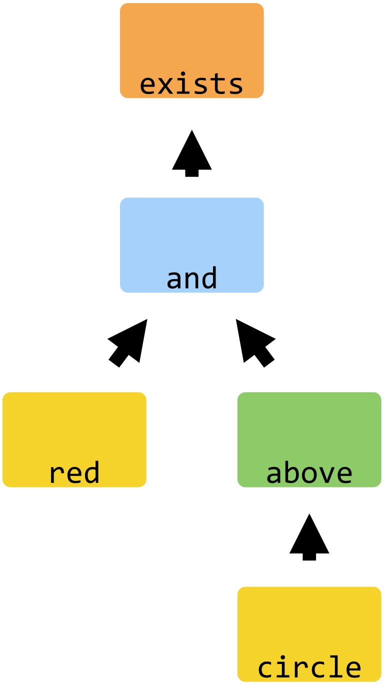
Is there a red shape above a circle?

Slide credit: Jacob Andreas

Sentence meanings are computations

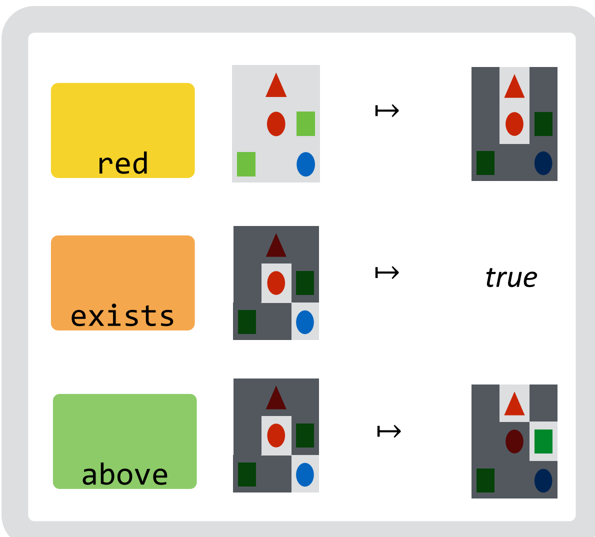
Is there a red shape above a circle?

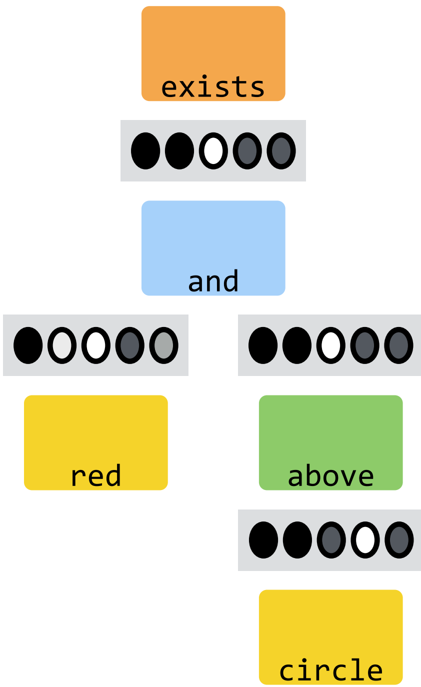




Slide credit: Jacob Andreas

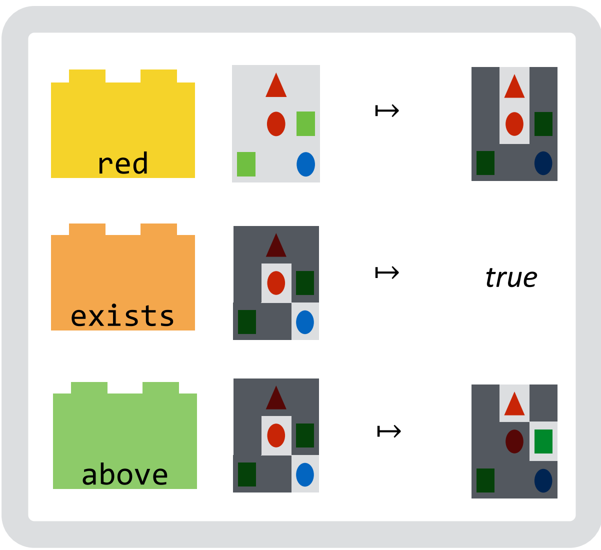
Composing vector functions

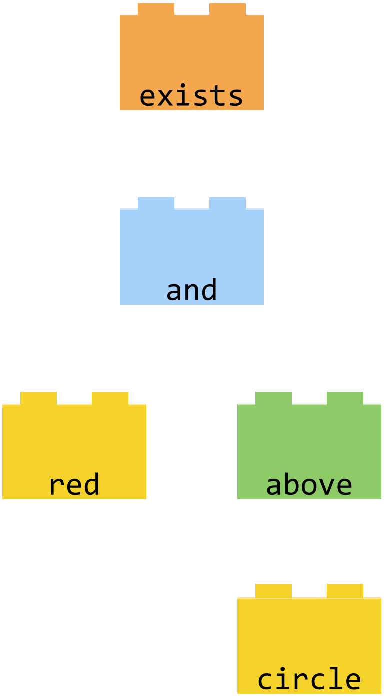




Slide credit: Jacob Andreas

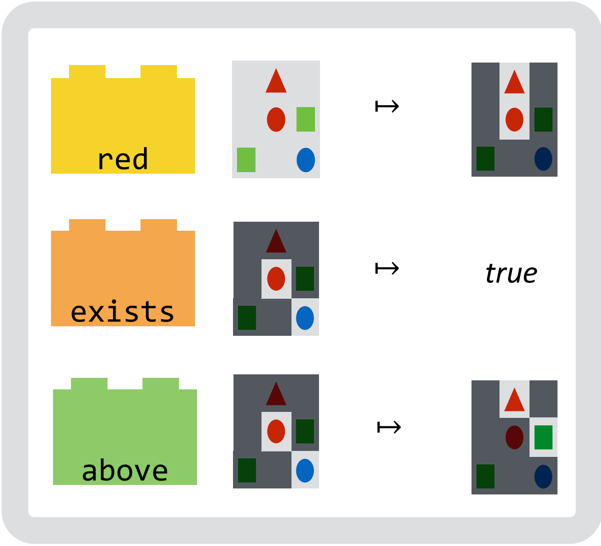
Composing vector functions

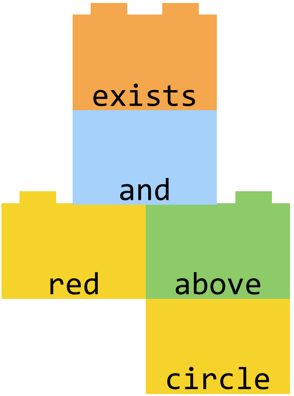




Slide credit: Jacob Andreas

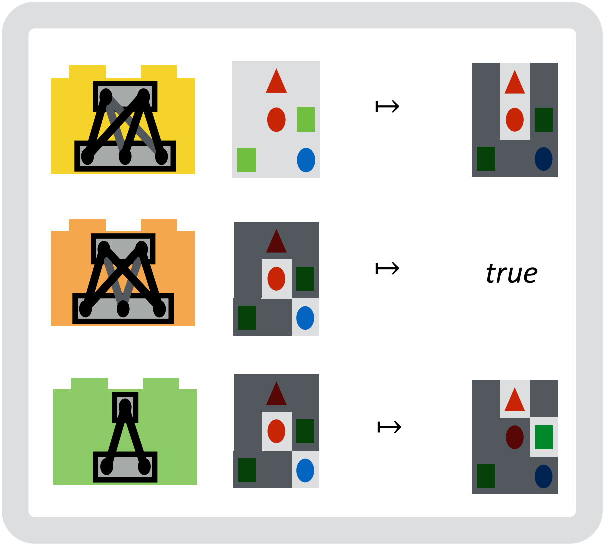
Composing vector functions

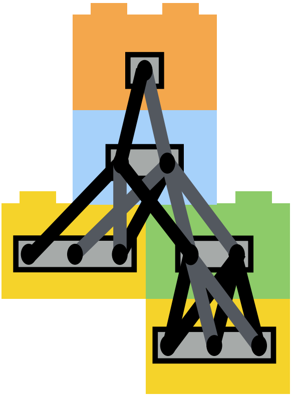




Slide credit: Jacob Andreas

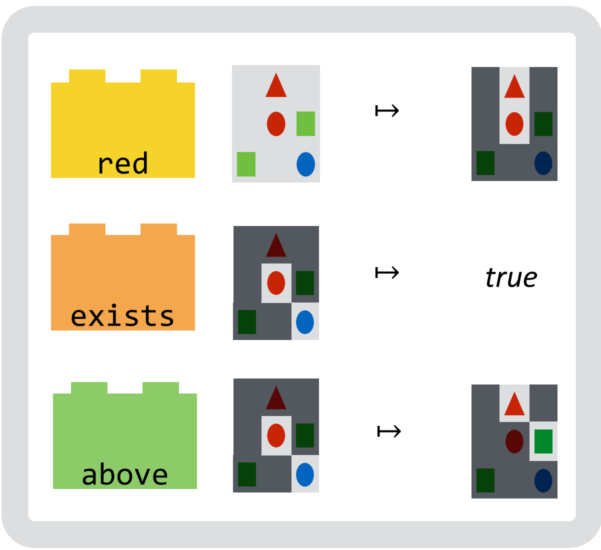
Compositions of vector functions are neural nets

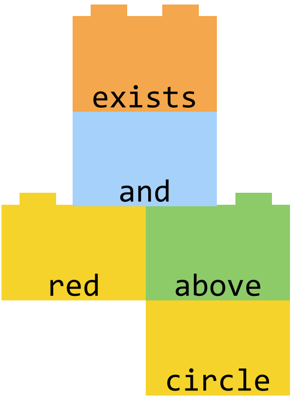




Slide credit: Jacob Andreas

Compositions of vector functions are neural nets





Slide credit: Jacob Andreas

What modules do we need?

Is there a red shape above a circle?

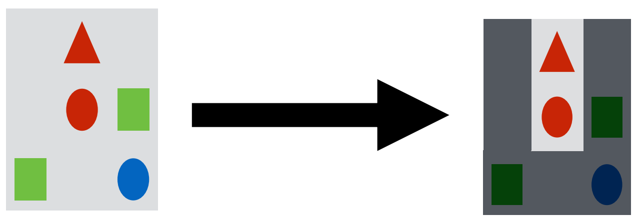
What color is the triangle?

How many goats are there?

What cities are south of San Diego?

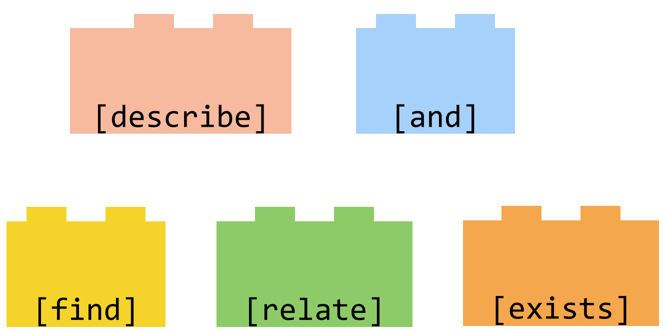
Slide credit: Jacob Andreas

Module inventory



Is there a red shape above a circle?

What color is the triangle?

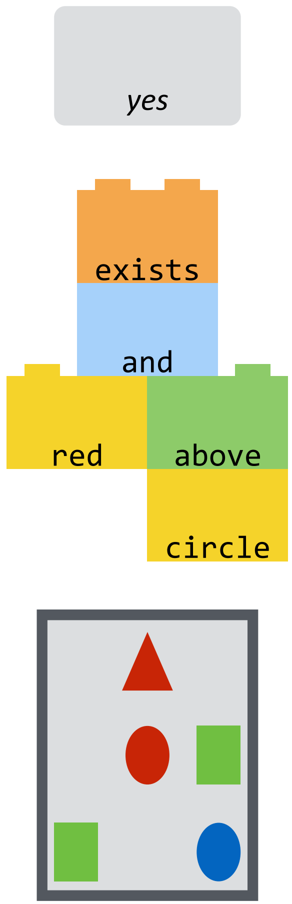


Who is running in the grass?

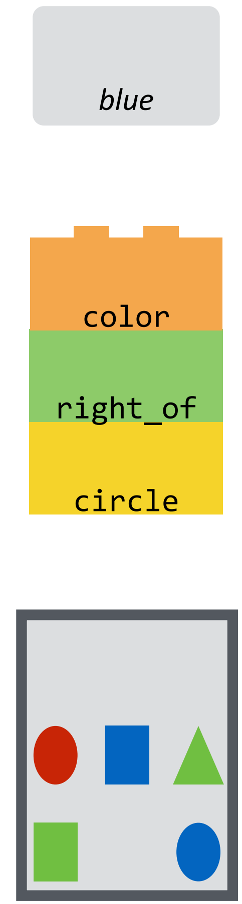
What cities are south of San Diego?

Slide credit: Jacob Andreas

Learning



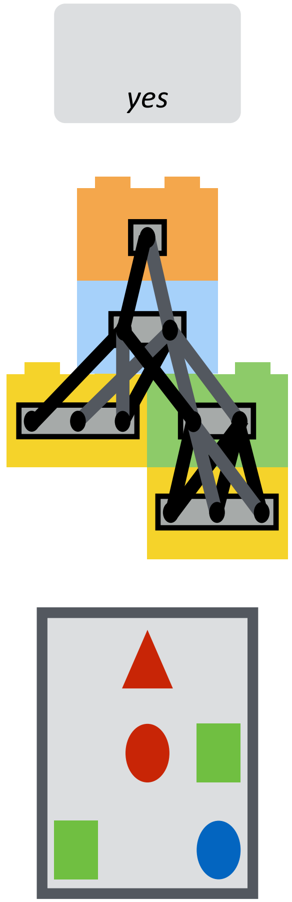
Is there a red shape above a circle?



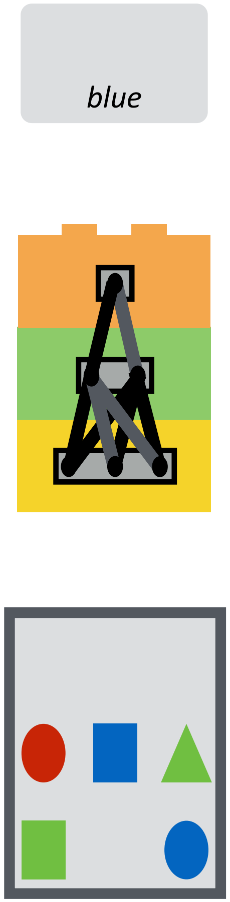
What color is the shape right of a circle?

Slide credit: Jacob Andreas

Learning



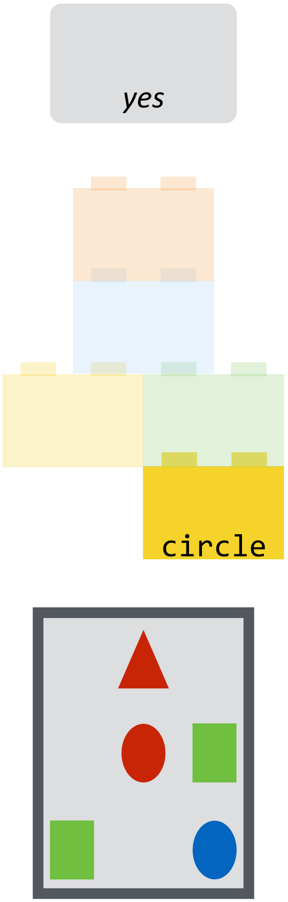
Is there a red shape above a circle?

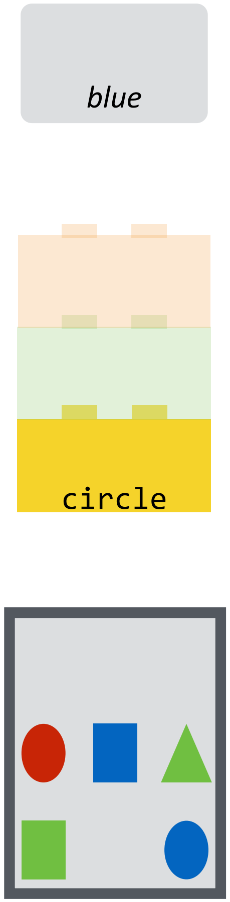


What color is the shape right of a circle?

Slide credit: Jacob Andreas

Parameter tying



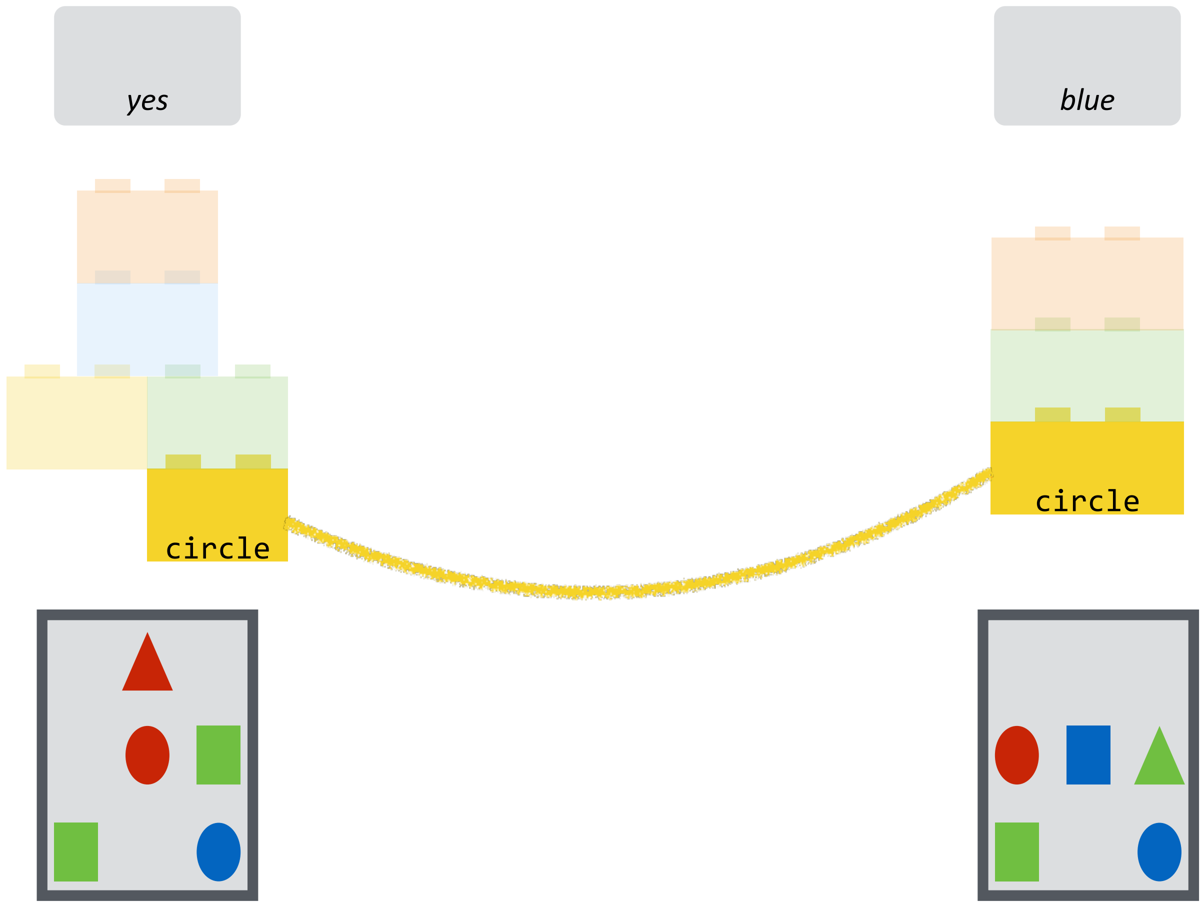


Is there a red shape above a circle?

What color is the shape right of a circle?

Slide credit: Jacob Andreas

Parameter tying

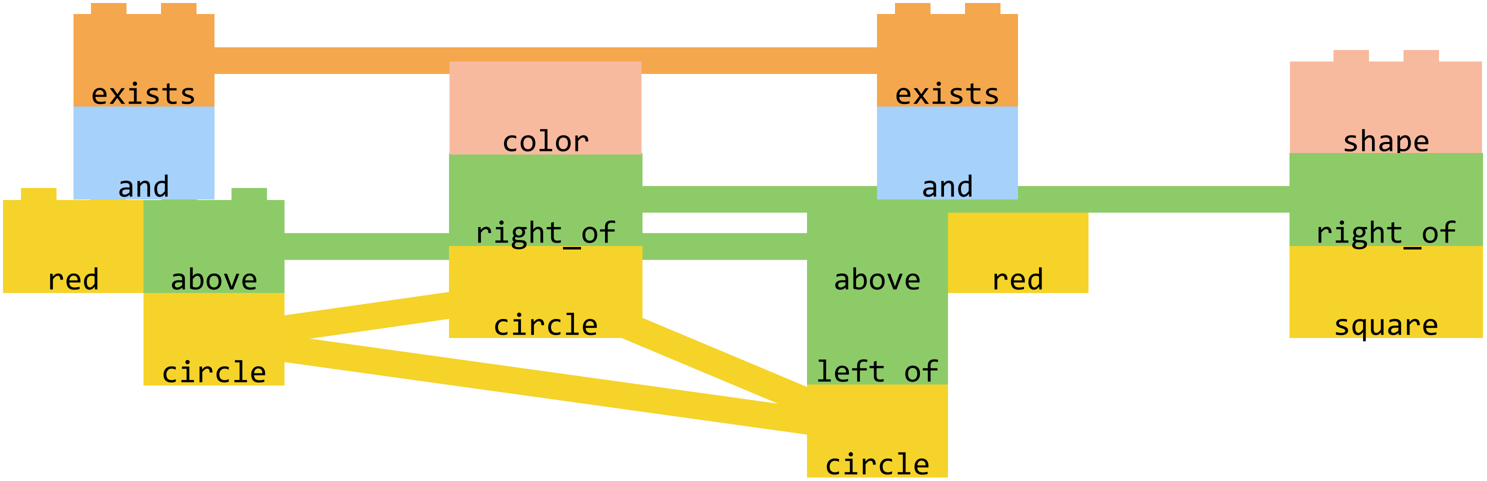


Is there a red shape above a circle?

What color is the shape right of a circle?

Slide credit: Jacob Andreas

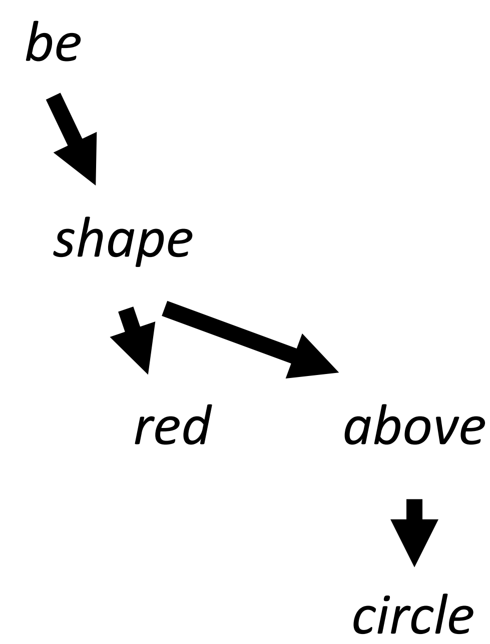
Extreme parameter tying



Slide credit: Jacob Andreas

Where do layouts come from?

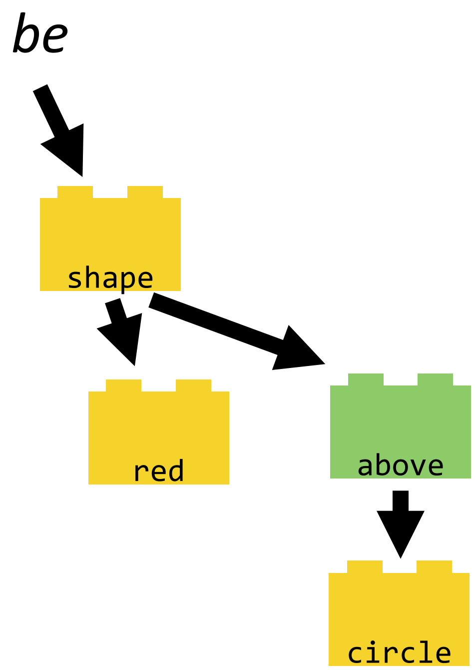
Is there a red shape above a circle?



Slide credit: Jacob Andreas

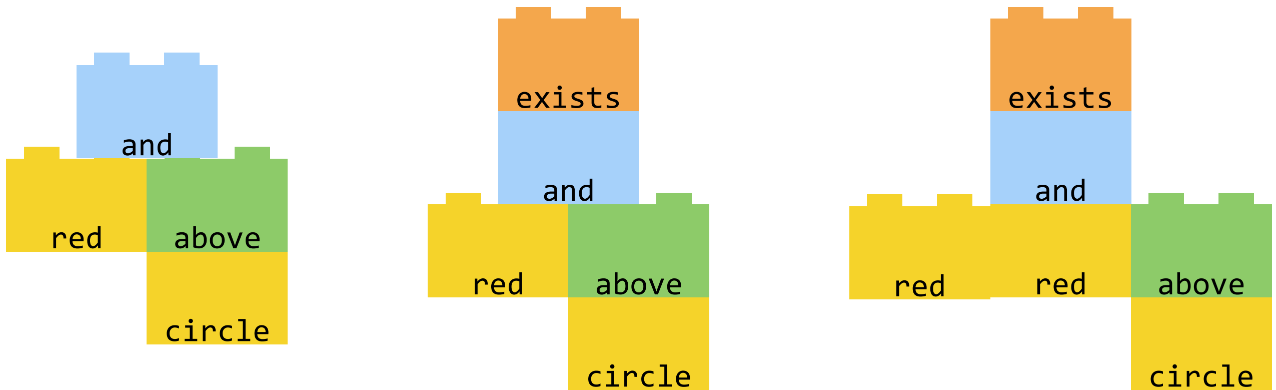
Where do layouts come from?

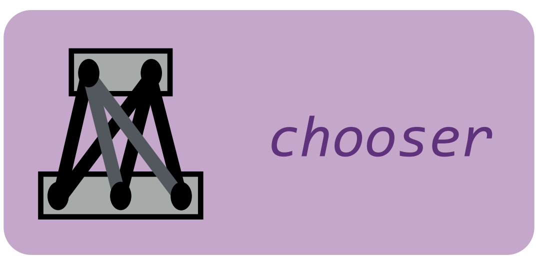
Is there a red shape above a circle?



Slide credit: Jacob Andreas

Choosing among layouts

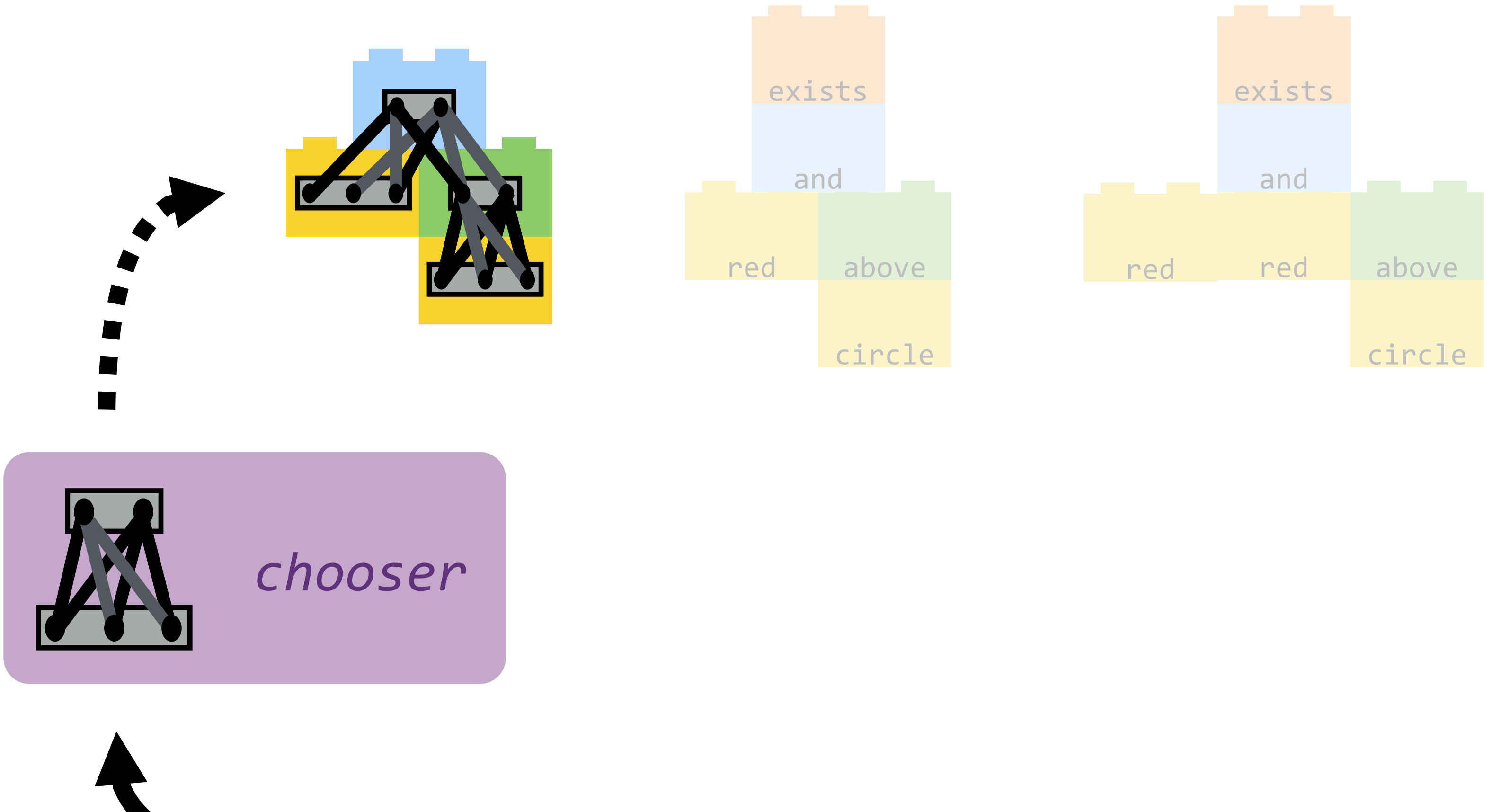




Is there a red shape above a circle?

Slide credit: Jacob Andreas

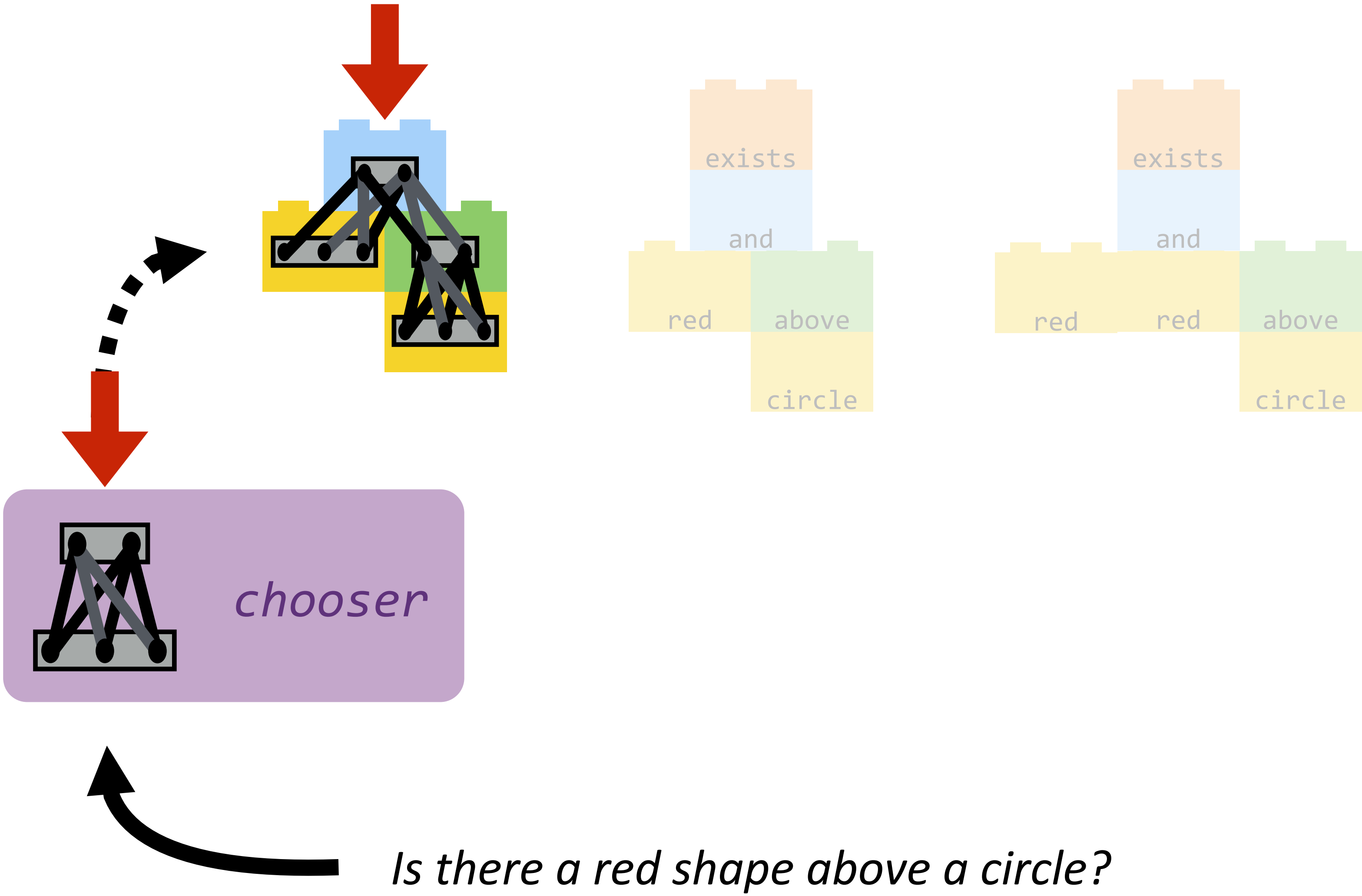
Learning to choose layouts



Is there a red shape above a circle?

Slide credit: Jacob Andreas

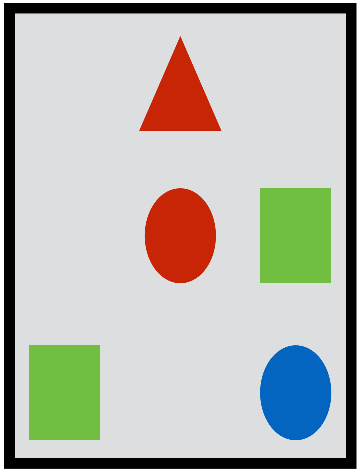
Learning with unknown layouts uses RL



[Williams 1992]

Slide credit: Jacob Andreas

Experiments



name	type	coastal
Columbia	city	no
Cooper	river	yes
Charleston	city	yes

Slide credit: Jacob Andreas

Experiments: VQA dataset

What color is the necktie?

yellow

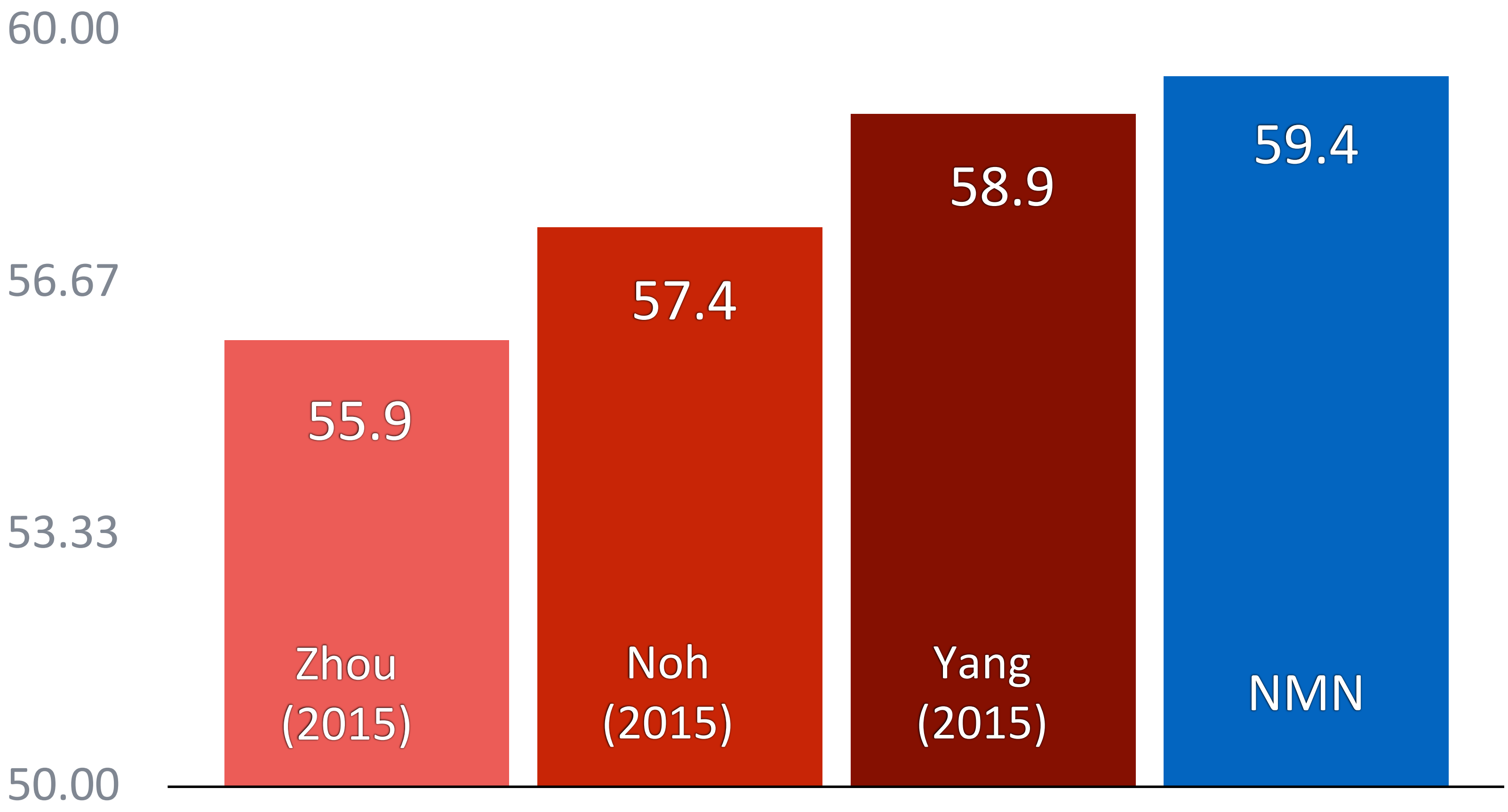
What is in the sheep's ear?

tag

[Antol et al. 2015]

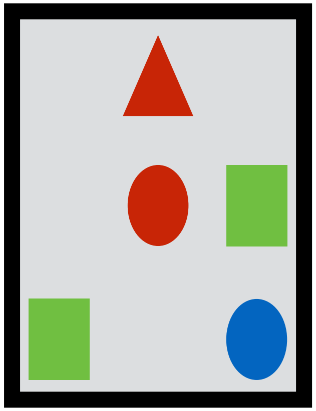
Slide credit: Jacob Andreas

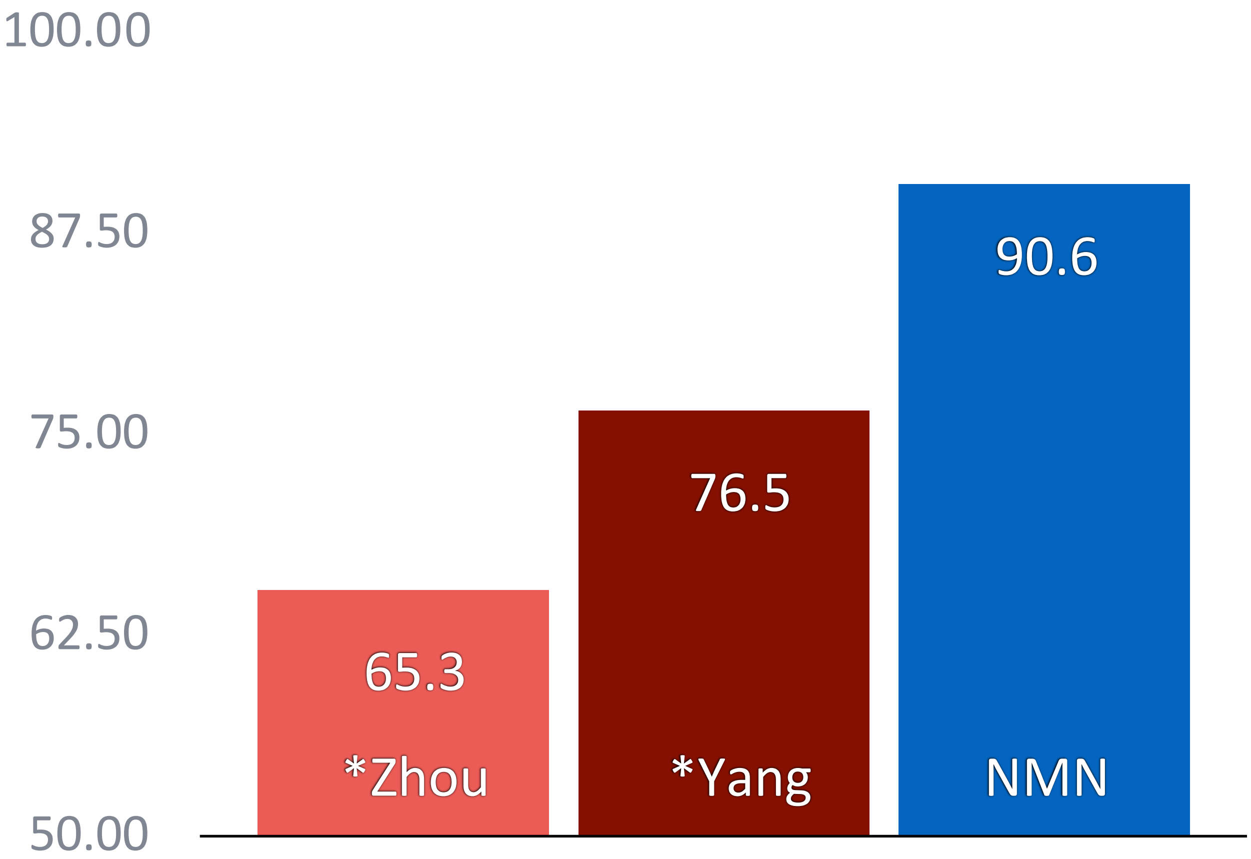
Experiments: VQA dataset



Slide credit: Jacob Andreas

Experiments: SHAPES dataset





Slide credit: Jacob Andreas

Experiments: VQA dataset

What color is she wearing?

color

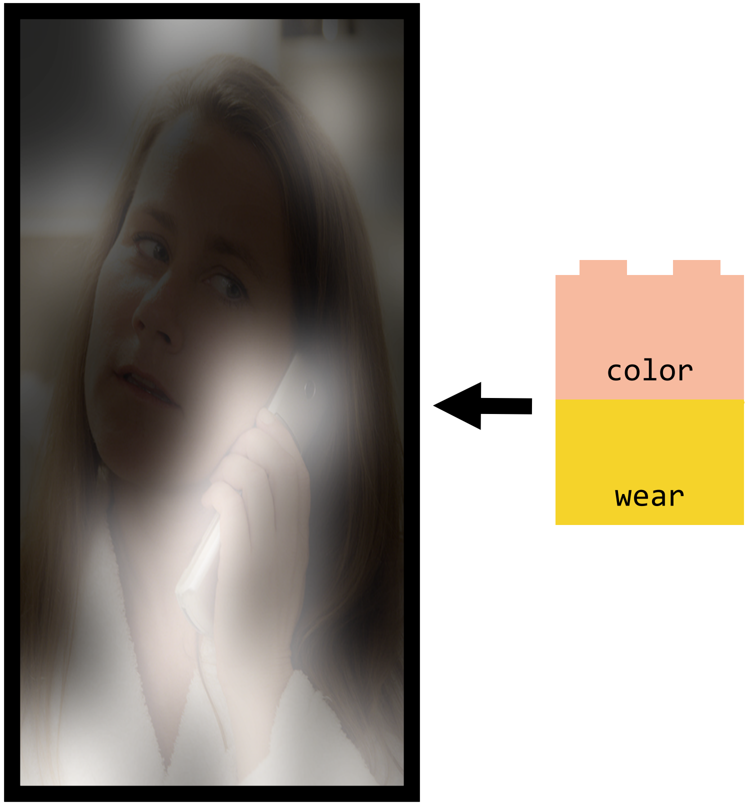
wear

white

Slide credit: Jacob Andreas

Experiments: VQA Dataset

What color is she wearing?

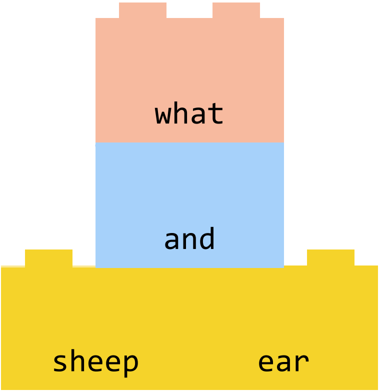


white

Slide credit: Jacob Andreas

Experiments: VQA Dataset

What is in the sheep's ear?



tag

Slide credit: Jacob Andreas

Experiments: VQA Dataset

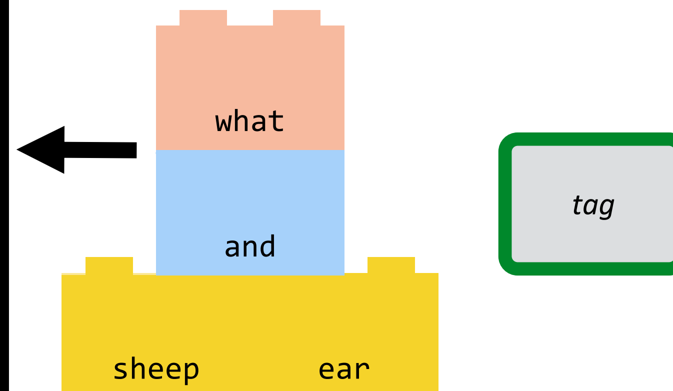
What
sheep

tag

Slide credit: Jacob Andreas

Experiments: VQA Dataset

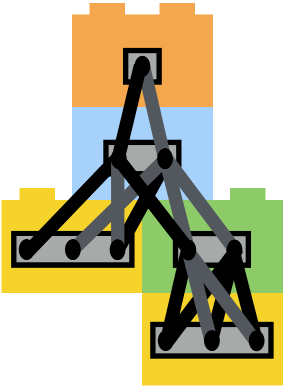
What is in the sheep's ear?



Slide credit: Jacob Andreas

Neural module networks

Linguistic structure dynamically generates model structure



Combines advantages of:

- Representation learning (like a neural net)
- Compositionality (like a semantic parser)

Slide credit: Jacob Andreas

Datasets

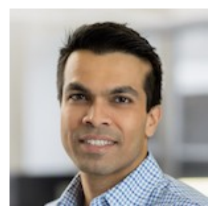
Models

Current Status

Ongoing Efforts

Slide credit: Devi Parikh

Visual Question Answering Challenge 2020



Ayush Shrivastava
(Georgia Tech)

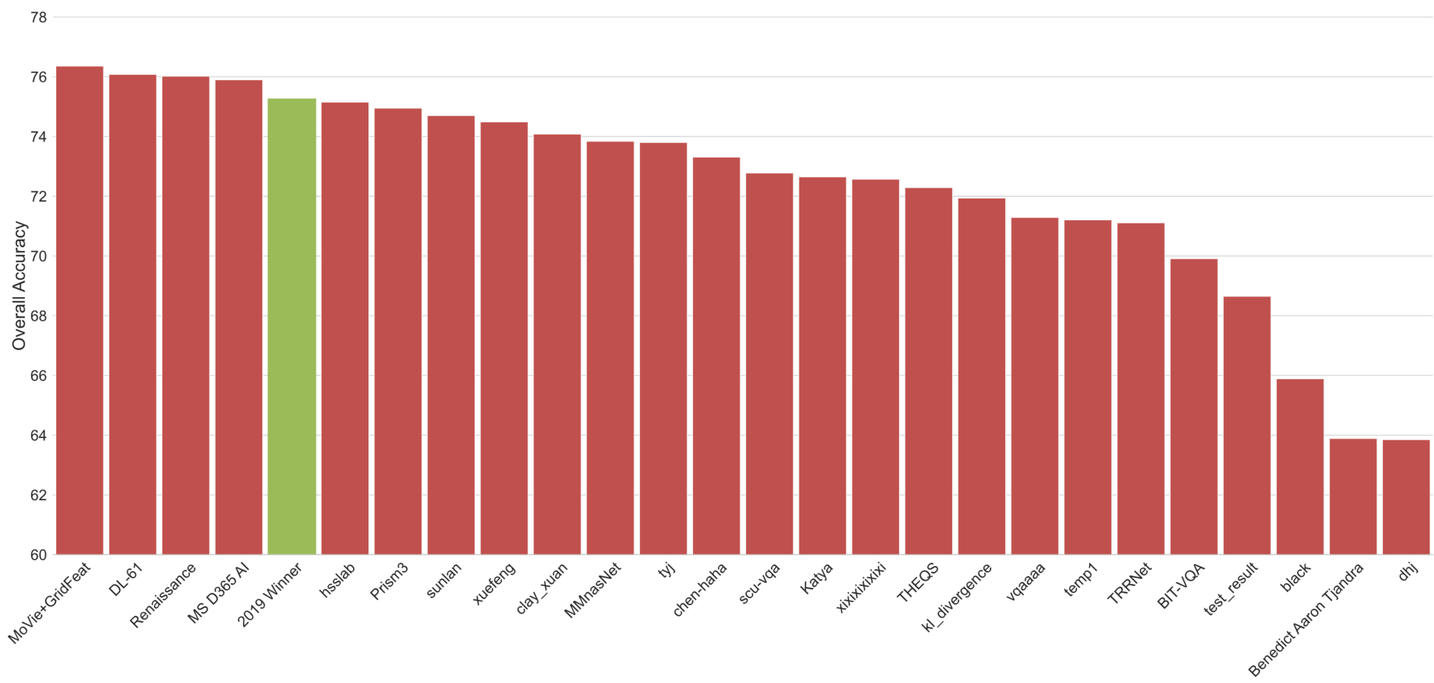
Yash Goyal
(Georgia Tech → SAIL Montreal)

Dhruv Batra
(Georgia Tech / FAIR

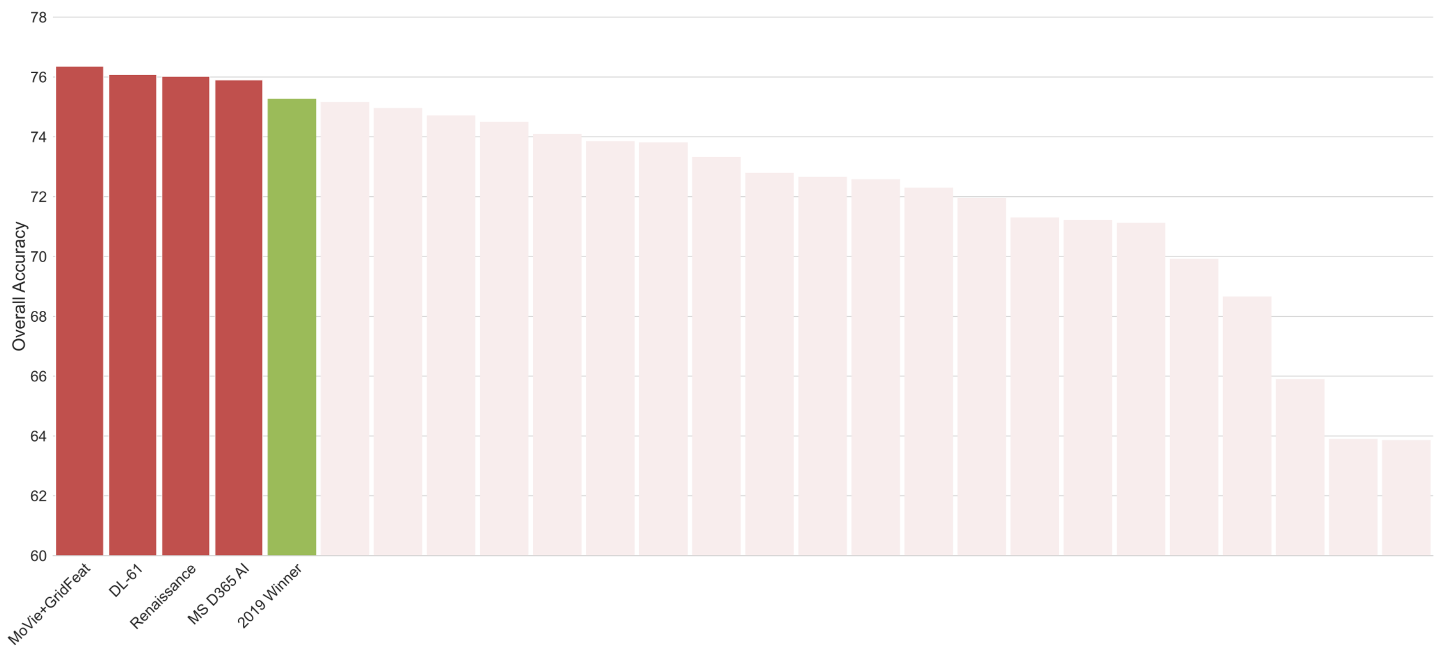
Devi Parikh
(Georgia Tech / FAIR)

Aishwarya Agrawal
(DeepMind)

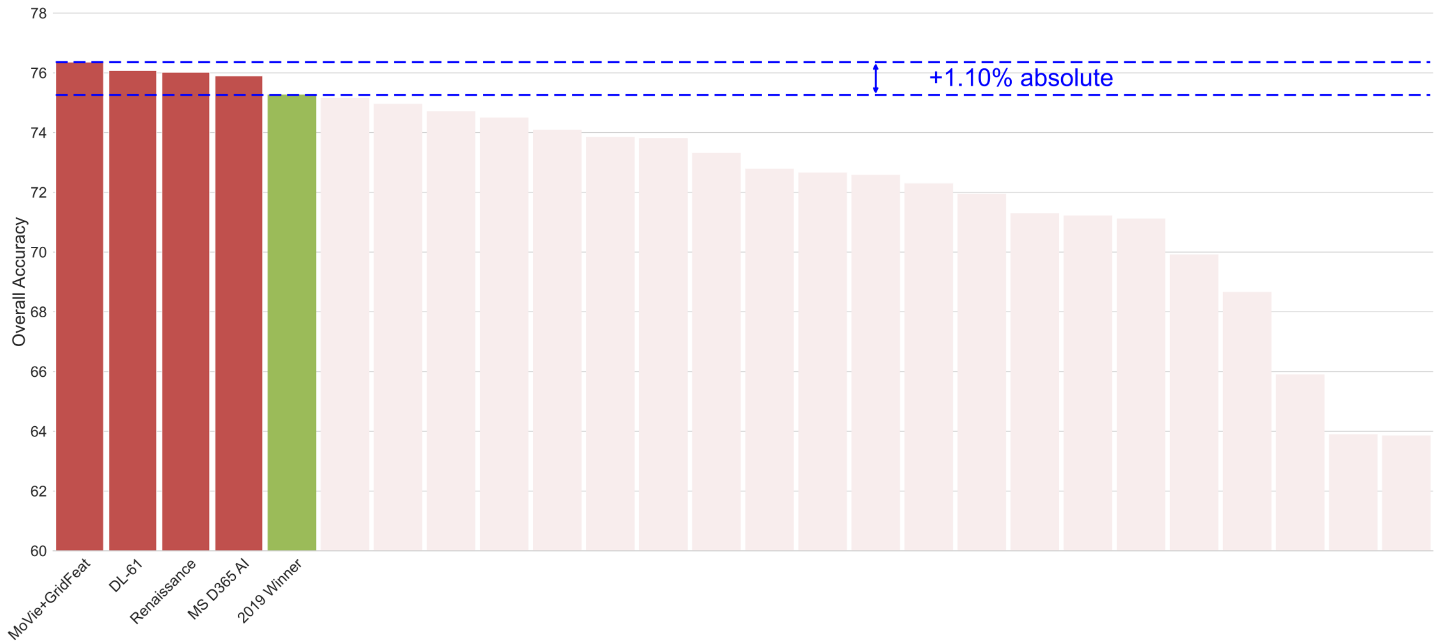
Challenge Results



Challenge Results



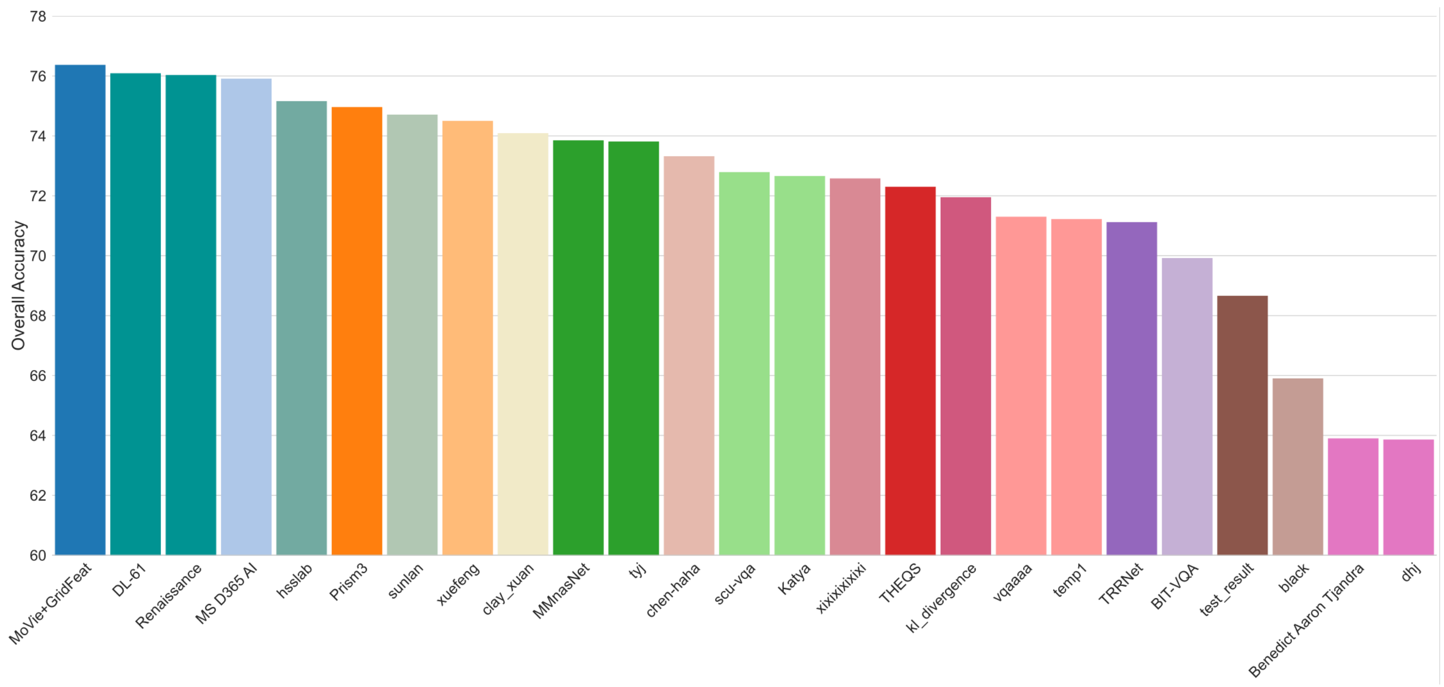
Challenge Results



Statistical Significance

- Performed Wilcoxon signed-rank test
- @ 95% confidence

Statistical Significance

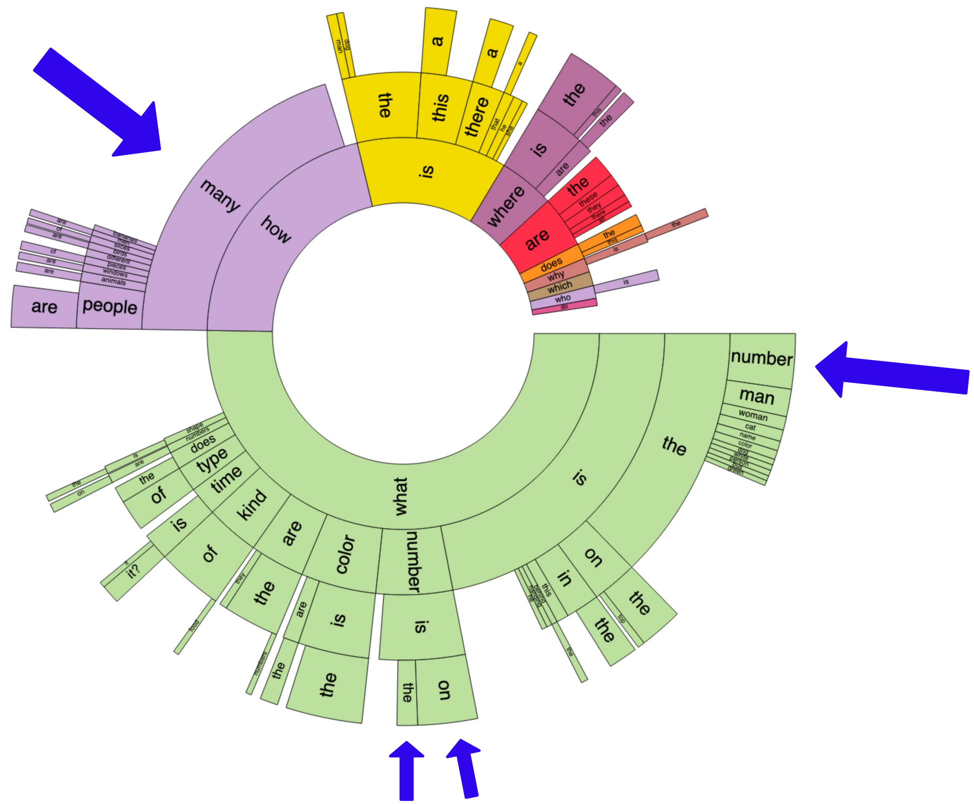


Easy and Difficult Questions

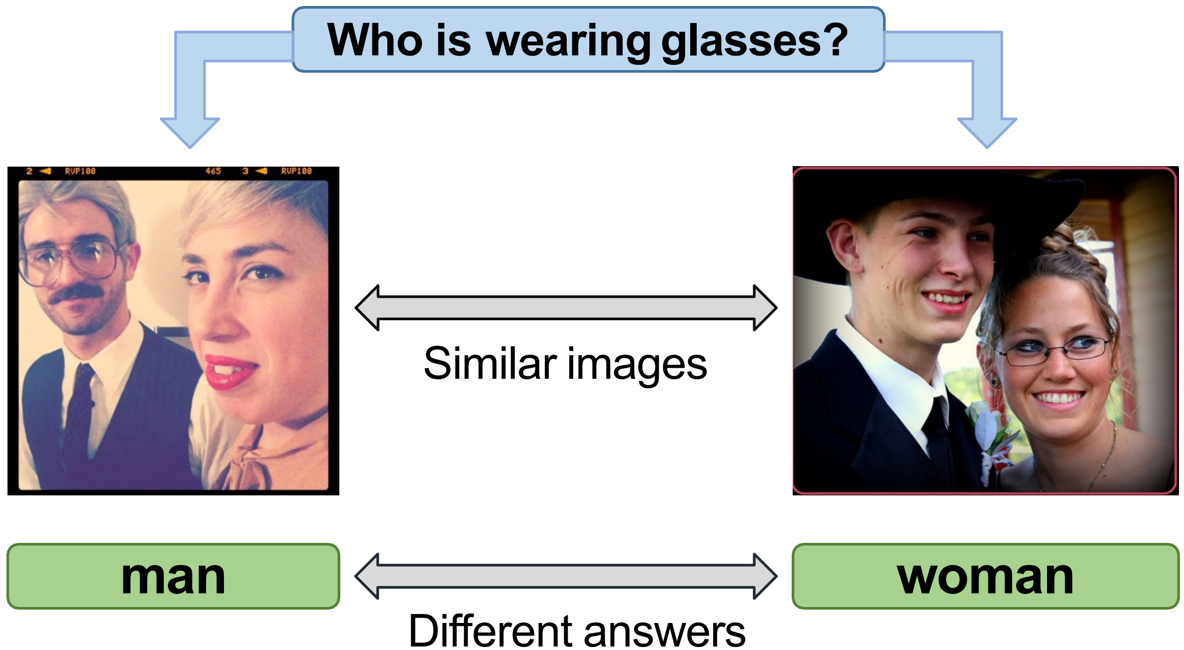
61% question answered by all top-10 teams → Easy questions

12.8% question not answered by any of top-10 teams → Difficult questions

Difficult Questions in 2019 (not in 2020)



Are models sensitive to subtle changes in images?

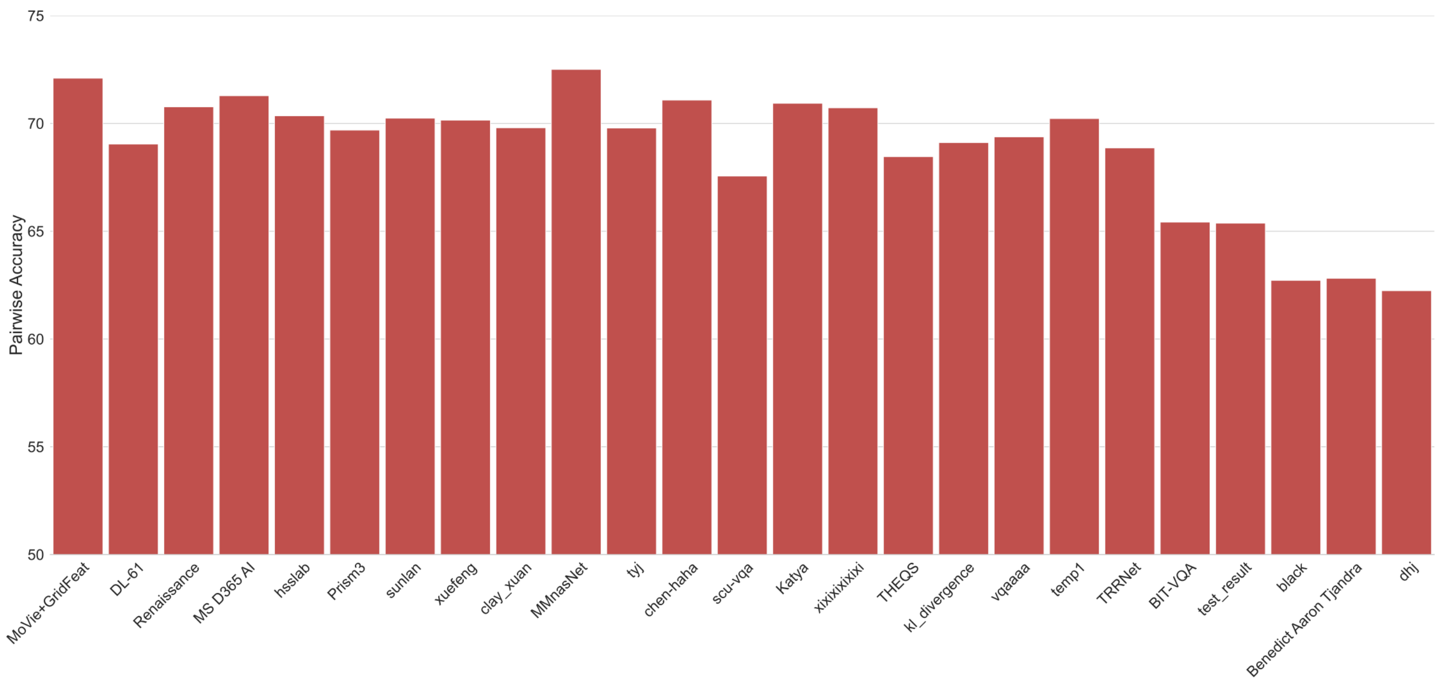


Are models sensitive to subtle changes in images?

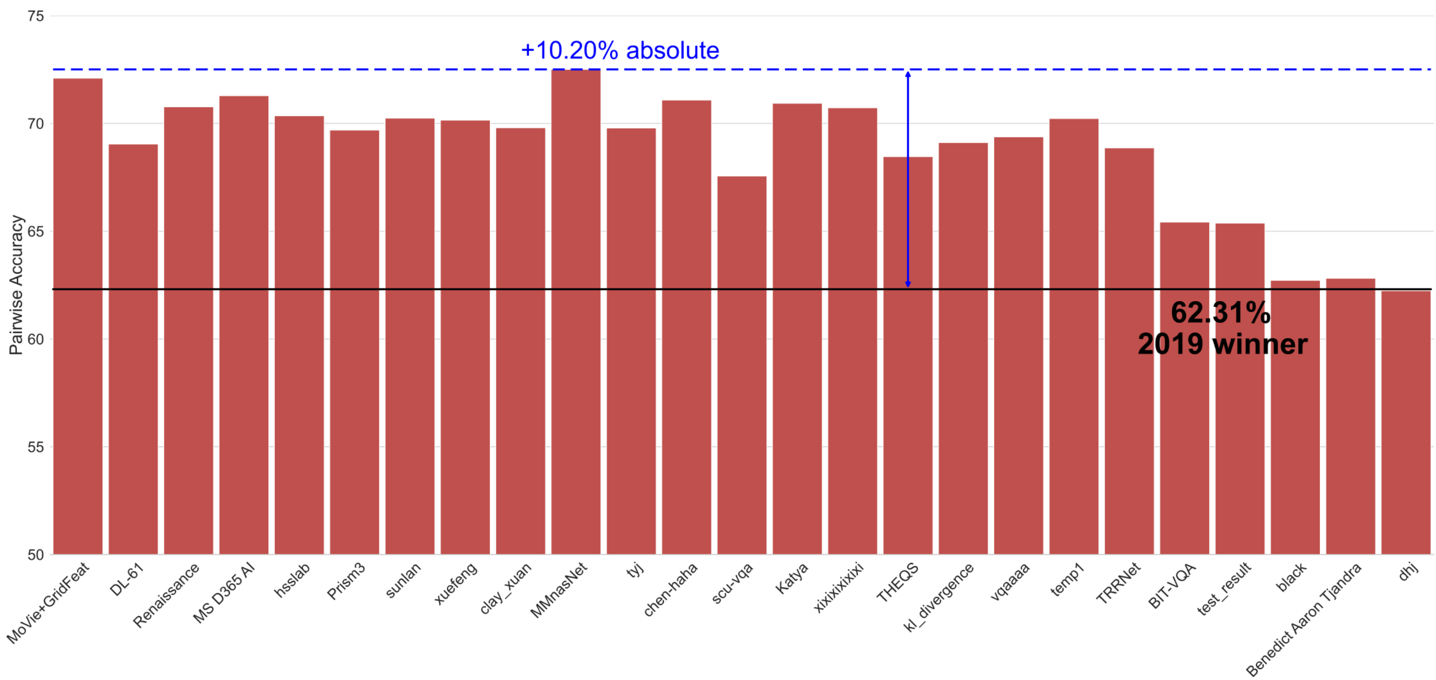
- Are predictions accurate for complementary images?

- Accuracy computed for each complementary pair:
  - 1 point: Predict correct answers for both images
  - 0 point, otherwise

Are predictions **accurate** for complementary images?



Are predictions **accurate** for complementary images?



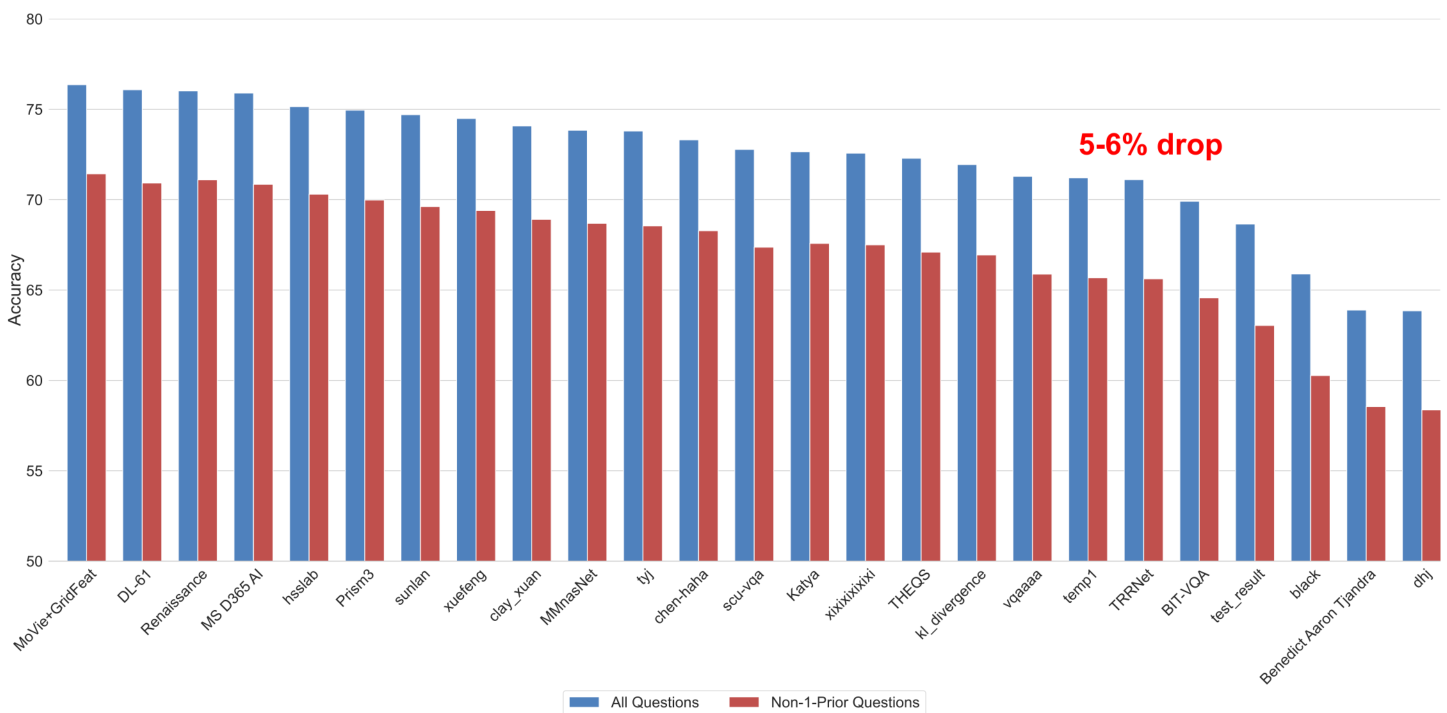
Are models driven by priors?

Non-1-Prior:

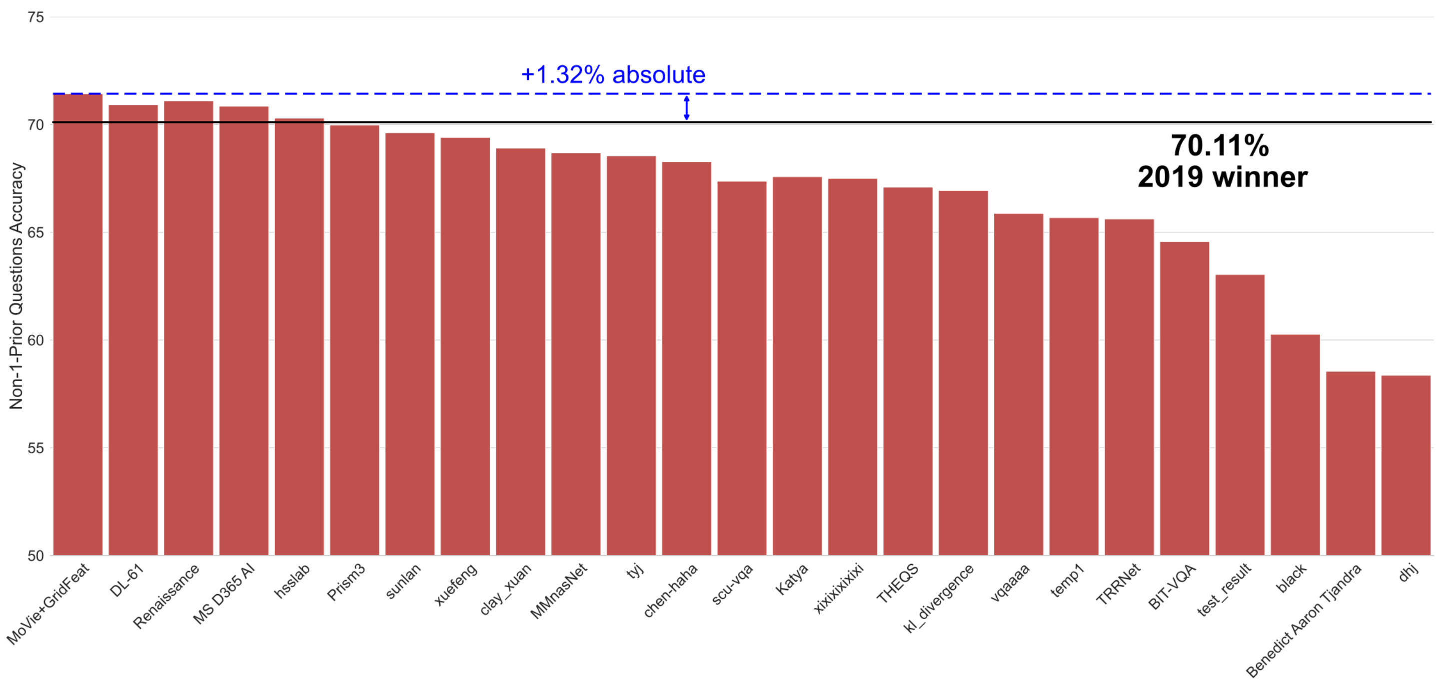
- Questions whose answers are not top-1 most common for the given question n-gram in training.
- Consists of 73% of all test-challenge questions.

Agrawal et al., CVPR 2018

Are models driven by priors?



Are models driven by priors?



Are models compositional?

Only consider those questions which are compositionally novel:

- QA pair is not seen in training
- Constituting concepts seen in training

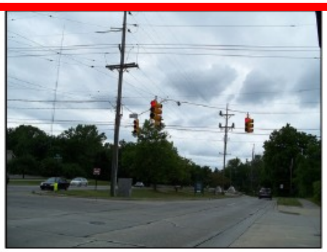
Agrawal et al., Arxiv 2018

Are models compositional?

Training

Q: What color is the plate?

A: Green



Q: What color are stop lights?

A: Red

Testing

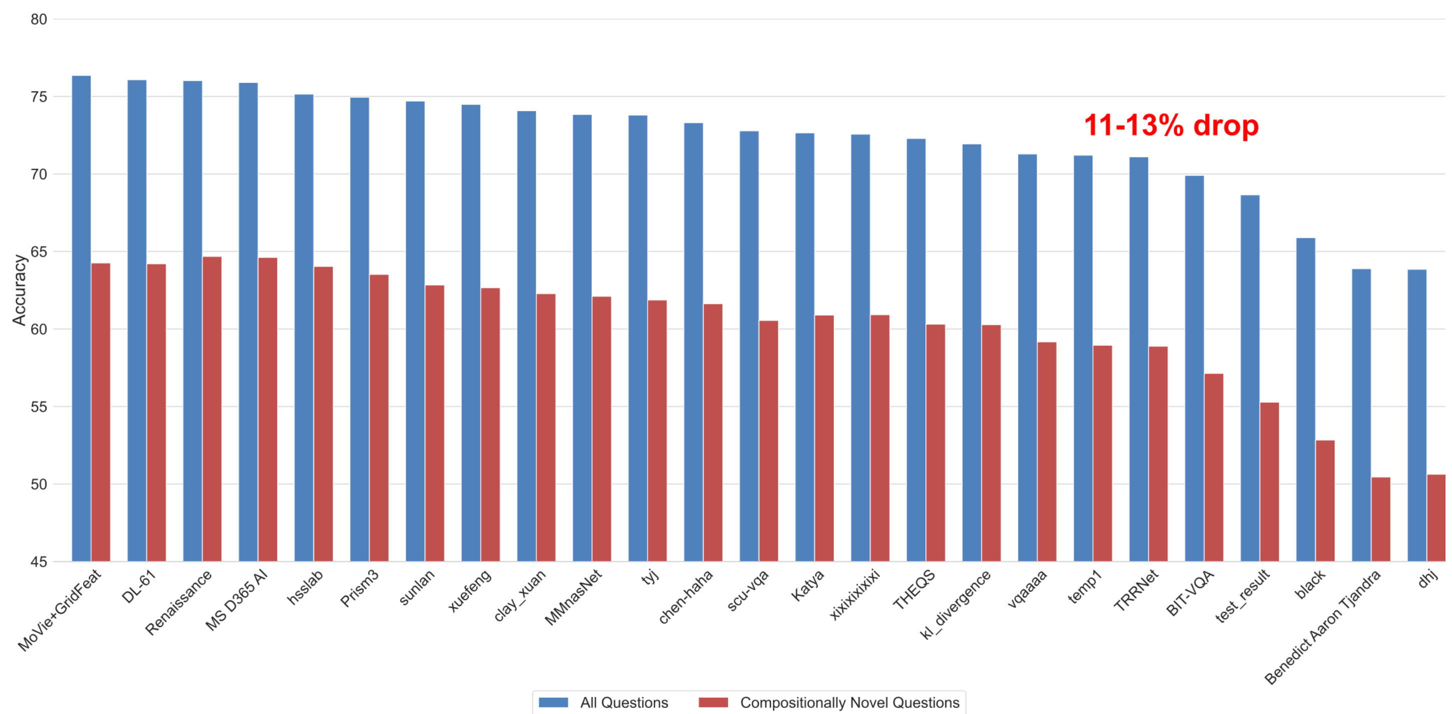
Q: What color is the stop light?

A: Green

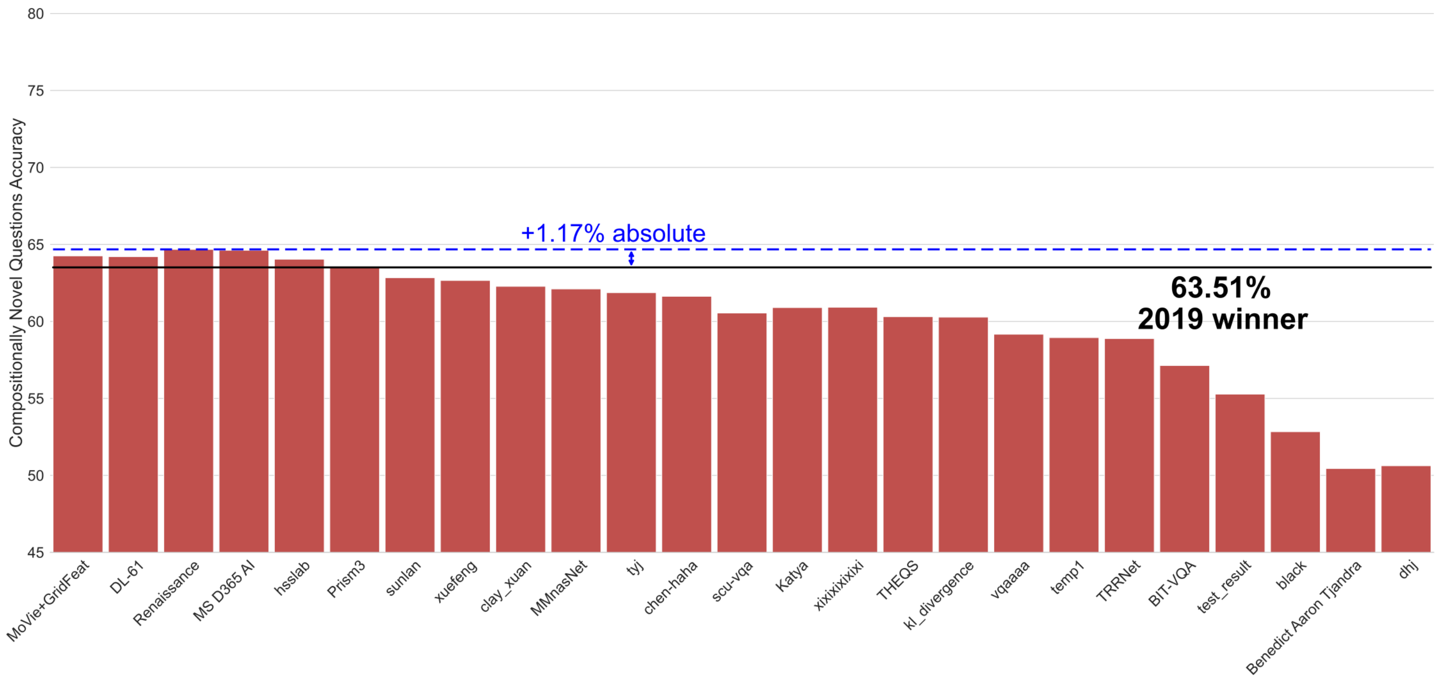
Q: What is the color of the plate?

A: Red

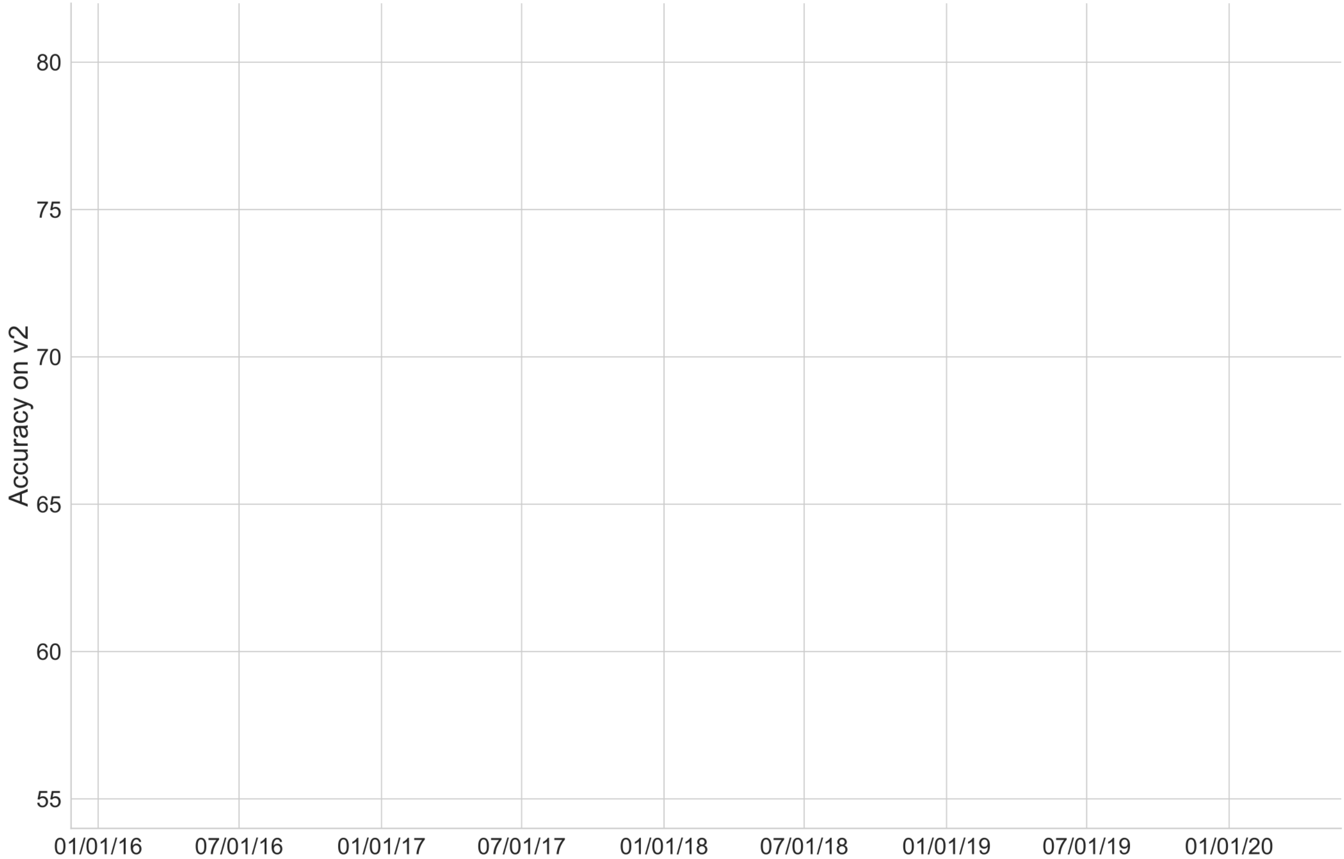
Are models compositional?



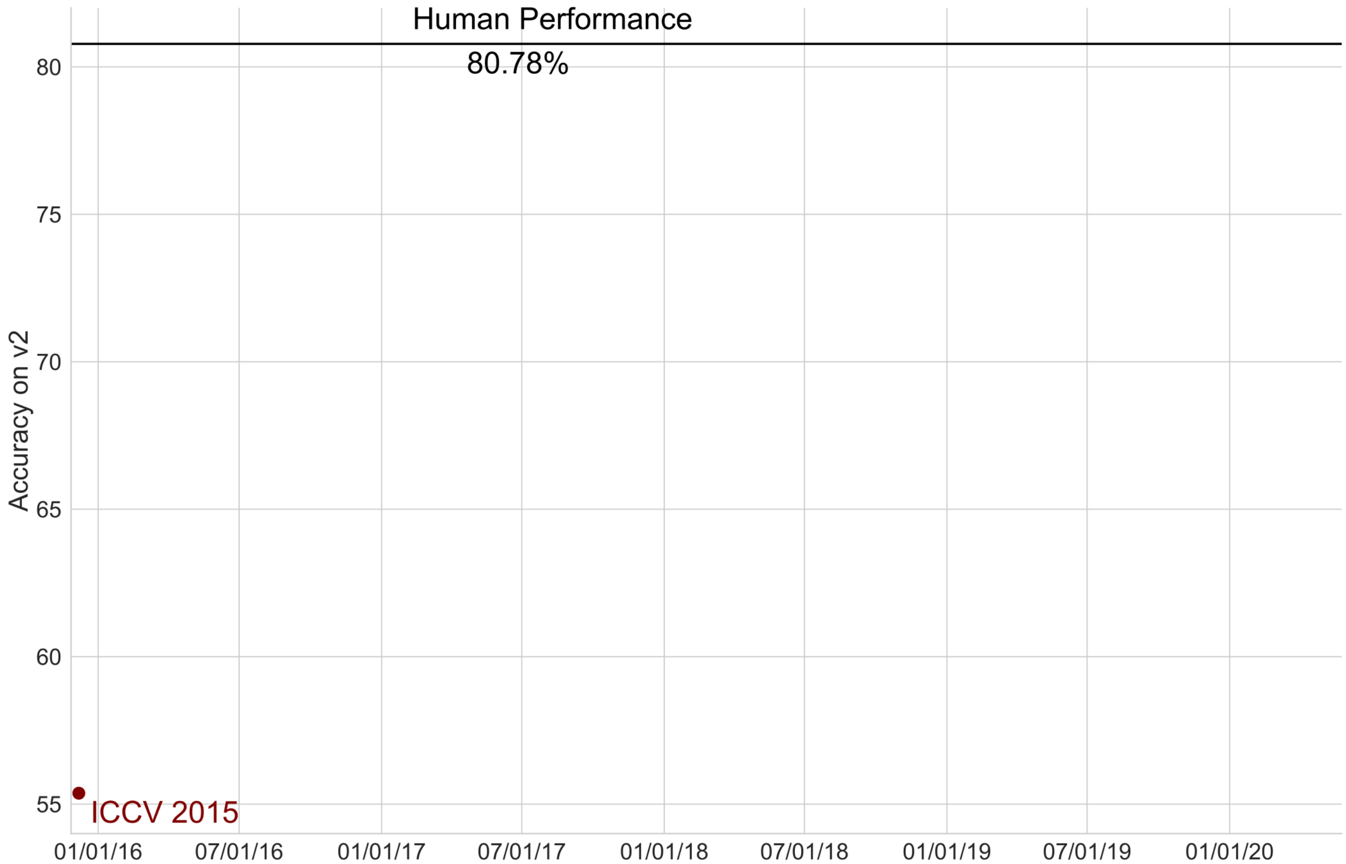
Are models compositional?



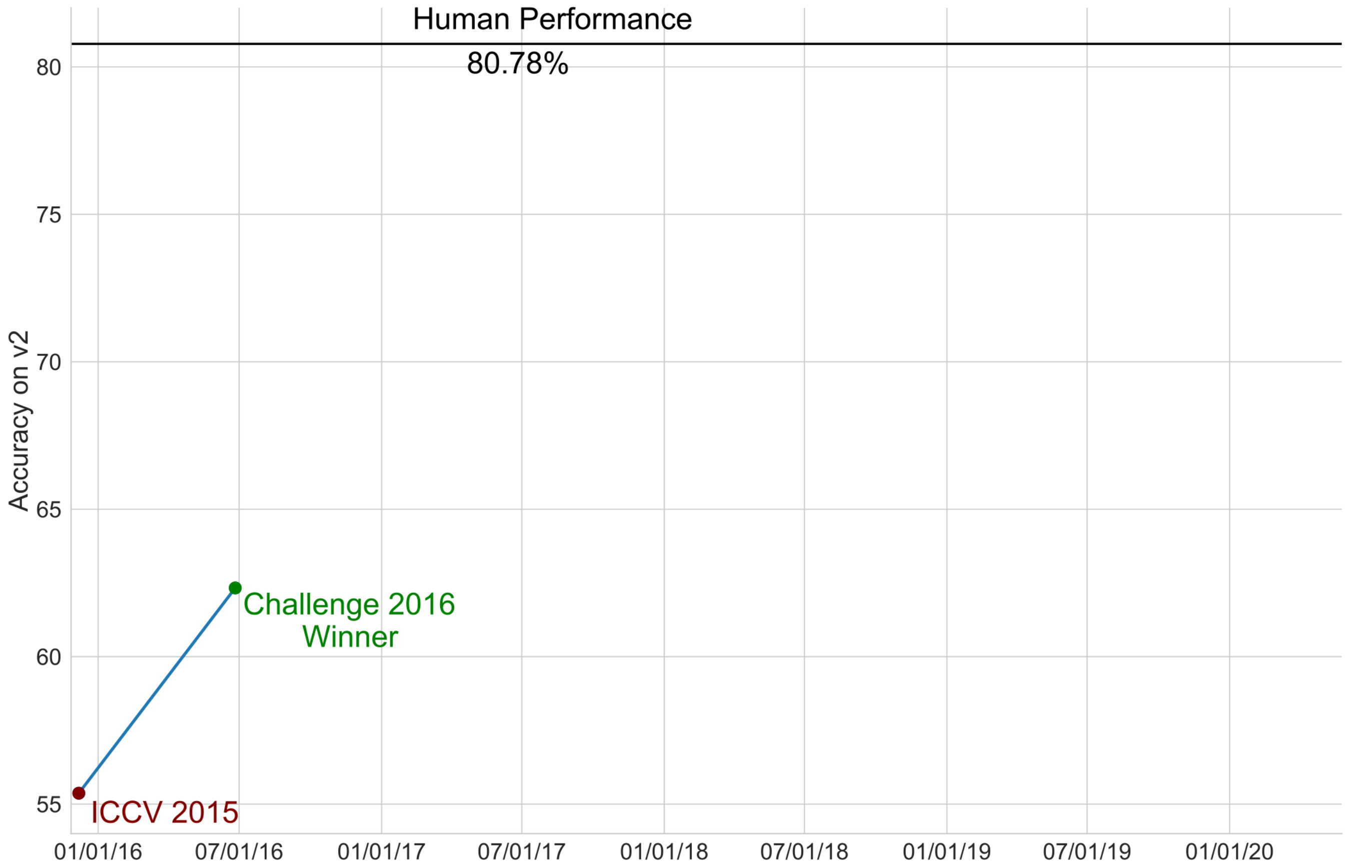
Progress in VQA



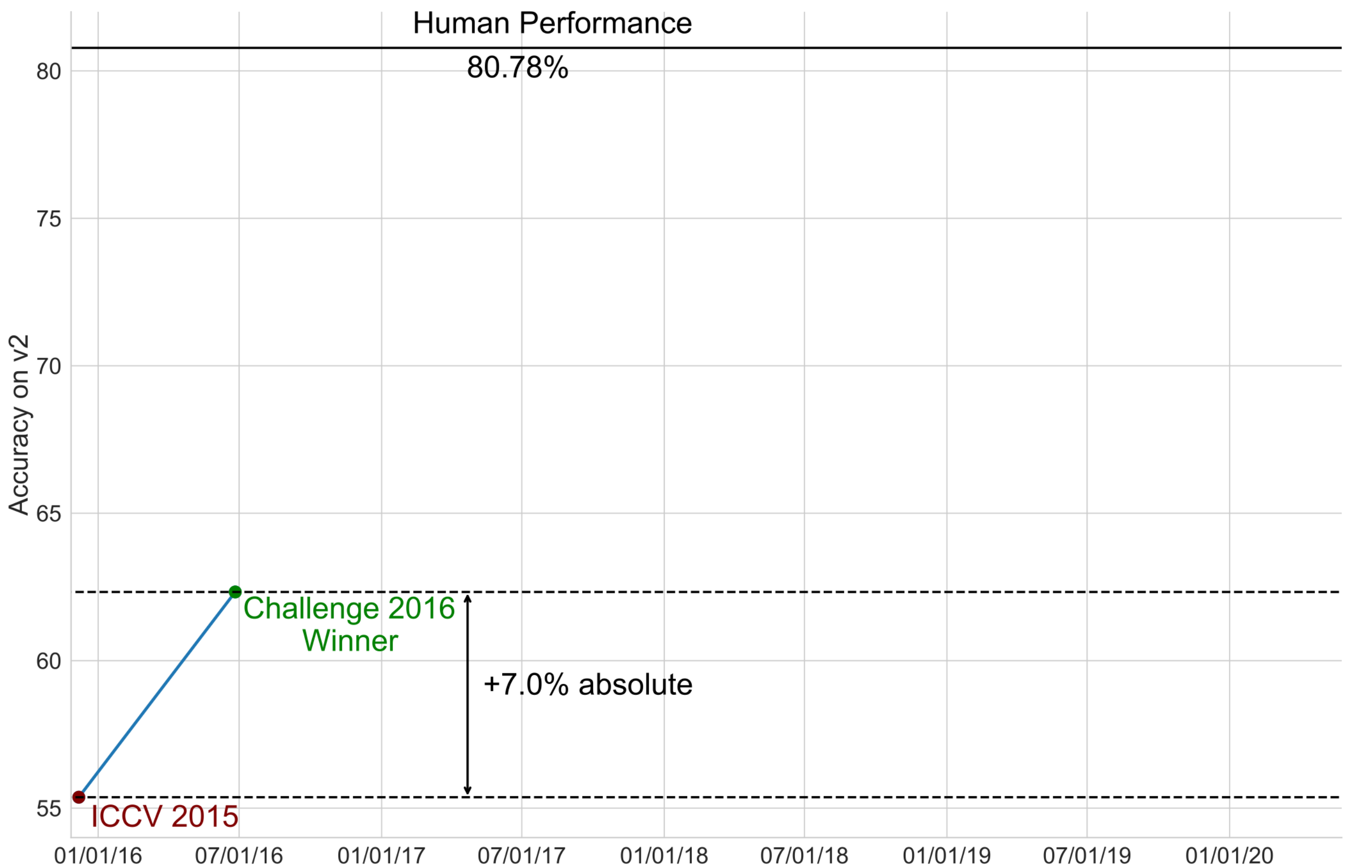
Progress in VQA



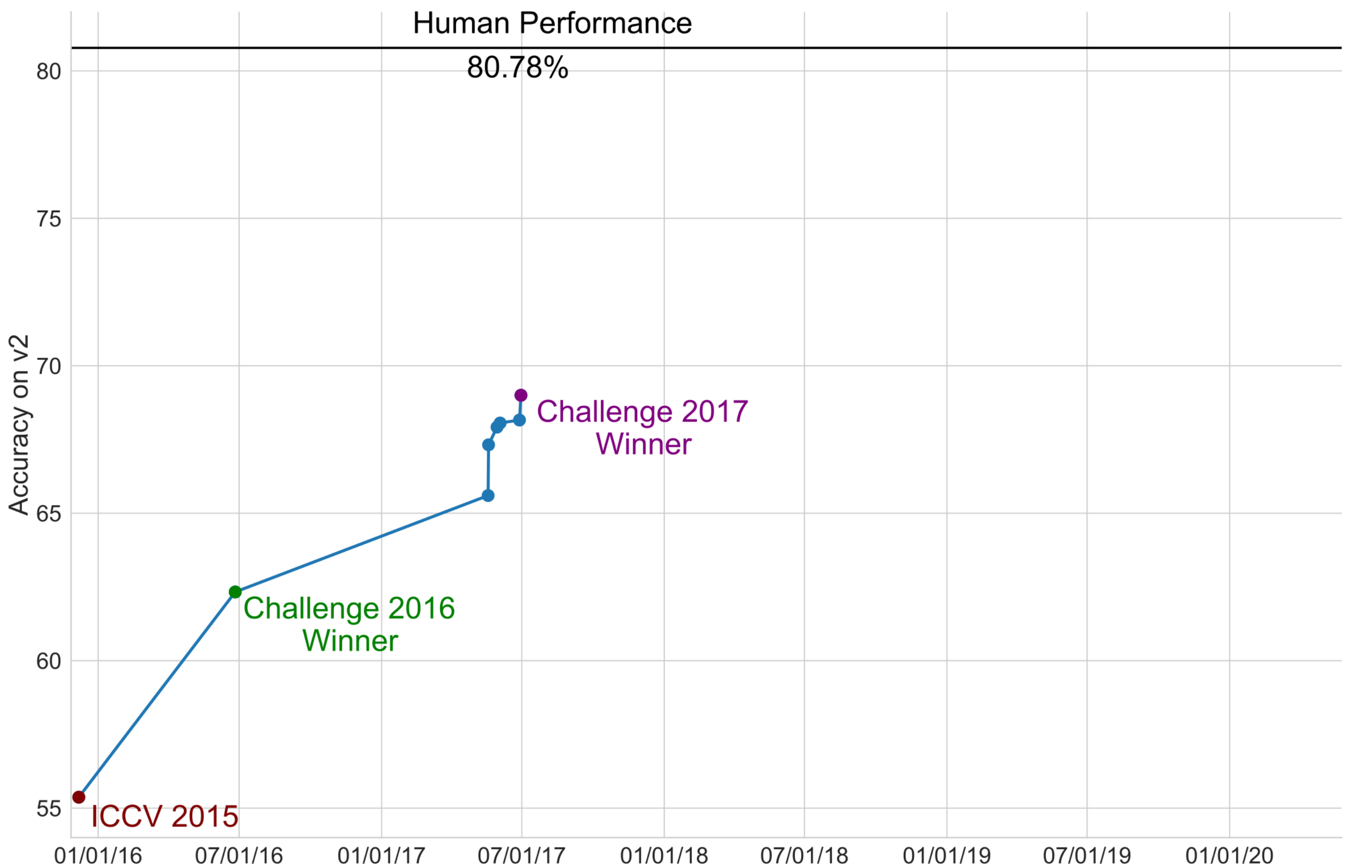
Progress in VQA



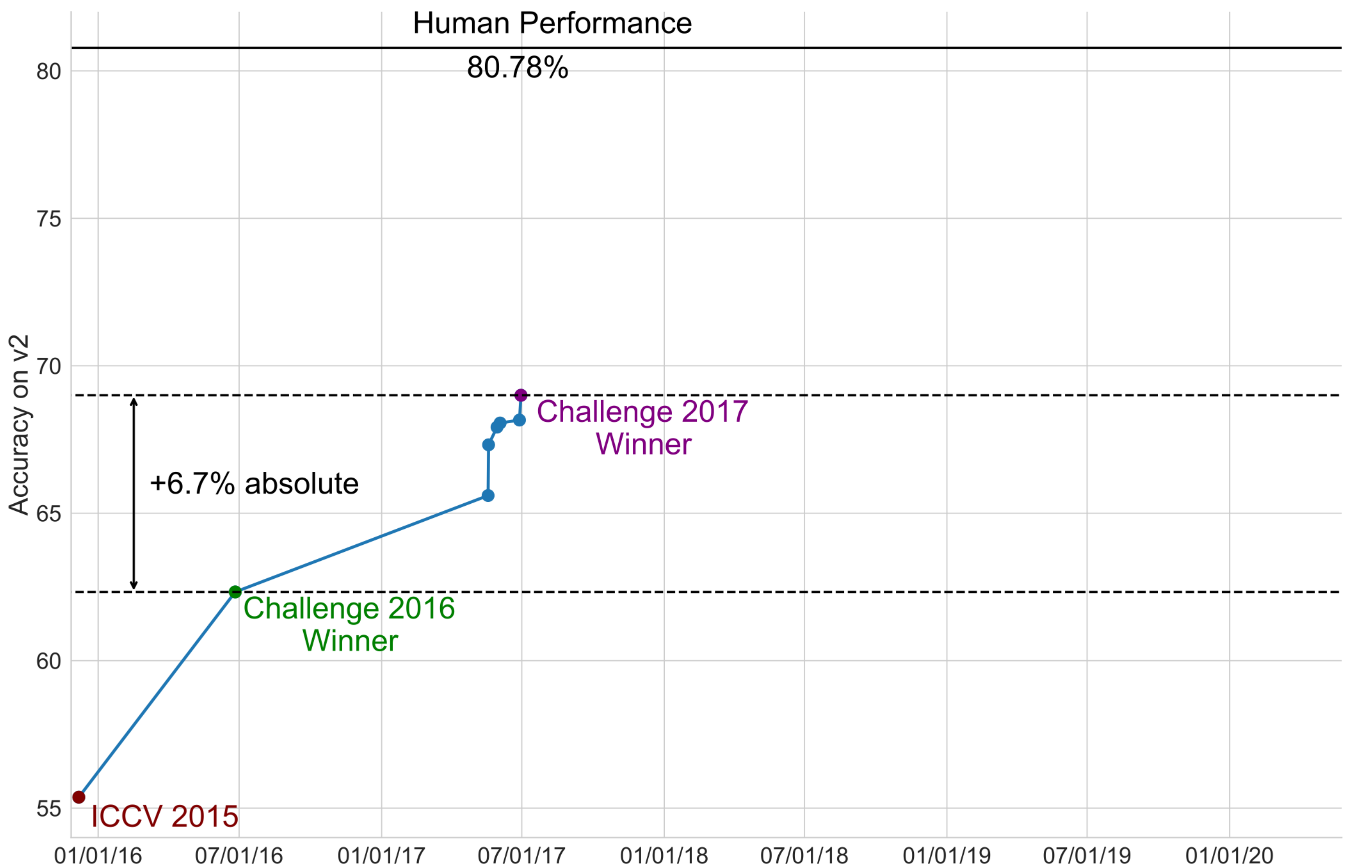
Progress in VQA



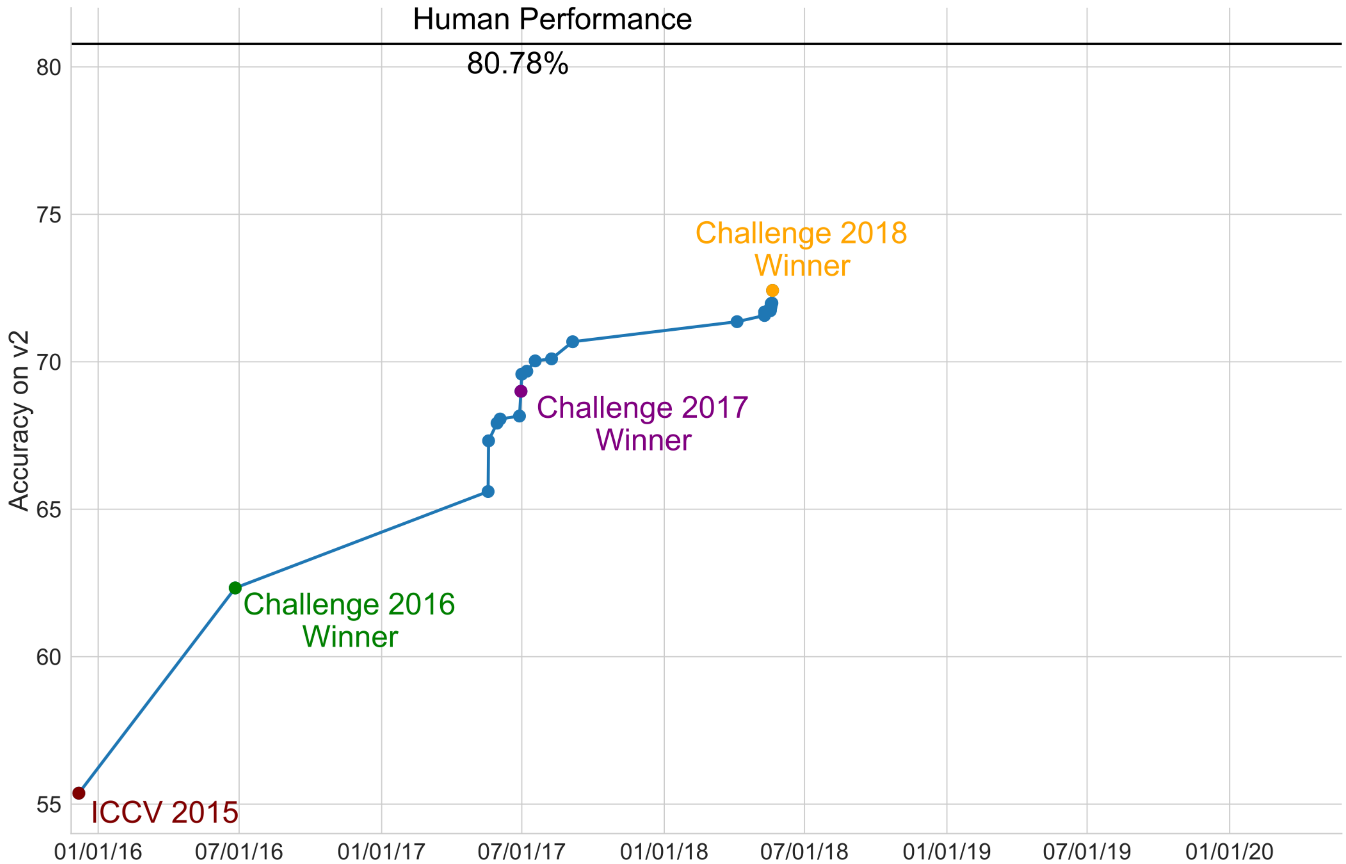
Progress in VQA



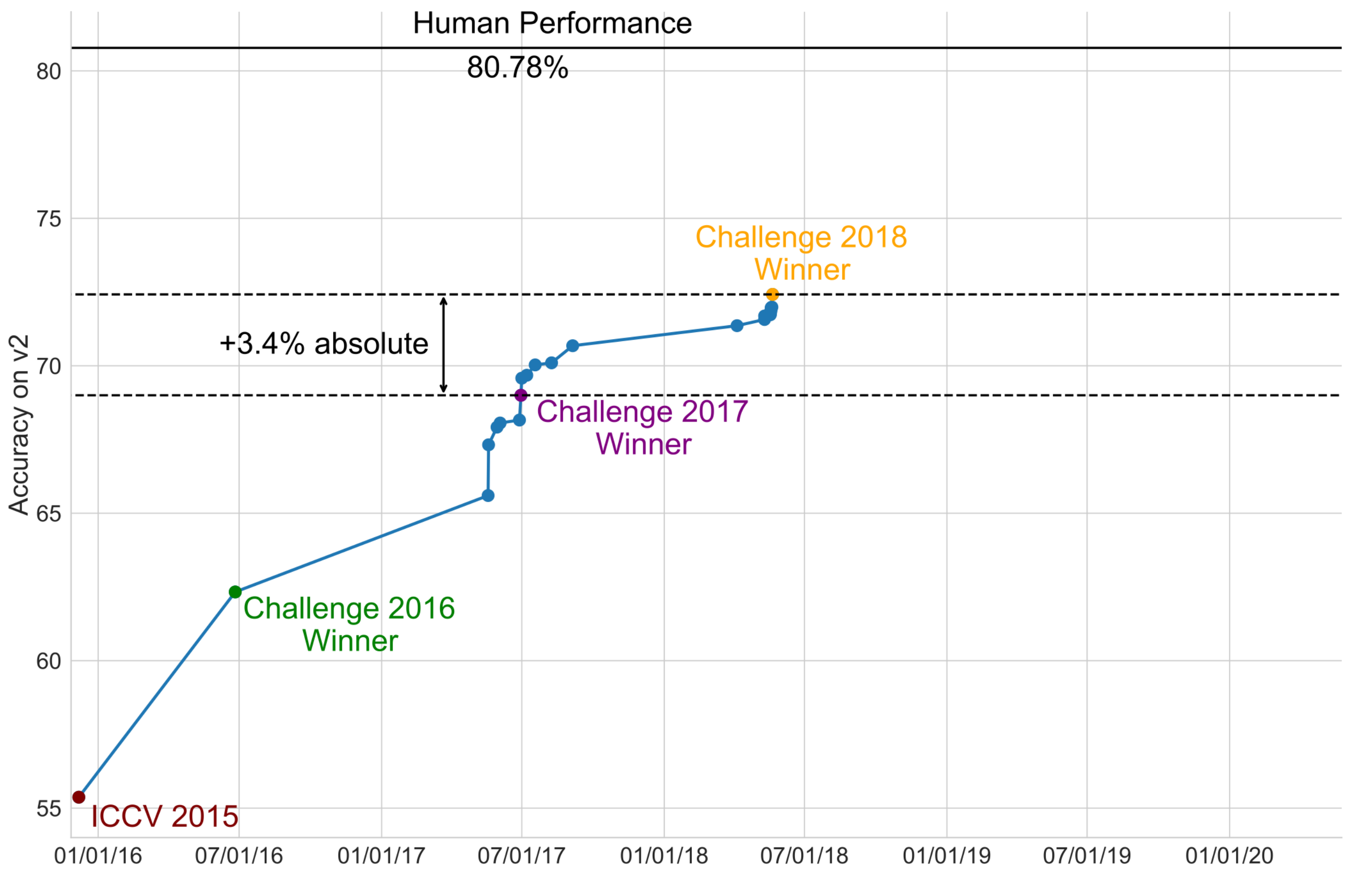
Progress in VQA



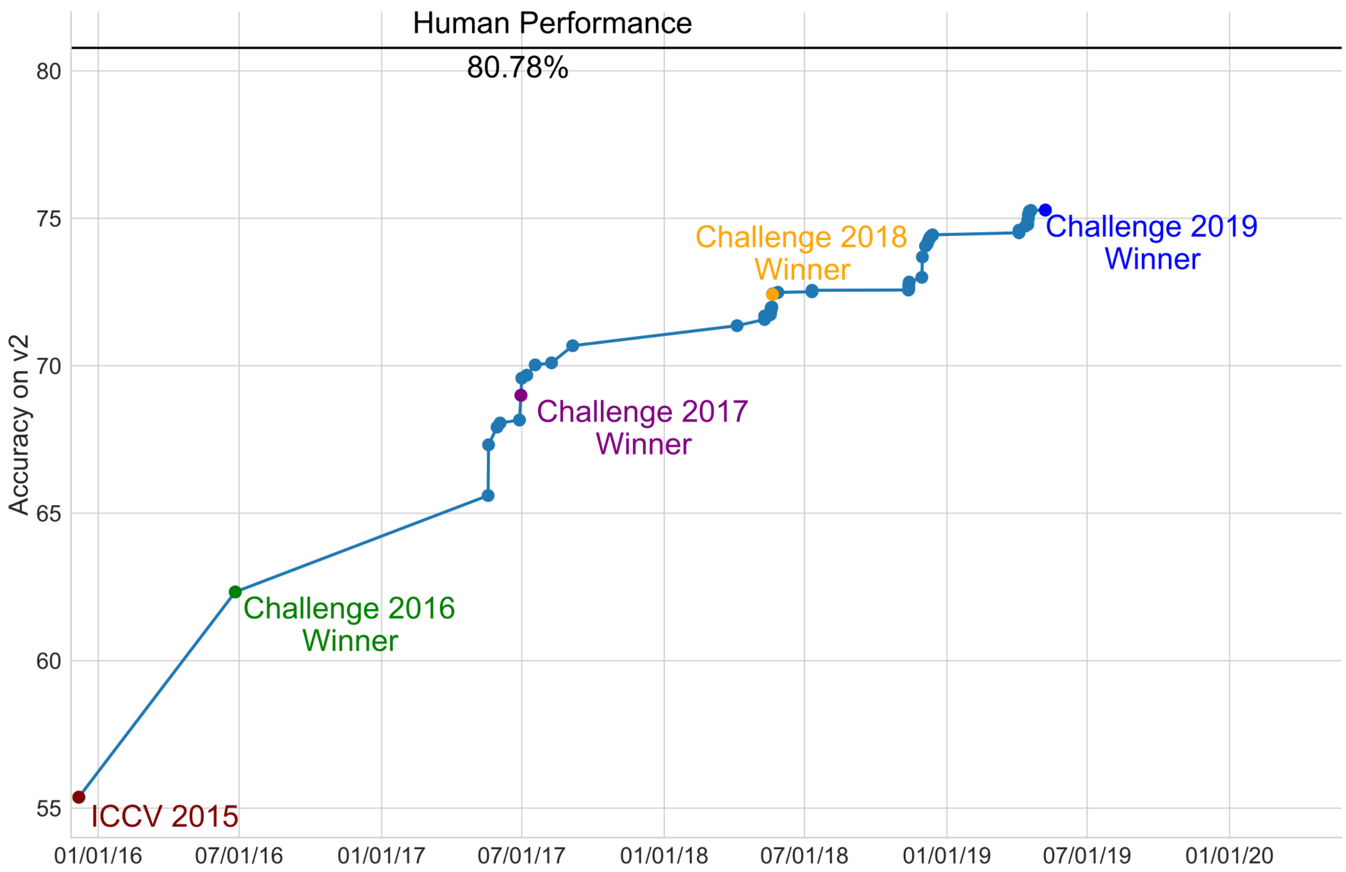
Progress in VQA



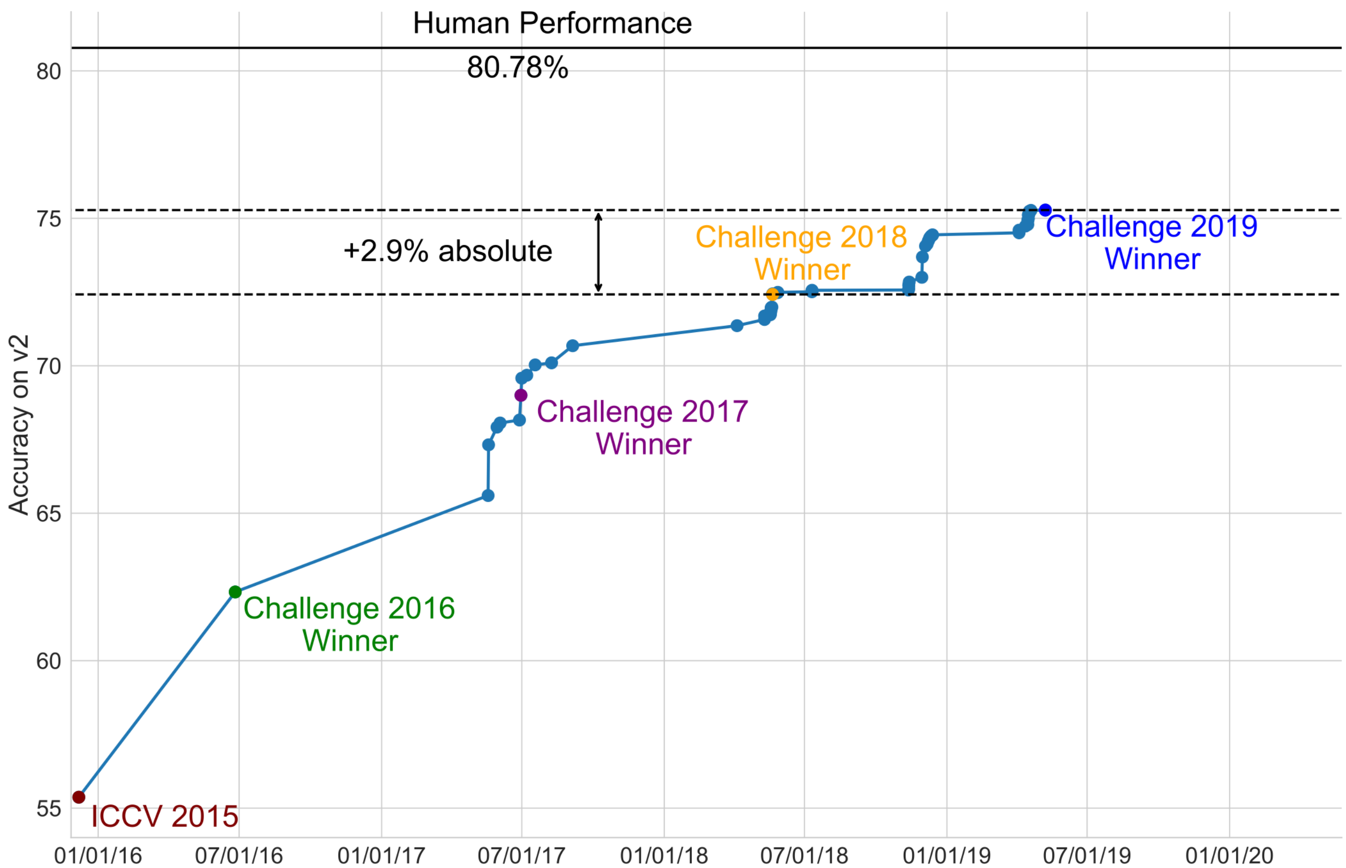
Progress in VQA



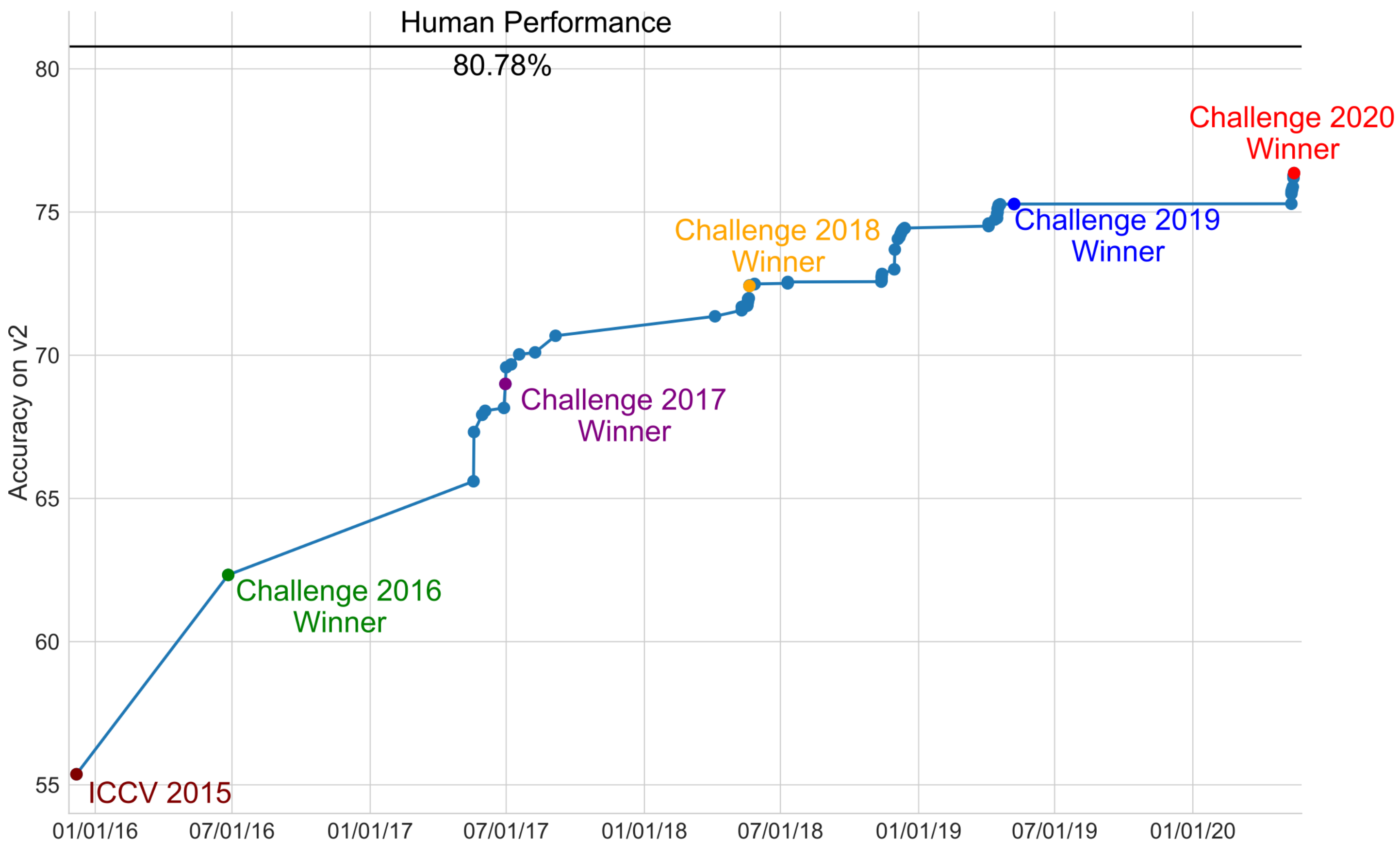
Progress in VQA



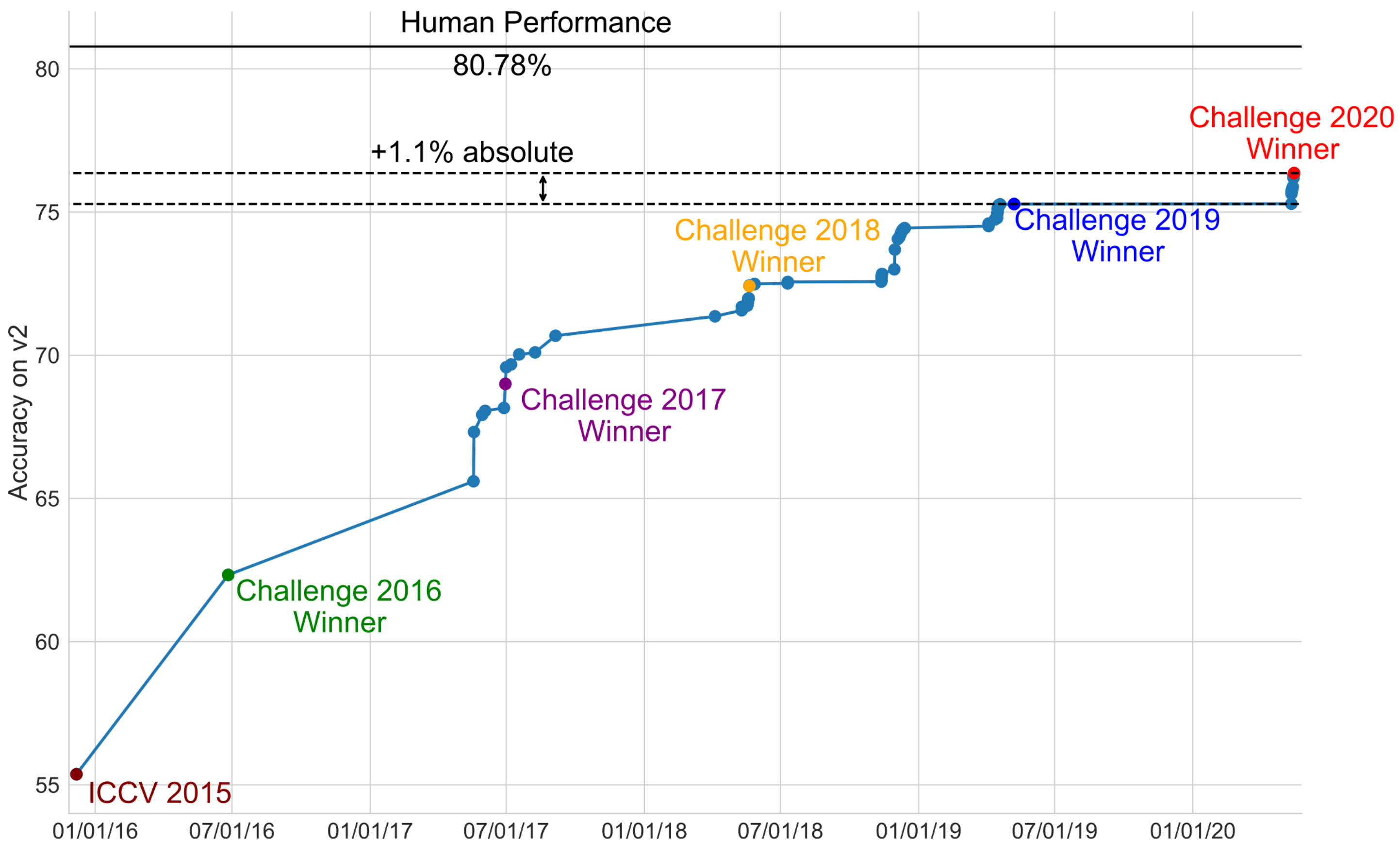
Progress in VQA



Progress in VQA



Progress in VQA



Future Directions

- Very large language models as common-sense priors
- Recent GPT-3 model from OpenAI
  - https://arxiv.org/pdf/2005.14165.pdf
  - 175B parameters, ~0.5T words

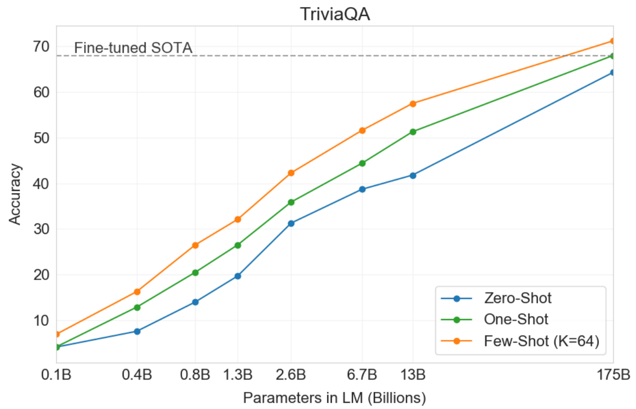


Figure 3.3: On TriviaQA GPT3's performance grows smoothly with model size, suggesting that language models continue to absorb knowledge as their capacity increases. One-shot and few-shot performance make significant gains over zero-shot behavior, matching and exceeding the performance of the SOTA fine-tuned open-domain model, RAG [LPP+20]

GPT-3 Generation

```
Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church
has agreed to a historic split - one that is expected to end in the
creation of a new denomination, one that will be "theologically and
socially conservative," according to The Washington Post. The majority of
delegates attending the church's annual General Conference in May voted to
strengthen a ban on the ordination of LGBTQ clergy and to write new rules
that will "discipline" clergy who officiate at same-sex weddings. But
those who opposed these measures have a new plan: They say they will form a
separate denomination by 2020, calling their church the Christian Methodist
denomination.
The Post notes that the denomination, which claims 12.5 million members, was
in the early 20th century the "largest Protestant denomination in the U.S.,"
but that it has been shrinking in recent decades. The new split will be the
second in the church's history. The first occurred in 1968, when roughly
10 percent of the denomination left to form the Evangelical United Brethren
Church. The Post notes that the proposed split "comes at a critical time
for the church, which has been losing members for years," which has been
"pushed toward the brink of a schism over the role of LGBTQ people in the
church." Gay marriage is not the only issue that has divided the church. In
2016, the denomination was split over ordination of transgender clergy, with
the North Pacific regional conference voting to ban them from serving as
clergy, and the South Pacific regional conference voting to allow them.
```