Forcing Vision and Language Models to Not Just Talk But Also Actually See

Devi Parikh

Georgia Tech

facebook Artificial Intelligence Research

Pythia for Vizwiz


Vivek Natarajan*


Tina Jiang*


Meet Shah


Xinlei Chen


Dhruv Batra


Devi Parikh


Marcus Rohrbach

* - indicates equal contribution

facebook Artificial Intelligence Research

facebook
Artificial Intelligence Research | A-STAR

Slide credit: Devi Parikh

Pythia

Pythia is a modular framework for Visual Question Answering research, which formed the basis for the winning entry to the VQA Challenge 2018 from Facebook AI Research (FAIR)'s A-STAR team. Please check our paper for more details.

(A-STAR: Agents that See, Talk, Act, and Reason.)

Q: What is the cat wearing?
A: Hat

Q: What is the weather like?
A: Rainy

Q: What surface is this?
A: Clay

Q: What is the weather like?
A: Sunny

Q: What color is the cat's eyes?
A: Green

Q: What toppings are on the pizza?
A: Mushrooms

Slide credit: Devi Parikh

VizWiz Challenge Entry: 10:50 am to 11:20 am

Pythia for Vizwiz


Vivek Natarajan*


Tina Jiang*


Meet Shah


Xinlei Chen


Dhruv Batra


Devi Parikh


Marcus Rohrbach

* - indicates equal contribution

facebook Artificial Intelligence Research

facebook
Artificial Intelligence Research | A-STAR

Slide credit: Devi Parikh

Slide credit: Devi Parikh

"Color College Avenue", Blacksburg, VA, May 2012

People coloring a street in rural Virginia.

Slide credit: Devi Parikh

"Color College Avenue", Blacksburg, VA, May 2012

It was a great event! It brought families out, and the whole community together.

Slide credit: Devi Parikh

"Color College Avenue", Blacksburg, VA, May 2012

Q. What are they coloring the street with?

A. Chalk

Slide credit: Devi Parikh

"Color College Avenue", Blacksburg, VA, May 2012

AI: What a nice picture! What event was this?

User: *"Color College Avenue". It was a lot of fun!*

AI: I am sure it was! Do they do this every year?

User: *I wish they would. I don't think they've organized it again since 2012.*

…

Slide credit: Devi Parikh

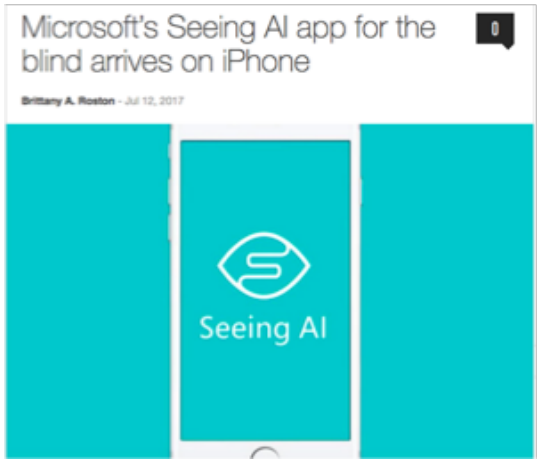
"Color College Avenue", Blacksburg, VA, May 2012

Applications

FACEBOOK'S AI CAN CAPTION PHOTOS FOR THE BLIND ON ITS OWN

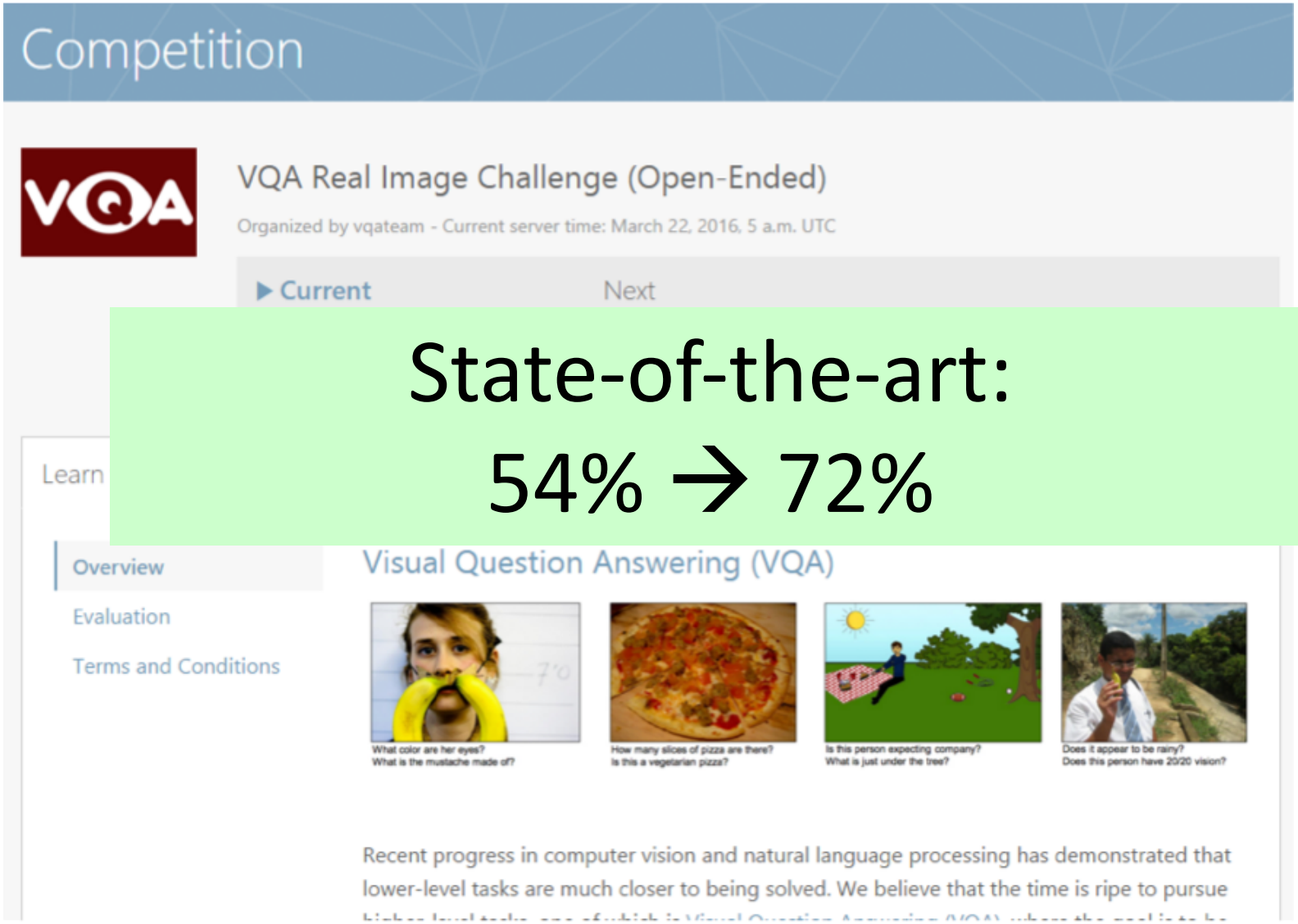
Microsoft's Seeing AI app for the blind arrives on iPhone

Brittany A. Roston - Jul 12, 2017



Slide credit: Devi Parikh

VQA Challenge @ CVPR16, 17, 18



Slide credit: Devi Parikh

Success Cases

Q: What is the woman holding?
GT A: laptop
Machine A: laptop

Q: What room is the cat located in?
GT A: kitchen
Machine A: kitchen

Q: Is it going to rain soon?
GT A: yes
Machine A: yes

Q: Is this a casino?
GT A: no
Machine A: no

Slide credit: Aishwarya Agrawal

Models affected by language priors

Slide credit: Devi Parikh

Test Sample

Q: What color are the safety cones?

Slide credit: Aishwarya Agrawal

Test Sample

Q: What color are the safety cones?

GT Ans: green

Predicted Ans: orange

Nearest Neighbor Training Samples

Slide credit: Aishwarya Agrawal

Test Sample

Q: What color are the safety cones?

GT Ans: green

Predicted Ans: orange

Nearest Neighbor Training Samples

Q: What color are the cones?

GT Ans: orange

Slide credit: Aishwarya Agrawal

Test Sample

Q: What color are the safety cones?

GT Ans: green

Predicted Ans: orange

Nearest Neighbor Training Samples

Q: What color are the cones?

GT Ans: orange

Q: What color is the cone?

GT Ans: orange

Slide credit: Aishwarya Agrawal

Test Sample

Q: What color are the safety cones?

GT Ans: green

Predicted Ans: orange

Nearest Neighbor Training Samples

Q: What color are the cones?

GT Ans: orange

Q: What color is the cone?

GT Ans: orange

Q: What color are the cones?

GT Ans: orange

Slide credit: Aishwarya Agrawal

Q: Are **A:** military

Q: Are they **A:** yes

Q: Are they playing **A:** yes

Q: Are they playing a **A:** yes

Q: Are they playing a game? **A:** yes

GT Ans: yes

Slide credit: Aishwarya Agrawal

Q: How **A:** no

Q: How many **A:** 2

Q: How many horses **A:** 2

Q: How many horses are **A:** 2

Q: How many horses are on **A:** 2

Q: How many horses are on the **A:** 2

Q: How many horses are on the beach? **A:** 2

GT Ans: 6

Slide credit: Aishwarya Agrawal

Q: Is **A:** kitchen

Q: Is the **A:** outside

Q: Is the bench **A:** no

Q: Is the bench made **A:** no

Q: Is the bench made of **A:** no

Q: Is the bench made of metal? **A:** no

GT Ans: yes

Slide credit: Aishwarya Agrawal

Q: What does the red sign say?

Predicted Ans: stop

Correct Response

Slide credit: Aishwarya Agrawal

Q: How many zebras?

Predicted Ans: 2

Correct Response

Slide credit: Aishwarya Agrawal

Q: What covers the ground?

Predicted Ans: snow

All Correct Responses

Slide credit: Aishwarya Agrawal

How do we force vision+language models to also *look*?

Slide credit: Devi Parikh

Outline

- Datasets
  - E.g., VQA v2.0
- Evaluation metrics
  - E.g., # pairs of images correctly answered
- Novel problem spaces
  - E.g., changing priors (VQA-CP), novel object captioning, robust captioning, visual coreference resolution
- Inductive biases in models
  - E.g., G-VQA, Neural Baby Talk
- Human-in-the-loop evaluation
  - E.g., accuracy is not the final metric, performance of human-AI team is

Slide credit: Devi Parikh

Outline

- Datasets
  - E.g., VQA v2.0
- Evaluation metrics
  - E.g., # pairs of images correctly answered
- Novel problem spaces
  - E.g., changing priors (VQA-CP), novel object captioning, robust captioning, visual coreference resolution
- Inductive biases in models
  - E.g., G-VQA, Neural Baby Talk
- Human-in-the-loop evaluation
  - E.g., accuracy is not the final metric, performance of human-AI team is

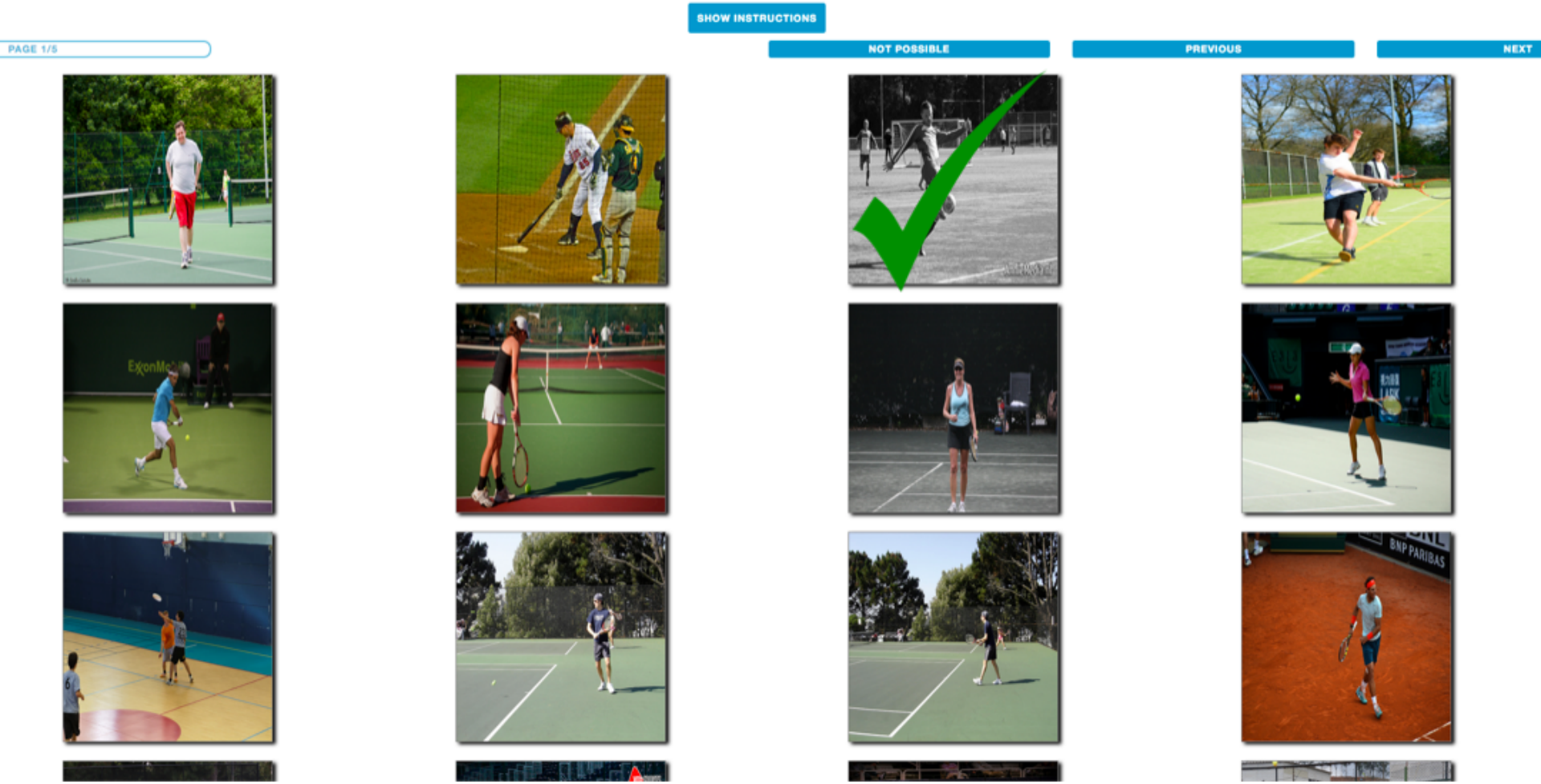
Slide credit: Devi Parikh

Balancing the VQA dataset

What game is this?
Tennis

Slide credit: Devi Parikh

Balancing the VQA dataset



Slide credit: Devi Parikh

Balancing the VQA dataset

Is the TV on?

yes

no

Slide credit: Devi Parikh

Balancing the VQA dataset

How many pets are present?

2

1

Slide credit: Devi Parikh

Balancing the VQA dataset

What sign is this?

handicap

one way

Slide credit: Devi Parikh

Balancing the VQA dataset

Where is the child sitting?

fridge

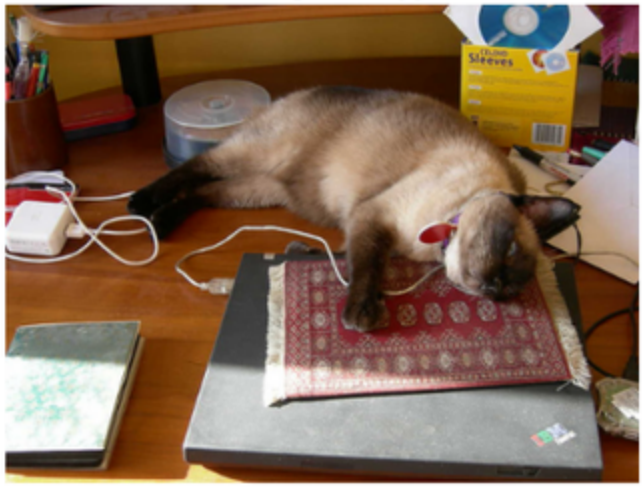
arms

Slide credit: Devi Parikh

Balancing the VQA dataset

What is the cat doing on the rug?

sleeping



sitting

Slide credit: Devi Parikh

Balancing the VQA dataset

What color are the pants?

orange

brown

Slide credit: Devi Parikh

VQA v2.0

- More balanced than VQA v1.0
  - Entropy of answers increases by 56%
- Bigger than VQA v2.0
  - ~1.8 times image-question pairs

Slide credit: Devi Parikh

Benchmarking SOTA VQA models

- SOTA VQA models
  - Drop in performance by 7-8%
  - Gain 1-2% back when re-trained on balanced dataset
- By answer types
  - Biggest drop in performance in yes/no (10-12%)
  - Biggest improvement gained by re-training in yes/no (3-4%) and number (2-3%)

Slide credit: Devi Parikh

Trends

	By Answer Type			Overall
	Yes/No	Number	Other	
UC Berkeley & Sony[14]	83.79	38.9	58.64	66.9
Naver Labs[10]	83.78	37.67	54.74	64.89
DLAIT[5]	83.65	39.18	52.62	63.97
snubi-naverlabs[25]	83.64	38.43	51.61	63.4

0.15% 1.51% 7.03% 3.5%

Slide credit: Devi Parikh

VQA v2.0

2nd and 3rd VQA Challenges @ CVPR17, 18.

Slide credit: Devi Parikh

Removing Language Priors

Slide credit: Yash Goyal and Peng Zhang

Removing Language Priors

Answer: No

Answer: Yes

complementary scenes

Question: Is the girl walking the bike?

Slide credit: Yash Goyal and Peng Zhang

Outline

- Datasets
  - E.g., VQA v2.0
- Evaluation metrics
  - E.g., # pairs of images correctly answered
- Novel problem spaces
  - E.g., changing priors (VQA-CP), novel object captioning, robust captioning, visual coreference resolution
- Inductive biases in models
  - E.g., G-VQA, Neural Baby Talk
- Human-in-the-loop evaluation
  - E.g., accuracy is not the final metric, performance of human-AI team is

Slide credit: Devi Parikh

Classifying a pair of complementary scenes

	Training set	
	Unbalanced	Balanced
Blind (no image features)		
Holistic image features		
Attention-based image features		

Slide credit: Devi Parikh

Outline

- Datasets
  - E.g., VQA v2.0
- Evaluation metrics
  - E.g., # pairs of images correctly answered
- Novel problem spaces
  - E.g., changing priors (VQA-CP), novel object captioning, robust captioning, visual coreference resolution
- Inductive biases in models
  - E.g., G-VQA, Neural Baby Talk
- Human-in-the-loop evaluation
  - E.g., accuracy is not the final metric, performance of human-AI team is

Slide credit: Devi Parikh

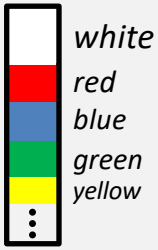
(A related) problem with existing setup

Train

Q: What color is the dog?

A: White

Training Prior



Slide credit: Aishwarya Agrawal

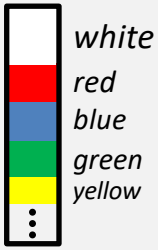
(A related) problem with existing setup

Train

Q: What color is the dog?

A: White

Training Prior



Test

Q: What color is the dog?

A: Black

Slide credit: Aishwarya Agrawal

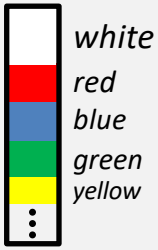
(A related) problem with existing setup

Train

Q: What color is the dog?

A: White

Training Prior



Test

Q: What color is the dog?

A: Black

Prediction: White

Slide credit: Aishwarya Agrawal

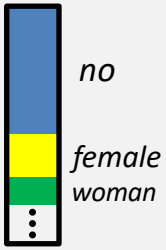
(A related) problem with existing setup

Train

Q: Is the person wearing shorts?

A: No

Training Prior



Test

Q: Is the person wearing shorts?

A: Yes

Prediction:
No

Slide credit: Aishwarya Agrawal

(A related) problem with existing setup

- Similar priors in train and test
- Memorization does not hurt as much
- Problematic for benchmarking progress

Slide credit: Aishwarya Agrawal

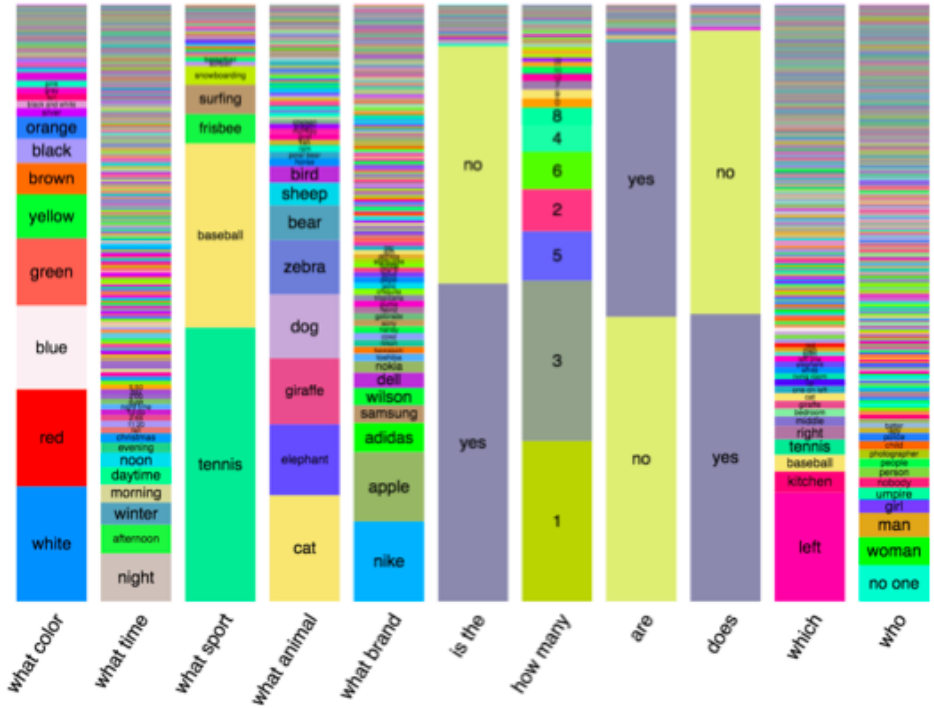
Meet VQA-CP!

- Visual Question Answering under Changing Priors

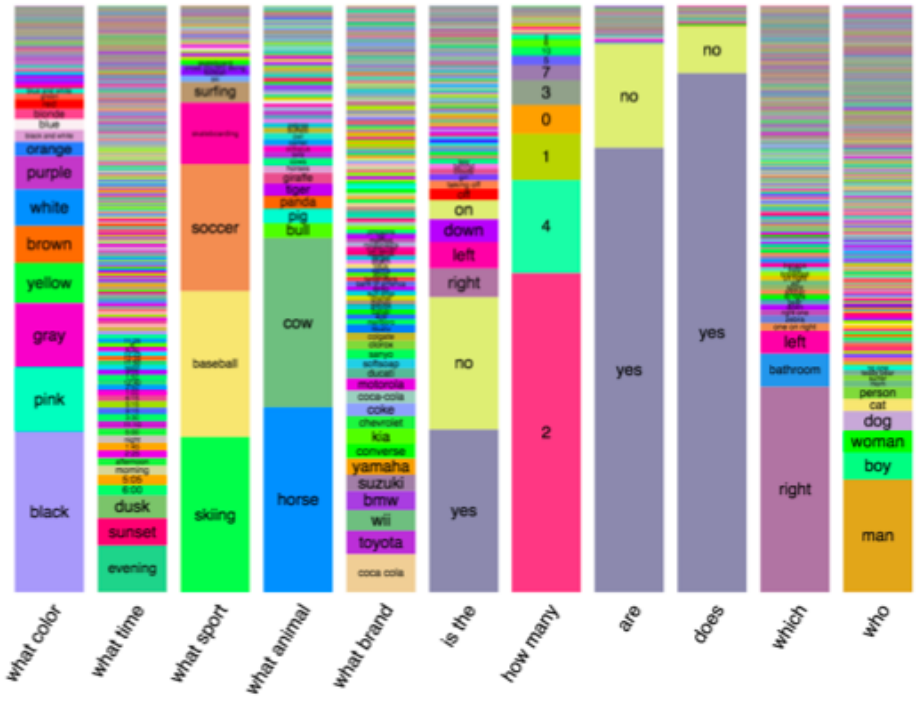
- A new split of the VQA v1.0 dataset (Antol et al., ICCV 2015)

Slide credit: Aishwarya Agrawal

VQA-CP Train Split



VQA-CP Test Split



Slide credit: Aishwarya Agrawal

Performance of VQA models on VQA-CP

Model	Dataset	Overall	Yes/No	Number	Other	
d-LSTM Q + norm I (Antol et al. ICCV15)	VQA	54.23	79.81	33.26	40.35	31% drop
	VQA-CP	23.51	34.53	11.40	17.42	
NMN (Andreas et al. CVPR16)	VQA	54.83	80.39	33.45	41.07	25% drop
	VQA-CP	29.64	38.85	11.23	27.88	
SAN (Yang et al. CVPR16)	VQA	55.86	78.54	33.46	44.51	29% drop
	VQA-CP	26.88	35.34	11.34	24.70	
MCB (Fukui et al. EMNLP16)	VQA	60.97	81.62	34.56	52.16	27% drop
	VQA-CP	34.39	37.96	11.80	39.90	

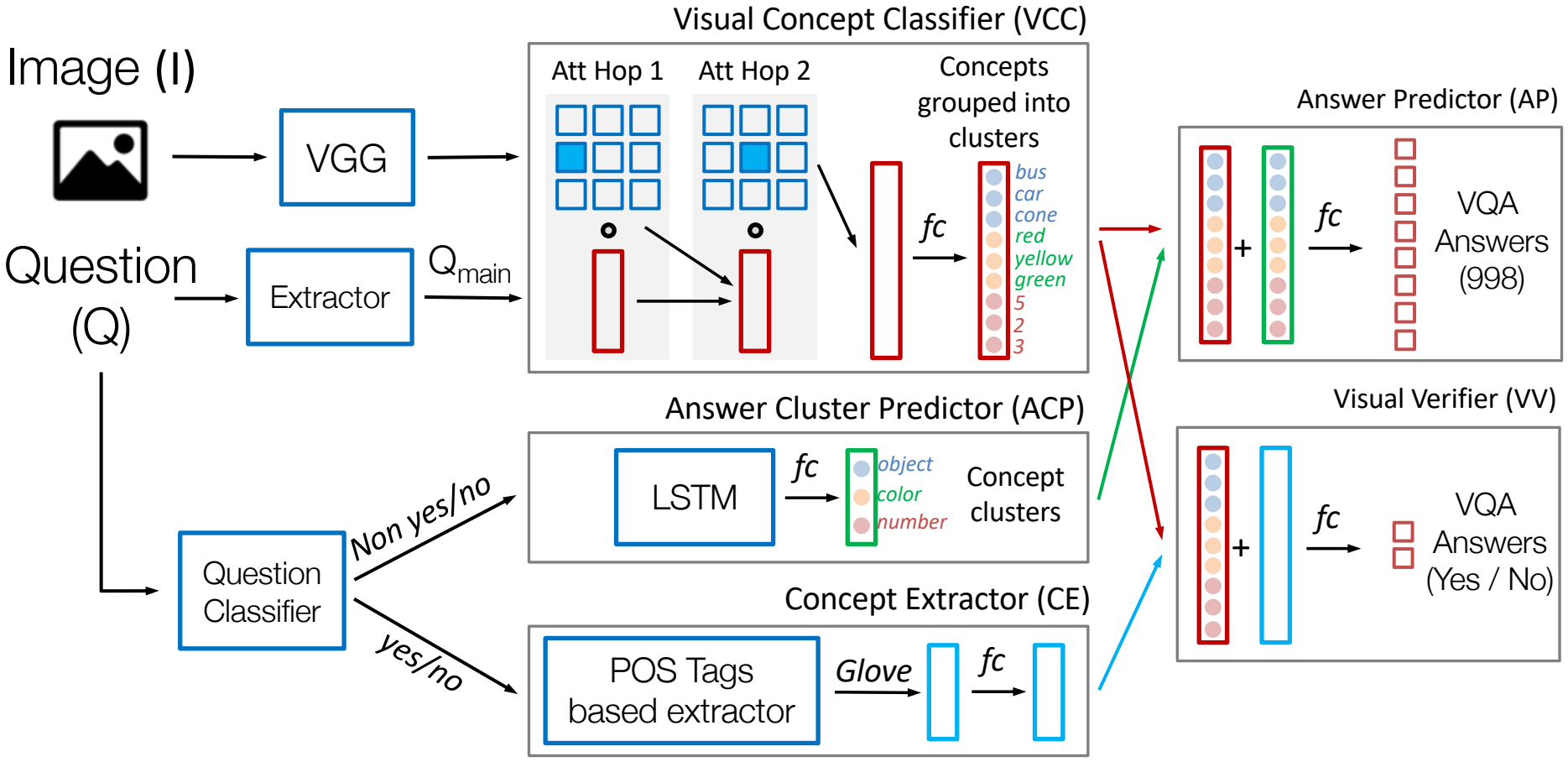
Slide credit: Aishwarya Agrawal

Outline

- Datasets
  - E.g., VQA v2.0
- Evaluation metrics
  - E.g., # pairs of images correctly answered
- Novel problem spaces
  - E.g., changing priors (VQA-CP), novel object captioning, robust captioning, visual coreference resolution
- Inductive biases in models
  - E.g., G-VQA, Neural Baby Talk
- Human-in-the-loop evaluation
  - E.g., accuracy is not the final metric, performance of human-AI team is

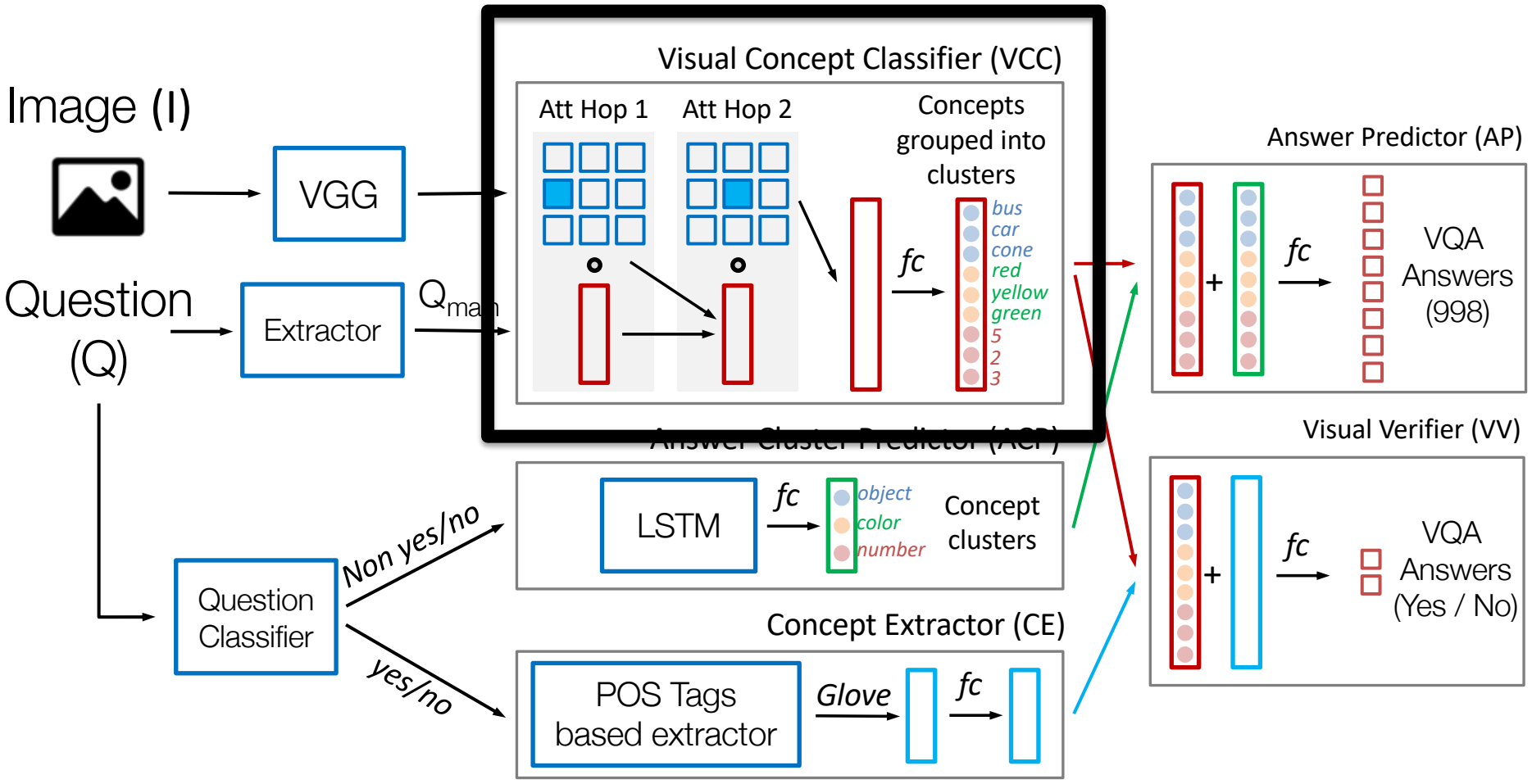
Slide credit: Devi Parikh

Grounded-VQA (GVQA)



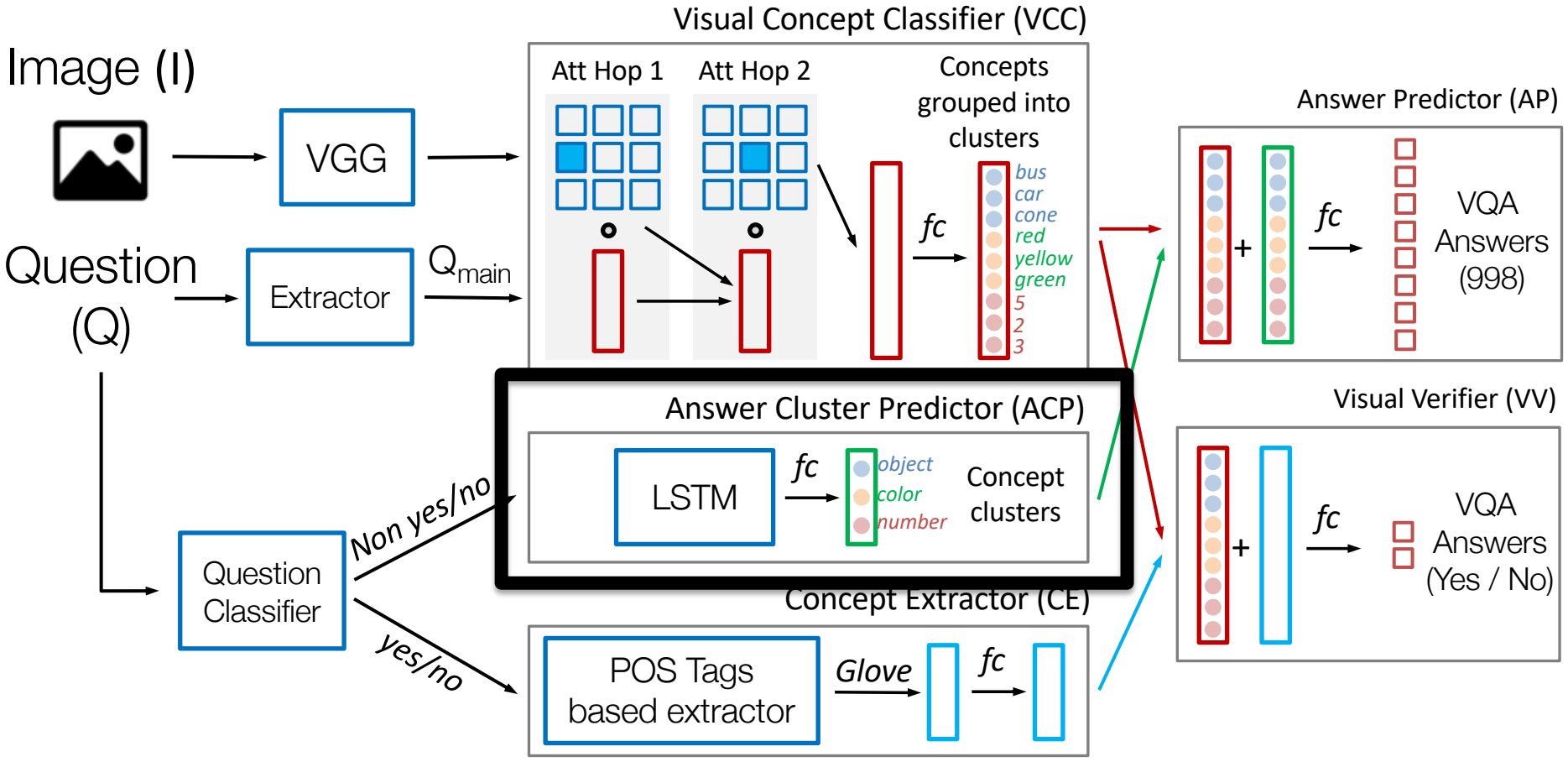
Slide credit: Aishwarya Agrawal

Grounded-VQA (GVQA)



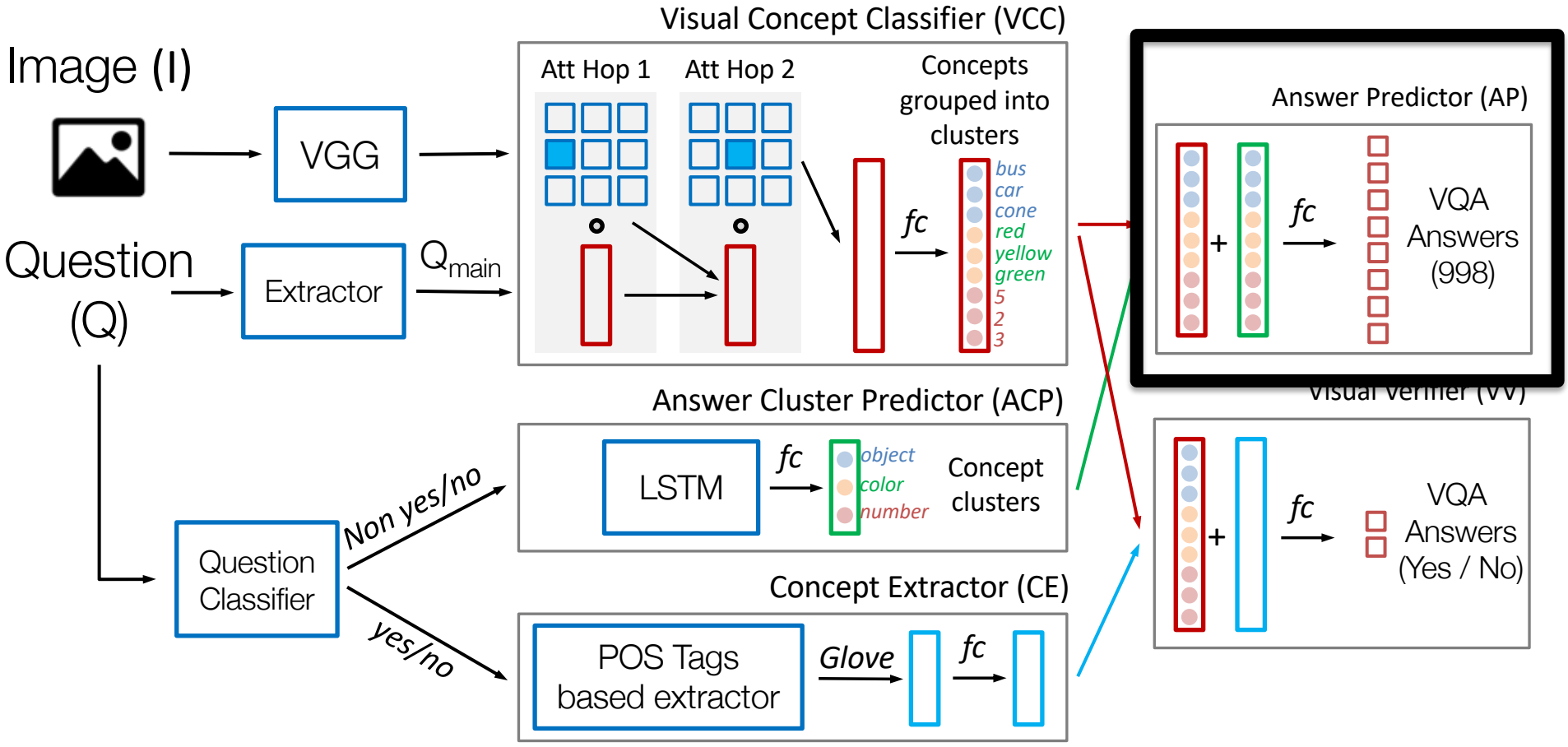
Slide credit: Aishwarya Agrawal

Grounded-VQA (GVQA)



Slide credit: Aishwarya Agrawal

Grounded-VQA (GVQA)



Slide credit: Aishwarya Agrawal

Visual Coreference Resolution in Visual Dialog using Neural Module Networks

C : The boat has a dragon head on the

Q1 : Is the boat on water ?

A1 : Yes

Q2 : What color is it ?

A2 : White

Q3 : Does the dragon have a body ?

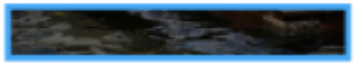
A3 : No, just the head

Q4 : What color is it ?

Known Entities

boat

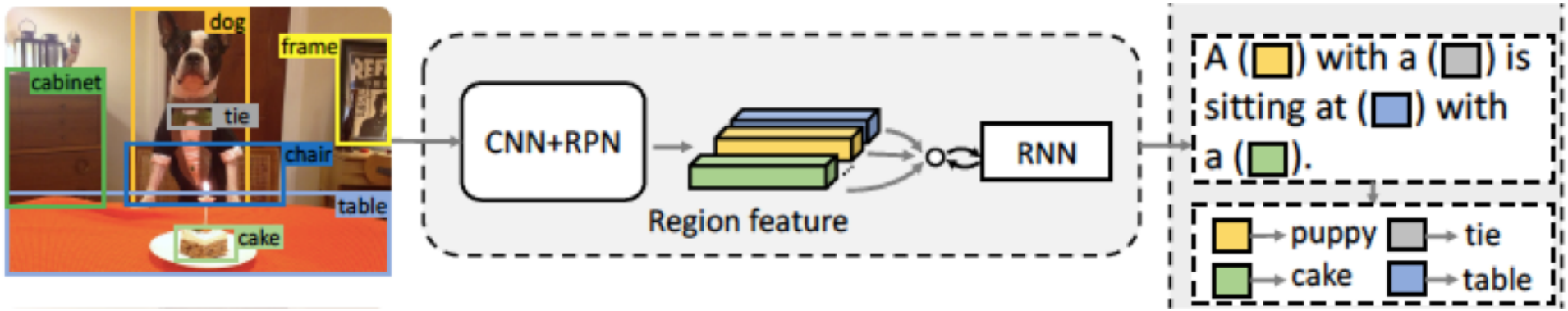
dragon head



water

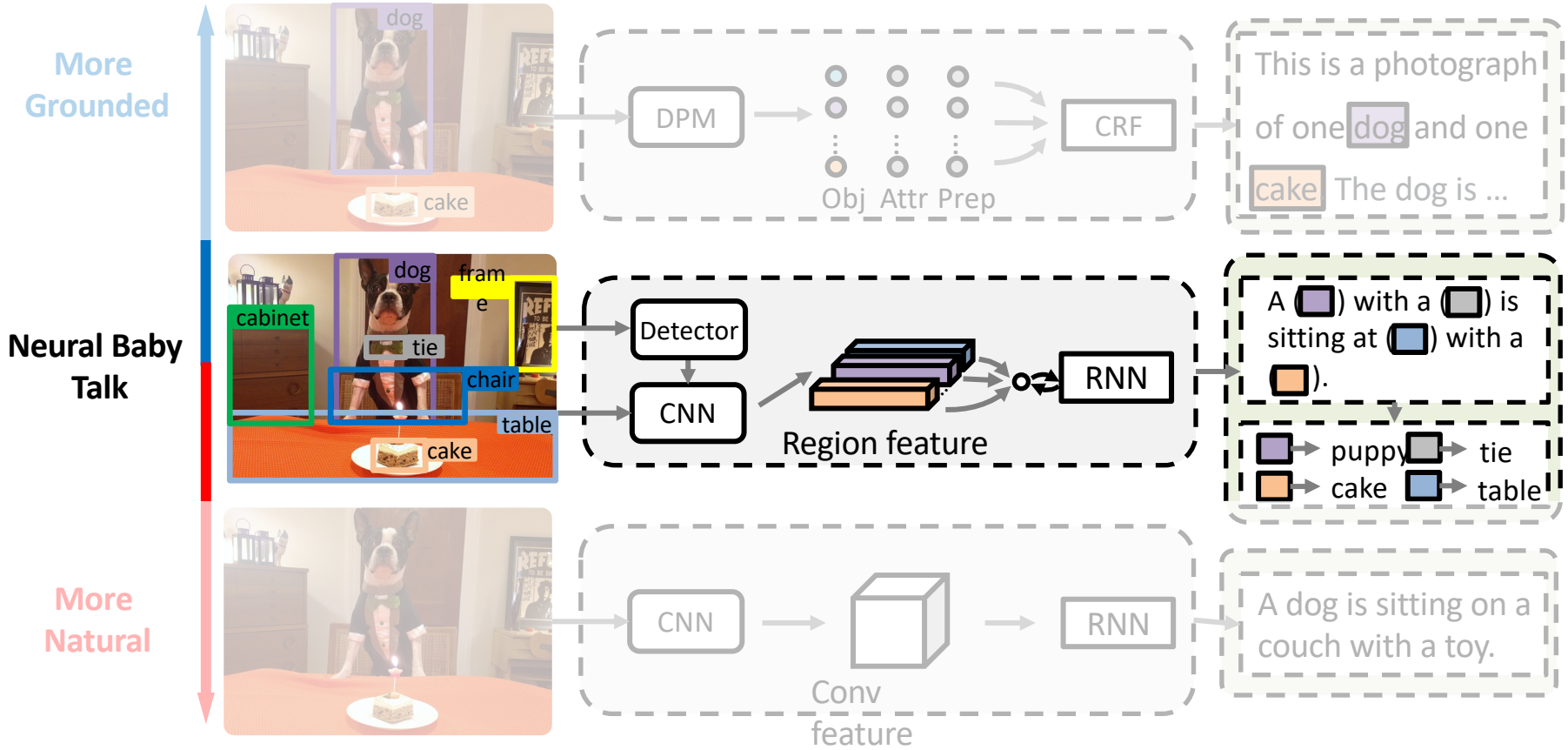
Slide credit: Devi Parikh

Neural Baby Talk



Slide credit: Devi Parikh

Framework

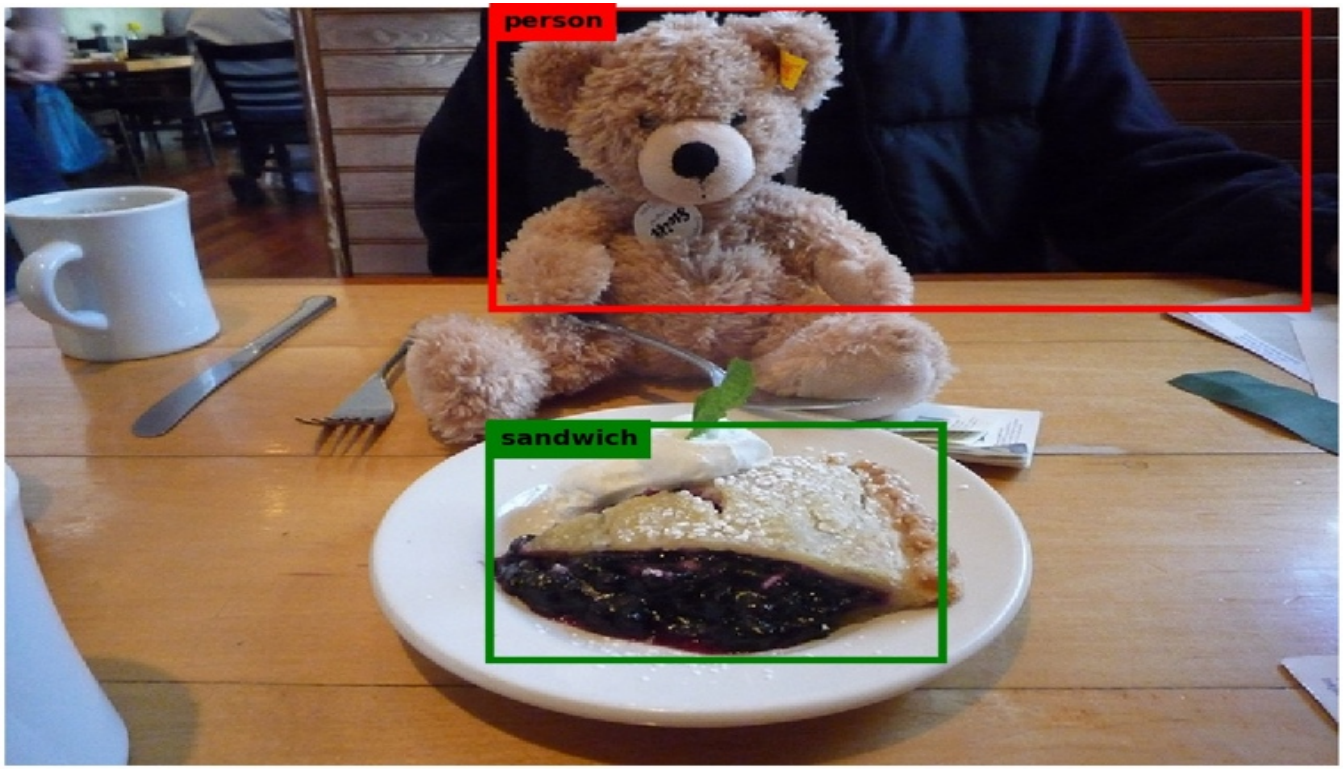


Slide credit: Jiasen Lu

A close up of a stuffed animal on a plate.

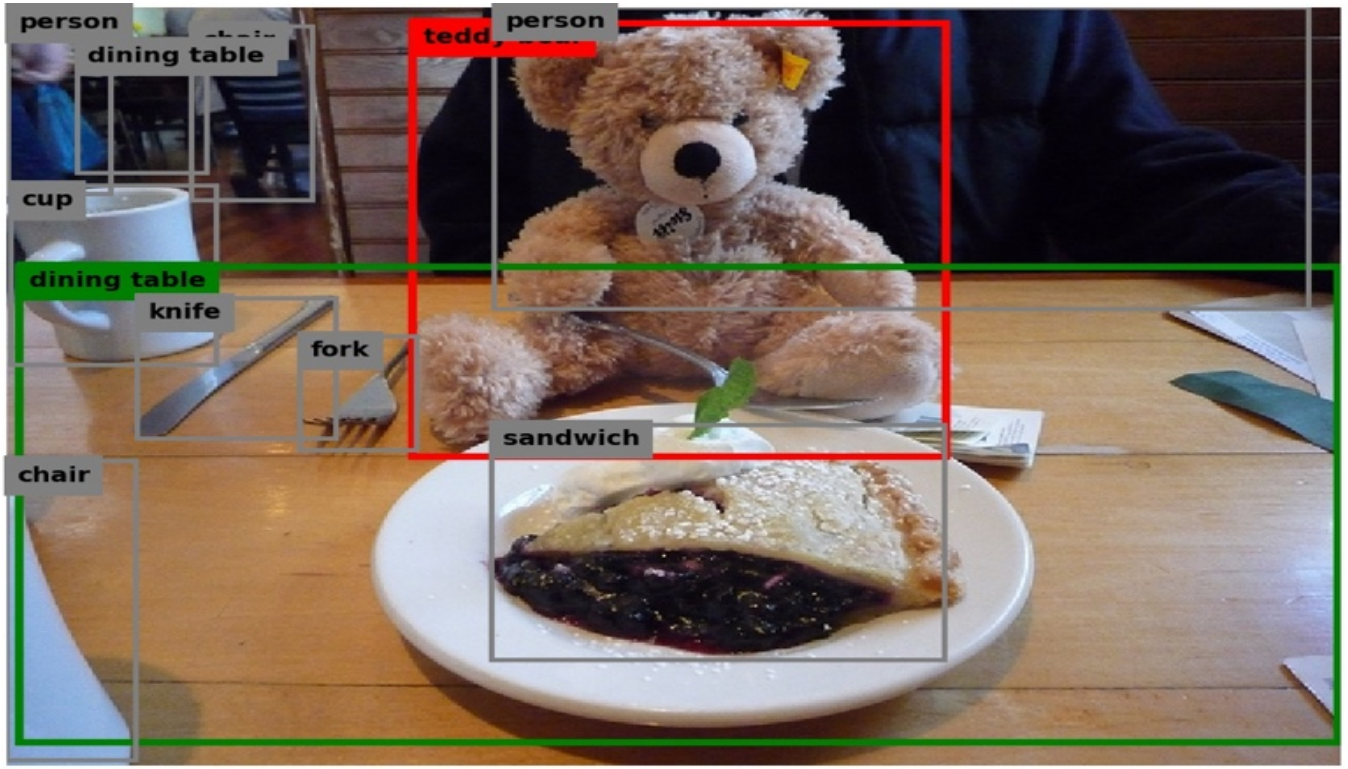
Slide credit: Jiasen Lu

A person is sitting at a table with a sandwich.



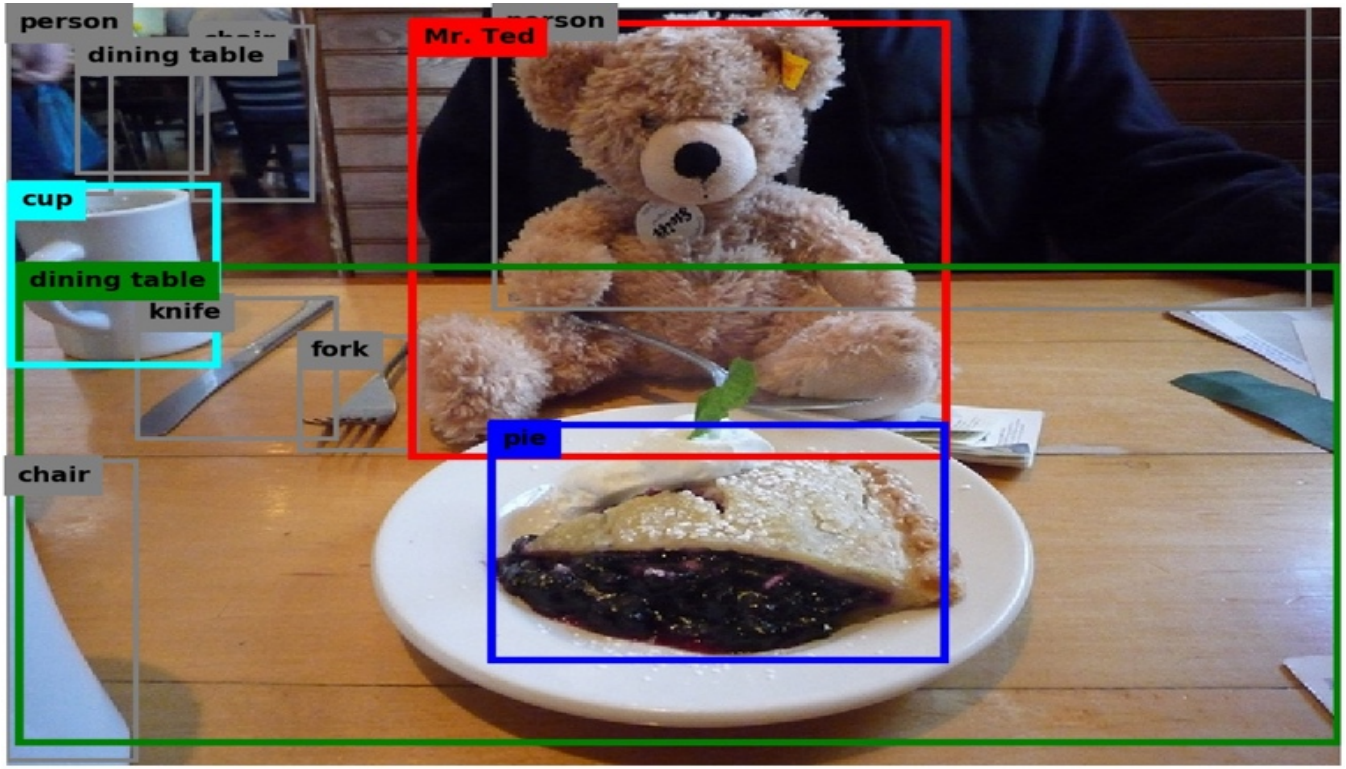
Slide credit: Jiasen Lu

A teddy bear sitting on a table with a plate of food.



Slide credit: Jiasen Lu

A Mr. Ted sitting at a table with a pie and a cup of coffee.



Slide credit: Jiasen Lu

Outline

- Datasets
  - E.g., VQA v2.0
- Evaluation metrics
  - E.g., # pairs of images correctly answered
- Novel problem spaces
  - E.g., changing priors (VQA-CP), novel object captioning, robust captioning, visual coreference resolution
- Inductive biases in models
  - E.g., G-VQA, Neural Baby Talk
- Human-in-the-loop evaluation
  - E.g., accuracy is not the final metric, performance of human-AI team is

Slide credit: Devi Parikh

GuessWhich Game

Total points earned: **0** Possible points for this game: **200**

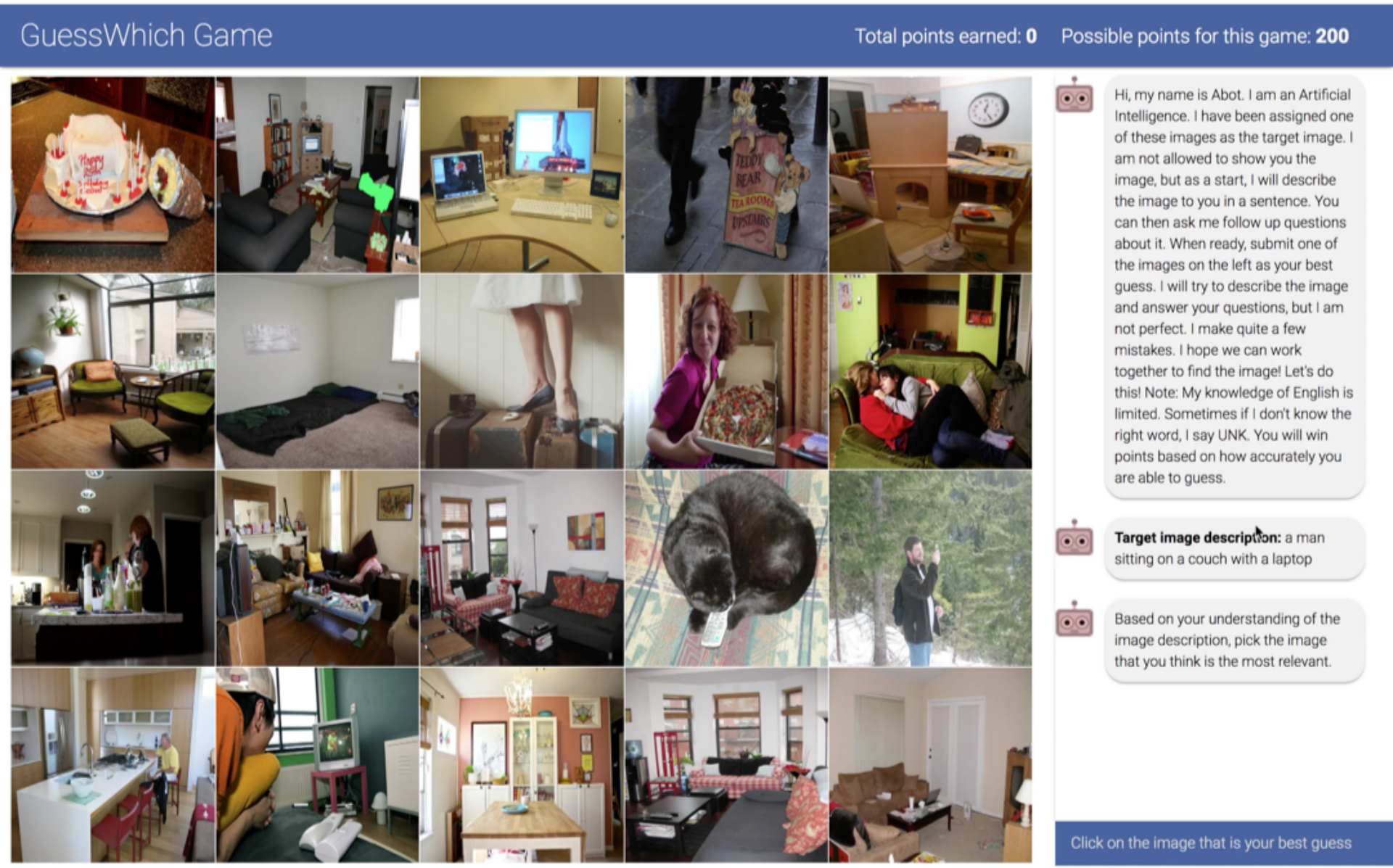
Hi, my name is Quincy. I am an Artificial Intelligence. You can see an image on your screen. I cannot see this image, but can only see a one line description of this image. My task is to find this target image from a large pool of images. Please help me! I will ask you questions about the image, and your answers will help me find the right image. I will try to ask reasonable questions, but I am not perfect. I make quite a few mistakes. And my memory is a little weak. So please be patient with me, I may ask the same question more than once. I hope we can work together to find the image! Let's do this! Note: My knowledge of English is limited. Sometimes if I don't know the right word, I say UNK.

Target image description: Some boats are going around the river near the docks.

Q: 1/9 what color are boats ?

Start typing answer here ...

Slide credit: Devi Parikh



Slide credit: Devi Parikh

Image Guessing Game

- Compared two bots
- One did better at this game in bot-bot evaluation
- Trend did not generalize when evaluated with humans*

* caveats apply

Slide credit: Devi Parikh

Summary

- VQA is a rich problem space
  - Vision, language, attention, reasoning, external knowledge, HCI, …
- Challenges:
  - Grounding and generalization
- Forcing models to look
  - Counter via datasets
    - E.g., VQA v2.0, VQA-CP
  - Counter via evaluation metrics
    - E.g., # pairs of images correctly answered
  - Counter via novel problem spaces
    - E.g., novel object captioning, robust captioning, visual coreference resolution
  - Counter via inductive biases in models
    - E.g., G-VQA, Neural Baby Talk
  - Counter via human-in-the-loop
    - E.g., accuracy is not the final metric, performance of human-AI team is

Slide credit: Devi Parikh

Thank you.

Slide credit: Devi Parikh