A tale of two negative results

Dhruv Batra

Georgia Tech

facebook research

A tale of ~~two~~ three negative results

Dhruv Batra

Georgia Tech

facebook research

three negative-ish

A tale of ~~two~~ ~~negative~~ results

Dhruv Batra

A tale of ~~two~~ three ~~negative~~ negative-ish results

and a rant about vision reviewing!

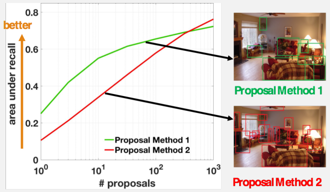
Dhruv Batra

Georgia Tech

facebook research

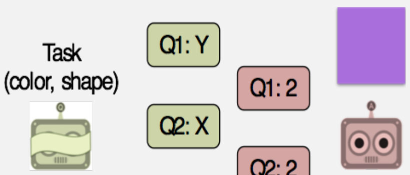
Outline

Object-Proposal Evaluation Protocol is ‘Gameable’ [CVPR ‘16]



Human Attention in VQA: Do Humans and Deep Networks Look at the Same Regions? [EMNLP ‘16]

Natural Language Does Not Emerge ‘Naturally’ in Multi-Agent Dialog [EMNLP ‘17]



Rant about Vision vs NLP reviewing!

Object-Proposal Evaluation Protocol is 'Gameable'

[CVPR '16]

Neelima Chavali* Harsh Agrawal* Aroma Mahendru* Dhruv Batra

What are object proposals?

A candidate set of regions or boxes that may contain an object.

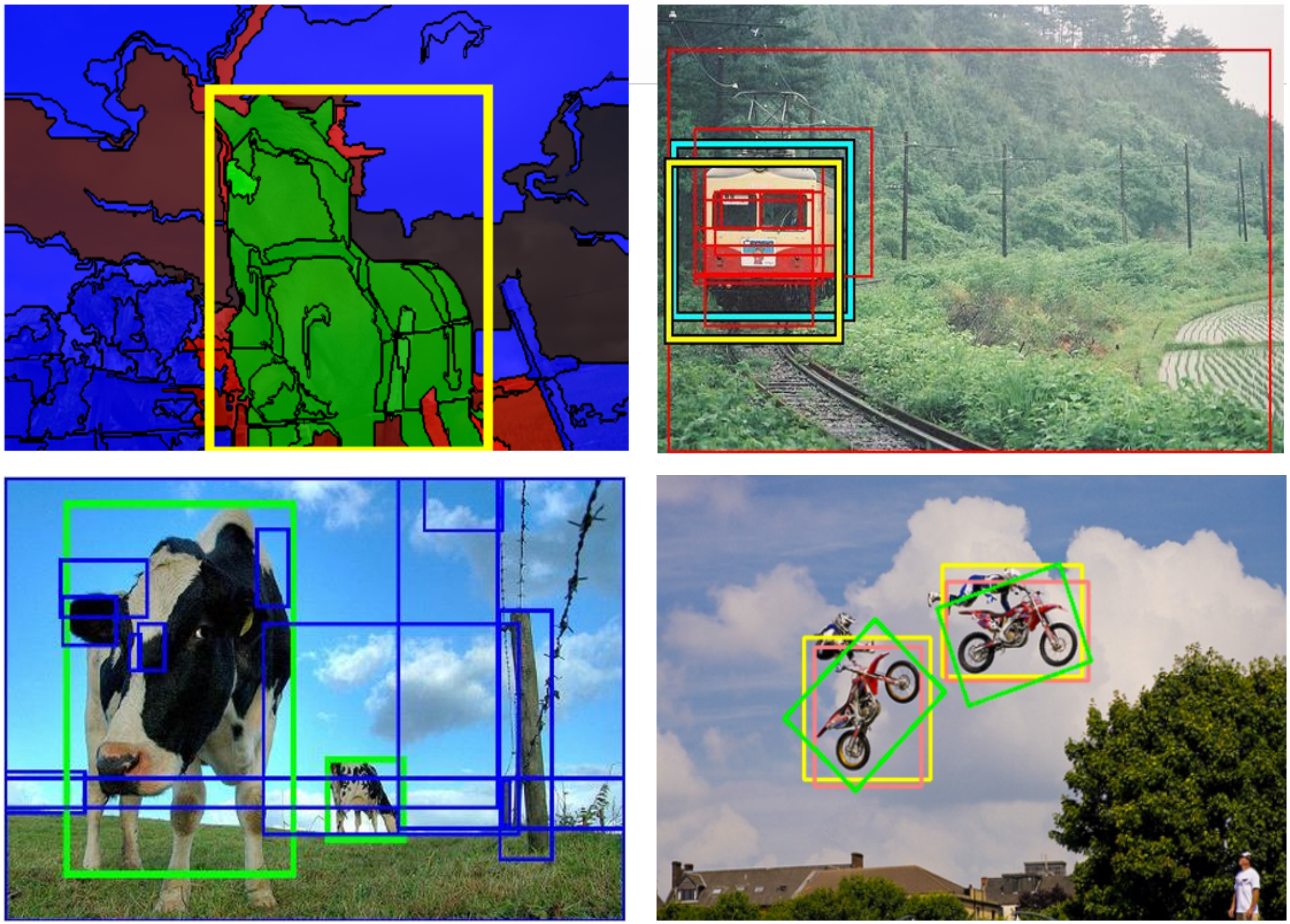
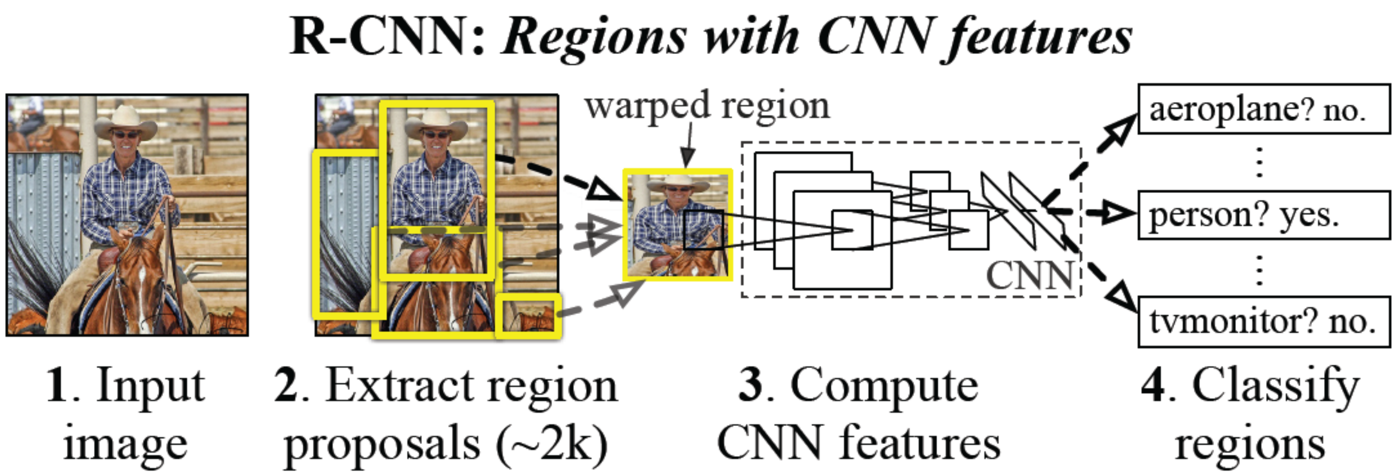
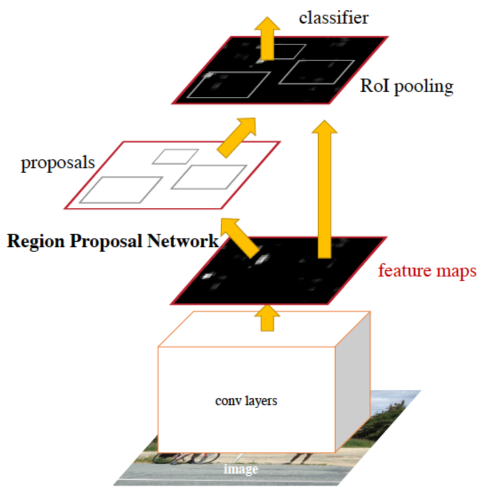


Image Credit: Piótr Dollar

Pipeline: R-CNN

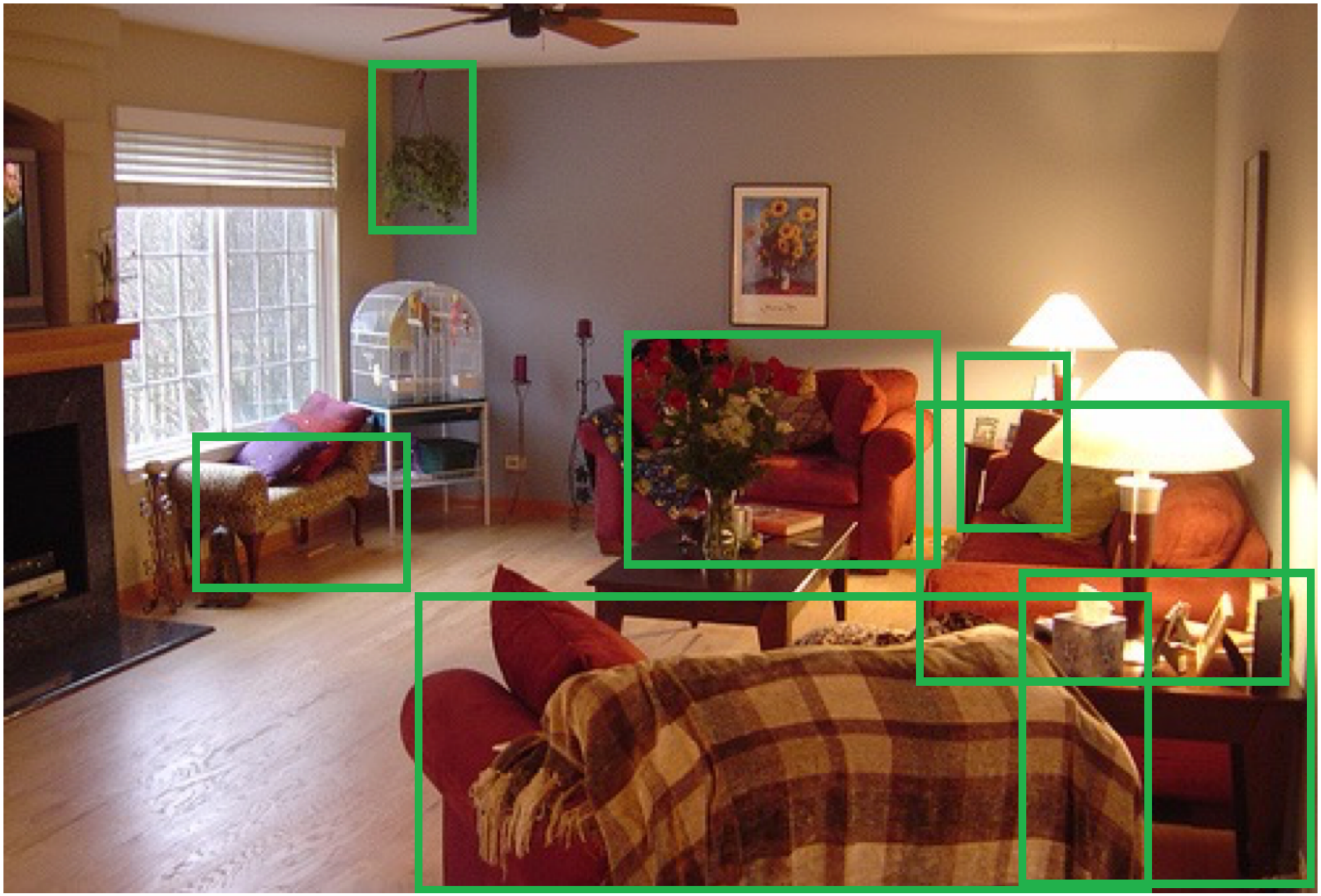


End-to-End: Faster R-CNN

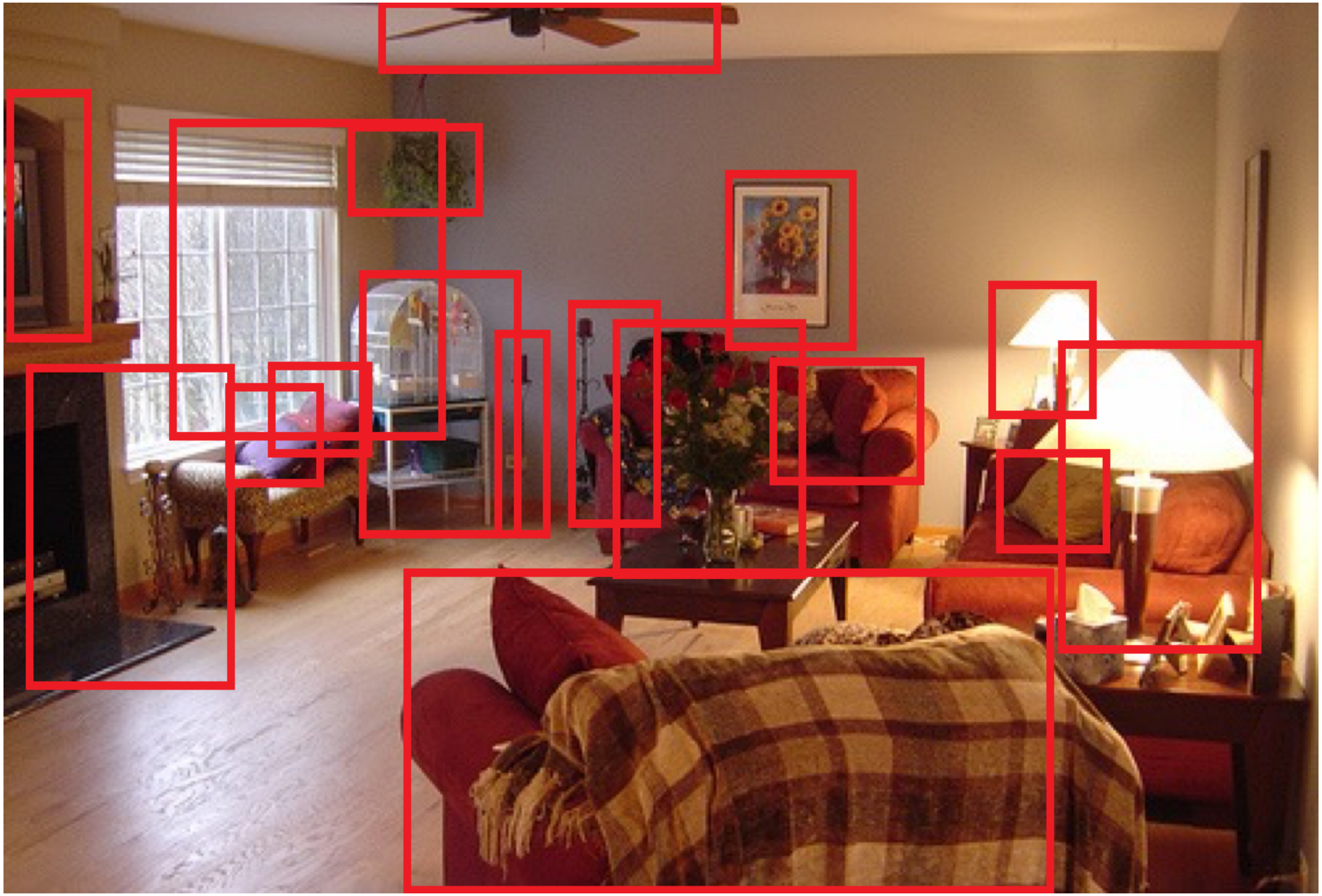


Thought Experiment

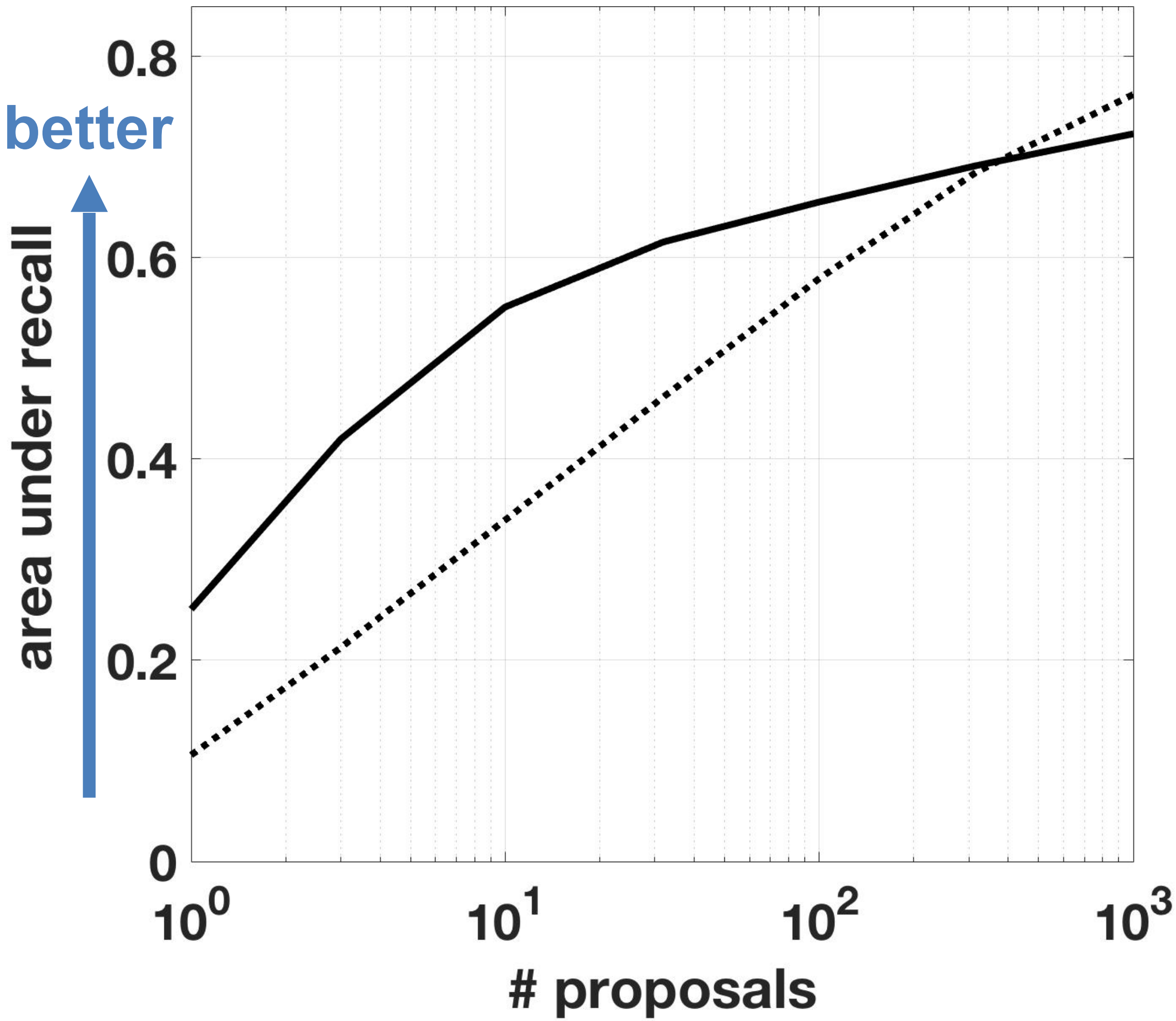
Thought Experiment: Proposal Method 1

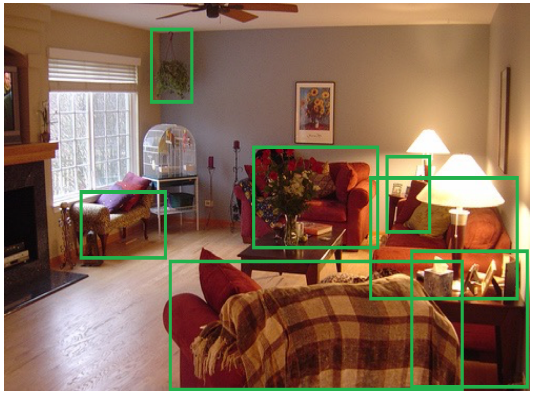


Thought Experiment: Proposal Method 2

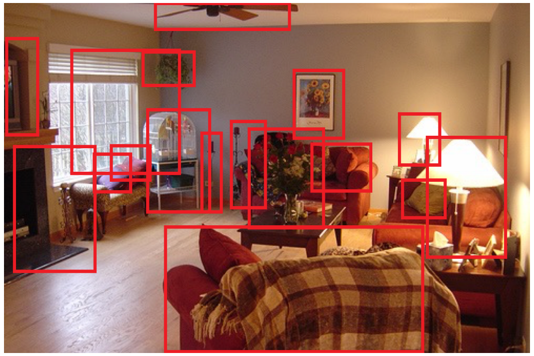


Which is better?



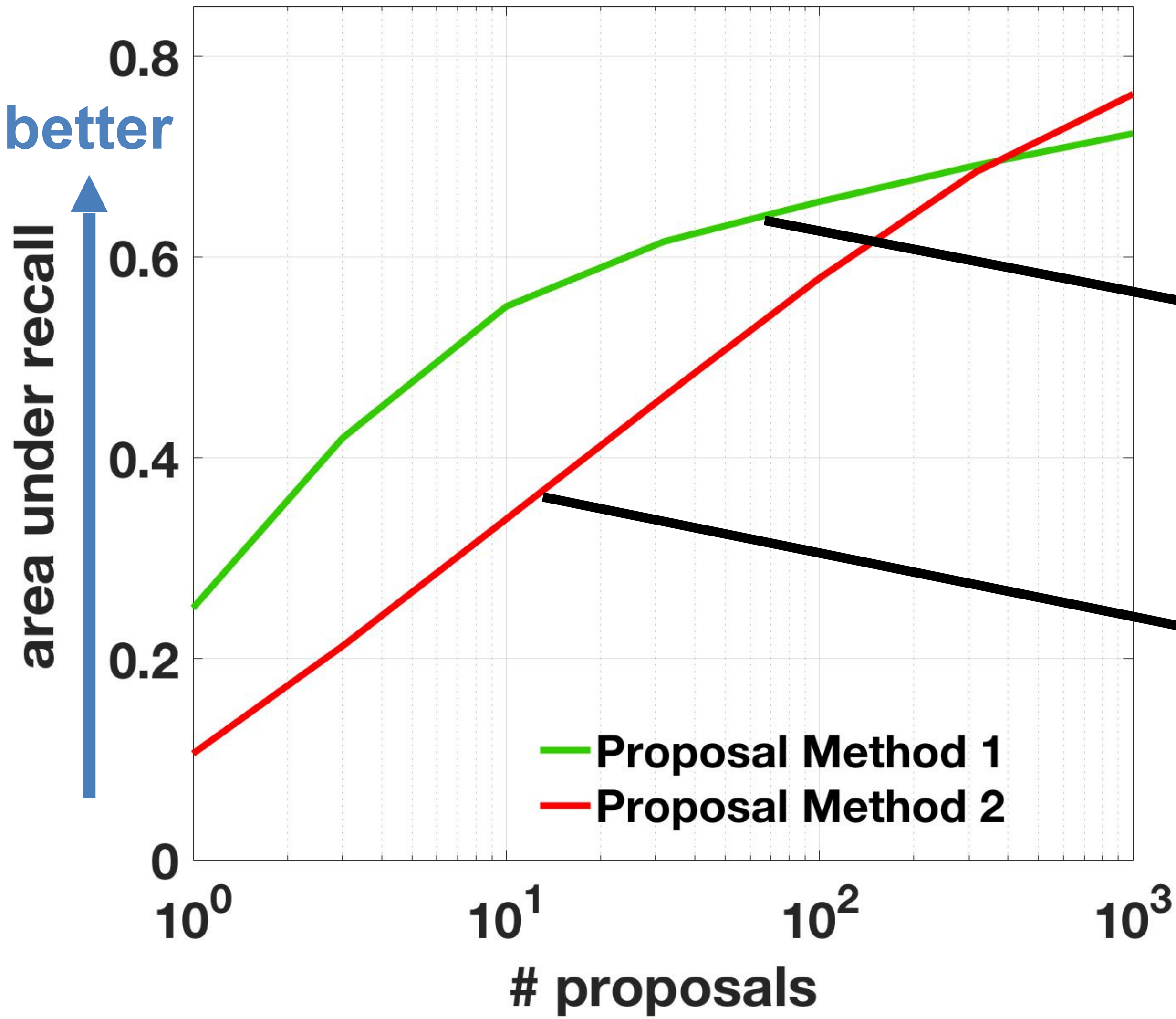


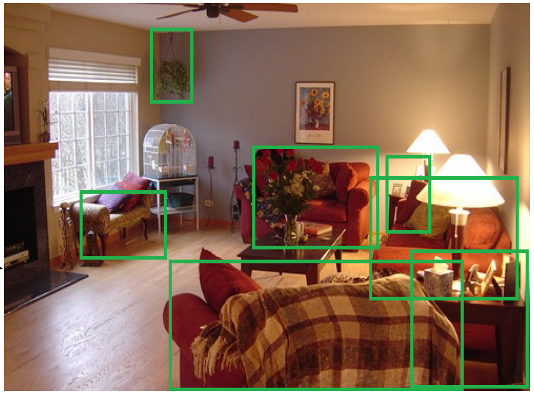
Proposal Method 1



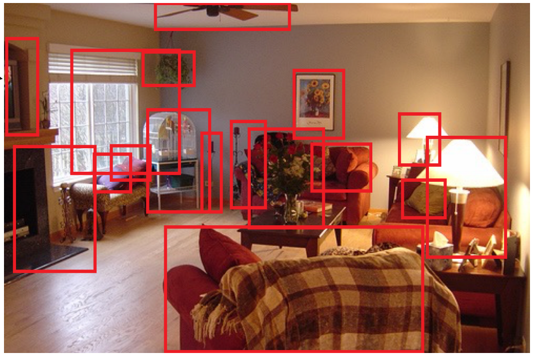
Proposal Method 2

Think Again!



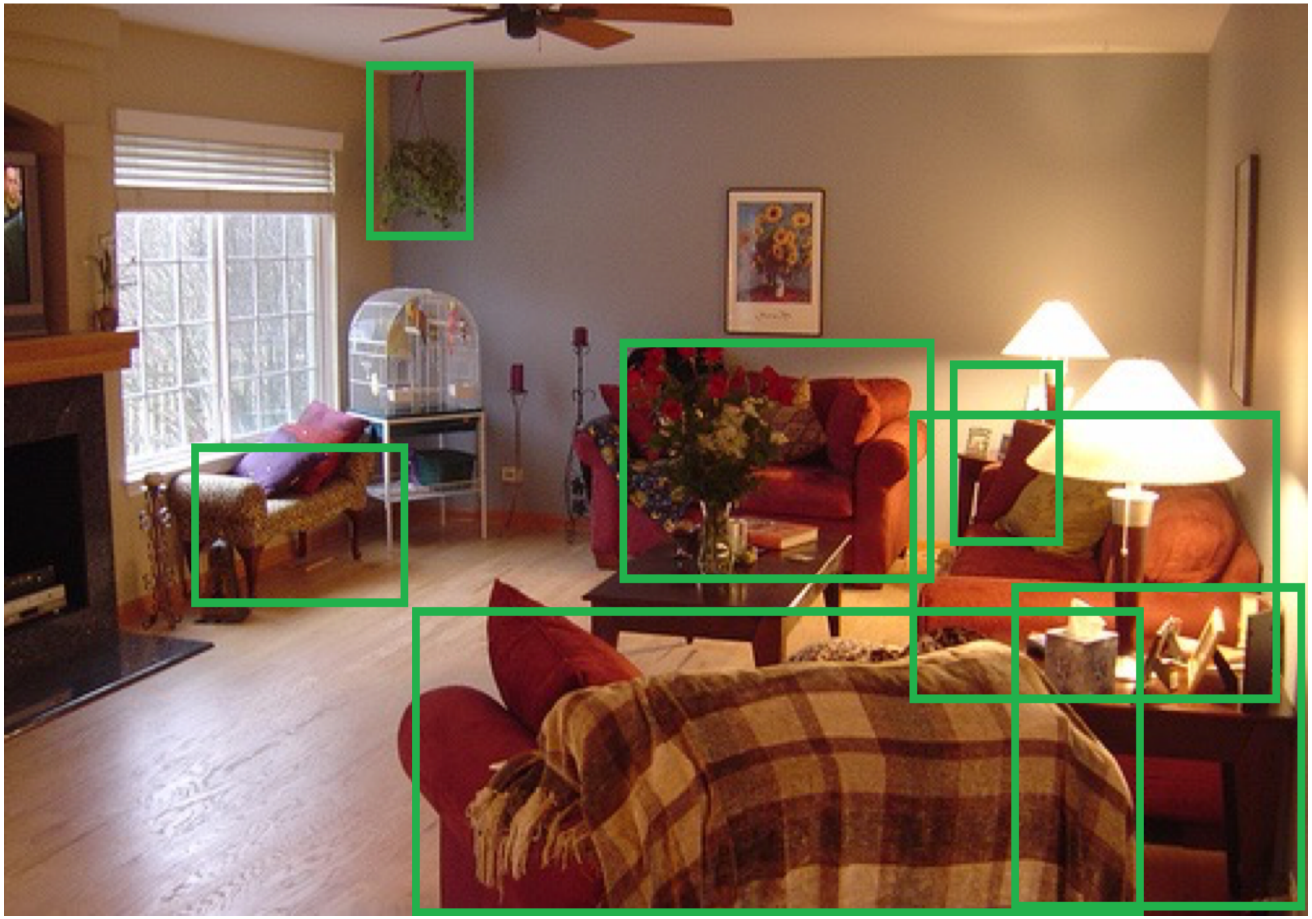


Proposal Method 1



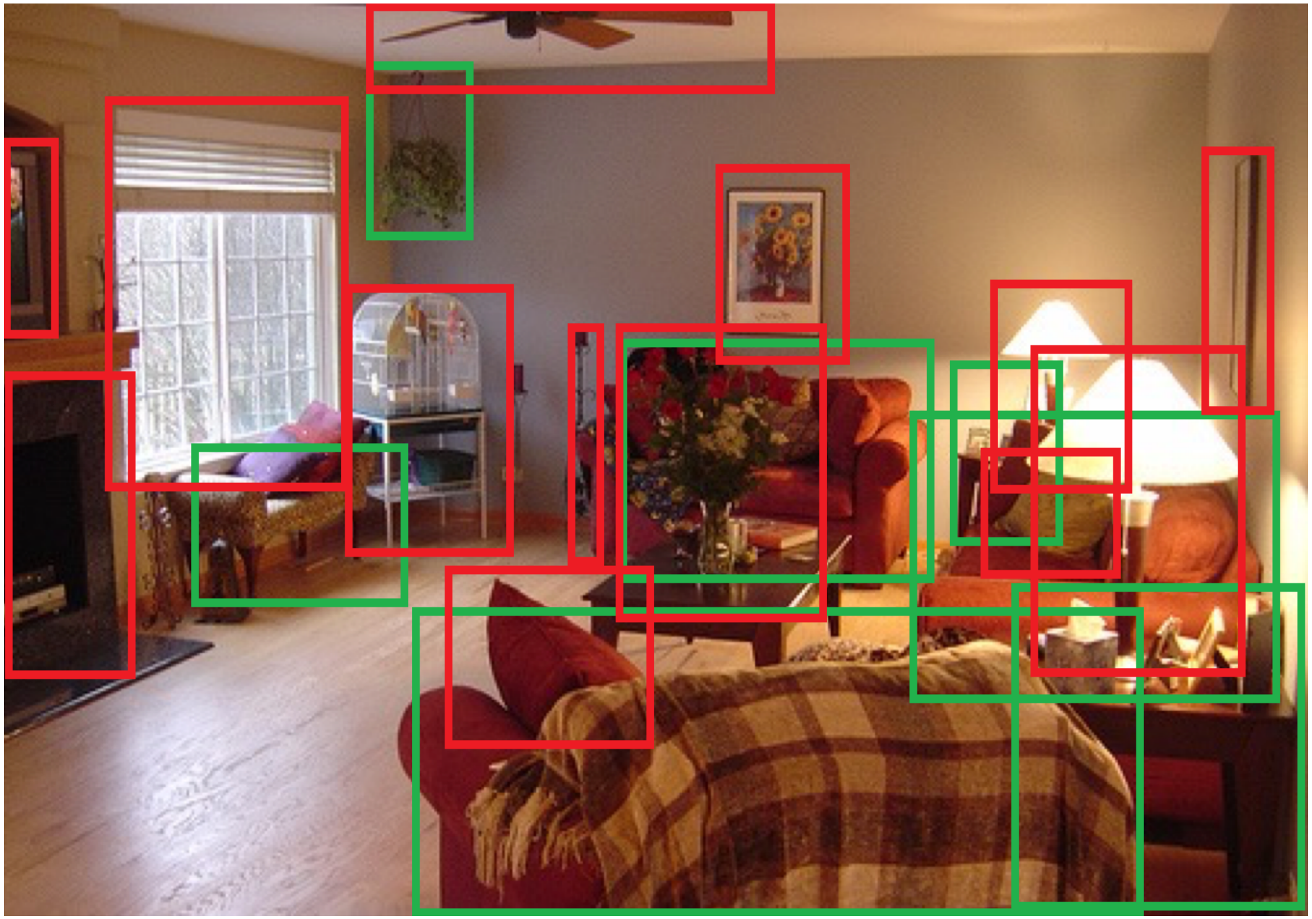
Proposal Method 2

What the ground-truth contains



- **Table**
- **Sofa**
- **Potted Plant**

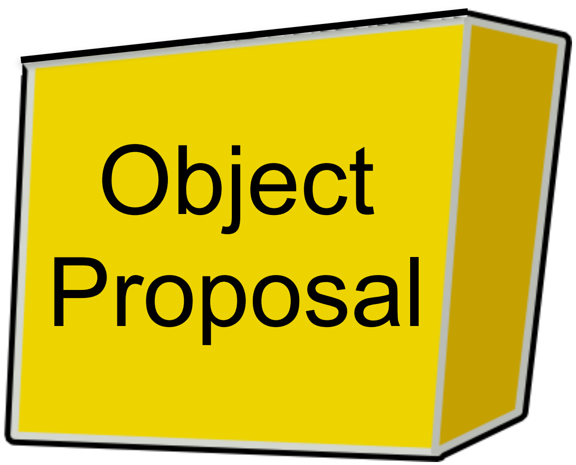
What the ground-truth is missing



- Table
- Sofa
- Potted Plant
- Picture Frame
- Lamp
- Fireplace
- Window
- Cushion
- Fan

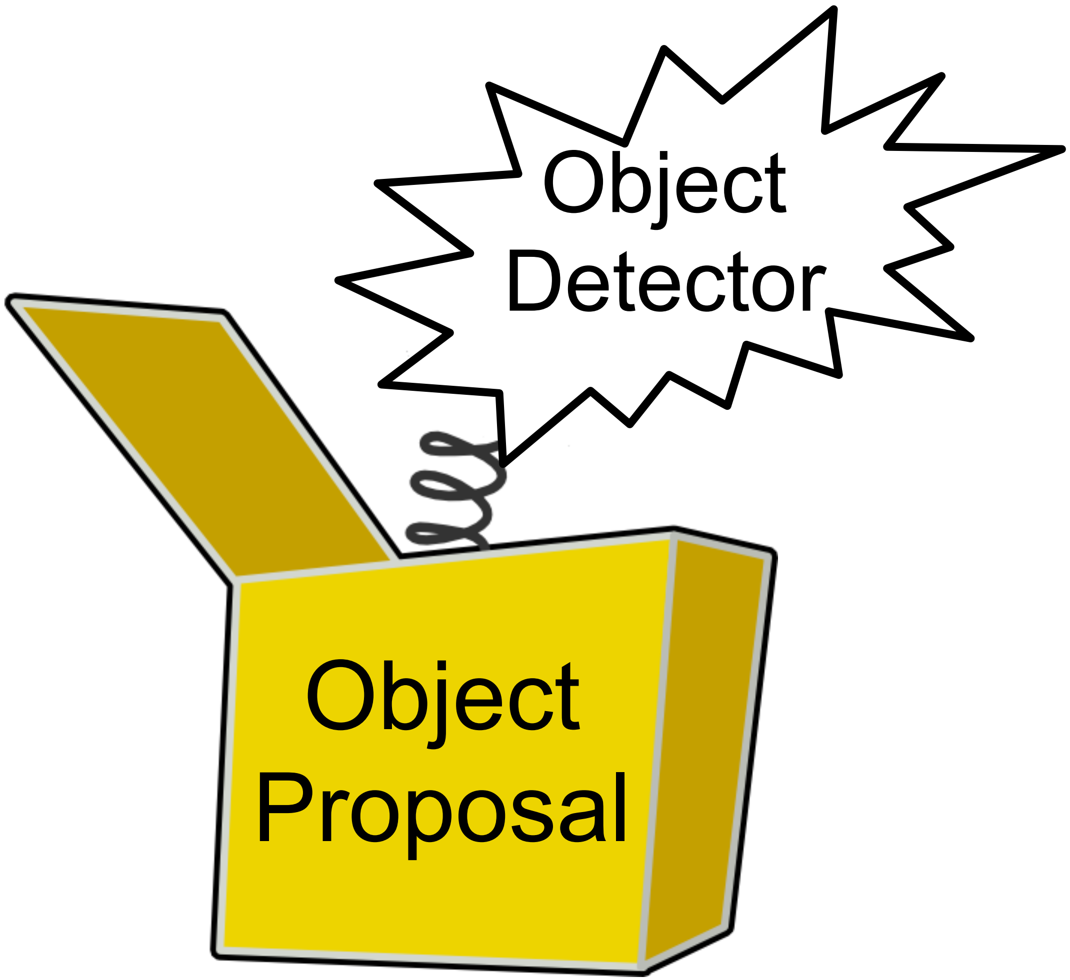
How to game the protocol?

Detector Masquerading as Proposals (DMP)



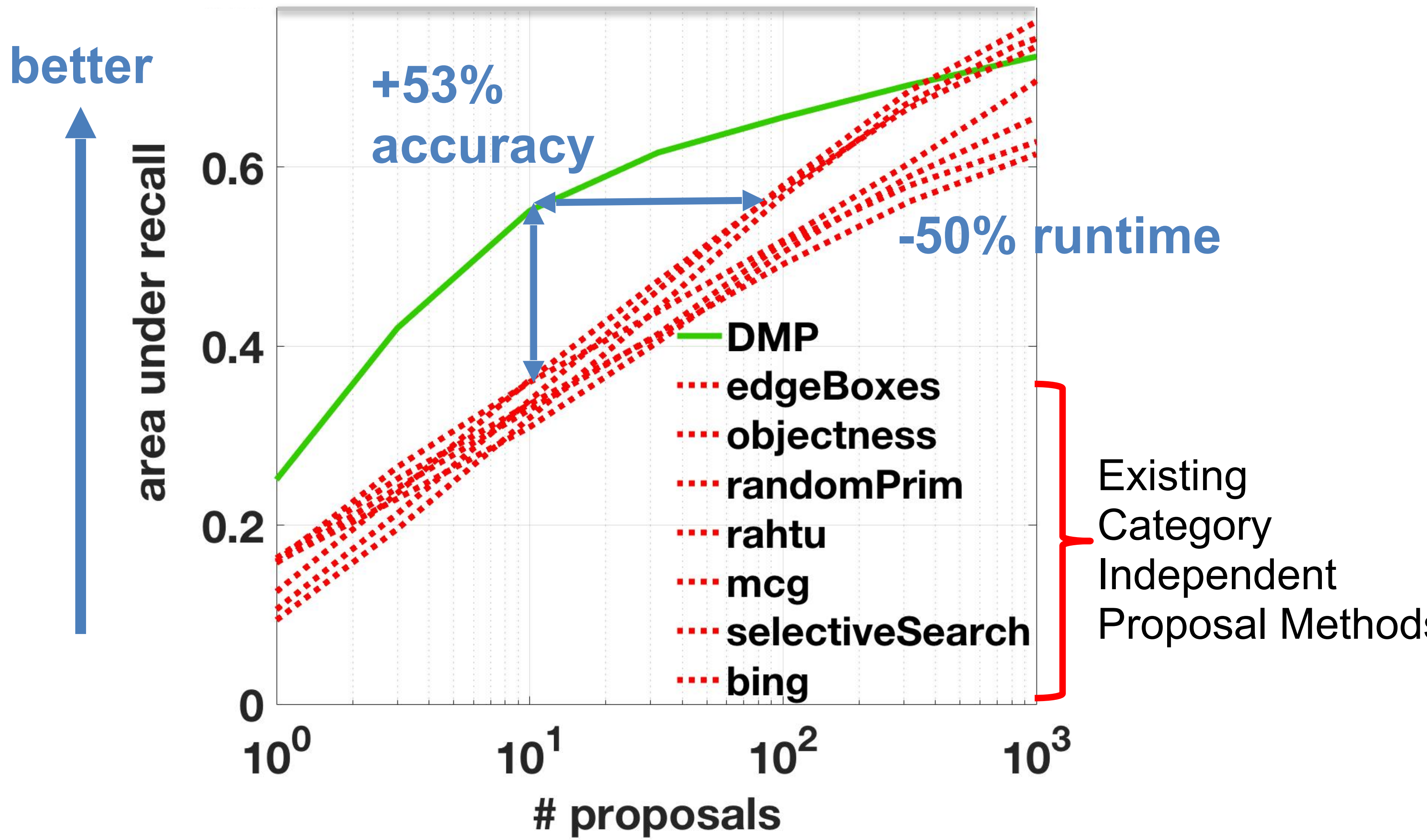
How to game the protocol?

Detector Masquerading as Proposals (DMP)



How to game the protocol?

Detector Masquerading as Proposals (DMP)

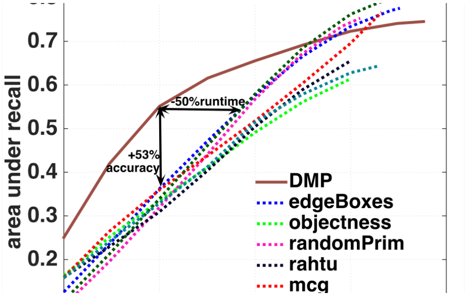


4. A Thought Experiment: How to Game the Evaluation Protocol

Let us conduct a thought experiment to demonstrate that the object proposal evaluation protocol can be 'gamed'.

Imagine yourself reviewing a paper claiming to introduce a new object proposal method – called DMP.

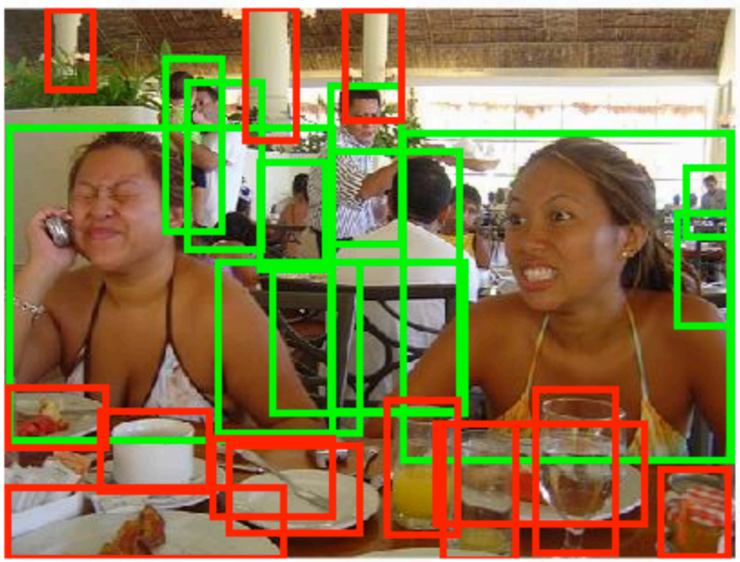
Before we divulge the details of DMP, consider the performance of DMP shown in Fig. 3 on the PASCAL VOC 2010 dataset, under the AUC-*vs.*-#proposals metric.



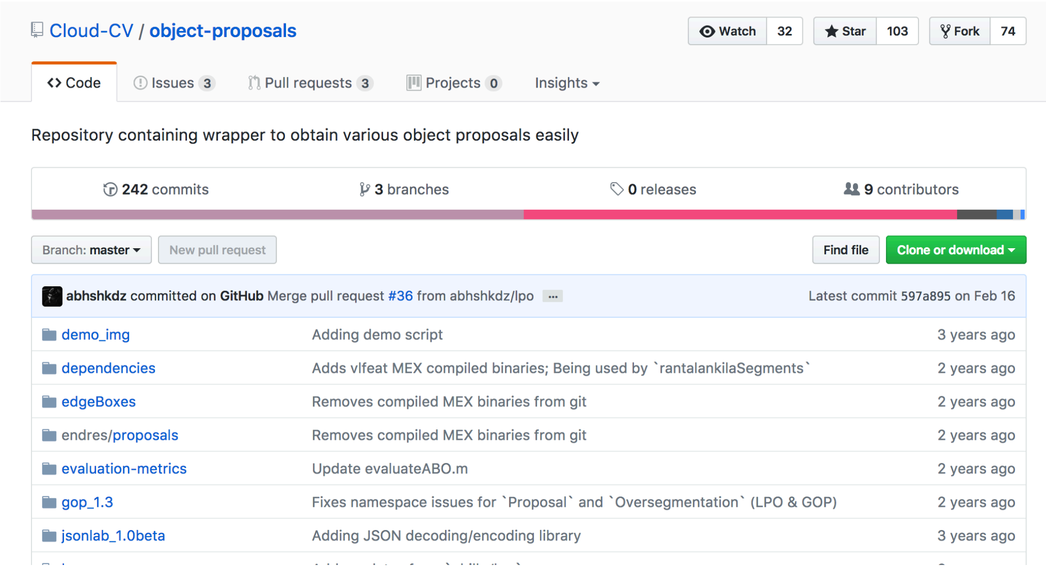
So what is our proposed state-of-art technique DMP?
It is a mixture-of-experts model, consisting of 20 experts, where each expert is a deep feature (fc7)-based [61] objectness detector. At this point, you, the savvy reader, are probably already beginning to guess what we did.

Contributions

- Alleviating the 'game-ability' by:
  - Evaluation on fully annotated dataset
  - Cross dataset evaluation on densely annotated datasets
  - Measuring bias capacity

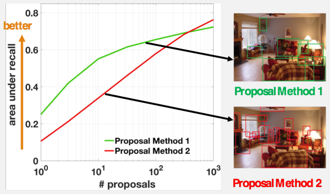


PASCAL-Context Instance Annotations

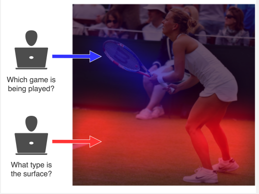


Outline

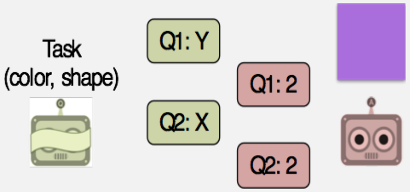
Object-Proposal Evaluation
Protocol is ‘Gameable’ [CVPR ‘16]



Human Attention in VQA:
Do Humans and Deep Networks
Look at the Same Regions? [EMNLP ‘16]



Natural Language Does Not Emerge
‘Naturally’ in Multi-Agent Dialog [EMNLP ‘17]



Rant about Vision vs NLP reviewing!

Human Attention in VQA: Do Humans & Deep Networks Look at the Same Regions?

[EMNLP '16]


Abhishek Das*


Harsh Agrawal*


Larry Zitnick


Devi Parikh


Dhruv Batra

VirginiaTech

facebook research

Georgia Institute of Technology

Visual questions target different image regions

Which game is being played?

What type is the surface?

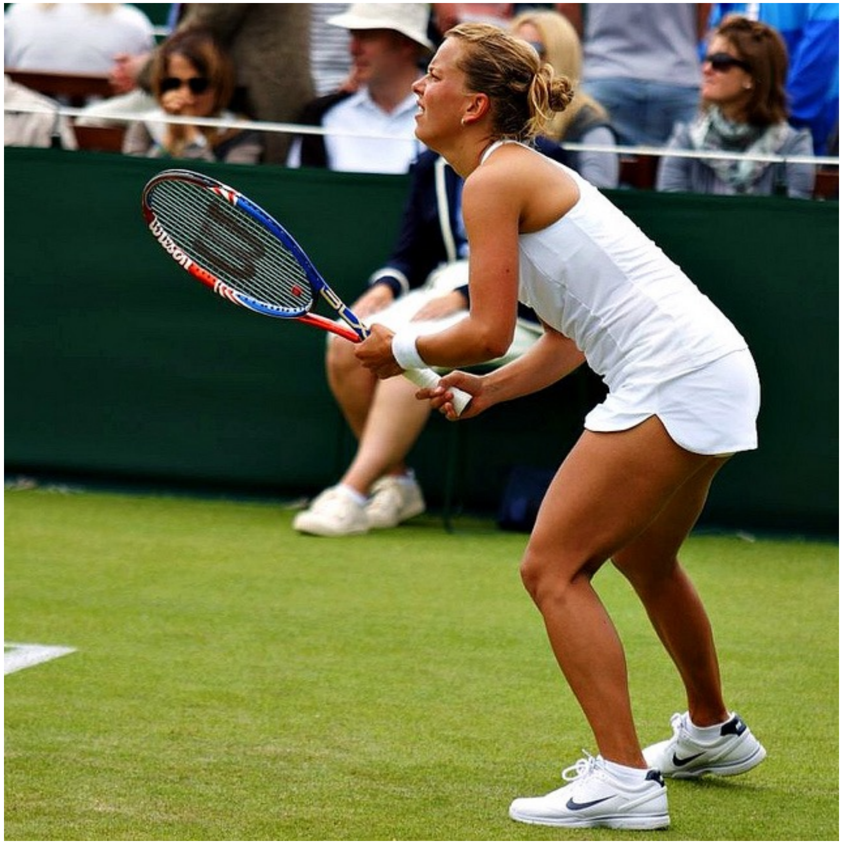


Image from MS-COCO [Lin et al., 2014]

Visual questions target different image regions

Which game is being played?

What type is the surface?

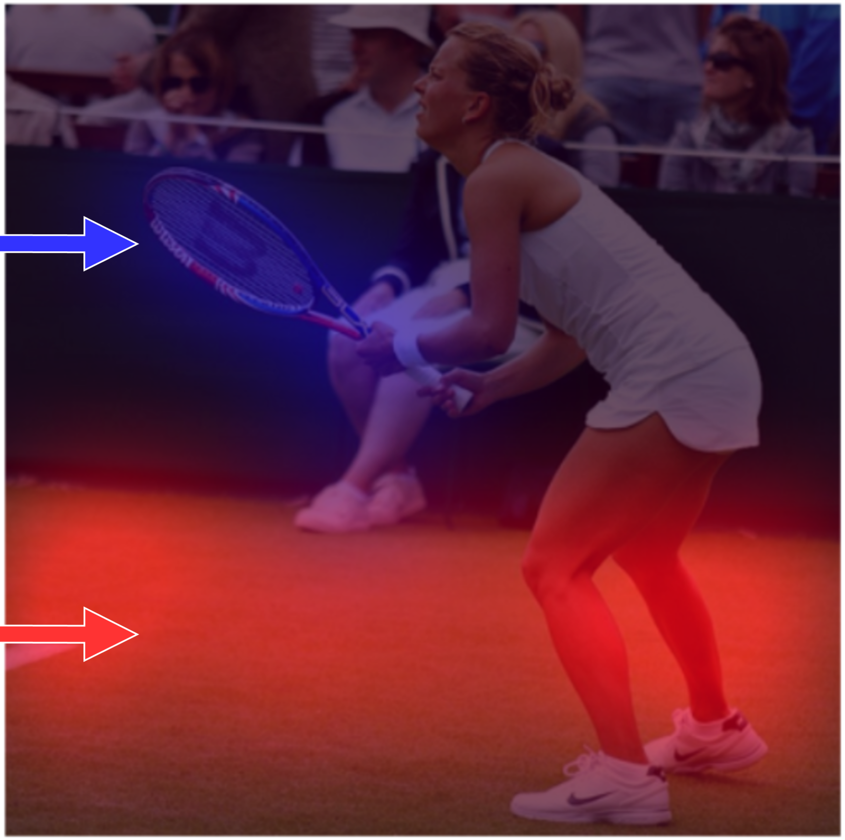
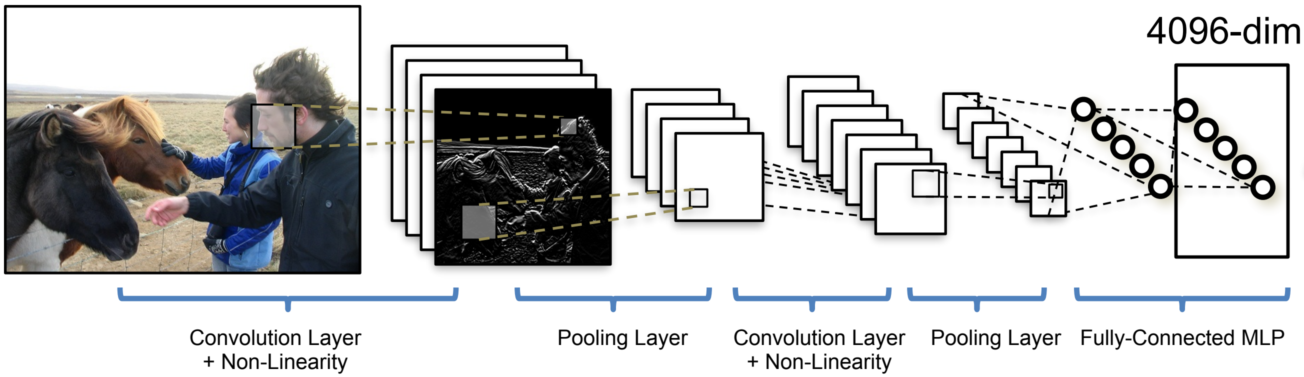


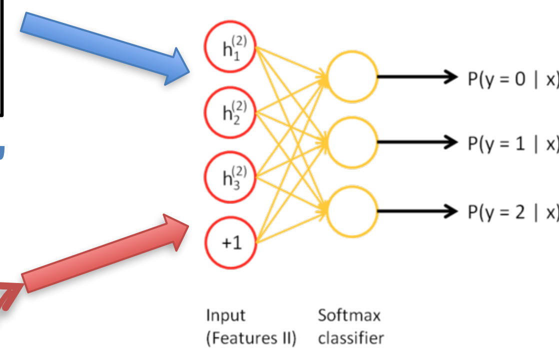
Image from MS-COCO [Lin et al., 2014]

2-Channel VQA Model

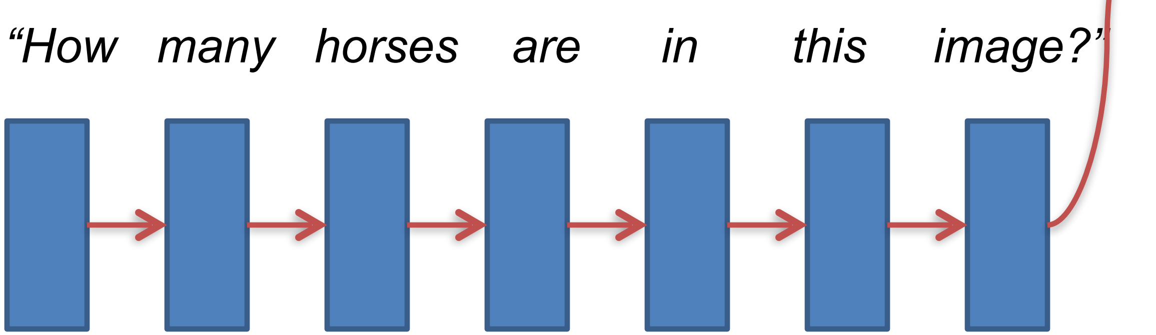
Image Embedding (VGGNet)



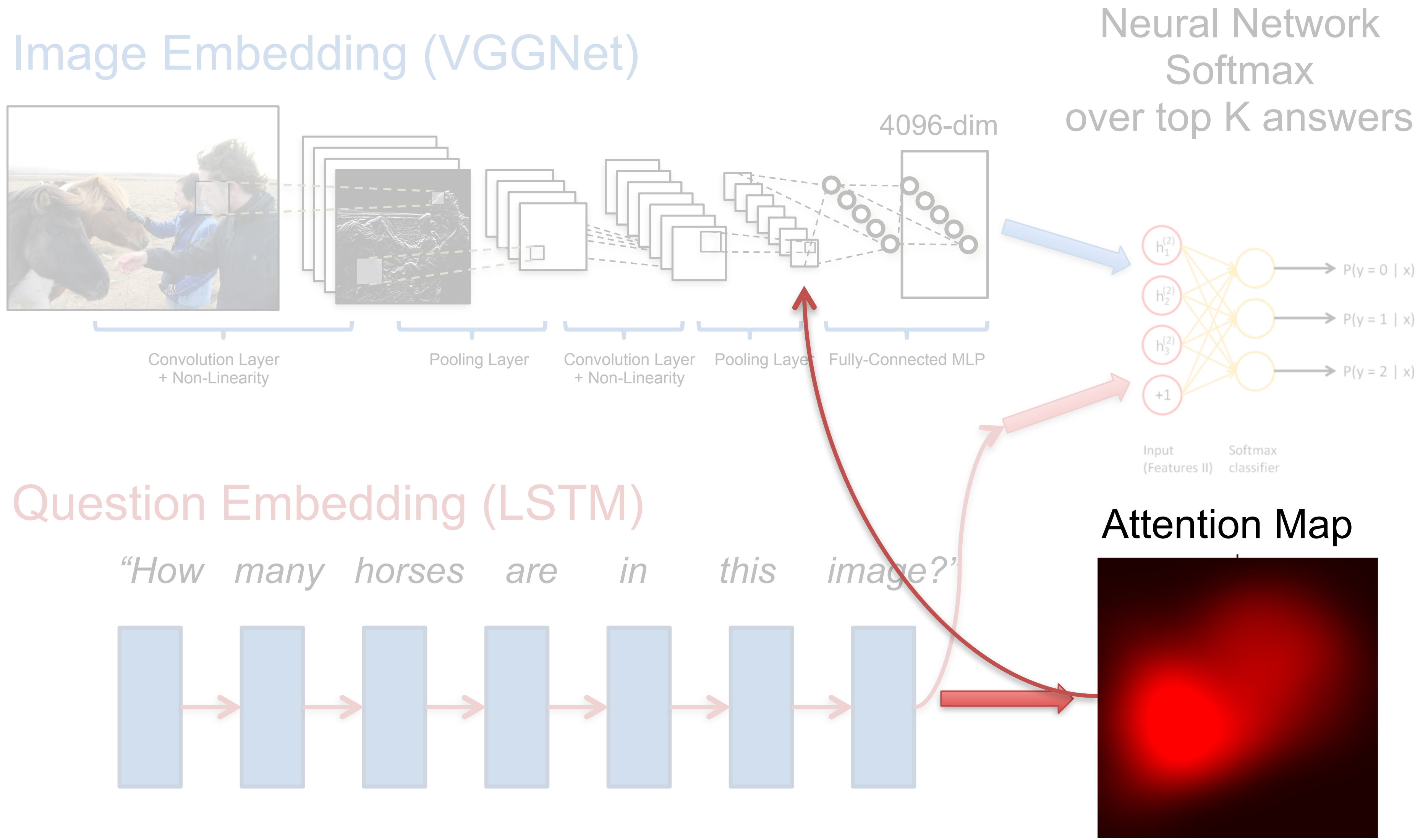
Neural Network
Softmax
over top K answers



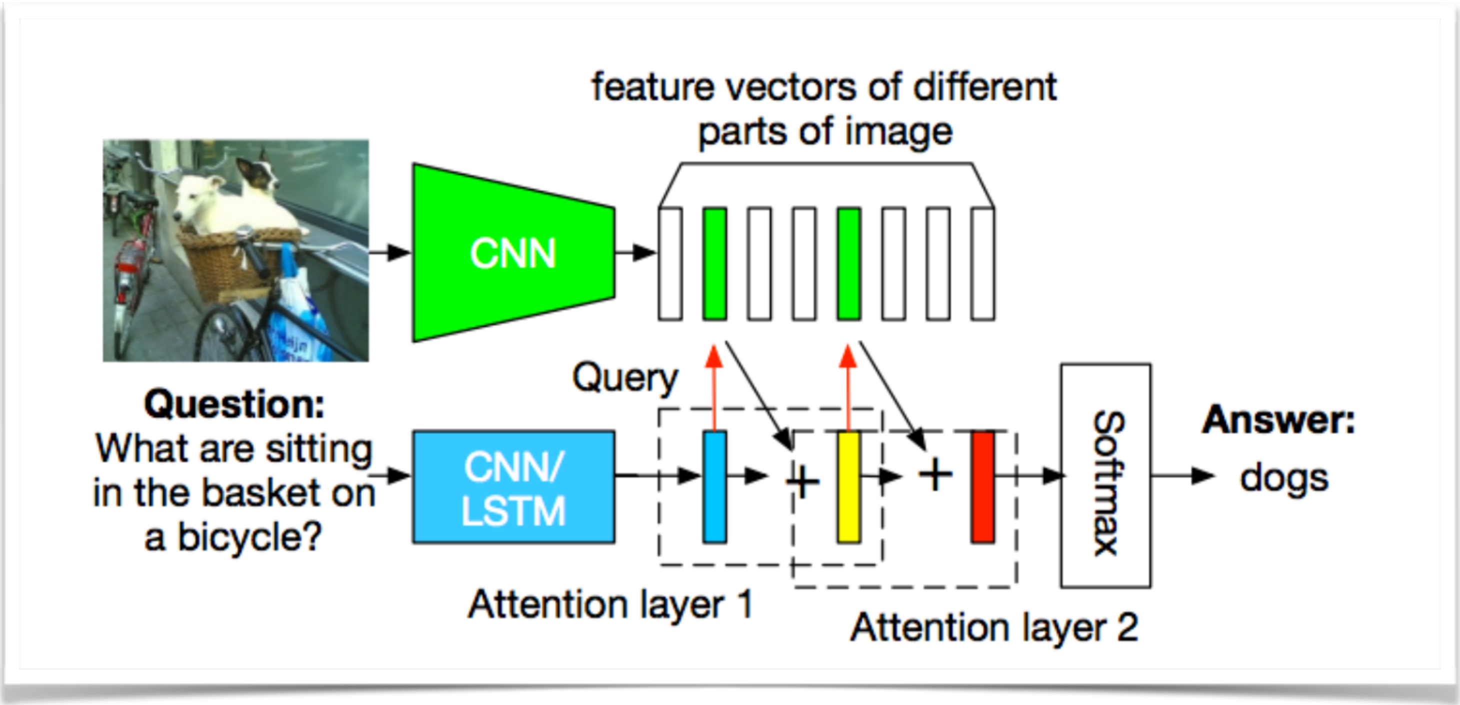
Question Embedding (LSTM)



2-Channel VQA Model

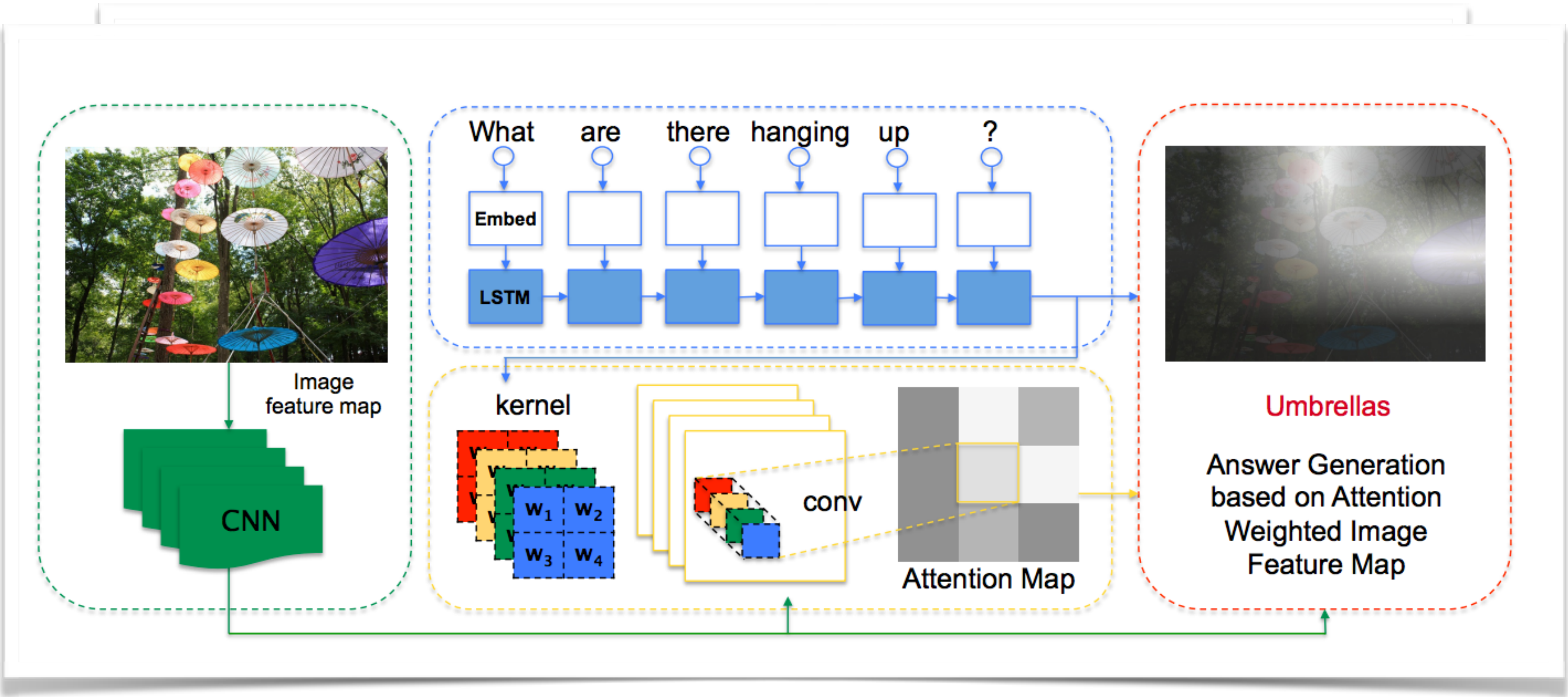


Attention-based VQA models



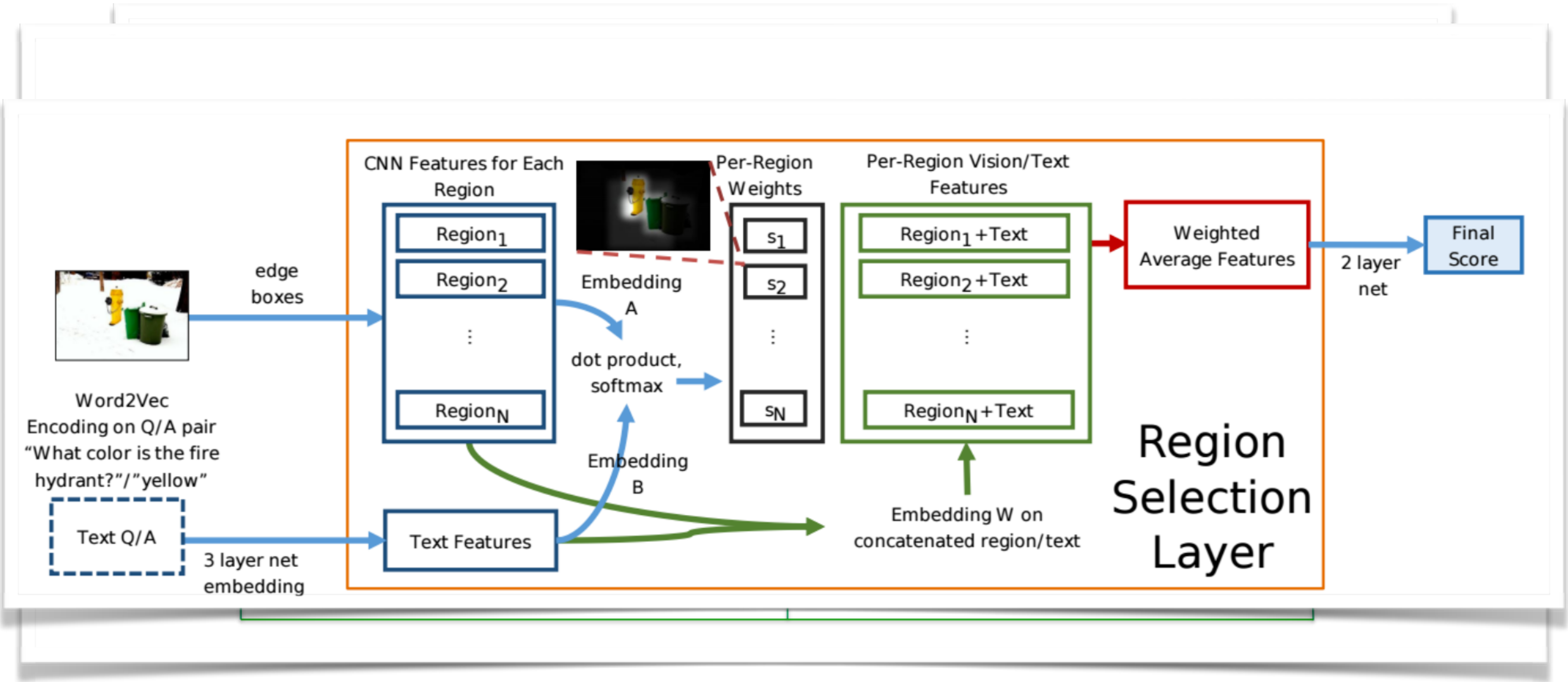
[Yang et al., CVPR16]

Attention-based VQA models



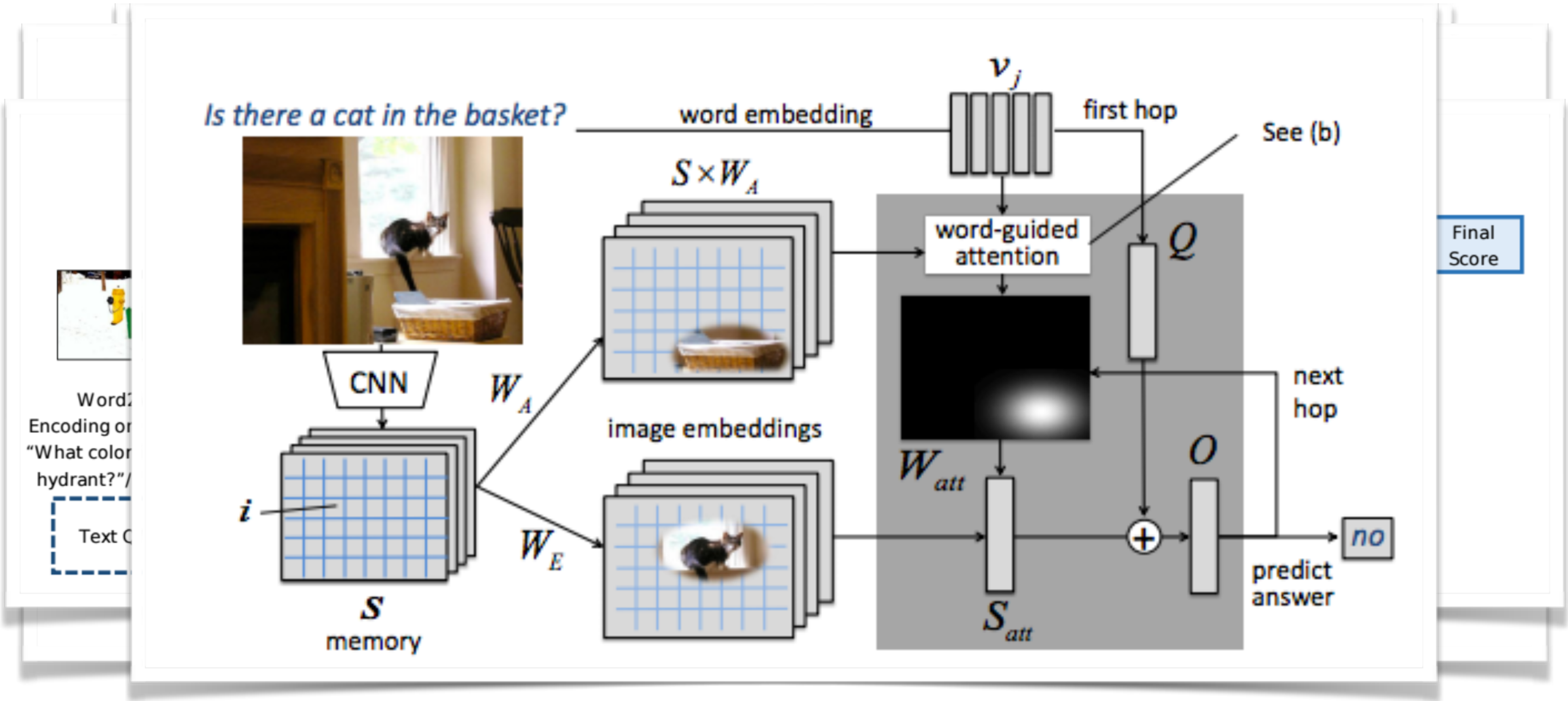
[Chen et al., 2015]

Attention-based VQA models



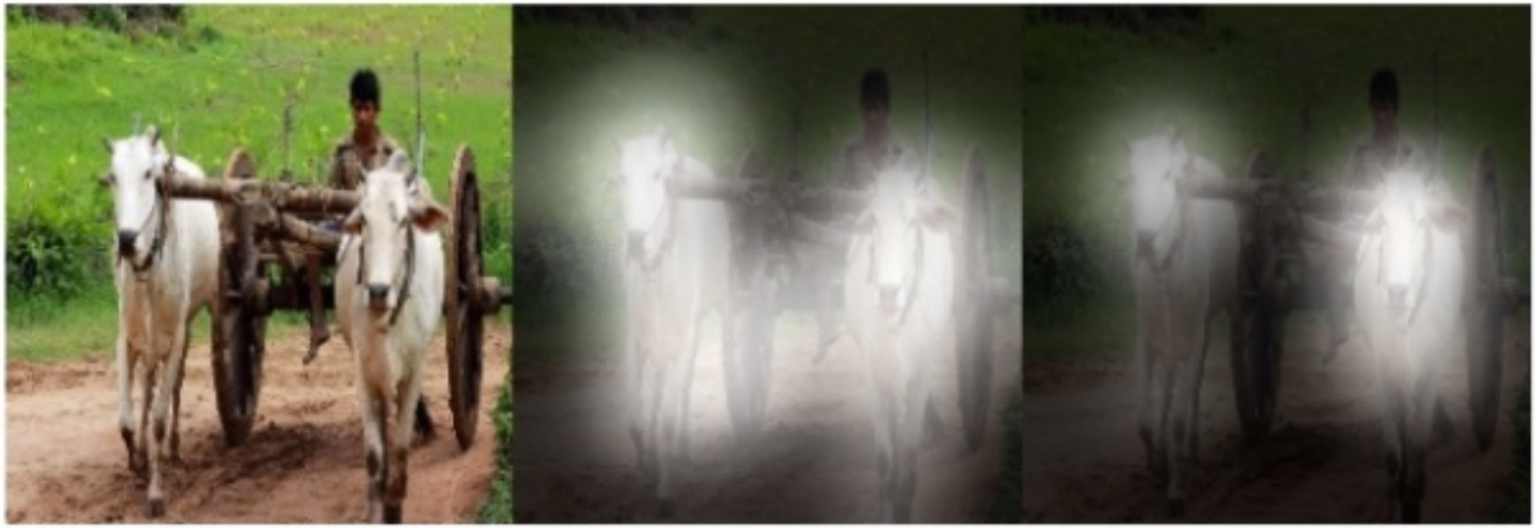
[Shih et al., CVPR16]

Attention-based VQA models



[Xu et al., 2015]

Attention-based VQA models



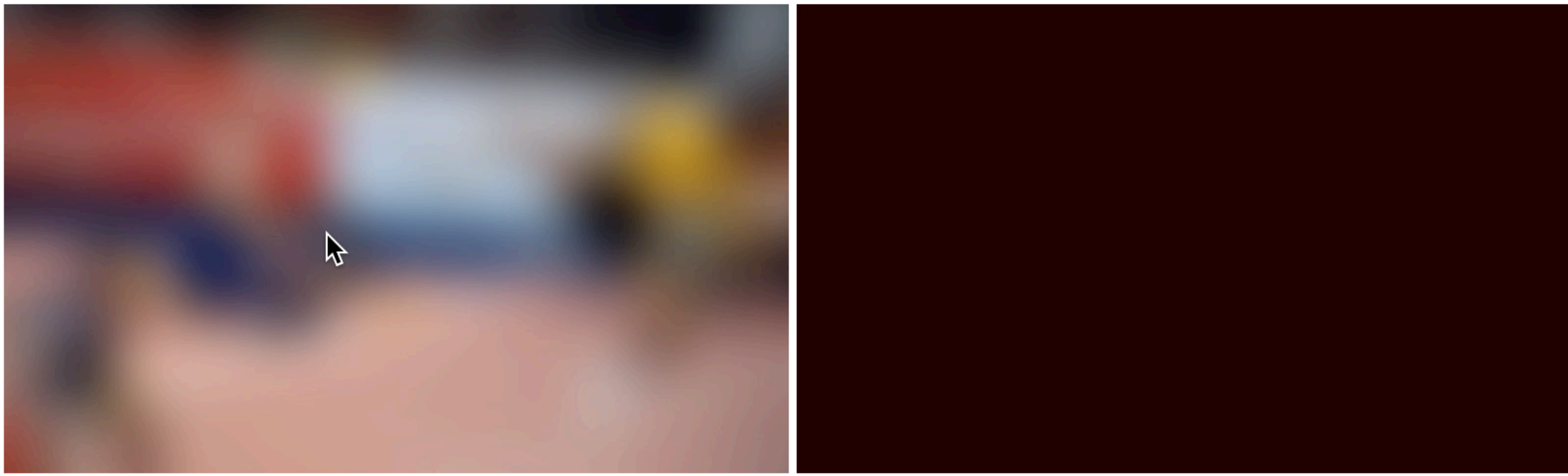
[Yang et al., CVPR16]

- Few qualitative examples
- Nothing to compare with
- No ‘gold standard’

Where do humans choose to look to answer visual questions?

VQA-HAT (Human ATtention)

Question: How many players are visible in the image?

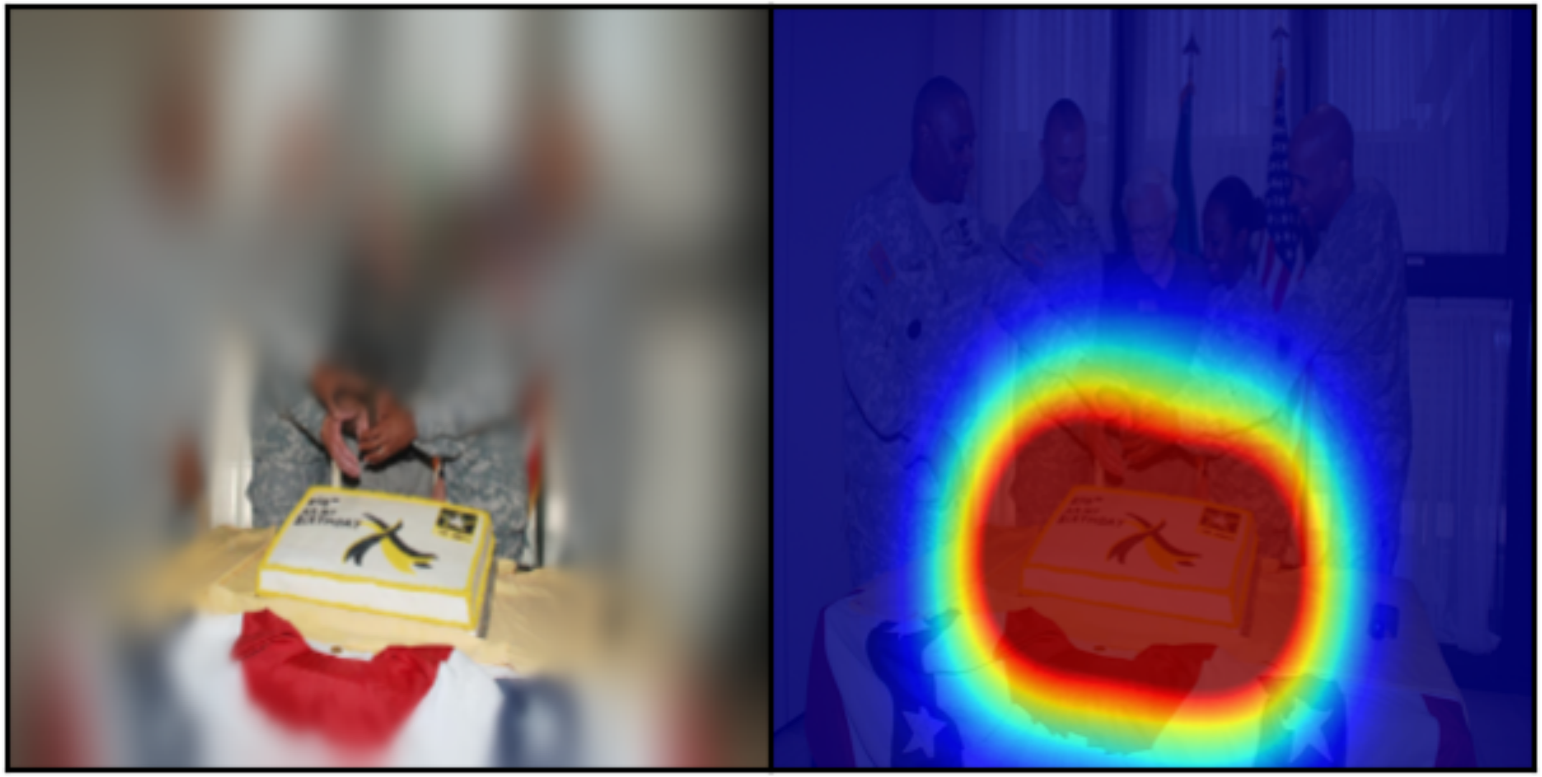


BLUR IMAGE

Answer:

3

VQA-HAT (Human ATtention)

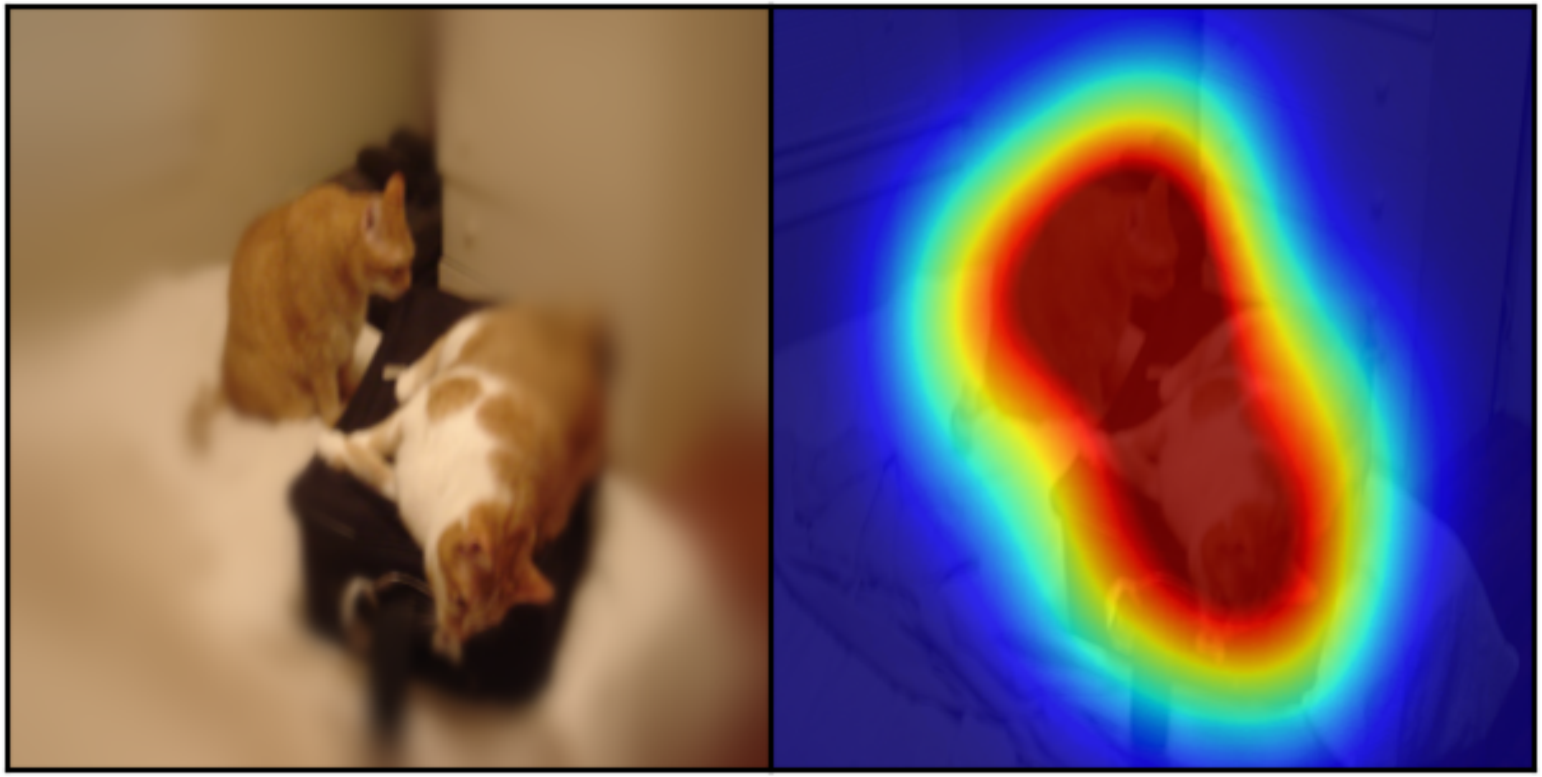


What food is on the table? Cake

VQA-HAT (Human ATtention)

What animal is she riding? Horse

VQA-HAT (Human ATtention)



What number of cats are laying on the bed? 2

VQA-HAT (Human ATtention)

- Attention maps for VQA dataset [Antol et al., ICCV15]
- 58k training I-Q pairs
- 1.4k validation I-Q pairs (3 maps per question)

VQA-HAT (Human ATtention)

Is it sunny? Is it storming outside? Is there a plane in this photo? Are there trees in the background?

What is the color of the train on the left? What is next to the house?

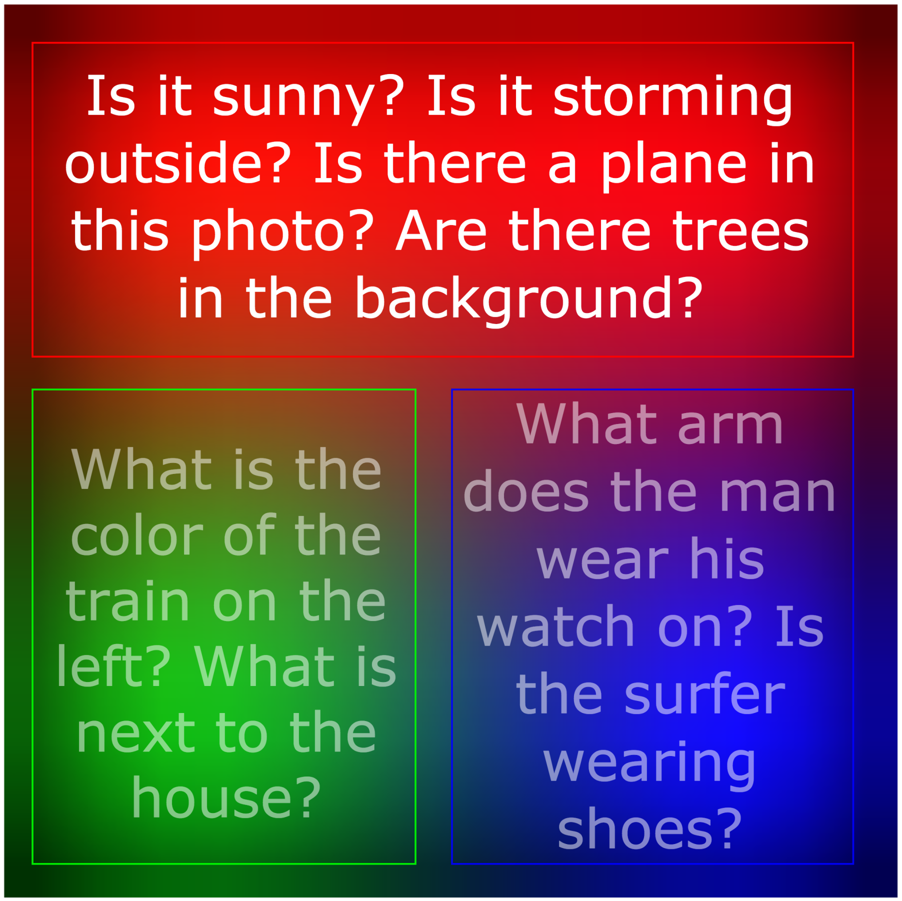
What arm does the man wear his watch on? Is the surfer wearing shoes?

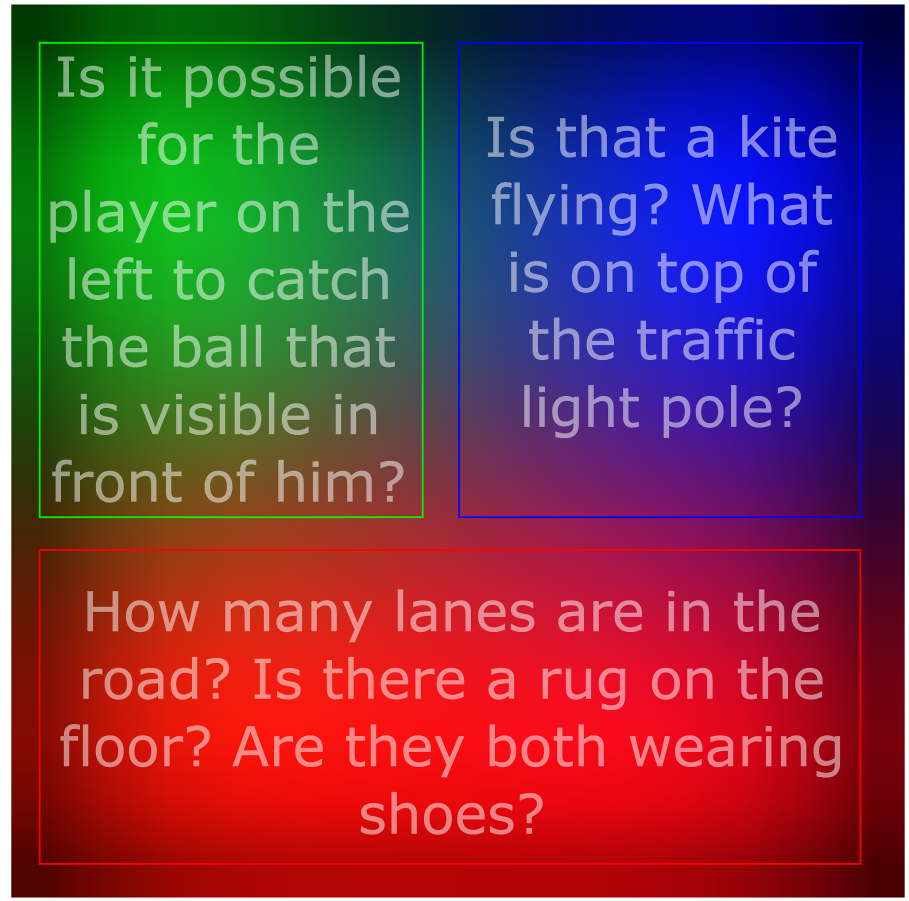
Is it possible for the player on the left to catch the ball that is visible in front of him?

Is that a kite flying? What is on top of the traffic light pole?

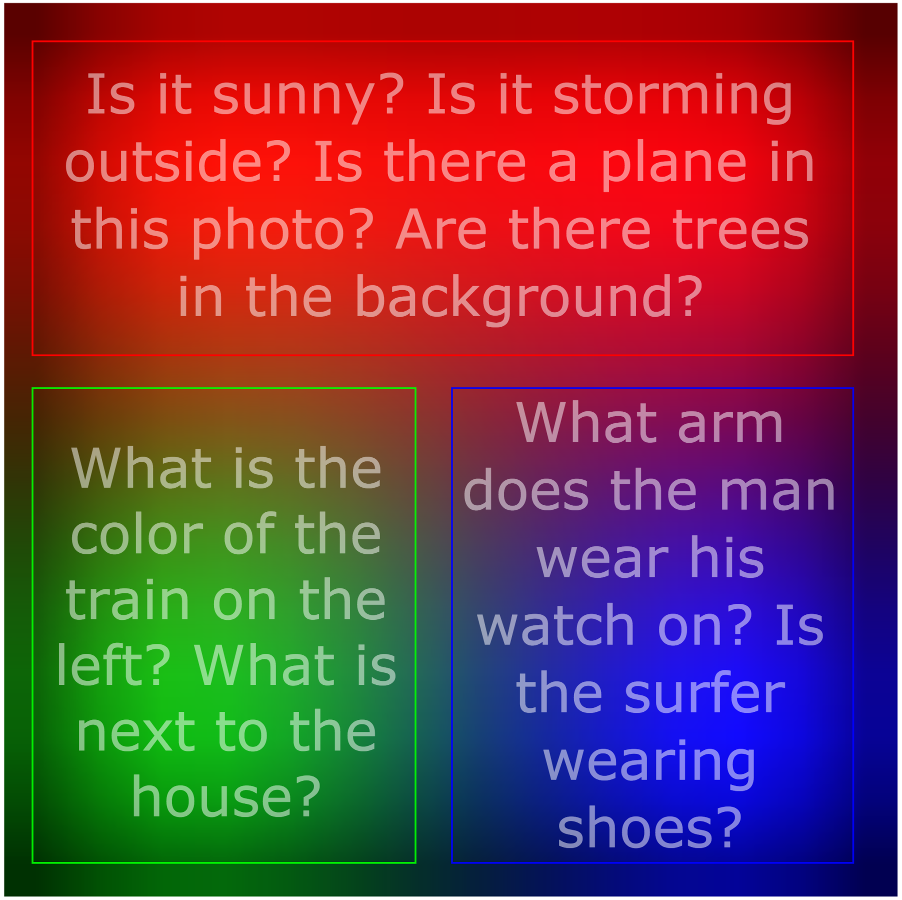
How many lanes are in the road? Is there a rug on the floor? Are they both wearing shoes?

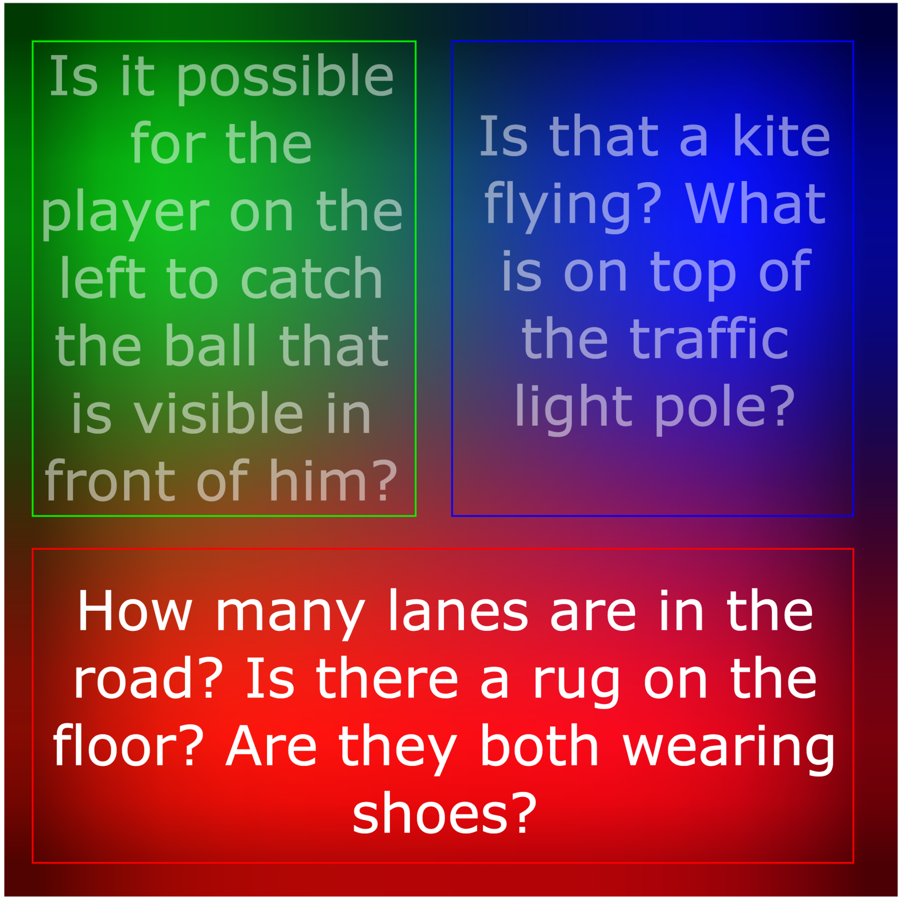
VQA-HAT (Human ATtention)





VQA-HAT (Human ATtention)





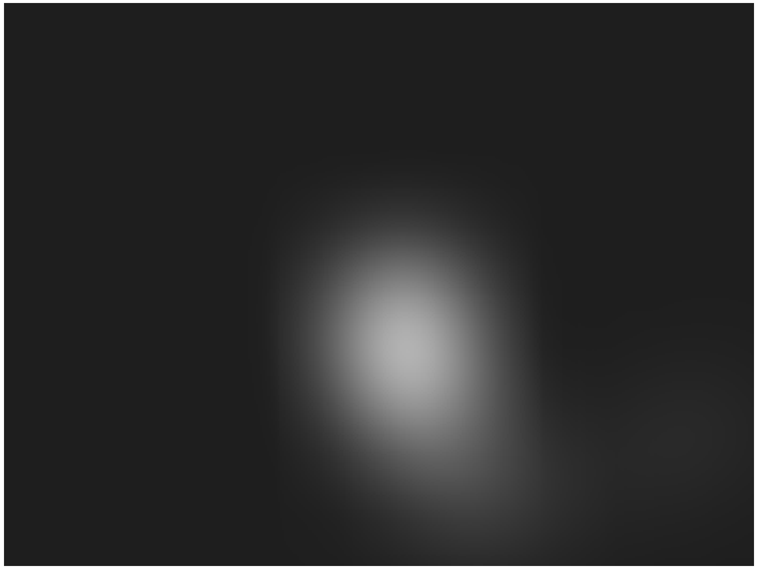
Do deep neural networks ‘look at’ the same regions as humans?

Evaluating attention maps

- Metric: Rank correlation

Evaluating attention maps

- Metric: Rank correlation

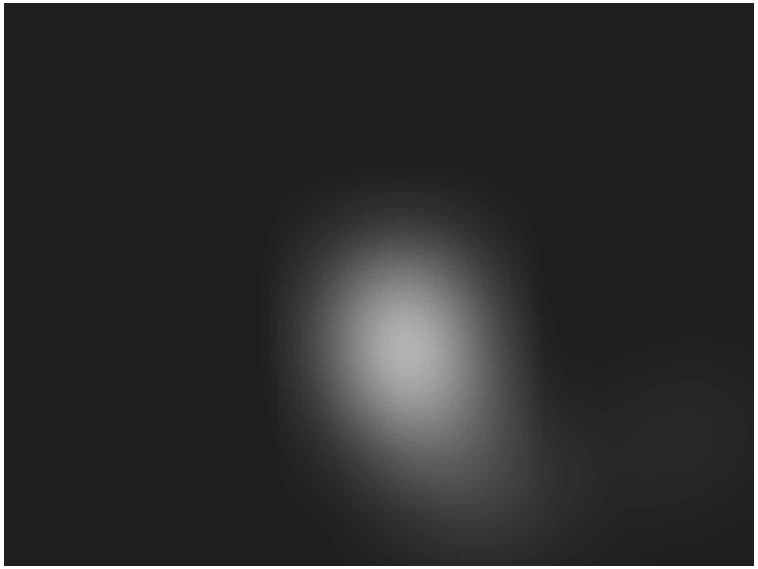


Attention map 1

Attention map 2

Evaluating attention maps

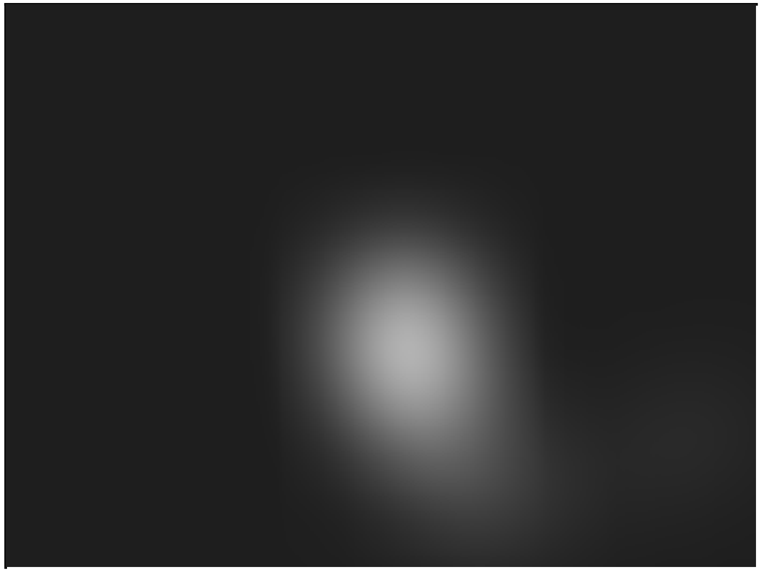
- Metric: Rank correlation



Scale to same size

Evaluating attention maps

- Metric: Rank correlation




Compute pixel intensities

Evaluating attention maps

- Metric: Rank correlation

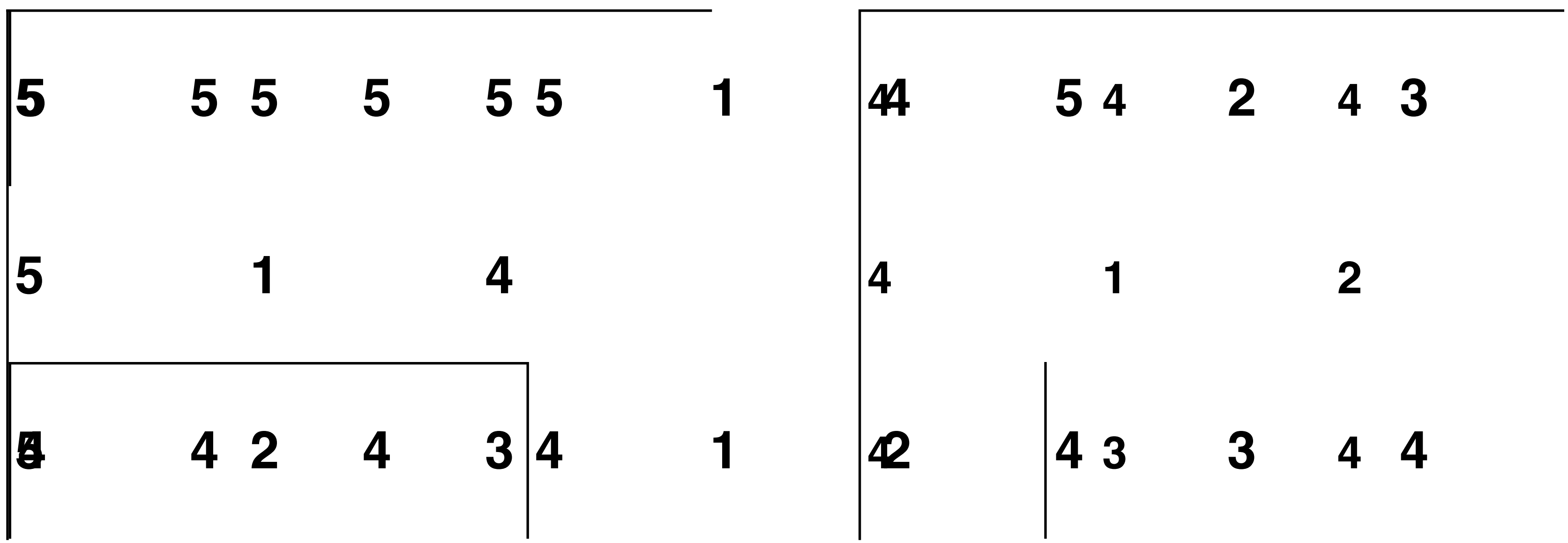
0	0	0
0	200	400
0	250	320

0	0	0
0	240	270
0	340	0

Rank the pixel intensities

Evaluating attention maps

- Metric: Rank correlation



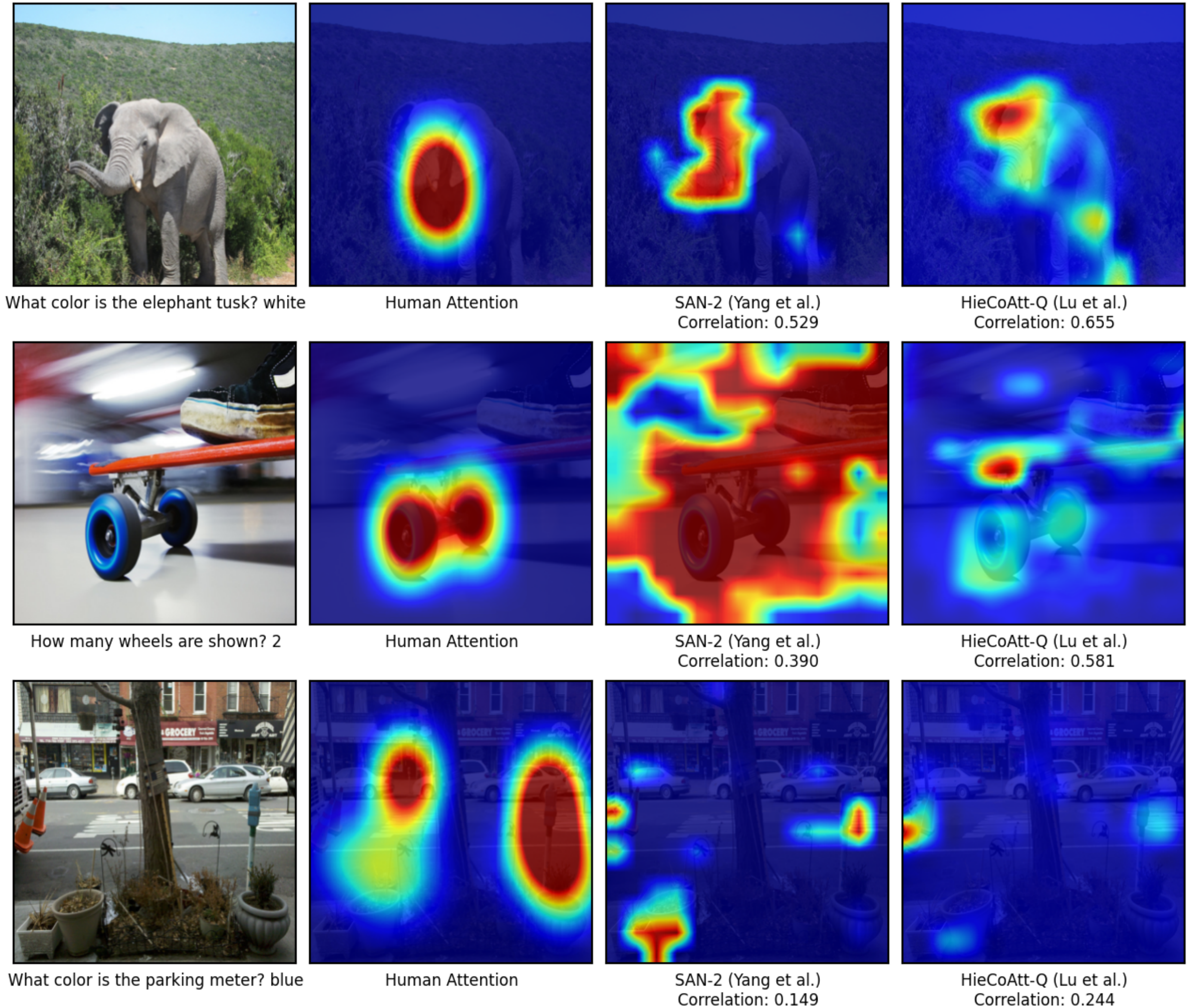
Flatten them into vectors

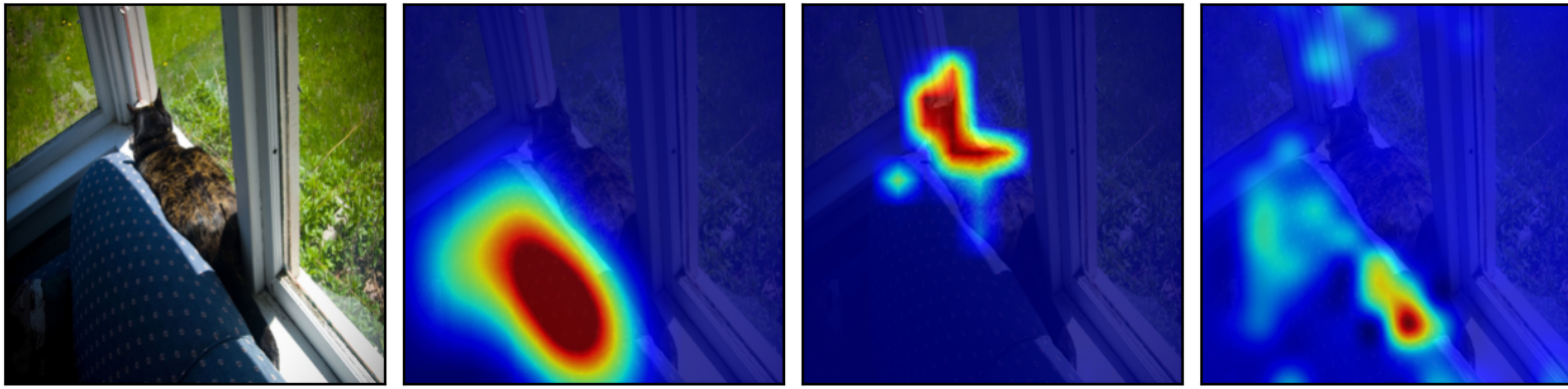
Evaluating attention maps

- Metric: Rank correlation

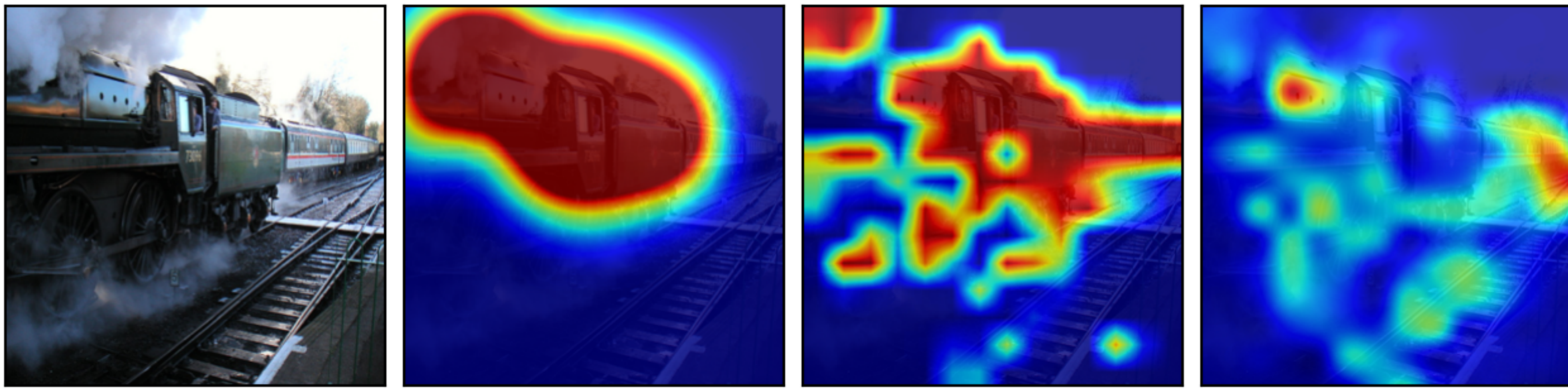
5	5	5	5	1	4	5	2	3
4	4	4	4	1	2	4	3	4

Compute correlation between the ranked lists

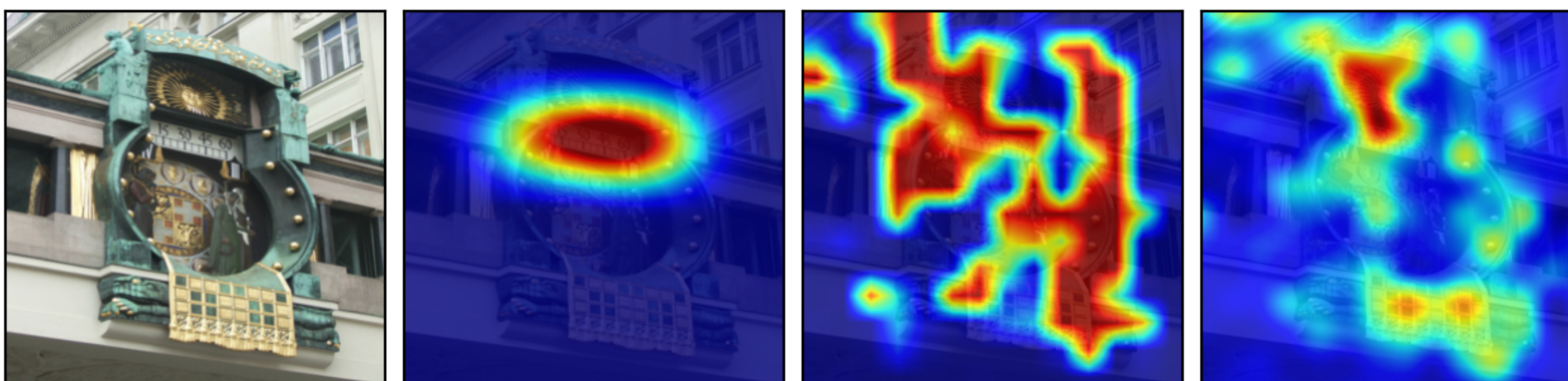

What color is the elephant tusk? white
Human Attention
SAN-2 (Yang et al.)
Correlation: 0.529
HieCoAtt-Q (Lu et al.)
Correlation: 0.655
How many wheels are shown? 2
Human Attention
SAN-2 (Yang et al.)
Correlation: 0.390
HieCoAtt-Q (Lu et al.)
Correlation: 0.581
What color is the parking meter? blue
Human Attention
SAN-2 (Yang et al.)
Correlation: 0.149
HieCoAtt-Q (Lu et al.)
Correlation: 0.244



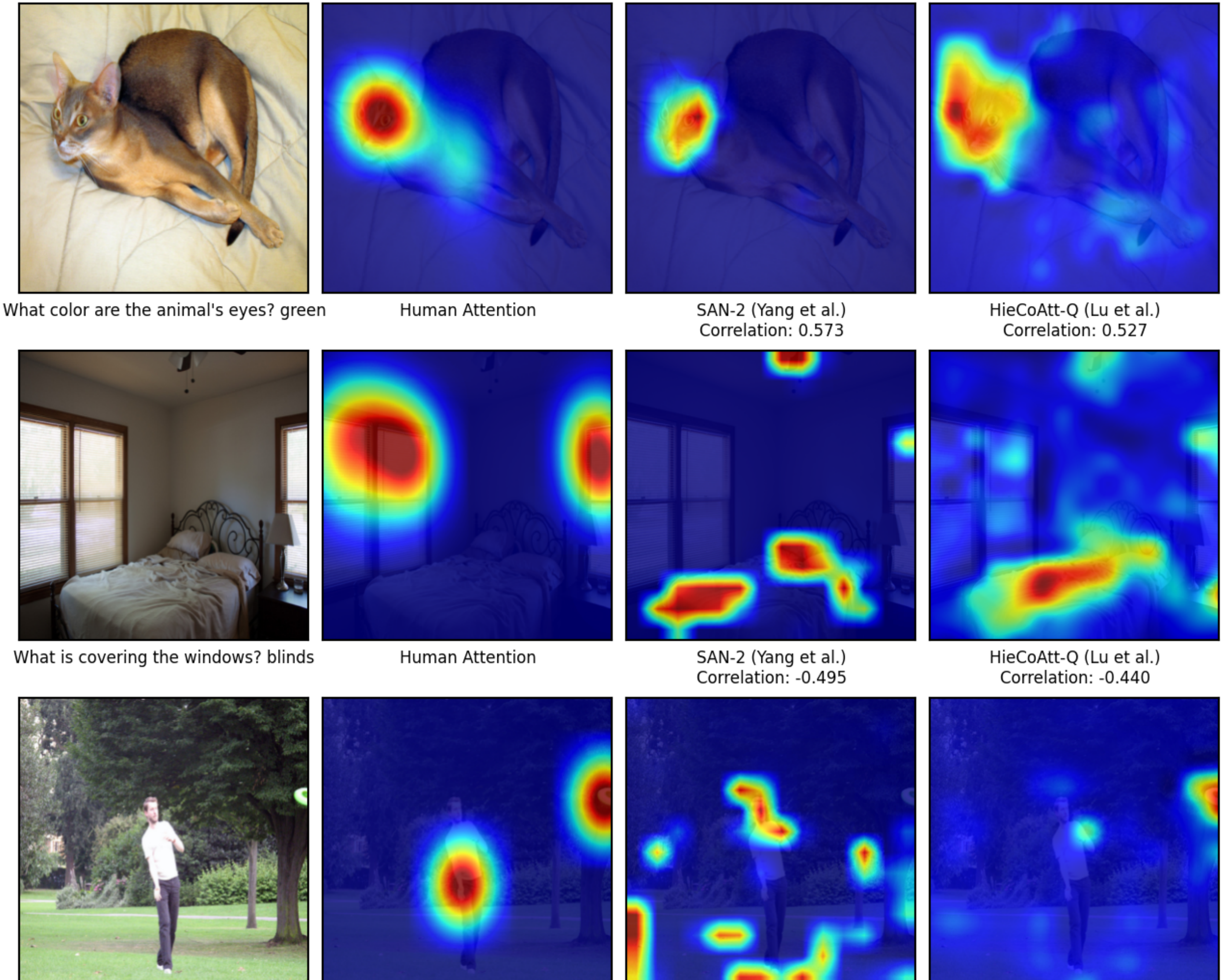
What color is the couch? blue | Human Attention | SAN-2 (Yang et al.) Correlation: 0.519 | HieCoAtt-Q (Lu et al.) Correlation: 0.538



What kind of engine is this? steam | Human Attention | SAN-2 (Yang et al.) Correlation: 0.468 | HieCoAtt-Q (Lu et al.) Correlation: 0.456



What is the highest number shown? 60 | Human Attention | SAN-2 (Yang et al.) Correlation: 0.460 | HieCoAtt-Q (Lu et al.) Correlation: 0.359



What color are the animal's eyes? green | Human Attention | SAN-2 (Yang et al.) Correlation: 0.573 | HieCoAtt-Q (Lu et al.) Correlation: 0.527

What is covering the windows? blinds | Human Attention | SAN-2 (Yang et al.) Correlation: -0.495 | HieCoAtt-Q (Lu et al.) Correlation: -0.440

What is the man doing? playing frisbee | Human Attention | SAN-2 (Yang et al.) Correlation: -0.060 | HieCoAtt-Q (Lu et al.) Correlation: 0.238

Evaluating attention maps

VQA Model	Mean Rank-Correlation
Human	0.62

Evaluating attention maps

VQA Model	Mean Rank-Correlation
Human	0.62
Implicit Attention: Occlusion	0.07
Implicit Attention: Guided Backprop [Springenberg et al., ICLR15]	0.12

Evaluating attention maps

VQA Model	Mean Rank-Correlation
Human	0.62
Implicit Attention: Occlusion	0.07
Implicit Attention: Guided Backprop [Springenberg et al., ICLR15]	0.12
Explicit Attention: SAN-2 [Yang et al., CVPR16]	0.25
Explicit Attention: HieCoAtt [Lu et al., 2016]	0.26

Evaluating attention maps

VQA Model	Mean Rank-Correlation
Human	0.62
Implicit Attention: Occlusion	0.07
Implicit Attention: Guided Backprop [Springenberg et al., ICLR15]	0.12
Explicit Attention: SAN-2 [Yang et al., CVPR16]	0.25
Explicit Attention: HieCoAtt [Lu et al., 2016]	0.26
Task-independent Saliency [Judd et al., ICCV09]	0.50

Task-specific

Slide Credit: Abhishek Das

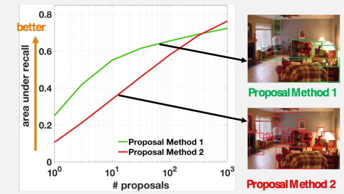
Evaluating attention maps

Task-specific

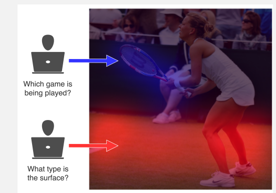
VQA Model	Mean Rank-Correlation: complete set	Mean Rank-Correlation: without center bias
Human	0.62	0.73
Implicit Attention: Occlusion	0.07	0.02
Implicit Attention: Guided Backprop [Springenberg et al., ICLR15]	0.12	-0.03
Explicit Attention: SAN-2 [Yang et al., CVPR16]	0.25	0.06
Explicit Attention: HieCoAtt [Lu et al., 2016]	0.26	0.12
Task-independent Saliency [Judd et al., ICCV09]	0.50	-0.07

Outline

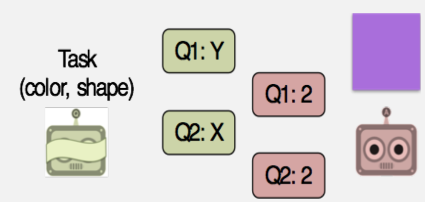
Object-Proposal Evaluation
Protocol is 'Gameable' [CVPR '16]



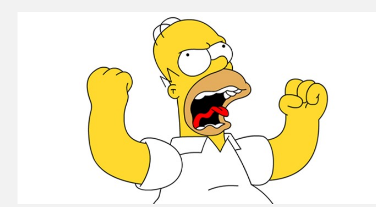
Human Attention in VQA:
Do Humans and Deep Networks
Look at the Same Regions? [EMNLP '16]



Natural Language Does Not Emerge
'Naturally' in Multi-Agent Dialog [EMNLP '17]



Rant about Vision vs NLP reviewing!



Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

[ICCV '17]


Abhishek Das*
(Georgia Tech)

Satwik Kottur*
(CMU)



José Moura
(CMU)

Stefan Lee
(Virginia Tech)

Dhruv Batra
(Georgia Tech / FAIR)

Visual Dialog: Task

- Given
  - Image I
  - History of human dialog
    $(Q_1, A_1), (Q_2, A_2), \ldots, (Q_{t-1}, A_{t-1})$
  - Follow-up Question Q_t

- Task
  - Produce free-form natural language answer A_t

Visual Dialog

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.

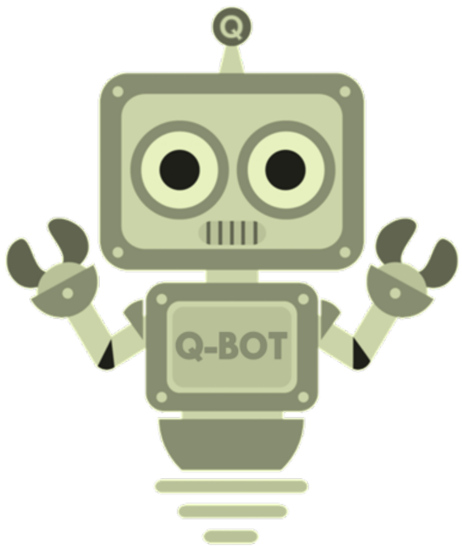
Problems

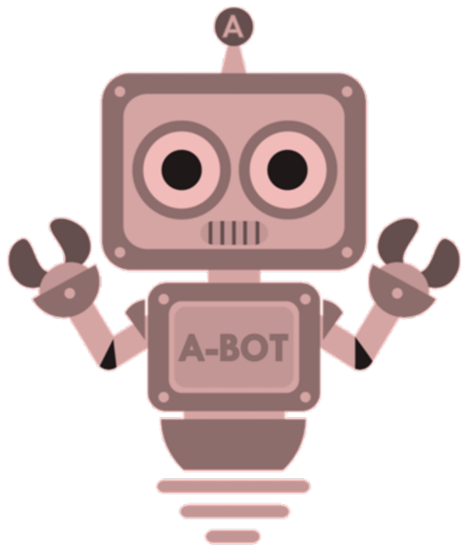
- No goal
  - Why are we talking?

- Agent not in control
  - Artificially injected at every round into a human conversation
  - Can't *steer* the conversation
  - Doesn't get to see its errors during training

- Can't learn the *meaning* of responses
  - Many equivalent responses that should be treated equally, but aren't!
  - Is log-likelihood of human response really a good metric?

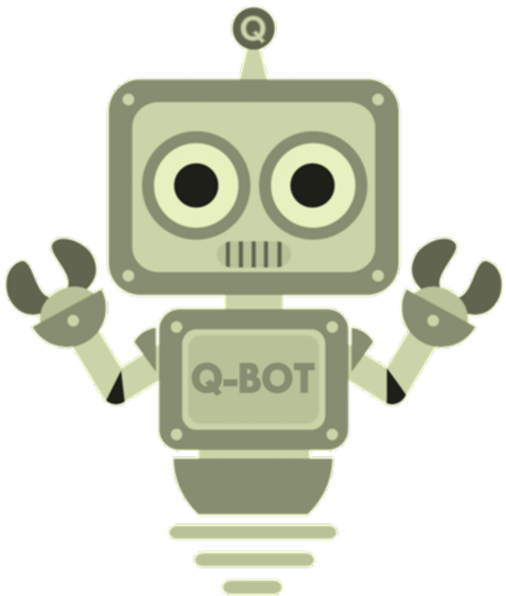
Image Guessing Game

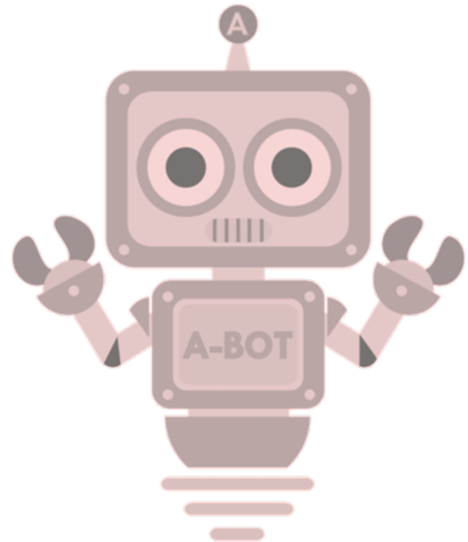




Slide Credit: Abhishek Das

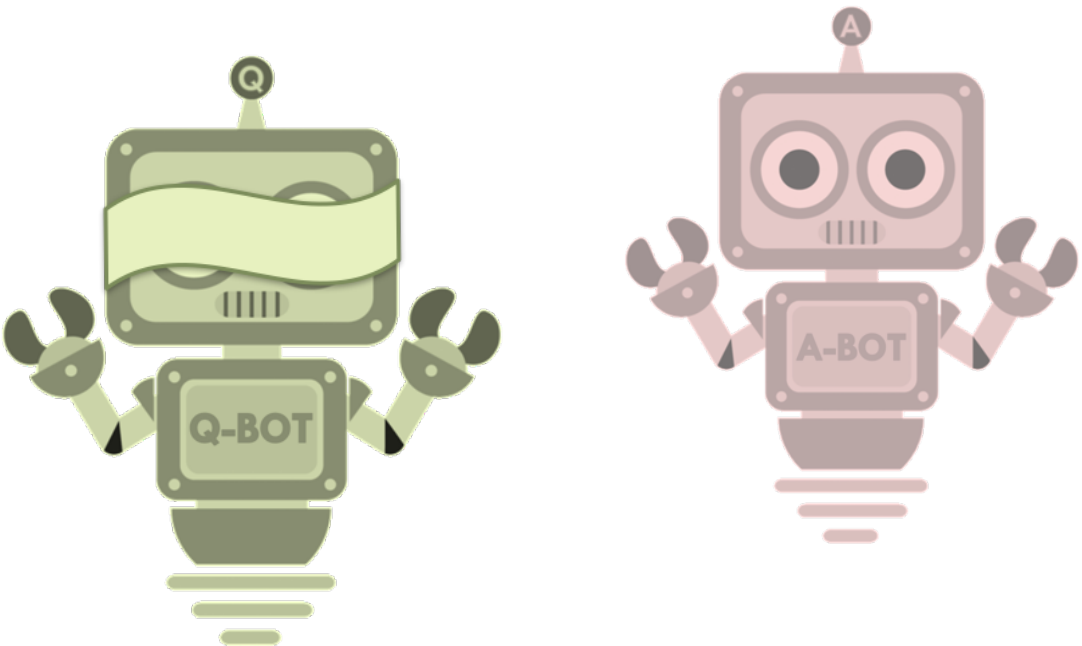
Image Guessing Game





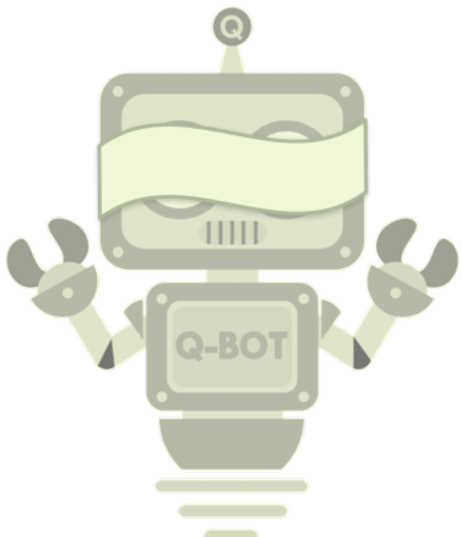
Q-Bot asks questions

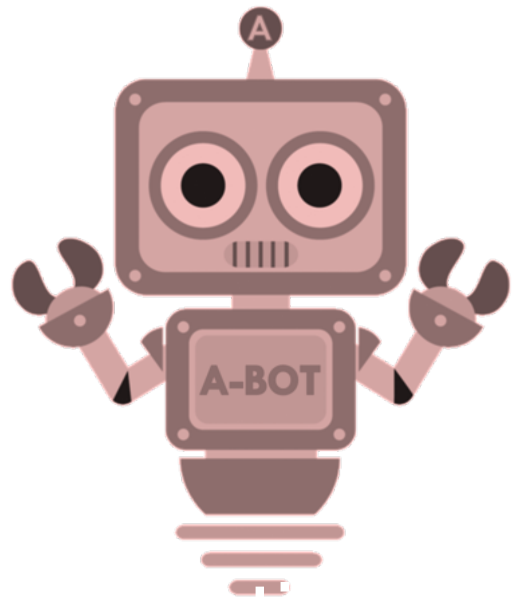
Image Guessing Game



Q-Bot is blindfolded

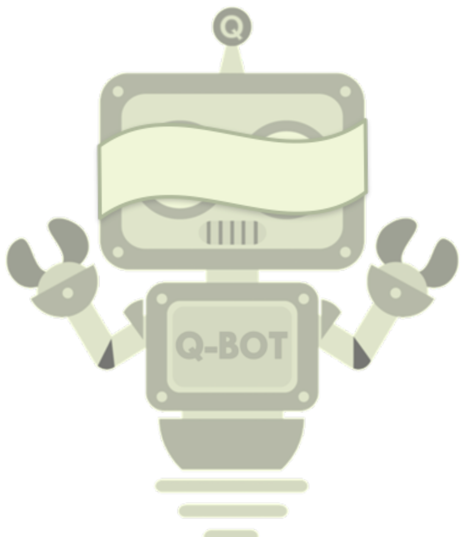
Image Guessing Game

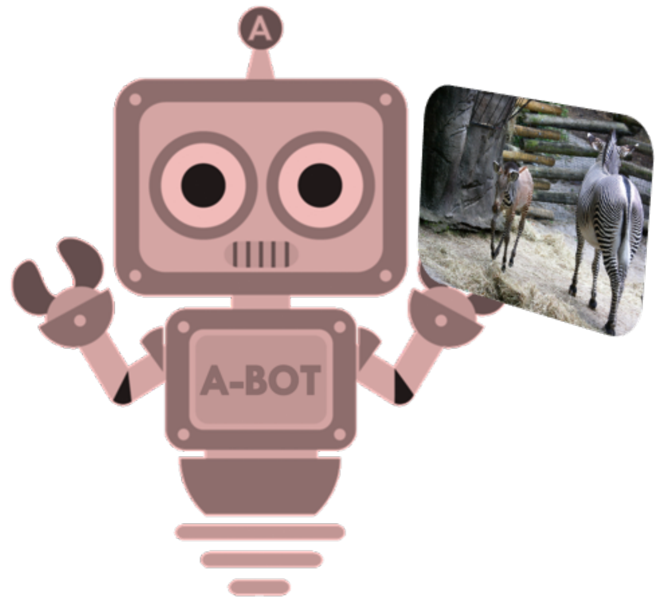




A-Bot answers questions

Image Guessing Game





A-Bot sees an image

Image Guessing Game

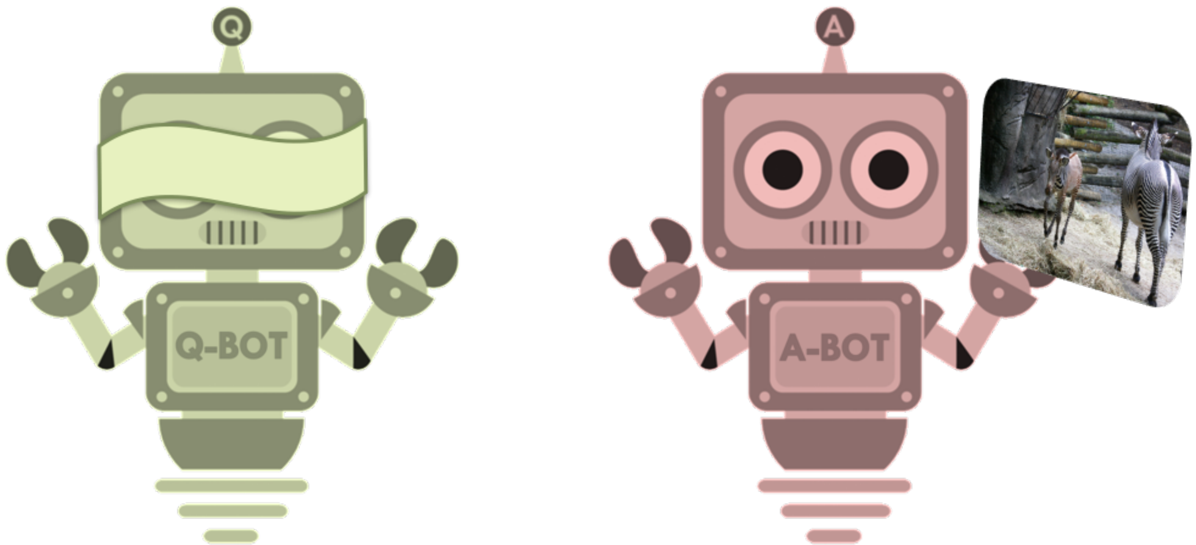


Image Guessing Game

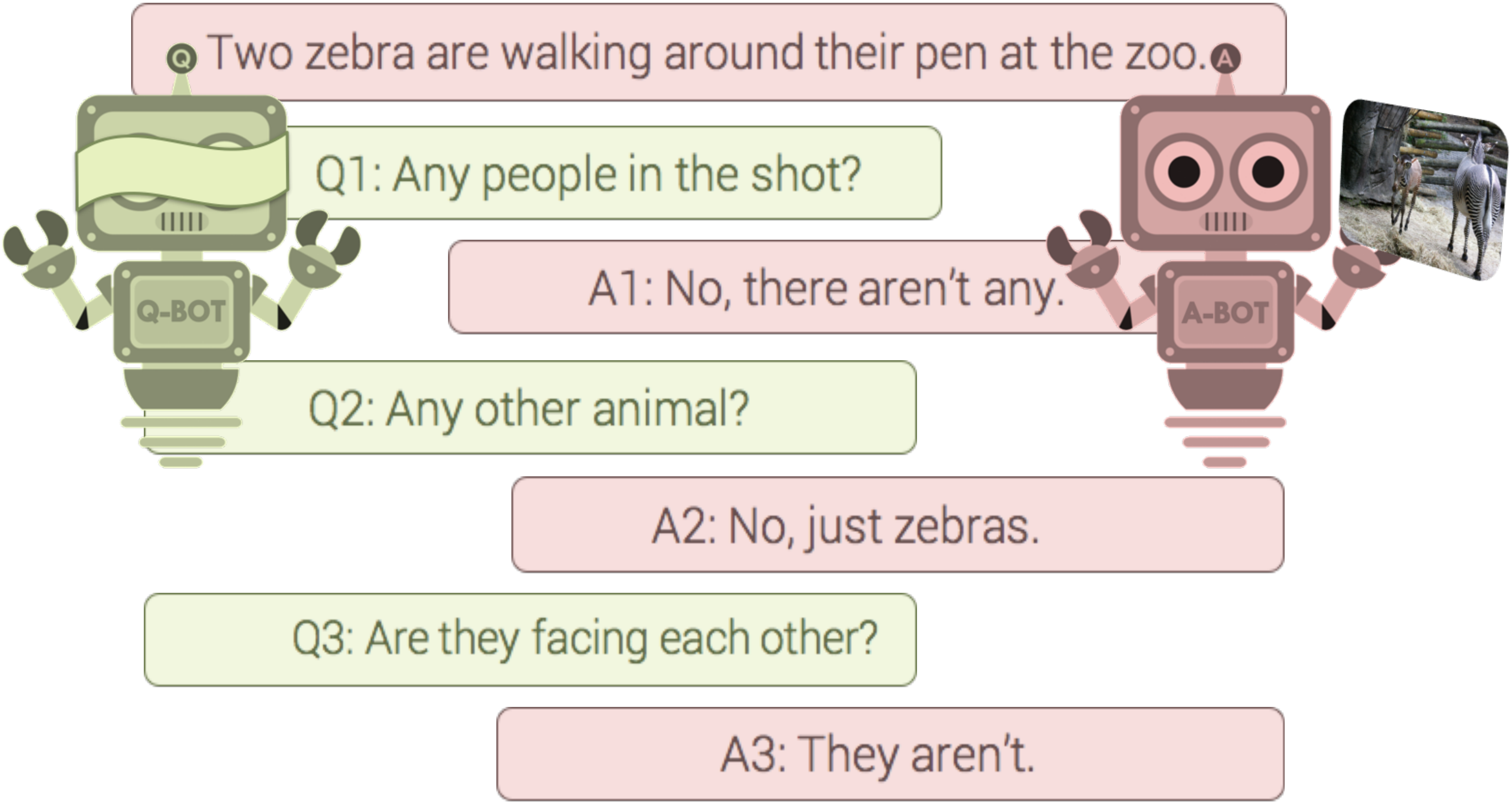
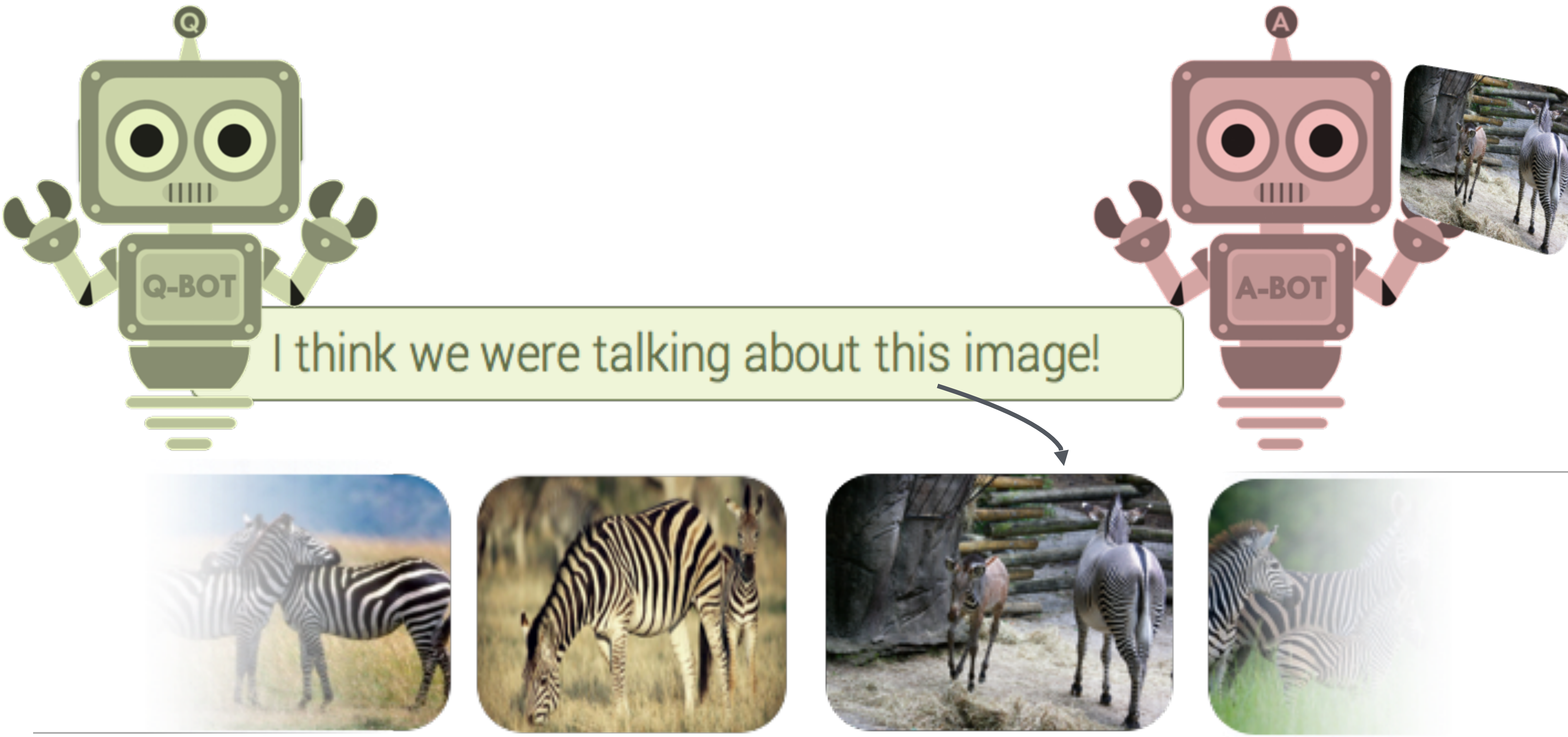
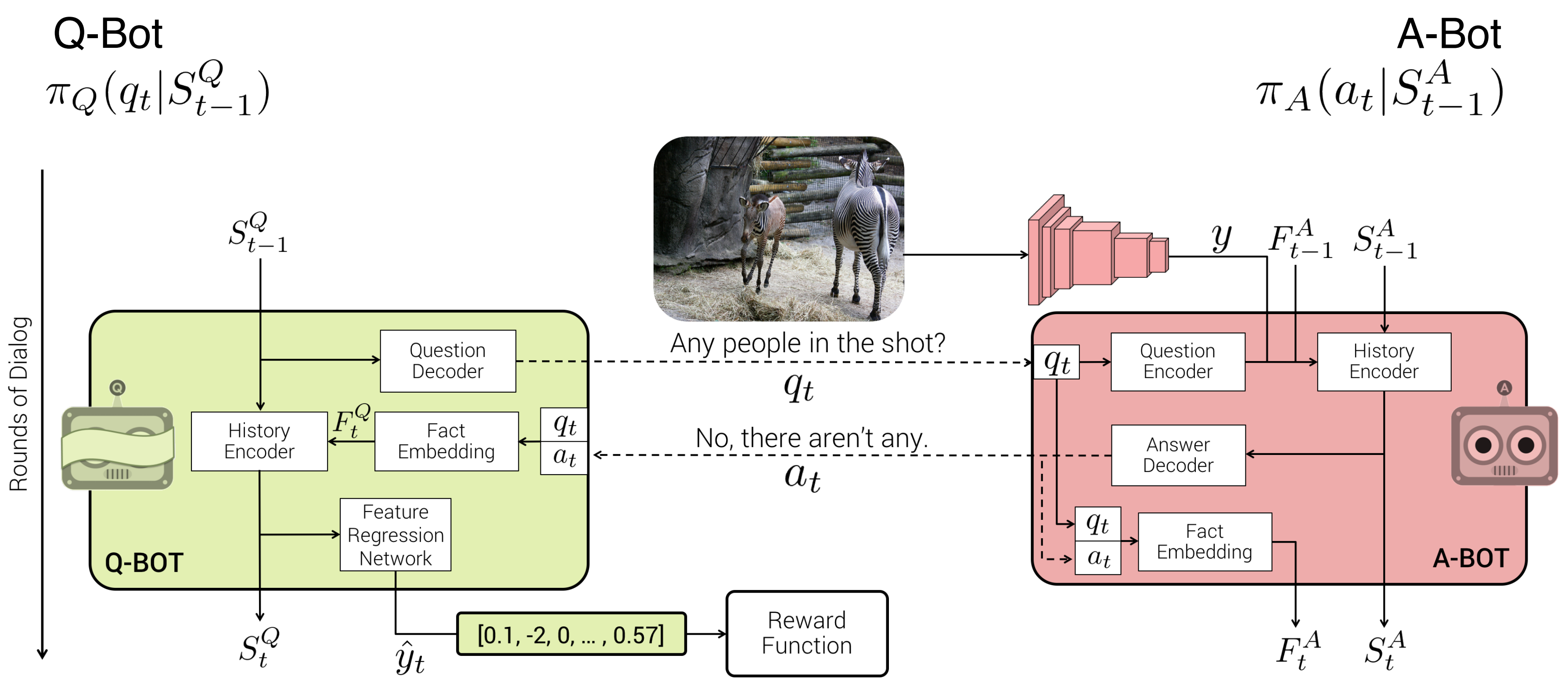


Image Guessing Game

A3: They aren't.



Policy Networks



Slide Credit: Abhishek Das

Turing Test

How old is this man?	He looks like late teens
Does he have short hair or long hair?	Medium
What color is his hair?	Brown
Is he wearing glasses?	No
Does he have facial hair?	No
How is he dressed?	He's white dress shirt with red tie . . . school uniform
Is he larger man or is he skinny?	He's skinny
Is this picture taken outside?	Inside
Can you tell if it's daytime?	I'm not sure
Can you tell what room he is standing in?	He's in hallway

A man making the live long and prosper sign from star trek.

Where is man located?	Looks like classroom of some sort
What is man wearing?	Black t shirt and jeans
Are there any other people in picture?	Yes, quite few people in background
What color is table?	Picture is black and white, but it's wood table
Is there anything else on table?	Not that I can see
What are other people doing?	They are standing in front of him, but I don't . . .
Are there any windows?	Not that I can see
What color are walls?	I can't see walls
What time of day do you think it is?	I can't tell
What is man wearing?	Black t shirt and jeans

How old is man?	I can't see his face, I don't know
What is he wearing?	shorts and t shirt
What color is his shirt?	white
what color are his pants?	black
what color is his shirt?	blue
What color are his shoes?	Black
What color are his shoes?	Black
What color are his shoes?	Black
What color are his shoes?	Black
What color are his shoes?	Black

What color is skateboard?	It is hard to tell but I think it's black
Is he wearing helmet?	Yes
Is he wearing knee pads?	No, no knee pads
Is it day or night?	It is daytime
Is it sunny?	It is bright out, but I see no shadows, so I can not
Can you see any other people?	There are 2 other people in background
Is it day or night?	It is daytime
Is it sunny?	Yes, it is sunny
What color is skateboard?	It is hard to tell but I think it's black
Is he wearing helmet?	Yes, he is wearing helmet

SL Agents

RL Agents

Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog

[EMNLP ‘17]


Satwik Kottur*
(CMU)


José Moura
(CMU)

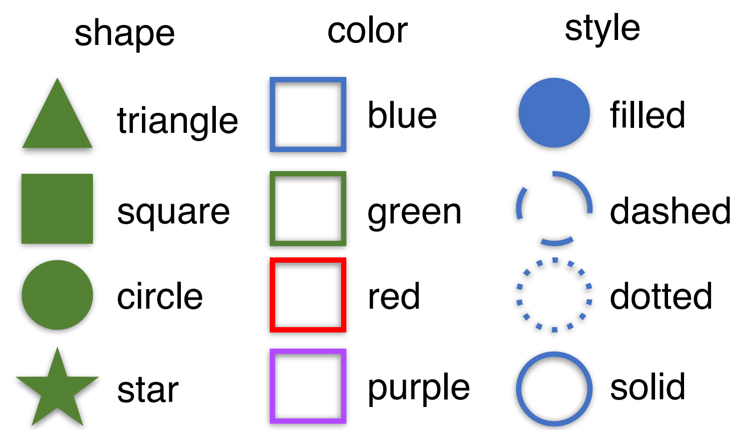

Stefan Lee
(Virginia Tech)


Dhruv Batra
(Georgia Tech / FAIR)

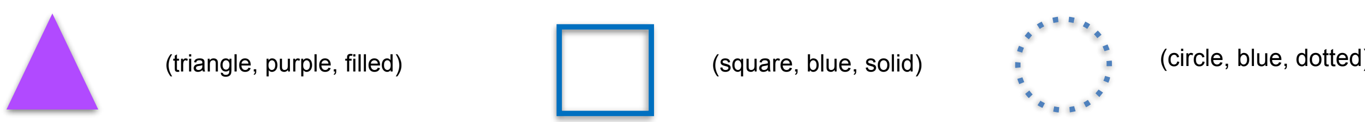
Toy World

- Sanity check

- Simple, synthetic world
  - Instances - (shape, color, style)
  - Total of $4^3 (64)$ instances

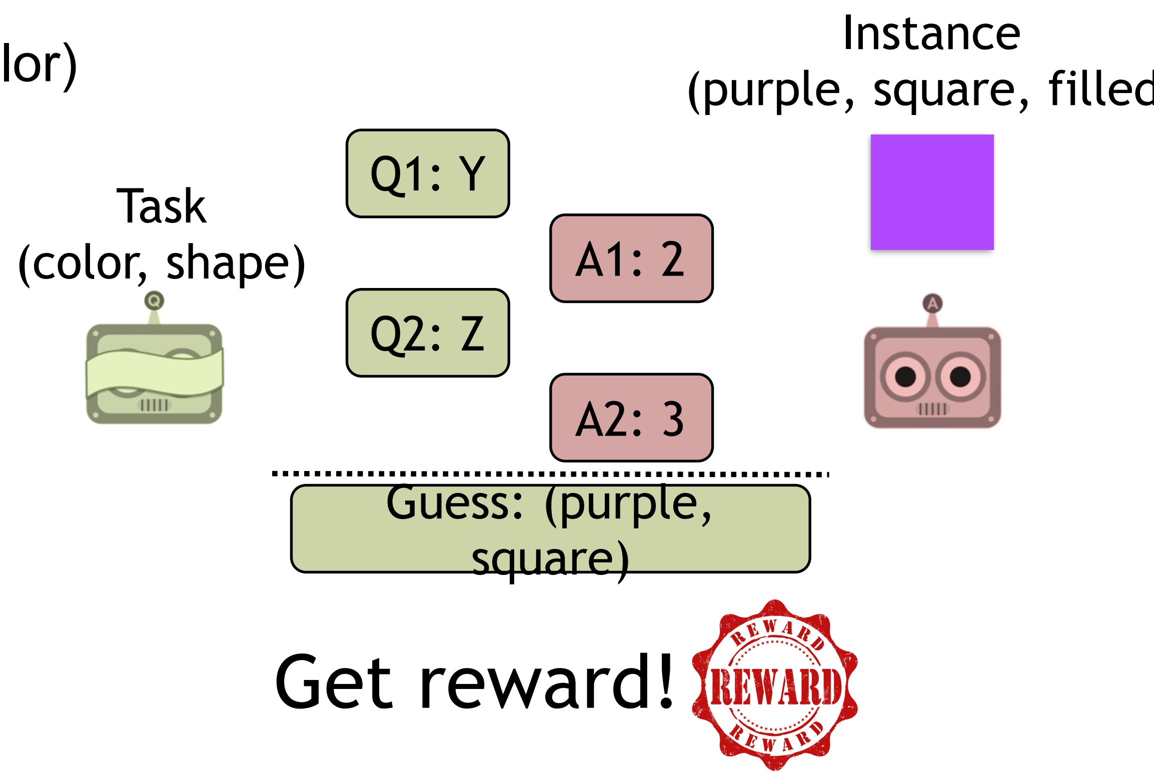


  - Example instances:

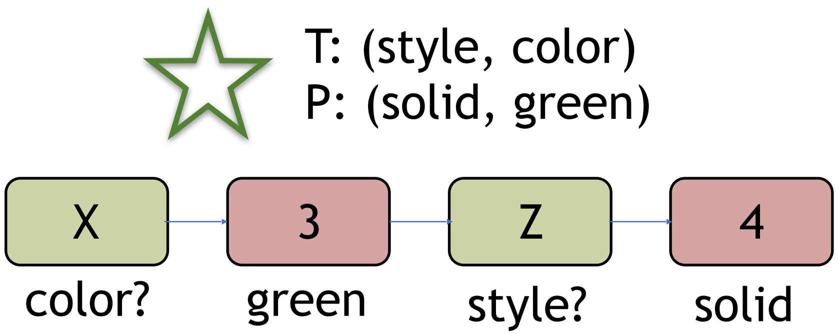


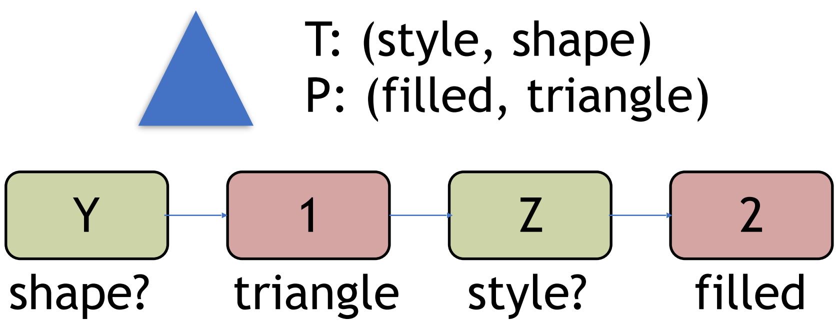
Task & Talk

- Task (G)
  - Inquire pair of attributes
  - (color, shape), (shape, color)
- Talk
  - Single token per round
  - Two rounds
- Q-bot guesses a pair
  - Reward : +1 / -1
  - Prediction order matters!



Emergence of Grounded Dialog





Emergence of Grounded Dialog

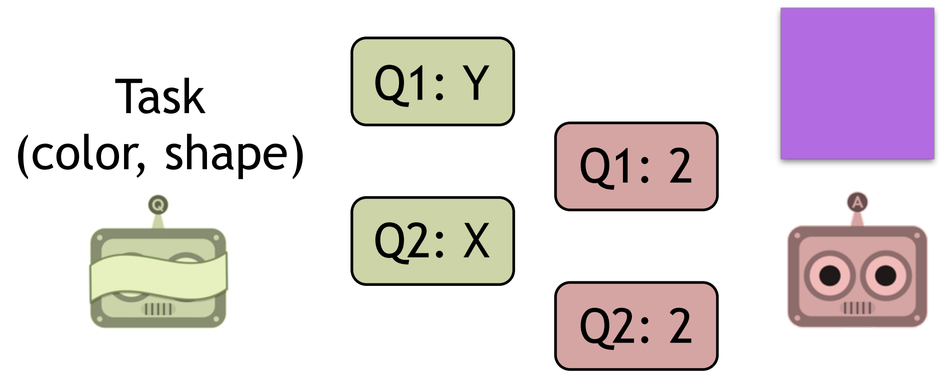
- Compositional grounding
- Predict dialog for unseen instances

	Attributes		
	color	*shape*	*style*
V_A	X	Y	Z
1	*blue*	*triangle*	*dotted*
2	*purple*	*square*	*filled*
3	*green*	*circle*	*dashed*
4	*red*	*start*	*solid*

(a) A-BOT

Task	q_1, q_2
(color, shape)	Y, X
(shape, color)	
(shape, style)	Y, Z
(style, shape)	
(color, style)	Z, X
(style, color)	X, Z

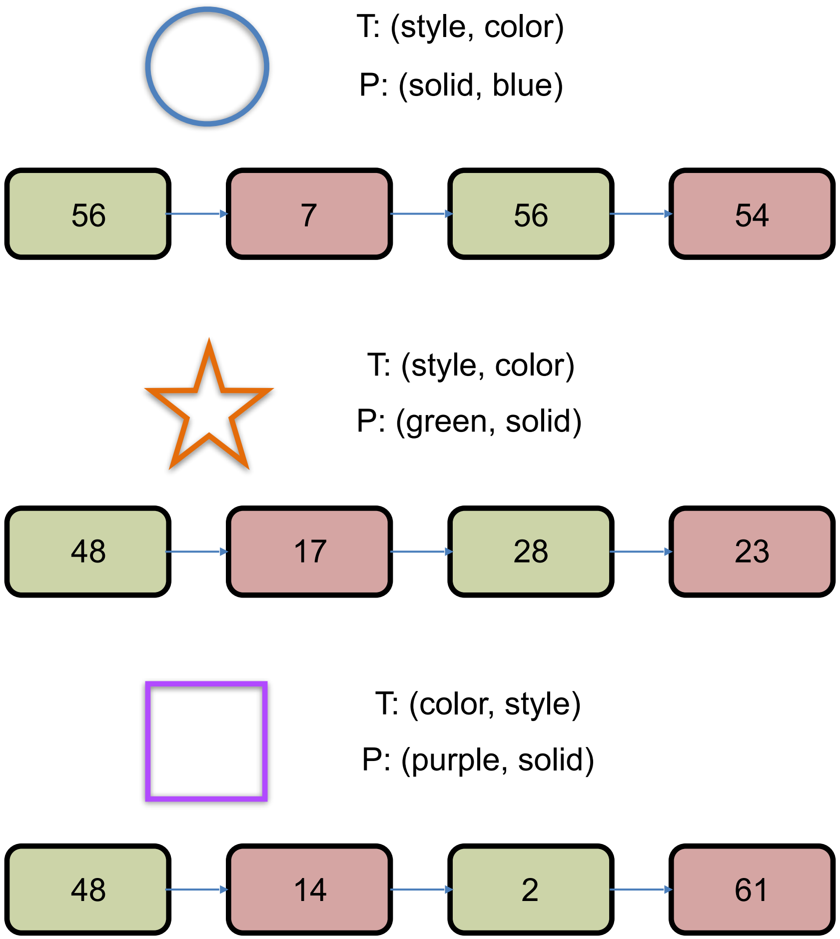
(b) Q-BOT



A. Over-complete vocabulary

-
  - $|V_Q|, |V_A| > 64$ (#instances)
    - More capacity than needed
    - Enough for A-bot to convey entire instance in a single round
  - Observations:
    - ABot learns a mapping between instance and (pairs) of V_A

A. Over-complete vocabulary



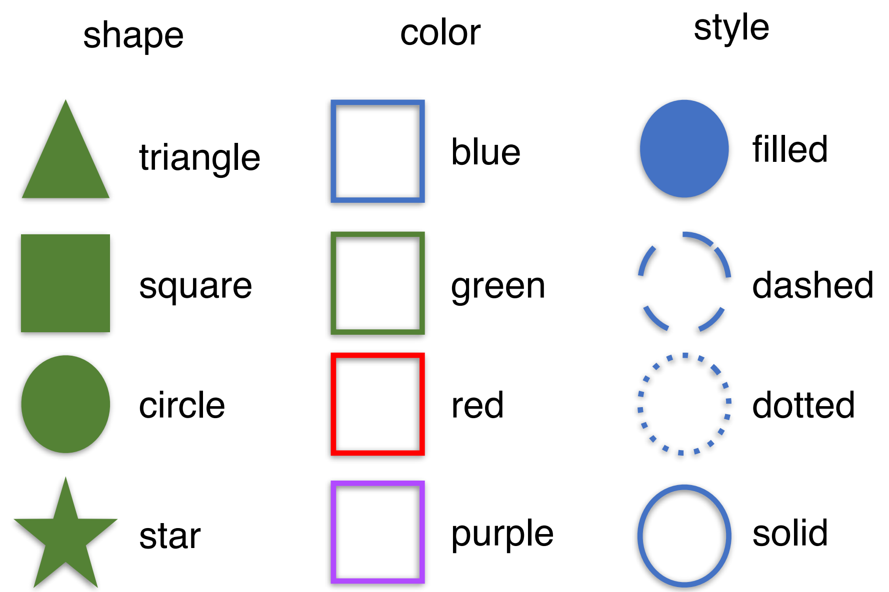
- Learn mapping from instance to pairs of V_A
- Two rounds unnecessary

- Q-bot immaterial
- Lewis Signaling Game
- No compositionality
- Poor generalization: 25.6%

- Let's limit capacity!

B. Attribute vocabulary

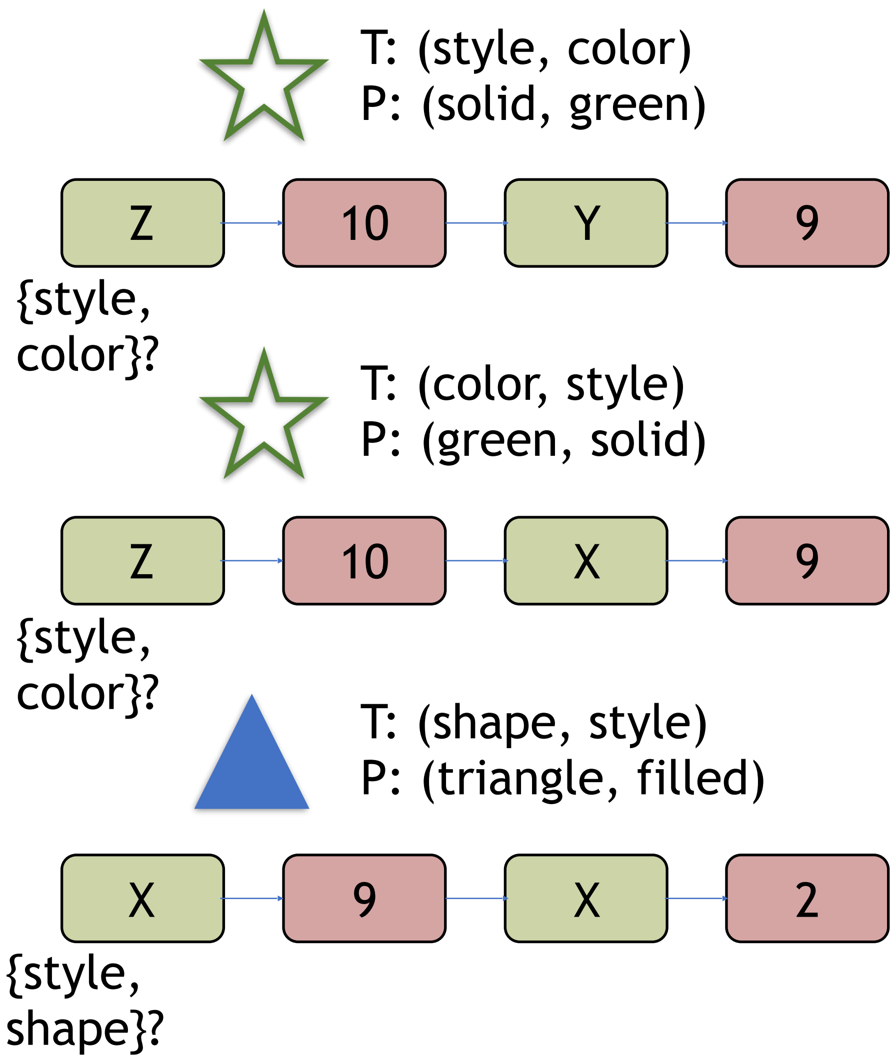
- $|V_Q| = 3$
  - Matches #attributes
  - (style, shape, color)



- $|V_A| = 12$
  - Matches #attribute values
  - 3 attributes x 4 values / attribute

- Can they learn individual concept grounding ?

B. Attribute vocabulary

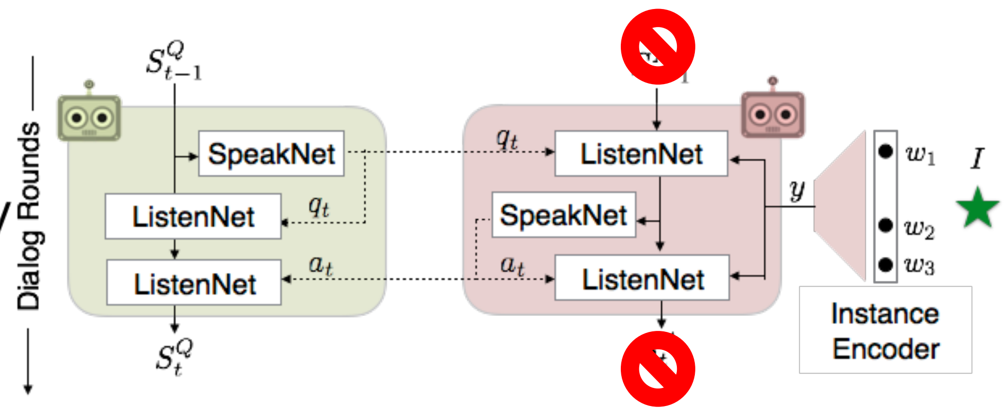


- $|V_Q| = 3, |V_A| = 12$
- Q-bot conveys task in first round
- Non-compositional
- Inconsistent A-bot grounding
- Poor generalization : 38.5%

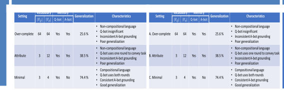
C. Minimal vocabulary

- $|V_Q| = 3, |V_A| = 4$
  - Minimum
  - Q-bot : tasks across two rounds
  - A-bot : values across two rounds

- Remove A-bot memory
  - A-bot cannot leverage history
  - Consistent groundings



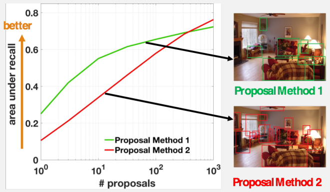
Summary of findings



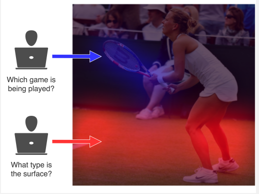
Setting	Vocabulary		Memory		Generalization	Charac
			Q-bot	A-bot		
A. Over-complete	64	64	Yes	Yes	25.6 %	• Non-compositio • Q-bot insignific • Inconsistent A- • Poor generaliza
B. Attribute	3	12	Yes	Yes	38.5 %	• Non-compositio • Q-bot uses one task • Inconsistent A- • Poor generaliza
C. Minimal	3	4	Yes	No	74.4 %	• Compositional • Q-bot uses bot • Consistent A-b • Good generaliz

Outline

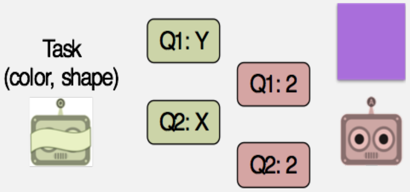
Object-Proposal Evaluation Protocol is 'Gameable' [CVPR '16]



Human Attention in VQA: Do Humans and Deep Networks Look at the Same Regions? [EMNLP '16]



Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog [EMNLP '17]



Rant about Vision vs NLP reviewing!

Object-Proposal Evaluation Protocol is ‘Gameable’

[CVPR ‘16]


Neelima Chavali*


Harsh Agrawal*


Aroma Mahendru*


Dhruv Batra

Rejected from ICCV 2015

- Weak Reject, Weak Reject, Strong Reject

- R1: Strengths:
  - *The authors convincingly show that object proposal generation is indeed "gameable".*
  - *The paper warns for dataset biases, which is an important issue.*
  - *The authors introduce an improved version of the fully annotated Pascal Context dataset [41] by providing instance-level annotations.*
  - *The paper is well written.*
  - *The authors release software which allows comparison of object proposal methods.*
  - *The main object proposal methods are compared.*
  - *Three different datasets are used with quite a variety of different object classes.*

Rejected from ICCV 2015

- Weak Reject, Weak Reject, Strong Reject

- R1: Weaknesses:
  - *I do not find many new insights in the paper. To my knowledge, methods on generating object proposals have always stressed the importance of being able to find any object, even though because of lack of datasets they were only evaluated on Pascal VOC. It is indeed true that one could learn class-specific classifiers to get better results on creating object proposals, but this is not very surprising and I think that such a method would not survive a good review process.*

Rejected from ICCV 2015

- Weak Reject, Weak Reject, Strong Reject

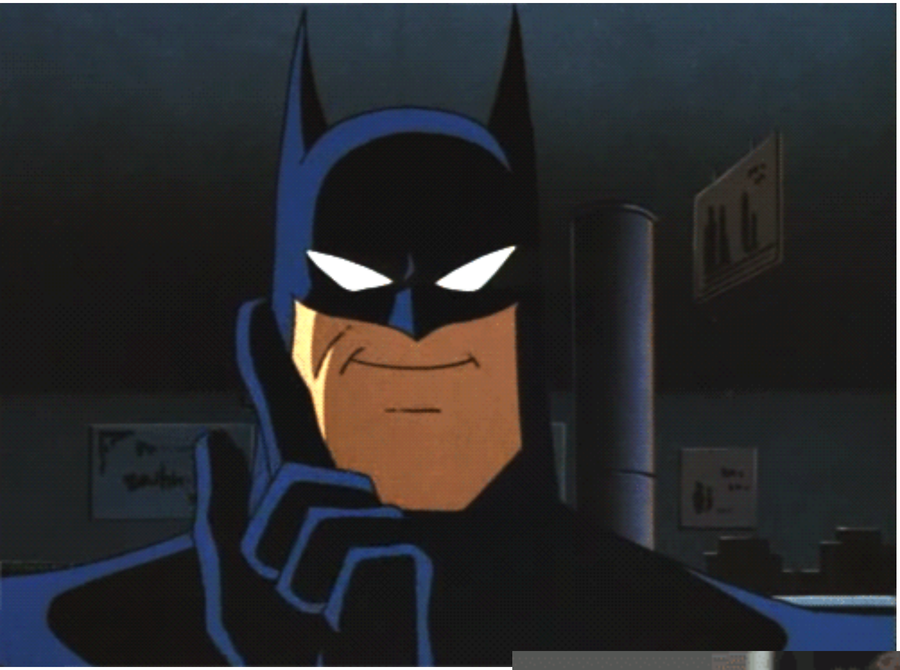
- R1: Weaknesses:
  - *The authors posit that there are two interpretations of object proposal: "category-independent object proposals" and "detection proposals", whereas the latter is to improve object detection only for the target classes in the dataset. However,* ***I have never seen a work on object proposals which uses the second interpretation because this is just a fundamentally wrong interpretation***

Rejected from ICCV 2015

- Consistent theme
  - *Hence I feel that this paper does not bring enough new insights to the community to* ***warrant publication at a top-tier conference like ICCV****.*
  - *The use of the Pascal dataset as a base for the experiments is in the reviewer's opinion* ***not enough in an current ICCV paper****.*
  - *All in all, the reviewer finds that the only contribution of the paper is an experiment to make researchers aware of the potential per-class bias of the current evaluations, which the* ***reviewer thinks is not enough for an ICCV paper****.*

CVPR Reviewers

From reviewers with love:
Not on my watch!



How to write a good CVPR submission

Bill Freeman
MIT CSAIL
Nov. 6, 2014

From an AC's point of view

- The types of papers in your pile:
  - About 1/3 are obvious rejects
  - In the whole set, maybe 1 is a really nice paper
    - well-written, great results, good idea.
  - The rest are borderline, and these fall into two camps..

Two types of "borderline" papers

- The Cockroach
  - You try, but you can't find a way to kill this paper. While there's nothing too exciting about it, it's pretty well written, the reviews are ok, the results show an incremental improvement. Yet another kind of boring CVPR paper.

Two types of “borderline” papers

- The Puppy with 6 toes
  - A delightful paper, but with some easy-to-point-to flaw. This flaw may not be important, but it makes it easy to kill the paper, and sometimes you have to reject that paper, even though it’s so fresh and wonderful.

Short Papers!

Long papers

EMNLP 2017 long paper submissions must describe substantial, original, completed and unpublished work. Wherever appropriate, concrete evaluation and analysis should be included. Each submission will be reviewed by at least three program committee members. Each long paper submission consists of a paper of up to **eight (8) pages of content, plus unlimited pages for references**; final versions of long papers will be given one additional page (up to nine pages with unlimited pages for references) so that reviewers' comments can be taken into account.

Short papers

EMNLP 2017 also solicits short papers. Short paper submissions must describe original and unpublished work. While a short paper is not a shortened long paper, the characteristics of short papers include a small, focused contribution; work in progress; a negative result; an opinion piece; an interesting application nugget. Each short paper submission consists of up to **four (4) pages of content, plus unlimited pages for references**; final versions of short papers will be given one additional page (up to five pages in the proceedings and unlimited pages for references) so that reviewers' comments can be taken into account. Each short paper submission will be reviewed by at least three program committee members.

Concluding thoughts

- Definitive conclusive negative results are hard
- Publishing such negative results is *even* harder!
- It seems to be particularly hard at Vision venues
- Please please please please please please
  try accepting the puppy with 6 toes!

Machine Learning & Perception Group

Dhruv Batra
Assistant Professor

PhD

Qing Sun

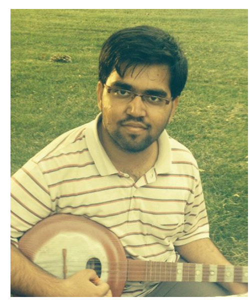
Aishwarya Agrawal

Yash Goyal

Michael Cogswell

Abhishek Das

Ashwin Kalyan



MS

Aroma Mahendru

Akrit Mohapatra

Postdoc

Stefan Lee

Interns

Deshraj Yadav

Tejas Khot

Viraj Prabhu

Thanks!