Visual Question and Answering

Shayan and Prem

Overview

- DAQUAR and the Visual Turing Challenge
- VQA dataset and methods
- VQA v2 and a model with explanation

Toward a Visual Turing Challenge

Malinowski et al.

Story so far

- Tremendous progress in machine perception and language understanding
- Motivation to attack visual question-answering
- Performance of different methods:
  - Measured against a crafted set of benchmarks
- Crowdsourcing generated curated datasets, with a unique ground truth

Challenges of Crafting Good Benchmarks

- As the complexity and the openness of the task grows, becomes more difficult
  - Interpreting and evaluating the answer of the system
  - Establishing and evaluation methodology that assigns scores
  - Inconsistency even in human answers
    - What are the “true” annotations?
- Instead, we can use “social consensus”
  - Multiple human answers as different interpretations

Challenges of dealing with difficult tasks

- Vision and Language
- Common sense knowledge
- Defining a benchmark dataset and quantifying performance

Vision and Language

- Scalability: scale up to thousands of concepts
- Concept ambiguity: As the number of categories increase, the semantic boundaries become more fuzzy
- Attributes: human concepts also includes, not only objects, but attributes as well
    - Gender, colors, states (lights can be on or off)
- Ambiguity in reference resolution
    - Object-centric, observer-centric, or world-centric frames of reference

Common sense knowledge

- Some question can be answered with high reliability with only access to common sense knowledge
  - “Which object on the table is used for cutting?” Probably knife or scissors
  - “What is in front of scissors?”
- Can be utilized to fulfill the task or limit the hypothesis space
- Reduce the computational complexity

Defining a benchmark dataset and quantifying performance

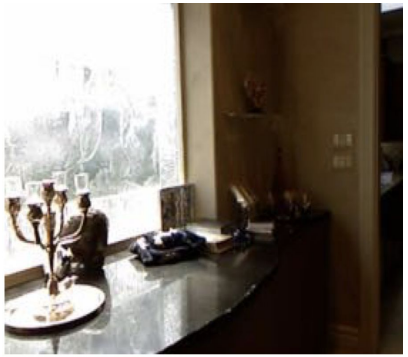
- VQA is about an end to end system
- Do not want to enforce any constraints for the internal representation
- Benchmark dataset for VQA similar to turing test is more tractable
- QA needs textual annotations for the aspects related to the questions

DAQUAR: dataset for Visual Turing Challenge

- Images present real-world indoor scenes
- Based on NYU-Depth v2 dataset, with fine-grained categories
- Questions are unconstrained natural language sentences
- Contains 1088 different nouns in the questions, 803 in the answers, 1586 altogether

QA: (Where is oven?, on the right side of refrigerator)

QA: (What is behind the table?, window)



QA: (what is beneath the candle holder, decorative plate)

Quantifying the Performance

- Complexity and openness of the task makes it challenging:
  - *Automation* for evaluating different architectures at scale
  - *Ambiguity* inherent in complex task that we are facing
    - Multiple interpretation, hence many correct answers
  - Coverage: Automatic performance metric should assign similar scores to different ways of expressing the same concept

WUPS Score

- Automatic metric that quantifies performance of the holistic architecture

$$\frac{1}{N}\sum_{i=1}^{N} min\{\Pi_{a\in A^i} max_{t\in T^i}\mu(a,t), \Pi_{t\in T^i} max_{a\in A^i}\mu(a,t)\}.100$$

Visual Question Answering

Agrawal et al.

Introduction

Motivation

- Spawning the next generation of AI algorithms requires:
  - Multi-modal knowledge beyond a single sub-domain
  - Well-defined quantitative evaluation metric
- This paper introduces the task of free-form and open-ended Visual Question Answering

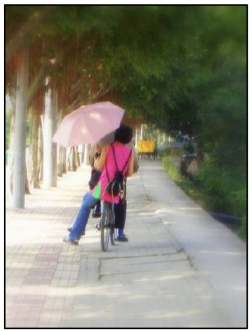
VQA System

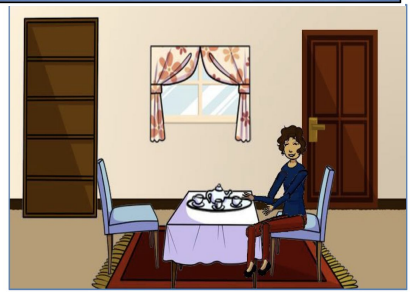
- Input:
  - Image
  - Free-form, open-ended, natural language question about the image
- Output:
  - Natural language answer as the output
- Requires a vast set of AI capabilities
  - Fine-grained recognition
  - Activity recognition
  - Knowledge base reasoning
  - Commonsense reasoning

Dataset

Dataset

- Consists of *real images* and *abstract scenes*
- *Real images*: 123k training images, 81k test images from MS COCO
  - Contains 5 image captions per image
- *Abstract scenes*:To explain the high-level reasoning for VQA, but not the low level vision tasks
  - Contains 50k scenes, 5 single captions for all abstract scenes





Questions

- Collecting interesting and diverse questions is challenging:
- Must require the image to answer the question
- Three question were gathered for each image/scene
- The subjects were shown previous questions
- 0.76M questions in total

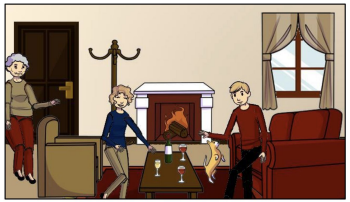
"We have built a smart robot. It understands a lot about images. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene (e.g., kitchen, beach), people's expressions and poses, and properties of objects (e.g., color of objects, their texture). Your task is to stump this smart robot!
Ask a question about this scene that this smart robot probably can not answer, but any human can easily answer while looking at the scene in the image."

Answers

Does this man have children?	yes yes yes	yes yes yes
Is this man crying?	no no no	no yes yes

uestion

e question



How many glasses are on the table?	3 3 3	2 2 6
s the woman ng for?	door handle glass wine	fruit glass remote

How many pickles are on the plate?	1 1 1	1 1 1
What is the shape of the plate?	circle round round	circle round round

Testing

- Two modalities in answering the question
  - Open-ended
  - Multiple choice
- Accuracy metric for open-ended task

$$min\left(\frac{\#\text{ of humans that provided that answer}}{3}, 1\right)$$

- Multiple choice task: 18 candidate answers were created for each question

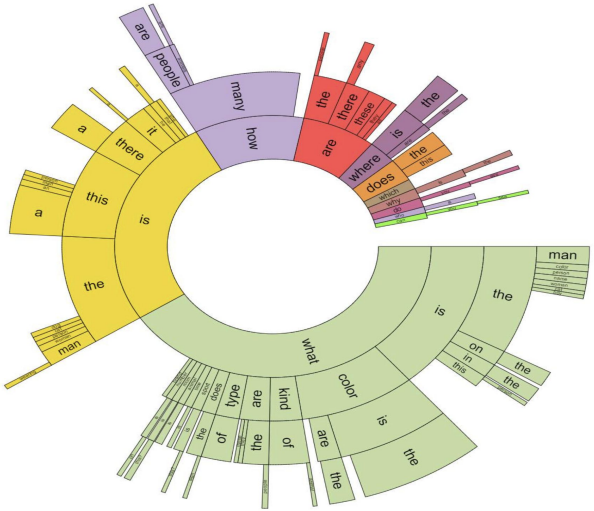
Testing (Cont.)

- Generated a candidate set of correct and incorrect answers from four sets of answers
- *Correct*: The most common out of 10 correct answers
- *Plausible*: three subjects answered the questions without seeing the image
    - Ensures that the image is necessary to answer the question
- *Popular*: 10 most popular answers
    - “Yes”, ”no, ”2”, “1”, “white”, “3”, ”red”, “blue”, “4”, “green” for real images
- *Random*: Correct answers from random questions in the dataset
- 18 candidate answers: union of correct, plausible, and popular, and then random answers added

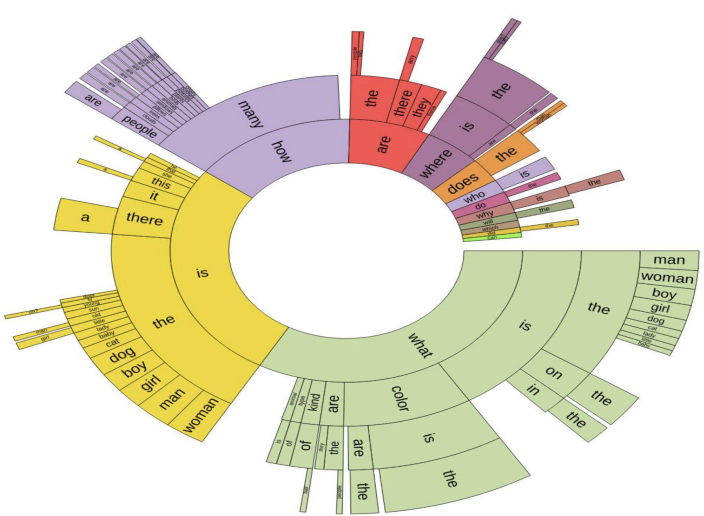
Questions Analysis

Questions can be clustered based on the words that start the question

Real Images

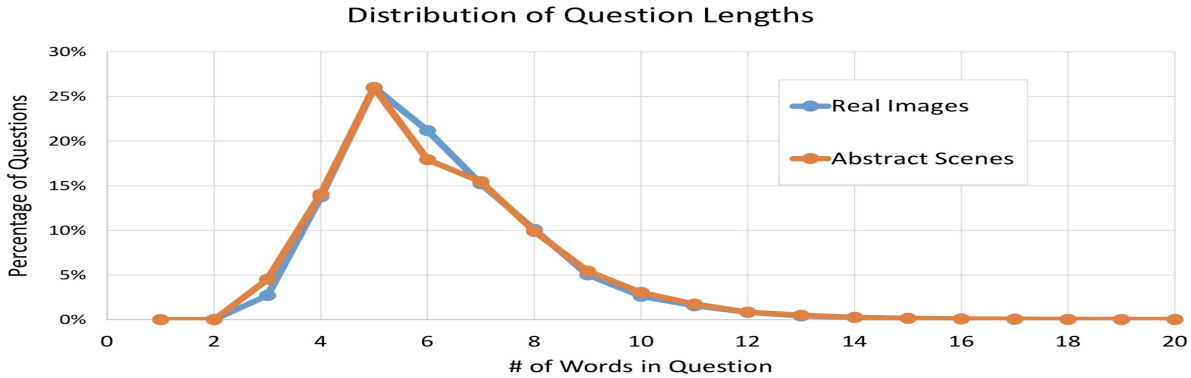


Abstract Scenes



Questions Analysis (Cont.)

- Length of the questions:



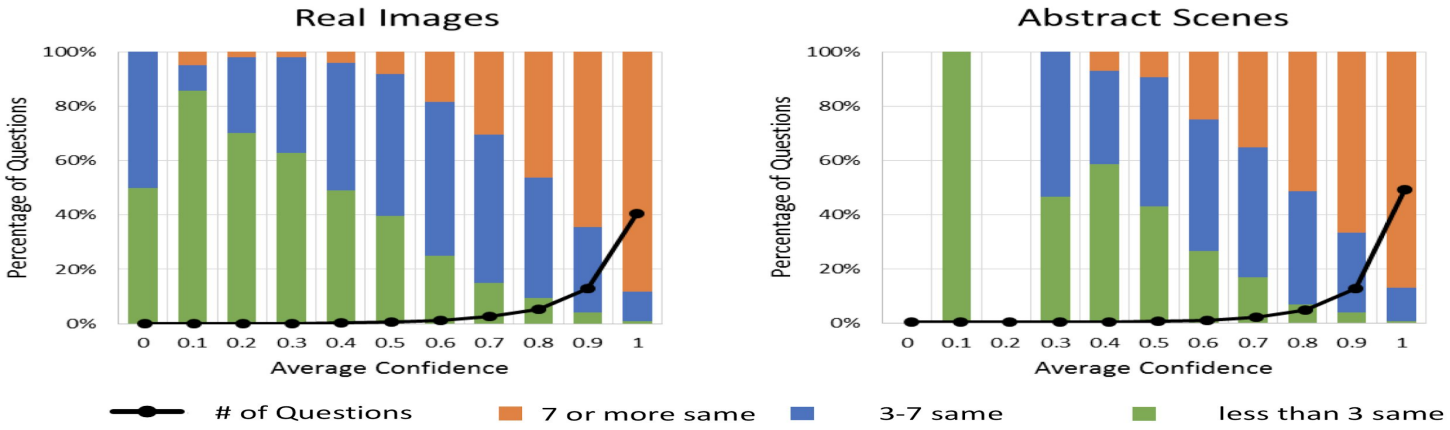
- Most questions range from 4 to 10 words

Answers Analysis

- Length of the answers:
  - Real images: one, two, three words are 89.32%, 6.91%, and 2.74% respectively
  - Abstract images: one, two, three words are 90.51%, 5.89%, and 2.94% respectively
- The brevity of the answers make automatic evaluation feasible
- Brevity of the answer does not necessarily make the problem easier
  - Open-ended answers to open-ended questions
- Questions require complex reasoning to arrive at these “simple” answers

Answer Analysis (Cont.)

- Subject confidence and inter-human agreement
- Self-judgement of confidence -> Answer agreement between subjects ??

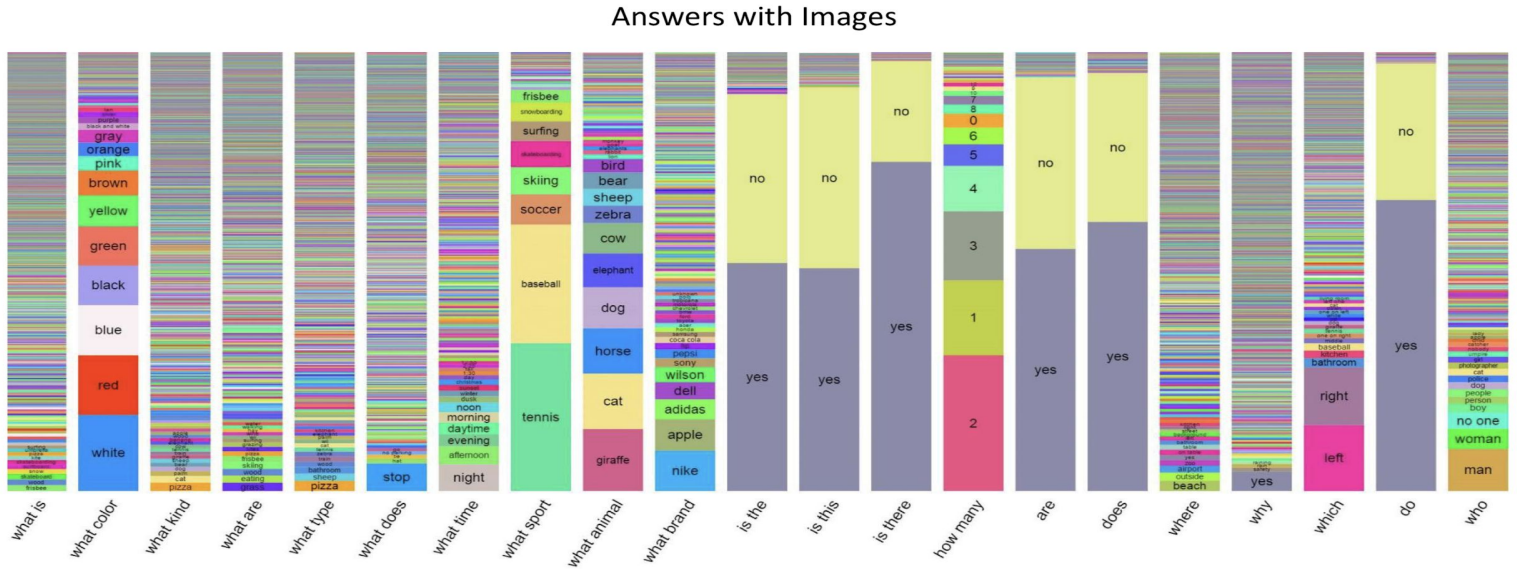


Is the image necessary?

Test accuracy of human subjects when asked to answer the question

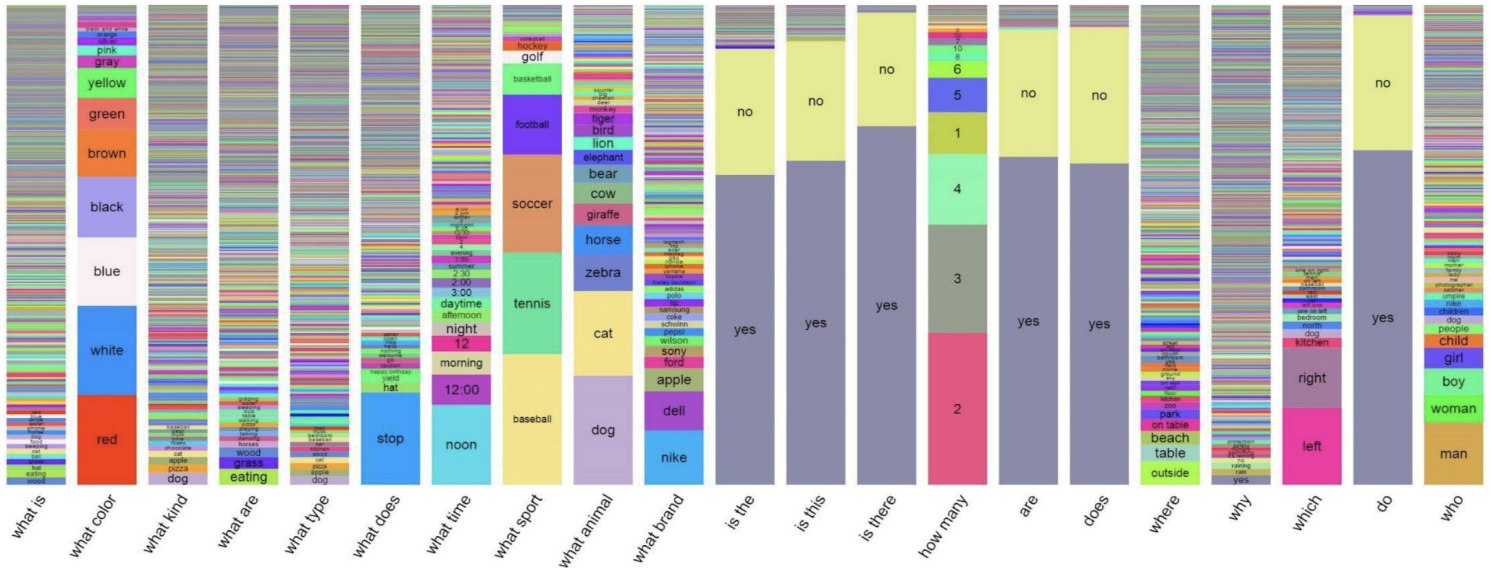
Dataset	Input	All	Yes/No	Number	Other
Real	Question	40.81	67.60	25.77	21.22
	Question + Caption*	57.47	78.97	39.68	44.41
	Question + Image	83.30	95.77	83.39	72.67
Abstract	Question	43.27	66.65	28.52	23.66
	Question + Caption*	54.34	74.70	41.19	40.18
	Question + Image	87.49	95.96	95.04	75.33

Answers Analysis



Answer Analysis (Cont.)

Answers without Images



Which Questions Require Common Sense?

- Two AMT studies on a subset of 10k images from the real images
- Subjects were asked:
  - Whether or not they believed a question required commonsense to answer the question
  - The youngest age group that can answer the question
- Each question was shown to 10 subjects
- For 47.43% of questions 3 or more subjects voted yes to commonsense
- Degree of commonsense to answer a question: percentage of subjects who voted yes

Evaluation

VQA Baselines

- Random: Randomly choose an answer from the top 1k answers
- Prior (“yes”): Most popular answer
- Per question type prior: Most popular answer per question type
- Nearest neighbor: find k nearest neighbors, pick the most frequent ground truth

Methods

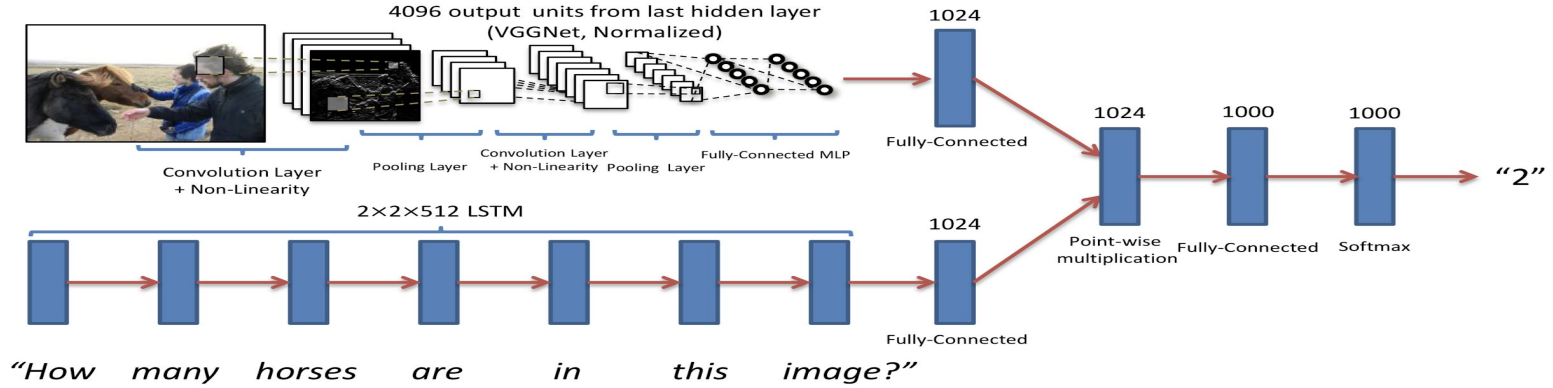
- 2-channel model: Vision (image) channel, and language (question) channel
- Top k=1000 most frequent answers are chosen as the possible outputs
- Covers 82.67% of the train+val answers
- Image Channel:
  - **I** : activation from the last layer of VGGNet are used as 4096-dim image embedding
  - **Norm I** : I2 normalized activations from the last hidden layer of VGGNet

Methods (Cont.)

- Question Channel
  - *Bag-of-words Question (BoW Q)*: top 1000 words in the question + top 10 first, second and third word of the question -> 1030 dim embedding of the question
  - *LSTM Q*: One hidden layer -> 1024 dim embedding of the question
  - *Deeper LSTM Q:* Two hidden layer

Methods (Cont.)

- Multilayer Perceptron: Two embeddings are combined to obtain a single embedding
  - BoW Q + I : concatenate BoW Q and I embeddings
  - LSTM Q + I , deeper LSTM Q+ norm I: image embedding transformation, and fusion with LSTM embedding



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

Results (Cont.)

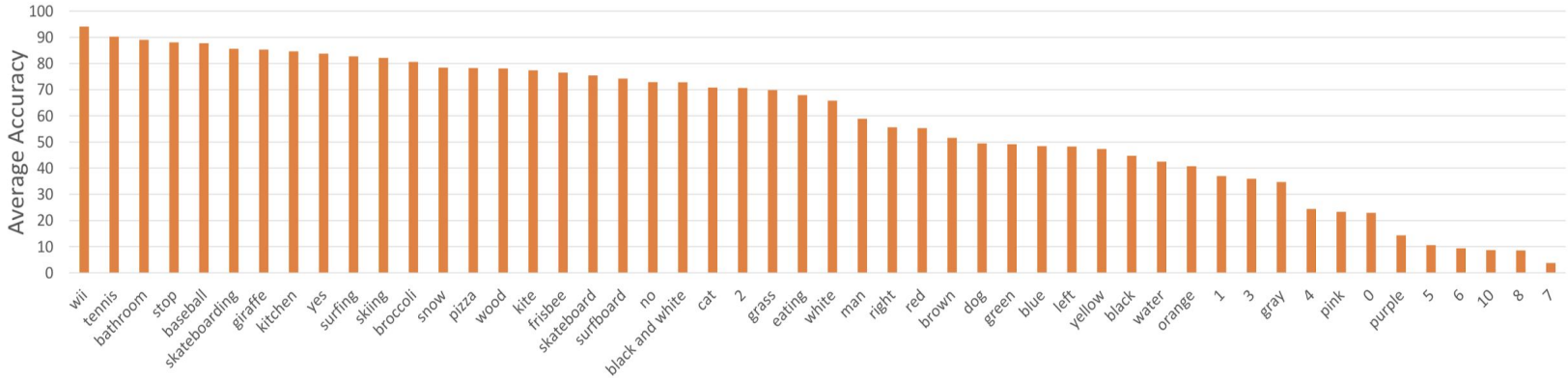


Fig. 9: Pr (system is correct | answer) for 50 most frequent ground truth answers on the VQA validation set (plot is sorted by accuracy, not frequency). System refers to our best model (deeper LSTM Q + norm I).

Results (Cont.)

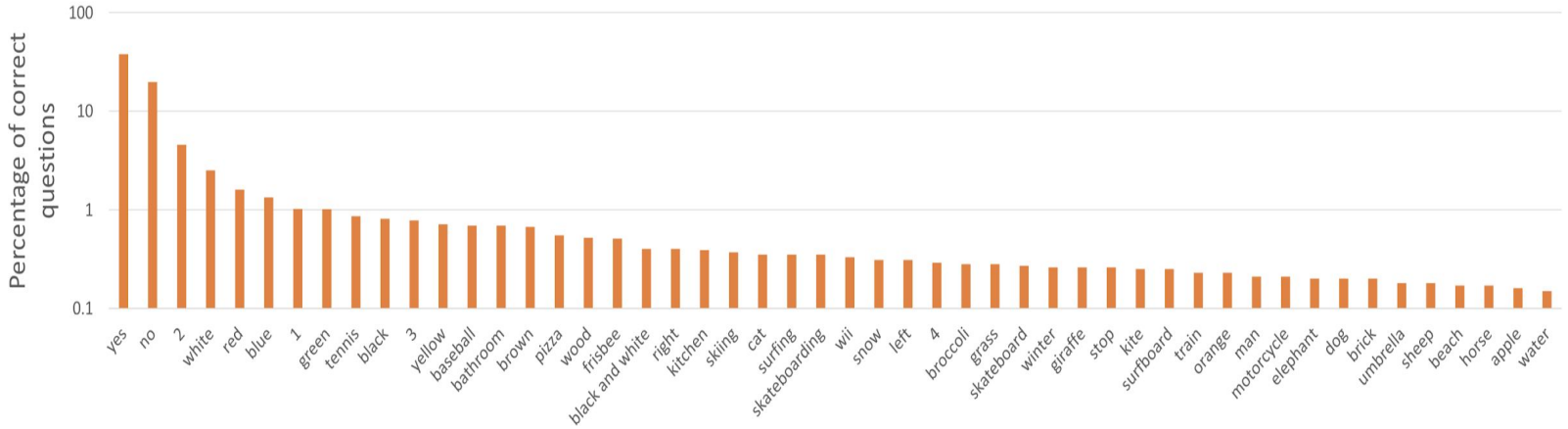
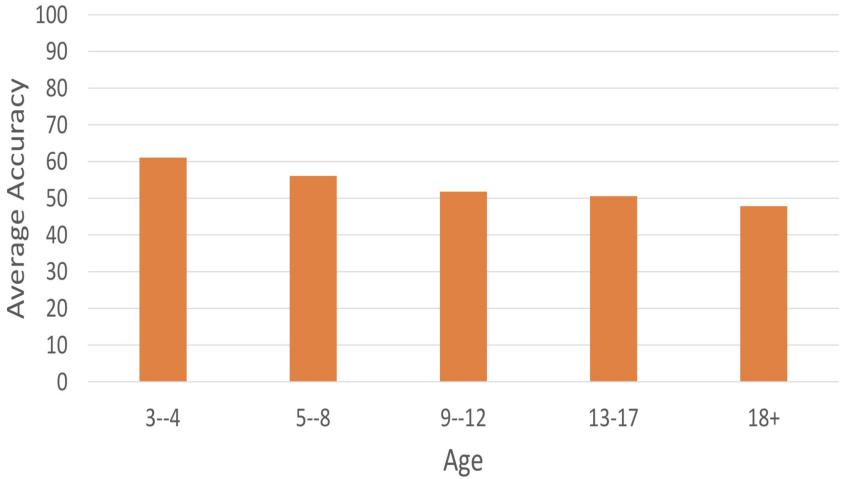
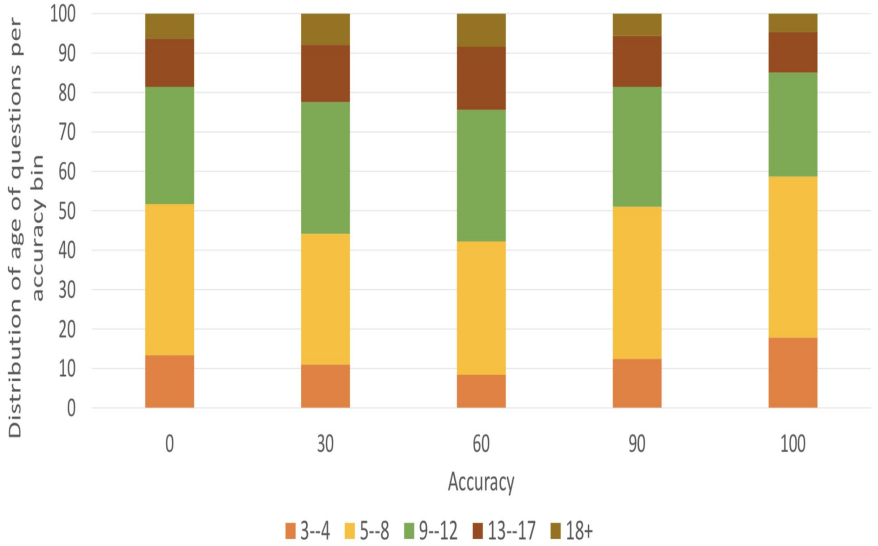


Fig. 10: Pr (answer | system is correct) for 50 most frequently predicted answers on the VQA validation set (plot is sorted by prediction frequency, not accuracy). System refers to our best model (deeper LSTM Q + norm I).

Results (Cont.)





Conclusion

Conclusion

- Images are important in answering the question (both for humans and the models)
- The accuracy of the best model is still far from human accuracy
- The best model performs as well as a 4.74 year old child !

Making the V in VQA Matter

Goyal et al.

Some problems with VQA

Language priors can overshadow visual information and lead to good performance.

- Tennis is the most popular sport (41%)
- 2 is the most popular number (39%)

“Visual priming bias”

- Do you see a...? Yes (87%)
- “Is there a clock tower?” only asked when there are clock towers

VQA v2

- For every (I, Q, A) triple have (I', Q, A') where A != A'
- Increases entropy of P(A|Q) to make vision more important
- SOTA v1 models do worse on v2
- The dataset formulation allows for a new kind of explainable model

Dataset creation

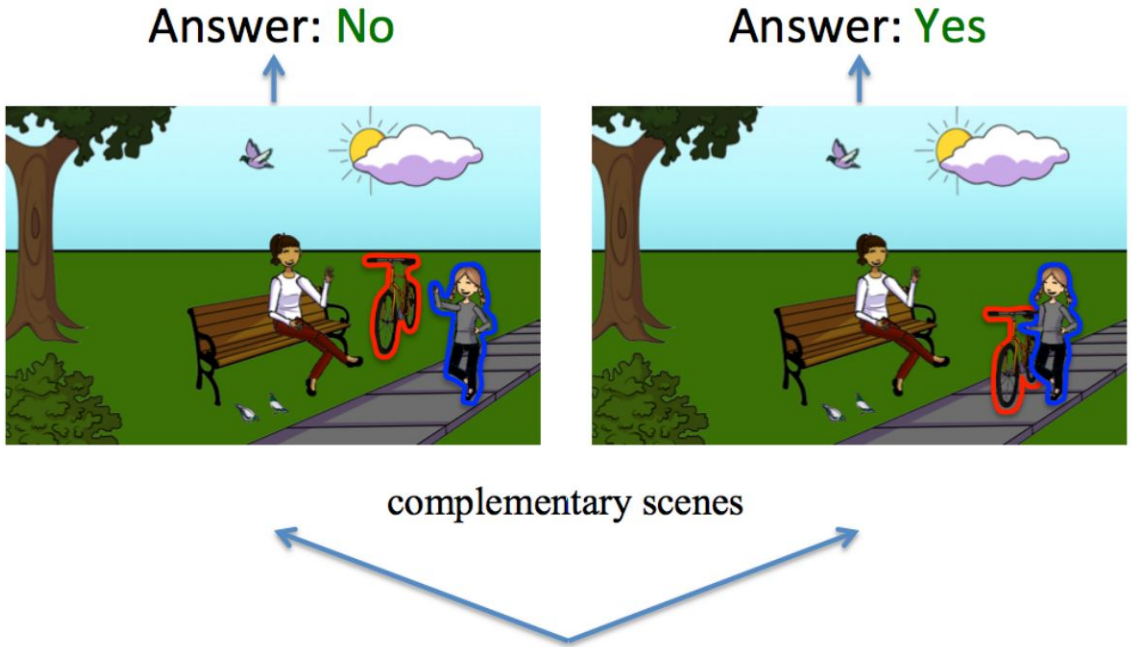
Expand VQA v1

Dataset: a large set of (I, Q, A) tuples

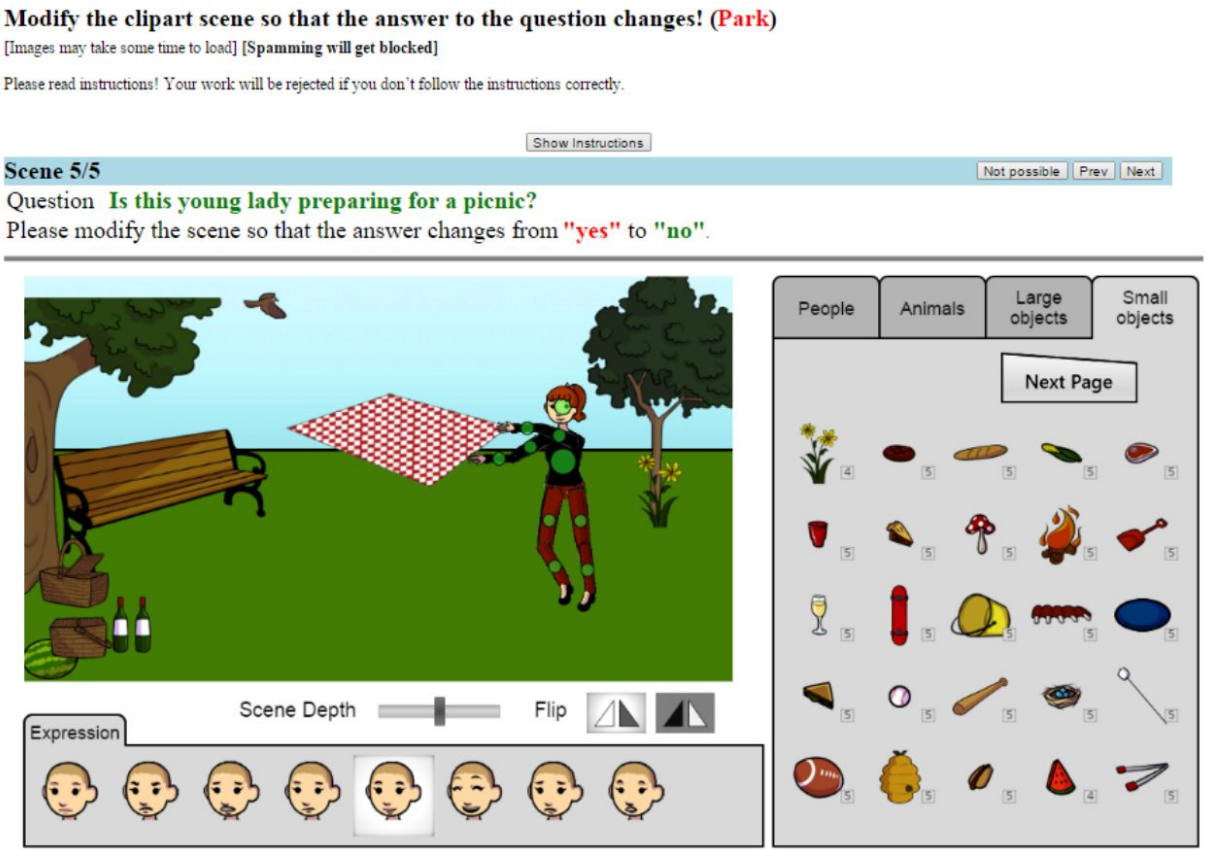
We want to find another relevant (I', Q, A') such that A' != A.

Additional requirements

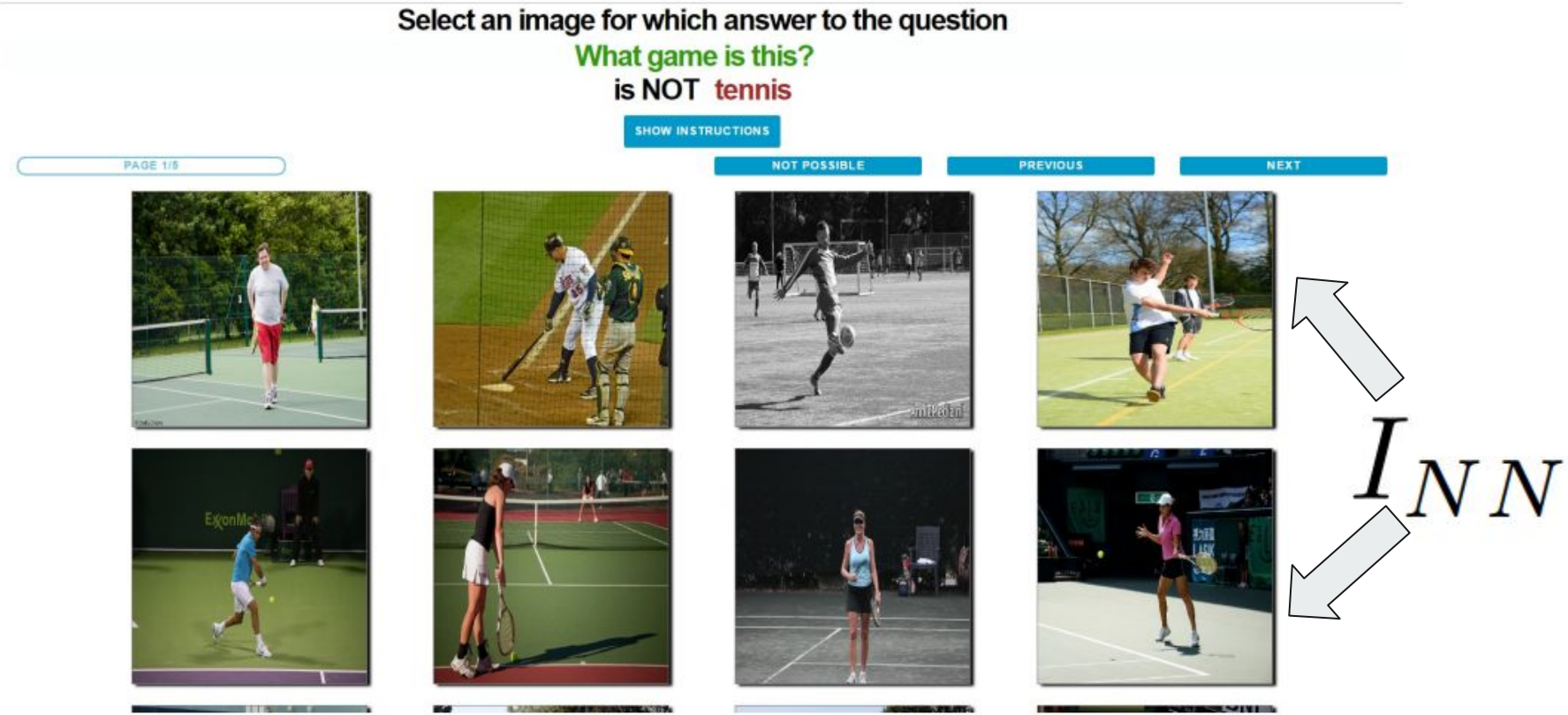
- Q should make sense for I' as well
- Method should work for entire dataset



I restricted to abstract scenes, Q restricted to binary questions, only one possible A' (Zhang et al.)



AMT interface from [Zhang et al.]



This paper's AMT interface to select I'

Statistics

- Not possible selected 22% of the time (135k questions)
  - Object too small, similar images don't have it
  - Any A' is rare (consider Q "What color is the banana?")
- A=A' about 9% of the time (worker was inaccurate?)
- 195k/93k/191k additional images for train/test/val
- Total is now 443k/214k/453k (question, image) pairs
- Entropy across answer distributions, weighted by question type is +56%

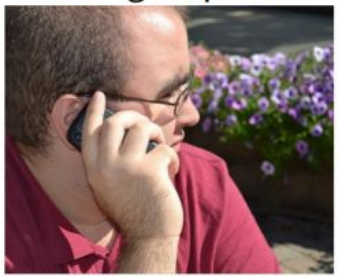
How many doughnuts have sprinkles?

3

2

What task is the man performing?

talking on phone



eating

What is this device?

train

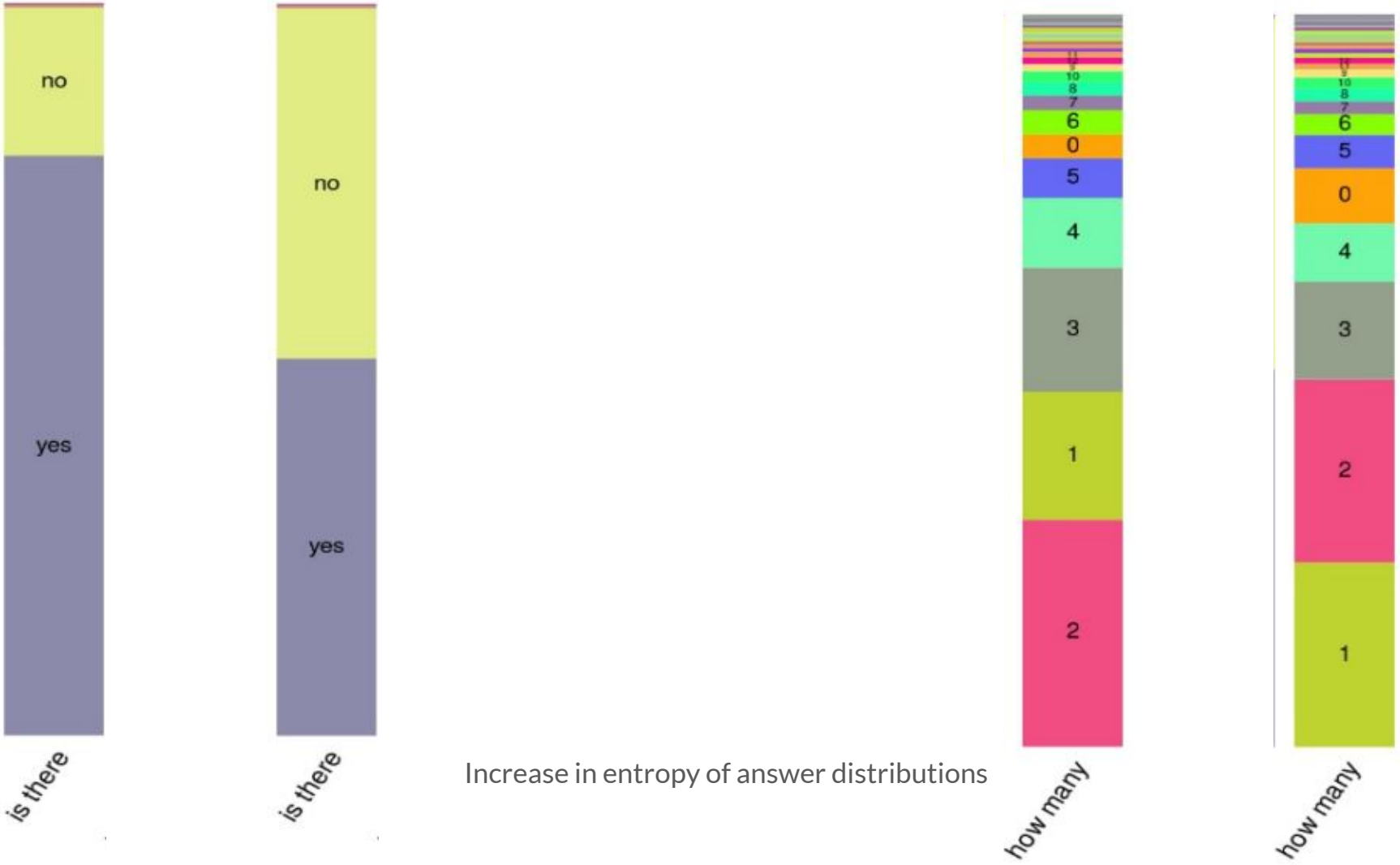
airplane

What is the girl reaching into?

bucket

apples

Sample complementary images


no
yes
is there
no
yes
is there
Increase in entropy of answer distributions
2
1
3
4
0
5
6
7
8
10
how many
2
1
3
4
0
5
6
7
8
10
how many

Benchmark

Models compared

- Deeper LSTM + norm I from VQA paper
- Hierarchical Co-Attention (previously SOTA)
- Multimodal Compact Bilinear Pooling (MCB) (2016 challenge winner)
- Baselines
  - Always yes (27% on unbalanced v1, 24% on balanced v2)
  - Similar to Deeper LSTM + norm I but without I (question only)

Approach	UU	UB	$B_{half}B$	BB
Prior	27.38	24.04	24.04	24.04
Language-only	48.21	41.40	41.47	43.01
d-LSTM+n-I [24]	54.40	47.56	49.23	51.62
HieCoAtt [25]	57.09	50.31	51.88	54.57
MCB [9]	60.36	54.22	56.08	59.14

First character is training, second character is testing

Approach	UU	UB	$B_{half}B$	BB
Prior	27.38	24.04	24.04	24.04
Language-only	48.21	41.40	41.47	43.01
d-LSTM+n-I [24]	54.40	47.56	49.23	51.62
HieCoAtt [25]	57.09	50.31	51.88	54.57
MCB [9]	60.36	54.22	56.08	59.14

Models trained on v1 do worse on v2

Approach	UU	UB	$B_{half}B$	BB
Prior	27.38	24.04	24.04	24.04
Language-only	48.21	41.40	41.47	43.01
d-LSTM+n-I [24]	54.40	47.56	49.23	51.62
HieCoAtt [25]	57.09	50.31	51.88	54.57
MCB [9]	60.36	54.22	56.08	59.14

When training on similar size v2, we improve

Approach	UU	UB	$B_{half}B$	BB
Prior	27.38	24.04	24.04	24.04
Language-only	48.21	41.40	41.47	43.01
d-LSTM+n-I [24]	54.40	47.56	49.23	51.62
HieCoAtt [25]	57.09	50.31	51.88	54.57
MCB [9]	60.36	54.22	56.08	59.14

Performance increases with more data

Approach	Ans Type	UU	UB	$B_{half}B$	BB
MCB [9]	Yes/No	81.20	70.40	74.89	77.37
	Number	34.80	31.61	34.69	36.66
	Other	51.19	47.90	47.43	51.23
	All	60.36	54.22	56.08	59.14
HieCoAtt [25]	Yes/No	79.99	67.62	70.93	71.80
	Number	34.83	32.12	34.07	36.53
	Other	45.55	41.96	42.11	46.25
	All	57.09	50.31	51.88	54.57

Table of two models divided into question types

Approach	Ans Type	UU	UB	$B_{half}B$	BB
MCB [9]	Yes/No	81.20	70.40	74.89	77.37
	Number	34.80	31.61	34.69	36.66
	Other	51.19	47.90	47.43	51.23
	All	60.36	54.22	56.08	59.14
HieCoAtt [25]	Yes/No	79.99	67.62	70.93	71.80
	Number	34.83	32.12	34.07	36.53
	Other	45.55	41.96	42.11	46.25
	All	57.09	50.31	51.88	54.57

Large decrease in yes/no

Approach	Ans Type	UU	UB	$B_{half}B$	BB
MCB [9]	Yes/No	81.20	70.40	74.89	77.37
	Number	34.80	31.61	34.69	36.66
	Other	51.19	47.90	47.43	51.23
	All	60.36	54.22	56.08	59.14
HieCoAtt [25]	Yes/No	79.99	67.62	70.93	71.80
	Number	34.83	32.12	34.07	36.53
	Other	45.55	41.96	42.11	46.25
	All	57.09	50.31	51.88	54.57

Large increase in Y/N and number

Counterexample explanation

Explainable models

Related work:

- Generate natural language explanation (Hendricks et al.)
- Heatmap of important image regions (many papers)

Explanation by counterexample: find a similar image which we would have returned a different answer

First attempt

Model takes in (Q, I) and produces A_{pred}.

Search through I_{NN} for I' with the lowest $P(A_{pred})$.

But Q might not apply to I'!

Two headed neural network

Shared trunk: Generate QI embeddings for I and I_{NN}

Answering head: Predict A_{pred} from QI embedding of I.

Explaining head: Score all of I_{NN} using their QI embeddings and A_{pred}.

Loss function is cross-entropy (for A) + sum of pairwise hinge losses (for I')

Q: Which way is its head turned?
A: left

Q: What color is the plate?
A: blue

Sample output of the model

	Random	Distance	VQA [3]	Ours
Recall@5	20.79	42.84	21.65	43.39

Nearly worse than a naive baseline!

Improvements?

Models

- Similar images I and I’ still have very high answer equivalence
- Limited to fixed number of answers

Explanation

- The fancy explanation model isn’t much better than a close image.

Dataset

- 22% “not possible”

Wrap up

Wrap up

Is VQA an ultimate Turing Test for AI? Perception, language, KBs, etc. all needed

VQA v2 is today's dataset of choice, accumulating improvements from DAQUAR and VQA v1

VQA challenge accuracy is around 69-70%, still has room for improvement

Why was Anton demoted on the arXiv VQA v1 paper author list?