VQA: Visual Question Answering
(ICCV 2015)

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, L. Zitnick, D. Parikh

Himani Arora & Hassan Akbari
Deep Learning for Computer Vision, Speech and Language (Spring 2017)

Introduction

- Proposed the task of *free-form* and *open-ended* Visual Question Answering:
  - Given an image and a natural language question about the image, the task is to provide an accurate natural language answer.
- Both the questions and answers are open-ended

AI-Complete Task

- Should require *multi-modal knowledge* beyond a single sub-domain (such as Computer Vision)
- Have a well-defined *quantitative evaluation metric* to track progress.

AI-Complete Task

- Image Captioning ?
  - A coarse scene-level understanding of an image paired with word n-gram statistics suffices to generate reasonable image captions
  - Automatic evaluation is still a difficult and open research problem
  - Suggests image captioning may not be as “AI-complete” as desired.


a man is playing tennis on a tennis court


a train is traveling down the tracks at a train station


a cake with a slice cut out of it

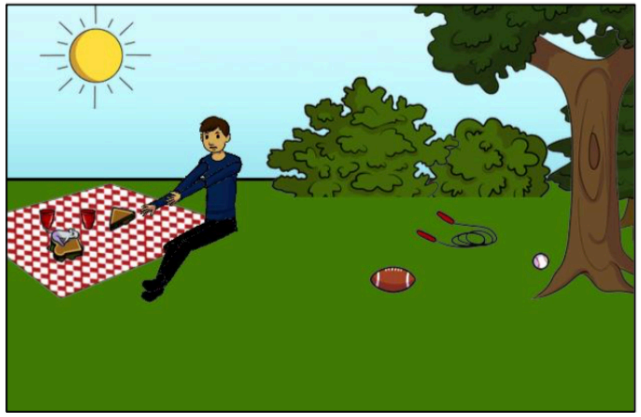

a bench sitting on a patch of grass next to a sidewalk

Sourec: https://github.com/karpathy/neuraltalk2

AI-Complete Task: VQA

- Visual questions selectively target different areas of an image , therefore, needs a more detailed understanding of the image
- Require a potentially vast set of AI capabilities to answer

Fine Grained Recognition:
What kind of cheese is on the pizza?



Object Detection:
How many balls are there on the ground?

Commonsense Reasoning :
Does this person have 20/20 vision?

Activity Recognition:
Is the woman sitting?

AI-Complete Task: VQA

- we can easily evaluate a proposed algorithm by the number of questions it answers correctly
  - Answers to many questions is simply “yes” or “no”
  - Since questions about images often tend to seek specific information, simple one- to-three word answers are sufficient for many questions.
  - Or a closed set of answers can be provided in a multiple-choice format

Related Work

- Questions and answers were generated from a limited predefined vocabulary and object categories. [1] [2]
- Proposed combining LSTM for the question with a CNN for the image to generate an answer. [3]
- Generated abstract scenes to capture visual common sense relevant to answering (purely textual) fill-in-the-blank and visual paraphrasing questions. [4]

[1] D. Geman, S. Geman, N. Hallonquist, and L. Younes. A Visual Turing Test for Computer Vision Systems. In *PNAS*, 2014

[2] M. Malinowski and M. Fritz. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *NIPS*, 2014

[3] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015.

[4] X. Lin and D. Parikh. Don't Just Listen, Use Your Imagination: Leveraging Visual Common Sense for Non-Visual Tasks. In *CVPR*, 2015.

Related Work

- Collected questions & answers in Chinese (later translated to English) for COCO images.[1]
- Automatically generated four types of questions (object, count, color, location) using COCO captions. [2]
- Text-based Q&A, sentence completion [3]
- Describing Visual Content: Image tagging, image captioning and video captioning [4][5]

[1] H. Gao, J. Mao, J. Zhou, Z. Huang, and A. Yuille. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*, 2015.
[2] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015.
[3] M.Richardson,C.J.Burges,andE.Renshaw.MCTest:AChallenge Dataset for the Open-Domain Machine Comprehension of Text. In *EMNLP*, 2013.
[4] J.Deng,A.C.Berg,andL.Fei-Fei.HierarchicalSemanticIndexing for Large Scale Image Retrieval. In *CVPR*, 2011.
[5] G.Kulkarni,V.Premraj,S.L.SagnikDharand,Y.Choi,A.C.Berg, and T. L. Berg. Baby Talk: Understanding and Generating Simple Image Descriptions. In *CVPR*, 2011

Contributions

- Prepared a dataset containing 0.25M images, 0.76M questions, and 10M answers and analyzed it.

- Provided numerous baselines and methods for VQA and compared them with human performance.

Part 1: Dataset

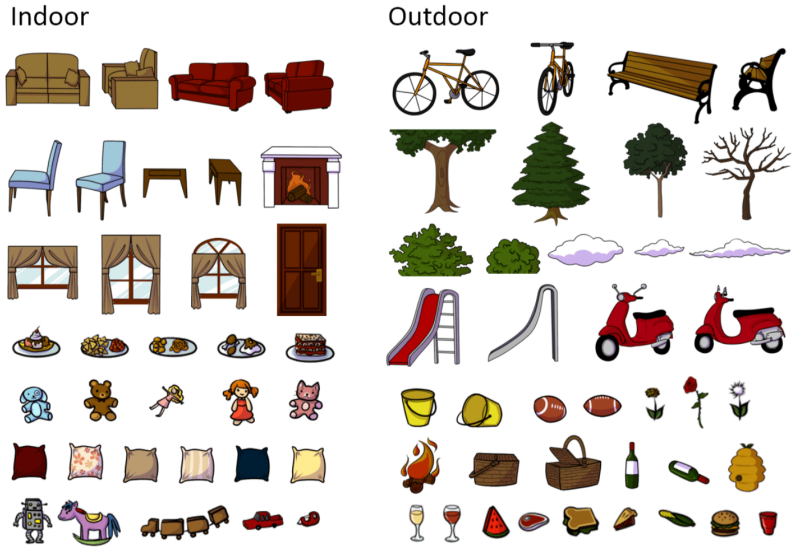
VQA Dataset: Images

- Real Images
  - Used 204,721 images from the Microsoft Common Objects in Context (MS COCO)
  - Selected images with multiple objects and rich contextual information.




VQA Dataset: Images

- Abstract Scenes
  - 50,000 scenes
  - To enable research focused on high-level reasoning required for VQA, but not the low-level vision tasks.
  - Contains 20 “paperdoll” human models with adjustable limbs and over 100 objects and 31 animals in various pose



VQA Dataset: Questions

- User Interface for collecting “interesting” questions
- For each image/scene: gathered 3 questions from unique workers
- Subjects were shown the previous questions already asked for that image to increase the question diversity.
- Subjects were instructed to ask questions that *require* the image to correctly answer and not be answerable using just commonsense information

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

Updated instructions: Please read carefully

Hide Show

We have built a smart robot. It understands a lot about scenes. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene type (e.g., kitchen, beach), people's expressions and poses, and properties of objects (e.g., the color of objects, their texture). Your task is to stump this smart robot! In particular, it already knows answers to some questions about this scene. We will tell you what these questions are.

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.
IMPORTANT: The question should be about this scene. That is, the human should need the image to be able to answer the question -- the human should not be able to answer the question without looking at the image.

Your work will get rejected if you do not follow the instructions below:

- Do not ask questions that are similar to the ones listed below each image. As mentioned, the robot already knows the answers to those questions for the scene in this image. Please ask about something different.
- 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 the scene in 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 scenes. Ask questions specific to the scene in each image.

Below is a list of questions the smart robot can already answer. Please ask a different question about this scene that a human can answer *if* looking at the scene in the image (and not otherwise), but would stump this smart robot:

Q1: What is unusual about this mustache? (The robot already knows the answer to this question.)

Q2: What is her facial expression? (The robot already knows the answer to this question.)

Q3: Write your question, different from the questions above, here to stump this smart robot.

prev next **Page 2/3**

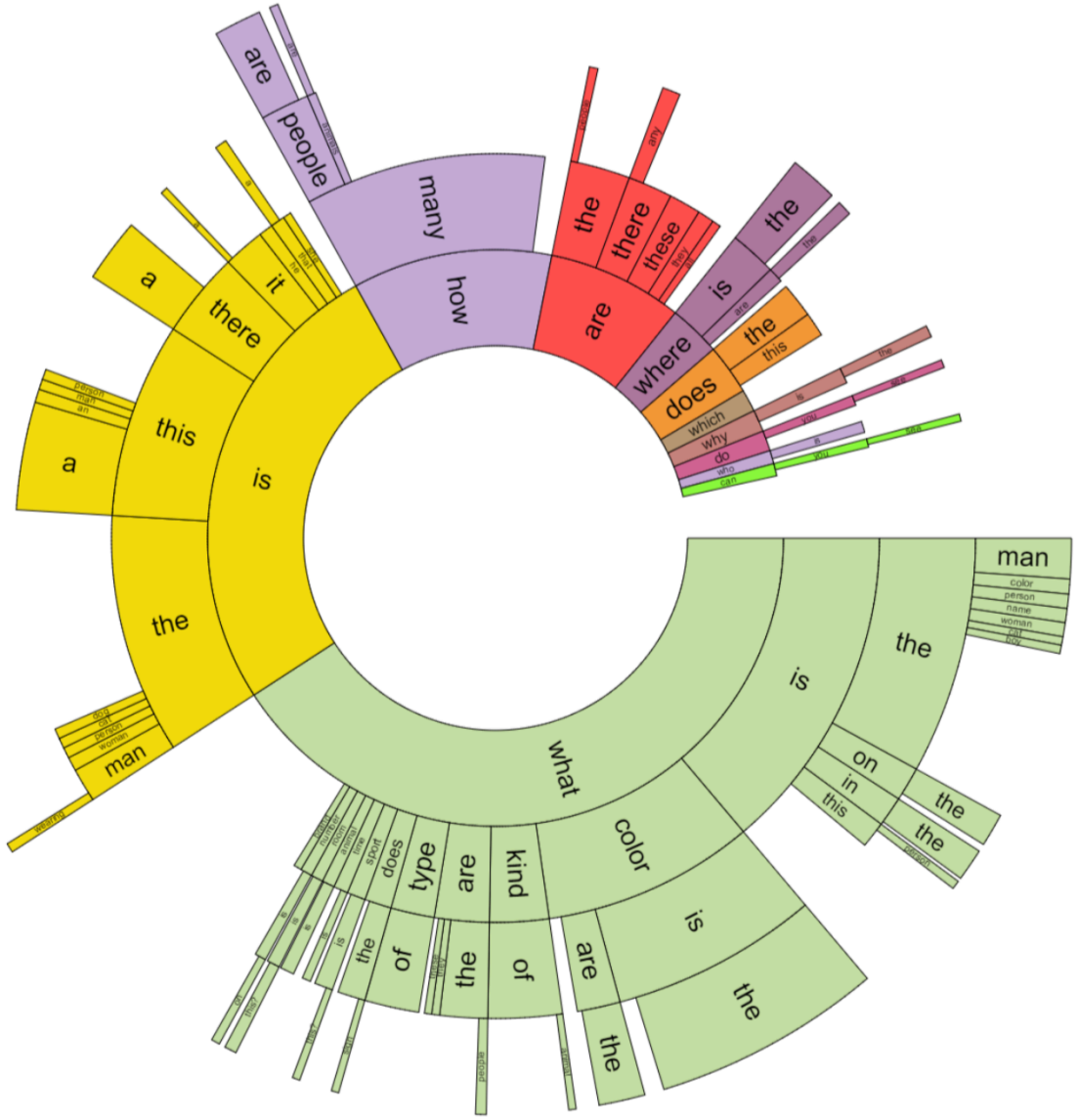
Answers

- To handle discrepancies, they gather *10 answers for each question from unique workers*
- Two modalities of answering questions:
  - Open Answer: Answer in short phrase
  - Multiple choice: 18 candidate answers are created for each question containing the correct answer + plausible + popular + random answer choices
- Selected **ground truth answers** using the following accuracy metric:

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

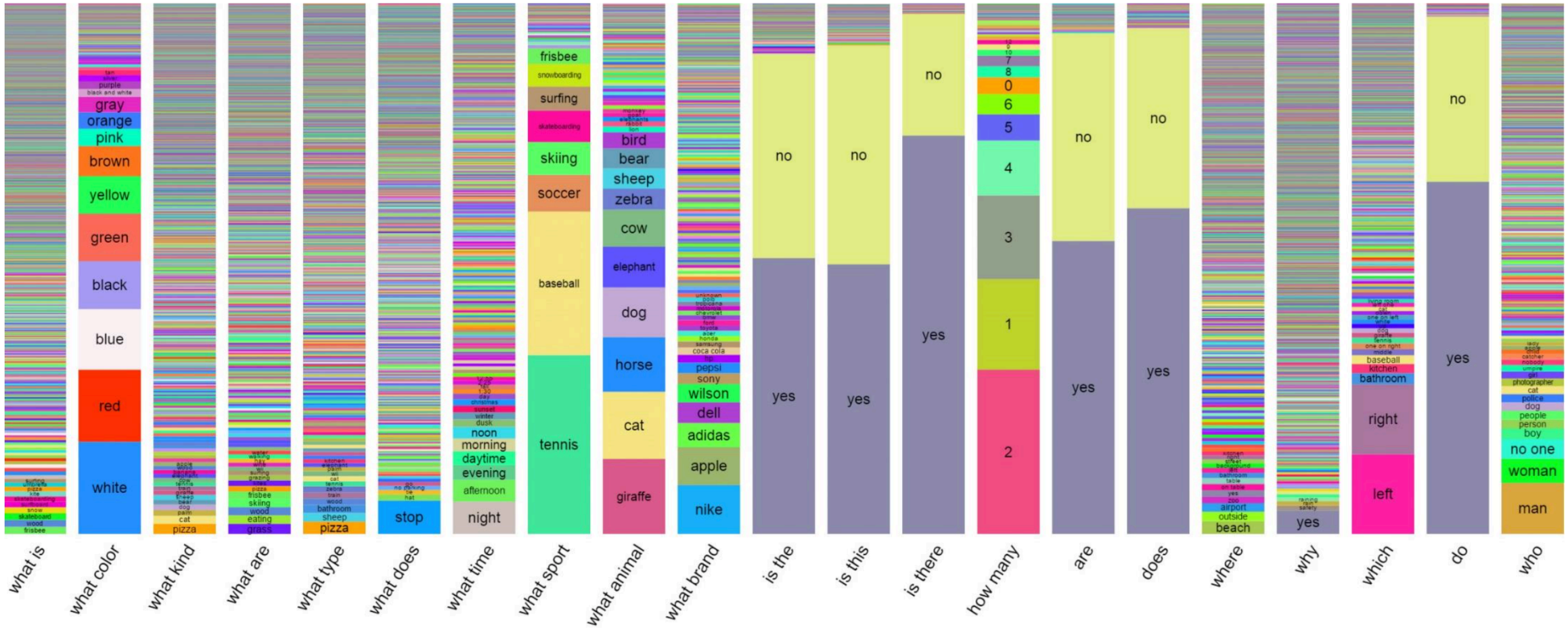
  - *i.e.*, an answer is deemed 100% accurate if at least 3 workers provided that exact answer

VQA Dataset Analysis: Types of Questions



- Distribution of questions by their first four words for a random sample of 60K questions
- The ordering of the words starts towards the center and radiates outwards.
- The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show.

VQA Dataset Analysis: Types of Answers



"Is the...", "Are...", "Does. . . " : typically answered using "yes" and "no" as answers.

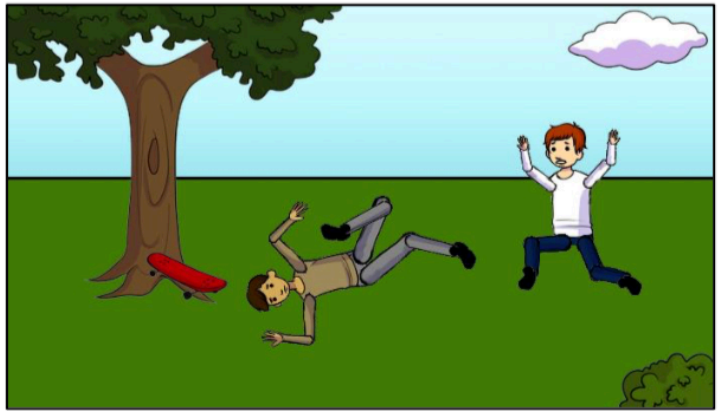
"What is. . . " and "What type. . . " : have a rich diversity of responses.

"What color..." or "Which..." have specialized responses, such as colors, or "left" and "right".

Commonsense Knowledge: Is the Image Necessary?

- Some questions can be answered correctly using commonsense knowledge alone without the need for an image
  - What is the color of banana?
- Asked 3 subjects to answer the questions *without seeing the image*

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



Do you think the boy on the ground has broken legs?	yes yes yes	no no yes
Why is the boy on the right freaking out?	his friend is hurt other boy fell down someone fell	ghost lightning sprayed by hose

Black: Question

Green: answers given when looking at the image

Blue: answers given when not looking at the image

Commonsense Knowledge: Is the Image Necessary?

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

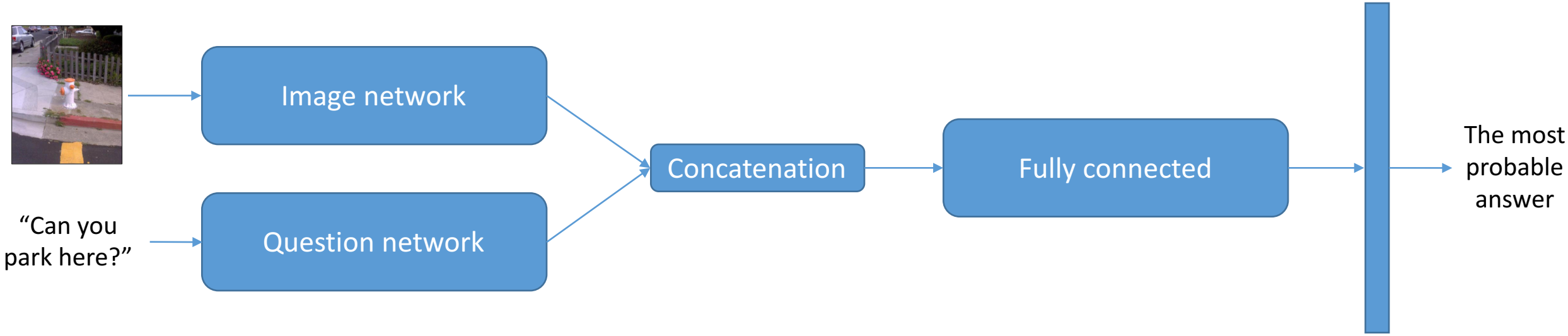
- For “yes/no” questions, the human subjects respond better than chance.
- For other questions, humans are only correct about 21% of the time.

- Demonstrates that understanding the visual information is critical to VQA and that commonsense information **alone is not sufficient**.

Part 2: Methods

Methods

- A 2-channel vision (image) + language (question) model
- Choosing the top K = 1000 most frequent answers as possible outputs



Methods – Image Network

- **I:** The activations from the last hidden layer of VGGNet [1] are used as 4096-dim image embedding, followed by an MLP with output layer of 1024-dim.
- **norm I:** These are $\ell 2$ normalized activations from the last hidden layer of VGGNet [1], followed by an MLP with output layer of 1024-dim.

[1] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Methods – Question Network

- **Bag-of-Words Question (BoW Q):**
  - [A bag-of-words representation based on the top 1,000 words in the questions] + [the top 10 first, second, and third words of the questions] ==> 1030-dim embedding

- **LSTM Q:**
  - Each question word is encoded with 300-dim embedding by a fully-connected layer + tanh non-linearity ==> fed into an LSTM with one hidden layer ==> 1024-dim embedding.

- **deeper LSTM Q:**
  - 300-dim embedding by a fully-connected layer + tanh non-linearity ==> fed in to an LSTM with two hidden layers: 2048-dim embedding ==> followed by a fully-connected layer + tanh non-linearity ==> 1024-dim embedding

Methods – Concatenation and MLP

- Concatenation for **BoW Q + I:**
  - Simple concatenation of the **BoW Q** and **I** embeddings.

- Concatenation for **LSTM Q** + **I**, **deeper LSTM Q** + **norm I**:
  - The image embedding is fed to a fully-connected layer + tanh non-linearity to match the LSTM embedding of the question. The transformed image and LSTM embeddings (being in a common space) are then fused via element-wise multiplication.

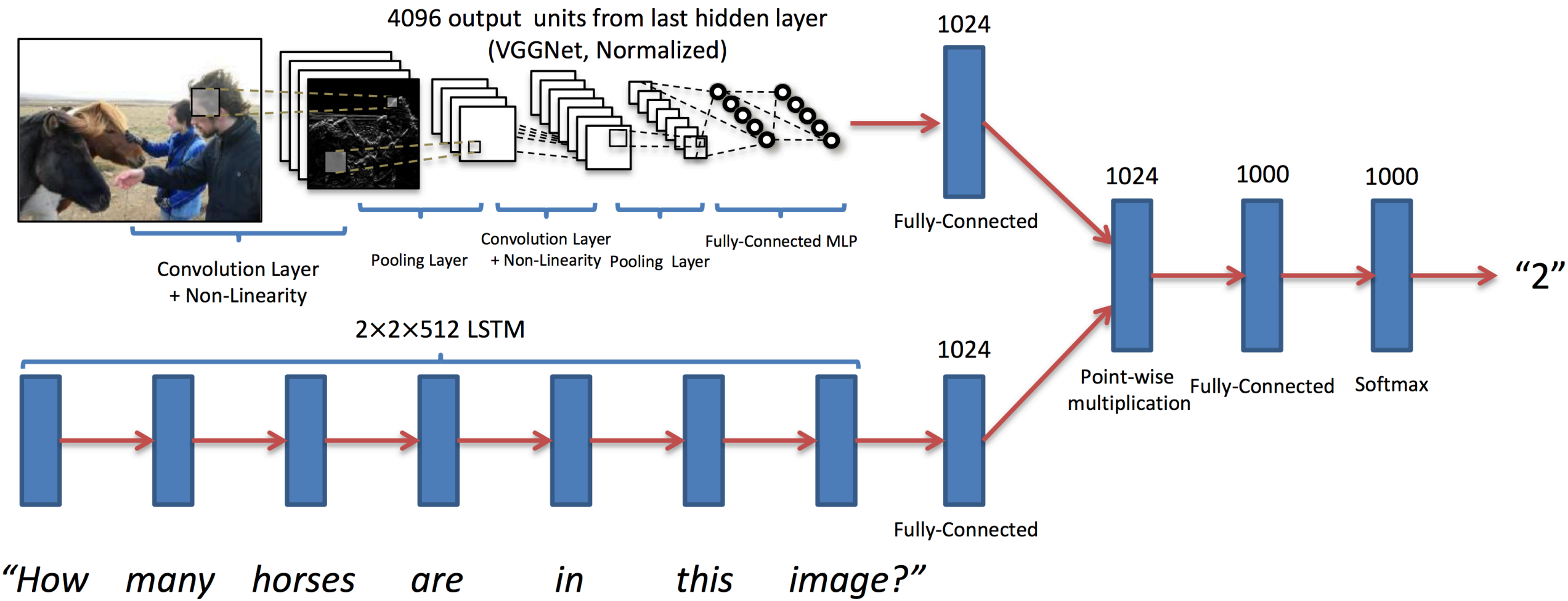
- After concatenation:
  - The combined embedding is passed to an MLP with 2 hidden layers and 1000 hidden units (dropout 0.5) in each layer with tanh non-linearity, followed by a softmax layer to obtain a distribution over K answers.

Methods – Output and Decision

- Output:
  - 1000 nodes followed by softmax non-linearity

- Decision:
  - open-ended:
    - Selecting the answer with highest activation from all possible K answers
  - multiple-choice:
    - Picking the answer that has the highest activation from the potential answers

Methods – **deeper LSTM Q + norm I**



Evaluation

- Accuracy:
  - $Acc = \frac{\#correct\ answers}{\#all\ the\ trials}$

- Baselines:
  - prior (“yes”):
    - answer to all the questions is “yes”
  - per Q-type prior:
    - clustering the questions to different types beforehand ==> classifying the incoming question and finding its type ==> finding the most popular answer to that type of question
  - nearest neighbor:
    - finding the k nearest neighbors of the incoming question in the training set. also, finding the most similar image with the incoming image (among all the k images associated with those k questions) ==> the most common ground truth answer for this selected question and image pair is selected as the answer

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

- Some points:
  - The vision-alone model (**I**) performs rather poorly
  - The language-alone methods (per Q-type prior, BoW Q, LSTM Q) perform surprisingly well ==> due to the language-model exploiting subtle statistical priors about the question types (“Color of banana” => Yellow)
  - **deeper LSTM Q + norm I** model is able to outperform both the vision-alone and language-alone baselines.
  - Results on multiple-choice are better than open-ended
  - All methods are significantly worse than human performance
  - “yes/no" has a big bias on “yes”

Q-type Results

Question Type	Open-Ended K = 1000 Q	Open-Ended K = 1000 Q + I	Open-Ended K = 1000 Q + C	Open-Ended Human Q	Open-Ended Human Q + I	Human Age To Be Able To Answer	Commonsense To Be Able To Answer (%)
what is (13.84)	23.57	34.28	43.88	16.86	73.68	09.07	27.52
what color (08.98)	33.37	43.53	48.61	28.71	86.06	06.60	13.22
what kind (02.49)	27.78	42.72	43.88	19.10	70.11	10.55	40.34
what are (02.32)	25.47	39.10	47.27	17.72	69.49	09.03	28.72
what type (01.78)	27.68	42.62	44.32	19.53	70.65	11.04	38.92
is the (10.16)	70.76	69.87	70.50	65.24	95.67	08.51	30.30
is this (08.26)	70.34	70.79	71.54	63.35	95.43	10.13	45.32
how many (10.28)	43.78	40.33	47.52	30.45	86.32	07.67	15.93
are (07.57)	73.96	73.58	72.43	67.10	95.24	08.65	30.63
does (02.75)	76.81	75.81	75.88	69.96	95.70	09.29	38.97
where (02.90)	16.21	23.49	29.47	11.09	43.56	09.54	36.51
is there (03.60)	86.50	86.37	85.88	72.48	96.43	08.25	19.88
why (01.20)	16.24	13.94	14.54	11.80	21.50	11.18	73.56
which (01.21)	29.50	34.83	40.84	25.64	67.44	09.27	30.00
do (01.15)	77.73	79.31	74.63	71.33	95.44	09.23	37.68
what does (01.12)	19.58	20.00	23.19	11.12	75.88	10.02	33.27
what time (00.67)	8.35	14.00	18.28	07.64	58.98	09.81	31.83
who (00.77)	19.75	20.43	27.28	14.69	56.93	09.49	43.82
what sport (00.81)	37.96	81.12	93.87	17.86	95.59	08.07	31.87
what animal (00.53)	23.12	59.70	71.02	17.67	92.51	06.75	18.04
what brand (00.36)	40.13	36.84	32.19	25.34	80.95	12.50	41.33

Q-type Results

- Some points:
  - For question types that require more reasoning (e.g. “Is the” or “How many”), the scene-level image features do not provide any additional information
  - For questions that can be answered using scene-level information (e.g. “What sport”), we do see an improvement.
  - For all question types, the results are worse than human accuracies.

Age-related Results

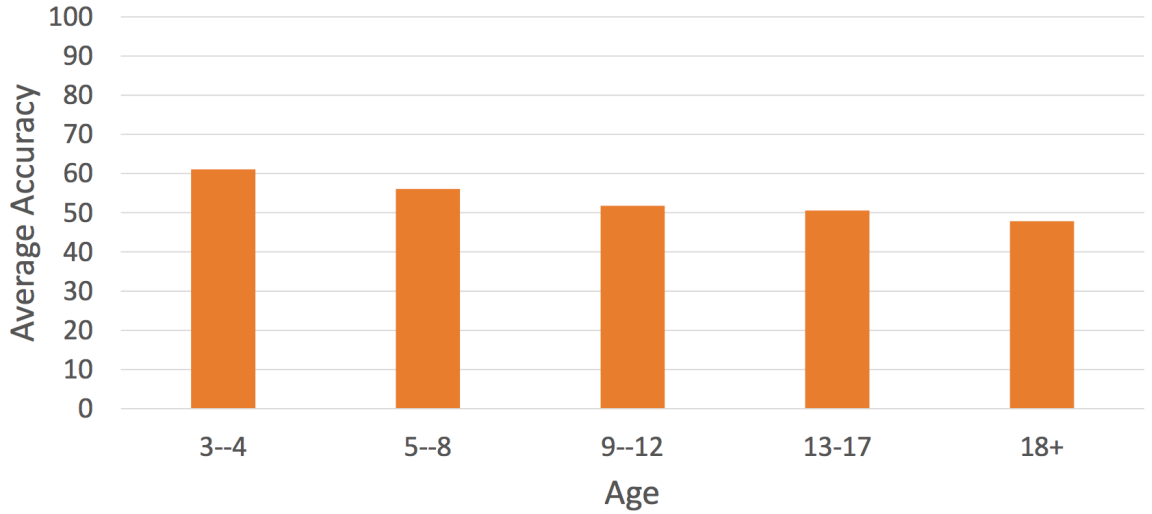


Fig. 11: Pr (system is correct | age of question) on the VQA validation set. System refers to our best model (deeper LSTM Q + norm I).

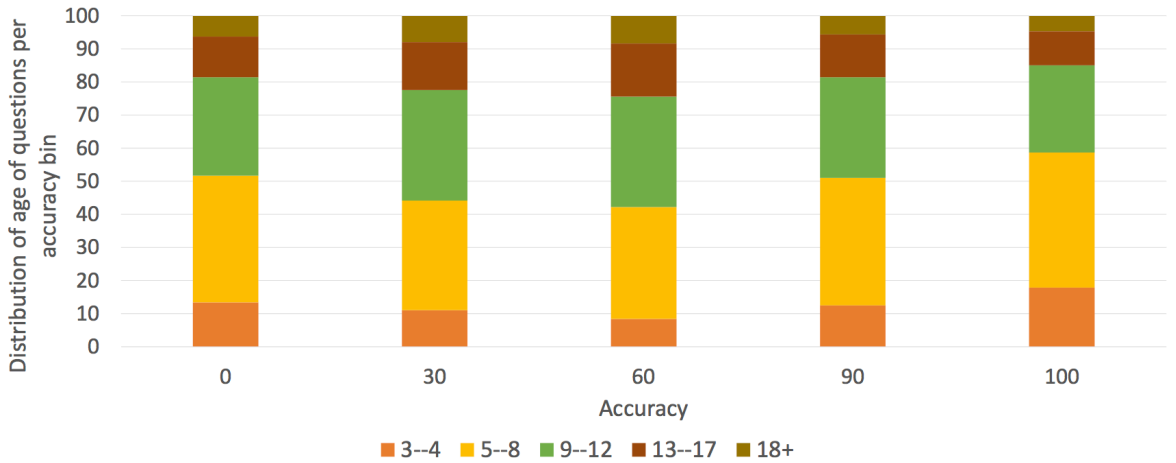


Fig. 12: Pr (age of question | system is correct) on the VQA validation set. System refers to our best model (deeper LSTM Q + norm I).

A-type Results

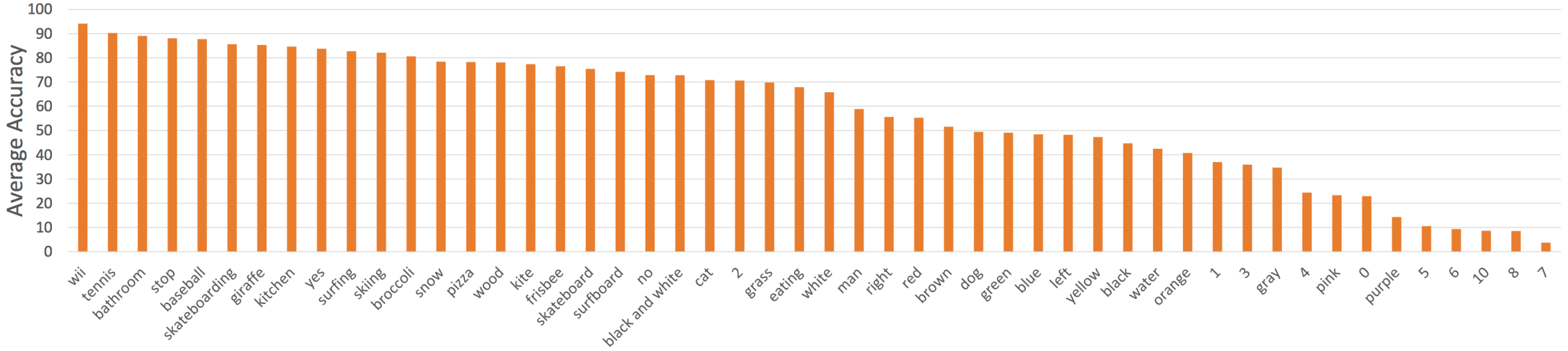


Fig. 9: Pr (system is correct | answer) for 50 most frequent ground truth answers on the VQA validation set (plot is sorted by accuracy, not frequency). System refers to our best model (deeper LSTM Q + norm I).

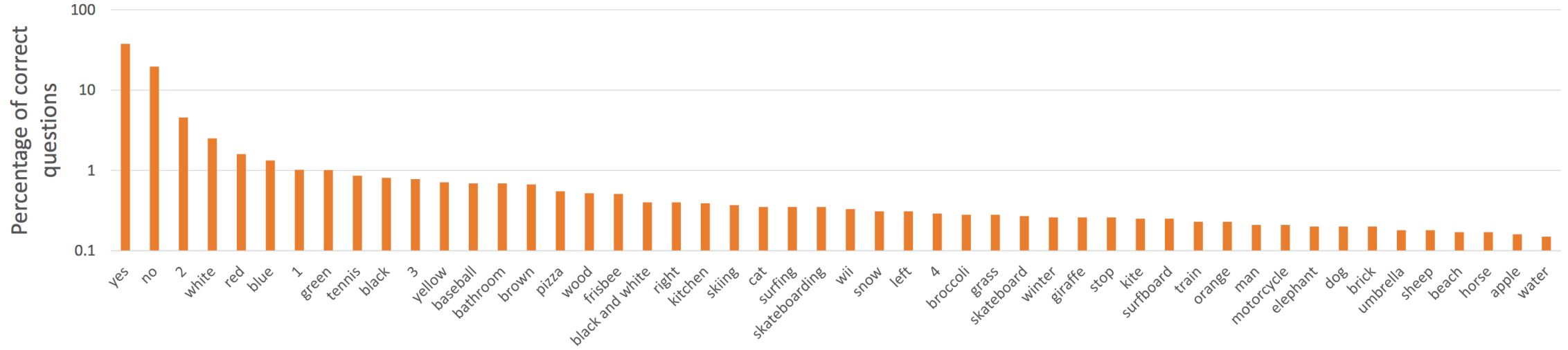


Fig. 10: Pr (answer | system is correct) for 50 most frequently predicted answers on the VQA validation set (plot is sorted by prediction frequency, not accuracy). System refers to our best model (deeper LSTM Q + norm I).

Part 3: Demo

Demo

- http://cloudcv.org/vqa/
- http://complianceandauditingservices.com/wp-content/uploads/2015/07/business-man-in-office-BE-1.png

What is she playing?
Tennis (0.9972)

What is she holding in her hand?
Tennis racket (0.7395)

What is the color of her shoes?
White (0.48520)

What is the color of her shoe?
Black (0.2661)

How many people are there in the image?
3 (0.2094)

Are they sad?
No (0.8915)

What are the people doing?
Playing frisbee (0.0210)

Where are the people standing?
Frisbee (0.0541)

Where are they?
Park (0.0669)

Are they happy?
Yes (0.9536)

Why are they upset?
Happy (0.1090)

What are they doing?
Skateboarding (0.0480)

Code and implementation

- https://github.com/abhshkdz/neural-vqa
- https://github.com/VT-vision-lab/VQA_LSTM_CNN

Suggestions

- Solving the bias on questions
- Balancing the dataset
- Fine-tuning the last layer of VGGNet to remove the bias on the original label
- Creating a generative network rather than a classification one

Any questions?