VQA: Visual Question Answering

Stanislaw Antol Aishwarya Agrawal Jiasen Lu Margaret Mitchell
Dhruv Batra C. Lawrence Zitnick Devi Parikh

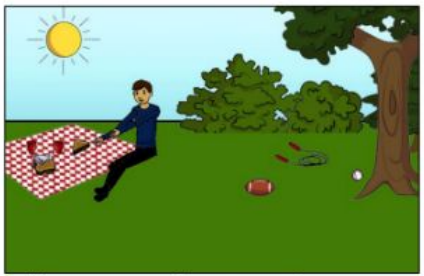
Presented by: Paul Choi

Visual Question Answering (VQA)

Task: Given an image and a natural language question about the image, provide an accurate natural language answer.

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

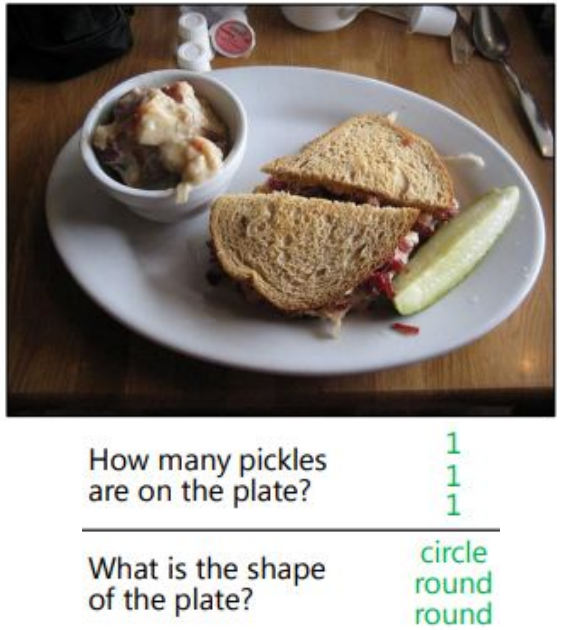
Does it appear to be rainy?
Does this person have 20/20 vision?

Motivation for VQA

- Push the AI state-of-the-art
- Requires multi-modal knowledge
  - Computer Vision
  - Natural Language Processing
  - Knowledge Representation and Reasoning
- Simple to quantitatively evaluate
- Useful applications
  - Helping the visually-impaired by answering questions about a scene
  - Assisting intelligence analysts extract visual information

Contributions

- A VQA dataset containing 250K images, 760K questions, and 10M answers.
  - Previous largest: 2.6K images
  - Open-ended, free-form questions and answers instead of fixed vocabulary
- Analysis of the dataset
  - What types of questions are being asked?
  - What types of answers are given?
- Baseline approaches



Dataset collection

Real Images: 200K images from MS COCO

- Multiple objects per scene, rich contextual information

Abstract Scenes: 50K scenes generated using clipart

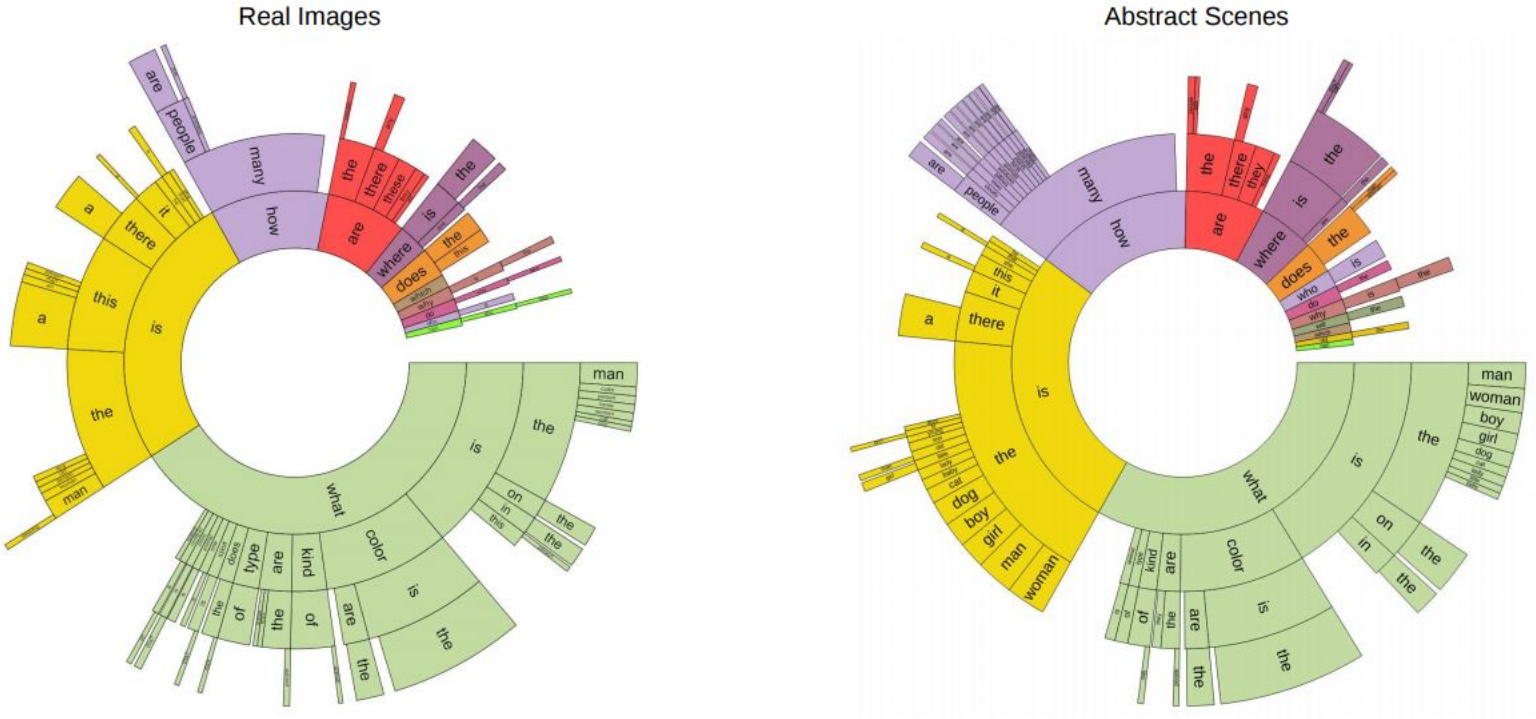
- Less visual noise, makes visual recognition part of task easier

Dataset collection

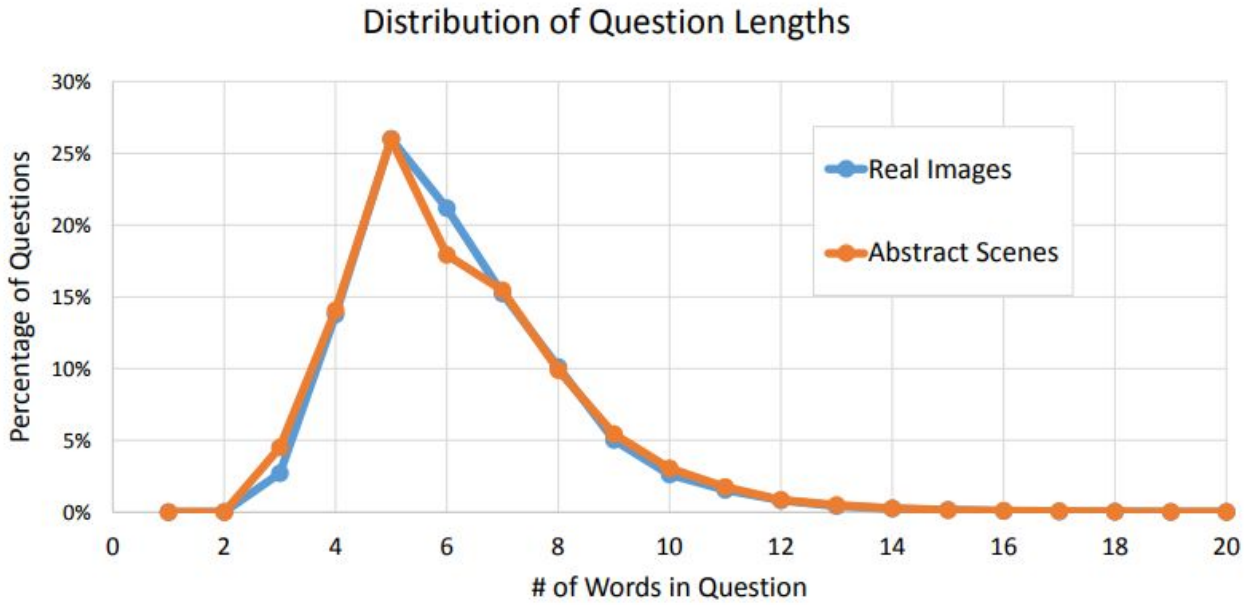
- **Captions**
  - Five single-sentence captions collected for each abstract scene
  - MS COCO already provides five captions for its images
  - *Ex: A child catches a disc.*
- **Questions**
  - Three unique questions gathered for each image from unique AMT workers
  - Questions must require the image to answer
  - *Ex: What color is the disc?*
- **Answers**
  - Ten answers for each question from unique workers
  - *Ex: Blue*

Dataset Analysis

Distribution of Questions

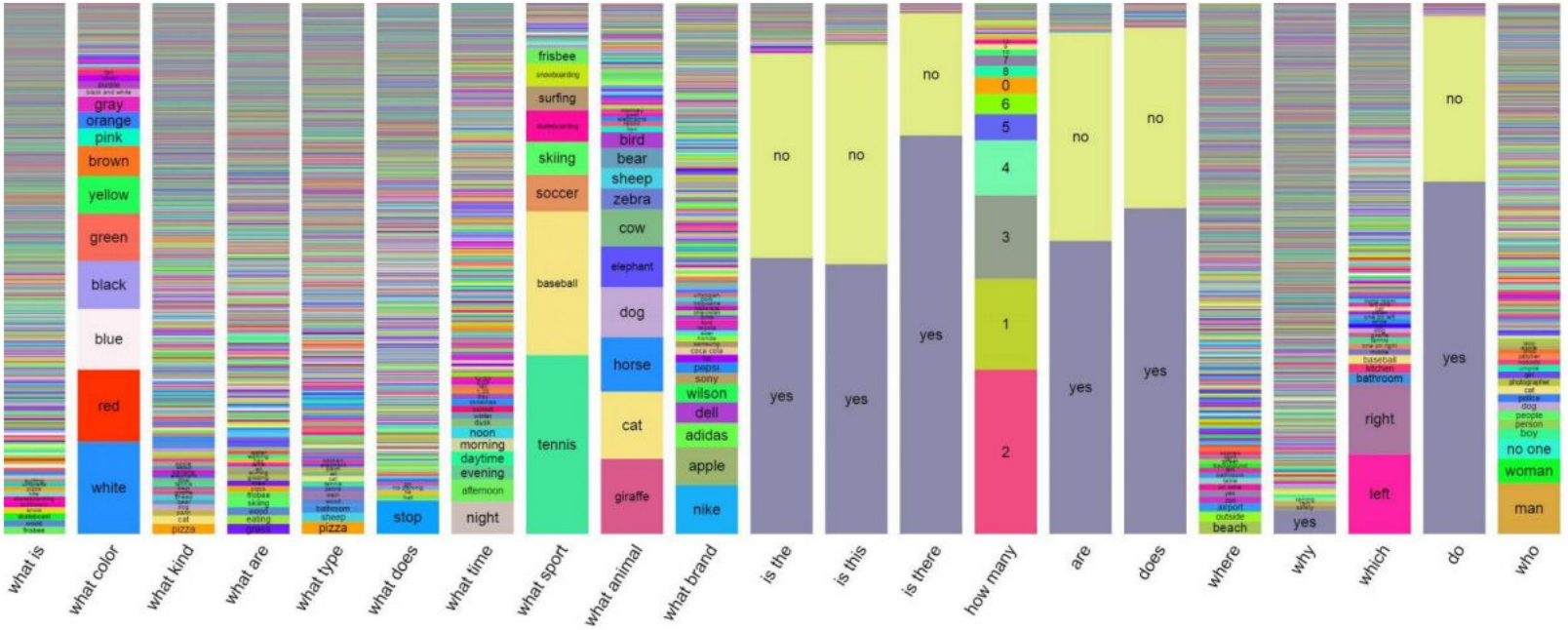


Both distributions are very similar, suggesting abstract scenes are a feasible replacement for real images in VQA.



Again, very similar distributions between real and abstract images..

Answers with Images



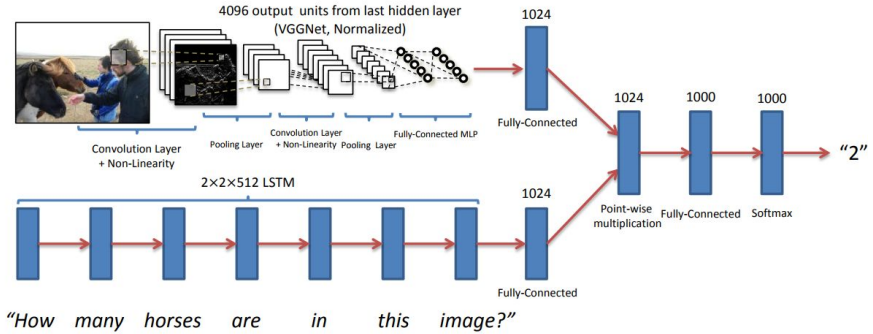
Baselines

Possible inputs

- **Question channel**
  a. Bag-of-words representation of top 1000 words in all questions
  b. LSTM embedding (1024-dim)
- **Caption channel:** Bag-of-words representations of top 1000 words in all captions
- **Image channel:** Last hidden layer of VGG (4096-dim)

The model

- Multi-layer perceptron with 2 hidden layers (1000 hidden units each)
- Input options
  - Concatenate **image** embedding and **bag-of-words question** embedding
  - Concatenate **image** embedding and **bag-of-words caption** embedding
  - Element-wise multiplication of **image** embedding with **LSTM question** embedding
- Output: softmax distribution over *K* possible answers
  - Authors use top *K* = 1000 answers in the dataset as possible answers



Baseline Results

	All	Yes/No	Number	Other
Question	48.09	75.66	36.70	27.14
Image	28.13	64.01	00.42	03.77
Q+I	52.64	75.55	33.67	37.37
LSTM Q	48.76	78.20	35.68	26.59
LSTM Q+I	53.74	78.94	35.24	36.42
Caption	26.70	65.50	02.03	03.86
Q+C	54.70	75.82	40.12	42.56

Even without the image, all methods perform far better than chance on Yes/No questions!

Question Type	K = 1000			Human	
	Q	Q + I	Q + C	Q	Q + I
what is (13.84)	23.57	34.28	43.88	16.86	73.68
what color (08.98)	33.37	43.53	48.61	28.71	86.06
what kind (02.49)	27.78	42.72	43.88	19.10	70.11
what are (02.32)	25.47	39.10	47.27	17.72	69.49
what type (01.78)	27.68	42.62	44.32	19.53	70.65
is the (10.16)	70.76	69.87	70.50	65.24	95.67
is this (08.26)	70.34	70.79	71.54	63.35	95.43
how many (10.28)	43.78	40.33	47.52	30.45	86.32
are (07.57)	73.96	73.58	72.43	67.10	95.24
does (02.75)	76.81	75.81	75.88	69.96	95.70
where (02.90)	16.21	23.49	29.47	11.09	43.56
is there (03.60)	86.50	86.37	85.88	72.48	96.43
why (01.20)	16.24	13.94	14.54	11.80	21.50
which (01.21)	29.50	34.83	40.84	25.64	67.44
do (01.15)	77.73	79.31	74.63	71.33	95.44
what does (01.12)	19.58	20.00	23.19	11.12	75.88
what time (00.67)	8.35	14.00	18.28	07.64	58.98
who (00.77)	19.75	20.43	27.28	14.69	56.93
what sport (00.81)	37.96	81.12	93.87	17.86	95.59
what animal (00.53)	23.12	59.70	71.02	17.67	92.51
what brand (00.36)	40.13	36.84	32.19	25.34	80.95

- Image features don’t help for questions which require deeper reasoning
- They do help for identification questions (“what…”)

Demo!

https://vqa.cloudcv.org/

Strengths and Weaknesses

- Pros
  - Challenging problem with useful applications
  - Trumps previous datasets in quantity of data and complexity of questions
  - Approaching human performance in VQA will lead us to more “complete”, human-assistant-like AIs
- Cons
  - Biases in the dataset can skew results
    - Language priors can give easy accuracy gains but mask deficiencies of the method
  - Doesn’t account for synonyms, pluralities in answers