Visual Question Answering

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

(*ICCV 2015)*

Yanan Liu

04/12/2018

Outline

- Problem Statement
- Motivations
- The VQA Dataset
- Discussion of the Dataset
- Baselines
- Conclusion

What is VQA

Given an image and a natural language question about the image, the task is to provide a natural language answer.

VQA Demo

what are in the picture

Submit

Predicted top-5 answers with confidence:

elephants	99.133%
elephant	0.707%
people	0.017%
trees	0.010%
zebras	0.008%

what is in the picture

Submit

Predicted top-5 answers with confidence:

elephant	89.543%
elephants	9.952%
horse	0.126%
man	0.053%
people	0.027%

what are the things in the picture doing

Submit

Predicted top-5 answers with confidence:

walking	61.979%
standing	26.749%
eating	4.109%
playing	3.278%
drinking	0.572%

is it rainy

Submit

Predicted top-5 answers with confidence:

no	82.649%
yes	17.351%
day	0.000%
cloudy	0.000%
0	0.000%

Motivations

- Applications: helping the visually impaired
- “AI completeness” (image captioning can be too simple)
- A task has to:
  - Be open-ended
  - Require multi-modal knowledge (CV, NLP, KR)
  - Suitable for quantitative evaluation

Multi-modal knowledge

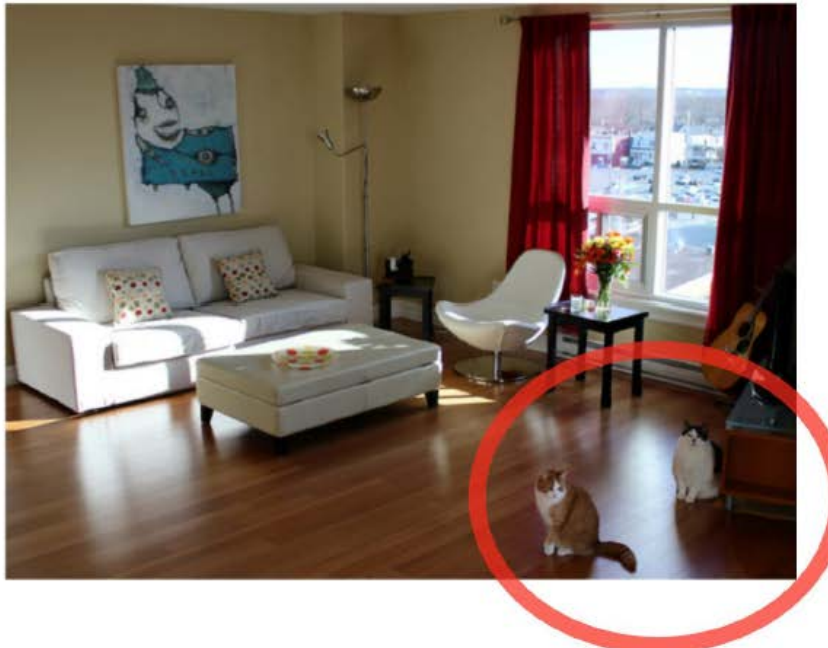
Quantitative Evaluation Metric

- Open Ended Answering V.S. Multiple Choice

- Suitable for Self-Evaluation:
  - Multiple choice
  - Simple answer
    - Yes or no
    - Number

How many cats are there?

(a) One
(b) Two
(c) Three
(d) Four



VQA Dataset

- 250k Images (200k MS COCO + 50k abstract images)
- Each image with 3 questions (750k questions)
- Each question with ~10 answers (10M answers)
- Train/Val/Test split: 2:1:2

VQA-real

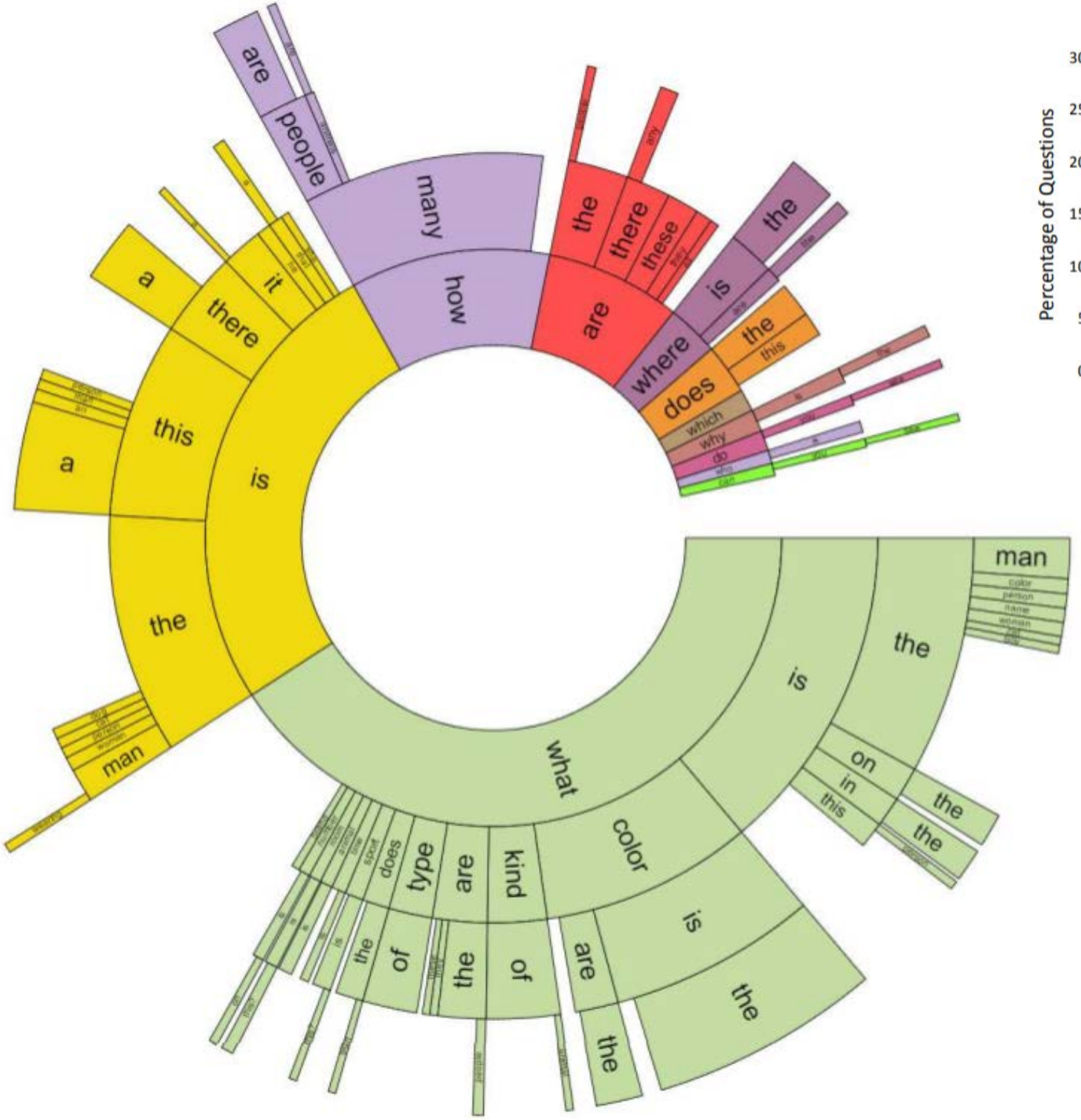
Q: Where are the magazines in this picture ?

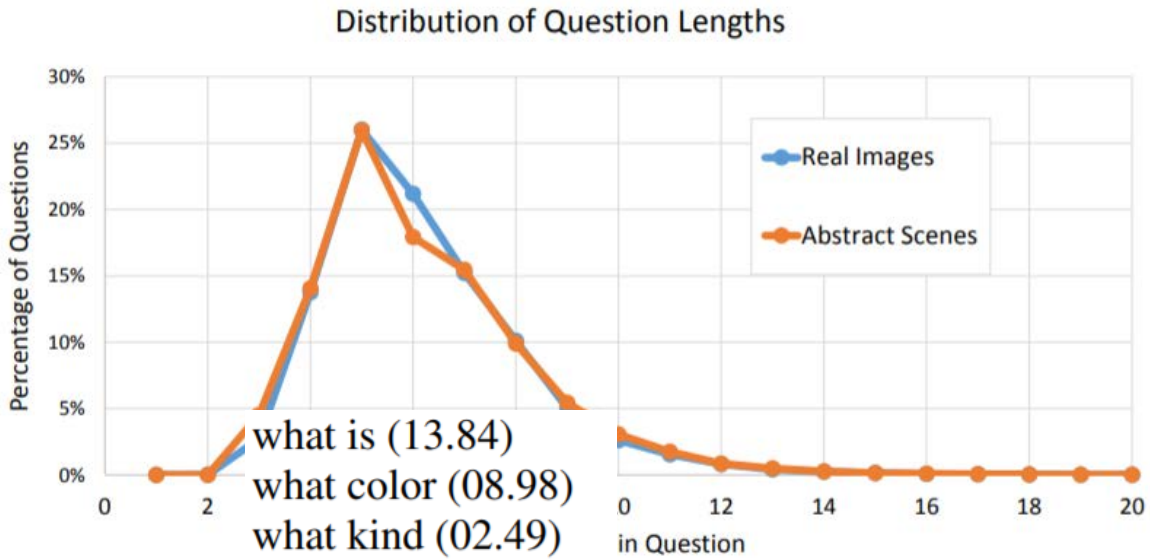
A: On stool, stool, on stool, on bar stool, on table, stool, on stool, on chair, on bar stool, stool

VQA-abstract

Q: Where are the flowers ?

A: near tree, tree, around tree, tree, by tree, around tree, around tree, grass, beneath tree, base of tree





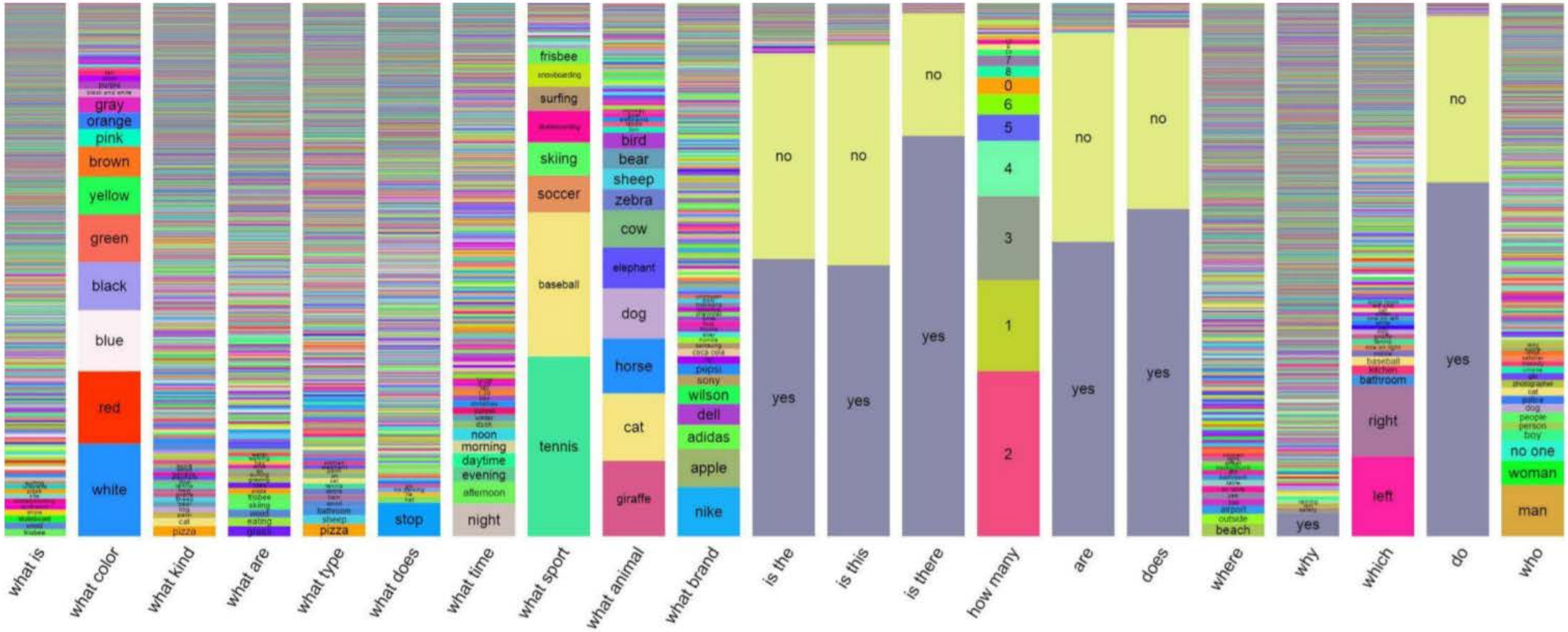
what is (13.84)
what color (08.98)
what kind (02.49)
what are (02.32)
what type (01.78)
is the (10.16)
is this (08.26)
how many (10.28)
are (07.57)
does (02.75)
where (02.90)
is there (03.60)
why (01.20)
which (01.21)
do (01.15)
what does (01.12)
what time (00.67)
who (00.77)
what sport (00.81)
what animal (00.53)

Distribution of Answers

- Most answers are short:
  - ~90% one word, ~6% two words, ~2.5% three words
- Brevity of the answers doesn’t mean the problem is easy:
  - 23,234 unique one-word answers

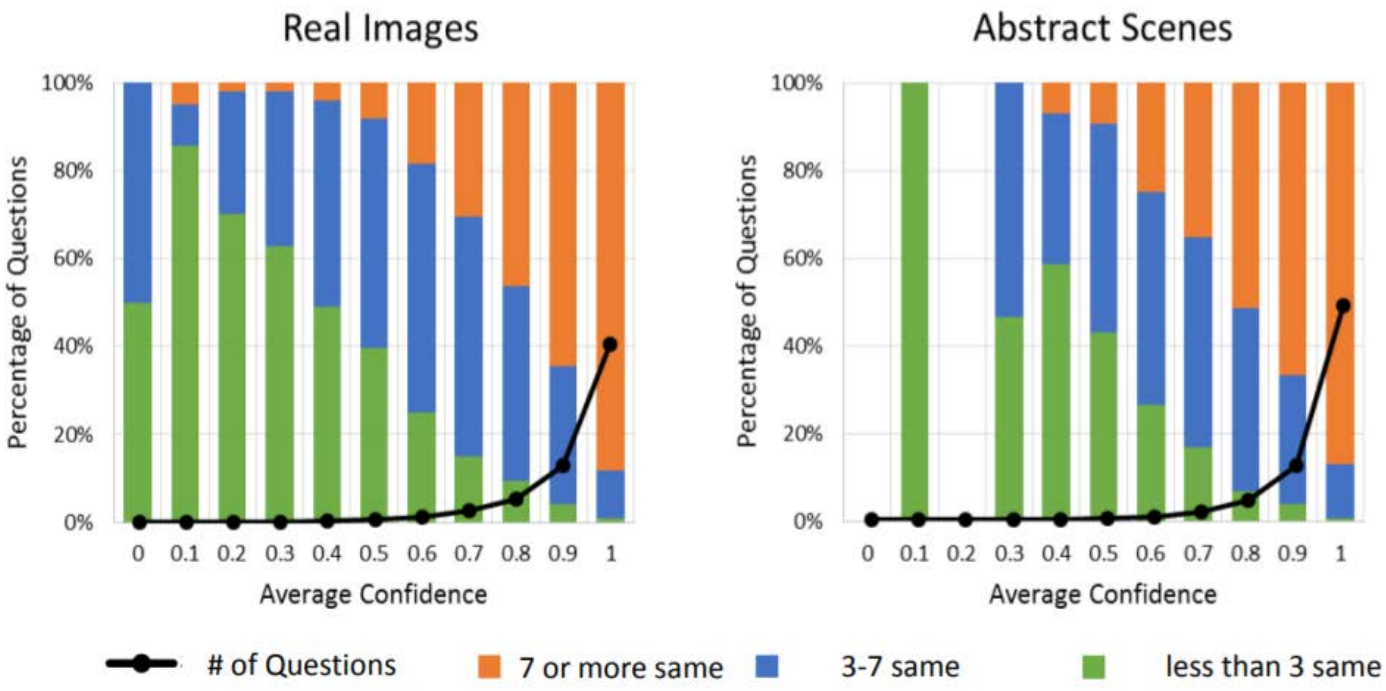
- ~40% are “yes”/”no” answers, among which ~56% are “yes”
- ~13% percent are numbers, among which “2” is the most popular

Distribution of Answers



More Discussion

- Subject has high confidence about their answers
- Inter-human agreement on the answers is high
- Common se
- Image is nee



Baselines

- Randomly choose an answer from top 1k answers: 0.12%
- Always select the most popular answer (“yes”): 29.72%
- Pick the most popular answer per question type: 36.18%

- Classification
  - Output: 1 of the top 1k answers(cover 82.67% answers in the train+val set)
  - Model: MLP, LSTM
  - Input: Question(bag of words), Caption(bag of words), Image(last hidden layer of VGGNet), Q+I, Q+C, Q+I+C

Baselines

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

	Open-Answer				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
Question	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
Image	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
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
Caption	26.70	65.50	02.03	03.86	28.29	69.79	02.06	03.82
Q+C	54.70	75.82	40.12	42.56	59.85	75.89	41.16	52.53

Conclusion

- The task of VQA is introduced.
  - Open-ended
  - Multi-modal knowledge
  - Suitable for self-evaluation
- A VQA dataset is provided.
- Annual challenge, workshop, etc.

Further Reading

- Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel. "Visual Question Answering: A Survey of Methods and Datasets." arXiv preprint arXiv:1607.05910 (2016).
- Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh. "VQA: Visual Question Answering." arXiv preprint arXiv:1505.00468 (2016).
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick. "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning." arXiv preprint arXiv:1612.06890 (2016).
- Kevin J. Shih, Saurabh Singh, Derek Hoiem. "Where To Look: Focus Regions for Visual Question Answering." arXiv preprint arXiv:1511.07394 (2016).
- Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Dan Klein. "Learning to Compose Neural Networks for Question Answering." arXiv preprint arXiv:1601.01705 (2016).
- Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, Marcus Rohrbach. "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding." arXiv preprint arXiv:1606.01847 (2016).
- Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh. "Hierarchical Question-Image Co-Attention for Visual Question Answering.". arXiv preprint arXiv:1606.00061 (2017).
- Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, Dhruv Batra. "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning." arXiv preprint arXiv:1703.06585 (2017).
- Yuke Zhu, Oliver Groth, Michael Bernstein, Li Fei-Fei. "Visual7W: Grounded Question Answering in Images." arXiv preprint arXiv:1511.03416 (2016).
- Allan Jabri, Armand Joulin, Laurens van der Maaten. "Revisiting Visual Question Answering Baselines." arXiv preprint arXiv:1606.08390 (2016).
- Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel. "Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources." arXiv preprint arXiv:1511.06973 (2016).