NEURAL NETWORK READING GROUP

VQA:
Visual Question Answering

E/15/043 Bhagya T.P.Y.
E/15/092 Ekanayake I.U.
E/15/187 Kulanjith G.D.
E/15/246 Opanayake R.L.

Group No: 02

PRESENTATION OUTLINE

KEY DISCUSSION POINTS

Overview of the research
Introduction
Related work
Dataset
Dataset Analysis
VAQ baseline and methods

WHAT?

IS THIS PAPER ?

FYP | Group No: 02

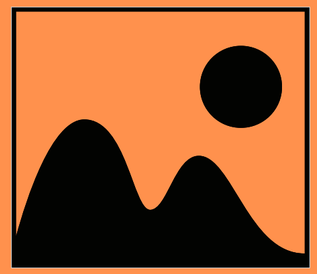
Name : **VQA: Visual Question Answering**

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

Institutes:
- **Virginia Tech**
- **Microsoft Research**

Publisher : **2015 IEEE International Conference on Computer Vision (ICCV)**

WHAT?



+ ?wh..?

✓ Answer ✓

THEY HAVE DONE ?

- **Open-ended Q&A**
- **Complex reasoning & detailed understanding**
- **Images=0.25M Q=0.75 A=10.0M**
- **Small Questions and Closed set of Answers (“yes” or “no” or small 1 to 3 words asnwers)**

INTRODUCTION

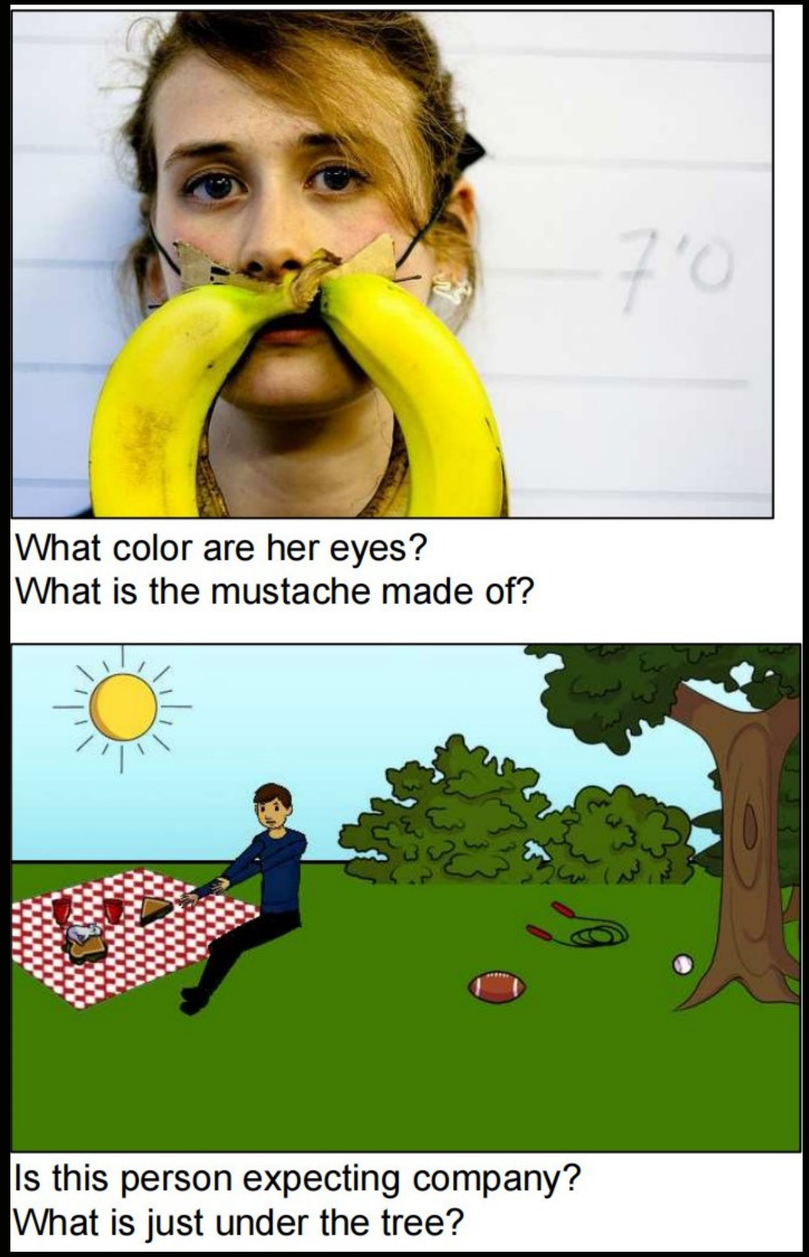
Multi-discipline Artificial Intelligence

- Computer Vision (CV)
- Natural Language Processing (NLP)
- Knowledge Representation & Reasoning (KR)

What is AI-complete?

Combination of human understanding and computer technology

multi-modal knowledge + quantitative evaluation metric



INTRODUCTION

Type of Answers

- Open-ended answering
- Multiple-choice

Evaluation?

Number of questions it answers correctly

Datasets

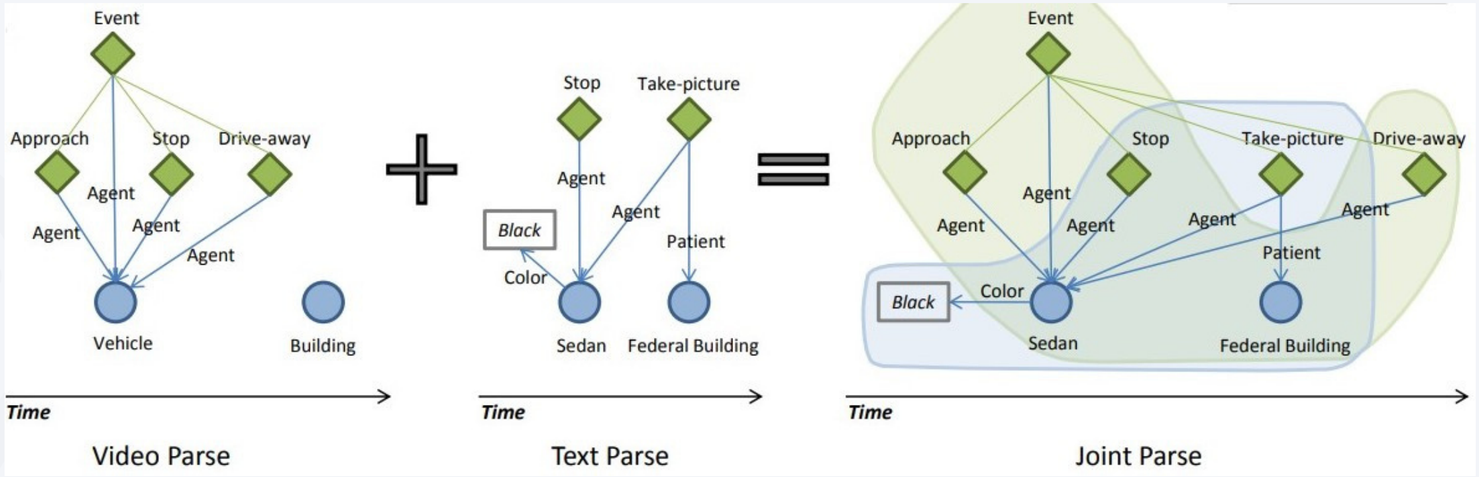
MS COCO - 204,721 images
abstract scene dataset - 50,000 scenes (3Qs)

RELATED WORK

Other VQA works

- M. Malinowski and M. Fritz - (Small data set/ Small range of Questions)
- D. Geman and the team (A Visual Turing Test for Computer Vision Systems)
- K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S. C. Zhu (Video VQA)
  By providing a text and a video answer

Not open-ended, Not free-form Qs & As



1. Q: Is there a person in the blue region? A: yes
2. Q: Is there a unique person in the blue region? A: yes
   (Label this person 1)
3. Q: Is person 1 carrying something? A: yes
4. Q: Is person 1 female? A: yes
5. Q: Is person 1 walking on a sidewalk? A: yes
6. Q: Is person 1 interacting with any other object? A: no

⋮

9. Q: Is there a unique vehicle in the yellow region? A: yes
   (Label this vehicle 1)
10. Q: Is vehicle 1 light-colored? A: yes
11. Q: Is vehicle 1 moving? A: no
12. Q: Is vehicle 1 parked and a car? A: yes

⋮

14. Q: Does vehicle 1 have exactly one visible tire? A: no
15. Q: Is vehicle 1 interacting with any other object? A: no
17. Q: Is there a unique person in the red region? A: no
18. Q: Is there a unique person that is female in the red region? A: no
19. Q: Is there a person that is standing still in the red region? A: yes
20. Q: Is there a unique person standing still in the red region? A: yes
    (Label this person 2)

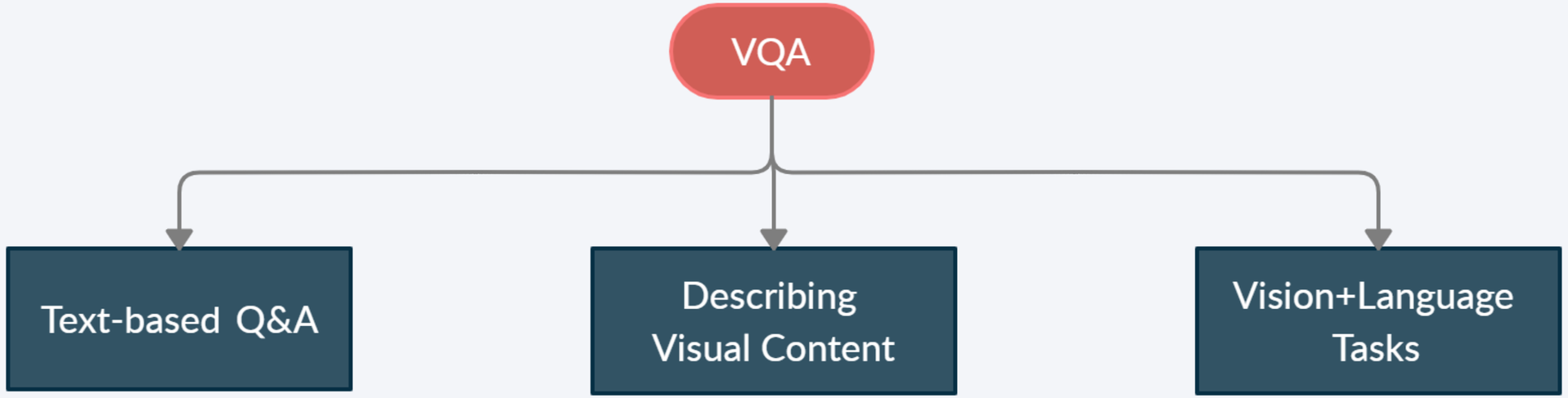
⋮

23. Q: Is person 2 interacting with any other object? A: yes
24. Q: Is person 1 taller than person 2? A: amb.
25. Q: Is person 1 closer (to the camera) than person 2? A: no
26. Q: Is there a person in the red region? A: yes
27. Q: Is there a unique person in the red region? A: yes
    (Label this person 3)

⋮

36. Q: Is there an interaction between person 2 and person 3? A: yes
37. Q: Are person 2 and person 3 talking? A: yes

RELATED WORK



sentence completion with multiple-choice answers.

grounding of questions (understand text + image)

commonsense, knowledge, and complex reasoning

Image tagging

Image captioning

Video captioning (Generate sentences to describe visual content.)

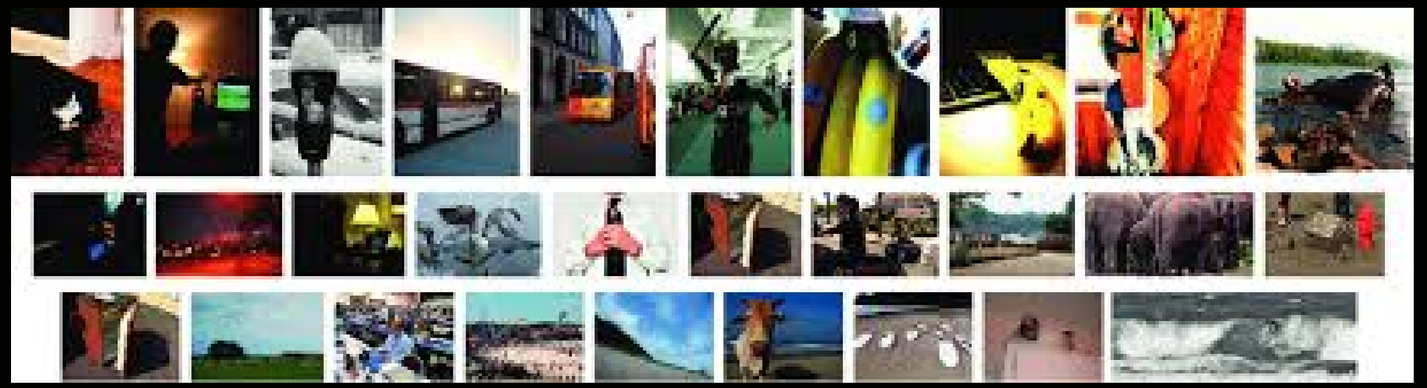
Detailed specific info

Evaluate image captioning

- coreference resolution
- generating referring expressions

DATASETS

- MS COCO - 204,721 images
- Abstract scene dataset - 50,000 scenes



VQA Dataset Collection

	Training and Validation set	Test set
Real Images (MS COCO)	123,287	81,434
Abstract Scene	30,000	20,000

- The MS COCO dataset already contains five single-sentence captions for all images.
- Abstract scene dataset
  - 20 “paperdoll” human models spanning genders, races, and ages with 8 different expressions
  - 100 objects and 31 animals in various poses

Collecting Questions

- Simple questions - require low-level computer vision knowledge.
  ex- “What color is the cat?”
- Questions that require commonsense knowledge about the scene.
  ex- “What sound does the pictured animal make?”
- Three questions for each image/scene.
- Dataset contains over ~0.76M questions.

Collecting Answers

- Open-ended questions result in a diverse set of possible answers.
- 10 answers for each question from unique workers.

Testing

Accuracy metric:
min((# humans that provided that answer)/3, 1)

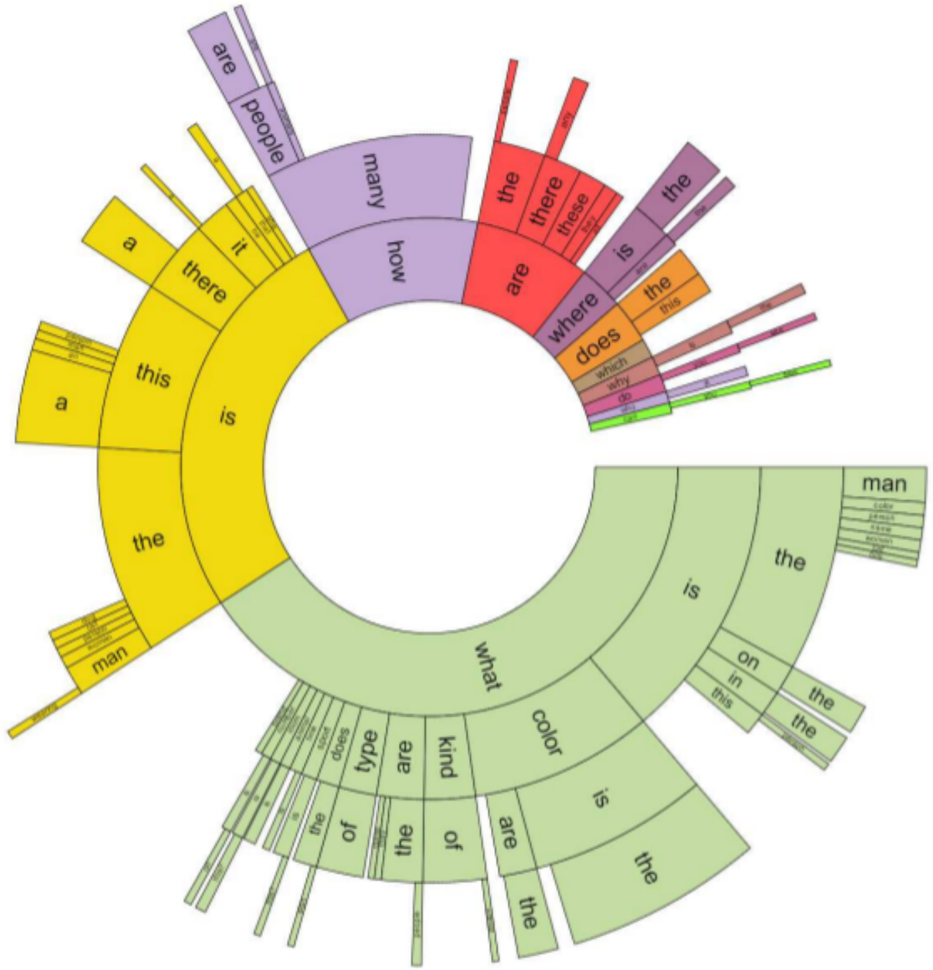
- 100% accuracy if at least 3 workers provided that exact answer.

VAQ Dataset Analysis

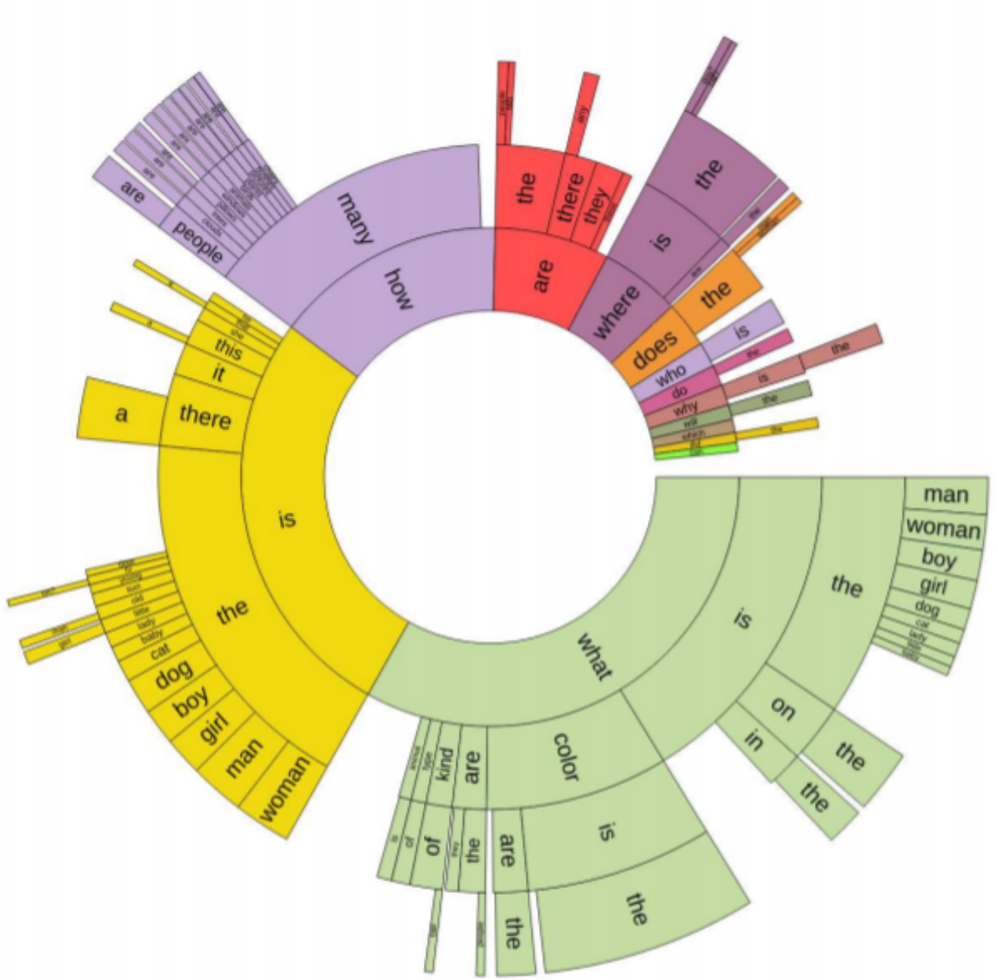
- Provide an analysis of the questions and answers in the VQA train dataset
    - To gain an understanding of the types of questions asked and answers provided following things can be done
        - Visualize the distribution of question types and answers
        - Explore how often the questions may be answered without the image using just common sense information
        - Analyze whether the information contained in an image caption is sufficient to answer the questions

Types of Questions

Real Images

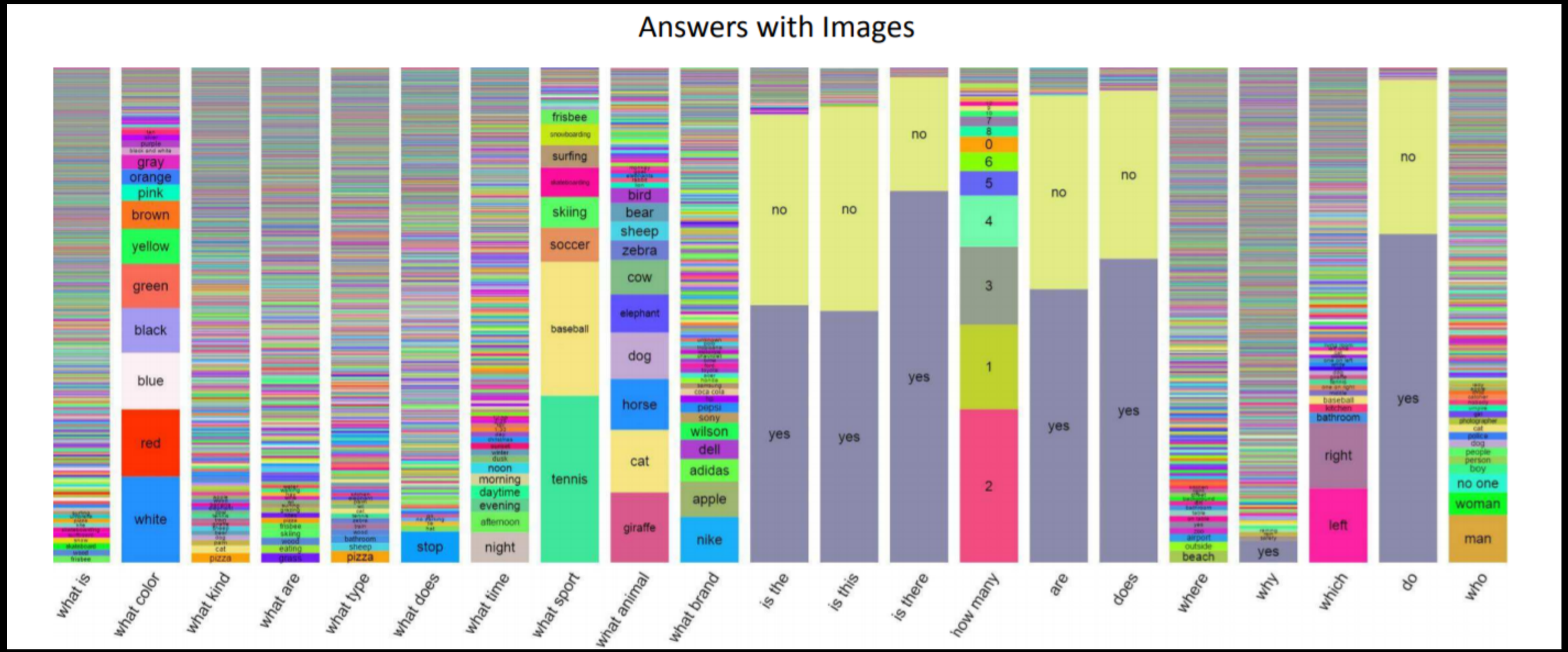


Abstract Scenes

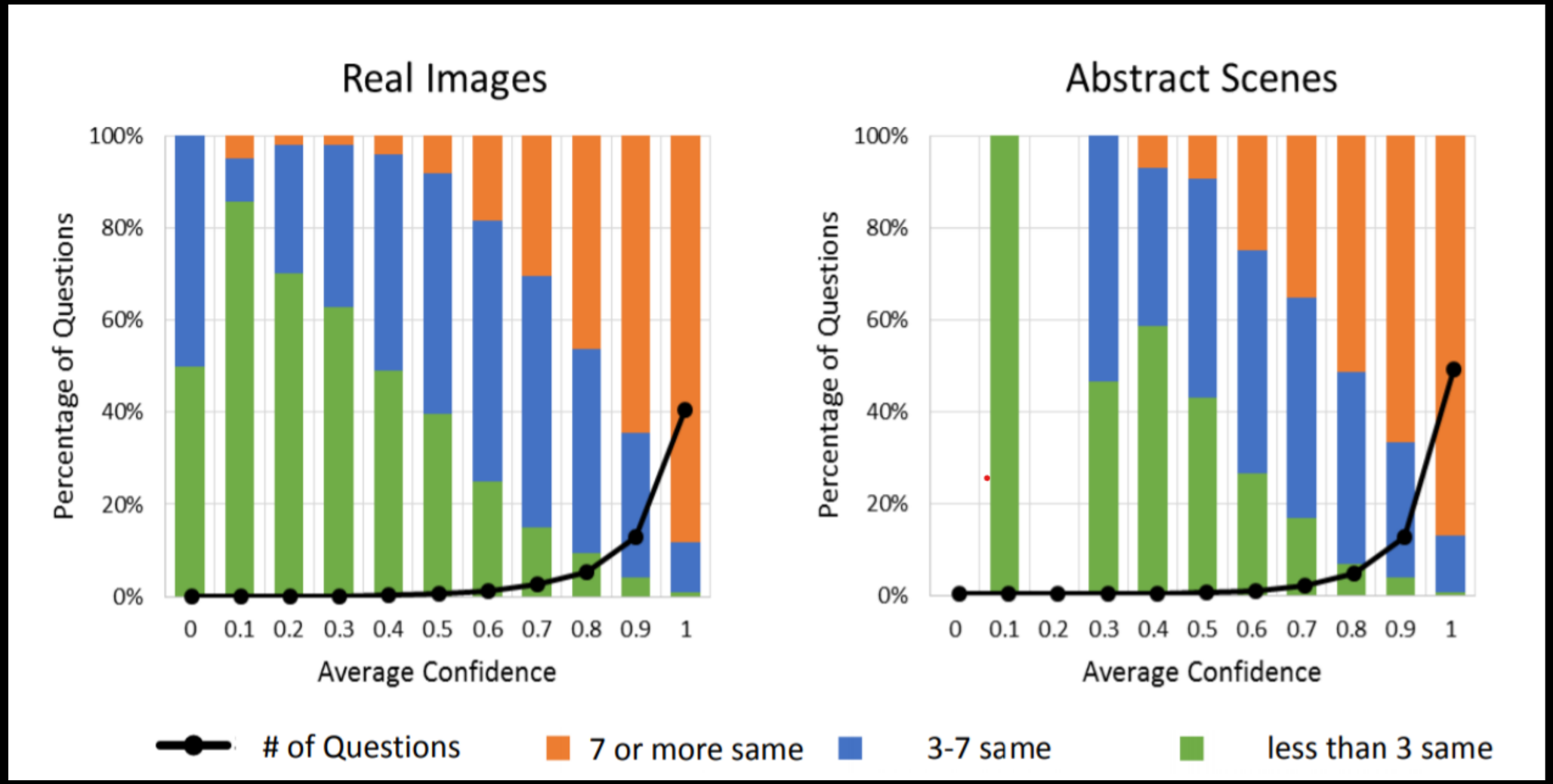


Answers

- Typical Answers
- Lengths
- 'Yes/No' and ‘Number’ Answers



- Subject Confidence
- Inter-human Agreement



- COMMON SENSE of KNOWLEDGE

- Is the Image Necessary?

e.g. - What is the colour of a fire hydrant?

- CAPTIONS vs. QUESTIONS

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

VQA BASELINES AND METHODS

Baselines

- random : randomly choose an answer from the top 1K answers of the VQA train/val dataset

- prior ("yes") : always select the most popular answer (“yes”) for both the open-ended and multiple-choice tasks.
- per Q-type prior :
  For the open-ended task :- pick the most popular answer per question type
  For the multiple-choice task:- pick the answer that is most similar to the picked answer in the open-ended task(cosine similarity in Word2Vec feature space)

- k nearest neighbor

Methods

2-channel vision (image) + language (question) model

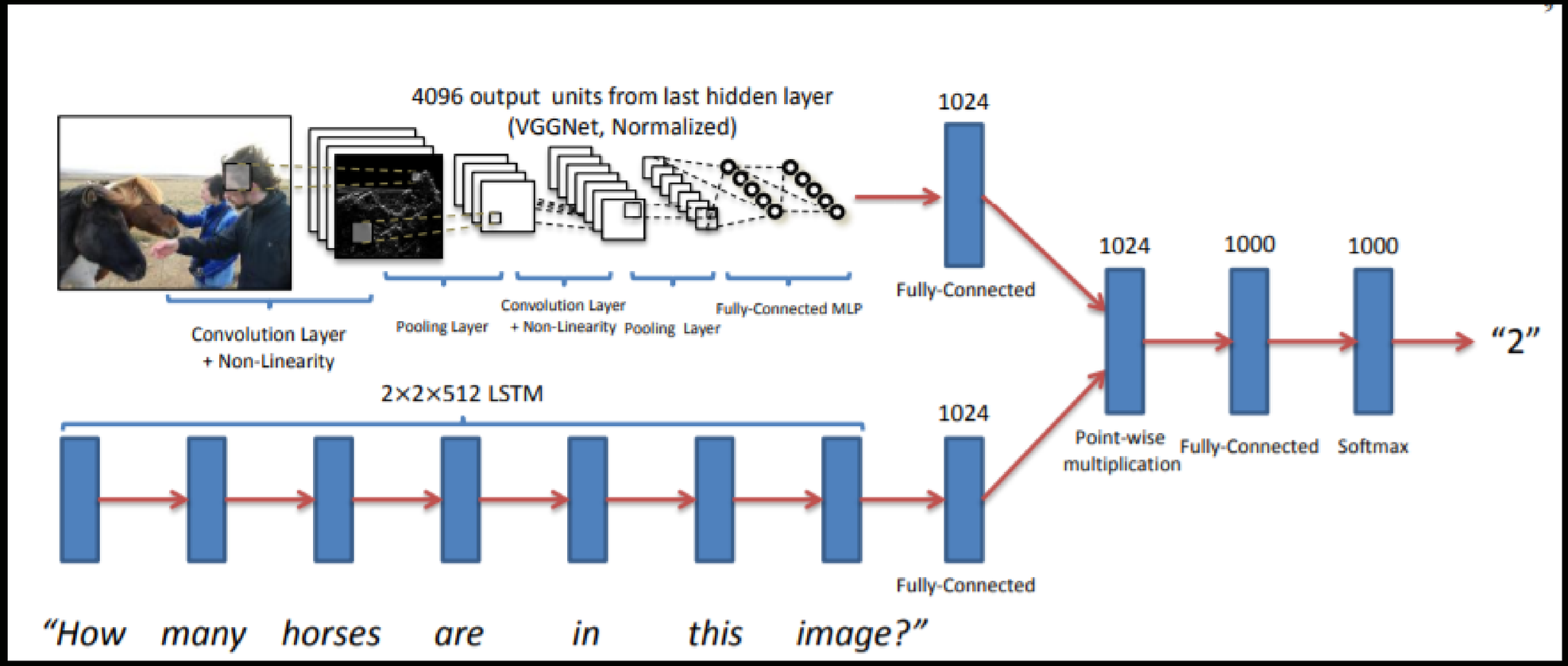


Image Channel: This channel provides an embedding for the image

1. The activations from the last hidden layer of VGGNet are used as 4096-dim image embedding.
2. norm I: These are l2 normalized activations from the last hidden layer of VGGNet.

Question Channel: This channel provides an embedding for the question.

1. Bag-of-Words Question (BoW Q)
2. LSTM Q
3. deeper LSTM Q

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

TABLE 2: Accuracy of our methods for the open-ended and multiple-choice tasks on the VQA test-dev for real images. Q = Question, I = Image, C = Caption. (Caption and BoW Q + C results are on val). See text for details.

vision-alone model that completely ignores the question performs rather poorly

best model (deeper LSTM Q + norm I)

Conclusion

- Large data set is used providing more generalization to the VQA Task
- Data obtained from real persons
- Contribution to the idea of "Ai complete"
- For some applictions Task specific question may improve performance

THANK YOU !

Q&A