COMP-SCI 5542 (SP17) - Big Data Analytics and Applications

Presentation of Paper 6

VQA: Visual Question Answering

Stanislaw Antol, *et al.*

(*Proceedings of the IEEE International Conference on Computer Vision*, 2015)

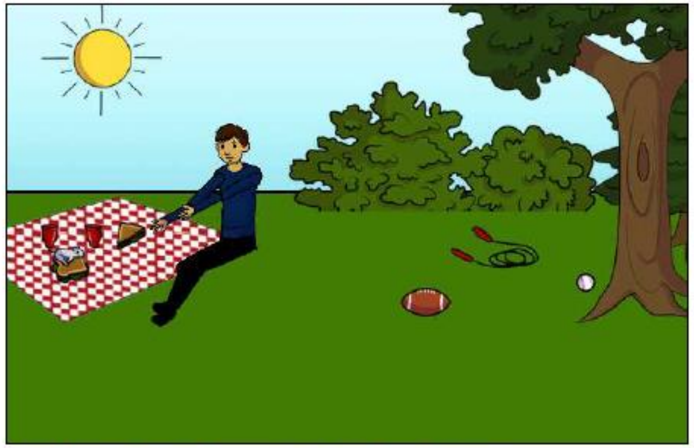
Team 9: ***Chen Wang*** **(44) - First Speaker (2-8)**
Yunlong Liu **(22) - (10-17)**
Dayu Wang **(45) - (19-23)**

Mar 21st, 2017

Introduction

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?

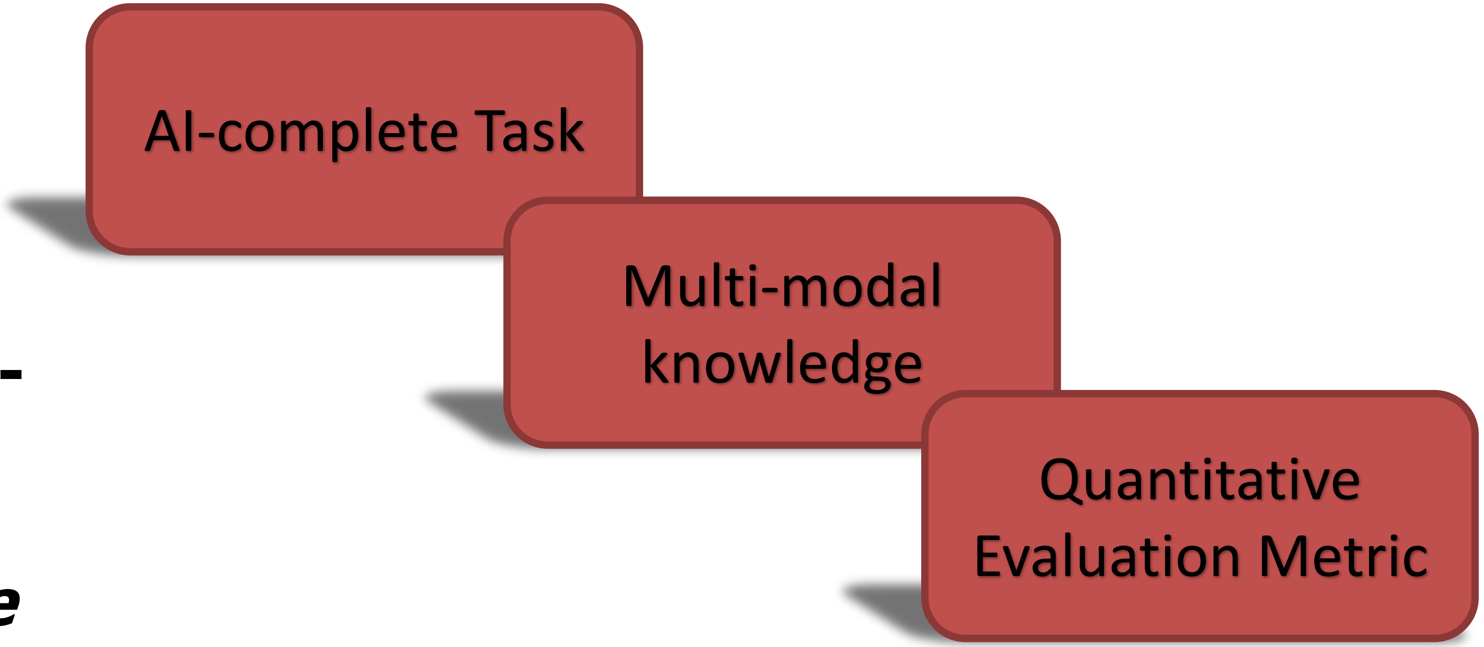
VQA:
Visual Question Answering

Ref: Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

Introduction

AI-complete Task:

Ideal task should
(i) require *multi-modal knowledge* beyond a single sub-domain
(ii) have a well-defined *quantitative evaluation metric* to track progress.



Introduction

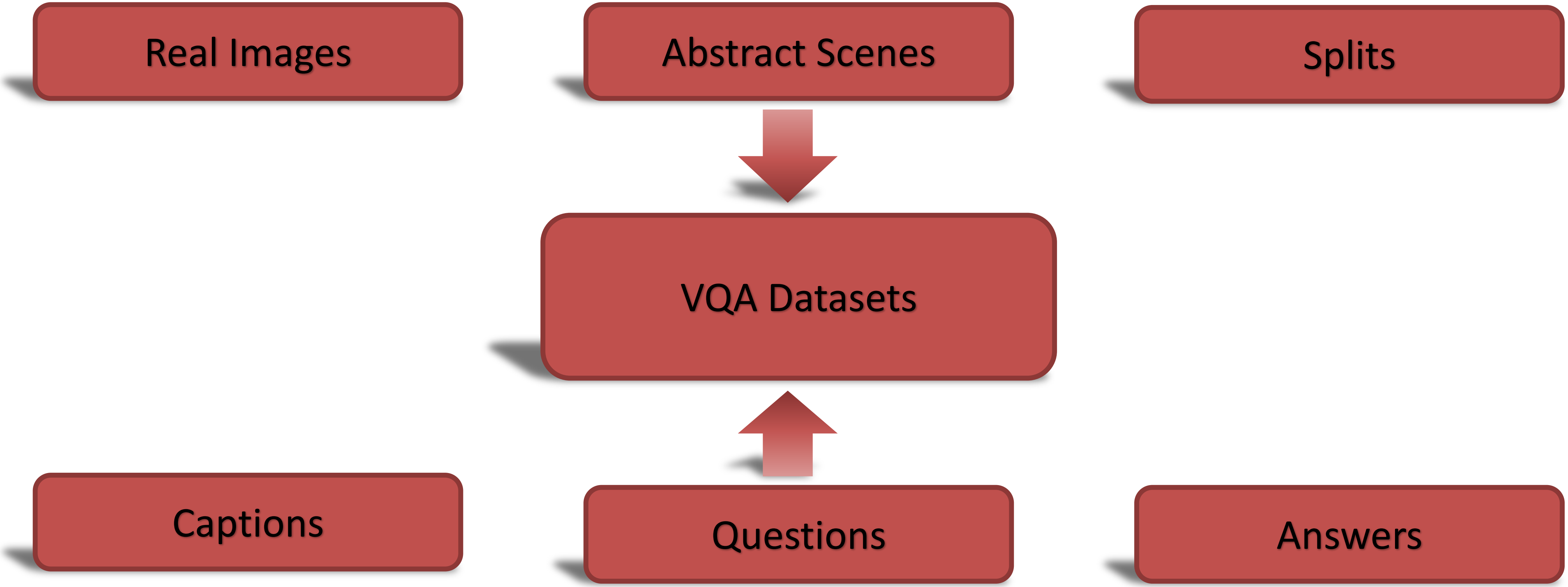
Answers of Free-form and open-ended VQA

Fine-grained recognition	**What kind of cheese is on the pizza?**
Object detection	**How many bikes are there?**
Activity recognition	**Is this man crying?**
Knowledge base reasoning	**Is this a vegetarian pizza?**
Commonsense reasoning	**Is this person expecting company?**

Related Work

Related Work	VQA Efforts	Text-based Q&A	Describing Visual Content	Other Vision+Language Tasks.
Use	Study visual question answering	Well studied problem in the NLP and text processing communities	Words or sentences are generated to describe visual content	Intersection of vision and language
Limited	fairly restricted settings with small datasets.	Text is the *grounding* of questions	Captions can often be non-specific	Limited set of visual concepts tend to be captured
Innovation	Involves *open-ended*, *free-form* questions and answers provided by humans	VQA requires the understanding of both text vision	VQA require detailed specific information about the image	Richer variety of visual concepts emerge from visual questions and their answers

VQA Dataset Collection



VQA Dataset Collection

Real Images → **Microsoft Common Objects in Context datasets which containing multiple objects and rich contextual information.**

Abstract Scenes → **A new abstract scenes dataset containing 50K scenes based on high-level tasks.**

Splits → **The same train/val/test split strategy as the MC COCO dataset.**

VQA Dataset Collection

Captions → The datasets contains five single-sentence captions for all images.

↓

Questions → Three questions from unique workers were gathered for each image/scene.

↓

Answers → Gather 10 answers for each question from unique workers, while

Questions?

The next speaker is ***Yunlong Liu*** **(22)**.

VQA Dataset Analysis

VQA Dataset Analysis

questions

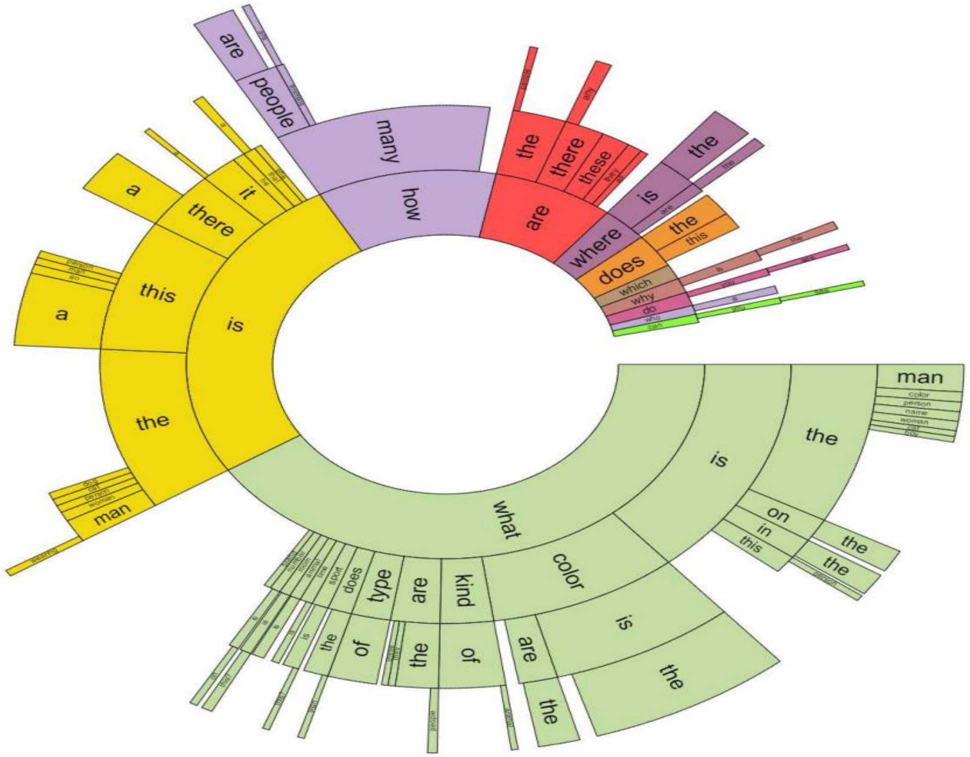
answers

commonsense knowledge

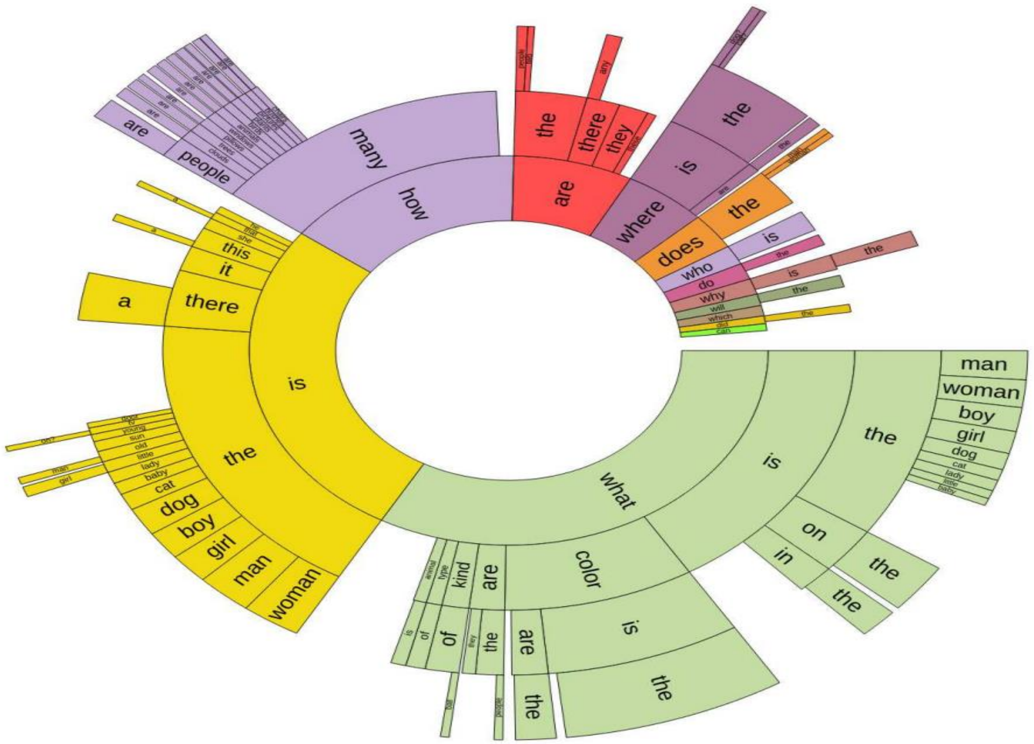
questions analysis

1. **Types of Questions**
   - **based on the words that start the question**

Real Images

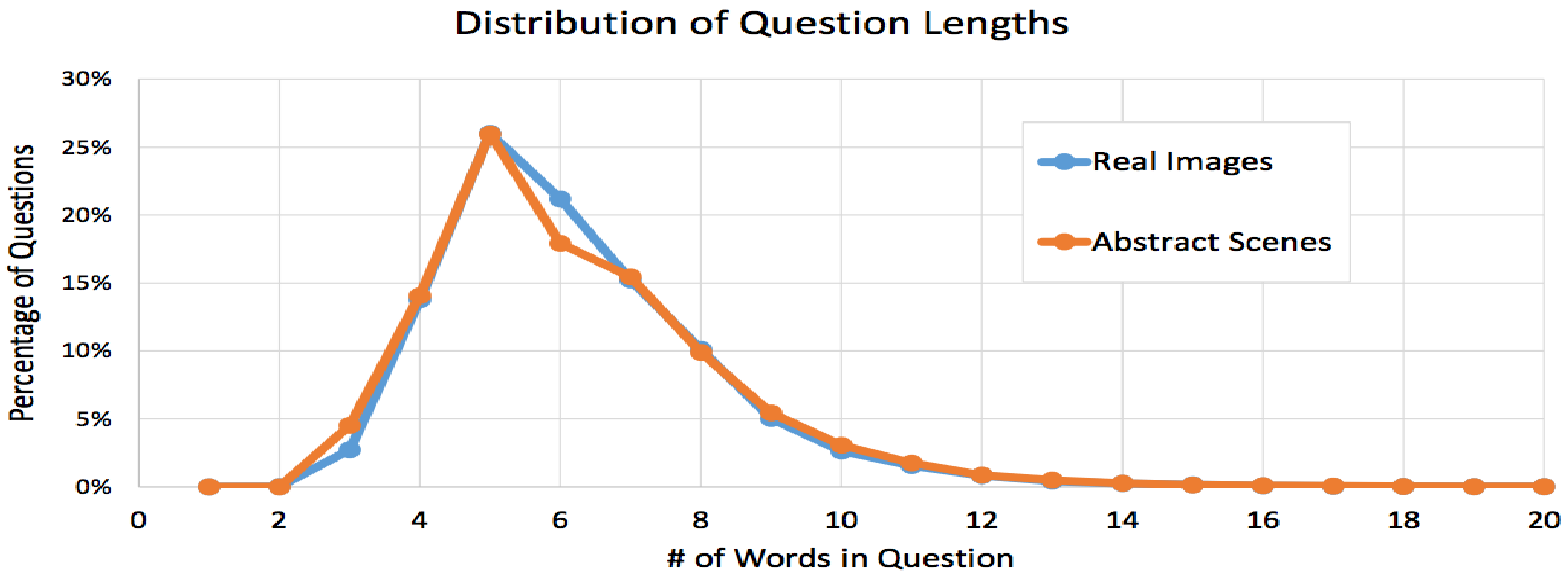


Abstract Scenes



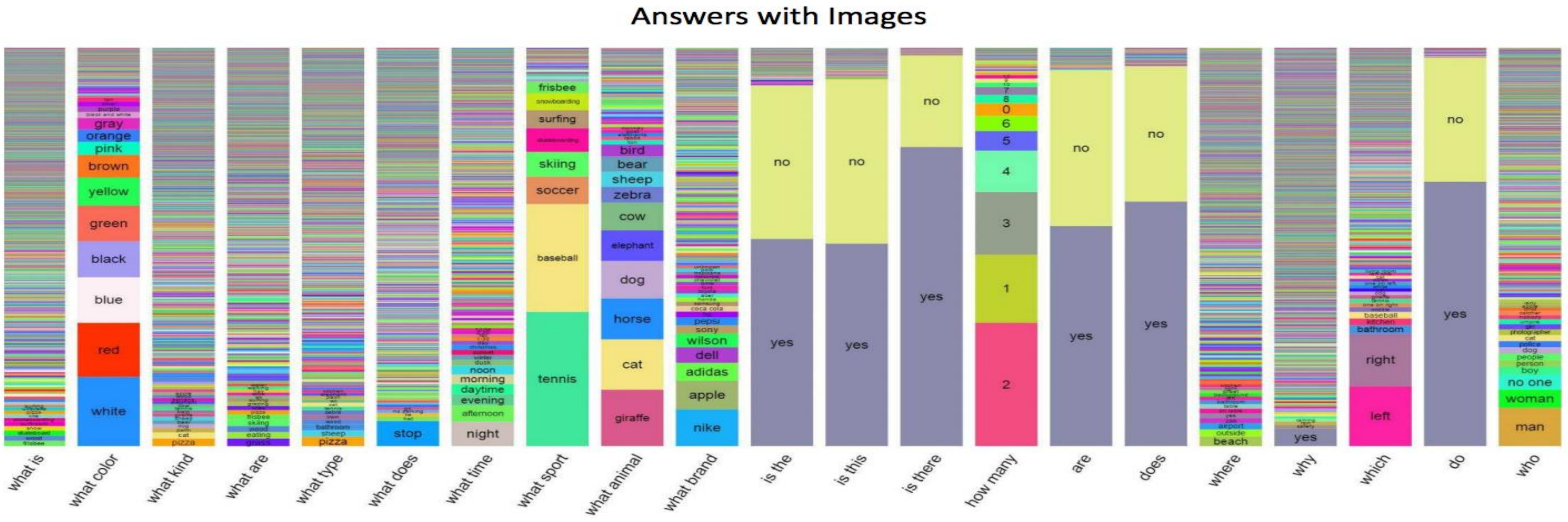
questions analysis

2. Lengths of Questions



Answers analysis

1. **typical Answers**
   - **typical answers using “yes” and “no”**
   - **rich diversity of responses**
   - **specialized responses**



Answers analysis

2. Lengths

	one word	two words	three words
real images	89.32%	6.91%	2.74%
abstract scenes	90.51%	5.89%	2.49%

percentage of the lengths

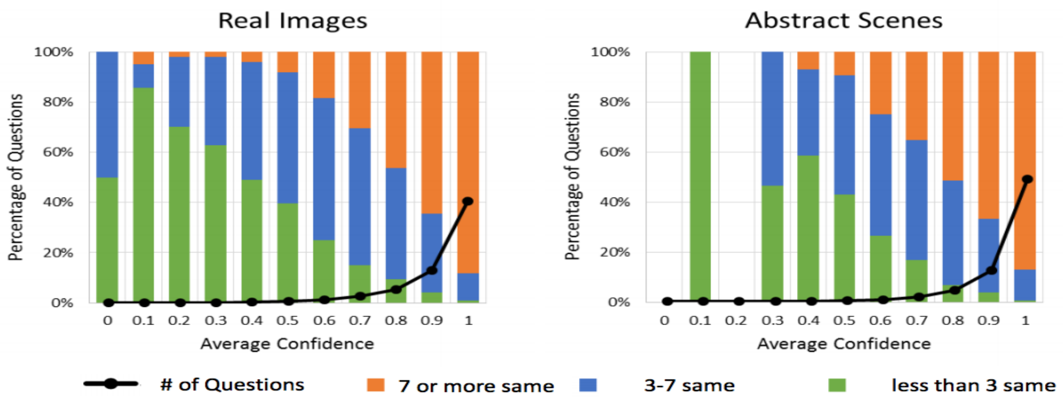
3. 'Yes/No' and 'Number' Answers

- ***Many questions are answered using either "yes" or "no" (or sometimes "maybe") – 38.37% and 40.66% of the questions on real images and abstract scenes respectively. Among these 'yes/no' questions, there is a bias towards "yes" – 58.83% and 55.86% of 'yes/no' answers are "yes" for real images and abstract scenes.***

Answers analysis

4. Subject Confidence

- ***When the subjects answered the questions, we asked "Do you think you were able to answer the question correctly?"***



Number of questions per average confidence score (0 = not confident, 1 = confident) for real images and abstract scenes (black lines). Percentage of questions where 7 or more answers are same, 3-7 are same, less than 3 are same (color bars).

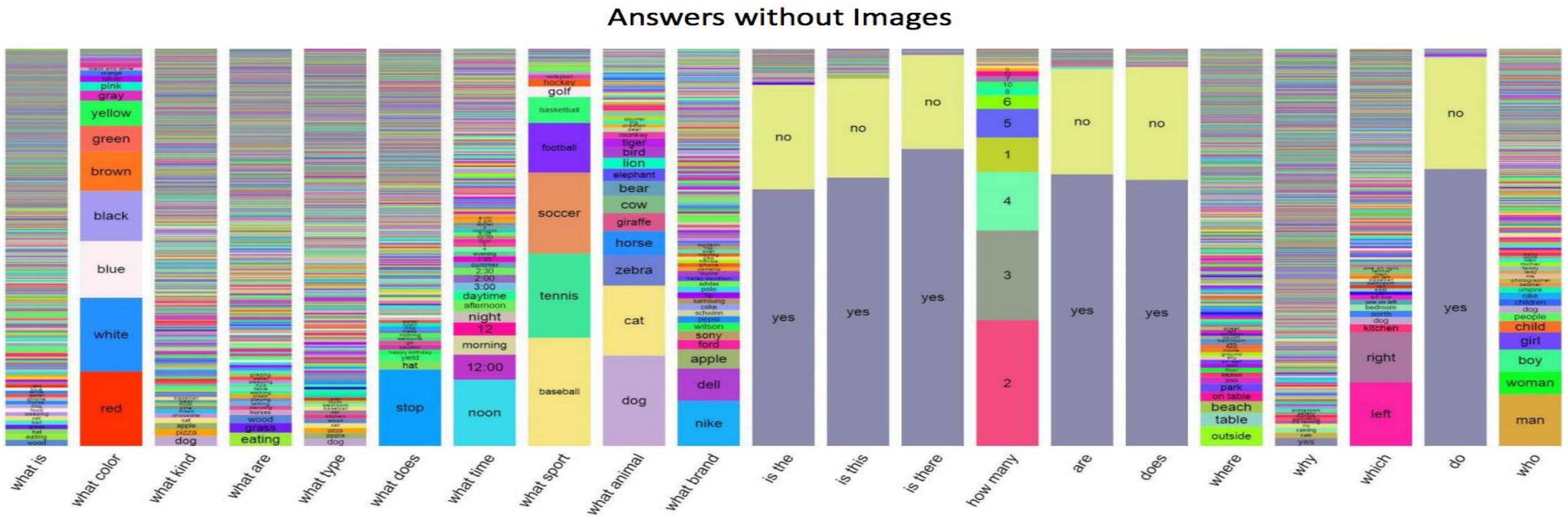
5. Inter-human Agreement

- ***the agreement between subjects increases with confidence.***

Commonsense Knowledge

1. answers without images

- ***some questions can sometimes be answered correctly using commonsense knowledge alone without the need for an image.***



Commonsense Knowledge

- ***the percentage of questions answered correctly when human subjects are given the question and a human-provided caption describing the image, but not the image.***

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

Questions?

The next speaker is ***Dayu Wang*** **(45)**.

- **VQA Baselines and Methods (Part 5)**

- **Preliminary Results - Microsoft COCO Dataset**

Method	Accuracy
Randomly Choose from Top 1K Answers	**0.12%**
Choose the Most Popular Answer	**29.72%**
Choose the Most Popular Answer per Question Type	**36.18%**
Nearest Neighbor Approach	**40.61%**

k nearest (question, image) pairs → based on the test question

Spark `Word2Vec` used to get neighbors → cosine similarity

Pick m $(1 \leq m \leq k)$ answers which have "**consensus**" answers.

Devlin, J., et al. (2015). Exploring nearest neighbor approaches for image captioning. arXiv preprint arXiv:1505.04467.
Lin, T. Y., et al. (2014). Microsoft COCO: Common Objects in Context. Euro Conf Computer Vision (740-755). Springer.
Mohapatra, A. (2015). Exploring Nearest Neighbor Approach on VQA. ECE 5554 (FS15) Class Project. Virginia Tech.

- **Training Baselines - $k = 1000$ → covers 82.67% of (train, val) answers.**

MLP Neural Network Classifier

```
hiddenLayers := 2
hiddenUnits := 1000
dropout := 0.5
(with tanh non-linearity)
```

LSTM Model + Softmax Layer → **One-Hot Encoding** → **Linear Transformation**

Question	Caption	Image
Bag-of-Words (1000)	**Human-Generated Inputs**	**Last Hidden Layer of VGGNet** (4096-Dimensional Feature)
30-Dimentional Rep. (first, second, third) × 10	**Bag-of-Words Rep.** (1000 Words in Caption)	**Q + I, Q + C, Q + I + C** (FHL Rep Trained on Features Ind)
1030-Dimentional Input (Concatenated)	**VGG = Visual Geometry Group (Univ. of Oxford)** Very Deep Convolutional Networks for Large-Scale Visual Recognition	

[Simonyan, K., et al. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv.]

- **Testing - "Open-Answer" and "Open-Choice" Tasks**

Open-Answer → Answer with **highest activation** from **all possible k answers**

Open-Choice → Answer with **highest activation** from the **potential answers**

	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

Study from the Results

Type of question matters.

LSTM is better.

Multiple-Choice is better.

All methods are **significantly worse** than human performance.

- **Further Insight**

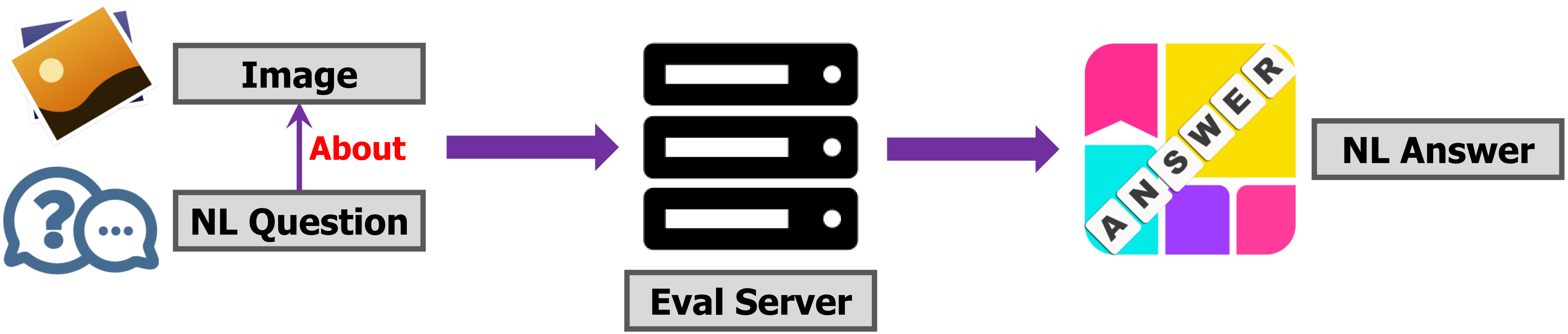
Question Type	Imp?
Requires more reasoning ("How many", "Is the")	N
Can be answered using scene-level information ("What sport")	Y
Answer contained in a generic caption ("What animal")	Y
For all question types, the results are worse than human accuracies.	

Best Model (54.06%): LSTM Q + I

Behaves like a 4.45-year-old child.

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

- **Conclusion and Discussion (Part 6)**



Dataset	AI Capabilities
250K Images	Computer Vision
760K Questions	Natural Language Processing
10M Answers	Common Sense Reasoning

- **The questions were open-ended and not task-specified.**

Domain-Specific Data Sets → Enable practical VQA Apps. → Training on task-specific datasets

- **This is the END of the presentation.**

Mar 21st, 2017

Paper 6 Presentation

"VQA: Visual Question Answering" - Stanislaw Antol, *et al*.
(*Proceedings of the IEEE International Conference on Computer Vision*, 2015)

Team 9: ***Chen Wang*** (44), ***Yunlong Liu*** (22), and ***Dayu Wang*** (45)

Questions?

Thank you!