

LECTURE 11

QUESTION ANSWERING

석사과정 장명준



KOREA
UNIVERSITY

DS·BA
Data Science &
Business Analytics

CONTENTS



1. READING COMPREHENSION

2. ANSWER SENTENCE SELECTION

3. VQA: VISUAL QUESTION ANSWERING

READING COMPREHENSION

Reading Comprehension

[News](#)
[Regions](#)
[Video](#)
[TV](#)
[Features](#)
[Opinions](#)
[More...](#)
[International Edition](#)

[World](#)
[Sport](#)
[Technology](#)
[Entertainment](#)
[Style](#)
[Travel](#)
[Money](#)

'Most Interesting Man' cutout doesn't pass in HOV lane

By [Todd Leopold, CNN](#)
 Updated 1859 GMT (0259 HKT) March 26, 2015

[Email](#) [Facebook](#) [Twitter](#) [More...](#)

Story highlights

- A driver was caught in the HOV lane with a cutout of "Most Interesting Man"
- He earned a ticket -- and a tweet admiring his tenacity

(CNN) Call it "The Most Interesting Traffic Ticket in the World."


A Washington state trooper caught a driver using a cardboard cutout of [Jonathan Goldsmith](#), the DosEquis beer pitchman known as "The Most Interesting Man in the World." The driver, who was by himself, was attempting to use the [HOV lane](#).

"The trooper immediately recognized it was a prop and not a passenger," Trooper Guy Gill told the [New York Daily News](#). "As the trooper approached, the driver was actually laughing."

Gill sent out a tweet with a photo of the cutout--who was clad in what looked like a knit shirt, a far cry from his usual attire -- and the unnamed laughing driver: "I don't always violate the HOV lane law... but when I do, I get a \$124 ticket! We'll give him an A for creativity!"

The driver was caught on Interstate 5 near Fife, Washington, just outside Tacoma.

"He could have picked a less recognizable face to put on his prop," Gill told the [Daily News](#). "We see that a lot. Usually it's a sleeping bag. This was very creative."



A driver was caught in the x with a cutout of "Most Interesting Man"

Promoted Stories



Reading Comprehension

Teaching Machines to Read and Comprehend

Karl Moritz Hermann[†] Tomáš Kočiský^{†‡} Edward Grefenstette[†]
Lasse Espeholt[†] Will Kay[†] Mustafa Suleyman[†] Phil Blunsom^{†‡}

[†]Google DeepMind [‡]University of Oxford

{kmh,tkocisky,etg,lespeholt,wkay,mustafasul,pblunsom}@google.com

Abstract

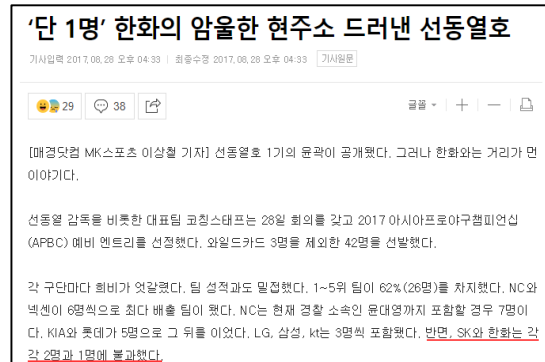
Teaching machines to read natural language documents remains an elusive challenge. Machine reading systems can be tested on their ability to answer questions posed on the contents of documents that they have seen, but until now large scale training and test datasets have been missing for this type of evaluation. In this work we define a new methodology that resolves this bottleneck and provides large scale supervised reading comprehension data. This allows us to develop a class of attention based deep neural networks that learn to read real documents and answer complex questions with minimal prior knowledge of language structure.

1 Introduction

Progress on the path from shallow bag-of-words information retrieval algorithms to machines capable of reading and understanding documents has been slow. Traditional approaches to machine reading and comprehension have been based on either hand engineered grammars [1], or information extraction methods of detecting predicate argument triples that can later be queried as a relational database [2]. Supervised machine learning approaches have largely been absent from this space due to both the lack of large scale training datasets, and the difficulty in structuring statistical models flexible enough to learn to exploit document structure.

Reading Comprehension

Entity replacement and permutation



Context

Reading
comprehension

x 는 1명만 예비 엔트리에 선정되었다

Query

Answer

한화

Hard to make

Lack of training data

Contribution of
this paper

Readily convert summary and paraphrase sentences using simple entity detection and anonymization algorithms

Reading Comprehension

Entity replacement and permutation

1) Data set

- CNN (93k articles)
- Daily Mail (220k articles)

2) Procedure

- Use **coreference** system to establish coreference in each data point
- Replace all entities with abstract entity markers according to coreference
- Randomly permute these entity markers whenever a data point is loaded

Original Version	Anonymised Version
Context The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the “Top Gear” host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon “to an unprovoked physical and verbal attack.” ...	the <i>ent381</i> producer allegedly struck by <i>ent212</i> will not press charges against the “ <i>ent153</i> ” host , his lawyer said friday . <i>ent212</i> , who hosted one of the most - watched television shows in the world , was dropped by the <i>ent381</i> wednesday after an internal investigation by the <i>ent180</i> broadcaster found he had subjected producer <i>ent193</i> “ to an unprovoked physical and verbal attack . ” ...
Query Producer X will not press charges against Jeremy Clarkson, his lawyer says.	producer X will not press charges against <i>ent212</i> , his lawyer says .
Answer Oisin Tymon	<i>ent193</i>

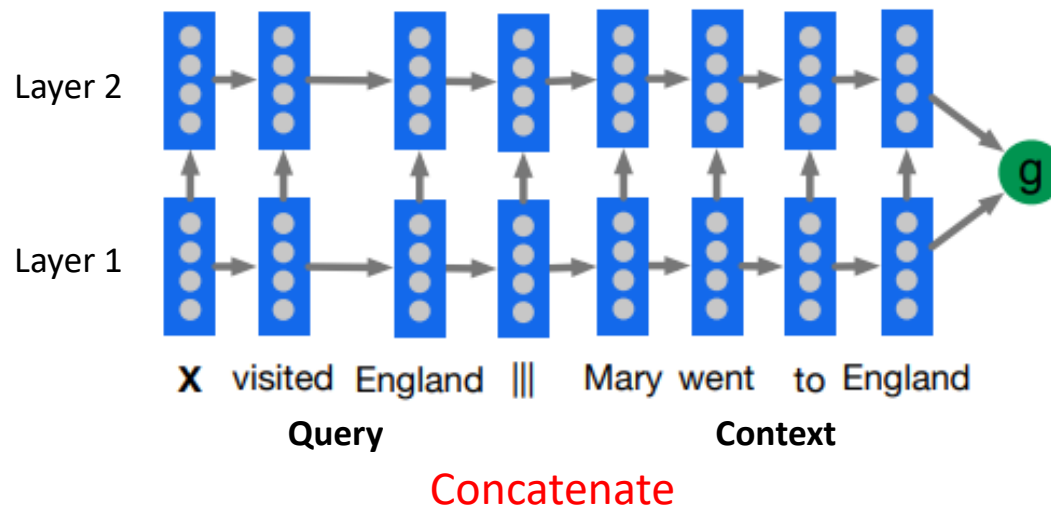
Table 3: Original and anonymised version of a data point from the Daily Mail validation set. The anonymised entity markers are constantly permuted during training and testing.

The object is to provide a corpus for evaluating a model’s ability to read and comprehend document, not world knowledge

Reading Comprehension

Models

1) The Deep LSTM Reader



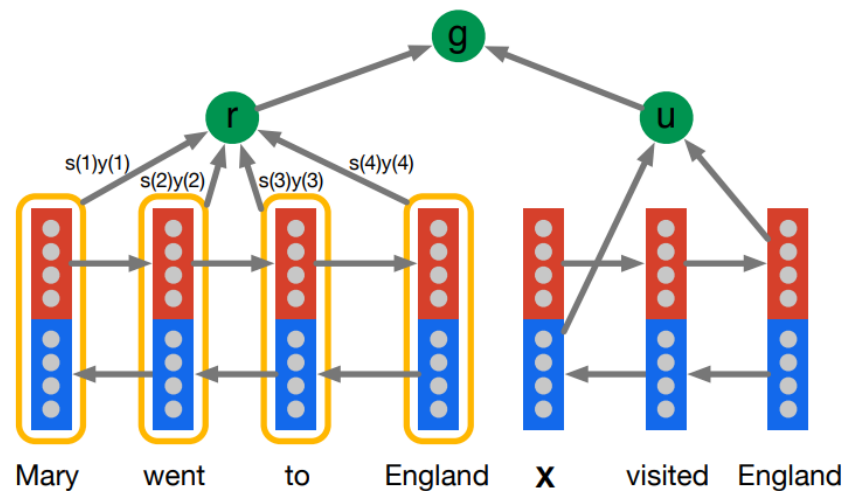
Problem : Context is long

➡ Use Attention

Reading Comprehension

Models

2) Attentive Reader



Query

$$u = \vec{y}_q(|q|) || \tilde{y}_q(1)$$

Context

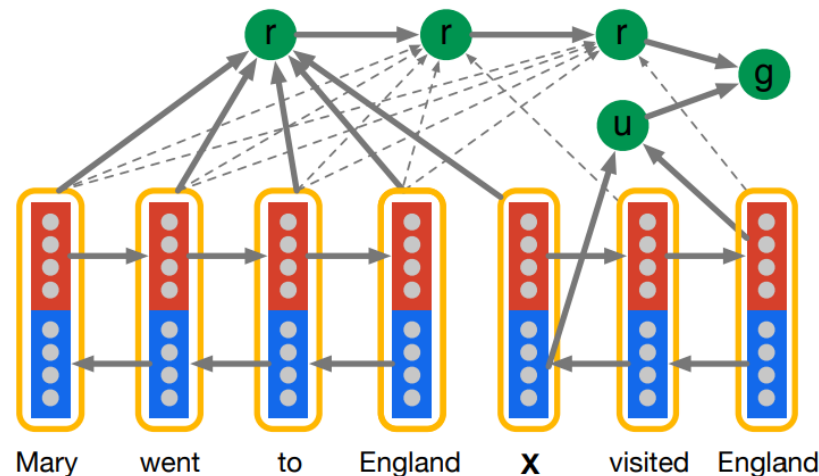
$$\begin{aligned} r &= y_d s \\ s(t) &\propto \exp(w_{ms}^T m(t)) \\ m(t) &= \tanh(W_{ym} y_d(t) + W_{um} u) \end{aligned}$$

Reading Comprehension

Models

3) Impatient Reader

- Reread from the document as each query token is read



Query

$$y_q(i) = \vec{y}_q(i) || \tilde{y}_q(i)$$

Context

$$r(0) = 0, r(i) = y_d^T s(i) + \tanh(W_{rr} r(i-1))$$

$$s(i, t) \propto \exp(w_{ms}^T m(i, t))$$

$$m(i, t) = \tanh(W_{dm} y_d(t) + W_{rm} r(i-1) + W_{qm} y_q(i))$$

Reading Comprehension

Result

	CNN		Daily Mail	
	valid	test	valid	test
Maximum frequency	30.5	33.2	25.6	25.5
Exclusive frequency	36.6	39.3	32.7	32.8
Frame-semantic model	36.3	40.2	35.5	35.5
Word distance model	50.5	50.9	56.4	55.5
Deep LSTM Reader	55.0	57.0	63.3	62.2
Uniform Reader	39.0	39.4	34.6	34.4
Attentive Reader	61.6	63.0	70.5	69.0
Impatient Reader	61.8	63.8	69.0	68.0

Table 5: Accuracy of all the models and benchmarks on the CNN and Daily Mail datasets. The Uniform Reader baseline sets all of the $m(t)$ parameters to be equal.

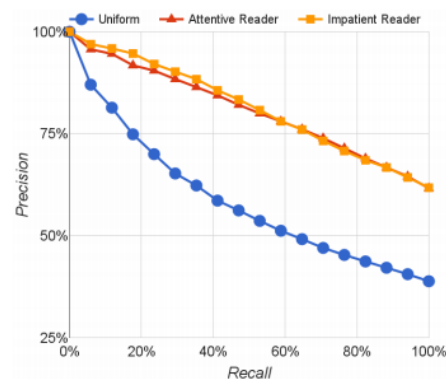


Figure 2: Precision@Recall for the attention models on the CNN validation data.

by ent423 ,ent261 correspondent updated 9:49 pm et ,thu march 19 ,2015 (ent261) a ent114 was killed in a parachute accident in ent45 ,ent85 ,near ent312 , a ent119 official told ent261 on wednesday . he was identified thursday as special warfare operator 3rd class ent23 ,29 , of ent187 , ent265 . `` ent23 distinguished himself consistently throughout his career . he was the epitome of the quiet professional in all facets of his life , and he leaves an inspiring legacy of natural tenacity and focused

...

ent119 identifies deceased sailor as X , who leaves behind a wife

by ent270 ,ent223 updated 9:35 am et ,mon march 2 ,2015 (ent223) ent63 went familial for fall at its fashion show in ent231 on sunday , dedicating its collection to `` mamma '' with nary a pair of `` mom jeans '' in sight . ent164 and ent21 , who are behind the ent196 brand , sent models down the runway in decidedly feminine dresses and skirts adorned with roses , lace and even embroidered doodles by the designers ' own nieces and nephews . many of the looks featured saccharine needlework phrases like `` i love you ,

...

X dedicated their fall fashion show to moms

ANSWER SENTENCE SELECTION

Answer Sentence Selection

Deep Learning for Answer Sentence Selection

Lei Yu¹ Karl Moritz Hermann² Phil Blunsom^{1,2} Stephen Pulman¹

¹Department of Computer Science, University of Oxford

²Google DeepMind

{lei.yu, phil.blunsom, stephen.pulman}@cs.ox.ac.uk
kmh@google.com

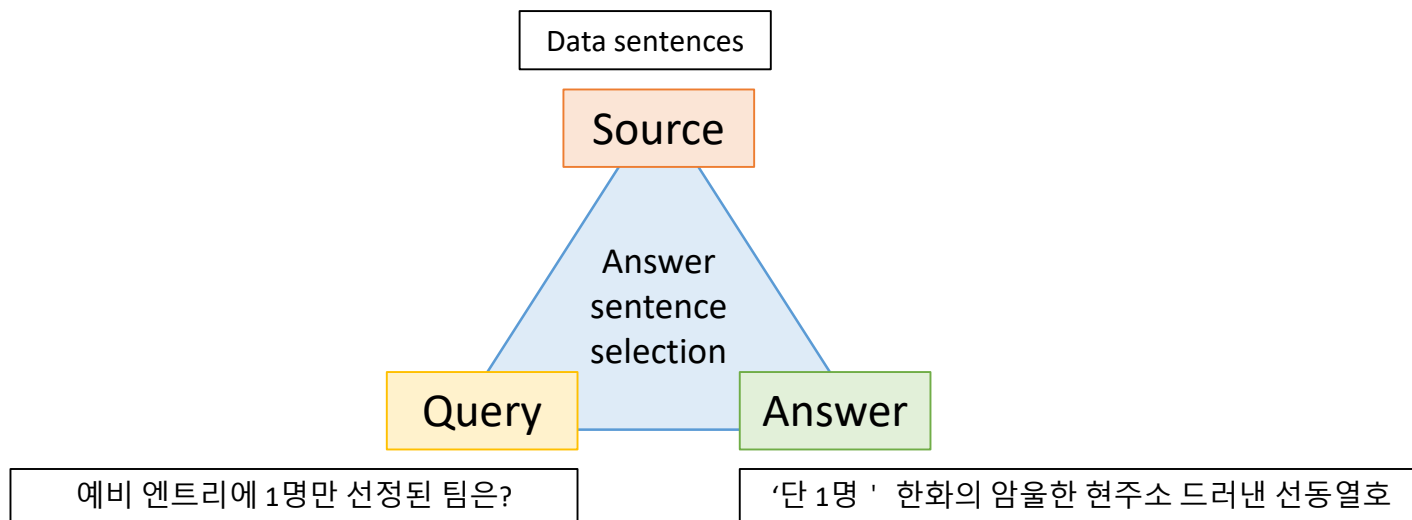
Abstract

Answer sentence selection is the task of identifying sentences that contain the answer to a given question. This is an important problem in its own right as well as in the larger context of open domain question answering. We propose a novel approach to solving this task via means of distributed representations, and learn to match questions with answers by considering their semantic encoding. This contrasts prior work on this task, which typically relies on classifiers with large numbers of hand-crafted syntactic and semantic features and various external resources. Our approach does not require any feature engineering nor does it involve specialist linguistic data, making this model easily applicable to a wide range of domains and languages. Experimental results on a standard benchmark dataset from TREC demonstrate that—despite its simplicity—our model matches state of the art performance on the answer sentence selection task.

Answer Sentence Selection

- What is Answer Sentence Selection ??

Answer sentence selection. Answer sentence selection denotes the task of selecting a sentence that contains the information required to answer a given question from a set of candidates obtained via some information extraction system.



- Difference with Reading comprehension

The answer is guaranteed to be extracted, while in reading comprehension it could be either generated or extracted

Answer Sentence Selection

Models

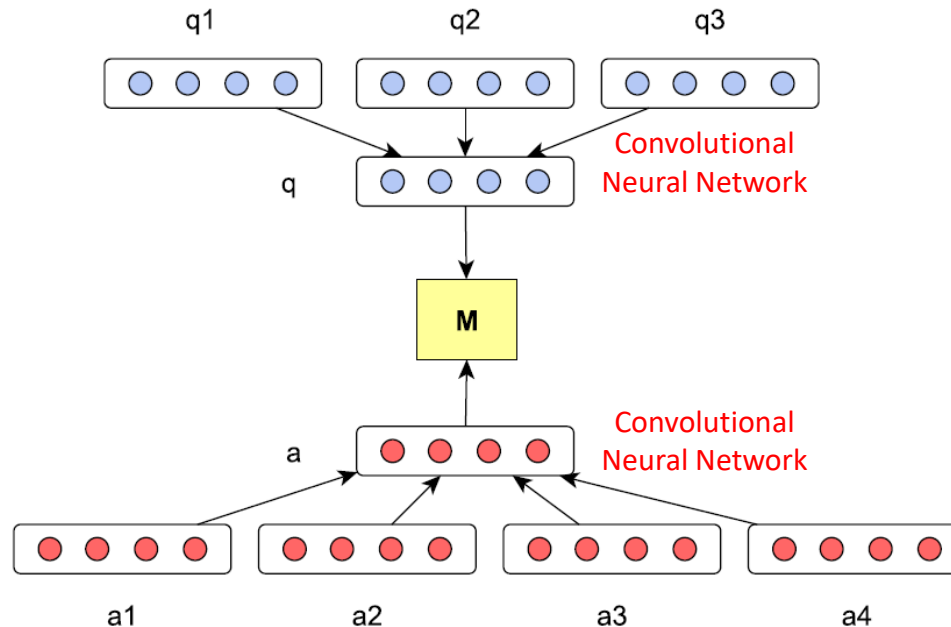
- Assume a set of questions Q , where each question $q_i \in Q$
- q_i is associated with a list of answer sentences $\{a_{i1}, a_{i2}, \dots, a_{im}\}$ and with their judgements $\{y_{i1}, y_{i2}, \dots, y_{im}\}$
- $y_{i1} = 0$ if the answer is correct
- By treating each data point as a triple (q_i, a_{ij}, y_{ij}) , they changed multi-labelling task to binary task

Data	# Questions	# QA Pairs	% Correct	Judgement
TRAIN-ALL	1,229	53,417	12.0	automatic
TRAIN	94	4,718	7.4	manual
DEV	82	1,148	19.3	manual
TEST	100	1,517	18.7	manual

Answer Sentence Selection

Models

Key idea : correct answers have high semantic similarity to questions



$$p(y = 1|q, a) = \sigma(q^T M a + b)$$

M : transformation matrix, $M \in R^d$

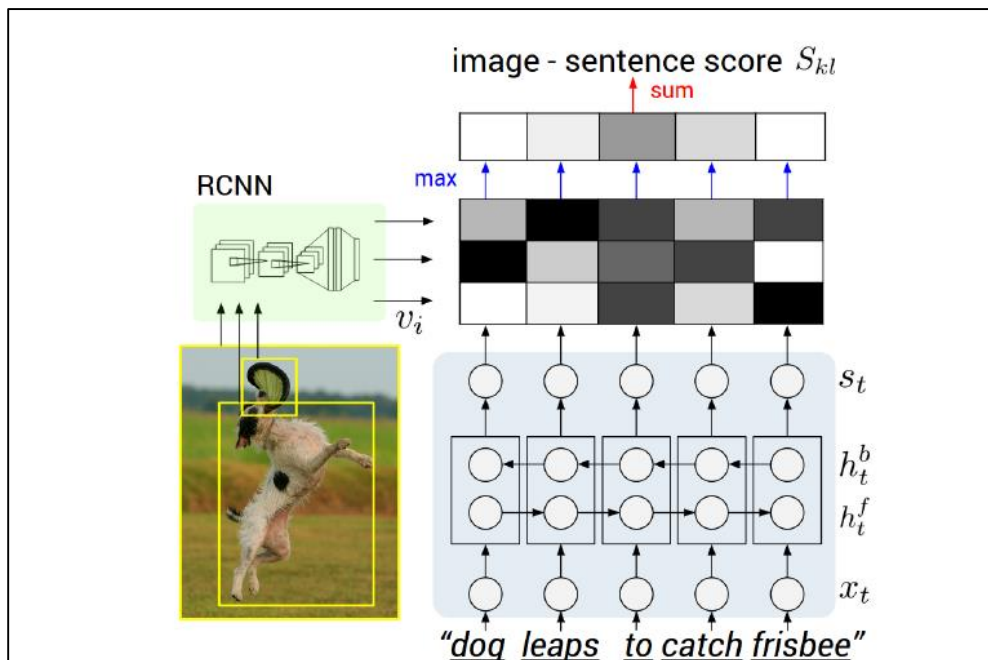
$$L = -\log \prod_n p(y_n|q_n, a_n) + \frac{\lambda}{2} \|\theta\|_F^2$$

- Sensitive to word ordering
- Able to capture features of n-grams
- Removing reliance on external resources such as parse trees
- Convolution and pooling layer helps to capture long-range dependencies

Answer Sentence Selection

Models

Key idea : correct answers have high semantic similarity to questions



$$S_{kl} = \sum_{t \in g_l} \max_{i \in g_k} v_i^T s_t.$$

$$p(y = 1|q, a) = \sigma(q^T M a + b)$$

M : transformation matrix, $M \in R^d$

$$L = -\log \prod_n p(y_n|q_n, a_n) + \frac{\lambda}{2} \|\theta\|_F^2$$

Answer Sentence Selection

Result

Model	MAP	MRR
TRAIN		
unigram	0.5387	0.6284
bigram	0.5476	0.6437
unigram + count	0.6889	0.7727
bigram + count	0.7058	0.7800
TRAIN-ALL		
unigram	0.5470	0.6329
bigram	0.5693	0.6613
unigram + count	0.6934	0.7677
bigram + count	0.7113	0.7846

System	MAP	MRR
Baselines		
Random	0.3965	0.4929
Word Count	0.5707	0.6266
Wgt Word Count	0.5961	0.6515
Published Models		
Wang et al. (2007)	0.6029	0.6852
Heilman and Smith (2010)	0.6091	0.6917
Wang and Manning (2010)	0.5951	0.6951
Yao et al. (2013)	0.6307	0.7477
Severyn and Moschitti (2013)	0.6781	0.7358
Yih et al. (2013) – LR	0.6818	0.7616
Yih et al. (2013) – BDT	0.6940	0.7894
Yih et al. (2013) – LCLR	0.7092	0.7700
Our Models		
TRAIN bigram + count	0.7058	0.7800
TRAIN-ALL bigram + count	0.7113	0.7846

VQA: VISUAL QUESTION ANSWERING

Visual Question Answering

VQA: Visual Question Answering

www.visualqa.org

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

Abstract—We propose the task of *free-form* and *open-ended* Visual Question Answering (VQA). Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. Moreover, VQA is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format. We provide a dataset containing $\sim 0.25\text{M}$ images, $\sim 0.76\text{M}$ questions, and $\sim 10\text{M}$ answers (www.visualqa.org), and discuss the information it provides. Numerous baselines and methods for VQA are provided and compared with human performance. Our VQA demo is available on CloudCV (<http://cloudcv.org/vqa>).

1 INTRODUCTION

We are witnessing a renewed excitement in multi-discipline Artificial Intelligence (AI) research problems. In particular, research in image and video captioning that combines Computer Vision (CV), Natural Language Processing (NLP), and Knowledge Representation & Reasoning (KR) has dramatically increased in the past year [16], [9], [12], [38], [26], [24], [53]. Part of this excitement stems from a belief that multi-discipline tasks like image captioning are a step towards solving AI. However, the current state of the art demonstrates that a coarse scene-level understanding of an image paired with word n -gram statistics suffices to generate reasonable image captions, which suggests image captioning may not be as “AI-complete” as desired.

What makes for a compelling “AI-complete” task? We believe that in order to spawn the next generation of AI algorithms, an ideal task should (i) require *multi-modal knowledge* beyond a single sub-domain (such as CV) and (ii) have a well-defined *quantitative evaluation metric* to track progress. For some tasks, such as image captioning, automatic evaluation is still a difficult and open research problem [51], [13], [22].



Fig. 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.

Visual Question Answering

VQA data collection

1) Images

- Real Images : MS COCO data set



Does this man have children?
yes yes yes
yes yes yes
Is this man crying?
no no yes
no no yes



Has the pizza been baked?
yes yes yes
yes yes yes
What kind of cheese is topped on this pizza?
feta feta mozzarella
ricotta ricotta mozzarella



How many pickles are on the plate?
1 1 1
1 1 1
What is the shape of the plate?
circle circle circle
round round round



What does the sign say?
stop stop stop
stop stop stop
What shape is this sign?
octagon octagon octagon
octagon octagon octagon

- Abstract Scenes : real images requires the use of complex and often noisy visual recognizers



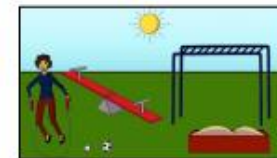
How many glasses are on the table?
3 3 2
3 3 2
What is the woman reaching for?
door handle glass wine
fruit glass remote



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



Are the kids in the room the grandchildren of the adults?
probably yes yes
yes yes yes
What is on the bookshelf?
nothing nothing nothing
books books books



How many balls are there?
2 2 1
2 2 3
What side of the teeter totter is on the ground?
right right left
right right right side

2) Captions

- Real Images : contained in MS COCO data set (5 sentences)
- Abstract Scenes : collected (5 sentences)

Visual Question Answering

VQA data collection

3) Questions

- Collecting interesting, diverse, and well posed questions is a significant challenge
- Questions should require the image to correctly answer
- Used “smart robot” interface

4) Answers

- ‘yes’ or ‘no’ questions are not matters
- Open-ended questions and multiple choice questions have a diverse set of possible answers



Answers:

‘dog’, ‘poodle’, ‘brown poodle’, ‘Andy’

- Open-ended : gathered 10 answers for each question from unique workers
- Multiple choices : gathered 18 answers for each question from unique workers
- Answers are evaluated using the following accuracy metric

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

Visual Question Answering

VQA data analysis

1) Questions

- Types of questions

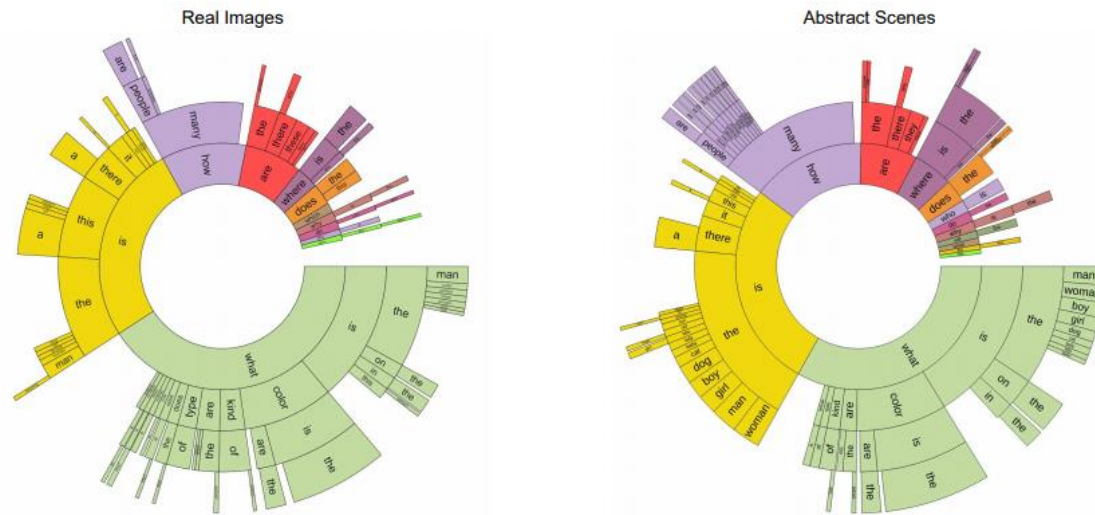
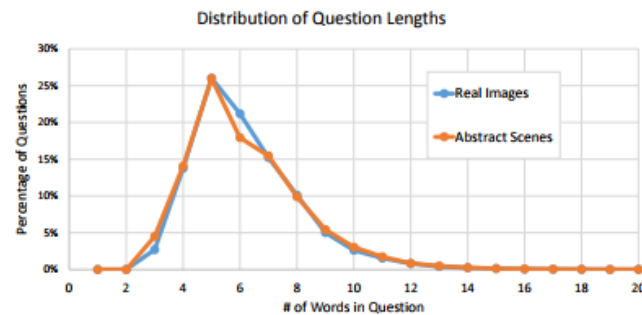


Fig. 3: Distribution of questions by their first four words for a random sample of 60K questions for real images (left) and all questions for abstract scenes (right). The ordering of the words starts towards the center and radiates outwards. The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show.

- Lengths

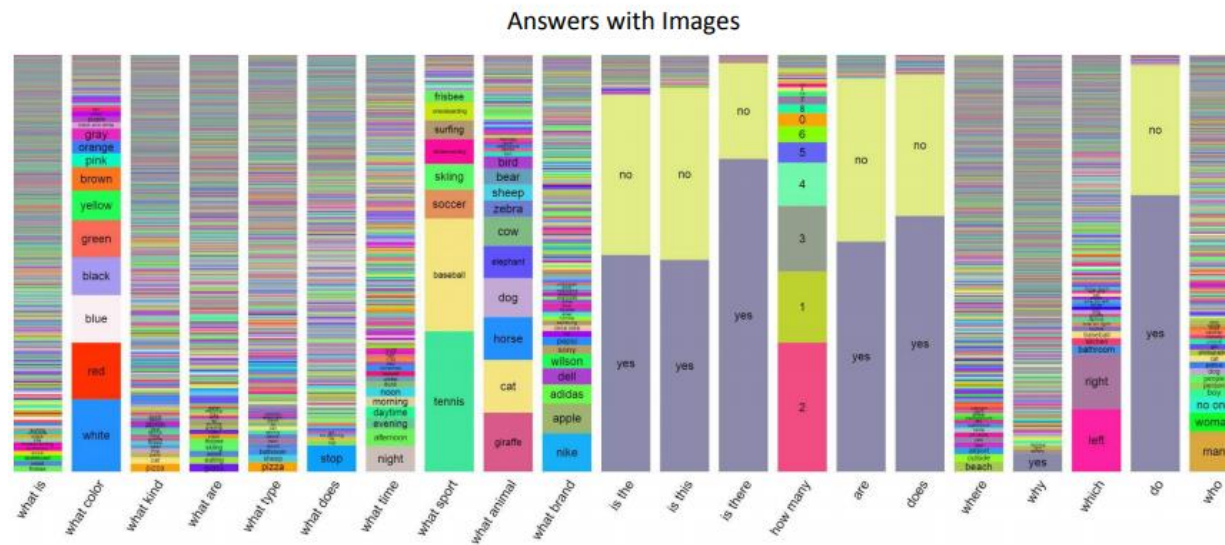


Visual Question Answering

VQA data analysis

2) Answers

- Typical Answers



- Lengths

Real Images

89.32%

One word

6.91%

Two words

2.74%

Three words

Abstract images

90.51%

One word

5.89%

Two words

2.49%

Three words

Visual Question Answering

VQA data analysis

3) Commonsense knowledge

Is the Image Necessary?

Q: 상대방 앞마당에 셔틀이 드랍한 유닛은?



Visual Question Answering

VQA data analysis

3) Commonsense knowledge

Which questions require common sense?

3-4 (15.3%)	5-8 (39.7%)	9-12 (28.4%)	13-17 (11.2%)	18+ (5.5%)
Is that a bird in the sky?	How many pizzas are shown?	Where was this picture taken?	Is he likely to get mugged if he walked down a dark alleyway like this?	What type of architecture is this?
What color is the shoe?	What are the sheep eating?	What ceremony does the cake commemorate?	Is this a vegetarian meal?	Is this a Flemish bricklaying pattern?
How many zebras are there?	What color is his hair?	Are these boats too tall to fit under the bridge?	What type of beverage is in the glass?	How many calories are in this pizza?
Is there food on the table?	What sport is being played?	What is the name of the white shape under the batter?	Can you name the performer in the purple costume?	What government document is needed to partake in this activity?
Is this man wearing shoes?	Name one ingredient in the skillet.	Is this at the stadium?	Besides these humans, what other animals eat here?	What is the make and model of this vehicle?

Fig. 7: Example questions judged by Mturk workers to be answerable by different age groups. The percentage of questions falling into each age group is shown in parentheses.

4) Caption vs. Questions

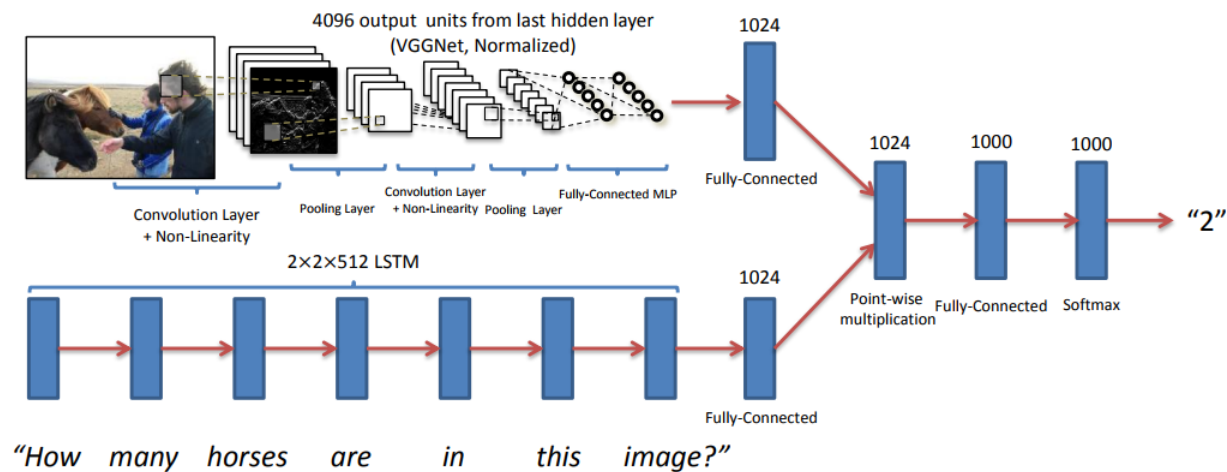
Do image captions provide enough information to answer the questions?

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

Visual Question Answering

Models

- 1) Question is language : RNN Encoder
- 2) Context is single picture : CNN
- 3) Answer is a (single) word : classification



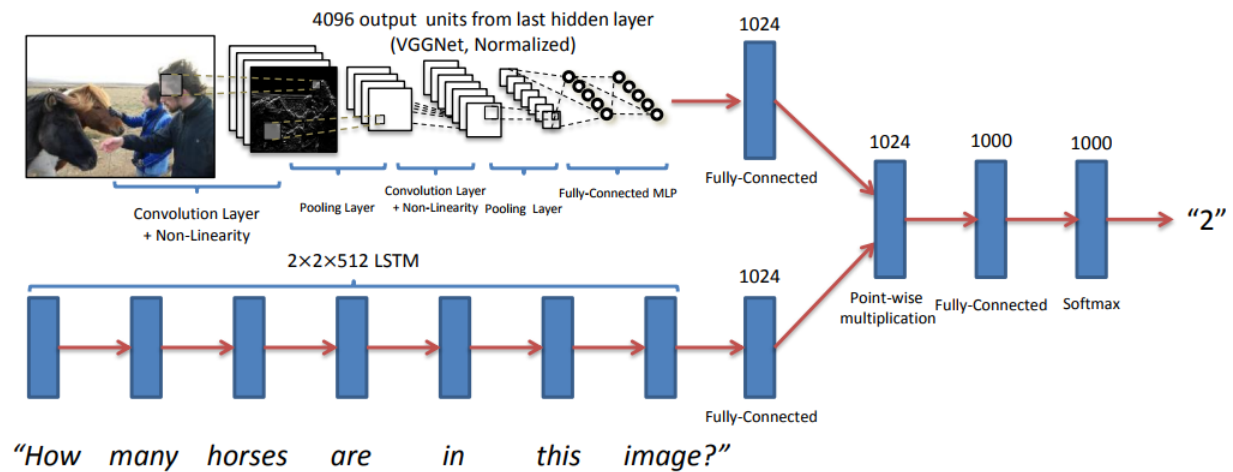
Visual Question Answering

Models

Variations

1) Image Channel

- I : use output of VGGNet(4096d) vectors as embedding
- Norm I: l_2 normalized output



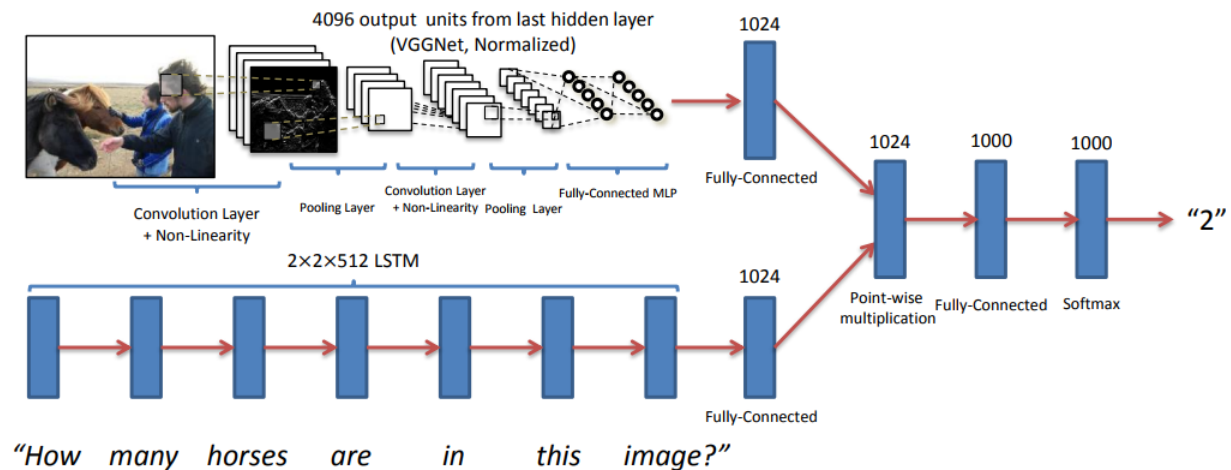
Visual Question Answering

Models

Variations

1) Image Channel

- I : use output of VGGNet(4096d) vectors as embedding
- Norm I: l_2 normalized output



2) Question Channel

- BoW Q: use top 1,000 words to create bag-of-words representations
- LSTM Q : use concatenation of last state (1024d)
- Deeper LSTM Q: with two layers (2048d)

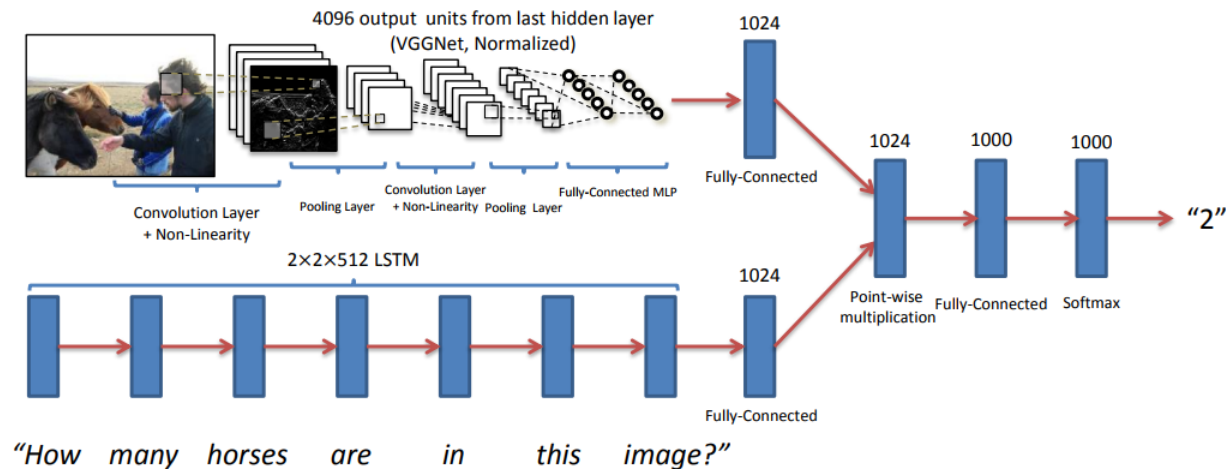
Visual Question Answering

Models

Variations

1) Image Channel

- I : use output of VGGNet(4096d) vectors as embedding
- Norm I: l_2 normalized output



2) Question Channel

- BoW Q: use top 1,000 words to create bag-of-words representations
- LSTM Q : use concatenation of last state (1024d)
- deeper LSTM Q: with two layers (2048d)

3) Multi-Layer Perceptron

- BoW Q+I: simply concatenate
- LSTM Q+I, deeper LSTM Q+I : transform image embeddings to 1024d by FC layer + tanh, then fused with question embeddings via element-wise multiplication

Visual Question Answering

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



1. Use bag-of-words representation of top 1,000 words in caption as caption embedding
2. Concatenate BoW Q and Caption embeddings

Visual Question Answering

Results

Question Type	Open-Ended			Human Age		Commonsense	
	K = 1000			Human		To Be Able	To Be Able
	Q	Q + I	Q + C	Q	Q + I	To Answer	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

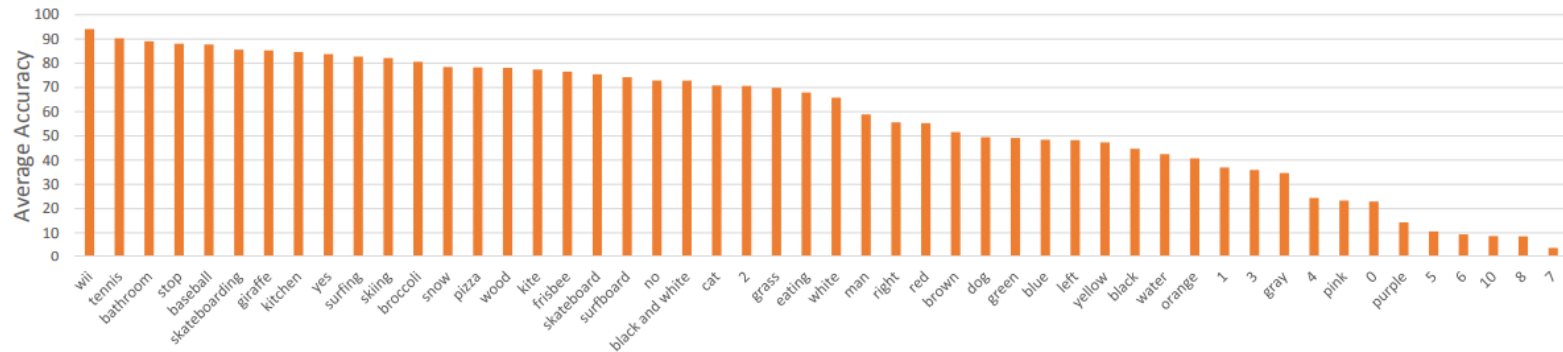
Questions that require
more reasoning

Questions that require
scene-level informations

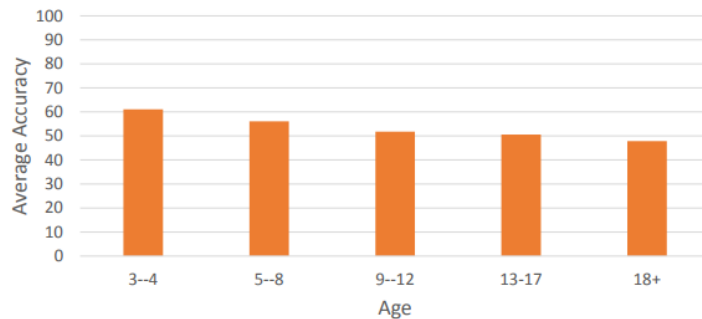
Visual Question Answering

Results

Accuracy on questions with 50 most frequent ground truth answers



Answers for Common visual object



Answers for counts

