VQA: Visual Question Answering

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

Presented by: Surbhi Goel

Note: Images and tables that have not been cited have been taken from the above-mentioned paper

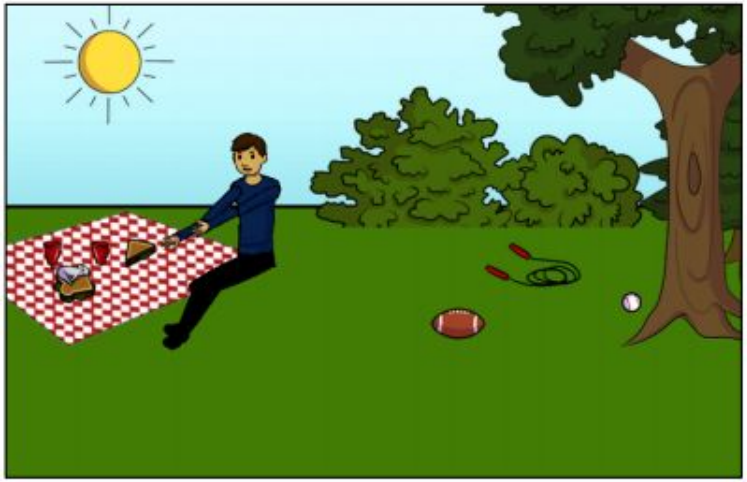
Outline

- VQA Task
- Importance of VQA
- Dataset Analysis
- Human Accuracy
- Model Comparison for VQA
- Common Sense Knowledge
- Conclusion
- Future Work
- Discussion

Visual Question Answering

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?

Importance of VQA

- Multi modal task - a step towards solving AI
- Allows automatic quantitative evaluation
- Useful applications eg. answer questions asked by visually-impaired users

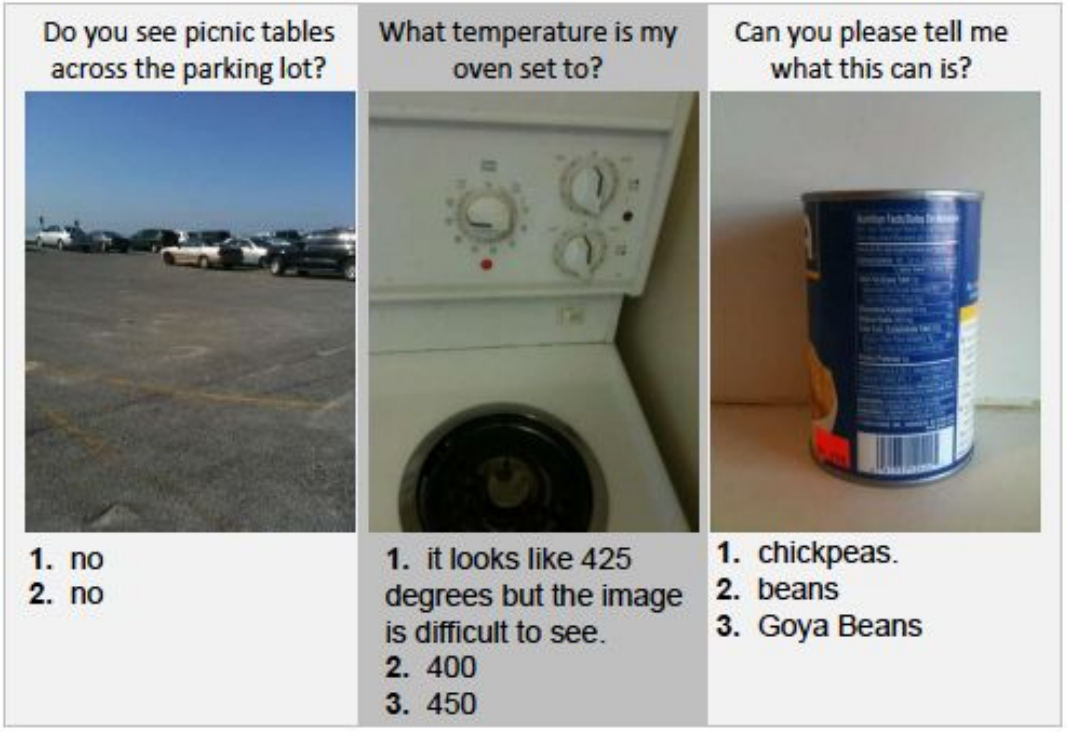


Image credits: Dhruv Batra

Dataset

- **>250K images**
  - 200K from MS COCO
    - 80 train / 40 val / 80 test
  - 50K from Abstract
- **QAs**
  - 3 questions/image
  - 10 answers/question
    - +3 answers/question without showing the image
- **>760K questions**
- **~10M answers**
  - will grow over the years

Stump a smart robot!

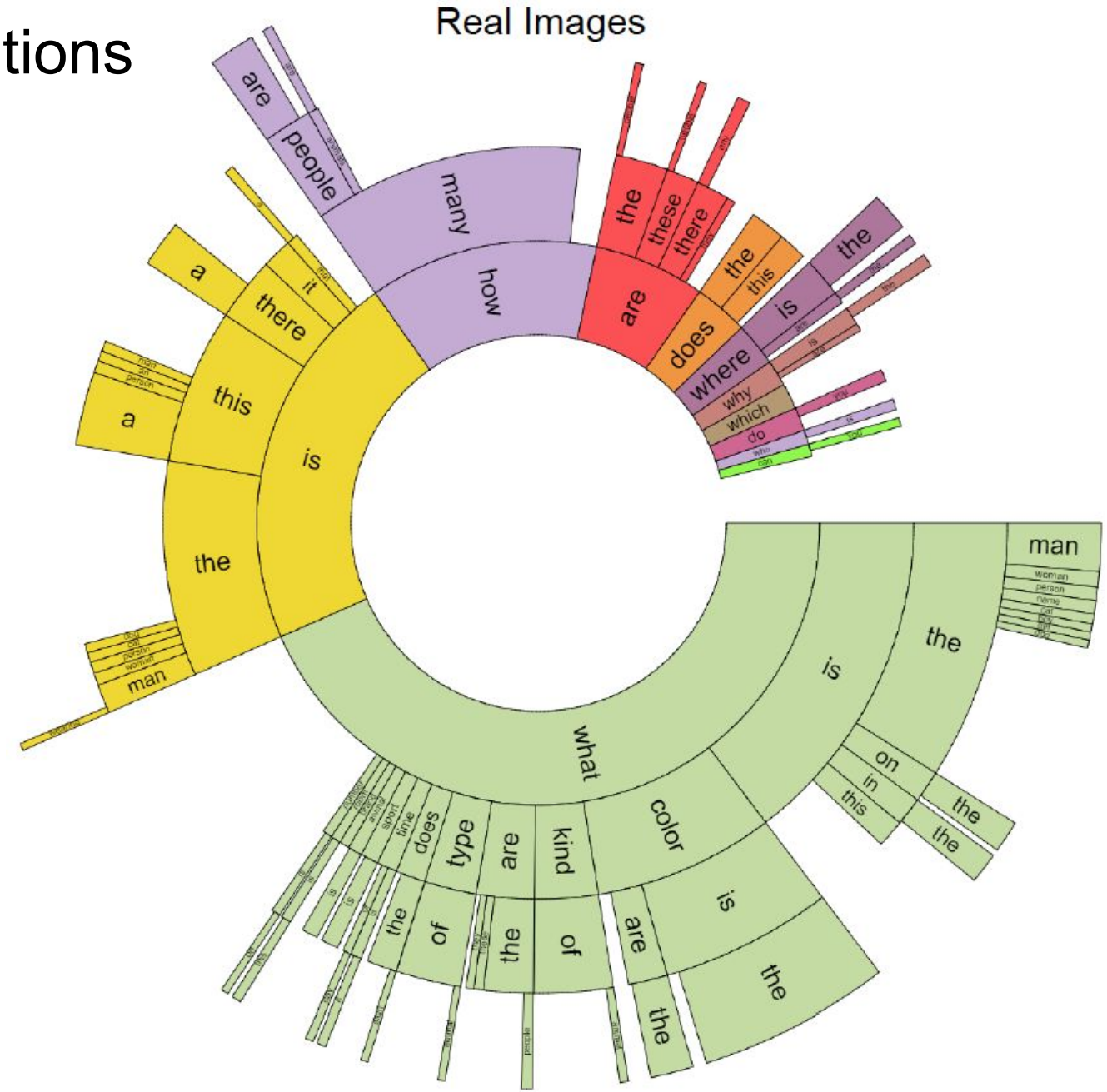
Ask a question that a human can answer,
but a smart robot probably can't!

- Mechanical Turk
- **>10,000 Turkers**
- **>41,000 Human Hours**

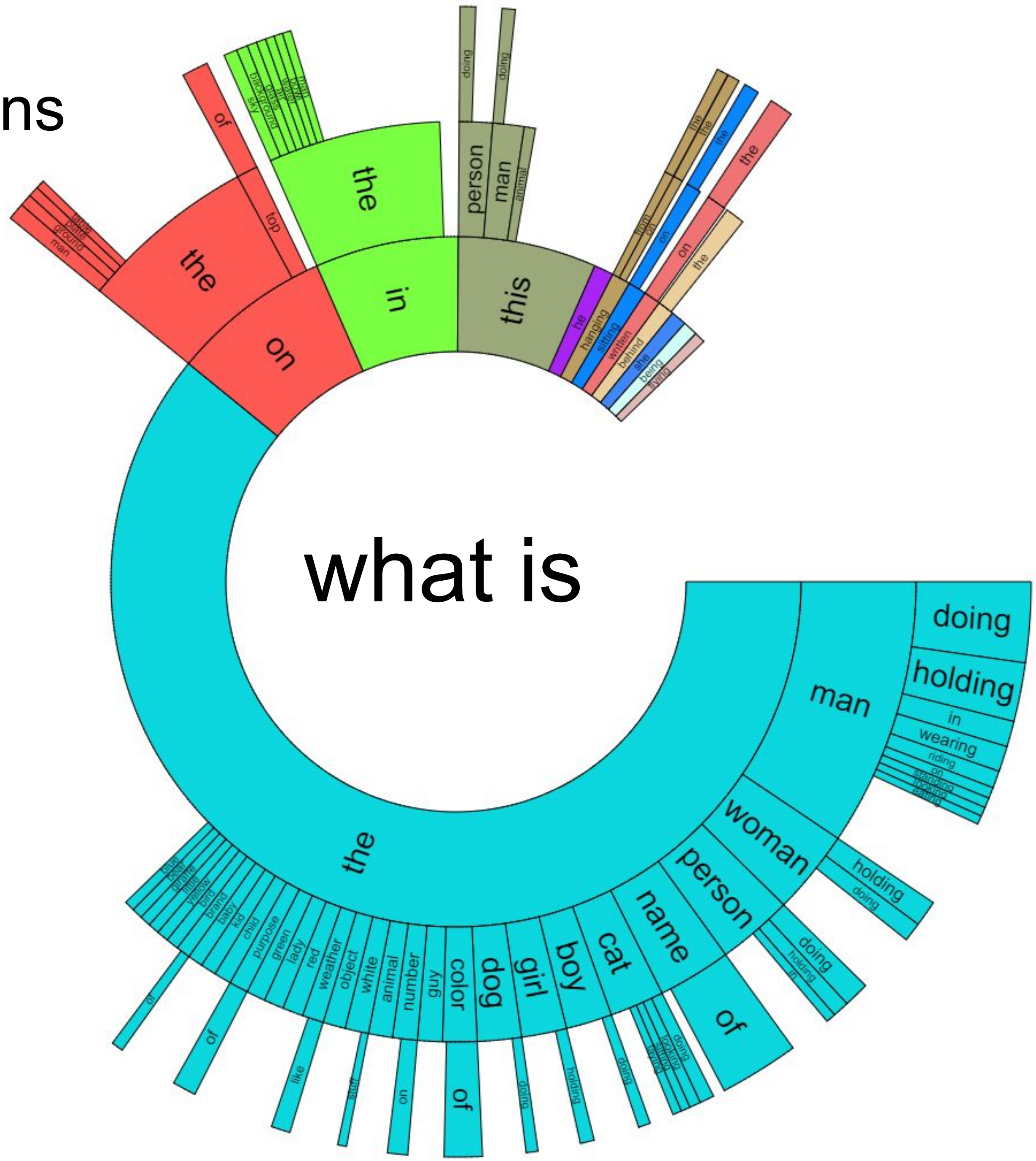
4.7 Human *Years!*
20.61 Person-Job-Years!

Slide credits: Dhruv Batra

Questions


Real Images
are
people
many
how
the
these
there
are
does
where
why
which
do
who
can
the
this
is
the
this
is
the
this
is
a
there
it
a
this
a
man
woman
person
what
is
the
man
woman
person
on
in
this
the
the
color
is
the
kind
are
the
type
are
the
of
the
of
does
time
short
the
of
are
the
is
the

Questions


what is
the
on
in
this
man
woman
person
name
cat
boy
girl
dog
color
doing
holding
in
wearing
of
the
the
person
man
doing
doing
on
the

Answers

- 38.4% of questions are binary yes/no
  - 39.3% for abstract scenes
- 98.97% questions have answers <= 3 words
  - 23k unique 1 word answers

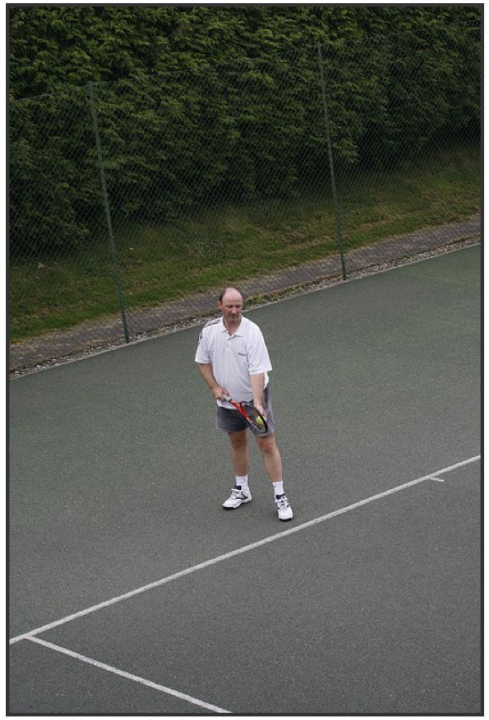
- Two evaluation formats:
  - Open answer
    - Input = question
  - Multiple choice
    - Input = question + 18 answer options
    - Options = correct / plausible / popular / random answers

Slide credits: Dhruv Batra

Answers

2627. COCO_train2014_000000044093

Image On/Off



Open-Ended Answers/Multiple-Choice Options

Q: what is he playing?

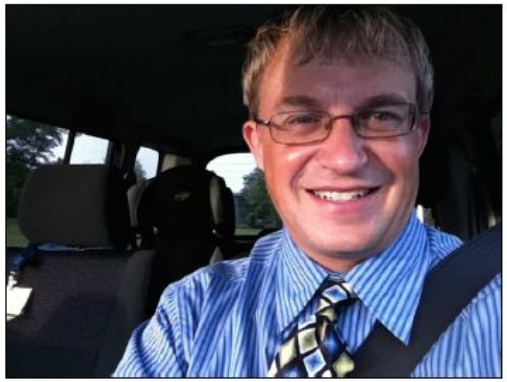
Answers with Image (*i.e.*, ground-truth):	Answers without Image (*i.e.*, commonsense):
(a) tennis	(a) guitar
(b) tennis	(b) drums
(c) tennis	(c) tennis

Q: is he standing outside the playing area?

Answers with Image (*i.e.*, ground-truth):	Answers without Image (*i.e.*, commonsense):
(a) yes	(a) no
(b) yes	(b) yes
(c) yes	(c) yes

Slide credits: Dhruv Batra

Answers



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

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

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

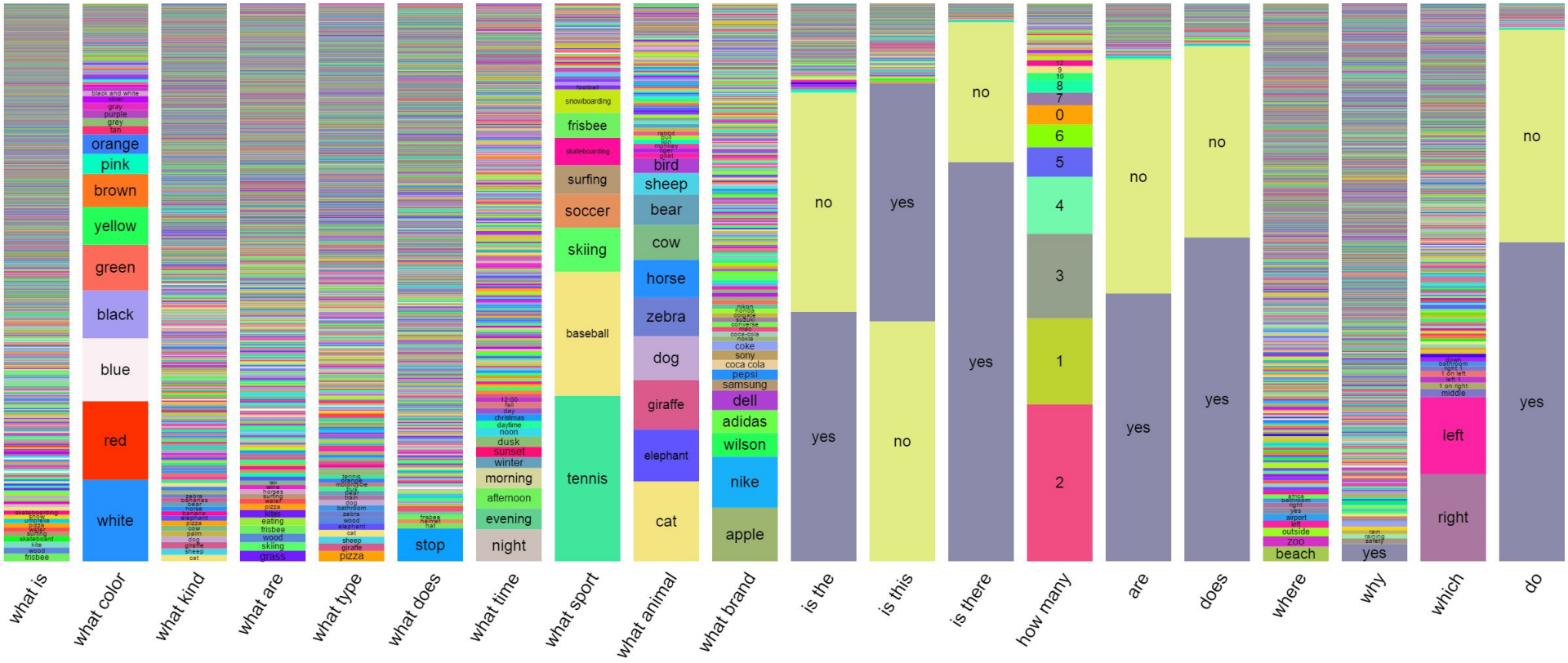
How many glasses are on the table?	3 3 3	2 2 6
What is the woman reaching for?	door handle glass wine	fruit glass remote

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

Slide credits: Dhruv Batra

Answers

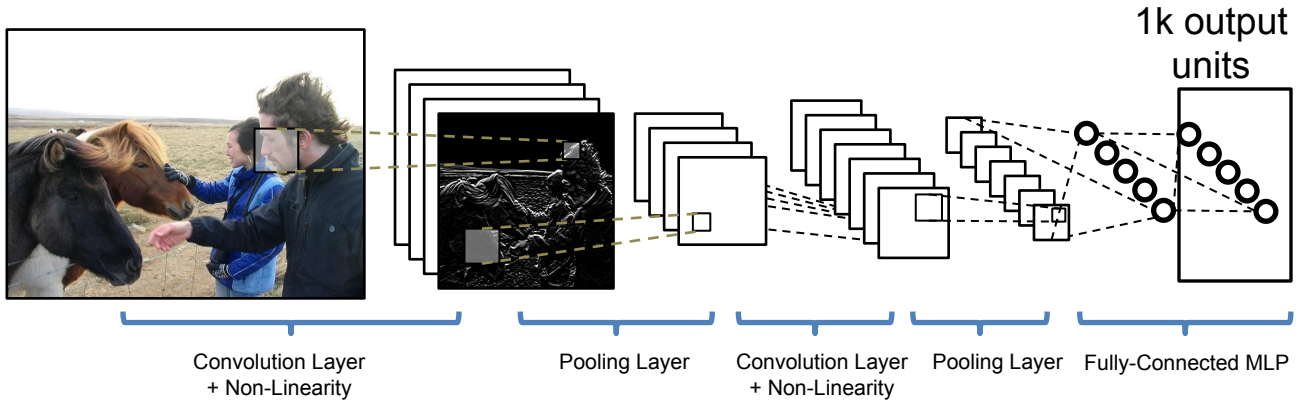

what is
what color
what kind
what are
what type
what does
what time
what sport
what animal
what brand
is the
is this
is there
how many
are
does
where
why
which
do
white
red
blue
black
green
yellow
brown
pink
orange
stop
night
tennis
cat
apple
yes
no
2
1
3
4
5
6
0
7
8
beach
left
right
middle

Human Accuracy

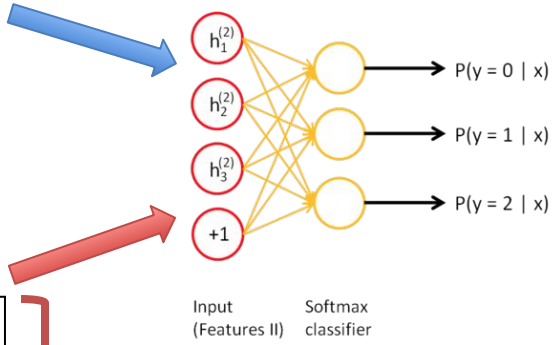
Dataset	Input	All	Yes/No	Other
Real	Question	40.81	67.60	21.22
	Question + Caption*	57.47	78.97	44.41
	Question + Image	83.30	95.77	72.67
Abstract	Question	43.27	66.65	23.66
	Question + Caption*	54.34	74.70	40.18
	Question + Image	87.49	95.96	75.33

VQA Model

Image Embedding



Neural Network Softmax over top K answers



Question Embedding (BoW)

"How many horses are in this image?"

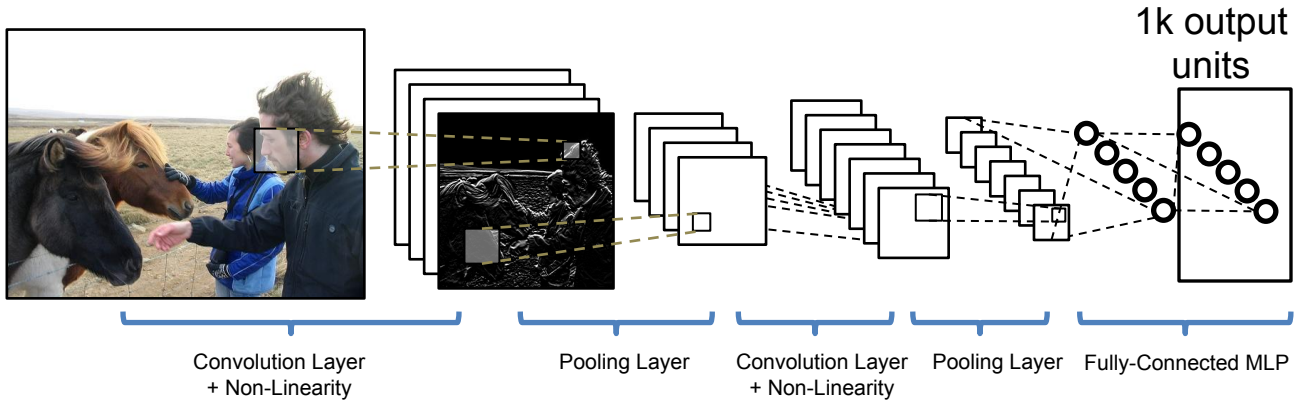
what	0
where	0
how	1
is	0
could	0
are	0
…	
are	1
…	
horse	1
…	
image	1

Beginning of question words

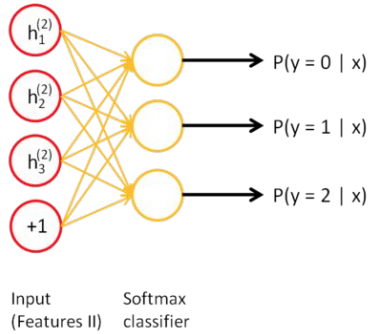
Slide credits: Dhruv Batra

VQA Model

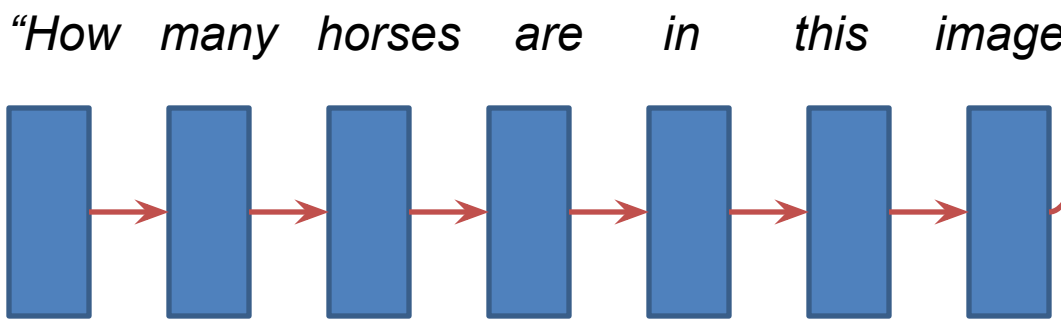
Image Embedding



Neural Network Softmax over top K answers



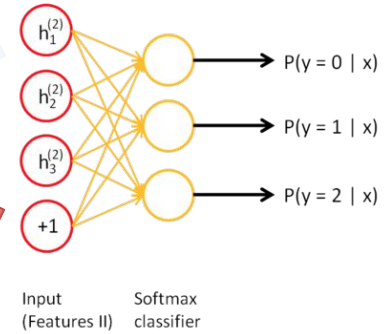
Question Embedding (LSTM)



Slide credits: Dhruv Batra

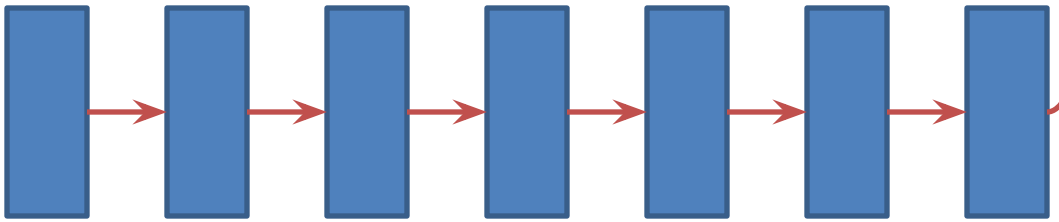
Baseline #1 - Language-alone

Neural Network
Softmax
over top K answers



Question Embedding (LSTM)

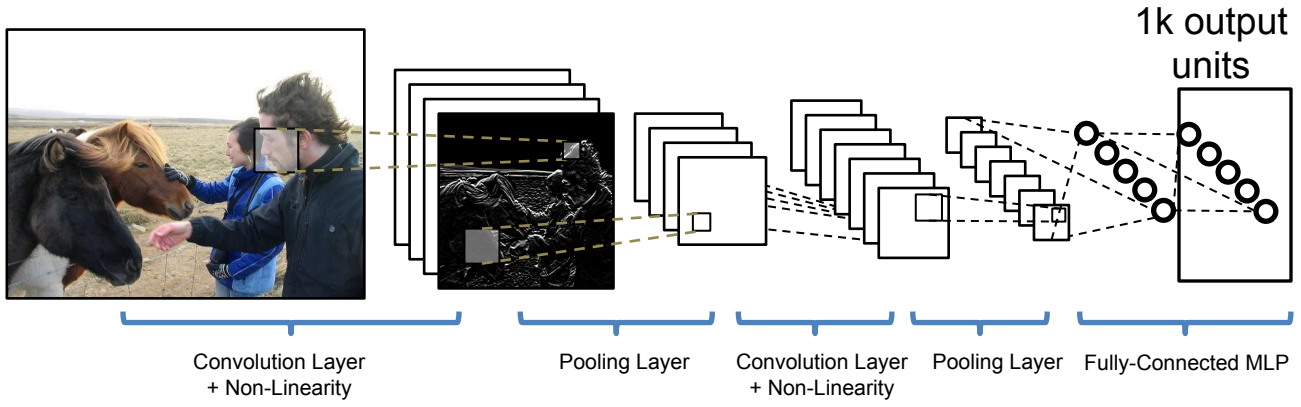
"How many horses are in this image?"



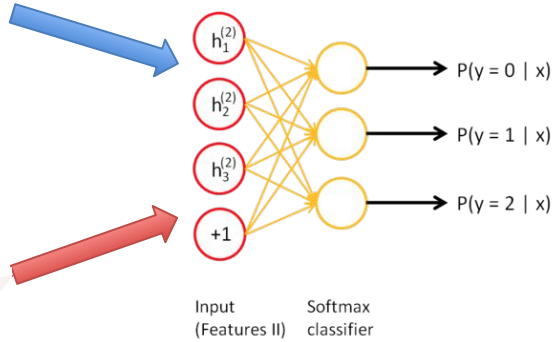
Slide credits: Dhruv Batra

Baseline #2 - Vision-alone

Image Embedding



Neural Network Softmax over top K answers



Question Embedding (LSTM)

"How many horses are in this

Slide credits: Dhruv Batra

Results

	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

Challenge: Common Sense

Does the person have perfect vision?

Evaluate Common Sense in the Dataset

Asked users:

- Does the question require common sense?
- How old should a person be to answer the question?

COCO_val2014_000000318571.jpg-3

Question: What color is the cat?

Average Commonsense:	Average Age: 4.100
0.000	toddler (3-4): 8
no: 10	younger child (5-8): 2

Image credits: Dhruv Batra

Evaluate Common Sense in the Dataset

Asked users:

- Does the question require common sense?
- How old should a person be to answer the question?

COCO_train2014_000000411247.jpg-3

Question: How many calories are in this pizza?

Average Age: 19.200

teenager (13-17): 3

adult (18+): 7

Average Commonsense: 0.900

yes: 9

no: 1

Image credits: Dhruv Batra

Conclusion

- Compelling ‘AI-complete’ task
- Combines a range of vision problems in one, such as
  - Scene Recognition
  - Object Recognition
  - Object Localization
  - Knowledge-base Reasoning
  - Commonsense
- Far from achieving human levels

Future Work

- **Dataset**
  - Extend the dataset
  - Create task-specific datasets eg. visually-impaired
- **Model**
  - Exploit more image related information
  - Identify the task and then use existing systems

Challenge and workshop to promote systematic research (www.visualqa.org)

Discussion Points (Piazza)

- Should a different evaluation metric (such as METEOR) be used?
- How to collect questions faster (compared to humans)?
- Is the length restriction on the answers limiting the scope of the task?
- Since distribution of question types is skew, will it bias a statistical learner to answer only certain types of questions?
- Why use ‘realistic’ abstract scenes for the task?
- Why does LSTM not perform well?
- Would using Question + Image + Caption give better results than using Question + Image ?
- Should we focus on task specific VQA?

Thank You!