embedded VISION SUMMIT 2018

Words, Pictures, and Common Sense

Devi Parikh
School of Interactive Computing, Georgia Tech
Facebook AI Research

"Color College Avenue", Blacksburg, VA, May 2012

People coloring a street in rural Virginia.

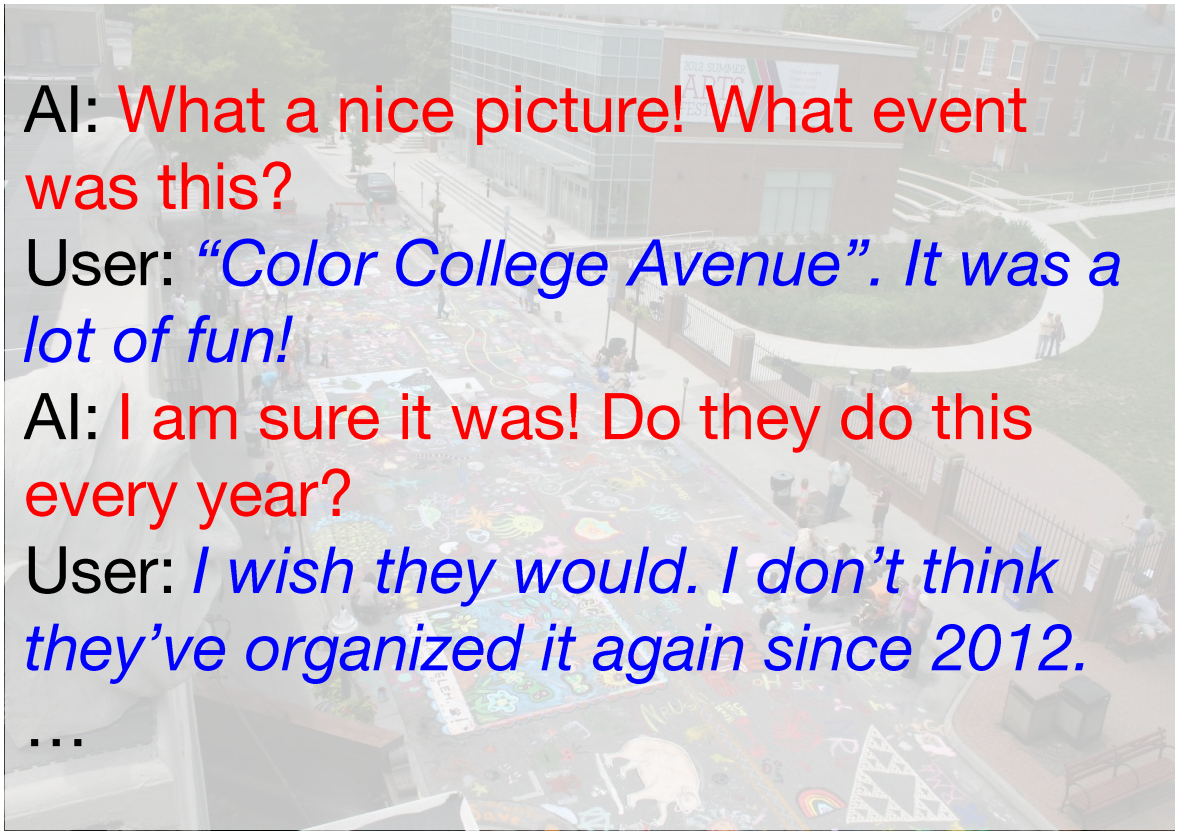
"Color College Avenue", Blacksburg, VA, May 2012

It was a great event! It brought families out, and the whole community together.

"Color College Avenue", Blacksburg, VA, May 2012

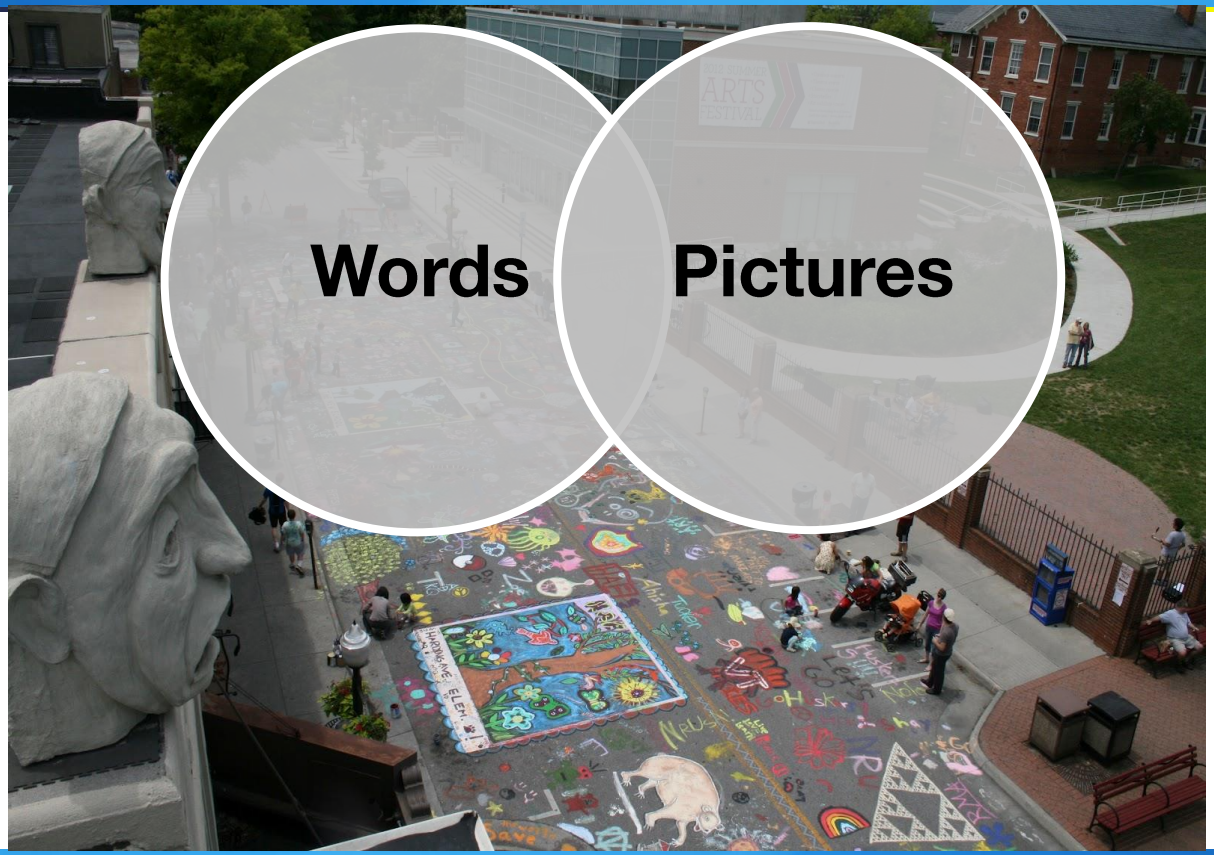
Q. What are they coloring the street with?
A. Chalk

"Color College Avenue", Blacksburg, VA, May 2012

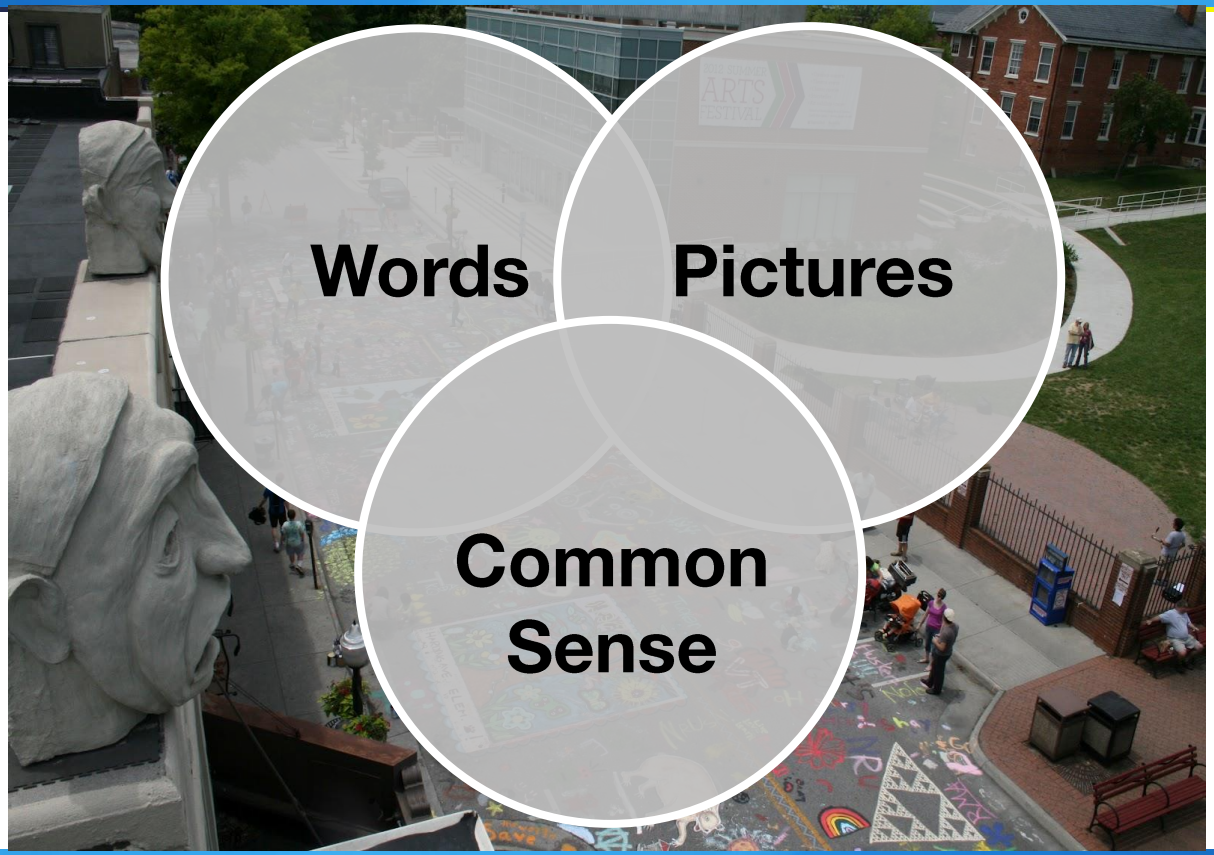


"Color College Avenue", Blacksburg, VA, May 2012

"Color College Avenue", Blacksburg, VA, May 2012



"Color College Avenue", Blacksburg, VA, May 2012



"Color College Avenue", Blacksburg, VA, May 2012

Why Words and Pictures? 1

Pictures are everywhere

Words are how we communicate

Why Words and Pictures? 1

Applications

Applications

Interact with, organize, and navigate visual data

Why Words and Pictures? 1

Applications

Leverage multi-modal information on the web

Applications

Aid visually-impaired users

Microsoft

Applications

Aid visually-impaired users

FACEBOOK'S AI CAN CAPTION PHOTOS FOR THE BLIND ON ITS OWN

Why Words and Pictures? 1

Applications

Summarize visual data for analysts

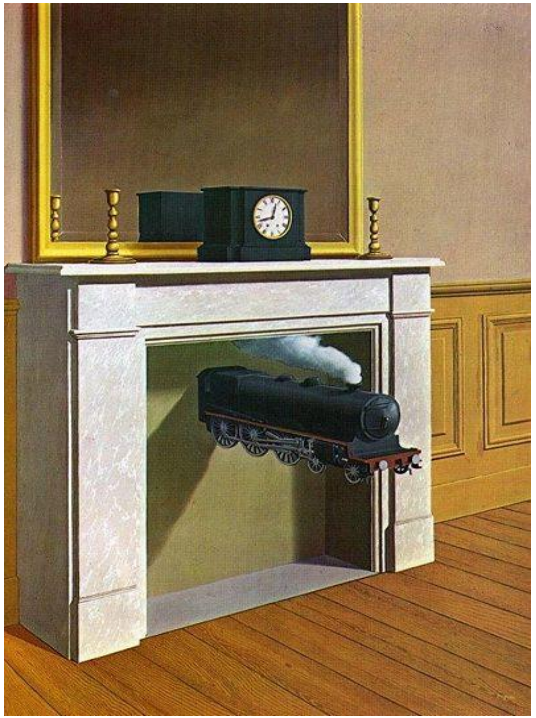
Why Words and Pictures? 2

- Measuring and demonstrating AI capabilities
  - Image understanding
  - Language understanding

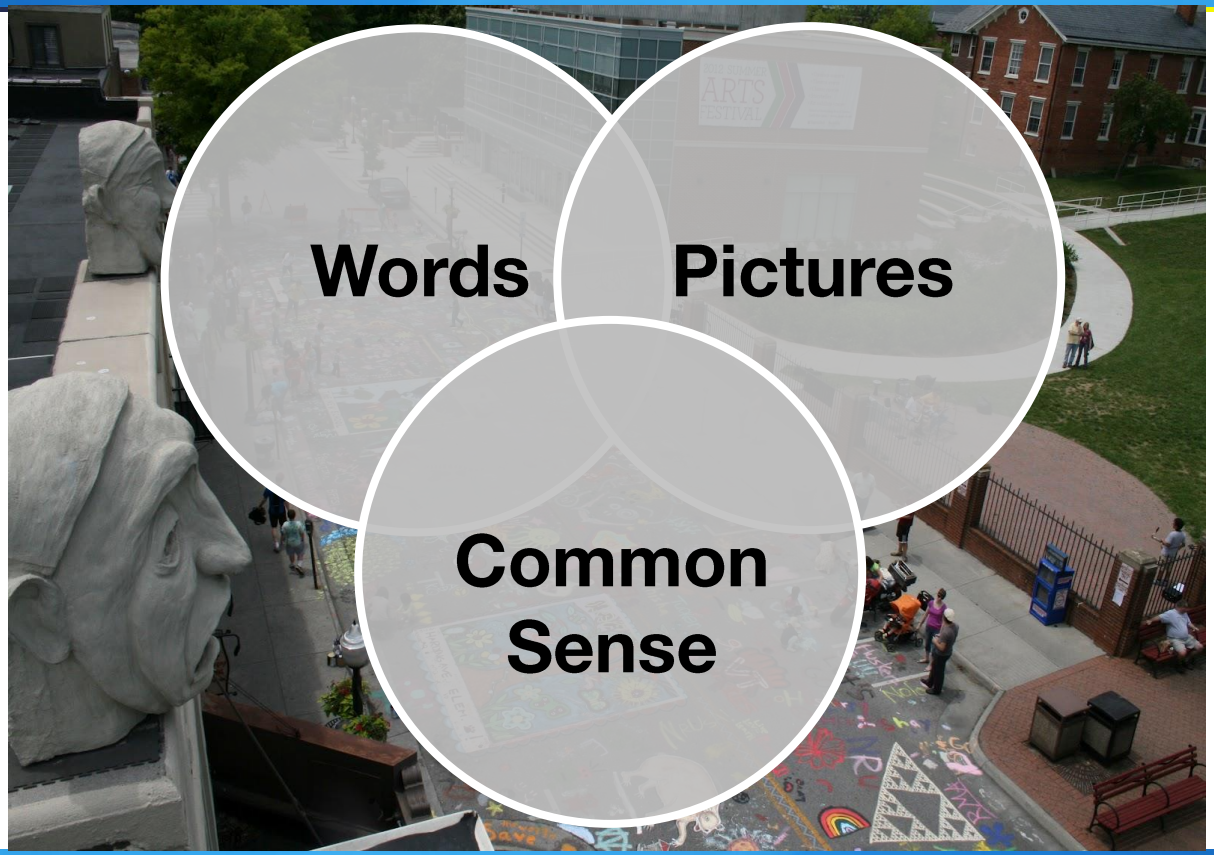
Why Words and Pictures? 3

- Beyond “bucket” recognition
- Language is compositional

“A steam engine is coming out of a fireplace.”



René Magritte (1938)



"Color College Avenue", Blacksburg, VA, May 2012

Visual Question Answering (VQA)

Visual Question Answering (VQA)

Visual Question Answering (VQA)

What is the mustache made of?

Visual Question Answering (VQA)

What is the mustache made of?

AI System

Visual Question Answering (VQA)

What is the mustache made of?

AI System

bananas

CloudCV: Large Scale Dist

cloudcv.org/vqa/

CloudCV | Image Stitching | Object Detection | Decaf-Server | Classification | VIP | Train a new category

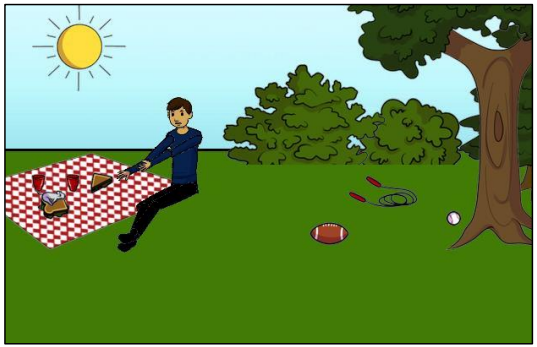
Ask any question about this image

Answer

Visual Question Answering (VQA)

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?

Visual Question Answering (VQA)

- Details of the image
- Common sense + knowledge base
- Task-driven
- Holy-grail of semantic image understanding

Visual Question Answering (VQA) Dataset

>0.25 million images

Visual Question Answering (VQA) Dataset

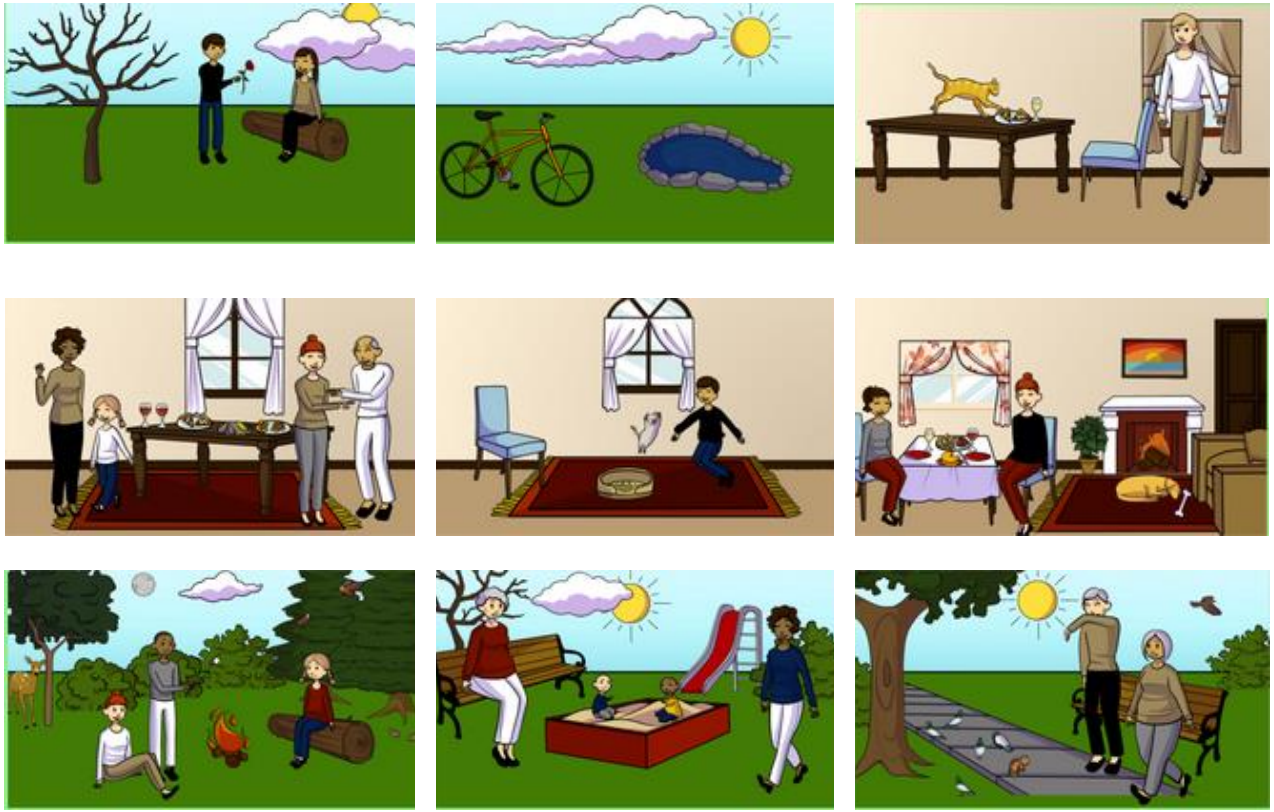







254,721 images (COCO)

Visual Question Answering (VQA) Dataset



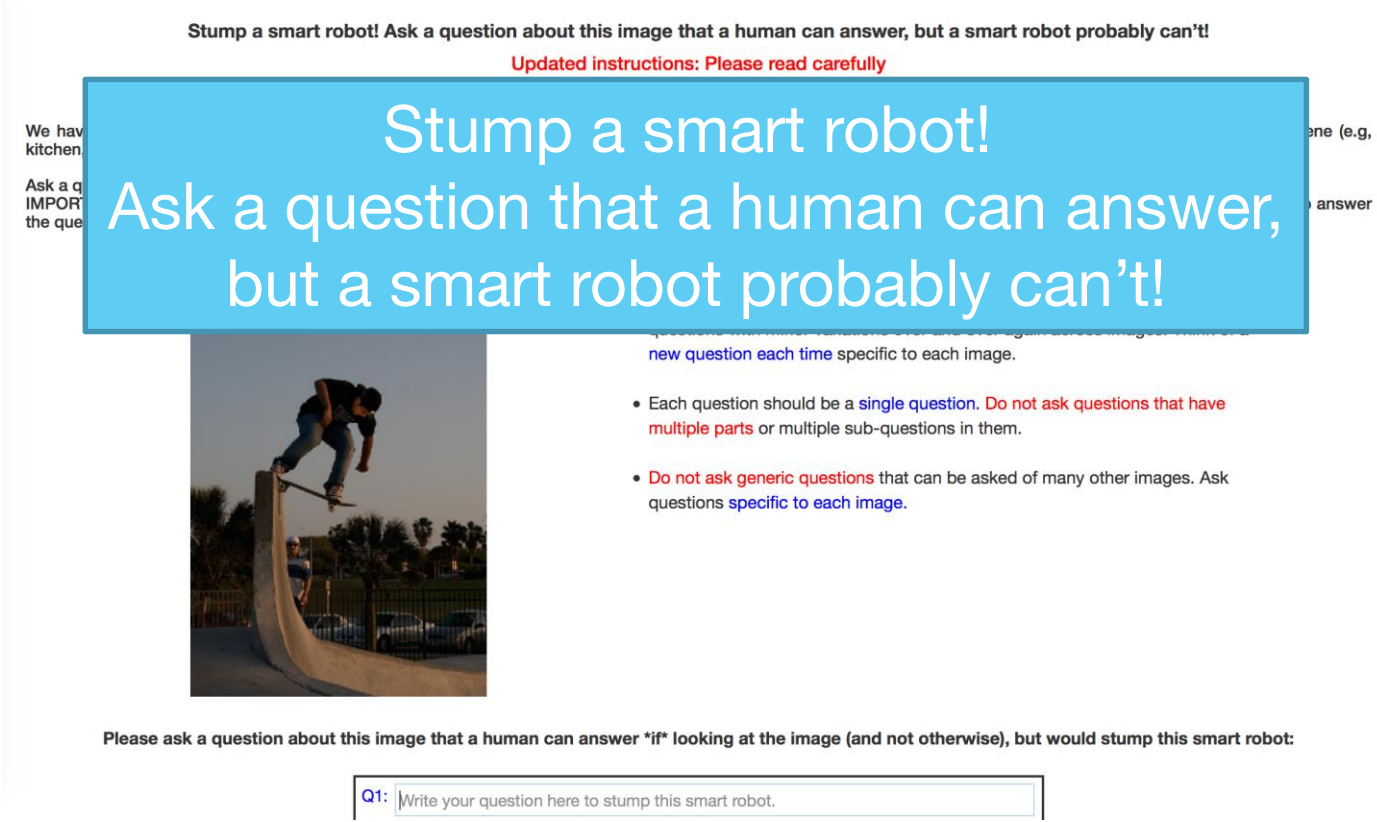
50,000 scenes

Visual Question Answering (VQA) Dataset

>0.25 million images

>0.76 million questions

Visual Question Answering (VQA) Dataset



Visual Question Answering (VQA) Dataset

>0.25 million images

>0.76 million questions

~10 million answers

Taxing the Turkers

- *Beware also the lasting effects of doing too many --for hours after the fact you will not be able to look at any photo without automatically generating a mundane question for it.*
- *If I were in possession of state secrets they could be immediately tortured out of me with the threat of being shown images of: skateboards, trains, Indian food and [long string of expletives] giraffes.*
- *(Please...I will tell you everything...just no more giraffes...)*

Reset

Top Answers

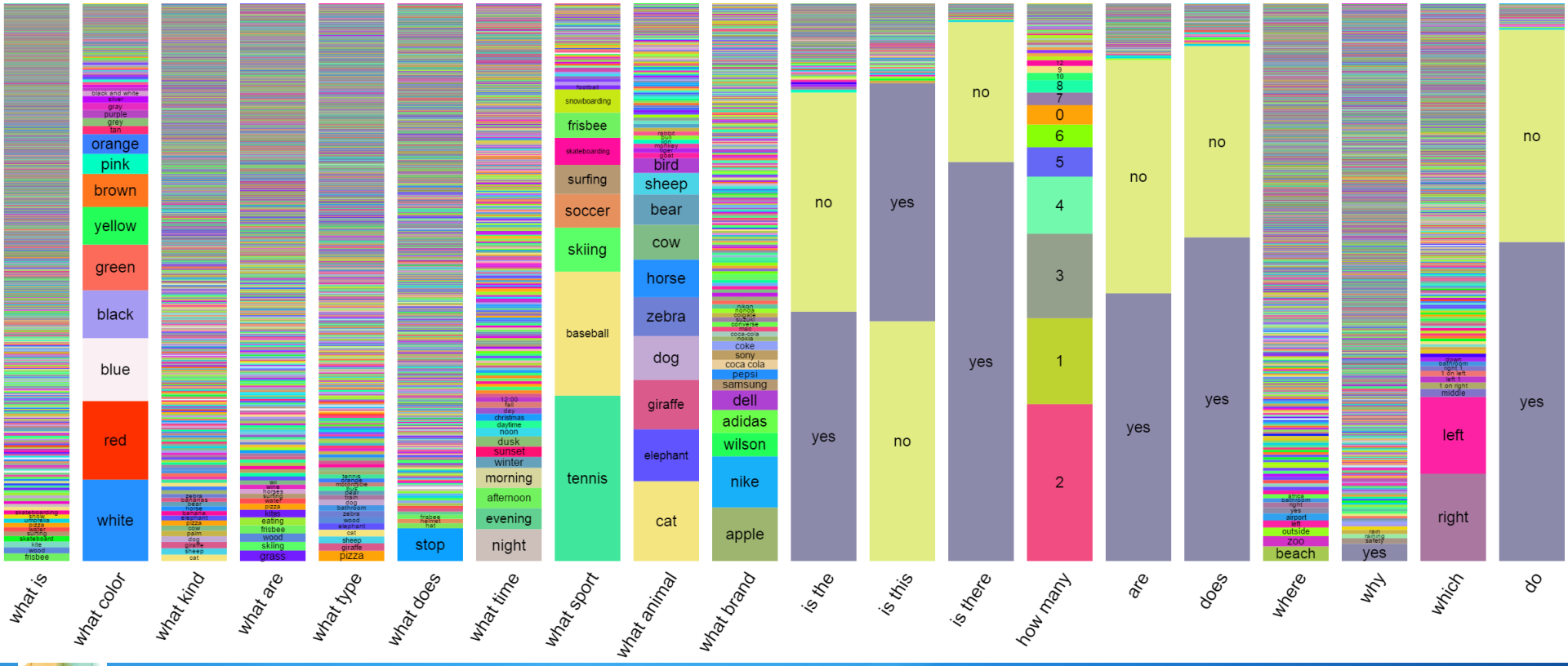
Answers

- 38% of questions are binary yes/no

- 99% questions have answers <= 3 words
  - Evaluation is feasible
  - 23k unique 1 word answers

CIDEr: Consensus-based Image Description Evaluation
[Vedantam, Zitnick, and Parikh, CVPR 2015]

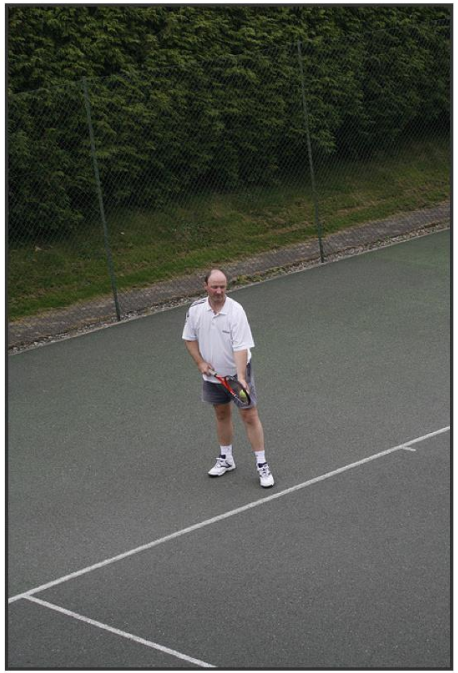
Answers



Evaluation Formats

- Open answer
  - Input = image, question
- Multiple choice
  - Input = image, question, 18 answer options
  - Avoids language generation
  - Evaluation (even more) feasible
  - Options = {correct, plausible, popular, random} answers

Plausible Answers



Q. What is he playing?

- guitar
- drums
- baseball

Accuracy Metric

$$\text{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

1940. COCO_train2014_000000012015

Open-Ended/Multiple-Choice/Ground-Truth

Q: WHAT OBJECT IS THIS

Ground Truth Answers:

(1) television	(6) television
(2) tv	(7) television
(3) tv	(8) tv
(4) tv	(9) tv
(5) television	(10) television

Q: How old is this TV?

Ground Truth Answers:

(1) 20 years	(6) old
(2) 35	(7) 80 s
(3) old	(8) 30 years
(4) more than thirty years old	(9) 15 years
(5) old	(10) very old

Q: Is this TV upside-down?

Ground Truth Answers:

(1) yes	(6) yes
(2) yes	(7) yes
(3) yes	(8) yes
(4) yes	(9) yes
(5) yes	(10) yes

Human Accuracy, Inter-Human Agreement

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

Human Accuracy, Inter-Human Agreement

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

Human Accuracy, Inter-Human Agreement

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

Human Accuracy, Inter-Human Agreement

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

Model

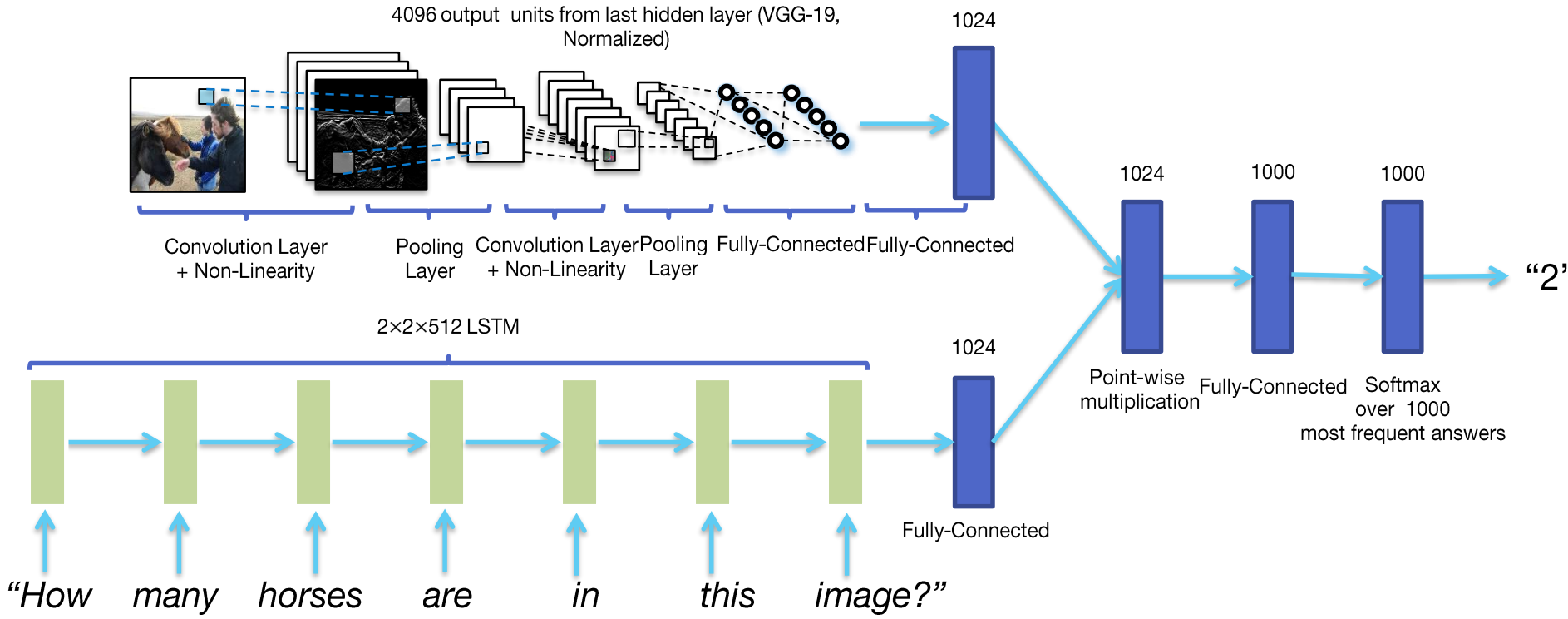
- Input: Image, Question
- Output: Answer
- Image:
  - Convolutional Neural Network (CNN)

    [Fukushima 1980, LeCun et al. 1989]
- Question:
  - Recurrent Neural Network
  - Specifically, a Long Short-Term Memory (LSTM)

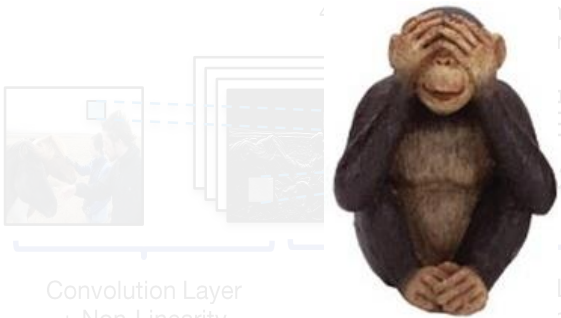
    [Hochreiter & Schmidhuber, 1997]
- Output: 1 of K most common answers

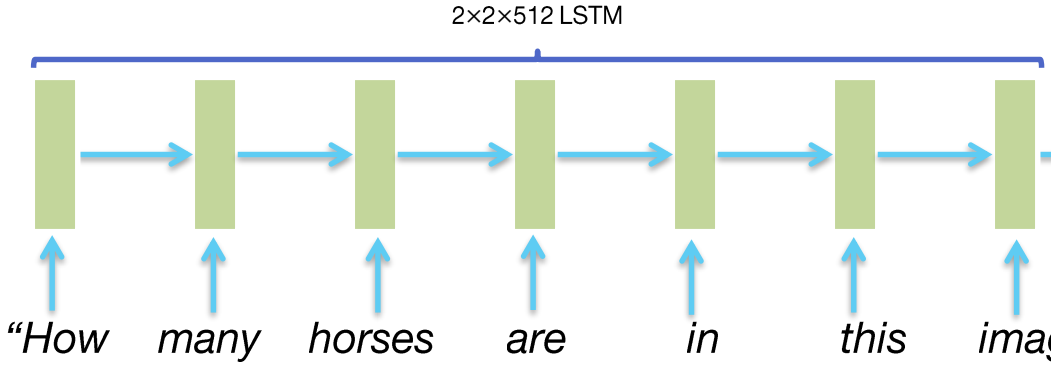
What color are her eyes?

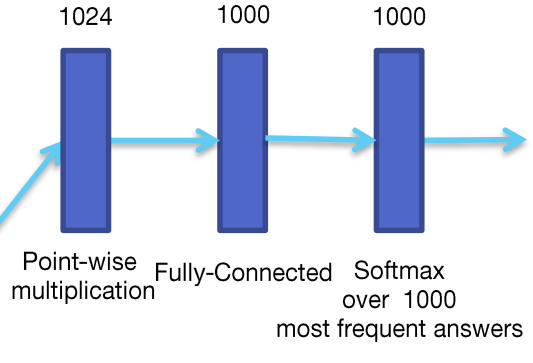
Model



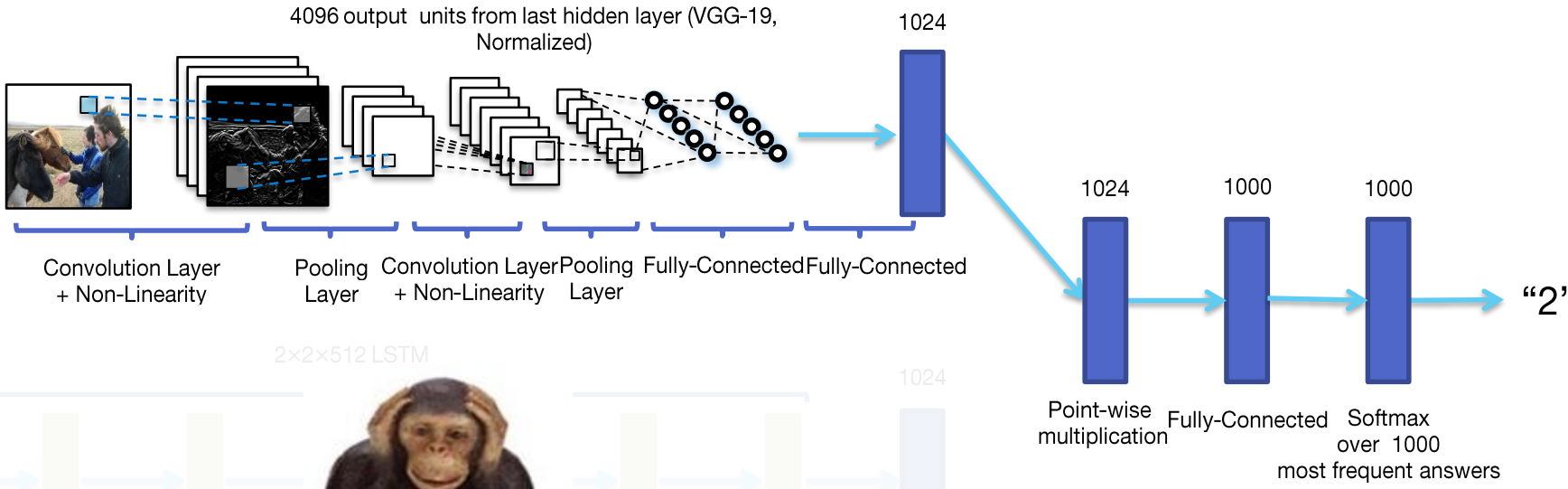
Model







Model



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	50.39	78.41	34.68	30.03	55.88	78.45	35.91	41.13
LSTM Q+I	57.75	80.5	36.77	43.08	62.7	80.52	38.22	53.01
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

VQA Age

How Old Do You Think a Person Needs to be to Answer These Questions?

We will present you with a series of questions about images. For each question, please select **the youngest age group** that you think a person must be in order to be able to correctly answer the question.

To answer this question, I would expect a person to have to at least be a:

- 1. toddler (3-4)
- 2. younger child (5-8)
- 3. older child (9-12)
- 4. teenager (13-17)
- 5. adult (18+)

VQA Age

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?

VQA Age

Question	Average Age
what brand	12.5
why	11.18
what type	11.04
what kind	10.55
is this	10.13
what does	10.06
what time	9.81
who	9.58
where	9.54
which	9.32
does	9.29
do	9.23
what is	9.11
what are	9.04
are	8.65
is the	8.52
is there	8.24
what sport	8.06
how many	7.67
what animal	6.74
what color	6.6

VQA Age

- This baseline model =* 4.74 years old!
- Average “age of questions” = 8.98 years.

* age as estimated by untrained crowd-sourced workers in uncontrolled environment

Papers Using VQA

Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources

Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, Anthony Dick
School of Computer Science, The U

Simple Baseline for Visual Question Answering

long Tian[2], Sainbayar Sukhbaatar[2], Arthur Szlam[2], and Rob Fergus[2]

ABC-CNN: An Attention Based Convolutional Neural Network for Visual

Academia, industry, start ups

Jacob Andreas
Department of
U
{jda,rohrbach,t

Zichao Yang[1], Xiaodong He[2], Jianfeng Gao[2], Li Deng[2], Alex Smola[1]
[1]Carnegie Mellon University, [2]Microsoft Research, Redmond, WA 98052, USA
zichaoy@cs.cmu.edu, {xiaohe, jfgao, deng}@microsoft.com, alex@smola.org

oiem
gn

Papers Using VQA

ORAL SESSION

Image Captioning and Question Answering

Monday, June 27th, 9:00AM - 10:05AM.

These papers will also be presented at the following poster session

1 **Deep Compositional Captioning: Describing Novel Object Categories Without Paired Training Data.**
Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell

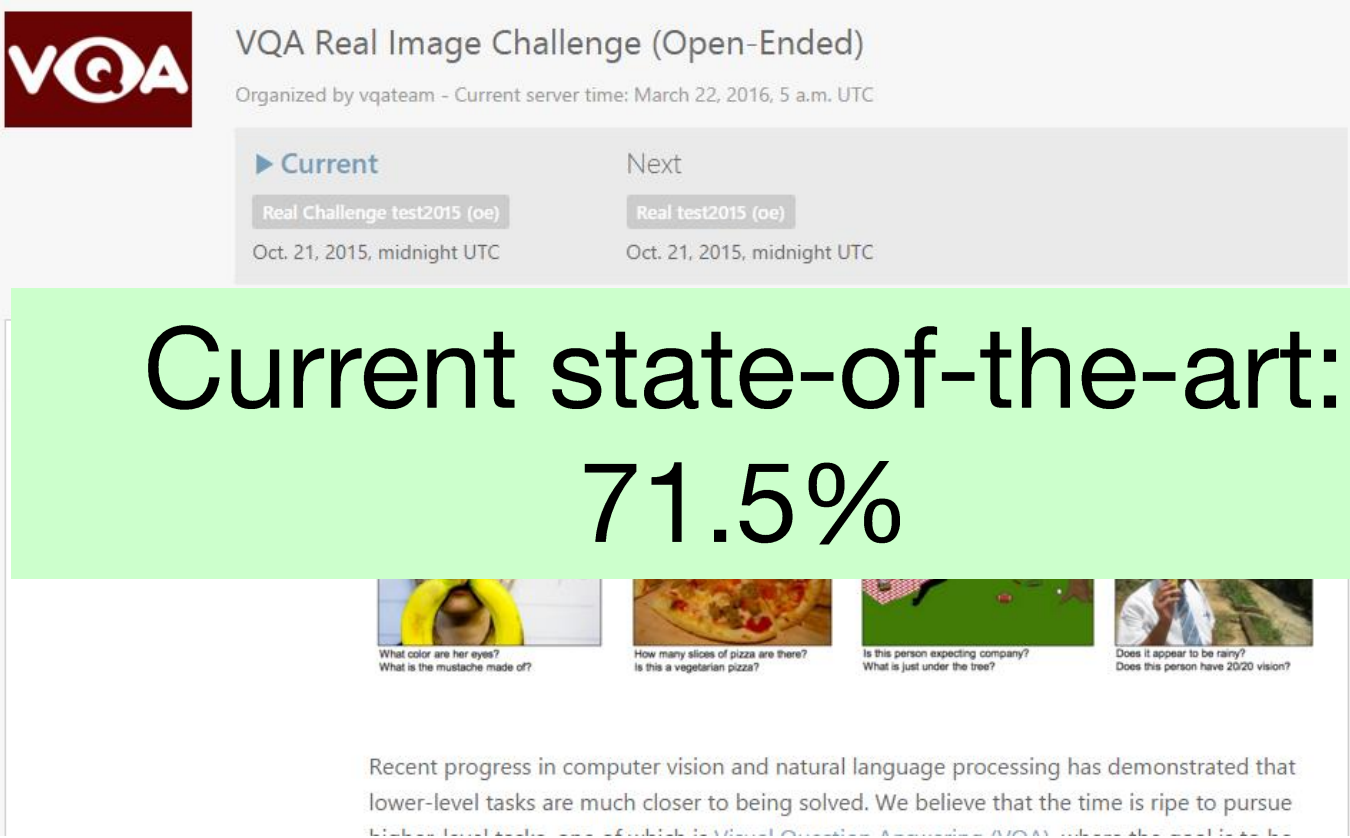
2 **Generation and Comprehension of Unambiguous Object Descriptions.**
Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, Kevin Murphy

3 **Stacked Attention Networks for Image Question Answering.**
Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola

4 **Image Question Answering Using Convolutional Neural Network With Dynamic Parameter Prediction.**
Hyeonwoo Noh, Paul Hongsuck Seo, Bohyung Han

5 **Neural Module Networks.**
Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Dan Klein

VQA Challenge @ CVPR 16, 17, 18



Current state-of-the-art:
71.5%

Hierarchical Co-Attention

Hierarchical Co-Attention

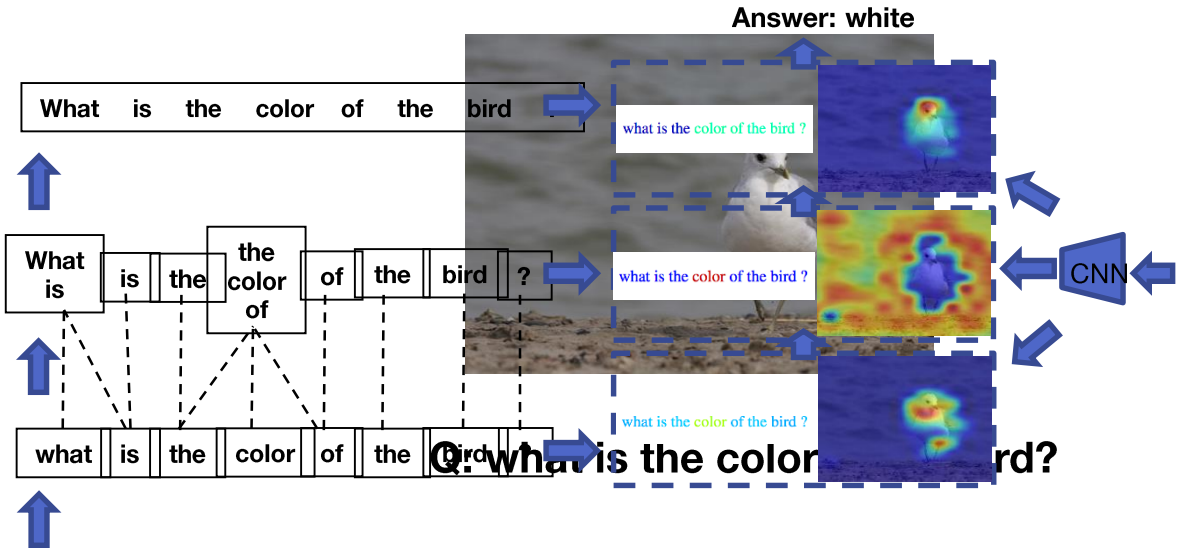
- Image attention: Decide where to look
- Question attention: Decide what to listen to

Alternating and Parallel

Hierarchical Co-Attention

- Hierarchical model of the question

Hierarchical Co-Attention



Slide credit: Jiasen Lu

What such a model can’t do

How many vegetarian slices are left in the pizza box?

It can’t count…

How many vegetarian slices are left in the pizza box?

It doesn’t have commonsense / knowledge…

How many vegetarian slices are left in the pizza box?

It can’t reason…

How many vegetarian slices are left in the pizza box?

It doesn't leverage compositionality…

How many vegetarian slices are left in the pizza box?

It lacks consistency…

How many vegetarian slices are left in the pizza box?

More in VQA

- Balancing (VQA v2.0) [CVPR 2017]

Who is wearing glasses?

man

woman

Where is the child sitting?

fridge

arms

Is the umbrella upside down?

yes

no

How many children are in the bed?

2

1

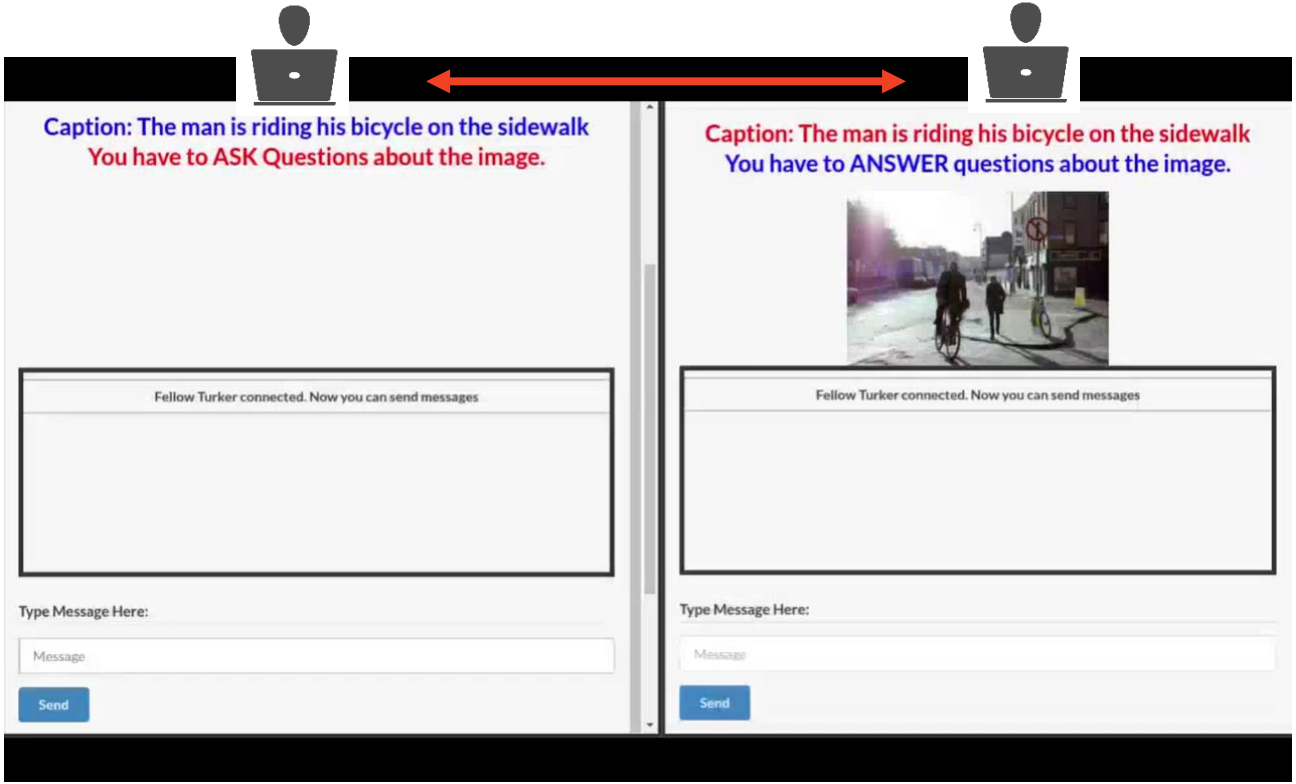
More in VQA

- Balancing (VQA v2.0) [CVPR 2017]
- Counting [CVPR 2017]
- VQA as multiple perspectives [ECCV 2016]
  - Image-caption ranking
- Human-like responses [EMNLP 2016]
- Just one-round of interaction
  - State-less, memory-less
  - Multiple rounds → Visual Dialog [CVPR 2017]

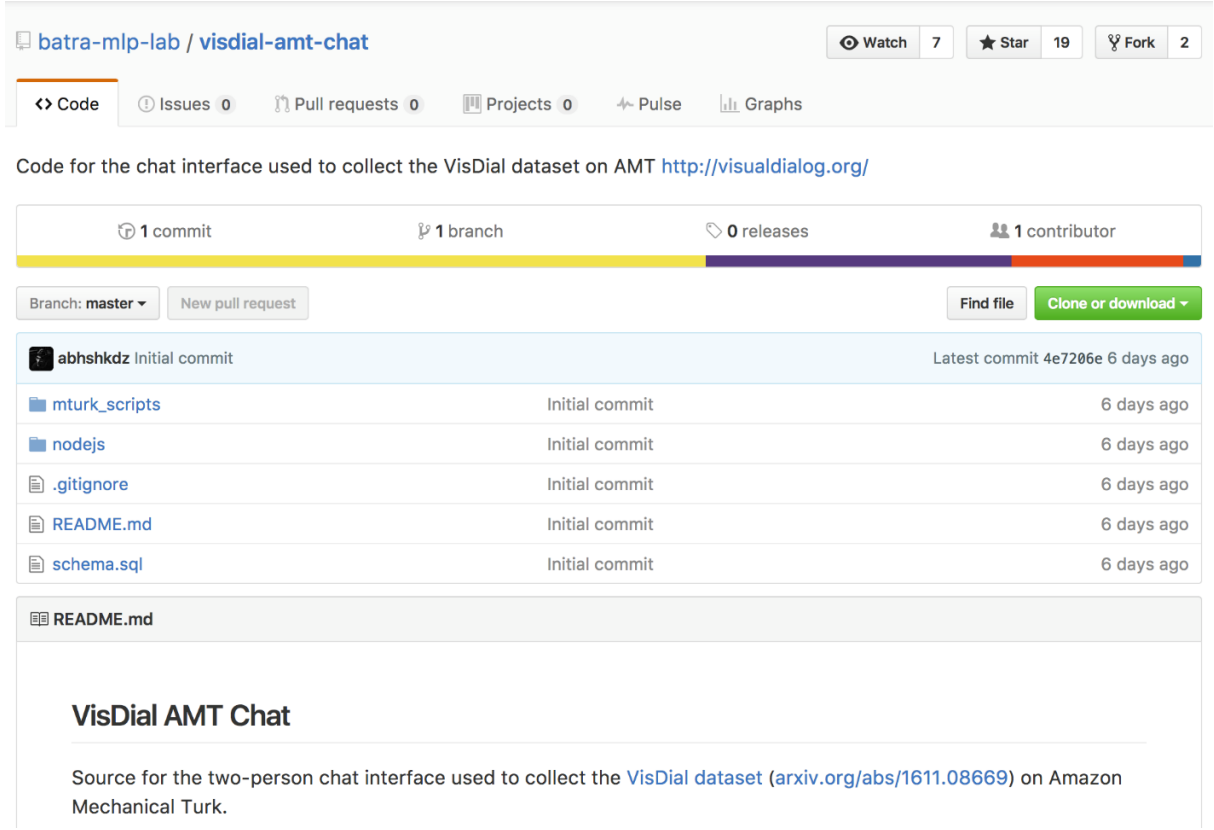
Visual Dialog

VisDial Dataset

VisDial Dataset



VisDial Dataset



batra-mlp-lab / visdial-amt-chat

Watch 7 | Star 19 | Fork 2

Code | Issues 0 | Pull requests 0 | Projects 0 | Pulse | Graphs

Code for the chat interface used to collect the VisDial dataset on AMT http://visualdialog.org/

1 commit | 1 branch | 0 releases | 1 contributor

Branch: master | New pull request | Find file | Clone or download

abhshkdz Initial commit — Latest commit 4e7206e 6 days ago

mturk_scripts	Initial commit	6 days ago
nodejs	Initial commit	6 days ago
.gitignore	Initial commit	6 days ago
README.md	Initial commit	6 days ago
schema.sql	Initial commit	6 days ago

README.md

VisDial AMT Chat

Source for the two-person chat interface used to collect the VisDial dataset (arxiv.org/abs/1611.08669) on Amazon Mechanical Turk.

VisDial v0.9 Stats

>120k images (from COCO)

1 dialog/image

10 question-answer rounds/dialog

Total of *>1.2 Million* dialog QA pairs

VisDial Dataset

Visual Dialog Overview People Data Bibtex Acknowledgements VirginiaTech Georgia Tech

VisDial Dataset

Code for the real-time chat interface used to collect the VisDial dataset on Amazon Mechanical Turk

VisDial v0.9

Training set (235M)
82,783 images

Validation set (108M)
40,504 images

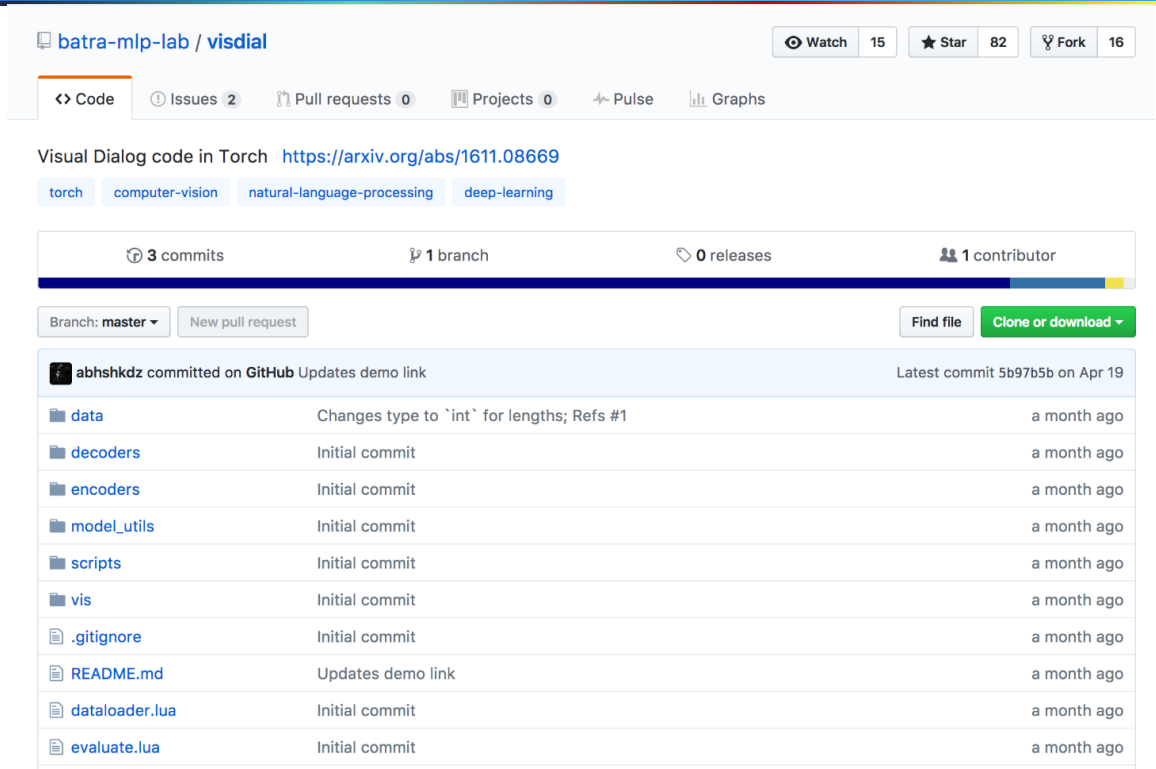
Readme

- v0.9 Training is from COCO Training and v0.9 Validation set is from COCO Validation
- Numbers (in papers, etc.) should be reported on v0.9 val

Format

```
[
  {
    'data': {
      'questions': [
        'does it have a doorknob',
        'do you see a fence around the bear',
        ...
      ],
      'answers': [
        'no, there is just green field in foreground',
        'countryside house',
        ...
```

www.visualdialog.org



batra-mlp-lab / visdial

Watch 15 | Star 82 | Fork 16

Code | Issues 2 | Pull requests 0 | Projects 0 | Pulse | Graphs

Visual Dialog code in Torch https://arxiv.org/abs/1611.08669

torch | computer-vision | natural-language-processing | deep-learning

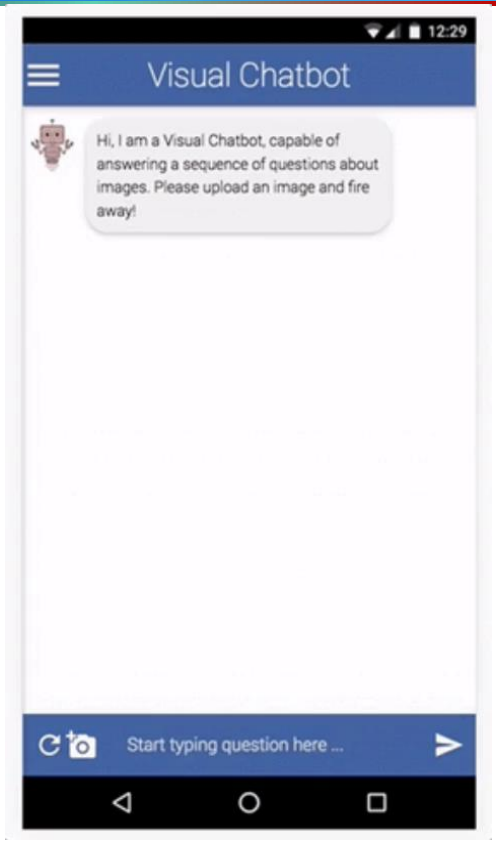
3 commits | 1 branch | 0 releases | 1 contributor

Branch: master | New pull request | Find file | Clone or download

abhshkdz committed on GitHub Updates demo link — Latest commit 5b97b5b on Apr 19

data	Changes type to `int` for lengths; Refs #1	a month ago
decoders	Initial commit	a month ago
encoders	Initial commit	a month ago
model_utils	Initial commit	a month ago
scripts	Initial commit	a month ago
vis	Initial commit	a month ago
.gitignore	Initial commit	a month ago
README.md	Updates demo link	a month ago
dataloader.lua	Initial commit	a month ago
evaluate.lua	Initial commit	a month ago

www.visualdialog.org



Embodied Question Answering

Embodied QA

Question: What color is the car?

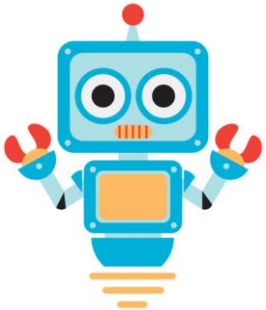
Type answer here

SUBMIT

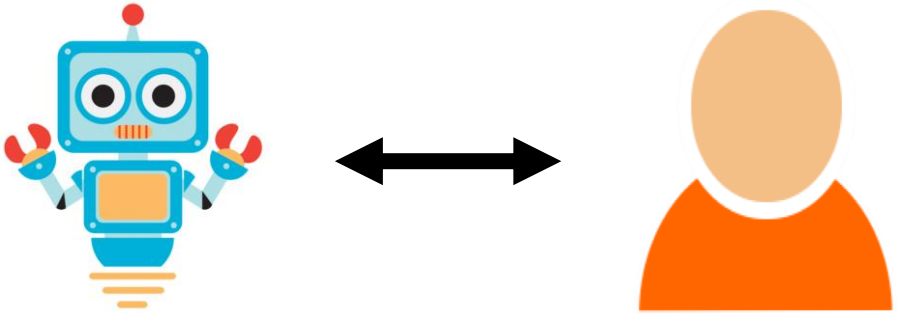
Human-AI Teams

As AI gets better…

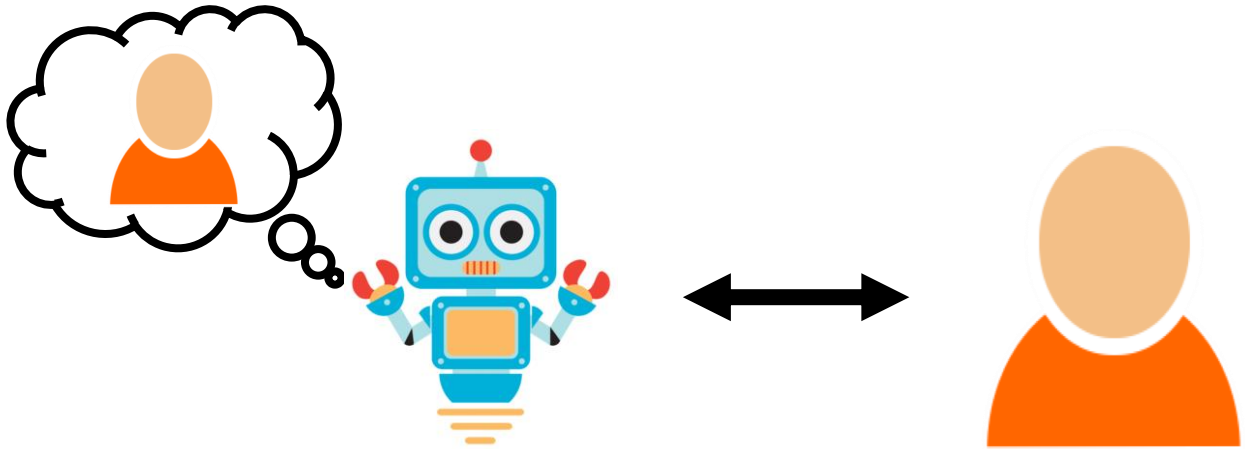
As AI gets better…



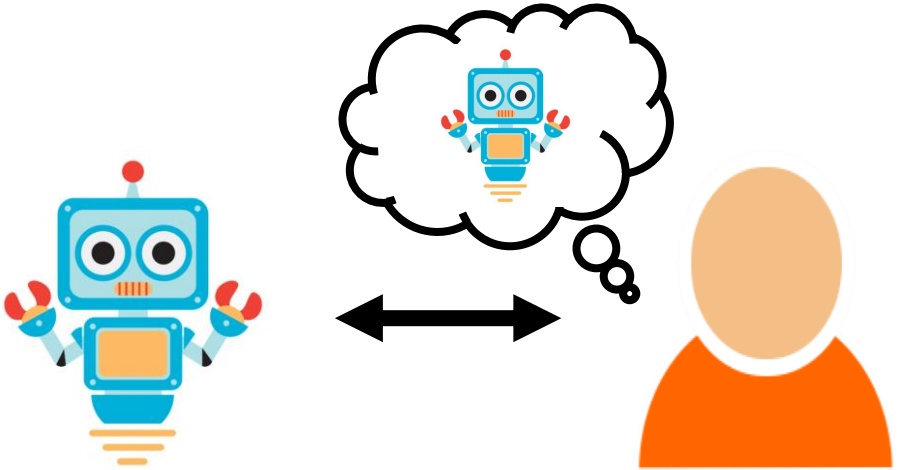
As AI gets better…



As AI gets better…

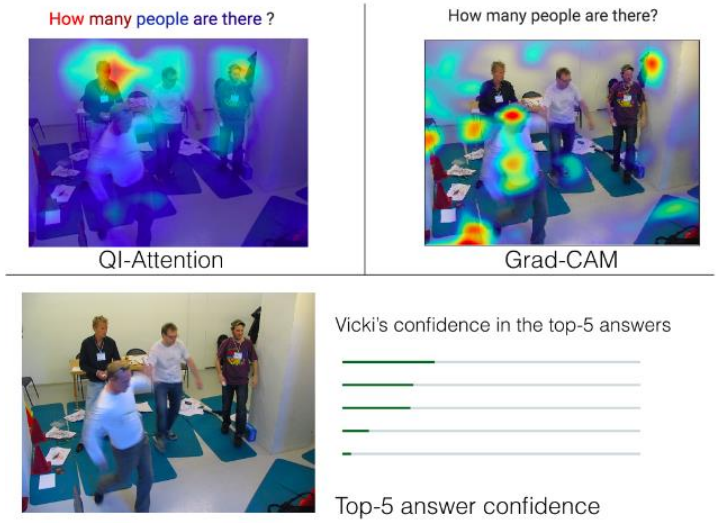


It takes two to tango…



Towards a Theory of AI's Mind

- Humans predict AI's success, failure, responses
  - Humans approximate a neural network!
- Role of explanation modalities
- Human-AI teams put to goal-driven tasks (games)
- Initial scope: Visual Question Answering



Common Sense

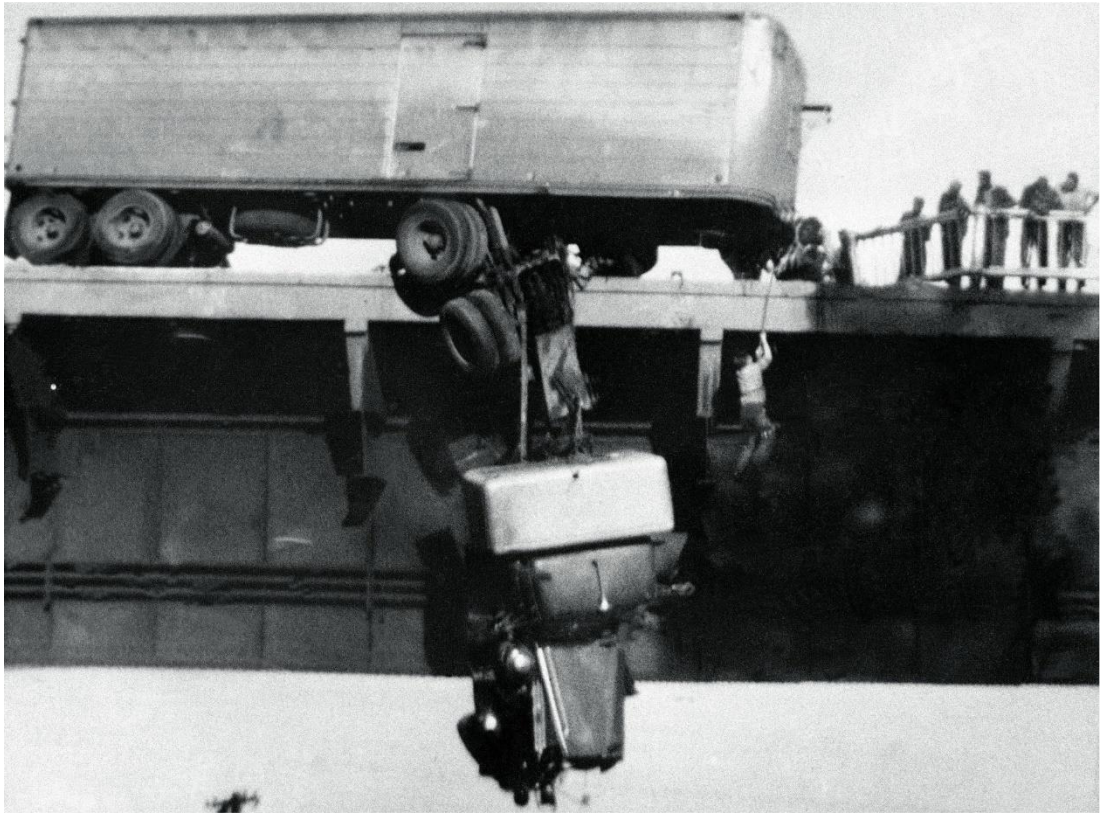
Man in blue wetsuit is surfing on wave
Karpathy and Fei-Fei (Stanford) 2015

A group of young people playing a game of Frisbee
Vinyals et al. (Google) 2015

A car is parked in the middle of nowhere
Kiros et al. (University of Toronto) 2015

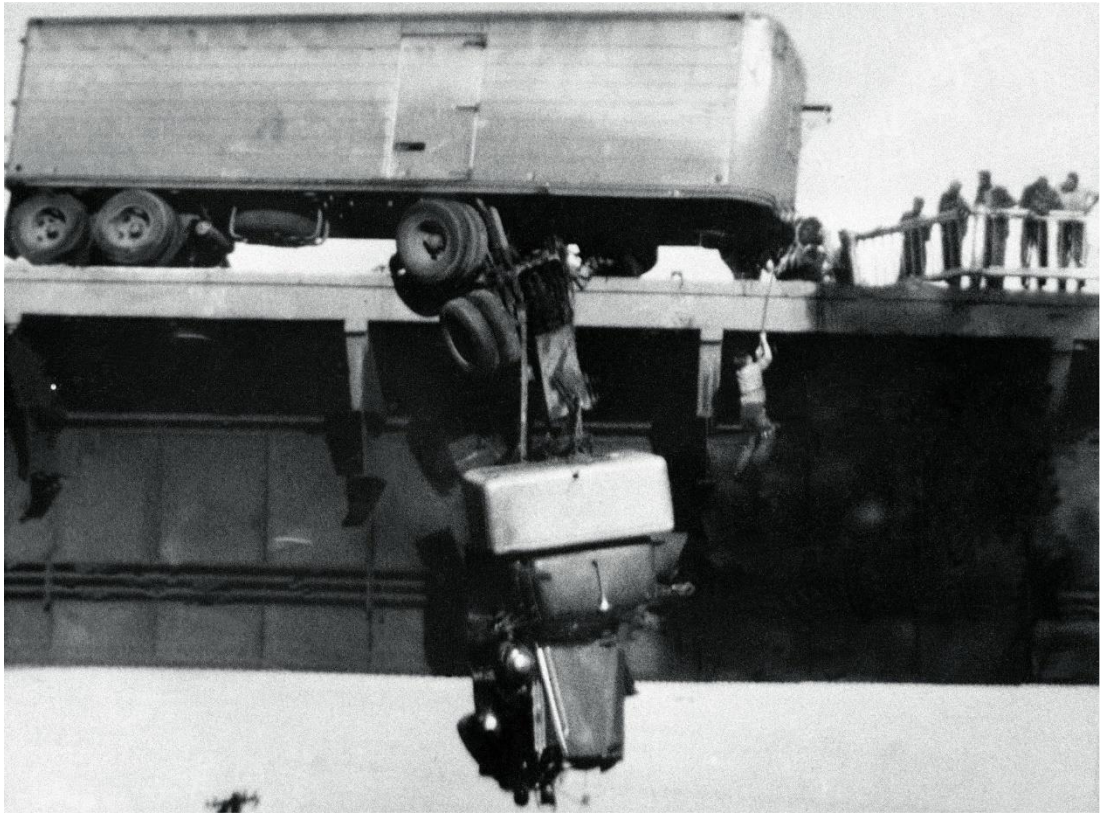
A pot of broccoli on a stove.
Fang et al. (Microsoft Research) 2015

A man is rescued from his truck that is hanging dangerously from a bridge.



Slide credit: Larry Zitnick

A man is *rescued* from his truck that is hanging *dangerously* from a bridge.



Slide credit: Larry Zitnick

Learning Common Sense

- Text
  - Reporting bias

Learning Common Sense

Word	*Teraword*	*Knext*
spoke	11,577,917	244,458
laughed	3,904,519	169,347
murdered	2,843,529	11,284
inhaled	984,613	4,412
breathed	725,034	34,912

Word	*Teraword*	*Knext*
hugged	610,040	10,378
blinked	390,692	20,624
was late	368,922	31,168
exhaled	168,985	3,490
was punctual	5,045	511

[Gordon et al. 2013]

Learning Common Sense

Word	Teraword	Knext	Word	Teraword	Knext
spoke	11,577,917	244,458	hugged	610,040	10,378
laughed	3,904,519	[illegible]	[illegible]	390,692	20,624
murdered	2,843,529	11,284	was late	368,922	31,168
inhaled	**984,613**	4,412	**exhaled**	**168,985**	3,490
breathed	725,034	34,912	was punctual	5,045	511

inhale:exhale = 6:1

[Gordon et al. 2013]

Learning Common Sense

Word	Teraword	Knext	Word	Teraword	Knext
spoke	11,577,917	244,458	hugged	610,040	10,378
laughed	3,904,519	169,347	blinked	390,692	20,624
murdered	2,843,529	11,284	was late	368,922	31,168
inhaled	984,613	4,412	exhaled	168,985	3,490
breathed	725,034	34,912	was punctual	5,045	511

murder:exhale = 17:1

[Gordon et al. 2013]

Learning Common Sense

Body Part	*Teraword*	*Knext*	*Body Part*	*Teraword*	*Knext*
Head	18,907,427	1,332,154	Liver	246,937	10,474
Eye(s)	18,455,030	1,090,640	Kidney(s)	183,973	5,014
Arm(s)	6,345,039	458,018	Spleen	47,216	1,414
Ear(s)	3,543,711	230,367	Pancreas	24,230	1,140
Brain	3,277,326	260,863	Gallbladder	17,419	1,556

[Gordon et al. 2013]

Learning Common Sense

Body Part	Teraword	Knext	Body Part	Teraword	Knext
Head	18,907,427	1,332,154	Liver	246,937	10,474
Eye(s)	18,455,030	1,090,640	Kidney(s)	183,973	5,014
Ear(s)	3,543,711	230,367	Pancreas	24,230	1,140
Brain	3,277,326	260,863	Gallbladder	17,419	1,556

People have heads:gallbladders = 1085:1

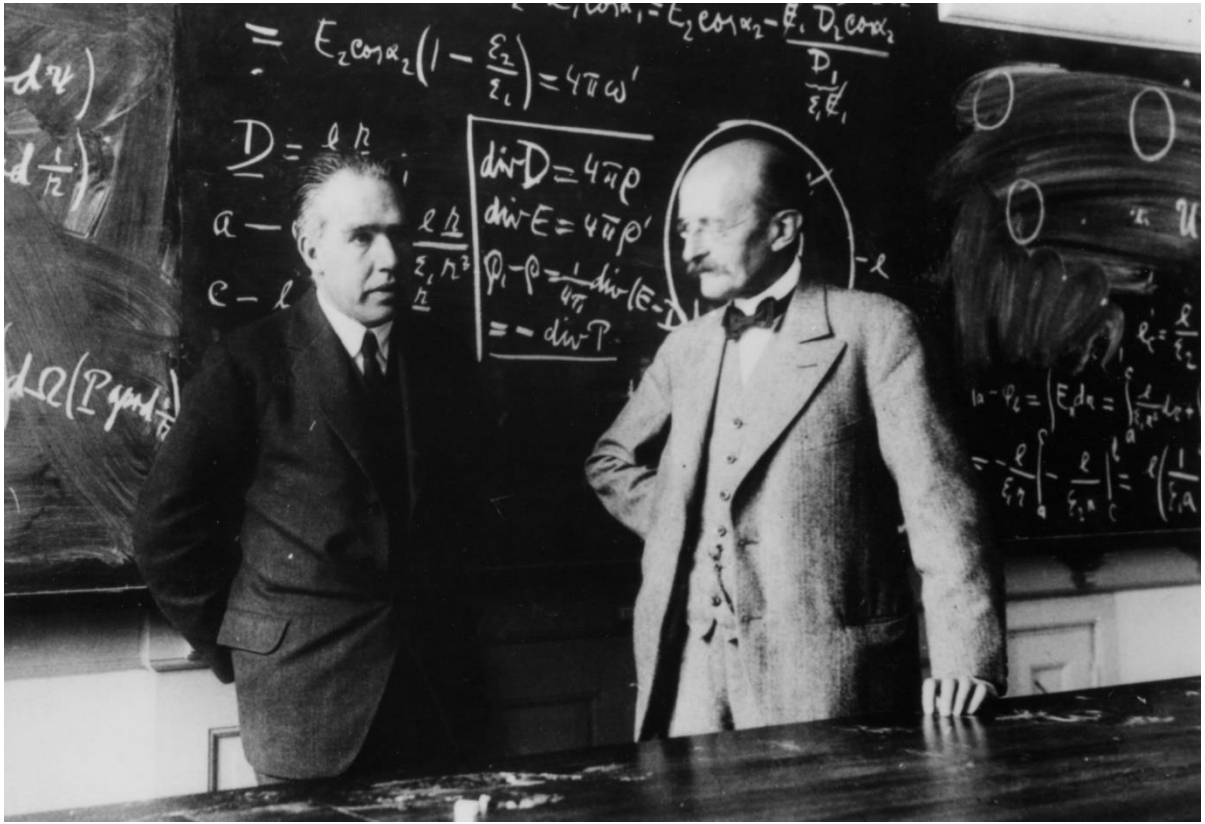
[Gordon et al. 2013]

Learning Common Sense

- Text
  - Reporting bias

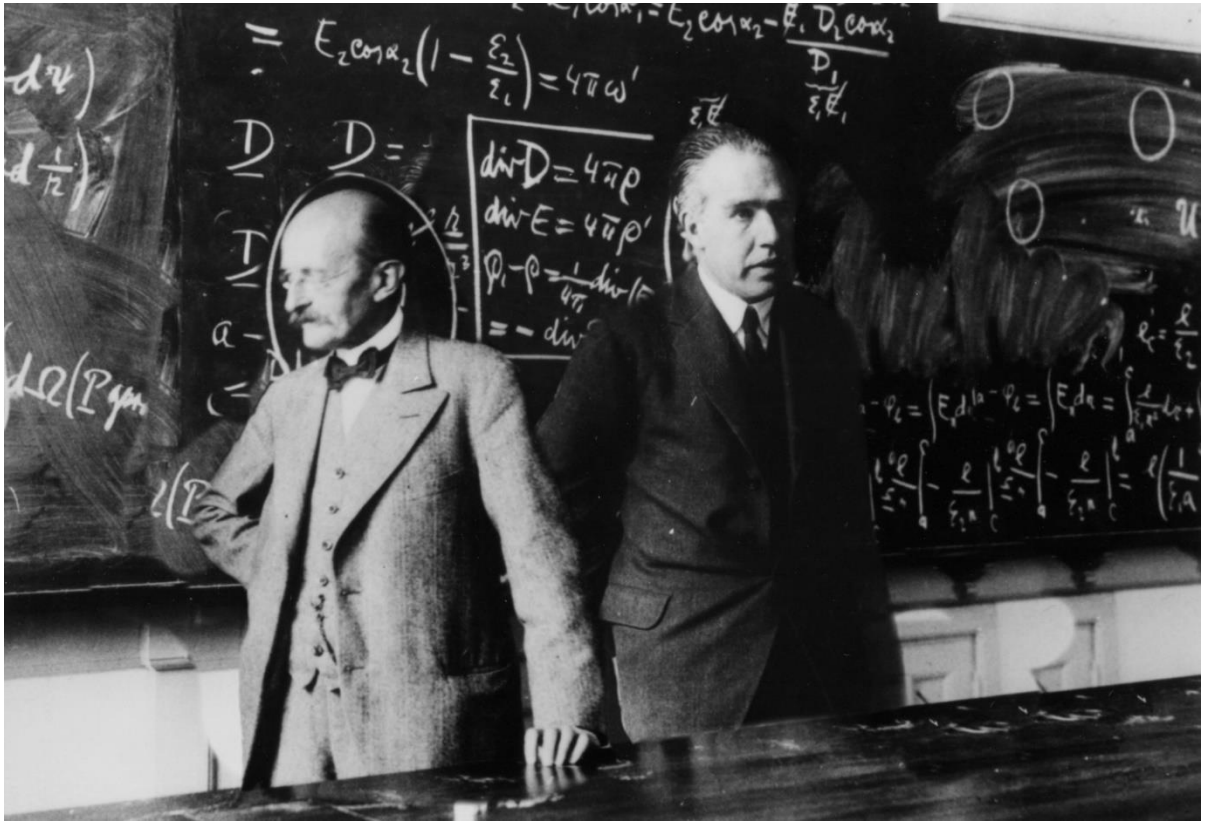
- From structure in our visual world?

Two professors converse in front of a blackboard.



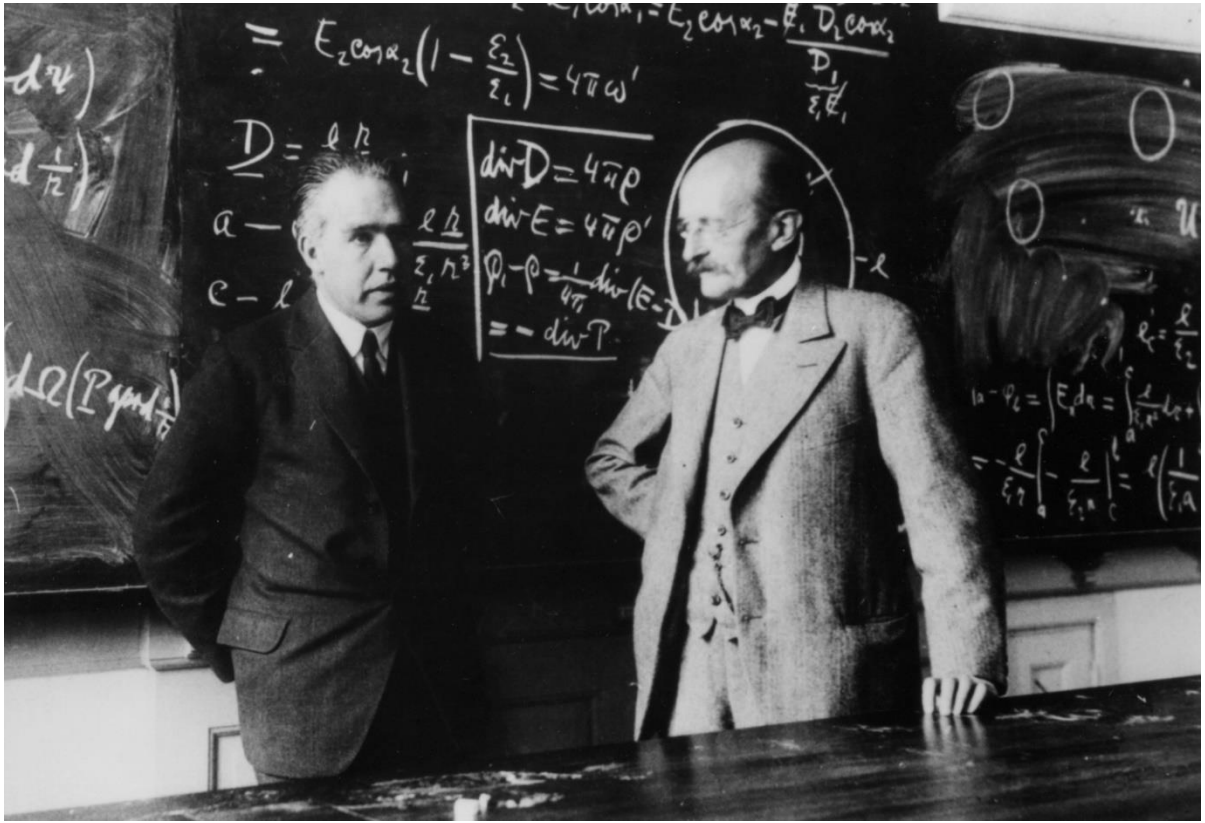
Slide credit: Larry Zitnick

Two professors stand in front of a blackboard.



Slide credit: Larry Zitnick

Two professors converse in front of a blackboard.



Slide credit: Larry Zitnick

Learning Common Sense

- Text
  - Reporting bias

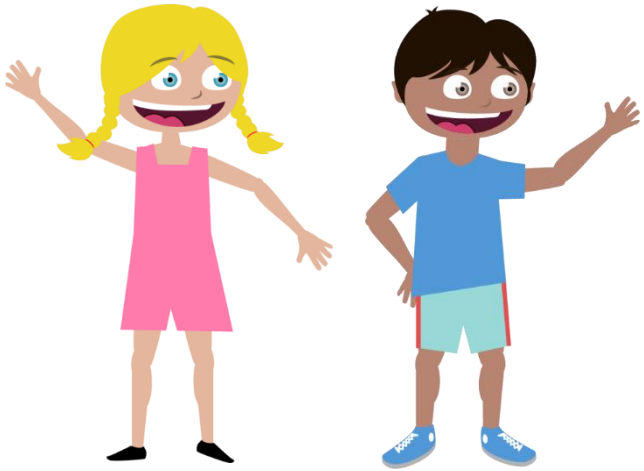
- From structure in our visual world?
  - Lacking visual density

Learning Common Sense

Is photorealism necessary?

Slide credit: Larry Zitnick

Abstract Scenes



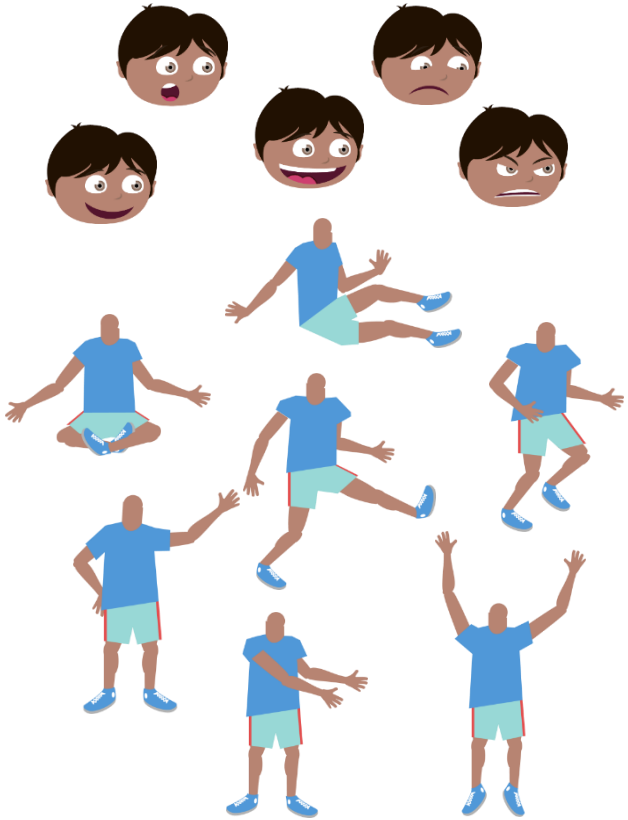
Jenny Mike

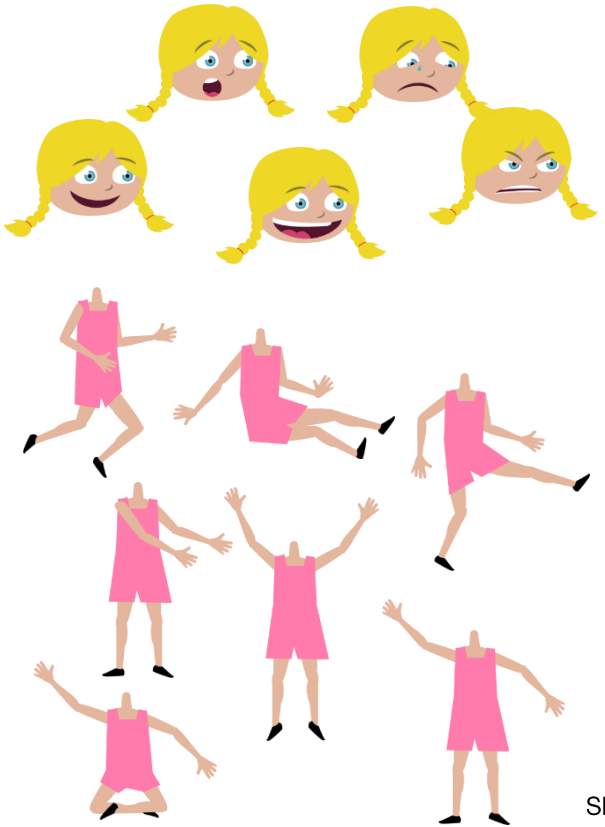
Slide credit: Larry Zitnick

Abstract Scenes

Slide credit: Larry Zitnick

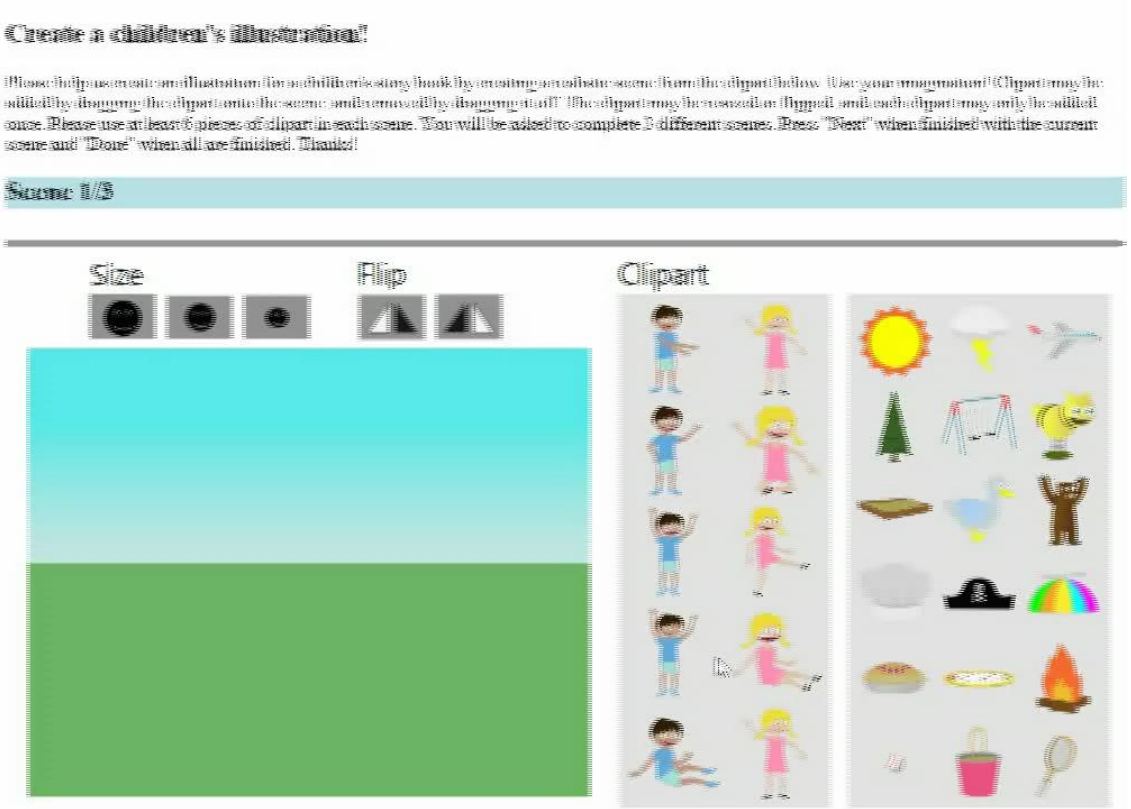
Abstract Scenes





Slide credit: Larry Zitnick

Abstract Scenes

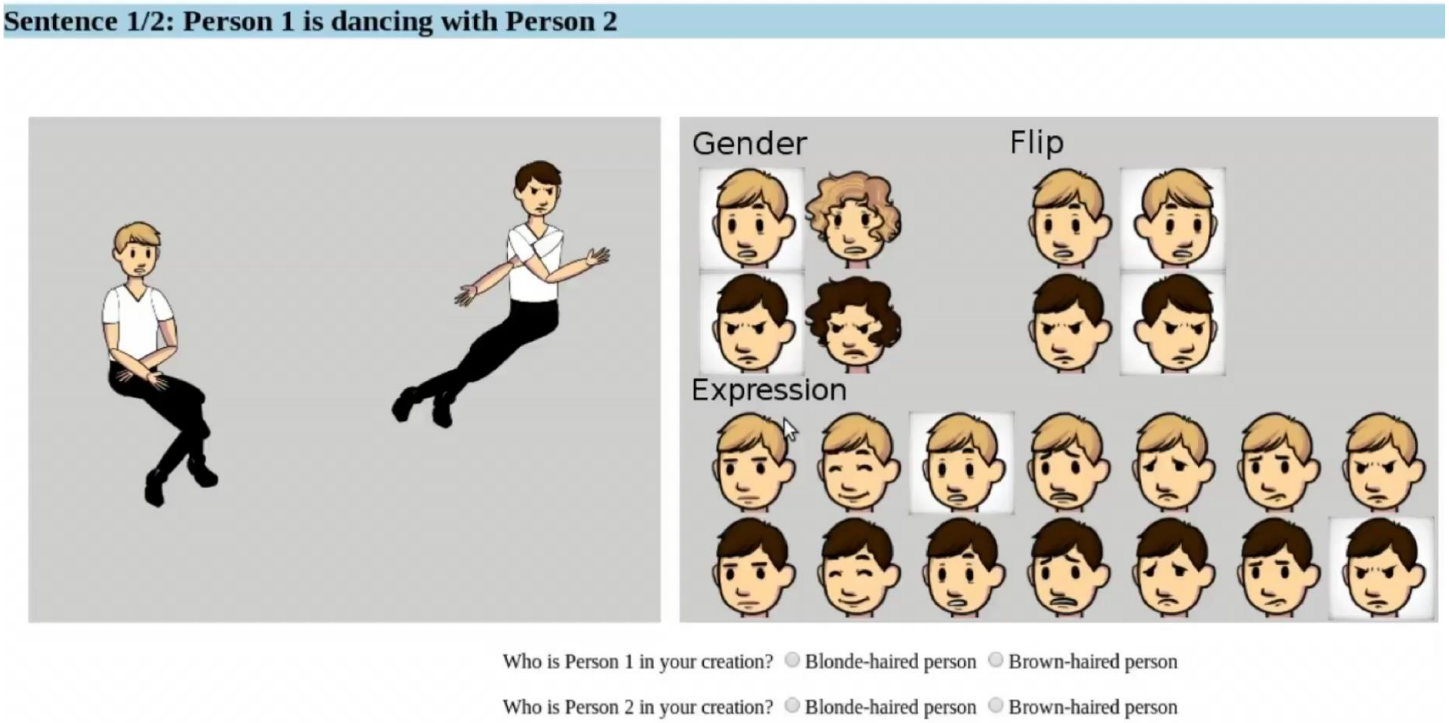


2x

Mike fights off a bear by giving him a hotdog while Jenny runs away.

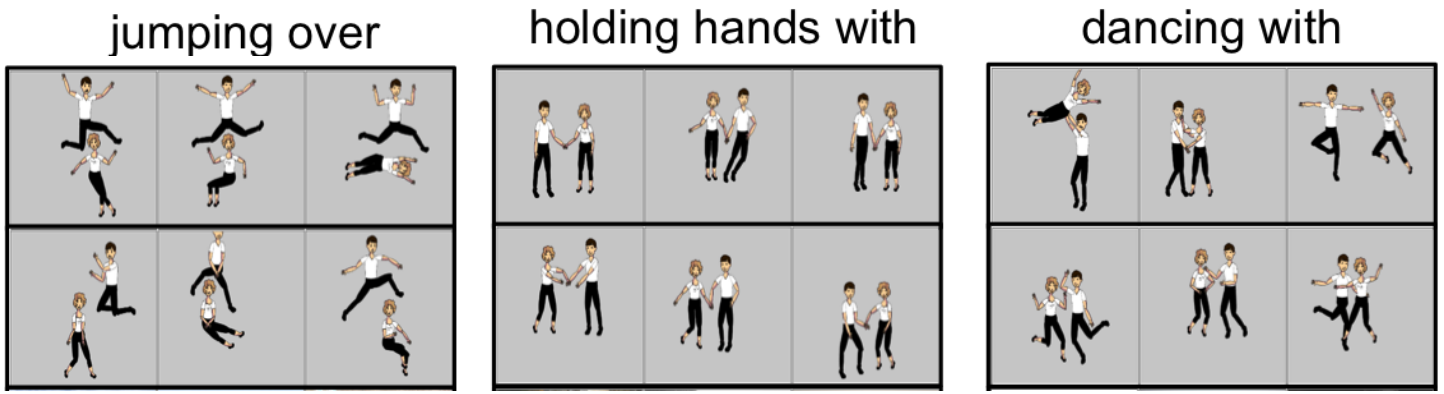
Slide credit: Larry Zitnick

Learning Fine-grained Interactions



3x

Learning Fine-grained Interactions



Commonsense Tasks

- Text-based tasks

Key idea

- Imagine the scene behind the text
- Reason about the visual interpretation of the text, not just the text alone

Commonsense Tasks

Visual Paraphrasing:
Are these two descriptions describing the same scene?

1. Jenny was going to throw her pie at Mike.

2. Jenny is very angry. Jenny is holding a pie.

Approach: Imagination

_____.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.

B. The brown dog is standing next to Mike.

C. The sun is in the sky.

D. Jenny is standing dangerously on the swing.

Slide credit: Xiao Lin

Approach: Imagination

There is a tree near a table.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A

A. There is a tree near a table.

B. The brown dog is standing next to Mike.

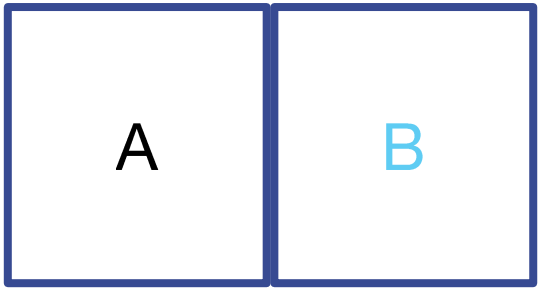
C. The sun is in the sky.

D. Jenny is standing dangerously on the swing.

Slide credit: Xiao Lin

Approach: Imagination

The brown dog is standing next to Mike.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.



A. There is a tree near a table.

B. The brown dog is standing next to Mike.

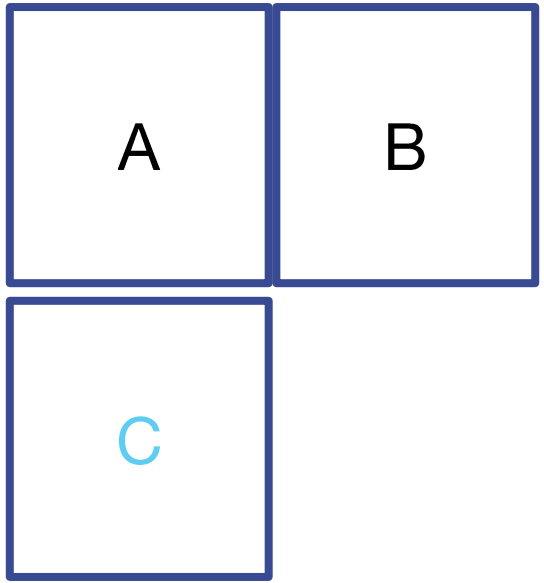
C. The sun is in the sky.

D. Jenny is standing dangerously on the swing.

Slide credit: Xiao Lin

Approach: Imagination

The sun is in the sky.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.



A. There is a tree near a table.

B. The brown dog is standing next to Mike.

C. The sun is in the sky.

D. Jenny is standing dangerously on the swing.

Slide credit: Xiao Lin

Approach: Imagination

Jenny is standing dangerously on the swing.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.

B. The brown dog is standing next to Mike.

C. The sun is in the sky.

D. Jenny is standing dangerously on the swing.

Imagined scenes need not be photorealistic but rich in semantics

Slide credit: Xiao Lin

Approach: Imagination

- Clipart Visual World

  [CVPR 2013]

  - Two children playing in the park
  - 58 objects

  - 7 poses and 5 expressions

Slide credit: Xiao Lin

Approach: Imagination

- Scene generation given description [ICCV 2013]

There is a tree near a table.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

Slide credit: Xiao Lin

Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples

<Tree, near table>
<Mike, wear, cap>
<Mike, tell, get> <Jenny, get off, swing>

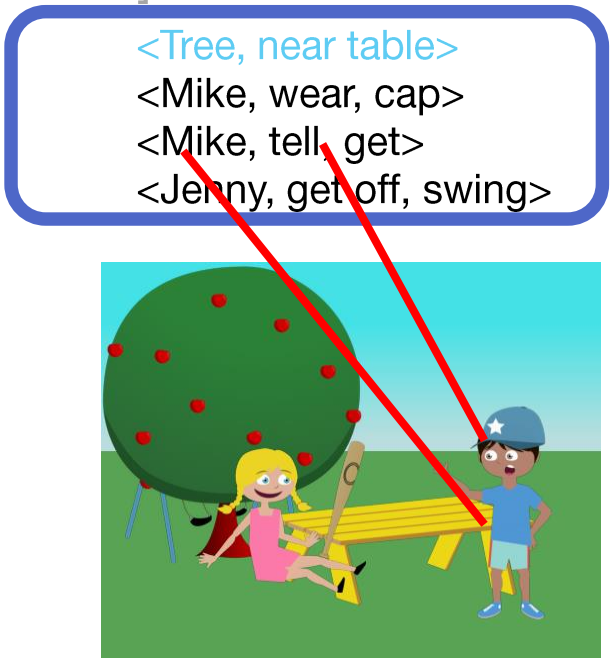
Slide credit: Xiao Lin

Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples
- Scene generation CRF

  Conditional Random Field (CRF)

$$p(\text{objects}|\text{tuples})$$



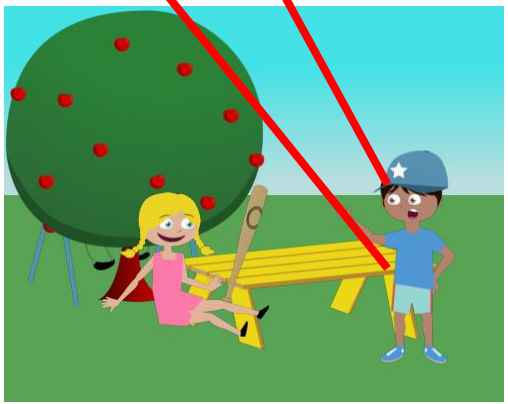
Slide credit: Xiao Lin

Approach: Imagination

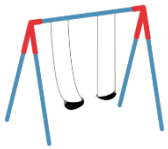
- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples
- Scene generation CRF

  Which objects are present

<Tree, near table>
<Mike, wear, cap>
<Mike, tell, get>
<Jenny, get off, swing>



mike
he
ground
shirt



swing
set
swingset
playground

tree
apple
fruit
grab

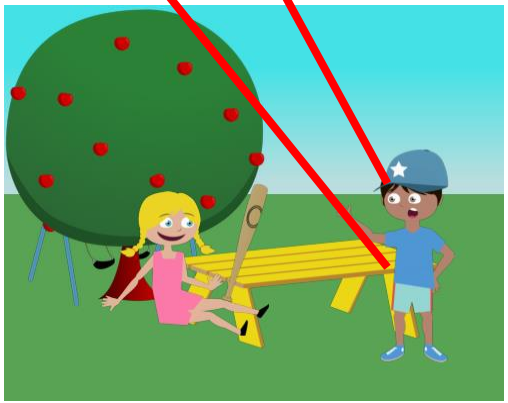
Slide credit: Xiao Lin

Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples
- Scene generation CRF

  Where objects are

<Tree, near table>
<Mike, wear, cap>
<Mike, tell, get>
<Jenny, get off, swing>

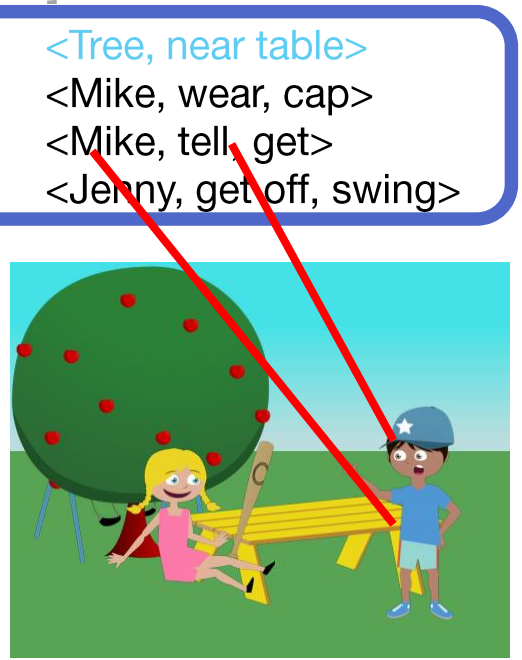


Slide credit: Xiao Lin

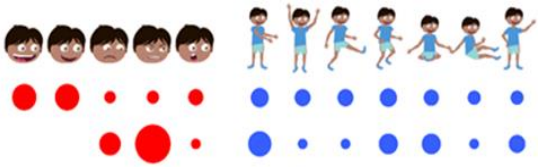
Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples
- Scene generation CRF

  What are the poses and expressions



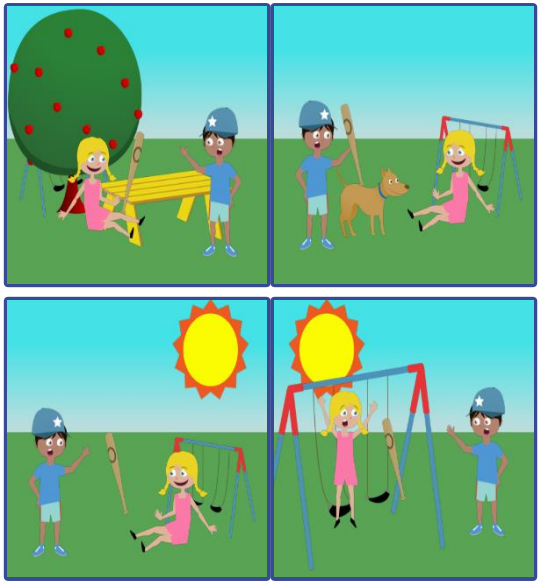
Play with



Slide credit: Xiao Lin

Approach: Imagination

_____.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.



A. There is a tree near a table.

B. The brown dog is standing next to Mike.

C. The sun is in the sky.

D. Jenny is standing dangerously on the swing.

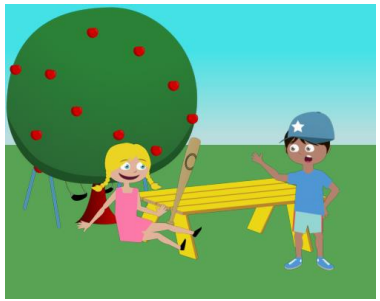
Slide credit: Xiao Lin

Approach: Joint Text + Visual Reasoning

Jenny is standing dangerously on the swing.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

$\geq$

There is a tree near a table.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.



$$w^T \phi_i^{\text{gt}} \geq w^T \phi_i^j + 1$$

Ranking Support Vector Machine (Ranking SVM)

Slide credit: Xiao Lin

Results: Fill-in-the-blanks

- Fill-in-the-blank dataset
  - 7,198 train
  - 1,761 test

Method	Accuracy
Random	25%
Text Baseline	44.97%
Human	52.87%

Slide credit: Xiao Lin

Results: Fill-in-the-blanks

- Fill-in-the-blank dataset
  - 7,198 train
  - 1,761 test

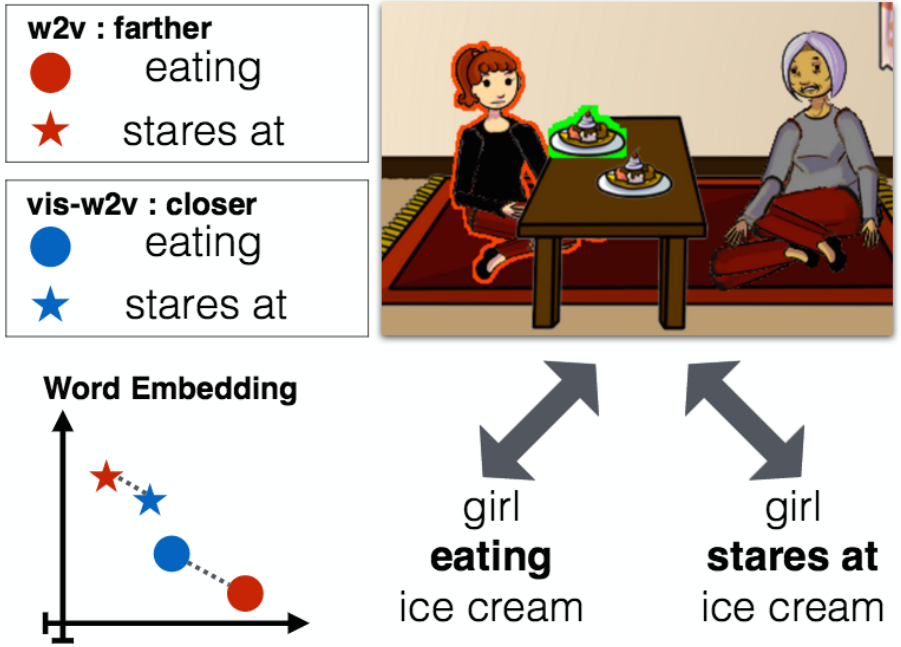
Method	Accuracy
Random	25%
Text Baseline	44.97%
Text + Visual	48.04%
Human	52.87%

+3.07%

Slide credit: Xiao Lin

Visual Word2Vec

- Learn word embeddings that respect visual (as well as textual) similarity



Visual Abstraction For…

- Studying mappings between images and text [CVPR 2013, ICCV 2013]
- Zero-shot learning [ECCV 2014]
- Studying
  - Image memorability [PAMI 2016]
  - Image specificity [CVPR 2015]
  - Visual humor [CVPR 2016]

“This terrified woman's home is being invaded by mice as the cat sleeps.”

“The man is about to trip on his child's car and spill wine on his wife.”

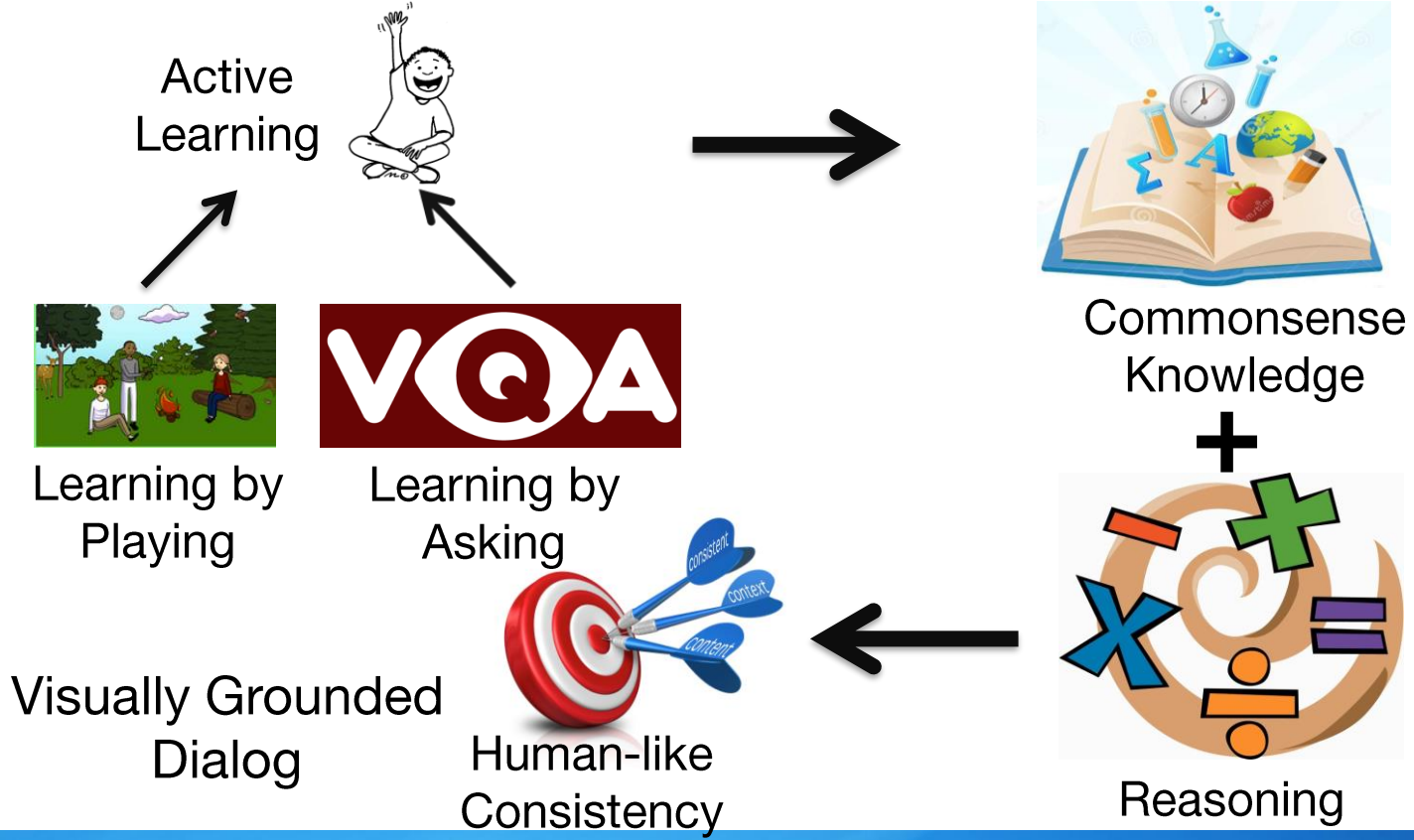
Visual Abstraction For…

- Studying mappings between ima
- Zero-shot learning [ECCV
- Studying
  - Image memorability [PA
  - Image specificity [CVPR
  - Visual humor [CVPR 2016]
- Learning common sense knowledge [CVPR 2015, ICCV 2015, CVPR 2016]
- Rich annotation modality
  - Ask for descriptions
  - Ask for scenes
  - Show scene and ask f
  - Perturb a scene and a descriptions
  - …

Study high-level image understanding tasks without waiting for lower-level vision tasks to be solved

Future work:
Learning by "playing"

Open Directions



Summary

- Natural progression in vision + language
  - Captioning → VQA → Visual Dialog
- Vision, language, action
- Common sense
- Interpretable AI
- Connecting humans and AI

My Lab

Devi Parikh
Assistant Professor

Ramakrishna Vedantam
Ph.D. Student

Arjun Chandrasekaran
Ph.D. Student

Jiasen Lu
Ph.D. Student

Jianwei Yang
Ph.D. Student

Ramprasaath Selvaraju
Ph.D. Student

Samyak Datta
Ph.D. Student

Prithvijit Chattopadhyay
M.S. Student

Viraj Prabhu
M.S. Student

Dhruv Batra's Lab

Qing Sun
(2012 — Present)

Aishwarya Agrawal
(2014 — Present)

Yash Goyal
(2014 — Present)

Michael Cogswell
(2015 — Present)

Abhishek Das
(2016 — Present)

Ashwin Kalyan
(2016 — Present)

Nirbhay Modhe
(2017 — Present)

Akrit Mohapatra

Deshraj Yadav

Dhruv Batra

Research Scientist II

Stefan Lee

Collaborators and Alumni

Mainak Jas
Telecom ParisTech

Tanmay Batra
CMU

Satwik Kottur
CMU

Stanislaw Antol
Think Tank

Xiao Lin
SRI

Khushi Gupta
CMU

Avi Singh
UC Berkeley

Georgia Gkioxari
Facebook AI Research

Meg Mitchell
Google Research

Larry Zitnick
Facebook AI Research

Lucy Vanderwende
Microsoft Research

Mohit Bansal
UNC Chapel Hill

Thank you.