Microsoft

Vision, Language and Commonsense

Microsoft Research
Faculty Summit
2015

Larry Zitnick
Microsoft Research

What does it mean to “understand”?

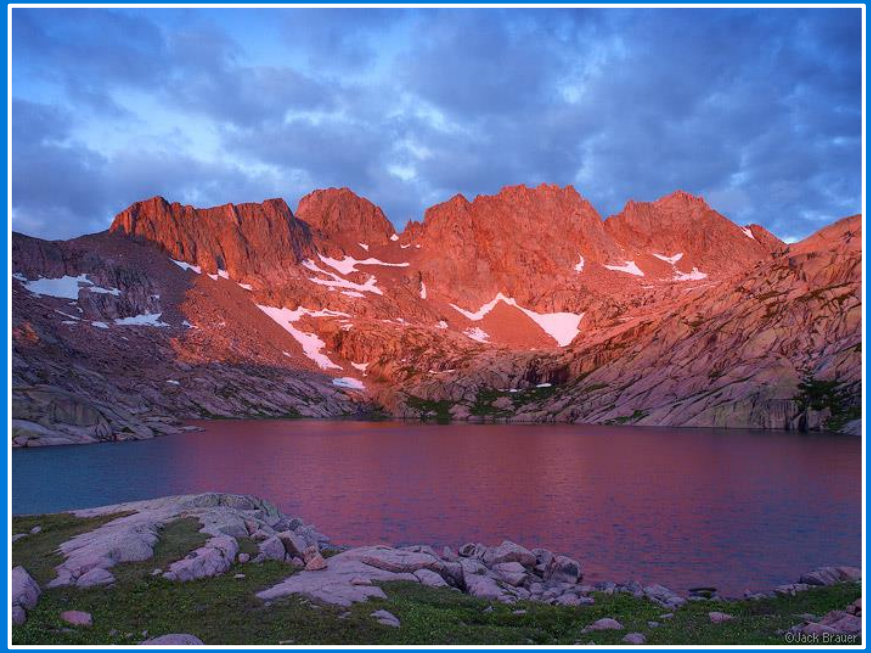
Red

Is a sheep fluffy?

A tree?

Case study:

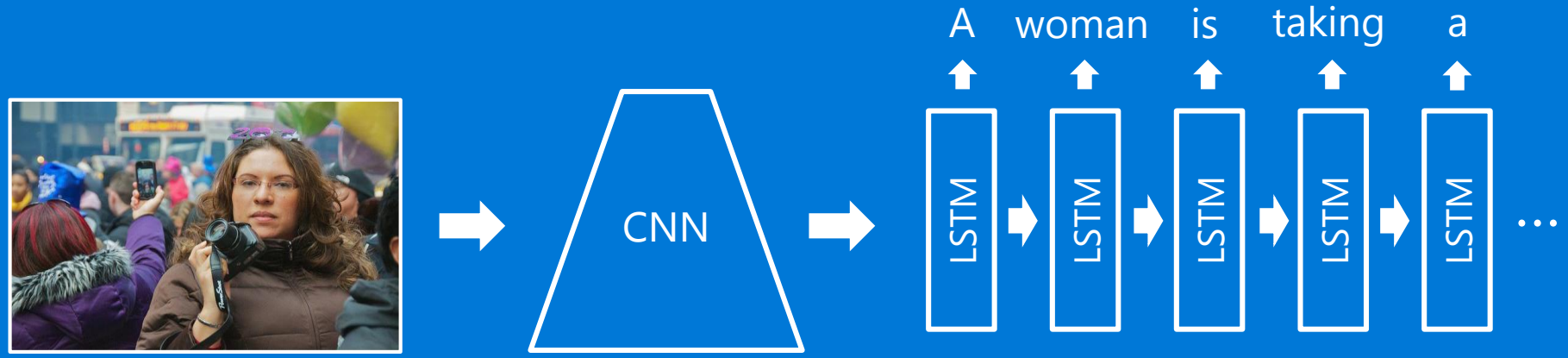
Image captioning



"Long, blue, spiky-edged shadows crept out across the snow-fields, while a rosy glow, at first scarce discernible, gradually deepened and suffused every mountain-top, flushing the glaciers and the harsh crags above them."
John Muir

Vision ➡ Representation ➡ Language

LSTM



Show and Tell: A Neural Image Caption Generator,
Vinyals et al., CVPR 2015

A man standing next to a fire hydrant in front of a brick building.

From Captions to Visual Concepts and Back,
Fang et al., CVPR 2015

35% - 85% of captions are identical to training captions.

Nearest Neighbor

Test

Train

Nearest Neighbor

A black and white cat sitting in a bathroom sink.

Two zebras and a giraffe in a field.

See mscoco.org for image information

Results

MS COCO Caption Challenge

	CIDEr-D	Meteor	ROUGE-L	BLEU-4
Google[4]	0.943	0.254	0.53	0.309
MSR Captivator[9]	0.931	0.248	0.526	0.308
m-RNN[15]	0.917	0.242	0.521	0.299
MSR[8]	0.912	0.247	0.519	0.291
Nearest Neighbor[11]	0.886	0.237	0.507	0.280
m-RNN (Baidu/ UCLA)[16]	0.886	0.238	0.524	0.302
Berkeley LRCN[2]	0.869	0.242	0.517	0.277
Human[5]	0.854	0.252	0.484	0.217
Montreal/Toronto[10]	0.85	0.243	0.513	0.268
PicSOM[13]	0.833	0.231	0.505	0.281
MLBL[7]	0.74	0.219	0.499	0.26
ACVT[1]	0.709	0.213	0.483	0.246
NeuralTalk[12]	0.674	0.21	0.475	0.224
Tsinghua Bigeye[14]	0.673	0.207	0.49	0.241
MIL[6]	0.666	0.214	0.468	0.216
Brno University[3]	0.517	0.195	0.403	0.134

Why does nearest neighbor work?

Microsoft COCO
Common Objects in Context

80-100 object categories
160k images
2 million segmentations
5 sentences per image

http://mscoco.org




Microsoft COCO
Common Objects in Context

Tsung-Yi Lin
Cornell Tech

Genevieve Patterson
Brown University

Serge Belongie
Cornell Tech

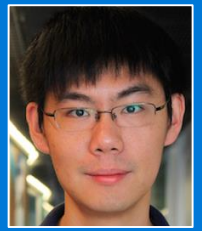
Pietro Perona
Caltech

James Hays
Brown University

Deva Ramanan
CMU

Michael Maire
TTI Chicago

Matteo Ronchi
Caltech



Yin Cui
Cornell

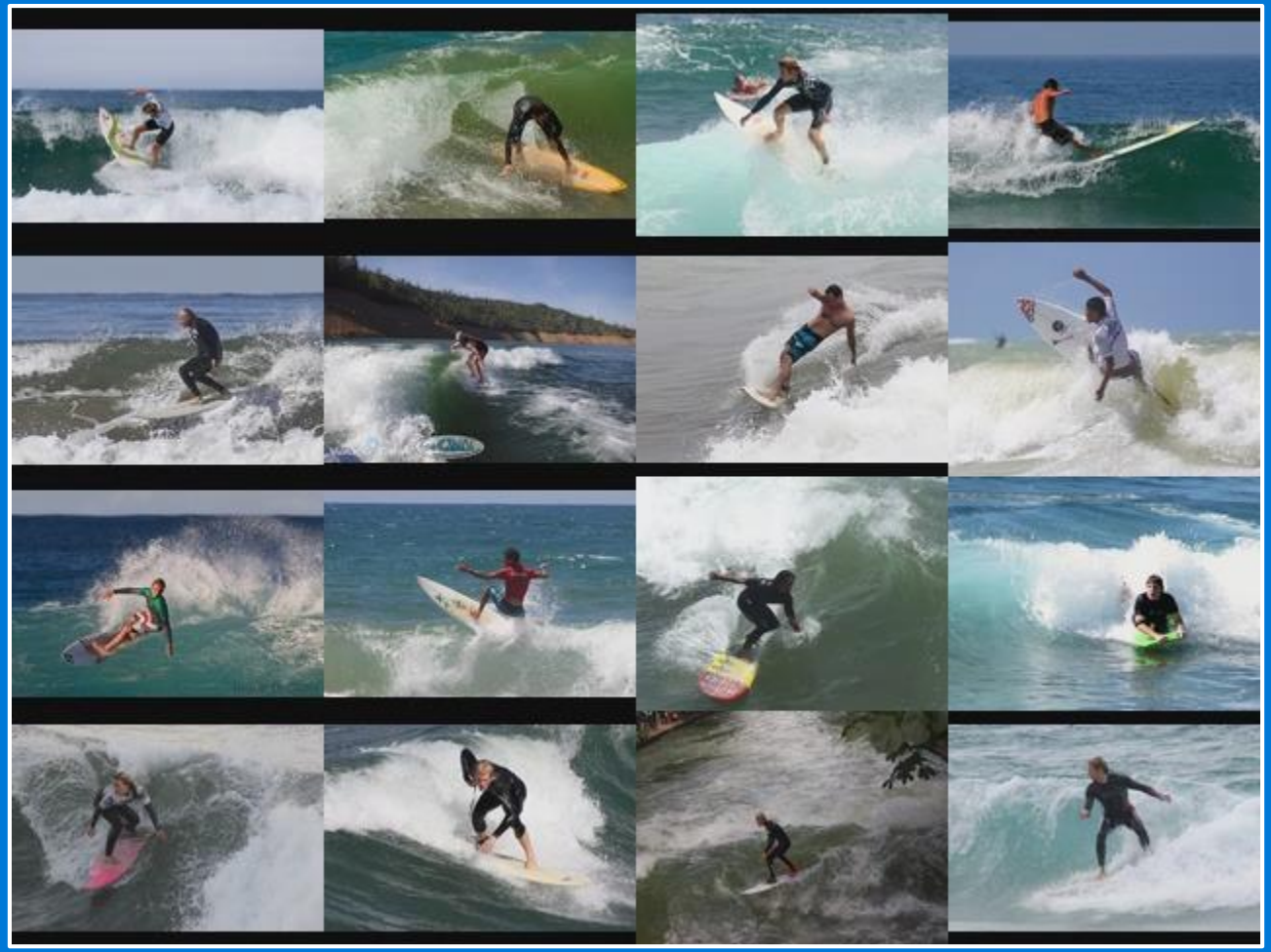
Lubomir Bourdev
Facebook

Piotr Dollar
Facebook

Ross Girshick
Microsoft Research

Larry Zitnick
Microsoft Research

A man riding a wave on a surfboard in the water.





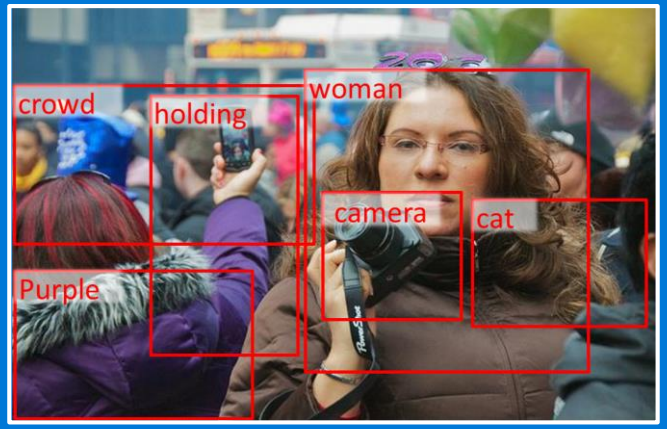





vemödalen - n. the frustration of photographing something amazing when thousands of identical photos already exist

Vemödalen: The Fear That Everything Has Already Been Done
https://www.youtube.com/watch?v=8ftDjebw8aA

Going beyond…



From Captions to Visual Concepts and Back, Fang et al., CVPR 2015.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, Xu et al., ArXiv, 2015.

Mind's Eye

Xinlei Chen

CMU

Mind's Eye: A Recurrent Visual Representation for Image Caption Generation,
Chen and Zitnick, CVPR 2015.

Microsoft

A girl

Mind's Eye: A Recurrent Visual Representation for Image Caption Generation,
Chen and Zitnick, CVPR 2015.

A girl and boy

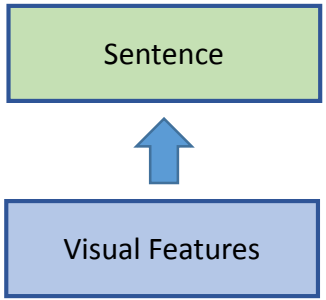
A girl and boy knocked

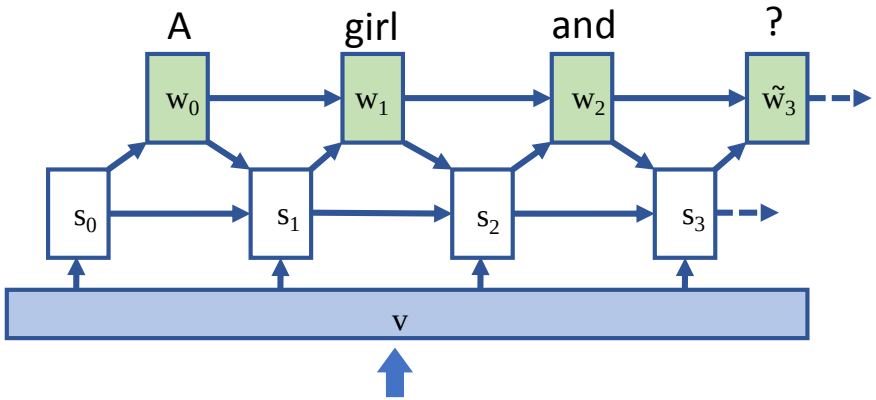
A girl and boy knocked down the tower.

A girl and boy knocked down the tower.

Vision ⟷ Representation ⟷ Language

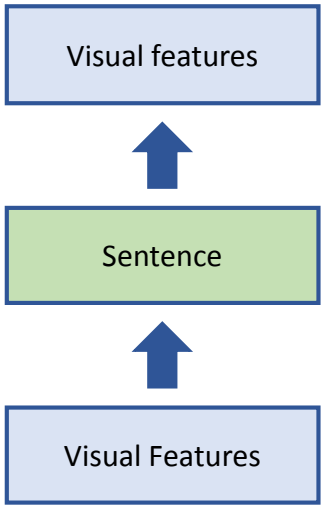
Recurrent Neural Networks



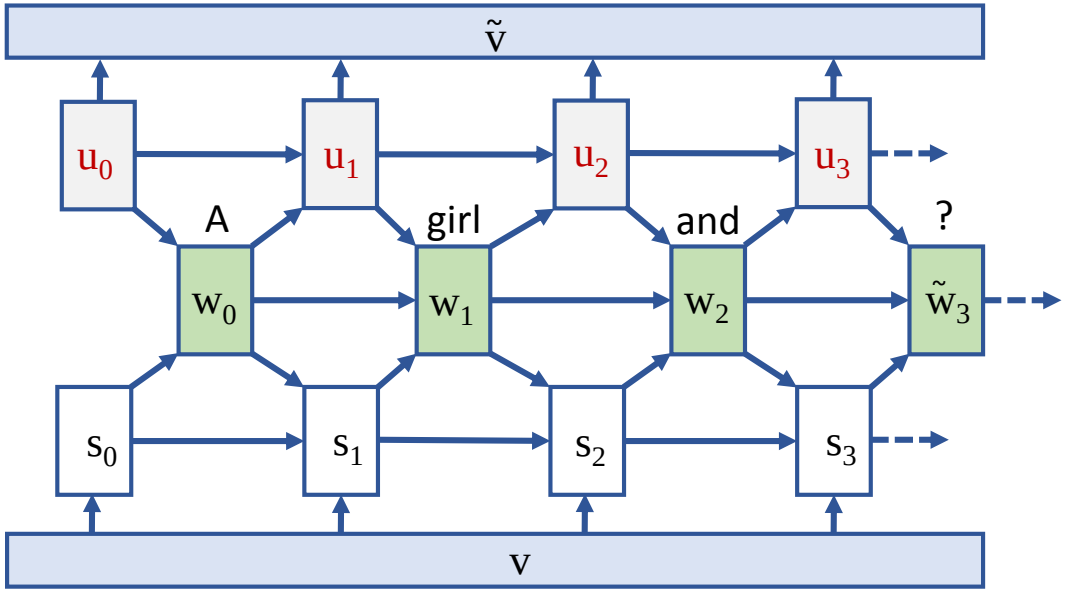


Context dependent recurrent neural network language model. T. Mikolov and G. Zweig, SLT 2012.

Our model



$\mathbf{U}$ = long-term visual memory

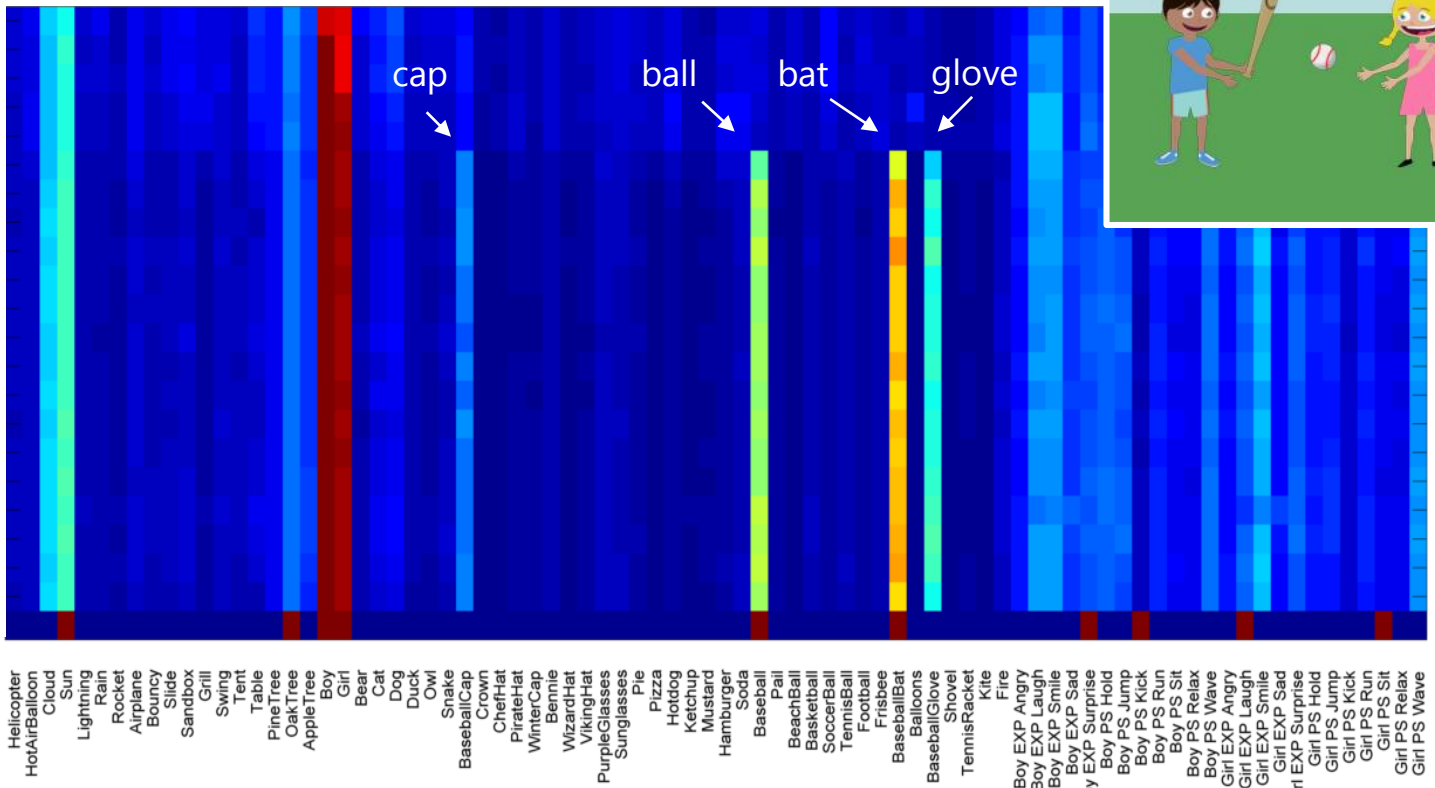


$$\tilde{\mathbf{V}} \approx \mathbf{V}$$

Learning a Recurrent Visual Representation for Image Caption Generation
Chen and Zitnick, CVPR 2015.

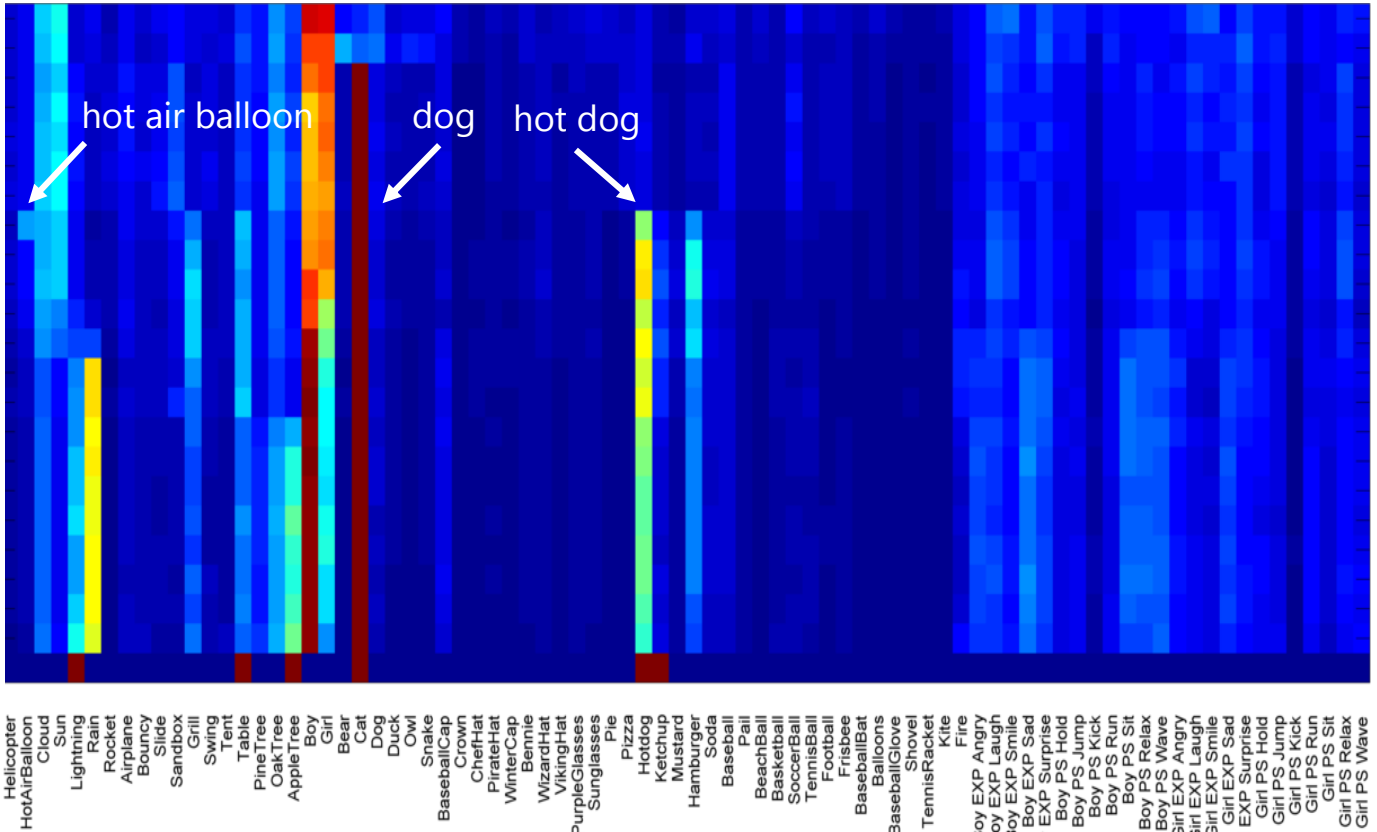
Sample Results

Mike is holding a baseball bat .
Jenny just threw the baseball .
Mike did not hit the the baseball .



Sample Results

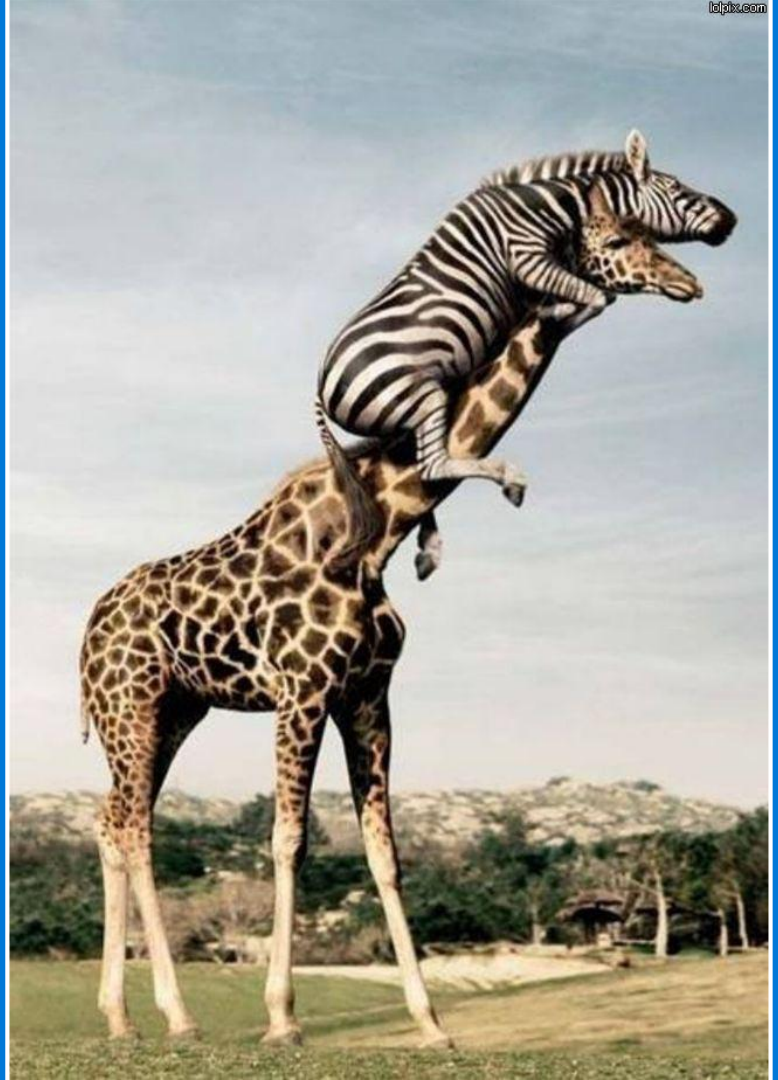
The cat is looking at the hot dog . Lightning is striking the tree . Apples are growing on the tree .



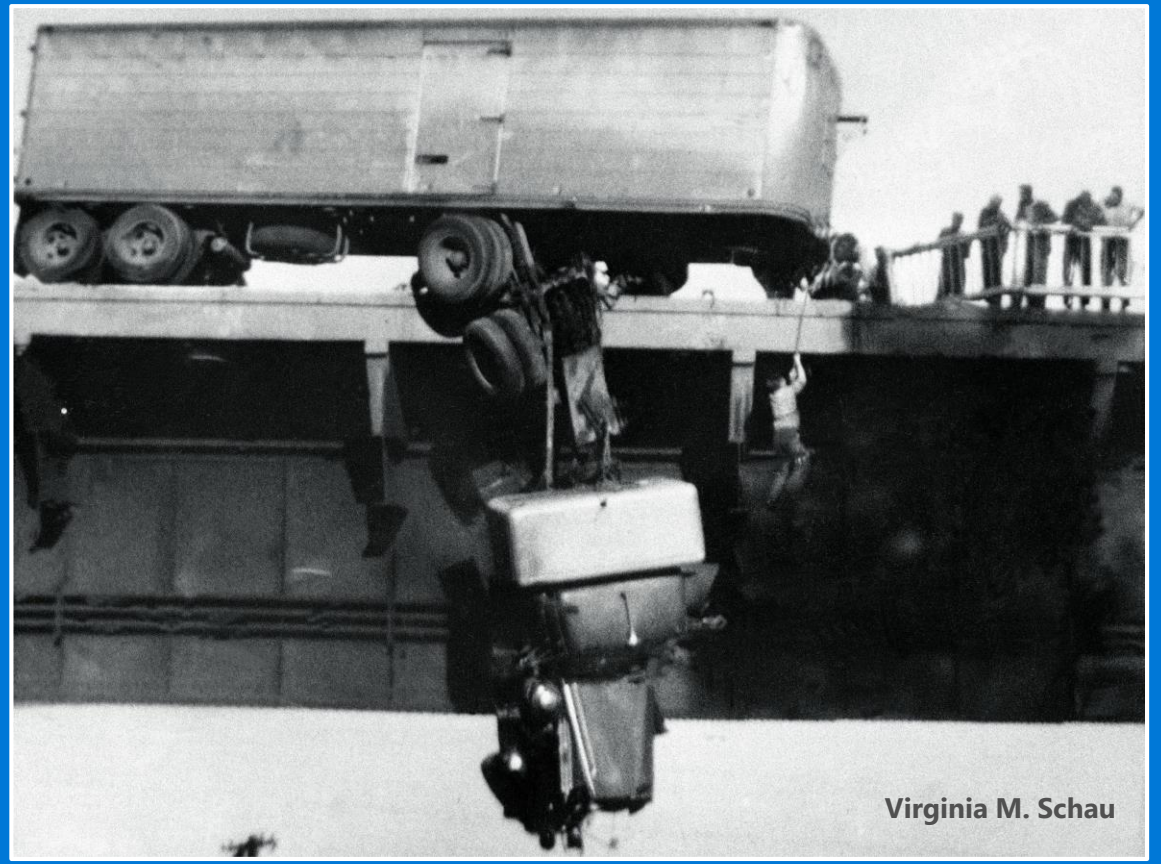
Limitations

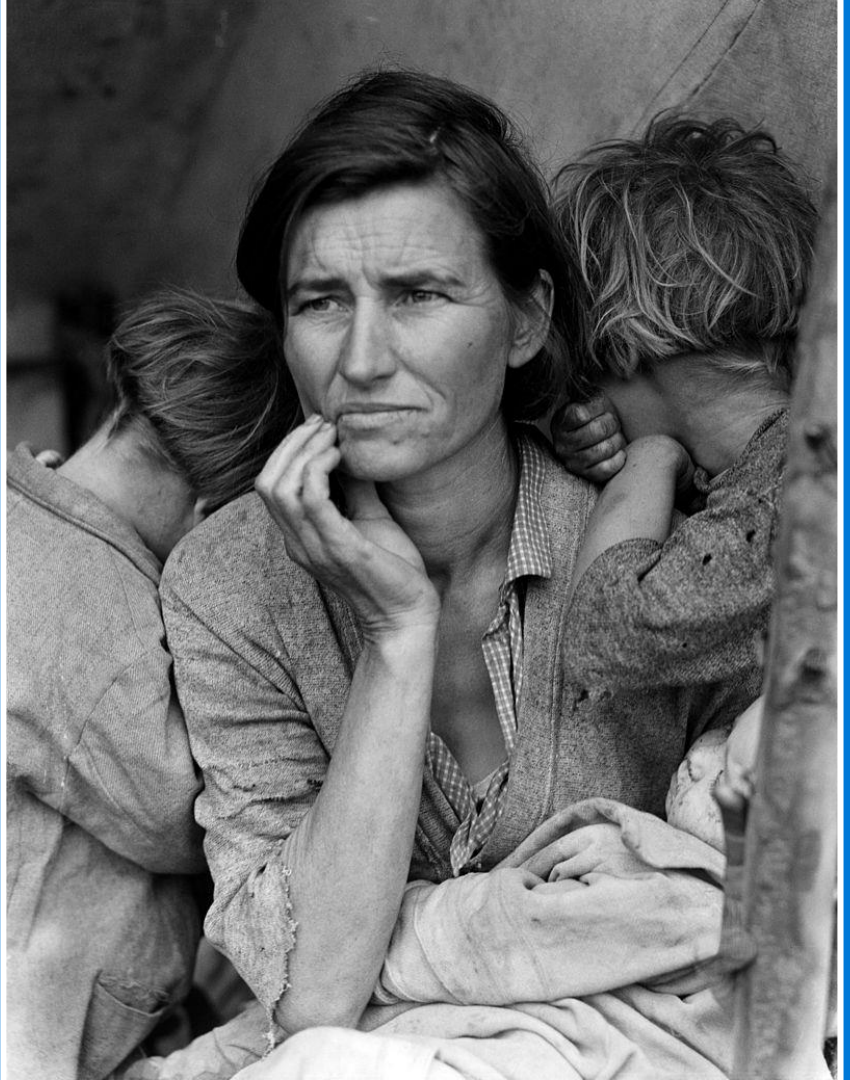
A crazy zebra climbing a giraffe to get a better view.

The limits of vision and language models…



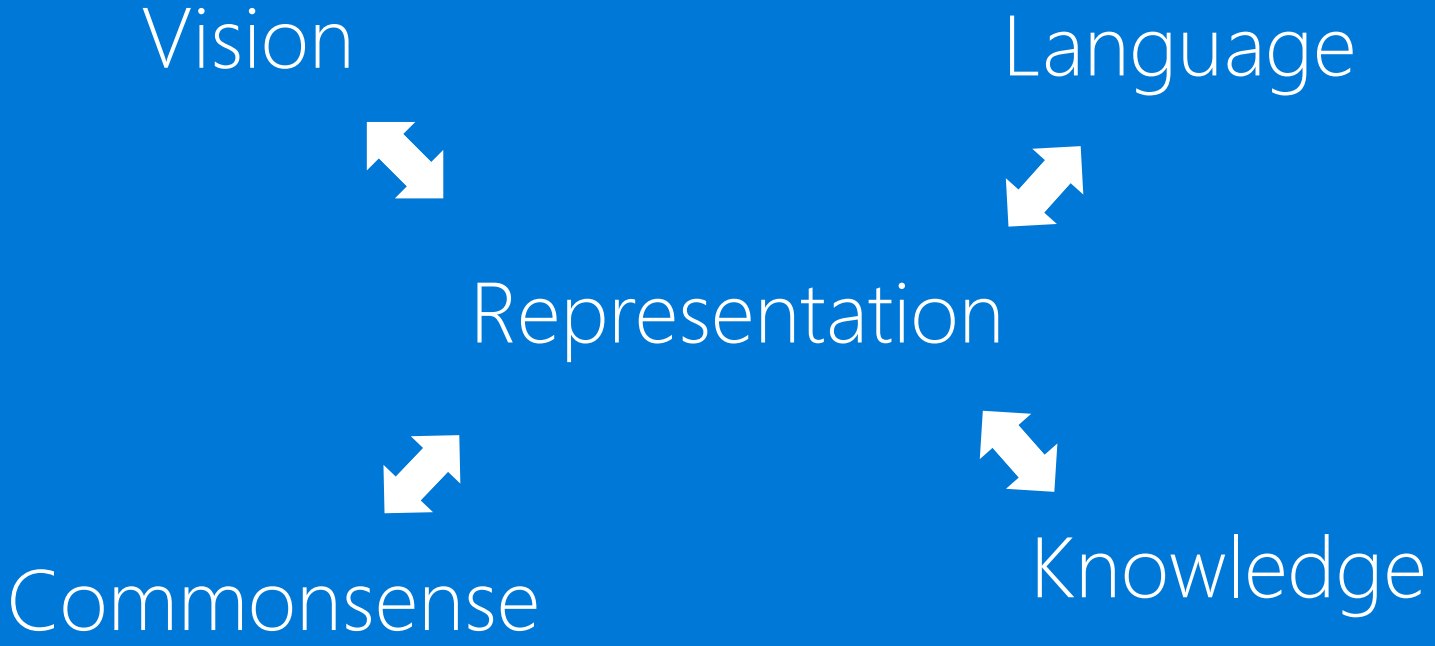
A man is rescued from his truck that is hanging dangerously from a bridge.





Migrant Mother, Dorothea Lange

Vision ➡ Representation ➡ Language


Vision
Language
Representation
Commonsense
Knowledge

Abstract Scenes

Devi Parikh,
VT

Xinlei Chen,
CMU

Rama Vedantam,
VT

David Fouhey,
CMU

Stan Antol,
VT

Xiao Lin,
VT

Lucy Vanderwende,
Microsoft Research

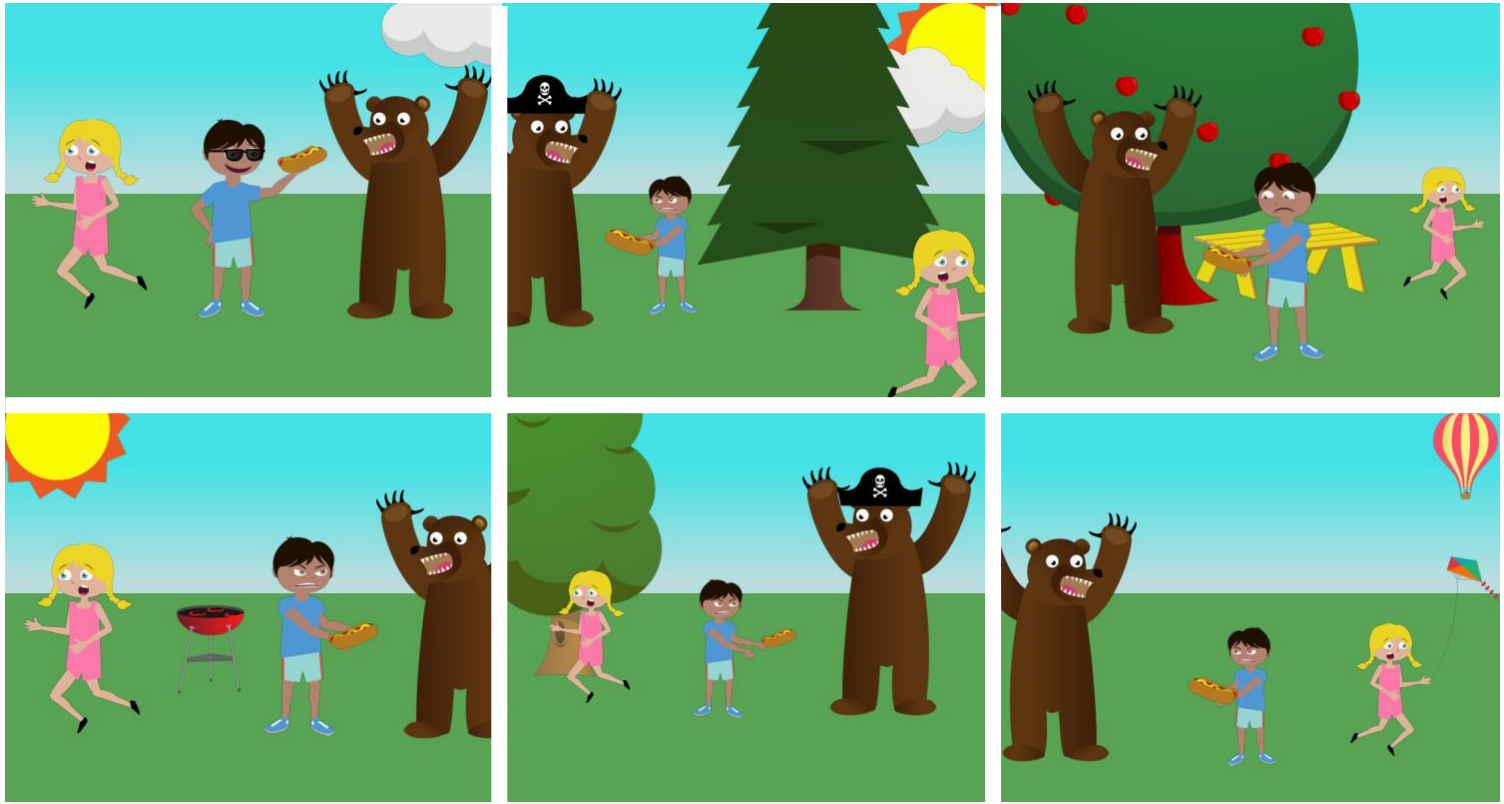
Is photorealism necessary?

Generating data

Jenny just threw the beach ball angrily at Mike while the dog watches them both.

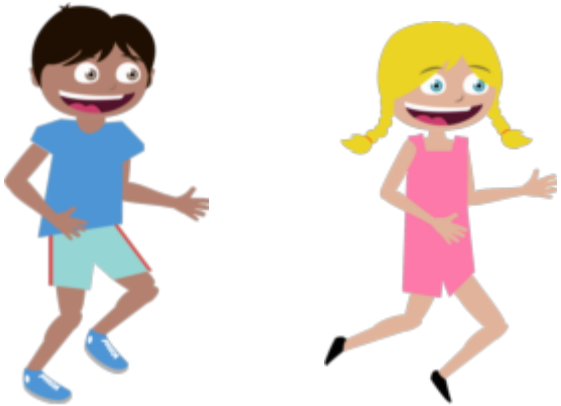
Bringing Semantics Into Focus Using Visual Abstraction, Zitnick and Parikh, CVPR 2013

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

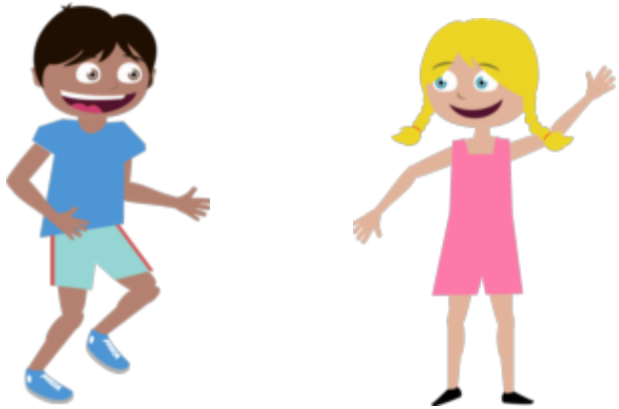


Bringing Semantics Into Focus Using Visual Abstraction, Zitnick and Parikh, CVPR 2013

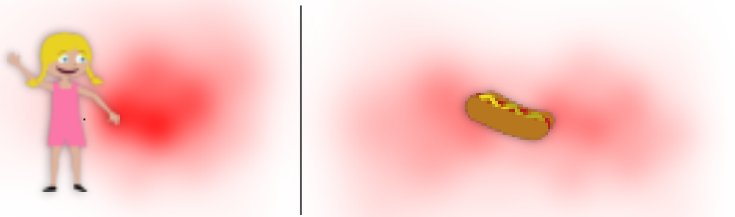
run after



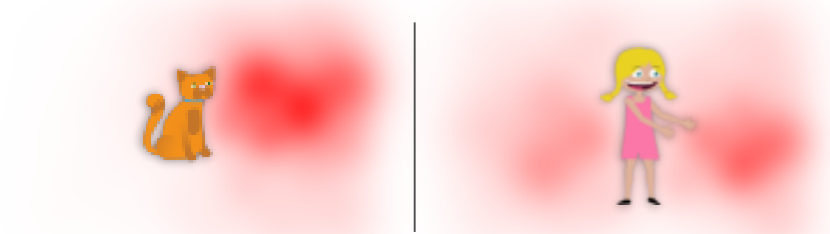
run to



want



watch



Learning the Visual Interpretation of Sentences,
Zitnick, Parikh, and Vanderwende, ICCV 2013.

Visual Question Answering

VQA: Visual Question Answering

Stanislaw Antol
Virginia Tech

Aishwarya Agrawal
Virginia Tech

Jiasen Lu
Virginia Tech

Meg Mitchell
Microsoft Research

Dhruv Batra
Virginia Tech

Larry Zitnick
Microsoft Research

Devi Parikh
Virginia Tech

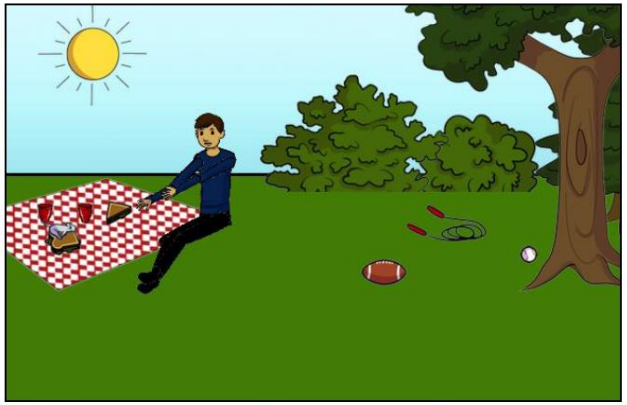
VQA: Visual Question Answering
Antol et al., ArXiv, 2015.

visualqa.org

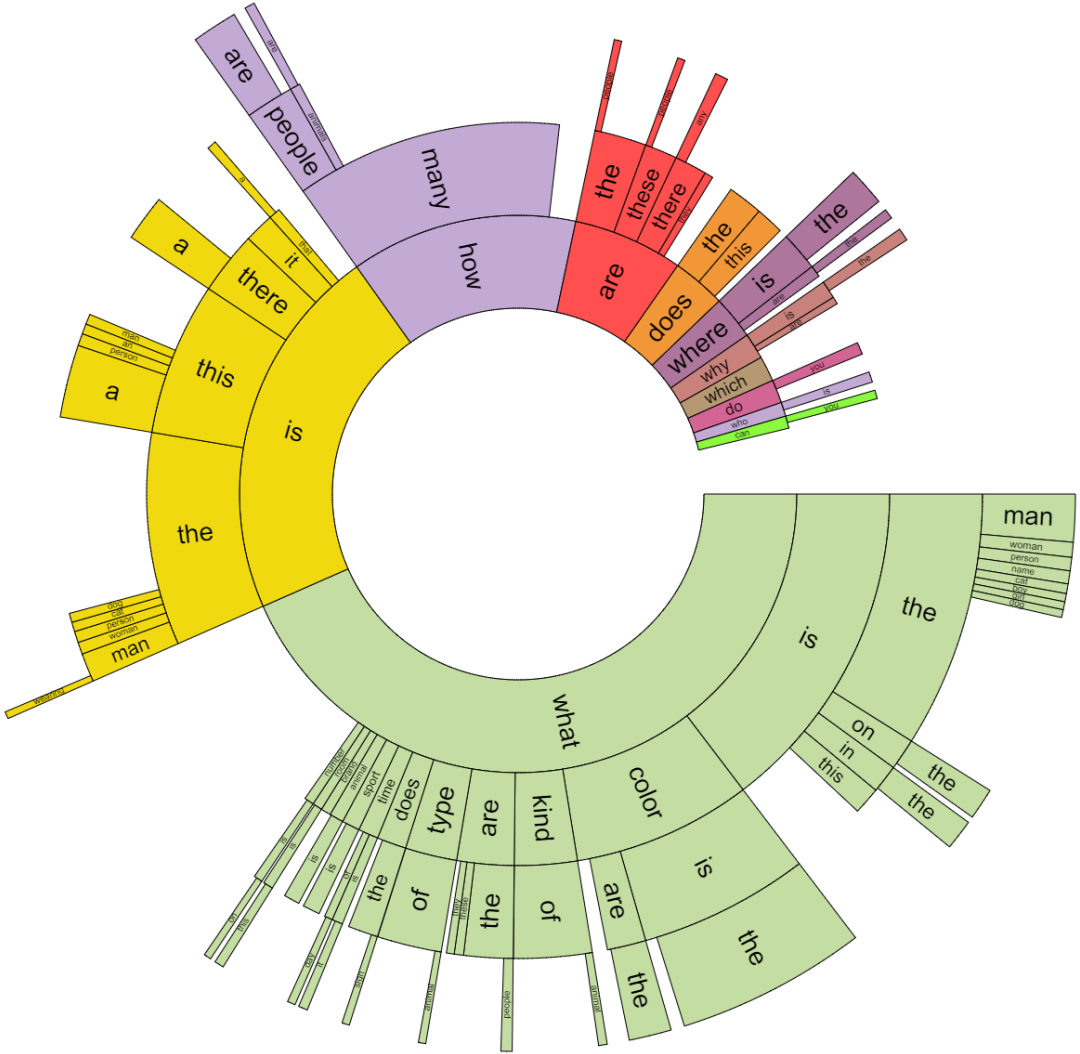
VQA: Visual Question Answering

How many slices of pizza are there?
Is this a vegetarian pizza?

Does it appear to be rainy?
Does this person have 20/20 vision?



Is this person expecting company?
What is just under the tree?


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

VQA: Visual Question Answering

July 2015 (Beta release)
- 120k images (360k questions, 3.6M answers)

September 2015 (Full release)
- 120k COCO train+val images
- 60k “random” images
- 50k abstract scenes

visualqa.org

Conclusion