

COMP 646: Deep Learning for Vision and Language

Referring Expressions and Visually Grounded Question Answering



Last Class

- Quiz
- Quiz Recap

Today

- Referring Expressions
 - Referring Expressions vs Image Captions
 - Generating Referring Expressions
 - Referring Expression Comprehension

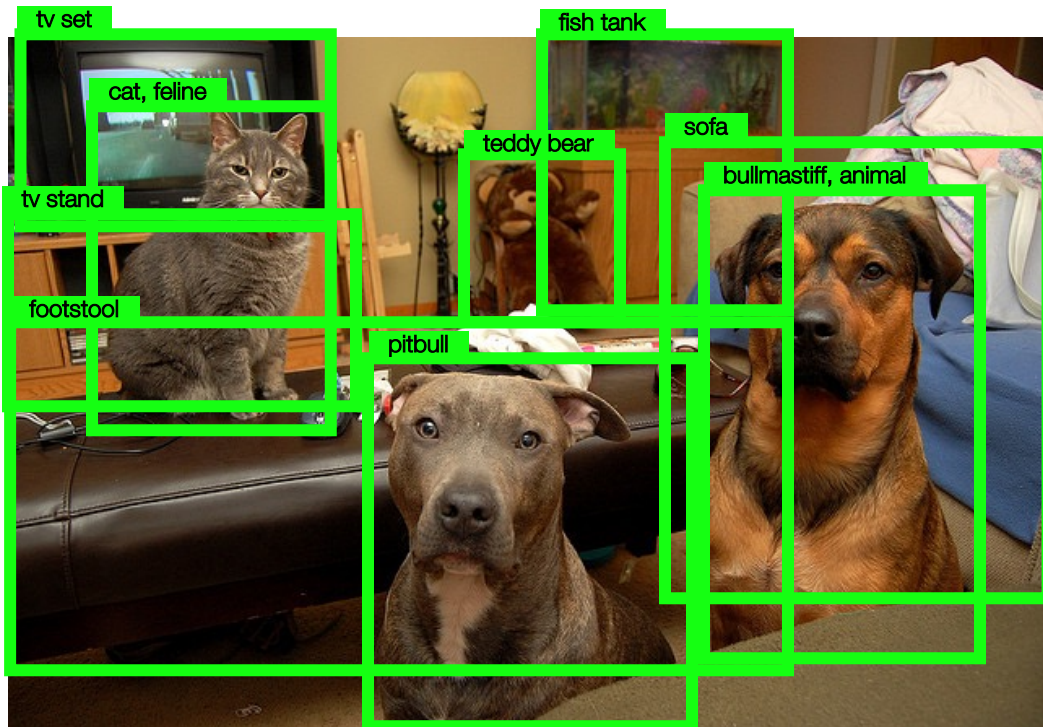
Computer Vision



Image tagging / Image classification

feline
tv set
teddy bear
pitbull
bullmastiff
cat
tv stand
group of dogs
fish tank
room
indoor
man-made
footstool
furniture

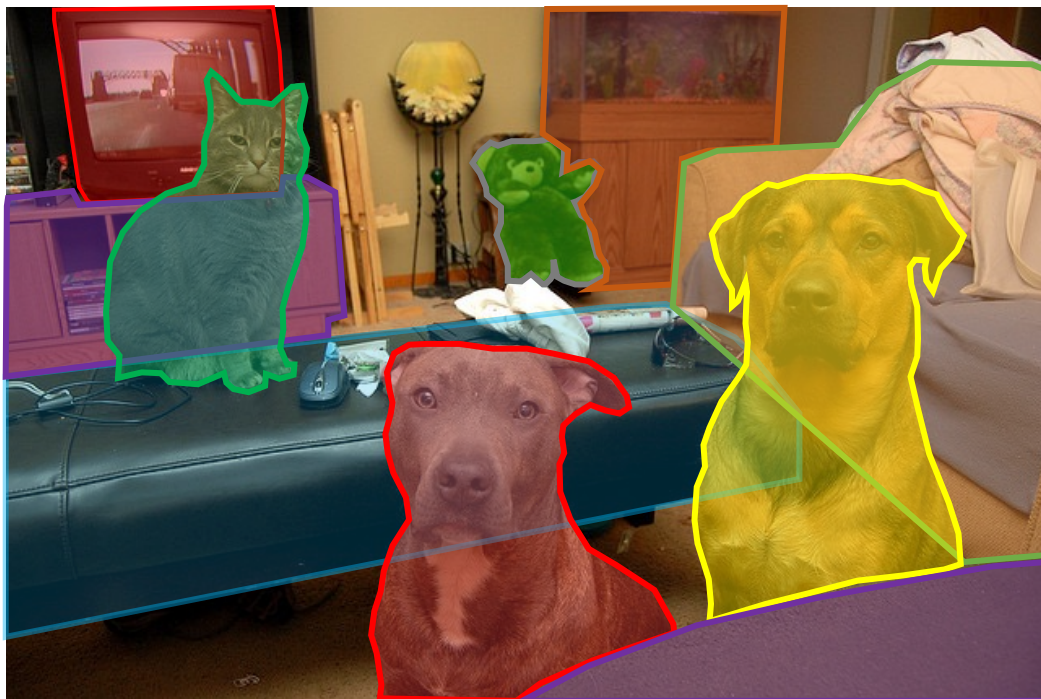
Computer Vision



Object Detection

feline
tv set
teddy bear
pitbull
bullmastiff
cat
tv stand
group of dogs
fish tank
room
indoor
man-made
footstool
furniture

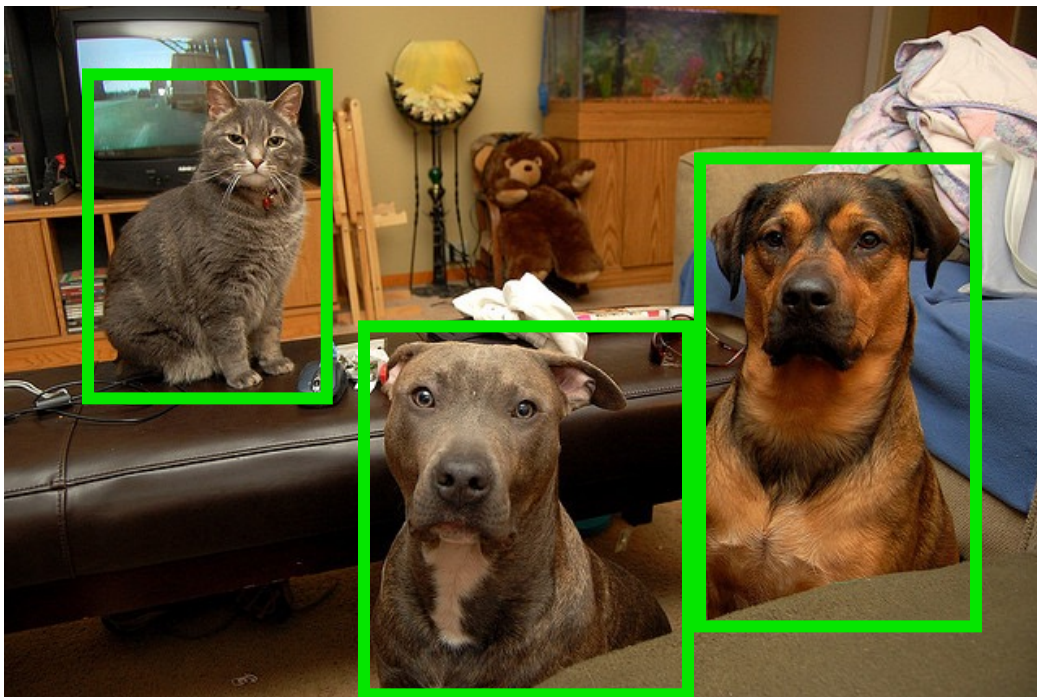
Computer Vision



- feline
- tv set
- teddy bear
- pitbull
- dog
- cat
- tv stand
- group of dogs
- fish tank
- room
- indoor
- man-made
- footstool
- furniture

Image Parsing / Image Segmentation

How do we describe images?



Object
Importance

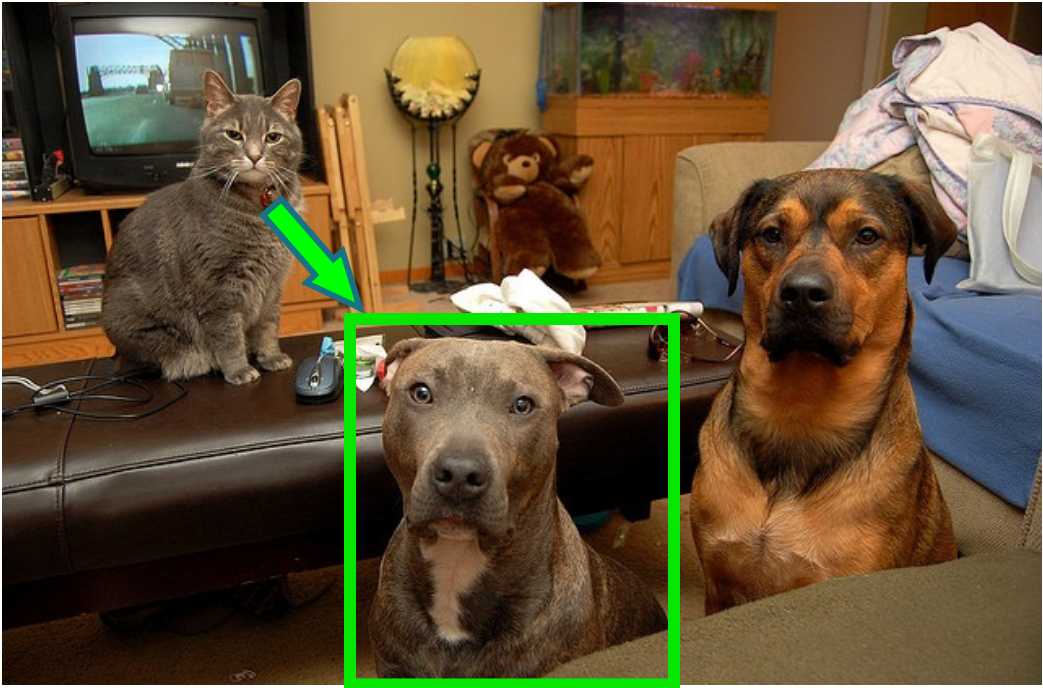
Attribute
Importance

Action
Importance

World
knowledge

A cat and two big dogs staring at the camera

Referring to objects



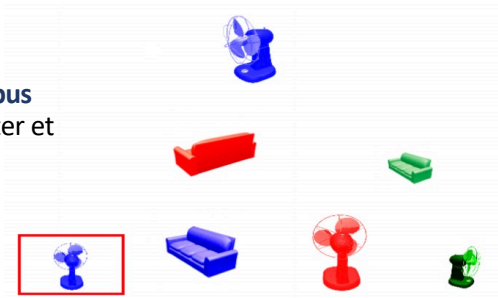
The dog
in the
middle

The gray
dog in the
middle

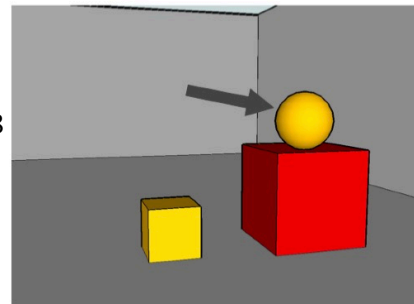
The gray
dog

Work on Referring Expression

TUNA Corpus
van Deemter et
al 2006



GRE3D3 Corpus
Viethen and Dale 2008
[20 scenes]



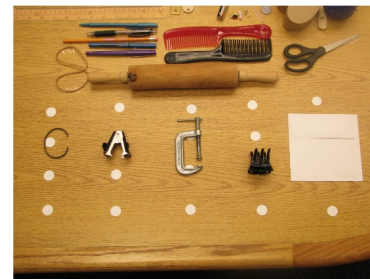
Size Corpus
Mitchell et al 2011
[96 scenes]



GenX Corpus
FitzGerald et al 2013
[269 scenes]



Typicality Corpus
Mitchell et al 2013
[35 scenes]



Referit Game


Player 1



✓ Like Share You, Nanxi Che and 56 others like this. 29692 Games Played Goal: 100,000

Time Elapsed
19

Score
38



Orange bottle on the right

Player 2



✓ Like Share You, Nanxi Che and 56 others like this. 29692 Games Played Goal: 100,000

Time Elapsed
19

Score
38

Orange bottle on the right

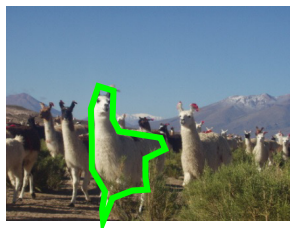


Submit

Referring Expressions for Natural Scenes

Diverse

Many real world objects



Complex

Many object instances



Big



Referit Game Dataset



Blue shirt man

Blue guy

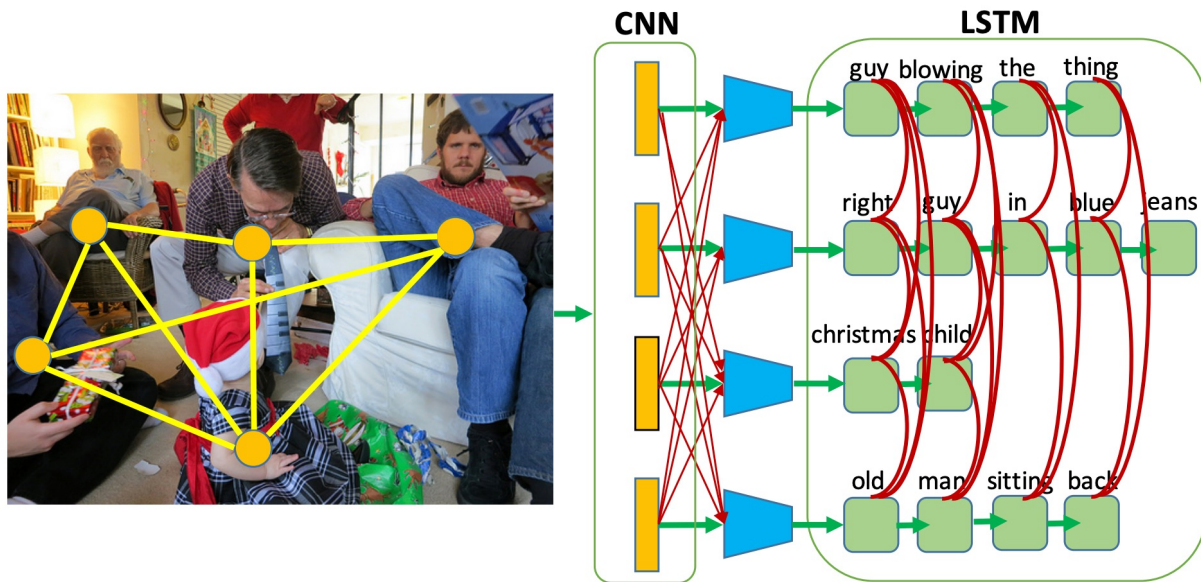
Second guy from left

ReferItGame Dataset

130k Referring expressions for **90k** Objects in **19k** images

ReferItGame: Referring to Objects in Photographs of Natural Scenes
Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, Tamara L. Berg.
Empirical Methods on Natural Language Processing. **EMNLP 2014**.

Deep Generation of Referring Expressions



Modeling Context in Referring Expressions

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, Tamara L. Berg

2016

Department of Computer Science,
University of North Carolina at Chapel Hill
{licheng,poirson,alexyang,aberg,tlberg}@cs.unc.edu

RefCOCO+ testA



Baseline: blue shirt

MMI: black shirt

visdif: person in striped shirt

visdif+tie: arm with striped shirt



Baseline: tennis player

MMI: girl

visdif: woman in white

visdif+tie: tennis player

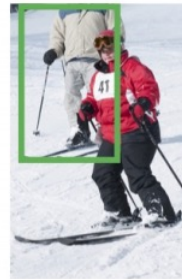


Baseline: man

MMI: man

visdif: man with glasses

visdif+tie: man with glasses



Baseline: red jacket

MMI: red jacket

visdif: skier in white

visdif+tie: man in white

RefCOCO+ testB

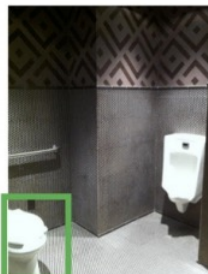


Baseline: plant

MMI: plant that is cut off

visdif: tall plant

visdif+tie: plant on screen side



Baseline: toilet

MMI: toilet

visdif: toilet with lid

visdif+tie: toilet with lid

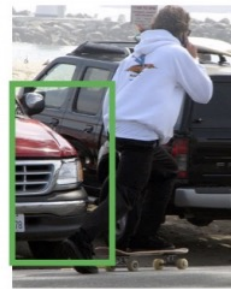


Baseline: donut at 3

MMI: glazed donut

visdif: donut with hole

visdif+tie: donut with hole



Baseline: car with red roof

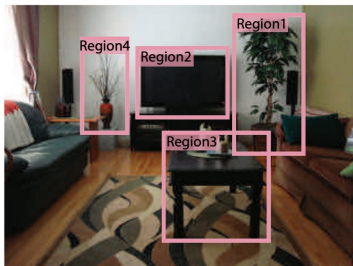
MMI: car

visdif: car with headlights

visdif+tie: car with headlights

Referring Expression Comprehension

The plant on the
right side of the TV

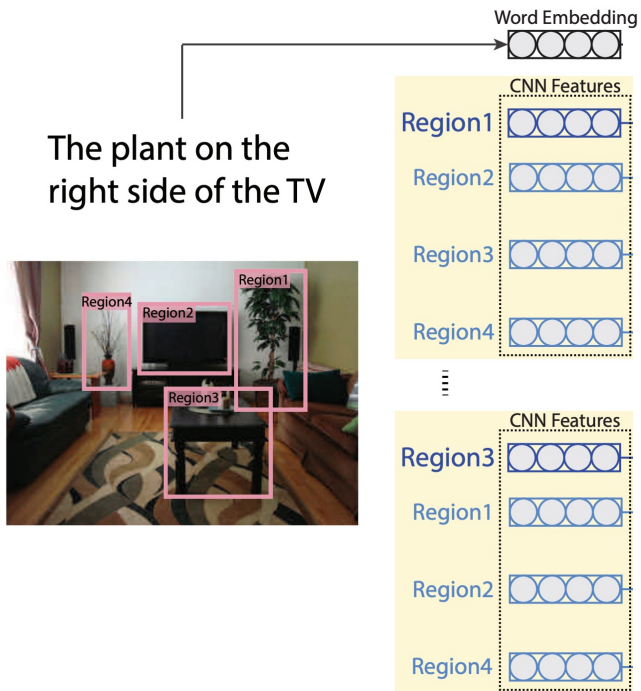


Modeling Context Between Objects for Referring Expression Understanding

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

Referring Expression Comprehension

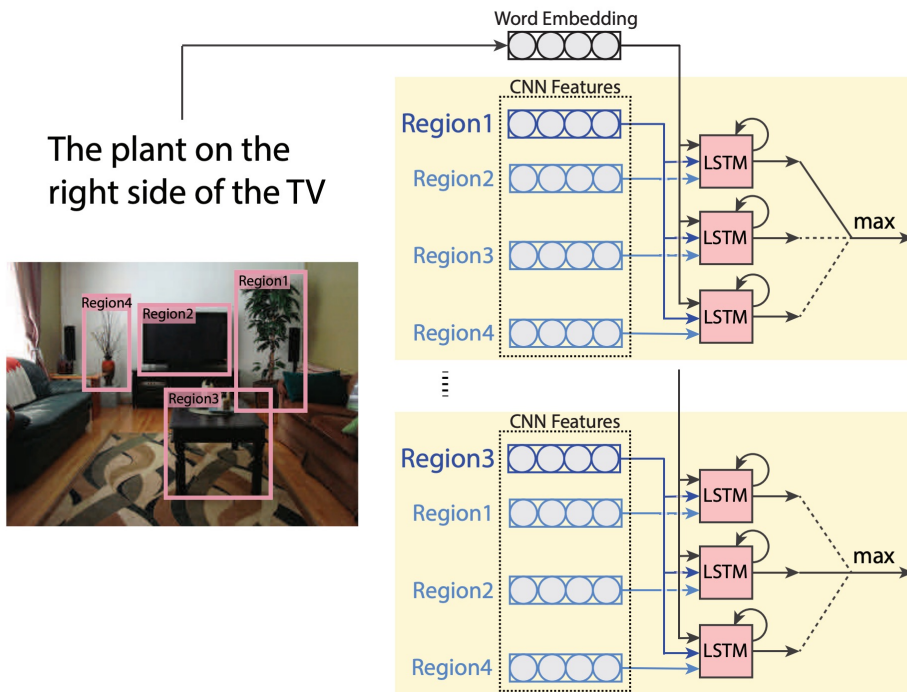


Modeling Context Between Objects for Referring Expression Understanding

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

Referring Expression Comprehension

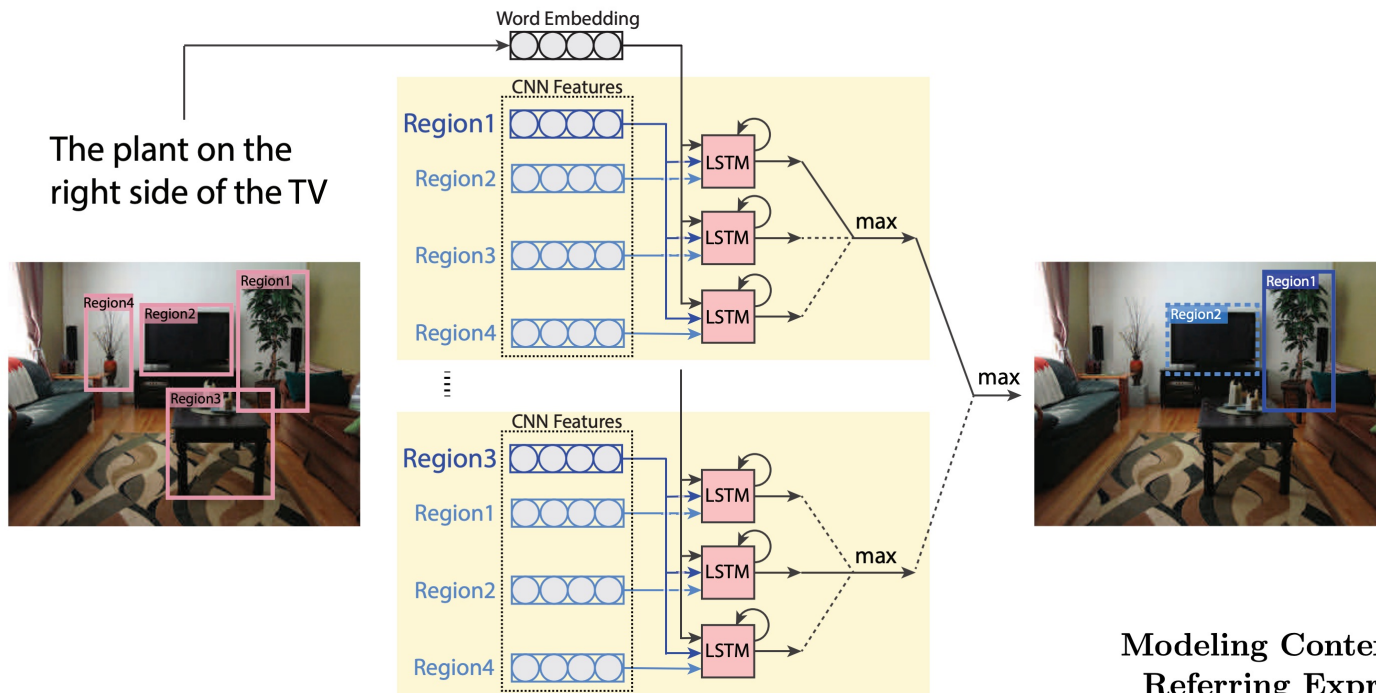


jects for
anding

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

Referring Expression Comprehension



Modeling Context Between Objects for Referring Expression Understanding

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

2016

Other important work

MattNet: Yu et al. <https://arxiv.org/abs/1801.08186>

Mao et al. <https://arxiv.org/abs/1511.02283>

Rohrbach et al. <https://arxiv.org/abs/1511.03745>

Visually Grounded Question Answering



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

Visually Grounded Question Answering



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

VQA: Visual Question Answering

www.visualqa.org

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



Is this person trying
to hit a ball?

yes
yes
yes

yes
yes
yes

What is the person
hitting the ball with?

frisbie
racket
round paddle

bat
bat
racket



What is the guy
doing as he sits
on the bench?

phone
taking picture
taking picture with phone

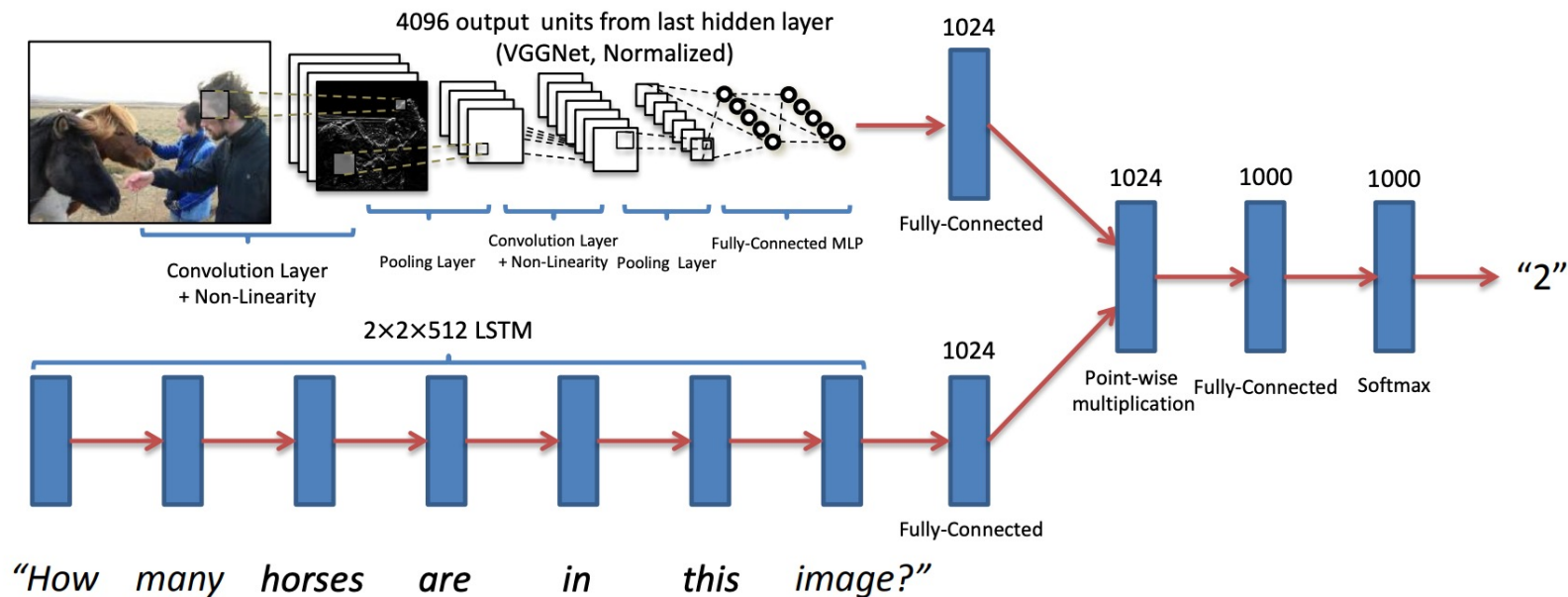
reading
reading
smokes

What color are
his shoes?

blue
blue
blue

black
black
brown

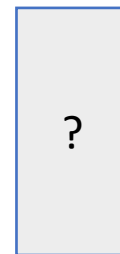
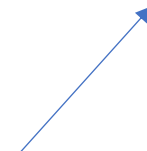
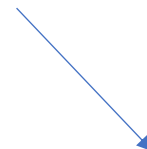
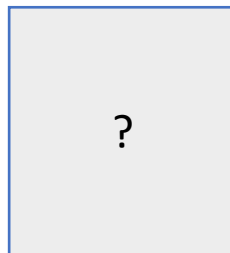
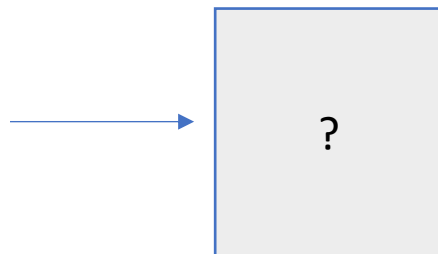
Visually Grounded Question Answering



VQA Solution today?



What is the color of the jacket of the man on this picture?

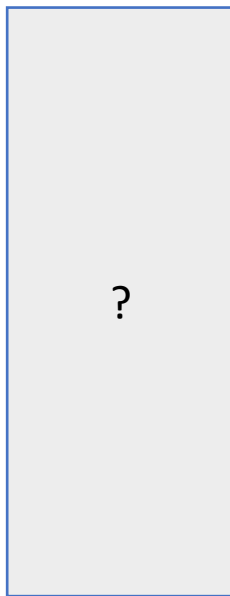


Cross Entropy Loss
Across 5000
possible answers

VQA Solution today?



What is the color of the jacket of the man on this picture?



Cross Entropy Loss
Across 5000
possible answers

Questions?