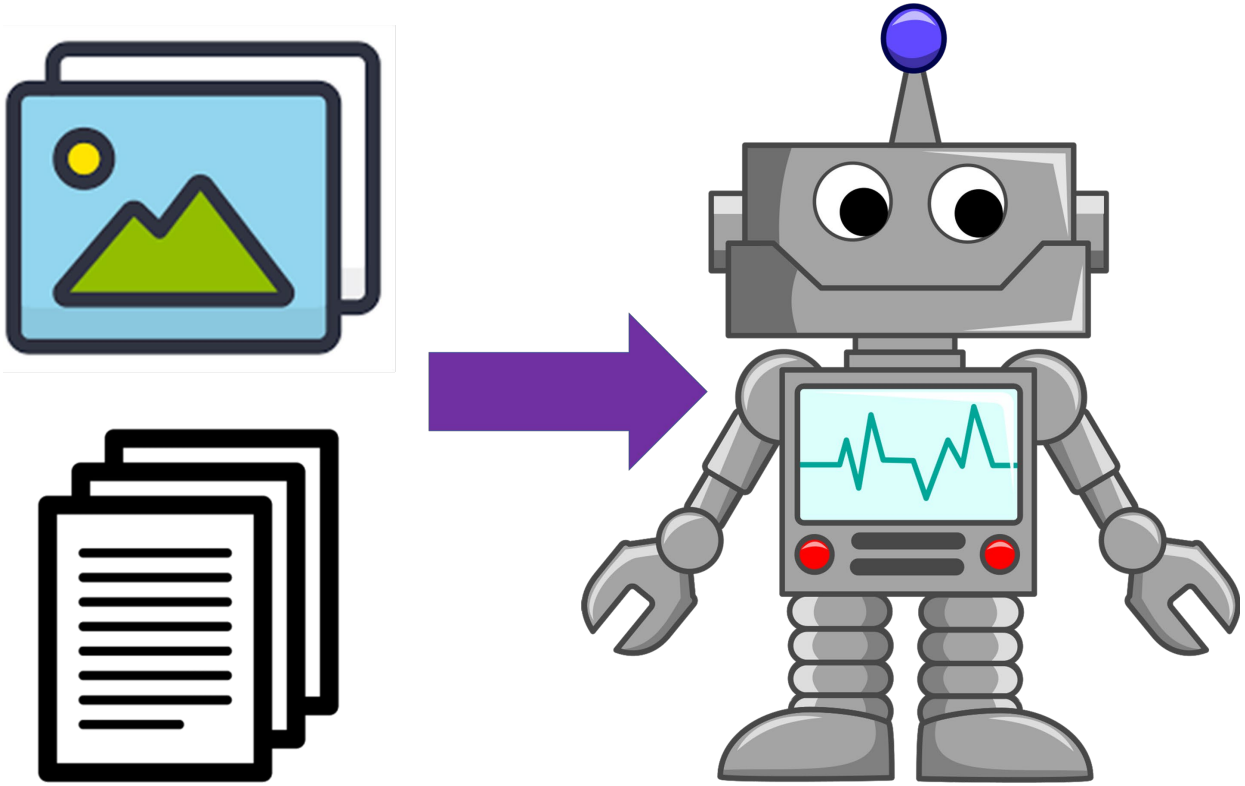
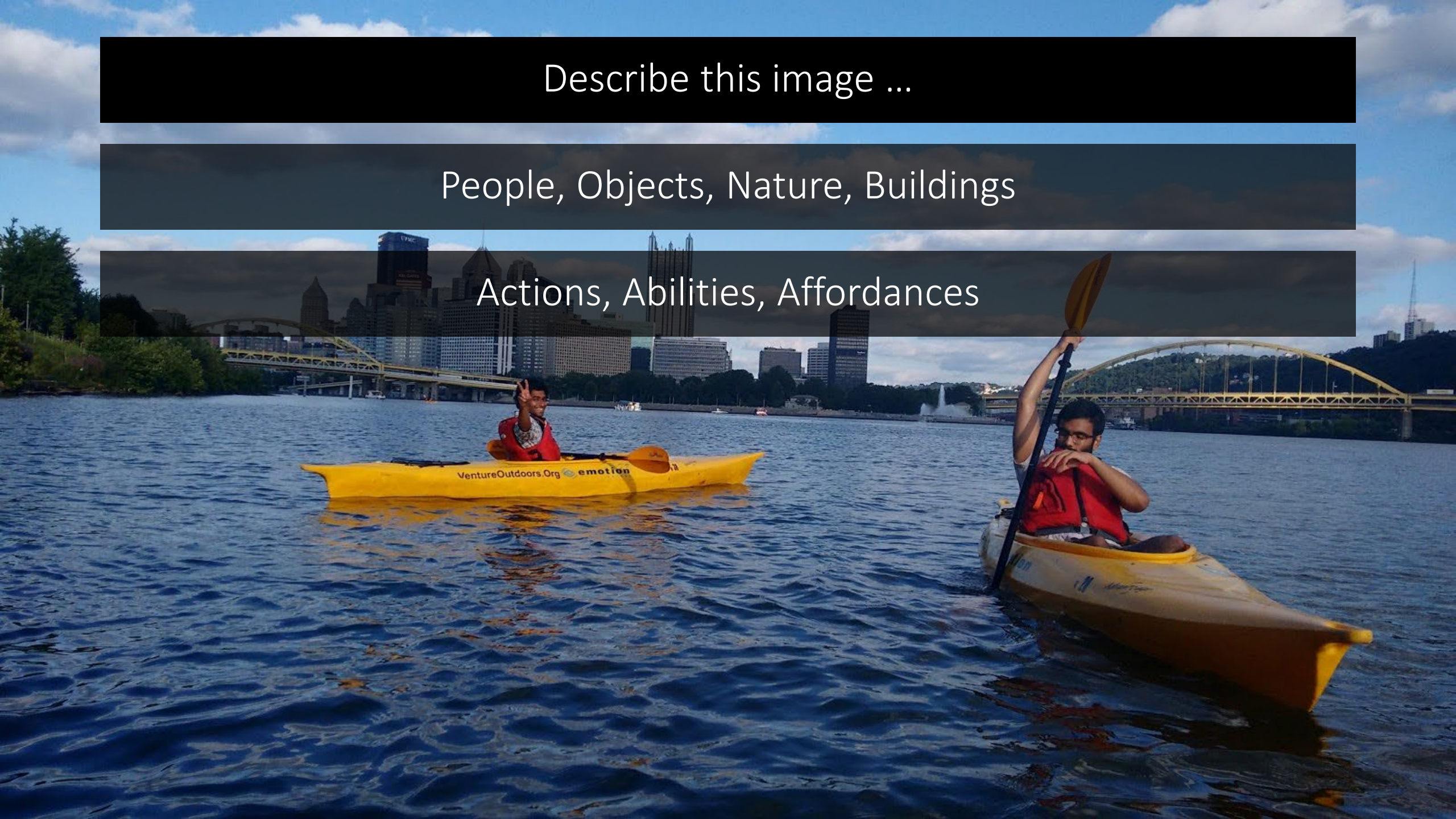
Lecture 20a: "Multimodal" Computer Vision

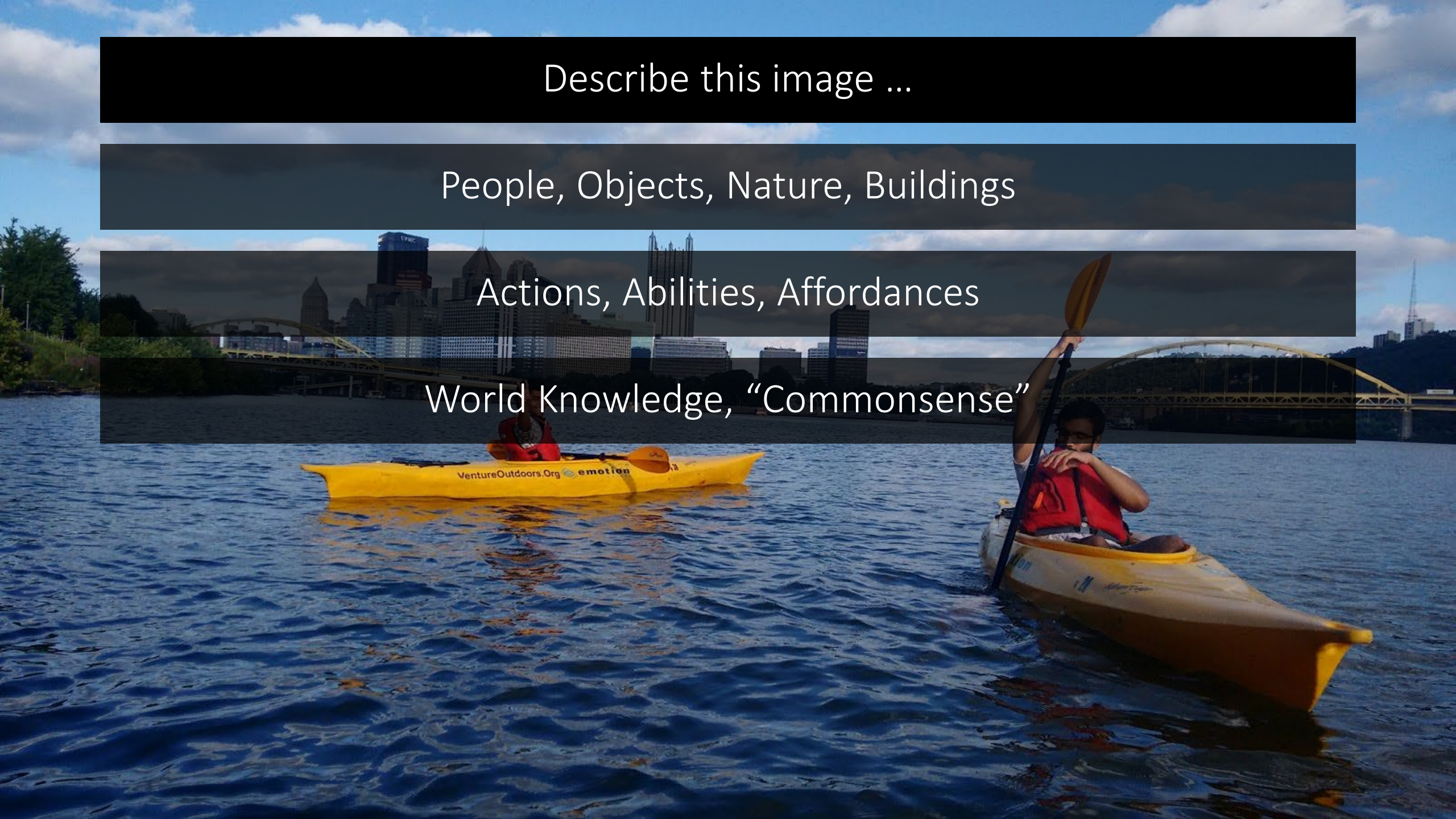
CMSC 472 / 672 Computer Vision

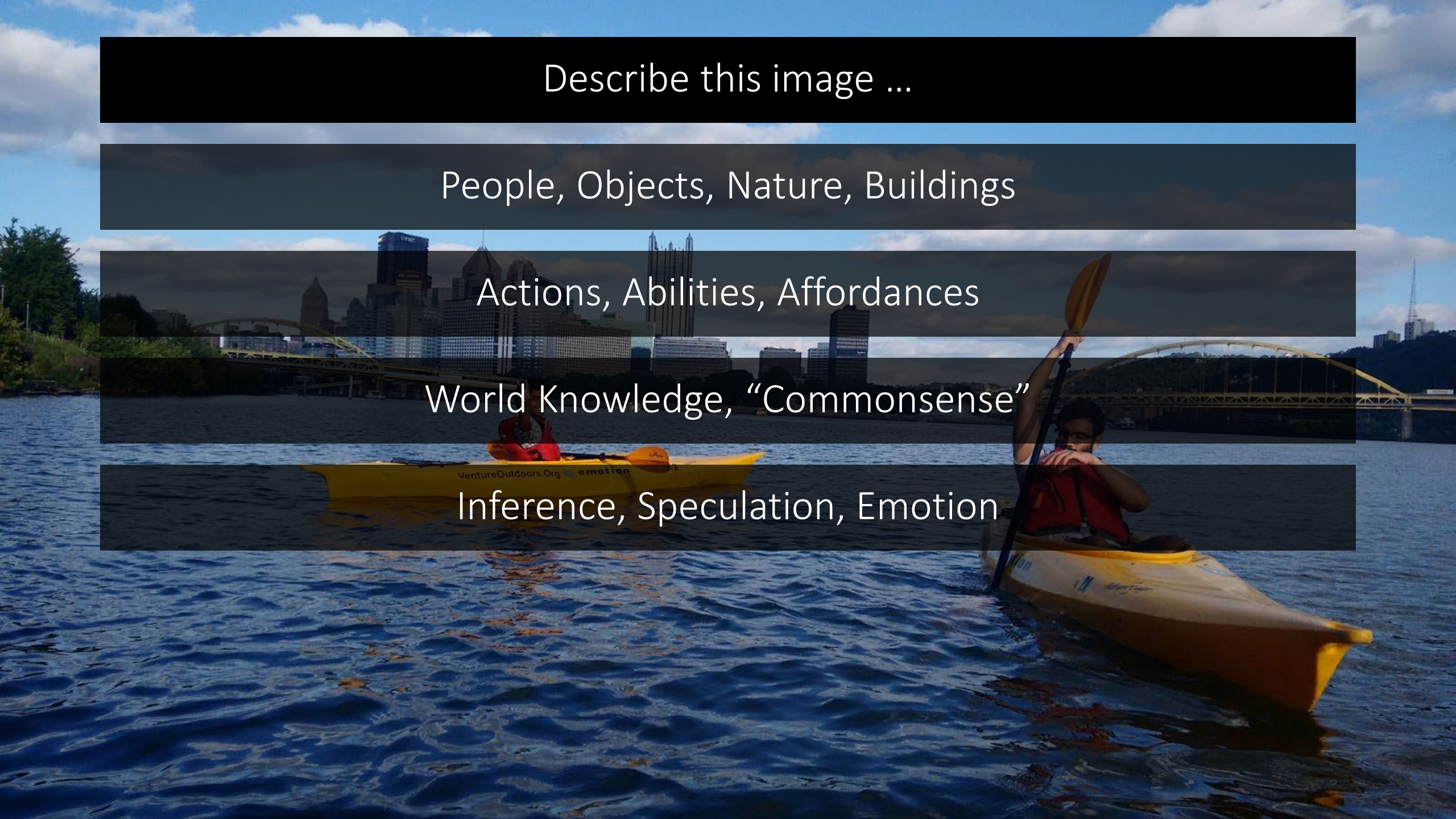


Describe this image …

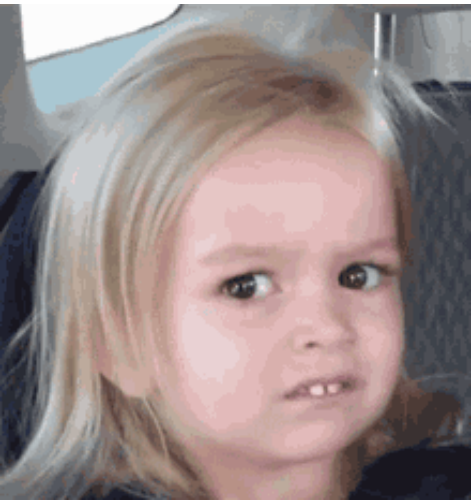

Describe this image …
People, Objects, Nature, Buildings
VentureOutdoors.Org
emotion


Describe this image …
People, Objects, Nature, Buildings
Actions, Abilities, Affordances
VentureOutdoors.Org
emotion


Describe this image …
People, Objects, Nature, Buildings
Actions, Abilities, Affordances
World Knowledge, “Commonsense”
VentureOutdoors.Org
emotion


Describe this image …
People, Objects, Nature, Buildings
Actions, Abilities, Affordances
World Knowledge, “Commonsense”
Inference, Speculation, Emotion

Images convey emotions



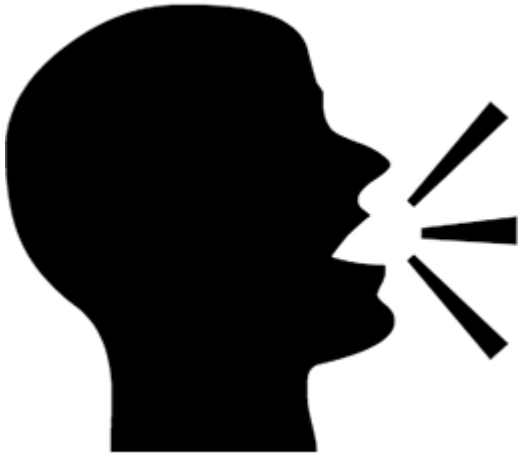
Vision + Language

A brand new era for computer vision!

Vision + Language

A brand new era for computer vision!

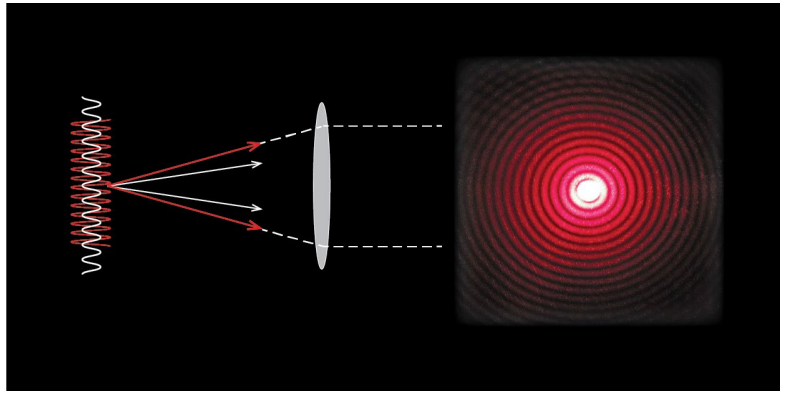
Perception+Reasoning needs Vision+Language



Computer Vision: A Pyramid

PHYSICS-BASED

- Optics
- Computational Imaging

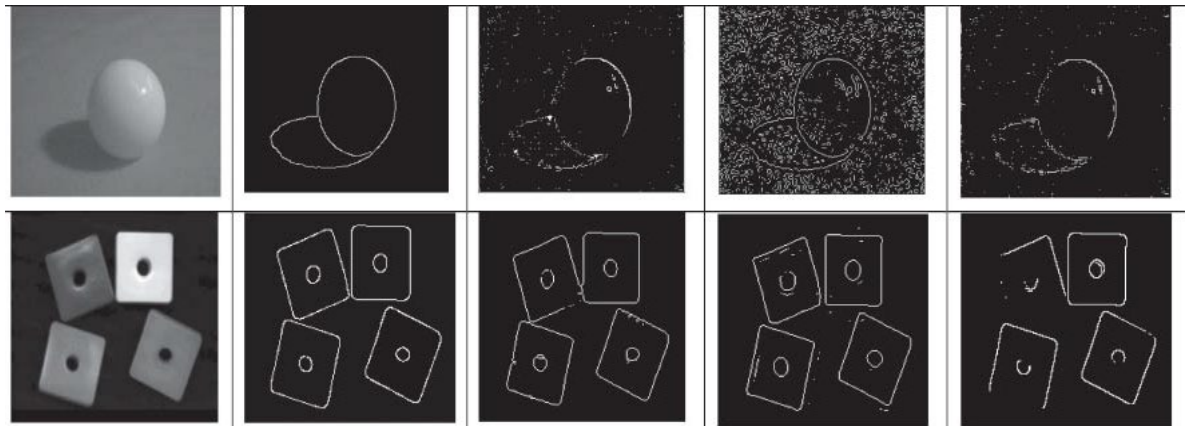


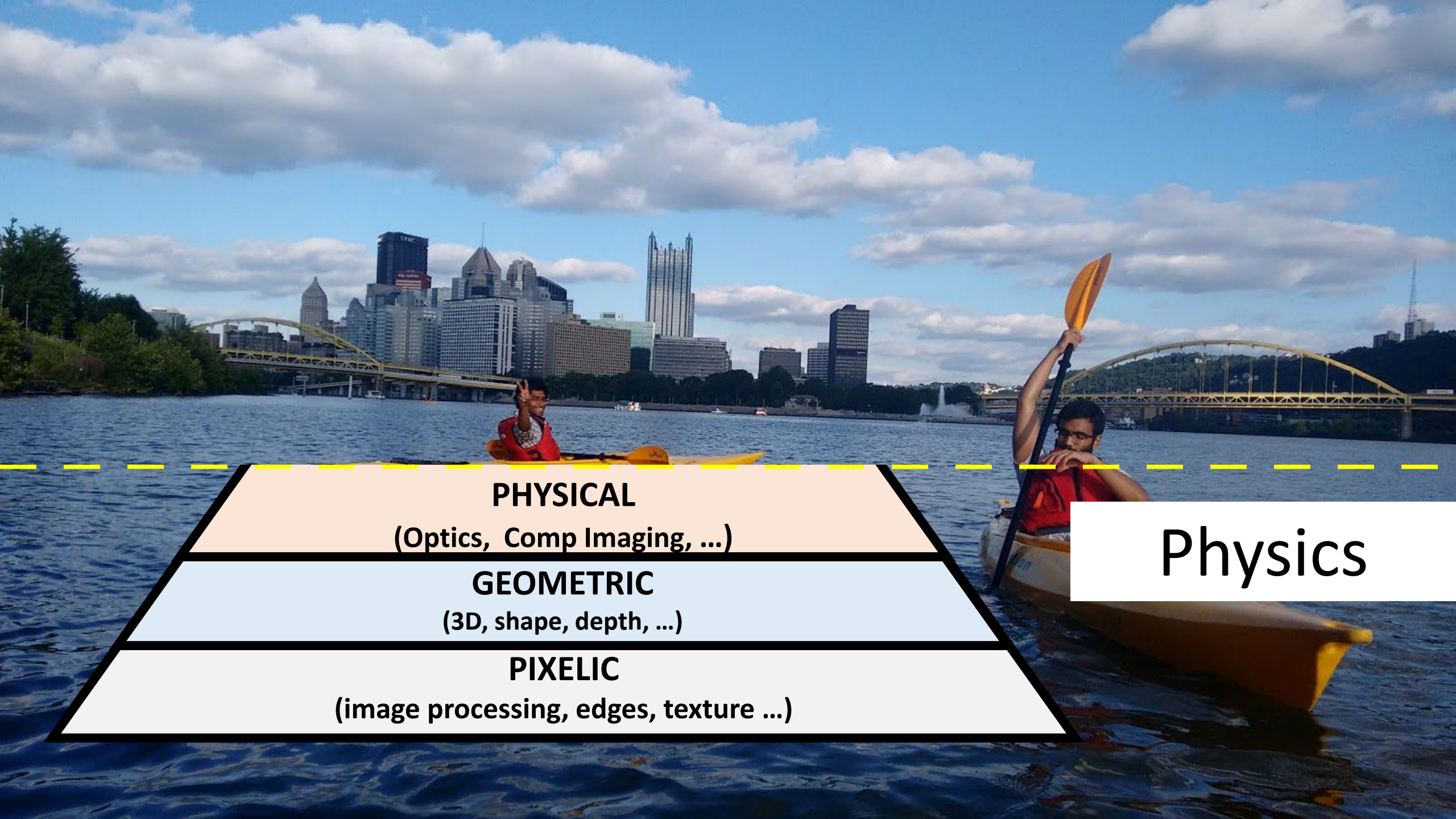
GEOMETRIC

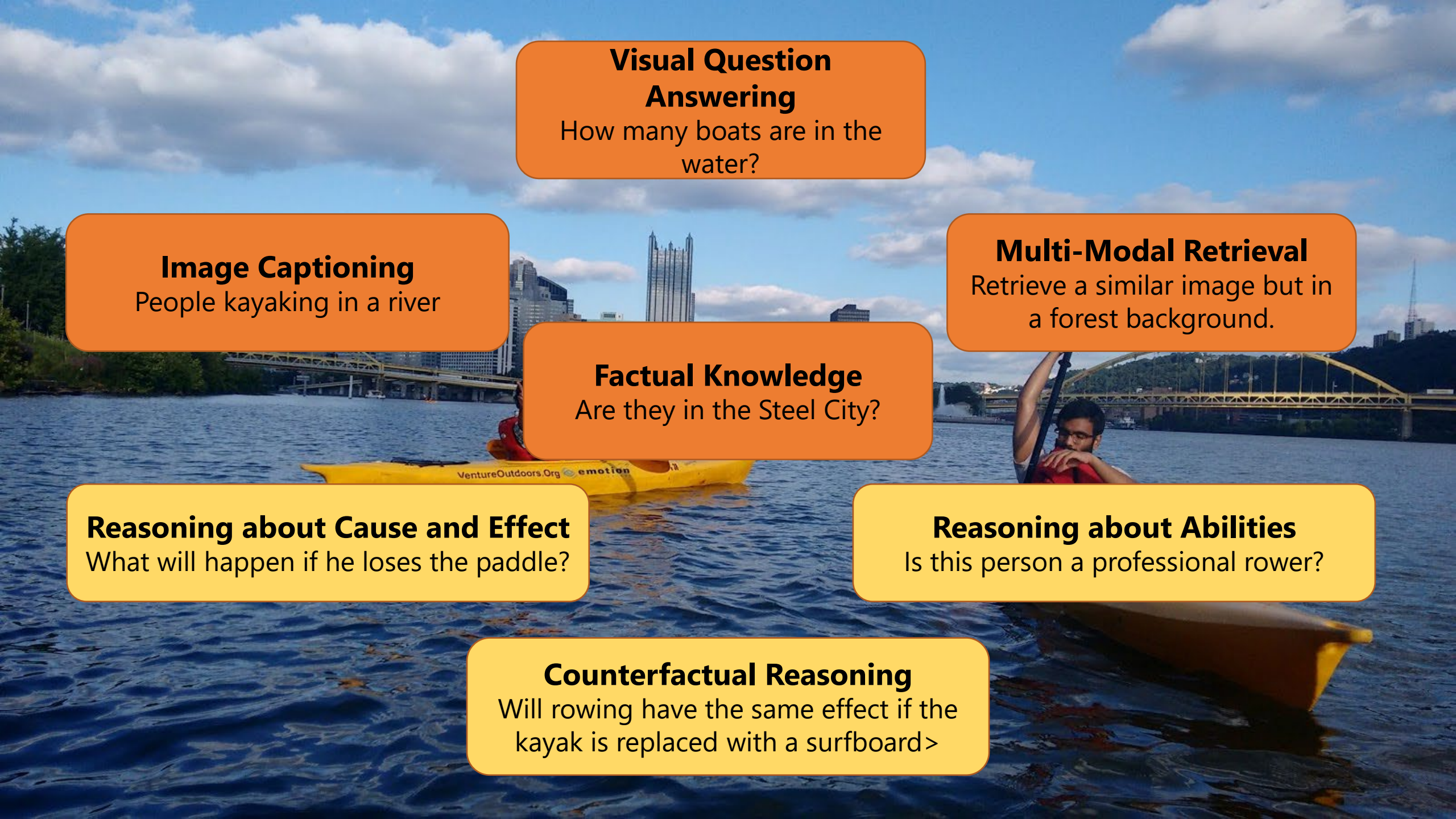
- 3D Reconstruction
- Shape, Depth, ...

PIXELIC

- Image Processing
- Edge Detection




PHYSICAL
(Optics, Comp Imaging, …)
GEOMETRIC
(3D, shape, depth, …)
PIXELIC
(image processing, edges, texture …)
Physics



Visual Question Answering
How many boats are in the water?

Image Captioning
People kayaking in a river

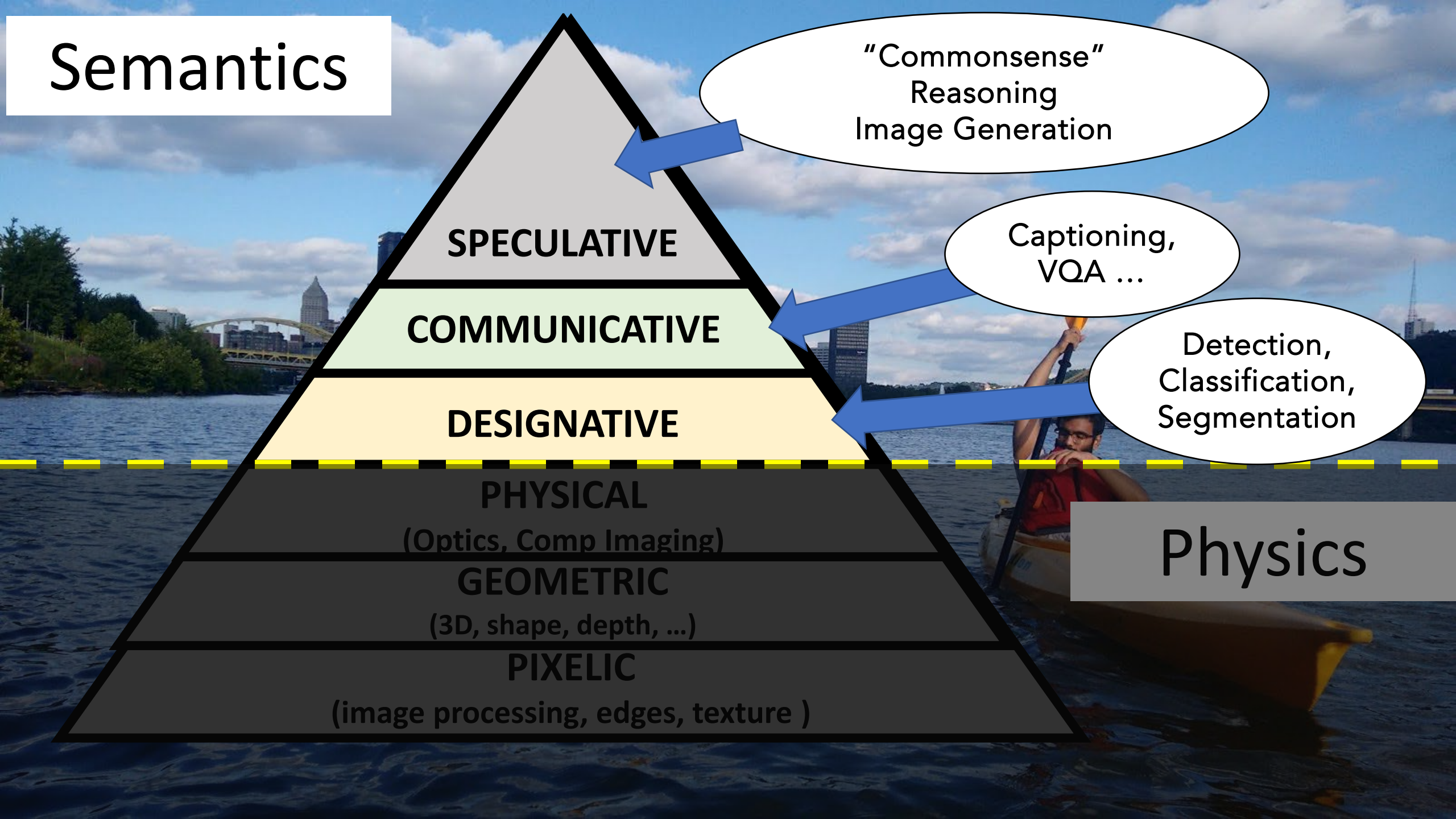
Multi-Modal Retrieval
Retrieve a similar image but in a forest background.

Factual Knowledge
Are they in the Steel City?

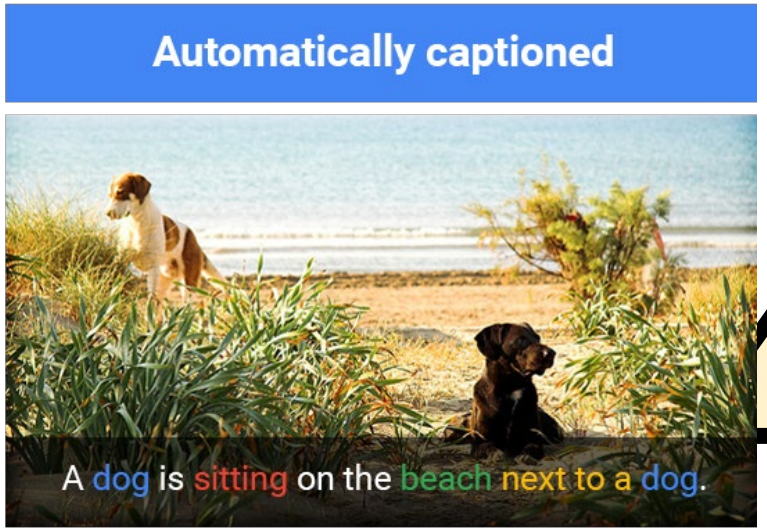
Reasoning about Cause and Effect
What will happen if he loses the paddle?

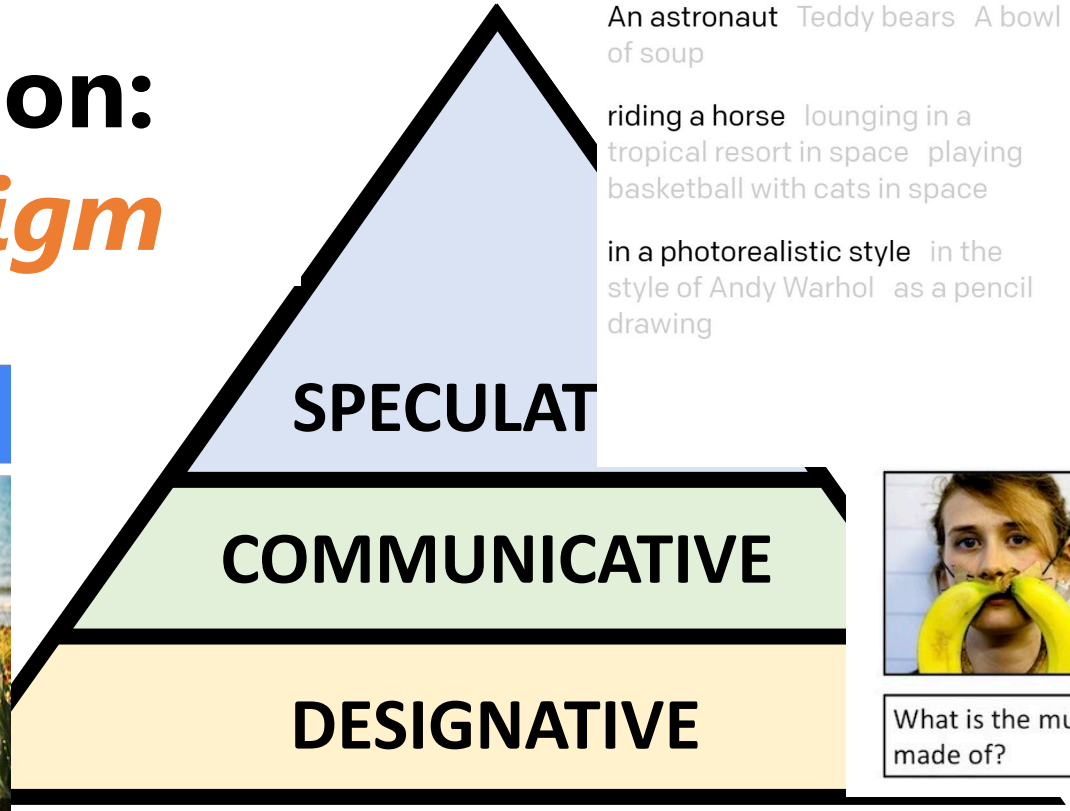
Reasoning about Abilities
Is this person a professional rower?

Counterfactual Reasoning
Will rowing have the same effect if the kayak is replaced with a surfboard>


Semantics
SPECULATIVE
COMMUNICATIVE
DESIGNATIVE
PHYSICAL
(Optics, Comp Imaging)
GEOMETRIC
(3D, shape, depth, …)
PIXELIC
(image processing, edges, texture)
"Commonsense"
Reasoning
Image Generation
Captioning,
VQA …
Detection,
Classification,
Segmentation
Physics

Semantic Vision:
A New Paradigm





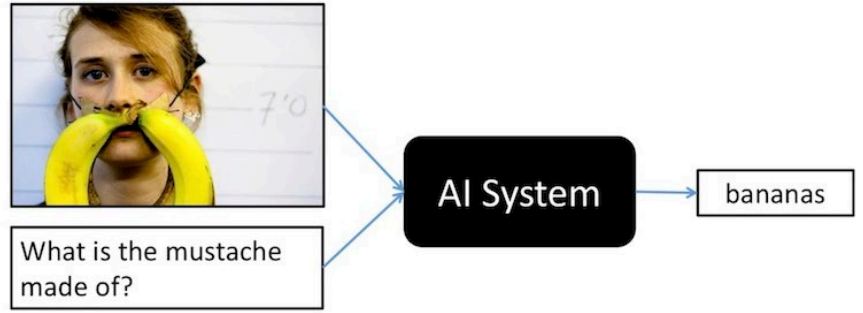
TEXT DESCRIPTION

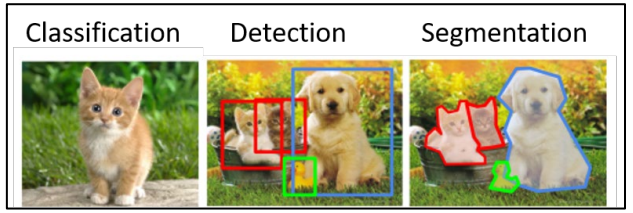
An astronaut Teddy bears A bowl of soup

riding a horse lounging in a tropical resort in space playing basketball with cats in space

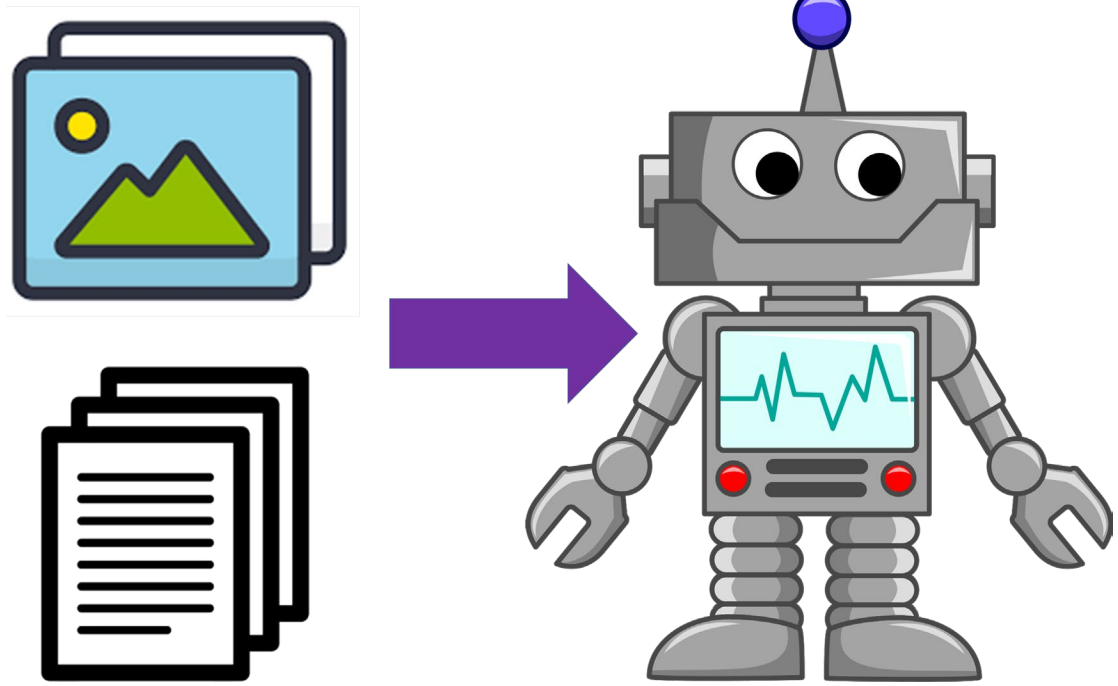
in a photorealistic style in the style of Andy Warhol as a pencil drawing

DALL·E 2





Multi-Modal (Vision + Language) Learning



Multimodal Learning:
Tremendous potential in

- Robotics
- Embodied AI
- Graphics , AR / VR
- Human-Computer Interaction

Learning jointly from images and text has caused a **paradigm shift** in AI

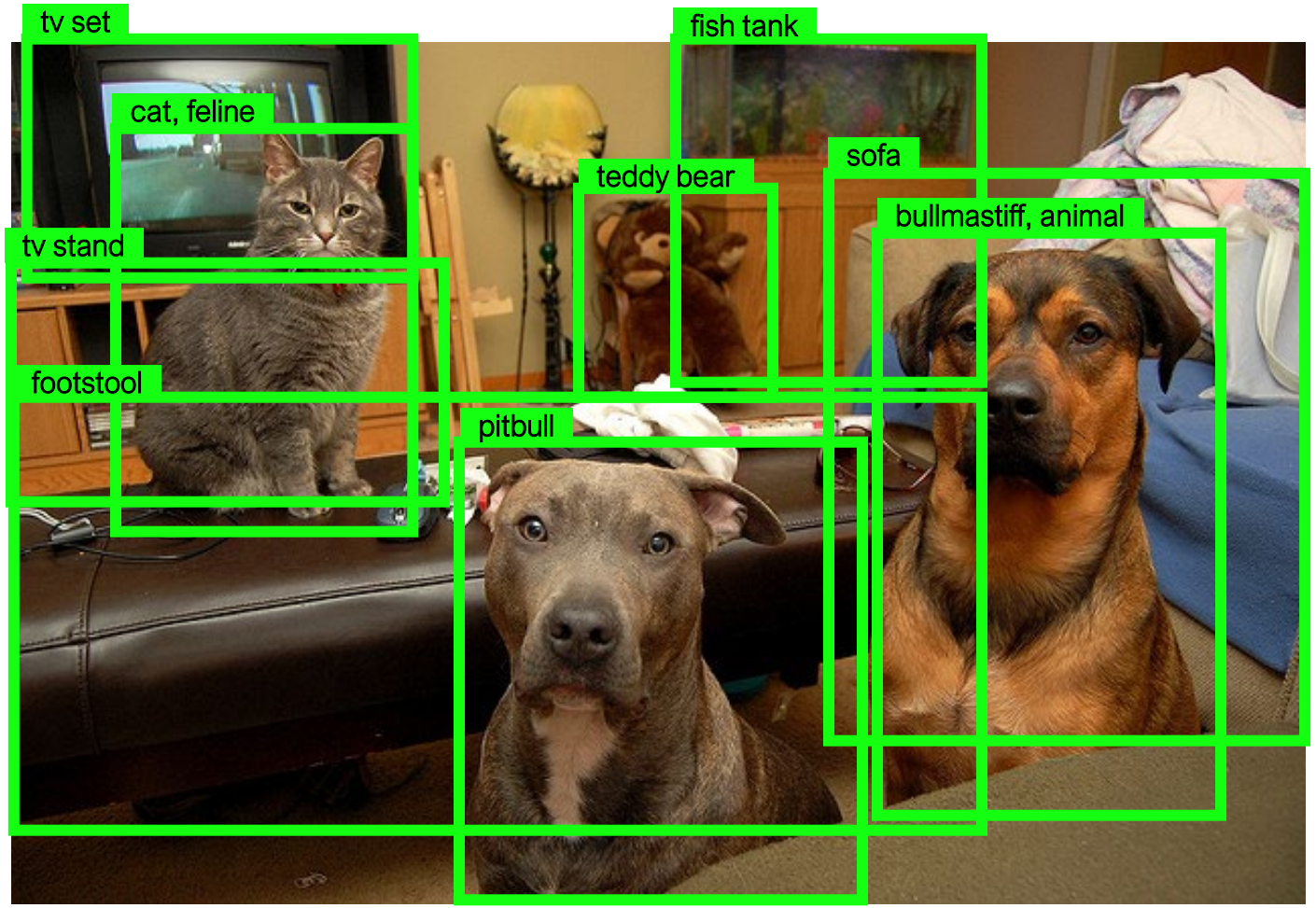

VentureOutdoors.Org
emotion

Old 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

Old Computer Vision



Object Detection

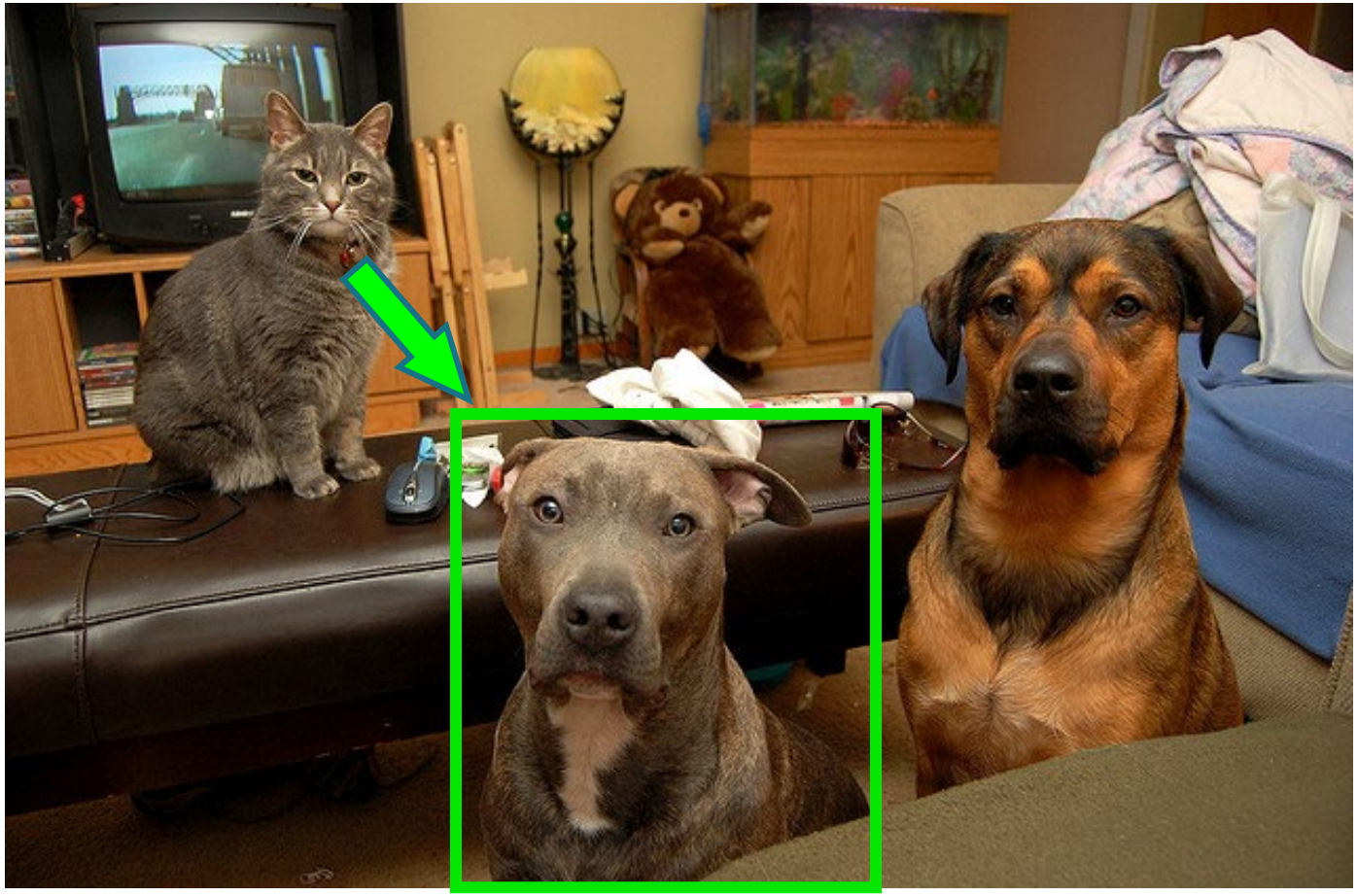
feline
tv set
teddy bear
pitbull
bullmastiff
cat
tv stand
group of dogs
fish tank
room
indoor
man-made
footstool
furniture

Old Computer Vision

Image Parsing / Image Segmentation

New tasks with Vision + Language

Referring to objects



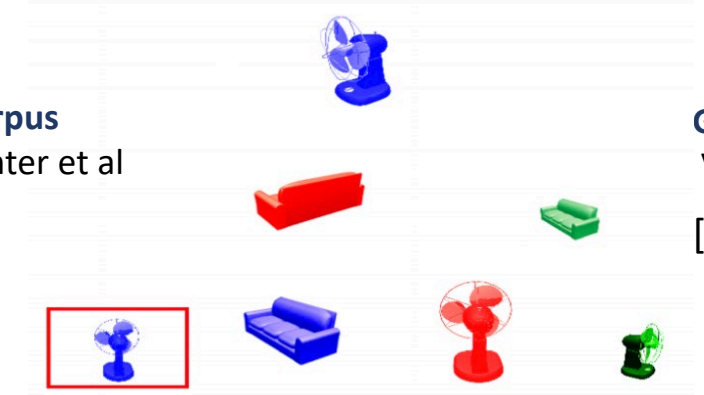
The dog in the middle

The gray dog in the middle

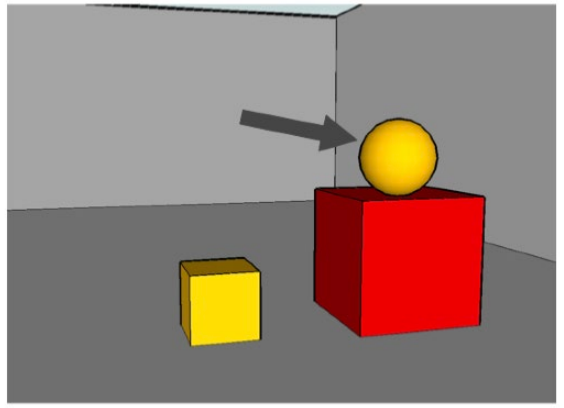
The gray dog

Work on Referring Expression

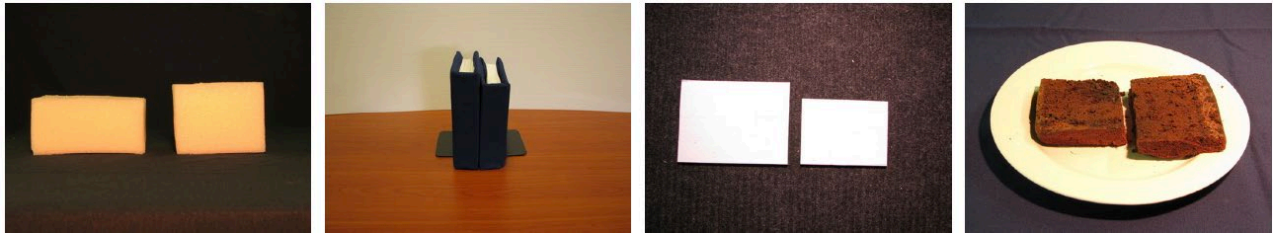
TUNA Corpus
van Deemter et al 2006



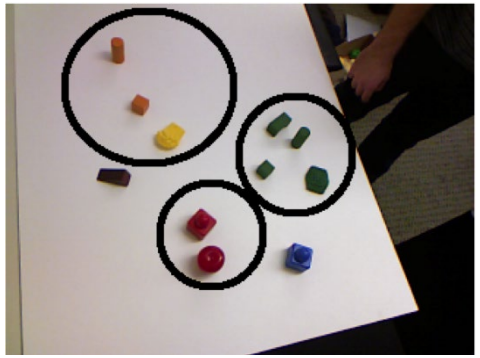
GRE3D3 Corpus
Viethen and Dale 2008
[20 scenes]



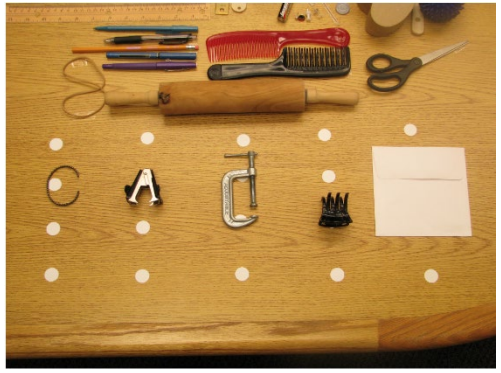
Size Corpus
Mitchell et al 2011
[96 scenes]



GenX Corpus
FitzGerald et al 2013
[269 scenes]

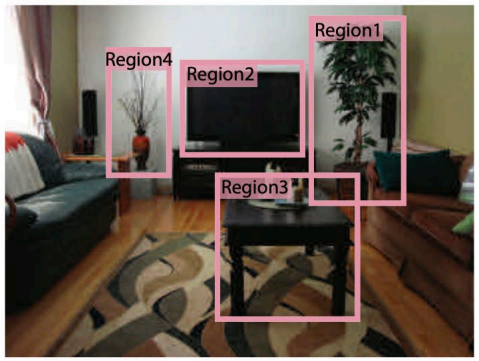


Typicality Corpus
Mitchell et al 2013
[35 scenes]



Referring Expression Comprehension

The plant on the
right side of the TV

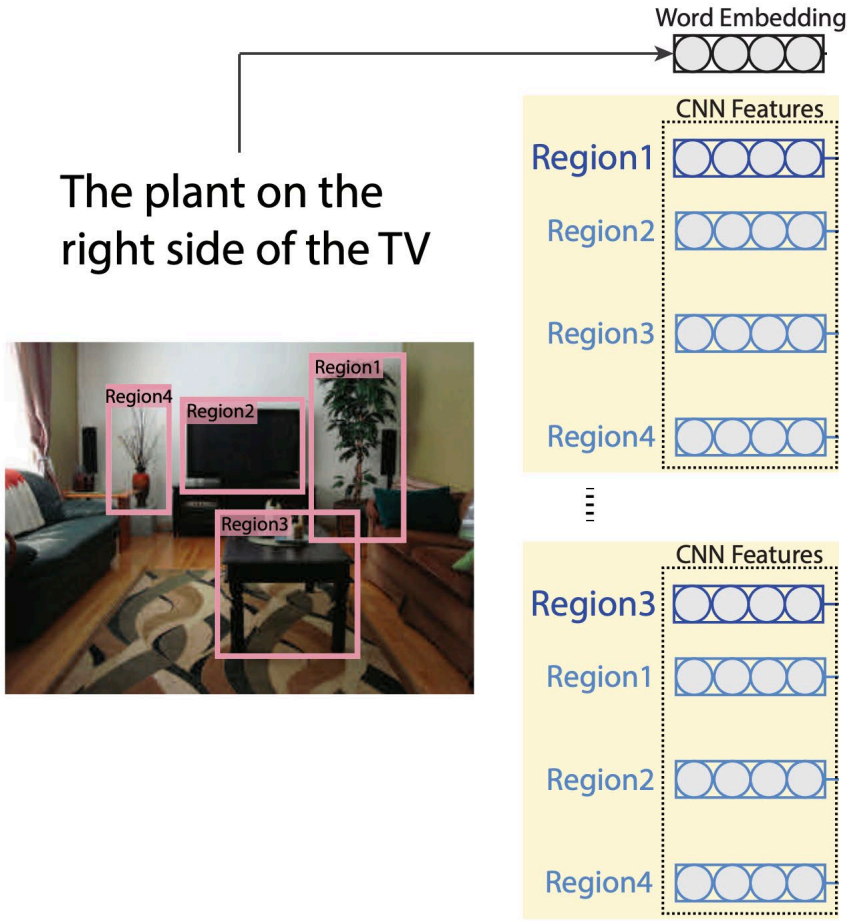


Modeling Context Between Objects for Referring Expression Understanding

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

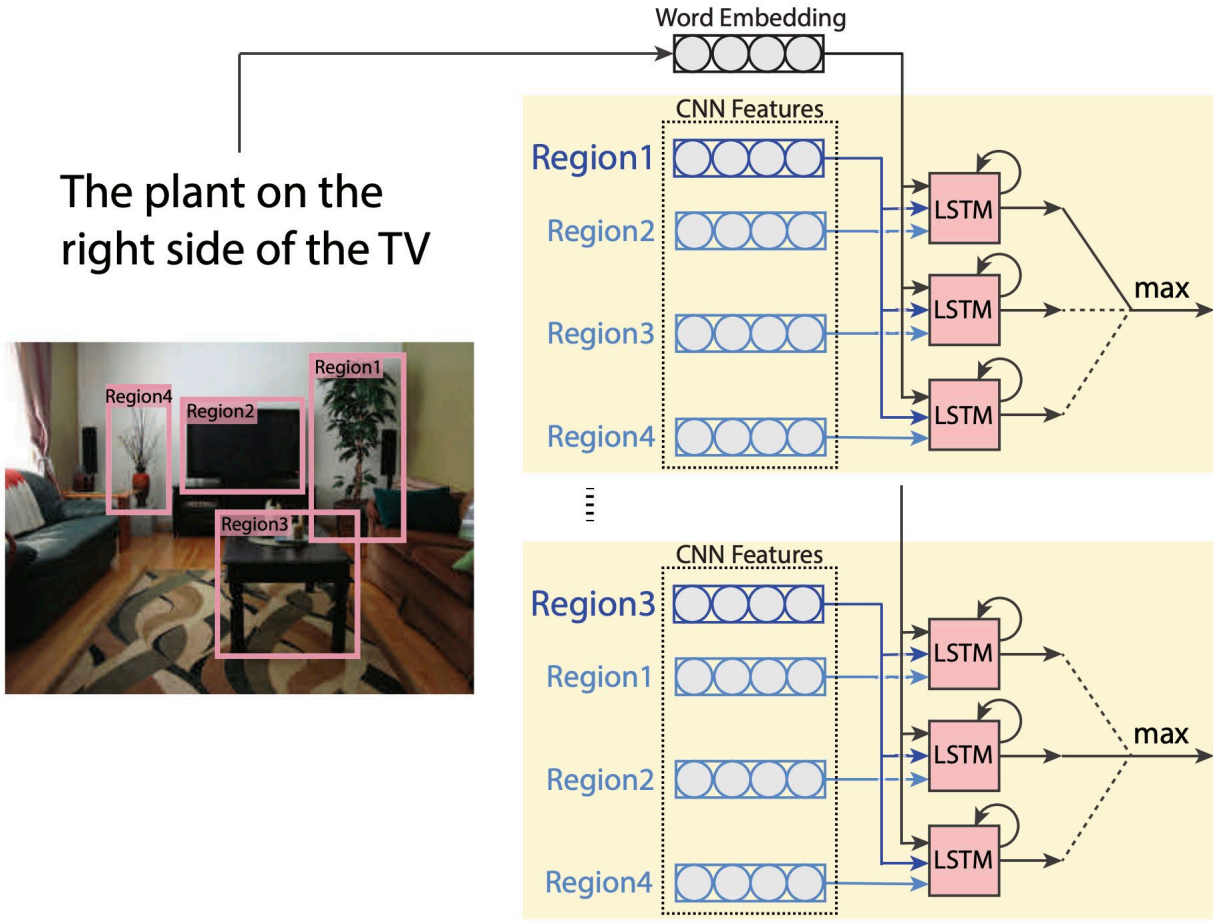


Modeling Context Between Objects for Referring Expression Understanding

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

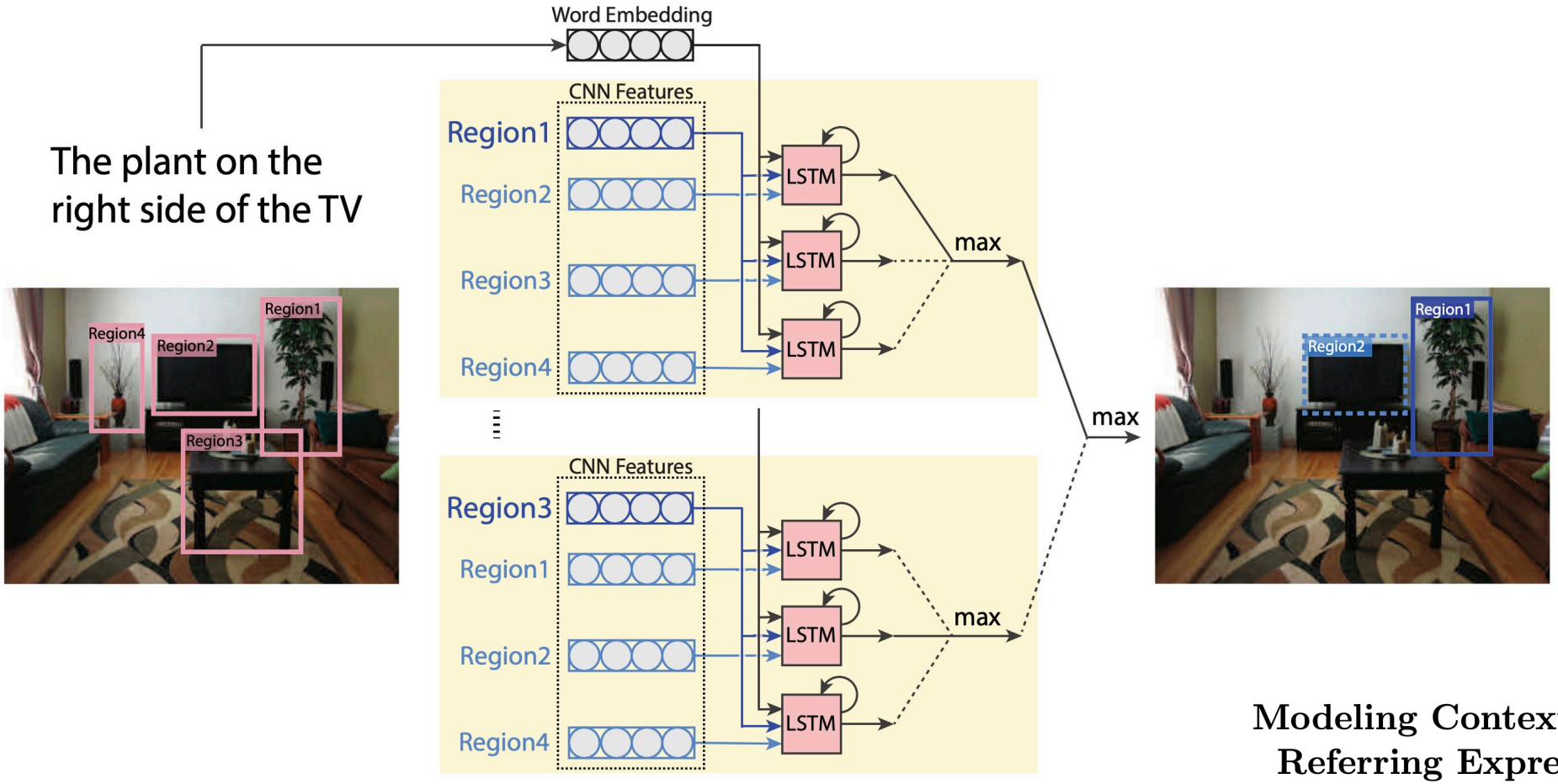


Modeling Context Between Objects for Referring Expression Understanding

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

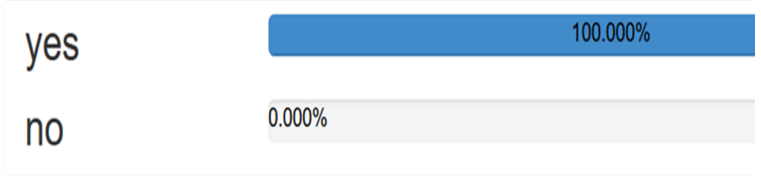
University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

2016

Visual Question Answering

Given an image and a question about it, produce an answer to that question.

Is the food made of eggs?



VQA: Visual Question Answering

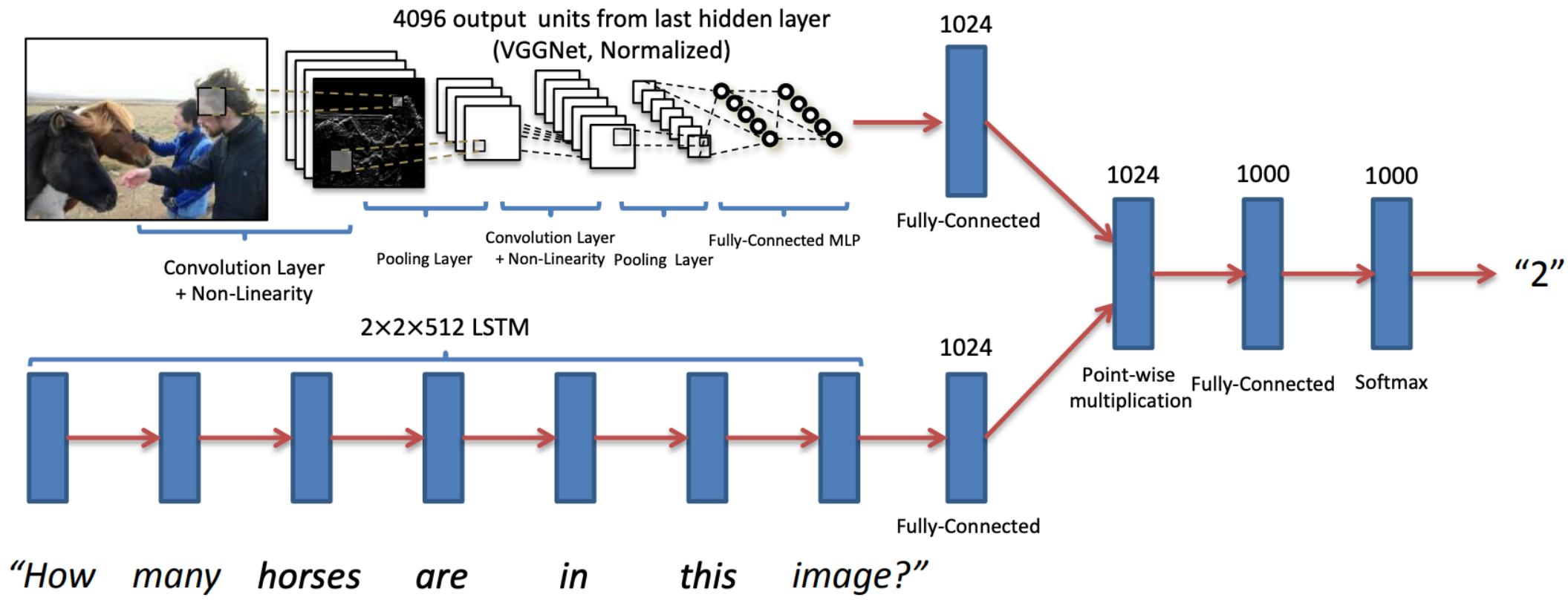
www.visualqa.org

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

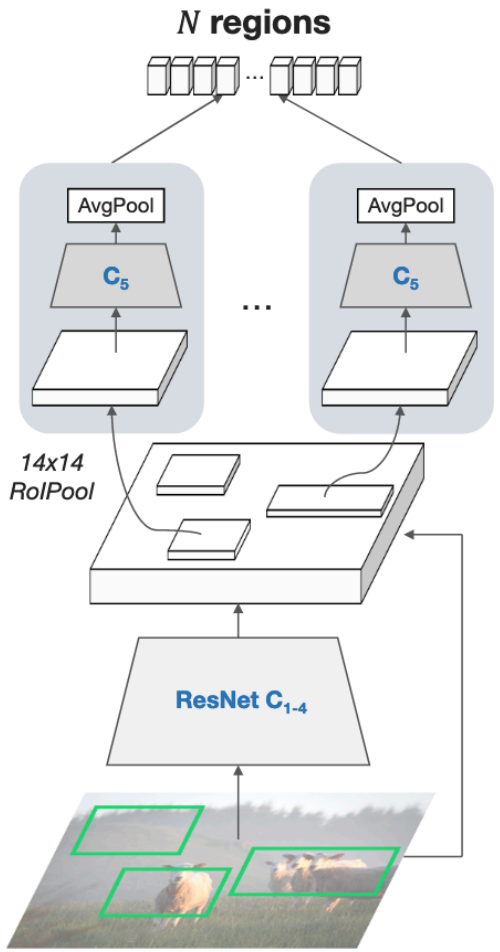
Is this person trying to hit a ball?	yes yes yes	yes yes yes
What is the person hitting the ball with?	frisbie racket round paddle	bat bat racket

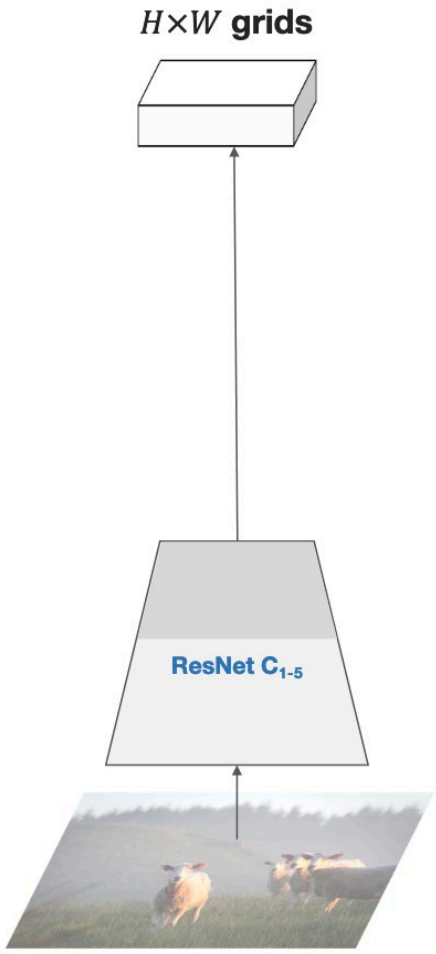
What is the guy doing as he sits on the bench?	phone taking picture taking picture with phone	reading reading smokes
What color are his shoes?	blue blue blue	black black brown

Visual Question Answering: Naïve Approach



What Features to use as input visual features?

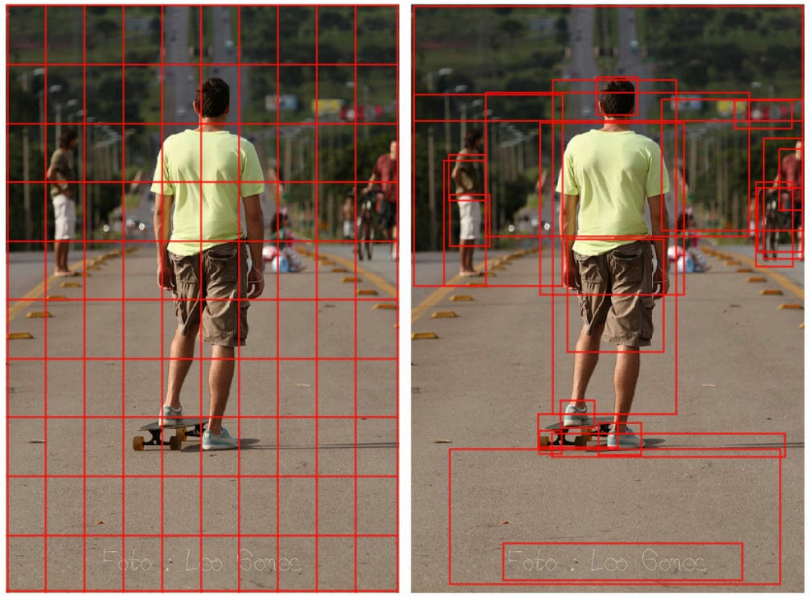


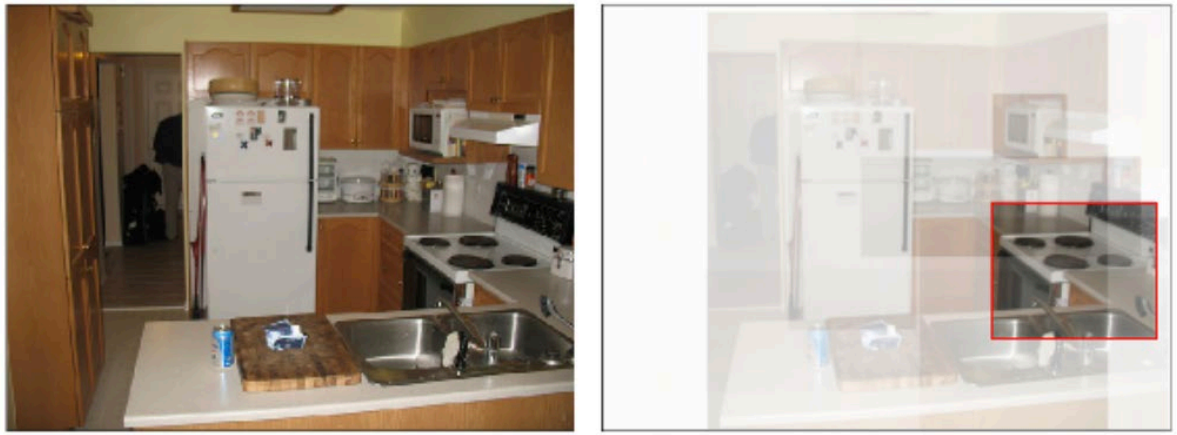


CVPR 2017

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson[1]* Xiaodong He[2] Chris Buehler[3] Damien Teney[4]
Mark Johnson[5] Stephen Gould[1] Lei Zhang[3]
[1]Australian National University [2]JD AI Research [3]Microsoft Research [4]University of Adelaide [5]Macquarie University
[1]firstname.lastname@anu.edu.au, [2]xiaodong.he@jd.com, [3]{chris.buehler,leizhang}@microsoft.com
[4]damien.teney@adelaide.edu.au, [5]mark.johnson@mq.edu.au





Question: What room are they in? Answer: kitchen

VQA Solution 5 years ago:
Learn V and L features separately, and fuse.

?

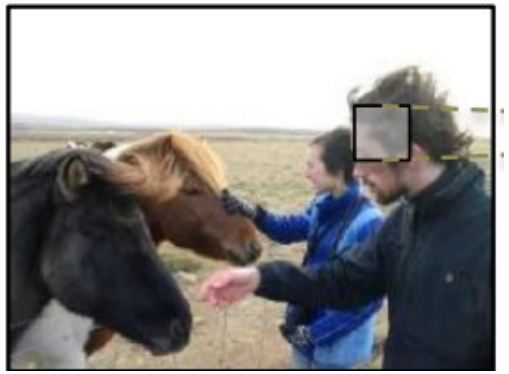
?

?

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

Cross Entroy Loss Across 5000 possible answers

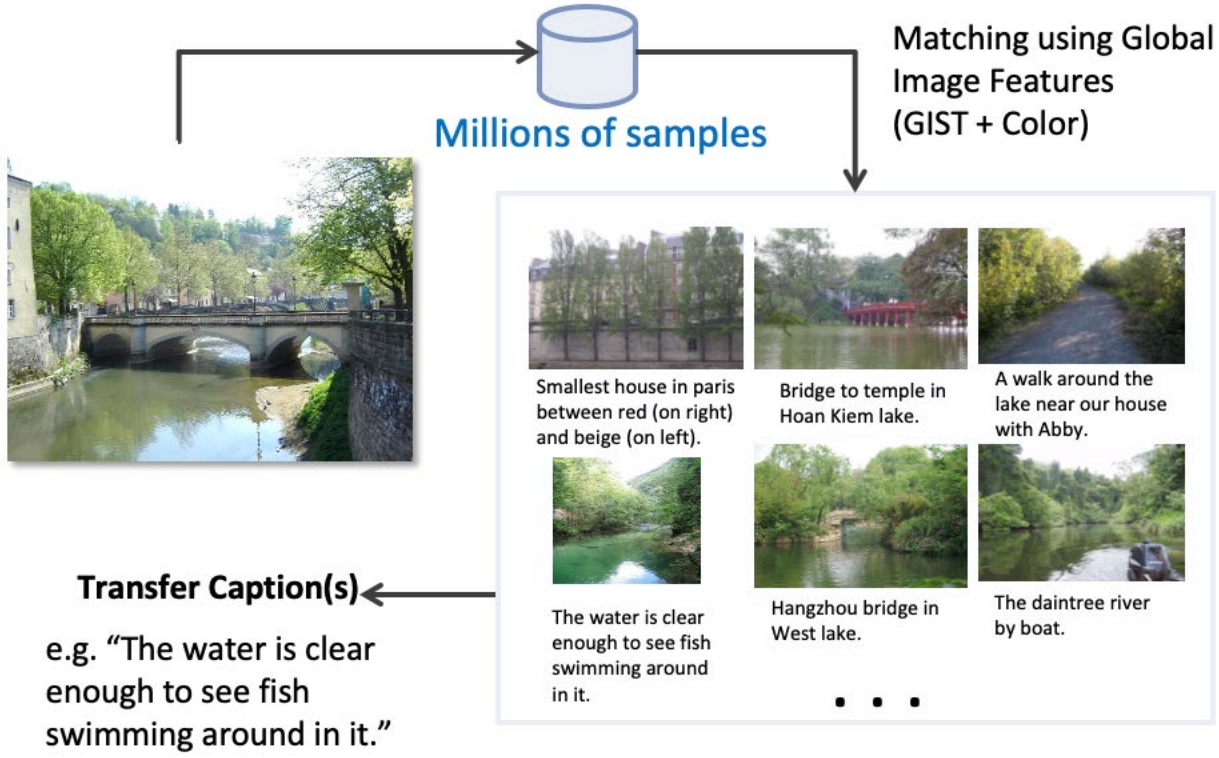
VQA Solution from 2019 … Multimodal Pretraining
(typically using masked language modeling)



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

?

Describing images with language (Image Captioning)



Im2Text: Describing Images Using 1 Million Captioned Photographs
Vicente Ordonez, Girish Kulkarni, Tamara L. Berg.
Advances in Neural Information Processing Systems. **NIPS 2011**. Granada, Spain.

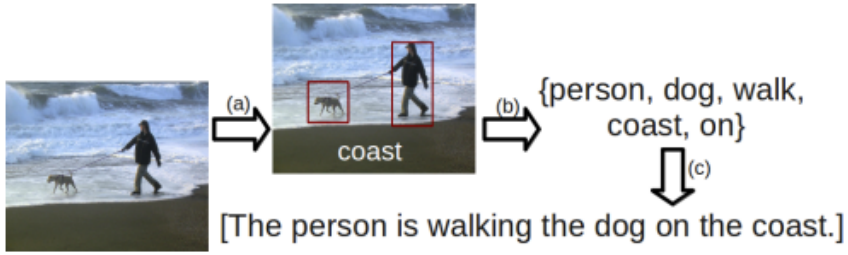


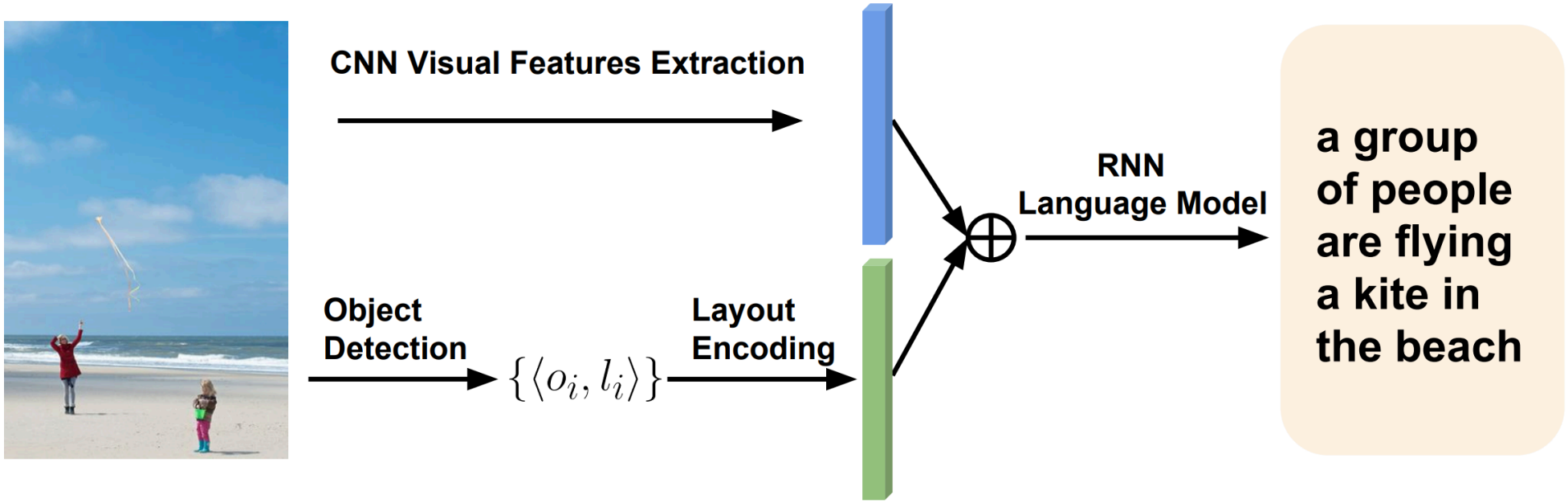
Figure 3: Overview of our approach. (a) Detect objects and scenes from input image. (b) Estimate optimal sentence structure quadruplet $\mathcal{T}^*$. (c) Generating a sentence from $\mathcal{T}^*$.

Corpus-Guided Sentence Generation of Natural Images

EMNLP 2011

Yezhou Yang † and **Ching Lik Teo** † and **Hal Daumé III** and **Yiannis Aloimonos**
University of Maryland Institute for Advanced Computer Studies
College Park, Maryland 20742, USA
{yzyang, cteo, hal, yiannis}@umiacs.umd.edu

One method for image captioning …

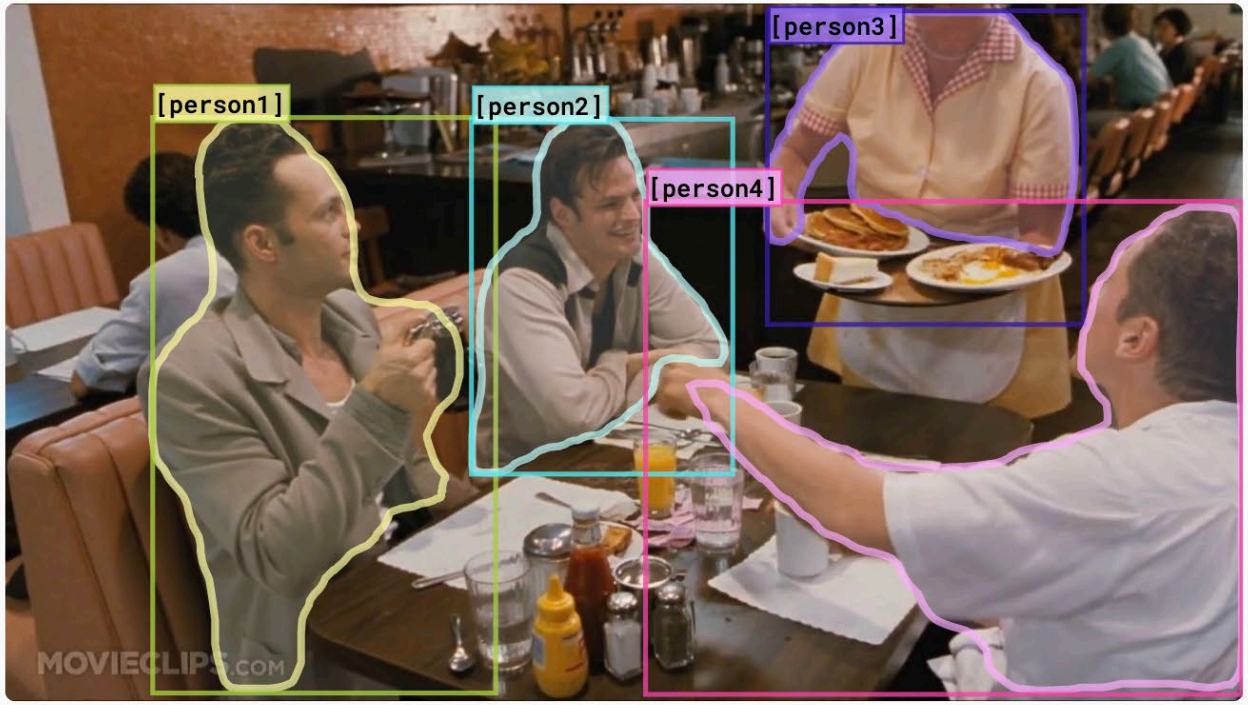


Obj2Text: Generating Visually Descriptive Language from Object Layouts

Xuwang Yin, Vicente Ordonez. Empirical Methods in Natural Language Processing. **EMNLP 2017**. Copenhagen, Denmark. September 2017. [pdf] [arxiv] [code] [bibtex]

(~Oral presentation)

Visual Common Sense Reasoning



hide all | show all | [person1] | [person2] | [person3] | [person4]

more objects »

Why is [person4] pointing at [person1]?

- a) He is telling [person3] that [person1] ordered the pancakes.
- b) He just told a joke.
- c) He is feeling accusatory towards [person1].
- d) He is giving [person1] directions.

Rationale: I think so because...

- a) [person1] has the pancakes in front of him.
- b) [person4] is taking everyone's order and asked for clarification.
- c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
- d) [person3] is delivering food to the table, and she might not know whose order is whose.

From Recognition to Cognition: Visual Commonsense Reasoning

https://visualcommonsense.com/

Enriching Video Captioning with Commonsense Descriptions

Standard Caption
A band is playing at a concert

Generated Commonsense Descriptions
Intention
to entertain the audience

Effect
will get standing ovation

Video2Commonsense Dataset

- Videos of agents doing actions
- Annotations for intentions of agents, effect of actions

Benchmarking Video Captioning

- Existing models found lacking
- Guidance from commonsense knowledge bases required

V2C
VIDEO TO COMMONSENSE

Fang*, Gokhale* Banerjee, Yang, Baral. "Generating Commonsense Descriptions to Enrich Video Captioning" *(EMNLP 2020)*

Video2Commonsense
Enriching Video Captioning with Commonsense Descriptions

Conventional Caption

Group of runners get prepared to run a race.

Commonsense-Enriched Caption

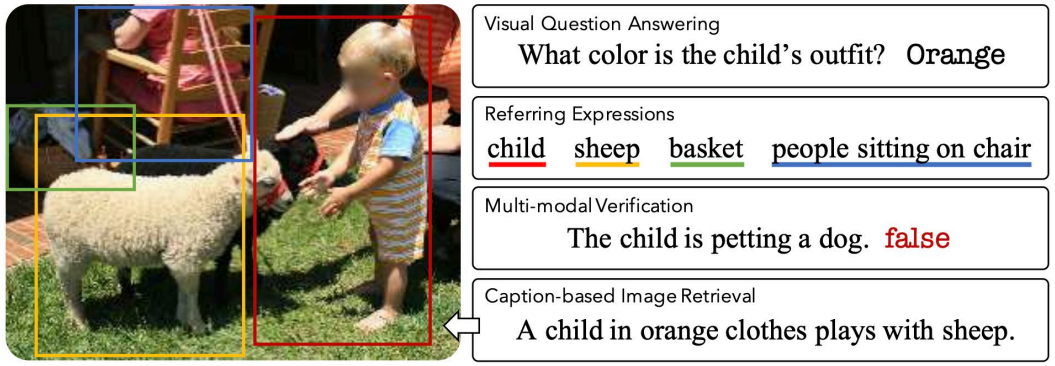
In order to win a medal, a group of runners get prepared to run a race. As a result they are congratulated at the finish line. They are athletic.

Commonsense Question Answering

What happens next to the runners?

{ Are congratulated at the finish line
become tired

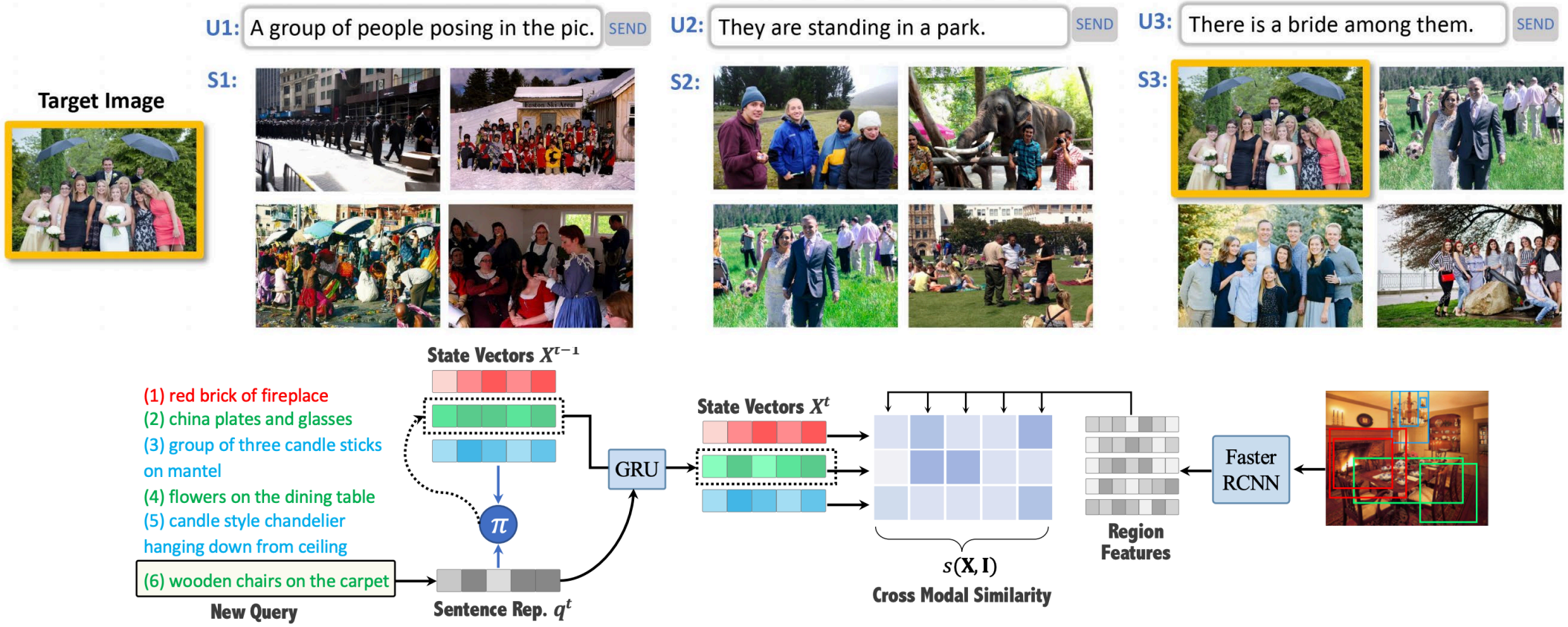
Multi-task Learning / More General Models



Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

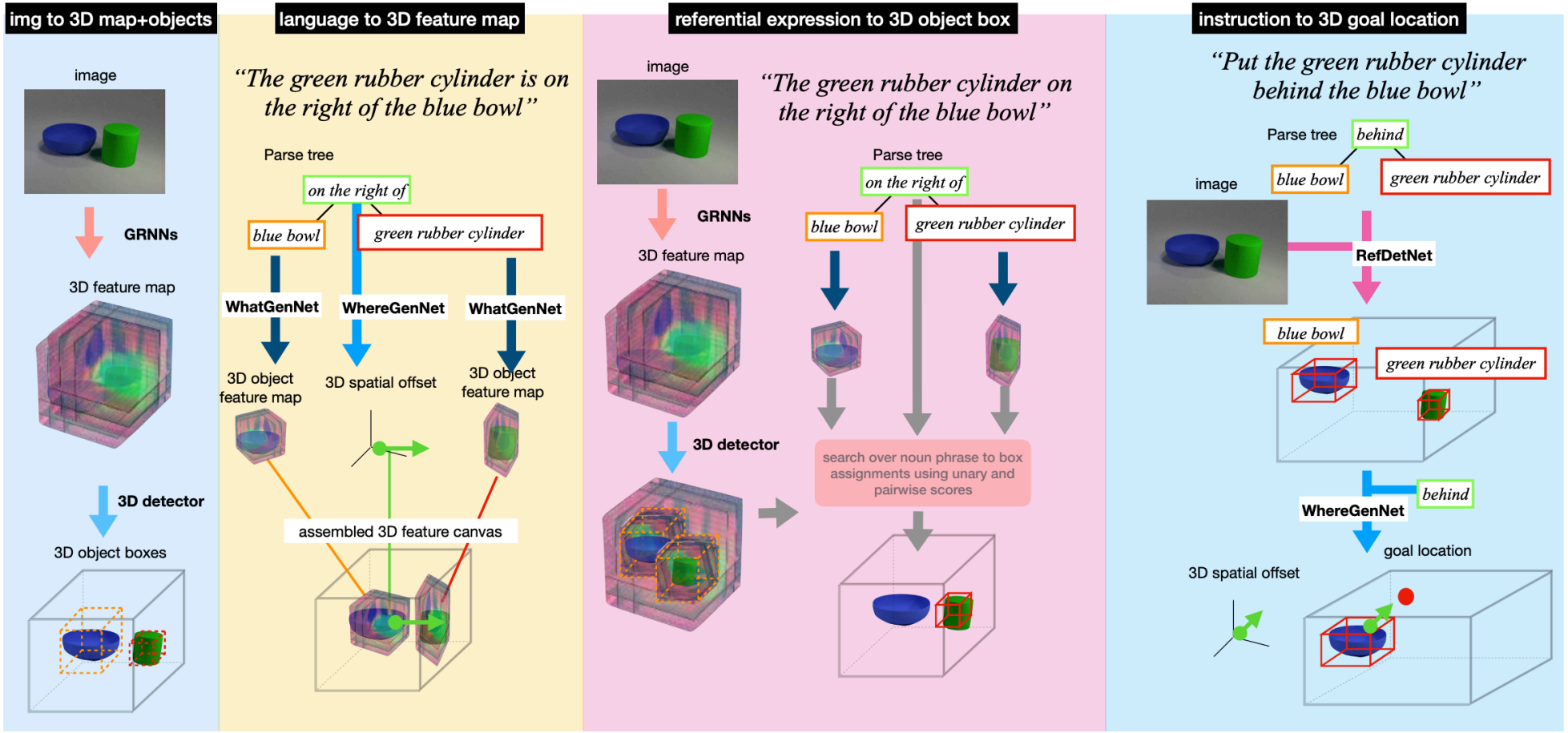
12-in-1: Multi-task Vision and Language Representation Learning
https://arxiv.org/abs/1912.02315
CVPR 2020

Interactivity + Language and Vision



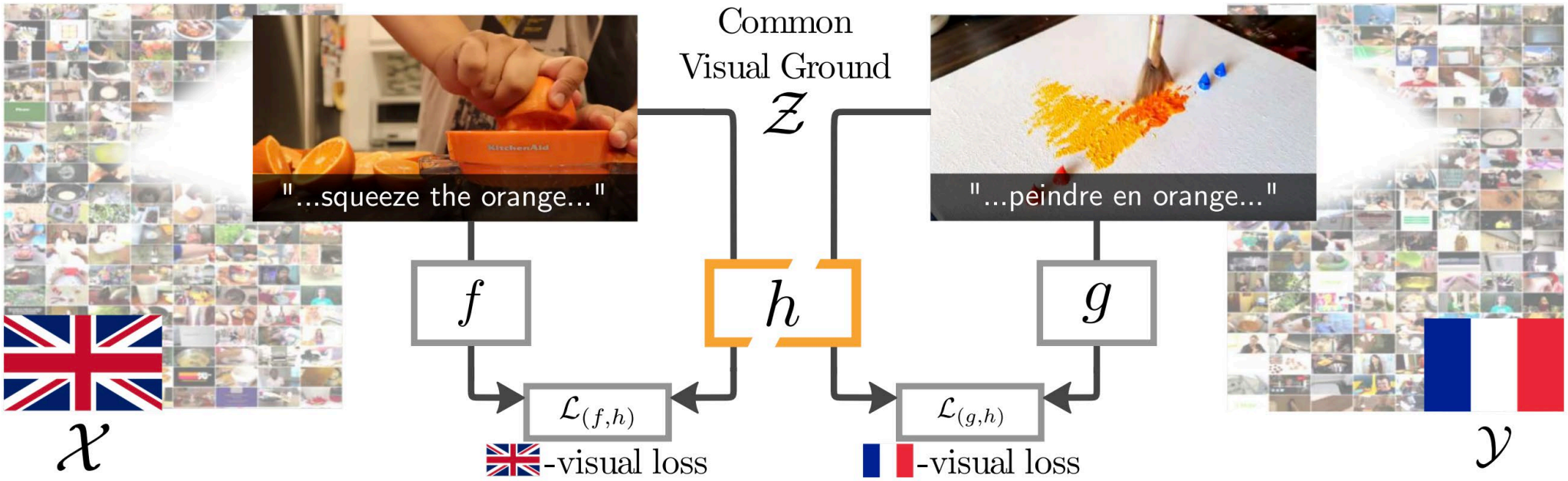
Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queries https://arxiv.org/abs/1911.03826 NeurIPS 2019

Vision + Language + 3D



https://arxiv.org/abs/1910.01210

Multiple Languages and Vision



https://arxiv.org/abs/2003.05078

Video + Language Tasks

00:00:03,576 --> 00:00:05,697
Gavin Mitchell's office.
Rachel Green's office.

00:00:05,870 --> 00:00:07,409
Give me that phone.

00:00:08,873 --> 00:00:12,293
Hello, this is Rachel Green.
How can I help you?

00:00:12,460 --> 00:00:17,629
Uh-huh. Okay, then.
I'll pass you back to your son.

00:00:18,800 --> 00:00:21,639
Hey, Mom. No, that's just my secretary.

(positive) The woman becomes upset when the man answers the phone because he pretends it is his own office.
(negative) The woman becomes upset when the man answers the phone because she is expecting a phone call from her mom.

Inferring reasons

(positive) The woman realizes it is the man's mother who is calling and she passes the phone back to the man.

(negative) The man realizes it is the woman's mother who is calling and he passes the phone back to the woman.

Identifying characters

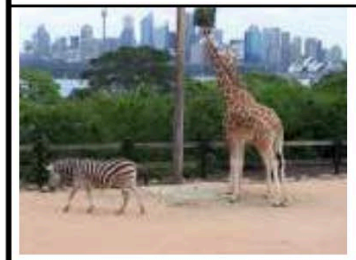
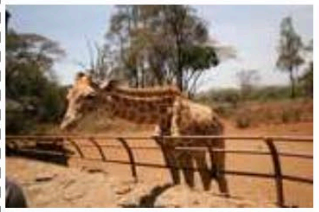
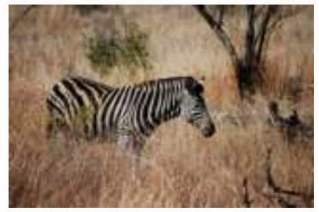
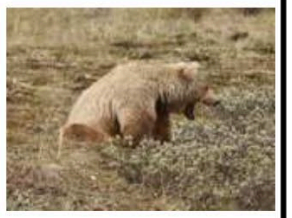
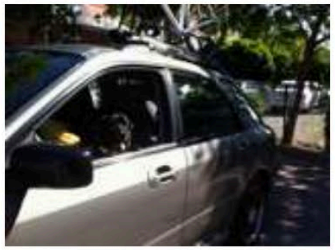
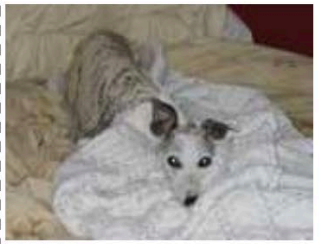
(positive) The phone rings, a man picks it up, and a woman slams her hand on the desk and demands the man give her the phone.

(negative) The two people that the man in the glasses is talking to need to be briefed on something.

Global video understanding

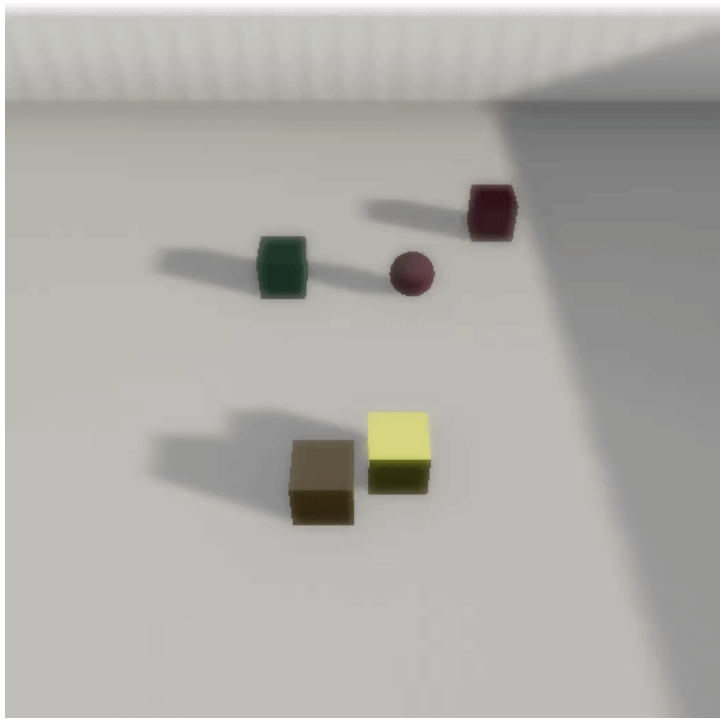
https://openaccess.thecvf.com/content_CVPR_2020/papers/Liu_Violin_A_Large-Scale_Dataset_for_Video-and-Language_Inference_CVPR_2020_paper.pdf

Counterfactuals in Vision and Language

Question Image	Counterfactual Questions	Counterfactual Images
 Is this in Australia?	1. Is the grass green? 2. Is there grass on the ground? 3. Are they standing on a green grass field? 4. Is the stop light green?	  
 What color is the person's helmet?	1. What color jacket is the girl wearing? 2. What color jacket is the person wearing? 3. What color is the jacket? 4. What color is the woman's jacket?	  
 Where did the shadow on the car come from?	1. What kind of dog is this? 2. What type of dog is this? 3. What kind of dog is shown? 4. What is the breed of dog?	  

https://openaccess.thecvf.com/content_CVPR_2020/papers/Abbasnejad_Counterfactual_Vision_and_Language_Learning_CVPR_2020_paper.pdf

Asking Counterfactual Questions to Reason about Physical Properties

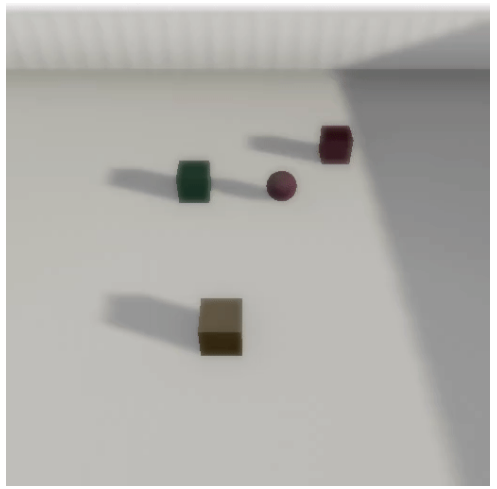


Input Video

Counterfactual Question

What will happen if the yellow cube is **removed** ?

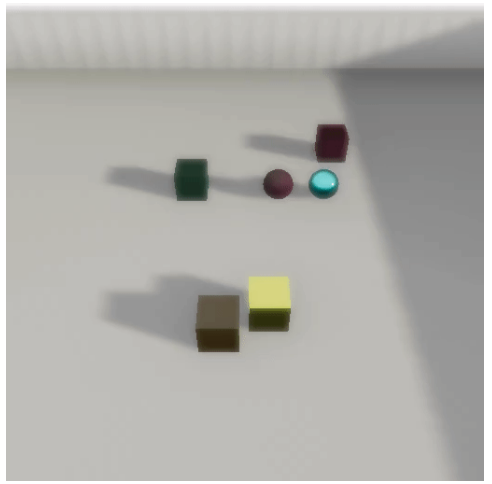
(A) Purple Cube will collide with brown cube



Planning Question

How can the collision between yellow and purple cube be stopped?

(A) **Add** teal sphere to the right of purple sphere



Effect of Action

My work: Perception & Reasoning with Robustness

Robust Visual Reasoning (Visual QA, Video Captioning, V&L Inference)

V&L Robustness: Logical, Semantic, Spatial (use additional knowledge sources and sensors)

Is the fork **NOT** on the plate? — Negation

yes 94.785%
no 5.215%

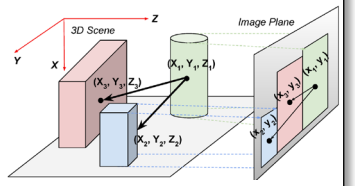
Is the fork on the plate **AND** is the food made of eggs? — Conjunction

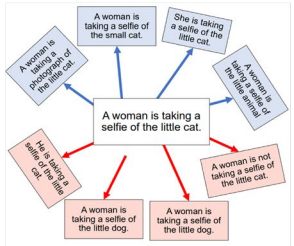
no 97.855%
yes 2.144%

Is the fork on the plate **OR** is the food made of eggs? — Disjunction

no 32.221%
scrambled 17.040%

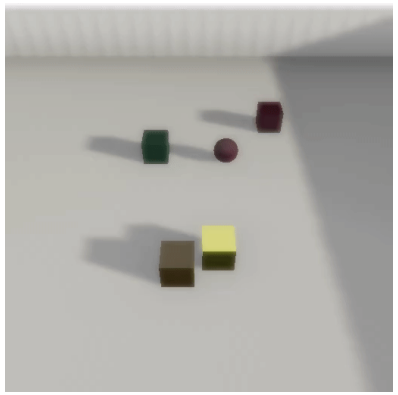
Question	Answer
Is that a giraffe or an elephant?	Giraffe
Who is feeding the giraffe *behind* the man?	Lady
Is there a fence *near* the animal *behind* the man?	Yes
On which side of the image is the man?	*Right*
Is the giraffe *behind* the man?	Yes





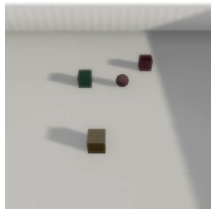
f → $\hat{y}_e$

Understanding Agent Actions in Videos with Commonsense, Counterfactual and Physics-Based Reasoning



Counterfactual Question
What will happen if the yellow cube is **removed** ?

(A) Purple Cube will collide with brown cube



Planning Question
How can the collision between yellow and purple cube be stopped?

(A) **Add** teal sphere to the right of purple sphere

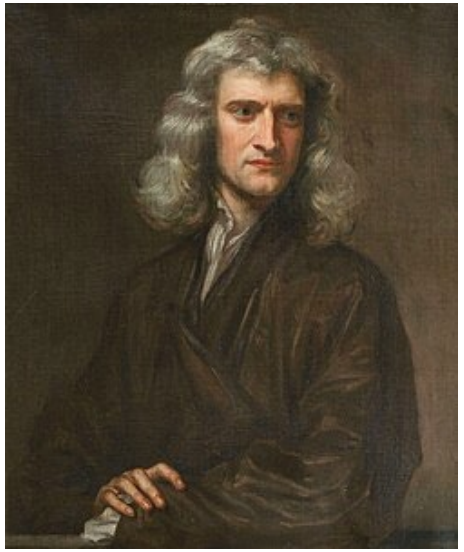
Conventional Caption	Group of runners get prepared to run a race.
Commonsense-Enriched Caption	In order to win a medal, a group of runners get prepared to run a race. As a result they are congratulated at the finish line. They are athletic.
Commonsense Question Answering	What happens next to the runners? { Are congratulated at the finish line / become tired

Gokhale ECCV '20; Gokhale EMNLP'20; Gokhale ACL'21; Fang EMNLP'20; Banerjee ICCV'21; Patel EMNLP'22

All of this research by hundreds of people in hundreds of places …

culminated in **“Vision-Language Models (VLMs)”**
that you have today (in ChatGPT, Gemini, etc.)

These ideas started as research projects led by **PhD students** and with enough academic evidence, led to companies and products.



Isaac Newton remarked in a 1675 letter to his rival Robert Hooke:

“What Des-Cartes did was a good step.
You have added much several ways, & especially in taking the colours of thin plates into philosophical consideration.

If I have seen further,
it is by standing on the shoulders of Giants. ”