DeepRob Discussion 0

1/13/2026

Session Objective

- Overview of ML and the path forward
- Overview of the tools and how they relate
- Overview of the autograder
- Solve a couple cases of P0 together
- Everyone leaves with at least 10 points of P0

Learning Objective

- Describe general structure of a machine learning models and how it relates to robot vision tasks.
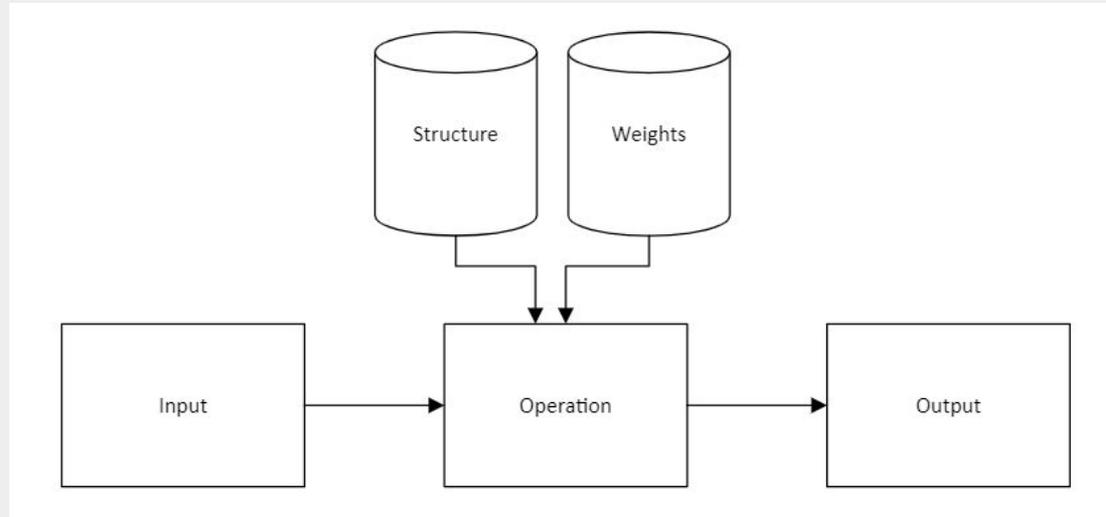
AI/ML Background

Input/Output Formats

- Text
- Numbers
- Images
- Video
- Multi-Modal

All are converted to numbers to feed into model

Operation Simple Case

$$(Input * 5.3 + 3.6) = Output$$

Structure:
(Input * p1 + p2) = Output

Weights/Parameters:
P1 = 5.3, P2 = 3.6

(Slightly less) Simple Case

(Input * 5.3 + 3.6) * 6.4 + 3.9 = Output

Structure:
(Input * P1 + P2) * P3 + P4 = Output

If f: Input * x + y = Output
f(f(Input)) = Output

However X's and Y's are different for each layer!

Weights/Parameters:
P1 = 5.3, P2 = 3.6, P3 = 6.4, P4 = 3.9
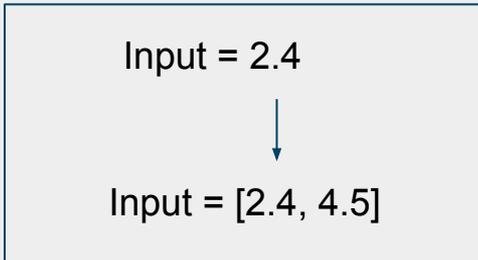
(Slightly, slightly less) Simple Case

(Input * 5.3 + 3.6) * 6.4 + 3.9 = Output

Structure:
(Input * P1 + P2) * P3 + P4 = Output

Input = 2.4

Input = [2.4, 4.5]

Weights:
P1 = 5.3, P2 = 3.6, P3 = 6.4, P4 = 3.9

Weights (Matrix):
P1 = [5.6, 5.3], P2 = [p21, p22], …

(Slightly, slightly less) Simple Case

(Input * 5.3 + 3.6) * 6.4 + 3.9 = Output

Structure:
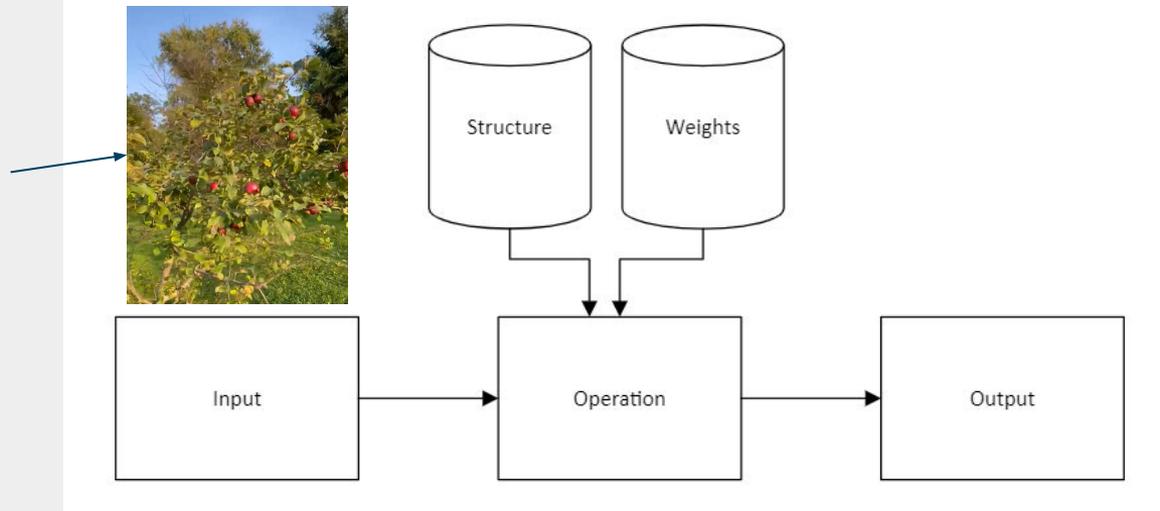(Input * P1 + P2) * P3 + P4 = Output

This operation is orders of magnitude
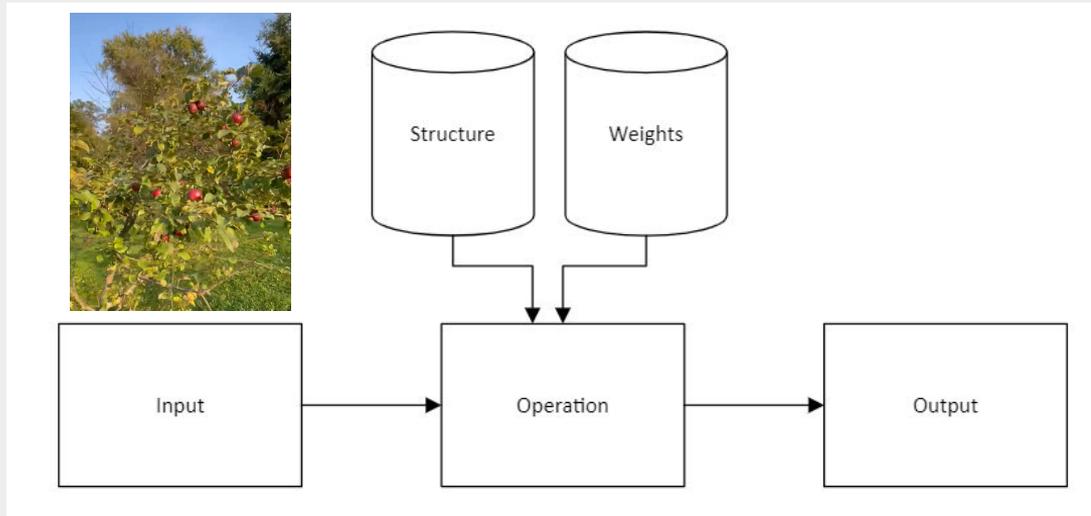more efficient on GPUs

Bringing It Back, Input and Output

- Input for vision models are images and video
- An image is represented by a matrix of pixels
- A video is a dimensional extension of images

Bringing It Back, Input and Output

- Output is determined by the structure of the function

- Structure of the function is chosen to facilitate a task

What vision tasks can we design for? How does it relate to output?

Input

Output Formats

Text	Numbers	Images	Video*
"apple, tree, leaf, outdoors" Example: CLiP	apple is located at [x,y] and is [w,z] in size Examples: YOLO, PoseCNN	Pixel segmentation, color correction Example: YOLO-Seg, Affinity-LCFCN	Video segmentation Example: SAM2

Potential Vision Outputs: Text

Youtube-BB

airplane, person (89.0%) Ranked 1 out of 23 labels
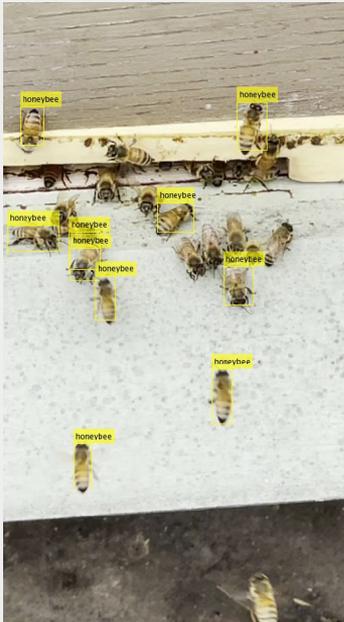
✓ a photo of a **airplane**.
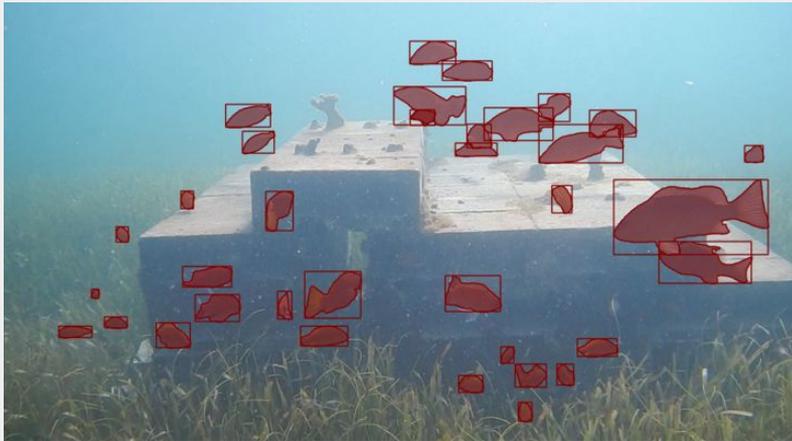
✗ a photo of a **bird**.

✗ a photo of a **bear**.

✗ a photo of a **giraffe**.

✗ a photo of a **car**.

Potential Vision Outputs: Numbers

Potential Vision Outputs: Images (Masks)

Back to Our Learning Objective

- Describe general structure of a machine learning models and how it relates to robot vision tasks.

- Machine learning paradigms are composed of inputs, outputs, structures and weights
- Inputs for machine vision are primarily images and video
- Structure is determined by task to be accomplished
- Weights are tuned via training
- Outputs are determined by structure: example outputs shown included recognition, detection, classification, and segmentation
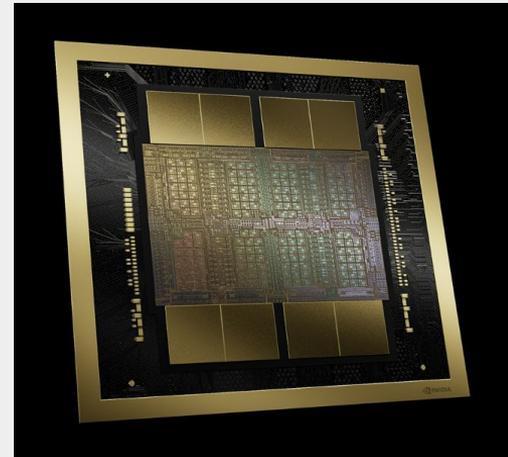
Demo - SAM3 Playground

Learning Objective

- Identify core tools used in this class
    - Hardware
    - Programming Languages
    - Libraries
    - Environments
    - Course Tools

Hardware

- GPUs
  - Specifically CUDA-enabled
- What are some of the advantages?
- What are some of the disadvantages?
- Specific to class?

Programing Languages

- Python!
- What are some of the advantages?
- What are some of the disadvantages?
- Specific to class?

Libraries

- PyTorch ecosystem
- What are some of the advantages?
- What are some of the disadvantages?
- Specific to class?

Environments

- Colab vs. Local
- What are some of the advantages?
- What are some of the disadvantages?
- Specific to class?

Learning Objective

- Identify core tools used in this class

    - Hardware - CUDA hardware
    - Programming Languages - Python
    - Libraries - PyTorch
    - Environments - Colab/VS Code

Quick break!

Learning Objective - Class Tools

- Class Specific Tools
  - Autograder
    - Demo Later
    - Nuances of submission
  - Piazza