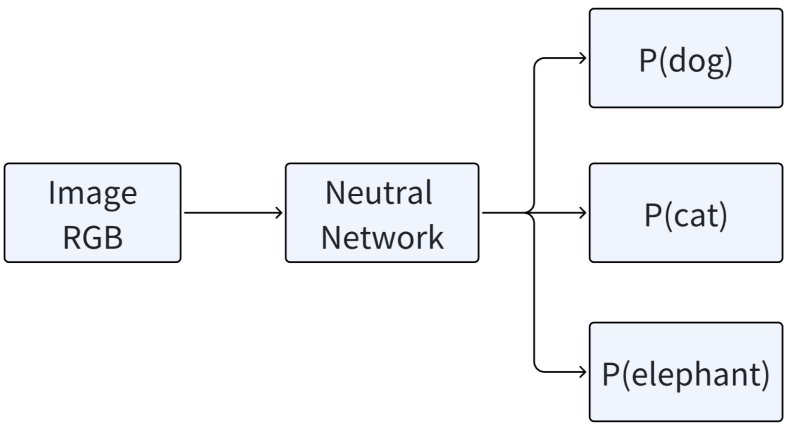
# 16. Deep Perception part 1

Perception with geometry done, even antipodal grasping unknown object case. Which requires cameras all around. You need to see both sides of objects

## What is Deep Perception?

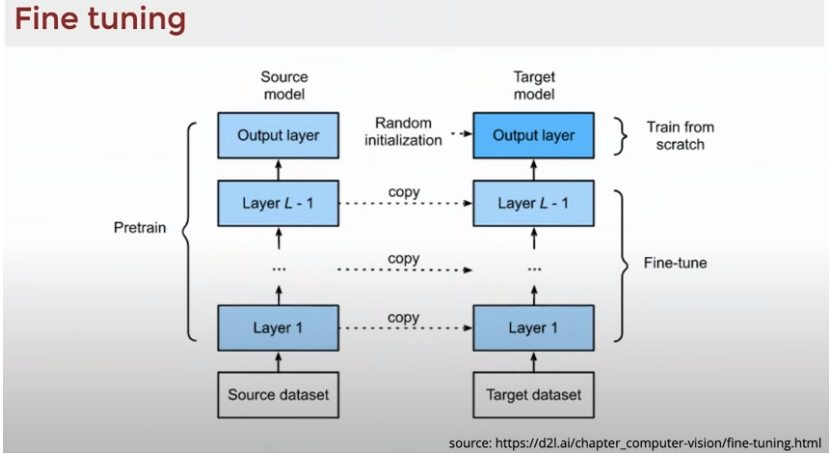
## Today's Topic outline

- Limitations of Using Geometry only:
  - No understanding of what an object is
  - Double picks; Might pick up a heavy object from one corner
  - Partial views; Depth returns don't work for transparent objects
  - Some tasks require object recognition! "pick up the mustard bottle"
  - Deep learning comes in, and works increbly well
- Basics of Computer Vision
  -




- level 1: Image recognition
  - level 2: object detection (tight bounding box of the location of sheep in the image)
  - Level 3: semantic segmentation
  - Level 4: instance segmentation
  - Segmentation + ICP very powerful;
  - MS COCO data set, image net

- Transfer learning: pretrainging on ImageNet/Coco makes it easier to 'learning' to recognize other objects.
- Fine tuning




- Option 1, pay a start up company to generate last set of segmentation data. Business: generating the last set of labeled image data at segmentation level
  - Option 2, label the data. Mentioned in the ICP, labelFusion, a pipeline for generating GT labels for real rgbd data of cluttered scenes. Human to click, 3points on model,
  - Option 3, generate images in simulation high resolution rendering + domain randomization
- Topic 3

## 1. Mask-RCNN, extremly good

- Architecture for instance segmentation, ex. Mask RCNN
- CNN, object classification -> detection(sliding window), returning hightest possbiliby box
- R-CNN: Regions with CNN features
- Faster R-CNN, adds a region proposal network
- Mask R-CNN -> Faster R-CNN + FCN
- FCN, fully convolutional Net
- The most important thing is architecture. They are carefully designed and simple
  - Pixelwisa labels after faster RCNN

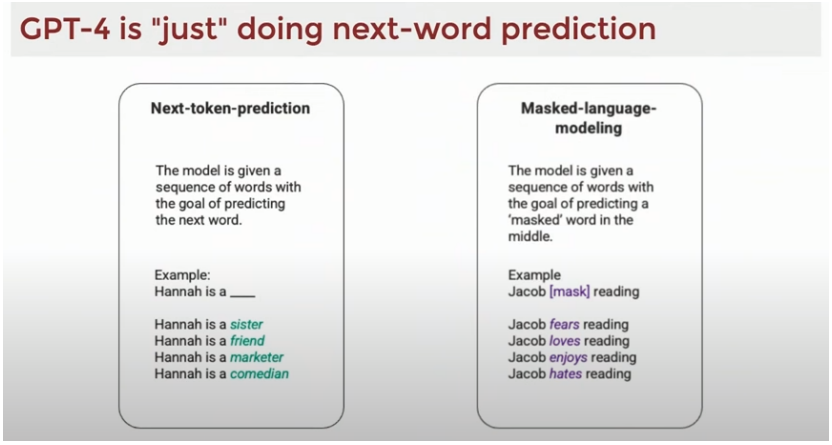
## 2. The input and output of Mask RCNN in drake

- Input: the data set, is the image from COCO,
- Output: The output label you have to give it are

        - The bounding box for every instance
        - The label for every instance
        - The segmentation for every instance
    - Segmentation + ICP -> model-based grasp selection
    - Segmentation -> antipodal grasp selection

## 3. Self-supervised learning

- Example: text completion, no labeling
- GPT-4 is 'just' doing next-word prediction, next token prediction
- For images, going through the pixels of an image and just trying to predict the next pixel in the image
- It turns out ***Masked-language-modeling*** is better than ***next token prediction*** in images

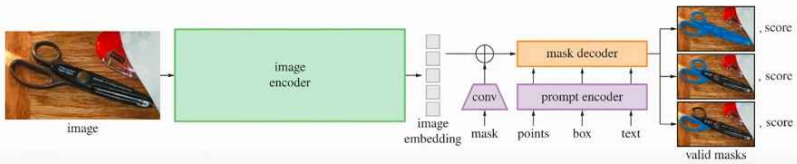



- Masked Auto-Encoders, DIDOv2: A Self-supervised Vision Transformer Model

## 4. Fundation models

- CLIP: Contrastive Language-Image Pre Training
- Segment Anything(Meta)
    - Always try to get the right segmentation out, not only
    - Have model that can work sort of zero-shot with no training on new data
    - But also models that replace fune-tuning with ***prompt Engineering***
    - First one worked on robot data
    - Try the online demos, works extreamly good: https://segment-anything.com/demo

# Segment Anything




Open-source release doesn't accept text. You need a wrapper... e.g. Grounded Segment Anything