CS195: Computer Vision

Image Segmentation

Foundation Models

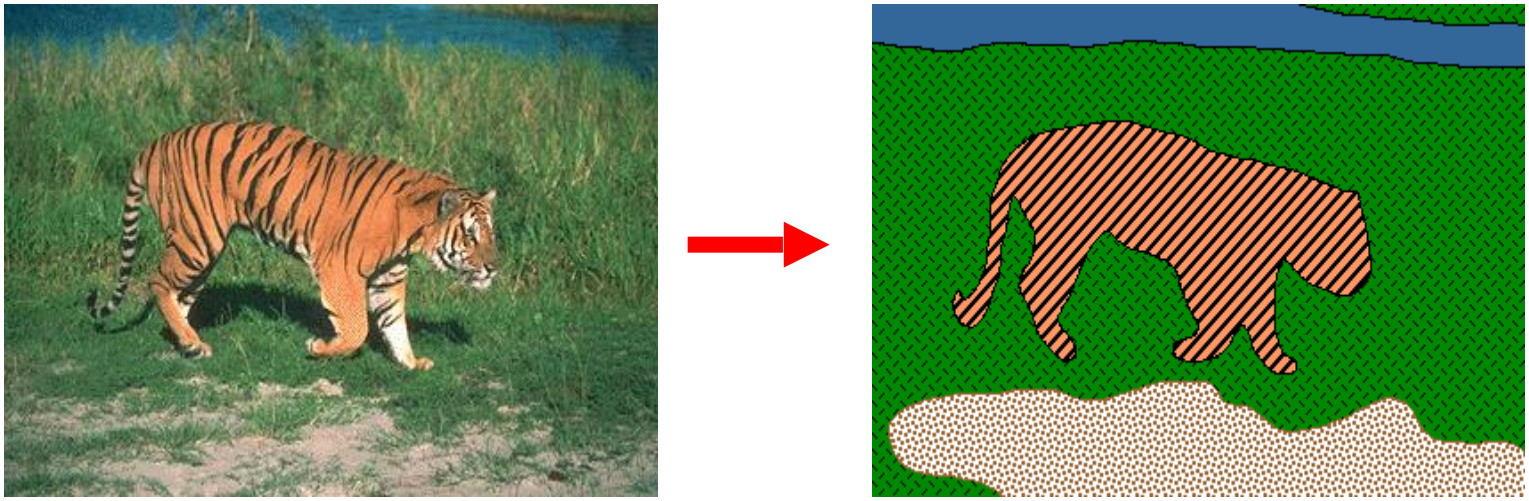
Segmentations using a Foundation Model

Monday, October 21st, 2024

Drake UNIVERSITY

Recap: Image segmentation

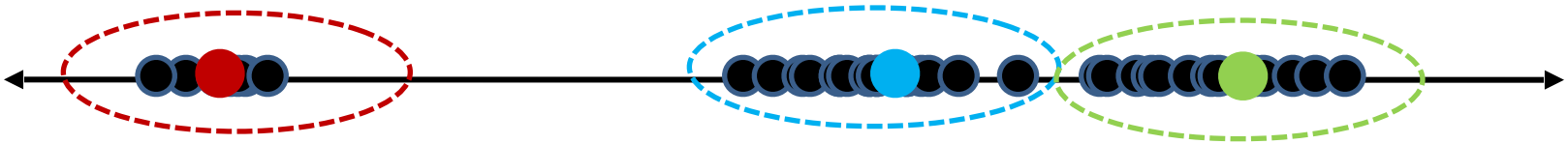
- Goal: break apart an image into simpler components



Classical: Clustering Based Segmentation

Recap: K-means clustering

1. Randomly initialize the cluster centers, c_1, ..., c_K
2. Given cluster centers, determine points in each cluster
3. Given points in each cluster, solve for c_i



If we run it for enough iterations, it will converge to this state, when the centers won't be changing anymore. Clustering mechanism stops.

Recap: Segmentation as clustering on a grayscale image

We can customize the clustering by changing the feature space.

e.g. Grouping pixels based on intensity similarity

x-axis denoting different grayscale values

Feature space: 1D gray-scale value

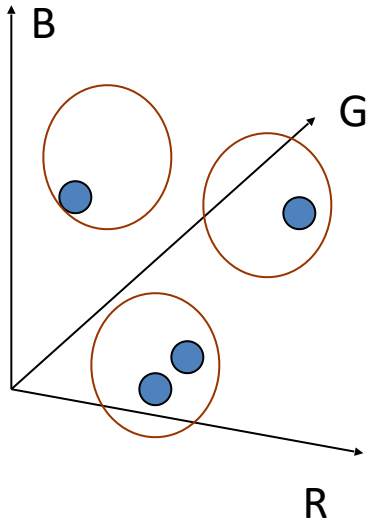
Recap: Segmentation as clustering on a grayscale image

K=2

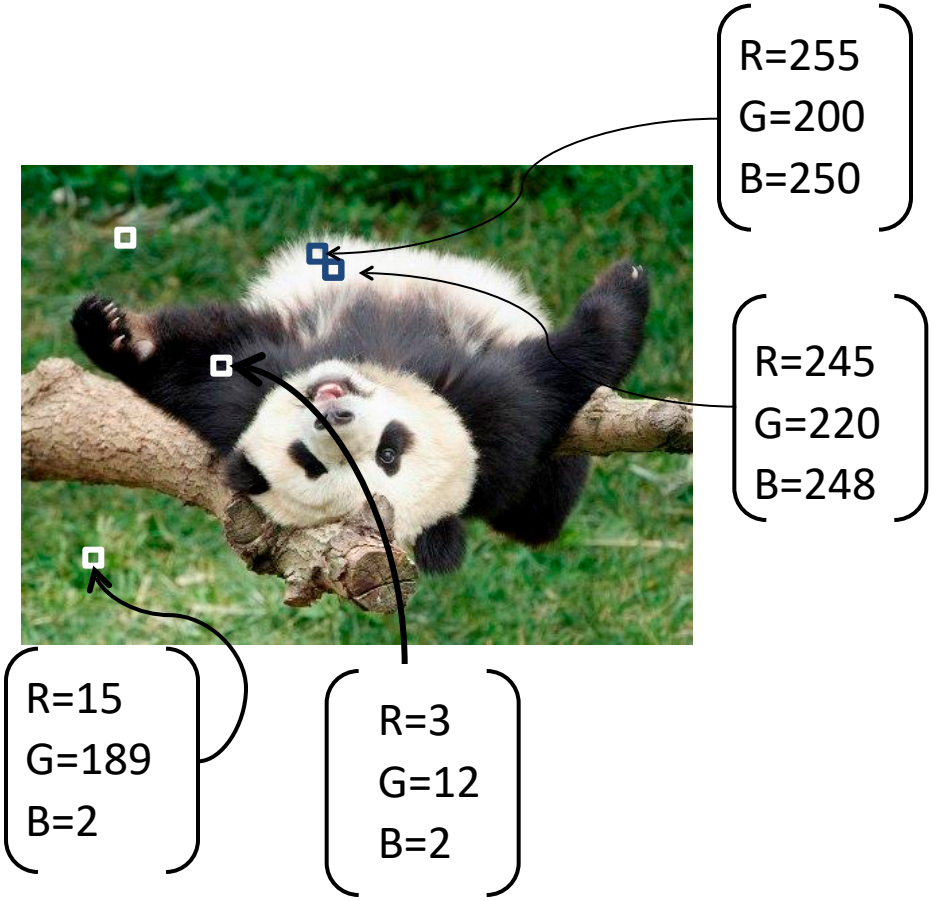
K=3

Recap: Segmentation as clustering on a color image

e.g. Grouping pixels based on color similarity



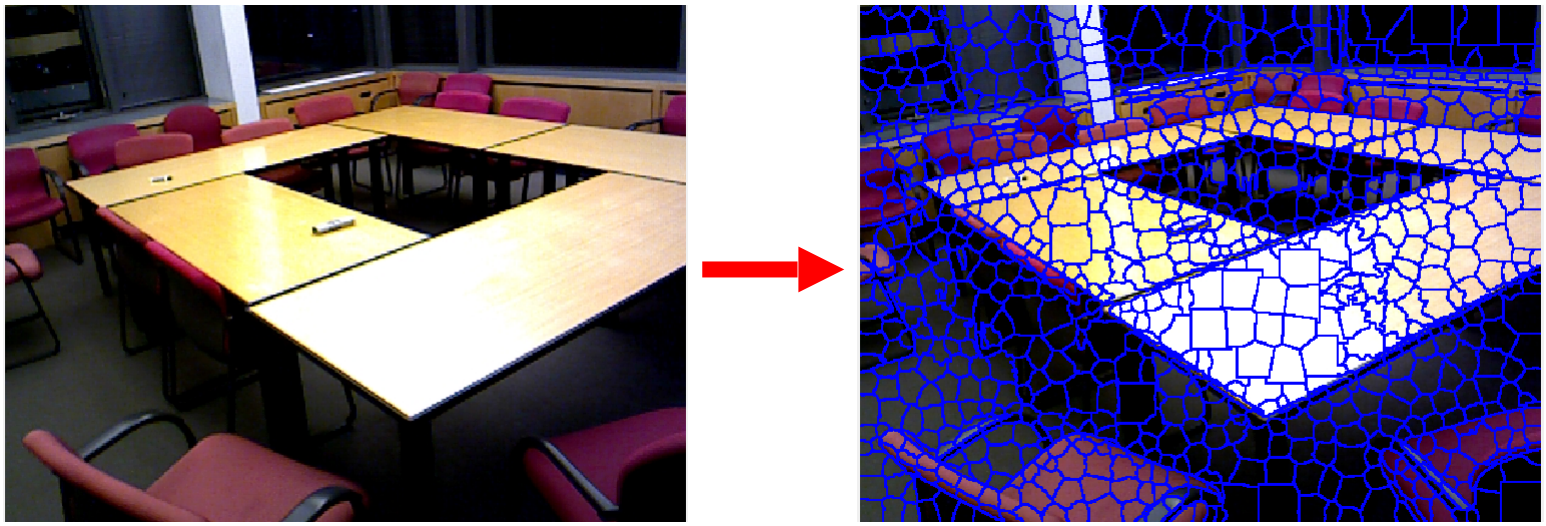
x-axis, y-axis, and z-axis denoting different color values R, G, B



Feature space: color value (3D)

Recap: Simple Linear Iterative Clustering (SLIC)

- SLIC clusters pixels in the combined five-dimensional color and image plane space to generate compact, nearly uniform segments (also called superpixels)
- Very fast, can generate superpixels in less than a second.

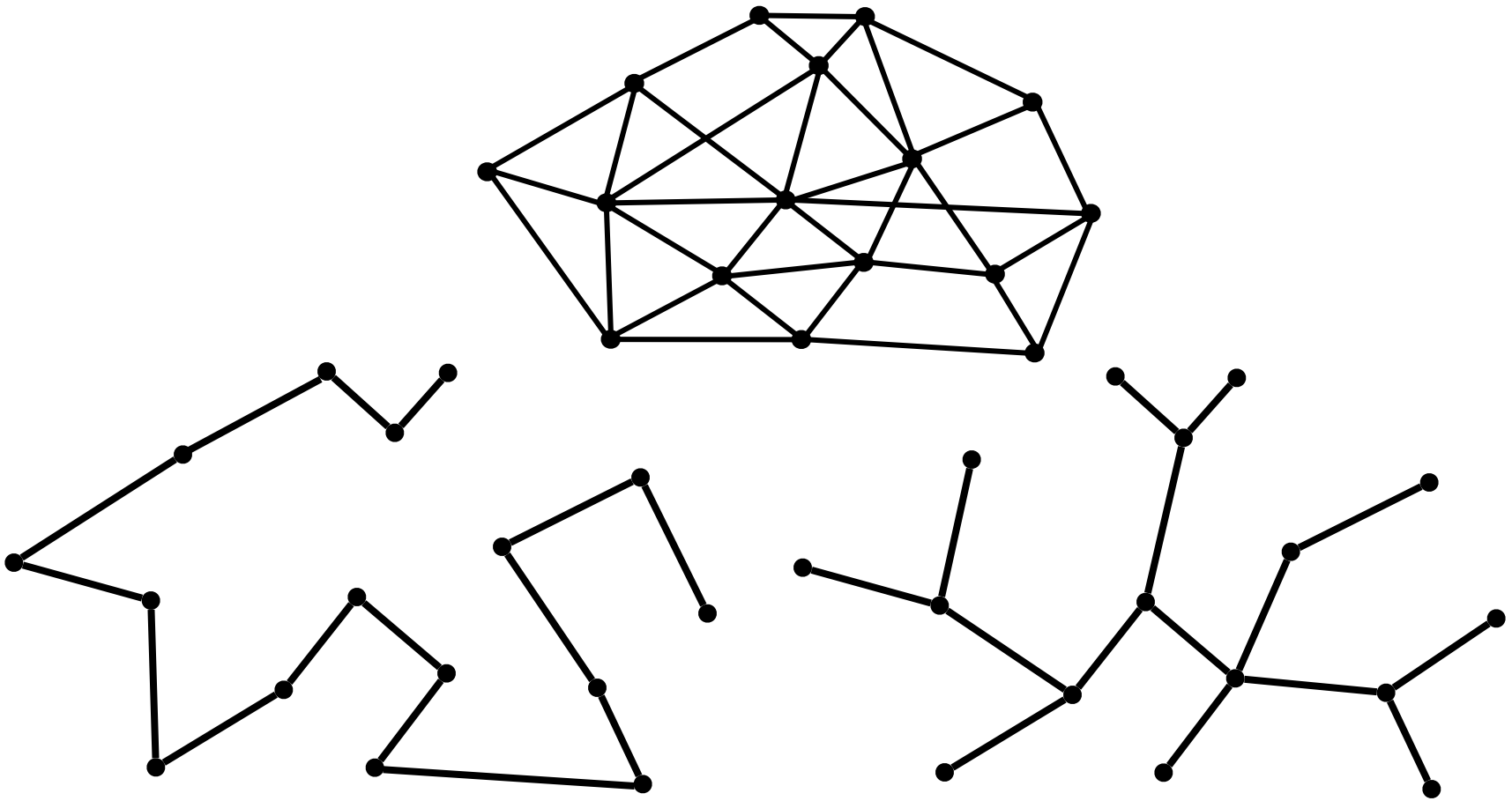


RGB

SLIC superpixels

Classical: Graph Based Segmentation

Recap: View image as a graph

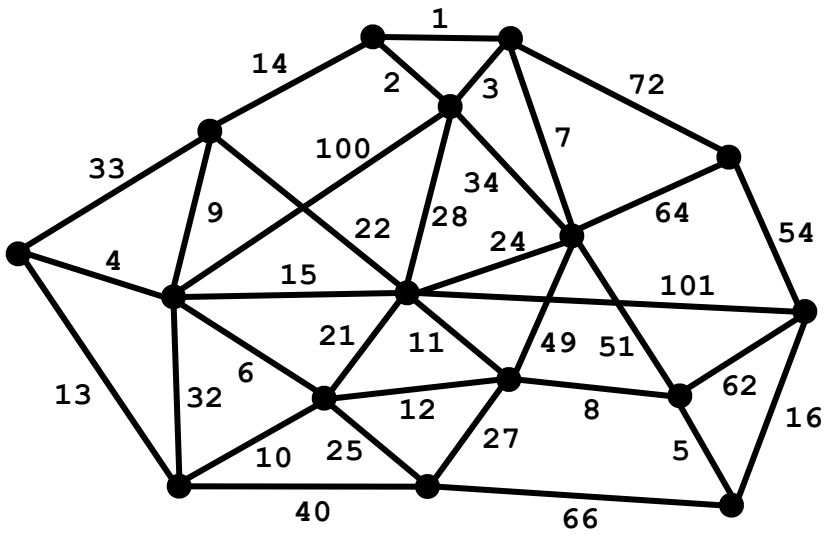


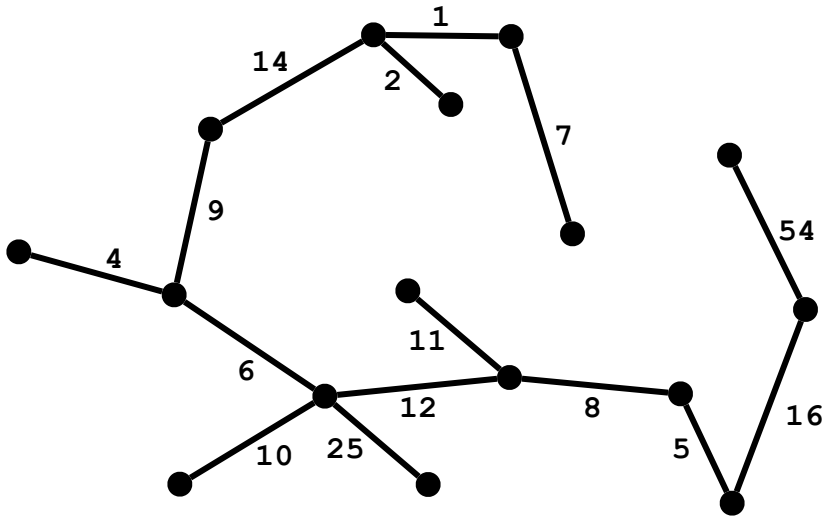
Spanning tree 1

Spanning tree 2

Recap: Minimum Spanning Trees

- Suppose edges are weighted, and we want a spanning tree of *minimum cost* (sum of edge weights)





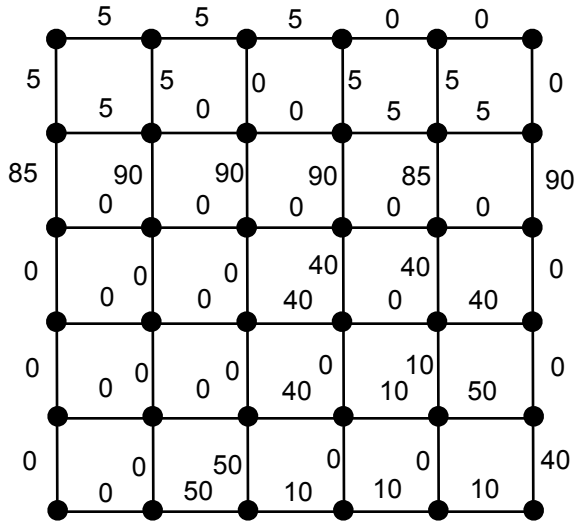
Recap: Back to image segmentation…

- Goal: reduce an image to a small number of homogeneous regions (“segments”)

Recap: Segmentation as a graph problem

- Represent an image as a graph
  - Vertices represent image pixels
  - Edges between adjacent pixels
  - Edge weights give difference in color between pixels

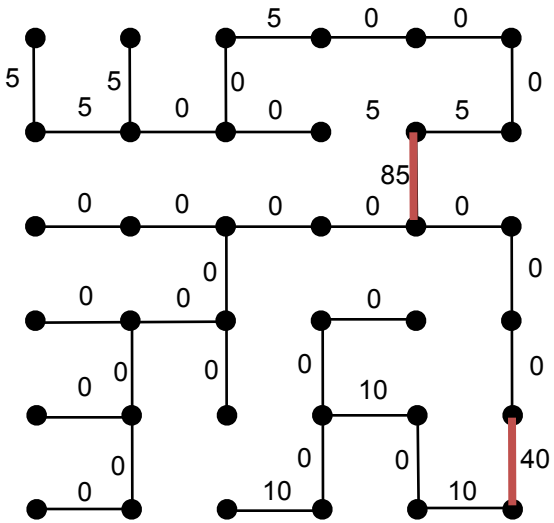
10	15	10	15	10	10
15	10	10	10	15	10
100	100	100	100	100	100
100	100	100	60	60	100
100	100	100	60	50	100
100	100	50	60	50	60



Recap: Segmentation as a graph problem

- Goal: Find a small number of homogeneous regions
  - Or: Find a set of connected components in the graph, such that the sum of edge weights in each component is low
  - We can do this by finding a minimum spanning tree, and then removing a few high-weight edges

10	15	10	15	10	10
15	10	10	10	15	10
100	100	100	100	100	100
100	100	100	60	60	100
100	100	100	60	50	100
100	100	50	60	50	60



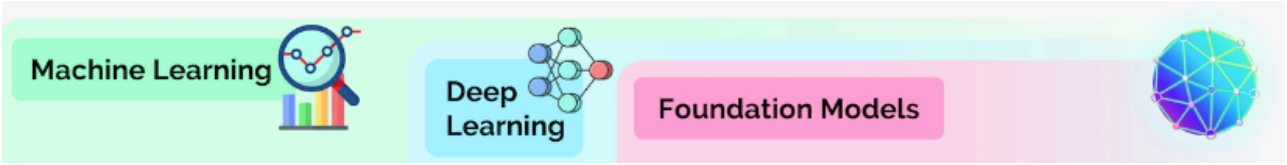
Deep Neural Network-Based Segmentation

Foundation Models

- A foundation model is any model that is trained on broad data (generally using self-supervision + very large scale) that can be adapted (eg, fine-tuned) to a wide range of downstream tasks
    - BERT, GPT, CLIP

- Foundation models term was used to fill a void in describing the paradigm shift we are witnessing:
    - Other names are large language models (LLMs), large multimodal models (LMMs)

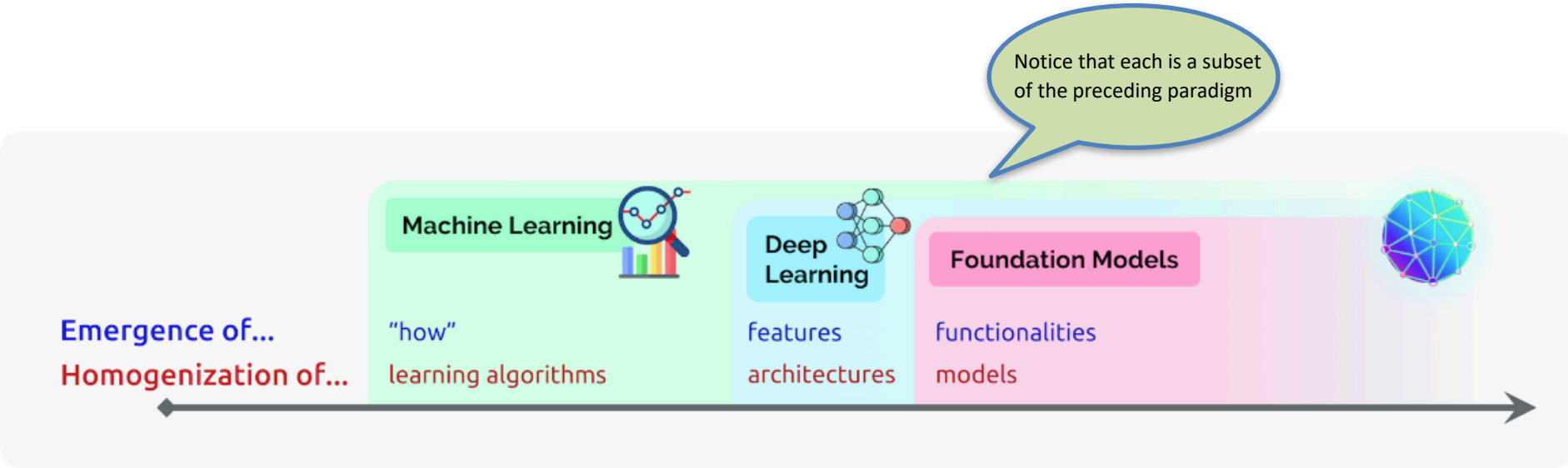
- Foundation models are not new — they are based on deep neural networks and self-supervised learning, both of which have existed for decades. What's powerful about foundation model is the fact despite not being trained explicitly to do many of those tasks



On the Opportunities and Risks of Foundation Models. Bommasani, R et al. 2021

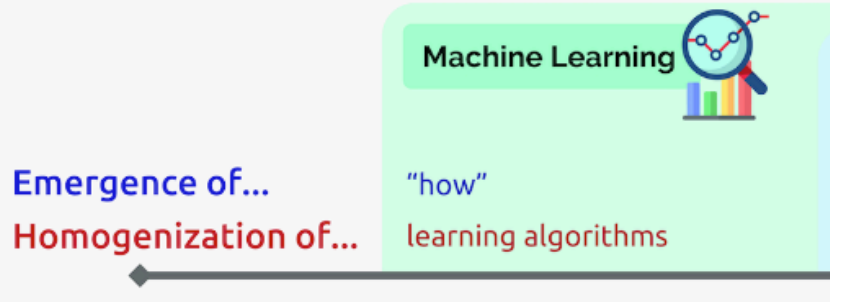
Foundation Models

- Foundation model is also a deep neural network based model
- Its significance can be summarized by two words:
  - Emergence: the behavior of a system which is *implicitly induced* rather than explicitly constructed
  - Homogenization: the *consolidation of methodologies* for building machine learning systems *across a wide range of applications*



On the Opportunities and Risks of Foundation Models. Bommasani, R et al. 2021

Machine Learning (1990-2010)



- **Machine Learning:** The rise of ML started in 1990s representing a marked shift from the way AI systems were built previously. Let's explain it in terms of emergence and homogenization

  - **Emergence**: rather than *how to solve a task (process)*, a learning algorithm would implicitly induce "**how**" based on the data. The *how (process)* **emerges** from the learning algorithm. Various machine learning methods were proposed each capable of inducing the "how"
    - Logistic regression method
    - Ensemble learning method
    - Kernel method

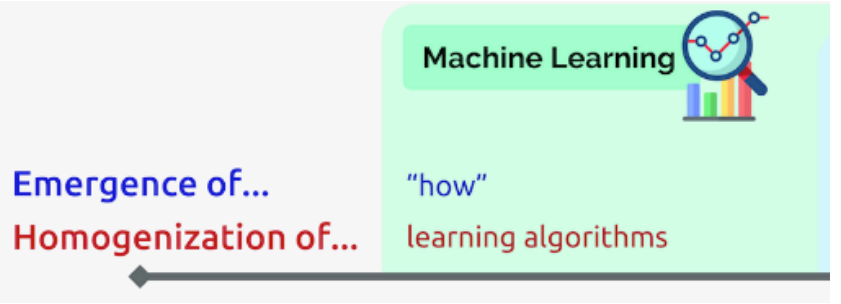
You don't have to explicitly tell how to solve each problem; it will emerge via learning algorithm

Homogenization is the consolidation of **learning algorithms**

  - **Homogenization**: wide range of applications could now be powered by a single generic **learning algorithm**. A single machine learning algorithm can solve a lot of applications.
    - Housing price prediction task: fit **a logistic regression** to the household data
    - Weather prediction task: fit **a logistic regression** to the weather data
    - …

On the Opportunities and Risks of Foundation Models. Bommasani, R et al. 2021

Machine Learning (1990-2010)

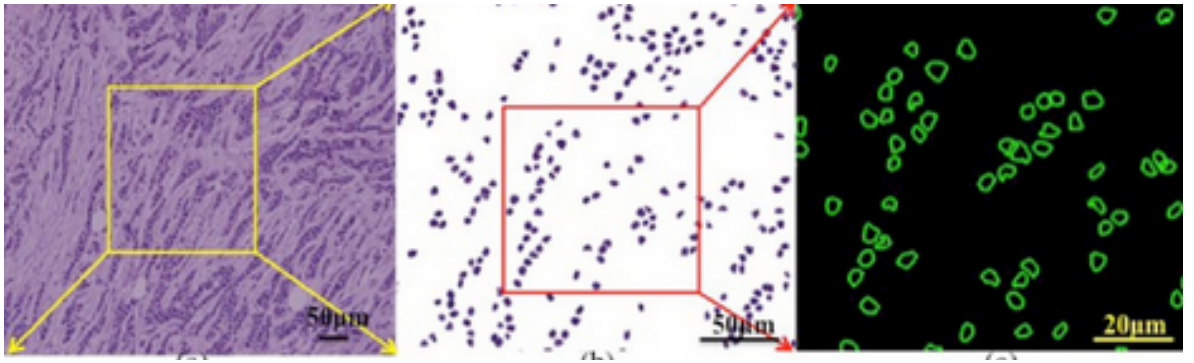


- **Machine Learning:** could not cope up with the complex task in computer vision (CV) and natural language processing (NLP). Those complex task still would require some domain expertise to do feature engineering by the humans
  - Convert raw data eg, image into features/attributes
    - Geometrical shape features
    - SIFT feature
    - HOG feature
    - Wavelet feature

Color-based features
- Hue
- Saturation
- Intensity

Shape-based features
- Area
- Perimeter
- Circularity
- Aspect ratio
- Distance transform
- Line sweep

Architectural features
- Distance transform (global)
- Line sweep (global)



Mark Zarella, David Breen, Alimoor Reza, Aladin Milutinovic, and Fernando Garcia. Lymph Node Metastasis Status in Breast Carcinoma Can Be Predicted via Image Analysis of Tumor Histology - Journal of Analytical Quantiative Cytopathology and Histopathology'2015

Deep Learning (2010-2018)

Computer vision popularized Deep learning

- **Deep Learning:** Around 2010, represents another shift from the way AI systems were built previously. Let's explain it in terms of emergence and homogenization

  - **Emergence**: rather than *feature engineering*, deep learning with deep neural network would implicitly induce good features based on the raw input data. The higher-level **features emerges** through the training process. Various types of deep neural networks (DNNs) were proposed each capable of automatically discovering good "high-level features"
    - Convolutional Neural Networks (CNN): **ResNet, AlexNet**
    - Long Short-Term Memory Networks
    - Graph Neural Networks

You don't have to explicitly tell predictors/features; it will emerge inside the neural network

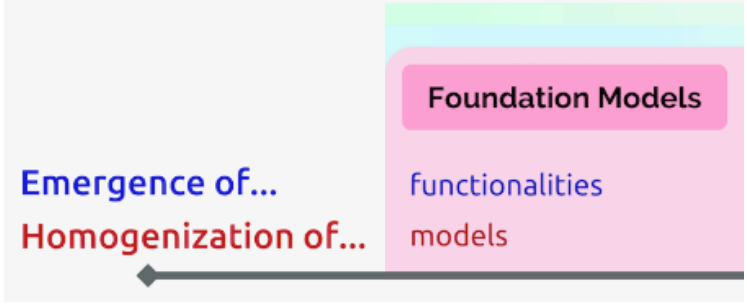
Homogenization is the consolidation of **network architectures**

  - **Homogenization**: rather than feature engineering pipeline, for example a single **ResNet** deep neural network **architecture** can solve a wide range of applications
    - Face recognition for employee authentication: train **a ResNet** to the employee images
    - Plant subspecies recognition: train **a ResNet** to the images of different subspecies of plants
    - …

On the Opportunities and Risks of Foundation Models. Bommasani, R et al. 2021

Foundation Models (2018-present)

NLP popularized Foundation Models

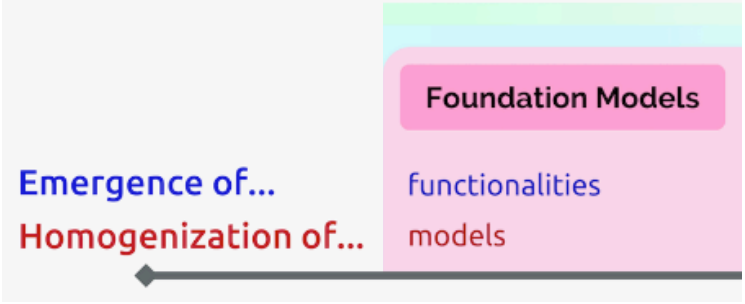


- **Foundation Model:** Around 2018, represents one more shift of general AI paradigm. Mainly the area of Natural language Processing (NLP) witnessed this seismic shift; however these change is applicable to Computer Vision and other areas of AI. Foundational models are enabled by two things:
  - **Scale**: this property made foundation model a powerful one. Scale was achieved by:
    - **Better GPU:** faster computation
    - **Better deep neural network:** eg, Transformer model
    - **Much more training data:** in the scale of billions and more
    - **Self-supervised learning:** suppress a small portion of the data and train a model to do the prediction (BERT, GPT models are learned via self-supervised method)
  - **Transfer learning**: take knowledge from one task *(object recognition of 1000 image category using ImageNet)* and apply it to another task *(21 underwater animal recognition task)*. It makes foundation model possible.
    - You have already done transfer learning when you took a *pre-trained* AlexNet/ResNet model for your homework#3 and fine-tuned it on another (adapting pretrained AlexNet on Drake's underwater images)

On the Opportunities and Risks of Foundation Models. Bommasani, R et al. 2021

Foundation Models (2018-present)

NLP popularized Foundation Models



- **Foundation Model:** Around 2018, represents one more shift of general AI paradigm. Mainly the area of Natural language Processing (NLP) witnessed this seismic shift; however these change is applicable to Computer Vision and other areas of AI. Let's explain it in terms of emergence and homogenization

  - **Emergence**: result from the scale. surprising emergence of functionality that was neither the AI system specifically trained for nor anticipated to arise

    - GPT-3 with 175 billion parameters can be adapted to a downstream task simply providing it with a prompt (description of the task)

You don't have to explicitly tell the function (translate, rephrase, paraphrase, question-answer); it will emerge from the model

For example, you could ask GPT-3 to generate Python code for building a board game LUDU with graphical visualization. While GPT-3 may not have been specifically trained for this task, as long as you provide the correct logic and clear expectations for the visualization, it can generate code snippets with a surprising degree of accuracy. Here the **function of LUDU emerges** automatically

  - **Homogenization**:

    - a single **model** such as BERT/RoBERTa/GPT can now solve a wide range of applications

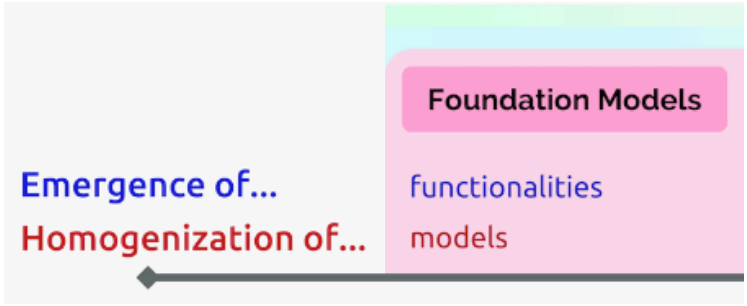
Homogenization is the consolidation of **extremely large models**

    - Besides homogenization of approaches, we observe *homogenization across multiple subfields* (such NLP, Computer Vision) in the form of ***multimodal models***: foundation models trained jointly on images and textual data
    - …

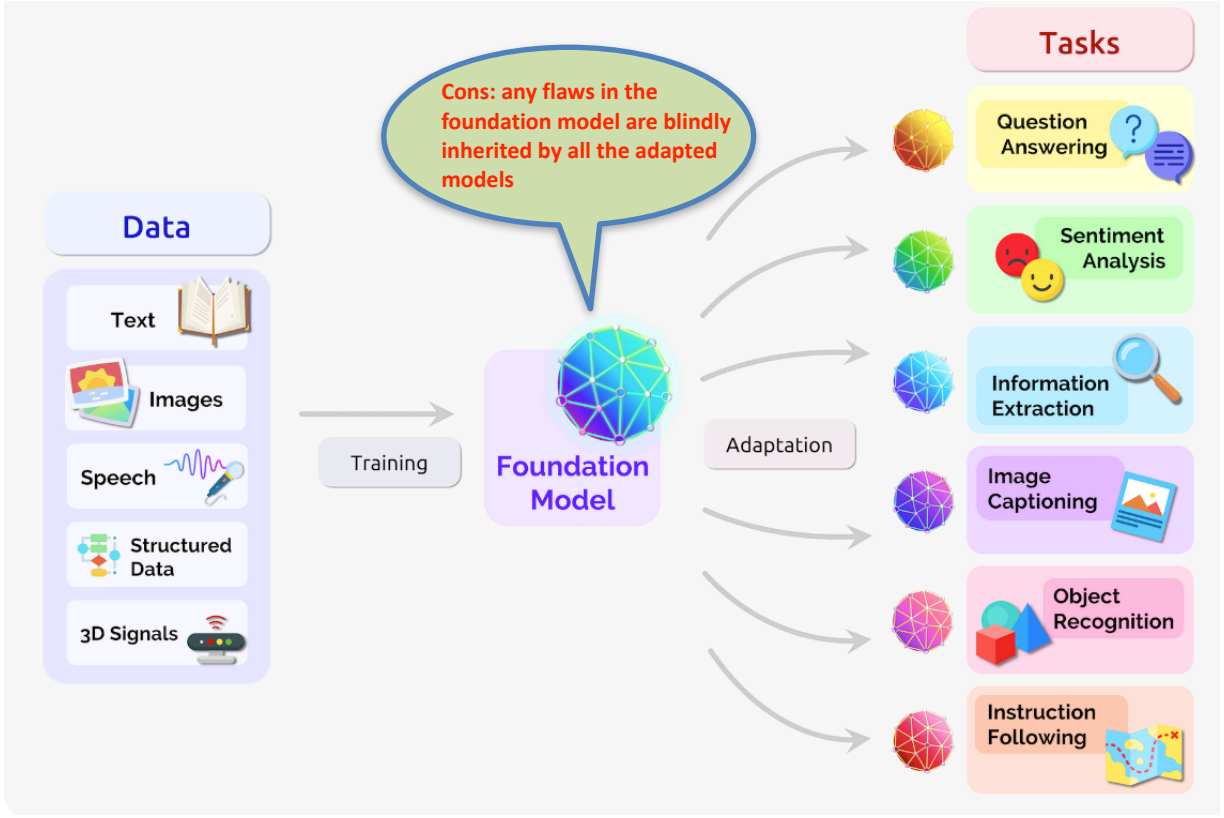
On the Opportunities and Risks of Foundation Models. Bommasani, R et al. 2021

Foundation Models (2018-present)

NLP popularized Foundation Models



- **Foundation Model:** Around 2018, represents one more shift of general AI paradigm. Foundation model is any model that is trained on broad data (generally using self-supervision + very large scale) that can be adapted (eg, fine-tuned) to a wide range of downstream tasks
  - BERT, GPT, CLIP

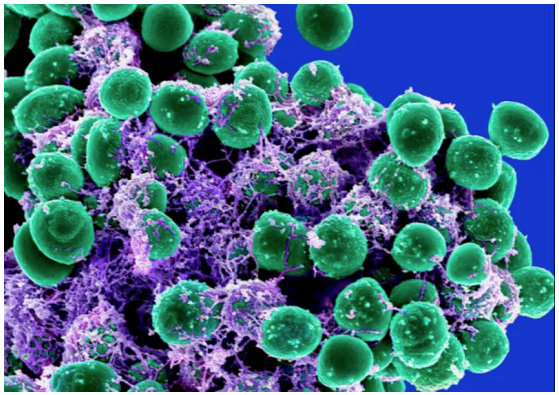


On the Opportunities and Risks of Foundation Models. Bommasani, R et al. 2021

Segment Anything (SAM)

- Segment Anything (SAM) is a foundation model for segmentation

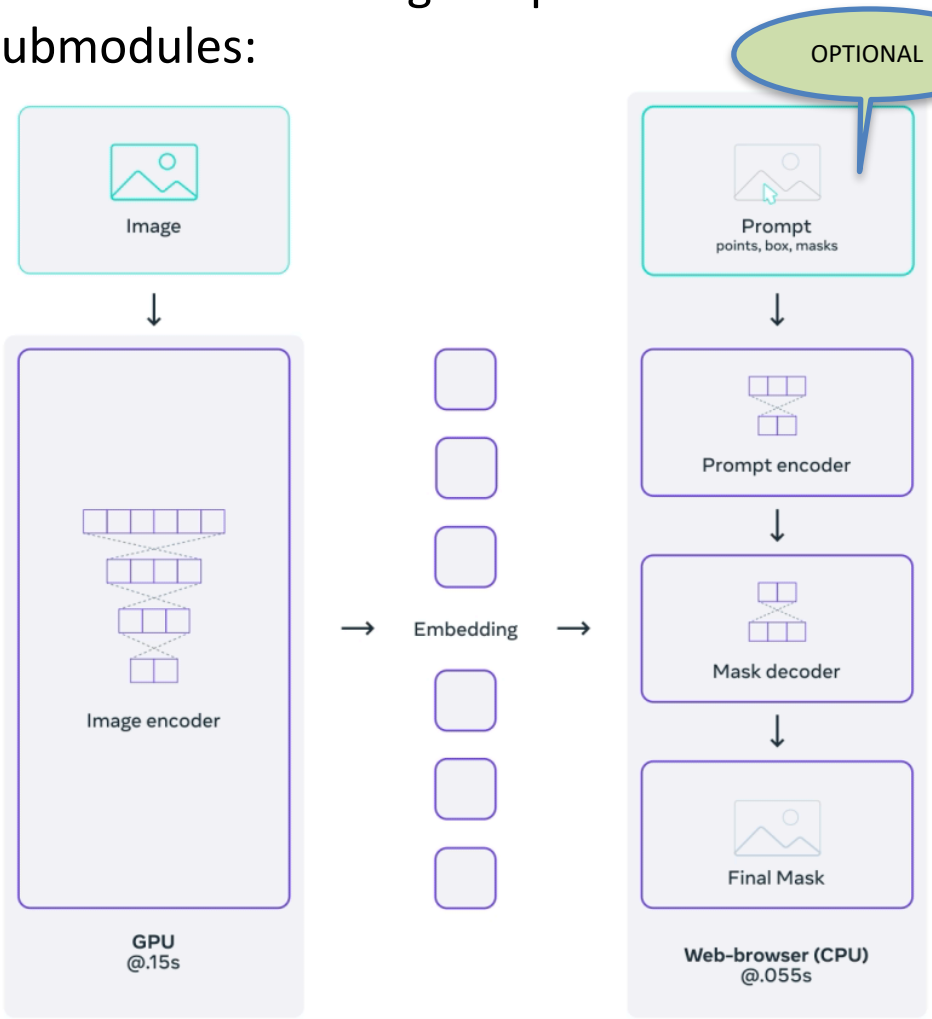
- It can segment unfamiliar objects and never-seen-before image without the need for additional need for training
    - zero-shot generalization property



Segment Anything - arxiv'23

Segment Anything (SAM)

- Segment Anything (SAM) is designed to be efficient enough to power its data engine. Model is decoupled into two submodules:
  - a one-time image encoder and
  - a lightweight mask decoder (that can run in a web-browser in just a few milliseconds per prompt)



- SAM is trained on more than 1.1 billion segmentation masks collected on 11 million images
  - authors used SAM and data to interactively annotate images and update the model. This cycle was repeated many times over to improve both the model and the dataset

Segment Anything - arxiv'23

Segment Anything (SAM)

- Segment Anything (SAM) can also segment images based on a variety of input prompts

Segment based on prompt with interactive points and boxes

Generate multiple valid masks for ambiguous prompts

Segment Anything - arxiv'23

Segment Anything (SAM)

- Segment Anything (SAM) can also segment images based on a variety of input prompts

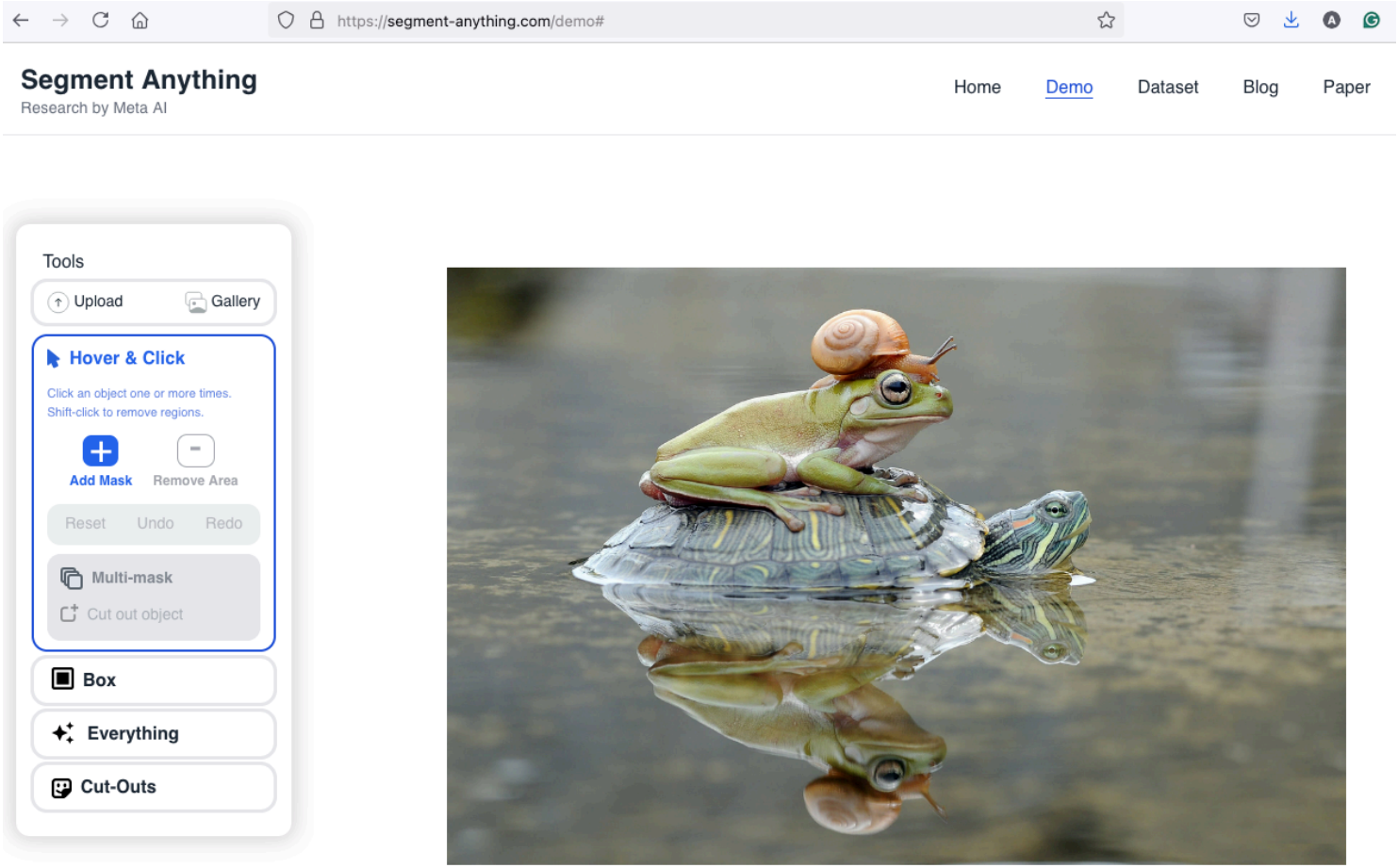
Cons: it may not work for a uncommon object such as sea-lion which remains In underwater environment

Bounding box prompts from an object detector can enable text-to-object segmentation (may work for popular objects)

Segment Anything - arxiv'23

Group Activity#1: Try Segment Anything Demo

- Try zero-shot generalization capability
- Try to segment images based on a variety of input prompts



https://segment-anything.com/demo#

Group Activity#2: Try Segment Anything Demo

- Try zero-shot generalization capability using PyTorch code

https://github.com/alimoorreza/cs195-fall24-notes/blob/main/cs195_segment_anything_SAM_i

cs195-fall24-notes / cs195_segment_anything_SAM_inference.ipynb

Blame 749 lines (749 loc) · 1.87 MB

```
cur_image = 'Crocodile_27.png'
#cur_image   = 'gridline.png'

image       = cv2.imread(root_dir + cur_image)
image       = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
masks2      = mask_generator_2.generate(image)

if is_display_enabled == True:
  plt.figure(figsize=(10,10))
  plt.imshow(image)
  show_anns(masks2)
  plt.axis('off')
  plt.show()


# ---- save the mask ----
# make a directory for each animal
cur_animal    = cur_image.split("/")[0]
dest_dir            = root_dir + "/"

# save the masks + image as pickle file
f1      = open(dest_dir + cur_image[0:-4] + '.pkl', 'wb')
my_dict = {'image': image, 'masks': masks2}
pickle.dump(my_dict, f1)
f1.close()
```

https://github.com/alimoorreza/cs195-fall24-notes/blob/main/cs195_segment_anything_SAM_inference.ipynb