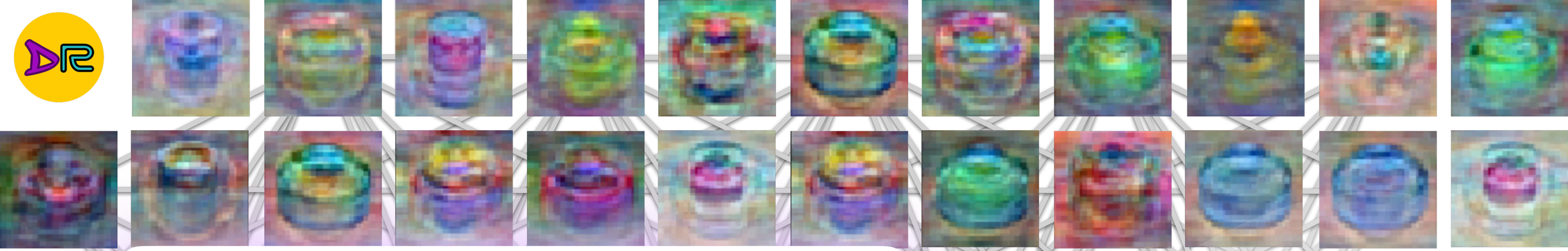


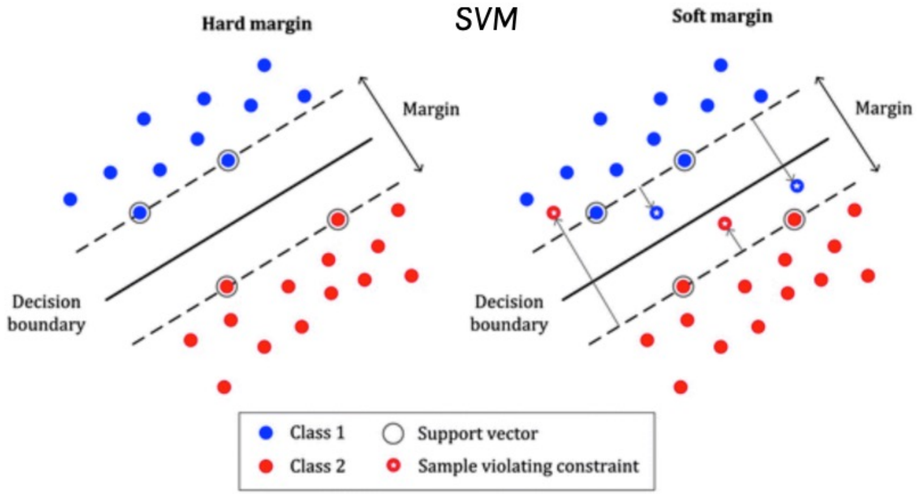
DeepRob

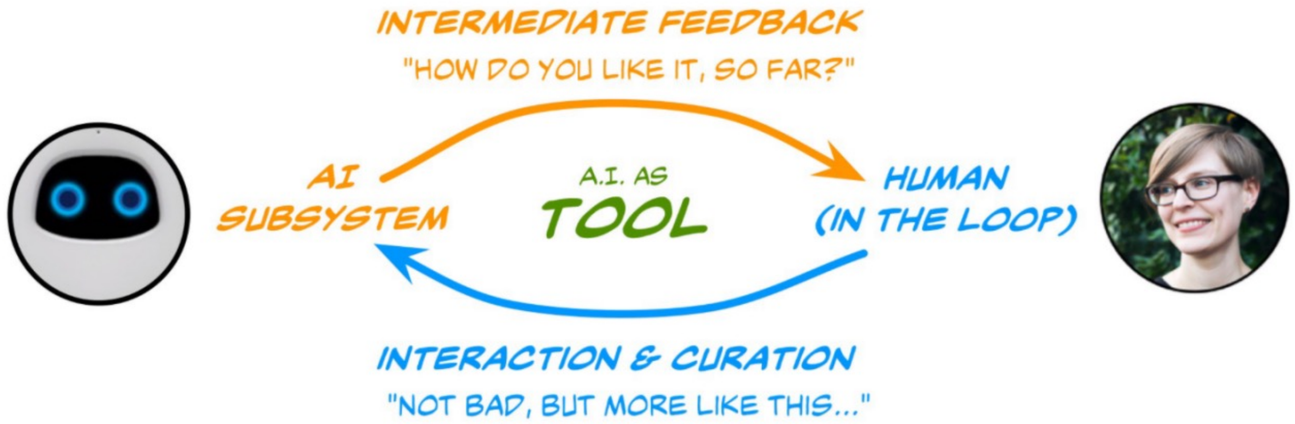
Lecture 18
Interpretability and Uncertainty
University of Michigan | Department of Robotics

DeepRob

Why Interpretability

- Trust
- Safety
- Contestability/Reproducibility





Stanford, Humans in the Loop General Bluprint
*Some figures adapted from Harvard AC295 Lecture 11

Transparent Model

- Visualize feature maps:

- Saliency Map
- Class Activation Maximization (CAM)

Answers this question: What made the network give a certain class as output?

e.g., Why is it classifying this image as “dog”?

Saliency Map

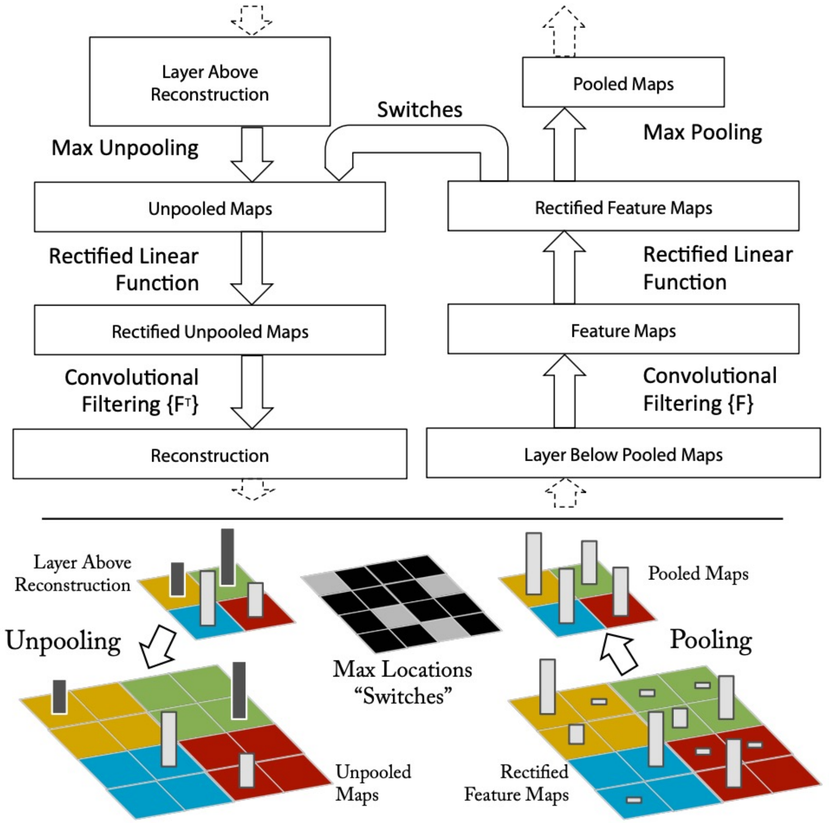
- (we didn't talk about this in class but good to know)

"a way to measure the spatial support of a particular class in each image"

Common method:

Deconvolution (Zeiler and Fergus, 2013)

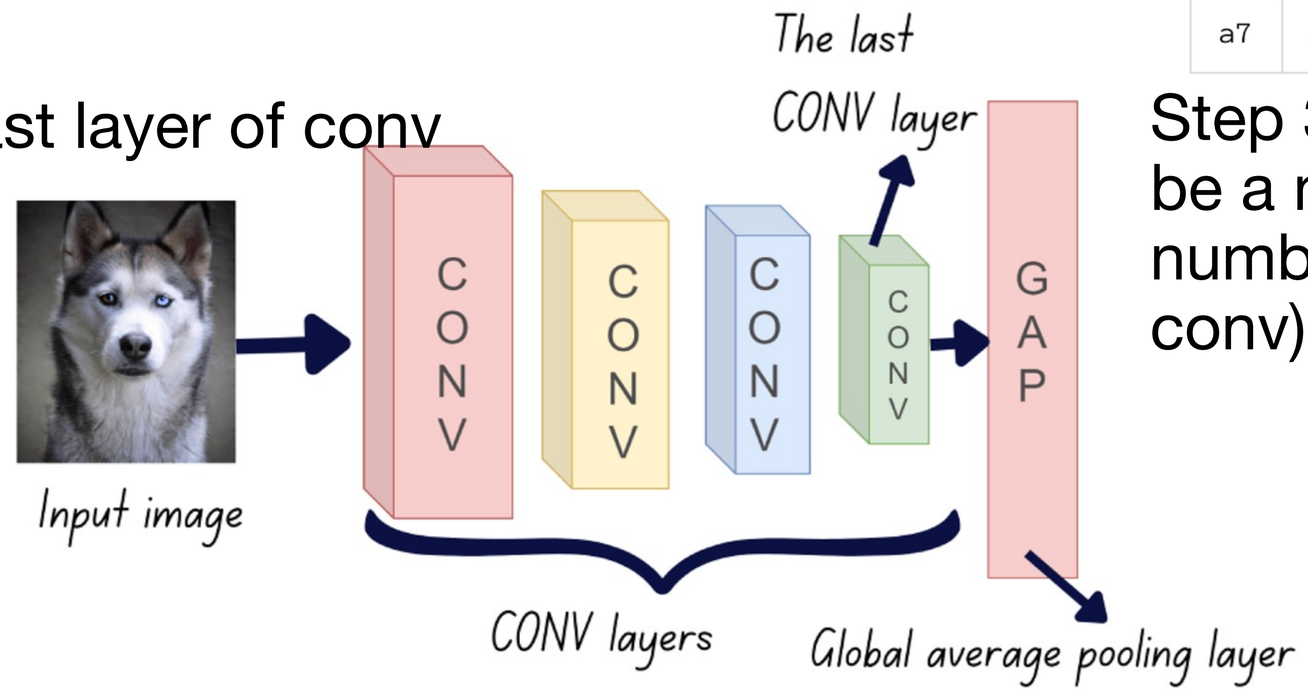
https://arxiv.org/abs/1311.2901



CAM

- Class Activation Map

Step 1: last layer of conv



Step 2: GAP (Global average pooling)

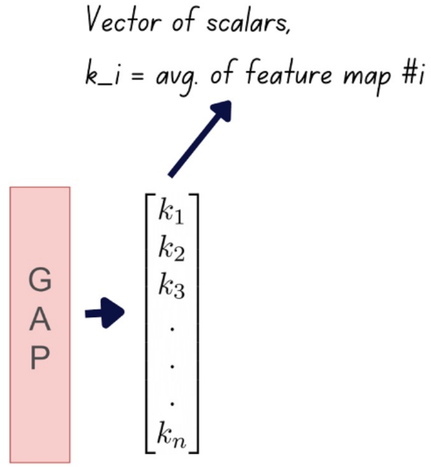
Take the average of feature map -> scalar

Feature map #1

a1	a2	a3
a4	a5	a6
a7	a8	a9

$$k1 = \frac{a1 + a2 + \ldots + a9}{9}$$

Step 3: The output of GAP will be a n-length vector, for n is the number of feature maps (from conv)



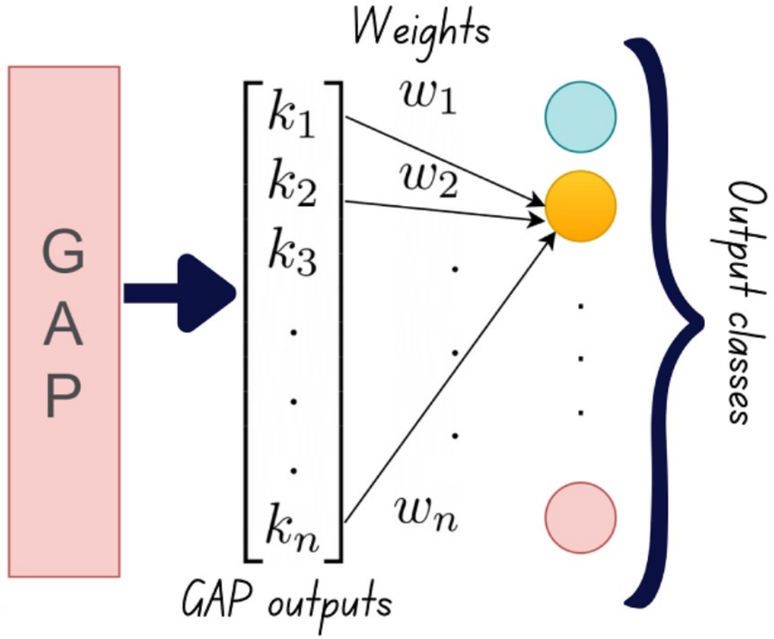
https://www.pinecone.io/learn/class-activation-maps/

CAM

- Class Activation Map

*This captures the relationship between feature maps learned from Conv to class labels!

Step 4: train a linear model to learn the weights between GAP vector outputs and class labels



https://www.pinecone.io/learn/class-activation-maps/

We will train this C times for C classes

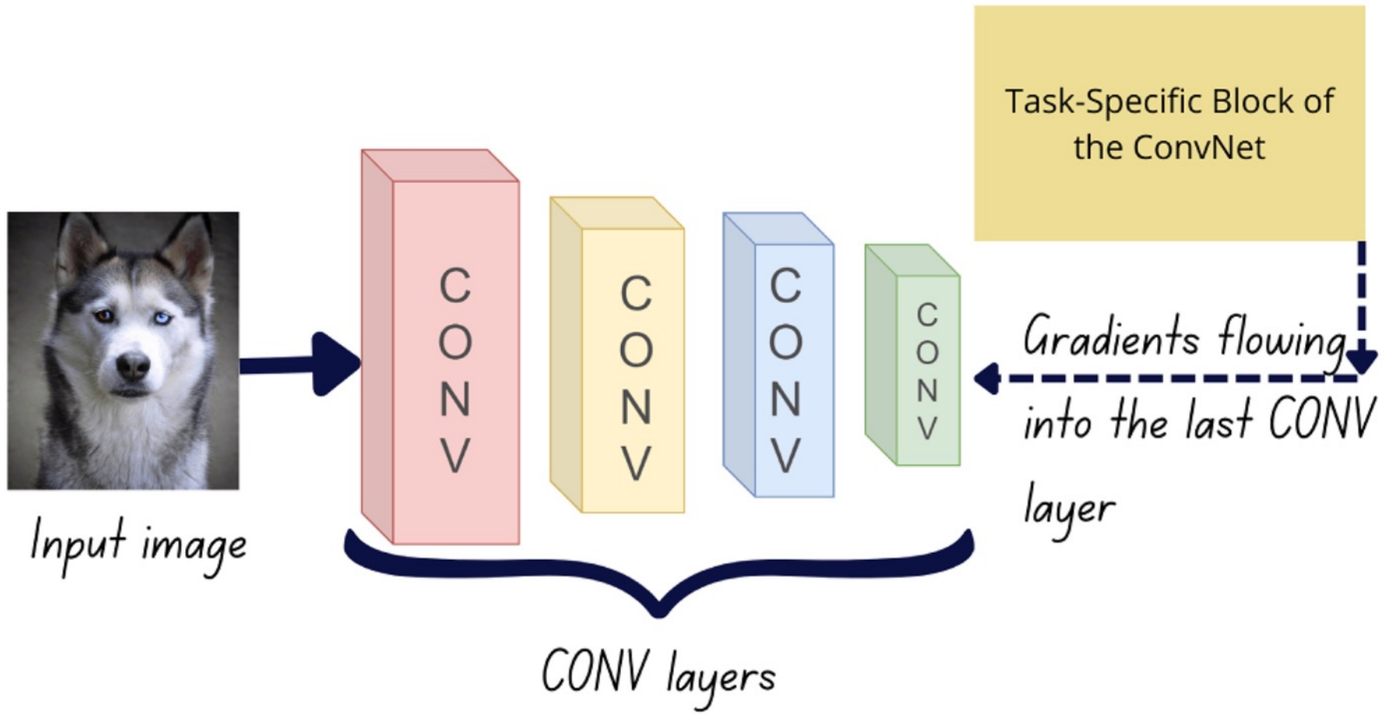
$$y^c = \sum_k w_k^c \frac{1}{Z} \sum_i \sum_j A_{ij}^k$$

c – class
k- k^{th} feature map, k=1,…,n
Z – total number of pixels
A_{ij}^k - the pixel value at (i,j) for the k^{th} feature map

Grad-CAM

*Use backprop to flow gradients from output class score y^c to the feature maps

- Gradient-Weighted Class Activation Maps

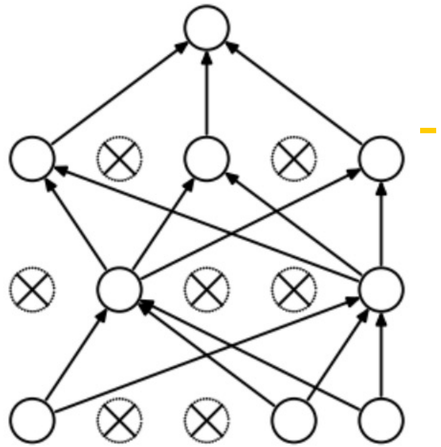


$$w_k^c = \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

$$L_{grad-cam}^c = ReLU\left(\sum_k w_k^c A^k\right)$$

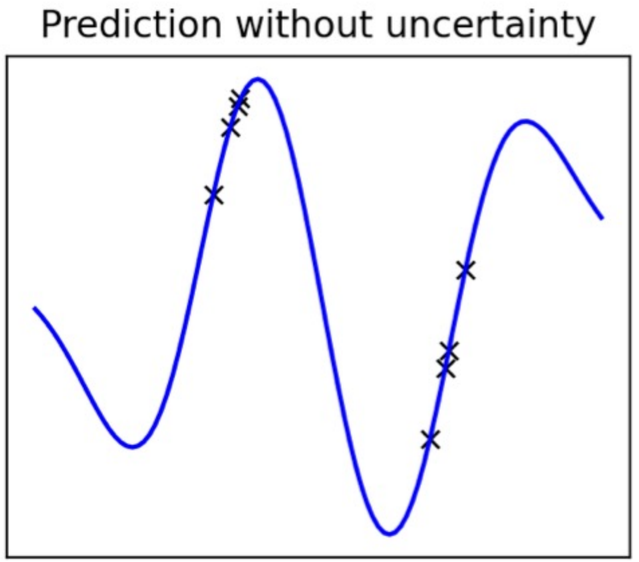
See https://www.pinecone.io/learn/class-activation-maps/ for derivations.

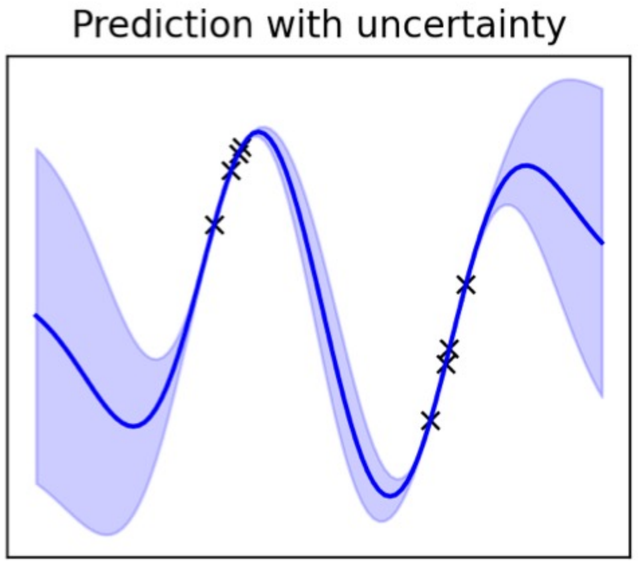
Model with Uncertainty



- Uncertainty in data, label/annotations, model, etc...

- Examples (not limited to):
- **1. Dropout**
- **2. Bayesian Neural Networks**

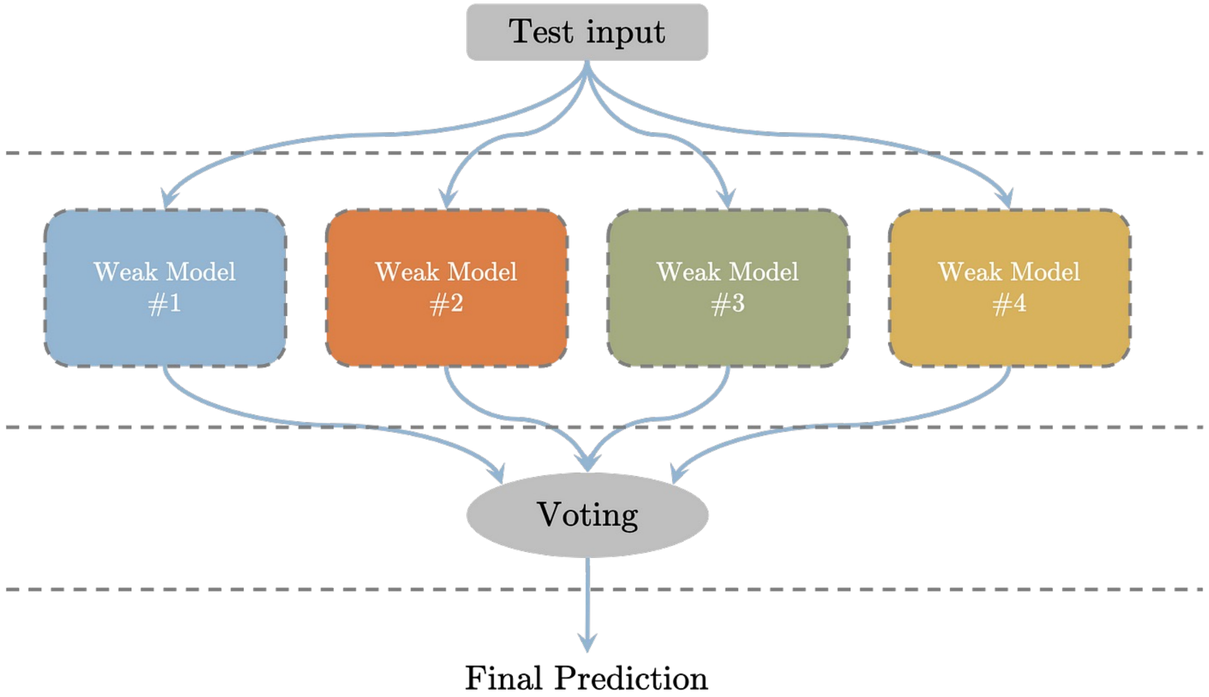




Model with Uncertainty

- Uncertainty in data, label/annotations, model, etc...

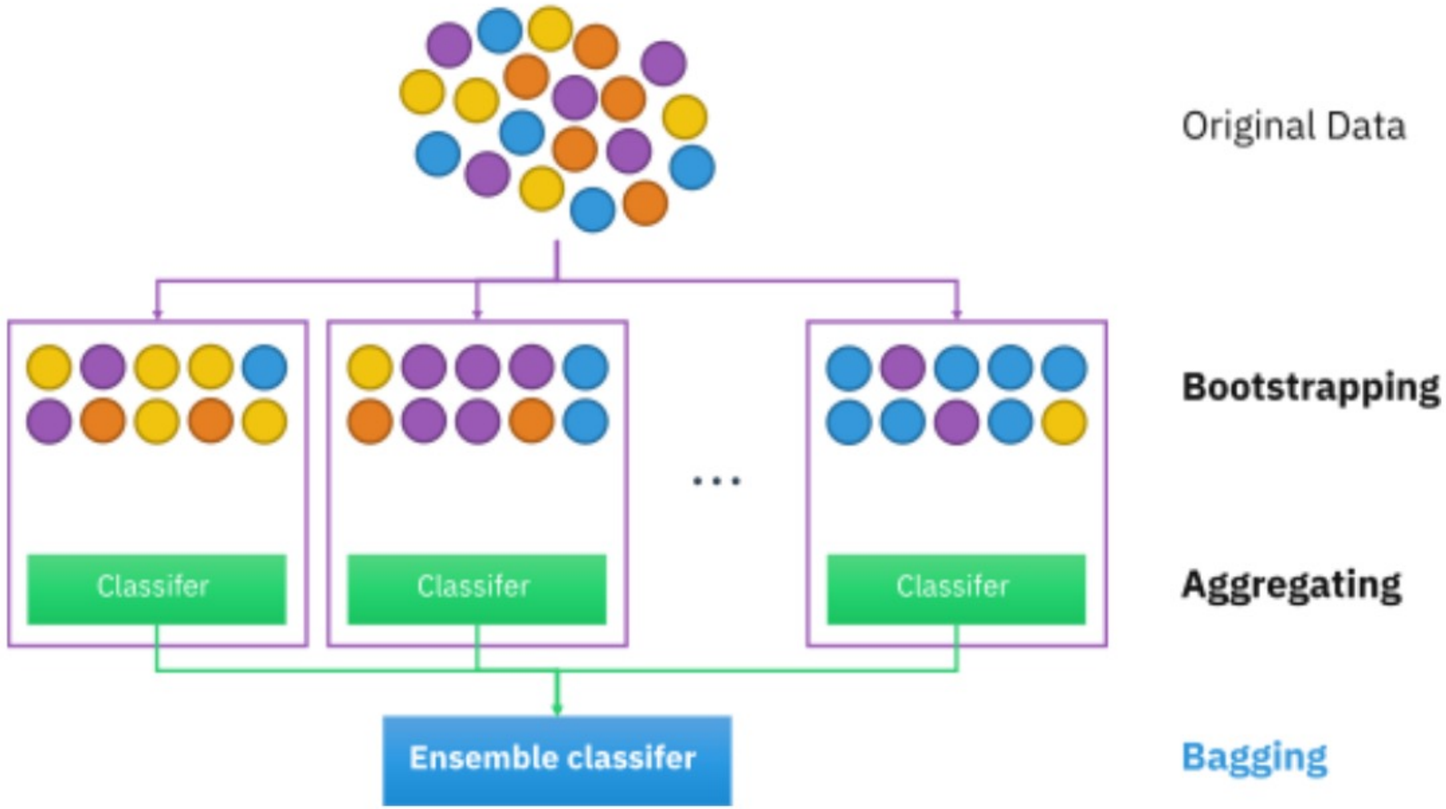
- Examples:
- Dropout
- Bayesian Neural Networks
- **3. Ensemble Methods**



Model with Uncertainty

- Uncertainty in data, label/annotations, model, etc...

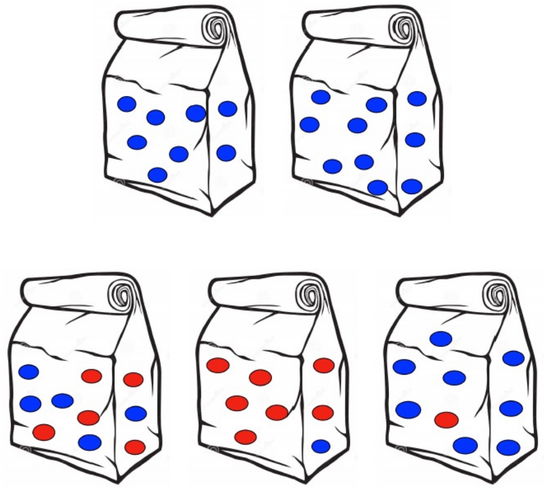
- Examples:
- Dropout
- Bayesian Neural Networks
- Ensemble Methods
- **4. Bootstrap Aggregating (Bagging)**
- Multiple Instance Learning

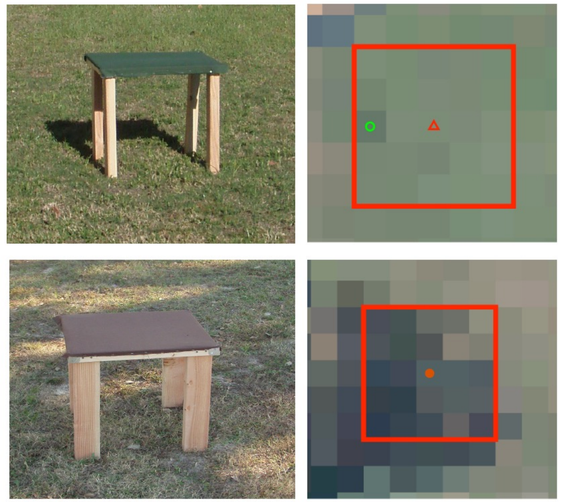


Model with Uncertainty

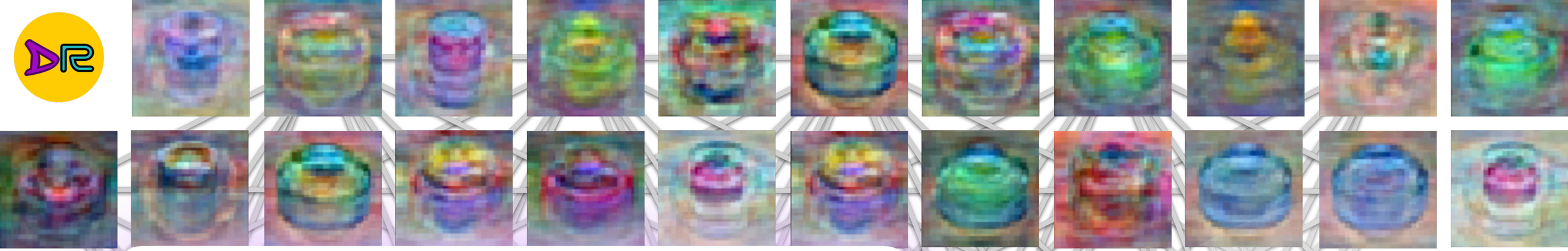
- Uncertainty in data, label/annotations, model, etc...

- Examples:
- Dropout
- Bayesian Neural Networks
- Ensemble Methods
- Bootstrap Aggregating (Bagging)
- **5. Multiple Instance Learning**





"label uncertainty"
Design loss function based on "bag-level" labels



DeepRob

Lecture 18
Interpretability and Uncertainty
University of Michigan | Department of Robotics

DeepRob