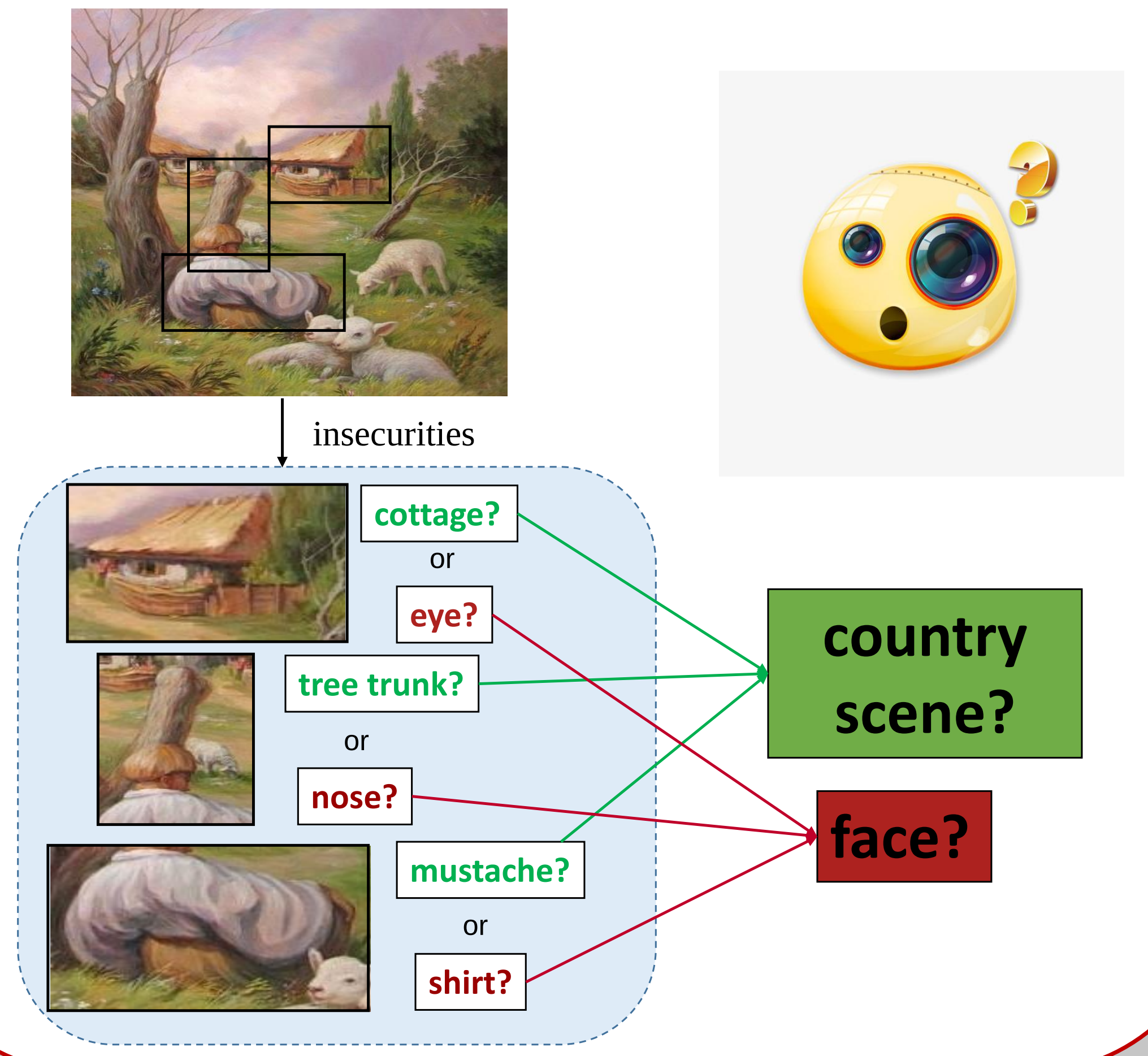
Deliberative Explanations: visualizing network insecurities

Pei Wang Nuno Vasconcelos

Statistical Visual Computing Laboratory, Dept. of Electrical and Computer Engineering, University of California, San Diego

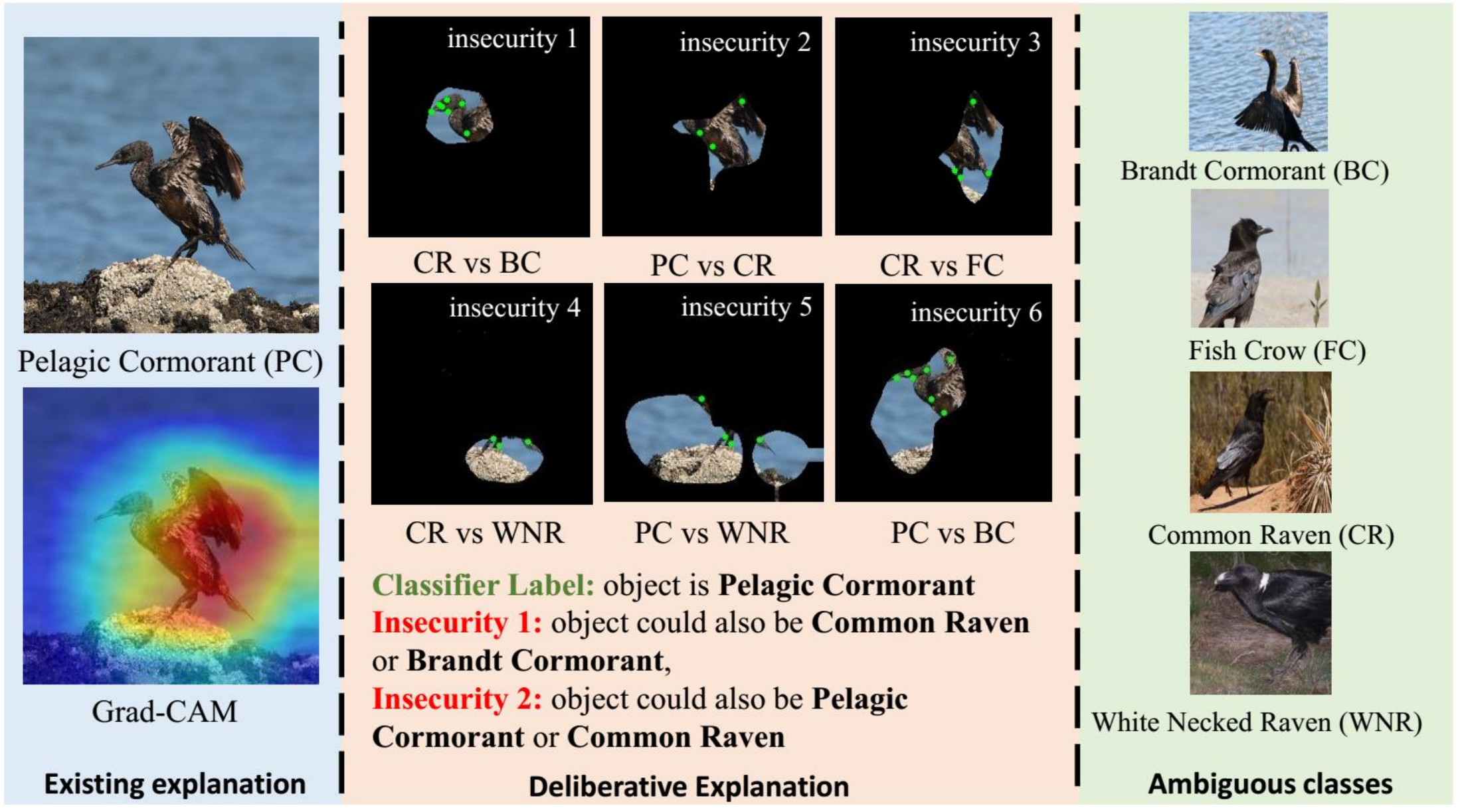
Motivation

- To visualize the deliberative nature of the inference process like humans for classifiers;
- Humans could reasonably oscillate between different interpretations;
- In the limit of highly ambiguous inputs it is even acceptable for different systems (or people) to make conflicting predictions, as long as they provide a convincing justification.



Explanation generation

The explanation consists of a set of insecurities.



Classifier Label: object is **Pelagic Cormorant**
Insecurity 1: object could also be **Common Raven** or **Brandt Cormorant**,
Insecurity 2: object could also be **Pelagic Cormorant** or **Common Raven**

Insecurity generation

- Construct a set of candidate class ambiguities;
- Combine attribution maps of two ambiguous classes and difficulty;

$$m_{i,j}^{(a,b)} = f(m_{i,j}^{a}, m_{i,j}^{b}, m_{i,j}^{s})$$

- Resize and threshold.

Attribution maps

$$m_{i,j}^{p} = [\nabla g_p(\mathbf{A})]_{i,j}^T \mathbf{a}_{i,j} + \frac{1}{2}\mathbf{a}_{i,j}^T [\mathbf{H}(\mathbf{A})]_{i,j} \mathbf{a}_{i,j}$$

Difficulty scores

- Hesitancy score

$$s^{he}(\mathbf{x}) = 1 - \max_y f_y(\mathbf{x})$$

- Entropy score

$$s^{e}(\mathbf{x}) = -\frac{1}{\log C}\sum_y f_y(\mathbf{x}) \log f_y(\mathbf{x})$$

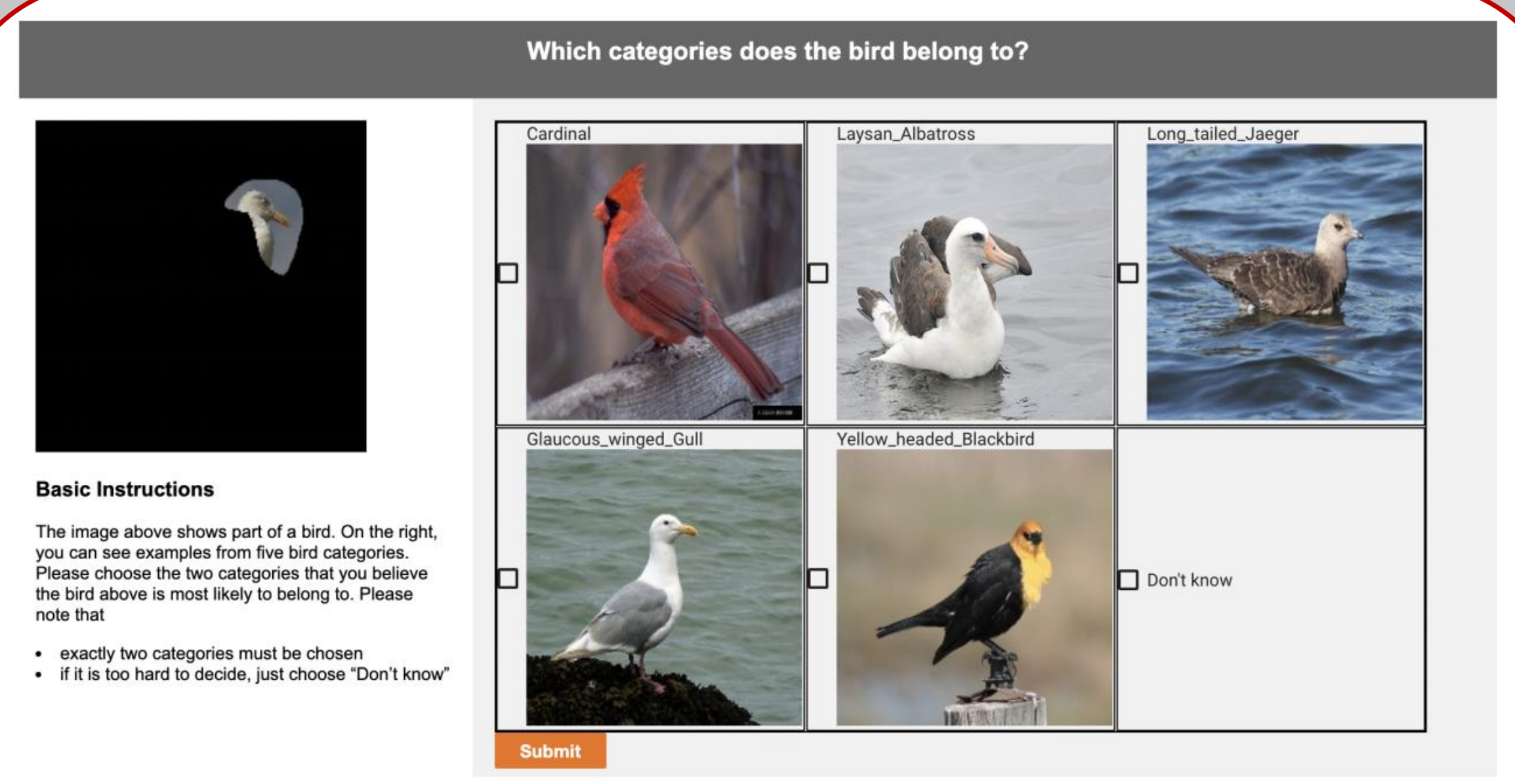
- Hardness score [2].

$$s^{ha}(\mathbf{x}) = s(\mathbf{x})$$

Evaluation

Explanations are usually difficult to evaluate, since explanation ground truth is usually not available.

Human evaluation



Which categories does the bird belong to?

Basic Instructions

The image above shows part of a bird. On the right, you can see examples from five bird categories. Please choose the two categories that you believe the bird above is most likely to belong to. Please note that

- exactly two categories must be chosen
- if it is too hard to decide, just choose "Don't know"

MTurk interface

- Contrast: randomly cropped regions with the same size as insecurities;
- Results: turkers agreed amongst themselves on a and b for 59:4% of the insecurities and 33.7% of randomly cropped regions. Turkers agreed with the algorithm for 51.9% of the insecurities and 26.3% of the random crops.

Evaluation by proxy tasks

- Define part, points on CUB200 and segments on segmentation datasets;
- Compute ambiguity strength (similarity) for all parts $\mathbf{p}$, class pairs (a, b),

$$\alpha_{a,b}^{k} = \gamma(\phi_a^k, \phi_b^k)$$

- Remain 20% strongest as ground truth set $\mathcal{G} = \{(\mathbf{p}_i, a_i, b_i)\}_{i=1}^{M}$;
- To evaluate each insecurity $\mathbf{r}(a, b)$
- On CUB200, precision and recall are used

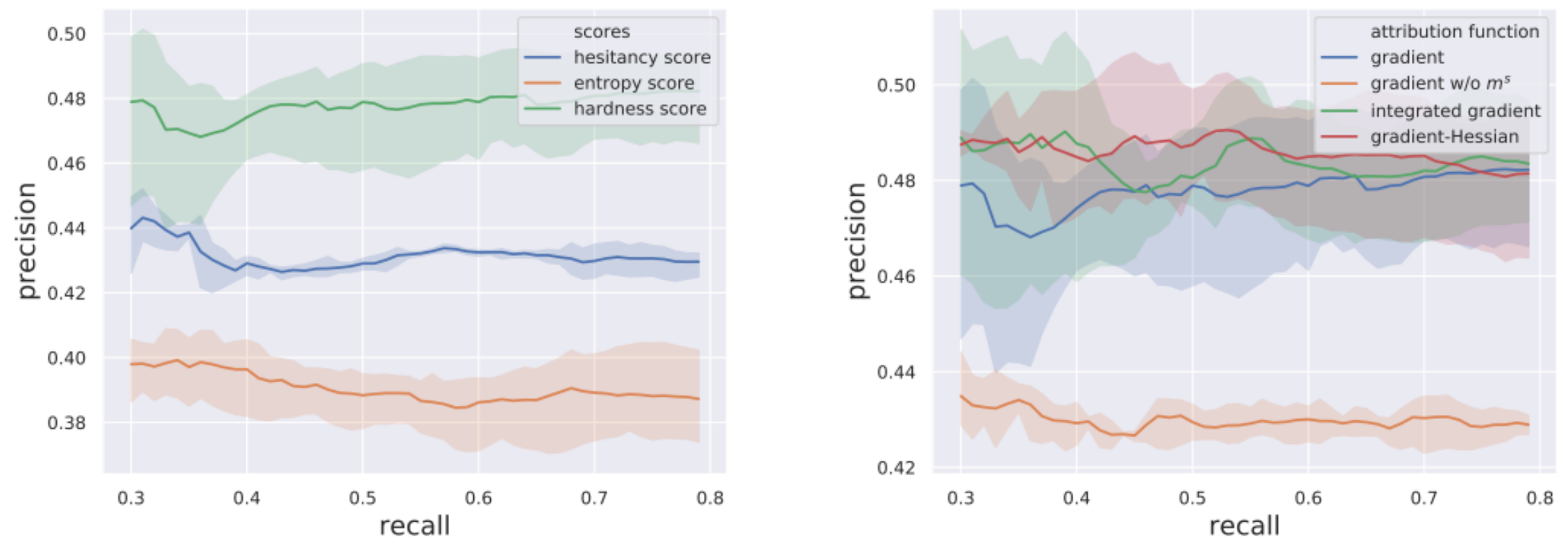
$$P = \frac{|\{i|\mathbf{p}_i \in \mathbf{r}, a_i = a, b_i = b\}|}{|\{k|\mathbf{p}_k \in \mathbf{r}\}|} \quad R = \frac{|\{i|\mathbf{p}_i \in \mathbf{r}, a_i = a, b_i = b\}|}{|\{i|(\mathbf{p}_i, a_i, b_i) \in \mathcal{G}, a_i = a, b_i = b\}|}$$

- On segmentation datasets, IoU metric is used

$$IoU = \frac{|\mathbf{r} \cap \mathbf{p}|}{|\mathbf{r} \cup \mathbf{p}|}$$

Results

Ablation study



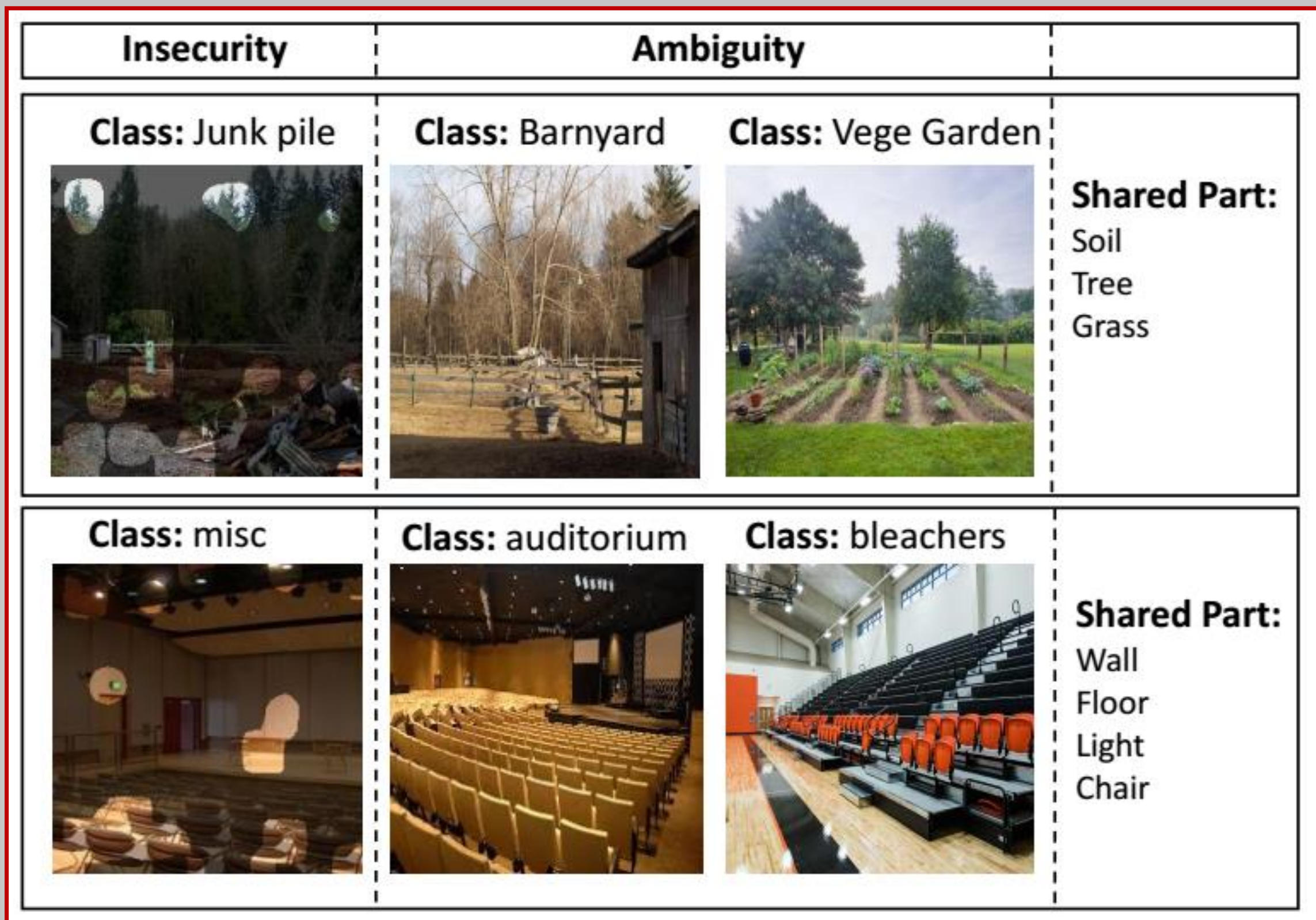
CUB200 [4]

Methods	10%	20%	30%	40%	50%	Avg.
Hesitancy score	8.32(0.05)	15.62(0.01)	22.25(0.02)	28.45(0.06)	34.31(0.11)	21.79(0.03)
Entropy score	8.16(0.06)	15.10(0.08)	21.26(0.07)	26.92(0.18)	32.23(0.30)	20.73(0.09)
Hardness score [2]	**8.63**(0.12)	**16.59**(0.16)	**24.14**(0.19)	**31.34**(0.22)	**38.29**(0.24)	**23.80**(0.19)
Gradient [6]	8.63(0.12)	16.59(0.16)	24.14(0.19)	31.34(0.22)	38.29(0.24)	23.80(0.19)
Gradient w/o m^s	8.54(0.17)	16.35(0.44)	23.70(0.77)	30.67(1.16)	37.39(1.59)	23.33(0.82)
Int. grad. [1]	8.70(0.12)	16.75(0.20)	24.37(0.27)	31.60(0.31)	38.56(0.30)	23.99(0.24)
Gradient-Hessian	**8.86**(0.20)	**17.00**(0.29)	**24.65**(0.32)	**31.92**(0.35)	**38.88**(0.34)	**24.26**(0.30)

ADE20K [5]

Impact of different difficulty scores and attribution functions

Visualization results



Reference

[1] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. ICML, 2017
[2] Pei Wang and Nuno Vasconcelos. Towards realistic predictors. ECCV, 2018.
[3] Ramprasaath R Selvaraju,et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. ICCV, 2017.
[4] P. Welinder,et al. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.
[5] Bolei Zhou, et al. Scene parsing through ade20k dataset. CVPR, 2017.
[6] Karen Simonyan et al. Deep inside convolutional networks: Visualising image classification models and saliency maps, ICLR, 2014