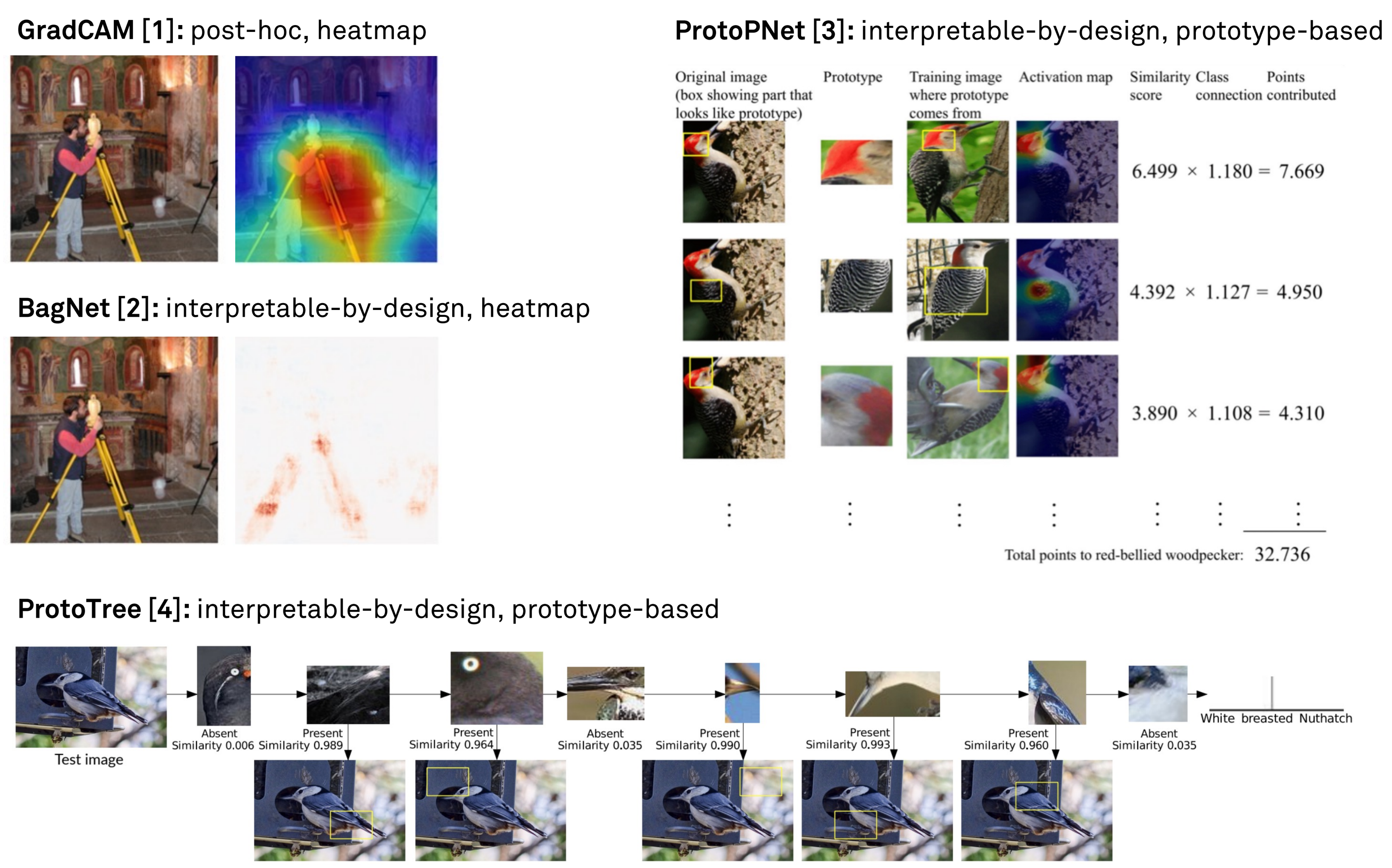
HIVE: Evaluating the Human Interpretability of Visual Explanations

Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, Princeton University

Overview

- Despite the growth of the interpretability/XAI field, **evaluating interpretability methods** remains a challenge, particularly due to the diverse range of proposed explanation types.

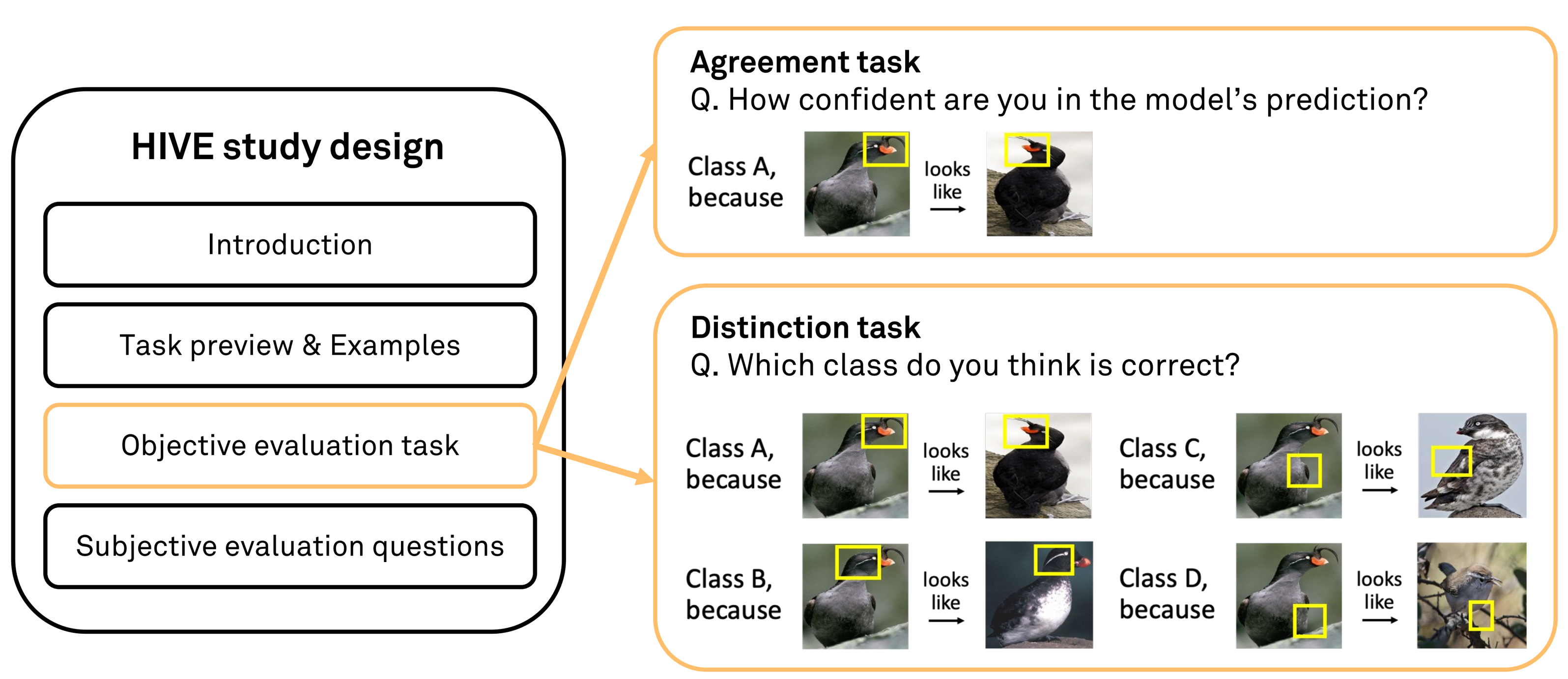


Our contributions:

- We present HIVE (Human Interpretability of Visual Explanations), a novel human evaluation framework for visual interpretability methods.
- We demonstrate HIVE's effectiveness and usefulness for evaluating a variety of interpretability methods, and open-source our UI code.
- We are the first to investigate the utility of visual explanations in distinguishing correct and incorrect predictions, conduct human studies for interpretable-by-design models, and study how users trade off interpretability and accuracy.

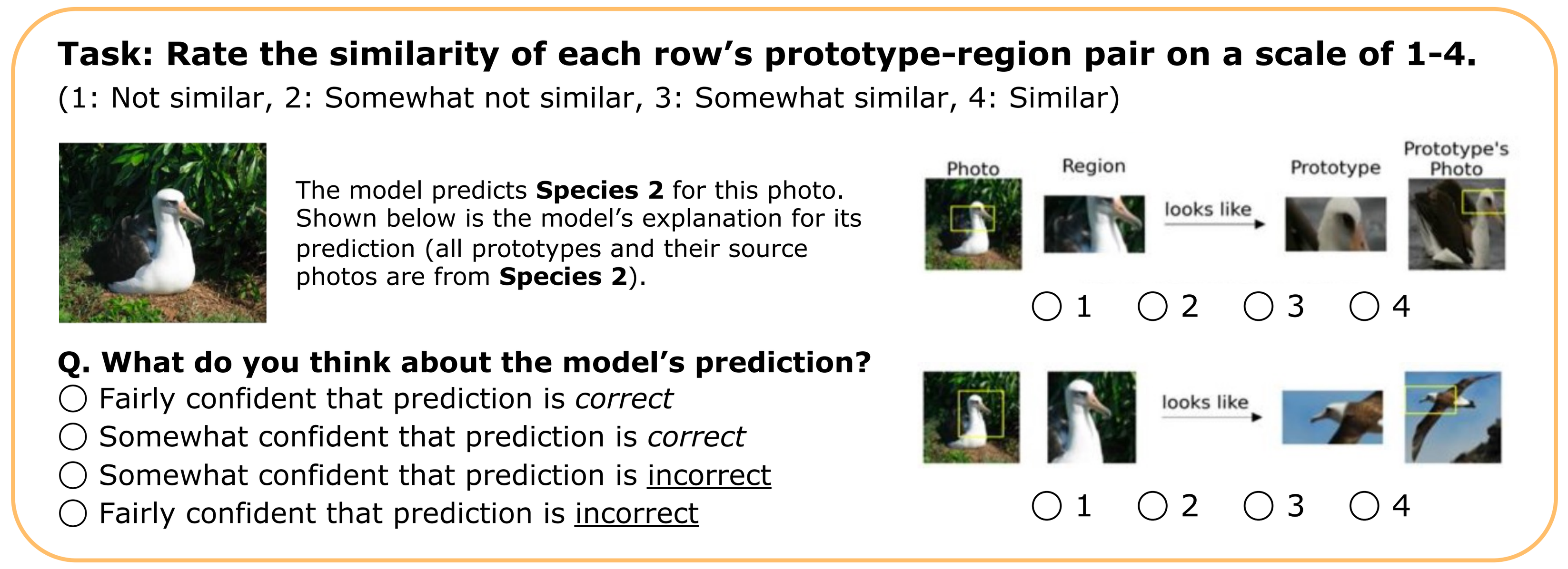
HIVE study design

- HIVE was designed to enable **cross-method comparison** by evaluating a variety of interpretability methods on a common task.
- For **human-centered evaluation**, we design these tasks to measure the utility of explanations to human users in AI-assisted decision making scenarios.
- These objective evaluation tasks enable **falsifiable hypothesis testing** about whether a given method has a certain property.

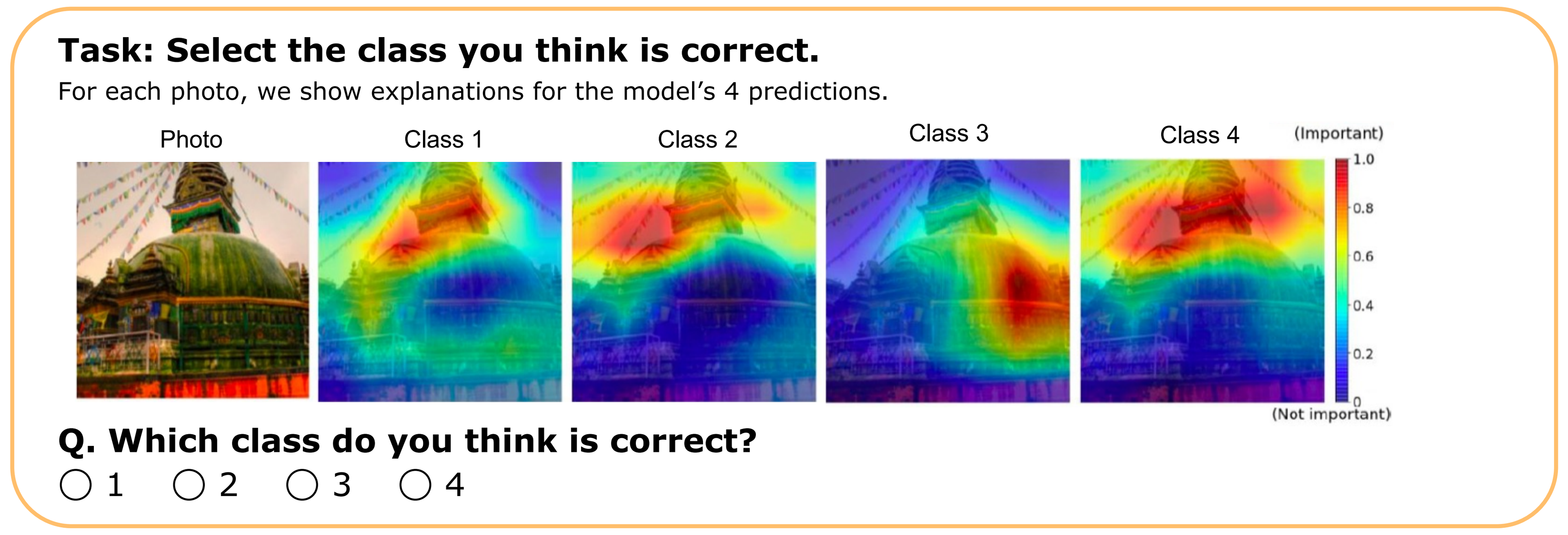


Evaluation UI examples

UI for ProtoPNet [3] agreement study



UI for GradCAM [1] distinction study



Experimental setup

- We conduct IRB-approved human studies of 4 methods that span the diversity of visual interpretability methods on CUB (birds) and ImageNet (objects) image classification tasks.
- We evaluate each method on the agreement and distinction tasks. Each study is completed by 50 participants recruited through Amazon Mechanical Turk.
- For each study, we report the mean accuracy and standard deviation of the participants' performance. We also compare the study result to random chance and compute the p-value from a 1-sample t-test.

Key findings

- The **agreement** task results reveal an issue of **confirmation bias**: Participants tend to believe that a model prediction is correct when given an explanation for it.

CUB	GradCAM [1]	BagNet [2]	ProtoPNet [3]	ProtoTree [4]
Correct	72.4% ± 21.5%	**75.6% ± 23.4%**	73.2% ± 24.9%	66.0% ± 33.8%
Incorrect	32.8% ± 24.3%	*42.4% ± 28.7%*	***46.4% ± 35.9%***	37.2% ± 34.4%

ImageNet	GradCAM [1]	BagNet [2]
Correct	**70.8% ± 26.6%**	66.0% ± 27.2%
Incorrect	***44.8% ± 31.6%***	35.6% ± 26.9%

- **Goal**: 100% accuracy, i.e., participants can perfectly identify whether or not a prediction is correct
- **Baseline**: 50% accuracy with random guessing

- **How to read the numbers:** For GradCAM on CUB, participants thought 72.4% of correct predictions were correct and 100 − 32.8 = 67.2% of incorrect predictions were correct.

Key findings (continued)

- The **distinction** task results reveal that participants struggle to identify the correct class based on explanations, especially when the model has made an incorrect prediction.

CUB	GradCAM [1]	BagNet [2]	ProtoPNet [3]	ProtoTree [4]
Correct	**71.2% ± 33.3%**	45.6% ± 28.0%	54.5% ± 30.3%	33.8% ± 15.9%
Incorrect	*26.4% ± 19.8%*	**32.0% ± 20.8%**		

ImageNet	GradCAM [1]	BagNet [2]
Correct	**51.2% ± 24.7%**	38.4% ± 28.0%
Incorrect	***30.0% ± 22.4%***	*26.0% ± 18.4%*

- **Goal**: 100% accuracy, i.e., participants can perfectly identify the correct class (the predicted class for the below table)
- **Baseline** : 25% accuracy with random guessing

- **How to read the numbers:** For GradCAM on CUB, participants were able to identify the correct class for 71.2% of the correct predictions and 26.4% of the incorrect predictions.

- For GradCAM [1] and BagNet [2], we also ask participants to select the class they think the model predicts (**output prediction** task) and find they struggle to identify the output based on explanations.

Dataset	CUB		ImageNet	
Method	GradCAM [1]	BagNet [2]	GradCAM [1]	BagNet [2]
Correct	**69.2% ± 32.3%**	50.4% ± 32.8%	**48.0% ± 28.3%**	46.8% ± 29.0%
Incorrect	**53.6% ± 27.0%**	30.0% ± 24.1%	**35.6% ± 24.1%**	34.0% ± 24.1%

- **How to read the numbers:** For GradCAM on CUB, participants were able to identify the class the model predicted for 69.2% of the correct predictions and 53.6% of the incorrect predictions.

- For ProtoPNet [3] and ProtoTree [4], we ask participants to rate the similarity of prototype-image pairs and empirically confirm prior work's [4, 5] anecdotal observation that **prototype-based models' notion of similarity sometimes doesn't align with that of humans**.

- Finally, we study the **interpretability-accuracy tradeoff** participants are willing to make under different risk settings. On average, participants require a baseline model to have +6.2% higher accuracy for low-risk (e.g., scientific or educational purposes), +8.2% for medium-risk (e.g., biodiversity and ecosystem monitoring), and +10.9% for high-risk (e.g., veterinary science or medical diagnosis) settings, to use it over a model with explanations.

More information

- **Project page:** https://princetonvisualai.github.io/HIVE/
- **Code:** https://github.com/princetonvisualai/HIVE

References

1. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." ICCV 2017.
2. Brendel & Bethge. "Approximating CNNs with Bag-of-Local-Features Models Works Surprisingly Well on ImageNet." ICLR 2019.
3. Chen*, Li* et al. "This Looks Like That: Deep Learning for Interpretable Image Recognition." NeurIPS 2019.
4. Nauta et al. "Neural Prototype Trees for Interpretable Fine-grained Image Recognition." CVPR 2021.
5. Hoffmann et al. "This Looks Like That... Does it? Shortcomings of Latent Space Prototype Interpretability in Deep Networks." ICML Workshops 2021.