Interpretable and Actionable Models using Attribute and Uncertainty Information

Yew Siang Tang, Mikaela Uy, Olivia Hsu
{yewsiang, mikacuy, owhsu} @stanford.edu
Computer Science Department, Stanford University

Stanford Computer Science

Introduction

Current deep-learning based classification or regression models achieve superior performance over humans in many benchmarks but suffer from both poor **interpretability** and **actionability** due to the black-box nature of these models and the lack of control of their decisions. We define interpretability as the ability to understand why a model arrived at its decisions and actionability as the ability to intervene when we believe the model made mistakes.

In this work, we show that
1) attribute models can improve the **interpretability** of our models while not compromising and even improving the performance of our models on the CUB and OAI datasets,
2) attribute models can also be used in a **test-time intervention** procedure that enables humans to make contributions to intermediate model outputs and improve overall target performance,
3) and **uncertainty modelling** of these attributes enables us to understand which attributes the model is having difficulty with and provide human intervention, allowing us to achieve better test-time intervention results.

Stanford University

Dataset

We used two image datasets, the Caltech-UCSD Birds 200 (CUB) dataset and the Osteoarthritis Initiative (OAI) dataset.

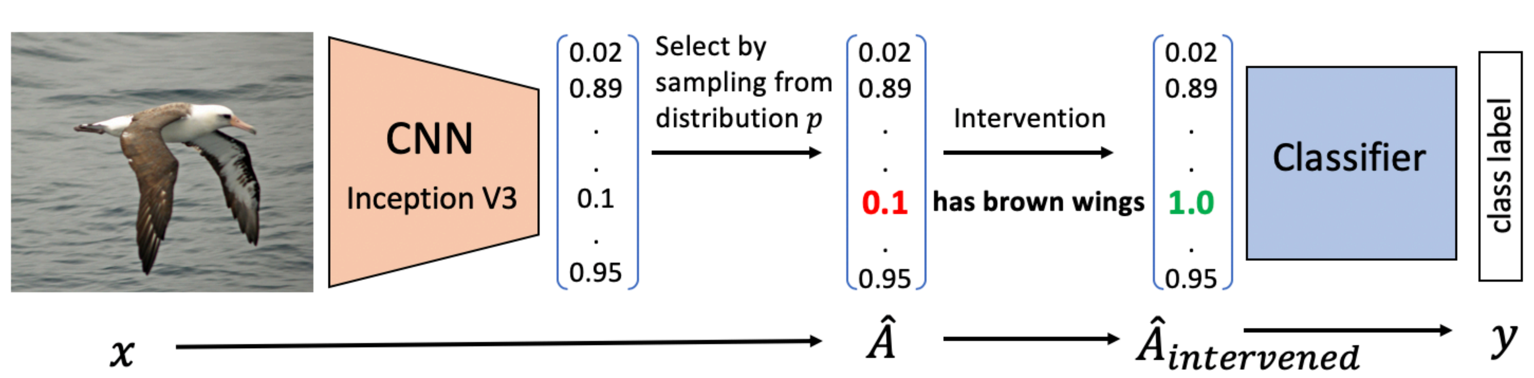
<u>CUB dataset (public)</u> [1]:
- **Colored** image **classification** of **200** different bird species.
- Pre-processed by removing attributes with counts < 10
- A total of **113 bird attributes** were used out of the original 312 hand-labelled **binary** attributes.
- Train examples: 5994, Test examples: 5794

<u>OAI dataset (private)</u>:
- **Grayscale** knee X-ray images **ordinal regression** of the Kellgren & Lawrence Grade (KLG), which indicates the severity of osteoarthritis.
- Selected the **10 least imbalanced attributes** of the original 18 attributes, each an **ordinal number** representing **clinical annotations**.
- Train examples: 21340, Test examples: 11320

Interpretability with Attribute Models

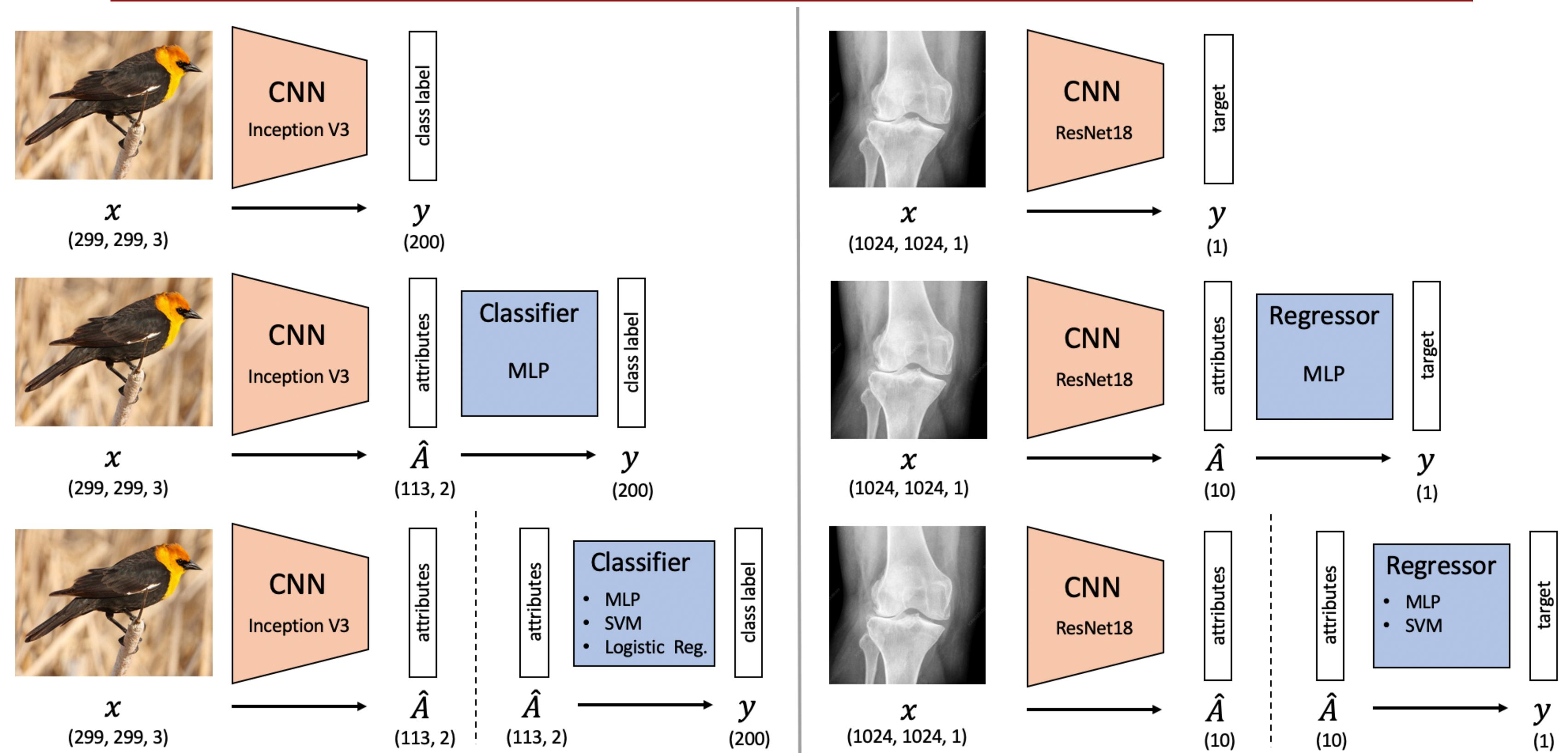
- Given image x and target y, we could use a deep network to model $x \rightarrow y$ directly, i.e. $\hat{y} = f(x)$.
- This is not interpretable nor actionable as users have to accept the predicted outputs.
- We introduce an intermediate attributes A layer which is a **bottleneck layer** that such that the final prediction only uses these attributes and not other parts of the input.

Actionability with Test-time Intervention



- A bottleneck attribute layer enables us to achieve interpretability and actionability by allowing us to ask **counterfactual questions** such as "What happens if the model sees brown wings instead?"
- As shown above, this enables us to examine model outputs and potential outcomes, improving target performance.
- We simulate a **human intervention** procedure where we get assistance on certain attribute values by sampling attributes according to some sampling distribution p.

Model Architectures



Loss Functions

CUB Dataset:

$$L^{CUB}_{x \rightarrow y} = L^{CUB}_{A \rightarrow y} = \frac{1}{n}\sum_{i=1}^{n} CE(y^{(i)}, \hat{y}^{(i)})$$

$$L^{CUB}_{x \rightarrow A} = \frac{1}{n}\sum_{i=1}^{n} CE(A^{(i)}, \hat{A}^{(i)})$$

$$L^{CUB}_{x \rightarrow A \rightarrow y} = L^{CUB}_{A \rightarrow y} + \lambda L^{CUB}_{x \rightarrow A}$$

OAI Dataset:

$$L^{OAI}_{x \rightarrow y} = L^{OAI}_{A \rightarrow y} = \frac{1}{n}\sum_{i=1}^{n} MSE(y^{(i)}, \hat{y}^{(i)})$$

$$L^{OAI}_{x \rightarrow A} = \frac{1}{n}\sum_{i=1}^{n} MSE(A^{(i)}, \hat{A}^{(i)})$$

$$L^{OAI}_{x \rightarrow A \rightarrow y} = L^{OAI}_{A \rightarrow y} + \lambda L^{OAI}_{x \rightarrow A}$$

Uncertainty-based Intervention

- Test-time intervention will be a realistic setting if there is assistance on only a small subset of attributes.
- We want to intervene on attributes that have **high uncertainty** and hence, are likely to give us more information.
- To understand the uncertainty of attributes, we implemented **Bayesian neural networks** through the use of dropouts [2] and compute the standard deviation of samples in Algorithm 1.
- We perform test-time intervention with the following **uncertainty-based sampling distribution**, $p_{dropout}$, where S is a weight that controls how deterministic we want to be.
- We compare with random and softmax schemes to show the effectiveness of our sampling uncertainty scheme.

$$p_{dropout}(a_i) = \frac{(\sigma_i)^S}{\sum_{j=1}^{K}(\sigma_j)^S}$$

Algorithm 1 Predicting uncertainty σ_i of attribute a_i of given input x

```
1: Generated samples G = ∅
2: for j in range(num_samples) do
3:     â_i = CNN(x)                ▷ dropout turned on
4:     Append â_i to G
5: σ_i = std(G)
```

Discussion

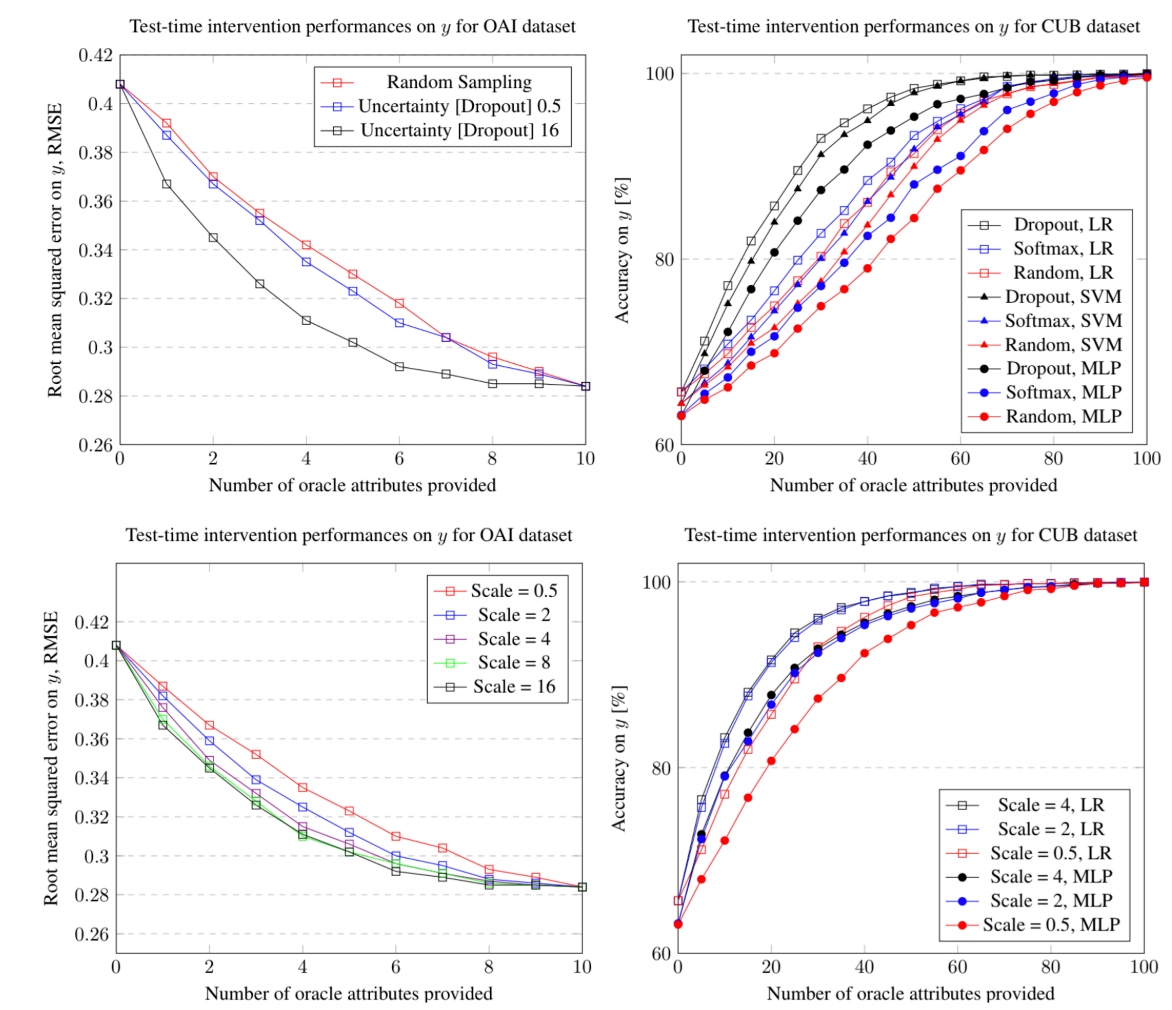
We were able to demonstrate interpretability and actionability for deep-learning image models using multiple datasets through attribute modeling and uncertainty-based test-time intervention. Test-time replacement of both random and highly uncertain attributes improved classification accuracy, which was expected. Our results not only show better overall classification performance, but also illustrate specific examples where humans intervened to correct the classification.

Results

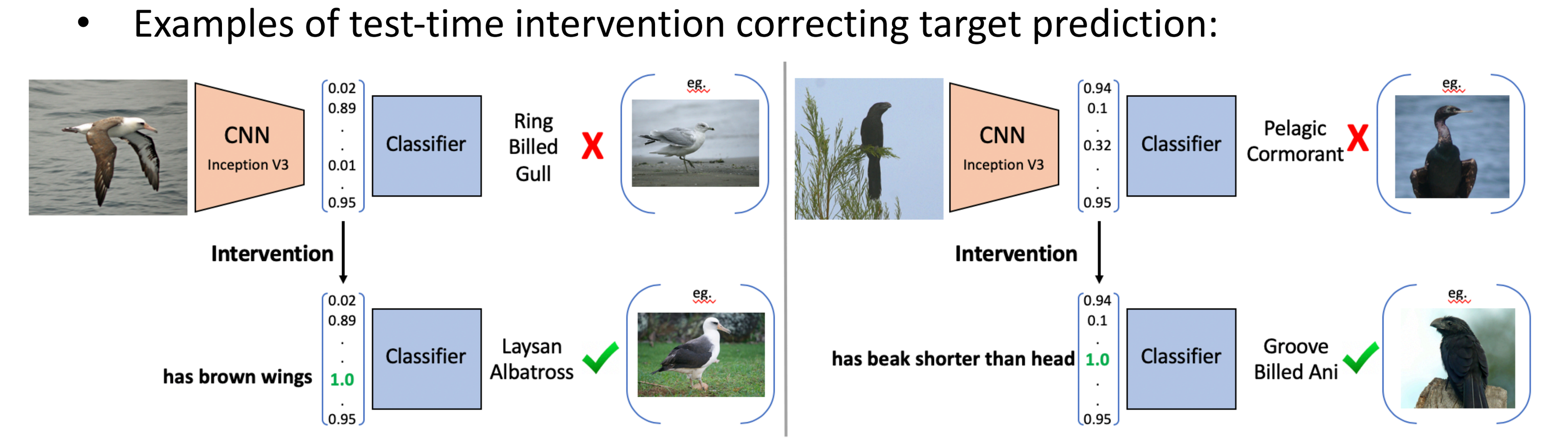
- Prediction performances on y and A for different model setups

	OAI		CUB	
Algorithm	RMSE of y	RMSE of A	Acc of y	Acc of A
Oracle A with RBF SVM $A \rightarrow y$	0.120	-	100.00	-
$X \rightarrow y$	0.428	-	74.64	-
$X \rightarrow \hat{A} \rightarrow y$	0.408	0.718	73.14	83.24
$X \rightarrow \hat{A},\ \hat{A} \rightarrow y$ with MLP	0.429	0.737	63.93	94.92
$X \rightarrow \hat{A},\ \hat{A} \rightarrow y$ with RBF SVM	-	-	64.43	94.92
$X \rightarrow \hat{A},\ \hat{A} \rightarrow y$ with LR	-	-	65.67	94.92

- Test-time intervention improves the performance on multiple datasets. Selection based on uncertainty achieves the best.



- Examples of test-time intervention correcting target prediction:



Future Work

- Use different ML models for $\hat{A} \rightarrow y$ on the OAI dataset.
- Look into intervening on attributes with the maximum increase in expected information gain on the output.
- Expand on uncertainty modeling using other methods, like Bayesian CNNs which place probability distributions on the weights.

Acknowledgements / References

We would like to thank Thao Nguyen and Pang Wei Koh for providing the initial codebase and data pre-processing for the CUB dataset.

[1] Wah C., Branson S., Welinder P., Perona P., Belongie S. "The Caltech-UCSD Birds-200-2011 Dataset." Computation & Neural Systems Technical Report, CNS-TR-2011-001.
[2] Gal, Y., and Ghahramani, Z. 2015a. Bayesian convolutional neural networks with bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158.