# Improving the trustworthiness of image classification models by utilizing bounding-box annotations

**Dharma KC**
University of Arizona
kcdharma@email.arizona.edu

**Chicheng Zhang**
University of Arizona
chichengz@cs.arizona.edu

## Abstract

We study utilizing auxiliary information in training data to improve the trustworthiness of machine learning models. Specifically, in the context of image classification, we propose to optimize a training objective that incorporates bounding box information, which is available in many image classification datasets. Preliminary experimental results show that the proposed algorithm achieves better performance in accuracy, robustness, and interpretability compared with baselines.

## 1 Introduction

Building reliable and trustworthy prediction models has long been a central topic of machine learning research. The reliability and trustworthiness of machine learning models can be characterized from many aspects, including test accuracy, robustness against adversarial attacks [8], interpretability [15], etc. To earn trust and adoption from human users, it is important to develop machine learning models that have good performance on all these aspects.

However, many works show that there might be fundamental tradeoffs between these performance metrics. For example, it is shown by [29] that adversarial robustness may be at odds with accuracy. As another example, decision trees or sparse linear models enjoy global interpretability, however their expressivity may be limited [1, 23].

On the other hand, in many real-world supervised machine learning applications, rich auxiliary information beyond (feature, label) pairs is available at training time. For example, many object detection benchmark datasets provide bounding-box annotations for images [27, 12]. As another example, in ECG-based heart disease prediction, doctors can highlight parts of signals most indicative of her diagnoses. This motivates the question: can learning algorithms benefit from these auxiliary information to train models with improved test accuracy, adversarial robustness and interpretability?

There have been considerable efforts addressing the above question, with most progress focusing on improving test accuracy (e.g. [30, 11, 13]). Although there are some recent efforts aiming to build models with improved accuracy and interpretability (e.g. [21, 14, 19]), a comprehensive understanding of when and how learning from auxiliary information help improve other aspects of trustworthiness of machine learning models (e.g. adversarial robustness) is still missing.

In this paper, we study learning from auxiliary information with the goal of simultaneously improving accuracy, adversarial robustness and interpretability. Specifically, we focus on image classification, and consider auxiliary information in the form of object bounding boxes [27, 12]. Inspired by works on gradient-based regularization [5, 21, 20], we propose a training objective that has different degrees of regularization on different parts of input data, taking advantage of bounding box auxiliary information. Experimentally, we demonstrate on the Caltech-UCSD Birds dataset [31] that our proposed algorithm outputs image classification models with improved accuracy, robustness and interpretability, both quantitatively and qualitatively. Our open source code is available at: https://github.com/ck-amrahd/birds.

1st NeurIPS workshop on Interpretable Inductive Biases and Physically Structured Learning (2020), virtual.

## 2 Related work

Advanced model training algorithms aiming at improving adversarial robustness have been proposed in the literature [28, 5, 20]. However, improvements on robust accuracy often comes at the price of lower standard accuracy [29].

Learning from auxiliary information beyond labels has been studied in various contexts in the literature, for example in text [32, 24, 34] and image [4] domains. It has also received more theoretical treatments in the works of [30, 17, 3]. In the context of image classification, several works aim at improving model accuracy and interpretability using bounding box-based auxiliary information. [14] penalizes the mismatch between the model-generated attention masks and bounding boxes to improve the accuracy and interpratability of convolutional neural networks (CNNs). [21] proposes a regularization term in the training objective that penalizes the gradients of cross entropy losses with respect to input features outside bounding boxes; our work can be seen as a generalization of that work, in that we additionally incorporate gradient-based penalty with respect to the features inside bounding boxes. [33] utilizes more refined part localization bounding box information to train CNNs and improves model accuracy in fine grained classification tasks. Recently, [19] utilizes the attribution algorithm of [26] to propose a new regularizer that can flexibly encourage or penalize different parts of input features.

Our work is also closely related to attribution map or saliency map generation for images [25, 35, 22, 2], in that one can propose training objectives that promote "alignment" between such attribution maps and bounding boxes. Although our work only focuses on regularization based on gradient-based explanations, we believe that regularizing based on these more sophisticated attribution maps are interesting avenues towards improving the trustworthiness of image classification models.

Finally, recent works empirically demonstrate that adversarial robustness and interpretability, two important performance metrics considered in this paper, are tightly connected. On one hand, adversarially robust models generate more interpretable explanations than non-robust ones [29, 6, 10]; on the other hand, models trained to mimic gradient-based explanations of adversarially robust models exhibit robustness [16], hinting at the possibility that robustness is a side-benefit of interpretability.

## 3 Algorithm

**Definitions and settings.** We study image classification with bounding box annotations being part of training data. We are given a set of $m$ training examples $\{(x_i, y_i, M_i)\}_{i=1}^m$, where for example $i$, $x_i \in \mathbb{R}^d$ is its feature part (image $i$'s pixel representation), $y_i \in [K]$ is its label part (the class of the object in the image), $M_i \subseteq [d]$ is the image's associated bounding box. An example of an image with bounding box information is given in Figure 1. Our goal is to train a neural network-based classification model such that, when predicting on test examples, it has high accuracy, robustness and good interpretability.
Formally, given an example $x$, our network outputs a prediction $f(x;\theta)$ that is a probability vector in $\Delta^{K-1}$, the $K$-dimensional probability simplex. Define the cross entropy loss of model $f(\cdot;\theta)$ on example $(x, y)$ as $\ell_{\mathrm{CE}}(\theta,(x,y)) \triangleq \ln \frac{1}{f^y(x;\theta)}$, where we use the notation $z^j$ to denote the $j$-th coordinate of vector $z$.

Figure 1: A Northern flicker image with bounding box shown in green, taken from [31].

**Training objective.** For model training, we propose to optimize the following objective function: $\min_\theta \sum_{i=1}^m \ell(\theta,(x_i,y_i,M_i))$, where

$$\ell(\theta,(x,y,M)) \triangleq \ell_{\mathrm{CE}}(\theta,(x,y)) + \lambda_1 \sum_{j\in M}\left(\frac{\partial \ell_{\mathrm{CE}}(\theta,(x,y))}{\partial x^j}\right)^2 + \lambda_2 \sum_{j\in[d]\setminus M}\left(\frac{\partial \ell_{\mathrm{CE}}(\theta,(x,y))}{\partial x^j}\right)^2 \quad (1)$$

for some $\lambda_1, \lambda_2 > 0$. The intuition behind this training objective is that, in addition to minimizing the usual cross entropy loss, we would like to ensure that the model's predictions have different degrees of sensitivity to different parts of the training images. Specifically, the magnitude of

$\frac{\partial \ell_{\mathrm{CE}}(\theta,(x,y))}{\partial x^j}$ characterizes the sensitivity of the cross entropy loss with respect to the $j$-th pixel. We would like to train a model whose sensitivity to input aligns with object bounding boxes as much as possible; formally, $\frac{\partial \ell_{\mathrm{CE}}(\theta,(x,y))}{\partial x^j}$ should be large for $j$ in $M$ and should be small otherwise.

**Comparison to prior works.** Our training objective generalizes those of [5, 21, 20]. Specifically, when $\lambda_1 = 0$, the objective function becomes that of [21], where only the sensitivity of the loss with respect to the irrelevant parts of the input (coordinates in $[d] \setminus M$) are penalized; [21] shows that this formulation promotes model interpretability. On the other hand, when $\lambda_1 = \lambda_2$, the objective function becomes the double backpropagation objective [5, 20], which is known to improve the generalization accuracy and adversarial robustness of models.

# 4 Experiments

We use the Caltech-UCSD Birds (abbrev. CUB) dataset [31] for experimental evaluation, which has 11,788 examples. We take the union of the training and test sets provided by the CUB dataset, permute the set, and perform a four-way split. The first split consists of $1/2$ of the data, which is used for training by optimizing our objective (1). The remaining data is divided into three sets of equal sizes: the first set is used to select the best model during training, the second set is for $\lambda_1$ and $\lambda_2$ hyperparameter selection, and the third set is used for testing. We choose ResNet50 [9] as our model architecture, and train with mini-batch stochastic gradient descent with a learning rate of 0.001. We consider training with the choices of $\lambda_1$ and $\lambda_2$ in a grid $\Lambda^2$, where $\Lambda = \{0\} \cup \left\{ (\sqrt[3]{10})^i : i \in \{-3, -2, \ldots, 9\} \right\}$. All experiments are repeated three times. We evaluate the following set of algorithms:

1. $\lambda$-VARY (our proposed approach): train a model for each $(\lambda_1, \lambda_2)$ in $\Lambda^2$, and use the validation set to select the best performing model.
2. $\lambda$-EQUAL: train a model for each $(\lambda_1, \lambda_2)$ in $\left\{(\lambda_1, \lambda_2) \in \Lambda^2 : \lambda_1 = \lambda_2\right\}$, and use the validation set to select the best performing model.
3. BLACKOUT: train a model that minimizes the cross entropy loss over modified training examples $(\tilde{x}_i, y_i)$'s; here $\tilde{x}_i$ is defined as $x_i$ with coordinates outside $M_i$ set to zero.
4. STANDARD: standard training that minimizes the cross entropy loss over $(x_i, y_i)$'s; this is also equivalent to setting $\lambda_1 = \lambda_2 = 0$.

## 4.1 Standard and robust accuracy comparison

The adversarial robustness of our trained models are tested for 10 values of adversarial perturbation radii $\epsilon$'s in $\left\{ \frac{0.2i}{9} : i \in \{0, \ldots, 9\} \right\}$ using the Fast Gradient Sign Method [8]. We choose max value of $\epsilon$ to be $0.2$ because beyond that perturbation, the images become unrecognizable by humans. Our adversarial examples were generated using the Foolbox library [18]. Recall that for $\lambda$-VARY and $\lambda$-EQUAL, for each value of $\epsilon$, we choose separate values of $(\lambda_1, \lambda_2)$ pairs using the validation set.
Our results are shown in Figure 2. It can be seen that $\lambda$-VARY trains models that have higher standard accuracy and also robust to adversarial attacks; the performance of the learned models beat those of $\lambda$-EQUAL (especially when $\epsilon$ is large), showing the utility of incorporating bounding box information in the training objective.

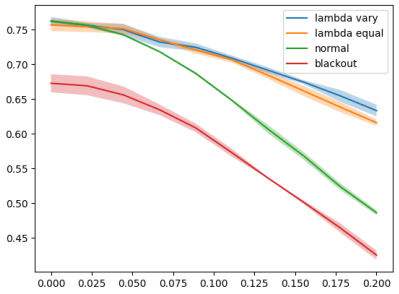
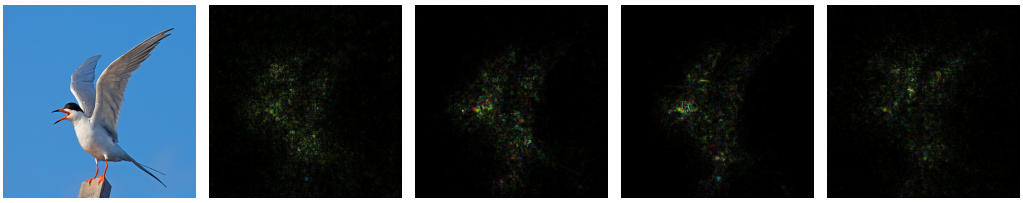



Figure 2: Test robust accuracy for different values of $\epsilon$'s, for the CUB dataset; the error bands here represent standard deviation.

## 4.2 Interpretability comparison

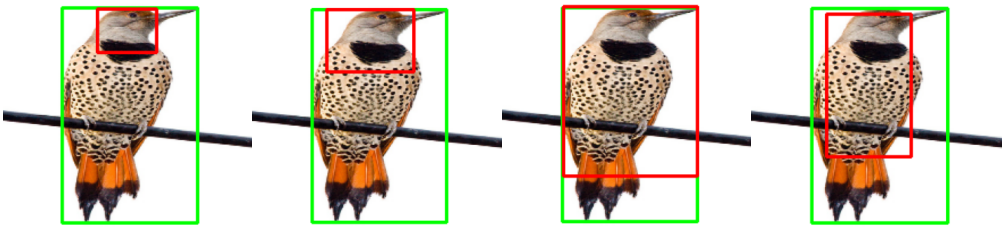
We compare the interpretability of the trained models qualitatively and quantitatively.

**Qualitative results.** We plot the gradient-based saliency map [25] generated by the model trained by each algorithm on a few bird images in the CUB dataset. We can see from Figure 3 that the



(a) original (b) (c) (d) (e)

Figure 3: Gradient-based saliency maps of a bird image (a) generated by models trained using different objectives: (b) STANDARD; (c) $\lambda$-EQUAL; (d) $\lambda$-VARY; (e) BLACKOUT.



(a) (b) (c) (d)

Figure 4: Localization results on the bird image in Figure 1, generated by models trained using different objectives: (a) STANDARD; (b) $\lambda$-EQUAL; (c) $\lambda$-VARY; (d) BLACKOUT. Here for each model, we extract a bounding box (shown in red) from its gradient-based saliency map.

saliency map of the STANDARD and BLACKOUT is dispersed and clearly not focusing on the bird body. The saliency map of $\lambda$-EQUAL is doing better than STANDARD and finally the model trained by $\lambda$-VARY even highlights subtle parts such as beaks and legs with the complete shape of bird.

**Quantitative results.** To quantitatively measure the interpretability of the gradient-based saliency maps output by different models, we extract bounding boxes from them and evaluate them in two ways: first, we use the saliency metric proposed in [2]; second, we compare the extracted bounding boxes with the ground truth bounding boxes using localization accuracy [7].

To generate a bounding box from a saliency map, we binarize the image by thresholding, and output the tightest rectangular box that contains the pixels whose grayscale is above the threshold.

**Saliency metric:** We follow [2] to perform the following calculation: after generating a bounding box, crop the corresponding region from the original image and pass it into the network to make prediction. The saliency metric of [2] is defined as: $s(a, p) = \log(a) - \log(p)$, where $a = \max(0.05, \hat{a})$, and $\hat{a}$ is the area fraction of the bounding box, and $p$ is the model's predictive probability for the correct label. The lower value the saliency metric the better. Table 1 shows the test saliency metric of models trained by all methods, where we can see that $\lambda$-VARY outperforms the baselines.

| STANDARD | BLACKOUT | $\lambda$-EQUAL | $\lambda$-VARY |
|---|---|---|---|
| $0.466 \pm 0.047$ | $0.396 \pm 0.033$ | $0.343 \pm 0.02$ | $\mathbf{0.283 \pm 0.03}$ |

Table 1: Saliency metric comparison among the evaluated methods on the CUB dataset.

**Localization accuracy:** The localization accuracy is defined as the fraction of examples where the model prediction is correct and the generated bounding box has intersection over union (IOU) value of $\geq 0.5$ with the ground truth bounding box. Table 1 shows the test localization accuracy of models trained by all methods, where we can see that $\lambda$-VARY outperforms the baselines.

| STANDARD | BLACKOUT | $\lambda$-EQUAL | $\lambda$-VARY |
|---|---|---|---|
| $0.236 \pm 0.02$ | $0.30 \pm 0.021$ | $0.30 \pm 0.169$ | $\mathbf{0.343 \pm 0.012}$ |

Table 2: Localization accuracy comparison among the evaluated methods on the CUB dataset.

## References


[1] Mark Craven and Jude W Shavlik. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*, pages 24–30, 1996.

[2] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6967–6976, 2017.

[3] Sanjoy Dasgupta, Akansha Dey, Nicholas Roberts, and Sivan Sabato. Learning from discriminative feature feedback. In *Advances in Neural Information Processing Systems*, pages 3955–3963, 2018.

[4] Jeff Donahue and Kristen Grauman. Annotator rationales for visual recognition. In *2011 International Conference on Computer Vision*, pages 1395–1402. IEEE, 2011.

[5] Harris Drucker and Yann Le Cun. Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, 3(6):991–997, 1992.

[6] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Beomsu Kim, Junghoon Seo, and Taegyun Jeon. Bridging adversarial robustness and gradient interpretability. *arXiv preprint arXiv:1903.11626*, 2019.

[11] John Lambert, Ozan Sener, and Silvio Savarese. Deep learning under privileged information using heteroscedastic dropout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2018.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[13] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[14] Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Embedding human knowledge in deep neural network via attention map. *arXiv preprint arXiv:1905.03540*, 5, 2019.

[15] Christoph Molnar. *Interpretable Machine Learning*. 2019. `https://christophm.github.io/interpretable-ml-book/`.

[16] Adam Noack, Isaac Ahern, Dejing Dou, and Boyang Li. Does interpretability of neural networks imply adversarial robustness? *arXiv preprint arXiv:1912.03430*, 2019.

[17] Stefanos Poulis and Sanjoy Dasgupta. Learning with feature feedback: from theory to practice. In *Artificial Intelligence and Statistics*, pages 1104–1113, 2017.

[18] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017.

[19] Laura Rieger, Chandan Singh, W James Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. *arXiv preprint arXiv:1909.13584*, 2019.

[20] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[21] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

[22] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[23] Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

[24] Manali Sharma and Mustafa Bilgic. Learning with rationales for document classification. *Machine Learning*, 107(5):797–824, 2018.

[25] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[26] Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*, 2018.

[27] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[28] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[29] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[30] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

[31] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[32] Omar F Zaidan, Jason Eisner, and Christine Piatko. Machine learning with annotator rationales to reduce annotation cost. In *Proceedings of the NIPS* 2008 workshop on cost sensitive learning*, pages 260–267, 2008.

[33] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[34] Ye Zhang, Iain Marshall, and Byron C Wallace. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 795. NIH Public Access, 2016.

[35] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.