Doubly Right Object Recognition

Revant Teotia [*1] **Chengzhi Mao** [*1] **Carl Vondrick** [1]

Abstract

Existing deep neural networks are optimized to predict the right thing, yet they may rely on the wrong evidence. Using the wrong evidence for prediction undermines out-of-distribution generalization, underscoring the gap between machine perception and human perception. In this paper, we introduce an overlooked but important problem: "doubly right object recognition," which requires the model not only to predict the right outcome, but also to use the right reasons that are aligned with human perception. The existing benchmarks fail to learn or evaluate the doubly right object recognition task, because both the right reason and spurious correlations are predictive of the final outcome. Without additional supervision and annotation for what is the right reason for recognition, doubly right object recognition is impossible. To address this, we collect a dataset, which contains annotated right reasons that are aligned with human perception and train a fully interpretable model that only uses the attributes from our collected dataset for object prediction. Through empirical experiments, we demonstrate that our method can train models that are more likely to predict the right thing with the right reason, providing additional generalization ability on ObjectNet, and demonstrating zero-shot learning ability.

1. Introduction

Deep neural networks focus on predicting the right "what" in the image, yet they often ignore the correctness of "why" for the predictions (He et al., 2016; Simonyan & Zisserman, 2015; Dosovitskiy et al., 2021; Tan & Le, 2019). Using the right reason for the prediction is important, especially in critical applications, such as medical imaging (Singh et al., 2020; Ancona et al., 2018; Eitel & Ritter, 2019; Pereira et al., 2018). Existing methods investigate how to get the reasons that contribute to the models' prediction (Nguyen et al., 2016; Olah et al., 2017; Simonyan et al., 2014; Shrikumar et al., 2017; Zeiler & Fergus, 2014; Smilkov et al., 2017; Selvaraju et al., 2016). However, due to the spurious correlations in the dataset, not every evidence is right (Singla & Feizi, 2022). Existing classifiers have no knowledge to discriminate which evidence is right or not, which causes the model to often predict the right thing for the wrong reason. While using the wrong evidence can help in-distribution accuracy, generalization gets difficult when the distribution shifts (Mao et al., 2021).

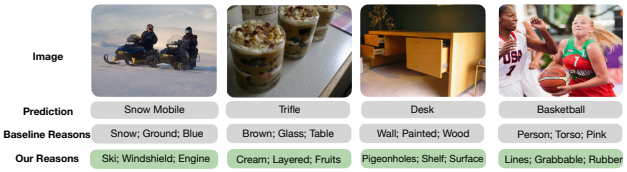

Figure 1. Doubly Right Object Recognition. Without the right inductive knowledge, the model often predicts the right thing for the wrong reasons. Our benchmark allows training doubly right classifiers that predict the right thing using the right reasons.

In this paper, we introduce the doubly right object recognition problem. As shown in figure 1, besides predicting the correct "what" is the image, we also desire the prediction to be based on the correct "why". We propose a dataset benchmark evaluating doubly right object recognition. Existing attribute datasets (Pham et al., 2021; Bau et al., 2017) contain untrimmed attributes, which include nuisance concepts that are spuriously correlated with the prediction. To address this, we collect a new attribute prediction dataset (section 2) that annotates the right attributes, which are aligned with human perception, for object classification. To ensure that the model only bases the prediction on the obtained reasons, our model first linearly transforms the learned representation to attributes that are aligned with human perception, and then learns a linear read-out layer using the transformed feature. The model's prediction is totally interpretable because it can only be based on the attribute features that correspond to a

[*]Equal contribution [1]Department of Computer Science, Columbia University, New York, NY, USA. Correspondence to: Revant Teotia <rt2819@columbia.edu>, Chengzhi Mao <cm3797@columbia.edu>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA.

semantic concept.

Empirical experiments and visualizations show that our approach can predict the right thing for the right reasons. On our annotated attributes of the validation set images, we obtain a recall@7 for the correct attribute of 68.78% when the prediction is correct. In addition, with the right prediction reason, we can improve generalization ability on the out-of-distribution benchmark as well as conducting zero-shot classification. We will release our data, code, and model.

The major contributions of our paper are: 1) we propose a new benchmark task "doubly right object recognition" and a new dataset that allows training and evaluation on our proposed task, 2) demonstrate the additional advantage that "doubly right object recognition" can improve the generalization in out-of-distribution generalization and zero-shot learning setting.

2. Doubly Right Object Recognition: Training Set and Benchmark

Relying on spurious attributes can harm generalization when the spurious correlations are changed in a new environment. However, without additional assumptions or constraints, the model will learn to use spurious correlations for the prediction. In other words, it is theoretically impossible to learn the right attribute for the prediction rather than the spurious ones without additional knowledge (Mao et al., 2022). Our goal is to use minimal human effort to enable the model to learn the right ones. For this, we first list the core attributes for each class, and then automatically collect images containing the attributes using Google image search.

Dataset collection details. We manually annotated a few descriptive attributes for each of the 40 object classes in our classification categories. Using our prior human knowledge, we ensured that we only annotate those visual attributes for an object that actually describe the object thus avoiding spurious attributes. For example, for 'bike' the annotated attributes are *wheel, seat, beam, handlebar, metal, etc.* Unlike general attributes like *tree, ground, road*, which could be spuriously correlated with 'bike' in images, our small set of causal attributes only describe the object and are not spurious. In total, we have 135 causal attributes with 6.2 attributes on average per object class.

Once we have the lists of causal attributes for object classes, we do image search over the internet to collect attribute training images. For example, we have *tire* attribute for 'bike' object class. We search 'bike tire' images over the internet and collect the top 50 image results. These images are then used to train attribute prediction model to predict the attribute *tire*. Similarly, for a 'car' too, we collect images by searching 'car tire' and add them to the *tire* attribute training set. This way we collect a total of 11,335 images for 135 attributes. Some of the example training images are shown in figure 4.

Benchmark Different images, even if they are from the same category, contain different sets of attributes. To validate the precision of the predicted reason from the model, we annotate each image in the validation set with attributes. We use Amazon Mechanical Turk to crowdsource the attribute annotation. We provide image-attribute pairs to the turkers and for each pair, we ask if the given attribute is applicable to the given object in the image or not. Note that, in the image-attribute pairs we show to turkers, we only use those causal attributes (from the lists we created earlier) that are useful for describing the object in the image. For example, when we give an image of a 'bike', we only ask for attributes (like *wheel, handlebar*) that describe a 'bike'. We do not ask turkers to annotate attributes like *headlight/windshield* for a 'bike' image. This way, we get a list of attributes for each image that actually describe the object in the image. We use annotations from 3 turkers for each image-attribute pair and take a majority vote to decide if the attribute is present in the image. Finally, we have 2000 images belonging to our 40 classification object categories annotated with causal attributes having an average of 5.8 attributes per image (some examples are shown in Fig. 4). We use this benchmark dataset to verify if our method is actually predicting right for the right reasons or is it using the wrong reasons to predict (see Fig. 3).

Evaluation metric. We use 3 attribute recall-based metrics to find if a classification model is able to predict the right reasons for object classification. 1. Average Recall@k: calculated as the fraction of annotated attributes in top-k predicted attributes for an image and averaged over all test set images. 2. *Strict DoublyRight score*: calculated as $\sum Recall_R/N$, where $Recall_R$ is attribute recall for rightly predicted image and N is the number of images in the test set. 3. *Relaxed DoublyRight score*: calculated as $(\sum Recall_R + \sum(1 - Recall_W))/N$ where $Recall_W$ is the attribute recall for wrongly predicted images. Both *DoublyRight score* metrics favor methods that predict both correct objects and their correct reasons. *Strict DoublyRight score* favors more accurate methods as it completely penalizes wrong predictions, while *Relaxed DoublyRight score* favors more interpretable methods because counting (1-recall) for wrongly predicted images tells that the method predicted wrong because it found wrong reasons in the image.

3. Experiments

We train an interpretable object classification model on our attribute prediction dataset that first learns to rotate the representation in a neural network into the space of interpretable attributes, and then train a linear classifier using the interpretable concepts. We first evaluate our method's perfor-

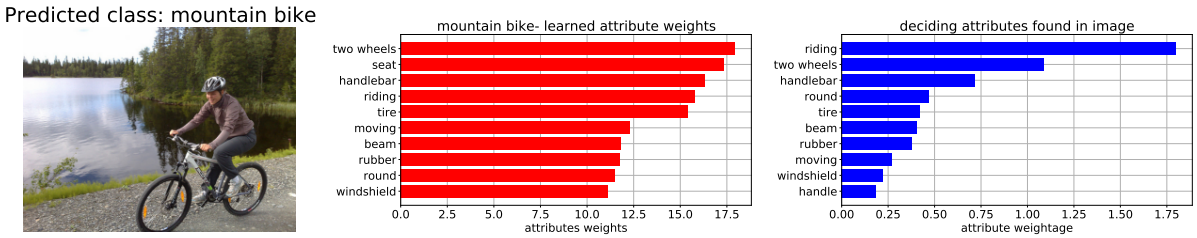

Figure 2. **Right reasons for right prediction** According to learned attribute weights by the model, a 'mountain bike' is described by attributes- *two wheels, seat, handlebar, tire, riding, beam*. The model predicts the above image as a 'mountain bike' because it finds these describing attributes in the image. The model provides explainable reason for its decision in terms of core visual attributes in the image.

Table 1. **Doubly Right evaluation.** High scores on the benchmark set show that our method predicts correct attributes as right reasons.

Method	Avg. Recall@7	Strict DoublyRight score	Relaxed DoublyRight score
DoublyRight	65.47%	63.04%	68.96%

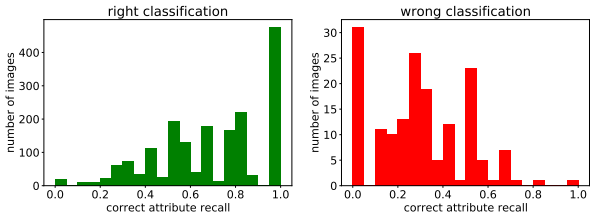

Figure 3. **Effectiveness of DoublyRight method.** We plot attribute recall@7 for both correct and incorrect classifications. For most of the correct classifications, the predicted top attributes are also correct and for the incorrectly classified images, the predicted attributes are also incorrect. It shows that our method makes right choices if the reasons (predicted attributes) are right and makes wrong choices if the reasons are wrong.

Table 2. **Generalization on ObjectNet images.** Our interpretable model using right reasons for prediction outperforms baseline ResNet50 by 2.6% in top-1 accuracy on out-of-distribution images.

Model	ObjectNet acc.	Imagenet acc.
ResNet50	49.98%	96.11%
DoublyRight	**52.58%**	91.84%

mance on our proposed doubly right object recognition task. We then validate on two established generalization tasks: out-of-distribution generalization and zero-shot learning, where we demonstrate that our approach improves generalization by capturing the right reason.

3.1. Interpretability

Once the object classification linear model is trained, we can analyze the learned attribute weights to find 'what does the model think about the object class?' By looking at the

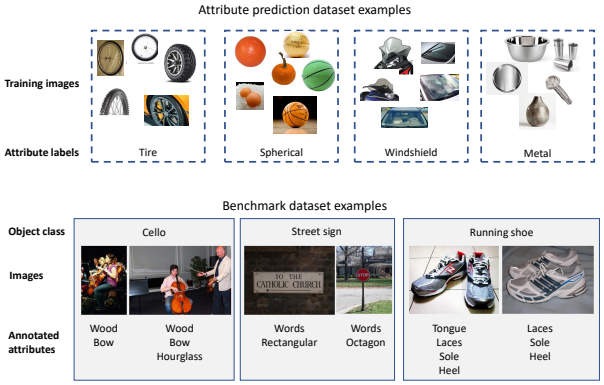

Figure 4. **Example images.** On top are some of the examples of images, collected using Google image search, used to train our attribute prediction model. Below are some examples of attribute annotation in our benchmark test set which is used to evaluate the effectiveness of our Doubly Right approach.

predicted attributes and the learned weights for the predicted object class, we can explain why the model has made its choice. As seen in figure 2, the model classifies the image as a 'mountain bike'. As we can see the learned attribute weights (in red) for 'mountain bike', the model thinks that a mountain bike image should have the following attributes- *two wheels, seat, handlebar, tire, riding, beam*. It predicts attributes like - *riding, two wheels, handlebar, round, beam* in the image. Since these attributes match the description of 'mountain bike' learned by the model weights, the model predicts the image is of 'mountain bike'. This way, the model gives its reasons for its decision in terms of the predicted attributes which makes the model interpretable and its predictions explainable.

Table 3. **Explainable zero-shot.** We provide explaining attributes (from training set attributes) for unseen classes and ask the classification model to classify images of the unseen class among the seen and unseen classes.

Unseen object class	Explaining attributes	Zero-shot accuracy
Green apple	*green, spherical, round, shiny*	76%
Bagel	*round, bread*	62%
Cheeseburger	*bread, sauce, meat, layered*	40%

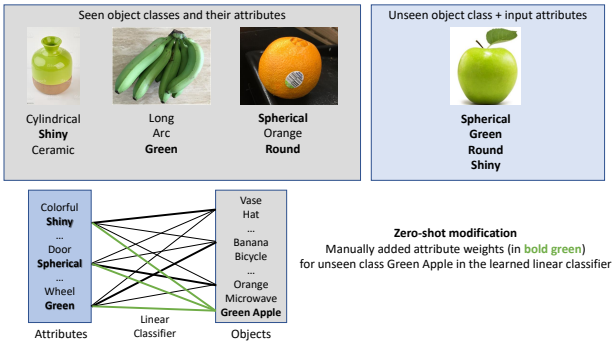

Figure 5. **Zero-shot adaptability.** We add new weights (shown in bold green lines) for attributes *Spherical, Green, Round, Shiny* to the classifier for the 'Green Apple' class and the classifier is now able to classify into one more class that is unseen during training.

Right reasons lead to right predictions. Wrong reasons lead to wrong predictions. Using the benchmark dataset of annotated attributes for the validation set images, we show that our model gives right predictions when it predicts right attributes and gives wrong predictions when it predicts wrong attributes. We plot attribute recall@7, calculated by finding the number of annotated attributes in top-7 predicted attributes, for both correctly and incorrectly classified images in figure 3. We find that, for most of the images that are correctly classified, the correct attribute recall is high. While for most of the images that are incorrectly classified, the correct attribute recall is low. We also evaluate our method on the evaluation metrics defined in section 2 and find that it performs well (table 1). This shows that our model is actually Doubly Right- gives right predictions when its reasons for the predictions are right.

3.2. Improved Generalization: ObjectNet Performance

Visual attributes of objects do not change much with changes in orientation, viewpoint and background. For example, a 'cup' would still be made of *ceramic* material, have a *cylindrical* shape and have a *handle* to hold even if it is placed upside down in an unusual surrounding like a bathroom. Since our model uses attributes for object classification, it is able to generalize on out-of-distribution dataset ObjectNet (Barbu et al., 2019) which has test images of objects with diverse viewpoints and backgrounds. We evaluate the performance of our method on 2,723 ObjectNet images belonging to 16 object classes that are common between ObjectNet and our 40 object classes. We found that our method outperforms baseline Resnet50 (trained to predict our 40 classes) by **2.6%** in terms of top-1 classification accuracy, thus showing better generalizability (figure 2). This shows that an interpretable model using right attributes for classification could help in generalizability for out-of-distribution images.

3.3. Zero-Shot Learning

As our object classification model is able to learn attributes for objects, we can manually provide attributes for unseen objects and our classification model can identify the object without even a single training example, thus having zero-shot learning ability. To enable our model to classify a new object class, we manually describe the new object class using our existing set of learned attributes, and add a new set of attribute weights for it to the existing model (see Fig. 5). For example, we manually provide attributes of a 'green apple' (an unseen category not in the 40 training classes) as a list of attributes = [*green, spherical, round, shiny*]. We then add a new weight vector for 'green apple' to the already trained weights of the classifier such that the weights of the attributes not in the above list are zero and the weights of the attributes present in the above list are equal to the mean of the weights of the top-7 attributes of the existing 40 categories. Thus the values in the new weight vector describe the attributes of 'green apple' numerically. We evaluated our approach on validation set images of a few of the unseen ImageNet classes (in table 3) and found that it performs well even with minimal information provided in terms of visual attributes. This shows that the in-built interpretability of our model could be easily adapted for zero-shot classification.

4. Conclusion

We study the problem of doubly right object recognition, where the classifier not only needs to classify the image correctly, but also needs to use the right reason to make the prediction. We collect a new dataset that allows the training and benchmarking on the doubly right prediction task. Our work is the first one that shows the importance of doubly right prediction, and connects models' better interpretation ability to improved generalization.

References

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sy21R9JAW.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Eitel, F. and Ritter, K. Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification. In Suzuki, K., Reyes, M., Syeda-Mahmood, T. F., Glocker, B., Wiest, R., Gur, Y., Greenspan, H., and Madabhushi, A. (eds.), *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support - Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings*, volume 11797 of *Lecture Notes in Computer Science*, pp. 3–11. Springer, 2019. doi: 10.1007/978-3-030-33850-3_1. URL https://doi.org/10.1007/978-3-030-33850-3_1.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Mao, C., Cha, A., Gupta, A., Wang, H., Yang, J., and Vondrick, C. Generative interventions for causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3947–3956, 2021.

Mao, C., Xia, K., Wang, J., Wang, H., Yang, J., Bareinboim, E., and Vondrick, C. Causal transportability for visual recognition. *arXiv preprint arXiv:2204.12363*, 2022.

Nguyen, A. M., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3387–3395, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/5d79099fcdf499f12b79770834c0164a-Abstract.html.

Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Pereira, S., Meier, R., Alves, V., Reyes, M., and Silva, C. A. Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment. In Stoyanov, D., Taylor, Z., Kia, S. M., Oguz, I., Reyes, M., Martel, A. L., Maier-Hein, L., Marquand, A. F., Duchesnay, E., Löfstedt, T., Landman, B. A., Cardoso, M. J., Silva, C. A., Pereira, S., and Meier, R. (eds.), *Understanding and Interpreting Machine Learning in Medical Image Computing Applications - First International Workshops MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings*, volume 11038 of *Lecture Notes in Computer Science*, pp. 106–114. Springer, 2018. doi: 10.1007/978-3-030-02628-8_12. URL https://doi.org/10.1007/978-3-030-02628-8_12.

Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., and Shrivastava, A. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13018–13028, June 2021.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL http://arxiv.org/abs/1610.02391.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings*

of Machine Learning Research, pp. 3145–3153. PMLR, 2017. URL http://proceedings.mlr.press/v70/shrikumar17a.html.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

Singh, A., Sengupta, S., and Lakshminarayanan, V. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.

Singla, S. and Feizi, S. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=XVPqLyNxSyh.

Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. URL http://arxiv.org/abs/1706.03825.

Tan, M. and Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pp. 818–833. Springer, 2014. doi: 10.1007/978-3-319-10590-1_53. URL https://doi.org/10.1007/978-3-319-10590-1_53.