Finding Spuriously Correlated Visual Attributes

Anonymous Authors[1]

Abstract

Deep neural models learn to use spurious features present in image datasets which hurts their out-of-distribution performance and makes them unreliable for critical application like medical imaging. To help develop robust models, it becomes essential to find spurious features in training datasets. Existing methods to find spurious features do not give any semantic meaning to the features and rely on human interpretation of the discovered correlated features to find if they are spurious or not. In this paper, we propose to first rotate the latent features into visual attributes and then learn correlation between the attributes and object classes by training a simple linear classifier. Correlated visual attributes are easily interpretable because they have well defined semantic meaning and makes it easier to find if they are spurious or not. Through visualizaions and experiments, we show how to find spurious visual attributes, their extent in existing dataset and failure mode examples showing negative impact of learned spurious correlations on out-of-distribution generalization.

1. Introduction

Existing image datasets contain both spurious and non-spurious features. Due to spurious correlations present in the training datasets, classification models learn to use the spurious features for predictions. While using the spurious features may help in-distribution accuracy, the classifier cannot generalize when the distribution shifts (Mao et al., 2021). Reliance on non-spurious features becomes specially important in critical applications, such as medical imaging (Singh et al., 2020; Ancona et al., 2018; Eitel & Ritter, 2019; Pereira et al., 2018). Thus it becomes important to find spurious features in datasets.

Existing methods of finding spurious correlations in image datasets (Singla et al., 2021; Mao et al., 2021; Wong et al., 2021b; Singla & Feizi, 2022) do not give semantic meaning to the discovered spurious features. (Wong et al., 2021b) first learn sparse linear classifier to find correlated neurons and then use images that activate the correlated neuron the most to ask human annotators to name the feature and say if it is spurious or not. (Singla et al., 2021) and (Singla & Feizi, 2022) develop a feature visualization method and ask humans if the visualized feature seem spurious or not. While (Mao et al., 2021) use generative models to find correlations between class labels and nuisance factors such as viewpoint and backgrounds. All of these methods rely on humans interpretation of the found correlated features to decide if they are spurious or not, and do not give any semantic meaning to the features on their own.

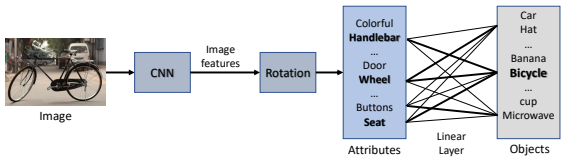

Figure 1. **Our framework.** We first rotate the representation in a neural network to the space of interpretable attributes, and then we train the linear classifier using our interpretable concepts in the latent space. The learned weights (highlighted in bold in the figure) give global level explanation of the learned attributes for an object class. For example, in this figure, the model has learned that visual attributes like {*handlebar, seat, wheel*} describe a 'bicycle'.

In this paper, we develop an interpretable object classification framework that we can use to find correlated visual attributes in image datasets. Our key insight is that the representations in deep models contains entangled spurious features and non-spurious features and if we can rotate the latent representation to the axis of interpretable visual attributes, we can disentangle the spurious attributes and access their association to the model output. To do this, our framework first learns a rotation to predict visual attributes present in the images and then uses those predicted attributes to classify the objects in the images. We can then analyze the learned attribute weights of the classifier model to find 'what visual attributes are associated by the model to an object category?'. Unlike existing methods (Singla et al., 2021; Wong et al., 2021b) where humans need to interpret

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

Table 1. **Spurious attributes.** Showing % of images in the imagenet validation set with at least one of the spurious attributes.

Object class	Spurious attributes	Images with spurious attributes
Snowmobile	*tree, snow, mountain*	96%
Rugby ball	*leg, torso, arm, person, hand*	80%
Street sign	*windowpane, building, tree, sidewalk, road*	92%
Tent	*mountain, tree, rock*	86%
Cowboy hat	*head, torso, arm, fence*	82%
Mountain bike	*tree, ground, leg, road*	88%
Traffic light	*building, street, road, sky*	94%
Airliner	*sky*	86%

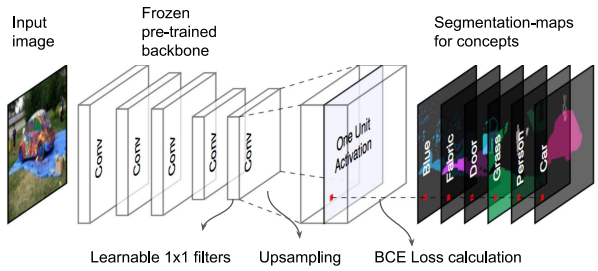

Figure 2. **Training attribute prediction model on Broden dataset:** We use segmentation maps of visual concepts from the Broden Dataset (Bau et al., 2017). We first use an Imagenet (Russakovsky et al., 2015) pretrained backbone (ResNet50 (He et al., 2015)) to extract feature maps of an input image. Then a 1x1 convolution filter is learned for each visual concept which operates on the feature penultimate layer resnet feature maps of input image to get concept activation map. The concept activation map is then upsampled to match the annotated segmentation map and used to calculate cross-entropy loss.

the meaning of a feature through most activating images or heat maps to figure out if the feature is spurious or not, our framework gives easily interpretable visual attributes with defined semantic meanings which makes it easier to decide if it is spurious or not. For example, our framework gives 'snow' attribute to be correlated with 'snowmobile' class, while existing methods would highlight snowy regions of 'snowmobile' images and leave it on human to interpret its meaning.

Through experiments and visualizations we show that our approach is effective in finding spuriously correlated features in terms of easily interpretable visual attributes. We use Broden (Bau et al., 2017) attributes to train our framework to classify ImageNet (Russakovsky et al., 2015) objects. We analyze the learned weights of the classifier (figure 3) to find spurious attributes and use the attribute predictions to find that 80%-90% of the ImageNet validation set images have spurious visual attributes (table 1). We also show a few failure model examples (figure 4) to reinforce the fact that learned spurious correlations by deep neural networks hurt their out-of-distribution generalization.

2. Framework

Deep models make predictions using the latent features. However, the features are black-box and often do not correspond to an interpretable concept. Our method first learns to rotate the representation in a neural network into the space of interpretable attributes, and then retrains the classifier using the interpretable concepts. We show our model pipeline in Fig. 1.

Attribute prediction. We interpret the latent representation in a deep model with our specified attributes. Given an image x, we first use the given CNN model F to extract its features $F(x)$ as f feature maps of size $H_f \times W_f$. We then learn to rotate the feature space $F(x)$ to find principal directions that correspond to interpretable visual attributes in the image. To do this, we learn a rotation model R which has A pointwise convolution filters operating on image features $F(x)$ to produce $(a_1, a_2, ..., a_A)$ attribute maps of size $H_f \times W_f$. Each attribute map a_i corresponds to an interpretable visual attribute. Finally, we average pool the attribute maps to get an A dimensional attribute feature vector $R(F(x)) \in \mathbb{R}^A$, such that each dimension of the vector corresponds to a human interpretable visual attribute. A simple object classifier could then be trained with these attribute features as inputs to make its prediction interpretable.

Object classification using attributes To make our classification method interpretable, we learn a simple linear model C that takes the predicted attribute vector $R(F(x))$ as input and predicts the object class among O object classes. We also softmax the attribute vector with a learnable temperature so that the input to the classification model could be treated as a probability distribution over attributes. To make the model weights more interpretable, we use Elastic-Net (Wong et al., 2021a) regularization so that its learned weights become sparse and only those attributes have some non-zero significant weights that are essential to describe

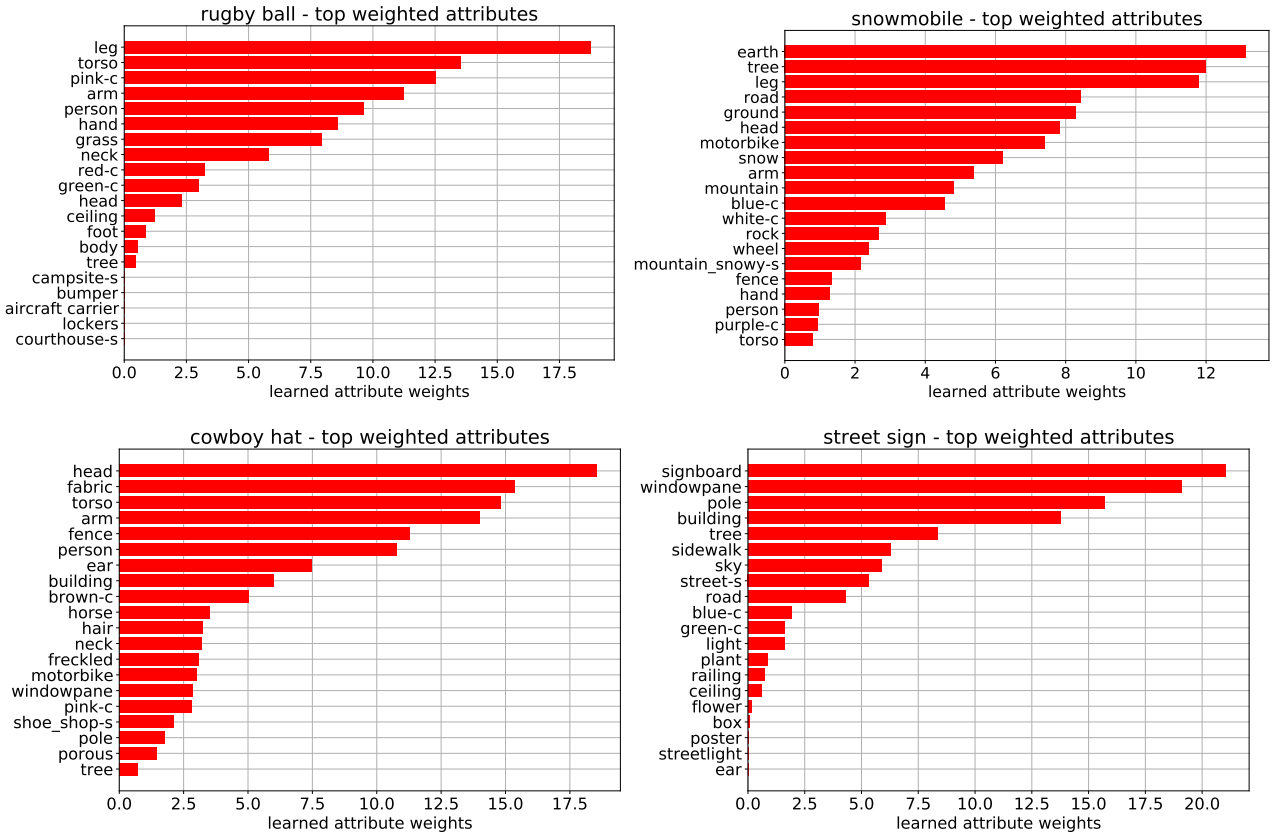

Figure 3. **Analyzing learned weights to find spurious attributes in Imagenet.** By analyzing top weighted attributes in the learned object classification weights, we find the spurious attributes in the Imagenet dataset. Here we find that attributes like {*head, torso, leg, arm*} are in the top weighted attributes for 'rugby ball', {*tree, snow, mountain*} for 'snowmobile', {*head, torso, arm*} for 'cowboy hat' and {*building, tree, sidewalk*} for 'street sign'. Though these attributes do not describe these objects, they are present in the training images and are learned by the classifier because they are spuriously correlated with the object classes.

the object class. The learned linear model is a weight matrix W of C is of shape $A \times O$, such that the value of weight $W_{i,o}$ gives the importance of attribute a_i for predicting the object o. By combining the results of the attributes and the linear layer, attribute weights can be used to explain 'what the model thinks about an object class?'. For example, for 'car' object class, the learned weights of concepts like- wheel, metal, windshield, door, street - would have higher values than concepts like - fabric, indoor, wood. Using this information, we can say that the model thinks a car looks like it is made of metal, has wheels, a windshield, doors, and is usually found on the streets. As the feature attributes are fully interpretable, we can interpret which attributes contribute the most to the predicted category.

3. Finding spuriously correlated visual concepts in ImageNet

Spurious correlations in datasets are known to harm the generalization ability of the learned models. Without specifying the core attribute to rely on, existing datasets may contain spurious correlated attributes. As spurious correlated attributes are also associated with the final prediction, the model will also learn to use it. The top weighted attributes in the learned W of classifier C in our framework can be reviewed to find spuriously correlated attributes.

We start our analysis with 40 object classes that are common in both MS COCO (Lin et al., 2014) and ImageNet (Russakovsky et al., 2015) dataset. We select these object classes because they represent objects from our day-to-day lives and have diverse set of attributes. We will introduce how to train our pipeline on existing attribute dataset, and show the spurious correlations used in classifying imagenet.

Broden visual concepts to find spurious attributes. Broden (Bau et al., 2017) dataset is an established large scale dataset that provides supervision for attributes in the image. Broden contains 63,305 images and 1197 attribute segmentation maps per image. The attributes include different kinds of colors, textures, materials, parts, objects and

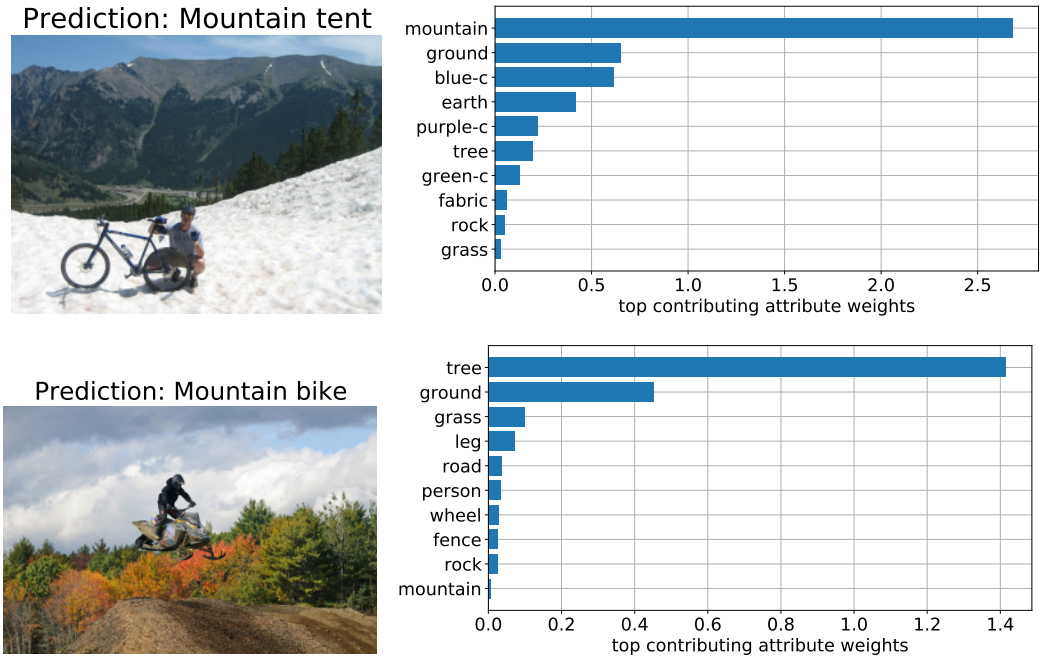

Figure 4. **Failure mode examples due to spurious correlations.** These examples show how spuriously correlated attributes can cause wrong object class prediction. Presence of *mountains, trees and rocks* causes the model to predict the top image of mountain bike as tent because these attributes are spuriously correlated with tent class (see table1). Similarly, presence of *tree, ground, leg and road* in the bottom image of snowmobile, combined with the absence of *snow and mountain*, causes the model to predict it as a mountain bike.

scenes. To train the pointwise convolution filters of the attribute rotation model R, we first compute the attribute maps $(a_1, a_2, ..., a_A)$ as described above in section 2. The images in Broden (Bau et al., 2017) dataset are of size 224×224 and the concept maps are of size 112×112. We use ResNet50 (He et al., 2015) to get image features maps of size 7×7 for the images. Then we learn 1197 1×1 filters for concepts to produce 7×7 concept activation maps which we then interpolate to 112×112 to match training concept maps for loss calculation. Refer to Figure 2 for details.We then use the predicted 1197 Broden attributes to train the linear object classification model using training images of the ImageNet dataset. We then manually probe the learned attribute weights for some of the object classes in the classifier to find their spurious attributes.

Visualizations and numerical results show that trained from Broden, the model's predictions are often based on spuriously visual attributes. For example in figure 3, we can see that attributes that contribute to the prediction of 'snowmobile' are 'snow,' 'tree' and 'mountain'. But none of these attributes describe a snowmobile. These attributes just happen to be present in the training images of 'snowmobile'. Similarly, we found such spurious attributes for a few more object classes and used the attribute predictions to find the number of images having such attributes. As seen in Tab 1, around 80-90% of images of the shown classes in the ImageNet validation set have at least one of the probed spurious attribute.

Relying on spurious attributes can harm model generalization when the spurious correlations are changed in a new environment. We show two example images from the ImageNet validation set in figure 4, one of bike over mountains and the other of a snowmobile in a snow-less forest track. In both of these cases, the model makes wrong predictions because the surrounding environments of the objects are slightly changed and the spuriously correlated attributes with the object classes are not present in the images anymore.

4. Conclusion

Existing computer vision dataset often contains undesired spurious correlations. Deep neural networks trained on these dataset will naturally learn those spurious correlations. Our work proposes a simple framework that allows to interpret the latent representation by rotating the representation to directions that directly correspond to semantic meanings. Our results demonstate rich spurious correlations leveraging existing Broden dataset as inductive bias for spurious correlations. Our findings highlight that existing deep models learn rich spurious correlations that are associated with the task output.

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.

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.

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. *arXiv preprint arXiv:1512.03385*, 2015.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

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.

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.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

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.

Singla, S., Nushi, B., Shah, S., Kamar, E., and Horvitz, E. Understanding failures of deep networks via robust feature extraction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*. Computer Vision Foundation / IEEE, 2021.

Wong, E., Santurkar, S., and Madry, A. Leveraging sparse linear layers for debuggable deep networks. *arXiv preprint arXiv:2105.04857*, 2021a.

Wong, E., Santurkar, S., and Madry, A. Leveraging sparse linear layers for debuggable deep networks. *arXiv preprint arXiv:2105.04857*, 2021b.