Frequency Shortcut Learning in Neural Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

The generalization of neural networks is harmed by shortcut learning: the use of simple non-semantic features may prevent the networks from learning deeper semantic and task-related cues. Existing studies focus mainly on explicit shortcuts, e.g. color patches and annotated text in images, that are visually detectable and may be removed. However, there exist implicit shortcuts determined by bias or superficial statistics in the data that neural networks can easily exploit. Mitigating the learning of implicit shortcuts is challenging due to the simplicity-bias and an intrinsic difficulty in identifying them. We empirically investigate shortcut learning in the frequency domain and propose a method to identify learned frequency shortcuts based on frequency removal. We found that frequency shortcuts often correspond to textures consisting of specific frequencies. We also investigate the influence of frequency shortcuts in Out-of-Distribution (OOD) tests.

1 Introduction

In classification, shortcuts are decision rules specific for a dataset, based on spurious correlations between data and ground truth, rather than on the correlation of semantic and task-related cues [1]. The training of NNs is affected by a biased learning behavior known as simplicity-bias, inducing the networks to learn simple but effective patterns, disregarding semantics related to the problem at hand [2]. For instance, in regression tasks, the networks are biased towards low-frequency components during learning [3] and favor fitting input data to a low-frequency function [4], because low frequencies carry most of the needed information to reconstruct signals and are easier to learn [5]. To improve the generalization and robustness of NNs, researchers investigate methods to mitigate shortcut learning. Most of them focus on shortcuts that are explicitly observable [6, 7] which are used to augment other classes, thus reducing their discriminative power. However, data can contain implicit shortcuts that are not always observable, e.g. spurious texture [8] or superficial statistics [9], which contribute to high performance on IID datasets but hamper generalization to OOD datasets. These are more difficult to identify.

Shortcut learning [1] is caused by a simplicity-bias of NNs [2]. Models trained for regression tasks tend to learn low-frequency components [3]. Classification models, instead, tend to learn spurious correlations between input images and ground truth labels to achieve classification with the least effort, e.g. a text manually embedded in images of a certain class [10], which the network uses as the most discriminant feature for classification. The presence of the text in images of other classes as well as the absence of the text in those of the intended class lead to incorrect predictions.

Avoiding shortcut learning is a promising way to improve the generalization of NNs. In [6], the authors identify the shortcuts for a class and embed them in another class during training. Identifying implicit shortcuts in the training data is much more difficult than identifying explicit visible shortcuts. Regularization was used to decouple feature learning dynamics, forcing NNs to learn from more features instead of a subset of shortcut features [11]. In [12], the authors evaluate a shortcut degree of

Submitted to Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

learned features, inspecting the gradient at early stage and at the end of the training. An auxiliary network with low capacity was applied to measure the shortcut degree of images in [7]. Implicit Feature Modification with adversarial perturbations was proposed to reduce learning shortcut features in contrastive learning [13]. The focus on mitigation of shortcut learning in image classification is limited to synthetic and visible shortcuts, rather than those implicitly present in the data [6, 7].

In this work, we empirically investigate shortcut learning in classification from a Fourier perspective, and show that simplicity-biased learning results in frequency-biased learning. NNs can reach their objective by finding implicit shortcut solutions in the Fourier domain, which we can identify with an iterative method based on the evaluation of the contribution of single frequencies to classification. We show that the identified shortcuts directly reveal the texture-bias of NNs.

2 Identifying frequency shortcuts

We investigate shortcut learning in the Fourier domain by an iterative method (see Algorithm 1) that evaluates the contribution of frequencies to classification results. In each iteration, the contribution of a single frequency is evaluated, and the frequency is discarded from subsequent iterations if it does not affect classification. At each iteration, the contribution of a certain frequency component is evaluated by measuring the performance degradation on the classification results induced by filtering it out from the test images. If the degradation is above a certain threshold, the specific frequency is deemed relevant. The threshold is dynamically updated based on the predictions after a frequency is removed. We do not use a fixed threshold as removing low frequencies may already reach the threshold, and subsequently, all high frequencies are retained, and thus every frequency removed early is considered unimportant. We use a margin m to control performance degradation and decrease the comparison threshold. The higher the value of m, the easier the frequencies are removed and thus more degradation of performance is allowed. The final outcome is a dominant frequency map (DFM), which shows the frequencies contributing the most to the classification of a specific class. Our hypothesis for frequency shortcut is that a small set of frequencies cannot represent properly task-related features, such as object shape. Thus, a DFM containing few frequencies that guarantee high classification performance is an indication of a learned frequency shortcut.

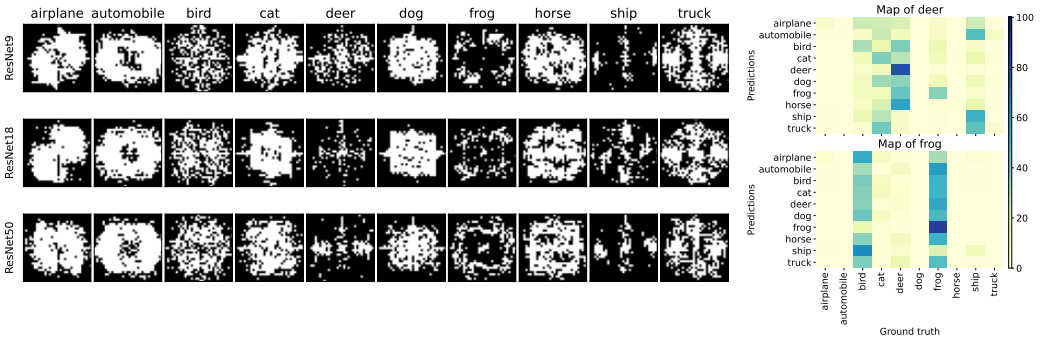
Algorithm 1 Culling irrelevant frequencies

Require: Test model M, Dataset D, Exploration class c, Predefined margin m, Frequency set ranked by importance R,
Ensure: Dominant frequency map M_c of class c

```
for (u, v) in R do
    M_c[u, v] = 1
end for
for (u, v) in R do
    Prediction_correct = 0
    for X_j, L_j in D_c do
        F'(X_j) = F(X_j) − (u, v)
        X'_j = F^{-1}(F'(X_j))
        Y_predict = M(X'_j)
        if Y_predict = L_j then
            Prediction_correct += 1
        end if
    end for
    if Prediction_correct >= Prediction_best − m then    ▷ Margin controls performance degradation
        M_c[u, v] = 0
        Prediction_best = Prediction_correct    ▷ Threshold is updated if a frequency is removed
    end if
end for
```

Experiment setup. We train different models (several ResNet and VGG configurations) on the CIFAR-10 dataset [14]. We inspect the trained models and the relevance of the frequency components learned during training for classification using Algorithm 1 with different values of m to generate class-wise DFMs. Then, we filter the images of each class using the DFMs (in Fig. 1a) as masks in the Fourier domain to retain only the dominant frequencies, and compare the classification performance on the filtered images with that obtained when using the original images. We measure the performance degradation as $\Delta = (acc(D_O) - acc(D_F))/acc(D_O)$, where $acc(D_O)$ is the accuracy on the original test set and $acc(D_F)$ is the accuracy on the test images filtered with the DFMs. With the resulting maps, the models can maintain roughly 70% of the prediction performance, compared to the performance on the original images. We evaluate how frequency shortcuts affect the generalization of models on OOD datasets, by testing on the ImageNet-10 dataset [15], a subset of ImageNet with similar characteristics as CIFAR-10. We apply the (inverse) DFMs to generate test sets with contributing frequencies only or without them.

Class-wise frequency shortcuts result in texture-biased classification. In Fig. 1a and Fig. A.1, we show the DFMs obtained for different architectures for each class in the CIFAR-10 dataset. We observed that, for instance, the maps of dominant frequency learned by ResNet-18 for the classes



(a) (b)

Figure 1: (a) Dominant frequency maps for the classes in CIFAR-10, and (b) confusion matrices when ResNet18 tested on CIFAR-10 by filtering images with the maps of the classes 'deer' and 'frog'.

Figure 2: A network trained on CIFAR-10 uses non-semantic features for classification. Correct predictions of '*frog*' images, even when only the texture is considered, indicate shortcuts learned during training. Extracting similar textures from images of '*dog*' results in false predictions as '*frog*'.

'deer' and 'frog' contain less than 20% of all frequencies. However, about 80% of the samples of these two classes are predicted correctly despite the limited amount of frequency information.

In Fig. 2, example images from the class 'frog' and 'dog', correctly classified, are shown. Filtering these images using the DFM of the class 'frog' results in only-texture images, which are both classified as 'frog'. This indicates a bias of the classifier that associates the class 'frog' with certain texture patterns which we can identify using our Algorithm 1. The very specific texture patterns used for classification are not clearly visible by observing the original images or not directly associated with task-related semantic cues of the object to recognize. We filter the test images from CIFAR-10 using the DFMs of the class 'deer' and 'frog', and report the test results in the confusion matrices in Fig. 1b. They highlight a classification bias towards the classes concerned, indicating a misuse of specific sets of frequency as discriminative features to recognize samples from these classes. The dependency on specific frequency patterns to achieve classification is in line with the observation about texture bias in NNs reported in [8].

How do frequency shortcuts affect generalization to OOD datasets? We expect that the frequency shortcuts specifically learned for CIFAR-10 would not generalize to OOD tests on ImageNet-10, thus impacting negatively the classification accuracy. We test on ImageNet-10 the models that we have trained on CIFAR-10, in order to evaluate OOD generalization and its relation with frequency shortcuts. We carry out tests using the original ImageNet-10 images, and two filtered versions of them using the corresponding class-wise DFMs of the models trained on CIFAR-10. In the first filtered version, we only keep the contributing frequencies of each class (referred to as 'w/ df'), while the other version only includes the complementary remaining frequencies (referred to as 'w/o df') to test the case when shortcuts are explicitly removed.

In Table 1, we report numerical results regarding the percentage of contributing frequencies from the overall Fourier spectrum according to Algorithm 1, the corresponding degradation Δ of performance measure as accuracy drop achieved when using only frequency from the DFMs to classify images of the different classes and the accuracy of OOD tests. Classes with a low number of dominant frequencies, relatively small degradation and many false positive predictions are considered using frequency shortcuts.

Table 1: No. of frequencies and performance degradation of ResNet(s) on CIFAR-10, and OOD test results on ImageNet-10. For OOD tests, we report accuracy on the original dataset, on images with the dominant frequencies removed (w/o df), and with only dominant frequencies retained (w/ df).

CIFAR-10: details of dominant frequency maps

model		airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
ResNet9	% freqs	42.38	51.76	29.88	43.55	23.93	38.48	16.6	46.58	10.06	41.76
	Δ	24.41	22.52	25.22	23.42	22.76	19.58	26.58	22.23	20.53	31.96
ResNet18	% freqs	44.82	54.79	35.94	38.57	10.94	40.04	15.43	45.70	19.73	40.33
	Δ	22.49	21.89	23.59	22.24	15.94	19.35	13.07	21.39	22.15	24.97
ResNet50	% freqs	46.97	55.96	38.96	42.48	15.23	32.23	22.95	44.43	12.69	34.77
	Δ	25.61	20.84	22.92	20.30	12.86	23.31	23.07	29.57	7.02	27.54

ImageNet-10: OOD tests

model		airliner	wagon	humming bird	Siamese cat	ox	golden retriever	tailed frog	zebra	container ship	tailer truck
ResNet9	acc	89.54	82.54	55.85	59.46	31.23	65.77	84.77	17.62	82.15	74.08
	acc w/o df	42.69	3.23	22.08	24.54	33.92	1.77	87.00	2.31	76.00	0.46
	acc w/ df	62.15	44.69	42.92	39.23	36.62	55.46	17.69	9.31	23.62	57.62
ResNet18	acc	91.31	84.00	54.62	69.31	35.46	70.23	84.69	20.08	86.00	76.23
	acc w/o df	79.8	16.1	32.8	42.2	55.69	12.1	85.1	7.3	53.7	2.9
	acc w/ df	74.31	54.77	54.08	47.46	8.54	59.54	4.92	15.31	31.38	63.69
ResNet50	acc	90.23	81.77	54.15	59.39	27.77	72.46	81.69	26.54	82.31	74.54
	acc w/o df	58.46	18.08	37.38	36.31	41.54	2.23	81.08	5.85	61.08	9.54
	acc w/ df	71.69	54.08	54.54	46.00	6.92	51.38	18.92	14.23	39.85	64.92

We observed a considerable drop in performance, especially on the class 'ox', which corresponds to class 'deer' (for which we identified as shortcut) according to [15]. For images of the classes 'ox' and 'tailed frog' of the ImageNet-10 dataset, corresponding to 'deer' and 'frog' of CIFAR-10, the use of the dominant frequency only to perform classification results in a considerable drop of performance (for ResNet18 and ResNet50 specifically), worse than random, so generalizing poorly. Interestingly, when forcing the classifier to ignore the shortcut learned on the CIFAR-10 dataset, namely filtering out the frequency contained in the DFMs of the classes with identified shortcuts, the results on the 'ox' and 'tailed frog' classes sharply increase. This indicates that the frequency shortcuts learned on the CIFAR-10 dataset no longer existing in the OOD dataset affect the extraction of semantic and task-related cues and have a negative impact on generalization.

Does higher model capacity mitigate frequency shortcut learning? In Fig. 1a, we observe that different models have similar frequency biases for classes 'deer', 'frog' and 'ship'. However, the values of Δ in Table 1 indicate that the models learn slightly different frequency shortcuts. ResNet50 uses more frequencies for classes 'deer' and 'frog' than ResNet18, showing a slight mitigation of learned frequency shortcuts as the capacity of the model increases. However, ResNet50 uses fewer frequencies to perform classification for class 'ship' with small performance degradation, intensifying the learning of a different frequency shortcut. On the contrary, ResNet9 learns to use a similar amount of frequencies to ResNet50 for the class 'ship' but with larger degradation of performance, indicating a less strong learned shortcut. It is worth noting that the number of false positive predictions of classes with frequency shortcuts decreases (see Fig. A.2 in Appendix A). We think that this is determined by the classification model paying attention to a larger set of frequencies to perform the classification for all classes. Although the experiments indicate a tendency of models with higher capacity to mitigate shortcut learning, we foresee the need for further investigation to better understand their relation. This would help to design specific procedures that increase generalization.

3 Conclusions

We empirically investigated frequency shortcut learning in deep neural networks and designed a method to identify shortcuts in the Fourier domain. We found that frequency shortcuts usually relate to textures consisting of a small set of frequencies that lead to biased predictions on certain classes. We show that frequency shortcut learning is common across different architectures, and investigate its relation with model capacity. Our preliminary results on CIFAR-10 and ImageNet-10 indicate that higher model capacity may contribute to mitigating the learning of frequency shortcuts, although deeper investigations are needed.

References

[1] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, pp. 665–673, nov 2020.

[2] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli, "The pitfalls of simplicity bias in neural networks," vol. 33, pp. 9573–9585, Curran Associates, Inc., 2020.

[3] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310, PMLR, 09–15 Jun 2019.

[4] Z.-Q. John Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," *Communications in Computational Physics*, vol. 28, no. 5, pp. 1746–1767, 2020.

[5] Z.-Q. J. Xu and H. Zhou, "Deep frequency principle towards understanding why deeper learning is faster," 2020.

[6] M. Nauta, R. Walsh, A. Dubowski, and C. Seifert, "Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis," *Diagnostics*, vol. 12, no. 1, 2022.

[7] N. Dagaev, B. D. Roads, X. Luo, D. N. Barry, K. R. Patil, and B. C. Love, "A too-good-to-be-true prior to reduce shortcut reliance," 2021.

[8] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.," in *International Conference on Learning Representations*, 2019.

[9] H. Wang, Z. He, Z. C. Lipton, and E. P. Xing, "Learning robust representations by projecting superficial statistics out," 2019.

[10] S. Lapuschkin, S. Wäldchen, A. Binder, *et al.*, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nat Commun 10*, vol. 1096, 2019.

[11] M. Pezeshki, S.-O. Kaba, Y. Bengio, A. Courville, D. Precup, and G. Lajoie, "Gradient starvation: A learning proclivity in neural networks," in *Advances in Neural Information Processing Systems* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), 2021.

[12] M. Du, V. Manjunatha, R. Jain, R. Deshpande, F. Dernoncourt, J. Gu, T. Sun, and X. Hu, "Towards interpreting and mitigating shortcut learning behavior of nlu models," 2021.

[13] J. Robinson, L. Sun, K. Yu, K. Batmanghelich, S. Jegelka, and S. Sra, "Can contrastive learning avoid shortcut solutions?," 2021.

[14] A. Krizhevsky, "Learning multiple layers of features from tiny images," tech. rep., 2009.

[15] H. Huang, X. Ma, S. M. Erfani, J. Bailey, and Y. Wang, "Unlearnable examples: Making personal data unexploitable," in *International Conference on Learning Representations*, 2021.

A Appendix

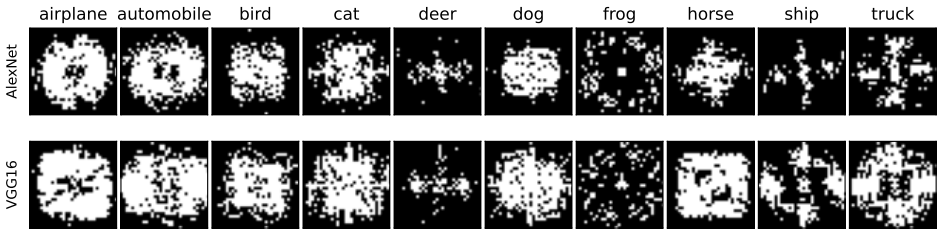


Figure A.1: Dominant frequency maps for the classes in CIFAR-10 (AlexNet and VGG16).

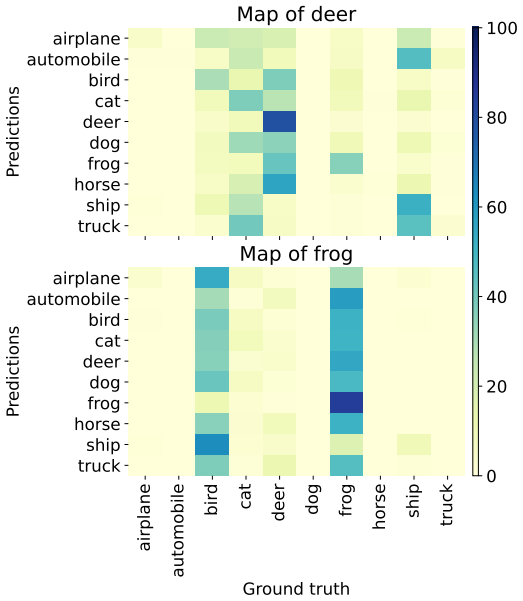


Figure A.2: Confusion matrices when ResNet50 tested on CIFAR-10 test set retaining only the dominant frequencies of the classes ‘deer’, and ‘frog’ respectively.