Redundant features can hurt robustness to distribution shift

Guillermo Ortiz-Jiménez [*1] **Apostolos Modas** [*1] **Seyed-Mohsen Moosavi-Dezfooli** [2] **Pascal Frossard** [1]

Abstract

In this work, we borrow tools from the field of adversarial robustness, and propose a new framework that permits to relate dataset features to the distance of samples to the decision boundary. Using this framework we identify the subspace of features used by CNNs to classify large-scale vision benchmarks, and reveal some intriguing aspects of their robustness to distributions shift. Specifically, by manipulating the frequency content in CIFAR-10 we show that the existence of redundant features on a dataset can harm the networks' robustness to distribution shifts. We demonstrate that completely erasing the redundant information from the training set can efficiently solve this problem. This paper is a short version of (Ortiz-Jimenez et al., 2020).

1. Introduction

Despite its tremendous success in controlled laboratory environments, deploying deep learning in the real world has turned to be a great challenge. One of the main reasons for this, is the extreme sensitivity of neural networks to small corruptions of the input data (Szegedy et al., 2014; Hendrycks & Dietterich, 2019) or to slight shifts on the testing distribution (Recht et al., 2019). One possible explanation for these two weaknesses is the over-reliance of deep networks on brittle and non-human aligned features of the training and validation sets, which might not be present in the real world distribution that they try to represent.

For this reason, it has become a pressing issue to identify which features do deep classifiers really use, as well as to describe the mechanisms that lead them to select certain features. In this sense, the decision boundary of a classifier encapsulates all the information required to interpret its decisions. However, due to its high complexity and dimensionality, obtaining a precise description of the decision boundary of a deep neural network is close to impossible.

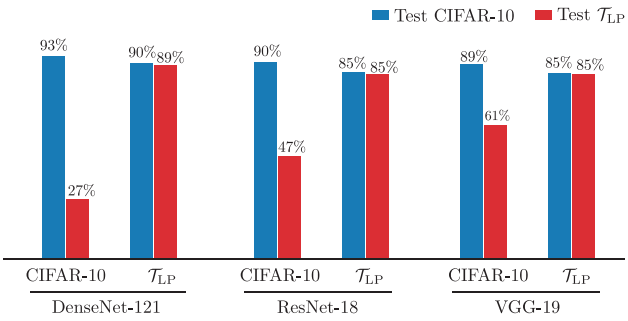


Figure 1. Test accuracy of CNNs trained and tested on combinations of CIFAR-10 and a low-pass version of CIFAR-10 ($\mathcal{T}_{\text{LP}}$).

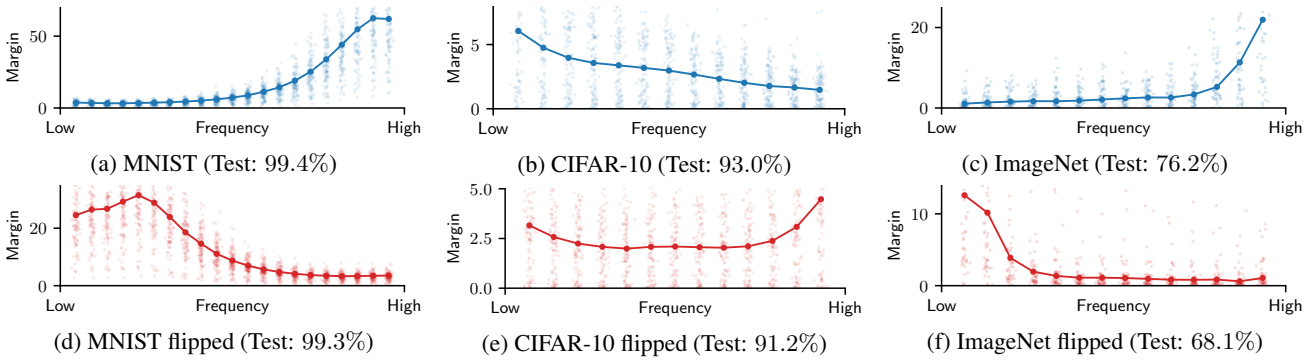
The main properties of these boundaries have been studied mainly from a robustness point of view (Fawzi et al., 2018; He et al., 2018; Ramamurthy et al., 2019). Interestingly, the extreme vulnerability of deep networks to adversarial perturbations (Szegedy et al., 2014; Goodfellow et al., 2015) implies that their boundaries still lie alarmingly close to any input sample. However, it seems that such perturbations are not irrelevant signals, but rather discriminative features of the training set (Jetley et al., 2018; Ilyas et al., 2019).

In this paper, we leverage this particular connection to develop a new framework that allows to discover the features used by a neural network. In particular, we use adversarial proxies to construct a local summary of the decision boundary of a deep classifier based on margin observations along a sequence of orthogonal directions. This framework permits to carefully shift the training distributions and measure the induced changes on the geometry of the boundary. This provides a new perspective on the relationship between margin and the discriminative features used by deep networks.

Furthermore, we can use this new tool to understand an intriguing aspect of the robustness of deep learning to distribution shifts; namely, its directionality. Let A and B be two data distributions differing only by a small shift. Surprisingly, the performance of a classifier trained on A and tested on B, can be significantly worse than the performance of the same classifier trained on B but tested on A.

[*]Equal contribution [1]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland [2]Eidgenössische Technische Hochschule Zürich (ETHZ), Switzerland. Correspondence to: Guillermo Ortiz-Jiménez <guillermo.ortizjimenez@epfl.ch>, Apostolos Modas <apostolos.modas@epfl.ch>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.



(a) MNIST (Test: 99.4%) (b) CIFAR-10 (Test: 93.0%) (c) ImageNet (Test: 76.2%)

(d) MNIST flipped (Test: 99.3%) (e) CIFAR-10 flipped (Test: 91.2%) (f) ImageNet flipped (Test: 68.1%)

Figure 2. Margin distribution of test samples in subspaces taken from the diagonal of the DCT (low to high frequencies). The thick line indicates the median values of the margin, and the shaded points represent its distribution. **Top**: (a) MNIST (LeNet) (Lecun et al., 1998), (b) CIFAR-10 (DenseNet-121) (Huang et al., 2017) and (c) ImageNet (ResNet-50) (He et al., 2016) **Bottom**: (d) MNIST (LeNet), (e) CIFAR-10 (DenseNet-121) and (f) ImageNet (ResNet-50) trained on frequency "flipped" versions of the standard datasets.

An example of the directionality of the robustness to distribution shift is illustrated in Fig. 1, where we can see a performance comparison of several convolutional neural networks (CNNs) trained and tested on combinations of the standard CIFAR-10 dataset and a slightly shifted version of it $\mathcal{T}_{\text{LP}}$, with low-pass filtered samples[1]. Clearly, the CIFAR-10 classifiers perform badly on $\mathcal{T}_{\text{LP}}$. Meanwhile, the classifiers trained on $\mathcal{T}_{\text{LP}}$ data achieve equal accuracy on both distributions, while still performing comparably to the original CIFAR-10 classifier[2].

In what follows, we will show that the reason for this strange behaviour is the strong inductive bias of neural networks to create boundaries only on discriminative features, regardless of their generalization performance. This means that, if a certain distribution contains too much unnecessary information, there is a high chance that a network will create boundaries around it. This will eventually harm its performance on samples from a slightly different distribution that does not contain this unnecessary information.

All in all, we demonstrate that one possible way to improve the robustness of deep learning to distribution shift is the identification and subsequent removal of all redundant features of a dataset to avoid their use by neural networks.

2. Margin and discriminative features

Let $F : \mathbb{R}^D \to \{1, \dots, L\}$ be a neural network, such that, for any input $\boldsymbol{x} \in \mathbb{R}^D$, it outputs a class label $y = F(\boldsymbol{x})$. Given a sub-region of the input space $\mathcal{S} \subseteq \mathbb{R}^D$, we define the ($\ell_2$) minimal adversarial perturbation of $\boldsymbol{x}$ in $\mathcal{S}$ as

$$\boldsymbol{\delta}_{\mathcal{S}}(\boldsymbol{x}) = \underset{\boldsymbol{\delta} \in \mathcal{S}}{\operatorname{argmin}} \|\boldsymbol{\delta}\|_2 \quad \text{s.t.} \quad F(\boldsymbol{x} + \boldsymbol{\delta}) \neq F(\boldsymbol{x}).$$

The magnitude $\|\boldsymbol{\delta}_{\mathcal{S}}(\boldsymbol{x})\|_2$ is the margin of $\boldsymbol{x}$ in $\mathcal{S}$.

In this work, we obtain a local summary of the decision boundary around a set of observation samples, by measuring their margin in a sequence of distinct subspaces $\{\mathcal{S}_j\}_{j=0}^{R-1}$. In practice, we use a subspace-constrained version of DeepFool (Moosavi-Dezfooli et al., 2016)[3] to approximate the margins in each $\mathcal{S}_j$. Having access to these new measurements, we can show that deep networks have a strong inductive bias towards invariance, which translates to small margins only along the direction of the discriminative features. As far as we know, we are the first to rigorously corroborate this property on state-of-the-art CNNs trained on standard computer vision datasets[4].

Let W, H, C denote the width, height, and number of channels of the images in those datasets, respectively. In our experiments we use the 2-dimensional discrete cosine transform (2D-DCT) (Ahmed et al., 1974) basis of size $H \times W$ to generate the observation subspaces. In particular, let $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{H \times W \times H \times W}$ denote the 2D-DCT generating tensor, such that $\text{vec}(\boldsymbol{\mathcal{D}}(i, j, :, :) \otimes \boldsymbol{I}_C)$ represents one basis element of the image space. We generate the subspaces by sampling $K \times K$ blocks from the diagonal of the DCT tensor using a sliding window with step-size T:

$$\mathcal{S}_j = \text{span}\{\, \text{vec}\left(\boldsymbol{\mathcal{D}}\left(j \cdot T + k, j \cdot T + k, :, :\right) \otimes \boldsymbol{I}_C\right) \\ k = 0, \dots, K-1\}.$$

In fact, previous studies on the robustness of CNNs have shown that these networks are more vulnerable to noise in certain frequency bands than others (Yin et al., 2019). Hence we can expect large differences in margin to appear in the

[1] The construction of this distribution is detailed in Sec. 2.

[2] A similar effect was shown on ImageNet (Yin et al., 2019), although the network was only tested on filtered data.

[3] We do not enforce the $[0, 1]^D$ box constraints on the adversarial images, as we are not interested in finding "plausible" adversarial perturbations, but in measuring the distance to the boundary.

[4] We provide an additional validation on a controlled synthetic example on Sec. A of the Appendix.

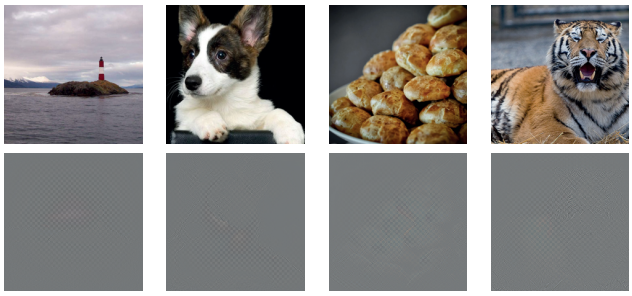


Figure 3. (**Top**) Original and (**bottom**) flipped ImageNet samples.

spectral domain. The sliding window on the diagonal of the DCT gives a good trade-off between visualization abilities in simple one-dimensional plots, and a diverse sampling of the spatial spectrum of natural images, with a well-defined gradient flowing from low to high frequencies.

The margin distribution of the evaluated test samples is presented in the top of Fig. 2. For MNIST and ImageNet, the networks present a strong invariance (i.e., high margin) along high frequency directions and small margin along low frequency ones. Notice, however, that for CIFAR-10 dataset the margin values are more uniformly distributed.

Towards verifying that these margin differences are associated to the data features exploited by the networks, we must first ensure that the directions of the observed invariance (large margin) are related to the features presented in the dataset, rather than being just an effect of the network itself. We do so, by first showing that the margin values follow the data representation, and later demonstrating that removing features in certain directions leads to a margin increase.

Adaptation to data representation Based on our observation that the margin tends to be small in low frequency directions and large in high frequency ones, we choose to carefully tweak the representation of the data, such that the low frequencies are swapped with the high frequencies. In practice, if $\mathfrak{D}$ denotes the forward DCT transform operator, the new image representation $\boldsymbol{x}'$ is expressed as

$$\boldsymbol{x}' = \mathfrak{D}^{-1}(\text{flip}(\mathfrak{D}(\boldsymbol{x}))),$$

where flip corresponds to one horizontal and one vertical flip of the DCT transformed image (see Fig. 3). Thus, if the direction of the resulting margin is strongly related to the data features, the constructed decision boundaries should also adapt to this new data representation, and the margin along the invariant directions (high frequencies) should swap with the margin of the discriminative ones (low frequencies). Informally speaking, the margin distribution should "flip".

We apply our framework on multiple networks trained on the "flipped" datasets, and the margin distribution is depicted at the bottom of Fig. 2. For both MNIST and ImageNet, the directions of the decision boundaries indeed *follow* the new data representation – although they are not an exact mirroring of the original representation. This indicates that the margin strongly depends on the data distribution, and it is not solely an effect of the network architecture. Note again that for CIFAR-10 the effect is not as obvious, due to the quite uniform distribution of the margin.

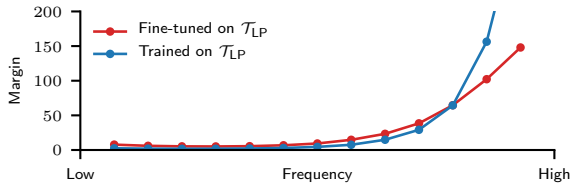


Figure 4. Median margin of test samples from CIFAR-10 for a DenseNet-121 (i) trained on CIFAR-10 and fine-tuned on $\mathcal{T}_{\text{LP}}$ (test: 90.79%), and (ii) trained on $\mathcal{T}_{\text{LP}}$ from scratch (test: 89.67%).

Removing features causes invariance The second property we need to verify is that the small margins reported in Fig. 2 do indeed correspond to directions containing discriminative features in the training set. For doing so, we exploit the flat margin of CIFAR-10 in Fig. 2b, and show that, by explicitly modifying the CIFAR-10 data, we can induce a high margin response in the measured curve in a set of selected directions.

In particular, we create a low-pass filtered version of CIFAR-10 ($\mathcal{T}_{\text{LP}}$), where we retain only the frequency components in a 16×16 square at the top left of the diagonal of the DCT-transformed images. This way we ensure that no training image has any energy, hence information, outside of this frequency subspace. The median margin[5] of CIFAR-10 test samples for a network trained on $\mathcal{T}_{\text{LP}}$ is illustrated in Fig. 4. Indeed, by eliminating the high frequency content, we have forced the network to become invariant along these directions. This clearly demonstrates that there existed discriminative features in the high frequency spectrum of CIFAR-10, and that by effectively removing these from all the samples, the inductive bias of training pushed the network to become invariant to them. All in all, we can say that small margins can only be identified in the discriminative directions used by the network.

Moreover, this effect can *also* be triggered during training. To show this, we start with the CIFAR-10 trained network studied in Fig. 2b and continue training it for a few more epochs with a small learning rate using only $\mathcal{T}_{\text{LP}}$. Fig. 4 shows the new median margins of this network. The fine-tuned network is again invariant on the high frequencies.

The elasticity to the modification of features during training gives a new perspective to the theory of catastrophic forgetting (McCloskey & Cohen, 1989), as it confirms that the

[5] We do not plot the full distribution to avoid clutter. The 5-perc. of the margin in the last subspace is 5.05.

Figure 5. (**Top**) CIFAR-10 and (**bottom**) $\mathcal{T}_{\text{LP}}$ samples.

decision boundaries of a neural network can only exist as long as the classifier is trained with the features that hold them together. In Sec. D of the Appendix we provide an additional experiment to further discuss this relation.

3. Redundant features and distribution shift

Now that we understand the relationship between discriminative features and margins, we can try to understand the intriguing generalization directionality exposed in Fig. 1. In this sense, note that the low-pass version of CIFAR-10 provides a good controllable model of distribution shift. Indeed, as we can see in Table 1, the average ℓ_2 distance between corresponding samples in CIFAR-10 and $\mathcal{T}_{\text{LP}}$ is very small (the standard ℓ_2 robustness threshold on CIFAR-10 is set at a norm of $\epsilon = 1$ (Madry et al., 2018)).

Moreover, we can clearly see in Fig. 5 that the semantic information on the samples of $\mathcal{T}_{\text{LP}}$ is perfectly retained and that the changes are barely perceptible. In fact, recall that the human visual system is mostly receptive to low frequencies (Gonzalez & Woods, 2017), and hence the human-assigned labels will necessarily correlate with the information in that frequency band. Nevertheless, there exist of course other features in natural images, which are not perceptible to humans, but that can be captured by neural networks. This is precisely what the low margins in the high frequency subspaces on Fig. 2b demonstrate.

However, it is clear that high frequency information is not necessary for the network to achieve good generalization. Actually, removing it as in Fig. 4, does not harm much the network's final performance. Yet, the existence of boundaries in this frequency band can heavily affect the robustness of the network to small distribution shifts in this spectral regime. Fig. 1 shows an example of such distribution shift, where we can interpret each sample of $\mathcal{T}_{\text{LP}}$ as a slightly perturbed version (in the high frequencies) of CIFAR-10.

The fact that the CIFAR-10 networks are vulnerable to high frequency perturbations means that they are exploiting some features in this frequency subspace (see Fig. 2b). Therefore, modifications of the input data in the high frequencies are likely to alter the networks' decisions. On the other hand, the networks trained directly on $\mathcal{T}_{\text{LP}}$ cannot exploit any information on the high frequencies to discriminate the training data. Hence, they do not create boundaries in this spectral regime, and treat samples coming from $\mathcal{T}_{\text{LP}}$ and CIFAR-10 equally.

Table 1. Average ℓ_2 distance between corresponding samples from CIFAR-10 and $\mathcal{T}_{\text{LP}}$ with an original pixel range of $[0, 1]$.

	TRAINING SET	TEST SET
AVG. ℓ_2 DIST.	0.011	0.025

4. Concluding remarks

In this paper, we proposed a new framework that permits to relate data features and margin along specific directions. We use this novel perspective to explain how redundant features on a training set make neural networks prone to suffer under distribution shifts, and show that removing these features from the training data can improve the networks' robustness and boost their invariance.

Note that augmenting the training data with certain corruptions can also improve robustness. In the case of CIFAR-10, introducing high frequency noise during training would probably achieve the same level of invariance. However, this would require many more training epochs and heavy fine-tuning. In general, we believe that directly removing redundant and imperceptible features from the training set can be an efficient preprocessing step to increase the robustness of deep networks to distribution shift. The vast literature in human perception and image compression could be a good inspiration to address this. Nevertheless, some sources of redundant features such as object positions, illumination, or color, are not related to perception. In those cases, data augmentation techniques combined with more sophisticated methods to address these shifts might be necessary.

Finally, we would like to draw the attention of the research community to an open question that stems from our observation, which is how neural networks select certain features. In particular, we still cannot explain why a CNN exploits high frequency information on CIFAR-10, if this is clearly not necessary for generalization. Furthermore, how is it possible that a small change in the position of the training samples, e.g., low-pass filtering them, can trigger such a dramatic change in the network geometry so that it becomes invariant to the high frequencies? We believe that answering these questions is necessary to understand the success (and limitations) of certain training regimes, like adversarial training, that slightly modify the position of the training samples to improve robustness.

Acknowledgments

We thank Maksym Andriushchenko, and Evangelos Alexiou for their fruitful discussions and feedback. This work has been partially supported by the CHIST-ERA program under Swiss NSF Grant 20CH21_180444, and partially by Google via a Postdoctoral Fellowship and a GCP Research Credit Award.

References

Ahmed, N., Natarajan, T., and Rao, K. R. Discrete Cosine Transform. *IEEE Transactions on Computers*, C-23(1): 90–93, 1974.

Fawzi, A., Moosavi-Dezfooli, S.-M., and Frossard, P. The Robustness of Deep Networks: A Geometrical Perspective. *IEEE Signal Processing Magazine*, 34(6):50–62, 2017.

Fawzi, A., Moosavi-Dezfooli, S.-M., Frossard, P., and Soatto, S. Empirical Study of the Topology and Geometry of Deep Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pp. 3762–3770. IEEE, 2018.

Gonzalez, R. C. and Woods, R. E. *Digital Image Processing*. Pearson, 4 edition edition, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR 2015)*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 770–778. IEEE, 2016.

He, W., Li, B., and Song, D. Decision Boundary Analysis of Adversarial Examples. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR 2019)*, 2019.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pp. 2261–2269. IEEE, 2017.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial Examples Are Not Bugs, They Are Features. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, pp. 125–136. Curran Associates, Inc., 2019.

Jetley, S., Lord, N. A., and Torr, P. H. S. With Friends Like These, Who Needs Adversaries? In *Advances in Neural Information Processing Systems (NeurIPS 2018)*, pp. 10749–10759. Curran Associates, Inc., 2018.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

McCloskey, M. and Cohen, N. J. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation*, 24: 109–165, 1989.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 2574–2582. IEEE, 2016.

Ortiz-Jimenez, G., Modas, A., Moosavi-Dezfooli, S.-M., and Frossard, P. Hold me tight! Influence of discriminative features on deep network boundaries. *arXiv:2002.06349*, 2020.

Ramamurthy, K. N., Varshney, K. R., and Mody, K. Topological Data Analysis of Decision Boundaries with Application to Model Selection. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pp. 5351–5360. PMLR, 2019.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pp. 5389–5400, Long Beach, CA, USA, 2019. PMLR.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR 2014)*, 2014.

Yin, D., Lopes, R. G., Shlens, J., Cubuk, E. D., and Gilmer, J. A Fourier Perspective on Model Robustness in Computer Vision. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, pp. 13255–13265. Curran Associates, Inc., 2019.

A. Validation of proposed framework on synthetic data

We want to show that neural networks only construct boundaries along discriminative features, and that they are invariant in every other direction[6]. To this end, we generate a balanced training set $\mathcal{T}_1(\epsilon, \sigma)$ by independently sampling N points $\boldsymbol{x}^{(i)} = \boldsymbol{U}(\boldsymbol{x}_1^{(i)} \oplus \boldsymbol{x}_2^{(i)})$ such that $\boldsymbol{x}_1^{(i)} = \epsilon y^{(i)}$ and $\boldsymbol{x}_2^{(i)} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{D-1})$, where $\oplus$ denotes the concatenation operator and $\epsilon > 0$ the feature size, and $D = 100$. The labels $y^{(i)}$ are uniformly sampled from $\{-1, +1\}$. The multiplication by a random orthonormal matrix $\boldsymbol{U} \in \mathrm{SO}(D)$ is performed to avoid any possible bias of the classifier towards the canonical basis. Note that this is a linearly separable dataset with a single discriminative feature parallel to $\boldsymbol{u}_1$ (i.e., first row of $\boldsymbol{U}$), and all other dimensions filled with non-discriminative noise.

To evaluate our hypothesis we train a heavily overparameterized multilayer perceptron (MLP) with 10 hidden layers of 500 neurons using SGD (test: 100.0%). Table 2 shows the margin statistics on the linearly separable direction $\boldsymbol{u}_1$; its orthogonal complement $\mathrm{span}\{\boldsymbol{u}_1\}^\perp$; a fixed random subspace of dimension S, $\mathcal{S}_{\text{rand}} \subset \mathbb{R}^D$; and a fixed random subspace of the same dimensionality, but orthogonal to $\boldsymbol{u}_1$, $\mathcal{S}_{\text{orth}} \subset \mathrm{span}\{\boldsymbol{u}_1\}^\perp$. From these values we can see that along the direction where the discriminative feature lies, the margin is much smaller than in any other direction. Therefore, we can see that the classification function of this network is only creating a boundary in $\boldsymbol{u}_1$ with median margin $\epsilon/2$, and that it is approximately invariant in $\mathrm{span}\{\boldsymbol{u}_1\}^\perp$.

Table 2. Margin statistics of an MLP trained on $\mathcal{T}_1(\epsilon = 5, \sigma = 1)$ along different directions ($N = 10,000$, $M = 1,000$, $S = 3$).

	$\boldsymbol{u}_1$	$\mathrm{span}\{\boldsymbol{u}_1\}^\perp$	$\mathcal{S}_{\text{ORTH}}$	$\mathcal{S}_{\text{RAND}}$
5-PERC.	1.74	4.85	30.68	17.21
MEDIAN	2.50	12.36	102.0	27.90
95-PERC.	3.22	31.60	229.5	80.61

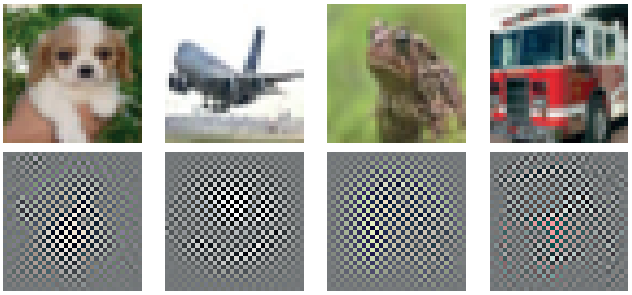
Comparing the margin values for $\mathcal{S}_{\text{orth}}$ and $\mathcal{S}_{\text{rand}}$ we see that, if the observation basis is not aligned with the features exploited by the network, the margin measurements might not be able to separate the small and large margin directions. Indeed, since $\mathcal{S}_{\text{orth}}$ is orthogonal to the only discriminative direction $\boldsymbol{u}_1$ we see that the margin values reported in this region are much higher than those reported in $\mathcal{S}_{\text{rand}}$. The reason for this is that the margin required to flip the label of a classifier in a randomly selected subspace is of the order of $\sqrt{S/D}$ with high probability (Fawzi et al., 2017), and hence the non-trivial correlation of a random subspace with the discriminative features will always hide the differences between small and large margin directions.

Finally, the fluctuations in the values and the fact that the classifier is not completely invariant on $\mathrm{span}\{\boldsymbol{u}_1\}^\perp$ might indicate that the network has built a complex boundary. However, similar fluctuations and finite values in $\mathrm{span}\{\boldsymbol{u}_1\}^\perp$ would also be expected, even if the model was linear by construction and was perfectly separating the training data.

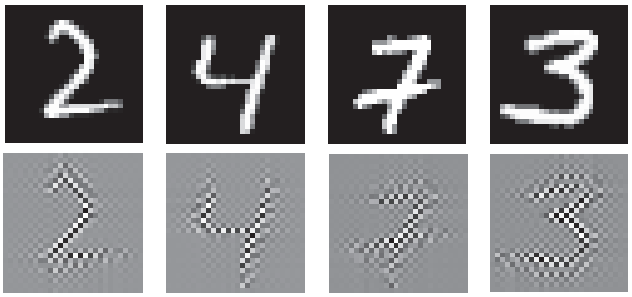
[6] This is indeed a desired property for any classification method, but note that for neural networks the existence of adversarial examples contests the idea of it being a reasonable assumption.

B. Examples of frequency "flipped" images

Figure 6 shows a few example images of the frequency "flipped" versions of the standard computer vision datasets.



(a) CIFAR-10



(b) MNIST

Figure 6. "Flipped" image examples. **Top** rows show original images and **bottom** rows the "flipped" versions.

C. MNSIT high-pass

We further validate our observation of Section 3 that small margin do indeed corresponds to directions containing discriminative features in the training set, but this time for a different dataset (MNIST), on a different network (ResNet-18), and using different discriminative features (high-frequency). In particular, we create a high-pass filtered version of MNIST (MNIST_{HP}), where we completely remove the frequency components in a 14×14 square at the top left of the diagonal of the DCT-transformed images (see Fig. 7 for some visual examples). This way we ensure that every pairwise connection between the training images (features) has zero components outside of this frequency subspace.

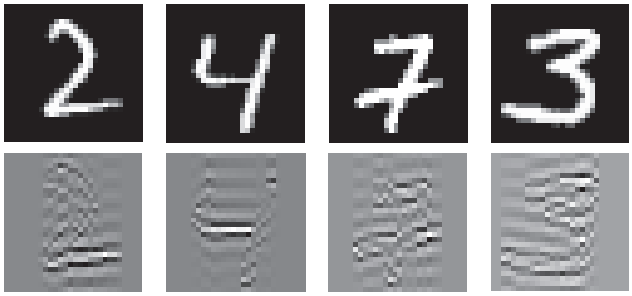


Figure 7. (**top**) MNIST and (**bottom**) high-pass MNIST examples. Notice that the digits can still be perceived, probably due to the contribution of the medium frequencies.

The margin distribution of $1,000$ MNIST test samples for a ResNet-18 trained on MNIST_{HP} is illustrated in Figure 8. Indeed, similarly to the observations on CIFAR-10, by eliminating the low frequency features, we have forced an increased margin along these directions, while forcing the network to focus on the previously unused high frequency features.

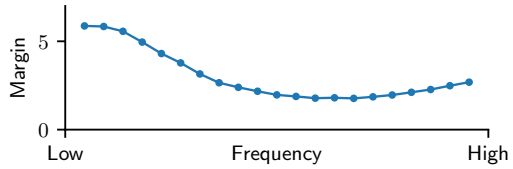


Figure 8. Median margin of test samples from MNIST for a ResNet-18 trained on MNIST_{HP} from scratch (test: 98.71%).

Finally, the directional robustness to this high-pass distribution shift is illustrated in Fig. 1, where we can see a performance comparison of several convolutional neural networks (CNNs) trained and tested on combinations of the standard MNIST dataset and a slightly shifted (high-passed) version of it, MNIST-HP. Similarly to what we observed in Fig. 1 for the CIFAR-10 dataset, MNIST classifiers do not generalize to MNIST-HP data. Meanwhile, a LeNet trained on MNIST-HP data achieves equal accuracy on both distributions, while still performing comparably to the original MNIST LeNet. Note though that, for the case of ResNet-18, we can see a slight performance drop when trained on MNIST-HP and evaluated on MNIST data.

In general, we cannot rule out the possible existence of a certain directional bias on these architectures. This is, they might be biased towards certain data representations in which the discriminative features align with certain frequency directions. We believe that this directional inductive bias might explain the drop in accuracy and margin values on the "flipped" distributions (see Fig. 2) and the perfomance decrease on the ResNet-18 when trained only on high-pass MNIST data.

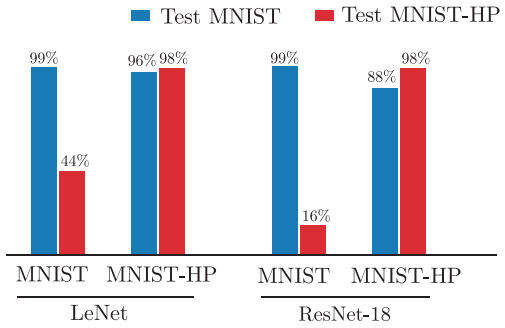


Figure 9. Test accuracy of CNNs trained and tested on combinations of MNIST and a high-pass version of MNIST (MNIST-HP).

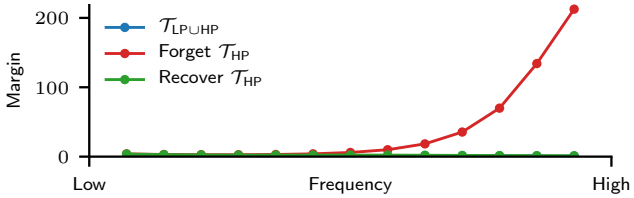
D. Connections to catastrophic forgetting

The elasticity to the modification of features during training gives a new perspective to the theory of catastrophic forgetting (McCloskey & Cohen, 1989), as it confirms that the decision boundaries of a neural network can only exist for as long as the classifier is trained with the samples (features) that hold them together. In particular, we demonstrate this by adding and removing points from a dataset such that its discriminative features are modified during training, and hence artificially causing an elastic response on the network.

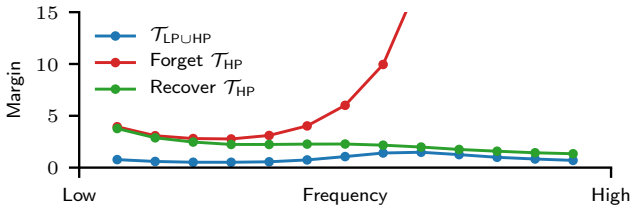
To this end, we train a DenseNet-121 on a new dataset $\mathcal{T}_{\text{LP}\cup\text{HP}} = \mathcal{T}_{\text{LP}} \cup \mathcal{T}_{\text{HP}}$ formed by the union of two filtered variants of CIFAR-10: $\mathcal{T}_{\text{LP}}$ is constructed by retaining only the frequency components in a 16×16 square at the top-left of of the DCT-transformed CIFAR-10 images (low-pass), while for $\mathcal{T}_{\text{HP}}$ only the frequency components in a 16×16 square at the bottom-right of the DCT (high-pass). This classifier has a test accuracy of 86.59% and 57.29% on $\mathcal{T}_{\text{LP}}$ and $\mathcal{T}_{\text{HP}}$, respectively. The median margin of $1,000$ $\mathcal{T}_{\text{LP}}$ test samples along different frequencies for this classifier is shown in blue in Figure 10. As expected, the classifier has picked features across the whole spectrum with the low frequency ones probably belonging to boundaries separating samples in $\mathcal{T}_{\text{LP}}$, and the high frequency ones separating samples from $\mathcal{T}_{\text{LP}}$ and $\mathcal{T}_{\text{HP}}$[7].

After this, we continue training the network with a linearly decaying learning rate (max. $\alpha = 0.05$) for another 30 epochs, but using only $\mathcal{T}_{\text{LP}}$, achieving a final test accuracy of 87.81% and 10.01% on $\mathcal{T}_{\text{LP}}$ and $\mathcal{T}_{\text{HP}}$, respectively. Again, Figure 10 shows in red the median margin along different frequencies on test samples from $\mathcal{T}_{\text{LP}}$. The new median margin is clearly invariant on the high frequencies – where $\mathcal{T}_{\text{LP}}$ has no discriminative features – and the classifier has completely *erased* the boundaries that it previously had in

[7] $\mathcal{T}_{\text{LP}}$ and $\mathcal{T}_{\text{HP}}$ have only discriminative features in the low-frequency and high-frequency part of the spectrum, respectively.



(a) Zoom-out axes for observing the general invariance.



(b) Zoom-in axes for a more detailed observation.

Figure 10. Median margin of $\mathcal{T}_{\text{LP}}$ test samples for a DenseNet-121. **Blue:** trained on $\mathcal{T}_{\text{LP}\cup\text{HP}}$; **Red:** after forgetting $\mathcal{T}_{\text{HP}}$; **Green:** after recovering $\mathcal{T}_{\text{HP}}$.

these regions, regardless of the fact that those boundaries did not harm the classification accuracy on $\mathcal{T}_{\text{LP}}$.

Finally, we investigate if the network is able to recover the forgotten decision boundaries that were used to classify $\mathcal{T}_{\text{HP}}$. We continue training the network ("forgotten" $\mathcal{T}_{\text{HP}}$) for another 30 epochs, but this time by using the whole $\mathcal{T}_{\text{LP}\cup\text{HP}}$. Now this classifier achieves a final test accuracy of 86.1% and 59.11% on $\mathcal{T}_{\text{LP}}$ and $\mathcal{T}_{\text{HP}}$ respectively, which are very close to the corresponding accuracies of the initial network trained from scratch on $\mathcal{T}_{\text{LP}\cup\text{HP}}$ (recall: 86.59% and 57.29%). The new median margin for this classifier is shown in green in Figure 10. As we can see by comparing the green to the blue curve, the decision boundaries along the high-frequency directions can be recovered quite successfully.