

IMPROVING ROBUSTNESS TO OUT-OF-DISTRIBUTION DATA BY FREQUENCY-BASED AUGMENTATION

Koki Mukai, Soichiro Kumano, and Toshihiko Yamasaki

The University of Tokyo

ABSTRACT

Although Convolutional Neural Networks (CNNs) have high accuracy in image recognition, they are vulnerable to adversarial examples and out-of-distribution data, and the difference from human recognition has been pointed out. In order to improve the robustness against out-of-distribution data, we present a frequency-based data augmentation technique that replaces the frequency components with other images of the same class. When the training data are CIFAR10 and the out-of-distribution data are SVHN, the Area Under Receiver Operating Characteristic (AUROC) curve of the model trained with the proposed method increases from 89.22% to 98.15%, and further increased to 98.59% when combined with another data augmentation method. Furthermore, we experimentally demonstrate that the robust model for out-of-distribution data uses a lot of high-frequency components of the image.

Index Terms— neural network, out-of-distribution, frequency, data augmentation

1. INTRODUCTION

In recent years, the accuracy of image recognition performance has been improving. In particular, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [1] in 2012, a competition for image recognition accuracy, saw a dramatic improvement in accuracy with the introduction of AlexNet [2], which uses Convolutional Neural Networks (CNNs). However, the vulnerability towards adversarial examples [3, 4, 5] and fooling images [6], overconfidence to out-of-distribution images have also been reported [7, 8, 9]. The generalization performance of such CNNs has been studied in relation to the frequency of the input images [10, 11, 12].

Wang et al. [10] argued that CNNs improve accuracy by using regions that are meaningful to humans and those with high-frequencies that cannot be perceived by humans. For this reason, they argued that images such as adversarial examples, which have relatively noisy high-frequency components, have a difference in recognition from humans. They also pointed out that there is a trade-off between the robustness to adversarial examples and the accuracy to normal images, and raised a question about the focusing on accuracy alone. In addition, Chen et al. [12] pointed out that the dif-

ference between humans and CNNs is that CNNs are sensitive to the amplitude component of the image, while humans are sensitive to the phase component. They also pointed out that adversarial examples and other examples show that the recognition of CNNs and humans are different because many changes are made to the amplitude component of an image. Based on their assumptions, they proposed an augmentation method, Amplitude-Phase Recombination (APR), to let CNNs pay more attention to the phase of the images.

For out-of-distribution detection, existing studies have proposed dedicated classifiers or attempted to improve the detection accuracy for pre-trained models in various ways [7, 8, 9, 13]. While these methods are effective, it would be beneficial if the robustness of the model itself to out-of-distribution data could be improved. In addition, existing data augmentation methods mainly focus on enhancing the accuracy, which may in turn sacrifice robustness. In this study, we propose a frequency-based data augmentation that improves robustness of the model to the out-of-distribution data in a confidence-based out-of-distribution detection method [8]. The contributions of our research are as follows.

- We propose a new frequency-domain data augmentation technique that enhances the robustness of the model to out-of-distribution data. Since it is a data augmentation method, it can be combined with other methods.
- We show that models with high robustness to out-of-distribution data pay more attention to high-frequency components of the images.

2. PROPOSED METHOD

To improve the robustness of CNNs, we propose a frequency-based data augmentation method, in which images are decomposed into low-frequency and high-frequency components and they are swapped with those of other images of the same class (here after, we call this procedure as Replacement of Frequency Component, RFC). While retaining the low and high-frequency information of each class, novel (augmented) images of each class can be generated. The flow of this data augmentation is shown in Figure 1 and the detailed procedure is described below.

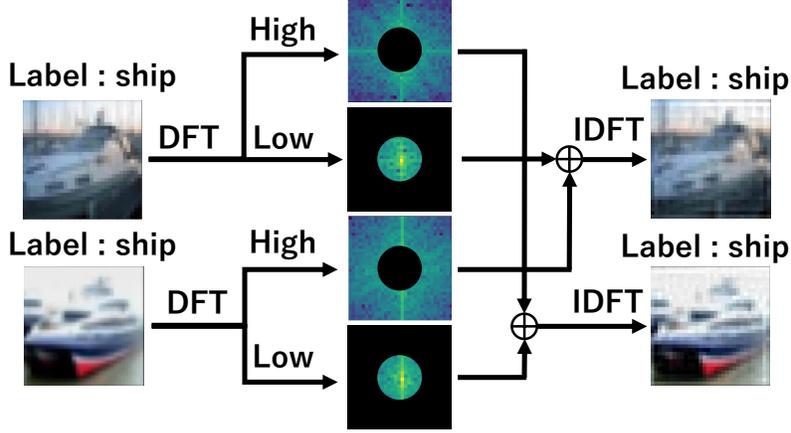


Fig. 1. The flow of RFC. Images taken from the same class are used and their frequency components are swapped.

Let \mathbf{x} be an image and let us denote its frequency component \mathbf{z} as $\mathbf{z} = \mathcal{F}(\mathbf{x})$ using the Discrete Fourier Transform (DFT) $\mathcal{F}(\cdot)$. We obtain the low-frequency and the high-frequency components of \mathbf{z} , \mathbf{z}_l and \mathbf{z}_h , respectively, by using the mask matrices to pass the low-frequency and high-frequency components M_l and M_h , respectively. With (c_i, c_j) as the indices of the frequency center (DC component) of M and using a radius r , M_l and M_h can be described as:

$$M_l(i, j) = \begin{cases} 1 & \left(\sqrt{(i - c_i)^2 + (j - c_j)^2} < r \right) \\ 0 & \left(\sqrt{(i - c_i)^2 + (j - c_j)^2} \geq r \right) \end{cases}, \quad (1)$$

$$M_h(i, j) = \begin{cases} 0 & \left(\sqrt{(i - c_i)^2 + (j - c_j)^2} < r \right) \\ 1 & \left(\sqrt{(i - c_i)^2 + (j - c_j)^2} \geq r \right) \end{cases}. \quad (2)$$

Then, \mathbf{z}_l and \mathbf{z}_h are defined as:

$$\mathbf{z}_l = M_l \otimes \mathbf{z}, \quad (3)$$

$$\mathbf{z}_h = M_h \otimes \mathbf{z}. \quad (4)$$

Here, \otimes is the product of each element. In the same way, let \mathbf{z}'_l and \mathbf{z}'_h be respectively the low and high-frequency components of \mathbf{x}' , a randomly selected image from the same class as \mathbf{x} . The frequency-component swapped images, \mathbf{x}_{mix} and \mathbf{x}'_{mix} , using the two images, \mathbf{x} and \mathbf{x}' , are obtained using the inverse Discrete Fourier transform (IDFT) $\mathcal{F}^{-1}(\cdot)$ as:

$$\mathbf{x}_{\text{mix}} = \mathcal{F}^{-1}(\mathbf{z}_l) + \mathcal{F}^{-1}(\mathbf{z}'_h), \quad (5)$$

$$\mathbf{x}'_{\text{mix}} = \mathcal{F}^{-1}(\mathbf{z}'_l) + \mathcal{F}^{-1}(\mathbf{z}_h). \quad (6)$$

Then, \mathbf{x}_{mix} and \mathbf{x}'_{mix} are added to the training data.

3. EXPERIMENTS

3.1. Experimental setup

In our experiments, the model is ResNet18 [14], the dataset is CIFAR10 [15], and the optimization method is Stochastic Gradient Descent (SGD), the learning rate started at 0.1, was multiplied by 0.2 at the 60th, 120th, 160th, and 190th epochs, and was continued until the 200th epoch. We use the basic data augmentation of RandomCrop and RandomHorizontalFlip as a baseline. We also employed CutMix [16], mixup [17], APR [12], RFC (proposed), and RFC+APR for comparison.

We used SVHN [18], LSUN [19], ImageNet [20], and CIFAR100 [15] as out-of-distribution data. For out-of-distribution detection, we use the Hendrycks's method [8]. It uses confidence scores of the model's output to distinguish whether the input image is in-distribution or out-of-distribution. By using True Positive Rate (TPR) and False Positive Rate (FPR) of this binary classification, we calculate the AUROC that is used as an evaluation metric.

3.2. Results of out-of-distribution detection

Table 1 shows the accuracy of each data augmentation method and the AUROC values when the models are trained using ResNet18 and CIFAR10 as the training dataset. The accuracy of the data augmentation methods for the spatial domain, such as mixup [17] and CutMix [16], is over 95%, while the baseline accuracy is 93.50%. However, the AUROC, which indicates the robustness to out-of-distribution data, drops for all out-of-distribution datasets, especially for SVHN, to around 83%, which is about 6% inferior to the baseline (89.22%). In comparison, the AUROC of our RFC is better than the baseline for all the out-of-distribution datasets, especially for SVHN, with 98.15%, an improvement of nearly 9%. In addi-

Table 1. Comparison of AUROC to out-of-distribution data (SVHN, LSUN, ImageNet, and CIFAR100) for models trained CIFAR10 on ResNet18 by each data augmentation. The best values are shown in bold.

method	Test acc.(%)	SVHN	LSUN	ImageNet	CIFAR100
baseline	93.50	89.22	88.61	82.68	84.88
CutMix [16]	95.00	83.74	87.26	79.24	83.18
mixup [17]	95.31	82.91	87.41	76.63	78.09
APR [12]	95.21	98.13	92.94	84.46	88.45
RFC (proposed)	94.07	98.15	91.03	83.04	85.83
RFC (proposed) + APR	94.71	98.59	93.82	85.17	89.02

tion, RFC+APR [12] yeilds the best AUROC for all the out-of-distribution datasets. The advantage of our proposed RFC is that, as demonstrated in this experiment, it can be combined with other data augmentation methods.

3.3. Accuracy for CIFAR10-C

In this section, we investigate the accuracy of each model in the CIFAR10-C dataset [21]. It consists of the CIFAR10 dataset plus 19 types of corruptions in five levels. The accuracy of each model for each corruptions (gaussian noise, gaussian blur, fog, and contrast) of level five intensities is shown in Table 2. The column of averages in the table shows the average accuracy of each model for all 19 types of corruptions in CIFAR10-C. From average accuracy of the table, we can see that RFC+APR is much more robust to corruptions with the accuracy of 75.86% than the baseline (57.52%). Although APR is better for fog and contrast with a small margin, RFC+APR is the best in terms of the average performance.

3.4. Investigation on how the models utilize the frequency components

We investigate how differently each model handles frequency components. For this purpose, we use ResNet18 trained with each data augmentation in CIFAR10 and examine the accuracy when the low and high-frequency components of the images are separately input to the model.

Here, we describe how to generate the image with only the phase component. Let \mathcal{D}_t be test datasets and \mathbf{x} is included in \mathcal{D}_t . The frequency component $\mathcal{F}(\mathbf{x})$ of \mathbf{x} is calculated using its amplitude \mathcal{A}_x and phase \mathcal{P}_x as shown below:

$$\mathcal{F}(\mathbf{x}) = \mathcal{A}_x \otimes e^{i \cdot \mathcal{P}_x}. \quad (7)$$

Using this, we can calculate the average amplitude \mathcal{A}_m as

$$\mathcal{A}_m = \frac{1}{|\mathcal{D}_t|} \sum_{x \in \mathcal{D}_t} \mathcal{A}_x. \quad (8)$$

For each image, the image with only the phase component \mathbf{x}^p is calculated as:

$$\mathbf{x}^p = \mathcal{F}^{-1}(\mathcal{A}_m \otimes e^{i \cdot \mathcal{P}_x}). \quad (9)$$

Using the mask matrices M_l and M_h defined in Eq. (2), the phase-only images of the low-frequency and high-frequency components, \mathbf{x}_l^p and \mathbf{x}_h^p , are computed as:

$$\mathbf{x}_l^p = \mathcal{F}^{-1}(\mathcal{A}_m \otimes M_l \otimes e^{i \cdot \mathcal{P}_x}), \quad (10)$$

$$\mathbf{x}_h^p = \mathcal{F}^{-1}(\mathcal{A}_m \otimes M_h \otimes e^{i \cdot \mathcal{P}_x}). \quad (11)$$

The accuracies of the models trained on each data augmentation for these images when $r = 4, 8$ are shown in Table 3. From Table 3, when $r = 4$, the accuracy of the baseline model is 10% for high-frequency and high-frequency of phase only components, which is equivalent to a random classifier since CIFAR10 is a 10-class classification. And the baseline model is close to a random classifier for low-frequency components by yielding the accuracy of 15.40%. Similarly, the performance of mixup is the same as that of the random classifier regardless of the presence or absence of amplitude components in both low and high-frequency components. APR, on the other hand, is slightly more accurate than the baseline by achieving the accuracy about 27% for images with only low-frequency components and about 89% for high-frequency components. Furthermore, even for the low-frequency and high-frequency components with only phase, the accuracy does not decrease significantly, and it is considered that a large percentage of the judgment is made based on the phase of the image. RFC has the same tendency to the high-frequency component as APR, but is the same as the random classifier when the low-frequency component is used. However, in RFC+APR, the accuracy using the low-frequency component is greatly increased.

Next, we compare the accuracy of each model when $r = 8$. Compared to $r = 4$, the amount of information in the low-frequency component is larger and that in the high-frequency component is smaller. Therefore, from Table 3, the overall trend is that the accuracy using the low-frequency component is higher and vice versa. For APR and RFC+APR, the accuracy using the low-frequency component increases considerably to more than 80%, and the accuracy using the phase of the low-frequency component alone is more than 70%. The decline of accuracies of APR and RFC+APR is less than that of baseline (31.95% to 17.47%) or mixup (45.59%

Table 2. Comparison of the accuracy of each data augmentation to CIFAR10-C (%). The best values are shown in bold.

method	Test acc.	gaussian noise	gaussian blur	fog	contrast	average
baseline	93.50	27.53	31.89	73.42	48.65	57.52
CutMix [16]	95.00	30.39	22.98	76.78	67.72	59.05
mixup [17]	95.31	41.42	50.64	79.33	72.18	68.10
APR [12]	95.21	44.24	85.65	90.67	74.78	73.27
RFC (proposed)	94.07	16.94	27.27	81.34	50.59	51.71
RFC (proposed) + APR	94.71	51.56	86.48	89.97	73.54	75.86

Table 3. Accuracy of the model trained with each data augmentation on various test data for $r = 4$ (%). Values in parentheses are for $r = 8$. Low, High means low-frequency and high-frequency respectively. P means using only phase component.

method	Original	Low	High	Low-P	High-P	Phase only
baseline	93.50	15.40 (31.95)	10.01 (10.00)	12.26 (17.47)	10.00 (10.00)	73.71
CutMix [16]	95.00	13.37 (12.91)	9.35 (7.49)	11.27 (9.56)	10.06 (9.98)	70.24
mixup [17]	95.31	14.45 (45.59)	9.89 (9.30)	11.42 (23.17)	10.00 (9.98)	79.61
APR [12]	95.21	26.71 (81.07)	88.77 (68.75)	20.34 (72.14)	80.41 (48.65)	92.61
RFC (proposed)	94.07	10.01 (17.33)	89.47 (65.71)	10.00 (11.54)	64.24 (26.46)	74.25
RFC (proposed) + APR	94.71	48.43 (85.19)	87.75 (80.15)	37.60 (74.36)	80.09 (66.29)	91.47

to 23.17%). This indicates that the information of the low-frequency phase is utilized more than that of amplitude. Furthermore, since the accuracy of the high-frequency component of RFC+APR is higher than that of APR by more than 10% and the accuracy of the low-frequency component is also higher in RFC+APR, the difference between APR and RFC+APR can be explained as follows: RFC+APR uses the high-frequency component above $r = 8$.

Here, let us discuss the relationship between the robustness to out-of-distribution data and the accuracy of low and high-frequency components. The robustness to out-of-distribution data is generally in the order of RFC+APR > APR > RFC > baseline > CutMix > mixup. Unlike the other models, the baseline and mixup models, which are less robust to out-of-distribution data, are equivalent to random classifiers in terms of accuracy for images with high-frequency components, while the RFC, APR, and RFC+APR models, which are more robust to out-of-distribution data, use more high-frequency components with $r = 8$ or higher than the other models. Furthermore, RFC+APR, which has the highest performance, especially utilizes more high-frequency components than the other models. Therefore, it is inferred that the robustness to out-of-distribution data depends on whether the model uses high-frequency components or not. Also, RFC+APR and APR have high accuracy for low-frequency and phase only images, and both models have both accuracy and robustness to out-of-distribution data. Thus, it is expected that the low-frequency or phase is necessary to improve both accuracy and robustness to out-of-distribution data. In ad-

dition, the fact that the RFC utilizes more high-frequency components explains the accuracy of the Section 3.3 with respect to CIFAR10-C. RFC is considerably less accurate against gaussian noise and gaussian blur than the baseline model. Considering the corruptions, low-frequency component of gaussian noise is not so different from the original image, but the high-frequency component deviates greatly from the original image due to noise. Gaussian blur, which is equivalent to a low-pass filter in terms of its operation on the image, is an operation in which the high-frequency components required by RFC are lost. Therefore, for image corruptions such as noise and blur, the high-frequency components are significantly changed or lost, and the accuracy of RFC is supposed to be greatly reduced. On the other hand, noises such as fog and contrast are regarded as operations in which a uniform change in the entire frequency range is applied, and as a result, RFC is expected to be as accurate as baseline in such corruptions.

4. CONCLUSIONS

In this paper, we proposed a frequency-based data augmentation that can enhance the robustness to out-of-distribution data. Furthermore, we experimentally showed that robust models mainly use high-frequency of images. It was also suggested that models that also use low-frequency or phase components are also more robust to corrupted data. Future research may include investigating the robustness to adversarial examples and further frequency-based data augmentation.

5. REFERENCES

- [1] “ImageNet Large Scale Visual Recognition Challenge (ILSVRC),” <http://www.image-net.org/challenges/LSVRC/>.
- [2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, “Intriguing properties of neural networks,” in *International Conference on Learning Representations (ICLR)*, 2014.
- [4] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference on Learning Representations (ICLR)*, 2015.
- [5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, “Towards deep learning models resistant to adversarial attacks,” in *International Conference on Learning Representations (ICLR)*, 2018.
- [6] Anh Nguyen, Jason Yosinski, and Jeff Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 427–436.
- [7] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira, “Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [8] Dan Hendrycks and Kevin Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” in *International Conference on Learning Representations (ICLR)*, 2017.
- [9] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 7167–7177.
- [10] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing, “High-frequency component helps explain the generalization of convolutional neural networks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8684–8694.
- [11] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer, “A fourier perspective on model robustness in computer vision,” in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2019.
- [12] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian, “Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain,” in *IEEE International Conference on Computer Vision*, 2021, pp. 458–467.
- [13] Ev Zisselman and Aviv Tamar, “Deep residual flow for out of distribution detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [14] He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [15] Alex Krizhevsky, “Learning multiple layers of features from tiny images,” Tech. Rep., University of Toronto, 2009.
- [16] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in *IEEE International Conference on Computer Vision*, 2019.
- [17] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz, “mixup: Beyond empirical risk minimization,” in *International Conference on Learning Representations (ICLR)*, 2018.
- [18] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bisacco, Bo Wu, and Andrew Y. Ng, “Reading digits in natural images with unsupervised feature learning,” in *Neural Information Processing Systems (NeurIPS) Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- [19] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao, “Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop,” *arXiv preprint arXiv:1506.03365*, 2015.
- [20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.
- [21] Dan Hendrycks and Thomas Dietterich, “Benchmarking neural network robustness to common corruptions and perturbations,” in *International Conference on Learning Representations (ICLR)*, 2019.