The Investigation on Adversarial Attacks of Adversarial Samples Generated by Filter Effects

Qincheng Yang[1, *] and Jianing Yao[2]

[1] International E-Commerce and Law, Beijing University of Posts and Telecommunications, Beijing, 100876, China

[2] Computer Science and Technology, Zhejiang University of Technology, Shaoxing, 312030, China

*2020213359@bupt.edu.cn

Abstract. In contemporary times, there has been a growing inclination among individuals to engage in photography and employ uncomplicated filters to enhance their visual outputs. Although these seemingly straightforward and aesthetically enhanced images are favored by many, they can inadvertently lead to erroneous interpretations by computer vision systems. Such misinterpretations often arise due to the presence of imperceptible image noise, which remains undetectable to the human eye. In this paper, we aim to add some filter effects to the image to verify the effectiveness of the classification results of the interference model, conduct black-box disturbance attacks on the model, and generate adversarial attack samples. For specific anti-attack implementation, we will use the following algorithms to filter the image, among which the contrast and brightness of the image are improved using the histogram equalization procedure; the blur filter algorithm is used to reduce the noise, texture or details in the image to make it more blurred; Utilize the sharpening algorithm to improve the image's edges and features for a crisper, sharper appearance; through the smoothing algorithm to make the image look smoother; through the edge enhancement algorithm to make it clearer. We will use the classic CNNs model to conduct experiments on two datasets of similar size and number but with large differences in image content. The final experimental findings demonstrate that filter interference does affect the model's categorization outcomes.

Keywords: Computer Vision, Adversarial Attack, Filter Effects.

1 Introduction

In the domain of machine vision, the process of deceiving machine learning models by fabricating, augmenting, or changing input data is known as an adversarial attack. These assaults are planned to provide inaccurate results or hinder machine learning models' ability to recognize

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_64

recognize input data accurately [1]. They can be connected to the identification of flaws in machine learning models. Despite the great outcomes of modern machine learning models on many tasks, research has shown that they are more sensitive to even the smallest changes in input data, which opens them up to the prospect of adversarial assaults [2].

Currently, adversarial attacks have emerged as a prominent area of investigation within the realm of machine vision Researchers are committed to developing effective adversarial attack algorithms and exploring defense mechanisms against adversarial attacks. The research on adversarial attacks covers many aspects, including the design and optimization of attack methods, the generation of technology of adversarial samples, the design of defense mechanisms, and the evaluation criteria of adversarial attacks [1]. Adversarial attacks come with a range of benefits and drawbacks. On the one hand, research on adversarial attacks drives the development of robustness and security of machine learning models. By creating adversarial perturbations during training, Madry et al.'s "Robust Optimization" training strategy helped the model better withstand adversarial attacks [3]. Athalye et al. propose a method to synthesize adversarial examples that are more robust against adversarial attacks [4]. A brand-new defense system dubbed "Defense-GAN" was proposed by Samangouei et al. The researchers pointed out that traditional adversarial attacks often deceive the classifier by introducing small perturbations in the input data, making it produce misleading prediction results. Defense-GAN is able to generate adversarial examples with high variation and diversity, making it difficult for classifiers to be fooled [5]. By discovering and exploiting model weaknesses, researchers can propose more robust model designs and defense strategies, thereby improving the reliability of models in real-world scenarios. On the other hand, adversarial attacks also raise some concerns. Attackers can use adversarial attack techniques to trick the system, such as by adding imperceptible perturbations to images so that they are misclassified [6]. Therefore, the importance of resisting attacks cannot be ignored. Studying adversarial attacks can help improve the robustness and security of machine learning models and protect users and systems from malicious attacks. At the same time, research on adversarial attacks also provides new perspectives and challenges for model design and development of defense strategies [1, 7, 8, 9].

To create interference with the data and use it for model training, this paper adds filters to the data. In the instance of filter interference, it is determined by contrasting the accuracy of the experimental results that the use of filters for adversarial attacks effectiveness. We will discuss the training of various data sets with different models under different filter effects to judge the classification accuracy of relevant models in the current mainstream machine vision field, and prove that the filter effect will indeed cause certain interference to the model [10, 11].

2 Method

2.1 Adversarial Attacks

Perturbation Attack. The Perturbation attack is a kind of attack method against the machine learning model, which aims to deceive the model and induce erroneous output by making slight modification to the input data and introducing imperceptible disturbances to human observers. These attack methods exploit the fragility of the model and its sensitivity to input perturbations. In a perturbation attack, the attacker endeavors to identify alterations to the original data that retain a semblance of similarity, yet possess the ability to deceive the model. These perturbations can be pixel changes in images, character substitutions or insertions in text, or small changes in speech signals [12, 13]. We will employ the method of black box attack to perturb the model to achieve adversarial attacks [14].

Proposed Histogram Equalization-based Perturbation Attacks. Histogram equalization is a technique used to adjust the brightness distribution of an image, aiming to enhance the contrast and visual effect of the image [15]. Specifically, the algorithm works by computing a grayscale histogram of the image (depicting the number of pixels at different brightness levels), and then normalizing the histogram so that it appears uniformly distributed. Next, contrast is enhanced by extending the luminance range of the original image to the full range of luminance levels by mapping the pixel values of the original image.

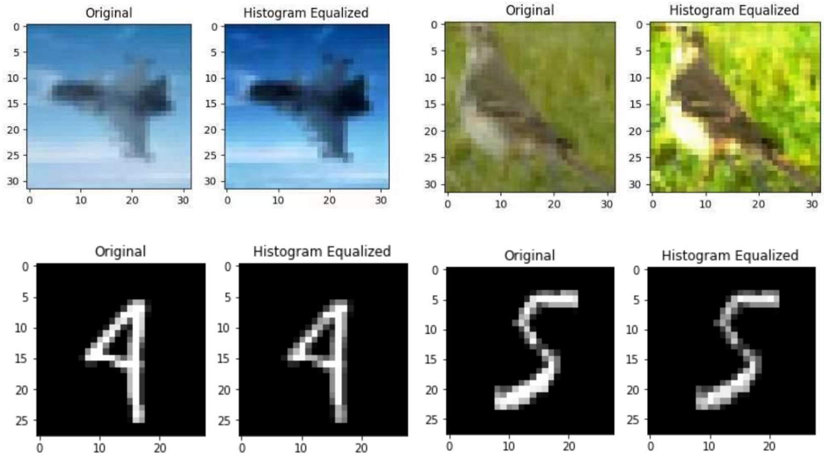


Fig. 1. The original and corresponding histogram equalized sample images [16, 17].

Each pixel in the image is subjected to mapping operations based on the normalized histogram as part of the process for mapping the pixel values of the original image. This entails replacing the original pixel value with a corresponding new value derived from the mapping procedure. Through histogram equalization, areas with darker or brighter brightness in the original image will be enhanced, thereby improving the contrast and visual effect of the image. However, it is important to acknowledge that histogram equalization may alter the overall tone of the image and, in some cases, may introduce noise or over-enhance image detail. Therefore, in practical applications, it may be necessary to combine other technologies or methods to further optimize and adjust the image quality. Fig. 1 shows the original and corresponding histogram equalized sample images.

Blur Filter-based Perturbation Attacks. Blur filtering is an image processing technique that works primarily to reduce noise, texture or detail in an image to make it smoother and blurrier. Specifically, there are three things: 1) Blur filtering can smooth out noise and reduce its effect on an image by averaging around pixels. 2)Blur filtering can blur edges and details to produce a smooth image. 3) Blur filtering can reduce the detail and size of an image by combining adjacent pixels into an average. Fuzzy filtering has various roles in image processing, including noise removal, image smoothing, size reduction, and image pre-processing. Fig. 2 shows the original and corresponding blur filtering-based sample images.

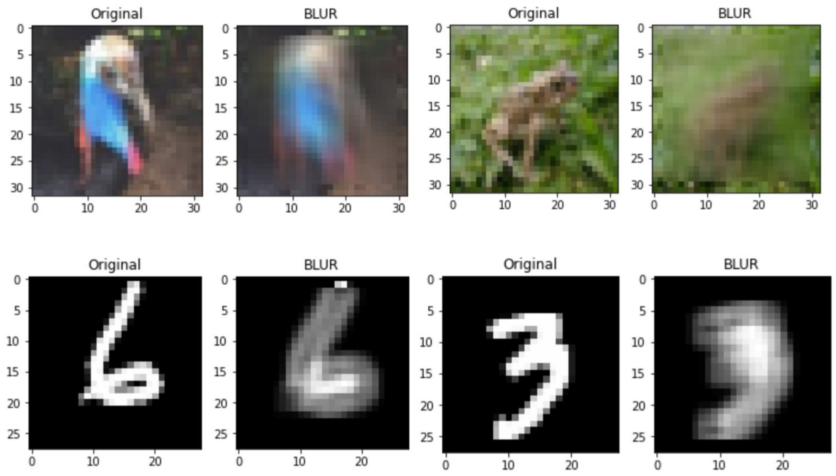


Fig. 2. The original and corresponding blur filtering-based sample images [16, 17].

Sharpen Filter-based Perturbation Attacks. Sharpening filter is a filtering method for image enhancement that enhances edges and details in an image, making the image look clearer and sharper. In terms of the practical implementation, the initial step involves the specification of a sharpening kernel, which serves as a defined filter. Subsequently, this sharpening kernel is applied to every individual pixel present within the image. Then the sharpening kernel is weighted and summed with the pixels of the image and its surrounding pixels, and the result of the weighted sum is used as the new value of the original pixels. By increasing the difference between the pixel values, sharpening filtering enhances the sharpness and detail of the image. Fig. 3 shows the original and corresponding sharpen filtering-based sample images.

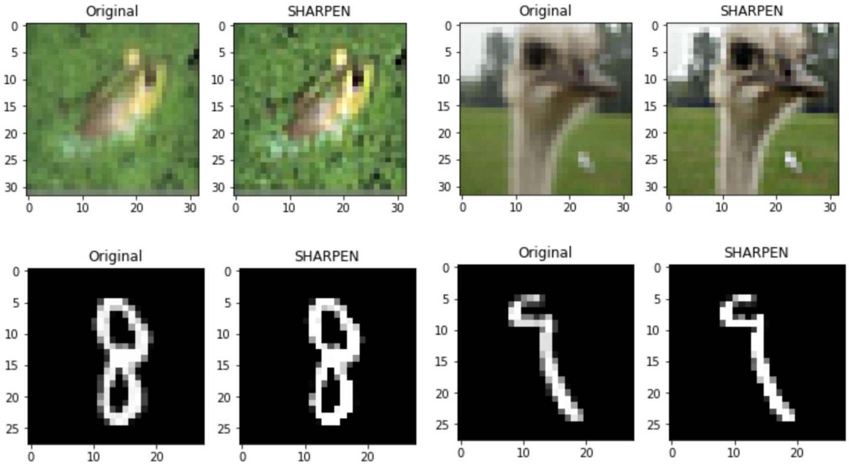


Fig. 3. The original and corresponding sharpen filtering-based sample images [16, 17].

Smooth Filter-based Perturbation Attacks. The objective of smoothing filtering is to reduce noise, texture or detail in an image, making it smoother and more blurred. It does this by performing a weighted average or other operations between the pixels of the image. First, a smoothing kernel needs to be defined and applied to each pixel in the image. This is done by weighing the smoothing kernel with the pixels of the image and its surrounding pixels. The result of the weighted summation will be the new value of the original pixel. The process of smoothing filtering entails the blurring of intricate details and textures within an image, resulting in a visually smoother appearance. By employing smoothing filters, the impact of noise is diminished, and the transitions present in the image become more seamless and gentle, promoting a sense of continuity. Fig. 4 shows the original and corresponding smooth filtering-based sample images.

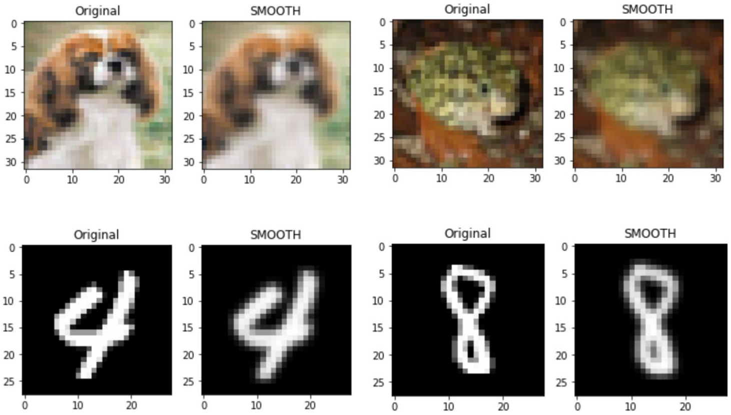


Fig. 4. The original and corresponding smooth filtering-based sample images [16, 17].

Edge Enhance Filter-based Perturbation Attacks. The objective of edge enhancement filtering is to enhance the edges in an image to make them sharper and more visible. It does this by highlighting the edge parts of the image and increasing their contrast and sharpness in order to make the edges more prominent and visible. First, an edge enhancement kernel is defined, and an edge enhancement kernel is applied to each pixel in the image by weighing the edge enhancement kernel with the pixels of the image and its surrounding pixels and summing them. The result of the weighted summation will be the new value of the original pixel. Edge enhancement filtering increases the contrast and definition of edges, thus making them more prominent and visible. Fig. 5 shows the original and corresponding edge enhanced filtering-based sample images.

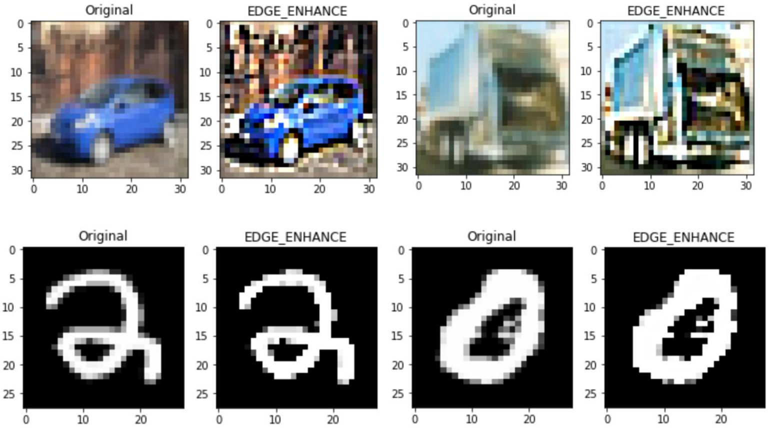
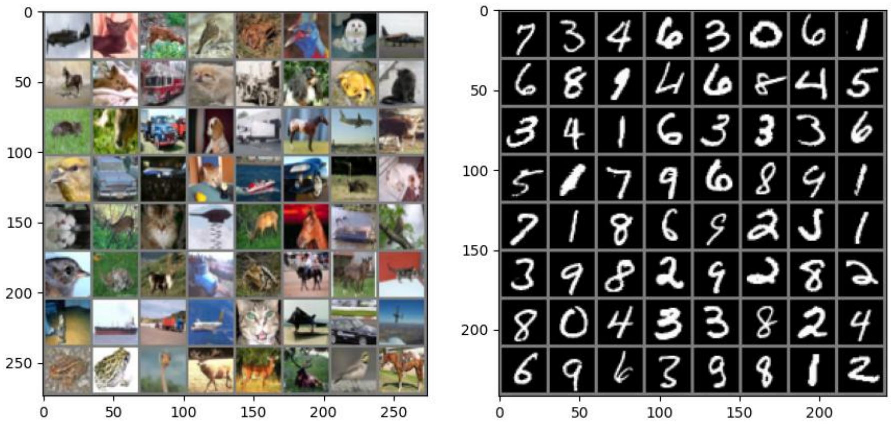


Fig. 5. The original and corresponding edge enhanced filtering-based sample images [16, 17].

3 Results and Discussion

3.1 Experimental Dataset

This experiment uses CIFAR-10, MNIST datasets and CIFAR-10 dataset after filter rendering. CIFAR-10 is a commonly used image classification data set for research in the domain of computer vision. It comprises ten different sorts of image samples, with a total of 60,000 images in ten categories. The size of these images is 32 × 32 pixels, with RGB channels. CIFAR-10 consists of 10 categories and each class image is an example of objects in the real world, with certain visual complexity and changes. The widespread application of the CIFAR-10 dataset has made it one of the benchmark data sets commonly used in machine learning communities. It offers a uniform testing environment that may be used to assess how well various algorithms and models work. Fig. 6 contains examples of the CIFAR-10 pictures.



(a)

(b)

Fig. 6. The sample images of the (a) CIFAR-10 and MNIST [16, 17].

With 60,000 handwritten digital image examples for training, MNIST is a traditional handwritten image dataset. The MNIST contains grayscale images with a 2828-pixel size. Each image depicts a handwritten number, and each of the numbers 0 through 9 has an associated image sample. The sample images of the MNIST can be found in Fig. 6.

3.2 Adversarial Attacks

In the method, we have mentioned several commonly employed filter methods, which form the basis for the specific filter effects observed in practical applications. We attacked the test set of CIFAR10 dataset using several filter methods separately. The counter samples of each filter were generated and tested with three models, LeNet, VGG and ResNet.

Table 1. Accuracy of perturbation attack in different filter effects on the CIFAR-10.

Test	LeNet	VGG	ResNet
Original	65%	79%	74%
Histogram Equalization	44%	21%	15%
Blur Filter	41%	17%	13%
Sharpen Filter	51%	22%	15%
Smooth Filter	51%	21%	14%
Edge Enhance Filter	42%	21%	15%

Table 1 shows that the generated adversarial samples significantly affect the attack effect of each of the three models. In particular, the effect of the attacks on two models, VGG and ResNet, is more significant. Among them, the adversarial attack has the least effect on LeNet, with at most a 24% decrease. The adversarial attack had the largest effect on ResNet, with a maximum drop in accuracy of 61% to only about 14%. This is unacceptable for CIFAR10, a dataset with only ten categories, as the correct rate would be around 10% if random classification was performed. VGG also dropped significantly, from a maximum of 79% correct to around 20%. However, from a model perspective both VGG and ResNet are more advanced models compared to LeNet, and both perform significantly better than LeNet on the original test set, but their performance in adversarial attacks is more susceptible. Both are clearly relatively vulnerable in our adversarial attack tests compared to LeNet. In particular, the most advanced ResNet model of the three showed the worst resistance to attack when attacked by our filter-generated adversarial samples, even with an accuracy rate close to the correct rate of random grouping. According to the model structure, the three models LeNet, VGG, and ResNet deepen in complexity and increase in convolutional layers sequentially. Therefore, the model's ability to extract details of things' features deepens in turn, and extracts more abstract features. However, our adversarial samples did not destroy the main structure and basic features of things during the attack, while some detailed features and abstract features were more severely damaged during the attack. As a result, the latter two models' accuracy suffers significantly, as does their capacity to defend against an attack. We infer that LeNet, which can only extract the basic features of things due to the small number of convolutional layers, is relatively resistant to interference and is not easily affected by filter effects. The abstract features of the object are severely damaged by the filter interference, resulting in the VGG and ResNet models can no longer recognize the abstract features of the object, which in turn leads to the poor anti-interference ability of the two models.

And we can observe that although the filters are different, the results against the attacks have similarity: the accuracy of LeNet floats between 40% and 50%, the accuracy of VGG stays around 20%, and the accuracy of ResNet stays around 14%. It seems that the models have similar judgments about the impact of the three filter attacks. In our human perspective, we observe several filter effects: Histogram Equalization makes the colors more vivid; objects stand out more relative to the background and have more distinctive features. Blur Filter makes the picture blurred and more difficult to identify. Sharpen Filter makes the outline of the object clearer and the object features more prominent. Smooth Filter reduces the noise of the image, but also makes the outline of the object more blurred and reduces the color vividness. Edge Enhance Filter makes the overall color of the image brighter and highlights the features and edge outline of the object. But the models have a clear tendency to differ from human judgment in their performance of several attacks. Even the two models, VGG and ResNet, showed convergent judgment accuracy when dealing with different filter confrontation samples, which clearly distinguished from the tendency of human judgment.

Table 2. Accuracy of perturbation attack in different filter effects on the MNIST.

Test	LeNet	VGG	ResNet
Original	98%	99%	99%
Histogram Equalization	45%	22%	19%
Blur Filter	73%	22%	18%
Sharpen Filter	98%	20%	25%
Smooth Filter	98%	19%	24%
Edge Enhance Filter	98%	19%	24%

We conducted further experiments with the MNIST dataset to verify our conjecture since we believe that the underlying features of the numeric symbols are obvious and relatively simple. It can be observed from Table 2 that all three models have a very high accuracy rate on the original dataset. Following the generation of adversarial samples utilizing similar attack methods, both the VGG and ResNet models exhibit a significant decline in accuracy, comparable to the accuracy rates observed after the CIFAR10 adversarial attack. VGG's accuracy experiences fluctuations hovering around 20%, while ResNet's accuracy remains within the range of 18% to 25%. Conversely, LeNet demonstrates a noteworthy ability to maintain a 98% accuracy rate, even after encountering certain interference, with only two filters successfully impairing its accuracy. Hence, our conjecture is substantiated, suggesting that LeNet exhibits greater resilience to interference due to its focus on extracting fundamental object features. Conversely, the VGG and ResNet models exhibit diminished resistance to interference, likely attributable to their extraction of a multitude of abstract object features.

4 Conclusion

This experimental study aims to investigate the efficacy of anti-attacks in the presence of filter interference by subjecting images to various filters. Through our experiments, we demonstrated that the introduction of filter interference enhances the generation of adversarial examples, subsequently leading to improvements in the performance and robustness of machine learning models. The findings indicate that filter interference significantly influences image classification, with the impact becoming more pronounced as the complexity of the model's extracted features increases. Furthermore, we observed that the effect of attacks is more prominent when the images in the dataset contain richer content. These results contribute to our understanding of the role of filter interference in enhancing the resilience of machine learning models against adversarial attacks.

Acknowledgment

All the authors contributed equally and their names were listed in alphabetical order.

References

1. Goodfellow, I. J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).
2. Szegedy, C., Zaremba, W., Sutskever, I., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013).
3. Madry, A., Makelov, A., Schmidt, L., et al.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017).
4. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-gan: Protecting classifiers against adversarial attacks using generative models. arXiv preprint arXiv:1805.06605, (2018).
5. Carlini, N., Athalye, A., Papernot, N., et al.: On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705, (2019).
6. Moosavi-Dezfooli, S. M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks, Proceedings of the IEEE conference on computer vision and pattern recognition. 2574-2582 (2016).
7. Papernot, N., McDaniel, P., Jha, S., et al.: The limitations of deep learning in adversarial settings. 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, 372-387 (2016).
8. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016).
9. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks, 2017 ieee symposium on security and privacy (sp). Ieee, 39-57 (2017).
10. Eykholt, K., Evtimov, I., Fernandes, E., et al.: Robust physical-world attacks on deep learning visual classification, Proceedings of the IEEE conference on computer vision and pattern recognition, 1625-1634 (2018).
11. Hayes, J., Melis, L., Danezis, G., et al.: Logan: Membership inference attacks against generative models. arXiv preprint arXiv:1705.07663 (2017).
12. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236, (2016).
13. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks, 2017 ieee symposium on security and privacy (sp). Ieee, 39-57 (2017).
14. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, International conference on machine learning. PMLR, 274-283 (2018).
15. Brown, T. B., Mané, D., Roy, A., et al.: Adversarial patch. arXiv preprint arXiv:1712.09665, (2017).
16. LeCun, Y., Bottou, L., Bengio, Y., et al.: Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324 (1998).

17. Krizhevsky. A., Nair, V., Hinton, G.: Cifar-10 (canadian institute for advanced research). URL http://www. cs. toronto. edu/kriz/cifar. html, 5(4): 1 (2010).