Adversarial Defense Networks via Gaussian Noise and RBF

Jingjie Li[1], Jiaquan Gao[1(✉)], Qi Jiang[1], and Guixia He[2]

[1] School of Computer and Electronic Information, Nanjing Normal University, Nanjing 210023, China

[2] Zhijiang College, Zhejiang University of Technology, Hangzhou 310024, China

Abstract. Convolutional Neural Networks (CNNs) have excellent representative power and are state-of-the-art classifiers on many tasks. However, CNNs are vulnerable to adversarial examples, which are samples with imperceptible perturbations while dramatically misleading the CNNs. It has been found that, in past studies, Radial Basis Function (RBF) network can effectively reduce the linearization of the neural networks model, and Gaussian noise injection can prevent the network from overfitting, all of which are conducive for defending against adversarial examples. In this paper, we propose an incorporated defense method with Gaussian noise injection and RBF network, and analytically investigate the robustness mechanism of incorporated defense method. For our proposed method, it has the following two advantages: (1) it has significant classification accuracy, and (2) it can resist to various adversarial attacks effectively. The experimental results show the proposed method achieves the performance of about 79.25% accuracy on MNIST dataset and 43.87% accuracy on Fashion-MNIST dataset, even in the full white-box attack where attackers can craft malicious adversarial examples from defense models.

Keywords: Adversarial examples · RBF · Gaussian noise

1 Introduction

CNNs have recently outperformed machine learning methods in various tasks, such as image classification [1–3], object detection [4–6], and speech recognition [7–9]. However, like many classification machine learning algorithms, deep learning methods can be easily fooled by small imperceptible perturbations in the input [10]. The main reason may be that the linear classifier is used at the last layer of deep learning methods. Although linear classifiers are very effective for the linear classification, they force the model to assign high confidence to these regions far from the decision boundary. Thus the adversarial attack can easily

The research has been supported by the Natural Science Foundation of China under grant number 61872422, and the Natural Science Foundation of Zhejiang Province, China under great number LY19F020028.

X. Sun et al. (Eds.): ICAIS 2021, LNCS 12736, pp. 494–504, 2021.
https://doi.org/10.1007/978-3-030-78609-0_42

make some changes in multiple dimensions of the input image, which makes the perturbed images cross the classification boundary. The perturbed images are misclassified by the network in the end.

In order to mitigate the effect of adversarial attacks, the following two kinds of defense techniques are proposed: data-level method and algorithmic-level method. The former includes adversarial training [11,12], pre-processing methods using basis functions [13] and noise removal [14]. The later can be seen in literatures [15–20], and the deep model is modified or the algorithm is trained by reducing the magnitude of gradients [17], or masking gradient [18]. However, these approaches are not completely effective against several different white-box and black-box attacks [14]. Similar to the methods based on pre-processing, they may decrease accuracy to defense some attacks. Generally, most of these defense strategies make the classification accuracy descend on clean data.

As mentioned above, successful adversarial attacks are mainly due to that fact that the models are linearly high in high dimension. This greatly decreases the flexibility of the models and makes the decision boundary close to the manifolds of the training data. In order to improve the nonlinearity of the model, Goodfellow et al. [11] have explored a variety of methods, including shallow and deep RBF networks. They used the shallow RBF network to achieve good performance against adversarial perturbations. However, they found that it was difficult to train the deep RBF network.

In this paper, we explore the incorporated network model with the deep neural network and RBF network, which not only ensures that the network model can effectively resist perturbations, but also makes the network model be trained easily. Meanwhile, small noise is added before the network input, which improves the robustness of the network attack and resists the white-box and black-box attack effectively.

2 Related Work

2.1 Adversarial Examples

The adversarial examples were first introduced by Szegedy et al. [12]. Szegedy et al. show that the prediction of the network can be changed arbitrarily by applying imperceptible non-random perturbations to the input image. The malicious input is $X' = X + \alpha$ and α is a lightly perturbation with $\|\alpha\| < \epsilon$, where ϵ is so small that it makes no visual difference between X and X' for human being but deep neural networks will be fooled.

In addition, Szegedy et al. point out that the adversarial examples are relatively robust and can be generalized in neural networks with different depths and activation functions. In other words, if we use one neural network to generate the adversarial examples, another neural network also misclassifies these examples even when it is trained with different hyper-parameters, or when it is trained on different subset of a dataset. This phenomenon makes the black-box attacks feasible.

Finally, Szegedy et al. also state that it is linear transformation of high-dimensional space of the model that leads to the phenomenon of adversarial examples, and proposes the Fast Gradient Sign Method (FGSM) for crafting adversarial examples. FGSM is an untargeted attack method and uses the same attack strength at every dimension:

$$X_{FGSM} = X + \varepsilon sign\left(\bigtriangledown_X J\left(X, y\right)\right). \tag{1}$$

In this equation, the adversarial examples are obtained by adding a transformed gradient to the input X, where ε is small enough to be undetectable.

2.2 Gaussian Noise

In the space domain and frequency domain, Gaussian noise (also known as normal noise) is commonly used. The probability density function of the Gaussian random variable Z is given by the following formula:

$$p\left(Z\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(Z-\bar{Z}\right)^2}{2\sigma^2}}, \tag{2}$$

where Z is the gray value, and its mean and deviation are $\bar{Z}$ and σ, respectively.

The perturbation is extremely small in general in order to get a normal example. In [21], Gu and Rigazio consider an alternative strategy by adding Gaussian noise damage into the adversarial examples. The aim of this additional noise strategy is to remove the adversarial examples "blind spot" areas of the classification space through adding extra "larger" interference noise to the input. Moreover, adding the ordinary tiny perturbation does not have a impact on the performance of the neural network. Experimental results show that the Gaussian noise injection can defend against samples to some extent.

2.3 RBF

RBF networks [22–24] are neural networks with one hidden layer of RBF units and a linear output layer. An RBF unit is a neuron with multiple real input $X = (X_1, \cdots, X_n)$ and one output y. Each unit is determined by n-dimensional vector C and the parameter $\beta > 0$. The output y is computed as:

$$y = \varphi\left(\xi\right); \xi = \beta\left\|X - C\right\|^2, \tag{3}$$

where $\varphi : \mathbf{R} \to \mathbf{R}$ is the suitable activation function. Typically, Gaussian $\varphi\left(Z\right) = e^{-z^2}$. Thus the network computes the following function $f : \mathbf{R}^n \to \mathbf{R}^n$:

$$f_s\left(X\right) = \sum_{j=1}^{h} \omega_{ij}\varphi\left(\beta_j\left\|X - C_j\right\|\right), \tag{4}$$

where $\omega_{ij} \in \mathbf{R}$ and f_s is the output of the sth output unit.

Compared with normal networks, RBF networks use radial basis function units at the last layer. Unlike the linear unit, RBF unit is activated in a well-defined region of its input space [25]. In this case, the goal of the feature extractor network is to map the data to a new representation, where each kind of data forms a cluster. Experimental results show that when RBF is combined with CNN, it can effectively resist the disturbance.

3 Method

3.1 Model

Inspired by the defense methods in [21,25], we combine the data-level defense method with the algorithm-level defense method, and propose an incorporated model with Gaussian noise injection and RBF network on neural networks. Noth that the network can be any CNN (see Fig. 1).

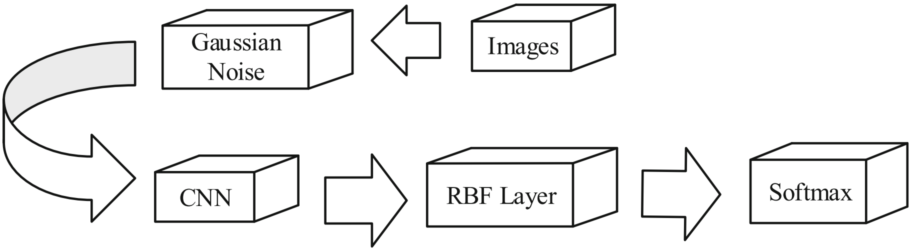

Fig. 1. The incorporated network model.

Here we take the classic Lenet-5 [26] network as the example to demonstrate the proposed model (see Fig. 2). In this figure, the LeNet-5 network structure is in the dashed box, and before the image input into the LeNet-5 network, the tiny Gaussian noise is added to the image. After the feature extraction and the classification through LeNet-5 network, the output of LeNet-5 is set as the input into the RBF network. The RBF output is the final classification result of our proposed model.

3.2 Loss Function

When training the proposed model, we calculate the loss of the final output and use the cross-entropy loss function, the loss is computed as:

$$Loss = loss\left(\hat{y}, y\right) \tag{5}$$

$$loss = -\sum_{i=1}^{N} y_i log\left(\hat{y_i}\right) + \left(1 - y_i\right) log\left(1 - \hat{y_i}\right), \tag{6}$$

where $\hat{y}$ is the final output of the proposed model and y is the true label.

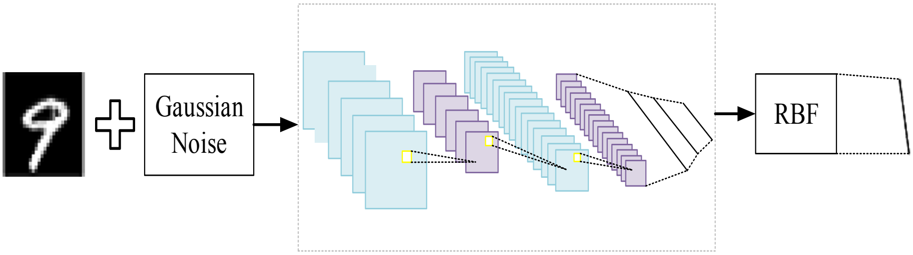

Fig. 2. The incorporated network model based on LeNet-5.

4 Experiments

These experiments in this paper are based on two benchmark datasets: MNIST [26] and Fashion-MNIST [27]. MNIST contains 60,000 images in the training set and 10,000 images in testing set. Each image is a grayscale image with 28×28 pixels, and the number of possible classes is 10. Fashion-MNIST is the standard data set of commodity classification. The color size of pictures and the scale of training set and the test set are all the same with MNIST, but the classification difficulty is higher than MNIST.

4.1 Experiment Setups

In the experiments, we consider two basic CNNs: Simple-CNN with two convolutional layers (S-CNN) and LeNet-5. The activation function in each network is RELU and loss function is the cross-entropy function. The detailed model structure and the parameter information are shown in Table 1, and Table 2 contains the chosen training hyper-parameters for all models.

In order to verify the superiority of the proposed model, the following three defense methods are chosen: the basic CNN model, the CNN model combined with RBF network (CNN_RBF) [25] and the centers are set as 300 in RBF network, and a CNN model incorporating Gaussian noise (Gauss_CNN) [21] with standard deviation $\sigma = 0.3$ on MNIST and Fashion-MNIST.

Moreover, five test sets are set in the defense experiments. For the network based on S-CNN, we first set the 10,000 clean test set images of the MNIST dataset as Test Set I. We then generate adversarial examples test sets by attacking four models(S-CNN, S-CNN_RBF, Gauss_S-CNN, the proposed model), which are set as Test Sets from II to V. For the network based on LeNet-5, two benchmark datasets, MNIST and Fashion-MNIST, are chosen. For each dataset, the same numbers of the clean test set images are set in Test Set I. Finally, we generate adversarial examples test sets by attacking four models (LeNet-5, LeNet-5_RBF, Gauss_LeNet-5, the proposed model), which are set as Test Sets from II to V.

Table 1. Network model parameters.

LeNet-5	Parameters	S-CNN	Parameters
Conv.	32, 5 × 5	Conv.	32, 3 × 3
Max Pooling	2 × 2	Max Pooling	2 × 2
Conv.	64, 5 × 5	Conv.	64, 3 × 3
Max Pooling	2 × 2	Max Pooling	2 × 2
Full Connection	1024	Full Connection	128
Softmax	10	Softmax	10

Table 2. Overview of training parameters.

	S-CNN	LeNet-5
Learning Rate	0.01	0.01
Optimizer	Adam	Adam
Batch Size	128	128
Epochs	20	20

4.2 Experiment Results

In this section, we report the results of several experiments for the task of classification. We first start with S-CNN model on MNIST. Next, the proposed method is applicable to another classical model LeNet-5 on MNIST and Fashion-MNIST. A series of adversarial examples of each test set are generated by FGSM attack, as shown in Fig. 3.

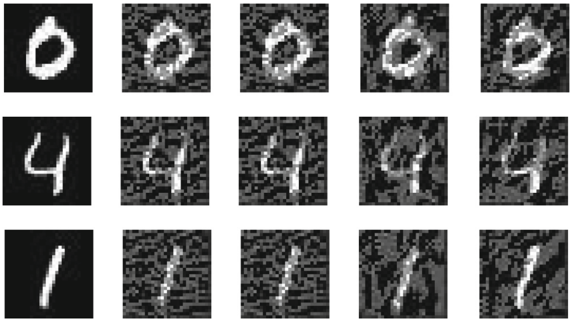

Fig. 3. An illustration of each test set. In this figure, the leftmost column displays the original images, and the next four columns show adversarial examples corresponding to Test Set II, III, IV and V from left to right.

S-CNN. In this section, we compare the recognition accuracy of the four network models in each dataset and the experimental results are shown in Table 3.

Observing the data of the first column (Test Set I) in Table 3, the proposed network model does not drop the accuracy significantly on clean images. That is, Gaussian noise can prevent the network model from overfitting, and the interference of Gaussian noise is equivalent to the data augmentation of the original dataset. From Test Set II to Test Set V, the diagonal data indicate the recognition accuracy of each network model under the white-box attack. Compared with these diagonal data, the performance of S-CNN_RBF and Gauss_S-CNN against white-box attack are improved, especially the improvement effect of S-CNN_RBF is obvious, which is more than 65% higher than S-CNN. Therefore, adding the RBF network after the softmax layer of basic network can greatly improve its robustness. This is due to that fact that the strong local approximation ability of RBF network successfully makes the basic model S-CNN better fit the real decision boundary and compress the adversarial space, and thus the basic mode with RBF can be against the white-box attack effectively. Furthermore, the data in the Test Set II show that defensive performance of four models can be against the black-box attack. Obviously, although S-CNN_RBF model has higher performance against the white-box attack, the defense performance against the black-box attack is still only about 30%. This is due to the fact that the adversarial examples of Test Set II are generated by the original model, and the adversarial examples have the ability to migrate across model. Therefore, utilizing RBF network can resist the white-box attack, though it cannot be against the migration attack effectively. The migration attack is defined as the attack that uses other network against vulnerabilities to generate adversarial examples of the space, and use its migration ability against unknown network.

Table 3. Experimental results based on S-CNN model on MNIST dataset.

	Test Set I	Test Set II	Test Set III	Test Set IV	Test Set V
S-CNN	99.14%	10.69%	84.70%	48.80%	79.16%
S-CNN+RBF	99.00%	38.08%	75.36%	39.70%	75.58%
Gauss+S-CNN	99.20%	74.90%	93.21%	40.23%	84.31%
Gauss+S-CNN+RBF	99.13%	83.80%	93.55%	68.77%	79.25%

In addition, though the network with only adding Gaussian noise is not ideal against the white-box attack (compared with the S-CNN, there is only an increase of about 30%), it has a high recognition accuracy against the black-box attack, which increases by more than 70%. This is because the perturbations are extremely small, and there is a "blind spot" area in the input space of the image. Using extra noise to images can undermine the dominance against perturbations. In this way, adversarial examples can be removed from the "blind spot" to obtain correct classification. Meanwhile, the extra noise will have less impact on the performance of neural network. From the last line in the Table 3, we can observe that the incorporated model with Gaussian noise and RBF network can promote each other. Compared with other defense model against the white-box attack and the black-box attack, the defense performance of the proposed model has improved effectively. From the Test Set I to Test Set V, the proposed model almost keeps the superiority. In conclusion, the proposed network structure has better robustness against the white-box and black-box attacks.

LeNet-5. Another classic network model, Lenet-5, is used in this section, and experiments are carried out on two datasets: MNIST and Fashion-MNIST datasets. The detailed defense performance on the MNIST and Fashion-MNIST is reported in Table 4 and Table 5, respectively.

Table 4. Experimental results based on LeNet-5 model on MNIST dataset.

	Test Set I	Test Set II	Test Set III	Test Set IV	Test Set V
LeNet-5	99.19%	16.86%	59.28%	33.67%	57.37%
LeNet-5+RBF	98.82%	34.57%	42.90%	31.37%	53.40%
Gauss+LeNet-5	98.67%	86.38%	92.23%	36.34%	83.16%
Gauss+LeNet-5+RBF	99.08%	82.25%	92.12%	69.03%	71.16%

Observing the data reported in Table 4, the same conclusions as in S-CNN are obtained. The diagonal data from Test Set II to Test Set V represent those defense models performance against the white-box attack. In addition, the data from Test Set II to Test Set V except diagonal are the performance of each defense model against the black-box attack. Though data analysis, LeNet-5_RBF has a better defense performance against the white-box attack, while Gauss_LeNet-5 has better performance against the black-box attack. The proposed model incorporated the merits of them has better defense performance both in two aspect of against white-box and black-box attacks. Obviously, the proposed model based on LeNet-5 has a recognition accuracy of over 70% on each adversarial examples test set.

Table 5. Experimental results based on LeNet-5 model on Fashion-MNIST dataset.

	Test Set I	Test Set II	Test Set III	Test Set IV	Test Set V
LeNet-5	91.43%	11.18%	31.90%	12.86%	28.46%
LeNet-5+RBF	90.72%	32.43%	30.20%	15.72%	33.98%
Gauss+LeNet-5	89.31%	57.88%	72.81%	8.94%	56.28%
Gauss+LeNet-5+RBF	88.96%	58.03%	73.07%	20.18%	43.87%

Observing data in Table 5, the improvement is not so obvious in the Fashion-MNIST dataset compared with the MNIST dataset, but the improvement trend in each defense model is consistent with MNIST dataset. Like LeNet-5_RBF, it is helpful to enhance the robustness of the model and the defense ability against the white-box attack. The average defensive performance of the Gauss_LeNet-5 against black-box attack is about 60%. And the proposed incorporated model on MNIST dataset is slightly better than it on Fashion-MNIST dataset. However, Gauss_LeNet-5 and the proposed model defense accuracy have a drop in the clean test set. This phenomenon is due to that fact that Fashion-MNIST dataset is more complicated than MNIST dataset. Adding Gaussian noise on Fashion-MNIST dataset has a impact on classification task. The recognition accuracy in the fourth column of the Table 5 on each network in the Test Set IV is less than 20%. The reason may be that images in Test Set I are generated by attacking the Gauss_LeNet-5 model, and the incorporated noise is larger compared to original image. Therefore these noises make the models have low accuracy in the classification task. However, the consistent trend of ascension on Test Set IV again indicates that the incorporated model structure is effective in defense adversarial examples.

In summary, compared the defense performance with other three models against white-box and black-box attacks, the proposed model can defend adversarial examples effectively on Fashion-MNIST.

5 Conclusion

In this paper, we propose an incorporated defense method with Gaussian noise and RBF network. The experimental results show that the proposed method can effectively be against the adversarial examples in the white-box and black-box attacks. Furthermore, compared with other methods, the proposed method effectively improves the classification accuracy on adversarial images, and does not drop the accuracy significantly on clean images.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE, Las Vegas (2016)
2. Zoph, B., Vasudevan, V., Shlens J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8697–8710. IEEE, Salt Lake City (2018)
3. Cao, Y.Q., Tan, C., Ji, G.L.: A multi-label classification method for vehicle video. J. Big Data **2**(1), 19–31 (2020)
4. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural. Inf. Process. Syst. **39**(6), 1137–1149 (2017)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988. IEEE, Venice (2017)
6. Qayyum, A., Ahmad, I., Iftikhar, M., Mazher, M.: Object detection and fuzzy-based classification using UAV data. Intell. Autom. Soft Co. **26**(4), 693–702 (2020)

7. Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., Bengio, Y.: End-to-end attention-based large vocabulary speech recognition. In: IEEE International Conference on Acoustics, pp. 4945–4949. IEEE, Shanghai (2016)
8. Chiu, C.C., et al.: State-of-the-art speech recognition with sequence-to-sequence models. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4774–4778. IEEE, Calgary (2018)
9. Park, J., Kim, S.: Noise cancellation based on voice activity detection using spectral variation for speech recognition in smart home devices. Intell. Autom. Soft Co. **26**(1), 149–159 (2020)
10. Nguyen, A.M., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 427–436. IEEE Computer Society, Boston (2015)
11. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations (ICLR). San Diego, CA, USA (2015) arXiv:1412.6572. Accessed 20 Mar 2015
12. Szegedy, C., et al.: Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations (ICLR). Banff, AB, Canada (2014). arXiv:1312.6199. Accessed 19 Feb 2014
13. Xu, W., Evans, D., Qi, Y.: Feature squeezing: detecting adversarial examples in deep neural networks. In: 25th Annual Network and Distributed System Security Symposium (NDSS). The Internet Society, San Diego, California, USA (2017). arXiv:1704.01155. Accessed 5 Dec 2017
14. Meng, D., Chen, H.: MagNet: a two-pronged defense against adversarial examples. In: Proceeding of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 135–147. ACM, New York (2017)
15. Folz, J., Palacio, S., Hees, J., Dengel, A.: Adversarial defense based on structure-to-signal autoencoders. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3568–3577. IEEE, Snowmass Village (2020)
16. Wong, E., Kolter, J.Z.: Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope. In: International Conference on Machine Learning (ICML), pp. 5283–5292. PMLR, Stockholmsmässan, Stockholm, Sweden (2018)
17. Papernot, N., McDaniel, P.D., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP), pp. 582–597. IEEE Computer Society, San Jose (2018)
18. Buckman, J., Roy, A., Raffel, C., Goodfellow, I.J.: Thermometer encoding: one hot way to resist adversarial examples. In: 6th International Conference on Learning Representations (ICLR). OpenReview.net, Vancouver (2018). https://openreview.net/forum?id=rJUYGxbCW. Accessed 16 Feb 2018
19. Zhang, J., Wang, J.: A survey on adversarial example. J. Inf. Hiding Privacy Protect. **2**(1), 47–57 (2020)
20. Chen, H., Zhu, H.Q., Yan, L.M., Wang, J.W.: A survey on adversarial examples in deep learning. J. Big Data **2**(2), 71–84 (2020)
21. Gu, S., Rigazio, L.: Towards deep neural network architectures robust to adversarial examples. In: 3rd International Conference on Learning Representations (ICLR). San Diego, CA, USA (2015). arXiv:1412.5068. Accessed 9 Apr 2015
22. Peng, J.X., Li, K., Irwin, G.W.: A novel continuous forward algorithm for RBF neural modelling. IEEE T. Automat. Contr. **52**(1), 117–122 (2007)

23. Moody, J.E., Darken, C.J.: Fast learning in networks of locally-tuned processing units. Neural Comput. **1**(2), 281–294 (1989)
24. Broomhead, D.S., Lowe, D.: Multivariable functional interpolation and adaptive networks. Complex Syst. **2**(3), 321–355 (1988)
25. Vidnerová, P., Neruda, R.: Deep networks with RBF layers to prevent adversarial examples. In: Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J.M. (eds.) ICAISC 2018. LNCS (LNAI), vol. 10841, pp. 257–266. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91253-0_25
26. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. P. IEEE **86**(11), 2278–2324 (1998)
27. Han, X., Kashif, R., Roland, V.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747. Accessed 15 Sept 2017