Deep Learning Based Adversarial Images Detection

Haiyan Liu[1], Wenmei Li[1,2(✉)], Zhuangzhuang Li[1], Yu Wang[1], and Guan Gui[1]

[1] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
liwm@njupt.edu.cn
[2] School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Abstract. The threat of attack against deep learning based network is gradually strengthened in computer vision. The adversarial examples or images are produced by applying intentional a slight perturbation, which is not recognized by human, but can confuse the deep learning based classifier. To enhance the robustness of image classifier, we proposed several deep learning based algorithms (i.e., CNN-SVM, CNN-KNN, CNN-RF) to detect adversarial images. To improve the utilization rate of multi-layer features, an ensemble model based on two layer features generated by CNN is applied to detect adversarial examples. The accuracy, detection probability, fake alarm probability and miss probability are applied to evaluate our proposed algorithms. The results show that the ensemble model based on SVM can achieve the best performance (i.e., 94.5%) than other methods for testing remote sensing image dataset.

Keywords: Adversarial detection · Deep learning · Ensemble model · Support vector machine (SVM) · K-nearest neighbors (KNN) · Random forest (RF)

1 Introduction

In the field of deep learning applications, it has been great progress in image classification, speech recognition and computer vision [1]. As a classical network model of deep learning, convolutional neural network (CNN) is indispensable and profoundly meaningful to the development of computer vision, especially in the field of massive image processing [2].

In the subsequent study of the adversary, adversarial training is found to be useful in improving the robustness of classifier based on neural network [3]. It is good news for the acquisition and storage of many precious datasets (e.g. HSRRS images, remote sensing images with high spatial resolution, abundant color, complex texture and similar shape). There are two types of adversarial training, one is the Black-box attack, and the other is White-box attack [4]. White-box attack is undoubtedly catastrophic for neural networks, as it masters almost overall structures and parameters of a particular neural network. Therefore, the performance of neural networks can be better improved by adversarial training [5, 6]. The Fast Gradient Sign Method (FGSM) is the most

G. Gui and L. Yun (Eds.): ADHIP 2019, LNICST 301, pp. 279–286, 2019.
https://doi.org/10.1007/978-3-030-36402-1_30

representative of White-box attack. It has the characteristics of specific adversarial perturbation and one-step target classification in computer vision. Kurakin et al. [6] noted that for ImageNet, the top-1 error rate for the adversarial examples generated by FGSM is approximately 63–69%. Therefore, the input image generated by FGSM makes sense to checkout and improve the performance of neural networks.

The detection is found to be as a technique for defensing against adversarial attacks [7, 8]. Literally, the 'detection only' approach refers to merely monitoring tar-get classification result of deep neural network on adversarial examples with no further processing (e.g. SafetyNet, Additional Class Augmentation). Li et al. [8] applied CNN-based neural networks into the convolution filters statistics to classify the test images. A cascaded classifier is designed to detect adversarial examples generated by the FGSM methods with the accuracy more than 85%. And Meng et al. proposed a framework that utilizes single or multiply external classifier to detect an adversarial image effectively. For the specified remote sensed dataset, Zhang et al. [9] proposed a method based on CNN to achieve classification by obtaining promising features in hyperspectral image. In brief, CNN and external classifier both have good performance in attack detection, but the detection performance is needed to be further improved.

Generally speaking, a single learner is more susceptible to erroneous predictions against adversary. To achieve complicated learning goals, ensemble model is de-signed to construct and combine multiple weak learners into strong learners [10, 11]. It greatly improves the robustness for integral neural network. To solve the limitation of single-dimensional feature representation, this paper is planned to integrate high level and low dimensional features through SVM to detect adversarial attack. The two dimensional features are extracted by CNN first or second full-connection layer. At the same time, CNN's output will be compared to verify the effectiveness of the integrated network (CNN-classifier). The objective of this paper is briefly summarized as the following two points:

(a) Multi-layer features fusion: It combines the features of different scales of the full connected layer, reduce the influence of feature positions and improve the robustness of the classification.
(b) Adversarial samples detection based on machine learning: Cooperation with three typical classification algorithms, convolutional neural network can obtain better performance in detecting adversarial attack.

The paper is organized as follows. Section 2 several classification algorithms and hybrid model cascaded based on CNN are presented. Section 3 the main experimental method is stated. Section 4 the experimental results are analyzed and in Sect. 5, conclusions and prospect are drawn.

2 Methodology

2.1 CNN

A CNN is a representative method for image classification in computer vision. The classical CNN mainly consists of the convolutional layer, pooling layer and fully

connected layer. Among them, the convolutional layer is to extract key features in images and to migrate learning of different features. The multi-use of the max pooling layer facilitates feature extraction, reduces feature dimension and prevents over fitting effectively. In addition, the fully connected layer integrates the feature representations, greatly reducing the impact of feature locations on the classification results. The classical structure of CNN is shown in Fig. 1.

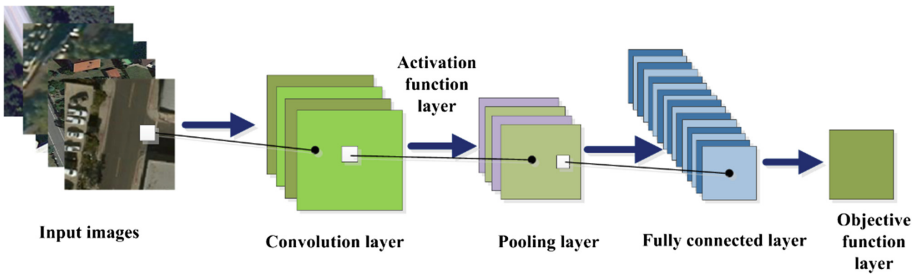

Fig. 1. The classical structure of CNN.

In our experiment, the input image is provided with the width and height of 128 and the color channel is 3. For feature extraction, the convolution layer is set as two 3 × 3 kernels with dimension 128 and 64, respectively. The ReLU is applied as the activation function of the convolutional layer. And the pooling layer is described as max pooling with 2 × 2 kernel. In addition, the Dropout is set to 0.25. For classification part, the 1024, 512 and 10 respectively indicate the number of neural units of the first (Dense1), second full connected layer (Dense2) and the number of final classification labels. The activation function is still ReLU, except for Softmax in the output.

2.2 Several Classification Algorithms

2.2.1 SVM

The SVM is a kind of supervised learning with stable effects and fast prediction capacity, which is applied widely in its application in two or multiple and linear or nonlinear classification. It is designed to find a hyperplane to segment the positive and negative examples and ensure the largest interval between the two classes. Given a linearly separable training dataset, the classification hyperplane is obtained by maximizing or equivalently solving the corresponding convex quadratic programming problem,

$$f(x) = sign(w \cdot x + b) \tag{1}$$

$$w \cdot x + b = 0 \tag{2}$$

where the Eqs. (1) and (2) is decision function and hyperplane representation, respectively. The value of function *sign(x)* takes −1 when the variable is less than 0; otherwise, +1. The hyperplane divides the feature space into two parts: a positive class

and a negative class. The points in side with positive normal vector belong to positive class, and the other side is negative class.

2.2.2 KNN

The KNN is a fundamental classification and regression method. For classification, its model is designed to divide the eigenvector space by training dataset. The KNN input is a vector and the output is category, a scalar. As shown in Eq. (3), the criteria for the division is the distance between different feature values. The distance L can be obtained by the following:

$$L = \left(\sum_{l=1}^{n} \left| x_i^{(l)} - x_j^{(l)} \right|^p \right)^{\frac{1}{p}} \tag{3}$$

It is called Manhattan distance when $p = 1$, the Euclid distance when $p = 2$ and the Maximum distance of each coordinate when p tends to infinity. The l presents instance vector dimension of vector space, the i and j, meaning i-th and j-th input trained instance vector. The values i, j, l are smaller than the dimension n of the vector space R^n. Known the relationship between the dataset and the category, and the label of the test dataset is compared with its corresponding feature. To predict the label, the KNN uses the majority voting rule to select the data of the similar feature.

$$f : R^n \rightarrow \{c_1, c_2, \cdots, c_k\} \tag{4}$$

$$P(Y \neq f(X)) = 1 - P(Y = f(X)) \tag{5}$$

$$\frac{1}{k} \sum_{x_i \in N_k(x)} I(y_i \neq c_j) = 1 - \frac{1}{k} \sum_{x_i \in N_k(x)} I(y_i = c_j) \tag{6}$$

where Eqs. (4, 5, 6) present classification function, misclassification probability and local misclassification probability, respectively. The function f exists in the n dimension vector space, which contains k categories. In the formula (5), the total probability that the instance X corresponds to any label Y equals to 1. In the formula (6), I presents indication function related label y_i and category c_j. The set $N_k(x)$ is constituted by k neighbor trained points x.

2.2.3 RF

As a representative model of ensemble learning, RF implements relatively simple. It is not sensitive to partial feature deletions on massive and highly parallel sample training. Due to the random sampling, the trained model has a small variance and a strong generalization ability. Since the decision tree node partitioning feature has been randomly selected, the model can be trained effectively when the dimension of sample feature is high.

RF uses the CART decision tree as a weak learner, and it uses the majority voting method of weak learners to determine the final classification result. For classification issues, RF utilizes the information gain to measure the amount of information,

and determine the direction of feature split. The Eq. 7 describes that D presents training data set, A_i is the i th feature of D, $H(D)$ denotes the information entropy, $H(D/A_i)$ means the mutual information.

$$g(D, A_i) = H(D) - H(D/A_i) \tag{7}$$

2.2.4 Ensemble Method

Considering the disadvantage of CNN in adversary, the hybrid model is developed to strengthen the robustness of classifier and detect adversarial images. The specific flowchart of the integrated model is shown in Fig. 2. The above classifier integrates different features for predictive classification. These features are extracted from the first and second full connected layer, respectively. Then the classifier gains a prediction result.

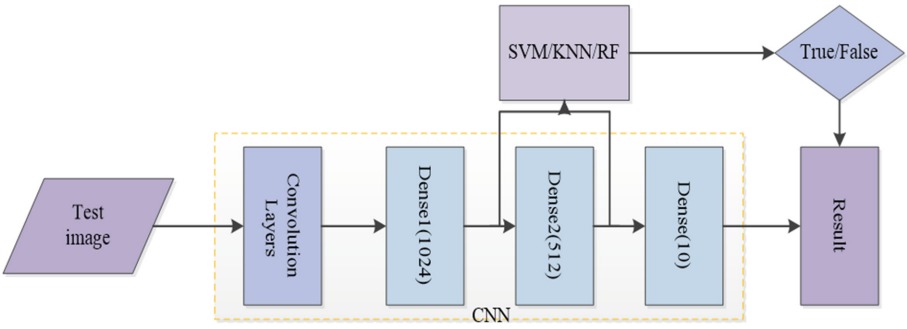

Fig. 2. Flow chart of ensemble model.

2.3 Metric Evaluation

A confusion matrix, where the positive class is set to the adversarial image, is the basic evaluation criteria of experimental performance for adversarial detection. The accuracy (Acc), detection probability (P_D), fake alarm probability (P_{FA}) and miss probability (P_M) are all applied to evaluate the performance of classifier. The TP, TN, FP, FN in confusion matrix are all key parameters for calculating the above metrics. The calculation process is carried out by following,

$$\begin{aligned} P_D &= \frac{TP}{TP + FN} \\ Acc &= \frac{TP + TN}{TP + TN + FP + FN} \\ P_{FA} &= \frac{FP}{TP + FP} \\ P_M &= \frac{FN}{TP + FN} \end{aligned} \tag{8}$$

3 Experiment

This experiment adopts a high spatial resolution remote sensing (HSRRS) image as the test data, with 10 labels, each containing 100 samples. The clean test images and the fake images generated by FGSM are feed into the classifier.

3.1 Adversarial Attack and Detection

The Fig. 3 shows the flowchart of adversarial attack and detection. The adversarial image is generated by adding perturbation or noise on clean image (original HSRRS image). The perturbation originated from FGSM. Then, we train CNN to achieve target prediction. The same feature extraction process by consecutive four convolutional layers from clean image. The fully connected layer integrates gradually features for classification. Then we replace clean image with adversarial image repeating the above steps.

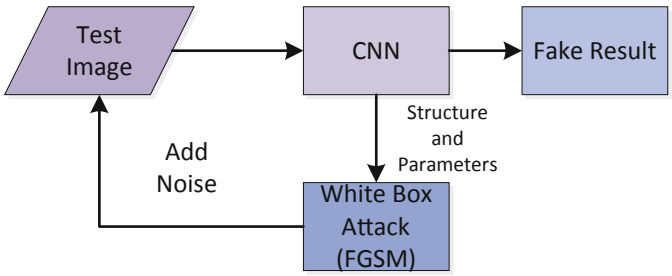

Fig. 3. The flowchart of adversarial attack and detection.

3.2 Adversarial Detection on Ensemble Model

The full connected layer is regarded as a classifier in above step. We design a framework that combines the fully connected layer with the real classifier to select promising features, then integrate these features and predict results. These promising features are extracted from the first and second full connected layer, respectively. They work together on the SVM classifier, then SVM makes the final prediction. Same process is repeated, just replacing the classifier SVM with KNN or RF. Finally, the experimental results are presented using the confusion matrix, also the relevant metric results are calculated at the same time.

4 Results and Analysis

Due to the FGSM perturbation, the classification accuracy of CNN has declined from 96.4% to 33.3%. The experimental results are carried out in the following two aspects: (1) for each ensemble model, the feasibility of each full connected layer, and (2) for each metric, the performance comparison of classifier. Table 1 illustrates the metrics of several ensemble methods.

With respect to first aspect, CNN-SVM gains a best performance of accuracy, the detection probability and the miss probability, especially in the first full connected layer (Dense1), followed by CNN-KNN and CNN-RF. Yet the CNN-KNN gets a terrible performance in detection probability and miss probability, 0.880 and 0.120, respectively. CNN-RF has a moderate advantage at each metric. In addition, the feature integration by classifier significantly improves the classification performance of each fully connected layer. As it gains a better performance than the best between Dense1 and Dense2. Above all, the robustness of ensemble model based on other classifier is improved to the greatest extent.

In terms of second aspect, the metric of ensemble model is described as following. For the accuracy, CNN-SVM is at a dominant position in the above three models, followed by CNN-KNN and CNN-RF. For the detection probability and the miss probability, CNN-SVM is best, CNN-RF is moderate and CNN-KNN is worst. For the fake alarm probability, CNN-KNN is best, followed by CNN-SVM and CNN-RF.

From the perspective of fake alarm probability, we can choose CNN-KNN for classification prediction. In addition, CNN-RF is the most suitable choice for the moderate performance requirements. It also can be summarized that CNN-SVM possesses a good advantage for its highest accuracy and highest detection probability, lowest miss probability in feature analysis and classification. What's more, CNN-KNN tends to identify clean image as adversarial one, especially for the features collected in Dense1.

Table 1. Four metrics of the three ensemble methods

Method		Accuracy	Detection Probability (P_D)	Fake Alarm Probability (P_{FA})	Miss Probability (P_M)
CNN-SVM	Dense1	94.0%	0.933	0.054	0.067
	Dense2	93.3%	0.923	0.058	0.077
	Dense1&2	94.5%	0.933	0.044	0.067
CNN-KNN	Dense1	92.1%	0.880	0.040	0.120
	Dense2	93.7%	0.917	0.045	0.083
	Dense1&2	93.8%	0.907	0.032	0.093
CNN-RF	Dense1	92.8%	0.923	0.067	0.077
	Dense2	92.5%	0.927	0.064	0.073
	Dense1&2	94.3%	0.927	0.058	0.073

5 Conclusion

The adversarial detection approaches based on machine learning are proposed, and the original classifier model is based on CNN. These methods have improved the precision of target prediction and classification. What's more, we verified the validity of the ensemble model and found its better performance for adversarial images detection.

The results show that CNN-SVM obtains the best performance in classification with the detection probability 0.933 and the accuracy 94.5% in Dense1&2. Considering the overall performance, CNN-SVM is the most appropriate choice for adversarial images detection.

There are still some promotions need to be done in the future work. For example, robust light network or other neural networks for HSRRS images classification. A better feature representation acquired by preprocessing data. And the defense or detection of adversarial examples in DNN based classifiers.

References

1. Gui, G., Huang, H., Song, Y., Sari, H.: Deep learning for an effective nonorthogonal multiple access scheme. IEEE Trans. Veh. Technol. **67**(9), 8440–8450 (2018)
2. Huang, H., Yang, J., Huang, H., Song, Y., Gui, G.: Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system. IEEE Trans. Veh. Technol. **67**(9), 8549–8560 (2018)
3. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR), pp. 1–11 (2015)
4. Tabacof, P., Valle, E.: Exploring the space of adversarial images. In: Proceedings of International Joint Conference on Neural Networks (IJCNN), pp. 426–433 (2016)
5. Liang, B., Li, H., Su, M., Li, X., Shi, W., Wang, X.: Detecting adversarial image examples in deep neural networks with adaptive noise reduction. IEEE Trans. Dependable Secur. Comput. (2018). https://doi.org/10.1109/TDSC.2018.2874243
6. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. In: International Conference on Learning Representations (ICLR), pp. 1–17 (2017)
7. He, W., Wei, J., Chen, X., Carlini, N., Song, D.: Adversarial example defenses: ensembles of weak defenses are not strong (2017). http://arxiv.org/abs/1706.04701
8. Li, X., Li, F.: Adversarial examples detection in deep networks with convolutional filter statistics. In: IEEE International Conference on Computer Vision (ICCV), pp. 5775–5783 (2017)
9. Zhang, M., Li, W., Du, Q.: Diverse region-based CNN for hyperspectral image classification. IEEE Trans. Image Process. **27**(6), 2623–2634 (2018)
10. Fawzi, A., Moosavi-Dezfooli, S.-M., Frossard, P.: Robustness of classifiers: from adversarial to random noise. In: 30th Conference on Neural Information Processing Systems (NIPS), pp. 1632–1640 (2016)
11. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy (SSP), pp. 39–57 (2017)