ENHANCES THE ROBUSTNESS OF DEEP LEARNING MODELS USING ROBUST SPARSE PCA TO DENOISE ADVERSARIAL IMAGES

Truong Phi Ho[1], Truong Quang Binh[2], Nguyen Vinh Quang[1], Nguyen Nhat Hai[2], Pham Duy Trung[1*]
[1]Vietnam Academy of Cryptography Techniques
[2]School of Information and Communication Technology - Hanoi University of Science and Technology

ARTICLE INFO

Received:	**08/11/2023**
Revised:	**07/12/2023**
Published:	**07/12/2023**

KEYWORDS

Deep learning
Adversarial examples
Image features
Sparse PCA
Robustness of model

ABSTRACT

Recent years have demonstrated the rapid development of artificial intelligence. Deep learning applications have been widely developed in life such as object recognition, face recognition, automatic vehicle operation, and even medicine, etc. However, these systems contain many risks from adversarial attacks on deep learning models. Attackers often use examples containing small perturbations that are barely perceptible to the naked eye and can fool even deep learning models. Many studies have shown that the creation of adversarial examples largely depends on adding perturbations to clean image. In this paper, the authors propose to use the Sparse Principal Component Analysis (PCA) method to denoise adversarial images. With the experimental results, the authors have demonstrated that the Robust sparse PCA method is effective in selecting and classifying key features of the image to remove unwanted noise present in the input image. The image after denoising has been accurately classified by machine learning model.

TĂNG CƯỜNG ĐỘ MẠNH MẼ MÔ HÌNH HỌC SÂU BẰNG CÁCH SỬ DỤNG PCA THƯA MẠNH KHỬ NHIỄU CHO HÌNH ẢNH ĐỐI KHÁNG

Trương Phi Hồ[1], Trương Quang Bình[2], Nguyễn Vinh Quang[1], Nguyễn Nhất Hải[2], Phạm Duy Trung[1*]
[1]Học viện Kỹ thuật Mật mã
[2]Trường Công nghệ Thông tin và Truyền thông - Đại học Bách khoa Hà Nội

THÔNG TIN BÀI BÁO

Ngày nhận bài:	**08/11/2023**
Ngày hoàn thiện:	**07/12/2023**
Ngày đăng:	**07/12/2023**

TỪ KHÓA

Học sâu
Mẫu đối kháng
Đặc trưng hình ảnh
PCA thưa
Sự mạnh mẽ của mô hình

TÓM TẮT

Những năm gần đây đã chứng kiến sự phát triển nhanh chóng của trí tuệ nhân tạo. Các ứng dụng học sâu đã được phát triển rộng rãi trong cuộc sống như nhận dạng đối tượng, nhận diện khuôn mặt, vận hành xe tự động, y học,… Tuy nhiên, các hệ thống này ẩn chứa nhiều rủi ro từ các cuộc tấn công đối kháng vào các mô hình học sâu. Kẻ tấn công sử dụng hình ảnh có chứa nhiễu loạn rất nhỏ khó có thể nhận thấy và có thể đánh lừa các mô hình học sâu. Nhiều nghiên cứu đã chỉ ra rằng mẫu đối kháng phần lớn được tạo ra từ việc thêm nhiễu loạn vào hình ảnh sạch. Trong bài báo này, các tác giả đề xuất sử dụng phương pháp phân tích thành phần đặc trưng hình ảnh theo cách thưa thớt để khử nhiễu cho các ảnh đối kháng. Với kết quả thực nghiệm, các tác giả đã chứng minh rằng phương pháp PCA thưa mạnh có hiệu quả trong việc lựa chọn và phân loại các đặc điểm chính của ảnh để loại bỏ nhiễu loạn không mong muốn có trong ảnh đầu vào. Hình ảnh sau khi khử nhiễu đã được mô hình học máy phân lớp chính xác.

DOI: https://doi.org/10.34238/tnu-jst.9166

[*] Corresponding author. *Email: trungpd@actvn.edu.vn*

1. Introduction

Neural networks (NNs), an idea dating back decades, are the driving force behind rapid advancement in the field of machine learning. Deep neural networks (DNNs) have gained significant popularity and extensive application in various computer vision fields. Notably, they have been widely used for tasks such as image recognition [1], image processing [2], image segmentation [3], and image merging [4]. This widespread adoption of DNNs signifies their success in these domains. A fairly common application of NN is image classification. NN is the most accurate machine learning method known to date. These models appear to be vulnerable to attacks to adversarial attacks that attempt to fool the classifier [5]. DNN improved and more developed than NN, but DNN can still be easily fooled by adversarial examples. In simpler terms, attacker can generate a visually similar image x' with a different classification than the original natural image x. Such a pattern x' is called the adversarial example (AE) [6]. AE has been shown to exist in most domains where NN/DNN is used. By a simple method of noise step size allocation based on gradient information, Shi et al. [7] have proposed an effective attack method to fool the image classifier. The application of DNN is very sensitive and needs high accuracy such as object recognition, face recognition, automatic driving or medical analysis, etc. Therefore, it is necessary to research to understand AE and its impact. The methods to combat AE are always a topical issue.

To combat adversarial attacks, many countervailing defense methods have been developed such as using trained AEs, counter-poisoning methods, etc. Previous methods [8], [9] to improve the robustness of the target model added adversarial examples to the training data but poor generalization to invisible attacks. To overcome this, Xie et al. [10] proposed to add feature-denoising blocks specific to the classifier. In contrast, another defensive approach by overcoming input poisoning [11], [12], does not require retraining or modifying the classifier. The primary objective of this technique is to eliminate any adversarial noise present in the input data before it is provided to the classifier. The authors used various input transformations including color bit depth, image blur, and JPEG compression, to obtain good protection performance (in [13], [14]). However, these methods suffer from loss of image information and do not work well with strong adversarial noise.

Other common techniques for defending against adversarial attacks include adversarial retraining [15], gradient regularization [16], and input transformation methods [17]. Contrast training and gradient regularization require retraining or editing of the classifier. Compared with the above two methods, the input transformation-based methods focus on tuning and modifying the input before entering the classifier, so this method is more applicable. Jia et al. [18] have proposed a method to compress and recover images from AEs using CNN models using large data sets for training. Similarly, the GAN defense [19] uses Generative Adversarial Networks (GAN) to reduce the impact of adversarial interference. However, these defensive methods are mainly based on the external preferences learned from the large training datasets, ignoring the rich internal preferences of the input. Statistical training data cannot be generalized for all types of attacks, so the application of these defense methods is also limited.

Several recent input transformation-based defensive approaches [7], [12] attempt to refine or modify the antagonistic samples into clean images using DNN. Liao et al. introduce a method called High-Level Representation Guided Denoiser (HGD) in their paper [20]. The HGD technique is designed specifically to eliminate counter noise from the input data. The authors took advantage of generalized models to turn adverse images into clear ones [12], [21] (using the general model to clean up antagonistic patterns). The proposed defense-generating network (Defend-GAN) [19] projects adversarial patterns into the space of a machine that models the distribution of clean images. Although Defense-GAN and ComDefend belong to unsupervised methods, there are some disadvantages: they require a large amount of unlabeled data for training; ComDefend and Defense-GAN require much more computing power. Therefore, it is possible to learn that external preferences are statistically skewed, thereby limiting their application in practice.

Through surveying many studies around the world, there exists a large amount of work in machine learning (ML) and statistics on how to exploit data sparsity and analyze data [22]. This is motivated by the fact of the increasing amount, dimension, and size of data. Another higher-level goal is to use the basic sparsity of natural data to extract meaningful features using a large number of samples and sub-linearities in the data dimensions. Among the defenses is a second (other) neural network to classify images as natural or antagonistic. The method of using PCA to detect statistical properties of images or network parameters in studies [23] – [25] has also achieved certain results.

Principal Components Analysis (PCA) takes a set of points in a space with multiple dimensions and restructures them into a new set of points in a space with fewer dimensions (where the new space can have the same or fewer dimensions as the original space). For instance, suppose we have a collection of points in an n-dimensional space. Our objective is to convert this set of points into a new set residing in a lower-dimensional space of k dimensions, where ($k < n$). This restructuring is achieved through a linear transformation. Hendrycks and Gimpel [24] used PCA to detect natural images from AE, finding that antagonistic patterns place higher weights on larger principal components than natural images (and lower weights for the previous principal components). Nevertheless, this defense strategy limits the attacker's ability to manipulate only the first k elements, while the classifier disregards the rest. If the adversarial examples are generated based on the last principal components, confining the attack to the first k components will considerably amplify the distortion required to create an AE.

The authors want to mention the Robust Sparse PCA method. We use norms generated in sparse PCA. We show that if we can find the weights on samples (like the classical PCA method) so that the experimental covariance of these samples has a minimum dual standard, then the value of the determinant is inferred. The double level gives us a distinction between the case where noise is added or not. To find this set of weights, we find that the standards are all convex, and therefore the results are affected.

As mentioned, there is a huge amount of work in the world on different ways to exploit sparseness in machine learning and statistics. Taking advantage of the heterogeneity, the authors demonstrate that this can be done through a program using the types of techniques described in the study, and the vertex eigenvectors are optimized to solve the problem. the problem to be solved is. The method under consideration is illustrated on a real data sample and compared with the results of the rate identified in the simulation experiment. From this, it is shown that the Robust sparse PCA method is efficient and provides protection against perturbation.

PCA is commonly used for analyzing multivariate data, but interpreting the results can be challenging because the components are a linear combination of variables. In order to resolve this problem, several methods have been developed to distribute non-zero coefficients across the components. These methods include rotation threshold methods and a more recent approach called PCA with sparse constraints [26].

A significant limitation of principal component analysis, especially with high-dimensional data, is that the resulting components are formed as a linear combination of all input variables. To enhance the interpretability of PCA, several methods that promote sparsity have been proposed recently. However, all of these methods are subject to outliers in the data. Todorov et al. [27] recently introduced an algorithm for calculating sparse and robust principal component analysis. This study compares the performance of our method with standard classical and robust PCA (non-sparse) as well as several other sparse methods.

In this study, the authors propose a method that is not like the previous defensive methods but focuses on using the preferences inside the image to analyze the image into its main components to remove unwanted noise following by the strategy of using the two-stage feature in the sparse estimation, reconstruction process. Furthermore, the proposed method only requires and applies to an adversarial example, the individual input image. Therefore, they can be more flexible to defend against different types of attacks with different attacks.

Currently, there is a lack of well-defined research on the properties and performance of different sparse PCA methods. This is partly due to the misconception that the formulations and equivalences observed in conventional PCA also hold true for sparse PCA. To demonstrate and show the potential of sparse PCA methods, the proposed method is theoretically grounded and substantiated by the application of dual norms, sparse mean estimation, and feature analysis. The theoretical foundation provides insights into the behavior of the algorithm, which can contribute to improved performance understanding. The approach is versatile and can be applied across various domains, including image recognition and feature analysis. This adaptability makes it useful in tackling noise-related challenges in diverse applications. We propose to use this method on total sparse average estimation to anti-noise by then feature analysis of the image and image recovery using this method will be described in detail in section 2.

In the next part of the paper, in section 2 we will present our proposed method. The authors provide experimental analysis and results in section 3 of the paper to showcase the efficacy of our proposed method. The author ends the paper with the conclusion in section 4.

2. Proposed method (Robust sparse PCA) in feature analysis and image recovery to remove adversarial noise

Sparse PCA is a method of matrix analysis used to find the principal components of data. With the assumption that the data is noisy by outliers, sparse PCA tries to find the principal components of the data more precisely and sparingly, while ignoring the outliers. A robust sparse mean estimation problem in which the input data is noisy (see in [28]). Thus, it is required to estimate the sparse mean of the input distribution, where only a small number of components make a significant contribution to the mean. In addition, the input value may be noisy by outliers or other outlier. The goal of the problem is to build a computationally efficient algorithm to estimate a stable sparse mean from this noisy data. This algorithm solves the problem of estimating the mean of a noisy sparse vector using the magnitude of the symmetry vector of the vector to be estimated and dual norms. This is the theoretical basis to prove the correctness of the research method in this paper. The proposed method feature analysis and image recovery will be presented in detail in sections 2.1 and 2.2.

2.1. Feature analysis of images using Robust sparse PCA

Consider a signal-to-noise ratio $\rho > 0$ and a sample set affected by noise ϵ from a distribution d-dimensional $\mathcal{D}$, where $\mathcal{D}$ can be $\mathcal{N}(0, I)$ or $\mathcal{N}(0, I + \rho etc^T)$ for a unit vector k-sparse v.

Given ρ, δ, $\epsilon > 0$ fixed. Set $\eta = \mathcal{O}\left(\epsilon\sqrt{\log(1/\epsilon)}\right)$. If $\eta = \mathcal{O}(\rho)$, and we have a sample set affected by noise ϵ from the distribution $\mathcal{N}(0, I)$ or $\mathcal{N}(0, I + \rho etc^T)$ for a unit vector k-sparse v according to Equation (1). The goal of this problem is to distinguish between two cases: data generated from a unit normal distribution $\mathcal{N}(0, I)$ and data generated from a normal distribution $\mathcal{N}(0, I + \rho etc^T)$ where v is a unit vector k-sparse. In the first case, the data is unaffected by any vector and its distribution is a unit normal distribution. In the second case, the distribution of the data is changed by a unit vector k-sparse v and it will have a different property from the unit normal distribution.

$$n = \Omega\left(\frac{\min(d, k^2) + \log\binom{d^2}{k^2} + \log(1/\delta)}{\rho^2}\right). \tag{1}$$

The goal of the problem is to find a way to distinguish these two cases using as few samples as possible. This can be achieved using an algorithm that matches a specified number of samples. This algorithm must have polynomial complexity and must be able to detect noise-affected data sets with probability $1 - \delta$ as follows Algorithm 1.

```
Algorithm 1: Feature analysis of images using Robust sparse PCA
Procedure FeatureAnalysis($(X_1, \ldots, X_n, w)$)
    Let $\hat{\mu} = \sum w_i X_i$
    Let
        $\hat{\Sigma} = \sum w_i (X_i - \hat{\mu})(X_i \hat{\mu})^T$
    Let $A = d_{\chi_k}(\hat{\Sigma})$
    If $|< A, \hat{\Sigma} - I| \geq 20\eta$ then
        Let $\sigma = sgn(< A, \hat{\Sigma} - I >)$
        return the hyperplane $l$ given by
            $l(w) = \sigma\left(\sum_{i=1}^{n} w_i \langle A, (X_i - \hat{\mu})(X_i - \hat{\mu})^T\rangle - 1\right) - |\langle A, \hat{\Sigma} - I\rangle|$
     else
     return $YES$.
End.
```

2.2. Image recovery using extracted features

Here, we are given a sample set affected by noise ε from the distribution $\mathcal{N}(0, I + \rho etc^\top)$, and our goal is to output a u that minimizes L(u, v), where $L(u, v) = \frac{1}{\sqrt{2}}\left||uu^\top - etc^\top|\right|_2$.

Given $\epsilon, \rho > 0$ fixed. Set $\eta = \mathcal{O}\left(\epsilon\sqrt{\log(1/\epsilon)}\right)$. There is an efficient algorithm, for a set of n samples affected by noise ϵ from the distribution $\mathcal{N}(0, I + \rho etc^\top)$, according to Equation (2):

$$n = \Omega\left(\frac{\min(d, k^2) + \log g\binom{d^2}{k^2} + \log g(1/\delta)}{\eta^2}\right), \tag{2}$$

and for the output is (3):

$$\mathcal{L}(u, v) = \mathcal{O}\left(\frac{(1+\rho)\eta}{\rho}\right). \tag{3}$$

In particular, observe that when $\eta = \mathcal{O}(\rho)$, then when $\varepsilon = \tilde{\mathcal{O}}(\rho)$ this implies that we recover $\boldsymbol{v}$ with a small constant error. Therefore, with the same number of samples as for Robust sparse PCA detection, algorithm starts to provide recovery commits of small enough value and a sufficiently large number of samples, according to Algorithm 2.

```
Algorithm 2: Image recovery using extracted features
Procedure RecoverSparsePCA($(X_1, \ldots, X_n, \epsilon, \delta, \rho)$)
    Let $w^*, A^*$ be the solution to
        $\underset{w \in S_{n,\varepsilon}, A \in \mathcal{X}_k}{\arg\min} \left\|\sum_{i=1}^{n} w_i (X_i X_i^T - I) - \rho A\right\|_{\mathcal{W}_{2k}}^{*}$

        return The $d_{\mathcal{U}_k}(u) \parallel u \parallel_{\mathcal{U}_k}^{*}$ the top vectors are removed except for $k$.

End.
```

3. Experiments and Results

3.1. Experiments

In the field of machine learning experiments, there are many shared and diverse data sets chosen by many research groups to be applied in the field of artificial intelligence, such as CIFAR-10 [29], Image-Net [30], etc. For this study, the authors used the CIFAR-10 dataset, which datasets of detection, segmentation, and captioning, of large-scale objects. CIFAR-10 dataset consists of 60,000 color images size at 32 × 32 pixels in 10 classes. Which used in experiments to evaluate the model's classification accuracy. We select three classes in the included dataset horse, car, airplane. Each class includes 500 images. We use 1500 selected images to create noise and using the proposed method is shown in the two- phase flowchart in Figure 1.

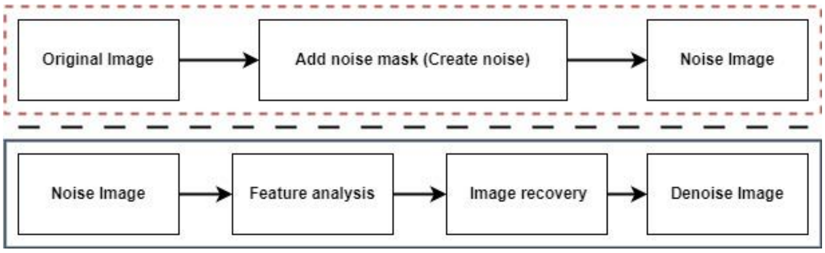

Figure 1. *Two phase flowchart: (Above) Generate noise; (Below) Use Robust sparse PCA to remove adversarial noise*

We experimented with the algorithm by using several different noise models, including Gaussian noise model, Salt and Pepper noise model:

- **Gaussian Noise models:** to generate random noise for an image that follows a Gaussian distribution. Given an image I and Gaussian noise ϵ, the noisy image I_{noisy} defined as (4):

$$I_{noisy}(x,y) = I(x,y) + \epsilon(x,y), \quad (4)$$

where (x,y) represents the pixel coordinates, and $\epsilon(x,y)$ is the Gaussian noise at each pixel.

- **Salt and Pepper Noise model:** introduces random black and white pixels into an image. Given an image I and the probability p of adding salt and pepper noise, the noisy image I_{noisy} can be defined as (5), where (x,y) represents the pixel coordinates.

$$I_{\text{noisy}}(x,y) = \begin{cases} 0, & \text{with probability } \frac{p}{2} \\ 255, & \text{with probability } \frac{p}{2} \\ I(x,y), & \text{otherwise} \end{cases} \quad (5)$$

We use Convolutional Neural Network (CNN) model to train and evaluate the algorithm by testing the dataset in three states: original image, image containing noise, denoised image show in Figure 2. CNNs are a cornerstone of modern computer vision, renowned for their ability to learn hierarchical patterns in image data. However, they are not without limitations; they often require large amounts of labeled data for effective training and can struggle with understanding spatial hierarchies or complex relationships in images, especially when context outside the immediate field of view is important.

3.2. Results and discussions

We conduct statistical results in Table 1 and Table 2 with 3 selected classes in the CIFAR-10 data set are Car, Horse, and Airplane. In Table 1, we compare the true and false labels on each class of data for images after noise generation and images after denoising.

Table 1. *Evaluate true (✓) and false (×) labels on each class using CNN models to class identification*

Class	Image with Gaussian noise	Denoised image	Image with Salt&Pepper noise	Denoised image
Car	×	✓	✓	✓
Horse	×	✓	×	×
Airplane	×	✓	×	✓

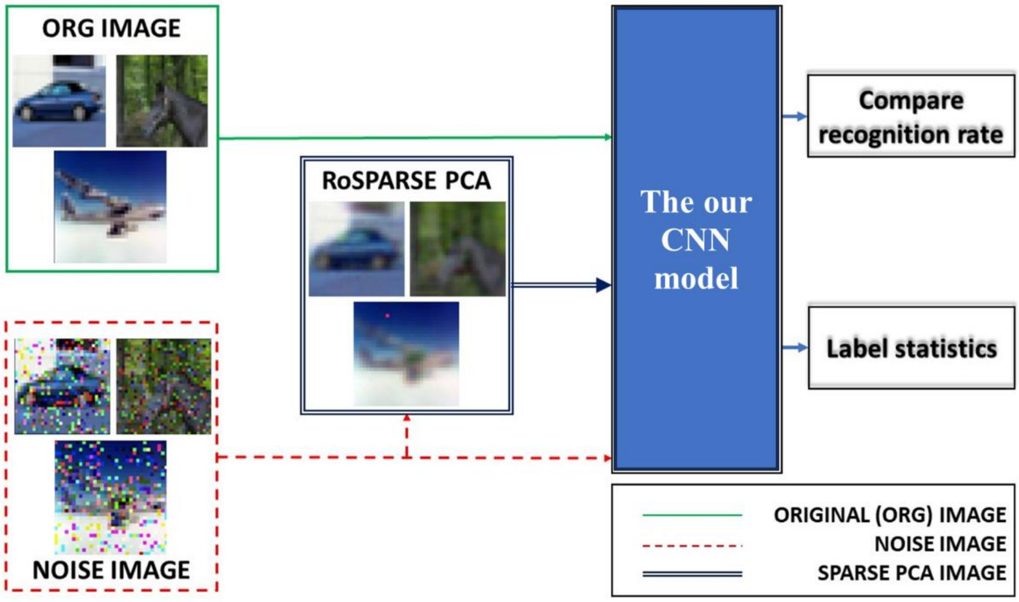

Figure 2. *The our experimental model*

Table 2 shows Average recognition rate (%) of original image - **R_{org}, R_{avg}** - average results of label recognition rate of images on each class and label recognized by our CNN model. Through the experimental results in Table 1 and Table 2, we find that the proposed method is effective in effective defense against adversarial attacks. The proposed method leverages Sparse PCA to effectively identify and classify key features within images, thus removing unwanted noise introduced by adversarial attacks. This robustness against adversarial examples enhances the model's ability to with stand attacks and maintain its accuracy.

Table 2. *Label and average recognition rate - R_{avg} (%) on three classifiers (false labels are bolded).*

Class	R_{org}	Image with Gaussian noise		Denoised image		Image with Salt&Pepper noise		Denoised image	
		Label	R_{avg}	Label	R_{avg}	Label	R_{avg}	Label	R_{avg}
Car	94.36	**Ship**	87.21	Car	62.34	Car	88.03	Car	11.48
Horse	70.87	**Dog**	43.39	Horse	39.43	**Truck**	90.03	**Deer**	32.43
Airplane	85.63	**Bird**	42.67	Airplane	80.07	**Horse**	19.66	Airplane	27.87

In addition to the image classes after denoising are accurately identified (see in Table 2), we see that the image after denoising is recognized with approximate labels, for example deer is almost similar to horse. This proves the effectiveness of the proposed algorithm. The results from applying the method showcase improved denoising performance, leading to images with reduced noise and preserved image details. This is especially evident in cases where images are affected by Salt and Pepper noise. Compared to traditional techniques such as PCA [24] and Sparse PCA, the Robust Sparse PCA (see in [27], [28]) consistently demonstrated superior denoising results.

This indicates that the Robust Sparse PCA method is more effective in removing noise and maintaining image integrity.

The experimental process also requires a lot of time, although experiments on a small image set, CIFAR-10, have shown that performance might be sensitive to parameter settings, requiring careful tuning for optimal results. This sensitivity could pose challenges, particularly when applying the method to different datasets or scenarios. While the method's effectiveness is showcased across domains such as CIFAR-10 dataset, its performance could vary based on dataset characteristics and noise distribution. The method's performance should be validated on diverse datasets to assess its generalizability. The complexity of the method could demand substantial computational resources, which could limit its application in resource-constrained environments or real-time applications.

4. Conclusion

In this paper, the authors present defense adversarial methods. The paper also points out the limitations of existing defensive methods. From there, a method of extracting dimensional features of the data is proposed to remove unnecessary disturbances. With this method, the authors found that the restored image retains the basic features and components. The image is put into the recognition process. It will give the recognition rate and the result of the high number of duplicate labels with the original image. However, the proposed method involves multiple steps and parameters, making it relatively complex to implement and tune. This complexity might hinder its adoption, especially in scenarios where simplicity is preferred.

In the future, the authors hope to experiment with the proposed method on different data sets with a larger number of images. The goal for future research is to make the algorithm faster and more efficient, so that it can be used as a preprocessing step for deep learning models. This would enhance the image classification performance, and make the deep learning models more secure and accurate.

Acknowledgments

Truong Phi Ho was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2023.TS.037.

REFERENCES

[1] L. Li, “Application of deep learning in image recognition,” *Journal of Physics: Conference Series*, IOP Publishing, vol. 1693, no. 1, 2020, Art. no. 012128.

[2] N. Xu, “The application of deep learning in image processing is studied based on the reel neural network model,” *Journal of Physics: Conference Series*, IOP Publishing, vol. 1881, no. 3, 2021, Art. no. 032096.

[3] J. Yang, Y. Sheng, Y. Zhang, W. Jiang, and L. Yang, “On-device unsupervised image segmentation,” *arXiv - CS - Computer Vision and Pattern Recognition*, 2023, doi: arxiv-2303.12753.

[4] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang, “Infrared and visible image fusion via detail preserving adversarial learning,” *Information Fusion*, vol. 54, pp. 85–98, 2020.

[5] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndiˇc, P. Laskov, ´G. Giacinto, and F. Roli, “Evasion attacks against machine learning at test time,” in *Proceedings of Machine Learning and Knowledge Discovery in Databases: European Conference*, Springer, 2013, Part III 13, pp. 387–402.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv - CS - Computer Vision and Pattern Recognition*, 2013, doi: 10.48550/arXiv.1312.6199.

[7] Y. Shi, Y. Han, Q. Zhang, and X. Kuang, “Adaptive iterative attack towards explainable adversarial robustness,” *Pattern recognition*, vol. 105, 2020, Art. no. 107309.

[8] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” *arXiv - CS - Computer Vision and Pattern Recognition*, 2016, doi: 10.48550/arXiv.1611.01236.

[9] F. Tramer, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *International Conference on Learning Representations (ICLR),* 2018, doi: 10.48550/arXiv.1705.07204.

[10] C. Xie, Y. Wu, L. V. D. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 501–509.

[11] J. Chen, X. Zhang, R. Zhang, C. Wang, and L. Liu, "De-pois: An attack agnostic defense against data poisoning attacks," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3412–3425, 2021.

[12] Y. Bai, Y. Feng, Y. Wang, T. Dai, S.-T. Xia, and Y. Jiang, "Hilbertbased generative defense for adversarial examples," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4784–4793.

[13] A. Shukla, P. Turaga, and S. Anand, "Gracias: Grassmannian of corrupted images for adversarial security," *arXiv - CS - Computer Vision and Pattern Recognition,* 2020, doi: 10.48550/arXiv.2005.02936.

[14] C. Guo, M. Rana, M. Cisse, and L. V. D. Maaten, "Countering adversarial images using input transformations," *arXiv preprint arXiv:1711.00117*, 2017.

[15] M. O. Mendonc¸a, J. Maroto, P. Frossard, and P. S. Diniz, "Adversarial training with informed data selection," in *2022 30th European Signal Processing Conference (EUSIPCO), IEEE,* 2022, pp. 608–612.

[16] E. C. Yeats, Y. Chen, and H. Li, "Improving gradient regularization using complex-valued neural networks," in *International Conference on Machine Learning*, *PMLR,* 2021, pp. 11 953–11 963.

[17] F. Nesti, A. Biondi and G. Buttazzo, "Detecting Adversarial Examples by Input Transformations, Defense Perturbations, and Voting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 3, pp. 1329-1341, 2023, doi: 10.1109/TNNLS.2021.3105238.

[18] X. Jia, X. Wei, X. Cao, and H. Foroosh, "Comdefend: An efficient image compression model to defend adversarial examples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6084–6092.

[19] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," *arXiv preprint arXiv:1805.06605*, 2018.

[20] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2018, pp. 1778–1787.

[21] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," *arXiv preprint arXiv:1710.10766,* 2017.

[22] C. Croux, P. Filzmoser, and H. Fritz, "Robust sparse principal component analysis," *Technometrics*, vol. 55, no. 2, pp. 202–214, 2013.

[23] Bhagoji, Arjun Nitin, Daniel Cullina, and Prateek Mittal, "Dimensionality reduction as a defense against evasion attacks on machine learning classifiers," *arXiv preprint arXiv:1704.02654,* 2017.

[24] D. Hendrycks and K. Gimpel, "Early methods for detecting adversarial images," *arXiv preprint arXiv:1608.00530,* 2016.

[25] P. Kaur and J. Singh, "A study on the effect of gaussian noise on psnr value for digital images," *International journal of computer and electrical engineering*, vol. 3, no. 2, p. 319, 2011.

[26] R. Guerra-Urzola, K. V. Deun, J. C. Vera, and K. Sijtsma, "A guide for sparse pca: model comparison and applications," *Psychometrika*, vol. 86, no. 4, pp. 893–919, 2021.

[27] V. Todorov and P. Filzmoser, "Comparing classical and robust sparse pca," in *Synergies of soft computing and statistics for intelligent data analysis*. Springer, 2013, pp. 283–291.

[28] J. Li, "Robust sparse estimation tasks in high dimensions," *arXiv preprint arXiv:1702.05860,* 2017.

[29] Y. Abouelnaga, O. S. Ali, H. Rady, and M. Moustafa, "Cifar-10: Knn- based ensemble of classifiers," in 2016 *International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2016, pp. 1192–1195.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.