Mathematical Modelling of Engineering Problems
Vol. 11, No. 12, December, 2024, pp. 3259-3268
Journal homepage: http://iieta.org/journals/mmep

Improving Robotic Vision Reduces Vulnerability to Attacks in Deep Learning Systems

Darin Shafek[1*], Zainab Mejeed Khadim[1], Mohammed Noori[2]

[1] Department of Computer Engineering Techniques, Al-Ma'moon University College, Baghdad 1004, Iraq
[2] Department of Communications Engineering, Al-Ma'moon University College, Baghdad 1004, Iraq

Corresponding Author Email: darin.s.salim@almamonuc.edu.iq

https://doi.org/10.18280/mmep.111206

Received: 30 March 2024
Revised: 3 September 2024
Accepted: 10 September 2024
Available online: 31 December 2024

Keywords:
vision of robots, image transformations, attacks, deep learning systems

ABSTRACT

This study deals with challenges facing the promotion of robot vision and alleviating restrictions on deep learning systems in processing visual information. In the context of our increasingly interconnected world, artificial intelligence (AI) transforms many industries, with a noticeable effect on independent vision through specialized nerve networks. However, these systems often face problems related to weakness and reliability. This research examines hostile attacks on vision systems, secure treatment processing strategies, and the protection of nerve networks. By using advanced technologies such as noise fillers, engineering transfers, and data increase, the strength of these networks against attacks has been significantly improved. The study deals with weaknesses in current defense technologies and explores the application of image transfers as a major strategy. In addition, concepts such as litigation, class learning, confidence improvement, and obstetric models are discussed. In affirming the development and publication of artificial intelligence, the study reviews realistic state studies to highlight the effectiveness and possible challenges of these innovative technologies. The results emphasize the importance of strong defense mechanisms in developing artificial intelligence applications.

1. INTRODUCTION

In our increasingly connected world, artificial intelligence revolutionized many industries, especially in the field of self-driving vehicles. These self-driving compounds now work as smart surveillance systems, which use advanced neurological networks specialized in artificial intelligence of night vision, which has received global attention despite the large costs of it. However, spreading these artificial intelligence systems, especially in automatic vision, represents unique weaknesses that need to be addressed [1].

The networks used for vision tasks vary and often operate in hostile environments, which makes choosing the network is very important and not trivial. As the opponents continue to develop their technologies, it becomes necessary to develop strong wireless communications for nerve networks to ensure the safety and reliability of these systems.

This research aims to explore the weaknesses of robotic vision systems, with a focus on the field of hostile attacks and their impact on communication strategies with pictures. It highlights the importance of protecting these networks from hostile attacks. By using advanced techniques such as noise filler, engineering shifts, and increasing data, nervous networks can be immunized against corruption and manipulation [2].

During this study, we will primarily study neural network systems in vision applications and various hostilities. We will discuss the restrictions imposed on current defense technologies and explore the application of photo transfers as a strategic defense mechanism. In addition, we will delve into concepts such as competitive training, augmented learning, improving confidence, and birth models [3].

The necessity of the responsible development of these technologies is emphasized and published through realistic status studies that present the effectiveness and potential challenges of innovative applications. The recent developments in the field of robots and computer vision have strengthened the ability of robots to realize and interact their environments. However, this progress led to the emergence of new security challenges and weaknesses [1].

Harmful attacks, a borrowed concept of cybersecurity, pose a major threat to vision systems. These attacks use perceptions of robots, which leads to wrong interpretations and decisions that may be dangerous. For example, changing visual data entry into the self-driving car may lead to incorrect explanations for the traffic light, causing accidents.

Understanding and measuring weaknesses in robotic vision systems is very important. Researchers explore various defenses such as training in competition, strong reinforcement and detecting abnormal conditions to enhance the flexibility of these systems. This study aims to highlight the diversity and influence of hostile attacks on vision systems and discuss the latest developments in defines technologies [2-4].

By providing a detailed analysis of competitive technologies and their effects on robotic vision, this research seeks to bridge the current gaps and suggest effective solutions.

The goals are to improve the safety of artificial intelligence-based vision systems and ensure their reliable operation in important applications such as self-driving and health care robots.

2. REFERENCE STUDIES

Deep networks have been widely adopted in many areas of application, but it turned out that they can be deceived using hostile examples-modified images with hidden and incomplete noise often designed to mislead the classification process. Bhatia et al. [5] investigated the effect of transforming the image on protecting the models of identification on the nerve networks of the nerve attacks from hostile attacks. This study found that hostile roller images can help restore the correct identification in nerve models, indicating that recycling transfers can provide some protection against hostile attacks in neural networks. However, this approach may have restrictions regarding the types of attacks that he can defend against, indicating that future studies must explore other transformations and evaluate their effectiveness in different practical scenarios.

Zhang et al. [6] addressed improving the robustness of the Vision Transformer (ViT) model against hostile attacks by integrating the ResNet-SE module. This improvement affects ViT's attention module, which not only learns edge and font information, but also progressively extracts complex features. The ResNet-SE module emphasizes critical information in each feature map while withholding secondary information, helping the model extract key features more effectively.

Guadarrama et al. [7] suggested a different way to counter hostile attacks on the classification of images, which depend heavily on deep nerve networks, where researchers used JPEG pressure and the GNT model (GAN) to remove the noise causing rivalry attacks. The GAN network has been trained on various pressures, allowing its teachers to change dynamically during every repetition of the defense process, making the approximation of the graduation difficult for the attacker. Experiments using the proposed defense method showed their effectiveness against three attacks in the White Fund and two attacks in the Black Fund, with a special focus on the BPDA attack.

This defense proposed by Kaur and Bhatia [8] provided the advantage of the lack of need of aggressive training and independence from the model and the type of attack. The dynamic amendment of the AR-Gan parameter plays a decisive role in achieving a high defense effectiveness. However, it is important to note that this study focuses only on protecting classification models from hostile attacks without assessing the general performance of the model or its ability to deal with regular use cases. These factors should be taken into account when assessing the effectiveness of the proposed defenses in practical contexts or real-world applications.

In recent years, competitive automatic learning field has witnessed significant developments. The researchers have explored different ways of defense against hostilities, including litigation training, defensive distillation, and input conversion techniques. Tang and Huang [9] provided the litigation training, where models are trained using examples of litigation, as a widely approved approach. However, it can be mathematically expensive and may not be well circulated to all kinds of attacks. Yang et al. [10] suggested defensive distillation, which aims to make models more powerful using cognitive distillation techniques. However, it has been proven that this approach has limits, especially against adaptive attacks. I have gained input transformation techniques, such as JPEG pressure and random size change, interest in its simplicity and effectiveness. These methods adjust the input data to remove hostile disorders before they reach the model, providing a layer of defense that does not depend on the model.

Our work aims to build on these current studies by exploring the common effects of image transfers and advanced nerve network engineering to enhance the power of automatic vision systems. By integrating technologies such as engineering transformations, confrontation training and obstetric models, we seek to develop a comprehensive defensive strategy that addresses restrictions on individual methods. This study contributes to the ongoing efforts in this field by providing a detailed analysis of the effectiveness of various defense mechanisms and proposing new methods to protect the systems of vision based on artificial intelligence against opponents' attacks. Through this comprehensive approach, we aim to bridge the current gaps and provide effective solutions to improve the safety and reliability of independent and robotic systems in important applications.

3. METHODOLOGY

3.1 Building a deep-based robotic vision system

In recent years, robots have witnessed great developments, enabling robots to perform a wide range of tasks with increasing independence. One of the most important aspects of robotic systems is its ability to realize and understand the surrounding environment, which is achieved through the computer vision. Image-based robotic vision has emerged, in particular, as a major search field, allowing robots to analyze visual data and take enlightened decisions based on what they see. Deep learning, a sub-field of artificial intelligence, has revolutionized the field of computer vision by enabling machines from learning and extracting meaningful information from huge amounts of visual data. Consequently, the robotic vision systems based on deep learning are necessary for understanding and development [11-13].

3.2 Understanding deep learning to build a deep learning model

Deep learning, in essence, is a technique of machine learning inspired by the structure and function of the human brain. The training of artificial nervous networks includes multiple layers to identify complex patterns and extract them from the data entered. In the context of image-based robotic vision, deep learning models can be trained to understand and interpret visual information, enabling robots to perform tasks such as identifying things, understanding scenes and navigation.

The first step is to build a deep-based robotic vision system in collecting and processing a variety of data in advance. This data collection should consist of classified images that cover a wide range of objects, viewers and lighting conditions that the robot is expected to face in its operational environment. Data increase techniques, such as rotation, measurement and heart, are used to increase the size and diversity of the data set [14]. Figure 1 shows the process of extracting features in the nervous network, where the ripples are used to make the

features of the features.

Once the data set is prepared, the next step is to design a deep learning model structure. The CNN (CNN) has proven its high effectiveness in the tasks based on images due to their ability to capture spatial hierarchical serials and extract related features of the images. The structure usually consists of multiple convictions, followed by assembly layers to reduce spatial dimensions and fully connected layers for classification or slope tasks [15].

Deep learning model training involves presenting frequently categorized images to the network, allowing it to learn and adjust its internal parameters to reduce the difference between expected and actual ratings. This process is accomplished using optimization algorithms such as Stochastic Gradient Descent (SGD) or its variants. To prevent over-processing, techniques such as leakage and regulation are used [16]. The model is built according to generalizations as in Eq. (1).

$$f(x,y) = \Sigma i \Sigma jki,j \times xi - j \tag{1}$$

where, f is the processed image, k is the candidate image, x is the original image, and the model is built according to generalizations such as $ai = \sigma(\Sigma jwijxj + bi)$, where a is the outputs of unit i, w weights, b biased edges, σ activation function, and the model is trained using inverse propagation and associative error equation such as Eq. (2).

$$\delta j = \frac{\partial E}{\partial aj}, \qquad \Delta wij = \eta \delta jxi \tag{2}$$

where, δ is the error, E is the total error, and η is the learning rate.

After training the model, it is necessary to evaluate its performance on a separate data set to verify its health. Measurements such as accuracy, control, summons and F1 are used to assess the effectiveness of the model. If the performance is not satisfactory, the form can be accurately adjusted by adjusting the super parameters, adjusting the structure, or collecting additional data. Once it is trained and evaluated, the deep learning model can be published and integrate into the automated system, analyzing visual data in the actual time picked up by robot-installed cameras, thus providing valuable information to make decisions and control [17].

Integration may include the development of software interfaces, the creation of communication protocols, and the guarantee of compatibility with automated devices. Building a deep learning system for photo-based automatic vision is a complex and recurrent process. By harnessing the power of deep learning, robots can gain a deeper understanding of what surrounds them, enabling them to intelligence intelligently with the environment. The main steps included in this process include data collection, pre-processing, form structure design, training, improvement, evaluation, careful control, and finally, publishing and complementarity.

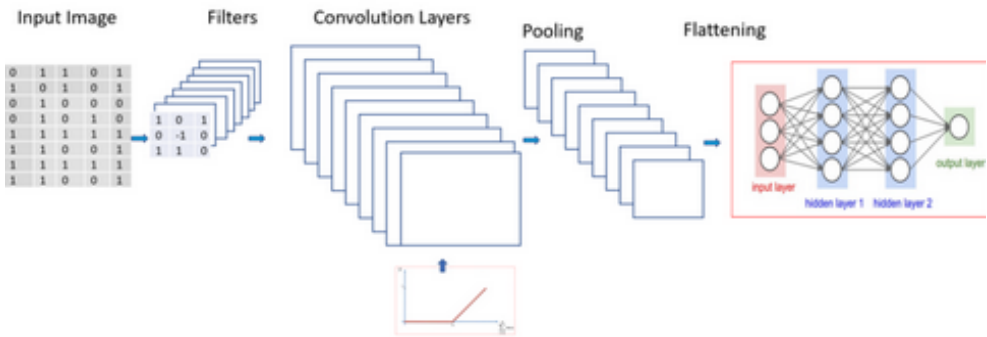

Figure 1. Feature extraction in neural network

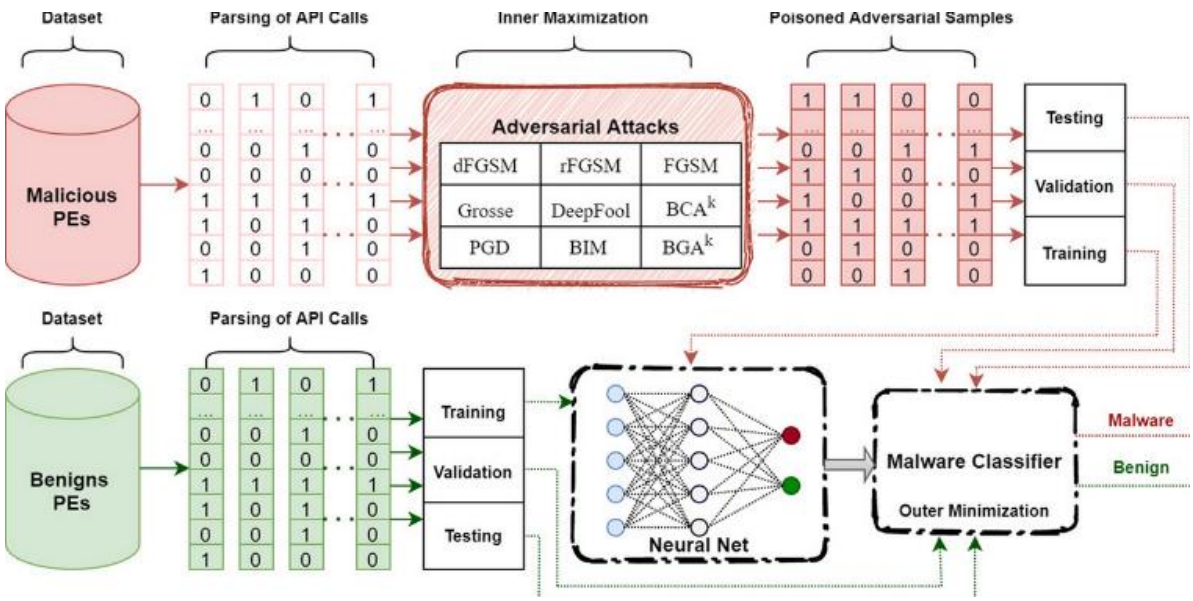

Figure 2. Attack mechanism

3.3 Functional attacks and rapid registration method (FGSM)

FGSM is a simple but strong technique for creating hostile cases. It takes advantage of the gradients of losing the form with regard to input data to determine the direction in which the input must be disturbed. FGSM algorithm is calculated and measured by a small fixed value, Epsilon (ε), to ensure accurate disorders remains [18].

To create a hostile example using FGSM, we start with a clean introduction and calculate the graduation function of the model in relation to the entry. The gradient represents the sensitivity of the model predictions of changes in input data. By annoying the inputs in the opposite direction of the gradual and expanding the scope of the disorder by ε, we can create a hostile example that misleads the form:

$$x' = x + \varepsilon \times sign\left(\nabla_x J\ (\theta, x, y)\right) \quad (3)$$

where, x' is the hostile counterpart, ε is the maximum permissible limit of disorder. In this way, the model can be misleading by poor classification of the rivalry.

Understanding the concepts behind hostile attacks and technologies such as FGSM is very important to develop strong defines mechanisms and ensure the safety of artificial intelligence systems in various fields. With the continued development of this field, more research and cooperation are needed to address ethical considerations and develop effective counter measures against hostile attacks. Figure 2 shows the attack mechanism used in female circumcision.

3.4 Defending against FGSM attacks

Several defense methods were proposed against female circumcision attacks, including training on hostile examples, adding random noise during the training or reasoning process, and re-training the model on hostile examples before classification operations.

3.4.1 Training on hostile examples

The hostile examples created using FGSM are added during the training process, which helps the model to learn to resist this type of attack. This technique, known as litigation training, is one of the most effective strategies to enhance the power of deep models against rivalry attacks. The function of the amending loss of litigants appears in the Eq. (4).

$$L'(\theta; x, y, x') = L(\theta; x, y) + L(\theta; x' + r, y) \quad (4)$$

where, x' represents the examples of the adversary generated, and r represents the amount of adversarial turbulence.

3.4.2 Adding random noise

This technique involves adding random noise to the data entered during training. By doing so, the model learns how to deal with annoying data and becomes more resistant to small changes in input. The process can be expressed as in Eq. (5).

$$x' = x + \varepsilon \quad (5)$$

where, ε is a matrix representing random noise.

3.4.3 Retraining method

Retraining method involves retraining the model on both original and opposing examples to improve its ability to distinguish between them. The loss function used in this approach is shown in Eq. (6).

$$L(\theta;\ x, y, x', y') = L(\theta; x, y) + L(\theta; x', y) \quad (6)$$

The aim is to improve the model's ability to distinguish between original examples and opposing counterparts, thereby minimizing the impact of disturbances on future taxonomic outcomes [19].

3.4.4 Retraining on original and adversarial data together

This approach combines the original examples x and the generated adversarial examples x' into one training set, using the joint loss function shown in Eq. (7).

$$L(\theta; D) = L(\theta; Doriginal) + L(\theta; Dadversarial) \quad (7)$$

where, D original is the original dataset and D adversarial is the hostile dataset. This method ensures that the model is trained on both types of data, increasing its resistance to hostile modifications.

3.4.5 Modified loss functions (L1/L2)

Using modified loss functions, such as L1 and L2, helps improve the resistance of deep models to hostile attacks. The loss function $L1$ is defined in Eq. (8).

$$L1\ (y, \hat{y}) = \sum i\ |yi - \hat{y}i| \quad (8)$$

And the $L2$ loss function is defined in Eq. (9).

$$L2\ (y, \hat{y}) = \frac{1}{2}\sum i\ (yi - \hat{y}i)^2 \quad (9)$$

These loss functions reduce the severity of gradients, making gradient-based attacks less effective [20].

The implementation of these defense mechanisms enhances the model's strength against hostile attacks, ensuring the reliability and security of AI systems in various applications.

4. SYSTEM AND RESULTS

We have developed a computer vision model classifying the Fashion Monist data collection using CNNS. The Fashion Monist Data collection is a common standard for assessing automated learning models, and it consists of gray gradient images of fashion elements such as shirts, pants, woolen jackets, dresses, coats, sandals, sports shoes, bags and long shoes. The goal is to accurately classify these images into their appropriate categories. In this section, we will discuss data exploration, models engineering, training, evaluation, preserving the form and downloading it for future use.

4.1 Data exploration

The Fashion MNIST dataset consists of two sets: training images and test images. The training kit contains 60,000 images, while the test kit consists of 10,000 images. Each image is 28 pixels long and wide, and pixel values range from 0 to 255, representing the intensity of grayscale.

4.2 Data pre-processing

To prepare data for the CNN model, we reshaped images

from three-dimensional arrays to four-dimensional matrices, allowing us to capture spatial information effectively. In addition, we normalized the pixel values by dividing them by 255, so that the values are between 0 and 1.

4.3 Model architecture

Our CNN model consists of several layers:

(1) Conflict layers: We start with a fodder layer that applies 64 candidates with a size of 3×3 to the input images. The ReLU activation function is not written, followed by an assembly layer by taking maximum values in each area of 2×2.

(2) The flat layer: This layer reinitializes the outputs of the ripples to a unilateral vector, to prepare it to enter into the full layers.

(3) Full class: Our dense layer contains 128 nerve cells and ReLU activation, learn sophisticated patterns of flat features. Finally, an output layer contains 10 nerve cells (one for each fashion category) uses SoftMax activation for categories.

4.4 Model training

After defining the structure of the model we assemble it by defining the optimizer, loss function and metrics. We use the Adam optimizer which is a popular choice for training neural networks. The loss function used is sparse class entropy and is suitable for multi-class classification tasks. During training the model is exposed to training images and labels. Corresponding to it for a certain number of epochs, the model learns to minimize the loss and improve its accuracy over time.

4.5 Model evaluation

Once the model is trained, we evaluate its performance on the test set. Test loss and accuracy are calculated using the evaluate function. The accuracy obtained is a measure of how well the model generalizes to unseen data. Higher accuracy indicates better performance. Figure 3 shows images generated by training a neural network.

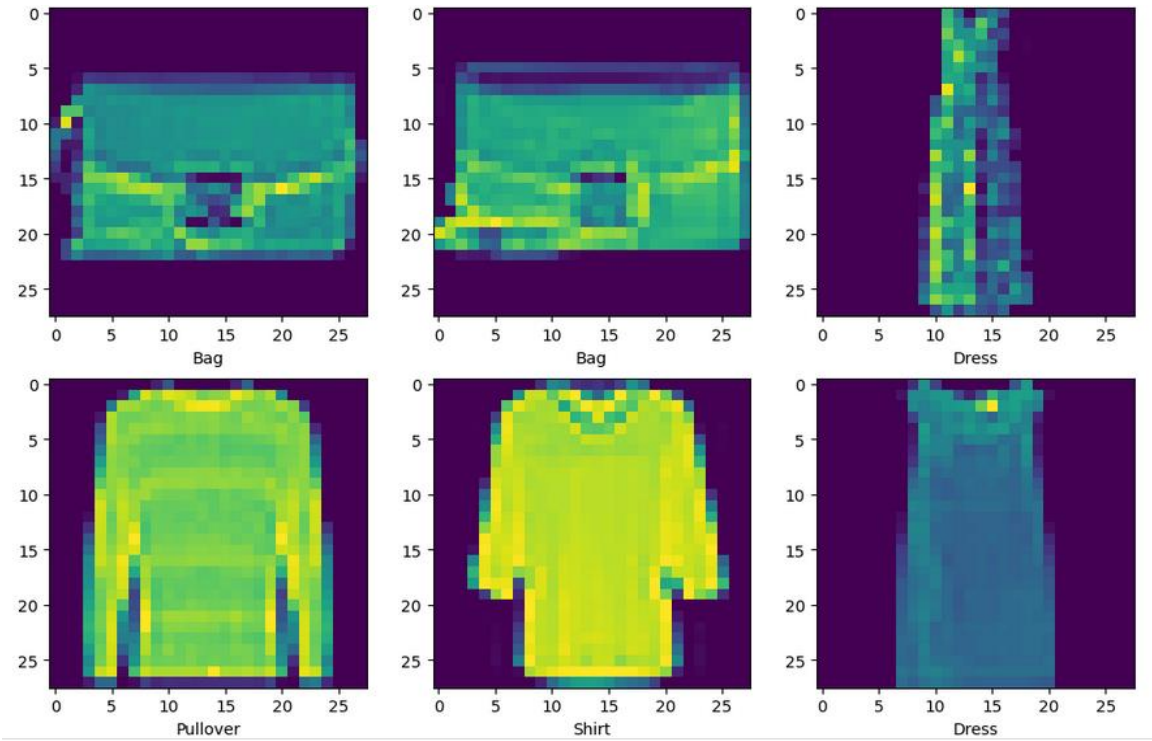

Figure 3. Images generated by training a neural network

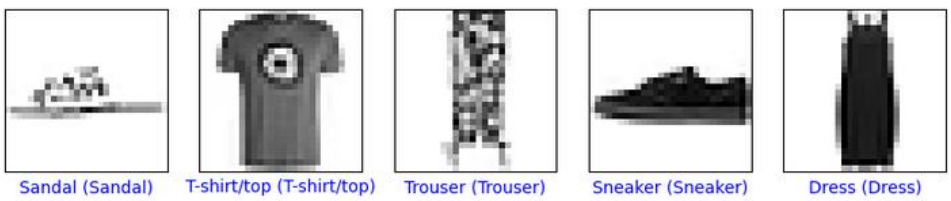

Figure 4. Images generated by training a neural network in gray scale

We convert images into gray shades instead of colors in automated learning applications and computer vision for multiple reasons:

(1) Less simplicity: The images in gray shades contain one channel (severity) for each pixel instead of 3 (RGB), which reduces the complex and calculation dimensions.

(2) Storage requirements: Gray-gradient images require less storage space, and this is important for large data groups.

(3) Stiffness: Discrimination may depend on essential patterns instead of absolute color values. Color removal removes potential problems from color lighting

differences and repercussions.

(4) Speed: Gray-shaded images faster, which speeds up the training process.

(5) Focus on the figure: For the tasks of focusing on shape, structure and fabric, gray grades maintain basic information while removing external dimensions.

(6) Compatibility: Many data collections, such as MNIST, are gray grades, which makes data trained on Gray-grades more diverse.

Figure 4 shows images created by training a neural network in grayscale.

5. DEFENSIVE AND OFFENSIVE ATTACKS

We have implemented offensive attacks on the trained Fashion-MNIST neural network model using the Python TensorFlow library and the Keras programming interface. The attack method follows the Fast Gradient Sign Method (FGSM). For each selected test image, a distorted image is generated by calculating the image gradients relative to the model loss function, as shown in Eq. (10).

$$[\nabla_X L = \{\partial L\}/\{\partial X\}] \quad (10)$$

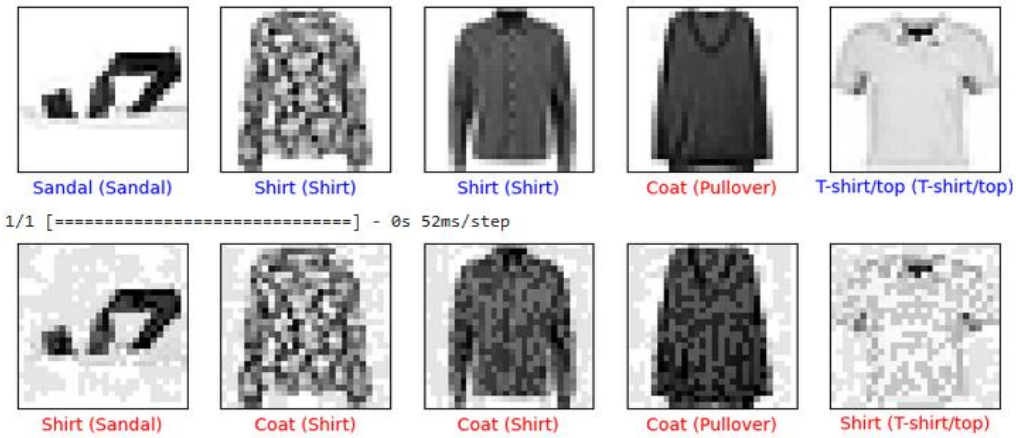

Figure 5. Images generated by grayscale neural network after being subjected to FGSM attack

Figure 5 shows images created by a neural network in grayscale after being attacked by FGSM.

Distorted pixel drawing shows the amount of distortion applied to each pixel in distorted images compared to the original images. By analyzing distortion values, we can understand which pixels have been modified and to what extent. This information is necessary to understand the rigidity of the model and its susceptibility to hostile attacks. Figure 6 shows the amount of distortion applied to each pixel during an FGSM attack.

We also implemented a Carlini Wagner (CW) attack, a Deep Fool attack, and a Projected Gradient Descent (PGD) attack. Figures 7-10 show images generated by the neural network in grayscale after experiencing these attacks, respectively.

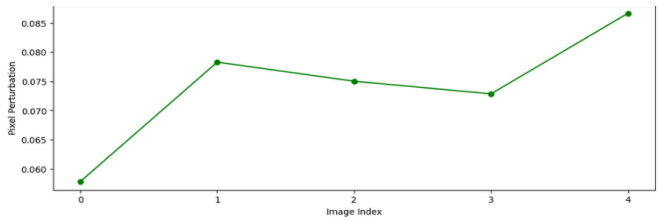

Figure 6. The amount of perturbation applied to each pixel in the images when an FGSM attack

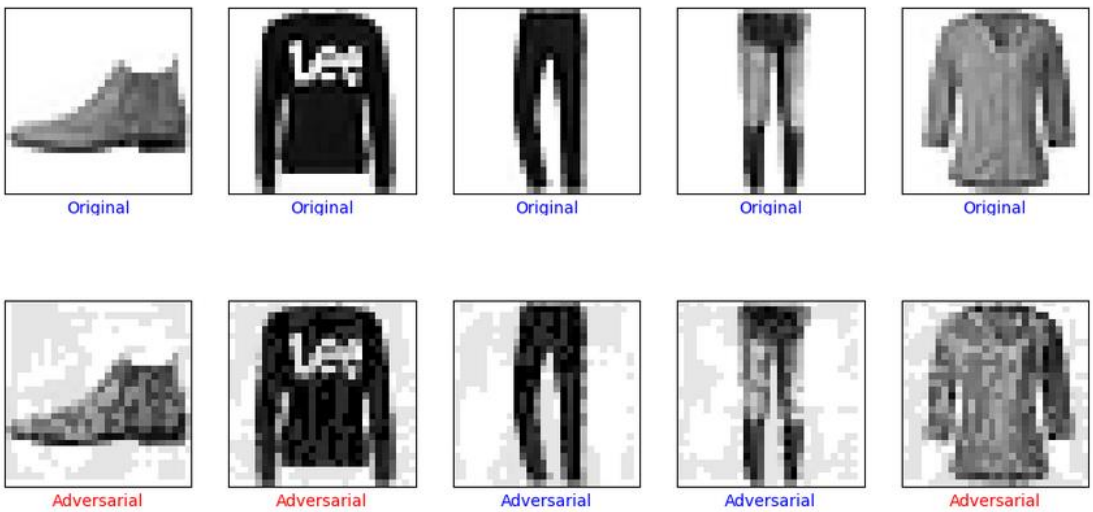

Figure 7. Images generated by a grayscale neural network after being subjected to a CW attack

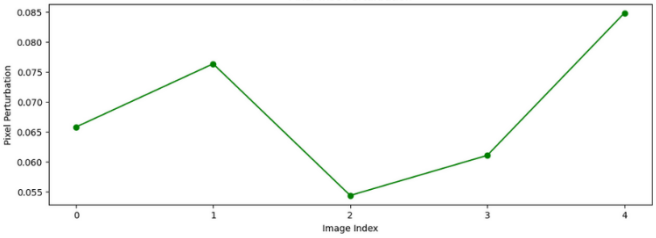

Figure 8. The amount of flicker applied to each pixel in images when chemical weapons are attacked

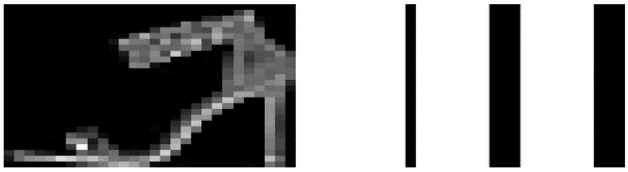

Figure 9. Images created by a grayscale neural network after being subjected to a deep foolish attack

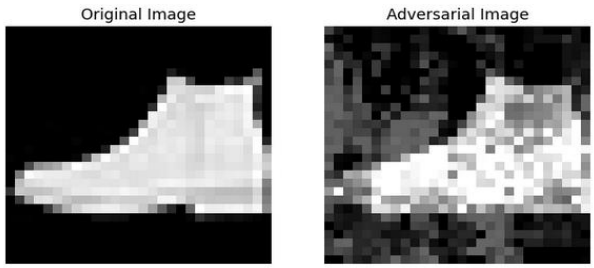

Figure 10. Images created by a grayscale neural network after being attacked by PGD

In order to improve the category accuracy after exposure to PGD, Deep Fool, Black Box and FSGM attacks, we have applied a set of transfers to the original image to improve the category accuracy:

(1) Change the image size: We changed the size of the original image with 28×28 pixels. This helps to unify the size of the image, reduce possible noise and excessive treatment.

(2) Employment of brightness and contrast: We modified the brightness of the image, increased contrast to highlight important parts of the image and improve the classification. This helps to improve the ability of the model to extract the features and patterns in the image.

(3) Improving contrast: We have improved light distribution in the image using a technique called improving contrast. This expands the range of colors and improving the details in the image.

(4) Victory homogeny: This removes noise and reduces excess detail in the image, thus enhancing the ability for accurate classification.

(5) Image normalization: We divide the image pixel values into 255 to convert the image into [0, 1] range. This makes it easier for the model to handle the normalized data correctly and after applying these transformations, we classified the transformed image using the model trained on the Fashion MNIST dataset.

Our goal is to improve the classification accuracy by improving the image quality and highlighting the important features in it. We have carried out the processes gradually according to the Table 1, where the accuracy of the different transformations is evaluated against different types of malicious attacks (FGSM, Deep Fool, PGD, and Black Box). Accuracy is measured based on the correct classification results of the model after image transformation.

Table 1. Accuracy of the classification

Transformation Type	FGSM	Deep Fool	PGD	Black Box
resized_image adjusted_image	0.82	0.831	0.835	0.862
resized_image adjusted_image enhanced_image	0.84	0.838	0.853	0.869
resized_image adjusted_image enhanced_image smoothed_image	0.86	0.841	0.859	0.872
resized_image adjusted_image enhanced_image smoothed_image normalized_image	0.91	0.85	0.87	0.88

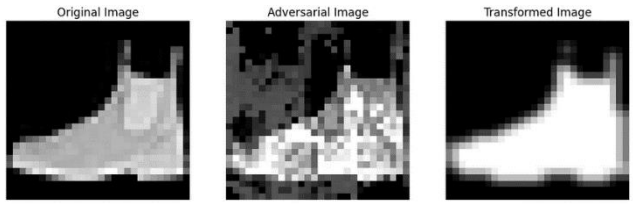

Figure 11. Images generated by grayscale neural network after applying transformations to the image in a PDG attack

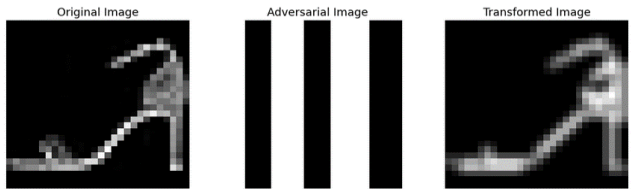

Figure 12. Images generated by grayscale neural network after applying transformations to the image in Deep Fool attack

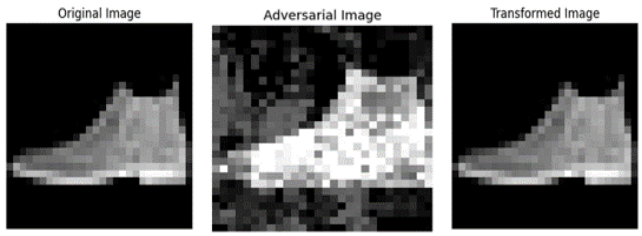

Figure 13. Images generated by a grayscale neural network after applying transformations to the image in a Black Box attack

From the Table 1, it can be seen that the use of different transformation operations contributes to enhancing the model's resistance to malicious attacks. Each time an additional transformation is applied, the accuracy achieved increases. For example, when using a sequential transformation of resized image, adjusted image, enhanced image, smoothed image, and normalized image, the highest

level of classification accuracy among all attack types is achieved. Overall, using a sequence of transformations appears to improve the overall performance of the model against malicious attacks, as higher classification accuracy is achieved with each additional transformation. Images generated by grayscale neural network after applying transformations to the image in a PDG attack, Deep Fool attack and Black Box attack are shown in Figures 11-13, respectively.

6. QUANTITATIVE RESULTS AND DISCUSSION

The quantitative results section deals with results including accuracy, predictive accuracy, retrieval, and F1 points, compared to our results with advanced methods.

When comparing the results, we find the following: For the basic network, the loss function was 0.1479 and the accuracy was 0.9452. The high values of accuracy and low values of loss show that the network is capable of achieving accurate classifications for the test set. While in the PGD attack, the loss function was 0.4914 and the accuracy was 0.8261. It is noted that the accuracy decreased and increased loss compared to the main network, and this means that the model has become less able to accurately recognize the misleading images created by the PGD attack. At the Deep Fool attack, the loss function was 0.4964 and the accuracy of 0.8245. Low accuracy and increased loss compared to the main network indicates that the model has become less able to identify misleading images resulting from the Deep Fool attack. In the Black box attack, the loss function was 0.3816 and the accuracy of 0.8648, as it aims to take advantage of the model without knowing its internal details. The decrease in loss and increase in accuracy compared to previous attacks shows that the model is able to better deal with misleading images generated by the black box attack. We find in the FGSM attack that the loss function was 0.4988 and accuracy was 0.8243, where the decrease in accuracy and increase in loss compared to the basic network shows that the model, it became less able to recognize images, and we find this according to the Table 2.

Table 2. A comparison between the results of the basic network without an attack and the network after being exposed to attacks

Attack	Loss	Accuracy
Baseline (no attack)	0.1479	0.9452
PGD	0.4914	0.8261
Deep Fool	0.4964	0.8245
Black Box	0.3816	0.8648
FGSM	0.4988	0.8243

High values of accuracy and low losses for the basic network indicate that the model achieves accurate classifications of the test group. However, the accuracy decreases and the loss increases when the model is exposed to hostile attacks, which shows its fragility.

The results show that applying the proposed sequence of transformations significantly enhances the accuracy of the model, increasing its resilience against hostile attacks.

From Table 3, the model's loss and accuracy are evaluated before and after image transformations for different types of malicious attacks (PGD, Deep Fool, Black Box, FGSM). The loss is measured in one standard unit (e.g., average loss) and the accuracy is measured by the correct classification rate of the model. From Table 3, it can be seen that malicious attacks cause increased loss and decreased accuracy of the base model before any image transformation. However, after applying progressive image transformations, an improvement in loss and an increase in accuracy of the model against malicious attacks are achieved. For example, when using a PGD attack, a higher loss and lower accuracy of the model are achieved. However, after image conversion, the loss level decreases and the resolution level increases.

Table 3. Comparison between network results before image transformations and after image transformations in the final stage

Attack	Before Image Conversions		After Image Transformations	
Classification Type	Accuracy	Loss	Accuracy	Loss
Core network (no attack)	0.9452	0.1479	0.9452	0.1479
PGD	0.8261	0.4914	0.87	0.31
Deep Fool	0.8245	0.4964	0.85	0.291
Black Box	0.8648	0.3816	0.88	0.25
FGSM	0.8243	0.4988	0.91	0.15

In general, it is shown that the use of sequential image transformations contributes to improving the model's performance and increasing its resistance to malicious attacks, as lower loss and higher accuracy are achieved after applying the transformations.

6.1 Discuss constraints, future research directions and practical applications

6.1.1 Study restrictions

Our study seeks to categorize Fashion MNIST accurately using the CNNS and assess the solidity of the model against hostilities. However, there are some restrictions that affect our study results.

6.1.2 Form complicated

The structure used for the model may be simple compared to advanced models such as Resnet or Efficiencies, which may achieve better performance in the classification of images and providing higher hardness against attacks.

6.1.3 Quality of attacks

We focused on a limited set of hostile attacks, such as FGSM, Deep Fool, PGD and CW. Our study does not include other types of attacks such as white opponent attacks or unreasonable attacks, which may lead to different results.

6.1.4 Defenses and defensive technologies

Despite the improvements made by transfers, these transfers may be ineffective against certain types of hostile attacks. Also, some defenses, such as neuron-trained, may require more verification.

6.1.5 Circular

The study is limited to the Fashion MNIST data collection, and does not deal with more complex or different patterns, which limits the ability of the results to generalize.

6.1.6 Future search trends

(1) Exploration of advanced models: The performance of

the model can be improved by exploring advanced neurological network models such as Deep Neural Networks or Deep Convolutional Networks that may offer higher accuracy and better performance against hostile attacks.

(2) More advanced defenses: Advanced defines methods such as weighted training, augmented training, and defensive encryption can be studied. In addition, methods such as the exchange of trained nervous networks can be explored for pre-disdain data.

(3) Analysis of the effect of transfers: Detailed studies must be conducted to assess the effect of different transfer components on the solidity of the model against hostile attacks. Study the interaction between transfers and how to improve them can provide new visions.

(4) Expanding the scope of the study: The study can be expanded to include other data groups such as CIFAR-10 or ImageNet, allowing the effectiveness of the effectiveness of models and defenses in various and complex contexts.

(5) Research in new attacks: Analyzing the effect of new and advanced attacks can provide valuable information about the solidity of the model and direct the development of future defenses.

6.1.7 Practical applications

Our study results applications include:

(1) Safety applications: Improving the solid of models against hostilities is of great importance in security applications such as face recognition systems, fraud, and text identification.

(2) Classification of images in realistic environments: Improved defense technologies can be applied in commercial systems that depend on the classification of images, such as quality control systems, and medical image analysis tools.

(3) Artificial Intelligence Research: The results provide a strong basis for developing new defense techniques and reviewing classification methods in artificial intelligence.

(4) Clarify the effectiveness of defenses: Our results showed an improvement in the performance of the model when applying the proposed defenses compared to traditional models. Defenses such as transfers (re-scaling, modification, improving contrast, camouflage, and normalization) have proven effective in improving the accuracy of the model in the face of hostile attacks.

The effectiveness of these defenses was experimentally verified by comparing the performance of the model against hostile attacks before and after the implementation of the defenses. Experimental data showed a remarkable improvement in accuracy and a decrease in loss, which demonstrates the effectiveness of defenses in reducing the impact of hostilities.

Table 4. Model pilot performance with and without impellers

Attack Type	Accuracy Without Defenses	Accuracy with Defenses	Improvement (%)
FGSM	0.8243	0.8685	+ 5.35
Deep Fool	0.8245	0.8510	+ 3.23
PGD	0.8261	0.8591	+ 4.00
Black Box	0.8648	0.8785	+ 1.58

Table 4 summarizes the experimental performance of the model with and without impellers, showing the observed improvement after the application of defenses.

Finally, it is worth to mention other applications of deep learning system including solution of fuzzy linear programming problem [21], applying numerical model for the prediction of aeration in mechanical systems [22] and simulation of plasma systems [23].

7. CONCLUSIONS

In this study, we explored the impact of hostile attacks on the CNN nervous network model (CNN) that was trained on the Fashion Monist data collection and evaluated the various defense mechanisms. The results we find reveals that the application of a series of photo transfers greatly enhances the elasticity of the model in the face of hostile disorders. Specifically, our results appear in accuracy with applied defenses, with increases of up to 5.35% for FGSM attacks, 3.23% for Deep Fool attacks, 4.00% for PGD attacks, 1.58% for black box attacks.

These technologies outperform traditional methods, which indicates their practical character and their effectiveness in preserving the durability of the model. However, there are restrictions, such as the difference in the effectiveness of these defenses depending on the type and intensity of the attack. Future research should focus on additional defense mechanisms, conducting detection studies to understand the contribution of each transformation better, and testing these technologies on other data collections and more advanced attacks. Practical applications for this study are important, as strong models are essential for highly reliable and security applications in areas such as independent systems, financial prediction and medical diagnosis. In general, this study highlights the importance of integrating effective defense mechanisms to protect automatic learning models from hostile threats, indicating that pre-treatment techniques can be a valuable strategy to enhance the safety of models in hostilities. Future work should continue to improve these methods and expand their scope to ensure their effectiveness in various and advanced contexts.

ACKNOWLEDGMENT

The authors are grateful to Al-ma'moon University College, for supporting this work.

REFERENCES

[1] Rahman, C.R., Arko, P.S., Ali, M.E., Khan, M.A.I., Apon, S.H., Nowrin, F., Wasif, A. (2020). Identification and recognition of rice diseases and pests using convolutional neural networks. Biosystems Engineering, 194: 112-120. https://doi.org/10.1016/j.biosystemseng.2020.03.020

[2] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[3] Simonyan, K., Zisserman, A. (2014). Very deep

convolutional networks for large-scale image recognition. Computer Science. https://doi.org/10.48550/arXiv.1409.1556

[4] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 2261-2269. https://doi.org/10.48550/arXiv.1608.06993

[5] Bhatia, A., Chug, A., Singh, A.P., Singh, D. (2022). A hybrid approach for noise reduction-based optimal classifier using genetic algorithm: A case study in plant disease prediction. Intelligent Data Analysis, 26(4): 1023-1049. https://doi.org/10.3233/IDA-216011

[6] Zhang, S., Kong, W., Wang, Z. (2017). Plant classification method based on dictionary learning with sparse representation. Acta Agriculturae Zhejiangensis, 29(2): 338-344. https://doi.org/10.3969/j.issn.1004-1524.2017.02.22

[7] Guadarrama, L., Paredes, C., Mercado, O. (2022). Plant disease diagnosis in the visible spectrum. Applied Sciences, 12(4): 2199. https://doi.org/10.3390/app12042199

[8] Kaur, M., Bhatia, R. (2019). Development of an improved tomato leaf disease detection and classification method. In 2019 IEEE Conference on Information and Communication Technology, Allahabad, India, pp. 1-5. https://doi.org/10.1109/CICT48419.2019.9066230

[9] Tang, W., Huang, Z. (2021). Lightweight model of tomato leaf diseases identification based on knowledge distillation. Journal of Jiangsu Agriculture, 3: 570-578. https://doi.org/10.3969/j.issn.1000-4440.2021.03.004

[10] Yang, L., Zhang, R.Y., Li, L., Xie, X. (2021). Simam: A simple, parameter-free attention module for convolutional neural networks. In International Conference on Machine Learning, pp. 11863-11874. https://proceedings.mlr.press/v139/yang21o.

[11] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 1625-1634. https://doi.org/10.1109/CVPR.2018.00175

[12] Lu, J., Sibai, H., Fabry, E., Forsyth, D. (2017). No need to worry about adversarial examples in object detection in autonomous vehicles. Preprint arXiv:1707.03501. https://doi.org/10.48550/arXiv.1707.03501

[13] Liu, Y., Zhang, W., Li, S., Yu, N. (2017). Enhanced attacks on defensively distilled deep neural networks. Preprint arXiv:1711.05934. https://doi.org/10.48550/arXiv.1711.05934

[14] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D. (1989). Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4): 541-551. https://doi.org/10.1162/neco.1989.1.4.541

[15] Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2: 1-21. https://doi.org/10.1186/s40537-014-0007-7

[16] Nguyen, L., Wang, S., Sinha, A. (2018). A learning and masking approach to secure learning. In Decision and Game Theory for Security: 9th International Conference, GameSec 2018, Seattle, WA, USA, pp. 453-464. https://doi.org/10.1007/978-3-030-01554-1_26

[17] Carlini, N., Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 3-14. https://doi.org/10.1145/3128572.3140444

[18] Wu, H., Han, H., Wang, X., Sun, S. (2020). Research on artificial intelligence enhancing Internet of Things security: A survey. IEEE Access, 8: 153826-153848. https://doi.org/10.1109/ACCESS.2020.3018170

[19] Liu, P., Xu, X., Wang, W. (2022). Threats, attacks and defenses to federated learning: Issues, taxonomy and perspectives. Cybersecurity, 5(1): 4. https://doi.org/10.1186/s42400-021-00105-6

[20] Bhandari, K., Kumar, K., Sangal, A.L. (2023). Artificial Intelligence in Software Engineering: Perspectives and Challenges. In 2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, pp. 133-137. https://doi.org/10.1109/ICSCCC58608.2023.10176436

[21] Prakash, Y., Appasamy, S. (2023). Optimal solution for fully Spherical Fuzzy Linear Programming Problem. Mathematical Modelling of Engineering Problems, 10(5): 1611-1618. https://doi.org/10.18280/mmep.100511

[22] Mastrone, M.N., Concli, F. (2023). Implementation of a numerical model for the prediction of aeration in mechanical systems. International Journal of Computational Methods and Experimental Measurements, 11(2): 65-71. https://doi.org/10.18280/ijcmem.110201

[23] Gutierrez, A.D., Alvarez, L.F. (2022). Simulation of plasma assisted supersonic combustion over a flat wall. Mathematical Modelling of Engineering Problems, 9(4): 862-872. https://doi.org/10.18280/mmep.090402