International Journal of Communication Networks and Information Security

ISSN: 2073-607X, 2076-0930
Volume 14 Issue 03 Year 2023

Study on Network Virtual Printing Sculpture Design using Artificial Intelligence

Wei Xu
Silpakorn University, Faculty of Decorative Art, 31 Naphalan Rd., Pra Ma-ha borom ratchawang phanakorn Bangkok
wei_x@su.ac.th
Veerawat Sirivesmas
Silpakorn University, Faculty of Decorative Art, 31 Naphalan Rd., Pra Ma-ha borom ratchawang phanakorn Bangkok
veerawatsi@gmail.com

Article History

Received: 24 February 2023
Revised: 07 April 2023
Accepted: 18 May 2023

Abstract

Sculptures are visionaries of a country's culture from time immemorial. Chinese sculptures hold an aesthetic value in the global market, catalysed by opening the country's gates. On the other hand, this paved the way for many duplicates and replicates of the original sculptures, defaming the entire artwork. This work proposes a defrauding model that deploys a Siamese-based Convolutional Neural Network (S-CNN) that effectively detects the mimicked sculpture images. Nevertheless, adversarial attacks are gaining momentum, compromising the deep learning models to make predictions for faked or forged images. The work uses a Simplified Graph Convolutional Network (SGCN) to misclassify the adversarial images generated by the Fast Gradient Sign Method (FGSM) to combat this attack. The model's training is done with adversarial images of the Imagenet dataset. By transfer learning, the model is rested for its efficacy in identifying the adversarial examples of the Chinese God images dataset. The results showed that the proposed model could detect the generated adversarial examples with a reasonable misclassification rate.

Keywords: Deep Learning, Siamese Neural Networks, Adversarial Attack, Transfer Learning, Imagenet, Chinese Sculpture, Graph Convolutional Networks

1. Introduction

The era of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) has acclaimed unprecedented attention from almost all fields, from manufacturing, maintenance, education, and health to fine arts [1]. This has made academicians and industrialists explore these contemporary technologies to build solutions to many real-world problems. Many ML models were developed and deployed for several use cases. However, compared with the classical ML models, the DL models are gaining more momentum as they are more accurate and highly adaptable [2]. This has made DL widely deployed in computer vision, natural language processing and speech recognition. It allows the models with multiple processing layers to perceive, learn and even represent the data with multiple abstraction levels. The inherent processing of DL models mimics the brain's perceptive power, which can comprehend multimodal information by capturing and learning intricate data structures. In addition, DL has various methods like hierarchical probabilistic models, deep neural networks, neural networks with memory, and supervised and unsupervised feature learning models. The surge of interest among the researchers in DL models is because they outperform other state-of-the-art techniques.

DL systems are prone to multiple types of attacks like backdoor attacks, model stealing attacks, adversarial examples, membership inference attacks, etc., which are severe threats to the privacy and security of these systems. An adversarial attack in image recognition is considered to modify the source image so that the changes are invisible to human vision [3]. The modified image is a negative image that will be treated as original by any AI-based model. Another legitimate application is a content modification to bypass the security measure by automatic moderation algorithms [4]. Technically, the attackers focus on the gradients of the loss function and change their direction, eventually decreasing the models' predictive power [5]. Because of these reasons, the applicability of the technology is limited in sensitive applications like the military. The CIA security triad categorises the attacks on DL models as breaches of Confidentiality, Integrity and Availability, as shown in Figure 1 [6].

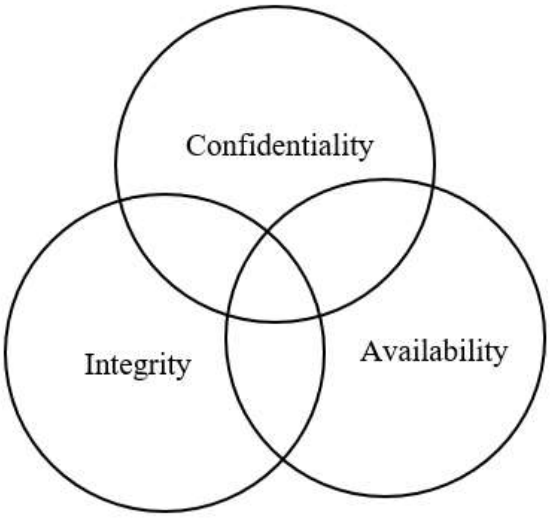


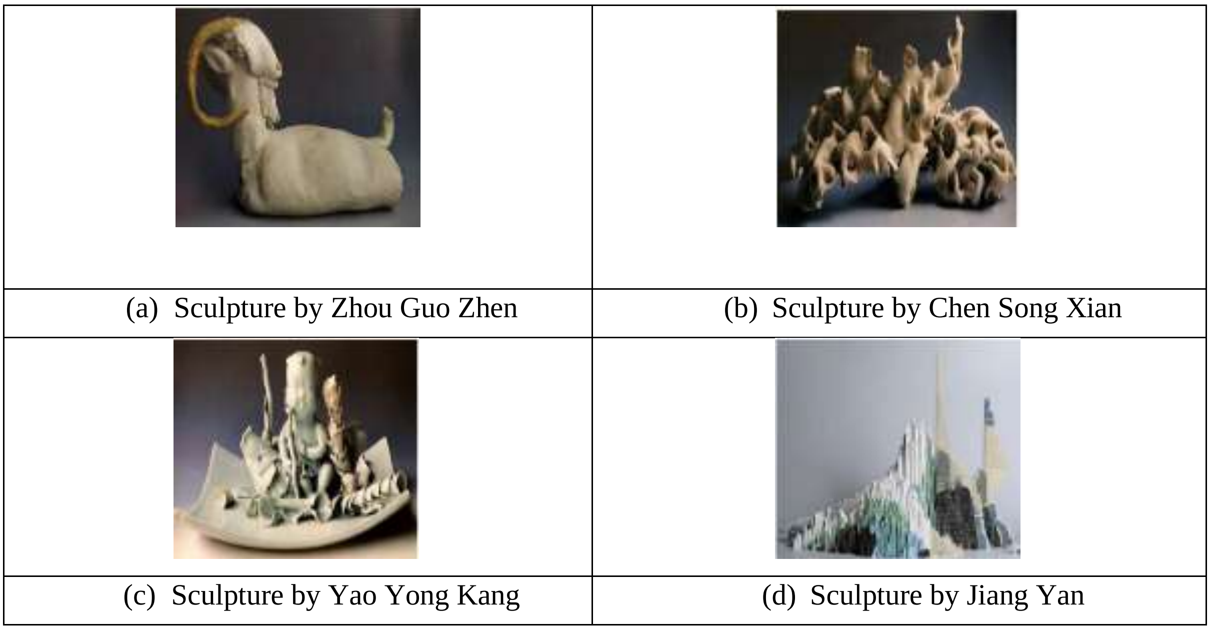
Figure 1. CIA Triad of Information Security

Technological advancements and globalisation have opened doors for unrestricted access to all resources, especially artistic works like sculptures, paintings and drawings [7]. Thus, art media classification has sought more attention due to the recent surge in adversarial attacks on these artistic elements. The artistic elements are more vulnerable to adversarial attacks because of their complex extraction process and the analysis of intricate features, especially high-valued unique artistic pieces. Sculptures hold a predominant place in fine arts. A sculpture is perceived as the representation of cultures' beliefs, practices and views. Chinese ceramics occupy a unique position in the world of sculptures. In contemporary sculpture making, the content and artistic values make them an international tool for communication. The ceramic sculpture is a 3D art form in which the clay is fired at high temperatures to get the desired shape and patterns to capture even the emotional content [8]. These visual art forms disclose the economic, social and even political issues of the period in which they were made.

Table 1. CIA triads in the Perspective of Information Security of Sculptures

CIA properties	**Explanation**
Confidentiality	• Assuring that the information is available only to legitimate users with proper access. • The confidentiality mechanisms protect the sculpture images from unauthorised access. • It must also ensure total protection of data during its transmission.
Integrity	• Assuring that the sculptures and sculpture images are kept intact in their correct state without tampering. • The completeness and accuracy of the sculptures should be ensured even during their processing and transmission. • Any alterations, modifications and other minor and significant changes made to the data during its transmission, processing and storage must be found using appropriate mechanisms.
Availability	• Assuring that only legitimate users can access legitimate assets only when necessary. • This includes building fault tolerance mechanisms, handling data loss, backup policies etc.

China's cultural-artistic developments were isolated from other parts of the world for centuries. The unprecedented cultural change experienced by the country in the last few decades has brought a cultural revolution [9]. In the early 1980s, Western art influenced Chinese artists, which caused a tremendous turnover of Chinese artistic styles. Contemporary Chinese ceramic sculpture is reborn and gained worldwide popularity, which was initially a ritual art of China [10]. One of the essential elements of any form of contemporary art is the subject matter or the theme on which it is based. The aesthetic beauty of the Chinese sculptures, influenced by Chinese culture and tradition, gives artistic appeal to the art pieces [11]. Today's ceramic artists in the country focus more on creating artworks with personal feelings than traditional craft. As a result, the sculpture art is geared towards aesthetic improvement like the country's economy. Fig 2 shows some of the sculpture works by famous Chinese artists.



(a) Sculpture by Zhou Guo Zhen

(b) Sculpture by Chen Song Xian

(c) Sculpture by Yao Yong Kang

(d) Sculpture by Jiang Yan

Figure 2. Sculpting Works by Famous Chinese Artists [12]

The aesthetic appeal of these works has a unique value in the global market. This paved the way for creating fake or copied versions of the works, affecting the original works' creditability. As the human perception of artistic works is often biased, the people interested in collecting the sculptures face challenges in determining the authenticity of the sculptures, which can be brought under the umbrella of adversarial attack. An adversarial attack generally is a means to generate adversarial examples. These examples are fake and closely resemble the original sculpture as they are designed to counterfeit the original works. Fig 3 shows the adversarial attack model, and Table 1 describes the CIA triads concerning maintaining the information security of Chinese sculptures and their images.

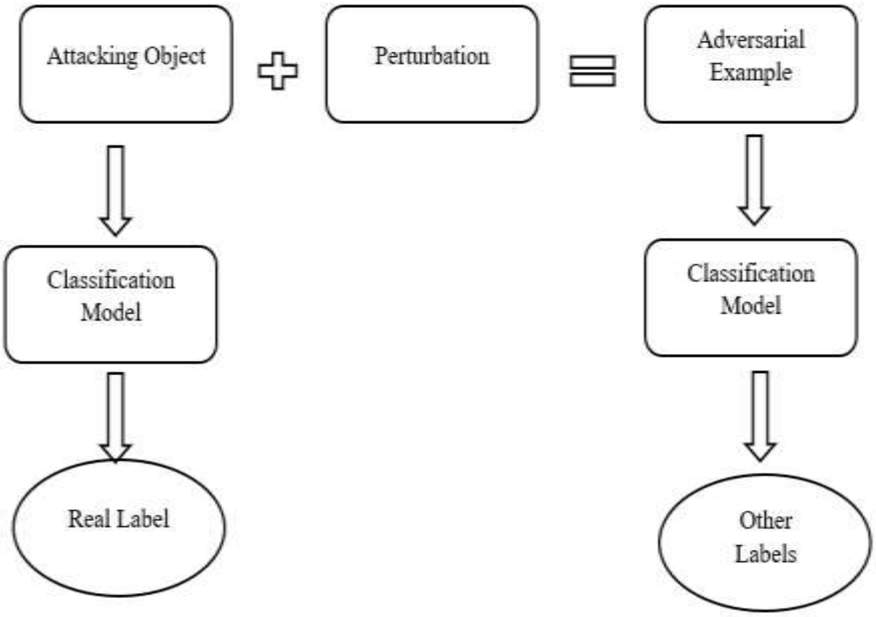


Figure 3. Adversarial Attack Model

1.1 Terminologies

This section elucidates the standard technical terms that are related to adversarial attacks on Chinese sculpture images:

The adversarial example sculptures are modified versions of original artwork, which are purposefully disturbed to fool the classification models [13].

Adversarial training is a practical approach to defending against adversarial examples by making the models learn the adversarial and original ones [14]. Black-box attacks occur when adversarial examples.

In White-box attacks, the attacker has complete knowledge of the model's parameters, training data, training method, and architecture.

In black box attacks, the attacker does not know the model's parameters. Hence the attacker employs a different model to generate adversarial examples.

The mechanism used by the predictor model to isolate the adversarial examples from the original ones is termed a detector.

The process of imperceptible perturbations generates adversarial examples less discernible to humans.

The performance of ML and DL models has surpassed many other techniques through their intricate feature extraction from training images that do not depend on handcrafted feature detectors used by other methods. But the most significant challenge is related to the reliability of the models as small-scale perturbations made on the input image of the sculpture in an adversarial attack can change the complete predictions. The integrity attacks on sculpture images impact the model's output, thus declining its overall performance. The fundamental aim of generating adversarial examples for sculptures is that the adversary examples consist of imperceptible perturbations that will act as testing examples for models developed to detect duplicates [15]. The adversarial examples often tap the DL model's inference or feature learning phase and do not entirely affect the model. These are generally called backdoor attacks, as they inject hidden backdoors into the DL models by corrupting or poisoning the data so that the model increases the rate of true negatives and false positives.

Multiple defence mechanisms exist to combat the adversarial examples to ensure the integrity of DL-based defrauding systems [16]. Some predominant methods include gradient masking, adversarial training and detecting adversaries [17]. These defence mechanisms are built by understanding the existing attacks during the inference phase and controlling the inducement of those images by changing the input pattern. Nevertheless, these mechanisms do not consider the similarity among these attacks affecting the integrity of the sculpture images, as the DL models cannot render a complete and comprehensive defence against the integrity. Integrating the existing defence mechanisms into the models to handle the integrity attacks will incur high computational costs, lowering its efficacy. This necessitates enforcing a universal defence mechanism to combat the integrity attacks against artistic works, which is still an open research issue.

This work proposes a Siamese detector network that combats the adversarial attacks on Chinese sculpture works. The work advances by cleaning the images to create a proper validation set. The model estimates the similarity between the negative and original images using the triplet loss function. Finally, the model isolates the fake or adversarial image from the authentic sculpture images. The primary contributions of the work are summarised as follows:

(1) Explaining the integrity attacks against DL defrauding models with a primary focus on backdoor adversarial examples.

(2) Designing a Sienese detector network for DL defrauding models that isolates the integrity attacks by estimating the similarity between the adversarial and genuine sculpture works.

(3) Empirically investigating the performance of the proposed model using custom-defined and synthetic Chinese sculpture images.

2. Literature Review

This section describes the essential works in the literature focused on combating adversarial attacks in various domains.

Sandy Huang et al. showed that adversarial attacks could also target neural network policies [18]. The adversaries induced minor perturbations to assess the performance, which was found to decline drastically as the number of perturbations increased. A much broader group of momentum-based iterative algorithms are proposed by Yinpeng Dong et al. to improve the adversarial attacks by stabilising the gradient update directions [19]. The results indicate that this method was very effective against black-box attacks.

Recently, medical images have been subjected to a high rate of adversarial attacks. Samuel Finlayson et al. briefed the motivations of adversarial attacks in the medical domain [20]. On the other hand, Han Xu et al. surveyed the mechanisms available to combat adversarial attacks on graphs and texts [21]. Sparse matrix representations of the input data are employed to combat adversarial attacks in work proposed by Soorya Gopalakrishnan et al. [22]. The results showed that this method was effective against attacks on DL models. Adversarial attacks on textual content are also increasing at an alarming rate. A Recurrent Neural Network (RNN) based model which imbibes new backoff strategies to handle rare words is developed by Danish Pruthi et al. [23].

Olga Taran et al. propose a novel Key-based Diversified Aggregation method as a defence strategy to handle both grey and black-box adversarial attacks [24]. The randomisation method used in this work prevents backpropagation in gradients, thus confining the adversarial attack to bypass the model. Another model is to prevent adversarial attacks on textual content by using an untrained iterative approach that integrates context-independent and dependent character-level features [25]. Erik Jones et al. introduced a robust encoding framework, which assured robustness in preventing adversarial attacks on textural content [26]. The encoding function in this work mapped the sentences to discrete but smaller spaces to develop a better-fidelity system.

Deep Armour is a malware cognitive system to perform classification against adversarial attacks by a voting system [27]. The work used ML classifiers such as random forest, structure2vec and multi-layer perceptron and was competent against white-box evasion attacks. Zhaoyu Chen et al. proposed adversarial watermarks to handle fake facial image recognition models [28]. The Cross-Model Universal Adversarial Watermark deployed in this work outperforms another state other state-of-the-art methods. Zach Jorgensen et al. proposed a counterattack method for spam detection using logistic regression [29]. The principle behind this work is transforming the mail into a bag of several segments and then applying multi-instance regression on each bag. Xiaohu Du et al. proposed a contemporary Robust Adversarial Training to defend against word-level adversarial attacks [30].

Most of the adversarial attack counter-measured focus on gradient-based methods. Mingyuan Fan et al. suggested a Non-Gradient Attack method to mitigate adversarial attacks [31]. Rajeev Sahay et al. proposes an efficient method for combating the Fast Gradient Sign type of attack [32]. This method compresses and denoises the data, which is then processed by a multi-layer Denoising Autoencoder, which acts as a defence mechanism. Jianpeng Ke et al. designed DisPAT, a framework to mitigate adversarial examples in text classifiers [33]. This method employed a deep neural network-based discriminator with neuron-salience-based pruning.

Sibo Song et al. proposed the Saak transform, a representation method for a spatial-spectral view of images [34] that efficiently combats adversarial attacks. The idea of including an anti-adversary layer in DL models is proposed by Motasem Alfarra et al. [35] to control the adversarial attack. This method was found helpful in isolating black box attacks. Defense-GAN was proposed by Pouya Samangouei et al. as a novel framework that uses generative models to prevent adversarial attacks [36]. Junyan Peng et al. deployed a revised Naive Bayes classifier to control adversarial attacks by including additional weight depending on the count spam and ham feature [37]. A two-phase spelling correction model, which comprises a detector and correcter mechanism to combat adversarial attacks, is proposed by Zhengxiao Liu et al. [38]. This work was more efficient when tested against Stanford Sentiment Treebank.

Thus, the brief literature shows that adversarial attacks are common in almost all fields. The inference from the survey suggests that only a little work is devoted to preserving the integrity of artistic works, which is quintessential for maintaining the aesthetic value of the works.

3. Methodology

This section explains the sculpture defrauding system, simulated adversarial attack on the DL system and its defence measure.

3.1 Sculpture Defrauding System Using Siamese Network

The sculpture defrauding system uses Siamese Convolutional Neural Network (S-CNN) augmented with a Triplet Loss function, as the model has a sound success track in one-shot learning [39]. The model uses three dependent CNNs that share their weights, generating Feature Encodings (FE) of the learned images. The S-CNN used in this work uses a triplet function against the conventional contrastive loss to learn the features represented as embedding vectors from three independent images. Triplet Loss ensures a significant margin between negative images' similarity distances and positive images' similarity distances. This is in contrast to the contrastive loss that considers the margin value between dissimilar pairs, which may converge too early local minima, thus failing to distinguish the adversarial image. On the other hand, the Triplet Loss function places the model to vector space in a better position.

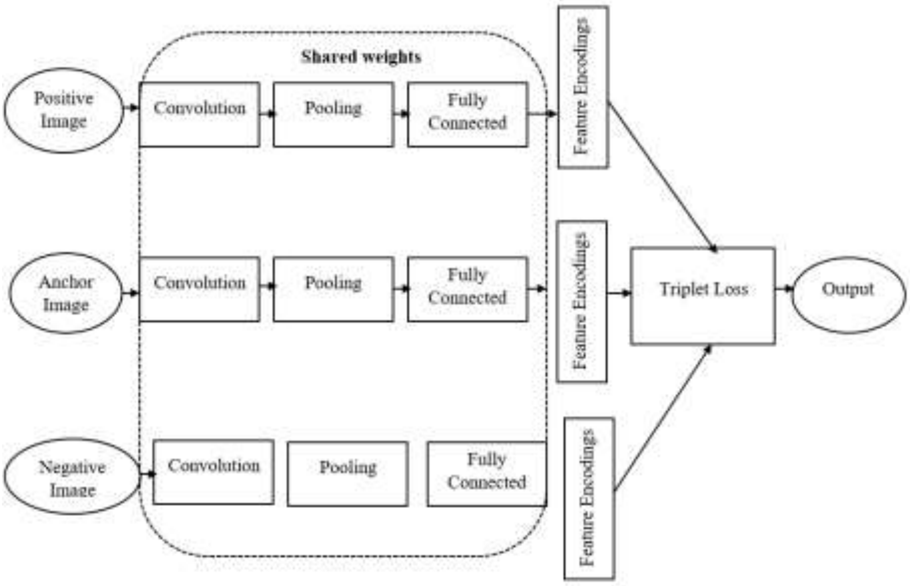


Figure 4. Siamese Network with Triplet Loss Function Used as Sculpture Defrauding System

The feature extraction happens in the convolution layers of the CNN by deploying filters on the input image, according to Equation 1.

$$f_{i,j}^{(c)} = h(\sum_{k=0}^{Cy-1} \sum_{l=0}^{cx-1} w_{k,l}^{c} f_{(i+j,k+l)}^{(c-1)} + b^{c} \qquad (1)$$

In the above equation, the term indicates the neuron's weight (i, j) at the cth convolutional layer, whereas signifies the bias value at the same layer. The filter size is set as cx x cx, and h indicates the activation function used in the layer. After the convolution layer, the pooling layer will be augmented to downsample the extracted features to mitigate the computational complexity. The fully connected layer implements the classifier for the defrauding system. This work uses Graph Convolutional Network, initially semi-supervised learning that relies on graph-structured data. The convolutions are focused towards establishing a localised first-order approximation of the graph of the images. As they can linearly scale the number of edges and graphs, these networks can capture even minute differences in the sculpture images.

In a trained S-CNN, the distance between the FE for dissimilar pairs will have higher values and vice versa. The design consideration of the S-CNN is that it can compare the distance between the sculpture images only in pairs. However, the triplet loss estimated in the proposed work is destined to learn the FE through triplets rather than as pairs. The set of triplets formed is <anchor (Ia), positive (Ip), negative (In)> images as mentioned in Table 2.

Table 2. Objects Considered in S-CNN Using Triplet Loss Function

Images	**Description**
Anchor sculpture	Reference sculpture image
Positive sculpture image	Image with the same label as the anchor sculpture image
Negative sculpture image	An image whose label is different from that of the anchor and positive sculpture image

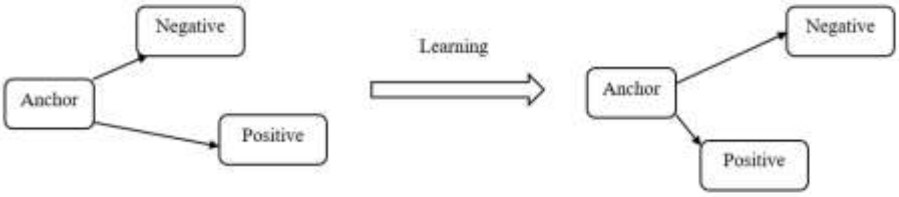


Figure 5. Learning Using Triplet Loss Function

The distance between the anchor sculpture images and positive sculpture images is much lesser than that between the negative and anchor sculpture images. The objective function for parameter training from the extracted feature encodings is governed by equation 2.

$$E = \sum_{(a,p,n)\epsilon\theta} \max(0, m + \left\|f(I_a) - f(I_p)\right\|_2^2 - \|f(I_a) - f(I_n)\|_2^2) \tag{2}$$

The FE is denoted as f(.), and the parameter m signifies the distance between the clusters formed.

3.2 Simulating the Adversarial Attack in the S-CNN

The attack of the S-CNN is simulated using the Fast Gradient Sign Method (FGSM), which is a successful white box attack that accesses the classification layers of the network [40]. The image perturbation method adopted in this work disturbs the sculpture images at the classification layer. In general, digital images use 8 bits to represent a pixel. Because of this, the models ignore the information which is lower than 1/255. But in the case of adversarial attacks, the perturbations in the image will be minuscule, which makes the model ignore them. To combat this challenge, the adversarial perturbations increase the activation by wT φ, as in Equation 3.

$$w^T x' = w^T x + w^T \varphi \tag{3}$$

This value of increase is done by estimating φ = sign (w). The FGSM used in work adds noise or perturbations to the images aligned to the direction of the loss function's gradient concerning the input data. This work focuses on using FGSM in both targeted as well as untargeted ways:

Targeted attack: This type of attack modifies the input so that the model outputs a different label, which is predefined. Its functionality is explained through Equation 4.

$$ASI=CSI-\varepsilon \text{ sign } (\Delta CSI\ J\ (\theta, CSI, label)) \tag{4}$$

Untargeted attack: This attack modifies the input so that the model outputs a different label, which is not predefined. Equation 5 describes the operation of an untargeted attack.

$$ASI=CSI+\varepsilon \text{ sign } (\Delta CSI\ J\ (\theta, CSI, label)) \tag{5}$$

In both above equations, AdversarialSculptureImage (ASI) is the perturbated image, while CleanSculptureImage (CSI) is the original, clean image expressed in three dimensions (width x length x depth). The label is the intended class which is the output value. The noise level is indicated by ε, and J estimates the cross-entropy loss, a function of model parameters (θ), CSI and the label. The method of attack through FGSM for the proposed work is shown in Figure 6.

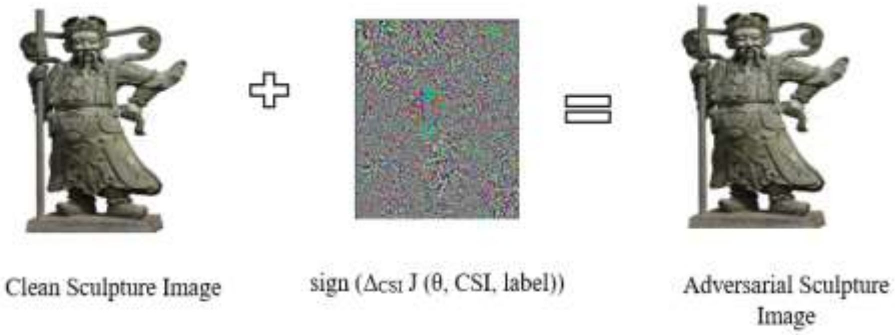


Figure 6. FGSM Method of Generating Adversarial Image

3.3 Preventing the Adversarial attack on sculpture images using Simplified Graph Convolution Networks

The attack is induced in the proposed work through the FGSM method, which is effectively handled by Simplified Graph Convolutional Networks (SGCN) [41], a variant of applying CNN in the perspective of graphs. The SGCN avoids adversarial attacks by stacking multiple layers of first-order spectral filters, followed by an utterly non-linear activation function. The working of SGCN is shown in Figure 7.

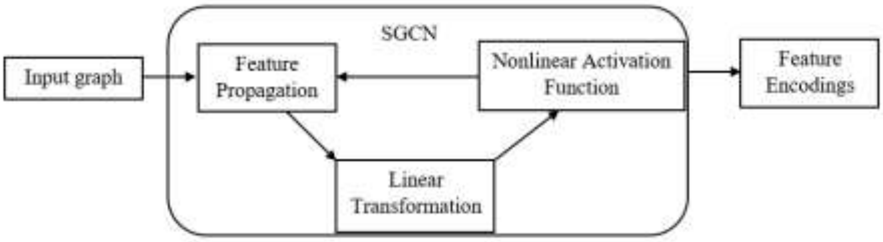


Figure 7. Generation of FE through SGCN

The proposed SGCN learns the representations of any feature xi in multiple layers (k). At any kth convolution layer, the input is represented as Hk-1 while the output is Hk. The input features for the first convolution layer are given in Equation 6.

$$H'(0)= X = [x1, x2, \ldots,xn]T \tag{6}$$

Feature Propagation is done at the beginning of every layer, where the mean of the feature (hi) of every node (vi) along with its neighbour nodes are found according to equation 7.

$$h_i^k = \frac{1}{d_i + 1} h_h^{k-1} + \sum_{j=1}^{n} \frac{I_{ij}}{\sqrt{(d_i + 1)(d_j + 1)}} h_j^{k-1} \tag{7}$$

This smoothens the representations along the graph edges and propagates similar predictions to all the locally connected nodes within the distance (d). The weight matric associated with each layer represented as θk is used for the linear transformation of the features. Then a non-linear transformation function like ReLU is applied, as mentioned in equation (8).

$$Hk \leftarrow ReLU\ (H'k\ \theta k) \tag{8}$$

The FE is obtained by Equation 8.

$$FE=SHk\text{-}1\theta k \tag{8}$$

The value of S is determined by Equation 9, which is the normalised adjacency matrix among the nodes with added self-loops to form the graphs.

$$S=D\text{-}0.5\ A'D\text{-}0.5 \tag{9}$$

The value of A' is computed by A'=A+I, where A denotes the adjacency matrix of weights between the edges vi to vj. D is the degree matrix computed as D=diag(d1, d2, …, dn) and each d value is the row-wise segregation of aij.

4. Results and Discussion

This section discusses the empirical analysis of the effectiveness of the proposed methodology. The experimental settings are elaborated, followed by the results.

4.1 Training the Model

The Chinese sculpture images are minimal and hard to obtain, so the model is trained on the Imagenet data [42]. The training of the proposed method is done with 289 244 adversarial examples generated through the FGSM method. The model is trained with 10, 20, 30, 40, and 50 epochs in targeted and untargeted attacks. The modified images are generated by varying the ε values. The adversarial images are generated with ε=0.01, ε=0.02, ε=0.03, ε=0.04 and ε=0.05. So a total of five tests were performed, and each test randomly selected the input.

Table 3 provides the training results in the Imagenet dataset with special mention of intra-collection misclassification rates. The adversarial attack using the FGSM method was imposed on the S-CNN model augmented with SGCN. The misclassification is a case of the defrauding systems should have high misclassification rates, as the adversarial example should not be labelled in the class in which the attacker has intended to place it. The training results show that the S-CNN model is 76.28% effective in combating adversarial attacks, while the S-CNN integrated with SGCN is 82.28% effective.

Table 3. Training Results of Adversarial Attack with and without SGCN Model

Name of the Image	**Number of Original images**	**Number of Images after Data Augmentation**	**Number of adversarial images**	**Misclassification rate in S-CNN**	**Misclassification rate in S-CNN augmented with SGCN**
Ma-zu	**324**	**2660**	**3023**	**73.98%**	**84.13%**
Sam-Tai-Zu	**126**	**1038**	**2835**	**75.83%**	**83.19%**
Earth God	**340**	**1968**	**2749**	**78.36%**	**81.74%**
Guan Yu	3116	4220	75.72%	87.42%	3116
Guanyin Bodhisattva	244	2004	2652	73.83%	85.28%
Average	***	***	***	75.54%	84.35%

4.2 Testing the Model with Chinese Sculpture Images

The sculpture images are very scanty, so training the proposed model uses the ImageNet dataset. Transfer Learning strategy is applied here, as the images of Chinese sculpture images are very limited in number. Training the model with insufficient or scanty data will not improve its efficacy. The transfer learning method used in this work is described in Figure 8.

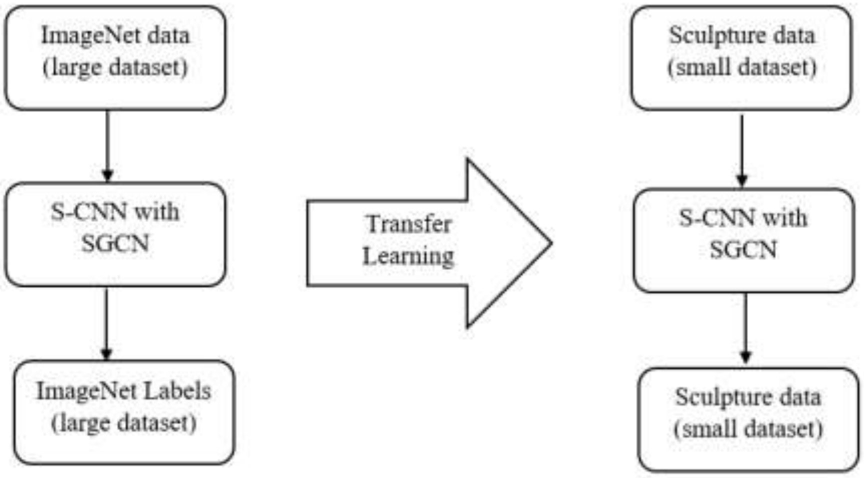

ImageNet data (large dataset)
S-CNN with SGCN
ImageNet Labels (large dataset)
Transfer Learning
Sculpture data (small dataset)
S-CNN with SGCN
Sculpture data (small dataset)

Figure 8. Transfer Learning to Handle Adversarial Attacks in Sculpture Images

The dataset for testing is extracted from the Traditional Chinese God Statue dataset [43], which comprises 11970 images after proper image augmentation. The Chinese god statue images were taken from famous Chinese temples and even from Google search engines. The images are in their raw digital format. The images are either partial or full-bodied. The low-quality images and cluttered backgrounds are cropped for perfection, while the noisy images are discarded. The images considered in this study are listed in Table 4. This dataset comprises five images extracted from Google: Ma-zu, Earth God, Sam-Tai-Zu, Guanyin Bodhisattva and Guan Yu. The images are pre-processed by weeding out undersized and blurry sculpture images. The image augmentation is done by rotating them by 300, 600, 900, 1200, 1500, and 1800.

Table 4. Testing Results of Adversarial Attack with and without SGCN Model Using Chinese Sculpture Images

Name of the Image	**Number of Original images**	**Number of Images after Data Augmentation**	**Number of adversarial images**	**Misclassification rate in S-CNN**	**Misclassification rate in S-CNN augmented with SGCN**
Ma-zu	324	2660	3023	73.98%	84.13%
Sam-Tai-Zu	126	1038	2835	75.83%	83.19%
Earth God	340	1968	2749	78.36%	81.74%
Guan Yu	380	3116	4220	75.72%	87.42%
Guanyin Bodhisattva	244	2004	2652	73.83%	85.28%
Average	***	***	***	75.54%	84.35%

The results indicate that preventing an adversarial attack on the proposed S-CNN with SGCN is much more effective than the vanilla S-CNN. The proposed model is 8.81% more accurate in misclassifying the negative image perturbated by the FGSM method. It can be inferred from the study that both in the training phase using Imagenet data and in the testing phase using the Chinese God images; the proposed model was able to show better performance which is shown graphically in Fig 9. The training for the model is given using a massively big Imagenet dataset, which comprises different types of image classes that effectively learns the images from varied backgrounds and quality levels. The number of images in the Chinese God dataset is relatively low, so it is tough to train the model using this dataset. Hence transfer learning has contributed significantly to enhancing the efficacy of the model.

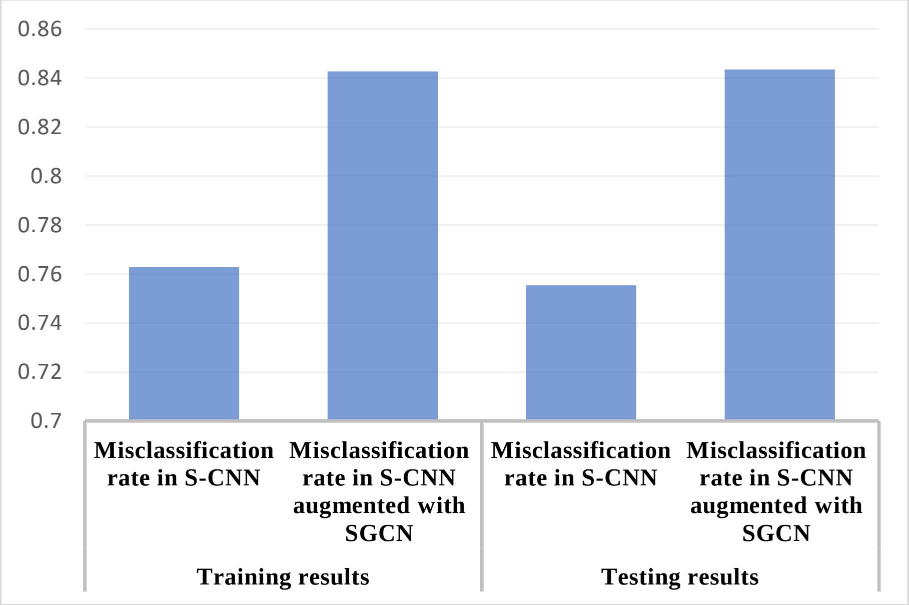


Figure 9. Performance Comparison of the Model in the Training and Testing Phase

This work is very robust and versatile and could be adopted to prevent adversarial attacks on any type of image. However, the dataset considered in this work primarily focuses on the Chinese Deity images and images of modern ceramic sculpture works as they have good market value. But this can be extended to landscaping images, which could be a future enhancement of the work.

5. Conclusion

Adversarial attack detection and prevention is now a widely researched topic in information security. This comprehensive work discusses the S-CNN defrauding model for Chinese sculpture images, which uses Triplet loss as the objective function to train the model. This defrauding system is attacked with adversarial images creating image perturbations using the FGSM method with various noise values. The image perturbation considered in this work focuses on targeted and non-targeted attacks. This adversarial attack is combated by deploying a new SGCN, which uses graph functions to learn the correct parameters and misclassify the negative image. The model's training is done using perturbations created on the Imagenet dataset. The model is tested against the adversarial images induced in the Chinese God images dataset by applying a transfer learning strategy. The results indicate that the proposed model has a higher misclassification rate of adversarial images, thus making the model qualify as a better defrauding system that could effectively combat the adversarial attack. The work can be extended to detect the presence of adversarial images in real-time scenarios in different applications. Also, other methods of generating negative images can be experimented with to validate the efficacy of the proposed model.

References

[1] Sharanya, S., Venkataraman, and Murali, G. "Estimation of remaining useful life of bearings using reduced affinity propagated clustering." J. Eng. *Sci. Technol*, vol.16, 2021, pp.3737-3756.

[2] Janiesch, C., Zschech, P. and Heinrich, K. "Machine learning and deep learning." *Electronic Markets*, vol.31, no.3, 2021, pp.685-695.

[3] Yuan, X., He, P., Zhu, Q. and Li, X. "Adversarial examples: Attacks and defences for deep learning. IEEE Transactions on neural networks and learning systems," vol.30, no.9, 2019, pp.2805-2824.

[4] Yuan, X., He, P., Zhu, Q. and Li, X. "Adversarial examples: Attacks and defences for deep learning." *IEEE Transactions on neural networks and learning systems*, vol.30, no.9, 2019, pp.2805-2824.

[5] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D. "Robust physical-world attacks on deep learning visual classification." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp.1625-1634.

[6] Samonas, S. and Coss, D. "The CIA strikes back: Redefining confidentiality, integrity and availability in security." *Journal of Information System Security*, 2014, vol.10, no.3.

[7] Harahap, S.H., Sunendar, D. and Damayanti, V.S. "Requirements Analysis: Drama Education in High School." *Educational Administration: Theory and Practice*, vol.28, no.2, 2022, pp.66-73.

[8] Nadelman, C. "Pot People Recent Figurative Ceramic Sculpture." *Sculpture Review*, vol.59, no.3, 2010, pp.18-23.

[9] Yang, G.M. and Suchan, T. "The cultural revolution and Contemporary Chinese Art." *Art Education*, vol.62, no.6, 2009, pp.25-32.

[10] Kharchenkova, S., "The market metaphors: Making sense of the emerging market for contemporary art in China." *Poetics*, vol.71, 2018, pp.71-82.

[11] Chen, Y. and Zhou, L. "Reflections on the Development of Contemporary Ceramic Aesthetics." In 7th International Conference on Education, Management, Information and Mechanical Engineering (EMIM 2017), Beijing, China: Atlantis Press, pp. 1899-1902, April, 2017.

[12] Haque, M. A. "A comparative study of contemporary ceramic sculptures between China and Bangladesh." *International Journal of Visual and Performing Arts*, vol.2, no.1, 2020, pp.42-58.

[13]Akhtar, N. and Mian, A. “The threat of adversarial attacks on deep learning in computer vision: A survey.” *IEEE Access*, vol.6, 2018, pp.14410-14430.

[14]Bai, T., Luo, J., Zhao, J., Wen, B. and Wang, Q., 2021. “Recent advances in adversarial training for adversarial robustness.” arXiv preprint arXiv:2102.01356.

[15]Soares, E. *Radnn: Robust to imperceptible adversarial attacks deep neural network*, 2021

[16]Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X. and Zhu, J. “Defence 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.

[17]Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D. and McDaniel, P., 2017. Ensemble adversarial training: Attacks and defences. arXiv preprint arXiv:1705.07204.

[18]Huang, S., Papernot, N., Goodfellow, I., Duan, Y. and Abbeel, P., 2017. “Adversarial attacks on neural network policies.” arXiv preprint arXiv:1702.02284.

[19]Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X. and Li, J., “Boosting adversarial attacks with momentum.” In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9185-9193.

[20]Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L. and Kohane, I. S. “Adversarial attacks on medical machine learning.” *Science*, vol.363, no.6433, 2019, pp.1287-1289.

[21]Xu, H., Ma, Y., Liu, H.C., Deb, D., Liu, H., Tang, J.L. and Jain, A.K. “Adversarial attacks and defences in images, graphs and text: A review.” *International Journal of Automation and Computing*, vol.17, 2020, pp.151-178.

[22]Gopalakrishnan, S., Marzi, Z., Madhow, U. and Pedarsani, R., 2018. “Combating adversarial attacks using sparse representations.” arXiv preprint arXiv:1803.03880.

[23]Pruthi, D., Dhingra, B. and Lipton, Z.C., 2019. “Combating adversarial misspellings with robust word recognition.” arXiv preprint arXiv:1905.11268.

[24]Taran, O., Rezaeifar, S., Holotyak, T. and Voloshynovskiy, S. “Machine learning through cryptographic glasses: combating adversarial attacks by key-based diversified aggregation.” *EURASIP journal on information security*, 2020, pp.1-18.

[25]Keller, Y., Mackensen, J. and Eger, S., “BERT-defence: A probabilistic model based on BERT to combat cognitively inspired orthographic adversarial attacks,” *arXiv preprint*, 2021 arXiv:2106.01452.

[26]Jones, E., Jia, R., Raghunathan, A. and Liang, P., 2020. Robust encodings: A framework for combating adversarial typos. arXiv preprint arXiv:2005.01229.

[27]Ji, Y., Bowman, B. and Huang, H.H., “Securing malware cognitive systems against adversarial attacks.” In 2019 IEEE international conference on cognitive computing (ICCC). *IEEE*, July, 2019, pp.1-9.

[28]Huang, H., Wang, Y., Chen, Z., Zhang, Y., Li, Y., Tang, Z., Chu, W., Chen, J., Lin, W. and Ma, KK, 2022, June. Cmua-watermark: A cross-model universal adversarial watermark for combating deep fakes. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, 2022, pp. 989-997.

[29]Jorgensen, Z., Zhou, Y. and Inge, M., 2008. “A Multiple Instance Learning Strategy for Combating Good Word Attacks on Spam Filters.” *Journal of Machine Learning Research*, vol. 9, no. 6.

[30]Zhou, Y., Jorgensen, Z. and Inge, M., 2007, October. “Combating good word attacks on statistical spam filters with multiple instance learning.” In 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), *IEEE*, vol. 2, pp. 298-305.

[31]Ke, J., Wang, L., Ye, A. and Fu, J., 2022, July. Combating Multi-level Adversarial Text with Pruning-based Adversarial Training. In 2022 International Joint Conference on Neural Networks (IJCNN), *IEEE*, pp.1-8.

[32]Sahay, R., Mahfuz, R. and Gamal, A.E., 2019. “A computationally efficient method for defending adversarial deep learning attacks.” arXiv preprint arXiv:1906.05599.

[33]Ke, J., Wang, L., Ye, A. and Fu, J., “Combating Multi-level Adversarial Text with Pruning-based Adversarial Training.” In 2022 International Joint Conference on Neural Networks (IJCNN). *IEEE*, 2022, pp.1-8.

[34]Song, S., Chen, Y., Cheung, N. M. and Kuo, C. C. J., “Defence against adversarial attacks with Saak transform,” *arXiv preprint*, 2018, arXiv:1808.01785.

[35]Alfarra, M., Pérez, J.C., Thabet, A., Bibi, A., Torr, P.H. and Ghanem, B., . “Combating adversaries with anti-adversaries.” In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, Jun. 2022, pp. 5992-6000.

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

[37]Peng, J. and Chan, P. P, “Revised Naive Bayes classifier for combating the focus attack in spam filtering.” In 2013 International Conference on Machine Learning and Cybernetics. *IEEE*, vol. 2, Jul. 2013, pp. 610-614.

[38]Liu, Z., Wang, F., Lin, Z., Wang, L. and Yin, Z., 2020, December. De-co: A two-step spelling correction model for combating adversarial typos. In 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom). *IEEE*, pp. 554-561.

[39]Shi, D., Orouskhani, M. and Orouskhani, Y. A conditional Triplet loss for few-shot learning and its application to image co-segmentation. *Neural Networks*, vol. 137, 2021, pp. 54-62.

[40]Yuan, X., He, P., Zhu, Q. and Li, X. “Adversarial examples: Attacks and defences for deep learning.” *IEEE Transactions on neural networks and learning systems*, vol. 30, no. 9, 2019, pp.2805-2824.

[41]Yusuf, A.A., Chong, F. and Xianling, M., “An analysis of graph convolutional networks and recent datasets for visual question answering.” *Artificial Intelligence Review*, vol.55, no. 8, 2022, pp. 6277-6300. Avaliable: https://www.image-net.org/download.php

[42]Huang, Mei-Ling; Shiau, Kai-Ling; Tseng, Yu-Lun; Liao, Yu-Chieh, “Dataset of traditional Chinese god statue,” *Mendeley Data*, vol.2, 2022, Avaliable: doi: 10.17632/z6t86sjwts.2