International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Disposed Watermarking to Attack Deep Neural Network

Ms. Gauri A. Bhosale, Mr. Devidas S. Thosar and Mr. Sharad M. Rokade

Department of Computer Engineering, SVIT, Chincholi, Nashik

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

I have proposed in this paper it show that the generated adversarial samples are not only capable of fooling the Inception of theV3 model with high success rates, but also shared to the other models, such as the Recognition developed by Amazon. In this paper, we proposed a visible adversarial attack approach utilizing watermarks, with two types of attack to simulate the real-world cases of watermarks and have successfully interfered the judgment from some state-of-the-art deep learning models. partial adversarial samples that shows great transferability onto other models with including the Recognition.In conclusion, we believe this work suggests that the robustness of current object recognition models are yet to be further improved, and more defense approaches shall be employed. We provide a comprehensive list of the requirements that the empowers of quantitative and qualitative assessment of an current and pending the DL watermarking approaches.

Keywords: *Adversarial attack, Watermark,OCR, DNN.*

Introduction

The watermark is mainly used approach for identification of the image source and avoid the replication or potential copyright infringements activities yet whether its introduction would affect the content and to which extent remains unanswered. In this work, we propose a visible adversarial attack method using watermarks as perturbations[6].

The input images, which originally being correctly classified, are is classified by the target model after being watermarked via our method, while the watermarks are constrained in either transparency or size. During the process of each attack, we iteratively adjust only 9 parameters to perform the transform of location, angle, transparency, color and size of the watermark while leaving the input image untouched[5].

Additionally, the transferability of adversarial samples on other models including Recognition for black box attack purposes. Our method present promising results in all experiments, and more importantly has identified the problem that inappropriate watermarks are able to perturb the judgment of state-of-the-art deep learning models. This detector can be an online platform that veri_es the authenticity of images or a media player that implements digital rights management.The attacker iteratively changes the signal by analyzing the binary responses until the watermark is not detected anymore. The necessary changes are minimized to preserve the signal. In the watermark the adversary aims at to recovering the watermark.

Similar to machine learning, watermarking methods are also used in an adversarial environment and thus should not only survive unintentional changes like compression, but also targeted attacks. We focus on the following two attack classes that correspond to black-box evasion and model-extraction attacks. In this way, she can embed or remove the watermark in a variety of new signals, thus determining the use case of achieving copyright protection. due to its black-box mechanism. For instance,an adversary may robustness and transferability of CNN have always been suspected the context of spam ltering, an adversary can try to evade detection by omitting words from spam emails indicative for unsolicited content. Depending on her knowledge, the adversary operates between a black box and white-box setting. In case of non blind watermarking both cover image and watermarked image are required during watermark extraction process[9].

Literature Survey

Y. Nagai et al.[1] specifies that disposal of watermarking to deep neural network In this paper,we extend this work and examine a novel black box attack against digital watermarking based on concepts from adversarial learning.In reality, watermark is one of the most frequently used approach for identification of the image source and revention of potential copyright infringement activities such as and, yet whether its introduction would affect the content and to which extent remains unanswered. In this work, we propose

a visible adversarial attack method using watermarks as perturbations.

E. L. Merrer et al.[3] The entire process of the proposed method can be divided into two parts: image recognition and adversarial watermarking,where the former is to feed the watermarked target image into a well-trained deep neural network T for image recognition,and the latter is to update the parameters to determine how the watermark is transformed and placed according to the recognition result. (RGB + alpha channel containing transparency information), is modified by the watermark transformation layer in terms of color, transparency, location, angle and size.

Kaiming He et al.[2] Express that the watermark extraction process extracts watermark from watermarked image and it is exactly reverse process as that of watermark embedding process. In case of non blind watermarking both cover image and watermarked image are required during watermark extraction process, while in blind image watermarking only watermarked image is required in watermark extraction process. The variety of attacks are applied such as noise addition, noise ltering, rotation, scaling, translation, Gamma correction, resizing, cropping,compression. these attacks are considered. The design specification describes the features of the system, the modules or elements of the system and their appearance to end-users. to evaluate the robustness of developed image watermarking techniques under proposed system. The proposed system also includes MEO based grey scale image watermarking techniques which is proposed for optimization of perceptual quality and robustness under high payload scenario. This optimization significantly reduces computation time with compared to existing optimization techniques under consideration.

Y. Adi et al.[5] During the process of each attack, we iteratively adjust only 9 parameters to perform the transform of location, angle, transparency,color and size of the watermark while leaving the input image untouched. Additionally, experiment the transferability of adversarial samples on other models including Recognition for black-box attack purposes. Our method present promising results in all experiments, and more importantly has identi_ed the problem that inappropriate watermarks are able to perturb the judgment of state-of-the-art deep learning models.

Y. Nagai et al.[1] In this paper, we propose a digital watermarking technology for ownership authorization of deep neural networks. First, we formulate a new problem.In addition, the Validation could be given by utilizing Message Authentication Code, and installing this code into the picture as well. The issue comes here is the computational cost and time many-sided quality of utilizing a vigorous cryptographic algorithm with some verification code algorithm.On the other hand the existing watermarking.

Y. Nagai et al.[1]This is a very innovative zone of research. The technique will have a significant on defense, business, copyright protection and other _elds where information needs to be protected at all costs from attacks.Limitation of existing work counts very less security and more attacks. For the most part watermarks are utilized where confirmation or possession is required.

Y.Chen et al[4] Watermarks utilized ought to be undetectable, as in, they are inserted into the picture in the wake of actualizing any cryptographic algorithm. Being the need of more security, the con_rmation can additionally be utilized. Those methods were guaranteeing that if there comes any change in the message, it will be gotten as, when the validation code ascertained with the ruined message, it won't match the particular case that is in picture. Assume the situation when interloper not has any desire to change the message, in disdain he simply needs to take the message, such as replicating watchword. At this point, the past procedure falls at,as the message is in plain content. Thus, a method ought to be proposed, which utilizes both message security, i.e., secrecy and message validation i.e., honesty.

The proposed digital image watermarking system includes multiple techniques those collectively full predefined research objectives. The system Design is defined as The method of implementing numerous principles and techniques for the significant process or a system to authorize of its physical realization,various design structures are there to develop the system.

Proposed Methodology

On the purpose of producing appropriate watermark for interfering the judgment of T, we perform three steps of transformation on W as illustrated in Fig. 1. Firstly, we perform rotation with parameter →ı, and each pixel originally at (x, y) is moved to We utilize the spatial transformation network (STN) [12], which employs bilinear interpolation to eliminate the problem that the result coordinates might not be integers, for actual implementation to ensure the module is differentiable. (x0,y0)where:

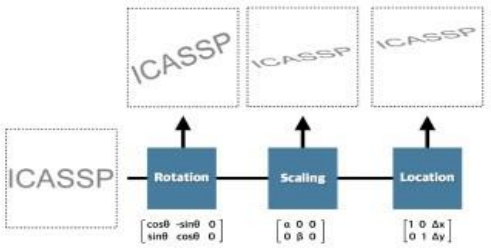


Fig. 1. Three steps of transformations of the watermark (see stacked transformation layer.

As described in, the parameters are adjusted in each iteration until an adversarial sample is produced or the number of iterations reaches limit.In addition to the transformation layers, the color/transparency adjusting layer and merging layer are also indispensable for our method[9].

In it is proposed that watermark takes input images, which originally being correctly classified,are misclassi_ed by the target model after being watermarked via our method, while the watermarks are constrained in either transparency or size.In it is proposed that watermark takes input images, which originally being correctly classified,are misclassi_ed by the target model after being watermarked via our method, while the watermarks are constrained in either transparency or size.

The digital watermarking it hides the copyright information into the digital data through that the certain algorithm. The secret information to be embedded and it can be some text,author serial number or unique number, company logo, images with some special importance. This secret information is embedded into the digital data like (images, audio, and video) to make sure the safety from the attacker,data authentication, identification of owner and copyright protection. The watermark are often hidden in digital data either visibly or invisibly.

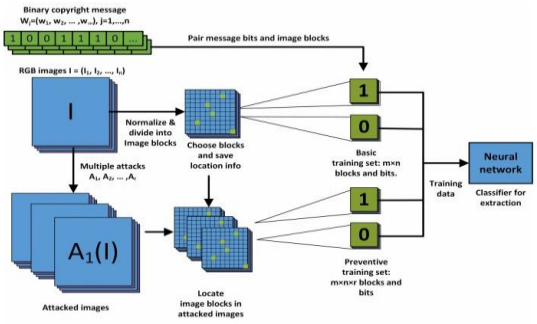


Fig. 2. Network training procedure.

The neural network in our scheme can be partly analogous to the hashing procedure in general digital signature systems, in the sense that the neural network serves as a mapping from images to discrete numeric values (digests). We then train a neural network to transform the image blocks into copyright message[8].

A Architecture

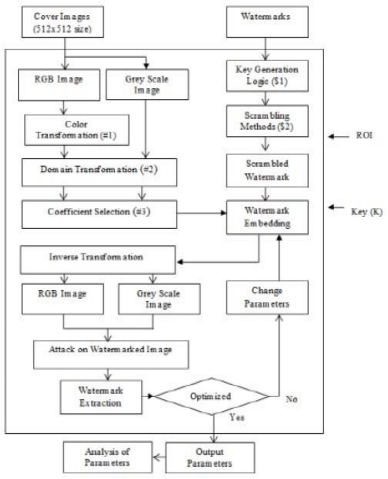


Fig. 3. Architecture of Extracting watermark and deep neural network

The watermark embedding process in proposed system includes two phases namely cover image processing phase and watermark processing phase. Both of the phases are shown in overall block diagram of proposed system. Cover Image Processing Phase handles cover images of size 512x512 including grey scale, color and medical images. If the cover is grey scale image, then it is directly taken into transform domain. If cover is color image, then color transformation like RGB to KLA, RGB to YUV, RGB to YIQ, RGB to XYZ, RGB to YCbCr, RGB to RcGcBc, RGB to RsGsBs, RGB to LUV is applied and then one or more color planes are taken into transform domain[1]. Watermark Processing Phase the grey scale watermark of sizes including 64x64, 128x128 and 256x256,512x512 are used as input to test proposed watermarking system. As part of watermark processing the key generation logic is applied. It includes methods like PN sequence generator, threshoding with sequence generator and Harris corner detection method to measure number of corners. The key generation logic generates key(K) which is further used in watermark scrambling. The watermark extraction process extracts watermark from watermarked image and it is exactly reverse process as that of watermark embedding process. In case of non blind watermarking both cover image and watermarked image are required during watermark extraction process, while in blind image watermarking only watermarked image is required in watermark extraction process. The variety of attacks are applied such as noise addition, noise ltering, rotation, scaling, translation, Gamma correction, resizing, cropping, compression. These attacks are considered to evaluate the robustness of developed image watermarking techniques under proposed system. The proposed system also includes MEO based grey scale image watermarking techniques which is proposed for optimization of perceptual quality and robustness under high payload scenario. This optimization signi_cantly reduces computation time with compared to existing optimization techniques under consideration. In this neural network prediction in the very last layer of a DLmodel needs to closely match the ground-truth data (e.g.,training labels in a classification task) in order to have the maximum possible accuracy. As such, instead of directly regularizing the activation set of the output layer, we choose to adjust the tails of the decision boundaries to incorporate a desired statistical bias in the network as an 1-bit of watermark.We focus on an, in particular, classification of tasks using an deep neural networks.

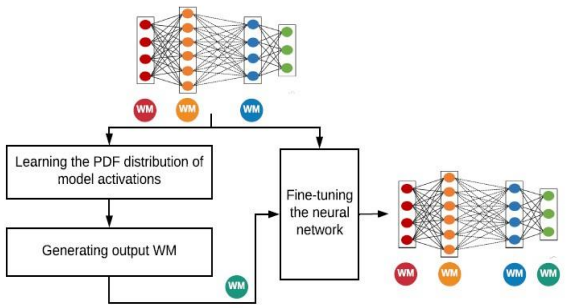


Fig. 4. watermarking the output layer.

Next phase is output of watermarking is a postrocessing step performed after embedding the chosen binary WMs within the intermediate (hidden) layers. Then we examine the the attacker is conscious of the watermarking technique, he may plan to damage the first watermark by embedding a replacement WM within the DL model. In practice, the attacker doesn't have any knowledge about the situation of the watermarked layers. However, in our experiments, we consider the worst-case scenario during which the attacker knows where the WM is embedded but doesn't know the first watermark information. To perform the overwriting attack, the attacker follows the protocol. a replacement set of watermark information (using a special projection matrix, binary vector, and input keys) summarizes the results of watermark overwriting for all three benchmarks within the black-box setting.As shown, DeepSigns is strong against the overwriting attack and may successfully detect the first embedded WM within the overwritten model.

B. Mathematical Model

Mathematical Model may be a system description that uses mathematical concepts and language. the tactic of making a mathematical model is named mathematical modeling. Let the device or the program be S.

S={I,P,O} where
I=Input
P= Process
O= Output

i = (i.a, i.b, • • • , I.n)
i1= series of RGB images;

Wj = (w1,w2, · · · ,wm),wi ∈{0, 1}, j = 1, 2, · · · , n;
W= binary sequences of copyright message for each image
A= expanded training data is yield from an attack set
A(Ik) = {A1(Ik), A2(Ik), · · · , Ar(Ik)}.;

P={S1, EF,S2,F,PI,NR}
Where,
S1= segmentation S1 mask;
S2= S2 mask;
F= Fusion the image and make it available for processing
PI= Process the image
NR= Remove noise

O={seg S}

C. Algorithm

The proposed algorithm includes three steps: that are the decomposing the cover image, an embedding the image, and extraction of that image.
Step 1: A grayscale cover image with pixel dimensions is first selected.
Step 2: A grayscale watermark image with pixel dimensions of 64 is used then converted into binary.
Step 3: The watermark binary image with pixel dimensions of 64 is split into small, non-overlapping blocks with pixel dimensions of 4 2, thus producing 16 32 blocks. These blocks are then embedded into the chosen wavelet coefficient blocks. However, the watermark is embedded block by block, instead of being converted into a vector, which facilitates embedding. This process includes a discount within the loop iteration and time interval . This step also enables simple control and follows the flow of the embedded data.
Step 4: The original coefficient of the chosen block, I0 (i,j) is that the watermarked coefficient like I(i,j), w is that the watermark bit, and a is that the embedding strength coefficient that controls the watermarking strength. the worth of a directly influences embedding effectiveness and is chosen experimentally.To ensure high watermarking quality, the watermark blocks (4096 pixels) are embedded within the three wavelet coefficients (cH3, cD3, and cV3) sequentially, as follows: 1. the primary 25% (1024 pixels) of the watermark pixel values are embedded into cH3. 2. The second 25% means the (1024 pixels) of the watermark image pixel values are embedded . 3. The remaining 50% of (2048 pixels) of the watermark pixel values are embedded into cV3.
Step 5: Inverse decomposition wavelet transform is performed on each coefficient to get the watermarked image. robustness property without the requiring prior knowledge of an possible distortions on the marked image. The proposed system constructs an unsupervised deep neural network structure with a completely unique loss computation for automated image watermarking. a clear adversarial attack approach utilizing watermarks, with two sorts of attack to simulate the real-world cases of watermarks and have successfully interfered the judgment from some state-of-theart deep learning models. Experimental results along side a
challenging application of watermark extraction from camera resampled marked-images have confirmed the prevalence performance of the proposed system.

Results

To check the expected output of the proposed system I've implemented in python with the operating system window.Initially, we give the normal input as image to the system as shown in fig 5.

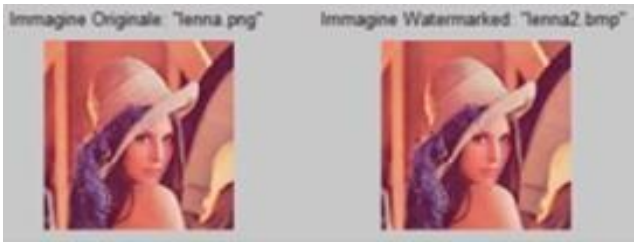


Fig. 5. Input Image

The admin panel has a authority to add users and also approved or disapproved the user.

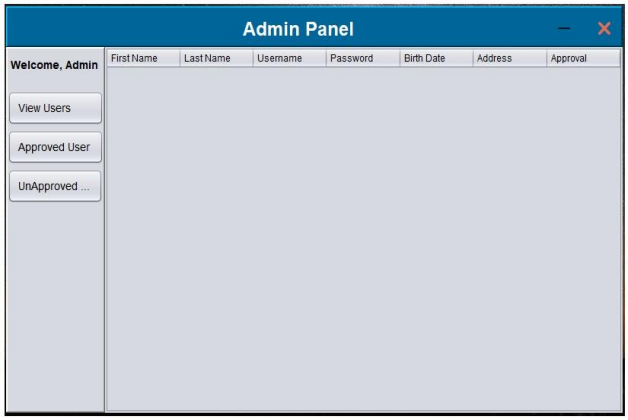
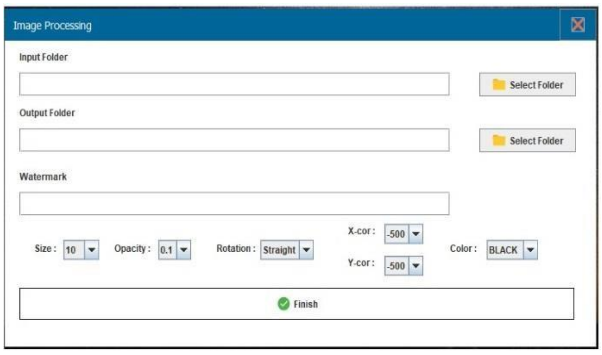


Fig.6. Admin Panel

The Users are get approved by the admin after creating the account or register, Then and then only the user have the autority to approved it.After approving the request the user can select the input as a image for adding the watermarkto it.



Acknowledgment

My sincere thanks go to SVIT College of Engineering for providing a strong forum to improve my skills and capabilities. I would like to thank all those who are helping us, directly or indirectly, to present the paper. I take this opportunity to express my heartfelt gratitude to the people whose support is very helpful in completing our project. I would really like to express my sincere thanks to my guide Prof. S.M. Rokade whose experienced guidance has become very important to me.

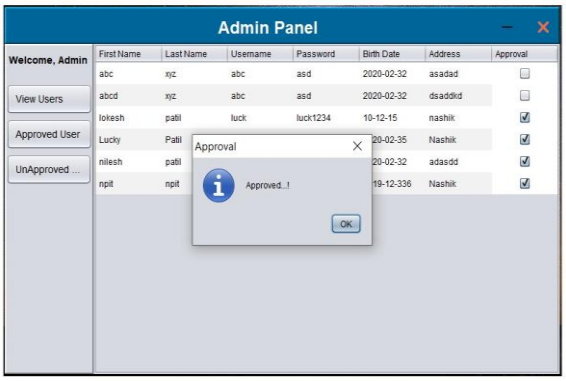


Fig.7. Approved User

The user can be disapproved by the admin if admin don't want to carry it then the admin take back his rights from the user.

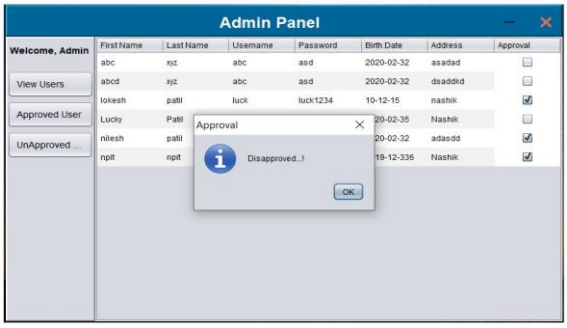


Fig.8. Disapproved the user.

After approving the user by admin.The user can choose the image from particular folder to add watermark on it. After all the processing is done it takes the input as image from system and inserted as a watermark.By Passing the parameters like Size, Opacity,Rotation, and Color.

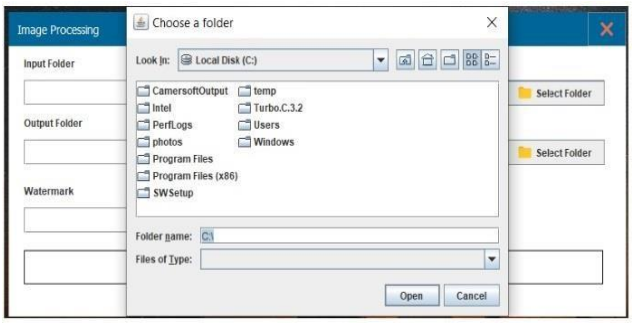


Fig. 10. Select Image to Insert Watermark.

Conclusion

In this paper, we propose an automated image watermarking system using deep convolutional neural networks. The variety of attacks are applied such as noise addition, noise _ltering, rotation, scaling, translation, Gamma correction, resizing, cropping, compression. These attacks are considered to evaluate the robustness of developed image watermarking techniques under proposed system.The admin has an authority to add the user and disapproved it if the he found any kind of misbehave then the user will be block by the admin.After approving the admin then user can add watermark on selected image to avoid the copyrights.The proposed blind image watermarking system achieves its.By exploring the ability of deep neural networks in the task of fusion between the cover-image and the latent spaces of the watermark, the proposed system has successfully developed an image fusion application on image watermarking.

References

[1]. Y. Nagai, Y. Uchida, S. Sakazawa, and S. Satoh, Digital watermarking for deep neural networks, International Journal of Multimedia Information Retriev, vol. 7,no. 1, 2018.

[2]. Xiangu Zhang, Shaoqing Ren, and Jian Sun, Deep residual learning for image recognition,2016.

[3]. "Embedding Watermarks into Deep Neural

[4]. Networks",https://arxiv.org/abs/1701.04082

[5]. E. L. Merrer, P. Perez, and G. Adversarial fo neural network watermarking, 2017.Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet, Turning your weakness into a strength: Watermarking deep neural networks bybackdooring, Security Symposium 2018.

[6]. K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, On the detection of an adversarial 2017.

[7]. Y.Chen, C. Qiao and X. Yu, Convolutional neural networks for medical image analysis: Full training or _ne tuning? IEEE transactions on medical imaging,vol. 35, no. 5, 2018.

[8]. B. D. Rouhani, M. Samragh, T. Javidi, and F. Koushanfar,Safe machine learning and defeat-ing adversarial attacks, IEEE Security and Privacy March 2018.

[9]. Laurens van der Maaten Gao Huang, Zhuang Liu and Kilian Weinberger,Densely connected convolutional networks, in CVPR, 2017.\Towards e_cient multi-keyword fuzzy search over encrypted outsourced data with accuracy improvement", IEEE Trans. Inf. Forens. Security, vol. 11, no. 12, Dec. 2016.

[10]. Ja-Ling Wu Yu-Hsun Lin, Unseen visible watermarking for color plus depth map 3d images,A Secure and High-Capacity Data-Hiding method Using Compression, Encryption and Optimized Pixel Value Di_erencing, IEEE Access,Vol. 6,October 2018.

[11]. Jianting Ning, Zhenfu Cao,Xiaolei Dong,Kaitai Liang, Hui Ma and Lifei Wei,Watermarking neural network, IEEE Transactions On Information Forensics And Security, Vol. 13, No., 1, January 2016.

[12]. V. Goyal, O. Pandey, A. Sahai, and B. Waters, Dynamic neural network in hospital management system, Vol. 29, 2006.