Review of visible watermark’s influence on private information-related machine learning techniques

Yizhe Lyu
Shenyuan Honors College, Beihang University, Beijing, 100191, China
Corresponding author’s e-mail: 16231249@buaa.edu.cn

Abstract. Under the increasing needs of sharing datasets and risks of leaking privacy during data sharing, there is growing attention of the owners of datasets to claim ownership and protect their privacy from being misused. To achieve this secure purpose, the visible watermark may be an effective way. However, due to the loss of visual information, it is still unclear the influence of these visible watermark methods on related machine learning techniques. To investigate these, this paper provides a review of visible watermark’s influence on private information-related machine learning techniques. This paper compared the performance of different CNN methods with or without visible watermarking. According to the results of the experiments, the accuracy of watermarked images’ identification decreased significantly. By average, the accuracy would decrease by 4.75%. And the performance changed when using different CNN structures, watermarked methods and datasets. After that, there are some experiments to explore the feasibility of using none-watermarked images trained network to identify the watermarked images. The results show that the accuracy of identification at least decreased by 25% when only one of the train set and test set are watermarked.

1. Introduction
Machine learning is a branch of artificial intelligence[1]. It learns the knowledge and laws from the sample data and then uses them for actual inference and decision making. A significant difference between machine learning and the normal program is the need for sample data, machine learning is a data-driven approach. After decades of development, machine learning has gradually become the cornerstone of solving the problem of artificial intelligence. After going through classical algorithms such as classification and regression tree[2], backpropagation[3], SVM[4]. In 2012, neural networks triumphed in the competition of SVM, it showed obvious advantages in challenging problems such as image and speech recognition.

It is not an accident that history has chosen neural networks as a solution to complex AI problems such as image, voice recognition. The neural network theoretically has the guarantee of universal approximation, as long as the number of neurons is sufficient, activate function satisfies certain mathematical properties, and a feedforward neural network containing at least one hidden layer can approximate any continuous function on a closed interval to any specified precision. So neural networks can simulate arbitrarily complex functions. The image and voice data we want to identify can be regarded as a vector or matrix, and the recognition algorithm is a mapping function of these data to the category value[5].

At present, machine learning’s researches have entered the period of deep learning, and deep learning has made rapid progress in various fields. In the field of face recognition[6], deep learning has achieved incredible success. Taking the LFW dataset as an example, the face++ network achieved an accuracy of
0.9950, Baidu achieved an accuracy of 0.9977, and FaceNet also achieved 0.9963[7]. In terms of speech recognition, deep learning has also made great progress[8]. In 2009, Geoffrey Hinton and Deng Li used DNN for acoustic model to replace GMM, and found that pretraining is unnecessary when training data is sufficient. After using DNN, the word error rate of speech recognition is relatively reduced by 30%.

Despite the achievements of machine learning, there are also numbers of problems. One of the most is the privacy of machine learning datasets especially face recognition and fingerprint recognition datasets. Figure 1 is an example of LFW dataset which is a large, real-world face dataset. Figure 2 is an example of NIST fingerprint dataset. They all contain a lot of privacy information. Take the face recognition datasets as an example, when these datasets are stolen by the unscrupulous, they may use these faces to do something illegal. First of all, they can use these faces to pretend as someone else on the Internet. In addition, 3D model methods can drive a single image to shake eyes and open mouth which are used for face-spoofing detection by most software.

Figure 1. An Example From LFW Dataset.  Figure 2. An Example From NIST Fingerprint Images.

The existing image watermarking scheme is mostly an invisible digital watermark, which is an information hiding technology based on content and non-cryptographic mechanism. It embeds some identification information into the digital carrier. It can be mainly divided into Spatial Domain Techniques and Frequency Domain Techniques.

In Spatial Domain Techniques, the watermark is inserted in the cover image changing pixels or image characteristics[9]. Mahfuzur Rahman and Koichi Harada proposed a method to insert information in objects with layered 3D triangular meshes such as those reconstructed from CT or MI data, a parity enhanced topology based spot area watermarking method[10]. Xiangui Kang proposed the data extraction process as one associated with a generalized channel of additive noise with a generally non-zero mean and fading by adaptively estimating the decision zone exploiting a training sequence and estimating the quantization step size using the Fourier analysis method[11]. C. Lu, H. Yuan and M. Liao presented a multipurpose watermarking scheme that can be applied to attain both authentication and protection of multimedia data[12].

In Frequency Domain Techniques, the target is to insert the watermarks in the spectral coefficients of the image. The most commonly used transforms are the Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), and Discrete Wavelet Transform (DWT). Discrete Cosine Transform is like as Discrete Fourier Transform[13]. It is a technique for converting a signal into elementary frequency components. The technique proposed by R. Mehul and R. Priti provide the watermark is inserted in four different frequency ranges by selecting coefficients in zigzag order[14]. The technique proposed by Y. Yang, X. Sun, H. Yang, and C.T. Li present a DCT domain based removable visible watermarking algorithm that moderately succeeds in defeating illegal removal and resisting compression[15]. Discrete Wavelet Transform is a mathematical tool for hierarchically decomposing an image. Digital watermarking in wavelet domain presented by Taskovski They implemented two watermarks using binary marks in LL2 and HH2 respectively, resulting in a mark which is robust against manipulations.
like compression and weak against cropping and rescaling. G. Hai-ying presented a watermark adapted to JPEG2000 using two algorithms to modify the wavelet coefficients of the LH2 band of the cover image[16], introducing only minimal differences between the watermarked image and the original.

Invisible digital watermarking can hide a large amount of information and has good concealment[17], but it is not robust to common image preprocessing techniques[18], and users must use corresponding watermark detection and extraction techniques to confirm copyright. Visible watermarks do not have a watermark extraction or detection process, which provides an intuitive copyright notice that can be better applied to copyright protection.

However, the visible watermark will destroy the visual information of the image. This is also the reason why the visible watermark is not widely used in machine learning. And almost no paper analyses the impact of visible watermarks on machine learning models. Take the face datasets Olivettifaces for example. This paper compares the watermarked dataset that is added watermark by OpenCV with original dataset, the rate of face recognition accuracy decreases by 7%. After that, add watermark to training set, but use the original test set, the recognition accuracy decreases by 20%; When the watermark is added to the test set, but the training set is original, the recognition accuracy is 16%. When mix the watermark image with the original, the training set and the test set are randomly selected, the face recognition accuracy is reduced by 12%.

2. METHODOLOGY

2.1. Convolutional neural network
Convolutional neural network is a kind of feedforward neural network with convolutional computation and deep structure. It has become a hotspot in the field of image recognition and is one of the representative algorithms of deep learning. It can reduce the complexity of the network structure and reduce the amount of weight. And this network structure is not deformed for translation, scaling, tilting or height.

Convolutional Neural Networks (CNN) mainly include the following: input layer, convolution layer, pooling layer, fully connected layer and output layer.

![Figure 3. Olivettifaces Dataset's Recognition CNN Network.](image)

2.1.1. Convolution layer
The function of the convolutional layer is to perform feature extraction on the input data, which contains multiple convolution kernels, and each element constituting the convolution kernel corresponds to a weight coefficient and a deviation amount. The size of the convolution kernel is usually smaller than the size of the input image, so the convolution will focus more on the extracted image.

2.1.2. Pooling layer
The pooling layer is also referred to as the down sampling layer. The pooling layer can reduce the size of features and reduce the number of parameters. It also has translation invariance. There are usually two types of pooling operations: max-pooling and mean-pooling.
2.1.3. Fully connected layer
Each node of the fully connected layer is connected to the node of the previous layer of all nodes. It is used to integrate features extracted from the previous layer. This output layer is a highly purified feature that is easy to judge using a classifier. The fully connected layer is located at the last part of the hidden layer of the convolutional neural network and only passes signals to other fully connected layers. The feature map loses the spatial topology in the fully connected layer, is expanded into a vector and passes the excitation function.

2.1.4. Output layer
The output layer of the convolutional neural network is the classifier layer. We use the softmax regression classifier as the output layer. The output layer uses a logic function or a normalized exponential function to output the classification label.

2.2. Datasets introduction
Information based on facial feature points is stored digitally, and databases are at risk of being compromised by hackers. In the past, passwords have been stolen and can be changed by resetting passwords. But biological information like faces is unique, along with lifelong features. So, once divulge, will incur people's individual wealth or privacy divulge, cause significant loss. Face datasets often require privacy and copyright protection, take two commonly used datasets: Olivettifaces and LFW for testing.

2.2.1. Olivettifaces Dataset
This dataset contains a set of face images taken between April 1992 and April 1994 at AT&T Laboratories Cambridge. Each of the 40 different subjects has ten different images. For some subjects, images were taken at different times, changing lighting, facial expressions and facial details. All images were taken in a dark, uniform background with the subject in an upright frontal position. The image is quantized to 256 gray levels and stored as an unsigned 8-bit integer. The "target" of the database is an integer between 0 and 39 representing the identity of the person depicted; however, there are only 10 examples per shift.

2.2.2. LFW Dataset
LFW is an unconstrained natural scene face recognition data set, which consists of more than 13,000 face-to-face images of different worlds of natural world scenes, expressions and lighting environments. There are more than 5,000 people, of which 1680 have 2 or more face pictures. Each face image has its unique name ID and serial number to distinguish it.

The LFW dataset mainly tests the accuracy of face recognition. The database randomly selects 6000 pairs of faces to form a face recognition image pair, of which 3000 pairs belong to the same person and 2 face photos, and 3000 pairs belong to different people. The test process LFW gives a pair of photos asking if the two photos in the system under test are the same person, and the system gives the answer of "yes" or "no". The face recognition accuracy rate can be obtained by the ratio of the system answer of the 6000 pairs of face test results to the real answer.
2.3. Face Recognition Model

Figure 3 is one of the structures of CNN network proposed in this paper. Take the Olivettifaces and three convolution layers network as an example. The size of the input image is 57*47 pixels. First, the data is normalized, and it is used as the input of the CNN. The first convolution layer has a kernel of 5*5. After convolution and Relu layer, the size of the image is unchanged, the output size is 57*47*32. After the pooling layer, the output becomes 29*24*32. Repeat the above process twice, the output of the last max-pooling layer is 8*6*128. In the fully connected layer, the data is stretched into a column vector with only one column, the dimension is 1024, and the last layer is the output layer whose vectors represents the number of categories.

3. Experimental results

During the experiment, the face dataset and CNN network for face recognition test are used, from recording the recognition accuracy, it comes a conclusion that watermark can significantly reduce the accuracy of face recognition. After that, different datasets, methods of watermark, CNN network structure, and different combinations of training sets and test sets with or without watermarks are tested. As for the watermark, OpenCV and Photoshop which are two most popular methods of watermark. Figure. 6, Figure. 8, and Figure. 9 show the result. While Set A means that both the train set and the test set are original. Set B means that the train set is watermarked but the test is original. Set C means that the test set is watermarked but the train is original. Set D means that both the train set and the test set are watermarked. Set E means that half of the train set is watermarked and half of the test set is watermarked.

3.1. With or without watermark in training set and test set

Combining the data in the three figures, the face recognition accuracy rate is 84% in the case of no watermark in training set and test set. The face recognition accuracy rate is 79.25% when both the training set and the test set are watermarked. The accuracy drop is 4.75%. When the training set and the test set are mixed with half of the watermark image, the face recognition accuracy is 74% the accuracy rate is reduced by 10%. When the training set is watermarked, the test set is without of watermark, the
accuracy rate decreased by 25.125%. In the case of the training set is without of watermark, the test set has a watermark, the accuracy rate decreased by 27%.

3.2. Network structures
Two different network structures’ performance are compared in Figure 6. The recognition accuracy of the 4 layers convolution network is higher than that of the 3 layer 2%. In the case of both watermarked, the 4 layers’ accuracy is lower by 2.5%. In the case of a watermark image mixed with a without-watermarked images, 4 layers' accuracy decreased by 3.875%. In the case of a watermarked test set, the 4 layers' accuracy is improved by 3%. In the case of a without-watermark training set, and watermarked test set, the 4 layers' accuracy is improved by 7.8%.

![Figure 6](image6.png)

Figure 6. This is a figure compares different network structures’ recognition accuracy; Figure (a) is based on Olivettifaces dataset and uses OpenCV as watermark method; Figure (b) uses Olivettifaces and Photoshop; Figure (c) uses LFW dataset and OpenCV; Figure (d) uses LFW and Photoshop.

3.3. Different datasets
As showed in Figure 8. In the case of no watermarking, compare the Olivettifaces to LFW, it is reduced by 1.25%. In the case of watermarking, it is decreased by 3.25%. In the case of a watermark image mixed with a watermark-free image, the accuracy is increased by 1.625%. In the case of watermark training set, without watermarking test set, it is increased by 4%. In the case of a watermark-free training set, the watermark test set is increased by 6.875%.

![Figure 7](image7.png)

Figure 7. This is a figure compares different datasets’ recognition accuracy; Figure (a) is based on 4 conv layers network and uses OpenCV as watermark method; Figure (b) uses 4 conv layers and Photoshop; Figure (c) uses 3 conv layers and OpenCV; Figure (d) uses 3 conv layers and Photoshop.

3.4. Different ways to add watermarks
It can be found in Figure 9 that compare OpenCV to photoshop, in the case of no watermarking, it is decreased by 1.25%; in the case of watermarking, it is increased by 0.625%. In the case of mixing watermark images and non-watermarking images, the accuracy is increased by 1.5%. In the case of watermark training set, without-watermark test set, it is increased by 1%. In the case of a without-watermark training set, watermarked test set, it is reduced by 0.9%.
Figure 8. This is a figure compares different watermark methods’ recognition accuracy. Figure (a) is based on 4 conv layers network and uses Olivettifaces dataset; Figure (b) uses 3 conv layers and Olivettifaces; Figure (c) uses 4 conv layers and LFW; Figure (d) uses 3 conv layers and LFW.

Figure 9. This is a figure shows part of Olivettifaces dataset which is added watermark using OpenCV.

4. Conclusion
According to the experiment results, the visible watermark will significantly influence the identification accuracy. Although the visible watermark’s loss of visual information doesn’t affect our cognition of certain image, but it will do disturb the accuracy in machine learning, that’s why the visible watermark hasn’t been the prime method to protect ownership in machine learning. And the effect differs because of the structure of CNN, dataset, method of watermark. Nowadays there are a lot of advanced methods to improve the accuracy of identification of visible watermarked images like GNN.

References
[1] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer, Verlag Berlin
[2] Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems, 9(2), 181-199.
[3] Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In: International Joint Conference on Neural Networks. Washington DC. pp. 65-93.
[4] Joachims, T. (1999) Making Large-Scale SVM Learning Practical. In: Scholkopf, B., Burges, C. and Smola, A. (Eds.), Advances in Kernel Methods Support Vector Learning, Chapter 11, MIT Press, Cambridge. pp.169-184.
[5] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. Lake Tahoe. pp. 1097-1105.)
[6] Parkhi, O.M., Vedaldi, A. and Zisserman, A. (2015) Deep Face Recognition. In: Proceedings of the British Machine Vision Conference (BMVC). Swansea. pp. 41.1-41.12.

[7] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Vancouver. pp. 815-823.

[8] Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. Boston. pp. 8201-8204.

[9] T. Kimoto.(2013) Spatial Image Watermarking by Error-Correction Coding in Gray Codes. Journal of Signal and Information Processing, 04(3): 259-273.

[10] Rahman, M., & Harada, K. (2006). Parity enhanced topology-based spot area watermarking method for copyright protection of layered 3D triangular mesh data. Int J Comput Sci Netw Secur, 6(2): 252-259.

[11] Kang, X., Huang, J., & Zeng, W. (2008). Improving robustness of quantization-based image watermarking via adaptive receiver. IEEE Transactions on Multimedia, 10(6): 953-959.

[12] Lu, C. S., & Liao, H. Y. (2001). Multipurpose watermarking for image authentication and protection. IEEE transactions on image processing, 10(10): 1579-1592.

[13] Wu, C. F., & Hsieh, W. S. (2000). Digital watermarking using zerotree of DCT. IEEE Transactions on Consumer Electronics, 46(1): 87-94.

[14] Raval, M. S., & Rege, P. P. (2003, October). Discrete wavelet transform based multiple watermarking scheme. In: TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region. Bangalore. pp. 935-938.

[15] Yang, Y., Sun, X., Yang, H., & Li, C. T. (2008). Removable visible image watermarking algorithm in the discrete cosine transform domain. Journal of Electronic Imaging, 17(3), 033008.

[16] Gao, H. Y., Liu, G. Q., Li, X., & Xu, Y. (2009, August). A robust watermark algorithm for jpeg2000 images. In: 2009 Fifth International Conference on Information Assurance and Security. Xian. Vol. 2, pp. 230-233.

[17] Mohanty, S. P., Ranganathan, N., & Namballa, R. K. (2003, August). VLSI implementation of invisible digital watermarking algorithms towards the development of a secure JPEG encoder. In: 2003 IEEE Workshop on Signal Processing Systems. SEOUL. pp. 183-188.

[18] Zhou, Y., Bao, L., & Chen, C. F. (2014). A new 1D chaotic system for image encryption. Signal processing, 97: 172-182.