Miroslaw Pawlak 05 November 2023

External Examiner's Report on the Doctoral Thesis of Marcin Plata of Polish Academy of Science

Thesis title: Zastosowanie zaawansowanych metod sztucznej inteligencji do wybranych problemów bezpieczenstwa informacji

I. General Comments

This thesis is concerned with the emerging topic in machine learning (ML) that can be called secure AI or adversarial learning. Here one relaxes the common assumption that environment is friendly during the training and testing periods of the trained model. In practice, however, malicious adversaries may interfere with the learning and testing processes. Hence, it is an important task to design ML systems being robust against adversarial attacks.

The submitted thesis, however, is not focusing on one particular problem on ML systems in the adversarial setting but rather examines several topics from watermarking and biometrics. The methodology used in the thesis is (almost) entirely based on deep neural networks algorithms. As such the thesis is of the implementation/design nature.

The first part of the thesis provides background to the area of study and an overview of the five papers that represent the most material presented in the dissertation. There are three conference papers and two journal ones including the paper in the highly ranked journal of *Pattern Recognition*. The core of the thesis consists of five chapters (Chapters 2-6) that present somehow connected material. Chapters 2 and 3 are strongly correlated, whereas the remaining chapters present separate studies on various aspects of secure biometrics. The common tool used throughout the thesis is the neural networks paradigm for designing the proposed learning algorithms.

In the rest of this report, I first overview each of the five chapters. This will be followed by detailed comments on the results. At the end of this report, I summarize my overall view on the thesis and provide final recommendations.

II. Thesis Overview

1. Watermarking Systems

Watermarking is the problem of designing of systems for embedding one signal, called a watermark within another signal, called a host signal. The embedding must be done such that the embedded signal is hidden and no causing serious degradation to its host. At the same time the embedding must be robust for possible attacks that can damage or even remove the watermark. Finally, the extraction (detection) of the watermark should be efficient and simple. Watermarking systems are examined in Chapters 2 and 3 and the description below will briefly detail the obtained results.

1.1. Watermarking Using Neural Networks

In Chapter 2 the convolution neural network architecture is utilized for the designing watermarking system based on the concept of spatial spreading of message bits. The latter is represented by the propagator block being the tuple based spatial design of the message. This is associated with translator at the decoder side that performs the inverse operation to the propagator. The system takes the form of the common communication structure, i.e.,

$$\text{encoder} \rightarrow \text{channel (called noiser)} \rightarrow \text{decoder.}$$

This system is augmented by adding the discriminator block that allows us to determine the transparency of the watermarked image. The encoder, decoder, and discriminator are trained blocks and they are modeled by convolution neural networks that are optimized with respect to a new mean-variance cost function. The system is trained using a sequence of input images and randomly selected class of channel distortions (attacks). The attacks that are taken into account are the following: cropping, dropout, Gaussian convolution, rotation (by a small angle), image size, sampling, and JPEG compression. The designed system was empirically assessed yielding the favourable bit accuracy uniformly with respect to the aforementioned image distortions.

1.2. Watermarking with the Discriminator-Detector Scheme

In Chapter 3 the refined version of the watermarking system from Chapter 2 is introduced and examined. This new architecture is also based on the deep neural network algorithm augmented by the double detector-discriminator scheme. This provides the joint testing (on the presence of a watermark) and detection strategy. As a result this yields an improved watermarking system that reveals the high encoded transparency, resistance to some attacks (JPEG compression, rotation), and the overall robustness. The author concludes that the proposed watermarking system outperforms the existing deep learning approaches in terms of the resilience and complexity, while maintaining the high transparency and capacity.

2. Biometric Systems Chapters 4, 5, and 6 examine the selected problems in the field of biometrics. Biometric systems use distinctive and measurable human characteristics to form identification marks for the purpose of personal authentication, privacy and security. Common biometric identifiers are based on face recognition, fingerprint, palm print, iris recognition, hand geometry, voice, and finger vein recognition. The thesis focuses on biometric systems utilizing face recognition (Chapter 4), voice recognition (Chapter 5) and hand geometry (Chapter 6).

2.1. Privacy in the Context of Facial Recognition

The facial recognition based biometric system is proposed in Chapter 4. The main goal is to design a system that provides a higher level of data privacy compared to the existing methods. This goal was achieved by selecting the critical subset of face landmark points utilizing facial expressions rather than the static face features. This has resulted in the increasing privacy of the solution. The selection process of face landmarks is based on the heuristic approach of removing features that are highly correlated. The designed system was tested on a sequence of videos from a set of different individuals proving its practical usefulness.

2.2. Preventing Spoofing Attack in Speaker Recognition

Chapter 5 examines another biometric problem using the voice recognition and speaker identification characteristics. In particular, the main objective is to obtain a time-efficient method to prevent spoofing attacks. This goal is examined by means of lightweight convolutional networks and Bayesian networks of the reduced complexity. The choice of the system parameters is based on the modified version of cross-validation called the attack-out cross-validation procedure. This strategy allows to estimate the effectiveness of the trained model against spoofing attacks which the model does not have access during the training phase. The designed system applicability was confirmed in the 2019 ASVspoof Challenge.

2.3. User Identification Based on the Hand Contour Geometry

Chapter 6 presents the final biometric system that utilizes the hand contour geometry. This is the computer vision driven system that extracts the initial 62 hand geometric features. Using the feature quality index the feature selection procedure is performed resulting in the final 19 biometric features. The decision process is based on the simple single threshold rule. This parsimonious strategy is motivated by the time-efficiency and hardware constrains. In fact, the system can be implemented using a standard office scanner.

III. Specific Comments

Chapter 2 In a watermarking system we are embedding a digital watermark within an input host signal to form the watermarked signal that is transmitted through the ”noisy” communication channel. The channel is a source of various degradation processes (attacks) that damage or even remove the watermark. The embedding is designed to achieve efficient tradeoffs among the three conflicting goals: maximizing information-embedding rate, minimizing distortion between the host signal and watermarked signal, and maximizing the robustness of the embedding. This description easily reveals that the watermarking problem is strictly related to the information theory. In Chapter 2 (and Chapter 3) the author employs the black-box strategy using deep neural networks (NN) for the watermarking problem applied to image data. The proposed system is well described and its accuracy is evaluated by means of simulation studies. Here are the specific comments related to the material presented in this chapter.

1. The watermarking problem has deep roots in the information theory, where some optimality results (for restricted attacks) were proved. This literature is entirely ignored in the thesis. Here are some classical references:
    - M. Costa. Writing on dirty paper. *IEEE Trans. Inf. Theory*, 1983. The classical paper with the comprehensive information theory based analysis of an additive distortion system.
    - P. Moulin and R. Koetter. Data-hiding codes. *Proceedings IEEE*, 2005. The review on the information theory approach to data hiding.
2. The class of considered attacks include: cropping, dropout, Gaussian convolution (low-pass filtering), rotation (by a small angle), image size, sampling, JPEG compression. In the pre-deep NN literature other important attacks were taken into account such as: histogram equalization, median filtering, noise (Gaussian, salt and pepper) just to name a few. Here are some references

    Y. Xin, S. Liao and M. Pawlak. Circularly orthogonal moments for geometrically robust image watermarking. *Pattern Recognition*, 2007. Technique tailored to geometric attacks, e.g., all affine transformations.

    Wenbo Wan et al. A comprehensive survey on robust image watermarking. *Neurocomputing*, 2022. This contribution already includes NN methods.

Yet another important attack is the following warping deformation that is common, e.g., in medical imaging (radiology)

$$I'(x,y) = I(\Theta(x,y)),$$

where $\Theta(x,y)$ is a continuous bijective (1-1 and onto) warping function. Can we extend the presented methods to the aforementioned attacks ?

3. The learning process assumes that the channel (noiser) parameters are known. Is it possible to jointly learn the source (encoder) and channel parameters ? In the communication theory this problem is referred to as the joint source-channel coding.
4. The distance measure defined in (2.6) for comparing images is the L_2 norm. It is well known in image analysis that such measures do not reflect the contextual and subjective difference between images. There is a well established theory on alternative image distances, see the seminal work by Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Trans. on Image Processing*, 2004. Clearly these new image distances are much harder to minimize.
5. The measure for comparing messages M and M' defined in (2.10) is a linear combination of the average and variance terms. I would use the square of the mean to have both terms of the same size. Then, this would be equivalent to the ratio $Var/Mean^2$ being the square of the so-called coefficient of variation that has some importance in statistics. Clearly the ultimate measure between M and M' would be the probability $Pr[M' \neq M]$.
6. The hyperparameters appearing in the final risk function in (2.11) are assumed to be known. Normally they should be determined during the validation process.
7. It is customary in ML to perform the ablation analysis, i.e., to remove a certain element of the designed system in order to observe its role in the overall performance. For example, we would like to see the importance of the discriminator block.
8. Some symbols are not defined: $\mathbb{E}$ in (2.11), PSNR.

Chapter 3

The watermarking system examined in this chapter is the augmented version of the design in Chapter 2. The additional problem that is addressed here is of the hypothesis testing nature, i.e., we wish to verify the null hypothesis $H_0 : I_e = I_c$ versus the alternative $H_1 : I_e \neq I_c$. This interesting statistical question is resolved by using the NN based discriminator.

1. All the concerns that were discussed in Chapter 2 also apply to the system proposed in Chapter 3.
2. The performance measures for the joint discriminator-detector system are identical as those used in the theory of statistical testing. The test statistic is: compare the normalized (to the interval $[0, 1]$) output of the discriminator ($F(I)$) to the threshold value t_F and if $F(I) \leq t_F$ then one declares that the image does not contain the watermark. In statistics t_F is selected by controlling the size of the probability of false rejection, i.e., $Pr[F(I) > t_F|H_0]$. How the threshold t_F is selected in the process of designing the discriminator-detector is not clear to me.
3. Some simple issues: Fig. 3.1, I do not see ”czerwone strzalki”. Page 60: ”W pracy”. Shouldn't be ”w tym rozdziale”. Page 38: nieurnowych ?

Chapter 4

Chapter 4 examines the secure biometric system based on dynamic face features. The selected features represent facial expressions rather than the static face features.

1. The key component of the proposed system is the extraction of face landmark points. The systematic methodology of representing of a planar object (an image) by landmark points has been developed in the statistical shape theory. This theory provides an elegant way of understanding the concept of shape, shape similarities, shape distance and much more. The theory maps the landmark points into the configuration space (not being the Euclidean space) and defines such concepts as size of the configuration, the distance between configurations, and probabilistic descriptors of the configuration. The most useful fact is the so-called Procrustes (Procrustes analysis) distance between configurations that is invariant with respect to object translation, rotation, and scale. Unfortunately, this methodology has been greatly ignored in computer vision and ML and this is also the case in this thesis. Here are some basic references:

- F. Bookstein. Morphometric Tools for Landmark Data: Geometry and Biolog, Cambridge Univ. Press, 1991.
- D. G. Kendall, D. Barden, T. K. Carne, H. Le. Shape and Shape Theory, Wiley, 1999.
- U. Grenander and M. Miller. Pattern Theory: From Representation to Inference, Oxford Univ. Press, 2007.

2. Page 71: it is said that the regression trees method was used. How do you define the regression function in this case, i.e., what is the independent variable ($\boldsymbol{X}$) and which one is the dependent variable (Y). We know the regression of Y on $\boldsymbol{X}$ is the conditional mean $\mathbb{E}[Y|\boldsymbol{X}]$.

3. The feature selection is based on the classical (Pearson's correlation) correlation function. This type of correlation often leads to misleading conclusions as it measures only the linear dependence and is sensitive to outliers. The rank correlation (Spearman's correlation) goes beyond the linearity, measures a monotone association between data and also is robust to outliers.
   The correlation based feature selection procedure eliminates the minimum number of features to ensure that all pairwise correlations are below a certain threshold. This clearly identifies only pairwise linear correlations. Nevertheless, the simplicity of the method is a positive fact and even may lead to the significant improvement on the performance of the model. This is confirmed in the thesis, where the 50 most correlated features were removed yielding the acceptable performance.
   It is also worth noting the PCA approach produces de-correlated features. However, in this case the relationship between the features and outcome is more complex.

4. The decision method used for the person identification problem is the SVM algorithm modified to have the probabilistic output. What type of the SVM algorithm is used, the classical linear, regularized, or SVMs with kernels ? The smoothed output version of SVM is quite artificial. There are ML techniques that naturally give the decision in the form of the probability, e.g., a class of Logistic regression classifiers.

Chapter 5

1. The formal model of the automatic speaker verification is introduced in Section 5.4. This model is equivalent to the so-called in

statistics two-sample testing, i.e., given two independent sets of data

$$X_i \quad \text{distributed according} \quad f_X, \quad i = 1, \ldots, n_1,$$

$$Y_i \quad \text{distributed according} \quad f_Y, \quad i = 1, \ldots, n_2$$

we wish to verify the null hypothesis $H_0 : f_X = f_Y$ versus $H_1 : f_X \neq f_Y$. The classical approach to this question is to use the Hotelling T^2 test statistic. The author applies a simple threshold rule using the output of the ASV system.

2. The validation of the joint ASV + CM system is performed utilizing the Bayes decision theory. Here we have the decision problem in which the parameter space contains 3 points and decision space contains 6 points. The author considers the randomized decisions and defines the corresponding risk in (5.6). It should be noted, however, that the randomized decisions do not reduce the risk and one can confine the decision space to pure decisions, see M. DeGroot, *Optymalne Decyzje Statystyczne*, PWN, 1981, pp. 110 - 117. The minimization of the risk can be efficiently obtained using the result of Theorem 1 (p. 115) in DeGroot's book.

3. In Section 5.6.6 the author introduces the new validation measure refereed to as Attack-Out Cross-Validation. This is the version of the standard k-fold Cross-Validation technique. The new method has been employed for the specific problem examined in the thesis. Can you suggest of a wider application of the method ?

4. The Bayesian neural networks are discussed in Section 5.6.7. Here one makes the randomization of the network weights $\boldsymbol{W}$. Then, the predictive probability $P[y|\boldsymbol{x}, \boldsymbol{W}]$ should be understand as the average $\mathbb{E}_{\boldsymbol{W}}[P[y|\boldsymbol{x}, \boldsymbol{W}]]$. As a result, the right-hand-side of the formula in (5.20) is the approximation of the average. The left-hand-side of (5.20) should be corrected.

   The section on the Flipout layer describes the practical way of representing random weights by the so-called perturbation process. In particular for a single-weight the following representation $\Delta W_{ij} = \widehat{\Delta W}_{ij} E_{ij}$ is recommended, where E_{ij} is a random variable taking values in $\{-1, 1\}$ and being independent on $\widehat{\Delta W}_{ij}$. It is said that if ΔW_{ij} has an even density then the density of $\widehat{\Delta W}_{ij}$ is the same. This is not generally true. Consider the simplified version of this problem. Hence, let X and ϵ be independent random variables, where $f_X(x)$ is the pdf of X, whereas $\epsilon \in \{-1, 1\}$ with probabilities $1-p, p$, respectively. Then, the pdf of $Y = X\epsilon$ is

$$f_Y(y) = f_X(y)p + f_X(-y)(1-p).$$

Clearly if $f_X(x)$ is even then $f_Y(y) = f_X(y)$ for any $0 < p < 1$. If, however, $f_Y(x)$ is even this does not imply that $f_X(x)$ is even for every $0 < p < 1$.

Hence, we must assume that $\widehat{\Delta W}_{ij}$ has the even symmetric density.

5. Some typos: in (5.10) it should be dr; in (5.11) it should be dr; in (5.12) it should be dr. Undefined symbol $\circ$ on p. 115.

Chapter 6

In Chapter 6 the problem of the similar nature as in Chapter 4 is examined, i.e., designing a biometric system for image data and in this chapter the hand contour geometry is taken account as the biometric mark. This is the elegant parsimonious design based on the geometric landmark points. As it was pointed out in the discussion of Chapter 4, one could also apply the well developed statistical shape theory, where the concept of landmarks points in the configuration space plays a key role. Analysis of the hand contour geometry utilizing this methodology is an interesting alternative for future research.

The theory of contour based image descriptors is described in A. Blake and M. Isard, *Active Contours*, where the analysis of dynamic (moving) contours was also examined. Here are some classical references on the contour based image analysis.

- A. Blake and M. Isard, Active Contours, Springer, 1998.
- U. Grenander, Y. Chow, and D. Keenan. Hands: A Pattern Theoretic Study of Biological Shapes, Springer, 1991.

Chapter 7 The final conclusions are lacking any description on future research and possible extensions.

References

The thesis has an impressive long number of references. This, however, is predominated by conference papers. There is the lack of any general references on deep NN, e.g., *Deep Learning*, I. Goodfellow, Y. Bengio and A. Courville, The MIT Press, 2016. It would also be useful to have a single monograph on state-of-the-art of ML, e.g., *Probabilistic Machine Learning: Advanced Topics*, K.P. Murphy, The MIT Press, 2023.

IV. Final Comments and Recommendations

Pros:

1. I admire the broad scope of topics that the candidate included in the thesis. This makes the reviewing process quite challenging.

2. In my view, the thesis is very nicely organised and presented, with helpful illustrative figures. The writing is of a very high standard, making the thesis easy to read and follow.

3. The thesis is of the very practical/engineering nature with a number of useful solutions that can find applications in the real-world systems.

4. Importantly, the clarity of explanations enables any empirical experiments to be reproduced by other authors. For example, there are clear explanations for the methods considered, and the empirically driven choices of the design parameters.

5. In terms of novel contributions, the thesis contains several interesting new empirical solutions and presents the state-of-the-art strategies for the neural networks based algorithms for watermarking and secure AI.

6. The thesis covers a very wide range of topics from watermarking, face recognition, voice processing to hand analysis. In each of these areas the candidate has made practical contributions by enlarging the scope of their applications.

7. The candidate demonstrates practical knowledge of the modern deep neural networks learning algorithms.

Cons:

1. The main drawback of this research is the lack of any statistical performance analysis. In fact, the thesis gives the state-of-the art design paradigm for secure ML systems without, however, any statistical/mathematical analysis on the accuracy of the proposed learning methods. I should admit that such analysis can be involved and complex.

2. In several cases there is the lack of systematic procedures for the data-driven choice of design parameters (hyperparameters).

3. It is customary in ML to perform the ablation analysis. There is no such analysis in the thesis.

4. There is a very little discussion on possible future research.

5. Some relevant classical contributions to watermarking and image analysis have not been acknowledge.

Final Conclusions: In summary, the thesis develops a number of practical and useful contributions to the important field of adversarial and secure machine learning. The thesis meets the statutory and customary requirements for doctoral dissertations specified in the Act on Scientific Degrees and Titles: article187 of the Act on higher education and science (Journal of Law of 2023, position 742, unified text). I recommend the thesis to be accepted for public defense.
Note: It would be useful to have the thesis translated into English for wider accessibility and impact. It is common to write PhD dissertations in English in many EU countries, e.g., Germany, Sweden and even France.
Signed:

MPawlak

Professor Miroslaw Pawlak
University of Manitoba, Canada
AGH Krakow, Poland.