Steganographer Detection based on Multiclass Dilated Residual Networks

Mingjie Zheng
College of Compute Science and Software Engineering
Shenzhen University
Shenzhen, China
zhengmingjie@email.szu.edu.cn

Sheng-hua Zhong*
College of Compute Science and Software Engineering
Shenzhen University
Shenzhen, China
csshzhong@szu.edu.cn

Songtao Wu
College of Compute Science and Software Engineering
Shenzhen University
Shenzhen, China
csstwu@szu.edu.cn

Jianmin Jiang*
College of Compute Science and Software Engineering
Shenzhen University
Shenzhen, China
jianmin.jiang@szu.edu.cn

ABSTRACT

Steganographer detection task is to identify criminal users, who attempt to conceal confidential information by steganography methods, among a large number of innocent users. The significant challenge of the task is how to collect the evidences to identify the guilty user with suspicious images, which are embedded with secret messages generating by unknown steganography and payload. Unfortunately, existing methods for steganalysis were served for the binary classification. It makes them harder to classify the images with different kinds of payloads, especially when the payloads of images in test dataset have not been provided in advance. In this paper, we propose a novel steganographer detection method based on multiclass deep neural networks. In the training stage, the networks are trained to classify the images with six types of payloads. The networks can preserve even strengthen the weak stego signals from secret messages in much larger receptive filed by virtue of residual and dilated residual learning. In the inference stage, the learnt model is used to extract the discriminative features, which can capture the difference between guilty users and innocent users. A series of empirical experimental results demonstrate that the proposed method achieves good performance in spatial and frequency domains even though the embedding payload is low. The proposed method achieves a higher level of robustness of inter-steganographic algorithms and can provide a possible solution to address the payload mismatch problem.

*Sheng-hua Zhong and Jianmin Jiang are the corresponding authors of this paper.

ICMR '18, June 11–14, 2018, Yokohama, Japan

ACM ISBN 978-1-4503-5046-4/18/06…$15.00
https://doi.org/10.1145/3206025.3206031

CCS CONCEPTS

• **Computing methodologies** → **Neural networks** • **Security and privacy** → **Software and application security**

KEYWORDS

Steganographer detection; multiclass classification; deep neural networks; multimedia security

ACM Reference format:

M. Zheng, S. Zhong, S. Wu, J. Jiang. 2018. Steganographer Detection based on Multiclass Dilated Residual Networks. In *Proceedings of ACM International Conference on Multimedia Retrieval 2018, Yokohama, Japan, June11–14, 2018 (ICMR'18)*, 9 pages.
DOI: https://doi.org/10.1145/3206025.3206031

1 INTRODUCTION

Image steganography is the technique of concealing secret messages within images, so that the existence of the hidden data is not obvious in covert communication [1]. In general, the image in which the secret messages are intended to be embedded is termed as the *cover* image, and the image which includes messages is referred to the *stego* image [2]. Image steganalysis, from an opponent's perspective, is an art of revealing the presence of secret messages in the images, which are hidden by the steganographer [3]. Steganography and steganalysis are in a hide-and-seek game [3]. They try not only to beat each other, but also to create strategic coalitions to develop.

Currently, most of image steganalytic techniques are devoted to separate a suspicious image as cover image or stego image. This problem is termed as *stego detection* problem. However, in the real-world situation, there exist multiple users, and each user will transmit vast numbers of images. Moreover, only some of users are guilty of using unknown and diverse steganography with unknown embedding parameters such as embedding payload. It's more practical than the "laboratory environment" in most of image steganalytic tasks. Image steganalytic techniques

are desired to solve more challenges beyond the laboratory conditions. Therefore, in this paper, we devote to investigate a completely different problem: identifying which users try to hide secret messages into the images with steganography among many innocent users [4]. This kind of user is called as the guilty actor (or guilty user). This problem is known as *steganographer detection* problem. It poses significant challenge in comparisons of stego detection problem, which is how to collect the evidences from multiple images of suspicion, and then to identify the guilty user. We believe that steganographer detection will play an imperative role in many significant multimedia security applications.

Most of existing steganographer detection methods mainly depend on the handcrafted features based on the traditional steganalytic algorithms. As the first attempt, in [5], Ker *et al.* extracted PEV-274 features [6] from each image and calculate the distance of each pair of users by maximum mean discrepancy (MMD) [7]. Finally, they formulated the steganographer detection task as a clustering problem to separate the guilty user from majority innocent users. In 2012 and 2014, Ker *et al.* considered the steganographer as the outlier among innocent users and proposed to replace hierarchical clustering by the local outlier factor (LOF) method [8] to rank user's possibility of being guilty according to the degree of anomaly [9, 10]. In 2016, Li *et al.* proposed a method that used high-order joint features and clustering ensembles [4]. The high-order joint features were procured from high-order joint density matrices of Discrete Cosine Transformation (DCT) coefficients from JPEG images. In 2017, Li *et al.* extended the work in the way of proposing a sampling construction strategy [11]. They designed an embedding probability calculation model and selected DCT blocks with higher embedding probability to reconstruct a sample image. Then, they extracted a 155-D reduced PEV feature set from each sample image. Finally, the agglomerative hierarchical clustering was used to identify the guilty user according to their corresponding MMD distances.

In these years, deep learning based methods have achieved great success in many multimedia tasks, such as image retrieval [12], face recognition [13], emotion recognition [14] and so on. One of the most important contributions of deep learning techniques is extracting more efficient features through features auto-learning rather than handcrafted. Currently, there is limited existing work based on deep learning to solve the steganographer detection task. However, many works [15–24] based on deep learning techniques have been proposed for steganalysis, which is close to steganographer detection task. These methods directly learn the discriminative features to separate stegos from covers, which are benefit from the process in deep learning techniques. For instance, in 2016, Xu *et al.* reported a well-designed convolutional neural network (CNN) architecture which took into account of the knowledge of steganalysis [18]. The results demonstrated that the proposed model was competitive compared with that achieved by the spatial rich model. In 2017, Xu presented an empirical study on applying CNNs to detect JPEG version of the UNIversal WAvelet Relative Distortion [23]. Wu *et al.* proposed the deep residual network for image steganalysis [20]. By virtue of the deep residual learning, the network could capture the weak signals from the secret messages which brought that the proposed model achieved better performance than the classical method spatial rich model and several other CNN-based models.

Inspired by these steganalytic methods based on deep learning, in 2017, we were first to propose the steganographer detection method based on deep residual network [25]. The effective features extracted by residual learning were used to calculate the distance between each pair of users by MMD. Finally, the guilty user was identified by the agglomerative hierarchical clustering algorithm. Although the proposed framework could achieve good detection accuracies in spatial domain when the payload was low, all of stego images must be paired with their corresponding cover images when extracting the features both in training and the inference stage.

Currently, existing methods for the steganographer detection task mainly depend on the features extracted from the steganalytic methods. In other words, the features are utilized to separate stego image from cover image. However, these existing works based on deep learning for steganalysis were served for the binary classification. Thus, these methods ignore the important information from stego images, such as the embedding payload. Let us think about the distortions caused by different payloads. We believe the distance between cover image and stego image with low payload, e.g. 0.1 payload, is smaller than the distance between stego images with 0.1 and 0.4 payloads. But in the binary classification for image steganalysis, this kind of differences has been ignored. This ignorance makes it harder to classify the images with different kinds of payloads, especially when the payloads of images in test dataset have not been provided in advance. This phenomenon belongs to the cover-source mismatch problem [26], which makes the detection accuracy much lower, especially when the payload is low. To solve this challenge, one kind of solution is to train each model for each payload, and then utilizes the ensemble strategy to obtain the final result based on all models. But it inevitably requires high computational complexity.

In our paper, we propose a novel steganographer detection method based on multiclass deep neural networks. In machine learning, multiclass is the problem of classifying instances into one of three or more classes. In our model, the dilated residual networks are trained for multiclass classification of images with six kinds of embedding payloads. In the inference stage, we employ the learnt model as the feature extractor to extract the discriminate features from each image of each user. Then, the agglomerative hierarchical clustering algorithm is used to identify the steganographer according to the corresponding distance metrics. To our best knowledge, we are the first to propose a steganographer detection framework based on multiclass neural networks. This multiclass neural networks are an attempt to solve the payload mismatch problem. In our experiments, we also find that the multiclass neural networks can learn effective features from the information of relatively high payloads and promote to separate the low payload embedded stegos from the covers.

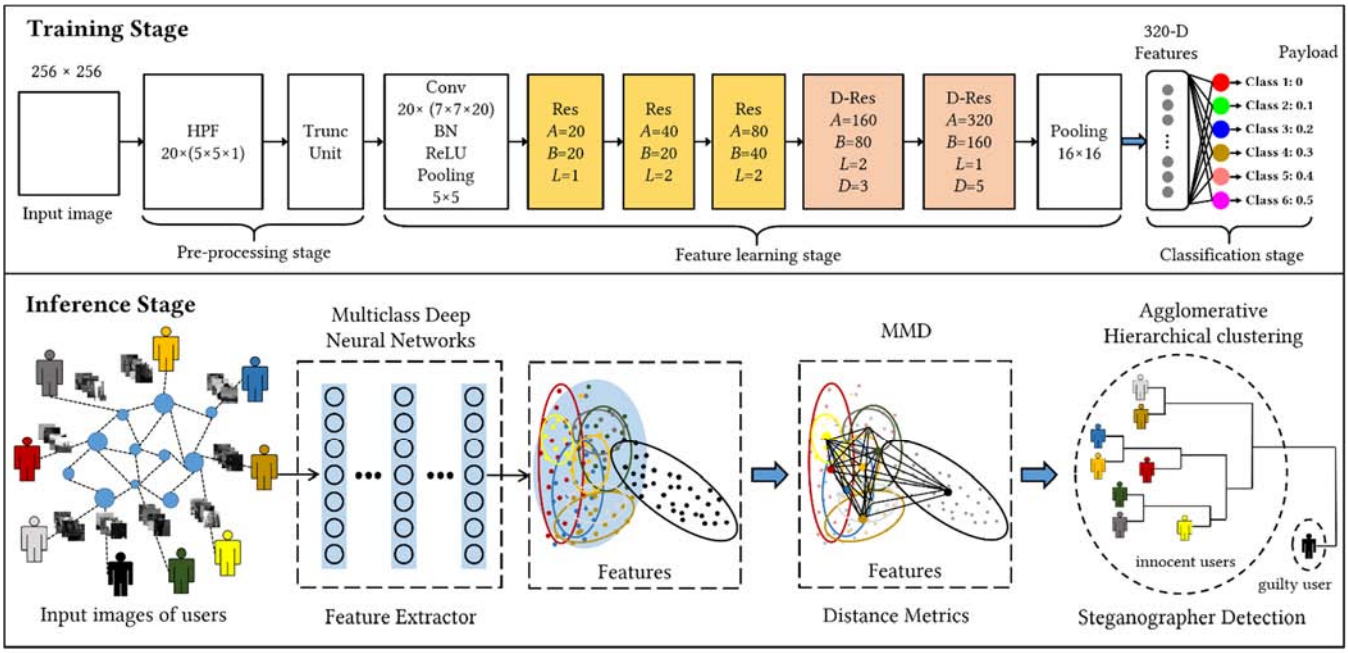

Figure 1: Multiclass deep neural networks based steganographer detection.

2 PROPOSED METHOD

In this paper, we propose a novel steganographer detection method based on multiclass deep neural networks. Fig. 1 illustrates the framework of Multiclass Deep Neural Networks based Steganographer Detection (MDNNSD). In the training stage, the multiclass dilated residual networks are trained for the classification of images with six types of embedding payloads. In the testing stage, the learnt model is utilized to extract features of each image from each user. Then, the distance between each pair of user is calculated based on MMD algorithm. Finally, the agglomerative hierarchical clustering algorithm is performed to identify the guilty user.

2.1 Feature Extraction via Multiclass Neural Networks

As illustrated in Fig. 1, in the learning part, the multiclass neural networks are learnt to classify images with six types of embedding payloads. In the inference part, the learnt model is utilized to extract the features of each image from each user. The proposed networks consist of three parts: the pre-processing stage, the feature learning stage and the classification stage.

The pre-processing stage aims to extract the message (noise) component, which includes two layers: the High-Pass-Filtering (HPF) layer and the truncation layer. As we known, the steganographic algorithm can be considered as adding low-amplitude additive noise to cover images. Hence, it means the weak stego signals have much lower amplitude in comparison of that of the image content. To suppress image content and extract effective information from low signal-to-noise ratio (SNR) stego signals, in the HPF layer, 20 high-pass filters are used to extract the high frequency components from the input. To guarantee the pre-processing stage has the ability to extract the high frequency component, each filter of HPF layer is initialized by a high-pass kernel. In the truncation layer, we try to use the truncation unit to constrain the dynamic range of input feature maps. This setting is also used to improve the convergence speed. The truncation unit is defined by Eq. 1 as follows:

$$Trunc(x) = \begin{cases} -T, & x < -T \\ x, & -T \le x \le T \\ T, & x > T \end{cases} \quad (1)$$

where T denotes the truncation threshold.

The feature learning stage is to extract the discriminative features. This sub-network mainly consists of two kinds of units: the Res and the D-Res units. The structure of the Res and the D-Res units are shown in Fig. 2. In the feature learning stage, the networks first use 20 convolutional kernels with the size of 7×7 to preserve the noise components generated in the pre-processing stage. Then, a series of units are utilized to extract the discriminative features, which include the residual learning units and dilated residual learning units. For the residual learning block in residual learning units, it fits the residual function $F(\mathbf{s}) := H(\mathbf{s}) - \mathbf{s}$ rather than approximating an underlying function $H(\mathbf{s})$ directly, which can be implemented by feedforward networks with the shortcut connections [27]. He *et al.* has proved that the networks could be easier to optimize by the short connections [27]. Some prior works have also proved that deep residual learning is benefit for extracting effective patterns from weak stego signals [20, 25]. Hence, in the feature learning stage, we take advantage of the residual learning to capture the weak stego signals from the steganographer.

Specifically, for the Res unit, it first uses the projection shortcut connection to increase the dimension of feature map, and then directly utilizes the identity shortcut connections in the following repeated building blocks.

The structure of D-Res unit is similar with that of the Res unit in addition to replacing all convolutional layers with the dilated convolutional layers. In our networks, we take advantage of the dilated residual unit with a larger receptive field. It is utilized to preserve the effective information from the weak signals generated by the guilty user. After a series of units processing, an average pooling layer with the size of 16×16 to transform the feature maps into feature vectors.

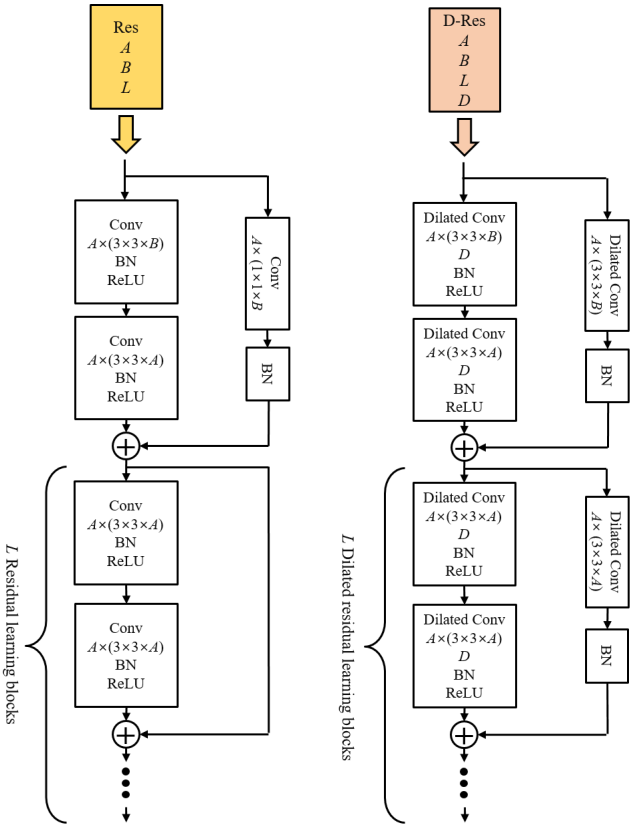

Figure 2: The structure of the Res and the D-Res units. Conv represents convolutional layers, with kernel sizes following (number of kernels) × (height × width × number of channels). *L* denotes the number of residual or dilated residual learning blocks. *D* denotes the dilation value in the dilated convolutional layers.

For the classification stage, in the training phase, the networks map extracted feature vectors into six labels. Six output nodes correspond to the embedding payload of the image, *i.e.*, 0, 0.1, 0.2, 0.3, 0.4 and 0.5, respectively. To train the networks, the feature vectors, which are generated by the last pooling layer in the feature learning stage, are fully connected to the six output nodes for multiclass classification. In the inference stage, we consider the learnt model as the feature extractor. Therefore, 320-D feature vector of each image from each user is generated from the last average pooling layer in the feature learning stage.

2.2 Distance Metrics based on MMD

By virtue of multiclass deep neural networks, the discriminative features of each image from each user are extracted via the learnt model. Then, the distance between each pair of users is calculated by maximum mean discrepancy (MMD) based on the extracted features. Mathematically, MMD is employed to measure the similarity between two probability distributions. In our work, we use MMD as the distance metrics to calculate the similarity between feature sets from any pair of users.

We denote a set of m users as U_1, U_2, ..., U_m, which include one guilty user and m-1 innocent users, and each of them transfers n images. The feature sets $\mathbf{F}_X = (\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n)$ and $\mathbf{F}_Y = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ are extracted from the images transferred by U_X and U_Y, where $\mathbf{x}_i$ and $\mathbf{y}_i$ $(1 \le i \le n)$ denote a 320-D feature vector of an image from U_X and U_Y, respectively. The sample estimate for the MMD distance of U_X and U_Y is defined by Eq. 2 as follows:

$$d(U_X, U_Y) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} k(\mathbf{x}_i, \mathbf{x}_j) - \frac{2}{n^2}\sum_{i,j=1}^{n} k(\mathbf{x}_i, \mathbf{y}_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(\mathbf{y}_i, \mathbf{y}_j)} \quad (2)$$

where $k(\mathbf{x}, \mathbf{y})$ represents a positive definite kernel function, including the linear kernel and the Gaussian kernel. In this paper, we employ the Gaussian kernel to calculate the MMD distance.

2.3 Steganographer Detection by Hierarchical Clustering

After the MMD distance between each pair of users is calculated, we identify the steganographer (guilty user) based on the MMD distance by using the agglomerative hierarchical clustering algorithm.

The agglomerative hierarchical clustering algorithm is a representative method of cluster analysis. Initially, each user is considered as a singleton cluster. Then, the nearest two clusters are combined by using the linkage function and form a new larger cluster. It is repeated until the last cluster is linked and a complete binary hierarchical tree is formed. Ideally, for the steganographer detection task, all the innocent users should be clustered as a cluster and the other cluster only consists of the guilty user.

In this paper, we use four kinds of linkage functions to calculate the distance between two clusters based on the MMD distance, including the average linkage, the single linkage, the complete linkage, and the weighted average linkage.

3 EXPERIMENTS

A series of empirical experiments are conducted on two standard datasets. In Section 3.1, we test the proposed method on a standard spatial domain dataset BOSSbase ver 1.01 [28]. Then, in Section 3.2, we further evaluate the proposed method in frequency domain. The images from BOSSbase ver 1.01 are compressed with JPEG quality factor 80, and they are used to test

the effectiveness of the proposed method. In the following sections, we call it as JPEG version of BOSSbase ver 1.01.

For the training stage of our proposed networks, parameters are updated by stochastic gradient descent. Each element $\mathbf{W}_{ij}$ in the weight matrix $\mathbf{W}$ is initialized by the improved "Xavier" method [29], *i.e.*, Gaussian distribution with zero mean and the standard deviation inversely proportional to the number of network's connections. The bias vectors are initialized to zeros. The threshold T in the truncation layer is set to 5. The momentum and weight decay in our networks are set to 0.9 and 0.0001, respectively. A mini-batch with 40 images (20 cover-stego pairs) is used as the input for training. We adaptively adjust the learning rate in the training phase. Specifically, the learning rate is initialized to 0.001 and scheduled to decrease 10% for every 50 training epochs. All of the experiments are conducted on a Tesla K80 GPU.

3.1 Steganographer Detection in Spatial Domain

In this section, we evaluate the proposed method for the steganographer detection task in spatial domain from easy to difficult cases.

3.1.1 Experimental Setting. The standard dataset BOSSbase ver 1.01 is utilized to validate our proposed method. The original BOSSbase ver 1.01 consists of 10,000 grayscale images with the size of 512×512. Following the setting in [16, 20, 25], each image of the original dataset is cropped into four non-overlapping sub-images with the size of 256×256. In the training stage, 20,000 cover images are selected randomly from all cropped images, and their five kinds of stego images are generated by Spatial version of the UNIversal WAvelet Relative Distortion (S-UNIWARD) steganography [30] at five kinds of payloads. The payloads are set from 0.1 to 0.5 bit-per-pixel (bpp) with step 0.1. Thus, these images are constructed 100,000 cover-stego pairs, and they are used for training the multiclass deep neural networks. The remaining 20,000 covers and their corresponding stegos are utilized to evaluate the performance of our model in detecting the guilty user. In our experiments, we randomly selected m users, which include one guilty user and m-1 innocent users, and each of them transfers n images. All the statistical experiments are repeated 100 times, and each time, m is set to 100 and n is set to 200. The average results are reported.

In spatial domain, we use five content-adaptive steganography to embed the messages into the images, *i.e.* S-UNIWARD [30], HIgh-pass Low-pass Low-pass (HILL) [31], Wavelet Obtained Weights (WOW) [32], Highly Undetectable steGO implemented using the Gibbs construction with Bounding Distortion (HUGO-BD) [33], and Minimizing the power of the most POwerful Detector (MiPOD) [34].

We compare our method with the conventional method SRMQ1_SD, and two CNN-based methods, including: ANSD and XuNet_SD. SRMQ1_SD is the steganographer detection method via SRMQ1 [35], which is a well-known spatial rich model with a single quantization step. ANSD is the steganographer detection method via a well-known deep CNN architecture AlexNet [36]. We modify the network slightly. First, the input size of the network is modified as 256×256×3. Second, the number of neurons in the first two fully-connected layers is modified as 1,000. The last fully-connected layer has 2 neurons to classify the covers and stegos. Each image is first filtered by the *KV* filter proposed by Qian *et al.* [16] and the input image of the network is the filtered image. In the training stage, the model is only trained on S-UNIWARD at 0.4bpp. Considering AlexNet is served for the classification of ImageNet Dataset and the number of images of ImageNet is far more than that of BOSSbase. Thus, the size of min-batch is reduced from 256 to 64 (32 cover-stego pairs). To other parameters, we just follow the general setting. XuNet_SD is the abbreviation of the steganographer detection method based on the network proposed by Xu *et al.* [18]. As a matter of convenience, we refer this network as XuNet. It is noticed that the input size of XuNet is modified as 256×256, which is in accord with the input size of our networks. The size of mini-batch is set to 40 (20 cover-stego pairs). In the training phase, the model is only trained on S-UNIWARD at 0.4bpp. In addition, we train the model once and use the learnt model for testing the steganographer detection task instead of training it five times for ensemble learning as described in [18].

3.1.2 A Single Steganographic Algorithm with A single payload. In this subsection, we aim to evaluate our proposed method under a simple condition. In detail, the stego images of the guilty user are generated by one steganographic algorithm (S-UNIWARD) at one payload. The payload is set to 0.05, 0.1, 0.2, 0.3 and 0.4. Here, we employ the single linkage as the clustering linkage function. We compare our proposed method with classical method SRMQ1_SD and two CNN-based methods including ANSD and XuNet_SD. The comparison results are shown in Table 1.

Table 1: The Detection Accuracy Comparisons of Different Methods on S-UNIWARD with a Single Payload.

Payload (bpp)	Method	Feature Dimension	Average Distance		Acc. (%)	STD
			AD1	AD2		
0.05	MDNNSD	320	0.0763	0.0836	6	0.24
	ANSD	1,000	0.0771	0.0807	5	0.22
	XuNet_SD	128	0.0757	0.0762	2	0.14
	SRMQ1_SD	12,753	0.0757	0.0753	0	0
0.1	MDNNSD	320	0.0763	0.1148	84	0.37
	ANSD	1,000	0.0773	0.0988	73	0.45
	XuNet_SD	128	0.0761	0.0801	2	0.14
	SRMQ1_SD	12,753	0.0753	0.0757	0	0
0.2	MDNNSD	320	0.0762	0.2028	100	0
	ANSD	1,000	0.0772	0.1610	100	0
	XuNet_SD	128	0.0753	0.0990	71	0.46
	SRMQ1_SD	12,753	0.0751	0.0748	0	0
0.3	MDNNSD	320	0.0763	0.2743	100	0
	ANSD	1,000	0.0771	0.2285	100	0
	XuNet_SD	128	0.0756	0.1324	100	0
	SRMQ1_SD	12,753	0.0747	0.0734	1	0.1
0.4	MDNNSD	320	0.0764	0.3476	100	0
	ANSD	1,000	0.0773	0.2756	100	0
	XuNet_SD	128	0.0754	0.1718	100	0
	SRMQ1_SD	12,753	0.0748	0.0772	2	0.14

From Table 1, we can find that three CNN-based methods can detect the steganographer accurately when the payload is greater than 0.2. However, the classical method, SRMQ1_SD cannot detect the guilty user in most of cases. For the average distance, AD1 and AD2 is the average distance between the features of innocent users, and the average distance between the features of the guilty user and innocent users, respectively. Obviously, the value of AD1/AD2 is high means that the guilty user is more deviated from innocent users. We can find that the value of AD1/AD2 of SRMQ1_SD is smaller than other CNN-based methods. Besides, we can observe that the feature dimension of SRMQ1_SD is 12,573, which is much larger than other CNN-based methods. Although the high-dimensional rich models with ensemble classifiers are benefit for the steganalysis task, the performance of SRMQ1_SD is not good for steganographer detection task due to the high dimension of the extracted features. Based on these results, we can make the conclusion that CNN-based methods can effectively capture features to distinguish the guilty user from innocent users. In addition, the detection accuracies of all methods decrease with the decrease of the payload. However, our proposed method outperforms other CNN-based methods when the payload is lower than or equal to 0.2. It means our proposed method can detect the guilty user more effectively when the payload is low.

3.1.3 Multiple Steganographic Algorithms with A Single Payload. In this subsection, we try to test the performance of the proposed method when the stego images of the guilty user are generated through data embedded by multiple steganographic algorithms (S-UNWARD, HILL, WOW, HUGO-BD, and MiPOD) with a single payload. That is, for the guilty user, the images are equally divided into five groups. The images of each group include messages embedded by each steganographic algorithm. The payload is set to 0.05, 0.1, 0.2 and 0.3, respectively. Here, we also use the single linkage function. The detection accuracies are displayed in Fig. 3.

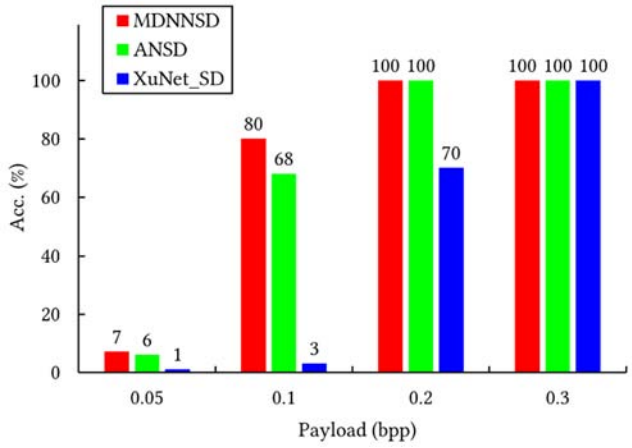

Figure 3: The detection accuracies of MDNNSD, ANSD and XuNet_SD. The models are tested on multiple steganographic algorithms with a single payload.

As Fig. 3 shows, although the condition is more difficult, all of the CNN-based methods can detect the guilty user successfully in most of cases. When the payload is 0.1, the accuracies of our proposed method and ANSD are higher than XuNet_SD. For AlexNet, which is a well-known CNN model for image classification, we can find the features extracted from it can effectively distinguish the stego images and cover images. In spite of this, our proposed method shows a higher level of robustness of inter-steganographic algorithms. However, when the payload is 0.05, all methods cannot detect the guilty user well. As we know, when the embedding payload is low, the stego image is similar to the cover image, making models harder to distinguish the guilty user from innocent users.

3.1.4 The Comparisons of Different Training Strategies. In this subsection, we test the performance of the proposed networks under different training strategies. In detail, the first training strategy is the networks are trained on S-UNIWARD at 0.1bpp, which is named as DNNSD_01. The second is that the networks are trained on S-UNIWARD at 0.4bpp, which is termed as DNNSD_04. Specifically, for these two strategies, the models are trained for binary classification (cover/stego), and the training networks map the input images into two labels. On the contrary, the networks in our proposed method are trained for multiclass classification. All of three models are tested on S-UNIWARD at 0.1 and 0.4bpp, respectively. We compare our proposed method, MDNNSD with these two other training strategies. The performance of the networks under different training strategies is shown in Fig. 4.

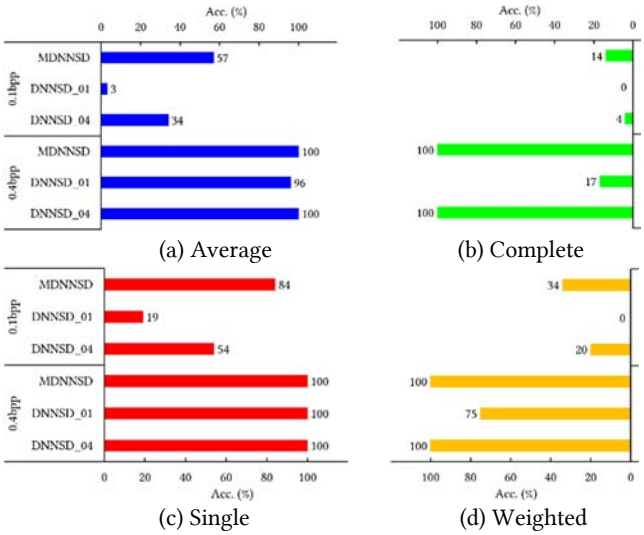

Figure 4: The comparison results of different models on S-UNIWARD at 0.1 and 0.4bpp, respectively. Different colors represent different linkage functions.

Fig. 4 shows that the performance of our proposed method outperforms other strategies. In other words, the model trained for multiclass classification is better than those trained for binary classification even though the payload used in training stage is the same as the testing payload. Moreover, compared

with DNNSD_04 testing on 0.1bpp, the performance of DNNSD_01 tested on 0.1bpp is worse. As we all know, when the payload is low, the stego image is much similar to the cover image. Thus, when DNNSD_01 is trained on the images with low payload, the learnt model cannot capture more useful information from the weak stego signal. For a fair comparison, we also test the performance of the similar model but training for binary classification on the images embedded by S-UNIWARD with six kinds of payloads. We can find the performance of that case is worse than the performance of our proposed method. For instance, under the single linkage function, the detection accuracy of this similar model tested on 0.1bpp is 29%, while the detection accuracy of our proposed model is 84%. Thus, we can make the conclusion that the outstanding performance of our method is mainly dependent on the proposed multiclass networks rather than more samples from different payloads.

3.2 Steganographer Detection in Frequency Domain

In Section 3.1, we have evaluated the effectiveness of our proposed method for the steganographer detection task in spatial domain. In this section, we extend the task to frequency domain, and devote to evaluate whether our proposed method can identify the guilty user in frequency domain or not.

3.2.1 Experimental Setting. Consistent with the setting in Section 3.1.1, each image from the original BOSSbase ver 1.01 is cropped into four non-overlapping sub-images. Then, each sub-image is compressed with JPEG quality factor 80 using Matlab's imwrite function. All of these cropped JPEG images compose the JPEG version of BOSSbase ver 1.01 dataset, which is utilized to validate the performance of our method in frequency domain. We also randomly select 20,000 images from the cropped JPEG images and their corresponding stego images generated by JPEG version of the UNIversal WAvelet Relative Distortion (J-UNIWARD) [30] at five kinds of payloads to train the model. The payloads are set from 0.1 to 0.5 bits per non-zeros Alternating Current DCT coefficient (bpnzAC) with step 0.1. The rest 20,000 covers and their corresponding stegos are used to validate our proposed method in steganographer detection task. Keeping the same setting in spatial-domain experiments, m users are randomly selected and each of them transfers n images. All the statistical experiments are repeated for 100 times, and each time, m is set to 100 and n is set to 200. The average accuracy is used as basis for evaluation.

In frequency domain, we use two JPEG-domain steganographic algorithms to embed messages, including: J-UNIWARD [30], and no-shrinkage F5 (nsF5) [37]. J-UNIWARD is a kind of content-adaptive steganographic algorithm, which is more difficult than nsF5. We compare our proposed method with the classical method PEV_SD, and the CNN-based method ANSD. PEV_SD is the steganographer detection framework based on the PEV-274 features. The parameters setting of ANSD is identical to that of ANSD in spatial domain. Similar to what has been done in JPEG-domain steganalytic methods [23, 38], each image in JPEG format is first decompressed into the spatial domain. Then, the decompressed JPEG images are used as the input of all models.

3.2.2 A Single Steganographic Algorithm with A single payload. In the first experiment in frequency domain, we try to compare our proposed method with two other methods under a simple condition. For the guilty user, two hundred stego images include message embedded by a single steganographic algorithm with a single payload. We test the performance of all models in four cases, including: J-UNIWARD at 0.1bpnzAC, J-UNIWARD at 0.4bpnzAC, nsF5 at 0.1bpnzAC, and nsF5 at 0.4bpnzAC, respectively. Here, we use the single linkage as the linkage function in the clustering algorithm. The experiment results are provided in Table 2.

Table 2: The Detection Accuracy Comparisons of Different Methods Tested on a Single Steganographic Algorithm with a Single Payload.

Case	Method	Feature Dimension	Acc. (%)	STD
J-UNIWARD at 0.1bpnzAC	MDNNSD	320	58	0.5
	ANSD	1,000	1	0.1
	PEV_SD	274	0	0
J-UNIWARD at 0.4bpnzAC	MDNNSD	320	100	0
	ANSD	1,000	100	0
	PEV_SD	274	4	0.2
nsF5 at 0.1bpnzAC	MDNNSD	320	65	0.48
	ANSD	1,000	8	0.27
	PEV_SD	274	2	0.14
nsF5 at 0.4bpnzAC	MDNNSD	320	100	0
	ANSD	1,000	100	0
	PEV_SD	274	90	0.3

From Table 2, we can observe that the classical method PEV_SD achieves good performance in the case of nsF5 at 0.4bpnzAC due to the lower dimension of features. However, in the case of J-UNIWARD at 0.4bpnzAC, the accuracy of PEV_SD is low. It is owing to J-UNIWARD is a content-adaptive steganography, which is more difficult than nsF5. For the CNN-based methods, both of them can achieve good performance when the payload is high. This illustrates the CNN-based method is effective for steganographer detection task in frequency domain. However, when the payload is low, *i.e.* 0.1bpnzAC, ANSD and PEV_SD are inferior to our proposed method. In other words, our proposed method can effectively solve the steganographer detection problem in spatial and frequency domain under low payload condition. In addition, compared with the performance of our proposed method in spatial domain, the detection accuracies of that in frequency domain are lower. It is due to these images loss detail information in the compression and decompression procedure, and this detail information are used to hide the weak stego signal.

4 CONCLUSIONS

In this paper, we propose a novel steganographer detection method based on multiclass deep neural networks. To our best

knowledge, we are the first to propose a steganographer detection framework based on multiclass neural networks. In the training stage, the proposed multiclass neural networks are trained to classify the images with six types of embedding payloads. In the inference stage, the learnt model is served as the feature extractor, which utilizes to extract the features of each image from each user. Finally, the agglomerative hierarchical clustering algorithm is utilized to identify the steganographer based on the MMD distance metric.

In the experiments, we evaluate the effectiveness of the proposed method on two standard datasets in different image domain. In spatial domain, the proposed method can accurately identify the guilty user from easy to difficult conditions, especially the low embedding payloads. Moreover, the proposed method demonstrates its robustness under inter-steganographic algorithms situations. In frequency domain, the performance of the proposed method outperforms other methods when the payload is low. In future work, we seek to apply our proposed method under the real-world large-scale social media networks.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 61502311, No. 61620106008), the Natural Science Foundation of Guangdong Province (No. 2016A030310053, 2016A030310039, 2017A030310521), the Science and Technology Innovation Commission of Shenzhen under Grant (No.JCYJ20160422151736824), Shenzhen Emerging Industries of the Strategic Basic Research Project under Grant (No. JCYJ20160226191842793), the Shenzhen high-level overseas talents program, and the Tencent "Rhinoceros Birds" - Scientific Research Foundation for Young Teachers of Shenzhen University (2016).

REFERENCES

[1] R. J. Anderson and F. A. P. Petitcolas. 1998. On the limits of steganography. *IEEE Journal on Selected Areas in Communications*. 16, 4 (May 1998), 474–481. DOI: 10.1109/49.668971

[2] T. Lin, C. Wang, W. Chen, F. Lin and W. Lin. 2017. A novel data hiding algorithm for high dynamic range images. *IEEE Transactions on Multimedia*. 19, 1 (Jan. 2017), 196-211. DOI: 10.1109/TMM.2016.2605499

[3] B. Li, J. He, J. Huang, and Y. Q. Shi. 2011. A survey on image steganography and steganalysis. *Journal of Information Hiding and Multimedia Signal Processing*. 2, 2 (Apr. 2011), 142-172.

[4] F. Li, K. Wu, J. Lei, M. Wen, Z. Bi, and C. Gu. 2016. Steganalysis over large-scale social networks with high-order joint features and clustering ensembles. *IEEE Transactions on Information Forensics and Security*. 11, 2 (Feb. 2016), 344-357. DOI: 10.1109/TIFS.2015.2496910

[5] A. D. Ker and T. Pevný. 2011. A new paradigm for steganalysis via clustering. In *Proceeding of SPIE, Media Watermarking., Security, and Forensics III.* San Francisco Airport, California, U.S. 78800U1–U13.

[6] T. Pevný and J. Fridrich. 2007. Merging Markov and DCT features for multi-class JPEG steganalysis. In *Proceeding of SPIE, Security, Steganography, and Watermarking of Multimedia Contents IX.* San Jose, CA, U.S. 650503–14.

[7] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schöölkopf, and A. J. Smola. 2007. A kernel method for the two-sample-problem. In *Proceeding of Advances in Neural Information Processing Systems (NIPS)*. Cambridge, MA, USA, 513–520.

[8] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. 2000. LOF: Identifying density-based local outliers. In *Proceeding of ACM SIGMOD International Conference on Management of Data (ACM SIGMOD'00)*. ACM New York, NY, USA. 93–104.

[9] A. D. Ker and T. Pevný. 2012. Identifying a steganographer in realistic and heterogeneous data sets. In *Proceeding of SPIE, Media Watermarking., Security, and Forensics 2012.* Burlingame, California. U.S. 83030N1–N13.

[10] A. D. Ker and T. Pevný. 2014. The steganographer is the outlier: Realistic large-scale steganalysis. *IEEE Transactions on Information Forensics and Security*. 9, 9 (Sep. 2014), 1424–1435.

[11] F. Li, M. Wen, J. Lei, and Y. Ren. 2017. Efficient steganographer detection over social networks with sampling reconstruction. *Peer-to-Peer Networking and Applications*. (Sep. 2017), 1-16. DOI: https://doi.org/10.1007/s12083-017-0603-3

[12] O. Seddati, S. Dupont, and S. Mahmoudi. 2017. Quadruplet Networks for Sketch-Based Image Retrieval. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR '17)*. ACM, New York, NY, USA, 184-191.

[13] Z. Wang, K. He, Y. Fu, R. Feng, Y. Jiang, and X. Xue. 2017. Multi-task Deep Neural Network for Joint Face Recognition and Facial Attribute Prediction. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR '17)*. ACM, New York, NY, USA, 365-374.

[14] J. Gao, Y. Fu, Y. Jiang, and X. Xue. 2017. Frame-Transformer Emotion Classification Network. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR '17)*. ACM, New York, NY, USA, 78-83.

[15] S. Tan and B. Li. 2014. Stacked convolutional auto-encoders for steganalysis of digital images. In *Proceeding of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. Siem Reap. 1-4.

[16] Y. Qian, J. Dong, W. Wang, and T. Tan. 2015. Deep learning for steganalysis via convolutional neural networks. In *Proceeding of SPIE, Media Watermarking. Security, and Forensics*, San Francisco, California, U.S. 9409J1–J10.

[17] Y. Qian, J. Dong, W. Wang, and T. Tan. 2016. Learning and transferring representations for image steganalysis using convolutional neural network. In *Proceeding of IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ. 2752-2756.

[18] G. Xu, H. Z. Wu, and Y. Q. Shi. 2016. Structural design of convolutional neural networks for steganalysis. *IEEE Signal Processing Letters*. 23, 5 (May 2016), 708-712. DOI: 10.1109/LSP.2016.2548421

[19] G. Xu, H. Z. Wu, and Y. Q. Shi. 2016. Ensemble of CNNs for steganalysis: An empirical study. In *Proceeding of ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'16)*. ACM New York, NY, USA, 103-107.

[20] S. Wu, S. Zhong, and Y. Liu. 2017. Deep residual learning for image steganalysis. *Multimedia Tools and Applications*. (2017), 1-17. DOI: https://doi.org/10.1007/s11042-017-4440-4

[21] J. Ye, J. Ni, and Y. Yi. 2017. Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security*. 12, 11 (Nov. 2017), 2545-2557. DOI: 10.1109/TIFS.2017.2710946

[22] J. Zeng, S. Tan, B. Li, and J. Huang. 2016. Large-scale JPEG image steganalysis using hybrid deep-learning framework. arXiv: 1611.03233. Retrieved from https://arxiv.org/abs/1611.03233

[23] G. Xu. 2017. Deep convolutional neural network to detect J-UNIWARD. In *Proceeding of ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'17)*. ACM New York, NY, USA, 103-107.

[24] J. Yang, Y. Shi, E. K. Wong, and X. Kang. 2017. JPEG steganalysis based on denseNet. arXiv: 1711.09335. Retrieved from https://arxiv.org/abs/1711.09335

[25] M. Zheng, S. Zhong, S. Wu, and J. Jiang. 2017. Steganographer detection via deep residual network. In *Proceeding of IEEE International Conference on Multimedia and Expo (ICME'17)*. Hong Kong, 235-240.

[26] J. Kodovský, V. Sedighi, and J. Fridrich. 2014. Study of cover source mismatch in steganalysis and ways to mitigate its impact. In *Proceeding of SPIE, Media Watermarking. Security, and Forensics*. San Francisco, California, U. S .90280J1–J12.

[27] K. He, X. Zhang, S. Ren, J. Sun. 2016. Deep residual learning for image recognition. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA. 770-778.

[28] P. Bas, T. Filler, and T. Pevný. 2011. "Break our steganographic system": The ins and outs of organizing BOSS. In *Proceeding of the 13th International Workshop on Information Hiding (IH)*. Springer, Berlin, Heidelberg, 59–70.

[29] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Delving deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceeding of IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile. 1026-1034.

[30] V. Holub and J. Fridrich. 2014. Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*. 2014, 1 (Dec. 2014), 1–13. DOI: https://doi.org/10.1186/1687-417X-2014-1

[31] B. Li, M. Wang, J. Huang, and X. Li. 2014. A new cost function for spatial image steganography. In *Proceeding of IEEE International Conference on Image Processing (ICIP)*. Paris, France, 4206-4210.

[32] V. Holub and J. Fridrich. 2012. Designing steganographic distortion using directional filters. In *Proceeding of IEEE International Workshop on Information Forensics and Security (WIFS)*. Tenerife, 234-239.

[33] T. Filler and J. Fridrich. 2010. Gibbs construction in steganography. *IEEE Transactions on Information Forensics and Security*. 5, 4 (Dec. 2010), 705-720. DOI: 10.1109/TIFS.2010.2077629

[34] V. Sedighi, R. Cogranne, and J. Fridrich. 2016. Content-Adaptive steganography by minimizing statistical detectability. *IEEE Transactions on Information Forensics and Security*. 11, 2 (Feb. 2016), 221-234. DOI: 10.1109/TIFS.2015.2486744

[35] J. Fridrich and J. Kodovsky. 2012. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*. 7, 3 (Jun. 2012), 868-882. DOI: 10.1109/TIFS.2012.2190402M.

[36] A. Krizhevsky, I. Sutskever, and G. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceeding of Advances in Neural Information Processing Systems (NIPS)*. 1097-1105.

[37] J. Fridrich, T. Pevný, and J. Kodovský. 2007. Statistically undetectable JPEG steganography: Dead ends challenges, and opportunities. In *Proceeding of ACM Workshop on Multimedia & security (MM&Sec'07)*. ACM New York, NY, USA. 20-21.

[38] V. Holub and J. Fridrich. 2015. Low-complexity features for JPEG steganalysis using undecimated DCT. *IEEE Transactions on Information Forensics and Security*. 10, 2 (Feb. 2015), 219-228. DOI: 10.1109/TIFS.2014.2364918