

RESEARCH

Open Access



Joint multi-domain feature learning for image steganalysis based on CNN

Ze Wang^{1,2}, Mingzhi Chen³, Yu Yang^{1,2*} , Min Lei^{1,2} and Zhexuan Dong⁴

*Correspondence:

yangyu@bupt.edu.cn

¹State Key Laboratory of Public Big Data, Guizhou University, 550025 Guizhou Guiyang, China

²Laboratory of Cyberspace Security, Beijing University of Posts and Telecommunications, 100876 Beijing, China

Full list of author information is available at the end of the article

Abstract

In recent years, researchers have been making great progress in the steganalysis technology based on convolution neural networks (CNN). However, experts ignore the contribution of nonlinear residual and joint domain detection to steganalysis, and how to detect the adaptive steganographic algorithms with low embedding rates is still challenging. In this paper, we propose a CNN steganalysis model that uses a joint domain detection mechanism and a nonlinear detection mechanism. For the nonlinear detection mechanism, based on the spatial rich model (SRM), we introduce the maximum and minimum nonlinear residual feature acquisition method into the model to adapt to the nonlinear distribution of steganography information. For the joint domain detection mechanism, we not only apply the high-pass filters from the SRM for spatial residuals, but also apply the patterns from the discrete cosine transform residual (DCTR) for transformation steganographic impacts, so as to fully capture the interference trace of spatial steganography to transform domain. We also apply a new transfer learning method to improve the model's performance. That is, we apply the low embedding rate steganography samples to initialize the model, because we think that the method makes the network more sensitive than applying high embedding rate steganography samples to initialize the model. The simulation results also confirm this assumption. Combined with the above improved methods, the detection accuracy of the model for WOW and S-UNIWARD is higher than that of SRM+EC, Ye-Net, Xu-Net, Yedroudj-Net and Zhu-Net, which is about 4~6% higher than that of the optimal Zhu-Net. The results can provide a certain reference for steganalysis and image forensics tasks.

Keywords: Image steganalysis, Convolutional neural networks, Feature learning, Joint domain, Nonlinear detection

1 Introduction

Steganalysis [1] and information hiding are mutually restricted and mutually promoted [2, 3]. And there is a more hopeful prospect to carry out the steganalysis work. Image steganography is a technique of hiding secret messages in images. In the transformation domain, images are converted by discrete cosine transform (DCT) [4], discrete wavelet transform (DWT) [5], and so on. And the secret messages are embedded in the transformation coefficients. The popular algorithms are J-UNIWARD [6], nsF5 [7], UED [8],

and UERD [9]. In the spatial domain, steganography algorithms are characterized by directly changing the pixels. The typical algorithms are the least significant bit (LSB) [10, 11], LSB matching [12], and pixel value differencing (PVD) [13]. There are also some steganography algorithms in the compression domain [14]. Those algorithms above can be regarded as the non-adaptive steganography algorithms. Compared with non-adaptive steganography algorithms, the adaptive steganography algorithms have been proved to have better performance. At present, the popular adaptive algorithms are edge adaptive image steganography (EA) [15], HUGO [16], HILL [17], MiPOD [18], and S-UNIWARD [6]. And since the security of steganographic algorithms keeps increasing [19–21], attempts to detect such data hiding methods encounter more challenges.

A traditional steganalysis method is often based on a manual designed feature. The most often adopted features include a gray level co-occurrence matrix (GLCM) [22], local binary patterns (LBP) [23], and Gaussian Markov random field. In addition, some perceptual hashing techniques can be applied for steganalysis tasks. The perceptual hash technology [24, 25] can be used to extract information closely related to human perception of image visual quality, so these perceptual description models can also detect the tampering trace of steganography. Since the spatial rich model (SRM) [26] algorithm was proposed, a lot of research has been focused on SRM algorithm, and many improved algorithms have been proposed. Although these algorithms have made performance improvements, they fail to solve the key shortcomings of the feature extraction method. The traditional feature extraction method relies on the characteristics of manual design, the design process depends on the expert experience, and the heuristic method is usually applied. It means that this kind of steganalysis algorithm is difficult to deal with the challenge brought by the rapid development of steganalysis algorithm.

Deep learning technology can effectively solve the problems caused by manual feature design and is widely used in the field of image perception [27, 28] and steganalysis. Deep learning technology can automatically recognize and extract features through deep network, which makes steganalysis technology possible to get rid of the dependence on expert experience. With the development of graphics processing unit (GPU) and parallel computing technology, this process has been accelerated. In 2014, Tan and Li [29] proposed the first steganalysis model that applied deep learning techniques. In 2015, Qian et al. [30] proposed the first convolution neural network (CNN) model using supervised learning methods, whose steganalysis performance surpasses SRM. In 2016, Xu et al. [31] proposed a CNN model similar to Qian's model, the so-called Xu-Net. The difference is that an absolute value layer (ABS) and a 1×1 convolution kernel are employed in the Xu-Net. Recently, Qian et al. [32] brought forward a creative concept, called the transfer learning, to improve steganalysis performance. The above models only capture spatial steganography features, so they are used to detect spatial steganography algorithm. Until 2017, the research results of detection for steganography algorithm in transformation domain gradually appeared. Zeng et al. [33, 34] proposed a JPEG-based steganalysis model. Xu et al. [35], inspired by ResNet [36], proposed a new CNN steganalysis model consisting of 20 convolutional layers with batch normalization (BN). Ye et al. [37] proposed a spatial domain CNN steganalysis model, and they added a truncated linear unit (TLU) activation function to the preprocessing layer. The main trend in 2017 was to optimize the convolution neural network architecture through ResNet

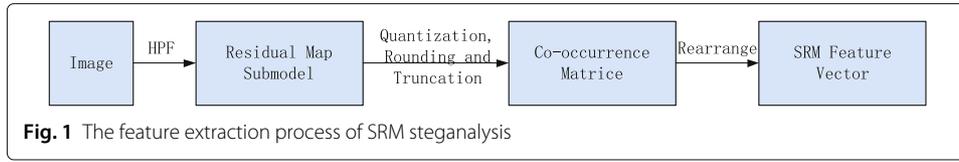
and draw on the feature extraction method of SRM. In 2018, Yedroudj et al. [38] proposed a spatial domain CNN steganalysis model consisting of five convolutional layers. In addition to the traditional image datasets BOSSBass [39], they added the BOWS2 [40] image datasets. Tsang et al. [41] improved Ye-Net, which made the model perform steganalysis on high-resolution images. Zhang et al. [42] proposed a new CNN steganalysis model, and they used the depth separable convolution network and spatial pyramid pooling (SPP) to obtain the channel correlation and adapt to different sizes of images. Deep steganalysis technology has made remarkable progress, but there is still much room for improvement. The existing deep steganalysis technology adopts single domain mode, that is, only spatial features are captured when detecting spatial steganography, and the same is true when detecting transformation domain steganography. However, steganography in transformation domain will destroy the spatial characteristics of image, and vice versa. So, the joint domain detection can better capture the trace of steganography.

In this paper, we propose a novel spatial domain steganalysis model called Wang-Net. It has the following characteristics:

- (1) A joint domain detection concept is brought forward. Joint domain detection is to capture steganography features in both spatial and transformation domain to complete steganography detection task. At present, the typical spatial and transformation domain steganalysis models are Zhu-Net and Xu-Net, which only extract single domain features. However, the steganography of one domain will affect other domains, so joint domain detection method can capture more comprehensive steganography information. We simulate SRM and discrete cosine transform residual (DCTR) feature extraction methods to detect steganography feature in both spatial and transformation domain.
- (2) The nonlinear feature detection mechanism is introduced. The nonlinear detection mechanism is to capture the steganographic features through nonlinear transformation. At present, the famous Ye-Net and Zhu-Net all simulate the linear feature extraction method of SRM to complete the steganalysis task. However, the embedding of steganography information is nonlinear, so it is necessary to introduce nonlinear detection mechanism. We simulate the nonlinear feature extraction of SRM to complete the design and implementation of the nonlinear detection mechanism.
- (3) A new transfer learning method is applied. For Zhu-Net and other steganalysis models using the transfer learning method, the authors use high embedding rate samples to initialize the model, in order to solve the problem that the model in the training stage is difficult to converge to the low embedding rate samples. Compared with high embedding rate samples, low embedding rate samples have less steganography information with the same steganography mode. For the transfer learning method in this paper, the low embedding rate samples are used to initialize the model, in order to enhance the sensitivity of the model to steganography information.

2 Preliminaries

We mainly extract feature information through high-pass filters (HPFs) from SRM and DCT patterns from DCTR. The contents of SRM and DCTR are as follows.



2.1 SRM

The feature extraction method of SRM steganalysis can be seen in Fig. 1. Firstly, the residual map sub-models are obtained by the high-pass filter, then the fourth-order co-occurrence matrix of each residual map sub-model is extracted by quantization, rounding, and truncation. Finally, the elements of these co-occurrence matrices are rearranged to form the steganalysis feature vector.

Scholars design various HPFs in SRM and use them to generate residual map sub-models. The original linear residual calculation formula is as follows.

$$R_{mn} = pred(N_{mn}) - cI_{mn} \quad (1)$$

where c is called the residual order, m and n represent the pixel coordinates, N_{mn} is the adjacent pixel of image I_{mn} , $pred(N_{mn})$ is the predictor of cI_{mn} , and R_{mn} is the residual of image I_{mn} . Generally, the number of pixels of N_{mn} is equal to c .

The residuals mainly include first-order, second-order, third-order, SQUARE, EDGE3x3, and EDGE5x5 six types, and each type of residuals is divided into linear filtering residuals and nonlinear filtering residuals. The typical residuals and high-pass filters can be seen in Table 1, Eqs.(2), (3), and (4). As shown in Eq.(2), the residuals in the horizontal, vertical, diagonal, and anti-angular directions are denoted as R^h , R^v , R^d , and R^m . The max nonlinear filtering residual is denoted as R_{max} , the min nonlinear filtering residual is denoted as R_{min} . As shown in Eq.(3), the left side is the SQUARE3x3 high-pass filter, and the right side is the SQUARE5x5 high-pass filter. As shown in Eq.(4), the left and right sides are the EDGE3x3 and EDGE5x5 high-pass filter respectively.

$$\begin{aligned} R_{ij}^{min} &= \min(R_{ij}^h, R_{ij}^v, R_{ij}^d, R_{ij}^m), \\ R_{ij}^{max} &= \max(R_{ij}^h, R_{ij}^v, R_{ij}^d, R_{ij}^m) \end{aligned} \quad (2)$$

$$\begin{bmatrix} -1 & 2 & -1 \\ 2 & -4 & 2 \\ -1 & 2 & -1 \end{bmatrix}, \begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix} \quad (3)$$

Table 1 First-order, second-order, and third-order linear residuals in the horizontal direction

Residual type	HPF	Linear residual
First-order	(1, -1)	$R_{ij}^h = y_{ij+1} - y_{ij}$
Second-order	(1, -2, 1)	$R_{ij}^h = y_{ij-1} - 2y_{ij} + y_{ij+1}$
Third-order	(1, -3, 3, -1)	$R_{ij}^h = y_{ij-1} - 3y_{ij} + 3y_{ij+1} - y_{ij+2}$

$$\begin{bmatrix} 2 & -1 \\ -4 & 2 \\ 2 & -1 \end{bmatrix}, \begin{bmatrix} -2 & 2 & -1 \\ 8 & -6 & 2 \\ -12 & 8 & -2 \\ 8 & -6 & 2 \\ -2 & 2 & -1 \end{bmatrix} \quad (4)$$

For the linear residual, Table 1 has given the calculation method of the first-order, second-order, and third-order linear residuals. It is not difficult to find that the linear residuals of SQUARE, EDGE3x3, and EDGE5x5 only apply more directional neighborhood pixels in the calculation. The SQUARE, EDGE3x3, and EDGE5x5 high-pass filters are shown in Eqs. (3) and (4). In fact, the linear residual calculation method can be converted to the convolution operation:

$$R = I * K = (R_{ij}) = \left(\sum_{r,c} x_{i,j}^{r,c} k^{r,c} \right) \quad (5)$$

where (i, j) is the pixel coordinates, and (r, c) is the index of the convolution kernel. $x_{i,j}^{r,c}$ denotes the pixel value of the fixed neighborhood window index (r, c) of the central pixel (i, j) . $k^{r,c}$ denotes the value of index (r, c) of the convolution kernel, which is the same size as the fixed neighborhood window. R_{ij} denotes the result of convolution operation for pixels (i, j) . I and K denotes the image and convolution kernel respectively. R denotes the residual for the whole image. $*$ denotes the convolution operation.

As shown in Eq.(2), the nonlinear residual can be obtained by finding the maximum or minimum of some linear filtering residuals. As shown in Table 1, we take the first-order linear residual as the residual prototype; there are totally eight first-order linear residuals:

$$\begin{aligned} R_{ij} = \{ & y_{i,j+1} - y_{ij}, y_{i+1,j+1} - y_{ij}, y_{i+1,j-1} - y_{ij}, \\ & y_{i,j+1} - y_{ij}, y_{i,j-1} - y_{ij}, y_{i-1,j+1} - y_{ij}, \\ & y_{i-1,j} - y_{ij}, y_{i-1,j-1} - y_{ij} \} \end{aligned} \quad (6)$$

Then, the first-order nonlinear residual is:

$$R_{ij}^{min} = \min\{R_{ij}\} \quad (7)$$

$$R_{ij}^{max} = \max\{R_{ij}\} \quad (8)$$

The nonlinear residuals combine the statistical characteristics of the same kind of linear residuals, which fully reflect the adjacent pixels changes in image caused by the steganography.

2.2 DCTR

In the transformation domain, the DCTR [43] is a general steganalysis algorithm. The steps of its feature processing are as follows:

- (1) Obtain $64 \times 8 \times 8$ DCT bases patterns by calculation, then obtain feature maps by convoluting the decompressed JPEG image with the DCT basis patterns.
- (2) Obtain the sub-feature maps by quantifying and truncating the raw feature maps.
- (3) Compress the sub-feature maps into an 8000-dimensional feature vector.

In the above steps, the DCT basic patterns are 8×8 matrices, $B^{(i,j)} = (B_{mn}^{(i,j)})$, $0 \leq m, n \leq 7$, and $B_{mn}^{(i,j)}$ is calculated as follows:

$$B_{mn}^{(i,j)} = \frac{u_i u_j}{4} \cos \frac{\pi i(2m+1)}{16} \cos \frac{\pi j(2n+1)}{16} \tag{9}$$

where $u_0 = \frac{1}{\sqrt{2}}$, $u_k = 1$ for $k > 0$, (i, j) is the pixel coordinates.

DCT is defined as the convolution operation of the image and 64 DCT basic patterns $B^{(i,j)}$. In order to understand DCT better, we set the length and width of all images to a multiple of 8. When given a grayscale image $I \in R^{M \times N}$ of size $M \times N$ (M, N is a multiple of 8):

$$U(I) = \{U^{(i,j)} | 0 \leq i, j \leq 7\} \tag{10}$$

$$U^{i,j} = I * B^{i,j} \tag{11}$$

where $U^{(i,j)} \in R^{(M-7) \times (N-7)}$, $*$ denotes a non-padded convolution operation.

3 The proposed method

As shown in Fig. 2, our model consists of preprocessing layer, feature extraction layer, and classification layer. For the preprocessing layer, we simulate the SRM feature extraction method in the spatial domain and simulate the DCTR feature extraction method in the transformation domain and added the nonlinear residual features extraction method. For the general feature extraction stage, we design eight different convolution layers, together with the fully connected layer as the tenth layer for steganographic detection. Nonlinear feature extraction method, joint domain detection mechanism, and detailed designs are introduced in the following sections.

3.1 Nonlinear feature extraction method

At present, the linear feature extraction method of steganalysis model cannot perfectly adapt to the nonlinear embedding state of steganalysis information, so we design a nonlinear feature extraction method.

For the linear feature extraction method, like Zhu-Net, we apply six types of SRM HPFs. All HPFs of the same type are composed of their basic “spams” filters and rotation variants, so as to capture multi-directional and comprehensive residual information in the same neighborhood. The pixel residual information captured by different types of HPFs has different statistical characteristics. And compared with lower-order HPFs,

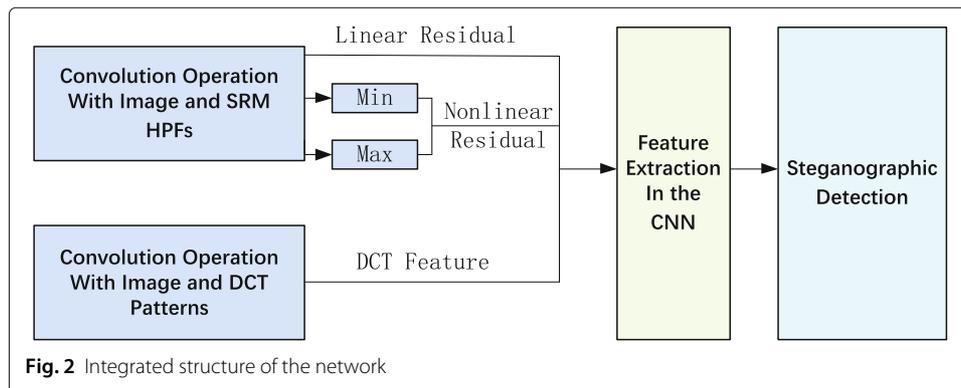


Fig. 2 Integrated structure of the network

higher-order HPFs can capture pixel residual information of larger neighborhood. These six types of HPFs contain the first-order, second-order, third-order, SQUARE, EDGE3x3, and EDGE5x5, and the number of filters are 8, 4, 8, 2, 4, and 4, respectively. We obtain 30 linear residual feature maps through these 30 high-pass filters.

For the nonlinear feature extraction method, we use SRM's nonlinear feature statistics method to capture the nonlinear residual feature map from these six types of HPFs. Specifically, SQUARE is divided into SQUARE3x3 and SQUARE5x5. SQUARE3x3, and EDGE3x3 belong to the same category, so there are two nonlinear residual feature maps in SQUARE3x3 and EDGE3x3. And there are also two nonlinear residual feature maps in SQUARE5x5 and EDGE5x5. Finally, we obtain a total of 10 nonlinear residual feature maps by statistics.

After simulating the linear and nonlinear feature extraction methods, we design two networks, called the single linear residual feature net (Linear Kernel-Net) and nonlinear residual feature net (Non-linear Kernel-Net), and carry out the steganography detection.

According to Table 2, for WOW (0.2 bpp), WOW (0.4 bpp), S-UNIWARD (0.2 bpp), and S-UNIWARD (0.4 bpp), the accuracy of Non-linear Kernel-Net are 0.697, 0.788, 0.609, and 0.734 respectively, and the accuracy of Linear Kernel-Net are 0.710, 0.819, 0.661, and 0.766 respectively. The accuracy of Non-linear Kernel-Net is about 2~6% lower than that of Linear Kernel-Net. For nonlinear residual features, we roughly calculate the maximum and minimum values of each type of linear residual features and do not consider the distribution characteristics of residual feature values. Therefore, the adjacent pixel changes in the image caused by steganography are not comprehensively reflected, that is, the advantage of the nonlinear residual feature is not fully utilized. However, the accuracy of Non-linear Kernel-Net is higher than that of CNN steganalysis Network Ye-Net, indicating that the Non-linear Kernel-Net still has good competition and can enhance the feature representation. Therefore, we add linear and nonlinear residual features to our network named All Kernel-Net to continue the steganalysis.

According to the information in Table 3, for WOW (0.2 bpp), WOW (0.4 bpp), S-UNIWARD (0.2 bpp), and S-UNIWARD (0.4 bpp), the accuracy of All Kernel-Net is 0.714, 0.844, 0.669, and 0.792, which is about 0.4~6% higher than that of Linear Kernel-Net and Non-linear Kernel-Net. All Kernel-Net combines the advantages of linear and nonlinear residual features and has a great steganalysis effect.

3.2 Joint domain detection mechanism

At present, Zhu-Net and other steganalysis models only capture the steganalysis features from a single domain, without considering the impact of steganalysis on other domains. And the steganography features captured by these models have the defect of singleness.

Table 2 The performance of Linear Kernel-Net, Non-linear Kernel-Net, and Ye-Net on resampled images

Algorithms	Linear Kernel-Net	Non-linear Kernel-Net	Ye-Net
WOW (0.2 bpp)	0.710	0.697	0.669
WOW (0.4 bpp)	0.819	0.788	0.768
S-UNIWARD (0.2 bpp)	0.661	0.609	0.600
S-UNIWARD (0.4 bpp)	0.766	0.734	0.688

The involved networks are trained and tested on BOSSBase

Table 3 The performance of Linear Kernel-Net, Non-linear Kernel-Net, and All Kernel-Net on resampled images

Algorithms	Linear Kernel-Net	Non-linear Kernel-Net	All Kernel-Net
WOW (0.2 bpp)	0.710	0.697	0.714
WOW (0.4 bpp)	0.819	0.788	0.844
S-UNIWARD (0.2 bpp)	0.661	0.609	0.669
S-UNIWARD (0.4 bpp)	0.766	0.734	0.792

The involved networks are trained and tested on BOSSBase

Therefore, we propose a joint domain detection mechanism based on All Kernel-Net and simulate the feature extraction method DCTR in the transformation domain.

Small convolution kernels can effectively reduce the parameters scale, and matrix operations can take full advantage of the parallel computing. Therefore, referring to Zhu-Net, we design the convolution kernel as a matrix with 94 channels and 5×5 size, which was initialized with DCT patterns and HPFs. At this point, the calculation formula for the new DCT patterns is as follows:

$$B_{mn}^{(i,j)} = \frac{u_i u_j}{5} \cos \frac{\pi i(2m+1)}{10} \cos \frac{\pi j(2n+1)}{10} \quad (12)$$

where $u_0 = 1$, $u_k = \sqrt{2}$ for $k > 0$, $0 \leq m, n \leq 4$, $0 \leq i, j \leq 7$.

In this way, we add the matrix initialized by DCT patterns and HPFs to the preprocessing layer. The network is called Wang-Net that combines the advantages of linear and nonlinear feature extraction in the spatial and transformation domains. We exert the steganalysis simulation.

According to the results in Table 4, the steganalysis accuracy of Wang-Net for WOW (0.2), WOW (0.4), S-UNIWARD (0.2), S-UNIWARD (0.4) are 0.749, 0.860, 0.691, and 0.819 respectively, which is about 2~3% higher than that of All Kernel-Net. It strongly shows that the joint domain detection mechanism can force the model to learn richer and more comprehensive steganography features and achieve better steganography detection performance.

3.3 Detailed design in network architecture

Our network receives an image of 256×256 size and outputs two types of labels. Wang-Net consists of 10 network layers, including a preprocessing layer, eight convolutional layers for feature extraction, and a fully connected layer for result classification. For preprocessing layer, we apply a convolution kernel with the channel number of 94 and the size of 5×5 , which is initialized by the SRM filters and DCT patterns. For the feature extraction process, 3×3 convolution kernels are applied in the layers 2, 3, 4, 8, and 9, and

Table 4 The performance of Wang-Net and All Kernel-Net on resampled images

Algorithms	All Kernel-Net	Wang-Net
WOW (0.2 bpp)	0.714	0.749
WOW (0.4 bpp)	0.844	0.860
S-UNIWARD (0.2 bpp)	0.669	0.691
S-UNIWARD (0.4 bpp)	0.792	0.819

The involved networks are trained and tested on BOSSBase

5×5 convolution kernels are applied in the layers 5, 6, and 7. In each convolution layer, we add BN, rectified linear unit (ReLU), and TLU nonlinear activation functions. And we also add average pooling to the 4, 5, and 6 convolutional layers.

4 Simulation configuration

We applied two well-known content adaptive steganography algorithms to evaluate the performance of the CNN models, which are WOW and S-UNIWARD. And we use a randomly embedded key when applying the steganography algorithm, which is also in line with the actual steganography situation. The datasets applied in the simulations is the BOSS-Base 1.01. BOSSBase 1.01 contains 10,000 512×512 natural grayscale cover images taken directly from the camera, which have different texture features and are widely used in steganalysis. Due to the limitations of GPU computing resources, in the simulation, like Zhu-Net, we scale the images of BOSSBase 1.01 to 256×256 (using “imresize()” in matlab, the function parameter remains the default configuration). In the simulation, we apply the steganography algorithm and cover images to generate 10,000 corresponding stego images. In order to prevent overfitting, we need to allocate as much data as possible in the training process for our complex model with strong learning ability. Therefore, the data ratio of training datasets, verification datasets, and test datasets is 8:1:1. We apply the AdaDelta [44] to train the network model, which accelerates the convergence of the model. Due to GPU memory limitations, we set the mini-batch size to 16. We apply an exponential decay method with a decay rate of 0.95, a decay step of 2000, and an initial learning rate of 0.4. We also apply Xavier [45] to initialize the weights and biases in all convolution layers.

5 Results and discussions

In this section, we compare Wang-Net with existing spatial domain steganalysis models, such as SRM+EC, Xu-Net, Ye-Net, Yedroudj-Net, and Zhu-Net. Then, we apply the new migration learning method to the model, hoping to enhance the model’s ability of steganography detection generalization.

We compare the detection performance of Wang-Net with other steganalysis algorithms. The results are shown in the Table 5. The performance of Wang-Net is significantly better than that of traditional steganalysis algorithm SRM+EC and deep learning steganalysis algorithms Xu-Net, Ye-Net, and Yedroudj-Net, but the detection accuracy is about 2~3% lower than that of Zhu-Net. It shows that Wang-Net has a good steganalysis performance.

Table 5 The performance of Wang-Net and other steganalysis models on resampled images

Algorithms	WOW (0.2 bpp)	WOW (0.4 bpp)	S-UNIWARD (0.2 bpp)	S-UNIWARD (0.4 bpp)
SRM+EC	0.635	0.745	0.634	0.753
Xu-Net	0.676	0.793	0.609	0.728
Ye-Net	0.669	0.768	0.600	0.688
Yedroudj-Net	0.722	0.859	0.633	0.772
Zhu-Net	0.766	0.882	0.719	0.847
Wang-Net	0.749	0.860	0.691	0.819

The involved networks are trained and tested on BOSSBase

Table 6 The performance of Wang-Net, Yedroudj-Net, and Zhu-Net on resampled images

Algorithms	Payload	Yedroudj-Net	Zhu-Net	Wang-Net
WOW	0.2	0.722	0.766	0.812
	0.4	0.859	0.882	0.920
S-UNIWARD	0.2	0.633	0.719	0.777
	0.4	0.772	0.847	0.888

The involved networks are trained and tested on BOSSBase

In order to improve the generalization ability of steganography detection of the model, inspired by transfer learning, we propose a novel transfer learning method. We apply the datasets with lower embedding rates for training and compare the performance again.

According to the results in Table 6, for WOW (0.2), WOW (0.4), S-UNIWARD (0.2), and S-UNIWARD (0.4), the accuracy of Wang-Net are 0.812, 0.920, 0.777, and 0.888 respectively, which surpasses the Zhu-Net. It shows that Wang-Net can capture key steganographic traces under multiple embedding rates and has a good ability to express features. Now, our CNN model has the best steganalysis detection performance.

To sum up, after applying nonlinear detection mechanism, joint domain detection mechanism, and new migration learning method, Wang-Net can capture more abundant and diversified steganography semantic information and has better steganography detection performance.

6 Conclusion

In the field of steganalysis, it is of great significance for applying the CNN. In this paper, we propose a CNN steganalysis model with three great advantages.

- (1) We creatively propose the nonlinear feature detection mechanism, and simulate the nonlinear features extraction method of SRM. For WOW and S-UNIWARD, the accuracy of the model is 0.3~6% higher than that of the basic model. It shows that the nonlinear detection mechanism forces the model to adapt to the nonlinear distribution of steganography.
- (2) We pioneer the joint domain detection mechanism and simulate the manual feature extraction method of SRM in the spatial domain and DCTR in the transformation domain. For WOW and S-UNIWARD, the accuracy of the model is increased by 2~3%. It shows that the joint domain detection mechanism can help the model capture more abundant steganography features.
- (3) We propose a model transfer learning method, which uses low embedding rate images to initial the model. For WOW (0.2), WOW (0.4), S-UNIWARD(0.2), and S-UNIWARD (0.4), the accuracy of Wang-Net are 0.812, 0.920, 0.777, and 0.888 respectively, which is higher than that of the current Zhu-Net and other steganalysis models. It shows that Wang-Net can capture more levels of steganography features, which is conducive to the feature expression.

Our model is not suitable for steganalysis of color image. This issue will be addressed in further research.

Acknowledgements

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Authors' contributions

Yu Yang and Ze Wang designed the algorithm and revised the article content. Ze Wang carried out the experiments and wrote the manuscript. Yu Yang, Mingzhi Chen, and Min Lei gave the suggestions on the structure of manuscript and participated in modifying. Zhexuan Dong approved the final manuscript and optimized the English language.

Funding

Supported by the National Key R&D Program of China (2017YFB0802703), Major Scientific and Technological Special Project of Guizhou Province (20183001), Open Foundation of Guizhou Provincial Key Laboratory of Public Big Data (2018BDKFJJ014), Open Foundation of Guizhou Provincial Key Laboratory of Public Big Data (2018BDKFJJ019), and Open Foundation of Guizhou Provincial Key Laboratory of Public Big Data (2018BDKFJJ022).

Availability of data and materials

The datasets generated or analyzed during the current study are available in the [BOSSBase] repository, [<http://agents.fel.cvut.cz/boss/index.php?mode=VIEW&tmpl=materials>].

Competing interests

The authors declare that they have no competing interests.

Author details

¹State Key Laboratory of Public Big Data, Guizhou University, 550025 Guizhou Guiyang, China. ²Laboratory of Cyberspace Security, Beijing University of Posts and Telecommunications, 100876 Beijing, China. ³College of New Media, Beijing Institute of Graphic Communication, 102600 Beijing, China. ⁴Department of Computer Science, University of California, Irvine, 92697 CA, USA.

Received: 10 March 2020 Accepted: 3 June 2020

Published online: 08 July 2020

References

1. Y. Kang, F. Liu, C. Yang, X. Luo, T. Zhang, Color image steganalysis based on residuals of channel differences. *Comput. Mater. Contin.* **59**, 315–329 (2019). <https://doi.org/10.32604/cmc.2019.05242>
2. L. Shi, Z. Wang, Z. Qian, N. Huang, P. Puteaux, X. Zhang, Distortion function for emoji image steganography. *Comput. Mater. Contin.* **58**, 943–953 (2019). <https://doi.org/10.32604/cmc.2019.05768>
3. R. Meng, S. G. Rice, J. Wang, X. Sun, A fusion steganographic algorithm based on faster R-CNN. *Comput. Mater. Contin.* **55**, 1–16 (2018). <https://doi.org/10.3970/cmc.2018.055.001>
4. V. Holub, J. Fridrich, Low-complexity features for JPEG steganalysis using undecimated DCT. *IEEE Trans. Inf. Forensics Secur.* **10**(2), 219–228 (2015)
5. B. Furht, *Discrete wavelet transform (DWT)* (Springer, Boston, 2006), pp. 205–207. <https://doi.org/10.1007/038730038462>
6. V. Holub, J. Fridrich, T. Denemark, Universal distortion function for steganography in an arbitrary domain. *EURASIP J. Inf. Secur.* **1**(1), 1 (2014)
7. J. Fridrich, T. Pevný, J. Kodovsky, in *Proceedings of the 9th Workshop on Multimedia & Security. MM & Sec '07*, Statistically undetectable JPEG steganography: dead ends challenges, and opportunities (Association for Computing Machinery, New York, 2007), pp. 3–14. <https://doi.org/10.1145/1288869.1288872>
8. L. Guo, J. Ni, Y. Q. Shi, Uniform embedding for efficient JPEG steganography. *IEEE Trans. Inf. Forensics Secur.* **9**(5), 814–825
9. L. Guo, J. Ni, W. Su, C. Tang, Y. Shi, Using statistical image model for JPEG steganography: uniform embedding revisited. *IEEE Trans. Inf. Forensics Secur.* **10**(12), 2669–2680
10. N. F. Johnson, S. Jajodia, Exploring steganography: seeing the unseen. *Computer.* **31**(2), 26–34 (1998)
11. J. Fridrich, M. Goljan, D. Rui, Detecting LSB steganography in color, and gray-scale images. *Multimedia IEEE.* **8**(4), 22–28 (2001)
12. J. Mielikainen, LSB matching revisited. *IEEE Signal Process. Lett.* **13**(5), 285–287
13. D. C. Wu, W.-H. Tsai, A steganographic method for images by pixel-value differencing. *Pattern Recogn. Lett.* **24**(9–10), 1613–1626
14. J. S. Pan, W. Li, C. S. Yang, L. J. Yan, Image steganography based on subsampling and compressive sensing. *Multimed. Tools Appl.* **74**(21), 9191–9205 (2015)
15. W. Luo, F. Huang, J. Huang, Edge adaptive image steganography based on LSB matching revisited. *IEEE Trans. Inf. Forensics Secur.* **5**(2), 201–214 (2010)
16. T. Pevný, T. Filler, P. Bas, Using high-dimensional image models to perform highly undetectable steganography. **6387**, 161–177 (2010)
17. B. Li, M. Wang, J. Huang, X. Li, A new cost function for spatial image steganography. 2014 IEEE Int. Conf. Image Process. ICIP 2014, 4206–4210 (2015). <https://doi.org/10.1109/ICIP.2014.7025854>
18. V. Sedighi, R. Cograñe, J. Fridrich, Content-adaptive steganography by minimizing statistical detectability. *IEEE Trans. Inf. Forensics Secur.* **11**(2), 221–234 (2016)
19. Z. Qu, S. Wu, M. Wang, L. Sun, X. Wang, Effect of quantum noise on deterministic remote state preparation of an arbitrary two-particle state via various quantum entangled channels. *Quantum Inf. Process.* **16**(12), 306 (2017). <https://doi.org/10.1007/s11128-017-1759-8>
20. Z. Qu, Z. Cheng, W. Liu, X. Wang, A novel quantum image steganography algorithm based on exploiting modification direction. *Multimed. Tools Appl.* **78** (2018). <https://doi.org/10.1007/s11042-018-6476-5>
21. Z. Qu, Z. Li, G. Xu, S. Wu, X. Wang, Quantum image steganography protocol based on quantum image expansion and grover search algorithm. *IEEE Access.* **7**, 50849–50857 (2019). <https://doi.org/10.1109/ACCESS.2019.2909906>

22. R. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification. *IEEE Trans. Syst. Man. Cybern.* **SMC-3**, 610–621 (1973)
23. T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern. Anal. Mach. Intell.* **24**(7), 971–987
24. Q.-Y. Zhang, Q.-Y. Dou, Z. Yang, Y. Yan, Perceptual hashing of color images using interpolation mapping and non-negative matrix factorization. *Inf. Hiding Multimed. Signal Process.* **8**(3), 525–535 (2017)
25. Q.-Y. Zhang, Z. Yang, Q.-Y. Dou, Y. Yan, Robust hashing for color image authentication using non-subsampled contourlet transform features and salient features, vol. 8, (2017), pp. 1029–1042
26. J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images. *IEEE Trans. Inf. Forensics Secur.* **7**(3), 868–882 (2012). <https://doi.org/10.1109/TIFS.2012.2190402>
27. C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval. *IEEE Trans. Patt. Anal. Mach. Intell.* (01). **5555**, 1–1 (2020)
28. C. Yan, B. Gong, Y. Wei, Y. Gao, 3D room layout estimation from a single RGB image. *IEEE Trans. Multimed.* **PP**, 1–1 (2020). <https://doi.org/10.1109/TMM.2020.2967645>
29. S. Tan, B. Li, in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Stacked convolutional auto-encoders for steganalysis of digital images, (2014), pp. 1–4. <https://doi.org/10.1109/APSIPA.2014.7041565>
30. Y. Qian, D. Jing, W. Wei, T. Tan, *Deep learning for steganalysis via convolutional neural networks*. 9409, 94090–9409010 (2015)
31. G. Xu, H. Z. Wu, Y. Q. Shi, Structural design of convolutional neural networks for steganalysis. *IEEE Signal Process. Lett.* **23**(5), 708–712 (2016)
32. Y. Qian, D. Jing, W. Wei, T. Tan, in *2016 IEEE International Conference on Image Processing (ICIP)*, Learning and transferring representations for image steganalysis using convolutional neural network (IEEE, Phoenix, 2016), pp. 2752–2756
33. J. Zeng, S. Tan, B. Li, J. Huang, Large-scale JPEG steganalysis using hybrid deep-learning framework. *IEEE Trans. Inf. Forensics Secur.* **13**(5), 1200–1214 (2016)
34. J. Zeng, S. Tan, B. Li, J. Huang, Pre-training via fitting deep neural network to rich-model features extraction procedure and its effect on deep learning for steganalysis. *Electron. Imaging.* **2017**(7), 44–49
35. G. Xu, in *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. IH & MMSec '17*, Deep convolutional neural network to detect J-UNIWARD (Association for Computing Machinery, New York, 2017), pp. 67–73. <https://doi.org/10.1145/3082031.3083236>
36. M. Boroumand, M. Chen, J. Fridrich, Deep residual network for steganalysis of digital images. *IEEE Trans. Inf. Forensics Secur.* **14**(5), 1181–1193 (2019). <https://doi.org/10.1109/TIFS.2018.2871749>
37. J. Ni, J. Ye, Y. Yi, Deep learning hierarchical representations for image steganalysis. *IEEE Trans. Inf. Forensics Secur.* **12**(11), 2545–2557 (2017)
38. M. Yedroudj, F. Comby, M. Chaumont, *Yedrouj-Net: an efficient CNN for spatial steganalysis*. pp. 2092–2096 (2018)
39. BossBase (2019). <http://agents.fel.cvut.cz/boss/index.php?mode=VIEW&tmpl=materials>. Accessed 15 June 2020
40. Bows2 (2019). <http://bows2.gipsa-lab.inpg.fr/>. Accessed July 2007
41. C. F. Tsang, J. Fridrich, Steganalyzing images of arbitrary size with CNNs. *Electronic Imaging* (2018). <https://doi.org/10.2352/ISSN.2470-1173.2018.07.MWSF-121>
42. R. Zhang, F. Zhu, J. Liu, G. Liu, Efficient feature learning and multi-size image steganalysis based on CNN (2018). 1807.11428
43. V. Holub, J. Fridrich, Low-complexity features for JPEG steganalysis using undecimated DCT. *IEEE Trans. Inf. Forensics Secur.* **10**(2), 219–228
44. M. D. Zeiler, ADADELTA: an adaptive learning rate method. *Comput. Sci.* (2012). 1212.5701
45. X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res. Proc. Track.* **9**, 249–256 (2010)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
