SCIENCE CHINA
Information Sciences

• LETTER •

December 2018, Vol. 61 129105:1–129105:3
https://doi.org/10.1007/s11432-018-9640-7

Deep learning for steganalysis based on filter diversity selection

Kai ZHONG[1,2], Guorui FENG[1,2*], Liquan SHEN[1,2] & Jun LUO[3]

[1]*Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai* 200444*, China;*
[2]*School of Communication and Information Engineering, Shanghai University, Shanghai* 200444*, China;*
[3]*School of Mechatronic Engineering and Automation, Shanghai University, Shanghai* 200444*, China*

Received 24 April 2018/Revised 1 September 2018/Accepted 14 November 2018/Published online November 20, 2018

Citation Zhong K, Feng G R, Shen L Q, et al. Deep learning for steganalysis based on filter diversity selection. Sci China Inf Sci, 2018, 61(12): 129105, https://doi.org/10.1007/s11432-018-9640-7

Dear editor,

Steganography is an important method of covert communication. Steganalysis, an advanced anti-steganographic technique can detect steganographic behavior and even reveal the stego key in a certain steganographic method [1]. Steganalysis includes specialized steganalysis [2] and universal steganalysis [3]. This study focuses on feature extraction [3] and feature post-processing [4]. Ensemble classifiers are proposed to address the complexity [5]. Recently, some scholars have conducted research on steganalysis using deep learning. Xu et al. proposed two important methods and outperformed the spatial rich model (SRM) method in [6, 7]. They presented an excellent convolution neural network (CNN) structure, proposed to use batch normalization (BN) and rectified linear units (ReLU) structure to enhance the effect in [6], and used an ensemble of five CNNs to obtain better accuracy than the SRM method on the spatial-universal wavelet relative distortion (S-UNIWARD) [8]. They then suggested several different ensemble methods [7].

Proposed method. We will introduce three ensemble methods. The accuracy of ensemble learning is related to the accuracy of a single classifier and the diversity between different classifiers. Thus, we attempted to increase the diversity between different classifiers.

Diversity in different training sets. Different training sets cause diversity, which indicates that the sample permutation can produce better results. Diversity can be increased in many ways, such as bagging and bootstrap methods. We used a sample selection method similar to the holdout method. The whole data set D is divided into two mutually exclusive sets, set T and a testing set $\bar{T}$, i.e., $D = T \cup \bar{T}$, $T \cap \bar{T} = \emptyset$. After training the model on T, the testing error is evaluated by $\bar{T}$. We randomly selected 4000 pairs of images as the training set from T. The rest of the images of T were used as a validation set to estimate the performance of models.

Diversity in different high-pass filters. The diversity in initialization parameters is exhibited in several ways, such as the different initialization weights of the convolution filter (different in each of training) and the difference between high-pass filters (HPFs). The noise residual produced by different filters will cause an obvious diversity because the layers behind HPF will have a great effect. In the proposed method, every HPF is a 5×5 matrix. According to [6], we produced the basic HPF based on the following expression:

$$\boldsymbol{H} = \frac{1}{12} \times \begin{pmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{pmatrix},$$

[*] Corresponding author (email: fgr2082@aliyun.com)

and define all raw filters as

$$\frac{1}{|z-12|}\begin{pmatrix} -1 & 2+m & -2+n & 2+m & -1 \\ 2+m & -6+k & 8+l & -6+k & 2+m \\ -2+m & 8+l & -12+z & 8+l & -2+n \\ 2+m & -6+k & 8+l & -6+k & 2+m \\ 1 & 2+m & -2+n & 2+m & -1 \end{pmatrix}.$$

We obtain

$$\begin{cases} n=-2m, \\ 2m+2k+l=0, \\ 2l-4m+z=0. \end{cases}$$

According to the above conditions, we have randomly set values of m and l to be natural numbers and generated ten HPFs known $\mathrm{HPF}_0, \mathrm{HPF}_1, \ldots, \mathrm{HPF}_9$. We used these ten HPFs to train the CNNs and obtained ten different CNN models. Four HPFs were selected with each of them having an enormous diversity as compared to the other HPFs. The four HPFs and the basic HPF will form five different filters to achieve diversity. We defined a metric to compare the diversity between the HPFs. Let $\boldsymbol{C}_k$ and $\boldsymbol{S}_k$ denote the kth original image and the corresponding image with steganography, respectively ($1 \leqslant k \leqslant s$). We calculated the convolution between $\boldsymbol{H}_p$ and $\boldsymbol{C}_k$ or $\boldsymbol{S}_k$ to obtain $\boldsymbol{I}^c_{k,p}$ and $\boldsymbol{I}^s_{k,p}$:

$$\boldsymbol{I}^c_{k,p} = \boldsymbol{C}_k * \boldsymbol{H}_p,\ \boldsymbol{I}^s_{k,p} = \boldsymbol{S}_k * \boldsymbol{H}_p, \tag{1}$$

where $\boldsymbol{H}_p$ denotes the pth HPF ($0 \leqslant p \leqslant s$). Then the difference between $\boldsymbol{I}^c_{k,p}$ and $\boldsymbol{I}^s_{k,p}$ is calculated as below:

$$\boldsymbol{I}_{k,p} = \boldsymbol{I}^c_{k,p} - \boldsymbol{I}^s_{k,p}. \tag{2}$$

To uniform the value range of different images, the element $I_{k,p}(i,j)$ of $\boldsymbol{I}_{k,p}$ is normalized as $\bar{I}_{k,p}(i,j)$:

$$\bar{I}_{k,p}(i,j) = \frac{I_{k,p}(i,j)}{\sum_i \sum_j |I^c_{k,p}(i,j)|}. \tag{3}$$

We calculated the Frobenius norm of $I^n_{k,p}$ as the difference between HPF_i and HPF_j:

$$C_{i,j} = \sum_{k=1}^{s} \|\bar{\boldsymbol{I}}_{k,i} - \bar{\boldsymbol{I}}_{k,j}\|. \tag{4}$$

All space models are named as S, wherein each S has s models. Our goal is to find s' sub-models, such that these sub-models have maximal diversity for further ensemble requirement. If we choose n ($n < s' < s$) models to be included in the model collection S', then we will select the $(n+1)$-th models. We define $\bar{S}'$ as the complementary set of S'. The model whose diversity is maximal compared to that of the other models, can be obtained by the following expression:

$$i_o = \arg\max_i \{D(i,S') | i \in \bar{S}'\}, \tag{5}$$

where

$$D(i,S') = \sum_{j_n} C_{i,j_n}, \quad j_n \in S'. \tag{6}$$

If $D(i_o,S')(i_o \in \bar{S}')$ is maximal, then HPF_{i_o} is the model selected to be included in S'. The above process is repeated until the models in S' are sufficient in number. In the following experiments, we have $s=10$ and $s'=5$, i.e., five HPFs are selected from the ten HPFs.

Diversity in CNN training process. When the CNN learning parameters are updated, the results generated by different iterations are not similar. If we can achieve diversity in the same training process, then we can improve the accuracy without increasing the cost. The proposed method saves a model every two epochs, therefore, we will obtain T models. In accordance with the accuracy of the models in the validation set, we sort all the models and select J suitable models (experimentally $J=10$).

Ensemble scheme. The entire ensemble stage has $J \times H \times R = N$ models totally, where J, H, and R are the numbers of the selected models, HPFs and permutated training sets, respectively. If we use all the models for selective integration, then the calculation would be costly. Therefore, we used the step-by-step ensemble method. We defined $h_{jhr}(\boldsymbol{X}_k)$ to represent the class probability of class 0 (the class probability of class 1 is $1-h_{jhr}(\boldsymbol{X}_k)$), where j, h, and r represent the jth chosen models, hth HPF kernel and rth training set in the training stage, respectively. In the experiments, ten models, five filters, and three permutations were selected. The process includes the following steps:

(1) We obtained J suitable models to execute the ensemble learning by using the prior method. We choose i models from J models ($i \leqslant J$); thus, we have a total of C^i_J combinations. We calculated the accuracy by comparing the class label $\hat{y}_n$ with the real class label y_n. Hence, the class label $\hat{y_n}$ is the following:

$$\hat{y}_k = \begin{cases} 0, & \sum_{j=1}^{i} h_{jhr}(\boldsymbol{X}_k) < 0.5i; \\ 1, & \sum_{j=1}^{i} h_{jhr}(\boldsymbol{X}_k) \geqslant 0.5i. \end{cases} \tag{7}$$

We need to determine the label of the i models.

(2) We selected ensemble methods with different HPFs. H represents the number of HPFs for

Table 1 Detection errors by compared with other methods

	Sel	Method in [6]	Ens	Method in [7]	SRM [3]
S-UNIWARD	17.98	19.18	17.53	18.44	20.35
WOW	16.92	–	16.23	–	20.59

the same training sets and C_H^i combinations are present.

(3) We have used R classifiers with different training sets. The classifiers are evaluated according to the accuracy.

In the proposed method, CNN is used to extract the feature and ensemble classifier in [5] is trained for the final decision because the ensemble classifier is stronger than the softmax classifier with regard to performing classification of steganalysis. We concatenated the output of every CNN's pooling layer as the ensemble classifier's input. The ensemble classifier chooses different combinations of features to achieve improved classification performance, such as the selective ensemble.

Experiments. All the experiments were performed on two spatial domain steganographic algorithms: spatial-universal wavelet relative distortion (S-UNIWARD) [8] and wavelet obtained weights (WOW) [9], with an embedding rate of 0.4 bpp. The well-known BOSSBase [3] is used to verify the proposed method. Both the sample sets include 5000 cover images and 5000 stego images in pairs. In the training phase, we used different random seeds to generate 4000/1000 splits as the training set from these 5000 pairs. The 1000 pairs are treated as the validation set.

In our experiments, the training epochs are set to 480. The parameters of the gradient descent are almost the same as those of the method in [7]. Learning rate is initialized to 0.001, which is scheduled to decrease by 10% every 5000 iterations and the momentum is fixed at 0.9. A mini-batch includes 64 pairs. We saved the weights models every two epochs, and ensured that the cover and stego pairs are in the same batch. The last convolution module has 128 convolution kernels; thus, the feature dimensions of all the ensemble methods are $150 \times 128 = 19200$. The error rate P_E, which is the average of false alarm and missing detection probabilities, is used to evaluate the classification performance. Table 1 shows a comparable result of the proposed method and other recent methods. Sel is a step-by-step selective ensemble method, the method in [6] is the simple ensemble of CNN, Ens is the method of incorporating features into the ensemble classifier, the method in [7] is of adding a new representative in the ensemble method of combinations. The error rate P_E of Ens is 1.65% lower than that of the method in [6] as well as lower than P_E of the method in [7].

Acknowledgements This work was supported by National Natural Science Foundation of China (Grants Nos. U1536109, 61671282, 61373151, 61525305).

References

1 Liu J, Tian Y, Han T, et al. Stego key searching for LSB steganography on JPEG decompressed image. Sci China Inf Sci, 2016, 59: 032105

2 Luo X Y, Song X F, Li X L, et al. Steganalysis of HUGO steganography based on parameter recognition of syndrome-trellis-codes. Multimed Tools Appl, 2016, 75: 13557–13583

3 Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. IEEE Trans Inform Forensic Secur, 2012, 7: 868–882

4 Ma Y Y, Luo X Y, Li X L, et al. Selection of rich model steganalysis features based on decision rough set α-positive region reduction. IEEE Trans Circ Syst Video Tech, 2018. doi: 10.1109/TCSVT.2018.2799243

5 Kodovsky J, Fridrich J, Holub V. Ensemble classifiers for steganalysis of digital media. IEEE Trans Inform Forensic Secur, 2012, 7: 432–444

6 Xu G, Wu H Z, Shi Y Q. Structural design of convolutional neural networks for steganalysis. IEEE Signal Process Lett, 2016, 23: 708–712

7 Xu G, Wu H Z, Shi Y Q. Ensemble of CNNs for steganalysis: an empirical study. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, 2016. 103–107

8 Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain. EURASIP J Info Secur, 2014, 2014: 1

9 Holub V, Fridrich J. Designing steganographic distortion using directional filters. In: Proceedings of IEEE International Workshop on Information Forensics and Security (WIFS), 2012. 234–239