IOP Conf. Series: Materials Science and Engineering **719** (2020) 012068 doi:10.1088/1757-899X/719/1/012068

Steganalysis of Intra Prediction Mode and Motion Vector-based Steganography by Noise Residual Convolutional Neural Network

Peng Liu and Songbin Li

Haikou Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Haikou 570105, China.
Email: liup@dsp.ac.cn

Abstract. In this paper, we present a universal steganalysis method for both intra prediction mode and motion vector-based steganography based on deep learning. Since the embedding process is eventually reflected in the modification of pixel values in decoded frames, we design a Noise Residual Convolutional Neural Network (NR-CNN) from the perspective of the spatial domain, which is the first CNN-based approach for this subject. In NR-CNN, feature extraction and classification modules are integrated into a unified and trainable network framework. It automatically learns features and implements classification in a data-driven manner, which effectively solves the existing problems. Experimental results show that NR-CNN has better performance of steganalysis than the related method.

1. Introduction

Steganography is the art and science of data hiding, which realizes covert communication by embedding secret data into an innocent-looking cover media, such as digital image, audio, video, et al., without arousing any suspicion. In contrast, steganalysis aims to expose the presence of hidden data. In this paper, we are focusing on the intra prediction mode [1-3] and motion vector-based [4-6] steganography.

To detect the intra prediction mode-based steganography, Li et al. [7] design a series of features based on Markov chain to quantify this correlation property. Zhao et al. [8] conduct the steganalysis based on intra prediction mode calibration; In the aspect of steganalysis for motion vector-based steganography, some feature-based methods have been presented in recent years. They can be divided into three major categories. The first category designs features based on neighboring motion vector difference [9-10] The second category uses calibrations to enhance the features [11-12]. The third category utilizes the statistics of Sum of Absolute Differences (SADs) to construct features [13-15].

In this paper, we propose a universal steganalysis method for both intra prediction mode and motion vector-based steganography based on deep learning. Since the process of intra prediction mode and motion vector-based steganography is eventually reflected in the modification of pixel values in decoded frames, we design a Noise Residual Convolutional Neural Network (NR-CNN) from the perspective of the spatial domain. Feature extraction and classification modules are integrated into a unified and trainable network framework. It automatically learns features and implements classification in a data-driven manner, which effectively solves the existing problems.

2. Network Architecture

The embedding process can be considered as adding low-amplitude noise in the cover image, so the secret information has an extremely low SNR compared to the image content. The low SNR mainly leads to two types of problems. The first is that commonly used activation functions such as ReLU do not fully apply to this type of task. Since the proportion of the useful signal in the input signal itself is already very low, forcibly losing half of the signal each time will cause a large number of invalid filters in the training process; the second is the problem of parameter initialization. The network usually cannot converge when using the most commonly used weights initialization method in the field of computer vision.

Recently, Ye et al. [16] give some ideas to solve these problems. They introduce an Image Steganalysis Network (IS-Net) based on CNN which uses 30 high-pass filters as weights to initialize the first convolutional layer. In addition, they introduce a new activation function called Truncated Linear Unit (TLU). The idea of IS-Net can help us to solve the problems in video steganalysis. But IS-Net is designed specifically for image steganalysis. In this paper, we present a Noise Residual Convolutional Neural Network (NR-CNN) based on the ideas of IS-Net. The main idea of the NR-CNN is to extract features from noise residuals, which are used for classification. The network structure is shown in figure 1. Improvements are made from three aspects: First, we add four global filters in the residual computation part to obtain more feature maps of steganographic noise residual signal; second, a new activation function called Parametric Truncated Linear Unit (PTLU) is proposed to better capture the structure of embedding signals; third, a steganalysis residual block structure is presented, which can improve the learning ability of steganographic noise residual signal. As we can see from figure 1, the network is divided into three main parts, which are used for residual computation, feature extraction, and binary classification.

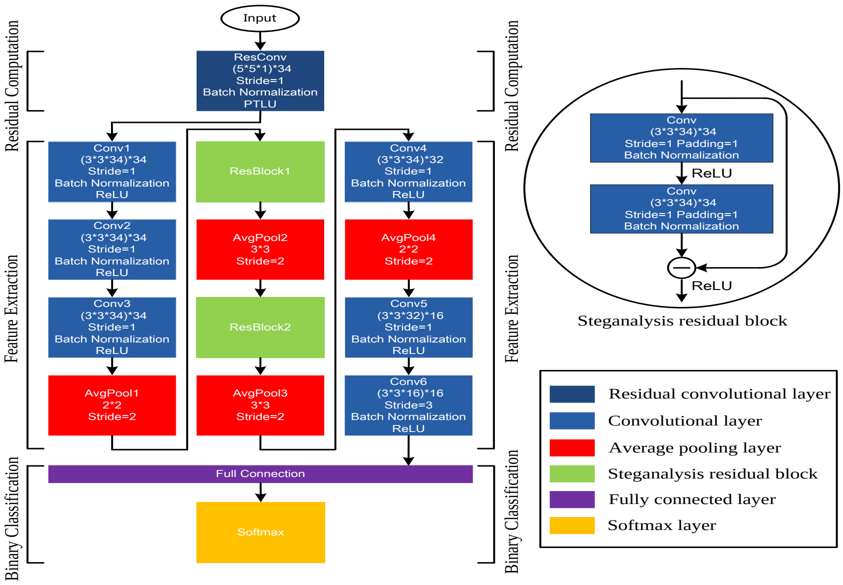

Figure 1. The architecture of the proposed convolutional neural network.

2.1. Residual Convolutional Layer

The residual convolutional layer "ResConv" is used to compute the steganographic noise residual. Current convolutional neural networks tend to learn features from the image content. However, the embedded secret information is independent from the image content. Thus, the residual convolutional layer is very important, whose role is to obtain steganographic noise residual features that are independent of the image content.

Figure 2 shows the visualizations of high-pass filters used in the first convolutional layers of IS-Net. Each filter corresponds to a noise residual feature, and we can see that only the filter in the fifth row and the second column is a global filter that covers every pixel in the 5×5 area. In the video steganography, information hiding is conducted in units of one block. Modification of the intra prediction mode or motion vector will changes the pixel value of one block. So, we consider that more global filters are needed to represent the steganographic noise residual signals. Figure 3 shows the new global filters introduced in this paper. Therefore, a total of 34 convolution kernels are used in the residual convolutional layer of NR-CNN. Although initialization using these 34 convolution kernel parameters is better than random initialization, these parameters are not the best. These parameters will be optimized by global optimization during training process.

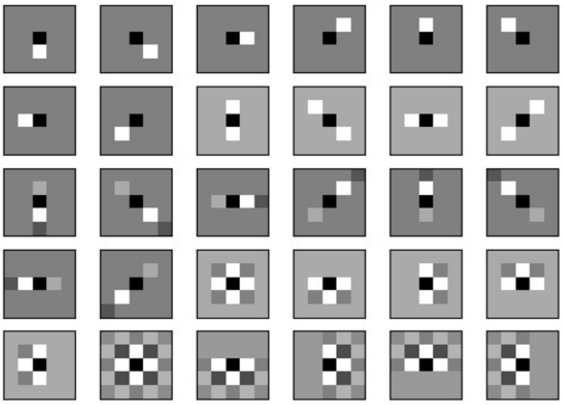

Figure 2. Visualization of high-pass filters used in the first convolutional layer of IS-Net.

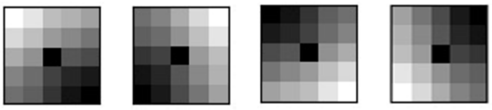

Figure 3. Visualizations of new global filters.

The difference between the residual convolutional layer and the common convolutional layer is that the parameters of residual convolutional layer are initialized with fixed values. The input data for this layer is 256×256 single-channel image data. This layer contains 34 filters of size 5×5×1, where 1 represents the number of channels. The convolution step size is 1 and the output of this layer is 34 feature maps with a size of 252×252.

2.2. Convolutional Layer

The feature extraction part contains six convolutional layers. The convolutional layers "Conv1", "Conv2", and "Conv3" all contain 34 filters of size 3x3x34 with a step size of 1, and use ReLu as the activation function. Among them, the output of the convolutional layer “Conv1” is 34 feature maps with a size of 250×250. The output of the convolutional layer “Conv2” is 34 feature maps with a size of 248×248, and the output of the convolutional layer “Conv3” is 34 feature maps with a size of 246×246. The convolutional layer "Conv4" contains 32 filters of 3x3x34, and uses ReLU as an activation function. "Conv4" outputs 32 feature maps with a size of 28x28. The convolutional layer "Conv5" contains 16 filters of 3x3x32 with a step size of 1, and uses ReLU as the activation function. The output is 16 feature maps with a size of 12x12. The convolutional layer “Conv6” contains 16 filters of 3×3×16 with a step size of 3. It uses ReLU as the activation function, and outputs 16 feature maps with a size of 4×4. It should be noted that we use the batch normalization operation to normalize the data before the activation function in each convolutional layer.

2.3. Activation Function

Ye el. [16] present a new activation function for image steganalysis called Truncated Linear Unit (TLU). They prove that TLU is better than ReLU for image steganalysis. We extend TLU and propose a new activation function called Parametric Truncated Linear Unit (PTLU), which is defined as

$$f(x)=\begin{cases} T, & x>T \\ x, & 0\le x\le T \\ \alpha x, & -T/\alpha \le x<0 \\ -T, & x<-T/\alpha \end{cases} \tag{1}$$

For PTLU, the coefficient of the negative part is not constant and can be learned adaptively. When $\alpha=1$, PTLU is equivalent to TLU.

2.4. Steganalysis Residual Block

Since the depth of CNN has a great influence on the classification performance, it is usually considered that deeper network result in the better result. However, a deeper network will lead to higher training error than the shallow network. This can be understood as data disappearing through too many layers of the network, which will lead to a worse result. To this end, He et al. [17] propose the ResNet, which contains several residual blocks as shown in figure 4(a).

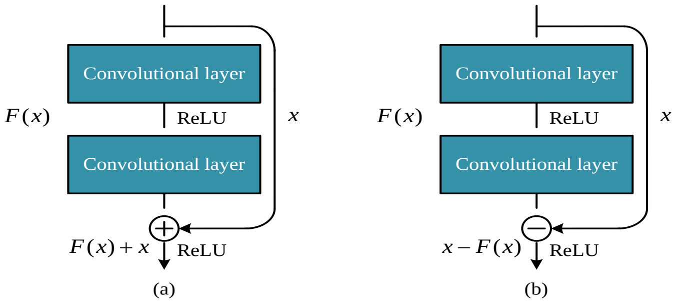

Figure 4. The comparison of residual blocks.
(a)The structure of traditional residual block. (b) The structure of steganalysis residual block.

In this paper, an improved structure of the residual block is proposed for steganalysis. As shown in figure 4(b), the improved structure is called steganographic residual block. The main improvement is to change the final mapping function from $F(\boldsymbol{x})+\boldsymbol{x}$ to $\boldsymbol{x}-F(\boldsymbol{x})$. In the steganalysis problem, the input data $\boldsymbol{x}$ can be regarded as the sum of the carrier image content $\boldsymbol{c}$ and the steganographic residual signal $\boldsymbol{m}$. Ideally, the content of the carrier image in the input data $\boldsymbol{x}$ has been already filtered out in the previous processing, and at this time $\boldsymbol{c}=0$. However, in the actual situation, the carrier image content cannot be filtered completely, which means $\boldsymbol{c}\neq 0$. The purpose of the steganographic residual block is to further suppress the carrier image content so as to reduce $\boldsymbol{c}$. In the steganographic residual block, $F(\boldsymbol{x})$ is used to filter out the steganographic residual signal $\boldsymbol{m}$ so as to only retain the carrier image content $\boldsymbol{c}$. Then the steganographic residual signal $\boldsymbol{m}$ can be retained as much as possible by $\boldsymbol{x}-F(\boldsymbol{x})$. Therefore, the steganographic residual block is very suitable for learning the steganographic residual signal ***m***. In this paper, two steganographic residual blocks are used for steganalysis.

Each steganographic residual block contains two convolutional layers. Each convolutional layer contains 34 filters with a size of 3×3×34 and uses zero padding. The step size is 1 and ReLU is used as the activation function. Batch Normalization is conducted before ReLU. Due to zero padding, the size of output feature maps is the same as that of input. The output of the steganographic residual block "ResBlock1" is 34 feature maps with a size of 123×123. The output of "ResBlock2" is 34 feature maps of a size of 61×61.

2.5. Pooling Layer

The main role of the pooling layer is to reduce the dimension of input feature, thereby reducing the parameters and computation of the entire network, and suppressing overfitting.

The feature extraction part contains four pooling layers, all using average pooling. The pooling layer “AvgPool1” has a kernel size of 2×2 and a step size of 2, which outputs 34 feature maps with a size of 123×123. The pooling layer "AvgPool2" has a kernel size of 3x3 and a step size of 2, which outputs 34 feature maps with a size of 61x61. The pooled layer “AvgPool3” has a kernel size of 3×3 and a step size of 2, which outputs 34 feature maps with a size of 30×30. The pooling layer "AvgPool4" has a kernel size of 2x2 and a step size of 2, which outputs 32 feature maps with a size of 14x14.

3. Experimental Results and Discussion

In this section, we will evaluate the availability and effectiveness of the proposed NR-CNN. Since

NR-CNN is inspired by IS-Net, we will compare with it in this section.

3.1. Experimental Settings

The dataset used in this paper contains three parts: training set, verification set and test set. There are 200,000 frames in the training set, 20,000 frames in the verification set, and 200,000 frames in the test set. The training set and the verification set are used to train the network. The test set is used to evaluate the information hiding detection accuracy.

In the experiment, we use the methods in [1] and [6] for intra prediction mode and motion vector-based steganography, respectively. For the training set and verification set, we use 100% embedding rate for intra prediction mode-based steganography and 20% embedding rate for motion vector-based steganography. For the test set, we use five embedding rates of 20%, 40%, 60%, 80%, and 100% for intra prediction mode-based steganography, and use 5%, 8%, 10%, 15%, and 20% for motion vector-based steganography. IS-Net and NR-CNN are implemented on the deep learning framework PyTorch, and the batch size is 32. The network optimizer is AdaDelta with a learning rate of 0.4, a momentum value of 0.95, a weight decay of 5×10^{-4}, and a "delta" parameter of 1×10^{-8}. The number of training iterations epoch is 150, and the goal of the training process is to minimize the cross-entropy cost function.

3.2. Comparison of NR-CNN and IS-Net

Previous experiments show that both NR-CNN and IS-Net have the best detection performance when T=7. Thus, we will compare the two networks under this threshold. Table 1 shows the detection accuracy for intra prediction mode-based steganography under different embedding rates. It can be seen from the table that NR-CNN performs better than IS-Net under each embedding rate. Table 2 shows the detection accuracy for motion vector-based steganography under different embedding rates. It can be seen from the table that NR-CNN performs better than IS-Net under all embedding rates.

Table 1. The detection accuracy for intra prediction mode-based steganography under different embedding rates.

Method	Embedding rates				
	20%	40%	60%	80%	100%
NR-CNN	59.82%	85.67%	98.2%	98.95%	99.74%
IS-Net	58.07%	83.41%	96.25%	97.69%	98.7%

Table 2. The detection accuracy for motion vector-based steganography under different embedding rates.

Method	Embedding rates				
	5%	8%	10%	15%	20%
NR-CNN	62.53%	73.82%	82.37%	91.48%	95.39%
IS-Net	60.39%	70.86%	77.69%	90.28%	93.45%

4. Conclusion

In this paper, a universal steganalysis method is proposed for both intra prediction mode and motion vector-based steganography based on deep learning. Since the embedding process is eventually reflected in the modification of pixel values in decoded frames, we design a Noise Residual Convolutional Neural Network (NR-CNN) from the perspective of the spatial domain. Feature extraction and classification modules are integrated into a unified and trainable network framework. It automatically learns features and implements classification in a data-driven manner, which effectively solves the existing problems in the framework of “feature extraction-feature classification”. Experimental results show that NR-CNN has better performance of steganalysis than IS-Net.

5. Acknowledgments

This work is supported partly by the Important Science & Technology Project of Hainan Province under grant ZDKJ201807, part by the National Natural Science Foundation of China under grant U1636113, part by the Hainan Provincial Natural Science Foundation of China under grant 618QN309, part by the Scientific Research Foundation Project of Haikou Laboratory, Institute of Acoustics, Chinese Academy of Sciences, and part by the IACAS Young Elite Researcher Project QNYC201829 and QNYC201747.

6. References

[1] HU Y, ZHANG C, SU Y. Information hiding based on intra prediction modes for H.264/AVC[C]. Proceedings of the 2007 IEEE International Conference on Multimedia and Expo. Beijing, China, July 2007: 1231-1234.
[2] YANG G, LI J, HE Y, et al. An information hiding algorithm based on intra-prediction modes and matrix coding for H.264/AVC video stream[J]. AEU-International Journal of Electronics and Communications, 2011, 65(4): 331-337.
[3] XU D, WANG R, WANG J. Prediction mode modulated data-hiding algorithm for H.264/AVC[J]. Journal of Real-Time Image Processing, 2012, 7(4): 205-214.
[4] ZHU H, WANG R, XU D. Information hiding algorithm for H.264 based on the motion estimation of quarter-pixel[C]. International Conference on Future Computer and Communication, IEEE. 2010:V1-423-V1-427.
[5] Yao Y, Zhang W, Yu N, et al. Defining embedding distortion for motion vector-based video steganography[J]. Multimedia Tools and Applications, 2015, 74(24), 11163-11186.
[6] ZHANG H, CAO Y, ZHAO X. Motion vector-based video steganography with preserved local optimality[J]. Multimedia Tools and Applications , 2016 , 75 (21):1-17.
[7] LI S, DENG H, TIAN H, et al. Steganalysis of prediction mode modulated data-hiding algorithms in H.264/AVC video stream[J]. annals of telecommunications - annales des télécommunications, 2014, 69(7-8):461-473.
[8] ZHAO Y, ZHANG H, CAO Y, et al. Video steganalysis based on intra prediction mode calibration[C]. International Workshop on Digital Watermarking. Springer, Cham, 2015: 119-133.
[9] SU Y, ZHANG C, ZHANG C. A video steganalytic algorithm against motion-vector-based stegangraphy[J]. Signal Processing, 2011, 91(8): 1901-1909.
[10] Hao-Tian Wu, Yuan Liu, Jiwu Huang, and Xin-Yu Yang. 2014. Improved steganalysis algorithm against motion vector based video steganography. In Proc. IEEE Int. Conf. Image Processing (ICIP). 5512–5516.
[11] CAO Y, ZHAO X, FENG D. Video steganalysis exploiting motion vector reversion-based features[J]. IEEE signal processing letters, 2012, 19(1): 35-38.
[12] Yu Deng, Yunjie Wu, and Linna Zhou. 2012. Digital video steganalysis using motion vector recovery-based features. Appl. Opt. 51, 20 (Jul. 2012), 4667–4677.
[13] WANG K, ZHAO H, WANG H. Video steganalysis against motion vector-based steganography by adding or subtracting one motion vector value[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(5): 741-751.
[14] Yanzhen Ren, Liming Zhai, Lina Wang, and Tingting Zhu. 2014. Video steganalysis based on subtractive probability of optimal matching feature. In Proc. 2nd ACM Workshop Inf. Hiding Multimedia Security. 83–90.
[15] ZHANG H, CAO Y, ZHAO X. A steganalytic approach to detect motion vector modification using near-perfect estimation for local optimality[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(2): 465-478.
[16] YE J, NI J, YI Y. Deep Learning Hierarchical Representations for Image Steganalysis[J]. IEEE Transactions on Information Forensics & Security, 2017, 12(11):2545-2557.
[17] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016:770-778.