RHFCN: FULLY CNN-BASED STEGANALYSIS OF MP3 WITH RICH HIGH-PASS FILTERING

Yuntao Wang, Xiaowei Yi, Xianfeng Zhao, Ante Su*

State Key Laboratory of Information Security, Institute of Information Engineering,
Chinese Academy of Sciences, Beijing, 100093
School of Cyber Security, University of Chinese Academy of Sciences, Beijing, 100093

ABSTRACT

Recent studies have shown that convolutional neural networks (CNNs) can boost the performance of audio steganalysis. In this paper, we propose a well-designed fully CNN architecture for MP3 steganalysis based on rich high-pass filtering (HPF). On the one hand, multi-type HPFs are employed for "residual" extraction to enlarge the traces of the signal in view of the truth that signal introduced by secret messages can be seen as high-pass frequency noise. On the other hand, to utilize the spatial characteristics of feature maps better, fully connected (Fc) layers are replaced with convolutional layers. Moreover, this fully CNN architecture can be applied to the steganalysis of MP3 with size mismatch. The proposed network is evaluated on various MP3 steganographic algorithms, bitrates and relative payloads, and the experimental results demonstrate that our proposed network performs better than state-of-the-art methods.

***Index Terms*—** CNNs, MP3, steganalysis, steganography, QMDCT coefficients

1. INTRODUCTION

MP3 is a kind of widely used audio compressed format. Due to the development of audio processing technology, complexity of MP3 encoding and its good concealment, MP3 has become an excellent carrier for the data hiding in audio steganography. Recently, MP3 steganography and steganalysis has drawn more attention than before and MP3 steganalysis is more urgent to be dealt with. Additionally, it's worthwhile to note that various steganalytic algorithms of MP3 could also be deployed on the steganalysis of AAC with little or no modification because of the similarity with MP3 which leads to a significance boost of MP3 steganalysis.

This work was supported by NSFC under U1636102 and U1736214, National Key Technology R&D Program under 2016QY15Z2500 and 2016YFB0801003, and Project of Beijing Municipal Science & Technology Commission under Z181100002718001.

E-mail: {wangyuntao2, yixiaowei, zhaoxianfeng, suante}@iie.ac.cn.

GitHub: https://github.com/Charleswyt/tf_audio_steganalysis.

* Corresponding author.

In recent years, various CNN-based methods [1]-[6] have been proposed for the steganalysis of spatial and JPEG images, while most of MP3 steganalytic algorithms are still based on the handcrafted features design [7]-[11]. Therein, Qiao et al. [9] presented a MP3 steganalytic algorithm based on frequency-based sub-band moment statistical features, accumulative Markov transition features, and accumulative neighboring joint density features on second-order derivatives. After that, Jin et al. [10] proposed to extract one-step transition probabilities between two immediately neighboring coefficients which named as ADOTP to steganalyze MP3Stego with low embedding-rate. Besides, Ren et al. [11] extracted the correlation characteristics, including Markov transition probability and accumulative neighboring joint density, between the multi-order differential coefficients of Intra and Inter frame (MDI2) to detect AAC steganography. Furthermore, the first CNN-based steganalytic algorithm for MP3 steganalysis is proposed in [12], which has a great detection performance on the steganalysis of equal length entropy codes substitution (EECS) [13] algorithm, and we name the network as WASDN.

In this paper, we propose a fully CNN structure for MP3 steganalysis which is named as RHFCN. First, the signal introduced by steganographic algorithms can be seen as additive high-frequency noise compared with audio content, thus we deploy a rich HPF module to capture the signal comprehensively and suppress the impact of audio signal on the steganalysis. And, the residual data and the original data are concatenated in depth to make full use of the information of the input data. Then, convolutional layers are applied to replace Fc layers for a better utilization of the spatial and structural correlation of feature maps. Last but not the least, due to the removal of Fc layers, our proposed network can be employed to deal with the steganalysis of MP3 with size mismatch. The experimental results have shown that RHFCN provides a better detection performance than state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 introduces the design of our proposed network. Section 3 shows the experimental settings and results. And, the conclusions are drawn in Section 4.

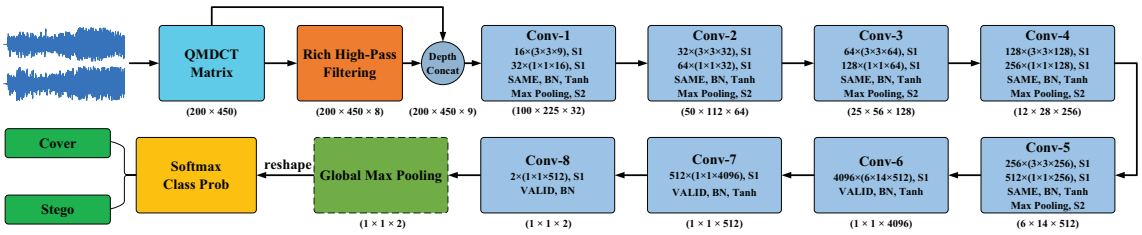

Fig. 1: Structure of RHFCN. "$16 \times (3 \times 3 \times 9)$" means that the kernel size of the convolutional layer is 3×3, the input channels are 9, and the output channels are 16. "S1" means the stride of slide window is 1. "SAME" and "VALID" are types of padding algorithms to use. "BN" and "Tanh" represent the batch normalization layer and the activation function respectively. The dimension of each block's output feature maps is presented below the convolutional boxes.

2. THE PROPOSED NETWORK

2.1. Architecture Overview

The whole architecture of RHFCN is shown in Fig. 1. In our network, the quantized modified discrete cosine transform (QMDCT) coefficients matrix of 200×450 is selected as the input data. Rich high-pass filtering module, including eight HPFs, is deployed to enhance the strength of the signal introduced by secret messages. And, the QMDCT coefficients matrix is concatenated with all differential matrices in depth to not lose the information of QMDCT coefficients matrix itself. Then, eight convolutional (Conv) blocks are connected in cascade for automatic feature extraction. The detailed diagram of Conv blocks is shown in Fig. 2. Each block is a combination of a 3×3 Conv layer, a 1×1 Conv layer, two Tanh activation functions, a batch normalization (BN) layer and a max pooling layer. The size of 6×14 convolutional kernel and valid mode of padding are deployed in Conv-6 for the dimension reduction of feature maps, which is used to replace Fc layers. And, the global max pooling (GMP) layer at the end of the network is used for the dimension unification of large size matrices. Feature maps are finally reshaped to a 1×2 vector for the final steganalysis. Thus, our network can be employed to steganalyze MP3 clips with size mismatch to some extends. What's more, in order to retain the zero mean value characteristics of the input data and reduce the inhibition effect of ReLu on the negative value of the QMDCT coefficients matrix, we employ activation function Tanh in our network.

2.2. Input Data – QMDCT Coefficients Matrix

The upper bound of the network performance depends on the input data and the fine-tuning of networks is a process of approaching the limit. In our network, the QMDCT coefficients matrix is selected as the input data. Firstly, QMDCT coefficients are important parameters of MP3 and arbitrary modification may lead to a failure of decode, which means potential characteristics can be captured based on QMDCT coefficients. Next, the influence of all MP3 steganographic algorithms on MP3 audio is reflected on the coefficients indeed. Finally, we can obtain the correlation proprieties of inter and intra frames through the forming of QMDCT coefficients matrix. Particularly, it is common to design efficient handcrafted features based on the QMDCT coefficients matrix. For stereo MP3 files we see commonly, a frame consists of two granules, a granule contains two channels and a channel includes 576 QMDCT coefficients. The coefficients at the end of each channel are all zero. Thus, to reduce redundant computation and save graphics memory, these coefficients will be removed directly. 50 frames (almost duration of 1.3s) of MP3 audio files are extracted as the basic analytic unit, and the size of the QMDCT coefficients matrix is 200×450. To facilitate the description better, the matrix is denoted as

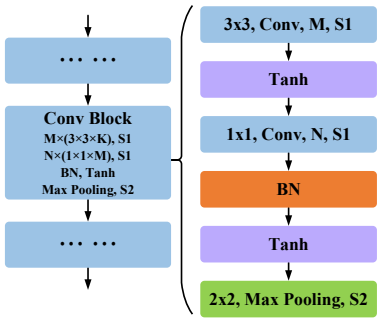

Fig. 2: Structure of convolutional block.

$$M_Q = \begin{bmatrix} Q_{1,1} & \cdots & Q_{1,j} & \cdots & Q_{1,450} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ Q_{i,1} & \cdots & Q_{i,j} & \cdots & Q_{i,450} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ Q_{200,1} & \cdots & Q_{200,j} & \cdots & Q_{200,450} \end{bmatrix} \quad (1)$$

where variable $i \in \{1, 2, ..., 200\}$ is the number of selected channels and $j \in \{1, 2, ..., 450\}$ is the index of QMDCT coefficients in a channel. The range of variable i depends on the selected frames N which satisfies $i \in [0, 4N]$.

Table 1: Percentages (%) of modified QMDCT coefficients via each high-pass filtering (Bitrate = 128 kbps, W = 4)

M_Q	$M^{\rightarrow}$	$M^{\downarrow}$	$M^{\rightrightarrows}$	$M^{\downdownarrows}$	$A^{\rightarrow}$	$A^{\downarrow}$	$A^{\rightrightarrows}$	$A^{\downdownarrows}$
1.45	2.00	2.82	2.87	4.14	2.17	2.81	2.88	4.14

2.3. Rich High-Pass Filtering Module

Different from many tasks in computer vision, steganalysis focuses on the data change instead of content itself. Compared with the audio content information, the signal introduced by secret messages can be seen as generalized additive high-frequency noise, which is formulated as

$$S_{i,j} = C_{i,j} + M_{i,j} \tag{2}$$

where $C_{i,j}$ and $S_{i,j}$ represent cover and stego signal separately, and $M_{i,j}$ is the signal introduced by hidden messages. i and j are the row and column indexes of the matrix.

Motivated by the spirit of Rich Models [14] in image steganalysis, we deploy rich high-pass filtering module to "enlarge" the trace of signal introduced by hidden messages in order to suppress the impact of audio signal on the steganalysis, and capture the minor modification introduced by secret messages better. In our paper, eight HPFs are used for MP3 steganalysis, including $M^{\rightarrow}$, $M^{\downarrow}$, $A^{\rightarrow}$, $A^{\downarrow}$, $M^{\rightrightarrows}$, $M^{\downdownarrows}$, $A^{\rightrightarrows}$ and $A^{\downdownarrows}$. Each filter of rich HPF module is interpreted respectively as:

$$M^{\rightarrow}_{m,n} = Q_{i,j} - Q_{i,j+1} \tag{3}$$

$$M^{\downarrow}_{m,n} = Q_{i,j} - Q_{i+1,j} \tag{4}$$

$$A^{\rightarrow}_{m,n} = |Q_{i,j}| - |Q_{i,j+1}| \tag{5}$$

$$A^{\downarrow}_{m,n} = |Q_{i,j}| - |Q_{i+1,j}| \tag{6}$$

$$M^{\rightrightarrows}_{m,n} = Q_{i,j} - 2 \times Q_{i,j+1} + Q_{i,j+2} \tag{7}$$

$$M^{\downdownarrows}_{m,n} = Q_{i,j} - 2 \times Q_{i+1,j} + Q_{i+2,j} \tag{8}$$

$$A^{\rightrightarrows}_{m,n} = |Q_{i,j}| - 2 \times |Q_{i,j+1}| + |Q_{i,j+2}| \tag{9}$$

$$A^{\downdownarrows}_{m,n} = |Q_{i,j}| - 2 \times |Q_{i+1,j}| + |Q_{i+2,j}| \tag{10}$$

where matrices $M^{\rightarrow}$, $M^{\downarrow}$, $A^{\rightarrow}$ and $A^{\downarrow}$ are obtained as the first order difference, and $M^{\rightrightarrows}$, $M^{\downdownarrows}$, $A^{\rightrightarrows}$ and $A^{\downdownarrows}$ represent the second order difference. m and n are the indexes of row and column in differential matrices. Indeed, as we can observe from Table 1, the percentage of modified coefficients is enlarged via each high-pass filtering method.

2.4. Removal of Fully Connected Layers

For typical CNNs, Fc layers are placed at the end of the network for classification. However, it is noteworthy that the feature maps are flattened into a one dimensional vector first when being fed into Fc layers. Spatial and structural correlation characteristics can not be utilized adequately. In our proposed network, Conv layers are deployed to replace Fc layers like FCN [15] so that spatial and structural properties of feature maps can be better captured. The comparison of the CNN with Fc layers and the fully Conv network (FCN) is shown in Fig. 3. Different from the filter kernels in first five Conv blocks, when the dimension of input data is 200×450, filters in Conv-6 block is the same size with feature maps, thus the dimension of output feature maps is 1×1. Finally, output feature maps of Conv-8 are reshaped as a vector of length 2 for the final classification.

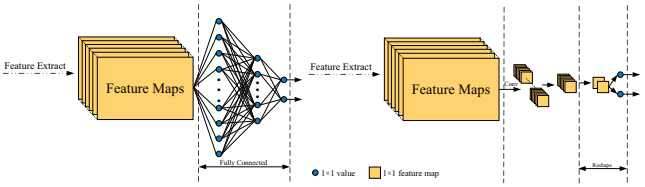

Fig. 3: Comparison between CNN with Fc layers and FCN.

What's more, for previous steganalytic networks like WASDN, if the dimension of the input data mismatches the parameters of trained models, the input data need cropping first and then are fed into the network. If the left size does not satisfy with the cropping rule, the data will be removed directly which leads to the loss of useful information. However, with the removal of Fc layers and the employment of GMP layer, steganalysis of MP3 with size mismatch can be dealt with. The GMP layer is used to resize feature maps to the size of 1×1, if the input data is larger than 200×450.

3. EXPERIMENTS

3.1. Experimental Settings

To evaluate the performance of our proposed network, a dataset which containing 33038 stereo WAV audio clips with a sampling rate of 44.1 kHz and duration of 10s is constructed. The WAV audios are encoded into MP3 files with two common bitrates of 128 kbps and 320 kbps. We carry out experiments using RHFCN to detect two typical MP3 steganographic algorithms: Huffman Codes Mapping (HCM) [16] and Equal Length Entropy Codes Substitution (EECS) [13].

For the non-adaptive steganographic algorithm HCM, the embedding capacity is measured with the relative embedding rate (RER) of 0.1, 0.3 and 0.5. While Syndrome-Trellis Code (STC) [17] is introduced in EECS algorithm, we use the variable W to represent the embedding capacity. W is the constraint width of parity-check matrix which satisfies $W = 1/\alpha$, and α is the relative payload. The secret information is embedded with W of 2, 3, 4, 5, 6 and the constraint height of parity-check matrix is fixed at 7. In every epoch,

Table 2: Description and detection accuracy (%) of each network variant. (EECS, Bitrate = 128 kbps, W = 4)

ID	Description of the network	Accuracy
a	Proposed Network – RHFCN	**80.44**
b	Remove rich HPF module	78.13
c	Quit removing Fc layers	79.09
d	Remove rich HPF module and quit removing Fc layers	77.36

19200 pairs are set for training, and the other 12800 pairs are for validation. The rest 1038 pairs are left for test in order to compare the detection performance of the network with other steganalytic algorithms.

For optimization, we use Adam [18] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The network is trained with an initial learning rate of 10^{-3}, and we use exponential decay function with a decay rate of 0.9 and decay steps of 10^5. The batch size of each iteration is set to 16 (8 cover/stego pairs). To further reduce the danger of over-fitting, we introduce the L_2 regularization with the gain of 10^{-3}.

3.2. Analyses of Network Structure

In this part, we would like to present the benefits of rich high-pass filtering module and the structure of fully CNN. Thus, we conduct experiments by fine-tuning each module separately. All experiments are conducted to detect EECS algorithm with the bitrate of 128 kbps and W of 4. The results are listed in Table 2. As we can see from the results, the modification of the network structure indeed improves the detection performance, especially the introduction of rich HPF module.

3.3. Steganalysis of MP3 with Size Mismatch

Now, we would like to elaborate the impact of input data size on the final detection performance. We train the network with the input data of 200×450 first, and apply the trained model to the MP3 steganalysis with the size of 230×450, 200×480 and 230×480. According to the experimental results shown in Table 3, the detection performance of the network can be maintained better. And, the accuracies drop more with the width increase than height. It is because that more useless values for steganalysis are introduced due to the increase of width dimension, which reduces the difference between cover and stego.

Table 3: Detection accuracy (%) of MP3 with size mismatch (Bitrate = 128 kbps, W = 4)

Size	**200×450**	230×450	200×480	230×480
Accuracy	**80.44**	78.22	77.07	75.53

Table 4: Detection accuracy (%) of HCM algorithm

Bitrate	RER	RHFCN	WASDN	MDI2	ADOTP
128	0.1	**87.18**	83.71	58.48	56.84
	0.3	**92.77**	88.05	68.11	65.13
	0.5	**95.18**	93.34	80.35	74.95
320	0.1	**98.84**	93.26	82.45	68.21
	0.3	**99.23**	94.99	88.44	80.44
	0.5	**99.51**	98.27	93.55	88.34

Table 5: Detection accuracy (%) of EECS algorithm

Bitrate	W	RHFCN	WASDN	MDI2	ADOTP
128	2	**93.26**	90.08	68.79	68.30
	3	**87.96**	82.17	60.79	60.30
	4	**80.44**	74.37	57.71	56.74
	5	**74.76**	64.55	54.72	54.34
	6	**68.50**	55.97	52.02	51.54
320	2	**98.46**	95.57	76.59	73.41
	3	**95.57**	90.17	66.86	61.66
	4	**88.63**	80.15	61.75	57.03
	5	**83.23**	72.54	58.86	54.82
	6	**78.71**	66.67	54.24	53.28

3.4. Comparison with Existing Steganalytic Algorithms

For a comprehensive assessment of the network performance, two state-of-the-art handcrafted features (ADOTP and MDI2) and a CNN-based steganalytic algorithm are compared with RHFCN. The results of each algorithm are shown in Table 4 and 5. Compared with handcrafted features with ensemble classifier [19], the detection of MP3 steganalysis via CNN-based algorithms brings a significant boost. Compared with WASDN, more effective potential properties can be extracted due to the introduction of rich HPF module and the improvement of the network structure, which leads to a boost on the MP3 steganalysis. Experiments demonstrate that our network outperforms in various bitrates, algorithms and payloads.

4. CONCLUSIONS

In this paper, we propose a novel CNN named RHFCN for MP3 steganalysis. For one thing, rich HPF module "enlarges" the traces of signal introduced by secret messages, so that the network is more sensitive to the existence of stego signal. For another thing, the design of fully CNN structure does not only improve the performance of the network due to the utilization of spatial and structural correlation of feature maps, but also contributes to the steganalysis of MP3 with size mismatch. What's more, RHFCN leads to a great boost on the detection of EECS at low relative payloads.

5. REFERENCES

[1] Guanshuo Xu, Hanzhou Wu, and Yunqing Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Processing Letters*, vol. 23, pp. 708–712, 2016.

[2] Yinlong Qian, Jing Dong, Wei Wang, and Tieniu Tan, "Learning and transferring representations for image steganalysis using convolutional neural network," in *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, pp. 2752–2756.

[3] Mo Chen, Vahid Sedighi, Mehdi Boroumand, and Jessica Fridrich, "JPEG-phase-aware convolutional neural network for steganalysis of JPEG images," in *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec 2017, Philadelphia, PA, USA, June 20-22, 2017*, pp. 75–84.

[4] Jian Ye, Jiangqun Ni, and Yang Yi, "Deep learning hierarchical representations for image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2545–2557, 2017.

[5] Jishen Zeng, Shunquan Tan, Bin Li, and Jiwu Huang, "Large-scale JPEG image steganalysis using hybrid deep-learning framework," *IEEE Transactions Information Forensics and Security*, vol. 13, pp. 1200–1214, 2018.

[6] Mehdi Yedroudj, Frédéric Comby, and Marc Chaumont, "Yedroudj-net: An efficient CNN for spatial steganalysis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pp. 2092–2096.

[7] Rongshan Yu, Xiao Lin, Susanto Rahardja, and Chichung Ko, "A statistics study of the MDCT coefficient distribution for audio," in *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, 27-30 June 2004, Taipei, Taiwan*, pp. 1483–1486.

[8] Mengyu Qiao, Andrew Sung, and Qingzhong Liu, "Feature mining and intelligent computing for MP3 steganalysis," in *International Joint Conferences on Bioinformatics, Systems Biology and Intelligent Computing, IJCBS 2009, Shanghai, China, 3-5 August 2009*, pp. 627–630.

[9] Mengyu Qiao, Andrew Sung, and Qingzhong Liu, "MP3 audio steganalysis," *Information Sciences*, vol. 231, pp. 123–134, 2013.

[10] Chao Jin, Rangding Wang, and Diqun Yan, "Steganalysis of MP3Stego with low embedding-rate using Markov feature," *Multimedia Tools and Applications*, vol. 76, pp. 6143–6158, 2017.

[11] Yanzhen Ren, Qiaochu Xiong, and Lina Wang, "A steganalysis scheme for AAC audio based on MDCT difference between intra and inter frame," in *Digital Forensics and Watermarking - 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23-25, 2017*, pp. 217–231.

[12] Yuntao Wang, Kun Yang, Xiaowei Yi, Xianfeng Zhao, and Zhoujun Xu, "CNN-based steganalysis of MP3 steganography in the entropy code domain," in *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, Innsbruck, Austria, June 20-22, 2018*, pp. 55–65.

[13] Kun Yang, Xiaowei Yi, Xianfeng Zhao, and Linna Zhou, "Adaptive MP3 steganography using equal length entropy codes substitution," in *Digital Forensics and Watermarking - 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23-25, 2017*, pp. 202–216.

[14] Jessica Fridrich and Jan Kodovský, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 868–882, 2012.

[15] Evan Shelhamer, Jonathan Long, Trevor Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[16] Diqun Yan, Rangding Wang, and Liguang Zhang, "A high capacity MP3 steganography based on Huffman coding," *Journal of Sichuan University (Natural Science Edition)*, vol. 6, pp. 1281–1286, 2011.

[17] Tomás Filler, Jan Judas, and Jessica Fridrich, "Minimizing embedding impact in steganography using trellis-coded quantization," in *Media Forensics and Security II, part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, January 18-20, 2010, Proceedings*, 2010, p. 754105.

[18] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *CoRR, abs/1412.6980*, 2014.

[19] Jan Kodovský and Jessica J. Fridrich and Vojtech Holub, "Ensemble Classifiers for Steganalysis of Digital Media," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, 2012.