STEGANALYSIS OF ADAPTIVE JPEG STEGANOGRAPHY BASED ON RESDET

Xiaosa Huang, Shilin Wang, Tanfeng Sun, Gongshen Liu and Xiang Lin*

School of Electric Information and Electronic Engineering, Shanghai Jiaotong University

ABSTRACT

With the development of the adaptive JPEG steganography, steganalysis has become much more difficult in recent years. In order to detect the adaptive JPEG steganography, a CNN based framework, i.e. the ResDet, is proposed in this paper, which is sensitive to the artifacts caused by the adaptive JPEG steganography. To avoid the influences caused by various image content, the JPEG image under investigation is preprocessed by being passed through a series of filters. Then the feature maps are put into multiple convolutional layers. Contributing to the combination of the shortcut connection and the dense connection, the proposed network can differentiate JPEG steganography artifacts accurately with much more compact feature. Experiment results on BossBase with J-UNIWARD have demonstrated that the proposed framework with 84-dimensional feature, which will remarkably improve the efficiency of steganalysis, outperforms several state-of-the-art approaches investigated.

INDEX TERMS Adaptive steganography; JPEG; steganalysis; CNN.

1. INTRODUCTION

Since JPEG is the most widely used format in image communication and storage, the steganography and steganalysis on JPEG images has become an import branch in information hiding. The existing JPEG steganography algorithms can be roughly divided into two categories based on whether the image content characteristics is considered: the non-adaptive steganography approaches and the adaptive steganography approaches. And the adaptive approaches have attracted more research interests in recent years because they are usually more difficult to detect.

In order to detect the adaptive steganography effectively, many researchers have proposed various methods, among which algorithms based on Rich Model is a hot research direction [1-4]. In [1], an 8000-dimensional DCTR (discrete cosine transform residual) feature is proposed, which uses only first-order statistics of noise residuals obtained with DCT bases as the pixel predictor kernels. The historical overview showed that high-dimensional feature can better capture statistical dependencies. Following this direction, the same authors proposed PHARM (PHase-Aware pRojection Model) in [2], in which residuals are represented using first-order statistics of their random projections. PHARM showed better performance, while at the expense of a higher dimension (12600). In [3], GFR (Gabor Filter Rich Model) is proposed to filter image with a set of 2D Gabor filters, which can capture feature from different scales and orientations. The SCA-GFR presented in [4] is a selective channel-aware variant of the JPEG Rich Models above. However, the computation costs on feature extraction with these methods are too high for practical application.

Recently, considering the superior performance of CNNs in image classification and understanding, some researchers have adopted CNN-based approaches in JPEG steganalysis. Zeng et al. [5] designed a hybrid CNN model for large-scale JPEG steganalysis. Three convolutional layers are adopted to make their model effective on ImageNet. Experiment results on 5 million images have demonstrated that their CNN outperformed the traditional feature-based methods in detecting JPEG steganography with large-scale dataset. In [6], A JPEG-Phase-Aware CNN model is proposed, in which four 5×5 high pass filters including Gabor filter were adopted in preprocessing phase. In PhaseSplit module, the input feature maps with size of 512×512 are split into 64 64×64 feature maps. Experiment results illustrated that the detection performance can be improved by splitting JPEG phase. In [7], Xu proposed a CNN model with 20 convolutional layers. He concluded that deep CNN can better capture steganography artifacts. In order to solve the problem of gradient vanishing, shortcut connection [8,9] is introduced. Experiment results showed that this method can get better results than traditional methods. However, there are still much redundancy in those CNN based features, which means most dimensions of the extracted feature are useless.

In this paper, a CNN based network named ResDet has been proposed to detect adaptive JPEG steganography. The major contributions of our approach can be summarized as follows: i) a new network structure is proposed to detect the adaptive JPEG stegonagraphy where the residual connection and dense connection are combined to better extract traces left by steganography; ii) the feature extracted by ResDet is much more compact than the other methods, including handcrafted features and the traditional CNN based features; iii) the experiment results have demonstrated that the pro-

Xiaosa Huang, Shilin Wang*, Tanfeng Sun, Gongshen Liu and Xiang Lin are with School of Electric Information and Electronic Engineering, Shanghai Jiao- tong University, 200240, Shanghai, China (e-mail: 070@sjtu.edu.cn, wsl@sjtu.edu.cn, tfsun@sjtu.edu.cn, lgshen@sjtu.edu.cn, lionel@sjtu.edu.cn). *Corresponding author

posed framework performed better than state-of-the-art methods when embedding rate is high.

2. PREPROCESSING

To avoid the influences caused by various image content, the following preprocessing steps are adopted. Considering JPEG steganography modifies the DCT coefficients, specific embedding artifacts will be introduced in the stego images. In a decompressed JPEG image, the statistical characteristics of pixels are not spatially invariant, which means that each coefficient depend on its neighborhood within the JPEG 8 × 8 pixel grid. It thus makes sense to collect statistical characteristic of pixels separately for each phase (i, j), $0 \le i, j \le 7$.

Inspired by Xu [7], the input image in the JPEG format are first decompressed to the spatial domain. Then the decompressed image is convolved by sixteen DCT basis patterns $B^{(k,l)} = \left(B_{mn}^{(k,l)}\right), 0 \le m, n \le 4$ to make the CNN architecture concentrate on the steganography artifacts rather than the image contents.

$$B_{mn}^{(k,l)} = \frac{w_k w_l}{4} \cos\left(\frac{k\pi(2m+1)}{8}\right) \cos\left(\frac{l\pi(2n+1)}{8}\right) \qquad (1)$$

where $w_0 = 1/2, w_k = 1$ for $k > 0$. Finally, the filtered image is quantized with a truncation threshold of 8. Images before and after filtering is illustrated in Fig.1.

Fig. 1 Images before and after filtering through one of the filters. ul is origin image, ur is filtered origin image, dl is filtered stego image, dr is difference of two filtered images.

3. THE ARCHITECTURE OF RESDET

3.1 The overall structure

The overall architecture of the proposed ResDet is illustrated in Fig.2. We take the BossBase [13] image with size of 512×512 as an example to illustrate the change of feature dimension of each layer.

The input of ResDet is the filtered and truncated image with shape of (511,511,16), since 16 DCT basis patterns are used to convolve the decompressed image. The backbone of ResDet is composed of three resent-like blocks (RBLOCK) and three densenet-like blocks (DBLOCK) alternately.

RBLOCK contains two branches: the body branch and the shortcut branch. Two convolutional layers are included in the body branch, the kernel size of which is 3×3. Each convolutional layer is followed by Batch-Normalization (BN) to reduce internal covariant shift [10] and the most widely used Rectified Linear Unit (ReLU) [11] is adopted as the nonlinear activation function. The convolutional layer in the shortcut path is used to resize the input x to a different dimension to match that of the body path. To reduce the activation dimension's height and width by a factor of 2, the stride of the convolutional layer is set to be 2, like the second convolutional layer in the body branch. The two branches are connected by the operation of "add". It is noteworthy that the number of the output feature maps increases by 12 through each RBLOCK.

DBLOCK contains one convolutional layer, followed by a BN layer and a Relu layer. The size of the convolution kernel is 3×3 and the number of output feature maps is fixed to 12. The learned feature maps are concatenated with the input feature maps by the operation of "concatenate", which is referred to as the dense connection in DenseNet [12]. In such a way, features learned by the previous RBLOCK and the DBLOCK can be passed into the next stage.

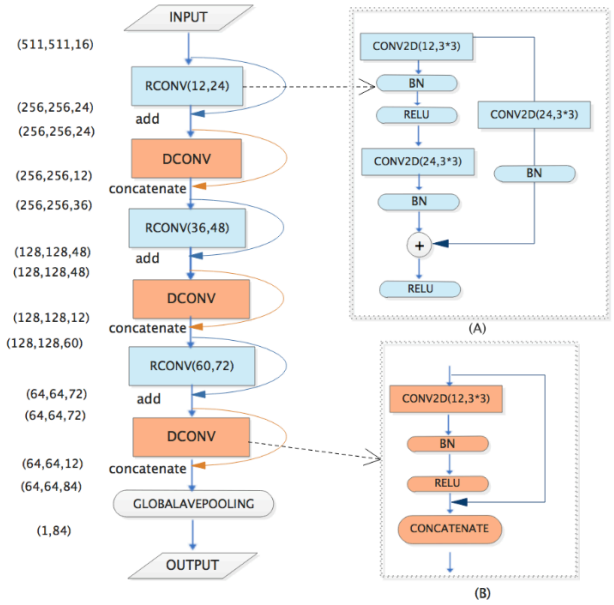

Fig. 2 Architecture of the proposed dense CNN.

A global average-pooling is then performed in which spatial average of each feature map is calculated, so that the output is more robust to the spatial diversity of image contents and has better generalization ability. And the softmax layer is adopted for classification.

3.2 Characteristic of the proposed structure

Different from CNNs in computer vision, the proposed ResDet is well designed to extract traces left by JPEG steganography. Taking advantage of residual connection and dense connection, the feature extracted from the filtered image can propagate more effectively as well as be better exploited. The major characteristics of our network are:

1)Residual connection:

Inspired by Xu [7], transformed residual connection is also adopted in ResDet to overcome the gradient vanishing problem.

2)Dense connection:

When steganography traces pass through many layers, some discriminative information may be lost. By dense connection, features extracted by early layers are directly used by deep layers, which makes the output features possess a hierarchical representation. Therefore, the artifacts caused by JPEG steganography can be depicted more comprehensively. Moreover, shallow and deep fields can be combined more freely, and the dependencies between different layers are smaller, making the model more compact and robust. It can take advantage of the low complexity of shallow features, making it easier to get a smooth decision function with better generalization performance.

3)The combination of two connections:

Summing up the output of two branches in residual connection may block the information flow if the feature maps of the two layers have very different distributions. Therefore, the cascading feature map (in dense connection) can preserve all feature maps and increase the variance of the output, thus encouraging feature reuse.

Furthermore, there are feature redundancy in residual connection. Each RBLOCK only extracts very few features (so-called residuals). While in dense connection, the latter layer has direct access to feature maps from the previous layer, which means that there is no need to re-learn redundant feature maps. In this way, the final extracted feature can be more compact.

However, since DBLOCK concatenates feature maps of its two branches, the number of the output feature maps increases while the size of the output feature maps keep unchanged, which account for the increase of network parameters. Thus, each DBLOCK is followed by a RBLOCK, which reduces the feature map dimension's height and width by a factor of 2, like what pooling layer does. In this way, main features from DBLOCK can be remained, parameters(dimensions) can be reduced at the same time. Moreover, it can also prevent overfitting and improve the generalization ability of the model.

4. EXPERIMENTS AND DISCUSSIONS

4.1 Experiment Setups

The BOSS dataset [13] containing 10,000 images with the size 512×512 is used to evaluate the effectiveness of ResDet. All these images are firstly JPEG compressed with a specific quality factor(QF) to generate negative-class samples. QF75 and QF95 are selected as representatives for low and high quality. J-UNIWARD is chosen as the steganography algorithm since it is the most secure method we investigated. The corresponding positive-class samples were generated through information embedding by J-UNIWARD with embedding rates of 0.1, 0.2, 0.3, and 0.4 bpnzAC. Similar to Xu [7], 50% of the origin and stego images are randomly selected as the training samples, and the remaining samples are used for validation/testing.

In order to make a fair comparison, transfer learning strategy is not adopted in the following experiments, which means CNN model is trained from randomly initialized weights over each embedding rate.

In all the following experiments, the detection accuracy is adopted as a judgement criterion, which is defined as,

$$\mathrm{ACC} = 1 - (\mathrm{FAR} + \mathrm{FRR})/2 \quad (2)$$

where FAR is the rate of origin images misclassified as stego ones and FRR is the rate of stego images misclassified as origin ones.

4.2 Training Process

Mini-batch gradient descent is applied to train the CNN model in all experiments. A batch size of 32 samples is chosen for weight and bias updates. The learning rate is initialized with 0.001, and the value of learning rate is reduced to 0.1 times 10 epochs after which there is no improvement of validation loss. CNN is trained for 120 epochs. For each convolutional layer, the parameters in the convolution kernels were randomly initialized from zero-mean Gaussian distribution and bias learnings were disabled. In the last dense layer, parameters were initialized using 'glorot_uniform' and weight decay was enabled. Keras is used as the deep learning framework.

Fig.3 Effect of Number of Epochs on Model Accuracy, QF=75, embedding rate=0.4 bpnzAC

To visualize the learning progress of ResDet, Fig.3 is depicted for demonstrating the accuracies for the training, validation by taking the QF=75 and embedding rate 0.4 for example. As shown in the figure, the two curves share similar convergent trend, which means the model is sensitive to the artifacts caused by the adaptive JPEG steganography.

4.3 Performance Comparisons with State-of-the-art Approaches

In this section, the detection performance on BOSS dataset [13] for J-UNIWARD with payloads ranging from 0.1 bpnzAC to 0.4 bpnzAC are given. Besides the proposed algorithm, two state-of-the-art CNN based approaches, including Xu's [7] and Chen et al.'s [6] and the conventional steganalysis methods SCA-GFR [4], are employed for comparison. Experiment results are showed in Table 1 and Table 2.

Table 1: Experiment results on BOSS with QF75

	Embedding rates(bpnzAC)				
	D	0.1	0.2	0.3	0.4
SCA-GFR	17000	0.6402	0.7684	0.8591	0.9193
CHEN	512	0.6425	0.7874	0.8772	0.9344
XU	384	**0.6717**	**0.8053**	0.8876	0.9359
RESDET	**84**	0.6523	0.7901	**0.8991**	**0.9581**

Table 2: Experiment results on BOSS with QF95

	Embedding rates(bpnzAC)				
	D	0.1	0.2	0.3	0.4
SCA-GFR	17000	**0.5371**	0.6002	0.6697	0.738
CHEN	512	0.515	0.5966	0.6785	0.7452
XU	384	0.511	**0.6025**	0.6894	0.7636
RESDET	**84**	0.512	0.5876	**0.7030**	**0.7853**

From the table, it can be observed that the extracted feature of ResDet is more compact than the others. The dimension of previous CNN based feature is at least 4x than that of ours, while Rich Model feature is at least 200x than ours. Because of the introduction of residual connection and dense connection, the features caused by JPEG steganography are reused more effectively, which makes the network more concise. With low dimension feature, our ResDet framework achieves an average 2% accuracy gain in high embedding rate(such as 0.4 and 0.3 bpnzAC). It can be explained as follows. Minor traces are left during adaptive JPEG steganography. Statistical features are too homogeneous to cope with the fluctuated distribution well. In such case, the CNN-based feature, with the hierarchical representation of the filtered image, can still extract useful information to differentiate the two classes. Moreover, with the mutual complement of residual connection and dense connection, feature with low dimension can characterize steganography traces adequately. However, when embedding rate is low (such as 0.1 bpnzAC), the performance of ResDet is slightly worse than the model of Xu. This is mainly because there are few differences between stego and origin images when embedding rate is low. Feature with low dimension may not well characterize these subtle differences.

5. CONCLUSIONS

In this paper, a novel CNN structure(ResDet) for detecting adaptive JPEG steganography is presented. Using high pass filtered images as the input, 84-D feature is extracted by ResDet which can differentiate JPEG steganography images. The most notable characteristic of ResDet is the combination of residual connection and dense connection, which makes the network more concise, as well as more sensitive to traces of JPEG steganography. For the experiment results, the proposed method is shown to outperform the state-of-the-art methods in most embedding rates.

6. ACKNOWLEDGMENT

The work described in this paper is supported by National Key Research and Development Program of China NO. 2016QY03D0604, NSFC Fund No. 61771310, Shanghai STCSM Fund No. 18511105902.

7. REFERENCES

[1] V. Holub and J. Frydrych. Low-complexity features for jpeg steganalysis using undecimated dct. IEEE Transactions on Information Forensics and Security, 10(2):219–228, Feb 2015.

[2] Vojtech Holub and Jessica J Fridrich. Phase-aware projection model for steganalysis of jpeg images. In Media Watermarking, Security, and Forensics, page 94090T, 2015.

[3] Xiaofeng Song, Fenlin Liu, Chunfang Yang, Xiangyang Luo, and Yi Zhang. Steganalysis of adaptive jpeg steganography using 2d gabor filters. In Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security, pages 15–23. ACM, 2015

[4] Toma´sˇ Denemark Denemark, Mehdi Boroumand, and Jessica Fridrich. Steganalysis features for content-adaptive jpeg steganography. IEEE Transactions on Information Forensics and Security, 11(8):1736–1746, 2016.

[5] Jishen Zeng, Shunquan Tan, Bin Li, and Jiwu Huang. Large-scale jpeg steganalysis using hybrid deep-learning framework. arXiv preprint arXiv:1611.03233, 2016.

[6] Mo Chen, Vahid Sedighi, Mehdi Boroumand, and Jessica Fridrich. Jpeg-phase-aware convolutional neural network for steganalysis of jpeg images. In 5th ACM Workshop Inf. Hiding Multimedia Secur. (IH&MMSec), 2017.

[7] Guanshuo Xu. Deep convolutional neural network to detect J-UNIWARD. In 5th ACM Workshop Inf. Hiding Multimedia Secur. (IH&MMSec), 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. 2016. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. 2016. Identity mappings in deep residual networks. In Proceeding of European Conference on Computer Vision (Oct. 2016). 630-645.

[10] Boureau Y L, Ponce J, Lecun Y. “A Theoretical Analysis of Feature Pooling in Visual Recognition,” Proceedings of the 27th International Conference on Machine Learning, pp. 111-118, 2010.

[11] S. Ioffe, C. Szegedy, “Batch normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” Proceedings of the 32nd International Conference on Machine Learning, France, pp. 448–456, 2015.

[12] G. Huang, Z. Liu, L. v. d. Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 2261-2269, 2017.

[13] Patrick Bas, Tomáš Filler and Tomáš Pevný. “Break our steganographic system – the ins and outs of organizing BOSS,” In Proceeding of 13th International Conference, Prague, Czech Republic, 2011.