Convolutional Neural Network-based Steganalysis on Spatial Domain

Dong-Hyun Kim, and Hae-Yeoun Lee

***Abstract*—**Steganalysis has been studied to detect the existence of hidden messages by steganography. However, human intervention is required to determine the flaws of the steganography and it is time consuming. This paper presents a steganalysis method using a convolutional neural network for spatial domain steganography whereby there is no human intervention is required. We have designed a convolutional neural network-based steganalysis model to have 5 convolutional layers and 2 full connected layers. Especially, binarized differential filter and high pass filter are applied to extract hidden messages. After the model is trained with cover images and LSB-based stego-images, unknown images are tested to determine if secret messages have been embedded. Experiments are performed using BOSS and SIPI database and the presented models show over 99% and 96% accuracy for stego-images with the same key and different keys.

***Keywords*—**Steganalysis, convolutional neural network, binarized differential filter, high pass filter, LSB steganography.

I. Introduction

STEGANOGRAPHY is the science of hiding information whose goal is to hide even the existence of secret messages in an innocent-like cover image, called the stego-image. Against steganography, the steganalysis is the science of detecting the existence of hidden messages in the stego-image [1].

By investigating the flaws of each steganography method, steganalysis researchers have studied to design steganalysis methods to defeat steganography. As a result, efficient steganalysis methods which are specific to each steganography method have been developed. However, these approaches have required human intervention to determine their flaws and its time consuming to analyze steganography methods. Also, universal steganalysis methods applicable to all kinds of steganography methods are still required to be studied.

Recently, the interest about deep learning has increased and many remarkable results are emerging in image and video processing applications. As a result, steganalysis researchers attempt to apply deep learning to detect the existence of the secret messages in the stego-images, which does not require human intervention.

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A1B03030432).

Dong-Hyun Kim is with Kumoh National Institute of Technology, 61 Daehak-ro, Gumi, Gyeongbuk, 39177, South Korea.

Hae-Yeoun Lee is with Kumoh National Institute of Technology, 61 Daehak-ro, Gumi, Gyeongbuk, 39177, South Korea (corresponding author to provide phone: +82-54-478-7549; e-mail: haeyeoun.lee@kumoh.ac.kr).

In this paper, we present a convolutional neural network-based steganalysis model against spatial domain steganography. The steganalysis model has 5 convolutional layers and 2 full connected layers. Also, two pre-processing filters are designed to extract hidden messages in the input image. After the model is trained with cover (original) images and stego-images, unknown images are tested to decide the existence of the secret message. Experiments are preformed using BOSS and SIPI databases and over 99% and 96% accuracy are achieved for LSB-based stego-images with the same key and the different keys.

The paper is organized as follows. Sec. II reviews related works, the presented steganalysis model is explained in Sec. III and experimental results are shown in Sec. IV and Sec. V concludes.

II. Related Works

Depending on the embedding domain, the steganography method can be categorized as spatial domain methods or frequency domain methods. In the spatial domain, secret messages are inserted at the pixel level, which affects statistical characteristics of pixel values. In the transform domain, secret messages are inserted by modifying coefficient values after transformation such as discrete cosine transform, discrete wavelet transform, or discrete Fourier transform.

Against steganography methods, steganalysis researches have studied to prevent the transmission of secret messages using the innocent-like cover. However, it is difficult to deal with any steganography methods by one steganalysis method called as universal steganalysis. Therefore, the target specific steganalysis methods have studied and performed well. However, there are limitations to determine the flaws of each steganography method with human intervention and it is time consuming to consider recent steganography methods.

Nowadays, deep learning has a great attention because of its remarkable results in many application fields. Especially, it can define features or patterns from big data without human intervention. In steganalysis researches, deep learning is studied to be applied for getting accurate results.

Sedighi and Fridrich have designed a convolutional neural network (CNN) model to analyze the features of images. After extracting features using a projection spatial rich model (PSRM) technology, the optimization filter is applied to minimize the number of projection kernels to optimize the deep learning model [2].

Without the preprocessing of the image, Bayar and Stamm use a local structural relationship between pixels. A prediction filter is used to predict the pixel value at the center of the filter and then difference values between the original and predicted pixel values are used in CNN based deep learning model [3].

Qian et al has designed a customized CNN model for steganalysis. The center pixel is estimated using neighboring pixels and difference values between the original and estimated pixel values are used for this CNN based deep learning model. Also, the amplitude of Gaussian output is limited by using non-linear activation function [4].

Using a large convolution filter, Salomon et al. has conducted a steganalysis research against steganography with the same embedding key [5].

These steganalysis researches using deep learning are at the initial stage and continuous researches are required to get satisfactory results in accuracy.

III. Proposed Deep Learning-Based Steganalysis

A universal steganalysis against all steganography methods is not feasible to design. We focus on a target specific steganalysis and present a convolutional neural network-based steganalysis method against spatial domain steganography.

The presented steganalysis model consists of two steps: training and testing. During the training step, the model is trained with cover images and stego-images. During the testing step, unknown images are tested whether secret messages are embedded.

A. Deep Learning

Deep learning is a new representation for the neural network which has been studied for a long time, not a new technology. With advances in computer hardware technology, the deeper levels of neural network can be computed efficiently and remarkable performances are achieved.

The most commonly used neural networks are Deep Neural Network (DNN), Convolutional Neural Network, Recurrent Neural Network (RNN), Restricted Boltzmann Machine (RBM), and Deep Belief Network (DBN). DNN, CNN, and RNN are the most frequently used methods.

DNN is modeled by a complex nonlinear relationship, whereby each object is hierarchically represented. The lower layer is a method of modeling complex data through a small number of units by integrating the features of the progressively gathered previous layers. CNN is a model composed of one or more neural networks with multiple product layers. The composite product layer consists of convolution, pooling, activation, etc., and can use 2-D data as input data. Therefore, it is applied in the fields of video and voice processing. RNN is a neural network that simultaneously considers current input data and past input data. Feedback loop is a deep neural network in which the current output value is used as the input value of the next layer. Previous results, such as time series data, are used to analyze data that affect subsequent outcomes.

Since steganalysis is to detect secret messages in images, we have adapted the CNN-based deep learning model. Also, the CNN-based deep learning model focuses on the recognition based on image contents. Hence, to expose the hidden message in the innocent-like cover, two filters have been adapted in the proposed steganalysis model such as; high pass filter and binarized differential filter.

B. High Pass Filter

Since embedding secret messages into a cover image distorts the locality of adjacent pixel values, the secret messages can be regarded as noise. To extract this noise, a high pass filter (HPF) as follows is applied to the input image [6]. Usually, convolutional layers without this filter have a limitation to extract the noise embedded in an image.

$$\mathrm{HPF} = \frac{1}{12}\begin{pmatrix} -1 & +2 & -2 & +2 & -1 \\ +2 & -6 & +8 & -6 & +2 \\ -2 & +8 & -12 & +8 & -2 \\ +2 & -6 & +8 & -6 & +2 \\ -1 & +2 & -2 & +2 & -1 \end{pmatrix} \quad (1)$$

C. Binarized Differential Filter

Generally, natural images have locality characteristics which mean that adjacent pixels have a high possibility to have the same value. Since spatial domain steganography methods modify pixel values to insert secret messages, the possibility to have the same values between adjacent pixels will be lower. Therefore, this paper presents the design of binarized differential filter (BDF) to calculate the binary value of pixel differences as follows; adjacent pixels in x direction are considered in the proposed model. However, it can be extended to consider the y direction and diagonal directions.

$$\mathrm{BDF} = \begin{cases} 1 & f(x-1,y) \le f(x,y) \\ -1 & otherwise \end{cases} \quad (2)$$

D. Convolutional Neural Network-based Steganalysis

As shown in Fig. 1, to detect the existence of the secret message, our CNN-based steganalysis model is designed to have 5 convolutional layers and 2 fully connected layers. Also, high pass filter or binarized differential filter is included to extract the secret message.

To extract features for steganalysis, convolutional layer is composed of 5 layers. The 1st layer performs convolution with 64 filters with 7x7 size and stride 2 to minimize the number of node. For the activation, ReLU (rectified linear unit) function is applied and results are normalized with local response normalization. The 2nd layer performs convolution with 16 filters with 5x5 size and stride 1. Then, ReLU and local response normalization are performed. The 3rd layer performs in similar to the 2nd layer. The 4th layer performs convolution with 32 filters with 5x5 size and stride 2 to minimize the number of node. For the activation ReLU function is applied without following local response normalization. For the 5th layer, convolution with 32 filters with 5x5 size and stride 1 is performed. Then, ReLU is applied as the activation function.

In the convolutional layer, pooling is not applied as possible because it has a possibility to remove or minimize the trace of the embedded secret messages.

The fully connected layer is also composed of 2 layers. In the 1st layer, the number of input nodes is 63x63x32 and the number of output nodes is set at 1,000. To overcome over- and under-fitting problems, the dropout rate is set at 0.5. In the 2nd layer, the number of input nodes is 1,000 and the number of output nodes is set at 2 to classify the cover image or the stego-image. Also, the dropout rate is set at 0.5. Finally, a SoftMax function is applied to normalize the possibility of the cover image or the stego-image.

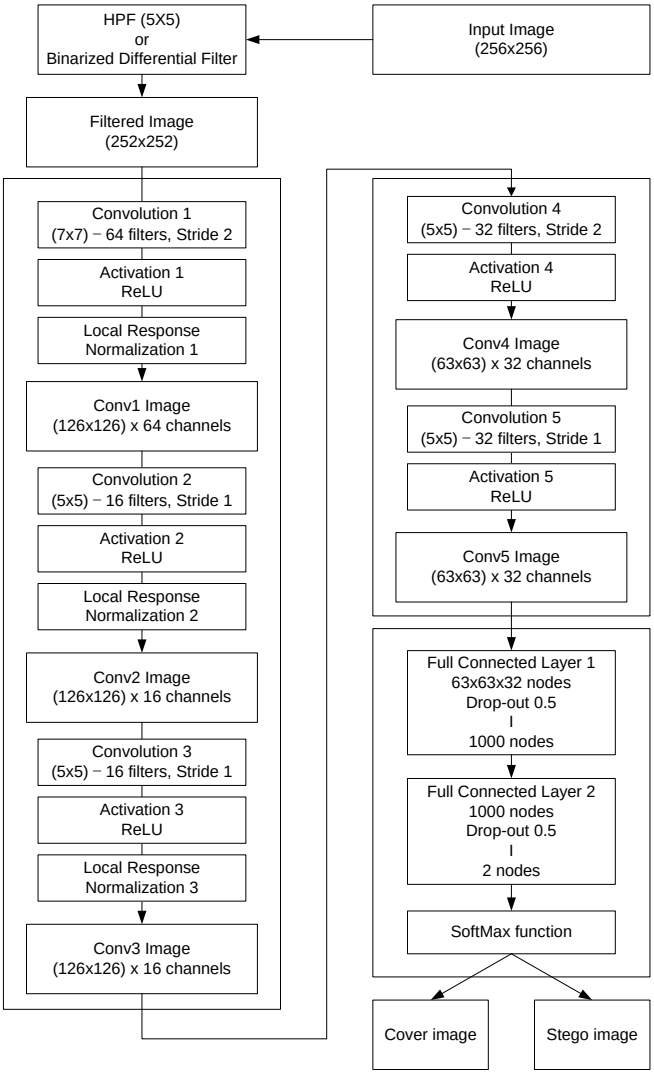


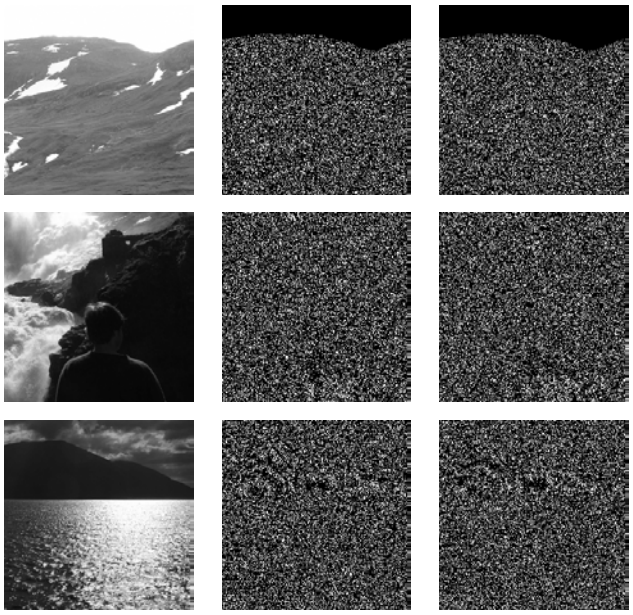
Fig.1. CNN-based steganalysis model

IV. Experimental Results

Experiments are performed on hardware configuration with Intel i7-7700 CPU, 16GB RAM, and nVidia Titan XP graphic card with 11GB memory. 10,000 cover images of 512x512 pixel sizes are collected from well-known database such as BOSS and SIPI. Because of the memory limitation of graphic card, we have clipped their size as 256x256 pixels and hence generated 40,000 images.

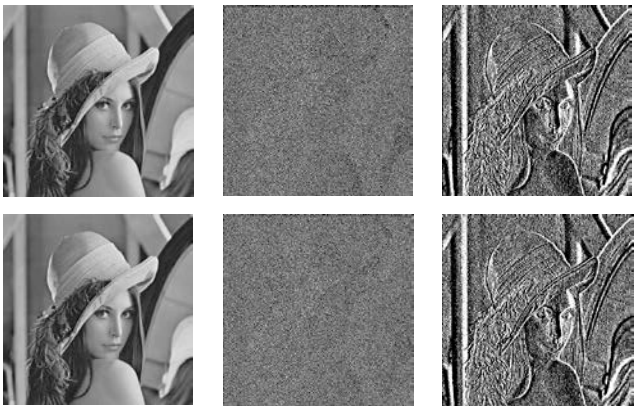
Using a LSB-based spatial domain steganography method, 80,000 stego-images are generated with the same key and the different keys, where 75 percent of images are used for the training of the model and 25 percent of images are used for the testing. Also, training images are repeatedly applied in our presented deep learning model to improve the accuracy.

Fig. 2 shows the sample of cover images, residual images between cover and stego-images with the same key and residual images between cover and stego-images with the different keys. Although the secret message with the same key is embedded, the residual image can be different because it depends on the pixel value of cover images.



(a) (b) (c)

Fig.2. (a) Cover image, (b) residual images with the same key and (c) residual images with the different keys



(a) (b) (c)

Fig.3. (a) Cover image, (b) extracted hidden messages with HPF and (c) extracted hidden messages with BDF

In the proposed steganalysis model, high pass filter and binarized differential filter are applied to extract hidden messages in the innocent-like cover. Fig. 3 shows the sample of

cover image and extracted hidden messages with HPF and BDF. Since the filter is not ideal, the extracted message cannot be exactly matched to the embedded messages, but it can help to extract the pattern of the embedded messages to increase the performance of the steganalysis model.

Using the BDF, Fig. 4 shows the accuracy of the CNN-based steganalysis model for the stego-images with the same key. Fig. 5 shows the accuracy of the CNN-based steganalysis model for the stego-images with the different key. X axis is the number of learning for each training data set (epoch) and Y axis is the detection accuracy.

Using the HPF, Fig. 6 shows the accuracy of the CNN-based steganalysis model for the stego-images with the same key. Fig. 7 shows the accuracy of the CNN-based steganalysis model for the stego-images with the different key.

In order to show the exact performance of the CNN-based steganalysis model, Table 1 summarizes the accuracy at 4,200 epochs.

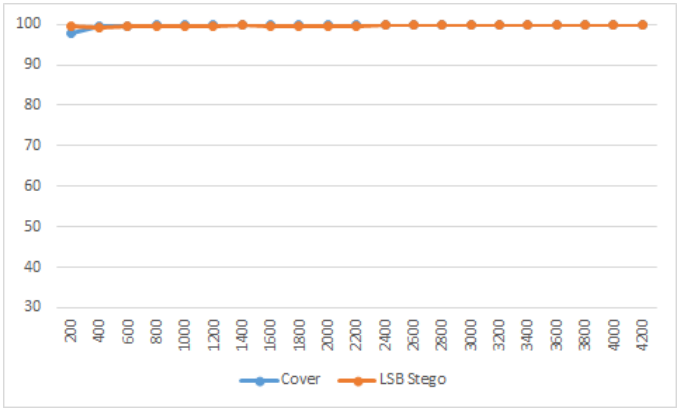


Fig.4. Accuracy for stego images with the same key and the BDF

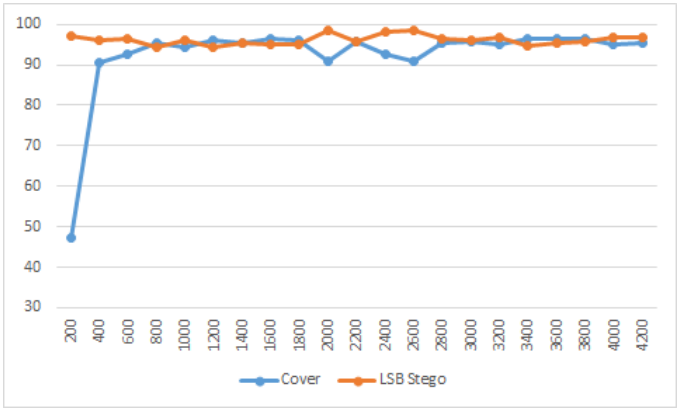


Fig.5. Accuracy for stego images with the different key and the BDF

By referring to these figures and table, we can know that the trends of the accuracy increases when the number of learning increases. For the same key, the detection accuracy of CNN-based steganalysis model is over 99% and the convergence speed is very high. For the different key, the detection accuracy is 96% which is relatively low against the same key. However, by adjusting this CNN-based steganalysis model, there is a possibility to increase the accuracy for the different keys.

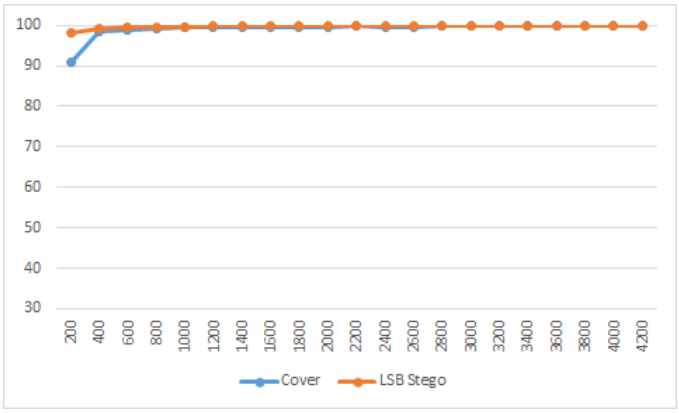


Fig.6. Accuracy for stego images with the same key and the HPF

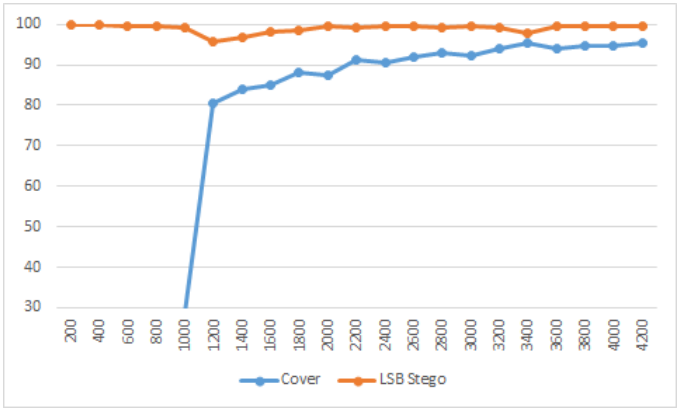


Fig.7. Accuracy for stego images with the different key and the HPF

Table 1. Accuracy of steganalysis model using BDF and HPF at 4,200 epoch

	BDF	HPF
Same Key	99.91	99.84
Different Key	96.01	97.62

V. Conclusion

Steganalysis was studied to detect the existence of hidden messages in the innocent-like cover image. However, previous studies have a limitation to determine the flaws of specific steganography with human intervention and it was crucial to the accuracy.

To defeat spatial domain steganography, this paper presented a CNN-based steganalysis model having 5 convolutional layers and 2 full connected layers. Also, two filters to extract hidden messages are designed and included. The weakness of previous methods requiring human intervention was overcome through deep learning.

Experiments were performed using 40,000 cover and 90,000 LSB stego-images from BOSS and SIPI databases. Promising results for steganalysis were achieved.

There are many possibilities to improve accuracy. Future works will be to increase the depth of deep learning model and tune activation and pooling functions. In addition, filters to enhance the existence of secret messages should be studied.

REFERENCES

[1] J.-C. Joo, H.-Y. Lee, and H.-K. Lee, "Improved Steganographic Method Preserving Pixel-Value Differencing Histogram with Modulus Function," EURASIP Journal on Advances in Signal Processing, vol. 2010, pp. 1--13, June 2010.

[2] V. Sedighi and J. Fridrich, "Histogram Layer, Moving Convolutional Neural Networks Towards Feature-Based Steganalysis," Electronic Imaging, Media Watermarking, Security and Forensics 2017, Jan. 29 2017.

[3] B. Bayar and M. C. Stamm, "A Deep Learning Approach To Universal Image Manipulation Detection Using A New Convolutional Layer," Information Hiding and Multimedia Security 2016, IH&MMSec, pp. 05--10(6), June. 2016.

[4] Y. Qian, J. Dong, W. Wang, and T. Tan, "Deep learning for steganalysis via convolutional neural networks," Proceeding of of IS&T International Syposium on Electronic Imaging: Media Watermarking, Security, and Forensics, vol. 9409, pp. 94090J, 2015.

[5] M. Salomon, R. Couturier, C. Guyeux, J.-F. Couchot, and J. M. Bahi, "Steganalysis via a Convolutional Neural Network using Large Convolution Filters for Embedding Process with Same Stego Key: A deep learning approach for telemedicine," European Research in Telemedicine, vol. 6(2), pp. 79--92, July 2017.

[6] B. Bayar and M. C. Stamm, "Design Principles of Convolutional Neural Networks for Multimedia Forensics," Proceedings of IS&T International Syposium on Electronic Imaging: Media Watermarking, Security, and Forensics 2017, pp. 77--86, Jan. 2017.

Dong-Hyun Kim (B'05) received the BS degree in Computer Software Engineering from Kumoh National Institute of Technology, South Korea, in 2005. Recently, he is pursuing a MS degree in Video and Image Processing Laboratory, Department of Software Engineering, Kumoh National Institute of Technology, South Korea. His research interests include multimedia forensics and information security.

Hae-Yeoun Lee (B'97-M'99-D'06) received the BS degree in information engineering from Sung Kyun Kwan University, South Korea, in 1997 and the MS degree and PhD degree in computer science from Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 1999 and 2006, respectively. From 2006 to 2007, he was a postdoctoral researcher at Weill Medical College, Cornell University, USA. Since 2008, He joined the Department of Computer Software Engineering, Kumoh National Institute of Technology, South Korea where he is currently a regular Professor. His current areas of research interest include multimedia forensics, digital watermarking, video and image processing, remote sensing, and digital rights management.