Deep Learning-based Steganalysis against Spatial Domain Steganography

Dong-Hyun Kim
dept. of software engineering
kumoh national institute of technology
Gumi, Gyeongbuk, Republic of Korea
knight2995@naver.com

Hae-Yeoun Lee
dept. of computer software engineering
kumoh national institute of technology
Gumi, Gyeongbuk, Republic of Korea
haeyeoun.lee@kumoh.ac.kr

***Abstract*—Against steganography to hide secret messages into an innocent-like cover, steganalysis was studied to detect the presence of hidden messages, and steganography flaws were determined by human intervention. In this paper, we present a steganalysis method using deep learning for spatial domain steganography which does not require human intervention. The deep learning-based steganalysis model is designed to have 1 high pass filter, 2 convolutional layers and 2 full connected layers. After being trained with cover images and LSB 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 model shows 98% and 90% accuracy for LSB stego-images with the same key and different keys.**

Keywords—steganalysis, deep learning, LSB steganography

I. Introduction

Steganography is to hide the existence of secret messages by embedding the secret messages into an innocent-like cover. On the contrary, steganalysis is the science of detecting the existence of hidden messages in the embedded image by steganography called the stego-image [1].

Steganalysis researchers try to design steganalysis methods to defeat steganography methods by exposing their flaws. As a result, efficient steganalysis methods which are specific to each steganography method have been developed. However, these previous approaches require human intervention to determine their flaws. Also, universal steganalysis methods applicable to all steganography methods are still required to be studied.

Recently, the interest about deep learning has increased and many remarkable results are emerging. Hence, steganalysis researchers attempt to apply deep learning to detect the existence of the secret messages in unknown images without human intervention.

In this paper, we present a deep learning-based steganalysis model against spatial domain steganography and show the preliminary results. The steganalysis model is designed to have 1 high pass filter, 2 convolutional layers and 2 full connected layers and 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 98% and 90% accuracy are achieved for LSB stego-images with the same key and the different keys.

II. Related Works

Steganography methods can be categorized depending on the embedding domain: spatial domain methods and frequency domain methods. In the spatial domain methods, secret messages are inserted at the pixel level, which affects statistical characteristics of pixel values. In the transform domain methods, secret messages are inserted by modifying coefficient values after transformation such as DCT, DWT, and DFT.

Against steganography methods, steganalysis researches have been studied to prevent the transmission of secret messages. However, it is difficult to deal with any steganography methods by one steganalysis method called as universal steganalysis. Therefore, target specific steganalysis methods have studied and performed well. Also, there are limitations to determine the flaws of steganography methods with human intervention.

In these days, 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 apply.

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 uses a local structural relationship between pixels. A prediction

978-1-5386-2085-4/17 $31.00 © 2017 IEEE
DOI 10.1109/EECS.2017.9

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 researches are at the initial stage and continuous researches are required to get satisfactory results.

III. Proposed Steganalysis Using Deep Learning

A universal steganalysis against all steganography methods is not feasible. This section presents a target specific deep learning-based steganalysis against LSB-based 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 another name for an existing neural network, 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, where 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.

B. Convolutional neural network-based steganalysis

To detect the existence of the secret message, our CNN-based steganalysis model is designed to have 1 high pass filter, 2 convolutional layers and 2 fully connected layers as shown in Fig. 1.

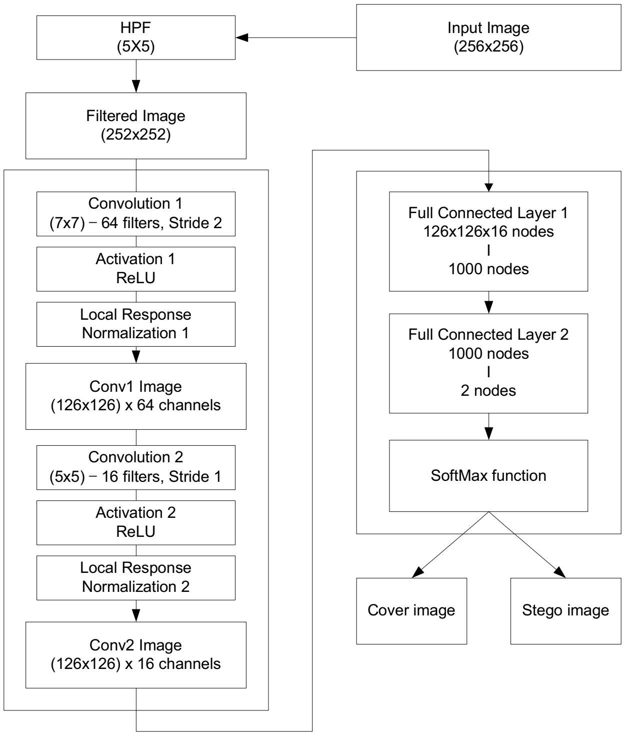

Fig. 1. CNN-based steganalysis model

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.

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

To extract features for steganalysis, convolutional layer is composed of 2 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. In the convolutional layer, pooling is not applied 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 are 126x126x16 and the number of output nodes are set at 1,000. In the 2nd layer, the number of input nodes are 1,000 and the number of output nodes are set at 2 to classify the cover image or the stego-image. Finally, a SoftMax function is applied to normalize the possibility of the cover image or the stego-image. For avoiding over-fitting, drop-out processing is not applied.

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 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.

Using a LSB-based steganography method, 20,000 stego-images are generated with the same key and the different keys, where 80 percent of images are used for the training of the model and 20 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.

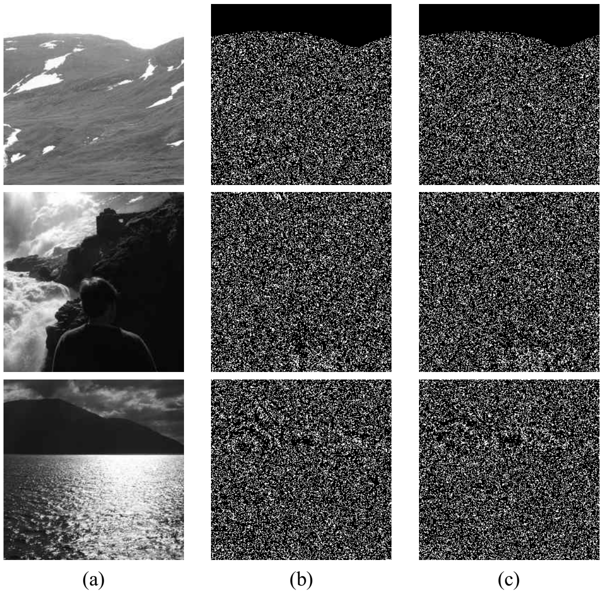
(a) (b) (c)

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

Fig. 3 shows the accuracy of the CNN-based steganalysis model for the stego-images with the same key. Fig. 4 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 and Y axis is the detection accuracy.

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 almost 98% and the convergence speed is very high. For the different key, the detection accuracy is 90% 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.

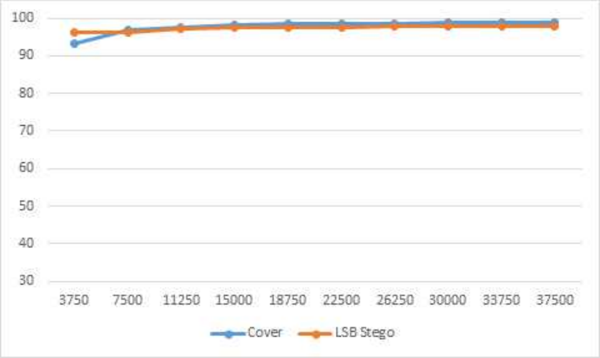

Fig. 3. Accuracy for the stego images with the same key

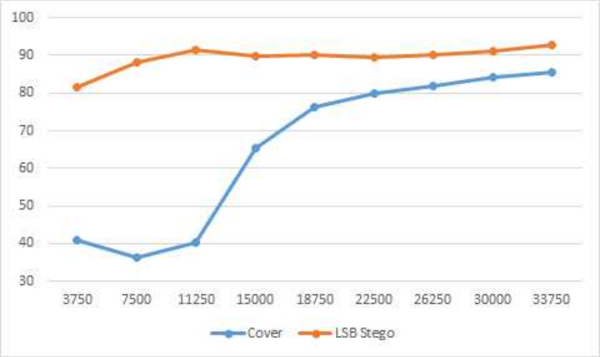

Fig. 4. Accuracy for the stego images with the different key

V. Conclusion

Steganalysis was studied to detect the existence of hidden messages in the 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 LSB-based steganography, this paper presented a CNN-based steganalysis model having 1 high pass filter, 2 convolutional layers and 2 full connected layers and showed preliminary results. No human intervention was required through deep learning. Experiments were performed using 10,000 cover and 20,000 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.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-2014-0-00639) supervised by the IITP(Institute for Information & communications Technology Promotion)

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.