DIRECT ADVERSARIAL ATTACK ON STEGO SANDWICHED BETWEEN BLACK BOXES

Hongyue Zha, Weiming Zhang, Chuan Qin, Nenghai Yu

University of Science and Technology of China, Hefei, 230027, China,
CAS Key Laboratory of Electro-magnetic Space Information
Email:zhahongyue@163.com

ABSTRACT

Due to the amazing progresses in deep learning techniques, steganography has now been challenged to tackle not only artificial feature-based but also effective deep-learning-based steganalysis. Recent steganographers have tried to conduct adversarial attacks to defend the steganalysis networks by fine-tuning the embedding details with the help of adversarial information, which, however, mostly are white-box attacks. This research studies a novel method to conduct steganographic adversarial attacks in practical scenario where stegos are sandwiched between black boxes. In our case, the toolboxes to generate stegos are steganographic black boxes where embedding adjustments are prohibited, and networks to detect stegos are semi-black boxes where most of the steganalysis networks' details are unavailable. By reforming few-pixel-attack into the form of extraction conservation noises and add them directly onto stegos, we ensure the message extraction and launch the attack in practical scenario. Experiments show that the proposed method can significantly boost the error rate of the deep-learning-based steganalysis and at the same time keep a comparable error rate when facing artificial feature-based steganalysis.

***Index Terms*—** adversarial attack, steganography, deep learning, steganalysis

1. INTRODUCTION

Steganography is the technique to embed secret messages into cover objects via introducing slight modifications that are indistinguishable from normal noises thus achieving covert communication. The approved embedding framework is the pipeline of distortion calculation plus STC(Syndrome-Trellis Codes) [1]. In spatial image domain, ever since the BOSS competition [2] in 2010, varieties of algorithms have been proposed to interpret better acquisitions of distortion, such as S-UNIWARD [3], HILL [4] etc. While in practice, steganographic algorithms are usually implemented as softwares or hardwares for efficiency and property rights.

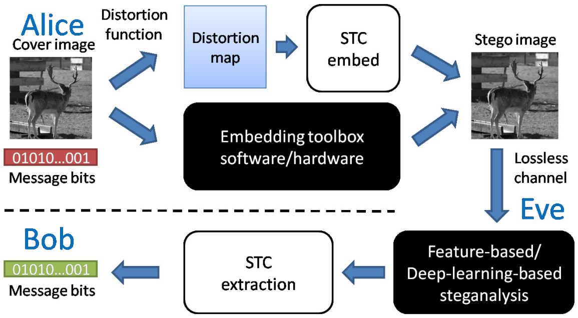

Fig. 1. A Practical Steganographic diagram

Conversely, steganalysis mostly aims at detecting the existence of steganography which can be regarded as a binary classification problem. Traditionally, types of artificial feature extraction methods (such as SPAM [5], versions of SRM (Spatial Rich Models) [6, 7]) and machine learning classifiers (such as Ensemble Classifier [8]) are introduced to fulfil the task. Meanwhile, with the development of deep learning techniques, some typical types of CNNs have been introduced to assist in steganalysis [9–13]. XuNet [11] is the first network that achieves comparative performance as artificial feature-based classifiers by constructing a 7-layered CNN. YeNet [12] even outperforms via introducing the Rich Model HPFs, Thresholded Linear Unit, and Selection-Channel-Aware messages. SRNet [13] removes the pooling layer and former hand-designed parts and provides good detection accuracy. State-of-the-art deep-learning based steganalysis methods indicate the effectiveness of neural network classifiers and the urgent demand for safer steganographic techniques.

However, the scheme of deep neural network is far from perfect reliability. Adversarial example attack has nowadays been a red-hot topic in the field of AI, as it shows the weakness of deep neural network that hasn't been conquered yet. Szegedy *et al.* [14] firstly found that adding well-designed small noises to the image context will dramatically mislead the image classification network with high confidence. From then on, varieties of adversarial examples have been proposed to launch attacks under varied situations [15].

Therefore, for steganographers, it is a natural thought to

This work was supported in part by the Natural Science Foundation of China under Grant U1636201 and 61572452, and by Anhui Initiative in Quantum Information Technologies under Grant AHY150400

conduct adversarial attacks to inactivate the steganalysis networks. Unfortunately, these attacks cannot be directly utilized under current STC framework because it might mess up the exact states of the stego bits which are indispensable for message extraction. To avoid this misfortune, researchers have tried to adjust the steganographic details to compromise. Li *et al.* [16] split the cover image into two parts thus seperating the embedding perturbations and adversarial noises. Zhang *et al.* [17] proposed a method to generate enhanced covers by iteratively adding adversarial noises to cover context so as to enhance the attacking noises' robustness to embedding modifications. Ma *et al.* [18] modified the pixel bits by ±1 according to the direction of adversarial noises under the framework of single-layered STC and introduced an unbalanced distortion function for ternary embedding according to the adversarial gradients. Tang *et al.* [19] proposed ADV-EMB which generates adversarial stego images with minimum amount of adjustable elements and achieved good performance.

All the prior arts need to fine-tune the details of the embedding strategies to absorb in the adversarial attacking abilities. However in practice, as is shown in **Fig.1**, some of the embedding techniques are packaged into black-toolboxes such as software programs or hardware chips for efficiency and property rights, therefore, no adjustment is allowed during the embedding procedure, and all the prior arts will fail. Moreover, most of the prior arts require the detailed structure and parameters of the attacked network and they are considered to be white-box attacks. However, most steganalysis networks are built by unexpected monitors like Eve in **Fig.1** and should be regarded as semi-black or even total-black boxes. In this paper, we propose a novel method to deal with the practical cases. By generating adversarial attacks directly on stegos and reform few-pixel-attack, we achieve the adversarial attack sandwiched between black boxes. We constrain the amplitudes of adversarial noises to ensure the message extraction, and limit the number of modified elements to resist the artificial feature-based steganalysis at the same time.

2. DIRECT ADVERSARIAL ATTACK ON STEGO

As we have mentioned above, a practical scenario is that steganography and steganalysis are both black boxes. Firstly, to deal with the steganographic black box, a straight-forward way is to conduct direct adversarial attacks on stegos.

2.1. Extraction Conservation Noises

To begin with, we have an assumption that all the steganographic algorithms packaged into black toolboxes are ternary embedding achieved by double-layered STC which modifies the cover elements by ±1 as is described in [1]. This assumption is reasonable because double-layered STC is still the dominating framework in steganography. In this case, the main problem for direct adversarial attack on stego is how to ensure the message extraction because there is a special rule for STC that the generated stego should be kept unchanged as we need to know the exact states of the stego bits when extracting the embedded message. Therefore, directly adding adversarial noises on stego images might mess up the stego bits and result in failure of message extraction. And that's the main reason why former researchers refuse to conduct a direct attack. But we find a special kind of noises that can be directly added onto stegos without failure of message extraction, and we call it Extraction Conservation Noises (ECN). And here we'll find the solution to ECN.

Actually, there's a common but important fact that the message extraction do not require the entire stego image but the Least Significant Bits (LSB) string and the 2nd LSB string of the stego image! As is explained in [1], message bits $\mathbf{m}$ are split into two parts *i.e.* $\mathbf{m}_{LSB}$ and $\mathbf{m}_{2ndLSB}$. The extractor Bob needs to know the exact length of the split messages to generate the party-check matrixes $\mathbf{H}_{LSB}$ and $\mathbf{H}_{2ndLSB}$, and extract the messages separately from the LSB string $\mathbf{y}_{LSB}$ and the 2nd LSB string $\mathbf{y}_{2ndLSB}$ of stego:

$$\mathbf{m}_{LSB} = \mathbf{H}_{LSB}\mathbf{y}_{LSB}, \mathbf{m}_{2ndLSB} = \mathbf{H}_{2ndLSB}\mathbf{y}_{2ndLSB},$$

Then we have $\mathbf{m} = cat(\mathbf{m}_{LSB}, \mathbf{m}_{2ndLSB})$ where $cat()$ is the splicing function. Therefore, we have the conclusion that noises can be directly added onto a stego without message extraction failure only if the LSB string and the 2nd LSB string of the stego are kept unchanged. To achieve this, we can easily get the solution to ECN as noises whose amplitudes are integer multiples of 4. After we get the solution to ECN, we need then to adjust the amplitudes of adversarial noises to the form of ECN so as to achieve direct adversarial attack on stego! No adjustment needs to be introduced into the embedding procedure thus tackling the steganographic black box. This conclusion is applicable for both spatial and JPEG images, but such significant modifications might result in complicated situations in JPEG domain, for simplicity and effectiveness, in this paper we only discuss about spatial cases.

2.2. ECN Few-pixel attack

Then we need candidate attacking methods and reform them to ECN and launch the direct adversarial attack. As we want to deal with the steganalysis black box, we prefer to select those black-box-attacks as candidates.

However, in our case, we have a special demand for our candidates. Although the target of our attack is deep-learning-based classifiers, the generated adversarial stegos still have to face the detection of artificial feature-based classifiers which are inevitable and still working as the mainstream methods in steganalysis. Note that according to our previous knowledge, the risk of steganographic exposure arises significantly with the relief of the limitations for perturbation levels when facing the artificial feature-based steganalysis. And this is an important reason why steganographic modifications have been constraint in range of ±1 so far. Therefore, the candidate attacking method should pay special attention to defend the artificial feature-based detections as well. To deal with this, we notice that the most effective artificial features like SPAM or SRM are mostly based on statistical frequencies. And we suggest a possible way to avoid their detection by limiting the number of modified pixels which might relieve the influence on statistical frequency changes.

Therefore, we finally select few-pixel-attack [20] as the winner, because it is a semi-black-box attack and is able to adjust both amplitude and scale of the adding noises. And then we'll illustrate the way to reform few-pixel-attack to ECN and launch the attack.

Few-pixel-attack is a semi-black box attack which means it only requires the prediction probabilities of each labels, in our case the $\mathbf{p}(\mathbf{I}) = [p_c(\mathbf{I}), p_s(\mathbf{I})]$, where $p_c(\mathbf{I})$ and $p_s(\mathbf{I})$ are the probabilities to classify image **I** as 'cover' and 'stego' respectively, with $p_c(\mathbf{I}) + p_s(\mathbf{I}) = 1$. Here we set an evaluation function as $f(\mathbf{I}) = p_c(\mathbf{I}) - p_s(\mathbf{I})$. Denote the stego image as **S**, the adversarial noises as **A**. And the goal of our attack goes to:

$$\underset{\mathbf{A}}{\text{maxmize}} \;\; f(\mathbf{S}+\mathbf{A}), \text{subject to } \|\mathbf{A}\|_0 \leqslant k, k \in \mathbb{N}.$$

$\|\mathbf{A}\|_0$ is the number of non-zero elements in **A**, hence k denotes the number of pixels to be modified, and the attack is a so-called k-pixel-attack.

To solve this problem, we need to apply the DE (Differential Evolution) algorithm. Similar to [20], we encode the state of **A** into an array **r** which consists of k tuples with each tuple containing 3 elements: x, y coordinates and amplitude of the noise. We set the population number to 400 which means we'll initial 400 arrays as the first generation. For initialization, in each tuples of the array, x, y coordinates are randomly selected legal candidates within the image size boundary and the amplitudes are random multiples of 4 constrained to the grayscale limit. Then we'll produce the next generations (children) by DE formula:

$$\mathbf{r}_{i,t+1} = \mathbf{r}_{n_1,t} + \beta(\mathbf{r}_{n_2,t} + \mathbf{r}_{n_3,t}), n_1 \neq n_2 \neq n_3.$$

$\mathbf{r}_{i,t}$ represents array of the i-th population in the t-th generation, and n_1, n_2, n_3 are randomly selected population indexes, β is the scale parameter set to 0.5. We set both the location boundaries and grayscale boundaries as periodic boundaries. And moreover, due to the scale parameter, the amplitude of the generated children might not be an integer multiple of 4, if that happen, we will tune it to its nearest smaller legal level. Once the children are generated, we'll calculate the evaluation function of both children and parents, and compete between each pair of the corresponding parent and child. Only the winner survives for the next iteration. The maximum number of iteration is 100. And we'll select the population with highest evaluation as the final **A**. When computed on a GTX1080, a total 100 iterations for a single image will take 15 minutes on average.

3. EXPERIMENTS

In this part, we'll verify our proposed ECN few-pixel-attack by attacking the approved deep-learning-based steganalysis XuNet [11] and YeNet [12].

3.1. Setup and Evaluation

Experiments are carried on the imagesets of BOSS [2] and BOWS [21], both containing 10000 spatial images. Resize the images to the size of 256×256 using the MATLAB **imresize()** function, and we get the original cover imageset with 20000 images. Then we embed the cover set with HILL and S-UNIWARD at payload ratio of 0.2bpp and 0.4bpp using the embedding-simulator, and we get four sets of stegos.

Each turn of the experiments, we select the cover set and one stego sets *i.e.* 20000 pairs of cover-stegos. We randomly select 14000 pairs of them for training, 1000 pairs for validation and the rest 5000 pairs for testing. The first layer of XuNet is slightly adjusted as we change the stride of the first convolution layer from 2 to 1 to deal with images of size 256×256 rather than 512×512 in [11]. YeNet is trained as the SCA version. All the unmentioned hyper parameters are set the same as the original paper.

Then we conduct the proposed ECN few-pixel-attack on the stegos that are correctly classified as 'stego' by the network. And then, we'll replace the original stegos with the corresponding adversarial stegos to update the testing sets. And we'll test the performance of target networks on the updated testing sets under the common evaluation of error rate:

$$P_E = \frac{P_{MD} + P_{FA}}{2},$$

where P_{MD} denotes the miss detection rate and P_{FA} denotes the false alarm rate.

3.2. Experimental results

We adjust the parameter k which represents the number of pixels to be modified *i.e.* k-pixel-attack. When $k = 0$, the result is therefore the original performance without adversarial attack. Moreover, the experiment under each conditions is repeated three times and the average P_E is calculated.

Firstly, we evaluate the performance when attacking XuNet. As is shown in **Fig. 2**, the proposed few-pixel-attack can fool XuNet successfully. The P_E goes higher with the increasement of k which indicates a growing strength of adversarial attack. To reach the same P_E, more pixels are required to be modified when the payload ratio is 0.4bpp than 0.2bpp. This is reasonable that more modifications are introduced during the STC embedding when the payload ratio is 0.4bpp, and therefore a stronger adversarial strength is required to cover up. The P_E is similar between the attacks on stegos embedded with HILL and S-UNIWARD. Note that the P_E after attack is higher than 0.5 which seems weird for

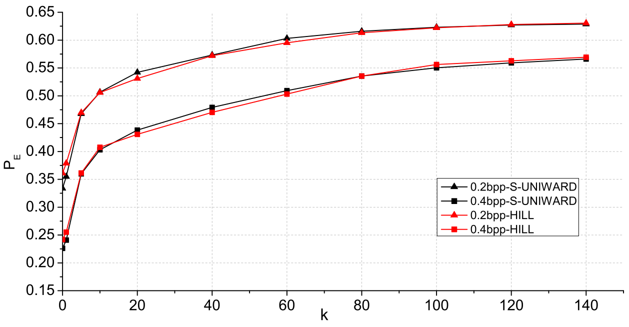

Fig. 2. P_E - k when attacking XuNet trained on imagesets embedded by S-UNIWARD and HILL at 0.2bpp and 0.4bpp

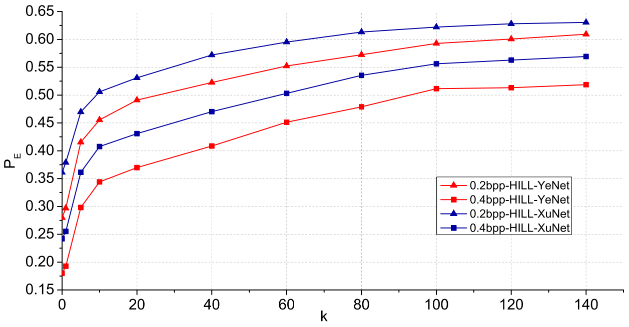

Fig. 3. P_E - k when attacking XuNet and YeNet trained on imagesets embedded by HILL at 0.2bpp and 0.4bpp.

a binary classifier. And this is caused by the setting that the proposed attack is targeted on the classifier trained on original set without adversarial modifications, and no retraining is considered.

Then we attack YeNet trained on only HILL and compare the performance with that of XuNet. As is shown in **Fig. 3**, the proposed attack is still efficient. When k and payload changes, we have similar trends on P_E with that of XuNet. But the final P_E grows weaker than XuNet which we suspect was caused by the selection-channel-aware mechanism in YeNet that makes its prediction more sensitive to embedding noises so that the influence of adversary is slighter.

To verify the extraction conservation effect, we compare the generated adversarial stegos with their corresponding original stegos in the above experiments. We find that the LSB and 2ndLSB strings of each pair of them are 100% the same. As STC extract messages purely on LSB and 2ndLSB strings, we conclude that the message extraction is ensured.

As we have mentioned before, we need to evaluate the adversarial stegos under the artificial feature-based steganalysis as well. Therefore, we extract the SRM features of the imagesets generated from HILL 0.2bpp and 0.4bpp and train ensemble-classifiers to evaluate the performance on the testing sets updated with adversarial stegos target at XuNet.

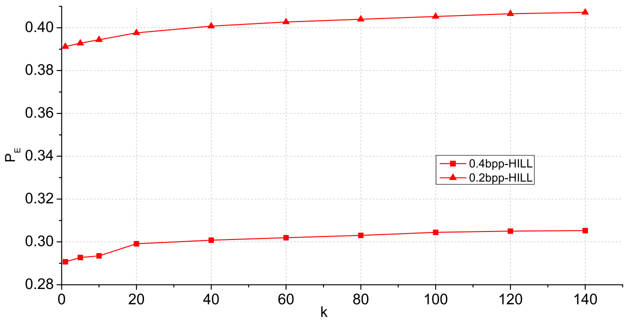

Fig. 4. P_E - k when attacking XuNet but tested by SRM+Ensemble Classifiers both trained on imagesets embedded by HILL at 0.2bpp and 0.4bpp

As we can see in **Fig. 4**, the proposed method can not only attack the target network but also resist the detection of artificial feature-based classifier. The P_E of artificial feature-based classifier even arise slightly compared with original testing set *i.e.* the case of $k = 0$. And combine with prior results, we suggest to set k to 100 because when k is bigger than 100 the P_E of both kinds of classifiers tend to level off.

The proposed method is the first trial to conduct adversarial attacks directly on stegos. And therefore it can deal with practical situations when steganography and steganalysis are both black boxes while previous arts cannot. Moreover in [19], the error rates when attacking XuNet trained with S-UNIWARD sets at 0.2bpp and 0.4bpp are 0.623 and 0.598, which in our case are 0.621 and 0.550. This means our proposed attack can achieve comparable attacking performance as previous white-box arts.

4. CONCLUSIONS

We successfully conduct steganographic adversarial attacks in practical scenario where steganography and steganalysis are both black boxes. This is a real start for direct adversarial attack against deep-learning based steganalysis. However there are still some defects in our method and we'd like to improve them in future works. This work discusses only about spatial images, and we will extend it to JPEG domain in future. Although the few-pixel-attack is a semi-black-box attack which is much closer to practical use than white-box methods, we will search for real black-box attacks to improve. Moreover, defending methods like retraining are not considered due to space limit and we'll further study them for complementation.

5. REFERENCES

[1] Tomáš Filler, Jan Judas, and Jessica Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 920–935, 2011.

[2] Patrick Bas, Tomáš Filler, and Tomáš Pevnỳ, "" break our steganographic system": The ins and outs of organizing boss," in *International Workshop on Information Hiding*. Springer, 2011, pp. 59–70.

[3] Vojtěch Holub and Jessica Fridrich, "Digital image steganography using universal distortion," in *Proceedings of the first ACM workshop on Information hiding and multimedia security*. ACM, 2013, pp. 59–68.

[4] Bin Li, Ming Wang, Jiwu Huang, and Xiaolong Li, "A new cost function for spatial image steganography," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4206–4210.

[5] Tomáš Pevny, Patrick Bas, and Jessica Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Transactions on information Forensics and Security*, vol. 5, no. 2, pp. 215–224, 2010.

[6] Jessica Fridrich and Jan Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[7] Tomas Denemark, Vahid Sedighi, Vojtech Holub, Rémi Cogranne, and Jessica Fridrich, "Selection-channel-aware rich model for steganalysis of digital images," in *Information Forensics and Security (WIFS), 2014 IEEE International Workshop on*. IEEE, 2014, pp. 48–53.

[8] Jan Kodovskỳ, Jessica J Fridrich, and Vojtech Holub, "Ensemble classifiers for steganalysis of digital media.," *IEEE Trans. Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, 2012.

[9] Yinlong Qian, Jing Dong, Wei Wang, and Tieniu Tan, "Deep learning for steganalysis via convolutional neural networks," in *Media Watermarking, Security, and Forensics 2015*. International Society for Optics and Photonics, 2015, vol. 9409, p. 94090J.

[10] Songtao Wu, Sheng-hua Zhong, and Yan Liu, "Steganalysis via deep residual network," in *Parallel and Distributed Systems (ICPADS), 2016 IEEE 22nd International Conference on*. IEEE, 2016, pp. 1233–1236.

[11] Guanshuo Xu, Han-Zhou Wu, and Yun-Qing Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 708–712, 2016.

[12] 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.

[13] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, May 2019.

[14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[15] Naveed Akhtar and Ajmal Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *arXiv preprint arXiv:1801.00553*, 2018.

[16] Shiyu Li, Dengpan Ye, Shunzhi Jiang, Changrui Liu, Xiaoguang Niu, and Xiangyang Luo, "Attack on deep steganalysis neural networks," in *International Conference on Cloud Computing and Security*. Springer, 2018, pp. 265–276.

[17] Yiwei Zhang, Weiming Zhang, Kejiang Chen, Jiayang Liu, Yujia Liu, and Nenghai Yu, "Adversarial examples against deep neural network based steganalysis," in *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*. ACM, 2018, pp. 67–72.

[18] Sai Ma, Qingxiao Guan, Xianfeng Zhao, and Yaqi Liu, "Adaptive spatial steganography based on probability-controlled adversarial examples," *arXiv preprint arXiv:1804.02691*, 2018.

[19] W. Tang, B. Li, S. Tan, M. Barni, and J. Huang, "Cnn-based adversarial embedding for image steganography," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2019, DOI: 10.1109/TIFS.2019.2891237.

[20] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, 2019.

[21] Alessandro Piva and Mauro Barni, "The first bows contest: break our watermarking system," in *Security, Steganography, and Watermarking of Multimedia Contents IX*. International Society for Optics and Photonics, 2007, vol. 6505, p. 650516.