Proceedings of CECNet 2022
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220536

A Generative Learning Architecture Based on CycleGAN for Steganalysis with Unpaired Training Images

Han ZHANG [a*], Zhihua SONG [b, 1*], Feng CHEN[b], Xiangyang LIN[b], Qinghua XING[b], Qingbo ZHANG[b] and Yongmei ZHAO[a]
[a] *Equipment Management and UAV Engineering College of Air Force Engineering University, Xi'an, China*
[b] *Air and Missile Defense College of Air Force Engineering University, Xi'an, China*

Abstract. Steganalysis based on deep learning has made noticeable progress over the past few years where the training is all based on paired images. However, scenes without paired training data exist. We present an architecture for learning to generate corresponding pseudo stego image from a cover-image in the absence of paired training images. We seek a mapping G that can generate pseudo stego images indistinguishable from the real but unpaired stego images using an adversarial loss. Because this mapping is highly under-constrained, we designed a CycleGAN and introduce spectrum of stego images to reinforce the adversarial loss. Qualitative comparisons demonstrate the superiority of our approach.

Keywords. Steganalysis, generative learning, residual, CycleGAN

1. Introduction

Steganalysis and steganography are two sides of a coin and cannot be studied separately. In this communication game, the steganography player attempts to achieve communication by hiding secret message or image in a carrier image, which we named it cover as shown in Figure 1, through the public communication channel. The steganalysis player tries to anticipate the risk of misusing of the public communication channel by steganography, i.e., to calculate the probability that the images on the public communication channel are embedded with secret information.

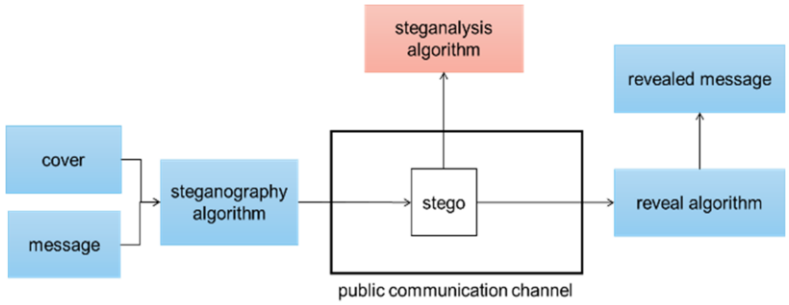


Figure 1: The architecture of the steganography and steganalysis game. The cover in the left stands for the carrier image to hide the secrete message. The stego in the middle stands for the image generated by a steganographic algorithm using the cover and the secret message.

[1] Corresponding Author: Zhihua Song, E-mail: szhele@163.com, * indicates equal contribution

For the steganalysis player, paired cover and stego images are ideal to train the steganalysis neural network. But what if there are no paired cover images? Compared with stego image, the paired cover image is more difficult to get on the public communication channel as the cover image usually does not need to be transmitted to the receiver.

In this study, we are seeking to generate paired cover and pseudo stego images all in the absence of any paired training images: capturing special spatial and spectrum characteristics of the stego image collection and translating them to a cover image collection to generate corresponding pseudo stego image collection.

This problem can be described as an unpaired image-to-image translation problem[1]. Unlike the other image-to-image translation applications, such as style transfer[1], object transfiguration[2], and image synthesis[3], the cover-to-stego image translation must work between two extremely similar image collections, while the minimization of distortion loss is a relentless pursuit goal for the steganography player.

We therefore focused on cycle-consistent generative adversarial network (CycleGAN) [1] which is a state-of-the-art unpaired image-to-image translation architecture. In theory, the CycleGAN prevents mode collapse, where all input images map to the same output image, and can add stego styles to a cover image to make a pseudo stego image if the styles are obvious enough to be captured. But in reality, it is usually difficult to distinguish the style of a brilliant stego image from that of an ordinary image. Our early experiments also proved this conjecture, the generated images were not good enough to train any of the baseline steganalysis neural networks and they tend to be the same as the cover. This indicates that the CycleGAN cannot grasp the style difference between the two collections. Such a result is consistent with the nature of the problem, the two image collections are highly similar in the texture and visual effect, while these are the features that the convolution layers at the CycleGAN are good at.

The next thing we need to do is add more detailed and distinguishable information to our network. Therefore, we exploit the spectrum disequilibrium property, that the steganography distribution over different frequency is generally uneven. The rest of the paper is arranged as follows: after a brief review of related works in Section 2, we describe the architecture of the proposed network in Section 3. In Section 4, the experimental results are presented. Finally, Section 5 discusses the conclusion and future work.

2. Related Works

CycleGAN[1] was presented by Jun-Yan Zhu in 2017 for the problem of unpaired image-to-image translation and achieved impressive results in object transfiguration, season transfer, collection style transfer, and photo enhancement. The key to CycleGAN's success is the idea of cycle-consistent loss that encourages the bijection mapping between the generated image and the real source image and forces the generated image to be indistinguishable from images in the target domain. We adopt a cycle-consistent loss to our network. To the best of our knowledge, there is no study has reported using CycleGAN-generated images for the extensive training of steganalysis models.

In [4], a deep residual steganalysis architecture called SRNet is proposed to minimize the use of heuristics and externally enforced elements and it provides state-of-the-art performance for both spatial-domain and JPEG steganography. We borrows this architecture directly in our network as the steganalysis block.

Baluja[5] present an image-into-image steganography network, which can embed a full-sized image inside another image with minimal quality loss. There are three

components in the system, i.e., the preparation network, the hiding network, and the reveal network. These three components are trained simultaneously as a single network and the reveal network uses the stego images only. We call the network Baluja-Net for convenience and choose it as one of our baseline for our experiments.

3. Model

The key is to learn mapping functions between the cover image collection C and the stego image collection S, given unpaired training samples $\{c_i\}_{i=1}^{N} \in C$ and $\{s_j\}_{j=1}^{M} \in S$, using the CycleGAN. However, such an approach does not guarantee that the output y can embody the subtle feature difference between the cover and its corresponding stego. The CycleGAN is good at transfer texture difference and the texture distortion is usually what the steganography player is trying to minimize.

Moreover, in practice, we have found it difficult to optimize the adversarial objective: standard CycleGAN leads to the problem of mapping collapse, where input image is mapped to the same image as itself. Therefore, we exploit the spectrum information of the training samples.

The proposed model is shown in Figure2. The model includes two generators $G_s: C \rightarrow S$ and $G_c: S \rightarrow C$, and two discriminators D_s and D_c.

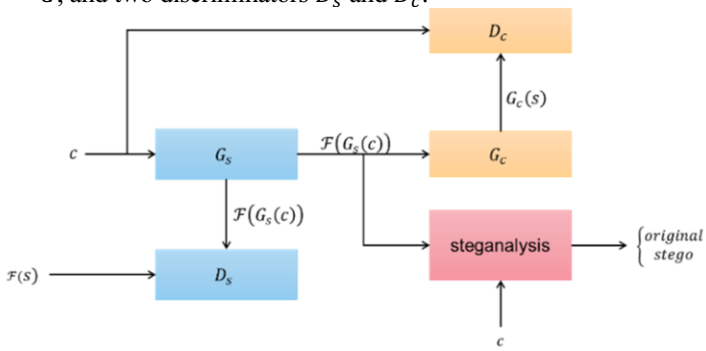


Figure2: Architecture of the proposed model

For the mapping function $G_s: C \rightarrow S$ and its discriminator D_s, we express the loss function:

$$L_{D_s} = \left[\mathcal{L}_{BCE}\left(D_s\left(\mathcal{F}\left(G_s(c)\right)\right), 0\right) + \mathcal{L}_{BCE}\left(D_s\left(\mathcal{F}(s)\right), 1\right)\right]/2$$

$$L_{G_s} = \mathcal{L}_{BCE}\left(D_s\left(F\left(G_s(c)\right)\right), 1\right)$$

where $\mathcal{L}_{BCE}(\cdot)$ is a binary cross entropy loss function, $\mathcal{F}(x)$ is the spectrum of image x. G_s tries to generate pseudo stego images that resemble unpaired real stego images in the frequency domain, while D_s aims to distinguish between the pseudo stego images and real stego images in the frequency domain.

For the mapping function $G_c: S \rightarrow C$ and its discriminator D_c, we express the loss function in the spatial domain as there is no need for the spectrum similar for the cover images:

$$L_{D_c} = \left[\mathcal{L}_{BCE}\left(D_c\left(G_c(s)\right), 0\right) + \mathcal{L}_{BCE}\left(D_c(c), 1\right)\right]/2$$

$$L_{G_c} = \mathcal{L}_{BCE}\left(D_c\left(G_c(s)\right), 1\right)$$

where G_c tries to generate pseudo cover images that resemble corresponding real cover images in the spatial domain, while D_c aims to distinguish between the pseudo cover images and real cover images in the spatial domain.

Similarly, we introduced identity losses in the frequency domain of stego images and in the spatial domain of cover images:

$$L_{I_s} = \mathcal{L}_{BCE}\left(D_s\left(\mathcal{F}\left(G_s(s)\right)\right), 0\right)$$

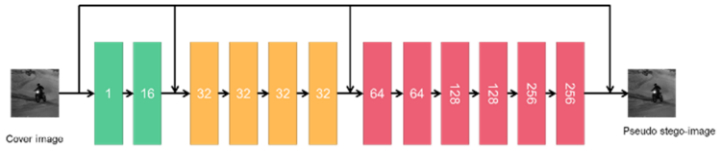
$$L_{I_c} = \mathcal{L}_{BCE}\left(D_c\left(G_c(c)\right), 0\right)$$

The total GAN loss is

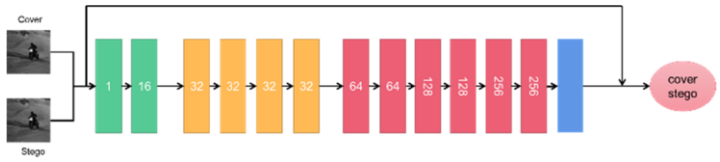
$$L_{GAN} = \alpha_1 L_{D_s} + \beta_1 L_{D_c} + \alpha_2 L_{G_s} + \beta_2 L_{G_c} + \alpha_3 L_{I_s} + \beta_3 L_{I_c}$$

where we can adjust their weightsα_i and β_i, $i \in \{1,2,3\}$, to strengthen or weaken the desired domain.

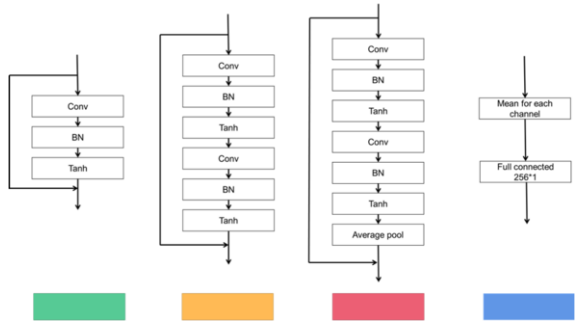
The structures of generator and discriminator are shown in Figure3.



(a)generator



(b)discriminator



(c)basic layer

Figure 3. Structure of the generator (a) and the discriminator (b). There are four types of layers. The first type of layers is residual convolution layer with 3 × 3-kernal as the first layer in (c). The second type of layer is convolution layer with 1 × 1 kernel as the second layer shown in (c). The third type of layer is convolution layer with 1 × 1 kernel as the third layer shown in (c). The number in each layer is its number of input channels. Unlike the SRNet, we use Tanh as the activation function. The fourth type of layer is a fully connected layer with a mean operation for each input channel as.

4. Experiments and Analysis

Our experiments are conducted on a commonly used publicly available sources BOSSBase v.1.01. The Baluja-Net [5] is used to embed a full-sized image into a cover image. We generate 2000 real stego images from the BOSSbase in such a way: for an image i in BOSSbase, we randomly select any other image $j \in \{k|1 \leq k \leq 10000, k \neq i, k \in Z\}$ as secret image for steganography.

We conducted several experiments for performance comparison and show the results in Table 1. The Non in the second column stands for the steganalysis is trained with the real unpaired cover and stego images only.

Table1. Performance comparisons of proposed architecture

Steganography	generator	Training size		
		200	400	800
Baluja-Net[5]	RES	86.33%	91.61%	96.31%
	Non	93.33%	89.46%	93.11%
S_UNIWARD (0.2bpp)	RES	61.05%	70.19%	73.28%
	Non	50.00%	50.00%	50.00%
S_UNIWARD (0.4bpp)	RES	59.83%	72.60%	77.89%
	Non	50.00%	50.13%	50.58%
S_UNIWARD (0.6bpp)	RES	65.39%	74.88%	79.15%
	Non	50.540%	51.03%	51.85%

For the unpaired dataset composed of stego images generated by Baluja-Net, the generative learning framework proposed in this paper has certain advantages over the Non mode when the training size is 400 and 800. However, for the unpaired dataset composed of stego images generated by S-UNIWARD algorithm, the generative learning framework proposed in this paper has obvious advantages. If there is no paired dataset, the stego image generated by S-UNIWARD algorithm can be recognized by steganalysis module at about 50%, which is basically equivalent to random guess.

5. Conclusions and Future Work

In this paper, we proposed a generative learning network for steganalysis. The experiment results showed that the generative learning architecture improves the detecting accuracy of the steganalysis when the training images are unpaired. The generative learning framework proposed in this paper is a feasible and effective strategy for steganalysis training in the case of unpaired training dataset.

Although the generative learning framework proposed in this paper has achieved good results in the case of unpaired datasets, there is still much work to be done This architecture performance bad for other classic steganography algorithms such as WOW [6] and HUGO [7]. Future work includes exploring more unpaired training sets generated by different steganography algorithms, and trying to improve the performance of the generative learning framework from the aspects of network structure and learning algorithms.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 71571190, 62002381.

References

[1] J. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244.

[2] Chen, X., Xu, C., Yang, X., Tao, D. (2018). Attention-GAN for Object Transfiguration in Wild Images. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11206. Springer, Cham. https://doi.org/10.1007/978-3-030-01216-8_11

[3] T. Park, M. -Y. Liu, T. -C. Wang and J. -Y. Zhu, "Semantic Image Synthesis with Spatially-Adaptive Normalization," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2332-2341, doi: 10.1109/CVPR.2019.00244.

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

[5] Shumeet Baluja. "Hiding images in plain sight: deep steganography". Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 2017. 2066–2076.

[6] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in 2012 IEEE International Workshop on Information Forensics and Security (WIFS), 2012, pp. 234–239. doi: 10.1109/WIFS.2012.6412655.

[7] T. Filler and J. Fridrich, "Gibbs Construction in Steganography," IEEE Trans. Inf. Forensics Secur., vol. 5, no. 4, pp. 705–720, 2010, doi: 10.1109/TIFS.2010.2077629.