Rate-Distortion-Based Steganography Via Generative Adversarial Network

Subjects: Computer Science, Information Systems
Contributor: Yi-Lun Pan, Ja-Ling Wu

Information hiding can imperceptibly transfer secret information into chosen cover media. It can ensure the origins of data and behave as a second channel for data transmission. Steganography is the art of covering or hiding extra data inside a chosen cover message, e.g., an image. The term itself dates back to the 15th century; in a typical scenario, the sender hides a secret message inside a cover image and transmits it to the receiver, who recovers the message. Even if eavesdroppers monitor or intercept the communication in-between, no one besides the sender and receiver should detect the presence of the hidden message.

Keywords: image steganography ; rate-distortion ; mutual information ; generative adversarial network

1. Introduction

Compared to cryptography, steganography has the advantage that non-target intermediaries will not suspect the existence of secret information itself. The media embedded within extra messages is called the stego media, and the media used to host the embedded messages are called the cover media. Attackers use steganalysis techniques to prevent the successful transmission of secret information. To conduct steganography is challenging because embedding extra messages can alter the cover's appearance and underlying statistical distribution.

The first common challenge in designing a steganography scheme is how to enlarge the amount of transmittable payload, named the scheme's capacity. Steganography capacity is usually measured in *bits-per-pixel* (bpp). The longer the embedded message, the larger the bpp and the more altered the cover. Suppose the visual appearance of the hidden-message embedded image (denoted as the stego-image) does not appear close to that of the cover images. In that case, non-photo-realistic issues may result in the associated synthesis-based applications, such as the anchor face generation application in the metaverse. Existing image steganography approaches are only practical for embedding a relatively low payload of around 0.4 bits per pixel [1]. With vigorous developments in generative adversarial networks (GANs), many works have applied GAN-based approaches to steganographic design methods [2][3][4], which saw a boom in image steganography's applicability. Reference [5] is the first article that attempts to address the application of GAN in image steganography with acceptable performance. Afterwards, with the help of GAN, Zhang et al. proposed the SteganoGAN [6], which achieves the embedding capacity with a payload of 4.4 bits per pixel. In 2020, Fu et al. improved the work of SteganoGAN; they proposed the HIGAN [7], which can handle a 24-bit-sized payload. Investigating the possibility of further increasing information capacity is one of the to-be-conquered challenges.

Furthermore, to enlarge the information embedding capacity to higher than 192+ bpp, inspired by the authors of [8], the rate-distortion loss functions was leveraged to ensure the visibility of the cover image and enhance the compressibility of the embedding image. In other words, the primary goal is to optimize the visual quality of the stego-image and hide as much secret-related information as possible at the same time.

The second challenge of steganography is its poor robustness against the chosen cover attacks [5]. When an attacker knows both the stego and the cover images, conducting a simple pixel differencing operation may leak secretly-related information. Although the recent work proposed by Lu et al. [9] can hide multiple secret images, low system security against simple pixel-differencing operations is still the main weak point of the approach, i.e., the confidential information will be exposed. In contrast, the proposed multiple-secret-image embedding scheme, besides enlarging the capacity, will also significantly improve the system's security.

The third challenge concerns the stability of the trained model. Most of the related works developed a supervised cover synthesis steganography, as addressed in [10], to face the model's training stability issues.

As for the state-of-the-art in the field of NN-based steganography published in the past three years, it is recommend that the following five highly related works: Duan et al. [11], SteganoGAN [6], HIGAN [7], SteganoCNN [12], and ISN [9]. Among them, [6][7][12] are limited in their model capabilities and can only process a singular secret image or text information. Nevertheless, it is worth noting that the quality of the images processed in [11] is superior. Inspired by [11], it was also tried to make the quality of the generated stegos and the reconstructed images as good as possible. SteganoCNN increased the embedding payload capacity to 47.92 bpp, while ISN considered how to handle multiple secret images hidden. Increasing the embedding capacity and relatively high computational complexity are still weaknesses of these proposals, which reduces the computational complexity from the perspective of network architecture. In summary, compared with the works mentioned above, the approach enlarged the payload capacity, enhanced the computational stability, and increased the computational efficiency simultaneously.

Besides, most of the above studies did not provide theoretical information-based analyses of their work, which might bring further insights for comprehending the approaches' physical meaning. To respond to this concern, it was not only proposed the RD-Stego system but also provided an informational-theoretic explanation of the design of the adopted cost functions. Shannon's mutual information (MI) was taken into the construction of the RD-Stego system's cost functions, including (a) visual acceptability—in maximizing the MI lower bound of the difference between the cover and the stego-images, which is equivalent to maximizing the acceptable perception range between them. (b) Recovery fidelity—maximizing the MI lower bound between the embedded secret and the reconstructed secret, which is equivalent to maximizing the retrieval fidelity related to the secret messages.

2. Steganography Based on GANs

With the great help of GAN, several researchers found that GAN-based steganography can solve the problem of non-photo-realistic appearance in cover synthesis. Abadi et al. [13] first applied this idea to steganography's cover synthesis and added an adversarial network to their algorithm. Zhu et al. [14] proposed an encoder-decoder network architecture to deal with the embedding and extraction of secret information. The shortcomings of [13][14] are the adopted loss functions, which complicate the system design and make the training process unstable. Zhang et al. [6] significantly improved the loss function design and presented an end-to-end GAN-based steganographic model. They used adversarial training to solve the steganography task and regarded message embedding and extraction as encoding and decoding problems. Tancik et al. [15] achieved robust decoding even under "physical transmission" by adding a set of differential image corruptions between the encoder and decoder that successfully approximate the space of distortions. However, the steganographic images generated by the neural network are highly correlated with the original cover.

Hu et al. [16] tried to accomplish the cover synthesis of steganography in an unsupervised manner. The key idea is finding a map from the noise to message and hiding messages into noises. A special extractor is then trained to extract messages from the noise. However, the high implementation cost of the latter training handicaps its value in practical usage. In response to unsupervised cover synthesis steganography being hard to use in practice, subsequent works redirect themselves toward the semi-supervised counterparts instead. Inspired by ACGAN, Liu et al. [17] proposed establishing a mapping relationship between the class label and noise first and then generating stego-images. The proposed RD-Stego model leverages the advantages of semi-supervised cover synthesis steganography algorithms. Herein, the encoder network comprises a convolution layer and the residual block. As a result, the generated steganography image has much lower distortion and closer distribution to the original carrier image. Moreover, it can smooth the discontinuity in gradient calculation during training. Such a smoothing gradient calculation characteristic provides reasonable training stability and conforms to steganographic basic conditions (BSC) [18][19].

3. The Limitations of the Current Steganography Works

At present, the most apparent limitations of GAN-based steganography algorithms are their low embedding capacity and low robustness against the chosen cover attacks. As for the embedding capacity, Baluja [5] presented an encoder–decoder network and tried to increase the amount of information it carried [5], successfully embedding a color image into another color image of the same size, yet the resulting stego-image may expose confidential information. Rehman et al. [20] tried to hide a gray-level picture into a color picture of the same size, but severe color distortion was observed in the resultant stego-image. Zhang et al. [21] proposed the ISGAN process, which hides a grayscale image into the Y channel of a color cover image and improves the security of the model through adversarial training between the encoder–decoder and steganalysis networks.

Zhang et al. [21] inspired us to use another channel to aggregate the information that needs to be protected. Besides traditional RGB color channels, an extra channel was used for hiding QR code/text information in the work. In this way, the

SteganoGAN [6] could be used to hide the color, grey-scale, and binary data in a hosted picture and enlarge the information capacity contained in the stego-image. In doing so, SteganoGAN achieves 4.4 bits-per-pixel embedding capacity; this is still not good enough. Fu et al. [7] enlarged the payload of [6] in 2020. Whether it is possible to continue to increase the embedding capacity is the main target herein. The lesson learned from [7] tells us that using other channels to handle non-color information, such as QR-coded messages, seems to be a good choice. In other words, if the designed RD-Stego can rebuild QR-coded messages perfectly, the embedding capacity issue will be completely solved.

Deep Steganography [5], proposed by Baluja, faces the problem of chosen cover attacks, especially when attackers have both the stego and cover images. The attackers can magnify the difference between the stego and the cover images and easily extract secret-related information. This shortage comes from the Deep Steganography method inputting both the cover and the secret images into its pre-trained model and then connecting them back into GAN in series. Therefore, an attacker can choose a specific cover image as input and subtract it from the associated stego-image to find their difference. To deal with this issue, Tang et al. [22] proposed an adversarial embedding scheme based on CNN-ADV-EMB architecture to resist the above-mentioned chosen cover attack. Unfortunately, this type of method is of a security concern. Instead of directly concatenating the cover and the stego-images, the proposed RD-Stego uses element-wise additions to perform perturbation, significantly enhancing system security. In 2021, although the method proposed by Lu et al. [9] can hide multiple secret images, the main weakness of the method is also apparent in terms of security, which requires a simple pixel-differencing operation for the secret information to be exposed. On the contrary, the proposed method also significantly improves security, especially for this problem.

References

1. Pevný, T.; Filler, T.; Bas, P. Using High-Dimensional Image Models to Perform Highly Undetectable Steganography. In L ecture Notes in Computer Science; Springer Science + Business Media: Berlin, Germany, 2010; Volume 6387, p. 161.
2. Hayes, J.; Danezis, G. Generating Steganographic Images via Adversarial Training. arXiv 2017, arXiv:stat.ML/1703.00 371.
3. Ke, Y.; Zhang, M.; Liu, J.; Su, T.; Yang, X. Generative Steganography with Kerckhoffs' Principle. arXiv 2021, arXiv:cs.M M/1711.04916.
4. Shi, H.; Dong, J.; Wang, W.; Qian, Y.; Zhang, X. SSGAN: Secure Steganography Based on Generative Adversarial Net works. arXiv 2018, arXiv:cs.CV/1707.01613.
5. Baluja, S. Hiding Images in Plain Sight: Deep Steganography. In Proceedings of the Advances in Neural Information Pr ocessing Systems; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30.
6. Zhang, K.A.; Cuesta-Infante, A.; Xu, L.; Veeramachaneni, K. SteganoGAN: High Capacity Image Steganography with GANs. arXiv 2019, arXiv:cs.CV/1901.03892.
7. Fu, Z.; Wang, F.; Xu, C. The Secure Steganography for Hiding Images via GAN. EURASIP J. Image Video Processing 2020, 2020, 1–18.
8. Ballé, J.; Minnen, D.; Singh, S.; Hwang, S.J.; Johnston, N. Variational Image Compression with a Scale Hyperprior. In P roceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 201 8.
9. Lu, S.P.; Wang, R.; Zhong, T.; Rosin, P.L. Large-capacity Image Steganography Based on Invertible Neural Networks. I n Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10811–10820.
10. Liu, J.; Ke, Y.; Zhang, Z.; Lei, Y.; Li, J.; Zhang, M.; Yang, X. Recent Advances of Image Steganography With Generative Adversarial Networks. IEEE Access 2020, 8, 60575–60597.
11. Duan, X.; Jia, K.; Li, B.; Guo, D.; Zhang, E.; Qin, C. Reversible Image Steganography Scheme Based on a U-Net Struc ture. IEEE Access 2019, 7, 9314–9323.
12. Duan, X.; Liu, N.; Gou, M.; Wang, W.; Qin, C. SteganoCNN: Image Steganography with Generalization Ability Based on Convolutional Neural Network. Entropy 2020, 22, 1140.
13. Abadi, M.; Andersen, D.G. Learning to Protect Communications with Adversarial Neural Cryptography. arXiv 2016, arXi v:cs.CR/1610.06918.
14. Zhu, J.; Kaplan, R.; Johnson, J.; Fei-Fei, L. HiDDeN: Hiding Data With Deep Networks. In Proceedings of the ECCV, M unich, Germany, 8–14 September 2018.

15. Tancik, M.; Mildenhall, B.; Ng, R. StegaStamp: Invisible Hyperlinks in Physical Photographs. In Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
16. Hu, D.; Wang, L.; Jiang, W.; Zheng, S.; Li, B. A Novel Image Steganography Method via Deep Convolutional Generativ e Adversarial Networks. IEEE Access 2018, 6, 38303–38314.
17. Liu, M.m.; Zhang, M.q.; Liu, J.; Zhang, Y.n.; Ke, Y. Coverless Information Hiding Based on Generative adversarial netw orks. arXiv 2017, arXiv:cs.CR/1712.06951.
18. Fridrich, J. Steganography in Digital Media: Principles, Algorithms, and Applications; Cambridge University Press: Cam bridge, UK, 2009.
19. Li, B.; He, J.; Huang, J.; Shi, Y.Q. A Survey on Image Steganography and Steganalysis. J. Inf. Hiding Multim. Signal Pr ocess. 2011, 2, 142–172.
20. Rehman, A.u.; Rahim, R.; Nadeem, M.S.; Hussain, S.u. End-to-End Trained CNN Encoder-Decoder Networks for Imag e Steganography. In Proceedings of the Computer Vision—ECCV 2018 Workshops—Munich, Germany, 8–14 Septemb er 2018, Proceedings, Part IV; Leal-Taixé, L., Roth, S., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Ger many, 2018; Volume 11132, pp. 723–729.
21. Dong, S.; Zhang, R.; Liu, J. Invisible Steganography via Generative Adversarial Network. arXiv 2018, arXiv:abs/1807.0 8571.
22. Tang, W.; Li, B.; Tan, S.; Barni, M.; Huang, J. CNN-Based Adversarial Embedding for Image Steganography. IEEE Tran s. Inf. Forensics Secur. 2019, 14, 2074–2087.
