Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230614

New encoders based on depthwise separable convolution for SteganoGAN to improve steganography efficiency

Ziliang Wang

Macau University of Science and Technology, Macau, China

wry.crappie.0s@icloud.com

Abstract. In recent years, deep learning has developed vigorously, including many applications in the field of image steganography, of which steganogan is a representative. However, its network structure is a generative countermeasure network based on convolutional neural network (CNN). Because the generative adversial network (GAN) is difficult to train, there are problems such as mode collapse and gradient disappearance. Training the generative countermeasure network often requires many rounds of training, consuming a lot of computing resources. In this paper, the idea of deep separable convolution of mobile net is used for reference, and the encryption and generator in the generation countermeasure network for encryption is changed into a depthwise separable convolution encoder, which is compared with the original model, improves the speed of training and reasoning, and facilitates steganography on mobile devices and some low-performance devices.

Keywords: deep learning, steganography, SteganoGan, CNN, MobileNet, Generative Adversarial Network, Depthwise separable convolution.

1. Introduction

Steganography is a technology that hides information into carrier files. As early as ancient Greece, some people shaved the hair of slaves, then wrote words on it, waited until the hair of slaves grew out, and then sent letters. This is a very early steganography [1]. From this example, we can see that the process of steganography is divided into encryption and decryption. In an ideal state, this process must be completed by the sender and the receiver. This process is very difficult [2]. How to prevent the opponent from discovering that the carrier has been steganographic, or even if it is found, it is impossible to extract secret information. This is a big problem.

With the development of computer networks, the growth of multimedia information is explosive [3]. If you want to transmit some information that you don't want others to find, you need to hide the information in the carrier. Among them, image is a key field. Image has a wide range of applications,
Image steganogan used in this paper is a steganogan steganography method based on carrier modification. Carrier modification requires transforming the information of the original image, and then generating a new encrypted image. The less easily the generated encrypted image is detected by others with steganography detection tools, the better [4]. Before 2010, the traditional method could only steganogan 0.4 bits per pixel. Steganogan used in this paper can achieve 4.4 bits per pixel. Traditional steganogan steganogan faces the problems of low steganographic capacity and easy detection by

DOI: 10.54254/2755-2721/6/20230614

steganographic detection tools. Steganogan can effectively improve the steganographic performance through deep learning [5]. Steganogan is composed of the generator encoder of four-layer convolutional neural network, the decoder of three-layer convolutional neural network and the critical of three-layer convolutional neural network, Decoder is used to extract secret information from the secret-containing image, and critical is used to evaluate the quality of the generated image and give a score [6,7]. The error of the decoded information by decoder and the score of critical will be used to optimize the objective function.

Deep separable convolution first appeared in an article called "rigid motion scattering for image classification" "In the doctoral thesis of xception and mobilenet, it is mainly used in lightweight networks. The channels are convoluted separately through the depth convolution, and each channel is combined through the pointwise convolution [8]. It is assumed that the image size is $w * h * c$, and the convolution kernel size is $a * a * n$. The number of operations required is $w * h * a * a * c * n$. however, for the depth convolution, because each layer is convoluted separately, only w is required $* H * a * a * c$ times of calculation, the number of pointwise convolution is $w * h * c * n$, and the total number of discrete discrete convolution is $w * h * a * c + w * h * c * n$, which is $1 / n + 1 / (a \text{ ^ } 2)$ compared with the number of operations of general convolution. We can find that discrete discrete convolution greatly improves the computational efficiency. This convolution method is mainly inspired by mobilenet, Mobile net is a powerful depthwise separate convolution method, which greatly saves computing resources and is an efficient method. This paper introduces this method into the design of encoder encryption network to improve the efficiency of the original steganogan network.

There were three different encoders in steganogan, namely, basic encoder, residual encoder and dense encoder. This paper introduces discrete separate convolution based on these three encoders to form three new encoders. These new encoders reduce the speed and cost of calculation and improve the efficiency of network encryption [9].

2. Related works

2.1. Generative adversarial network [10]

The generative adversarial network (GAN) was proposed by Ian J. Goodfellow, Jean Pouget Abadie and others in 2014. The idea of the generative adversary network comes from game theory, in which the generator and the discriminator game with each other and eventually reach Nash equilibrium. The basic structure of the generative adversary network has the disadvantage of slow convergence, or even non-convergence. The generator and discriminator play infinitely.

The generative countermeasure network is composed of a generator g and a discriminator D. the input of the generator g is random noise, and the input of the discriminator G is the real picture and the random input of the generator g. then the discriminator verifies the difference between the output and the real distribution through cross-entropy. The generator and the discriminator game with each other until they reach a Nash equilibrium. What is a Nash equilibrium, That is, both generator g and discriminator d have reached the strategy of maximizing their own interests. Discriminator D cannot distinguish between true and false, and generator g cannot generate a more realistic image. At this time, the distribution of the image generated by generator and the real image is almost the same.

DOI: 10.54254/2755-2721/6/20230614

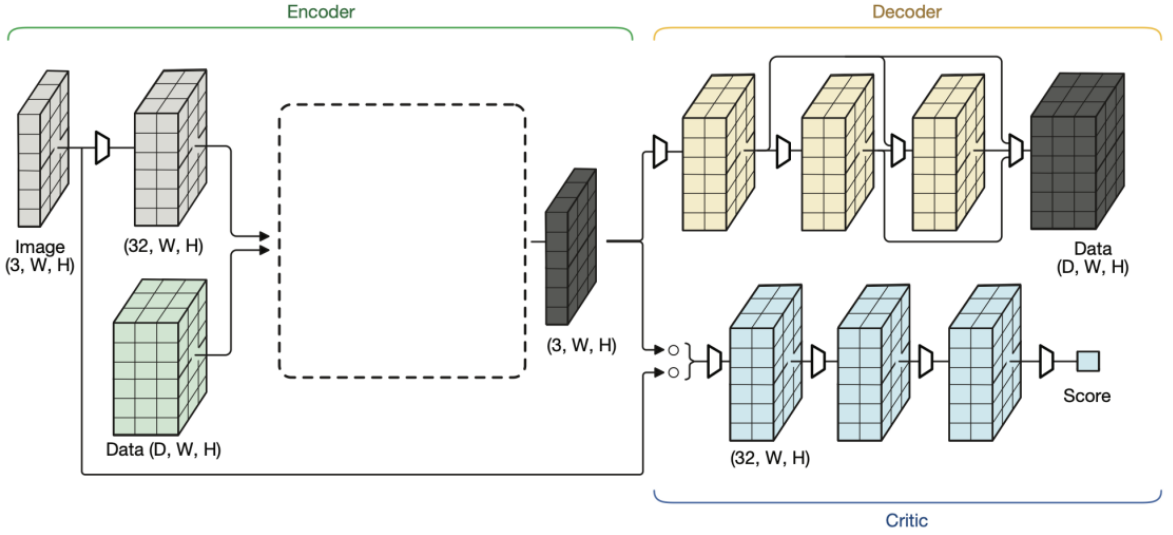

Figure 1. SteganoGAN.

Figure 1 shows the structure of the SteganoGAN, the picture is from the original paper, We can see that if we want to generate high-resolution images that confuse the real with the false, we need to improve the depth of the model, which will use a lot of computing resources and storage resources. In some scenarios with limited performance, such as the scenario of mobile devices, encryption and decryption will take a long time, which is inefficient. So we introduce the idea of profound separable evolution, which reduces training, The occupation of computing resources and time for encryption and decryption.

The following function is the objective function used for optimization of the generative countermeasure network. The following is the optimization function of generator g and discriminator D in the original paper. We can see that this is a form of cross-entropy, which is used to measure the difference between the distribution of generated images and the distribution of real images, and then minimize this difference.

$$\min_{G}\max_{D} V(D,G) = E_{X\sim Pdata(x)}[\log D(x)] + E_{z\sim P_z(z)}[\log(1 - D(G(z)))]$$

Finally, as shown in the original paper in the figure below, the distribution of the gradually generated image is the same as that of the real image. The blue dotted line is the discriminator

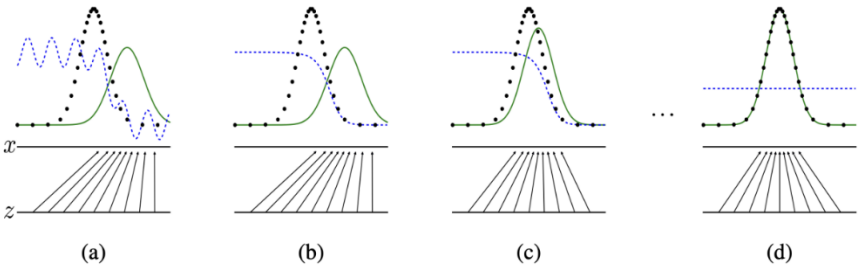

Figure 2. GAN training process.

DOI: 10.54254/2755-2721/6/20230614

2.2. SteganoGAN [11]

Steganogan is a steganography method used to carry modified steganography. The data is steganographed by concatenating with the original data in the middle layer of the encoder and continuing to process. When deep learning was not used before, it was only 0.4 bits per pixel. After deep learning, steganogan structure realized 4.4 bits per pixel, reaching the SOTA at that time. The following is steganogan's network structure. We can see that it is a very clear method.

SteganoGAN optimizes the generated image through the confrontation between Encoder, Decoder and Critical. The Decoder is used to decode information. Input the image generated by the Encoder, and then output the decoded information. It can be seen that the Decoder in the figure is composed of three layers of dense neural networks, which is a very simple structure. The structure of the Encoder is a little more complex than that of the Decoder, and the image without information is input. In addition, the image information is converted into 32 channels through a layer of convolutional neural network, then the information is connected into 33 channels, and then processed into 3 channels through two layers of 32 channel convolutional neural network. Critical is a neural network used for scoring. This neural network scores the model and the generated image containing information, and minimizes this score.

The optimization function of the model consists of the following three parts. The first part is the accuracy of a decoder. The decoded information is compared with the original information, and a cross-entropy formula is used

$$\theta_d = E_{X\sim P_C} CrossEntropy(D(\varepsilon(X,M)),M)$$

The second part is to measure the similarity of image generation. Here, a mean square error is used to measure

$$\theta_s = E_{X\sim P_C} \frac{1}{3 * W * H} ||X - \varepsilon(X,M)||_2^2$$

The third part is the evaluator, which is used to calculate a score. The lower the score, the more authentic the evaluator critical thinks the data is.

$$\theta_r = E_{X\sim P_C} C(\varepsilon(X,M))$$

The final objective function is to minimize the sum of these three. We can find that we should not only achieve similarity with the original image, but also make the effect of the decoder as good as possible

$$minimize\ \theta_d + \theta_s + \theta_r$$

The following is the objective function of critical's update. The smaller the difference between X and the steganographic image, the better

$$\theta_c = E_{X\sim P_C} C(X) - E_{X\sim P_C} C(\varepsilon(X,M))$$

2.3. MobileNet [9]

Mobilenet is a method proposed by Google for computing on mobile devices, because it convolves separately through multiple channels and unifies each layer with pointwise. This model reduces the number of weights of the model through layered convolution, so that the model shows strong performance

DOI: 10.54254/2755-2721/6/20230614

3. Method

In this paper, the blank part of the encoder is replaced by a depth wise convolution layer and a pointwise convolution layer, and a depth separate convolution encoders is proposed, shown on figure 3

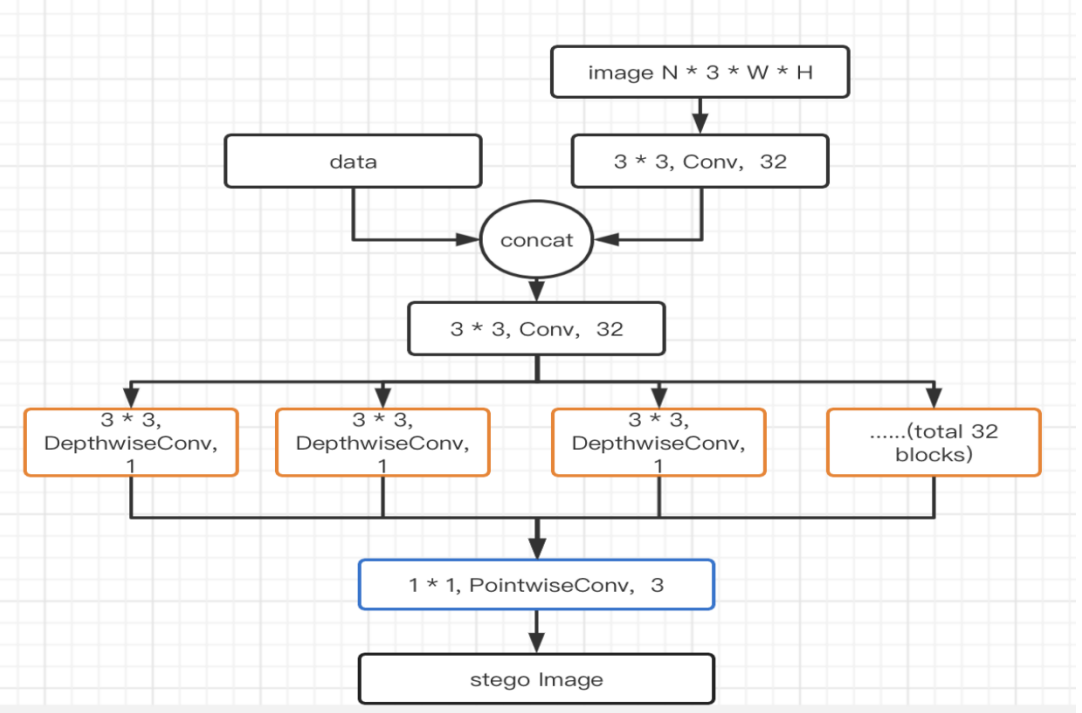

Figure 3. Depthwise Separable Convolution Encoder.

3.1. D epthwise convolution

3.1.1. What is depthwise convolution? The depthwise convolution layer will convolute each input channel separately, and the convolution kernel size used for each channel is $3 * 3$, but it is worth noting that the convolution kernel used for each channel is different, which is equivalent to two-dimensional convolution for each channel.

3.1.2. What is the pointwise convolution layer? The pointwise convolution layer corresponds to the depthwise convolution layer. Since the depthwise convolution layer convolutes each layer separately, there is no relationship between the data between the channels, so it needs the pointwise layer to uniformly convolute them with the pointwise convolution layer for each channel of a pixel point. The convolution kernel size is $1 * 1$. This convolution kernel, which is only $1 * 1$ in size, considers the relationship between the data of the channels in depth, thus unifying the data of all channels.

3.2. Method effect analysis [9]

This kind of convolution will greatly reduce the calculation amount and improve the calculation speed. Assuming that the image size is $w * h * c$ and the convolution kernel size is $a * a * n$, the number of operations required is $w * h * a * a * c * n$. however, for depthwise convolution, because each layer is convoluted separately, only $w * h * a * a * c$ calculations are required, and the number of pointwise convolutions is $w * h * c * n$, The total number of convolutions of the discrete discrete convolution is $w * h * a * a * c + w * h * c * n$, which is $1 / n + 1 / (a \wedge 2)$ compared with the number of operations of the general convolution. We can find that this is a very efficient method, which improves the efficiency without significantly reducing the output effect of the model [12]. Moreover, the following experiments will show that the dense encoder takes the most complex training

DOI: 10.54254/2755-2721/6/20230614

time, The corresponding DSC dense encoder also improves the efficiency the most. Since there is no obvious connection between the input channels, the accuracy of the model will decline to a certain extent, which is indistinguishable to the naked eye [13]. Moreover, there is no network imbalance, making the decoder far better than the encoder, making the model unable to converge, and still able to encrypt and decrypt normally [14,15]. Therefore, this method is effective.

3.3. SteganoGAN's Encoder

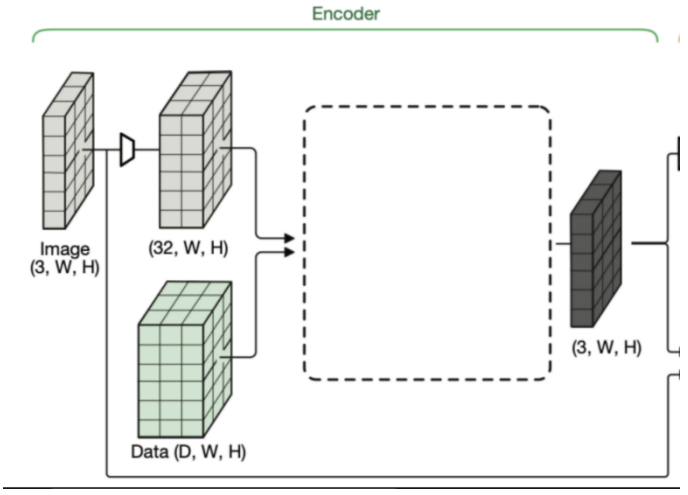

Figure 4. Encoder structure.

The structure of the original encoder given by the author is as Figure 4 and 5. There are three types, namely, basic encoder, residual encoder and dense encoder.

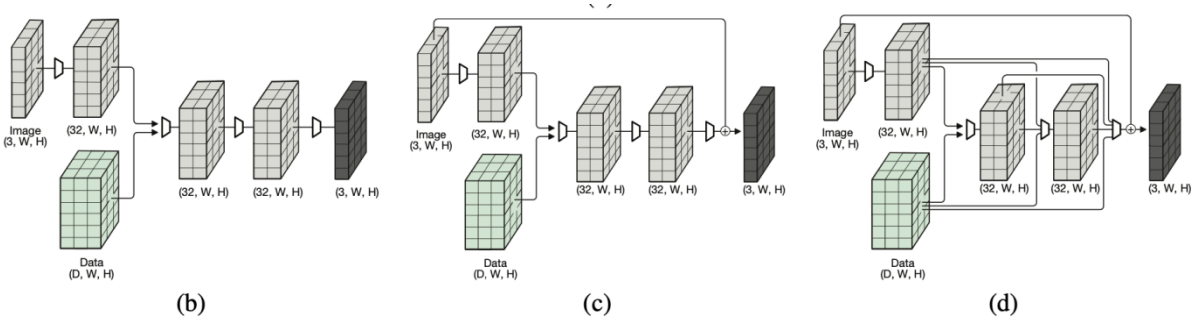

Figure 5. SteganoGAN's original encoders.

3.4. Depthwise separable convolution encoder

As shown in the figure at the previous page, this is the new encoder basic structure based on the discrete separate evolution proposed in this paper, and three different encoders are designed according to this basic structure.

The orange part is depthwise convolution layers, the input channel is 1, and each layer is convoluted separately. The blue part is pointwise convolution layer, and the input channel is the output of depthwise convolution layer. The output is the final image, corresponding to steganogan's cover image, and the output is also $n * 3 * w * h$ steganographic image.

4. Experiments

Experimental environment:

DOI: 10.54254/2755-2721/6/20230614

Ubuntu 18.04, NVIDIA Tesla T4

Our experimental data set adopts the div2k data set, which is composed of 1000 high-definition pictures. In this paper, 800 pictures are used for training and 100 pictures are used for verification. They are basic encoder, residual encoder and dense encoder. The network structure proposed in this paper has three corresponding encoders designed according to the network structure proposed in this paper, namely depthwise basic encoder and depthwise residual encoder, Depthwise dense encoder trained for 10 rounds using the training code given by steganogan's author, and recorded the time spent in training and encode.

Here are my experiment results:

Table 1. SteganoGAN's model.

	Basic Encoder	Residual Encoder	Dense Encoder
Trinning speed	17min14.682826s	17min18.342445s	19min09.1110s
Encoding speed	1.102814s	1.128548s	1.225709s

Table 2. Depthwise encoders.

	Depthwise Basic Encoder	Depthwise Residual Encoder	Depthwise Dense Encoder
Trinning speed	17min07.993978s	17min06.71.774s	18min4.905766s
Encoding speed	1.088569s	1.103925s	1.189193s

5. Conclusions

SteganoGAN is a technology that hides information in multimedia files. Deep neural networks not only improve the quality of steganography, but also reduce the probability of being detected by steganoGAN. SteganoGAN is a high-capacity steganography framework. By introducing Deepwise Separable Convolution, the speed of steganography is improved. This paper also finds that the more complex the network, the better the effect of Deepwise Convolution. This is helpful to the research of improving the efficiency of complex steganography networks [16]. Because steganogan has a simple structure, the improved efficiency is not obvious. Introducing this method to more complex large networks can better reflect the advantages of this method [17]. This paper improves the speed of model training and encoding and decoding by introducing the depthwise separate evolution in mobilenet into steganogan, and analyzes the impact of model complexity on the speed. Denseencoder, which has the best encoding effect, has the largest improvement, so that steganogan can be better applied to mobile devices, reducing the number of weights of models, and providing a reference for the design of lightweight steganography networks.

References

[1] Yang J, Liu K, Kang X, et al. CNN based adversarial embedding with minimum alteration for image steganography[J]. arXiv: Multimedia, 2018.

[2] Tang W, Tan S, Li B, et al. Automatic steganographic distortion learning using a aenerative adversarial network [J]. IEEE Signal Processing Letters, 2017, 24(10):1547-1551.

[3] Ronneberger O, Fischer P, Brox T. U -Net: Convolutional networks for biomedical image segmentation[C]//International conference on medical image computing and computer-assisted intervention. Springer, 2015: 234–241.

DOI: 10.54254/2755-2721/6/20230614

[4] Shi H, Dong J, Wang W, et al. SSGAN: Secure steganography based on generative adversarial networks[J]. IEEE Access, 2018: 38303-38314.
[5] Qian Y, Dong J, Wang W, et al. Deep learning for steganalysis via convolutional neural networks[C]//Media Watermarking, Security, and Forensics 2015. International Society for Optics and Photonics, 2015, 9409: 94090J.
[6] Rubner Y , Tomasi C , Guibas L J. The Earth Mover\"s Distance as a Metric for Image Retrieval [J]. 2000, 40(2):99-121.
[7] Yang J, Liu K, Kang X, et al. CNN based adversarial embedding with minimum alteration for image steganography[J]. arXiv: Multimedia, 2018
[8] Tang W, Tan S, Li B, et al. Automatic steganographic distortion learning using a aenerative adversarial network [J]. IEEE Signal Processing Letters, 2017, 24(10):1547-1551.
[9] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications” https://arxiv.org/abs/1704.04861
[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, “Generative Adversarial Nets”
[11] Kevin Alex Zhang, Alfredo Cuesta-Infante, Lei Xu, Kalyan Veeramachaneni “SteganoGAN: High Capacity Image Steganography with GANs” https://arxiv.org/abs/1901.03892
[12] Kong X Z, Yao Z Q. A novel double 3D digital wa-termarking scheme[A].Piscataway: IEEE Computer Society Press, 2009: 553-556.
[13] Han D Z, Yang X Q, Zhang C M.A novel robust 3D mesh watermarking ensuring the human visual system[A]. Pis-cataway: IEEE Computer Society Press,2009:705-709.
[14] Li S and Zhang X. Towards construction based data hiding: From secrets to fingerprint images[J]. IEEE Transactions on Image Processing, 2019, 38(3): 1482-1497.
[15] Hu D, Wang L, Jiang W, et al. A novel image steganography method via deep convolutional embedding based on adversarial training [C]//ACM Turing Celebration Conference, 2020.
[16] Pantic N, Husain M I. Covert botnet command and control using twitter[C]//The 31st Annual Computer Security Applications Conference. ACM, 2015: 171-180.
[17] Zhang Z, Liu J, Ke Y, Lei Y, Li J, Zhang M and Yang X. Generative steganography by sampling[J]. IEEE Access, 2019: 118586-118597.