A NOVEL APPROACH BASED ON MODIFIED CYCLE GENERATIVE ADVERSARIAL NETWORKS FOR IMAGE STEGANOGRAPHY

P.G. KUPPUSAMY, K.C. RAMYA, S. SHEEBA RANI, M. SIVARAM, AND VIGNESWARAN DHASARATHAN

Abstract. Image steganography aims at hiding information in a cover medium in an imperceptible way. While traditional steganography methods used invisible inks and microdots, digital world started using images and video files for hiding the secret content in it. Steganalysis is a closely related field for detecting hidden information in these multimedia files. There are many steganography algorithms implemented and tested but most of them fail during Steganalysis. To overcome this issue, in this paper, we are proposing to use generative adversarial networks for image steganography which include discriminative models to identify steganography image during training stage and that helps us to reduce the error rate later during Steganalysis. The proposed modified cycle Generative Adversarial Networks (Mod Cycle GAN) algorithm is tested using the USC-SIPI database and the experimentation results were better when compared with the algorithms in the literature. Because the discriminator block evaluates the image authenticity, we could modify the embedding algorithm until the discriminator could not identify the change made and thereby increasing the robustness.

Key words: Generative Adversarial Network (GAN), PSNR, Steganalysis, LSB replacement, Cryptography, generator, discriminator.

AMS subject classifications. 94A60

1. Introduction. Steganography deals with embedding data in some cover medium. The content that is getting hidden should be not visible both perceptually as well as statistically. The embedding capacity also needs to be as high as possible with best robustness as well. Fig. 1.1 depicts a basic image steganography model.

The steganography process involves a cover image and a stego-key to hide the secret data in to the cover medium and produces a stego image as output. There are many algorithms discussed in the literature including those in spatial domain and transformed spectral domain. Though image steganography has many advantages compared to cryptography as they do not attract attention directly as in the case of the latter, they are still not very robust with the present day algorithms. There are many challenges including those through the Steganalysis methods that can identify the presence of secret data. This includes identification of visual attacks, statistical attacks and structural changes through different Steganalysis approaches. To overcome these issues, we propose a new methodology which includes inbuilt discriminator network to identify the above mentioned attacks during the initial stages of steganography process before producing the output stego image. Generative Adversarial Networks aims to generate data from zero and through reinforcement learning, the learning rate is higher. Since the generator alone can produce or create only some random noise, we need a mechanism to provide feedback to the generator. The discriminator portion of the GAN model will provide inputs to the generator model on what and how to create. So, when we have some secret data to be hided in a cover image, the discriminator will provide feedback to the generator on the best possible way to hide the data in an imperceptible way. We use real and generated images together for GAN’s to learn initially on what is real and
what is fake as shown in Fig 1.2 and the proposed algorithm is tested using the USC-SIPI image dataset and the results are encouraging.

2. Related Work. Steganography, the method of concealing the secret data like an image or a text within another plain image in the most imperceptible manner possible. Modern steganography techniques exists as early as from 1985 by concealing messages in image files, chaffing and winnowing, pictures hided in video materials, changing the elements order in a set and so on. R. Chandramouli and N. Memon presented a study method that uses LSB technique for image steganography [1] through which large capacity of around 25% of the compressed data size were obtained at minimal degradation of image quality. J. Spaulding, M. N. Shirazi and E. Kawaguchi discussed about information hiding through various techniques [2]. They have discussed various applications of information hiding along with concealing techniques in this paper. Abbas Cheddad, Joan Condell, Kevin Curran and Paul Mc Kevitt also made a survey much later [3] and analysed the methods at that point of time. The common standards of image steganography and guidelines drawn from the literature are discussed in depths.

Wojciech Mazureczyk, Wendzel, Sebastian Zander, Houmansadr and Szczypiorski, detailed about information hiding in communication networks, the fundamentals, mechanisms, applications, and countermeasures [10] in their Wiley published IEEE book. This book covers the network steganography in detail along with traffic types, network flow and steganography counter measures. Prosanta Gope, Anil Kumar and Gaurav Luthra, have brought out the image steganography in compressed domain [4] with encryption technique. They have proposed an enhanced JPEG image steganography methodology added with a suitable encryption technique for achieving better results in data hiding mechanism.

Anjali and Umesh have again used wavelets for a secure skin tone based image steganography [5]. They describe on how Discrete Wavelet Transform (DWT) can outperform the Discrete Cosine Transform (DCT) when the secret message is hided on one of the high frequency DWT sub band. M.Vijay and V.VigneshKumar brought in image steganography method using integer wavelet transform [6]. They have used wavelets to hide secret data in the source image and found that higher embedding capacity with lower distortion rate can be
achieved through embedding the secret data on the LSB’s of the wavelet coefficients.

Deepika Kulshreshtha and Kamal Kumar in their research have used Discrete Cosine Transform for high capacity and lossless steganography [9]. Their results indicate that the mean square error almost reaches zero even after embedding the message onto the source image. Though there are multiple techniques discussed in the literature on both spatial and transformed domain, most of them fail during Steganalysis and the need for robust image steganography algorithm still remains. We started looking in the direction of techniques that has built-in Steganalysis methods to overcome these challenges and that made us to use Generative Adversarial Networks with built-in discriminative networks for robust image steganography.

Ian J. Goodfellow and his team first proposed a framework for generative model using an adversarial process [10] that includes both a generator network and discriminator network in it. They proposed a new framework for finding generative model through an adversarial process. They had two models defined in their architecture, a generative model for data distribution and a discriminative model for estimating the output data. Their results were outstanding when tested with various data sets. Their framework can also be further extended for a conditional generative model, learned approximate inference, semi-supervised learning and efficiency improvements. They have used MNIST data set for training adversarial nets and also the Toronto Face Database, and CIFAR.

Denis Volkonskiy, Ivan Nazarov, Boris Borisenko and Evgeny Burnaev have discussed about detection of payloads through binary classifier [11] and also proposed a new model based on deep convolutional generative adversarial networks. Their research work also showcased the viability of the adversarial framework and the challenges involved in generative modeling. Kevin Alex Zhang, Alfredo Cuesta-Infante, Lei Xu and Kalyan Veeramachaneni have used GANs for high capacity image steganography [18]. Their approach could achieve payloads of 4.4 bpp and is found to be efficient when tested on multiple data sets.

Ru Zhang, Shiqi Dong and Jianyi Liu further used these Generative Adversarial Networks [15] for image steganography. They had three major contributions in their work including a) hiding data only on the luminance channel of the grey image, b) Presence of GAN architecture strengthen their security and c) construction of a mixed loss function which helps to generate more realistic stego images. Though they have proposed a robust method of image steganography, it can be applied only on stego images that are lossless otherwise the secret data portions gets lost. To overcome these short comings and to implement a robust algorithm, we proposed to use a modified cycle GAN approach which gives better results even in case of slightly lossy images.

3. Generative Adversarial Networks in image steganography. For hiding the secret data in images, we need to select the right set of images for training and testing purposes. There are many times where do not have sufficient data to create a model. Generative Adversarial Networks (GAN’s) can learn about the data and to synthesize or generate never before seen data to augmented data set. So, they can be considered as an unsupervised learning technique or semi-supervised learning.

While machine learning helps in the study of statistical models and algorithms that systems use to perform a function without explicit instructions, they require training before the task can be accomplished. Supervised, semi-supervised and unsupervised machine learning algorithms are the different types present in the literature while the GAN’s fall under semi-supervised learning methodology where the model uses a small amount of labelled data along with large quantity of unlabelled data for training.

Adversarial can be considered as a competition between two components including the generator and the discriminator. The generator replicates real data to produce fake data while the discriminator helps to distinguish between the real data and the fake generated data. The discriminator loss will then be passed as feedback to generator to produce a new set of data as an objective function. We make use of this principle in our image steganography technique to improve the security by minimizing the error between the stego images and natural images until the discriminator can no longer identify the hidden data.

GANs were first introduced by Ian Goodfellow and his team at the University of Montreal and their applications are huge as they can mimic any data distribution. Fig 3.1 below depicts the basic GAN:

The first step to train a GAN is to define the problem which in our case is to hide the secret data on the cover image. We define the GAN architecture next describing the tasks of the generator and the discriminator. The discriminator model then needs to be trained for classifying the real and fake data. The training here uses both the real images and the fake images produced by the generator and labelled as fake. We then train the
generator using the loss of the discriminator as the objective function. The generator parameters are fine-tuned on every iteration to maximize the loss of discriminator. The generator and discriminator training is continued in loop until the discriminator cannot classify between the real and fake data as shown in Fig 3.2. Once the training is complete, we synthesize data from generator and the system is ready for live testing.

The discriminator uses cross entropy loss as we are working on a classification model and also they give a better performance than misclassification rate and MSE. The predictions are also better with entropy loss as compared to other techniques. Learning also does not get stalled at any stage of the training with this method. The cross entropy loss between the true distribution \( p \) and the estimated distribution \( q \) is given by:

\[
H(p, q) = - \sum p_i \log(q_i) \tag{3.1}
\]

The discriminator used here is just a binary classifier and the output can hence be either true or false. So equation (1) gets reduced to:

\[
H(p, q) = -(1 - p_i) \log(1 - q_i) - p_i \log(q_i) \tag{3.2}
\]

Equation 3.2 represents a loss for single sample. The overall loss for \( N \) samples can be computed as well. Considering that almost half our samples comes from the real images while the other half are just imaginary or fake, we can write the predicted output of discriminator as follows:

\[
E_x P_{\text{generator}[D(x)]=0} \tag{3.3}
\]

\[
E_x P_{\text{data}[1-D(x)]=0} \tag{3.4}
\]

There are many different types of GAN’s available including the basic GAN, deep convolutional GAN, conditional GAN, Info GAN, Wasserstein GAN, Attention GAN and cycle GAN. We discuss in detail about cycle GAN and our proposed algorithm in the coming sections.

4. Proposed Cycle GAN Method. Image to image translation aims at learning the mapping between an input source image and the output image using multiple training data sets. There are cases or applications where paired training data sets are not available. So, there is a need for an efficient approach to produce an image with data hided in it using a source image. Let the source image be represented by \( X \) while the output stego image be represented by \( Y \) and the goal here would be to map \( G : X \rightarrow Y \) in such a way that images from \( G(X) \) is indistinguishable from \( Y \) using an adversarial technique. The transfer happens without one-to-one mapping between images from source to output in the training data set. Cycle GAN is capable of achieving
A Novel Approach Based on Modified Cycle Generative Adversarial Networks for Image Steganography

The need for pairing is eliminated due to the two step transformation: a) mapping it to the target domain and b) back to the source image. This mapping or hiding the data is performed using the generator. The quality of the newly generated image is then improved by using the discriminator network.

Both the generator and discriminator are trained together and over a period of time, the generator produces samples using the underlying distribution while the discriminator tries to guess randomly. It is a kind of graphics problem in which the goal is to learn the mapping of an input image to output. There are cases or exceptions during which the adversarial loss alone is not sufficient enough to confirm that the learned function can map an input to desired output. So, there is a necessity of consistency required and hence cycle-consistent GANs were required to overcome this issue.

Cycle GANs look at meaning mapping between the input image and the generated image. So, cycle GANs works by first taking an input image and fed to the generator to produce a new image which in turn is fed to another generator to make sure that the new image matches to the one in the original domain without much deviation. So, it contains 3 parts namely the encoder, the transformator and the decoder. The encoder takes care of feature extraction from the input image, the transformator in turn uses these features and combines them together while the decoder uses deconvolution methods for reverse encoding.

If the discriminator outputs $D(x)$ predicting the chance of $x$ being a real image, then the objective is maximum likelihood of the data observed. This objective is written as follows:

$$
\max_D V(D) = E_{x \sim p_{data}} \log D(x) + E_{z \sim p(z)} \log(1 - D(G(z)))
$$

Similarly the objective function on the generator side is defined as follows:

$$
\min_G V(G) = E_{z \sim p(z)} \log(1 - D(G(z)))
$$

Here $G$ needs to be optimized to the extent that can fool the discriminator.

4.1. Proposed Algorithm (Modified cycle GAN for data hiding and producing stego image). The USC-SIPI dataset which is a collection of digitized images for supporting research in image analysis, segmentation, processing and machine vision is used for image steganography as well in our work. Most of the images in the data set used for training are color images. In order to embed the secret data in to the image set, we prefer to use only the luminance part of the image. For this purpose, e fist convert the RGB images in to grey image and pull out the luma part from it as follows:

$$
\begin{bmatrix}
Y \\
Cb \\
Cr
\end{bmatrix} =
\begin{bmatrix}
0.256 & 0.504 & 0.098 \\
-0.148 & -0.292 & 0.441 \\
0.441 & -0.369 & -0.071
\end{bmatrix}
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix} +
\begin{bmatrix}
16 \\
128 \\
128
\end{bmatrix}
$$

Here $Y$ represents the luminance information of the image while $Cb$ and $Cr$ represent the color. We embed the secret bits in the LSB's of this $Y$ plane and not on the color data as color distortions are human perceptible easily while brightness changes are less visible. Also, the other bits in the pixel are adjusted during every iteration of the discriminator output in order to produce less error thereby at some point of time the discriminator will not be able to identify the secret data present in the source image.

The hiding capacity of this modified cycle consistent GAN algorithm depends on the number of training data set and the learning rate. The number of bits that gets adjusted in the identified pixel values also improves the proposed techniques embedding capacity. The steps are listed as follows:

1. The color images are converted in to grey scale images and the LSB bits of the $Y$ plane pixel are substituted with the secret bits.
2. The newly generated image is fed as input to the discriminator along with the original images.
3. The discriminator takes both the sets of inputs then returns the output as 0 or 1.
4. The generator and discriminator neural networks run in competition during the training phase in loop. The discriminator is first trained (forward propagated) when the generator is idle and vice versa.
5. During every iteration, the bits that are present in the $Y$ plane before the hidden bits are adjusted according to the feedback from the discriminator.
If \( d \) is the pixel value, then \( d_1 \) represents the value of last \( n \) bits while \( d_2 \) shows value of hidden bits. If \( (d_1 \sim d_2) \leq (2^n)/2 \), then no adjustments are required for that pixel else it requires modification during every iteration. The data hiding process using this modified cycle GAN algorithm is depicted in Fig 4.1.

Cross entropy is the discriminator loss function. One of the frequently encountered problem during training is the gradient saturation problem where the gradients are very small for any type of learning. A poorly designed generator loss function can lead to this issue and we have addressed this problem in our research work. The generators as seen in fig 5 above does the density estimation from the stego data to real data and then provides it as input to the discriminator. We wanted the discriminator to maximize the log probability in prediction and we achieve this by adjusting the MSB pixel values along with LSB values based on the class output from the discriminator. Due to this additional substitution, the pixels gets normalized and the noise is not observed either visually or through any steg analysis algorithms. Cycle GAN learns to perform translations of images without the need for explicit pairs. In other words, there is no need for one-to-one image pairs. Once the training is completed, the generator present in the GAN model should be able to produce stego images that the discriminator or any steg-analysis algorithm will not be able to identify. Through this cycle GAN approach, we are able to define two types of loss:

1. Adversarial loss;
2. Cycle-consistency loss.

The adversarial loss matches the distribution of stego images and the ones in the target domain. The cycle consistent loss adds stability to the training process and helps in producing a more robust system as compared to the traditional GAN networks. The proposed bit manipulation technology overall improves the embedding capacity and the robustness.

5. Results and discussions. Steganography is the science that works towards transmitting the secret data with the help of a suitable carrier normally a multimedia transporter like image, audio or video files. In our work as detailed in the previous sections, we have used image carrier to hide the secret data using GAN algorithm. In order to train and test the proposed modified cycle GAN algorithm, we have used the USC-SIPI database provided by the signal and image processing institute of the Southern California University. The data sets were first published in 1977 and are updated on a periodic basis. The image details and the results are detailed in the coming sections.

The dataset is organized across various types based on the characteristics of the image. They are of type having different textures, aerials, miscellaneous and sequences. The user has to identify the images for reading and access them within programs. The image size in the database varies between 256 * 256 to 512 * 512 and 1024 * 1024. The images present are either color images or grey images. Some of them are black and white images and are 8 bit depth while color images are of 24 bit depth. Fig 5.1 gives the glimpse of the same. The data set used in this work is originally present in TIFF format.
As a first step of our programming, we have to first uncompress the TIFF image. A sample c program is also available for this purpose along with the data set. But we have used MATLAB which is a high level programming language and simple to use for machine learning and image processing applications. While the grey images can be used as such without any conversion, in case of RGB color images, we convert the color images in to grey scale image as shown in figure 5.2 below before proceeding with data hiding techniques using the proposed algorithm. We then select the source image and the secret data for training and image steganography.

We evaluate our proposed GAN method based image steganography through secrecy (difficulty in identifying the stego images), capacity (size of hiding message), and robustness (PSNR comparison). In terms of security, the methods discussed in the literature uses GAN to learn a modification probability matrix have higher visual effects because they depend on the original carrier and the output image quality decides the security. In our method, the key space is too small and the quality does not deter much. In terms of capacity, there is a big gap between the cover synthesis steganography and traditional cover modification steganography.

We embed the secret bits in to the source image and provide it as input to the discriminator to get its feedback. Based on the feedback, the other bits in every pixel of the y plane is adjusted and provided as input again to the discriminator. The generator loss function using different methods are compared in fig 5.3. The discriminator takes both the source image and the stego image and tries to classify which is real and which has hidden data in it. So, there is a double feedback loop and this framework hence resists all types
of steg analysis attacks. 10% of the data set is considered for Steganalysis purposes. The training set from the database is denoted by \( A \), the testing set by \( B \) and the proposed steganography algorithm by stego\((x)\). We embed secret information to the data set and the training set becomes \( A + \text{stego}(A) \) while the testing set becomes \( B + \text{stego}(B) \). The accuracy of the steganalysis trained using real data set is represented in table 5.1.

Both SGAN and cyclic GAN are capable of reducing the detection accuracy of any Steganalysis method but our proposed modified cyclic GAN algorithm is capable of giving even more robust results and with a better embedding capacity. Table 5.2 shows the capacities of various non-modification methods and its seen that the proposed method is better here as well. Table 5.3 represents the PSNR values observed across different image steganography algorithms. Even this indicates that the proposed algorithm presents state-of-the-art results as compared with the other techniques in the literature.

Fig 5.4 below represents the Steganalysis detection values comparison across different methods. It is evident that with the proposed method, the chances of stego image detection is almost impossible whereas with the regular LSB substitution method it is quite easy to identify. Also Figure 8 shows the different generator loss functions where minimax outweighs the other two methods.

It is always possible for the distributions derived from training to overlap as shown in Fig 5.5. In such scenarios, the feature space decision boundaries that are proportional to the covariance presented by each class take a quadratic form and decide the output class. So, our proposed classification method has less chances of failing as compared to other methods.
A Novel Approach Based on Modified Cycle Generative Adversarial Networks for Image Steganography

<table>
<thead>
<tr>
<th>Different Methods in Image Steganography</th>
<th>PSNR value in dB across different images</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCT</td>
<td>Lena 47.11, Barbara 46.9, Boat 46.73</td>
</tr>
<tr>
<td>LSB Technique</td>
<td>Lena 53.54, Barbara 53.65, Boat 52.96</td>
</tr>
<tr>
<td>Regular DWT</td>
<td>Lena 46.09, Barbara 46.43, Boat 47.72</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>Lena 64.7, Barbara 58.3, Boat 60.1</td>
</tr>
</tbody>
</table>

Fig. 5.4. Steganalysis detection values comparison across different methods

Fig. 5.5. Discriminator Output classifications

6. Conclusion and future directions. Steganography and also cryptography aims at protecting information from unwanted parties but both the methods have their issues to be resolved. We have studied about various existing image steganographic algorithm present in the literature and came up with a new approach of modified cycle GAN for robust image steganography. The PSNR values witnessed across different techniques illustrates undoubtedly the benefits of the proposed method. Moreover, the stego images produced through the proposed algorithm has got better visual quality as compared to the other methods thereby making the stego image not to be discovered easily. We have also tested our algorithm on the MNIST database [13] containing over 60,000 examples for training and 10,000 samples for testing. The classifier model produced great accuracy
both on the real images as well as on the steganographic images produced by the proposed method. There are both merits and demerits to the proposed framework. Though the system has improved payload capacity, better hiding technique and integrity check for issues encountered in LSB method, the system cannot be overly trained and that can lead to the Helvetica scenario which corresponds to mode collapse in case of generative adversarial networks. Also, the proposed algorithm can be optimized for even better and improved results in terms of capacity and robustness. We aim to address these issues in our next research paper.

7. Acknowledgement. The authors also would like to express their sincere thanks to Prof. Dr. Truong Khanh Nguyen, Division of Computational Physics, Institute for Computational Science, Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam for giving his valuable suggestion, comments and support to complete this work as effective.

REFERENCES


Edited by: Swaminathan JN
Received: Sep 30, 2019
Accepted: Jan 7, 2020