论文题目： Steganographic Generative Adversarial Networks

----from NIPS 2016

一、创新点

1、提出了一种基于深度卷积生成对抗网络（DCGAN）生成图像类容器的新模型。

2、允许使用标准隐写算法生成更多的 setganalysis-secure 消息嵌入。

二、Steganography

1、作者采用±1-embedding 算法，将复杂的信息嵌入到 raster image（位图）。该算法源于LSB算法的关键思想（将秘密消息存储在给定图像容器中每个像素的某个颜色通道的最低有效位（最后位）中），但利用了更具战略性的像素处理技术：对于原始图像X及其带有secret message $\hat{X}$ 的final version，以这样的方式拾取像素以最小化失真函数

$$D(X,\hat{X}) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\rho(X_{ij},\hat{X}_{ij})\left|X_{ij}-\hat{X}_{ij}\right|,$$

$\rho(X_{ij},X_{ij})$代表改变 X 的像素的代价，特定于每个特定的隐写算法.

2、为了检测容器中隐藏信息的存在，通常使用 Steganalysis，将图像与一些隐藏消息与空区分开的阶段通常通过二进制分类来执行，作者提出使用深度卷积神经网络（CNN）进行隐写分析，并表明在使用 CNN 作为分类器时，分类精度会显著提高。

三、Steganographic Generative Adversarial Networks （SGAN）

作者基于DCGAN网络提出SGAN网络，模型架构如下

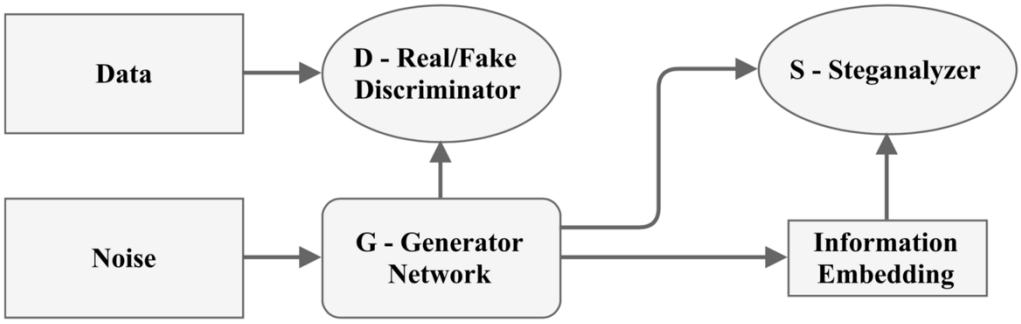

Figure 1: SGAN information flow diagram

模型结构：

•生成器网络 G，可从噪声中产生逼真的图像;

•判别网络 D，其分类图像是合成的还是真实的;

•判别网络 S，即 steganalyser，用于确定图像是否包含隐藏的秘密消息。

作者通过希望生成器生成可用作 secure message 嵌入容器的真实图像， G 同时与模型 D 和 S 竞争。

$$L = \alpha\Big(\mathbb{E}_{x\sim p_{data}(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim p_{noise}(z)}\left[\log(1 - D(G(z)))\right]\Big) + \\ +(1-\alpha)\mathbb{E}_{z\sim p_{noise}(z)}\left[\log S(Stego(G(z))) + \log(1 - S(G(z)))\right] \rightarrow \min_{G}\max_{D}\max_{S}. \quad (4)$$

S(x)表示 x 有 secret message 的概率，作者使用 D 和 S 的误差的凸组合与参数α∈[0,1]，控制生成图像的真实性的重要性和作为容器的质量与隐写分析之间的权衡。对初步实验结果的分析表明，对于α≤0.7，生成的图像是不现实的，类似于噪声。下面列出了 SGAN 组件的随机小批量梯度下降更新规则：

- for D the rule is $\theta_D \leftarrow \theta_D + \gamma_D \nabla_G L$ with

$$\nabla_G L = \frac{\partial}{\partial \theta_D} \Big\{ \mathbb{E}_{x \sim p_{data}(x)} [\log D(x, \theta_D)] + \mathbb{E}_{z \sim p_{noise}(z)} [\log(1 - D(G(z, \theta_G), \theta_D))] \Big\};$$

- for S (it is updated similarly to D): $\theta_S \leftarrow \theta_S + \gamma_S \nabla_S L$ where

$$\nabla_S L = \frac{\partial}{\partial \theta_S} \mathbb{E}_{z \sim p_{noise}(z)} [\log S(Stego(G(z, \theta_G)), \theta_S) + \log(1 - S(G(z, \theta_G), \theta_S))];$$

- for the generator G: $\theta_G \leftarrow \theta_G - \gamma_G \nabla_G L$ with $\nabla_G L$ given by

$$\begin{aligned} \nabla_G L = & \frac{\partial}{\partial \theta_G} \alpha \mathbb{E}_{z \sim p_{noise}(z)} [\log(1 - D(G(z, \theta_G), \theta_D))] \\ & + \frac{\partial}{\partial \theta_G} (1 - \alpha) \mathbb{E}_{z \sim p_{noise}(z)} [\log(S(Stego(G(z, \theta_G), \theta_S)))] \\ & + \frac{\partial}{\partial \theta_G} (1 - \alpha) \mathbb{E}_{z \sim p_{noise}(z)} [\log(1 - S(G(z, \theta_G), \theta_S))]. \end{aligned}$$

作者提出与 GAN 模型的主要区别在于我们更新 G 不仅最大化 D 的误差，而且最大化分类器 D 和 S 的线性组合的误差。

四、实验

1、Data Description

数据集：Celebrities dataset [ZLT15]

图像大小 reshape　64*64

作者将 10%的数据视为测试集，用 A 代表训练集，由 B 代表测试集和用于通过 Stego(x)隐藏信息的隐写算法。嵌入一些信息后获得了 A + Stego(A)和测试集 B + Stego(B)的训练集，最终得到了 38 万张用于隐写分析训练的图像和 20000 张用于测试的图像。如下图所示

Figure 2: Examples of images, generated by SGAN after training for 8 epochs on the Celebrities dataset

对于信息嵌入，作者使用±1-embedding 算法，其有效载荷大小等于每像素 0.4 位，仅三个中的一个通道。 作为嵌入文本，作者使用来自纽约时报的一些文章中随机选择的摘录。

2、Experimental Setup

The structure of the individual steganalyser S* has the form : F(0) filter → Conv2D →Conv2D →Max Pooling →Conv2D →Conv2D →Max Pooling → Fully connected layer (1024 neurons) →Fully connected layer (1 neuron) → Sigmoid function for output.

其中F(0) filter是作者定义的一个特殊用于隐写分析应用的滤波器，如下图所示

$$F^{(0)} = \frac{1}{12}\begin{pmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{pmatrix}.$$

作者对学习率、损失使用的算法等进行设置，并给出了实验的流程：

For training of this steganalyser we use the Adam optimization algorithm on the loss (4) with the learning rate equal to 5^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$. As a loss function we use a binary cross-entropy.

The setup of experiments can be described as follows:

- We train and use the SGAN and/or DCGAN model to generate images to be used as containers;
- We train the independent steganalyser S^* using either real images (sec. 5.3) or generated images (sec. 5.4);
- We measure the accuracy of the steganalyser S^*.

3、Result

作者通过实验给出的结果：

Table 1: Accuracy of the steganalyser S^* trained on real images

Type of a test set \ Image generator	SGANs	DCGANs
Real images	0.962	
Generated images	0.501	0.522

Table 2: Training/testing on generated images according to experimental conditions **C1**-**C3**

Experimental Conditions	Accuracy
C1	0.982
C2	0.517
C3	0.499

Table 3: Training/testing on generated images according to experimental conditions **C4**-**C6**

Experimental Conditions	Accuracy
C4	0.649
C5	0.630
C6	0.581

五、总结

1.作者为 Generative Adversarial Networks 的应用开辟了一个新的领域，即隐写术应用的容器生成;

2.作者采用±1-embedding 算法和测试新方法来进行更多的隐写分析 - 安全信息嵌入：

a）作者证明了 SGAN 和 DCGAN 模型都能够将隐写分析方法的检测精度几乎降低到随机分类器的检测精度;

b）如果初始化具有不同随机种子值的容器的发生器，甚至可以进一步降低隐写检测精度。

改进点：

1、 在更高级的隐写算法上测试 SGAN 方法

2、 采用 CGAN、StarGAN 网络等网络应用在信息隐藏领域