基于CNN对抗嵌入的图像隐写——CNN-Based Adversarial Embedding for Image Steganography

原创

Jhouery 于 2021-10-09 14:13:37 发布 220 收藏 1

文章标签： cnn 深度学习 人工智能

本文链接：https://blog.csdn.net/weixin_48654804/article/details/120668541

CNN-Based Adversarial Embedding for Image Steganography

> https://ieeexplore.ieee.org/abstract/document/8603808

> 这篇论文发表在TIFS（ IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY），TIFS是网络与信息安全领域的国际两大顶级期刊之一，中国计算机学会（CCF）推荐的A类期刊，SCI一区TOP期刊，目前影响因子有7多了。

1 概述

近些年来，深度学习被引入隐写分析领域。基于深度学习的隐写分析技术相比传统的隐写分析，有更高的检出性能。基于深度学习的隐写分析成为隐写术面临的巨大挑战。

这篇论文提出了一种名为对抗嵌入（adversarial embedding ，ADV-EMB）的隐写方案。这个方式在实现了隐藏隐写信息的同时，欺骗基于CNN网络的隐写分析网络。

对抗嵌入根据目标CNN隐写分析网络的反向传播的梯度，挑战图像元素修改的成本。其实就是和对抗样本生成原理相同。

但是直接对抗嵌入整张图像，会导致隐写信息嵌入部分的像素也被干扰，从而可能导致隐写信息的损坏或者丢失。但是所幸，对抗样本的生成并不需要整个图像的参与，还有论文（one pixel attack）只修改一个像素就完成对抗攻击。因此本文的解决方式就是将图像分成两个部分，一个负责隐写信息的嵌入，一个负责添加扰动实现对抗攻击。

2 技术基础

一个分类器

隐写分析的基本要求是区分隐写图像（stego image）和覆盖图像（cover image）。
一个方式是通过一个有监督地机器学习。训练一个二分类器。
分类器有训练数据C和S，并获得一个决策函数：

$$\begin{cases} \mathbf{X} \text{ is a cover image}, & \text{if } \phi_{\mathcal{C},\mathcal{S}}(\mathbf{X}) = 0, \\ \mathbf{X} \text{ is a stego image}, & \text{if } \phi_{\mathcal{C},\mathcal{S}}(\mathbf{X}) = 1. \end{cases} \tag{1}$$

这个函数就是隐写分析

隐写的失真最小化框架

$$\min_{\mathbf{S}} \; D(\mathbf{C},\mathbf{S}), \quad \text{s.t. } \psi(\mathbf{S}) = k, \qquad (5)$$

D代表一种距离度量。这个目标函数说明要求在嵌入前后，C和S之间的距离尽可能小。约束是从S中提取的有效信息载荷

3 技术实现

基本想法：
图像的元素被随机分成两组。一组为common groups，一组为adjustable groups。
隐写嵌入也分为两个阶段。
第一个阶段：使用传统的隐写嵌入方式，在common group中嵌入隐写信息；
第二个阶段：利用对抗嵌入方式，将剩余的隐写信息嵌入adjustable group。 adjustable group嵌入信息后，隐写分析网络将给出一个错误的分类结果。

隐写信息的嵌入方案，两者是相同的。只不过一个是通过传统方式嵌入，一个还需要通过对抗样本方式，通过优化方式嵌入。
两者的隐写信息的提取方式因此也是一样的。所以不需要确定嵌入比率β，直接提取即可。

$$\min \; \beta, \quad \text{s.t. } \phi_{\mathcal{C},\mathcal{S}}(\mathbf{Z}) = 0 \;\; \text{and} \;\; \psi(\mathbf{Z}) = k, \qquad (11)$$

这一个目标函数，β表示adjustable group的元素，占整个图像元素的比率。也就是减少对抗扰动的成本，因为对抗嵌入会更大地修改图像。
约束的话，一个是使得分类为0（cover image），一个是保证隐写信息的有效载荷。

实现

使用JPEG图像作为cover image。
使用Xu-CNN作为目标隐写分析器。
使用J-UNIWARD作为传统隐写嵌入的baseline方案。

步骤

> 计算嵌入损失。
>
> 通过嵌入损失，首先在原始图像C上的common elements上进行嵌入操作。
>
> 加入隐写网络的信息。计算梯度方向。
>
> 嵌入信息。

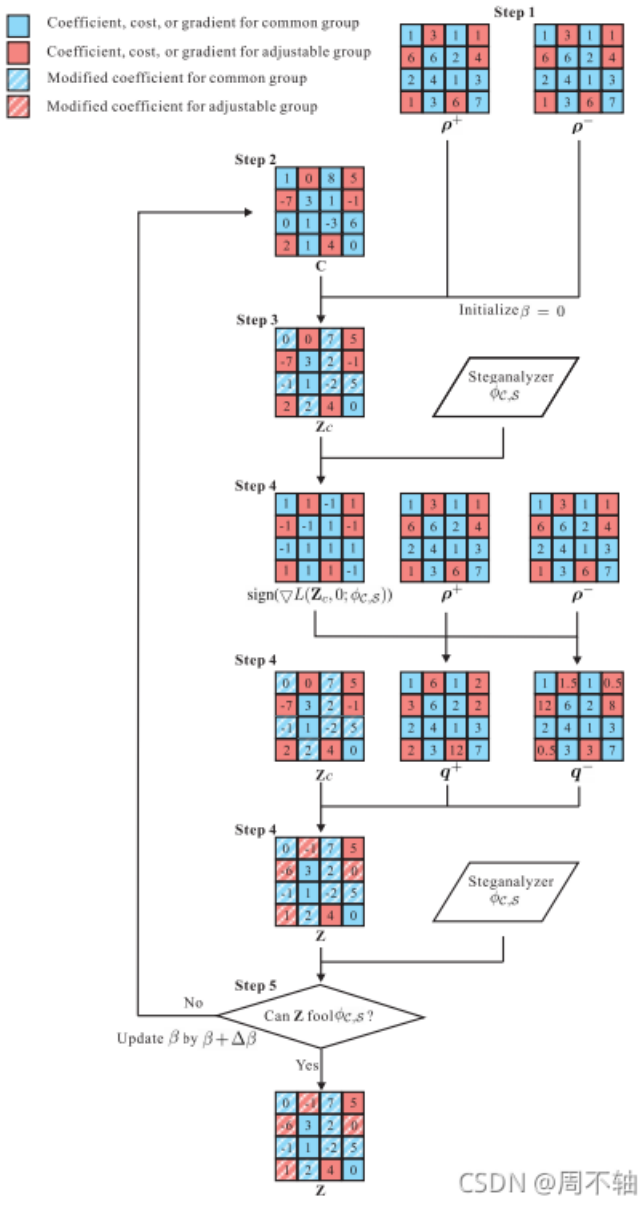
$$L(\mathbf{X}, y; \phi_{\mathcal{C},\mathcal{S}}) = -y \log\big(F(\mathbf{X})\big) - (1-y)\log\big(1 - F(\mathbf{X})\big) \qquad (9)$$

$$q^{+}_{i,j} = \begin{cases} \rho^{+}_{i,j}/\alpha, & \text{if } -\nabla_{z_{i,j}} L(\mathbf{Z}_c, 0; \phi_{\mathcal{C},\mathcal{S}}) > 0, \\ \rho^{+}_{i,j}, & \text{if } -\nabla_{z_{i,j}} L(\mathbf{Z}_c, 0; \phi_{\mathcal{C},\mathcal{S}}) = 0, \\ \rho^{+}_{i,j}\cdot\alpha, & \text{if } -\nabla_{z_{i,j}} L(\mathbf{Z}_c, 0; \phi_{\mathcal{C},\mathcal{S}}) < 0. \end{cases}$$

$$(\rho_{i,j}^{-}, \quad \text{if} \quad \nabla_{z_{i,j}} L(\mathbf{Z}_c, 0; \phi_{\mathcal{C},\mathcal{S}}) < 0, \tag{12}$$

$$q_{i,j}^{-} = \begin{cases} \rho_{i,j}^{-}/\alpha, & \text{if} \quad -\nabla_{z_{i,j}} L(\mathbf{Z}_c, 0; \phi_{\mathcal{C},\mathcal{S}}) < 0, \\ \rho_{i,j}^{-}, & \text{if} \quad -\nabla_{z_{i,j}} L(\mathbf{Z}_c, 0; \phi_{\mathcal{C},\mathcal{S}}) = 0, \\ \rho_{i,j}^{-}.\alpha, & \text{if} \quad -\nabla_{z_{i,j}} L(\mathbf{Z}_c, 0; \phi_{\mathcal{C},\mathcal{S}}) > 0, \end{cases} \tag{13}$$

判断是否能够欺骗分类器，否则，调整β大小，再次进行隐写。



> 在几乎所有的隐写方案中，图像隐写信息的嵌入顺序，都是将原始图像的信息进行随机置乱，这个置乱的规律通过接收和发送双方制定的密钥确定。为了保证安全，本文也建议这么做。

> 隐写算法这部分，详细得要看J-UNIWARD的论文。

实验设置

论文设置了几组实验

1. 使用没有经过对抗隐写样本训练的隐写分析网络进行评估；
2. 使用经过对抗隐写样本训练的隐写分析网络进行评估；
3. 假设隐写术人员和隐写分析人员的知识交替更新；
4. 使用实验证明，为什么由梯度和最小adjustable elements引导的对抗嵌入是有必要的
5. 随机化图像元素的重要性
6. 在另一个图像集上进行实验
7. 评估空域图像性能。

具体实验结果略

结论

很多对抗样本的内容也许可以应用在此。