|  |  |  |  |
| --- | --- | --- | --- |
| **文章信息** | **摘要** | **方法概述及实验结果** | **相关算法介绍** |
| **阅读程度：**  方法精读  **文章标题：**  On Visible Adversarial Perturbations & Digital Watermarking  **中文标题：**  关于可视对抗扰动和数字水印  **发表于：**  CVPR2018  **作者：**  Jamie Hayes  **单位：**  当年在伦敦大学，现在在deepmind | 给定一个机器学习模型，对抗性扰动变换图像，使模型的输出被归类为攻击者选择的类。这一领域的大多数研究都集中在人眼无法察觉到的对抗性扰动上。  然而，最近的研究考虑了可感知的攻击，但局限于图像的一个小区域。在这个威胁模型下，我们讨论了消除这种对抗性干扰的防御，以及可以绕过这些防御的攻击。 | None-blind方法：即事先知道补丁的准确位置然后进行防御，这通常是理想中的情况，毕竟事实通常是不了解位置的。   1. 公式中，指的是补丁内部，指整个图片，指补丁边缘的一个像素点的位置，那么B（u）指在像素点附近的一些像素点的集合，集合中的像素点在的邻域内，在图片上，补丁外。      1. 这个函数提供了由像素坐标到像素值的映射，公式表示点像素值加上改变量得到的是以为基准对点处像素值的估计      1. 显然这样的点有很多，B(u)就是这些点的集合，每一个点都预测一个u点处的像素值，经过加权平均就得到u点处像素值的估计。      1. 对的边缘上每一个位置的像素值进行预测，之后重复上述过程，直到中所有值都被预测。   出自A. Telea. An image inpainting technique based on the fast marching method.  Blind方法：    1: 先找saliency map(这个方法后面记录)，  3: 对于map中大于阈值u的值记为0，其他的记为1，如图二所示  5: 进行几次erosion,dilation操作，并且检测map中像素值为1的这些个连通域的面积，如果面积小于某个阈值，就把整个连通域归0，得到图三  6：再进行一波erosion dilation操作，然后翻转0,1像素值，生成图四作为map  然后用map覆盖原图即可。  实验：三个实验，第一个是用none-blind方法进行的实验，第二个使用blind方法进行实验，都是测试攻击成功率（被成功攻击的样本个数）与防御成功率。第三个是将补丁扩散开来以绕过防御。 | 关于saliency map：  是接受回传的梯度，然后可视化梯度来展示原图中的那些梯度变化最大，认为梯度大的地方是网络关注较多的地方。关于relu层梯度的回传，有三种说法：  **20200826215558698**  第一种比较好理解：正向传递中为负值的输入值，在反向传递中接收到的梯度是0  第二种是反向传递过程中收到负值时，relu将传递0，收到正值的位置将传递此正值。这种思想的意思是（以上图第四行的6这个值为例），右侧的6意味着正向传递过程中，6这个位置的值变化1，会引起最终的输出改变6，如果优化的目标是最大化最终输出，那么这个节点对优化最终目标是正向的作用，因此应该将此节点的梯度应该继续回传。  第三种要求，只有正向输入大于0，反向输入也大于0时才回传这个梯度，**文中选用第三种，并且效果挺好。**  关于erosion与dilation，例如用这个滑动窗口，erosion是腐蚀操作，窗口中数值为1对应的像素值（左上右上左下）将被替换为三者中最小的那个像素值，右下不发生改变。Dilation操作相反，换成三者中最大的那个像素值。  8789591-d7df75e79e9d3058 |