|  |  |  |  |
| --- | --- | --- | --- |
| **文章信息** | **摘要** | **实验一** | **实验二** |
| **阅读程度：**  精读  **文章标题：**  LaVAN: Localized and Visible Adversarial Noise  **中文标题：**  LaVAN：局部可视的对抗性噪声  **发表于：**  ICML2018  **作者：**  Danny Karmon, Daniel Zoran  **单位：**  Microsoft，Deepmind | 大多数研究于基于深度学习的图像分类器的对抗性例子，使用的噪声，虽然很小，但覆盖了整个图像。  我们探索的情况下，噪声是允许是可见的，但限制在一个小的，局部的补丁的图像，而不覆盖图像中的任何主要对象。  我们表明，可以生成局部的对抗噪声只覆盖2%的像素图像，没有一个主要对象，可以在图像中迁移，可以在不同位置生效，并成功地欺骗了先进的Iceptionv3模型以非常高的成功率。 | 针对固定图片的固定位置生成对抗补丁，不考虑通用性：    损失函数上选择最大化target类（攻击目标类）的logits值，最小化source类（原本类别）的logits值，m是一个mask，贴补丁位置是1，不贴补丁位置是0，这一坨简单来说就是在图像的固定位置贴上补丁，使用mask可能是方便梯度回传（直接对进行梯度回传，的形状与图片是一样的，由于m的存在，只有m为1的位置有梯度，因此只有部分区域有梯度更新，以此将梯度更新限制在局部。）噪声初始化是全0  效果：  简单来说，Network domain优于Image Domain（废话），  分三个难度评测：1.攻击为target类且概率大于90%，2.攻击为target类即可，3.攻击为非source类即可。  缺陷：  没有任何通用性，补丁向任何方向移动一个像素就会导致图片被重新分类为source类别，在图像之间的迁移性更是完全没有。**但是有一点，当补丁连带着其四周的原始图片内容一起迁移到新图片上时，还保留了一定攻击效果。**  **一些缺陷的思考**  首先：作者没有考虑补丁攻击在不同网络之间的迁移性  其次：作者只考虑了补丁对于位置的鲁棒性，而没有其他的变换  其三：实验一中**补丁连带着其四周的原始图片内容一起迁移到新图片上时，还保留了一定攻击效果，这是一个值得思考的现象。**  **其四：补丁从0进行梯度累积，最后在语义上与target类很相近，这也是一个有意思的现象。**  **其五：Network Domain这种攻击，对图片的局部修改超出了像素的范畴，文中应该是直接将这样的数值送入神经网络，但是如果在物理世界中是不可能产生这样的补丁的（物理世界的图片当然是不能超出像素的范畴），因此如果将这种补丁归一化到像素范围内（这其实相当于Network Domain是累积梯度最后归一化，Image Domain是每次更新后都clip到像素范围内），是否会大幅度削弱攻击的成功率。** | 可以在图片和位置之间进行迁移的补丁的生成，即考虑通用性。  只有100张图片，每次抽取一张并且选择四个角中的某一个角打上补丁，然后对补丁朝着最大化target类的方向进行更新，注意整个过程只有一个补丁，这个补丁从全0开始叠加梯度回传，最后能够在不同图片不同位置之间生效。  1 测试补丁对位置的鲁棒性：    将补丁从图片左上移动到右下，记录了补丁在每个位置时，图片被分类为source类（中间图）和分类为target类（右边图）的概率。显然补丁在几乎所有位置都能将source类概率降为0，把target类概率抬高。  2 测试补丁对图片的鲁棒性：    首先，同一个补丁可以对不同的图片生效，其次在训练时没见过的图片上补丁也能生效。  3 类独立性：即有没有哪个target类容易被攻击到，有没有哪个source类不容易被攻击，作者尝试将各种source类攻击为各种target类，发现确实以有的类为target，攻击难以成功，有的source类难以被攻击为其他类。    此外作者发现，补丁显然跟target有很大程度上的视觉相似性，为了证明生成补丁的有效性作者还进行了两个实验：1.直接把一个Vulture的图片缩小作为补丁贴上去，发现并没有攻击成功，2.不使用图片，直接在白色的图片上贴上补丁，然后朝着Vuture这个方向对补丁进行梯度上升，然后将补丁放入图片中，发现同样攻击失败。说明补丁的生成确实不能独立于图片，也不是简简单单拿个Vulture缩小就能够充当补丁的。  **实验三**  作者试图将被攻击的图片，1. 通过对source类进行梯度上升，重新将图片分类为source类，2. 通过对target类进行梯度下降，将图片的分类结果改为target以外的其他类。统计这个过程中的对图片添加的梯度信息。    发现这两种情况进行梯度更新时，网络都没有对补丁进行过多的关注。  Adversarial patch中认为，补丁攻击成功的机理在于补丁吸引了网络的注意力，导致网络错误分类，本文认为，梯度信息并没有关注补丁位置，因此Ad patch中的观点是错误的 |
| **介绍** |
| 第一段：先介绍神经网络的脆弱性。（大环境）  第二段：说以前的工作，要么是全局添加细微扰动，要么是在主要位置添加局部扰动（比如针对人脸识别器，在人眼周围生成眼镜来干扰人脸识别器）这样生成的局部扰动遮盖了原有的特征信息。而本文工作在局部生成扰动，并且完全不遮盖图片中的主要内容（与其他工作对比）  第三段：撇清与adversarial patch这篇工作的关系，表达我们的工作是互补的，他关注物理世界，我更关注对先进的神经网络进行攻击，（其实感觉没说），并且我的扰动区域更小。（与最接近的工作对比）  第四段：开始介绍自己的方法：两个域（网络域与图片域），网络域不考虑补丁的像素值，直接对网络回传的梯度进行叠加，图片域每次在补丁上叠加回传的梯度后把补丁的像素值恢复到0-1之间，再进行下次梯度叠加。  第五段：介绍进一步的实验成果：用梯度回传验证网络对区域的关注程度，然后推翻adversarial patch认为网络更关注补丁区域这一观点（我不认可本文的观点）  第六段：局限性，虽然补丁能在不同图片与不同位置之间迁移，但是针对的是单一的白盒的网络（作者并没有进行网络之间迁移性的实验）。 |