|  |  |  |  |
| --- | --- | --- | --- |
| **文章信息** | **摘要** | **动机及方法概述** | **实验** |
| **阅读程度**：  精读  **文章标题：**  Shape Matters:Deformable Patch Attack  **中文标题：**  形状很重要：可变性对抗补丁  **发表于：**  ECCV2022  **作者：**  Zhaoyu Chen  **单位：**  Academy for Engineering and Technology, Fudan University  复旦大学工程技术学院 | 虽然深度神经网络（DNNs）在计算机视觉中表现出了出色的性能，但它们很容易受到精心设计的对抗性例子的影响，这些例子可能会误导DNNs产生不正确的输出。（对抗样本大环境，套话）。  补丁攻击是最具威胁性的形式之一，它有可能威胁到现实世界系统的安全。（引出补丁攻击的优势）  以前的工作总是假设补丁具有固定的形状，如圆形或矩形，并且不认为补丁的形状是补丁攻击的一个因素。（引出本文核心）。  为了探讨这一问题，我们提出了一种新的可变形补丁表示（DPR），它可以利用三角形的几何结构来支持轮廓建模和掩模之间的可微映射。（对理论进行抽象俗称讲故事）。  此外，我们还引入了一种联合优化算法，称为可变形对抗补丁（DAPatch），该算法允许同时有效地优化形状和纹理，以提高攻击性能。（实际的东西）  我们证明，即使面积很小，一个特定的形状也可以提高攻击性能。因此，DAPatch通过在各种网络架构上在GTSRB和ILSVRC2012数据集上改变补丁形状，实现了最先进的攻击性能，生成的补丁可能在现实世界中构成威胁。 | DPR技术，O（AP的中心点）给定，然后给**n个值**，这n个值代表的是射线的长度。  1：射线对360度进行等分，可以算出每个三角形的顶角。  2：找出每个三角型中的像素集合（三角形中有几个像素，这个集合中就有几组x,y坐标）。  3：计算射线的末端p的坐标。  4,5,6,7,8,9,10：计算OC与AB的交点D的坐标  11：通过函数获得C点对应位置的掩膜值（其实不用算也知道是接近0），但是通过这样的计算形成了一个从**n个值**到掩膜M的可微映射，掩膜M将被用于计算损失函数，那么就可以通过反向传递来优化这**n个值**，从而达到优化形状的目的**。**  损失函数：      Ladv就是交叉熵损失函数，初始化噪声，初始化掩膜，原始样本准备好后，使用公式生成初始化，然后送入神经网络后使用交叉熵损失函数最大化target目标的分类概率。这个过程中将同时优化和（代表长度的n个数）：    Lshape主要是控制补丁的面积：    显然最小化这个损失可以减小补丁的面积，当面积已经小于给定面积时，就不再优化这个损失了。 | 第一个实验：只用纯白补丁，即只优化补丁的形状，发现这种情况下经过形状优化的纯白补丁比一个纯白的方框与圆形补丁有更好的攻击效果。  第二个实验：作者根据某个论文把网络分为shape biased network和shape-texture debiased，结果发现攻击效果差别不大。（感觉是在凑工作量）  在数字世界：  第一个实验自然是比较三个网络，两个数据集上，6种不同方法（AP两种 LaVAN两种，PS-GAN和ours）的攻击效果。  第二个实验是在对抗训练网络上再比较。  第三个实验是在两种emprical defense与两种certifiable defense防御方法下比较。  在物理世界：  打印对抗补丁然后进行实验。  问题：  1：没有考虑位置，位置是定死的。  2：关于形状的建模我觉得可以进一步优化。  3：这个关于的优化过程多少有点不对。    对的他的优化多少有点问题。 |
| **介绍与相关工作** |
| 第一段：介绍对抗样本  第二段：介绍了已有的两个工作：AP与LaVAN（之前看过）。  第三段：说前面的工作都不关注形状，然后引了一些文献来表达神经网络除了纹理texture也关注形状shape，来引出自己的工作。  第四段：介绍本工作如何对补丁的形状进行建模，使得形状可以被优化。  第五段：再次强调DPR（可变性补丁表征，用于对补丁形状进行建模，使得形状可以被优化）和DAP（可变性对抗补丁）。  **贡献：**   1. 引入DPR构建了形状参数与掩膜之间的可微映射，这样可以通过神经网络学习到合适的形状参数。 2. 基于DPR，使用损失函数同时优化shape和texture，来生成DAP（可变性对抗补丁）。 3. 在两个数据集，CNN与ViT两个网络结构上以及数字与物理世界中进行实验，证明对形状的关注确实有助于提高攻击能力。 4. 再次强调DAP首次引入形状信息，从一个新的角度去衡量网络的鲁棒性。   相关工作介绍了三点：   1. 对抗补丁的其他工作，包括AP与LAVAN已经其他使用GAN生成补丁的工作，并表明他们都没有关注形状。 2. 介绍了关于补丁攻击的防御，从emprical防御角度有几种预处理的手段进行防御，第二种是对抗训练，还有几种certifiable防御方法。 3. 介绍了神经网络对texture与shape的感知的一些工作。 |