|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  方法精读  **文章标题：**  Spatially transformed adversarial examples  **中文标题：**  空间转换对抗样本  **发表于：**  ICLR2018  **作者：**  Chaowei Xiao（小超威？？）  **单位：**  University of Michigan, Ann Arbor, USA | 生成对抗样本是使用双线性差值生成的，这个过程中有一个参数叫流量场。  流量场的形状与图片相同，为（h, w），表示第i行第j列像素点坐标的偏移量，其中真实像素点的坐标是1 2 3 这种整数，而是0~1之间的小数，真实像素点（i,j）经过偏移量得到虚拟像素点，这个虚拟像素点是个小数。下面介绍给定原始图像x与流量场f后，如何得到xadv:  记是原始图像x，坐标为k=(i,j)的像素点的像素值，是坐标为k1=（i+1，j+1）k2=（i+1，j-1）k3=（i-1，j+1）k4=（i-1，j-1）即右上，左上，右下，左下像素点的像素值，通过（i，j）处的偏移量得到虚拟像素点的坐标为。是对抗样本坐标为（i,j）的像素点的像素值，其值为：，其中是使用虚拟坐标k’与像素点的坐标计算而来，距离度量公式是。总结来说就是以原图像的一个像素点周围的四个像素点的值value，以虚拟坐标与四个像素点的坐标之间的距离度量为权重得到对抗样本的一个像素值。  上述过程中可优化的对象就是流量场f，因为f给出了虚拟坐标相对于真实坐标的偏移量，f确定之后，根据原图x生成的对抗样本也确定了，因此我们要找到一个f，满足两个条件，1.根据f生成的样本可以引导神经网络产生误判 2.根据f生成的对抗样本与原图应该具有更好的视觉上的一致性。因此设置了损失函数：    第一项就是来优化第一个条件：用的是cw的损失函数，保证对抗样本攻击成功：  第二项就是来优化第二个条件：，第一层是要对图像上的所有像素点进行遍历，第二层是对一个像素点的邻域内的像素点进行遍历，优化的目标是，像素点及其邻域内的所有像素点的偏移量的欧式距离较小，换言之，希望一个像素点及其邻域内的像素点的偏移向量相互平行，，类似于这样一个效果，从直观角度来讲，这样的一个流量场施加在原图x上得到的对抗样本xadv在视觉上与x会有较高的相似度。  （与基于lp范数的方法的区别，lp范数的方法是优化一个，的形状也是(h,w)，最后通过得到对抗样本，等于直接优化一个像素值的增量添加到原图的像素值上，本文提出的是基于空间变换的方法优化一个，对抗样本某点的像素值是**以原图该点周围的四个像素点的像素值值为value**，**以通过生成的虚拟坐标与这四个像素点的坐标之间的距离度量为权重**，加权平均得到的。Lp的方法优化的是一个像素值增量，本文的方法优化的是四个像素值的权重） |
| **摘要** | **实验** |
| 最近的研究表明，广泛使用的深度神经网络（DNNs）很容易受到精心设计的对抗性例子的影响。（大背景）  许多先进的算法被提出通过利用Lp距离来惩罚扰动来产生对抗的例子。研究人员探索了不同的防御方法来抵御这种对抗性攻击。虽然Lp距离作为感知度量的有效性仍然是一个活跃的研究领域，但在本文中，我们将专注于一种不同类型的扰动，即空间转换，而不是像之前的工作那样直接操纵像素值。（本文与前者不同）  通过空间转换产生的扰动可能导致大的Lp距离测量，但我们进行的广泛实验表明，这种空间转换的对抗性例子在感知上是现实的，而且更难用现有的防御系统来防御。这可能为对抗性实例的生成和相应防御的设计提供了一个新的方向。（简述本文工作的优缺点和意义）  我们将不同的例子的基于空间变换的扰动可视化，并表明我们的技术可以产生真实的具有光滑的图像变形的对抗性例子。最后，我们用不同类型的对抗性例子来可视化深度网络的注意力，以更好地理解如何解释这些例子。（一些实验） | 略，18年的实验太简单了，拴条狗都能做 |