|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  粗读  **文章标题：**  On the Robustness of Deep K-Nearest Neighbors  **中文标题：**  关于深度KNN算法的鲁棒性研究  **发表于：**  2019 IEEE Security and Privacy Workshops (SPW)  **作者：**  Chawin Sitawarin, David Wagner  **单位：**  EECS Department, UC Berkeley | 没细看，就说说这篇论文看完我了解到了什么东西叭。   1. 有个哥们把KNN算法推广为了D-KNN算法，大概就是KNN不用训练直接比对原始样本之间的欧氏距离或余弦相似度，D-KNN算法会对原始样本提取特征，然后比对每层之间的的余弦相似度。 2. 有几种很nt的针对KNN的对抗样本生成方法，比如mean attack就是找到离被攻击样本x最近的被knn分类为另一个类别的样本x’，然后利用类似于mixup的方法，(1-m)\*x + m\*x’将m从0开始增加到1，找最小的m使得分类错误获得对抗样本。（我觉得可以更改为梯度方法：对于knn聚类的moco模型，要攻击自然样本x，先找到离它最近的与x类别不同的样本x'，注意我们并不关心对x'的分类结果是否正确，只要模型对x'的预测结果与x不同即可，然后以x'的特征向量为目标，通过梯度信息对x进行加噪生成对康样本，不断拉近对抗样本的特征向量与x'的特征向量之间的距离，来引导对抗样本被错分） 3. 作者提出了基于梯度的一种对抗样本生成方法，拉开对康样本与相似样本的距离，拉近对康样本与其他类别样本的距离。 4. 实验在mnist数据集上进行。 |
| **摘要** |  |
| 尽管人们对敌对的例子进行了大量的关注，但很少有作品证明它能有效地防御这种威胁。  我们研究了深度k-最近邻（DkNN），这是一种结合了k-最近邻（kNN）和深度学习的防御方法，以提高模型对对抗性例子的鲁棒性。  由于对具有大k值和高维数据的kNN分类算法缺乏有效的攻击，因此评估该方案的鲁棒性具有挑战性。（缺乏有效的攻击）  我们提出了一种启发式攻击，允许我们使用梯度下降来寻找kNN分类器的对抗性例子，然后将其应用于攻击DkNN防御。  结果表明，我们的攻击比任何对kNN的幼稚攻击都要强，并且明显优于对DkNN的其他攻击。 |  |