附录

1. ID模型训练算法伪码

|  |
| --- |
| **算法1 ID模型训练算法** |
| **输入：**目标模型，制作对抗样本的模型，ID模型，训练轮数*T*，学习率，为增强型干净样本数据集。  **输出：**收敛的ID模型。 |
| 1. **begin** 2. //初始化抗样本数据集为空 3. **for** each in **do** //采用*FGSM*算法（或采用其它攻击算法）制作对抗样本数据集 4. //计算对抗扰动，为损失函数,，为扰动值 5. //生成原始样本的对抗样本及类标签 7. **end for** 8. **for** t = 1 to **do //** 进行轮训练 9. **for** each in **do //** 10. //，采用反向传播算法更新模型参数 11. **end for** 12. **end for** 13. **return** 14. **end** |

B. 定理1及证明：

**定理1**：对于第类干净/对抗样本隐层向量集合/，其中的任意两点的线性组合均在凸包中。

**证明：** 定理1等价为：给定凸包，对于任意正整数，任意，任意给出非负实数且，恒有。

由数学归纳法证明：

(1) 当，由凸集的定义[22]可知结论成立。

(2) 假设对个点时结论成立，证明个点时结论成立。

，

则：

* 1. 当 时，即时：
  2. 当时，

由上述假设(2)可知： ，

又∵，且系数满足：，。

由凸集的基本定义可知：时结论成立，即：

C. HIR模型训练算法伪码

|  |
| --- |
| **算法**2 **HIR模型训练算法** |
| **输入**：目标模型，目标模型的卷积层和分类层，已训练的ID模型，未训练的HIR模型，训练轮数*T*，原始数据集，对抗样本数据集，学习率  **输出**：已训练好的HIR模型 |
| 1. **begin** 2. **for** t = 1 to *T* **do** 3. **for** each class samples in **do** 4. //取的值为0到1之间的随机数 5. //，计算第类干净样本去噪后的隐层向量 6. //，计算第类对抗样本去噪后的隐层向量 7. // *Mixup*取出第类凸包内的隐层向量 8. //前向传播计算损失函数值 9. //利用反向传播计算参数的梯度 10. **end for** 11. **end for** 12. **return** 13. **end** |

D. ID+HIR算法在CIFAR100数据集上的实验结果及分析

表 4 ID+HIR算法在CIFAR100数据集上的对抗鲁棒性对比结果表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **样本**  **防御算法** | **干净样本** | | **FGSM ()** | | **CW ()** | | **BIM ()** | |
| VGG16 | ResNet18 | VGG16 | ResNet18 | VGG16 | ResNet18 | VGG16 | ResNet18 |
| **无防御** | 72.6 | 76.4 | 52.1/45.3/39.6 | 51.3/42.4/37.9 | 63.2/61.4/54.8 | 62.0/59.3/51.3 | 53.4/50.1/29.3 | 52.6/48.4/28.2 |
| NRP[16] (2020) | 65.2 | 65.4 | 65.4/62.1/61.2 | 69.2/68.5/68.3 | 68.1/68.7/66.3 | 71.8/70.9/70.3.8 | 69.3/67.7/59.3 | 72.1/70.7/66.4.3 |
| Fast AT[28] (2020) | 65.8 | 70.1 | 70.2/62.3/55.1 | 74.1/70.2/65.2 | 72.1/70.9/69.2 | 74.2/74.1/73.6.3.4 | **72.0/70.3/68.2** | 75.5/**73.7**/**72.9** |
| ComDefend[10] (2019) | 68.7 | 71.4 | 66.2/61.4/60.2 | 70.2/69.8/67.2 | 70.0/69.7/65.2 | 73.8/71.2/69.1 | 68.2/66.3/60.1 | 73.2/71.7/67.5 |
| FS[14] (2018) | 71.3 | 74.1 | 67.2/57.1/54.2 | 54.7/44.5/40.1 | 70.2/70.1/70.5 | 70.4/67.2/65.6.7 | 68.3/62.5/57.3 | 71.3/69.5/64.1.9 |
| JPEG[15] (q=25) (2018) | 55.4 | 55.7 | 47.2/42.3/39.1 | 49.2/41.3/39.1 | 57.2/57.1/56.80.5 | 56.1/54.1/50.4 | 57.7/55.4/50.0 | 58.2/56.2/51.2.4 |
| JPEG[15] (q=50) | 58.8 | 60.5 | 49.2/43.7/38.29.1 | 51.5/44.7/42.7 | 64.2/63.4/62.7 | 65.2/62.6/62.1.4 | 66.0/61.6/58.2 | 65.1/62.4/59.7.3 |
| JPEG[15] (q=75) | 68.2 | 71.8 | 58.9/47.2/45.2 | 55.2/47.5/45.4 | 69.3/68.9/67.2 | 68.3/67.3/63.5 | 67.3/58.2/45.5 | 63.3/59.2/47.5 |
| TVM[15] (2018) | 69.1 | 72.3 | 61.7/51.2/49.7 | 57.1/49.3/51.6 | 70.1/68.2/67.7 | 69.8/68.2/63.1 | 67.9/61.3/50.2 | 64.2/60.8/51.7 |
| **ID+HIR (Our)** | **71.8** | **74.8** | **71.2/68.3/63.7** | **75.5**/**72.3**/**68.7** | **72.2/71.2/70.6** | **74.9**/**74.7**/**74.2** | 71.5/69.8/64.5 | **75.5**/73.1/69.3.1 |

表 4的实验结果表明：在CIFAR10数据集上，本文提出的IHD算法在VGG16模型与ResNet18模型上展现了近似的对抗鲁棒性。从上述实验结果得出以下结论：

1）网络模型在不采取任何防御手段的场景下，对ResNet56模型上制作出的对抗样本有着较差的鲁棒性，在使用IHD算法作为防御算法时，VGG16模型和ResNet18模型均对对抗样本有着较高的分类精度；

2）IHD算法对原始样本的识别精度为91.2%，虽然相比无防御场景下模型对干净样本92.6%的分类精度有所降低，但是相比其他传统的对抗样本去噪类算法，如ComDefend算法对干净样本的分类精度为85.2%，IHD均有着较大的提升，且IHD对对抗样本的分类精度在各种对抗样本上也远高于ComDefend等去噪算法；

3）在ResNet18模型上，IHD算法对原始样本识别精度略低于FS（Feature Squeezing）算法，但是FS算法对FGSM、CW和BIM等对抗样本展现出的对抗鲁棒性均大幅度低于IHD算法。JPEG-Compression、TVM等算法同样在干净样本和对抗样本上的分类精度低于IHD。本实验选择的对抗训练类算法Fast AT在面对BIM对抗样本时略优于IHD，其原因在于BIM算法是一种和PGD类似的迭代攻击的算法，对模型具有更强的干扰性，对抗训练算法在训练模型时直接更新目标模型的参数，具有更好的适应能力，但是在面对FGSM和CW对抗样本时，IHD算法具有更好的性能。

E. ID+HIR的消融实验及分析

训练数据为扰动值分别采用4、8、16的FGSM和CW对抗样本。目标模型选择ResNet56模型，该模型对CIFAR10干净样本的识别精度为92.6%，对SVHN的识别精度为96.5%。针对MNIST数据集选择具有4层卷积层的卷积神经网络，识别精度为99.2%。目标模型对对抗样本的识别精度以及两种防御场景下的实验结果如表 3所示。

表 3 HIR模型消融实验结果表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **样本 防御算法** | **FGSM ()** | | | **CW ()** | | |
| CIFAR10 | SVHN | MNIST | CIFAR10 | SVHN | MNIST |
| **无防御** | 52.7/44.0/34.0 | 52.1/40.0/27.4 | -/-/41.2 | 17.1/14.2/9.0 | 37.1/17.2/9.5 | -/-/74.3 |
| **ID** | 86.7/90.5/92.2 | 92.6/95.3/96.1 | -/-/98.9 | 81.8/82.1/73.2 | 95.8/93.5/94.6 | -/-/98.0 |
| **ID+HIR** | **88.3/91.5/92.2** | **93.5/96.2/96.4** | -/-/**99.1** | **85.0/85.2/90.0** | **96.2/94.8/95.7** | -/-/**98.7** |

注：由于MNIST数据集为小尺度黑白数据，当扰动值较低时，“-”表示无明显的攻击效果。

表3实验结果表明：

1）在CIFAR10数据集上，在仅使用ID模型做防御时，针对各扰动值的对抗样本训练ID模型后，均能大幅提升模型对对抗样本的识别精度，说明ID模型有效地去除了对抗样本上的扰动噪声；

2）在CW（）作为训练数据时，可以看到本文算法下目标模型对对抗样本识别精度为90.0%，而ID为73.2%，证明了HIR模型在隐层的Mixup凸包训练方式同时可以有效地将对抗样本的隐层向量拉回到分类器所能正确识别的空间中；

3）在SVHN数据集上的实验，有着和CIFAR10数据集相似的特点，但是针对该数据集训练得到的模型在对抗样本上均有着比CIFAR10更高的精度。通过MNIST数据集上的实验可以发现，在本文ID+HIR算法防御下，目标模型对对抗样本的识别精度维持在98.7%，与之前的99.2%相差无几，证明HIR防御模型几乎不影响对抗鲁棒性。

总之，通过在CIFAR10、SVHN和MNIST等数据集上对HIR模型的性能测试，证明了本文提出的基于凸包训练的隐层恢复模型HIR可以有效地将对抗样本和干净样本的隐层向量拉回到分类器所能正确识别的空间中。ID去噪器在提升模型的对抗鲁棒性的同时，HIR能够维持深度模型对干净样本的分类精度，很好地解决了去噪算法导致干净样本信息损失的缺陷，同时也说明了本文基于凸包理论构建的HIR模型的训练数据及算法的正确性与有效性。