击神经网络,相对于黑盒攻击而言具有更强的攻击能力.因此,使用白盒攻击来验证提高模型鲁棒性带来的防御能力具有更大的说服力.本文选择了 7 种流行的白盒攻击方式:快速梯度符号法(fast gradient sign method,简称 FGSM)[22]、基本迭代法(basic iterative method,简称 BIM)[23]、投影梯度下降(project gradient descent,简称 PGD)[24]、动量迭代法(momentum iterative method,简称 MIM)[25]、基于雅可比矩阵的显著图攻击(Jacobian-based saliency map attack,简称 JSMA)[26]、Carlini & Wagner(C&W)[27]、弹性网络攻击(elastic-net attack,简称 EAD)[28].此外,为了观察模型在不同攻击力度下防御表现的变化,每种攻击方式都设置了 3 种参数.

接下来选择实验使用的数据集.我们选取了图像分类领域两个最常用的数据集 MNIST 和 CIFAR-10.MNIST 是一个黑白手写数字数据集,包含 0~9 这 10 类来自 250 人的手写数字,图片尺寸为 28×28.其中,训练集图片数量为 60 000 张,测试集图片数量为 10 000 张;CIFAR-10 是一个更接近于普适物体的彩色图像数据集,包含飞机、汽车、鸟类、猫、鹿、狗、蛙类、马、船和卡车这 10 类数据,图片尺寸为 32×32.其中,训练集图片数量为 50 000 张,测试集图片数量为 10 000 张.

本文认为:使用这两种数据集,能成功验证本方法的可行性和有效性.

为了回答问题 1 和问题 3,我们设计了 2(数据集)×7(攻击方式)×3(攻击参数)×4(模型)=168 组对比实验,在两个数据集上分别使用上述 7 种攻击方式,对每种攻击方式设置 3 种攻击参数调节攻击力度,对比上述 4 种模型的分类正确率.

为了回答问题 2 和问题 3,我们设计了 2(数据集)×4(模型)=8 组对比实验,在两个干净的数据集上对比上述 4 种模型的分类准确率.同时,使用 T-SNE 视图[29]对各个模型的输出降维可视化,进一步佐证了本方法的可用性.

为了回答问题 4,我们设计了 2(数据集)×4(攻击方式)×3(攻击参数)×2(模型)=48 组对比实验,在两个数据集上分别使用 FSGM,BIM,MIM,PGD 这 4 种攻击方式.之所以使用这 4 种方式,是因为它们具有相似的攻击原理,同样,对每种攻击方式设置 3 种攻击参数,对比 F+D 模型和 AdvT+F+D 模型的防御表现.

3.2 实验过程

本次实验所使用的 CPU 型号为 Intel i7 9700k,使用的图形处理器型号为 Nvidia RTX 2080Ti,操作系统为 Linux 18.04,Python 版本为 3.7,机器学习平台为 Tensorflow v1.12[30]以及 Keras v2.4.

在训练模型之前,我们首先对数据集进行了归一化预处理:将所有的训练样本都归一化到 0-1 范围内.同时,为达到更好的训练效果并降低训练出来的模型的过拟合程度,在训练过程中也使用了数据增广技术,对原始图像样本分别进行水平翻转、水平平移和竖直平移操作.在平移过程中,使用常量 0 填充超出边界的部分.

实验选择的 B 模型为深度残差网络 Resnet-32,它的基本结构如图 2(a)所示,它包含 3 组通道数依次为 16,32,64 的残差块:第 1 组残差块由 5 个恒等残差块构成,第 2 组、第 3 组残差块均由 1 个卷积残差块和 4 个恒等残差块构成.恒等残差块和卷积残差块的差别在于块中残差分支是否做卷积操作,所以卷积残差块会改变特征图的大小,从而满足每个阶段特征图尺寸缩小的需求.这 3 个残差块对应的特征图尺寸分别为 32,16,8.最后,通过一个全连接层输出 10 分类结果.

F 模型的结构如图 2(b)所示,它是在 B 模型的基础上,在第 1 组和第 2 组残差块之后分别加入额外分支,算上原出口,改动后的模型存在 3 个出口.在模型中加入特征金字塔,实现特征融合的过程分为 3 步.

- 第 1 步在第 3 组残差块的特征图做上采样后,达到与第 2 组残差块的特征图同一尺寸,再采用横向连接的方法与第 2 组的特征图融合;
- 第 2 步对第 1 步得到的融合后特征图再做上采样到与第 1 组残差块的特征图同一尺寸,再次进行特征融合,并在融合过程中使用小卷积,以保证 3 部分特征图的通道数一致;
- 最后对 3 个分支都进行小尺寸卷积核的特征压缩,统一生成 8×8 尺寸 128 通道的特征图,然后对其进行全局平均池化,生成 1×128 的向量,分别经过全连接层得到最后的 10 分类预测(包含 softmax 变换),最终输出结果是 3 个预测结果的平均值.

D 模型的结构如图 2(c)所示,它同样在 B 模型的第 1 组和第 2 组残差块之后加入额外分支,与 F 模型不同的是:它并不会对各个出口输出的特征图做特征融合,而是直接通过全连接层得到一个 10 分类预测结果.为了

削减共享层的影响,使训练出来的各个分支的预测结果不过于趋同,我们添加了整体多样性计算:使用了一个带权重的组合交叉熵来保证了各分支的预测准确率,根据出口深度,由浅至深权重依次设置为 1~3;最后,使用第 3.2 节提出的单模型内多分支预测的整体多样性 L_{ED} 的计算公式为输出结果添加了整体多样性计算,公式中,超参数γ和μ的值分别设为 1 和 0.01.

F+D 模型的结构如图 2(d)所示,它综合了 F 模型与 D 模型的改造方法,对各个出口输出的特征图做特征融合之后再输入到全连接层,得到一个 10 分类预测结果;然后,对 3 个出口的预测结果进行整体多样性计算,参数设置与 D 模型相同.AdvT+F+D 模型的结构与 F+D 模型一致,它与 F+D 模型的区别只在于训练过程中动态地生成对抗样本用作数据扩容,设置对抗样本和正常样本的比例为 1:1,计算模型参数梯度的误差由对抗样本误差和正常样本误差累加获得.

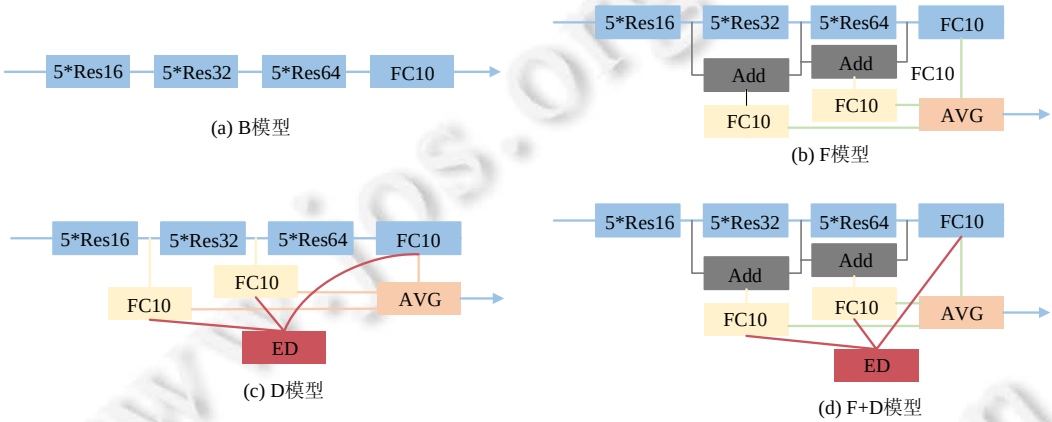


Fig.2 Four models used in the experience

图 2 实验中使用的 4 种模型

在上述模型的训练过程中,我们将初始学习率α设置为 0.001.指数衰减率β_1,β_2 分别控制之前的时间步的梯度动量和梯度平方动量的影响情况.为了使对比结果更公平,将它们设置为领域内默认值.其中,β_1 设置为 0.9,β_2 设置为 0.999.在 MNIST 训练集上的训练轮数为 40,但会在 20 轮之后,将学习率降低为初始的十分之一,直至降到 $10e^{-4}$ 数量级;在 CIFAR-10 数据集上训练轮数为 180,分别在 80,120,160 轮之后降低学习率为前一时刻的十分之一,直至降到 $10e^{-6}$ 数量级.批量(batchsize)设置为 128.此外,由于在训练过程中损失函数表示的误差不可能归为 0,若出现,则意味着模型过拟合了,所以我们使用了基于验证集准确率的模型保存机制.

实验中,对抗样本生成使用框架为 cleverhans v2.1.0[31],cleverhans 是谷歌基于 Tensorflow 开发的,集成了大多数现有对抗样本生成方法.攻击过程为针对要攻击的模型图结构,cleverhans 按照所需攻击方法和设定的攻击力度,生成对应的数据流图,输入原始样本后,输出生成对抗样本.防御实验中,CIFAR-10 数据集上 FGSM 扰动力度为 0.01~0.04;BIM 等 3 种迭代方法扰动力度设计为 0.01~0.03;JSMA 设计扰动力度为 0.1,而攻击像素点占比为 5%~15%;C&W 攻击力度为 0.001,0.01 和 0.1;EAD 下,$L1$ 正则的超参为 0.01,攻击力度为 0.1,1,5.MNIST 数据集上,FGSM 扰动力度为 0.1~0.3;BIM 等 3 种迭代方法扰动力度设计为 0.05~0.15;JSMA 设计扰动力度为 0.2,而攻击像素点占比为 10%~40%;C&W 攻击力度为 0.1,1 和 5;EAD 下,$L1$ 正则的超参为 0.01,攻击力度为 1,5,10.

3.3 实验结果分析

3.3.1 对于对抗样本的处理能力

为回答问题 1 和问题 3,首先在 CIFAR-10 数据集和 MNIST 数据集上测试 B 模型、F 模型、D 模型以及 F+D 模型对于 7 种常用对抗样本生成方法下白盒攻击的防御效果.除每种攻击方法的扰动参数设定之外,设置 BIM,MIM 和 PGD 这 3 种攻击的迭代次数为 10,设置 C&W 和 EAD 攻击的迭代次数为 1 000,且学习率为 0.01.实验结果记录在表 1 中.

Table 1 Comparison of classification accuracy against adversarial examples on CIFAR-10 and MNIST (%)

表 1 CIFAR-10 和 MNIST 数据集上对于对抗样本分类正确率比较 (%)

数据集	攻击类型	攻击参数	B 模型	F 模型	D 模型	F+D 模型
CIFAR-10	FGSM	ε=0.01	20.62	33.28	46.51	**64.32**
		ε=0.02	13.64	24.21	32.52	**60.28**
		ε=0.04	9.84	19.25	18.12	**49.44**
	BIM	ε=0.01	6.5	9.35	18.79	**42.26**
		ε=0.02	5.76	5.74	10.46	**32.42**
		ε=0.03	5.75	5.48	8.95	**27.58**
	MIM	ε=0.01	7.5	10.98	24.74	**47.82**
		ε=0.02	5.8	5.84	11.47	**38.07**
		ε=0.03	5.76	5.52	8.93	**32.33**
	PGD	ε=0.01	7.78	12.01	22.82	**43.96**
		ε=0.02	5.37	5.8	10.43	**31.57**
		ε=0.03	4.84	5.01	7.47	**24.14**
	JSMA	θ=0.1, γ=0.05	11.1	18	38.6	**45.3**
		θ=0.1, γ=0.1	3.1	7.8	17.2	**32.4**
		θ=0.1, γ=0.15	2.3	7.1	8.6	**26.4**
	C&W	c=0.001	38.2	28.3	**69.05**	66.3
		c=0.01	5.75	5.6	**48.75**	47.9
		c=0.1	5.5	5.4	23.1	**30.8**
	EAD	β=0.01, c=0.1	73.7	36.9	88	**89.9**
		β=0.01, c=1	5.55	4.75	61.8	**69.2**
		β=0.01, c=5	2.3	2.75	15.65	**37.15**
MNIST	FGSM	ε=0.1	49.69	71.61	26.11	**94.84**
		ε=0.2	13.21	20.68	11.03	**65.46**
		ε=0.3	5.42	11.65	9.77	**20.68**
	BIM	ε=0.05	91.4	95.42	80.18	**95.76**
		ε=0.1	21.99	54.6	15.67	**87.8**
		ε=0.15	1.28	10.88	7.47	**72.84**
	MIM	ε=0.05	92.46	95.7	84.41	**96.38**
		ε=0.1	32.84	62.83	17.68	**90.89**
		ε=0.15	4.3	17.84	9.41	**79.84**
	PGD	ε=0.05	91.74	96.31	69.02	**95.98**
		ε=0.1	7.31	51.02	7.17	**75.89**
		ε=0.15	0.18	8.04	1.82	**38.61**
	JSMA	θ=0.2, γ=0.1	71.4	78.2	54.2	**86.8**
		θ=0.2, γ=0.2	30.6	52.4	32.4	**62.7**
		θ=0.2, γ=0.4	15.6	28.8	16.2	**30.6**
	C&W	c=0.1	60.9	89.4	93.1	**97.4**
		c=1	0.55	2.8	30.05	**87.55**
		c=5	0.55	0.8	3.25	**38.35**
	EAD	β=0.01, c=1	77.1	82.65	98.3	**98.8**
		β=0.01, c=5	0.65	6.45	69.35	**95.35**
		β=0.01, c=10	0.55	2.4	36.55	**93.5**

表 1 的第 1 部分为在 CIFAR-10 数据集上的实验结果.

- B 模型在这 7 种攻击方式下的分类表现都受到了很大的影响,在扰动较低的情况下,准确率大幅下降;扰动较高的情况下,准确率甚至只有个位数水平;
- F 模型仅在 FGSM 和 JSMA 这两种攻击方式下,分类准确率略有提升;但是对于其他的攻击方式,防御效果并不明显.在 C&W 和 EAD 这两种比较相似的攻击方式下,准确率下降的跨度甚至超过了 B 模型;
- D 模型相比于前两种模型而言,对所有类型对抗样本都有提高.在 7 种攻击方式下,准确率都达到了 B 模型的两倍以上.其中:对 JSMA 和 C&W 攻击的防御表现提升了 3~4 倍,对高扰动的 EAD 攻击防御效果甚至达到了 B 模型的 7 倍左右;
- 最后一列记录了应用本文提出的单模型鲁棒性提高方法后形成的 F+D 模型的防御结果.在前 4 种攻击方法下,该模型准确率相比于 D 模型都有成倍的提高;同时,3 种不同扰动值间的下降幅度也远小于 D 模型;后 3 种攻击方式下的防御表现完美继承了 D 模型的优势,JSMA 和 EAD 攻击下,面对各种扰动值都进一步提高,面对 EAD 高扰动攻击的分类准确率更是达到了 B 模型的 15 倍以上;对 C&W 两种小扰

动的防御表现略差于 D 模型,但是实验中的最高扰动下出现了反超,表明 F+D 模型在扰动提高时防御表现的下降趋势比较平缓.

表 1 中,第 2 部分为 MNIST 数据集上实验结果.在进行此部分实验时,由于 MNIST 数据集的样本结构比较简单,所以各种攻击方式的扰动范围较于 CIFAR-10 数据集上有大幅提高,而迭代次数和 C&W 的学习率和前部分实验相同.在如此高的扰动范围下.

- B 模型对对抗样本的分类准确率都大幅下降,特别在实验中设定的第 1 种、第 2 种扰动值间,出现了 3~4 倍的极速下滑;甚至在 C&W 和 EAD 这两种有目标的攻击方式下,受较高扰动攻击后的分类准确率下降到了 1 以下;
- F 模型的防御表现在前 5 种攻击方式下都有较大提升,特别是面对扰动幅度变大情况,下降趋势相对平缓;而对 C&W 和 EAD 两种相似攻击的防御表现只有略微提高;
- D 模型对 FGSM 等前 5 种对抗样本的防御表现虽然略高于 B 模型,却比 F 模型又有下降;而在 C&W 和 EAD 两种攻击下表现良好,抑制住了不同幅度间快速下降的趋势;
- 而本文方法得到的 F+D 模型在所有攻击方式下的防御表现都得到了非常大的提升,多种攻击下的表现提高了超过 60 个点;甚至在 C&W 和 EAD 这两种 B 模型表现极差的情况下,几乎保持了对原始样本的分类准确率,不同扰动幅度间的下降趋势也更加平缓.

观察整张表发现:面对 C&W 和 EAD 攻击,整体多样性可以提供更好的防御效果;而 FGSM 等前 5 种攻击方法则会受到样本复杂程度的影响.在相对复杂的 CIFAR-10 数据集上,整体多样性带来的提升高于特征融合;而在相对简单的 MNIST 数据集上,特征融合会提供更好的帮助.本文方法结合特征融合和整体多样性,最终达到了 1+1 大于 2 的优秀表现.在整个实验中,一直保持较优的防御表现,完美解答了问题 3.在两种测试集下进行对 4 种模型的白盒攻击防御实验中,回答了问题 1,证明本文提出的方法可以大幅提升模型的鲁棒性,对常见对抗样本攻击方法均可做出有效防御.

3.3.2 对于干净样本的处理能力

为回答问题 2,实验评估了 B 模型、F 模型、D 模型以及 F+D 模型在干净测试集上面的表现,实验结果记录在表 2 中.其中,

- B 模型在 MNIST 和 CIFAR-10 的识别率分别为 99.59%和 91.17%,达到了现有的深度神经网络分类器的基本水平;
- F 模型在两个数据集上的分类准确率则是 99.65%和 91.41%,可以看出:特征融合使得最终分类可以同时考虑语义特征和细节特征,模型能达到更好的精度;
- D 模型没有加入特征融合,考虑了整体多样性,但由于是在一个模型内,各分支路线有共享层,可以预见,分类准确率会受到一定的影响,最终结果 99.53%和 89.14%,也符合预期;
- F+D 模型作为本文方法改进并训练的模型,在 CIFAR-10 数据集上的分类准确率达到了 91.05%,比 D 模型表现优秀,较于 B 模型也基本没有准确率的下降,甚至在 MNIST 数据集上的表现超过了 F 模型,达到了 99.7%.上述结果表明,本文方法改进并训练的模型仍能保证对原始样本的分类精度.

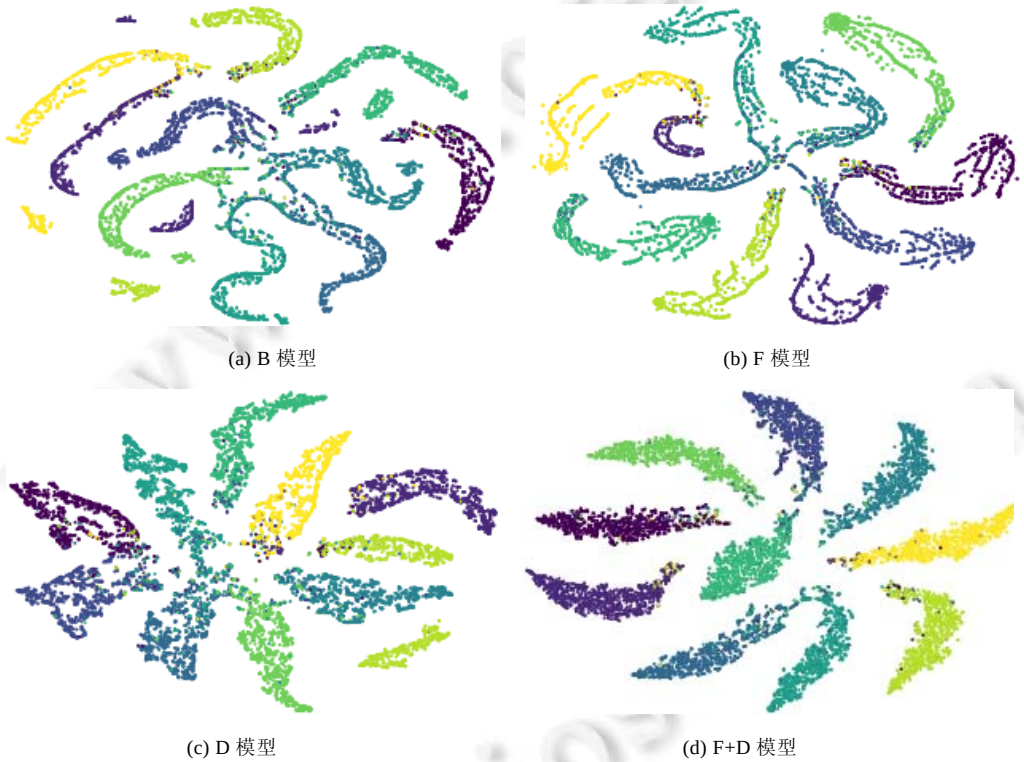
Table 2 Comparison of classification accuracy on clean examples from MNIST and CIFAR-10 (%)

表 2 MNIST 和 CIFAR-10 上干净样本分类准确率比较 (%)

数据集	B 模型	F 模型	D 模型	F+D 模型
CIFAR-10	91.17	**91.41**	89.14	91.05
MNIST	99.59	99.65	99.53	**99.7**

为更好地展现整体多样性的效果,实验打印了各模型最终输出的 T-SNE 视图.T-SNE 利用条件概率表示相似性,使用相对熵训练,可将高维分布的点映射到低维空间中,明确地显示出输入的聚类状况.图 3 绘制了对比的 4 种模型最终输出在低维的分布情况.图中每一种颜色代表一种分类,此实验在 CIFAR-10 验证集(10 000 张)上进行,所以颜色数目为 10.从图中可以看出:在同类的点中间参杂其他颜色的点,表示这些点是分类出错的部分.

因为本实验输入是原始样本,可以看出,图中的错误点仅是少数.图 3(a)是 B 模型输出的 T-SNE 视图,可以看出:每种分类并没有完全聚集在一起,会有部分零散分布在其他位置,总体分布得杂乱无章,甚至有几类出现了交融.从这里可以认为 B 模型对对抗样本很敏感,符合前一实验的结果.图 3(b)是 F 模型输出的视图,可以看出:在同时利用了语义特征和细节特征之后,各类自己的聚集相对基本模型已经有了改善,但是依旧存在不同类交叉的问题,分离程度不足.图 3(c)是 D 模型输出的视图,可以看出:各分类之间已经有了分离的趋势,但错误的点也明显增加.推测是由于在一个模型内,因为各分支存在共同层,多样性的加入可能影响了拟合,所以准确率有所下降.图 3(d)则是本文方法改进并训练的模型,同时引入了特征金字塔和改进的整体多样性,相比于前几种的结果,同类之间的聚集程度得到了极大的提高;且不同类之间有明显的分离,错误的点基本没有增加,符合表 2 中的实验结果.这部分同样证明了本文方法成功地提高了模型的鲁棒性,并且基本没有影响模型对原始样本精度.



(a) B 模型 (b) F 模型

(c) D 模型 (d) F+D 模型

Fig.3 T-sne views of the final output from each model on CIFAR-10

图 3 CIFAR-10 测试集上各模型最终输出的 T-SNE 视图

3.3.3 与对抗训练方法组合使用的效果

为回答问题 4,测试了 F+D 模型以及额外加入了对抗训练后的 AdvT+F+D 模型在受到相同攻击下的分类正确率,结果记录在表 3 中.CIFAR-10 和 MNIST 上的对抗模型训练都使用 PGD 方法,对抗训练过程中,CIFAR-10 数据集下设置 PGD 的扰动值为 0.01~0.05 随机采样,MNIST 数据集下设置 0.05~0.2 随机采样.随后测试了与前部分相同参数的 FGSM,BIM,MIM 和 PGD 这 4 种攻击.实验结果表明:在使用对抗训练之后,模型的防御表现进一步提高.其中,使用 PGD 作为训练中的扩容方式时提高最为明显:CIFAR-10 下,准确率都提升了近 1 倍;而 MNIST 下,对于 0.15 扰动攻击,更是近 3 倍的提高.BIM 和 MIM 除基本原理相似外,与 PGD 同样使用迭代方法,测试中,防御表现都有一定程度的提高.

- MNIST 下,基本达到了对干净样本的辨识水平;FGSM 攻击下,对于前两种较小的扰动情况都有提高;
- 但是同时,在 CIFAR-10 和 MNIST 下,对第 3 种较大扰动出现准确率下降的情况.我们认为:相比于迭代

的攻击方式,FGSM 在大扰动下对图片的破坏情况相对严重;而使用 PGD 攻击方式做对抗训练,模型达到的局部恒定比较适合图片未被严重破坏的情况.

总体上,本文方法改进并训练的模型在对抗训练前后防御表现有提高,可证明本文方法不与对抗训练冲突.

Table 3 Comparison of classification accuracy against corresponding adversarial examples before and after using adversarial training with PGD (%)

表 3 PGD 对抗训练前后对相似对抗样本的分类正确率的比较 (%)

数据集	攻击类型	攻击参数	F+D 模型	AdvT+F+D 模型
CIFAR-10	FGSM	ε=0.01	64.32	**75.82**
		ε=0.02	60.28	**62.79**
		ε=0.04	**49.44**	46.17
	BIM	ε=0.01	42.26	**75.17**
		ε=0.02	32.42	**58.81**
		ε=0.03	27.58	**45.62**
	MIM	ε=0.01	47.82	**75.53**
		ε=0.02	38.07	**60.23**
		ε=0.03	32.33	**48.33**
	PGD	ε=0.01	43.96	**78.4**
		ε=0.02	31.57	**66.33**
		ε=0.03	24.14	**54.61**
MNIST	FGSM	ε=0.1	94.84	**98.89**
		ε=0.2	65.46	**97.59**
		ε=0.3	**20.68**	12.36
	BIM	ε=0.05	95.76	**99.05**
		ε=0.1	87.8	**98.85**
		ε=0.15	72.84	**98.16**
	MIM	ε=0.05	96.38	**99.05**
		ε=0.1	90.89	**98.85**
		ε=0.15	79.84	**98.2**
	PGD	ε=0.05	95.98	**99.09**
		ε=0.1	75.89	**98.96**
		ε=0.15	38.61	**98.69**

4 结论与展望

针对深度神经网络对于对抗样本的脆弱性问题,本文提出了一种基于特征融合和整体多样性的单模型鲁棒性提升方法.该方法受组合模型防御效果优于单模型的启发,依据分支网络中浅层出口也可以达到较好的预测准确率理论,在现有模型基础上添加额外的分支模拟组合模型效果,同时在分支之间加入特征融合实现特征金字塔,并引入改进后的多分支单模型整体多样性计算辅助训练,以提高模型鲁棒性,使其具有更好的防御能力.通过在 MNIST 和 CIFAR-10 两种数据集上的实验结果表明:本文方法改进并训练的模型防御效果显著,对对抗样本的防御能力比改进前的原模型在 FGSM 等 4 种基于梯度的攻击下有 5 倍以上的提高,JSMA,C&W 以及 EAD 攻击下可达到 10 倍的提升;同时不干扰对干净样本的分类精度,也与对抗训练方法不抵触,可以联合使用,获得更好的防御效果.证明了本文提出的提升鲁棒性方法是可行且有效的.此外,实验中还发现:在不同复杂度的样本上,特征融合和整体多样性带来的鲁棒性影响不同.在今后的工作中,我们会对此方面做深入的研究,以改进本文提出的方法,获得更好的效果.

References:

[1] Lueth KL. State of the IoT 2018: Number of IoT devices now at 7B—Market accelerating. IOT ANALYTICS. https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/

[2] Dourado Jr CM, da Silva SP, da Nóbrega RV, Barros AC, Rebouças Filho PP, de Albuquerqu VH. Deep learning IoT system for online stroke detection in skull computed tomography images. Computer Networks, 2019,152:25–39.

[3] Mookherji S, Sankaranarayanan S. Traffic data classification for security in IoT-based road signaling system. In: Proc. of the Soft Computing in Data Analytics. 2019. 589–599.

[4] Rodrigues JD, Rebouças Filho PP, Peixoto Jr E, Kumar A, de Albuquerque VH. Classification of EEG signals to detect alcoholism using machine learning techniques. Pattern Recognition Letters, 2019,125:140−149.

[5] Zhang Y, Li PS, Wang XH. Intrusion detection for IoT based on improved genetic algorithm and deep belief network. IEEE Access, 2019,7:31711−31722.

[6] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: Proc. of the Int'l Conf. on Machine Learning (ICML). 2018. 274−283.

[7] Lu J, Issaranon T, Forsyth D. Safetynet: Detecting and rejecting adversarial examples robustly. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). 2017. 446−454.

[8] Metzen JH, Genewein T, Fischer V, Bischoff B. On detecting adversarial perturbations. In: Proc. of Int'l Conf. on Learning Representations (ICLR). 2017.

[9] Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proc. of the 10th ACM Workshop on Artificial Intelligence and Security. 2017. 3−14.

[10] Liao FZ, Liang M, Dong YP, Pang TY, Zhu J, Hu XL. Defense against adversarial attacks using high-level representation guided denoiser. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2018. 1778−1787.

[11] Pang TY, Xu K, Du C, Chen N, Zhu J. Improving adversarial robustness via promoting ensemble diversity. In: Proc. of Int'l Conf. on Machine Learning (ICML). 2019. 4970−4979.

[12] Teerapittayanon S, McDanel B, Kung H. BranchyNet: Fast inference via early exiting from deep neural networks. In: Proc. of the IEEE Int'l Conf. Pattern Recognition (ICPR). 2016. 2464−2469.

[13] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017. 936−944.

[14] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2016. 770−778.

[15] Ranjan R, Sankaranarayanan S, Castillo CD, Chellappa R. Improving network robustness against adversarial attacks with compact convolution. arXiv preprint arXiv:1712.00699, 2017.

[16] Miyato T, Maeda SI, Koyama M, Ishii S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018,41(8):1979−1993.

[17] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2017.

[18] Kurakin A, Goodfellow I, Bengio S, Dong YP, Liao FZ, Liang M, Pang TY,Zhu J, Hu, XL, Xie CH, *et al*. Adversarial attacks and defences competition. In: Proc. of the NIPS 2017 Competition: Building Intelligent Systems. Cham: Springer-Verlag, 2018. 195−231.

[19] Samangouei P, Kabkab M, Chellappa R. Defense-Gan: Protecting classifiers against adversarial attacks using generative models. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2018.

[20] Guo C, Rana M, Cisse M, Van Der Maaten L. Countering adversarial images using input transformations. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2018.

[21] Lamb A, Binas J, Goyal A, Serdyuk D, Subramanian S, Mitliagkas I, Bengio Y. Fortified networks: Improving the robustness of deep networks by modeling the manifold of hidden representations. arXiv preprint arXiv:1804.02485, 2018.

[22] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2015.

[23] Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. In: Proc. of the Int'l Conf. on Learning Representations (ICLR) Workshop. 2017.

[24] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2018.

[25] Dong YP, Liao FZ, Pang TY, Su H, Hu XL, Li JG, Zhu J. Boosting adversarial attacks with momentum. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2018. 9185−9193.

[26] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proc. of the 2016 IEEE European Symp. on Security and Privacy (EuroS&p). 2016. 372−387.

[27] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the 2017 IEEE Symp. on Security and Privacy (S&P). 2017. 39−57.

[28] Chen PY, Sharma Y, Zhang H, Yi JF, Hsieh CJ. Ead: Elastic-net attacks to deep neural networks via adversarial examples. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI). 2018. 10−17.

[29] Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008,9:2579−2605.

[30] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M. Tensorflow: A system for large-scale machine learning. In: Proc. of the 12th USENIX Symp. on Operating Systems Design and Implementation (OSDI). 2016. 265−283.

[31] Papernot N, Faghri F, Carlini N, Goodfellow I, Feinman R, Kurakin A, Xie C, Sharma Y, Brown T, Roy A, Matyasko A. Technical report on the cleverhans v2.1.0 adversarial examples library. arXiv preprint arXiv:1610.00768, 2016.

韦璠(1996－),男,硕士生,CCF 学生会员,主要研究领域为深度学习,对抗样本防御.

宋云飞(1994－),男,硕士生,CCF 学生会员,主要研究领域为人工智能安全.

邵明莉(1997－),女,硕士生,CCF 学生会员,主要研究领域为深度学习.

刘天(1988－),男,博士生,CCF 学生会员,主要研究领域为新型非易失性存储,嵌入式系统,机器学习.

陈小红(1982－),女,博士,副教授,CCF 专业会员,主要研究领域为需求工程,形式化方法.

王祥丰(1987－),男,博士,副教授,CCF 专业会员,主要研究领域为分布式优化,多智能体强化学习.

陈铭松(1982－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为信息物理融合系统设计自动化,计算机体系结构,物联网技术,形式化方法.