软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.000000]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

预测不确定性与对抗鲁棒性的关系研究*

陈思宏[1], 沈浩靖[1], 王 冉[2], 王熙照[1,3]

[1](深圳大学 计算机与软件学院,广东 深圳 518060)

[2](深圳大学 数学与统计学院,广东 深圳 518060)

[3](深圳大学 广东省智能信息处理重点实验室,广东 深圳 518060)

通讯作者: 王熙照, E-mail: wizhaowang@ieee.org

摘 要: 对抗鲁棒性指的是模型抵抗对抗样本的能力,对抗训练是提高模型对抗鲁棒性的一种常用方法.然而,对抗训练会降低模型在干净样本上的准确率,这种现象被称为 accuracy-robustness problem.由于在训练过程中需要生成对抗样本,这个过程显著增加了网络的训练时间.在本文中,我们研究了预测不确定性与对抗鲁棒性的关系,得出以下结论:预测不确定性越大,则模型对抗鲁棒性越大.结论解释为:用交叉熵训练得到的模型边界并不完美,为了使得交叉熵最小化,可能使得一些类的分类面变得狭隘,导致这些类的样本容易受到对抗攻击.如果在训练模型的同时最大化模型输出的信息熵,可以使得模型的分类面更加平衡,模型分类面边界与每一类数据的距离尽可能一样远,从而提高攻击难度.在此基础上,提出一种新的增强对抗鲁棒性的方法,通过增加模型预测的不确定性,以达到提高鲁棒性的目的;它在保证模型准确率的同时,使得模型预测的信息熵达到更大.在 MNIST、CIFAR-10 和 CIFAR-100 数据集上的大量实验和简化的模型推导都证实了对抗鲁棒性随模型预测不确定性的增加而增加的统计关系.本文的方法也可结合对抗训练,进一步提高模型对抗鲁棒性.

关键词: 对抗样本;不确定性;对抗防御;深度学习;对抗鲁棒性

中图法分类号: TP311

中文引用格式: 陈思宏,沈浩靖,王冉,王熙照.预测不确定性与对抗鲁棒性的关系研究.软件学报.

英文引用格式: Chen SH, Shen HJ, Wang R, Wang XZ. Research on the relationship between prediction uncertainty and adversarial robustness. Ruan Jian Xue Bao/Journal of Software, 2020 (in Chinese).

Research on the Relationship between Prediction Uncertainty and Adversarial Robustness

CHEN Si-Hong[1], SHEN Hao-Jing[1], WANG Ran[2], WANG Xi-Zhao[1]

[1](College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518000, China)

[2](College of Mathematics and Statistics, Shenzhen University, Shenzhen 518000, China)

Abstract: Adversarial robustness describes the ability of the model to resist adversarial examples and adversarial training is a common method to improve the model's adversarial robustness. However, adversarial training will reduce the accuracy of the model on clean samples. This phenomenon is called accuracy-robustness problem. Due to the need to generate adversarial examples during the adversarial training, this process significantly increases the training time of the network. In this article, we study the relationship between prediction uncertainty and adversarial robustness, and draw the following conclusions: the greater the prediction uncertainty, the greater the adversarial robustness.

* 基金项目: 国家自然科学基金重点项目(61732011); 国家自然科学基金项目(61976141,61772344,61732011); 深圳大学自然科学基金(827-000230); 深圳大学跨学科创新小组

Foundation item: Key Project of Natural Science Foundation of China (61732011); National Natural Science Foundation of China (61976141,61772344,61732011); Natural Science Foundation of Shenzhen University (827-000230); Interdisciplinary Innovation Team of Shenzhen University

收稿时间: 2020-08-08; 修改时间: 2020-09-30; 采用时间: 2020-10-09; jos 在线出版时间: 0000-00-00

CNKI 在线出版时间: 0000-00-00

The conclusion is explained as: the boundary of the model obtained by cross-entropy is not perfect. In order to minimize the cross-entropy, the classification surface of some classes may become narrow, which makes the samples of these classes vulnerable to adversarial attacks. And if we maximize the output's information entropy while training the model, we can make the classification surface of the model more balanced, that is, make the distance between boundary and data as far as possible, which makes it more difficult for the attacker to attack the samples. Based on this finding, we propose a new method to improve the adversarial robustness of the model, by increasing the uncertainty of the model's prediction to improve the adversarial robustness of the model. While ensuring the accuracy of the model, we make the prediction's information entropy larger. Extensive experiments and simplified model derivations on the MNIST, CIFAR-10 and CIFAR-100 datasets have confirmed the statistical relationship that the adversarial robustness increases with the increase of the model's prediction uncertainty. The method in this paper also can be combined with adversarial training to further improve the model's adversarial robustness.
Key words: adversarial training; uncertainty; adversarial defense; deep learning; adversarial robustness

深度学习已经被广泛应用于实际生活中,包括图像分类、人脸识别、语音识别等,甚至在有些任务中能超出人类的表现.但最近研究[4,23]表明,深度学习很容易受到对抗样本的攻击.对于攻击者精心设计的用于欺骗分类器的样本,模型很容易做出错误的分类,而且这类样本对于人类来说是不可察觉的.在某些强调安全的相关领域,比如人脸识别、自动驾驶等,如果无法抵抗住对抗攻击,例如:攻击者可以很容易地伪装成别人,自动驾驶中的汽车没有判断到前方的行人,这会造成巨大的安全隐患.对抗样本在这些应用的推广中带来了巨大的阻碍,因此模型对于对抗样本的鲁棒性就显得至关重要.我们希望我们的模型准确率更高,同时也要有更好的对抗鲁棒性,其训练时间也希望可以尽可能地短,也就是训练时间在一定范围内,准确率与普通训练相差无几,同时使得我们模型对抗鲁棒性更好.

目前来说, 对抗训练[1,4,6,13]是最常用的一种提高模型对抗鲁棒性的方法.对抗训练是一种数据增强的方法,主要做法就是在训练集中添加对抗样本,并重新训练模型.直观上来说,这样做可以使得模型学习到对抗样本的特征,对于这种攻击方法产生的对抗样本会有更好的鲁棒性.通过添加足够多的样本到训练集中,对于其他攻击方法产生的对抗样本也能获得一定的鲁棒性.Madry 等人[6]将对抗训练过程公式化为一个鞍点问题(公式(1)),现有的对抗训练方法都是基于该最小-最大化公式.然而,对抗训练会降低模型在干净样本上的准确率,这种现象被称为 accuracy-robustness problem. Tramer 等人[1]也做了大量的实验证明了对抗训练的确会降低模型预测的准确率.此外,因为在训练过程中需要生成对抗样本,即公式(1)中的最大化过程,使得整个训练过程的时间大大增加.

$$\min_{\theta} E_{(x,y)\in D}\left[\max_{\delta\in S} L(\theta, \boldsymbol{x}+\boldsymbol{\delta}, y)\right] \tag{1}$$

其中L指的是损失函数,$\boldsymbol{\delta}$指的是加到样本上的扰动,S是对扰动所加的限制.最大化过程的目的是找到使得模型做出错误判断最严重的对抗样本,之后最小化的过程可以视为一个矫正过程,让模型学习犯错最严重的的对抗样本的特征,并训练模型使其正确分类.

我们将深度学习中的不确定性分为三类,分别为模型的不确定性、数据的不确定性以及预测的不确定性,其具体含义将会在 2.3.1 节中进行阐述.在本文中,我们主要研究预测的不确定性与模型对抗鲁棒性的关系.预测的不确定性反映了模型预测的置信度,如果模型对于预测结果的置信度越低,则不确定性越大,一般我们使用的是信息熵[14]来度量预测的不确定性.据我们所知,研究预测不确定性与对抗鲁棒性关系的相关文章仅有 Guided Complement Entropy(GCE)[3],chen 等人通过摊平不正确类的概率以使得模型对于正确类的结果更加置信,从而提高模型的对抗鲁棒性,其做法就是最大化不正确类预测概率的信息熵,但他们只是提出了一个直观的想法.我们从理论上推导出了预测不确定性与对抗鲁棒性的关系,并在这基础上提出了一个新的提高模型鲁棒性的方法,本文将以 GCE 方法作为对照组,在实验阶段用于比较.

在本文中,我们研究了预测不确定性与对抗鲁棒性的关系,并得出以下结论:预测不确定性越大,则模型的对抗鲁棒性越好.我们认为,用交叉熵训练得到的模型边界并不完美,为了使得交叉熵最小化,可能使得一些类的分类面变得狭隘,导致这些类的样本容易受到对抗攻击的影响.而如果我们在训练模型的同时最大化输出的信息熵,可以使得模型的分类面更加平衡,模型分类面边界与每一类数据的距离都尽可能远,从而使得攻击者更难对样本进行攻击.在此基础上,我们提出一种新的增强对抗鲁棒性的方法,通过增加模型预测的不确定性,以达到

提高鲁棒性的目的.我们在保证模型准确率的同时,使得模型预测的信息熵变得更大.我们在 MNIST,CIFAR-10 和 CIFAR-100 数据集上都做了相应的实验,并采用白盒攻击来评估模型的鲁棒性.为了更为全面地评估我们的方法,我们用了三种类型的 p-范数距离来控制扰动大小. 大量实验和简化的模型推导都证实了对抗鲁棒性随模型预测不确定性的增加而增加的统计关系,验证了本文提出的方法的有效性.最后,本文提出的方法也可结合对抗训练,进一步提高模型对抗鲁棒性.

本文的贡献可总结如下:

(1) 我们研究了不确定性与对抗鲁棒性的关系,并得出结论:模型预测不确定性越大,则对抗鲁棒性越好.

(2) 从不确定性和鲁棒性的关系出发,我们提出一种新的度量模型鲁棒性的方法.

(3) 基于上面的发现,我们提出了一个新的方法用于提高模型鲁棒性.该方法能在不降低模型预测干净样本能力的同时,提高模型的信息熵,从而达到提高模型对抗鲁棒性的效果.此外,将我们的方法与对抗训练结合,也能比正常对抗训练得到更好的对抗鲁棒性.

(4) 我们从实验上证明了所提出方法的有效性.在 MNIST,CIFAR-10 和 CIFAR-100 数据集上做的大量实验结果显示,我们的方法的确能显著提高模型的对抗鲁棒性.

本文在第 1 节介绍了对抗样本的概念并总结了一些对抗攻击以及对抗防御的方法.在第 2 节中,简单介绍不确定性的相关概念,并进一步细化到深度学习中的不确定性以及不确定性对模型鲁棒性的影响.在第 3 节研究预测不确定性与对抗鲁棒性的关系,并从理论上推导出它们之间的关系,基于不确定性给出了一个提高模型鲁棒性的方法,同时,我们用一个简单的例子解释其原理.第 4 节中我们分别利用交叉熵、GCE 以及我们的方法训练模型,通过比较多种攻击下模型的准确率验证了我们所提方法的有效性.最后,对全文进行总结.

1 相关工作

Szegedy 等人[16] 首次发现神经网络存在对抗样本后,许多学者发布了相关研究[4,5,6,9,24,25,26].对抗样本指的是攻击者将对于人类无法察觉的扰动添加到样本上,使得模型对于原本能正确分类的样本进行错误分类[28].为了量化这种不可察觉性,我们一般将扰动的不可察觉性转换为对于所加扰动$\boldsymbol{\delta}$的L_p范数上的限制ϵ.对抗样本是在原样本$\boldsymbol{x}$上加入了一定限制的扰动,而且在加了扰动后,会导致模型f进行误分类.我们用以下集合表示对抗样本:

$$A(\boldsymbol{x}) = \{\boldsymbol{x}+\boldsymbol{\delta} | f(\boldsymbol{x}+\boldsymbol{\delta}) \neq f(\boldsymbol{x}), ||\boldsymbol{\delta}|| \leq \epsilon\} \tag{2}$$

其中ϵ表示最大扰动限制.

1.1 对抗攻击

在对抗攻击中,我们把对抗样本的不可察觉性转化为对于扰动大小的限制,通过p-范数距离来控制扰动的大小,常见的有L_0, L_2和L_∞三种常用的L_p度量[22].其中L_0距离表示对抗样本中被修改的像素的数量;L_2距离为原样本与对抗样本之间的欧氏距离;L_∞距离表示对抗样本中所有像素最大的改变值.

现有对抗攻击基本都是基于上面所述的度量方式,FGSM[4]就是一种典型的L_∞度量的攻击,在原样本的基础上, 沿 Loss 函数的梯度的符号函数方向加入扰动,并用L_∞范数控制该扰动大小,从而使模型对于对抗样本的原标签的损失变大的目的.JSMA[9]是一种L_0度量的攻击,该方法使用雅可比矩阵构造显著图,以选择每次迭代时要修改的像素,该方法根据修改的像素个数来度量扰动大小.Carlini 等人[5]中提出了基于几种距离度量方式的优化式用于产生对抗样本,这里我们使用的是其中一种,在优化式中对于扰动的限制用的是L_2距离.此外,还有一些方法基于这几种度量方式都能进行攻击的,比如 BIM[7],MIM[8],PGD[6]攻击,这些方法都是一些迭代的攻击方法,每次的迭代步伐用的是L_∞距离度量,它们的每次迭代都是一次小步伐的 FGSM 攻击.BIM 就是一种最基本的小步伐的迭代 FGSM.MIM 在 BIM 的基础上加入了动量,以提高攻击的成功率.PGD 在 BIM 的基础上加入了 random start,即每次攻击前在样本中加入一些噪音再进行攻击,并被证明是一阶导攻击方法中最强大的一种.这几种方法可以使用三种度量方式中的一种来控制总的扰动.

1.2 对抗防御

提高模型的对抗鲁棒性,要比产生对抗样本更加困难.许多研究提出了防御方法以提高模型的对抗鲁棒性.常见的方法有对抗训练,修改输入,添加正则项等[27].

对抗训练是一种数据增强的方式,主要做法就是在训练集中添加对抗样本,并重新训练模型.直观上来说,这样做可以使得模型学得对抗样本的特征,对于这种攻击方法产生的对抗样本会有更好的鲁棒性,对于其他攻击方法的对抗样本也有一定的泛化能力.Goodfellow 等人[4]首次将用 FGSM 攻击生成的对抗样本加入到训练样本中,并证实了这种做法的确能提高鲁棒性.Madry 等人[6]提出了 Projected Gradient Descent(PGD)攻击方法,并将其用于对抗训练,并证明只要能抵抗该种攻击方法,则能防御住其余利用 first-order 信息的攻击方法,该方法也是目前唯一没有被攻破的防御方法,此外,该文章还将对抗训练转化成一个鞍点问题.之后的对抗训练方法大部分都是基于该方法,比如 cheng 等人[13]在对抗训练中自适应地调整扰动大小以及训练的标签,解决干净样本到对抗样本的泛化问题.Zhang 等人[17]中得到了对抗样本的预测误差上界,并将它用于改进对抗训练方法,并能很好地在对抗鲁棒性以及干净样本准确率之间取的平衡.

但是,如果我们需要更好的对抗鲁棒性,则需要的内部优化时间会更长,现在也有大量的工作致力于减少对抗训练的时间.Shafahi 等人[18]通过回收在更新模型参数时计算出的梯度信息来减少生成对抗样本的开销成本,并证明这种方法能得到与 madry 等人[6]所提方法得到差不多的性能,同时时间更短.Wong 等人[19]证明可以使用更弱,代价更小的攻击来训练更为鲁棒的模型,文中证明结合了随机初始化的 FGSM 对抗训练与基于 PGD 的训练一样有效,但却可以大大地降低时间成本.

虽然前面所提的方法的确能减少对抗训练的时间,但相比正常训练,对抗训练所需要的时间还是要远远超过正常训练的.而添加正则项是通过在成本函数中添加类似于惩罚项的正则项来提高目标模型的泛化能力,并使模型具有良好的适应性,以抵抗预测中未知数据集的攻击.Ma 等人[11]提出的正则项是鲁棒优化目标式的最大上界,即该正则项为公式(1)最小-最大化问题中最大化问题的上界,这篇文章用正则项来代替产生对抗样本的过程,从而大大节约了训练的时间.Jin 等人[12]发现了局部稳定与 manifold regularization 的关系,并提出了一个新的正则化项来训练对一类局部扰动稳定的深度神经网络.这种防御方法跟对抗训练相比好处在于：因为对抗训练需要生成对抗样本来加入到训练集中,而且普遍来说越强大的攻击得到的样本加入到训练中效果会越好,但越强的攻击同时也代表需要耗费更多的时间.如果仅是添加正则项用于训练,其时间复杂度与正常训练差不多,但也能得到提高鲁棒性的效果,而且原则上能提高所有攻击方法下的鲁棒性.

Chen 等人[3]通过最大化不正确类的信息熵,即最大化不正确类之间预测的不确定性,达到摊平不正确类间概率的效果,使得模型对于正确类的预测更加置信,从而提高模型的鲁棒性.他们提出了一个 Guided 项来达到最大化不正确类间的熵的目的,该方法在损失函数中加入了惩罚项以提高模型对于对抗样本的泛化能力.但只是直观上描述这样可以提高对抗鲁棒性,本文从理论上推导出预测不确定性与对抗鲁棒性的关系,并在这基础上提出了一个新的提高模型鲁棒性的方法.与该方法类似,我们也是通过正则项使得模型的信息熵变大,从而达到提高模型对抗鲁棒性的效果.但我们的方法区别在于,我们不仅仅是摊平不正确类的概率,同时也允许正确类的预测的概率变低.相比于模型对正确分类的置信度,我们更着重于提高模型的对抗鲁棒性.

2 不确定性

在本节中,我们会介绍一些不确定性的定义,并对机器学习中的不确定性进行分类.之后我们再研究不确定性在深度学习中的应用,并将深度学习中的不确定性再次细分.最后,我们会介绍一种预测不确定性的度量以及该度量在深度学习和提高对抗鲁棒性方面的应用.

2.1 不确定性的定义

不确定性指的是涉及不完全或未知信息的认知情况.它适用于对未来事件的预测,已经进行的物理测量或未知的事物.

2.2 不确定性的分类

机器学习中有多种不同形式的不确定性.某些不确定性的概念描述了我们可预期的事件固有的随机性,比如抛硬币的结果,而另一些概念则描述了模型推理结果的不可靠性[29].一般来说,我们将不确定性分为以下几种类型:

(1) 偶然事件不确定性:在收集数据的过程中包含着随机性,且这些随机性无法通过收集更多的数据来消除[29].比如我们目前有的数据是温度,但这些数据收集过程,可能由于仪器精度问题,会与真实气温有一点偏差.这就是偶然事件不确定性.

(2) 认知不确定性: 认知不确定性衡量了我们模型预测的信息缺乏程度.如果测试数据与训练数据的距离越远,则这些数据的预测应该有更高的标准差.与偶然事件不确定性的差异是,通过收集更多数据和减少模型缺乏知识的区域可以降低认知不确定性[29].比如在气温预测中,一般温度都是在 20-30 度,但是有几天的数据为 15-20 度,模型可能无法处理这样的信息.我们可以通过收集更多包含 15-20 度的数据来训练模型.

(3) 超出分布的不确定性:一般来说,训练的模型是用于处理专门的任务的.如果我们将湿度的数据提供给用于预测气温的模型,那么预测结果很可能不如气温预测那么准确,甚至毫无意义.

2.3 深度学习中的不确定性

2.3.1 不确定性在深度学习中的分类

不确定性在深度学习中到处可见,在我们训练模型的各个阶段都存在不确定性,比如数据预处理时,我们无法确定应该如何填充数据,选择什么特征或者是否应该正则化.在设计模型的时候,我们应该选择什么样的模型结构,什么样的方式用于训练.即使是在测试时,如果我们有多个训练好的模型,选择哪一个模型用于测试也是充满了不确定性.最后,对于模型预测的结果,我们也并不是百分百确信,这里也包含着不确定性[21].

因此,在深度学习中,我们将不确定性再次细分为以下几类:

(1) 模型的不确定性:为了训练同一个任务,有多种模型可以用于选择.比如训练 CIFAR-10 数据集时可以选择 VGG 或者 ResNet 模型.都可以达到很不错的效果,但选择哪一个模型,其泛化能力或者其它方面的性能要更好一点,模型的选择包含了不确定性.其次,即使我们选定了模型,对于同样的任务,用同一个模型训练的好几次,得到的模型参数也可能会完全不同,如何选择参数用于测试也包含不确定性.

(2) 数据的不确定性:训练中所收集到的数据可能是包含噪音的.正如上文所说,数据收集过程中,本身就包含着一些随机性,在训练过程中这些随机性可能成为噪音,影响我们最终的训练结果.

(3) 预测的不确定性:模型观察到的标签也是不确定的,对于一张猫的图片,预测的结果可能是 60%可能性是猫,40%可能性是狗.虽然最后预测的结果是猫,这里预测的不确定性很大.

2.3.2 信息熵

信息熵最早是由香农[14]提出的,用于解决对于信息的量化度量问题.当我们不知道某事物具体状态,却知道它有几种可能性时,显然,可能性种类越多,不确定性也会越大,因为我们更难确认其状态.当我们最终得知了该事物的状态,相当于从中得到了关于该事物的信息.对于越不确定的事物,我们从中能得到的信息则越多,从这事物能获得的信息量越大,其中某个问题的所有可能取值的信息量的期望就称为信息熵.Seedat 等人[20]提到了四种评估模型不确定性的方式,其中我们可以用信息熵来度量预测的不确定性,对预测得到的向量,计算其信息熵,用信息熵表示该向量的不确定性.本文我们主要考虑预测不确定性的影响,对于一个向量$\boldsymbol{x}$,其信息熵计算公式如下:

$$H(\boldsymbol{x}) = -\sum_{i=1}^{N} p_i \log(p_i) \tag{3}$$

其中,p_i指模型预测样本为类别i的置信度.

2.3.3 信息熵对模型的影响

在深度学习中,我们一般用交叉熵来训练模型.交叉熵是由信息熵引申出的一个概念,因为仅仅最小化信息熵只能最小化向量的不确定性,但却无法保证输出的结果收敛到正确的标签,因此我们常常用交叉熵训练模型,其计算公式如下:

$$CEH(\boldsymbol{x}) = -\sum_{i=1}^{N} p_i \log(q_i) \tag{4}$$

其中$p_i \sim P$,P表示真实分布,即样本的 label.$q_i \sim Q$,Q表示预测的分布.我们用预测的分布Q来拟合真实分布P,交叉熵衡量了真实分布和预测分布之间的差异.

在一般的训练过程中,需要最小化交叉熵以最小化它们之间的差异,在这个过程中,预测的分布会慢慢接近真实分布,而真实分布仅仅有一个标签为 1,其余标签都是 0,由信息熵公式可知真实分布的信息熵为 0.所以在最小化交叉熵的过程中,预测结果的信息熵会慢慢变小,模型输出更为确信,准确率也会逐渐变高.

2.3.4 Guided Complement Entropy

Guided Complement Entropy(GCE) 是由 Chen 等人[3]提出的,他们提出了一个“guide”项对不正确类的模型概率进行中和,以平衡正确类以及不正确类.为了使得对于预测的正确类更加确信,他们将不正确类的概率拉平,尽可能使得不正确类分布更加均匀.他们认为,对于预测结果更加确信可以使得预测的准确率提升,而且还可以达到提高模型对抗鲁棒性的目的.而使得不正确类分布更加均匀最理想的情况就是他们的概率都为$1/n$,这种情况对应于这些类预测的信息熵最大化的情况.为此,他们提出以下 Loss 函数用于训练模型:

$$-\frac{1}{M}\sum_{i=1}^{M} \hat{y}_{ig}^{\alpha} \frac{1}{\log(N-1)} \sum_{j=1, j\neq g} \left(\frac{\hat{y}_{ij}}{1-\hat{y}_{ig}}\right) \log\left(\frac{\hat{y}_{ij}}{1-\hat{y}_{ig}}\right) \tag{5}$$

其中第一项用于提高模型的准确率,第二项称为 complement loss factor,通过最小化不正确类的信息熵来摊平不正确类输出的概率,M为训练时一个 batch 中样本的数量,N为类别数目,g表示正确类.

可以看到公式(5)的第二项就是公式(3)的信息熵公式,只是这里没有将正确类的概率加入运算,该项的目的是最大化不正确类的熵.此外,该表达式用$\hat{y}_{ig}$作为指导因素,根据模型的预测质量来控制 complement loss factor 的影响.当$\hat{y}_{ig}$比较小时,即模型对于预测不是很确信时,这时减少 complement loss factor 的影响,主要提升模型的预测能力.而当$\hat{y}_{ig}$比较大时,对应于模型对正确类比较确信的时候,此时 complement loss factor 的影响也会变大,这时优化器会倾向于调整不正确类的概率,训练出对抗鲁棒性更强的模型.

3 一种新的提高鲁棒性的方法

本节主要研究神经网络的不确定性与对抗鲁棒性的关系.通过提高模型预测不确定性,提高模型的对抗鲁棒性,在阐述该方法有效性的同时给予严格的证明.第 3.1 节对比了正常训练以及对抗训练的不确定性,并从实验中观察得到:模型输出的不确定性越大,则对抗鲁棒性越好.第 3.2 节中从不确定性与鲁棒性的关系出发,定义一种新的度量模型鲁棒性的方法.第 3.3 节在 3.2 节中提出的新度量的基础上,提出一种新的训练方法以提高鲁棒性.最后,在 3.4 节中,给出一个简单的例子解释我们的方法的有效性.

3.1 预测不确定性与对抗鲁棒性定义

如 2.3.1 节所述,预测不确定性反映了模型对于预测结果的置信度,对于最终的预测结果越确信,则模型的预测不确定性越低,反之则不确定性越高.文中我们用信息熵表示预测不确定性,信息熵越大则预测不确定性越大.对抗鲁棒性表示模型抵抗对抗样本的能力,受对抗样本影响程度越小,则对抗鲁棒性越高,文中我们用多种对抗攻击下模型的识别准确率来评估模型的对抗鲁棒性,如果准确率越高,则对抗鲁棒性越高.

3.2 正常训练以及对抗训练的不确定性对比

从 2.3.3 节可知,最小化交叉熵会使模型的输出结果的信息熵越来越小.而且,直观上来说,预测更加确信,显得模型更加稳定.但是,在实际中发现,从对抗鲁棒性的角度上看,信息熵并不是越小越好,在保证两个模型准确率相差不多的情况下,输出结果的信息熵越大的模型,其对抗鲁棒性往往要更优.

图 1 和图 2 分别为 LeNet-5 对于 MNIST 数据集中正常样本以及用 FGSM 生成的对抗样本的输出结果分布(为了更直观地看,我们把正确类的概率省略了),图中橙色代表的是对抗训练模型的输出分布,蓝色的是正常训练模型分布.可以看到正常训练的模型其不正确类概率大部分分布在 0 附近.经过对抗训练后的模型,无论是对于正常样本或者是对抗样本,其预测都会变得更加保守,对于模型输出结果中正确的类,其置信度要稍微低于正常训练的模型.用熵的角度来说,就是模型输出结果的信息熵会更大一些.因此我们用公式(3)计算了一下对抗训练以及正常训练的样本平均信息熵,用 PGD 进行对抗训练后模型预测平均信息熵为 0.5067,相比于正常训练的平均信息熵 0.0637,对抗训练后信息熵要比正常训练后的模型更大,模型的预测结果会更加保守.而且对抗训练中学习了对抗样本的特征,其对抗鲁棒性当然也会更好.根据这个现象,我们提出以下两个观点:

(1) 对抗训练会使得模型的输出更加保守,使得模型输出分布更加均匀,对抗训练不会像正常训练得到的模型一样,对于模型输出有很高的置信度.这意味着,对抗训练会提高模型输出结果的熵,因此,如果我们不通过对抗训练而是通过别的方法提高模型输出的信息熵,是否也能达到提高鲁棒性的目的.

(2) 在训练中,并不是使得熵越小越好的.因为训练过程中,为了最小化信息熵,可能使得某个类或者几个类的分类面变得狭隘,这样对这些类进行攻击会变得更加容易.因此我们可以使用一些相对更小的扰动,使得模型被误分类为另一个类.具体的我们会在 3.4 节进行阐述.

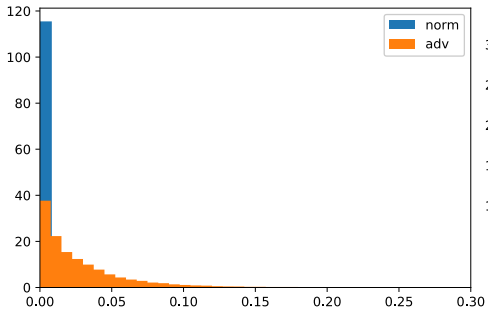

Fig.1 The output probatility distribution of clean samples by normal training and adversarial training

图 1 正常训练以及对抗训练下的 LeNet 模型预测正常样本输出概率分布图

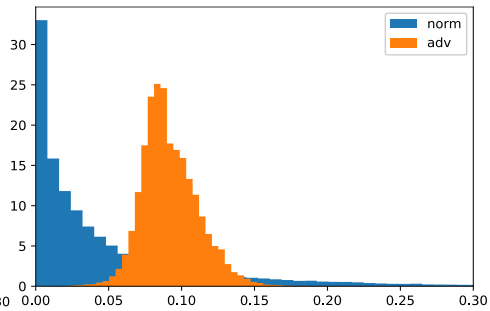

Fig.2 The output probatility distribution of adversarial eamples by normal training and adversarial training

图 2 正常训练以及对抗训练下的 LeNet 模型预测对抗样本输出概率分布图

3.3 一种新的度量模型对抗鲁棒性的方法

正如 3.2 节中所述,我们观察到模型具有更好的鲁棒性,则其输出结果的信息熵越大.接下来,我们将从不确定性与鲁棒性的关系出发,定义一种新的度量模型鲁棒性的方法.

我们用不确定性去衡量样本点到决策边界的距离程度.如果样本点距离某一类的边界越近,则表示样本点对该类的输出的概率越高,且样本被错分成该类的可能性越大,也就是可以用更小的扰动ϵ使得模型误分类.因此,我们希望样本点对所有错误类的预测的置信度越低越好,使得样本点距离分类边界都有足够长的距离.假设样本相对位置是不变的,分类面边界的位置也相对确定,但我们可以对分类面边界做出小范围的改变.为了保证样本距离所有分类边界都有足够长的距离,一般我们常用的做法会最小化信息熵,但这样得到的分类面并不完美.为了使得信息熵最小化,可能使得某些类的样本分类面变得狭隘,样本距离某个分类边界距离太近,因此容易被攻击,如图 3 中中间类的样本. 我们用信息熵表示距离,样本距离某一些不正确类边界较近时,模型输出的信

息熵较低,但容易受到对抗攻击.与之相反,当某个样本与其它类分类边界距离越远,此时模型输出的信息熵越大,样本也不容易受到对抗攻击.因此我们通过最大化不正确类间的信息熵,来保证分类面边界不会产生偏移,即牺牲样本距离一些分类边界的距离,使得样本距离所有边界都足够远,从而达到提高鲁棒性的效果.定义

$$U(\boldsymbol{x}) = -\sum_{i \neq y} f_i(\boldsymbol{x}) log f_i(\boldsymbol{x}) \tag{6}$$

其中$U(\boldsymbol{x})$表示样本$\boldsymbol{x}$对不正确类的不确定性程度.其中y表示样本$\boldsymbol{x}$的标签.f_i表示模型f输出$\boldsymbol{x}$为类别i的概率.保持正确类的概率不变时,$U(\boldsymbol{x})$越大,则样本$\boldsymbol{x}$越不容易受到对抗攻击.

如上文所说,我们不仅要考虑样本点与边界的问题,同时还需要考虑扰动ϵ的大小,ϵ越大,产生对抗样本越容易.定义$r_i = r(\boldsymbol{x}^{(i)}, y^{(i)}; f)$为模型$f$在样本$(\boldsymbol{x}^{(i)}, y^{(i)})$上的对抗鲁棒性,考虑所有被正确预测的点,我们有

$$r_i = \frac{U(\boldsymbol{x}^{(I)})}{\epsilon} I\left(argmax_k f_k(\boldsymbol{x}^{(i)}) == y^{(i)}\right), \qquad i = 1, \dots, M \tag{7}$$

显然,r_i越大,则模型的对抗鲁棒性越好.为了简洁,记$H(f) = -\sum_i f_i(\boldsymbol{x}^{(i)}) \log f_i(\boldsymbol{x}^{(i)})$.将(7)式展开,得到以下形式:

$$r_i = \frac{-\sum_{i \neq y} f_i(\boldsymbol{x}^{(i)}) \log f_i(\boldsymbol{x}^{(i)})}{\epsilon} I\left(argmax_k f_k(\boldsymbol{x}^{(i)}) == y^{(i)}\right) \tag{8}$$

$$= \frac{-\sum_i f_i(\boldsymbol{x}^{(i)}) \log f_i(\boldsymbol{x}^{(i)}) + f_y(\boldsymbol{x}^{(i)}) \log f_y(\boldsymbol{x}^{(i)})}{\epsilon} I\left(argmax_k f_k(\boldsymbol{x}^{(i)}) == y^{(i)}\right) \tag{9}$$

$$= \frac{H(f) \cdot I\left(f_{pred}(\boldsymbol{x}^{(i)}) == y^{(i)}\right) + f_y(\boldsymbol{x}^{(i)}) \log f_y(\boldsymbol{x}^{(i)}) \cdot I\left(y_{pred}(\boldsymbol{x}^{(i)}) == y^{(i)}\right)}{\epsilon} \tag{10}$$

$$= \frac{H(f) \cdot I\left(f_{pred}(\boldsymbol{x}^{(i)}) == y^{(i)}\right) - f_y(\boldsymbol{x}^{(i)}) Entropy\left(f(\boldsymbol{x}^{(i)}), y^{(i)}\right)}{\epsilon} \tag{11}$$

则模型对所有样本的鲁棒性和为$r = \sum_i r_i$.

于是,寻找一个具有对抗鲁棒性的模型可描述成如下优化问题:

$$\max r \tag{12}$$

$$s.t.\ \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in D \tag{13}$$

因为ϵ是个定值,所以公式(12)可以改写成:

$$\max \sum_i^M H\left(f^{(i)}\right) - f_{y^{(i)}}(\boldsymbol{x}^{(i)}) Entropy\left(f(\boldsymbol{x}^{(i)}), y^{(i)}\right) \tag{14}$$

$$s.t.\ \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in D \tag{15}$$

$$argmax_k f_k(\boldsymbol{x}^{(i)}) == y \tag{16}$$

将公式(14)改成极小化问题:

$$\min \sum_i^M f_{y^{(i)}}(\boldsymbol{x}^{(i)}) Entropy\left(f(\boldsymbol{x}^{(i)}), y^{(i)}\right) - H\left(f^{(i)}\right) \tag{17}$$

$$s.t.\ \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in D \tag{18}$$

$$argmax_k f_k(\boldsymbol{x}^{(i)}) == y \tag{19}$$

其中第一项是关于交叉熵的一个函数,第二项是模型输出的信息熵.从公式(17)明显可见,信息熵$H(f^{(i)})$越大,目标函数越小,即鲁棒性越强.

因为f通常是一个非凸函数,因此求解式子(17)是一件比较困难的事情.下面,我们将给出一个定理.根据该定理,可以得到公式(17)的一个下确界.

定理 1:设N为问题的类别数目,M为样本数量,$p = f_t(\boldsymbol{x}),\ t = argmax_k f_k(\boldsymbol{x})$.不妨令 $p = \delta + \frac{1}{N}, \delta > 0$,则当

$N > 2, \delta \to 0$时,公式(17)存在下确界.下确界为$\sum_i^M \frac{N-1}{N} \cdot \log\frac{1}{N}$.

证：当正确类概率p确定时,有

$$\sum_i^M f_{y^{(i)}}(\boldsymbol{x}^{(i)}) Entropy\big(f(\boldsymbol{x}^{(i)}), y^{(i)}\big) - H\big(f^{(i)}\big) \tag{20}$$

$$= \sum_i^M \sum_{k \neq y^{(i)}}^N f_k(\boldsymbol{x}^{(i)}) \log\Big(f_k(\boldsymbol{x}^{(i)})\Big) \tag{21}$$

$$\leq \sum_i^M (N-1) \cdot \frac{1-p}{N-1} \log\Big(\frac{1-p}{N-1}\Big) \tag{22}$$

$$\leq \sum_i^M (1-p) \cdot \log\Big(\frac{1-p}{N-1}\Big) \tag{23}$$

记 $T = (1-p) \cdot \log\left(\frac{1-p}{N-1}\right)$,则：

$$\frac{dT}{dp} = -\log\Big(\frac{1-p}{N-1}\Big) - 1 \tag{24}$$

$$= \log\Big(\frac{N-1}{1-p}\Big) - 1 \tag{25}$$

因为$p > \frac{1}{N}$,有$\frac{1}{1-p} > \frac{N}{N-1}$,于是：

$$\frac{dT}{dp} = \log\Big(\frac{N-1}{1-p}\Big) - 1 \tag{26}$$

$$> \log\left((N-1) \cdot \frac{N}{N-1}\right) - 1 \tag{27}$$

$$= \log(N) - 1 \tag{28}$$

又因为$K > 2$,所以$\frac{dT}{dp} > 0$.于是T是递减函数.当$\delta \to 0$时,T取得最小值.此时公式(17)趋近最小值：

$$(17) = \sum_i^M (1-p) \cdot \log\Big(\frac{1-p}{N-1}\Big) \tag{29}$$

$$> \sum_i^M \frac{N-1}{K} \cdot \log\frac{1}{N} \tag{30}$$

证毕.

3.4 基于不确定性增大的鲁棒性提高方法

根据定理 1 知道,(17)存在下确界,故用公式(17)作为损失函数最终能使网络收敛.但我们在实验中发现,直接用公式(17)作为网络的损失函数会导致网络收敛速度慢甚至很难收敛.为了加快收敛速度,我们提出一个改进的损失函数:

$$\sum_i^M Entropy\big(f(\boldsymbol{x}^{(i)}), y^{(i)}\big) - \alpha \times H\big(f^{(i)}\big) \tag{31}$$

其中α是超参数,显然有公式(31) ≥ 公式(17),即公式(31)同样存在一个下确界.第一项为公式(4)中提到的交叉

熵,用于训练模型使其能正确分类干净样本.第二项为我们所提出的正则项,即公式(3)的信息熵计算公式,通过添加该项,在训练中我们可以提高所训练的模型预测的信息熵.并用α控制信息熵对于模型的影响,M表示每个batch 中的样本数量,N表示分类任务类别数.其训练过程伪代码如下所示.

算法 1. 训练过程

Input: 模型f, batchsize m,学习率λ,正则项系数α,以及训练的周期数 n.

Output: 模型f的权重θ.

随机初始化模型f的权重θ;

初始化数据集$D = \{x^{(i)}, y^{(i)}\}_i^{len(D)}$; //$len(D)$表示数据集中样本的数目

for $i \leftarrow 1$ to n do

  for $j \leftarrow 1$ to B do //B表示 batch 的数目,即$B = \lceil len(D)/M \rceil$

    从数据集D中取出M个样本,$Q = \{x^{(i)}, y^{(i)}\}_{i=1}^{M}$;

    计算交叉熵$l_1 = \sum_{k=1}^{M} L(f(x_i), y)$;

    计算正则项$l_2 = \sum_{k=1}^{M} \sum_{l=1}^{N} -f_l(x_k) \log(f_l(x_k))$;

    $L = l_1 - \alpha \times l_2$;

    更新网络参数$\theta \leftarrow \theta - \lambda \nabla L$;

  end for

end for

return(θ)

利用算法 1 在 MNIST 和 CIFAR-10 数据集上分别训练 LeNet-5 和 ResNet-18 模型,得到的模型结果如表(1)所示.可以看到我们的方法能在保证模型训练准确率的同时,能够大大提高模型预测的熵.

Table 1 Comparison of models training by cross entropy and our method

表 1 分别使用交叉熵以及我们的方法训练得到的模型对比

MNIST			CIFAR-10		
Loss	Avg entropy	ACC	Loss	Avg entropy	ACC
CE	0.0119	99.00%	CE	0.0637	94.94%
Ours	1.8279	99.22%	Ours	1.8695	94.56%

3.5 不确定性对鲁棒性的影响:一个例子

在本节中,会关于信息熵对于对抗鲁棒性的影响给出一个直观的解释.我们在一个二维平面上,随机生成两类样本,中间蓝色的圆点为正类,周围橙色的三角点为负类,如图 3 所示.之后我们用一个包含一层隐藏层的全连接网络训练这些数据.为了阐述我们所提出方法与交叉熵的不同,我们用交叉熵以及我们的方法分别训练了一个模型.其对应分类边界如图 4 和 5 所示.

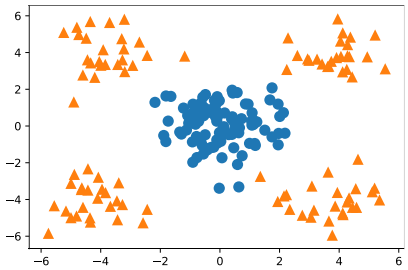

Figure.3 Samples distribution diagram

图 3:随机生成的样本示例图

对比图 4 和图 5 可以看到,在最小化交叉熵的过程中可能会使得某个类的分类面变得狭隘,这样虽然整体的交叉熵很小,结果看起来拟合得很好.但实际上,如果我们加入对样本加入一些微小的扰动,则样本很容易跨过分类面边界,导致模型误分类.而如图 5 所示,我们的方法在保证模型分类正确的同时,最大化模型预测的信息熵,在保证模型分类任务的同时,尽可能使得模型的分类边界处于类别间的一个平衡位置,分类边界与不同类的样本间的距离尽可能差不多.因此,在我们攻击的时候,在样本中加入一些微小的扰动,在交叉熵训练的分类器下,可能会造成误分类,但是我们的方法可以使得模型边界更均衡,更难受到对抗攻击影响,或者说,需要更大的扰动才能造成模型误分类.同时,由图 4 和 5 可以看到,更加均衡的分类面,可以使得模型将原本错误分类的样本正确分类,这也是我们在一些情况下,准确率要比交叉熵训练要高的原因.

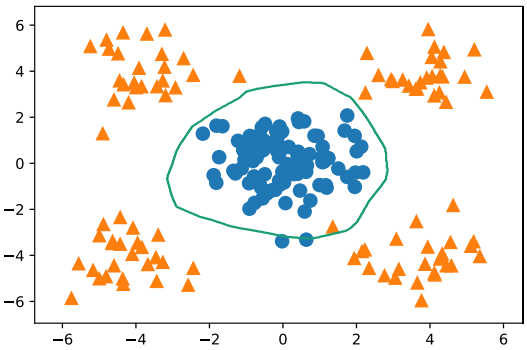

Figure 4. Classification surface trained by cross entropy
图 4:用交叉熵训练得到的分类面示意图

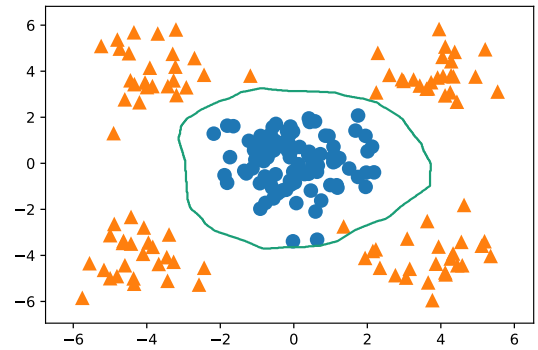

Figure 5. Classification surface trained by our method
图 5:我们的方法训练得到的分类面示意图

4 实验设计与比较

我们用 MNIST[30],CIFAR-10[31]和 CIFAR-100[31]数据集验证我们所提出的方法,在多种攻击方法下,查看模型的准确率,并以准确率作为鲁棒性的估计.并采用了白盒攻击用于评估我们的模型,因为我们认为这是所有攻击中最具有挑战性同时也是最难防御的攻击模式.攻击者能获取到模型的所有信息,比如参数,梯度,输出等等.

4.1 数据集介绍

MNIST 数据集包含 60,000 个训练样本和 10,000 个测试样本.每个样本都是28×28的灰度手写数字图像,数据集中共包含 10 类样本,分别为 0-9.CIFAR-10 数据集是一个更接近普适物体的彩色图像数据集,其中包含 10 类的样本,每一类的样本包含 5,000 个训练样本以及 1,000 个测试样本,共有 50,000 个训练样本以及 10,000 个测试样本,每个样本都是$32 \times 32 \times 3$的彩色图像,CIFAR-100 与 CIFAR-10 相似,但其包含 100 类的样本,每类包含 500 个训练以及 100 个测试样本.并选用 LeNet-5[32]来训练 MNIST 数据集,ResNet18[33]来训练 CIFAR 训练集.同时为了增强数据,以得到更好的结果,对于两个 CIFAR 数据集在训练时加入了水平翻转以及随机裁剪的操作.

4.2 攻击方法

为了验证模型的性能,我们选用了三种范数限制的攻击方式,即L_0,L_2和L_∞攻击.对于L_∞攻击,我们使用 FGSM[4],BIM[7],MIM[8],PGD[6],对于L_2攻击,我们使用了 CW 攻击[5].对于L_0攻击,我们使用的是 JSMA 攻击[9].攻击参数设置方面,对于 MNIST 数据集和 CIFAR-10 和 CIFAR-100 数据集,我们都按照文献[6]中提及的参数进行攻击.在 MNIST 数据集中,ϵ设置为 0.3, 对于迭代的攻击, 我们将每次迭代的步伐设置为 0.01,在 BIM 中,将迭代次数设置为 10,MIM 和 PGD 中设置为 40.对于 CIFAR-10 和 CIFAR-100 数据集,我们将ϵ设置为$8/255$,对于迭代的攻击,我们将迭代的步伐设置为 1/255,在 BIM 中,迭代次数设置为 10,MIM 和 PGD 中设置为 40.在 CW 攻击中,我们统一将 confidence 设置为 0,学习率设为 0.01,初始常数设置为 0.001,且最大的迭代次数设置为 1000.为了便于复现,文中所有的攻击方法我们使用文献[2]中的攻击源码.

4.3 超参数设置

我们选用随机梯度下降(Stochastic Gradient Descent)方法用于优化,对于 LeNet-5 模型,训练 20 个周期,学习率设为 0.01,动量设为 0.9.对于 ResNet-18 模型,我们选用了学习率衰减的方式,共训练 240 周期,其中第 1-135 周期学习率为 0.1,136-185 周期学习率为 0.01,186-240 周期学习率为 0.001,动量设置为 0.9,并将 weight decay 设置为 0.0005.

关于正则项系数的选择,我们比较了不同系数在 MNIST 和 CIFAR-10 数据集中几种攻击方法下的表现.实验中所使用的模型以及攻击方法参数的设置与前文所述一致.由表 2 可以得到以下结论:加入了正则项后的模型,对抗鲁棒性总是优于正常训练的模型,而且在普通样本上能保持与正常训练相近的准确率.但由于α的设置是一个很复杂的问题,理论上尚不能明确给出其最优值,我们通过实验比较得到.实验结果表明,在 MNIST 数据集中,当$\alpha = 1$时,尽管在 PGD 攻击下提升效果并不是最优,但在大部分攻击下都能得到更好的鲁棒性,因此文中我们把α设置为 1.而在 CIFAR-10 数据集中,实验结果表明当$\alpha = 4$时,其对抗鲁棒性要更好,因此实验中我们将其设置为 4.

Table 2: Adversarial robustness of MNIST and CIFAR-10 dataset under various setting of α

表 2: MNIST 和 CIFAR-10 数据集中在不同α设置下的对抗鲁棒性

α	FGSM	BIM	MIM	PGD	FGSM	BIM	MIM	PGD
0	37.99%	8.62%	4.31%	2.62%	37.99%	8.62%	4.31%	2.62%
1/4	64.67%	57.36%	57.02%	5.29%	42.40%	15.48%	13.58%	8.95%
1/3	61.38%	47.69%	47.56%	2.60%	51.87%	27.88%	14.30%	12.33%
1/2	61.84%	49.11%	48.50%	11.43%	48.95%	18.82%	5.38%	4.48%
1	**74.85%**	**66.30%**	**64.02%**	8.32%	51.71%	25.30%	10.19%	7.30%
2	70.27%	47.10%	49.22%	**14.44%**	53.17%	40.43%	24.62%	16.66%
3	69.96%	42.53%	44.66%	10.46%	**57.86%**	48.81%	39.14%	28.21%
4	61.99%	43.64%	45.00%	10.92%	56.44%	**51.10%**	**41.11%**	**29.30%**

4.4 对比实验

我们训练出用交叉熵作为损失函数,以及用 2.3.4 节中提到的 GCE 作为损失函数,分别训练出模型用于对比,以此显示出我们所提出方法的有效性.训练中参数都如 4.3 节中所提到的,GCE 中的α我们用的是文献[3]实验中所使用的 0.333.

4.5 实验结果

4.5.1 正常训练

我们用所提出的损失函数来训练 MNIST、CIFAR-10 和 CIFAR-100 数据集上的模型.得到的结果分别如表 3、表 4 以及表 5 所示:

Table 3: Accuracy of MNIST dataset under various attack methods of normal training

表 3:正常训练 MNIST 数据集中在各种攻击方法下的识别准确率

Attacks	Param	CE	GCE	Ours
FGSM	$\epsilon = 0.3$	37.99%	46.90%	**74.85%**
BIM	$\epsilon = 0.3$	8.62%	35.13%	**66.30%**
MIM	$\epsilon = 0.3$	4.31%	36.18%	**64.02%**
PGD	$\epsilon = 0.3$	2.62%	7.73%	**8.32%**
JSMA	$\gamma = 0.25$	0.52%	**20.31%**	13.85%
C&W	$c = 0$	0.00%	**3.30%**	1.34%

正如表 3 所示,在 MNIST 数据集下,所提出的方法基本在所有攻击方法上都提高了一定的抵抗能力,特别是 FGSM 等用L_∞距离限制的攻击方法,在这些方法中,性能不但优于正常训练的模型,对比于 GCE 损失函数训练的模型也有更优的性能.在 FGSM 以及 BMI 攻击下,我们的方法对于对抗样本识别的准确率接近于 GCE 模型的两倍.在 MIM 攻击下,普通训练基本无法抵抗这个攻击,但我们的方法较好地识别这些样本,远超过普通训练的模型,也远远要比 GCE 训练的模型要好.PGD 是利用一阶导信息最强的其中一种攻击,我们的模型能保证一定的准确率,也要比 GCE 鲁棒性要更好一点.这里证明了我们方法的有效性.但是在 JSMA 和 C&W 两个方法下,我们的方法没有 GCE Loss 要好,但是对比于正常训练的模型,我们的方法还是能大大提升模型的对抗鲁棒性,而且 JSMA 与 C&W 两种攻击方法所需要的时间也远远大于前面的方法,攻击效果自然也会越好.

Table 4: Accuracy of CIFAR-10 dataset under various attack methods of normal training

表 4:正常训练 CIFAR-10 数据集中在各种攻击方法下的识别准确率

Attacks	Param	CE	GCE	Ours
FGSM	$\epsilon = 8/255$	31.47%	35.94%	**56.44%**
BIM	$\epsilon = 8/255$	0.00%	1.40%	**51.10%**
MIM	$\epsilon = 8/255$	0.00%	0.60%	**41.11%**
PGD	$\epsilon = 8/255$	0.00%	0.44%	**29.30%**
JSMA	$\gamma = 0.07$	0.12%	8.20%	**21.15%**
C&W	$c = 0$	0.00%	**1.23%**	0.62%

我们的方法在 CIFAR-10 数据集中表现要更加优秀,如表 4 所示.在大部分攻击下与 MNIST 数据集中一致,但在这个数据集下,无论是L_0,L_2或是L_∞的限制的方法,我们的方法都能有效地提高模型的对抗鲁棒性.且在大部分情况下,都要远远优于 GCE 训练的模型.在 FGSM 方法下,GCE 所得到的模型也仅仅只能提高一点的准确率,但我们的方法提升了 25%的识别准确率.BIM、MIM 以及 PGD 方法下,普通训练完全无法抵抗该攻击,GCE 也仅仅只能提高一点点的准确率,甚至可以说无法提升,但我们的方法依然能大大地提高识别准确率,即使是在 PGD 方法下,也能提高 20%的识别准确率.而且,在 JSMA 下,与 MNIST 数据集不同,我们的方法要比 GCE 方法鲁棒性要更强.但是在 C&W 攻击下,与 MNIST 数据集结果一致,我们的方法还是没有 GCE 得到的模型要好.

Table 5: Accuracy of CIFAR-100 dataset under various attack methods of normal training

表 5:正常训练 CIFAR-100 数据集中在各种攻击方法下的识别准确率

Attacks	Param	CE	GCE	Ours
FGSM	$\epsilon = 8/255$	8.92%	20.57%	**28.52%**
BIM	$\epsilon = 8/255$	0.11%	18.60%	**32.33%**
MIM	$\epsilon = 8/255$	0.04%	18.58%	**32.29%**
PGD	$\epsilon = 8/255$	0.00%	12.38%	**19.86%**
C&W	$c = 0$	0.00%	0.42%	**1.05%**

在 CIFAR-100 上样本类别数变多,防御也变得更加困难,在大多数情况下我们的方法都能很好地提高模型的对抗鲁棒性.但由于攻击时间过长,CIFAR-100 中并没有加入 JSMA 攻击的实验结果.从表 5 中已有的数据可以仍然能得到以下结论:对于更复杂的数据集,我们的方法能保持有效性.

4.5.2 对抗训练

我们在第 1.2 节中介绍了对抗训练,对抗训练的初衷是产生一些对抗样本,并作为训练样本重新训练模型,希望模型学到对抗样本的特征并对这些样本正确分类,同时获得对其他对抗样本的对抗鲁棒性.目前已经提出了很多对抗训练的框架[6,15,17],这里我们选择的是文献[6]提出的攻击方法 PGD,并用该方法产生对抗样本用于对抗训练,因为 PGD 方法是 first-order 攻击中最强的一种攻击.我们将我们所提出的方法与用 PGD 进行对抗训练的框架结合,显示出我们的方法可以比正常对抗训练得到更好的结果.

这里对于我们方法中(公式 (6))的正则项系数α,在 MNIST 的对抗训练中,我们设为了 0.8.在 CIFAR-10 中,与 4.3 中提到的不同,我们设置为 1.

对于训练中的其他细节,我们使用的模型以及参数如 4.3 节中所提到的.对抗训练过程与文献[6]中的训练过程,我们加载了 4.5.1 节中的训练的模型,并在这些模型的基础上,产生对抗样本并重新训练模型.对于 MNIST 数据集:我们继续训练了 10 周期,并将学习率设为 0.01,动量设为 0.9.PGD 攻击方法中,我们把迭代次数设为 40,迭代步伐设为 0.1,对于总的扰动限制设为 0.3.对于 CIFAR-10 数据集:我们训练了 40 个周期,学习率设为 0.001,动量设为 0.9 并将 weight decay 设为 0.0005.对于 CIFAR-10 的 PGD 攻击,我们把迭代次数设为了 7,迭代步伐设为 2/255,总的扰动限制设为 8/255.实验结果显示,我们的方法结合对抗训练能得到更强的对抗鲁棒性,如表 6 所示.

Table 6: Accuracy of MNIST and CIFAR-10 under various attack methods of adversarial training

表 6:对抗训练 MNIST 和 CIFAR-10 数据集在各种攻击方法下的识别准确率

	MNIST			CIFAR-10		
Attacks	Param	CE	Ours	Param	CE	Ours
FGSM	$\epsilon = 0.3$	92.94%	**93.93%**	$\epsilon = 8/255$	50.27%	**52.00%**
BIM	$\epsilon = 0.3$	89.50%	**90.50%**	$\epsilon = 8/255$	46.02%	**47.37%**
MIM	$\epsilon = 0.3$	87.42%	**88.89%**	$\epsilon = 8/255$	40.93%	**41.82%**
PGD	$\epsilon = 0.3$	89.52%	**90.08%**	$\epsilon = 8/255$	39.80%	**40.55%**
JSMA	$\gamma = 0.25$	25.44%	**63.24%**	$\gamma = 0.07$	81.27%	**83.69%**
C&W	$c = 0$	9.25%	**40.72%**	$c = 0$	0.02%	**0.80%**

从表 6 中可以看出,在对抗训练中虽然我们的方法不如正常训练中有那么大的提升,但相对于用交叉熵进行的对抗训练,我们的方法结合对抗训练还是能得到更好的对抗鲁棒性.特别是对于 JSMA 以及 C&W 两个方法,在正常训练中我们的方法虽然能提高一点鲁棒性,但还是无法抵抗这么强的攻击,这里 MNIST 数据集中,与正常训练不同,我们的方法在对抗训练下对 JSMA 以及 C&W 这两种攻击鲁棒性大大提高了,且要远优于用交叉熵对抗训练的模型.在 CIFAR-10 数据集中也是如此,在所有的攻击下,我们的方法都能得到更好的鲁棒性.但是在 CIFAR-10 数据集,即使是进行了对抗训练,模型还是无法抵抗 C&W 攻击,这里我们认为是 C&W 攻击太强了,CIFAR-10 的特征维度要比 MNIST 更多,因此要防御会更难,从其他攻击方法的识别准确率也可以看出,CIFAR-10 对于对抗样本的准确率总是会低于 MNIST 数据集.此外,对抗训练对于我们的方法提升并不大,但对抗训练大大提高了模型在 JSMA 以及 PGD 攻击方法下的对抗鲁棒性.

5 总结

本文中,我们研究了模型预测不确定性与对抗鲁棒性的关系,我们认为,传统的训练方法虽然对输出置信度很高,但得到的分类边界并不完美,容易受到对抗样本的攻击.而如果在训练模型的同时提高模型输出的信息熵,使得模型预测不确定性变大,则可以使得模型的分类面更加平衡,使得模型分类面边界与每一类数据的距离尽可能远,这样在对样本进行攻击时,不会因分类面狭隘导致某类样本极容易受到对抗样本攻击.基于上面的结论,文中提出了一个新的提高模型鲁棒性的方法.本文通过在 MNIST,CIFAR-10 和 CIFAR-100 数据集上的大量实验和简化的模型推导都证实了对抗鲁棒性随模型预测不确定性的增加而增加的统计关系,同时验证了本文提出的方法的有效性.最后,本文的方法也可结合对抗训练,进一步提高模型对抗鲁棒性.

References:

[1] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses. In: 2018 International Conference on Learning Representations (ICLR). 2018.

[2] Ding G W, Wang L, Jin X. AdverTorch v0. 1: An adversarial robustness toolbox based on pytorch. arXiv:1902.07623, 2019.

[3] Chen H Y, Liang J H, Chang S C, et al. Improving adversarial robustness via guided complement entropy. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2019: 4881-4889. [doi: 10.1109/ICCV.2019.00498]

[4] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: 2015 International Conference on Learning Representations (ICLR). 2015.

[5] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of the 2017 IEEE Symp. on Security and Privacy (SP). 2017. [doi: 10.1109/SP.2017.49]

[6] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083, 2017.

[7] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. In: 2017 International Conference on Learning Representations (ICLR). 2017.

[8] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2018: 9185-9193.

[9] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2016: 372-387.

[10] Vivek B S, Babu R V. Regularizers for Single-step Adversarial Training. arXiv:2002.00614, 2020.

[11] Ma A, Faghri F, Farahmand A. Adversarial Robustness through Regularization: A Second-Order Approach. arXiv preprint arXiv:2004.01832, 2020.

[12] Jin C, Rinard M. Manifold regularization for locally stable deep neural networks. arXiv:2003.04286, 2020.

[13] Cheng M, Lei Q, Chen P Y, et al. Cat: Customized adversarial training for improved robustness. arXiv:2002.06789, 2020.

[14] Shannon C E. A mathematical theory of communication. The Bell system technical journal, 1948, 27(3): 379-423.

[15] Liu G, Khalil I, Khreishah A. Using Single-Step Adversarial Training to Defend Iterative Adversarial Examples. arXiv:2002.09632, 2020.

[16] Xie C, Yuille A. Intriguing properties of adversarial training at scale. In: 2020 International Conference on Learning Representations (ICLR). 2020.

[17] Zhang H, Yu Y, Jiao J, et al. Theoretically Principled Trade-off between Robustness and Accuracy. In: International Conference on Machine Learning (ICML). 2019: 7472-7482.

[18] Shafahi A, Najibi M, Ghiasi M A, et al. Adversarial training for free!. In: Advances in Neural Information Processing Systems (NIPS). 2019: 3358-3369.

[19] Wong E, Rice L, Kolter J Z. Fast is better than free: Revisiting adversarial training. In: International Conference on Learning Representations (ICLR). 2020.

[20] Seedat N, Kanan C. Towards calibrated and scalable uncertainty representations for neural networks. arXiv preprint arXiv:1911.00104, 2019.

[21] Wang X, He Y. Learning from uncertainty for big data: future analytical challenges and strategies. IEEE Systems, Man, and Cybernetics Magazine, 2016, 2(2): 26-31.

[22] Yuan X, He P, Zhu Q, et al. Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems, 2019, 30(9): 2805-2824.

[23] Duan GH, Ma CG, Song L, et al. Research on structure and defense of adversarial example in deep learning. Chinese Journal of Network and Information Security, 2020, 6(2): 1-11

[24] Zhang WX. Research on the generation of adversarial example based on batch gradient. Huazhong University of Science & Technology, 2019. [doi: 10.27157/d.cnki.ghzku.2019.001227]

[25] Wang SY, Jin H, Sun JZ. A Method for Image Adversarial Samples Generating Based on GAN. Journal of Frontiers of Computer Science and Technology

[26] Fan CL, Su T, Teng YP, et al. Black box attack optimization algorithm based on differential evolution. Application Research of Computers. [doi: 10.19734/j.issn.1001-3695.2019.10.0623]

[27] He ZB, Huang XL. Adversarial Attacks and Defenses Against Neural Networks. Aero Weaponry,2020, 27(3): 1－19.

[28] Pan WW, Wang XY, Song ML, Chen C. Survey on generating adversarial examples. Ruan Jian Xue Bao. Journal of Software, 2020, 31(1): 67-81 (in Chinese) [doi: 10.13328/j.cnki.jos.005884]

[29] Gal Y. Uncertainty in deep learning. University of Cambridge, 2016, 1(3).

[30] L. Deng, The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web], in IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 141-142, Nov. 2012. [doi: 10.1109/MSP.2012.2211477]

[31] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. 2009.

[32] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[33] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2016: 770-778.

附中文参考文献:

[23]段广晗,马春光,宋蕾,武朋.深度学习中对抗样本的构造及防御研究.网络与信息安全学报,2020,6(02):1-11.

[24]张文翔. 基于批量梯度的对抗样本生成方法的研究.华中科技大学,2019. [doi: 10.27157/d.cnki.ghzku.2019.001227]

[25]王曙燕,金航,孙家泽.GAN图像对抗样本生成方法.计算机科学与探索:1-12[2020-07-25]. http://kns.cnki.net/kcms/detail/11.5602.TP.20200722.1433.004.html.

[26]范纯龙,宿彤,滕一平,王翼新,丁国辉.基于差分进化的黑盒攻击优化算法.计算机应用研究:1-5[2020-07-25]. https://doi.org/10.19734/j.issn.1001-3695.2019.10.0623. [doi: 10.19734/j.issn.1001-3695.2019.10.0623]

[27] 何正保,黄晓霖.针对神经网络的对抗攻击及其防御.航空兵器,2020,27(03):11-19.

[28] 潘文雯,王新宇,宋明黎,陈纯.对抗样本生成技术综述.软件学报,2020,31(01):67-81. [doi: 10.13328/j.cnki.jos.005884]