编号_____

南京航空航天大学

毕业论文

题　目　结合度量学习的鲁棒神经网络研究和实现

学生姓名　陶略

学　　号　031510319

学　　院　计算机科学与技术学院

专　　业　计算机科学与技术

班　　级　1615102

指导教师　陈松灿

二〇一九年五月

南京航空航天大学
本科毕业论文诚信承诺书

本人郑重声明：所呈交的毕业论文（题目：结合度量学习的鲁棒神经网络研究和实现）是本人在导师的指导下独立进行研究所取得的成果。尽本人所知，除了毕业论文中特别加以标注引用的内容外，本毕业论文不包含任何其他个人或集体已经发表或撰写的成果作品。

作者签名：　　年　月　日

（学号）：

结合度量学习的鲁棒神经网络研究和实现

摘 要

由于在自然图像、自然语言等数据中普遍存在的层次性和非线性性，深度神经网络在诸多现实任务上都取得了巨大的进展。从已大规模部署的人脸识别系统，到正在发展阶段的自动驾驶汽车，深度学习技术正逐渐走进并将改变人类生活的各方面。虽然深度学习在通常情况下都具有较好的泛化性能，但近年来，神经网络被发现在对抗样本的攻击下，表现出了非常糟糕的性能，而这些对抗样本对人类来说与正常样本无异。这引发了人们对基于深度网络等智能系统的安全性担忧，同时也揭示了目前深度网络对周围世界的理解能力远远不及人类智能的事实。对抗训练作为一个有效的提高神经网络鲁棒性和安全性的方法已开始倍受广泛关注，但其泛化性能却难尽人意。本文工作旨在尝试结合度量学习思想，利用相关的正则化技术，以提高对抗训练的泛化性能，并通过大量的实验获得了有效性验证。

关键词：深度学习，神经网络，对抗样本，鲁棒学习，度量学习，正则化

Research and Implementation of Robust Neural Network Based on Metric Learning

Abstract

Because of the hierarchy and non-linearity in the natural image and natural language data, deep learning has made great progress in many practical tasks. From the face recognition system that has been deployed on a large scale, to the self-driving car in the developing stage, deep learning is stepping into and changing all aspects of human life. Although deep learning usually has good generalization performance in most cases, in recent years, neural networks have been found to exhibit very poor performance against adversarial examples, which are nearly indistinguishable from the clean examples. These raise concerns about the security of intelligent systems such as deep network, and also reveal the fact that deep networks are far less capable of understanding the surrounding world than human intelligence. As an effective method to improve the robustness and security of neural networks, adversarial training has attracted much attention, but its generalization performance is unsatisfactory. The purpose of this paper is to try to improve the generalization performance of adversarial training by combining the idea of metric learning and using relevant regularization techniques. We acquire the effectiveness of regularizations by extensive experiments.

Key Words: Deep learning, Neural network, Adversarial example, Robust learning, Metric learning, Regularization

目录

第一章 引言

深度学习（Deep Learning）已在诸多机器学习任务上取得了巨大进展：物体识别[1]、目标检测[2]、语音识别[3]、语言翻译[4]、社交网络分类[5]等。在大数据和硬件加速的驱动下，未经加工的特征能被深度网络用多层次方式和抽象的特征表示出来，减轻了传统机器学习方法中对手工设计的特征和专家知识的依赖。

深度学习技术正逐渐走进并将改变人类生活的各方面，世界上越来越多的应用和系统被其所赋能。例如，Google、Telsa、Uber 等公司正在测试的自动驾驶汽车，它需要用到物体识别、目标检测、强化学习等大量现代深度学习驱动的技术。作为一种生物认证的方式，人脸识别系统被应用到移动设备的解锁、自助取款机、火车站安检机等处，甚至被执法部门用来追捕嫌疑犯。一些基于行为的恶意软件检测和异常检测解决方案也利用深度学习来寻找语义特征[6]。

深度神经网络（Deep Neural Network, DNN）具有强大的拟合能力，并在许多现实任务中表现出较好的泛化性能。即使模型的参数数远远大于训练样本数（即过参数化），也没有发生严重的过拟合问题；即使模型优化时的高度非凸、非光滑的损失函数并不能被保证找到全局最优解，也能在工程应用中相对放心地的使用局部最优解；即使模型的可解释性差，也没有妨碍学术界和工业界将它视为一个黑盒广泛研究和应用。

然而，近年来，通过反向传播算法优化的深度网络被发现具有一个反直觉的特性和内在的盲点：即使一个在正常场景下表现良好的深度模型，也非常容易被加了一些对抗扰动的样本所欺骗，而这些样本对人类的感知系统来说却与正常样本无异[7]，如图 1.1所示。Szegedy 等人首先在图片上加入一些微小的扰动，使得最先进的做图片分类任务的深度神经网络能够以较大的概率被欺骗，即样本被错误分类。这些被错分类的不怀好意的样本被称为对抗样本（Adversarial examples）[8]。大量的基于深度学习技术的系统已经在现实世界被应用，或正有待去应用。另一方面对抗样本也能被应用于现实世界中。例如，一个敌人通过对公路上的停车标志进行一些人类察觉不到的微小改动，便可以迷惑自动驾驶汽车的交通信号识别系统，使其无视停车信号标志[9,10]；也可以通过类似的方式构

造对抗样本，使得目标检测系统无法检测出路上的行人[11]。攻击者也可以生成对抗指令去欺骗被应用到 Apple Siri、Amazon Alexa、Microsoft Cortana 上面的语音识别模型[12,13]。

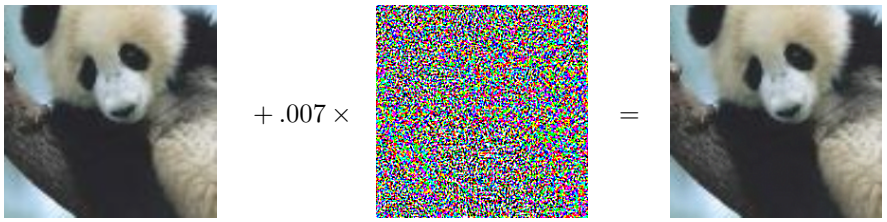

图 1.1　对抗样本的一个演示，干净的样本能够被 GoogLeNet[14] 正确识别成熊猫（Panda），而添加了不可察觉的扰动的对抗样本却被识别成了长臂猿（Gibbon）。左边：干净样本；中间：扰动；右边：对抗样本。

在对抗样本的攻击下，深度网络表现出了相当不理想的鲁棒性，使得应用深度学习的系统都不可避免地存在着潜在的风险，尤其在安全性攸关的任务上。另一方面，深度网络如此反直觉的特点，让我们不得不反思：那些号称在某项任务上深度学习模型超越人类水平的标语真的有意义吗？那些深度网络的智能水平真的比得上人类智能吗？从对抗鲁棒性的语境下看，答案当然是否定的。对抗鲁棒性问题如此具有现实意义，同时又带有一定的哲学内涵，吸引了学者们前赴后继地进行研究，大量的防御方法被提出，目的是为了能够让智能系统正确地认知对抗样本[8,15–19]。然而，这些已提出的防御措施大部分都不够有效，它们可以被更强大的敌人成功攻击[20]。同时也有一些工作旨在验证和训练可证明鲁棒的神经网络，但是这些方法只能提供单点的保证，而且它们需要巨大的计算代价，因此无法被推广到更为现实的数据集中[21–23]。另外，基于检测的防御方法也被大量提出[24–28]，然而，Carlini 等人的研究[29] 表明这些防御措施都可以被使用具有针对性的攻击机制的敌人轻易地规避。

许多的防御方法被提出不久即被更强的攻击方法击败，唯有对抗训练（Adversarial training）这个简单且有效的方法经受住了时间的考验。对抗训练的过程是将对抗样本作为训练集的一部分，让模型去学习，目的是使深度神经网络对未来（即测试时）样本的攻击具有更好的鲁棒性。Madry 等人[16] 从鲁棒优化（Robust optimization）的角度解释了对抗训练，表明对抗训练的过程是在每个训练样本的周围空间里寻找最坏的样本用以训

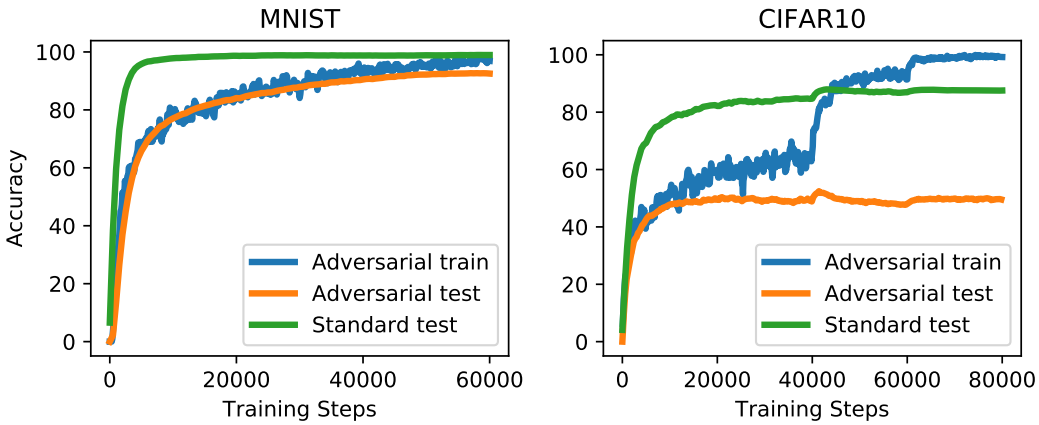

图 1.2　在 MNIST 和 CIFAR10 上进行鲁棒优化（即对抗训练）的分类正确率

练最优的分类器，并建议在对抗训练时使用一阶敌手（First-order adversary）作为自然且普适的安全性保证。

然而对抗训练本身的性能并不够理想。虽然通过合适的模型和适当的优化过程，对抗训练所获模型可以将训练集样本较好地拟合，但是在测试集上的泛化性能却很差[30]。如图 1.2所示，在 MNIST 上，对抗测试集精度接近对抗训练集精度（但还是有一些差距）；在 CIFAR10 上，虽然模型在标准（非对抗）的测试集上有较高的精度，但在鲁棒精度上有一个非常大的鸿沟，即对抗测试集上的精度远低于对抗训练集上的精度。

对抗训练泛化能力的不足，暗示着我们应该为对抗训练设计更为有效的正则化方法，尤其是在深度神经网络过参数化的情况下。本文的主要任务，就是尝试结合度量学习的思想，利用相关的正则化方法，以提高对抗训练的泛化性能，并通过大量的实验进行有效性验证。

第二章 国内外研究现状

本章介绍与本文相关的三个主题：深度学习、度量学习、对抗样本问题。本文研究的是如何结合度量学习的思想以缓和深度学习中的对抗样本问题，我们针对的主要是比较基础的图像分类任务，其它任务中的对抗样本问题不在本文讨论范畴，但值得进一步推广和探索。

2.1 深度学习

2006 年，多伦多大学教授 Hinton 等人发表在《Science》上的一篇论文引发了深度学习在学术界和工业界的发展热潮[31]。这篇文献抛出了两个结论：（1）、深度神经网络具备优异的特征表示能力，深度模型学习到的特征比原始数据维度更低，且不失判别性，非常有利于分类等任务；（2）、一层一层训练的方式可以有效地训练一个深度模型。

深度学习发展到今天，已经在许多领域大放异彩，例如：计算机视觉、自然语言处理、医学影像分析、语音、游戏、图网络、时间序列、机器人控制、计算机编程等。本节主要介绍深度学习在图像分类（Image classification）和目标检测 (Object detection) 上的进展。

2.1.1 图像分类

早在 1989 年，受到生物学家 Hubel 和 Wiesel 的动物视觉模型[32] 的启发，Yann LeCun 等人就提出了一个多层的卷积神经网络（Convolutional Neural Network，CNN）[33]。但由于优化方法、数据量和计算能力等限制，直到 2012 年，Hinton 等人才采用更深的卷积神经网络模型 AlexNet 在著名的 ImageNet 问题上取得当时世界上最好的成果，使得对于图像识别的研究工作前进了一大步[1]。

2014 年，更深的深度网络 VGGNet[34] 和 GoogLeNet[14] 被成功训练，分别为 16 层和 22 层。2015 年，He 等人提出 ResNet[35]，利用残差单元（Residual unit）有效地缓和了梯度消失问题，成功训练了一个 152 层的深度神经网络，进一步提高了在 ImageNet 上的分类性能。2017 年，注意力机制被引入深度网络架构中，Hu 等人提出 SENet[36]，利用通道间的自注意力机制，赢得了最后一届 ImageNet 竞赛的冠军。随后 Woo 等人提出 CBAM

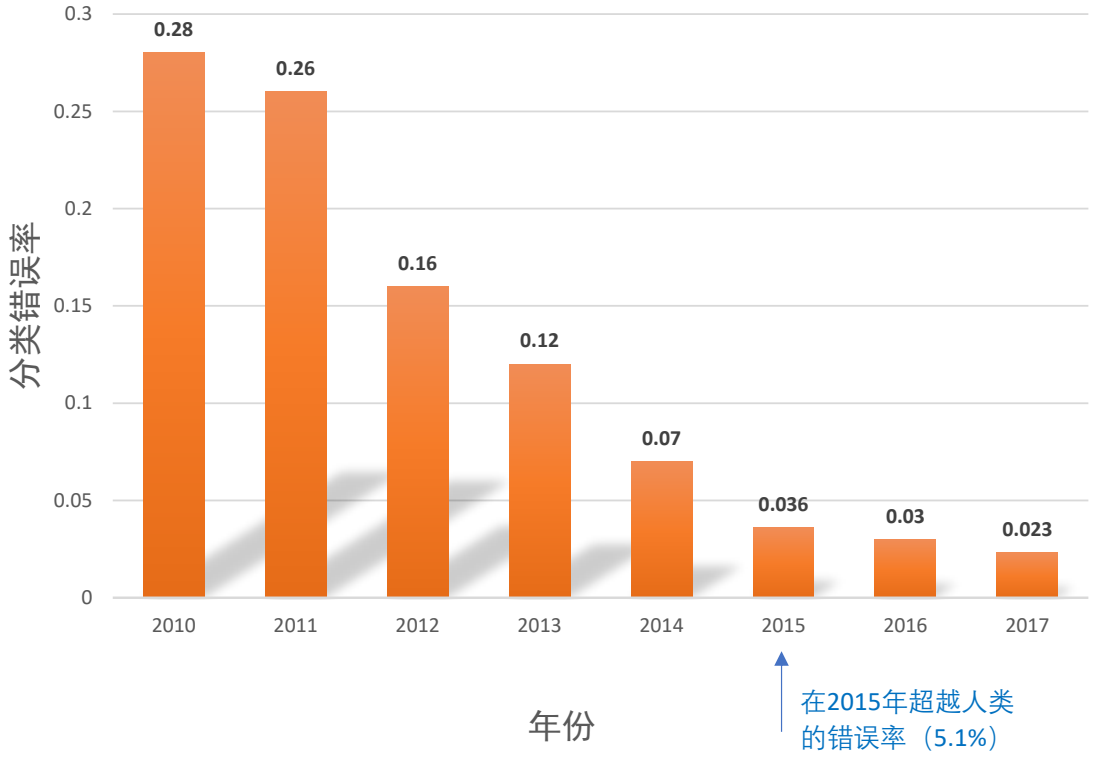

图 2.1　2010 ~ 2017 年 ImageNet 分类任务错误率

模块[37]，结合通道间和空间的自注意力，进一步提升了性能。图 2.1展示了 2010 ~ 2017 年 ImageNet 分类错误率变化情况，从错误率的指标上看，深度网络已经超越人类。

2.1.2　目标检测

自从 AlexNet 获得 ILSVRC 2012 挑战赛冠军后，用 CNN 进行特征提取便流行起来。2014 年，Girshick 等人提出 R-CNN[38]，利用候选区域方法（Region proposal method）生成检测感兴趣的区域（Resion of Interest, ROI），再送入 CNN 进行分类，此方法每预测一张图片需要 49 秒。2015 年，Ross Girshick 提出 Fast R-CNN[39]，利用特征图代替原图生成 ROI，显著地减少处理时间至每张图片 2.3 秒。2016 年，Ren 等人提出 Faster R-CNN[40]，使用可微的候选区域网络代替了不可微分的候选区域方法，使得 ROI 的生成效率进一步提高，每张图片的处理时间减小到 0.2 秒。

以上的目标检测器由于候选区提取和分类是分开做的，被称为二阶段检测器（Two-stage detector），另一种更直接的方式被称为一阶段检测器（One-stage detector）。Liu 等人提出的 SSD 模型[41] 和 Redmon 等人提出的 YOLO 模型[2,42,43] 将候选区提取和分类用

端到端（End-to-end）的方式融合到一起训练，使得目标检测器真正达到了实时检测的程度，但对小目标的检测效果较差。Lin 等人提出特征金字塔网络（Feature Pyramid Network, FPN）[44]，利用多尺度特征图提高准确率。

2.2 度量学习

过去的十几年里，大量的度量学习 (Metric Learning) 算法被提出, 其中最典型的是对比损失（Contrastive loss）[45,46] 和三元组损失（Triplet loss）[47,48]。它们的目标通常是学习一个良好的距离度量，使得正样本对（Positive pair）之间的距离尽量小，负样本对（Negative pair）之间的距离尽量大。这些度量学习方法被提出时，只能够学习一个线性变换，无法很好地刻画现实数据中的非线性流形结构。为了解决这个问题，核技巧通常被采用，先将样本映射到一个高维的特征空间中，然后再在这个高维空间里学习一个距离度量[49,50]。然而，这些方法无法显式地得到一个非线性映射函数，因此无法被应用到更大规模的数据上。

自深度学习表现出强大的表示学习能力之后，深度神经网络便被引入到度量学习之中，以学习一个非线性映射。Hu 等人将深度网络应用到对比损失中[51]，Hoffer 等人将深度网络应用到三元组损失中[52]。然而，将深度网络直接地应用到度量学习中，带来的问题是：这些方法随机采样样本对或三元组以构成训练批次，无法在小批次随机梯度下降的训练过程中充分利用一批次里的所有训练样本。Kihyuk 和 Oh Song 等人分别提出一种新颖的采样方式[53,54]，能够充分利用小批次里的所有样本，提升了深度度量学习 (Deep Metric Learning) 的性能。

2.3 对抗样本问题

早在十几年前，对抗样本问题就在传统机器学习中被讨论过。2004 年，Dalvi 等人首次讨论了对抗样本，把这个问题看做敌人和分类器之间的一个博弈问题[55]，在对抗样本上的攻击和防御变成了一场迭代式的博弈。通过增加字符以避免检测，对抗攻击曾被应用到垃圾邮件过滤系统中[55,56]。2013 年，Biggio 等人首先提出一个基于梯度的方法生成对抗样本以攻击浅层的分类器，比如支持向量机、一个两层的神经网络[57]。与深度学习中的对抗样本相比，传统机器学习中的方法在修改数据上拥有更大的自由度，第一个被用来评估攻击方法的数据集是 MNIST，但此时生成的对抗样本能被人类轻易地分辨出来。

在传统机器学习中，攻击和防御方法都在特征提取过程中下了很大功夫，甚至可以影响数据采集过程，在对人类的影响上并没有太多的考虑[58]；而在深度学习中，只需要关心原始数据的输入，且非常强调对人类视觉系统的影响。

本文重点研究图像分类任务中的对抗样本：使用一个第三方发布的预训练的分类器，用户输入一张图片以得到分类器的预测结果。对抗样本是在干净的原始样本上添加一些微小的扰动，这些扰动通常无法被人类感知到。然而，这样的扰动可以误导图片分类器，用户将得到一个不正确的预测结果。记 f 为一个训练好的深度学习模型，x 为一个原始的输入样本数据，一个对抗样本 x' 的生成过程可以被描述为一个带约束的优化问题：

$$\begin{aligned} \min_{x'} \quad & \|x' - x\| \\ s.t. \quad & f(x') = l', \\ & f(x) = l, \\ & l \neq l', \\ & x' \in [0,1], \end{aligned} \tag{2.1}$$

其中 l 和 l' 表示 x 和 x' 的输出标记，$\|\cdot\|$ 表示两个样本点之间的距离，$\eta = x' - x$ 表示添加在 x 上的扰动。这个优化问题在输入数据有范围限制的情况下，误导模型预测结果的同时最小化扰动。

2.3.1 攻击方法

2014 年，Szegedy 等人首先提出了针对深度神经网络的对抗样本，提出了一种名为 L-BFGS 的攻击方法。L-BFGS 方法使用代价较高的线性搜索方法去寻找参数的最优值，这非常耗时且实用性不强，因此 Goodfellow 等人提出了一个快速的生成对抗样本的方法，称作快速梯度符号方法（Fast Gradient Sign Method, FGSM）, 此方法仅仅只需要沿着每个像素的梯度符号方向执行一步梯度更新，便可生成如图 1.1所示的对抗样本[7]。2016 年，Kurakin 等人为了将对抗样本应用到物理世界中，他们扩展了 FGSM 方法，使用更细化的优化过程和多次的迭代以生成更具有迁移能力的对抗样本，称作基本迭代方法（Basic Iterative Method, BIM），成功欺骗了一个运行在手机上的 ImageNet 分类器[9]。Moosavi-Dezfooli 等人提出 DeepFool 方法以寻找与原始样本距离最小的对抗样本[59]。2017 年，Carlini 和 Wagner 提出了一个非常强大的攻击方法 C&W's Attack，使得许多防御方法失效[20]。Chen 等人提出一种基于零阶优化（Zeroth Order Optimization, ZOO）的攻击方法，

这个方法不需要梯度信息，所以可以被直接应用到黑盒攻击上，无需迁移[60]。Moosavi-Dezfooli 等人基于 DeepFool 提出了一种普适性的对抗攻击（Universal adversarial attack），能够找到攻击尽量多样本的一个普适性的扰动。为了让人类无法察觉对抗扰动，Su 等人提出单像素攻击（One pixel attack），利用进化算法寻找一个像素点进行扰动[61]。

在视觉应用方面。2016 年，Sharif 等人设计了一个眼镜形状的对抗扰动，以攻击基于深度神经网络的人脸识别系统[62]。2017 年，Xie 等人提出密集敌人生成（Dense Adversary Generation, DAG）算法，能够生成针对目标检测和语义分割系统的对抗样本[11]。2019 年，腾讯科恩实验室对特斯拉 Autopilot 自动驾驶系统的安全性进行研究，成功攻击了 Autopilot 系统的外部天机状况识别系统和道路交通标线识别系统，分别使得车辆雨刷误启动和驶入反向车道[63]。

在自然语言处理应用方面。2017 年，针对阅读理解系统，为了产生与正确答案一致且不混淆人的对抗性例子，Jia 等人在段落末尾增加了对抗性字段[64]。2019 年，针对语音识别系统，Qin 等人提出了一种有效的人类不可察觉的音频对抗样本生成方法，此方法充分利用了听觉掩模的心理声学原理，对任意的全语句目标都达到了百分之百的攻击成功率，进一步，他们发现在物理世界中通过空气传播此对抗音频后，攻击仍然有效[65]。

2.3.2 防御方法

2016 年，Papernot 等人提出使用网络蒸馏（Network distillation）作为一个防御对抗样本的手段[17]，但随后被 Carlini 等人提出的 C&W's Attack 成功击破。2017 年，许多检测方法被用来检测对抗样本。有的工作提出使用一个基于深度网络的二元分类器用于检测并过滤对抗样本[25,26,66]，有的工作提出添加一个离群类到原始的深度学习模型中以检测并过滤对抗样本[24]，还有的工作发现对抗样本的分布与干净样本的不同，提出用基于 PixelCNN[67] 排序的 p 值进行对抗样本的检测和拒绝[68]。然而，Carlini 和 Wagner 总结了以上对抗性的检测方法，展示了这些检测防御方法都可以被更具有针对性的攻击轻易规避[29,69]。2018 年，许多梯度掩码（Gradient masking）的方法被提出用来增强深度网络的鲁棒性。Jacob 等人提出用温度计编码（Thermometer encoding）来提高鲁棒性[70]；Guo 等人提出将输入的图片进行五个预变换，以消除对抗扰动，提升模型鲁棒性[71]；Dhillon 等人提出随机激活剪枝（Stochastic Activation Pruning, SAP），随机抛弃部分神经元以消除对抗扰动的影响[72]；Xie 等人提出在输入样本被送入模型前对样本进行随机变换，以

消除对抗扰动的影响[73]；Song 等人提出使用 PixelCNN 模型将潜在的对抗样本投影到真实的数据流形上以消除对抗扰动[74]；Samangouei 等人提出使用生成对抗网络将样本投影到生成器的数据流形中以消除对抗扰动[75]；然而，以上所有梯度掩码的方法都被 Athalye 等人使用具有针对性的攻击方法一一破解[76]。

同时，Athalye 等人也指出，目前只有对抗训练这一种防御方式无法被完全击破，但对抗训练也有自身的问题[76]。虽然对抗训练在 MNIST 上的鲁棒性比较令人满意[16]，但在 ImageNet 这样的数据集上的鲁棒性就差强人意了[77]。而且对抗训练在 CIFAR10 和 ImageNet 上的泛化性能比较差, 这个泛化性能包括在测试时对同一种攻击方式（与训练时使用的攻击方式相同）的鲁棒性和测试时对其它攻击方式（与训练时使用的攻击方式不同）的鲁棒性，通常认为对抗训练的样本复杂度远高于自然训练的样本复杂度[30]。

为了提高对抗训练的泛化能力，Kannan 等人提出对抗逻辑配对（Adversarial Logit Pairing），提高了对抗训练在 ImageNet 上的泛化能力[78]，并被 NIPS 2018 接收，但随后被人指出并实验上证明了该方法并没有真正地提在 ImageNet 上的泛化能力[79]，于是 ALP 又被作者从 NIPS 上撤稿。本文也对 ALP 的有效性进行了一些探究，详见章节4.2。Song 等人提出使用将干净样本分类和对抗样本分类看做一个域适应的问题，采用最大均值散度（Maximum Mean Discrepancy, MMD）和相关性匹配（CORrelation ALignment, CORAL）作为正则化项提高对抗训练的泛化能力[80]。Farnia 等人提出使用谱归一化来提高对抗训练的泛化能力[81]。

第三章 对抗训练及其正则化方法

本文专注于对基于深度学习的模型的攻击和防御，我们假设对手仅仅能够在模型的测试阶段实施攻击，对手只能在模型训练之后篡改模型的输入数据。已训练的模型和训练集数据均不可被篡改，但对手拥有已训练的模型的所有知识（模型架构和参数），这是符合目前大部分在线机器学习服务系统现状的假设。本节先介绍与对抗训练相关的几个攻击方法，再讨论对抗训练的泛化能力，最后介绍并提出相关的正则化方法。

3.1 攻击方法

本小节介绍三种攻击方法：快速梯度符号方法 (Fast Gradient Sign Method, FGSM)[7]、投影梯度下降（Projected Gradient Descent, PGD）[16]、Carlini 和 Wagner 的攻击（C&W's Attack）[20]。

记干净样本集合为 $\mathcal{D}$，对抗样本集合为 $\mathcal{A}$。我们有一个基于神经网络的分类器 $f(x): \mathbb{R}^d \to \mathbb{R}^k$，对每一个输入 $x \in [0,1]^d$，$f(x)$ 输出对应的概率分布，k 表示分类任务的类别总数。令 φ 表示从输入层到逻辑层（在最后一个 softmax 函数之前的一层）的映射，于是 $f(x) = softmax(\varphi(x))$。记 ϵ 为扰动的幅度，记 x^{adv} 为对抗样本，它是由原始样本 x 受到扰动而得到的。图片分类的损失函数记为 $J(x, y)$。

3.1.1 FGSM

Goodfellow 等人介绍了用 FGSM 生成对抗样本，通过在梯度的方向上添加扰动。

$$x^{adv} = x + \epsilon \cdot \operatorname{sign}\left(\nabla_x J\left(x, y_{true}\right)\right) \tag{3.1}$$

与其它方法比起来，FGSM 是一个简单、快速且有效的敌人，因此 FGSM 非常适合用于对抗训练。

3.1.2 PGD

PGD 是 Madry 等人介绍的一种攻击方式，是 FGSM 的迭代变种。此方法迭代地执行 k 次 FGSM，每次的步长为 α。

$$\begin{aligned} x^{adv_0} &= x \\ x^{adv_{t+1}} &= x^{adv_t} + \alpha \cdot \operatorname{sign}\left(\nabla_x J\left(x^{adv_t}, y_{true}\right)\right. \\ x^{adv_{t+1}} &= \mathbf{clip}\left(x^{adv_{t+1}}, x^{adv_{t+1}} - \epsilon, x^{adv_{t+1}} + \epsilon\right) \\ x^{adv} &= x^{adv_k} \end{aligned} \tag{3.2}$$

这里 $\mathbf{clip}(\cdot, a, b)$ 函数的功能是将它的输入裁剪到 $[a, b]$ 范围内。在白盒攻击下，PGD 通常拥有比 FGSM 更高的攻击成功率。

3.1.3 C&W's Attack

C&W's Attack 是 Carlini 和 Wagner 提出用来攻击防御性蒸馏（Defensive distillation）[17] 的一种强大的攻击方法。对抗扰动 δ 通过以下的优化过程得到：

$$\begin{aligned} \min_{\boldsymbol{\delta} \in \mathbb{R}^n} \quad & \|\boldsymbol{\delta}\|_p + c \cdot f(\mathbf{x} + \boldsymbol{\delta}) \\ \text{s.t.} \quad & \mathbf{x} + \boldsymbol{\delta} \in [0,1]^n \end{aligned} \tag{3.3}$$

其中 $c > 0$ 是一个合适的常数，ℓ_2、ℓ_0 和 ℓ_∞ 范数都可以考虑。

3.2 对抗训练

在本文研究的分类任务中，我们将模型的泛化能力细分为两种：

1. 标准泛化能力（Standard generalization），即表示为模型在标准的测试集（不含对抗样本）上的分类准确率。
2. 鲁棒泛化能力（Robust generalization），即表示为模型在对抗的测试集（不含干净样本）上的分类准确率。

防御深度模型的一个非常直觉的方式是对抗训练，它在模型训练过程中向训练集里注入对抗样本。Goodfellow 等人首次提出向模型同时送入原始样本和用 FGSM 生成的对抗样本以增加鲁棒性[7]，目标函数为：

$$\hat{J}\left(x, y_{true}\right) = \alpha J\left(x, y_{true}\right) + (1-\alpha) J\left(x + \epsilon \operatorname{sign}\left(\nabla_x J\left(x, y_{true}\right), y_{true}\right)\right. \tag{3.4}$$

Kurakin 等人在 ImageNet 上实现了对抗训练，发现了标签泄露（Label leaking）效应，并建议在对抗训练过程中不要使用基于真实标签 y_{true} 的 FGSM 方法，而应该使用基于最

大可能标签的 FGSM 方法[77]，本文的所有实验中对抗训练过程便采用了避免标签泄露的 FGSM 方法。其它的一些方法在对抗训练过程中使用 PGD 方法或更加复杂的优化方法产生最坏情况的对抗样本，然而这样的方式时间复杂度过高，导致推广到更大规模的神经网络上非常困难[16,22]。

以上所有的对抗训练方法都会存在鲁棒泛化能力不佳的问题。本文考虑的是在对抗训练过程中用最简单的 FGSM 方法产生对抗样本，但使用额外的正则化项以提高鲁棒泛化能力，这样的好处是不需要在生成对抗样本的优化过程花费太多时间，更容易拓展到更大规模的数据集和模型上。

3.3 对抗训练中的正则化方法

深度学习中对抗样本的存在性是一个反直觉的特性，一个被普遍接受的解释是：对抗样本是模型的内在的盲点区域，这个盲区的结构与数据分布紧密相关，以一种不那么明显的方式[8]。这种不明显体现在，对抗样本和干净样本在我们人类看来区别并没有那么地明显，甚至不可察觉，但事实上，对于模型来说，这种不明显的区别是致命的。这种致命体现在，两个在原始特征空间中非常邻近的样本点，经过深度网络一系列非线性变换后，两个样本点在隐空间中的距离竟然会变得很远，以至于最终落在了模型的两个不同的判决区域。

一个直觉的方案是在隐空间中尽量拉近对抗样本与干净样本的距离，期望使得模型不再将两个样本点（对抗样本和原始样本）判为不同类别。这个思想就是度量学习中的对比损失，学习一个良好的深度网络，使得在隐空间中对抗样本和干净样本之间的距离尽量小。

对比损失可以形式化为：

$$L = \frac{1}{(|\mathcal{D}| + |\mathcal{A}|)^2} \sum_{x_i, x_j \in \mathcal{D} \cup \mathcal{A}} w_{ij} \left\| \varphi(x_i) - \varphi(x_j) \right\|_2^2, \tag{3.5}$$

其中 w_{ij} 定义为：

$$w_{ij} = \begin{cases} 1 & \text{if } (x_i, x_j) \in \mathcal{P}_{sim}, \\ -1 & \text{if } (x_i, x_j) \in \mathcal{P}_{dis}, \\ 0 & otherwise. \end{cases} \tag{3.6}$$

其中 $\mathcal{P}_{sim}$ 表示相似样本对集合，$\mathcal{P}_{dis}$ 表示不相似样本对集合。这里的式(3.5)在度量学习

中只是一个最基本的形式，本文这样写是为了阐明本文思路，对比损失的其它变体形式在这里不予赘述。

为了将对比损失的思想作为一个正则化项，合适地融入对抗训练的损失中，下面介绍三种构造方式：对抗逻辑配对（Adversarial Logit Pairing, ALP）[78]、结合域适应的对抗训练（Adversarial Training with Domain Adaption, ATDA）[80]、对抗流形正则化（Adversarial Manifold Regularization, AMR）。

3.3.1 ALP

ALP 可以定义为：

$$L_{alp} = \frac{1}{|\mathcal{D}|}\sum_{x_i \in \mathcal{D}} \left\|\varphi(x_i) - \varphi(x_i^{adv})\right\|_2^2, \tag{3.7}$$

其中 x_i^{adv} 是 x_i 对应的使用 FGSM 攻击方法生成的对抗样本。式(3.7)旨在缩小每个样本点和它对应的对抗样本点间的欧式距离，即在式(3.6)中，$\mathcal{P}_{sim} = \{(x_i, x_i^{adv}) | x_i \in \mathcal{D}\}$，$\mathcal{P}_{dis} = \varnothing$。

Kannan 等人在提出 ALP 的文章中，他们仅仅在 ImageNet 上进行了有效性验证，在 FGSM 攻击的对抗训练过程中增加 ALP 正则化项，发现这样训练的模型对 FGSM 敌人和 PGD 敌人的鲁棒泛化能力都得到了提高[78]，随后却被人发现通过简单的增加 PGD 攻击的迭代次数便可将 PGD 鲁棒精度降为 0.6%[79]，这篇文章因此从 NIPS 2018 撤稿。本文将在 MNIST、SVHN 等数据集上对 ALP 的有效性进行验证，并画出 PGD 鲁棒精度随迭代次数增多而下降的收敛曲线，以保证结论的准确性，详见章节4.5。

3.3.2 ADTA

把对抗样本和干净样本看成两个不同的域，结合域适应的思想，ATDA 提出用最大均值散度（Maximum Mean Discrepancy, MMD）[82] 和相关性匹配（CORrelation ALignment, CORAL）[83] 作为正则化项提高对抗训练的泛化能力[80]。这和对比损失中点到点的拉近目标是有所区别的，MMD 和 CORAL 是一种集合到集合的拉近。

MMD 的目标是最小化干净样本和对抗样本的均值向量间的距离：

$$L_{mmd} = \frac{1}{k}\left\|\frac{1}{|\mathcal{D}|}\sum_{x \in \mathcal{D}} F(x_i) - \frac{1}{|\mathcal{A}|}\sum_{x^{adv} \in \mathcal{A}} F(x_i^{adv})\right\|_1. \tag{3.8}$$

CORAL 的目标是最小化干净样本和对抗样本间的协方差矩阵间的距离：

$$L_{coral} = \frac{1}{k^2}\left\|C_{\varphi(\mathcal{D})} - C_{\varphi(\mathcal{A})}\right\|_{\ell_1}, \tag{3.9}$$

其中 $C_{\varphi(\mathcal{D})}$ 和 $C_{\varphi(\mathcal{A})}$ 分别是干净样本和对抗样本在逻辑空间（Logit sapce）上的协方差矩阵，$\|\cdot\|_{\ell_1}$ 表示矩阵的 L_1 范数。

3.3.3 AMR

流形正则化（Manifold Regularization, MR）在早期被提出，主要是希望通过这种方式去利用大量的未标记样本，来提高分类或回归任务的性能。它的主要思想是利用数据集的几何性状去对模型空间产生约束。该正则化项可以被估计为[84]：

$$L_{mr} = \frac{1}{|\mathcal{D}|^2} \sum_{x_i, x_j \in \mathcal{D}} w_{ij} \left\| \varphi(x_i) - \varphi(x_j) \right\|_2^2, \tag{3.10}$$

其中 w_{ij} 定义为：

$$w_{ij} = \begin{cases} \exp\left(\frac{-\|x_i - x_j\|_2^2}{t} \right) & \text{if } (x_i, x_j) \in \mathcal{P}_k, \\ 0 & otherwise. \end{cases} \tag{3.11}$$

其中温度参数 $t \in \mathbb{R}$，$\mathcal{P}_k$ 表示集合 $\mathcal{D}$ 中所有样本间的 k 近邻构成的样本对集合。该正则化项的直觉是：在原空间中非常接近的样本对，经过变换后，尽量在隐空间中也相近；在原空间不那么靠近的样本对，在隐空间中的紧密度也就不那么重要了（体现在 w_{ij} 的相对大小中）。

本文将流形正则化适应到对抗训练的环境中，提出对抗流形正则化（Adversarial Manifold Regularization, AMR），其主要思想是在原空间中非常接近的干净样本和对抗样本（k 近邻），在隐空间中也应该尽量靠近。AMR 可以被形式化为：

$$\begin{aligned} L_{amr} = & \frac{1}{|\mathcal{D}| \cdot |\mathcal{A}|} \sum_{x_i^{cln} \in \mathcal{D}} \sum_{x_j^{adv} \in \mathcal{A}} w_{ij}^{(1)} \left\| \varphi(x_i^{cln}) - \varphi(x_j^{adv}) \right\|_2^2 \\ & + \frac{1}{|\mathcal{D}| \cdot |\mathcal{A}|} \sum_{x_i^{adv} \in \mathcal{A}} \sum_{x_j^{cln} \in \mathcal{D}} w_{ij}^{(2)} \left\| \varphi(x_i^{adv}) - \varphi(x_j^{cln}) \right\|_2^2, \end{aligned} \tag{3.12}$$

其中 $w_{ij}^{(l)} (l \in \{1,2\})$ 定义为：

$$w_{ij}^{(l)} = \begin{cases} \exp\left(\frac{-\|\Phi(x_i) - \Phi(x_j)\|_2^2}{t} \right) & \text{if } (x_i, x_j) \in \mathcal{P}_k^{(l)}, \\ 0 & otherwise. \end{cases} \tag{3.13}$$

其中温度参数 $t \in \mathbb{R}$，$\mathcal{P}_k^{(1)}$ 表示集合 $\mathcal{D}$ 中所有样本在集合 $\mathcal{A}$ 中的 k 近邻构成的样本对集合，$\mathcal{P}_k^{(2)}$ 表示集合 $\mathcal{A}$ 中所有样本在集合 $\mathcal{D}$ 中的 k 近邻构成的样本对集合。$\Phi(\cdot)$ 是一个

特征提取器（自编码器、自然训练的分类器或对抗训练的分类器），之所以使用一个特征提取器来计算 k 近邻的距离，是因为在高维的图像数据中欧式距离已失效。

第四章 实验与分析

本章对第四章介绍的四个正则化方法进行评估，它们分别是 ALP、MMD、CORAL、AMR。我们将在四个数据集上对它们进行测试，并采用三种攻击方法和一种攻击无关的鲁棒性指标进行有效性验证。我们将对 AMR 方法中的 k 进行取值分析。另外，ALP 由于已经被人发现在 ImageNet 上可以通过增加 PGD 攻击的迭代次数使得鲁棒精度降为 0.6%[79]，但我们认为这只是在一个数据集上的结果，并不一定在别的数据集上成立，因此我们对 ALP 方法进行了 PGD 攻击的收敛性分析；同理，AMR 方法可能会存在和 ALP 同样的问题，且 AMR 方法是我们提出来的，没有被测试过，故我们对 AMR 方法也进行 PGD 攻击的收敛性分析。

4.1 实验设置

我们在四个流行的数据集上进行测试，分别是 MNIST[85]、SVHN[86]、CIFAR10 和 CIFAR100[87]，它们的详细信息如下：

MNIST 是对黑白图像中阿拉伯数字进行识别的数据集，来自美国国家标准与技术研究所。训练集 (training set) 由来自 250 个不同人手写的数字构成，其中 50% 是高中学生，50% 来自人口普查局的工作人员。测试集也是同样比例的手写数字数据。每张图片大小为 $28 \times 28 \times 1$ 像素，测试集包含 50000 个图片，我们将其中 5000 分出来作为验证集，测试集包含 10000 个图片。

SVHN 是对彩色图像中阿拉伯数字进行识别的数据集，该数据集中的图像来自真实世界的门牌号数字，图像来自 Google 街景中所拍摄的门牌号图片，每张图片中包含一组'0-9' 的阿拉伯数字。每张图片大小为 $32 \times 32 \times 3$ 像素，训练集中包含 73257 个图片，测试集中包含 26032 个图片，另有 531131 个附加图片。在我们的实验中，将训练集的 73257 个图片分成了 53257 + 20000，前者为训练集，后者为验证集，不使用附加数字。

CIFAR10 是八十万小图片数据集的子集。由 10 个类的 60000 个 $32 \times 32 \times 3$ 的彩色图像组成，每个类有 6000 个图像。总共有 50000 个训练图像和 10000 个测试图像，

我们将训练图像中的 5000 分出来作为验证集。

CIFAR100 也是八十万小图片数据集的子集。它有 100 个类，每个类包含 600 个 $32\times32\times3$ 的彩色图像。每类各有 500 个训练图像和 100 个测试图像。总共有 50000 个训练图像和 10000 个测试图像，我们将训练图像中的 5000 分出来作为验证集。

对于所有的实验，我们将图片的像素值归一化到 [0, 1] 之间。除了四个正则化方法间的对比，我们还与普通的非对抗训练（Normal Training, NT）和普通的对抗训练（Adversarial Training, AT）进行对比。为了公平起见，我们将所有正则化方法的超参数 λ 设置为 0.1。AMR 方法中特征提取器 $\Phi(\cdot)$ 采用的是自然训练的分类器，温度参数 t 的设置是从 1、10 和 100 三个数字中选取效果最好的，k 是 [1, 10] 之间选择的。所有的实验都是在一个 Titan X GPU 上进行的。在我们的实验中，采用的都是 ℓ_∞ 范数，PGD 攻击的迭代次数 $k = 50$，步长 $\alpha = \epsilon/10$，CW 攻击的常数 $c = 0.001$，迭代次数 $k = 10$。用于训练 MNIST 的模型是一个五层的全连接网络，隐层结点个数依次为 256、256、64 和 10，用于训练 SVHN、CIFAR10 和 CIFAR100 的模型都是 VGG16[34]。我们调整模型使得它们有效，并没有把重心放在最优化这些设置上。

4.2 泛化能力分析

我们在四个数据集上评估模型的鲁棒泛化能力，并进行比较分析。

在 MNIST 上的实验。标准泛化能力和鲁棒泛化能力见表4.1a。在干净的样本上，ALP 表现得最好，AMR 次之。在对抗样本上，NT 的鲁棒泛化能力非常差，CORAL 在 FGSM 攻击下的鲁棒精度上表现最好，而 MMD 在 PGD 和 CW 攻击下的鲁棒精度上表现得最好。在这个数据集上，表现得最好的是 CORAL，表现得最差的是 NT 和 AT，AMR 总体上表现得比 ALP 和 MMD 差。

在 SVHN 上的实验。标准泛化能力和鲁棒泛化能力见表4.1b。在干净样本上，CORAL 表现得最好，NT 次之。在对抗样本上，NT 的鲁棒泛化能力非常差，CORAL 在 FGSM 攻击下的鲁棒精度上表现最好，而 ALP 在 PGD 和 CW 攻击下的鲁棒精度上表现得最好。在这个数据集上，表现得最好的是 CORAL，最差的是 NT 和 AT，AMR 总体上比 ALP 差，但是比 MMD 要好。

在 CIFAR10 上的实验。标准泛化能力和鲁棒泛化能力见表4.1c。在干净的样本上，NT 表现得最好，AT 次之，然而 NT 和 AT 的鲁棒泛化能力都不及其它方法。在对抗样本

上，在三种方法的攻击下，CORAL 均表现出了最好的泛化能力。在这个数据集上，表现得最好的是 CORAL，最差的是 NT 和 AT，总体上 AMR 比 ALP 和 MMD 都要好。

表 4.1 不同方法在测试集上的鲁棒精度

(a) 在 MNIST 上的实验，扰动的幅度为 0.1。

Method	Clean (%)	FGSM(%)	PGD(%)	CW(%)
NT	98.37	29.67	13.38	47.47
AT	98.95	94.16	91.95	94.77
ALP	**99.05**	94.89	91.65	95.08
MMD	98.96	94.52	**92.58**	**95.28**
CORAL	98.93	**95.08**	89.75	94.49
AMR	99.03	94.96	90.24	94.65

(b) 在 SVHN 上的实验，扰动的幅度为 0.02。

Method	Clean (%)	FGSM(%)	PGD(%)	CW(%)
NT	94.36	27.86	2.95	2.77
AT	93.66	93.87	10.18	39.19
ALP	93.63	92.49	**42.31**	**62.76**
MMD	94.00	92.88	20.78	45.18
CORAL	**94.66**	**94.01**	39.96	57.84
AMR	93.90	92.11	40.53	54.76

(c) 在 CIFAR10 上的实验，扰动的幅度为 4/255。

Method	Clean (%)	FGSM(%)	PGD(%)	CW(%)
NT	**87.19**	16.65	1.14	1.03
AT	82.65	56.26	52.59	52.68
ALP	78.85	57.95	55.40	54.64
MMD	82.51	56.70	53.51	53.32
CORAL	80.26	**60.75**	**58.45**	**57.36**
AMR	78.55	58.15	55.78	54.69

(d) 在 CIFAR100 上的实验，扰动的幅度为 4/255。

Method	Clean (%)	FGSM(%)	PGD(%)	CW(%)
NT	**62.55**	11.92	1.20	1.25
AT	56.13	28.16	25.50	25.31
ALP	51.47	30.61	28.74	27.14
MMD	55.90	27.83	24.99	24.89
CORAL	53.92	**32.10**	**30.19**	**28.13**
AMR	51.00	31.50	29.90	27.96

在 CIFAR100 上的实验。标准泛化能力和鲁棒泛化能力见表4.1d。在干净样本上，NT 表现得最好，AT 次之，然而 NT 和 AT 的鲁棒泛化能力都不及其它方法。在对抗样本上，在三种方法的攻击下，CORAL 均表现出了最好的泛化能力。在这个数据集上，表现得最好的是 CORAL，最差的是 NT 和 AT，总体上 AMR 比 ALP 和 MMD 都要好。

总的来说，CORAL 都是效果最好的方法，NT 都是效果最差的方法。AT 这种不加额外正则化项的方法基本上都比其它四种加了正则化的对抗训练方法要差，这证明了四种正则化方法的有效性。另外，注意到 AMR 方法的优劣与数据集相关，在 MNIST 和 SVHN 上，AMR 方法效果没有 ALP 好；而在 CIFAR10 和 CIFAR100 上，AMR 方法比 ALP 和 MMD 都要好。MNIST 和 SVHN 都是十个阿拉伯数字，模式相对简单；CIFAR10 和 CIFAR100 都是自然彩色图像，模式相对复杂，分类任务难度更大，此时体现出了 AMR 相对于 ALP 和 MMD 的优势。

4.3 损失敏感度分析

局部损失敏感度（The local loss sensitivity）是量化模型对扰动的光滑性和泛化性的一种方法。它可以被下式计算。它的值越小，表示损失函数越光滑。

$$\mathcal{S} = \frac{1}{m}\sum_{i=1}^{m} \|\nabla_x J(x_i, y_i)\|_2 \tag{4.1}$$

前文中已训练的模型的局部损失敏感度计算结果在表4.2中。结果显示，与自然训练相比，对抗训练的方法的确增加了模型的光滑度。且大部分加了正则化方法的对抗训练得到了比单纯对抗训练更好的结果，但不同的正则化方法在不同的数据集上各有优劣。ALP 方法在 SVHN 上表现最好，MMD 方法在 MNIST 上表现最好，CORAL 方法在 CIFAR10 上表现最好，AMR 方法在 CIFAR100 上表现最好。

表 4.2 正则化方法的损失敏感度分析

Dataset	NT	AT	ALP	MMD	CORAL	AMR
MNIST (10^{-4})	3.38	1.16	1.76	**1.15**	3.05	1.93
SVHN (10^{-3})	3.22	3.92	**2.70**	4.23	3.14	3.58
CIFAR10 (10^{-3})	6.72	1.75	1.45	1.74	**1.29**	1.36
CIFAR100 (10^{-3})	11.1	2.42	1.48	2.41	1.45	**1.30**

4.4 对抗流形正则化的 k 近邻分析

AMR 方法中，希望在原空间中非常接近的干净样本和对抗样本，在隐空间中也尽量靠近。这里我们选取的是在原空间中最接近的 k 个样本，拉近当前样本和这 k 个样本间的距离，所以 k 的取值决定了数据流形结构的稀疏程度，对最终结果会有所影响。由于显存限制，k 的取值不能过大，我们的实验将 k 在 2 ~ 10 的取值范围内进行分析。

在四个数据集上 k 的取值结果如图4.1、图4.2、图4.3和图4.4所示。可以看出，在四个数据集上，k 的取值对模型的标准泛化能力和 FGSM 的鲁棒泛化能力影响都不大，但是在 MNIST 和 SVHN 上，PGD 的鲁棒泛化能力有所起伏，分别在 k 为 7 和 5 的时候取到了最好的效果。而在 CIFAR10 和 CIFAR100 上，PGD 的鲁棒泛化能力呈现了微弱的下降趋势，分别在 k 为 2 和 3 的时候得到了最好的效果。

4.5 PGD 攻击的收敛性分析

ALP 由于已经被人发现在 ImageNet 上可以通过增加 PGD 攻击的迭代次数使得鲁棒精度从宣称的 27.9% 降为 0.6%[79]，但我们认为这只是在 ImageNet 上的结果，并不一定在本文的四个小型数据集上成立，因此我们对 ALP 方法进行了 PGD 攻击的收敛性分析；同理，AMR 方法可能会存在和 ALP 同样的问题，且 AMR 方法是我们提出来的，没有被测试过，故我们对 AMR 方法也进行 PGD 攻击的收敛性分析。

实验结果如图4.5、图4.6、图4.7和图4.8所示。可以看到，ALP 方法和 AMR 方法在四个数据集上均收敛到较高的一个正确率，并不像在 ImageNet 上那样精度降低至几乎为零。其中在 MNIST、CIFAR10 和 CIFAR100 数据集上，PGD 攻击的迭代次数大约在 50 次时便收敛；在 SVHN 数据集上，PGD 攻击的迭代次数在 100 次左右才收敛。可以看出，在 MNIST 和 SVHN 上，ALP 方法的正确率总是要高于 AMR 方法；而在 CIFAR10 和 CIFAR100 上，AMR 方法总是优于 ALP 方法。

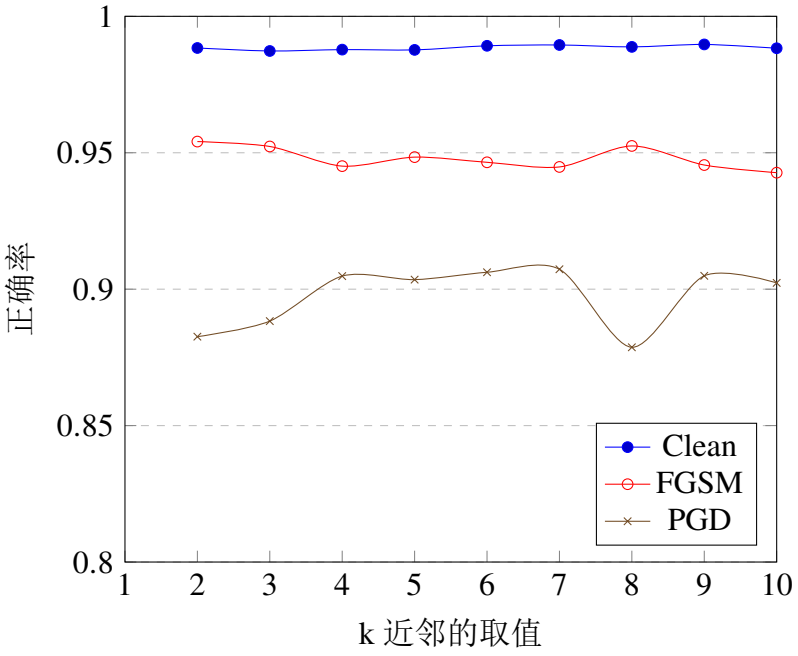

图 4.1　MNIST 的 k 近邻取值分析

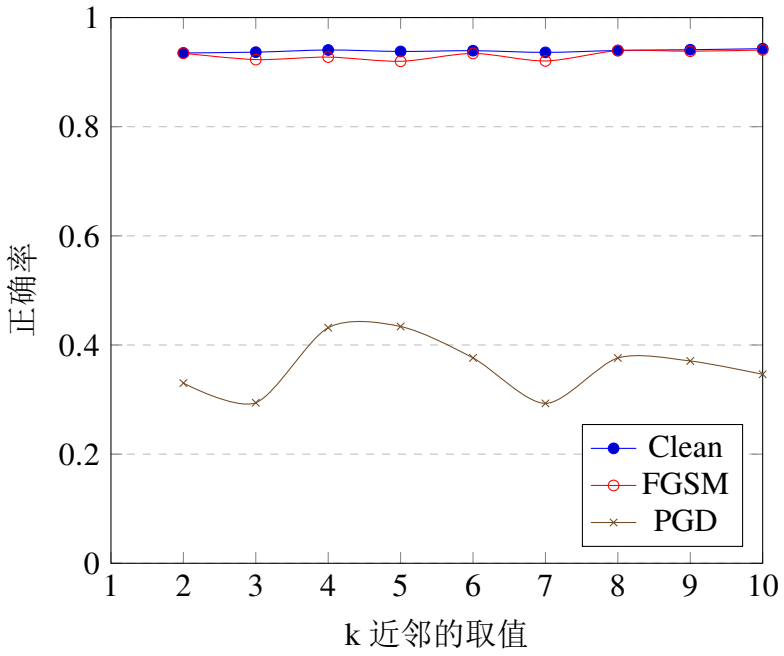

图 4.2　SVHN 的 k 近邻取值分析

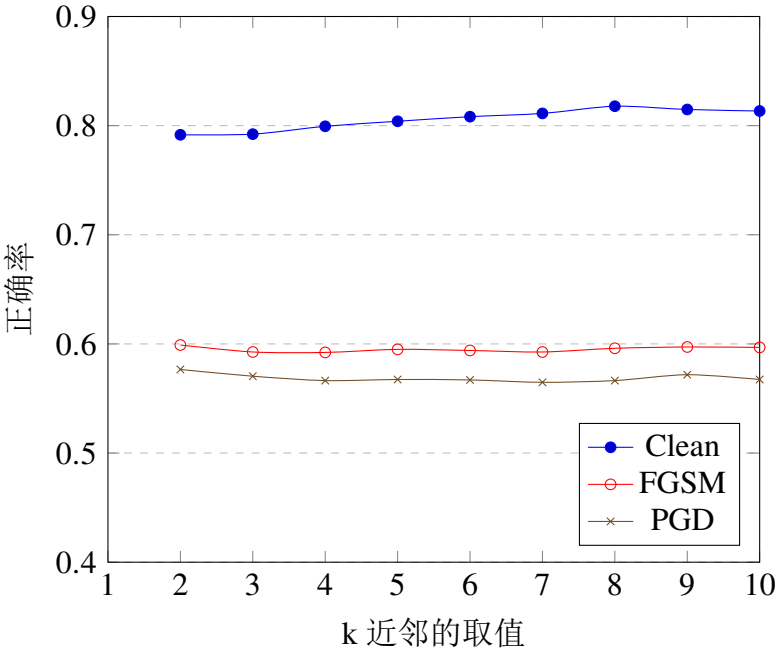

图 4.3　CIFAR10 的 k 近邻取值分析

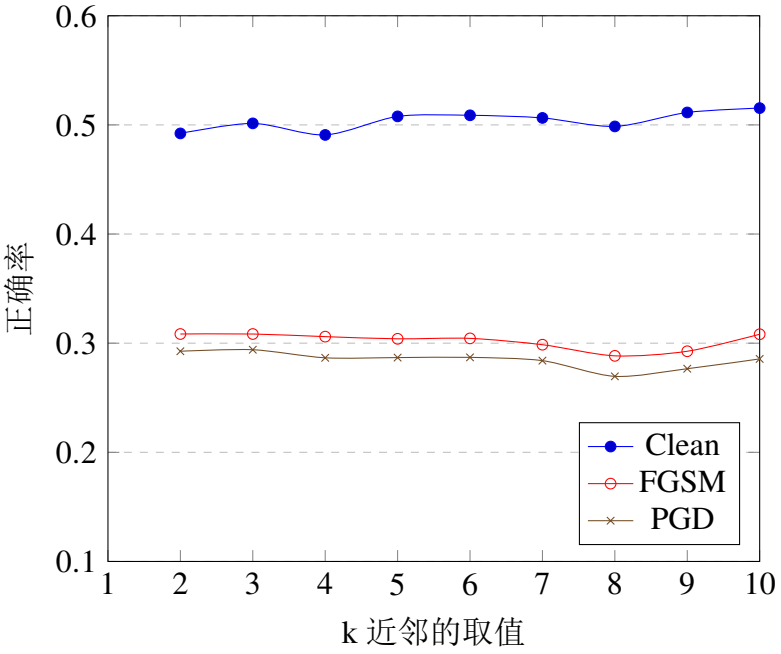

图 4.4　CIFAR100 的 k 近邻取值分析

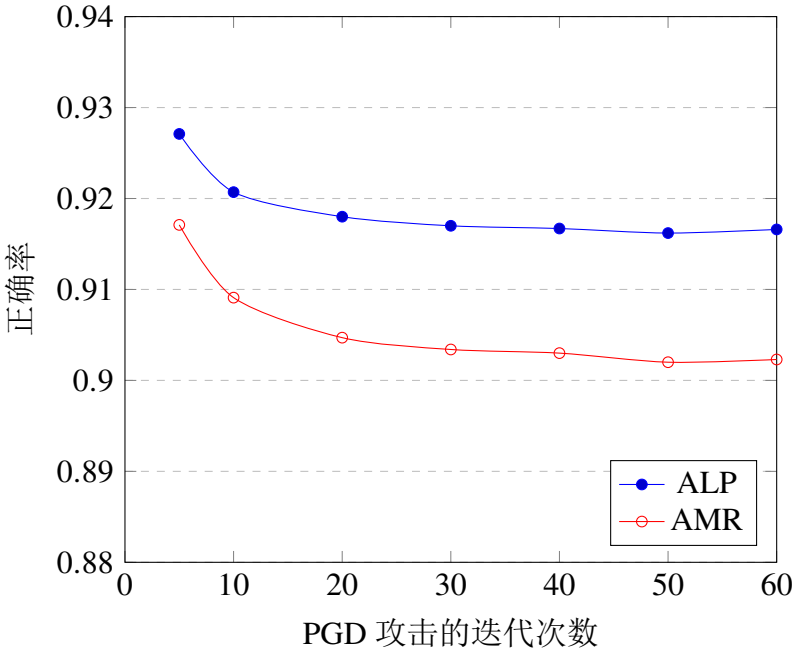

图 4.5　MNIST 上 PGD 攻击的收敛性分析

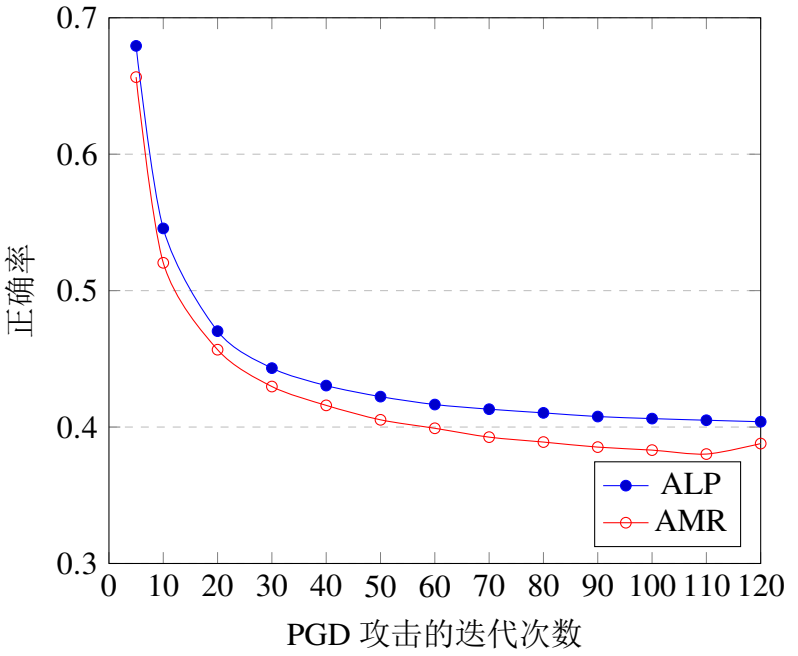

图 4.6　SVHN 上 PGD 攻击的收敛性分析

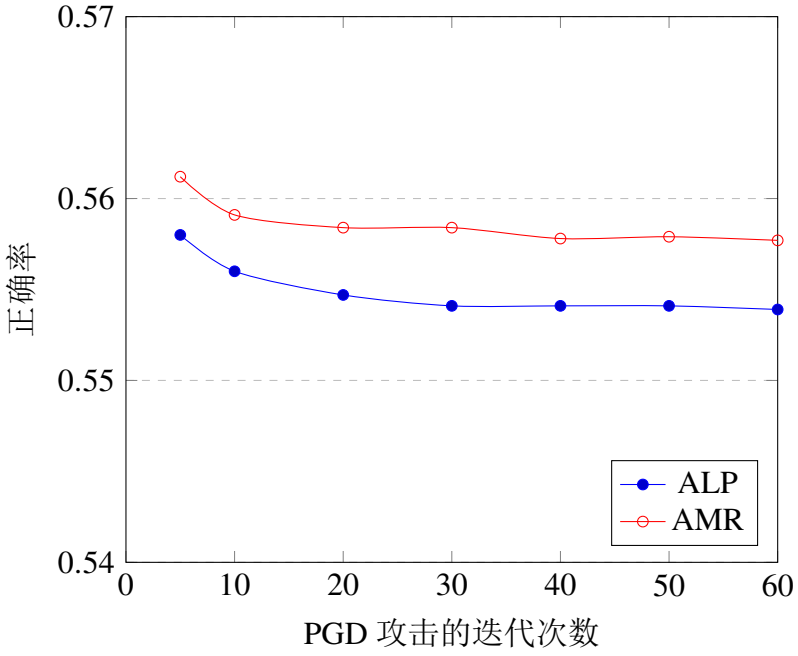

图 4.7　CIFAR10 上 PGD 攻击的收敛性分析

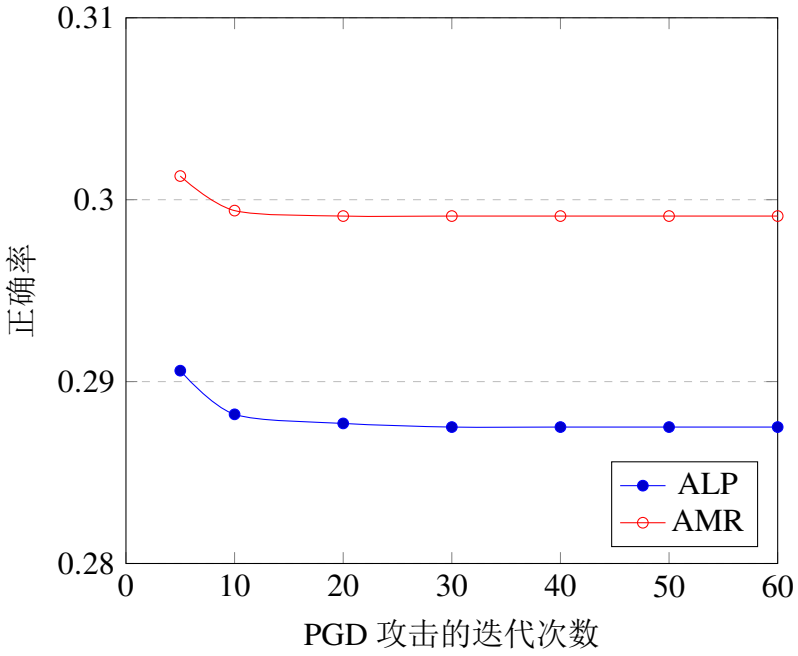

图 4.8　CIFAR100 上 PGD 攻击的收敛性分析

第五章　总结与展望

本文基于对抗训练，即用对抗样本并入深度神经网络的训练集，联合自然样本一起训练以增强神经网络的鲁棒性。为了进一步增加对抗训练的泛化能力，本文结合度量学习的概念，引申出了四种正则化项，其中对抗逻辑配对方法、最大均值散度方法和相关性匹配方法是已经被人应用到对抗训练中，而对抗流形正则化方法则是我们首次应用到对抗训练中。我们在实验部分探究了它们对模型的鲁棒泛化能力和损失敏感度的影响，结果现实正则化方法均有效地提升了神经网络的鲁棒泛化能力，并减小了损失敏感度，证明了我们方法的有效性。

未来的研究可以从以下几个方向入手：

（1）在对抗训练的语境下，探究四种正则化项之间的相互影响，探究多个正则化项的组合是不是可以进一步提升模型的鲁棒泛化能力。

（2）探究模型的深度和宽度等因素对对抗训练泛化能力的影响。

（3）探究多视图任务中对抗训练对模型鲁棒性的影响。

（4）探究对抗训练对迁移学习性能的影响。

（5）探究目标检测、语音识别等任务中对抗训练的实现方式。

参考文献

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Proceedings of Advances in neural information processing systems, 2012. 1097–1105.

[2] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. 779–788.

[3] Saon G, Kuo H K J, Rennie S, et al. The IBM 2015 English conversational telephone speech recognition system[J]. arXiv preprint, 2015. arXiv:1505.05899.

[4] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]. Proceedings of Advances in neural information processing systems, 2014. 3104–3112.

[5] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks[C]. Proceedings of International Conference on Learning Representations, 2017.

[6] Saxe J, Berlin K. Deep neural network based malware detection using two dimensional binary program features[C]. Proceedings of 2015 10th International Conference on Malicious and Unwanted Software (MALWARE). IEEE, 2015. 11–20.

[7] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]. Proceedings of International Conference on Learning Representations, 2015.

[8] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint, 2013. arXiv:1312.6199.

[9] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world[J]. arXiv preprint, 2016. arXiv:1607.02533.

[10] Eykholt K, Evtimov I, Fernandes E, et al. Robust Physical-World Attacks on Deep Learning Visual Classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[11] Xie C, Wang J, Zhang Z, et al. Adversarial examples for semantic segmentation and object detection[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017. 1369–1378.

[12] Carlini N, Mishra P, Vaidya T, et al. Hidden voice commands[C]. Proceedings of 25th {USENIX} Security Symposium ({USENIX} Security 16), 2016. 513–530.

[13] Zhang G, Yan C, Ji X, et al. Dolphinattack: Inaudible voice commands[C]. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017. 103–117.

[14] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 1–9.

[15] Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples[J]. arXiv preprint, 2014. arXiv:1412.5068.

[16] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[C]. Proceedings of International Conference on Learning Representations, 2018.

[17] Papernot N, McDaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]. Proceedings of 2016 IEEE Symposium on Security and Privacy (SP). IEEE, 2016. 582–597.

[18] Rozsa A, Rudd E M, Boult T E. Adversarial diversity and hard positive generation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016. 25–32.

[19] Zheng S, Song Y, Leung T, et al. Improving the robustness of deep neural networks via stability training[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 4480–4488.

[20] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]. Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017. 39–57.

[21] Raghunathan A, Steinhardt J, Liang P. Certified Defenses against Adversarial Examples[C]. Proceedings of International Conference on Learning Representations, 2018.

[22] Wong E, Kolter Z. Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope[C]. In: Dy J, Krause A, (eds.). Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, Stockholmsmässan, Stockholm Sweden: PMLR, 2018. 5286–5295.

[23] Wong E, Schmidt F, Metzen J H, et al. Scaling provable adversarial defenses[C]. Proceedings of Advances in Neural Information Processing Systems, 2018. 8400–8409.

[24] Grosse K, Manoharan P, Papernot N, et al. On the (statistical) detection of adversarial examples[J]. arXiv preprint, 2017. arXiv:1702.06280.

[25] Gong Z, Wang W, Ku W S. Adversarial and clean data are not twins[J]. arXiv preprint, 2017. arXiv:1704.04960.

[26] Metzen J H, Genewein T, Fischer V, et al. On detecting adversarial perturbations[C]. Proceedings of International Conference on Learning Representations, 2017.

[27] Li X, Li F. Adversarial examples detection in deep networks with convolutional filter statistics[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017. 5764–5772.

[28] Feinman R, Curtin R R, Shintre S, et al. Detecting adversarial samples from artifacts[J]. arXiv preprint, 2017. arXiv:1703.00410.

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

[30] Schmidt L, Santurkar S, Tsipras D, et al. Adversarially robust generalization requires more data[C]. Proceedings of Advances in Neural Information Processing Systems, 2018. 5014–5026.

[31] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786):504–507.

[32] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. The Journal of physiology, 1962, 160(1):106–154.

[33] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4):541–551.

[34] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint, 2014. arXiv:1409.1556.

[35] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. 770–778.

[36] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. 7132–7141.

[37] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018. 3–19.

[38] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic

segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014. 580–587.

[39] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2015. 1440–1448.

[40] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. Proceedings of Advances in neural information processing systems, 2015. 91–99.

[41] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. Proceedings of European conference on computer vision. Springer, 2016. 21–37.

[42] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. 7263–7271.

[43] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint, 2018. arXiv:1804.02767.

[44] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2117–2125.

[45] Chopra S, Hadsell R, LeCun Y, et al. Learning a similarity metric discriminatively, with application to face verification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005. 539–546.

[46] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, volume 2. IEEE, 2006. 1735–1742.

[47] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of Machine Learning Research, 2009, 10(Feb):207–244.

[48] Chechik G, Sharma V, Shalit U, et al. Large scale online learning of image similarity through ranking[J]. Journal of Machine Learning Research, 2010, 11(Mar):1109–1135.

[49] Tsang I W, Kwok J T, Bay C, et al. Distance metric learning with kernels[C]. Proceedings of the International Conference on Artificial Neural Networks. Citeseer, 2003. 126–129.

[50] Yeung D Y, Chang H. A kernel approach for semisupervised metric learning[J]. IEEE Transactions on Neural Networks, 2007, 18(1):141–149.

[51] Hu J, Lu J, Tan Y P. Discriminative deep metric learning for face verification in the wild[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014. 1875–1882.

[52] Hoffer E, Ailon N. Deep metric learning using triplet network[C]. Proceedings of International Workshop on Similarity-Based Pattern Recognition. Springer, 2015. 84–92.

[53] Oh Song H, Xiang Y, Jegelka S, et al. Deep metric learning via lifted structured feature embedding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 4004–4012.

[54] Sohn K. Improved deep metric learning with multi-class n-pair loss objective[C]. Proceedings of Advances in Neural Information Processing Systems, 2016. 1857–1865.

[55] Dalvi N, Domingos P, Sanghai S, et al. Adversarial classification[C]. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004. 99–108.

[56] Biggio B, Fumera G, Roli F. Multiple classifier systems for robust classifier design in adversarial environments[J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1-4):27–41.

[57] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time[C]. Proceedings of Joint European conference on machine learning and knowledge discovery in databases. Springer, 2013. 387–402.

[58] Roli F, Biggio B, Fumera G. Pattern recognition systems under attack[C]. Proceedings of Iberoamerican

Congress on Pattern Recognition. Springer, 2013. 1–8.

[59] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. 2574–2582.

[60] Chen P Y, Zhang H, Sharma Y, et al. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. ACM, 2017. 15–26.

[61] Su J, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019..

[62] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[C]. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016. 1528–1540.

[63] Lab T K S. Experimental Security Research of Tesla Autopilot[R]. Technical report, 2019. `https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf`.

[64] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017.

[65] Qin Y, Carlini N, Goodfellow I, et al. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition[C]. Proceedings of the 36th International Conference on Machine Learning, 2019.

[66] Lu J, Issaranon T, Forsyth D. Safetynet: Detecting and rejecting adversarial examples robustly[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017. 446–454.

[67] Salimans T, Karpathy A, Chen X, et al. PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications[C]. Proceedings of International Conference on Learning Representations, 2017.

[68] Song Y, Kim T, Nowozin S, et al. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples[J]. arXiv preprint arXiv:1710.10766, 2017..

[69] Carlini N, Wagner D. Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples[J]. arXiv preprint arXiv:1711.08478, 2017..

[70] Buckman J, Roy A, Raffel C, et al. Thermometer Encoding: One Hot Way To Resist Adversarial Examples[C]. Proceedings of International Conference on Learning Representations, 2018.

[71] Guo C, Rana M, Cisse M, et al. Countering Adversarial Images using Input Transformations[C]. Proceedings of International Conference on Learning Representations, 2018.

[72] Dhillon G S, Azizzadenesheli K, Bernstein J D, et al. Stochastic activation pruning for robust adversarial defense[C]. Proceedings of International Conference on Learning Representations, 2018.

[73] Xie C, Wang J, Zhang Z, et al. Mitigating Adversarial Effects Through Randomization[C]. Proceedings of International Conference on Learning Representations, 2018.

[74] Song Y, Kim T, Nowozin S, et al. PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples[C]. Proceedings of International Conference on Learning Representations, 2018.

[75] Samangouei P, Kabkab M, Chellappa R. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models[C]. Proceedings of International Conference on Learning Representations, 2018.

[76] Athalye A, Carlini N, Wagner D. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples[C]. In: Dy J, Krause A, (eds.). Proceedings of Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*,

Stockholmsmässan, Stockholm Sweden: PMLR, 2018. 274–283.

[77] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale[C]. Proceedings of International Conference on Learning Representations, 2017.

[78] Kannan H, Kurakin A, Goodfellow I. Adversarial logit pairing[J]. arXiv preprint arXiv:1803.06373, 2018..

[79] Engstrom L, Ilyas A, Athalye A. Evaluating and understanding the robustness of adversarial logit pairing[J]. arXiv preprint arXiv:1807.10272, 2018..

[80] Song C, He K, Wang L, et al. Improving the Generalization of Adversarial Training with Domain Adaptation[C]. Proceedings of International Conference on Learning Representations, 2019.

[81] Farnia F, Zhang J, Tse D. Generalizable Adversarial Training via Spectral Normalization[C]. Proceedings of International Conference on Learning Representations, 2019.

[82] Borgwardt K M, Gretton A, Rasch M J, et al. Integrating structured biological data by kernel maximum mean discrepancy[J]. Bioinformatics, 2006, 22(14):e49–e57.

[83] Sun B, Saenko K. Deep coral: Correlation alignment for deep domain adaptation[C]. Proceedings of European Conference on Computer Vision. Springer, 2016. 443–450.

[84] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples[J]. Journal of machine learning research, 2006, 7(Nov):2399–2434.

[85] LeCun Y. The MNIST database of handwritten digits[J]. http://yann. lecun. com/exdb/mnist/, 1998..

[86] Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning[C]. Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.

[87] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[R]. Technical report, Citeseer, 2009.

致　谢

光阴似箭，日月如梭。仿佛昨天刚刚拖着行李箱踏入南航大西门的广场，今天就要完成毕业前最后一门功课，曾经过往犹如黄粱一梦。只有这大学期间印在脑海的种种酸甜苦辣，才让人真真切切地感叹到：噢！原来真的四年过去了。

这四年里，我做的最大的一个决定，也是最无悔的一个决定，就是从工研班肄业后选择进入计算机科学与技术专业学习了吧。感谢一路上给予我鼓励和帮助的老师和同学们。

首先，非常感谢我的导师陈松灿教授。陈老师对待学术严谨又充满激情，待人平和又不失活泼，是一位难得的好老师。陈老师不仅仅是我本科毕业设计的导师，也是我的本科学术导师，感谢工研班给我提供了这样的机会，让我在本科低年级时就有幸与陈老师交流并获得学术指导，也是因此坚定了我进入计科的决心。

其次，非常感谢孙涵副教授。在我转入计算机科学与技术专业之后，给予了我许多支持和鼓励，让我能够在修读核心专业课程的同时，还参与了多项科创活动，极大地提高了我的编程能力和对深度学习的理解程度。同时，还要感谢这一路上同甘共苦的郭越超同学和杨昊同学，和你们一起在实验室熬过的那些夜晚是最珍贵的回忆。

再次，感谢实验室的师兄师姐们：黄飞虎、李平、朱颖雯、马忠臣、马迪、耿传兴、冯泉、林云霞、李伟凯、李想、胡梦磊、徐丹丹、余欢欢、葛尧、刘颉羲、王玮皓、谭正豪、史小艳、钟颖宇、夏笑秋等。感谢你们营造的良好的实验室氛围，和你们一起学习、交流让我受益匪浅，祝福大家今后都有美好的前程。

最后，特别感谢我亲爱的父母，是你们让我得以一路求学而无后顾之忧，鼓励我求知上进，在我遇到困难时做我最坚强的后盾，还有一直陪伴支持我的女朋友，你们是我人生中最大的动力，谢谢你们。