Adversarial attack and interpretability of the deep neural network from the geometric perspective

夏萌霏, 叶子鹏, 赵旺, 易冉 and 刘永进

Citation: 中国科学: 信息科学 **51**, 1411 (2021); doi: 10.1360/SSI-2020-0169

View online: https://engine.scichina.com/doi/10.1360/SSI-2020-0169

View Table of Contents: https://engine.scichina.com/publisher/scp/journal/SSI/51/9

Published by the 《中国科学》杂志社

中国科学: 信息科学 2021 年 第 51 卷 第 9 期: 1411–1437
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
评述

几何视角下深度神经网络的对抗攻击与可解释性研究进展

夏萌霏, 叶子鹏, 赵旺, 易冉, 刘永进*
清华大学计算机科学与技术系, 北京 100084
* 通信作者. E-mail: liuyongjin@tsinghua.edu.cn
收稿日期: 2020–06–08; 修回日期: 2020–08–06; 接受日期: 2020–10–14; 网络出版日期: 2021–09–14

国家杰出青年科学基金 (批准号: 61725204) 和国家重点研发计划项目 (批准号: 2016YFB1001200) 资助

摘要 随着深度神经网络在机器学习的各个领域获得广泛成功, 其自身所存在的问题也日益尖锐和突出, 例如可解释性差、鲁棒性弱和模型训练难度大等. 这些问题严重影响了神经网络模型的安全性和易用性. 因此, 神经网络的可解释性受到了大量的关注, 而利用模型可解释性改进和优化模型的性能也成为研究热点之一. 在本文中, 我们通过几何中流形的观点来理解深度神经网络的可解释性, 在通过流形视角分析神经网络所遇到的问题的同时, 汇总了数种有效的改进和优化策略并对其加以解释. 最后, 本文对深度神经网络流形解释目前存在的挑战加以分析, 提出将来可能的发展方向, 并对今后的工作进行了展望.

关键词 深度学习, 对抗攻击, 可解释性, 流形

1 引言

深度学习近些年来在诸多机器学习领域都取得了巨大的成功, 包括图像分类 [1]、物体检测 [2]、物体识别 [3]、图像生成 [4] 等方面. 而随着深度学习的蓬勃发展, 其安全性也得到了越来越多的关注. 尽管目前主流的深度学习模型都取得了出色的准确率, 但近期的研究也表明深度学习大都存在鲁棒性差的问题: 比如, 使用经过处理、人眼难以区分的对抗样本, 就可轻易使得性能良好的模型得到错误的结果.

对于对抗攻击的研究和理解, 能够提高深度学习这一机器学习中的 “黑盒” 的可解释性, 从而能够更加合理地设计和搭建深度学习的网络结构、并提高模型的性能和鲁棒性; 另一方面, 在对于网络有更加深刻的理解和认识后, 也能更具有针对性地设计和生成深度学习网络的对抗样本. 因此在 Szegedy 等 [5] 第 1 次提出了针对深度神经网络的对抗攻击算法并取得了显著的效果后, 有关对抗攻击的研究也得到了迅速的发展. Yuan 等 [6] 汇总分类了目前主流的对抗攻击方法, 常见的攻击类别有: 黑盒攻击、白盒攻击、l^0 攻击、l^1 攻击、l^∞ 攻击等. 图 1 为使用 FGSM [7] 算法的对抗攻击结果.

引用格式: 夏萌霏, 叶子鹏, 赵旺, 等. 几何视角下深度神经网络的对抗攻击与可解释性研究进展. 中国科学: 信息科学, 2021, 51: 1411–1437, doi: 10.1360/SSI-2020-0169
Xia M F, Ye Z P, Zhao W, et al. Adversarial attack and interpretability of the deep neural network from the geometric perspective (in Chinese). Sci Sin Inform, 2021, 51: 1411–1437, doi: 10.1360/SSI-2020-0169

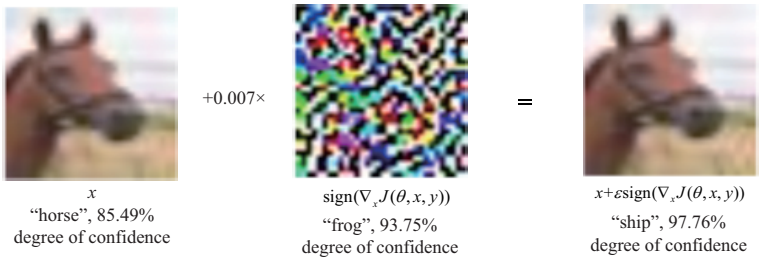

图 1 (网络版彩图) 利用 FGSM [7] 算法使得分类器将人眼识别为 "马" 的图片识别为 "船", 并有着很高的置信度

Figure 1 (Color online) FGSM [7] algorithm misleads the classifier to recognize the picture as a ship with a high degree of confidence, which the human eye recognizes as a horse

1.1 对抗攻击与深度学习可解释性

自从深度神经网络和对抗攻击算法提出以来, 关于其可解释性的研究也随之蓬勃发展. 传统的神经网络考虑一个如下的损失函数:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(x,y;\theta)], \tag{1}$$

其中 $\mathcal{D}$ 是数据的分布, L 是损失函数, θ 是模型参数. 深度学习的训练旨在找到全局最优的模型参数 θ. 但对抗攻击算法的提出使得模型的鲁棒性越发受到关注, 因此 Tsipras 等 [8] 在式 (1) 的基础上提出了如下的对抗损失函数:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\Delta(x)} L(x+\delta,y;\theta)\right], \tag{2}$$

其中 $\Delta(x)$ 是某些特定类别的、针对 x 的有效扰动. 此外, Tsipras 还尝试了将对抗攻击算法所生成的对抗样本加入神经网络的训练数据, 从而达到增强模型鲁棒性的目的. Szegedy 等 [5] 的研究也表明, 将对抗样本加入训练过程可以起到增加模型正则化 (regulization) 的作用, 因此对抗训练也被视作一种提高模型泛化性能的数据增强 (data augmentation) 方法. 而 Tramér 等 [9] 提出对抗样本具有一定的迁移性, 这一定程度上表明各种神经网络模型的决策边界是类似的, 给神经网络的可解释性提供了新的突破口. 另一方面, Zhang 等 [10] 认为对于对抗攻击有着较高鲁棒性的神经网络模型具有更高的可解释性. 因此, 对于对抗攻击的理解和分析, 能够提高神经网络的可解释性.

另一方面, 对于神经网络更加深刻的理解则提供了更有效的对抗攻击算法. 从最初计算复杂度很高的 L-BFGS [5], 到基于梯度的 FGSM [7] 和较高迁移性的 MI-FGSM [11], 再到利用对抗样本迁移性从白盒攻击迁移至黑盒攻击 [12], 对抗攻击背后的理论框架在逐步深化、并更加具有针对性, 这些进步都得益于对于神经网络越发深刻的理解和分析.

1.2 本文工作

本文调研了从几何观点下理解和分析对抗攻击与深度神经网络可解释性的研究工作, 并将其总结提炼成基于流形观点的理论框架. 在这一理论框架下, 本文将对深度神经网络的工作原理、结构缺陷进行阐述与分析, 并从理论角度分析多种对抗样本算法的理论基础和合理性. 更进一步地, 本文将在同一理论框架下针对深度神经网络领域中的生成对抗网络 (generative adversarial network) 进行分析和解释. 本文的主要结构如下:

(1) 第 2 节主要介绍深度学习中流形理论的基本假设, 即假设数据样本集合具有流形的结构, 还将介绍在这一假设下对于深度神经网络工作原理的理解和流形框架相比于传统优化或统计等观点的

优势.

(2) 第 3 和 4 节将从流形观点详细分析深度神经网络中出现的种种问题和对抗攻击的理论依据. 特别地, 将对生成对抗网络进行解释, 尤其是对生成对抗网络中最为典型的模式崩塌问题进行分析. 另外, 还将结合针对深度神经网络的优化、对抗攻击的防御, 对常见的优化方法进行总结和分析.

(3) 第 5 节将对于流形理论框架目前面临的挑战进行阐述, 以及对未来基于流形角度的深度神经网络理解分析的工作进行展望, 而第 6 节对全文进行了总结.

2 深度神经网络的流形框架

在本节, 首先对深度学习模型已有的传统分析观点进行了总结, 然后介绍基本的流形理论和假设, 并对该几何理论模型下的深度学习理解进行了阐述.

2.1 传统深度学习解释框架

随着深度神经网络这一领域的蓬勃发展, 神经网络这一黑箱的可解释性也越来越受到关注. 本小节主要汇总了目前较为主流的 3 种解释观点并对其进行简要的介绍.

2.1.1 优化观点

传统神经网络基于如下的万能逼近定理:

定理1 (万能逼近定理, Cybenko [13]) 令 $\varphi:\mathbb{R}\to\mathbb{R}$ 为非常值有界连续函数, 令 $C([0,1]^m)$ 为 $[0,1]^m$ 上的实值连续函数. 则对于任意 $\epsilon>0$ 和 $f\in C([0,1]^m)$, 存在 $N\in\mathbb{N}$, $v_i,b_i\in\mathbb{R}$, $w_i\in\mathbb{R}^m$, $i=1,2,\ldots,N$, 定义

$$F(x):=\sum_{i=1}^{N}v_i\varphi(w_i^{\mathrm{T}}x+b_i),$$

则有

$$|F(x)-f(x)|<\epsilon,\quad \forall x\in[0,1]^m,$$

即形如 $F(x)$ 的函数在 $C([0,1]^m)$ 中稠密.

因此损失函数 (1) 可以看作神经网络所生成的函数与真实映射之间的距离, 而神经网络的训练过程则可视为以这一损失函数为目标函数的优化求解过程. 在线性分类器中通常考虑的 L^2 度量下, 式 (1) 化为

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[|w^{\mathrm{T}}x-y|^2]. \tag{3}$$

这是一个凸函数, 因此可以使用凸优化理论针对这一模型进行细致的分析. 但在绝大多数的任务中, 损失函数往往不具有凸性. 因此尽管目前已经有了较为成熟的优化理论, 但由于神经网络的优化问题参数多、网络结构复杂、维数大等, 神经网络的训练过程仍然难以控制, 并往往容易陷入局部最优解.

2.1.2 微分方程观点

尽管目前的深度神经网络具有很高的层数, 但每一层都可以看作一个动力系统:

$$X_{n+1}=f_n(X_n), \tag{4}$$

这里 f_n 代表神经网络在第 n 层的函数, X_n, X_{n+1} 分别代表神经网络在第 n 层的输入和输出. Lu 等 [14] 将包括 ResNet [15] 在内的若干最新神经网络模型的动力系统分别化为了一阶常微分方程 (ODE) 的前向 Euler 格式、后向 Euler 格式和 Runge-Kutta 格式, 以 ResNet 为例,

$$X_{n+1} = X_n + f(X_n) \Longrightarrow X_t = f(X),$$

这里 X_t 为对一个虚拟时间 t 的导数. 因此神经网络可以视为一个数值 ODE, 从而神经网络的训练过程也可通过控制论来理解和分析. 更进一步地, 使用随机的 dropout 函数的神经网络也可被视为一个随机动力系统 [16]. 因此 ODE 的不同离散化方法则对应了不同的网络结构. Lu 等 [14] 通过线性多步法

$$X_{n+1} = kX_n + (1-k)X_{n-1} + f(X_n),$$

设计出新的网络结构, 有效地提高了模型的性能.

2.1.3 统计观点

同传统的机器学习理论一样, 统计理论对于解释生成模型也有着重要的意义. 在 1994 年, Cheng 等 [17] 就使用统计观点对深度神经网络进行了细致的分析. 本文将在第 4 节中以生成对抗网络 (GAN) 为例进行分析, 而本小节将使用 Bayes 理论对于分类问题进行分析.

考虑一个 $\{0,1\}$ 的二分类问题, 记数据集为 $\mathcal{X}$, 神经网络分类器对应函数 $f(x) \in [0,1]$, $x \in \mathcal{X}$, 记 $\mathbb{P}[x]$ 为样本 x 的标签为 1 的概率, 则损失函数可写为

$$\mathcal{E} = \sum_{x\in\mathcal{X}} \left(\mathbb{P}[x](1-f(x))^2 + (1-\mathbb{P}[x])f^2(x)\right). \tag{5}$$

由于损失函数为凸函数, 通过损失函数的最小化可以得到方程

$$-2\mathbb{P}[x](1-f(x)) + 2(1-\mathbb{P}[x])f(x) = 0.$$

因此有 $f(x) = \mathbb{P}[x]$, 也即最优的神经网络分类器能够学到后验分布, 而期望的损失变为

$$\mathbb{P}[x](1-\mathbb{P}[x])^2 + (1-\mathbb{P}[x])\mathbb{P}[x]^2 = \mathbb{P}[x](1-\mathbb{P}[x]),$$

而总损失函数 $\mathcal{E}$ 变为期望的方差.

2.2 基本流形理论

本小节将介绍基本的流形理论, 并将讨论流形上的概率空间.

定义1 (拓扑流形, Lee [18]) 一个 n 维拓扑流形 M 是一个第二可数的 Hausdorff 空间, 并有一族开覆盖 $\{(U_\alpha,\phi_\alpha)\}_{\alpha\in I}$ 满足:

- $\bigcup_{\alpha\in I} U_\alpha = M$;
- $\phi_\alpha : U_\alpha \to \mathbb{R}^n$ 是同胚;
- $\forall \alpha,\beta$, $U_\alpha \cap U_\beta \neq \emptyset$, 都有 $\phi_{\alpha\beta} = \phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta)$ 是 $\mathbb{R}^n$ 中的连续函数. 其中, $\{(U_\alpha,\phi_\alpha)\}_{\alpha\in I}$ 被称为 M 的图册, 每个 (U_α,ϕ_α) 被称作 M 的图卡.

定义2 (微分流形, Milnor [19]) 一个 n 维 C^i 流形 M 是一个第二可数的 Hausdorff 空间, 并有一族开覆盖 $\{(U_\alpha,\phi_\alpha)\}_{\alpha\in I}$ 满足:

• $\bigcup_{\alpha\in I} U_\alpha = M$;

• $\phi_\alpha : U_\alpha \to \mathbb{R}^n$ 是 C^i 微分同胚;

• $\forall \alpha, \beta$, $U_\alpha \cap U_\beta \neq \emptyset$, 都有 $\phi_{\alpha\beta} = \phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta)$ 是 $\mathbb{R}^n$ 中的 C^i 光滑函数. 其中, $\{(U_\alpha, \phi_\alpha)\}_{\alpha\in I}$ 被称为 M 的图册, 每个 (U_α, ϕ_α) 被称作 M 的图卡.

定义3 (流形间的映射, Lee [18]) 令 M, N 分别是 m, n 维流形, $f : M \to N$, 则对 $x \in M$, 称 f 在 x 处 C^i, 若有 M 的图卡 (U_α, ϕ_α) 和 N 的图卡 (V_β, ψ_β) 使得

$$x \in U_\alpha \subset M, \quad f(x) \in V_\beta \subset N,$$

且复合映射

$$\psi_\beta \circ f \circ \phi_\alpha^{-1} : \mathbb{R}^m \to \mathbb{R}^n,$$

在 $\phi_\alpha(x) \in \mathbb{R}^m$ 处 C^i. 称 f 在 M 上 C^i, 如果 $\forall x \in M$ 都有 f 在 x 处 C^i.

注释1 一般地, 对于拓扑流形只能谈论其间映射的连续性; 而对于 C^j 光滑性的讨论只适用于 $i \geqslant j$ 的 C^i 流形.

定义4 (Hausdorff 测度, Evans 等 [20]) 设 $n \in \mathbb{N}, k \geqslant 0$, 记 $\Gamma(x)$ 为 Gamma 函数, 并令 $\alpha_k = \frac{\pi^{\frac{k}{2}+1}}{\Gamma(\frac{k}{2}+1)}$, 对任意 $\delta > 0$ 和任意集合 $\mathbb{E} \subset \mathbb{R}^n$, 定义

$$H^*_{k,\delta}(E) = \inf\left\{\sum_{j=1}^{\infty} \alpha_k \left(\frac{\mathrm{diam}B_j}{2}\right)^k \,\middle|\, E \subset \bigcup_{j=1}^{\infty} B_j, \mathrm{diam}B_j \leqslant \delta\right\},$$

其中每个 B_j 非空, $\mathrm{diam}B_j$ 为集合的直径, 也即,

$$\mathrm{diam}B = \sup_{x,y\in B} \|x - y\|_2^2.$$

令

$$H^*_k(E) = \lim_{\delta\to 0^+} H^*_{k,\delta}(E),$$

称 $H^*_k(E)$ 为 E 的 k 维 Hausdorff 外测度. 令 $(\mathbb{R}^n, \mathcal{M}_k, H_k)$ 是从外测度 H^*_k 通过 Caratheodory 定理得到的完备度量空间, 称 $\mathcal{M}_k$ 的元素 E 为 H_k – 可测集, 称 $H_k(E)$ 为 E 的 k 维 Hausdorff 测度.

定义5 (流形上的概率分布, Evans 等 [20]) 对于 k 维流形 M 定义 σ – 代数为

$$\Sigma = \{E \cap M \mid E \in \mathcal{M}_k\},$$

再定义 M 上的非负 H_k – 可测函数 $p : M \to \mathbb{R}$, 使得其在 M 上的积分满足

$$\int_M p(x)\mathrm{d}H_k(x) = 1,$$

则可定义概率测度

$$\mathbb{P}[E] = \int_E p(x)\mathrm{d}H_k(x), \quad \forall E \in \Sigma,$$

也即定义了概率空间 $(M, \Sigma, \mathbb{P})$, 并称 p 为 M 上的概率密度函数.

定理2 (Villani [21]) 令 M, N 分别是 m, n 维流形, (M, Σ, μ) 是概率空间. 令 $f : M \to N$, 则可定义 N 上的 σ – 代数为 $\Sigma' = \{B \mid f^{-1}(B) \in \Sigma\}$, 那么 $\forall V \in \Sigma'$, 定义

$$f_\#(\mu)(V) = \mu(f^{-1}(V)). \tag{6}$$

那么 $f_\#(\mu)$ 是 N 上的概率分布, 称为 μ 的推前测度 (push-forward).

2.3 流形观点基本假设

本文主要关注的是深度神经网络在图像领域的应用, 而图像领域也是流形假设最广为接受的一个领域. Seung [22] 在 2000 年提出, 一族图像具有某一坐标系下的笛卡尔坐标, 从而构成一个图像空间. 而随着观测角度、物体旋转角度连续的变化, 图像也会连续的变化, 从而构成图像空间的一个流形.

更进一步地, 以尺寸为 256×256 的彩色真实图像为例, 所有的真实图像构成了 $[0,255]^{3\times256\times256}$ 中的一个子集, 其中 $[0,255]$ 代表了 RGB 值的范围, 3 代表了 RGB 的 3 个通道. 由于真实图像在视觉上具有特殊的语义信息, 因此真实图像集合满足某种几何分布上的限制条件, 即可以合理地假设这样的集合同样具有流形的结构. 同样地, 根据不同真实图像的出现频率也可以假定在真实图像流形上具有某种概率分布函数.

通过将流形框架与 2.1 小节中提到的深度神经网络解释框架相对比, 可以发现:

• 相比于优化观点来说, 虽然在优化理论下可以通过改进优化算法来改进模型的训练过程, 但由于神经网络的损失函数往往是非凸的, 这样的改进往往十分困难; 另一方面, 优化观点只关注了模型在整个训练集的损失函数而无法涉及到模型的泛化性能 (即模型在测试集上的性能表现) 这一重要指标, 因此也就无法指导和解决模型所遇到的性能较差的问题. 由于优化观点基于万能逼近定理, 即神经网络的连续函数能够逼近任何函数, 因此从流形角度考虑神经网络模型逐层所使用的函数, 能够更深入地分析模型结构对于逼近函数的影响, 也即对神经网络学习能力和神经网络泛化性能的影响.

• 相比于微分方程观点来说, 微分方程观点通过将神经网络模型逐层的函数看作动力系统, 能够更加深刻地理解模型结构和模型性能之间的关系, 并且通过微分方程数值解的离散化理论, 能够对于模型结构的改进和设计起到直观的指导作用. 但目前为止这一理论仅适用于某一部分神经网络模型, 也并不能像流形框架直观地考虑和分析数据集对于模型训练的影响, 并且由于缺乏对于数据分布的考虑可能难以适用于复杂的生成模型.

• 相比于统计观点而言, 虽然统计观点能够从模型学习数据分布的能力来分析模型的性能, 但在这一框架下分析模型训练的过程对于性能的影响却较为困难. 并且, 统计观点也同优化观点一样无法分析模型结构对于模型性能的影响. 除此以外, 因为流形框架将数据集看作嵌入空间的一个低维子流形, 这不仅考虑了流形上的分布, 更涉及到了模型逐层所使用的函数, 所以在兼具统计观点的优势的同时还能够考虑数据集本身的几何特征. 因此流形框架能够更精细地考虑模型的优化问题.

2.4 几何观点下深度学习的两种基本理解

在数据集合构成嵌入空间的子流形这一假设下, 深度学习的模型和深度学习的工作原理主要有两种理解方式.

2.4.1 神经网络的流形降维理解方式

Brahma 等 [23] 认为, 深度学习模型之所以有着良好的性能是因为模型给复杂的数据集合做了数据降维. 以基于深度神经网络的分类器为例, 由于一般数据集数据量大、维度高、特征数多, 且同一标签的数据流形几何结构复杂并往往伴随着扭曲和缠绕, 不同标签的数据流形之间也会有大量的耦合. 因此传统机器学习中针对少量特征分类任务的、基于线性分类器的算法, 甚至层数较少的多层感知机, 在复杂的深度学习任务中往往都表现较差. 这正是因为其处理数据流形的能力较差, 往往只能应对较为简单的几何结构 (例如超平面、圆柱面). 因此 Brahma 认为, 基于深度神经网络的分类器具有将数据流形进行降维、解耦和平展的能力, 从而将复杂的数据流形降维成为简单的几何结构并完成分类任务.

这样的想法十分直观并且容易通过实验验证. Brahma 用一个简单样例模型验证了: 通过神经网络的学习数据流形的曲率逐渐下降, 并且不同标签的数据流形的耦合程度也在减轻. 但这一理解方式只能定性地去描述神经网络, 而缺乏对于数据流形的几何特征以及神经网络结构相关性质的描述, 因此难以对模型结构起到指导作用.

2.4.2 神经网络的学习流形间映射的理解方式

我们更进一步, 可将数据集和目标集一并视作流形, 并将神经网络视作数据流形到目标流形的映射. 在这一框架下, 神经网络的训练过程即为其拟合正确的从数据流形到目标流形的映射的过程, 而万能逼近定理给这一过程提供了原理性的保证. 与传统的解释所不同的是, 流形框架下额外考虑了数据流形和目标流形的拓扑结构, 而这样的拓扑信息会影响到神经网络的拟合过程甚至最终性能.

除此以外, 传统的神经网络逼近理论所基于的万能逼近定理 (定理 1) 虽然保证了深度神经网络可以逼近任意指定函数, 也足以应对大多数物体识别和物体检测的任务; 但是对于生成模型而言, 神经网络模型还需要学习到目标流形的分布. 以生成手写数字图像为例, 大多数深度学习模型往往会遇到模式崩塌 (mode collapse) 的问题, 即网络主要生成了某几类数字的图像而很少生成其他数字的图像, 这就说明神经网络仅仅逼近到了目标流形的映射, 而没有学习到其分布. 我们通过构造如下反例来证明传统的逼近理论不能完成这一任务.

定理3 令 μ 是 $[0,1]^m$ 上的 Lebesgue 测度, 存在形如定理 1 中 $F(x)$ 的一列函数 $\{f_n\}_{n=1}^{\infty}$ 和 $f\in C([0,1]^m)$, 使得 f_n 一致收敛到 f, 即 $f_n\rightrightarrows f$, 但存在 $B\subset\mathbb{R}$ 使得

$$f_{n\#}\mu(B)\nrightarrow f_{\#}\mu(B),\quad n\to\infty.$$

证明 不妨令 $m=1$. 取 $f(x)\equiv 0$, $f_n(x)=\frac{1}{n}\sin 2\pi x$, 故 f_n 形如定理 1 中的 $F(x)$, 且 $f_n\rightrightarrows f$. 取 $B=\{0\}$, 可见

$$f_{n\#}\mu(B)=\mu(f_n^{-1}(0))=0\nrightarrow f_{\#}\mu(B)=1.$$

定理获证.

而在流形框架下, 深度神经网络任务可以描述如下:

定义6 (深度神经网络) 令 $\mathcal{Z}$, $\mathcal{X}$ 分别为数据流形和目标流形, 分别有分布 μ, ν. 真实映射为 $f:\mathcal{Z}\to\mathcal{X}$ 并使得 $f_{\#}\mu=\nu$. 深度神经网络的参数集合记为 Θ, 神经网络对应 $f_\theta:\mathcal{Z}\to\mathcal{X}$, $\theta\in\Theta$. 通过训练神经网络使得

$$f_\theta\rightrightarrows f,\quad f_{\theta\#}\mu\to f_{\#}\mu=\nu. \tag{7}$$

定义 6 在保证逼近真实映射函数的同时, 在原理上保证了模型可以学习到真实数据分布. 并且因为这一框架考虑到了数据流形和目标流形的几何信息, 因此能够分析数据流形和网络结构对于模型性能的影响, 并能够直观分析神经网络模型的合理性. 因此本文主要通过神经网络学习流形间的映射这一观点, 来对深度神经网络的可解释性和对抗攻击进行原理的分析. 如下给出了两个深度神经网络通过流形框架解释的实例.

在流形间映射的理论框架下, Vincent 等 [24] 对于深度神经网络输入图像的降噪预处理进行了这样的解释:

例1 (图像降噪, Vincent 等 [24]) 神经网络的输入图像往往具有一定的噪声, 因此可以将具有噪声的图像 $\tilde{x}$ 看作真实图像 x 某邻域内的随机变量, 并且具有分布 $q_{\mathcal{D}}(\tilde{x}|x)$. 而图像的降噪过程则可视为将 $\tilde{x}$ 拉回数据流形 M.

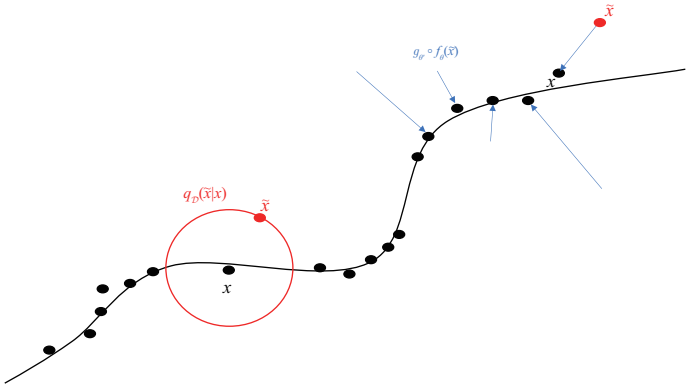

图 2 (网络版彩图) 神经网络降噪操作的过程 [24]. 真实数据 (黑点) 分布在一个低维的流形上, 具有噪声的输入图像 (红色圆) 是真实数据邻域内的随机变量. 降噪操作通过映射 $g_{\theta'} \circ f_\theta$ 将具有噪声的输入图像 $\tilde{x}$ (右侧红点) 映射到数据流形上 (蓝色箭头)

Figure 2 (Color online) The process of the denoising of neural networks [24]. Training data (black dot) lies near a low-dimensional manifold, input image with noise (red circle) is a random variable in the neighborhood of the training data. Denoising operation maps the input image $\tilde{x}$ (red dot on the right) with noise onto the manifold (blue arrow) by $g_{\theta'} \circ f_\theta$

具有噪声的输入图像往往距离真实数据和数据流形有一定距离. 且在降噪操作中噪声严重的图像要求更大的跳跃, 而噪声较小的图像则只需要小幅度的跳跃, 如图 2 所示.

除此以外, 这一理论框架还可以进行进一步的细化. 如 Lei 等 [25] 针对自动编码器 (autoencoder) 这一生成模型提出了如下观点:

例2 (自动编码器, Lei 等 [25]) 自动编码器分为编码器 (encoder) 和解码器 (decoder) 两部分. 由于流形局部同胚于一个欧氏空间, 因此编码器和解码器二者通过训练学习到数据流形和目标流形的参数表示和其逆映射, 再通过局部参数化实现数据流形和目标流形之间的映射. 具体而言, 编码器会通过对流形进行胞腔分解 (cell decomposition) 来实现数据流形的局部参数化, 随之得到其参数表示; 而解码器则会使用这一参数表示重构目标流形, 也即进行了胞腔分解的细分 (refinement).

如图 3 所示, 编码器在学习流形 M 的参数表示时对整个空间进行了胞腔分解, 解码器在重构目标流形时则对胞腔分解进行了细分. 因此自动编码器模型的性能不仅与编码器、解码器的胞腔分解能力有关, 更与流形所在空间的复杂程度有关. 据此, Lei 对于模型的学习能力和流形的学习难度进行了分析, 并给出了一定的判断依据.

3 流形框架下的神经网络分类器的可解释性与对抗攻击

本节将在流形框架下从 3 个方面对于深度神经网络的可解释性和对抗攻击进行原理分析, 分别是数据流形的几何结构、神经网络中的函数性质和当前神经网络的理论框架. 除此以外, 还将特别对生成对抗网络 (GAN) 进行分析和解释. 最后还将通过流形观点阐述数种模型的可能改进方法.

3.1 从流形的几何性质解释

定义 6 在流形框架下重新叙述了深度神经网络的学习任务, 即通过训练使神经网络逼近数据流形和目标流形的映射. 那么流形自身的几何性质将会影响最终神经网络的逼近结果, 也即影响到神经网络的性能.

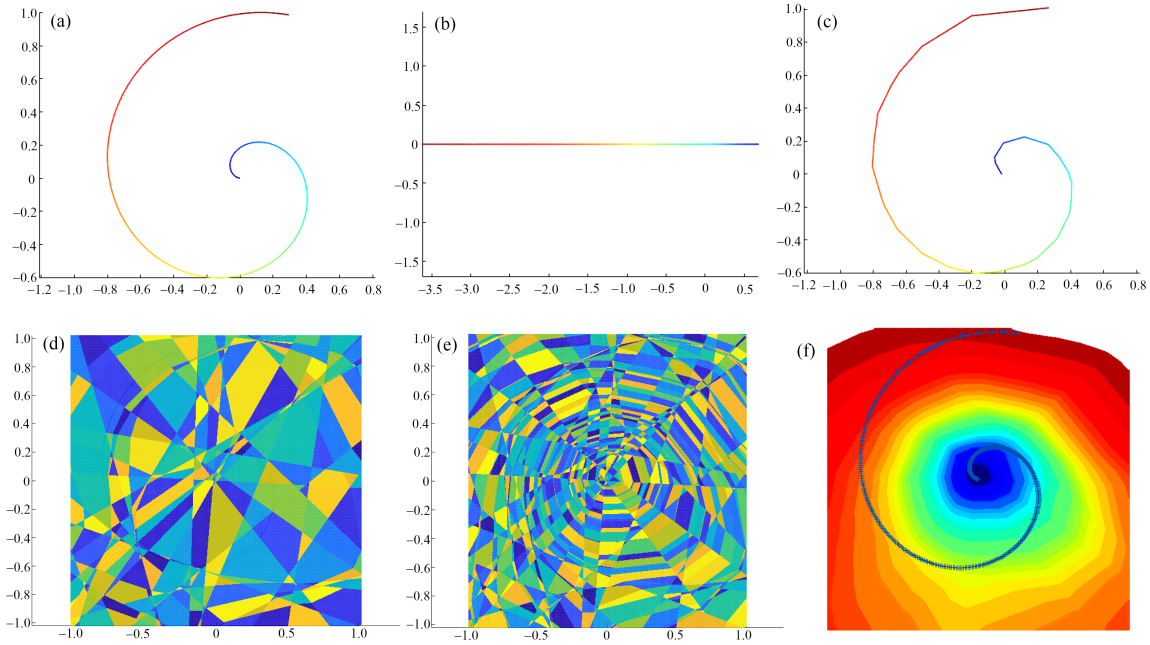

图 3 (网络版彩图) Lei 等 [25] 文中所展示的解码器和编码器学习螺旋线的参数表示的过程. (a) 输入的数据流形; (b) 流形的参数表示; (c) 解码器重构的目标流形; (d) 编码器学习参数表示时对空间的胞腔分解; (e) 解码器重构时对空间胞腔分解的细分; (f) 水平集

Figure 3 (Color online) The process of learning the parameterization of a helix of Autoencoders by Lei et al. [25]. (a) The input manifold; (b) the parameterization of the manifold; (c) the target manifold reconstructed by the decoder; (d) the cell decomposition of the encoder; (e) the refinement of the cell decomposition; (f) the level

3.1.1 数据流形光滑性差

真实数据所构成的流形往往有局部非光滑点出现, 并且在 2.4.1 小节中提到, 数据流形可能出现扭曲甚至自交, 而这对于神经网络分类器会产生极大的影响. 构造如下例子来展示非光滑点的影响.

例3 (流形非光滑点对函数光滑性的影响) 令 γ 为 1 – 流形 $\{(x,y)\in\mathbb{R}^2 \mid y=|x|\}$, 则 γ 分片光滑. 令 $f:\gamma\to\mathbb{R},(x,y)\mapsto y$, 则 f 是 γ 上的连续函数. 且 γ 有参数表示

$$\phi(t)=\begin{cases}(t,-t), & t\leqslant 0,\\ (t,t), & t>0.\end{cases}$$

记 $f(t)=f\circ\phi(t)$, 通过计算 f 在 $t=0$ 处的光滑性:

$$\lim_{t\to 0^-} f'(t)=-1,\quad \lim_{t\to 0^+} f'(t)=1,$$

有 f 在 $(0,0)$ 不光滑.

并且由于 f 的非光滑性, 模型在非光滑点附近往往有着较差的表现. 因此在目前的深度神经网络中仅仅通过对于测试数据的随机扰动即可起到降低模型正确率的攻击效果. Xie 等 [26] 根据这一观察, 在主流的对抗攻击算法中加入了随机扰动, 进一步提高了对抗攻击的效果和成功率. 这一想法虽然简单并且有着良好的表现, 但这样的假设却难以得到验证. 这是因为, 数据流形往往结构复杂, 且数据集只是这一流形上的离散采样, 而仅仅通过采样难以直接得到流形的几何性质, 也就无法更进一步去验证流形的光滑性.

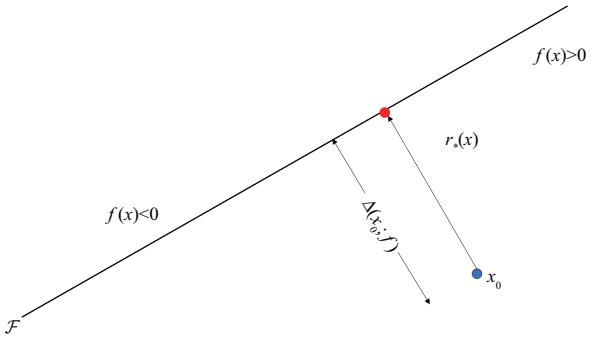

图 4 (网络版彩图) DeepFool [28] 算法生成对抗样本的过程

Figure 4 (Color online) The process of generating adversarial examples of DeepFool [28] algorithm

事实上, 随机扰动后的图像仍然满足数据集合所应具有的一定语义信息, 且的确可能出现在测试集中, 因此可以被视为数据流形中未被数据集所采样到的点. 由于嵌入空间的维度过大, 神经网络模型难以对这一部分未采样的数据进行泛化, 因此表现出鲁棒性差的问题. 而对于大多数对抗样本, Stutz 等 [27] 认为它们并没有分布在数据流形上, 而是分布在数据点的一些小邻域内. 因此将这些未在数据流形上的对抗样本 (off-manifold adversarial example) 投影在数据流形所获得的对抗样本, 也可以视作数据流形中未被采样到的点, 从而可以被用于数据增强来提高模型的泛化性能. 这一想法十分新颖地将对抗样本和流形理论相结合, 并通过简单的模型证明了神经网络的泛化性能和对数据流形上对抗样本的鲁棒性正相关. 但这一想法同样涉及到对数据流形的采样问题, 即不能证明对抗样本是否在流形之上.

3.1.2 数据流形维度高

根据流形观点的基本假设, 数据流形和目标流形是高维欧氏空间的嵌入流形. 由于嵌入空间的维度往往达到上万维, 流形的结构也会十分复杂. 以图片大小为 256×256 的图像分类问题为例, 数据流形可看作 $\mathbb{R}^{3\times256\times256}$ 中的嵌入流形. 但嵌入流形却往往有着较低的维度, 因此嵌入空间过高的维度会导致深度神经网络模型出现各种问题.

Moosavi-Dezfooli 等 [28] 根据这一特性从最朴素的角度提出了 DeepFool 的对抗攻击算法, 这一算法旨在通过对数据流形上的点添加扰动, 使其落在错误标签的决策边界内, 即可达到对抗攻击的目的. 具体而言, Moosavi-Dezfooli 对于 $\{\pm1\}$ 二分类问题的线性分类器给出了这样例子.

例4 (Moosavi-Dezfooli 等 [28]) 设有线性分类器函数 $f:\mathbb{R}^n\to\mathbb{R}, x\mapsto w^{\mathrm{T}}x+b$, 其预测标签为 $\hat{k}(x)\mathrm{sign}(f(x))$, 定义 $\mathcal{F}=\{x\in\mathbb{R}^n \mid f(x)=0\}$ 是水平集. 如图 4 所示, x_0 为数据流形中的某点, $\Delta(x_0,f)$ 等于 x_0 到决策边界 $\mathcal{F}$ 这一超平面的距离, 而改变分类器判断的最小扰动即为样本 x_0 在决策边界的垂直投影. 因此有

$$r_*(x_0)=\arg\min\{\|r\|_2 \mid \mathrm{sign}(f(x_0+r))\neq\mathrm{sign}(f(x_0))\}=-\frac{f(x_0)}{\|w\|_2^2}w.$$

更进一步地, Moosavi-Dezfooli 将这一模型推广至一般的分类器和任意的 l^p 模并得到了显著效果. 而从流形观点来解释这一算法可以看出, 正是数据流形相对于嵌入空间过低的维度, 导致分类器模型难以得到有效且显著的决策边界 (类比在 $\mathbb{R}^3$ 中学习到一个点的决策边界); 再加上高维空间所带来的“维度灾难”, 决策边界将有更大的概率出现过度靠近流形内数据点的情形, 也即更容易受到对抗样本

的影响. 因此, 通过加入对抗样本的对抗训练算法就显得尤为重要. 尽管对抗训练算法有诸如难以收敛、计算复杂度高等问题, 但如今仍成为提高模型鲁棒性的主要方法.

但 Moosavi-Dezfooli 等 [29] 从理论角度进一步指出, 对抗攻击并不会随着对抗训练中所使用的对抗样本的增加而失效; 相反地, 对抗样本会持续存在. Moosavi-Dezfooli 指出, 数据流形上的点在一个邻域内的决策边界之间存在一定的相关性, 而这也是随机扰动的对抗攻击算法性能劣于主流基于特定扰动的对抗攻击算法性能的主要原因. 因此对于数据流形上 n 个数据点 $\{x_1,\dots,x_n\}$, Moosavi-Dezfooli 考虑了如下矩阵:

$$N=\left[\frac{r(x_1)}{\|r(x_1)\|_2}\ \cdots\ \frac{r(x_n)}{\|r(x_n)\|_2}\right].$$

在二分类问题的线性分类器模型下, 矩阵 N 即为一个秩为 1 的矩阵. 在一般情形下通过对 N 进行奇异值分解可以得到嵌入空间中的一个低维子空间 $\mathcal{S}$, 而 $\mathcal{S}$ 包含了点集 $\{x_i\}_{i=1}^n$ 所在邻域内绝大多数垂直于决策边界的向量. 仅通过使用这一子空间 $\mathcal{S}$, 随机扰动的对抗攻击算法的成功率即可获得大幅提高. 因此 Moosavi-Dezfooli 认为, 通过使用这一集合, 即可针对模型持续地构造出对抗样本.

3.1.3 数据流形的特征空间复杂

由于数据流形的较高维度, 因此传统的分类需要通过神经网络对数据集进行特征的抽象和提取, 并根据所提取的特征来进行分类. 而由于嵌入空间的维度过大, 对于图像的特征划分方式纷繁复杂, 因此数据流形的特征空间十分复杂. Ilyas 等 [30] 对于二分类问题的特征进行了重新定义.

定义7 (分类问题的特征, Ilyas 等 [30]) 设数据流形为 $\mathcal{X}$, 标签集为 $\{\pm1\}$. 称 $f:\mathcal{X}\to\mathbb{R}$ 是一个特征, 如果 f 满足

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[f(x)]=0,\quad \mathbb{E}_{(x,y)\sim\mathcal{D}}[f^2(x)]=1,$$

其中 $\mathcal{D}$ 为数据对 $(x,y)\in\mathcal{X}\times\{\pm1\}$ 上的分布.

定义8 (有用特征、鲁棒特征和非鲁棒特征, Ilyas 等 [30]) 设 f 为定义 7 中的特征, 分别定义有用特征、鲁棒特征和非鲁棒特征如下:

• 称 f 是 ρ – 有用的 $(\rho>0)$, 如果 f 和正确的标签期望上正相关, 也即

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[y\cdot f(x)]\geqslant\rho.$$

• 记 $\Delta(x)$ 为某些特定类别的、针对 x 的有效扰动. 对 ρ – 有用特征 f, 称之为 γ – 鲁棒的, 如果有

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\inf_{\delta\in\Delta(x)} y\cdot f(x+\delta)\right]\geqslant\gamma,$$

也即特征 f 在一定的对抗扰动下仍然保持 γ – 有用.

• 称 f 为非鲁棒特征, 如果对某个 $\rho>0$ 而言 f 是 ρ – 有用的, 但对任意 $\gamma\geqslant0$ 来说 f 都不是 γ – 鲁棒的.

正是特征空间的复杂性, 才会导致无用特征、甚至有用但非鲁棒特征的出现. 这也解释了基于神经网络的分类器模型性能表现未达预期的原因: 一是因为模型没有很好地提取有用特征来进行分类; 二是因为模型过多地提取了非鲁棒特征导致鲁棒性较差. 而由非鲁棒特征的定义可知, 对抗样本即为对样本的非鲁棒特征进行扰动的数据点. 这一想法巧妙地将分类问题转化为线性分类器, 将问题简化的同时还对数据特征进行了分类和理解, 能够更好地帮助我们理解神经网络的工作原理. 但由于鲁棒

特征的提取需要用到一个训练好的鲁棒模型, 因此从严谨性的角度考虑, 使用这种方法提取出的鲁棒特征并不一定是 “鲁棒” 的.

Inkawhich 等 [12] 从类似的角度设计出针对基于神经网络的分类器的对抗攻击算法. 在神经网络的分类器中, 特征提取主要通过激活函数 (activation function) 来实现. 利用特征空间的复杂性、尤其是大量非鲁棒特征的存在, 算法使得标签为 A 的数据在某一激活函数下表现出标签为 B 的数据的特征, 并通过这样的特征 “攻击” 即可达到对抗攻击的目的. 更进一步地, 由于特征空间是数据流形的固有属性, 且基于神经网络的分类器拥有类似的特征提取层, Inkawhich 巧妙地通过已知的白盒模型进行特征 “攻击”, 再将生成的对抗样本直接用于攻击内部结构未知的黑盒模型, 并获得了较高的成功率. 这说明了对抗样本具有较高的迁移性, 神经网络会抽象并提取近似的特征, 也即不同的神经网络分类器有着类似的决策边界.

3.2 从神经网络表征能力解释

在传统的神经网络中, 大量使用的卷积函数、乘法函数、ReLU 函数均为线性函数或分段线性函数. 以 ReLU 函数和乘法函数为例:

$$\begin{aligned}&\mathrm{ReLU}:\mathbb{R}^n\to\mathbb{R}^n,\quad (x_1,\dots,x_n)\mapsto(\chi_{\{x_1>0\}}x_1,\dots,\chi_{\{x_n>0\}}x_n),\\&w:\mathbb{R}^n\to\mathbb{R},\quad x\mapsto w^{\mathrm{T}}x.\end{aligned}$$

而这样的线性性会导致模型缺乏正则化 (regulization): 设 $x\in\mathcal{X}$ 是数据流形上的点, η 是关于 x 一个有效扰动且满足 $\|\eta\|_\infty<\epsilon$, 新的输入图像记为 $\tilde{x}=x+\eta$. 当 ϵ 充分小时, 模型对于二者的判断应当是一致的. Goodfellow 等 [7] 以线性模型为例提出, 由于数据流形的嵌入空间往往维度较高, 在权重向量 w 的作用下有如下形式:

$$w^{\mathrm{T}}\tilde{x}=w^{\mathrm{T}}x+w^{\mathrm{T}}\eta.$$

因此激活函数 ReLU 的输入会改变 $w^{\mathrm{T}}\eta$, 而通过令 $\eta=\mathrm{sign}(w)$ 则这一改变将取到最大值, 也即 $\|w\|_1$. 所以同 3.1.2 小节类似的是, 当 x 的维度较大时, 即使对输入图像进行微小的改变也足以极大地影响激活函数 ReLU 的输入, 最终影响分类器的输出.

对于一般的神经网络分类器模型, Goodfellow 提出了一种基于 l^∞ 度量的对抗攻击算法 —— 快速梯度符号方法 (fast gradient sign method, FGSM). 与线性分类器模型基于乘法函数的 Jacobi 矩阵 w 类似, FGSM 通过对输入图片 x 求导, 得到模型的 Jacobi 矩阵 w 并计算其符号 $\eta=\mathrm{sign}(w)$, 即可生成输入图片 x 的对抗样本.

而由于神经网络的函数都是连续函数或分段光滑函数, 因此神经网络通过有限次训练得到的仍然是连续函数或分段光滑函数. 具体来说, 对于一个线性分类器在极端情形下会有这样的反例:

定理4 (连续函数分类器) 设 $\mathcal{X}$ 为数据流形, 称连续函数 f 为一个 $\{\pm1\}$ 分类器, 如果 $f:\mathcal{X}\to[0,1]^2, x\mapsto(f_{-1}(x),f_1(x))$, 其中 $f_i(x)$ 表示 x 属于标签 i 的概率且满足

$$f_{-1}(x)+f_1(x)=1,\quad \forall x\in\mathcal{X},$$

那么对于 1 – 流形 $\mathcal{X}=[0,1]$ 的二分类 $\{\pm1\}$ 问题, 存在可测集 $I\subset[0,1]$, 使得对于真实类别

$$y(x)=\begin{cases}1, & x\in I,\\ -1, & x\notin I,\end{cases}$$

和任意分类器函数 f, 都存在 $A\subset[0,1]$ 使得 $f_{y(x)}(x)\leqslant\frac{1}{2}$, 即 f 在 A 上分类错误, 且满足 A 的 Lebesgue 测度 $\mu(A)>0$.

证明 构造 $I\subset[0,1]$ 如下: 第 1 步从 $[0,1]$ 中间挖去长度为 $\frac{1}{4}$ 的区间, 以此类推在第 k 步从剩余的 2^{k-1} 个区间的中间分别挖去长度为 4^{-k} 的区间. 无限操作下去最终得到集合 I. 注意到 I 不包含任何区间, 且满足 $\mu(I)=\frac{1}{2}$.

反设存在在 $\mathcal{X}$ 上分类正确的分类器函数 f, 由 $[0,1]$ 区间紧致知 $f_{\pm1}(x)$ 在 $[0,1]$ 上一致连续. 由假设知 f 仅在零测集 Z 上分类错误. 记 $I'=I-Z$, 取 $x_0\in I'$, 由定义可知 x_0 属于标签 1, 则有 $f_1(x_0)>\frac{1}{2}$. 取 $\epsilon\in(0,f_1(x_0)-\frac{1}{2})$, 由一致连续知有 $\delta>0$ 使得 $\forall x\in(x_0-\delta,x_0+\delta)$ 都有 $|f_1(x)-f_1(x_0)|<\epsilon$, 也即 $f_1(x)>\frac{1}{2}$, 故 $(x_0-\delta,x_0+\delta)\subset I$, 与 I 不包含任何区间矛盾.

注释2 通常而言基于神经网络的二分类器 $f(x)=(f_{-1}(x),f_1(x))$ 在 $f_1(x)=f_{-1}(x)$ 时会采用随机的方法决定 x 的预测, 因此定理 4 将 $f_{y(x)}(x)=\frac{1}{2}$ 也称为分类错误.

定理 4 表明, 从理论上不能保证使用连续函数的分类器达到百分之百的正确率, 因此一味地只追求正确率并没有实际意义甚至适得其反. 且由于现有的基于连续函数的神经网络分类器存在理论的性能上限, 神经网络分类器的优化工作应集中于针对模型结构的优化. 但由于分类器模型的象空间较为简单, 且在实际应用中出现定理 4 这样极端情形的概率较低, 因此基于神经网络的分类器往往能表现出尚可的性能. 但对于象空间复杂的神经网络模型 (如生成模型), 过强的连续性和光滑性会极大地影响神经网络的性能, 本文将在第 4 节中详细讨论连续性对于生成对抗网络 (GAN) 的影响.

3.3 从深度学习理论框架解释

传统的对抗训练方法如数据增强、将对抗样本加入训练集等方法, 在提高模型鲁棒性的同时都会一定程度导致模型准确性的下降, 因此有人提出了正确性和鲁棒性相互制衡的观点. 根据 3.1.3 小节中的分析可知, 神经网络在进行对抗训练的过程中, 减少了对于数据流形非鲁棒特征的使用. 因此虽然模型的鲁棒性得到了提高, 但由于缺少了非鲁棒特征这类的 "有用" 信息, 分类器的性能自然会有所下降. Tsipras 等 [8] 在非鲁棒特征的基础上, 从理论角度证明了正确性和鲁棒性相互制约, 也即从理论上说明了现有深度神经网络模型在架构上不能同时保证高准确率和强鲁棒性.

Tsipras 考虑了 $\mathcal{X}=\mathbb{R}^{d+1}$ 上的 $\{\pm1\}$ 二分类问题. 假设数据对 $(x,y)\in\mathcal{X}\times\{\pm1\}$ 满足如下分布 (记为 $\mathcal{D}$):

$$\begin{aligned}&\mathbb{P}[y=+1]=\mathbb{P}[y=-1]=\frac{1}{2},\\&\mathbb{P}[x_1=+y]=p,\quad \mathbb{P}[x_1=-y]=1-p,\quad x_2,\dots,x_{d+1}\ \text{i.i.d.}\sim\mathcal{N}(\eta y,1),\end{aligned}\tag{8}$$

这里 $p\geqslant\frac{1}{2}$, $\eta=O(3/\sqrt{d})$. 可以看出, x_1 可以视作一个鲁棒特征, $x_2,\dots,x_{d+1}$ 均为非鲁棒特征. 考虑一个线性分类器

$$f(x)=\operatorname{sign}(w^{\mathrm{T}}x),\quad w=\left(0,\frac{1}{d},\dots,\frac{1}{d}\right),$$

则有

$$\mathbb{P}[f(x)=y]=\mathbb{P}\left[\frac{y}{d}\sum_{i=1}^{d}\mathcal{N}(\eta y,1)>0\right]=\mathbb{P}\left[\mathcal{N}\left(\eta,\frac{1}{d}\right)>0\right],$$

不妨取 $p=0.99$, $\eta=3/\sqrt{d}$, 因此有 $\mathbb{P}[f(x)=y]>0.99>p$, 即用非鲁棒特征能得到更高的正确率. 但反过来, 若在 x 上添加 l^∞ 度量下长度不超过 2η 的扰动则会得到

$$\min_{\|\delta\|_\infty\leqslant2\eta}\mathbb{P}[\operatorname{sign}(x+\delta)=y]\leqslant\mathbb{P}\left[\mathcal{N}\left(\eta,\frac{1}{d}\right)-2\eta>0\right]=\mathbb{P}\left[\mathcal{N}\left(-\eta,\frac{1}{d}\right)>0\right].$$

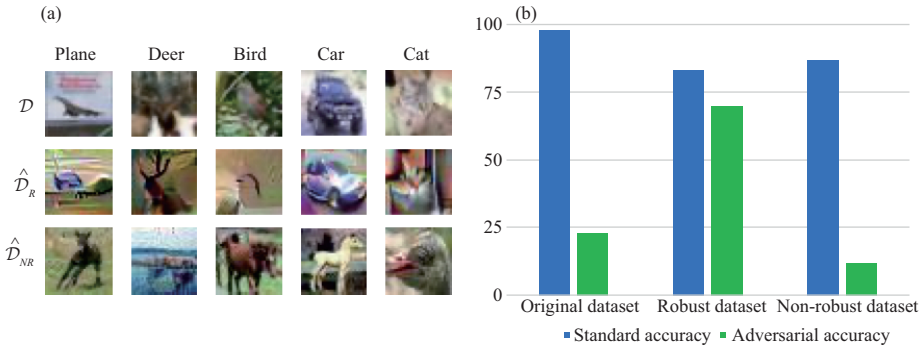

图 5 (网络版彩图) (a) 通过 CIFAR-10 [31] 提取鲁棒特征、非鲁棒特征所得到的数据集 [30], 从上到下依次为: 原始数据分布 $\mathcal{D}$; "鲁棒" 数据分布 $\widehat{\mathcal{D}}_R$; "非鲁棒" 数据分布 $\widehat{\mathcal{D}}_{NR}$. (b) 蓝色柱状图表示标准正确率, 绿色柱状图表示对抗攻击下的正确率, 从左到右分别为: 使用 $\mathcal{D}$ 的标准训练; 使用 $\widehat{\mathcal{D}}_R$ 的标准训练; 使用 $\widehat{\mathcal{D}}_{NR}$ 的标准训练

Figure 5 (Color online) (a) Dataset generating from CIFAR-10 [31] by extracting robust features and non-robust features: original dateset $\mathcal{D}$, robust dataset $\widehat{\mathcal{D}}_R$, and non-robust dataset $\widehat{\mathcal{D}}_{NR}$. (b) Standard (blue) and robust (green) accuracy trained with: $\mathcal{D}$, $\widehat{\mathcal{D}}_R$, and $\widehat{\mathcal{D}}_{NR}$

因此在对抗攻击下的正确率不超过 0.01. 更进一步地, Tsipras 证明了如下定理:

定理5 (正确性和鲁棒性折衷, Tsipras 等 [8]) 在式 (8) 中的分布 $\mathcal{D}$ 下, 任何正确率高于 $1-\epsilon$ 的分类器在 l^∞ 度量下长度不超过 2η 的扰动下, 其正确率至多是 $\frac{p}{1-p}\epsilon$.

注释3 值得注意的是, 当分类器的准确率趋近 1 ($\epsilon \to 0$), 对抗攻击下的准确率则会趋近 0, 也即对于现有的神经网络分类器, 不能同时保证高准确率和强鲁棒性.

与 Stutz 等 [27] 的观点所不同的是, Tsipras 整体考虑了对抗样本与模型正确率的关系. 由于对于一般分类任务而言无法判断对抗样本是否分布在流形上, 这一理论结果无疑更加传统也更具有指导作用. 对于基于神经网络的分类器模型而言, 现有的框架已经能较好地处理现实生活中的任务并且已经大规模地投入使用; 而对于更为复杂的生成模型, 理论框架的局限性则表现得更为明显和尖锐. 在第 4 节中会详细讨论生成对抗网络 (GAN) 下的理论框架挑战.

3.4 神经网络分类器的改进方法和其流形理解

3.4.1 使用鲁棒特征

3.1.3 小节提到, 尽管非鲁棒特征能够对分类器提供帮助, 但在一定的扰动下却缺乏鲁棒性. 因此 Ilyas 等 [30] 通过对于鲁棒特征的提取构建了一组 "鲁棒" 数据集, 并使用鲁棒数据重新训练模型; 作为对照, Ilyas 还通过提取非鲁棒特征构造了 "非鲁棒" 数据集并进行了同样的训练. 如图 5 [30,31] 所示, 虽然缺少非鲁棒特征对于分类器的帮助导致模型的准确性有所下降, 但是模型在对抗攻击下的正确率大幅提高, 这表明模型的鲁棒性得到了提高; 反观通过 "非鲁棒" 数据集训练的模型尽管仍具有较高的正确率, 但在对抗攻击的表现下甚至不如原始数据集. 这一实验验证了仅靠非鲁棒数据集也能够完成分类任务, 但模型的鲁棒性更多是来自数据集的鲁棒特征.

3.4.2 增加随机扰动

鉴于 3.1.1 小节中数据流形存在非光滑点的影响, 本小节考虑对训练数据增加随机扰动来提高光滑性. 我们构造如下例子来展示扰动的作用.

例5 (扰动对自交曲线的作用) 代数曲线 $\Gamma = \{(x,y) \in \mathbb{R}^2 \mid y^2 = x^3 + x^2\}$ 如图 6 所示, 在 $(0,0)$

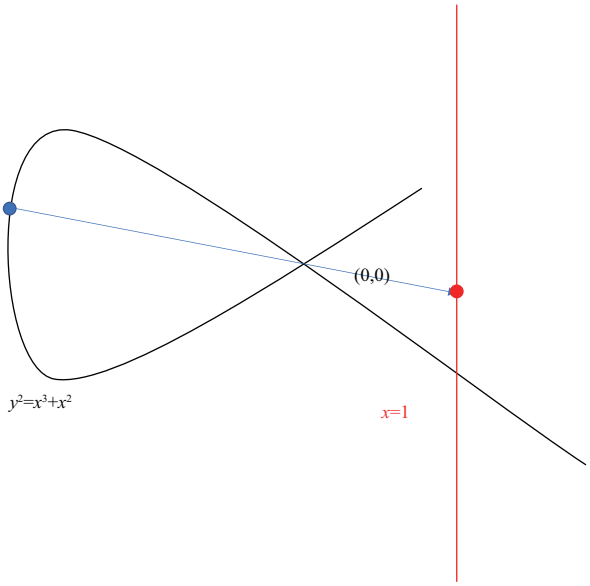

图 6 (网络版彩图) 扰动对自交曲线的作用

Figure 6 (Color online) Disturbation acts on a self-intersection curve

处有两条切线 $y \pm x = 0$. 通过扰动将曲线两支在 $(0,0)$ 处解开得到 Γ', 即可构造微分同胚

$$\phi(x,y) = \frac{y}{x}, \quad \phi^{-1}(t) = (t^2-1, t(t^2-1)),$$

即对任意 $(x,y) \in \Gamma$, 可以将其通过 $(0,0)$ 投影至直线 $x=1$ 上, 因此 Γ' 微分同胚于 $\mathbb{RP}^1$.

Xie 等 [32] 和 Laidlaw 等 [33] 根据这一观察提出了针对神经网络分类器的防御算法, 其本质思路均为通过对于输入图像的随机扰动来消除这些非光滑点的干扰. 在神经网络分类器中, 通过对训练数据的裁剪 (cropping)、翻转 (flipping)、填充 (padding)、调整尺寸 (resizing) 等方法, 即可提高模型的鲁棒性. 但从几何观点来看, 对于数据流形的随机扰动也可能带来新的自交点和非光滑点, 因此这一思路只能定性地描述随机扰动对于数据流形的作用, 而缺乏严谨的理论证明.

3.4.3 数据压缩

由于数据流形的特征空间极其复杂, 深度神经网络难以对整个特征空间进行精准的处理, 而诸多基于梯度扰动的对抗攻击方法也正是利用了这一特性. 并且根据 Goodfellow 等 [7] 的研究, 对于高维空间中的数据流形, 微小扰动会对神经网络产生巨大的影响. 因此 Gu 等 [34] 和 Dziugaite 等 [35] 分别提出了针对数据流形的压缩算法, 来达到减少扰动对于数据样本的影响.

从流形角度来看, 数据压缩过程即为 2.4.1 小节中所介绍的流形降维和平展的过程. 通过降低数据流形嵌入空间的维度并减少数据流形复杂程度, 其特征空间也能得以降维处理. 从另一方面考虑, 由于针对数据流形的扰动可以视作一个高维的随机变量. 因此根据强大数定理和中心极限定理, 数据压缩能够有效地降低扰动的幅度和随机性. 这里通过构造如下示例来进行说明:

例6 考虑一列随机变量 $\xi_1, \xi_2, \ldots, \xi_n$ i.i.d.$\sim \mathcal{N}(0,1)$, 并进行如下的数据压缩:

$$\eta_1 = \frac{1}{n}\left(\sum_{k=1}^{n} \xi_k\right), \quad \eta_2 = \frac{1}{\sqrt{n}}\left(\sum_{k=1}^{n} \xi_k\right),$$

则由强大数定理有如下结论:

$$\lim_{n\to\infty}\mathbb{P}[\eta_1=0]=1,$$

且由中心极限定理, 当 n 趋于无穷时, $\eta_2\sim\mathcal{N}(0,1)$. 因此可以表明, 数据压缩可以有效地降低扰动对于数据样本的影响, 从而得以提高模型的鲁棒性.

除去针对数据流形的数据压缩, 直接对于特征空间的压缩和提取也能够降低对抗攻击的成功率. Hinton 等 [36] 首先提出了 "馏化 (distillation)" 的方法, 将复杂的神经网络和其所学到的特征迁移至简单神经网络. Papernot 等 [37] 借助这一思想, 实现了对于特征的压缩, 并成功防御了微小扰动的对抗攻击.

3.4.4 减少线性函数的使用

为减轻 3.2 和 4.2 小节中过多使用线性函数对于神经网络鲁棒性的影响, Nayebi 等 [38] 仿照生物大脑的非线性树突的工作原理, 提出了使用高度非线性的激活函数以防御对抗攻击的策略. 具体而言, 为代替常用的 ReLU 激活函数, Nayebi 使用了如下的激活函数:

$$\phi(x)=\frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2}.$$

而 Krotov 等 [39] 也采用类似机制设计了对抗攻击的防御策略. 更进一步地, Lu 等 [40] 的研究表明, ReLU 激活函数的使用导致深度神经网络在进行例 2 的胞腔分解时, 会分割出测度小且几乎不包含数据点的区域 $\mathcal{D}$. 当神经网络的分类器函数在区域 $\mathcal{D}$ 上满足一定条件时, 这样的区域容易构造出对抗样本, 并可以通过 L-BFGS 方法进行对抗样本的构造和对神经网络的对抗攻击. 这也从理论上证明了 Nayebi 的防御策略的合理性和可行性.

但从流形观点来看, 根据 2.4.1 小节提出神经网络分类器所使用的卷积函数和线性函数起到了对数据流形的降维和平展的作用, 因此减少线性函数的使用与数据流形的降维平展是互相矛盾的. 而缺少降维操作将导致神经网络分类器难以取得令人满意的正确率, 因此这一思路并不能从原理上解决分类器模型鲁棒性差的问题.

3.4.5 降噪处理

3.4.3 小节介绍了通过数据压缩以降低扰动对于神经网络影响的防御策略, 本小节将介绍从另一角度降低扰动对模型鲁棒性的影响的策略. 由于 FGSM [7], L-BFGS [5] 这样基于扰动的对抗攻击算法需要严格控制扰动的幅度 (例如控制扰动在 l^0 范数、l^2 范数或 l^∞ 范数下的长度), 因此这样的扰动可被视作原始数据流形上的噪声. 那么通过使用如例 1 的图像降噪手段, 即可消除对抗样本的扰动噪声. 根据这一观察, Meng 等 [41] 和 Xie 等 [42] 分别提出通过神经网络重构原始数据流形和原始数据特征的防御方法. 而鉴于 GAN 相比传统深度神经网络在图像生成方面更为出色的表现, Lee 等 [43] 提出使用 GAN 来完成扰动后数据样本的鉴别的防御策略.

事实上, 如果将含有扰动噪声的对抗样本和真实样本视为两种图像风格, 并将用于消除扰动噪声的神经网络视为在进行两种图像风格的风格迁移, 我们可以将这一 "降噪" 处理方法拓展到所有类型对抗攻击的防御. Jin 等 [44] 通过这一想法设计出防御多种对抗攻击算法的数据预处理模型.

与 3.4.3 小节相比, 降噪处理在理论上不会导致新的自交点或非光滑点的出现. 这是因为, 当用于降噪的神经网络的能力足够强时, 这一神经网络只会处理数据点的噪声部分并重构出数据流形. 但受限于目前神经网络的学习和泛化能力, 神经网络的降噪能力还有待提高.

4 流形框架下神经网络生成模型的分析

与分类器模型所不同的是, 传统的基于神经网络的生成模型在学习隐空间到目标流形映射的同时还需要学习目标流形上的分布. 目前常见的神经网络生成模型包括: 变分自编码器 (VAE) [45]、生成对抗网络 (GAN) [46] 等. 而本文将研究和分析的重点放在目前最广为使用的 GAN 上. 相比于传统的生成模型, GAN 在使用生成器 (generator) 的神经网络的同时加入了另一个神经网络, Goodfellow 称之为判别器 (discriminator). 生成器在学习目标流形的分布的同时, 判别器会对生成器所学习的效果进行判定, 这种生成 – 对抗的过程, 能够有效地提高生成器的生成质量.

4.1 GAN 的基本理论

用流形观点来看 GAN 模型有如下定义.

定义9 记隐空间 (latent space) 和底空间 (ambient space) 分别为 $\mathcal{Z}$, $\mathcal{X}$, 记目标流形为 $M \subset \mathcal{X}$ 且满足 $\dim M \ll \dim X$. 定义 $\mathcal{Z}$, $\mathcal{X}$ 上的密度函数 $p_z(z)$, $p_x(x)$, 其中 p_x 满足 $\mathrm{supp}p_x \subseteq M$ 为真实分布, 记对应的分布 μ, ν, 即

$$\mu(U) = \int_U p_z(z)\mathrm{d}z, \quad \nu(V) = \int_V p_x(x)\mathrm{d}x.$$

记生成器的参数集合为 Θ, 生成器构造了映射 $g_\theta : \mathcal{Z} \to \mathcal{X}$, $\theta \in \Theta$, 生成器的目标为找到合适的参数 θ 使得 $g_{\theta\#}\mu = \nu$.

而对于给定生成器 G 和判别器 D, GAN 优化如下的损失函数:

$$V(G, D) = \mathbb{E}_{x\sim p_x(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]. \tag{9}$$

而 GAN 的训练过程则进行一个 minimax 过程, 即

$$\min_G \max_D V(G, D). \tag{10}$$

更进一步, Goodfellow 证明了定理 6 和 7.

定理6 (Goodfellow 等 [46]) 记 $p_g(x)$ 为生成器从 p_z 生成的分布, 则对于固定的 G, 最优判别器 D^* 满足

$$D_G^*(x) = \frac{p_x(x)}{p_x(x) + p_g(x)}.$$

定理7 (Goodfellow 等 [46]) 取定最优判别器 D_G^* 后, 记 $C(G) = V(G, D_G^*)$. 则 $C(G)$ 取到全局最小当且仅当 $p_g = p_x$, 且此时 $C(G)$ 达到最小值 $-\log 4$.

如图 7 所示, 从 (a) 到 (d) 的过程中, 判别器 D 从不具备分辨能力, 到收敛至最优判别器 $D_G^*(x) = \frac{p_x(x)}{p_x(x)+p_g(x)}$; 而随着判别器能力的增强, 生成器得到来自判别器的反馈, 其生成质量也在逐步提高. 当二者都达到最优时, 由于 $p_x = p_g$, 判别器不能区分样本来自生成器还是真实数据, 而生成器也会因得不到来自判别器的反馈停止更新其分布 p_g.

因此 GAN 在理论上能够完成针对目标流形上分布的学习任务, 而由定理 3, 这是传统的万能逼近定理 1 所不能保证的. 因此 GAN 模型一经提出就得到了广泛的关注和应用, 但随之而来 GAN 模型也暴露出了包括模式崩塌 (mode collapse)、训练过程中梯度消失等一系列问题. 后续章节将通过流形框架对 GAN 模型的若干问题进行理论解释和分析.

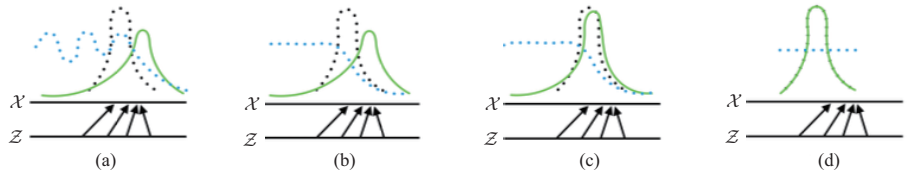

图 7 (网络版彩图) GAN 的训练过程. 通过更新判别器 D 的分布 (蓝色虚线) 使得其能够判别样本来自真实数据分布 p_x (黑色点线) 还是生成器生成分布 p_g(绿色实线). 下方的一组平行直线代表了在隐空间的采样 p_z, 在这里为均匀分布, 而向上的箭头则表明生成器 G 如何将 $z \in \mathcal{Z}$ 映射到 $x \in \mathcal{X}$, 同时生成分布 p_g

Figure 7 (Color online) The training process of GAN. By updating the discriminative distribution of D (blue dotted curve) so that it discriminates between the data distribution (black dotted curve) and the generating distribution p_g (green curve). The horizontal line below shows the sample of p_z in latent space, which is a uniform distribution in this case. The upward arrows imply how generator G maps $z \in \mathcal{Z}$ to $x \in \mathcal{X}$ and generates p_g at the same time

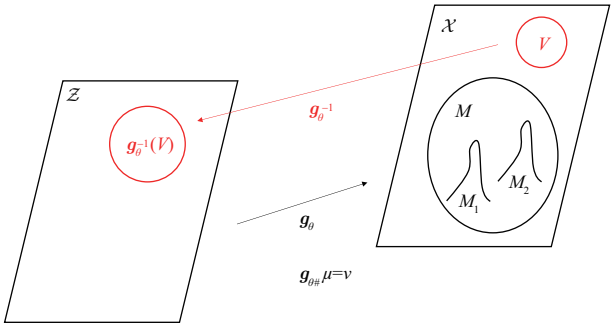

图 8 (网络版彩图) 连续函数难以学习多峰分布示意图

Figure 8 (Color online) Continuous function can hardly learns the multi-modal distribution

4.2 从 GAN 的表征能力解释

3.2 小节中已经提到神经网络过强的光滑性会影响分类器模型的性能, 而对于生成模型而言也有相同的问题. Yi 等 [47] 根据神经网络的连续性证明了如下结论, 证明的示意图参考图 8.

定理8 (多峰分布, Yi 等 [47]) 考虑目标流形 M 的多峰分布, 即 $M = M_1 \cup M_2$, $M_1 \cap M_2 = \emptyset$ 且 $\mathrm{supp}p_x \cap M_1 \neq \emptyset$, $\mathrm{supp}p_x \cap M_2 \neq \emptyset$. 不妨设 $\mathrm{supp}p_z = \mathcal{Z}$, 则不存在 Lipschitz 连续函数 $g_\theta : \mathcal{Z} \to \mathcal{X}$ 使得 $g_{\theta\#}\mu = \nu$.

注释4 定理中考虑 Lipschitz 连续函数是因为在深度神经网络中所用到的卷积函数、线性函数、ReLU 函数、tanh 函数、sigmoid 函数等都是 Lipschitz 连续的.

上述定理在理想情况下证明了, 生成器网络由于过强的连续性无法学到多峰分布. 这导致在目标流形有多个类别时 (如生成手写数字图像), 生成器所得到的分布 $g_{\theta\#}\mu$ 会无法遍历每一类别 (未生成某一数字类别的图像), 或是出现了目标流形分布 ν 以外的数据 (生成非数字的图像). 这种情况被称为模式崩塌 (mode collapse).

在介绍一般情形之前, 先给出如下定义.

定义10 (KL – 散度、JS – 散度) 给定概率密度函数 p_x, p_g, 定义其 KL – 散度 (Kullback-Leibler divergence) 为

$$\mathrm{KL}(p_x \| p_g) = \int p_x(x) \log \frac{p_x(x)}{p_g(x)} \mathrm{d}x. \tag{11}$$

定义其 JS – 散度 (Jensen-Shannon divergence) 为

$$\mathrm{JS}(p_x, p_g) = \mathrm{KL}(p_x \| p_A) + \mathrm{KL}(p_g \| p_A), \tag{12}$$

其中 $p_A = \frac{1}{2}(p_x + p_g)$.

那么在 GAN 模型中, 给定了最优判别器 D^* 后, 其损失函数 (9) 可化为

$$\begin{aligned} V(G, D^*) &= \mathbb{E}_{x\sim p_x(x)}[\log D^*(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1 - D^*(x))] \\ &= \int p_x(x) \log \frac{p_x(x)}{p_x(x) + p_g(x)} \mathrm{d}x + \int p_g(x) \log \frac{p_g(x)}{p_x(x) + p_g(x)} \mathrm{d}x \\ &= \int p_x(x) \left(\log \frac{p_x(x)}{p_A(x)} - \log 2 \right) \mathrm{d}x + \int p_g(x) \left(\log \frac{p_g(x)}{p_A(x)} - \log 2 \right) \mathrm{d}x \\ &= \mathrm{JS}(p_x, p_g) - \log 4. \end{aligned}$$

Arjovsky 等 [48,49] 表示, 由于底空间维数巨大, p_x 和 p_g 的支集在底空间交集测度过小, 则当 $p_x(x) \neq 0$, $p_g(x) = 0$ 时有

$$\begin{aligned} p_x(x) \log \frac{p_x(x)}{p_A(x)} &= p_x(x) \log \frac{p_x(x)}{\frac{1}{2} p_x(x)} = p_x(x) \log 2, \\ p_g(x) \log \frac{p_g(x)}{p_A(x)} &= p_g(x) \log \frac{p_g(x)}{\frac{1}{2} p_x(x)} = 0, \end{aligned}$$

同理当 $p_x(x) = 0$, $p_g(x) \neq 0$ 时有

$$\begin{aligned} p_g(x) \log \frac{p_g(x)}{p_A(x)} &= p_g(x) \log \frac{p_g(x)}{\frac{1}{2} p_g(x)} = p_g(x) \log 2, \\ p_x(x) \log \frac{p_x(x)}{p_A(x)} &= p_x(x) \log \frac{p_x(x)}{\frac{1}{2} p_x(x)} = 0, \end{aligned}$$

因此 $\mathrm{JS}(p_x, p_g) = \log 4$ 为常值. 定义判别器 D 的 Sobolev 范数为

$$\|D\| = \sup_{x\in\mathcal{X}} |D(x)| + \|\nabla_x D(x)\|_2.$$

则有结论:

定理9 (生成器梯度消失, Arjovsky 等 [49]) 令 $g_\theta : \mathcal{Z} \to \mathcal{X}$ 是光滑函数, D 是光滑的判别器函数, 记 p_g 的支集为 $\mathcal{P}$. 若 $\mathcal{P}$, M 不交且紧或存在 $x \in \mathcal{P} \cap M$ 使得在 x 处不横截, 则当 $\|D - D^*\| < \epsilon$ 且 $\mathbb{E}_{z\sim p_z}[\|J_\theta g_\theta(z)\|_2^2] \leqslant M^2$ 时有

$$\|\nabla_\theta \mathbb{E}_{z\sim p_z}[\log(1 - D(g_\theta(z)))]\|_2 < \frac{M\epsilon}{1 - \epsilon}. \tag{13}$$

从流形角度来看, 由于目标流形的维度远远小于底空间, 因此在绝大多数时候, GAN 都会出现生成器梯度消失的现象. Arjovsky 将传统 GAN 模型的 JS – 散度换作 Wasserstein 距离提出了 WGAN, 即

$$W(\mu_1, \mu_2) = \inf_{\gamma\in\prod(\mu_1,\mu_2)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|], \tag{14}$$

这里 $\prod(\mu_1,\mu_2)$ 为 $\mathcal{X}\times\mathcal{X}$ 上的联合分布, 且满足两个边缘分布分别为 μ_1 和 μ_2. Arjovsky 证明使用 Wasserstein 距离能够解决 GAN 中梯度消失的问题, 即优化的损失函数变为

$$V(G,D)=W(g_{\theta\#}\mu,\nu).$$

尽管 WGAN 改进和优化了传统的 GAN 模型, 但 Lei 等 [50,51] 通过最优传输理论指出, 由于理论上隐空间到目标流形的映射存在不连续的部分, 因此基于 GAN 的模型必然会出现模式崩塌的问题. 具体而言有:

定义11 (保测度映射, Villani 等 [21]) $X,Y\subset\mathbb{R}^d$ 是两个子集, 分别配有概率测度 μ, ν, 对应的密度函数为 $f(x)$, $g(y)$, 映射 $T:X\to Y$ 称为保测度的, 如果对 Y 中任意 ν – 可测集 B, $T^{-1}(B)$ 是 μ – 可测的, 且有 $\mu(T^{-1}(B))=\nu(B)$, 也即有 $T_{\#}\mu=\nu$.

定义12 (Monge 最优质量传输问题, Villani 等 [21]) 给定运输代价函数 $c:X\times Y\to\mathbb{R}$, Monge 最优质量传输问题需找到保测度映射 T, 使得最小化如下运输总代价:

$$\min_{T_{\#}\mu=\nu}\int_X c(x,T(x))\mathrm{d}\mu(x),$$

此时 T 被称为最优传输, 其运输总代价被称为 μ, ν 的 Wasserstein 距离, 也即

$$W_c(\mu,\nu)=\min_{T_{\#}\mu=\nu}\int_X c(x,T(x))\mathrm{d}\mu(x). \tag{15}$$

则传统 GAN 模型可归结为

定义13 (GAN 模型的最优传输, Lei 等 [51]) 给定参数集合 Θ, 求生成器映射 $g_\theta:\mathcal{Z}\to\mathcal{X}$, $\theta\in\Theta$, 使得 g_θ 是 $(\mathcal{Z},\mu)$ 到 $(\mathcal{X},\nu)$ 的最优传输.

Lei 进一步指出如下定理:

定理10 (最优传输的正则性, Lei 等 [51]) 令 $\Omega,\Lambda\subset\mathbb{R}^d$ 是两个有界开集, 令 $f,g:\mathbb{R}^d\to\mathbb{R}^+$ 为支集分别落在 Ω, Λ 的概率密度函数, 且满足

$$0<l_1<f(x)<r_1<\infty,\quad\forall x\in\mathrm{supp}f,$$
$$0<l_2<g(x)<r_2<\infty,\quad\forall x\in\mathrm{supp}g,$$

则对于最优传输 $T:\Omega\to\Lambda$, 存在相对闭集 $\Sigma_\Omega\subset\Omega$, $\Sigma_\Lambda\subset\Lambda$ 为零测集, 且满足对某个 $\alpha>0$ 而言, $T:\Omega\backslash\Sigma_\Omega\to\Lambda\backslash\Sigma_\Lambda$ 是 $C^{0,\alpha}_{\mathrm{loc}}$ 的同胚. 而 Σ_Ω 被称为 Ω 的奇异集.

如图 9 所示, Ω 的奇异集由两部分组成, 分别为

$$\Sigma_0=\{x_0,x_1\},\quad\Sigma_1=\bigcup_{k=0}^{3}\gamma_k.$$

最优传输映射 T 将 x_0 映射为 Λ 内部一个洞的边界, 将 x_1 映射成阴影三角形的 3 个顶点, 将每个 γ_k 映射成 Λ 凹陷部分的边界, 且 T 在 Σ_1, Σ_2 上不连续. 但由于神经网络只能生成连续函数, 因此这会导致以下 3 种情况出现:

(1) 训练过程不稳定且难以收敛;

(2) 若连续性生成器收敛到多峰分布中的某一连通部分, 此时生成器只会生成这一部分的样本;

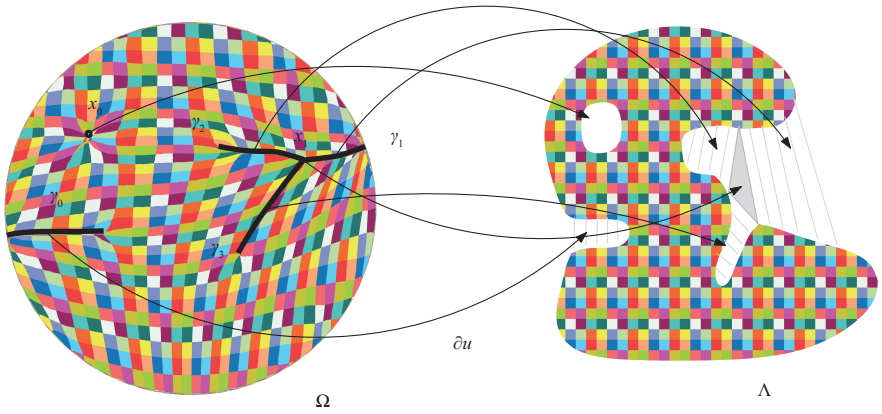

图 9 (网络版彩图) Lei 等 [51] 文中所展示的最优传输的奇异集 (左图的黑色部分). 由于象空间 Λ 非凸, 最优传输映射存在奇异集, 且最优传输映射在奇异集处不连续

Figure 9 (Color online) The singularity set of optimal transportation map shown by Lei et al. [51]. Because of the non-convexity of Λ, the optimal transportation map is discontinuous at the singularity set

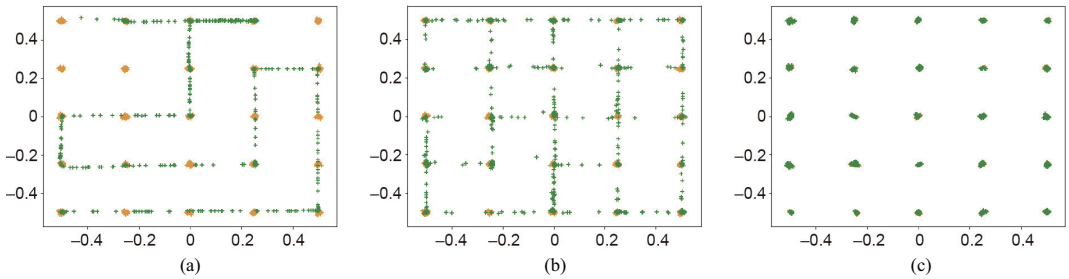

图 10 (网络版彩图) Lei 等 [51] 文中所展示的模式崩塌情形. 橙色点为真实分布, 绿色点为生成器所生成的分布. (a) GAN [46] 所生成的分布; (b) PacGAN [52] 所生成的分布; (c) 使用最优传输的 AE-OT [50] 所生成的分布

Figure 10 (Color online) Mode collapse shown by Lei et al. [51]. Orange dots are the data distribution, green dots are the generating distribution. The generating distribution by (a) GAN [46], (b) PacGAN [52], and (c) AE-OT [50]

(3) 虽然生成器遍历了目标流形, 但生成器会由于连续性生成目标流形以外的样本, 如图 10 [46, 50∼52] 所示.

最优传输理论首次完整地建立了 GAN 的几何理论框架, 对于 GAN 的任务目标、训练过程和模式崩塌问题都进行了系统的阐述和分析, 并从理论上证明了在现有 GAN 的框架内模式崩塌不能从根本上得到解决. 而使用最优传输理论对 GAN 进行分析与改进, 也成为现如今的热门方向.

4.3 从 GAN 的理论框架解释

尽管 GAN 在绝大多数应用中都表现出了优秀的性能, 但是 Arora 等 [53] 认为, 无论是使用 JS – 散度的传统 GAN 模型、还是使用 Wasserstein 距离的 WGAN 模型, 这二者的损失函数都会使模型缺少泛化性能. Arora 定义 GAN 的泛化性能如下:

定义14 (GAN 的泛化性能, Arora 等 [53]) 设 $g_{\theta\#}\mu, \nu$ 分别是生成器的分布和真实数据分布, $\hat{\nu}$ 为 ν 采样 m 个样本的观测分布 (empirical distribution). 如果下式在关于 $\widehat{g_{\theta\#}\mu}$ 的选择上大概率成立, 则

称 $g_{\theta\#}\mu$ 在度量 $d(\cdot,\cdot)$ 下具有泛化误差 ϵ:

$$|d(\widehat{g_{\theta\#}\mu},\widehat{\mu})-d(g_{\theta\#}\mu,\nu)|<\epsilon, \tag{16}$$

这里 $\widehat{g_{\theta\#}\mu}$ 为 $g_{\theta\#\mu}$ 在采样多项式级别个数的样本后的观测分布.

因此泛化性能是衡量观测分布和真实分布差距的一项重要指标, 对于生成模型来说, 它代表了模型生成新数据的能力强弱. Arora 证明, 使用 JS – 散度和 Wasserstein 距离会影响泛化性能.

定理11 (Arora 等 [53]) 令 $p(x)\sim\mathcal{N}(0,\frac{1}{d}I)$, $\widehat{p}(x)$ 是 $p(x)$ 采样 m 个样本的观测分布, 则有 $\mathrm{JS}(p,\widehat{p})=\log 2$, $W(p,\widehat{p})\geqslant 1.1$.

推论1 (Arora 等 [53]) 考虑 ν, $g_{\theta\#}\mu$ 具有概率密度函数 $p(x)$, 而 $\widehat{\nu}$, $\widehat{g_{\theta\#}\mu}$ 具有概率密度函数 $\widehat{p}(x)$, 则当取多项式级别个数的样本后有

$$W(g_{\theta\#}\mu,\nu)=0,\quad W(\widehat{g_{\theta\#}\mu},\widehat{\mu})>1,$$

这和式 (16) 矛盾.

推论2 (Arora 等 [53]) 考虑 ν 具有概率密度函数 $p(x)$, 而 $\widehat{\nu}$, $g_{\theta\#}\mu$ 具有概率密度函数 $\widehat{p}(x)$, 即生成器 G 记住了 $\widehat{\nu}$ 中的所有样本. 在此情况下, 由于 G 的样本空间为至多可数集, 即具有离散分布, 在足够的采样下其观测分布 $\widehat{g_{\theta\#}\mu}$ 可近似看作 $g_{\theta\#}\mu$, 因此有

$$W(g_{\theta\#}\mu,\nu)>1,\quad W(\widehat{g_{\theta\#}\mu},\widehat{\mu})\approx 0,$$

这同样和式 (16) 矛盾.

上述两个推论说明: 即使在采样样本足够多的情况下, 使用 Wasserstein 距离和 JS – 散度仍可能导致生成器 G 缺乏泛化能力. 而这一由采样导致的问题在流形的几何结构复杂时则更为尖锐. 因此这一结果从另一方面表明了包括数据增强、数据正则化等预处理操作的重要性.

4.4 GAN 的基于最优传输的改进策略

通过最优传输理论, 并借助如下定理:

定理12 (Brenier [54]) 设 $X,Y=\mathbb{R}^n$ 且分别具有分布 μ, ν, 运输损失函数为 $c(x,y)=|x-y|^2$. 如果分布 μ 绝对连续且 μ, ν 有有限二阶中心矩, 即 $\int_X|x|^2\mathrm{d}\mu(x)+\int_Y|y|^2\mathrm{d}\nu(y)<\infty$. 则存在凸函数 $u:X\to\mathbb{R}$ 使得梯度函数 ∇u 给出 Monge 最优质量传输问题 (定义 12) 的唯一解. 这里 u 被称为 Brenier 势能函数. 设 u 是 C^2 光滑, 则 Monge 最优质量传输问题可化为如下偏微分方程 (PDE):

$$\det\left(\frac{\partial^2 u}{\partial x_i\partial x_j}\right)(x)=\frac{\mu(x)}{\nu\circ\nabla u(x)}. \tag{17}$$

定理13 (Gu 等 [55]) 令 $\Omega\subset\mathbb{R}^n$ 紧凸区域并有分布 μ, $\{y_i\}_{i=1}^k$ 是 $\mathbb{R}^n$ 中两两不同的点集. 则对任意 $v_1,v_2,\ldots,v_k>0$, $\sum_{i=1}^k v_i=\mu(\Omega)$, 都存在 $(h_1,\ldots,h_k)\in\mathbb{R}^k$, 在相差常向量 $(c,\ldots,c)$ 意义下唯一, 使得函数

$$u_h(x)=\max\{\langle x,y_i\rangle+h_i\mid i=1,\ldots,k\}.$$

满足: 记 $w_i(h)=\mu(\{x\in\mathbb{R}^n\mid\nabla u_h(x)=y_i\}\cap\Omega)$, 则 $w_i(h)=v_i$. 此时 ∇u_h 最小化下述二次误差:

$$\min_{T_\#\mu=\nu}\int_\Omega|x-T(x)|^2\mathrm{d}\mu(x),$$

其中 ν 为 Dirac 测度 $\nu=\sum_{i=1}^k v_i\delta(y-y_i)$.

Lei 等 [50,51] 将神经网络的训练问题转化为一个偏微分方程或是凸几何问题. 而通过计算式 (17) 或解定理 13 中的凸优化问题, 即可得到隐空间到目标流形的最优传输, 也即得到了最优的生成器 G, 见图 10. 这一思路不仅无需针对 GAN 模型进行复杂繁琐的训练, 更重要的是能够得到非连续的最优传输映射. 因此理论上使用最优传输理论能够彻底解决 GAN 模型训练难度大和模式崩塌的问题. 但就目前而言, GAN 模型到最优传输问题的转化仍难以处理, 且使用现有理论只能得到最优判别器而非最优生成器. 因此尽管最优生成器和最优判别器在理论上是等价的, 但这一方法目前还无法应用在实际任务中.

5 深度模型流形解释的挑战和发展方向

2.3 小节提到将数据集看作一个高维欧氏空间的低维子流形, 并使用流形的诸多结论来描述深度学习的模型. 这样的理论框架相比于传统的深度学习解释理论, 能够更深入地分析数据流形和模型结构对于模型性能的影响, 由此可对对抗攻击算法背后的理论依据和深度学习现如今所遇到的种种问题进行更深入地剖析和理解. 但与此同时, 这样的流形框架还不完善, 还有不少挑战需要进一步解决, 也暴露出诸多的缺陷. 本节将对于流形框架需要进一步解决的挑战和可能的拓展进行阐述.

5.1 流形性质的验证

Seung [22] 提出从流形视角来理解图像集以来, 这一观点逐渐成为了大家的共识. 但一直以来, 这一假设都缺乏足够的实验论证. 比如考虑尺寸为 256×256 的真实图像, Seung 以相机的位置连续变化、光线的连续变化得到了一族真实图像为例, 说明这一族照片构成了一个低维流形. 但目前的流形框架更加庞大且复杂, 它将所有的真实图像视为 $\mathbb{R}^{3\times256\times256}$ 中的一个低维子流形, 并在此基础上借由流形理论对深度神经网络进行原理分析.

由于对于绝大多数的图像任务, 其嵌入空间的维度较大 (如 CIFAR-10 [31] 和 ImageNet [56]), 现如今还难以通过实验验证光滑性、连续性等这些基本的几何结构和几何性质, 因此现如今在流形框架下的实验 [23,25,50,51] 都是通过人为构造出的流形实例来佐证其理论结果. 所以现如今深度学习流形框架的研究重点之一, 就是通过流形理论来探索主流数据集和数据流形的几何性质, 如流形维度、光滑性、曲率等基本信息. 更进一步地, 参考本文所提到的数据流形的特征空间和 GAN 模型中非凸性导致的模式崩塌, 我们可以通过流形理论对数据流形有关鲁棒性和凸性进行定量的评估与分析, 再借由几何方法对数据集进行预处理操作和修复工作.

5.2 流形框架对模型结构的指导作用

自从深度神经网络提出以来, 关于参数的调整和模型的选择就一直是一大难题. 由于传统的深度学习理论解释难以对于模型结构和参数调整进行有效的指导和分析, 因此如今人们往往依赖经验性的选择和调整, 并形成了一系列的关于参数调整和模型选择技巧. 虽然本文所介绍的流形框架通过将神经网络的函数视为流形间的映射、将损失函数视为分布间的距离, 一定程度上能够对模型和参数选择的合理性进行判断和指导, 但由于目前理论尚未完善以及数据构成的流形无法定量描述等原因, 目前深度学习的流形理论还不能对于模型选择和参数的优劣进行定量的判断.

因此现如今, 除了借助深度学习流形解释框架来改进深度神经网络模型的结构, 研究重点也应放在用流形的观点去指导设计对抗攻击的鲁棒性、准确率等系统指标. 而由于本文中所提到的诸多现有深度学习理论框架的固有弊病, 建立流形视角下神经网络模型的统一评价体系, 也将推动用几何框架

去设计新一代可解释的深度网络.

5.3 流形理论的进一步完善

现如今的流形理论虽然仅仅是起步阶段, 但相比于之前的深度学习理论解释已经体现出其优势所在. 流形框架通过研究数据流形的几何性质和神经网络的函数性质来分析神经网络模型的最终性能, 并在流形结构上融合了统计的观点, 因而能够在传统的万能逼近定理上进一步研究和分析神经网络模型. 但目前流形框架也存在着一定的问题和缺陷, 例如缺乏对卷积函数、ReLU 函数等网络逐层函数的分析, 以及缺乏对模型训练过程的解释与优化, 这些问题可能仅凭流形理论难以得到很好的解决.

因此, 为了进一步发展和细化神经网络的流形理论, 我们需要参考传统的机器学习和深度学习的理论解释模型, 并且结合包括 2.1.2 小节动力系统的理论、计算数学中各种数值算法和优化的思想, 取长补短地建立起一套成熟且完备的深度学习流形理论体系. 只有这样才能更深入地探究深度学习这一“黑盒” 工具, 才能反作用于神经网络模型的结构设计. 并且考虑到现如今学界对于深度学习流形框架的理解各不相同甚至互相冲突, 因此, 十分有必要建立一个统一的理论解释框架和测评平台.

6 总结

本文通过总结和完善现有的深度学习的流形解释思想构建了其流形解释框架, 在介绍目前主流的深度神经网络的 3 种理解框架的同时, 通过构造定理 3 的反例, 阐述了流形框架相比于传统基于万能逼近定理的优化视角的优势所在.

通过深度学习的流形解释框架, 本文对于深度神经网络任务进行重新整理、叙述和分析 (参见定义 6), 即深度学习通过训练, 在使得神经网络逼近流形间的真实映射的同时还需保证这一映射是保测度的. 在这一理论基础上, 本文总结了当今较为成功的主流对抗攻击算法, 并基于流形视角和构造例 3 和定理 4 这样的示例来理解这些对抗攻击算法背后所蕴含的流形观点, 即真实数据流形具有难以被神经网络模型有效处理的复杂特征空间; 并且由于数据流形光滑性较差但神经网络使用的函数光滑性较好, 因此模型往往会出现正确率低于预期和较差的鲁棒性. 在分析神经网络的工作原理和所遇到诸多问题的成因后, 本文还对于目前广泛应用的对抗生成网络模型进行了流形框架下的理解 (参见定义 9), 并通过最优传输理论对这一模型进行了重新的阐述 (参见定义 13). 针对生成对抗网络所遇到的以模式崩塌为代表的尖锐问题, 本文主要介绍了通过最优传输及其正则性理论的理论分析成果, 并得出这一问题通过现有深度学习理论和神经网络模型无法彻底解决的结论.

除此以外, 本文整理了目前较为成功的深度神经网络的改进方法, 并基于流形框架给出了包括例 5 在内的理论解释和原理分析, 即通过对于数据流形的随机扰动能够提高数据流形的光滑性, 以达到提高模型性能和鲁棒性的目的. 本文还简要介绍了在最优传输理论中, 通过计算生成对抗网络中隐空间到目标流形的最优传输映射的方法. 这一方法无需训练模型即可直接得到最优生成器, 并且从理论上能够避免模式崩塌问题.

本文的最后还对于深度神经网络流形框架目前所面临的挑战进行了阐述, 并对未来可能的拓展进行了展望. 现如今流形框架的基本假设还无法得到验证, 并且由于这一理论仍缺乏更多更深刻的理论成果, 还难以对于优化和设计神经网络模型起到重大的指导作用.

致谢 感谢大连理工大学的雷娜教授提供本文中使用的图 3, 9, 10.

参考文献

1 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Commun ACM, 2017, 60: 84–90
2 Ren S Q, He K M, Girshick R B, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 1137–1149
3 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations, San Diego, 2015
4 Reed S, Akata Z, Yan X, et al. Generative adversarial text-to-image synthesis. In: Proceedings of the 33rd International Conference on Machine Learning, New York, 2016. 1060–1069
5 Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. In: Proceedings of the 2nd International Conference on Learning Representations, Banff, 2014
6 Yuan X Y, He P, Zhu Q L, et al. Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Netw Learn Syst, 2019, 30: 2805–2824
7 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the 3rd International Conference on Learning Representations, San Diego, 2015
8 Tsipras D, Santurkar S, Engstrom L, et al. Robustness may be at odds with accuracy. In: Proceedings of the 7th International Conference on Learning Representations, New Orleans, 2019
9 Tramér F, Papernot N, Goodfellow I J, et al. The space of transferable adversarial examples. 2017. ArXiv:1704.03453
10 Zhang T Y, Zhu Z X. Interpreting adversarially trained convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, Long Beach, 2019. 7502–7511
11 Dong Y P, Liao F Z, Pang T Y, et al. Discovering adversarial examples with momentum. 2017. ArXiv:1710.06081
12 Inkawhich N, Liang K, Carlin L, et al. Transferable perturbations of deep feature distributions. In: Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, 2020
13 Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control Signal Syst, 1989, 2: 303–314
14 Lu Y P, Zhong A X, Li Q Z, et al. Beyond finite layer neural networks: bridging deep architectures and numerical differential equations. In: Proceedings of the 35th International Conference on Machine Learning, Stockholm, 2018. 3282–3291
15 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 770–778
16 Gastaldi X. Shake-shake regularization. 2017. ArXiv:1705.07485
17 Cheng B, Titterington D M. Neural networks: a review from a statistical perspective. Statist Sci, 1994, 9: 2–30
18 Lee M J. Introduction to Topological Manifolds. Berlin: Springer, 2006
19 Milnor W J. Topology from the Differentiable Viewpoint. Berlin: Springer, 1998
20 Evans L C, Gariepy R F. Measure Theory and Fine Properties of Functions. Roca Raton: CRC Press, 1992
21 Villani C. Topics in Optimal Transportation. Providence: AMS, 2003
22 Seung H S. COGNITION: the manifold ways of perception. Science, 2000, 290: 2268–2269
23 Brahma P, Wu D P, She Y Y. Why deep learning works: a manifold disentanglement perspective. IEEE Trans Neural Netw Learn Syst, 2016, 27: 1997–2008
24 Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference, Helsinki, 2008. 1096–1103
25 Lei N, Luo Z X, Yau S T, et al. Geometric understanding of deep learning. 2018. ArXiv:1805.10451
26 Xie C H, Zhang Z S, Wang J Y. Improving transferability of adversarial examples with input diversity. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 2730–2739
27 Stutz D, Hein M, Schiele B. Disentangling adversarial robustness and generalization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 6976–6987
28 Moosavi-Dezfooli S, Fawzi A, Frossard, P. DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 2574–2582
29 Moosavi-Dezfooli S, Fawzi A, Fawzi O. Universal adversarial perturbations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 1765–1773
30 Ilyas A, Santurkar S, Tsipras D, et al. Adversarial examples are not bugs, they are features. In: Proceedings of

Advances in Neural Information Processing Systems, Vancouver, 2019. 125–136
31 Krizhevsky A. Learning Multiple Layers of Features from Tiny Images. University of Toronto Technical Report TR-2009, 2012
32 Xie C H, Wang J Y, Zhang Z S. Mitigating adversarial effects through randomization. In: Proceedings of the 6th International Conference on Learning Representations, Vancouver, 2018
33 Laidlaw C, Feizi S. Functional adversarial attacks. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2019. 10408–10418
34 Gu S X, Rigazio L. Towards deep neural network architectures robust to adversarial examples. 2015. ArXiv:1412.5068
35 Dziugaite G K, Ghahramani Z, Roy D M. A study of the effect of JPG compression on adversarial images. 2016. ArXiv:1608.00853
36 Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network. 2015. ArXiv:1503.02531
37 Papernot N, McDaniel P D, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proceedings of IEEE Symposium on Security and Privacy, San Jose, 2016. 582–597
38 Nayebi A, Ganguli S. Biologically inspired protection of deep networks from adversarial attacks. 2017. ArXiv:1703.09202
39 Krotov D, Hopfield J J. Dense associative memory is robust to adversarial inputs. Neural Comput, 2018, 30: 3151–3167
40 Lu J J, Issaranon T, Forsyth D A. SafetyNet: detecting and rejecting adversarial examples robustly. In: Proceedings of ACM SIGSAC Conference on Computer and Communications Security, Dallas, 2017. 446–454
41 Meng D Y, Chen H. MagNet: a two-pronged defense against adversarial examples. In: Proceedings of ACM SIGSAC Conference on Computer and Communications Security, 2017. 135–147
42 Xie C H, Wu Y X, Maaten L, et al. Feature denoising for improving adversarial robustness. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 501–509
43 Lee H, Han S, Lee J. Generative adversarial trainer: defense to adversarial perturbations with GAN. 2017. ArXiv:1705.03387
44 Jin G Q, Shen S W, Zhang D M, et al. APE-GAN: adversarial perturbation elimination with GAN. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, 2019. 3842–3846
45 Diederik P K, Max W. Auto-encoding variational Bayes. In: Proceedings of the 2nd International Conference on Learning Representations, Banff, 2014
46 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 2672–2680
47 Yi R, Xia M F, Liu Y J, et al. Line drawings for face portraits from photos using global and local structure based GANs. IEEE Trans Pattern Anal Mach Intell, 2020. doi: 10.1109/TPAMI.2020.2987931
48 Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks. In: Proceedings of the 5th International Conference on Learning Representations, Toulon, 2017
49 Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. 2017. ArXiv:1701.07875
50 Lei N, An D S, Guo Y, et al. A geometric understanding of deep learning. Engineering, 2020, 6: 361–374
51 Lei N, Guo Y, An D S, et al. Mode collapse and regularity of optimal transportation maps. 2019. ArXiv:1902.02934
52 Lin Z, Khetan A, Fanti G, et al. PacGAN: the power of two samples in generative adversarial networks. In: Proceedings of Advances in Neural Information Processing Systems, Montréal, 2018. 1505–1514
53 Arora S, Ge R, Liang Y Y, et al. Generalization and equilibrium in generative adversarial nets (GANs). In: Proceedings of the 34th International Conference on Machine Learning, Sydney, 2017. 224–232
54 Brenier Y. Polar factorization and monotone rearrangement of vector-valued functions. Comm Pure Appl Math, 1991, 44: 375–417
55 Gu X F, Luo F, Sun J, et al. Variational principles for minkowski type problems, discrete optimal transport, and discrete monge-ampere equations. 2013. ArXiv:1302.5472
56 Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis, 2015, 115: 211–252

Adversarial attack and interpretability of the deep neural network from the geometric perspective

Mengfei XIA, Zipeng YE, Wang ZHAO, Ran YI & Yongjin LIU*

Department of Computer Science and Technology, Tsinghua University, Beijing 100084*, China*
* Corresponding author. E-mail: liuyongjin@tsinghua.edu.cn

Abstract Deep learning has achieved significant success in various engineering fields. However, its drawback has also received considerable attention recently, i.e., it suffers from poor interpretability, weak robustness and difficulty for network training, which seriously affect the security and usability of deep neural networks. Therefore adversarial attacks and interpretability become the focuses of the next generation of artificial intelligence research. In this paper, we survey recent works on them from a novel geometric perspective. We reformulate the problems in traditional deep learning models from the viewpoint of manifold theory, and summarize several strategies for possible optimization of the deep networks based on interpretability. Finally, we state several challenges on the interpretability from manifold theory and outline possible future directions.

Keywords deep learning, adversarial attack, interpretability, manifold

Mengfei XIA is a Ph.D. student at the Department of Computer Science and Technology, Tsinghua University. He received his B.Sci. degree from the Department of Mathematical Science, Tsinghua University, China, in 2020. He won the silver medal twice in 30th and 31st Chinese Mathematical Olympiad. His research interests include mathematical foundation in deep learning, image processing, and computer vision.

Zipeng YE is a Ph.D. student at the Department of Computer Science and Technology, Tsinghua University. He received his B.Eng. degree from Tsinghua University, China, in 2017. His research interests include computational geometry and computer vision.

Wang ZHAO is a Ph.D. student at the Department of Computer Science and Technology, Tsinghua University. He received his B.Eng. degree from Tsinghua University, China, in 2019. His research interest focuses on 3D computer vision.

Yongjin LIU is a professor at the Department of Computer Science and Technology, Tsinghua University, China. He received his B.Eng. degree from Tianjin University, China, in 1998, and his Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong, China, in 2004. His research interests include computational geometry, computer vision, and computer graphics.