中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2023)04-0963-21

论文引用格式:Jiang L Y, Zheng Y F, Chen C, Li G H and Zhang W J. 2023. Review of optimization methods for supervised deep learning. Journal of Image and Graphics, 28(04):0963-0983(江铃燚, 郑艺峰, 陈澈, 李国和, 张文杰. 2023. 有监督深度学习的优化方法研究综述. 中国图象图形学报, 28(04):0963-0983)[DOI:10.11834/jig.211139]

有监督深度学习的优化方法研究综述

江铃燚[1,2],郑艺峰[1,2*],陈澈[1,2],李国和[3],张文杰[1,2]

1. 闽南师范大学计算机学院,漳州　363000; 2. 数据科学与智能应用福建省高校重点实验室,漳州　363000;
3. 中国石油大学信息科学与工程学院,北京　102249

摘　要: 随着大数据的普及和算力的提升,深度学习已成为一个热门研究领域,但其强大的性能过分依赖网络结构和参数设置。因此,如何在提高模型性能的同时降低模型的复杂度,关键在于模型优化。为了更加精简地描述优化问题,本文以有监督深度学习作为切入点,对其提升拟合能力和泛化能力的优化方法进行归纳分析。给出优化的基本公式并阐述其核心;其次,从拟合能力的角度将优化问题分解为3个优化方向,即收敛性、收敛速度和全局质量问题,并总结分析这3个优化方向中的具体方法与研究成果;从提升模型泛化能力的角度出发,分为数据预处理和模型参数限制两类对正则化方法的研究现状进行梳理;结合上述理论基础,以生成对抗网络(generative adversarial network,GAN)变体模型的发展历程为主线,回顾各种优化方法在该领域的应用,并基于实验结果对优化效果进行比较和分析,进一步给出几种在GAN领域效果较好的优化策略。现阶段,各种优化方法已普遍应用于深度学习模型,能够较好地提升模型的拟合能力,同时通过正则化缓解模型过拟合问题来提高模型的鲁棒性。尽管深度学习的优化领域已得到广泛研究,但仍缺少成熟的系统性理论来指导优化方法的使用,且存在几个优化问题有待进一步研究,包括无法保证全局梯度的Lipschitz限制、在GAN中找寻稳定的全局最优解,以及优化方法的可解释性缺乏严格的理论证明。

关键词: 机器学习;深度学习;深度学习优化;正则化;生成对抗网络(GAN)

Review of optimization methods for supervised deep learning

Jiang Lingyi[1,2], Zheng Yifeng[1,2*], Chen Che[1,2], Li Guohe[3], Zhang Wenjie[1,2]

1. College of Computer Science, Minnan Normal University, Zhangzhou　363000, China; 2. Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Zhangzhou　363000, China;
3. College of Information Science and Engineering, China University of Petroleum, Beijing　102249, China

Abstract: Deep learning technique has been developing intensively in big data era. However, its capability is still challenged for the design of network structure and parameter setting. Therefore, it is essential to improve the performance of the model and optimize the complexity of the model. Machine learning can be segmented into five categories in terms of learning methods: 1) supervised learning, 2) unsupervised learning, 3) semi-supervised learning, 4) deep learning, and 5) reinforcement learning. These machine learning techniques are required to be incorporated in. To improve its fitting and

收稿日期:2021-12-16;**修回日期**:2022-05-20;**预印本日期**:2022-05-27

*通信作者:郑艺峰　zyf@mnnu.edu.cn

基金项目:国家自然科学基金项目(62141602);福建省自然科学基金项目(2021J011004, 2021J011002);克拉玛依科技发展计划项目(2020CGZH0009);教育部产学研创新计划(2021LDA09003)

Supported by: National Natural Science Foundation of China (62141602); Natural Science Foundation of Fujian Province, China (2021J011004, 2021J011002)

generalization ability, we select supervised deep learning as a niche to summarize and analyze the optimization methods. First, the mechanism of optimization is demonstrated and its key elements are illustrated. Then, the optimization problem is decomposed into three directions in relevant to fitting ability: 1) convergence, 2) convergence speed, and 3) global-context quality. At the same time, we also summarize and analyze the specific methods and research results of these three optimization directions. Among them, convergence refers to running the algorithm and converging to a synthesis like a stationary point. The gradient exploding/vanishing problem is shown that small changes in a multi-layer network may amplify and stimuli or decline and disappear for each layer. The speed of convergence refers to the ability to assist the model to converge at a faster speed. After the convergence task of the model, the optimization algorithm to accelerate the model convergence should be considered to improve the performance of the model. The global-context quality problem is to ensure that the model converges to a lower solution (the global minimum). The first two problems are local-oriented and the last one is global-concerned. The boundary of these three problems is fuzzy, for example, some optimization methods to improve convergence can accelerate the convergence speed of the model as well. After the fitting optimization of the model, it is necessary to consider the large number of parameters in the deep learning model as well, which can cause poor generalization effect due to overfitting. Regularization can be regarded as an effective method for generalization. To improve the generalization ability of the model, current situation of regularization methods are categorized from two aspects: 1) data processing and 2) model parameters-constrained. Data processing refers to data processing during model training, such as dataset enhancement, noise injection and adversarial training. These optimization methods can improve the generalization ability of the model effectively. Model parameters constraints are oriented to parameters-constrained in the network, which can also improve the generalization ability of the model. We take generative adversarial network (GAN) as the application background and review the growth of its variant model because it can be as a commonly-used deep learning network. We analyze the application of relevant optimization methods in GAN domain from two aspects of fitting and generalization ability. Taking WGAN with gradient penalty (WGAN-GP) as the basic model, we design an experiment on MNIST-10 dataset to study the applicability of the six algorithms (stochastic gradient method(SGD), momentum SGD, Adagrad, Adadelta, root mean square propagation(RMSProp), and Adam) in the context of deep learning based GAN domain. The optimization effects are compared and analyzed in relevant to the experimental results of multiple optimization methods on variants of GAN model, and some GAN-based optimization strategies are required to be clarified further. At present, various optimization methods have been widely used in deep learning models. Various optimization methods to improve the fitting ability can improve the performance of the model. Furthermore, these regularized optimization methods are beneficial to alleviate the problem of model overfitting and improve the robustness of the model. But, there is still a lack of systematic theories and mechanisms for guidance. In addition, there are still some optimization problems to be further studied. The Lipschitz limitation of global gradients is not guaranteed in deep neural networks due to the gap between theory and practice. In the field of GAN, there is still a lack of theoretical breakthroughs to find the stable global optimal solution, that is, the optimal Nash equilibrium. Moreover, some of the existing optimization methods are empirical and its interpretability is lack of clear theoretical proof. There are many and complex optimization methods in deep learning. The use of various optimization methods should be focused on the integrated effect of multiple optimizations. Our critical analysis is potential to provide a reference for the optimization method selection in the design of deep neural network.

Key words: machine learning; deep learning; deep learning optimization; regularization; generative adversarial network (GAN)

0 引 言

随着智能技术的发展,深度学习备受青睐,广泛应用于计算机视觉(Agbo-Ajala 和 Viriri, 2021; Abdolahnejad和Liu,2020)、图异常检测(陈波冯 等,2021)、推荐系统的数据分析(Khan等,2021)和自然语言处理(Torfi等,2020)等领域。研究人员主要致力于如何更好地提高深度学习模型的性能。Hinton和Salakhutdinov(2006)最早初步解决了“梯度消失”问题。首先通过无监督的学习方法逐层训练模型,每训练一层隐藏节点就作为下一层隐藏节点的输入

（该过程称为预训练），再使用有监督的反向传播（Rumelhart 等，1986）进行调优，以逐层预训练的方式提取数据的高维特征，初步解决梯度消失的问题。Glorot 和 Bengio（2010）提出 Xavier 初始化，使状态方差和梯度方差保持不变，进而提升模型分类性能。2011 年，ReLU（rectified linear unit）激活函数被证明可以针对性控制梯度消失的情况（Glorot 等，2011），此时深度学习仍处于理论研究阶段。直至 2012 年，Hinton 团队在 ILSVRC（ImageNet large scale visual recognition challenge）大赛上通过结合 ReLU 激活函数构建 AlexNet（Krizhevsky 等，2012），以碾压性的分类性能夺冠，进一步推动深度学习成为研究热点。此后，深度学习在其他领域也得以蓬勃发展。Girshick（2015）提出 R-CNN（region convolutional neural network）将深度学习引入目标检测领域。

作为一种机器学习方法，深度学习可与其他机器学习方法相结合，如图 1 所示。为了更加精简地描述优化问题，仅以有监督深度学习的优化作为切入点进行分析。监督学习的目标是通过得到的样本找到一个近似底层函数的函数，主要由“表示”（representation）、“优化”（optimization）和“泛化”（generalization）3 个步骤组成（Sun，2020）。“表示”即找到一个丰富的函数族用以表示目标函数；“优化”即通过最小化损失函数以确定函数参数；“泛化”指用得到的目标函数进行预测，产生的误差称为测试误差，包括表示误差、优化误差和泛化误差。一般默认已经找到适合的目标函数再进行优化，因此不考虑表示误差。

神经网络本质上是一种对网络参数优化变量的方法。因此，在确定适合的目标函数后，深度学习的核心问题可归结为一个优化问题，其强大性能高度依赖经验，研究人员需要经过训练大量模型才能得到适合的参数。此外，在训练模型的过程中，现有的理论无法严谨地分析所用方法的有效性。因此，深度学习的优化问题可概括为：设计有效的优化方法来提升模型的拟合能力，降低训练误差；同时，还要考虑通过正则化方法提升模型的泛化能力，还能降低模型的复杂度，更加高效地训练神经网络。

首先，介绍深度学习优化的理论基础，从拟合能力的角度将深度学习优化问题划分为 3 个具体研究方向，即收敛性、收敛速度和全局质量问题，并对每个方向进行总结分析。其次，针对集成多种优化方法可能存在过拟合进而降低模型泛化能力的问题，以提升模型泛化能力的正则化方法作为切入点，详细阐述分析不同正则化方法的作用。接着，讨论上述优化方法在生成对抗网络（generative adversarial network，GAN）中的使用。最后，在现有深度学习优化理论的基础上，分析目前深度学习领域仍存在的问题并分析未来研究方向。

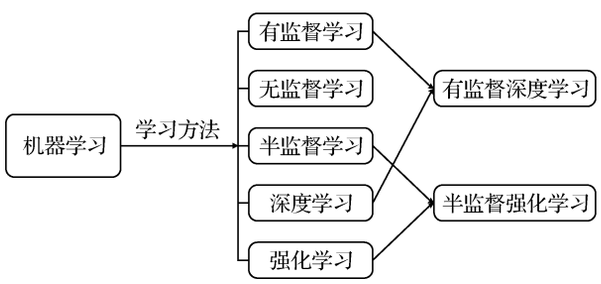

图 1　机器学习方法相结合

Fig. 1　Combination of machine learning methods

1　理论基础

1.1　优化的形式化描述

对于有监督学习而言，给定数据点 $x_i \in \mathbf{R}$，$y_i \in \mathbf{R}$，$i = 1, \cdots, n$（n 表示输入样本的个数），构建深度模型对 x_i 进行预测 $\hat{y}_i$。首先，构造一个神经网络用以近似底层映射函数 $f_\theta(x)$，将 x_i 映射到 y_i。为了简化表述，忽略神经网络表达式中的偏差，将 $f_\theta(x)$ 形式化表示为

$$f_\theta(x) = \boldsymbol{W}^L \varphi\left(\boldsymbol{W}^{L-1} \varphi\left(\boldsymbol{W}^{L-2} \cdots \varphi\left(\boldsymbol{W}^2 \varphi\left(\boldsymbol{W}^1 x\right)\right)\right)\right) \tag{1}$$

式中，$\varphi(\cdot)$ 是 $\mathbf{R} \to \mathbf{R}$ 的神经元激活函数，$\boldsymbol{W}^j$ 是维数 $d_j \times d_{j-1}$ 的矩阵，$j = 1, \cdots, L$，$\theta = \left(\boldsymbol{W}^1, \cdots, \boldsymbol{W}^L\right)$ 表示所有参数的集合。

在模型训练过程中，需最优的网络参数以提高模型的性能，即预测输出值 $\hat{y}_i = f_\theta(x_i)$ 更逼近于真实值 y_i。这意味着需要最小化 $\hat{y}_i$ 与 y_i 之间的距离，故可将寻找最优参数的问题（即损失函数）表示为

$$\min_\theta F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, f_\theta(x_i)\right) \tag{2}$$

式中，ℓ 为距离度量单位。对于回归问题，ℓ 通常为二次损失函数 $\ell(y, z) = \|y - z\|^2$；对于二分类问题，则采用交叉熵公式 $\ell(y, z) = \log\left(1 + \mathrm{e}^{-yz}\right)$。

从式（2）可以看出，$F(\theta)$ 由 3 部分组成，即输入样本 x_i、参数矩阵 $\boldsymbol{W}^j$ 以及真实值 y_i。式中矩阵 $\boldsymbol{W}^j$ 是唯一不确定参数。因此，可对参数 $\boldsymbol{W}^j$ 求导以求解函

数的极值点，间接找到模型的最优参数。然而，在实际操作中，损失函数的导函数多为不可逆，无法求解，从而导致深度学习的发展陷入瓶颈。

1.2 优化的核心：梯度下降

在实际应用中的深度学习模型结构更为复杂，为了便于理解优化问题，主要以全连接神经网络作为切入点。

虽然损失函数的极值点不易得到，但只需将参数矩阵 $\boldsymbol{W}$ 代入即可获得极值点的导数值（即梯度）。梯度方向是指曲线切线向上的方向，沿着其反方向则为函数值下降最快的方向。因此，可按照给定步长沿梯度的反方向不断迭代，逐步逼近极值点。现阶段，研究人员以梯度下降作为优化最基本思想，其表示为

$$\theta_{t+1} = \theta_t - \alpha_t \nabla F_i(\theta_t) \tag{3}$$

式中，α_t 表示第 t 轮迭代步长（或学习率），$\nabla F_i(\theta_t)$ 为第 t 轮迭代的第 i 小批次损失函数的梯度。值得注意的是，只需随机初始化 $\boldsymbol{W}^0$，就可计算下一轮迭代的 $\boldsymbol{W}^j, j = 1, \cdots, L$。

此外，在模型训练过程中，需使用链式法则对损失函数的每一个参数求偏导。在计算资源不足情况下，求解效率非常低。直至 Hinton 和 Salakhutdinov（2006）提出无监督逐层训练的方法，将预训练结合反向传播算法（Werbos，1974），仅需一次正向传播就可计算其损失，再通过一次反向传播更新其网络参数，从而有效降低模型计算量。上述方法的核心思想在于先求解局部最优解，再将所有的局部最优解进行整合，从而获取全局最优解。

2 优化拟合能力

仅考虑拟合能力的情况下，将复杂的优化问题分为3步。首先运行算法并收敛到一个合理解，如一个驻点；然后使算法能够更快收敛；最后确保算法收敛于一个较低解，即全局最小值。因此，将优化问题归纳为收敛性（局部问题）、收敛速度（局部问题）和全局质量问题（全局问题）。注意，上述3个问题的边界较为模糊。例如，收敛性讨论所涉及的技术也可解释收敛速度的提高。

2.1 收敛性问题

在模型训练过程中，由于收敛速度缓慢可能导致神经网络无法在多项式时间内收敛，也称为梯度爆炸/消失问题。上述情况的原因在于多层网络中微小的变动可能在每一层放大而爆炸（也可能在每一层衰减而消失），即神经网络中的权值由于梯度反向传播中的连乘效应导致更新不稳定，尤其随着网络层数的增加，将会变得愈加严重。更加具体的解释，梯度爆炸/消失影响收敛速度缓慢可进一步归结为两点。其一，收敛速度通常是由 Hessian 矩阵的条件数决定，梯度爆炸/消失意味着梯度的每个分量可以非常大（或非常小），进而使得 Hessian 对角矩阵的 Lipschitz 常数也随之变化。因此，Hessian 对角矩阵可能具有高度动态的范围，从而导致条件数呈指数级增大；其二，在实际应用中，更新学习率一般使用恒定步长或固定步长。而当 Lipschitz 常数发生急剧变化时，恒定步长或固定步长远小于理论步长，从而导致收敛缓慢。

为了更直观地阐述，引用 Sun（2020）中的例子。$F(w) = (w^7 - 1)^2$，式中 $w \in \mathbf{R}$。函数的曲线图如图2所示。假设 c 为常数（以0.2为例），$[-1+c, 1-c]$ 区域曲线平缓，找不到下降梯度（即梯度消失）；$(-\infty, -1-c] \cup [1+c, +\infty)$ 区域则过于陡峭（即梯度爆炸）。假如在全局极小值 $w = 1$ 附近开始初始化，则梯度下降可快速收敛，但如果在 $w = -1$ 附近，会由于梯度消失而需更长的收敛时间。

综上所述，如何解决梯度爆炸/消失是模型训练过程中亟需解决的问题。正如上述例子所示，倘若选择盆地内接近全局极小值的位置作为初始点，则可有效减少迭代时间。因此，如何使深度模型能在一个接近全局极小值的盆地位置初始化，需从精细初始化、归一化和改变神经结构3方面入手。

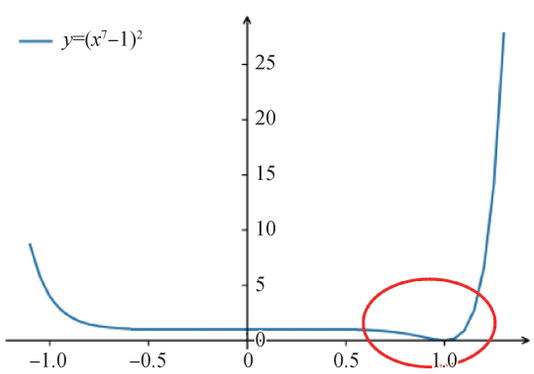

图2 $F(w) = (w^7 - 1)^2$ 曲线图

Fig. 2 Function graph of $F(w) = (w^7 - 1)^2$

2.1.1　精细初始化

研究初期，研究人员通常采用简单的点进行初始化（naive initialization）。例如，全0或仅一小部分权值为非0的稀疏初始点，或从随机分布中抽取权重。然而上述方法在实际应用中普适性较差。随后，Bottou（1988）和LeCun等人（2012）先后提出的具有特定方差的随机初始化，调整随机初始点的缩放来实现更快初始化，但不足之处在于缩放因子无法适用于其他神经网络。Glorot和Bengio（2010）通过分析深度神经网络的信号传播，提出预训练和Xavier初始化，可同时处理前馈和反向信号的传播。但两种初始化方法均基于sigmoid激活函数。此后，随着ReLU激活函数的普及，He等人（2016）提出凯明初始化。上述讨论的方法均与数据无关，直至Mishkin和Matas（2016）提出依赖数据的LSUV（layer sequential unit-variance）初始化，在一些问题上具有经验性的知识。由于凯明初始化无法推广到一般非线性网络，Poole等人（2016）提出采用平均场近似的具有一般非线性激活函数的无限宽网络。Hanin和Rolnick（2018）通过分析具有ReLU激活函数的有限宽网络，从而解释部分训练深度网络困难的原因，以及部分关于深度神经网络训练困难的猜测。随着研究的深入，Dauphin和Schoenholz（2019）提出元初始化，可不使用归一化方法发现效果最好的初始点。此外，还可采用基于动态等距的方法。Saxe等人（2014）研究深度线性网络的正交初始化。Pennington等人（2017，2018）对无限宽深度非线性网络进行形式化分析。Xiao等人（2018）提出卷积神经网络（convolutional neural networks，CNN）的动态等距，可训练10 000层CNN而无需批归一化或跳跃连接等技巧。Li和Nguyen（2019）以及Gilboa等人（2019）分别将动态等距分析应用于其他网络当中，针对网络提出新的初始化方案。

2.1.2　归一化

在模型中进行归一化，可看做精细初始化的扩展。归一化方法不仅修改初始点，而且可为接下来的所有迭代网络修改初始点。最具代表性的方法是批归一化方法（Ioffe和Szegedy，2015），其对神经网络的中间层也进行归一化处理，从而解决网络经过多层运算后使得数据分布发生较大改变的问题，可获得更好的训练效果。

随着研究的深入，研究人员对批归一化的理解不断发生变化。Santurkar等人（2018）认为批归一化的成功与内部协变量无关，是减小了梯度的Lipschitz常数。Bjorck等人（2018）认为批归一化允许更大的学习率，讨论其与初始化方案的关系。此外，Arora等人（2019）、Cai等人（2019）和Kohler等人（2019）进一步分析批归一化在各种设置下的理论效益。Ghorbani等人（2019）发现对于未进行批归一化的网络，可能存在较大的孤立特征值，反之则不会出现该现象。

批归一化的关键问题在于每个小批次的平均值和方差是作为所有样本的平均值和方差的近似值来计算。若不同小批次的数据存在不相似性，则会导致批归一化的效果不佳。基于这个问题，研究人员提出许多归一化方法。层归一化（Ba等，2016）对特征通道进行归一化，适用于样本较少的小批次训练；组归一化（Wu和He，2020）则不再按单个特征通道来归一化，而是将特征通道分组后以组为单位来归一化；实例归一化（Ulyanov等，2016）依据图像像素进行归一化，在风格迁移领域应用较广；可切换归一化（Luo等，2021）通过加权平均的方式将批归一化、实例归一化和层归一化结合起来，在对象检测等任务中表现出色；由于批归一化对所有的样本使用统一的仿射变换参数，所以条件归一化（Dumoulin等，2017）对每个类别设置不同的变换参数；权重归一化（Salimans和Kingma，2016）对网络中的连接权重进行归一化，由于共享权重，开销较批归一化小得多；谱归一化（Miyato等，2018）将每层的参数除以参数矩阵的最大奇异值，限制函数满足Lipschitz限制来提升模型的拟合能力。

2.1.3　改变神经架构

随着神经网络层数不断增加，模型性能越来越好。然而，在训练过程中发现，即使通过精细初始化和批归一化，也难以训练一个20～30层的网络。为此，研究人员倾向于改变神经结构以解决上述问题。例如，深度多达152层的ResNet（residual network）（He等，2016），其核心是引入residual representation的概念，通过使用多个有参层来学习输入输出之间的残差表示，并设计恒等跳跃连接层用于直接跳过一个或多个层。

假设连续的两层权重分别为 $\boldsymbol{W}^j$ 和 $\boldsymbol{W}^{j+1}$，$j=1,\cdots,L$。令 $\varphi(\cdot)$ 表示激活函数，则给定输入 x_i 时，$i=$

1, …, n, 残差学习过程如图3所示。若下一层激活函数前的输出最优为原输入x_i时, 无需优化权重层$\boldsymbol{W}^j$和$\boldsymbol{W}^{j+1}$, 直到$\boldsymbol{W}^{j+1}\varphi(\boldsymbol{W}^j x_i)=x_i$, 通过残差学习简单优化$\boldsymbol{W}^{j+1}\varphi(\boldsymbol{W}^j x_i)-x_i$为0即可。更加高效的同时, 确保在深度神经网络迭代的过程中, 至少不会因为网络层的堆叠而导致模型退化。

此外, 研究人员也相继提出与之相关的流行结构, 包括high-way networks(Srivastava等, 2015)、DenseNet(dense convolutional network)(Huang等, 2017)、ResNeXt(Xie等, 2017)和自动搜索神经结构(Zoph和Le, 2017)。目前, 在图像分类网络结构中, 可借助额外的技巧达到比ResNet高出7%的准确率, 例如EfficientNet(Tan和Le, 2019)。

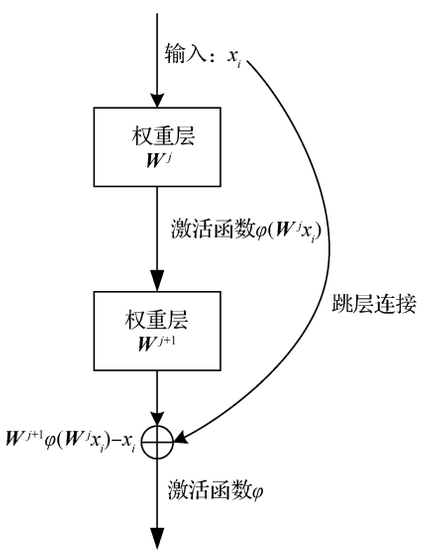

图3 ResNet中残差学习示意图

Fig. 3 Schematic diagram of residual learning in ResNet

2.2 收敛速度问题

上一节主要分析如何采用优化方法以提高模型收敛性问题。本节主要从收敛速度角度作为切入点进行分析。

2.2.1 随机梯度下降

随机梯度下降法(stochastic gradient method, SGD)(Robbins和Monro, 1951)的工作原理旨在每轮开始时, 随机打乱样本集合顺序, 再将其划分为多个小批次, 迭代时仅计算每一批次的梯度并更新权重。

假设第t轮中随机选择第i个批次, 其更新参数θ_t定义为

$$\theta_{t+1}=\theta_t-\alpha_t\nabla F_i(\theta_t) \tag{4}$$

式中, $\nabla F_i(\theta_t)$为第t轮的第i小批次训练损失的梯度, α_t为步长。

从式(4)可以看出步长的选择至关重要。为有效提升SGD的性能, 研究人员提出3种不同的方法。1)Vanilla朴素SGD(Robbins和Monro, 1951), 在训练过程中每迭代几次就将步长除以一个固定的常数; 2)预热学习率(Srivastava等, 2015), 在迭代开始时使用非常小的步长, 后逐步增加到常规的步长; 3)周期学习率(Smith, 2017; Loshchilov和Hutter, 2017), 使步长在一个较低阈值和较高阈值之间跳动, 有助于模型逃离鞍点。

2.2.2 带动量的SGD

带动量的SGD方法(Qian, 1999)是重球法和加速梯度法的随机版, 工作原理为在第t轮迭代中随机选择第i个批次, 进而更新一阶动量项m_t和参数θ_t, 具体定义为

$$m_t=\beta m_{t+1}+(1-\beta)\nabla F_i(\theta_t) \tag{5}$$

$$\theta_{t+1}=\theta_t-\alpha_t m_t \tag{6}$$

式中, β为初始动量参数, $0\leqslant\beta\leqslant1$。

一般认为历史梯度的滑动均值是朝向最优方向, 而当前的小批次的方差可能较大, 可能使梯度下降方向发生偏移。引入动量的目的在于减小梯度方差, 纠正方向, 以此加速收敛。但在实际应用中, 并不能保证该方法能适用不同的网络模型。

2.2.3 自适应梯度法

自适应梯度法是SGD算法的改进版。以AdaGrad(adaptive gradient)(Duchi等, 2011)为例, 工作原理为添加对步长的约束。离最优点越远, 则使用较大步长, 反之步长较小, 能有效处理稀疏和不平衡的数据。

假设在第t轮迭代中随机选择第i个批次, 其参数θ_t更新为

$$\theta_{t+1}=\theta_t-\alpha_t v_t^{-1/2}\circ g_t \tag{7}$$

式中, $g_t=\nabla F_i(\theta_t)$, $v_t=\sum_{k=1}^{t}g_k\circ g_k$, "$\circ$"表示矩阵的点积。

但AdaGrad存在一个明显的弊端, 即采用同样的方法处理所有的梯度, 在迭代后期由于步长过小, 可能较难找到有用的解。针对上述问题, 研究人员提出RMSProp(root mean square propagation)(Tieleman和Hinton, 2012), 新增衰减系数用于控制历史信息的获取比例, 结合历史梯度来分配学习率, 加快收敛速度。

假设第 t 轮迭代中随机选择第 i 个批次并计算 g_t，其二阶动量项 v_t 和参数 θ_t 更新式定义为

$$v_t = \beta v_{t-1} + (1-\beta) g_t \circ g_t \tag{8}$$

$$\theta_{t+1} = \theta_t - \alpha_t v_t^{-1/2} \circ g_t \tag{9}$$

Adam（advanced digital asset management）则是将 RMSProp 与带动量加速的 SGD 相结合，对超参数相对不敏感，且在训练过程中收效较快。但其不足之处在于，模型的训练误差和测试误差相比 SGD 和带动量加速的 SGD 较大（Kingma 和 Ba，2015）。

假设在第 t 轮迭代中随机选择第 i 个批次并计算 g_t，其一阶动量项 m_t、二阶动量项 v_t 和参数 θ_t 更新为

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{10}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t \circ g_t \tag{11}$$

$$\theta_{t+1} = \theta_t - \alpha_t v_t^{-1/2} \circ m_t \tag{12}$$

2.3　全局质量问题

上述内容注重于解决模型训练的“局部问题”，可以保证模型有效地收敛到局部最小值。但问题往往具有非凸性，如何有效地找到全局最小值值得进一步研究。损耗面是指 $\mathbf{R}^{D+1}$ 中的高维面 $(\theta, F(\theta))$（D 为参数的总数），也称为优化“图景”（landscape）。本节将详细分析优化远景的有效性。

随着图景的深入研究，研究人员发现以下 3 个重要性质。1）深度神经网络存在一种称为“模式连通性”的几何特性。网络中两个局部极小值可以通过几乎无损失的路径进行连接，也称“子级集的连通性”（Draxler 等，2018；Garipov 等，2018）。2）彩票假设（lottery ticket hypothesis，LTH）（Frankle 和 Carbin，2019）。研究人员发现若对大网络进行剪枝，并继承大网络已训练好的权重，则测试精度与原有网络基本相同。LTH 假设的核心是根据一个半随机初始点训练出来的小网络可取得和大网络基本等效的性能。LTH 假设不足之处在于，其工作大部分都是经验性的。3）已知测试损失函数与训练损失函数基本相似，在平且宽的曲线上极值点则相差不大，而在锐利的曲线则会产生较大的差别。Dinh 等人（2017）认为可对锐利曲线进行重新参数化，使其变宽，从而达到与在平且宽曲线上类似的效果，如图 4 所示。

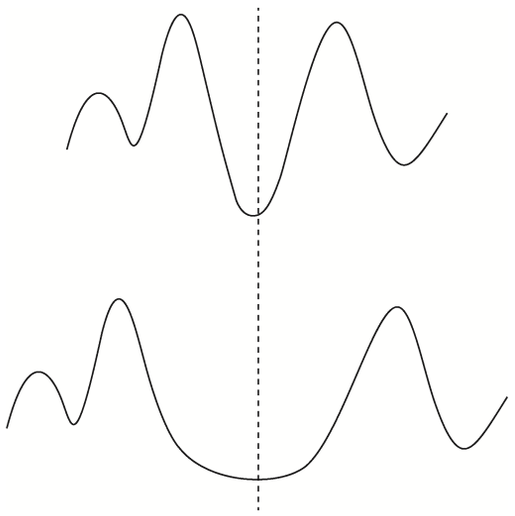

图 4　重新参数化锐利曲线

Fig. 4　Reparameterize sharp curve

3　优化泛化能力

上述方法可有效降低模型的训练误差，有助于提升模型性能。但训练深度模型的目的在于提高模型在测试集上的泛化效果，而深度学习模型由于参数量庞大，容易因过拟合从而导致泛化效果较差。可见，提升模型的泛化效果也是优化问题中非常关键的部分。

正则化能有效保证算法泛化能力，是深度学习领域中非常重要的研究课题之一。本节主要从数据预处理、模型参数限制两个方面进行分析。

3.1　数据预处理

3.1.1　数据集增强

过拟合可视为对数据集中噪声点的过度捕捉，其防止过拟合最行之有效的方法就是增加更多的数据用于训练模型。在实际过程中，对于数据集进行标记成本昂贵。因此，可采用数据增强（data agumentation，DA）方法往数据集中人工增加假数据以进一步提升模型泛化效果。该方法已广泛应用于自然语言处理（Yu 等，2018；Wei 和 Zou，2019）、语音识别（Hinton 等，2012；Ko 等，2015）和目标识别（Zoph 等，2020）等领域。

在实际应用过程中，根据是否提前设立变换规则来扩增数据集，具体可分为有监督的 DA 和无监督的 DA。有监督的 DA 仅围绕所需变换样本本身进行操作（也称单样本 DA），常见操作包括几何变换、

颜色变换和仿射变换等。此外，还可结合多个样本用以产生新数据的多样本DA（Chawla等，2002）。而无监督的DA方法产生的数据更具有随机性，模型可将数据分布作为先验知识，随机生成与先验分布一致的新样本，生成对抗网络（generative adversarial networks，GAN）（Goodfellow等，2020）就应用了该方法。此外，Cubuk等人（2018）提出的AutoAugment也可以通过模型自主学习，产生一种适合当前任务的DA策略，以及能加快搜索时间的改进版Fast Auto-Augment（Lim等，2019）。

3.1.2　噪声注入

在训练期间对神经网络进行噪声注入（noise injection，NI），相当于将一个额外的项添加到误差函数中，可显著提高泛化性能。在噪声注入过程中，需注意噪声的大小，过小的噪声对网络几乎没有影响，而过大的噪声会使得学习映射函数过程变得更加困难。噪声注入方式可分为以下6种：

1）对输入数据NI。训练时对输入数据加入随机噪声以进一步提高模型的健壮性。Vincent等人（2008）提出降噪自编码器对输入数据NI，提高编码器的鲁棒性。

2）对隐藏层NI。将噪声注入到隐藏层中（在多个抽样层上DA），一种通过加入噪声重构新输入的正则化策略，相比简单收缩参数，可获得更好的效果。Chu等人（2021）提出一种对跳跃连接层NI，在基本不添加额外计算开销的情况下解决跳跃连接层富集和性能损失问题。

3）对权重NI。将噪声注入到权重中，在某种程度上可以等价地解释为传统的范数惩罚，是一种有效的正则化措施。Burton和Mpitsos（1992）证明在权重更新时增加噪声并将噪声的大小应用于输出误差可以实现更快的学习率。Hanson（1990）提出一个适应权重均值和标准差的随机增量规则，提高了收敛率。

4）对梯度NI。在梯度中注入噪声则更适用于非凸问题。Neelakantan等人（2015）发现训练具有随机梯度下降的深度神经网络时注入退火高斯噪声到梯度成效明显，注入的噪声可通过鼓励模型积极探索参数空间来实现更低的训练损失。Zhuo等人（2019）提出一种用于深度神经网络优化的校正随机梯度下降算法，更容易建立CNN，并为深度学习范式中的无偏变量估计提供理论框架以优化模型参数的计算。

5）对激活函数NI。激活函数用于提高神经网络对模型的表达能力，注入噪声可解决硬饱和问题。Gulcehre等人（2016）提出一种在激活函数的饱和部分注入噪声并学习噪声规模的训练技术，在解决优化和测试时激活函数存在硬饱和问题的同时，还发现可使用更加广泛的激活函数种类来训练神经网络。

6）对标签NI。直接对数据集中的标签进行NI。Xie等人（2016）也提出在对抗训练每次迭代过程中将部分标签随机替换为错误标签。Vaswani等人（2017）则在机器翻译领域，使用标签平滑在BLEU（bilingual evaluation understudy）上获得较小但非常重要的提高。

3.1.3　对抗训练

现阶段，研究人员还提出一类利用反向思维作为切入点的正则化方法，即分析模型分类错误的样本。最早是在2014年，Szegedy等人（2014）提出对抗样本，并指出即使用不同超参数在训练集的不同子集上得到的模型，都会对同一组对抗样本产生误判，即DNN具有非直观的特征和固有的盲点。与传统的NI不同，对抗样本是通过对输入数据注入一些人类无法察觉的细微噪声，误导模型判断错误来最大化预测误差。

正则化背景下，在模型训练过程中添加对抗样本的正则化方法称为对抗训练，可在灵活捕捉数据集线性趋势的同时，进一步提升模型对于局部扰动的鲁棒性。Madry等人（2018）提出的PGD（projected gradient descent）方法，通过多次迭代使损失函数局部线性假设基本成立。但在进行前向和后向计算时，PGD会计算出每次的参数梯度和输入梯度，前者仅用于梯度下降过程，后者仅用于梯度上升过程，存在计算效率低的问题。针对上述问题，Shafahi等人（2019）提出FreeAT（free adversarial training）用每次计算得到的梯度，同步更新噪声和模型参数，比传统对抗训练快3～30倍。Zhang等人（2019a）利用DNN的结构以及PMP（pontryagin's maximum principle），提出仅求第1层的梯度来更新参数的YOPO（you only propagate once）方法。FreeLB（free large-batch）（Zhang等，2019a）累积多步的参数梯度和输入梯度，综合起来再进行更新，虽然未明显提升计算效率，但能更精确地更新参数。

3.2　模型参数限制

3.2.1　参数范数惩罚

参数范数惩罚正则化在损失函数中加入对参数的范数惩罚项，通过限制模型的学习能力降低其过拟合的可能性。典型方法包括L_2正则化、L_1正则化、Elastic-Net正则化（Zou和Hastie，2005）和L_0正则化。

1）L_2正则化。在原始的损失函数后加上L_2正则化项。令L_0代表原始的损失函数，增加L_2正则化的损失函数定义为

$$L = L_0 + \frac{\lambda}{2n}\sum_j (\boldsymbol{W}^j)^2 \tag{13}$$

$$L_0 = \min_{\theta} F(\theta) \tag{14}$$

式中，n为训练样本的总数，λ为正则化系数，$\boldsymbol{W}^j$为权重矩阵，$j = 1, \cdots, L$。λ越小则正则化作用越小，模型侧重于优化原本的损失函数；而λ越大则表示正则化作用越明显，权重矩阵$\boldsymbol{W}^j$则会趋于0。

值得注意的是，L_2正则化对偏差无影响。通常只对权重做惩罚而不对偏置做惩罚。

2）L_1正则化。在原始的损失函数L_0后加上L_1正则化项，损失函数可重定义为

$$L = L_0 + \frac{\lambda}{n}\sum_j \left|\boldsymbol{W}^j\right| \tag{15}$$

L_1正则化在权重矩阵$\boldsymbol{W}^j$大于0时起抑制作用，小于0时反而增大$\boldsymbol{W}^j$，整体上使$\boldsymbol{W}^j$趋于0来降低模型复杂度。一般情况下，会更偏向于使用L_2正则化，因为其便于求导，从而方便优化。

3）Elastic-Net正则化。Elastic-Net正则化结合L_1正则化和L_2正则化，即在L_1正则化的基础上加上L_2正则化项，损失函数可重定义为

$$L = L_0 + \frac{\lambda_1}{n}\sum_j \left|\boldsymbol{W}^j\right| + \frac{\lambda_2}{2n}\sum_j (\boldsymbol{W}^j)^2 \tag{16}$$

式中，λ_1为L_1正则化系数，λ_2为L_2正则化系数。

由此可见，Elastic-Net正则化结合了L_1正则化更易获得稀疏解和L_2正则化约束参数量级的优点。

4）L_0正则化。L_0正则化主要对非0参数起限制作用，达到约束模型学习能力的目的。从理论上L_0正则化可用于求稀疏解（Natarajan，1995），但L_0范数的优化求解是一个NP（non-deterministic polynomial）难问题。

3.2.2　作为约束的范数惩罚

参数范数惩罚通过在损失函数中加入惩罚项以达到正则化目的，每个惩罚项可表示为KKT（Karush-Kuhn-Tucker）乘子和约束目标函数的表达式之间的乘积，可构建广义Lagrange函数为

$$L = L_0 + \alpha\Omega(\theta) \tag{17}$$

式中，$\theta = \left(\boldsymbol{W}^1, \cdots, \boldsymbol{W}^L\right)$为所有参数集合，$\Omega(\theta)$是参数范数惩罚；$\alpha \in [0, \infty)$表示控制惩罚项作用大小的超参数，即KKT乘子，$\alpha$越大惩罚程度越高。因此参数范数惩罚会强行将权重约束在一个区域中，如L_2正则化会将权重约束在一个L_2球区域中。

但是，以惩罚强加约束并不适用于所有情况，还有一种通过显式限制来寻找最佳参数的方法，即作为约束的范数惩罚。该方法通过在参数范数惩罚的基础上，进一步限制参数范数惩罚项小于某个常数来正则化，即$L = L_0 + \alpha\left(\Omega(\theta) - k\right)$，其中$k$为用于约束$\alpha$的常数。

作为约束的范数惩罚被认为是参数范数惩罚优化问题的显式形式，可先用梯度下降法计算L_0的下降梯度，再投影到$\Omega(\theta) < k$区域的最近点。使用惩罚强加约束时，较高的学习率易诱导梯度迅速增大，导致权重远离原点并溢出，而显式约束中的常数k通过直接鼓励权重接近原点，避免了权重无限制增加的情况。此外，显式约束仅在权重逃离约束区域时起作用，有效避免损失函数非凸情况，进而规避将模型陷入局部极小值。

3.2.3　稀疏表示

函数稀疏表示最初主要用于信号处理领域以解决自然界中低频信号和高频噪声的问题（Donoho，1995）。Candès等人（2006）发现稀疏度和L_1范数之间存在联系，即当采样维度和信号的稀疏度能满足一定关系时，采样后的信号可以通过最小化L_1范数来恢复。Donoho（2006）将这个过程归纳为压缩感知，直接推动Candès等人（2011）提出人脸识别的鲁棒主成分分析。

与参数惩罚方法不同的是，稀疏表示是通过惩罚DNN的激活单元以提高隐藏层的稀疏性的，是一种间接的参数惩罚方法。与L_1正则化类似，稀疏表示则是限制隐藏层的输出趋于或等于0，目的是希望输出更加稀疏。可认为对损失函数添加一个稀疏表示惩罚项，损失函数可重定义为

$$L = L_0 + \alpha\left(\Omega(h)\right), h = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X} \tag{18}$$

式中，$\boldsymbol{X}$为稀疏矩阵，h是权重参数的转置矩阵和稀疏矩阵的矩阵乘积，即稀疏编码。

计算稀疏表示惩罚的方法可分为两种。1)使用与L_1正则化相同的惩罚方法$\Omega(h)=\sum|h|$;2)使用KL(Kullback-Leible)散度即相对熵来计算。

Hinton和Salakhutdinov(2006)提出自编码器(AutoEncoder),即将数据样本输入到编码器中得到的特征再送进解码器中,得到的输出与原数据样本之间的误差称为重构误差,通过稀疏表示惩罚来调整编码器与解码器的参数,进而使得重构误差最小。Ng(2011)在自编码器的基础上增加L_1正则化,构建稀疏自编码器(sparse AutoEncoder)。此外,Vincent等人(2008)在自编码器的基础上引入训练数据NI,构建降噪自编码器(denoising AutoEncoder)。

3.2.4 参数绑定

当任务之间足够相似时,如具有相似的数据分布和相同的特征空间,则可认为模型参数是相似的。将已训练模型中的参数作为未训练模型参数的先验知识的正则化方式就称为参数绑定。与针对偏离0的权重参数进行惩罚的L_2正则化相比,参数绑定是惩罚偏离先验参数的权重参数。参数绑定的惩罚项定义为

$$\Omega=\left\|\boldsymbol{W}^A-\boldsymbol{W}^B\right\|_2^2 \tag{19}$$

式中,$\boldsymbol{W}^A$为已训练模型的参数,$\boldsymbol{W}^B$为未知模型的参数。

最广为使用的参数绑定方法就是参数共享,即强迫某些参数与先验参数相等,多用于CNN和循环神经网络中。CNN当中主要通过卷积核的空间处理来共享参数,循环神经网络的参数共享则是利用一个类似“滑窗”的过程进行时间处理。参数共享使得不同长度的样本都能够在模型上扩展并进行泛化,尤其当多次出现某部分数据信息时,这样的共享可以避免重复学习(Goodfellow等,2016)。

3.2.5 半监督学习

随着智能应用的普及,信息量呈爆发式增长,数据获取变得极为容易,但对数据进行标记的代价昂贵。对于这些数量巨大且蕴含较大有价值信息的未标记数据,最早由Shahshahani和Landgrebe(1994)开始对其优化模型进行研究。半监督学习假设未标记样本和标记样本具有相似的数据分布,则可利用未标记数据来帮助模型学习,提高泛化能力。

半监督学习作为一种介于监督学习和无监督学习之间的学习方式,其核心在于添加约束未标记样本的正则项到目标函数当中。目前主要有两种设置正则项的方法。

1)熵最小化(Grandvalet和Bengio,2004)。熵可以衡量信息的不确定性,熵越小则表明两个数据分布的重叠程度越高。因此,在半监督学习中,通过最小化模型对未标记数据的预测结果(伪标签)的熵,增强模型的泛化性能(Lee,2013)。

2)一致性正则化。在半监督学习背景下,对于受到扰动的未标记数据,模型的预测输出应当是一致的。Laine和Aila(2017)及Tarvainen和Valpola(2017)选取L_2正则化来计算一致性正则项,Miyato等人(2019)和Sohn等人(2020)则选取KL散度来计算。不仅如此,还有一部分研究者将上述两种方法融合使用(Berthelot等,2019a,2019b;Xie等,2020)。

4 GAN中的优化方法

GAN已作为一种典型的深度学习模型得到广泛应用(段宾 等,2020)。与传统深度学习不同之处在于,其目标是学习真实数据分布后生成逼真的新样本。GAN模型由生成网络(G)和判别网络(D)组成,G通过噪声(z)来生成假样本,而D则致力于将真假样本区分开来,二者在迭代过程中博弈并进化。但GAN的训练过程存在挑战性,主要包括Lipschitz限制和模式坍塌现象(2018)。

1)Lipschitz限制。Arjovsky等人(2017)证明D的目标函数的梯度满足Lipschitz连续是稳定训练的关键。

2)模式坍塌现象(Yoshida和Miyato,2017)。输入任何噪声,都生成相似假样本导致训练失败的情况。

上述问题分别对应着优化问题中的拟合能力与泛化能力。因此,研究学者充分利用各种优化方法,研究适应GAN模型的变体模型。

4.1 拟合能力

4.1.1 收敛性

若D的目标函数的梯度满足Lipschitz限制,任意的数据点$x_i\in\mathbf{R},x_j\in\mathbf{R},i,j=1,\cdots,n$,存在一个大于等于0的Lipschitz常数$K$满足以下等式,即

$$\|D\|_{\text{Lip}}=\frac{\left|D(x_i)-D(x_j)\right|}{\left|x_i-x_j\right|}=K \tag{20}$$

目前，主要有3种方法来达到满足Lipschitz限制的目的。第1种方法是Arjovsky等人(2017)直接通过权重裁剪令权重参数有界，间接令梯度满足Lipschitz限制。然而该方法会使得神经网络中的权重参数大部分都趋于边界值，同时这个边界值的选取也是一个难题。因此，Gulrajani等人(2017)提出第2种方法，对梯度进行惩罚可以直接将相对于样本空间的梯度限制在一定范围内。但对于K的取值仍存在争议，Gulrajani等人(2017)和Kodali等人(2017)认为K值应取1，即满足1-Lipschitz限制；而一部分研究者则认为K值应取0(Mescheder等，2018；Zhou等，2019；Thanh-Tung等，2019)，Wu等人(2018)也认为K值应取0，更是通过数学理论推翻Lipschitz限制这一说法，认为该方法有效的根本原因是由于满足Wasserstein散度公式；Petzka等人(2018)则宽泛地认为$K\leqslant 1$即可。第3种方法是对网络中的每层输出使用归一化技术。目前，许多常见的归一化方法均已用于GAN的变体当中，如批归一化(Miyato等，2018；Kurach等，2019)、层归一化(Miyato等，2018；Kurach等，2019)、权重归一化(Miyato等，2018)、实例归一化(Karras等，2019)和条件归一化(Miyato等，2018；Zhang等，2019b)。Chen等人(2019)基于批归一化提出Self-Modulation对输出进行归一化，不以外部信息为条件，仅以生成器的输入为条件。

GAN的变体模型通常在CIFAR-10(Canadian Institutes for Advanced Research)(Krizhevsky，2009)、STL-10(self-taught learning)(Coates等，2011)、ImageNet(Deng等，2009)、CelebA(large-scale CelebFaces attributes)(Liu等，2015)以及LSUN(large-scale scene understanding)的bedroom(Yu等，2015)数据集上进行训练和测试，并以IS(inception score)(Salimans等，2016)和FID(Fréchet inception distance)(Heusel等，2017)作为衡量指标。其中，IS越高表明生成的样本清晰且具有多样性，FID衡量生成样本分布与真实样本分布之间的差距，越低则说明生成的样本更加逼真。表1和表2分别展示为了满足Lipschitz限制的各优化方法在这5个数据集上的IS结果以及FID结果。

表1　各优化方法在不同数据集上的IS结果

Table 1　IS results of each optimization method on different datasets

	方法	CIFAR-10	STL-10	ImageNet	CelebA	LSUN
权重裁剪	Arjovsky等人(2017)	6.41±0.11(−)	−	7.57±0.10(−)	−	−
梯度惩罚	$K\to1$(Gulrajani等，2017)	6.68±0.06	8.42±0.13	12.17±0.09	2.80±0.08	4.72±0.11
	$K\to1$(Kodali等，2017)	6.11±0.05	−	−	−	−
	$K\to0$(Wu等，2018)	−	−	−	−	−
	$K\to0$(Mescheder等，2018)	6.45±0.12(−)	−	17.18±0.06(−)	−	−
	$K\to0$(Zhou等，2019)	8.03±0.03	−	8.67±0.04	−	−
	$K\to0$(Thanh-Tung等，2019)	11.42±0.09(−)	−	−	−	−
	$K\leqslant1$(Petzka等，2018)	7.99±0.12	−	−	−	−
对每层输出归一化	批归一化(Miyato等，2018)	6.27±0.10	−	−	−	−
	层归一化(Miyato等，2018)	7.19±0.12	7.61±0.12	−	−	−
	权重归一化(Miyato等，2018)	6.84±0.07	7.16±0.10	−	−	−
	谱归一化(Miyato等，2018)	7.58±0.12	8.79±0.14	11.29±0.12	3.26±0.16	4.32±0.17
	条件归一化(Zhang等，2019b)	−	−	12.52±0.03(−)	−	−
	自调制归一化(Chen等，2019)	7.71±0.59	−	11.52±0.07	2.92±0.13	5.28±0.18

注："−"表示对应文献中无数据，"(−)"表示对应原文献无数据，数值由本文利用文献代码得到。

表2　各优化方法在不同数据集上的FID结果

Table 2　FID results of each optimization method on different datasets

	方法	CIFAR-10	STL-10	ImageNet	CelebA	LSUN
权重裁剪	Arjovsky等人(2017)	42.6(-)	-	64.2(-)	-	24.6
梯度惩罚	$K\to1$(Gulrajani等,2017)	40.2(-)	55.1(-)	86.23(-)	30.02(-)	26.8(-)
	$K\to1$(Kodali等,2017)	-	-	-	-	-
	$K\to0$(Wu等,2018)	18.1	-	-	21.5	15.9
	$K\to0$(Mescheder等,2018)	-	-	-	-	-
	$K\to0$(Zhou等,2019)	15.64	-	14.9	-	-
	$K\to0$(Thanh-Tung等,2019)	-	-	-	-	-
	$K\leqslant1$(Petzka等,2018)	-	-	-	-	-
对每层输出归一化	批归一化(Miyato等,2018)	56.3	-	-	-	-
	层归一化(Miyato等,2018)	33.9	75.6	-	-	-
	权重归一化(Miyato等,2018)	34.7	73.4	-	-	-
	谱归一化(Miyato等,2018)	25.5	43.2	27.62	26.15	17.1
	条件归一化(Zhang等,2019b)	-	-	18.65	-	-
	自调制归一化(Chen等,2019)	26.93	-	78.31	24.5	14.32

注:"-"表示对应文献中无数据,"(-)"表示对应原文献无数据,数值由本文利用文献代码得到。

从表1—表2中数据可以看出,在训练GAN变体模型时,对梯度进行惩罚或对每层输出采取归一化策略可以明显改善生成样本的效果,因此现已衍生出许多相关的优化方法。权重裁剪也有优化效果,但鉴于其存在严重的参数集中化问题,现已基本弃用该优化方法。

除了令D的梯度满足Lipschitz限制外,不平衡训练同样能稳定训练以确保收敛性,如表3所示。Goodfellow等人(2020)在实验中已经证明多重更新D后更新一次G的不平衡训练更适合GAN,Arjovsky等人(2017)在实验中也同样使用了多重更新D的训练方法。Heusel等人(2017)提出TTUR(two time-scale update rule),控制G和D的学习率来达成不平衡学习。当然,也有研究学者将这两种不平衡训练方法同时使用(Karras等,2019),但对于参数的确定仍有待进一步研究。

表3　各不平衡训练方法在GAN变体模型中的应用

Table 3　Application of unbalanced training methods in variants of GAN model

不平衡训练方法	方法介绍
多重更新(Goodfellow等,2020)	提出参数num用于不平衡训练,更新num次D后更新1次G,实验部分选定num为1。
多重更新(Arjovsky等,2017)	实验部分选定num为5效果最佳,即更新5次D后更新1次G。
不平衡学习率(Heusel等,2017)	G和D使用不同的学习率进行训练,即两个时间尺度的更新规则(TTUR),证明该方法比使用多重更新的训练方法提高了效率。
两种不平衡训练方法相结合(Brock等,2019)	实验部分更新2次D后更新1次G,同时将G的学习率改为D的一半,将两个不平衡训练方法结合使用。

4.1.2　收敛速度

鉴于优化算法的更新梯度的作用,Goodfellow等人(2020)在设计原始GAN时使用SGD算法。而GAN的目的是通过G和D博弈达到纳什均衡状态,即找到鞍点。因此用于跳出鞍点的优化算法就不适用于GAN,如带动量的SGD。目前,自适应梯度的优化算法仍是训练GAN时的最佳选择,如RMSProp(Arjovsky等,2017)和Adam(Gulrajani等,2017),但

注意使用Adam时要将动量项的参数几乎降为0。

为可视化各优化算法的效果，设计以下实验进行比较。MNIST（mixed national institute of standards and technology）数据集（LeCun等，1998）包含0～9共10类28×28像素的图像，每类有6 000幅图像，其中5 000幅为训练样本，剩余为测试样本。考虑到WGAN-GP（WGAN with gradient penalty）（Gulrajani等，2017）具有较高的稳定性，常作为GAN变体模型的基线来比较，故本实验将通过WGAN-GP在MNIST数据集上比较6种优化算法的效果，即SGD、带动量的SGD、Adagrad、Adadelta、RMSProp以及Adam。同时为了保证实验的单一变量原则，除优化算法外其余变量保持不变，且不使用其他优化方法，共训练50 000次迭代，将梯度惩罚的大小作为衡量指标，越小则表明模型的收敛程度越好，间接说明对应优化算法的效果越好。收敛曲线如图5所示，模型应用Adam优化算法生成的样本如图6所示。

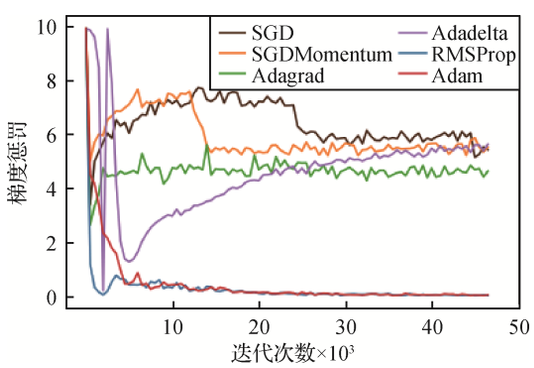

图5　各优化算法在WGAN-GP上的收敛曲线

Fig. 5　The convergence curve of each optimization algorithm on WGAN-GP

从实验结果来看，SGD很难收敛到全局极小值，而带动量的SGD同样没有逃脱鞍点，Adagrad和Adadelta算法作为SGD改进版，直接累加梯度的平方，易收敛于局部极小值，导致训练失败。而RMSProp优化算法则在AdaGrad的基础上添加了一个衰减系数来控制历史信息的获取，在非凸条件下有良好的表现。兼顾Adagrad和RMSPSprop优点的Adam算法能在训练时快速看到效果，同时对超参数相对没有那么敏感，因此在近几年GAN变体模型训练中广泛使用。

4.1.3　全局质量问题

由于GAN并非像传统深度神经网络那样寻找全局最优解，反而是寻找纳什均衡点。神经网络的

图6　应用Adam优化算法的WGAN-GP生成样本

Fig. 6　Using WGAN-GP of Adam optimization algorithm to generate samples

解空间中可能存在着多个纳什均衡点，如何在多次迭代过程中达到稳定状态才是GAN的首要任务。

因此，在设计GAN模型时，应将对收敛性的优化放在首位，保证模型的收敛性使得判别器中目标函数的梯度能满足Lipschitz限制，这是稳定训练的关键。但对收敛速度的优化也是一个不可或缺的部分，一个合适的优化算法能避免模型限于局部极小值，辅助模型寻找最优解。例如，ImprovedGAN（Salimans等，2016）通过消融实验说明，当去除用于提升收敛速度的历史平均梯度更新算法时，IS从6.86±0.06下降至4.36±0.04；而删除保证收敛性的小批次批归一化方法后，IS从8.09±0.12下降至3.87±0.03。故在确保模型稳定训练的基础上，辅以合适的优化算法，才能高效提升模型的拟合能力。

4.2　泛化能力

在GAN模型中，验证模型的拟合能力时更注重数据集的样本，而泛化性更强调面对未知数据时模型的预测能力。当生成样本的数据分布与真实样本的数据分布足够相近时，具有较高的泛化能力。模式坍塌现象是指模型生成的样本缺乏多样性，所以避免模式坍塌问题可以有效提高模型的泛化能力。类似地，用于提升深度神经网络的泛化能力的正则化方法同样广泛应用于GAN的当中，来避免模式坍塌问题。现简述GAN领域目前已有的正则化方法，通过评估模型在测试集上的效果，对比模型使用提升拟合能力的优化方法前后评估指标的变化。表4展示了正则化方法的细节以及优化效果。

表4 GAN领域不同正则化方法的应用及其优化效果

Table 4 Application and optimization effect of different regularization methods in GAN field

	方法	方法介绍及优化效果
限制参数	权重惩罚(Brock 等,2017)	添加$\sum_j \left\| \boldsymbol{W}^j (\boldsymbol{W}^j)^{\mathrm{T}} - \boldsymbol{I} \right\|_p^2$到目标函数中,利用正交正则化对权重进行惩罚,其中$\boldsymbol{I}$是单位矩阵,$\|\cdot\|_p$是$L_p$范数。在SVHN数据集(Netzer等,2011)上训练模型后,在SVM分类器上进行验证,分类错误率由22.18%(DCGAN)降为18.5%。
	权重惩罚(Zhou 等,2018)	添加$\sum_j \left(\left\| \boldsymbol{W}^j \right\|_p - \boldsymbol{I} \right)^2$到目标函数中,利用正交正则化对权重进行惩罚。在CIFAR-10数据集上,对比IS为2.68±0.03的WGAN,正则化后的IS为3.09±0.02,但不及WGAN-GP的效果(3.10±0.03)。
	权重惩罚(Brock 等,2019)	添加$\sum_j \left\| (\boldsymbol{W}^j)^{\mathrm{T}} \boldsymbol{W}^j \odot (1 - \boldsymbol{I}) \right\|_p^2$到目标函数中,利用正交正则化对权重进行惩罚,其中$\odot$为Hadmard乘积。在Google TPUv3 Pod数据集(Google,2018)上以SAGAN为基线模型进行训练,正则化后的IS由98.76±2.84升为99.31±2.10,FID由8.73±0.45降为8.51±0.32。
	权重惩罚(Kurach 等,2019)	添加$\sum_j \left\| \boldsymbol{W}^j \right\|_2$到目标函数中,利用$L_2$正则化对权重进行惩罚。在CelebA数据集上,正则化后FID由40.92升为47.17;在LSUN的bedroom数据集上,FID由136.09升为141.85。
	半监督学习(Arjovsky 等,2017)	通过$\mathrm{clip}\left(\boldsymbol{W}^j, [-0.01, 0.01]\right)$调整权重矩阵。在CIFAR-10数据集上,IS为6.41±0.11,FID为42.6;在STL-10数据集上,IS为7.57±0.10,FID为64.2。
	半监督学习(Miyato 等,2018)	通过$\boldsymbol{W}^j / \sigma(\boldsymbol{W}^j)$调整权重矩阵,$\sigma$为最大奇异值。在CIFAR-10数据集上,IS为7.42±0.08,FID为29.3;在STL-10数据集上,IS为8.28±0.09,FID为53.1。
	半监督学习(Liu 等,2020)	通过$\boldsymbol{W}^j / \sigma(\boldsymbol{W}^j) + \nabla \boldsymbol{W}^j / \sigma(\boldsymbol{W}^j)$调整权重矩阵。与SNGAN进行对比,在CIFAR-10数据集上,IS由8.61升为8.70,FID由21.01降为19.28;在STL-10数据集上,IS由9.10升为9.17,FID由40.11降为39.46;在ImageNet数据集上,IS由21.78升为31.58,FID由71.37降为61.08。
	半监督学习(Zhang 等,2020)	通过$\boldsymbol{W}^j / \sqrt{\left\| \boldsymbol{W}^j \right\|_1 \left\| \boldsymbol{W}^j \right\|_\infty}$调整权重矩阵。与SNGAN进行对比,在CIFAR-10数据集上,IS由3.88±0.14升为4.00±0.17;在ImageNet数据集上,IS由3.78±0.22升为3.87±0.23。
	半监督学习(Odena,2016)	采样未标记样本,与标记样本及生成样本共同输入D中。在MNIST数据集上训练模型后通过D的判别结果进行验证:准确率在采样25个样本时,由0.750升为0.802;采样100个样本时,由0.895升为0.928;采样1 000个样本时,由0.965降为0.964。
处理数据	NI(Sønderby 等,2017)	在真实数据分布中添加实例噪声。在CIFAR-10数据集上对GAN进行NI,FID由34.25降为32.28;在CelebA数据集上对GAN进行NI,FID由34.63降为14.23。
	对抗训练(Zhou和Krähenbühl,2019)	添加一个额外的对抗正则项$\left[D(G(z)) - D(G(z) + \boldsymbol{v})\right]^2$到目标函数中,其中$\boldsymbol{v}$为对抗向量。在CIFAR-10数据集上对GAN进行对抗训练,FID由34.25降为24.06;在CelebA数据集上对GAN进行对抗训练,FID由34.63降为12.04。

权重惩罚与归一化方法作为GAN中一种限制模型参数的正则化方法,已广泛用于解决模态坍塌问题。一般可分为权重惩罚和权重归一化两类。权重惩罚方法将权值矩阵的范数作为正则项增加到GAN的目标函数当中(Brock等,2017;Zhou等,2018;Kurach等,2019;Brock等,2019),这与L_1正则化和L_2正则化的方法相似。权重归一化则主要通过特定的归一化方法不断更新GAN中的权重矩阵(Arjovsky等,2017;Miyato等,2018;Liu等,2020;Zhang等,2020)。

充分利用未标记数据的半监督学习也可用于提升GAN模型的泛化能力。Odena(2016)提出的半监督生成对抗网络(semi-supervised GAN,SGAN)将采样一部分未标记样本,共3部分数据输入到D中。所以判别结果不再仅真假样本两类,而是$N+1$类,其中N为多分类任务的总类数,SGAN的模型如图7所示。

此外,还可通过数据预处理方式对GAN进行正则化。最为常见的方法就是利用噪声对数据进行处理,Sønderby等人(2017)是在去噪过程中反向推导

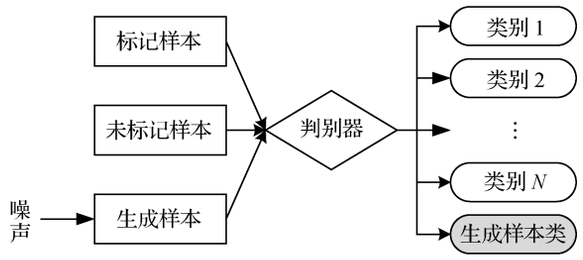

图7　SGAN模型

Fig. 7　Model of SGAN

去噪的梯度估值，可以减小真实数据分布和生成数据分布之间的交叉熵，Jenni 和 Favaro（2019）则证明使用滤波版本的真实数据分布能辅助稳定训练并提高性能。Zhou 和 Krähenbühl（2019）将对抗模型引入到 GAN 当中，使 D 面对对抗攻击具有较强的鲁棒性。

表5展示了近年来效果较好的前沿GAN变体模型所用的优化方法，均出自论文的实验设置部分。不难发现，从拟合能力来看，进行梯度惩罚时大部分研究者更倾向于 Gulrajani 等人（2017）提出的 1-Lipschitz 限制，并采取不平衡训练方法，同时针对具体的模型对每层输出选用不同的归一化方法，其中谱归一化的效果极好，但存在耗时长的缺点。从提升泛化能力来看，Adam 优化算法已经是大部分研究学者的默认选择，同时大部分模型普遍采取权重归一化和 NI 措施来提升模型的鲁棒性。

表5　不同优化方法在前沿GAN变体模型中的应用

Table 5　Application of different optimization methods in state-of-the-art variants of GAN

方法	拟合能力				泛化能力	
	收敛性			收敛速度	限制参数	处理数据
	梯度惩罚	对每层输出归一化	不平衡训练			
PGGAN（Karras 等，2018）	$K\to1$	像素归一化	–	Adam	权重归一化	NI
StarGAN（Choi 等，2018）	$K\to1$	实例归一化	多重更新	Adam	–	对抗训练
SNGAN（Miyato 等，2018）	$K\to1$	谱归一化+批归一化	多重更新	SGD	权重归一化	–
SAGAN（Zhang 等，2019b）	$K\to1$	谱归一化+批归一化	TTUR	Adam	权重归一化	–
BigGAN（Brock 等，2019）	$K\to1$	谱归一化+批归一化	多重更新+TTUR	Adam	权重惩罚	NI
StyleGAN（Karras 等，2019）	$K\to1$	实例归一化	–	Adam	权重归一化	NI
MSGGAN（Karnewar 和 Wang，2020）	$K\to1$	像素归一化+实例归一化	–	RMSProp	权重归一化	NI
MFFGAN（Zhang 等，2021）	$K\to1$	批归一化	多重更新	Adam	权重惩罚	NI

注：“–”表示对应文献中未使用该优化方法。

5　分析与展望

深度学习优化发展迅速，现已成为设计深度神经网络时不可或缺的一部分，但仍存在一些问题有待更进一步的研究。

1）梯度的 Lipschitz 限制。在深度学习优化的收敛性分析中，默认式（2）的全局梯度均满足 Lipschitz 连续限制，如梯度下降理论的基础假设。但现实与理论存在差距，无法保证在全局都能找到 Lipschitz 常数满足限制要求。目前大多数的做法是假设迭代有界，则选定适当步长的梯度下降得以收敛；或添加一个球约束到网络参数中，并使用梯度投影法，但上述方法在理论上无法证明。不仅如此，若找到的 Lipschitz 常数为指数级大/小，会导致梯度爆炸/消失现象。

2）GAN 的全局最优解。由于在 GAN 中寻找博弈后的纳什均衡状态极为困难，所以此前大多数研究都致力于寻找稳定的纳什均衡状态。但 GAN 的解空间中仍存在全局最优解，当生成样本分布收敛到完全与真实样本分布一致时，此时的纳什均衡点即为全局最优解。所以，目前还需要新的理论突破来找到稳定的全局最优解。

3）深度学习优化的可解释性。深度学习中各类

优化方法繁多而复杂,在设计网络时需充分考虑其拟合能力与泛化能力,这也意味着各类优化方法的组合效果是一个值得推敲的问题。但目前仍有许多优化方法属于经验性结论,缺乏严格的理论证明,深度学习优化的可解释性还处于一个初级阶段。在未来的工作中,亟需加强对优化方法的可解释性证明,辅以数学理论的推导可以科学提升模型的优化效果,同时有助于启发新的研究思路。

6 结 语

深度学习中的优化问题是提升模型性能的一个重要研究领域,从模型的拟合能力出发,通过模型优化过程的3个步骤分解优化问题为收敛性问题、收敛速度问题和全局质量问题,并以这3个具体的优化角度来概述现有研究成果。其次,考虑到模型的泛化效果,依据正则化方法的作用原理,分为处理数据和限制模型参数两个方向进行梳理。接着,在GAN的研究背景下,介绍上述优化方法在实际中的应用,基于实验分析给出几种能较好辅助训练的优化方法。最后,对深度学习优化领域中仍存在的问题进行分析,并提出未来的研究方向。通过梳理各种优化方法并在GAN中具体举例,可为设计深度神经网络时选择优化方法提供参考。

参考文献(References)

Abdolahnejad M and Liu P X. 2020. Deep learning for face image synthesis and semantic manipulations: a review and future perspectives. Artificial Intelligence Review, 53(8): 5847-5880 [DOI: 10.1007/s10462-020-09835-4]

Agbo-Ajala O and Viriri S. 2021. Deep learning approach for facial age classification: a survey of the state-of-the-art. Artificial Intelligence Review, 54(1): 179-213 [DOI: 10.1007/s10462-020-09855-0]

Arjovsky M, Chintala S and Bottou L. 2017. Wasserstein generative adversarial networks//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR. org: 214-223 [DOI: 10.5555/3305381.3305404]

Arora S, Li Z Y and Lyu K. 2019. Theoretical analysis of auto rate-tuning by batch normalization//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Ba J L, Kiros J R and Hinton G E. 2016. Layer normalization [EB/OL]. [2021-12-01]. https://arxiv.org/pdf/1607.06450.pdf

Berthelot D, Carlini N, Cubuk E D, Kurakin A, Zhang H and Raffel C. 2019a. ReMixMatch: semi-supervised learning with distribution alignment and augmentation anchoring [EB/OL]. [2021-12-01]. https://arxiv.org/pdf/1911.09785.pdf

Berthelot D, Carlini N, Goodfellow I, Oliver A, Papernot N and Raffel C. 2019b. MixMatch: a holistic approach to semi-supervised learning//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS: #32

Bjorck J, Gomes C, Selman B and Weinberger K Q. 2018. Understanding batch normalization//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 7705-7716 [DOI: 10.5555/3327757.3327868]

Bottou L. 1988. Reconnaissance de la parole par reseaux connexionnistes. Proceedings of Neuro Nimes, 88: 197-218 [DOI: 10.51257/a-v2-h3728]

Brock A, Donahue J and Simonyan K. 2019. Large scale GAN training for high fidelity natural image synthesis//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Brock A, Lim T, Ritchie J M and Weston N. 2017. Neural photo editing with introspective adversarial networks//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net [DOI: 10.48550/arXiv.1609.07093]

Burton Jr R M and Mpitsos G J. 1992. Event-dependent control of noise enhances learning in neural networks. Neural Networks, 5(4): 627-637 [DOI: 10.1016/S0893-6080(05)80040-1]

Cai Y Q, Li Q X and Shen Z W. 2019. A quantitative analysis of the effect of batch normalization on gradient descent//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 882-890

Candès E J, Li X D, Ma Y and Wright J. 2011. Robust principal component analysis? Journal of the ACM, 58(3): #11 [DOI: 10.1145/1970392.1970395]

Candès E J, Romberg J and Tao T. 2006. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory, 52(2): 489-509 [DOI: 10.1109/TIT.2005.862083]

Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P. 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357 [DOI: 10.1613/jair.953]

Chen B F, Li J D, Lu X J, Sha C F, Wang X L and Zhang J. 2021. Survey of deep learning based graph anomaly detection methods. Computer Research and Development, 58(7): 1436-1455 (陈波冯, 李靖东, 卢兴见, 沙朝锋, 王晓玲, 张吉. 2021. 基于深度学习的图异常检测技术综述. 计算机研究与发展, 58(7): 1436-1455) [DOI: 10.7544/issn1000-1239.2021.20200685]

Chen T, Lucic M, Houlsby N and Gelly S. 2019. On self modulation for

generative adversarial networks//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Choi Y, Choi M, Kim M, Ha J W, Kim S and Choo J. 2018. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8789-8797 [DOI: 10.1109/CVPR.2018.00916]

Chu X X, Zhang B and Li X D. 2021. Noisy differentiable architecture search//Proceedings of the 32nd British Machine Vision Conference. Addis Ababa, Ethiopia: BMVA Press

Coates A, Lee H and Ng A Y. 2011. An analysis of single-layer networks in unsupervised feature learning. Journal of Machine Learning Research, 15: 215-223

Cubuk E D, Zoph B, Mané D, Vasudevan V and Le Q V. 2018. AutoAugment: learning augmentation strategies from data//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 113-123 [DOI: 10.1109/cvpr.2019.00020]

Dauphin Y N and Schoenholz S S. 2019. Metainit: initializing learning by learning to initialize//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc.: #1133

Deng J, Dong W, Socher R, Li L J, Li K and Li F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Dinh L, Pascanu R, Bengio S and Bengio Y. 2017. Sharp minima can generalize for deep nets//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR: 1019-1028 [DOI: 10.48550/arXiv.1703.04933]

Donoho D L. 1995. De-noising by soft-thresholding. IEEE Transactions on Information Theory, 41 (3): 613-627 [DOI: 10.1109/18.382009]

Donoho D L. 2006. Compressed sensing. IEEE Transactions on Information Theory, 52(4): 1289-1306 [DOI: 10.1109/TIT.2006.871582]

Draxler F, Veschgini K, Salmhofer M and Hamprecht F A. 2018. Essentially no barriers in neural network energy landscape//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 1309-1318

Duan B, Fu X, Jiang Y and Zeng J X. 2020. Lightweight blurred car plate recognition method combined with generated images. Journal of Image and Graphics, 25(9): 1813-1824 (段宾, 符祥, 江毅, 曾接贤. 2020. 结合GAN的轻量级模糊车牌识别算法. 中国图象图形学报, 25(9): 1813-1824) [DOI: 10.11834/jig.190604]

Duchi J, Hazan E and Singer Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. The Journal of Machine Learning Research, 12: 2121-2159 [DOI: 10.5555/1953048.2021068]

Dumoulin V, Shlens J and Kudlur M. 2017. A learned representation for artistic style//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net: 1-9

Frankle J and Carbin M. 2019. The lottery ticket hypothesis: finding sparse, trainable neural networks//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Garipov T, Izmailov P, Podoprikhin D, Vetrov D and Wilson A G. 2018. Loss surfaces, mode connectivity, and fast ensembling of DNNs//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 8803-8812 [DOI: 10.5555/3327546.3327556]

Ghorbani B, Krishnan S and Xiao Y. 2019. An investigation into neural net optimization via hessian eigenvalue density//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 2232-2241

Gilboa D, Chang B, Chen M M, Yang G, Schoenholz S S, Chi E H and Pennington J. 2019. Dynamical isometry and a mean field theory of LSTMs and GRUs [EB/OL]. [2021-12-01]. https://arxiv.org/pdf/1901.08987.pdf

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Glorot X and Bengio Y. 2010. Understanding the difficulty of training deep feedforward neural networks//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: JMLR.org: 249-256

Glorot X, Bordes A and Bengio Y. 2011. Deep sparse rectifier neural networks//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA: JMLR.org: 315-323

Goodfellow I, Bengio Y and Courville A. 2016. Deep Learning. Cambridge: The MIT Press: 1-100

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2020. Generative adversarial networks. Communications of the ACM, 63(11): 139-144 [DOI: 10.1145/3422622]

Grandvalet Y and Bengio Y. 2004. Semi-supervised learning by entropy minimization//Proceedings of the 17th International Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press: 529-536 [DOI: 10.5555/2976040.2976107]

Gulcehre C, Moczulski M, Denil M and Bengio Y. 2016. Noisy activation functions//Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR.org: 3059-3068 [DOI: 10.48550/arXiv.1603.00391]

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A. 2017. Improved training of wasserstein GANs//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5769-5779

[DOI: 10.5555/3295222.3295327]

Hanin B and Rolnick D. 2018. How to start training: the effect of initialization and architecture//Proceedings of the 32nd Conference on Neural Information Processing Systems. Montréal, Canada: NeurIPS: 571-581

Hanson S J. 1990. A stochastic version of the delta rule. Physica D: Nonlinear Phenomena, 42(1/3): 265-272 [DOI: 10.1016/0167-2789(90)90081-Y]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6629-6640 [DOI: 10.5555/3295222.3295408]

Hinton G, Deng L, Yu D, Dahl G E, Mohamed A R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T N and Kingsbury B. 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Processing Magazine, 29(6): 82-97 [DOI: 10.1109/MSP.2012.2205597]

Hinton G E and Salakhutdinov R R. 2006. Reducing the dimensionality of data with neural networks. Science, 313(5786): 504-507 [DOI: 10.1126/science.1127647]

Huang G, Liu Z, van der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2261-2269 [DOI: 10.1109/CVPR.2017.243]

Ioffe S and Szegedy C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France: JMLR.org: 448-456 [DOI: 10.5555/3045118.3045167]

Jenni S and Favaro P. 2019. On stabilizing generative adversarial training with noise//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 12137-12145 [DOI: 10.1109/CVPR.2019.01242]

Karnewar A and Wang O. 2020. MSG-GAN: multi-scale gradients for generative adversarial networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7796-7805 [DOI: 10.1109/CVPR42600.2020.00782]

Karras T, Aila T, Laine S and Lehtinen J. 2018. Progressive growing of GANs for improved quality, stability, and variation//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4396-4405 [DOI: 10.1109/CVPR.2019.00453]

Khan Z Y, Niu Z D, Sandiwarno S and Prince R. 2021. Deep learning techniques for rating prediction: a survey of the state-of-the-art. Artificial Intelligence Review, 54(1): 95-135 [DOI: 10.1007/s10462-020-09892-9]

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR

Ko T, Peddinti V, Povey D and Khudanpur S. 2015. Audio augmentation for speech recognition//Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany: ISCA: 3586-3589 [DOI: 10.21437/interspeech.2015-711]

Kodali N, Abernethy J, Hays J and Kira Z. 2017. On convergence and stability of GANs [EB/OL]. [2021-12-01]. https://arxiv.org/pdf/1705.07215v1.pdf

Kohler J, Daneshmand H, Lucchi A, Hofmann T, Zhou M and Neymeyr K. 2019. Exponential convergence rates for batch normalization: the power of Length-direction decoupling in non-convex optimization//Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. Naha, Japan: PMLR: 806-815

Krizhevsky A. 2009. Learning Multiple Layers of Features from Tiny Images. Technical Report TR-2009. University of Toronto

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]

Kurach K, Lučić M, Zhai X H, Michalski M and Gelly S. 2019. A large-scale study on regularization and normalization in GANs//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 3581-3590

Laine S and Aila T. 2017. Temporal ensembling for semi-supervised learning//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net: #7

LeCun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of 1998 IEEE, 86(11): 2278-2324 [DOI: 10.1109/5.726791]

LeCun Y A, Bottou L, Orr G B and Müller K R. 2012. Efficient backprop//Montavon G, Orr G B and Müller K R. Neural Networks: Tricks of the Trade. Berlin: Springer: 9-48 [DOI: 10.1007/978-3-642-35289-8_3]

Lee D H. 2013. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks//Proceedings of 2013 Workshop: Challenges in Representation Learning. Atlanta, USA: [s.n.]: 1-6

Li P and Nguyen P M. 2019. On random deep weight-tied autoencoders: exact asymptotic analysis, phase transitions, and implications to

training//Proceedings of the 7th International Conference on Learning Representations. New Orleans, Louisiana, USA: OpenReview.net

Lim S, Kim I, Kim T, Kim C and Kim S. 2019. Fast autoAugment//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS

Liu K L, Qiu G P, Tang W M and Zhou F. 2020. Spectral regularization for combating mode collapse in GANs. Image and Vision Computing, 104: #104005 [DOI: 10.1016/j.imavis.2020.104005]

Liu Z W, Luo P, Wang X G and Tang X O. 2015. Deep learning face attributes in the wild//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 3730-3738 [DOI: 10.1109/ICCV.2015.425]

Loshchilov I and Hutter F. 2017. SGDR: stochastic gradient descent with warm restarts//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net

Luo P, Zhang R M, Ren J M, Peng Z L and Li J Y. 2021. Switchable normalization for learning-to-normalize deep representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2): 712-728 [DOI: 10.1109/TPAMI.2019.2932062]

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2018. Towards deep learning models resistant to adversarial Attacks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Mescheder L, Geiger A and Nowozin S. 2018. Which training methods for GANs do actually converge?//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 3481-3490

Mishkin D and Matas J. 2016. All you need is a good init//Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico, USA: ICLR

Miyato T, Kataoka T, Koyama M and Yoshida Y. 2018. Spectral normalization for generative adversarial networks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Miyato T, Maeda S I, Koyama M and Ishii S. 2019. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8): 1979-1993 [DOI: 10.1109/TPAMI.2018.2858821]

Natarajan B K. 1995. Sparse approximate solutions to linear systems. SIAM Journal on Computing, 24(2): 227-234 [DOI: 10.1137/S0097539792240406]

Neelakantan A, Vilnis L, Le Q V, Sutskever I, Kaiser L, Kurach K and Martens J. 2015. Adding gradient noise improves learning for very deep networks [EB/OL]. [2021-12-01]. https://arxiv.org/pdf/1511.06807.pdf

Netzer Y, Wang T, Coates A, Bissacco A, Wu B and Ng A Y. 2011. Reading digits in natural images with unsupervised feature learning//Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Granada, Spain: [s.n.]: 1-9

Ng A. 2011. Sparse autoencoder [EB/OL]. [2021-12-16]. https://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf

Odena A. 2016. Semi-supervised learning with generative adversarial networks [EB/OL]. [2021-12-01]. https://arxiv.org/pdf/1606.01583.pdf

Pennington J, Schoenholz S S and Ganguli S. 2017. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 4788-4798 [DOI: 10.5555/3295222.3295232]

Pennington J, Schoenholz S S and Ganguli S. 2018. The emergence of spectral universality in deep networks//Proceedings of the 21st International Conference on Artificial Intelligence and Statistics. Playa Blanca, Spain: PMLR: 1924-1932

Petzka H, Fischer A and Lukovnikov D. 2018. On the regularization of wasserstein GANs//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Poole B, Lahiri S, Raghu M, Sohl-Dickstein J and Ganguli S. 2016. Exponential expressivity in deep neural networks through transient chaos//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 3368-3376 [DOI: 10.5555/3157382.3157474]

Qian N. 1999. On the momentum term in gradient descent learning algorithms. Neural Networks, 12(1): 145-151 [DOI: 10.1016/S0893-6080(98)00116-6]

Robbins H and Monro S. 1951. A stochastic approximation method. The Annals of Mathematical Statistics, 22(3): 400-407 [DOI: 10.1214/aoms/1177729586]

Rumelhart D E, Hinton G E and Williams R J. 1986. Learning representations by back-propagating errors. Nature, 323(6088): 533-536 [DOI: 10.1038/323533a0]

Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A and Chen X. 2016. Improved techniques for training GANs//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 2234-2242 [DOI: 10.5555/3157096.3157346]

Salimans T and Kingma D P. 2016. Weight normalization: a simple reparameterization to accelerate training of deep neural networks//Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 901-909 [DOI: 10.5555/3157096.3157197]

Santurkar S, Tsipras D, Ilyas A and Madry A. 2018. How does batch normalization help optimization? //Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 2488-2498 [DOI: 10.5555/3327144.3327174]

Saxe A M, McClelland J L and Ganguli S. 2014. Exact solutions to the

nonlinear dynamics of learning in deep linear neural networks//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR

Shafahi A, Najibi M, Ghiasi A, Xu Z, Dickerson J, Studer C, Davis L S, Taylor G and Goldstein T. 2019. Adversarial training for free!//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #302 [DOI: 10.5555/3454287.3454589]

Shahshahani B M and Landgrebe D A. 1994. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. IEEE Transactions on Geoscience and Remote Sensing, 32(5): 1087-1095 [DOI: 10.1109/36.312897]

Smith L N. 2017. Cyclical learning rates for training neural networks//Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision. Santa Rosa, USA: IEEE: 464-472 [DOI: 10.1109/WACV.2017.58]

Sohn K, Berthelot D, Li C L, Zhang Z Z, Carlini N, Cubuk E D, Kurakin A, Zhang H and Raffe C. 2020. FixMatch: simplifying semi-supervised learning with consistency and confidence//Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS

Sønderby C K, Caballero J, Theis L, Shi W Z and Huszár F. 2017. Amortised map inference for image super-resolution//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net

Srivastava R K, Greff K and Schmidhuber J. 2015. Highway networks [EB/OL]. [2021-05-03]. http://arxiv.org/pdf/1505.00387.pdf

Sun R Y. 2020. Optimization for deep learning: an overview. Journal of the Operations Research Society of China, 8(2): 249-294 [DOI: 10.1007/s40305-020-00309-6]

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I J and Fergus R. 2014. Intriguing properties of neural networks//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR

Tan M X and Le Q V. 2019. EfficientNet: rethinking model scaling for convolutional neural networks//Proceedings of 2019 International Conference on Machine Learning. Long Beach, USA: PMLR: 6105-6114

Tarvainen A and Valpola H. 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 1195-1204

Thanh-Tung H, Tran T and Venkatesh S. 2019. Improving generalization and stability of generative adversarial networks//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Tieleman T and Hinton G. 2012. Lecture 6.5-RMSProp: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 4(2): 26-31

Torfi A, Shirvani R A, Keneshloo Y, Tavaf N and Fox E A. 2020. Natural language processing advancements by deep learning: a survey [EB/OL]. [2021-12-01]. https://arxiv.org/pdf/2003.01200v1.pdf

Ulyanov D, Vedaldi A and Lempitsky V. 2016. Instance normalization: the missing ingredient for fast stylization [EB/OL]. [2021-12-01]. https://arxiv.org/pdf/1607.08022v3.pdf

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010 [DOI: 10.5555/3295222.3295349]

Vincent P, Larochelle H, Bengio Y and Manzagol P A. 2008. Extracting and composing robust features with denoising autoencoders//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM: 1096-1103 [DOI: 10.1145/1390156.1390294]

Wei J and Zou K. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics: 6382-6388 [DOI: 10.18653/v1/d19-1670]

Werbos P J. 1974. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Cambridge, USA: Harvard University

Wu J Q, Huang Z W, Thoma J, Acharya D, and van Gool L. 2018. Wasserstein divergence for GANs//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 673-688 [DOI: 10.1007/978-3-030-01228-1_40]

Wu Y X and He K M. 2020. Group normalization. International Journal of Computer Vision, 128(3): 742-755 [DOI: 10.1007/s11263-019-01198-w]

Xiao L C, Bahri Y, Sohl-Dickstein J, Schoenholz S S and Pennington J. 2018. Dynamical isometry and a mean field theory of CNNs: how to train 10,000-layer vanilla convolutional neural networks//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 5393-5402

Xie L X, Wang J D, Wei Z, Wang M and Tian Q. 2016. DisturbLabel: regularizing CNN on the loss layer//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4753-4762 [DOI: 10.1109/CVPR.2016.514]

Xie Q Z, Dai Z H, Hovy E, Luong M T and Le Q V. 2020. Unsupervised data augmentation for consistency training//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc: #3496249 [DOI: 10.5555/3495724.3496249]

Xie S N, Girshick R, Dollár P, Tu Z W and He K M. 2017. Aggregated residual transformations for deep neural networks//Proceedings of

2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5987-5995 [DOI: 10.1109/CVPR.2017.634]

Yoshida Y and Miyato T. 2017. Spectral norm regularization for improving the generalizability of deep learning [EB/OL]. [2021-12-01]. https://arxiv.org/pdf/1705.10941v1.pdf

Yu A W, Dohan D, Luong M T, Zhao R, Chen K, Norouzi M and Le Q V. 2018. QANET: combining local convolution with global Self-attention for reading comprehension//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Yu F, Seff A, Zhang Y D, Song S R, Funkhouser T and Xiao J X. 2015. LSUN: construction of a Large-scale image dataset using deep learning with humans in the loop [EB/OL]. [2021-12-01]. https://arxiv.org/pdf/1506.03365.pdf

Zhang D H, Zhang T Y, Lu Y P, Zhu Z X and Dong B. 2019a. You only propagate once: accelerating adversarial training via maximal principle//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS

Zhang H, Goodfellow I, Metaxas D and Odena A. 2019b. Self-attention generative adversarial networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 7354-7363

Zhang H, Le Z L, Shao Z F, Xu H and Ma J Y. 2021. MFF-GAN: an unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. Information Fusion, 66: 40-53 [DOI: 10.1016/j.inffus.2020.08.022]

Zhang Z H, Zeng Y B, Bai L, Hu Y Q, Wu M H, Wang S and Hancock E R. 2020. Spectral bounding: strictly satisfying the 1-lipschitz property for generative adversarial networks. Pattern Recognition, 105: #107179 [DOI: 10.1016/j.patcog.2019.107179]

Zhou B and Krähenbühl P. 2019. Don't let your discriminator be fooled//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Zhou C S, Zhang J S and Liu J M. 2018. Lp-WGAN: using Lp-norm normalization to stabilize wasserstein generative adversarial networks. Knowledge-Based Systems, 161: 415-424 [DOI: 10.1016/j.knosys.2018.08.004]

Zhou Z M, Liang J D, Song Y X, Yu L T, Wang H W, Zhang W N, Yu Y and Zhang Z H. 2019. Lipschitz generative adversarial nets//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 7584-7593

Zhuo L A, Zhang B C, Chen C, Ye Q X, Liu J Z and Doermann D. 2019. Calibrated stochastic gradient descent for convolutional neural networks//Proceedings of the 31st Innovative Applications of Artificial Intelligence Conference. Honolulu, USA: AAAI: 9348-9355 [DOI: 10.1609/aaai.v33i01.33019348]

Zoph B, Cubuk E D, Ghiasi G, Lin T Y, Shlens J and Le Q V. 2020. Learning data augmentation strategies for object detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 566-583 [DOI: 10.1007/978-3-030-58583-9_34]

Zoph B and Le Q V. 2017. Neural architecture search with reinforcement learning//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net

Zou H and Hastie T. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2): 301-320 [DOI: 10.1111/j.1467-9868.2005.00503.x]

作者简介

江铃燚，女，硕士研究生，主要研究方向为深度学习。
E-mail: lorrainejiang970@163.com
郑艺峰，通信作者，男，讲师，主要研究方向为数据挖掘、特征工程、机器学习和深度学习。E-mail: zyf@mnnu.edu.cn
陈澈，男，硕士研究生，主要研究方向为移动边缘计算和深度强化学习。E-mail: cc5551@foxmail.com
李国和，男，教授，博士生导师，主要研究方向为人工智能、机器学习和知识发现。E-mail: lgh102200@sina.com
张文杰，男，教授，主要研究方向为移动边缘计算和深度强化学习。E-mail: zhan0300@ntu.edu.sg