Learning Not to Learn: Training Deep Neural Networks With Biased Data

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, Junmo Kim

摘要

论文提出了一种新的正则化算法来训练深度神经网络，其中训练时的数据是存在严重偏置的。由于神经网络可以有效地学习数据分布，因此网络很可能会学习偏置信息来对输入数据进行分类。如果偏置实际上与分类无关，则会导致测试时表现不佳。本文提出了一个基于特征嵌入和偏置之间的互信息的正则化损失。基于互信息最小化的思想，我们提出了一种迭代算法来消除偏差信息。我们使用一个额外的网络来预测偏置分布，并与特征嵌入网络进行对抗性训练。在学习结束时，偏置预测网络不能预测偏置，不是因为它训练得不好，而是因为特征嵌入网络成功地取消了偏置信息。我们还展示了定量和定性的实验结果，表明我们的算法有效地去除了特征嵌入中的偏置信息。。

关键词：神经网络；数据偏置

1 引言

机器学习算法和人工智能已被广泛应用于各个领域。越来越多的应用程序导致了对健壮算法的巨大需求。最理想的训练神经网络的方法是使用合适的无偏置的数据。然而，收集分布良好的数据往往需要付出很大的努力。此外，对于什么是分布良好的数据，还缺乏共识。

除了哲学问题，数据分布还显著影响网络的特征，因为目前基于深度学习的算法直接从输入数据中学习。如果在训练过程中提供有偏置的数据，机器将会将有偏置的分布视为有意义的信息。这种情况是不希望出现的，因为它削弱了算法的鲁棒性，并可能引入不公平的歧视。类似的概念在文献中也有探讨，被称为 unknowns[1]。作者将 unknows 分类如下: 已知的未知和未知的未知。区分这些类别的关键标准是训练过的模型所作预测的置信度。未知未知对应的是模型预测错误且置信度高的数据点，如 softmax 得分高，而已知未知对应的是模型预测错误且置信度低的数据点。由于分类器的置信度较低，已知的未知有更好的机会被检测到，而由于分类器产生较高的置信度分数，未知的未知很难被检测到。在本文中，我们考虑的数据偏置与 [1] 中的未知的未知类似。但是与 [1] 中的未知的未知不同，偏置并不代表数据本身。相反，偏置代表一些属性，例如肤色、种族或性别。

我们的主要贡献可以总结如下: 首先，我们提出了一个新的基于互信息的正则化项，以消除给定数据中的目标偏置。其次，我们通过实验表明，提出的正则化项最小化了数据中偏置的有害影响。通过从特征嵌入中去除与目标偏置相关的信息，网络能够学习更多信息的特征进行分类。在所有实验中，用提出的正则化损失训练的网络表现出性能的提高。

2 相关工作

在本节中，我们将介绍在防止模型学习偏置的最先进技术。这些技术可以分为 (但不限于) 三种主要方法: 直接从源数据去偏，使用 GANs/集成来进行数据去偏，以及在训练过的模型中直接学习去偏。

2.1 基于数据源去除偏置

众所周知，数据集通常都存在着偏置。在 Torralba 和 Efros[2]的工作中，他们展示了偏置如何影响一些最常用的数据集，并考虑了训练过的 ANN 模型的泛化性能和分类能力。采用类似的方法，Tommasi[3]等人进行了实验，报告了多个数据集之间的差异，并验证了在应用不同的去偏策略以平衡数据时，最终的性能如何变化。在数据集级别上工作通常是一个关键方面，它极大地有助于理解数据及其结构[4]。Gupta 等人在实践和经验的背景下探索了通过使用不同来源的数据来消除偏见的概念[5]。特别是，他们设计了一种去偏策略，通过减少不平衡数据的影响，将不完善的执行和校准误差的影响降到最低，显示出最终模型在泛化方面的改进。

2.2 基于对抗训练去除偏置

对损失函数中的偏置分布给出一个明确的公式通常是困难的。一种可能的方法是使用额外的模型来学习数据中的偏置，并使用它们来调整主要模型，使其避免偏置。另一种可能是使用灰度共现矩阵提取无偏特征并训练模型，正如 Wang 等人用 HEX[6]提出的那样。Alvi 等人提出了 BlindEye[7]技术，他们在提取的深度特征上训练分类器以从偏置中检索信息: 然后，他们迫使“偏置分类器”不再能够检索与偏差相关的信息，相应地修改深度特征。Bahng[8]等人开发了一种基于集成的技术，称为 ReBias。它是解决一个最小-最大问题，目标是提高网络预测与所有有偏差预测之间的独立性。

2.3 基于深度模型去除偏置

从数据集中去除偏置有助于训练过程，因为训练是在无偏置的数据上地执行的。然而，使用这种方法，我们通常无法直接控制从数据集本身删除的信息，或者我们包含了极高的计算复杂度，就像训练 GANs 时一样。相反，亨德里克斯等人提出了一个我们可以直接接触到这些偏见的环境[9]。在这些工作中，他们明确地引入一个校正损失项 (与 Vinyals 等人提出的公式一致[10])，目的是帮助 ANN 模型专注于正确的特征。类似地，Cadene 等人提出了 RUBi[11]，其中他们使用 logit 重加权来降低学习过程中的偏差影响，而 Sagawa 等人提出了 Group-DRO[12]，通过定义先验数据子组并控制其泛化来避免偏差过拟合。

3 本文方法

3.1 符号定义

在本节中，我们提出了一种新的正则化损失，它最大限度地减少了有偏置数据的不良影响。在引入公式之前，我们先定义以下符号。假设我们有一个图像 $x \in X$ 和对应的标签 $y_x \in Y$，我们定义一个偏置集合 B，它包含 $x \in X$ 可以拥有的所有可能的目标偏置。我们还定义了一个潜在函数 $b : X \in B$，其中 $b(x)$ 表示 x 的目标偏置。我们还定义了随机变量 X 和 Y，它们的值分别为 x 和 y_x

图像 x 输入到特征提取网络 $f : X \rightarrow R^k$，其中 K 为 f 嵌入特征的维数。随后，提取的特征 $f(x)$ 通过预测网络 $g : R^k \rightarrow Y$ 和偏置预测网络 $h : R^k \rightarrow B$ 向前反馈。每个网络的参数定义为 θ_f，θ_g，和 θ_h 下标表示其具体的网络。图 1描述了神经网络的总体架构。我们没有明确指定一个详细的体系结构，因为我们的正则化损失适用于任意的网络体系结构。

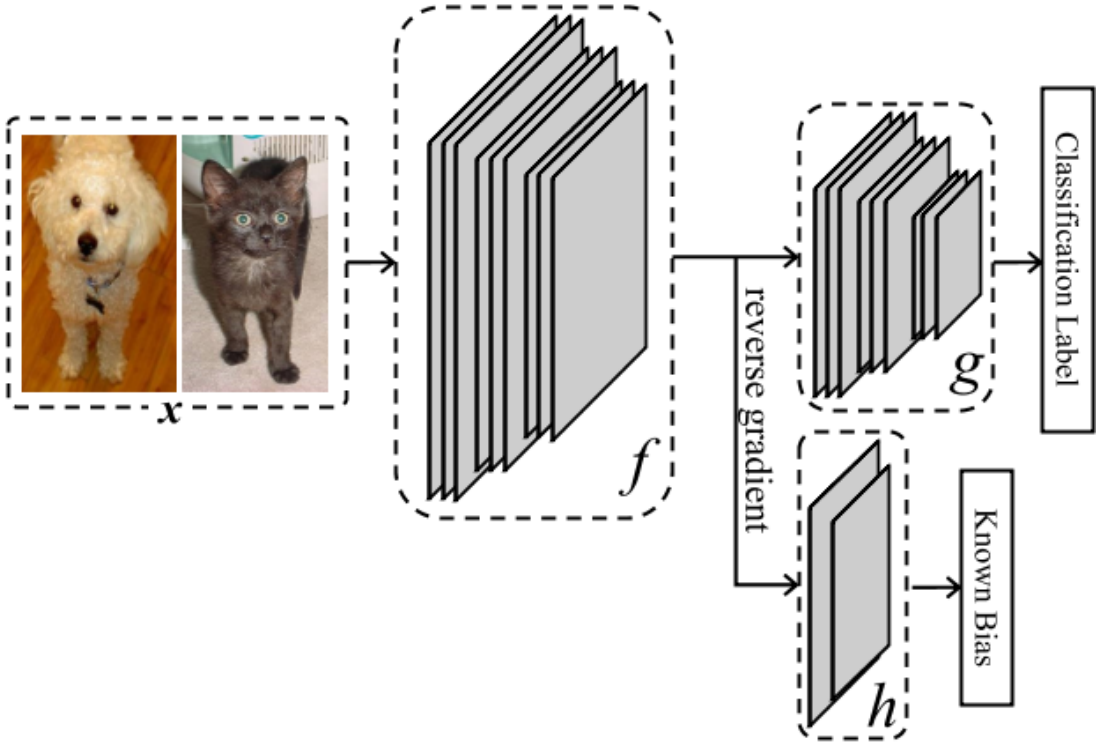

图 1: 模型示意图

3.2 方法概述

我们工作的目标是训练一个在测试期间使用无偏数据也能够稳健执行的网络，即使网络是用有偏数据训练的。数据偏置有着以下特征：

$$I\left(b\left(X^{train}\right);Y\right) \gg I\left(b\left(X^{test}\right);Y\right) \approx 0 \tag{1}$$

其中 X^{train} and X^{test} 分别表示训练和测试过程中采样的随机变量, 有偏置的训练数据导致有偏置的网络，因此网络严重依赖于数据的偏置:

$$I\left(b\left(x\right);g\left(f\left(x\right)\right)\right) \gg 0 \tag{2}$$

为此，我们在训练网络的目标函数中加入互信息。我们最小化了 $f\left(x\right)$ 的互信息, 而不是 $g\left(f\left(x\right)\right)$。这是因为标签预测网络,$g$, 以 $f\left(x\right)$ 为输入。从预测网络 g 的角度看, 如果网络 f 不提取目标偏差的信息，则训练数据是无偏差的。换句话说，提取的特征 $f\left(x\right)$ 应该不包含目标偏差 $b\left(x\right)$ 的信息。因此，训练过程是优化以下问题:

$$\min_{\theta_f,\theta_g} \mathbb{E}_{\tilde{x}\sim P_X(\cdot)}\left[\mathcal{L}_c\left(y_{\tilde{x}}, g(f(\tilde{x}))\right)\right] + \lambda\mathcal{I}(b(X); f(X)) \tag{3}$$

其中 $\mathcal{L}_c(\cdot,\cdot)$ 表示交叉熵损失，λ 是平衡项的超参数。式 (3) 中的互信息可等效表示为:

$$\mathcal{I}(b(X); f(X)) = H(b(X)) - H(b(X) \mid f(X)) \tag{4}$$

式中，$H(\cdot)$ 和 $H(\cdot|\cdot)$ 分别为边际熵和条件熵。由于偏置的边际熵是恒定的，不依赖于 θ_f 和 θ_g, $H\left(b\left(x\right)\right)$ 可以从优化问题中省略，我们试图最小化负熵，$-H(b(X) \mid f(X))$. 式 (4) 很难直接最小化，因为它需要后验分布 $P(b(X) \mid f(X))$ 由于它在实践中难以处理，将式 (4) 最小化使用辅助分布 Q 重新表述，并附加了一个等式约束:

$$\min_{\theta_f} \mathbb{E}_{\tilde{x}\sim P_X(\cdot)}\left[\mathbb{E}_{\tilde{b}\sim Q(\cdot|f(\tilde{x}))}[\log Q(\tilde{b} \mid f(\tilde{x}))]\right] \tag{5}$$

使用分布 Q 的好处是我们可以直接计算目标函数。因此，我们可以在等式约束下训练特征提取网络 f。

由于式 (5) 中的等式约束很难满足 (特别是在训练过程的开始阶段)，我们将等式约束修改为 P，Q 之间的 KL 散度最小，使 Q 随着学习的进行越来越接近 P。我们放松式 (5)，使辅助分布，Q 可以用来近似后验分布。松弛正则化损失如下：

$$\begin{aligned}\mathcal{L}_{MI} = \mathbb{E}_{\tilde{x}\sim P_X(\cdot)}\left[\mathbb{E}_{\tilde{b}\sim Q(\cdot|f(\tilde{x}))}[\log Q(\tilde{b}\mid f(\tilde{x}))]\right] \\ + \mu D_{KL}(P(b(X)\mid f(X))\|Q(b(X)\mid f(X)))\end{aligned} \tag{6}$$

其中 D_{KL} 表示 kL 散度，μ 是平衡这两项的超参数。与 Chen $et\ al$[13]提出的方法类似，我们将辅助分布 Q 参数化为 p 偏置预测网络 h。注意，我们将训练网络 h，使 KL 散度最小化。假设网络 h 实现的分布 Q 收敛于 $P(b(X)\mid f(X))$，我们只需要训练网络 f，使式 (6) 中的第一项最小化。

虽然后验分布 $P(b(X)\mid f(X))$ 是不可处理的，但如果我们用 SGD 优化器训练以 $b: X\in B$ 为标签的网络，则偏置预测网络 h 可被训练成随机近似 $P(b(X)\mid f(X))$。因此，我们以 $b: X\in B$ 和 $h\left(f\left(x\right)\right)$ 之间的交叉熵损失为期望，放松式 (6) 的 KL 散度，并训练网络 h，使偏置预测损失 $\mathcal{L}_{\mathcal{B}}$ 最小化。

$$\mathcal{L}_{\mathcal{B}}\left(\theta_f,\theta_h\right)=\mathbb{E}_{\tilde{x}\sim P_X(\cdot)}\left[\mathcal{L}_c(b(\tilde{x}),h(f(\tilde{x})))\right] \tag{7}$$

虽然单独训练网络 h 来最小化式 (7) 足以使 Q 更接近 P，但以对抗的方式训练 f 来最大化式 (7) 更好，即让网络 f 和 h 进行极大极小博弈。直观地说，由网络 f 提取的特征使偏差预测变得困难。由于 f 被训练为最小化式 (6) 中的第一项，我们可以用 $\mathcal{L}_{\mathcal{B}}$ 代替 KL 散度重新表述式 (6)，如下所示:

$$\begin{aligned}\min_{\theta_f}\max_{\theta_h}\mathbb{E}_{\tilde{x}\sim P_X(\cdot)}\left[\mathbb{E}_{\tilde{b}\sim Q(\cdot|f(\tilde{x}))}[\log Q(\tilde{b}\mid f(\tilde{x}))]\right] \\ -\mu\mathcal{L}_{\mathcal{B}}\left(\theta_f,\theta_h\right)\end{aligned} \tag{8}$$

我们训练 f 从其特征嵌入 $f\left(x\right)$ 中正确预测偏差 $b\left(x\right)$。我们训练 f 最小化负条件熵。在最小化负条件熵的情况下，网络 h 是固定的。网络 f 也被训练为最大化交叉熵以抑制 h 预测 $b\left(x\right)$。结合原始分类问题，极小极大对策公式如下:

$$\begin{aligned}\min_{\theta_f,\theta_g}\max_{\theta_h}\mathbb{E}_{\tilde{x}\sim P_X(\cdot)}\left[\mathcal{L}_c\left(y_{\bar{x}},g(f(\tilde{x}))\right)\right. \\ \left.+\lambda\mathbb{E}_{\tilde{b}\sim Q(\cdot|f(\tilde{x}))}[\log Q(\tilde{b}\mid f(\tilde{x}))]\right] \\ -\mu\mathcal{L}_{\mathcal{B}}\left(\theta_f,\theta_h\right)\end{aligned} \tag{9}$$

在实验中，深度神经网络 f, g 和 h 同时使用对抗策略[14]和梯度反转技术[15]进行训练。在训练的初期，$g\circ f$ 迅速的学会使用偏置特征进行分类。然后 h 学会了预测偏置，f 开始学习如何提取独立于偏置的特征嵌入。在训练结束时，h 回归到表现较差的网络，不是因为偏置预测网络 h 没有训练好，而是因为 f 不学习偏置，所以特征嵌入 $f\left(x\right)$ 没有足够的信息来预测目标偏置。

4 复现细节

4.1 与已有开源代码对比

本次实验采用了原作者提供的代码和数据集，网络模型采用的是原作者提供的代码，在源代码的基础上我们改进了原文中提到的梯度反转层，加入了一个适应性的参数，使得网络结构更为合理，并且在偏置严重时取得了更好的效果。

4.2 创新点

原文采用了对抗训练和梯度反转的思想来去除偏置，详细的网络结构如图 2所示，梯度反转的目的是为了让特征提取网络学会不提取偏置信息，但是原文中衡量特征提取网络效果的标准是按照偏置识别网络是否能够很好的识别偏置来判定的，这样即使偏置识别网络不能很好的识别偏置也不能保证是因为提取的特征中不包含偏置信息，因为这有可能是偏置识别网络没收敛导致的，因此我们在原文的梯度反转层中加入了一个随训练轮数变化的参数 λ, 用来改善这种情况，确保特征提取网络提取无偏置的特征，具体公式如下：

$$\lambda = \frac{2}{1 + \exp(-\gamma)} - 1 \tag{10}$$

其中 γ 为训练轮数，λ 随着训练轮数的增加逐渐由 0 变为 1，保证了偏置识别网络能够收敛，实验结果表明在偏置严重时，该改进措施提高了预测网络的准确率。

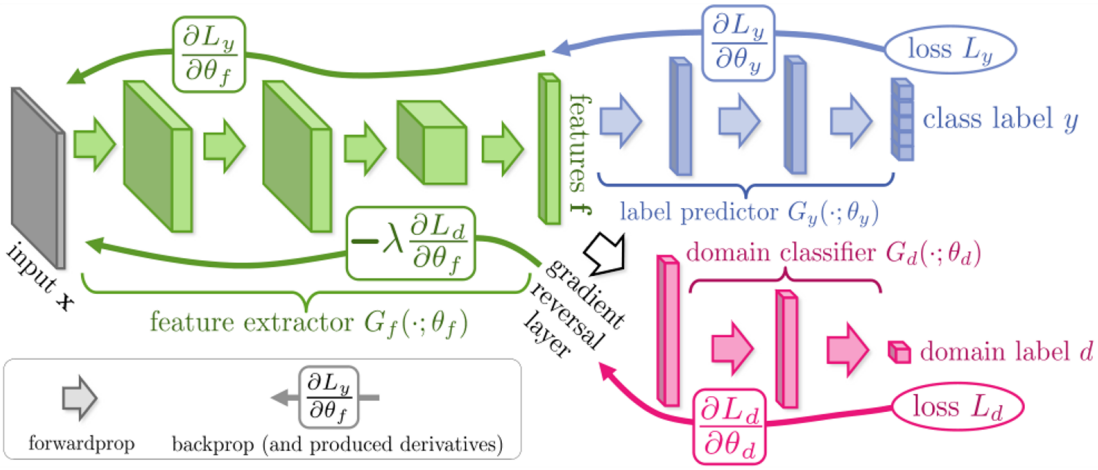

图 2: 网络详细结构

5 实验结果分析

5.1 数据集

大多数现有的基准测试旨在评估特定的问题。收集者经常巧妙地将数据集分成训练/测试集，并且保持训练集和测试集的数据分布相似，这会对我们的实验结果产生影响。因此，我们有意地在平衡的数据集中中植入偏置，以确定我们的算法是否可以消除偏置。

我们在 MNIST[16]中植入了颜色偏置。为了综合颜色偏置，我们选择了十种不同的颜色，并将它们分配到每个数字类别中作为它们的平均颜色。然后，对每个训练图像，从对应的平均颜色的正态分布中按照给定的方差随机抽样一种颜色对数字进行着色。方差越小代表偏置越严重，测试集则是无偏的数据。

5.2 实验结果

我们将方差从 0.02 变化到 0.05，区间为 0.005。因此，Train-0.02 是偏差最大的集合，而 Train-0.05 是偏差最小的集合。图 3显示了有偏置 MNIST 的实验结果。基线模型代表一个没有额外正则化训练的网络，基线模型性能大致可以用作训练数据偏置的基准。BlindEye 算法代表了一个用混淆损失[7]训练的网络，而不是我们的正则化。另一种算法，记为“Gray”，表示用灰度图像训练的网络。对于给定的彩色偏置数据，我们将彩色数字转换为灰度。转换成灰度是一种可以用来减轻颜色偏置的简单方

法。我们假设转换为灰度并不会显著减少信息，因为 MNIST 数据集最初是灰度提供的。在所有的方差下，本文算法的结果优于 BlindEye[7]和基线模型。值得注意的是，我们获得了与使用灰度图像训练和测试的模型相似的性能。由于我们在训练和测试时都转换了图像，因此网络的偏差要小得多。在大多数实验中，我们的模型的表现略优于灰色算法，这表明我们的算法可以有效地消除目标偏差，并鼓励网络提取更多信息的特征。

σ^2	**0.02**	**0.025**	**0.03**	**0.035**	**0.04**	**0.045**	**0.5**
Baseline	0.4055	0.4813	0.5996	0.6626	0.7333	0.7973	0.8450
BlindEye	0.6741	0.7123	0.7883	0.8203	0.8638	0.8927	0.9159
Gray	0.8374	0.8751	0.8996	0.9166	0.9325	0.9472	0.9596
Unlearn	0.8185	0.8854	0.9137	0.9306	0.9406	0.9555	0.9618
Ours	0.8393	0.8956	0.9197	0.9324	0.9388	0.9469	0.9512

图 3: 实验结果

6 总结与展望

本文介绍了什么是数据偏置和它对神经网络训练过程会造成的影响，以及几种常见的去除偏置算法，之后提出了一个新的正则化项来训练深度神经网络使其可以使用有偏置的数据进行训练。为了去除偏置，我们最小化偏置与特征提取网络输出特征之间的互信息。通过让网络玩极大极小游戏，让网络学会分类，同时消除偏置。实验结果表明，使用所提出的正则化训练的网络可以提取出与偏置无关的特征嵌入，并且在大多数实验中都取得了较好的性能。此外，我们的模型比用几乎无偏数据训练的“Gray”模型表现得更好，这表明特征嵌入变得更有信息量。总之，我们在本文中证明了所提出的正则化项提高了用有偏置数据训练的神经网络的性能。我们希望这项研究能够扩展各种数据的使用，并为特征解缠领域做出贡献。

参考文献

[1] ATTENBERG J, IPEIROTIS P, PROVOST F. Beat the machine: Challenging humans to find a predictive model's “unknown unknowns”[J]. Journal of Data and Information Quality (JDIQ), 2015, 6(1): 1-17.

[2] TORRALBA A, EFROS A A. Unbiased look at dataset bias[C]//CVPR 2011. 2011: 1521-1528.

[3] TOMMASI T, PATRICIA N, CAPUTO B, et al. A deeper look at dataset bias[J]. Domain adaptation in computer vision applications, 2017: 37-55.

[4] CUBUK E D, ZOPH B, MANE D, et al. Autoaugment: Learning augmentation strategies from data[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 113-123.

[5] GUPTA A, MURALI A, GANDHI D P, et al. Robot learning in homes: Improving generalization and reducing dataset bias[J]. Advances in neural information processing systems, 2018, 31.

[6] WANG H, HE Z, LIPTON Z C, et al. Learning robust representations by projecting superficial statistics out[J]. arXiv preprint arXiv:1903.06256, 2019.

[7] ALVI M, ZISSERMAN A, NELLÅKER C. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings[C]//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018.

[8] BAHNG H, CHUN S, YUN S, et al. Learning de-biased representations with biased representations[C]//International Conference on Machine Learning. 2020: 528-539.

[9] HENDRICKS L A, BURNS K, SAENKO K, et al. Women also snowboard: Overcoming bias in captioning models[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 771-787.

[10] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3156-3164.

[11] CADENE R, DANCETTE C, CORD M, et al. Rubi: Reducing unimodal biases for visual question answering[J]. Advances in neural information processing systems, 2019, 32.

[12] SAGAWA S, KOH P W, HASHIMOTO T B, et al. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization[J]. arXiv preprint arXiv:1911.08731, 2019.

[13] CHEN X, DUAN Y, HOUTHOOFT R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets[J]. Advances in neural information processing systems, 2016, 29.

[14] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[15] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. The journal of machine learning research, 2016, 17(1): 2096-2030.

[16] LECUN Y, CORTES C, BURGES C, et al. MNIST handwritten digit database[Z]. 2010.