中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

基于生成对抗网络的分类模型研究

作者姓名：胡兵兵

指导教师：吴幼龙 助理教授 上海科技大学

学位类别：工学硕士

学科专业：通信与信息系统

培养单位：中国科学院上海微系统与信息技术研究所

2020 年 6 月

Classification Models Based on Generative Adversarial Networks

A thesis submitted to the
University of Chinese Academy of Sciences
in partial fulfillment of the requirement
for the degree of
Master of Engineering
in Communication and Information System

By

Hu Bingbing

Supervisor: Professor Wu Youlong

Shanghai Institute of Microsystem and Information
Technology, Chinese Academy of Sciences

June, 2020

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关保存和使用学位论文的规定，即中国科学院大学有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

随着科技的发展，数据量日益增长，人们希望从这些原始数据中挖掘出感兴趣的信息。从一方面来看，这些数据数量庞大，所以靠人工去标注成本太高；另一方面这些高维数据复杂度高，所以常规的数据预处理方法也会失效。近年来深度学习领域研究活跃，其中生成对抗网络因其新颖的训练方式和超强的可扩展性受到了广泛关注。

本文研究了基于生成对抗网络的分类模型。传统的分类方法需要有标注的数据，而且容易过拟合，面对大量高维数据，这些方法不再适用。生成对抗网络可以无监督地进行训练，而且在模型稳定之后，生成的新数据可以用来扩充数据集。本文提出了两种基于生成对抗网络的分类模型：C-InfoGAN 和 InfoCatGAN。前者将 InfoGAN 扩展为分类模型，利用模型中的辅助网络做分类，能够在生成高质量图片的同时，达到较好的分类准确率；后者在 CatGAN 的基础上增加了互信息约束，使得生成的图片更加逼真。二者均能通过隐变量控制生成图片的类别，这对数据增强具有较大意义。此外，在加入少量标签信息之后，模型的准确率能大幅提高。

关键词：生成对抗网络，分类，无监督学习，半监督学习

Abstract

As the science and technology grow, the amount of data is increasing day by day. People want to mine valuable information from these raw data. On the one hand, the amount of these data is too large, so manual labeling is unrealistic; on the other hand, these high-dimensional data have high complexity, so classic data preprocessing methods will also fail. In recent years, the deep learning community is very active, and generative adversarial network has received extensive attention for its novel training methods and super-scalability.

This thesis studies the classification models based on generative adversarial networks. Traditional classification methods require labeled data and are easy to overfit. Facing a large number of high-dimensional data, these methods are no longer applicable. Generative adversarial networks can be trained in an unsupervised manner, and the generated data from the trained model can be used to expand the data set. This thesis proposes two classification models based on generative adversarial networks: C-InfoGAN and InfoCatGAN. The former extends InfoGAN into a classification model, and uses the auxiliary network in the model for classification, which can achieve high classification accuracy while generating high-quality pictures; the latter adds mutual information constraints to CatGAN, making the generated pictures more realistic. Additionally, both two models can control the category of generated pictures through latent variables, which has great significance for data augmentation. In addition, after adding a small amount of label information, the accuracy of the model can be greatly improved.

Keywords: GAN, Classification, Unsupervised Learning, Semi-supervised Learning

目 录

图形列表

表格列表

符号列表

字符

DUnif(a,b)	区间 [a,b] 上的离散均匀分布
Unif(a,b)	区间 [a,b] 上的连续均匀分布
p	概率分布
Pr	事件概率，概率测度
$\mathbf{x}$	数学加粗表示矢量
Bern(p)	参数为 p 的 Bernoulli 分布
X	大写字母表示随机变量
Supp(X)	随机变量 X 的支撑集，即 X 可能的取值

算子

Symbol	**Description**
∇	gradient operator

缩写

AI	Artificial Intelligence
GAN	Generative Adversarial Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbors

第 1 章　绪论

近年来，人们的生活方式随着人工智能技术的快速发展已经发生了许多变化：从现金结账到刷脸支付，从出租车到无人驾驶，从人工咨询到问询机器人，从营销员推销式购物到无人零售，从自己动手做家务到智能家居。这些都是过去几年科技的发展给人类生活带来的影响。科技的背后往往都是技术的发展，这其中尤为重要的一个就是人工智能技术。其实“人工智能”的概念并不是近几年才出现的，早在 1956 年，就有科学家想用当时才刚刚出现不久的计算机来构建出具有人类智能的机器。人工智能的概念就诞生于此，之后也没有引起多大的反响，只有少数人将它当作研究目标默默在实验室耕耘。再后来的几十年间，人工智能就这样慢慢酝酿，不温不火，它一方面代表了科学家们的美好愿景，一方面被技术人员视为空中楼阁。直到 2012 年深度学习出现，伴随着逐年增长的数据量，以及越来越容易获得的计算资源，人工智能才算步入一个新时代。据领英《全球 AI 领域人才报告》显示，截至 2017 年第一季度，全球人工智能相关专业技术人员的人才总量超过 190 万，中国地区人工智能领域专业人才超过 5 万人，并且还存在 500 多万的人才缺口。可见人工智能领域的发展造成了相关专业人才短缺，供不应求。与此同时，人工智能所涵盖的范围也在不断扩张：如机器学习、专家系统、推荐系统等等。图 1.1给出了人工智能的一些分支。其中机器学习可以看作实现人工智能的一种方法，它通过特定算法在数据集上进行训练，进而学习到数据的特征，最后达到对未知样本做出决策或预测的效果。

机器学习的概念起源于早期的人工智能，研究至今已有不少经典的算法：如聚类(Clustering)、决策树(Decision Tree)、朴素贝叶斯(Naive Bayes)、Expectation Maximization (EM)、支持向量机（Support Vector Machine, SVM)、AdaBoost 等。此外，根据算法的学习方法不同又可分为无监督学习（如聚类)、监督学习(如分类)、半监督学习、集成学习、深度学习和强化学习。其中深度学习是最近几年来的热门研究领域。从上面的划分来看，深度学习其实不是一种独立的学习方法，而是属于机器学习的一个分支。只是近年来深度学习领域的创新方法层出不穷，使得越来越多的学者都将其看作为一种独立的学习方法。总而言之，越来越多的学者都踊跃加入深度学习研究的行列，为深度学习活跃发展及应用

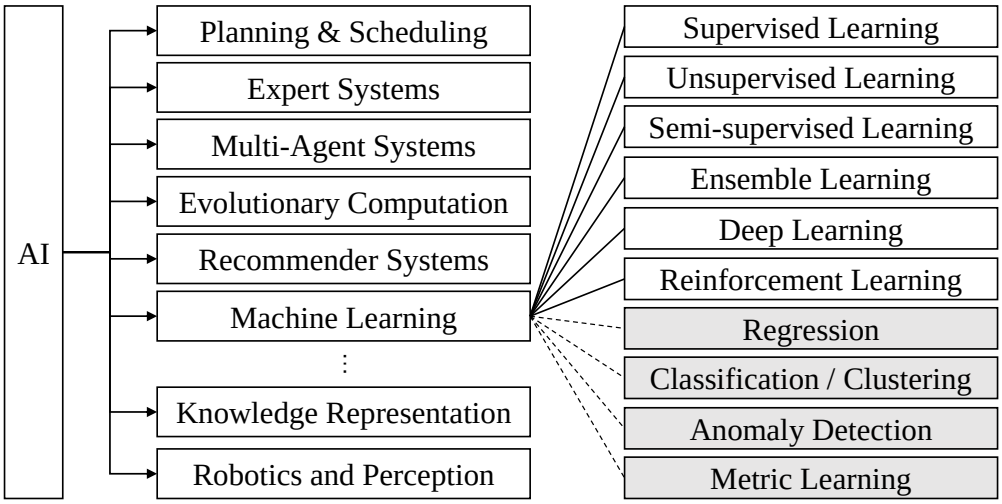


图 1.1 人工智能研究分支

Figure 1.1 Research branches of AI

做出极大的贡献。

1.1　研究背景

近些年，机器学习在人们生活的方方面面发挥着极大的作用：语音识别、交通预测、视频监控、垃圾邮件过滤、智能客服、商品推荐等等。然而，机器学习算法需要从大量数据集中提取有用的特征，进而做出决策。但是目前并没有一个通用的特征提取方法适用于不同场景。因此学者们提出了一个新的方法，称为表征学习（Representation Learning），它能够在分类和检测任务中自动提取有用的特征[1]。深度学习[2] 也属于一种表征学习方法，它可以提取出高维度、更抽象的数据表征。然而，这些方法都属于监督学习，需要数据集具有多样化的特征，并且每个数据样本都需要被标注。通常获取这些标注好的数据集或者去标注数据集的难度很大。所以，越来越多的学者将重心放在无监督学习上。在无监督学习中，生成式模型（见2.2节）是相对有效的方法，早期的生成式模型（如 Restricted Boltzmann Machines (RBM)[3], Deep Belief Networks (DBN)[4], Deep Boltzmann Machines[5] ）通常基于马尔可夫链、极大似然和近似推断等工具。然而这些方法的计算复杂度都很高，并且无法保证良好的泛化性能。近年来，生成对抗网络（Generative Adversarial Networks, GAN)[6] 作为新兴的生成式模型，因其新颖的训练方式和超强的可扩展性而收到广泛关注。该模型包含两个网络：生成器和判别器。在对抗训练的过程中，生成器会学习到数据的概率分布，同时生成逼真的数据去迷惑判别器，而判别器的目标则是从虚假数据中区分出真实数据。两个网络之间通过相互对抗，最终达到一个平衡点。值得一提的是，GAN 不需要使用复杂的近似推断或马尔可夫链，并且在计算机视觉，自然语言处理及其他众多领域都有较好的生成效果。

GAN 在图像超分辨率方面有很好的效果。SRGAN[7] 是第一个使用生成对抗网络实现图像超分辨率的模型，ESRGAN[8] 则通过对 SRGAN 的三个主要组成部分进行改进，利用 Jolicoeur-Martineau[9] 中提出的方法令判别器输入相对的真实距离而非绝对距离。基于 CycleGAN[10]，Yuan 等[11] 提出 Cycle-in-Cycle GAN 用于无监督的图像超分辨率。Ding 等[12] 提出 TGAN，通过探索张量结构来生成大型高质量图像。

GAN 在图像生成方面也有很多应用。Tran 等[13] 提出的 DR-GAN 可用于姿态鲁棒的人脸识别。Huang 等[14] 提出 TP-GAN 模型，通过同时感知整体结构和局部细节，该模型可以生成逼真的正面视图。Ma 等[15] 提出了一种使用姿

势指导生成人体图像的生成网络（PG^2），该网络基于一个新的姿势和该人体的图像来合成任意姿势的该人体图像。虽然大多数论文使用 GAN 来合成二维图像[16,17]，但 Wu 等[18] 将三维卷积应用于 GAN，生成了许多三维样本包括沙发、椅子、汽车和桌子。Im 等[19] 使用循环对抗网络生成图像。Yang 等[20] 提出了分层循环生成对抗网络（LR-GAN），该模型也可以生成图像。

GAN 也可以用于目标检测。考虑学习一个对变形和遮挡鲁棒的物体检测器，一种方法是使用具有变形和遮挡的数据集来训练。但是一些变形和遮挡非常罕见，它们在现实中可能不会发生。此时可以用 GAN 来生成这样的数据以参与训练。Wang 等[21] 使用 GAN 生成具有变形和遮挡的实例，对手的目的是生成难以被对象检测器分类的样本。通过将 segmentor 和 GAN 相结合，Segan[22] 能够检测到图像中被其他物体遮挡的物体。为了解决小物体检测问题，Li 等[23]] 提出感知生成对抗网络模型；Bai 等[24] 提出了一种端到端的多任务 GAN（MTGAN）。

1.2 研究内容与意义

本文研究基于生成对抗网络的分类模型。传统的分类方法如逻辑回归、支持向量机、朴素贝叶斯、K 近邻都属于监督学习，需要数据标签，且只能应对数据量较小的情况。随着科技的飞速发展，各行各业每时每刻都在产生海量数据，对于大体量、高维度、高差异的数据，传统分类方法不免有些捉襟见肘。因此，伴随着深度学习领域的崛起，许多学者开始使用深度神经网络去做分类研究[25,26]。在复杂数据面前，神经网络的超强拟合能力使其成为了挖掘数据价值的不二之选。以上这些方法的都包含以下三个过程：数据压缩，特征提取和模型预测。这些过程往往依赖于大量的数据标注，但现实生活中标注好的数据十分稀缺：有些涉及隐私，有些获取代价高昂，有些甚至无法获得。因此人们开始研究无监督分类。

无监督分类通常建模为聚类问题，并且已经具有一些经典的方法：K-means、Gaussian mixture model、Density estimation，这些方法均是针对数据分布进行建模。此外，一些判别式方法比如 Maximum margin clustering (MMC)[27]、Regularized information maximization (RIM)[28]，则是将数据划分到某个类别，而无须估计数据分布。尽管判别式方法更为直接，但是它们容易受一些虚假相关性的影响而产生过拟合[29]。当与深度神经网络这种拟合能力很强的模型相结合

的时候，过拟合现象尤为显著。随着深度学习的发展，基于深度模型的无监督或半监督方法也渐渐出现。它们通常训练一个生成式模型（如波尔兹曼机[5,30]、前馈神经网络[31,32] 以及自编码器[33,34]），通过重建输入样本来学习数据特征，刻画数据分布。因此，它们避免了因直接划分数据而产生的过拟合问题。但是这些方法在重建训练样本的过程中，并没有额外的约束，所以会保留原始数据的所有信息，这和训练分类器的目标相背[1]。

生成对抗网络结合了生成式和判别式模型，一定程度上克服了上述问题。通过将判别器扩展为分类器，对抗的训练机制可以让判别器无监督或半监督地学习到数据的类别特征，从而达到分类的效果。与此同时，生成器生成的数据一方面可以提高判别器的鲁棒性，一方面可以用于数据增强，进一步提升判别性能。

1.3　研究现状

CatGAN[29] 将判别器从二分类输出扩展为多分类输出，同时设计了新的对抗目标，在 MNIST[35] 和 CIFAR-10[36] 上均取得了十分可观的分类准确率。Salimans 等[37] 提出了几种训练 GAN 的技巧，以及给出了一种半监督分类方法。不同于 CatGAN，他们令判别器输出 $K+1$ 个类别，其中 K 为数据集原始的类别个数，而多出来的类别就代表着生成器生成的虚假数据，也就是把生成器的所有输出都归为一类。这显然忽略了生成器生成的不同类别的数据。SGAN[38] 提出了类似的网络模型。Dai 等[39] 指出给定判别器的目标，良好的半监督性能需要一个较差的生成器，即对一个判别器来说，分类和区分真假是不兼容的目标。TripleGAN[40] 也指出了这一点，同时提出了改进方法：将分类任务和判别任务分配给两个不同的模块。具体来说，TripleGAN 增加了一个分类网络，令判别器专注于区分真假，同时对训练目标做了相应改动，使得生成器和分类器对应的概率分布同时收敛到真实数据分布。Wu 等[41] 在 TripleGAN 的基础上应用了 feature matching[37]，并且引入了两个分类网络协同工作，达到了更高的性能。

[1]在训练分类器时，通常只希望保留和分类目标相关的信息，从而使得模型对其他不重要的信息更加鲁棒。

1.4 本文工作

本文主要研究两个基于生成对抗网络的分类模型，基于这两个模型提出一些算法改进。第一个模型是 InfoGAN[42]，严格来说它并不属于基于 GAN 的分类模型，但是它包含一个辅助网络，契合了分类器的特征，因此我们可以对辅助网络添加用于分类的正则项，将辅助网络训练成一个分类器。第二个模型是 CatGAN[29]，它是完全以分类为目标而设计的，因此其中有很多思想值得我们参考。例如，对于目标函数的改变，CatGAN 使用判别器输出的条件分布的熵作为判别真假的依据。它创新性地引入了信息论尺度作为训练目标，在具备实践效果的同时具备一定的可解释性。

对于 InfoGAN，我们几乎没有改变模型结构，通过在辅助网络的目标函数添加了一个分类损失，就达到了比较可观的分类准确率。称该模型为 Classifier InfoGAN (C-InfoGAN)，它保留了 InfoGAN 的大部分优点：比如通过离散隐变量控制生成图片的类别，以及通过连续隐变量调整生成图片的局部细节，同时还具有可观的分类性能。此外，本文还提出了半监督版本的训练方法。在拥有少量标签的情况下，可将标签直接放入辅助网络中，从而充分利用标签信息，将真实标签和隐变量相互绑定，解决了 InfoGAN 中离散隐变量和真实标签不对应的问题。

对于 CatGAN，本文受 InfoGAN 启发，将生成器的输入噪声分解为无意义噪声：提供模型的容量，使得模型具有足够的自由度去学习数据的细节（高度耦合的特征）；和有意义隐变量：用于在学习过程中绑定到数据类别；同时优化隐变量和生成图片之间的互信息，提出了 InfoCatGAN 模型。该模型可以在牺牲少量分类准确率的情况下，大大提高生成器生成图片的质量。这是由于 CatGAN 使用条件熵作为目标函数，没有类别指向性，因为无法从条件熵的高低推知条件分布的形态。InfoCatGAN 模型通过在隐空间构造一维隐变量，在训练过程中将生成数据的类别标签与之绑定，使得可以通过该隐变量来控制生成数据的类别。改进的主要思想就是在生成虚假数据的时候，令生成器生成具体类别的数据，而不是笼统地最小化条件熵。此外，提出的模型能够很好的兼容半监督学习，并且在少量标签信息的帮助下，隐变量和真实类别一一对应，也就是说，可以通过隐变量控制生成图片的类别。

本文提出了两个改进模型都使用了基于互信息的正则项，这从侧面说明了

信息论在未来的深度学习领域拥有值得探索的价值。

1.5　文章结构

本文整体章节结构安排如下:

- 第一章为绪论。主要介绍本文的研究背景即当前深度学习领域的蓬勃发展, 接着介绍了本文的研究内容与意义, 然后介绍了生成对抗网络的发展及众多应用, 引入使用生成对抗网络模型做分类的动机和好处, 介绍了国内外关于该问题的研究现状。最后介绍了本文的主要工作。
- 第二章为预备知识。首先介绍了什么是生成式模型，什么是判别式模型，接着介绍了生成对抗网络的核心思想和基本原理。最后详细阐述了 InfoGAN 和 CatGAN 的模型结构及各自的原理和目标函数。
- 第三章为本文工作。分别详细介绍了本文提出的两个模型 C-InfoGAN 和 InfoCatGAN，阐明了它们的原理以及目标函数，同时给出了训练方法。
- 第四章为模型评估。主要介绍了评估模型的方法和尺度，以及实验细节。接着给出了两个模型在 MNIST 和 FashionMNIST 上的结果，以及对结果做了简单分析，同时给出了模型的收敛速度。
- 第五章为总结与展望。总结全文内容，并对未来的研究方向进行展望。

第 2 章　预备知识

2.1　引言

在开始介绍本文工作之前，首先需要了解一些背景知识。本章首先介绍什么是生成式模型和判别式模型，然后给出生成对抗网络的详细介绍及其数学模型，最后介绍 InfoGAN 和 CatGAN 模型。

2.2　生成式模型和判别式模型

在介绍生成对抗网络之前，我们有必要了解两个概念：生成式模型（Generative Modeling, GM）和判别式模型（Discriminative Modeling，DM）。我们先用简单地一句话说明，接着再详细阐述它们区别。简而言之，生成式模型是针对数据的联合分布进行建模，而判别式模型则是针对数据的条件分布进行建模。

在一个基本的机器学习问题中，通常有输入 $x \in \mathcal{X}$ 和输出 $y \in \mathcal{Y}$ 两个部分。通俗的说，DM 关注 x 和 y 的内在联系，即在给定 x 的条件下，y 的分布应该满足什么样的性质；而 GM 更关注于 (x, y) 的联合分布。模型训练完成之后，判别式模型将训练数据集的信息提取，模型本身拟合了训练样本中的经验条件分布 $P(y|x)$，因此对于测试样本 x^*，判别式模型直接输出对应的条件概率 $P(y^*|x^*)$，给出分类结果。生成式模型则对训练集的联合分布 $P(x, y)$ 进行建模，对于未见样本 x^*，选择使得 $P(x^*, y)$ 最大的 y 作为预测值。由贝叶斯公式知

$$\begin{aligned} y^* &= \underset{y}{\operatorname{argmax}} P(y|x^*) \\ &= \underset{y}{\operatorname{argmax}} \frac{P(x^*|y)P(y)}{P(x^*)} \\ &= \underset{y}{\operatorname{argmax}} P(x^*|y)P(y) \\ &= \underset{y}{\operatorname{argmax}} P(x^*, y). \end{aligned}$$

因此，这等价于寻找使得 $P(y|x^*)$ 最大的 y 作为预测值。

常见的生成式模型有：朴素贝叶斯、隐马尔可夫模型（Hidden Markov Model, HMM）、高斯混合模型（Gaussian Mixture Model，GMM）、受限玻尔兹曼机（Restricted Boltzmann Machine，RBM）；常见的判别式模型有：KNN（K-Nearest

Neighbours)、感知机 (Perceptrons)、决策树 (Decision Tree)、逻辑回归 (Logistic Regression)、最大熵模型 (Maximum Entropy Model)、支撑向量机 (Support Vector Machine, SVM)、条件随机场 (Conditional Random Field)、神经网络 (Neural Networks) 等。在实际使用中，两者各有优劣。生成式模型可以建模联合分布，但联合分布本身难以估计，所以会需要较大地数据量和计算量。判别式模型具有更低地渐近误差，但其收敛到渐近误差的速度比生成式模型更慢；而生成时模型的渐近误差往往高于判别式模型[43]，而且可以处理隐变量和半监督甚至无监督学习。

2.3 生成对抗网络

2.3.1 基本思想

Goodfellow 等[6] 在 2014 年提出生成对抗网络 (Generative Adversarial Networks, GAN)，该模型使用博弈理论中的 minimax 博弈机制来训练两个模块：生成器和判别器。其目标是通过对抗训练使得生成器对应的分布 p_g 和真实数据的分布 p_{data} 尽量接近。生成式模型不直接估计每个数据样本的概率，而是利用生成网络 G 将噪声变量 $\boldsymbol{z} \sim p_z(\boldsymbol{z})$ 映射到虚假样本 $G(\boldsymbol{z})$。在训练过程中，判别器的目标是将虚假样本和真实样本区分开，生成器的目标则是生成让判别器无法区分的虚假样本。训练过程可以类比为两个玩家博弈：判别器读取一个数据希望能够分别真假，而生成器希望生成以假乱真的数据从而让判别器判定为真。整个过程大致如下：生成器将一个随机噪声映射到数据空间，形成一个虚假数据。判别器接受真实样本或虚假样本作为输入，并输出当前样本来自真实数据的概率。训练的目标是一方面让判别器能够区分真假，一方面使生成器能够以假乱真，这就形成了两方对抗，最终二者的能力越来越强，无法进一步提高自己的性能，博弈达到平衡点。

2.3.2 数学模型

设 $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ 为含有 N 个样本的数据集，$\boldsymbol{x}_i = (x_{i1}, \cdots, x_{in}) \in \mathbb{R}^n$ 为单个数据样本，标准生成对抗网络模型可被正式定义如下：生成器 $G : \mathbb{R}^m \mapsto \mathbb{R}^n$，判别器 $D : \mathbb{R}^n \mapsto [0,\ 1]$。其中 m 为隐空间维度，即噪声维度，n 为数据维度。给定输入噪声 $\boldsymbol{z} \sim p_z, \boldsymbol{z} = (z_1, \cdots, z_m) \in \mathbb{R}^m$，$G(\boldsymbol{z}) \in \mathbb{R}^n$ 为生成器生成的虚假样本；给定真实数据样本 $\boldsymbol{x}$ 或虚假样本 $G(\boldsymbol{z})$，$D(\cdot) \in [0,\ 1]$ 表示当前输入来自真

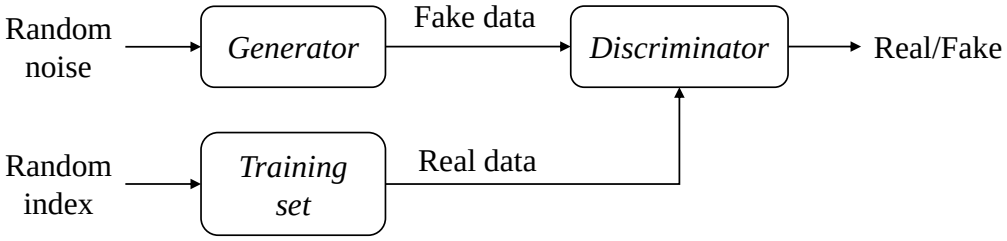


图 2.1 朴素 GAN 结构示意

Figure 2.1 The architecture of vanilla GAN

实数据的概率。接着训练 D 使得对于真实数据 $\boldsymbol{x}$，$D(\boldsymbol{x})$ 接近于 1；对于虚假数据 $\tilde{\boldsymbol{x}} = G(\boldsymbol{z})$，$D(\tilde{\boldsymbol{x}})$ 接近于 0。同时训练 G 使得 $\log(1 - D(G(\boldsymbol{z}))$ 达到最小。换言之，D 和 G 类似两个玩家博弈，其价值函数如下：

$$\min_G \max_D \; V(D,G) = \mathrm{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}}[\log D(\boldsymbol{x})] + \mathrm{E}_{\boldsymbol{z}\sim p_z}[\log(1 - D(G(\boldsymbol{z})))]. \tag{2.1}$$

在实际应用中，GAN 通常被实现为可微神经网络：生成器网络 $G(\cdot;\boldsymbol{\theta}_g)$ 和判别器网络 $D(\cdot;\boldsymbol{\theta}_d)$，其中 $\boldsymbol{\theta}_g$ 和 $\boldsymbol{\theta}_d$ 分别为生成器和判别器的网络参数[1]。此外，生成器和判别器在实现中也需要采用交替迭代的机制，使用基于梯度的优化方法来训练。在训练初期，G 生成的加虚假样本比较容易被 D 识别，从(2.1)式可以看出，$\log(1 - D(G(\boldsymbol{z})))$ 的梯度变化很小，生成器难以得到优化。因此，在实现中通常训练 G 使得 $\log D(G(\boldsymbol{z}))$ 达到最大。

2.4　InfoGAN

2.4.1　基本思想

朴素 GAN 通常使用高维高斯噪声作为生成器的输入，并没有规定生成器如何利用这个噪声，也没有对隐空间作任何限制。因此，噪声的利用可以是任意的，生成器可能生成具有高度耦合特征的数据。这从图像上来看，也就是模型生成的虚假图片高度抽象，没有明显轮廓。从底层来看，$\boldsymbol{z}$ 的每一维度并没有对应到生成数据的某个特征。然而，一个好的模型很自然地可以将数据的各个特征分离出来，以提供给用户调整。比如，当要求 GAN 从 MNIST[35] 生成图片时，理想的模型可以利用一个离散随机变量来表示数字类别特征（0-9），利用两个连续随机变量来分别表示数字的角度和笔画的粗细。

[1]为了简便起见，在不产生歧义的情况下通常省略网络参数。

Mirza 和 Osindero[44], Odena 等[45], Miyato 和 Koyama[46] 均指出通过调整隐空间结构，可以控制生成器生成具有某个特征的图片。InfoGAN[42] 将隐变量分解为两部分，一部分仍然是无结构的噪声 $\boldsymbol{z} \sim p_z$，为模型提供足够的容量；另一部分则作为隐变量 $\boldsymbol{c} \sim p_c$，用于学习数据特征。

2.4.2 互信息约束

设 $\boldsymbol{c} = (c_1, c_2, \cdots, c_L)$ 表示输入空间中的 L 个隐变量，最简单的情况是各隐变量之间互相独立，即 $p_{\boldsymbol{c}}(c_1, c_2, \cdots, c_L) = \prod_{i=1}^{L} p_{c_i}(c_i)$。InfoGAN 基于朴素 GAN 模型提出了一种在无监督条件下，利用隐变量学习到数据特征的方法。该方法将隐空间噪声分解为两部分，包含一个生成器 $G(\boldsymbol{z}, \boldsymbol{c})$ 和一个判别器 D。由于在朴素 GAN 中，生成器可以忽略隐空间结构，此时 $p_g(x|c) = p_g(x)$，其中 p_g 为生成器对应的概率分布。为了解决这个问题，InfoGAN 为模型增加了互信息正则项，并指出隐变量 $\boldsymbol{c}$ 和生成数据 $G(\boldsymbol{z}, \boldsymbol{c})$ 之间的互信息应该很高。

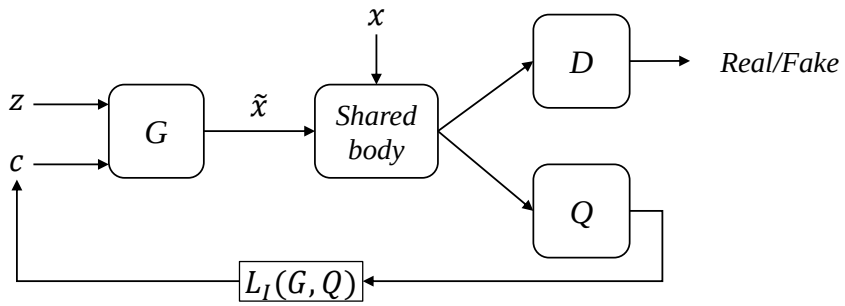


图 2.2 InfoGAN 结构示意。其中 x 表示真实样本，$\tilde{x} = G(z, c)$ 表示由生成器生成的虚假样本。在实现中，通常将辅助网络 Q 和判别器 D 共享一部分网络结构。

Figure 2.2 The architecture of InfoGAN. x denotes the real data and $\tilde{x} = G(z, c)$ denotes fake data. In practice, Q and D share a body.

在信息论中，熵、相对熵以及互信息的定义如下[47]：

定义 2.1. 熵 设有离散随机变量 $X \sim p_X(x)$ 及其支撑集 $\mathcal{X}$，则它的熵 $H(X)$ 定义为

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x). \tag{2.2}$$

定义 2.2. KL 距离 设有定义在同一个支撑集 $\mathcal{X}$ 上的两个概率分布 $p(x)$ 和 $q(x)$，

则它们的相对熵（或称 Kullback-Leibler 距离）定义为

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}. \tag{2.3}$$

定义 2.3. 互信息 设有两个离散随机变量 X, Y 及其支撑集 $\mathcal{X}$, $\mathcal{Y}$，它们的联合分布为 $p_{X,Y}(x,y)$，边缘分布分别为 $p_X(x)$ 和 $p_Y(y)$，则 X 和 Y 之间的互信息 $I(X;\ Y)$ 定义为它们联合分布和边缘分布乘积之间的相对熵：

$$\begin{aligned} I(X;\ Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x) \cdot p_Y(y)} \\ &= D_{\mathrm{KL}}\left(p_{X,Y}(x,y) \parallel p_X(x) p_Y(y)\right) \\ &= \mathrm{E}_{p_{X,Y}}\left[\log \frac{p_{X,Y}}{p_X \cdot p_Y}\right]. \end{aligned} \tag{2.4}$$

在信息论中，随机变量的熵是其不确定性的度量。由以上定义，容易得到

$$\begin{aligned} I(X;\ Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X). \end{aligned} \tag{2.5}$$

可知随机变量 X, Y 的互信息度量着观测到 Y 之后，X 的不确定性的减少量。从(2.4)式可以看出，互信息越大表明两个随机变量之间的相关性越大，反之互信息为零则说明变量间相互独立。于是，最大化隐变量 $\boldsymbol{c}$ 和生成器输出 $\tilde{\boldsymbol{x}} = G(\boldsymbol{z}, \boldsymbol{c})$ 的互信息 $I(\boldsymbol{c};\ \tilde{\boldsymbol{x}}) = H(\boldsymbol{c}) - H(\boldsymbol{c}|\tilde{\boldsymbol{x}})$ 等价于最小化 $H(\boldsymbol{c}|\tilde{\boldsymbol{x}})$，这里由于 $\boldsymbol{c}$ 的分布在整个过程中是确定的，所以 $H(\boldsymbol{c})$ 可以视为常数。换句话说，隐变量 $\boldsymbol{c}$ 的信息量在生成过程中应该流向 $\tilde{\boldsymbol{x}}$，即给定 $\tilde{\boldsymbol{x}}$ 之后，$\boldsymbol{c}$ 的信息量应该很小。类似的互信息约束也可以应用到聚类中[28,48,49]。InfoGAN 在朴素 GAN 的基础上增加了互信息约束，其价值函数 $V_I(D,G)$ 如下：

$$\min_G \max_D \ V_I(D,G) = V(D,G) - \lambda I(\boldsymbol{c};\ G(\boldsymbol{z},\boldsymbol{c})). \tag{2.6}$$

2.4.3 互信息的变分下界及其估计

在实际应用中，由于依赖后验信息 $p_{\boldsymbol{c}|\tilde{\boldsymbol{x}}}(c|x)$，互信息 $I(\boldsymbol{c};\ G(\boldsymbol{z},\boldsymbol{c}))$ 难以直接计算。Poole 等[50] 提出互信息的变分下界

$$\begin{aligned} I(\boldsymbol{c};\tilde{\boldsymbol{x}}) &= \mathrm{E}_{p_{\boldsymbol{c},\tilde{\boldsymbol{x}}}}\left[\log\frac{p_{\boldsymbol{c},\tilde{\boldsymbol{x}}}(c,x)}{p_{\boldsymbol{c}}(c)p_g(x)}\right] = \mathrm{E}_{p_{\boldsymbol{c},\tilde{\boldsymbol{x}}}}\left[\log\frac{p_{\boldsymbol{c}|\tilde{\boldsymbol{x}}}(c|x)}{p_{\boldsymbol{c}}(c)}\right] \\ &= \mathrm{E}_{p_{\boldsymbol{c},\tilde{\boldsymbol{x}}}}\left[\log\frac{Q(c|x)}{p_{\boldsymbol{c}}(c)}\right] + \mathrm{E}_{p_g(x)}\left[D_{\mathrm{KL}}\left(p_{\boldsymbol{c}|\tilde{\boldsymbol{x}}}(c|x)\ ||\ Q(c|x)\right)\right] \\ &\geq \mathrm{E}_{p_{\boldsymbol{c},\tilde{\boldsymbol{x}}}}\left[\log Q(c|x)\right] + H(\boldsymbol{c}) \triangleq L_I(Q(c|x)), \end{aligned} \tag{2.7}$$

其中 $p_{\boldsymbol{c}}$ 为隐变量分布，p_g 为生成器对应的概率分布，$Q(c|x)$ 为 InfoGAN 模型的辅助网络用以估计真实后验概率 $p_{\boldsymbol{c}|\tilde{\boldsymbol{x}}}(c|x)$，$L_I(Q(c|x))$ 即为互信息的变分下界。在训练过程中，隐变量的分布是固定的，所以 $H(\boldsymbol{c})$ 可视为常量。事实上，下界 L_I 与生成器和辅助网络都有联系：

$$L_I(G,Q) = \mathrm{E}_{p_{\boldsymbol{c},\tilde{\boldsymbol{x}}}}\left[\log Q(c|G(\boldsymbol{z},\boldsymbol{c}))\right] + H(\boldsymbol{c}).$$

从(2.7)式可以看出，随着辅助分布 Q 接近真实后验分布，即

$$\mathrm{E}_{p_g(x)}[D_{\mathrm{KL}}(p_{\boldsymbol{c},\tilde{\boldsymbol{x}}}(\cdot|x)\ ||\ Q(\cdot|x))] \to 0,$$

变分下界越来越紧。综上所述，InfoGAN 的目标函数如下：

$$\min_{G,Q}\max_{D}\ V_{\mathrm{InfoGAN}}(D,G,Q) = V(D,G) - \lambda L_I(G,Q), \tag{2.8}$$

其中 λ 为正则系数。

2.5 CatGAN

2.2节提到生成式模型适合用于半监督甚至无监督学习，CatGAN[29] 就是研究无监督和半监督分类的模型。

2.5.1 基本思想

朴素 GAN 模型的判别器的输出是一个概率，代表着当前输入来自真实数据的概率。如前所述，GAN 的训练过程可以表述如下。设有真实数据集 $\mathcal{X} = \{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$，$\boldsymbol{x}_i \in \mathbb{R}^n$，$N$ 为数据样本个数。生成器 G 将噪声 $\boldsymbol{z} \in \mathbb{R}^m$ 映射到数

据空间，生成虚假样本 $\tilde{\boldsymbol{x}} = G(\boldsymbol{z})$。判别器 D 输出样本 x 来自真实数据集 $\mathcal{X}$ 的概率：$\Pr(y=1|x) = \frac{1}{1+e^{-D(x)}}$。(2.1)式可以写为如下等价形式：

$$\min_G \max_D \ \mathrm{E}_{\boldsymbol{x}\sim p_{\text{data}}} \left[\log \Pr(y=1|\boldsymbol{x})\right] + \mathrm{E}_{\boldsymbol{z}\sim p_z} \left[\log\left(1 - \Pr(y=1|G(\boldsymbol{z}))\right)\right], \quad (2.9)$$

其中 p_{data} 和 p_z 分别表示数据分布和噪声分布。朴素 GAN 没有对噪声分布作任何限制，CatGAN 令 $p_z = \text{Unif}(0,1)$，即连续均匀分布。

基于以上陈述，CatGAN 提出了一种新的生成对抗网络结构来作无监督和半监督学习。首先考虑无监督的设定，无监督条件下的结构是针对朴素 GAN 的推广。

2.5.2　问题建模

如前所述，设 $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ 为 N 个无标注数据样本。考虑以无监督的方式，从数据 $\mathcal{X}$ 学习一个判别器 D，使得 D 将给定数据划分到事先确定好的 K 个类别之一。进一步地，我们要求 $D(\boldsymbol{x})$ 给出 $\boldsymbol{x}$ 属于每个类别的概率：$\sum_{k=1}^{K} \Pr(y=k|\boldsymbol{x}) = 1$。CatGAN 的目标是训练这样一个概率分类器，使其在分配概率时满足某种要求。注意到真实的标签分布是未知的，无法直接最大化对数似然函数，所以 CatGAN 了设计一个鉴定分类器性能的度量方式。具体说来，如果对于给定样本 $\boldsymbol{x}$，D 输出的条件分布 $p(y|\boldsymbol{x})$ 具有很高的确定性，而边缘分布 $p(y)$ 对于所有 $k \in \{1,2,\cdots,K\} \triangleq [K]$ 都接近某个先验分布 p_Y。在整个推导过程中，CatGAN 假设标签的先验分布 p_Y 为类别均匀分布，也就是说 $\mathcal{X}$ 中每个类别的样本数量大致相等：$\forall k_1,\ k_2 \in [K],\ \Pr(y=k_1) = \Pr(y=k_2)$。

上述问题的建模可以自然联想到软聚类问题，所以理论上来说可以用概率聚类算法如：Regularized information maximization (RIM)[28]，或者最小化相对熵[51]。然而这些方法都容易对数据集中的一些虚假相关性产生过拟合。第二个直观感觉是朴素 GAN 的结构无法直接用来解决这个问题，因为朴素 GAN 的判别器只能输出一个概率以判断真假，而不能判断输入到底来自哪个类别。原则上，我们希望一个分类器可以建模数据分布，同时学习到数据特征加以利用进行下一步操作，比如判别式聚类。然而，判别器并不一定只能区分真假（二分类），它也应该能够将输入分配到多个类别。

尽管可能存在一些问题，但是一个非常简单的做法是将朴素 GAN 结构扩展一下使得判别器可被用于多分类任务。一旦 D 的结构改变，那么朴素 GAN 模

型的对抗机制则需要重新设计。CatGAN 将博弈机制改变为：要求判别器将每个真实数据样本划分到 K 个类别中的一个，而对于虚假样本，保持一个较高的不确定性。这样可以让分类器更加鲁棒。类似地，要求生成器生成具体到某个类别的图片而不是仅仅生成和真实数据类似的图片。

2.5.3 目标函数

如前所述，CatGAN 所定义的优化问题与朴素 GAN 的不同之处主要在于：CatGAN 想要学习一个判别器，它能够为每个真实数据样本 $\boldsymbol{x}$ 附加一个标签 $y \in [K]$ 而非学习一个二分类的判别器。定义判别器 $D(\boldsymbol{x})$ 为可微函数，它输出一个 K 维对数概率向量（logits）：$D(\boldsymbol{x}) \in \mathbb{R}^K$。样本 $\boldsymbol{x}$ 属于 K 个互斥类别中的某一个的概率可以通过对判别器输出施加 softmax 变换获得：

$$\Pr(y = k|\boldsymbol{x}) = \frac{e^{D_k(\boldsymbol{x})}}{\sum_{k=1}^{K} e^{D_k(\boldsymbol{x})}}, \tag{2.10}$$

其中 $D_k(\boldsymbol{x})$ 表示 $D(\boldsymbol{x})$ 的第 k 个分量。和朴素 GAN 一样，生成器 $G(\boldsymbol{z})$ 定义为将噪声 $\boldsymbol{z} \in \mathbb{R}^m$ 映射到数据空间 $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ 的函数：

$$\tilde{\boldsymbol{x}} = G(\boldsymbol{z}), \quad \boldsymbol{z} \sim p_z, \tag{2.11}$$

其中 p_z 表示噪声分布。

如2.5.2节所说，如何度量分类器性能呢？CatGAN 本意其实是让生成器作为判别器的正则项，使得判别器对于虚假样本更加鲁棒。基于这个想法，CatGAN 提出了 3 条判别器需要满足的要求和 2 条生成器需要满足的要求：

判别器 (i) 对真实数据样本具有很高的确定性，(ii) 对于虚假样本具有很高的不确定性，(iii) 均匀使用所有类别[2]。

生成器 (i) 生成虚假样本使得判别器对其具有较高的确定性，(ii) 生成的样本均匀的分布在 K 个类别中。

CatGAN 将这些要求都转化为可优化的条件概率，以此设计出目标函数。注意，在 K 个类别信息未知的情况下，我们无法直接优化条件概率 $\Pr(y = k|\boldsymbol{x})$ 以满足判别器的第一个要求。因此，只能通过信息论中的一些尺度直接对分布信息进行刻画。最常用的信息论尺度就是熵。直观来看，对于真实样 $\boldsymbol{x}$，我们想让条件分布 $p(y|\boldsymbol{x})$ 呈现单峰趋势（即 D 很确定将当前样本划分到哪个类别），

[2]因为 CatGAN 假设标签的先验分布是类别均匀的。

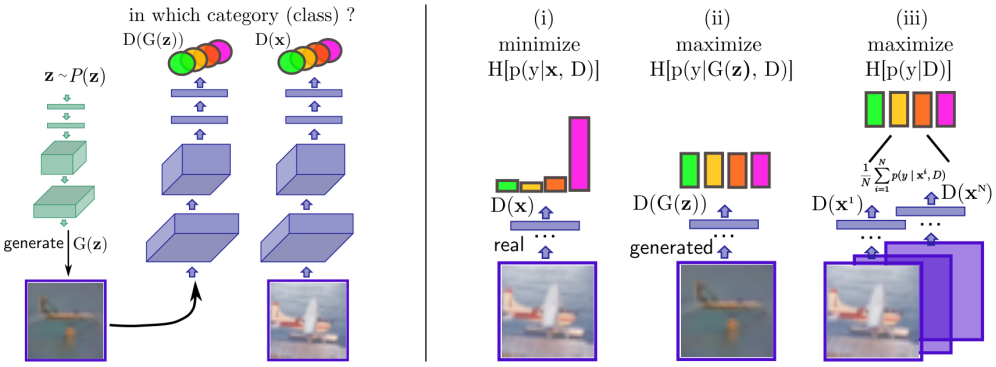


图 2.3 左图为 CatGAN 结构示意，绿色为生成器，紫色为判别器。右图为 CatGAN 判别器目标 L_D^{cat} 的直观解释。给定真实样本，判别器的输出呈单峰分布；给定虚假样本，判别器的输出呈均匀分布；对于所有真实图片，标签 y 的边缘分布呈均匀分布。

Figure 2.3 The architecture of CatGAN is shown on the left, where generator is green and discriminator is violet. An intuitive interpretation of the objective function L_D^{cat} for the discriminator is shown on right. For real samples, miminizing $H(p(y|\mathbf{x}))$ leads to single peak distribution; for fake samples, maximizing $H(p(y|G(\mathbf{z})))$ leads to uniform distribution. Finally, maximizing the marginal class entropy over all data-points leads to uniform usage of all classes.

可以通过最小化 $H(p(y|\boldsymbol{x}))$ 来满足判别器要求 (i)。另一方面，对于虚假样本 $\tilde{\boldsymbol{x}}$，我们想让条件分布 $p(y|\tilde{\boldsymbol{x}})$ 呈现扁平化趋势（即 D 不确定将它划分到哪个类别），此时可以通过最大化 $H(p(y|\tilde{\boldsymbol{x}}))$ 来满足判别器的要求 (ii)，当 $H(p(y|\tilde{\boldsymbol{x}}))$ 达到最大时，$p(y|\tilde{\boldsymbol{x}})$ 在所有类别中均匀分布。因此，对于真实样本 CatGAN 定义其条件熵的估计

$$\begin{aligned}\mathrm{E}_{\boldsymbol{x}\sim p_{\text{data}}}[H(p(y|\boldsymbol{x})] &= \frac{1}{N}\sum_{i=1}^{N} H(p(y|\boldsymbol{x}_i)) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\Pr(y=k|\boldsymbol{x}_i)\log\Pr(y=k|\boldsymbol{x}_i).\end{aligned} \tag{2.12}$$

对于虚假样本，条件熵的估计可以采用 Monte-Carlo 采样，使用 $H(p(y|G(\boldsymbol{z})))$ 对噪声分布 p_z 取期望

$$\mathrm{E}_{\boldsymbol{z}\sim p_z}[H(p(y|G(\boldsymbol{z})))] \approx \frac{1}{M}\sum_{i=1}^{M} H(p(y|G(\boldsymbol{z}_i))), \quad \boldsymbol{z}_i \sim p_z, \tag{2.13}$$

其中 M 表示独立采样的噪声样本个数（CatGAN 令 $M = N$）。为了满足判别器和生成器的最后一个要求：均匀使用所有类别（对应均匀的边缘分布），CatGAN

分别在真实数据集 $\mathcal{X}$ 以及虚假样本集上定义边缘分布的熵估计：

$$\begin{aligned} H_{\mathcal{X}}(p(y)) &= H\left(\frac{1}{N}\sum_{i=1}^{N} p(y|\boldsymbol{x}_i)\right), \\ H_G(p(y)) &\approx H\left(\frac{1}{M}\sum_{i=1}^{M} p(y|G(\boldsymbol{z}_i))\right), \quad \boldsymbol{z}_i \sim p_z. \end{aligned} \tag{2.14}$$

判别器和生成器分别最大化上面两个熵，其物理意义是对于判别器，希望均匀使用所有类别；对于生成器，希望生成的数据是类别均匀的。此时，判别器的要求 (iii) 和生成器的要求 (ii) 即可满足。对于生成器的要求 (i)，直接最小化(2.13)式即可。注意，对于判别器，需要最大化(2.13)式，其目的对虚假样本保持较高的不确定性；对于生成器，需要最小化(2.13)式，其目的是希望判别器对于虚假样本具有较高的确定性，以判别为真。这样一来，就形成了对抗。CatGAN 模型结构见图 2.3，图中给出了 CatGAN 目标函数的形象解释[3]。

结合(2.12)(2.13)(2.14)式我们可以得到 CatGAN 的目标函数 L_D^{cat} 和 L_G^{cat}：

$$\begin{aligned} L_D^{cat} &= \max_D \; H_{\mathcal{X}}(p(y)) - \mathrm{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[H(p(y|\boldsymbol{x}))\right] + \mathrm{E}_{\boldsymbol{z}\sim p_z}\left[H(p(y|G(\boldsymbol{z})))\right], \\ L_G^{cat} &= \min_G \; -H_G(p(y)) + \mathrm{E}_{\boldsymbol{z}\sim p_z}\left[H(p(y|G(\boldsymbol{z})))\right]. \end{aligned} \tag{2.15}$$

以上给出的目标函数满足所有的设计要求，并且具有一定的可解释性：(2.15)式中 L_D^{cat} 的前两项可以合并为 $I(p_{\text{data}};\, p(y))$，其中 $p(y)$ 预测标签分布。也就是说，判别器在最大化真实数据分布和预测标签分布之间的互信息，同时最小化虚假样本分布和预测标签分布之间的互信息。同理，对于 L_G^{cat} 前两项也可以合并为虚假样本和预测分布之间的互信息，这说明生成器想要最大化 $I(p_g;\, p(y))$。很多判别式分类模型如感知机、逻辑回归等，都可以用信息论的角度解释为优化数据和标签之间的互信息，提取数据和标签相关的部分，寻找使得 $\Pr(y|\theta(x)) = \Pr(y|x)$ 的充分统计量 $\theta(x)$，作为模型所提取到的特征。这里，CatGAN 也具有类似的可解释性。

2.5.4 半监督学习

如果拥有少量标签信息，CatGAN 就可以扩展到半监督学习模型，此时模型的性能可以达到进一步提升。考虑2.5.2节所描述的问题，现在在原有的 N 个无标注数据样本上，额外增加 L 个带标签的数据 $\mathcal{L} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{L}$，使用 $\boldsymbol{y}_i \in \mathbb{R}^K$

[3]CatGAN 模型结构图摘自 Springenberg[29]。

表示标签 y_i 经过 one-hot 编码之后的标签向量，即若 $y_i = k$，则 $\boldsymbol{y}_{ik} = 1$ 且对任意 $j \neq k$, $\boldsymbol{y}_{ij} = 0$。通过计算预测标签分布 $p(y|\boldsymbol{x})$ 和 $\mathcal{L}$ 上的真实标签分布之间的交叉熵，可以将这些有标签数据整合进(2.15)式中。给定一组 $(\boldsymbol{x}, \boldsymbol{y})$，则预测标签分布和真实分布之间的交叉熵具有如下形式:

$$\mathrm{CE}[\boldsymbol{y}, p(y|\boldsymbol{x})] = -\sum_{i=1}^{K} y_i \log \Pr(y = y_i|\boldsymbol{x}), \tag{2.16}$$

其中 y_i 是标签向量 $\boldsymbol{y}$ 的第 i 个分量。综上所述，可知半监督版本的 CatGAN 目标函数 L_D^{sscat} 和 L_G^{sscat} 具有如下形式:

$$\begin{aligned} L_D^{sscat} &= L_D^{cat} + \lambda \mathrm{E}_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{L}} \left[\mathrm{CE}[\boldsymbol{y}, p(y|\boldsymbol{x})]\right], \\ L_G^{sscat} &= L_G^{cat}, \end{aligned} \tag{2.17}$$

其中 λ 为正则化系数。

2.6　本章小结

本章介绍了生成对抗网络的基本思想及其数学模型。接着详细阐述了两种 GAN 的变体：InfoGAN 和 CatGAN。其中 InfoGAN 将噪声空间分解为无意义的噪声和有意义的隐变量，通过加入互信息正则项可以让隐变量学习到数据特征，并可以通过隐变量控制生成器的输出，为本文后续工作做了良好的铺垫。CatGAN 则对朴素 GAN 中的判别器进行扩展，将原来的二分类扩展为多分类判别器，并利用判别器做无监督和半监督的多分类任务。

第 3 章　基于互信息正则的分类模型

3.1　引言

在原始的生成对抗网络中，生成器通过将噪声映射成数据，然后由对抗损失函数驱动，进而能够在训练过程中慢慢学习到真实数据的分布信息，最终生成较为逼真的虚假数据。生成对抗网络的特殊性在于它创新地结合了生成式模型和判别式模型。我们既可以认为它训练了一个判别式模型，也可以认为它训练了一个生成式模型。它的最大贡献在于提出了一个对抗训练的机制，而且本身没有过多的约束，这为后续的研究提供了巨大的可扩展性。Goodfellow 等[6] 在给出生成对抗网络的模型以及结果之后，在文章最后关于 GAN 的扩展工作给几点建议：

1. 可以为生成器和判别器添加条件信息，此时生成器可以学习到对应的条件分布；

2. 通过增加一个辅助网络来估计 $p(\boldsymbol{z}|\boldsymbol{x})$，可以进行进一步的统计推断；

3. 半监督学习：当拥有少量标签信息时，判别器学到特征可以用于提高分类器的性能。

鉴于生成对抗网络具有如此良好的可扩展性，越来越多的学者开始研究该模型[40,44,52]。尽管如此，关于生成器如何将噪声映射成数据的细节仍有待探索。许多研究表明，通过对隐变量连续插值，会在生成的图片上得到连续平滑的变化[42,46,52,53]。然而，大多数变化无法解释并且没有明确的意义。Chen 等[42] 提出将隐变量分解，通过在训练过程中增加隐变量和生成数据之间的互信息，达到了将特征解耦的效果。此时，InfoGAN 的隐变量可以明确对应到一个个有意义的数据特征（如 MNIST 中手写数字的角度、笔画粗细等）。这证明了互信息约束在生成对抗网络中有值得探究的作用。

3.2　C-InfoGAN

在传统的监督分类方法中，模型从训练集 $\mathcal{L}_n = \{\mathbf{x}_i, y_i\}_{i=1}^n$ 学习一个决策边界，其中 $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \Omega = \{\omega_1, \ldots, \omega_K\}$。对于未见样本，模型通过自身的决策边界给出预测值。现在我们考虑无监督情况，即对于所有训练样本 $\mathbf{x}_i$，其对应标

签 y_i 都是未知的。换句话说，训练集只含有大量未标注的原始数据。这种情况通常无法分类，因为连目标类别都是未知的。上述问题通常定义为聚类更合适，此处考虑添加一个额外信息：类别总数 K 已知，但具体类别未知。这时，对于给定数据输入，模型可以生成 K 个虚假类别，然后为每个输入分配一个虚假类别。在测试集上评估模型的时候，可以利用有限的标签将虚假标签和真实标签对应（具体方法参见第 4章开头部分），从而得到一个分类器。

如 2.4节所述，InfoGAN 通过最大化隐变量 $\boldsymbol{c}$ 和生成数据 $G(\boldsymbol{z}, \boldsymbol{c})$ 之间的互信息，可以无监督地学习到数据的解耦合特征（disentangled representation），这在一定程度上解释了隐空间的结构变化对生成图片的影响。在模型收敛之后，隐变量 $\boldsymbol{c}$ 的每一维度都能绑定到数据的某个特征。比如对于 MNIST 数据集 $\mathbf{c} = (c_1, c_2, c_3)$，其中 $c_1 \sim \mathrm{DUnif}(0, 9)$，$c_1$，$c_2 \sim \mathrm{Unif}(-1, 1)$，离散隐变量 c_1 绑定到数字的类别，连续隐变量 c_1，c_2 绑定到数字的倾斜角度和笔画粗细。本节基于 InfoGAN 的特性，使用辅助网络 Q 来做分类，提出 Classifier InfoGAN (C-InfoGAN) 模型。

3.2.1 无监督分类方法

InfoGAN 通过引入互信息约束探究了隐空间和数据空间的联系，在精心设计之下，能够达到每个隐变量对应生成数据一个特征的效果。值得注意的是，在 MNIST 数据集上，隐变量 c_1 绑定到了数字的类别特征，加上辅助网络 Q 是对后验概率 $p(\boldsymbol{c}|\boldsymbol{x})$ 的估计，这天然地为分类任务提供了基础。本文基于 InfoGAN 的特点，利用 InfoGAN 的 Q 网络作为分类器，提出 Classifier InfoGAN (C-InfoGAN) 模型。

具体来说，本文在 InfoGAN 的目标函数上添加一个正则项 $L(c, \hat{c})$，其中 $\hat{c} = Q(c|\tilde{\boldsymbol{x}}) \in \mathbb{R}^K$ 是 Q 网络的输出，这里的 c 仅代表绑定到类别特征的一维离散隐变量。在训练过程中，该正则项可以驱使 Q 网络的输出与输入隐变量尽可能接近。这实际上是让隐变量 c 充当虚假标签，虽然在训练初期这个虚假标签没有任何意义，但是通过生成对抗网络的对抗机制，在生成器能够生成逼真数据 $G(\mathbf{z}, \mathbf{c})$ 的时候，此时的 c 就具有一定的意义。这是因为 InfoGAN 在训练过程中最大化互信息

$$I(\mathbf{c};\ \tilde{\mathbf{x}}) = H(\boldsymbol{c}) - H(\boldsymbol{c}|\tilde{\boldsymbol{x}}).$$

在整个训练过程中，c 的先验分布不变，所以 $H(\boldsymbol{c})$ 可视为常量，最大化互信息意味着最小化 $H(c|\tilde{\boldsymbol{x}})$。当生成器能够生成逼真数据的时候 $\tilde{\boldsymbol{x}} = G(\boldsymbol{z}, \boldsymbol{c}) \approx \boldsymbol{x}$，所以此时 $H(c|\boldsymbol{x})$ 应该也较小。这背后的物理意义就是，在给定真实数据 $\boldsymbol{x}$ 之后，离散类别隐变量 c 的不确定性较低，即 $p(c|\boldsymbol{x})$ 呈现单峰分布。因此用 c 来作为 $\boldsymbol{x}$ 的虚假标签是有意义的。为了简便起见本文将 C-InfoGAN 模型简称为 CIG，其目标函数如下：

$$\min_{G,Q} \max_{D} V_{\text{CIG}}(G, D, Q, \lambda_1, \lambda_2) = V_{\text{InfoGAN}}(G, D, Q, \lambda_1) + \lambda_2 L(c, Q(c|\tilde{\boldsymbol{x}})), \quad (3.1)$$

其中 λ_2 是正则化系数，$L(c, \hat{c}) = L(c, Q(c|\tilde{\boldsymbol{x}}))$ 在实现中一般采用均方误差或交叉熵，参见 (3.10)式，模型结构见图 3.1。

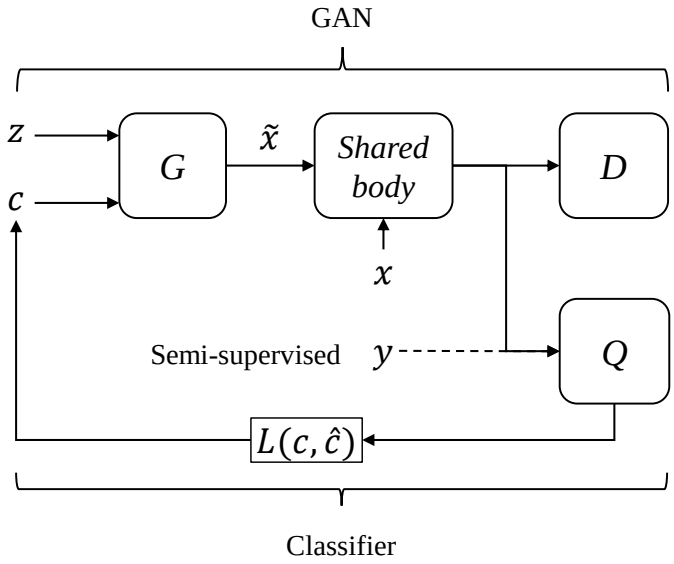


图 3.1 C-InfoGAN 模型结构。无监督情况下，生成数据 $\tilde{x}$ 和真实数据 x 参与训练，通过和 D 共享部分结构，Q 网络可以将 GAN 模型学习到的特征加以利用，实现分类任务；在半监督情况下，一部分真实标签 y 会直接被 Q 网络利用，以得到更好的效果。优化 Q 网络的输出 $\hat{c}$ 和隐变量 c 构成的损失函数 $L(c, \hat{c})$ 来增加 Q 的分类准确率。

Figure 3.1 The architecture of C-InfoGAN. In unsupervised case, the generated data $\tilde{x}$ and the real data x are used for training. By sharing the body with D, Q is able to using features learned by GAN framework to perform classification. In semi-supervised case, labels y is directly fed into Q to get better performance. We optimize the loss function $L(c, \hat{c})$ of latent code c and the output $\hat{c}$ of Q to improve its accuracy.

使用 InfoGAN 做分类并不是一个新的想法，Zhang 等[54] 基于 InfoGAN 提出了一种无监督分类方法。他们生成对抗网络训练的同时，训练一个 Parital

Inverse Filter (PIF)，它接受一个样本作为输入，输出一个和隐变量同维度的向量。之所以叫做 PIF，是因为它可以看作生成器 G 的逆映射，将数据映射到隐空间，但又不是完全还原输出噪声，它只输出噪声中的隐变量部分。训练过程中，将 PIF 的输出与隐变量作均方误差，使得 PIF 的输出和隐变量的取值尽可能接近。事实上，这个 PIF 和 InfoGAN 中的辅助网络具有类似的作用，都是和输入噪声中的隐变量发生联系，而实践发现，使用 Q 网络做分类已经具有可观的效果，而且对计算量的增加较小，模型结构也相对简单。

算法 1给出了 C-InfoGAN 的训练步骤，其中 θ_g, θ_d, θ_q 分别是 G, D 和 Q 的网络参数。

算法 1 Training procedure for C-InfoGAN

1: **for** numbers of training iterations **do**

2: Sample a batch of $\boldsymbol{x} \sim p_{\text{data}}(x)$ of size m.

3: Sample a batch of noise $\boldsymbol{z} \sim p_z$, $\boldsymbol{c} \sim p_c$ of size m.

4: Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \left[\frac{1}{m} \sum_{i=1}^{m} \Big(\log D\left(\boldsymbol{x}_i\right) + \log\left(1 - D\left(G(\boldsymbol{z}_i, \boldsymbol{c}_i)\right)\right) \Big) \right].$$

5: Update G and Q by descending along its stochastic gradient:

$$\nabla_{\theta_g, \theta_q} \left[\frac{1}{m} \sum_{\tilde{\boldsymbol{x}}} \Big(\log(1 - D(G(\boldsymbol{z}, \boldsymbol{c}))) - p(\boldsymbol{c}) \log Q(G(\boldsymbol{z}, \boldsymbol{c})) \Big) \right].$$

6: **end for**

3.2.2 半监督分类方法

当拥有少量标签信息时，C-InfoGAN 可以利用这些标签进一步提升分类准确率和生成效果。同时将隐变量 c 直接绑定到真实的标签，实现精准调控。针对少量标注信息，Spurr 等[55] 提出半监督的 InfoGAN 模型，称为 ss-InfoGAN。图 3.2给出了其模型结构[1]，ss-InfoGAN 将隐变量 c 进一步分解为无监督部分 c_{us}，负责捕捉大量无标注数据的潜在特征；和有监督部分 c_{ss}，负责捕捉已有标签 y。同时他们设置了两组隐变量对应的先验分布，以及对应的辅助网络 Q_{us} 和 Q_{ss}，使用隐变量 c_{ss} 和辅助网络 Q_{ss} 专门处理那部分有标注信息。本文直接将标签信息加入 Q 网络，先用真实数据和标签训练，接着用生成数据和虚假标签

[1]ss-InfoGAN 模型结构图摘自 Spurr 等[55]。

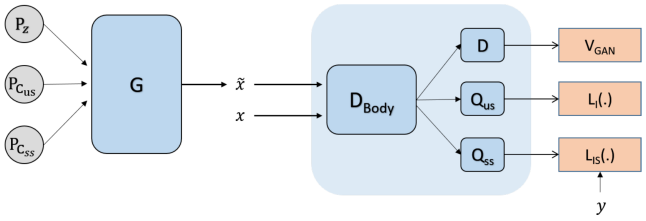


图 3.2 ss-InfoGAN 结构示意

Figure 3.2 The architecture of ss-InfoGAN

(即隐变量 c) 来训练，这样做的目的是为了使真实标签的信息流入隐变量 c 中，或者可以说是用真实标签指导 c 绑定到正确的类别特征。经过实践发现，简单地使用上述方法也能达到同样的效果，而且模型更为简单。使用和2.5.4节中类似的方法，我们给出半监督 C-InfoGAN 的目标函数:

$$\begin{aligned}\min_{G,Q}\max_{D}\ & V_{\text{ss-CIG}}(G,D,Q,\lambda_1,\lambda_2,\lambda_3)=V_{\text{CIG}}(G,D,Q,\lambda_1,\lambda_2)+\\ & \lambda_3\mathrm{E}_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{L}}\left[\mathrm{CE}[\boldsymbol{y},Q(y|\boldsymbol{x})]\right].\end{aligned}\tag{3.2}$$

其中 λ_3 为正则化系数，模型结构参见图 3.1。

算法 2给出了半监督版本的 C-InfoGAN 的训练步骤，其中 θ_g, θ_d, θ_g 分别为 G, D 和 Q 的网络参数。可以看到，在第11行我们直接将有标签数据放入 Q 网络中训练，通过 Q 网络让标签信息流入虚假标签。在这种情况下，训练稳定之后绑定到类别的隐变量能够和真实标签一一对应（参见 4.3节）。在半监督学习中，如何充分利用有标签数据是一个很重要的问题。在算法 2中，我们采用投掷硬币的方式，决定如何采样。假设有一枚非均匀硬币，投掷一次正面向上的概率为 p。设用随机变量 X 表示投掷一次该硬币产生的结果，如正面向上，则 $X=1$，反之 $X=0$，显然 $X\sim\mathrm{Bern}(p)$。每一次采样前，投掷一次硬币，如果正面向上，则从有标签的数据中采样，否则从无标签数据采样。随着迭代次数的增加，逐渐减小正面向上的概率 p 至某个确定的值。这样做的好处是当 p 减小到最小值之后，虽然由无标签数据占据主导，但还是会偶尔出现一两个有标签的数据批次参与训练，给予无监督训练一定的指导。

算法 2 Training procedure for semi-supervised C-InfoGAN

for numbers of training iterations **do**

  Sample `flag` from Bern(p). ▷ Toss a coin to decide whether to use labels

  if `flag` is 1 **then**

    Sample a batch of labeled samples $(\boldsymbol{x}, y) \sim p_{\text{data}}(x, y)$ of size m.

  else

    Sample a batch of unlabeled samples $\boldsymbol{x} \sim p_{\text{data}}(x)$ of size m.

  end if

  Sample a batch of noise $\boldsymbol{z} \sim p_{z},\ \boldsymbol{c} \sim p_{c}$ of size m.

  Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d}\left[\frac{1}{m}\sum_{i=1}^{m}\Big(\log D\left(\boldsymbol{x}_i\right)+\log\left(1-D\left(G(\boldsymbol{z}_i,\boldsymbol{c}_i)\right)\right)\Big)\right].$$

  if `flag` is 1 **then**

    Update Q by ascending its stochastic gradient: ▷ Bind real labels to fake

$$\nabla_{\theta_q}\left[\frac{1}{m}\sum_{(\boldsymbol{x},\,y)}p(y|\boldsymbol{x})\log Q(\mathbf{x})\right].$$

  end if

  Update G and Q by descending along its stochastic gradient:

$$\nabla_{\theta_g,\theta_q}\left[\frac{1}{m}\sum_{\tilde{\boldsymbol{x}}}\Big(\log(1-D(G(\boldsymbol{z},\boldsymbol{c})))-p(\boldsymbol{c})\log Q(G(\boldsymbol{z},\boldsymbol{c}))\Big)\right].$$

  $p \leftarrow \max(0.01, \text{Annealing}(p, \mathit{iterations}))$ ▷ Gradually anneals p to 0.01

end for

3.3　InfoCatGAN

生成对抗网络具有很高的可扩展性，自提出以来应用广泛。Salimans 等[37], Odena[38] 提出使用生成对抗网络模型做半监督分类。在他们的模型中，判别器需要输出 $K+1$ 个类别。考虑一个传统分类模型，给定输入 x，分类器需要将其分类到 K 个类别中的一个。这样的模型通常接受 x 为输入，输出一个 K 维向量 $(\ell_1,\dots,\ell_K)$。通过 softmax 变换可以将这个向量转化为类别概率 $p(y=j|x)=\frac{\exp(\ell_j)}{\sum_{k=1}^{K}\exp(\ell_k)}$。现考虑使用生成对抗网络做分类，可以通过增加一个类别来对应生成器生成的虚假数据。也就是说，判别器将所有数据分为 $K+1$ 个类别 $\Omega=\{\omega_1,\dots,\omega_K,\omega_{K+1}\}$，其中 ω_{K+1} 对应虚假数据类别。此时，可以用 $p(y=\omega_{K+1}|x)$ 来表示 x 为虚假数据的概率，对应朴素 GAN 模型中的 $1-D(x)$。对于大量无标签数据，可以通过最大化 $\log p(y\in\{\omega_1,\dots,\omega_K\}|x)$ 来训练分类器。这样一来，在拥有少量标签的情况下，对于真实数据和虚假数据都有了对应的损失函数

$$L_{\text{ss}}=-\mathrm{E}_{x,y\sim p_{\text{data}}(x,y)}[\log p(y|x)], \tag{3.3}$$

$$L_{\text{us}}=-\mathrm{E}_{x\sim p_{\text{data}}(x)}\log[1-p(y=\omega_{K+1}|x)], \tag{3.4}$$

$$L_{\text{fake}}=-\mathrm{E}_{x\sim p_g}[\log p(y=\omega_{K+1}|x)], \tag{3.5}$$

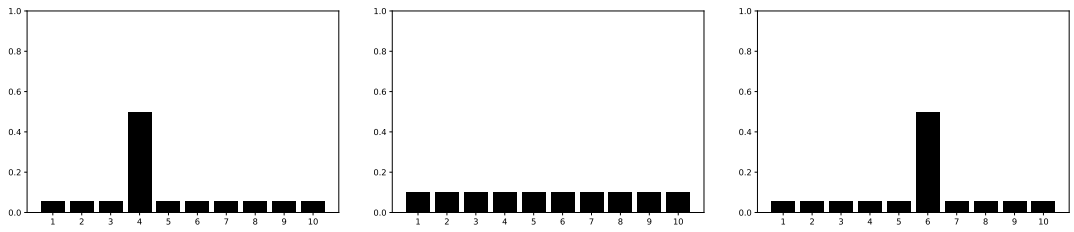
其中 L_{ss}，L_{us}，L_{fake} 分别表示针对有标签数据，无标签数据以及虚假数据的损失函数函数。对于有标签数据，直接利用标签信息优化预测标签和真实标签之间的交叉熵；对于无标签数据，由于标签信息无从得知，所以只能笼统地增加 $1-p(y=\omega_{K+1}|x)$ 来减小被误判的概率；对于虚假数据，都将它判为 ω_{K+1}。

Springenberg[29] 提出的 CatGAN 模型既可以作无监督分类，也可以做半监督分类。与 Salimans 等[37], Odena[38] 类似，CatGAN 将判别器扩展为多分类器，令其输出 K 个类别的概率。不同的是，CatGAN 重新设计了基于条件熵的损失函数（参见 2.5）。本节在 CatGAN 的基础上，添加互信息约束，提出 InfoCatGAN 模型。该模型同样支持无监督和半监督分类，但是其在生成图片的质量上优于 CatGAN。

3.3.1　无监督分类方法

在训练概率分类模型的过程中，通过优化条件熵可以将分类边界调整到更自然的位置（数据分散区域）[51]，因此 CatGAN 使用条件熵作为判别器判断真

假数据的依据。但是，使用熵作为目标函数的一个缺点是没有类别指向性。考虑一个离散随机变量 Y 表示给定输入 $\mathbf{x}$ 对应的标签，Y 可能的取值为 $\text{Supp}(Y) = \Omega = \{\omega_1, \ldots, \omega_K\}$。如果模型仅仅最小化 $H(Y|\mathbf{x})$，那么它无法预测 $\mathbf{x}$ 的具体类别。因为任何一个单峰分布都可以使得条件熵达到最小，即 K 个类别中任意一个都可以使 $p(y|\mathbf{x})$ 呈单峰分布，见图 3.3。对于一个分类器，我们希望对



(a) $H = 2.58$, **predict** $y = 4$ **(b)** $H = 3.32$ **(c)** $H = 2.58$, **predict** $y = 6$

图 3.3 不同分布的熵

Figure 3.3 Entropy of different distributions

于给定输入 $\mathbf{x}$，有且仅有一个 $k \in [K]$，使得 $p(y = \omega_k|\mathbf{x})$ 最大，而对于任意 $k' \neq k$, $p(y = \omega_{k'}|\mathbf{x})$ 均很小。然而问题在于训练数据集没有标注，每个数据样本对应的标签无从获得。

对于上述问题本文从 InfoGAN 中获得启发，提出 InfoCatGAN 模型。InfoGAN 将输入噪声划分为 $\mathbf{z}$ 和 $\mathbf{c}$，实际上是对隐空间的结构进行了人为划分。一部分提供模型的容量，使得模型具有足够的自由度去学习数据的细节（高度耦合的特征）；一部分提供隐变量，用于在学习过程中绑定到数据的显著特征（如：MNIST 中的数字类别、笔画粗细、角度）。模型的核心思想如下：通过在隐空间构造一维隐变量 c，在训练过程中将生成数据的类别标签与之绑定，使得可以通过 c 来控制生成数据的类别。CatGAN 对 GAN 的扩展主要在于改变了判别器的输出结构：为所有真实数据分配一个类别标签而对于虚假数据则保持一个不确定的状态。类似的，生成器应该致力于生成某个具体类别的数据而不是仅仅生成足够逼真的图片。因此，我们构造的隐变量实际上充当了虚假图片的标签，它在训练过程中约束着生成器生成指定类别的图片。

下面给出 InfoCatGAN 的损失函数：设 $\mathbf{x} \in \mathcal{X}$ 为一个真实数据样本，$\tilde{\mathbf{x}} = G(\mathbf{z}, c)$ 为一个生成数据，其中 $z \sim p_z$ 为噪声，$c \sim p_c$ 为隐变量。为了简单起见，这里只考虑 c 为一维离散随机变量，p_c 为离散均匀分布。生成器 $G = G(\mathbf{z}, c;\ \theta_G)$

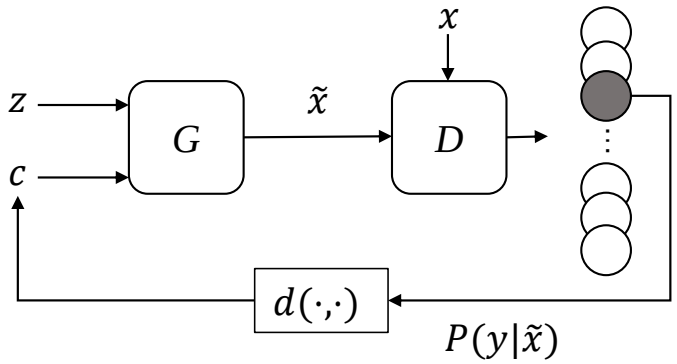


图 3.4 InfoCatGAN 模型结构。图中 D 的输出为 $P(y|\cdot)$。在训练生成器的时候，将判别器的输出 $P(y|\tilde{\mathbf{x}})$ 和隐变量 c 通过某种度量 $d(\cdot,\cdot)$ 建立联系使得条件概率的峰值与 c 的取值对应。

Figure 3.4 The architecture of InfoCatGAN. The discriminator outputs $P(y|\cdot)$. When training generator, we add a regularizer $d(\cdot,\cdot)$ between the ouput of discriminator and the latent code c to match the peak of $P(y|\tilde{\mathbf{x}})$ with c.

和判别器 $D = D(\mathbf{x};\ \theta_D)$ 均为可微深度神经网络，其中 θ_G, θ_D 分别为生成器和判别器的参数[2]。通过在 D 网络的最后一层做 softmax 变换，可以直接将 $D(x)$ 作为条件概率 $p(y|x)$ 的估计。注意到(2.15)式可以重写为：

$$L_D^{\text{cat}} = -I(X;Y) - \mathrm{E}_{\tilde{\mathbf{x}}\sim p_g}[H(p(y|\tilde{\mathbf{x}}))], \tag{3.6}$$

$$L_G^{\text{cat}} = -I(\tilde{X};Y), \tag{3.7}$$

其中 $X \sim p_{\text{data}}$, $\tilde{X} \sim p_g$ 分别表示真实数据和虚假数据对应的随机变量，Y 表示未知标签对应的随机变量。从(3.6)、(3.7)式可以看出，CatGAN 在优化数据与标签之间的互信息。互信息是常用的变量间相关性的衡量标准，所以本文用它作为生成器损失函数的正则项，由此得到 InfoCatGAN 的损失函数如下：

$$\begin{aligned} L_D &= L_D^{\text{cat}}, \\ L_G &= L_G^{\text{cat}} - \lambda_1 I(c;\tilde{\mathbf{x}}), \end{aligned} \tag{3.8}$$

其中 λ_1 为正则系数，可知当 $\lambda_1 = 0$ 时，InfoCatGAN 退化为 CatGAN，模型结构见图 3.4。参考(2.8)式，$I(c;\tilde{\mathbf{x}})$ 可以放缩为 $\mathrm{E}_{p(\mathbf{c},\tilde{\mathbf{x}})}[\log p(c|\tilde{\mathbf{x}})]$，在实现中通常使用交叉熵

$$CE[\mathbf{c}, p(c|\tilde{\mathbf{x}})] = -\sum_{i=1}^{K} c_i \log p(c = c_i|\tilde{\mathbf{x}}) \tag{3.9}$$

[2]为了简便起见，在无歧义的情况下通常省略网络参数。

来优化此项，这里的 $\mathbf{c} \in \mathbb{R}^K$ 是隐变量 c 经过 one-hot 编码之后的向量，$p(c|\tilde{\mathbf{x}})$ 可以用 $D(\tilde{\mathbf{x}})$ 来近似。

算法 3给出了 InfoCatGAN 的训练步骤，其中 θ_g, θ_d 分别为 G, D 的网络参数。第4、5行的更新公式可以参考(3.8)式和(2.15)式。从第 5行可以看出，我们将隐变量有意识地和虚假图片的类别绑定，这样做的效果是训练稳定后，可以通过隐变量控制生成图片的类别。注意到第4、5行的最后一项分别对应着(2.15)中的 $H_{\mathcal{X}}(p(y))$ 和 $H_G(p(y))$，Springenberg[29] 指出边缘分布的熵的估计方法应当视情况而定。原本 $H_{\mathcal{X}}$ 应当针对整个训练数据集来计算边缘分布，然后在计算熵；而 H_G 也不能单单只用一个批次的虚假样本计算，应当生成远大于批处理大小的虚假样本来计算边缘分布，再计算熵。但是在批处理大小 m 远大于类别总数 K（比如 $K = 10$, $m = 100$）的时候，算法 3中对于边缘分布的熵的估计是合理的[3]。

算法 3 Training procedure for InfoCatGAN

1: **for** numbers of training iterations **do**

2: Sample a batch of $\boldsymbol{x} \sim p_{\text{data}}(x)$ of size m.

3: Sample a batch of noise $\boldsymbol{z} \sim p_z$, $\boldsymbol{c} \sim p_c$ of size m.

4: Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d}\left[\frac{1}{m}\sum_{i=1}^{m}\Big(-H(D(\boldsymbol{x}_i)) + H(D(G(\boldsymbol{z}_i,\boldsymbol{c}_i)))\Big) + H\left(\frac{1}{m}\sum_{i=1}^{m}D(\boldsymbol{x}_i)\right)\right].$$

5: Update G and D by descending along its stochastic gradient:

$$\nabla_{\theta_g,\theta_d}\left[\frac{1}{m}\sum_{\tilde{\boldsymbol{x}}=G(\boldsymbol{z},\boldsymbol{c})}\Big(H(D(\tilde{\boldsymbol{x}})) - p(\boldsymbol{c})\log D(\tilde{\boldsymbol{x}})\Big) - H\left(\frac{1}{m}\sum_{i=1}^{m}D(G(\boldsymbol{z}_i,\boldsymbol{c}_i))\right)\right].$$

6: **end for**

3.3.2 半监督分类方法

作为 CatGAN 的扩展，InfoCatGAN 能够很自然地适用于半监督的情况。假设 $\mathcal{L} = \{\mathbf{x}_i, y_i\}_{i=1}^m$ 为 m 个有标签的样本，$\mathbf{y}_i \in \mathbb{R}^K$ 表示标签 y_i 经过 one-hot 编码之后的向量：即如果 $\mathbf{x}_i$ 的标签是 ω_k，则 $\mathbf{y}_{ik} = 1$ 且对于所有的 $j \neq k$, $\mathbf{y}_{ij} = 0$。对于有标签的样本，$D(\mathbf{x})$ 的分布信息可以明确获得，所以可以通过计算 $\mathbf{y}$ 和

[3]本文在实验中也采用了校正后的边缘分布的熵的估计方法，发现这对实验结果影响并不大。

$p(y|\mathbf{x})$ 之间的交叉熵:

$$\mathrm{CE}[\mathbf{y}, p(y|\mathbf{x})] = -\sum_{i=1}^{K} y_i \log p(y = y_i|\mathbf{x}) \tag{3.10}$$

来辅助判别器做出更精确的判断。半监督版本的 InfoCatGAN 损失函数如下:

$$L_D^L = L_D + \lambda_2 \mathrm{E}_{(\mathbf{x},\mathbf{y})\in\mathcal{L}} \left[CE[\mathbf{y}, p(y|\mathbf{x})] \right], \tag{3.11}$$

其中 λ_2 为正则系数而生成器的损失函数同(3.8)式: $L_G^L = L_G$.

算法 4 Training procedure for semi-supervised InfoCatGAN

1: **for** numbers of training iterations **do**

2: Sample `flag` from Bern(p).

3: **if** `flag` is 1 **then**

4: Sample a batch of labeled samples $(\boldsymbol{x}, y) \sim p_{\text{data}}(x, y)$ of size m.

5: **else**

6: Sample a batch of unlabeled samples $\boldsymbol{x} \sim p_{\text{data}}(x)$ of size m.

7: **end if**

8: Sample a batch of noise $\boldsymbol{z} \sim p_z,\ \boldsymbol{c} \sim p_c$ of size m.

9: Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \left[\frac{1}{m} \sum_{i=1}^{m} \Big(-H(D(\boldsymbol{x}_i)) + H(D(G(\boldsymbol{z}_i, \boldsymbol{c}_i))) \Big) + H\left(\frac{1}{m} \sum_{i=1}^{m} D(\boldsymbol{x}_i) \right) \right].$$

10: **if** `flag` is 1 **then**

11: Update D by ascending its stochastic gradient:

$$\nabla_{\theta_d} \left[\frac{1}{m} \sum_{(\boldsymbol{x}, y)} p(y|\boldsymbol{x}) \log D(\mathbf{x}) \right].$$

12: **end if**

13: Update G and D by descending along its stochastic gradient:

$$\nabla_{\theta_g, \theta_d} \left[\frac{1}{m} \sum_{\tilde{\boldsymbol{x}}=G(\boldsymbol{z},\boldsymbol{c})} \Big(H(D(\tilde{\boldsymbol{x}})) - p(\boldsymbol{c}) \log D(\tilde{\boldsymbol{x}}) \Big) - H\left(\frac{1}{m} \sum_{i=1}^{m} D(G(\boldsymbol{z}_i, \boldsymbol{c}_i)) \right) \right].$$

14: $p \leftarrow \max(0.01, \mathrm{Annealing}(p, iterations))$

15: **end for**

算法 4给出了半监督版本的 InfoCatGAn 的训练步骤，其中 θ_q, θ_d 分别为 G, D 的网络参数。从第11行可以看出，如果当前批次是有标签的数据，则直接最小化真实标签和预测概率之间的交叉熵。这样做的目的通过判别器 D 将真实

标签信息流入虚假标签，即隐变量中。训练稳定之后，虚假标签和真实标签会一一对应（参加4.3)。在半监督版本的训练中，首先将数据集分为有标签部分和无标签部分，这里同样采用投掷硬币的方式，决定从哪个数据集采样。

3.4 本章小结

本章首先介绍了生成对抗网络的可扩展性：可以添加条件信息学习条件分布；可以增加逆向网络，通过数据空间推断隐空间；可以加入少量标签与半监督学习结合。接着介绍了本文提出的两个模型：C-InfoGAN 和 InfoCatGAN，这两个模型都可以做到无监督和半监督的分类，并且都是用了互信息约束。第一节详细阐述了无监督和半监督 C-InfoGAN 的模型结构及理论，并给出了具体的训练算法。第二节详细阐述了无监督和半监督 InfoCatGAN 的模型结构及理论，并给出了相应的训练算法。具体的实现细节和实验结果将在第4章给出。

第 4 章　模型评估

4.1　引言

本章将在 MNIST[35] 和 FashionMNIST[56] 数据集上评估模型。MNIST 是深度学习广泛使用的评价模型的基础数据集，它包含 60000 个训练样本，10000 个测试样本，每个样本均为 28×28 的灰度手写数字图片。FashionMNIST 具有和 MNIST 相同的数据结构，同样是 60000 个训练样本和 10000 个测试样本，每个样本均为 28×28 灰度图片，也具有 10 个类别。然而相较于 MNIST，FashionMNIST 的图片构成更为复杂，因此对模型具有更大的挑战。

在所有实验中，本文主要考察两个指标：分类准确率和图片生成质量。对于分类准确率，计算模型预测值并不像一般分类器那样直接。隐变量虽然可以学习到数据类别的特征，但是其取值并不一定和真实标签正确对应（例如 $c = 1$ 对应生成真实标签为 ω_2 的数据）。因此不能直接使用隐变量的取值作为模型的预测值，必须将隐变量的取值与真实标签之间做一个映射。

对于这个问题，本文采取与 Springenberg[29] 类似的做法。首先在测试集上选取 t 个测试样本 $\mathcal{X}_t = \{\mathbf{x}_i, y_i\}_{i=1}^t$，计算模型在这批数据上的预测概率矩阵

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{t1} & p_{t2} & \cdots & p_{tk} \end{bmatrix},$$

其中 $p_{ij} = \Pr(y = \omega_j | \mathbf{x}_i)$ 表示给定输入样本 $\mathbf{x}_i$ 时，模型预测 $y = \omega_j$ 的概率。然后对每一行选取最大概率的索引作为模型在这批样本上的预测值

$$\mathbf{P}^\star = (\ell_1, \ell_2, \cdots, \ell_t)^T, \quad \ell_i = \operatorname*{argmax}_j p_{ij}.$$

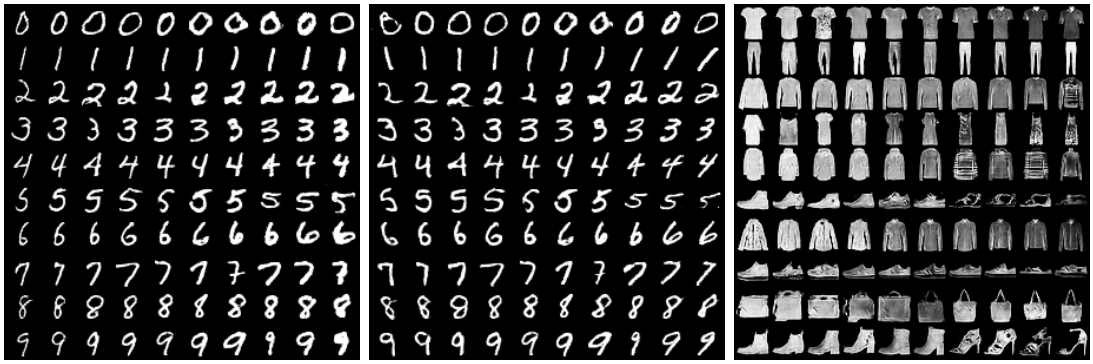
显然 $\ell_i,\ y_i \in \Omega = \{\omega_1, \omega_2, \cdots, \omega_K\}$。定义映射 $f : \Omega \to \Omega$,

$$f(\ell) = \operatorname*{argmax}_y \left(\sum_{\mathbf{x}:\ell(\mathbf{x})=\ell} \mathbf{1}\Big(y(\mathbf{x}), y\Big) \right),$$

其中 $\ell(\cdot)$ 表示模型预测值，$y(\cdot)$ 表示数据真实标签，$\mathbf{1}(a, b) = 1$ 当且仅当 $a = b$, 否则为零。此时，给定模型预测的虚假标签 ℓ，映射 f 都可以将它映射为真实标

签 $f(\ell)$。这就完成了虚假标签到真实标签的转换[1]。简单来说，模型为每一个数据 $\mathbf{x}_i$ 输出对应的概率向量 $p(y|\mathbf{x}_i)$，从而分配虚假标签 ℓ_i（概率向量中最大概率对应的索引），然后将预测值和真实标签对比：将虚假标签落入最多的真实标签的取值作为该虚假标签的取值。比如在所有 10 个被分类为虚假标签 ℓ_1 的样本中，有 9 个真实标签为 ω_3，则将虚假标签 ℓ_1 映射到真实类别 ω_3。

对于图片生成质量，本文采用 Fréchet Inception Distance (FID)[57] 来进行衡量[2]。相较于 Inception Score[37] 只考虑生成数据，FID 还利用了真实数据，因此更能反映生成数据和真实数据的差异。FID 越小代表生成的图片和真实图片越接近，生成质量越好。



(a) c_2 调控生成数字的粗细　(b) c_3 调控生成数字的角度　(c) c_2 调控生成图片的亮度

图 4.1 隐变量对生成图片的调控。图中每一行对应类别隐变量的一个取值，每一列对应其他某个隐变量（固定剩余隐变量）的连续变化。

Figure 4.1 The control of latent variables on the generated pictures. Each row in the figure corresponds to a value of the category latent code, and each column corresponds to variants of a continuous code with other codes fiexd.

4.2 实现细节

在实践中，模型中的所有模块（生成器、判别器、辅助网络等）均被实现为深度神经网络。具体实现可在 Github 上查看[3]。

[1]在实际应用中，可能会出现某一行的最大值不唯一的情况，此时选择第一个最大值索引作为输出。值得注意的是，这种情况在实验中并不多见。

[2]FID 一般用于彩色图片，而 MNIST 数据集是单通道的灰度图片，本文的做法是将单通道复制 3 份，形成一张 RGB 彩色图片然后再计算其 FID 值。

[3]实现代码地址：https://github.com/guyueshui/baGAN

4.2.1　C-InfoGAN

对于 C-InfoGAN 模型，我们的设定和 InfoGAN 保持大致相同。辅助网络 Q 和判别器 D 共享大部分网络结构，Q 在判别器的尾部（通常是倒数第二层）分离出一个全连接网络，输出后验概率的估计 $Q(\boldsymbol{c}|\boldsymbol{x})$。因此，C-InfoGAN 带来的计算复杂度增加较小。对于离散隐变量 c_i，在输出之前施加一个 softmax 激活层，得到的输出代表 $Q(c_i|\boldsymbol{x})$。对于连续隐变量 c_j，我们假设其先验分布为高斯分布，然后利用 Q 网络输出它的均值和方差，放到高斯分布中去拟合，计算差值。由于 GAN 在训练时容易坍塌，DCGAN[52] 提出了一种训练稳定的生成对抗网络结构，所以我们采用的网络结构对此有很多参考。关于超参数的选择我们也沿用 InfoGAN 的设定，对于离散隐变量设置为 1；对于连续隐变量，设置为较小的值 0.1。

表 4.1给出了 C-InfoGAN 在 MNIST 上的网络结构，其中各缩写的意义为：FC 表示全连接层；ReLU 代表 ReLU 激活函数；lReLU 代表 Leaky ReLU 激活函数; Tanh 代表双曲正切激活函数；sigmoid 表示 sigmoid 激活函数；softmax 表示 softmax 变换；conv 表示卷积层；upconv 表示反卷积层；bn 表示 batch normalization。如前所述，判别器 D 和辅助网络 Q 共享大部分网络结构。在 MNIST 数据集上，噪声空间的构成为一个 10 维离散隐变量，两个连续隐变量和 62 维高斯噪声，最终得到的噪声维度为 74。对于 FashionMNIST，我们在实践中发现使用同样的网络结构也能达到很好的效果。因此本文对于两个数据集采用同样的网络结构。

4.2.2　InfoCatGAN

通常情况下 GAN 的训练很不稳定。首先，如果判别器学习得太快，则(2.1)式可能变得不稳定（此时生成器的损失函数梯度消失）。其次，生成器可能对于某种特定的模式无法生成；或者是只能生成某种特定的模式而无法生成其它模式。首先我们在生成器和判别器加入 batch normalization，这样可以限制激活输出中的每个值的大小，并且实践证明加入 batch normalization 可以让生成器生成的图像更稳定，在给定少量标签的情况下，还可以提高判别器的泛化能力。值得一提的是，我们发现在输入层加噪声并没有对实验结果产生明显影响，这一点和 Springenberg[29] 不一致。表 4.2给出了 InfoCatGAN 的网络结构。在 MNIST 数

表 4.1 C-InfoGAN 在 MNIST 上的网络结构

Table 4.1 Architectures of C-InfoGAN used for MNSIT dataset

生成器 G	判别器 D/辅助网络 Q
Input $\in \mathbb{R}^{74}$	Input 28×28 Gray image
FC. 1024 ReLU. bn	4×4 conv. 64 lReLU. bn
FC. $7 \times 7 \times 128$ ReLU. bn	4×4 conv. 128 lReLU. bn
4×4 upconv. 64 ReLU. stride 2. bn	FC. 1024 lReLU. bn
4×4 upconv. 1 Tanh. stride 2	FC. 1 sigmoid. output for D, FC. 128-bn-lReRL-FC. output for Q

据集上，InfoCatGAN 的噪声空间的构成和 C-InfoGAN 类似，不同的没有连续隐变量，即包含一个 10 维离散隐变量和一个 62 维高斯噪声，最终得到的噪声维度为 72。对 FashionMNIST 也使用同样的配置。

表 4.2 InfoCatGAN 在 MNIST 上的网络结构

Table 4.2 Architectures of InfoCatGAN used for MNSIT dataset

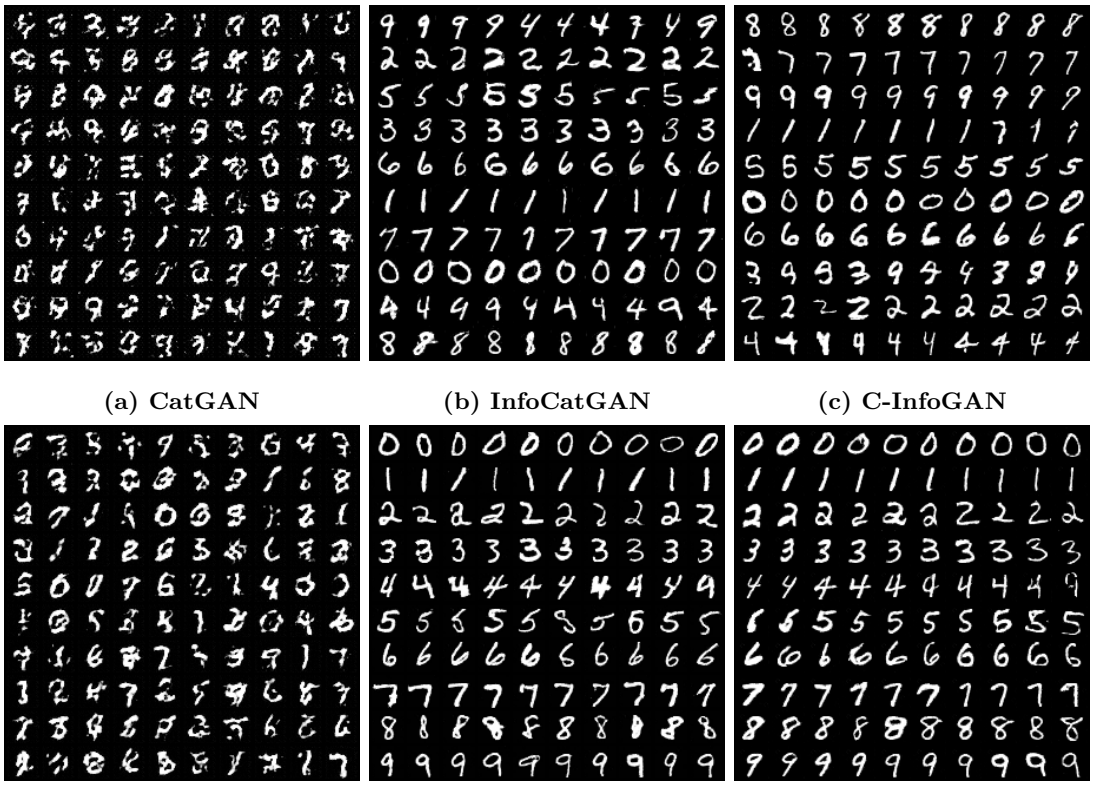
生成器 G	判别器 D
Input $\in \mathbb{R}^{72}$	Input 28×28 Gray image
FC. 1024 ReLU. bn	4×4 conv. 64 lReLU. stride 2
FC. $7 \times 7 \times 128$ ReLU. bn	4×4 conv. 128 lReLU. stride 2. bn
4×4 upconv. 64 ReLU. stride 2. bn	FC. 1024 lReLU. bn
4×4 upconv. 1 Tanh. stride 2	FC. 10-way softmax

4.3 实验结果

4.3.1 MNIST

图 4.2a和图 4.2b是在无监督情况下 CatGAN 和 InfoCatGAN 的生成效果，可以看到，InfoCatGAN 的生成效果明显高于 CatGAN，并且每一行基本是一种数字类别，对应隐变量的不同取值。半监督情况下有类似的结果，不同的是在少量标签信息的辅助下，InfoCatGAN 可以将隐变量 c 和真实标签正确绑定，

$c = 1$ 对应生成数字 '1'，见图 4.2e。CatGAN 生成的图片质量很差，原因在于其目标函数是为了分类而设计的。生成器的作用只是为了判别器能够更加鲁棒，如3.3.1节所述，从(2.15)式中可以看到，G 的目标函数只有条件熵，无法针对性地生成图片，从而会降低生成图片的质量。而 InfoCatGAN 由于增加了隐变量 c，并在训练过程中有意识地将生成数据的类别与之绑定，所以生成的图片质量较好。



(a) CatGAN　**(b) InfoCatGAN**　**(c) C-InfoGAN**

(d) CatGAN(132 labels)　**(e) InfoCatGAN(132 labels)**　**(f) C-InfoGAN(132 labels)**

图 4.2 模型在 MNIST 上的生成效果。在 InfoCatGAN 和 C-InfoGAN 的生成结果中，每一行对应隐变量 c 的一个取值，从 0 到 9。

Figure 4.2 Generated images on MNIST. Each row corresponds to a value of the categorical latent code c, from 0 to 9.

表 4.3给出了无监督和半监督情况下的分类准确率[4]。从表中看出，在无监督的情况下 InfoCatGAN 的分类准确率相较于 CatGAN 有所提升，而 FID 从 236.75 降低到 8.04，图片生成质量有了极大的提升，见图 4.2b。在半监督的情况下，CatGAN 的准确率达到了 96.05%，而 InfoCatGAN 随着正则化系数 λ_1 的变

[4]表中有关 CatGAN 的数据来自与本文复现的结果，与 Springenberg[29] 有所差距。经过多次尝试，我们仍无法完美复现出原文中的结果。

化呈现出不同的效果。当系数较小时，分类准确率较高，但生成图片的质量非常差；当系数较大时，生成的图片效果很好，但分类准确率有所降低。通过调节参数 λ_1，可以实现生成效果和分类准确率之间的折中。实验使用的默认值是 $\lambda_1 = 1.1$，当 λ_1 减小时，生成图片的质量开始下降，同时分类准确率也会相应增加。值得注意的是，λ_1 越小并不意味着分类准确率越高。当 $\lambda_1 = 0$ 时，InfoCatGAN 退化为 CatGAN；而从表 4.3可以看出，当 $\lambda_1 = 0.02$ 时，InfoCatGAN 的分类准确率高于 CatGAN，达到 98.15%。

表 4.3 MNIST 分类准确率对比

Table 4.3 Classification accuracy on MNIST

模型	准确率（%）	FID
CatGAN	64.41	236.75
InfoCatGAN	69.04	8.04
C-InfoGAN	77.65	7.55
CatGAN(132 labels)	96.05	127.92
InfoCatGAN(132 labels, $\lambda_1 = 0.02$)	**98.15**	169.72
InfoCatGAN(132 labels, $\lambda_1 = 0.03$)	80.43	**6.59**
C-InfoGAN(132 labels)	93.70	9.16

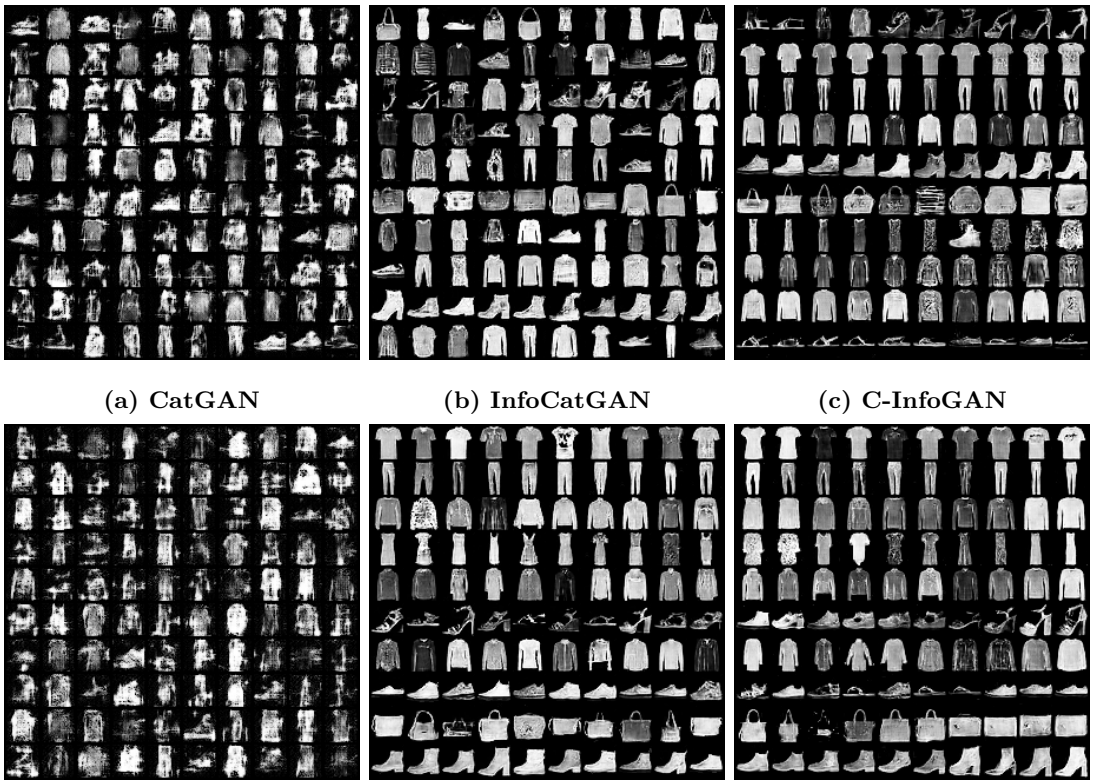
图 4.2c和图 4.2f给出了无监督和半监督情况下 C-InfoGAN 的生成结果。从图中可以看出无监督情况下，模型已经达到了很好的生成效果，隐变量 c 基本可以控制生成图片的类别，但是仍有部分类别未能精确控制（图 4.2c）；在半监督情况下，隐变量达到了精确的绑定，每一行对应生成一种类别的数字，而且顺序和真实标签是对应的。另外从图 4.1可以看出，C-InfoGAN 模型不仅可以生成指定类别的图片，并且可以通过额外的隐变量调节图片局部特征，如手写数字的粗细，角度等，这对指定特征的数据补足具有很大意义。

表 4.3给出了 C-InfoGAN 的准确率和 FID 及其与 CatGAN 模型的性能对比。可以看出，相较于 CatGAN 模型，C-InfoGAN 模型可以获得更高的准确率和生成质量，而且隐变量的绑定效果也更好。而在半监督情况下，C-InfoGAN 在保证生成质量的前提下，仍然能够达到 93.7% 的分类准确率。这是因为 InfoGAN 模型使用的是一个辅助网络 Q 来做类别绑定和分类任务，训练过程中并没有判

别器做过多约束，所以无论如何调整分类网络或更改分类约束，也不会对生成效果产生很大影响。这使得模型可以进一步利用生成的图片和标签扩充数据集，以达到更进一步的性能提升。

4.3.2　FashionMNIST

FashionMNIST[56] 是一个类似 MNIST 的数据集，二者拥有同样的结构，图像大小，同样的类别数目。但是相对于 MNIST，FashionMNIST 拥有更复杂的图像结构，以及更难获得非常高的分类准确率，所以对模型更具有检验性。



(a) CatGAN　(b) InfoCatGAN　(c) C-InfoGAN

(d) CatGAN(100 labels)　(e) InfoCatGAN(100 labels)　(f) C-InfoGAN(100 labels)

图 4.3 模型在 FashionMNIST 上的生成效果。在 InfoCatGAN 和 C-InfoGAN 的生成结果中，每一行对应隐变量 c 的一个取值，从 0 到 9。

Figure 4.3 Generated images on FashionMNIST. Each row corresponds to a value of the categorical latent code c, from 0 to 9.

图 4.3中给出了所有模型的生成结果。值得一提的是，加入互信息约束后的半监督版本模型从上往下每一行都对应同一个类别，并且顺序和训练数据的真实标签正确对应（图 4.3e、4.3f)。这说明了隐变量正确绑定到类别特征，并且可以精准调控生成图片的类别。

表 4.4给出了模型在 FashionMNIST 的数值结果，其中 InfoCatGAN 的正则系数 $\lambda_1 = 0.03$。从表中可以看出，无论是在无监督还是半监督情况下，CatGAN 的分类准确率都是比较高的，但从生成效果来看，它却是最差的。在无监督情况下，InfoCatGAN 提高了生成图片的质量，但与此同时，牺牲了分类准确率，而 C-InfoGAN 在一定程度上二者兼顾，不仅生成质量最优，而且具有相对较高的分类准确率，此外其模型复杂度也较低。在半监督情况下，InfoCatGAN 在两个方面均体现出优势，分类准确率达到了 74.21%，FID 为 16.92，生成效果见图 4.3e，这说明增加标签信息对模型的增益很大。

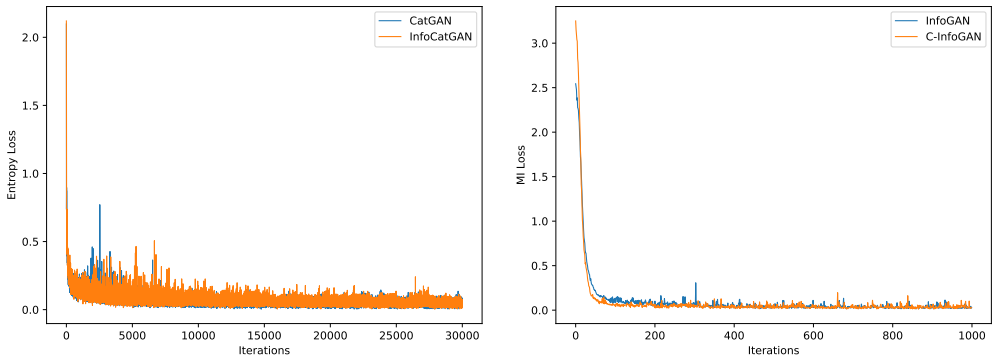
表 4.4 FashionMNIST 分类准确率对比

Table 4.4 Classification accuracy on FashionMNIST

模型	准确率（%）	FID
CatGAN	64.28	120.04
InfoCatGAN	57.66	26.43
C-InfoGAN	60.50	15.97
CatGAN(100 labels)	73.34	119.16
InfoCatGAN(100 labels)	**74.21**	16.92
C-InfoGAN(100 labels)	68.94	**15.94**

4.3.3 收敛速度分析

本文提出的两个模型在原理上都属于正则化生成对抗网络，因此相较于原先的两个模型 CatGAN 和 InfoGAN，增加的计算复杂度很小。由于 GAN 的训练方式特殊，训练的过程是生成器和判别器的对抗，因此目前没有一个统一的评判收敛性的标准。针对 InfoCatGAN 和 C-InfoGAN 两种模型，本文分别用条件熵损失（即判别器输出的概率分布对应的熵）以及互信息损失（实际采用交叉熵估计，详见3.2节）来作为模型收敛的佐证，见图 4.4。



(a) Entropy loss of CatGAN and InfoCatGAN

(b) Mutual information loss of InfoGAN and C-InfoGAN

图 4.4 模型在 MNIST 上的收敛速度

Figure 4.4 Convergence speed on MNIST

4.4　本章小结

本章分别给出了 InfoCatGAN 和 C-InfoGAN 在两个数据集 MNIST 和 FashionMNIST 上的实验结果。从结果来看，InfoCatGAN 相比于 CatGAN 在图片生成质量上有很大的提升，分类准确率在多数情况下略优于 CatGAN。该模型对正则化系数较为敏感，并且可以通过调节正则化系数达到生成质量和分类准确率之间的折中。C-InfoGAN 具有良好的生成质量和可观的分类准确率，并且可以通过隐变量控制图片的局部细节。值得一提的是，两个模型均能通过隐变量控制生成图片的类别，这对数据增强具有很大的意义。

第 5 章　总结与展望

5.1　全文总结

本文主要研究了两种基于生成对抗网络的分类模型：C-InfoGAN 和 InfoCatGAN。在科技飞速发展的现代，大量未处理的原始数据正等待人们去研究，去挖掘，而数据预处理和数据标注是非常耗费精力的，所以如何无监督处理和挖掘这些数据中的价值已成为当下亟待解决的问题。本文研究的基于 GAN 的分类模型可以无监督或半监督地对数据分类，同时还能生成指定类别的数据样本，具有相当程度的研究意义。现将全文内容总结如下：

第一章作为绪论，介绍了当下人工智能、深度学习领域的快速发展，奠定了本文的研究背景。接着介绍了生成对抗网络及其在各种场景下的应用，引出分类问题。最后介绍了 GAN 在分类领域的研究现状，提出本文的主要工作并给出文章结构。

第二章介绍了一些预备知识，包括生成式模型和判别式模型的区别，生成对抗网络基本思想和数学模型，详细介绍了两篇前人的工作并以此阐述了使用生成对抗网络做分类的理论推导和实际方法。

第三章介绍了本文的主要工作：C-InfoGAN 和 InfoCatGAN。详细分析了第二章提出的两种模型的特点并有针对性地提出改进方案，将互信息的物理意义与当前损失函数结合，分别给出了无监督和半监督条件下的损失函数推导、具体训练方法和模型结构图，得到的损失函数具有一定的可解释性。

第四章给出了模型在 MNIST 和 FashionMNIST 数据集上的评估结果。首先给出了实验设定和实现细节以并详细介绍了评估方法，接着给出了评价指标。然后分别给出了两个模型在两个数据集上的结果，最后给出了收敛速度分析。

本文的研究兼顾理论与实践，既有问题的形式化定义和损失函数的理论推导，也给出了详实的实验细节和实验步骤。本文将信息论中的互信息与生成对抗网络联系起来，从而得到了生成效果和分类性能的提升，是对信息论与深度学习交叉结合的一次探索。从本文的结果来看，信息论在深度学习领域具有值得探索的价值。

5.2 未来展望

限于本文作者的能力，时间和精力，本文所做的研究工作虽有一定贡献，但仍存在一些值得继续探索和深入挖掘的方向，具体如下：

- 两个模型的分类准确率均较低，如何进一步提高准确率仍需要研究；
- 本文仅在 CatGAN 的基础上添加了互信息正则，就能够显著提升生成效果，可见互信息在其中起了重要作用，其中相关原理尚待探索；
- 在 InfoCatGAN 模型中，互信息正则项仅体现在隐变量和生成数据之间，而没有对真实数据和判别器做相应的设计，对于判别器一方，原则上也需要设计相应的正则项；
- 本文在实验中均假设训练数据集是类别均匀的，对于非均衡数据集的研究仍是需要解决的问题。

参考文献

[1] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8):1798-1828.

[2] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. nature, 2015, 521(7553): 436-444.

[3] SMOLENSKY P. Information processing in dynamical systems: Foundations of harmony theory[R]. Colorado Univ at Boulder Dept of Computer Science, 1986.

[4] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7):1527-1554.

[5] SALAKHUTDINOV R, HINTON G. Deep boltzmann machines[C]//Artificial intelligence and statistics. 2009: 448-455.

[6] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.

[7] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.

[8] WANG X, YU K, WU S, et al. Esrgan: Enhanced super-resolution generative adversarial networks[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 0-0.

[9] JOLICOEUR-MARTINEAU A. The relativistic discriminator: a key element missing from standard gan[J]. arXiv preprint arXiv:1807.00734, 2018.

[10] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.

[11] YUAN Y, LIU S, ZHANG J, et al. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018: 701-710.

[12] DING Z, LIU X Y, YIN M, et al. Tgan: Deep tensor generative adversarial nets for large image generation[J]. arXiv preprint arXiv:1901.09953, 2019.

[13] TRAN L, YIN X, LIU X. Disentangled representation learning gan for pose-invariant face recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1415-1424.

[14] HUANG R, ZHANG S, LI T, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2439-2448.

[15] MA L, JIA X, SUN Q, et al. Pose guided person image generation[C]//Advances in Neural Information Processing Systems. 2017: 406-416.

[16] BAO J, CHEN D, WEN F, et al. Cvae-gan: fine-grained image generation through asymmetric training[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2745-2754.

[17] DONG H, YU S, WU C, et al. Semantic image synthesis via adversarial learning [C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 5706-5714.

[18] WU J, ZHANG C, XUE T, et al. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling[C]//Advances in neural information processing systems. 2016: 82-90.

[19] IM D J, KIM C D, JIANG H, et al. Generating images with recurrent adversarial networks[J]. arXiv preprint arXiv:1602.05110, 2016.

[20] YANG J, KANNAN A, BATRA D, et al. Lr-gan: Layered recursive generative adversarial networks for image generation[J]. arXiv preprint arXiv:1703.01560, 2017.

[21] WANG X, SHRIVASTAVA A, GUPTA A. A-fast-rcnn: Hard positive generation via adversary for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2606-2615.

[22] EHSANI K, MOTTAGHI R, FARHADI A. Segan: Segmenting and generating the invisible[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6144-6153.

[23] LI J, LIANG X, WEI Y, et al. Perceptual generative adversarial networks for small object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1222-1230.

[24] BAI Y, ZHANG Y, DING M, et al. Sod-mtgan: Small object detection via multi-task generative adversarial network[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 206-221.

[25] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[26] TAIGMAN Y, YANG M, RANZATO M, et al. Deepface: Closing the gap to human-

level performance in face verification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1701-1708.

[27] XU L, NEUFELD J, LARSON B, et al. Maximum margin clustering[C]//Advances in neural information processing systems. 2005: 1537-1544.

[28] KRAUSE A, PERONA P, GOMES R G. Discriminative clustering by regularized information maximization[C]//Advances in neural information processing systems. 2010: 775-783.

[29] SPRINGENBERG J T. Unsupervised and semi-supervised learning with categorical generative adversarial networks[J]. arXiv preprint arXiv:1511.06390, 2015.

[30] GOODFELLOW I, MIRZA M, COURVILLE A, et al. Multi-prediction deep boltzmann machines[C]//Advances in Neural Information Processing Systems. 2013: 548-556.

[31] BENGIO Y, LAUFER E, ALAIN G, et al. Deep generative stochastic networks trainable by backprop[C]//International Conference on Machine Learning. 2014: 226-234.

[32] KINGMA D P, MOHAMED S, REZENDE D J, et al. Semi-supervised learning with deep generative models[C]//Advances in neural information processing systems. 2014: 3581-3589.

[33] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786):504-507.

[34] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th international conference on Machine learning. 2008: 1096-1103.

[35] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4):541-551.

[36] KRIZHEVSKY A, HINTON G, et al. Learning multiple layers of features from tiny images[J]. 2009.

[37] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training gans[C]//Advances in neural information processing systems. 2016: 2234-2242.

[38] ODENA A. Semi-supervised learning with generative adversarial networks[J]. arXiv preprint arXiv:1606.01583, 2016.

[39] DAI Z, YANG Z, YANG F, et al. Good semi-supervised learning that requires a bad gan[C]//Advances in neural information processing systems. 2017: 6510-6520.

[40] CHONGXUAN L, XU T, ZHU J, et al. Triple generative adversarial nets[C]//Advances in neural information processing systems. 2017: 4088-4098.

[41] WU S, DENG G, LI J, et al. Enhancing triplegan for semi-supervised conditional instance synthesis and classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 10091-10100.

[42] CHEN X, DUAN Y, HOUTHOOFT R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets[C]//Advances in neural information processing systems. 2016: 2172-2180.

[43] NG A Y, JORDAN M I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes[C]//Advances in neural information processing systems. 2002: 841-848.

[44] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[45] ODENA A, OLAH C, SHLENS J. Conditional image synthesis with auxiliary classifier gans[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 2642-2651.

[46] MIYATO T, KOYAMA M. cgans with projection discriminator[J]. arXiv preprint arXiv:1802.05637, 2018.

[47] COVER T M, THOMAS J A. Elements of information theory[M]. John Wiley & Sons, 2012.

[48] BRIDLE J S, HEADING A J, MACKAY D J. Unsupervised classifiers, mutual information and'phantom targets[C]//Advances in neural information processing systems. 1992: 1096-1101.

[49] BARBER D, AGAKOV F V. Kernelized infomax clustering[C]//Advances in neural information processing systems. 2006: 17-24.

[50] POOLE B, OZAIR S, OORD A V D, et al. On variational bounds of mutual information[J]. arXiv preprint arXiv:1905.06922, 2019.

[51] GRANDVALET Y, BENGIO Y. Semi-supervised learning by entropy minimization [C]//Advances in neural information processing systems. 2005: 529-536.

[52] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[53] DUMOULIN V, BELGHAZI I, POOLE B, et al. Adversarially learned inference[J]. arXiv preprint arXiv:1606.00704, 2016.

[54] ZHANG S, FENG X, JI Y, et al. The cramér-infogan and partial inverse filter system for unsupervised image classification[C]//2018 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, 2018: 348-353.

[55] SPURR A, AKSAN E, HILLIGES O. Guiding infogan with semi-supervision[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2017: 119-134.

[56] XIAO H, RASUL K, VOLLGRAF R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms[EB/OL]. (2017-08-28).

[57] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[C]//Advances in neural information processing systems. 2017: 6626-6637.

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历

胡兵兵，安徽马鞍山人，中国科学院上海微系统与信息技术研究所硕士研究生。

已发表 (或正式接受) 的学术论文:

1. **胡兵兵**，唐华，吴幼龙. 基于互信息约束的生成对抗网络分类模型研究，中国科学院大学学报，2020. (尚在评审)

致　谢

不知不觉又过了三个年头，依稀记得三年前我刚来到学校的时候还对硕士生涯充满着未知，不知道这三年又会是什么样的生活。如今回过头来发现，越是学习，越是发现自己不懂的东西很多。犹记得在学习课程时期，经常熬夜到凌晨写作业，做实验的情景；犹记得每天读论文，每周开组会，上台做报告情景；犹记得某些问题弄不懂，抓耳挠腮和同学讨论，请教老师的情景。这些经历都是我宝贵的回忆，以后也会在我前行的道路上给予我动力。

首先我要感谢我的导师吴幼龙。作为一名年轻的教授，他具有严谨的学术态度、专业的知识素养、灵活的思维方式、以及浓烈的科研热情。更难能可贵的是，他平易近人，丝毫没有老师的架子，与学生相处亦师亦友。在求学途中，我的导师总是孜孜不倦、不厌其烦地教导我，经常与我讨论问题到很晚，无论是学习还是生活上，都给予了我莫大的关怀。在完成毕业论文期间，我更是频繁地向他请教问题，讨论实验方案，汇报论文进展，他也非常耐心地为我解答，和我讨论，帮我修改论文。我在此向他表示由衷的感谢！我还要感谢我的同窗们：汪科、马莹莹，以及同组的学弟学妹陈家慧、唐华，他们陪我一起度过了求学的时光，我们经常一起讨论学习和生活上的各种问题。我还要感谢我的学长学姐们：黄曦、高欣、边思梦，他们在我刚来学校的那段时间，给了我很多的帮助，让我能快速地融入学校，适应研究生阶段的学习模式。我也要感谢我的父母，是他们给了我求学的机会，给我提供了生活的基础。此外，我还要感谢我的女朋友尚玉静，她一直默默支持我，鼓励我，陪伴我。都说每个成功男人的背后都有一个伟大的女人，我可能不够成功，她可能不够伟大，但我觉得我们能走到一起这件事就是我最大的成功。

最后，感谢审阅论文的老师们，感谢百忙之中出席答辩的老师们！我在此向所有关心、支持和帮过我的人表示由衷的感谢！