自编码神经网络理论及应用综述

袁非牛[1),2)] 章 琳[1),3)] 史劲亭[1),4)] 夏 雪[1)] 李 钢[1),5)]

[1)](江西财经大学信息管理学院 南昌 330032)

[2)](上海师范大学信息与机电工程学院 上海 201418)

[3)](江西科技师范大学数学与计算机科学学院 南昌 330038)

[4)](江西农业大学职业师范技术学院 南昌 330045)

[5)](宜春学院数学与计算机科学学院 江西 宜春 336000)

摘 要 自编码器是深度学习中的一种非常重要的无监督学习方法,能够从大量无标签的数据中自动学习,得到蕴含在数据中的有效特征.因此,自编码方法近年来受到了广泛的关注,已成功应用于很多领域,例如数据分类、模式识别、异常检测、数据生成等.该文对传统自编码基础理论、自编码方法、改进技术以及应用领域进行了比较全面的综述.首先,该文介绍传统自编码基础理论与实现方法,分析自编码器的一般处理框架.然后,讨论现有各种改进的自编码器,分析这些方法的创新点、所要达成的目的和可能存在的问题.随后,该文介绍自编码器的实际应用领域,分析这些领域的代表性自编码算法,并详细地分析、比较和总结这些方法的特点.最后,总结现有方法存在的问题,并探讨了自编码器的将来发展趋势和可能挑战.

关键词 自编码器;深度学习;无监督学习;特征学习;约束

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.00203

Theories and Applications of Auto-Encoder Neural Networks: A Literature Survey

YUAN Fei-Niu[1),2)] ZHANG Lin[1),3)] SHI Jin-Ting[1),4)] XIA Xue[1)] LI Gang[1),5)]

[1)](*School of Information Technology, Jiangxi University of Finance and Economics, Nanchang* 330032)

[2)](*College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai* 201418)

[3)](*School of Mathematics and Computer Science, Jiangxi Science and Technology Normal University, Nanchang* 330038)

[4)](*Vocational School of Teachers and Technology, Jiangxi Agricultural University, Nanchang* 330045)

[5)](*College of Mathematics and Computational Science, Yichun University, Yichun, Jiangxi* 336000)

Abstract An auto-encoder is one of unsupervised learning methods in the deep learning community. It can automatically learn effective and robust features from a large amount of unlabeled data. It has been received a lot of attention in recent years. Therefore, auto-encoders and their variants have successfully been applied in a wide range of fields, such as image classification, pattern recognition, anomaly detection, and data generation. To provide researchers a quick overview of auto-encoders, this paper makes a full survey on the basic theory of traditional auto-encoders, related algorithms, improved techniques and several applications in detail. We first introduce the fundamental theory and common implementations of auto-encoder methods, and then we focus on analyzing the general processing framework of auto-encoder methods. A general auto-encoder method usually contains an encoding structure and a decoding one. The encoding structure is

收稿日期:2017-12-21;在线出版日期:2018-09-21.本课题得到国家自然科学基金(61862029)、江西省高校科技落地计划(KJLD12066)和江西省教育厅科技项目(GJJ170317)资助.**袁非牛**,男,1976 年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为图像处理、模式识别、三维可视化.E-mail:yfn@ustc.edu.**章 琳**(通信作者),女,1982 年生,博士研究生,副教授,主要研究方向为图像处理、模式识别.E-mail:zymm_nc@163.com.**史劲亭**,女,1982 年生,博士研究生,讲师,主要研究方向为图像处理、模式识别.**夏 雪**,女,1990 年生,博士研究生,主要研究方向为图像处理、模式识别.**李 钢**,男,1980 年生,博士研究生,讲师,主要研究方向为图像处理、模式识别.

often a contracting path that provides abundant context information for each pixel, while the decoding one is just an expanding path for localization information. Then, we discuss traditional auto-encoder methods and a series of improved methods, and analyze the innovation, motivation and existing problems of these methods. Specifically, we successively introduce some representative methods of improved auto-encoder methods, and we also point out the advantages and drawbacks of these auto-encoder methods. These methods analyzed in this paper mainly include denoising auto-encoders, marginalized denoising auto-encoders, sparse auto-encoders, contractive auto-encoders, saturating auto-encoders, convolutional auto-encoders, transforming auto-encoders, and other variants of auto-encoders. In addition, we specially discuss two kinds of marginalized auto-encoder methods, which are marginalized denoising auto-encoders for domain adaption and marginalized denoising auto-encoders for nonlinear representation, respectively. Afterwards, we introduce the major application fields of these methods based on auto-encoders, and then we also analyze the representative auto-encoder methods in each application field and discuss their advantages and drawbacks. In addition, we compare these methods with each other in each application field, and summarize the strong and weak points of these methods. The applications of auto-encoder methods mainly include data classification, anomaly detection, pattern recognition, data generation, product recommendation and other fields. Overall, the most important application of auto-encoder methods is the whole image classification, such as ImageNet Large Scale Visual Recognition Competition (ILSVRC). We also introduce another important application that is all kinds of object and event detections, so we briefly describe anomaly detection and fault diagnosis using convolutional auto-encoders. This paper also discusses dense classification of each pixel, which is also known as image segmentation. Finally, after we point out some issues commonly existing in auto-encoder methods, we discuss the possible future developing trends and challenges of auto-encoder methods. These issues existing in auto-encoder methods include gradient diffusion that limits the depth of auto-encoders, hyper-parameters that are hard to be set, classical overfitting problems existing in all neural networks, and so on. Aiming at overcoming these challenging issues, we summarize possible solutions, which may be investigation of training methods with small data sets, automatic learning methods of hyper-parameters, supervised generative models of variational auto-encoders, novel sparse theory for auto-encoders and so on.

Keywords auto-encoder; deep learning; unsupervised learning; feature learning; constraint

1 引 言

近年来,深度学习得到了史无前例的关注和发展,其最大的优势[1]在于能够自动学习数据中好的特征,因此几乎被广泛地应用于所有与特征提取相关的研究领域,比如图像分类、模式识别、异常检测等.为什么自动学习特征如此重要?我们知道,在传统机器方法中,为了能够很好地完成模式识别任务,特征提取是其中非常关键的步骤,但是特征提取方法的设计受到很多因素的影响,需要非常丰富的工程和领域相关知识,这严重限制了研究领域的发展,而深度学习所具备的性能很好地解决了这个问题.

深度学习的概念最早是由 Hinton 等人[2]提出的,他们证明了相比浅层神经网络,多层神经网络具有更加优秀的特征学习能力;同时当网络参数通过预训练过程被初始化到一个比较合适的值时,网络训练非常容易完成.根据学习方式的不同,深度学习可分为有监督和无监督学习.目前用的最多的是有监督学习,有监督学习是指送入深度学习网络中进行学习的不仅是数据,还有与数据相对应的标签,然后通过反向传播算法和优化算法来最小化实际输出与实际标签之间的误差来调整网络参数.在实际的深度学习网络中,可能有成千上万的参数需要学习,因此就需要非常大量的带有标签的训练数据.但是,标签是通过人工方式在进行数据学习之前就标记好

的.可想而知,如果数据量非常大的情况下,对数据打标签将是非常耗时耗力的工作.此时,自然的想法就是需要一个可以直接处理没有标签的数据的网络,这就促使了无监督学习的产生.无监督学习与有监督学习是相对的,即送入深度学习网络的只有数据本身,没有与数据相对应的标签,其主要目的是预训练一个能够用于其他任务的模型.一般而言,在人类学习的早期都以有监督学习为主,而在学习能力得到一定提升之后,自学成为人类学习知识的主要途径,这其实是一种无监督学习,与人类大脑的思维方式更加接近[3].

无监督学习方法中,比较著名的深度学习方法有受限波尔兹曼机[4-5]、自编码器[6-7]以及生成式对抗网络[8].受限玻尔兹曼机是一种随机神经网络模型,只包括可见层和隐藏层,和普通前馈神经网络一样,同层神经元之间无连接,相邻层神经元之间全连接,其中的隐藏层通常可以看作特征提取层.受限玻尔兹曼机得到真正的关注是在 2006 年,Hinton 等人[9]提出将受限玻尔兹曼机堆叠形成一个深度信念网络,通过对受限玻尔兹曼机进行逐层训练来完成整个网络的训练.目前,受限玻尔兹曼机主要用于两方面:特征提取和预训练,用于网络参数初始化.生成式对抗网络由一个生成网络和一个判别网络组成,生成网络用于数据生成,判别网络用于判别生成数据的真假,目前最广泛的应用是各种复杂数据的生成.

传统自编码器(Auto-Encoder)的概念最早来自于 Rumelhart 等人[6]在《Nature》上发表的论文.随后,Bourlard 等人[7]对其进行了详细的阐述.自编码器具有重建过程简单、可堆叠多层、以神经科学为支撑点的优点①.近 10 年来很多改进版本被相继提出,并且被广泛用于各种研究领域,取得了令人瞩目的成绩,其中应用非常成功的有图像分类[10-20]、视频异常检测[21-26]、模式识别[27-29]、数据生成[30-33]等.通过文献阅读和梳理,我们整理了自编码器发展历程,如表 1 所示.可以发现,自编码器在近 5 年内得到高速的发展.这是因为在自编码器发展早期,理论研究占主要地位,因此新型自编码器的提出比较缓慢,后期由于理论基础的成熟,各种针对研究领域的自编码器被相继提出,并取得了令人满意的效果.目前,自编码器和玻尔兹曼机的用法非常类似.自编码器由于内部层数不能太深,因此单个自编码器通常被逐个训练,然后堆叠多个自编码器的编码层,以完成深度学习的训练过程.

表 1　自编码器发展历程

自编码器名称	提出年份
传统自编码器	1986
降噪自编码器	2008
稀疏自编码器	2011
收缩自编码器	2011
卷积自编码器	2011
变换自编码器	2011
领域适应性边缘降噪自编码器	2012
k-稀疏自编码器	2013
饱和自编码器	2013
非线性表示边缘降噪自编码器	2014
高阶收缩自编码器	2014
变分自编码器	2014
协同局部自编码器	2014
张量自编码器	2014
条件变分自编码器	2015
变分公平自编码器	2015
区分自编码器	2015
局部约束稀疏自编码器	2015
协同稳定非局部自编码器	2016
损失变分自编码器	2016
大边缘自编码器	2017
信息最大化变分自编码器	2017
最小二乘变分自编码器	2017
循环通道变分自编码器	2017
多阶段变分自编码器	2017

从现有文献来看,目前国内外还没有对自编码器非常系统的综述类文献.为此,本文对各种自编码器的理论进行详细梳理、介绍和分析,同时结合各个应用领域的最新研究文献,对目前自编码器的发展进行非常全面的归纳和总结.最后,对目前自编码器还存在的问题进行分析,提出一些以后可能的发展方向和趋势.

2　传统自编码器

一般来说,传统自编码器主要包括编码阶段和解码阶段,且结构是对称的,即如果有多个隐层时,编码阶段的隐层数量与解码阶段相同.其结构模型如图 1 所示.传统自编码器的目的就是要在输出层重建输入数据,最完美的情况就是输出信号 $\boldsymbol{y}$ 与输入信号 $\boldsymbol{x}$ 完全一样.按照图 1 所示结构,传统自编码器的编码、解码过程可描述为[34]:

$$\text{编码过程：}\quad \boldsymbol{h}_1 = \sigma_e(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) \tag{1}$$

$$\text{解码过程：}\quad \boldsymbol{y} = \sigma_d(\boldsymbol{W}_2\boldsymbol{h}_1 + \boldsymbol{b}_2) \tag{2}$$

其中 $\boldsymbol{W}_1$,$\boldsymbol{b}_1$ 为编码权重和偏置,$\boldsymbol{W}_2$,$\boldsymbol{b}_2$ 为解码权重和偏置,σ_e 为非线性变换,目前比较常用的有 sigmoid、

① 一篇文章带你进入无监督学习:从基本概念到 4 种实现模型. https://www.jiqizhixin.com/articles/2016-10-30-3, 2016.10.30.

tanh、Relu 等，σ_d 可以是与编码过程中相同的非线性变换或者仿射变换. 因此，传统自编码器的损失函数就是要最小化 $\boldsymbol{y}$ 和 $\boldsymbol{x}$ 之间的误差：

$$J(\boldsymbol{W},\boldsymbol{b})=\sum(L(\boldsymbol{x},\boldsymbol{y}))=\sum\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \quad (3)$$

编码阶段可以看成是通过一种确定性的映射将输入信号转换为隐层表达，而解码阶段则是尽量将隐层表达重新映射为输入信号. 传统自编码器中的参数，即权重和偏置，通过最小化目标函数来学习获得. 除了式(3)所给出的均方误差，损失函数还可以选择交叉熵，具体表示为

$$J(\boldsymbol{W},\boldsymbol{b})=\sum(L(\boldsymbol{x},\boldsymbol{y}))=-\sum_{i=1}^{n}(\boldsymbol{x}_i\log(\boldsymbol{y}_i)+(1-\boldsymbol{x}_i)\log(1-\boldsymbol{y}_i)) \quad (4)$$

我们发现，以上编解码过程没有涉及输入数据的标签信息. 因此，传统自编码器被看作是一种无监督学习方法. 针对图 1 所示的结构，传统自编码器的隐层表达有 3 种不同的形式[35]：压缩结构、稀疏结构和等维结构. 当输入层神经元数量大于隐藏层时，称为压缩结构. 反之，当输入层神经元数量小于隐藏层时，称为稀疏结构. 而如果输入层与隐藏层神经元数量相等时则为等维结构.

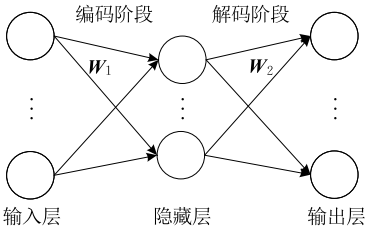

图 1 传统自编码器网络结构图

在传统自编码器中，编码和解码阶段的权重是分别单独训练的，并没有什么实际的联系. 但是如果令 $\boldsymbol{W}_2=\boldsymbol{W}_1^{\mathrm{T}}$，即编码和解码阶段使用相同的权重，这种自编码器被称为绑定权重自编码器(Tied weight Auto-Encoder，TAE).

对于传统自编码器，还可以通过在损失函数上增加一个权重衰减项来控制权重的减小程度，我们称之为传统正则自编码器，损失函数为

$$J_{\mathrm{CoAE}}(\boldsymbol{W},\boldsymbol{b})=\sum(L(\boldsymbol{x},\boldsymbol{y}))+\lambda\|\boldsymbol{W}\|_2^2 \quad (5)$$

参数 λ 用来控制正则化的强度，一般取 0～1 之间的值. 权重衰减项的增加能有效抑制静态噪声对目标和权重向量中不相关成分的影响，显著提升网络的泛化能力，并有效避免过拟合现象的产生[36].

3 改进的自编码器

传统自编码器的目的是使输出与输入尽量相同，这完全可以通过学习两个恒等函数来完成，但是这样的变换没有任何意义，因为我们真正关心的是隐层表达，而不是实际输出. 因此，针对自编码器的很多改进方法都是对隐层表达增加一定的约束，迫使隐层表达与输入不同. 如果此时模型还可以重建输入信号，那么说明隐层表达足以表示输入信号，这个隐层表达就是通过模型自动学习出来的有效特征. 接下来，我们将对几种比较典型的改进自编码器进行详细介绍.

3.1 降噪自编码器(Denoising Auto-Encoders)

前面提到，在自编码器中我们真正关心的其实是隐层表达，那么到底什么才是好的表达？Vincent 等人[37]认为一个好的表达应该能够捕获输入信号的稳定结构，具有一定的鲁棒性，同时对重建信号是有用的. 因此，Vincent 等人[37]从鲁棒性着手，于 2008 年提出了降噪自编码器.

降噪自编码器的提出是受到一个现象的启发，那就是对于部分被遮挡或损坏的图像，人类仍然可以进行准确的识别. 因此，降噪自编码器的主要研究目标是：隐层表达对被局部损坏的输入信号的鲁棒性. 也就是说，如果一个模型具有足够的鲁棒性，那么被局部损坏的输入在隐层上的表达应该与没有被破坏的干净输入几乎相同，而利用这个隐层表达就完全可以重建干净的输入信号.

因此，降噪自编码器通过对干净输入信号人为加入一些噪声，使干净信号受到局部损坏，产生与它对应的一个损坏信号，然后将这个损坏信号送入传统自编码器，使其尽量重建一个与干净输入相同的输出. 其中损坏输入信号 $\tilde{\boldsymbol{x}}$ 通过一个随机映射从干净输入 $\boldsymbol{x}$ 获得：$\tilde{\boldsymbol{x}}\sim q_D(\tilde{\boldsymbol{x}}\,|\boldsymbol{x})$. 降噪自编码器的基本结构如图 2 所示.

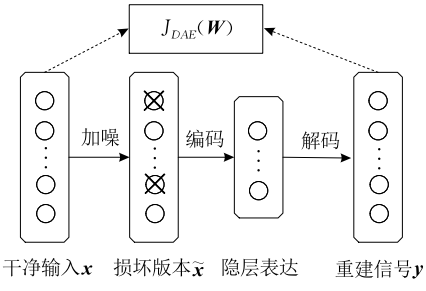

图 2 降噪自编码器网络结构图

为了使重建信号与干净信号的误差尽可能小，降噪自编码器的目标就是最小化损失函数：

$$J_{DAE}(\boldsymbol{W}) = \sum E_{\tilde{\boldsymbol{x}}\sim q_D(\tilde{\boldsymbol{x}}|\boldsymbol{x})}[L(\boldsymbol{x},\boldsymbol{y})] \tag{6}$$

紧接着，在 2010 年 Vincent 等人[38]采用和深度信念网络中一样的方法，将降噪自编码器进行堆叠，构造了一个堆叠降噪自编码网络并将其用于图像分类. 其基本结构如图 3 所示.

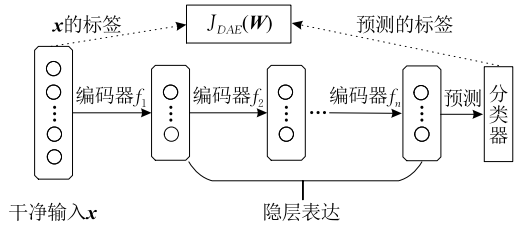

图 3　堆叠降噪自编码器网络结构图

编码器 f_1 到 f_n 分别对应的是预先训练好的降噪自编码器 D_1 到 D_n 的编码函数. 需要特别说明的是，该网络采用降噪自编码器 D_1 对应的编码函数 f_1 对干净输入 $\boldsymbol{x}$ 进行非线性变换，再把变换结果作为 D_2 的输入信号去训练 D_2，以获得 D_2 的编码函数 f_2，并以此类推. 最后利用 $\boldsymbol{x}$ 的真实标签和预测获得的标签去进行有监督学习，对整个网络的参数进行进一步的微调.

综上所述，降噪自编码器通过对输入信号人为地进行损坏，主要是为了达到两个目的，首先是为了避免使隐层单元学习一个传统自编码器中没有实际意义的恒等函数，其次就是为了使隐层单元可以学习到一个更加具有鲁棒性的特征表达. 因此降噪自编码器最大的优点在于，重建信号对输入中的噪声具有一定的鲁棒性，而最大的缺陷在于每次进行网络训练之前，都需要对干净输入信号人为地添加噪声，以获得它的损坏信号，这无形中就增加了该模型的处理时间.

3.2　边缘降噪自编码器(Marginalized Denoising Auto-Encoders)

边缘降噪自编码器最初是由 Chen 等人[39]提出的，其主要思想是在降噪自编码器的基础上，对噪声干扰进行边缘化处理，因此取名为边缘降噪自编码器. Chen 针对线性和非线性问题，分别于 2012 年和 2014 年提出了两种边缘降噪自编码器：领域适应性边缘降噪自编码器[39](Marginalized Denoising Auto-Encoders for Domain Adaption)和非线性表示边缘降噪自编码器[40](Marginalized Denoising Auto-Encoders for Nonlinear Representation).

3.2.1　领域适应性边缘降噪自编码器

通过上一节的介绍，可以发现降噪自编码器存在两个关键缺陷：计算代价高，缺少对高维特征的可伸缩性. 而领域适应性边缘降噪自编码器的出发点就是要解决以上两个缺陷. 领域适应性降噪自编码器通过将噪声边缘化，使得在算法中不需要使用任何优化算法来学习模型中的参数. 这一改进在很大程度上提升了网络的训练速度，一般来说，相比降噪自编码器，速度可以提升两个数量级.

领域适应性边缘降噪自编码器是由一些单层降噪自编码器组合而成，因此，它的输入信号与降噪自编码器一样，是经过加噪处理的损坏信号. 如果采用平方重建误差作为损失函数，那么领域适应性边缘降噪自编码器的目标就是要将下式最小化：

$$J_{DAE}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i - \boldsymbol{W}\tilde{\boldsymbol{x}}_i\|^2 \tag{7}$$

通常为了让模型更加具有普遍性，会将实验过程重复多次，因此 Chen 等人[39]重复以上过程 m 次，每次通过添加不同的噪声，产生 m 个不同的损坏信号作为输入. 因此式(7)中的目标函数可改写为最小化整体平方重建误差：

$$J_{DAE'}(\boldsymbol{W}) = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\|\boldsymbol{x}_i - \boldsymbol{W}\tilde{\boldsymbol{x}}_{i,j}\|^2 \tag{8}$$

其中 $\tilde{\boldsymbol{x}}_{i,j}$ 为原始输入 $\boldsymbol{x}_i$ 的第 j 个损坏版本.

根据范数与矩阵之间的关系，式(8)可进一步改写为

$$J_{DAE'}(\boldsymbol{W}) = \frac{1}{mn}tr[(\bar{\boldsymbol{X}} - \boldsymbol{W}\tilde{\boldsymbol{X}})^{\mathrm{T}}(\bar{\boldsymbol{X}} - \boldsymbol{W}\tilde{\boldsymbol{X}})] \tag{9}$$

其中 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]$，$\bar{\boldsymbol{X}}=[\boldsymbol{X},\cdots,\boldsymbol{X}]$，$\bar{\boldsymbol{X}}$ 具有 m 个元素，$\tilde{\boldsymbol{X}}$ 是与 $\bar{\boldsymbol{X}}$ 相对应的损坏版本，$tr(\boldsymbol{A})$ 表示求矩阵 $\boldsymbol{A}$ 的迹. 然后根据普通最小二乘法，最终，最小化领域适应性边缘降噪自编码器的损失函数可转化为求解以下问题：

$$\boldsymbol{W} = \boldsymbol{P}\boldsymbol{Q}^{-1} \tag{10}$$

其中 $\boldsymbol{P}=\bar{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{\mathrm{T}}$，$\boldsymbol{Q}=\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{\mathrm{T}}$.

根据弱大数定律，当实验次数 m 足够大时，式(10)中的 $\boldsymbol{P},\boldsymbol{Q}$ 最终会分别收敛于它们的期望值，因此令 $m\to\infty$，式(10)可转化为

$$\boldsymbol{W} = E[\boldsymbol{P}]E[\boldsymbol{Q}]^{-1} \tag{11}$$

其中

$$E[\boldsymbol{P}]_{\alpha,\beta} = \boldsymbol{S}_{\alpha,\beta}q_\alpha \tag{12}$$

$$E[\boldsymbol{Q}]_{\alpha,\beta} = \begin{cases}\boldsymbol{S}_{\alpha,\beta}q_\alpha q_\beta, & \alpha\neq\beta\\ \boldsymbol{S}_{\alpha,\beta}q_\alpha, & \alpha=\beta\end{cases} \tag{13}$$

在式(12)、(13)中 $\boldsymbol{S}=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$，$q_\alpha$ 表示特征 α 不被损坏的概率，q_β 表示特征 β 不被损坏的概率.

最终,通过式(11)可以算出网络参数 $\boldsymbol{W}$. 通过推导过程,我们发现,求解网络参数的过程中没有使用任何的优化算法,仅需要遍历训练数据一次就可以根据式(12)和(13)求得 $E[\boldsymbol{P}]$和 $E[\boldsymbol{Q}]$的值. 这就是领域适应性边缘降噪自编码器能够在训练时间上取得巨大提升的根本原因.

领域适应性边缘降噪自编码器的提出有效证明了线性降噪器可以作为学习特征表达的基本模块. 将线性降噪器作为基本模块可以带来两方面的优势,首先线性表示能够显著地简化参数估计. 其次在保证分类性能的基础上,极大缩短了训练时间.

综上所述,领域适应性边缘降噪自编码器具有以下优点:具有比较强的特征学习能力;训练速度快;更少的中间参数,更快的模型选择以及基于逐层训练的凸性. 但缺点在于只能用于线性表示.

3.2.2 非线性表示边缘降噪自编码器

领域适应性边缘降噪自编码能够克服降噪自编码器计算强度大,处理时间长的缺点,但是它仅适用于线性表示. 因此为了使边缘降噪自编码能被用于非线性表示,Chen 等人[40]在此基础上进行了扩展,于 2014 年提出了非线性表示边缘降噪自编码器. 该方法与前者最大的区别在于,在进行边缘化处理时,改为利用降噪自编码器损失函数的泰勒展开式来近似表示其期望损失函数.

除了式(7)的表示方法,降噪自编码器的目标函数还可以写成一种更加普遍的形式:

$$J_{DAE'}(\boldsymbol{W})=\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}L(x_i,f_\theta(\tilde{\boldsymbol{x}}_{i,j})) \tag{14}$$

其中 $\tilde{\boldsymbol{x}}_{i,j}$为原始输入 $\boldsymbol{x}_i$的第 j 个损坏版本,L 可以是平方重建误差损失函数,或者交叉熵损失函数.

与领域适应性边缘去噪自编码器的策略一样,这里也采用了弱大数定律,利用无穷多个损坏数据将优化问题转化为期望问题. 因此在满足损失分布 $p(\tilde{\boldsymbol{x}}|\boldsymbol{x})$的条件下,式(14)可表示为期望问题:

$$J_{DAE'}(\boldsymbol{W})=\frac{1}{n}\sum_{i=1}^{n}E_{p(\tilde{\boldsymbol{x}}_i|\boldsymbol{x}_i)}[L(\boldsymbol{x}_i,f_\theta(\tilde{\boldsymbol{x}}_i))] \tag{15}$$

在 $E_{p(\tilde{\boldsymbol{x}}|\boldsymbol{x})}[\tilde{\boldsymbol{x}}]=\mu_x$上利用泰勒展开式,对式(15)中的损失函数 L 进行展开并加上期望可得:

$$\begin{aligned}E[L(\boldsymbol{x},f_\theta(\tilde{\boldsymbol{x}}))]\approx{}&E[L(\boldsymbol{x},f_\theta(\mu_x))]+\\&E[(\tilde{\boldsymbol{x}}-\mu_x)^{\mathrm{T}}\nabla_{\tilde{x}}L]+\\&E\left[\frac{1}{2}(\tilde{\boldsymbol{x}}-\mu_x)^{\mathrm{T}}\nabla_{\tilde{x}}^2L(\tilde{\boldsymbol{x}}-\mu_x)\right]\end{aligned} \tag{16}$$

其中,$\nabla_{\tilde{x}}L$ 和 $\nabla_{\tilde{x}}^2L$ 分别是损失函数 L 关于 $\tilde{\boldsymbol{x}}$ 的一阶和二阶导数. 由于 $E_{p(\tilde{\boldsymbol{x}}|\boldsymbol{x})}[\tilde{\boldsymbol{x}}]=\mu_x$,因此 $E[(\tilde{\boldsymbol{x}}-\mu_x)^{\mathrm{T}}\nabla_{\tilde{x}}L]=0$,同时令 $\boldsymbol{\Sigma}_x=E[(\tilde{\boldsymbol{x}}-\mu_x)(\tilde{\boldsymbol{x}}-\mu_x)^{\mathrm{T}}]$,式(16)可进一步改写为

$$E[\ell(\boldsymbol{x},f_\theta(\tilde{\boldsymbol{x}}))]\approx L(\boldsymbol{x},f_\theta(\mu_x))+\frac{1}{2}tr(\boldsymbol{\Sigma}_x\nabla_{\tilde{x}}^2L) \tag{17}$$

为了方便求解,给模型加入一个假设:噪声被独立添加到输入信号的每一维,则 $\boldsymbol{\Sigma}_x$是一个对角矩阵. 问题就简化为只需要计算黑塞矩阵 $\nabla_{\tilde{x}}^2L$ 的对角项,这大大减少了模型的计算开销,尤其是对高维数据.

黑塞矩阵对角线上的第 d 项可表示为

$$\frac{\partial^2L}{\partial\tilde{\boldsymbol{x}}_d^2}=\left(\frac{\partial\boldsymbol{z}}{\partial\tilde{\boldsymbol{x}}_d}\right)^{\mathrm{T}}\frac{\partial^2L}{\partial\boldsymbol{z}^2}\frac{\partial\boldsymbol{z}}{\partial\tilde{\boldsymbol{x}}_d}+\left(\frac{\partial L}{\partial\boldsymbol{z}}\right)^{\mathrm{T}}\frac{\partial^2\boldsymbol{z}}{\partial\tilde{\boldsymbol{x}}_d^2} \tag{18}$$

其中 $\boldsymbol{z}$ 为 $\tilde{\boldsymbol{x}}$ 的隐层表达.

为了进一步简化计算问题,根据 LeCun 等人[41]提出的建议,式(18)中的第 2 项可以舍弃,同时由于 $\nabla_z^2L=\dfrac{\partial^2L}{\partial\boldsymbol{z}^2}$ 通常是正定的,因此根据正定矩阵的性质,式(18)可简化为

$$\frac{\partial^2L}{\partial\tilde{\boldsymbol{x}}_d^2}\approx\sum_{h=1}^{D_h}\frac{\partial^2L}{\partial\boldsymbol{z}_h^2}\left(\frac{\partial\boldsymbol{z}_h}{\partial\tilde{\boldsymbol{x}}_d}\right)^2 \tag{19}$$

因此,非线性表示边缘降噪自编码器的目标函数为

$$J_{MDAE}(\boldsymbol{W})=L(\boldsymbol{x},f_\theta(\mu_x))+\frac{1}{2}\sum_{d=1}^{D}\sigma_{xd}^2\sum_{h=1}^{D_h}\frac{\partial^2L}{\partial\boldsymbol{z}_h^2}\left(\frac{\partial\boldsymbol{z}_h}{\partial\tilde{\boldsymbol{x}}_d}\right)^2 \tag{20}$$

其中 σ_{xd}^2是 $\boldsymbol{\Sigma}_x$对角线上第 d 项,即第 d 维输入的噪声损坏方差. 不难发现,式(20)的损失函数相当于在降噪自编码器目标函数的基础上增加了一个正则项,该正则项既考虑了重建函数对隐层表达的敏感度,同时还考虑了隐层表达对输入的损坏信号的敏感度.

3.3 稀疏自编码器(Sparse Auto-Encoders)

3.3.1 稀疏自编码器

自编码器最初提出是基于降维的思想,但是当隐层节点比输入节点多时,自编码器就会失去自动学习样本特征的能力,此时就需要对隐层节点进行一定的约束. 与降噪自编码器的出发点一样,Ng[42]认为高维而稀疏的表达是好的,因此提出对隐层节点进行一些稀疏性的限制. 稀疏自编码器就是在传统自编码器的基础上通过增加一些稀疏性约束得到的. 这个稀疏性是针对自编码器的隐层神经元而言的,通过对隐层神经元的大部分输出进行抑制使网络达到一个稀疏的效果. 根据所选激活函数的不同,神经元被抑制的概念有些许区别. 如果激活函数为 sigmoid,输出接近 0 表示被抑制,如果激活

函数为 tanh，那么神经元被抑制是其输出在−1 附近.

为了实现抑制效果，稀疏自编码器通过对隐层神经元输出的平均激活值进行约束，利用 KL 散度(KL divergence)迫使其与一个给定的稀疏值相近，并将其作为惩罚项添加到损失函数中，因此，稀疏自编码器的损失函数可表示为

$$J_{SAE}(\boldsymbol{W})=\sum(L(\boldsymbol{x},\boldsymbol{y}))+\beta\sum_{j=1}^{h}KL(\rho\,\|\,\hat{\rho}_j)\quad(21)$$

其中，$\hat{\rho}_j=\frac{1}{m}\sum_{i=1}^{m}(a_j(x_i))$，代表所有训练样本在隐层神经元 j 上的平均激活值，a_j 为隐层神经元 j 上的激活值. $\sum_{j=1}^{h}KL(\rho\,\|\,\hat{\rho}_j)=\sum_{j=1}^{h}\left(\rho\log\frac{\rho}{\hat{\rho}_j}+(1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}\right)$，$\beta$ 用于控制稀疏惩罚项的权重，可取 0～1 之间的任意值. 为了达到大部分神经元都被抑制的效果，ρ 一般取接近于 0 的值. 如果 ρ 取值 0.02，那么通过这个约束，自编码器的每个隐层神经元 j 的平均激活值都会接近于 0.02.

使用 KL 散度是因为它可以很好地度量两个不同分布之间的差异. 当 $\hat{\rho}_j=\rho$ 时，$KL(\rho\,\|\,\hat{\rho}_j)=0$；而当 $\hat{\rho}_j$ 与 ρ 差异较大时，KL 散度会呈现单调增加的规律. 因此为了使 $\hat{\rho}_j$ 与给定的 ρ 尽量相同，采用两者之间的 KL 散度作为惩罚项.

如果通过隐层神经元的稀疏表达可以完美重建输入信号，那么说明这些稀疏表达已经包含了输入信号大部分主要特征，可以看作是对输入数据的一种简单表示，这样就在保证模型重建精度的基础上，极大地降低了数据的维度，使模型的性能得到了很大的提升.

3.3.2　k-稀疏自编码器(k-Sparse Auto-Encoders)

由上一节可以知道，稀疏自编码器可以使模型通过抑制隐层神经元达到一个很好的稀疏效果，但是无法指定哪些神经元处于激活状态，哪些神经元被抑制，使其无法获得一个准确的稀疏度. 而 k-稀疏自编码器[43]针对这个问题，对稀疏自编码器进行了改进. 两者之间最本质的区别在于 k-稀疏自编码器仅使用线性激活函数，同时在隐层只保留 k 个最大激活值. 仅使用线性激活函数是为了有效提升模型的训练速度，而通过指定 k 的取值，可以避免使用 KL 散度，同时使模型获得一个准确的稀疏度. 通过以上两点改进，k-稀疏自编码器能很好的处理大数据量问题，这是稀疏自编码器无法做到的.

k-稀疏自编码器的基本思想为：(1) 在神经网络前馈阶段，首先对输入进行线性变化，计算其激活值，然后利用排序算法或者 Relu 函数选取 k 个最大的激活值，k 的设置可以看成一个规则化，防止使用过多的隐层单元去重建输入；(2) 利用隐层的稀疏表示计算输出和误差，最后使用反向传播算法对网络参数进行优化. 算法整体流程描述如下.

训练阶段：

(1) 前馈阶段，利用线性变化计算激活值

$$\boldsymbol{z}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}\quad(22)$$

(2) 选取 k 个最大激活值构成集合 Γ，其余全部设置为 0.

$$\Gamma=\mathrm{supp}_k(\boldsymbol{z})\quad(23)$$

$$\boldsymbol{z}_{(\Gamma)^c}=0\quad(24)$$

其中 $\mathrm{supp}_k(\boldsymbol{z})$ 用于产生 k 个最大激活值，该步骤被看作该算法中唯一的非线性变换.

(3) 计算输出值和重构误差：

$$\hat{\boldsymbol{x}}=\boldsymbol{W}\boldsymbol{z}+\boldsymbol{b}'\quad(25)$$

$$e=\sum\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2^2\quad(26)$$

其中 $\boldsymbol{z}$ 为第(2)步中更新的 $\boldsymbol{z}$.

(4) 反向传播重建误差，迭代更新网络参数.

测试阶段：

(1) 利用训练好的参数计算特征：

$$\boldsymbol{h}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}\quad(27)$$

(2) 选取 $\boldsymbol{h}$ 前 αk 个最大激活值，其余均设为 0.

$$\Gamma=\mathrm{supp}_{\alpha k}(\boldsymbol{h})\quad(28)$$

$$\boldsymbol{h}_{(\Gamma)^c}=0\quad(29)$$

对于 k-稀疏自编码器，有几点需要特别说明：第一，在编码和解码阶段使用的是相同的权重，也就是我们前面提到的绑定权重自编码器，据我们分析，这样做是为了加快网络的训练速度. 第二，在测试阶段中，算法选取 αk 个最大激活值来代替训练阶段的 k 个，其中 $\alpha\geq1$. 这种改进是基于 Coates 和 Ng[44]在 2011 年提出的思想：当用于分类的稀疏编码与用于训练的编码不完全相同时，反而可以获得更好的结果.

相比稀疏自编码器，k-稀疏自编码器更容易训练，编码速度更快，而且在隐层可以获取准确的稀疏度. 但是对于如何选取 k 和 α 的值没有给出具体指导方法，而且对于不同数据集，k 的不同取值对结果影响较大.

3.4　收缩自编码器(Contractive Auto-Encoders)

3.4.1　收缩自编码器

衡量一个模型的效果如何可以通过两个标准来评判：(1) 模型是否可以很好地重建输入信号；(2) 模型对输入数据一定程度下的扰动是否具有不变性.

对于第 1 点，大部分自编码器都能很好地完成，但是第 2 点就无法达到很好的效果. 然而，对于分类任务而言，第 2 个标准却更加重要，这就促使了收缩自编码器的产生. 收缩自编码器最初是由 Rifai 等人[34]在 2011 年提出的，其主要目的是为了抑制训练样本在所有方向上的扰动，通过在传统自编码器的目标函数上增加一个惩罚项来达到局部空间收缩的效果. 该惩罚项是关于输入的隐层表达的 Jacobian 矩阵的 F 范数，其目的是为了使特征空间在训练数据附近的映射达到收缩效果，具体表示为

$$\|J_f(\boldsymbol{x})\|_F^2=\sum_{ij}\left(\frac{\partial h_j(\boldsymbol{x})}{\partial \boldsymbol{x}_i}\right)^2 \tag{30}$$

根据模型评判标准 2，一个好的模型对输入的微小变化应该具有一定的鲁棒性. 也就是说，当输入信号出现微小变化，比如在输入信号中加入一点噪声，它的隐层表达应该和干净输入信号的隐层表达非常接近，不应该发生很大的变化. 这其实和优化问题中的良态、病态系统非常类似，所谓病态系统就是当输入出现非常微小的变化，其输出结果变化非常大，说明该系统对输入太敏感，根本不具有任何的实用性. 所以，收缩自编码器的出发点就是为了使其自身成为一个良态系统.

收缩自编码器通过将式(30)作为惩罚项添加到损失函数中，减少模型对输入微小变化的敏感程度，从而达到良态系统的目的. 将式(30)作为惩罚项是因为它具有如下特性：当惩罚项具有比较小的一阶导数时，说明与输入信号对应的隐层表达比较平滑，那么当输入出现一定变化时，隐层表达不会发生很大的改变，这就达到了对输入变化不敏感的目的. 因此，收缩自编码器的损失函数可表示为

$$J_{CAE}(\boldsymbol{W})=\sum(L(\boldsymbol{x},\boldsymbol{y}))+\lambda\|J_f(\boldsymbol{x})\|_F^2 \tag{31}$$

其中 λ 是用于控制惩罚项强度的超参数，可选择 0～1 之间的任意值. 损失函数的第 1 项是让重建误差尽可能小，使收缩自编码器可以尽可能地捕获到输入信号的全部信息. 而第 2 项又可以看作是收缩自编码器在尽可能的丢弃所有信息. 因此，收缩自编码器最终只会捕获到训练数据上出现的扰动信息，使模型对扰动具有不变性.

在线性编码的情况下，收缩自编码器的损失函数与传统正则自编码器完全相同，两者都通过保持小的权重达到收缩的目的. 而在非线性情况下，收缩自编码器的收缩性与稀疏自编码器的稀疏性非常类似，两者都鼓励稀疏表示，稀疏自编码器通过使其大部分隐层神经元受到抑制，隐层输出对应于激活函数的左饱和区域，而收缩自编码器通过将隐层神经元的输出推向它的饱和区域来达到收缩性. 同样的，收缩自编码器的鲁棒性与降噪自编码器也是如出一辙，两者都对输入噪声具有鲁棒性，其主要区别在于作用对象不同，降噪自编码器是针对重建信号的鲁棒性，而收缩自编码器则是针对隐层表达的鲁棒性.

3.4.2 高阶收缩自编码器(High Order Contractive Auto-Encoder)

为了将收缩自编码器对于输入微小变化的鲁棒性做进一步的提升，2014 年 Rifai 等人[45]在收缩自编码器的基础上提出了高阶收缩自编码器. 该模型对收缩自编码器的目标函数做了进一步改进，在其基础上又增加了一个二阶惩罚项，该惩罚项是关于输入的隐层表达的 Hessian 矩阵的 F 范数，具体表示为

$$\|H_f(\boldsymbol{x})\|_F^2=\left\|\frac{\partial J_f(\boldsymbol{x})}{\partial \boldsymbol{x}}\right\|_F^2 \tag{32}$$

由此可获得高阶收缩自编码器的目标函数为

$$J_{CAE+H}(\boldsymbol{W})=\sum(L(\boldsymbol{x},\boldsymbol{y}))+\lambda\|J_f(\boldsymbol{x})\|_F^2+\gamma\|H_f(\boldsymbol{x})\|_F^2 \tag{33}$$

一阶导数和重建误差的计算复杂度其实差异并不大，但是二阶导数的计算会极大地增加模型的复杂度，因此，Rifai 等人[45]提出利用随机近似方法，通过将二阶导数转化为一阶导数来降低其计算复杂度：

$$\|H_f(\boldsymbol{x})\|_F^2=\lim_{\sigma\to 0}\frac{1}{\sigma^2}\mathrm{E}[\|J_f(x)-J(x+\varepsilon)\|_F^2] \tag{34}$$

因此，高阶收缩自编码器的最终目标就是最小化以下目标函数：

$$J_{CAE+H}(\boldsymbol{W})=\sum(L(\boldsymbol{x},\boldsymbol{y}))+\lambda\|J_f(\boldsymbol{x})\|_F^2+\gamma\mathrm{E}[\|J_f(x)-J(x+\varepsilon)\|_F^2] \tag{35}$$

3.5 饱和自编码器(Saturating Auto-Encoders)

从一定角度上来说，饱和自编码器[46]和使用非线性变换的收缩自编码器具有非常类似的地方，两者都鼓励隐层神经元的输出落在激活函数的饱和区域. 不同之处在于，饱和自编码器是通过在传统自编码器的隐层神经元上引入一个正则项，使隐层神经元的激活函数至少包含一个零梯度区域，即饱和区域，而该正则项的目的就是尽量使激活值落在相应激活函数的饱和区域.

为了使激活函数至少包含一个零梯度区域，Goroshin 等人[46]提出利用分段函数 f 作为饱和自编码器的激活函数，被选择的每个分段函数都至少包含一个平坦区域，而每个激活函数都对应一个互

补函数 f_c，具体定义为

$$f_c(\boldsymbol{z}) = \inf_{\boldsymbol{z}' \in S} |\boldsymbol{z} - \boldsymbol{z}'| \tag{36}$$

其中 $S=\{\boldsymbol{z}|f'(\boldsymbol{z})=0\}$，通过最小化这个互补函数可以使 $\boldsymbol{z}$ 与 $f(\boldsymbol{z})$ 的平坦点更接近. 因此，饱和自编码器的损失函数定义为

$$\begin{aligned} J_{SATAE}(\boldsymbol{W},b) &= \sum(L(\boldsymbol{x},\boldsymbol{y})) + \alpha\sum_{i=1}^{h} f_c(\boldsymbol{W}_i^e\boldsymbol{x}+b_i^e) \\ &= \sum_{\boldsymbol{x}\in D}\frac{1}{2}\left\|\boldsymbol{x}-(\boldsymbol{W}^d f(\boldsymbol{W}^e\boldsymbol{x}+b^e)+b^d)\right\|^2 + \\ &\quad \alpha\sum_{i=1}^{h} f_c(\boldsymbol{W}_i^e\boldsymbol{x}+b_i^e) \end{aligned} \tag{37}$$

其中 h 为隐层神经元的数量，$\alpha\in[0,1]$ 是一个超参数，用于控制重建和饱和之间的权衡. 很明显，这个正则项能明确地提升不在数据流形附近的输入信号的重建误差.

3.6　卷积自编码器(Convolutional Auto-Encoders)

近年来，卷积神经网络所取得的各种优异表现，直接推动了卷积自编码器的产生. 严格上来说，卷积自编码器[47]属于传统自编码器的一个特例，它使用卷积层和池化层替代了原来的全连接层. 传统自编码器一般使用的是全连接层，对于一维信号并没有什么影响，而对于二维图像或视频信号，全连接层会损失空间信息，通过采用卷积操作，卷积自编码器能很好的保留二维信号的空间信息.

卷积自编码器与传统自编码器非常类似，其主要差别在于卷积自编码器采用卷积方式对输入信号进行线性变换，并且其权重是共享的，这点与卷积神经网络一样. 因此，重建过程就是基于隐藏编码的基本图像块的线性组合.

卷积自编码器的损失函数与传统正则自编码器一样，具体可表示为

$$J_{\mathrm{CoAE}}(\boldsymbol{W}) = \sum(L(\boldsymbol{x},\boldsymbol{y})) + \lambda\|\boldsymbol{W}\|_2^2 \tag{38}$$

3.7　变分自编码器(Variational Auto-Encoders)

3.7.1　变分自编码器

变分自编码器[48-49]是由 Kingma 等人[48]在 2014 年提出的一种模型，目前主要用于数据生成. 当作为生成模型时，首先利用一组数据训练变分自编码器，然后只需要使用变分自编码器网络的解码部分，就可以自动生成与训练数据类似的输出.

介绍变分自编码器原理之前，先介绍两个变量：$\boldsymbol{z}$ 和 $\boldsymbol{x}$. $\boldsymbol{z}$ 称为隐变量，与传统自编码器的隐层输出非常类似，$\boldsymbol{x}$ 是最后想要生成的数据. 假设有一组函数 $f(\boldsymbol{z};\theta)$ 用于由 $\boldsymbol{z}$ 产生 $\boldsymbol{x}$，每个函数由 θ 唯一的确定. 而变分自编码器的目标就是通过优化 θ，使得在采样为 $\boldsymbol{z}$ 的前提下，最大化 $\boldsymbol{x}$ 最后产生的概率 $P(\boldsymbol{x})$. 根据贝叶斯公式，$P(\boldsymbol{x})$ 可表示为

$$P(\boldsymbol{x}) = \int f(\boldsymbol{z};\theta)P(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \tag{39}$$

利用 $P(\boldsymbol{x}|\boldsymbol{z};\theta)$ 替代 $f(\boldsymbol{z};\theta)$，使 $\boldsymbol{x}$ 对 $\boldsymbol{z}$ 的依赖更加明确：

$$P(\boldsymbol{x}) = \int P(\boldsymbol{x}|\boldsymbol{z};\theta)P(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \tag{40}$$

这个替换非常必要，因为这样就可以使用优化算法使 $P(\boldsymbol{x}|\boldsymbol{z};\theta)$ 在某些采样 $\boldsymbol{z}$ 的情况下尽量接近 $\boldsymbol{x}$，进而最大化 $\boldsymbol{x}$ 的产生概率 $P(\boldsymbol{x})$. 在变分自编码器中，一般选择输出的分布为高斯分布，即 $P(\boldsymbol{x}|\boldsymbol{z};\theta)=\mathcal{N}(\boldsymbol{x}|f(\boldsymbol{z};\theta),\sigma^2\times\boldsymbol{I})$. 当然如果训练样本是二值的，也可以选择伯努利分布作为输出分布.

然而想要最大化 $P(\boldsymbol{x})$ 的前提是必须知道隐变量 $\boldsymbol{z}$ 的分布，这通常是未知的，并且还可能是一个复杂的分布. 但是，任何复杂的分布都可以通过对简单分布，比如 $\mathcal{N}(0,\boldsymbol{I})$，进行一个映射获得，而这个映射可以通过一个神经网络来实现. 假设 $f(\boldsymbol{z};\theta)$ 是一个多层神经网络，那么该神经网络前几层所要完成的工作就是将一个简单分布映射为隐变量的分布，而后几层则作为生成模型，将隐变量作为输入用来生成数据. 基于此思想，为了简化问题，直接令 $P(\boldsymbol{z})=\mathcal{N}(0,\boldsymbol{I})$.

变分自编码器最关键的问题在于尝试去采样可能生成 $\boldsymbol{x}$ 的 $\boldsymbol{z}$，同时计算 $P(\boldsymbol{x})$. 因此如果想要实现变分自编码器，首先需要解决的问题就是怎样定义隐变量 $\boldsymbol{z}$，其次是如何处理隐变量 $\boldsymbol{z}$ 的积分. 显然随机选择一个隐变量肯定是行不通的，一个可行的方法就是前面所说的多层神经网络，通过在生成模型前添加一个编码网络，训练一些样本来获得隐变量的分布. 因此这里需要引入一个新的函数 $Q(\boldsymbol{z}|\boldsymbol{x})$ 来完成编码网络的功能，通过该函数，模型可以在一个给定的 $\boldsymbol{x}$ 的前提下，获得一个能使最终输出为 $\boldsymbol{x}$ 的关于 $\boldsymbol{z}$ 的分布.

此时，我们并不知道编码函数 $Q(\boldsymbol{z}|\boldsymbol{x})$ 该如何定义，但是有一点很明确，就是希望 $Q(\boldsymbol{z}|\boldsymbol{x})$ 与理想的 $P(\boldsymbol{z}|\boldsymbol{x})$ 尽量接近. 很明显这可以通过 KL 散度(式(41)中表示为 $\mathcal{D}$)来完成，即要最小化式(41)：

$$\mathcal{D}[Q(\boldsymbol{z}|\boldsymbol{x})\|P(\boldsymbol{z}|\boldsymbol{x})] = E_{\boldsymbol{z}\sim Q}[\log(Q(\boldsymbol{z}|\boldsymbol{x}) - \log P(\boldsymbol{z}|\boldsymbol{x})] \tag{41}$$

使用贝叶斯公式将 $P(\boldsymbol{z}|\boldsymbol{x})$ 展开，式(41)可进一步写为

$$\mathcal{D}[Q(z|x)\|P(z|x)]=E_{z\sim Q}[\log(Q(z|x)-\log P(x|z)-\log P(z)+\log P(x)] \quad (42)$$

由于 $P(x)$与 z 无关,然后将式(42)右边中括号里的第 1 项和第 3 项结合,就可得出变分自编码器的核心公式:

$$\log P(x)-\mathcal{D}[Q(z|x)\|P(z|x)]=E_{z\sim Q}[\log(P(x|z)]-\mathcal{D}[Q(z|x)\|P(z)] \quad (43)$$

由于变分自编码器最终的目标是最大化 $P(x)$以及最小化$\mathcal{D}[Q(z|x)\|P(z|x)]$. 因此变分自编码器的目标函数为

$$J_{VAE}=E_{z\sim Q}[\log(P(x|z)]-\mathcal{D}[Q(z|x)\|P(z)] \quad (44)$$

此时,变分自编码器的目标就变成利用优化算法最大化其目标函数,这可以通过最小化式(44)右边第 2 项来达到,即让 $Q(z|x)$与真实的 $P(z)$尽可能接近. 如果对传统自编码器比较熟悉的人会发现,式(44)右边第 2 项可以看作传统自编码器的编码部分,而第一项可以看作它的解码部分. 这就是为什么即使变分自编码器在数学基础上与传统自编码器没有什么关系,但还是被称为自编码器的原因. 因此,如果想要 $Q(z|x)$与 $P(z)$尽可能接近,可以通过在传统自编码器的编码部分增加一个约束,使它所产生数据的分布与 $P(z)$尽可能相同.

那么,剩下的问题就是如何利用优化算法对式(44)的右边进行优化,以获取一个合适的函数 Q. 在进行优化操作时,通常会先给 $Q(z|x)$一个特殊的分布表达:$Q(z|x)=\mathcal{N}(z|\mu(x;\vartheta),\Sigma(x;\vartheta))$,其中 $\mu(x;\vartheta)$和 $\Sigma(x;\vartheta)$是任意确定性的函数,ϑ 是通过数据学习的参数. 那么优化的目的就转化为使 $Q(z|x)$的分布$\mathcal{N}(z|\mu(x;\vartheta),\Sigma(x;\vartheta))$尽可能接近 $P(z)$的分布$\mathcal{N}(0,\boldsymbol{I})$. 为了简化计算过程,对 Σ 增加一个约束,令其为一个对角矩阵. 因此变分自编码器的整个训练网络如图 4 所示.

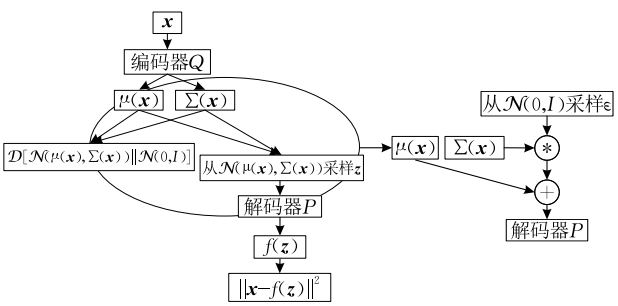

图 4 变分自编码器网络结构图

需要特别注意的是图4的右边部分,由于该网络采用的是随机采样,无法使用反向传播算法,因此引入一个称为“重新参数化”的技巧. 首先按照分布 $\varepsilon\sim\mathcal{N}(0,\boldsymbol{I})$对 ε 进行采样,然后通过一个变换 $z=\mu(x)+\Sigma^{1/2}(x)\times\varepsilon$ 来获得 z 的采样.

当变分自编码器使用图 4 的结构完成训练后,就可以利用模型中的解码部分来生成与训练数据类似的数据,生成数据模型具体结构如图 5 所示.

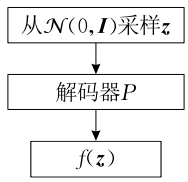

图 5 变分自编码器产生模型

变分自编码器的优点在于不需要很强的假设,通过反向传播算法可以快速训练,缺点在于只能生成与训练数据相似的输出.

3.7.2 条件变分自编码器(Conditional Variational Auto-Encoders)

由上一节可知,变分自编码器可以产生与训练样本非常类似的输出,但是如果我们想要产生一个特定的数据呢?它似乎无法完成. 基于这个目的,Sohn 等人[50]在 2015 年提出了条件变分自编码器. 为了能够产生特定的数据,Sonh 等人在变分自编码器的基础上引入了一个输出变量 y,因此条件变分自编码器总共包含 3 种变量:输入变量 x、隐变量 z 及输出变量 y. 由于 y 为需要生成的特定数据,因此条件变分自编码器的目标就转变为最大化 $P(y|x)$. 与变分自编码器的原理类似,条件变分自编码器也令 $P(z)=\mathcal{N}(0,\boldsymbol{I})$,同时引入一个函数 Q,使 $Q(z|y,x)$与理想的 $P(z|y,x)$尽量接近,因此,条件变分自编码器的核心公式为

$$\log P(y|x)-\mathcal{D}[Q(z|y,x)\|P(z|y,x)]=E_{z\sim Q(\cdot|y,x)}[\log(P(y|z,x)]-\mathcal{D}[Q(z|y,x)\|P(z|x)] \quad (45)$$

式(45)与变分自编码器核心公式非常相似,唯一的差别在于输出变量 y. 因此,条件变分自编码器采用和变分自编码器一样的优化方法对式(45)进行优化,而且也利用了变分自编码器中的“重新参数化”技巧. 条件变分自编码器的整个训练网络以及用于产生数据的网络如图 6、图 7 所示.

由于输出变量 y 的引入,使得条件变分自编码器可以产生与 y 相匹配的指定数据,这是与变分自编码器最大的不同之处,也是其最大优势所在.

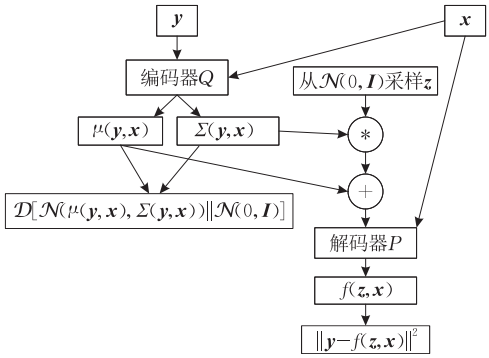

图 6　条件变分自编码器网络结构图

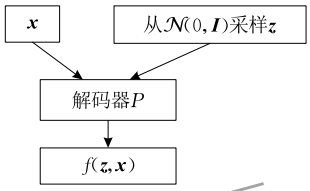

图 7　条件变分自编码器产生模型

3.7.3　其它变分自编码器

由于变分自编码器在数据生成方面所取得的成绩，近 2 年来，很多文献[51-56]提出了对变分自编码器的改进方法，通过分析发现，这些改进方法可归纳为以下几种：(1) 在变分自编码器的损失函数上增加一个约束[51-53]；(2) 将变分自编码器与自回归模型相结合[54-55]，比如循环神经网络(Recurrent Neural Network，RNN)，像素卷积神经网络(Pixel-CNN)等；(3) 对变分自编码器的结构进行修改，将解码阶段分为两部分，产生由粗到细的图像[56].

3.8　变换自编码器(Transforming Auto-Encoders)

变换自编码器[57]是由 Hinton 等人在 2011 年提出的，他们认为人工神经网络应该使用一种被称为“胶囊”的基础单元去执行一些非常复杂的内部计算，然后把计算结果封装到一个比较小的向量里，该向量中包含了关于输入的大量信息.

胶囊的主要优势在于提供了一种通过识别部分进而识别整体的简单方法. 图 8 是变化自编码器用于图像变换的一个结构图，虚线框中包含的是一个胶囊，一个变换自编码器可以由多个胶囊并行组合而成. 图 8 实现的功能是：给定一副图像以及一组确定的 Δx 和 Δy，那么输出图像会按照 Δx 和 Δy 对输入图像进行位置上的变换.

每个胶囊的输出数据包括两部分：可视化实体出现在其限制域上的概率 p 以及一组实例化参数，这组参数中的信息都与可视化实体的一个标准版本相对应，主要包含其准确位置、光照和变换等信息，相当于我们平时所讲的特征. 不难发现，可视化实体

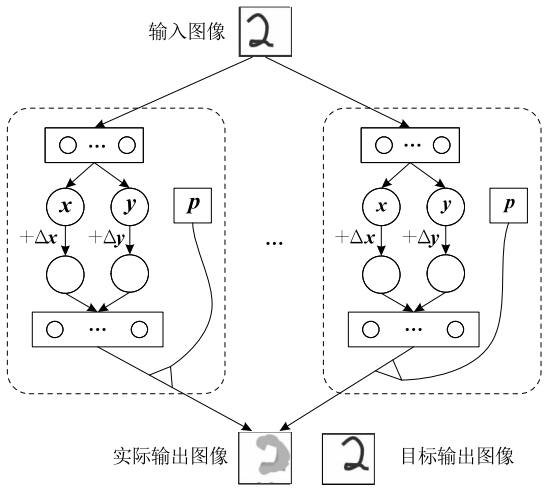

图 8　变换自编码器网络结构图

出现的概率是局部不变的，但实例化参数却是等变的. 因为当实体随着流形移动时，它出现的概率不会发生变化，而实例化参数表达的是实体在流形上的内在坐标，会随着观测条件的改变而变化.

以图 8 为例，当胶囊正常工作时，识别单元首先对输入图像进行识别，获取它的位置信息 (x,y)，然后胶囊根据事先指定的位置偏移变量 Δx 和 Δy，将 $\Delta x+x$ 和 $\Delta y+y$ 作为新的变量送入生成单元，最后输出的图像由每个胶囊的输出数据共同决定. 然而，每个胶囊在模型中的价值并不等价，因此，概率 p 主要用于反应每个胶囊的实例化参数对目标输出的贡献程度. 那么，那些无效的，没有位置变化的胶囊对最后的输出不会有任何影响. 对于变换自编码器，除了概率的影响，最重要和必不可少的一个条件是每个活动胶囊计算出来的位置信息 (x,y) 都必须与可视化实体真实的位置信息 (x,y) 相一致，这是保证变换自编码器能够输出正确图像的重要前提.

其实胶囊与卷积神经网络中的局部单元非常相似，但是一个最重要的区别在于，卷积神经网络池化层通常只输出那个网络中最活跃的神经元的结果，而胶囊会将识别单元的所有输出封装到一个简单向量中并用于后续更高级别的活动. 此外，由于识别单元计算出来的位置信息非常精确，因此胶囊可以更好地利用这些信息，甚至可以精确到图像的每一个像素.

变换自编码器的最大优点在于，可以通过任何一种已知的方式去强制规定一个胶囊的输出，这些输出代表的是图像的任意性质. 但是它也有一个非常严重的缺陷，那就是每个胶囊在同一时间只能代表可视化实体的一个实例.

3.9 其它自编码器

除了前面介绍的通用自编码器，还有一类自编码器是根据具体的研究领域所具有的特点进行改进的，我们会选择几类比较有代表性的进行详细介绍，具体包括基于可区分性的自编码器，基于局部的自编码器以及张量自编码器.

3.9.1 基于可区分性的自编码器

(1) 区分自编码器(Discriminative Auto-Encoder)

Xie 等人[58]在 2015 年提出了一种新的自编码器——区分自编码器，它通过对隐层神经元增加一个 Fisher 区分准则，使隐层特征对几何结构变化具有可区分性和不敏感性. 区分自编码器最关键的技术在于它对传统正则自编码器的损失函数所做出的改进，其损失函数具体可描述为

$$J_{DisAE}(\boldsymbol{W},\boldsymbol{b})=\sum(L(\boldsymbol{x},\boldsymbol{y}))+\frac{\lambda}{2}\|\boldsymbol{W}\|_F^2+\frac{\gamma}{2}(tr(S_w(\boldsymbol{z}^t))-tr(S_b(\boldsymbol{z}^t))) \tag{46}$$

该损失函数前两项与传统正则自编码器一样，第 3 项是区分正则项，$S_w(\boldsymbol{z}^t)$和 $S_b(\boldsymbol{z}^t)$分别为

$$S_w(\boldsymbol{z}^t)=\sum_{i=1}^{C}\sum_{\boldsymbol{z}_{i,j}^t\in i}(\boldsymbol{z}_{i,j}^t-\boldsymbol{m}_i^t)(\boldsymbol{z}_{i,j}^t-\boldsymbol{m}_i^t)^{\mathrm{T}} \tag{47}$$

$$S_b(\boldsymbol{z}^t)=\sum_{i=1}^{C}n_i(\boldsymbol{m}_i^t-\boldsymbol{m}^t)(\boldsymbol{m}_i^t-\boldsymbol{m}^t)^{\mathrm{T}} \tag{48}$$

其中 $\boldsymbol{z}^t=[\boldsymbol{z}_1^t,\boldsymbol{z}_2^t,\cdots,\boldsymbol{z}_C^t]$，$\boldsymbol{z}_i^t$是第 i 类样本在自编码器的第 k 个隐层的特征表达，即 $\boldsymbol{z}_i^t=[\boldsymbol{z}_{i,1}^t,\boldsymbol{z}_{i,2}^t,\cdots,\boldsymbol{z}_{i,J}^t]$，$J$ 为每类样本的样本数量，$\boldsymbol{m}_i^t$和 $\boldsymbol{m}^t$分别表示 $\boldsymbol{z}_i^t$和 $\boldsymbol{z}^t$的平均向量. 因此 $S_w(\boldsymbol{z}^t)$和 $S_b(\boldsymbol{z}^t)$所表达的意义是 $\boldsymbol{z}^t$的类内离散度和类间离散度. 而区分自编码器的最终目标就是使类内离散度尽量小，同时类间离散度尽量大，很明显这对于分类任务非常有利.

(2) 大边缘自编码器(Large Margin Auto-Encoders)

受到区分自编码器最小化类内离散度和最大化类间离散度思想的启发，Liu 等人[36]在 2017 年提出了大边缘自编码器，通过使属于不同类的样本在隐层空间具有更大的边缘距离来进一步提升自编码器的区分能力.

与区分自编码器的思想类似，大边缘自编码器也是对传统正则自编码器的损失函数进行了改进，通过对损失函数增加一个大边缘惩罚项，使带有不同类标签的样本在 k 最近邻中保持一个安全距离. 因此，大边缘自编码器的损失函数可表示为

$$J_{LMAE}(\boldsymbol{W},\boldsymbol{b})=\sum(L(\boldsymbol{x},\boldsymbol{y}))+\frac{\lambda}{2}\|\boldsymbol{W}\|_F^2+J_{LM}(\boldsymbol{W}) \tag{49}$$

$$J_{LM}(\boldsymbol{W})=\sum_{k_1=1}^{m}\sum_{k_2=1}^{m}\eta_{k_1k_2}\|\boldsymbol{W}(\boldsymbol{x}_{k_1}-\boldsymbol{x}_{k_2})\|^2+\sigma\sum_{k_1=1}^{m}\sum_{k_2=1}^{m}\sum_{k_3=1}^{m}\eta_{k_1k_2}(1-\tau_{k_1k_3})h(1+\|\boldsymbol{W}(\boldsymbol{x}_{k_1}-\boldsymbol{x}_{k_2})\|^2-\|\boldsymbol{W}(\boldsymbol{x}_{k_1}-\boldsymbol{x}_{k_3})\|^2) \tag{50}$$

其中 $\eta_{k_1k_2}=\begin{cases}1, & \boldsymbol{x}_{k_2}\text{ 是 }\boldsymbol{x}_{k_1}\text{ 的一个目标邻居}\\ 0, & \text{否则}\end{cases}$ (51)

$$\tau_{k_1k_3}=\begin{cases}1, & x_{k_3}\text{ 与 }x_{k_1}\text{ 属于同一个类}\\ 0, & \text{否则}\end{cases} \tag{52}$$

σ 用于控制大边缘惩罚项中两项之间的相对重要性. 该惩罚项与大边缘 k 最近邻分类算法的优化问题类似，主要区别在于所使用的 $h(\cdot)$不同，前者为 $h(s)=\frac{1}{\gamma}\ln(1+\exp(\gamma s))$，后者为 $h(s)=\max(s,0)$.

3.9.2 基于局部的自编码器

(1) 协同局部自编码器(Collaborative Local Auto-Encoder)

许多研究发现，同一副图像中会存在一些具有相似结构的局部图像块，并且会在图像的不同尺度上重复出现多次，因此图像超分重建可以通过在测试数据本身进行自相似性采样来提升重建效果. 针对此思想，文献[59]提出先使用非局部自相似性搜索对图像进行预处理，然后将其结果作为自编码器的输入信号进行图像超分重建.

由于非局部自相似性搜索的结果中会出现很多重叠的块，而为了协调重叠块之间的兼容性，产生更加平滑和自然的纹理，Cui 等人[59]在稀疏自编码器的基础上提出了协同局部自编码器. 由此分析，协同局部自编码器应该包括两个约束：稀疏约束和兼容性约束.

假设 $\hat{\boldsymbol{x}}_i$是由非局部自相似搜索所产生的重叠块，那么协同局部自编码器的损失函数可表示为

$$J_{CLAE}(\boldsymbol{W},\boldsymbol{b})=\frac{1}{2n}\sum_{i=1}^{n}\|\boldsymbol{y}_i-\hat{\boldsymbol{x}}_i\|^2+\frac{\lambda}{n}\sum_{i=1}^{n}\|\boldsymbol{u}_i\|_1+\frac{\beta}{2n}\sum_{i=1}^{n}\|\boldsymbol{y}_i-\mathcal{F}_i\tilde{\boldsymbol{x}}\|^2 \tag{53}$$

其中，$\boldsymbol{u}_i=\boldsymbol{z}_i/\sqrt{\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i}$，$\boldsymbol{z}_i$为 $\hat{\boldsymbol{x}}_i$的隐层表达 $\sigma(\boldsymbol{W}\hat{\boldsymbol{x}}_i+\boldsymbol{b})$，$\boldsymbol{y}_i=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{u}_i+\boldsymbol{b}'$是协同局部自编码器在重建阶段产生的块，$\tilde{\boldsymbol{x}}$是最终重建的超分图像，$\mathcal{F}_i$是提取块操作.

损失函数的第 2 项就是稀疏约束，使用对 $\boldsymbol{u}_i$ 的 L_1 范数能在获取稀疏表示的同时极大的简化计算过程. 第 3 项是兼容性约束，这里所说的兼容性是指在将重建块组合成完整图像时，如何避免图片产生锯齿形的边缘，使最后产生的图像具有更加平滑和自然的纹理特征. 为了实现兼容性，最理想的状态就是由协同局部自编码器重建的块 $\boldsymbol{y}_i$ 与从重建的超分图像中提取的块越接近越好.

(2) 局部约束稀疏自编码器(Locality-Constrained Sparse Auto-Encoder)

Luo 等人[60] 针对图像分类问题，在 2015 年提出了一种局部约束稀疏自编码器. 他们认为，对于分类问题，局部性比稀疏性更重要，因此在稀疏自编码器的基础上引入了局部的概念. 改进后的自编码器能利用相似的特征对相似的输入进行编码.

严格来说，局部约束稀疏自编码器的思想与 k-稀疏自编码器更加接近，它在编码层仅保留输入信号 k 个最近邻信号的特征，其他都设置为 0，解码器仅利用这 k 个特征进行信号重建. 具体描述如下

$$\text{编码过程：}\quad \boldsymbol{z}=s(\widetilde{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}) \tag{54}$$

$$\text{解码过程：}\quad \boldsymbol{y}=\boldsymbol{W}g(\boldsymbol{z};\widetilde{\boldsymbol{W}},\boldsymbol{x})+\boldsymbol{b}' \tag{55}$$

其中 $\widetilde{\boldsymbol{W}}=[\widehat{\boldsymbol{w}}_1,\widehat{\boldsymbol{w}}_2,\cdots,\widehat{\boldsymbol{w}}_N]$，$\widehat{\boldsymbol{w}}_i=\dfrac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|_2}$，函数 g 用来保留 $\boldsymbol{z}$ 的 k 个特征，具体表示为

$$g(\boldsymbol{z}_i;\boldsymbol{W},\boldsymbol{x})=\begin{cases}\boldsymbol{z}_i, & \text{对应输入}\boldsymbol{x}\text{的}k\text{最近邻}\\ 0, & \text{否则}\end{cases} \tag{56}$$

输入 $\boldsymbol{x}$ 的 k 最近邻选择方法为先计算 $\dfrac{\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}}{\|\boldsymbol{w}_i\|_2}$ 的值，然后选取其中值最大的 k 个. 同时，为了使损失函数对局部性进行约束，在稀疏自编码器的基础上对正则项进行了修改：

$$J_{NCAE}(\boldsymbol{W})=\sum(L(\boldsymbol{x},\boldsymbol{y}))+\lambda\sum_{i=1}^{N}|\boldsymbol{z}_i| \tag{57}$$

这个正则项的目的是使每个特征具有相近的概率被更新. 因为在基于局部的学习模型中，越早被更新的特征被选为输入最近邻的概率远远大于随机特征，这样会导致学习主要集中在早期更新的特征，从而导致其他特征不再被更新，这就是所谓的死特征现象. 通过对稀疏自编码器损失函数的修改，可以有效缓解地这种现象.

(3) 协同稳定非局部自编码器(Non-Local Auto-Encoder with Collaborative Stabilization)

大脑具有一个非常重要的特性，那就是类似的视觉线索可以刺激相同的神经元去引导相似的神经信号，传统神经网络一般没有考虑这个特性，因此 Wang 等人[61] 从这一关键性质出发，提出了一种基于协同稳定的非局部自编码器. 该自编码器提出相似的输入应该引导相似的网络传播，并通过两部分去达成这个目标. 首先对稀疏自编码器的损失函数进行相似性约束以及噪声约束，提出了非局部自编码器，然后再通过一个协同稳定的方法进一步提升这种特性.

假设有 3 个实体对象 $\boldsymbol{x}_1$、$\boldsymbol{x}_2$ 和 $\boldsymbol{x}_3$，其相应的隐层表达分别为 $\boldsymbol{h}_1$、$\boldsymbol{h}_2$ 和 $\boldsymbol{h}_3$，相似性是指如果 $\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_p<\|\boldsymbol{x}_1-\boldsymbol{x}_3\|_p$，那么 $\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_p<\|\boldsymbol{h}_1-\boldsymbol{h}_3\|_p$，这里将 $\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_p$ 称为一个扰动. 由于目前所采用的激活函数一般是非线性的，所以这个关系并不一定能成立，而这明显与大脑的工作原理是相违背的. 因此，为了能够满足相似性，非局部自编码器提出对扰动进行惩罚，在稀疏自编码器的损失函数上增加了一个相似性正则项：

$$\lambda\sum_i\omega_i\|\boldsymbol{h}-\boldsymbol{h}_i\|_p \tag{58}$$

其中 λ 用来控制模型对扰动的抑制程度. 由于不同的扰动对模型具有不同的贡献，因此为每个扰动分配不同的权重 ω_i. 同时为了快速移除权重随机初始化时所产生的噪声对模型的影响，因此，该模型的损失函数可表示为

$$J_{NLAE}(\boldsymbol{W},\boldsymbol{b})=\sum(L(\boldsymbol{x},\boldsymbol{y}))+\lambda\sum_i\omega_i\|\boldsymbol{h}-\boldsymbol{h}_i\|_p+\beta\sum_{j=1}^{k}KL(\rho\|\hat{\rho}_j)+\eta(\|\boldsymbol{W}\|_1+\|\boldsymbol{W}'\|_1) \tag{59}$$

$\boldsymbol{W}$ 和 $\boldsymbol{W}'$ 分别表示编码权重和解码权重.

由于自编码器自身的缺陷，其优化参数很容易陷入局部最优解，因此，仅通过一个训练好的非局部自编码器是无法完全消除网络中的扰动的. 所以 Wang 等人在测试阶段又增加了一个协同稳定算法来进一步消除这种扰动.

通过对损失函数的相似性约束以及后续的协同稳定性操作，协同稳定非局部自编码器不仅能够提取输入信号的有用特征，同时有效地消除了数据传播过程中的扰动，保证了网络内部信号传输时的稳定性.

3.9.3　张量自编码器(Tensor Auto-Encoder)

前面所介绍的自编码器模型都是基于向量空间的，因此无法学习大数据的特征，因为一个向量无法对大数据的高度非线性分布建模，尤其是异类数据. 因此，2014 年 Zhang 等人[62] 提出了张量自编码器的概念，同时通过堆叠多个张量自编码器去构建一个深度计算模型用于大数据的特征提取. 张量自编

码器和传统自编码器非常类似,同样包括一个输入层、一个隐层和一个输出层.唯一的不同在于张量自编码器的每一层都用一个张量表示.

张量自编码器的编码过程可描述为

$$H = f_\theta(\boldsymbol{W}^{(1)} \odot X + b^{(1)}) \tag{60}$$

解码过程为

$$Y = h_{\boldsymbol{W},b}(X) = f_\theta(\boldsymbol{W}^{(2)} \odot H + b^{(2)}) \tag{61}$$

张量自编码器的编解码过程和传统自编码器基本完全一样.不同之处在于 $\boldsymbol{W}^{(1)}$,$\boldsymbol{W}^{(2)}$ 是 $(N+1)$ 阶张量,$b^{(1)}$ 和 $b^{(2)}$ 是 N 阶张量,而 $\odot$ 为一个多点积过程,一个 $(N+1)$ 阶和 N 阶张量的多点积操作结果为一个 N 阶张量.

假设有一个训练样本集 $\{(X^{(1)},Y^{(1)}),\cdots,(X^{(m)},Y^{(m)})\}$,张量自编码器首先根据张量距离将每个训练样本 $(X^{(i)},Y^{(i)})(i=1,\cdots,m)$ 从张量形式转换为向量形式 $(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})(i=1,\cdots,m)$.而为了从复杂数据中尽可能地获取未知分布,张量自编码器采用张量距离作为重建误差中的平均平方和误差项,因此张量自编码器的损失函数可表示为

$$J_{TAE}(\boldsymbol{W},b) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}(h_{\boldsymbol{W},b}(\boldsymbol{x})-\boldsymbol{y})^{\mathrm{T}}G(h_{\boldsymbol{W},b}(\boldsymbol{x})-\boldsymbol{y})\right)+\frac{\lambda}{2}(\|\boldsymbol{W}^{(1)}\|_2^2+\|\boldsymbol{W}^{(2)}\|_2^2) \tag{62}$$

其中 G 是一个度量系数.

4 自编码器对比与分析

通过对大量自编码器的研究发现,大部分自编码器的创新点都在于对基础的损失函数增加各种正则项约束,使模型达到不同的效果,比如稀疏性、鲁棒性、收缩性、兼容性等.同时,大部分自编码器进行约束的对象都是隐层表达,因为对于大部分研究者而言,自编码器最重要的性质还是在于其能自动学习有效特征,因此大部分自编码器所做的改进都是针对隐层表达.而其中研究最多的是隐层表达对输入噪声的鲁棒性,包括边缘降噪自编码器、收缩自编码器、饱和自编码器、协同稳定非局部自编码器.除此之外,还有一些其他的创新方向,比如领域自适应边缘降噪自编码器和 k-稀疏自编码器,在满足各种约束的前提下,大幅度减少了训练时间.卷积自编码器是为了有效利用二维数据的空间相关信息,变换自编码器是为了使网络更加简单,信息高度集中,而张量自编码器则是为了能够实现对高维数据的建模.表 2 对前面介绍的自编码器进行了一个归纳总结,特别指出了各类自编码器的改进点以及所要达到的目的.

表 2 各类自编码器分析与对比

自编码器名称	改进点	目的
传统正则自编码器	在传统自编码器的损失函数上增加权重衰减项 $\lambda\|\boldsymbol{W}\|_2^2$	提升网络泛化能力,防止过拟合
降噪自编码器	将带有噪声的损坏信息作为输入信号	使重建信号对输入信号中的噪声具有一定的鲁棒性
领域适应性边缘降噪自编码器	利用 $\boldsymbol{W}=E[\boldsymbol{P}]E[\boldsymbol{Q}]^{-1}$ 直接计算权重,无需使用优化算法	大幅度缩减训练时间
非线性表示边缘降噪自编码器	在传统自编码器的损失函数上增加正则项 $\frac{1}{2}\sum_{d=1}^{D}\sigma_{xd}^2\sum_{h=1}^{D_h}\frac{\partial^2 L}{\partial \boldsymbol{z}_h^2}\left(\frac{\partial \boldsymbol{z}_h}{\partial \tilde{\boldsymbol{x}}_d}\right)^2$	既考虑重建函数对隐层表达的敏感度,又考虑隐层表达对输入信号的敏感度
稀疏自编码器	在传统自编码器的损失函数上增加稀疏规则项 $\beta\sum_{j=1}^{h}KL(\rho\|\hat{\rho}_j)$	获取输入信号的稀疏表达
k-稀疏自编码器	丢弃非线性激活函数,利用排序算法或 Relu 函数选取 k 个最大激活值	获取隐层准确的稀疏度,缩减训练时间
收缩自编码器	在传统自编码器的损失函数上增加正则项 $\lambda\|J_f(\boldsymbol{x})\|_F^2$	保证隐层表达对输入微小变化的鲁棒性
高阶收缩自编码器	在收缩自编码器的损失函数上增加一个二阶正则项 $\gamma\|H_f(\boldsymbol{x})\|_F^2$	在收缩自编码器基础上进一步延伸隐层表达对输入微小变化的鲁棒性
饱和自编码器	在传统自编码器的损失函数上增加正则项 $\alpha\sum_{i=1}^{h}f_c(W_i^e x+b_i^e)$	鼓励激活值落在相应激活函数的饱和区域
卷积自编码器	使用卷积层和池化层代替传统自编码器的全连接层	保留二维信号的空间信息
变分自编码器	相当于在传统自编码器的隐层表达上增加一个对隐变量的分布约束 $\mathcal{D}[\mathcal{N}(\mu(\boldsymbol{x}),\Sigma(\boldsymbol{x}))\|\mathcal{N}(0,\boldsymbol{I})]$	使编码器产生的隐层表达满足 $\mathcal{N}(0,\boldsymbol{I})$ 分布
条件变分自编码器	在变分自编码器的基础上引入标签信息	产生与标签相匹配的数据
变换自编码器	引入胶囊的概念,同时增加额外信息输入	每个胶囊的输出可以代表数据的任意性质,同时使网络学习更加简单
区分自编码器	在传统正则自编码器的损失函数上增加正则项 $\frac{\gamma}{2}(tr(S_w(\boldsymbol{z}^t))-tr(S_b(\boldsymbol{z}^t)))$	使类内离散度尽量小,类间离散度尽量大
大边缘自编码器	在传统正则自编码器的损失函数上增加正则项 $J_{LM}(\boldsymbol{W})$	增加不同类样本在隐层空间的边缘距离,进一步提升自编码器的区分能力

（续　表）

自编码器名称	改进点	目的
协同局部自编码器	将稀疏自编码器损失函数的稀疏项改为$\frac{\lambda}{n}\sum_{i=1}^{n}\|\boldsymbol{u}_i\|_1$，同时增加兼容性约束$\frac{\beta}{2n}\sum_{i=1}^{n}\|\boldsymbol{y}_i-\mathcal{F}_i\tilde{\boldsymbol{x}}\|^2$	使用简单方式获取稀疏性，重建平滑和自然的图像纹理
局部约束稀疏自编码器	将稀疏自编码器的稀疏约束项改为$\lambda\sum_{i=1}^{N}\|\boldsymbol{z}_i\|$，同时保留$\boldsymbol{z}$的$k$个隐层表达	利用相似特征对相似输入进行编码，缓解死特征现象
协同稳定非局部自编码器	在稀疏自编码器的损失函数上增加正则项$\lambda\sum_i\omega_i\|h-h_i\|_p+\eta(\|\boldsymbol{W}\|_1+\|\boldsymbol{W}'\|_1)$	消除网络传播中的扰动和随机初始化权重时所产生的噪声
张量自编码器	每一层都用一个张量表示	对大数据的高度非线性分布建模

此外，我们从文献中提取了部分自编码器在MNIST和CIFAR-10数据库上的实验结果，并进行对比和分析，如表3、表4所示.需要说明的是：对于有多个实验结果的算法，我们只选取其最优的结果.通过观察表3和表4，我们发现卷积自编码器在两个数据库上都取得了最好的结果.这说明对图像而言，二维空间邻域信息对图像的分类效果具有非常重要的影响.此外，由于降噪自编码器考虑了鲁棒性，性能提升比较明显.而收缩自编码器由于考虑的是隐层表达对噪声的鲁棒性，所以效果优于降噪自编码器，这与收缩自编码器的理论分析相一致.高阶收缩自编码器中二阶惩罚项的加入，延续了模型的鲁棒性，性能得到进一步的提升.另外，值得一提的是，对于MNIST数据库，k-稀疏自编码器也表现出非常好的效果，因为k-稀疏自编码器选择隐层表达激活值中最大的k个神经元，而这k个特征值已经包含了输入信号中的大部分信息，基本可以完全表达输入信息.基于以上实验结果，我们得出以下结论：对于图像数据而言，二维空间信息最重要.其次，如果能在隐层中找到最具代表性的特征，对于最后的分类性能有很大帮助.最后，模型对噪声的鲁棒性非常重要，而隐层表达对噪声的鲁棒性明显比重建信号对噪声的鲁棒性更加重要.

表3　部分自编码器在MNIST库上的实验对比

自编码器名称	错误率
传统自编码器	1.78
传统正则化自编码器	1.68
降噪自编码器	1.28
非线性表示边缘降噪自编码器	1.37
收缩自编码器	1.14
高阶收缩自编码器	1.04
k-稀疏自编码器	0.97
卷积自编码器	0.71

表4　部分自编码器在CIFAR-10库上的实验对比

自编码器名称	错误率
传统自编码器	27.7
传统正则化自编码器	27.3
降噪自编码器	26.1
收缩自编码器	25.6
高阶收缩自编码器	24.6
稀疏自编码器	25.2
卷积自编码器	21.8

5　实际应用

5.1　数据分类

近年来，由于自编码器所具有的特征学习能力，被广泛地应用于数据分类领域.数据分类主要指的是根据数据内在的一些特征对数据进行归类.对于分类任务，无论是传统方法还是深度学习方法，其基本原理十分相似，首先提取对象的有效特征，然后将特征送入分类器进行分类.那么当分类器相同时，提取特征的质量将直接影响最后分类结果的准确率.相比传统分类方法，深度学习得到快速发展的原因就是它能自动学习到对象比较好的特征，甚至不需要任何领域知识和相关技能.而自编码器作为深度学习中一种无监督学习方法，已经被证明能够逐层自动学习到对象的有效特征，因此在分类任务中被广泛应用，并且取得了很好的效果.

根据对大量文献的研究发现，利用自编码器进行分类的方法有一个类似的通用模型，如图9所示.该模型首先采用无监督方法对多个自编码器进行学习，然后堆叠各种自编码器的编码部分用于特征提取，最后采用有监督学习对网络参数进行微调.

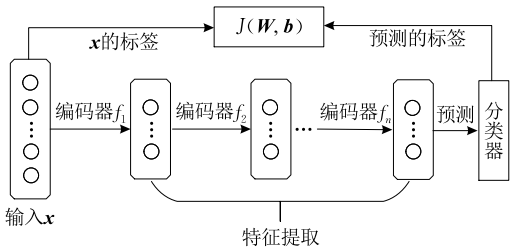

图9　基于自编码器的数据分类模型

基于以上模型的研究方法最大的区别在于所采

用的自编码器、损失函数以及送入自编码器网络的数据不同.接下来,我们将对目前自编码器应用比较多的几个数据分类研究领域进行详细介绍.

(1)图像分类

在分类任务中,图像分类是研究最多的一个分支,非常有名的 ImageNet 计算机视觉挑战赛就是为图像分类专门举办的一个赛事,涉及到 1000 类图像的分类问题,其中所有模型的提出全部都是基于卷积神经网络.之所以受到如此巨大的关注,其根本原因在于卷积神经网络能够自动学习有效的特征,同时还考虑了图像的空间信息.相比卷积神经网络,自编码器除了可以自动学习信息的内在特征,它还能够处理大量无标签信息,因此被大量用于完成图像分类任务中的特征提取工作.

Wang 等人[10]利用传统自编码器的降维特性去提取特征,并重点分析了网络结构中隐层节点的数量对分类精度的影响,实验结果显示在图像结构比较简单的情况下,比如 MNIST 图像,当隐层节点的数量接近原始数据的本质维度时,分类精度就能达到 93%以上,此时即使大幅度增加隐层节点数量也对精度没有很大帮助,而在图像结构比较复杂的情况下,比如人脸图像,这个规律就不太适用.但是不管图像结构如何,一个普遍存在的规律是随着隐层节点数量的增加,分类精度保持上升趋势.这个结论为以后网络结构的设计提供了一定的理论基础支持.

在基于自编码器的图像分类任务中,目前研究比较多的有遥感图像和雷达图像.Chen 等人[11]首次将深度学习方法引入高光谱遥感图像分类,首先利用主成分分析(Principal Component Analysis,PCA)提取图像的空间域信息,然后提取像素的邻域信息并将它拉直为一个一维向量,最后利用堆叠自编码器进一步提取图像的深层特征,实验结果显示,该方法在 KSC 和 Pavia 两个数据集上的整体准确率分别达到了 98.76%、98.52%,其平均准确率分别为 97.9%、97.82%,分类效果优于很多后期提出的研究方法.该方法不仅将自编码器引入遥感图像分类,还将它与传统特征提取方法相结合,为以后的研究奠定了很好的基础.接着出现了很多将传统特征提取方法与自编码器相结合的研究,Cheng 等人[12]使用方向梯度直方图(Histogram of Oriented Gradient,HOG)去提取图像的初步特征,然后训练一个单隐层稀疏自编码器去学习一个多层可视化字典,最后利用一个有监督单隐层神经网络去微调这个字典,实验显示在 LULC 数据集上的准确率为 87.75%.Li 等人[13]则先使用稀疏自编码器去提取初始特征,然后再利用多重归一化差分特征提取方法对特征进行扩展,共同进行特征提取,在 UCM 数据集上的准确率达到了 91.29%.Geng 等人[14]利用灰度共生矩阵和 Gabor 滤波器去提取雷达图像的初步特征,然后将初步特征送入稀疏自编码器进行学习,接着 Geng 等人[15]又在前者的基础上,将传统特征提取方法扩展为灰度梯度共生矩阵,Gabor 和 HOG 滤波器,然后使用收缩自编码器对初步提取的特征进行学习,相比前者,该方法在相同雷达图像上的准确率最大提升了大约 9%.分析原因可能是由于 HOG 滤波器能很好的捕获物体的轮廓信息,同时对光照和小的位移不敏感,因此能较大地提升模型对雷达图像的分类性能.其次,相比稀疏自编码器,收缩自编码器对隐层表达所具备的鲁棒性对模型的整体性能也具有一定的帮助.文献[16]与前几个方法略有不同,它没有将传统方法与自编码器串联使用,而是改为并联使用,提出一个基于 WPCA(Whitened Principal Component Analysis)和 ICA(Independent Component Analysis)的传统特征学习框架,与堆叠卷积自编码器网络同时进行特征提取,然后将特征一起送入分类器,该方法在 Indian Pine 和 PaviaU 数据集上的整体准确率分别达到了 97.11%、96.08%,其平均准确率分别为 96.78%、97.36%.

Kasun 等人[35]认为传统的线性自编码器无法单独表现一个数据的局部,比如人脸图像上的眼睛,而传统的非线性自编码器又存在学习速度慢的问题,因此针对这些问题,提出将极限学习机(Extreme Learning Machine)分别与传统自编码器、稀疏自编码器相结合,提出了线性和非线性维度减少框架:极限学习机自编码器(Extreme Learning Machine Auto-Encoder,ELM-AE)和稀疏极限学习机自编码器(Sparse Extreme Learning Machine Auto-Encoder,SELM-AE).这两种框架在某种程度上可以表现出数据的局部,具有对噪声鲁棒,训练时间短,能学习类内散布子空间的优点.与传统自编码器最大的区别在于,这两种模型分别使用正交和稀疏随机权重对编码阶段所产生的随机隐藏神经元的参数进行初始化,并且不再需要后期调整,因此模型仅需要学习解码阶段的权重.由于不用学习编码阶段的权重,因此模型的计算代价低,训练速度快.基于以上优点,Lv 等人[17]提出将堆叠传统自编码器与极限学习机相结合用于图像分类,在 Indian Pines 上的整体准确率达到了 94.67%.而为了考虑模型的稀疏性和

鲁棒性，Tao 等人[18]、Wu 等人[19]分别使用了堆叠稀疏自编码器和稀疏降噪自编码器，后者还结合 Fisher 向量去学习更加具有代表性和可区分性的特征，在 UCM 数据集上的准确率达到了 93.5%. Hou 等人[20]、Zhang 等人[63]则将图像的局部空间信息分别引入堆叠自编码器和堆叠稀疏自编码器，在 FLEVOLAND 数据集上的整体准确率分别为 98.61%和 95.02%. 而 Xie 等人[64]并没有额外引入局部空间信息，他们直接利用卷积自编码器考虑图像的局部空间信息，同时将 Wishart 距离与卷积自编码器相结合，在 FLEVOLAND 数据集上的整体准确率为 93.31%. 显然这种方法的准确率不如前两种直接引入局部空间信息的方法，分析原因应该是卷积自编码器虽然也考虑了局部空间信息，但是由于所有的局部采用的是共享权重，没有考虑到每个局部对整体具有不同的贡献程度，因此比直接使用局部空间信息方法的分类效果要差.

以上几种方法都只是将现有的一些自编码器与其他技术相结合，并没有涉及自编码器本身的改进. 因此，文献[65-67]都对传统自编码器的损失函数进行了修改. 第 1 种方法通过增加一个正则项来考虑采样的相似性，并结合上下文信息进一步更新特征，将 Indian Pine 的整体准确率和平均准确率提升到 99.22%和 98.57%，与前面提到的方法相比，其准确率在同一数据库上达到了最佳，主要原因在于该正则项同时考虑了数据的全局和局部相似性，使得类似的数据获取相同标签的概率大幅度增加，这非常符合客观现实，因此效果得到了非常明显的提升. 第 2 种方法则提出一个带有 L_0 范数稀疏约束的单隐层自编码器，并将它与一个单层神经网络相结合用于特征提取，在 LULC 数据集上的准确率达到了 92.44%，同样是采用单隐层自编码器，但该结果在文献[41]的基础上提升了 5%左右，其根本原因是稀疏规则化能实现特征的自动选择，筛除那些对正确预测产生干扰的信息，将这部分信息对应的权重设置为 0，因此在损失函数中增加了稀疏约束能有效提升模型的预测结果. 最后一种方法则增加了一个基于欧几里得距离的有监督约束，用于提取雷达图像的深度特征.

另外，Elshamli 等人[68]还将降噪自编码器引入领域自适应问题中，由于遥感图像中的高光谱图像和多光谱图像可以看成是来自不同领域的数据，因此分别使用降噪自编码器和领域对抗神经网络来学习来自不同领域数据的不变表达.

除了以上两种研究比较多的图像，自编码器还被用于生物医学图像[69-71]，织物图像[72]的分类. 其中比较特别的是，文献[71]提出了一种基于传统自编码器的框架用于同时重建和分类图像. 据了解这是第一个将重建与分类同时进行的研究，该方法使用自编码器去重建信号，然后将一个学习好的线性分类映射添加到自编码器后用于分类. 实验结果显示同等条件下，重建效果优于压缩感知技术，而且运行时间短，具有很好的实时性.

前面讨论的图像分类问题都是基于平衡数据集，但在实际应用中，数据集往往都是不平衡的. 目前比较常用的提高不平衡数据分类准确率的方法有数据再采样，阈值移动，组合技术等，但是这些方法对多类不平衡问题都表现欠佳. 因此 Ng 等人[73]将自编码器用于不平衡数据分类问题. 该方法使用两个带有不同激活函数的堆叠自编码器去分别获取数据的两个不同特征集，然后将这两个特征集相结合形成一个双自编码器特征集，最后利用这个新的特征集去进行不平衡数据的分类. 实验结果显示其效果优于数据再采样方法.

除了对单独图像进行分类，文献[74-75]将两个自编码器进行组合用于多视图问题. 前者先分别建立两个堆叠降噪自编码器网络，然后通过在每个对应层进行相互耦合用于分类. 相比前者只用于单个领域分类问题，后者将一个耦合边缘降噪自编码器(Coupled Marginalized Auto-Encoders)框架用于跨领域多视图学习. 跨领域经常出现的问题是源域和目标域数据具有不同的分布[76]，为解决这个问题，通常是寻找一个中间域作为桥梁来建立两者的联系. 该方法设计了两种类型的边缘降噪自编码器，一个用于目标域，另一个用于源域和中间域. 为了更好地结合两个自编码器，该模型还采用了一个特征映射方案以缓解不同领域之间的分歧因素. 此外，Hayat 等人[77]、Cheng 等人[78]还将自编码器和投票策略相结合用于图像集分类. 前者将 LBP(Local Binary Patterns)，PCA 与增加了权重衰减项的稀疏自编码器相结合，然后利用高斯限制波兹曼机对网络参数进行预训练，最后使用投票策略进行分类，实验结果显示，在 Honda、Mobo、YTC、PubFig 和 ETH 数据集上的准确率分别达到了 100%、98.33%、72.55%、89.9%和 98.25%. 而后者则重点考虑了图像集分类问题中的两个关键性问题：类内多样性和类间相似性，并在此基础上提出了一种双度量学习模型. 该模型在区分堆叠自编码器的基础上，给

隐层神经元逐层强加一个度量学习正则项以获取新的特征映射,使相似样本在映射空间距离更近,相异样本距离更远.然后训练一个分类器并通过优化一个目标函数来微调区分堆叠自编码网络,该目标函数包含一个分类误差项和一个度量学习正则化项.最后采用两个简单的投票策略进行最终的分类,在相同的数据集上,准确率分别为 100%、98.8%、82.8%、84.9%和 100%.相比前者,该方法除了第 4 个数据集略有下降,其他准确率都得到了一定程度的提升,这说明对于数据集的分类,类内相似和类间相异是非常重要的特性,构建模型时考虑这个特性对数据集的分类效果有很大帮助.

以上图像分类任务存在一个共同的缺点就是网络中的超参数都需要通过手工方式设置,这对研究者的专业技能有比较高的要求,而且结果不一定是最优的.因此针对这个问题,Sui 等人[79]首次将边缘降噪自编码器与粒子群优化算法相结合,用于网络中最优参数的自动选择.实验结果显示,在不使用任何先验知识和经验进行参数手工设置的前提下,由网络自动选择的参数在 MNIST 数据库上达到了 98.7%的准确率.

通过分析以上文献的实验结果我们可以发现以下规律:分类效果最好的是对自编码器损失函数有所改进的方法,即根据研究领域先验知识增加各种惩罚项,其次是将传统方法与未经改进的自编码器相结合的方法,即将比较优秀的传统特征提取方法与自编码器相结合共同用于特征提取,而传统特征提取方法的优劣会对最终结果产生直接的影响,效果最差的是仅采用未改进的自编码器的方法,在这类方法中,堆叠自编码器的效果要优于单个自编码器,因为堆叠自编码器可以提取图像更多的深层特征,对分类效果有很大帮助,此外,在同等条件下,由于卷积自编码器考虑了图像的二维空间信息,对图像的特征提取效果普遍要优于其他自编码器.这与我们所预期的结果是一致的,因为分类任务中最重要的就是特征提取部分,自编码器特征提取的效果是优于传统特征学习方法的,因此,针对研究领域对自编码器进行改进,必定可以达到令人满意的效果,尽管传统方法与未经改进的自编码器的结合可以达到一定的补充效果,但在模型中起主导作用的仍然还是自编码器.

(2)其它数据分类

除了研究比较多的图像分类,自编码器还被用于 3D 模型分类[80]、无线信号分类[81]、AMR(Adaptive Multi-Rate)音频分类[82]、软件分类[83]、大数据分类[62,84]等.其中文献[80]将卷积自编码器与 ELM 相结合,在 ModelNet10 和 ModelNet40 数据库上的准确率达到了 91.41%和 84.35%,为了进一步证明模型的效果,还在 MNIST 和 Norb 库上进行了验证,准确率达到了 98.87%和 94.5%,说明该模型不仅适用于 3D 数据,在 2D 图像上也具有一定的推广性.

5.2 异常检测

在很多研究领域中,相对于正常情况而言,人们关注的往往是异常情况的发生,所谓的异常情况是指不符合期望行为的数据[85],而异常检测技术就是用于发现这些异常情况.按照学习方法的不同,异常检测技术一般可分为有监督和无监督.有监督方法需要通过手工方法标记大量的行为序列以获取足够的训练样本[86],非常的浪费人力物力.因此基于无监督的异常检测受到了广泛地关注.而基于自编码器的异常检测由于其良好的性能,被广泛地应用于视频异常检测、故障检测[87,88]、网上欺凌信息检测[89]等,其中视频异常检测尤为突出.

通过对大量文献的研究,我们发现基于自编码器的视频异常检测也具有一个比较类似的框架,即先利用自编码器去学习正常视频帧的各种特征,建立正常行为模型,然后在测试阶段,重建误差比较高的视频帧就被认为包含异常行为.因此,正常视频帧的特征学习成为了其中非常关键的步骤.

文献[21-22]都是直接利用上面提到的框架进行异常检测,不同之处在于前者使用的训练网络为卷积自编码器,而后者利用了两个自编码器:传统自编码器和稀疏自编码器,去分别训练视频正常块,对于传统自编码器,测试阶段异常块的重建误差会高于正常块,而对于稀疏自编码器,表达不够稀疏的块就会被认为是异常块的候选,实验结果显示,在 UMN 数据库 3 个场景的平均 AUC(Area Under the roc Curve)达到了 99.6%,相比前者的 81%有非常显著的提升.效果提升的主要原因在于后者将两个自编码器串联使用,具有很好的互补作用,即使前面存在漏检现象,第 2 个自编码器也可以将其检测出来.

而文献[23-24]认为仅利用自编码器去进行正常视频的特征学习性能还不够,因此提出将效果比较好的手工特征提取方法与卷积自编码器相结合.前者首先利用方向梯度直方图(HOG)和方向光流直方图(Histogram of Oriented Optical Flow,

HOF)提取正常视频的时空局部特征，然后将提取的特征送入传统自编码器进行训练，在 Avenue、UCSD Ped1 和 Ped2 数据集上的 AUC 分别为 70.2%、81%和 90%. 后者利用 Canny 边缘检测算子和光流(Optical Flow)算法分别提取正常视频帧的外貌和运动特征，然后将当前正常视频帧与两种特征一起送入卷积自编码器进行训练，在相同数据集上，其 AUC 分别为 77.2%、89.5%和 84.7%. 实验结果说明，由于不同视频的特征存在较大差异，不存在一种能使所有数据库都达到最佳效果的模型，因此研究时应该针对视频内在特征的不同选择合适的特征提取方法，以提高模型的整体性能.

文献[25-26]虽然也采用了类似的框架，但与前面的方法有较大差异，前者利用了 3 个堆叠稀疏降噪自编码器分别去学习视频的外观、运动以及两者的融合特征，然后使用多个 SVM 分类器去预测每种特征的异常分数，最后将 3 个分数进行后期融合来完成异常检测，该方法在 UCSD Ped1 和 Ped2 上的 AUC 分别达到了 92.1%和 90.8%. 同样是提取外观和运动特征，相比文献[80]，该方法性能有较大提升，再次说明相比手工特征提取算法，自编码器提取的特征更加优秀. 后者则将稀疏自编码器与深度卷积网络相结合，首先利用一个训练好的 2 层堆叠稀疏自编码器网络快速过滤掉很多正常块，留下少量疑似异常块，然后再利用一个深度卷积神经网络进行进一步的检测，在相同的数据集上的 AUC 分别为 90.9%和 91.8%.

循环神经网络由于其自身的特点，非常适合于处理各种序列数据，比如音频[90]、视频[91-92]、自然语言[93]等. 回声状态网络[94]是一种隐层具有稀疏连接的循环神经网络，由于其隐层神经元的生成过程与回声状态网络的训练过程相互独立，因此只要采用线性方法就可以对隐层到输出层的权重进行训练，极大的简化了网络训练过程，同时保证了全局最优性以及良好的泛化能力[95]，因此 Suh 等人[96]将回声状态网络与变分自编码器相结合用于序列数据的异常检测. 而 Lu 等人[90]则首先对降噪自编码器进行扩展，然后将训练好的降噪自编码器集成到循环神经网络中用于无监督视频异常点检测.

前几种方法都没有对自编码器进行改进，Yuan 等人[92]提出在稀疏降噪自编码器的损失函数上增加一个梯度差信息约束项，用于强化输入与输出之间的局部相似性来提升自编码器的解码效果，这个梯度差信息同时考虑了图像在输入层和输出层上水平和垂直方向的灰度梯度图像的梯度差图 $\boldsymbol{D}$ 和 $\boldsymbol{D}'$，并将它作为惩罚项加入到损失函数中，具体可描述为

$$J(\boldsymbol{W},\boldsymbol{b})=\sum(L(\boldsymbol{x},\boldsymbol{y}))+\beta\sum_{j=1}^{h}KL(\rho\,\|\,\hat{\rho}_j)+\lambda_1\boldsymbol{W}+\lambda_2(\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{A}\hat{\boldsymbol{x}}\|^2) \tag{63}$$

其中 $\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{A}\hat{\boldsymbol{x}}\|^2=(vec(\boldsymbol{D}-\boldsymbol{D}')^{\mathrm{T}}vec(\boldsymbol{D}-\boldsymbol{D}'))$.

该方法在 UMN 数据集上的平均 AUC 为 98.8%，其中场景 2 的 AUC 高达 99.8%. 通过对以上方法实验结果的分析，我们可以得出与图像分类非常相似的结论，模型中最重要的还是自编码器的性能，因此，方法的改进应该集中在对自编码器性能的提升上. 此外，多个自编码器的串联或并联使用能有效提升模型的性能.

5.3　模式识别

(1) 人脸识别

文献[27-28]都将自编码器用于单样本人脸识别. 对于单样本问题，最主要的问题是缺少足够的样本，因此 Gao 等人[27]提出了一种有监督自编码器方法，首先通过一个有监督自编码器使不规则人脸与他相应的标准人脸建立一个对应关系，同时使同一人脸的特征彼此相似，然后堆叠这个有监督自编码器用于特征提取. 这个模型对光照，姿势，表达的变化具有很强的鲁棒性，在扩展的 Yale B 和 AR 上的准确率分别为 82.22%和 85.21%. 而 Zhang 等人[28]通过深度自编码器将多样本推广到单样本，同时重建新的样本. 首先利用一个深度自编码器训练所有的样本数据，然后利用单个样本对一个特殊类深度自编码器进行微调，在同样的数据库上的准确率分别达到了 79.66%和 94.68%. 通过对两个数据库的分析发现，第 1 个模型对于不同角度的人脸具有比较好的识别效果，而第 2 个模型的优势在于不同尺度的人脸图像识别.

而 Xu 等人[29]则提出将两个传统自编码器与两个浅层神经网络相结合用于人脸识别和检索的方法. 该网络通过传统自编码器去重建同一个人不同年龄段的一对图像，使用非线性因素分析方法对传统自编码器的隐层表达进行分解，获取身份特征、年龄特征和噪声；然后利用 2 个浅层神经网络连接两个传统自编码器分别拟合老化和去老化过程，以达到具有年龄不变性的人脸识别和检索，在 FGNET 和 CACD-VS 上的识别率达到了 86.5%和 92.3%.

(2) 情感识别

语音情感识别主要包括语音情感特征提取、特征降维和识别三部分[97]，其中特征提取的质量直接

影响到最后的识别效果.因此,文献[98-99]都将自编码器用于提取语音数据的情感特征.同时,语音数据由于其获取设备和条件的不同,比如来自不同演讲者、音响环境、语音内容和领域条件等,会造成训练和测试样本之间存在内在差异,而这也会给情感识别带来非常不利的影响.所以 Deng 等人在 2014 年[98]和 2017 年[99]先后针对这个问题提出了改进措施.前者提出一种基于无监督领域自适应的模型:自适应降噪自编码器,它利用从目标域数据集上学到的先验知识去调整源域数据集的训练,为目标域和源域数据集获取一个相匹配的特征空间表达,同时确保目标域知识的迁移,在 ABC 和 SUSAS 数据集上的未加权平均召回率(Unweighted Average Recall,UAR)分别为 64.41%和 63.01%.而后者也提出了一种基于无监督领域自适应的模型:Universum 自编码器,不同之处在于,它不仅要从有标签数据中学习可区分的信息,还要将从无标签数据中学习到的先验知识合并到学习中去提升当训练条件与测试条件不匹配情况下的系统性能.在相同数据库上,UAR 大约有 0.5%的提升.除了可以利用语音数据进行情感识别,文献[100]还利用脑电波信号进行情感识别.由于情感的变化与时间存在很大的关系,因此该方法引入了 LSTM,在多模信号情感识别中考虑时间信息,提出一个双模 LSTM 模型,并对降噪自编码器进行扩展,利用一个双模深度降噪自编码器对其进行建模.

5.4 数据生成

在深度学习中,目前比较流行的数据生成基础模型有变分自编码器和生成式对抗网络(Generative Adversarial Networks,GAN)[8].和变分自编码器一样,生成式对抗网络也出现了很多改进模型,其中比较著名的有条件生成式对抗网络(Conditional GAN)[101],深度卷积生成式对抗网络(Deep Convolutional GAN)[102]以及 2017 年最新提出的 Wasserstein 生成式对抗网络(Wasserstein GAN)[103-104].作为两大主流数据生成模型,他们到底有什么区别?其实两者在数据生成时具有非常相似的部分,比如都会利用符合$\mathcal{N}(0,\boldsymbol{I})$分布的噪声,但是两者又存在很大的差异,变分自编码器通过在编码阶段增加一个对分布的约束来使解码器产生与训练数据相似的输出,而生成式对抗网络则是另外再构造一个区分网络去分辨生成数据与真实数据之间的差异.另外两者的关注点也不一样,变分自编码器本质上还是对数据进行压缩,同时用于数据生成,而生成式对抗网络的目的是还原真实数据的分布,通过采样来完成数据的生成.在理论方面,变分自编码器利用了变分推论的近似,而生成式对抗网络不需要这种近似.除此之外,变分自编码器由于具有编码和解码部分,因此可以直接对重建数据和原始数据进行比较,这是生成式对抗生成网络无法做到的.

目前,利用变分自编码器及其改进模型生成的数据类型有图像[30-31]、音乐[32]和自然语言[33]等.

Sun 等人[30]将变分自编码器和循环神经网络相结合用于多数位图像合成.模型中的变分自编码器成功地从给定数量的连续输入中产生了多种多样的多数位图像,并保持了修复不同类型背景的泛化能力.Sabathé 等人[32]利用变分自编码器去生成音乐数据,只需要给定某个作家或某种音乐的风格,就可以生成类似风格的音乐片段.

前几种方法都是直接应用变分自编码器,没有针对研究领域进行任何改进.而 Yan 等人[31]结合研究领域,将条件变分自编码器中的输出变量改为视觉属性.该模型将图像分为前景和背景图像,提出了一个新的图像产生模型——分离条件变分自编码器(Disentangling CVAE).与条件变分自编码器的区别在于,由于该方法将图像分解为前景和背景图像,因此隐变量 $\boldsymbol{z}$ 也被分解为前景和背景隐变量 $\boldsymbol{z}_F$ 和 $\boldsymbol{z}_B$,这两个隐变量分别通过两个自编码器去近似获得.在图像生成阶段,首先通过背景变量去产生完整的背景,然后将视觉属性与前景隐变量相结合,利用一个解码器产生前景图像以及用于决定背景可见度的形状图谱,最后利用另一个解码器产生最终的完整图像.其生成模型如图 10 所示.

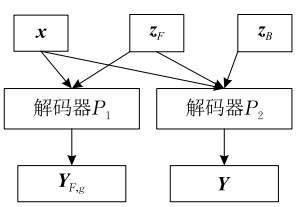

图 10 分离条件变分自编码器生产模型

$\boldsymbol{x}$ 是额外引入的视觉属性,条件变分自编码器中的 $\boldsymbol{z}$ 对应于该模型中的 $\boldsymbol{z}=[\boldsymbol{z}_F,\boldsymbol{z}_B]$,$\boldsymbol{Y}_F$ 是生成的前景图像,$g\in[0,1]$是用于确定背景可见度的一个控制函数.因此最终生成的图像由以下式(64)获得

$$\boldsymbol{Y}=\boldsymbol{Y}_F+\boldsymbol{Y}_B\odot g \tag{64}$$

文献[105]中提到,在获取隐变量 $\boldsymbol{z}$ 时,相比生成式对抗网络,变分自编码器在编码阶段所采用的分布约束效果更好,而在区分重建数据和原始数据

方面，生成式对抗网络又优于变分自编码器. 因此 Bao 等[105]提出将条件变分自编码器的编码器与条件生成式对抗网络相结合，提出一种变分生成式对抗网络用于生成指定类型的高分辨率图像. 与条件生成式对抗网络相比，该模型将由编码器产生的隐变量和对应的条件信息一起送入生成网络，这个条件信息可以是类标签或属性值，然后将生成的图像送入判别网络和分类器. 其结构如图 11 所示.

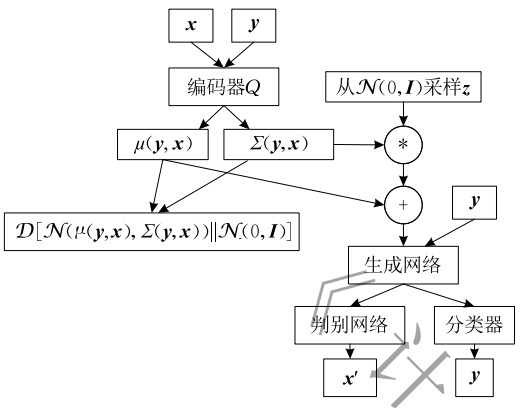

图 11　变分生成式对抗网络结构图

5.5　产品推荐

协同过滤技术[106]的提出使推荐系统真正受到了大家的关注. 协同过滤的主要理念[107]是相似用户之间具有相似的偏好，主要思想是首先对用户-产品评分矩阵(Rating Matrix)$\boldsymbol{R}$进行矩阵因式分解，来学习用户和产品的特征表达，然后基于相似性计算结果进行产品推荐，其中矩阵因式分解是协同过滤方法中的核心步骤[108]. 此外还有一些改进的方法，通常是引入一些额外的支持信息，比如用户与产品属性之间的社会关系，作为规则化去丰富特征表达的先验知识. 但是由于评分矩阵中的评分数据通常远远小于产品数量，因此会出现数据稀疏的问题，而一些新出现的用户或产品，由于缺乏评分数据，又存在冷启动问题[107].

因此为了克服这些困难，有些文献提出采用自编码器去自动学习用户和产品的特征表达. 据分析，这些方法存在一个共同点，那就是都采用两个独立的堆叠自编码器网络去分别学习用户和产品的特征$\boldsymbol{U}$和$\boldsymbol{V}$，然后通过$\boldsymbol{R}'\approx\boldsymbol{U}\times\boldsymbol{V}$获得一个新的评分矩阵用于后续工作，其基本框架如图 12 所示，而不同点在于其中采用的堆叠自编码器不同.

文献[109-111]都是使用上述框架进行用户和产品特征学习，分别采用了两个堆叠传统自编码器，两个堆叠降噪自编码器和两个堆叠边缘降噪自编码器去训练用户和产品信息，获取它们的特征表达. 第 2 种方法没有对自编码器进行修改，直接采用了原始的损失函数进行训练，在 MovieLens 1M 和 MovieLens 100k 数据集上的 RMSE(Root Mean Square Error)分别为 0.8498 和 0.8929. 而第 1 种和第 3 种方法为了进一步提升系统性能，分别对整体框架的损失函数进行了改进. 第 1 种方法在损失函数上增加了两个堆叠传统自编码器的权重和偏置衰减项，在相同数据库上的 RMSE 为 0.8474 和 0.9114. 而第 3 种方法则比较复杂，该模型提出的损失函数为

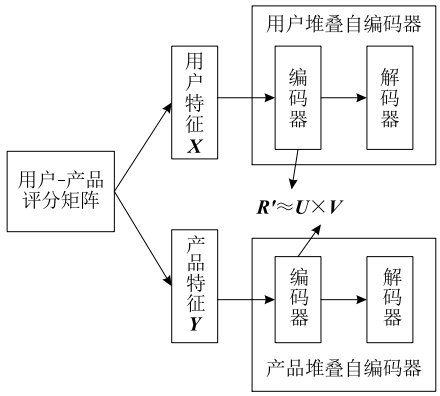

图 12　产品推荐框架

$$J=l(\boldsymbol{R},\boldsymbol{U},\boldsymbol{V})+\beta(\|\boldsymbol{U}\|_F^2+\|\boldsymbol{V}\|_F^2)+\gamma\mathcal{L}_U(\boldsymbol{X},\boldsymbol{U})+\delta\mathcal{L}_V(\boldsymbol{Y},\boldsymbol{V})) \tag{65}$$

第 1 项是为了使新生成的评分矩阵与原始评分矩阵之间的误差尽可能小，第 2 项是对$\boldsymbol{U},\boldsymbol{V}$的$F$范数约束，目的和传统正则自编码器中的权重衰减项一样，最后两项分别对应的是两个堆叠边缘降噪自编码器的损失函数，具体描述为

$$\mathcal{L}_U(\boldsymbol{W}_1,\boldsymbol{P}_1,\boldsymbol{U})=\|\boldsymbol{X}-\boldsymbol{W}_1\tilde{\boldsymbol{X}}\|_F^2+\lambda\|\boldsymbol{P}_1\boldsymbol{U}^{\mathrm{T}}-\boldsymbol{W}_1\boldsymbol{X}\|_F^2 \tag{66}$$

$$\mathcal{L}_V(\boldsymbol{W}_2,\boldsymbol{P}_2,\boldsymbol{V})=\|\boldsymbol{Y}-\boldsymbol{W}_2\tilde{\boldsymbol{Y}}\|_F^2+\lambda\|\boldsymbol{P}_2\boldsymbol{V}^{\mathrm{T}}-\boldsymbol{W}_2\boldsymbol{Y}\|_F^2 \tag{67}$$

$\boldsymbol{X},\boldsymbol{Y}$为对应的用户和产品矩阵，由于采用了边缘降噪自编码器，因此输入信号为$\boldsymbol{X}$、$\boldsymbol{Y}$的损坏矩阵$\tilde{\boldsymbol{X}}$、$\tilde{\boldsymbol{Y}}$，$\boldsymbol{P}_1$、$\boldsymbol{P}_2$是两个低维映射矩阵. 实验结果显示在相同数据库上，该模型的 RMSE 分别达到了 0.8304 和 0.8849. 实验再次证明，针对研究领域先验知识所改进的自编码器能取得最好的结果.

文献[112]虽然也使用了类似的整体框架，但它与前面三种方法有较大的不同. 该方法使用连续限制玻尔兹曼机对网络参数进行预训练，然后利用用户之间的相似度以及用户彼此的信赖程度对新的评分矩阵以及损失函数进行改进，新的评分矩阵获取

方式具体描述为

$$\boldsymbol{R}'_{a,i}=\boldsymbol{U}_a^{\mathrm{T}}\boldsymbol{V}_i+\sum_{b\in S(a)}T_{a,b}(\boldsymbol{U}_b^{\mathrm{T}}\boldsymbol{V}_i+\Delta_{a,b}-\boldsymbol{U}_a^{\mathrm{T}}\boldsymbol{V}_i)+avg+bias_a+bias_b \tag{68}$$

其中 $T_{a,b}=\dfrac{sim(a,b)\times trust(a,b)}{\sum_{s\in S_a}sim(a,s)\times trust(a,s)}$ 是模型的信任度，$sim(a,b)$ 为用户 a,b 之间的相似性，$trust(a,b)$ 表示用户 a 对 b 的信任值. 改进的损失函数具体表示为

$$J=l(\boldsymbol{R},\boldsymbol{U},\boldsymbol{V})+\frac{\lambda_1}{2}\|\boldsymbol{U}\|_F^2+\frac{\lambda_2}{2}\|\boldsymbol{V}\|_F^2+\frac{\beta}{2}\sum_{a=1}^{m}\left\|\boldsymbol{U}_a-\frac{1}{|N(a)|}\sum_{b\in N(a)}T_{a,b}\boldsymbol{U}_b\right\|_F^2 \tag{69}$$

不难发现，该方法与文献[110]最本质的区别在于损失函数所增加的正则项. 式(69)中的正则项的目的是将用户 a 和与它相邻的用户之间的偏好最小化到平均水平. 即用户 a 与它所有邻居的整体偏好应该相似. 该方法在 Epinions 和 Flixster 数据库上的 RMSE 分别为 1.0736 和 0.7853.

5.6 其它应用

除了以上几类研究比较多的应用领域，自编码器还被应用于图像重建[113-114]、图像配准[115]、人脸对齐[116-117]、数据增强[118-120]、目标检测[121-122]、目标追踪[123]、血管分割[124-125]、数字水印[126]、股票市场预测[127]等.

通过对以上各研究领域大量文献的研究，我们发现各种改进方法的提出可归纳为以下几个方向：(1)将比较优秀的传统特征提取方法与各种自编码器相结合，共同用于特征的提取；(2)结合各研究领域的先验知识，对自编码器的损失函数进行修改，一般都是通过增加能反映该先验知识的惩罚项来达成；(3)将各种自编码器与一些比较好的方法相结合，比如 ELM、k 最近邻、GAN 等，以提升模型的整体效果；(4)将两个自编码器进行组合，并通过一些方法建立两个自编码器之间的联系.

6 总 结

近年来，深度学习在各领域取得的巨大成功，受到了广泛关注. 自编码器作为深度学习中无监督学习的代表，由于其所具有的良好的特征学习能力以及可以处理大量无标签数据、节省人力物力的优点，受到了很多研究者的青睐. 本文对自编码器及其改进方法进行了详细阐述，重点介绍了各种改进自编码器的基本理论以及特点. 然后结合各种应用领域，对各种自编码器的应用进行了分类梳理和总结.

尽管自编码器得到了非常广泛的应用，但还存在很多有待解决的问题：(1)作为无监督学习中的一种，由于没有标签数据的辅助，所以在实际应用中的性能与有监督学习还存在一定差距；(2)对于自编码器自身，仍然需要贪婪地训练每一层；(3)由于采用的是逐层训练方式，因此只能达到局部最优；(4)由于梯度扩散问题，单个自编码器网络内部的隐层数量无法达到很深；(5)训练一个自编码器需要大量的数据，耗费很长时间，同时模型中存在很多超参数，很难设置这些参数的取值.

针对自编码器以上存在的问题，我们认为未来的研究方向可以归纳为以下几点：

(1)大部分基于自编码器的方法都存在训练时间过长的问题，那么想要选取合适的网络参数就需要耗费很长的时间，同时对硬件设施也提出了比较高的要求. 针对这个问题，可以考虑从以下方面解决：① 利用小样本进行网络训练. 无论对于无监督学习还是有监督学习，实现小样本训练网络是必然的趋势；这样可以极大地节省人力物力，降低网络训练时间；② 利用分布式优化算法来减少模型的计算复杂度[111]；③ 效仿 k-稀疏自编码器和领域适应性边缘降噪自编码器，对自编码器自身算法进行改进.

(2)目前模型中超参数的设置主要依靠经验，需要通过反复试验来确定其性能优劣，而且当存在多个超参数时，情况更加复杂，即使单个参数达到了最优效果，但是对多个参数进行组合之后，模型的效果并不能保证是最优的. 在我们收集的文献中，目前只有一篇是利用优化算法对超参数进行自动学习的. 因此，将性能卓越的优化算法和自编码器网络相结合，自动学习网络超参数是一个发展趋势.

(3)如何对自编码器网络参数进行有效地初始化. 当前研究方法大部分采用的是随机初始化的方法，这会给网络产生额外的噪声，虽然可以通过在损失函数中增加 L_1 范数来移除噪声的影响，但是如何利用有效的手段将网络参数初始化到一个相对满意的值仍是一个值得研究的问题.

(4)实现真正的无监督学习. 通过对大量文献的研究发现，目前绝大部分基于自编码器的方法最后都需要增加一个有监督学习步骤对网络参数进行微调，而该步骤会受到反向传播算法中梯度消失的影响，因此如何实现真正的无监督学习值得思考.

(5)相反的，对于作为生成模型的变分自编码

器，则可以考虑采用有监督模型，来学习更加复杂的噪声分布.另外，目前大部分生成模型可以生成图像、音频等，而视频生成仍然是一个难题，这也是值得研究的一个方向.

参 考 文 献

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444

[2] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504-507

[3] Yang Sheng-Chun, Jia Lin-Xiang. Comparison between supervised learning and unsupervised learning in neural networks. Journal of Xuzhou Institute of Architectural Technology, 2006, 6(3): 55-58(in Chinese)
(杨盛春，贾林祥. 神经网络内监督学习和无监督学习之比较. 徐州建筑职业技术学院学报, 2006, 6(3): 55-58)

[4] Fischer A, Igel C. An introduction to restricted Boltzmann machines//Proceedings of the Lberoamerican Congress on Pattern Recognition. Berlin, Germany, 2012: 14-36

[5] Zhang Chun-Xia, Ji Nan-Nan, Wang Guan-Wei. Restricted Boltzmann machine. Chinese Journal of Engineering Mathematics, 2015, 32(2): 159-173(in Chinese)
(张春霞，姬楠楠，王冠伟. 受限玻尔兹曼机. 工程数学学报, 2015, 32(2): 159-173)

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

[7] Bourlard H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics, 1988, 59(4): 291-294

[8] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

[9] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527-1554

[10] Wang Ya-Si, Yao Hong-Xun, Zhao Si-Cheng. Auto-encoder based dimensionality reduction. Neurocomputing, 2016, 184: 232-242

[11] Chen Yu-Shi, Lin Zhou-Han, Zhao Xing, et al. Deep learning-based classification of hyperspectral data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014, 7(6): 2094-2107

[12] Cheng Gong, Zhou Pei-Cheng, Han Junwei, et al. Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images. IET Computer Vision, 2015, 9(5): 639-647

[13] Li Er-Zhu, Du Pei-Jun, Samat A, et al. Mid-level feature representation via sparse autoencoder for remotely sensed scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(3): 1068-1081

[14] Geng Jie, Fan Jian-Chao, Wang Hong-Yu, et al. High-resolution SAR image classification via deep convolutional autoencoders. IEEE Geoscience and Remote Sensing Letters, 2015, 12(11): 2351-2355

[15] Geng Jie, Wang Hong-Yu, Fan Jian-Chao, et al. Deep supervised and contractive neural network for SAR image classification. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(4): 2442-2459

[16] Kemker R, Kanan C. Self-taught feature learning for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(5): 2693-2705

[17] Lv Fei, Han Min, Qiu Tie. Remote sensing image classification based on ensemble extreme learning machine with stacked autoencoder. IEEE Access, 2017, 5: 9021-9031

[18] Tao Chao, Pan Hong-Bo, Li Yan-Sheng, et al. Unsupervised spectral-spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification. IEEE Geoscience and Remote Sensing Letters, 2015, 12(12): 2438-2442

[19] Wu Hang, Liu Bao-Zhen, Su Wei-Hua, et al. Deep filter banks for land-use scene classification. IEEE Geoscience and Remote Sensing Letters, 2016, 13(12): 1895-1899

[20] Hou Biao, Kou Hong-Da, Jiao Li-Cheng. Classification of polarimetric SAR images using multilayer autoencoders and superpixels. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2016, 9(7): 3072-3081

[21] Gutoski M, Aquino N M R, Ribeiro M, et al. Detection of video anomalies using convolutional autoencoders and one-class support vector machines//Proceedings of the XIII Brazilian Congress on Computational Intelligence. Rio de Janeiro, Brazil, 2017: 1-12

[22] Sabokrou M, Fathy M, Hoseini M. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. Electronics Letters, 2016, 52(13): 1122-1124

[23] Hasan M, Choi J, Neumann J, et al. Learning temporal regularity in video sequences//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 733-742

[24] Ribeiro M, Lazzaretti A E, Lopes H S. A study of deep convolutional auto-encoders for anomaly detection in videos. Pattern Recognition Letters, 2018, 105: 13-22

[25] Xu Dan, Yan Yan, Ricci E, et al. Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding, 2017, 156: 117-127

[26] Sabokrou M, Fayyaz M, Fathy M, et al. Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Transactions on Image Processing, 2017, 26(4): 1992-2004

[27] Gao Sheng-Hua, Zhang Yu-Ting, Jia Kui, et al. Single sample face recognition via learning deep supervised autoencoders. IEEE Transactions on Information Forensics and Security, 2015, 10(10): 2108-2118

[28] Zhang Yan, Peng Hua. Sample reconstruction with deep autoencoder for one sample per person face recognition. IET Computer Vision, 2017, 11(6): 471-478

[29] Xu Chen-Fei, Liu Qi-He, Ye Mao. Age invariant face recognition and retrieval by coupled auto-encoder networks. Neurocomputing, 2017, 222: 62-71

[30] Sun Hao-Ze, Xu Wei-Di, Deng Chao, et al. Multi-digit image synthesis using recurrent conditional variational autoencoder//Proceedings of the International Joint Conference on Neural Networks. Vancouver, Canada, 2016: 375-380

[31] Yan Xin-Chen, Yang Ji-Mei, Sohn K, et al. Attribute2image: Conditional image generation from visual attributes//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 776-791

[32] Sabathé R, Coutinho E, Schuller B. Deep recurrent music writer: Memory-enhanced variational autoencoder-based musical score composition and an objective measure//Proceedings of the International Joint Conference on Neural Networks. Anchorage, USA, 2017: 3467-3474

[33] Shen Xiao-Yu, Su Hui, Niu Shu-Zi, et al. Wake-sleep variational autoencoders for language modeling//Proceedings of the International Conference on Neural Information Processing. Guangzhou, China, 2017: 405-414

[34] Rifai S, Vincent P, Muller X, et al. Contractive auto-encoders: Explicit invariance during feature extraction//Proceedings of the 28th International Conference on Machine Learning. Washington, USA, 2011: 833-840

[35] Kasun L L C, Yang Yan, Huang Guang-Bin, et al. Dimension reduction with extreme learning machine. IEEE Transactions on Image Processing, 2016, 25(8): 3906-3918

[36] Liu Wei-Feng, Ma Teng-Zhou, Xie Qiang-Sheng, et al. LMAE: A large margin auto-encoders for classification. Signal Processing, 2017, 141: 137-143

[37] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders//Proceedings of the 25th International Conference on Machine learning. Helsinki, Finland, 2008: 1096-1103

[38] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 2010, 11(Dec): 3371-3408

[39] Chen Min-Min, Xu Zhi-Xiang, Weinberger K, et al. Marginalized denoising autoencoders for domain adaptation//Proceedings of the International conference on machine learning. Edinburgh, Scotland, 2012: 1627-1634

[40] Chen Min-Min, Weinberger K, Sha Fei, et al. Marginalized denoising auto-encoders for nonlinear representations//Proceedings of the 31st International Conference on Machine Learning. Beijing, China, 2014: 1476-1484

[41] LeCun Y, Bottou L, Orr G B, et al. Efficient BackProp in Neural Networks: Tricks of the Trade. Berlin, Germany: Springer, 1998

[42] Ng A. Sparse autoencoder. CS294A Lecture Notes, 2011, 72(2011): 1-19

[43] Makhzani A, Frey B. K-sparse autoencoders//Proceedings of the International Conference on Learning Representation. Banff, Canada, 2014: 1-9

[44] Coates A, Ng A Y. The importance of encoding versus training with sparse coding and vector quantization//Proceedings of the 28th International Conference on Machine Learning. Bellevue, USA, 2011: 921-928

[45] Rifai S, Mesnil G, Vincent P, et al. Higher order contractive auto-encoder. Machine Learning and Knowledge Discovery in Databases, 2011: 645-660

[46] Goroshin R, LeCun Y. Saturating auto-encoders//Proceedings of the International Conference on Learning Representation. Scottsdale, USA, 2013: 1-9

[47] Masci J, Meier U, Cireşan D, et al. Stacked convolutional auto-encoders for hierarchical feature extraction//Proceedings of the International Conference on Artificial Neural Networks. Espoo, Finland, 2011: 52-59

[48] Kingma D P, Welling M. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114, 2013

[49] Doersch C. Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908, 2016

[50] Sohn K, Yan Xin-Chen, Lee H. Learning structured output representation using deep conditional generative models//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2015: 3483-3491

[51] Louizos C, Swersky K, Li Y, et al. The variational fair autoencoder. arXiv preprint arXiv:1511.00830, 2015

[52] Zhao Sheng-Jia, Song Jia-Ming, Ermon S. InfoVAE: Information maximizing variational autoencoders. arXiv preprint arXiv:1706.02262, 2017

[53] Ramachandra G. Least square variational Bayesian autoencoder with regularization. arXiv preprint arXiv:1707.03134, 2017

[54] Chen Xi, Kingma D P, Salimans T, et al. Variational lossy autoencoder. arXiv preprint arXiv:1611.02731, 2016

[55] Shang Wen-Ling, Sohn K, Akata Z, et al. Channel-recurrent variational autoencoders. arXiv preprint arXiv:1706.03729, 2017

[56] Cai Lei, Gao Hong-Yang, Ji Shui-Wang. Multi-stage variational auto-encoders for coarse-to-fine image generation. arXiv preprint arXiv:1705.07202, 2017

[57] Hinton G E, Krizhevsky A, Wang S D. Transforming auto-encoders//Proceedings of the International Conference on Artificial Neural Networks. Berlin, Germany, 2011: 44-51

[58] Xie Jin, Fang Yi, Zhu Fan, et al. Deepshape: Deep learned shape descriptor for 3D shape matching and retrieval// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1275-1283

[59] Cui Zhen, Chang Hong, Shan Shi-Guang, et al. Deep network cascade for image super-resolution//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 49-64

[60] Luo Wei, Yang Jian, Xu Wei, et al. Locality-constrained sparse auto-encoder for image classification. IEEE Signal Processing Letters, 2015, 22(8): 1070-1073

[61] Wang Ru-Xin, Tao Da-Cheng. Non-local auto-encoder with collaborative stabilization for image restoration. IEEE Transactions on Image Processing, 2016, 25(5): 2117-2129

[62] Zhang Qing-Chen, Yang L T, Chen Zhi-Kui. Deep computation model for unsupervised feature learning on big data. IEEE Transactions on Services Computing, 2016, 9(1): 161-171

[63] Zhang Lu, Ma Wen-Ping, Zhang Dan. Stacked sparse autoencoder in POLSAR data classification using local spatial information. IEEE Geoscience and Remote Sensing Letters, 2016, 13(9): 1359-1363

[64] Xie Wen, Jiao Li-Cheng, Hou Biao, et al. POLSAR image classification via Wishart-AE model or Wishart-CAE model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(8): 3604-3615

[65] Ma Xiao-Rui, Wang Hong-Yu, Geng Jie. Spectral-spatial classification of hyperspectral image based on deep auto-encoder. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2016, 9(9): 4073-4085

[66] Cheng Gong, Han Jun-Wei, Guo Lei, et al. Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(8): 4238-4249

[67] Deng Sheng, Du Lan, Li Chen, et al. SAR automatic target recognition based on euclidean distance restricted autoencoder. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(7): 3323-3333

[68] Elshamli A, Taylor G W, Berg A, et al. Domain adaptation using representation learning for the classification of remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(9): 4198-4209

[69] Xu Jun, Xiang Lei, Liu Qing-Shan, et al. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. IEEE Transactions on Medical Imaging, 2016, 35(1): 119-130

[70] Zhang Xiao-Fan, Dou Hang, Ju Tao, et al. Fusing heterogeneous features from stacked sparse autoencoder for histopathological image analysis. IEEE Journal of Biomedical and Health Informatics, 2016, 20(5): 1377-1383

[71] Majumdar A, Gogna A, Ward R. Semi-supervised stacked label consistent autoencoder for reconstruction and analysis of biomedical signals. IEEE Transactions on Biomedical Engineering, 2016, 64(9): 2196-2205

[72] Li Yun-Dong, Zhao Wei-Gang, Pan Jia-Hao. Deformable patterned fabric defect detection with Fisher criterion-based deep learning. IEEE Transactions on Automation Science and Engineering, 2017, 14(2): 1256-1264

[73] Ng W W Y, Zeng Guang-Jun, Zhang Jiang-Jun, et al. Dual autoencoders features for imbalance classification problem. Pattern Recognition, 2016, 60: 875-889

[74] Wang Wen, Cui Zhen, Chang Hong, et al. Deeply coupled auto-encoder networks for cross-view classification. arXiv preprint arXiv:1402.2031, 2014

[75] Wang Shu-Yang, Ding Zheng-Ming, Fu Yun. Coupled marginalized auto-encoders for cross-domain multi-view learning// Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA, 2016: 2125-2131

[76] Zhang Bo, Shi Zhong-Zhi, Zhao Xiao-Fei, et al. A transfer learning based on canonical correlation analysis across different domains. Chinese Journal of Computers, 2015, 38(7): 1326-1336(in Chinese)
(张博，史忠植，赵晓非等. 一种基于跨领域典型相关性分析的迁移学习方法. 计算机学报，2015, 38(7): 1326-1336)

[77] Hayat M, Bennamoun M, An S. Deep reconstruction models for image set classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(4): 713-727

[78] Cheng Gong, Zhou Pei-Cheng, Han Jun-Wei. Duplex metric learning for image set classification. IEEE Transactions on Image Processing, 2018, 27(1): 281-292

[79] Sui Chao, Bennamoun M, Togneri R. Deep feature learning for dummies: A simple auto-encoder training method using particle swarm optimisation. Pattern Recognition Letters, 2017, 94: 75-80

[80] Wang Yue-Qing, Xie Zhi-Ge, Xu Kai, et al. An efficient and effective convolutional auto-encoder extreme learning machine network for 3D feature learning. Neurocomputing, 2016, 174: 988-998

[81] Long Fei, Yan Xin. ELM-based signal detection scheme of MIMO system using auto encoder//Proceedings of the International Conference on Neural Information Processing. Guangzhou, China, 2017: 499-509

[82] Luo Da, Yang Rui, Li Bin, et al. Detection of double compressed AMR audio using stacked autoencoder. IEEE Transactions on Information Forensics and Security, 2017, 12(2): 432-444

[83] Kim J Y, Bu S J, Cho S B. Malware detection using deep transferred generative adversarial networks//Proceedings of the International Conference on Neural Information Processing. Guangzhou, China, 2017: 556-564

[84] Kasun L L C, Zhou Hong-Ming, Huang Guang-Bin, et al. Representational learning with ELMs for big data. IEEE Intelligent Systems, 2013, 28(6): 31-34

[85] Chen Bin, Chen Song-Can, Pan Zhi-Song, et al. Survey of outlier detection technologies. Journal of Shandong University (Engineering Science), 2009, 39(6): 13-23(in Chinese)
(陈斌, 陈松灿, 潘志松等. 异常检测综述. 山东大学学报(工学版), 2009, 39(6): 13-23)

[86] Li He-Ping, Hu Zhan-Yi, Wu Yi-Hong, et al. Behavior modeling and abnormality detection based on semi-supervised learning method. Journal of Software, 2007, 18(3): 527-537 (in Chinese)
(李和平, 胡占义, 吴毅红等. 基于半监督学习的行为建模与异常检测. 软件学报, 2007, 18(3): 527-537)

[87] Lee H, Kim Y, Kim C O. A deep learning model for robust wafer fault monitoring with sensor measurement noise. IEEE Transactions on Semiconductor Manufacturing, 2017, 30(1): 23-31

[88] Qi Yu-Mei, Shen Chang-Qing, Wang Dong, et al. Stacked sparse autoencoder-based deep network for fault diagnosis of rotating machinery. IEEE Access, 2017, 5: 15066-15079

[89] Zhao Rui, Mao Ke-Zhi. Cyberbullying detection based on semantic- enhanced marginalized denoising auto-encoder. IEEE Transactions on Affective Computing, 2017, 8(3): 328-339

[90] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks//Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 6645-6649

[91] Lu Wei-Ning, Cheng Yu, Xiao Cao, et al. Unsupervised sequential outlier detection with deep architectures. IEEE Transactions on Image Processing, 2017, 26(9): 4321-4330

[92] Yuan Jing, Zhang Yu-Jin. Application of sparse denoising auto encoder network with gradient difference information for abnormal action detection. Acta Automatica Sinica, 2017, 43(4): 604-610(in Chinese)
(袁静, 章毓晋. 融合梯度差信息的稀疏降噪自编码器网络在异常行为检测中的应用. 自动化学报, 2017, 43(4): 604-610)

[93] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model//Proceedings of the Conference of the International Speech Communication Association. Makuhari, Japan, 2010: 1045-1048

[94] Jaeger H, Haas H. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Science, 2004, 304(5667): 78-80

[95] Guo Jia. Research on Echo State Networks Classification Algorithm and Application [Ph. D. dissertation]. Harbin Institute of Technology, Harbin, 2017(in Chinese)
(郭嘉. 回声状态网络分类方法及其应用研究[博士学位论文]. 哈尔滨工业大学, 哈尔滨, 2017)

[96] Suh S, Chae D H, Kang H G, et al. Echo-state conditional variational autoencoder for anomaly detection//Proceedings of the International Joint Conference on. Neural Network. Vancouver, Canada, 2016: 1015-1022

[97] Liu Zhen-Tao, Xu Jian-Ping, Wu Min, et al. Review of emotional feature extraction and dimension reduction method for speech emotion recognition. Chinese Journal of Computers, 2018, 41(12): 2849-2867(in Chinese)
(刘振焘, 徐建平, 吴敏等. 语音情感特征提取及其降维方法综述. 计算机学报, 2018, 41(12): 2849-2867)

[98] Deng Jun, Zhang Zi-Xing, Eyben F, et al. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. IEEE Signal Processing Letters, 2014, 21(9): 1068-1072

[99] Deng Jun, Xu Xin-Zhou, Zhang Zi-Xing, et al. Universum autoencoder-based domain adaptation for speech emotion recognition. IEEE Signal Processing Letters, 2017, 24(4): 500-504

[100] Tang Hao, Liu Wei, Zheng Wei-Long, et al. Multimodal emotion recognition using deep neural networks//Proceedings of the International Conference on Neural Information Processing. Guangzhou, China, 2017: 811-819

[101] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014

[102] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks//Proceedings of the International Conference on Learning Representation. San Juan, Puerto Rico, 2016

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

[104] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of Wasserstein Gans. Neural Information Processing Systems, 2017: 5767-5777

[105] Bao Jian-Min, Chen Dong, Wen Fang, et al. CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training//Proceedings of the International Conference on Computer Vision. Venice, Italy, 2017: 2764-2773

[106] Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992, 35(12): 61-70

[107] Huang Li-Wei, Liu Yan-Bo, Li De-Yi. Deep learning based recommender systems. Chinese Journal of Computers, 2018, 41(7): 1619-1647(in Chinese)
(黄立威, 刘艳博, 李德毅. 基于深度学习的推荐系统. 计算机学报, 2018, 41(7): 1619-1647)

[108] Sun Guang-Fu, Wu Le, Liu Qi, et al. Recommendation based on collaborative filtering by exploiting sequential behaviors. Journal of Software, 2013, 24(11): 2721-2733 (in Chinese)
(孙光福, 吴乐, 刘淇等. 基于时序行为的协同过滤推荐算法. 软件学报, 2013, 24(11): 2721-2733)

[109] Zhuang Fu-Zhen, Zhang Zhi-Qiang, Qian Ming-Da, et al. Representation learning via dual-autoencoder for recommendation. Neural Networks, 2017, 90: 83-89

[110] Barbieri J, Alvim L G M, Braida F, et al. Autoencoders and recommender systems: COFILS approach. Expert Systems with Applications, 2017, 89: 81-90

[111] Li Sheng, Kawale J, Fu Yun. Deep collaborative filtering via marginalized denoising auto-encoder//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne, Australia, 2015: 811-820

[112] Deng Shui-Guang, Huang Long-Tao, Xu Guan-Dong, et al. On deep learning for trust-aware recommendations in social networks. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(5): 1164-1177

[113] Mehta J, Majumdar A. Rodeo: Robust de-aliasing autoencoder for real-time medical image reconstruction. Pattern Recognition, 2017, 63: 499-510

[114] Zeng Kun, Yu Jun, Wang Ru-Xin, et al. Coupled deep autoencoder for single image super-resolution. IEEE Transactions on Cybernetics, 2017, 47(1): 27-37

[115] Wu Guo-Rong, Kim M, Wang Qian, et al. Scalable high-performance image registration framework by unsupervised deep feature representations learning. IEEE Transactions on Biomedical Engineering, 2016, 63(7): 1505-1516

[116] Zhang J, Shan S, Kan M, et al. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 1-16

[117] Weng R, Lu J, Tan Y P, et al. Learning cascaded deep auto-encoder networks for face alignment. IEEE Transactions on Multimedia, 2016, 18(10): 2066-2078

[118] Sun Meng, Zhang Xiong-Wei, Hamme H V. Unseen noise estimation using separable deep auto encoder for speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(1): 93-104

[119] Lore K G, Akintayo A, Sarkar S. LLNet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition, 2017, 61: 650-662

[120] Dai Jie-Jie, Song Hui, Sheng Ge-Hao, et al. Cleaning method for status monitoring data of power equipment based on stacked denoising autoencoders. IEEE Access, 2017, 5: 22 863-22 870

[121] Su Sheng-Ran, Gao Zhi-Fan, Zhang He-Ye, et al. A detection of lumen and media-adventitia borders in ivus images using sparse auto-encoder neural network//Proceedings of the IEEE 14th International Symposium on Biomedical Imaging. Melbourne, Australia, 2017: 1120-1124

[122] Han Jun-Wei, Zhang Ding-Wen, Hu Xin-Tao, et al. Background prior-based salient object detection via deep reconstruction residual. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(8): 1309-1321

[123] Kuen J, Lim K M, Lee C P. Self-taught learning of a deep invariant representation for visual tracking via temporal slowness principle. Pattern Recognition, 2015, 48(10): 2964-2982

[124] Zhao Bo-Wen, Cao Zhu-Lou, Wang Si-Cheng. Lung vessel segmentation based on random forests. Electronics Letters, 2017, 53(4): 220-222

[125] Fan Zhun, Mo Jia-Jie. Automated blood vessel segmentation based on de-noising auto-encoder and neural network//Proceedings of the International Conference on Machine Learning and Cybernetics. Jeju, South Korea, 2016, 2: 849-856

[126] Haribabu K, Subrahmanyam G, Mishra D. A robust digital image watermarking technique using auto encoder based convolutional neural networks//Proceedings of the IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions. Kanpur, India, 2015: 1-6

[127] Sun Hao-Nan, Rong Wen-Ge, Zhang Jia-Yi, et al. Stacked denoising autoencoder based stock market trend prediction via K-nearest neighbour data selection//Proceedings of the International Conference on Neural Information Processing. Guangzhou, China, 2017: 882-892

YUAN Fei-Niu, born in 1976, Ph.D., professor, Ph.D. supervisor. His research interests include image processing, pattern recognition and 3D visualization.

ZHANG Lin, born in 1982, Ph. D. candidate, associate professor. Her research interests include image processing and pattern recognition.

SHI Jin-Ting, born in 1982, Ph. D. candidate, lecturer. Her research interests include image processing and pattern recognition.

XIA Xue, born in 1990, Ph. D. candidate. Her research interests include image processing and pattern recognition.

LI Gang, born in 1980, Ph. D. candidate, lecturer. His research interests include image processing and pattern recognition.

Background

Auto-encoder is a shallow artificial neural network used for learning effective features. It is one of unsupervised learning methods in deep learning. The aim of an autoencoder is usually to learn an intrinsic representation for data. It should automatically learn efficient features from a large amount of unlabeled data, thus we can avoid boring, troublesome, time-consuming work of labelling data. Hence it has attracted a lot of attention in recent years, and it has successfully been applied in many research fields, such as data classification, pattern recognition, anomaly detection, data generation. However, there are new papers to offer a comprehensive reviewing of auto-encoders.

In order to provide a systematic and comprehensive understanding of auto-encoders for researchers, it is necessary for us to perform a full survey on the basic theory of traditional auto-encoders, related algorithms, improved techniques and several applications in detail. In this paper, we first introduce the basic theory and implementation of auto-encoders, and analyze the general processing framework of auto-encoders. Then, we discuss existing improved methods of auto-encoders, and analyze the innovation, motivation and existing problems of these methods. Afterwards, we introduce the application fields of auto-encoders, and we also analyze, compare and summarize the representative auto-encoders of each application field. Finally, after we point out some issues in existing auto-encoders, we discuss possible developing trends and challenges of auto-encoders.

This work was partially supported by the National Natural Science Foundation of China (61862029), the Science Technology Application Projects of Jiangxi Province (KJLD12066), and the Science Technology Projects of Education Department of Jiangxi Province (GJJ170317).