自编码器及其应用综述

来杰[1]，王晓丹[1]，向前[1]，宋亚飞[1]，权文[2]
（1. 空军工程大学防空反导学院，陕西 西安 710051；2. 空军工程大学空管领航学院，陕西 西安 710051）

摘 要：自编码器作为典型的深度无监督学习模型，能够从无标签样本中自动学习样本的有效抽象特征。近年来，自编码器受到广泛关注，已应用于目标识别、入侵检测、故障诊断等众多领域中。基于此，对自编码器的理论基础、改进技术、应用领域与研究方向进行了较全面的阐述与总结。首先，介绍了传统自编码器的网络结构与理论推导，分析了自编码器的算法流程，并与其他无监督学习算法进行了比较。然后，讨论了常用的自编码器改进算法，分析了其出发点、改进方式与优缺点。接着，介绍了自编码器在目标识别、入侵检测等具体领域的实际应用现状。最后，总结了现有自编码器及其改进算法存在的问题，并展望了自编码器的研究方向。

关键词：自编码器；深度学习；无监督学习；特征提取；正则化
中图分类号：TP183
文献标识码：A
DOI: 10.11959/j.issn.1000−436x.2021160

Review on autoencoder and its application

LAI Jie[1], WANG Xiaodan[1], XIANG Qian[1], SONG Yafei[1], QUAN Wen[2]
1. School of Air and Missile Defense, Air Force Engineering University, Xi’an 710051, China
2. School of Air Traffic Control and Navigation, Air Force Engineering University, Xi’an 710051, China

Abstract: As a typical deep unsupervised learning model, autoencoder can automatically learn effective abstract features from unlabeled samples. In recent years, autoencoder has been widely used in target recognition, intrusion detection, fault diagnosis and many other fields. Thus, the theoretical basis, improved methods, application fields and research directions of autoencoder were described and summarized comprehensively. At first, the network structure, theoretical derivation and algorithm flow of traditional autoencoder were introduced and analyzed, and the difference between autoencoder and other unsupervised learning algorithms was compared. Then, common improved autoencoders were discussed, and their innovation, improvement methods and relative merits were analyzed. Next, the practical application status of autoencoder in target recognition, intrusion detection and other fields were introduced. At last, the existing problems of autoencoder were summarized, and the possible research directions were prospected.

Keywords: autoencoder, deep learning, unsupervised learning, feature extraction, regularization

1 引言

深度学习作为机器学习领域的研究热点，自被 Hinton 等[1]提出后，便深刻影响着机器学习的发展走向。深度学习通过构建多层神经网络模型，逐层提取样本的高级抽象特征，而后通过分类器或回归算法完成抽象特征到期望输出的映射。与浅层神经网络相比，深度学习模型的多层网络结构拥有更强的特征提取能力，避免了传统机器学习算法需要人工选取特征的局限[2]，同时采用贪婪预训练方式，

收稿日期：2021−04−29；**修回日期**：2021−07−29
通信作者：王晓丹，afeu_wxd@163.com
基金项目：国家自然科学基金资助项目（No.61876189, No.61806219, No.61703426）；陕西省自然科学基础研究计划基金资助项目（No.2021JM—226）
Foundation Items: The National Natural Science Foundation of China (No.61876189, No.61806219, No.61703426), The Natural Science Basic Research Plan in Shaanxi Province (No.2021JM—226)

逐层初始化网络参数，加快了网络收敛速度[3]。深度学习模型的优异性能得益于其复杂的网络结构，而复杂的网络结构则需要大量样本进行训练。然而在有监督学习模式下，大量样本标签的人工标注是非常困难的，这促进了无监督深度学习模型的发展[4]。典型的深度无监督学习模型有自编码器（AE, autoencoder）[5-6]、受限波尔兹曼机（RBM, restricted Boltzmann machine）[7-8]与生成对抗网络（GAN, generative adversarial network）[9]。

自编码器作为典型的无监督深度学习模型，旨在通过将网络的期望输出等同于输入样本，实现对输入样本的抽象特征学习。Rumelhart 等[5]最早提出了自编码器的概念，Bourlard 等[6]对其进行了详细的阐释。随着深度学习得到空前的关注，AE 也被广泛研究与改进[10-13]。为获得高维且稀疏的抽象特征表示，Ng[10]通过在隐含层输出中引入稀疏性限制，迫使网络使用较少神经节点提取有效特征，提出了稀疏自编码器（SAE, sparse autoencoder）。Vincent 等[11]提出了去噪自编码器（DAE, denoising autoencoder）。DAE 在 AE 中引入退化过程，运用添加噪声后的样本重构无噪声样本，使提取的抽象特征不易受噪声影响，具有更强的稳健性。为抑制输入样本中的微小扰动，Rifai 等[12]提出了收缩自编码器（CAE, contractive autoencoder）。CAE 通过在 AE 损失函数中添加收缩正则化项，以达到局部空间收缩效果。Kingma 等[13]提出了变分自编码器（VAE, variational autoencoder），并将其用于数据生成。凭借训练过程简单、多层堆栈容易、泛化性能优秀的特点，AE 及其改进算法被成功应用于目标识别[14-15]、入侵检测[16-17]与故障诊断[18-19]等领域。

本文将着重介绍 AE 网络结构及其算法流程，梳理自编码器改进算法的创新点与实现方式。同时，结合最新文献，总结 AE 在多个领域的研究进展。最后，通过分析当前 AE 及其改进算法存在的问题，讨论其进一步的研究方向。

2　自编码器

作为在无监督学习中使用的神经网络，自编码器的输入与期望输出均为无标签样本，而隐含层输出则是样本的抽象特征表示。AE 首先接收输入样本，将其转换成高效的抽象表示，而后再输出原始样本的重构。

2.1　网络结构

自编码器通常包括两部分：编码器和解码器。编码器将高维输入样本映射到低维抽象表示，实现样本压缩与降维；解码器则将抽象表示转换为期望输出，实现输入样本的复现。自编码器结构如图 1 所示。

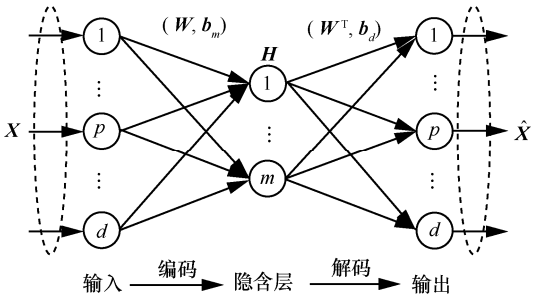

图 1　自编码器结构

与全连接神经网络相同，AE 节点连接方式也为全连接，但由于采用无监督学习范式，AE 的输入输出均为无标签样本，不需要标签信息，旨在学习样本的内在结构，提取抽象特征。传统全连接网络采用有监督学习范式，其输出为样本标签，旨在完成特征到标签的映射。

AE 的训练过程包括编码和解码 2 个阶段。编码过程，对输入样本进行编码得到编码层；解码过程，对编码层进行解码，得到输入样本的重构，并通过调整网络参数使重构误差达到最小值，以获得输入特征的最优抽象表示。

假设给定输入样本 $\boldsymbol{X}=\boldsymbol{R}^{d\times n}$，编码层与输入层间的权值矩阵 $\boldsymbol{W}$，编码层节点偏置 $\boldsymbol{b}_m$，解码层偏置 $\boldsymbol{b}_d$，节点激活函数 $g(\cdot)$，自编码器首先通过线性映射和非线性激活函数完成对样本的编码

$$\boldsymbol{H}=g\left(\boldsymbol{W}\boldsymbol{X}+\boldsymbol{b}_m\right) \tag{1}$$

然后，解码器完成对编码特征的解码，得到输入样本的重构 $\hat{\boldsymbol{X}}$。当给定编码 $\boldsymbol{H}$ 时，$\hat{\boldsymbol{X}}$ 也可以看作对 $\boldsymbol{X}$ 的预测，与 $\boldsymbol{X}$ 的维度相同。解码过程与编码过程类似

$$\hat{\boldsymbol{X}}=g\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}+\boldsymbol{b}_d\right) \tag{2}$$

AE 的训练旨在使损失函数达到最小值

$$\arg\min_{\boldsymbol{W},\boldsymbol{b}} J\left(\boldsymbol{W},\boldsymbol{b}\right) \tag{3}$$

在 AE 中，损失函数通常可取平方误差损失函数或交叉熵损失函数。对于输入样本

$\boldsymbol{X}=\left\{\boldsymbol{x}_i\in\boldsymbol{R}^d\right\}_{i=1}^n$ 与重构 $\hat{\boldsymbol{X}}=\left\{\hat{\boldsymbol{x}}_i\in\boldsymbol{R}^d\right\}_{i=1}^n$，平方误差与交叉熵损失函数分别为

$$J\left(\boldsymbol{X},\hat{\boldsymbol{X}}\right)=\frac{1}{2}\sum_{i=1}^{n}\left\|\hat{\boldsymbol{x}}_i-\boldsymbol{x}_i\right\|_2^2 \tag{4}$$

$$J\left(\boldsymbol{X},\hat{\boldsymbol{X}}\right)=-\sum_{i=1}^{n}\left[\boldsymbol{x}_i\log\left(\hat{\boldsymbol{x}}_i\right)+\left(1-\boldsymbol{x}_i\right)\log\left(1-\hat{\boldsymbol{x}}_i\right)\right] \tag{5}$$

AE 通常利用梯度下降算法，反向传播误差以调整网络参数，通过迭代微调逐步使重构误差函数达到最小值，以学习样本数据中的关键抽象特征。当采用梯度下降算法时，假设学习速率为 η，AE 的连接权值与偏置更新式分别为

$$\boldsymbol{W}=\boldsymbol{W}-\eta\frac{\partial J(\boldsymbol{W},\boldsymbol{b})}{\partial\boldsymbol{W}} \tag{6}$$

$$\boldsymbol{b}=\boldsymbol{b}-\eta\frac{\partial J(\boldsymbol{W},\boldsymbol{b})}{\partial\boldsymbol{b}} \tag{7}$$

研究者通常将多个 AE 进行级联，构建堆栈自编码器，采用逐层贪婪训练方式，将上一级 AE 的隐含层输出作为下一级 AE 的输入，进行层次化特征提取，使最终提出的特征更具代表性，且维数通常较低[20]。堆栈自编码器通常具有对称的多隐含层结构，对应于解码与编码过程，但被应用于分类或回归问题时，一般将解码部分舍去，将最终的编码用于分类[21]或回归[22]。

2.2 算法流程

对于特定的输入样本，自编码器通过特征的正向传播与误差的反向传播，采用梯度下降算法更新网络连接权值与节点偏置，使重构样本逐步逼近输入样本，进而提取样本的抽象特征。算法流程如算法 1 所示。

算法 1 自编码器算法

输入 训练样本 $\boldsymbol{X}=\{\boldsymbol{x}_i\in\boldsymbol{R}^d\}_{i=1}^n$，隐含层节点数 m，激活函数 $g(\cdot)$，学习速率 η，最大迭代次数 $N_{\max}$

输出 映射函数 $f:\boldsymbol{R}^d\to\boldsymbol{R}^d$

1) 初始化，迭代次数 $N=0$，随机赋值连接权值与节点偏置 $\boldsymbol{W}$，$\boldsymbol{b}_m$，$\boldsymbol{b}_d$；

2) while $N<N_{\max}$

3) 更新迭代次数 $N=N+1$；

4) 正向传播，利用式(1)和式(2)计算隐含层输出 $\boldsymbol{H}$ 和样本重构 $\hat{\boldsymbol{X}}$；

5) 利用式(4)或式(5)计算网络误差 $J\left(\boldsymbol{W},\boldsymbol{b}\right)$；

6) 反向传播，利用式(6)和式(7)分别更新权值 $\boldsymbol{W}$ 和偏置 $\boldsymbol{b}_m$、$\boldsymbol{b}_d$；

7) end while

8) 返回映射函数 $f\left(\boldsymbol{X}\right)=g(\boldsymbol{W}^{\mathrm{T}}g(\boldsymbol{W}\boldsymbol{X}+\boldsymbol{b}_m)+\boldsymbol{b}_d)$

以上流程中，算法停止条件仅为最大迭代次数。在实际应用中，还可设定期望误差等其他停止条件。

除 AE 以外，基于神经网络的无监督学习算法还包括 RBM、GAN 与自组织映射（SOM, self-organization mapping）[23]。表 1 归纳了上述无监督学习算法的目的、实现方式与特点。

与主成分分析、独立成分分析等常用无监督学习算法相比，AE 通过将特征进行加权融合，转化为维数更低、更具代表性的高级抽象，能够有效利用次要特征中的重要信息，而不是一味地舍弃次要特征，因此采用 AE 进行特征提取或降维后的抽象特征，更有助于分类与回归任务。与 RBM 相比，AE 不需要对比散度算法中的采样运算，训练时间更短。与 SOM 相比，AE 通过改变隐含层节点数，可以完成任意维度特征空间的映射。

表 1 基于神经网络的无监督学习算法分析与比较

算法	目的	实现方式	特点
AE	特征提取 特征降维	由编码器与解码器组成，前者完成样本的抽象特征表示，后者完成输入样本的重构，通过梯度下降算法完成迭代微调	确定型无监督学习模型，能有效提取高级抽象特征，但所提取特征的稀疏性、稳健性不足
RBM	特征提取 特征降维	由可见层与隐含层组成，旨在最大化可见变量与隐含变量联合概率分布的似然函数，通过对比散度算法完成迭代微调	概率型无监督学习模型，能有效提取高级抽象特征，但易产生过拟合现象
GAN	数据生成	由生成器与判别器组成，旨在通过生成器与判别器的二元博弈提升生成样本质量，网络参数的优化由梯度下降算法完成	无监督生成模型，能生成高质量样本，但易发生模式崩溃问题
SOM	聚类分析 特征提取 特征降维	由输入层与竞争层组成，旨在通过竞争学习策略逐步优化网络，近邻关系函数维持输入样本的拓扑结构，进而完成样本的低维映射	浅层无监督学习模型，具有独特的竞争学习机制，能有效生成样本的低维映射，但不适用于高维特征的提取

3　自编码器的改进

传统自编码器仅通过最小化输入样本与重构样本之间的误差来获取输入样本的抽象特征表示，这可能导致自编码器学习到的特征仅仅是原始输入的恒等表示，不能保证提取到样本的本质特征。为避免上述问题，需要对传统自编码器添加约束或修改网络结构，进而产生了 SAE、DAE、CAE 与 VAE 等改进算法。

3.1　稀疏自编码器

稀疏自编码器在自编码器中添加稀疏性限制，以发现样本中的特定结构，避免网络对恒等函数的简单学习。在 SAE 中，稀疏性限制迫使隐含层节点大部分时间被抑制，即隐含层节点输出接近于 0（因激活函数不同而不同），使网络仅依赖少量隐含层节点进行编码和解码，提取到的特征稀疏性更强。

稀疏性限制需要在损失函数上添加关于激活度的正则化项，对过大的激活度加以惩罚，以降低隐含层节点激活度。通常采用 L_1 范数或 KL 散度（Kullback-Leibler divergence）正则化项。

当采用 L_1 范数正则化项时，给定 $a_j(\boldsymbol{x}_i)$ 表示隐含层节点 j 对输入 $\boldsymbol{x}_i$ 的激活值，λ 表示控制惩罚程度的 L_1 正则化系数，则 SAE 的损失函数为

$$J_{\mathrm{SAE}}(\boldsymbol{W},\boldsymbol{b})=J\left(\boldsymbol{X},\hat{\boldsymbol{X}}\right)+\lambda\sum_{i,j}\left|a_j(\boldsymbol{x}_i)\right| \tag{8}$$

当采用 KL 散度正则化项时，给定稀疏性参数 ρ，隐含层节点 j 的平均激活度为 $\hat{\rho}_j$，KL 正则化系数 β，SAE 的损失函数为

$$J_{\mathrm{SAE}}(\boldsymbol{W},\boldsymbol{b})=J\left(\boldsymbol{X},\hat{\boldsymbol{X}}\right)+\beta\sum_{j=1}^{m}\mathrm{KL}\left(\rho\,\|\,\hat{\rho}_j\right) \tag{9}$$

KL 散度计算式为

$$\mathrm{KL}\left(\rho\,\|\,\hat{\rho}_j\right)=\rho\log\frac{\rho}{\hat{\rho}_j}+(1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j} \tag{10}$$

稀疏性参数 ρ 通常为一个较小的值（如 ρ=0.05），表示隐含层节点的理想平均激活度，为满足这一条件，隐含层节点的激活度必须接近 0。$\hat{\rho}_j$ 的计算式为

$$\hat{\rho}_j=\frac{1}{n}\sum_{i=1}^{n}\left[a_j(\boldsymbol{x}_i)\right] \tag{11}$$

KL 散度随着 ρ 与 $\hat{\rho}_j$ 之间差的增加而单调递增，这使 SAE 的训练会强迫隐含层节点的平均激活度 $\hat{\rho}_j$ 接近 ρ，更多的节点激活度接近 0，以增强所提取特征的稀疏性。

与其他自编码器相比，SAE 能够有效学习重要特征，抑制次要特征，提取的抽象特征维度更低，更具稀疏性。但 SAE 无法指定特定节点处于激活或抑制状态，且稀疏性参数的设置缺乏指导，通常需要额外的参数影响实验进行确定。

3.2　去噪自编码器

为避免传统自编码器学习到无编码功能的恒等函数，去噪自编码器在 AE 的基础上引入了退化过程。DAE 在退化过程中对输入样本添加噪声，以改变输入样本的数据分布，而后通过训练重构无噪声的样本，防止 AE 简单地将输入复制到输出，迫使 AE 提取的抽象特征更加反映样本本质、更具稳健性。DAE 网络结构如图 2 所示。

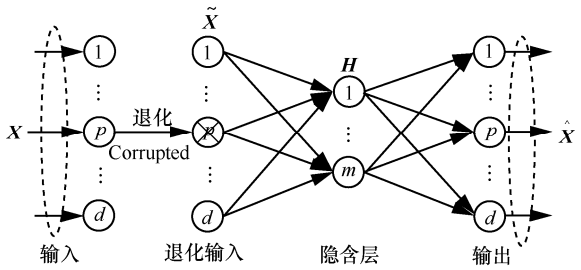

图 2　DAE 网络结构

DAE 中，退化过程是指对于每一个输入样本，按照一定比例 v 将其特征值置为 0 或其他值，这个比例 v 被称作退化率。退化过程如图 3 所示（对于灰度图像，置 0 意味着置黑）。

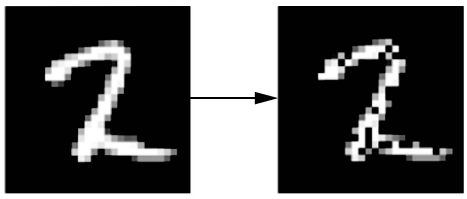

图 3　退化过程

DAE 加入退化过程的自然原理是人眼在看物体时，如果物体某一小部分被遮住了，人眼依然能将其识别出来。该现象说明人所带有的“生物”自编码器所提取的特征更具有代表性与稳健性，对于输入中含有一定噪声的样本数据，它经过编码、解码后仍能得到纯净无噪的样本。这要求自编码器不仅要有编码功能，还要有去噪作用[8]。然而，即使

样本中含有噪声，AE 却只能重构含有噪声的输入样本。所以，对原始样本进行适当的退化处理，再让自编码器重构原始样本，这样所提取的特征更本质、更抗干扰[11]。

与其他自编码器相比，DAE 能够在一定程度上克服输入样本中存在的噪声干扰，提取的抽象特征更具有代表性与稳健性。但 DAE 引入了额外的退化过程，增加了模型的训练时间，且算法对退化率敏感，过小的退化率难以有效提升算法性能，而过大的退化率会使输入样本严重失真，降低算法性能。

3.3 收缩自编码器

收缩自编码器在传统自编码器的基础上，通过在损失函数上添加收缩正则化项，迫使编码器学习到有更强收缩作用的特征提取函数，提升对输入样本小扰动的稳健性，防止对恒等函数的学习。

假设收缩正则化系数为 λ，隐含层输出关于输入样本的雅可比矩阵为 $J_f(\boldsymbol{x})$，CAE 的损失函数为

$$J_{\mathrm{CAE}}(\boldsymbol{W},\boldsymbol{b})=J\left(\boldsymbol{X},\hat{\boldsymbol{X}}\right)+\lambda\left|J_f(\boldsymbol{x})\right|_{\mathrm{F}}^2 \tag{12}$$

其中，$\left|J_f(\boldsymbol{x})\right|_{\mathrm{F}}^2$ 为雅可比矩阵 Frobenius 范数的平方，其计算式为

$$\left|J_f(\boldsymbol{x})\right|_{\mathrm{F}}^2=\sum_{i,j}\left(\frac{\partial h_j(\boldsymbol{x})}{\partial \boldsymbol{x}_i}\right)^2 \tag{13}$$

从损失函数看，CAE 通过重构误差与收缩正则化项的平衡以提取样本的抽象特征。收缩正则化项使 CAE 学习到的函数对于输入的梯度都较小，而重构误差迫使 CAE 保留完整的信息。在两者共同作用下，特征提取函数关于输入的梯度大都较小，只有少部分梯度较大。这样在输入具有小扰动时，较小的梯度会削弱这些扰动，从而提升 CAE 对输入小扰动的稳健性。

需要注意的是，CAE 与 DAE 存在差别。DAE 通过对输入样本添加噪声，经过编码与解码获得样本的稳健性重构；CAE 通过对损失函数添加正则化项，提升特征提取函数稳健性。CAE 是通过内部因素提升特征提取稳健性，而 DAE 则是通过外部因素提高特征提取稳健性。

与其他自编码器相比，CAE 能够抵抗输入样本存在的一定扰动，提取到的抽象特征具有更强的稳健性。但收缩正则化项在具有多个隐含层的自编码器中难以计算，因此 CAE 通常仅包含单一隐含层。

3.4 变分自编码器

作为特殊的自编码器，变分自编码器并非判别式模型，而属于生成模型。VAE 旨在通过对样本分布的学习，采用估计分布近似逼近样本真实分布，进而由估计分布生成原始样本的类似样本[24]。VAE 结构如图 4 所示。

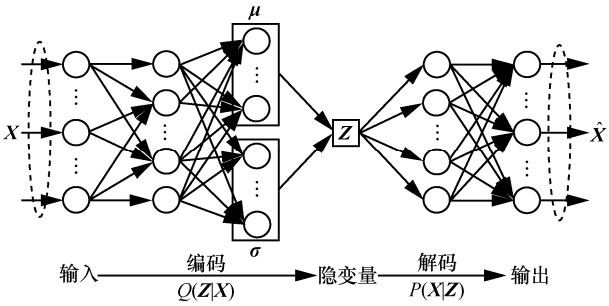

图 4 VAE 结构

图 4 中，$\boldsymbol{Z}$ 为隐变量，$\boldsymbol{\mu}$ 与 $\boldsymbol{\sigma}$ 为隐变量 $\boldsymbol{Z}$ 的均值与标准差，$Q(\boldsymbol{Z}|\boldsymbol{X})$ 与 $P(\boldsymbol{X}|\boldsymbol{Z})$ 分别为编码过程与解码过程学习到的条件分布，对应识别与生成模型[25]。其中，为使 VAE 具有样本生成能力，而非确定的映射过程，隐变量 $\boldsymbol{Z}$ 需为随机变量，且为简化计算，通常假设隐变量 $\boldsymbol{Z}$ 服从多元正态分布，即 $P(\boldsymbol{Z})\sim N(0,\boldsymbol{I})$。$Q(\boldsymbol{Z}|\boldsymbol{X})$ 为近似后验分布，旨在逼近未知的真实先验分布 $P(\boldsymbol{Z}|\boldsymbol{X})$，通常假设为正态分布。而 $P(\boldsymbol{X}|\boldsymbol{Z})$ 需被提前定义，针对二值与实值样本，通常分别选择伯努利分布与正态分布。

VAE 的训练目标旨在最小化输入样本分布 $P(\boldsymbol{X})$ 和重构样本分布 $P(\hat{\boldsymbol{X}})$ 距离，通常采用 KL 散度进行分布之间的距离衡量，即

$$D_{\mathrm{KL}}\left(P(\boldsymbol{X})\middle|P\left(\hat{\boldsymbol{X}}\right)\right)=\int P(\boldsymbol{X})\frac{P(\boldsymbol{X})}{P\left(\hat{\boldsymbol{X}}\right)}\mathrm{d}\boldsymbol{X} \tag{14}$$

但由于真实分布的未知性，KL 散度不可直接计算，因此 VAE 引入近似后验分布 $Q(\boldsymbol{Z}|\boldsymbol{X})$，并采用极大似然法优化目标函数，推导出其对数似然函数

$$\log P(\boldsymbol{X})=D_{\mathrm{KL}}\left(Q(\boldsymbol{Z}|\boldsymbol{X})\middle|P(\boldsymbol{Z}|\boldsymbol{X})\right)+L(\boldsymbol{X}) \tag{15}$$

由于 KL 散度非负，因此 $L(\boldsymbol{X})$ 称为似然函数的变分下界，其计算式为

$$L(\boldsymbol{X})=E_{Q(\boldsymbol{Z}|\boldsymbol{X})}\left(-\log Q(\boldsymbol{Z}|\boldsymbol{X})+\log P(\boldsymbol{Z})+\log P(\boldsymbol{X}|\boldsymbol{Z})\right) \tag{16}$$

由于 VAE 旨在同时最大化 $P(\boldsymbol{X})$ 与最小化 $D_{\mathrm{KL}}\left(Q(\boldsymbol{Z}|\boldsymbol{X})\middle|P(\boldsymbol{Z}|\boldsymbol{X})\right)$，因此由式(15)与式(16)可推导出其损失函数为

$$J_{\mathrm{VAE}}=D_{\mathrm{KL}}\left(Q(\boldsymbol{Z}|\boldsymbol{X})\middle|P(\boldsymbol{Z})\right)-\mathrm{E}_{Q(\boldsymbol{Z}|\boldsymbol{X})}\left(\log\left(P(\boldsymbol{X}|\boldsymbol{Z})\right)\right) \tag{17}$$

式(17)中等号右边第一项为正则化项，第二项为 VAE 期望重构误差的负值。VAE 通过最小化损失函数，使估计分布 $Q(\boldsymbol{Z}|\boldsymbol{X})$ 接近 $P(\boldsymbol{Z})$，且期望重构误差接近 0。

需要注意的是，在 VAE 训练过程中，需要对隐变量 $\boldsymbol{Z}$ 进行随机采样，无法求导，导致无法采用反向传播算法优化参数。为克服该问题，VAE 提出了重参数技巧，引入参数 $\boldsymbol{\varepsilon}\sim N(0,\boldsymbol{I})$，通过抽取 L 个样本 ε^i，将隐变量 $\boldsymbol{Z}$ 的直接采样变换为 $\boldsymbol{z}^i=\boldsymbol{u}^i+\varepsilon^i\boldsymbol{\sigma}^i$ 的线性运算，使其能够采用梯度下降算法进行优化。

VAE 的训练可以分为 3 个阶段：编码、采样和解码。编码阶段，对于输入样本 $\boldsymbol{X}$，VAE 通过识别模型 $Q(\boldsymbol{Z}|\boldsymbol{X})$ 产生隐变量 $\boldsymbol{Z}$ 分布的均值 $\boldsymbol{\mu}$ 与标准差 $\boldsymbol{\sigma}$；采样阶段，对于均值 $\boldsymbol{\mu}$ 与标准差 $\boldsymbol{\sigma}$，VAE 通过重参数化技巧，生成隐变量 $\boldsymbol{Z}$ 的随机采样样本；解码阶段，对于隐变量 $\boldsymbol{Z}$ 的采样样本，VAE 通过生成模型 $P(\boldsymbol{X}|\boldsymbol{Z})$ 生成新样本。

与其他自编码器相比，VAE 属于生成模型，能够估计样本的真实分布，进而生成类似样本，但也因此不能直接应用于分类与回归任务中。并且由于需要最小化重构误差，与 GAN 相比，生成的样本更加自然，却也更加模糊。

3.5　自编码器的其他变体

1) 卷积自编码器

传统自编码器通常采用全连接层，这会造成图像空间信息的损失，而卷积自编码器（CoAE, convolutional autoencoder）[26]受卷积神经网络启发，采用卷积层与池化层取代全连接层，以更好地保留图像的空间信息。在 CoAE 中，编码过程由卷积层和池化层（下采样层）组成，解码过程由上采样层和卷积层组成，其中上采样层为池化层的逆过程。

假设 $\boldsymbol{X}$ 表示输入样本，$\boldsymbol{W}_k$ 与 $\boldsymbol{b}_k$ 分别表示第 k 个卷积核的权值与偏置，$*$ 表示卷积运算，$g(\cdot)$ 表示池化函数，$\boldsymbol{h}_k$ 表示第 k 个卷积核所提取的抽象特征，则 CoAE 的编码过程为

$$\boldsymbol{h}_k=g\left(\boldsymbol{W}_k*\boldsymbol{X}+\boldsymbol{b}_k\right) \tag{18}$$

假设 $\boldsymbol{W}_k'$ 与 $\boldsymbol{b}_k'$ 分别表示解码器中第 k 个卷积核的权值与偏置，$g'(\cdot)$ 表示上采样函数，$\boldsymbol{H}$ 表示抽象特征集合，则解码过程为

$$\hat{\boldsymbol{X}}=\sum_{\boldsymbol{H}}\boldsymbol{W}_k'*g'\left(\boldsymbol{h}_k\right)+\boldsymbol{b}_k \tag{19}$$

解码器将各个卷积核提取的抽象特征进行解码重构，并将其合并为最终的重构样本。

CoAE 的损失函数为

$$J_{\mathrm{CoAE}}\left(\boldsymbol{W},\boldsymbol{b}\right)=J\left(\boldsymbol{X},\hat{\boldsymbol{X}}\right)+\lambda\left|\boldsymbol{W}\right|_2^2 \tag{20}$$

式(20)中等号右边第二项为权值 L_2 范数正则化项，与传统正则化自编码器相同，用于控制权值的衰减程度，以降低噪声影响，提升网络的泛化性能，并改善过拟合现象。

与其他自编码器相比，CoAE 能够直接应用于图像样本的处理，提取到的特征能够保留更多的图像空间信息。但由于涉及卷积、池化及其逆操作，CoAE 的实现更复杂。

2) 极限学习机−自编码器

传统自编码器需要误差的反向传播，通过迭代微调修改网络参数，这使其易陷入局部最优，且需要较多的训练时间。为克服局部最优问题并减少训练时间，极限学习机−自编码器（ELM-AE, extreme learning machine autoencoder）[27]将极限学习机（ELM, extreme learning machine）与 AE 相结合，随机赋值隐含层输入权值与偏置，并通过求取隐含层输出权值的最小二乘解，完成网络训练，使网络参数不需要迭代微调，极大增加了网络训练速度，而且最小二乘解为全局最优解，保证了算法的泛化性能。

ELM-AE 网络结构与 AE 相同，但其隐含层输出权值通常以 $\boldsymbol{\beta}$ 表示，并且为保证 ELM-AE 的泛化性能，隐含层输入权值与偏置需要进行正交化，即 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}$，$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}=\boldsymbol{I}$。ELM-AE 隐含层输出 $\boldsymbol{H}$ 与重构样本 $\hat{\boldsymbol{X}}$ 的关系为

$$\boldsymbol{H}\boldsymbol{\beta}=\hat{\boldsymbol{X}}\approx\boldsymbol{X} \tag{21}$$

ELM-AE 隐含层输出权值 $\boldsymbol{\beta}$ 的求解方法与网络各层节点数有关。当输入层节点数 N 与隐含层节点数 L 不同时，$\boldsymbol{\beta}$ 计算式为

$$\boldsymbol{\beta}=\begin{cases}\left(\dfrac{\boldsymbol{I}}{C}+\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{X}, & N>L\\ \boldsymbol{H}^{\mathrm{T}}\left(\dfrac{\boldsymbol{I}}{C}+\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{X}, & N<L\end{cases} \tag{22}$$

当输入层节点数 N 与隐含层节点数 L 相同时，$\boldsymbol{\beta}$ 计算式为

$$\boldsymbol{\beta} = \boldsymbol{H}^{-1}\boldsymbol{X} \tag{23}$$

与其他自编码器相比，ELM-AE 不需要采用梯度下降算法进行迭代微调，极大缩减了训练时间。但由于随机赋值的隐含层输入权值与偏置无法通过迭代进行调整，导致 ELM-AE 的算法性能通常具有较大程度的波动。

3) 自编码器改进算法比较

除上述算法以外，研究者还提出了其他自编码器改进算法[28-35]，包括区分自编码器[28]、L_{21} 范数自编码器[29]、对抗自编码器[30]等。表 2 归纳总结了上述自编码器改进算法的出发点与改进方式。

这些改进算法主要通过 3 种方式对自编码器进行创新：1)对损失函数增加特定的正则化约束，改善所提取特征的特定性质；2)优化网络结构，保留有效信息或增加数据生成能力；3)与其他算法相结合，改善训练方法，减少训练时间。

4 自编码器的应用

作为典型的深度学习模型，自编码器凭借其优异的特征提取能力，已被应用于目标识别、入侵检测、故障诊断、文本分类、图像重建等诸多领域中。

4.1 目标识别

目标识别作为机器学习技术的热点应用领域，一直备受关注，而提升目标识别性能的关键是有效的特征提取与分类。随着传感器性能的增强，其能够获取的目标信息及其种类也在增加，这使传统人工特征提取方式难以深入挖掘目标的潜在本质特征，进而影响目标识别性能[36]。而以自编码器为代表的深度学习技术能够实现目标特征的自动提取，摆脱人工提取的局限，有利于目标识别性能的提升。自编码器在目标识别上的应用，根据数据源种类的不同，大致可分为基于自编码器的高分辨距离像（HRRP, high resolution range profile）识别与合成孔径雷达（SAR, synthetic aperture radar）图像识别。

对于 HRRP 识别，Mian 等[14]将稀疏自编码器应用于小样本识别中，通过堆栈稀疏自编码器（SSAE, stacked SAE）逐层提取样本抽象特征，softmax 分类器完成类别标签的映射与网络参数的迭代微调，实现了小训练样本条件下的目标 HRRP 识别，并验证了自编码器的特征提取性能优于 PCA

表 2 自编码器改进算法的分析与比较

算法	出发点	改进方式
传统正则化自编码器	调整权值衰减程度，改善过拟合现象	在自编码器损失函数中添加权值的 L_2 范数正则化项
稀疏自编码器	增强所提取特征的稀疏性	在自编码器损失函数中添加隐含层输出的 KL 散度正则化项
去噪自编码器	改善噪声影响，提升所提取特征的稳健性	引入退化过程，用添加噪声的输入样本重构原始纯净样本
收缩自编码器	提升所提取特征对输入小扰动的稳健性	在自编码器损失函数中添加隐含层输出的收缩正则化项
变分自编码器	迫使隐变量满足特定分布，具备数据生成能力	添加隐变量分布的 KL 正则化项，引入重参数技巧将采样计算变为数值计算
卷积自编码器	保留图像的空间信息	将自编码器的全连接层替换为卷积层与池化层
极限学习机−自编码器	改善泛化性能，提升训练效率	随机赋值隐含层输入权值与偏置，并计算输出权值的最小二乘解
区分自编码器	增强所提取特征的可区分性	在自编码器损失函数中添加隐含层输出的类内距离与类间距离正则化项
L_{21} 范数自编码器	兼顾所提取特征的稀疏性与稳健性	在自编码器损失函数中添加隐含层输出的 L_{21} 范数正则化项
对抗自编码器	迫使隐变量满足特定分布，提升数据生成能力	引入对抗学习策略，完成隐变量向特定分布的逼近
图自编码器	保持低维流形的局部一致性	通过引入迹函数，在自编码器损失函数中添加图正则化项
二次自编码器	提升所提取抽象特征的稳健性	采用输入样本的二次运算代替传统神经节点的内积，提出了二次神经节点，并构建其自编码器结构
狄利克雷变分自编码器	提升生成样本的对数似然性，避免解码器权值崩溃	将变分自编码器假设的高斯分布替换为狄利克雷分布，利用随机梯度下降拟合其分布函数
注意力协同自编码器	提升局部关键信息的重要程度	将注意力机制引入自编码器中，在反向传播中实现关键特征的重新加权
同构自编码器	降低重构误差，使抽象特征更能描述输入样本数据结构	在堆栈自编码器中，将各层自编码器期望输出均替换为原始输入样本

等传统方法。Feng 等[15]为突破传统浅层模型特征提取能力的局限，通过引入平均像正则化项，提出了矫正自编码器及其深度模型。矫正自编码器将 HRRP 与其平均像之间的马氏距离作为正则化项添加到 AE 损失函数中，迫使特征提取过程考虑 HRRP 的结构相似性与振幅波动性，使提取的抽象特征能够有效缓解斑纹效应与异常值影响。为保持 SAE 对 HRRP 识别的泛化性能，提升训练速度，Zhao 等[37]将 SAE 与 ELM 相结合，提出了 SAE-ELM 方法。该方法首先通过 SAE 逐层提取抽象特征，而后使用 ELM 完成分类，不需要参数的迭代微调，极大减少了训练时间。为实现多角度 HRRP 目标的角度特征提取与分析，Chen 等[38]首先采用 SSAE 逐层提取低维抽象特征，然后通过流形学习，利用低维空间映射完成目标角度特征的提取与可视化，并分别在仿真与实测数据上验证了该方法的有效性。

对于 SAR 图像识别，Kang 等[39]提出了基于堆栈自编码器的 SAR 目标识别特征融合算法，该方法首先提取 SAR 图像的基线特征与纹理特征，而后将通过零成分分析法降维后的特征代入堆栈自编码器中进行抽象特征提取与融合，最后采用 softmax 分类器完成抽象特征到类别标签的映射。为提升 AE 在小样本条件下的特征提取能力，Deng 等[40]将暗含样本类别信息的欧氏距离约束添加到 AE 中，迫使其提取的抽象特征具有更强的类间可区分性，并将 Dropout 正则化技术应用到改进 AE 的深度模型中，防止出现过拟合现象。Dong 等[41]首先综合分析了自编码器及其改进算法，然后通过具体的 SAR 图像识别实验，验证了自编码器进行提取特征的有效性，分析了不同参数对 AE 泛化性能的影响，并比较了堆栈自编码器与其他典型算法的分类性能。为提升 SAR 图像的特征提取性能，增强抽象特征的类内聚集性，Guo 等[42]通过对卷积自编码器施加紧凑性约束，提出了紧凑卷积自编码器。该方法将类内样本距离正则化项添加到损失函数中，同时最小化重构误差与类内样本距离，以生成更具区分性的抽象特征，并通过在 MSTAR 数据集上的实验证实了方法的有效性。

4.2　入侵检测

入侵检测旨在通过分析网络数据分组中的协议类型、服务类型以及持续时间等特征，识别其中的恶意攻击行为，为应对非法入侵提供预警，保障网络安全[43]。然而，随着大数据、云计算等技术的突飞猛进，网络安全威胁日益复杂，问题复杂度不断升高，数据维度不断增加，这使传统机器学习方法难以有效提取特征，存在学习效率低、误报率高的现象，而自编码器高效的特征提取能力，有利于发现潜在安全威胁，为解决复杂的入侵检测问题提供可能。

Li 等[16]提出一种将 AE 与深度置信网络相结合的入侵检测方法，该方法首先采用 AE 进行特征降维，然后采用深度置信网络对降维后的数据进行分类。相较于传统深度置信网络，该方法的检测准确率得到一定提升。Javaid 等[17]与 Al-qatf 等[44]将 SAE 应用于入侵检测中，2 种方法首先采用独热编码处理符号特征，然后通过 SAE 进行层次化抽象特征提取，最后分别使用 softmax 分类器与支持向量机（SVM, support vector machine）完成类别映射。2 种方法均在 NSL-KDD 数据集上进行实验，通过准确率、精准率、召回率与 F 值等指标的变化，验证算法有效性。为改进 AE 与 VAE 中的抽象特征表示，迫使抽象特征向原点聚集，Cao 等[45]通过增加 AE 损失函数中的抽象特征 L_2 范数正则化项，重构 VAE 损失函数中的 KL 散度项，提出了缩小自编码器与狄拉克变分自编码器，并将其应用于异常检测中。Chouhan 等[46]将堆栈自编码器与改进的卷积网络相结合，提出了一种新的入侵检测方法。该方法将多个堆栈自编码器用于原始特征空间到抽象特征空间的转换，得到多个不同的抽象特征空间，并采用信道增强方法将不同的抽象特征空间进行叠加，而后代入改进的卷积神经网络中进行分类，完成异常行为的检测。对于入侵检测数据集中存在的类别不平衡问题，通过将类别标签作为 VAE 的额外输入，Lopez-martin 等[47]提出了一种新的变分生成模型，并将其用于少数类的生成中，有效提升了检测准确率。通过将统计分析方法与自编码器相结合，Ieracitano 等[48]提出了一种新的入侵检测方法。该方法先后采用异常值分析剔除异常值、最小最大归一化统一数值范围、独热编码数值化符号特征，而后通过数值 0 的比例进行特征剔除，接着将剩余特征分别通过 AE 与 softmax 分类器完成特征降维与分类。Tang 等[49]为进一步提升特征提取能力，通过在输入层与隐含层间新增用于计算特征注意力向量的注意力机制层，将注意力机制引入 AE 中，提出了注意力自编码器（AAE, attention autoencoder），并将堆栈注意力自编码器（SAAE, stacked AAE）与

深度神经网络（DNN, deep neural network）相结合应用于入侵检测中。

4.3 故障诊断

机械故障诊断通过对获取的机械运行状态信息进行分析比较，旨在及时发现机器异常或故障，从而减少故障或事故的发生[50]。传统故障诊断方法基于对振动信号的分析与处理，通过经验知识进行特征提取与选择，并选择浅层分类器完成故障类别的判定。然而，随着现代化机电设备发展，传感器数量增多、采样频率升高、数据量加大，且振动信号的非线性、非高斯分布性等特性凸显，传统方法难以实现故障的快速准确判断，这促使包括自编码器在内的深度学习技术应用于故障诊断中。

Zhang 等[18]与 Lu 等[19]分别将 AE 与 DAE 应用于滚动轴承故障诊断中，使用其深度结构实现振动信号的抽象特征提取，克服了传统人工特征提取的局限，极大提升了故障诊断准确率。为实现特征的自动提取，克服训练样本与测试样本间的差异性，Wen 等[51]将 SAE 与迁移学习相结合，提出了基于深度迁移学习的故障诊断方法。该方法将训练样本与测试样本均采用 SAE 逐层提取样本抽象，并在网络损失函数中加入最大平均差异正则化项，最小化训练和测试样本抽象特征之间的差异，使提取的抽象特征能同时有效表征训练与测试样本。在常用故障诊断数据集上的实验表明，该方法的预测准确率高于深度置信网络、SVM 等算法。Li 等[52]将稀疏与邻域原理应用于 ELM-AE 中，通过在损失函数中增加稀疏与邻域正则化项，更新隐含层输出权值的最小二乘法，迫使抽象特征保留样本的全局与局部流形结构，提升其可区分性，并与 ELM 及其深度模型进行对比分析，验证了所提方法在故障诊断上的有效性。为提取含噪声振动信号的有效故障特征，Yu[53]提出了一种基于负相关学习的选择性堆栈去噪自编码器集成模型。该模型首先将 bagging 算法应用于堆栈去噪自编码器（SDAE, stacked DAE）中，通过 bootstrap 采样训练样本，使用不同的采样样本进行 SDAE 抽象特征提取，然后利用负相关学习进行微调，构建分类器，最后采用粒子群算法对 SDAE 进行选择性集成，得到稳定性与泛化性能最优的模型。Zhao 等[54]为解决故障样本少所导致的类不平衡问题，将 VAE 引入故障诊断框架中，通过增扩少数类的振动信号样本，构建出类别平衡的训练样本，并代入 CNN 中进行分类。实验结果表明，与真实信号相比，VAE 生成的振动信号具有相似的时频特性，能够促进诊断准确率的提升。Yu 等[55]为提升对一维振动信号的特征提取能力，将一维卷积自编码器应用于齿轮故障诊断中，并引入残差学习对其进行改进。Gao 等[56]提出了一种基于半监督堆栈自编码器与集成极限学习机相结合的高压开关故障诊断方法。该方法首先采用自适应噪声完备经验模态分解对针对信号进行处理，得到时频能量矩阵，然后对能量矩阵采用半监督堆栈自编码器进行自动特征提取，接着采用集成极限学习机建立两级分类器，第一级用于正常或异常状态识别，第二级用于异常状态的具体故障类型识别。在验证实验中，该方法的分类准确率可达到 99.5%。

4.4 其他领域

1) 文本分类

文本分类旨在通过文档的标题、关键词、正文等特征信息，对其所属类别进行判定，从而代替人工完成文本信息的分类管理。随着深度学习技术的发展，AE、CNN 等算法已逐步应用于文本分类中[57-59]。

许卓斌等[57]为提升 AE 在词嵌入中的效果，通过在 AE 隐含层中加入全局调整函数，实现特征的合并，增强特征向量的稀疏性，并在 20 News Groups 数据集上验证了该改进方法的有效性。为提升高维度文本的特征提取能力，减少训练时间，冀俊忠等[60]提出了基于 ELM-AE 的文本分类方法。该方法首先利用 ELM-AE 对高维度文本进行特征降维，而后通过堆栈 ELM-AE 实现文本抽象特征的层次化提取，并计算输出层权值的最小二乘解进行文本分类。Xu 等[61]针对半监督文本分类问题，提出了一种半监督序列变分自编码器。该方法通过将未标记样本的类别标签作为离散潜变量，最大化样本的似然变分下界，从而隐式推导出未标记样本的潜在类别分布，并通过解决序列解码器的自回归问题，使其能够应用于文本分类。

2) 图像重建

图像重建技术旨在根据物体测量数据，通过数据处理重新建立物体图像。但是常见的压缩感知[62-63]、字典学习[64-65]等图像重建方法具有重建时间过长与超参数选择困难的问题[66]。而基于深度学习的图像重建方法，能够学习样本的高级抽象特征，避免了传统方法的人工特征提取，在重建精度与速度上实现了突破[67]。

Tan 等[68]将 AE 应用于图像重建与识别中，通

过重建误差指标，比较了堆栈自编码器与主成分分析法、深度置信网络的性能。Mehta 等[69]为提升 DAE 对异常值的稳健性，实现实时医学影像重建，将原有的欧氏范数（L_2 范数）损失函数替换为 L_1 范数损失函数，降低了异常值影响，提升了网络参数的稀疏性，同时采用少量的矩阵乘积运算，极大搞高了重建速度。为保留更多的图像细节，Zhou 等[70]通过在卷积自编码器中引入结构相似度与多尺度结构相似度指标，构成结构增强损失项，添加到损失函数中，提出了结构增强卷积自编码器，并将其作为生成器与对抗生成网络相结合，用于高度欠采样样本的图像重建。在不同欠采样率与采样类型下的对比实验表明，该方法能够以较少的模型参数重建更高质量的图像。

5　存在的问题及研究方向

尽管近年来研究者对自编码器及其改进算法进行了深入研究，但现阶段仍存在以下问题亟须解决。

1) 无监督学习模式对特征提取能力的限制

与有监督学习相比，无监督学习模式摆脱了对样本标签的依赖、避免了人工标注的困难，但也因此失去了样本标签的辅助，标签信息难以有效应用于特征提取中，使自编码器性能与有监督学习存在一定差距。因此，研究半监督[71]或有监督条件下的自编码器[72]，合理运用标签信息提升自编码器特征提取能力，是一个需要重点关注与解决的问题。

针对此问题，一方面可以通过在自编码器输入层或输出层中直接添加样本标签，同时重构输入样本及其标签，强迫自编码器在编码与解码过程中考虑到标签损失，使提取的特征更加符合不同样本的类本质。另一方面，可以通过在损失函数上添加暗含标签信息的类内离散度或类间离散度正则化项，在最小化损失函数的过程中，减少抽象特征的类内距离，增加类间距离，增强抽象特征的类可区分性，提升自编码器的特征提取能力，使抽象特征更适用于分类任务。

2) 硬件要求高，训练时间长

复杂的网络结构依赖大量的训练样本，以自编码器为代表的深度学习模型具有较高的时空复杂度，需要消耗巨大的计算与存储资源，这对硬件设备提出了更高要求，往往导致训练时间过长[73]。

针对此问题，一方面可以将模型压缩技术应用于自编码器中，采用剪枝算法剔除冗余节点或通道[74]，实现网络结构的精简，或对权值进行稀疏化，抑制部分神经节点，完成对网络参数的压缩。另一方面可以研究轻量化自编码器算法，借鉴 ELM-AE 算法，对自编码器的训练方式进行改进[75]，减少参数迭代微调次数，提升算法训练效率。此外，还可以通过研究分布式优化算法来降低模型的计算复杂度[76]，或研究并行计算方法以充分利用现有计算资源。这些方法有助于降低自编码器的结构复杂度，降低软硬件要求，减少训练时间。

3) 缺乏有效超参数设置方法

以自编码器为代表的深度学习模型具有隐含层层数、节点数等众多的超参数，这些超参数对模型泛化性能有重大影响，因此如何合理设置超参数是一个重要问题。

目前，超参数的设置一般采用试错法，通过比较超参数不同排列组合下的模型性能，选出最优的超参数设置，然而这并不适用于超参数数量较多的情形。针对此问题，一个可行的方法是将遗传算法[77]、粒子群算法[78]、蝙蝠算法等算法应用于超参数优化中，将超参数取值作为搜索目标，自编码器泛化性能作为评价标准，通过上述搜索算法，寻找能够满足最优泛化性能条件下的超参数取值，实现自编码器超参数的自动学习与设置。

4) 随机初始化引入额外噪声

目前，绝大多数自编码器及其改进算法对网络参数均采用随机初始化，这不可避免地引入了额外噪声，影响算法的收敛速度与泛化性能。因此，如何有效地进行网络初始化是一个值得深入研究的问题。

针对此问题，一方面可以通过在损失函数中添加 L_1 或 L_2 范数正则化项，以降低随机初始化导致的噪声影响，另一方面可以采用 Glorot 初始化方法[79]、He 初始化方法[80]等其他改进初始化方法，在缓解噪声影响的同时，使自编码器的训练过程更加稳定，避免出现梯度消失或爆炸现象。

5) 难以适应小样本条件，易产生过拟合

自编码器及其深度结构由于模型结构复杂，需要大量样本进行训练，在小样本条件下训练自编码器极易产生过拟合，进而降低模型泛化性能。因此小样本条件已成为制约自编码器应用的关键因素。

针对此问题，可从样本扩充与模型优化 2 个方面加以解决。在样本扩充方面，既可通过平移、旋转、过采样等传统方法对有限样本进行数据扩充，

也可通过 VAE 或 GAN 等深度生成模型学习真实样本的估计分布，生成有限样本的类似样本，解决小样本条件下的样本稀缺问题。在模型优化方面，可将迁移学习应用于自编码器中，通过在相似充足数据集上的预训练阶段与在小样本数据集上的迭代微调阶段，完成对网络参数的优化，解决小样本条件下的训练不足问题。

6　结束语

随着在各领域的成功应用，深度学习受到了广泛关注，而自编码器作为典型的无监督深度学习模型，凭借训练过程简单、多层堆栈容易、泛化性能突出等优点，成为近年来的研究热点。本文详细介绍了自编码器及其改进算法，阐述了其基本理论及算法流程，梳理与分析了自编码器在多个具体应用领域的研究进展，最后总结了现有自编码器算法在学习模式、训练时间与超参数设置等方面所存在问题，给出了可行的解决方法，展望了自编码器的研究方向。希望本文能为今后自编码器相关研究提供一定的参考。

参考文献：

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

[2] 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.

[3] SCHOLKOPF B, PLATT J, HOFMANN T. Greedy layer-wise training of deep networks[J]. Advances in Neural Information Processing Systems, 2007, 19: 153-160.

[4] 袁非牛, 章琳, 史劲亭,等. 自编码神经网络理论及应用综述[J]. 计算机学报, 2019, 42(1): 203-230.
YUAN F N, ZHANG L, SHI J T, et al. Theories and applications of auto-encoder neural networks: a literature survey[J]. Chinese Journal of Computers, 2019, 42(1): 203-230.

[5] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[6] BOURLARD H, KAMP Y. Auto-association by multilayer perceptrons and singular value decomposition[J]. Biological Cybernetics, 1988, 59(4/5):291-294.

[7] SALAKHUTDINOV R, HINTON G. Deep boltzmann machines[C]//Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics. Clearwater Beach: MLRP, 2009: 448-455.

[8] FISCHER A, IGEL C. Training restricted Boltzmann machines: an introduction[J]. Pattern Recognition, 2014, 47(1): 25-39.

[9] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2014: 2672-2680.

[10] NG A. Sparse autoencoder[J]. CS294A Lecture Notes, 2011, 72(1): 1-19.

[11] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010, 11(12): 3371-3408.

[12] RIFAI S, VINCENT P, MULLER X, et al. Contractive auto-encoders: explicit invariance during feature extraction[C]//Proceedings of the 28th International Conference on Machine Learning. Bellevue: Omnipress, 2011: 833-840.

[13] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. arXiv Preprint, arXiv:1312.6114, 2013.

[14] MIAN P, JIE J, ZHU L, et al. Radar HRRP recognition based on discriminant deep autoencoders with small training data size[J]. Electronics Letters, 2016, 52(20): 1725-1727.

[15] FENG B, CHEN B, LIU H W. Radar HRRP target recognition with deep networks[J]. Pattern Recognition, 2017, 61: 379-393.

[16] LI Y C, MA R, JIAO R H. A hybrid malicious code detection method based on deep learning[J]. International Journal of Security and Its Applications, 2015, 9(5): 205-216.

[17] JAVAID A, NIYAZ Q, SUN W Q, et al. A deep learning approach for network intrusion detection system[C]//Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies. New York: ACM Press, 2016: 21-26.

[18] ZHANG Y Y, GAO L, LI X Y, et al. A novel data-driven fault diagnosis method based on deep learning[C]// International Conference on Data Mining and Big Data. Berlin: Springer, 2017: 442-452.

[19] LU C, WANG Z Y, QIN W L, et al. Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification[J]. Signal Processing, 2017, 130: 377-388.

[20] ZHANG J, YU J, TAO D C. Local deep-feature alignment for unsupervised dimension reduction[J]. IEEE Transactions on Image Processing, 2018, 27(5): 2420-2432.

[21] WANG J L, HOU B, JIAO L C, et al. POL-SAR image classification based on modified stacked autoencoder network and data distribution[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(3): 1678-1695.

[22] LIU C, TANG L X, LIU J Y. A stacked autoencoder with sparse Bayesian regression for end-point prediction problems in steelmaking process[J]. IEEE Transactions on Automation Science and Engineering, 2020, 17(2): 550-561.

[23] KOHONEN T, HONKELA T. Kohonen network[J]. Scholarpedia, 2007, 2(1): 1568.

[24] 胡铭菲, 刘建伟, 左信. 深度生成模型综述[J]. 自动化学报, 2020, 41: 1-35.
HU M F, LIU J W, ZUO X. Survey on deep generative model[J]. Acta Automatica Sinica, 2020, 41: 1-35.

[25] 翟正利, 梁振明, 周炜, 等. 变分自编码器模型综述[J]. 计算机工程与应用, 2019, 55(3): 1-9.
ZHAI Z L, LIANG Z M, ZHOU W, et al. Research overview of variational auto-encoders models[J]. Computer Engineering and Applications, 2019, 55(3): 1-9.

[26] MASCI J, MEIER U, CIREŞAN D, et al. Stacked convolutional auto-encoders for hierarchical feature extraction[C]//Lecture Notes in

Computer Science. Berlin: Springer, 2011: 52-59.

[27] KASUN L L C, ZHOU H, HUANG G B, et al. Representational learning with extreme learning machine for big data[J]. IEEE Intelligent Systems, 2013, 28(6): 31-34.

[28] XIE J, FANG Y, ZHU F, et al. Deepshape: deep learned shape descriptor for 3D shape matching and retrieval[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 1275-1283.

[29] LI R, WANG X D, LEI L, et al. L_{21} -norm based loss function and regularization extreme learning machine[J]. IEEE Access, 2019, 7: 6575-6586.

[30] MAKHZANI A, SHLENS J, JAITLY N, et al. Adversarial autoencoders[J]. arXiv Preprint, arXiv:1511.05644, 2015.

[31] MAJUMDAR A. Graph structured autoencoder[J]. Neural Networks, 2018, 106: 271-280.

[32] FAN F L, SHAN H M, KALRA M K, et al. Quadratic autoencoder (Q-AE) for low-dose CT denoising[J]. IEEE Transactions on Medical Imaging, 2020, 39(6): 2035-2050.

[33] JOO W, LEE W, PARK S, et al. Dirichlet variational autoencoder[J]. Pattern Recognition, 2020, 107: 107514.

[34] CHEN S, WU M. Attention collaborative autoencoder for explicit recommender systems[J]. Electronics, 2020, 9(10): 1716.

[35] YUAN X F, WANG Y L, YANG C H, et al. Stacked isomorphic autoencoder based soft analyzer and its application to sulfur recovery unit[J]. Information Sciences, 2020, 534: 72-84.

[36] WANG X D, LI R, WANG J, et al. One-dimension hierarchical local receptive fields based extreme learning machine for radar target HRRP recognition[J]. Neurocomputing, 2020, 418: 314-325.

[37] ZHAO F X, LIU Y X, HUO K, et al. Radar HRRP target recognition based on stacked autoencoder and extreme learning machine[J]. Sensors, 2018, 18(2): 173.

[38] CHEN X Y, PENG X Y, LI J B, et al. Sparse autoencoder based manifold analyzer model of multi-angle target feature[J]. IEEE Access, 2020, 8: 153250-153263.

[39] KANG M, JI K F, LENG X G, et al. Synthetic aperture radar target recognition with feature fusion based on a stacked autoencoder[J]. Sensors, 2017, 17(12): 192-197.

[40] DENG S, DU L, LI C, et al. SAR automatic target recognition based on euclidean distance restricted autoencoder[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(7): 3323-3333.

[41] DONG G G, LIAO G S, LIU H W, et al. A review of the autoencoder and its variants: a comparative perspective from target recognition in synthetic-aperture radar images[J]. IEEE Geoscience and Remote Sensing Magazine, 2018, 6(3): 44-68.

[42] GUO J, WANG L, ZHU D Y, et al. Compact convolutional autoencoder for SAR target recognition[J]. IET Radar, Sonar & Navigation, 2020, 14(7): 967-972.

[43] MISHRA P, VARADHARAJAN V, TUPAKULA U, et al. A detailed investigation and analysis of using machine learning techniques for intrusion detection[J]. IEEE Communications Surveys & Tutorials, 2019, 21(1): 686-728.

[44] AL-QATF M, YU L S, AL-HABIB M, et al. Deep learning approach combining sparse autoencoder with SVM for network intrusion detection[J]. IEEE Access, 2018, 6: 52843-52856.

[45] CAO V L, NICOLAU M, MCDERMOTT J. Learning neural representations for network anomaly detection[J]. IEEE Transactions on Cybernetics, 2019, 49(8): 3074-3087.

[46] CHOUHAN N, KHAN A, KHAN H U R. Network anomaly detection using channel boosted and residual learning based deep convolutional neural network[J]. Applied Soft Computing, 2019, 83: 105612.

[47] LOPEZ-MARTIN M, CARRO B, SANCHEZ-ESGUEVILLAS A. Variational data generative model for intrusion detection[J]. Knowledge and Information Systems, 2019, 60(1): 569-590.

[48] IERACITANO C, ADEEL A, MORABITO F C, et al. A novel statistical analysis and autoencoder driven intelligent intrusion detection approach[J]. Neurocomputing, 2020, 387: 51-62.

[49] TANG C F, LUKTARHAN N, ZHAO Y X. SAAE-DNN: deep learning method on intrusion detection[J]. Symmetry, 2020, 12(10): 1695.

[50] 樊红卫, 张旭辉, 曹现刚, 等. 智慧矿山背景下我国煤矿机械故障诊断研究现状与展望[J]. 振动与冲击, 2020, 39(24): 194-204.
FAN H W, ZHANG X H, CAO X G, et al. Research status and prospect of fault diagnosis of China's coal mine machines under background of intelligent mine[J]. Journal of Vibration and Shock, 2020, 39(24): 194-204.

[51] WEN L, GAO L, LI X Y. A new deep transfer learning based on sparse auto-encoder for fault diagnosis[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, 49(1): 136-144.

[52] LI K, XIONG M, LI F C, et al. A novel fault diagnosis algorithm for rotating machinery based on a sparsity and neighborhood preserving deep extreme learning machine[J]. Neurocomputing, 2019, 350: 261-270.

[53] YU J B. A selective deep stacked denoising autoencoders ensemble with negative correlation learning for gearbox fault diagnosis[J]. Computers in Industry, 2019, 108: 62-72.

[54] ZHAO D F, LIU S L, GU D, et al. Enhanced data-driven fault diagnosis for machines with small and unbalanced data based on variational auto-encoder[J]. Measurement Science and Technology, 2020, 31(3): 035004.

[55] YU J B, ZHOU X K. One-dimensional residual convolutional autoencoder based feature learning for gearbox fault diagnosis[J]. IEEE Transactions on Industrial Informatics, 2020, 16(10): 6347-6358.

[56] GAO W, QIAO S P, WAI R J, et al. A newly designed diagnostic method for mechanical faults of high-voltage circuit breakers via SSAE and IELM[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-13.

[57] 许卓斌, 郑海山, 潘竹虹. 基于改进自编码器的文本分类算法[J]. 计算机科学, 2018, 45(6): 208-210,240.
XU Z B, ZHENG H S, PAN Z H. Improved autoencoder based classification algorithm for text[J]. Computer Science, 2018, 45(6): 208-210,240.

[58] JIANG M Y, LIANG Y C, FENG X Y, et al. Text classification based on deep belief network and softmax regression[J]. Neural Computing and Applications, 2018, 29(1): 61-70.

[59] LI Q, LI P F, MAO K Z, et al. Improving convolutional neural network for text classification by recursive data pruning[J]. Neurocomputing, 2020, 414: 143-152.

[60] 冀俊忠, 庞皓明, 杨翠翠, 等. 基于多隐层极限学习机的文本分类方法[J]. 北京工业大学学报, 2019, 45(6): 534-545.
JI J Z, PANG H M, YANG C C, et al. Text classification method based

on multi-layer extreme learning machine[J]. Journal of Beijing University of Technology, 2019, 45(6): 534-545.

[61] XU W D, TAN Y. Semisupervised text classification by variational autoencoder[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(1): 295-308.

[62] DONOHO D L. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306.

[63] WANG M D, XIAO D, XIANG Y. Low-cost and confidentiality-preserving multi-image compressed acquisition and separate reconstruction for Internet of multimedia things[J]. IEEE Internet of Things Journal, 2021, 8(3): 1662-1673.

[64] RAVISHANKAR S, BRESLER Y. MR image reconstruction from highly undersampled k-space data by dictionary learning[J]. IEEE Transactions on Medical Imaging, 2011, 30(5): 1028-1041.

[65] TAO L, JIANG X, LIU X Z, et al. Multiscale supervised kernel dictionary learning for SAR target recognition[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(9): 6281-6297.

[66] QIN C, SCHLEMPER J, CABALLERO J, et al. Convolutional recurrent neural networks for dynamic MR image reconstruction[J]. IEEE Transactions on Medical Imaging, 2019, 38(1): 280-290.

[67] 张帅勇, 刘美琴, 姚超, 等. 分级特征反馈融合的深度图像超分辨率重建[J]. 自动化学报, 2020, 46: 1-12.
ZHANG S Y, ZHANG M Q, YAO C, et al. Hierarchical feature feedback network for depth super-resolution reconstruction[J]. Acta Automatica Sinica, 2020, 46: 1-12.

[68] TAN C C, ESWARAN C. Reconstruction and recognition of face and digit images using autoencoders[J]. Neural Computing and Applications, 2010, 19(7): 1069-1079.

[69] MEHTA J, MAJUMDAR A. RODEO: robust DE-aliasing autoencoder for real-time medical image reconstruction[J]. Pattern Recognition, 2017, 63: 499-510.

[70] ZHOU W Z, DU H Q, MEI W B, et al. Efficient structurally-strengthened generative adversarial network for MRI reconstruction[J]. Neurocomputing, 2021, 422: 51-61.

[71] DABIRI S, LU C T, HEASLIP K, et al. Semi-supervised deep learning approach for transportation mode identification using GPS trajectory data[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(5): 1010-1023.

[72] YUAN X, GU Y, WANG Y, et al. A deep supervised learning framework for data-driven soft sensor modeling of industrial processes[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(11): 4737-4746.

[73] LIU Z, LI J G, SHEN Z Q, et al. Learning efficient convolutional networks through network slimming[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 2755-2763.

[74] XIANG Q, WANG X D, SONG Y F, et al. One-dimensional convolutional neural networks for high-resolution range profile recognition via adaptively feature recalibrating and automatically channel pruning[J]. International Journal of Intelligent Systems, 2021, 36(1): 332-361.

[75] 王晓丹, 来杰, 李睿, 等. 多层去噪极限学习机[J]. 吉林大学学报(工学版), 2020, 50(3): 1031-1039.
WANG X D, LAI J, LI R, et al. Multilayer denoising extreme learning machine[J]. Journal of Jilin University (Engineering and Technology Edition), 2020, 50(3): 1031-1039.

[76] LI S, KAWALE J, FU Y. Deep collaborative filtering via marginalized denoising auto-encoder[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2015: 811-820.

[77] LAMATA L, ALVAREZ-RODRIGUEZ U, MARTÍN-GUERRERO J D, et al. Quantum autoencoders via quantum adders with genetic algorithms[J]. Quantum Science and Technology, 2018, 4(1): 014007.

[78] SUN Y N, XUE B, ZHANG M J, et al. A particle swarm optimization-based flexible convolutional autoencoder for image classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(8): 2295-2309.

[79] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. New York: ACM Press, 2010: 249-256.

[80] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 1026-1034.

[作者简介]

来杰（1994− ），男，四川简阳人，空军工程大学博士生，主要研究方向为机器学习及其在目标识别、入侵检测等领域中的应用。

王晓丹（1966− ），女，陕西汉中人，博士，空军工程大学教授，主要研究方向为机器学习及其在目标识别、入侵检测等领域中的应用。

向前（1995− ），男，四川广元人，空军工程大学博士生，主要研究方向为机器学习及其在目标识别、入侵检测等领域中的应用。

宋亚飞（1988− ），男，河南汝州人，博士，空军工程大学副教授，主要研究方向为机器学习及其在目标识别、入侵检测等领域中的应用。

权文（1988− ），女，陕西蒲城人，博士，空军工程大学讲师，主要研究方向为机器学习及其在空管领航、目标识别等领域中的应用。