计算机科学
COMPUTER SCIENCE

基于记忆增强 GAN 的异常检测

周士金, 邢红杰

引用本文

周士金, 邢红杰. 基于记忆增强 GAN 的异常检测[J]. 计算机科学, 2022, 49(11A): 211000202-9.

ZHOU Shi-jin, XING Hong-jieHebei. Memory-augmented GAN-based Anomaly Detection[J]. Computer Science, 2022, 49(11A): 211000202-9.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000202

基于记忆增强 GAN 的异常检测

周士金　邢红杰
河北大学数学与信息科学学院河北省机器学习与计算智能重点实验室　河北 保定 071002
(549409090@qq.com)

摘　要　基于生成式对抗网络(Generative Adversarial Networks,GAN)的异常检测方法在训练阶段训练集仅由正常数据构成,当训练数据较为充分时,它在该训练集上能够取得较小的重构误差。然而在测试阶段,正常数据的重构误差和部分异常数据的重构误差之间的差别很小,使得基于 GAN 的异常检测方法的判别性能较差。为了解决该问题,提出了基于记忆增强 GAN 的异常检测方法。在基于 GAN 的异常检测方法中加入记忆增强模块,使模型能够记忆正常数据的特征,从而使得异常数据的重构误差变大,该方法的判别性能得到增强。在 MNIST,Fashion-MNIST 和 CIFAR-10 上的实验结果表明,与相关方法相比,所提方法具有更优的检测性能。

关键词:异常检测;生成式对抗网络;记忆增强;MNIST
中图法分类号　TP391.4

Memory-augmented GAN-based Anomaly Detection

ZHOU Shi-jin and XING Hong-jie
Hebei Key Laboratory of Machine Learning and Computational Intelligence,College of Mathematics and Information Science,Hebei University,Baoding,Hebei 071002,China

Abstract　In the training stage of the generative adversarial networks(GAN) based anomaly detection method,its training set consists of only normal data. When training data are sufficient,the GAN based anomaly detection method may obtain smaller reconstruction error. However,in the testing stage,the difference between the reconstruction errors of normal data and those of part novel data is too small,which makes the discriminant performance of the GAN based anomaly detection method become poor. To solve this problem,a memory-augmented GAN based anomaly detection method is proposed. A memory-augmented module is introduced into the proposed method to make it remember the characteristic of normal data. Hence,the reconstruction error of novel data becomes larger and thus the discriminant ability of the proposed method is enhanced. In comparison with the related approaches,experimental results on MNIST,Fashion-MNIST and CIFAR-10 verify that the proposed method has better detection performance.

Keywords　Anomaly detection,Generative adversarial networks,Memory-augmented,MNIST

1　引言

生成式对抗网络(Generative Adversarial Networks,GAN)[1]已被成功地用于处理大规模数据集,如自然图像[2]、监控视频[3]和网络攻击[4]等。近年来出现了大量 GAN 的改进模型[5]。为了控制生成数据的类型,Mirza 等[6]在 GAN 的生成器和判别器中加入某些额外信息作为条件,可以生成指定条件的数据,提出了条件生成式对抗网络(Conditional GAN,CGAN)。为了使 GAN 学习可解释和有意义的表现形式,Chen 等[7]将输入的噪声分解为不可压缩噪声和潜在代码,提出了信息最大化生成式对抗网络(Information GAN,InfoGAN)。以上两种方法在输入噪声中加入潜在的信息,使生成图像带有指定的特征,增加了 GAN 网络的可解释性。为了解决使用 KL 散度训练 GAN 时易于出现的模式崩溃问题,Arjovsky 等[8]利用推土机距离(Earth Mover's Distance,EMD)代替 KL 散度,提出了 Wasserstein 生成式对抗网络(Wasserstein GAN,WGAN),提高了 GAN 训练过程的稳定性。为了克服 GAN 训练时容易出现梯度消失的问题,Mao 等[9]采用最小二乘损失代替交叉熵损失,提出了最小二乘生成式对抗网络(Least Squares GAN,LSGAN)。Berthelot 等[10]使用自编码器(Autoencoder,AE)作为判别器,利用比例控制理论来平衡生成器和判别器,提出了边界平衡生成式对抗网络(Boundary Equilibrium GAN,BeGAN)。以上方法通过损失函数有效地提高了 GAN 训练过程中的稳定性。由于卷积神经网络比多层感知器更擅于进行图像处理,Radford 等[2]提出了深度卷积生成式对抗网络(Deep Convolutional GAN,DCGAN),取得了更优的表示性能,使得到的特征更具有通用性,该结构成为了处理图像数据最常用的模型之一。

基金项目:国家自然科学基金(61672205);河北省自然科学基金(F2017201020);河北大学高层次人才科研启动项目(521100222002)
This work was supported by the National Natural Science Foundation of China(61672205),Natural Science Foundation of Hebei Province(F2017201020) and High-Level Talents Research Start-Up Project of Hebei University(521100222002).
通信作者:邢红杰(hjxing@hbu.edu.cn)

Donahue 等[11]在 GAN 的模型中添加了一个编码器，提出了双向生成式对抗网络(Bidirectional GAN，BiGAN)，使得输入数据不仅作为判别器的输入，还通过编码器将其映射到潜在空间，极大地缩短了输入到潜在空间的映射时间。针对非配对数据，Zhu 等[12]提出了循环生成式对抗网络(Cycle GAN，CycleGAN)，在 GAN 中加入了循环一致性条件，完成了非配对数据的训练任务。以上方法在 GAN 网络中加入编码器结构，提高了网络的循环一致性，缩短了图像空间到潜在空间的映射时间。

最近，Schlegl 等[13]首次提出了基于 GAN 的异常检测(AnoGAN)方法，该方法仅使用正常数据进行训练，使 GAN 能够获得正常数据的潜在空间分布，测试时通过反向传播，找到测试数据的潜在空间分布，并通过潜在空间分布来判定异常。然而，测试时的反向传播是一个代价非常大的优化过程，这对于大型数据集或实时应用程序来说是不切实际的。Zenati 等[14]以 BiGAN 为参考，在 AnoGAN 的模型结构中增加了一个编码器，提出了基于高效 GAN 的异常检测(Efficient GAN-Based Anomaly Detection，EGBAD)方法，极大地缩短了测试时间，但是存在训练不稳定的问题。为了使 GAN 的训练过程更为稳定，Zenati 等[15]提出了 EGBAD 的改进模型，即对抗学习异常检测(Adversarially Learned Anomaly Detection，ALAD)方法，该方法在 EGBAD 的模型结构中增加了两个判别器，以此增强模型的周期一致性，从而有效提高了 ALAD 在训练过程中的模型稳定性[16]。Akcay 等[17]将编码器-解码器-编码器子网络与 GAN 相结合，并将编码器-解码器-编码器子网络用作 GAN 的生成器，提出了 GANomaly。该模型将测试数据先编码后解码再编码，以潜在向量表示的重构误差作为异常得分。与 AnoGAN 和 ALAD 相比，GANomaly 在基准图像数据集上展示了更优的检测性能[18]。为了获得正常数据在高维空间中的多尺度分布，Akcay 等[19]使用了跳跃连接的编码器-解码器卷积神经网络，提出了 skip-GANomaly。生成器中的跳跃连接可以从多尺度中捕获图像空间的细节，从而生成质量更高的图像。上述基于 GAN 的异常检测方法均使用重构误差作为异常得分，而非使用 GAN 中的判别器的输出作为异常得分，原因是 GAN 的训练目标是鼓励生成数据的分布与正常数据的分布重叠，训练完成时判别器输出趋近于 0.5，无法区分输入数据的是否为正常数据，因此判别器是无效的。Ngo 等[18]通过修改 GAN 的目标函数，使生成数据的分布与正常数据的分布不再重叠，让生成数据位于正常数据分布的边缘，进而提出了 Fence GAN(FGAN)，仅通过修改损失函数，使得原始的 GAN 模型也可以进行异常检测。

随着数据形式的多样化，对数据间依赖关系的研究得到了广泛关注。Hochreiter 等[20]在长短期记忆网络(Long Short-term Memory，LSTM)模型中引入记忆单元，以使 LSTM 捕获数据中的长期依赖关系，但是由于其记忆单元的规模较小，且存储的是压缩知识，因此该模型的记忆性能较差。Sukhbaatar 等[21]在 RNN 中引入了具有注意力机制的记忆模块，提出了端到端记忆网络，在记忆模块中，通过将内存和查询之间的内积与内存槽中的知识进行加权，完成查询操作。Gulcehre 等[22]将可训练的记忆寻址模块引入神经图灵机(Neural Turing Machine，NTM)中，提出了动态神经图灵机(Dynamic NTM，D-NTM)。该模块中的记忆单元由内容向量和地址向量组成。Miller 等[23]在键值对结构的启发下提出了键值对记忆网络(Key-Value Memory Network，KV-MemNN)，该方法将知识存储到键值对结构的记忆模块中，利用键的信息来匹配相应的问题，找到键对应的值。Santoro 等[24]提出了记忆增强神经网络(Memory-Augmented Neural Network，MANN)，该方法依据存储的内容进行寻址完成查询操作。Gong 等[25]提出了记忆增强自编码器(Memory-Augmented Autoencoder，MemAE)，提高了异常检测的精度。Park 等[26]使用特征紧致性损失和特征分离损失来训练记忆模块，提高了记忆模块的多样性和辨别能力。

上述基于 GAN 的异常检测方法中，GANomaly 表现出很好的异常检测性能。GANomaly 的训练集仅由正常数据构成，当正常数据较为充分时，其生成器可以较好地生成正常数据。在测试阶段，将异常数据输入 GANomaly 则会产生较大的重构误差，从而将它们判定为异常数据。然而在实际应用中，部分异常数据的重构误差往往较小，GANomaly 则会将它们误判为正常数据，从而降低检测性能。受 MemAE 的启发，本文提出了基于记忆增强 GAN 的异常检测方法。所提方法的主要贡献如下：

(1)所提模型在 GANomaly 模型中加入记忆增强模块。记忆增强模块记忆正常数据的特征，利用稀疏化的注意力向量，剔除测试数据重构图像中的异常特征，使重构图像获取到更多的正常数据信息，从而增大异常数据的重构误差，解决了 GANomaly 在异常数据上重构误差较小的不足。

(2)将权重熵损失、编码损失、语境损失和对抗损失组合在一起，使正常数据产生的瓶颈特征更具表示性能，且使记忆增强模块能够最大限度地存储瓶颈特征的表示信息，有效降低了模型的误报率和漏报率等指标，从而提高异常检测的性能。

(3)在 MNIST，Fashion-MNIST 和 CIFAR-10 数据集上与相关方法进行了实验比较，验证了所提方法的有效性。

2 相关工作

本节对 GANomaly 和 MemAE 的模型结构进行了简要描述。

2.1 GANomaly

GANomaly[17]的模型结构如图 1 所示。它由编码器 G_E、解码器 G_D、编码器 E 和判别器 D 组成。

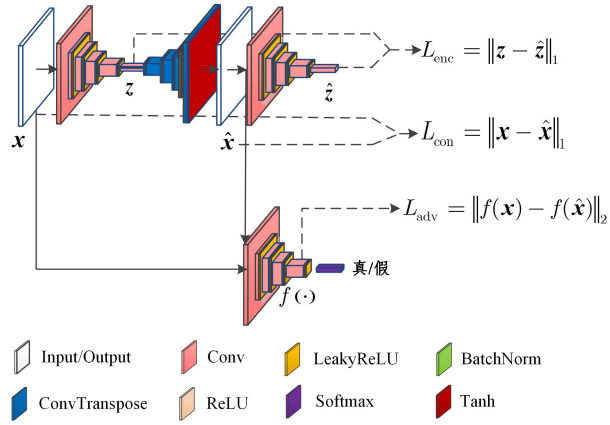

图 1 GANomaly 的模型结构

Fig. 1 Model structure of GANomaly

编码器 G_E：通过卷积、批量归一化和激活函数等运算，将输入数据 $\boldsymbol{x}(\boldsymbol{x}\in\mathcal{R}^{W\times H\times S})$ 从高维图像映射为低维行向量 $\boldsymbol{z}(\boldsymbol{z}\in\mathcal{R}^{C})$，$\boldsymbol{z}$ 称为 $\boldsymbol{x}$ 的瓶颈特征向量，则有：

$$\boldsymbol{z}=\boldsymbol{G}_{\mathrm{E}}(\boldsymbol{x}) \tag{1}$$

解码器 G_D：解码器 G_D 采用与 $DCGAN$[2] 生成器相同的结构，通过转置卷积、批量归一化、激活函数等运算，将 $\boldsymbol{z}'$ 变换回图像空间，得到重构数据 $\hat{\boldsymbol{x}}$，如式(2)所示：

$$\hat{\boldsymbol{x}}=G_{\mathrm{D}}(\boldsymbol{z}') \tag{2}$$

编码器 E：编码器 E 与编码器 G_E 的结构相同，但是参数不同。编码器 E 将 $\hat{\boldsymbol{x}}$ 映射成瓶颈特征 $\hat{\boldsymbol{z}}$，即：

$$\hat{\boldsymbol{z}}=E(\hat{\boldsymbol{x}}) \tag{3}$$

训练过程中最小化 $\hat{\boldsymbol{z}}$ 和 $\boldsymbol{z}$ 之间的距离，为了便于计算，$\hat{\boldsymbol{z}}$ 和 $\boldsymbol{z}$ 的维度相同。

判别器 D：解码器 G_D 通过转置卷积、批量归一化、激活函数等运算，依据激活函数的输出来判定输入数据是否为正常数据。

编码器 G_E 将真实图像映射为低维向量，然后解码器 G_D 将得到的低维向量重构为重构图像。最后，增加的编码器 E 将重构图像再映射为低维向量。将重构图像和真实图像输入判别器 D 进行判别。在训练过程中，GANomaly 有 3 个损失函数，分别为对抗损失 L_{adv}、语境损失 L_{con} 和编码损失 L_{enc}，分别表示为：

$$L_{\mathrm{adv}}=\mathscr{E}_{\boldsymbol{x}\sim p_x}\left\| f(\boldsymbol{x})-\mathscr{E}_{\boldsymbol{x}\sim p_X} f(G(\boldsymbol{x}))\right\|_2 \tag{4}$$

$$L_{\mathrm{con}}=\mathscr{E}_{\boldsymbol{x}\sim p_x}\left\| \boldsymbol{x}-G(\boldsymbol{x})\right\|_1 \tag{5}$$

和

$$L_{\mathrm{enc}}=\mathscr{E}_{\boldsymbol{x}\sim p_x}\left\| G_E(\boldsymbol{x})-E(G(\boldsymbol{x}))\right\|_2 \tag{6}$$

其中，p_x 表示正常数据的分布，$\boldsymbol{x}$ 为服从 p_x 的正常数据，$\|\cdot\|_1$ 和 $\|\cdot\|_2$ 分别表示 ℓ_1 和 ℓ_2 范数，$f(\cdot)$ 表示判别器中最后一个特征层的输出。

GANomaly 的目标函数为：

$$\mathscr{L}=w_{\mathrm{adv}}L_{\mathrm{adv}}+w_{\mathrm{con}}L_{\mathrm{con}}+w_{\mathrm{enc}}L_{\mathrm{enc}} \tag{7}$$

其中，w_{adv}，w_{con}，w_{enc} 均为权重参数。

GANomaly 在训练阶段仅使用正常数据作为训练集，且训练集的数量远大于测试集的数量。训练完成时生成器可以很好地重构正常数据。在测试阶段，正常数据进入模型可以很好地进行重构，但当异常数据进入模型后，大部分异常数据不能很好的重构，因此其重构误差较大。使用上述的编码损失作为异常得分的评价指标，异常得分 $\mathscr{A}(\hat{\boldsymbol{x}})$ 如下：

$$\mathscr{A}(\hat{\boldsymbol{x}})=\left\| G_{\mathrm{E}}(\hat{\boldsymbol{x}})-E(G(\hat{\boldsymbol{x}}))\right\|_1 \tag{8}$$

其中，$\hat{\boldsymbol{x}}$ 表示测试数据，G 由编码器 G_E、解码器 G_D 组成。

为了更好地评估模型的性能，依据评价指标计算出每一个测试数据 $\hat{\boldsymbol{x}}$ 的异常得分，将异常得分存在集合 S 中，然后将异常得分的范围缩放到[0，1]之间。

$$s_i'=\frac{s_i-\min(S)}{\max(S)-\min(S)} \tag{9}$$

其中，s_i 表示第 i 个测试数据的异常得分，s_i' 表示缩放后的第 i 个测试数据的异常得分，以 s_i' 作为最终的异常得分。

2.2 MemAE

MemAE[25] 的模型结构如图 2 所示。它由编码器-记忆增强模块-解码器组成。在训练阶段，记忆增强模块记录正常数据的典型特征；在测试阶段，测试数据的瓶颈特征 $\boldsymbol{z}$ 不会直接进入解码器中进行重构，而是在记忆增强模块中检索与该瓶颈特征 $\boldsymbol{z}$ 最相似的典型特征 $\hat{\boldsymbol{z}}$，再将 $\hat{\boldsymbol{z}}$ 输入解码器中进行重构，得到重构图像。

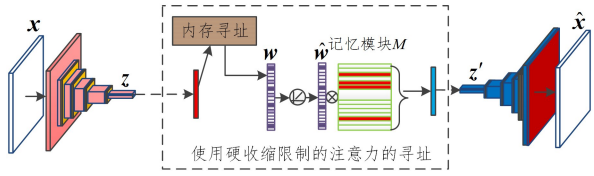

图 2 MemAE 的模型结构

Fig. 2 Model structure of MemAE

记忆增强模块利用内存项和瓶颈特征 $\boldsymbol{z}$ 的余弦相似度来计算注意力权重 $\boldsymbol{W}$，将注意力权重向量 $\boldsymbol{W}$ 作为寻址依据，其元素表示为：

$$w_i=\frac{\exp((d(\boldsymbol{z},\boldsymbol{m}_i)))}{\sum_{j=1}^{N}\exp((d(\boldsymbol{z},\boldsymbol{m}_j)))} \tag{10}$$

其中，$\boldsymbol{z}$ 表示输入数据编码后的瓶颈特征，$d(\cdot,\cdot)$ 表示余弦相似度，$\boldsymbol{m}_i$ 表示第 i 个存储项，w_i 表示第 i 个存储项与瓶颈特征 $\boldsymbol{z}$ 的注意力权重。

某些异常数据利用 $\boldsymbol{W}$ 与存储项的复杂组合使之很好地进行重构。为了解决该问题，MemAE 利用硬收缩操作来提高 $\boldsymbol{W}$ 的稀疏性。

$$\hat{w}_i=\frac{\max(w_i-\lambda,0)\cdot w_i}{|w_i-\lambda|+\varepsilon} \tag{11}$$

其中，λ 表示收缩阈值，ε 表示一个极小的正数。

模型训练过程中存在两个损失函数：重构误差 R 和熵损失 E。损失函数的表达形式为：

$$R(\boldsymbol{x}^t,\hat{\boldsymbol{x}}^t)=\left\| \boldsymbol{x}^t-\hat{\boldsymbol{x}}^t\right\|_2^2 \tag{12}$$

其中，$\boldsymbol{x}^t$ 和 $\hat{\boldsymbol{x}}^t$ 分别表示输入数据和输出数据。熵损失由式(13)计算得到：

$$E(\hat{\boldsymbol{W}}_i)=\sum_{i=1}^{T}-\hat{\boldsymbol{W}}_i\cdot\log(\hat{\boldsymbol{W}}_i), \tag{13}$$

其中，$\hat{\boldsymbol{W}}_i$ 表示记忆增强模块中第 i 个训练数据的注意力权重。

最终，MemAE 的目标函数为：

$$L(\theta_e,\theta_d,M)=\frac{1}{T}\sum_{i=1}^{T}(R(\boldsymbol{x}^t,\hat{\boldsymbol{x}}^t)+\alpha E(\hat{\boldsymbol{W}}_i)) \tag{14}$$

其中，α 为超参数，T 表示输入数据的数量。

3 基于记忆增强 GAN 的异常检测方法

本节将从模型结构、损失函数和构造过程 3 个方面详细介绍所提基于记忆增强 GAN 的异常检测方法。为了叙述方便，将所提方法简称为 Mem_GANomaly。

3.1 模型结构

Mem_GANomaly 的模型由 5 部分构成：编码器 G_E、记忆增强模块 M、解码器 G_D、编码器 E 和判别器 D。编码器 G_E-记忆增强模块 M-解码器 G_D-编码器 E 4 部分构成 Mem_GANomaly 模型的生成器。模型框架图如图 3 所示。下面对

模型组成进行详细描述。

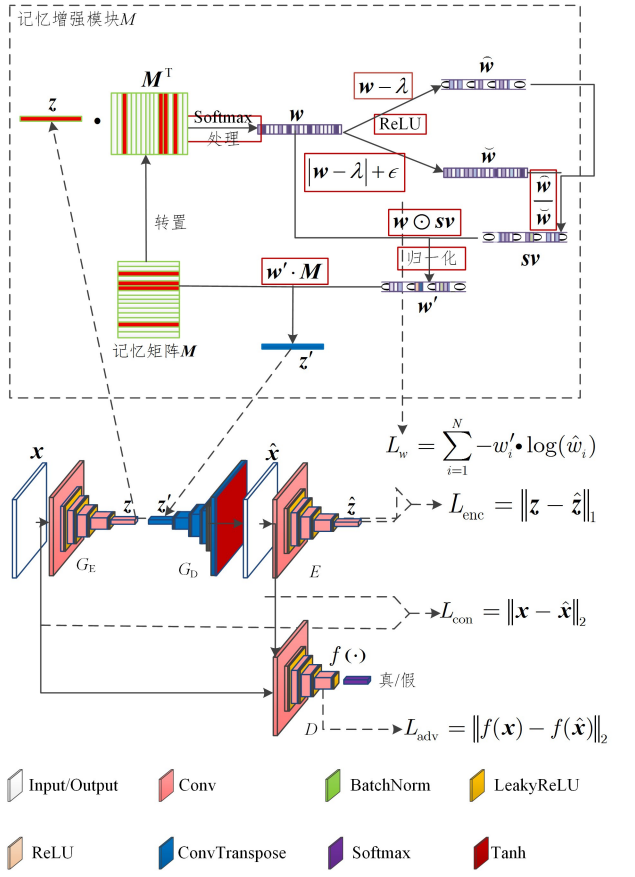

图 3 Mem_GANomaly 的模型结构

Fig. 3 Model structure of Mem_GANomaly

记忆增强模块 $\boldsymbol{M}$：由记忆矩阵 $\boldsymbol{M}$ 和注意力权重向量 $\boldsymbol{w}$ 组成。$\boldsymbol{M}=[\boldsymbol{m}_1,\boldsymbol{m}_2,\cdots,\boldsymbol{m}_N]^{\mathrm{T}}\in\mathcal{R}^{N\times C}$ 表示记忆矩阵，用于存储 N 个 C 维的瓶颈特征，记忆模块中的瓶颈特征称为典型特征，$\boldsymbol{m}_i$ 表示记忆矩阵 $\boldsymbol{M}$ 中存储的第 i 个典型特征。输入数据 $\boldsymbol{x}$ 经过编码器 G_{E} 得到瓶颈特征 $\boldsymbol{z}$ 后进入记忆增强模块，瓶颈特征 $\boldsymbol{z}$ 与记忆矩阵 $\boldsymbol{M}$ 的转置相乘得到匹配向量 $\boldsymbol{v}=(v_1,v_2,\cdots,v_N)$，计算式如下：

$$\boldsymbol{v}=\boldsymbol{z}\boldsymbol{M}^{\mathrm{T}} \tag{15}$$

通过 softmax 运算，将匹配向量 $\boldsymbol{v}$ 中的每个元素变换至(0,1)范围内，并且所有元素的和为 1，得到注意力权重向量 $\boldsymbol{w}=(w_1,w_2,\cdots,w_N)$，其第 i 个元素为：

$$w_i=\frac{\exp(v_i)}{\sum_{j=1}^{N}\exp(v_j)} \tag{16}$$

注意力权重向量 $\boldsymbol{w}$ 与记忆矩阵 $\boldsymbol{M}$ 相乘，即可得到由 $\boldsymbol{M}$ 中典型特征进行线性组合产生的瓶颈特征 $\boldsymbol{z}'$ 为：

$$\boldsymbol{z}'=\boldsymbol{w}\boldsymbol{M} \tag{17}$$

然而，记忆矩阵中典型特征仅由正常数据的瓶颈特征组成，因此异常数据的注意力权重向量中很多元素的值非常小。当利用注意力权重向量与记忆矩阵构造异常数据的瓶颈特征时，受这些值非常小的元素的影响，部分异常数据能够很好地重构，因此重构误差较小，无法使异常数据的重构误差增大。因此，剔除注意力权重向量中值非常小的元素，可以减少注意力权重向量 $\boldsymbol{w}$ 与记忆矩阵 $\boldsymbol{M}$ 组合产生异常数据瓶颈特征的可能性，使异常数据的重构数据与正常数据更相似。此外，记忆矩阵 $\boldsymbol{M}$ 的容量有限，不可能记录全部训练数据的瓶颈特征。为了增加异常数据的重构误差并且提高记忆矩阵的存储效率，可以将注意力权重向量稀疏化。设置稀疏阈值 λ，把注意力权重向量中小于阈值的元素置零，并归一化，得到由非负值组成的稀疏向量 $\boldsymbol{sv}$，其元素表示为：

$$sv_i=\frac{\max((w_i-\lambda),0)}{|w_i-\lambda|+\epsilon} \tag{18}$$

其中，$\epsilon>0$ 非常小，以避免分母为零。

利用稀疏向量 $\boldsymbol{sv}$ 对 $\boldsymbol{w}$ 进行稀疏化处理，可得：

$$\tilde{\boldsymbol{w}}=\boldsymbol{w}\odot\boldsymbol{sv} \tag{19}$$

其中，运算符⊙表示两个向量的哈达玛积，即两个向量的对应位置上的元素相乘。进一步对 $\tilde{\boldsymbol{w}}$ 中的元素进行归一化处理，得到注意力权重向量 $\boldsymbol{w}'$，即：

$$\boldsymbol{w}_i'=\frac{\tilde{w}_i}{\sum_{j=1}^{N}|\tilde{w}_j|} \tag{20}$$

将稀疏化的注意力权重向量 $\boldsymbol{w}'$ 与记忆矩阵 $\boldsymbol{M}$ 相乘，得到与输入数据的瓶颈特征 $\boldsymbol{z}$ 最相似的典型特征的组合 $\boldsymbol{z}'$，即：

$$\boldsymbol{z}'=\boldsymbol{w}'\boldsymbol{M} \tag{21}$$

3.2 损失函数

Mem_GANomaly 的损失函数分为两部分：生成器的损失函数和判别器的损失函数。其中生成器的损失函数由 4 部分组成，即对抗损失函数 L_{aav}、语境损失函数 L_{con}、编码损失函数 L_{enc} 和权重熵损失函数 L_{w}。

对抗损失函数：将判别器 D 中最后一个卷积层输出的特征差异作为对抗损失函数，这样便可以增强 GAN 的稳定性[27]。因此，将正常数据与其重构数据特征差异的 ℓ_2 范数用作对抗损失函数，则对抗损失函数的表达式为：

$$L_{\mathrm{adv}}=\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}}\|f(\boldsymbol{x})-f(\hat{\boldsymbol{x}})\|_2 \tag{22}$$

其中，$\hat{\boldsymbol{x}}$ 表示正常数据经过生成器生成的重构数据，表示如下：

$$\hat{\boldsymbol{x}}=\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}}G_{\mathrm{D}}(M(G_{\mathrm{E}}(\boldsymbol{x}))) \tag{23}$$

语境损失函数：将正常数据和重构数据尽可能保持一致，可以提高 GAN 在训练阶段的稳定性和循环一致性，同时使记忆增强模块 M 获得更好的典型特征，故将正常数据和其重构数据之间误差的 ℓ_2 范数作为语境损失函数，表示如下：

$$L_{\mathrm{con}}=\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}}\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2 \tag{24}$$

编码损失函数：为了使模型更好地捕获正常数据的特征，使正常数据被更好地重构，Mem_GANomaly 通过最小化正常数据的瓶颈特征 $\boldsymbol{z}$ 和重构数据的瓶颈特征 $\hat{\boldsymbol{z}}$ 之间的差异，反向传播对生成器进行优化。使用训练数据的瓶颈特征 $\boldsymbol{z}$ 和重构数据的瓶颈特征 $\hat{\boldsymbol{z}}$ 之间误差的 ℓ_1 范数作为编码损失函数，其表达式为：

$$L_{\mathrm{enc}}=\|\boldsymbol{z}-\hat{\boldsymbol{z}}\|_1 \tag{25}$$

其中，$\boldsymbol{z}$ 和 $\hat{\boldsymbol{z}}$ 分别为：

$$\boldsymbol{z}=\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}}G_{\mathrm{E}}(\boldsymbol{x}) \tag{26}$$

$$\hat{\boldsymbol{z}}=\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}}E(G_{\mathrm{D}}(M(G_{\mathrm{E}}(\boldsymbol{x})))) \tag{27}$$

权重熵损失函数：因为训练数据仅由正常数据组成，使记忆增强模块的记忆矩阵 $\boldsymbol{M}$ 内存储的典型特征尽可能的不同，

就可以使记忆增强模块的记忆矩阵 $\boldsymbol{M}$ 最大限度地存储有代表性的瓶颈特征，因此使用注意力权重向量 $\boldsymbol{w}'=(w_1', w_2', \cdots, w_N')$ 的交叉熵来表示权重熵损失，使记忆增强模块更加高效，则有：

$$L_{\mathrm{w}}=\sum_{i=1}^{N}-w_i'\log(w_i') \tag{28}$$

其中，N 表示训练数据的数量。

由式(18)可知，$\boldsymbol{w}'$中的元素非负且和为 1，且由式(16)可知，$\boldsymbol{w}'$可能包含多个零元素，使得 $\log(w'_i)$奇异。因此，需要对 $\boldsymbol{w}'$中的元素稍加修改，得到 $\hat{w}$，即：

$$\hat{w}_i=\begin{cases}w_i' & w_i'\neq 0\\ 1 & w_i'=0\end{cases} \tag{29}$$

权重熵损失函数则更新为：

$$L_{\mathrm{w}}=\sum_{i=1}^{N}-w_i'\log(\hat{w}_i) \tag{30}$$

最小化权重熵损失可以消除记忆增强模块中匹配度较小的典型特征，使记忆增强模块更加高效。

综上，Mem_GANomaly 中生成器的损失函数为：

$$L_{\mathrm{G}}=\eta_{\mathrm{adv}}L_{\mathrm{adv}}+\eta_{\mathrm{con}}L_{\mathrm{con}}+\eta_{\mathrm{enc}}L_{\mathrm{enc}}+\eta_{\mathrm{w}}L_{\mathrm{w}} \tag{31}$$

其中，η_{adv}，η_{enc}，η_{con}，η_w 为权重参数。通过最小化生成器损失函数，可以使模型更加稳定、记忆增强模块更高效，并能将正常数据较好地进行重构。

判别器损失函数：Mem_GANomaly 的判别器损失函数参照 GAN[1] 的表示方法，由极大极小博弈形式表示，即：

$$\min_{G_{\mathrm{E}},M,G_{\mathrm{D}},E}\max_{D}V(G_{\mathrm{E}},M,G_{\mathrm{D}},E,D)=\mathbb{E}_{x\sim p_{\mathrm{data}}}[\log D(\boldsymbol{x})]+\mathbb{E}_{x\sim p_{\mathrm{data}}}[\log(1-D(G_{\mathrm{D}}(M(G_{\mathrm{E}}(\boldsymbol{x})))))] \tag{32}$$

其中，G_{E}表示生成器中第一个编码器的输出，M 表示记忆增强模块的输出，G_{D}表示生成器中解码器的输出，E 表示生成器中第二个编码器的输出。判别器损失函数给正常数据较高的置信度，给生成数据较低的置信度，在更新判别器参数的同时，可以促使整个网络进行优化，生成网络和判别网络交替训练，直到生成数据可以以假乱真。

3.3 构造过程

所提方法的训练集仅由正常数据构成，训练阶段如图 3 所示。训练过程中存在由对抗损失函数 L_{adv}、语境损失函数 L_{con}、编码损失函数 L_{enc}和权重熵损失函数 L_w 构成的生成器损失函数和判别器损失函数。优化生成器损失函数可以增加模型训练的稳定性和循环一致性，同时优化记忆矩阵 $\boldsymbol{M}$ 的存储项。此外，优化判别器损失函数可以使生成器和判别器可以获得更优的参数。

训练过程结束后，模型可以很好地重构正常数据，因此在测试时正常数据可以很好地重构。当异常数据进入模型后，由于记忆增强模块的记忆矩阵 $\boldsymbol{M}$ 仅存储了正常数据的典型特征，异常数据的瓶颈特征经过记忆增强模块得到新的瓶颈特征，将该瓶颈特征输入解码器 G_{D}得到重构数据，该瓶颈特征由正常数据瓶颈特征的组合构成，因此异常数据的重构数据包含更多正常数据的特征，导致异常数据与其重构数据的重构误差增大，因此可以利用重构误差来检测异常数据。因为使用图像的重构误差计算量较大并且存在噪声干扰，为了获得更稳定的重构误差且更加高效的异常检测性能，使用编码损失函数 L_{enc}作为异常得分的依据，表示如下：

$$\mathscr{A}(\boldsymbol{t})=\|G_{\mathrm{E}}(\boldsymbol{t})-E(G_{\mathrm{D}}(M(G_{\mathrm{E}}(\boldsymbol{t}))))\|_1 \tag{33}$$

其中，$\boldsymbol{t}$ 为测试数据。

对于复杂的图像数据集，部分正常数据的重构误差也会很大，为了可以找到一个相对稳定的阈值，故将重构误差进行离差标准化，即计算出每个测试数据 $\boldsymbol{t}$ 的异常得分并存入集合 $A=\{a_i:\mathscr{A}(\boldsymbol{t}_i),\boldsymbol{t}_i\in D_{\mathrm{tst}}\}$并将异常得分压缩至[0，1]范围内，即可得到离差标准化后的异常得分集合 A'：

$$\boldsymbol{a}_j'=\frac{\boldsymbol{a}_i-\min(A)}{\max(A)-\min(A)} \tag{34}$$

最后，所提方法的具体实现过程如算法 1 所示。

算法 1 基于记忆增强 GAN 的异常检测

输入：训练集 $X=\{x_i\}_{i=1}^{M}$，测试集 $T=\{t_i\}_{i=1}^{N}$

输出：异常得分集合 A′

初始化：编码器 G_{E}、记忆增强模块 M、解码器 G_{D}、编码器 E 和判别器 D 的参数。

1. for i=1，2，…，do
2. $\quad\mathbf{z}_i\leftarrow G_{\mathrm{E}}(\mathbf{x}_i)$
3. $\quad\mathbf{z}_i'\leftarrow M(\mathbf{z}_i)$；
4. $\quad\hat{\mathbf{x}}_i\leftarrow G_{\mathrm{D}}(z_i')$；
5. $\quad\hat{\mathbf{z}}_i\leftarrow E(\hat{\mathbf{x}}_i)$；
6. $\quad$由上述参数计算损失函数 L_{enc}，L_{adv}，L_{con}，L_{w}
7. 使用随机梯度更新判别器 D。
8. $\nabla_{\theta_{\mathrm{D}}}\frac{1}{\mathrm{minibatch}}\sum_{i=1}^{\mathrm{minibatch}}[-\log D(\mathbf{x}_i)-\log(1-D(G_{\mathrm{D}}(M(G_{\mathrm{E}}(\mathbf{x}_i)))))]$
9. $\quad$使用随机梯度更新生成器中的编码器 G_{E}和解码器 G_{D}，同时优化记忆增强模块 M 的记忆矩阵。
10. $\nabla_{\theta_{G_{\mathrm{E}},G_{\mathrm{D}}}}\frac{1}{\mathrm{minibatch}}\sum_{i=1}^{\mathrm{minibatch}}[-\log D(G_{\mathrm{D}}(M(G_{\mathrm{E}}(\mathbf{x}_i))))+L_{\mathrm{adv}}(\mathbf{x}_i)+L_{\mathrm{con}}(\mathbf{x}_i)+L_{\mathrm{enc}}(\mathbf{x}_i)+L_{\mathrm{w}}(\mathbf{x}_i)]$
11. $\quad$使用随机梯度更新生成器中的编码器 E，同时优化记忆增强模块 M 的记忆矩阵。
12. $\nabla_{\theta_{\mathrm{E}}}\frac{1}{\mathrm{minibatch}}\sum_{i=1}^{\mathrm{minibatch}}[-\log D(G_{\mathrm{D}}(M(G_{\mathrm{E}}(\mathbf{x}_i))))+L_{\mathrm{enc}}(\mathbf{x}_i)]$
13. end for

测试开始

14. for j=1 to N do
15. $\quad A(\mathbf{t})\leftarrow\|G_{\mathrm{E}}(\mathbf{t})-E(G_{\mathrm{D}}(M(G_{\mathrm{E}}(\mathbf{t}))))\|_1$
16. $\quad A'\leftarrow A$ 离差归一化
17. end for
18. return 异常得分集合 A′。

4 实验

为了验证所提 Mem_GANomaly 模型的有效性，将它与 6 种相关方法在 3 个基准数据集上进行了实验比较。

4.1 数据集及其参数设置

下面对实验中所使用的 3 个基准数据集进行简单的介绍。

MNIST[28]：由 70000 幅大小为 28×28 的灰度手写数字图像组成，有 0～9 共 10 个类别，训练集中有 60000 幅图像，测试集中有 10000 幅图像。

Fashion-MNIST[29]：图像大小和格式及训练集/测试集的划分与 MNIST 完全相同。所有图像按照不同的商品类型划分为 10 个类别。

CIFAR-10[30]：由 60000 幅大小为 32×32 的彩色照片组成。依据不同的实体分为 10 类。训练集包含 50000 幅图像，测试集中有 10000 幅图像。

以下实验中，针对每个数据集，依次选取 10 个类别图像中的某一类图像作为正常数据，其余 9 类图像作为异常数据，组成 10 个异常检测数据集。将原训练集中的正常数据作为训练数据，将原训练集中的异常数据和原测试集的所有数据作为测试数据。

对于所提 Mem_GANomaly 模型，使用学习率为 2×10^{-3} 的 Adam 优化器对模型中的网络进行优化。在实验中发现，Mem_GANomaly 的性能受记忆增强模块中记忆矩阵的维度的影响较为明显，利用实验经验对不同的数据集设置了不同的参数。MNIST 数据集上记忆矩阵的维度为 100，迭代次数取为 60；Fashion-MNIST 数据集上记忆矩阵的维度为 1600，迭代次数取为 100；CIFAR-10 数据集上记忆矩阵的维度为 1400，迭代次数为 150。Mem_GANomaly 模型中所包含的网络均在 PyTorch 框架下搭建，编程语言 Python 的版本为 3.6.10，GPU 型号为 NVIDIA GeForce GTX TITAN X。

此外，为了衡量各种异常检测模型的性能，本文使用 ROC 曲线下的面积（AUROC）作为性能度量指标。ROC 的横坐标为假正率（FPR），表示预测为正常数据但实际为异常数据占所有异常数据的比例；ROC 的纵坐标为真正率（TPR），表示预测为正常数据且实际为正常数据占所有正常数据的比例。AUROC 是判断预测模型优劣的标准，其特性是无论数据集正常数据和异常数据是否存在不平衡，随机猜测的基线始终是 0.5[31]。AUROC 是最常用的性能度量之一[32]。但是当测试高度不平衡时，AUROC 可能产生过于乐观的结果[33]。因此还使用几何均值（gmean）、误报率（FPR）和漏报率（FNR）作为度量指标，比较模型的性能，其表达式如下：

$$gmean=\sqrt{\frac{TP*TN}{(TP+FN)*(TN+FP)}} \tag{35}$$

$$FPR=1-\frac{TP}{TP+FN} \tag{36}$$

$$FNR=\frac{FP}{TN+FP} \tag{37}$$

其中，TP 表示预测为正常数据实际是正常数据的个数，TN 表示预测为异常数据实际是正常数据的个数，FP 表示预测为正常数据实际是异常数据的个数，FN 表示预测为异常数据实际是异常数据的个数。

分类的阈值通过使用约登指数寻找最佳 ROC 的阈值来确定。

4.2 实验结果

为了比较不同异常检测方法的性能，将 AUROC 作为主要评价指标。6 种相关方法分别为 iForest[34]，DAGMM[35]，DSVDD[36]，f-AnoGAN[37]，GANomaly 和 MemAE。所有异常检测方法在 MNIST，Fashion-MNIST 和 CIFAR-10 上的 AUROC 测试结果分别概括在表 1－表 3 中。在工业异常检测中对误报率和漏报率等指标较为重视，因此表 4 所列为在 MNIST 上 gmean、误报率和漏报率的平均测试结果。

表 1 7 种不同方法在 MNIST 图像数据集上的测试性能

Table 1 Testing results of 7 different methods on MNIST

数据集	iForest	DAGMM	f-AnoGAN	MemAE	GANomaly	DSVDD	Mem_GANomaly
MNIST(0)	0.8816	0.7827	0.9763	0.9690	0.9890	0.9862	**0.9910**
MNIST(1)	0.9622	0.9331	0.9965	0.9880	0.9990	0.9948	**0.9990**
MNIST(2)	0.6617	0.6908	0.8362	0.8110	0.8980	0.8628	**0.9260**
MNIST(3)	0.7303	0.6180	0.8307	0.8470	0.9460	0.8966	**0.9470**
MNIST(4)	0.7620	0.7698	0.8938	0.8880	**0.9710**	0.9581	0.9700
MNIST(5)	0.6653	0.6755	0.9208	0.8080	**0.9750**	0.8421	0.9730
MNIST(6)	0.7727	0.6514	0.9676	0.9130	0.9910	0.9739	**0.9950**
MNIST(7)	0.8047	0.7234	0.9616	0.9160	0.9660	0.9505	**0.9750**
MNIST(8)	0.6663	0.6864	0.8182	0.7960	0.8770	**0.9341**	0.9290
MNIST(9)	0.7707	0.8037	0.9316	0.9010	0.9790	0.9505	**0.9820**
平均值	0.7678	0.7335	0.9133	0.8837	0.9591	0.9350	**0.9687**

注：最优结果加粗显示

表 2 7 种不同方法在 Fashion-MNIST 图像数据集上的测试性能

Table 2 Testing results of 7 different methods on Fashion-MNIST

数据集	iForest	DAGMM	f-AnoGAN	MemAE	GANomaly	DSVDD	Mem_GANomaly
F-MNIST(T-shirt)	0.8044	0.7399	0.8645	0.8180	0.9100	0.8741	**0.9180**
F-MNIST(Trouser)	0.9323	0.9418	0.9831	0.9080	0.9800	0.9816	**0.9840**
F-MNIST(Pullover)	0.8117	0.8193	0.9025	0.8670	0.8250	0.8446	**0.9080**
F-MNIST(Dress)	0.8642	0.8237	0.9308	0.8420	**0.9480**	0.9243	0.9350
F-MNIST(Coat)	0.8271	0.7622	0.9016	0.8790	0.8980	**0.9107**	0.9050
F-MNIST(Sandal)	0.8668	0.8734	0.9280	0.8870	**0.9280**	0.8779	0.9110
F-MNIST(Shirt)	0.7049	0.5890	0.8240	0.8120	0.8050	0.7692	**0.8310**
F-MNIST(Sneaker)	0.9361	0.9502	**0.9838**	0.9210	0.9820	0.9837	0.9610
F-MNIST(Bag)	0.7976	0.7690	0.9011	0.8190	0.8810	0.9209	**0.9330**
F-MNIST(Ankle boot)	0.9203	0.9249	0.9744	0.8980	0.9700	0.9854	**0.9820**
平均值	0.8465	0.8193	0.9194	0.8651	0.9127	0.9072	**0.9268**

表 3　7 种不同方法在 CIFAR-10 图像数据集上的测试性能

Table 3　Testing results of 7 different methods on CIFAR-10

数据集	iForest	DAGMM	f-AnoGAN	MemAE	GANomaly	DSVDD	Mem_GANomaly
CIFAR-10(Plane)	0.5922	0.6438	0.7135	0.8240	**0.9750**	0.6530	0.9120
CIFAR-10(Car)	0.4921	0.6055	0.4803	0.4540	0.6810	0.6118	**0.7090**
CIFAR-10(Bird)	0.5775	0.5715	0.6686	**0.6760**	0.6310	0.4763	0.6570
CIFAR-10(Cat)	0.5137	0.5121	0.5951	0.5340	0.6400	0.5760	**0.6430**
CIFAR-10(Deer)	0.6494	0.5872	0.7533	0.7940	0.7830	0.5205	**0.8070**
CIFAR-10(Dog)	0.5370	0.5590	0.5982	0.5780	0.6170	0.6476	**0.7500**
CIFAR-10(Frog)	0.6390	0.5424	0.7337	0.7610	0.9180	0.5810	**0.9820**
CIFAR-10(Horse)	0.5337	0.5942	0.5194	0.5500	0.6070	0.6079	**0.7800**
CIFAR-10(Ship)	0.6203	0.6339	0.7369	0.8230	**0.8880**	0.7681	0.8670
CIFAR-10(Truck)	0.5430	0.6984	0.4807	0.5370	0.6730	0.6997	**0.7900**
平均值	0.5698	0.5948	0.6280	0.6531	0.7413	0.6142	**0.7897**

表 4　6 种不同方法在 MNIST 图像数据集上 3 种不同评价标准的平均测试性能

Table 4　Average test results of 6 different methods for 3 different evaluation criteria on MNIST

评价标准	f-AnoGAN	DAGMM	MemAE	GANomaly	DSVDD	Mem_GANomaly
gmean	0.8432	0.6858	0.8069	0.9013	0.8712	**0.9188**
误报率	0.1745	0.2356	0.1963	0.1018	0.1090	**0.0790**
漏报率	0.1373	0.3788	0.1878	0.0950	0.1475	**0.0832**

由表 1 可知，除 MNIST(4)，MNIST(5)和 MNIST(8)外，Mem_GANomaly 在其余 7 个数据集上取了最优的测试结果；由表 2 可知，除 F-MNIST(Dress)，F-MNIST(Coat)，F-MNIST(Sandal)和 F-MNIST(Sneaker)外，Mem_GANomaly 在其余 6 个数据集上取了最优的测试结果；由表 3 可知，除 CIFAR-10(Plane)，CIFAR-10(Bird)和 CIFAR-10(Ship)外，Mem_GANomaly 在其余 7 个数据集上取了最优的测试结果。此外，由图像数据集上的平均值可以发现，Mem_GANomaly 优于其他 6 种方法，尤其是在 CIFAR-10 数据集上，Mem_GANomaly 的性能优势更为明显。由表 4 可知，Mem_GANomaly 在 MNIST 上的 gmean、漏报率和误报率 3 种评价指标有显著优势，证明了其在工业领域的有效性。

(1)f-AnoGAN 是提出了一种 Encoder 可以将测试数据快速映射到 GAN 潜在空间并使用 WGAN 进行异常检测的方法，与 AnoGAN 相比有显著的加速，但其只能检测出部分异常，存在较高的漏报率。与 GANomaly 和 f-AnoGAN 相比，Mem_GANomaly 引入了记忆增强模块，能有效过滤异常数据的特征并增大异常数据的重构误差，从而取得更优的检测性能。

(2)MemAE 将记忆模块与 AE 相结合，提高了异常检测的准确率。与 MemAE 相比，Mem_GANomaly 中的编码器-记忆增强模块-解码器-编码器子网能有效捕获正常数据的典型特征，同时，生成器与判别器之间的对抗机制可使所提模型获取更优的网络参数，进而有效提升所提模型的判别性能。

(3)DSVDD 使用神经网络进行特征提取，同时将输出空间优化为包含数据的超球，有效地提升了其性能，但是其使用待测数据与超球中心的距离作为异常检测的判定准测，对特征提取的质量有较高的要求。与 DSVDD 不同，Mem_GANomaly 将重构误差用作判定准则比基于距离的判定准则更具优势。

(4)iForest 是基于 Ensemble 的快速异常检测方法，具有线性时间复杂度和高精准度，但其不适用于特别高维的数据且离群点敏感。DAGMM 巧妙地将 DAE 的降维和密度估计过程结合到一起，通过多轮训练达到最大似然函数的效果。由表 4 可知，该方法的漏报率较高，即部分异常无法识别。与 iForest 和 DAGMM 相比，Mem_GANomaly 的性能优势非常明显，原因在于 Mem_GANomaly 采用的是基于卷积神经网络的模型，iForest 和 DAGMM 则是基于非卷积神经网络的模型，而卷积神经网络对图像数据集具有优秀的特征表示能力。

此外，需要指出的是，Mem_GANomaly 在一些数据集上并未取得最优的测试结果，原因在于针对每个图像数据集，Mem_GANomaly 在所构造的 10 个异常检测数据集上均使用了相同的参数设置，而不是根据正常数据的不同进行相应的参数调整，因此 Mem_GANomaly 在一些数据集上的判别性能较差。

异常得分依据编码损失函数求得。为了直观地展示正常数据与异常数据的异常得分之间差异，可将测试数据的瓶颈特征与其重构瓶颈特征之间的差进行 T-SNE 可视化。针对 MNIST(1)，仍采用其训练集，测试集则采取两种构造方式：第一个测试集记为 Test1，仍使用 MNIST(1)的测试集；第二个测试集记为 Test2，由 MNIST(1)的测试集中数字 1 和数字 7 两个类别的测试数据构成。两个测试集上的 T-SNE 可视化结果如图 4 所示。

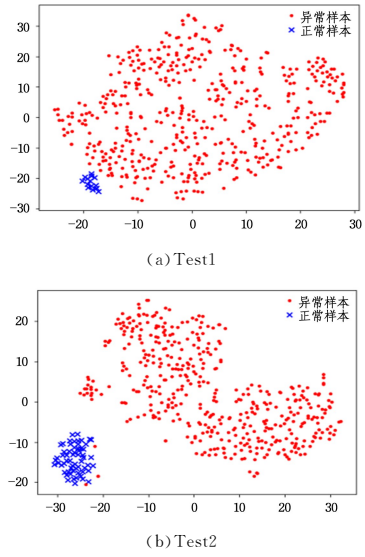

图 4　不同测试集上瓶颈特征与其重构瓶颈特征之差的 T-SNE 可视化效果

Fig. 4　T-SNE visualization of difference between bottleneck feature on different test sets and its refactored bottleneck feature

由图 4(a)的 T-SNE 可视化结果可以看出，Test1 中正常

数据和异常数据之间的边界非常明显。尽管在 MNIST 图像数据集中数字 1 和数字 7 容易混淆，但是由图 4(b)可以发现，数字 1 和数字 7 之间的边界依然非常明显。因此，Mem_GANomaly 中的记忆增强模块能够剔除异常数据的瓶颈特征中的部分特征，并使异常数据的重构图像带有正常数据的特征，从而增大异常数据的重构误差。

为了进一步展示记忆增强模块在 Mem_GANomaly 中的作用，对 MNIST(0)数据集的部分测试数据及其重构数据进行可视化，如图 5 所示。正常数据(数字 0 的图像)的重构图像与其输入图像非常相似，而异常数据(除数字 0 外的其他图像)的重构图像与其输入图像相差较大且具有数字 0 的特征。

(a)真实图像 (b)重构图像

图 5 MNIST(0)数据集中部分测试数据及其重构数据的可视化效果(红框为正常数据)(电子版为彩图)

Fig. 5 Partial test data in MNIST(0) data set and its refactoring data visualization(red box is normal data)

实验发现，Mem_GANomaly 的性能受记忆增强模块中记忆矩阵的维数的影响较为明显，下面利用 CIFAR-10(Plane)数据集，对比在不同维数的记忆矩阵下，Mem_GANomaly 的性能变化情况。如图 6 所示，当记忆矩阵的维数为 1 400 时，Mem_GANomaly 取得最优的 AUROC 值；当维数在 50～1 300 范围内变化时，AUROC 随着迭代次数的增加，在一个较小的范围内震荡；当记忆矩阵的维数超过 1 400 后，AUROC 值逐渐降低。因此可知，当记忆矩阵的维数处于一定范围内时，Mem_GANomaly 受记忆矩阵的维数的影响较小，当记忆矩阵的维数超过一定的值时，Mem_GANomaly 的性能则会下降。若记忆矩阵的维数过大，则会存储大量的典型特征，部分正常数据的特征会被稀疏的注意力权重向量过滤掉，从而使部分正常数据的重构误差增大，继而引起性能下降。

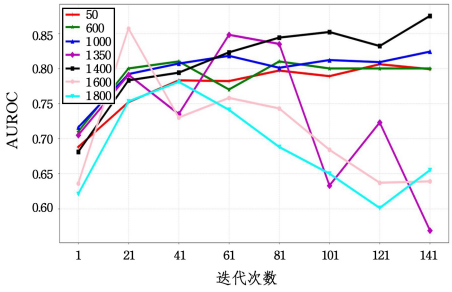

图 6 Mem_GANomaly 在 CIFAR-10(Plane)数据集上的性能随其记忆矩阵维数的变化情况

Fig. 6 Mem_GANomaly performance on CIFAR-10(Plane) dataset varies with its memory matrix dimensions

结束语 在基于 GAN 的异常检测方法中，由于测试数据中部分异常数据的重构误差较小，导致这些数据被错分为正常数据，从而取得较差的检测性能。针对上述问题，本文提出了一种基于记忆增强 GAN 的异常检测方法。该方法通过引入记忆增强模块，能够剔除异常数据所对应瓶颈特征中的部分特征，从而增大异常数据的重构误差，提升检测性能。此外，所提方法引入了权重熵损失使记忆增强模块存储最具代表性的瓶颈特征，提高了记忆增强模块的存储效率。相比基于 GAN 的异常检测的方法，所提方法有效地解决了部分样本重构误差小所导致错分的常见问题，并且有效地降低了误报率和漏报率，在工业生产应用中有显著优势。受记忆增强模块存储方式的限制，且没有单独的读入写出的操作，存储的典型瓶颈特征并不清晰。

在未来的工作中，将会从两个方面对所提方法的记忆增强模块进行探索：1)寻找更加高效的存储结构，通过提高查询效率使记忆增强模块更加高效；2)探索更优的记忆项特征提取方式和记忆项选取策略，通过提高记忆项的质量使记忆增强模块更加高效，从而提升所提方法的检测性能。

参 考 文 献

[1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial net [C]// Neural Information Processing Systems. MIT Press, 2014.

[2] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv: 1511.06434, 2016.

[3] CHEN D, YUE L, CHANG X, et al. NM-GAN: Noise-modulated generative adversarial network for video anomaly detection [J]. Pattern Recognition, 2021, 116: 107969.

[4] SIDDIQUI M A, STOKES J W, SEIFERT C, et al. Detecting cyber attacks using anomaly detection with explanations and expert feedback[C]// IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2019). IEEE, Brighton, UK, 2019.

[5] GUI J, SUN Z, WEN Y, et al. A review on generative adversarial networks: Algorithms, theory, and applications[J]. arXiv: 2001.06937, 2020.

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

[7] CHEN X, DUAN Y, HOUTHOOFT R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016.

[8] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]// Proceedings of the International Conference on Machine Learning. PMLR, Sydney, 2017.

[9] MAO X, LI Q, XIE H, et al. Least squares generative adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Beijing, China, 2017.

[10] BERTHELOT D, SCHUMM T, METZ L. Began: Boundary equilibrium generative adversarial networks [J]. arXiv: 1703.10717, 2017.

[11] DONAHUE J, KRÄHENBÜHL P, DARRELL T. Adversarial feature learning [J]. arXiv: 1605.09782, 2016.

[12] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017.

[13] SCHLEGL T, SEEBÖCK P, WALDSTEIN S M, et al. Unsuper-

vised anomaly detection with generative adversarial networks to guide marker discovery[C]// Proceedings of the International Conference on Information Processing in Medical Imaging. Springer: Boone, NC, USA, 2017.

[14] ZENATI H, FOO C S, LECOUAT B, et al. Efficient gan-based anomaly detection [J]. arXiv: 1802.06222, 2018.

[15] ZENATI H, ROMAIN M, FOO C S, et al. Adversarially learned anomaly detection [C]// Proceedings of the 2018 IEEE International conference on data mining(ICDM). IEEE, Sentosa, Singapore, 2018.

[16] HOU Y, CHEN Z, WU M, et al. Mahalanobis distance based adversarial network for anomaly detection [C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Barcelona, Spain, 2020.

[17] AKCAY S, ATAPOUR-ABARGHOUEI A, BRECKON T P. Ganomaly: Semi-supervised anomaly detection via adversarial training[C]// Proceedings of the Asian Conference on Computer Vision. Springer: Kyoto, 2018.

[18] NGO P C, WINARTO A A, KOU C K L, et al. Fence GAN: Towards better anomaly detection[C]// 31st International Conference on Tools with Artificial Intelligence(ICTAI). IEEE, Portland, OR, USA, 2019.

[19] AKÇAY S, ATAPOUR-ABARGHOUEI A, BRECKON T P. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection[C]// proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, Budapest, 2019.

[20] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-80.

[21] SUKHBAATAR S, WESTON J, FERGUS R. End-to-end memory networks[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada, 2015.

[22] GULCEHRE C, CHANDAR S, CHO K, et al. Dynamic neural turing machine with continuous and discrete addressing schemes [J]. Neural Computation, 2018, 30(4): 857-884.

[23] MILLER A, FISCH A, DODGE J, et al. Key-value memory networks for directly reading documents [J]. arXiv: 1606.03126, 2016.

[24] SANTORO A, BARTUNOV S, BOTVINICK M, et al. Meta-learning with memory-augmented neural networks [C]// Proceedings of the International Conference on Machine Learning. New York, USA, 2016.

[25] GONG D, LIU L, LE V, et al. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, CA, USA, 2019.

[26] PARK H, NOH J, HAM B. Learning memory-guided normality for anomaly detection[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA, 2020.

[27] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training gans[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016.

[28] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[29] XIAO H, RASUL K, VOLLGRAF R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms [J]. arXiv: 1708.07747, 2017.

[30] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images [J]. Journal of Software Engineering and Applications, 2009, 11(2): 1-60.

[31] FAWCETT T J P R L. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

[32] CAMPOS G O, ZIMEK A, SANDER J, et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study [J]. Data Mining Knowledge Discovery, 2016, 30(4): 891-927.

[33] AHMED F, COURVILLE A. Detecting semantic anomalies [C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York, 2020.

[34] LIU F T, TING K M, ZHOU Z H. Isolation forest[C]// Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. Pisa, Italy, 2008.

[35] ZONG B, SONG Q, MIN M R, et al. Deep autoencoding gaussian mixture model for unsupervised anomaly detection[C]// Proceedings of the International Conference on Learning Representations. Vancouver, BC, Canada, 2018.

[36] RUFF L, VANDERMEULEN R, GOERNITZ N, et al. Deep one-class classification [C]// Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018.

[37] SCHLEGL T, SEEBÖCK P, WALDSTEIN S M, et al. fAnoGAN: Fast unsupervised anomaly detection with generative adversarial networks[J]. Medical Image Analysis, 2019, 54: 30-44.

ZHOU Shi-jin, born in 1997, postgraduate. His main research interests include novelty detection and GAN.

XING Hong-jie, born in 1976, Ph.D, professor, master supervisor. His main research interests include kernel methods, neural networks, novelty detection, and ensemble learning.