基于 GAN 的异常检测

摘要

异常检测是对测试数据进行分类的任务，这些数据在某些方面不同于训练期间可用的数据。这可以被视为“一类分类”，其中构造了一个模型来描述“正常”训练数据。一般的分类任务包括从关键系统的数据集中进行推理，在这些系统中，可用的正常数据量非常大，因此可以准确地对“正常性”进行建模。但当可用“异常”数据的数量不足以为非正常类构造显式模型时（数据过少），通常使用异常检测方法。

例如在医学影像方面的癌细胞检测任务，由于癌细胞的样本过少，不足以支撑起作为一个单独的类别进行训练，所以一种经典的做法是利用自编码器训练正常的细胞，假设该模型不能对癌细胞进行稀疏表示，故在检测是，能够根据重建误差来检测出癌细胞。近年来由于 GAN 的蓬勃发展，异常检测任务也引入 GAN 的方法来解决，如 AnoGAN、OCGAN 等。本文整合了多种基于 GAN 的异常检测方法，并在多个公开数据集上实现了提及的几种方法。

1.引言

由用户生成的数据量在网络和移动应用程序增加每年以指数的速度，根据这些数据，系统正在开发学习数据模式和采取一些措施来改善用户体验，为业务决策，避免陷阱通过垃圾邮件过滤器，等等。如果用于创建这些系统的一些训练数据包含某种异常，就会影响这些系统的性能。这些系统的性能在很大程度上取决于数据的质量，因此必须识别恶意或损坏的数据，这些不一致性也称为异常。

异常是不符合预期行为的模式或数据点[1]。行为可以显式模型的一组规则(专家系统)也可以被建模基于数据集，例如基于聚类方法[2]可以直接用于表格数据，但是他们的斗争与图像数据使用时，因为直接使用像素的距离不收益率上下文信息，如对象的形象。

异常的检测可以应用在各种各样的应用程序也可以发挥重要作用的产品，如银行可以检测用户是否有他的信用卡被盗或克隆通过分析最后购买，如果不遵循用户购买模式可以被认为是一个潜在的欺诈性操作和世行可以取消购买，让他们的客户更安全。另一个例子是计算机网络中的异常数据流量模式，这可能意味着被黑客攻击的计算机试图窃取机密数据。

深度学习算法可以对复杂的非线性数据模式进行建模，并且已经在多个领域呈现出惊人的结果[3,4,5]。生成对抗网络是深度学习的子领域之一[6]，它通过对抗学习来学习对生成器建模，该算法在捕获数据分布和再现数据分布方面表现出了很好的效果，如[7]。由于 GANs 可以学习如何表示数据分布，一些最近的研究提出了一些方法来使用这种表示来通过大量不同的方法检测异常。本文对近年来提出的方法进行了阐述，并对其有效性进行了讨论。

2. 相关工作

异常或新颖性检测的研究课题已经从许多不同的角度进行了研究。文献[8]对利用统计和机器学习识别离群值的经典技术进行了全面的研究。文献[9]对使用深度学习检测计算机网络中的入侵者的网络安全进行了广泛的综述，文献[10]也研究了欺诈检测问题中的深度学习方法。通过图像数据，文献[11]展示了深度学习方法，包括对基于 GAN 的方法的一个小综述。

虽然有一些调查简要地引用了基于 GAN 的方法，但缺乏对这些方法的全面研究，介绍了它们的优点和缺点以及用于基准测试的方法过程。这项工作的目的是填补这一空白，并为谁想要在自己的问题上应用这一技术提供全面的参考。

2.1 生成对抗网络

生成式对抗网络[6]提出了一个神经网络训练框架，用于用对抗式训练估计生成式模型。该算法同时训练两个神经网络，生成网络在训练过程中学习捕获数据分布，并输出一个人工样本，鉴别器学习识别来自原始数据集或生成器输出的输入。

同时在这个训练过程中,发电机获得一个随机的高斯噪声向量也命名为一个潜在的向量,并输出一个矩阵代表一个新的数据点,然后鉴别器收到假发生器产生的样品和实际样本的原始数据集预测如果来样的训练数据。这个过程可以在图 1 中显示出来。

鉴别器的目标是最大限度地提高鉴别器做出错误预测的概率，鉴别器的目标是正确识别样本是原始的还是人工生成的。这个过程等于博弈论[6]中的一个极大极小博弈，最优情况是找到纳什均衡，其中生成器产生高质量的数据，使得鉴别器能够以 50%的概率对它们进行鉴别。

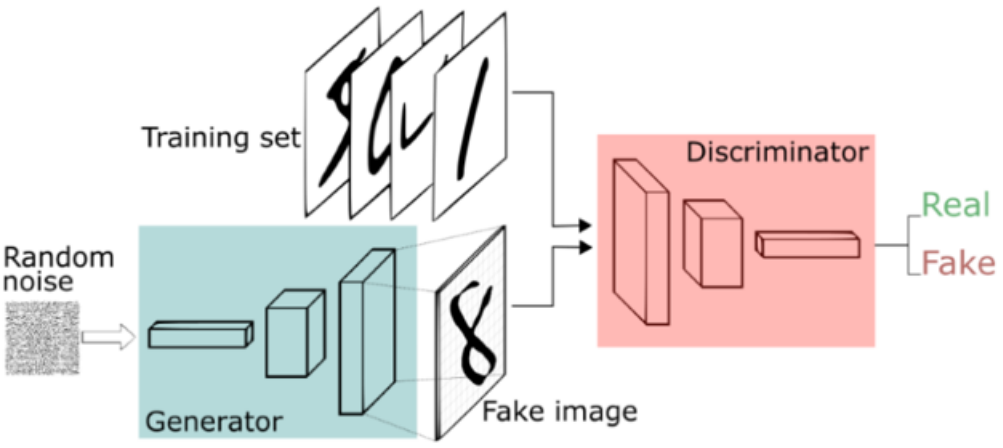

图 1. GAN 的网络结构图

3.利用生成对抗网络进行异常检测

在这一节中，我们将讨论所提出的各种方法，并说明 GAN 技术如何应用于异常检测问题。

所研究的方法都依赖于生成器网络，该网络基于输入潜向量对数据分布进行建模，潜向量是潜空间中人工样本的一种表示。换句话说，生成器学习如何将潜在的表示转换成图像。更改这个潜在向量中的值将导致修改输出的某个方面。例如，矢量的一个位置可以表示场景中一个物体的大小。神经网络能够学习如何进行这些转换。如果我们对一个存在的原始样本和一个异常样本进行相同的处理，我们可能会发现这些向量的值有所不同。

GAN 的原始结构并没有提供将图像转换为潜在向量的直接方法，只有另一种方法。在下面的小节中，我们将讨论每种方法，并介绍它们如何将此图像转换为潜在空间作为潜在向量。

3.1. AnoGAN

AnoGAN [12]是最早将 GANs 引入异常检测任务的方法之一，该方法是利用深度卷积生成对抗网络[13]架构，在常规 GAN 训练的基础上，只使用非异常样本进行训练。在测试时间每个图像样本需要映射到潜在的空间通过一个优化过程,最大限度地减少损失函数是一个加权平均相似度生成的图像和原始,和一个鉴别器的分数,计算与 L2distance 中间层的鉴别器网络。这个平均值是使用的异常分数。一个高的分数是一个异常，一个低的分数是一个非异常的样本。由于该工作依赖于测试时间的优化问题，因此计算成本较高。

3.2. Adversarially Learned Anomaly Detection

在 Adversarially Learned Anomaly Detection (ALAD) [14]的文章中，使用双向生成对抗网络(BiGANs) [15],使模型的学习样本的映射到特征空间,类似于标准的潜在空间甘斯,这个功能空间与 autoencoders 学到 BiGAN 架构。根据这种特征表示，作者提出了一些计算这种表示中样本间距离的方法，这些方程定义如下:

$$L_1 = \| x - x' \|_1 \tag{1}$$

$$L_2 = \| x - x' \|_2 \tag{2}$$

其中，L1,L2 的含义是计算输入图像和生成图像的 L1 和 L2 距离

$$Logits = \log(D_{xx}(x, x')) \tag{3}$$

其中，Logits 表示鉴别器输出的对数。

$$Features = \| f_{xx}(x,x) - f_{xx}(x,x') \|_1 \quad (4)$$

其中，Features 代表鉴别器对于输入为真实图片 x 或生成图片 x’的 L1 距离

作者阐述了使用鉴频器直接输出将导致随机预测，因为在最优情况下，编码器和解码器将完美地捕获数据分布，在这种情况下，鉴频器将无法区分真实样本和重建样本。

作者提出了一种新的自然测量方法，该方法给出了最佳的异常评分，而不是直接使用鉴别器的输出结果。

3.3. Fence GAN

在 Fence GAN 的工作中[16]等人。作者强调，GAN 的标准目标是输出的分布与原始数据的分布相等，这有利于重新创建与训练数据相似的新样本，但不足以解决异常检测问题。

假设在训练过程中，生成器只在数据分布的边界内产生样本，那么鉴别器将作为异常分类器。为了实现这个约束，作者修改了生成器和鉴别器的损失函数。

$$L_{generator}^{FGAN}(G\theta, D\phi, Z) = EL + \beta \times DL \quad (5)$$

$$EL(G_\theta, D_\phi, Z) = \frac{1}{N}\sum_{i=1}^{N}[\log(|a - D_\phi(G_\theta(z_i))|)] \quad (6)$$

$$DL(G_\theta, Z) = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}(\| G\theta(z_i) - \mu \|_2)} \quad (7)$$

由上面的公式可以看到，FenceGAN 的生成器 loss 函数，描述了 EL 和 DL 损失。EL 损失帮助生成器产生基于参数 a 的边缘分布的数据，而 DL 损失有助于强化所有生成点都位于数据分布的边界。

当生成器生成位于数据分布边界上的数据点时，鉴别器损失函数也需要修改，因为它将原始数据和位于边界上的数据点进行分类。

$$L_{generator}^{FGAN}(G\theta, D\phi, X, Z) = \frac{1}{N}\sum_{1}^{N}[-\log(D_\phi(x_i)) - \gamma \log(1 - D_\phi(G_\theta(z_i)))] \quad (8)$$

优先考虑原始数据正确分类,超参数 γ 介绍了鉴频器损失函数。γ 的值范围从[0,1],较低的值函数将产生高值时,正确鉴别器不真实的数据点进行分类,因此,将学习正常的数据分布。

3.4. GANomaly

GANomaly[17]方法类似于 ALAD [14]方法，其中生成器由编码器-

解码器-编码器子网络组成，从而产生直接的特征空间表示。

在计算异常分数时，必须考虑重构损失和特征匹配损失，分别计算输入样本与生成样本的差值，以及特征空间中的样本与生成图像的差值，其表达式如下:

$$A(\hat{x}) = \| G_E(\hat{x}) - E(G(\hat{x})) \|_1 \tag{9}$$

其中，$\hat{x}$ 是输入的数据样本，GE 代表了第一个解码器得到的结果 E 代表了第二个解码器得到的结果。理想的情况为：在归一化之后，正常样本 A（$\hat{x}$）的值接近于 0，而异常样本的值接近于 1。

3.5. OCGAN

OCGAN [18]假设在训练多个类的自动编码器时，潜在空间中的直线路径会产生不同的类，因此将潜在空间限制在一个特定的类上，在评估非目标类的样本时，会产生很高的重构误差。甘限制潜在空间作者提出一个架构,使用了两个鉴别器,一个潜在的空间网络,输出的概率样本来自限制潜在的空间,和一个视觉鉴别器,帮助模型生成高质量的数据从而提升异常检测任务的性能。

为了检测异常，分类器网络试图捕获输入图像与目标类内容的相似性，因此输入是编码图像与目标类的组合。

4. 实验

为了评估所选方法，我们选择了两个数据集，MNIST 包含 70.000 个手写数字，CIFAR-10 包含 60.000 个不同的对象，它们都有 10 个类。

为了解决异常检测问题，每个实验都使用一个类作为异常，另外 9 个类作为正常数据，因此与异常类相比，非异常样本要多得多。对两个数据集中所有类执行的过程。

Cifar-10 具有均匀分布的样本，因此每个实验有 6000 个训练异常样本，54000 个训练非异常样本，1000 个异常测试样本和 9000 个非异常测试样本。

对于每一种评估方法，超参数的使用是最好的由他们的作者报告。每个实验使用不同的随机种子执行三次，最后的结果是这三次执行的平均值。ALAD、AnoGAN、FenceGAN 和 GANomaly 的实现由各自的作者 Github 上发布，我们在相同环境下实现。OCGAN 和 KDE 的结果 [18,19,20,21]。

我们使用接收机工作特性(AUROC) [22]下的面积作为每种方法的性能测量。AUROC 评分衡量的是一个模型区分两个类别的能力，评分范围从 0 分(较差的结果)到 1 分(最好的结果)不等。

5. 实验结果

在表 1 中，可以观察到使用 MNIST 数据集的每个实验的 AUROC 评分的平均值，以及所有类的最终平均值。对于每个实验和总体平均值，突出显示最佳值。

在 MNIST 的结果中，可以用 OCGAN [18]的工作来观察与经典的 KDE [8]方法相比的巨大改进，其他方法产生了具有竞争力的结果，ALAD 方法除外。

Table 2. MNIST 数据集上的检测结果

Method	0	1	2	3	4	5	6	7	8	9	Mean
KDE	0.855	0.996	0.71	0.693	0.844	0.776	0.861	0.844	0.669	0.825	0.835
ALAD	0.743	0.296	0.684	0.525	0.456	0.432	0.579	0.38	0.553	0.363	0.491
AnoGAN	0.966	0.992	0.85	0.887	0.894	0.883	0.947	0.935	0.849	0.924	0.909
FenceGAN	0.78	0.959	0.784	0.81	0.665	0.714	0.686	0.844	0.605	0.602	0.747
GANomaly	0.892	0.675	0.946	0.793	0.81	0.847	0.821	0.689	0.838	0.65	0.816
OCGAN	0.998	0.999	0.942	0.963	0.975	0.98	0.991	0.981	0.939	0.981	0.981

CIFAR-10 实验结果表 3 中可以观察到,显示是一个更具挑战性的任务导致类似的最终成绩,与 ALAD [14]方法呈现增加 3 3%,其他方法都没有得分低于 KDE[8]方法。而且从图 2 可以看出，GAN 的方法相比 KDE 等传统方法，能够更好地重建图片，从而获得更精确的检测结果。

Table 3. CIFAR-10 数据集上的检测结果

Method	plane	car	bird	cat	deer	dog	frog	horse	ship	truck	Mean
KDE	0.658	0.52	0.657	0.497	0.727	0.496	0.758	0.564	0.68	0.54	0.611
ALAD	0.681	0.453	0.646	0.641	0.665	0.542	0.771	0.522	0.762	0.418	0.644
AnoGAN	0.671	0.547	0.529	0.545	0.651	0.603	0.585	0.625	0.758	0.655	0.614
FenceGAN	0.629	0.596	0.5	0.5	0.606	0.584	0.686	0.647	0.626	0.641	0.616
GANomaly	0.613	0.627	0.505	0.579	0.581	0.623	0.679	0.612	0.622	0.617	0.615
OCGAN	0.757	0.531	0.64	0.62	0.723	0.62	0.723	0.575	0.82	0.554	0.63

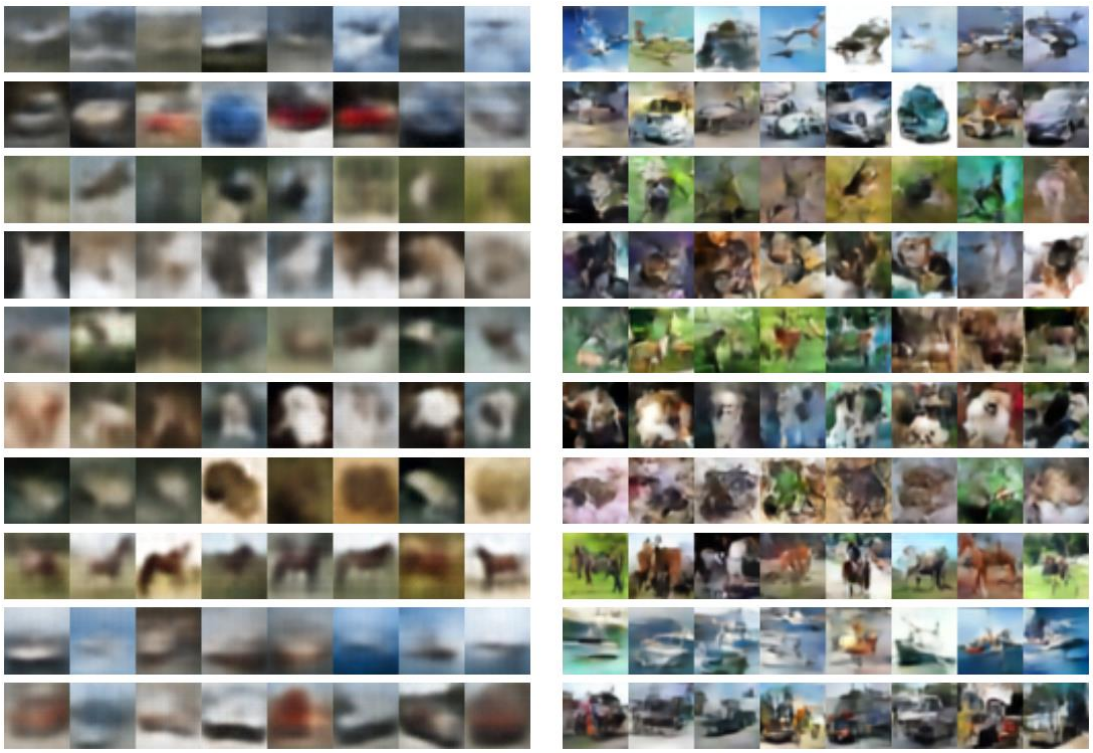

图 2. KDE 的结果(左)和 OCGAN 的结果(右)

结果表明，基于 GAN 的方法是一种很好的异常检测方法，结果表明，在这两种数据集中，GAN 方法都优于经典的 KDE 方法[8]。

6. 总结

这项工作提出了最近的努力，以带来更好的异常检测系统使用生成对抗网络。该任务的主要挑战是将测试数据转换为计算异常分数的潜在空间。AnoGAN [12]表明，这可以通过在测试时间内以较高的计算成本进行优化来实现，之后的方法依靠 OCGAN [18]等自动编码器在训练时解决了这个问题。

潜在空间表示是一种很好的图像上下文信息表示方法。最好的结果来自于使用这种表示来计算异常分数的方法。由于 KDE [8]方法依赖于基于像素值估计数据分布，因此不能对上下文信息建模，从而影响了算法的性能。

由于异常检测任务显示出良好的结果，未来的工作目标依赖于在具有更高分辨率图像的数据集上测试 GAN 方法。另一种方案是在对抗性攻击问题中使用生成的样本进行数据扩充，增加攻击者的攻击能力。

参考文献

[1] Chandola, V ., Banerjee, A., and Kumar, V . (2009). Anomaly detection: A survey. ACM Comput. Surv., 41(3):15:1 – 15:58.
[2] Alguliyev, R., Aliguliyev, R., and Sukhostat, L. (2017). Anomaly detection in big data based on clustering. Statistics, Optimization Information Computing, 5.
[3] Brock, A., Donahue, J., and Simonyan, K. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv e-prints, page arXiv:1809.11096.
[4] Shetty, R., Fritz, M., and Schiele, B. (2018). Adversarial Scene Editing: Automatic Object Removal from Weak Supervision. arXiv e-prints, page arXiv:1806.01911.
[5] Zhang, C., Song, D., Chen, Y ., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., and Chawla, N. V . (2018). A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. arXiv e-prints, page arXiv:1811.08055.
[6] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y . (2014). Generative Adversarial Networks. arXiv e-prints, page arXiv:1406.2661.
[7] Curto, J. D., Zarza, H. C., and Kim, T. (2017). High-Resolution Deep Convolutional Generative Adversarial Networks. arXiv e-prints, page arXiv:1711.06491.
[8] Hodge, V . and Austin, J. (2004). A survey of outlier detection methodologies. Artif. Intell. Rev., 22(2):85 – 126.
[9] Kwon, D., Kim, H., Kim, J., C. Suh, S., Kim, I., and Kim, K. (2017). A survey of deep learning-based network anomaly detection. Cluster Computing.
[10] Adewumi, A. O. and Akinyelu, A. A. (2017). A survey of machine-learning and nature inspired based credit card fraud detection techniques. International Journal of System Assurance Engineering and Management, 8(2):937 – 953.
[11] Kiran, B. R., Thomas, D. M., and Parakkal, R. (2018). An overview of deep learning based methods for pervised and semi-supervised anomaly detection in videos. arXiv e-prints, page arXiv:1801.03149.
[12] Schlegl, T., Seeböck, P ., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. page arXiv:1703.05921.
[13] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. page arXiv:1511.06434.
[14] Zenati, H., Romain, M., Foo, C. S., Lecouat, B., and Ramaseshan Chandrasekhar, V . (2018). Adversarially Learned Anomaly Detection. arXiv e-prints, page arXiv:1812.02288.
[15] Donahue, J., Krähenbühl, P ., and Darrell, T. (2016). Adversarial Feature Learning. arXiv e-prints, page arXiv:1605.09782.
[16] Phuc Ngo, C., Aristo Winarto, A., Kou Khor Li, C., Park, S., Akram, F., and Lee, H. K. (2019). Fence GAN: Towards Better Anomaly Detection. arXiv e-

prints, page arXiv:1904.01209.
[17] Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P . (2018). GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. arXiv e-prints, page arXiv:1805.06725.
[18] Perera, P ., Nallapati, R., and Xiang, B. (2019). OCGAN: One-class Novelty Detection Using GANs with Constrained Latent Representations. arXiv e-prints, page arXiv:1903.08550.
[19] Silva, T. (2018). An intuitive introduction to generative adversarial networks (gans).
[20] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara.(2019) AND: Autoregressive Novelty Detectors. In 2019 IEEE Conference on Computer Vision and Pattern Recognition.
[21] P. Perera and V. M. Patel.(2018) Learning Deep Features for OneClass Classification. ArXiv e-prints, page arXiv:1801.05365.
[22] Bradley, A. P . (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recogn., 30(7):1145 - 1159.