浙江大学

题　　目：**OCGAN：One-class Novelty Detection**

姓名学号：左同斌/**21821172**

邮　　箱：**zuotongbin@gmail.com**

电　　话：**15620864995**

老　　师：王东辉

专　　业：计算机科学与技术

2019 年 6 月 1 日

OCGAN：One-class Novelty Detection

摘要

这篇文章来自于 CVPR2019，针对经典的异常检测问题，提出了一种新的模型 OCGAN。给定一个特定的实例集，可以查询实例是否来自于同一个类。解决方案是使用去噪自编码器网络学习类内实例的潜在表示。工作主要贡献在于约束潜在空间，使其只表示给定的类。为了实现这一目标，通过在编码器的输出层引入 tanh 激活函数来使得获得的潜在空间是有限的。第二，在潜在空间下，使用判别器对抗训练，确保类内样例的编码表示类似于从同一空间随机获得的均匀随机样本。第三，在输入空间使用第二个判别器，确保随机潜在样本生成的样例看起来都是真实的。最后引入一种基于梯度下降的采样技术，该技术探索潜在空间中的点，这些点生成潜在的类外实例，反馈给网络，进一步训练网络这些点产生类内样例。这篇文章使用四个公开的数据集合和两个单类异常检测原则来证明该方法的有效性，结果显示达到了最先进的效果。

关键字：one-class、异常检测、GAN;

1 引论

单类异常检测解决了量化测试样本属于训练样本所定义分布概率的问题。不同于其它机器学习任务，在一个类的异常检测中，在训练时只观察到一个类的例子。在推理过程中，期望得到训练后的模型。接受类内示例并拒绝类外样例。由于问题方法假设没有任何负面的训练数据，所以很难在实践中解决。然而，它在异常检测入侵者检测、生物医学数据处理学习不平衡等问题中有许多应用。随着深度学习的发展，单类异常检测受到了国内外学者的广泛关注。在单类异常检测任务中，重点为学习给定类的代表性的潜在空间，一旦这样的一个空间被学习，异常检测将图像映射到学习的潜在空间上来执行。在以往的文献中，有两种不同的策略通常用来达到这一目的，第一种策略，区别查询图像和它的重构图像被用作一种新颖的检测器。第二种策略是使用分布，学习潜在空间。这篇文章使用了前一个异常检测策略。文章研究了现有的表示学习技术的局限性，建议只学习在单类样例产生的潜在空间，以提高异常检测的能力。

现有的工作重点都是生成一个保留给定类的细节的潜在表示。这样做时，假设出现类外对象时，对于网络来说，它在描述对象方面做得很差，因此报告的重建误差相对较高。文献 [1] 对自动编码器等网络进行训练，对于具有简单形状（如 0 和 1）的数字上具有异常检测精度。相反，复杂的形状，如数字 8，具有相对较弱的异常检测精度。

这是因为为具有复杂形状的类，学习的潜在空间本质上也学习表示一些类外对象。例如，在数字 8 上学习的潜在空间也可以表示其他数字，如 1、3、6、7。

异常检测的要求不只是为了确保类内样本得到很好的表示；它也是为了确保不合格样品的表示很差。至今没有工作已经解决了后一个要求。在这项工作中，提出一类 OCGAN，双重潜在空间，需要同时考虑这些要求的学习过程。

在高层次上，学习到一个潜在的空间，它代表给定类的对象。第二，保证从所学的潜在空间中产生的例子确实是来自已知道的类。换句话说，如果网络经过 8 这个数字的训练，可以确保当用于生成图像时，抽取的样本来自潜在空间的图像对应于数字 8 的图像。这样可以确保类外样本不能很好地由网络表示。因为整个潜在空间对应于来自数字 8，所有投射到潜在空间的反射都会产生数字 8 的图像。

2 相关工作

单类异常检测是一个定义明确的研究问题，具有标准的评价指标。传统上，它被视为一个表征学习问题。单类异常检测的最早方法是主成分分析法以及它的核扩展来找到一个子空间描述给定的概念。随着神经网络和深度学习的出现，开始使用自动编码网络寻求类似的映射。一旦学习了这种映射，就可以进行异常检测，要么基于重构误差，要么通过显式建模已知类的正常行为空间。在 [2] 中，采用了前一种策略使用均方误差作为异常函数。在 [3] 中，一个生成对抗网络对给定的噪声样本进行去噪。这里，判别器对图像空间进行预测用于量化重构误差。后另一种稍微不同的策略是，[4] 提出学习属于给定的类随机分布和图像流形之间的映射。在 [4] 中，最接近查询的图像是通过反向传播寻找，其中异常检测是基于两幅图像之间的差异来执行的。异常检测和单分类都是与单类异常检测相关的问题。两者都有相似的目标——在给定一组类内样本的情况下检测类外样本。一个硬标签被期望分配给一个给定的图像类; 因此，采用检测精度和 F1 评分来衡量其性能。相比之下，新奇

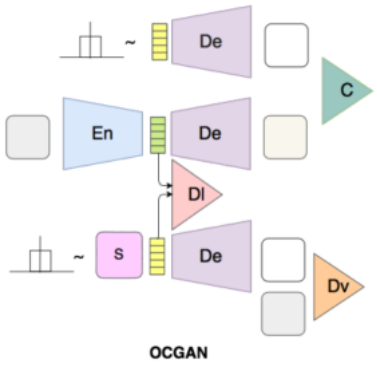
度检测只期望将一个新奇度得分与给定的图像相关联; 因此，异常检测的性能是通过工作特性 (ROC) 曲线来测量的。而基于边界的单类分类方法，如单类 SVM、SVDD 等，则可以将距离决策边界的距离作为异常得分，相比之下，异常检测 (也称为离群点检测) 是一种无监督学习任务。对于类内未标记和类外未标记的混合实例，异常检测的目标是类内分离类外的实例。由于异常检测和新奇性检测遵循不同的协原则，所以我们注意，这两个任务是不可比较的。因此, 用于异常检测和新奇性检测的工具不能互换使用。

介绍下 GAN，在给定一组图像的情况下，引入的生成对抗性网络在生成器网络和鉴别器网络之间进行博弈。这里，生成器网络试图从给定的图像分布生成真实的图像 (假图像)，而鉴别器网络则试图将假图像与真实图像区分开来。在平衡状态下，生成网络学习给定图像集的分布。为了达到这种状态，GAN 理论指出，两个网络的容量之间应该保持平衡。后来，GAN 被扩展到条件设置。在此基础上，GAN 算法被广泛应用于图像到图像的转换中。在 [5] 中，GANs 甚至可以使用深度卷积网络来学习稳定的表示，前提是要做出某些设计选择。受 [3] 网络架构的启发，遵循 [5] 中概述的原则，这篇文章提出了一个深度卷积 GAN 架构作为解决方案的主干。

3 OCGAN

OCGAN 主要由四个部分组成, 去噪编码器，隐判别器，图像判别器，和分类器。一一讲解它们的作用。去噪编码器。AE 的主要作用就是学习特征表示 (feature representation)。其瓶颈层的输出即为

图 1: OCGAN 整体结构图



表示，其所在空间即为隐空间。为了使得隐空间有界，作者使用了 tanh，将其空间的值限定在 [-1,1]。该 AE 的 loss 即为均方误差 MSE。使用去噪 AE 的原因是因为去噪 AE 可以降低过拟合，提高泛化能力。

隐判别器。如前所述，该动机是获得一个隐空间，空间中的每个实例表示给定类的图像。如果给定类的表示仅限于潜在空间的子区域，则无法实现此目标。因此，显式地强制给定类的表示均匀的分布在整个隐空间。做法是构建一个判别器 D_l , 来判别给定类的表示和来在 $U(-1,1)^d$ 的样本。

图像判别器，动机是隐空间的所有样本通过 decoder 生成的图像应该来自于给定类的图像的空间。为了满足这个约束，构建第二个判别器 D_v , 来判别给定类的图像和从隐空间随机采样通过 decoder 之后生成的图像。至此构成整个论文的核心。但是作者发现，即使这样，从隐空间中采样生成的图像有时候也很难对应于给定类图像。这是因为隐空间太大了，完全采样到是不可能的。于是不

如主动去发现隐空间中的那些产生 poor 的图像的区域。

分类器。分类器的作用是判别生成的图像和给定类的图像的相似度。使用给定类图像作为正样本，生成图像作为负样本。该分类器的损失函数为二类交叉熵 (BCE)。整体架构如下图 1 所示。

训练方式如图二所示。交替优化的方，第一步，固定住除分类器之外的所有部件，优化分类器。第二步，固定分类器，优化 AE 和判别器。

图 2: 训练算法

Input : Set of training data x, iteration size N, parameter λ
Output: Models: En, De, C, D_l , D_v

for *iteration 1* **to** $\rightarrow N$ **do**
  Classifier update: keep D_l, D_v, *En, De fixed.*
  $n \longleftarrow \mathcal{N}(0, I)$
  $l_1 \longleftarrow En(x + n)$
  $l_2 \longleftarrow \mathbb{U}(-1, 1)$
  $l_{classifier} \longleftarrow C(De(l_2), 0) + C(De(l_1), 1)$
  Back-propagate $l_{classifier}$ *to change* C
  Discriminator update:
  $l_{latent} \longleftarrow D_l(l_1, 0) + D_l(l_2, 1)$
  $l_{visual} \longleftarrow D_v(De(l_2), 0) + D_v(x, 1)$
  Back-propagate $l_{latent} + l_{visual}$ *and change* D_l, D_v
  Informative-negative mining : Keep all networks fixed.
  for *sub-iteration 1* **to** $\rightarrow 5$ **do**
    $l_{classifier} \longleftarrow \mathrm{C}(De(l_2), 1)$
    Back-propagate $l_{classifier}$ *to change* l_2
  end
  Generator update: keep D_l, D_v,*C fixed.*
  $l_{latent} \longleftarrow D_l(l_1, 1) + D_l(l_2, 0)$
  $l_{visual} \longleftarrow D_v(De(l_2), 1) + D_v(x, 0)$
  $l_{mse} \longleftarrow ||x - De(l_1)||^2$
  Back-propagate $l_{latent} + l_{visual} + \lambda l_{mse}$ *to change En, De*
end

4 实验结果

图 3 和图 4 分别展示了在 MNIST 数据集和 CIFAT10 数据集上 OCGAN 的效果，平均性能达到了 State of the art。消融实验验证了各个模块的作用。

图 3: MNIST 上的结果

	0	1	2	3	4	5	6	7	8	9	MEAN
OCSVM [24]	0.988	**0.999**	0.902	0.950	0.955	0.968	0.978	0.965	0.853	0.955	0.9513
KDE [2]	0.885	0.996	0.710	0.693	0.844	0.776	0.861	0.884	0.669	0.825	0.8143
DAE [4]	0.894	**0.999**	0.792	0.851	0.888	0.819	0.944	0.922	0.740	0.917	0.8766
VAE [6]	0.997	**0.999**	0.936	0.959	0.973	0.964	0.993	0.976	0.923	0.976	0.9696
Pix CNN [26]	0.531	0.995	0.476	0.517	0.739	0.542	0.592	0.789	0.340	0.662	0.6183
GAN [23]	0.926	0.995	0.805	0.818	0.823	0.803	0.890	0.898	0.817	0.887	0.8662
AND [1]	0.984	0.995	**0.947**	0.952	0.960	0.971	**0.991**	0.970	0.922	0.979	0.9671
AnoGAN [23]	0.966	0.992	0.850	0.887	0.894	0.883	0.947	0.935	0.849	0.924	0.9127
DSVDD [19]	0.980	0.997	0.917	0.919	0.949	0.885	0.983	0.946	**0.939**	0.965	0.9480
OCGAN	**0.998**	**0.999**	0.942	**0.963**	**0.975**	**0.980**	**0.991**	**0.981**	**0.939**	**0.981**	**0.9750**

图 4: CIFAR10 上的结果

	PLANE	CAR	BIRD	CAT	DEER	DOG	FROG	HORSE	SHIP	TRUCK	MEAN
OCSVM [24]	0.630	0.440	0.649	0.487	0.735	0.500	0.725	0.533	0.649	0.508	0.5856
KDE [2]	0.658	0.520	0.657	0.497	0.727	0.496	**0.758**	0.564	0.680	0.540	0.6097
DAE [4]	0.411	0.478	0.616	0.562	0.728	0.513	0.688	0.497	0.487	0.378	0.5358
VAE [6]	0.700	0.386	**0.679**	0.535	**0.748**	0.523	0.687	0.493	0.696	0.386	0.5833
Pix CNN [26]	**0.788**	0.428	0.617	0.574	0.511	0.571	0.422	0.454	0.715	0.426	0.5506
GAN [23]	0.708	0.458	0.664	0.510	0.722	0.505	0.707	0.471	0.713	0.458	0.5916
AND [1]	0.717	0.494	0.662	0.527	0.736	0.504	0.726	0.560	0.680	0.566	0.6172
AnoGAN [23]	0.671	0.547	0.529	0.545	0.651	0.603	0.585	0.625	0.758	0.665	0.6179
DSVDD [19]	0.617	**0.659**	0.508	0.591	0.609	**0.657**	0.677	**0.673**	0.759	**0.731**	0.6481
OCGAN	0.757	0.531	0.640	**0.620**	0.723	0.620	0.723	0.575	**0.820**	0.554	**0.6566**

图 5: 消融实验

Without any Discriminators	0.957
With latent Discriminator	0.959
With two Discriminators	0.971
Two Discriminators + Classifier	0.975

5 总结

只用单类样本训练 AE 寻找隐空间，会使得只有隐空间的一些子域表示原类样本，即使类内的样本很丰富。这篇文章的亮点在于将隐空间限制后，从隐空间中采样，要求隐空间所生成的样本都为原类，这样类外样本在重构后就会获得高的重构误差。OCGAN 源码还没有公布，对 OCGAN 作了实现，代码上传到 github。此外，在 FMNIST 上做了实验，效果不错，实验结果如图 6。

图 6: FMNIST 上的结果

0	1	2	3	4	5	6	7	8	9
0.915	0.954	0.85	0.936	0.9	0.842	0.779	0.984	0.824	0.992

参考文献

[1] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara. AND: Autoregressive Novelty Detectors. In 2019 IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[2] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In Proceedings of the MLSDA 2014 2Nd Workshop on Machine Learning for Sensory Data Analysis, 2014.

[3] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3379–3388, 2018.

[4] Thomas Schlegl, Philipp Seebock, Sebastian M. Waldstein, ¨Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In IPMI, 2017.

[5] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR, abs/1511.06434, 2015.