Towards Visually Explaining Variational Autoencoders 复现

摘要

最近，卷积神经网络（CNN）模型可解释性方面的最新进展在可视化和理解模型预测方面取得了显著的进展。具体而言，基于梯度的视觉注意方法推动了在可解释性方面使用视觉注意图的研究。然而，一个关键问题是这些方法是专为分类和分级任务设计的，而将它们扩展到解释生成模型，如变分自编码器（VAE），仍然具有挑战。

本研究复现了 Towards Visually Explaining Variational Autoencoders 一文中的实验，并对其进行了详细的分析和验证。该论文旨在填补这一关键差距，提出了一种基于梯度的注意力手段，首次实现了对 VAE 进行视觉解释的技术。研究介绍了从学习到的潜在空间生成视觉注意力的方法，并展示了这些注意力解释不仅仅用于解释 VAE 的预测。研究展示了这些注意力图如何用于定位图像中的异常，通过在 MVTecAD 数据集上展示最先进的性能，证明了其有效性。

此外，研究还探讨了将这些注意力图融入模型训练的可能性，以助力引导 VAE 学习更好的潜在空间解缠。研究在 Dsprites 数据集上的实验证明了这一点，展示了注意力机制在模型训练中的潜在价值。这一复现研究为在生成模型可解释性方面的进一步研究提供了重要的实证基础。

关键词： 视觉解释；变分自编码器；异常定位；空间解缠

1 引言

在深度学习的推动下，计算机视觉取得了巨大的进步 [8, 11, 14]，导致了相关算法在现实世界任务中的广泛采用，包括医疗保健、机器人和自动驾驶 [12, 15, 29] 等。在许多这样的安全关键和消费者关注领域的应用程序需要清楚地了解算法预测背后的推理，当然还有鲁棒性和性能保证。因此，最近人们对设计理解和解释底层的方法产生了浓厚的兴趣。

近年来，随着深度学习在计算机视觉和人工智能领域的广泛应用，对于深度模型的可解释性需求日益迫切。特别是在卷积神经网络（CNN）模型中，解释模型的预测过程对于理解模型行为、提高模型的可信度以及在实际应用中的可靠性至关重要。在这个背景下，变分自编码器（VAE）等深度生成模型的可解释性问题引起了研究者的极大兴趣从这种分类模型的可解释性出发，人们自然会想要解释更广泛的神经网络模型和架构。虽然算法生成建模的进展迅速，但解释这种生成算法仍然是一个相对未开发的研究领域。在生成模型中使用视觉注意的概念确实有一些正在进行的努力，但这些方法的重点是将注意力作为感兴趣的特定任务的辅助信息源，而不是直观地解释生成模型本身。

当前，基于梯度的视觉注意力方法在解释分类任务上的 CNN 模型取得了显著进展。然而，将这些方法扩展到解释生成模型，特别是 VAE，却面临着挑战。这一问题的存在使得在生成模型中解释潜在表示的有效性受到限制。因此，对于开发能够解释 VAE 的新技术成为当前深度学习研究中的重要议题。

本研究选取了 Towards Visually Explaining Variational Autoencoders [17] 这一具有挑战性的问题作为研究对象，旨在弥合现有关于深度生成模型可解释性的差距。通过开发新的技术，研究人员有望提供一种直观、可视化的方式来解释 VAE 的预测过程，从而深化对生成模型内部运作的理解。此外，通过将新的学习目标（注意力解纠缠损失）引入标准 VAE 模型，本研究探索了提高潜在表示解缠能力的可能性，为深度学习模型的改进和应用提供新的思路。在这一背景下，本研究对于推动深度学习模型在实际应用中更可信、可解释的发展具有重要意义。

2 相关工作

2.1 视觉解释生成模型

最近，人们花了很多精力来解释 cnn 在大多数视觉任务中的表现。一些被广泛采用的试图可视化中间 CNN 特征层的方法包括 Zeiler 和 Fergus [25] 以及 Mahendran 和 Vedaldi [18] 的工作，他们提出了理解卷积网络层内活动的方法。这一领域的一些最新扩展包括基于视觉注意力的方法 [7, 23, 28]，其中大多数方法可以分为基于梯度的方法或基于响应的方法。基于梯度的方法，如 GradCAM [23] 计算并可视化从决策单元反向传播到特征卷积层的梯度。另一方面，基于响应的方法 [7, 26, 28] 通常会在原始 CNN 架构中添加额外的可训练单元来计算注意力图。在这两种情况下，目标都是定位对模型预测贡献最大的关注和信息图像区域。然而，这些方法及其扩展 [7, 16, 24] 虽然能够解释分类/分类模型，但不能简单地扩展到解释深层生成模型 (如 vae)。在这项工作中，论文提出了使用基于梯度的网络注意原理的方法，直接从 VAE 的学习潜在嵌入中计算和可视化注意图，朝着解决相对未被探索的视觉解释生成模型的问题迈出了一步。具体来说，给定学习到的高斯分布，使用重新参数化技巧对潜在编码进行采样。然后，将潜在编码的每个维度的激活反向传播到模型中的卷积特征层，并聚合所有结果的梯度以生成注意图。此外, 还是使用注意地图进行端到端的训练，并显示这种变化如何导致改进的潜在空间解纠缠。

2.2 异常检测

异常检测中的无监督学习 [1] 仍然具有挑战性。最近的异常检测工作是基于分类 [4, 22] 或基于重建的方法。基于分类的方法旨在逐步学习具有代表性的一类决策边界，如正态类输入分布周围的超平面 [4] 或超球 [22]，以区分异常值/异常值。然而，研究也表明 [3]，这些方法在处理高维数据时存在困难。在这项工作中，我使用作者提出的 VAE 视觉解释生成方法生成的注意图作为线索来定位异常。直觉是，异常数据的表示应该在潜在嵌入中反映为异常，并且从这样的嵌入中生成输入视觉解释为研究提供了定位特定异常所需的信息。

2.3 VAE 解纠缠

在理解生成模型的潜在空间解纠缠方面已经有许多学者做出努力。最近一些用于解缠的无监督方法包括 β-VAE [10]，它试图探索观测数据变化的独立潜在因素。虽然仍然是一种流行的无监督框架，但 β-VAE 为了获得更好的解纠缠而牺牲了重建质量。Chen 等人 [5] 通过引入基于总相关性的目标将 β-VAE 扩展到 β-TCVAE，而 Mathieu 等人 [19] 探索了将潜在表征分解为两个因子以解缠，Kim 等人 [13] 提出了 FactorVAE，鼓励表征在各维度上的分布是阶乘的和独立的。虽然这些方法侧重于分解每个潜在神经元提供的潜在表征，但论文采用了不同的方法。通过论文基于提出的视觉解释 (即视觉注意图) 制定解纠缠约束来强制学习解纠缠空间。最后，根据作者提出一种新的注意力解缠学习目标，我定量地表现其解纠缠能力，与现有工作相比，该目标提供了更好的性能。

3 本文方法

在本部分，我介绍了一种使用基于梯度的注意力生成变分自编码器（VAE）解释的方法。首先在第 3.1 节中对 VAE 进行了简要回顾，然后详细介绍了作者提出的激发 VAE 注意力的方法。我深入讨论了使用这些注意图来定位图像异常的框架，并在广泛的实验中，特别是在 MVTec-AD 异常检测数据集上，建立了最先进的异常定位性能。接下来，我展示了生成的注意力可视化如何通过优化新的注意力解纠缠损失来协助学习潜在空间的解缠。在这一方面，我在 Dsprites[29] 数据集上进行了实验，并用量化的结果证明了相较于现有方法，解缠性能得到了明显的提高。

3.1 一类变分自编码器

一个基本的变分自编码器（VAE）本质上是一个使用标准的自编码器重构目标进行训练的自编码器，该目标在输入数据和解码或者重构数据之间。此外，它还包括一个变分目标项，试图学习一个标准正态潜在空间分布。变分目标通常是通过计算潜在空间分布和标准高斯分布之间的 Kullback-Leibler 分布度量来实现的。给定输入数据 x、编码器的条件分布 $q(z|x)$、标准高斯分布 $p(z)$ 和重构的数据 $\hat{x}$，基本的 VAE 进行了优化：

$$L = L_r(x, \hat{x}) + L_{KL}(q(z|x), p(z)) \tag{1}$$

3.2 生成 VAE 注意力

论文提出了一种基于梯度注意力计算的 VAE 视觉注意力生成方法，与现有的工作 [34,47,45] 有很大的不同，现有的工作是通过从分类模型中反向传播分数来计算注意力图。而这种方法不受这些要求的限制，直接使用学习到的潜在空间开发注意机制，从而不需要额外的分类模块。如图 1 所示，从潜在空间中计算一个分数，然后用它来计算梯度并获得注意图。

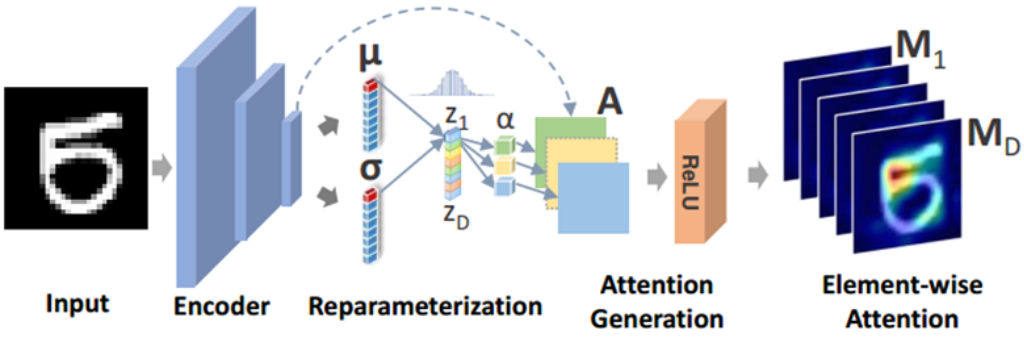

图 1. 注意力生成图

具体来说，给定经过训练的 VAE 对数据样本 x 推断的后验分布 $q(z|x)$，使用重参数化的技巧，以获得一个潜在的向量 z。对于每个元素 z_i，我们将梯度反向传播到最后一个卷积特征映射 $A \in \mathbb{R}^{n\times h\times w}$，得到 z_i 对应的注意映射 M^i。M^i 计算为线性组合:

$$M^i = ReLU(\sum_{k=1}^{n} \alpha_k A_k) \tag{2}$$

其中，α_k 是一个标量，是关于潜在编码 z_i 中的一个特征通道 A_k 的偏导数做全局平均池化（GAP）操作后的结果，GAP 操作就是计算矩阵的平均值，将所有元素相加并除以元素的总数。这一过程旨在提取特征通道 k 对输出的影响程度，用于生成注意力图。

$$\alpha_k = \frac{1}{T}\sum_{p=1}^{h}\sum_{q=1}^{\omega}\frac{\partial z_i}{\partial A_k^{pq}} \tag{3}$$

式中 $T = h\times w$,A_k^{pq} 为 $h\times w$ 矩阵 A_k 在位置 (p,q) 处的像素值。现在，我们对 D 维潜在空间的所有元素 z_1，z_2，$\cdots$，z_D 重复此操作，得到 M^1 ，$\cdots$，M^D。在这种情况下，整体注意图为 $M = \frac{1}{D}\sum_i^D M^i$

3.3 产生异常注意解释

用一个单类 VAE 与它所训练的数据，即正态数据 (例如数字“1”) 进行推理，理想情况下应该得到代表标准正态分布的学习潜空间。因此，给定来自不同类别的测试样本 (例如异常数据，数字“5”)，与学习到的正态分布相比，学习编码器推断的潜在表示应该有很大的差异。这种直觉可以通过很多方式体现出来。一种直接的方法是采用推断的平均向量并生成结果的注意图。另一种方法是使用正态差异分布。给定用于训练 VAE 的所有正态图像，我们可以推断出代表所有正态图像 x∈X 的嵌入分布的总体 μ^x 和 σ^x。现在，给定为异常样本 y 推断出的每个潜在变量 z_i 的 μ_i^y 和 σ_i^y，我们可以定义正态差分布为:

$$P_{q(z_i|\ x)-q(z_i|\ y)}(u) = \frac{e^{-\left[u-(\mu_i^x-\mu_i^y)\right]^2/\left[2\left((\sigma_i^x)^2+(\sigma_i^y)^2\right)\right]}}{\sqrt{2\pi\left((\sigma_i^x)^2+(\sigma_i^y)^2\right)}} \tag{4}$$

对于每个潜在变量 z_i。给定从 $P_{q(z_i|X)\ -\ q(z_i|Y)}$ 采样的潜在代码 z，可以按照上面描述的过程计算异常注意图 M。图 2 直观地总结了这一点。

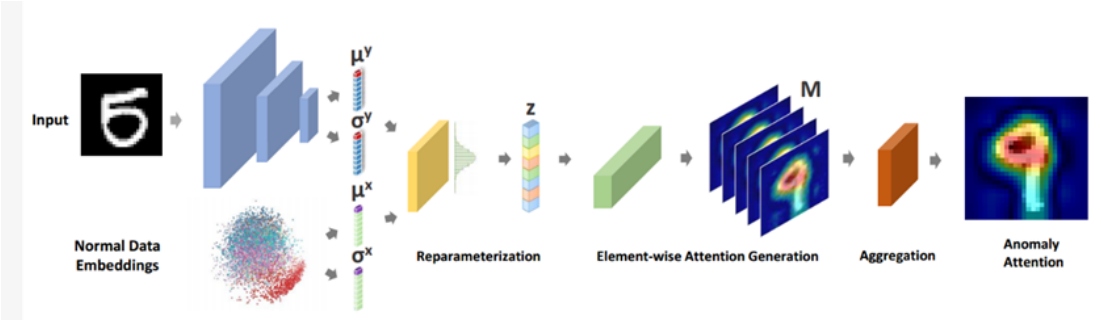

图 2. 异常注意图

3.4 注意力解纠缠

现有的学习深度生成模型的解纠缠表示的方法侧重于制定分解的、独立的潜在分布，以学习可解释的数据表示。例如，β-VAE [10]、InfoVAE [27] 和 FactorVAE [13] 等，它们都试图用因子概率分布对潜在先验进行建模。在这项工作中，作者提出了一种替代技术，基于作者提出的 VAE 注意力，称为注意力解缠损失。研究展示了如何将其与现有的基线 (例如，FactorVAE) 集成，并通过使用标准解缠度量的定性注意图和定量性能表征的方法演示了由此产生的影响。研究的大致思路是使用注意力图作为可训练的约束，明确地迫使从潜在空间的各个维度计算的注意力尽可能地分离或分离。论文的假设是，如果能够做到这一点，我们将能够学习一个改进的解纠缠潜在空间。为了实现这一目标，作者提出了一种新的损失，称为注意力解纠缠损失 (attention disentanglement loss, LAD)，它可以很容易地与现有的 VAEtype 模型集成。本文提出的 L_{AD} 以两个注意图 A^1 和 A^2 作为输入，并试图尽可能地分离其中的高响应像素区域。这可以在数学上表示为:

$$L_{AD} = 2 \cdot \frac{\sum_{ij} min(A^1_{ij}, A^2_{ij})}{\sum_{ij} A^1_{ij} + A^2_{ij}} \tag{5}$$

其中 $\cdot$ 为标量积运算，A^1_{ij} 和 A^2_{ij} 分别为注意图 A^1 和 A^2 中的第 (i,j) 个像素。建议的 L_{AD} 可直接与标准 FactorVAE 训练目标 L_{FV} 结合，从而得到一个整体的学习目标，可表示为:

$$L = L_{FV} + \lambda L_{AD} \tag{6}$$

解纠缠过程如图 3 所示。

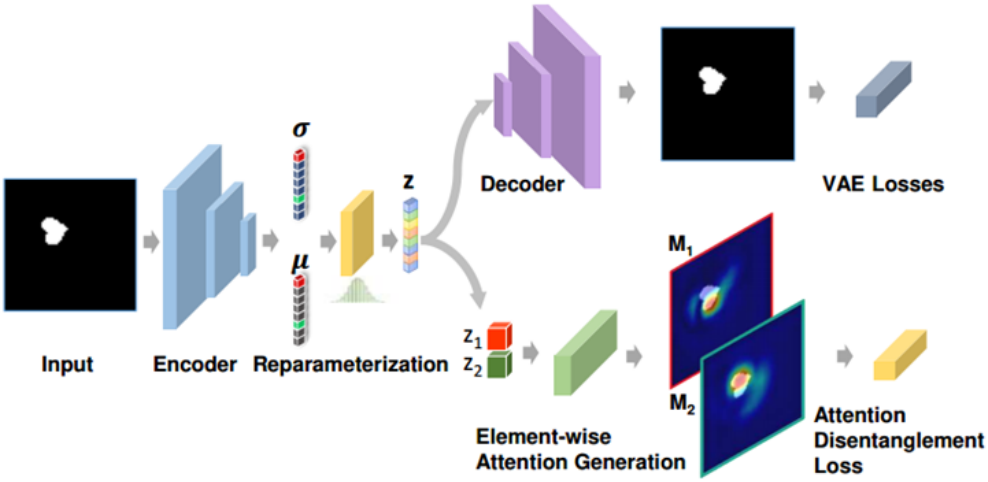

图 3. 解纠缠过程图

4 复现细节

4.1 与已有开源代码对比

使用了 GitHub 上两位作者的复现代码作为论文内容复现的代码参考。首先使用前一位作者代码对从定性地评估 MNIST 数据集上的视觉注意力图开始。使用来自单个数字类的训练图像，我训练了一个类 VAE 模型，该模型将用于对所有数字的测试图像进行测试。我将所有的训练和测试图像重塑为 28 × 28 像素的分辨率。

然后在另一位作者的代码基础，更改了测试数据集进行异常检测的分析，将原本的 UCSD Ped1 数据集替换成 MVTec-AD 数据集。不仅进行了定性的视觉评估，而且进行了定量的分析。我使用 ResNet18 作为特征编码器和 32 维潜在空间来训练 VAE。并且，我进一步使用原始作品中的随机镜像和随机旋转来生成增广训练集。给定一个测试图像，推断其潜在表征 z 来生成异常注意图。给定异常注意图，我使用像素响应值的各种阈值生成二元异常定位图，并将其封装在 ROC 曲线中。然后，我们计算并报告 ROC 曲线下的面积 (ROC AUC)，并根据 ROC 曲线上的 FPR 和 TPR 为研究的方法生成最佳借据数。

此外，我参考作者代码对不同的和注意力地图结合的 VAE 模型进行了测试并比较它们的解纠缠性能，同时将其可视化。

两位作者代码链接如下：

https://github.com/liuem607/expVAE

https://github.com/FrankBrongers

4.2 实验环境搭建

使用的实验环境为 anaconda 创建的虚拟环境，其各个工具包的版本如下：python 3.8.5 pytorch 1.7.0 torchvision 0.8.1 opencv 4.5.0 matplotlib 3.3.3 tqdm 4.56.0

4.3 使用说明

4.3.1 异常检测（MNIST）

使用 MNIST 数据集 [6] 进行异常检测和对 VAE 模型进行可视化解释。

训练 VAE 模型，在 expVAE_1 文件夹下使用 python train_expVAE.py 命令，其他可选选项为 [-h] [–result_dir DIR] [–ckpt_dir DIR] [–batch_size N] [–epochs N] [–seed S] [–resume PATH] [–latent_size N] [–one_class N]。这里 result_dir 保存验证可视化;Ckpt_dir 在每个 epoch 之后保存最佳模型;latent_size 决定了 VAE 的潜在向量形状;One_class 决定使用哪个数字作为训练的初始值。

测试 VAE 模型，并显示异常。在 expVAE-master 文件夹下使用 python test_expVAE.py 命令，其他可选选项为 [-h] [–result_dir DIR] [–batch_size N] [–seed S] [–latent_size N] [–model_path DIR] [–one_class N]。这里 one_class 决定离群数，expVAE 生成异常定位注意图，高响应区域表示离群数与内位数的差值。

4.3.2 异常检测（MVTec-AD）

使用 MVTec-AD 数据集 [2] 进行异常检测 。

训练 VAE 模型，在 expVAE_2/Anomaly_Detection 文件夹下使用 python code/train_expVAE.py –dataset=mvtec_ad –model= resnet18_3 –batch_size=8 –one_class=5 命令，即使用 mvtec_ad 数据集训练 resnet18 模型，批量为 8，训练的图片种类是第 5 种榛子。

测试模型，在 expVAE_2/Anomaly_Detection 文件夹下使用 python code/test_expVAE.py –dataset mvtec_ad –model resnet18_3 –batch_size 8 –model_path ./ckpt/resnet18_3_mvtec-Class_5_checkpoint.pth –one_class 5 –target_layer encoder.layer2.1.conv1 命令。

4.3.3 注意力解纠缠

使用相应的笔记本文件运行所有需要的实验。参见 Anomaly_Detection/了解异常检测的实现，以及 Latent_Space_Disentanglement/了解潜在空间解纠缠的实现，以及更详细的描述。

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。

5.1 可视化生成模型

通过注意力图直观地解释变分自编码器。潜在向量 (这里是 z_1-z_3) 中的每个元素都可以用注意力图单独解释，从而在不同样本中可视化一致的解释。图 4 显示了每个 M_i 所代表的示例，其中我们看到了跨多个数据样本的每个潜在维度的一致高响应区域。

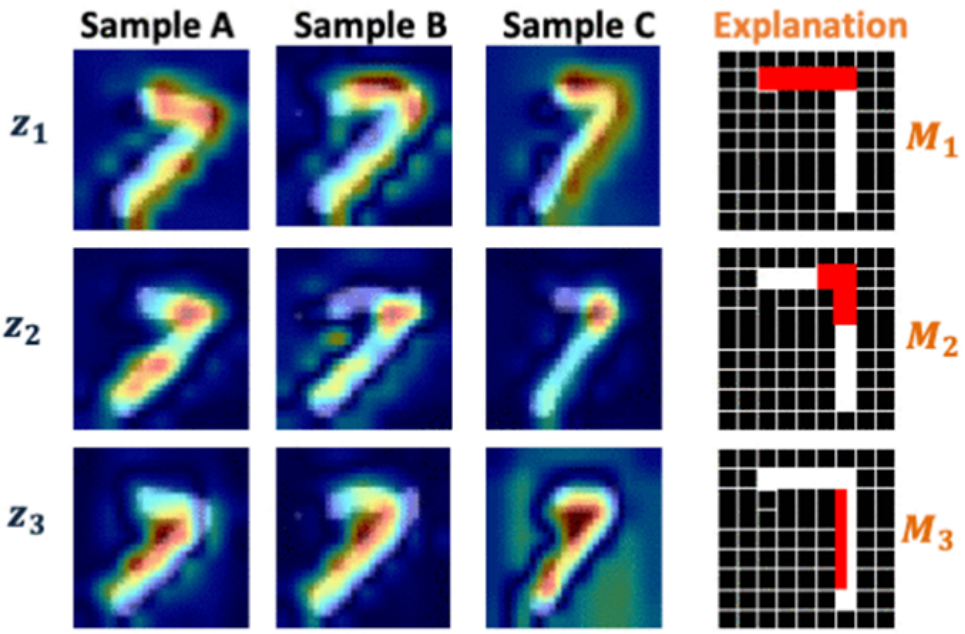

图 4. 潜在维度响应图

5.2 异常检测（MNIST）

如图 5 所示，我使用数字“1”和“3”(正常类) 训练的模型显示结果，并对不同数字 (成为异常类) 进行测试。对于每个测试图像，我们使用训练好的编码器推断潜在向量并生成注意图。从结果中可以看出，使用该方法计算的注意图在直观上是令人满意的。例如，让我考虑以数字“7”作为测试图像生成的注意图。我们的直觉告诉我们，“1”和“7”之间的关键区别在于“7”的顶部水平条，我们生成的注意力地图确实突出了这个区域。考虑以数字“8”作为测试图像生成的注意图，将其与以“3”作为正常类的模型测试。直觉告诉我们，“3”和“8”之间的关键区别在于“8”的左侧部分，可以看到，我们生成的注意力地图也突出了这个区域。

图 5. 异常检测（MNIST）图

5.3 异常检测（MVTec-AD）

我考虑了综合异常检测数据集:MVTec 异常检测 (MVTec AD)，它提供了多目标、多缺陷的自然图像。该数据集包含 5354 张不同物体/纹理的高分辨率彩色图像，在测试集中提供了正常和缺陷 (异常) 图像。我们将所有图像调整为 256 × 256 像素用于训练和测试。我进行了广泛的定性和定量实验，并将结果总结如下。

我选择使用 ResNet18 [9] 作为特征编码器，并采用 32 维潜在空间来训练变分自动编码器 (VAE)。对于给定的测试图像，通过推断其潜在表示 z 来生成异常关注图。在得到异常关注图后，我们使用不同的阈值生成二元异常定位图的像素响应值，并将其整合到 ROC 曲线中。然后，我计算并报告 ROC 曲线下面积 (ROC AUC)，并根据 ROC 曲线上的假正例率 (FPR) 和真正例率 (TPR) 确定我们方法的最佳阈值。图 6 以榛子作为类别的特征曲线图，AUROC 分数为 0.83。

Category	AnoGAN	CNN Feature Dictionary	ours
Hazelnut	0.87	0.72	0.83
Leather	0.64	0.87	0.82
Bottle	0.86	0.78	0.87

表 1. 复现结果

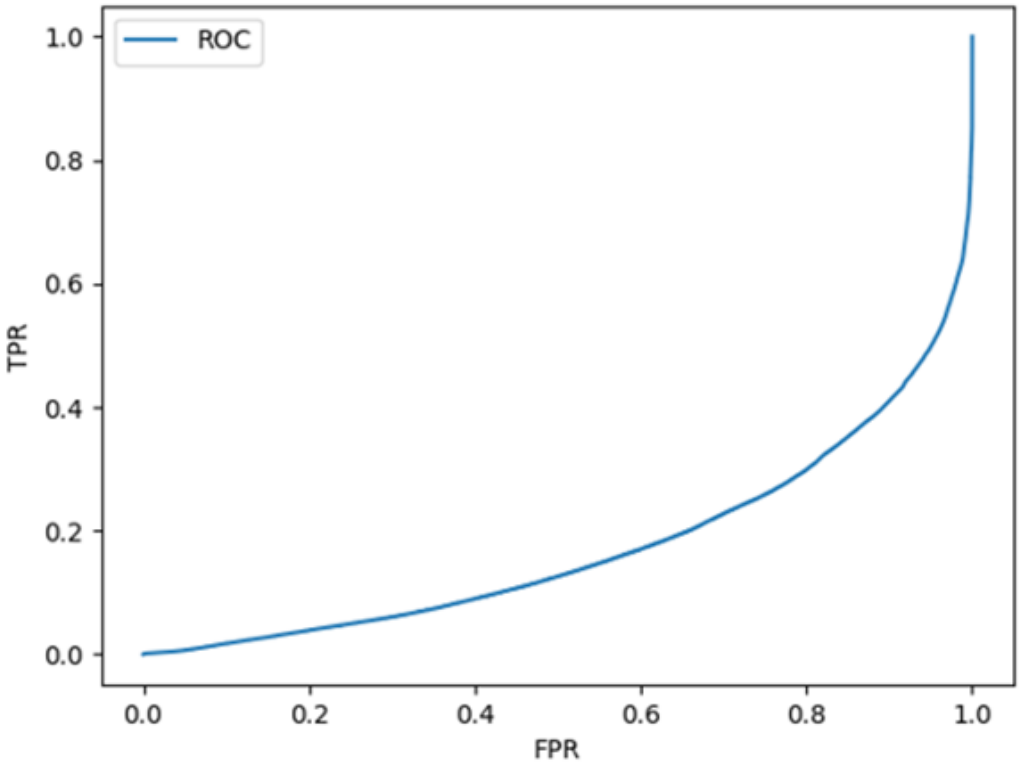

图 6. 接收器工作特征曲线图

表 1 显示了我的复现结果，并将其与 Bergmann 等人 [21] 在其基准论文中评估的技术进行了比较（这里的基线与 [21] 中的方法相同）。从复现结果中，我们注意到，使用作者提出的 VAE 关注的异常定位方法，在多数对象类别上我们获得了与竞争方法相差不大，甚至更好的结果。值得注意的是，其中一些方法是专门设计用于异常定位任务的，而作者则是训练了一个标准的 VAE，并生成了用于定位的 VAE 关注图。尽管方法相对简单，但它实现了有竞争力的性能，证明了这种关注生成技术的潜力，可用于除了模型解释之外的任务。

此外，图 7 展示了一些定性结果。对于榛子类别，我展示了原始图像和生成的异常关注图。我们可以观察到，我们的关注图能够准确地定位这些不同缺陷类别中的异常区域。

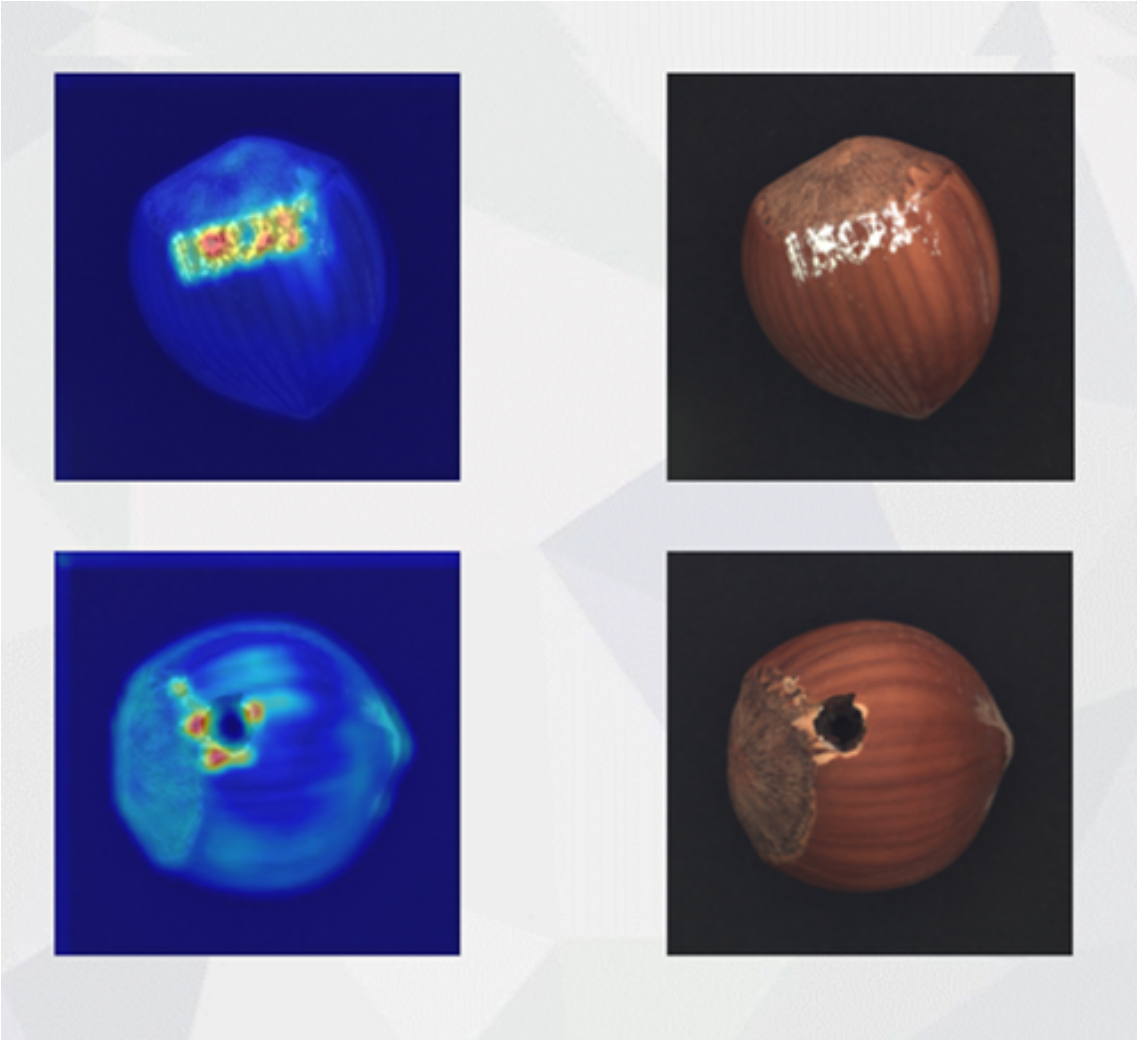

图 7. 榛子异常检测图

5.4 注意力解纠缠

我使用 Dsprites 数据集 [20]，该数据集提供 737,280 张二进制 64 × 64 2D 形状图像。我将提出的方法 (称 ADFactorVAE) 与其他竞争方法 (基线 FactorVAE(仅使用 L_FV 进行训练) 的最佳解纠缠性能 (根据重建误差绘制) 进行了比较。注意到，在相同的实验设置下，使用论文提出的 L_AD 进行训练可以获得更高的解纠缠分数，给出的最佳解纠缠分数约为 0.86，而基线 FactorVAE (γ= 40) 给出的解纠缠分数约为 0.82，两者的重建误差均在 40 左右，如图 8 所示。

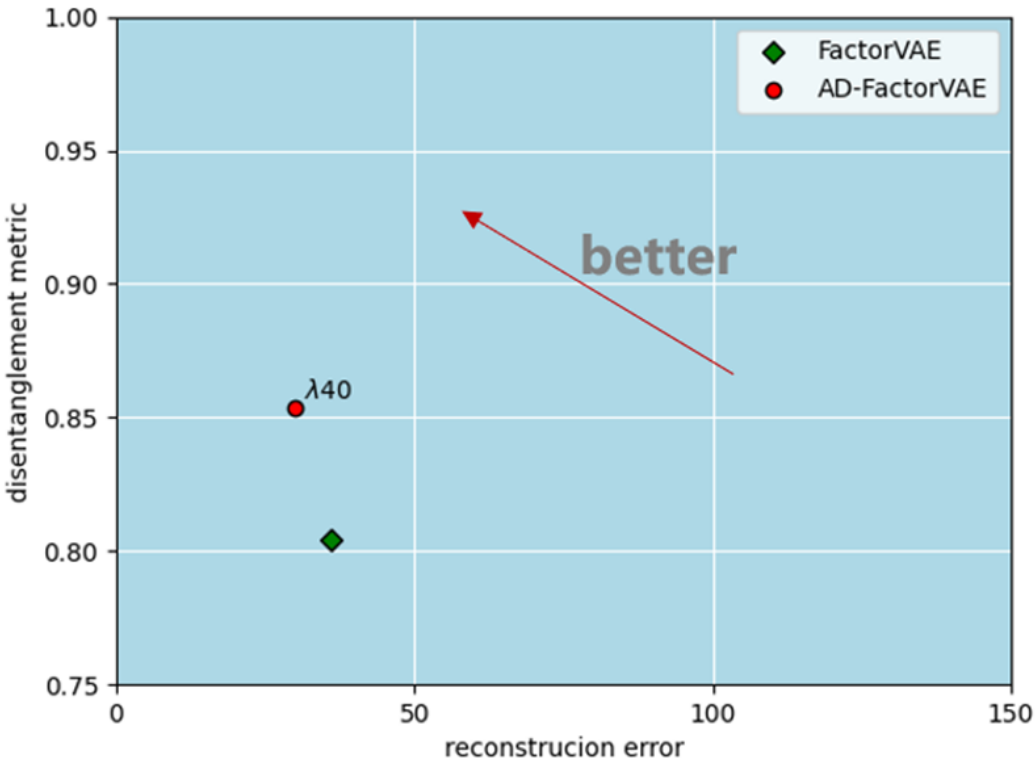

图 8. 解纠缠比较图

此外，我还比较了两种方法的解纠缠能力和重建损失随迭代次数的变化情况。如图 9 和 10 所示。

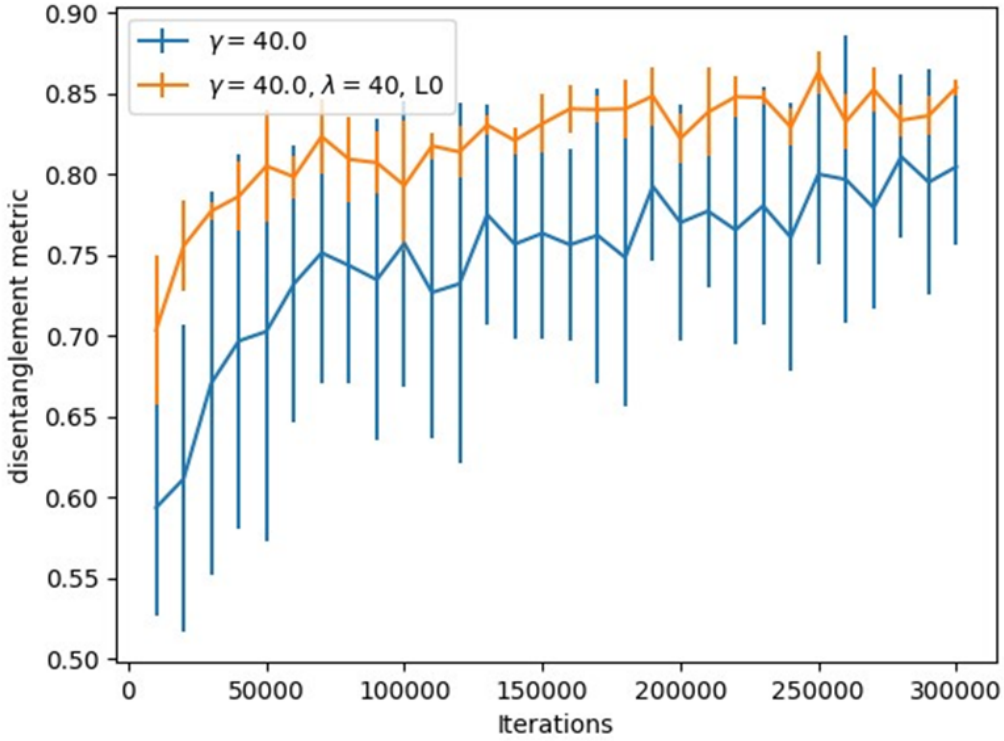

图 9. 解纠缠能力变化图

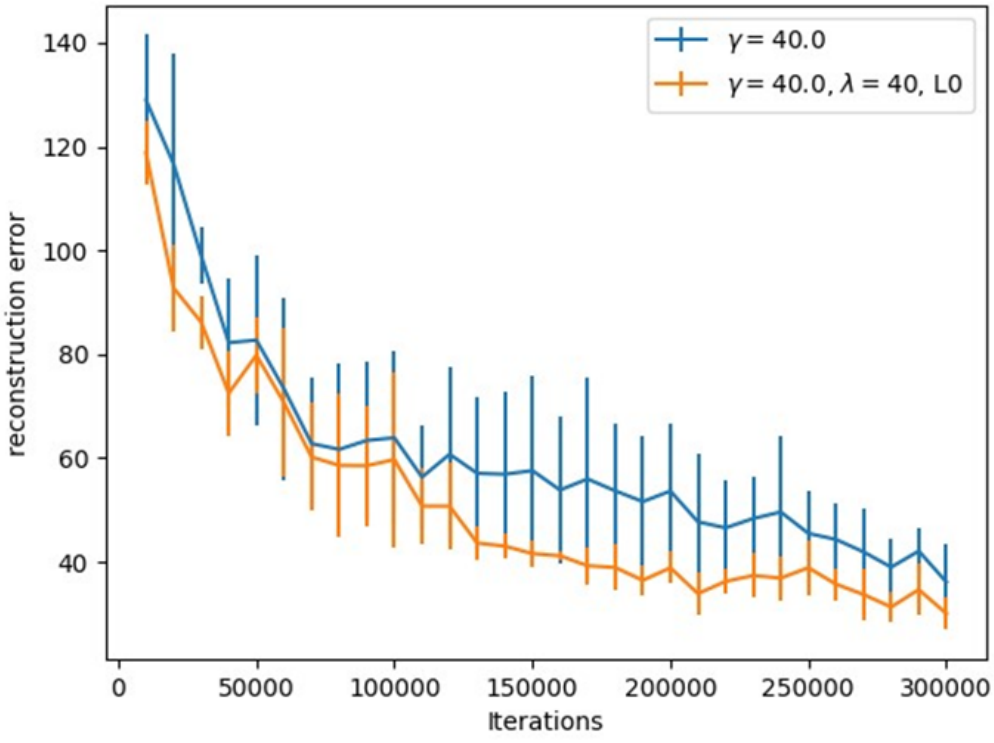

图 10. 重建损失变化图

6 总结与展望

作者提出了一种新的技术，通过基于梯度的网络关注机制，对变分自动编码器（VAE）进行视觉解释，迈出了解释深度生成模型的第一步。根据论文内容，我学会了如何利用学到的潜在表示计算梯度并生成 VAE 注意力图，而无需依赖于分类模型。我演示了该 VAE 注意力在异常定位和潜在空间解缠两个任务上的应用。在异常定位中，我们利用异常输入导致梯度反向传播和注意力生成中的潜在变量不符合标准高斯分布的事实。这些异常关注图被用作生成像素级二进制异常掩模的线索。在潜在空间解缠中，我展示了如何使用每个潜在维度的 VAE 关注来强制新的关注解缠学习约束，从而提高关注分离性和解缠性能。由于 VAE 能够推断完整的后验分布，使用作者的方法，可以通过重复抽样获得关注矩阵（图）的分布。虽然可以使用样本均值可视化这个分布，但对于整个矩阵分布生成更通用的视觉解释是未来研究的有趣主题。

通过这次复现，我不仅提升了实际应用技能，还深入理解了深度生成模型的解释性问题，并在解决技术挑战的过程中增强了问题解决的能力。然而，未来的研究还可以朝着更全面的方向发展。对于这项工作的未来展望包括：

进一步研究如何生成更通用的视觉解释，不仅仅局限于使用样本均值。可能的研究方向包括其他统计手段或可视化技术，以更全面地理解关注矩阵的分布。

在深度生成模型中，探索 VAE 注意力的更广泛应用，例如在其他深度生成模型中的解释性应用，以及在不同领域的任务中的适用性。

对于通过重复抽样获得的关注矩阵（图）的分布，研究如何生成更一般化的视觉解释，而不仅仅是使用样本均值。这将为深度生成模型的解释性提供更深入的理解。

总体而言，这次经验让我认识到深度生成模型解释性研究的潜在挑战，并鼓励我在未来的研究中进一步推动这一领域的发展。

参考文献

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. *ACCV*, 2018.

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. *CVPR*, 2019.

[3] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[4] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*, 2018.

[5] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *NeurIPS*, 2018.

[6] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[7] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. *CVPR*, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[10] Irina Higgins, Lo¨ıc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. *CVPR*, 2017.

[12] Dakai Jin, Dazhou Guo, Tsung-Ying Ho, Adam P. Harrison, Jing Xiao, Chen-Kan Tseng, and Le Lu. Accurate esophageal gross tumor volume segmentation in pet/ct using two-stream chained 3d deep network fusion. *MICCAI*, 2019.

[13] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *ICML*, 2018.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

[15] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. *CVPR*, 2019.

[16] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Guided attention inference network. *IEEE T-PAMI*, 2019.

[17] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. Towards visually explaining variational autoencoders. *arXiv preprint arXiv:1911.07389*, 2019.

[18] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. *CVPR*, 2015.

[19] Emile Mathieu, Tom Rainforth, Siddharth Narayanaswamy, and Yee Whye Teh. Disentangling disentanglement. *ArXiv,abs/1812.02833*, 2018.

[20] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. *https://github.com/deepmind/dsprites-dataset/*, 2017.

[21] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attentionguided image-to-image translation. *NeurIPS*, 2018.

[22] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Muller, and Marius Kloft. Deep one-class classification. *ICML*, 2018.

[23] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *ICCV*, 2017.

[24] Lezi Wang, Ziyan Wu, Srikrishna Karanam, Kuan-Chuan Peng, Rajat Vikram Singh, Bo Liu, and Dimitris N. Metaxas. Sharpen focus: Learning with attention separability and consistency. *ICCV*, 2019.

[25] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *ECCV*, 2014.

[26] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *IJCV*, 126:1084 – 1102, 2016.

[27] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae:information maximizing variational autoencoders. *ArXiv,abs/1706.02262*, 2017.

[28] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CVPR*, 2016.

[29] Yiming Zuo, Weichao Qiu, Lingxi Xie, Fangwei Zhong, Yizhou Wang, and Alan L. Yuille. Craves: Controlling robotic arm with a vision-based economic system. *CVPR*, 2019.