Automatic Correction of Internal Units in Generative Neural Networks

Ali Tousi1,*, Haedong Jeong1,2,*, Jiyeon Han1 , Hwanil Choi1 , Jaesik Choi1,3,†

摘要

通过设计复杂的网络结构和对抗性训练方案，生成对抗网络（GANs）在合成图像生成方面表现出了令人满意的性能。尽管 GANs 能够合成真实的图像，但存在许多生成的图像具有有缺陷的视觉模式，这被称为失真。虽然最近的大多数工作都试图通过干扰潜在变量来修复图片生成，但很少有人研究生成器的内部神经元来修复它们。在这项工作中，我们设计了一种方法，自动识别生成各种类型失真图像的内部单元。我们进一步提出了一种序列校正算法，该算法通过修改检测到的失真单元来调整生成流程，以在保持原始轮廓的同时提高生成质量。我们的方法在 FID 评分方面优于对照组的方法，并在人工评价中显示了令人满意的结果。

关键词：计算机视觉，模式识别

1 引言

在生成对抗网络中，已产生了非常多生成模型结构，其性能也达到了令人满意的地步。且训练一个模型所需要的人力物力达到企业级。人们开始研究如何基于已有的模型进行性能的改进。本人的研究方向为人脸编辑，为了深入了解此领域的前沿技术，以对 GAN 网络和人脸的编辑和生成有更多的了解，积累更多相关技术，在未来更好地从事相关工作。

2 相关工作

2.1 生成对抗网络

自引入 GAN 以来，生成器的输出图像的真实性和多样性稳步增加。一般情况下，GAN 模型输入一个采样的潜在向量，并输出一个合成的图像。虽然来自原始 GAN 模型的样本可以很容易地识别出来，但最近的模型产生了与真实数据难以区分的样本。尽管最近取得了进展，但很少有研究来了解 GAN 模型的内部机制。

2.2 描述深度网络的单元

人们提出了各种技术来检查和理解深度网络的内部表示。解释性热图可以用来解释单个网络决策。热图可视化了哪些输入区域对网络给出的分类预测贡献最大。

2.3 生成模型中的失真

通过深度神经网络，可以在合成图像中观察到缺陷区域。[1] 发现有/没有人工监督的人工诱导神经元。他们将每个神经元的最高激活图像可视化，并将其标记为正常单位或伪影单位。然后，他们消融有缺陷的神经元，以修复生成器。

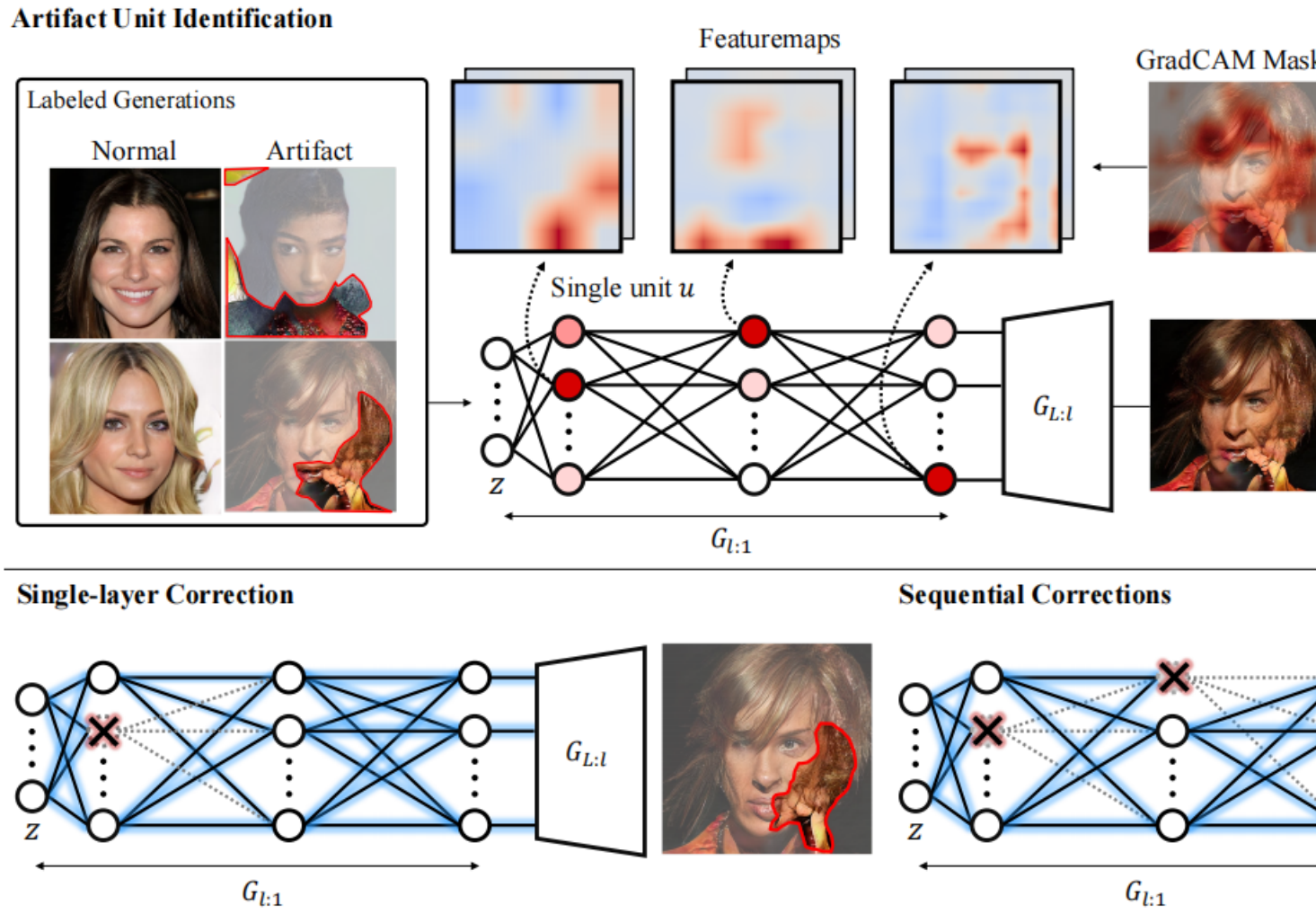

图 1: 方法示意图

3 本文方法

3.1 本文方法概述

在论文的方法中，首先根据预定义的标准将随机抽样的生成图像标注为两个类别，正常的和失真的。然后对所有已经标注的生成图像和一些随机采样的真实图像进行分类器训练，将图像分类为相应的类别。我们训练的分类器可以通过使用解释方法 Grad-cam 来生成缺陷区域的估计掩模。通过测量单个内部神经元的激活和缺陷区域的分割掩模之间的对齐，我们识别了诱发失真的神经元。为了纠正生成图像中的伪影区域，我们消融了重叠得分最高的神经元。如图 1所示：

3.2 基于分类器的工件单元识别

为了识别导致高级语义失真的内神经元，我将 2k 张生成的图像手工标记为正常或失真图像。然后，建立一个模型，将数据集分为三类，即：伪样本、正常和随机选择的真实样本。我们使用一个图像分类器（如 ResNet-18）作为我们的特征提取模块，并在其之上引入一个完全连接的层进行分类。在训练过程中，我们保持特征提取模块的参数固定，并且只优化分类器全连接层的权值。为了获得与图像标签相比更精细的注释，我们应用 GradCAM 为我们提供了一个缺陷区域的掩模。这样的掩模突出了对模型决策有效的区域。这种操作比手工标记缺陷区域更有效，因为后者需要很多时间。现在我们有了一个失真区域的掩模，比较特征激活映射图和掩模可以揭示失真诱导神经元。为此，对于生成器第 l 层的每一个单元 u，我们计算 $A_u(z_x) \in \mathbb{R}^{H_l \times W_l}$, 这是一个给定生成图像 x 和相应的潜在变量 z_x

的激活值。我们用分位数 T_u 作为激活阈值，使 $P\left(A_u\left(z_x\right)>T_u\right)=\tau$。这是根据所有图像的每个神经元的特征图分布来计算的。在应用阈值后，我们将结果双向上采样到失真掩模 $L_a(x)$ 的大小，并使用 $A_u(z_x)$ 计算 IoU。为了使识别过程全局化，我们定义了缺陷分数，即每个失真生成图像的平均值。

3.3 缺陷分数定义

生成图像为 X_a, 失真掩模为 $L_a(x)$, 对于生成图像 $x \in Xa$，第 l 层神经元 u 的激活值为 $A_u(z_x)$。神经元 u 的缺陷分数定义为：

$$DS_{l,u,a}=\frac{1}{|X_a|}\sum_{x\in X_a}\frac{|A_u(z_x)\cap L_a(x)|}{|A_u(z_x)\cup L_a(x)|}$$

第 l 层中所有单元的缺陷分数被定义为 $DS_{l,a}=\{DS_{l,1,a},DS_{l,2,a},...DS_{l,D_l,a}\}$。其中 D_l 是第 l 层的神经元数。最后，我们可以对分数进行排序，选择分数较高的神经元作为消融候选。

4 复现细节

4.1 与已有开源代码对比

由于原论文没有附带代码，因此本工作主要为复现论文中的代码。其中，该论文又与相关参考文献[1]具有类似功能，因此本工作是基于该参考文献的代码进行改写，实现本论文的功能。参考论文主要包括一下几个实现步骤：1. 计算指定层所有神经元的激活特征映射图（引用参考文献的代码）。

Procedure 1 Activation Computation

Input: latent code z
Output: Activation h_{k+1}
for $k\leftarrow$ 0 to l **do**
| $h_{k+1}=f_{k+1:k}(h_k)$ $h_{k+1}=upsample(h_{k+1})$
end

2. 计算指定层的所有神经元的阈值（引用参考文献的代码）。

Procedure 2 Threshold Computation

Input: Activation h_{k+1}
l: a stopping layer,n:number of units,$\tau=0.005$
Output: Threshold T_u
for $k\leftarrow$ 0 to l **do**
  for $u\leftarrow$ 0 to n **do**
  | $T_u \leftarrow P\left(h_{k+1}>T_u\right)=\tau$
  end
end

3. 计算通过阈值后的的激活映射图（自己编写）。

Procedure 3 Threshold Activation Computation

Input: Activation map h_{k+1} Threshold T_u
Output: Threshold Activation A_u
$A_u=h_{k+1}>T_u$

4. 计算随机生成图像的 masks（自己编写）。

Procedure 4 Mask Computation

Input: pretrained Resnet18 *Resnet18()* Generated image I_G
Gradcam: extract the mask from the network about the specified image
Output: Mask *mask*
$mask=Gradcam(Resnet18(I_G))$

5. 计算通过各层各神经元的 DS（自己编写）。

图 2: Caption

Procedure 5 Threshold Activation Computation

Input: Threshold Activation A_u Mask *mask* the set of artifact generations X_a
l: layer,u:unit **Output:** (Defective Score *DS*

$$DS_{l,u,a} = \frac{1}{|X_a|} \sum_{x \in X_a} \frac{|A_u(z_x) \cap L_a(x)|}{|A_u(z_x) \cup L_a(x)|}$$

4.2 实验结果分析

4.2.1 Resnet18 分类准确度

利用在 CelebA-HQ 数据集上训练过的 PGGAN 生成的 2k 张图像，并把它们手工便签为 1k 张正常图像和 1k 张失真图像，再加上 1k 张真实图像，作为数据集投入到 Resnet18 中训练，模型使用 Adam 优化器，在经过 40 个 epoches 的训练之后，得到了如下是训练结果：

model	loss	accuracy
resnet18	0.00337	0.927

4.2.2 GAN 序列消融实验分析

我对在 CelebA-HQ 数据集上训练的 PGGAN 进行校正。首先为该模型收集 100 张失真图像及其潜在代码。我们通过模型获得特征映射图，并消融所选择的特征映射图神经元。在这个实验中，我们使用了消融停止层 l = 6，消融神经元各层比例 n = 20%，以及比例因子 λ = 0.9。部分消融结果如图 2：

4.3 创新点

本实验对原论文中计算 DS 的方法进行了改进，由原公式（1）变为公式（2）, 实验结果表明，优化的效果有所提高。

$$DS_{l,u,a} = \frac{1}{|X_a|} \sum_{x \in X_a} \frac{|A_u(z_x) \cap L_a(x)|}{|A_u(z_x) \cup L_a(x)|} \quad (1)$$

$$DS_{l,u,a} = \frac{1}{|X_a|} \sum_{x \in X_a} |A_u(z_x) \cap L_a(x)| \quad (2)$$

5 总结与展望

本实验的消融时间有待提高，且消融效果也需要提高，对于一些更加明显到失真的图像，Resnet18 模型没能精确获取其失真区域的热图，导致在消融过程中也未能消融正确的部分。同时在计算 DS 的问题上，希望未来能设计出一个更好的方案，使得消融效果再提高。

参考文献

[1] BAU D, ZHU J Y, STROBELT H, et al. Understanding the role of individual units in a deep neural network[J]. Proceedings of the National Academy of Sciences, 2020, 117(48): 30071-30078. DOI: 10.1073/pnas.1907375117.