基于 BEGAN 的复现、学习率调优和各类 GAN 模型的分析比较

Learning rate tuning of BEGAN and comparison of different GANs

Yiran Hu
Tianjin University
jslshyr307@tju.edu.cn

摘要 ABSTRACT

本技术报告总结了对抗神经网络的发展过程、理论与应用领域，首先完成了对 BEGAN 的复现和训练。BEGAN 相比于 GAN 而言，其使用一种新的评价生成器生成质量的方式，使 GAN 即使使用较为简单的网络，不加训练 trick 或使用 SELU 激活函数也能实现较好的训练效果，不必再担心模式崩溃和训练不平衡问题。在成功复现并训练 BEGAN 之后，设计实验，对 BEGAN 中判别器和生成器的初始生成率进行参数调优。综合 Tensorboard 可视化的收敛性结果和生成图片的质量比较，发现模型的初始学习率设置为 0.0001 左右可以使模型的训练得到理想的结果。最后，以 FID、Precision、Recall 等为评价指标，从理论和实践出发，完成了对 BEGAN 模型和原始 GAN 模型之间的分析比较。结论发现，模型之间并没有绝对的优势，后续在改进模型时应该更多对结果的稳健性进行更认真的研究，而不是只观察抽样的误差。同时也对此比较模型提出了批判性的思考。

This technical report summarizes the development process, theory and application fields of adversarial neural networks, and firstly completes the reproduction and training of BEGAN. Compared with GAN, BEGAN uses a new way of evaluating the quality of generator generation, so that BEGAN no longer has to worry about mode collapse and training imbalance problems, even with simpler networks, or without training tricks or using SELU The activation function can also achieve better training results. After successfully reproducing and training BEGAN, experiments are designed to tune the parameters of the initial generation rate of the discriminator and generator in BEGAN. Based on the comparison of the convergence results visualized by Tensorboard and the quality of the generated images, it is found that the initial learning rate of the model is set to about 0.0001, so that the training of the model can obtain ideal results. Finally, the analysis and comparison between the BEGAN model and the original GAN model are completed with FID, Precision, Recall, etc. as evaluation indicators. It is concluded that there is no absolute advantage between the models, and more careful research should be conducted on the robustness of the results when improving the model, rather than just observing the sampling error. It also provides critical thinking on related papers published by the Google team.

关键词— BEGAN, 学习率调优, GAN, FID

1. 引言

生成式对抗网络 GAN (Generative adversarial networks) 在 2014 年一经提出，很快成为人工智能学界热门的研究方向之一。近年来，随着深度学习及移动设备的快速发展，图像处理[1]、图像风格迁移[2]、基于图像内容的检索与分类[3]、图像生成[4]等领域已经成为有巨大应用价值的课题。GAN 能够生成目标数据集，以弥补训练数据不足的缺陷，因而对深度学习意义重大; 此外 GAN 在场景生成、图像翻译、文本与图像的相互生成、视频预测等领域都发挥了独特的作用。而由于 GAN 本身自身存在的一些缺陷，原始 GAN 模型发生了许多的衍化。

由于数据集中图像内容复杂，规模较大，使用简单 GAN 模型很难控制生成的结果，机器理解的重点与人类理解存在偏差，最终导致生成结果与目标并不一致。一个自然的想法是增加约束条件，给生成器制定目标，文献[5]提出 CGAN；而 GAN 发展的一个里程碑，是 Radford 等人[6]提出的深度卷积生成对抗网络 DCGAN，将计算机视觉领域表现很好的卷积神经网络 CNN 与 GAN 结合起来，其为 CNN 的网络拓扑结构设置了一系列的限制来使得它可以稳定地训练; 而为了解决 GAN 训练中可能出现的梯度消失问题，文献[7]提出了 Wasserstein GAN(WGAN)，该模型使用 Wasserstein 距离 (又称为 Earth-Mover(EM) 距离)代替 JS 散度对真实样本和生成样本之间的距离进行度量，用一个批评函数 f 来对应 GAN 的判别器, 而且批评函数 f 需要建立在 Lipschitz 连续性假设上. 另外, GAN 的判别器 D 具有无限的建模能力, 无论真实样本和生成的样本有多复杂, 判别器 D 都能把它们区分开, 这使模型容易导致过拟合问题。

本文要介绍和复现的 BEGAN(Boundary Equilibrium Generative Adversarial Networks)，由 Google 在 2017 年提出，其和 GAN 的最大的区别在于他可以解决传统 GAN 存在的模式崩溃，难以训练，难以控制生成器和判别器的平衡等问题。DCGAN,WGAN,WGAN-GP 等都是使用了概率估计的方法，BEGAN 的做法相比之下更为

特殊，不去估计生成分布和真实分布的差距，而是估计分布的误差的分布之间的差距，此模型的发布者认为，若分布之间的误差分布也相近的话，也可认为这些分布是相近的[8]。

本技术报告，将会首先详细地介绍 BEGAN 的相关架构和原理，并选取 CelebA 数据集中的部分数据对 BEGAN 进行复现、训练和测试。在得到训练结果之后，对 BEGAN 中的 learning rate 参数进行调优，分别将判别器和生成器的 learning rate 的初始参数设置为 0.00008、0.001、0.0005、0.0001、0.00002，在保证训练集测试集验证集不变以及模型其他参数不变的情况下重复训练和测试，从人工检视生成图片质量和损失率收敛性变化的比较分析中，发现初始学习率设置为 0.0001 时模型的评估最优。

最终，由于计算资源、时间、疫情的影响，本人从谷歌团队发表的论文：Are gans created equal? a large-scale study[9]中获取灵感，借由谷歌团队开源的实验环境和测试平台完成了对 BEGAN 模型和 GAN 模型区别的分析比较。确实得到了论文中所描述的结果，即在谷歌团队设计的比较分析实验中，模型之间并没有绝对的优势。更进一步，本文批判性地对该论文的比较模型提出了思考，即该实验并不能完全反应不同 GAN 模型的特点，存在片面之处。

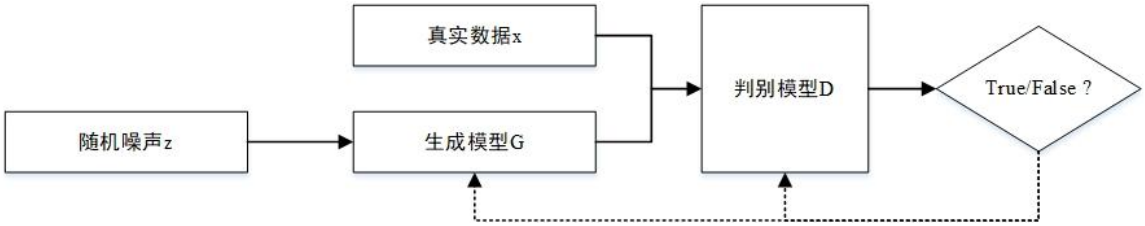

图 1：GAN 的计算流程和结构

在本节的最后，我将对本技术报告的主要内容与贡献总结如下：

1) 成功复现和训练 BEGAN 模型，掌握了模型的设计思想和原理，并利用 Tensorboard 进行了可视化呈现。

2) 设计实验对 BEGAN 中判别器和生成器的初始学习率进行实验和调优，发现在其他实验条件保持不变时，学习率设置为 0.0001 左右可以得到理想的训练结果。

3) 理论和实践相结合，对 BEGAN 和 GAN 两个模型进行比较分析，发现模型之间并没有绝对的优势，后续在改进模型时应该更多对结果的稳健性进行更认真的研究，而不是只观察抽样的误差。

4) 进一步对谷歌团队提出的比较不同 GAN 模型优劣的比较模型提出了批判性反思，指出了该比较模型的不足之处。

2. 研究方法

2.1. BEGAN 的复现与原理

BEGAN 的复现代码在此不做呈现和阐述。本节主要阐释 BEGAN 的原理。想要熟悉 BEGAN 的原理首先要从最基本的 GAN 的基本思想开始。

2.1.1. GAN 的基本原理与面临的问题

GAN 的核心思想来源是博弈论的纳什均衡。它设定参与游戏双方分别为一个生成器 G(Generator)和一个判别器 D(Discriminator)，生成器的目的是尽量去学习真实的数据分布，而判别器的目的是尽量正确判别输入数据是来自真实数据(true)还是来自生成器（false）；为了取得游戏胜利。两个游戏参与者需要不断优化，各自提高生成能力和判别能力，此优化过程就是寻找二者之间的一个纳什均衡[10]。

GAN 的计算流程与结构如图 1 所示。任意可微分函数都可表示 GAN 的生成器和判别器，由此,我们用可微分函数 D 和 G 来分别表示判别器和生成器，它们的输入分别为真实数据 x 和随机变量 z。 G(z) 则为由 G 生成的尽量服从真实数据分布的样本。如果判别器的输入来自真实数据，标注为 true(1)；如果输入样本为 G(z)，标注为.false(1)。D 的目标是实现对数据来源的二分类判别：true(来源于真实数据 x 的分布) 或者 false(来源于生成器的伪数据 G(z))，而 G 的目标是使自己生成的伪数据 G(z)在 D 上的表现 D(G(z))和真实数据 x 在 D 上的表现 D(x)一致，这两个相互对抗并迭代优化的过程使得 D 和 G 的性能不断提升，当最终 D 的判别能力提升到一定程度，并且无法正确判别数据来源时，可以认为这个生成器 G 已经学到了真实数据的分布。

GANs 确实可以生成十分逼真的图像，甚至比使用像素级损失的自动编码生成器生成的图像更清晰（BEGAN 的原英文论文），但是仍然有许多未解决的困难。首先，纵使使用了许多的技巧，模型仍然相比而言难以训练；其次，控制生成样本的图像的多样性是十

分困难的，因为正确的超参数选择十分重要；再次，比较难以平衡鉴别器和生成器的收敛性，并且很容易出现模式坍塌的现象。

2.1.2. BEGAN 设计架构

BEGAN 的开发团队给模型的全称是 Boundary Equilibrium Generative Adversarial Networks。BEGAN 中，使用一个自动编码器作为判别器 D。与普遍的 GANs 直接匹配数据分布不同，BEGAN 旨在匹配自动编码器的损失分布使用的 Wasserstein 距离。为了实现目标。模型使用典型的 GAN 目标并加上一个用于平衡生成器 G 和判别器 D 的平衡项。从而与典型的 GANs 模型相比，BEGAN 的训练过程更加简单，并且使用的申请网络模型也更加简洁。

如图 2 所示，判别器 D 是一个作为自编码器构建的卷积深度神经网络。$N_x = H \times W \times C$ 是 x 的尺寸缩写，其中 H 是高度，W 是宽度，C 表示颜色。为了防止典型的 GANs 模型中的问题，模型使用一个自动编码器与一个深度编码器和解码器。

模型使用 3×3 的卷积，并使用 ELUs 激活函数。通常每一个卷积层都会重复多次（通常为 2 次）。因为根据模型的开发者的说法，更多的重复会得到更好的视觉效果。在每次的降采样之后，卷积滤波器会以线性的方式增长。降采样是通过步长为 2 的子采样实现的；而上采样是通过最近的邻居实现的。

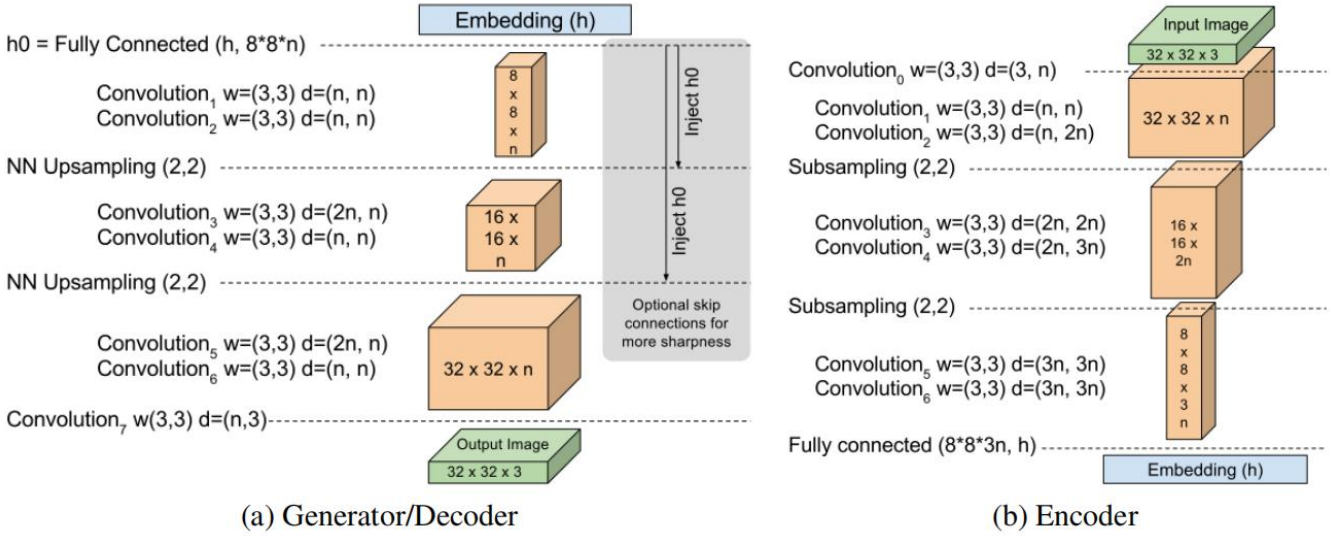

图 2：BEGAN 的设计架构

在判别器 G 和生成器 D 的边界之间，处理后的数据的张量通过完全连接层进行映射，从嵌入状态 $h \in R^{N_h}$ 之后没有任何非线性。其中 N_h 是自动编码器的隐藏状态的维度。

对于生成器 G，其使用的结构与判别解码器相同（虽然其中的权重有所差别）。而之所以采取这个模型，其中输入状态为 $z \in [-1, 1]^{N_x}$。

2.1.2.1. 自动编码器的 Wasserstein 距离下界

为了研究匹配误差分布的影响，BEGAN 模型首先引入自动编码器损失的概念（auto-encoder loss）之后计算证实样本和生成样本自动编码器的损失分布之间的 Wasserstein 距离下界。

定义训练一个像素级自动编码器的损失为：

$$\mathcal{L} : \mathbb{R}^{N_x} \mapsto \mathbb{R}^{+}$$

计算公式为：

$$\mathcal{L}(v) = |v - D(v)|^{\eta}$$

其中的参数解释如下：

$$\begin{cases} D : \mathbb{R}^{N_x} \mapsto \mathbb{R}^{N_x} & \text{is the autoencoder function.} \\ \eta \in \{1, 2\} & \text{is the target norm.} \\ v \in \mathbb{R}^{N_x} & \text{is a sample of dimension } N_x. \end{cases}$$

设 $\mu_{1,2}$ 为自动编码器损耗的两个分布；$\Gamma(\mu_1, \mu_2)$ 为 μ_1 和 μ_2 的耦合集合，以及 m_1 和 m_2 为其各自的均值。那么 Wasserstein 距离可以被表示为：

$$W_1(\mu_1, \mu_2) = \inf_{\gamma \in \Gamma(\mu_1, \mu_2)} \mathbb{E}_{(x_1, x_2) \sim \gamma}[|x_1 - x_2|]$$

并根据 Jensen 不等式，可以得到 Wasserstein 距离的下界为：

$$\inf \mathbb{E}[|x_1 - x_2|] \geqslant \inf |\mathbb{E}[x_1 - x_2]| = |m_1 - m_2| \quad (1)$$

2.1.2.2. GAN 的目标函数

并根接下来，要进一步设计判别器 D 将公式(1)最大化。不妨令 μ_1 是损失函数 L(x) 的分布，μ_2 是损失函数

L(G(z))的分布，其中 G 表示的是生成函数，$z \in [-1,1]^{N_x}$表示纬度为N_x的随机样本。由于$m_1, m_2 \in R^+$，有两种可能的表示将$|m_1 - m_2|$去掉绝对值，分别记为表示(a)与表示(b)：

$$(a)\begin{cases} W_1(\mu_1,\mu_2) \geqslant m_1 - m_2 \\ m_1 \to \infty \\ m_2 \to 0 \end{cases}$$

$$(b)\begin{cases} W_1(\mu_1,\mu_2) \geqslant m_2 - m_1 \\ m_1 \to 0 \\ m_2 \to \infty \end{cases}$$

BEGAN 选择表示(b)作为目标函数。因为在最小化m_1的过程中会导致真实图像的自动编码。之后，通过最小化损失得到判别器和生成器的参数θ_D和θ_G，由此可以将模型要解决的问题表示为 GAN 的目标函数，其中z_D和z_G都是 z 中的样本：

$$\begin{cases} \mathcal{L}_D = \mathcal{L}(x;\theta_D) - \mathcal{L}(G(z_D;\theta_G);\theta_D) & \text{for } \theta_D \\ \mathcal{L}_G = -\mathcal{L}_D & \text{for } \theta_G \end{cases} \quad (2)$$

为了简洁，在之后的原理介绍中，将使用 $G(\cdot)$代替$G(\cdot,\theta_D)$和 $G(\cdot,\theta_G)$。公式(2)的创新性在于，其匹配的是损失之间的分布而非样本之间的分布。

同时，因为神经网络与深度学习的应用领域当中，可能出现判别器和生成器之间不能很好的平衡，通常鉴别器很容易产生明显的优势（即生成器生成的图片很难迷惑判别器），因此 BEGAN 中还引入的平衡的概念。

图 3：CelebA 数据集示例

2.1.2.3. 平衡

在 GANs 模型当中，保证生成器和判别器的平衡是十分重要的：

$$\mathbb{E}[\mathcal{L}(x)] = \mathbb{E}[\mathcal{L}(G(z))] \quad (3)$$

如果生成器生成的图片可以通过判别器的识别，被识别为真实样本，那图片的误差分布应该是相同的。基于此原理，可以让生成器和判别器之间相互对抗促进互相的优化。

在 BEGAN 模型中，通过引入一个新的超参数$\gamma \in [0,1]$放松平衡，定义如下：

$$\gamma = \frac{\mathbb{E}[\mathcal{L}(G(z))]}{\mathbb{E}[\mathcal{L}(x)]} \quad (4)$$

在 BEGAN 模型中，判别器 D 有两个目标（这两个目标是相互竞争的）：自动编码的真实图像以及区分真实图像和生成的图像。超参数γ可以帮助平衡这两个目标。当γ值较低时，会导致较低的图样多样性，从而γ也被称为多样性比率。对于图像的清晰度和细致性是存在天然的边界的。

在 BEGAN 模型当中，其目标函数为：

$$\begin{cases} \mathcal{L}_D = \mathcal{L}(x) - k_t.\mathcal{L}(G(z_D)) & \text{for } \theta_D \\ \mathcal{L}_G = \mathcal{L}(G(z_G)) & \text{for } \theta_G \\ k_{t+1} = k_t + \lambda_k(\gamma\mathcal{L}(x) - \mathcal{L}(G(z_G))) & \text{for each training step } t \end{cases}$$

为了满足等式(4)，可以使用$k_t \in [0,1]$在梯度下降的过程中控制 LGzD 的权重。在 BEGAN 中，将k_0初始化为 0。λ_k是 k 的学习速率，在此模型中设置为 0.001，从而可以帮助等式(4)的成立。

与传统 GANs 模型相比，传统 GANs 需要教体训练判别器 D 和生成器 G，或者预训练判别器 D。而在 BEGAN 模型当中，在模型训练开始时不需要稳定的训练。此模型根据论文[10]使用了默认的超参数。参数θ_D和θ_G根据论文中提出的 Adam 优化器进行更新。模型的 batch size 为 n=16。

2.1.2.4. 收敛性测量

收敛性的测量可以利用平衡的概念。使用了如下的算法：

$$\mathcal{M}_{global} = \mathcal{L}(x) + |\gamma\mathcal{L}(x) - \mathcal{L}(G(z_G))|$$

此度量方法可用于确定网络何时达到最终的状态，或者确定此模型是否已经发生崩溃。

2.1.3. 模型的训练

以上就是 BEGAN 模型的基本原理和训练方式。在详细了解了 BEGAN 模型的原理之后，实验进一步对 BEGAN 模型进行了复现和训练。使用的数据集为 CelebA。CelebA 数据集中有 202599 张脸的照片，以及 10177 个名人身份。特征的示例如图 3 所示，包括 Eyeglasses、Bangs、Pointy Nose 等。数据集中每张图都做好了特征的标记，包括人脸的 bbox 的标注框、人脸特征点坐标以及 40 个人脸属性标记。由于算力的问题和本身时间的限制，本次模型的训练仅从数据集中选取其中的 50000 张图作为训练集。BEGAN 模型的训练结果在 3.1：BEGAN 的复现和结果呈现中，会具体呈现实验的参数配置和实验结果。

2.2 BEGAN 参数调优

在 BEGAN 的原始论文中，直接将学习率设置为了 0.0001，更多的是基于模型训练的经验，而没有给予过多的解释。为了深入了解在实际训练过程中学习率对模型训练和测试结果的影响，本实验聚焦学习率（learning rate）的调优，进一步从实验的角度体会其对 BEGAN 模型训练结果的影响，即不同的学习率在训练效果上会有什么不同。

理论上，一个合适的学习率对于网络的训练十分重要。如果学习率过高会导致梯度在最优解处来回震荡，甚至越过最优值。学习率太小，优化的效率可能过低，则会导致网络的收敛速度较为缓慢。一般而言，模型都会先采取较大的学习率进行训练，然后在训练的过程中不断衰减学习率。

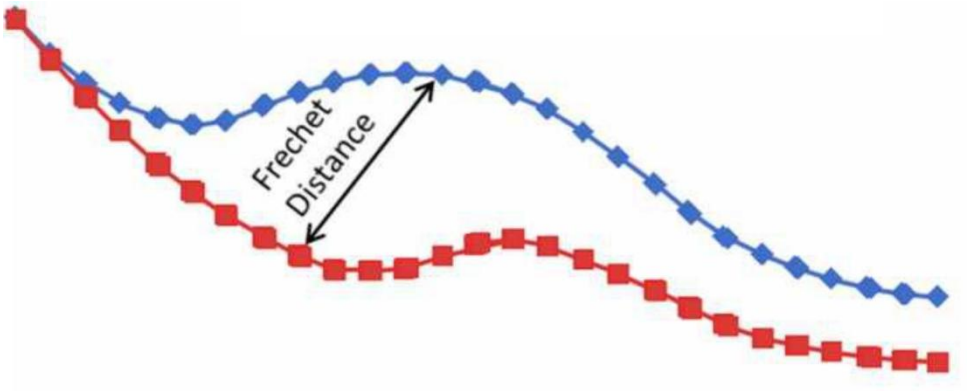

图 4：Frechet Distance 示例图

在自己复现的 BEGAN 当中，学习率优化使用常见的 Adam 优化器[11]，在此不对 Adam 优化方法进行过多介绍，详情见论文的引用[11]。在 tensorboard 的帮助下，可以可视化损失函数的收敛。

本报告将按照如下思路对学习率设计如下实验进行调参(在复现中，模型的生成器和判别器最初学习率参数均为 0.0001）：

(1) 模型中初始的学习率为 0.001，保证其他参数和训练模型相同，进行训练和测试；

(2) 将学习率调整为 0.0005，保证其他参数不变的情况下，进行模型的训练和测试；

(3) 将学习率调整为 0.0001，保证其他参数不发生改变的情况下，进行训练和测试；

(4) 将学习率调整为 0.00008，保证其他参数不发生改变的情况下，进行训练和测试；

(5) 将学习率调整为 0.00002，调整学习率的最底下界，保证其他参数不发生变化的情况下进行训练和测试。

在完成 5 轮训练和测试之后，从损失率的变化和生成的可视化图像进行比较，确认最佳的学习率的大致范围。具体的实验结果呈现在 3.2 节查看。

2.3 BEGAN 与原始 GAN 的分析比较

本部分将在阅读大量论文、参考他人的研究并自主实践的基础上，对 BEGAN 和 GAN 模型进行分析比较。

谷歌大脑团队的研究者发表的论文：Are gans created equal? a large-scale study[9] 中，对 MM GAN、NS GAN、WGAN、WGAN GP、LS GAN、BEGAN 等优秀的 GAN 模型与原始的 GAN 模型进行客观的比较，可是却发现这些模型并没有如不同论文中预期的那样，相比于原始的 GAN 有明显的优势。而本实验重点聚焦 BEGAN 和原始 GAN 模型，借用该论文的比较思路进行实践。

在该研究中，研究者介绍了两个评估指标来评估模型的性能。

第一个为 Inception Score(IS)。其提供了一种定量评价生成样本质量的方法。IS 的提出基于以下因素：（1）包含有意义对象的样本的条件标签分布应该具有比较低的熵；（2）生成器的生成的样本应该具有多样性，应该

尽量包含所有的类。谷歌的大脑团队将这些需求组合之后，得到如下 IS 的公式:

$$IS(G) = \exp(E_{x\sim G}[d_{KL}(p(y|x), p(y))])$$

团队发现 IS 也存在一些缺陷，比如对标签上的先验分布不够敏感。因此将指标 IS 排除在候选评价指标之外。

第二个指标为 Frechet Inception Distance(FID)。为了量化生成样本的质量，首先将 FID 嵌入到特征空间中。之后将嵌入层视为连续的多元高斯分布，对生成的数据和真实的数据同时估计其均值和协方差。之后使用两个高斯分布之间的 Frechet 距离对样本的质量进行量化，公式如下：

$$FID(x, g) = ||\mu_x - \mu_g||_2^2 + Tr(\sum_x + \sum_g - 2(\sum_x \sum_g)^{\frac{1}{2}})$$

其中，$(\mu_x, \sum_x)$和$(\mu_g, \sum_g)$是数据分布和模型分布的样本嵌入的均值和协方差。此指标在作者的测试中认为比 IS 对噪声更加稳定。但是这两种指标的共性缺点是没有办法检测出过拟合。FID 是发过数学家 Maurice René Fréchet 在 1906 年提出的一种路径空间相似形描述，类似于狗绳距离。描述来说，如图 4 所示，红色为主人的行走路径，蓝色为狗的行走路线，而 FID 距离则表示为各自走完这两条路径过程中所需的最短狗绳的长度。FID 与生成图像的质量呈负相关。

在使用 IS 和 FID 的基础上，需要进一步给出比较模型的方式。因为当改变超参数，随机种子或者数据集时，给定的评价指标也可能得到不同的分数和结果。根据论文，选取如下角度：

（1）对所有的模型采用相同的体系结构。

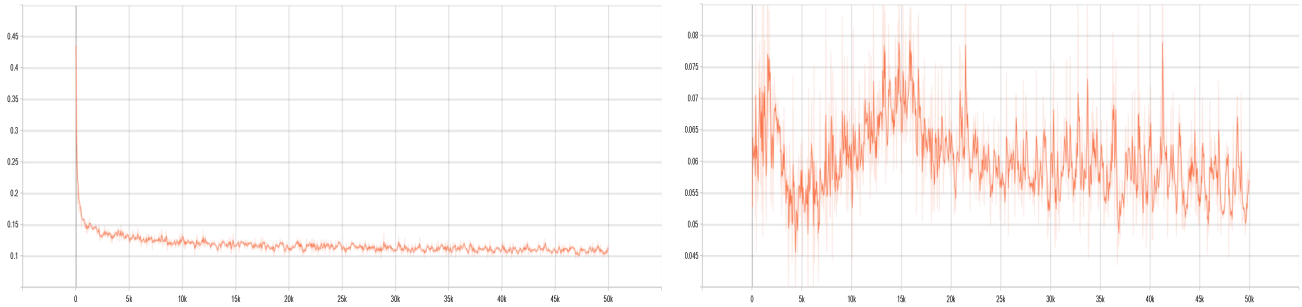

图 5：复现过程中损失函数变化曲线（左侧为判别器的损失函数收敛曲线，右侧为生成器损失函数的收敛曲线）

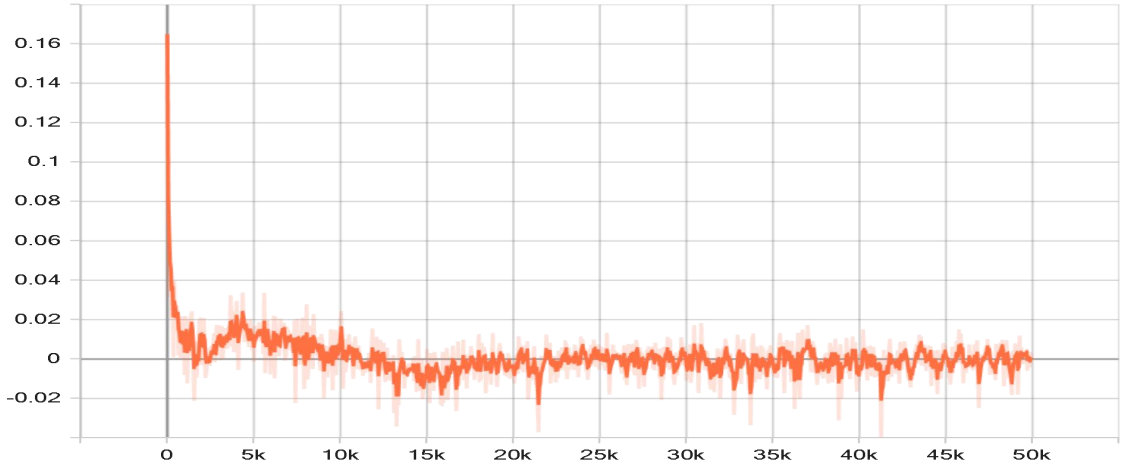

图 6：复现中 balance 指标的变化曲线

（2）对不同的模型进行相投算法的参数优化策略，因为不同的优化策略可能会影响最后的结果。

（3）从不同弄个 GAN 文献中选取 4 个流行常用的数据集，并为每个数据集报告结果。

除了文章中提出的 FID，常用的 Precision、Recall 和 F1 得分也广泛用于评价模型的质量。其中 F1 得分是表示精确率和召回率的调和平均值。而 IS 和精确率联系紧密，FID 则会同时捕获精确率和召回率。

实验搭建：为了保证比较的公平性，参照论文中的说法，原始 GAN 的生成器和判别器的结构都按照 INFO GAN 中的模型设计[12]。而 BEGAN 的判别器中仍然会使用自动编码器。将潜在代码空间（latent code size）设置为 64，并使潜在空间上的先验分布均匀分布在$[-1, 1]^{64}$上。两个模型中的优化算法都使用较为常见的 Adam 优化算法。在开始训练时都把 BEGAN 和 GAN 模型判别器和生成器的学习率设置为相同值：0.0001。

其余实验比较细节详见 3.3 的叙述。

3. 实验与结果分析

3.1 BEGAN 的复现

此部分基于之前介绍的 BEGAN 原理，主要介绍模型复现框架和模型中的参数呈现。代码主要使用 Python 编写。其中的一些重要的文件说明如下:layers.py 和 models.py 文件中实现了 BEGAN 的网络架构以及其中的激活函数损失函数等等，具体参照前文对 BEGAN 模型架构的介绍；config.py 文件中主要保存的是 BEGAN 网络架构中的各种参数；data_loader.py 文件主要负责将用于训练和测试的图样本集输入模型，以帮助模型进行测试和训练；trainer.py 文件主要帮助实现模型的训练过程，包括过程图片文件的生成和保存，过程参数的输出和保存等等。

BEGAN 模型中的各种参数设置见表 1。利用 Tensorboard，判别器和生成器的损失函数变化如图 5 所示。从展示的图以及判别器、生成器的损失变化来看，可见损失函数成功收敛。显然从纵坐标来看，随着训练过程的推进，损失函数逐渐降低并趋于稳定。以及图 6 呈现了生成器和判别器之间的平衡关系，从趋势上看,二者逐渐趋于平衡，在 0 上下有轻微的波动，体现了 BEGAN 中 Boundary Equilibrium 的核心内涵。在引文后的附录 A 中，呈现了模型训练的不同阶段所生成的图像，很明显可以观察到人脸图像变得逐渐清晰的趋势。

部分参数符号	参数值
batch_size	16
conv_hidden_num	128
discriminator_lr	0.00008
generator_lr	0.00008
gamma	0.5
grayscale	false
input_scale_size	64
lambda_k	0.001
lr_lower_bounary	0.00002
max_step	50000
lr_update_step	100000
optimizer	adam
randow_seed	123
sample_per_image	64
z_num	64
save_step	5000

表 1：BEGAN 复现部分参数示例

该模型的测试结果见引文最后的附录 B。可见，该模型成功的避免了模式坍塌的现象，生成的人脸较为多样。但是由于计算资源的问题，仅进行了 50000 轮的训练，如果将训练轮次调整为 200000 次或者 300000 次或更多，应该能得到更丰富的人脸生成图像和更清晰的图像。因为从目前的生成图中可看出部分图像中存在轮廓不够明显的现象。

3.2 BEGAN 参数调优

为了呈现不同初始学习率设置导致的训练结果的差别，本报告在保持其他参数和结构不变的情况下，分别将判别器和生成器的初始学习率设置为 0.001、0.0005、0.0001、0.00008、0.00001，分别进行模型的训练，分别对训练的模型进行图片的视觉比对，在无法通过人工分析出明显差别时，再借用 tensorboard 工具生成器损失函数变化、判别器损失函数变化、二者平衡指标等进行进一步模型优劣的比较。

学习率设置为 0.001 时，训练过程中生成器在训练 10k、20k、30k、40k、50k steps 之后生成的结果如图 7：

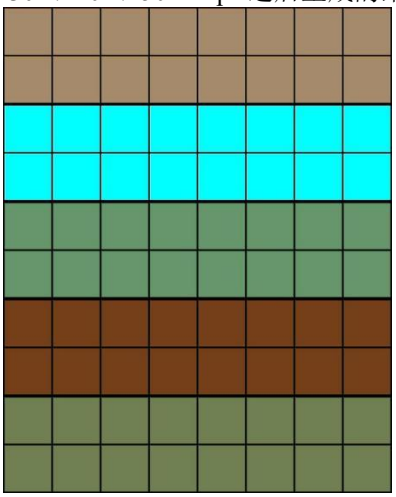

图 7：learning_rate = 0.001(图像从上往下，每两列分别表示在训练 10k、20k、30k、40k、50k steps 之后生成的结果)

很明显，对比 BEGAN 复现时 learning rate 为 0.0001 的结果，明显初始学习率设置为 0.001 时完全无法训练出可使用的模型，也无法生成出人脸，甚至只能生成纯色的图片。

学习率为 0.0005 时，训练过程中生成的结果如图 8：

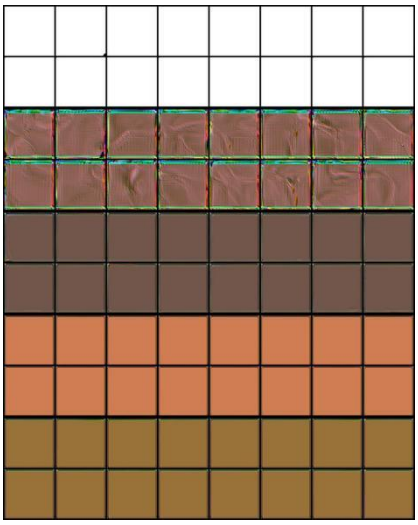

图 8：learning_rate = 0.0005(图像从上往下，每两列分别表示在训练 10k、20k、30k、40k、50k steps 之后生成的结果)

同样，很明显，对比 BEGAN 复现时 learning rate 为 0.0001 的效果，明显初始学习率设置为 0.0005 时也无法训练出可使用的模型。

学习率设置为 0.00008 时，参数和 BEGAN 的复现中初始学习率相同，训练和测试结果在 3.1 中呈现的比较详细，在此不再赘述。

学习率设置为 0.0001 时，训练过程生成器在训练到 10k、20k、30k、40k、50k steps 之后生成的结果见引文后的附录 C。

发现生成的图像质量和最初学习率设置为 0.00008 的生成图像质量无法用肉眼看出差别，因此进一步呈现损失函数和平衡指标的变化，如图 9 和图 10。

比较学习率为 0.0001 和 0.00008 两模型的损失函数和平衡指标，发现整体上依然十分接近。至此大概可以判断，使用 adam 优化器时，学习率设置为 0.0001~0.00008 之间，保持其他结构和参数不变时，基本不会影响最终模型的质量。

学习率设置为 0.00001 时，训练和测试结果如图 11。发现相比与初始学习率设置为 0.0001、0.00008 的模型而言，最后生成的图像五官更加模糊，明显可以看出人脸的轮廓不够清晰和明朗，甚至有些变形。

因此，基于上述的五组实验比较，发现将判别器和生成器的初始学习率设置为 0.0001 左右时，在其他指标和参数不发生改变的情况下，训练之后可以得到较为理想的模型。

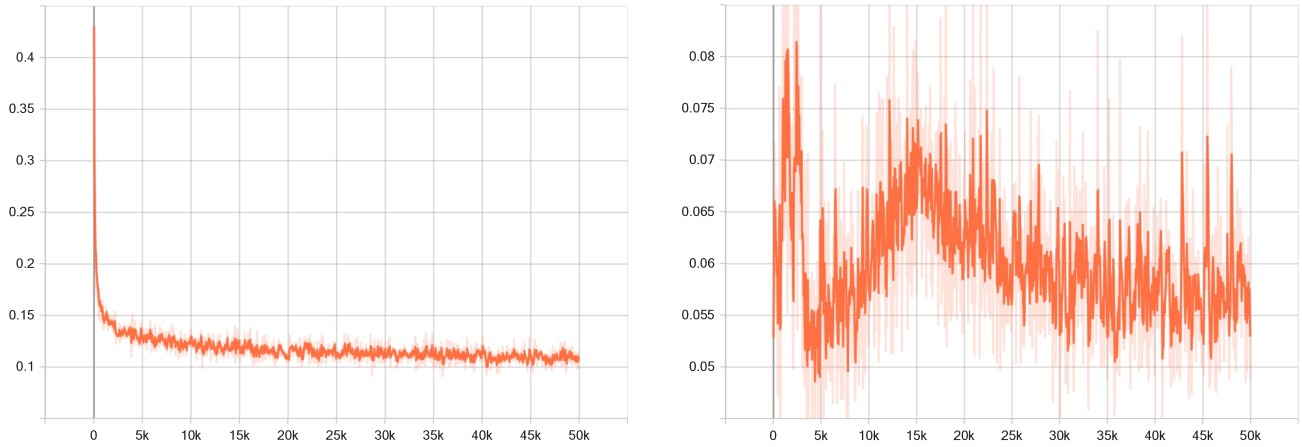

图 9：learning rate=0.0001 时损失函数的变化(左侧为判别器的损失函数收敛曲线，右侧为生成器损失函数的收敛曲线)

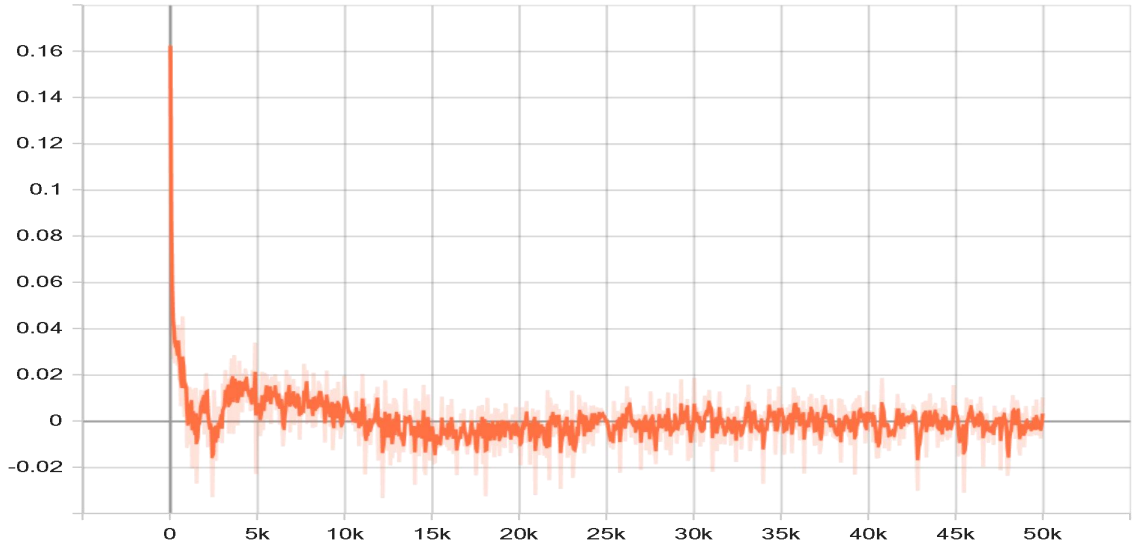

图 10：learning rate = 0.0001 时 balance 指标的变化曲线

3.3 BEGAN 和 GAN 的比较结果

根据论文，分别将原始 GAN 中的损失函数调整为 minimax 损失函数和 nonsaturating 损失函数，分别记为 MM GAN 和 NS GAN，与 BEGAN 模型进行对比，比较的指标包括：FID 得分、F1、Precision、Recall。F1 的含义是精准度（precision）和召回率（recall）的平均数。

使用的数据集为研究者自制的 mnist 数据集，数据集中都是灰度图，图像的目标是不同形状的三角形。数据集示例如图 12。

从实验结果来看，不同 GAN 模型图像的生成质量都类似，解释如下。

首先是 FID 得分的比较。表 2 呈现了 MMGAN、NSGAN、BEGAN 在数据集 MNIST 上的最佳 FID 区间。最佳 FID 的计算分为两个阶段：首先，对各种超参数进行大规模搜索，并选择最佳模型。然后，我们使用不同的初始化种子重新运行所选模型的训练 50 次，以估计训练的稳定性并报告平均 FID 和标准偏差，排除异常值。

Data Set: MNIST	
Model Type	Best FID
MM GAN	9.8±0.9
NS GAN	6.8±0.5
BEGAN	13.1±1.0

表 2：MMGAN、NSGAN、BEGAN 在 MNIST 上的最佳 FID 区间

从表中可见 BEGAN 模型比 MM GAN 和 NS GAN 模型更优。但是论文也发现，如果使用不同的数据集进行测试，会得到不同的得分情况，完整情况见表 2。该表中进一步使用了 GAN 模型训练中常用的数据集。除了团队自制的 MNIST 数据集之外，还有之前单独训练 BEGAN 模型的 CELEBA 数据集、FASHION 数据集、CIFAR 数据集。

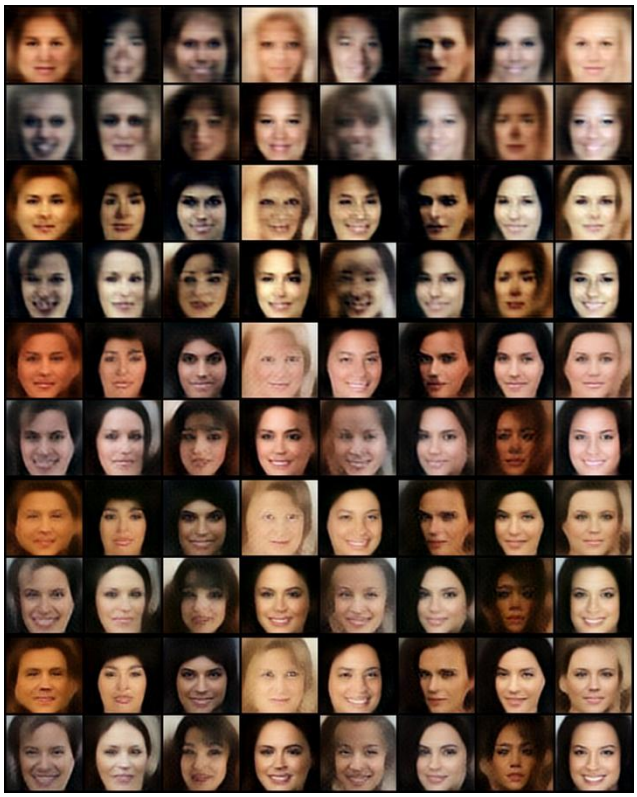

图 11：learning_rate = 0.00001(图像从上往下，每两列分别表示在训练 10k、20k、30k、40k、50k steps 之后生成的结果)

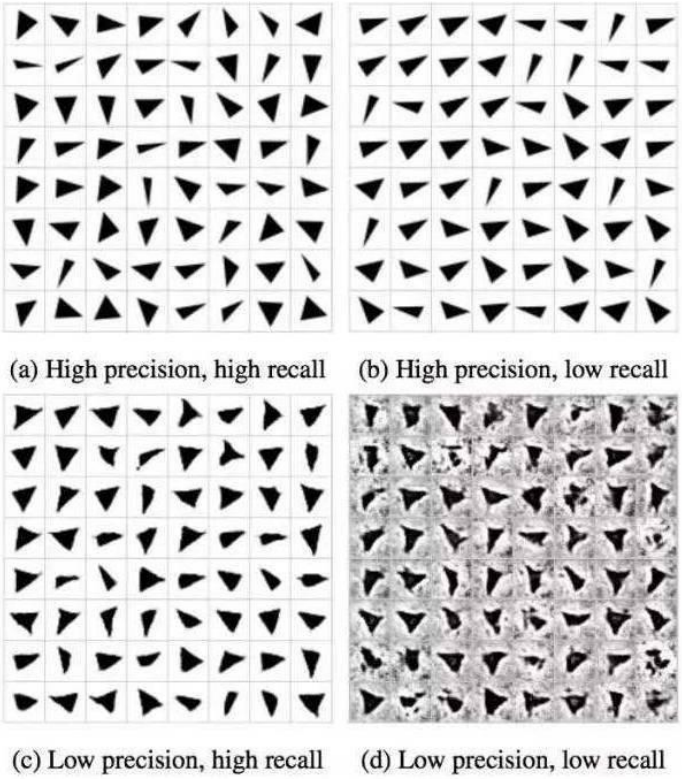

图 12：mnist 数据集示例

由下表 3 可见，并没有一个模型可以在所有的数据集上均有最突出的表现。

Model Type	MMGAN	NSGAN	BEGAN
MNIST	9.8±0.9	6.8±0.5	13.1±1.0
FASHION	29.6±1.6	26.5±1.6	22.9±0.9
CIFAR	72.7±3.6	58.5±1.9	71.4±1.6
CELEBA	65.6±4.2	55.0±3.3	38.9±0.9

表 3：不同数据集三个模型的最佳 FID 分布

在呈现 FID 的结果分布之后，接下来呈现 precision、recall 和 F1 的比较。如图 13 所示，该图显示了在使用 MINIST 数据集时，在 95%的置信区间的固定 budget 下最佳 F1 的分布、precision 分布和 recall 召回率的分布。发现在优化 F1 得分时，MMGAN 和 NSGAN 都有较高的 precision 和 recall。然而 BEGAN 在图中的表现十分平庸，原因需要进一步的深究。

结合原论文以及本报告中呈现的部分结果，按照此比较模式，并不能完全体现 BEGAN 模型比原始 GAN 模型有所提升之处。

但是这种比较方式也存在局限性。根据论文的比较思路，核心其实是“当各个模型都能稳定巡礼拿到收敛的情况下，谁的效果更好”。但是论文本身忽略了，有很多的 GAN 模型本身被设计出来，优势就是“在训练过程中能更容易地训练到收敛”，比如 WGAN 模型。同时虽然得到的结论是没有明显的证据说明 GAN 的衍生模型比原始 GAN 更优，但是从实际运用来说，让原始 GAN 模型生成 2K 清晰度的高清图是十分困难的，但是对于 BEGAN 来说这是一个可实现的结果，是一个十分简单而合适的过程。

4. 结论

4.1 课题内容小结

本技术报告在充分查阅文献和自主动手实践的基础上，对 BEGAN 模型进行了复现和训练。在训练的过程中，发现学习率的设置和优化过程十分重要。如果学习率的下降过快或者学习率太小，会导致模型的训练结果。

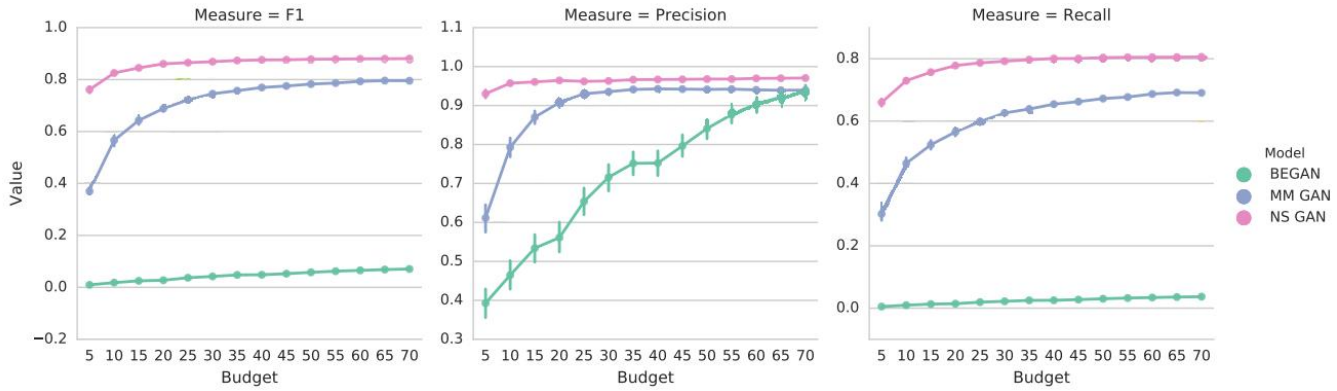

图 13：BEGAN、MMGAN、NSGAN 固定 budget 下 precision、recall 和 F1 的比较

质量下降甚至出现模式崩溃的情况。因此应当高度重视对学习率的调整和优化策略。

于是本报告进一步对学习率的调优设计实验，在进行 5 组实验的对比之后，不仅真实的感受到不同学习率对模型训练结果的巨大影响，也发现当初始学习率设置为 0.0001 左右时效果会比较理想。

进一步，为了验证 BEGAN 模型相比于 GAN 模型的理论分析的优势是否可以实验复现，本报告参照谷歌团队的论文与开源的实验平台，进一步从 FID、F1、precision、recall 等角度对比了原始 GAN 模型和 BEGAN，结果发现 BEGAN 在实验条件下并没有展现出明显比原始 GAN 模型更加突出的地方。而虽然此模型能从一定侧面反映问题，例如有部分模型的优势是在观察抽样的误差下得到的，后续在改进模型时应该更多对结果的稳健性进行更认真的研究。

但该论文的比较方式仍然不够全面，因为论文本身忽略了，有很多的 GAN 模型本身被设计出来，优势就是“在训练过程中能更容易地训练到收敛”，比如 WGAN 模型。同时虽然得到的结论是没有明显的证据说明 GAN 的衍生模型比原始 GAN 更优，但是从实际运用来说，让原始 GAN 模型生成 2K 清晰度的高清图是十分困难的，而对于 BEGAN 而言则更为简单和容易。

4.2 课题不足与分析

（1） 在 learning rate 的调参过程中，由于训练时间和计算资源的限制，只做了 5 组训练的对比，并且五组数据的选择没有固定的依据，参数设计比较随意。

（2） 由于知识的欠缺，虽然了解了论文（are GANs all equal）中比较不同那个模型的评价指标，但是难以使用代码复现，并体会论文中的所有技术原理。因此只是参照论文和给出的项目进行了程序的运行和体会，并没有从头到尾进行亲手实践。

（3）计算资源的缺乏对原始 GAN 和 BEGAN 模型的比较也有影响。

（4）在比较原始 GAN 模型和 BEGAN 模型时，其实只使用 MNIST 模型作为数据集是不够的。例如，

（5）比较 BEGAN 和原始 GAN 模型的方法是存在缺陷的。因为研究者为了保证所有的模型训练之后都可以成功收敛而固定了要生成的图片规格，但是很多模型的优势恰恰是“可以在训练过程中更快的收敛”，而论文的比较策略无法体现这一优势。同时例如 BEGAN 模型和 PG-GAN 模型可以较容易的生成 2K 高清图而原始 GAN 模型则十分困难，这类优势在此实验中业无法很好地体现。

4.3 未来任务规划

需要进一步加强知识学习和文献阅读。尝试从其他论文中找寻灵感，在已有的基础上进一步探究不同 GAN 模型应该从哪些指标进行进一步评价，而更客观全面的体现不同模型的不同特征。因为本实验基于谷歌团队所使用的的评价模型仅仅探讨了多种比较维度中的一个子集，显然在实际的分析中是不能排除某些模型在目前尚未开发的情况下显着优于其他模型的可能性的。

改变的策略不仅可以是比较模型的改变，比较中所使用的的数据集也可以有所改变，例如，在论文中所使用的 MNIST 数据集中：

（1）一次引入多个凸多边形；

（2）在多边形内提供颜色或纹理；

（3）逐渐增加分辨率等[13]。

这些手段都可以使当下的比较模型更加全面客观。当然，计算资源的分配也是很重要的部分。

5. 参考文献

[1] Sonka M，Hlavac V，Boyle R． Image Processing，Analysis and Machine Vision［M］． Boston，MA: Springer，1993． ［DOI: 10．1007 /978-1-4899-3216-7］

[2] Li C，Wand M． Precomputed real-time texture synthesis with Markovian generative adversarial networks［C］/ /Proceedings of the 14th European Conference on Computer Vision． Amsterdam，The Netherlands: Springer，2016: 702-716．［DOI: 10．1007 /978-3-319-46487-9_43］
[3] Cappelli R，Erol A，Maio D，et al． Synthetic fingerprint-image generation［C］/ /Proceedings of the 15th International Conferenceon Pattern R ecognition． Barcelona，Spain: IEEE，2000: 471-474． ［DOI: 10． 1109 /icpr． 2000． 903586］
[4] 王万良,李卓蓉.生成式对抗网络研究进展[J].通信学报,2018,39(02):135-148.
[5] Mirza M，Osindero S． Conditional generative adversarial nets[J]．arXiv preprint arXiv: 1411．1784，2014．
[6] Radford A，Metz L，Chintala S． Unsupervised representation learning with deep convolutional generative adversarial networks［EB /OL］．2016-12-20［2018-02-28］．https: / /arxiv．org /pdf /1511．06434. pdf.
[7] Arjovsky M，Chintala S，Bottou L． Wasserstein GAN[EB /OL] .2017-12-06[2018-02-23]. https: / /arxiv. org / pdf /1701. 07875. pdf.
[8] Berthelot D, Schumm T, Metz L. Began: Boundary equilibrium generative adversarial networks[J]. arXiv preprint arXiv:1703.10717, 2017
[9] Lucic M, Kurach K, Michalski M, et al. Are gans created equal? a large-scale study[J]. arXiv preprint arXiv:1711.10337, 2017
[10] 王坤峰,苟超,段艳杰,林懿伦,郑心湖,王飞跃.生成式对抗网络 GAN 的研究进展与展望[J].自动化学报,2017,43(03):321-332.DOI:10.16383/j.aas.2017.y000003.
[11] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
[12] Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in Neural Information Processing Systems (NIPS), 2016.
[13] Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. The GAN Landscape: Losses, architectures, regularization, and normalization. arXiv preprint arXiv:1807.04720, 2018

附录：

A. BEGAN 复现中，不同的训练阶段所生成的图像呈现：

图 14: BEGAN 复现的训练过程中生成的图像(本图中按照从左到右从上到下的顺序，每 2*8 个人脸图像分别代表第 5k、10k、15k、20k、25k、30k、35k、40k、45k、50k steps 之后生成器生成的人脸图像)

从上图可以看出，模型的人脸学习和生成经历了明显的从模糊逐渐清晰的过程。但是具体细节来看，由于训练迭代次数较少，计算资源和时间有限等问题，生成的人脸的轮廓并没有十分明显。但显然相比于初始学习率设置为 0.00001 时轮廓更加清晰。并且整体来看，生成的人脸样式略有丰富性和多样性，但是与许多理想状态相比，不够丰富和多元。除了调整训练迭代次数和训练时长，是否有其他策略可以使 BEGAN 模型生成更优质清晰的图像是本人未来可以进一步学习和探索的方向。

B. BEGAN 复现中，测试结果生成图像呈现：

图 15：BEGAN 复现模型依据表 1 训练完成后训练该模型生成的图像

C. 将 BEGAN 模型的判别器和生成器的学习率分别调整为 0.0001 时，训练过程在训练到 10k、20k、30k、40k、50k steps 时生成器生成的图像：

图 16：learning_rate = 0.0001(图像从上往下，每两列分别表示在训练 10k、20k、30k、40k、50k steps 之后生成的结果)