论文阅读报告：Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks [1]

纳文琪*

* 信息学院，学号：22018000164

1 引言

图像翻译是一种转换图像的表示的过程。大多数传统方法都是针对特定领域的图像翻译问题设计专门的方法来实现图像翻译。生成对抗网络兴起后，大量研究试图构建对任何训练数据都通用的框架来实现图像翻译。然而，这些方法要么生成的图片质量不高，要么需要获取过程困难且昂贵的配对训练数据。本文所阐述的论文提出了一种不需要昂贵的配对数据，同时生成图片的质量也比较高的方法，其被称为 CycleGAN。

2 相关工作

图像翻译 图像翻译（image-to-image translation）指的是，在给定充分训练数据的情况下，将图片的一种表示转换成另外一种。虽然这类任务的目的都差不多，仅是训练数据不同，但传统的方法都必须根据特定目标设计专门的方法来完成。

生成对抗网络 生成对抗网络 (Generative Adversarial Networks, GANs) 是最近今年的研究热点之一，它在图像生成、图像编辑、表示学习等方面取得

了瞩目的成绩。GAN 由生成器和判别器两部分组成。生成器用于生成期望的数据（假数据），并尽量让生成的数据骗过判别器。而判别器则负责尽可能准确地识别出哪些是假数据（由生成器生成），哪些是真数据。

GAN 生产的数据可能满足特定的分布，但却是无法控制的，而有条件的生产对抗网络（Conditional Generative Adversarial Networks, cGAN）的出现则解决了 GAN 无法控制数据的生成这个问题。cGAN 通过在生成器和判别器的输入端同时加入标签数据来控制数据的生成和判别。

pix2pix　pix2pix 是有 Isola 等 [2]提出的一个图像翻译系统。它针对图像翻译需要根据不同的训练数据设计专门的方法的问题，提出了一套通用的框架。它将配对数据作为训练集，通过以 cGAN 为基础设计的模型来实现图像翻译。这个框架适用于任何分布的数据集，只需要能够提供配对训练数据就行。

3　无配对数据情况下的图像翻译

3.1　概述

在计算机视觉领域，已经存在大量的方法可以实现图像翻译。在此论文之前，pix2pix 算是其中的佼佼者，但它需要使用配对数据对模型进行训练。然而，获取配对数据常常是及其困难的，且获取的代价也相对昂贵。因此，本文所讲述的论文试图找到一种无须配对训练数据的图像翻译方法。论文构建了这样一个系统：在仅有非配对训练数据的情况下，学习一组图片的特征，并能够将这些特征应用到另外一组图片上，使后者成为与前者同一类型的图片。例如，通过学习一组斑马的图片，将一张图片上的马转换为斑马。

为了达到上述目的，一般需要构建一个映射函数（生成器）$G : X \to Y$，其输出 $\hat{y} = G(x), x \in X$，且 $\hat{y} \approx y, y \in Y$。也就是说，G 的输出 $\hat{y}$ 要能骗过判别器。然而，在构建生成器的时候，常常遇到的模型坍塌的问题。当存

在一个 $\hat{y} \approx y$ 时，生成器最安全的做法就是将任意的输入 x 都映射到这个 $\hat{y}$。

文章提出用“循环连续性”来处理模型坍塌的问题。在构建生成器 $G : X \to Y$ 的同时，构建另外一个生成器 $F : Y \to X$。如果有 $\hat{y} = G(x) \approx y$ 时，应当同时有 $\hat{x} = F(\hat{y}) \approx x$，即 $F(G(x)) \approx \hat{x}$。通过最小化“循环一致性损失”优化生成器 G 和 F。

3.2 方法

模型 文章通过给定两个集合：$\{x_i\}_{i=1}^{N}$ 和 $\{y_j\}_{i=j}^{M}$，来构建两个生成器：$\hat{y} = G(x), x \in X$ 和 $\hat{y} = G(x), x \in X$。与此相对应，还需要构建两个判别器：D_X 和 D_Y，分别用于对 x 与 $\hat{x}$ 以及 y 与 $\hat{y}$ 做判别。

损失函数 本文模型的损失函数由两部分组成：对抗损失和循环一致性损失。

对抗损失函数有两个，分别对应生成器 G 和判别器 D_Y，以及生成器 F 和判别器 D_X。具体表示为：

$$\mathfrak{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_Y(G(x)))] \tag{1}$$

$$\mathfrak{L}_{GAN}(F, D_X, Y, X) = \mathbb{E}_{x \sim p_{data}(x)}[\log D_X(x)] + \mathbb{E}_{y \sim p_{data}(y)}[\log(1 - D_X(F(y)))] \tag{2}$$

文章将循环一致性分为两个部分：输入为 x 的过程 $x \to G(x) \to F(G(x)) \to \hat{x}$，这称为前向循环一致性，以及输入为 y 的过程 $y \to F(y) \to G(F(y)) \to \hat{y}$，这称为后向循环一致性。这两种一致性组合就得到循环一致性损失：

$$\mathfrak{L}_{cyc}(G, F) = E_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1] + E_{y \sim p_{data}(y)}[\|G(F(y)) - y\|_1] \tag{3}$$

模型总的损失函数就是：

$$\mathfrak{L}(G, F, D_X, D_Y) = \mathfrak{L}_{GAN}(G, D_Y, X, Y) + \mathfrak{L}_{GAN}(F, D_X, Y, X) + \lambda \mathfrak{L}_{cyc}(G, F) \tag{4}$$

其中 λ 是一个超参数，表示循环一致性损失的影响系数。

3.3 实现

3.3.1 网络架构

生成器网络 文章采用“卷积-残差-微步卷积”[3]的结构来构造生成器。与一般的 GAN 不同的是，此处的生成器输入的是 X 集合中的元素，即一张图片，而不是随机噪声。生成器网络首先对图片进行两次步长为 2 的卷积操作，其输出的结果再经过一组残差块，最后的结果再进行两次步长为 $\frac{1}{2}$ 的微步卷积，最终得到输入的图片。

判别器网络 对于判别器网络，文章使用 70×70 的 PatchGAN 来实现，将输入分割成很小的部分分别进行判别。这样的网络产生的参数更少，而且适用于任意尺寸的输入图片。

3.3.2 训练

训练的不稳定性是 GAN 总所周知的一个问题。文章采用两种技术来稳定模型的训练。首先，文章采用 LSGAN 来替代常规的 GAN 的损失函数，这在稳定训练过程的同时还能够提供训练的质量。其次，为避免模型抖动，训练判别器时，不单单使用刚刚生成的图片，而是连同之前生成的图片一起进行训练。

3.4 实验结果

文章将 CycleGAN 与 CoGAN、Pixel+GAN、Feature loss+GAN、BiGAN、pix2pix 进行了比较。文章首先进行了几个系统之间的“真假”检验实验，通过将系统生成的图片以众包（Amazon Mechanical Turk, AMT）的方式分发给真人，并让其判断是否是真实图片的方式，来评估系统的能力。经过对

比，在“Map → Photo”实验中，CycleGAN 生成的图片有超过 26% 的被标记为真，远远超过第二名的 2%。文章之后还做了 Cityscapes 数据集上的“Photo → Label”实验和“Label → Photo”实验，都取得了最好的成绩。

3.5 应用

CycleGAN 现在已经被用在了多个应用领域中，包括：

- 物体转换 在两类相似的物品之间进行转换，例如，将苹果转换成桔子。
- 四季转换 将冬天和夏天的风景图片进行转换。
- 画册风格转换 将一组图片的风格转换为另外一组图片的风格。转换模仿的是一组图片（例如梵高）的风格，而不是一张图片（例如《星空》）的风格。
- 由油画生成照片
- 照片增强 将景深较深的照片转换成景深较浅的照片。

3.6 局限和结论

虽然本文的方法取得了一些成就，但离完美还很远。这个方法针对颜色和纹理的转换效果不错，但如果存在几何形状上的转换，效果就不太好。还有些失败则是由训练图片的特征分布所决定的。例如，人骑着马的图片转换成斑马时就会失败，那是因为训练集中所有的马都没有人骑着。

本文的方法已经在很多方面展开了应用，本文也扩展了此类方法在“无监督”学习方面的边界。

参考文献

[1] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks”, arXiv preprint, 2017. 1

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks", arXiv preprint, 2017. 2

[3] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution", in European Conference on Computer Vision. Springer, 2016, pp. 694–711. 4