计算机视觉课程报告———StarGAN 解析

StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation

杨海宏 11721014

April 24, 2019

Contents

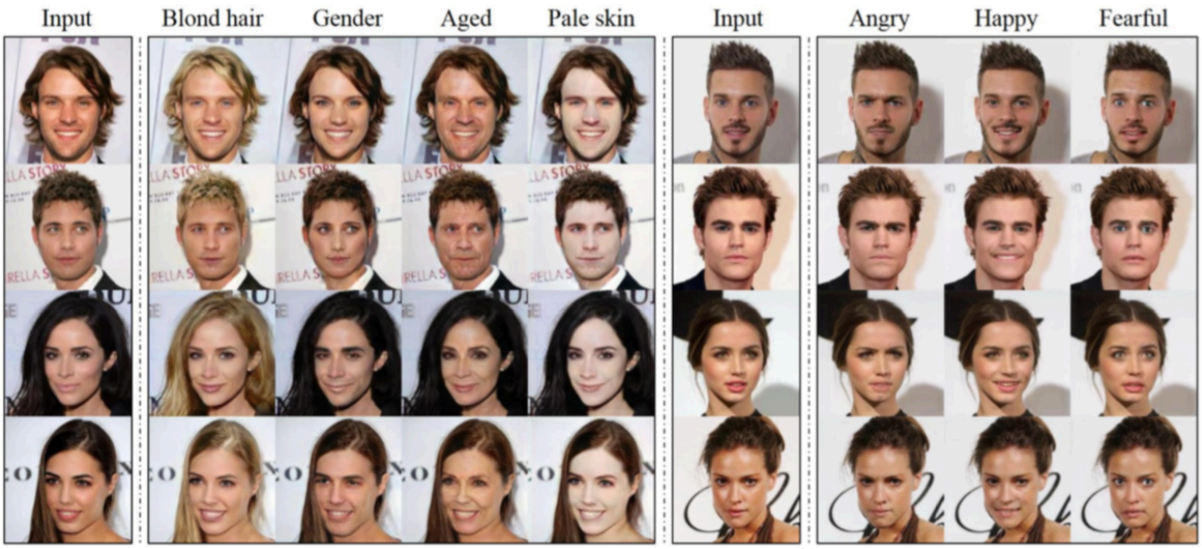

Figure 1: StarGAN 的结果示意图。除输入外，每一列代表一个特殊的类目（域）。

1 论文动机

StarGAN 所要解决的问题是：无监督跨多域的图像（到图像）翻译任务。这里的“跨多域”并非指图像域到文本域（如图像标题生成，Image Captioning），而是指数据集中标记的类目（Class）。如图 1。

在图 1中，既有关于发色，性别，年龄，肤色等较为粗粒度的类目，也有关于面部表情（生气，开心和恐惧）的细粒度类目。而 StarGAN 能实现的效果，就是在给定原始输入图像和目标域标签时，生成目标域图像，通过生成对抗网络实现对图像的属性和属性值（如属性为“性别”，属性值为“男”或“女”）的改变。

在无监督图像翻译任务中，已有较为著名的工作是 CycleGAN 和 DiscoGAN。它们均利用了“重构（Reconstruction）”的思想，来实现无监督地学习两个域的支撑集之间的对应关系。CycleGAN 因充满想象力地提出“循环一致性损失（Cycle Consistency Loss）”而在两年内被引用超过 1600 次[1]。类似的思想在无监督机器翻译中被建模为“逆向翻译（Back Translation）”。

但是，它们的不足之处在于效率低下（inefficient and ineffective）：

- 前人模型仅仅支持两个域之前的图像翻译。如果要进行跨多（k 个）域图像翻译任务，直接套用将会需要训练 $k(k-1)$ 个生成器，如图 2所示；
- 多个生成器之间无法复用图像中领域无关的基础特征，如脸型。

因此，这一篇论文提出了 StarGAN，一个专为跨多域图像翻译设计的模型，期望用一个统一的模型来解决跨多域关系的捕获问题，同时充分利用多个域或多个数据集的图像特征，来提高图像生成的质量。

2 数据与模型

本文所使用的数据集是 CelebA 和 RaFD。CelebA 包含 202,599 张名人面部图像。每张图像都标注了 40 个 0-1 标签，表示是否对应某种属性值。RaFD 包含 67 位志愿者的 3 个不同角度的 8 种面部表情，共 4,824 张图像。

[1] 至写作日 04/22/2019

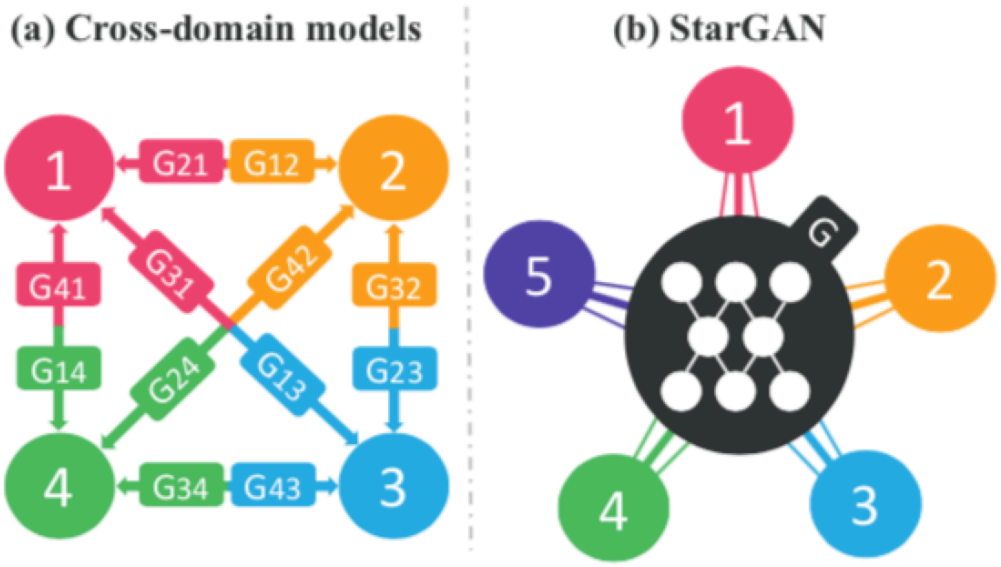

Figure 2: 针对跨多域图像翻译任务，跨双域模型与 StarGAN 模型对比示意图。直接应用跨双域模型 (a)，如 CycleGAN，则必须在多个域中针对每一对域训练一组生成器，表明提出一个统一的跨多域图像翻译模型 (b) 的必要性。

模型的主要贡献有两个部分：

- 对重建损失（或称循环一致性损失）进行略微的修改——在输入图像时附带域的类别标签，即 $\mathcal{L}_{rec} = E_{x,c,c'}[||x - G(G(x,c),c')||_1]$，注意到，StarGAN 无论是在源域到目标域生成，还是目标域到源域重建时，都使用同一个生成器 G；
- 引入遮挡向量（mask vector），使 StarGAN 模型可在多个数据集上同时训练。在测试阶段，StarGAN 根据多个数据集上的类标签来控制生成图像的效果。

遮挡向量 $\hat{c}$ 的设计是非常符合直觉的。

$$\hat{c} = [c_1, c_2, \ldots, c_n, m] \tag{1}$$

其中，m 是一个 n 维的独热向量，表示当前样本所属的数据集，c_i 表示第 i 个数据集中的类标签独热向量。假设模型同时在两个数据集上训练，且要表示某一图像属于第一个数据集中的第二个域标签，则 $\hat{c}$ 为

$$\hat{c} = [0, 1, 0, \ldots, 0, 1, 0] \tag{2}$$

引入遮挡向量是为了给模型提供标签信息，这是一个值得学习的技巧。这个标签信息不仅在图像生成时用于引导生成图像所属的域，而且在判别器中，也需要利用标签信息来评价生成的质量。

3 实验结论与不足之处

文中主要使用多个基线模型和新模型对比进行两种类型的实验——单属性的转换和多属性的转换。文中所有实验的定量评价均使用第三方人工（亚马逊众包平台）标注并统计结果。结论是 StarGAN 的生成效果都比基线模型好，尤其是在多属性转换的实验中，StarGAN 显著地超越了所有的基线模型。

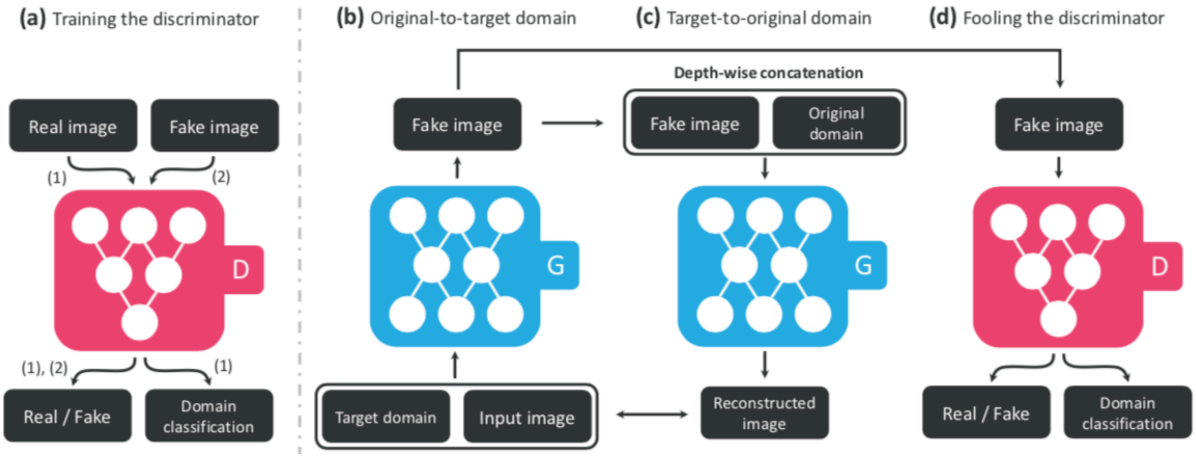

Figure 3: StarGAN 的概览图，包括一个判别器 D 和一个生成器 G。(a)D 学习如何区分真假图像以及将真实图像划分到对应的域（类别）中。(b)G 输入一张图像和一个对应的目标域标签，生成一张图像。(c)G 输入一张生成图像和一个对应的源域标签，尝试重建原始图像。(d) G 尝试生成以假乱真的图像，且要求判别器 D 输出的结果为目标域标签。

Table 1: 亚马逊众包平台评估多属性转换实验。H 代表头发颜色，G 代表性别，A 代表年龄。

Method	H+G	H+A	G+A	H+G+A
DIAT	20.4%	15.6%	18.7%	15.6%
CycleGAN	14.0%	12.0%	11.2%	11.9%
IcGAN	18.2%	10.9%	20.3%	20.3%
StarGAN	47.4%	61.5%	49.8%	52.2%

作者认为：这是由于 StarGAN 可以在多个数据集上同时训练，能学习到通用的基础图像特征，通过多任务学习的设定，不仅增大了训练集，还能防止模型过拟合。

而实际上，对于基线模型，要实现多属性转换的效果，必须是经过多次单属性的转换。在连续几次的单属性转换生成图像的过程中，基线模型会逐次引入误差并将这些误差传递，从而直接影响最终图像的生成质量。

4 总结与展望

个人认为，图像生成，风格迁移等是近两年非常成功的一个应用方向。对于自然语言处理，这些方法和思想都是值得借鉴的。最直接的例子之一是对抗样本。在这一系列的论文中，循环一致性损失，（或更一般地，重建损失）是无监督学习中一个重要的思想。这在自然语言处理中也有所体现。

在机器翻译中，小语种的平行语料往往是很难获取的，即使是“英语-> 小语种”的翻译任务。这时就出现了，跨域关系（翻译句子对）很难获取，但单域数据（英语、小语种）容易获取的情况。无监督机器翻译面临的挑战之一，就是自动地学习跨域关系。在机器学习中，本文所讨论的任务和思想，实际上从属于无监督领域适应问题 (Unsupervised Domain Adaption)。

在未来，我个人将从这一方面入手，关注如何无监督地学习到知识图谱中节点和边与给定文本之间的对应关系。最后，感谢王东辉教授为我们上课，介绍计算机视觉的前沿知识以及基本的科学思想。