Repurposing GANs for One-shot Semantic Part Segmentation

摘要

尽管GAN在真实图像生成方面取得了成功，但是将GAN用于其他与生成无关的任务的这一思想并未得到充分的使用。

那么GAN能不能在他们尝试复制目标的时候能够学习到目标模型中有意义的结构呢？

这篇文章就是围绕这个思想展开，测试了上面这个假说并提出了一种基于GAN网络简单并且有效的语义分割方法，该方法只需要一个标签示例和一个未标记的数据集

核心思想是利用一个训练好的GAN模型从输入图像中提取的最终表示并用这个结果作为语义分割网络的特征向量

而最终的实验表明，GAN的表示是容易辨别的，并产生了非常好的结果，可以与训练了更多标签的监督学习基线的结果相媲美

作者也相信这个新的GAN的用法会引起一种新的无监督的表示学习而且这种学习能够适用很多其他的任务

引言

如今，在计算机视觉方面，一个机器如何在只观看很少甚至一个样例的情况下去识别一个物体或者其部分。

一个孩子能够做到，但是需要几年持续的先验的视觉信息积累，并且他能够快速识别一个人的耳朵大概是利用率之前看人脸的经验。

这篇文章就建立在上面的场景之下，并且解决了这个问题。在给定大量的人脸照片集或者其他对象类的情况下，我们目标是识别未见过的带有部分注释的面部图像中每个语义部分对应的像素。

这个问题的的设定不同于经典的少镜头学习的设定，后者描述的是用多对象类训练的学习算法需要少量的新类的监督样本就能对新类进行分类或操作。与之相反，这篇文章的设定是只包括少量注释样本而且没有来自其他类的训练数据的单对象类。

针对少镜头学习，提出了许多方法，一般的思想是将外部学习的先验知识应用到少镜头任务中。然而这些通常需要带昂贵的标签或者部分注释的监督任务中学习，即成本较高。作者引入了一个新的方向，即使用一个生成模型，尤其是对抗生成网络（GAN）。GAN网络在建立数据分布和生成真实图像时取得了极大的成功。假设GAN需要学习对象中有意义的结构信息，来正确地合成结构他们，而且合成对象的不同部分所需的生成计算可以为其他任务提供有用的判别信息。

作者的主要贡献就是利用训练过的GAN从图像中提取有意义的像素表示，这些表示可以直接用于语义部分分割。又有实验表明，GAN对于学习这种表示是非常有效的，并且只有一个示例标签就可以实现非常好的效果。

尽管有显著的结果，但是这个核心思想在很大程度上依赖于耗时的潜在优化并且需要测试图像接近GAN学到的图像分布。文中又展示了一个简答的扩展，即自动镜头分割，可以绕过潜在的优化，会有更快、更有效的预测。更重要的是通过自动拍摄训练中执行的几何数据增强，我们可以同时分割具有不同大小和方向的多个对象，然而在训练过程中得不到真实的场景。

总而言之，作者的主要贡献是提出GAN在无监督像素表示学习的新用法，这个在少镜头的语义分割上取得了很好的性能。实验表明，这种表示很容易被区分出来，而且演示了如何将这种思想扩展到显示场景，来解决GAN的训练数据和真实图像之间的一些领域的差距。

相关工作

表征学习

表征学习的目标是从原始数据中捕获有用且更方便处理下游任务的底层信息。目前又很多方法能够从任务中学习到这些表征，并使用它们来帮助提高另一项任务的性能。本文则使用了一种从生成任务中学习表征的方法

生成模型

深度生成模型在建模图像分布方面展现了良好的结果，从而使合成真实图像成为可能。几种典型的视觉生成模型，如自回归模型，编码解码器架构的自动编码器（VAE以及其衍生）和生成对抗网络GAN。本文则使用GAN来进行表征学习，利用GAN内部表征与生成输出紧密耦合，并且他们可以保存有用的语义信息。

语义部分的分割

与语义分割不同，本文问题的目标是分割对象中的部分，而不是场景中的对象。因为在现实中，两个部分没有可见的边界，如鼻子和脸。在语义部分已经又 相当大的进展，但是这些技术需要大量的像素级的注解。

为了避免使用像素级的注解，一些方法依赖于获得成本更便宜的其他类型的注解，例如关键点、主体姿势或边缘贴图。然而这种方法只适用于特定领域工作，如人体部位。其他方法则尝试完全用自我监督的技术放弃了注释。然而这种方法的一个主要缺点是对对象部分的分割几乎没有控制，从而导致任意分割。作者的方法只需要少量带注释的示例就可以完全控制对象部分的分区。

少镜头语义分割

以前的研究就一直试图用很少的注释来解决分割问题。元学习方法首先在一个带注释的数据集上训练一个分割网络，然后在目标类的一个注释上微调网络参数；原型方法使用一个支持集来学习每个对象类的原型向量。两个方法都构造了两个训练分支，都支持分支对非目标类的注释或图像级注释进行训练，然后查询分支采用输入图像和提取的特征来预测分割掩码。相似度引导网络使用分割掩码来屏蔽支持图像中的背景，然后仅使用前景对象的特征来引导查询分支定位与支持分支的特征高度相似的像素。

有些工作在所有视频帧的片段对象只有第一帧有注释，然而这些方法在语义分割中没有表现出成功。元学习需要类似对象类的注释掩码，因此学习特定部分的原型是不可行的。由于缺乏部分级别的注释，也很难利用从支持集中获得的信息。因此，从GAN中提取的表征包含部分级信息，可以在没有监督的情况下学习。

方法

文中的问题涉及到语义部分分割有以下新的设定，给定单个对象类的图像数据集，我们的目标是通过从少数带有部分注释的图像中学习，从同一类中分割一个未见过的对象图片。可以由用户使用二进制掩码指定这些部分注释。而且，语义部分分割也可以看作是一个n路的像素级分类问题，其中n是各部分总数。

如果存在一个函数f能够映射每一个像素值到它包含对部分分类的鉴别信息的特征向量，而且它本身没有语义信息。作者建议从训练GAN来合成目标类的图像的模型中提取出一个函数。

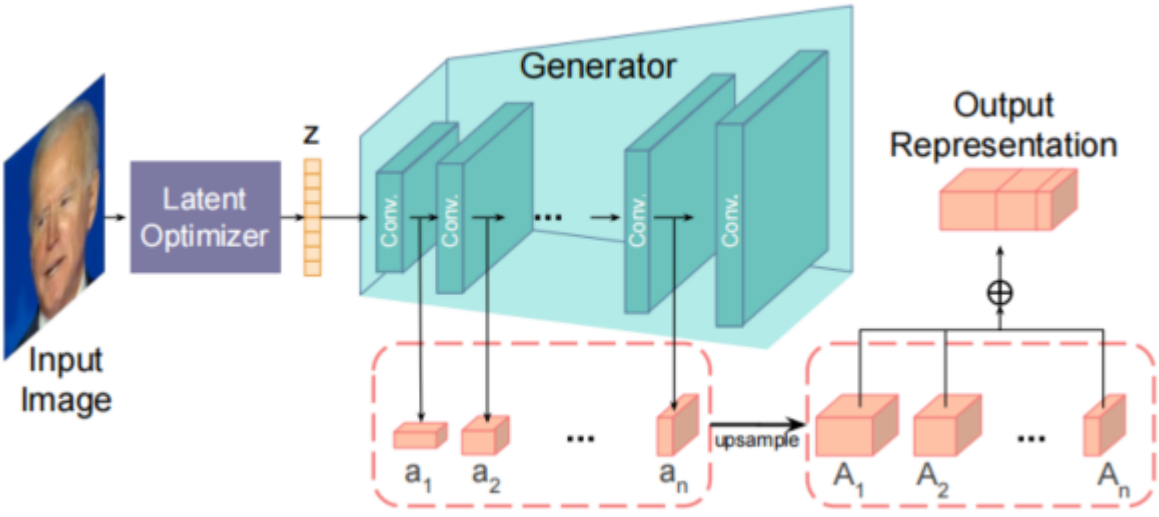
下面是详细的过程：

从GAN中进行表征提取

使用GANs座位映射函数并不简单，因为GANs的输入是有一个随机的隐藏特征，而不是映射的图像像素输入。

先通过一个随机的隐藏特征输入投喂给GAN模型生成一个图像。合成的输出图像由生成器通过一系列的空间卷积生成，每个输出像素都是唯一的生成的计算结果，可以通过每一个卷积层追溯到初始的特征向量。核心思想就是使用这些独一无二的计算路径找到的特征表征。总的来说，用与生成像素的计算路径形成了一个有向的无环图，其节点表示网络参数或输入该像素计算所涉及的隐藏特征。然而在工作中，这些节点表示激活值，我们简单地用生成器内所有层与该像素空间一致的单一激活序列来表示路径。

如下图所示：



我们从生成器的每一层提取激活切片 $a_1, a_2, ..., a_n$，每一项都有维数(h_i, w_i, c_i)，然后计算它们的像素表征为：

$$F = U(a_1) \oplus U(a_2) \oplus ... \oplus U(a_n)$$

其中$U(\cdot)$是对输入进行空间上采样直到最大的激活层(h_n, w_n) $\oplus$是通道维度的连接，这个过程将每个三维的RGB像素映射到一个C −维的特征向量中。其中$C = \sum_{i=1}^{n} c_i$ 通常情况下，这个提取的过程只对生成器合成的图像有作用，而无法直接对真实的图像工作。然而，人们都可以通过任何基于梯度的优化或更复杂的方案来优化给定测试图像的隐藏特征。 由此产生的隐藏特征允许类似的方式构造特征映射

带提取表征的分割

为了解决少镜头分割的问题，首先在目标类的图像上训练一个GAN，并通过输入随机隐藏特征生成k个随机图像。 然后计算特征映射并手动注释这些k个图像的对象部分。k个特征映射和注释一起形成了我们的监督训练对，可用于训练分割模型，如多层感知器或卷积网络。 为了分割一个测试图像，使用上述隐藏特征优化对测试图像计算像素级特征映射，然后将其提供给训练过的分割网络

扩展：自动镜头分割网络

使用GAN计算像素级特征向量有许多限制。

1. 测试图像需要靠近GAN建模的图像分布，否则隐藏特征优化再现测试图像的效果可能较差，导致特征向量较差。
2. 依赖GAN通过隐藏特征的优化过程生成特征向量很耗时，而且如果GAN的模型很大，那么开销很大

为了克服这些限制，使用训练好的GAN合成一组图像，并通过网络预测分割映射这些图像生成数据对。为了保留网络预测过程中所有的概率信息，分割映射中每个像素都由所有部分标签的一组日志记录值保存，而不是一个简单的部分ID。

因此，不能使用softmax和argmax来生成分割映射。利用训练数据来训练另一个网络，解决新图像的分割不依赖GAN或它的特征映射，称这个过程为自动镜头分割。另外，使用数据增强能够检测不同尺度和方向的对象。

实施

1. 在目标类上训练一个GAN，生成一些带像素级表征的图像，然后手动地分割注释这些图像
2. 训练一个少镜头分割网络接受上面地像素级表征，预测一个分割输出
3. 自动镜头分割，使用相同地GAN生成一个大的图像数据集，使用训练过地少镜头网络来预测这些图像地分割映射。这些生成地图像和相应地分割映射用来训练自动镜头分割网络

因此，我们一共需要训练三个网络

1. 对抗生成网络 Generative Adversarial Network

   > 使用StyleGAN2 生成图像

2. 少镜头分割网络 Few-shot Segmentation Network

   > 以多通道像素作为输入，分割映射图作为输出
   > 使用两种不同地架构：全卷积网络CNN或者多层感知器MLP

3. 自动镜头分割网络 Auto-shot Segmentation Network

   > 使用GAN生成地图像及其相应地分割图进行训练

实验结果

评估内容：

1. 在三个对象类上少镜头和自动镜头地性能对比，并与基线做比较
2. 评估少镜头分割网络的替代结构
3. 展示视频分割结果
4. 研究用于特征选择的层的选择对分割性能的影响
5. 测试是否可以分割任意或者不规则的形状，不对应语义
6. 探索其他生成模型的性能，如VAE

实验如下：

1. 人脸部分分割

2. 汽车部分分割

3. 马部分分割

4. GANs的性能分析

> 三种大小的多层感知器：
> 0个隐藏层 合理的分割掩码
> 1个2000个节点的隐藏层
> 2个2000和200个节点的隐藏层 需要一个非线性的MLP分类器才能在复杂区域获得更准确的边界

结论

这篇文章提出了一个简单却强大的方法，重新利用GANs进行合成。利用GANs提取易于识别的像素的特点，实现了在有很少图像注释的情况下进行部分分割，与需要多标签的完全监督学习基线具有竞争力