A paper a day keeps trouble away

论文地址：https://arxiv.org/abs/1611.07004

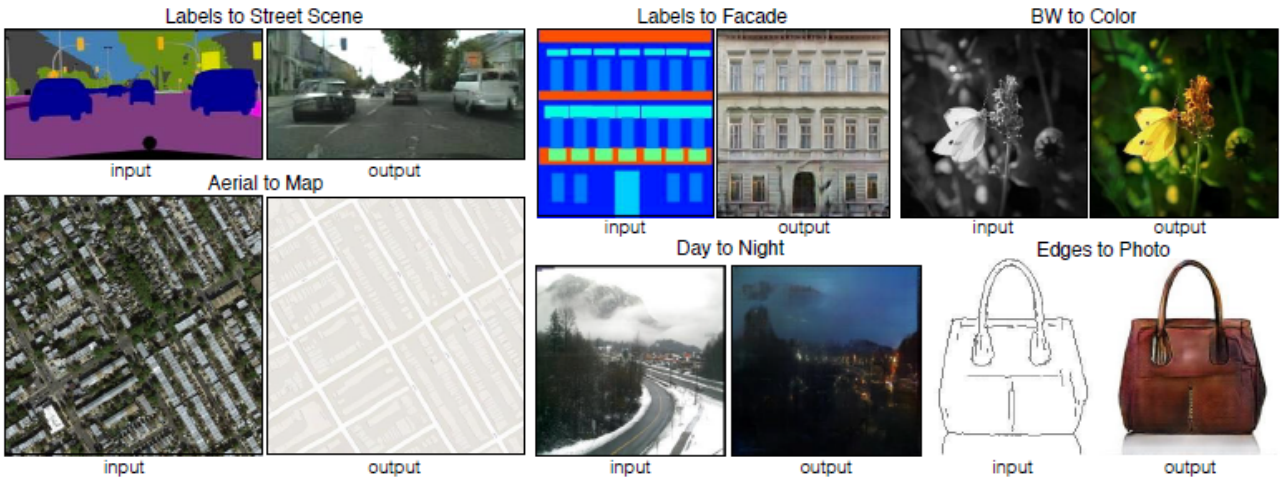
论文GitHub：https://phillipi.github.io/pix2pix/

Image-to-Image Translation with Conditional Adversarial Networks

Phillip Isola　　Jun-Yan Zhu　　Tinghui Zhou　　Alexei A. Efros

Berkeley AI Research (BAIR) Laboratory
University of California, Berkeley
{isola,junyanz,tinghuiz,efros}@eecs.berkeley.edu

这篇论文讲的主要内容是使用条件对抗生成网络来实现图像转换，其中涉及到很多东西很具有启发意义，是一篇很好的文章。首先给出它可以实现的效果图，有一个直观的认识。



摘要

图像转换是很多图像问题的基本抽象，在这篇文章之前已有很多的模型可以用来处理相关的问题。但是它们有一个共同的不足之处，虽然可以一定程度上完成转换的工作，但是需要人工的设计有效的损失函数，而且需针对不同的问题设计不同的损失函数，某一个具体问题的解决方法很难再用到其他的问题上。

本文作者所提出的这种条件网络就很好的解决了这个难题，它不仅可以很好的学习到输入图像到输出图像之间的映射，同样可以自动的学习到损失函数，省去人为的精心设计麻烦。这样得到的转换后的图像不仅效果好，而且学习到的模型的通用性更强。

在处理图像问题的模型中，CNNs是一个很重要同时也是使用很普遍的一种方法。它通过不断的学习，使得损失函数最小化，来达到有效处理的目的。尽管网络学习的过程是自动的，但是它同样需要我们设计有效的损失函数，而且如果使用欧式距离来评估将得到很模糊的结果。

而GAN的提出就可以有效的缓解这种尴尬的情况，它不仅可以区分图象是真实的还是生成的，还可以自动学习到一个损失函数，训练的过程就是使得损失函数值最小的过程。

本文的主要贡献在于两点：

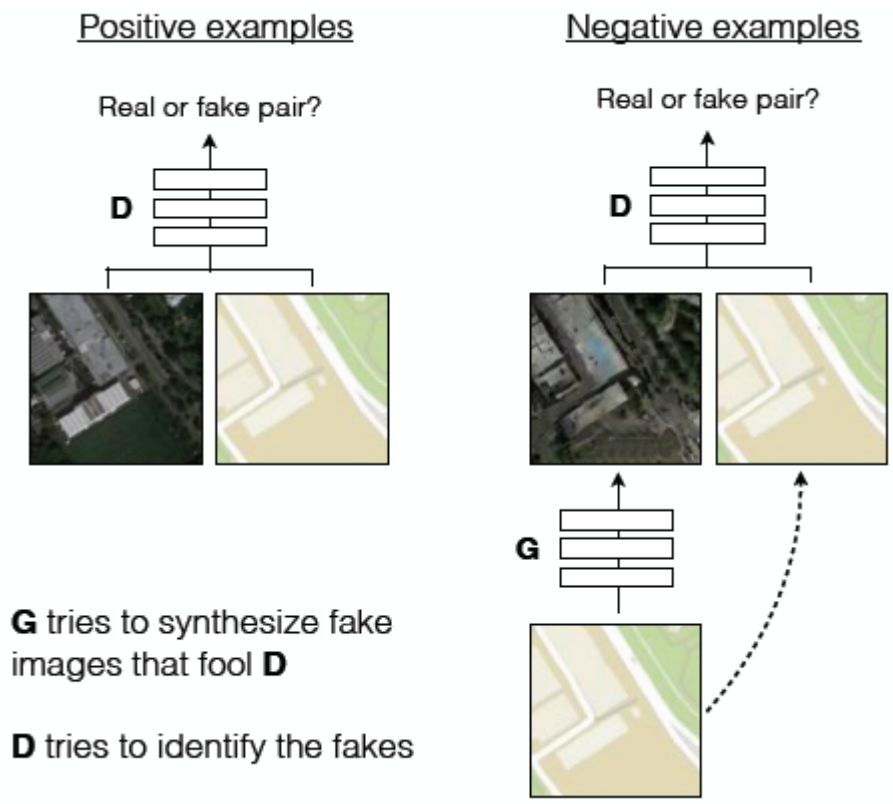
- 阐述了cGANs在很多的图像问题上都可以取得不错的效果
- 提出一个简单有效的框架，并对比了几种不同的架构的效果

相关工作

这一部分最重要的就是作者在生成器和判别器架构上所做的不同的选择：生成器G使用U-Net结构，判别器D使用卷积PatchGAN分类器，PatchGAN这种结构可以很好的捕捉图像局部的特征信息，而且不同尺寸的patch会有不同的效果。

方法

在《Coditional Generative Adversarial Nets》中，将随机噪声z 和条件x输入到生成器中产生y ,判别器接收y 和条件x 作为输入，来判别输入的图像是真实的还是生成的。它的基本框架如下所示：



根据G和D之间的对抗过程，CGANs的目标函数可以表示为：

$$\mathcal{L}_{cGAN}(G,D)=\mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x,y)]+\mathbb{E}_{x\sim p_{data}(x),z\sim p_z(z)}[\log 1-D(x,G(x,z))]$$

最后得到的最优生成器G^* 就是最大最小化目标函数的结果，即

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D)$$

而为了确定条件对于D的重要程度，将其去掉后得到的目标函数就类似于标准GAN的形式：

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_{y \sim p_{data}(y)}[\log D(y)] + \mathbb{E}_{x \sim p_{data}(x), z \sim p_z(z)}[\log 1 - D(G(x, z))]$$

这里创新的一点是**使用L1距离代替L2距离来评估生成图像的真实性**，这样得到的图像就不会那么模糊：

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y \sim p_{data}(x,y), z \sim p_z(z)}[\|y - G(x, z)\|_1]$$

则目标函数就变成了：

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

在CGAN中添加噪声z在一定程度上缓解了模式崩溃问题，但是这里作者使用了一种更为巧妙的方法，他**以dropout的形式提供噪声z**，并且在训练和测试中只应用于生成器的某几层。但是尽管这里使用dropout形式的噪声，输出之间的随机性仍然很好，并不能很好的解决模式崩溃这个问题。

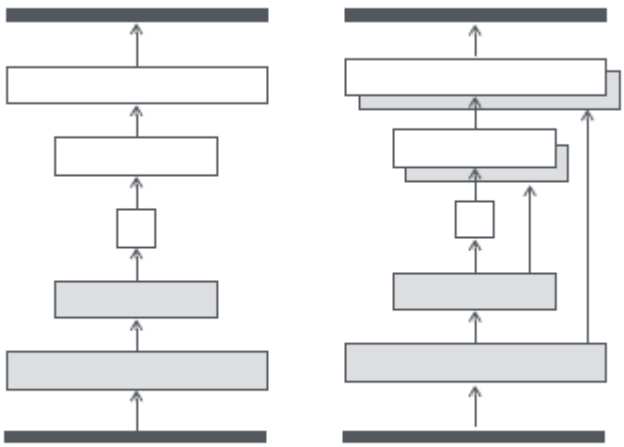
所以如何设计一个可以产生随机输出的GAN，来捕获模型完整的熵，即更好的理解模型的随机性，将是一个很重要的急需解决的事情。

在网络架构方面，G和D都使用conv-BatchNorm-Relu的卷积单元形式。

在图像转换问题中，一个很重要的特征就是将高分辨率的输入网格映射到一个高分辨率的输出网格。此外，对于我们考虑的问题，输入和输出在表象上虽然不一样，但是在底层结构上却是一致的。基于这个重要的特征，之前使用encoder-decoder network的形式来解决，但是它在所有的信息的传输过程中，需要经过所有的层。

而且对于许多图像转换问题，在输入和输出之间存在很多可以共享的低级信息，在网络中直接传递这些信息可能会有所帮助

为了更好的利用上面提到的这个重要的特征，同时避免encoder-decoder network中的问题，作者模仿U-Net的结构，增加一种**跳线连接（skip connection）**，特别的，我们在每第i 层和第$n-i$ 层之间添加跳线，其中n 是网络的总层数,每根跳线简单的将第i层和第n−i层的特征通道连接在一起。具体形式如下所示



Encoder-decoder　　U-Net

上面提到使用L1距离可以更加清晰的图像，同时它在很多情况下也可以很好的捕捉到图像的低频信息，这样的话设计GAN时只需要考虑如何只对高频信息建模即可。而且在对高频信息建模时，关注局部的图像就足够了，所以使用PatchGAN就很合适，我们只需要在整张图像上运行这个判别器，最后将平均值作为D的输出。

而且很小的patch size仍然可以获得高质量的结构，同时PatchGAN的参数更少，训练速度更快，可以用在任意大的图像上。如果假设通过patch size分割后的像素之间相互独立，那么使用PatchGAN将相当于将图像建模为一个马尔科夫随机场。

这里使用和《Generative Adversarial Nets》中相同的方法，交替训练G和D，在优化方法方面，使用miniBatchSGD并应用Adam优化器。

在推理的时候，使用和训练阶段相同的方式来运行生成器。在测试阶段使用$dropout$ 和$batch\ normalization$ ，而且使用测试批数据的统计值来执行批正则化，而不使用训练数据整合的统计值。

实验

在实验部分，作者使用了很多的数据集做了很多方面的实验来证明所提出框架的通用性。这部分占了很大的篇幅，但是并不难理解，具体来说就是对之前的一些验证，这里不打算每个实验都列出，具体内容可看原论文，主要提一些相对重要的东西。

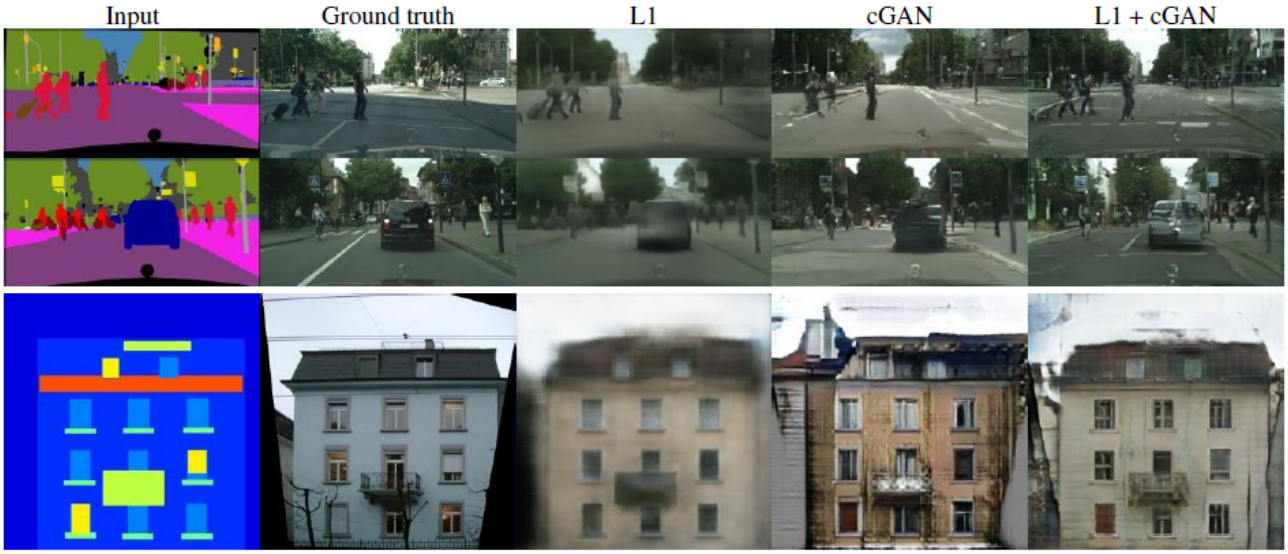
首先是他所使用的评估指标，有如下两个：

- AMT感知调查：向Turkers提供一系列真实的图片和我们算法生成的“假冒”的图片。在每次试验中，每幅图片出现1s，图片消失后，Turkers需要在规定时间内指出哪一副图片是假的。每轮实验前10张图片用于练习，Turkers会得到正确答案。在主实验中的40次试验中，不会给出答案。
- FCN分数：从直观上理解，用真实图片训练的分类器有能力对合成的图片进行正确的分类。为了达到这个目的，这里使用的是FCN-8s结构做语义分割，并在CityScapes数据集上训练，然后我们根据这些照片合成的标签，通过合成图片的分类准确性为合成照片进行评分。

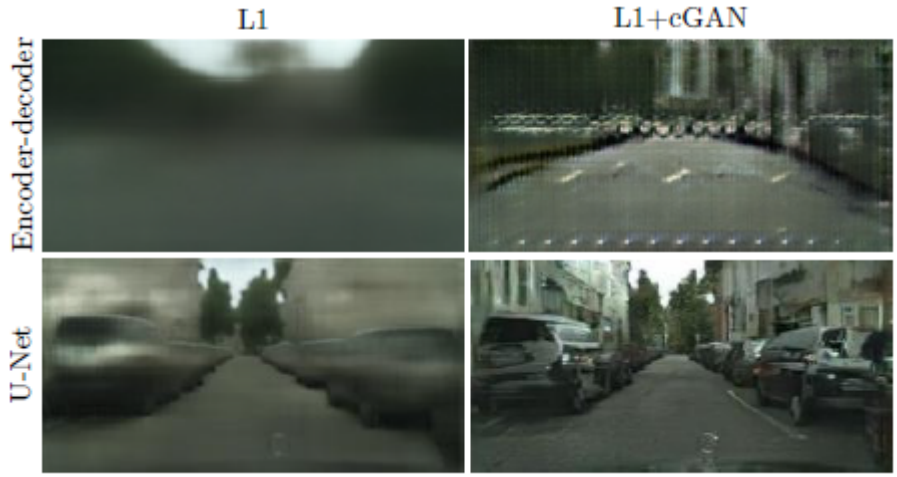
通过实验证明了使用L1+CGAN的形式的效果更好

Loss	Per-pixel acc.	Per-class acc.	Class IOU
L1	0.44	0.14	0.10
GAN	0.22	0.05	0.01
cGAN	0.61	**0.21**	**0.16**
L1+GAN	**0.64**	0.19	0.15
L1+cGAN	0.63	**0.21**	**0.16**
Ground truth	0.80	0.26	0.21

证明了使用CGAN可以得到色彩更加丰富的图像，同时它的输出也更加符合真实分布。



而且证明了U-Net的结构可以得到更真实的图像。



最后在patch size的选择上，证明70×70 得到的效果是最好的。

Discriminator receptive field	Per-pixel acc.	Per-class acc.	Class IOU
1×1	0.44	0.14	0.10
16×16	0.62	0.20	**0.16**
70×70	**0.63**	**0.21**	**0.16**
256×256	0.47	0.18	0.13

综上，这篇文章使用的都是70×70 de PatchGAN，而且都使用L1+CGANs形式的损失函数。

更多的实验结果可见原论文和它的GitHub。

最后提到的一段话：To our knowledge, this is the first demonstration of GANs successfully generating “labels”, which are nearly discrete, rather than “images”, with their continuous valued variation。他说了可以使用GANs来生成离散的标签，具体怎么实现的很值得接下来寻找答案。

个人感悟

- 如何设计噪声z zz或者说如何设计GAN的结构来使的生成的额图像获得更大的随机性，是一个很重要的事情，可以思考一下；
- 如何使用GANs来生成离散的标签，如果真实可行，那么将可以应用到很多的场景中
- 如何将不同的神经网络的架构和GAN进行结合，取长补短，是解决某些复杂问题的一个很好的思
- cGANs可以产生引人注目的颜色，但如何使用它来产生灰度或去饱和的图像仍是一个需要思考的问题
- 能否将其应用到监控视频领域来帮助警察更好的破案呢

更多

https://www.jianshu.com/p/a1058084288c

https://affinelayer.com/pixsrv/

https://www.microsoft.com/developerblog/2017/06/12/learning-image-image-translation-cyclegans/

https://cloud.tencent.com/developer/article/1089012

https://becominghuman.ai/unsupervised-image-to-image-translation-with-generative-adversarial-networks-adb3259b11b9