参考

论文：[Conditional Generative Adversarial Nets](https://arxiv.org/abs/1411.1784)

网站：<http://blog.csdn.net/solomon1558/article/details/52555083>

<http://blog.csdn.net/wspba/article/details/54666907>

GAN不需要预先建模就可进行训练，训练太过自由，对于像素较较多的图片，过于庞大的数据库，该种训练方法不太合适，要通过给GAN中的G和D增加一些约束性条件来解决训练过于自由的问题。

CGAN是给生成模型G和判别模型D的建模中引入条件变量y，y的类型据情况而定，这个改进被证明是非常有效的。

典型应用是图像的一对多标签问题。一个图像可以有多个描述标签，不同的人标注的结果不同，使用CGAN也可对图像进行多个标签的描述， 以输入的图片作为条件，使标注的词语更加全面，人工标注的结果可视为辅助的tag，不是groundtruth。

1. 目标函数

Y通过一个额外的输入层分别输入到G和D中。在生成模型中，y和z先经过一个隐藏层进行组合后再输入到生成模型中；在判别模型中，x和y也通过一个隐藏层合并后输入到判别模型中，如图所示：

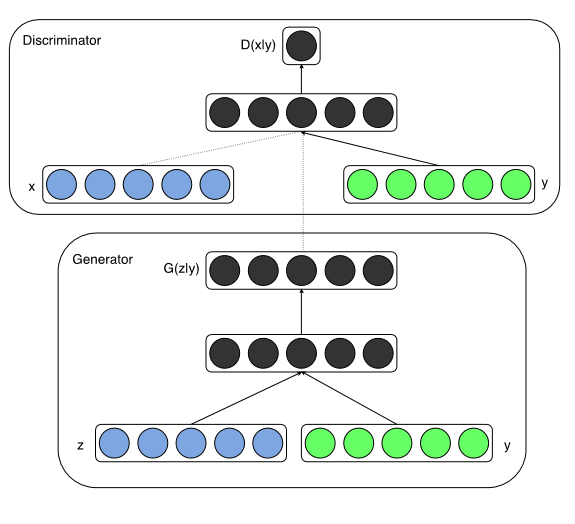


图1 Conditional GAN

1. 两个实验
2. MNIST数据集实验

在MINIST数据集上以类别标签(class labels)作为条件，标签被one-hot编码。根据标签条件信息，生成对应的数字。生成模型的输入为100维服从均匀分布的噪声数据z，z和y分别映射到隐藏层(单元数为200和1000)，用ReLU激活，再将这1200个单元联合，输入到第二层中，经过sigmoid归一化处理后得到784维的生成输出(28\*28单通道图像)。判别模型的输入为784维的输入数据x和one-hot编码的类别标签y。分别映射到隐藏层(240 units and 5 pieces, 50 units and 5 pieces)，maxout层激活。这两个隐藏层再映射到联合maxout层(240 units and 4 pieces)。最后经过sigmoid归一化处理得到输出。

用Parzen window-based log-likelihood estimates做结果评估，数值越大越好。结果表明，条件生成模型的效果与某些模型的效果有可比性，但比GAN要差，后续在超参数空间或条件模型结构上改进后可能会得到更好的结果。

1. 多模态学习(multi-modal)用于图像自动标注

在数据集Flickr 25,000上进行多标签标注，该数据集中包含大量拥有标签的图片，且很多标签是user-generated metadata(UGM, 用户自己产生的标签)，这类标签的特点是更加具有描述性，更加接近自然语言，对于同一幅图片，不同的人会给出同义标记。CGAN以图片特征为条件，产生一个标签向量分布。

对于image representation，用Alexnet提取，以最后一层fcn的4096维输出作为图片特征。对于word representation，先从YFCC100M中获得一个语料库，用它训练一个skip-gram模型，生成200维的语义向量。用卷积模型和skip-gram模型提取Flicker数据集中的图片和tag特征，用来训练条件生成对抗网络。

产生模型的输入为100维的噪声数据和4096维的图片特征数据。噪声数据映射到500维输出的ReLU层，特征数据映射到2000维输出的ReLU层，这两个隐藏层联合映射到一个200维输出的线性层，得到词语向量。

判别模型分别以500维和1200维的ReLU隐藏层作为词语向量和图像特征的隐藏层，并将它们的输出联合输入到一个maxout层(1000 units and 3 pieces)，经过sigmoid单元归一化得到100维的词语输出，从中选出10维最为最终结果。

从实验结果可以看出，如果图片中有很明显的目标，CGAN会更关注于目标的特点来形容，忽略了背景(特征提取导致)。是具象的描述，很少掺杂情感描述词。而图片的人工标注中，既关注了背景信息，又加入了感情色彩。

1. 模型的展望
2. 根据实际应用开发出更加复杂的模型，进行更加细致和全面的分析。
3. 在训练时，以tag为单元进行训练，每回训练一个tag，如果图片有多个tag，就要反复训练多次。作者希望模型可以一次训练多个tag，使训练更加高效。
4. 希望可以用一个统一的模型来做language model，而不是将训练分成中断的几部分。