今天是年二八，先给大家拜个早年，祝大家新年快乐！

新年在即，学习不能停。在进入本期话题InfoGAN[1]和improved GAN[2]之前，我们先来聊聊GANs有哪些可研究的方向。

下面说说我的看法。我把研究的方向分为三大块：theory、optimization、application。

在theory上的研究可以分为博弈论视角和非博弈论视角。收敛性是theory的一个很重要的研究方向。然而，理论研究非常难，一方面GANs一般是基于deep learning的，而deep learning的理论根基尚且未知；另一方面，有理论收敛性保证的零和博弈GANs在实际应用中训练效果并不好，实际应用的启发式GANs并非零和博弈，很难从理论上做分析。

至于optimization，我们知道GANs很难训练，容易发生模式坍塌（mode collapse），从这点上可以找到两个研究课题：如何找到合适的优化方法以稳定训练过程、发生模式坍塌的原因何在。正如Ian Goodfellow所言，寻找纳什均衡比优化loss function要难得多。经验上或者理论上找到一类相对容易寻找纳什均衡的GANs，也是一个不错的研究方向，DCGAN做的正是这方面的工作。对GANs提出合适的优化方法目前还没看到相关工作，多数optimization的文章还是在研究训练GANs的一些tricks，如DCGAN、improved GAN。总的来说，optimization上的研究，可以从loss function、training procedure、flamework上入手。

还有一类是研究GAN的application，这也是一个相对容易一些的研究方向。关于GAN的应用，在推出GANs系列的第一期我们就提到过，如根据描述生成图像、卫星图像与地图的转换、草图生成真实图像等。后续也会解读这方面的文章。

我尝试在这上面做些研究，一开始有一堆想法，然后一个个地思考、否定、心寒。看懂别人的工作似乎不难，但是自己研究还是很难的。But不要气馁，一切才刚刚开始。GANs的研究也才2年。

好了，进入正题。今天要介绍的两篇论文都来自于OpenAI，第一篇介绍了一种新的GAN，叫InfoGAN，它对generator和generator的loss做了一些修改，并取得了很好的效果。第二篇对GAN提出了一些改进，并将GANs应用到半监督学习中。

**InfoGAN**

一般的GAN的generator输入是“噪声”（noise，也称为latent code，姑且称为“图编码”），输出是根据“图编码”产生的图像【PS：我觉得这种做法有问题，但是想不到改进的方法】。我们完全不知道generator会怎样利用“图编码”，“图编码”可能完全交织在一起，很难确定“图编码”每个维度的具体语义，或者说，不知道“图编码”的每个维度分别控制着图像的哪些特征。一般的GANs求解的是以下的minmax博弈：



作者认为这种generator的输入是unstructured latent code，并不能很好地解释图像产生的参数，应该将generator的输入分为两部分，一部分还是noise，不可压缩的噪声向量，记为z；另一部分是structured latent code，控制数据分布的语义特征，如视角、光照等，记为c。简而言之，noise负责控制生成图像的内容，structure latent code负责控制图像的一些属性特征。

一般来说，一张图像具有很多属性特征，设有L个属性，记为



不妨假设它们是相互独立的，即



控制图像属性的structure latent code应该是与图像内容无关的变量，也就是说，p(c|x)的熵应该尽可能地小，也就是说，c与G(z,c)的互信息I(c, G(z,c))应该尽可能地大。换句话说，当c给定以后，图像x的不确定性减少量应该尽可能地大。

为了实现最大化互信息，以使structured latent code能够控制图像的属性，作者在loss function中引入互信息正则化项，因此，InfoGAN是如下的minmax博弈：



从上式可以看出，互信息正则项只对generator起作用。

然而，通过上式我们并不能启动训练，因为它实际上需要知道后验概率p(c|x)，而这是我们无法得知的。没办法得到后验概率，那我们就让模型自己学后验概率...

作者从理论上推导了，如果用Q(c|x)来逼近后验概率P(c|x)，那么互信息I(c, G(z,c))具有如下的下界：



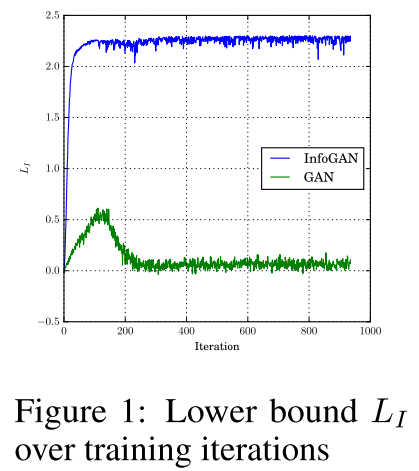
更详细的讨论可以参看原文[1]。

实际操作中，Q(c|x)的计算与Discriminator共享网络，除了最后一层全连接层不同，不妨称为Q网络。这样做可以减少计算代价。如果structured latent code是categorical变量，采用softmax的概率输出作为Q(c\_i|x)即可；如果是连续变量，一般采用Q网络输出构造一个高斯分布。

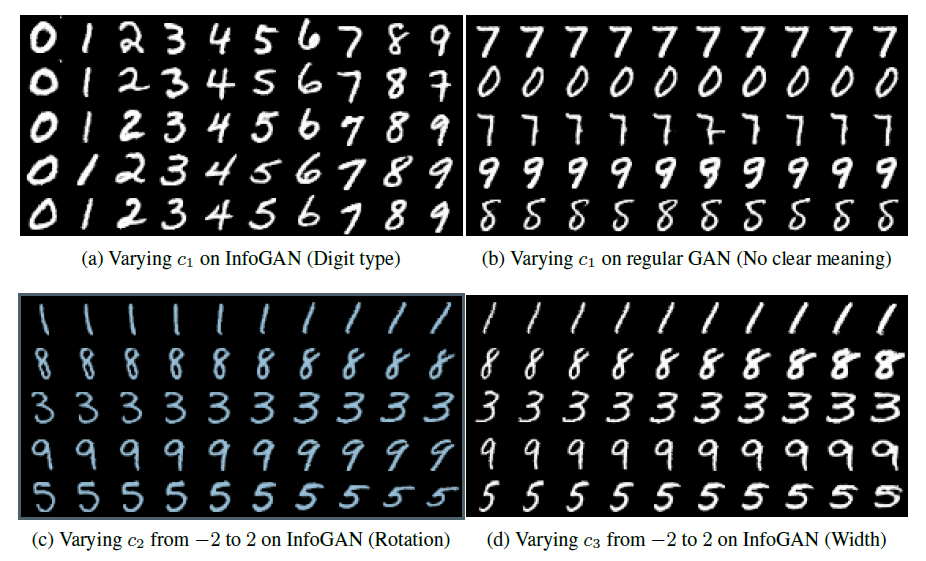
**InfoGAN实验**

InfoGAN采用DCGAN提出的限制设计generator和discriminator网络。每个实验具体的网络架构、实验参数可以在文献[1]的Appendix中找到。Generator的学习率设置为0.001，discriminator的学习率设为0.0002，正则化项参数\lambda设置为1（face dataset上除外）。

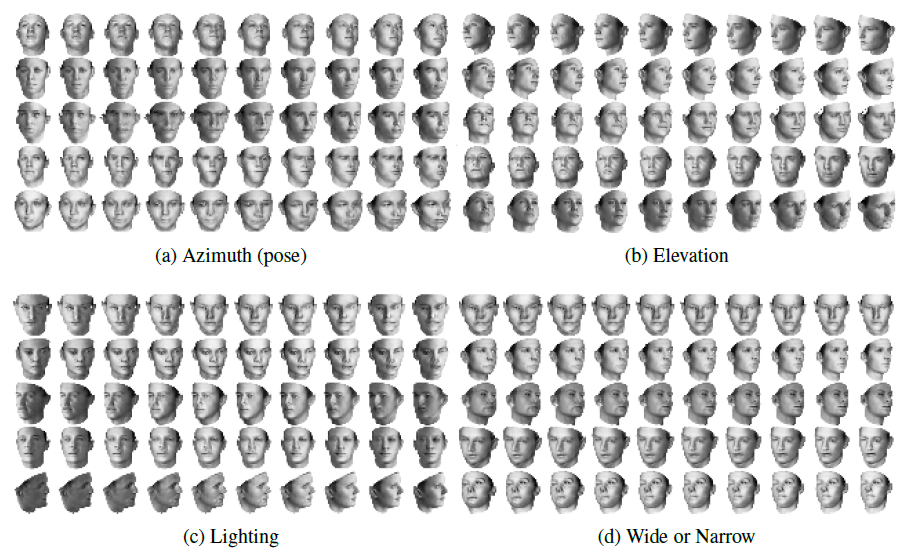
实验发现，互信息很快就达到了最大，表明文中给出的最大化互信息方法是有效的。



MNIST数据集上，InfoGAN学到的3个structured latent code具有明确的含义：其中c\_1表示数字类型；c\_2表示旋转角度；c\_3表示笔画宽度。如下图所示：



此外，在生成人脸和桌椅上也有类似效果：





**Improved GAN**

Improved GAN实际上是提出了几个使得GANs更稳定的技术，包括：feature matching、minibatch discrimination、historical averaging、one-sided label smoothing、virtual batch normalization。下面我们一一介绍：

* **Feature matching**

GANs的loss是建立在discriminator的输出上的，即discriminator输出的交叉熵，这可能导致GANs训练不稳定，毕竟给予generator的信息太少了，而图像空间又太大了。为了让训练更稳定，作者提出了feature matching的方法。所谓的feature matching，即是要求generator产生的图像在经过discriminator时，提取的特征尽可能地接近（匹配）自然图像经过discriminator时提取的特征。设discriminator的某个中间层对于输入x的输出为f(x)，作者提出的feature matching，实际上是用下面的loss function替换以前的交叉熵loss：



这个loss比交叉熵loss更加直接，对generator的指导作用更大一些。

* **Minibatch discrimination**

作者提出， generator的输出并没有相互作用，发生模式坍塌（mode collapse）时，没有任何信息能够引导它走出模式坍塌。为了防止这个情况发生，作者提出了generator的每个minibatch输出都应该具有足够的分辩能力，也就是说，每个minibatch不应该太相似，惩罚minibatch的相似性，能够在模式坍塌发生时，产生足够的梯度信息，引导它走出模式坍塌。设discriminator的某个中间层对于输入x的输出为f(x)，对于generator产生的图像x\_i，它的特征是



设generator的一个minibatch输出为x\_1, x\_2, ..., x\_n。Minibatch discrimination（可以看成是一个layer，称为minibatch layer）具体计算如下：

1. 对每个f(x\_i)，右乘以下张量



得到矩阵



1. 计算cross-sample distance



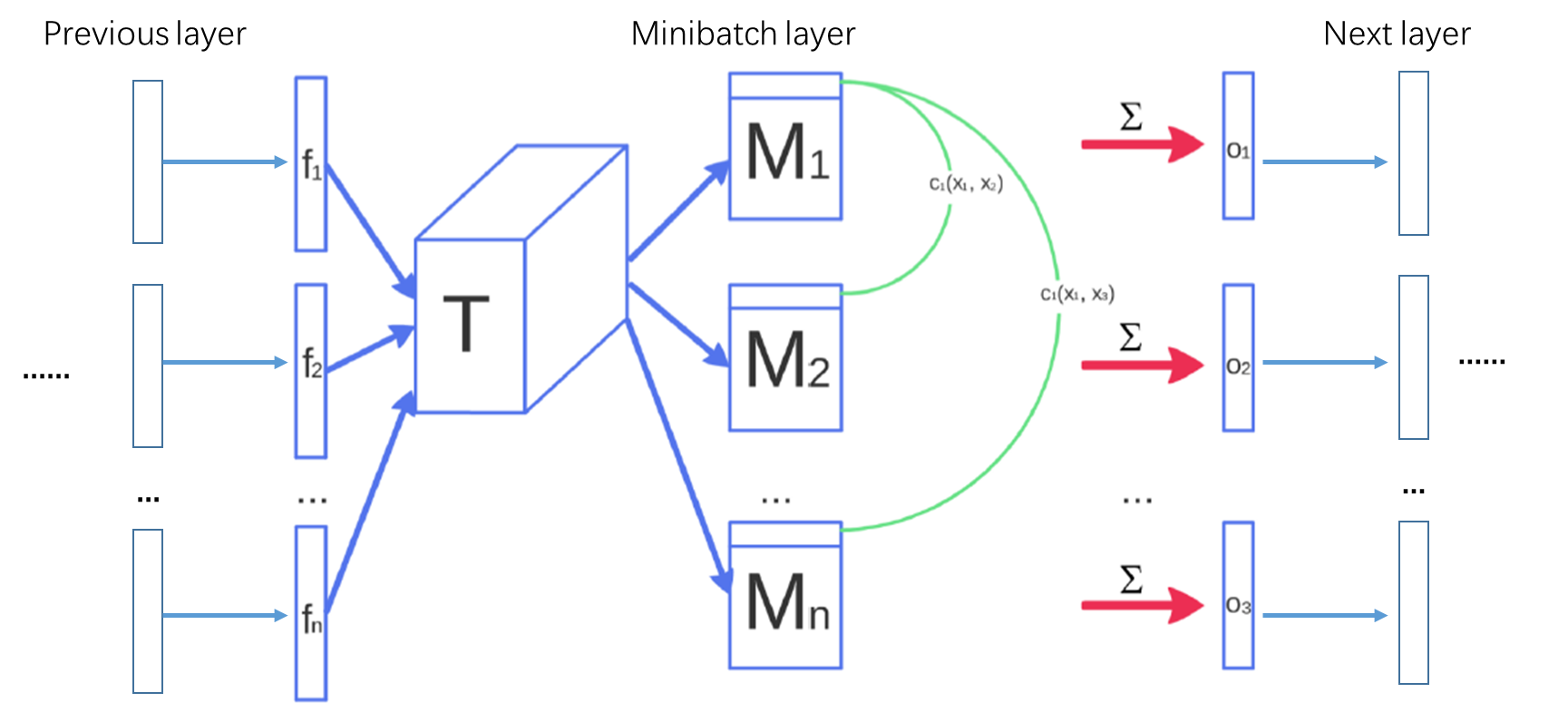
1. 计算minibatch layer的输出：







需要指出的是，作者提出的minibatch discrimination是在discriminator原来的网络的某个中间层后面插入minibatch layer，minibatch layer的输出作为原来discriminator下一层的输入。示例图解如下：



* **Historical averaging**

受到虚拟博弈算法（fictitious play algorithm, George W Brown, 1951）[3]的启发，作者提出，对博弈双方的loss function均加入历史信息项，即



其中，θ[i]表示历史时刻i的参数值，t为已经进行过的博弈轮数（迭代次数）。

实验发现，这种historical averaging的方法能够求解低维、连续非凸博弈的均衡点。

* **One-sided label smoothing**

标签平滑处理（label smoothing）将分类器的两类输出0,1替换为平滑后的结果a,b（其中: 0<a<b<1）。它能够降低神经网络对于对抗样本的脆弱性。然而，标签平滑处理也存在问题，由于训练集是有限的，在p\_data接近0，p\_model很大的地方，p\_model主导了discriminator，这将导致没有足够的梯度引导p\_model向p\_data收敛。因此，作者提出one-sided label smoothing，将正类标签平滑处理为\alpha，对负类标签不做处理，仍置为0。

* **Virtual batch normalization**

DCGANs将batch normalization引入到GANs中，并取得了巨大成功。但这种batch normalization也存在问题，每个样本对应的网络输出都高度依赖于同个batch的其他样本，为了避免这种依赖性，作者提出了virtual batch normalization，即每次做batch normalization时，用到的统计信息\mu, \sigma并不是当前batch的统计信息，而采用一批固定样本的统计信息。即，每次迭代需要计算两个batch，一个是训练集的输入batch，另一个是固定的batch（与迭代次数无关，每次均采用该batch），在需要batch normalization时，用固定batch的统计信息\mu，\sigma对训练集输入batch和固定的batch分别进行normalization。需要注意的是，VBN使得每次迭代的前向计算部分计算量是以前的两倍。实际操作时，一般只对generator使用VBN，以降低计算的代价。

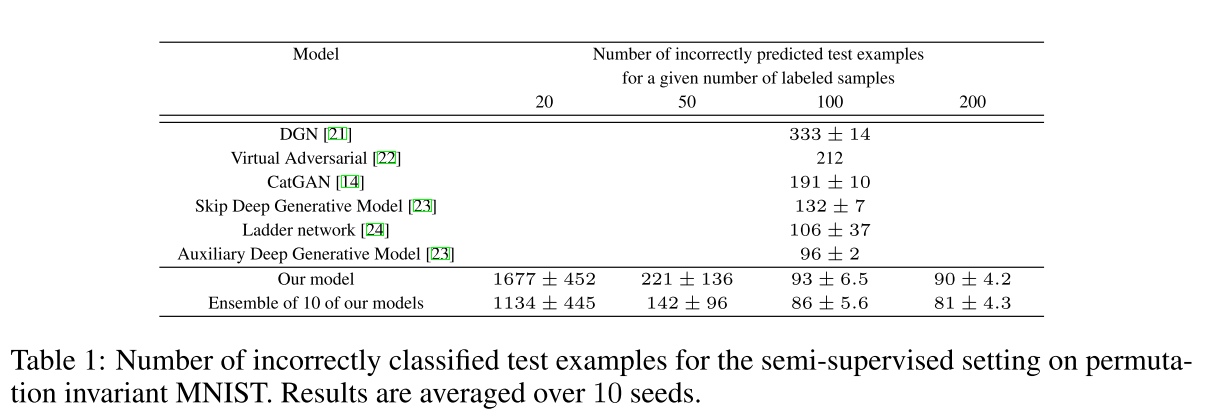
**Semi-supervised GANs**

GANs属于非监督学习，improved-GAN的作者还提出了半监督版本的GANs。假设我们需要对图像进行分类，分为K类，半监督的GANs将generator产生的图像作为第K+1类：fake image。一般来说，我们将discriminator设计为半监督学习的分类器，将样本分为K+1类。GANs在训练时只考虑fake images和real images两类，fake images为第K+1类，discriminator输出的第K+1类的概率即为图像属于fake image的概率，real images为前K类，discriminator输出的前K类的概率之和即为图像属于real image的概率。

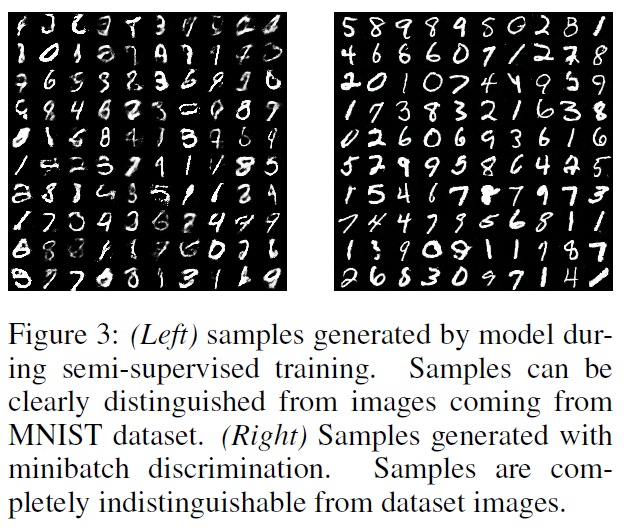
作者发现，前面提到的minibatch discrimination对于这种半监督的GANs训练并没有效果，而feature matching则对半监督GANs的提升效果非常明显！

**Improved GAN实验**

作者设计的GANs也是基于DCGAN的。在MNIST数据集上做半监督GANs训练，用上feature matching或者minibatch discrimination，发现用feature matching的半监督GANs效果更好，在MNIST的分类上达到了半监督方法的state-of-the-art。



如果考虑产生的图像质量，半监督GANs产生的图像质量并不是很高，容易区分，如果非监督GANs采用minibatch discrimination，产生的图像质量非常高，人眼很难判断是图像是否是generator产生的。图像质量实验结果参见下图：



此外，作者还在CIFAR-10、SVHN等数据集上做了半监督GANs分类，同样达到了非监督分类上state-of-the-art的效果。

**代码**

InfoGAN代码：<https://github.com/openai/InfoGAN>

Improved GAN代码：<https://github.com/openai/improved-gan>

**参考文献**

1. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *arXiv:1606.03657* [cs.LG], 1–14.
2. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved Techniques for Training GANs. *Nips*, 1–10.
3. George W Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374-376, 1951.

**下期预告**

下期在春节期间，我们将介绍LAPGAN，它将拉普拉斯金字塔模型引入到GAN中。此外，如无意外，我们还将介绍GANs的其中一种很有意思的应用，根据文字产生图像。参考文献：

1. Denton, E., Chintala, S., Szlam, A., & Fergus, R. (2015). Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. Arxiv, 1–10.
2. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative Adversarial Text to Image Synthesis. Icml, 1060–1069.