InfoGAN

Dec 20, 2016

1. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

contributions

- We propose and evaluate a set of constraints on the architectural topology of Convolutional GANs that make them stable to train in most settings. We name this class of architectures Deep Convolutional GANs (DCGAN)
- We use the trained discriminators for image classification tasks, showing competitive performance with other unsupervised algorithms.
- We visualize the filters learnt by GANs and empirically show that specific filters have learned to draw specific objects

总结了许多对于GAN的网络结构设计和针对CNN这种网络的训练经验。比如，他们用strided convolutional networks替代传统CNN中的pooling层，从而将GAN中的生成模型（G）变成了fully differentiable的，结果使得GAN的训练更加稳定和可控。

https://github.com/Newmu/dcgan_code

2. InfoGAN: using the variational bound on mutual information

- to learn disentangled representations in a completely unsupervised manner
- a generative adversarial network that also maximizes the mutual information between a small subset of the latent variables and the observation
- InfoGAN learns interpretable representations that are competitive with representations learned by existing supervised methods

What are disentangled representations?

- explicitly represents the salient attributes of a data instance
  - For example, for a dataset of faces, a useful disentangled representation may allocate a separate set of dimensions for each of the following attributes: facial expression, eye color, hairstyle, presence or absence of eyeglasses, and the identity of the corresponding person.
- disentangled representation can be useful for natural tasks that require knowledge of the salient attributes of the data, which include tasks like face recognition and object recognition

Related paper:

Disentangling factors of variation in deep representations using adversarial training

What are interpretable representations?

- interpretable representation is one that is easy for humans to understand
- Interpretability is closely related with disentanglement. This is because, in “human” domains of data like vision and audition, humans are remarkably good at inferring generative structure, and tend to internally use highly disentangled representations.

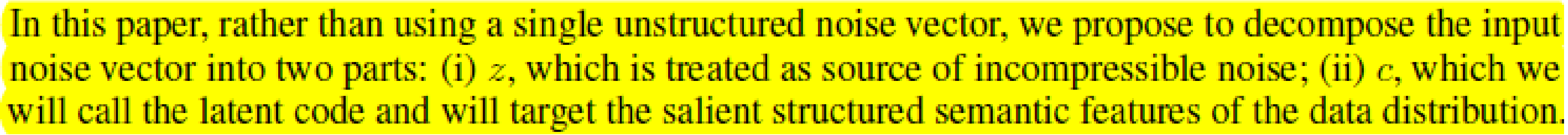
InfoGAN 的主要思想

- InfoGAN 的出发点是，GAN 的自由度是由于仅有一个 noise z，而无法控制， GAN 如何利用这个 z。于是，将 z 做了拆解，认为 GAN 中生成模型（G）应该包含的 “先验” 分成两种： （1）不能再做压缩的 noise z； （2）和可解释地、有隐含意义的一组隐变量 c_1, c_2, …, c_L， 简写为 c 。
- 主要思想是，当我们学习生成图像时，图像有许多可控的有含义的维度，比如笔划的粗细、图片的光照方向等等，这些便是 c ； 而剩下的不知道怎么描述的便是 z 。这样一来，InfoGAN实际上是希望通过拆解先验的方式，让 GAN 能学出更加 disentangled 的数据表示，从而既能控制 GAN 的学习过程，又能使得学出来的结果更加具备可解释性。为了引入这个 c ，InfoGAN利用了互信息的建模方式，即 **c 应该和生成模型 （G） 基于 z 和 c 生成的图片，即 G (z,c)， 高度相关** —— 互信息大。

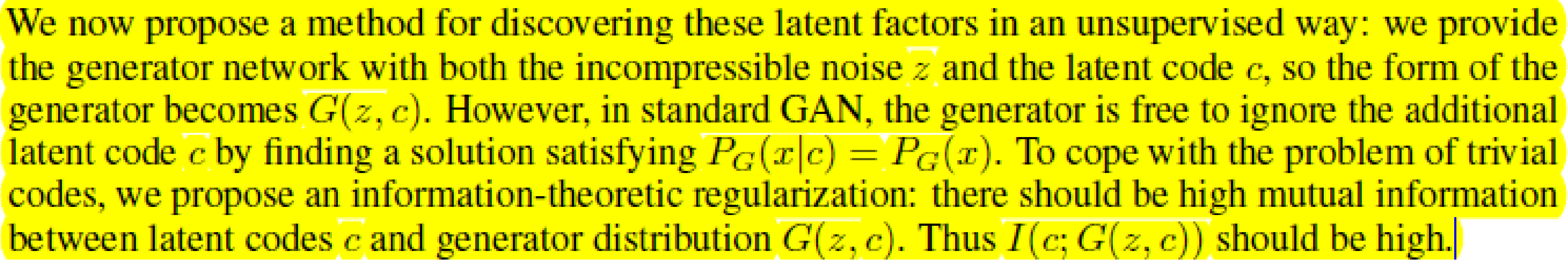
Mutual Information for Inducing Latent Codes

- The GAN formulation uses a simple factored continuous input noise vector z, while imposing no restrictions on the manner in which the generator may use this noise. It is possible that the noise will be used by the generator in a highly entangled way, causing the individual dimensions of z to not correspond to semantic features of the data.
- However, many domains naturally decompose into a set of semantically meaningful factors of variation. For instance, when generating images from the MNIST dataset, it would be ideal if the model automatically chose to allocate a discrete random variable to represent the numerical identity of the digit (0-9), and chose to have two additional continuous variables that represent the digit’s angle and thickness of the digit’s stroke. These attributes are both independent and salient, and it would be useful if we could recover these concepts without any supervision, by simply specifying that an MNIST digit is generated by an independent 1-of-10 variable and two independent continuous variables

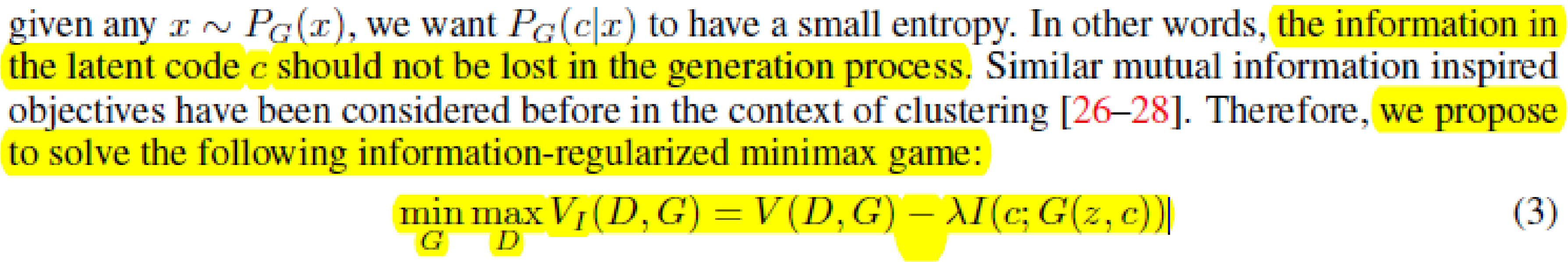
Mutual Information for Inducing Latent Codes



In this paper, rather than using a single unstructured noise vector, we propose to decompose the input noise vector into two parts: (i) z, which is treated as source of incompressible noise; (ii) c, which we will call the latent code and will target the salient structured semantic features of the data distribution.



We now propose a method for discovering these latent factors in an unsupervised way: we provide the generator network with both the incompressible noise z and the latent code c, so the form of the generator becomes $G(z, c)$. However, in standard GAN, the generator is free to ignore the additional latent code c by finding a solution satisfying $P_G(x|c) = P_G(x)$. To cope with the problem of trivial codes, we propose an information-theoretic regularization: there should be high mutual information between latent codes c and generator distribution $G(z, c)$. Thus $I(c; G(z, c))$ should be high.



given any $x \sim P_G(x)$, we want $P_G(c|x)$ to have a small entropy. In other words, the information in the latent code c should not be lost in the generation process. Similar mutual information inspired objectives have been considered before in the context of clustering [26–28]. Therefore, we propose to solve the following information-regularized minimax game:

$$\min_G \max_D V_I(D, G) = V(D, G) - \lambda I(c; G(z, c)) \tag{3}$$

利用这种更加细致的隐变量建模控制，infoGAN 将 GAN 的发展又推动了一步。首先，它们证明了 infoGAN 中的 c 对于 GAN 的训练是有确实的帮助的，即能使得生成模型（G）学出更符合真实数据的结果。其次，他们利用 c 的天然特性，控制 c 的维度，使得 infoGAN 能控制生成的图片在某一个特定语义维度的变化。

Mutual Information for Inducing Latent Codes

- A new term encourages high mutual information between generated samples and a small subset of latent variables cc. The hope is that by forcing high information content, we cram the most interesting aspects of the representation into c.
- If we were successful, cc ends up representing the most salient and most meaningful sources of variation in the data, while the rest of the noise variables z will account for additional, meaningless sources of variation and can essentially be dismissed as uncompressible noise.
- In order to maximise the mutual information, the authors make use of a variational lower bound. This, conveniently, results in a recognition model, similar to the one we see in variational autoencoders. The recognition model infers latent representation ccfrom data.

3. UNSUPERVISED LEARNING USING GENERATIVE ADVERSARIAL TRAINING AND CLUSTERING

- we propose an unsupervised learning approach that makes use of two components; a deep hierarchical feature extractor, and a more traditional clustering algorithm.
- We train the feature extractor in a purely unsupervised manner using generative adversarial training and, in the process, study the strengths of learning using a generative model as an adversary.
- We also show that adversarial training as done in Generative Adversarial Networks (GANs) is not sufficient to automatically group data into categorical clusters. Instead, we use a more traditional grouping algorithm, k-means clustering, to cluster the features learned using adversarial training

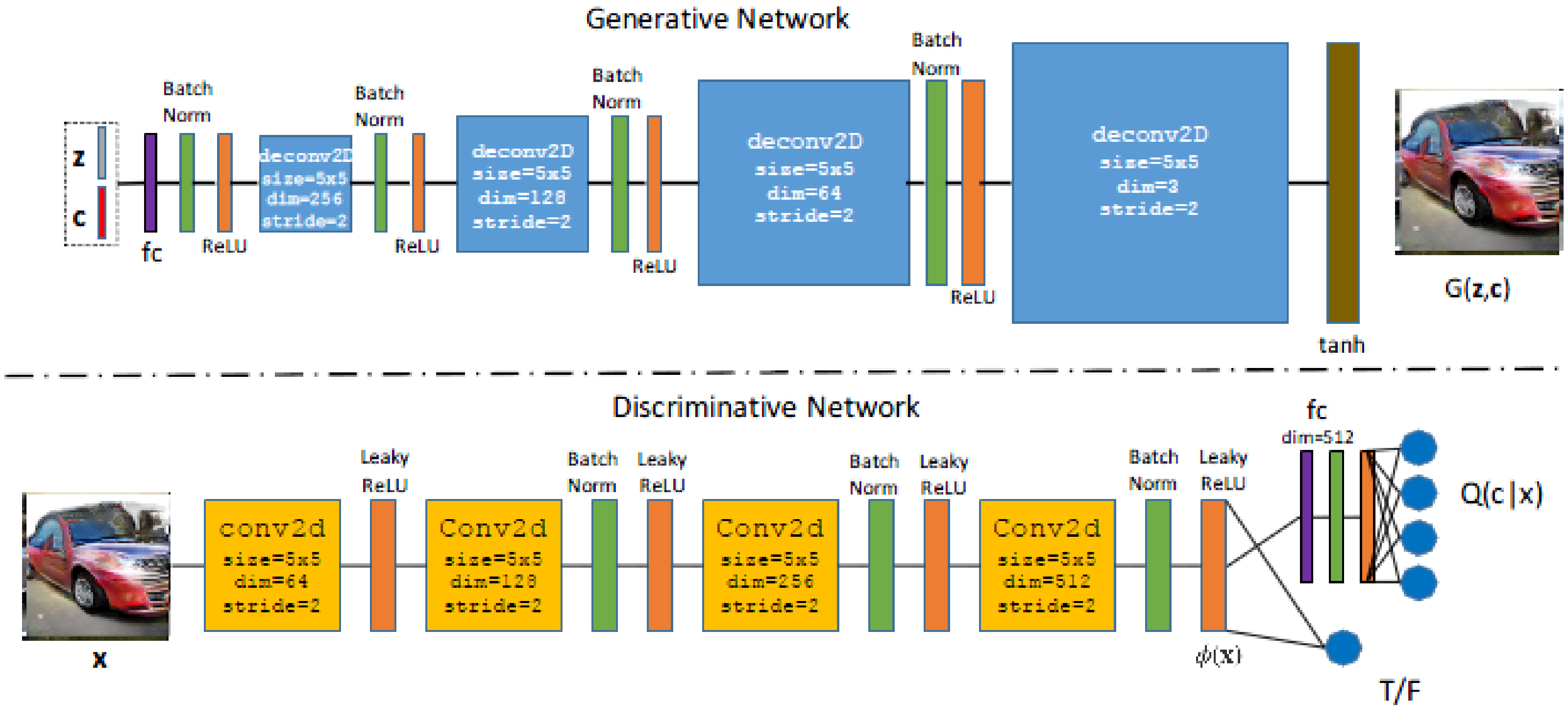

Generative Network
z
c
fc
Batch Norm
ReLU
deconv2D size=5x5 dim=256 stride=2
Batch Norm
ReLU
deconv2D size=5x5 dim=128 stride=2
Batch Norm
ReLU
deconv2D size=5x5 dim=64 stride=2
Batch Norm
ReLU
deconv2D size=5x5 dim=3 stride=2
tanh
G(z,c)
Discriminative Network
x
conv2d size=5x5 dim=64 stride=2
Leaky ReLU
Conv2d size=5x5 dim=128 stride=2
Batch Norm
Leaky ReLU
Conv2d size=5x5 dim=256 stride=2
Batch Norm
Leaky ReLU
Conv2d size=5x5 dim=512 stride=2
Batch Norm
Leaky ReLU
$\phi(x)$
fc
dim=512
Q(c|x)
T/F

Figure shows the InfoGAN architecture that was used in all our experiments. Notice that the input to G(.) is a combination of z and c. Also notice that most of the parameters are shared between the Q(.) network and the D(.) network, thus improving the computational efficiency.

UNSUPERVISED LEARNING WITH K-MEANS++

- while InfoGAN has the ability to group data into multiple groups automatically, there is no constraint to enforce that the groups need to correspond to the various object-level categories that are present in the dataset.

Therefore, **we employ a hybrid approach to unsupervised clustering**. We first **train the discriminative network** using either the vanilla GAN objective or the InfoGAN objective, until convergence. Upon convergence, we extract features for each image in the training set, from the top of the shared network, labeled as (x) in Fig 1, and do average pooling across the spatial resolution, for each feature channel. We then cluster these features using k-means++ into a discrete set of k categories. We set k to be the number of object classes that are present in the respective dataset. The cluster centers learned by k-means++ clustering act as the templates for the k categories that are present in the dataset.

During testing, we extract the feature representation of the test images by passing them through the discriminative network trained using the generator as an adversary, do average pooling on (x), and compute the distance of the test feature vector to each of the centers learnt by kmeans++ clustering during the training phase. The test image is assigned an index corresponding to the index of the closest center. Our experiments show that clustering on (x) produces better results than directly using the recognition model of InfoGAN. Note that while we use the simple kmeans++ algorithm for clustering, it could be replaced by more sophisticated unsupervised learning algorithms. We do not explore further down this route since the scope of this work is to study the strength of the features learned by adversarial training.

学习要点

1. Variational methods

2. Links between information theory and machine learning

https://www.ece.uic.edu/~devroye/courses/ECE534/project/Xiaokai.pdf

https://arxiv.org/pdf/1501.04309v1.pdf

https://github.com/mtomassoli/papers/blob/master/inftheory.pdf

3. Disentangled representation (latent factors) /interpretable representation