InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Presenter: Ye Tao (yt114)

Introduction

Learning Representation: Recently, learning interpretable representation reflecting meaningful semantic features of data has drawn attention of deep learning community. Good representation is essential to perceptual domains (e.g. computer vision) which demands the ability to learn representation of high-dimensional data, interpret and use the learned representation to perform certain tasks (e.g image reconstruction and transfer learning) [1].

Disentanglement Representation Disentanglement Representation learning aims to learn representation where one-dimension change of input results in one factor of variation of data whereas the rest of factor remains the invariant.

Application in generative mode For example, when we generate new object, it would be idea if we can generate digits with different angle by adjusting one input dimension. However, other features such as the categorical of digits remains untouched.

Learn in unsupervised manner Unsupervised learning is to learn from unlabeled data. The reason why it is useful:

- Labels are cost to obtain by human labour work
- human works might be inconsistent or certain variants are not identifiable to human.

The main challenge of representation learning is the difficulty in establishing a clear objective, or target for training. In this poster, we introduce InfoGAN [2], a generative adversarial network (GAN) model learning disentangled representation in unsupervised manner. The key idea is that InfoGAN maximizes object mutual information between latent code (as part of input to generator) and the output of generated fake data.

Generative Adversarial Network

The typical GAN architecture is shown in Fig.2a

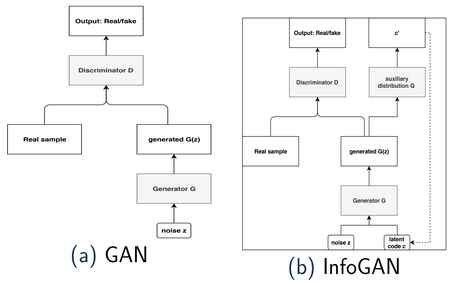
- Discriminator: $D(x)$ a neural network learning the possibility that input x is real or fake (i.e generated by generator)
- Generator: $G(z)$ a neural network learning to generate authentic data from noise sample z.

The parameters of D, G are learning by gradient descent with following **cost functions**

$$\min_G \max_D V(D,G) \tag{1}$$

where

$$V(D,G) = E_{x\sim P_{\text{data}}}[\log D(x)] \tag{2}$$
$$+ E_{z\sim \text{noise}}[\log(1-D(G(z)))]. \tag{3}$$



(a) GAN (b) InfoGAN

Figure: Architerture Diagram: GAN VS InfoGAN

InfoGAN Overview

InfoGAN encodes variants factors of generated data into latent code . It does so by maximize the mutual information beteern latet code and generated data $I(c; G(z, c)$.

The input of infoGAN's generator is:

1. z, which is treated as source of incompressible noise.
2. $c = [\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_l}]$ which are drawn from priors.

The object of infoGAN is

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c)) \tag{4}$$

where $V(D,G)$ is the object function of original GAN

Maximize Mutual Information

In [2] the author propose to estimate and maximum mutual information by variational lower bound. To maximum a lower bound of mutual information, a neural network Q estimates posterior distribution $P(c|x)$.

The family of functions $Q_\theta : \mathcal{X} \times \mathcal{C} \to \mathbb{R}$, where $\mathbf{c}$ is drawn from the prior distribution of latent code $\mathbb{P}_c$ and $\mathbf{x}$ is fake data generated by $G(\mathbf{c}, z)$.

lower bound of mutual information $I_\theta(X; C)$ [2]

$$I(C;X) = H(C) - H(C|X) \tag{5}$$
$$= H(C) + \mathbb{E}_X[\mathbb{E}_{C|x}[\log P(c|x)]] \tag{6}$$
$$\geq H(C) + \mathbb{E}_{c\sim P(c), x\sim G(z,c)}[\log Q(c|x)]] \tag{7}$$
$$= I_\theta(X;C) \tag{8}$$

training object of InfoGAN

$$\min_{G,Q} \max_D V_I(D,G) = V(D,G) - \mathbb{E}[\log Q(c|x)]] \tag{9}$$

Experiment

In this project, I implement the InfoGAN[a] on MNIST which demonstrate its ability to disentangle digit shape. In the experiment, chose one discrete latent code $c_1 \sim \text{Cat(K=10, p=0.1)}$ to control the digits category. Then two continuous latent code $c_2, c_3 \sim U(-1, 1)$.

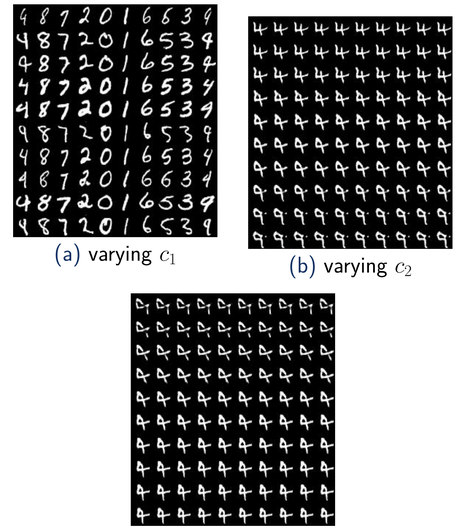
REFERENCES

[1] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[2] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[a] source code: `https://github.com/yt114/Infomation_Theory_project_InfoGAN`

Experiment Result



(a) varying c_1 (b) varying c_2 (c) varying c_3

Figure: (a)Varying discrete latent code: each column digits are generated with same discrete code.Varying first dimension of continuous latent code (sharpness): each row are with same continuous code (dimension 1). Varying second dimension of continuous latent code (rotation): each row are with same continuous code (dimension 2)

Conclusion

In this project, I present the InfoGAN which can learn disentangled representations by maximizing the mutual information (MI) between the generated samples and the latent code. Moreover, I demonstrate its performance on small dataset MNSIT. Through the experiment, we notice that the proposed architecture indeed generates data with only one factor varying. This demonstrates the possibility of using mutual information as a metrice to learning representation. However, it is reported that InfoGAN fails to capture representation in more complex dataset. One possibility is to use other variations approximation of MI which might outperforms the one in InfoGAN such as [1].