1 Title:

InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

2 Summary:

This paper extends GANs to also disentangle factors of variation by splitting the noise term into a latent code c and incompressible noise z, and maximizing mutual information between the generated image and the latent code coupled with regular GAN training framework. Since the posterior on c is unknown, a variational lower bound on mutual information is proposed and maximized through a variational distribution predicted by neural network from generated images. Qualitative results on MNIST, Celeba and 3D Chairs datasets show reasonable disentanglement and the categorical latent code for MNIST is shown to achieve $<$5% error despite using no labels.

3 Strengths:

i) Disentanglement is achieved in an unsupervised way and hence very generalizable. ii) The discriminator and variational distribution predictor share the network weights, resulting in minimal overhead over typical GAN. iii) Flexibility on choosing the prior for latent code allows to incorporate knowledge about the data in an easy way, especially categorical data, without explicitly labelling each and every image. iv) Theoretically grounded approach.

4 Weaknesses:

i) Requires Monte Carlo sampling to get the mutual information loss and hence requires multiple forward passes ii) If intrinsic property of data is unknown, such as in genetic sequences, it's difficult to come up with reasonable priors for latent code. For example, the generative distribution might depend on a multi-variate distribution which is unknown apriori and InfoGAN doesn't figure out such relationships automatically. iii) Problems with GAN training translate. iv) The disentanglement can become category specific as in Figure 4., if the variations in the dataset are large.

5 Analysis of Experiments:

i) Figure 1. verifies the hypothesis that the mutual information is actually maximized. ii) Figure 2. and Section 7.2. suggest that with carefully designed priors to reflect dataset, good disentanglement is achieved (On MNIST, $<$5% error rate from categorical latent code along with thickness and width disentanglement is shown) iii) Figure 5,6 further verify the disentanglement hypothesis by disentangling pose, lighting and width on 3D faces dataset and hairstyles, pose, emotion on Celeba dataset. iv) Figure 4. shows that the learnt pose mapping is different for different 3D shapes. This appears to be a failure case of InfoGAN.

6 Possible Extensions:

Domain transfer can be explored by learning two different domains from the same latent code and applying inductive bias through parameter sharing. Once trained, the same latent code represents two similar looking images belonging to respective domains. The variational approximator that predicts latent code from images can now be used for domain transfer. Note that domain adaptation can be achieved in a similar fashion by additionally learning a down-stream task from the source domain images (domain mapping).