InfoGAN: Information-Theoretic Generative Adversarial Networks

Behrooz Azarkhalili[1]

[1]Life Language Processing Lab, University of California, Berkeley
[1]azarkhalili@behrooz.tech

Introduction

InfoGAN is an extension of the original Generative Adversarial Networks (GANs) that improves interpretability and control over the generated data. It achieves this by maximizing the mutual information between a subset of the latent variables and the generated samples. This allows InfoGAN to discover disentangled representations in an unsupervised manner.

Architecture

InfoGAN consists of three main components:

- **Generator** $G(z, c)$: Generates samples from random noise z and structured latent code c.
- **Discriminator** $D(x)$: Distinguishes between real and fake samples.
- **Q Network** $Q(c|x)$: Approximates the posterior distribution of latent code c given the generated data x.

Mutual Information

The key concept in InfoGAN is maximizing the mutual information $I(c; x)$ between the structured latent code c and the generated data x. The mutual information is defined as:

$$I(c; x) = H(c) - H(c|x)$$

where $H(c)$ is the entropy of c, and $H(c|x)$ is the conditional entropy of c given x.

Objective Function

The total objective function of InfoGAN is composed of two parts:

1. **Adversarial Loss**: Similar to the original GAN's adversarial loss, where the discriminator tries to distinguish real from fake samples, and the generator tries to fool the discriminator:
$$\min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p(z),c\sim p(c)}[\log(1 - D(G(z, c)))]$$
2. **Mutual Information Loss**: To maximize the mutual information between c and $G(z, c)$, InfoGAN adds a term that maximizes $\mathbb{E}_{x\sim G(z,c)}[\log Q(c|x)]$:
$$\min_Q \mathbb{E}_{x\sim G(z,c)}[\log Q(c|x)]$$

Preprint. Under review.

Thus, the overall objective of InfoGAN is:

$$\min_G \max_D V(D, G) - \lambda I(c; G(z, c))$$

where λ controls the tradeoff between the adversarial loss and the mutual information regularization.

Training Procedure

The training procedure for InfoGAN is as follows:

1. Sample noise z from a prior distribution and latent codes c from a structured distribution.
2. Generate samples $G(z, c)$ using the generator.
3. Train the discriminator D to distinguish between real and fake samples.
4. Train the generator G to fool the discriminator while maximizing mutual information between c and $G(z, c)$.
5. Update Q to maximize the mutual information between c and the generated data.

Advantages of InfoGAN

- **Interpretability**: InfoGAN learns interpretable representations by maximizing mutual information.
- **Unsupervised Learning**: No labels are needed to discover interpretable features.
- **Disentangled Representations**: The latent variables c control specific interpretable aspects of the generated data.

Disadvantages of InfoGAN

- **Training Complexity**: InfoGAN introduces an additional network Q and a mutual information objective, making the training process more complex.
- **Sample Quality**: There may be a tradeoff between the interpretability of the generated data and the overall quality of the samples.