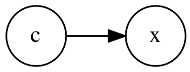
1 InfoGAN

InfoGAN extends the GAN objective to include a new term which encourages high mutual information between generated data and a subset of latent codes [1]. Let (c, z) be latent variable, where c are latent codes capturing semantic features of the data distribution and z are source of incompressible noise.

1.1 Probabilistic Interpretation



A simpler view of method presented in the paper is to consider the above generative model. The joint density can be factorized as follows

$$p_{\mathsf{c},\mathsf{x}} = p_{\mathsf{c}}(c)p_{\mathsf{x}|\mathsf{c}}(x|c) = \prod_{l=1}^{L} p_{\mathsf{c}_l}(c_l)p_{\mathsf{x}|\mathsf{c}}(x|c)$$

The paper implicitly model $p_{\mathsf{x}|\mathsf{c}}$ by using a combination of 1) a deterministic generator $G : \mathcal{C} \times \mathcal{Z} \to \mathcal{X}$ and 2) a stochastic noise sampler $\mathsf{z} \sim p_{\mathsf{z}}$. In particular, $f : \mathcal{C} \to \mathcal{X}; c \mapsto G(c, z)$ for some $z \sim p_{\mathsf{z}}$ is trained to sample from $p_{\mathsf{x}|\mathsf{c}}(\cdot|c)$ using the adversarial loss [2].

1.2 Variational Maximization of Mutual Information

The paper is motivated to construct latent code in such a way such that when given a sample, we would be quite certain what the latent codes are. In other words, we are interested in the following optimization problem

$$\min_G H(\mathsf{c}|\mathsf{x}) \qquad \text{where} \qquad \mathsf{x} = G(\mathsf{c}, \mathsf{z}) \tag{1}$$

If we know the parametric family of distribution c is in, this is equivalent to maximizing mutual information between latent codes and generated sample. Given $H(\mathsf{c}|\mathsf{x}) = H(\mathsf{c}) - I(\mathsf{c}; \mathsf{x})$, we can rewrite (1) as

$$\max_G I(\mathsf{c}; \mathsf{x}) = \mathbb{E}_{\mathsf{c},\mathsf{x}} \left[\log \frac{p_{\mathsf{c},\mathsf{x}}(c, x)}{p_{\mathsf{c}}(c)p_{\mathsf{x}}(x)} \right]$$

which is intractable, since we do not know the implicit likelihood $p_{\mathsf{x}|\mathsf{c}}$ nor the posterior $p_{\mathsf{c}|\mathsf{x}}$. Instead we approximate $p_{\mathsf{c}|\mathsf{x}}$ with using $q_{\mathsf{c}|\mathsf{x}}$, parameterize by a neural network, and derive a lower bound for the objective [3, 4],

$$\begin{aligned}
I(\mathsf{c}; \mathsf{x}) &= H(\mathsf{c}) - H(\mathsf{c}|\mathsf{x}) \\
&= \sum_x p_{\mathsf{x}}(x) \sum_c p_{\mathsf{c}|\mathsf{x}}(c|x) \log p_{\mathsf{c}|\mathsf{x}}(c|x) + H(\mathsf{c}) \\
&= \sum_x p_{\mathsf{x}}(x) \sum_c p_{\mathsf{c}|\mathsf{x}}(c|x) \log \frac{p_{\mathsf{c}|\mathsf{x}}(c|x)}{q_{\mathsf{c}|\mathsf{x}}(c|x)} + \sum_x p_{\mathsf{x}}(x) \sum_c p_{\mathsf{c}|\mathsf{x}}(c|x) \log q_{\mathsf{c}|\mathsf{x}}(c|x) + H(\mathsf{c}) \\
&= \mathbb{E}_{\mathsf{x}} \left[KL(p_{\mathsf{c}|\mathsf{x}}(c|x) || q_{\mathsf{c}|\mathsf{x}}(c|x)) \right] + \mathbb{E}_{\mathsf{c},\mathsf{x}} \left[\log q_{\mathsf{c}|\mathsf{x}}(c|x) \right] + H(\mathsf{c}) \\
&\geq \mathbb{E}_{\mathsf{c},\mathsf{x}} \left[\log q_{\mathsf{c}|\mathsf{x}}(c|x) \right] + H(\mathsf{c}) && (KL \geq 0) \\
&= \mathbb{E}_{\mathsf{c},\mathsf{z}} \left[\log q_{\mathsf{c}|\mathsf{x}}(c|G(c, z)) \right] + H(\mathsf{c})
\end{aligned}$$

1.3 Gradient Estimator

This lower bound can be optimized using stochastic gradient via Monte Carlo estimation,

$$\begin{aligned}
\nabla_\theta I(\mathsf{c};\mathsf{x}) &= \nabla_\theta \mathbb{E}_{\mathsf{c},\mathsf{x}}\left[\log q_{\mathsf{c}|\mathsf{x}}(c|x)\right] \\
&= \mathbb{E}_{\mathsf{c},\mathsf{z}}\left[\nabla_\theta \log q_{\mathsf{c}|\mathsf{x}}(c|G(c,z))\right] \\
&\approx \sum_{i=1}^{N} \nabla_\theta \log q_{\mathsf{c}|\mathsf{x}}(c^{(i)}|G(c^{(i)},z^{(i)})) \\
&\qquad \text{where} \quad c^{(i)} \sim p_{\mathsf{c}} \quad z^{(i)} \sim p_{\mathsf{z}} \qquad i = 1, \cdots, N
\end{aligned}$$

We could also interpret the idea of randomizing the generator using a noise sampler as performing the reparameterization trick [5]. We avoid taking gradient of expectation with respect to $p_{\mathsf{x}|\mathsf{c}}$; Instead, we take sample from a known distribution $z \sim p_{\mathsf{z}}$ and then compute the desired sample $x = G(c, z)$ via a deterministic function.

1.4 Optimization

Note $q_{\mathsf{c}|\mathsf{x}}$ is parameterized by a neural netowrk Q. Given the loss

$$\begin{aligned}
\mathcal{L}_{GAN}(D,G) &= \mathbb{E}_{x\sim p_{data}}\left[\log D(x)\right] + \mathbb{E}_{z\sim p_{niose}}\left[\log(1-D(G(z)))\right] \\
\mathcal{L}_{MI}(Q,G) &= \mathbb{E}_{\mathsf{c}\sim(c),\mathsf{x}\sim}
\end{aligned}$$

References

[1] Xi Chen et al. "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets". In: *arXiv:1606.03657 [cs, stat]* (June 11, 2016). arXiv: 1606.03657. URL: http://arxiv.org/abs/1606.03657 (visited on 12/04/2019).

[2] Ian J. Goodfellow et al. "Generative Adversarial Networks". In: *arXiv:1406.2661 [cs, stat]* (June 10, 2014). arXiv: 1406.2661. URL: http://arxiv.org/abs/1406.2661 (visited on 12/12/2019).

[3] David Barber and Felix Agakov. "The IM Algorithm: A Variational Approach to Information Maximization." In: Jan. 1, 2003.

[4] Ben Poole and Sherjil Ozair. "On Variational Bounds of Mutual Information". In: (2019), p. 10.

[5] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *arXiv:1312.6114 [cs, stat]* (May 1, 2014). arXiv: 1312.6114. URL: http://arxiv.org/abs/1312.6114 (visited on 11/13/2019).