InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

1 Idea

The motivation behind this work is to be able to learn interpretable disentangled representation from the latent space that otherwise would not exhibit these properties. This is achieved by maximizing the mutual information between a subset of the latent variables and the observable (known) variable.

2 Background

Representation learning learns a dense embedding of the entities in our data set which can then be used for downstream tasks. It is an unsupervised form of feature extraction.

2.1 Generative Adversarial Networks

- Typically, generative models i.e. decoders that are capable of producing a reasonably good approximation for the distribution of the source data, are good indicators of a well-learned representation. eg. GANs[1] and VAEs[2].
- This paper builds off the idea of GANs. The min-max game played by the generator and discriminator in GANs is given by the equation

$$min_G max_D V(D,G) = \mathbb{E}_{x\sim P_{data}}[\log D(x)] + \mathbb{E}_{z\sim noise}[\log(1 - D(G(z)))]$$

- The first term is the expected number of times the discriminator successfully identifies the true data distribution and the second term is the expected number of times the discriminator successfully identifies the data generated from noise sampling. From the perspective of the discriminator, ideally $D(\cdot) = 1$ if the data point is from the true distribution and $D(\cdot) = 0$ if the data point is from the sampled distribution.

2.2 Mutual Information

- Mutual information between two variables X and Y is defined as the information learned about one random variable, say Y from the other, say X. This can be expressed as the difference of 2 entropy terms like so:
$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$
- Maximizing the mutual information here, in the context of the first expression would mean that the distribution $P(X|Y)$ is a lot less uncertain than the distribution $P(X)$ and hence can be, on average, expressed in fewer information bits.

3 Method

- In addition to the standard latent variable z the authors introduce a separate latent code c that is meant to be the interpretable part of the latent space.
- The author propose an information theoretic regularization to prevent the explicitly modeled latent code c from being bypassed by the generator $G(z, c)$. They formulate this constraint by stating that the mutual information between the latent code and the generator distribution must be high. i.e. $I(c; G(z, c))$ must be maximized, This would imply that c and $G(z, c)$ would be highly entangled and c would not be lost in the generation process. $I(c; G(z, c))$ can be rewritten such that
$$I(c; G(z, c)) = H(c) - H(c|G(z, c))$$
- Since $H(X) = \mathbb{E}[-\log P(X)]$, we can say that
$$I(c; G(z, c)) = H(c) + \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c,x)}[\log P(c'|x)]] + H(c)$$

- Now, since we don't know the posterior $P(c, x)$, we can use an auxiliary distribution to approximate it and use the KL-divergence between the true and auxiliary distributions to derive a lower bound for the mutual information, which is

$$I(c; G(z, c)) \geq \mathbb{E}_{c \sim P(c), x \sim G(z,c)}[\log Q(c|x)] + H(c)$$

  where $Q(c|x)$ is the distribution that approximates the posterior $P(c'|x)$

- Let $L_I(G, Q)$ be the lower bound of the mutual information $I(c; G(z, c))$. We want to maximize this (or, minimize the negative lower bound, in practice).

- The authors derive the variational information maximization lower-bound to entangle c with $G(z, c)$ and add it to the minimax game of a vanilla GAN, with a hyperparameter λ

$$\min_{G,Q} \max_{D} V_{InfoGAN}(D, G, Q) = V_{GAN}(D, G) - \lambda L_I(G, Q)$$

- The implementation uses the training techniques introduced by DCGAN [3] to stabilize training.

- The experiments use 3 latent factors, c_1 $Cat(K = 10, p = 0.1)$ which is a factored distribution to represent the one-of-ten possible values of the digit in MNIST, and c_2, c_3 $unif(-1, 1)$ which hope to capture other semantics of the digits.

4 Observations

- The representation learning doesn't seem to be completely unsupervised, as we need prior information about the latent code to determine the conditional distribution (e.g. softmax distribution for categorical labels), but this can just be set to a factored Gaussian

- The results seem very convincing as varying each latent code independently causes changes in digit value, rotation and width, when changing c_1, c_2 and c_3 respectively.

- However, it is not evident how the network learns such a neat separation of rotation and width and attributes them independently to c_2 and c_3.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[3] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.