InfoVAE: Information Maximizing Variational Autoencoders

1 Idea

The authors indicate that the variational inference training objectives as defined in the original paper [1] is not expressive enough for a good generative model, but more expressive conditional distributions end up ignoring the latent space altogether. The authors wish to address this by proposing a new training objective for Variational Autoencoders.

2 Background

The variational inference lower bound is derived in the original paper as

$$\begin{aligned} \mathcal{L}_{ELBO} &= -D_{KL}(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[logp_\theta(x|z)] \\ &\leq logp_\theta(x) \end{aligned}$$

where the first term is the KL divergence loss that encourages the inferred latent space to be similar to a prior usually a Gaussian distribution and the second term minimizes the negative log likelihood of observing the data point x given the inferred latent variable z.

3 Method

The authors cite 2 problems with the ELBO objective, 'information preference' and 'exploding latent space':

- **Information Preference** The original ELBO term can be re-written as a sum of 2 divergences.

$$\mathcal{L}_{ELBO} = -D_{KL}(p_{data}(x)||p_\theta(x)) - \mathbb{E}_{p_{data}}[D_{KL}(q_\phi(z|x)||logp_\theta(z|x))]$$

  The first divergence becomes 0 when the reconstruction is perfect, and the second becomes 0 when x and z are independent under p_θ and q_ϕ and no information is gained from the latent code.

- **Exploding latent space** The learned distribution $q_\phi(z|x_i)$ could be a δ-distribution centered at x_i, making this seem optimal for the reconstruction loss. However, this is a case of extreme over-fitting, because a p_θ mapping could be learned for every $q_\phi(z|x_i)$ that could lead to good reconstruction. This is not beneficial, because we want the $q_\phi(z)$ to be almost the same as the prior $p(z)$ and this causes this learning algorithm to learn a bijection, instead of a generalized representation.

3.1 Proposed Solution

Instead of minimizing the previous KL-divergence $-D_{KL}(q_\phi(z|x)||p_\theta(z))$, try to minimize $-D_{KL}(q_\phi(z)||p_\theta(z))$ where

$$q_\phi(z) = \int_x q_\phi(z|x)p_{data}(x)dx$$

Since this cannot be computed directly, we need to use a likelihood-free optimization technique. The InfoVAE objective can thus be written as

$$\mathcal{L}_{InfoVAE} = -\lambda D_{KL}(q_\phi(z)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[logp_\theta(x|z)]$$

for any $\lambda > 0$

3.2 Optimization Techniques

- Adversarial training to minimize the Jensen-Shannon divergence between $q_\phi(z)$ and $p(z)$.
- Stein variational gradient that descends $D_{KL}(q_\phi(z)||p_\theta(z))$
- Maximum mean discrepancy (MMD), computed by comparing all the moments of a distribution. This can be done using the kernel trick.

3.3 Experiments

The authors use 2 strategies to empirically measure the distance between $q_\phi(z)$ and $p(z)$.

- Calculate the MMD statistic over the entire data
- Calculate the log determinant of the covariance matrix of the distribution $q_\phi(z)$. Since When $p(z)$ is the standard Gaussian and $q_\phi(z)$ is trying to emulate the distribution, $log[det(\sum_{q_\phi})]) = 0$

4 Observations

- The improvements suggested are only for the more expressive generators and don't apply to simple conditional distribution families for $p_\theta(x|z)$ like a Gaussian distribution.
- The MMD optimization seems to perform best empirically.

References

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