Deep Information Bottleneck

论文列表

- Diederik P Kingma and Max Welling. **Auto-encoding variational Bayes**. In ICLR, 2014.
- Fischer, I., Dillon, J.V., and Murphy, K. "**Deep Variational Information Bottleneck** [arXiv]." ArXiv (2016): 13 Pp. Web.
- Peng X B , Kanazawa A , Toyer S , et al. **Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow**[J]. 2018.on Bottleneck [arXiv]." ArXiv (2016): 13 Pp. Web.

目录

- 基本概念
  - 互信息
  - ***KL***散度
  - 信息瓶颈
- 深度变分自编码器（varitional auto-encoder）
- 深度变分信息瓶颈（deep varitional information bottleneck）
- 信息瓶颈判别器（varitional discriminator bottleneck）

互信息

- 设两个随机变量(X,Y)的联合分布为$p(x,y)$,边缘分布为$p(x)$,$p(y)$，互信息$\mathrm{I}(X;Y)$是联合分布$p(x,y)$与边缘分布$p(x)p(y)$的相对熵。

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

- 互信息可以看成是一个随机变量中包含的关于另一个随机变量的信息量，或者说是一个随机变量由于已知另一个随机变量而减少的不肯定性。

$$I(X;Y)=H(X)-H(X\mid Y)$$

KL散度与相对熵

- 设$P(x), Q(x)$是随机变量X上的两个概率分布，则在离散和连续随机变量的情形下，相对熵的定义分别为：

$$\mathrm{KL}(P\|Q) = \sum P(x) \log \frac{P(x)}{Q(x)}$$

$$\mathrm{KL}(P\|Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

- 非负性，非对称性

$$\mathrm{KL}(P\|Q) \geq 0$$

$$\mathrm{KL}(P\|Q) \neq \mathrm{KL}(Q\|P)$$

信息瓶颈

- 相关信息
  - 输入信号$x \in X$的相关信息指信号$x \in X$提供的和另一个信号$y \in Y$有关的信息$\bar{x}$。

$$x \rightarrow \bar{x} \qquad \bar{x} \rightarrow y$$

- 一方面，算法应该尽可能压缩原始信息，以互信息来衡量，即$I(x, \bar{x})$应该尽可能小；另一方面，算法应该使得码本保留尽可能多的相关信息，即$I(\bar{x}, y)$应该尽可能大。

$$I(x, \bar{x}) - \beta I(\bar{x}, y)$$

变分自编码器（VAE）

- 。VAE 假设隐向量z从先验 分布$p(z)$中采样得到，输入数据x从分布$p(x|z)$中产生，通过变分方法最大化似然 函数从而实现对输入数据的拟合。似然函数如下：

$$\log p(x_1, \ldots, x_n) = \sum_{i=1}^{n} \log p(x_i)$$

变分自编码器（VAE）

- 单个样本x

$$\begin{aligned}\log p(x) &= \log p(x,z) - \log p(z \mid x) \\ &= \log \frac{p(x,z)}{q(z)} - \log \frac{p(z \mid x)}{q(z)} \\ &= \int q(z) \log \frac{p(x,z)}{q(z)} dz - \int q(z) \log \frac{p(z \mid x)}{q(z)} dz \\ &= L(q,x) + D_{KL}(q(z) \| p(z \mid x))\end{aligned}$$

- 在 VAE 中， $q(z)$被构造为一个包含编码参数φ的后验概率，并服从高斯分布：

$$q(z) = q_\varphi(z \mid x) = N\left(z; \mu_z(x,\varphi), \sigma_z^2(x,\varphi)\right)$$

变分自编码器（VAE）

$$log\ p(x) \geq\ L(q,x)$$

$$\int q(z) \log \frac{p(x,z)}{q(z)} dz$$
$$= E_{q(z)}[\log p(x \mid z)] - D_{KL}(q(z) \| p(z))$$
$$= -D_{KL}(q_\varphi(z \mid x) \| p(z)) + E_{q_\varphi(Z|x)}[\log p(x \mid z)]$$

深度变分信息瓶颈（DIB）

- 将神经网络中间层的输出Z当作输入X的编码，以θ表示神经网络的参数，则这一编码过程可以表示为$p(z|x;\theta)$。以Y表示神经网络的输出，则根据信息瓶颈理论，神经网络的训练过程就可以最大化如下函数

$$R_{IB}(\theta) = I(Z, Y; \theta) - \beta I(Z, X; \theta)$$

- 假设：X,Y,Z的联合概率分布

$$p(X, Y, X) = p(Z \mid X, Y)p(Y \mid X)p(X) = p(Z \mid X)p(Y \mid X)p(X)$$

深度变分信息瓶颈（DIB）

$(1) I(Z, Y; \theta)$

$$I(Z,Y) = \int dydzp(y,z)\log\frac{p(y,z)}{p(y)p(z)} = \int dydzp(y,z)\log\frac{p(y \mid z)}{p(y)}$$

- $p(y|z)$即是神经网络的解码器。通常情况下，$p(y|z)$是难以直接计算的，使用变分方法来估计$I(Z,Y)$的下界。使用$q(y|z)$作为 $p(y|z)$ 的变分估计，则根据 KL 散度的非负性，有

$$\int dyp(y \mid z)\log p(y \mid z) \geq \int dyp(y \mid z)\log q(y \mid z)$$

深度变分信息瓶颈（DIB）

- 得到$I(Z,Y)$的一个下界:

$$
\begin{aligned}
I(Z,Y) &\geq \int dydzp(y,z)\log\frac{q(y\mid z)}{p(y)} \\
&= \int dydzp(y,z)\log q(y\mid z) - \int dyp(y)\log p(y) \\
&= \int dydzp(y,z)\log q(y\mid z) + H(Y)
\end{aligned}
$$

$$
p(y,z) = \int dxp(x,y,z) = \int dxp(x)p(y\mid x)p(z\mid x)
$$

$$
I(Z,Y) \geq \int dxdydzp(x)p(y\mid x)p(z\mid x)\log q(y\mid z)
$$

深度变分信息瓶颈（DIB）

(2) $I(Z, X; \theta)$

$$I(Z, X) = \int dzdxp(x, z) \log \frac{p(z|x)}{p(z)}$$

$$= \int dzdxp(x, z) \log p(z \mid x) - \int dzp(z) \log p(z)$$

- 通常情况下，$p(z)$是比较难以计算的，因此使用$r(z)$作为$p(z)$的变分估计，根据 KL 散度的非负性，有：

$$\int dzp(z) \log p(z) \geq \int dzp(z) \log r(z)$$

- 得到$I(Z,X)$的一个上界：$I(Z, X) \leq \int dxdzp(x)p(z \mid x) \log \frac{p(z \mid x)}{r(z)}$

深度变分信息瓶颈（DIB）

- 结合$I(Z,Y)$的下界和$I(Z,X)$的上界可以得到$RIB(\theta)$的一个下界

$$R_{IB}(\theta) = I(Z,Y;\theta) - \beta I(Z,X;\theta)$$

$$\geq \int dxdydzp(x)p(y \mid x)p(z \mid x)\log q(y \mid z)$$

$$-\beta \int dxdzp(x)p(z \mid x)\log \frac{p(z \mid x)}{r(z)} = L$$

- 在神经网络的训练过程中，输入X和输出Y都是已知的。假设训练数据表示为 $\{X,Y\}$，每个样本表示为$\{x_i,y_i\}$，则X,Y的联合概率密度

$$p(x,y) = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}(x)\delta_{y_i}(y)$$

$$\delta_{x_i}(x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{else} \end{cases}$$

深度变分信息瓶颈（DIB）

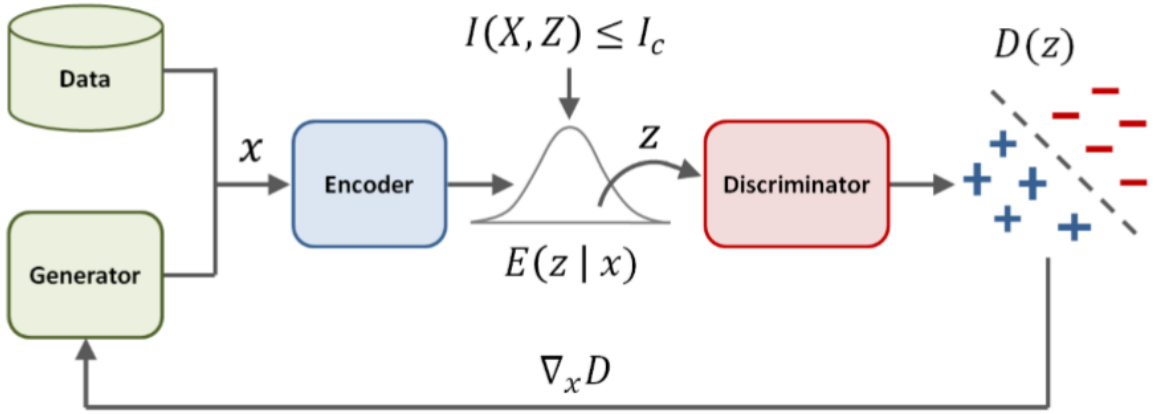
$$L = \frac{1}{N}\sum_{i=1}^{N}\int dz p(z \mid x_i)\log q(y_i \mid z) - \beta p(z \mid x_i)\log\frac{p(z|x_i)}{r(z)}$$

$$= \frac{1}{N}\sum_{i=1}^{N} E_{z\sim p(z|x_i)}\log q(y_i \mid z) - \beta D_{KL}(p(z \mid x_i)||r(z))$$

- 和VAE相比

$$L(q, x, \varphi) = -D_{KL}(q_\varphi(z \mid x)\|p(z)) + E_{q_\varphi(z|x)}[\log p(x \mid z)]$$

信息瓶颈判别器（VDB）



- 带约束的变分判别器

$$\max_G \min_D \quad \mathbb{E}_{\mathbf{x}\sim p^*(\mathbf{x})}[-\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{x}\sim G(\mathbf{x})}[-\log(1-D(\mathbf{x}))]$$

$$J(D,E) = \min_{D,E} \quad \mathbb{E}_{x\sim p^*(\mathbf{x})}\left[\mathbb{E}_{\mathbf{z}\sim E(\mathbf{z}|\mathbf{x})}[-\log(D(\mathbf{z}))]\right] + \mathbb{E}_{\mathbf{x}\sim G(\mathbf{x})}\left[\mathbb{E}_{\mathbf{z}\sim E(\mathbf{z}|\mathbf{x})}[-\log(1-D(\mathbf{z}))]\right]$$

$$\text{s.t.} \quad \mathbb{E}_{\mathbf{x}\sim \tilde{p}(\mathbf{x})}[\mathrm{KL}[E(\mathbf{z}\mid\mathbf{x})\|r(\mathbf{z})]] \leq I_c$$

- 引入拉格朗日算子β,

$$J(D,E) = \min_{D,E}\max_{\beta\geq 0} \quad \mathbb{E}_{\mathbf{x}\sim p^*(\mathbf{x})}\left[\mathbb{E}_{\mathbf{z}\sim E(\mathbf{z}|\mathbf{x})}[-\log(D(\mathbf{z}))]\right] + \mathbb{E}_{\mathbf{x}\sim G(\mathbf{x})}\left[\mathbb{E}_{\mathbf{z}\sim E(\mathbf{z}|\mathbf{x})}[-\log(1-D(\mathbf{z}))]\right]$$

$$+\beta\left(\mathbb{E}_{\mathbf{x}\sim \tilde{p}(\mathbf{x})}[\mathrm{KL}[E(\mathbf{z}\mid\mathbf{x})\|r(\mathbf{z})]] - I_c\right)$$

信息瓶颈判别器（VDB）

- 自适应更新β

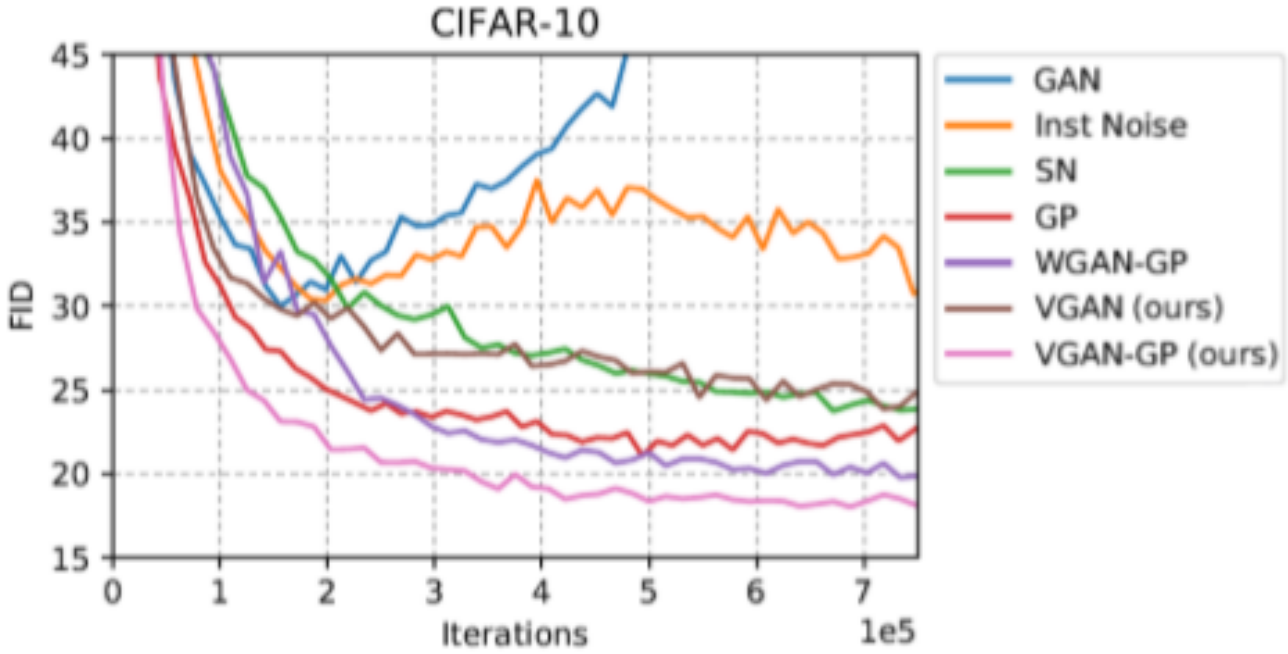
$$D, E \leftarrow \arg\min_{D,E} \mathcal{L}(D, E, \beta)$$

$$\beta \leftarrow \max\left(0, \beta + \alpha_\beta\left(\mathbb{E}_{\mathbf{x}\sim\tilde{p}(\mathbf{x})}[\mathrm{KL}[E(\mathbf{z}\mid\mathbf{x})\|r(\mathbf{z})]] - I_c\right)\right)$$

- 优化生成器

$$\max_G \quad \mathbb{E}_{\mathbf{x}\sim G(\mathbf{x})}[-\log(1 - D(\mu_E(\mathbf{x})))]$$

DVB用于图像生成（VGAN）



Method	FID
GAN	63.6
Inst Noise	30.7
SN	23.9
GP	22.6
WGAN-GP	19.9
VGAN (ours)	24.8
VGAN-SN (ours)	71.8
VGAN-GP (ours)	**18.1**

Figure 8: Comparison of VGAN and other methods on CIFAR-10, with performance evaluated using the Fréchet Inception Distance (FID).

DVB用于图像生成（VGAN）

Figure 9: VGAN samples on CIFAR-10, CelebA 128×128, and CelebAHQ 1024×1024.

稳定性分析

Method	Local convergence (a.c. case)	Local convergence (general case)
unregularized (Goodfellow et al., 2014)	✓	✗
WGAN (Arjovsky et al., 2017)	✗	✗
WGAN-GP (Gulrajani et al., 2017)	✗	✗
DRAGAN (Kodali et al., 2017)	✓	✗
Instance noise (Sønderby et al., 2016)	✓	✓
ConOpt (Mescheder et al., 2017)	✓	✓
Gradient penalties (Roth et al., 2017)	✓	✓
Gradient penalty on real data only	✓	✓
Gradient penalty on fake data only	✓	✓

Table 1. Convergence properties of different GAN training algorithms for general GAN-architectures. Here, we distinguish between the case where both the data and generator distributions are absolutely continuous (a.c.) and the general case where they may lie on lower dimensional manifolds.

Mescheder L , Geiger A , Nowozin S . Which Training Methods for GANs do actually Converge?[J]. 2018.