Information Bottleneck

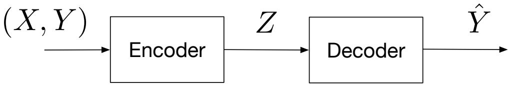
Youlong Wu

ShanghaiTech University

wuyl1@shanghaitech.edu.cn

- Vanilla Information Bottleneck
- Supervised Variational Information Bottleneck
- Unsupervised Variational Information Bottleneck v

Vanilla Information Bottleneck[1]

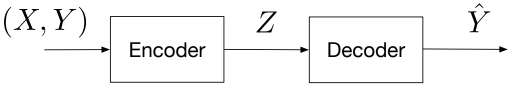


- Source input X, Label Y
- Extracted Feature Z; Estimated label $\hat{Y}$
- Information complexity: $I(X; Z)$, information utility $I(Z; Y)$

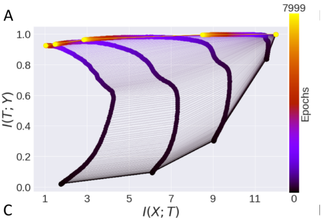
$$\boxed{L = \min_{p(z|x)} I(X; Z) - \beta I(Z; Y)}$$

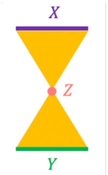
[1] N. Tishby, F.C. Pereira, and W. Biale. The information bottleneck method. In The 37th annual Allerton Conf. on Communication, Control, and Computing. pp. 368–377, 1999

Usefulness of Information Bottleneck



- Explain the learning process of DNN (T: layer's output of DNN)
- Guide the feature extraction of learning process





Challenges: 1) How to calculate/estimate mutual information; 2) Optimal stochastic information $p(x, z)$ are hard to unknown

Deep Variational Information Bottleneck[2]

$$\boxed{L = \min_{p(z|x)} I(X;Z) - \beta I(Z;Y)}$$

- $I(Y;Z)$ is a function of $p(z,y)$, $p(z,y)$ is hard to know
- Let decoder $q(y|z)$ be a variational approximation to $p(y|z)$

$$\begin{aligned}
I(Z;Y) &= H(Y) + \int p(y,z)\log p(y|z)dydz \\
&= H(Y) + \int p(y,z)\log \frac{p(y|z)}{q(y|z)}dydz + \int p(y,z)\log q(y|z)dydz \\
&\geq H(Y) + \int p(y,z)\log q(y|z)dydz = H(Y) + E_{p(y,z)}\log q(y|z) \\
&= H(Y) + \int p(x,y)p(z|x)\log q(y|z)dxdydz
\end{aligned}$$

where the inequality holds by $\mathrm{KL}(p||q) \geq 0$. Note $H(Y) = const$.

[2] Alemi, Alexander A. et al. Deep Variational Information Bottleneck. ICLR, 2017

Deep Variational Information Bottleneck

$$\boxed{L = \min_{p(z|x)} I(X;Z) - \beta I(Z;Y)}$$

- $I(X;Z)$ is a function of $p(z,x)$, where computing $p(z)$ is difficult
- Let $r(z)$ be be a variational approximation to $p(z)$

Since KL$(p(z)||r(z)) = \int p(z) \log p(z)/r(z) dz \geq 0$, we have

$$\int p(z) \log p(z) dz \geq \int p(z) \log r(z) dz \tag{1}$$

Thus,

$$\begin{aligned}
I(Z;X) &= H(Z) - H(Z|X) \\
&= -\int p(z) \log p(z) dz + \int p(x,z) \log p(z|x) dx dz \\
&\leq -\int p(z) \log r(z) dz + \int p(x,z) \log p(z|x) dx dz \\
&= \int p(x,z) \log \frac{p(z|x)}{r(z)} dx dz
\end{aligned}$$

Deep Variational Information Bottleneck

$$\boxed{L = \min_{p(z|x)} I(X;Z) - \beta I(Z;Y)}$$

Rewrite: Need $p(x,y), r(z), p(z|x), q(y|z)$ to compute

$$I(Z;Y) \geq H(Y) + \int p(x,y)p(z|x)\log q(y|z)dxdydz$$

$$I(Z;X) \leq \int p(x,z)\log\frac{p(z|x)}{r(z)}dydz$$

Use Monte Carlo sampling: $p(x,y) = \frac{1}{N}\sum_{n=1}^{N}\delta_{x_n}(x)\delta_{y_n}(y)$, we have

$$\min_{p(z|x)} I(X;Z) - \beta I(Z;Y) \leq \min_{p(z|x)} \int p(x,z)\log\frac{p(z|x)}{r(z)}dxdz - \beta E_{p(y,z)}\log q(y|z)$$

$$\approx \frac{1}{N}\sum_{n=1}^{N}\Big[\int p(z|x_n)\log\frac{p(z|x_n)}{r(z)}dz - \beta\int p(z|x_n)\log q(y_n|z)dz\Big]$$

Question: How to compute the integration over RV Z?

The Reparameterization Trick[3]: Computing $E_{p(z|x)}[f(Z)]$

- Let a deterministic mapping $z = g_\phi(x, \epsilon)$
- ϵ is an RV with independent marginal $p(\epsilon)$
- The noise ϵ is independent of f, so it is easy to take gradients
- g_ϕ is some vector-value function parameterzied by ϕ
- we have

$$g_\phi(z|x)dz = p(\epsilon)d\epsilon$$

$$\int g_\phi(x,\epsilon)f(z)dz = \int p(\epsilon)f(z)d\epsilon = \int p(\epsilon)f(g_\phi(\epsilon, x))d\epsilon$$

$$\int g_\phi(x,\epsilon)f(z)dz = \frac{1}{N}\sum_{n=1}^{N} f(g_\phi(\epsilon_n, x)), \text{ where } \epsilon_n \sim p(\epsilon)$$

- E.g., Let $z = (\mu + \sigma\epsilon) \sim p(z|x) = \mathcal{N}(\mu, \sigma^2)$, where $\epsilon \sim \mathcal{N}(0, 1)$. Then

$$E_{\mathcal{N}(z;\mu,\sigma^2)}[f(Z)] = E_{\mathcal{N}(z;0,1)}[f(\mu + \sigma\epsilon)] \approx \frac{1}{N}\sum_{n=1}^{N} f(\mu + \sigma\epsilon_n)$$

[3] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In ICLR, 2014.

The Reparameterization Trick for Deep IB

Rewrite Deep IB:

$$L = \frac{1}{N}\sum_{n=1}^{N} E_{p(z|x_n)}\Big[\log\frac{p(z|x_n)}{r(z)} - \beta\log q(y_n|z)\Big]$$

- Assume $p(z|x) = \mathcal{N}(z|f_e^{\mu}(x, f_e^{\Sigma}(x)))$
- f_e: MLP encoder that outputs J-dimensional mean μ and variance matrix $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_J^2)$ for z.
- Let $p(z|x)dz = p(\epsilon)d\epsilon$ where $z = \mu + \Sigma\epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
- Let $r(Z) = \mathcal{N}(0, \mathbf{I})$ ($r(z)$ is the variational approximation to $p(z)$)

$$\begin{aligned} L &= \frac{1}{N}\sum_{n=1}^{N} E_{p(z|x_n)}\Big[\log\frac{p(Z|x_n)}{r(Z)} - \beta\log q(y_n|z\Big] \\ &= \frac{1}{N}\sum_{n=1}^{N}\Big[\sum_{j=1}^{J}(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2) - \beta\log q(y_n|z)\Big] \end{aligned}$$

Deep IB v.s. Variational Autoencoder

$$\boxed{L_{\text{Unsuper-IB}} = \max_{p(z|x)} I(X;Z) - \beta I(Z;i),}$$

Consider unsupervised versions of IB:

$$I(Z,X) - \beta I(Z,i) \leq \int dx\, p(x) \int dz\, p(z|x) \log q(x|z) - \beta \frac{1}{N} \sum_i \mathrm{KL}[p(Z|x_i), r(Z)].$$

- Different from $L = \min_{p(z|x)} I(X;Z) - \beta I(Z;Y)$
- Maximize the mutual information contained in some encoding Z
- Restrict how much information we allow our representation to contain about the identity of each data element in our sample (i)
- The upper bound is exactly the same as VAE

Upper Bound of Unsupervised IB

$$I(Z,X) - \beta I(Z,i) \leq \int dx\, p(x) \int dz\, p(z|x) \log q(x|z) - \beta \frac{1}{N} \sum_i \mathrm{KL}[p(Z|x_i), r(Z)].$$

Proof:

$$\begin{aligned}
I(Z,X) &= \int dx\, dz\, p(x,z) \log \frac{p(x|z)}{p(x)} \\
&= H(x) + \int dz\, p(x) \int dx\, p(x|z) \log p(x|z) \\
&\geq \int dz\, p(x) \int dx\, p(x|z) \log q(x|z) \\
&= \int dx\, p(x) \int dz\, p(x|z) \log q(x|z).
\end{aligned}$$

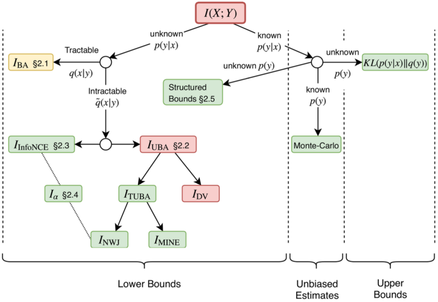
Since $p(z|i) = \int p(z|x)p(x|i) = \int p(z|x)\delta(x - x_i) = p(z|x_i)$ with $p(i) = \frac{1}{N}$.

$$\begin{aligned}
I(Z,i) &= \sum_i \int dz\, p(z|i)p(i) \log \frac{p(z|i)}{p(z)} \\
&= \frac{1}{N} \sum_i \int dz\, p(z|x_i) \log \frac{p(z|x_i)}{p(z)} \\
&\leq \frac{1}{N} \sum_i \int dz\, p(z|x_i) \log \frac{p(z|x_i)}{r(z)},
\end{aligned}$$

Applications of Deep IB

- Joint source-channel coding + IB
- Reinforcement Learning + IB
- Multimodal Data Learning + IB
- Graph Nural Leanring + IB
- etc.

On Variational Bounds of Mutual Information



- Upper bound

$$
\begin{aligned}
I(X;Y) &\equiv \mathbb{E}_{p(x,y)}\left[\log \frac{p(y|x)}{p(y)}\right] \\
&= \mathbb{E}_{p(x,y)}\left[\log \frac{p(y|x)q(y)}{q(y)p(y)}\right] \\
&= \mathbb{E}_{p(x,y)}\left[\log \frac{p(y|x)}{q(y)}\right] - KL(p(y)\|q(y)) \\
&\leq \mathbb{E}_{p(x)}\left[KL(p(y|x)\|q(y))\right] \triangleq R, \qquad (1)
\end{aligned}
$$

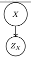
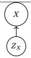
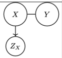
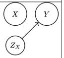
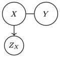
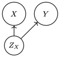
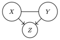
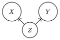
On Variational Bounds of Mutual Information

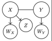
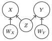
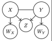
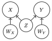
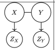
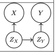
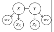
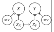
- Lower bound: $p(y|x)$ is unknown and its approximation $q(x|y)$ is tractable

$$
\begin{aligned}
I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{q(x|y)}{p(x)}\right] \\
+ \mathbb{E}_{p(y)}\left[KL(p(x|y)||q(x|y))\right] \\
\geq \mathbb{E}_{p(x,y)}\left[\log q(x|y)\right] + h(X) \triangleq I_{\text{BA}},
\end{aligned}
$$

Deep Variational Multivariate Information Bottleneck

- Applications in different methods[4].

Method Description	G_{encoder}	G_{decoder}
beta-VAE (Kingma & Welling, 2014; Higgins et al., 2016): Two independent Variational Autoencoder (VAE) models trained, one for each view, X and Y (only X graphs/loss shown). $L_{\text{VAE}} = \tilde{I}^E(X;Z_X) - \beta \tilde{I}^D(X;Z_X)$		
DVIB (Alemi et al., 2017): Two bottleneck models trained, one for each view, X and Y, using the other view as the supervising signal. (Only X graphs/loss shown). $L_{\text{DVIB}} = \tilde{I}^E(X;Z_X) - \beta \tilde{I}^D(Y;Z_X)$		
beta-DVCCA: Similar to DVIB (Alemi et al., 2017), but with reconstruction of both views. Two models trained, compressing either X or Y, while reconstructing both X and Y. (Only X graphs/loss shown). $L_{\text{DVCCA}} = \tilde{I}^E(X;Z_X) - \beta(\tilde{I}^D(Y;Z_X) + \tilde{I}^D(X;Z_X))$ **DVCCA** (Wang et al., 2016): a special case of β-DVCCA with $\beta = 1$.		
beta-joint-DVCCA: A single model trained using a concatenated variable $[X,Y]$, learning one latent representation Z. $L_{\text{jDVCCA}} = \tilde{I}^E((X,Y);Z) - \beta(\tilde{I}^D(Y;Z) + \tilde{I}^D(X;Z))$ **joint-DVCCA** (Wang et al., 2016): a special case of β-jDVCCA with $\beta = 1$.		

beta-DVCCA-private: Two models trained, compressing either X or Y, while reconstructing both X and Y, and simultaneously learning private information W_X and W_Y. (Only X graphs/loss shown). $L_{\text{DVCCA-p}} = \tilde{I}^E(X;Z) + \tilde{I}^E(X;W_X) + \tilde{I}^E(Y;W_Y) - \beta(\tilde{I}^D(X;(W_X,Z)) + \tilde{I}^D(Y;(W_Y,Z)))$ **DVCCA-private** (Wang et al., 2016): a special case of β-DVCCA-p with $\beta = 1$.		
beta-joint-DVCCA-private: A single model was trained using a concatenated variable $[X,Y]$, learning one latent representation Z, and simultaneously learning private information W_X and W_Y. $L_{\text{jDVCCA-p}} = \tilde{I}^E((X,Y);Z) + \tilde{I}^E(X;W_X) + \tilde{I}^E(Y;W_Y) - \beta(\tilde{I}^D(X;(W_X,Z)) + \tilde{I}^D(Y;(W_Y,Z)))$ **joint-DVCCA-private** (Wang et al., 2016): β-jDVCCA-p with $\beta = 1$.		
DVSIB: A symmetric model trained, producing Z_X and Z_Y. $L_{\text{DVSIB}} = \tilde{I}^E(X;Z_X) + \tilde{I}^E(Y;Z_Y) -\beta\left(\tilde{I}^D_{\text{MINE}}(Z_X;Z_Y) + \tilde{I}^D(X;Z_X) + \tilde{I}^D(Y;Z_Y)\right)$		
DVSIB-private: A symmetric model trained, producing Z_X and Z_Y, while simultaneously learning private information W_X and W_Y. $L_{\text{DVSIBp}} = \tilde{I}^E(X;W_X) + \tilde{I}^E(X;Z_X) + \tilde{I}^E(Y;Z_Y) + \tilde{I}^E(Y;W_Y) - \beta\left(\tilde{I}^D_{\text{MINE}}(Z_X;Z_Y) + \tilde{I}^D(X;(Z_X,W_X)) + \tilde{I}^D(Y;(Z_Y,W_Y))\right)$		

[4] Abdelaleem, Eslam et al. "Deep Variational Multivariate Information Bottleneck - A Framework for Variational Losses." ArXiv abs/2310.03311 (2023)