On Variational Bounds of Mutual Information

-Summary-

Introduction

Introduction

- Mutual information (MI) : $I(X;Y) = \mathbb{E}_{p(x,y)}[p(x,y)\log\frac{p(x,y)}{p(x)p(y)}] = \mathbb{E}_{p(x,y)}[\log(\frac{p(y|x)}{p(x)})] = \mathbb{E}_{p(x,y)}[\log\left(\frac{p(x|y)}{p(y)}\right)]$
- KL – divergence : $KL(P||Q) = \int_{\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$ when $P \ll Q$
- **Properties on MI required on paper**
  1. $I(X;Y) = KL(P(x,y)||p(x)p(y))$
  2. $I(X;Y) \geq 0$
  3. When $X_1 \rightarrow X_2 \ldots \rightarrow X_n$ forms Markov chain, $I(X_1;X_2,\ldots X_n) = I(X_1;X_2)$
  4. $I(X,Z;Y) = I(X;Y)$ if $p(x,y,z) = p(x,y)p(z)$ (i.e : $Z \perp (X,Y)$)
- MI : measure independence between X, Y (i.e : $I(X;Y) = 0 \; iff \; X \perp Y$)

- **Problem : estimating MI is challenging as we don't have access to underlying distributions, but only the samples.**

Introduction

Introduction

- Usages of MI :
    1. Just estimation of MI
    2. Limit upper bound of MI to restrict the capacity or contents of representations.
    3. Maximize MI between a learned representation and an aspect of the data. (representation learning)

       (Given $x \sim p(x)$, learn a stochastic representation of the data $p_\theta(y|x)$ that maximize MI subject to constraints on the mapping)

Note : To maximize MI, we need to find a lower bound on MI with respect to parameter θ (Variational lower bound of MI), and use Gradient Descent to tighten the lower bound to actual MI.

Variational bounds of MI

Normalized upper bound (what does 'normalized' means?)

- Upper bounding MI is challenging, but is possible when $p(y|x)$ is known

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log\frac{p(y|x)}{p(y)}\right] = \mathbb{E}_{p(x,y)}[\log\frac{p(y|x)}{q(y)}] - KL(p(y)||q(y)) \leq \mathbb{E}_{p(x)}[KL(p(y|x)||q(y)]$$

(Note : drop the $KL(p(y)||q(y)$ term)

where $q(y)$ is variational approximation of $p(y)$, which is intractable.

- $\boldsymbol{R} \triangleq \mathbb{E}_{\boldsymbol{p}(\boldsymbol{x})}[\boldsymbol{KL}(\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x})||\boldsymbol{q}(\boldsymbol{y})]$ is one of variational upper bound of MI

Variational bounds of MI

Normalized lower bounds (I_{BA})

- $I(X;Y) = \mathbb{E}_{p(x,y)}[\log\frac{q(x|y)}{p(x)}]$+$\mathbb{E}_{p(y)}[KL\big(p(x|y)\big|\big|q(x|y)\big)\big] \geq \mathbb{E}_{p(x,y)}[\log(q(x|y)] + h(X)$

  (drop the $\mathbb{E}_{p(y)}[KL\big(p(x|y)\big|\big|q(x|y)\big)\big]$ term)
- $I_{BA} \triangleq \mathbb{E}_{p(x,y)}[\log(q(x|y)] + h(X)$ is a variational lower bound of MI

- Note : $h(X)$ is differential entropy defined by $h(X) = \mathbb{E}_{p(x)}[-\log\big(p(X)\big)\big]$ => constant
- The bound is tight when $q(x|y) = p(x|y)$ and $\mathbb{E}_{p(x,y)}[\log(q(x|y)]$= $-h(X|Y)$
- Note : $h(X|Y) = \mathbb{E}_{p(x,y)}[-\log\big(p(X|Y)\big)\big]$

- Evaluating $h(X)$ is intractable (and often unknown), but gradient of I_{BA} is tractable if $q(x|y)$ [decoder on representation learning] is tractable (challenging when X is high dimensional and $H(X|Y)$ is large)

Variational bounds of MI

Unnormalized lower bounds – backgrounds

- To avoid 'tractable' decoder, we turn to 'unnormalized' distributions for the variational family of $q(x|y)$.
- **Here, we choose energy-based variational family using 'critic' $\mathbf{f}(\mathbf{x}, \mathbf{y})$ and scaled by $\mathbf{p}(\mathbf{x})$:**

$$q(x|y) = \frac{p(x)}{Z(y)} e^{f(x,y)}, \text{where } Z(y) = \mathbb{E}_{p(x)}[e^{f(x,y)}]$$

- Critic acts as loss function (if discrepancy between x, y is big, then gives high critic value $f(x, y)$)

Variational bounds of MI

Unnormalized lower bounds ($\boldsymbol{I_{UBA}}$)

- Recall $I_{BA} = \mathbb{E}_{p(x,y)}[\log(q(x|y)] + h(X)$, HERE, put $q(x|y)$ as a energy based variational family.

  (Again, recall energy based variational family : $q(x|y) = \frac{p(x)}{Z(y)} e^{f(x,y)}$, where $Z(y) = \mathbb{E}_{p(x)}[e^{f(x,y)}]$)

- Then, we get $I(X;Y) \geq \mathbb{E}_{p(x,y)}[f(x,y)] - \mathbb{E}_{p(y)}[logZ(y)] \triangleq I_{UBA}$

- Bound is tight when $f(x,y) = \log p(y|x) + \mathrm{c(y)}$ where $c(y)$ is solely a function of y.

  (actually using condition $q(x|y) = p(x|y)$, we can deduce $c(y) = \log \frac{Z(y)}{p(y)}$)

- Note : By scaling $q(x|y)$ by $p(x)$, we could remove intractable $h(X)$ term.

- **Problem : log partition function $\mathbf{\log Z(y)}$ is intractable.**

Variational bounds of MI

Unnormalized lower bounds (I_{DV})

- To avoid intractable log partition function $\log Z(y)$ in I_{UBA}, We use Jensen's inequality to $\mathbb{E}_{p(y)}[\log Z(y)]$ term on I_{UBA}.
- $\mathbb{E}_{p(y)}[\log Z(y)] \leq \log(\mathbb{E}_{p(y)}\,[Z(y)])$ using concavity of log and Jensen's inequality.
- Then, we get $I(X,Y) \geq I_{UBA} \geq \mathbb{E}_{p(x,y)}[f(x,y)] - \log(\mathbb{E}_{p(y)}\,[Z(y)]) \;\triangleq I_{DV}$
- **Problem : Achieving $\mathbf{I_{DV}}$ is still intractable in practice.**
- One may use the inequality $\log Z(y) = \log \mathbb{E}_{p(x)}[e^{f(x,y)}] \geq \mathbb{E}_{p(x)}\left[\log\left(e^{f(x,y)}\right)\right] = \mathbb{E}_{p(x)}[f(x,y)]$ to upper bound the $\log Z(y)$ term in I_{UBA} AND use MC approximation. but this gives neither an upper and lower bound.

  (Since this becomes upper bound of I_{UBA}, which is an lower bound of MI)

Variational bounds of MI

Unnormalized lower bounds (I_{TUBA})

- To form a tractable bound (especially deal with log partition), we use following inequality

$$\log(x) \leq \frac{x}{a} + \log(a) - 1 \text{ for all } x, a > 0$$

(this get tight when $x = a$)

- Then, $\log Z(y) \leq \frac{Z(y)}{a(y)} + \log\big(a(y)\big) - 1$, for some function $a(y) > 0$ and this get tight when $a(y) = Z(y)$
- Applying this inequality on $I_{UBA} = \mathbb{E}_{p(x,y)}[f(x,y)] - \mathbb{E}_{p(y)}[logZ(y)]$, we get following :

$$\mathbb{E}_{p(x,y)}[f(x,y)] - \mathbb{E}_{p(y)}[\frac{\mathbb{E}_{p(x)}\left[e^{f(x,y)}\right]}{a(y)} + \log\big(a(y)\big) - 1] \triangleq I_{TUBA}$$

To tighten this lower bound, we can maximize this bound w.r.t variational parameters $a(y)$ and f.

Variational bounds of MI

Unnormalized lower bounds (I_{NWJ})

- To simplify the I_{TUBA}, put $a(y) = e$,which yields I_{NWJ} :

$$I_{NWJ} = \mathbb{E}_{p(x,y)}[f(x,y)] - e^{-1}\mathbb{E}_{p(y)}[Z(y)]$$

Note : there exists a unique optimal critic $f^*(x,y) = 1 + \log(\frac{p(x|y)}{p(x)})$ such that $I_{NWJ} = I(X;Y)$.

Note : We can also choose $a(y) = \frac{1}{K}\sum_{i=1}^{K} e^{f(x_i,y_i)}$ (scalar exponential moving average, EMA), where $K =$ minibatch size, then, the gradient of I_{TUBA} yields the 'improved MINE gradient estimator'

Variational bounds of MI

Multi-sample unnormalized lower bounds

- To reduce variance of variational bounds, we extend the unnormalized bounds to depend on multiple samples.
- Goal : estimate $I(X_1;Y)$ given samples from $p(x_1)p(y|x_1)$ and access to $K-1$ additional samples $x_{2:k}\sim r^{K-1}(x_{2:K}) = \Pi_{j=2}^{K} p\left(x_j\right)$.
- Note : We assume $X_1, \dots X_K$ are independent (not sure…or assuming Markov chain) , so using fact : $I(X,Z;Y) = I(X;Y)$ if $Z \perp (X,Y)$, we get $I(X_1;Y) = I(X_1, \dots X_K; Y)$
- Recall : $f^*(x,y) = 1 + \log(\frac{p(x|y)}{p(x)})$ on $I_{NWJ} = \mathbb{E}_{p(x,y)}[f(x,y)] - e^{-1}\mathbb{E}_{p(y)}[Z(y)]$

  Using this, the optimal critic for multi sample case : $f^*(x_{1:k},y) = 1 + \log\left(\frac{p(y|x_{1:k})}{p(y)}\right) = 1 + \log\left(\frac{p(y|x_1)}{p(y)}\right)$

  => Critics now also depends on the additional samples $\boldsymbol{x_{2:K}}$

Variational bounds of MI

Multi-sample unnormalized lower bounds (I_{NWJ})

- By setting, the critic $f(x_{1:k}, y) = 1 + \log \frac{e^{f(x_1,y)}}{a(y;x_{1:K})}$ and $r^{K-1}(x_{2:K}) = \Pi_{j=2}^{K} P(x_j)$, the multi-sample I_{NWJ} becomes (the RHS term) :

$$I(X_1;Y) \geq 1 + \mathbb{E}_{p(x_{1:K})p(y|x_1)}\left[\log \frac{e^{f(x_1,y)}}{a(y;x_{1:K})}\right] - \mathbb{E}_{p(x_{1:K})p(y)}\left[\frac{e^{f(x_1,y)}}{a(y;x_{1:K})}\right]$$

- One way to exploit additional sample $x_{2:K}$ from $p(x)$ is to use MC estimate of the partition function $Z(y)$:

=> Set $a(y;x_{1:K}) = m(y;x_{1:K}) = \frac{1}{K}\left(\sum_{i=1}^{K} e^{f(x_i,y)}\right) \cong Z(y)$

- Then, the last term $\mathbb{E}_{p(x_{1:K})p(y)}\left[\frac{e^{f(x_1,y)}}{m(y;x_{1:K})}\right] = \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_{p(x_{1:K})p(y)}\left[\frac{e^{f(x_i,y)}}{m(y;x_{1:K})}\right] = 1$, Thus, we get I_{NCE} :

$$I(X;Y) \geq \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K} \log \frac{e^{f(x_i,y_i)}}{\frac{1}{K}\sum_{j=1}^{K} e^{f(x_i,y_i)}}\right]$$, where expectation is taken over $\Pi_j p(x_j, y_j)$

Variational bounds of MI

Multi-sample unnormalized lower bounds ($\boldsymbol{I_{NWJ}}, \boldsymbol{I_\alpha}$)

- Note : $I_{NCE} < \log K$ and the optimal critic for I_{NCE} is $f(x, y) = \log p(y|x) + c(y)$

  Therefore, if $I(X;Y) > \log K$, then the lower bound (I_{NWJ}) may be loose.

- Note : I_{NWJ} : low-bias, high variance estimator ⇔ I_{NCE} : high-bias, low-variance estimation.

- To get a continuum between I_{NWJ} and I_{NCE}, set $f(x_{1:k}, y) = 1 + \log \frac{e^{f(x_1,y)}}{\alpha m(y;x_{1:K})+(1-\alpha)q(y)}$ with $\alpha \in [0,1]$, then we get following lower bound I_α :

$$1 + \mathbb{E}_{p(x_{1:K})p(y|x_1)}\left[\log \frac{e^{f(x_1,y)}}{\alpha m(y;x_{1:K}) + (1-\alpha)q(y)}\right] - \mathbb{E}_{p(x_{1:K})p(y)}\left[\frac{e^{f(x_1,y)}}{\alpha m(y;x_{1:K}) + (1-\alpha)q(y)}\right]$$

- Note : $I_{NWJ}(\alpha = 0)$ and $I_{NCE}(\alpha = 1)$ and $I_\alpha < \log \frac{K}{\alpha}$

Variational bounds of MI

Structured bounds with tractable encoders

- When the conditional distribution $p(y|x)$ is known (This case is common in representation learning), we can use previous bounds to find upper bound.

- Recall $R \triangleq \mathbb{E}_{p(x)}[KL(p(y|x)||q(y)]$, which is an upper bound of MI.

  Given a minibatch of K (x_i, y_i) pairs, we can approximate $p(y) \cong \frac{1}{K} \sum_{i=1}^{K} p(y|x_i)$ and $q_i(y) = \frac{1}{K-1} \sum_{j \neq i} p(y|x_j)$.

  Using this, we can upper bound MI by following :

$$I(X;Y) \leq \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\left[\log \frac{p(y_i|x_i)}{\frac{1}{K-1}\sum_{j\neq i} p(y_i|x_i)}\right]\right]$$

  where the expectation is over $\Pi_i p(x_i, y_i)$.

- Using I_{NCE} and this upper bound , we can sandwich MI without introducing learned variational distribution.

Experiments

Comparing estimates across different lower bounds

- Experiment environment :
  - (x, y) are drawn from 20-dim Gaussian distribution with correlation $\rho(t)$, where t = step
  - $\left(x, (Wy)^3\right)$ are also prepared using $W_{ij} \sim N(0,1)$ and cubic exponent done by element-wise.
  - Note : $I(X;Y) = I(X;(WY)^3)$ (full rank linear transformation does not change MI)