

Mutual Information Neural Estimation

Mohamed Ishmael Belghazi¹ Aristide Baratin^{1,2} Sai Rajeswar¹ Sherjil Ozair¹ Yoshua Bengio^{1,3,4}
Aaron Courville^{1,3} R Devon Hjelm^{1,4}

【深度学习 111】MINE



张楚珩

清华大学 交叉信息院博士在读

32 人赞同了该文章

全称为 mutual information neural estimation。

原文传送门

[Belghazi, Mohamed Ishmael, et al. "Mutual information neural estimation." International Conference on Machine Learning. 2018.](#)

特色

这篇文章讲如何用神经网络来估计互信息（mutual information）。信息论在深度学习中应用非常广泛，在强化学习领域，本专栏之前就有讲过基于信息论来无监督学习地（不需要 reward）做探索的方法。这篇文章最核心的思想就是基于互信息的 Donsker-Varadhan 下界，这样使用神经网络来估计互信息就转化为了一个优化问题，可以通过梯度上升算法来实现。

互信息可以看做加强版的 correlation，correlation 只能反映变量之间的线性关系，但是互信息可以进一步反映变量之间的非线性关系；这一点在量化金融里面也非常有用。

本文归纳如下。给定一堆相互关联的数据 x 和 z ，需要给出这两组数据之间互信息的估计。方法就是训练一个神经网络（statistic network） T ，训练目标就是对 Donsker-Varadhan 下界做梯度上升，最后的估计值就是 Donsker-Varadhan 下界在样本上的估计。

当然，仅仅找出两组数据之间的互信息看起来也并不是特别有用。但是很多应用中会把互信息放入目标函数中，可以使用如下的 statistic network 来得到目标函数相对于输入（ x 或者 z ）的导数，从而完成梯度的传递。

Sketch

- Objective: NN to estimate mutual information
 - Definition of mutual information: $I(X, Z) = D_{KL}(\mathbb{P}_{XZ} || \mathbb{P}_X \otimes \mathbb{P}_Z)$
- Approach: Donsker-Varadhan representation

$$T^*(x, z) = \log \frac{\mathbb{P}(x, z)}{\mathbb{Q}(x, z)}$$

$$\mathbb{E}_{\mathbb{P}}[T^*] = I(X, Z)$$

$$\mathbb{E}_{\mathbb{Q}}[e^{T^*}] = 1$$

$$D_{KL}(\mathbb{P} || \mathbb{Q}) = \sup_{T: \Omega \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T])$$
- Training data $\{(x_i, z_i)\}_{i=1}^n$
- Neural network T_{θ}
- Maximize $\mathbb{E}_{\mathbb{P}}[T_{\theta}(xz)] - \log(\mathbb{E}_{\mathbb{Q}}[e^{T_{\theta}(xz)}])$
- Mutual information neural estimator (MINE)

$$I(\bar{X}; \bar{Z})_n = \hat{\mathbb{E}}_{XZ}[T_{\theta^*}] - \log(\hat{\mathbb{E}}_{X \otimes Z}[e^{T_{\theta^*}}])$$

知乎 @张楚珩

过程

1、Donsker-Varadhan Representation

从上一张 slide 中可以看到，互信息可以表示为 KL 散度，而 KL 散度可以写出 Donsker-Varadhan 表示形式；可以看出对于任意的一个 $T: \mathbf{X} \times \mathbf{Z} \rightarrow \mathbb{R}$ 函数，都对应了互信息的一个下界。具体证明方法如下。注意最优的 T 函数的形式。

Proof of Donsker-Varadhan

$$D_{KL}(\mathbb{P} || \mathbb{Q}) = \sup_{T: \Omega \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T]),$$

$$\text{PDF: } \mathbb{G}(x) = \frac{1}{Z} e^{T(x)} \mathbb{Q}(x)$$

Proof. A simple proof goes as follows. For a given function T , consider the Gibbs distribution \mathbb{G} defined by $d\mathbb{G} = \frac{1}{Z} e^T d\mathbb{Q}$, where $Z = \mathbb{E}_{\mathbb{Q}}[e^T]$. By construction,

$$\mathbb{E}_{\mathbb{P}}[T] - \log Z = \mathbb{E}_{\mathbb{P}} \left[\log \frac{d\mathbb{G}}{d\mathbb{Q}} \right] \quad (23)$$

Let Δ be the gap,

$$\Delta := D_{KL}(\mathbb{P} || \mathbb{Q}) - (\mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T])) \quad (24)$$

Using Eqn 23, we can write Δ as a KL-divergence:

$$\Delta = \mathbb{E}_{\mathbb{P}} \left[\log \frac{d\mathbb{P}}{d\mathbb{Q}} - \log \frac{d\mathbb{G}}{d\mathbb{Q}} \right] = \mathbb{E}_{\mathbb{P}} \log \frac{d\mathbb{P}}{d\mathbb{G}} = D_{KL}(\mathbb{P} || \mathbb{G}) \quad (25)$$

The positivity of the KL-divergence gives $\Delta \geq 0$. We have thus shown that for any T ,

$$D_{KL}(\mathbb{P} || \mathbb{Q}) \geq \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T]) \quad (26)$$

and the inequality is preserved upon taking the supremum over the right-hand side. Finally, the identity (25) also shows that this bound is tight whenever $\mathbb{G} = \mathbb{P}$, namely for optimal functions T^* taking the form $T^* = \log \frac{d\mathbb{P}}{d\mathbb{Q}} + C$ for some constant $C \in \mathbb{R}$. \square

Optimal T gives the log ratio between \mathbb{P} and \mathbb{Q} : $T^*(x) = \log \frac{\mathbb{P}(x)}{\mathbb{Q}(x)}$

知乎 @张楚珩

2、算法

算法很直接，就是最大化下界，只不过有一个小问题。如果目标是 $\mathbb{E}[f(\cdot)]$ （比如 $\mathbb{E}[\log(\cdot)]$ ）对应的随机梯度算法都是真实梯度的无偏估计；但是这里的目标是 $f(\mathbb{E}[\cdot])$ ，因此无法得到一个无偏的梯度估计（考虑 Jensen 不等式）。换句话说，无偏的梯度估计是下面红框里面给出的式子；用样本来代替期望的时候，分子分母上的样本需要独立同分布地采样，不能直接用同一个 minibatch 上的样本。文章这里用的方法是对于分母做 EMA，相当于减小分子分母在同一个 minibatch 上的关联性；这种情况下只要 learning rate 足够小，那么就能够恢复无偏估计。

Algorithm

- Practical problem: biased gradient on a minibatch

$$\hat{G}_B = \mathbb{E}_B[\nabla_{\theta} T_{\theta}] - \frac{\mathbb{E}_B[\nabla_{\theta} T_{\theta} e^{T_{\theta}}]}{\mathbb{E}_B[e^{T_{\theta}}]}.$$

- Solution: use the EMA for the denominator and a small learning rate

Algorithm 1 MINE

$\theta \leftarrow$ initialize network parameters

repeat

Draw b minibatch samples from the joint distribution:

$(\mathbf{x}^{(1)}, \mathbf{z}^{(1)}), \dots, (\mathbf{x}^{(b)}, \mathbf{z}^{(b)}) \sim \mathbb{P}_{XZ}$

Draw n samples from the Z marginal distribution:

$\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(b)} \sim \mathbb{P}_Z$

Evaluate the lower-bound:

$V(\theta) \leftarrow \frac{1}{b} \sum_{i=1}^b T_{\theta}(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) - \log(\frac{1}{b} \sum_{i=1}^b e^{T_{\theta}(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})})$

Evaluate bias corrected gradients (e.g., moving average):

$\tilde{G}(\theta) \leftarrow \tilde{\nabla}_{\theta} V(\theta)$

Update the statistics network parameters:

$\theta \leftarrow \theta + \tilde{G}(\theta)$

until convergence

知乎 @张楚珩

理论上文章说明了互信息的这种估计是 consistent 的（即，样本足够多的时候，它是无偏估计）。

Theoretical Properties

- I_{Θ} is the optimal over the function family
- Lemma 1 **proof sketch**: gap
 $\leq \mathbb{E}_{\mathbb{P}}[T^* - T] + \mathbb{E}_{\mathbb{Q}}[e^T - e^{T^*}] \leq$
 $\mathbb{E}_{\mathbb{P}}|T^* - T| + e^M \mathbb{E}_{\mathbb{Q}}|T^* - T|$, then apply
 universal approximation theorem (Hornik, 1989). Choose M s. t. $T^* > M$ is not
 significant under \mathbb{Q} .
- Lemma 2 **proof sketch**: similar process
 with uniform law of large numbers (Van
 de Geer, 2000).

Theorem 2. *MINE is strongly consistent.*

Definition 3.2 (Strong consistency). The estimator $\widehat{I}(\bar{X}; \bar{Z})_n$ is strongly consistent if for all $\epsilon > 0$, there exists a positive integer N and a choice of statistics network such that:

$$\forall n \geq N, \quad |I(X, Z) - \widehat{I}(\bar{X}; \bar{Z})_n| \leq \epsilon, \text{ a.e.}$$

where the probability is over a set of samples.

Lemma 1 (approximation). Let $\epsilon > 0$. There exists a neural network parametrizing functions T_{θ} with parameters θ in some compact domain $\Theta \subset \mathbb{R}^k$, such that

$$|I(X, Z) - I_{\Theta}(X, Z)| \leq \epsilon, \text{ a.e.}$$

Lemma 2 (estimation). Let $\epsilon > 0$. Given a family of neural network functions T_{θ} with parameters θ in some bounded domain $\Theta \subset \mathbb{R}^k$, there exists an $N \in \mathbb{N}$, such that

$$\forall n \geq N, \quad |I(\bar{X}; \bar{Z})_n - I_{\Theta}(X, Z)| \leq \epsilon.$$

知乎 @张楚珩

同时，文章还给出了 sample complexity。不过需要注意的是，这个 sample complexity 只是对于 lemma 2（相当于是做了 asymptotic 的分析）的进一步细化，并不是神经网络训练的 sample complexity。换句话说，这里没有说明需要多少样本能够训练得到一个 epsilon 精度的估计。

Theoretical Properties

- Sample complexity in terms of \mathbb{P}, \mathbb{Q}
 estimation, **not in terms of training**
- A refinement of Lemma 2
- Proof sketch: concentration

Theorem 3. Given any values ϵ, δ of the desired accuracy and confidence parameters, we have,

$$\Pr \left(|\widehat{I}(\bar{X}; \bar{Z})_n - I_{\Theta}(X, Z)| \leq \epsilon \right) \geq 1 - \delta, \quad (14)$$

whenever the number n of samples satisfies

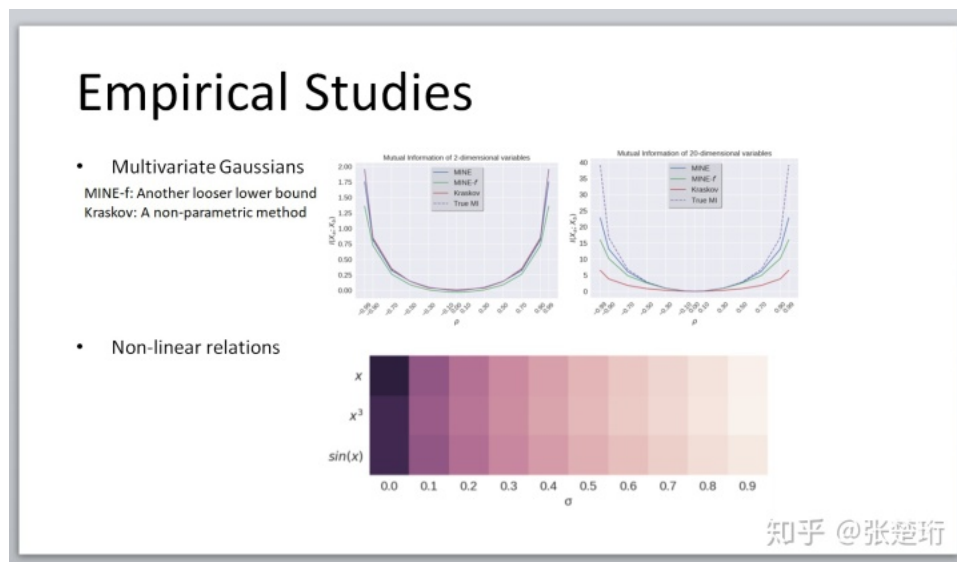
$$n \geq \frac{2M^2(d \log(16KL\sqrt{d}/\epsilon) + 2dM + \log(2/\delta))}{\epsilon^2}.$$

	Distribution	T
$I(X; Z)$	\mathbb{P}, \mathbb{Q}	True T^*
$I_{\Theta}(X; Z)$	\mathbb{P}, \mathbb{Q}	Optimal $T_{\theta^*}, \theta^* \in \Theta$
$I(\bar{X}; \bar{Z})_n$	$\mathbb{P}_n, \mathbb{Q}_n$	Optimal $T_{\theta^*}, \theta^* \in \Theta$

知乎 @张楚珩

3、实验结果

做了两个实验，一个实验说明了这个估计相比 non-parametric 的方法和基于另外一种下界的方法更好；另一个实验说明了互信息不仅能够反映线性的信息，而且也能很好地反映变量之间的非线性信息。



4、应用：GAN

回顾一下，GAN 的目标是给定一组数据 $\mathbf{x} \sim p(\mathbf{x})$ （比如一些照片），训练得到一个生成器 $G: \mathbf{z} \rightarrow \mathbf{x}$

，使得该生成器从一个 $z \sim p(z)$ 出发得到一个和原始数据分布类似的数据（比如机器生成的假图片）。这中间需要联合训练一个分别假图片和真图片的判别器 $D: \mathbf{x} \rightarrow [0, 1]$ 。

GAN 有一个问题是 mode-dropping，简单说来就是，生成器很容易只学习到一种能够有效“欺骗”判别器的模式；这样生成器生成出来的数据（比如图片）看起来很真，但是它只会生成这一种（比如实际图片分布有各种图片，但是生成器最后只会生成包含狗的图片）。这里的做法就是把 z 分为两个部分 $z=[e, c]$ ，然后同时最大化生成图片和 c 之间的互信息。这相当于要求生成的图片尽可能反映 c 的信息（比如给定的 c 如果是某个数字，那么 G 就生成狗；如果是另外的数字， G 就生成猫）。

通过文中的 MINE 方法，可以有效地对于互信息这一项关于 G 求导。实验结果表明，这种方法对应生成出来的数据点更能够有效覆盖整体的原始数据分布。具体实验设定参看原文。

Application: GAN

- MI regularized GAN to alleviate mode-dropping:

$$\max_G \mathbb{E} \left[\log \left(D(G([e, c])) \right) \right] + \beta I(G([e, c]); c)$$

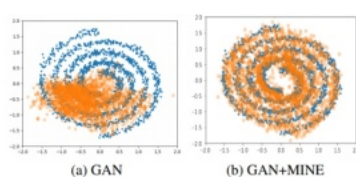


Figure 3. The generator of the GAN model without mutual information maximization after 5000 iterations suffers from mode collapse (has poor coverage of the target dataset) compared to GAN+MINE on the spiral experiment.

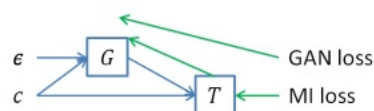
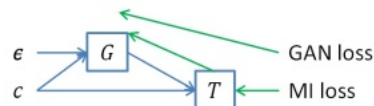


Figure 4. Kernel density estimate (KDE) plots for GAN+MINE samples and GAN samples on 25 Gaussians dataset.

Application: GAN

- MI regularized GAN to alleviate mode-dropping:
- $\max_G \mathbb{E} \left[\log \left(D(G([\epsilon, c])) \right) \right] + \beta I(G([\epsilon, c]); c)$



	Stacked MNIST	
	Modes (Max 1000)	KL
DCGAN	99.0	3.40
ALI	16.0	5.40
Unrolled GAN	48.7	4.32
VEEGAN	150.0	2.95
PacGAN	1000.0 \pm 0.0	0.06 \pm 1.0e ⁻²
GAN+MINE (Ours)	1000.0 \pm 0.0	0.05 \pm 6.9e ⁻³

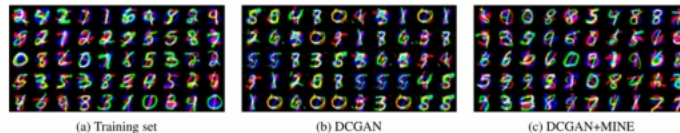


Figure 5. Samples from the Stacked MNIST dataset along with generated samples from DCGAN and DCGAN with MINE. While DCGAN only shows a very limited number of modes, the inclusion of MINE generates a much better representative set of modes.

知乎 @张楚珩

5、应用：Reconstruction (bi-directional adversarial models)

仍然考虑 GAN 类似的网络结构，不过现在加入一个 reverse model F。任务目标是给定一个图片（可能被污染了），重构出来这张图片。一个最明显的目标就是最小化重构误差

$$\mathcal{R} = \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} [-\log p(\mathbf{x}|\mathbf{z})]$$

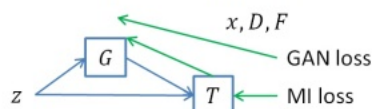
文章通过推到得到该重构误差的上界，发现重构误差的上界包含互信息项。接着仍然使用 MINE 来做优化。

Application: Reconstruction

- Reconstruction: generator $G: z \rightarrow x$, reverse model $F: x \rightarrow z$, discriminator $D: x \rightarrow [0,1]$
- Reconstruct: $G(F(x))$
- Reconstruction error: $\mathcal{R} \leq D_{KL}(q(x, z) || p(x, z)) - I_q(x, z) + H_q(z)$
- True joint distr. $p(x, z) = p(z | x)p(x)$ Reconstructed joint distr. $q(x, z) = q(x | z)p(z)$
- Mutual information regularization

$$\arg \max_D \mathbb{E}_{q(x, z)} [\log D(x, z)] + \mathbb{E}_{p(x, z)} [\log (1 - D(x, z))]$$

$$\arg \max_{F, G} \mathbb{E}_{q(x, z)} [\log (1 - D(x, z))] + \mathbb{E}_{p(x, z)} [\log D(x, z)] + \beta I_q(x, z).$$



知乎 @张楚珩

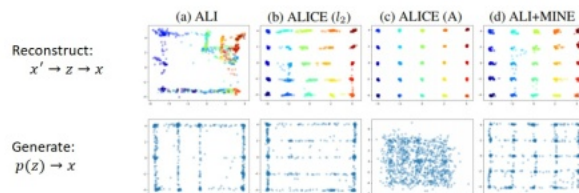
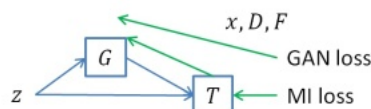
实验效果上，发现虽然在重构精度上没有之前的一个方法 ALICE 好，但是利用训练得到的生成器直接生成的时候，结果看起来比较好。

Application: Reconstruction

- Mutual information regularization

$$\arg \max_D \mathbb{E}_{q(x, z)} [\log D(x, z)] + \mathbb{E}_{p(x, z)} [\log (1 - D(x, z))]$$

$$\arg \max_{F, G} \mathbb{E}_{q(x, z)} [\log (1 - D(x, z))] + \mathbb{E}_{p(x, z)} [\log D(x, z)] + \beta I_q(x, z).$$



Model	Recons. Error	Recons. Acc. (%)	MS-SSIM
MNIST			
ALI	14.24	45.95	0.97
ALICE(l2)	3.20	99.03	0.97
ALICE(Adv.)	5.20	98.17	0.98
MINE	9.73	96.10	0.99
CelebA			
ALI	53.75	57.49	0.81
ALICE(l2)	8.01	32.22	0.93
ALICE(Adv.)	92.56	48.95	0.51
MINE	36.11	76.08	0.99

知乎 @张楚珩

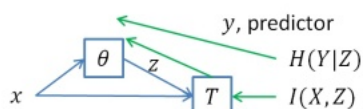
6、应用：Information Bottleneck

Information bottleneck 的目标是找到 x 的一个好的表示，使得这个表示比较 compact 但是也能够很好的预测另外一个变量 y 。经过推导，可以得到相应的目标函数，目标函数包含互信息项目。

注：这里实验中的任务叫做 permutation invariant MNIST，有点费解。其含义是，把 MNIST 数据集做某一种固定的像素点之间的随机交换（相当于变成一个一维数组），然后把它当做输入来做预测。这相当于强制性不准使用像素点之间的位置信息，比普通的 MNIST 任务更难。

Application: Information Bottleneck

- Information bottleneck is to learn a good (compact and predictive) representation of x about $y: x \rightarrow z \rightarrow y$
- Maximize: $R_{IB}(\theta) = I(Z, Y; \theta) - \beta I(Z, X; \theta)$.
- Or minimize: $\mathcal{L}[q(Z | X)] = H(Y|Z) + \beta I(X, Z)$,



Model	Recons. Error	Recons. Acc.(%)	MS-SSIM
MNIST			
ALI	14.24	45.95	0.97
ALICE(l_2)	3.20	99.03	0.97
ALICE(Adv.)	5.20	98.17	0.98
MINE	9.73	96.10	0.99
CelebA			
ALI	53.75	57.49	0.81
ALICE(l_2)	8.01	32.22	0.93
ALICE(Adv.)	92.56	48.95	0.51
MINE	36.11	76.08	0.99

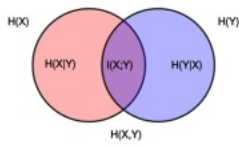
Table 2. Comparison of MINE with other bi-directional adversarial models in terms of euclidean reconstruction error, reconstruction accuracy, and MS-SSIM on the MNIST and CelebA datasets. MINE does a good job compared to ALI in terms of reconstructions. Though the explicit reconstruction based baselines (ALICE) can sometimes do better than MINE in terms of reconstructions related tasks, they consistently lag behind in MS-SSIM scores and reconstruction accuracy on CelebA.

知乎 @张楚珩

补充一下信息论相关量之间的关系

Connections in Information Theory

- Venn diagram
 - Area: amount of information
 - Relationship: additive and subtractive relationship
 - Left circle $H(x)$; right circle $H(y)$; Both $H(x, y)$
- Definition
 - Entropy: $H_p = H(x) = -\sum_x p(x) \log p(x)$
 - Joint entropy: $H(x, y) = -\sum_{x, y} p(x, y) \log p(x, y)$
 - Conditional entropy: $H(x, y) = -\sum_{x, y} p(x, y) \log \frac{p(x, y)}{p(x)}$
 - Mutual information: $I(x; y) = -\sum_{x, y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = D_{KL}(p(x, y) || p(x)p(y))$
 - KL divergence (relative entropy): $D_{KL}(p || q) = -\sum_x p(x) \log \frac{q(x)}{p(x)}$
 - Cross entropy: $H(p, q) = -\sum_x p(x) \log q(x) = D_{KL}(p || q) + H_p$



知乎 @张楚珩

发布于 2020-03-16

机器学习

深度学习 (Deep Learning)

计算机视觉

赞同 32



2 条评论

分享

喜欢

收藏



文章被以下专栏收录



强化学习前沿
读呀读paper

进入专栏