Mutual Information Neural Estimation

Paper by Belghazi et al.
Slides by Cameron Loewen for CSc 696h

Why use MINE

- Versatile: Able to calculate mutual information for any distribution that we can sample from (joint and marginal)
- Scalable: Boasts a (almost) linear scalability with dimensionality and sample size
- Ease of use: Model is a typical neural network

Layout of Presentation

Equations

Mutual Information

Donsker-Varadhan

F-divergence

Neural Information

MINE

Gradient Estimation

Properties

Lemma: Approximation

Lemma: Estimation

Theorem: Strongly Consistent

PAC-Learning bound

Sample Complexity

Application

Empirical Results on estimation

Mutual Information with GAN

Improve inference in adversarial models

Information Bottleneck

Equations: Mutual Information

Mutual information:

$$I(X;Z) = \int_{X \times Z} p(x,z) \log \frac{p(x,z)}{p(x) * p(z)} dxdz$$

- Mutual Information quantifies the dependence of two random variables
- If the random variables are independent, I is zero

Equations: Mutual Information

- Reduction in uncertainty given Z
- Rewritten as KL Divergence

$$I(X;Z) := H(X) - H(X|Z)$$

$$I(X;Z) = D_{KL}(p(x,z)||p(x) * p(z))$$

Equations: Dual Representations

Donsker-Varadhan Representation:

$$D_{KL}(p||q) = \sup_{T:\Omega \to \mathbb{R}} \mathbb{E}_p[T] - \log(\mathbb{E}_q[e^T])$$

- Supremum is taken over all functions T such that the two expectations are finite
- We are not going to worry about how we get to this form, only about how to use it

Equations: Dual Representations

Using a class of functions F with DK:

Original: $$D_{KL}(p||q) = \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_p[T] - \log(\mathbb{E}_q[e^T])$$

Modified: $$D_{KL}(p||q) \geq \sup_{T\in F} \mathbb{E}_p[T] - \log(\mathbb{E}_q[e^T])$$

- T satisfies the integrability constraints of the theorem

Equations: Dual Representations

F-divergence representation:

$$D_{KL}(p||q) = \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_p[T] - (\mathbb{E}_q[e^{T-1}])$$

- Both dual representations are tight under optimal function T
- However, the DK representation is a stronger bound (greater right hand side)

Equations: Dual Representations

Note, Gibbs density:

$$p(x) = \frac{q(x)}{Z} e^{T*}, \text{where } Z = \mathbb{E}_q[e^{T*}]$$

- When the bound is tight with optimal function T*, the marginal can be expressed as above

Equations: Key points

- We have two representations:
  - DK
  - F-divergence
- Both have tight bounds when optimal function T^* can be found
- Next, we will see how using a neural network to approximate T has desirable properties

Properties: Statistics Network

Idea (Statistics Network): choose F to be the family of functions $T_\theta : X \times Z \to R$ parameterized by a deep neural network with parameters $\theta \in \Theta$.

We then get the following bound:

$$I(X;Z) \geq I_\Theta(X,Z)$$

Where the RHS is called the neural information measure (more on this on the next slide)

Properties: Neural Information

The neural information measure is defined as:

$$I_\Theta(X;Z) = \sup_{\theta \in \Theta} \mathbb{E}_{p_{x,z}}[T_\theta] - \log(\mathbb{E}_{p_x p_z}[e^{T_\theta}])$$

Recall the DK representation (and def of mutual info):

$$D_{KL}(p||q) = \sup_{T:\Omega \to \mathbb{R}} \mathbb{E}_p[T] - \log(\mathbb{E}_q[e^T])$$

Properties: Empirical Neural Information

We then sample to estimate the neural information:

$$\hat{I}_{\Theta}(X;Z)_n = \sup_{\theta \in \Theta} \mathbb{E}_{p^n_{x,z}}[T_\theta] - \log(\mathbb{E}_{p^n_x \hat{p}^n_z}[e^{T_\theta}])$$

Neural information definition below for comparison:

$$I_{\Theta}(X;Z) = \sup_{\theta \in \Theta} \mathbb{E}_{p_{x,z}}[T_\theta] - \log(\mathbb{E}_{p_x p_z}[e^{T_\theta}])$$

Properties: Gradient

The following follows from a simple application of stochastic gradient estimation:

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

This is unfortunately a biased gradient
Authors claim: bias can be made arbitrarily low with an exponential moving average + small learning rates

Properties: Algorithm

Algorithm 1 MINE

$\theta \leftarrow$ initialize network parameters

repeat

Draw b minibatch samples from the joint distribution:
$(\boldsymbol{x}^{(1)}, \boldsymbol{z}^{(1)}), \ldots, (\boldsymbol{x}^{(b)}, \boldsymbol{z}^{(b)}) \sim \mathbb{P}_{XZ}$

Draw n samples from the Z marginal distribution:
$\bar{\boldsymbol{z}}^{(1)}, \ldots, \bar{\boldsymbol{z}}^{(b)} \sim \mathbb{P}_Z$

Evaluate the lower-bound:
$\mathcal{V}(\theta) \leftarrow \frac{1}{b}\sum_{i=1}^{b} T_\theta(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) - \log(\frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(\boldsymbol{x}^{(i)}, \bar{\boldsymbol{z}}^{(i)})})$

Evaluate bias corrected gradients (e.g., moving average):
$\widehat{G}(\theta) \leftarrow \widetilde{\nabla}_\theta \mathcal{V}(\theta)$

Update the statistics network parameters:
$\theta \leftarrow \theta + \widehat{G}(\theta)$

until convergence

Properties: Checkpoint

- We defined a lower bound, the neural information measure, in terms of a new class of functions
- We showed that this class can be approximated by a parameterized neural network
- We showed a way to optimize the neural network

Properties: Consistency

An estimator is strongly consistent if for all $\epsilon > 0$, there exists a positive integer N and a choice of statistics network such that:

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

Properties: Lemma 1

Lemma 1: Let $\epsilon > 0$. There exists a neural network parameterized function with parameters θ in some compact domain $\Theta \subset R^k$, s.t.:

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

Lemma 1 is covered by universal approximation theorems for neural networks

Properties: Lemma 2

Lemma 2: Let $\epsilon > 0$. Given a family of neural network functions with parameters θ in some bounded domain $\Theta \subset R^k$, there exists an $n \in N$ s.t.:

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

Lemma 2 is covered by the classical consistency theorems for extremum estimators

Properties: Consequences of Lemmas

Lemma 1 + Lemma 2 = MINE is strongly consistent (proof by triangle inequality)

Lemma 1 $$|I(X,Z) - I_\Theta(X,Z)| \leq \epsilon, \text{ a.e.}$$

Lemma 2 $$\forall n \geq N, |\hat{I}(X;Z)_n - I_\Theta(X;Z)| \leq \epsilon, \text{ a.e.}$$

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

Properties: PAC

PAC, refinement of lemma 2. Following assumptions: the functions T_θ are M-bounded and L-Lipschitz with respect to the parameters. Further, the domain $\Theta \subset R$ is bounded, so that $||\theta|| \leq K$ for some constant K.

$$P(|\hat{I}(X;Z)_n - I_\Theta(X;Z)| \leq \epsilon) \geq 1 - \delta$$

The above holds only when n is greater than the SC

Properties: Analyzing the sample complexity

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

M - The size of the function family
d - the dimensionality of Big-Theta
K - The boundedness of little-theta
L - the Lipschitz constant with respect to little-theta

This is how they get (almost) linear with respect to dimensionality and sample complexity

Results: Overview

- Empirical results, test the theory and accuracy
- GAN (we have seen this with InfoGAN) to help combat against mode dropping
- ALI-Improve inference and reconstruction
- Allows for tractable application of information bottleneck methods

Results: Empirical

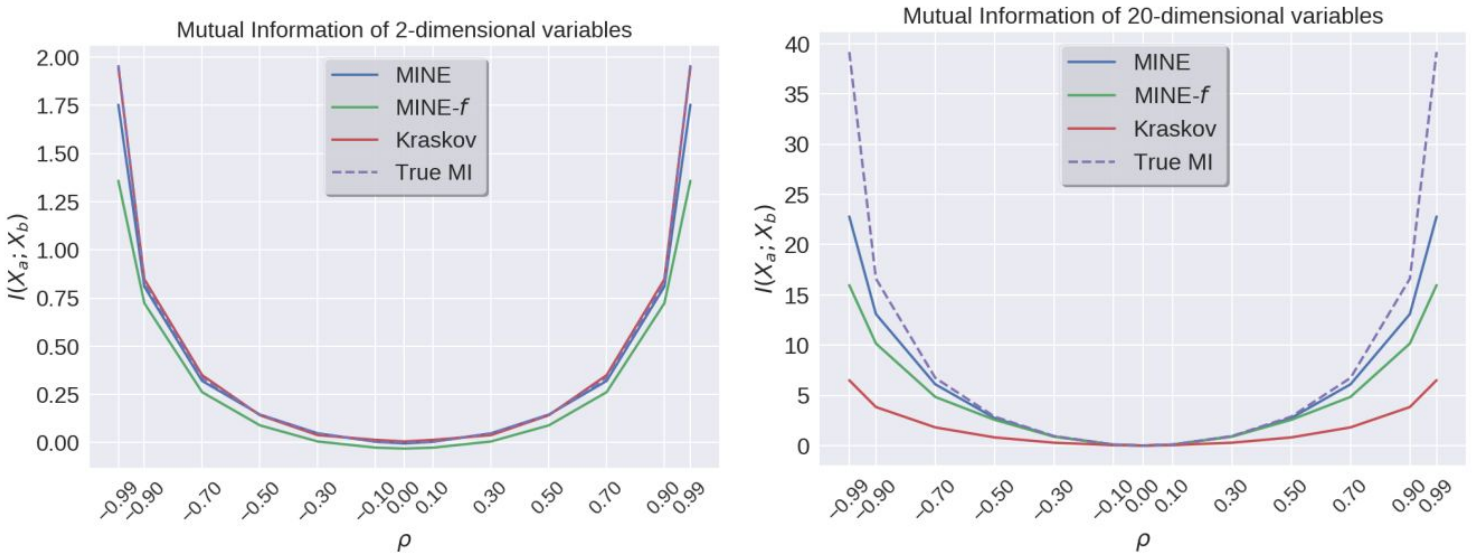

Figure 1. Mutual information between two multivariate Gaussians with component-wise correlation $\rho \in (-1, 1)$.

Results: Empirical

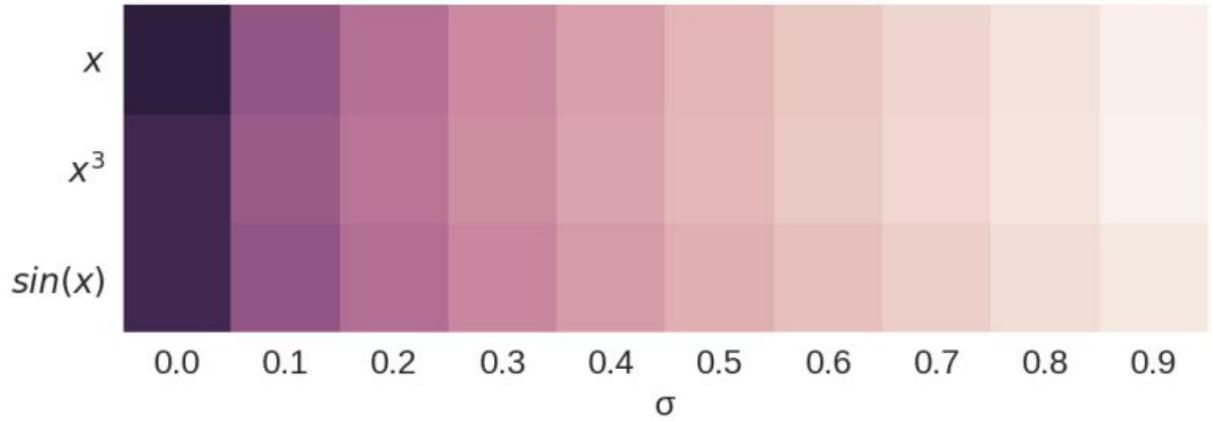

Figure 2. MINE is invariant to choice of deterministic nonlinear transformation. The heatmap depicts mutual information estimated by MINE between 2-dimensional random variables $X \sim \mathcal{U}(-1, 1)$ and $Y = f(X) + \sigma \odot \epsilon$, where $f(x) \in \{x, x^3, sin(x)\}$ and $\epsilon \sim \mathcal{N}(0, I)$.

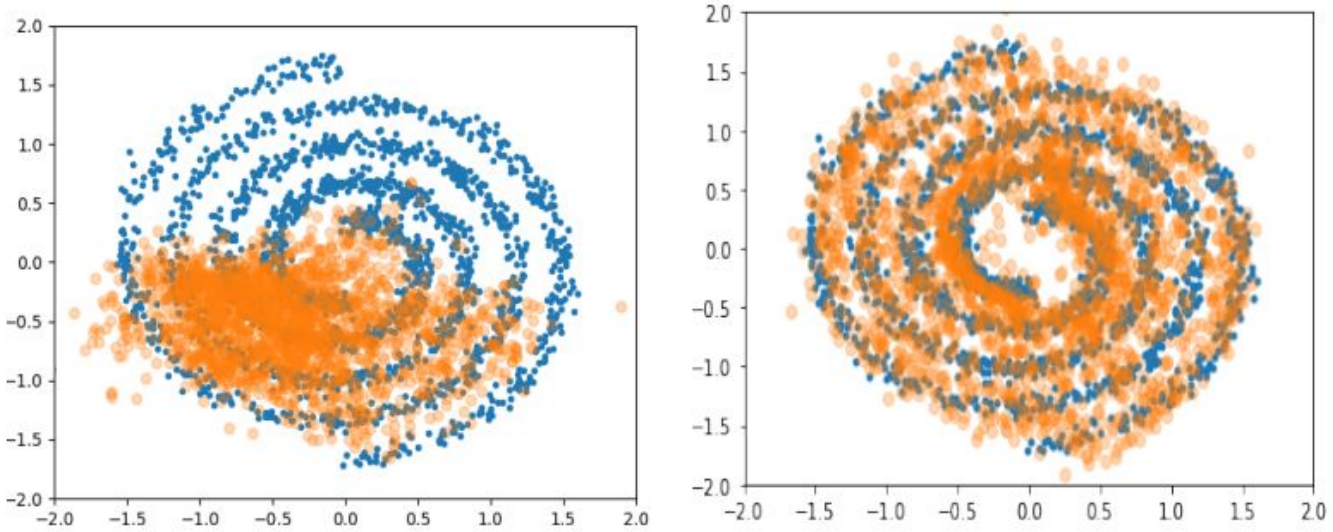

Figure 3. The generator of the GAN model without mutual information maximization after 5000 iterations suffers from mode collapse (has poor coverage of the target dataset) compared to GAN+MINE on the spiral experiment.

(a) Original data (b) GAN (c) GAN+MINE

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

	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 ± 0.0	$0.06 \pm 1.0\text{e}^{-2}$
GAN+MINE (Ours)	1000.0 ± 0.0	$0.05 \pm 6.9\text{e}^{-3}$

Table 1. Number of captured modes and Kullblack-Leibler divergence between the training and samples distributions for DCGAN (Radford et al., 2015), ALI (Dumoulin et al., 2016), Unrolled GAN (Metz et al., 2017), VeeGAN (Srivastava et al., 2017), PacGAN (Lin et al., 2017).

Results: ALI

- Basically a GAN that also does inference
- The paper basically just builds on GAN, but now we focus on the reconstruction loss:

$$\mathcal{R} \leq D_{KL}(q(\boldsymbol{x}, \boldsymbol{z}) \,||\, p(\boldsymbol{x}, \boldsymbol{z})) - I_q(\boldsymbol{x}, \boldsymbol{z}) + H_q(\boldsymbol{z})$$

- Tying this back to the whole model, we want to maximize mutual information

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.

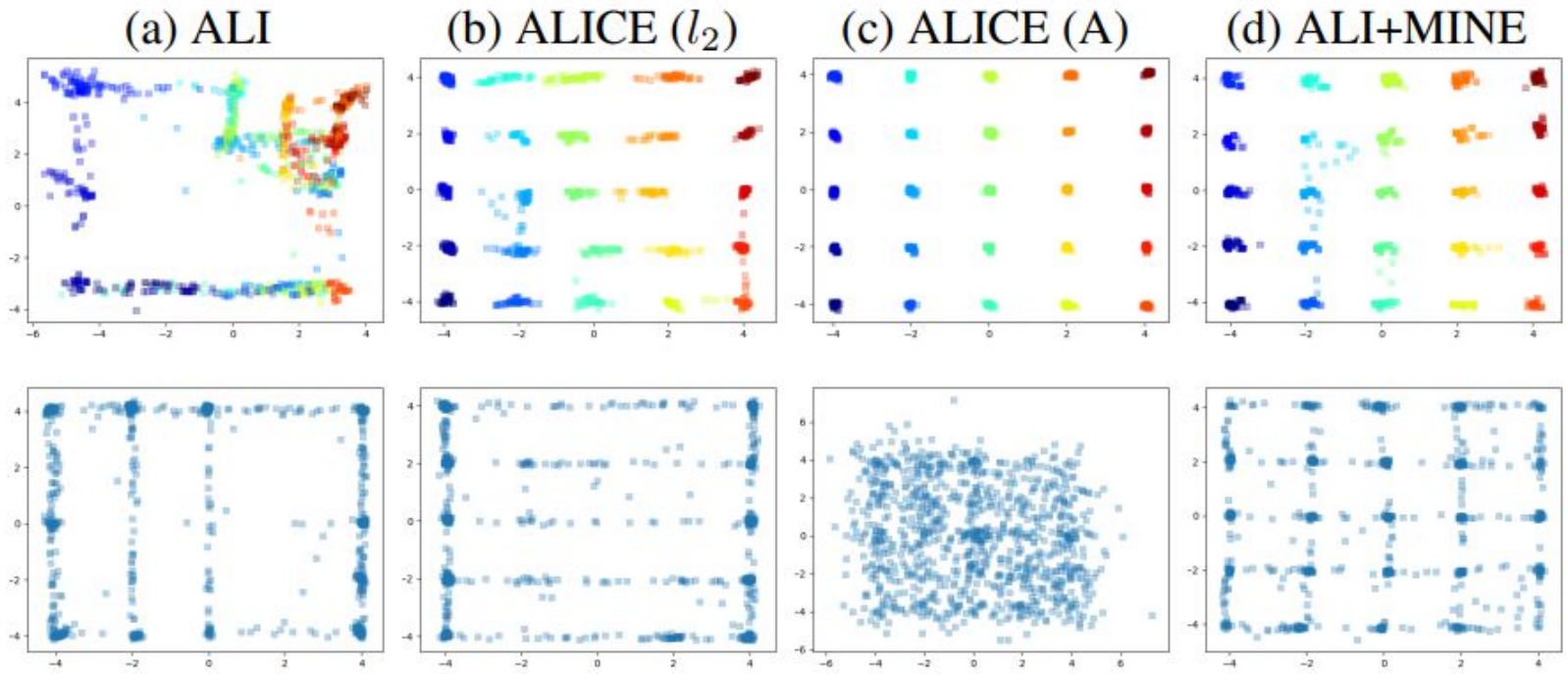

Figure 6. Reconstructions and model samples from adversarially learned inference (ALI) and variations intended to increase improve reconstructions. Shown left to right are the baseline (ALI), ALICE with the l2 loss to minimize the reconstruction error, ALICE with an adversarial loss, and ALI+MINE. Top to bottom are the reconstructions and samples from the priors. ALICE with the adversarial loss has the best reconstruction, though at the expense of poor sample quality, where as ALI+MINE captures all the modes of the data in sample space.

Results: Information Bottleneck

- MINE offers versatility in the distributions that can be used for models based on information bottleneck
- This is due to the intractability of the mutual information in the continuous setting, which MINE does well

$$L[q(Z|X)] = H(Y|Z) + \beta I(X;Z)$$

Model	Misclass. rate(%)
Baseline	1.38%
Dropout	1.34%
Confidence penalty	1.36%
Label Smoothing	1.40%
DVB	1.13%
DVB + Additive noise	1.06%
MINE(Gaussian) (ours)	1.11%
MINE(Propagated) (ours)	1.10%
MINE(Additive) (ours)	1.01%

Table 3. Permutation Invariant MNIST misclassification rate using Alemi et al. (2016) experimental setup for regularization by confidence penalty (Pereyra et al., 2017), label smoothing (Pereyra et al., 2017), Deep Variational Bottleneck(DVB) (Alemi et al., 2016) and MINE. The misclassification rate is averaged over ten runs. In order to control for the regularizing impact of the additive Gaussian noise in the additive conditional, we also report the results for DVB with additional additive Gaussian noise at the input. All non-MINE results are taken from Alemi et al. (2016).