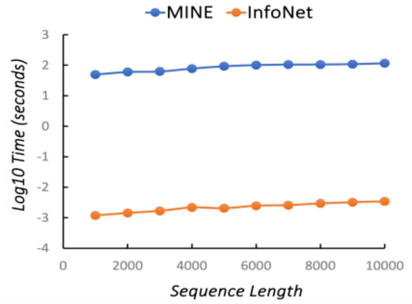
InfoNet: Neural Estimation of Mutual Information without Test-Time Optimization

Zhengyang Hu, Song Kang, Qunsong Zeng, Kaibing Huang, Yanchao Yang

THE UNIVERSITY OF HONG KONG

Contribution

- We propose InfoNet, the first mutual information model pre-learning from various synthetic distributions. Thus only one feed-forward pass can get the estimation of mutual information.



Extra Fast!

Motivation

- Mutual Information(MI):

$$\mathbb{I}(X;Y) \overset{\text{def}}{=} \sum p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

  is a good measure of the similarity between two variables.
- It is robust and can capture nonlinear relationships between features.
- Current neural MI estimation methods are not time efficient enough [1], statistical estimators are not differentiable [2], can not be used in modern learning frameworks.

Preliminary

- **Donsker-Varadhan Representation of MI**: $\mathbb{I}(x;y) = \sup_{\theta} E_{p_{x,y}}[\theta] - log\, E_{p_x p_y}[\exp(\theta)]$.
- θ is a scalar function $\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$, can be represented by a network or **a lookup table**.
- MINE[1] trains an MLP for each pair of x and y
  from scratch and do gradient ascend to optimize this lower bound until convergence to get the estimation of MI.

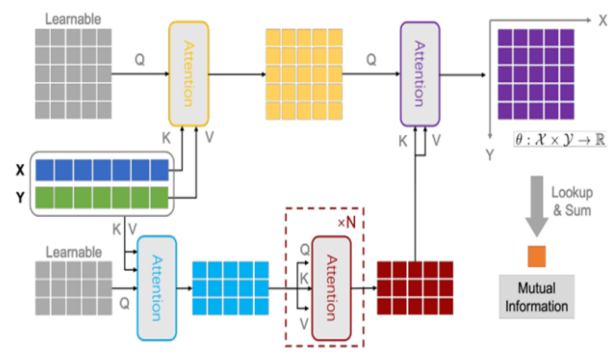
Method

We train a simulation-based model that can directly output the mutual information estimate of a sequence of samples within **only one feed-forward pass**. Detailed points are listed below:

- **Lookup table:** using a 2D lookup table $\mathbb{R}^{L\times L}$ as discretization representation of θ in D-V formular, then $\theta(x,y)$ can be directly read from it using bilinear interpolation.
- **Training data:** sampled from Gaussian Mixture Models (GMMs), representing distributions as weighted Gaussian sums.
- **Loss:** ϕ means model parameter, $\mathcal{D}$ is a dataset of N different distributions:

$$\mathcal{L}_{\text{MI}}(\phi,\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{T}\sum_{t=1}^{T}\theta_{\mathbf{x}^i,\mathbf{y}^i}(\mathbf{x}_t^i,\mathbf{y}_t^i) - \log\left(\frac{1}{T}\sum_{t=1}^{T}\exp(\theta_{\mathbf{x}^i,\mathbf{y}^i}(\mathbf{x}_t^i,\mathbf{y}_t^i))\right)\right\}.$$

Model Architecture



- Any-length input, permutation invariance
- Separate process joint and marginal samples

Important Techniques

Sliced Mutual Information

- Estimate high-dimensional mutual information by randomly projecting data onto lower(one) dimensional subspaces and aggregating the results [4]:

$$SI(X;Y) = \frac{1}{S_{d_x-1}S_{d_y-1}}\int_{S_{d_x-1}}\int_{S_{d_y-1}} I(\theta^T X;\phi^T Y)\,d\theta\, d\phi$$

- Preserve many properties and orders of original MI.

Copula Transformation

- Transform the original sample into uniform marginals on the interval [0,1] before training and testing.
- Similar to applying rank data on X and Y separately. Using softrank[5] in training tasks.
- MI is invariant during the transformation.
- Only need to consider the relative position relationship. Reduce data complexity and improve generalization.

Experiment Results

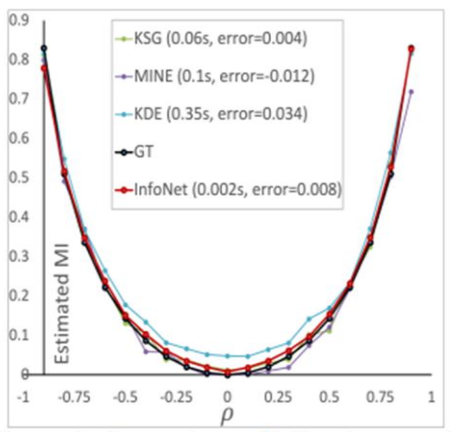
Time Complexity Comparing (seconds)

SEQ. LEN GTH	200	500	1000	2000	5000
KSG	0.009	0.024	0.049	0.098	0.249
KDE	0.004	0.021	0.083	0.32	1.801
MINE-2000	3.350	3.455	3.607	3.930	4.157
MINE-500	0.821	0.864	0.908	0.991	1.235
MINE-10	0.017	0.017	0.019	0.021	0.027
InfoNet-16	**0.001**	**0.002**	**0.002**	**0.002**	**0.003**

MINE-500 means train MINE for 500 iterations. InfoNet-16 means InfoNet estimates 16 different sequences in one forward pass.

Evaluate on Gaussian

We perform a check on the Gauss distributions, that has analytical ground truth MI.



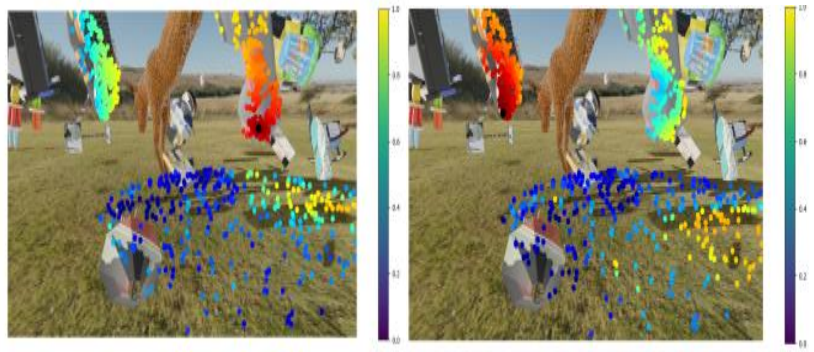
GMM Correlation Order Accuracy

In practice, correlation order is more critical for decision-making. Given one control variable A, and two observation variables B & C, $\mathbb{I}(A,B) > \mathbb{I}(A,C)$ or $\mathbb{I}(A,B) < \mathbb{I}(A,C)$?

No. of Comps.	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
KSG	98.7	99.0	98.2	98.0	97.9	97.7	97.6	97.5	97.0	97.3
KDE	97.4	97.7	97.9	97.5	97.9	97.8	97.0	97.4	97.4	97.4
MINE-500	98.5	91.2	90.8	87.2	84.5	83.7	81.2	79.6	81.3	78.1
MINE-100	94.6	77.1	75.4	71.6	67.5	69.4	66.5	66.3	68.7	66.4
MINE-10	60.9	56.1	55.1	54.3	52.4	54.9	53.7	50.4	53.1	52.5
InfoNet	**99.8**	**99.5**	**99.0**	**99.2**	**99.1**	**99.2**	**99.0**	**99.2**	**99.3**	**99.5**

Validation on Motion Data

Estimate mutual information between point trajectories in the Pointodyssey dataset [3]:



Left: Estimated MI with point in object 1 (black)

Right: Estimated MI with point in object 2 (black)

References

1. Belghazi et al, "Mutual Information Neural Estimation" ICML 2018
2. A Kraskov et al "Estimating mutual information" 2004
3. Zheng Y et al. "Pointodyssey: A large-scale synthetic dataset for long-term point tracking" ICCV 2023
4. Z Goldfeld et al "Sliced mutual information: A scalable measure of statistical dependence" Neurips 2021
5. M Blondel "Fast Differentiable Sorting and Ranking" ICML 2020

Welcome to our paper to find more experiments and details!

Arxiv

Github