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

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

The University of Hong Kong

Electrical and Electronic Engineering

Institute of Data Science

Mutual Information

A Mathematical Theory of Communication

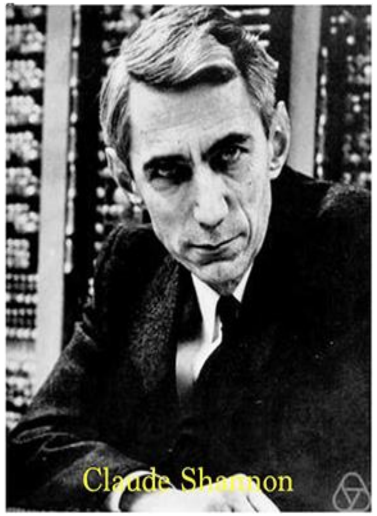
By C. E. SHANNON

9. THE FUNDAMENTAL THEOREM FOR A NOISELESS CHANNEL

We will now justify our interpretation of H as the rate of generating information by proving that H determines the channel capacity required with most efficient coding.

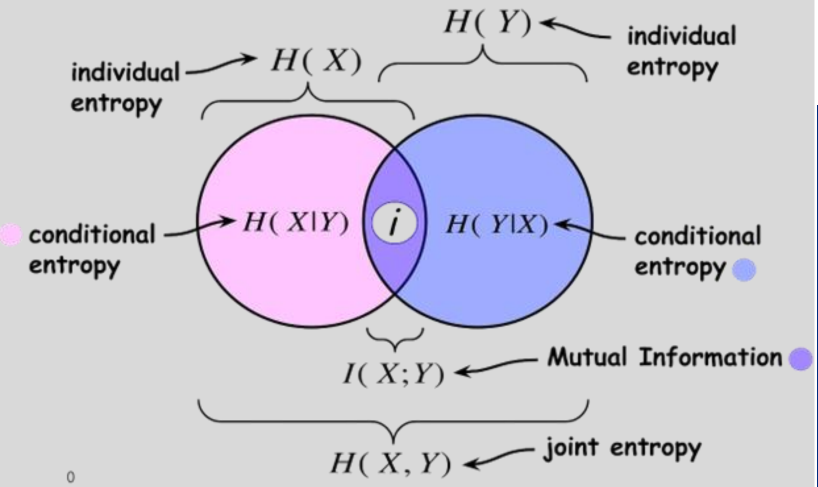
Theorem 9: Let a source have entropy H (bits per symbol) and a channel have a capacity C (bits per second). Then it is possible to encode the output of the source in such a way as to transmit at the average rate $\frac{C}{H} - \epsilon$ symbols per second over the channel where ϵ is arbitrarily small. It is not possible to transmit at an average rate greater than $\frac{C}{H}$.

1948



Mutual Information:

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



Properties:
- Non-negativity
- Transformation Invariance
- Data Processing Inequality
- Chain Rule
- …

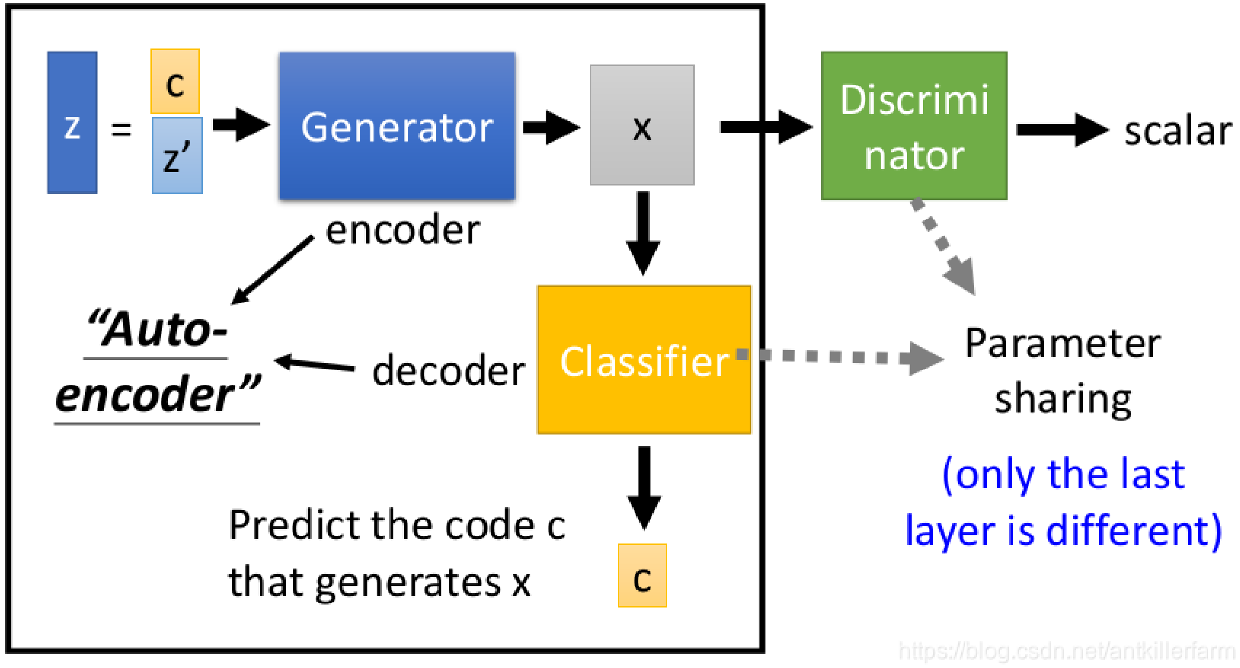
Advantages:
- Robustness
- Comprehensive Dependence Measure
- Nonlinear Sensitivity
- …

Mutual Information in Adversarial Learning

InfoGAN

Enhances GAN by maximizing the mutual information between the generated samples and the interpretable latent variables.

- Improves Disentanglement
- Enhances Data Generation
- Better Interpretability



InfoGAN architecture.

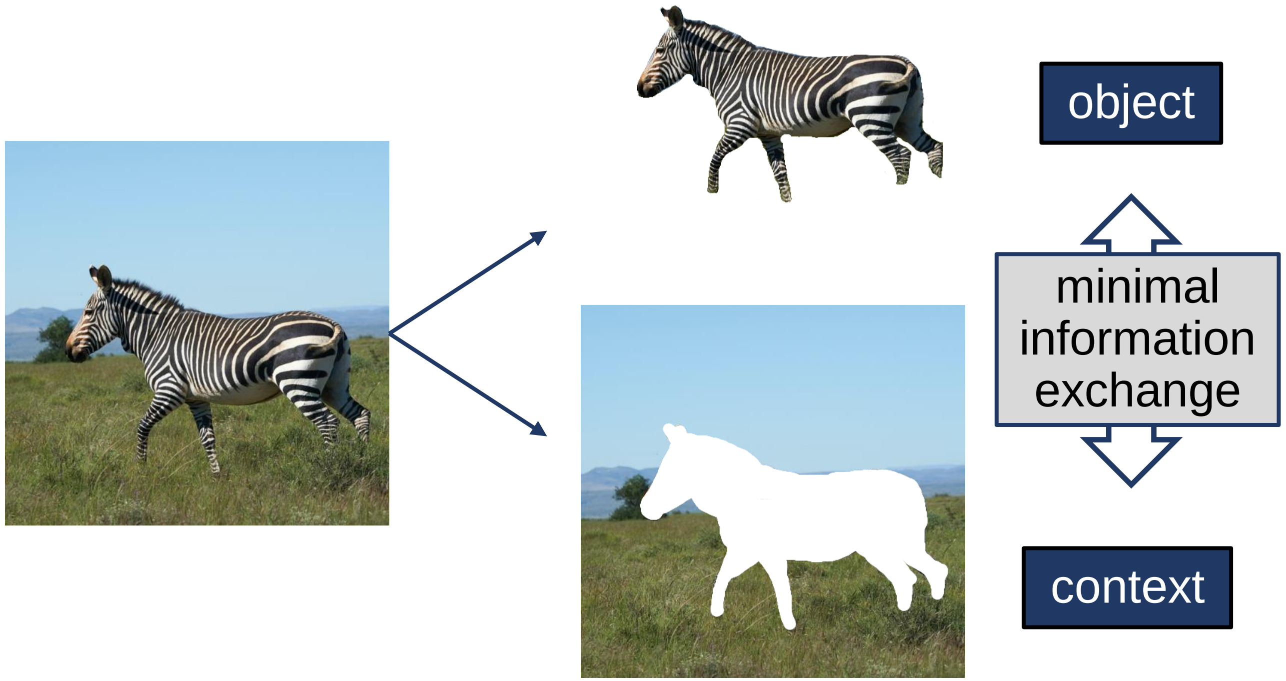
$$\min_G \max_D V_{\mathbb{I}}(D,G) = V(D,G) - \lambda \mathbb{I}(c; G(z,c))$$

Chen et al. "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets" Neurips16

Mutual Information in Video Segmentation

Video Object Segmentation

Minimize the mutual information between the pixels within and outside the region.

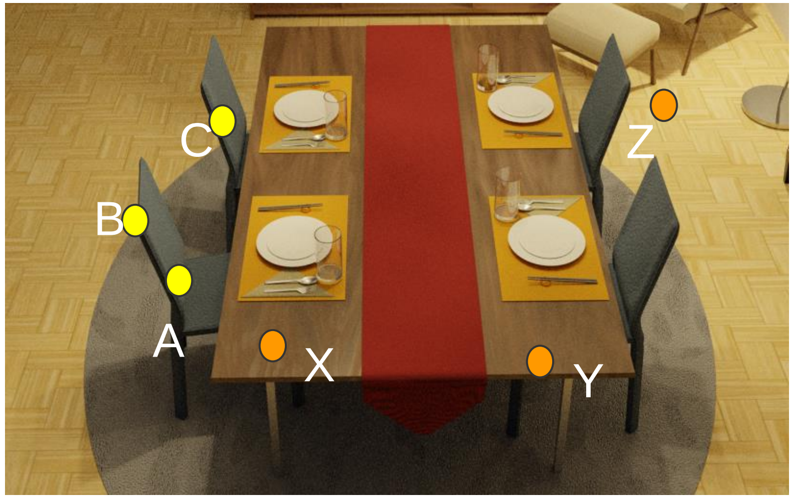


- Self-supervised object segmentation
- No need of explicit regularizers
- Improves generalizability

$$\underbrace{\mathcal{L}(\mathrm{m})}_{\text{region}} = \frac{\mathbb{I}(\overbrace{\mathrm{m}\odot\mathbf{u}}^{\text{object}} \,|\, \overbrace{(1-\mathrm{m})\odot \mathrm{u}}^{\text{context}})}{\mathbb{H}(\mathrm{m}\odot\mathbf{u})} + \frac{\mathbb{I}((1-\mathrm{m})\odot\mathbf{u} \,|\, \mathrm{m}\odot \mathrm{u})}{\mathbb{H}((1-\mathrm{m})\odot\mathbf{u})}$$

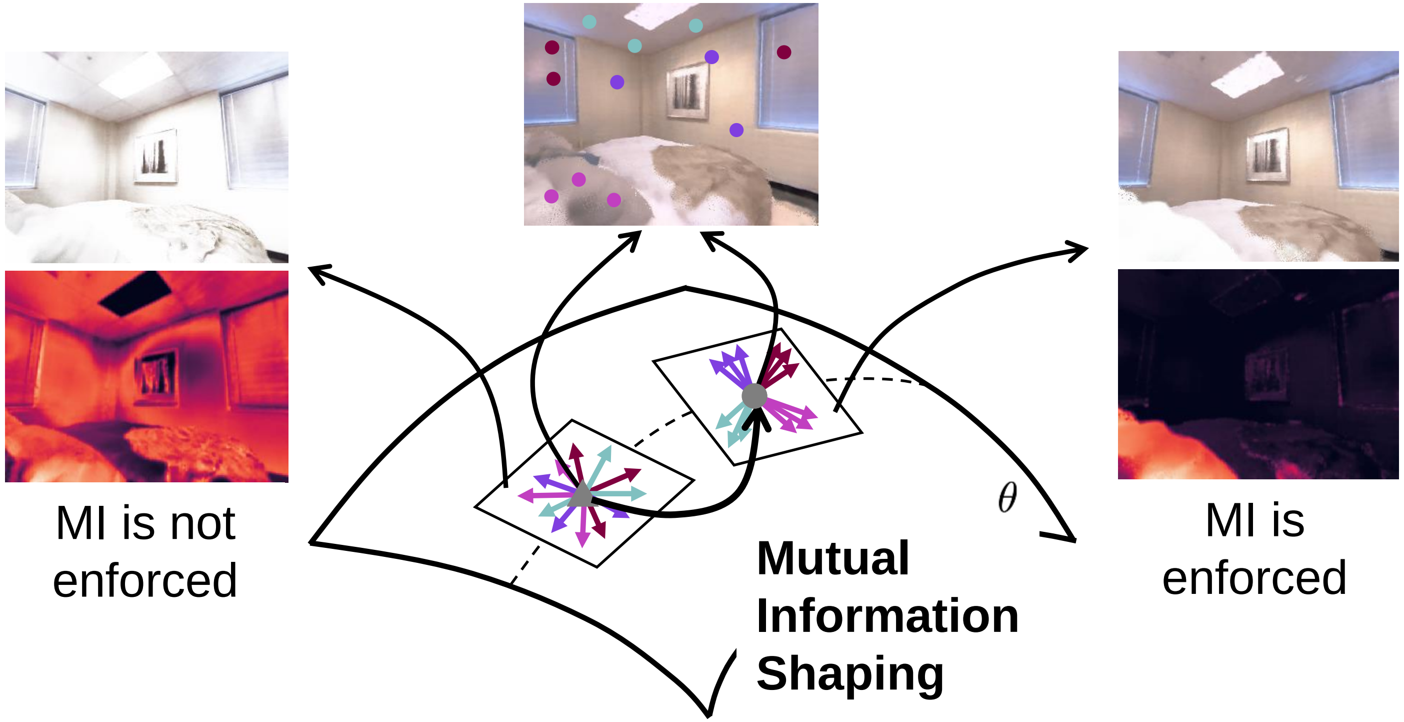
Mutual Information for Controllable Scene Representation

Encode Mutual Information correlation into NeRFs



A is more correlated with B than with C
X is more correlated with Y than with Z

$$\mathbb{I}(A,B) > \mathbb{I}(A,C) \qquad \mathbb{I}(X,Y) > \mathbb{I}(X,Z)$$

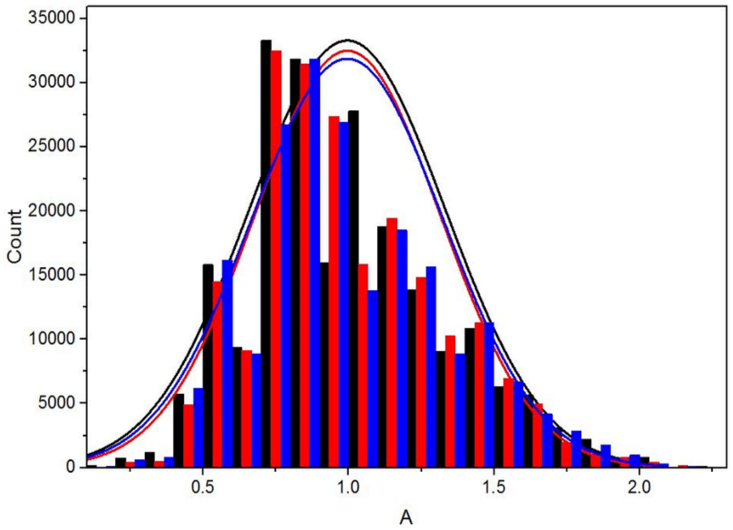


Xu & Yang et al. "JacobiNeRF: NeRF Shaping with Mutual Information Gradients" CVPR 23

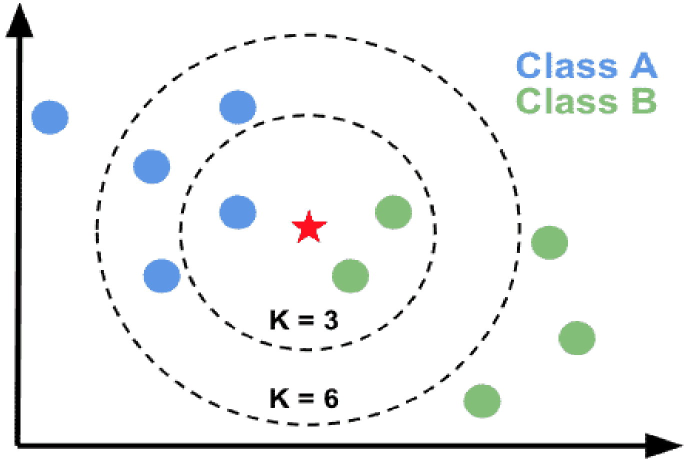
Traditional Mutual Information Estimators

- **Histogram**
- **K-Nearest Neighbor**
- **Kernel Density Estimation**

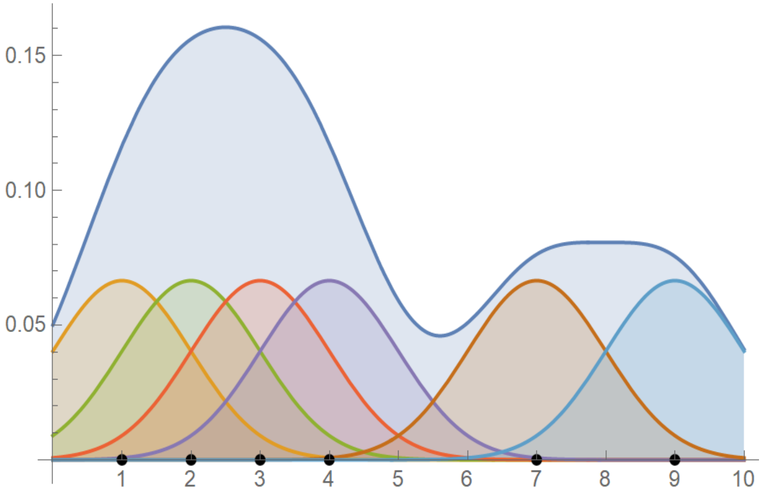
- Non-differentiable
- Inefficiency
- Curse of dimensionality



Histogram



K-Nearest Neighbors



Kernel Density Estimation

A Kraskov et al "Estimating mutual information" Physical Review E 2004

Mutual Information Neural Estimation (MINE)

MINE

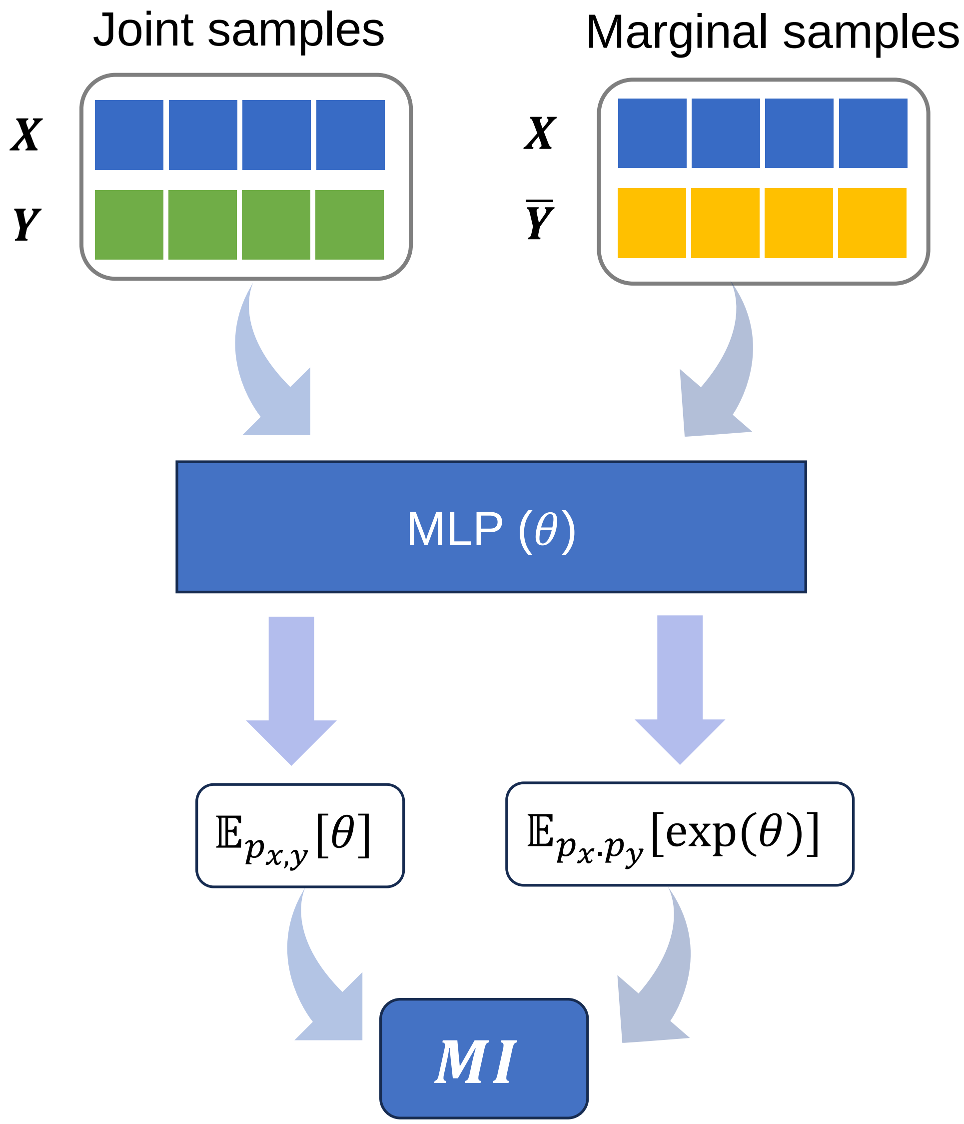
MI estimation as *functional optimization*

Donsker-Varadhan Representation:

$$\mathbb{I}(\mathbf{x},\mathbf{y}) = \sup_{\theta} \mathcal{J}^{\text{info}}(\theta;\mathbf{x},\mathbf{y})$$
$$= \sup_{\theta} \mathbb{E}_{\mathrm{p}_{\mathbf{x},\mathbf{y}}}[\theta] - \log(\mathbb{E}_{\mathrm{p}_{\mathbf{x}}\cdot \mathrm{p}_{\mathbf{y}}}[\exp(\theta)])$$

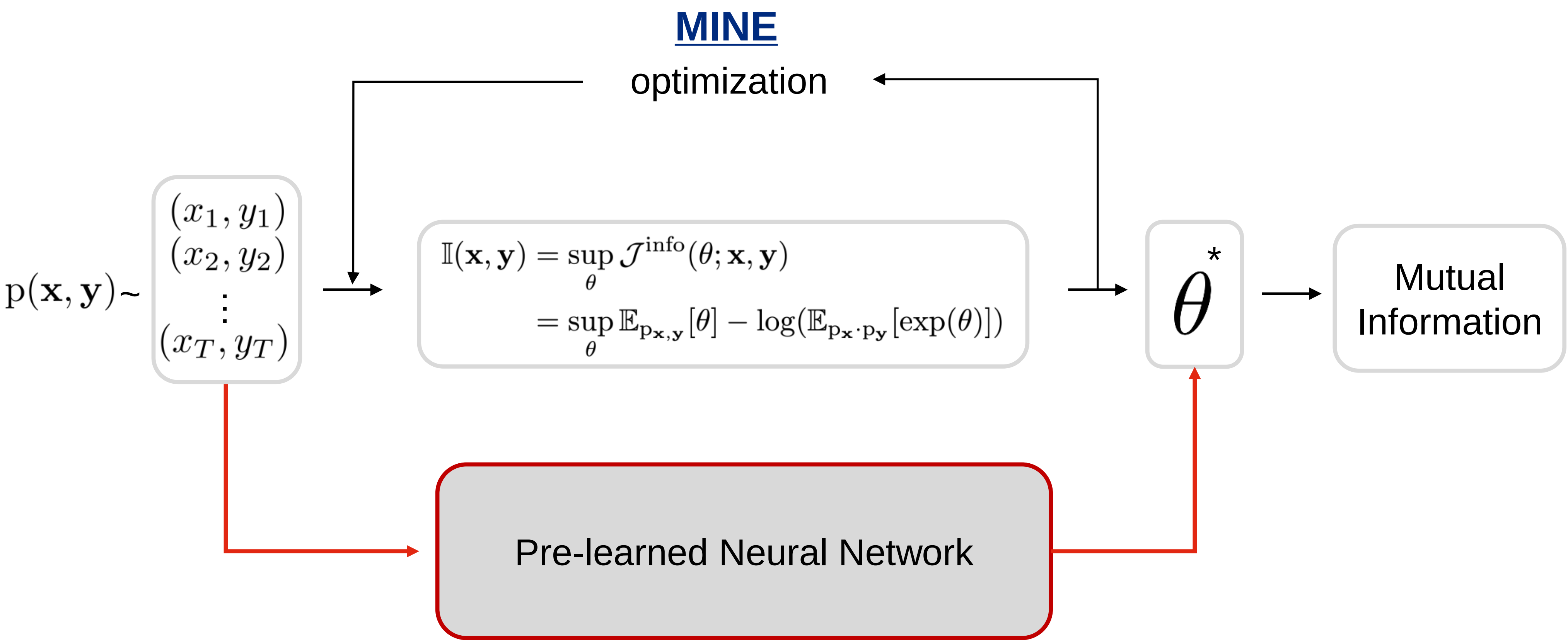
where $\theta : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$

However, for each pair of X and Y, a new MLP must be trained from scratch. Time-consuming and unstable.



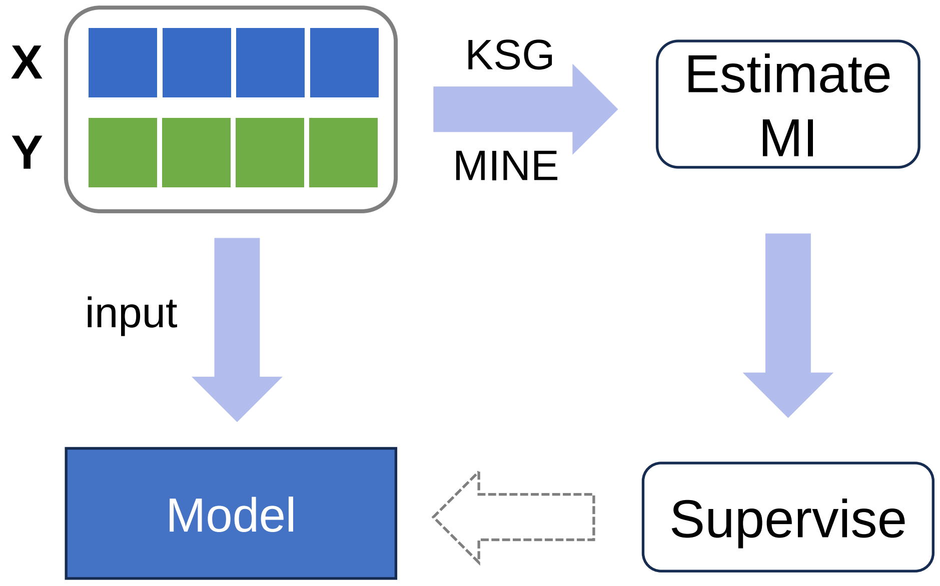
Belghazi et al, "Mutual Information Neural Estimation" ICML 2018

Can we design a neural network that could pre-learn the mutual information from all distributions?



No need to perform optimization, fast, and, differentiable!

How? Can we use pre-computed MI to supervise?



Generalization Ability

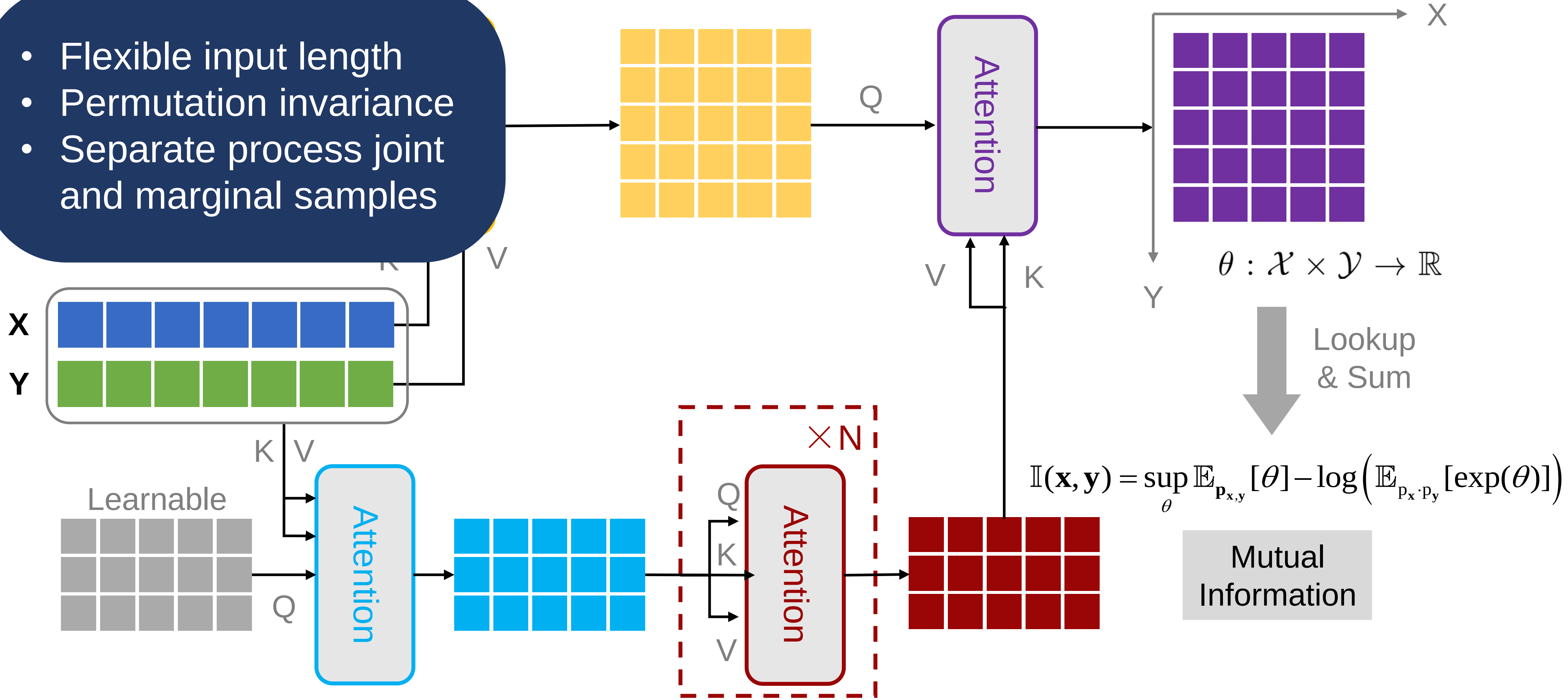
Difficult to predict MI on unseen distributions.

Efficiency

- Precomputing MI on various sequences is **time-consuming.**
- Performance will be upper-bounded by these precomputing methods.

InfoNet Architecture

- Flexible input length
- Permutation invariance
- Separate process joint and marginal samples

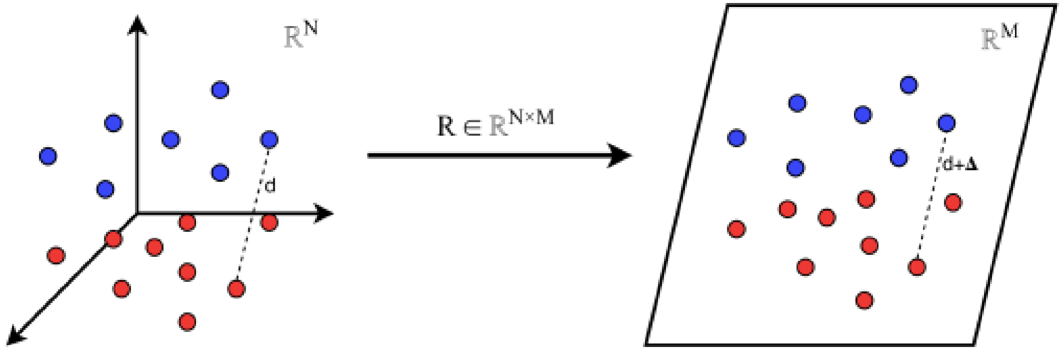


Sliced Mutual Information(SMI)

Estimate high-dimensional mutual information by randomly projecting data onto lower-dimensional subspaces and aggregating the results.

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

S^{d-1} denotes the d-dimensional sphere (its surface area is designated by S_{d-1}).



Random projection

Using SMI, we can focus on the MI between all **one-dimensional** XY pairs.

Z Goldfeld et al "Sliced mutual information: A scalable measure of statistical dependence" Neurips 2021

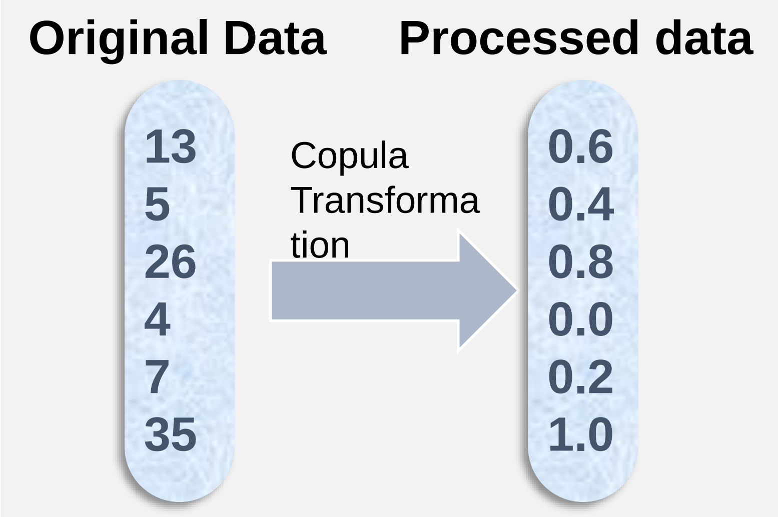
Inject Normalization

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
- Mutual Information is invariant during the transformation

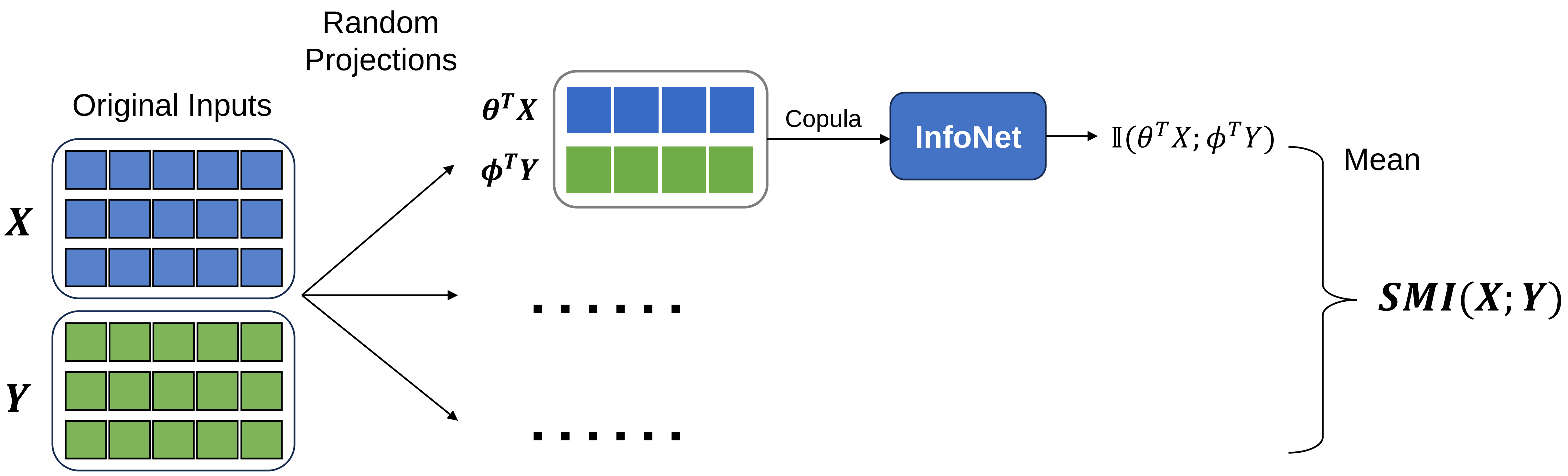
Advantages

- Only need to consider the relative position relationship.
- Reduce data complexity and improve the generalization ability of the model.



Undifferentiable?
Using SoftRank instead in training tasks.

Estimation Pipeline

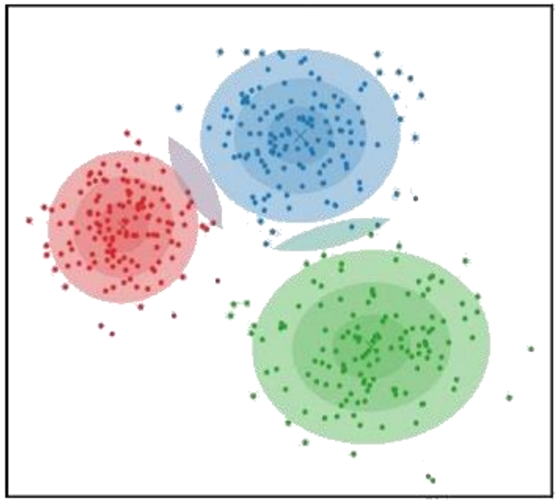


Training Data and Time

Gaussian Mixture Models

A weighted sum of multiple Gaussian distributions, each defined by its own mean and variance.

$$p(z) = \sum_{i=1}^{K} \pi_i \mathcal{N}(z|\mu_i, \Sigma_i)$$



GMM with three Gauss components

- Strong generalization ability
- Approximate any arbitrary distribution well with a sufficient number of Gauss components

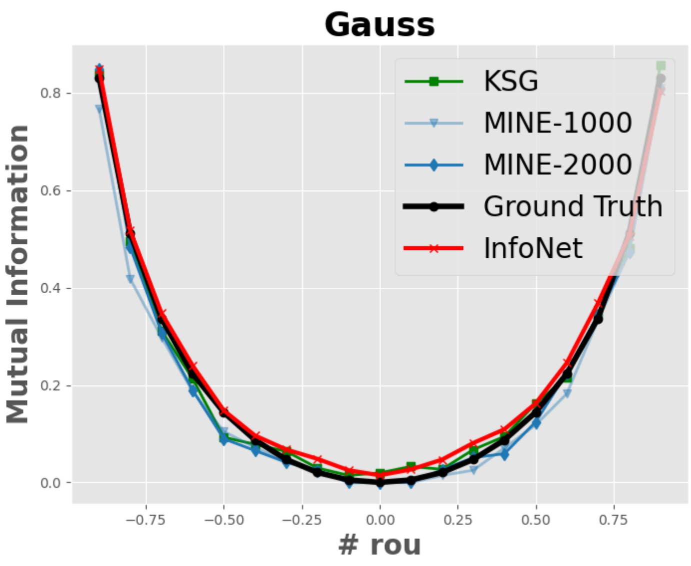
Training Time

5 hours to converge on RTX 4090

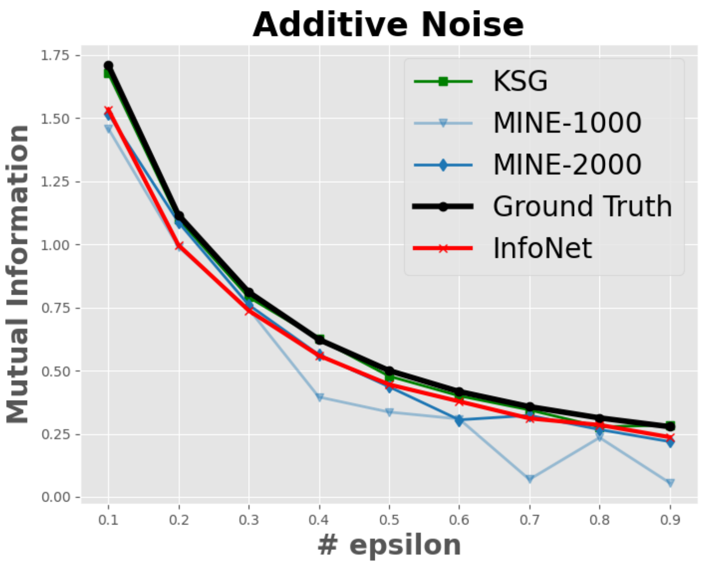
Much faster than estimating the MI of all training data using MINE individually.

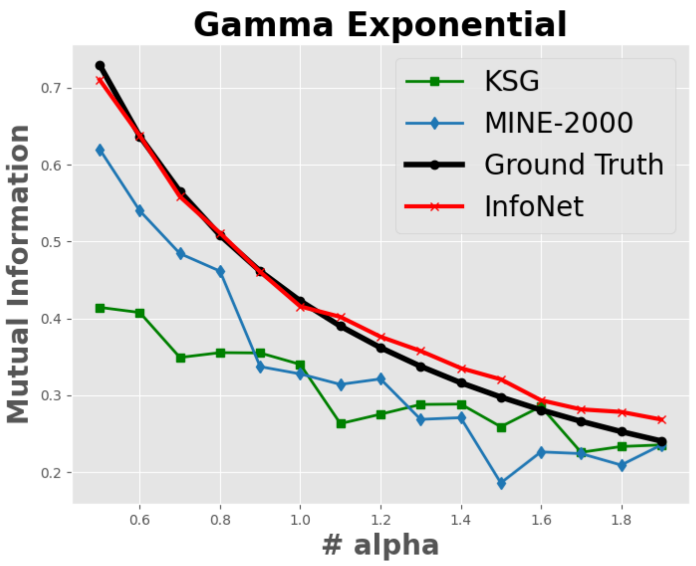
InfoNet performs well on seen and unseen distributions

Seen Distributions



Unseen Distributions





InfoNet is test-time efficient

SEQ. LENGTH	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:** train MINE for 500 iterations.
- **InfoNet-16:** estimate 16 distributions using InfoNet simultaneously (batchsize=16)

InfoNet is robust in correlation order prediction

In practice, correlation order is more critical for decision making

Given one reference variable A, and two test variables B & C, $\mathbb{I}(A,B) > \mathbb{I}(A,C)$ or $\mathbb{I}(A,B) < \mathbb{I}(A,C)$?

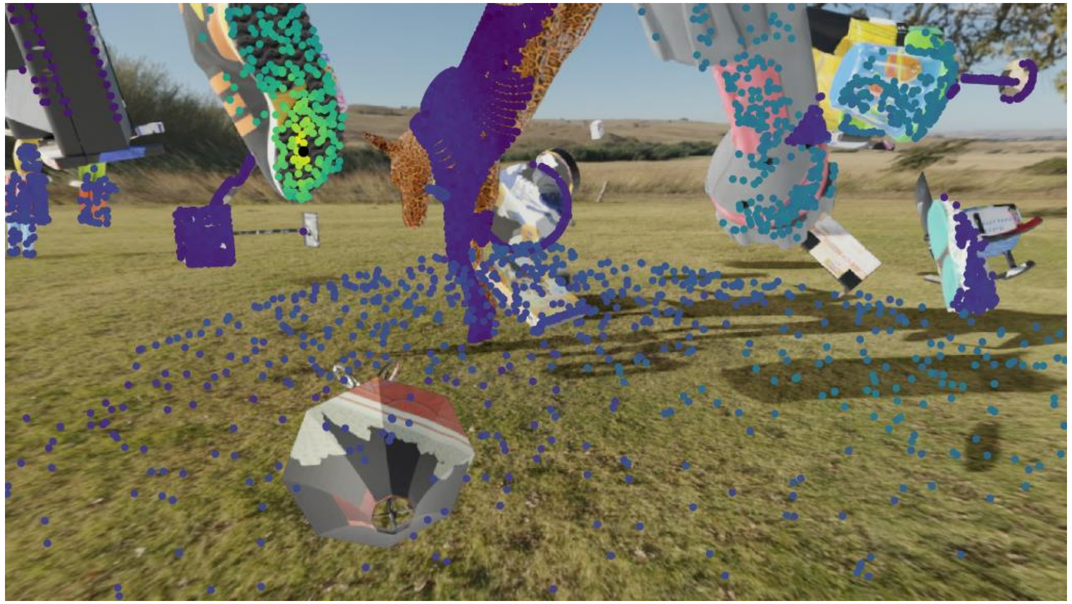
No. of Comps.	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
KSG	98.7	99.0	98.2	98.0	97.9	97.7	97.6	97.5	97.0	97.3
KDE	97.4	97.7	97.9	97.5	97.9	97.8	97.0	97.4	97.4	97.4
MINE-500	98.5	91.2	90.8	87.2	84.5	83.7	81.2	79.6	81.3	78.1
MINE-100	94.6	77.1	75.4	71.6	67.5	69.4	66.5	66.3	68.7	66.4
MINE-10	60.9	56.1	55.1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

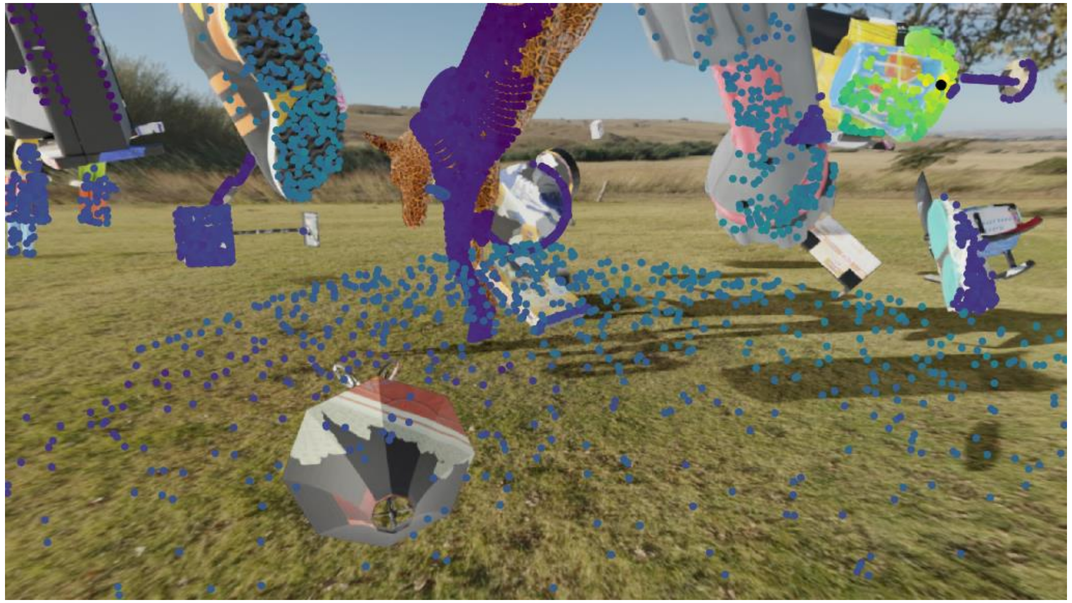
Mutual Information between point trajectories

Mutual information of trajectories between motion of video points.

T represents point trajectory in the video.

$$\mathbb{I}\left(T_{selected\ point}, T_{point\ from\ same\ object}\right) > \mathbb{I}\left(T_{selected\ point}, T_{point\ from\ other\ object}\right)$$





Zheng et al. "PointOdyssey: A Large-Scale Synthetic Dataset for Long-Term Point Tracking" ICCV 23

Conclusion and Discussion

1. InfoNet is the first mutual information estimate model pre-learns from various different distributions.

2. It has extra fast estimation speed and strong generalization ability, and numerous potential applications in the future.

3. Estimating high-dimensional MI requires more slices, reducing speed. Our current research aims to design a new architecture to address this issue.

InfoNet

Quick Mutual Information Estimator

Conclusion and Discussion

Thanks!
Q & A

Arxiv

Github