NEURAL INFORMATION PROCESSING SYSTEMS

A Benchmark Suite for Evaluating Neural Mutual Information Estimators on Unstructured Datasets

Kyungeun Lee, Wonjong Rhee

LG AI Research

Introduction

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

The exact calculation of MI is impossible
when we can only access the examples sampled from joint and marginals
but not the underlying distribution functions.

➔ We often rely on sample-based MI estimators.

Introduction

Estimation accuracy of sample-based MI estimators

Gaussian datasets

Tractable distributions
→ Tractable true MI

Complex unstructured datasets
(e.g., images, texts)

Intractable distributions
→ Intractable true MI

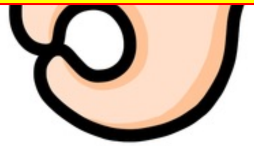
Introduction

Estimation accuracy of sample-based MI estimators

Gaussian datasets

Complex unstructured datasets
(e.g., images, texts)

**Do estimators that perform well on Gaussian datasets
also excel with more complex datasets like images or texts?**



Tractable distributions
→ Tractable true MI

Intractable distributions
→ Intractable true MI

Main contributions

We present a method for evaluating MI estimators on any dataset in the absence of underlying distribution functions.

- Same-class sampling as positive pairing
- Binary symmetric channels

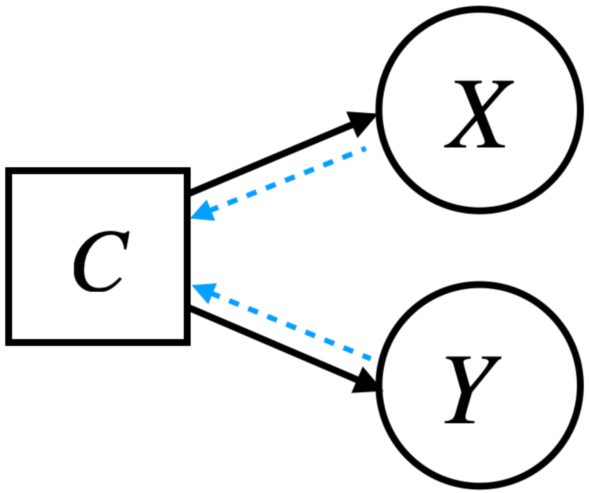
We suggest a benchmark suite based on our method, encompassing three data domains for Gaussian multivariates, images and sentence embeddings.

We examine performance of several neural MI estimators over seven key aspects:
critic architecture, critic capacity, choice of neural MI estimator, number of information sources, representation dimension, strength of nuisance, and layer-dependency.

Our benchmark suite

Same-class sampling for positive pairing

- When only the class information is shared between two random variables X and Y, the true MI is proven to be the same as the entropy of the class variable C.
- $I(X;Y)=H(C)$ for any choice of data domain.

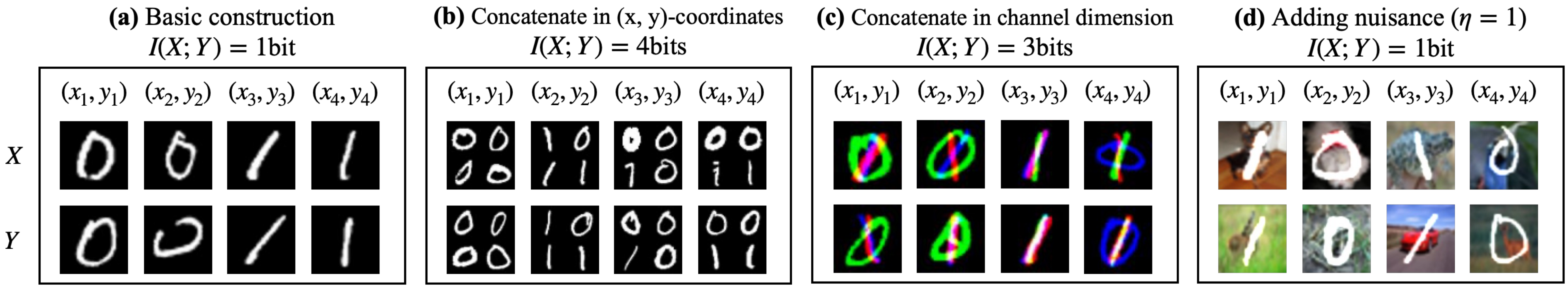


[Reference] Lee et al., Towards a rigorous analysis of mutual information in contrastive learning, 2024.

Our benchmark suite

Generating datasets with adjustable true MI values

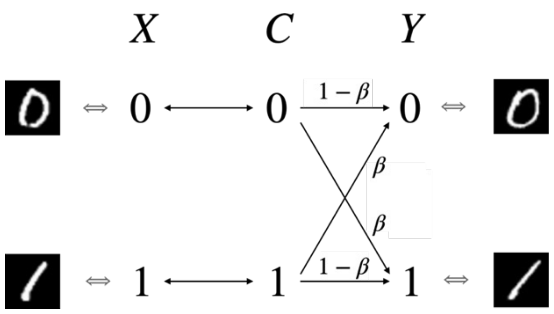
- Plain setup：Using a binary random variable C where p(0)=p(1)=0.5, I(X;Y)=1(bit)
  - Images：MNIST 0/1 images
  - Texts：BERT fine-tuned sentence embeddings of IMDB datasets
- Larger MI：Concatenating the samples of I(X;Y)=1
- Nuisance：Inserting random samples from CIFAR-10 in the background



Our benchmark suite

Manipulating MI to non-integer values: Binary symmetric channel

- To manipulate the true MI and construct a dataset with a non-integer MI value, we adopt the concept of binary symmetric channel (BSC).
- With BSC, X is always consistent with the class variable C but Y is noisy where it is different from C with a crossover probability of β.

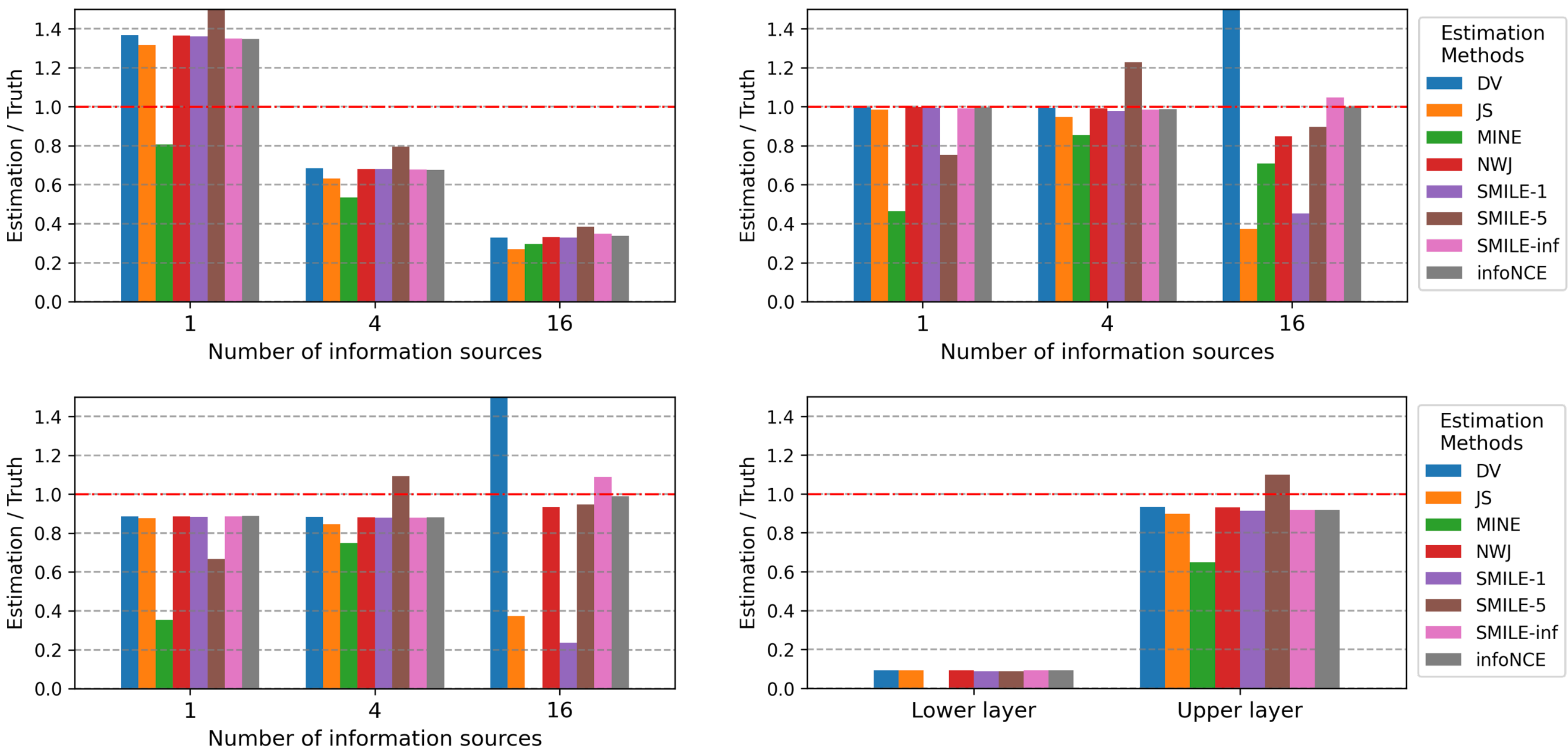


Theorem 4.4 (Manipulating MI to be non-integer). *When the information source C is transmitted perfectly to X, while it is transmitted to Y over a binary symmetric channel (BSC) with a crossover probability $\beta \in [0, 0.5]$, the mutual information $I(X;Y)$ between X and Y is determined as follows.*

$$I(X;Y) = H(C) \times (1 - H(\beta)) \quad (1)$$

[Reference] T. M. Cover, Elements of information theory, 1999.

Empirical investigations



Empirical investigations

- **Choice of critic architecture**: superiority of joint critic for unstructured datasets
- **Choice of critic capacity**: larger capacity does not ensure a higher estimation accuracy
- **Choice of MI estimator**: no universal winner exists across the three data domains
- **Number of information sources**: unstructured datasets outperform Gaussian in handling larger d_s
- **Representation dimension**: it does not affect the estimation accuracy
- **Nuisance**: MINE turns out to be relatively robust
- **Network and layer dependency**: estimation holds for invertible networks and upper layers

Thank you ☺

Visit our paper at Poster Session 1

Paper

Github