

# Information Bottleneck Analysis of Deep Neural Networks

Aleksandr Tolmachev<sup>1, 2</sup>

Research advisor: Dr.Sc. Alexey Frolov<sup>2</sup>

<sup>1</sup>Moscow Institute of Physics and Technology,

<sup>2</sup>Skolkovo Institute of Science and Technology

June 20, 2024

# Introduction: main definitions and preliminaries

Consider random vectors, denoted as  $X : \Omega \rightarrow \mathbb{R}^n$  and  $Y : \Omega \rightarrow \mathbb{R}^m$ , where  $\Omega$  represents the sample space. Let's assume that these random vectors have absolutely continuous probability density functions (PDF) denoted as  $\rho(x)$ ,  $\rho(y)$ , and  $\rho(x, y)$ , respectively.

## Differential entropy

- differential entropy of  $X$ :  $h(X) = -\mathbb{E} \log \rho(x)$
- conditional entropy:  $h(X | Y) = -\mathbb{E} \log \rho(X|Y) = -\mathbb{E}_Y (\mathbb{E}_{X|Y=y} \log \rho(X | Y = y))$
- joint differential entropy:  $h(X, Y) = -\mathbb{E} \log \rho(x, y)$

## Mutual Information

Mutual Information (MI) between variables  $X$  and  $Y$  is defined as

$$I(X; Y) = h(X) + h(Y) - h(X, Y) = \mathbb{E}_{\mathbb{P}_{(X,Y)}} \log \frac{d\mathbb{P}_{(X,Y)}}{d\mathbb{P}_X \otimes \mathbb{P}_Y} = D_{KL}(\mathbb{P}_{(X,Y)} || \mathbb{P}_X \otimes \mathbb{P}_Y)$$

Besides, the following equations holds:  $I(X; Y) = h(X) - h(X | Y) = h(Y) - h(Y | X)$

# Information Bottleneck principle (IB)

This concept was applied to DNNs by Shwartz-Ziv & Tishby (2017). The major idea of the IB approach is to **track the dynamics of two MI values**:

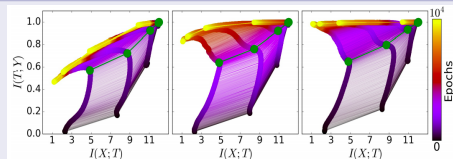
- $I(X; T)$  between the hidden layer output (T) and the DNN input (X)
- $I(Y; T)$  between the hidden layer output (T) and the target of the model (Y)

The **fitting-compression hypothesis** divides the learning process into two consequent phases:

- feature-extraction “**fitting**” phase: both MI values grow
- representation “**compression**” phase:  $I(Y; T)$  grows while  $I(X; T)$  decreases

## Fitting-Compression hypothesis: Tishby & Shwartz-Ziv conclusions

Firstly, classifier's construction based on the most significant features, next the internal representation is being compressed



# Aim and Objectives

## Problem Statement

Due to the **challenging nature of estimating MI between high-dimensional random vectors**, this hypothesis has only been verified for NNs of tiny sizes or specific types, such as quantized NNs

## Research goals

- 1 create the approach for the MI estimation that outperform previous methods in case of MI measurements between high-dimensional random variables
- 2 provide the Information Bottleneck analysis for close-to-real scale neural networks via the suggested approach

## Method: proposed ideas

**Manifold Hypothesis:** Real-world data usually lies (or close to) a low-dimensional manifold

### Compression is the main contribution

Our main goal is to precisely estimate MI in the high-dimensional case. To overcome the curse of dimensionality, we suggest to **COMPRESS the data before the MI estimation**:

- 1 learning the manifold with autoencoders
- 2 applying conventional estimators (KDE, KL, WKL, ...) to the compressed representations

### Loseless case: MI can be measured between loseless compressed representations

**Theorem 1.** Let  $\xi: \Omega \rightarrow \mathbb{R}^{n'}$  be an absolutely continuous random vector, let  $g: \mathbb{R}^{n'} \rightarrow \mathbb{R}^n$  be an injective piecewise-smooth mapping with Jacobian  $J$ , satisfying  $n \geq n'$  and  $\det(J^T J) \neq 0$  almost everywhere. Let  $h(\xi)$  and  $h(\xi | \eta)$  be defined. Then

$$I(\xi; \eta) = I(g(\xi); \eta) = I((g^{-1} \circ g)(\xi); \eta)$$

## Method: lossy compression case

Generally, **MI can get arbitrarily low** due to the imperfect (lossy) compression. However, additional assumptions allow for the following bounds:

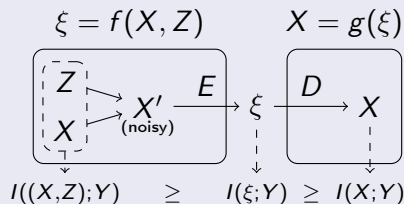
**Theorem 2.** Let  $X$ ,  $Y$ , and  $Z$  be random variables such that  $I(X; Y)$  and  $I((X, Z); Y)$  are defined. Let  $f$  be a function of two arguments such that  $I(f(X, Z); Y)$  is defined. If there exists a function  $g$  such that  $X = g(f(X, Z))$ , then the following chain of inequalities holds:

$$I(X; Y) \leq I(f(X, Z); Y) \leq I((X, Z); Y) \leq I(f(X, Z); Y) + h(Z) - h(Z | X, Y)$$

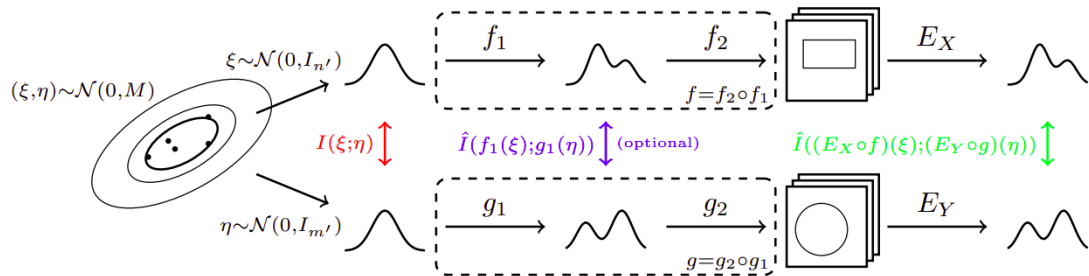
### Lossy compression via an autoencoder $A = D \circ E$

Here quantities can be interpreted as follows:

- 1  $f(X, Z)$  as compressed noisy data,
- 2  $X$  as denoised data,
- 3  $g$  as a perfect denoising decoder,
- 4  $Z$  controls the deviation from the manifold.

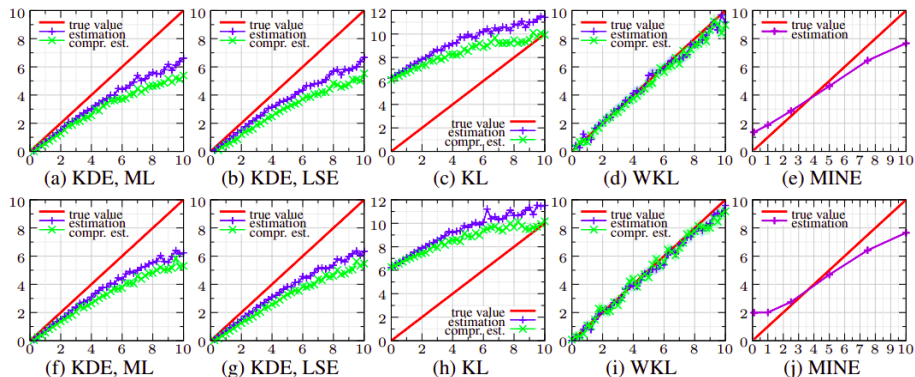


# Experiments: Measure mutual information estimation quality on high-dimensional synthetic datasets



In order to observe and quantify the loss of information caused by the compression step, we split  $f: \mathbb{R}^{n'} \rightarrow \mathbb{R}^n$  into two functions:  $f_1: \mathbb{R}^{n'} \rightarrow \mathbb{R}^{n'}$  maps  $\xi$  to a structured latent representation of  $X$  (e.g., parameters of geometric shapes), and  $f_2: \mathbb{R}^{n'} \rightarrow \mathbb{R}^n$  maps latent representations to corresponding high-dimensional vectors (e.g., rasterized images of geometric shapes). The same goes for  $g = g_2 \circ g_1$ .

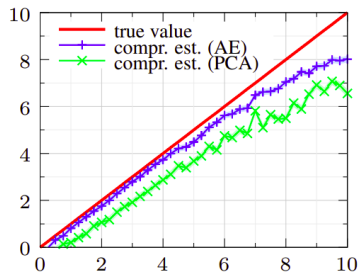
# Results: comparison of different estimators on synthetic image datasets



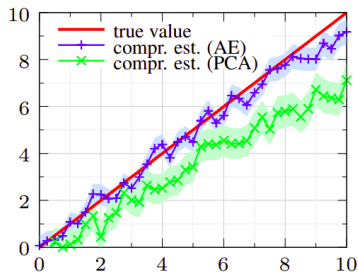
Maximum-likelihood and Least Squares Error KDE, Non-weighted and Weighted Kozachenko-Leonenko, MINE for  $16 \times 16$  (first row) and  $32 \times 32$  (second row) images of rectangles ( $n = m = 4$ ),  $5 \cdot 10^3$  samples. Along x axes is  $I(X; Y)$ , along y axes is  $\hat{I}(X; Y)$ .



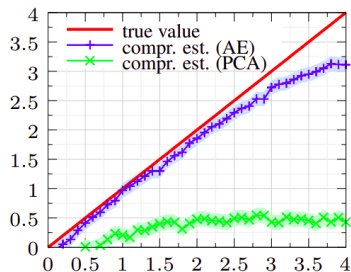
# Results: linear vs nonlinear compression



(a)  $32 \times 32$  images of 2D Gaussians ( $n' = m' = 2$ )



(b)  $32 \times 32$  images of rectangles ( $n' = m' = 4$ )



(c) Highly-nonlinear manifold in  $\mathbb{R}^{32}$  ( $n' = m' = 2$ )

**Figure:** Comparison of nonlinear AE and linear PCA performance in task of MI estimation via lossy compression:  $5 \cdot 10^3$  samples. Along x axes is  $I(X; Y)$ , along y axes is  $\hat{I}(X; Y)$ . WKL entropy estimator is used in these experiments

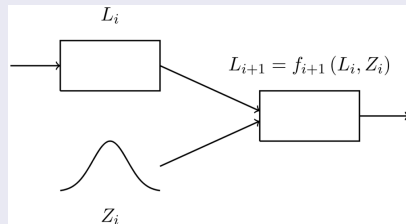
The experiments mentioned above confirm that the non-linearity of the encoder  $E$  is more versatile compared to the the linear compression

# IB Analysis: MI estimation between neural network layers

## The architecture of the MNIST convolution-DNN classifier

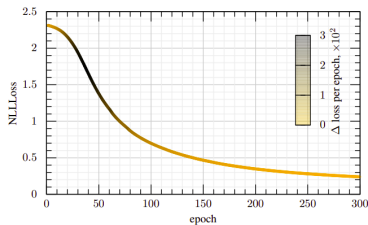
The stochastic modification of a network serves as a proxy to determine the information-theoretic properties of the original model. The stochasticity enables proper MI estimation between layers of the network

- $L_1$ : Conv2d(1, 8, ks=3), LeakyReLU(0.01)
- $L_2$ : Conv2d(8, 16, ks=3), LeakyReLU(0.01)
- $L_3$ : Conv2d(16, 32, ks=3), LeakyReLU(0.01)
- $L_4$ : Dense(32, 32), LeakyReLU(0.01)
- $L_5$ : Dense(32, 10), LogSoftMax

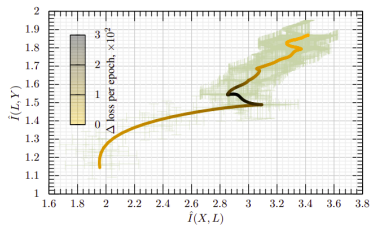


Let's observe corresponding information plane plots for this network...

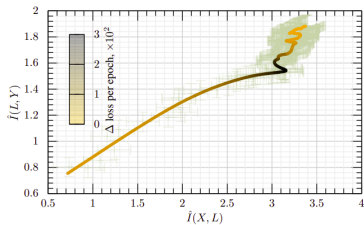
# Results: Information Bottleneck Analysis for the MNIST classifier



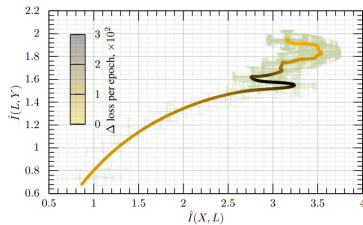
(a) Negative log likelihood loss (train data)



(b)  $L_3$  (convolutional, LeakyReLU)



(c)  $L_4$  (fully-connected, LeakyReLU)



(d)  $L_5$  (fully-connected, LogSoftMax)

Dynamics of information-theoretic quantities during the training of DNNs are indeed non-trivial

# Conclusion & Scientific novelty

## Conclusion

- ① theoretical and practical justifications of the MI estimation via compressed representations have been obtained
- ② the general framework to test conventional mutual information estimators complemented with the proposed lossy compression step and performing IB analysis have been developed
- ③ information plane experiment with the MNIST dataset classifier has been carried out

## Scientific Novelty

- ① the idea of compression is the key novelty of this research
- ② proposed method outperforms existing approaches for the MI evaluation
- ③ Information Bottleneck hypothesis was deeply explored and new MI dynamics dependencies were observed

## Papers

- 1 I. Butakov, A. Tolmachev, S. Malanchuk, A. Neopryatnaya, A. Frolov, K. Andreev **Information Bottleneck Analysis of Deep Neural Networks via Lossy Compression** (published at the ICLR 2024, Poster, A\* Core conference)
- 2 I.D. Butakov, S.V. Malanchuk, A.M. Neopryatnaya, A. D. Tolmachev, K. V. Andreev, S. A. Kruglik, E. A. Marshakov, A. A. Frolov **High-Dimensional Dataset Entropy Estimation via Lossy Compression** // Journal of Communications Technology and Electronics, 2021, № 66, pp. 764–768

## Conferences

- 1 66th All-Russian Scientific Conference of MIPT, April 2024 (oral talk)
- 2 All-Russian Summer School on Machine Learning SMILES-2023, Altai, August 20-31, 2023 (poster session, received “*Best poster*” prize)
- 3 65th All-Russian Scientific Conference of MIPT, April 2023 (oral talk)

## Future plans

- ① our paper devoted to the MI estimation via Normalizing Flows have been submitted to the NeurIPS 2024; the rebuttal phase are expected in July 2024
- ② provide additional theoretical bounds for the MI estimation methods
- ③ explore the Information Bottleneck hypothesis for a broader set of neural networks

## Acknowledgements

I thank my colleagues and research advisor A.A. Frolov for the fruitful and exciting work!