Information Bottleneck Analysis of Deep Neural Networks

Aleksandr Tolmachev

Skolkovo Institute of Science and Technology, Moscow Institute of Physics and Technology

Research advisor: Dr.Sc. Alexey Frolov

June 8, 2024

Introduction/Background

Consider random vectors, denoted as $X : \Omega \to \mathbb{R}^n$ and $Y : \Omega \to \mathbb{R}^m$, where Ω 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.

Entropy definitions

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

Mutual Information definition

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

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

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

Information Bottleneck principle

Information Bottleneck (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:

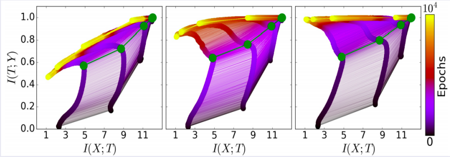
- $I(X;T)$ between the hidden layer output (T) and the DNN input (X)
- $I(Y;T)$ between the hidden layer output (T) and the target of the model (Y)

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

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

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:

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

Loseless case: MI can be measured between loseless compressed representations

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

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

Method: lossy compression case

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

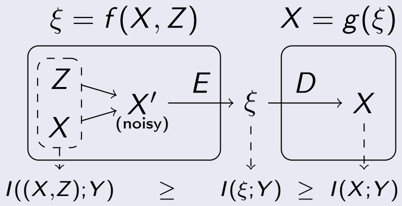
Theorem 2. Let X, Y, and Z be random variables such that $I(X;Y)$ and $I((X,Z);Y)$ are defined. Let f be a function of two arguments such that $I(f(X,Z);Y)$ is defined. If there exists a function g such that $X = g(f(X,Z))$, then the following chain of inequalities holds:

$$I(X;Y) \le I(f(X,Z);Y) \le I((X,Z);Y) \le I(f(X,Z);Y) + h(Z) - h(Z \mid X,Y)$$

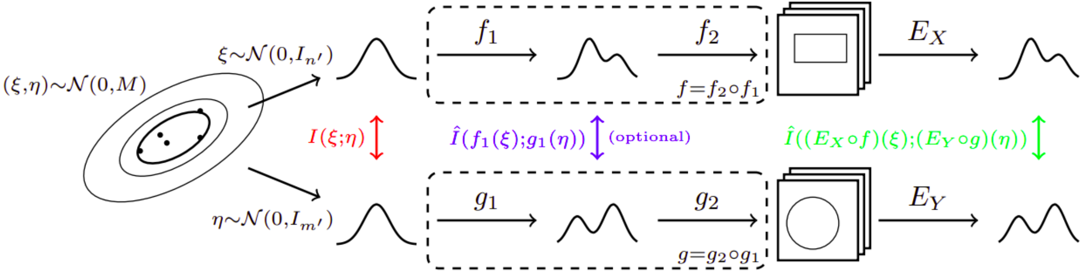
Lossy compression via an autoencoder $A = D \circ E$

Here quantities can be interpreted as follows:

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

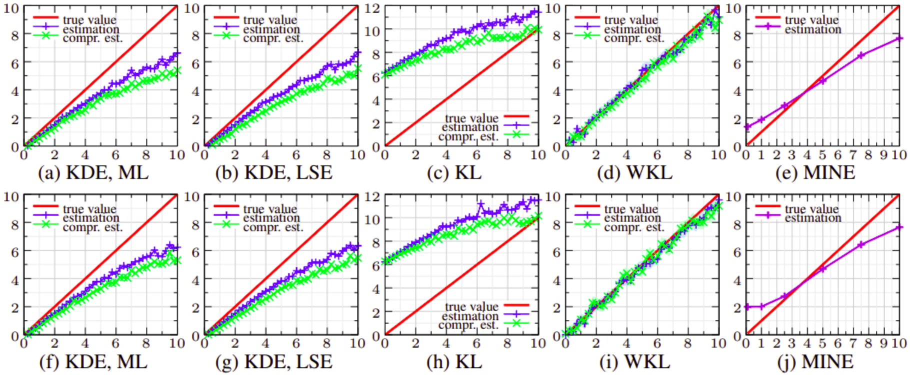


Experiments: Measure mutual information estimation quality on high-dimensional synthetic datasets



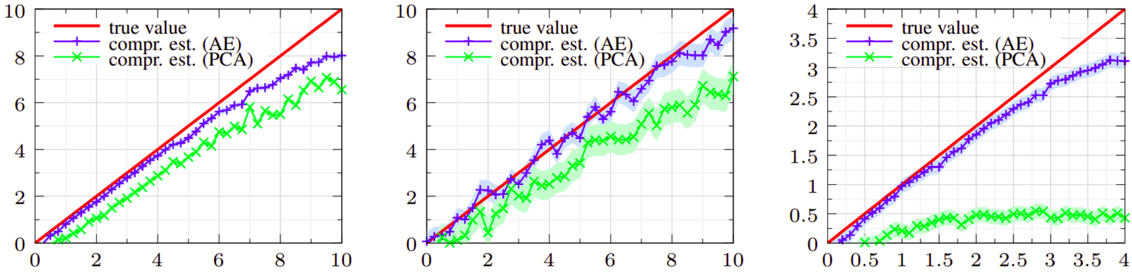
In order to observe and quantify the loss of information caused by the compression step, we split $f : \mathbb{R}^{n'} \to \mathbb{R}^n$ into two functions: $f_1 : \mathbb{R}^{n'} \to \mathbb{R}^{n'}$ maps ξ to a structured latent representation of X (e.g., parameters of geometric shapes), and $f_2 : \mathbb{R}^{n'} \to \mathbb{R}^n$ 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 16×16 (first row) and 32×32 (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 × 32 images of 2D Gaussians ($n' = m' = 2$)

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

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

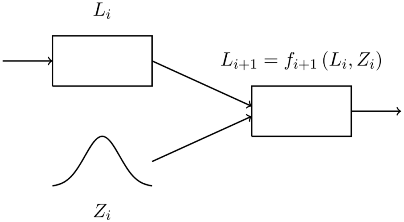
Figure: Comparison of nonlinear AE and linear PCA performance in task of MI estimation via lossy compression: $5 \cdot 10^3$ samples. Along x axes is $I(X;Y)$, along y axes is $\hat{I}(X;Y)$. WKL entropy estimator is used in these experiments

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

The architecture of the MNIST convolution-DNN classifier

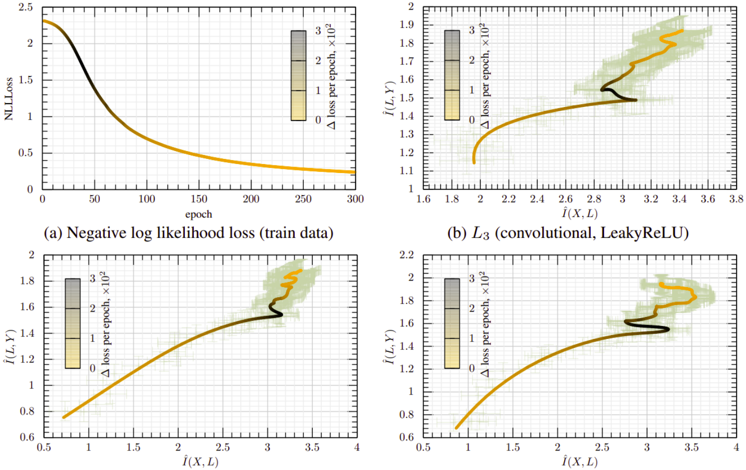
The stochastic modification of a network serves as a proxy to determine the information-theoretic properties of the original model. The stochasticity enables proper MI estimation between layers of the network

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

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

Scientific Novelty

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

Outcomes

Papers

- 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)
- 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

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

Outlook & Acknowledgements

Future plans

- 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
- provide additional theoretical bounds for the MI estimation methods
- 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!

Thank you for your attention!