# Теоретико-информационный анализ нейронных сетей с помощью сжатия данных

Александр Толмачев

Московский физико-технический институт, Сколковский институт науки и технологий,

Научный руководитель: д.ф.-м.н. Алексей Фролов

16 декабря 2023 г.

# Постановка задачи

## Information Bottleneck

Теоретико-информациооный анализ предлагает оценивать динамику двух значений взаимной информации в нейронных сетях: между входом и скрытым слоем, и между скрытым слоем и истинной меткой класса.

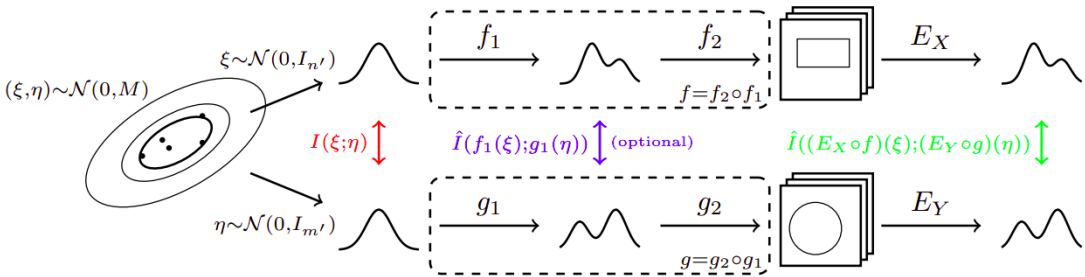
## Проблема

Взаимная информация между многомерными случайными величинами трудно оценивать, поэтому предлагаемые ранее гипотезы об упомянутых выше величинах были проверены либо на маленьких “игрушечных“ сетях, либо на нейронных сетях опредeнного вида.

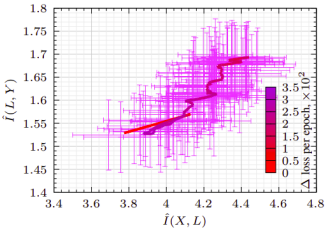
## Предлагаемое решение

В нашем подходе предлагается оценивать взаимную информацию по **сжатым** представлениям, которые получены с помощью обученных автоэнкодеров.

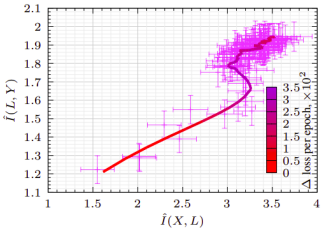
# Схема эксперимента для синтетических данных



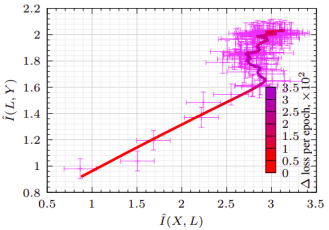
# Результаты анализа обучения CNN-классификатора на датасете MNIST



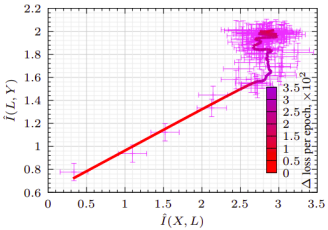

(a) $L_2$ (convolutional, LeakyReLU)




(b) $L_3$ (convolutional, LeakyReLU)




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



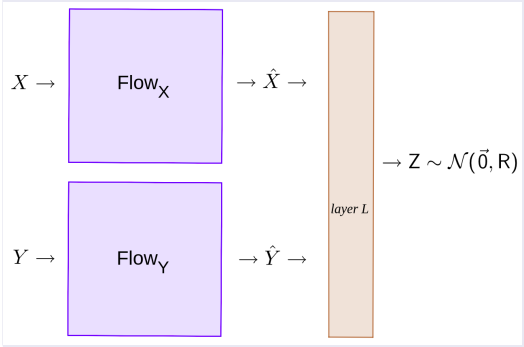

(d) $L_5$ (fully-connected, LogSoftMax)

# Нормализационные потоки для оценки взаимной информации

## Предлагаемый подход

Предлагается оценивать взаимную информацию между случайными величинами $I(X, Y)$ с помощью модели на основе двух норм. потоков и доп. слоя $L$ в "голове" модели

## Предлагаемая модель (в случае $X, Y \in \mathbb{R}^n$)




$$R = \left[\begin{array}{ccc|ccc} 1 & & & \rho & & \\ & 1 & & & \rho & \\ & & \ddots & & & \ddots \\ \hline \rho & & & 1 & & \\ & \rho & & & 1 & \\ & & \ddots & & & \ddots \end{array}\right] \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix}$$

$$\underbrace{\qquad}_{n} \qquad \underbrace{\qquad}_{n}$$

## Модификации для слоя $L$

Слой $L$ добавляется для того, чтобы преобразовать векторы $\hat{X}$ и $\hat{Y}$ так, чтобы вся взаимная информация содержалась только между соответствующими компонентами этих векторов, а компоненты каждого из векторов были бы попарно некоррелированны. Таким образом, выход слоя $L$ можно представить так: $Z = (\hat{X}, A\hat{Y})$, где матрицу $A$ можно подбирать разными способами:

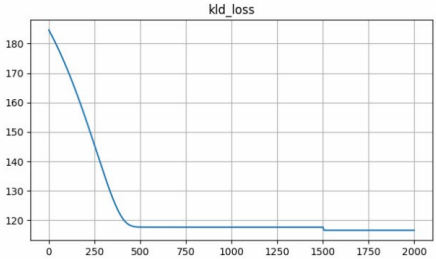
- $cov(\hat{x}, A\hat{y}) = \rho \cdot I_{2n}$
- на основе максимизации правдоподобия: $(\hat{x}, A\hat{y}) \sim \mathcal{N}(0, R)$

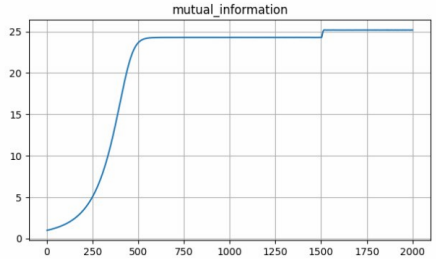
В обоих случаях выражение для матрицы $A$ можно получить аналитически

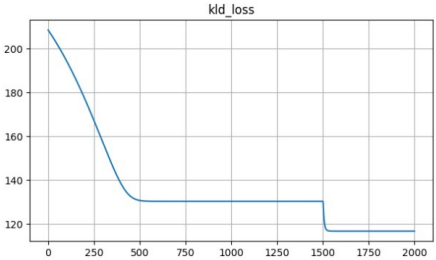
## Преимущества модели

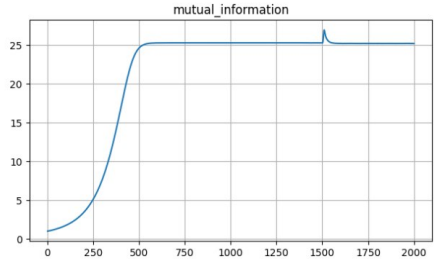
- В таком значение взаимной информации $I(X, Y)$ определяется только параметром $\rho$, который является обучаемым параметром модели
- Стратегия выбора матрицы $A$, описанная выше, применяется только на начальном этапе обучения модели, затем используется обновление весов с помощью Adam-оптимизатора

# Эксперименты: $X, Y \sim (0, I_{50})$; $I(X, Y) = 25$

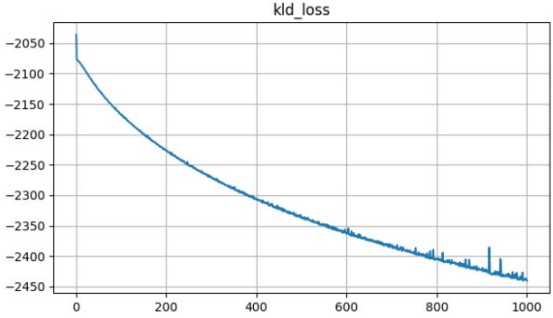


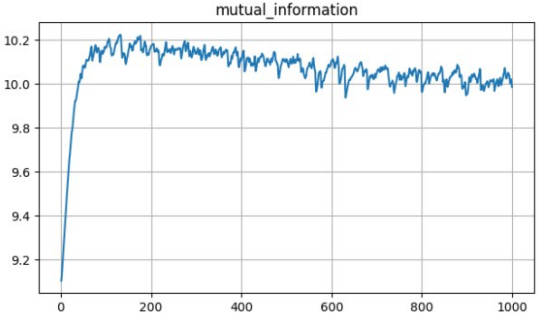



# Эксперименты с изображениями: изображения гауссиан размера $16 \times 16; I(X, Y) = 10; \text{latent dim} = 2$



Спасибо за внимание! Однако, это еще не всё...

# Проект про поиск оптимальных разбиений множеств на части меньшего диаметра

- V.A. Voronov, A.D. Tolmachev, D.S. Protasov, A.M. Neopryatnaya **Searching for distance graph embeddings and optimal partitions of compact sets in Euclidean space** // Mathematical Optimization Theory and Operations Research: Recent Trends. MOTOR 2023. Communications in Computer and Information Science, vol 1881. Springer
- D.S. Protasov, A.D. Tolmachev, V.A. Voronov "Optimal partitions of the flat torus into parts of smaller diameter" (готова, почти подана в журнал)