Информационные процессы, Том 20, № 4, 2020, стр. 397–404.
© 2020 Бутаков, Маланчук, Неопрятная, Толмачев, Андреев, Круглик, Маршаков, Фролов.

МАТЕМАТИЧЕСКИЕ МОДЕЛИ, ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

Оценка энтропии набора данных большой размерности с помощью сжатия с потерями [1]

И. Д. Бутаков[*], С.В. Маланчук[*], А.М. Неопрятная[**], А. Д. Толмачев[*], К.В. Андреев[*,****], С.А. Круглик[***,****], Е. А. Маршаков[***,****], А. А. Фролов[***,****,‡]**

[*] *Московский физико-технический институт (национальный исследовательский университет), Москва, Россия*

[**] *Адыгейский государственный университет, Майкоп, Россия*

[***] *Сколковский институт науки и технологий, Москва, Россия*

[****] *Научно-технологический университет "Сириус", Сочи, Россия*

[‡] *Институт проблем передачи информации, Российская академия наук, Москва, Россия*

Поступила в редколлегию 14.12.2020

Аннотация—В статье предлагается методика оценки энтропии многоразмерного набора данных, основанная на сжатии с помощью автокодировщика. Приведено теоритическое обоснование релевантности данной оценки, а также проведена серия экспериментов на наборах синтетических данных и на данных рукописных цифр MNIST.

КЛЮЧЕВЫЕ СЛОВА: нейронные сети, энтропия набора данных, автокодировщик, ядерная оценка плотности

1. ВВЕДЕНИЕ

В настоящее время область применения искусственного интеллекта и нейронных сетей значительно расширяется. В частности, нейронные сети активно применяются в области компьютерного зрения, улучшения качества изображений и видео, анализа текстов. При этом нейронные сети используются в большинстве случаев как «черный ящик» без какого-либо понимания внутренних механизмов их работы, что в условиях решения нейронными сетями критически важных задач просто недопустимо. Например, для выбора архитектуры модели приходится либо руководствоваться опытом и результатами предыдущих попыток, либо использовать генетические алгоритмы или другие методы машинного обучения [1–3]. В результате обучения нейронной сети выходы промежуточных слоев образуют некоторое внутреннее представление данных, на которых эта нейронная сеть была обучена. Зачастую анализ внутреннего представления данных производится визуально, вручную. К сожалению, существующие подходы к обучению нейронных сетей не дают возможности конструктивно описать критерии качества выбранной архитектуры или информативности внутреннего представления. Таким образом, задача теоретического анализа процесса работы нейронных сетей является крайне важной.

Есть два направления теоретического анализа нейронных сетей: анализ ошибки аппроксимации при использовании глубоких архитектур (см., например, [4, 5]) и анализ процесса обучения. При исследовании процесса обучения нейронной сети в работах [6, 7] строились оценки двух взаимных информаций: между выходом скрытого слоя нейронной сети и меткой класса (такая оценка показывает, насколько хорошо обучился скрытый слой) и между выходом скрытого слоя и входным набором данных (данная оценка показывает, насколько хорошо сеть научилась отбрасывать несущественные признаки).

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов 19-37-51036 и 19-37-90022.

Авторы работы [6] выдвинули гипотезу о том, что обучение нейронной сети проходит в две фазы: подгонки и сжатия. Во время первой фазы обе величины взаимной информации растут, что соответствует построению классификатора на основе наиболее существенных признаков. Во второй фазе вторая величина взаимной информации уменьшается – происходит сжатие внутреннего представления.

Однако, в работах [6, 7] есть принципиальные недочёты. Авторы измеряют взаимную информацию после некоторого количества эпох обучения при фиксированных обучаемых параметрах. В этом случае нейронная сеть является детерминированной функцией, а взаимная информация в таком случае может быть или константой (в случае дискретных распределений входных данных), или бесконечной (в случае непрерывных распределений входных данных). Также отметим, что оценка энтропии производилась путем дискретизации данных, что, в свою очередь, не позволяет достичь нужного уровня точности.

Описанные недостатки были устранены в работе [8], где было предложено, во-первых, добавлять аддитивный гауссовский шум к выходам слоев нейронной сети и, во-вторых, оценивать энтропию не с использованием дискретизации (некорректность данного подхода была проиллюстрирована в той же работе), а с использованием метода ядерной оценки плотности [9], для которого были доказаны оценки сходимости [10]. Отметим, что введенный шум не влияет на качество классификации, но делает вычисления взаимной информации теоретически обоснованным, а зашумленные нейронные сети – это вспомогательный объект, необходимый лишь для анализа исходных нейронных сетей. Авторами работы [8] на примере экспериментов с небольшими нейронными сетями было показано, что упомянутое ранее «сжатие» происходит не во всех случаях. В связи с этим авторы статьи предложили интерпретировать наблюдаемые изменения величин взаимной информации как следствие кластеризации внутреннего представления. В частности, было показано, что при наличии регуляризации, препятствующей кластеризации внутреннего представления, «сжатие» не наблюдается. Стоит также отметить, что метод оценки энтропии, основанный на дискретизации данных, позволяет дать представление о кластеризации оценок меток классов, происходящей в процессе обучения нейронных сетей. Таким образом, было показано, что теоретико-информационный анализ нейронных сетей позволяет получать сведения о том, насколько сильно кластеризовано внутреннее представление нейронной сети и какой вклад в кластеризацию вносят отдельные её компоненты. Тем не менее, авторы работы [8] не смогли наблюдать уменьшение оцениваемой энтропии в процессе обучения свёрточной нейронной сети – классификатора рукописных цифр MNIST. Возникшую проблему авторы объяснили большой размерностью как самого набора данных, так и выходных значений каждого из внутренних слоёв сети. Однако в отношении отдельных нейронов оценка энтропии действительно уменьшалась в процессе обучения. Таким образом, можно сделать вывод, что теоретико-информационный подход применим и в отношении больших нейронных сетей, однако предложенная на данный момент методика не позволяет производить полноценный анализ наборов данных большой размерности.

Известно, что взаимная информация может быть вычислена как разность двух энтропий (см. раздел 2), поэтому в данной работе мы рассматриваем задачу оценки энтропии набора данных большой размерности и предлагаем использовать алгоритмы сжатия с потерями и строить оценки энтропии исходного распределения с помощью оценок энтропии сжатых представлений.

2. ПРЕДВАРИТЕЛЬНЫЕ СВЕДЕНИЯ

В данном разделе мы приведем основные определения и введем необходимые для исследования объекты. Начнем с определения используемых теоретико-информационных величин.

Определение 1. *Дифференциальной энтропией* непрерывного случайного вектора $X \in \mathbb{R}^n$ и имеющего плотность распределения $\rho(x)$, называется величина

$$h(X) = -\mathbb{E}\log\rho(X) = -\int\limits_{suppX} \rho(x) log\rho(x)\,dx, \tag{1}$$

где основание логарифма задаёт единицу измерения введённой величины. Здесь и далее $suppX \subseteq \mathbb{R}^n$ – *носитель* $\rho(X)$.

В дальнейшем изложении под *энтропией* непрерывного случайного вектора будем подразумевать дифференциальную энтропию,причем тут и далее логарифмы имеют натуральное основание.

Определение 2. Пусть $X \in \mathbb{R}^n$ и $Y \in \mathbb{R}^m$ - случайные вектора, имеющие совместную плотность распределения $\rho(x, y)$, и носитель Y не пуст. Тогда *условной дифференциальной энтропией* X при условии Y называется величина

$$h(X \mid Y) = -\mathbb{E}\log\rho(X \mid Y) = -\int\limits_{suppX}\int\limits_{suppY} \rho(x, y)\log\rho(x \mid y)\,dx\,dy. \tag{2}$$

При этом

$$\forall y \in suppY \quad \rho(x \mid y) = \frac{\rho(x, y)}{\rho(y)}. \tag{3}$$

Определение 3. Пусть случайные векторы X и Y удовлетворяют условиям определения 2. Тогда *взаимной информацией* между X и Y называется следующая величина:

$$I(X;Y) = \mathbb{E}\left[\log\frac{\rho(X, Y)}{\rho(X)\rho(Y)}\right] = \int\limits_{suppX}\int\limits_{suppY} \frac{\rho(x, y)}{\rho(x)\rho(y)}\,dx\,dy. \tag{4}$$

Замечание 1. Пусть случайные векторы X и Y удовлетворяют условиям определения 2. Тогда, используя равенства (2), (3) и (4), а также свойство линейности математического ожидания и свойства логарифма, получаем [11]

$$I(X;Y) = h(X) - h(X \mid Y) = h(Y) - h(Y \mid X). \tag{5}$$

3. АВТОКОДИРОВЩИК

Теперь перейдем непосредственно к автокодировщику [12]. Для того, чтобы получить границу на энтропию изначального датасета, мы будем добавлять шум к сгенерированным латентным представлениям. Автокодировщик можно представить как следующую марковскую цепь

$$X \xrightarrow{enc} T \to T' \xrightarrow{dec} X', \tag{6}$$

где X – пространство изначального набора данных; T – пространство латентных представлений, полученное в результате сжатия; $T' = T + Z$, где Z – аддитивный гауссовский шум; X' – пространство восстановленных представлений изначального набора данных, а преобразования enc и dec задаются кодирующей и декодирующей частями автокодировщика. Теперь сформулируем теоретическую оценку, на которую мы будем опираться в рамках данной работы.

Лемма 1. *Пусть в обозначениях* (6) $\dim X = \dim X' = d$, $\dim T = \dim T' = \dim Z = d_l^*$. *Более того, пусть* $Z \sim \mathcal{N}(0, \sigma^2 I_{d_l^*})$ *Тогда верна следующая оценка на* $h(X)$*:*

$$h(X) \leq h(T') - d_l^* \left(c + \frac{\log \sigma^2}{2} \right) + d \left(c + \frac{\log D}{2} \right), \tag{7}$$

где $c = \frac{\log(2\pi e)}{2}$ *и* $D = \frac{1}{d}\mathbb{E}\left[\|X - X'\|^2\right]$.

Доказательство. Используя классическое неравенство обработки информации для (6), заключаем что

$$I(X, T') \leq I(X, X'). \tag{8}$$

Распишем сначала левую часть

$$I(X, T') = h(T') - h(T'|X) = h(T') - h(Z), \tag{9}$$

так как enc – детерминированное преобразование, заданное нейронной сетью. Теперь правая часть

$$I(X, X') = h(X) - h(X|X'). \tag{10}$$

Комбинируя все неравенства выше, получаем

$$h(X) \leq h(T') - h(Z) + h(X|X'). \tag{11}$$

Осталось оценить $h(Z)$ и $h(X|X')$. Так как $Z \sim \mathcal{N}(0, \sigma^2 I_{d_l^*})$, то заключаем, что $h(Z) = \frac{d_l^*}{2}\log(2\pi e)\sigma^2$. Перейдем к оценке $h(X|X')$

$$h(X|X') = h(X - X'|X') \leq h(X - X') \leq \sum_{i=1}^{d} h\left(X_i - X_i'\right). \tag{12}$$

Для оценки последнего выражения воспользуемся следующей цепочкой неравенств.

$$\sum_{i=1}^{d} h\left(X_i - X_i'\right) \leq \sum_{i=1}^{d} h\left(\mathcal{N}\left(0, D_i\right)\right) = \sum_{i=1}^{d} \frac{1}{2}\log\left(2\pi e D_i\right) \leq \frac{d}{2}\log\left(2\pi e D\right),$$

где $D_i = \mathbb{E}\left[(X_i - X_i')^2\right]$, а $D = \frac{1}{d}\sum_{i=1}^{n} D_i$. Таким образом, величину $h(X|X')$ можно оценить сверху следующим образом.

$$h(X|X') \leq \frac{d}{2}\log\left(2\pi e D\right). \tag{13}$$

Комбинируя выражения (8)–(13), получаем требуемый результат.

Таким образом, Утверждение 1 показывает, что при помощи сжатия можно построить верхнюю оценку энтропии случайной величины, из которой получена выборка. Более того, при подходящем выборе функции потерь, получится минимизировать последний член в правой части (7).

Algorithm 1 Оценка энтропии набора данных посредством сжатия автокодировщиком

Вход: $S = \{s_i\}_{i=1}^N \subseteq \mathbb{R}^d$ — набор данных из N элементов, $\mathcal{A} = \mathcal{D} \circ \mathcal{E}$ — архитектура автокодировщика (здесь $\mathcal{E}$ и $\mathcal{D}$ обозначают кодирование и декодирование исходного набора данных соответственно; далее d_l^* — размерность внутреннего представления автокодировщика), $\mathcal{P}_{AE}$ — параметры обучения автокодировщика, $\mathcal{P}_{CV} = (b_1, b_2, k, \varepsilon)$ — параметры поиска окна для ядерной оценки плотности, σ — стандартное отклонение каждой из компонент шума, применяемого к внутреннему представлению данных автокодировщиком.

Получить обученную по $(\mathcal{A}, \mathcal{P}_{AE}, S)$ модель: $A = \texttt{fit}(\mathcal{A}, \mathcal{P}_{AE}, S)$.

Получить кодировщик $E : \mathbb{R}^d \to \mathbb{R}^{d_l^*}$, $A = D \circ E$.

Получить множество кодов входных данных: $C = \{c_i\}_{i=1}^N = E(S) \subseteq \mathbb{R}^{d_l^*}$.

Линейно отцентрировать и отнормировать распределение кодов: $C \to C' = \{c_i'\}_{i=1}^N$, $\mathbb{E}C' = 0$, $cov(C') = I$.

Получить гауссовский шум: $N_l = \{\xi_i\}_{i=1}^N \subseteq \mathbb{R}^{d_l^*} : \forall i\ \xi_i \sim \mathcal{N}(0, \sigma^2 \cdot I_{d_l^*})$.

Получить набор зашумлённых кодов $C'' = \{c_i'' \mid c_i'' = c_i' + \xi_i\}_{i=1}^N$

Для множества C'' подобрать оптимальную ширину окна для ядерной оценки плотности с гауссовским ядром методом k-кратной перекрёстной проверки с начальным предположением — отрезком $[b_1, b_2]$ и точностью ε.

$$b^* = \texttt{search_cv}(C'', [b_1, b_2], k, \varepsilon)$$

По полученной оценке оптимальной ширины окна b^* построить ядерную оценку плотности с гауссовским ядром:

$$\hat{\rho}(x) = \frac{1}{N} \sum_{i=1}^{N} (\delta_{c_i''} * \mathcal{N}(0, b^*))(x)$$

$$\hat{\rho}_k(x) = \frac{1}{N-1} \sum_{\substack{i=1 \\ i \neq k}}^{N} (\delta_{c_i''} * \mathcal{N}(0, b^*))(x)$$

Произвести оценку энтропии кодов методом «убрать один элемент»:

$$\hat{h} = -\frac{1}{N} \sum_{k=1}^{N} \ln \hat{\rho}_k(c_k'')$$

Выход: $\hat{h}$.

4. АЛГОРИТМ ОЦЕНКИ ЭНТРОПИИ НАБОРА ДАННЫХ БОЛЬШОЙ РАЗМЕРНОСТИ

В рамках задачи оценки энтропии набора данных большой размерности нами предложен следующий метод: обучение автокодировщика, получение набора кодов – внутренних представлений исходных данных, центрирование и нормирование полученного набора, добавление к нему гауссовского шума, подбор окна для ядерной оценки плотности с гауссовским ядром посредством перекрёстной проверки, построение ядерной оценки плотности с подобранным окном, оценка энтропии методом «убрать один элемент»– Leave One-Out (LOO) [10, 13]. Формализуя, получаем Алгоритм 1.

Описание конкретного вида функций `fit` и `search_cv` выходит за рамки данной работы. Выбор архитектуры автодировщика $\mathcal{A}$ при этом основывается на предполагаемых свойствах набора данных, а параметры b_1, b_2 и ε , в свою очередь, при правильном выборе реализации `search_cv` влияют лишь на скорость и точность поиска b^*.

Параметр k следует брать равным N, а при выборе параметра d_l^* исходить из истинной внутренней размерности набора данных d_l. Оставшийся параметр σ подбирается эмпирически на основе визуального анализа работы автокодировщика и значениях функции потерь.

5. ЭКСПЕРИМЕНТЫ

В данной секции представлены результаты экспериментов работы Алгоритма 1 на различных наборах данных. Также тут представлены наблюдения о связи между изломами на графиках оценки энтропии и подсчета фрактальной размерности набора данных.

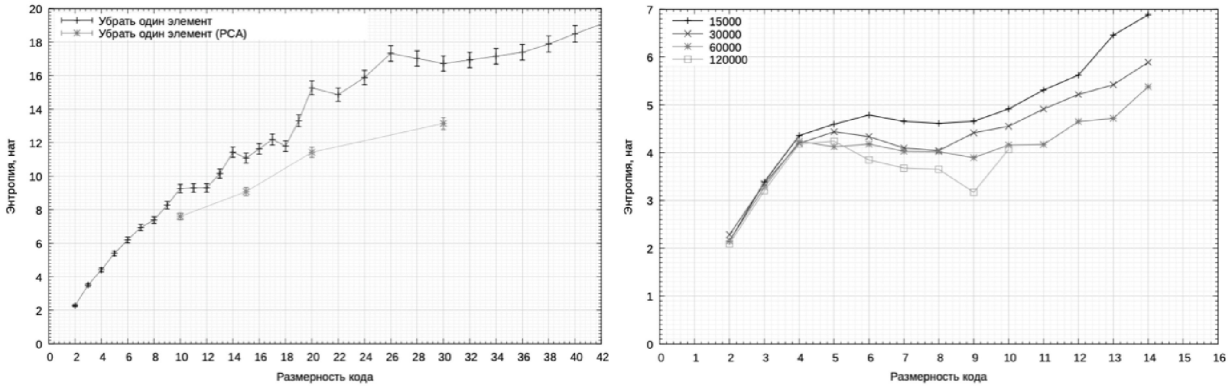

Рис. 1. Зависимость энтропии от размерности внутреннего представления. Слева – набор данных MNIST, справа – набор синтетических данных с разными размерами выборки N.

5.1. Синтетические данные

Алгоритм генерации синтетического набора данных можно описать следующим образом: происходит получение выборки объема N из гауссовского распределения размерности d_l с единичной матрицей ковариации. Далее к полученным точкам применяется некоторая гладкая функция $f : \mathbb{R}^{d_l} \to [-1, 1]^d$. После к полученным значениям добавляется некоторый гауссовский шум размерности d с матрицей ковариации $\sigma_{fn}^2 \cdot I_d$, где параметр σ_{fn} мал ($\sim 10^{-3}$–10^{-1}) или равен нулю. На Рис. 1 (справа) представлены результаты работы Алгоритма 1 для синтетических данных с внутренней размерностью $d_l = 4$, вложенных в пространство размерности $d = 16$ с помощью следующей функции f:

$$y = f(x): \quad y_i = \sin\left(\pi \cdot \tanh\left[\left(-3x^3_{(i/d_l^2) \bmod d_l} + 2x^2_{(i/d_l) \bmod d_l} + 4x_{i \bmod d_l}\right)/10\right]\right), \qquad (14)$$
$$x \in \mathbb{R}^{d_l},\ y \in \mathbb{R}^d.$$

Стоит отметить, что на графике наблюдается излом именно в окрестности $d_l^* = d_l$, и что более того, графики роста энтропии до этого излома идентичны для различных размеров набора данных.

5.2. Набор данных MNIST

Помимо синтетических данных была проведена серия экспериментов на наборе данных рукописных цифр MNIST. На Рис. 1 (слева) представлены результаты работы классического метода главных компонент (PCA) и предложенного Алгоритма 1. Отметим, что оценка энтропии с помощью метода главных компонент оказывается систематически меньше, чем оценка с помощью сжатия автокодировщиком, хотя и сравнительна близка с ней. Можно прелположить, что это связано с качеством сжатия, а как следствие с потерей информации о внутренней структуре данных.

5.3. Фрактальная размерность

В данной секции представлены результаты экспериментов оценки внутренней размерности набора данных. Для начала дадим определение фрактальной размерности.

Определение 4. Фрактальная (корреляционная) размерность некоторого множества X определяется следующим образом:

$$d_{corr} = \lim_{\varepsilon \to 0} \frac{\ln C(\varepsilon)}{\ln \varepsilon}, \tag{15}$$

$$C(\varepsilon) = \lim_{N \to \infty} \frac{2P_N(\varepsilon)}{N(N-1)},$$

где $C(\varepsilon)$ — корреляционный интеграл, $N = |S_N|$ — число точек множества $S_N \subseteq X$, а $P_N(\varepsilon)$ — число пар точек множества S_N, находящихся на расстоянии не более ε.

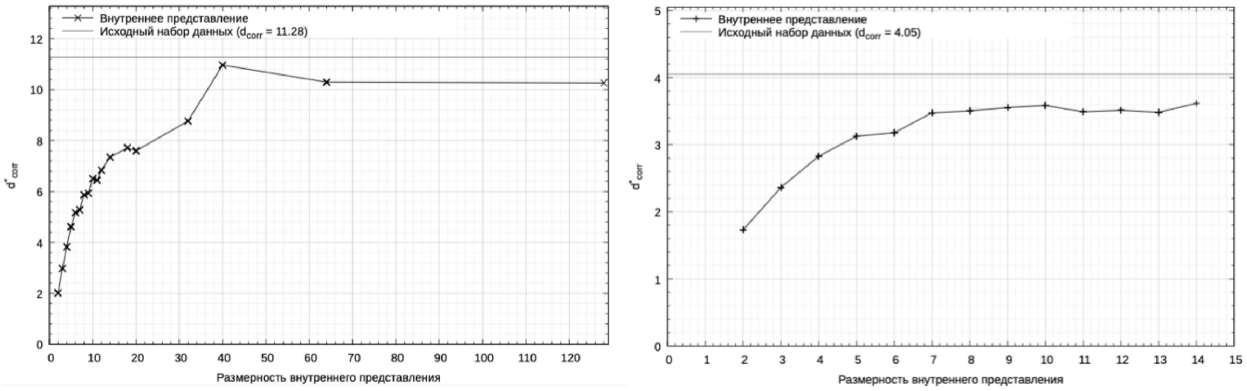

Рис. 2. Фрактальная размерности исходного и сжатых представлений набора данных. Слева – набор данных MNIST, справа – набор синтетических данных.

На Рис. 2 (справа) предаставлены результаты оценки внутренней размерности синтетического набора данных. Из графика видно, что фрактальная размерность сжатого представления стремится к истинной фрактальной размерности набора данных. Кроме того, можно пронаблюдать связь между скоростью роста графика оценки энтропии (Рис. 1, справа) и скоростью роста графика оценки внутренней размерности набора данных.

Похожие результаты можно видеть и на эксперименте с набором данных MNIST. Фрактальная размерность сжатого представления стремится к фрактальной размерности изначального набора данных (Рис. 2, слева), а изломы в графиках происходят в соответствующих друг другу размерностях сжатых представлений.

6. ЗАКЛЮЧЕНИЕ

В работе предложен алгоритм оценки энтропии набора данных большой размерности с помощью сжатия с потерями. Проведены серия экспериментов для оценки энтропии и внутренней размерности для синтетических данных, а также для набора данных рукописных цифр MNIST. В ходе дальнейшей работы планируется исследовать возможность использования данного алгоритма для оценки внутренней размерности набора данных.

Работа основана на результатах, полученных в рамках образовательного модуля “Современные методы теории информации, оптимизации и управления”, прошедшего со 2 по 23 августа 2020 в НТУ Сириус, Сочи, Россия.

СПИСОК ЛИТЕРАТУРЫ

1. H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean, Efficient neural architecture search via parameters sharing, *Proceedings of Machine Learning Research*, vol. 80, pp. 4095-4104, 2018.

2. H. Liu, K. Simonyan, and Y. Yang, DARTS: Differentiable architecture search, *International Conference on Learning Representations*, 2019.

3. B. Wu, X. Dai, P. Zhang, Y. Wang, F. Sun, Y. Wu, Y. Tian, P. Vajda, Y. Jia, and K. Keutzer, Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10726-10734, 2019.

4. P. Grohs, D. Perekrestenko, D. Elbrachter, and H. Bolcskei, Deep neural network approximation theory, *CoRR, vol. abs/1901.02220*, 2019.

5. D. Yarotsky, Error bounds for approximations with deep relu networks,*Neural Networks*, vol. 94, pp. 103-114, 2017.

6. N. Tishby and N. Zaslavsky, Deep learning and the information bottleneck principle, *2015 IEEE Information Theory Workshop (ITW)*, pp. 1-5, 2015.

7. R. Shwartz-Ziv and N. Tishby, Opening the black box of deep neural networks via information,*CoRR, vol. abs/1703.00810*, 2017.

8. Z. Goldfeld, E. Van Den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, Estimating information flow in deep neural networks, *Proceedings of Machine Learning Research*, vol. 97, pp. 2299-2308, 2019.

9. A. B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer Publishing Company, Incorporated, 1st ed., 2008.

10. Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy, Convergence of smoothed empirical measures with applications to entropy estimation, *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4368-4391, 2020.

11. T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*, USA: Wiley-Interscience, 2006.

12. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

13. L. F. Kozachenko and N. N. Leonenko, Sample estimate of the entropy of a random vector, *Problems Inform. Transmission*, vol. 23, pp. 95-101, 1987.

High-dimensional dataset entropy estimation via lossy compression

I.D. Butakov, S.V. Malanchuk, A.M. Neopryatnaya, A.D. Tolmachev, K.V. Andreev, S.A. Kruglik, E.A. Marshakov, A.A. Frolov

This article proposes a method for estimating a high-dimensional dataset entropy based on compressing via an autoencoder and working with the data's resulting code representation. The paper discusses this approach's correctness and hypotheses about using this method for the related problem of estimating the internal dimension of a dataset. The hypotheses are illustrated both by experiments with synthetic data and experiments with the MNIST dataset.

KEYWORDS: neural networks, dataset entropy, autoencoder, kernel density estimation

ACKNOWLEDGMENTS: The research was funded by RFBR, projects numbers 19-37-51036 and 19-37-90022.