Санкт-Петербургский Государственный Университет

«Изучение нейронных сетей с помощью современных методов оценки взаимной информации»

Курсовая работа
студента 3 курса
направления «Математика»
01.03.01
группы 17.Б02-мм
очной формы обучения
Тыщука Кирилла Ильича

Научный руководитель:
Доцент,
Кандидат физико-математических наук,
Николенко Сергей Игоревич

Санкт-Петербург
2020 г.

Содержание

1 Введение

1.1 Нейронные сети

Нейронные сети — это широко используемые в настоящее время модели машинного обучения. В простейшем случае они состоят из слоёв, которые последовательно преобразуют входной вектор X в промежуточные представления T_i, чтобы на выходе получить предсказание $\hat{Y}$. Преобразования данных между каждой парой последовательных слоёв параметризуются весами сети. Веса изменяются так, чтобы минимизировать заданную функцию ошибки, к примеру, показывающую, насколько далеки предсказания сети $\hat{Y}$ от правильного ответа Y. Как правило, оптимизация проводится вариантами стохастического градиентного спуска.

При фиксировании конкретного слоя T, его выход можно рассматривать как специальное представление (эмбеддинг) исходных данных $X \to T$, которое затем используется для получения предсказания $T \to \hat{Y}$. Например, если сеть используется для задачи классификации, то часть $X \to T$ может быть рассмотрена как извлечение из данных некоторого набора признаков (encoder, feature extractor), а часть $T \to \hat{Y}$ — как классификация по этому набору признаков (decoder, classifier). Такое разделение может быть условным, но также может быть продиктовано архитектурой сети. В данной работе рассмотрены сети ResNet[1][2], Inception[3] и Xception[4], в которых используются специальные свёрточные слои для извлечения признаков, специализированные на обработке изображений, и затем полносвязные слои для классификации. Такая архитектура позволяет использовать подход transfer learning, при котором признаки, извлекаемые сетью, натренированной решать одну задачу (например, классификацию изображений Imagenet), используются для решения другой, но схожей задачи.

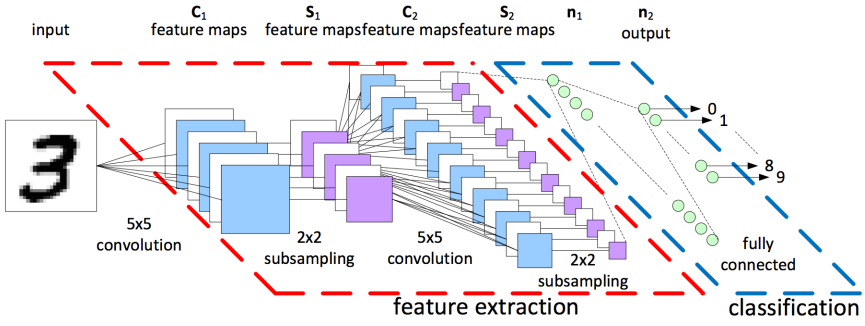


Рис. 1: Пример архитектуры сети с двумя выделенными частями — извлечением признаков из изображения и классификацией

1.2 Взаимная информация

Исследование нейронных сетей в данной работе проводится методами теории информации. Ключевым понятием является взаимная информация (mutual information, MI) между двумя случайными величинами. Она может быть описана в терминах энтропии для дискретных величин и дифференциальной энтропии — для непрерывных, и показывает степень зависимости.

$$I(X;Z) = H(Z) - H(Z|X)$$

Кроме того, взаимная информация выражается через расстояние (расхождение) Кульбака — Лейблера произведения маргинальных распределений величин относительно совместного.

$$I(X;Z) = D_{\mathrm{KL}}(P_{(X,Z)}|P_X \otimes P_Z)$$

Взаимная информация улавливает более тонкие зависимости между величинами, чем коэффициент корреляции, но на практике её значительно сложнее вычислить, особенно в случае высокой размерности изучаемых случайных величин.

1.3 Информация в нейронных сетях

Данная работа также опирается на статьи[5][6] Тишби, в которых взаимная информация используется для характеризации нейронных сетей. Для каждого выхода промежуточного слоя вводятся две характеристики — $I(X;T)$ и $I(T;Y)$, где X, T и Y — случайные величины, соответствующие входным данным сети, выходу слоя и настоящим ответам. Таким образом, $I(X;T)$ показывает, насколько сильно сеть сжала подаваемую ей информацию, а $I(T;Y)$ — сколько нужной для ответа на задачу информации при этом было сохранено. Оптимальным представлением входных данных X с этой точки зрения является такое T, которое бы убирало всю лишнюю информацию, то есть минимизировало $I(X;T)$, но при этом оставляло нужную, то есть максимизировало $I(T;Y)$.

Сложность применения такого подхода заключается в трудоёмкости измерения взаимной информации для больших нейронных сетей. Кроме того, задача нахождения $I(X;T)$ нуждается в уточнении, поскольку в случае непрерывной X

$$I(X;T) = h(T) - h(T|X) = h(T) - (-\infty) = +\infty$$

из-за того, что T является детерминированной функцией от X.

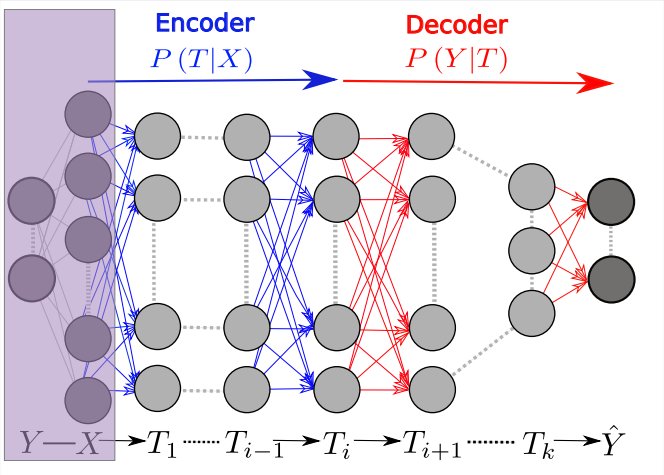


Рис. 2: Рассмотрение сети как совокупности кодировщика и раскодировщика

1.4 Поставленная задача

Основной задачей данной работы является поиск и реализация метода оценки взаимной информации для данных высокой размерности, который позволил бы измерять данную величину для крупных нейронных сетей на примере задач компьютерного зрения.

Работа включает в себя изучение предметной области, реализацию выбранного метода на Tensorflow 2, библиотеке для языка Python, проверку реализации метода и его применение.

2 Обзор статей

2.1 Статьи на смежные темы

В ходе работы были просмотрены статьи, имеющие отношение к измерению взаимной информации, особенно в нейронных сетях. Далее следует кратций обзор некоторых из них.

Saxe et al., On the Information Bottleneck Theory of Deep Learning [7]

В статье утверждается, что выводы Тишби не подтверждаются для сетей с другими функциями нелинейности.

Kolchinsky et al., Nonlinear Information Bottleneck [8]
Предложен новый непараметрический метод для оценки верхней грани на взаимную информацию. Но основе этого метода тренируется нейронная сеть. В сеть добавлен шум, что делает её стохастической, что, в свою очередь, гарантирует ограниченность величины $I(X;T)$.

Amjad, Geiger, How (Not) To Train Your Neural Network Using the Information Bottleneck Principle [9]
Статья также обращает внимание на проблемы с бесконечной информацией. Одним из предложенных решений так же является добавление шума. Кроме того, статья указывает на то, что, инвариантная относительно биекций, взаимная информация может не отражать качество представления данных, полученных нейронной сетью.

Nguyen, Choi, Layer-wise Learning of Stochastic Neural Networks with Information Bottleneck [10]
Послойная тренировка стохастической нейронной сети на основе взаимной информации.

Nash, Kushman, Williams, Inverting Supervised Representations with Autoregressive Neural Density Models [11]
Статья рекомендована соавтором Тишби. Оценка взаимной информации получается как побочный продукт оценки распределения на X при условии T. Открытия Тишби подтверждаются.

Löwe, O'Connor, Veeling, Putting An End to End-to-End: Gradient–Isolated Learning of Representations [12]
Послойная тренировка сети без обратного распространения ошибок, использующая оценки на взаимную информацию.

Goldfeld et al., Estimating Information Flow in Deep Neural Networks [13]
Во вступлении указываются проблемы с измерением взаимной информации, а также критикуется подход Тишби с округлением.

2.2 Избранные статьи

В итоге было решено остановиться на двух статьях: Belghazi et al., MINE: Mutual Information Neural Estimation[14] и Elad, Haviv, Blau, Michaeli, Direct Validation of the Information Bottleneck Principle for Deep Nets[15]

2.2.1 MINE

Избранным методом для измерения взаимной информации стал MINE. В нём используется двойственное представление Донскера — Варадхана для расстояния Кульбака — Лейблера между двумя распределениями:

$$D_{\mathrm{KL}}(\mathbb{P} \mid\mid \mathbb{Q}) = \sup_{F:\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}}[F] - \log(\mathbb{E}_{\mathbb{Q}}[e^F]$$

где супремум берётся по всем функциям, таким, что математические ожидания конечны.

В качестве $\mathbb{P}$ и $\mathbb{Q}$ выступают совместное распределение X и Z и произведение их маргинальных распределений.

В качестве функции F рассматривается дополнительная нейронная сеть, принимающая на входе один экземпляр X и один экземпляр Z и выдающая одно число. При подстановке получается нейронная информационная мера, являющаяся нижней оценкой на взаимную информацию.

$$I_\Theta(X;Z) = \sup_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}_{XZ}}[F_\theta] - \log(\mathbb{E}_{\mathbb{P}_X\otimes\mathbb{P}_Z}[e^{F_\theta}]$$

где F_θ — дополнительная нейронная сеть, параметризованная весами θ. Математические ожидания по распределениям берутся как эмпирические средние по соответствующим выборкам. Выборки из маргинальных распределений могут быть получены отбрасыванием одной из величин из выборки совместного распределения.

Поскольку нас интересует супремум величины, зависящей от F, дополнительная сеть тренируется так, чтобы максимизировать эту величину. Таким образом, по мере обучения дополнительной сети, нейронная информационная мера растёт и всё точнее приближает взаимную информацию.

Algorithm 1 MINE

$\theta \leftarrow$ initialize network parameters
repeat
  Draw b minibatch samples from the joint distribution:
  $(\boldsymbol{x}^{(1)}, \boldsymbol{z}^{(1)}), \dots, (\boldsymbol{x}^{(b)}, \boldsymbol{z}^{(b)}) \sim \mathbb{P}_{XZ}$
  Draw n samples from the Z marginal distribution:
  $\bar{\boldsymbol{z}}^{(1)}, \dots, \bar{\boldsymbol{z}}^{(b)} \sim \mathbb{P}_Z$
  Evaluate the lower-bound:
$\mathcal{V}(\theta) \leftarrow \frac{1}{b}\sum_{i=1}^{b} T_\theta(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) - \log(\frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(\boldsymbol{x}^{(i)}, \bar{\boldsymbol{z}}^{(i)})})$
  Evaluate bias corrected gradients (e.g., moving average):
  $\widehat{G}(\theta) \leftarrow \widetilde{\nabla}_\theta \mathcal{V}(\theta)$
  Update the statistics network parameters:
  $\theta \leftarrow \theta + \widehat{G}(\theta)$
until convergence

Рис. 3: Алгоритм обучения MINE

2.2.2 Direct Validation of the Information Bottleneck Principle for Deep Nets

В данной статье MINE используется для измерения информации в нейронных сетях. Проблема с бесконечной взаимной информацией решена заменой измерения $I(X,T)$ на измерение $I(X;T+\epsilon)$, где ϵ — нормально распределённый шум.

Величина $I(X;T+\epsilon)$ на самом деле показывает компактность представления T:

$$I(X;T+\epsilon) = h(T+\epsilon) - h(T+\epsilon \mid X) = h(T+\epsilon) - h(\epsilon) = h(T+\epsilon) - const$$

поскольку T — детерминировнная функция от X, и X и ϵ независимы.

Для вычисления $I(X;T+\epsilon)$ сети MINE необходимо различать пары $(X, T'+\epsilon)$, где X и T' берутся либо из совместного, либо из маргинальных распределений. Поскольку T, промежуточный слой — это детерминированная функция от X, то наивным методом сделать это было бы сравнение величин $T'+\epsilon$ и $T(X)$, где $T(X)$ — выход промежуточного слоя сети при подаче на вход X. Поэтому авторы статьи предлагают модификацию MINE: Architecture Aware MINE (AA-MINE), которому на вход подаются не X и $T'+\epsilon$, а сразу $T(X)$ и $T'+\epsilon$, или даже $T(X)$ и $T'+\epsilon-T(X)$, что приводит к лучшим оценкам. По сути, это означает копирование первых слоём изучаемой сети ($X \to T$) в начало MINE.

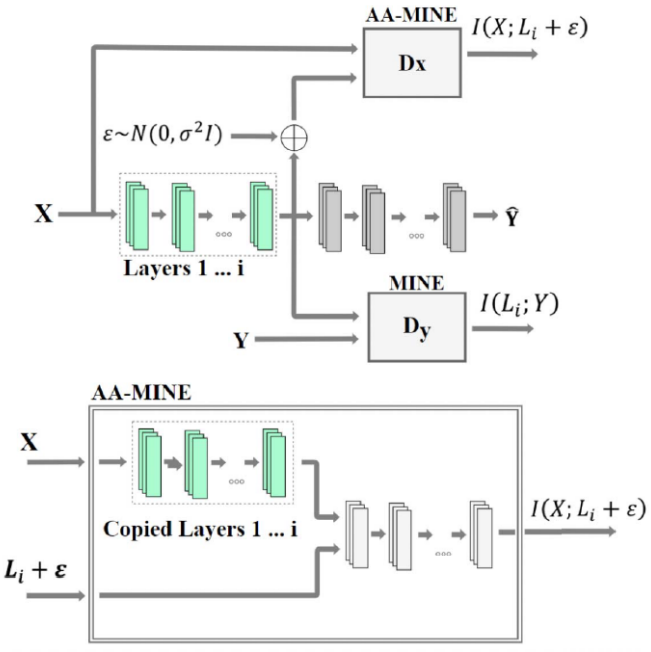


Рис. 4: Схема измерений информации в сети с помощью MINE и AA-MINE. В наших обозначениях L_i — это T_i

3 Работа с кодом

Код для статьи Direct Validation of the Information Bottleneck Principle for Deep Nets был любезно предоставлен её авторами, однако оказался непригоден для нужных экспериментов, поскольку был написан на Tensorflow 1 и таким образом, что для каждой конкретной задачи пришлось бы переписывать существенную его часть. Поэтому было принято решение реализовать MINE и AA-MINE с нуля. Это также позволило добавить корректировку градиентов от смещения (см. секцию 3.2 [14]), предложенную авторами MINE.

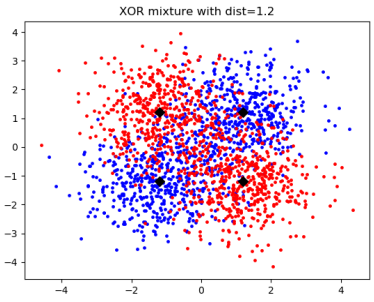
3.1 Смесь

Для проверки работоспособности реализации MINE необходимо было построить величины X и Z таким образом, чтобы плотность их совместного распределения явно задавалась формулой. После этого взаимную информацию можно было найти по определению, используя численное интегрирование.

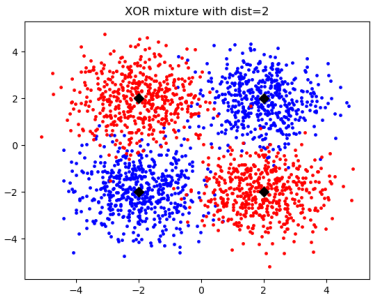
Для этой цели была написана программа, позволяющая генерировать выборки и считать взаимную информацию между величинами X и Z, где X — координаты точки на плоскости, а Z — её метка класса. В каждом классе распределение точек является смесью нормальных распределений.

Нетрудно понять, что чем сильнее перемешаны классы, тем меньше значение $I(X;Z)$. Если же классы практически не перемешаны и имеют равные пропорции, то значение $I(X;Z)$ приближается к $H(Z) = \log(2) \approx 0.693$ нат.

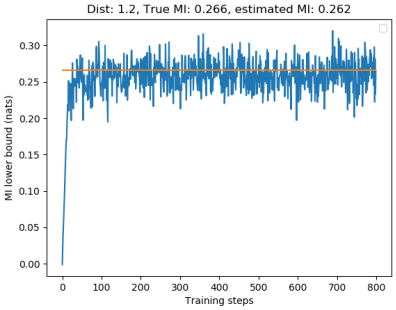
Ниже представлены примеры выборок точек для двух различных смесей, а также процесс обучения MINE, при которой нейронная информационная мера устремляется (в среднем) к правильному значению взаимной информации.



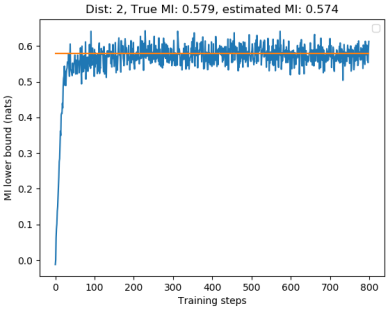
Смесь (а)



Смесь (б)



MINE (а)



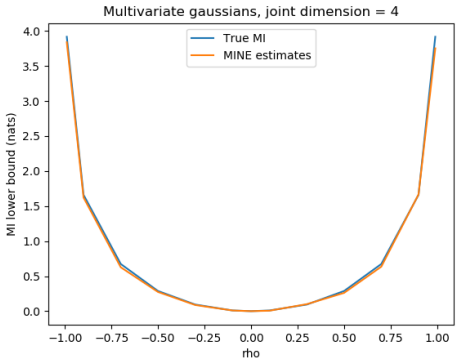
MINE (б)

Рис. 5: Чёрными ромбами отмечены центры нормальных распределений. Оранжевая линия показывает настоящее значение информации. Хотя оценки MINE имеют разброс, в среднем предскаания получаются достаточно точными.

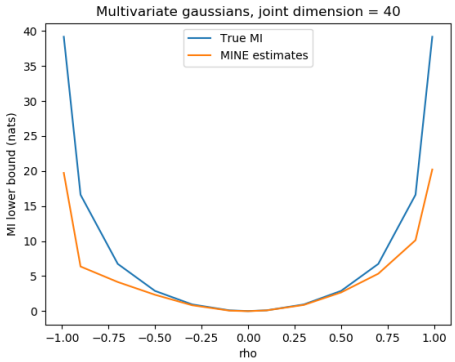
3.2 Тест из статьи MINE

В статье MINE данный метод тестировался на двух многомерных нормальных распределениях, X_a and X_b, с покомпонентными коэффициентами корреляции, $corr(X_a^i, X_b^j) = \delta_{ij}\,\rho$, где $\rho \in (-1, 1)$, а $\delta_i j$ — дельта Кронекера.

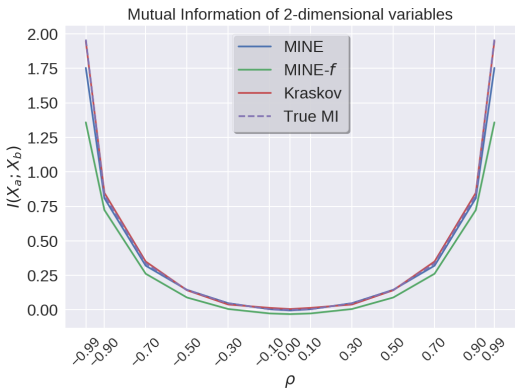
Данный эксперимент был воспроизведён и дал схожие результаты. Ниже представлены графики.



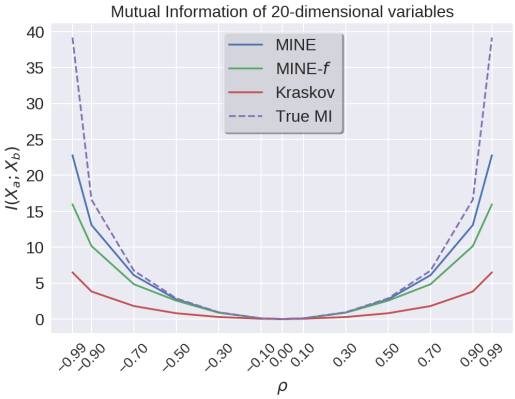
Низкая размерность, мой эксперимент



Высокая размерность, мой эксперимент



Низкая размерность, эксперимент авторов статьи

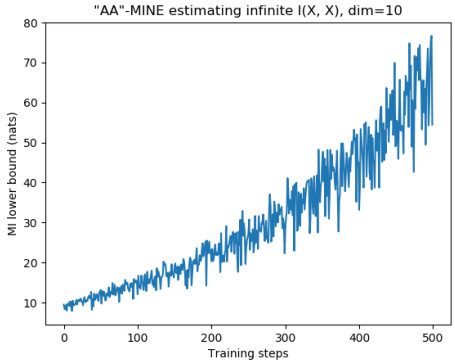


Высокая размерность, эксперимент авторов статьи

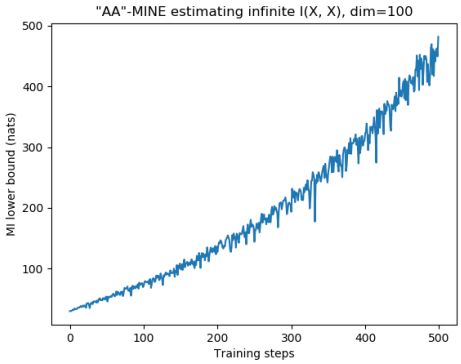
Рис. 6: Сравнение экспериментов с нормальными распределениями. При сильной корреляции точность MINE падает

3.3 Измерение бесконечной информации

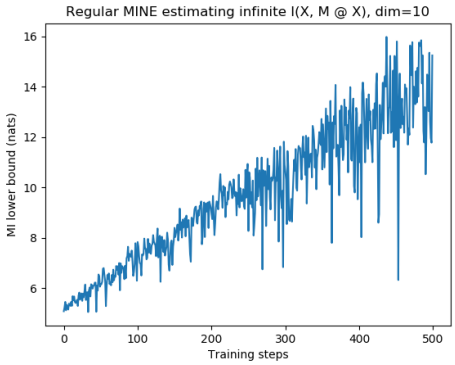
Чтобы продемонстрировать преимущество AA-MINE, был проведён эксперимент, в котором с помощью MINE измерялось значение информации, равное плюс бесконечности. В случае измерения информации между нормально распределённым вектором и им же самим, и в размерности 10. и в размерности 100 оценки MINE уверенно растут. Однако в случае измерения информации между вектором и им же, но домноженным на фиксированную матрицу, в высокой размерности оценки MINE расходятся.



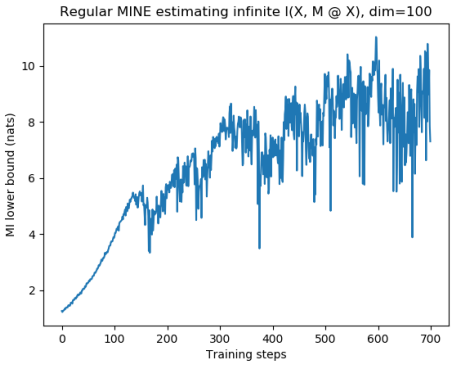
Низкая



Высокая размерность, мой эксперимент



Низкая размерность, эксперимент авторов статьи



Высокая размерность, эксперимент авторов статьи

Рис. 7: Расхождение оценок MINE для бесконечной информации

3.4 Измерение информации в сетях

Проверенная реализация MINE была использована для измерения информации в нейронных сетях.

Сначала информация измерялась на небольшой сети из статьи Тишби. Замеры производились по ходу обучения сети для каждого слоя. Поскольку такие измерения требуют существенных вычислительных ресурсов, большого разрешения добиться не удалось. Однако на графике видно, как по ходу тренировки сеть учится сохранять всё больше информации о правильном ответе. График представлен ниже.

Однако больший интерес представляет следующий вопрос: связаны ли значения взаимной информации с качеством эмбеддингов, полученных сетями? И если связаны, то какие их них более важные? В качестве второго эксперимента были рассмотрены сети ResNet50[1], ResNet50V2[2], ResNet101[1], ResNet101V2[2], InceptionV3[3] и Xception[4], предобученные на Imagenet, с помощью которых были получены эмбединги для датасета «cats vs dogs». На полученных эмбеддингах были обучены одинаковые классификаторы. Затем была измерена взаимная информация между парами «входное изображение,

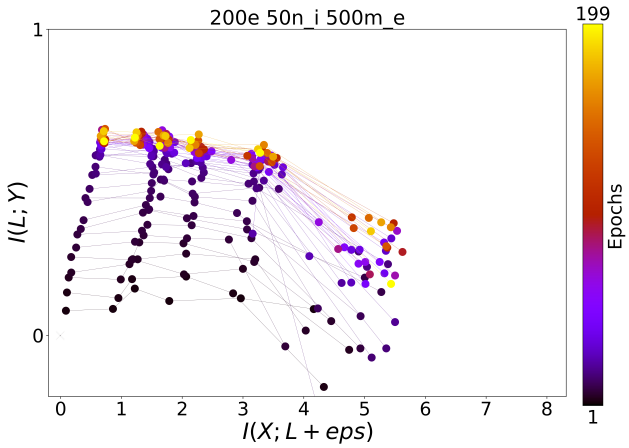


Рис. 8: Эпохи обучения показаны цветом. Для каждой из выбранных эпох на графике отмечена ломаная из пяти звеньев, где каждая вершина соответствует слою сети. Каждая точка имеет координаты $I(X;T)$ и $I(T;Y)$

эмбеддинг плюс шум» (с помощью AA-MINE), «предсказание классификатора и настоящая метка класса», «эмбеддинг и настоящая метка класса». Измерения проводились в трёх сценариях: недообученные сети, сети с оптимальным результатом, переобученные сети. Эмбеддинги и предсказания были также нормализованы по L_2 и L_∞ нормам соответственно.

Ниже представлены матрицы корреляций между точностью классификации (метрика accuracy), функцией ошибки сети (бинарная перекрёстная энтропия) и измеренными величинами взаимной информации по разным архитектурам сетей. Здесь E — эмбеддинг, $\hat{E}$ — эмбеддинг с шумом, P — предсказание классификатора.

	accuracy	loss	$I(X;\hat{E})$	$I(P;Y)$	$I(E;Y)$
accuracy	1.000000	-0.955969	-0.574027	0.633033	0.557338
loss	-0.955969	1.000000	0.496188	-0.400506	-0.526888
$I(X;\hat{E})$	-0.574027	0.496188	1.000000	-0.603023	-0.156921
$I(P;Y)$	0.633033	-0.400506	-0.603023	1.000000	0.581017
$I(E;Y)$	0.557338	-0.526888	-0.156921	0.581017	1.000000

Таблица 1: Корреляции после 1 эпохи обучения.

	accuracy	loss	$I(X;\hat{E})$	$I(P;Y)$	$I(E;Y)$
accuracy	1.000000	-0.387479	-0.603636	-0.145025	0.057250
loss	-0.387479	1.000000	0.602016	0.382073	-0.289438
$I(X;\hat{E})$	-0.603636	0.602016	1.000000	0.210973	-0.156921
$I(P;Y)$	-0.145025	0.382073	0.210973	1.000000	-0.669992
$I(E;Y)$	0.057250	-0.289438	-0.156921	-0.669992	1.000000

Таблица 2: Корреляции после 1 эпох обучения.

	accuracy	loss	$I(X;\hat{E})$	$I(P;Y)$	$I(E;Y)$
accuracy	1.000000	-0.858936	0.093812	0.898035	0.297414
loss	-0.858936	1.000000	0.081825	-0.634049	-0.320510
$I(X;\hat{E})$	0.093812	0.081825	1.000000	-0.112589	-0.156921
$I(P;Y)$	0.898035	-0.634049	-0.112589	1.000000	0.195581
$I(E;Y)$	0.297414	-0.320510	-0.156921	0.195581	1.000000

Таблица 3: Корреляции после 200 эпох обучения.

Наиболее устойчивым показателем качества оказалась величина $I(X,\hat{E})$, показывающая компактность полученного эмбеддинга. Ниже также показано изменения данных величин по мере тренировки сетей.

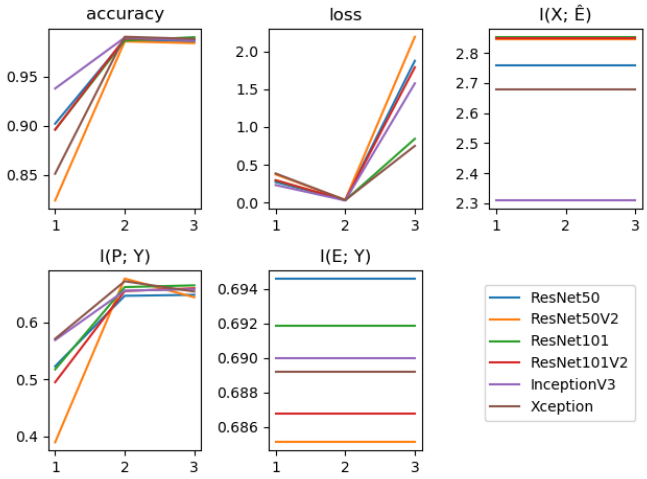


Рис. 9: Величины, не зависящие от предсказаний классификатора, не меняются по мере его обучения и представлены для сравнения. Величина $I(E,Y)$ измерена с погрешностью в силу существенно разных размерностей E (2048) и Y (1).

4 Заключение

Теория информации в глубоком обучении является достаточно запутанной темой с множеством противоречий и отсутствием единого подхода для решения задач. Не смотря на это, в результате работы был найден и реализован метод, позволяющий с достаточной точностью оценивать взаимную информацию в сценариях, которые представляют интерес. Метод удалось протестировать на специально сконструированных примерах и применить к реальной задаче. В результате была сформирована гипотеза о том, что самой информативной метрикой качества эмбеддинга является его компактность, выраженная величиной $I(X, \hat{E})$.

Список литературы

[1] He et al., Deep Residual Learning for Image Recognition

[2] Tumer et el., Identity Mappings in Deep Residual Networks

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, Rethinking the Inception Architecture for Computer Vision

[4] François Chollet Xception: Deep Learning with Depthwise Separable Convolutions

[5] Ravid Shwartz-Ziv, Naftali Tishby, Opening the Black Box of Deep Neural Networks via Information

[6] Naftali Tishby, Noga Zaslavsky, Deep Learning and the Information Bottleneck Principle

[7] Saxe et al., On the Information Bottleneck Theory of Deep Learning

[8] Kolchinsky et al., Nonlinear Information Bottleneck

[9] Amjad, Geiger, How (Not) To Train Your Neural Network Using the Information Bottleneck Principle

[10] Nguyen, Choi, Layer-wise Learning of Stochastic Neural Networks with Information Bottleneck

[11] Nash, Kushman, Williams, Inverting Supervised Representations with Autoregressive Neural Density Models

[12] Löwe, O'Connor, Veeling, Putting An End to End-to-End: Gradient-Isolated Learning of Representations

[13] Goldfeld et al., Estimating Information Flow in Deep Neural Networks

[14] Belghazi et al., MINE: Mutual Information Neural Estimation

[15] Elad, Haviv, Blau, Michaeli, Direct Validation of the Information Bottleneck Principle for Deep Nets