Математика и ИИ: проблемы и задачи, которые требуют решения

И. В. Оселедец

CEO, AIRI

Профессор и руководитель лаборатории @Skoltech

План доклада

- Краткие определения
- Выразительность (expressive power) и Проблема обобщаемости (generalization)
- Теория аппроксимации с помощью нейросетей
- Теория оптимизационных методов
- Геометрические подходы в ИИ (гиперболическая геометрия, спектральные методы)
- Большие языковые модели и их свойства

Что такое искусственный интеллект

ИИ-технология создания интеллектуальных машин компьютерных (Джон Маккарти).

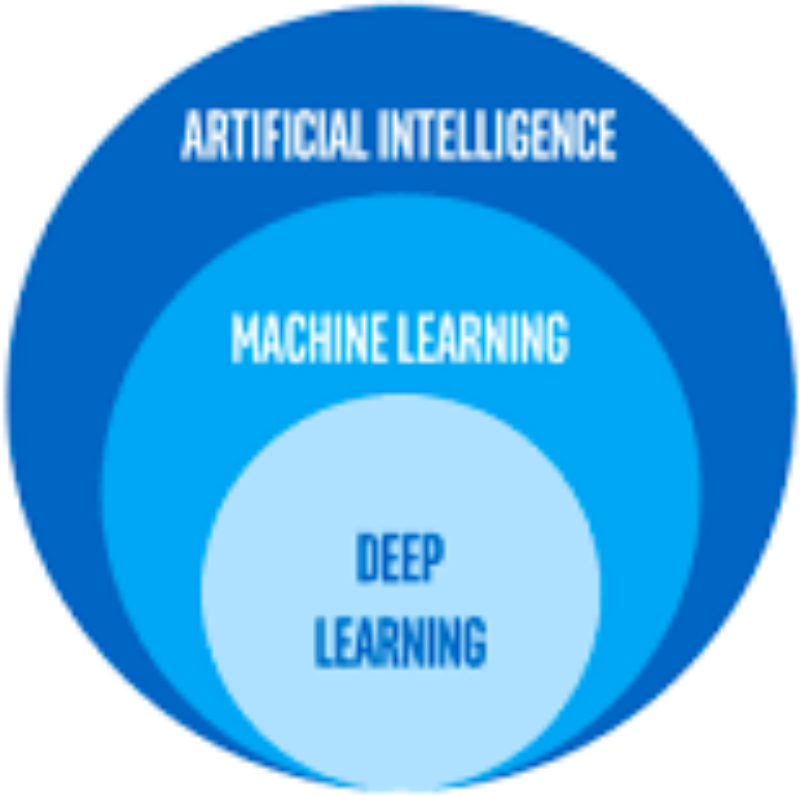
Следует отличать ИИ от общего искусственного интеллекта (AGI, Artificial General Intelligence).

Что такое искусственный интеллект

Искусственный интеллект: любая технология, использующая машины для имитации человеческого интеллекта (включая логику наборы правил, машинное обучение и глубокое обучение.

Машинное обучение: подмножество технологий ИИ, основанное на статистических техниках. Включает глубокое обучение.

Глубокое обучение: алгоритмы на глубоких (многослойных) нейросетях.



Как обучают ИИ

Стандартный подход: обучение с учителем.

Задана обучающая выборка $\{x_i, y_i\}, i = 1, \dots, m$, мы хотим выучить модель (алгоритм),

который аппроксимирует

эту зависимость:

$$f(x_i, \theta) = \hat{y}_i \approx y_i$$

Необходимо задать класс моделей

$f(x, \theta)$ и функцию потерь

$l(y, \hat{y})$ которая определяет, насколько мы ошиблись (риск).

Общая задача оптимизации имеет вид

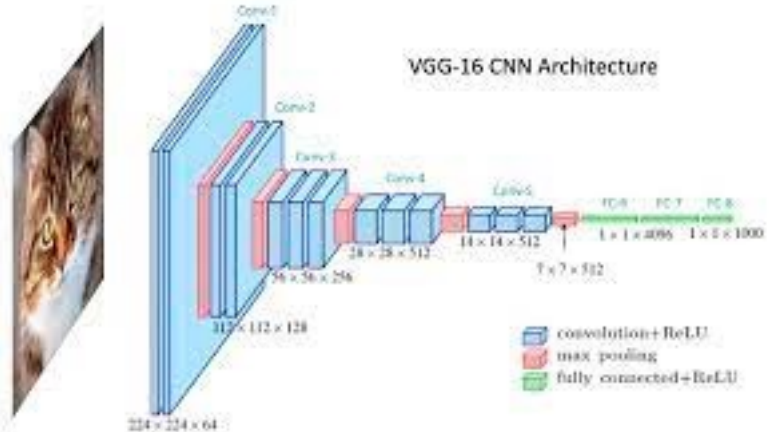
$$L(\theta) = \sum_{i=1}^{m} l(y_i, \hat{y}_i) \to min$$

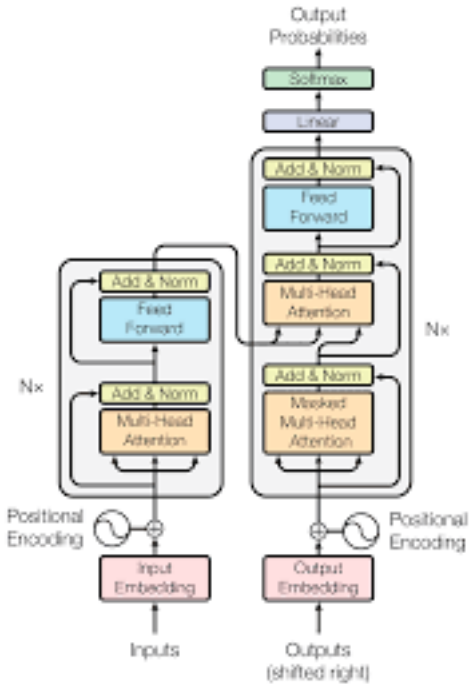
Что используется в качестве моделей ИИ

На данный момент, стандартным классом моделей являются **глубокие нейросетевые модели** (deep neural network models).

Они прекрасно работают на задачах классификации изображений.

Для задачи генерации текстов наибольшую популярность получили модели на основе архиктектуры **трансформеров.**



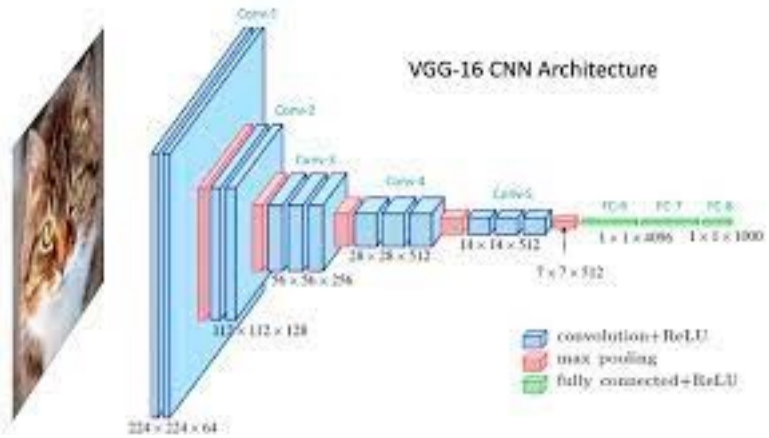


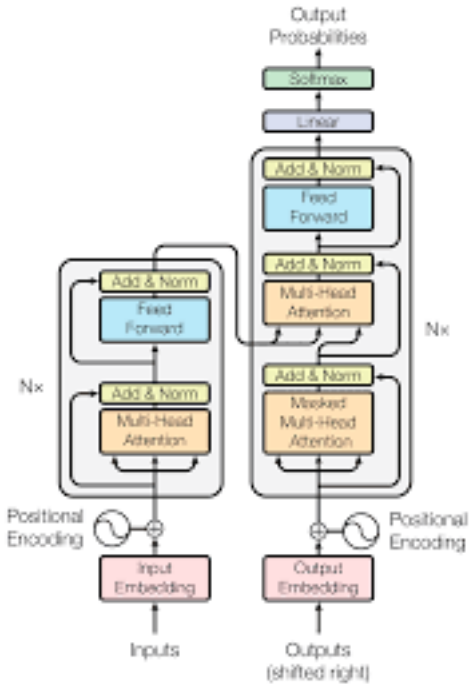
Идея: композиция простых функций

В основе идеи построения «общих» моделей лежит идея композиции функций:

$$f(x) = f_k(f_{k-1})f_{k-2} \cdots f_1(x_1))$$

В качестве блоков используются линейные преобразования, поточечные нелинейные преобразования и более сложные, такие как **нормализация** и блок внимания.





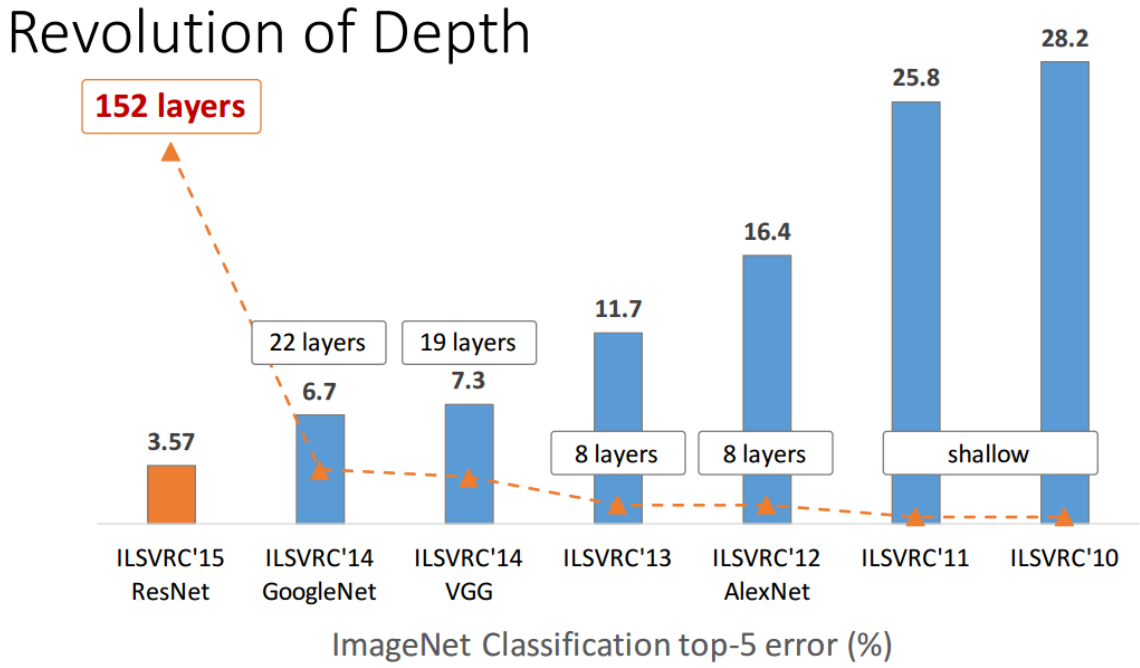
Классическая теория статистического машинного обучения

Если у нас зафиксирован класс модели $f(x, \theta)$ мы можем рассмотреть класс функций, который можно представить (выразительность, expressive power).

Первый вопрос: Верно или нет, что более глубокие архитектуры более выразительные? Ответ во многих постановках «Да».

Второй вопрос: Если размер обучающей выборки фиксирован, то как ведет себя предсказание модели?

Классическая теория говорит «Нет», но все эксперименты показывают обратное!



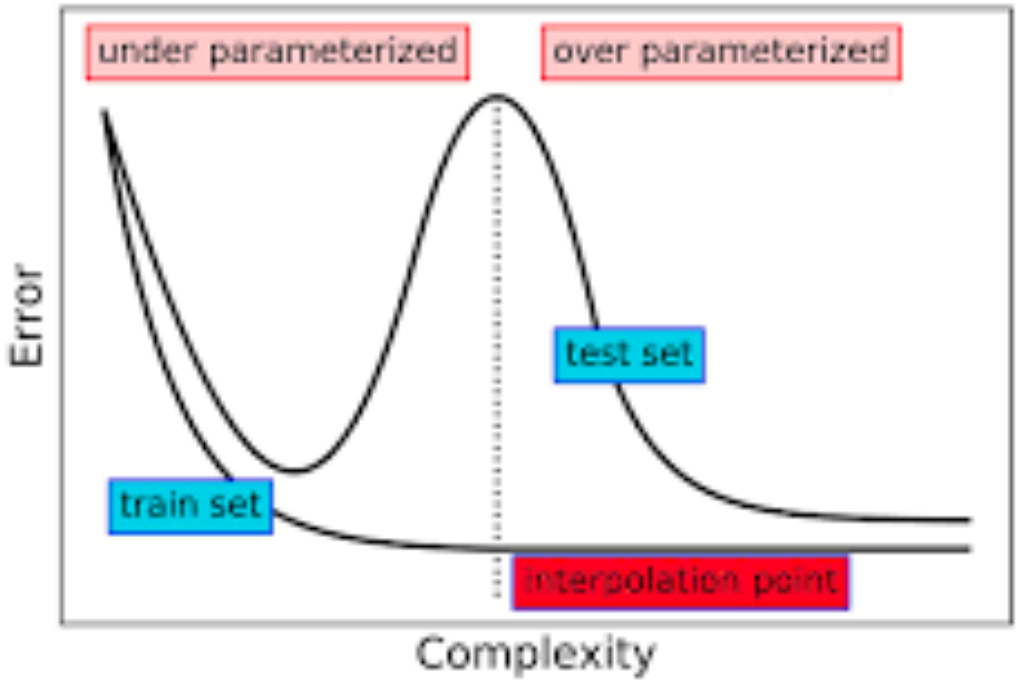
Эффект double descent

Достаточно уверенно в современных моделях наблюдается эффект double descent — зависимость точности на тестовой выборки от числа обучаемых параметров.

Кажется, что слишком много параметров — это не так плохо!

Более того, часто это помогает.

Основной способ исследования:
Neural Tangent Kernel



Neural Tangent Kernel

Из спектральных свойств ядра, можно получить асимптотику сходимости оптимизационных методов!

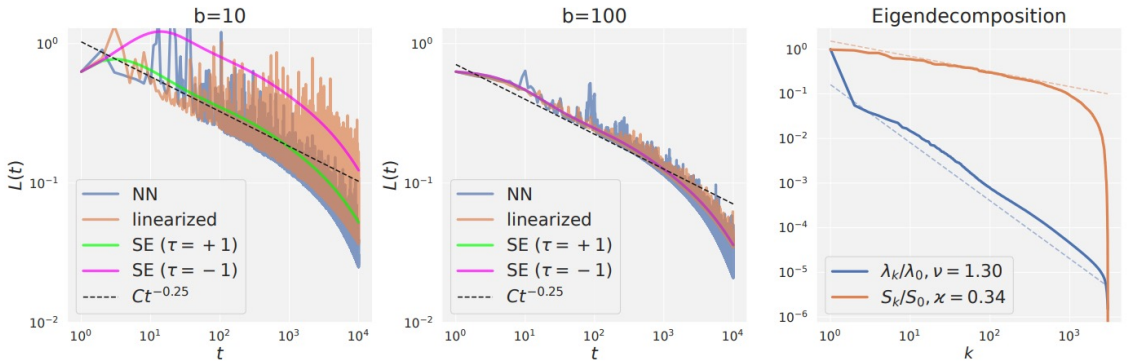


Figure 1: Optimization trajectories and spectral distributions for MNIST. **Left, center**: Loss trajectories of different training regimes for batch sizes 10 and 100 (see Sec. 3). We see good agreement of NN with SE approximation at $\tau = 1$. **Right**: Eigenvalues λ_k and partial sums $S_k = \sum_{l=k}^{N} \lambda_l C_{ll,0}$ characterizing $\mathbf{H}$ and $\mathbf{C}_0$. We see good agreement with assumed spectral power laws $\lambda_k \approx \Lambda k^{-\nu}$, $S_k \approx K k^{-\varkappa}$ as in Eq. (25). We estimate the values $\Lambda, K, \nu, \varkappa$ from the curves in the right figure. The dotted lines in the left and central figures then show $L_{\text{approx}}(t) = Ct^{-\zeta}$, with C computed as C_{signal} in Eq. (28a) and $\zeta = \frac{\varkappa}{\nu} \approx 0.25$. We see good agreement between Theorem 2 and experiment. Similar experiments for CIFAR10 with ResNet/MobileNet are given in Fig. 9 and also agree well with theory.

A view of mini-batch SGD via generating functions: conditions of convergence, phase transitions, benefit from negative momenta

M Velikanov, D Kuznedelev, D Yarotsky - arXiv preprint arXiv:2206.11124, 2022 - arxiv.org

Mini-batch SGD with momentum is a fundamental algorithm for learning large predictive models. In this paper we develop a new analytic framework to analyze mini-batch SGD for …

☆ Save Cite Cited by 2 Related articles »

AIRI

Связь с тензорными разложениями

Если рассмотреть нейросети со специальной нелинейностью, то можно показать, что

а) сверточные сети связаны с иерархическим разложением Такера

б) рекуррентные сети связаны с разложением тензорного поезда

$f(x) = \langle \Phi(x), W \rangle$ (аналог классического kernel trick,

$\Phi(x) = \Phi_1(x_1) \otimes \cdots \Phi_d(x_d)$)

W — представляется тензором малого ранга!

Для того, чтобы представить **случайную глубокую сеть** с помощью неглубокой, необходимо **экспоненциальное количество нейронов!**

Аппроксимация с помощью нейросетей

Классический результат:

(Cybenko): двух слоев достаточно, чтобы приблизить любую непрерывную функцию

Но как ошибка падает с числом параметров.

Есть ряд результатов, но не все так просто...;

Аппроксимация с помощью нейросетей

Рассмотрим DeepReLU сети:
разрешено умножение
на фиксированные матрицы,
и операция $ReLU(x) = max(x, 0)$

Можем ли мы приблизить функцию $y = x^2$?

Ответ: да! (Яроцкий, 2017)

Ошибка падает экспоненциально с числом параметров!

Theorem 2.1. *Suppose*

$$g = \begin{cases} 2x, & 0 \le x < 1/2, \\ 2(1-x), & 1/2 \le x \le 1, \\ 0, & otherwise, \end{cases}$$

$$g_s = \underbrace{g \circ g \circ .. \circ g}_{s},$$

then function

$$f_m = x - \sum_{s=1}^{m} \frac{g_s(x)}{2^{2s}}$$

approximates function $f(x) = x^2$ *with the following error bound:*

$$|f(x) - f_m(x)| \le 2^{-2m-2}.$$

g can be also rewritten as:

$$g(x) = 2\max(x, 0) - 2\max(x - 1/2, 0) + 2\max(x - 1, 0),$$

that exactly represents one layer on a fully-connected neural network with 3 neurons and ReLU activation function.

Следствие

Рассмотрим DeepReLU сети:
разрешено умножение
на фиксированные матрицы,
и операция $ReLU(x) = max(x, 0)$

Можем ли мы приблизить функцию
$y = x^2$?

Ответ: да! (Яроцкий, 2017)

Ошибка падает экспоненциально с числом параметров!

Мы можем показать, что
для широкого класса функций
(например, решений уравнений
в частных производных) ошибка
падает экспоненциально с числом
параметров.

Проблема: это никогда
не подтверждается на практике!

Стандартные оптимизационные
методы не находят этот минимум

Проблема DNN в аппроксимации гладких функций

Глубокие нейросети для аппроксимации даже простых гладких функций — **неустойчивая, плохообусловленная параметризация.**

Именно за счет композиции.

Для решения задач классификации точность $10^{-2} - 10^{-3}$ достаточна, для аппроксимации — нет

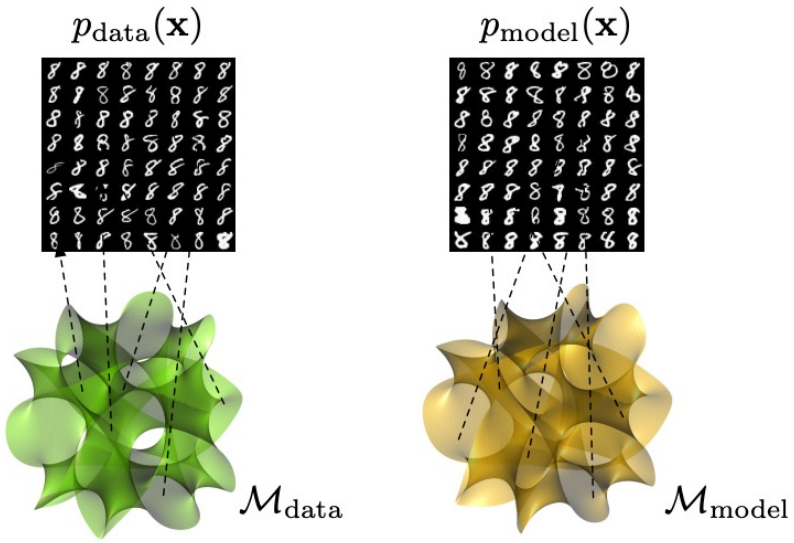
Необходим поиск альтернативных, более устойчивых, архитектур!

Геометрические и топологические подходы в DL

- Геометрические подходы очень важны для понимания структуры данных
- Важное направление — **topology data analysis:** вычисляем некоторые топологические характеристики и используем их для характеризация свойств фдатасета
- Например, анализ качества генеративных моделей
- **Направление** 2 — спектральные и графовые методы. Мы можем строить графы по данным и анализировать их свойства
- **Направление** 3 — гиперболические вложения для представления объектов

Geometry Score

Идея состоит в выделении топологических свойств



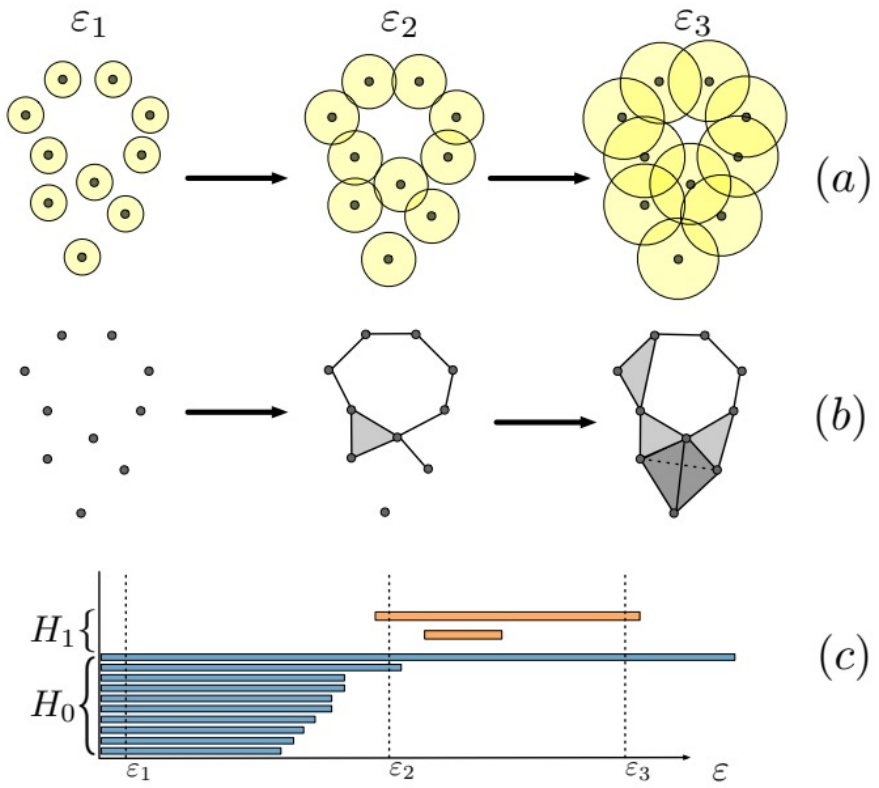
Geometry score: A method for comparing generative adversarial networks

V Khrulkov, I Oseledets - International conference on …, 2018 - proceedings.mlr.press

… we define a measure of their topological similarity (**Geometry Score**) in the following way: …

We also hypothesize that in the case of the large dimensionality of data **Geometry Score** of the …

☆ Save 99 Cite Cited by 103 Related articles All 11 versions »



AIRI

IMD Score (1)

- Пусть есть два многообразия можем определить расстояние Громова-Вассерштайна между ними
- Есть вычислимая верхняя оценка:

$$d_{\mathrm{GW}}(\mathcal{M},\mathcal{N}) \geq \sup_{t>0} e^{-2\left(t+t^{-1}\right)} |\mathrm{hkt}_{\mathcal{M}}(t) - \mathrm{hkt}_{\mathcal{N}}(t)|$$

  где **Heat Trace kernel** можно оценить через Лапласиан Графа:

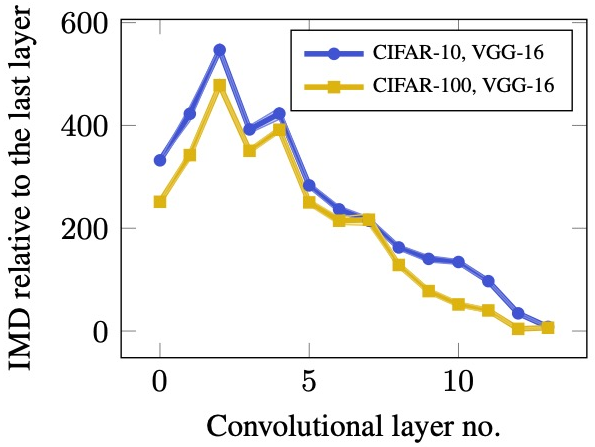
$$\mathrm{hkt}(t) = \mathrm{Tr}(\exp(t\mathcal{L}))$$

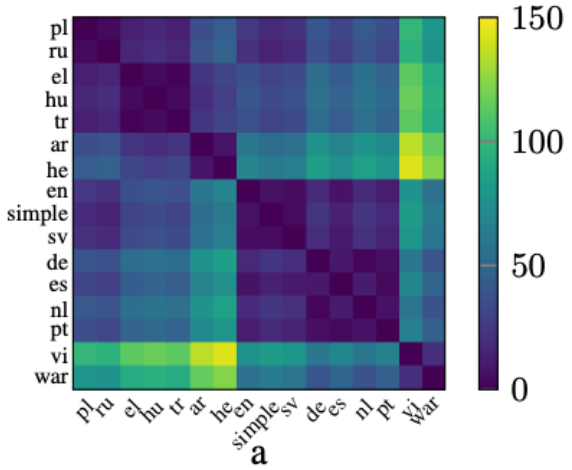
- Необходимо уметь все это эффективно вычислять.;

The **shape** of **data**: Intrinsic distance for **data** distributions
A Tsitsulin, M **Munkhoeva**, D Mottin, P Karras… - arXiv preprint arXiv …, 2019 - arxiv.org
… comparing **data** manifolds, … **data** moments. In a thorough experimental study, we demonstrate that our method effectively discerns the structure of **data** manifolds even on unaligned **data** …
☆ Save 99 Cite Cited by 30 Related articles All 9 versions »

IMD Score (2)

- IMD может отличать разные распределения
- Пример: активации с разных слоев
- Пример: распределение эмбеддингов слов





The **shape** of **data**: Intrinsic distance for **data** distributions

A Tsitsulin, M **Munkhoeva**, D Mottin, P Karras… - arXiv preprint arXiv …, 2019 - arxiv.org

… comparing **data** manifolds, … **data** moments. In a thorough experimental study, we demonstrate that our method effectively discerns the structure of **data** manifolds even on unaligned **data** …

☆ Save 99 Cite Cited by 30 Related articles All 9 versions »

AIRI

Связь самобучения с задачей восполнения пропусков в матрице

- В нашей недавне работе показана связь между **matrix completion** и **self-supervised learning**
- Идея SSL основана на использование связей между **положительными** и **отрицательными** примерами
- Результат: при некоторых предположениях, мы делаем лоу-ранг восполнение для Лапласиана графа
- Почему это важно: есть большая теория, которая исследует, сколько данных нужно для того, чтобы восстановить матрицу точно

$$\max_{\mathbf{Z}\in\mathbb{R}^{n\times d}} \quad \operatorname{Tr}\mathbf{Z}^\top\widehat{\mathbf{H}}\mathbf{Z}$$
$$\text{subject to} \quad \mathbf{Z}^\top\mathbf{Z}=\mathbf{I}$$

Is a Langrangian dual of

$$\min_{\mathbf{X}\in\mathcal{S}_+} \quad \|\mathbf{X}\|_*$$
$$\text{subject to} \quad \operatorname{vec}(\mathbf{W}\odot\mathbf{X})=\operatorname{vec}(\widehat{\mathbf{H}})$$

Bridging Spectral Embedding and Matrix Completion in Self-Supervised Learning
M Munkhoeva, I Oseledets
arXiv preprint arXiv:2305.19818, 2023 • arxiv.org

Связь самобучения с задачей восполнения пропусков в матрице

- Мы видим только часть пар
- Число сэмплов, необходимых для обучения пропорционально incoherence матрицы
- Информация должна быть равномерно распределена между столбами
- Можно понять, почему одни эмбеддинги лучше других!

M — матрица данных

$M = USV^\top$ — ее сингулярное разложение

Параметр некогерентность μ

$$\max_{1\le i\le n_1} \| \mathbf{U}^\top \mathbf{e}_i \|_2 \le \sqrt{\frac{\mu_0 r}{n_1}}, \max_{1\le j\le n_2} \| \mathbf{V}^\top \mathbf{e}_j \|_2 \le \sqrt{\frac{\mu_0 r}{n_2}}$$

Число сэмплов:
$p^* \ge c_0 \mu_0 r \log^2(n)/n$

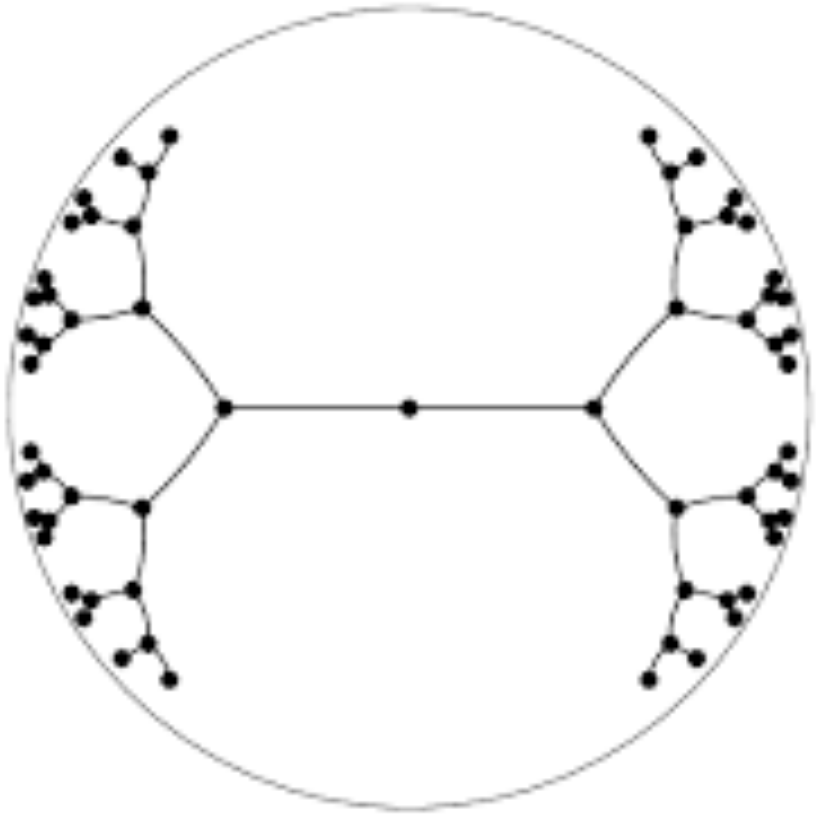
Bridging Spectral Embedding and Matrix Completion in Self-Supervised Learning
M Munkhoeva, I Oseledets
arXiv preprint arXiv:2305.19818, 2023 • arxiv.org

Гиперболическая геометрия

- Большинство методов Deep Learning отображают объекты в некоторое единое векторное пространство.
- Если объекты близки, то их векторные представления тоже близки.
- Как измерять близость?
- Три пространства постоянной кривизны: евлидово пространство, сфера, гиперболическое пространство.
- Использование гиперболической геометрии очень популярно!
- Почему?

Почему гиперболическое пространство?

- Гиперболическая геометрия хорошо представляет иерархию.
- Метрика: $g_{\mathcal{P}}(\mathbf{x}) = \lambda_x^2 g^E \lambda_x = \frac{2}{1-\|x\|^2}$
- Можно обобщить основные операции в этот формат.



Как реализовать гиперболические эмбеддинги

- Впервые — в NLP (Natural Language Processing)
- Наша работа — первая по использованию гиперболической геометрии для изображений
- Во второй работе получены наилучшие результаты для целого ряда задач
- Основная идея: нейросеть для выделения признаков + отображение в гиперболическое пространство

Hyperbolic image embeddings

V Khrulkov, L Mirvakhabova… - Proceedings of the …, 2020 - openaccess.thecvf.com

… that in **image** classification which construct **embeddings** in the Poincare model of **hyperbolic** … input **images** which are more familiar to the model get mapped closer to the boundary, and …

☆ Save 99 Cite Cited by 185 Related articles All 14 versions »

Hyperbolic vision transformers: Combining improvements in metric learning

A Ermolov, L Mirvakhabova… - Proceedings of the …, 2022 - openaccess.thecvf.com

… with output **embeddings** mapped to **hyperbolic** space. These **embeddings** are directly … Note that in our case, this token is used to obtain the **image embedding**, but we do not train …

☆ Save 99 Cite Cited by 31 Related articles All 6 versions »

Дельта-гиперболичность по Громову

- Можно измерить степень гиперболичности!
- Произведение Громола: $(y,z)_x = \frac{1}{2}(d(x,y) + d(x,z) - d(y,z))$
- $(x,z)_w \geq min((x,y)_w, (y,z)_w) - \delta$

  Если $\delta = 0$ получим дерево
- Мы всегда считаем δ
- Недавно обнаружили, что стандартный алгоритм вычисления имеет проблемы!

Encoder	Dataset			
	CIFAR10	CIFAR100	CUB	*Mini*ImageNet
Inception v3 [49]	0.25	0.23	0.23	0.21
ResNet34 [14]	0.26	0.25	0.25	0.21
VGG19 [42]	0.23	0.22	0.23	0.17

Гиперболическая геометрия для рекомендательных систем

- Заменяем SVD
- Оказалось, что гиперболический вариационный автокодировщик лучше всех
- Работает в индустрии, быстро учится

	@50		@100	
	REC	NDCG	REC	NDCG
$EASE^R$	0.3801	0.2978	0.5072	0.3510
SLIM	0.4002	**0.3203**	0.5299	0.3752
PureSVD	0.3593	0.2840	0.4784	0.3342
iALS	0.3138	0.2410	0.4216	0.2862
MultVAE	**0.4127**	0.3167	0.5456	0.3730
AE	0.1994	0.1333	0.3176	0.1744
HAE(H, 1)	0.3197	0.2623	0.4695	0.3282
HAE(H, 0.004)	0.3498	0.2801	0.5149	0.3516
HAE(M, 1)	0.3278	0.2694	0.4823	0.3375
HAE(M, 0.004)	0.3565	0.2872	0.5211	0.3586
H-VAE(0.004)	0.3793	0.3122	**0.5526**	**0.3891**

Performance of **hyperbolic** geometry models on top-N **recommendation** tasks
L Mirvakhabova, E **Frolov**, V Khrulkov… - … **Recommender Systems**, 2020 - dl.acm.org
… are known to be tightly connected to a **hyperbolic** geometry [14]. This fundamental …
recommender systems [5, 27]. In this work, we tackle the problem of **hyperbolic recommender** …
☆ Save 99 Cite Cited by 21 Related articles All 4 versions

Гиперболические вложения для моделирования последовательностей

- Текущая работа: применить к моделям типа «трансформер».
- Довольно успешно, меньше размерность векторного пространства при той же точности.
- Много открытых вопросов.

Большие языковые модели

- Большая часть исследований посвящена большим языковым моделям (LLM).
- LLM учится моделировать вероятность последовательности токенов (слов, букв): $p(x_1, \dots, x_d) = p(x_d | x_{d-1} \dots x_1) p(x_{d-1} | \dots)$
- Благодаря использованию внимания, LLM моделирует вероятностной распределение на последовательностях произвольной длины.
- Модель обучается на огромном количестве текстов.
- Наблюдается большое количество явлений, которое требует объяснения и новой аксиоматики.

Alignment

Пример теоретического результата (неформально)

Если языковая модель может генерировать с ненулевой вероятностью «плохие» тексты (заданные с помощью индикатора B(x))

то можно подобрать такое начало последовательности («промт»)

что в вероятность $p(B(y) = 1) \approx 1, y \sim p(\cdot\,|t)$

Также можно оценить длину «промта»

Fundamental limitations of **alignment** in large language models

Y Wolf, N Wies, Y Levine, A **Shashua** - arXiv preprint arXiv:2304.11082, 2023 - arxiv.org

… 3 order to assert several important statements regading LLM **alignment**: • **Alignment** impossibility: We show that an LLM **alignment** process which reduces undesired behaviors to a …

☆ Save 99 Cite Cited by 27 Related articles All 2 versions »

AIRI

Self-instruct

Очень интересный эффект — self-instruct.

Мы подаем в языковую модель «примеры» в начало текста.

И она начинает генерировать новые, которых не было в обучающей выборке.

Такое свойство «самоподобия» требует какого-то объяснения, «аксиоматики».

Применения языковых моделей

Интересное направление — использование LLM для генерации доказательств.

Наоборот — тестирование возможностей LLM на математических задачах/поиск проблем и бенчмарков.

Довольно много статей на эту тему, школьную математику +- современные модели научились решать (неясно, «научились» или запомнили).

Но это довольно интересное направление исследований.

Итоги

Перспективные направления: теория аппрокимации функций, геометические/топологическое подходы, теория представлений (не успел поговорить о disentanglement), теория вероятностей.

На мой взгляд, необходима некоторая аксиоматика для больших языковых моделей, чтобы объяснить их очень интересные свойства.

AIRI

airi.net

airi_research_institute

AIRI Institute

AIRI Institute

AIRI_inst

artificial-intelligence-research-institute

Telegram

Artificial Intelligence
Research Institute