

О сложности моделей и данных в параметрических моделях глубокого обучения

А.В. Грабовой

Диссертация на соискание ученой степени
доктора физико-математических наук

1.2.1 — Искусственный интеллект и машинное обучение

Научный консультант: профессор РАН К. В. Воронцов

2026 г.

Актуальность

- ▶ Экспоненциальный рост сложности моделей: от тысяч до сотен миллиардов параметров, с прогнозируемым переходом к триллионам.
- ▶ Особую остроту проблема приобретает при разработке больших языковых моделей (LLM), требующих огромных вычислительных, энергетических и финансовых ресурсов.
- ▶ Отсутствие строгой теоретической основы для предсказания поведения моделей при масштабировании делает процесс разработки экономически и энергетически неэффективным.
- ▶ Современные исследования преимущественно опираются на эмпирические корреляции, что приводит к необоснованным и противоречивым результатам.
- ▶ Классические подходы (VC, PAC, Радемахер) дают слишком консервативные оценки для перепараметризованных нейронных сетей и не учитывают специфику современных архитектур.

Ключевая проблема: Отсутствует единый теоретический аппарат для описания сложности моделей и данных с формальными критериями их соответствия.

Цель и задачи исследования

Цель: Построение единого теоретического аппарата для оценки сложности моделей глубокого обучения и сложности данных, а также установление формальных критериев соответствия между сложностью модели и сложностью выборки, необходимой для ее обучения.

Задачи:

- введение формальных определений мер сложности моделей и данных в рамках теории мер и установление критерия обучаемости модели на выборке;
- получение теоретических оценок ландшафтных мер на основе матриц Гессе для полносвязных, сверточных и трансформерных архитектур моделей глубокого обучения;
- установление связи ландшафтной меры с условной сложностью выборки;
- построение методов оценки достаточного объема выборок из простой генеральной совокупности;
- построение методов снижения сложности моделей глубокого обучения;
- демонстрация практического применения построенного теоретического аппарата в прикладных задачах.

Методы исследования

Для решения поставленных задач в диссертации используются:

- ▶ Методы теории мер, линейной алгебры, матричного анализа;
- ▶ Методы теории вероятностей и математической статистики;
- ▶ Методы теории оптимизации и анализа функций многих переменных;
- ▶ Методы статистической теории обучения.

Ключевой инструмент: анализ матриц Гессе функции потерь, содержащих информацию о локальной кривизне оптимизационного ландшафта.

Научная новизна

- ▶ Введены формальные определения меры сложности выборки $\mu_D(D)$ и меры сложности модели $\mu_f(f)$ в рамках теории мер, установлен критерий обучаемости $\mu_f(f) \leq \mu_D(D)$.
- ▶ Введена ландшафтная мера сложности модели $\mu_f(f|D) = \mathbb{E}\|\mathbf{H}_i - \mathbb{E}\mathbf{H}_i\|$, определяемая через спектральные свойства матриц Гессе.
- ▶ Впервые получены строгие теоретические оценки ландшафтной меры на основе спектральных норм матриц Гессе для полносвязных, сверточных и трансформерных архитектур.
- ▶ Доказана связь между достаточным объемом выборки и ландшафтной сложностью моделей.
- ▶ Предложены методы снижения сложности: прореживание на основе ковариации градиентов, мультидоменная дистилляция и анти-дистилляция.
- ▶ Доказаны теоремы о состоятельности LoRA-адаптеров и снижении сложности Радемахера в многозадачном обучении.

Теоретическая и практическая значимость

Теоретическая значимость:

- ▶ Разработан единый формальный язык для описания сложности моделей и данных, расширяющий классическую теорию статистического обучения на случай перепараметризованных нейронных сетей.
- ▶ Установлены прямые связи между архитектурными параметрами моделей (глубина, ширина, тип слоев) и сложностью оптимизационного ландшафта через спектр Гессе.

Практическая значимость:

- ▶ Методы оценки достаточного объема выборки позволяют экономить вычислительные ресурсы при планировании экспериментов.
- ▶ Методы снижения сложности обеспечивают сжатие моделей без потери качества для развертывания в ресурсно-ограниченных средах.
- ▶ Разработанный аппарат применен к реальным задачам: многозадачное обучение, декодирование фМРТ-изображений, детекция машинно-генерированного контента.
- ▶ Экспериментально подтверждена эффективность всех предложенных методов на общедоступных наборах данных.

Диссертация в контексте теории сложности

1. Эволюция теории сложности
2. Ограничения существующих подходов
3. Место диссертационной работы

Эволюция теории сложности обучения

Статистическая теория сложности (1970–2010)

- ▶ Вапник, Червоненкис, 1974 — введение VC-размерности как меры емкости класса; равномерная сходимость частот к вероятностям.
- ▶ Valiant, 1984 — формализация понятия обучаемости; гарантии с заданной точностью и надежностью.
- ▶ Воронцов, 2010 — развитие VC-теории; учет структуры данных и локализации алгоритмов для снижения консервативности оценок.

Аппроксимационная теория (1989–2015)

- ▶ Cybenko, 1989 — однослойная сеть с сигмоидой может аппроксимировать любую непрерывную функцию.
- ▶ Hastad, 1987; Bengio, 2007; Cohen, 2015 — экспоненциальный рост выразительной способности с увеличением числа слоев.

Современные эмпирические подходы (2017–2022)

- ▶ Sagun, 2017; Keskar, 2016 — эмпирическое исследование спектра Гессе; связь “плоских” минимумов с обобщением.
- ▶ Kaplan, 2020; Hoffmann, 2022 — степенные зависимости качества от числа параметров и объема данных; оптимальные соотношения для вычислительных бюджетов.

Направления развивались независимо. Отсутствует единая теория, связывающая сложность модели и сложность данных.

Ограничения существующих подходов

Статистическая теория сложности

- ▶ Ориентация на анализ худшего случая \Rightarrow оценки существенно завышены для реальных данных.
- ▶ Не объясняет феномен перепараметризации: модели с высокой VC-размерностью успешно обобщают данный.
- ▶ Сложность модели и данных рассматриваются изолированно, без формального критерия соответствия.

Аппроксимационная теория

- ▶ Характеризует принципиальную возможность представления функций, но не дает количественных оценок сложности обучения.
- ▶ Не учитывает влияние конечности выборки и процедуры оптимизации.

Эмпирический анализ ландшафта

- ▶ Результаты носят описательный характер, отсутствуют строгие аналитические оценки для конкретных архитектур.

Законы масштабирования

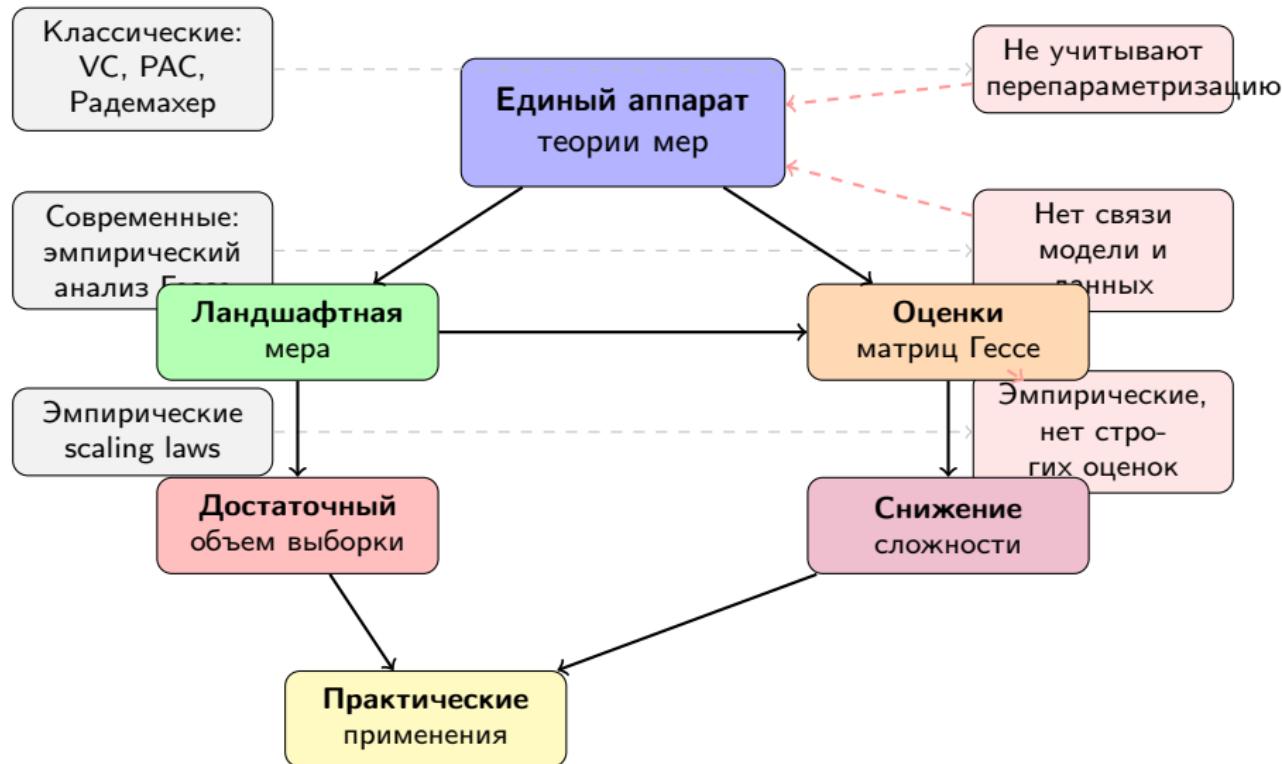
- ▶ Эмпирические зависимости не имеют строгого теоретического обоснования.
- ▶ Не объясняют механизмы, определяющие оптимальные соотношения между параметрами и данными.

Диссертация заполняет следующие пробелы

- ▶ **Отсутствие единого формального языка.** Введение мер сложности $\mu_D(D)$ и $\mu_f(f)$; критерий обучаемости $\mu_f(f) \leq \mu_D(D)$.
- ▶ **Нет аналитических оценок сложности.** Ландшафтная мера $\mu_f(f|D) = \mathbb{E}\|\mathbf{H}_i - \mathbb{E}\mathbf{H}_i\|$; строгие оценки $\|\mathbf{H}\|_2$ для полно связных, сверточных и трансформерных сетей.
- ▶ **Отсутствие теоретически обоснованных методов снижения сложности.** Прореживание на основе ковариации градиентов; мультидоменная дистилляция; антидистилляция.
- ▶ **Эмпирический характер законов масштабирования.** Теоремы о состоятельности LoRA; снижение сложности Радемахера в многозадачном обучении в $1/\sqrt{T}$ раз.

Работа создает единый теоретический аппарат, связывающий классическую теорию сложности с современными архитектурами и эмпирическими наблюдениями.

Диссертация в контексте теории сложности



Общее введение теории ландшафтной меры

1. Меры сложности выборки и модели
2. Критерий обучаемости
3. Условная сложность и достаточный объем данных
4. Ландшафтная мера: определение и интерпретация

Мера сложности выборки

Определение

Генеральной совокупностью данных Γ назовем произвольное множество объектов, которые исследуются в рамках той либо иной задаче.

Определение

Генеральную совокупность Γ назовем однородной, если все объекты генеральной совокупности порождаются из одного распределения. В противном случае выборку назовем k -родной.

Рассмотрим кольцо выборок: $\mathfrak{D} = \{D_\Gamma^i\}$, $D_\Gamma^i \subset \Gamma$.

Определение

Мерой сложности выборки назовем отображение μ_D , такое, что:

$$\mu_D(D_i) : \mathfrak{D}_\Gamma \rightarrow \mathbb{R}_+,$$

удовлетворяющее свойству:

$$\mu_D(D_i \cup D_j) \leq \mu_D(D_i) + \mu_D(D_j),$$

где равенство достигается при условии $D_i \cap D_j = \emptyset$.

Мера сложности модели

Определение

Для параметрического семейства функций $\mathfrak{F} = \{f_i\}$ вводится отображение

$$\mu_f(f_i) : \mathfrak{F} \rightarrow \mathbb{R}_+,$$

называемое **мерой сложности модели**.

Примеры мер:

- ▶ Число параметров модели.
- ▶ Норма вектора весов $\|w\|$.
- ▶ Эффективная емкость, оцениваемая через след гессиана.
- ▶ В дальнейшем будет введена ландшафтная мера $\mu_f(f|D)$, учитывающая данные.

Примечание: Конкретный вид меры зависит от задачи и модели; важно лишь, что она задает числовую характеристику сложности.

Критерий обучаемости и дообучение

Определение

Модель f **обучаема** на выборке D , если

$$\mu_f(f) \leq \mu_D(D).$$

Интуиция: сложность модели не должна превышать сложность данных, иначе неизбежно переобучение.

Теорема (Грабовой)

Если модель уже обучаема на исходной выборке D , то для добавления новой выборки D' достаточно выполнения условия

$$\mu_f(f) - \mu_D(D) \leq \mu_D(D').$$

Это означает, что остаточная емкость модели должна покрываться сложностью новых данных.

Примечание: Критерии дают формальную основу для планирования экспериментов и оценки необходимости сбора дополнительных данных.

Условная сложность выборки

Мера $\mu_D(D)$ оценивает абсолютную сложность данных, но для конкретной модели важна *относительная сложность*.

Определение

Условная сложность выборки D относительно модели f :

$$\mu_D(D|f) = \inf\{\mu_D(D') : D' \subseteq D, \mu_f(f) \leq \mu_D(D')\}.$$

Это минимальная сложность подвыборки, достаточная для выполнения критерия обучаемости.

Свойства

- ▶ $\mu_D(D|f) \leq \mu_D(D)$
- ▶ Если $\mu_f(f) \leq \mu_D(D)$, то $\mu_D(D|f) \leq \mu_D(D)$
- ▶ Зависит как от данных, так и от архитектуры модели

Простая генеральная совокупность

Определение

Однородную генеральную совокупность Γ_C назовем **простой**, если все ее объекты имеют одинаковую сложность C :

$$\forall \gamma \in \Gamma_C \mu_D(\{\gamma\}) = C.$$

Теорема (Грабовой)

Для простой совокупности мера сложности любой конечной выборки пропорциональна ее размеру:

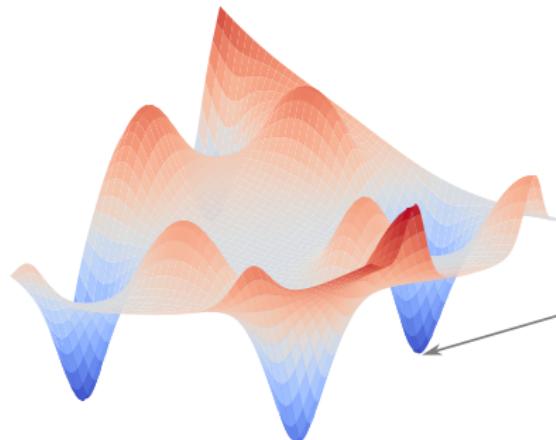
$$\mu_D(D) = C \cdot |D|.$$

Примечание: Условная сложность превращается в минимальный размер подвыборки:

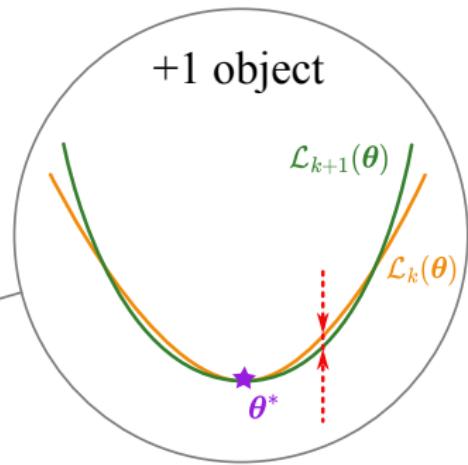
$$\mu_D(D|f) = C \cdot m^*, \quad m^* = \min\{k : \mu_f(f) \leq Ck\}.$$

m^* — минимальное количество объектов для обучения модели f .

От общей теории к конкретной мере сложности



(a) Loss function landscape



(b) Losses difference

Если добавление нового объекта данных существенно изменяет ландшафт оптимизации, то модель недостаточно обучена на текущей выборке.

Изменения ландшафта при вариации выборки

Пусть задана выборка из Γ_C :

$$D = \{(x_i, y_i)\}, \quad i = 1, \dots, m, \quad x \in \mathcal{X}, \quad y \in \mathcal{Y}, \quad D \subset \Gamma_C.$$

Рассмотрим эмпирический риск на выборках размера k и $k + 1$:

$$\mathcal{L}_k(\theta) = \frac{1}{k} \sum_{i=1}^k \ell(f_\theta(x_i), y_i), \quad \mathcal{L}_{k+1}(\theta) = \frac{1}{k+1} \sum_{i=1}^{k+1} \ell(f_\theta(x_i), y_i).$$

Их разность:

$$\mathcal{L}_{k+1}(\theta) - \mathcal{L}_k(\theta) = \frac{1}{k+1} \left(\ell(f_\theta(x_{k+1}), y_{k+1}) - \mathcal{L}_k(\theta) \right).$$

Примечание: при $k \rightarrow \infty$ эта разность стремится к нулю. Скорость убывания определяется тем, насколько новый объект “не похож” на уже имеющиеся. Далее эта скорость оценивается через спектральные свойства матрицы Гессе.

Предположение о стационарности точки минимума

Предположение

Пусть θ^* — локальный минимум обеих функций потерь $\mathcal{L}_k(\theta)$ и $\mathcal{L}_{k+1}(\theta)$, т.е.
 $\nabla \mathcal{L}_k(\theta^*) = \nabla \mathcal{L}_{k+1}(\theta^*) = 0$.

В окрестности минимума функции аппроксимируются квадратично:

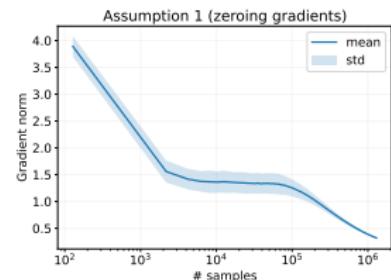
$$\mathcal{L}_k(\theta) \approx \mathcal{L}_k(\theta^*) + \frac{1}{2}(\theta - \theta^*)^\top H^{(k)}(\theta^*)(\theta - \theta^*).$$

Тогда разность потерь принимает вид:

$$\begin{aligned}\mathcal{L}_{k+1}(\theta) - \mathcal{L}_k(\theta) &\approx \frac{1}{k+1} \left(\ell(f_{\theta^*}(x_{k+1}), y_{k+1}) - \mathcal{L}_k(\theta^*) \right) \\ &+ \frac{1}{k+1} (\theta - \theta^*)^\top \left(H_{k+1}(\theta^*) - \frac{1}{k} \sum_{i=1}^k H_i(\theta^*) \right) (\theta - \theta^*).\end{aligned}$$

Замечание

Первое слагаемое ограничено константой, не зависящей от k . Второе содержит разность матриц Гессе — именно она определяет скорость сходимости ландшафта и используется в основе ландшафтной меры.



Ландшафтная мера сложности модели

Определение

Ландшафтной мерой сложности параметрической функции f называется

$$\mu_f(f|D) = \mathbb{E}_{x_i \in D} \|H_i(\theta^*) - \mathbb{E}_{x_i \in D} H_i(\theta^*)\|_2,$$

где $H_i(\theta^*)$ — матрица Гессе для i -го объекта в точке минимума θ^* .

Ландшафтная мера характеризует **вариативность** гессианов по объектам выборки:

- ▶ Малая вариативность — объекты одинаково влияют на ландшафт, модель насытилась данными.
- ▶ Большая вариативность — разные объекты по-разному искривляют ландшафт, данных недостаточно.

Связь с изменением функции потерь

$$|\mathcal{L}_{k+1}(\theta) - \mathcal{L}_k(\theta)| \leq \frac{M_l}{k+1} + \frac{\mu_f(f|D)}{k+1} \|\theta - \theta^*\|_2^2.$$

Ландшафтные меры для разных архитектур

1. Полносвязные нейронные сети
2. Сверточные нейронные сети
3. Экспериментальное подтверждение

Ландшафтная мера полносвязной сети

Теорема (Грабовой-Киселев)

Для L -слойной полносвязной сети с $ReLU$ верна оценка скорости сходимости ландшафта:

$$|\mathcal{L}_{k+1}(\theta) - \mathcal{L}_k(\theta)| \leq \frac{C_1}{k+1} + \frac{C_2 R^2}{k+1} \cdot L(hM)^{2L},$$

где C_1, C_2 — константы, R — радиус окрестности оптимума, h — ширина слоя, M — граница параметров.

Следствие

Ландшафтная мера сложности полносвязной модели: $\mu_f(f|D) \propto L(hM)^{2L}$.

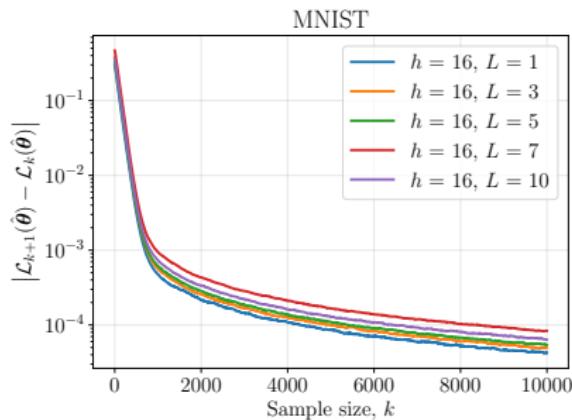
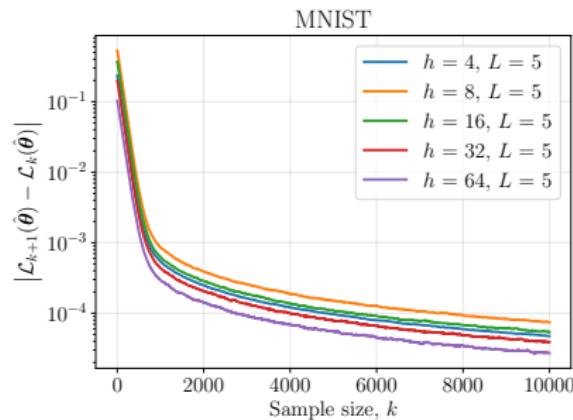
Интерпретация: сложность ландшафта растет **экспоненциально** с глубиной L и **полиномиально** с шириной h . Это объясняет, почему глубокие сети требуют больше данных для стабилизации.

Замечание

Для однослоиной сети ($L = 1$) имеем $\mu_f \propto h^2$ — квадратичная зависимость от ширины.

Эксперимент по полносвязной сети

Цель: проверить, что разность потерь $|\mathcal{L}_{k+1} - \mathcal{L}_k|$ убывает с ростом выборки k , а скорость убывания зависит от архитектуры (как предсказывает ландшафтная мера $\mu_f(f|D)$).



- Монотонное убывание подтверждено для MNIST, FashionMNIST, CIFAR10.
- Увеличение глубины L дает более медленную сходимость (больше μ_f).
- Увеличение ширины h слабее влияет на сходимость — согласуется с асимптотикой $\mu_f \propto L(hM)^{2L}$.

Ландшафтная мера сверточной сети

Теорема (Грабовой-Мешков-Киселев)

$\mu_f(f|D) \propto L(C^2 w^2 kd)^L$, где C — число каналов, k — размер ядра, d — длина последовательности.

Теорема (Грабовой-Мешков-Киселев)

$\mu_f(f|D) \propto C^2 k^2 L(C^2 k^2 w^2 mn)^L$, где $m \times n$ — пространственные размеры входа.

Теорема (Грабовой-Мешков-Киселев)

Операция пулинга (MaxPool/AvgPool) снижает сложность мультипликативно на $\left(\frac{1}{k_{pool}^2}\right)^{L-l+2}$, где l — позиция слоя пулинга.

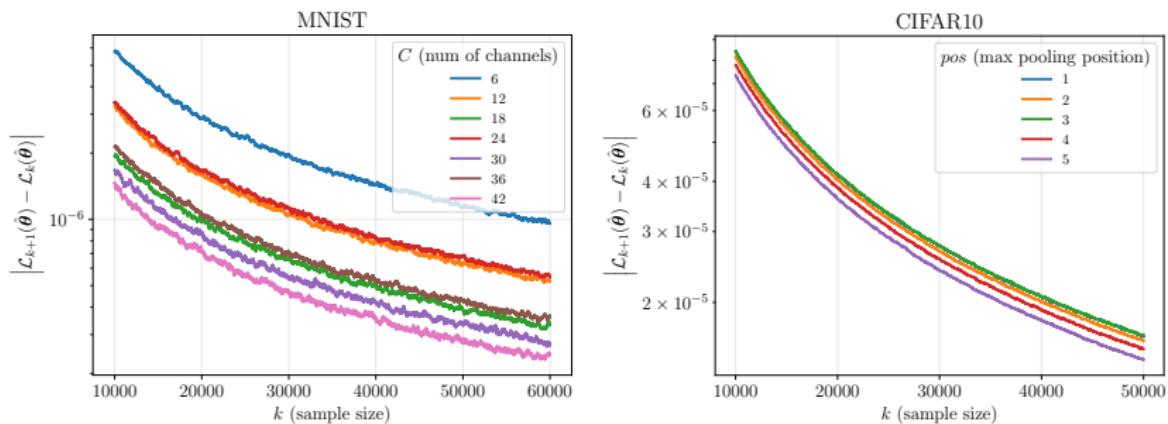
Интерпретация: сложность ландшафта растет **экспоненциально** с глубиной L и **полиномиально** с числом каналов, размером ядра и разрешением. Пулинг действует как эффективный регуляризатор, уменьшая ландшафтную меру.

Замечание

Полученные оценки объясняют, почему увеличение числа каналов или глубины требует больше данных, а пулинг облегчает обучение.

Экспериментальная проверка для сверточных сетей

Цель: исследовать влияние параметров сверточной архитектуры (количество каналов C , позиция пулинга) на разность потерь $|\mathcal{L}_{k+1} - \mathcal{L}_k|$ и сопоставить с теоретической асимптотикой $\mu_f(f|D)$.



- ▶ Увеличение числа каналов C монотонно увеличивает разность потерь, что согласуется с полиномиальным ростом сложности.
- ▶ Раннее применение пулинга снижает разность потерь — подтверждает теоретическую оценку о регуляризующем эффекте пулинга.
- ▶ Эксперименты на MNIST и CIFAR10 подтверждают теоретические асимптотики для сверточных архитектур.

Оценка матриц Гессе для нейросетевых моделей

1. Декомпозиция на G и H компоненты
2. Полносвязные сети: оценки нормы
3. Матричная факторизация
4. Сверточные сети и влияние пулинга
5. Сводная таблица асимптотических оценок

Декомпозиция матрицы Гессе на G и H компоненты

Матрица Гессе для функции $\mathcal{L}(\theta)$ от n параметров:

$$H(\mathcal{L})_{ij} = \frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j} = \sum_{i=1}^l \frac{\partial^2 \ell(z_i, y_i)}{\partial \theta_i \partial \theta_j},$$

Используя цепное правило, матрица Гессе представима в виде двух слагаемых:

$$H_i(\theta) = \underbrace{\nabla_{\theta} z_i \frac{\partial^2 \ell(z_i, y_i)}{\partial z_i^2} \nabla_{\theta} z_i^T}_{G\text{-компоненты}} + \underbrace{\sum_{k=1}^K \frac{\partial \ell(z_i, y_i)}{\partial z_{ik}} \nabla_{\theta}^2 z_{ik}}_{H\text{-компоненты}}.$$

Следствие

Эмпирически установлено, что H -компоненты близка к нулю, а особый интерес представляет G -компоненты:

$$H_i(\theta) \approx \nabla_{\theta} z_i \frac{\partial^2 \ell(z_i, y_i)}{\partial z_i^2} \nabla_{\theta} z_i^T.$$

Оценка матрицы Гессе для полносвязной сети

Теорема (Киселев-Грабовой)

Для L -слойной полносвязной нейронную сеть с функцией активации $ReLU$, применяемую для решения задачи классификации на K классов. Пусть: $\|W^{(p)}\|_2 \leq M_W$, $\|x_i\|_2 \leq M_x$, для всех слоев $p = 1, \dots, L$ в сети и для всех объектов $i = 1, \dots, m$. Тогда для любого объекта $i = 1, \dots, m$ выполняется следующее неравенство: $\|\mathcal{H}_i(\theta)\|_2 \leq L\sqrt{2}M_x^2M_W^{2L} + \sqrt{2}\frac{M_W^2(M_W^{2L}-1)}{M_W^2-1}$.

Теорема (Киселев-Грабовой)

Пусть все параметры модели ограничены некоторой константой $M > 0$, то есть для всех $i, j = 1, \dots, h$ и для всех слоев $p = 1, \dots, L$ выполняется условие $|w_{ij}^{(p)}| \leq M$, тогда при выполнении условий предыдущей теоремы справедлива оценка: $\|\mathcal{H}_i(\theta)\|_2 \leq L\sqrt{2}M_x^2(hM)^{2L} + \sqrt{2}\frac{(hM)^2((hM)^{2L}-1)}{(hM)^2-1}$.

Следствие

Частным случаем является оценка на матрицу Гессе для однослоиной сети $L = 1$ вида: $\|\mathcal{H}_i(\theta)\|_2 \leq \sqrt{2}M_W^2(M_x^2 + 1)$.

Факторизация матричных моделей глубокого обучения

Пусть $f_\theta(x)$ является суперпозицией линейных операторов:

$$f_\theta(x) = T^{(L+1)} \Lambda^{(L)} \dots \Lambda^{(1)} T^{(1)} x,$$

где $\theta = \text{col}(W^{(L+1)}, \dots, W^{(1)})$.

Теорема (Киселев-Грабовой-Мешков)

Пусть функция нейронной сети $f_\theta(x)$ представима в виде суперпозиции линейных операторов, тогда матрица Гессе функции потерь относительно параметров модели представляется в факторизованной форме: $H(\theta) \approx Q^T F^T A F Q$, описывающие G -компоненту матрицы Гессе.

Теорема (Киселев-Грабовой-Мешков)

Пусть функция нейронной сети $f_\theta(x)$ представима в виде суперпозиции линейных операторов. Пусть для всех p выполняется: $\|Q^{(p)}\| \leq q$, $\|T^{(p)}\|^2 \leq w_T^2$. Тогда справедлива оценка: $\|H_O\| \leq \sqrt{2}q^2 \|x\|^2 (L+1)w_T^{2L}$.

Сверточные модели глубокого обучения

Теорема (Грабовой-Мешков-Киселев)

Пусть задана функция нейронной сети $f_\theta x = C_{W^{(L+1)}} \circ \sigma \circ \dots \circ \sigma \circ C_{W^{(1)}}$, где $C_{W^{(i)}}$ — одномерная свертка с ядром $W^{(i)}$. Пусть заданы следующие верхние оценки на параметры: $C_l \leq C$, $k_i \leq k$, $d_i \leq d_1 := d$, $|W_{i,j,k}^{(p)}|^2 \leq w^2$. Тогда норма матрицы Гессе имеет следующую верхнюю оценку: $\|\mathbf{H}\| \leq \sqrt{2} \|x\|^2 d^2(L+1)(C^2 w^2 kd)^L$.

Теорема (Мешков-Киселев-Грабовой)

Пусть задана функция нейронной сети $f_\theta x = C_{W^{(L+1)}} \circ \dots \circ C_{W^{(1)}}$, где $C_{W^{(i)}}$ — двумерная свертка с ядром $W^{(i)}$. Пусть заданы следующие верхние оценки на параметры: $C_l \leq C$, $k_i \leq k$, $m_i \leq m_1 := m$, $n_i \leq n_1 := n$, $|W_{i,j,k}^{(p)}|^2 \leq w^2$. Тогда норма матрицы Гессе имеет следующую верхнюю оценку: $\|\mathbf{H}_0\| \leq \sqrt{2} \|x\|^2 C^4 k^4 m^2 n^2 (L+1)(C^2 k^2 w^2 mn)^L$.

Замечание

Полученные оценки имеют недостаток связанный с тем, что они не учитывают уменьшение размеров после сверточных операций и зависят только от верхних границ параметров.

Снижение сложности в сверточных моделях

Теорема (Мешков-Киселев-Грабовой)

Для сверточных сетей с операциями пулинга (*MaxPool2D* или *AvgPool2D*) норма матрицы Гессе уменьшается за счет множителя:

$$\left(\frac{1}{k_{\text{pool}}^2} \right)^{L-l+2}$$

где k_{pool} — размер ядра пулинга, L — глубина сети, l — позиция слоя пулинга.

Практический вывод

Операции пулинга снижают сложность модели и действуют как **механизм регуляризации** в глубоких сверточных нейронных сетях.

Замечание

Это объясняет, почему пулинг-слои улучшают обобщающую способность моделей: они уменьшают сложность оптимизационного ландшафта.

Сводная таблица асимптотических оценок

Архитектура	Асимптотика нормы $\ \mathbf{H}\ _2$
Полносвязная сеть	$L(hM)^{2L}$
1D свертка	$L(C^2 w^2 kd)^L$
2D свертка	$L(C^2 k^2 w^2 mn)^L$
С пулингом	$L \cdot \left(\frac{1}{k_{\text{pool}}^2}\right)^{L-l+2} \cdot (\text{базовая оценка})$

Ключевые наблюдения

- ▶ Экспоненциальный рост с глубиной сети L
- ▶ Полиномиальный рост с размером слоев
- ▶ Пулинг снижает сложность
- ▶ Оценки позволяют предсказывать сложность без обучения модели

Достаточный размер выборки

1. Статистические и байесовские методы
2. Сэмплирование эмпирической функции ошибки (D- и M-достаточность)
3. Близость апостериорных распределений (KL- и S-достаточность)
4. Экспериментальное сравнение

Сэмплирования эмпирической функции ошибки

Определение

Пусть задано некоторое $\varepsilon > 0$. Размер выборки m^* назовем D -достаточным, если для всех $k \geq m^*$ выполняется условие: $D(k) = \mathbb{D}_{\hat{w}_k} L(\mathcal{D}_m, \hat{w}_k) \leq \varepsilon$.

Определение

Пусть задано некоторое $\varepsilon > 0$. Размер выборки m^* назовем M -достаточным, если для всех $k \geq m^*$ выполняется условие: $M(k) = |\mathbb{E}_{\hat{w}_{k+1}} L(\mathcal{D}_m, \hat{w}_{k+1}) - \mathbb{E}_{\hat{w}_k} L(\mathcal{D}_m, \hat{w}_k)| \leq \varepsilon$.

Теорема (Киселев-Грабовой)

Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ и $\|\Sigma_{k+1} - \Sigma_k\|_F \rightarrow 0$ при $k \rightarrow \infty$. Тогда в модели линейной регрессии определение M -достаточного размера выборки корректно. А именно, для любого $\varepsilon > 0$ существует такой m^* , что для всех $k \geq m^*$ выполняется $M(k) \leq \varepsilon$.

Следствие

Пусть $\|\mathbf{m}_k - \mathbf{w}\|_2 \rightarrow 0$ и $\|\Sigma_k - [k\mathcal{I}(\mathbf{w})]^{-1}\|_F \rightarrow 0$ при $k \rightarrow \infty$. Тогда в модели линейной регрессии определение M -достаточного размера выборки корректно.

KL-близость апостериорных распределений

Определение

Подвыборки \mathcal{D}^1 и \mathcal{D}^2 назовем близкими, если \mathcal{I}_2 может быть получено из \mathcal{I}_1 путем удаления, замены или добавления одного элемента, то есть

$$|\mathcal{I}_1 \Delta \mathcal{I}_2| = |(\mathcal{I}_1 \setminus \mathcal{I}_2) \cup (\mathcal{I}_2 \setminus \mathcal{I}_1)| = 1.$$

Определение

Пусть задано некоторое $\varepsilon > 0$. Размер выборки m^* называется KL-достаточным, если для всех $k \geq m^*$

$$KL(k) = D_{KL}(p_k \| p_{k+1}) = \int p_k(w) \log \frac{p_k(w)}{p_{k+1}(w)} dw \leq \varepsilon.$$

Теорема (Киселев-Грабовой)

Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ и $\|\Sigma_{k+1} - \Sigma_k\|_F \rightarrow 0$ при $k \rightarrow \infty$. Тогда в модели с нормальным апостериорным распределением параметров определение KL-достаточного размера выборки корректно. А именно, для любого $\varepsilon > 0$ существует такой m^* , что для всех $k \geq m^*$ выполняется $KL(k) \leq \varepsilon$.

S -близость апостериорных распределений

Определение

Пусть задано некоторое $\varepsilon > 0$. Размер выборки m^* называется S -достаточным, если для всех $k \geq m^*$

$$S(k) = s\text{-score}(p_k, p_{k+1}) \geq 1 - \varepsilon.$$

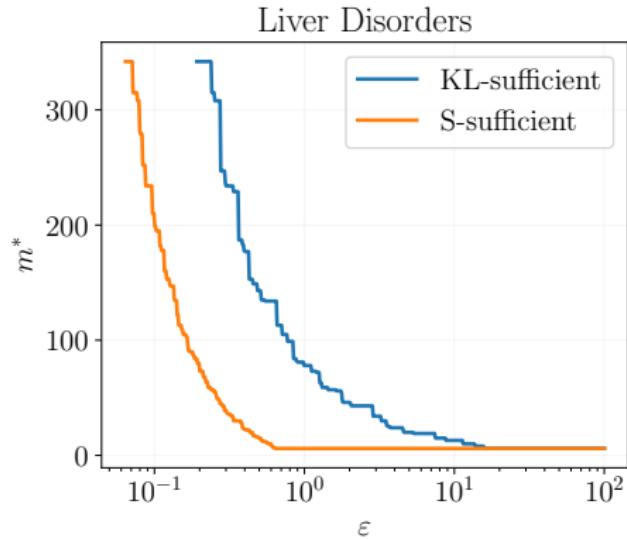
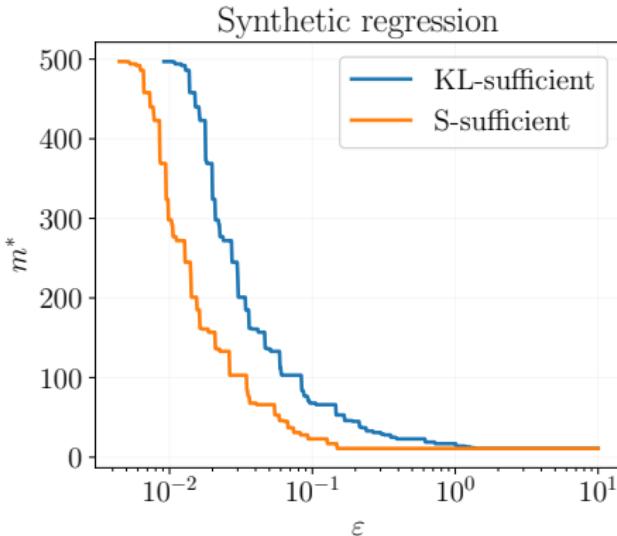
Теорема (Киселев-Грабовой)

Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ при $k \rightarrow \infty$. Тогда в модели с нормальным апостериорным распределением параметров определение S -достаточного размера выборки корректно. А именно, для любого $\varepsilon > 0$ существует такой m^* , что для всех $k \geq m^*$ выполняется $S(k) \geq 1 - \varepsilon$.

Теорема (Грабовой-Киселев)

Пусть множества значений признаков и целевой переменной ограничены, то есть существует константа $M \in \mathbb{R}$ такая, что $\|\mathbf{x}\|_2 \leq M$ и $|y| \leq M$ для всех объектов выборки. Если $\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k) = \omega(\sqrt{k})$ при $k \rightarrow \infty$, то в модели линейной регрессии с нормальным априорным распределением параметров $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ и $\|\Sigma_{k+1} - \Sigma_k\|_F \rightarrow 0$ при $k \rightarrow \infty$.

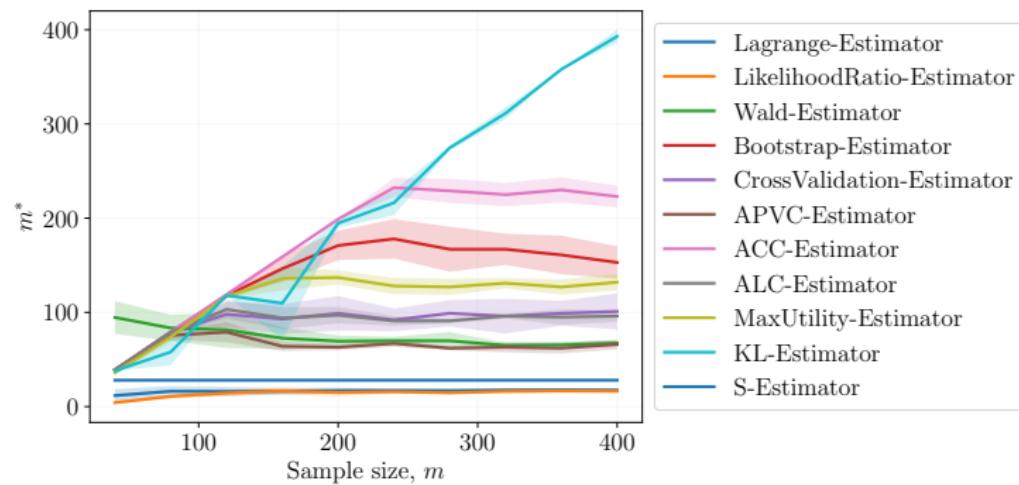
KL-достаточность против S-достаточности



Зависимость достаточного размера выборки от порогового параметра. Для S-достаточного размера выборки требуются более низкие значения порога. Таким образом, он оказывается более требовательным к этому значению.

Экспериментальная проверка KL- и S-достаточности

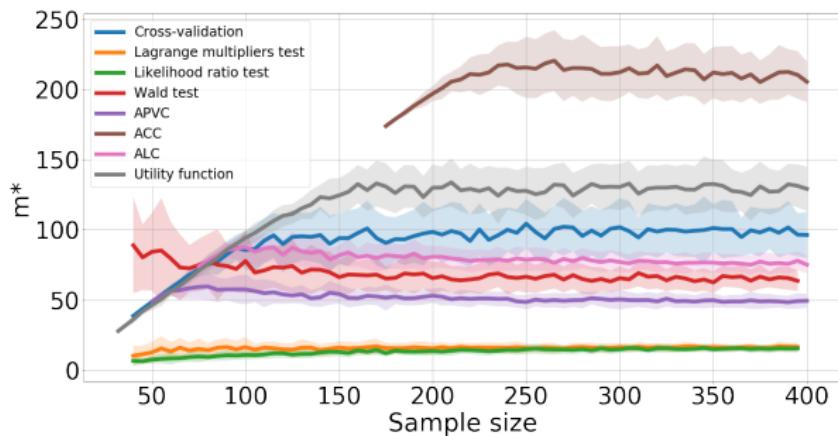
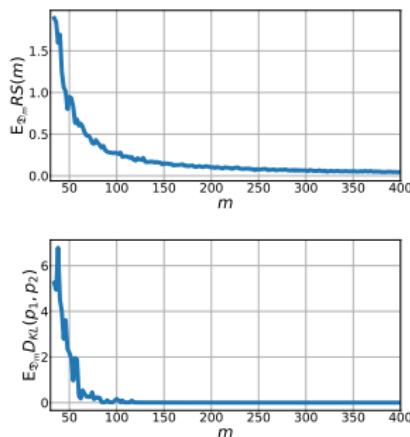
Цель: исследовать сходимость функций $KL(k)$ и $S(k)$ к предельным значениям и сопоставить с классическими методами оценки достаточного объема выборки.



- ▶ Функции $KL(k)$ и $S(k)$ сходятся к предельным значениям.
- ▶ KL-достаточность дает наиболее консервативные оценки.
- ▶ S-достаточность указывает на меньшие достаточные объемы, оставаясь строго обоснованной.
- ▶ Наблюдается согласие с классическими методами при больших m .

Экспериментальное сравнение методов оценки

Цель: сопоставить поведение различных методов определения достаточного объема выборки на реальных данных.

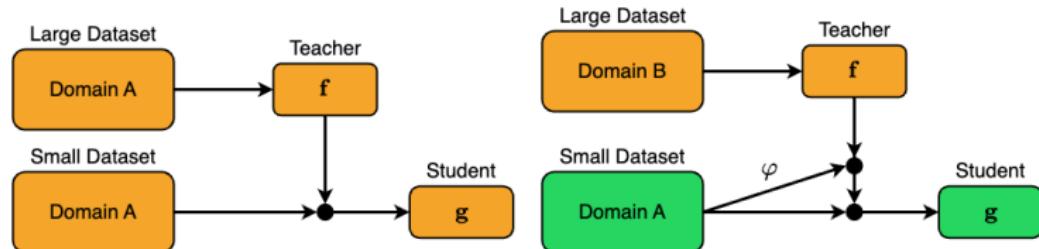


- ▶ Все исследованные методы демонстрируют **монотонное** поведение.
- ▶ Значения функций с ростом k асимптотически стремятся к константе, что подтверждает корректность определений достаточного объема.
- ▶ Дисперсия оценок m^* мала, методы устойчивы к выбросам.
- ▶ Полученные результаты согласуются с теоремами о сходимости предложенных критериев.

Снижение сложности моделей

1. Удаление параметров (pruning) на основе анализа градиентов
2. Мультидоменная дистилляция знаний
3. Анти-дистилляция: наращивание сложности с регуляризацией
4. Экспериментальная оценка методов

Дистилляция на многих генеральных совокупностях



Определение

Генеральная совокупность B называется близкой к генеральной совокупности A , если существует инъективное отображение $\varphi : A \rightarrow B$.

Заданы выборки из двух близких генеральных совокупностей отображением φ :

$$\mathcal{D}_A = \{(x_i, y_i)\}_{i=1}^n, x_i \in \mathbb{X}_A, y_i \in \mathbb{Y}, \quad \mathcal{D}_B = \{(x_i, y_i)\}_{i=1}^m, x_i \in \mathbb{X}_B, y_i \in \mathbb{Y}.$$

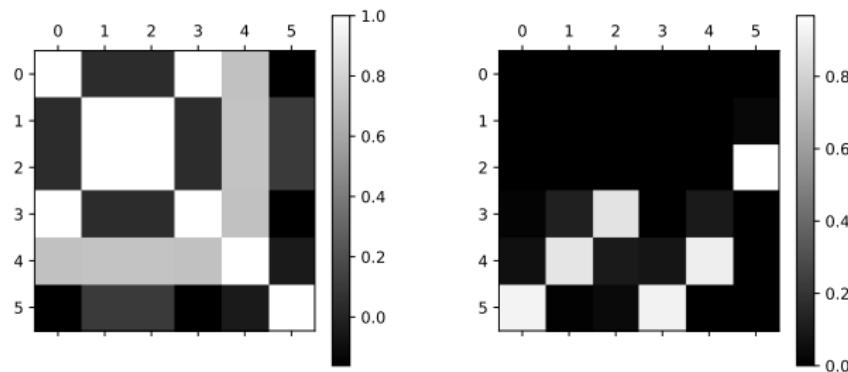
Оптимизационная задача для дистилляции:

$$\begin{aligned} \mathcal{L}(w, X, Y, f, \varphi) = & -\lambda \sum_{i=1}^m \sum_{r=1}^R \mathbb{I}[y_i = r] \log g^r(x_i, w) - \\ & - (1 - \lambda) \sum_{i=1}^m \sum_{r=1}^R (f \circ \varphi)^r(x_i) \log g^r(x_i, w). \end{aligned}$$

Прореживание нейросетей на основе анализа градиентов

Идея: оценить важность каждого параметра и удалить наименее значимые без потери качества.

- ▶ По **дисперсии градиента** — параметры с малой дисперсией слабо влияют на обучение.
- ▶ По **мультиколлинеарности** — параметры, линейно зависимые с другими, избыточны.



- ▶ Удаление до 80% параметров не снижает качества на простых моделях классификации.
- ▶ Метод Белсли эффективен при наличии мультиколлинеарности.

Антидистилляция: передача знаний от простой модели к сложной

Идея: модель-учитель (g_{tr}) обучена на простой задаче \mathcal{D}_1 ; нужно обучить более сложную модель-ученика (f_{st}) на сложной задаче \mathcal{D}_2 , используя знания учителя.

Проблема: размерности моделей разные ($N_{\text{tr}} < N_{\text{st}}$). Решение — составная функция потерь при инициализации:

$$\mathcal{L}(w) = \lambda_1 \mathcal{L}_{\text{ce}}(w, \mathcal{D}_1) + \lambda_2 \|u - \text{Pr}[w]\|^2 + \lambda_3 \mathcal{L}_3^\delta(w, \mathcal{D}_1) + \lambda_4 \text{tr}(\nabla^2 \mathcal{L}_{\text{ce}}).$$

Метод инициализации	Точность	FSGM-атака	Шум в параметрах
Xavier	0.68±0.08	0.42±0.04	0.58±0.06
Zero Pad	0.86±0.02	0.50±0.01	0.71±0.03
Uniform Pad	0.85±0.04	0.52±0.03	0.73±0.03
Transfer Learning	0.74±0.09	0.50±0.06	0.53±0.05
Net2Net	0.85±0.04	0.51±0.02	0.70±0.03
With Data Noise	0.81±0.07	0.51±0.03	0.70±0.05
Anti-Distillation ($\lambda_4 = 0$)	0.86±0.05	0.53±0.03	0.73±0.04
Anti-Distillation	0.86±0.05	0.57±0.03	0.67±0.03

Результаты: Anti-Distillation дает лучшую точность и устойчивость к адвокарным атакам (FSGM) и шуму в параметрах.

Радамахеровская сложность многозадачного обучения

Теорема (Грабовой-Грицай-Ремизова)

Пусть $S_t = \{x_i\}_{i=1}^n$ — выборка фиксированной целевой задачи t с $\|x_i\|_2 \leq R$. Пусть $\phi(\cdot; w)$ — энкодер, и рассмотрим линейные головы $f_{w,h}(x) = h^\top \phi(x; w)$ с $\|h\|_2 \leq B_{\text{head}}$. Предположим:

1. Ограничение на признаки: для всех x, w $\|\phi(x; w)\|_2 \leq L \|w\| \|x\|_2$.
2. Энкодер STL удовлетворяет $\|w_{\text{enc}}\| \leq B_{\text{enc}}$, общий энкодер MTL удовлетворяет $\|w_{\text{shared}}\| \leq B_{\text{shared}}$.
3. Многозадачное масштабирование: $B_{\text{shared}} \leq B_{\text{enc}}/\sqrt{T}$.

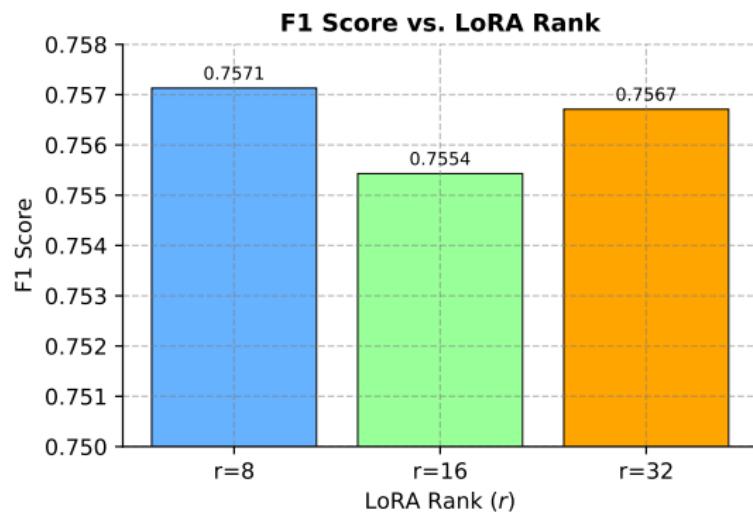
Тогда справедлива оценка сложности: $\widehat{\mathfrak{R}}_n(\mathcal{F}_{\text{MTL}}^{(t)}; S_t) \leq \frac{1}{\sqrt{T}} \widehat{\mathfrak{R}}_n(\mathcal{F}_{\text{STL}}^{(t)}; S_t)$.

Замечание

В постановке Бакстера, количество задач T само по себе способствует оценке общего индуктивного смещения: с ростом T сложность данных на задачу уменьшается пропорционально $1/T$. В отличие от этого, данный анализ сохраняет размер выборки целевой задачи n фиксированным и сравнивает STL и MTL на одном и том же n , поэтому улучшение проявляется как множитель $1/\sqrt{T}$ в сложности Радемахера на задачу.

LoRA-адаптеры в многозадачном обучении

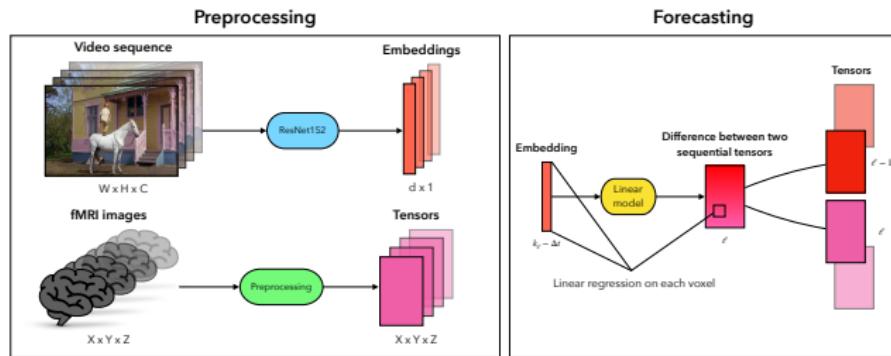
Цель: исследовать зависимость качества модели от ранга r LoRA-адаптеров и сопоставить с теоретическими предсказаниями.



- ▶ Наилучшая производительность достигается при $r = 8$; дальнейшее увеличение ранга не дает выигрыша.
- ▶ Подтверждает эффективность низкоранговой параметризации и теорему о состоятельности LoRA.
- ▶ Согласуется с теоретическим снижением сложности Радемахера в многозадачном обучении.

Декодирование фМРТ-изображений

Цель: применить методы снижения сложности данных для реконструкции фМРТ-снимков по видеоряду и оценить влияние сжатия на качество и скорость.



- ▶ Сжатие фМРТ-изображений в 2, 4 и 8 раз сокращает время обучения с 36 с до 6.7 с, 1.6 с и 1.4 с соответственно.
- ▶ Качество реконструкции (MSE) не ухудшается; оптимальный коэффициент регуляризации $\alpha \approx 1000$ сохраняется при всех степенях сжатия.
- ▶ Метод демонстрирует эффективность управления сложностью данных в реальной задаче нейровизуализации.

Положения, выносимые на защиту

1. Разработанный единый теоретический аппарат оценки сложности моделей глубокого обучения и сложности данных на основе теории мер и анализа ландшафта оптимизационной задачи, включающий формальные определения мер сложности и критерий обучаемости модели на выборке.
2. Введенная ландшафтная мера сложности модели, определяемая через спектральные свойства матриц Гессе функции потерь, задает количественную связь между архитектурными характеристиками модели и условной сложностью выборки.
3. Полученные теоретические оценки ландшафтной меры сложности для полно связных, сверточных и трансформерных архитектур раскрывают характер зависимости сложности моделей от их глубины, ширины и иных структурных параметров.
4. Получены теоретические оценки сходимости методов определения достаточного объема выборки, для которых установлена связь со сложностью моделей
5. Предложенные методы снижения сложности моделей глубокого обучения на основе анализа ковариационной матрицы градиентов, дистилляции на многодоменных данных и анти-дистилляции обеспечивают эффективное сокращение числа параметров и передачу знаний между моделями различной сложности и доменами данных.

Соответствие паспорту специальности 1.2.1

Пункт паспорта	Положения
2. Исследования в области оценки качества и эффективности алгоритмических и программных решений для систем искусственного интеллекта и машинного обучения. Методики сравнения и выбора алгоритмических и программных решений при многих критериях.	1, 4, 5
4. Разработка методов, алгоритмов и создание систем искусственного интеллекта и машинного обучения для обработки и анализа текстов на естественном языке, для изображений, речи, биомедицины и других специальных видов данных.	5
16. Исследования в области специальных методов оптимизации, проблем сложности и элиминации перебора, снижения размерности.	1, 2, 4, 5
17. Исследования в области многослойных алгоритмических конструкций, в том числе – многослойных нейросетей.	2, 3

Публикации по теме диссертации I

- Грицай Г. М., Грабовой А. В. Интерпретация классификаторов на основе архитектуры трансформер с помощью кластеризации // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2025. — Т. 527. — С. 432–448.
- Дорин Д. Д., Варламова К. Д., Грабовой А. В. Попарное сравнение изображений для обнаружения плагиата // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2025. — Т. 527. — С. 68–83.
- Дорин Д. Д., Грабовой А. В., Стрижов В. В. Улучшение декодирования данных ФМРТ с использованием пространственно-временных характеристик в условиях ограниченного набора данных // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2025. — Т. 527. — С. 11–30.
- Варламова К. Д., Дорин Д. Д., Грабовой А. В. За чертой знакомых доменов: исследование обобщающей способности детекторов машинно генерированных изображений // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2025. — Т. 527. — С. 103–116.
- Зверева А. К., Грабовой А. В., Каприелова М. С. Динамическое разделение труда в гибридном ии: стратегии кодировщиков и их воздействие на модуляторы на основе сетей долгой краткосрочной памяти // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2025. — Т. 527. — С. 117–133.
- Kiselev N. S., Grabovoy A. V. Sample size determination: Likelihood bootstrapping // Computational Mathematics and Mathematical Physics. — 2025. — Vol. 65, no. 2. — P. 416–423.
- Kiselev N., Grabovoy A. Sample size determination: posterior distributions proximity // Computational Management Science. — 2025. — Vol. 22, no. 1. — P. 1.
- Zvereva A. K., Kaprielova M., Grabovoy A. Anomlite: Efficient binary and multiclass video anomaly detection // Results in Engineering. — 2025. — Vol. 25. — P. 104162.
- G. Gritsai, A. Voznyuk, I. Khabutdinov, A. Grabovoy. Advacheck at genai detection task 1: Ai detection powered by domain-aware multi-tasking // Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect). — Abu Dhabi, UAE: International Conference on Computational Linguistics, 2025. — P. 236–243.
- Voznyuk A., Gritsai G., Grabovoy A. Advacheck at semeval-2025 task 3: Combining ner and rag to spot hallucinations in llm answers // Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025). — Vienna, Austria: Association for Computational Linguistics, 2025. — P. 1204–1210.
- Voznyuk A., Gritsai G., Grabovoy A. Team advacheck at pan: multitasking does all the magic // Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum. — Vol. 4038. — CEUR-WS.org, 2025. — P. 4007–4014.
- A. Levikin, I. Khabutdinov, A. Grabovoy, K. Vorontsov. The methodology of multi-criteria evaluation of text markup models based on inconsistent expert markup // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2025". — Moscow: JINR, 2025.

Публикации по теме диссертации II

13. I. Kopanichuk, A. Chashchin, A. Grabovoy et al. Structure extractor: Multilingual extraction of sections from scientific document // 2025 37th Conference of Open Innovations Association (FRUCT). — IEEE, 2025. — P. 122–128.
14. G. M. Gritsay, A. V. Grabovoy, A. S. Kildyakov, Y. V. Chekhovich Artificially generated text fragments search in academic documents // Doklady Mathematics. — 2024.
15. Gritsay G., Grabovoy A. Automated text identification on languages of the iberian peninsula: Llm and bert-based models aggregation // Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024). — Vol. 3756. — CEUR-WS.org, 2024.
16. Meshkov V., Kiselev N., Grabovoy A. Convnets landscape convergence: Hessian-based analysis of matricized networks // 2024 Ivannikov Ispras Open Conference (ISPRAS). — IEEE, 2024. — P. 1–10.
17. D. Dorin, N. Kiselev, A. Grabovoy, V. Strijov. Forecasting fmri images from video sequences: linear model analysis // Health Information Science and Systems. — 2024. — Vol. 12, no. 1. — P. 55.
18. Chekhovich Y., Grabovoy A., Gritsai G. Generative ai models with their full reveal* // 2024 4th International Conference on Technology Enhanced Learning in Higher Education (TELE). — Vol. 1. — IEEE, 2024. — P. 17–22.
19. M. Kaprielova, A. Grabovoy, K. Varlamova et al. Image plagiarism detection pipeline for vast databases // 2024 35th Conference of Open Innovations Association (FRUCT). — IEEE, 2024. — P. 328–335.
20. Gritsai G., Khabutdinov I., Grabovoy A. Multi-head span-based detector for ai-generated fragments in scientific papers // Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024). — Bangkok, Thailand: Association for Computational Linguistics, 2024. — P. 220–225.
21. Asvarov A., Grabovoy A. Neural machine translation system for lezgian, russian and azerbaijani languages // 2024 Ivannikov Ispras Open Conference (ISPRAS). — IEEE, 2024. — P. 1–7.
22. Poimanov D., Mestetsky L., Grabovoy A. N-gram perplexity-based ai-generated text detection // 2024 Ivannikov Ispras Open Conference (ISPRAS). — IEEE, 2024. — P. 1–8.
23. I. A. Khabutdinov, A. V. Chashchin, A. V. Grabovoy et al. Rugetor: Rule-based neural network model for russian language grammatical error correction // Programming and Computer Software. — 2024. — Vol. 50, no. 4. — P. 315–321.
24. Грицай Г. М., Хабутдинов И. А., Грабовой А. В. Stack more llms: эффективное обнаружение машинно-генерированных текстов с помощью аппроксимации значений перплексии // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2024. — Т. 520, № 2. — С. 228–237.
25. Boeva G., Gritsay G., Grabovoy A. Team ap-team at pan: Llm adapters for various datasets // Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). — Vol. 3740. — CEUR-WS.org, 2024. — P. 2527–2535.
26. A. V. Grabovoy, M. S. Kaprielova, A. S. Kildyakov et al. Text reuse detection in handwritten documents // Doklady Mathematics. — 2024.

Публикации по теме диссертации III

27. Asvarov A., Grabovoy A. The impact of multilinguality and tokenization on statistical machine translation // 2024 35th Conference of Open Innovations Association (FRUCT). — IEEE, 2024. — P. 149–157.
28. Киселев Н. С., Грабовой А. В. Раскрытие Гессиана: ключ к плавной сходимости поверхности функции потерь // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2024. — Т. 520, № 2. — С. 57–70.
29. G. Gritsay, A. Grabovoy, A. Kildyakov, Y. Chekhovich. Automated text identification: Multilingual transformer-based models approach // Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023). — Vol. 3496. — CEUR-WS.org, 2023.
30. Varlamova K., Khabutdinov I., Grabovoy A. Automatic spelling correction for russian: Multiple error approach // 2023 Ivannikov Ispras Open Conference (ISPRAS). — IEEE, 2023. — P. 169–175.
31. O. Bakhteev, Y. Chekhovich, A. Grabovoy et al. Cross-language plagiarism detection: A case study of european languages academic works // Academic Integrity: Broadening Practices, Technologies, and the Role of Students. — Vol. 4 of Ethics and Integrity in Educational Contexts. — New York: Springer International Publishing, 2023. — P. 143–161.
32. Avetisyan K., Gritsay G., Grabovoy A. Cross-lingual plagiarism detection: Two are better than one // Programming and Computer Software. — 2023. — Vol. 49, no. 4. — P. 346–354.
33. D. Shodiev, I. Kopanichuk, A. Chashchin et al. Ensembling models for the generation of queries to an altering search engine using reinforcement learning // 2023 Ivannikov Ispras Open Conference (ISPRAS). — IEEE, 2023. — P. 144–149.
34. I. Potyashin, M. Kaprieva, Y. Chekhovich et al. Hwr200: New open access dataset of handwritten texts images in russian // Proceedings of the International Conference "Dialogue 2023". — 2023.
35. Grashchenkov K., Grabovoy A., Khabutdinov I. A method of multilingual summarization for scientific documents // 2022 Ivannikov Ispras Open Conference (ISPRAS). — IEEE, 2022.
36. K. Petrushina, O. Bakhteev, A. Grabovoy, V. Strijov Anti-distillation: Knowledge transfer from a simple model to the complex one // 2022 Ivannikov Ispras Open Conference (ISPRAS). — IEEE, 2022.
37. Gritsay G., Grabovoy A., Chekhovich Y. Automatic detection of machine generated texts: Need more tokens // Ivannikov Memorial Workshop Proceedings 2022. — 2022.
38. A. V. Grabovoy, T. T. Gadaev, A. P. Motrenko, V. V. Strijov. Numerical methods of sufficient sample size estimation for generalised linear models // Lobachevskii Journal of Mathematics. — 2022. — Vol. 43, no. 9. — P. 2453–2462.
39. Грабовой А. В., Стрижков В. В. Вероятностная интерпретация задачи дистилляции // Автоматика и телемеханика. — 2022. — № 1. — С. 150–168.

Публикации по теме диссертации IV

40. Базарова А. И., Грабовой А. В., Стрижов В. В. Анализ свойств вероятностных моделей в задачах обучения с экспертом // Автоматика и телемеханика. — 2022. — № 10.
41. Grabovoy A. V., Strijov V. V. Bayesian distillation of deep learning models // Automation and Remote Control. — 2021. — Vol. 82, no. 11. — P. 1846–1856.
42. Grabovoy A. V., Strijov V. V. Prior distribution selection for a mixture of experts // Computational Mathematics and Mathematical Physics. — 2021. — Vol. 61, no. 7. — P. 1140–1152.
43. Grabovoy A., Bakhteev O., Chekhovich Y. The automatic approach for scientific papers dating // Proceedings of the 2020 Ivannikov Ispras Open Conference. — Los Alamitos, CA: IEEE Computer Society Press, 2021.
44. Грабовой А. В., Стрижов В. В. Анализ выбора априорного распределения для смеси экспертов // Журнал вычислительной математики и математической физики. — 2021. — Т. 61, № 7. — С. 1149–1161.
45. Грабовой А. В., Стрижов В. В. Байесовская дистилляция моделей глубокого обучения // Автоматика и телемеханика. — 2021. — № 11. — С. 16–29.
46. Грабовой А. В., Бахтеев О. Ю., Стрижов В. В. Введение отношения порядка на множестве параметров аппроксимирующих моделей // Информатика и ее применения. — 2020. — Т. 14, № 2.
47. Grabovoy A. V., Strijov V. V. Quasi-periodic time series clustering for human activity recognition // Lobachevskii Journal of Mathematics. — 2020. — Vol. 41, no. 3. — P. 333–339.
48. Грабовой А. В., Бахтеев О. Ю., Стрижов В. В. Определение релевантности параметров нейросети // Информатика и ее применения. — 2019.

По теме диссертации опубликовано 56 научных работ, из которых 17 статей в научно-технических журналах, входящих в перечень ВАК, 32 — в изданиях, входящих в международные научометрические базы Scopus и Web of Science. В трудах российских и международных конференций опубликовано 39 работ. Также на основе работ автора зарегистрировано 13 программ для ЭВМ.

Основные результаты, выносимые на защиту

1. Разработан единый формальный аппарат для оценки сложности моделей и данных на основе теории мер и анализа ландшафта оптимизационной задачи.
2. Введены формальные определения меры сложности выборки и меры сложности модели, установлен критерий обучаемости модели на выборке.
3. Введена ландшафтная мера сложности модели, определяемая через спектральные свойства матриц Гессе функции потерь.
4. Теоретически доказана связь ландшафтной меры со сложностью выборки.
5. Получены теоретические оценки ландшафтных мер на основе спектральных норм матриц Гессе для основных архитектур глубокого обучения: полносвязных, сверточных и трансформерных сетей.
6. Доказаны теоремы о зависимости ландшафтных мер от глубины сети, размеров слоев и других структурных параметров архитектур.
7. Доказана сходимость методов оценки достаточного объема выборок из простой генеральной совокупности.
8. Разработаны методы мультидоменной дистилляции и анти-дистилляции для передачи знаний между моделями различной сложности и между различными доменами данных.
9. Для многозадачного обучения доказаны теоремы о статистической состоятельности низкоранговых адаптеров и снижении эмпирической сложности.