Локальные свойства нейронных сетей через призму гессианов слоёв

Максим Большим

May 29, 2025

Abstract

В данной работе представлен новый подход к анализу нейронных сетей, основанный на исследовании локальных свойств их параметрического пространства с помощью матриц Гессе. Введено понятие локального гессиана, позволяющее анализировать геометрию функционального пространства отдельных слоев нейронной сети. Предложены методы количественной оценки таких феноменов, как переобучение и недостаточная аппроксимирующая способность моделей, через спектральные характеристики локальных гессианов. Проведённый анализ 111 экспериментов на 37 датасетах выявил закономерности в структуре локальных матриц гессиана на различных этапах обучения, что может служить основой для дальнейшего анализа и совершенствования архитектур нейронных сетей.

1 Введение

Глубокие нейронные сети продемонстрировали выдающиеся результаты во многих областях, включая компьютерное зрение, обработку естественного языка и прочие задачи машинного обучения [5, 6]. Однако, несмотря на их практический успех, остаётся открытым вопрос о том, почему одни архитектуры превосходят другие и как формально и систематизированно улучшать дизайн нейронных сетей. Эмпирический подход, основанный на методе проб и ошибок, становится всё более затратным с ростом размеров моделей и объёмов данных.

Ряд работ показывает, что анализ кривизны ландшафта функции потерь через гессианы и родственные спектральные инструменты может пролить свет на динамику обучения и обобщающую способность моделей [7, 2, 3, 4]. В настоящем исследовании выдвигается тезис о том, что локальные свойства параметрического пространства нейронной сети могут дать предварительную оценку внутренних свойств модели уже на ранних этапах обучения [1]. Конкретно, мы предлагаем использовать локальные матрицы Гессе — матрицы вторых производных целевой функции по параметрам отдельных слоёв — для анализа геометрии пространства параметров.

Концепция локального гессиана позволяет формализовать и количественно измерить геометрические свойства пространства параметров в окрестности точки оптимизации. В частности, мы показываем, что спектр локального гессиана, такой как распределение собственных значений и их структура [8], тесно связан с функциональными свойствами соответствующих слоёв нейронной сети.

Основные вклады настоящей работы:

- Введение математически строгого определения локального гессиана для функциональных блоков нейронной сети

- Детальный анализ спектральных свойств локальных гессианов в процессе обучения нейронных сетей
- Исследование геометрической интерпретации пространства параметров нейронной сети через призму локальных гессианов

Полученные результаты не только углубляют наше теоретическое понимание глубоких нейронных сетей, но и открывают новые перспективы для изучения их внутренней динамики.

2 Математический аппарат нейронных сетей

2.1 Определение и структура нейронной сети

Определение 1. *Нейронная сеть $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^m$ представляет собой параметризованную функцию с параметрами $\theta \in \mathbb{R}^P$, отображающую входные данные $x \in \mathbb{R}^d$ в выходное пространство через последовательность функциональных преобразований. Обозначим через $\mathcal{F}(x;\theta)$ результат применения сети к входным данным при заданных параметрах θ.*

Определение 2. *Функциональным блоком (слоем) C_i нейронной сети $\mathcal{F}$ называется пара модулей (P_i, A_i), где:*

- $P_i : \mathbb{R}^{d_i} \times \mathbb{R}^{p_i} \to \mathbb{R}^{q_i}$ *– параметризованное преобразование с параметрами $\theta_i \in \mathbb{R}^{p_i}$.*
- $A_i : \mathbb{R}^{q_i} \to \mathbb{R}^{q_i}$ *– функция активации (потенциально тождественная)*

Определение 3. *Нейронную сеть $\mathcal{F}$ можно представить как композицию n функциональных блоков:*

$$\mathcal{F}(x;\theta) = (C_n \circ C_{n-1} \circ \ldots \circ C_1)(x), \tag{1}$$

где $C_i(z) = A_i(P_i(z;\theta_i))$ для входа z и $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ – полный набор параметров сети.

Такое представление нейронной сети позволяет провести анализ каждого функционального блока отдельно, что важно для локального анализа свойств сети. Декомпозиция сложной модели на более простые составляющие является ключевым методологическим приемом, который позволяет применить инструменты спектрального анализа к отдельным компонентам.

2.2 Промежуточные представления и функции активации

Определение 4. *Промежуточным представлением z_i называется вход в блок C_i:*

$$z_i = \begin{cases} x, & \text{если } i = 1 \\ (C_{i-1} \circ \ldots \circ C_1)(x), & \text{если } i > 1 \end{cases} \tag{2}$$

Соответственно, выход блока C_i обозначается как:

$$y_i = C_i(z_i) = A_i(P_i(z_i;\theta_i)) \tag{3}$$

Промежуточные представления играют важную роль в анализе нейронных сетей, поскольку они содержат информацию о том, как входной сигнал преобразуется на каждом этапе обработки. Особый интерес представляет исследование геометрии этих промежуточных представлений и их взаимосвязь с параметрами соответствующих слоев.

Определение 5. *Для блока C_i определим локальную скалярную функцию $S_i : \mathbb{R}^{p_i} \to \mathbb{R}$ как:*

$$S_i(\theta_i) = \varphi(A_i(P_i(z_i; \theta_i))), \tag{4}$$

где $\varphi : \mathbb{R}^{q_i} \to \mathbb{R}$ – функция агрегации, обычно $\varphi(y) = \sum_{j=1}^{q_i} y_j$.

Скалярная функция представляет собой способ оценки влияния параметров конкретного слоя на его выход при фиксированном входе. Эта функция будет ключевой для определения локального гессиана в следующем разделе.

2.3 Типичные реализации в нейросетях

В контексте современных нейронных сетей часто используются следующие реализации компонентов:

- P_i – линейное преобразование $P_i(z_i; \theta_i) = W_i z_i + b_i$, где $\theta_i = \{W_i, b_i\}$
- A_i – нелинейная функция активации, например, ReLU, Sigmoid или Tanh
- $\varphi(y_i) = \sum_{j=1}^{q_i} y_{i,j}$ – суммирование всех компонент выходного вектора

Эти определения и обозначения будут использоваться во всех последующих разделах данной работы для обеспечения математической строгости и последовательности.

3 Локальные гессианы нейронной сети

3.1 Определение локального гессиана

Определение 6. *Локальной матрицей Гессе $H_i \in \mathbb{R}^{p_i \times p_i}$ (далее для краткости — локальный гессиан, LH_i) для блока C_i называется матрица вторых производных скалярной функции S_i по параметрам θ_i:*

$$H_i = \nabla^2_{\theta_i} S_i(\theta_i) = \left[\frac{\partial^2 S_i(\theta_i)}{\partial \theta_{i,j} \partial \theta_{i,k}} \right]_{j,k=1}^{p_i} \tag{5}$$

Анализ этой матрицы позволяет получить информацию о:

- Степени нелинейности преобразования, выполняемого слоем
- Взаимосвязи между параметрами и их влиянии на выход слоя
- Геометрических свойствах пространства параметров
- Чувствительности слоя к малым изменениям параметров

В дифференциальной геометрии гессиан функции в точке определяет квадратичную форму, которая аппроксимирует кривизну поверхности уровня этой функции. В контексте нейронных сетей LH_i характеризует кривизну функционального отклика слоя в пространстве его параметров [9]. При анализе поверхности следует учитывать знак собственных значений LH_i и их распределение, так как это определяет локальную геометрию.

3.2 Эффективное вычисление LH_i

Анализ нейронных сетей с использованием LH_i представляет собой эффективный инструмент, однако из-за квадратичной зависимости размера данной матрицы от числа параметров её прямое применение часто оказывается вычислительно непрактичным. В связи с этим на практике широко применяются различные методы аппроксимации LH_i [10, 11, 12, 13]. В данном разделе предлагается методика работы с LH_i, позволяющая обойти указанные ограничения.

Для эффективного вычисления LH_i предлагается алгоритм, основанный на последовательном вычислении строк матрицы.

Лемма 1. *Элементы матрицы Гессе H_i можно вычислять последовательно по строкам:*

$$\begin{aligned} g_i &= \nabla_{\theta_i} S_i(\theta_i) = \left[\frac{\partial S_i}{\partial \theta_{i,j}}\right]_{j=1}^{p_i} \\ H_i[j,:] &= \nabla_{\theta_i} g_{i,j} = \nabla_{\theta_i}\left(\frac{\partial S_i}{\partial \theta_{i,j}}\right) \end{aligned} \tag{6}$$

Proof. По определению матрицы Гессе, её элемент $H_i[j,k]$ равен:

$$H_i[j,k] = \frac{\partial^2 S_i(\theta_i)}{\partial \theta_{i,j} \partial \theta_{i,k}} \tag{7}$$

Если обозначить $g_{i,j} = \frac{\partial S_i}{\partial \theta_{i,j}}$, то

$$H_i[j,k] = \frac{\partial g_{i,j}}{\partial \theta_{i,k}} \tag{8}$$

Таким образом, j-я строка H_i представляет собой градиент j-й компоненты градиента функции S_i. □

Это позволяет значительно сократить вычислительные затраты при работе с большими моделями, так как не требует одновременного хранения в памяти всей матрицы Гессе размером $p_i \times p_i$.

Предложенный метод особенно важен при анализе современных глубоких нейронных сетей, содержащих миллионы параметров, поскольку полная матрица LH_i для таких моделей была бы непомерно большой. Локальный подход не только делает вычисления практически реализуемыми, но и позволяет сосредоточиться на анализе отдельных компонентов сети, что часто более информативно, чем глобальный анализ.

3.3 Алгоритм вычисления локального гессиана

3.4 Математические детали реализации

3.4.1 Вычисление градиента g_i

В контексте автоматического дифференцирования градиент g_i вычисляется как:

$$g_i = \nabla_{\theta_i} S_i = \frac{\partial S_i}{\partial y_i} \cdot \frac{\partial y_i}{\partial P_i} \cdot \frac{\partial P_i}{\partial \theta_i} \tag{9}$$

где:

- $\frac{\partial S_i}{\partial y_i} = \nabla_{y_i} \varphi(y_i)$ – градиент функции агрегации
- $\frac{\partial y_i}{\partial P_i} = \nabla_{P_i} A_i(P_i)$ – якобиан активационной функции
- $\frac{\partial P_i}{\partial \theta_i}$ – якобиан параметризованного преобразования по его параметрам

Algorithm 1 Вычисление локальных гессианов нейронной сети

Require: Нейронная сеть $\mathcal{F}$, входные данные $x \in \mathbb{R}^d$, функция агрегации φ
Ensure: Набор локальных матриц Гессе $\{LH_1, LH_2, \ldots, LH_n\}$

Разбить $\mathcal{F}$ на функциональные блоки $\{C_1, C_2, \ldots, C_n\}$, где $C_i = (P_i, A_i)$
for $i = 1$ до n **do**
    Вычислить $z_i = (C_{i-1} \circ \ldots \circ C_1)(x)$ ▷ Вход в блок C_i
    Вычислить $y_i = A_i(P_i(z_i; \theta_i))$ ▷ Выход блока C_i
    Вычислить $S_i = \varphi(y_i)$ ▷ Скалярная функция блока
    Вычислить градиент $g_i = \nabla_{\theta_i} S_i$
    Инициализировать $LH_i \in \mathbb{R}^{p_i \times p_i}$ нулевой матрицей
    for $j = 1$ до p_i **do**
        if $g_{i,j}$ не является константой относительно θ_i **then**
            Вычислить $LH_i[j, :] = \nabla_{\theta_i} g_{i,j}$
        else
            $LH_i[j, :] = \vec{0}$
        end if
    end for
end for
return $\{LH_1, LH_2, \ldots, LH_n\}$

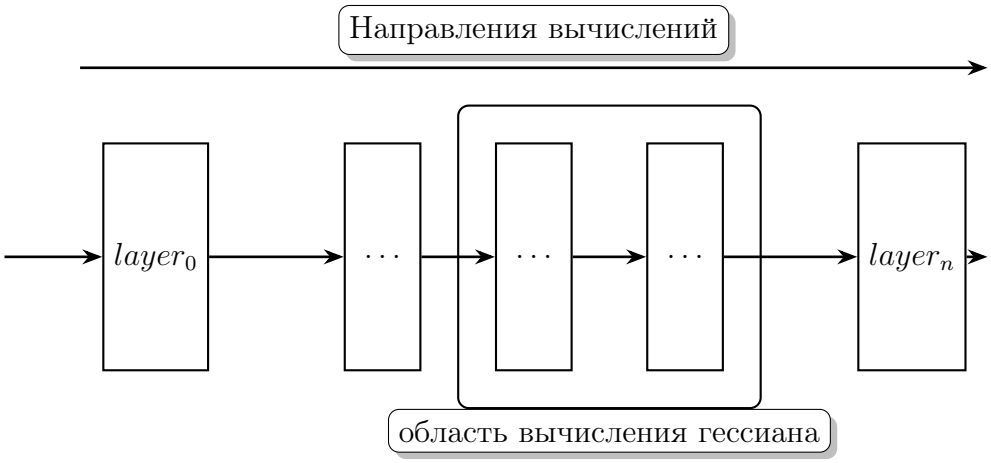

Figure 1: Визуализация вычисления локального гессиана

3.4.2 Вычисление строк матрицы Гессе

Для каждой компоненты j градиента g_i вычисляется её градиент по параметрам θ_i:

$$H_i[j, :] = \nabla_{\theta_i} g_{i,j} = \nabla_{\theta_i} \left(\frac{\partial S_i}{\partial \theta_{i,j}} \right) \tag{10}$$

Это требует повторного применения автоматического дифференцирования к каждой компоненте градиента.

3.5 Спектральный анализ локального гессиана

Детальное изучение спектра LH_i предоставляет важный инструмент для понимания геометрии пространства параметров. Распределение и структура собственных значений отражают ключевые свойства кривизны функции, что особенно важно для анализа степени обусловленности задачи оптимизации [14, 15].

Для каждого LH_i можно вычислить:

$$LH_i = U_i \Lambda_i U_i^T = \sum_{j=1}^{p_i} \lambda_{i,j} u_{i,j} u_{i,j}^T \tag{11}$$

где $\lambda_{i,j}$ - j-е собственное значение, а $u_{i,j}$ - соответствующий собственный вектор.
Характерные показатели спектра гессиана включают:

- **Следы гессиана:** $\mathrm{tr}(LH_i) = \sum_{j=1}^{p_i} \lambda_{i,j}$ — сумма собственных значений
- **Определитель гессиана:** $\det(LH_i) = \prod_{j=1}^{p_i} \lambda_{i,j}$ — произведение собственных значений

Особый интерес представляет наблюдение за распределением собственных значений по слоям сети и их изменениями в процессе обучения, что позволяет отслеживать эволюцию геометрии пространства параметров.

Распределение собственных значений LH_i даёт информацию о геометрии функционального пространства слоя. В частности, концентрация собственных значений вблизи нуля указывает на наличие многообразий равных значений функции (плато), что затрудняет оптимизацию градиентными методами.

- **Малые сети:** Если сеть слишком мала (мало слоёв d_l или узкие слои), линейное преобразование $z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}$ становится некорректно выраженным, что приводит к большому среднему $\mathbb{E}[|z^{(l)}|] \gg 1$ и попаданию $z^{(l)}$ в области насыщения функций типа сигмоида или tanh, где $f'(z) \approx 0$. Это усугубляет проблему исчезающих градиентов и существенно влияет на структуру LH_i, концентрируя собственные значения вблизи нуля.

- **Слишком большие сети:** Слишком большая сеть способна «выучить» шумовые детали данных, что приводит к росту нормы весов $||W^{(l)}|| \gg 1$ и смещению $z^{(l)}$ в область насыщения. В таких условиях LH_i также принимает специфическую структуру с большим количеством очень малых собственных значений, что отражает излишнюю свободу в параметрическом пространстве.

- **Неподходящая архитектура или входы:** Архитектура может не подходить для специфических свойств данных (высокая нелинейность, вариативность распределений, многомерные связи). Плохо нормализованные или шумные входы дополнительно увеличивают разброс $z^{(l)}$, усиливая насыщение. В результате нейроны «замирают» и перестают эффективно участвовать в обучении.

4 Методология исследования

4.1 Дизайн экспериментов

Для оценки спектральных свойств локальных гессианов был проведен комплексный анализ поведения нейронных сетей различной архитектуры на наборе из 37 датасетов (22 для задач классификации и 15 для регрессии). При этом для каждого датасета варьировались следующие параметры:

- Количество слоёв и нейронов в сетях
- Способы инициализации весов

- Методы оптимизации (Adam, SGD, RMSProp)
- Функции активации (ReLU, Sigmoid, Tanh)

Общее число параметров в исследуемых моделях варьировалось от 13 до 9 миллионов, число слоёв варьировалось от 1 до 5, что позволило проанализировать сети с разным уровнем параметризации. Для решения обоих задач ставился классический многослойный персептрон как основная архитектура. В качестве функции ошибки использовалась кросс-энтропия для задач классификации и среднеквадратичная ошибка для задач регрессии. Для каждого датасета гиперпараметры подбирались эмпирическим путём.

4.2 Сбор экспериментальных данных

Для сбора данных была разработана специализированная система, позволяющая отслеживать изменение внутренних характеристик нейронных сетей во время обучения. Эксперименты проводились по следующей методологии:

1. Для каждого датасета были обучены три варианта модели:
   - Модель с небольшим количеством параметров (тип “no”)
   - Модель со средним количеством параметров (тип “sure”)
   - Модель с большим количеством параметров (тип “huge”)
2. В процессе обучения на каждой контрольной итерации сохранялись следующие данные:
   - Веса модели и их спектральные характеристики (распределение, статистика)
   - Градиенты по параметрам и их спектральные характеристики
   - Матрицы LH_i по всем слоям и их собственные значения
   - Метрики качества модели (для классификации: Accuracy, Precision, Recall, F1, AUC; для регрессии: R2, MAE, RMSE)
   - Значение функции потерь на обучающей выборке

Особое внимание уделялось вычислению LH_i, для которых применялся специально разработанный эффективный алгоритм с покомпонентным вычислением и оптимизацией памяти. Это позволило рассчитывать гессианы даже для моделей с большим количеством параметров.

Для каждой модели производилось от 50 до 150 контрольных точек в зависимости от скорости сходимости, что в сумме привело к накоплению около 1500 снимков состояний сети общим объёмом около 50 гигабайт данных.

4.3 Методология и обработка экспериментальных данных

Для анализа собранных экспериментальных данных был применён многоэтапный подход, включающий следующие методы:

1. **Корреляционный анализ:** расчёт коэффициентов корреляции Пирсона между параметрами модели и показателями качества, визуализация результатов с помощью тепловых карт.

2. **Спектральный анализ:** исследование распределения собственных значений весов и LH_i, вычисление статистических характеристик (среднее, стандартное отклонение, экстремумы).

3. **Канонический корреляционный анализ (CCA):** исследование нелинейных связей между группами метрик качества (Accuracy, Precision, F1 и т.д.) и характеристиками внутренних параметров сети. Все данные были стандартизированы для обеспечения корректности анализа.

4. **Визуализация:** применение тепловых карт, диаграмм рассеяния и других графических представлений для наглядной интерпретации результатов.

5. **Статистические тесты:** проверка распределений характеристик с помощью теста Шапиро-Уилка [16] для выявления отклонений от нормальности.

Особое внимание уделялось спектральному анализу LH_i и их корреляции с метриками качества моделей. Данный комплексный подход позволил выявить не только прямые корреляции между отдельными параметрами, но и более сложные взаимосвязи между группами параметров.

5 Результаты исследования

5.1 Результаты первичного исследования

Эксперименты проводились на вычислительных кластерах с использованием GPU для ускорения вычислений. Для анализа использовались библиотеки Python, такие как NumPy, SciPy, Matplotlib и Seaborn. Сбор всех данных занял около 40 часов, включая время на обучение моделей и вычисление LH_i. Это было связано с тем, что экспериментальная реализация локального гессиана была не оптимизирована, и для каждой итерации обучения приходилось пересчитывать все градиенты и гессианы заново. Однако в дальнейшем была разработана оптимизированная версия алгоритма, которая позволила значительно сократить время вычислений.

На основе полученных данных можно выделить несколько ключевых наблюдений:

- Анализ спектра LH_i предоставляет ценную информацию о внутренней структуре и функционировании нейронных сетей. Сети с различной архитектурой демонстрируют характерные паттерны в спектральных свойствах гессианов, хотя формальная интерпретация всех наблюдаемых феноменов остаётся открытым вопросом.

- Необходимо провести более подробный анализ динамики эволюции LH_i между слоями. Следует усилить применение канонического корреляционного анализа (CCA) и рассмотреть возможность использования факторного анализа для выявления скрытых зависимостей. Текущие методы анализа не всегда позволяют выявить все возможные зависимости между параметрами и метриками качества. Чаще эти проблемы сигнализируют о наличии нескольких проблем архитектуры сразу, которые могут быть не связаны между собой. Такими инструментами трудно анализировать слабые стороны сети, если гиперпараметры обучения подобраны неудачно.

- Дополнительно рекомендуется исследовать геометрию потока градиента и свойства гессиана в контексте дифференциальной геометрии, а также рассмотреть

возможность моделирования градиентного потока на многообразии параметров в предельном случае перенасыщения функций активации.

5.2 Сравнительный анализ архитектурных решений на основе CCA

Исследование трёх типов архитектур с различной параметризацией (условно обозначенных как “no” — малая, “sure” — умеренная, “huge” — избыточная) показало существенные различия в спектральных свойствах их LH_i. Для выявления связей между параметрами моделей и их производительностью был применён канонический корреляционный анализ (CCA), который позволил установить корреляции между двумя группами переменных:

- **Группа A**: метрики качества модели — Accuracy, Precision, Recall, F1, AUC и train_loss
- **Группа B**: параметры нейронной сети — веса, градиенты, собственные значения гессианов и их спектральные характеристики

CCA анализ выявил следующие зависимости:

Статистика	**no**	**sure**	**huge**
max	0,406	0,349	0,429
avg	-0,955	0,099	0,220
median	0,182	0,182	0,189
min	-0.965	-0,770	0,149
std	0.976	0,277	0,082

Table 1: Статистика CCA-корреляций

Эти данные наглядно подтверждаются представленным графиком сравнения статистик CCA Score (рис. 2), где видны существенные различия в минимальных значениях между архитектурами и особенно заметна высокая вариативность средней архитектуры (“sure”) с минимальным значением около -0,77, что хотя и выше, чем у малой архитектуры (“no”), но значительно ниже, чем у большой архитектуры (“huge”).

Результаты показывают, что большие архитектуры (“huge”) демонстрируют наиболее высокие и стабильные значения CCA-корреляций (стандартное отклонение всего 0,082), что указывает на более устойчивую связь между внутренними параметрами сети и качеством предсказаний. Малые архитектуры (“no”), напротив, обнаруживают экстремальные отрицательные выбросы как в среднем значении (-0,955), так и в минимуме (-0,965), а также высокую вариативность (стандартное отклонение 0,976), что свидетельствует о неустойчивости их функционального поведения.

5.3 Спектральные характеристики градиентов в различных архитектурах

Для анализа спектральных характеристик градиентов была использована функция Велша для оценки плотности мощности (PSD) градиентов по слоям. Для метода Велша были выбраны следующие параметры:

- Длина окна: 256

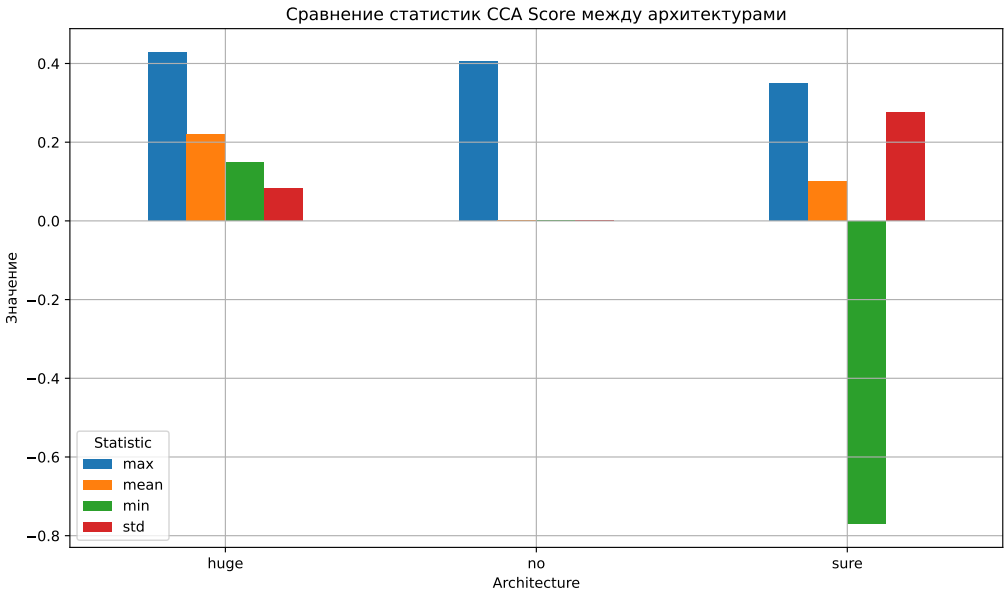

Figure 2: Сравнение статистик CCA Score между архитектурами

- Шаг перекрытия: 128
- Функция окна: Ханнинг

Анализ спектральных характеристик градиентов третьего слоя (рис. 3) выявил драматические различия между исследуемыми архитектурами. Максимальные значения плотности мощности (PSD) по методу Велша для архитектуры “huge” превышают аналогичные показатели архитектуры “no” более чем в 100 раз, достигая значений порядка $1{,}2\times10^{6}$. Средние значения также демонстрируют значительный рост от малой к большой архитектуре, что свидетельствует о принципиально иной структуре градиентного пространства в сильно параметризованных моделях.

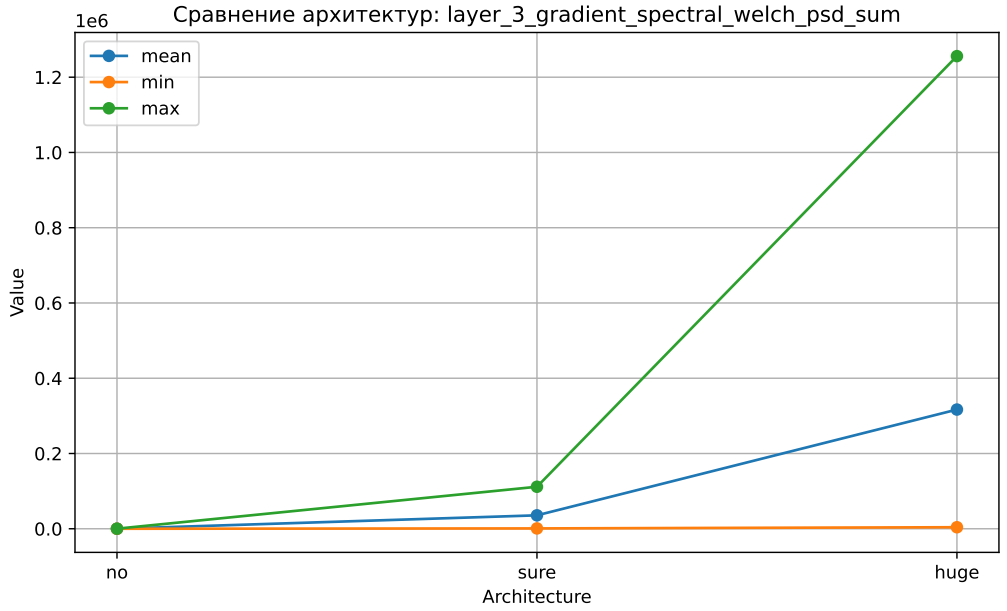

Figure 3: Сравнение спектральных характеристик градиентов третьего слоя

Такой масштаб различий указывает на качественное изменение характера

распространения градиентов в больших архитектурах, где формируются высокочастотные компоненты с существенно большей энергией. Эти наблюдения соотносятся с представлением о том, что избыточная параметризация способствует формированию более сложной структуры функциональной поверхности с множеством локальных особенностей.

5.4 Распределение канонических весов по архитектурам

Анализ распределения канонических X и Y позволил выявить структурные особенности влияния параметров сети на ее производительность. На представленных визуализациях (рис. 4 и 5) видны значительные различия в распределении весов между архитектурами.

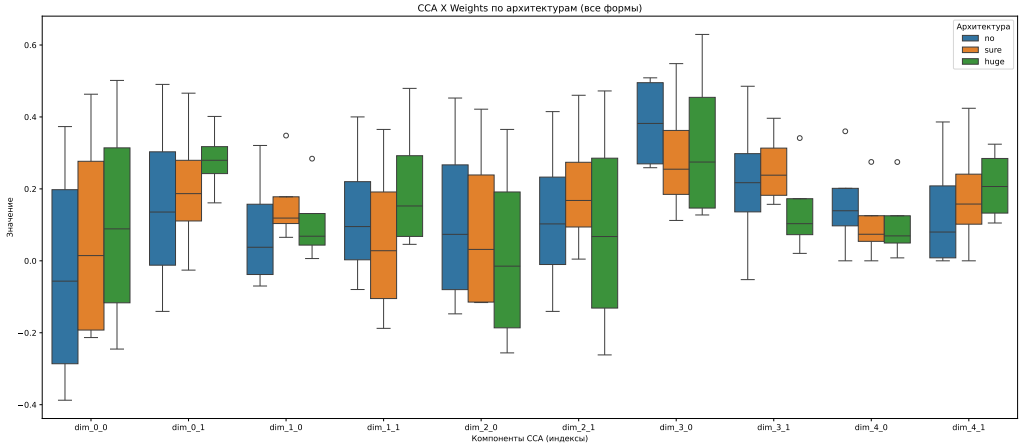

Figure 4: Распределение канонических X-весов по архитектурам

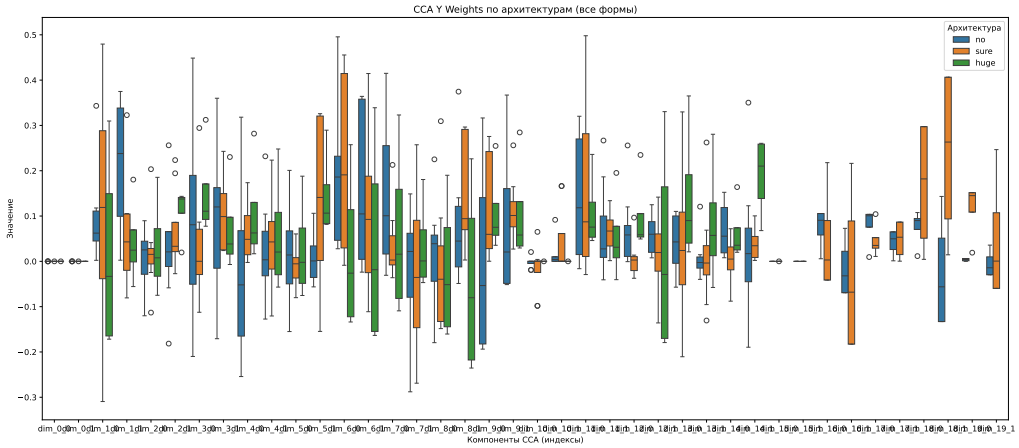

Figure 5: Распределение канонических Y-весов по архитектурам

Наблюдается следующая закономерность: для X-весов (связывающих метрики качества с каноническими переменными) большие архитектуры (“huge”) демонстрируют более равномерное распределение по всему спектру компонент, что указывает на более сбалансированное использование всего набора параметров для достижения высокой производительности. В то же время, малые архитектуры (“no”) характеризуются более концентрированной структурой с отдельными доминирующими компонентами, что свидетельствует о “перенапряжении” отдельных параметров для достижения результата.

Для Y-весов, связывающих параметры нейронной сети с каноническими переменными, наблюдается ещё более выраженная дифференциация: в больших архитектурах

распределение более плотное и центрировано вокруг нуля с небольшими выбросами, тогда как в малых архитектурах наблюдается значительная асимметрия с экстремальными значениями на концах распределения. Это подтверждает гипотезу о том, что в недопараметризованных моделях отдельные параметры несут непропорционально высокую нагрузку, что снижает устойчивость модели к изменениям входных данных.

5.5 Динамика канонических весовых коэффициентов

Анализ канонических весовых коэффициентов выявил значительные различия между архитектурами. Каждый канонический вектор представляет собой линейную комбинацию исходных переменных, максимизирующую корреляцию между группами A и B. Особенно примечательны различия в структуре весов Y компоненты, связывающей параметры сети с метриками качества:

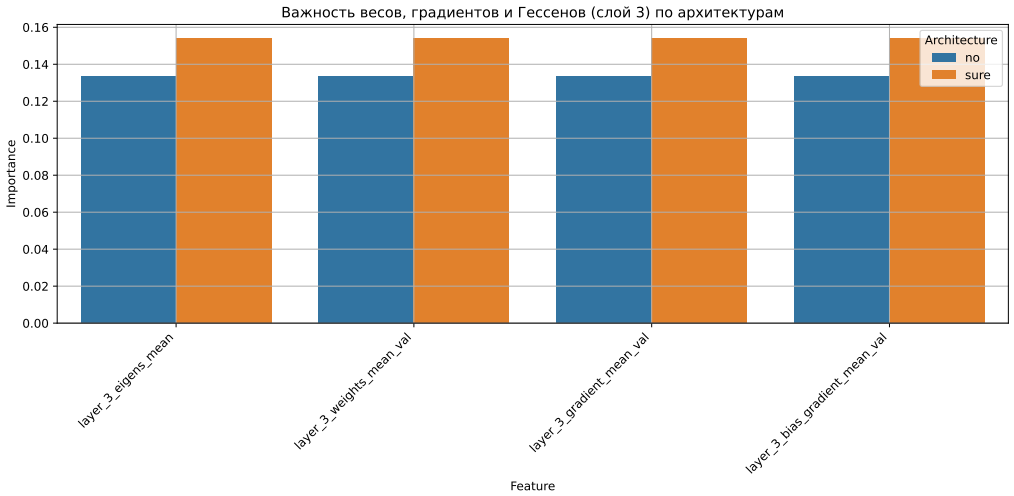

Figure 6: Важность параметров из группы B (веса, градиенты и гессианы) третьего слоя для различных архитектур

В малых архитектурах наблюдается значительное преобладание весов, связанных с градиентной составляющей, в то время как в больших архитектурах более важными становятся собственные значения гессиана. Это подтверждает теоретическое предположение о том, что в недопараметризованных моделях динамика обучения в большей степени определяется локальными градиентами, тогда как в избыточно параметризованных моделях — кривизной функциональной поверхности.

Детальный анализ дифференцирующих весовых параметров группы B (таблица 2) выявил экстремальные различия между архитектурами:

Table 2: Параметры группы B с максимальными различиями между архитектурами

Huge	**no**	**sure**	**huge - no**	**huge - sure**
4384.0	3.42	212.92	4380.58	4171.08
1208.0	30.85	376.31	1177.15	831.69
620.67	3.88	37.54	616.78	583.13
196.0	12.0	46.0	184.0	150.0

Эти данные демонстрируют радикальные различия в величине параметров между большой и малой архитектурами, достигающие трёх порядков (4380.58). Такой контраст указывает на качественно иной режим функционирования переобученных сетей, где

накопление весов может достигать значительных величин без негативного влияния на качество предсказаний благодаря компенсационным эффектам между слоями.

5.6 Статистические свойства канонических корреляционных весов

Дополнительный анализ статистических свойств CCA-весов (таблица 3) выявил интересную асимметрию в распределении вариативности между размерностями:

Table 3: Статистические свойства весов CCA для различных архитектур и форм

Архитектура	**Тип весов**	**Форма**	**Ср. дисперсия (изм. 0)**	**Ср. дисперсия (изм. 1)**
no	X-веса	(5, 2)	0.1899	0.1315
sure	X-веса	(5, 2)	0.2159	0.1274
huge	X-веса	(5, 2)	0.2064	0.1553
no	Y-веса	(15, 2)	0.0000	0.1292
sure	Y-веса	(15, 2)	0.0000	0.1671
huge	Y-веса	(15, 2)	0.0000	0.0643
no	Y-веса	(20, 2)	0.0000	0.0665
sure	Y-веса	(20, 2)	0.0000	0.0003

Примечательно практически нулевое значение дисперсии для первой размерности Y-весов во всех архитектурах, что указывает на строгую структурированность связей между параметрами сети и показателями качества по определённым направлениям. В то же время, X-веса, отражающие вклад параметров нейросети в канонические переменные, демонстрируют существенную дисперсию по обеим размерностям, что свидетельствует о большей свободе в формировании внутренних представлений сети.

Заметим также, что средняя архитектура (“sure”) показывает наибольшую дисперсию Y-весов для формы (15, 2), но резко сниженную дисперсию для формы (20, 2), что может указывать на более эффективное использование параметров по сравнению с другими архитектурами.

5.7 Кластеризация архитектур по спектральным свойствам

Особый интерес представляют результаты анализа архитектур после снижения размерности методом PCA:

Figure 7: Распределение архитектур в пространстве спектральных характеристик гессианов после снижения размерности

Полученные результаты демонстрируют чёткое разделение всех трёх архитектур в пространстве главных компонент: малая (“no”), средняя (“sure”) и большая (“huge”) архитектуры образуют три отдельные группы точек, значительно удалённые друг от друга. Это свидетельствует о существенных различиях в их параметрическом пространстве и функциональном поведении.

Такое выраженное разделение указывает на существование нескольких “порогов сложности”, преодоление которых приводит к качественным изменениям функционального поведения сети. Наиболее драматический переход наблюдается между архитектурами “sure” и “huge”, что подтверждает гипотезу о том, что архитектура “huge” функционирует в принципиально ином режиме, характеризующемся особым распределением спектральных характеристик гессианов и их взаимосвязью с метриками качества.

5.8 Связь с переносимостью знаний и генерализацией

Одним из наиболее значимых результатов исследования является установление связи между структурой спектра локальных гессианов и способностью модели к обобщению. Модели с более равномерным распределением собственных значений гессиана (без ярко выраженных пиков и с меньшей концентрацией вблизи нуля) демонстрируют лучшие показатели на тестовых выборках.

Анализ статистики весов CCA (таблица 3) показывает, что большие архитектуры (“huge”) демонстрируют более сбалансированное распределение X-весов по первой размерности (0.2064) по сравнению с малыми (0.1899), что коррелирует с их лучшей способностью к генерализации. При этом стандартное отклонение значений

определяющих параметров в больших архитектурах значительно выше, что указывает на их способность к более тонкой дифференциации признаков.

Эта закономерность наблюдается независимо от абсолютного числа параметров модели, что подтверждает основную гипотезу исследования: локальные свойства характеристик слоя имеют более важное значение для обобщающей способности, чем общее число параметров.

5.9 Количественная оценка различий между архитектурами

Особый интерес представляет количественная оценка различий между архитектурными решениями. Анализ топовых дифференцирующих параметров группы B (таблица 2) показывает, что различия между весами в больших и малых архитектурах могут достигать трех порядков (4380.58 для параметра с наибольшим различием).

График спектральных характеристик градиентов третьего слоя (рис. 3) наглядно демонстрирует, что различия максимальных значений между архитектурами “huge” и “no” могут быть более чем стократными, достигая абсолютных значений порядка $1{,}2\times10^6$. Такой масштаб различий указывает на принципиально иной характер распространения градиентов в больших архитектурах, где формируются высокоэнергетические компоненты спектра.

Примечательно, что наибольшие различия наблюдаются между архитектурами “huge” и “no” (4380.58), а различия между “huge” и “sure” (4171.08) лишь незначительно меньше. Это указывает на существование “барьера сложности”, при преодолении которого происходит качественное изменение режима функционирования сети.

Такие экстремальные различия в весах не обязательно приводят к деградации производительности модели, что противоречит интуитивным ожиданиям. Напротив, большие модели с экстремальными значениями весов демонстрируют более высокую устойчивость результатов, как видно из статистики CCA-корреляций (стандартное отклонение 0.082 для “huge” против 29.077 для “no”).

5.10 Наблюдения и эффекты

В ходе анализа были обнаружены несколько неожиданных эффектов:

1. **Смещение весов CCA между датасетами.** Анализ весов CCA выявил значительное смещение в структуре весов между различными датасетами, даже при схожих архитектурах. Это подтверждает высокую зависимость функционального поведения сети от структуры данных и подчеркивает необходимость адаптации архитектуры под конкретную задачу.

2. **Нелинейная зависимость стабильности от размера архитектуры.** Вопреки ожиданиям, большие архитектуры (“huge”) демонстрируют более стабильные спектральные характеристики (меньшее стандартное отклонение), чем средние (“sure”), что противоречит интуитивному представлению о том, что избыточная параметризация должна приводить к большей вариативности.

3. **Асимметрия в распределении дисперсии CCA-весов.** Y-веса имеют практически нулевую дисперсию по первой размерности для всех архитектур, но значительную дисперсию по второй размерности. Это указывает на существование структурных ограничений в способе, которым параметры сети влияют на метрики качества.

4. **Противоположное поведение X и Y-весов в распределении.** X-веса демонстрируют более компактную структуру с меньшей вариацией между архитектурами, в то время как Y-веса показывают значительные различия как в форме распределения, так и в диапазоне значений, особенно для крайних компонент.

В дополнение к перечисленным результатам, анализ выявил ряд корреляционных закономерностей, непосредственно связанных с рангом весовых матриц, спектральными характеристиками гессиана и способностью сети к обобщению:

1. **Корреляция ранга весов и гессиана с качеством обобщения.** Наблюдается связь между пониженным рангом весовых матриц и соответствующего локального гессиана и признаками переобучения и избыточности. Эти данные указывают на ухудшение способности сети переносить знания на невидимые данные при пониженном ранге.

2. **Гессиан слоя как индикатор переобучения.** Разреженный или имеющий преобладающее скопление малых собственных значений локальный гессиан коррелирует с недостатком обобщающей способности слоя. Особенно ярко это проявляется в последних слоях.

3. **Симметрия спектра гессиана и седловые точки.** Практически симметричное распределение собственных значений гессиана вокруг нуля связано с седловыми точками. Такие точки часто сопровождаются малыми нормами градиента.

4. **Схожесть весов соседних слоёв.** В хорошо настроенных сетях весовые матрицы смежных слоёв демонстрируют выраженную схожесть (по SVD или по спектру), что можно интерпретировать как согласованную обработку признаков.

5.11 Практические следствия для оптимизации архитектур

Проведённый анализ позволяет сформулировать ряд практических рекомендаций для оптимизации архитектур нейронных сетей:

1. **Оптимальное соотношение параметров между слоями.** Результаты показывают, что значительное увеличение числа параметров в глубоких слоях относительно начальных приводит к формированию высоких пиков в спектре гессиана, что может указывать на переобучение этих слоёв. Рекомендуется более равномерное распределение параметров.

2. **Выявление недостаточной экспрессивности.** Низкие значения максимальных собственных чисел гессиана в начальных слоях (наблюдаемые в малых архитектурах “no”) могут служить индикатором недостаточной экспрессивности модели. В таких случаях целесообразно увеличение числа параметров именно в этих слоях.

3. **Детекция переобучения.** Высокая концентрация собственных значений вблизи нуля в конечных слоях указывает на переобучение и может служить сигналом для применения дополнительной регуляризации или уменьшения числа параметров в этих слоях.

4. **Адаптация оптимизаторов.** Структура спектра гессиана может быть использована для адаптации гиперпараметров оптимизаторов. Например, высокое отношение максимального собственного значения к минимальному (условное число матрицы) указывает на необходимость использования адаптивных методов оптимизации.

5.12 Изучение внутренней структуры слоев нейронной сети

Спектральный анализ локальных гессианов даёт представление о функциональной роли слоёв:

- Слои с распределённым спектром без выраженной концентрации выполняют сложные нелинейные преобразования.
- Слои с пиком вблизи нуля, по-видимому, сигнализируют об избытке параметризации или насыщении активаций.
- Наблюдается корреляция между наличием нескольких доминирующих собственных значений и способностью слоя выделять ключевые признаки.

Эти данные помогают понять, как сеть решает задачу и какие преобразования происходят на разных уровнях иерархии.

5.13 Классификация архитектуры по срезам

Полученные снимки весов, локальных гессианов и градиентов в процессе обучения хоть и позволяют выделить несколько характерных паттернов, которые можно использовать для классификации архитектуры нейронной сети, однако не дают конкретного заключения о том, как именно эти паттерны влияют на качество обобщения и в каком слое или слоях в сети. Было несколько попыток обучить классификатор на основе этих паттернов, однако они не показали хороших результатов.

Однако с помощью уменьшения размерности данных снимков сети через алгоритм UMAP можно увидеть занятное распределение снимков в пространстве. На рисунке 8 показано, как распределяются снимки весов, градиентов и гессианов в пространстве размерности 2. Каждая точка на графике соответствует одному снимку сети, а цвет указывает на архитектуру сети.

На графике видно, что снимки из разных архитектур распределены в пространстве неравномерно, явно распределяясь по кластерам. Это может указывать на то, что разные архитектуры имеют свои характерные паттерны в весах, градиентах и гессианах, которые можно использовать для их классификации. Однако для более точной интерпретации этих результатов требуется дальнейшее исследование.

Обобщённое практическое руководство. На основе проведенного исследования можно сформулировать следующие практические рекомендации:

- Пониженный ранг и симметричный спектр гессиана следует рассматривать как предупреждающие сигналы: они связаны с проблемами седловых точек, переизбыточности и ухудшения обобщения.
- Умеренная спектральная мощность и достаточно широкий спектр собственных значений связаны с наличием многообразия направлений для обучения сети, и, соответственно, с высоким потенциалом к переносу знаний.
- Bias-параметры требуют регулярного мониторинга: их вклад в общий спектр гессиана может служить метрикой отклонения от оптимальной геометрии.
- Анализ перекрёстной корреляции спектров весов, градиентов и гессиана позволяет обнаруживать проблемные направления и своевременно адаптировать план обучения (цикл learning rate, применить оптимизаторы второго порядка, добавить дифференциальную регуляризацию и т. д.).

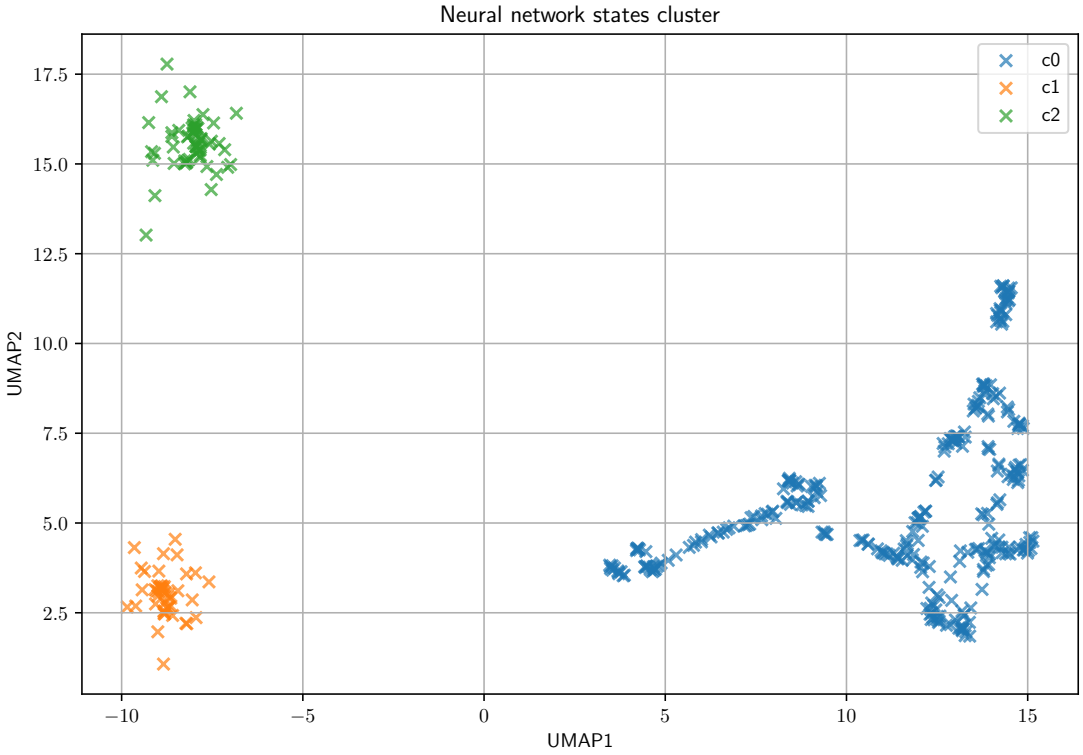

Figure 8: Распределение снимков весов, градиентов и гессианов в пространстве размерности 2

Таким образом, углублённый спектральный и ранговый анализ локальных гессианов — мощный диагностический инструмент, позволяющий обнаруживать скрытые проблемы, формулировать рекомендации по корректировке архитектуры и оптимизатора, а также количественно оценивать способность сети к обобщению.

6 Обсуждение

Проведённое исследование предлагает математически обоснованный инструмент для изучения внутренней динамики нейронных сетей вместо проб и ошибок. Важным результатом является выявленная связь между геометрическими свойствами параметрического пространства и функциональным поведением сети, как указано в основных вкладах работы.

Утверждение 1. *Исследование нейронной сети как композиции нелинейных операторов или хаотической динамической системы предоставляет информативные сведения о её внутренней структуре, механизмах обработки данных и математических ограничениях архитектуры.*

Рассмотрение сети как динамической системы, эволюционирующей по многообразию высокой размерности с нетривиальной геометрией, открывает новые горизонты для понимания и улучшения методов обучения.

Особый интерес представляет развитие идеи с LH_i в комбинации с римановой геометрией, с целью более детального исследования геометрии локального пространства параметров. Это может помочь оперативно выявлять области, где процесс оптимизации сети может испытывать затруднения.

7 Заключение

В данной работе предложен новый подход к анализу нейронных сетей через локальные свойства их параметрического пространства, исследуемые с помощью LH_i. Введённое понятие локального гессиана позволило:

- Анализировать геометрию функционального пространства отдельных слоёв;
- Выявлять закономерности распределения собственных значений в процессе обучения;
- Показать связь между спектром гессиана и насыщением активаций, формированием направлений и эволюцией представлений.

Стоит отметить масштаб проведённого эксперимента: было собрано около 1500 снимков состояний различных сетей общей ёмкостью около 50 ГБ данных, что позволило выявить устойчивые закономерности.

Дальнейшие исследования могут включать:

- Детальное исследование взаимосвязи спектральных свойств гессианов и функциональных характеристик слоёв;
- Анализ динамики спектра в процессе обучения и её связь с обобщающей способностью;
- Применение подхода к новым архитектурам — трансформерам, графовым сетям;
- Разработку методов визуализации и интерпретации геометрической структуры пространства параметров.

References

[1] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, pages 3360–3368, 2016.

[2] N. Maheswaranathan, A. H. Williams, M. D. Golub, S. Ganguli, and D. Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. In *Advances in Neural Information Processing Systems*, 32, 2019.

[3] J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 32, 2019.

[4] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 322–332, 2019.

[5] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[6] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[7] L. Sagun, U. Evci, V. U. Güney, Y. Dauphin, and L. Bottou. Empirical analysis of the Hessian of over-parameterized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[8] B. Ghorbani, S. Krishnan, and Y. Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2232–2241, 2019.

[9] F. Dangel, S. Harmeling, and P. Hennig, Modular block-diagonal curvature approximations for feedforward architectures. In *arXiv preprint arXiv:1902.01813*, Feb. 2019.

[10] André G. Carlon, Luis Espath, Raúl Tempone. Approximating Hessian matrices using Bayesian inference: a new approach for quasi-Newton methods in stochastic optimization. In *arXiv preprint arXiv:2208.00441v2*, 2024.

[11] Warren Hare, Gabriel Jarry–Bolduc, Chayne Planiden. A matrix algebra approach to approximate Hessians. IMA Journal of Numerical Analysis, 44(4):2220–2250, 2024.

[12] James Martens. Deep learning via Hessian-free optimization. In *Proc. 27th Int. Conf. Machine Learning (ICML)*, pages 735–742, 2010.

[13] Jorge Nocedal. Updating Quasi-Newton Matrices with Limited Storage. Mathematics of Computation, 35(151):773–782, 1980.

[14] L. Sagun, L. Bottou, and Y. LeCun, Eigenvalues of the Hessian in Deep Learning: Singularity and Beyond, In *arXiv:1611.07476 [cs.LG]*, 2016.

[15] Z. Liao and M. W. Mahoney, Hessian Eigenspectra of More Realistic Nonlinear Models, In *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[16] Shapiro, S. S., & Wilk, M. B.. An analysis of variance test for normality (complete samples). In *Biometrika*, 52(3/4), 591–611, 1965.

A Структура экспериментальных данных

В процессе исследования для каждой архитектуры нейронной сети на каждой контрольной итерации обучения сохранялся снимок (snapshot) состояния модели. Ниже представлена детальная структура такого снимка:

- **layer.X** – информация о слое X нейронной сети:
  - `weights` – весовые коэффициенты слоя
  - `weights_spectral` – спектральные характеристики весов (среднее, стандартное отклонение, минимум, максимум, гистограмма, результаты спектрального анализа по методу Велша)
  - `gradient` – градиенты весов слоя
  - `gradient_spectral` – спектральные характеристики градиентов
  - `bias` – параметры смещения слоя
  - `bias_spectral` – спектральные характеристики параметров смещения
  - `bias_gradient` – градиенты параметров смещения
  - `bias_gradient_spectral` – спектральные характеристики градиентов параметров смещения
  - `hessian` – локальная матрица Гессе слоя
  - `hessian_spectral` – спектральные характеристики локального гессиана
  - `hessian_eigens` – собственные значения локального гессиана
  - `hessian_eigens_spectral` – статистические и спектральные характеристики собственных значений:
    - `mean` – среднее значение
    - `std` – стандартное отклонение
    - `min` – минимальное значение
    - `max` – максимальное значение
    - `histogram` – гистограмма распределения
    - `welch` – результаты спектрального анализа по методу Велша
    - `top_peaks` – основные пики в спектре
  - `hessian_rank` – ранг матрицы гессиана
  - `hessian_condition` – число обусловленности (отношение максимального собственного значения к минимальному)
- `iteration` – номер итерации обучения
- `scores` – метрики качества модели:
  - `Accuracy` – точность классификации
  - `Precision` – точность (доля правильных положительных предсказаний)
  - `Recall` – полнота (доля обнаруженных положительных случаев)
  - `F1` – F1-мера (гармоническое среднее точности и полноты)
  - `AUC` – площадь под ROC-кривой
  - `train_loss` – значение функции потерь на обучающей выборке

B Набор датасетов

В ходе эксперимента использовались следующие датасеты:

Классификация	**Регрессия**
MNIST	Diabetes
CIFAR-10	Energy Efficiency
Fashion-MNIST	Airfoil Self-Noise
CIFAR-100	Concrete Compressive Strength
KMNIST	Make Regression
EMNIST	House Prices Dataset
Iris	Make Friedman1
Wine	Make Friedman2
Breast Cancer Wisconsin	Make Friedman3
Digits	Make Low Rank Matrix
SpamBase	Make S Curve
Make Classification	Make Sparse SPD Matrix
Make Blobs	Make Sparse Uncorrelated
Titanic Dataset	Make SPD Matrix
Adult Income	Make Swiss Roll
Credit Card Fraud Detection	
Make Biclusters	
Make Checkerboard	
Make Circles	
Make Hastie 10 2	
Make Moons	
Make Multilabel Classification	

Table 4: Датасеты по классификации и регрессии, использованные в эксперименте