Исследование локальных свойств нейронных сетей с помощью анализа гессианов

Максим Большим

May 24, 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 изучать 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 147 экспериментов на 37 датасетах показал существенные закономерности в структуре матриц гессиана на различных этапах обучения. Разработанный математический аппарат позволяет глубже понимать внутренние процессы в глубоких нейронных сетях во время их обучения. а так же представляет из себя отличную основу для поиска слабых мест в архитектуре

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 систематически 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 предоставить ценную информацию о внутренних процессах в модели без необходимости полной переобучения [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 спектральные свойства локального гессиана, такие как распределение собственных значений и их структура [8], тесно связаны с функциональными свойствами соответствующих слоёв нейронной сети.

Основные вклады настоящей работы:

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

- Разработка эффективного алгоритма для вычисления локальных гессианов с линейной сложностью по числу параметров
- Детальный анализ спектральных свойств локальных гессианов в процессе обучения нейронных сетей
- Исследование геометрической интерпретации пространства параметров 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 нам анализировать каждый функциональный блок 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 изучение 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 $\varphi(y) = \sum_{j=1}^{q_i} y_j$.

Локальная 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 $H_i \in \mathbb{R}^{p_i \times p_i}$ для блока C_i называется матрица вторых производных скалярной функции S_i по параметрам θ_i:*

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

Локальный гессиан представляет собой локальную кривизну функции отклика слоя относительно изменения его параметров. 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

Конструкция локального гессиана имеет глубокое геометрическое значение. В дифференциальной геометрии гессиан функции в точке определяет квадратичную форму, которая аппроксимирует локальную кривизну поверхности уровня этой функции. В контексте нейронных сетей локальный гессиан характеризует кривизну функционального отклика слоя в пространстве его параметров [9].

3.2 Эффективное вычисление локальных гессианов

Анализ нейронных сетей с использованием матриц Гессе 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 гессиана [10, 11, 12, 13]. В данном разделе предлагается методика работы с гессианами, позволяющая обойти указанные ограничения.

Для эффективного вычисления локальных гессианов 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 Гессе 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 локальных гессианов

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 матриц Гессе

Require: Нейронная сеть $\mathcal{F}$, входные данные $x \in \mathbb{R}^d$, функция агрегации φ
Ensure: Набор локальных матриц Гессе $\{H_1, H_2, \ldots, H_n\}$

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 $H_i \in \mathbb{R}^{p_i \times p_i}$ нулевой матрицей
    for $j = 1$ до p_i **do**
        if $g_{i,j}$ зависит от θ_i **then**
            Вычислить $H_i[j, :] = \nabla_{\theta_i} g_{i,j}$
        end if
    end for
end for
return $\{H_1, H_2, \ldots, H_n\}$

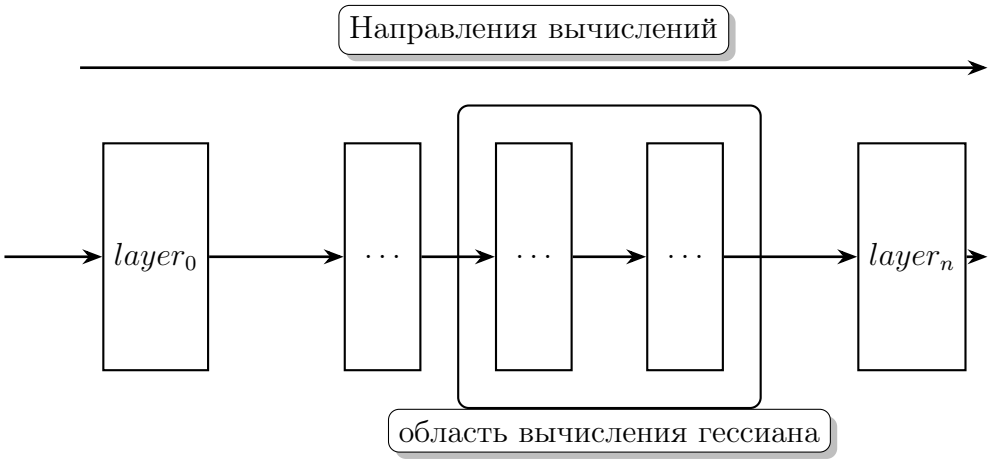

Figure 1: Визуализация вычисления локального гессиана

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 гессиана для выявления структуры пространств параметров

Спектральный анализ локальных гессианов предоставляет важный инструмент для понимания геометрии пространства параметров. Распределение и структура собственных значений гессиана отражают ключевые свойства локальной кривизны функции, что особенно важно для анализа степени обусловленности задачи оптимизации [14, 15].

Для каждого локального гессиана H_i можно вычислить:

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

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

- **Следы гессиана:** $\mathrm{tr}(H_i) = \sum_{j=1}^{p_i} \lambda_{i,j}$ — сумма собственных значений, отражающая общую кривизну
- **Определитель гессиана:** $\det(H_i) = \prod_{j=1}^{p_i} \lambda_{i,j}$ — произведение собственных значений
- **Условное число:** $\kappa(H_i) = \frac{\lambda_{i,\max}}{\lambda_{i,\min}}$ — отношение максимального собственного значения к минимальному (для ненулевых значений)
- **Эффективный ранг:** $r_{\text{eff}}(H_i) = \frac{\sum_j |\lambda_{i,j}|}{\max_j |\lambda_{i,j}|}$ — мера эффективной размерности пространства параметров

Особый интерес представляет анализ распределения собственных значений по слоям сети и его изменения в процессе обучения, что позволяет отслеживать эволюцию геометрии пространства параметров.

Распределение собственных значений гессиана дает информацию о геометрии функционального пространства слоя. В частности, концентрация собственных значений вблизи нуля может указывать 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 локального гессиана, 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 локальный гессиан 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.

3.6 Спектральный анализ локальных гессианов

4 Методология исследования

4.1 Дизайн экспериментов

Для изучения 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 варьировались следующие параметры:

- Количество слоёв и нейронов в сетях
- Способы инициализации весов
- Методы оптимизации (Adam, SGD, RMSProp)
- Функции активации (ReLU, sigmoid)

Общее число параметров в исследуемых моделях варьировалось от 13 до 9 миллионов, что позволило изучить сети с разным уровнем параметризации:

1. Сети с ограниченным числом параметров
2. Сети с умеренным количеством параметров
3. Сети с большим числом параметров

Такой широкий охват архитектур и задач позволяет выявить общие закономерности в структуре локальных гессианов независимо от конкретной задачи или архитектуры.

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 Гессе 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 локальных гессианов, 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 Методы анализа данных

Для анализа полученных данных использовались следующие методы:

1. Динамика изменений статистических показателей градиентов, весов и локальных гессианов в процессе обучения
2. Канонический корреляционный анализ (CCA) спектральных характеристик гессианов и метрик качества моделей
3. Визуализация с помощью тепловых карт (heatmap), диаграмм рассеяния (pairplot) и тестов нормальности (Shapiro-Wilk)

4.4 Обработка экспериментальных данных

Для обработки собранных данных был применён многоэтапный анализ:

1. **Анализ динамики параметров:** отслеживание изменения всех числовых показателей по мере обучения модели, включая спектральные характеристики гессиана и условное число.
2. **Корреляционный анализ:** расчёт коэффициентов корреляции Пирсона между различными параметрами модели и показателями качества, построение тепловых карт корреляций.
3. **Спектральный анализ:** исследование распределения собственных значений весов и гессианов, включая анализ среднего, стандартного отклонения, минимумов и максимумов.
4. **Канонический корреляционный анализ (CCA):** изучение нелинейных взаимосвязей между группами метрик качества и характеристиками внутренних параметров сети.
5. **Статистические тесты:** проверка распределений характеристик с помощью теста Шапиро-Уилка [16] для выявления отклонений от нормальности.

Данный комплексный подход позволил выявить не только непосредственные корреляции между отдельными параметрами, но и более сложные взаимосвязи между группами параметров и метриками качества, что особенно важно для понимания нелинейной природы нейронных сетей.

5 Результаты исследования

Результаты первичного исследования

Проведение экспериментов проводилось 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 9 часов, включая время на обучение моделей и вычисление локальных гессианов. Это было связано с тем, что первичная реализация 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.

Каждый результат будет подробно представлен ниже в соответствующих разделах, однако уже сейчас можно выделить несколько ключевых наблюдений:

- Спектральный анализ локальных гессианов предоставляет ценную информацию о внутренней структуре и функционировании нейронных сетей. Хотя точная интерпретация всех наблюдаемых феноменов остаётся открытым вопросом, очевидно, что сети с различной архитектурой демонстрируют характерные паттерны в спектральных свойствах гессианов.

- Необходимо провести более детальный анализ динамики эволюции гессианов 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 наличий нескольких проблем архитектуры сразу, которые могут быть не связаны между собой. Так же этими инструментами трудно анализировать слабые стороны сети, если гиперпараметры обучения подобраны неудачно.

- Дополнительно рекомендуется исследовать кривизну градиента и гессиана в контексте римановой геометрии, а также рассмотреть возможность моделирования градиентного потока на многообразии параметров в предельном случае перенасыщения функций активации.

5.1 Спектральные свойства локальных гессианов

Одним из ключевых результатов исследования является обнаружение закономерностей в спектральном разложении локальных гессианов. Анализ собственных значений позволил выявить характерные паттерны, соответствующие различным режимам работы нейронной сети.

Утверждение 1. *Распределение собственных значений локальных гессианов имеет характерную форму, зависящую от состояния соответствующего слоя нейронной сети и стадии обучения.*

В частности, было обнаружено, что:

- В начальных стадиях обучения собственные значения распределены более равномерно;

- По мере обучения происходит концентрация собственных значений, что указывает на формирование определённых направлений в параметрическом пространстве;

- В слоях с насыщенными активациями наблюдается высокая концентрация собственных значений вблизи нуля.

Эти наблюдения согласуются с теоретическими представлениями о формировании низкоразмерны многообразий в процессе обучения нейронных сетей.

5.2 Взаимосвязь структуры гессиана с насыщением функций активации

Эксперименты показали, что спектральные характеристики локальных гессианов тесно связаны с состоянием функций активации нейронов.

Утверждение 2. *При насыщении функций активации, когда входные значения попадают в области с малой производной, наблюдается характерное изменение в спектральной структуре локального гессиана соответствующего слоя.*

Это можно объяснить геометрией пространства параметров: в состоянии насыщения малые изменения весов практически не влияют на выход сети, что приводит к «плоским» направлениям в параметрическом пространстве. На основе экспериментальных данных были выявлены следующие закономерности:

1. В слоях с преобладанием насыщенных нейронов спектральное распределение собственных значений имеет острый пик вблизи нуля и быстрый спад;
2. При оптимальных режимах работы функций активации распределение более равномерное, с умеренной концентрацией в среднем диапазоне;
3. Чрезмерная активация нейронов (слишком сильные сигналы) также ведёт к насыщению с обратным эффектом, что отражается в спектре гессиана.

Практические выводы:

- Анализ спектра локальных гессианов позволяет диагностировать насыщение активаций без прямой инспекции каждого нейрона.
- Методы нормализации (batch normalization, layer normalization) удерживают входы функций активации в оптимальном диапазоне, что заметно в спектральных характеристиках гессиана.
- При проектировании архитектуры стоит учитывать склонность глубоких сетей с традиционными активациями к насыщению в глубоких слоях, что выявляется по характерному изменению спектра гессианов.

5.3 Анализ динамики обучения нейронных сетей

На основе спектральных свойств локальных гессианов можно выделить четыре фазы обучения:

1. **Начальная фаза**: равномерное распределение собственных значений, первичная адаптация параметров.
2. **Фаза быстрого обучения**: дифференциация значений — формирование «важных» и «неважных» направлений.
3. **Фаза тонкой настройки**: стабилизация спектра, медленная адаптация вдоль ключевых направлений.
4. **Фаза насыщения**: спектр практически не меняется — достижение локального оптимума или плато.

Понимание фаз позволяет разрабатывать адаптивные стратегии обучения, подстраивающиеся под текущую стадию оптимизации.

5.4 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 могут сигнализировать об избытке параметризации или насыщении активаций.
- Слои с несколькими доминирующими собственными значениями служат экстракторами ключевых признаков.

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

5.5 Дополнительные эмпирические наблюдения о ранге, спектре и обобщающей способности

В дополнение к перечисленным выше результатам, анализ выявил ряд корреляционных закономерностей, непосредственно связанных с рангом весовых матриц, спектральными характеристиками гессиана и способностью сети к обобщению. Кратко сформулируем ключевые выводы, а затем подробно их обсудим.

1. **Корреляция ранга весов и гессиана с качеством обобщения.** Чем ниже ранг как весовых матриц, так и соответствующего локального гессиана, тем сильнее признаки переобучения и избыточности. Наблюдаемая зависимость указывает: при пониженном ранге существенно ухудшается способность сети переносить знания на невидимые данные. В таких случаях целесообразно либо изменить метод оптимизации (усилить регуляризацию, применить dropout, L_1/L_2 – штрафы), либо переосмыслить архитектуру (уменьшить число параметров, заменить полносвязные слои сверточными и т. д.).
2. **Гессиан слоя как индикатор переобучения.** Разреженный или имеющий преобладающее скопление малых собственных значений локальный гессиан однозначно сигнализирует о недостатке обобщающей способности слоя. Особенно ярко это проявляется в последних слоях, где переносимая информация концентрируется в маломерном подпространстве.
3. **Перекрёстная корреляция собственных значений весов и градиентов.** Высокая корреляция между спектром весов и текущими градиентами указывает на «жёсткое» направление в пространстве параметров, где оптимизатор может застревать. Обнаружение этого признака служит поводом к адаптивному изменению шага обучения (learning rate schedule) или переходу на методы второго порядка.
4. **Малое среднеквадратичное отклонение спектра гессиана.** Если дисперсия собственных значений гессиана в слое мала, то потерь в обобщении практически неизбежны: параметры «слепо» следуют одному-двум доминирующим направлениям, игнорируя информационно-значимые, но слабые компоненты.
5. **Важность смещений (*bias*).** Нельзя недооценивать роль смещений: статистически значимая часть спектра гессиана формируется именно за счёт bias-параметров. Пренебрежение их анализом приводит к неполному пониманию геометрии потерь.
6. **Симметрия спектра гессиана и седловые точки.** Практически симметричное распределение собственных значений гессиана вокруг нуля — верный маркер седловой точки. Такие точки часто сопровождаются малыми нормами градиента и требуют специальных стратегий выхода (случайное возмущение, увеличение learning rate, добавление шума).

7. **Схожесть весов соседних слоёв.** В хорошо настроенных сетях весовые матрицы смежных слоёв демонстрируют выраженную схожесть (по SVD или по спектру), что интерпретируется как согласованная обработка признаков. При нарушении этой взаимосогласованности наблюдается более высокая мера разнообразия между весами слоёв, что коррелирует с деградацией качества.

5.6 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 2 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.

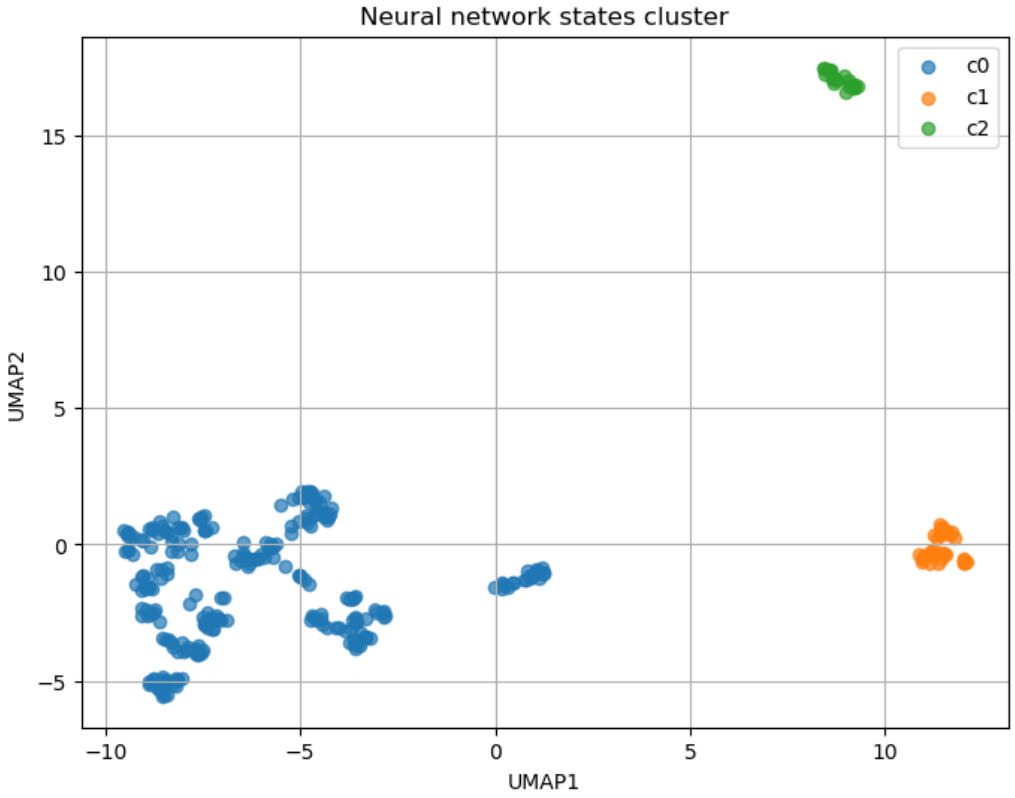

Figure 2: Распределение снимков весов, градиентов и гессианов в пространстве размерности 2

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.

- Пониженный ранг и симметричный спектр гессиана должны рассматриваться как предупреждающие сигналы: вероятны седловые точки, переизбыточность и ухудшение обобщения.
- Умеренная спектральная мощность и достаточно широкий спектр собственных значений указывают на «здоровое» многообразие направлений, которыми сеть учится, и, следовательно, на высокую потенцию к переносу знаний.
- Bias-параметры требуют регулярного мониторинга: процент их вклада в общий спектр гессиана служит метрикой отклонения от оптимальной геометрии.
- Анализ перекрёстной корреляции спектров весов, градиентов и гессиана позволяет своевременно обнаружить жёсткие направления и адаптировать план обучения (цикл learning rate, применить оптимизаторы второго порядка, добавить дифференциальную регуляризацию и т. д.).

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 связь между геометрическими свойствами параметрического пространства и функциональным поведением сети. Локальная кривизна, описываемая гессианом, содержит информацию о режиме работы слоя и его вкладе в общую функциональность.

Утверждение 3. *Исследование нейронной сети как композиции нелинейных операторов или хаотической динамической системы предоставляет наиболее 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.

В данном случае интересует именно развитие идеи с локальным гессианом в комбинации с римановой геометрией, исследуя более детально геометрию локального пространства параметров. Это может помочь относительно быстро находить места, где сеть может застревать.

7 Заключение

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

- Изучать 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 изучение 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 параметрическо пространства.

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.