ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2023, том 514, № 2, с. 187–195

УДК 004.8

ОДНОМЕРНЫЕ ТОПОЛОГИЧЕСКИЕ ИНВАРИАНТЫ ДЛЯ ОЦЕНКИ НЕВЫПУКЛОСТИ ФУНКЦИИ ПОТЕРЬ

© 2023 г. Д. С. Воронкова[1,*], С. А. Баранников[1,3], Е. В. Бурнаев[1,2]

Представлено академиком РАН А.Л. Семеновым
Поступило 05.09.2023 г.
После доработки 15.09.2023 г.
Принято к публикации 18.10.2023 г.

В данной работе рассматривается применение топологического анализа данных к исследованию геометрических свойств ландшафта функции потерь. Используя топологию и теорию Морса, мы строим одномерные топологические инварианты для оценки близости функции потерь к выпуклой функции с точностью до произвольной перепараметризации. Предложенный подход использует оптимизацию двумерных симплексов в пространстве параметров нейросети и позволяет проводить как качественную, так и количественную оценку ландшафта функции потерь для получения новых представлений о поведении и оптимизации нейронных сетей. Мы предоставляем геометрическую интерпретацию топологических инвариантов и описываем алгоритм их вычисления. Мы ожидаем, что предложенный подход может дополнить существующие инструменты для анализа ландшафта функции потерь и пролить свет на нерешенные вопросы в области глубокого обучения.

DOI: 10.31857/S2686954323601999, **EDN:** ZEYNUX

1. ВВЕДЕНИЕ

Исследование свойств ландшафта функции потерь стало быстро развивающимся направлением исследований, направленным на улучшение понимания поведения нейронных сетей. Среди многочисленных направлений выделяются недавние работы, которые исследовали влияние дизайна архитектуры на структуру поверхности функции потерь [1, 2], взаимосвязь размера батча во время обучения и узость наблюдаемого локального минимума и его способность к обобщению [3, 4]. Однако эти работы фокусируются на локальном поведении функции потерь в окрестности локального минимума. Другое направление исследований изучает свойство связности мод [5–8], т.е. возможность соединить два локальных минимума с помощью пути или подпространства с низкой ошибкой. Наконец, существует направление исследований, которое анализирует как локальную, так и глобальную структуру ландшафта функции потерь для получения представления о процессе оптимизации и характеристиках успешных решений [9, 10].

Следуя этой тенденции, наша работа исследует глобальную топологическую структуру поверхности функции потерь, измеряя одномерные топологические инварианты. Мы предполагаем, что топологическая стабильность ландшафта функции потерь улучшает способность нейронной сети к обобщению. Предложенный подход не зависит от вида параметризации весов нейронной сети, что является препятствием для некоторых подходов [4]. С помощью топологии можно характеризовать глобальное отклонение функции потерь от выпуклой функции с точностью до произвольной непрерывной перепараметризации. Мы эмпирически наблюдаем, что это свойство коррелирует с обобщаемостью модели нейронной сети.

Вклады данной работы следующие:

• Мы формулируем теоретические основы для изучения геометрии ландшафта функции потерь, основанные на топологии и теории Морса. Как следует из теоретических исследований, топологическая структура данной функции может быть полностью описана через набор ее топологических инвариантов, бар-кодов.

• Мы предоставляем геометрическую интерпретацию 1-мерных топологических инвариантов и алгоритм их вычисления. Мы исследуем практическое использование этих инвариантов и

[1]*Сколковский институт науки и технологий, Москва, Россия*
[2]*Научно-исследовательский институт искусственного интеллекта AIRI, Москва, Россия*
[3]*CNRS, IMJ, Paris Cité University, Франция*
*****E-mail: Darya.Voronkova@skoltech.ru*

их применение к задачам анализа поверхности функции потерь.

• Мы эмпирически показываем значимость предложенного подхода с помощью ряда экспериментов, проведенных на глубоких нейронных сетях.

• Мы публикуем код для воспроизведения результатов этой статьи по ссылке https://github.com/VoronkovaDasha/1-dim-Topological-Invariants.

2. ОБЗОР ЛИТЕРАТУРЫ

Функция потерь и ее ландшафт являются основным объектом в дизайне и оптимизации глубоких нейронных сетей, и существует множество работ, исследующих различные свойства функции потерь. В этом разделе мы рассматриваем несколько направлений, которые имеют наиболее тесную связь с текущей работой.

Наглядный подход к получению представления об обучении глубоких нейронных сетей заключается в визуализации ландшафта функции потерь. Одномерные визуализации часто используются для изучения траектории процесса оптимизации [12], широты найденных локальных минимумов [3], существования барьеров с высокими значениями функции потерь между локальными минимумами [12, 13], направлений асимметричного роста функции потерь в окрестности локальных минимумов [14]. С использованием двумерных визуализаций, работа [1] показывает, что наличие резидуальных соединений предотвращает переход ландшафта функции потерь к хаотическому поведению по мере увеличения глубины. Также двумерные визуализации поверхности функции потерь позвляют показать [2], что визуальные трансформеры обычно имеют более широкие локальные оптимумы, чем резидуальные сети. Несмотря на популярность различных техник по визуализации ландшафта функции потерь, они используют значительное уменьшение размерности и требуют тщательной интерпретации. В частности, выпуклость на двумерной визуализации поверхности функции потерь не обязательно подразумевает выпуклость истинной функции потерь в оригинальной размерности [1]. Кроме того, хотя техники визуализации подходят для качественного анализа, они не предоставляют числовую характеристику общей выпуклости поверхности функции потерь.

В общем случае линейная интерполяция выявляет наличие барьеров с высокими значениями функции потерь между двумя локальными минимумами, полученными с помощью градиентной оптимизации [5, 6, 12]. Однако [5, 6] эмпирически наблюдают существование нелинейных обучаемых путей с низкими значениями функции потерь между двумя локальными минимумами. Это свойство часто называется “связность мод”. [19] дает теоретическое объяснение этого явления, предполагая наличие дропаута и стабильность шума глубокой нейронной сети. Как показано в [15], свойство связности мод сохраняется даже тогда, когда два локальных минимума получены с помощью различных процедур обучения. Более того, [8] экспериментально подтверждает существование подпространств с низкими значениями функции потерь, соединяющих набор локальных минимумов. Работа [6] предлагает новые методы для ансамблирования глубоких нейронных сетей на основе наблюдаемого свойства связности мод. Другие работы [16, 17] исследуют структуру поверхности функции потерь через поиск различных паттернов в разных условиях. Несмотря на то что эти подходы являются мощными инструментами для исследования поверхности функции потерь, им не хватает наличия численной характеристики, количественно оценивающей сложность поверхности функции потерь глубокой нейронной сети.

Продолжая идею связности мод, [9, 10] утверждают, что локальные минимумы, полученные при помощи стохастического градиентного спуска, вероятно, не будут иметь барьеров при линейной интерполяции с точностью до перестановки весов глубокой нейронной сети. Это подразумевает, что ландшафт функции потерь по существу имеет одну моду по модулю перестановок.

Другое направление исследований связывает обобщаемость с узостью локальных минимумов. Общие подходы к измерению узости [1, 3] имеют определенные недостатки [4] и не могут полностью объяснить обобщаемость. Работа [14] показывает, что в окрестности локального минимума существует множество асимметричных направлений с разной скоростью роста функции потерь. Тем не менее эти подходы измеряют узость каждого локального минимума независимо [11], т.е. они не предоставляют характеристику для всего ландшафта функции потерь.

3. ГЕОМЕТРИЧЕСКОЕ ОПИСАНИЕ И АЛГОРИТМ ВЫЧИСЛЕНИЯ

В этом разделе мы приводим геометрическое описание исследуемых топологических инвариантов и алгоритм их вычисления в контексте глубоких нейронных сетей.

Рассмотрим обучение произвольной нейронной сети f с параметрами $\theta \in \Theta \subset \mathbb{R}^d$ на наборе данных X с минимизацией эмпирической функции потерь L. Хотя функция потерь L зависит как от параметров θ, так и от данных X, в этой работе мы предполагаем, что набор данных фиксирован, и рассматриваем только зависимость функции потерь от параметров: $L = L(\theta)$. Рассмотрим на-

бор из трех предельных точек $C_0 = \{\theta_{i_1}, \theta_{i_2}, \theta_{i_3} | \theta_{i_k} \in \Theta, \nabla L(\theta_k) = 0\}$, к которым сходится процесс градиентной оптимизации при различных инициализациях.

Для каждой пары точек $\theta_{i_j}, \theta_{i_k}$ определим значение

$$h_{j,k} = \min_{\substack{\gamma:[0,1]\to\Theta, \\ \gamma(0)=\theta_{i_j}, \gamma(1)=\theta_{i_k}}} \max_t L(\gamma(t)). \quad (1)$$

Неформально, $h_{j,k}$ определяет барьер с наименьшим значением функции потерь среди всех путей, соединяющих две критические точки. Для набора C_0 определим $h^1_{C_0} = \max_{j,k} h_{j,k}$.

Пусть σ обозначает стандартный двумерный симплекс, заданный уравнением $t_1 + t_2 + t_3 = 1$, $t_i \geq 0$, и e_i, $i = 1,2,3$, обозначают три вершины стандартного симплекса σ.

Для набора C_0 определим значение

$$h^2_{C_0} = \min_{\substack{\zeta:\sigma\to\Theta \\ \zeta(e_1)=\theta_{i_1}, \zeta(e_2)=\theta_{i_2}, \zeta(e_3)=\theta_{i_3}}} \max L(\zeta(t)). \quad (2)$$

Таким образом, $h^2_{C_0}$ обозначает наименьший пик L на двумерных симплексах, граница которых состоит из цикла трех наименьших кривых потерь, соединяющих точки θ_{i_1}, θ_{i_2}, θ_{i_3} в циклическом порядке. С топологической точки зрения, точки C_0 составляют топологический признак в многомасштабной подуровневой топологии функции потерь L. Этот признак, цикл, появляется на уровне $h^1_{C_0}$ и исчезает на уровне $h^2_{C_0}$. Следовательно, для набора C_0 сегмент $s_{C_0} = [h^1_{C_0}; h^2_{C_0}]$ записывает времена “рождения” и “смерти” этого топологического признака. Длина сегмента (“продолжительность жизни”) определяет важность признака: большая продолжительность жизни указывает на важные признаки, в то время как признаки с маленькой продолжительностью жизни похожи на шум. Когда эти сегменты агрегируются по всем тройкам минимальных критических точек функции L, получается оценка для одномерного баркода функции L:

$$Barcode^1(L) = \sqcup_{C_0} [h^1_{C_0}, h^2_{C_0}]. \quad (3)$$

Точное определение $Barcode^1(L)$ включает в себя фильтрованный симплициальный комплекс и представлено в разделе 4. Индекс 1 указывает на то, что бар-код вычисляется для топологических признаков, полученных как отображение из двумерного стандартного симплекса, граница которого является одномерным циклом. Аналогично можно определить и вычислить $Barcode^0(L)$, вычисление которого включает только пути между минимальными критическими точками, как это было показано в [18]. Однако данная работа направлена на изучение бар-кодов с индексом 1.

Бар-коды можно использовать для оценки топологического сходства двух функций. В этой работе мы изучаем отклонение функции потерь от выпуклой функции, так как градиентная оптимизация для выпуклой функции становится простой.

Бар-код индекса 1 выпуклой функции L_{ideal} представляет собой пустое множество. Бар-коды инвариантны относительно произвольной непрерывной перепараметризации.

Следовательно, если L может быть представлена как выпуклая функция после перепараметризации, то бар-коды L в каждом измерении, большем нуля, должны быть пустыми. Если L близка к выпуклой функции с точностью до перепараметризации, то эти бар-коды содержат только сегменты малой длины, т.е. соответствующие топологические признаки имеют малую продолжительность жизни.

Персистентные диаграммы, которые являются эквивалентными представлениями бар-кодов через множества точек (рождение, смерть), можно использовать для качественного анализа бар-кодов: чем ближе точки (рождение, смерть) к диагонали, тем меньше их продолжительность жизни, тем менее важны соответствующие признаки. Количественное сравнение двух функций возможно через расстояние между их бар-кодами. Одним из стандартных способов выбрать расстояние в пространстве бар-кодов является bottleneck-расстояние, также известное как $\mathbb{W}_\infty$, расстояние Вассерштейна-∞:

$$\mathbb{W}_\infty(D, D_{ideal}) = \inf_{\pi\in\Gamma(D, D_{ideal})} \sup_{a\in D\cup\Delta} \|a - \pi(a)\|, \quad (4)$$

где D – это бар-код $Barcode^1(L)$ функции потерь L, представленный как облако точек, D_{ideal} – это бар-код $Barcode^1(L_{ideal})$ идеальной выпуклой функции потерь L_{ideal}, Δ обозначает диагональ в $\mathbb{R}^2$, $\Gamma(D, D_{ideal})$ определяет множество биекций между $D \cup \Delta$ и $D_{ideal} \cup \Delta$. Для функций с похожей топологией расстояние между бар-кодами мало. Можно показать, что если $D = \{(b_i, d_i)\}_{i\geq 1}$ – это множество точек (рождение, смерть), то $\mathbb{W}_\infty(D, D_{ideal}) = \max_i \frac{d_i - b_i}{\sqrt{2}}$.

Алгоритм вычисления. На основе геометрического описания мы предлагаем следующую процедуру для вычисления бар-кодов индекса 1. Следуя [6], мы обучаем кривые $\phi : [0,1] \to \Theta$ между парой минимумов в виде полигональной цепи:

$$\phi_\theta(t) = \begin{cases} 2(t\theta + (0.5 - t)\theta_1), & 0 \le t \le 0.5, \\ 2((t - 0.5)\theta_2 + (1 - t)\theta), & 0.5 \le t \le 1, \end{cases} \quad (5)$$

где θ_1, θ_2 – это (фиксированные) конечные точки кривой, θ – это (обучаемые) параметры параметризации кривой. Мы определяем аналогичную параметризацию для отображения $\psi : \sigma \to \Theta$ из стандартного двумерного симплекса.

$$\psi_{\bar\theta}(t) = \begin{cases} (t_3 - t_2)\theta_1 + 3t_1\theta_{123} + (2t_2 - 2t_1)\theta_{12}, & t_1 \le t_2 \le t_3, \\ (t_2 - t_3)\theta_2 + 3t_1\theta_{123} + (2t_3 - 2t_1)\theta_{12}, & t_1 \le t_3 \le t_2, \\ (t_1 - t_2)\theta_3 + 3t_3\theta_{123} + (2t_2 - 2t_3)\theta_{23}, & t_3 \le t_2 \le t_1, \\ (t_3 - t_1)\theta_1 + 3t_2\theta_{123} + (2t_1 - 2t_2)\theta_{13}, & t_2 \le t_1 \le t_3, \\ (t_2 - t_1)\theta_2 + 3t_3\theta_{123} + (2t_1 - 2t_3)\theta_{23}, & t_3 \le t_1 \le t_2, \\ (t_1 - t_3)\theta_3 + 3t_2\theta_{123} + (2t_3 - 2t_2)\theta_{13}, & t_2 \le t_3 \le t_1, \end{cases} \quad (6)$$

где $\theta_1, \theta_2, \theta_3$ – это (фиксированные) вершины треугольника, $\theta_{12}, \theta_{13}, \theta_{23}, \theta_{123}$ – обучаемые параметры параметризации треугольника. В экспериментах мы проводим обучение в три этапа. Сначала мы получаем $\theta_1, \theta_2, \theta_3$ стандартным обучением глубокой нейронной сети. На втором шаге для каждой пары минимумов мы обучаем соединяющий путь с полигональной параметризацией согласно Уравнению (5) и получаем $\theta_{12}, \theta_{13}, \theta_{23}$. Наконец, для этого набора локальных минимумов мы обучаем соединяющий треугольник и получаем θ_{123}. При обучении треугольника мы инициализируем $\theta_1, \theta_2, \theta_3$ весами, соответствующими локальным минимумам. $\theta_{12}, \theta_{13}, \theta_{23}$ инициализируются весами, соответствующими обученным соединяющим кривым. Инициализация θ_{123} – линейная интерполяция $\theta_1, \theta_2, \theta_3$, $\bar\theta$ – это множество обучаемых и фиксированных параметров: $\bar\theta = \theta \sqcup \theta_f$. В наших экспериментах мы фиксируем $\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23}$, а θ_{123} – обучаемые параметры, т.е. $\theta = \{\theta_{123}\}$ и $\theta_f = \{\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23}\}$. Правило обновления параметров во время обучения треугольника представлено в Алгоритме 1.

Algorithm 1. Оптимизация треугольника

Input: N – количество итераций, $\bar\theta = \theta \sqcup \theta_f$ – множество обучаемых и фиксированных параметров для треугольника $\psi_{\bar\theta}$ (6), η – скорость обучения, L – функция потерь

$\theta^0 = \theta$

For $i \leftarrow 1, N$:

$t_1^i, t_2^i \sim U[0,1]$

if $t_1^i + t_2^i > 1$ **then**

$t_1^i, t_2^i = 1 - t_2^i, 1 - t_1^i$

$t_3^i = 1 - t_1^i - t_2^i$

$t^i = (t_1^i, t_2^i, t_3^i)$, выборка из двумерного стандартного симплекса

$\theta^i \leftarrow \theta^{i-1} - \eta \mathrm{grad}_\theta L(\psi_{\bar\theta}(t^i))$

Return: θ^N

Для вычисления бар-кода на основе обученных треугольников мы делаем дискретизацию T двумерного стандартного симплекса σ, состоящую из n точек. Затем, для каждого обученного соединяющего треугольника, мы вычисляем продолжительность жизни соответствующего топологического признака с помощью Алгоритма 2. Оценка бар-кода представляется как объединение этих сегментов.

Algorithm 2. Вычисление продолжительности существования признака

Input: $\overline{\theta} = \theta \sqcup \theta_f$ – набор обученных параметров, L – функция потерь, T – дискретизированный двумерный стандартный симплекс

$h_{2,3} = \max_{(0,t_2,t_3)\in T} L(\psi_{\overline{\theta}}(t_1,t_2,t_3))$

$h_{1,3} = \max_{(t_1,0,t_3)\in T} L(\psi_{\overline{\theta}}(t_1,t_2,t_3))$

$h_{1,2} = \max_{(t_1,t_2,0)\in T} L(\psi_{\overline{\theta}}(t_1,t_2,t_3))$

$h^1 = \max\{h_{1,2}, h_{1,3}, h_{2,3}\}$ – время рождения топологического признака

$h^2 = \max_{(t_1,t_2,t_3)\in T} L(\psi_{\overline{\theta}}(t_1,t_2,t_3))$ – время исчезновения топологического признака

Return: $[h^1, h^2]$

4. ПЕРСИСТЕНТНОСТЬ ТОПОЛОГИЧЕСКИХ ПРИЗНАКОВ РАЗМЕРНОСТИ r, ВЫЧИСЛЕНИЕ БАР-КОДА ИНДЕКСА r С ПОМОЩЬЮ ГРАДИЕНТНОГО СПУСКА НА СИМПЛЕКСАХ

В данной секции мы описываем алгоритм для вычисления бар-кодов топологических признаков произвольной размерности $r \geq 1$. Эти бар-коды количественно оценивают “продолжительности жизни” критических точек индекса r. Общее определение бар-кодов более высокого индекса основано на эволюции топологических признаков более высокой размерности (циклы, двумерные пустоты и т.д.) в подуровневых множествах функции потерь. r-й бар-код записывает интервалы появления-исчезновения r-мерных циклов.

Для расчета бар-кодов более высоких индексов сначала оптимизируются $(r+1)$-симплексы в пространстве параметров, аналогично оптимизации треугольников, применяя градиент к случайной точке из симплекса.

Множество оптимизированных симплексов строится индуктивно, начиная с множества оптимизированных 0-симплексов (минимумы), затем множества оптимизированных 1-симплексов для каждой пары минимумов, затем оптимизированные 2-симплексы для каждой тройки выбранных минимумов и т.д.

После оптимизации множества симплексов мы вычисляем фильтрацию на этом множестве, определенную как максимальное значение функции потерь на данном симплексе. Следующий шаг – построение фильтрованного симплициального комплекса из оптимизированных симплексов. Линейный граничный оператор ∂_r действует на линейное пространство, формально порожденное $(r+1)$-симплексами, отправляя симплекс в чередующуюся сумму его r-граней. Затем фильтрация используется для вычисления времен появления и исчезновения топологических признаков, приводя множество граничных линейных операторов ∂_r к простой форме [21–23]. Затем бар-код индекса r задается множеством времен рождения и смерти r-мерных топологических признаков в фильтрованном симплициальном комплексе.

Расстояние $\mathbb{W}_\infty$ для бар-кодов индекса r определяется той же формулой (4). Нулевые расстояния для всех индексов указывают на то, что функция потерь, после изменения параметризации в Θ, возможно, может быть сделана выпуклой.

5. ЭКСПЕРИМЕНТЫ

В этом разделе мы применяем предложенный алгоритм для изучения топологии функций потерь нейронных сетей.

Детали экспериментов. Мы обучаем архитектуру LeNet-like на FashionMNIST в течение 50 эпох с размером батча 128, коэффициентом регуляризации 10^{-3}, скоростью обучения 10^{-2} и расписанием скорости обучения согласно [6]. Мы воспроизводим постановку эксперимента из [1] и аналогично рассматриваем нейронные сети ResNet-like и VGG-like, обученные на наборе данных CIFAR10. Сети ResNet-like представляют собой ResNet-20, ResNet-56, ResNet-110, которые имеют соответственно 20, 56, 110 слоев. Сети VGG-like получены из сетей ResNet-like путем удаления резидуальных соединений, и именуются с дополнительным суффиксом ’no short’: ResNet-20-NS, ResNet-56-NS, ResNet-110-NS. Мы также используем WideResNet-56-k, где $k = 1, 2$ означает в k раз больше фильтров на слой. WideResNet-56-1(-NS) совпадает с ResNet-56(-NS). Для вычисления бар-кода мы следуем процедуре, описанной в разделе 3. Для обучения всех промежуточных структур мы используем процедуры обучения идентичные процедуре обучения минимумов.

Результаты. Во-первых, мы проводим эксперименты на сети LeNet-like и наборе данных FashionMNIST как доказательство концепции

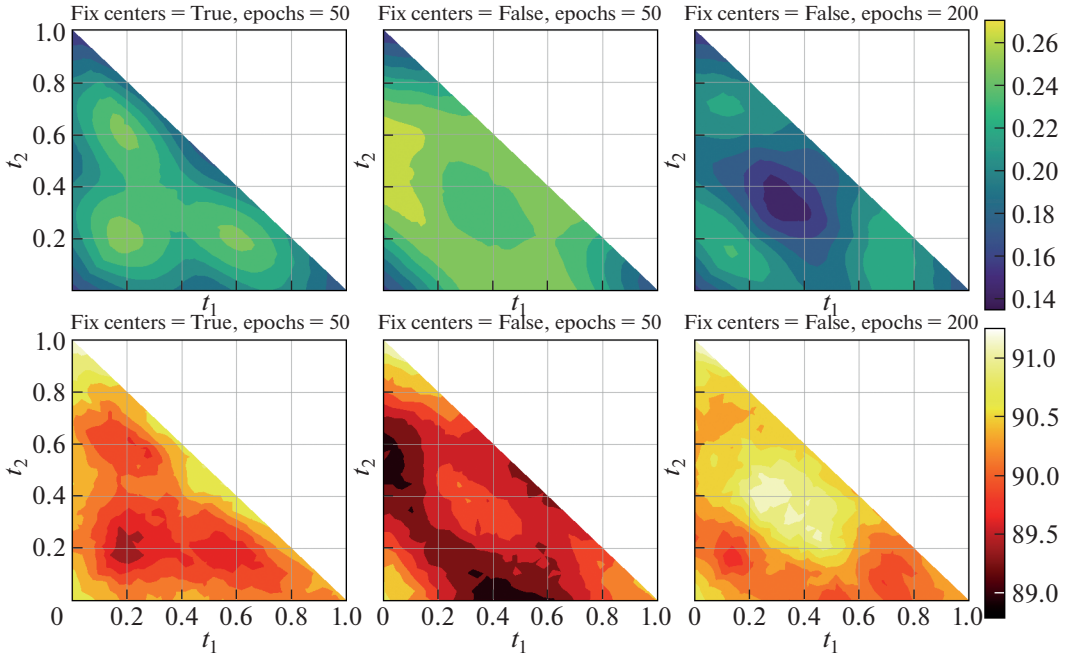

Рис. 1. Визуализация функции потерь (сверху) и метрики точности (снизу) внутри обученного треугольника для LeNet-like и набора данных FashionMNIST. Заголовок каждого графика содержит указание режима обучения – обучаемые или фиксированные центры сторон – и количество эпох обучения.

метода обучения треугольников, соединяющих тройки минимумов в пространстве весов сети. На рис. 1 мы экспериментируем с обоими режимами работы (фиксированными и обучаемыми центрами сторон) и с разными вычислительными бюджетами для модели LeNet-like и набора данных FashionMNIST. Можно заключить, что более гибкая модель приводит к лучшей производительности, но требует больших вычислительных затрат.

Затем мы проводим эксперименты с глубокими нейронными сетями. Аналогично [1], мы исследуем влияние резидуальных соединений, когда глубина или ширина сети увеличивается. Эффект увеличения глубины на сетях ResNet-like и VGG-like, показан на рис. 2, 4а и в табл. 1. Мы можем отметить, что функции потерь сетей с резидуальными соединениями обычно имеют топологические признаки с меньшим временем жизни, следовательно, эти функции потерь ближе к выпуклой функции. Также мы можем отметить, что для ResNet110-NS время жизни топологического признака равно нулю. В этом случае дополнительно следует рассматривать топологические инварианты в более низких/высоких размерностях. Для эффекта увеличения ширины на WideResNet-56-1, WideResNet-56-2 и WideResNet-56-1-NS, WideResNet-56-2-NS см. рис. 3, 4б и табл. 1. Функции потерь сетей ResNet-like ближе к выпуклой функции, что можно увидеть как на рис. 2, так и в табл. 1. В целом увеличение глубины или ширины в сетях ResNet-like приводит не только к уменьшению расстояния $\mathbb{W}_\infty$ на соответствующих бар-кодах, но и к уменьшению барьеров вокруг локальных минимумов в ландшафте функции потерь.

6. ОБСУЖДЕНИЕ

В этом разделе мы обсуждаем практические вопросы и область применения предложенного подхода.

Согласно определению, бар-код включает в себя вычисления для каждой критической точки функции потерь. В случае нейронных сетей прак-

Таблица 1. Расстояние $\mathbb{W}_\infty$, вычисленное для моделей ResNet-like и VGG-like

Тип	ResNet-20	ResNet-56 (WideResNet-56-1)	ResNet-110	WideResNet-56-2
с shortcut	0.083	0.061	0.034	0.025
без shortcut (NS)	0.109	0.078	0.0	0.082

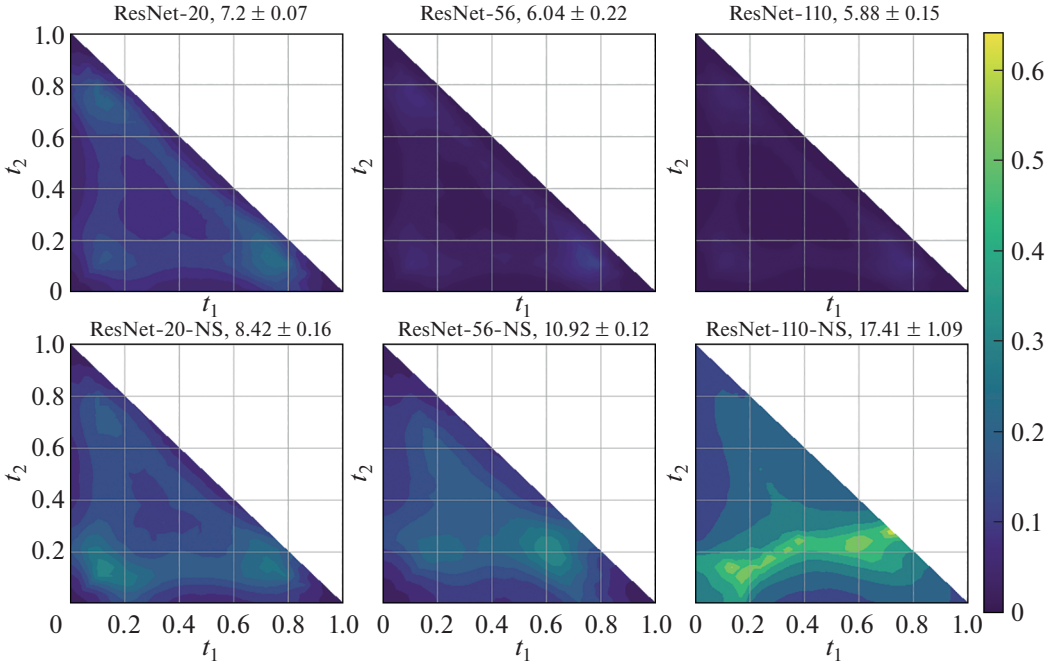

Рис. 2. Визуализация функции потерь внутри обученного треугольника для ResNet и ResNet-noshort с разной глубиной. Заголовок каждого графика содержит спецификацию модели и тестовую ошибку.

тически невозможно достичь всех критических точек во время оптимизации. Стандартное решение в топологическом анализе данных заключается в использовании выборок критических точек. Стабильность персистентных диаграмм, вычисленных с помощью облака точек, подробно изучена в [20]. Кроме того, мы предлагаем выполнять оценку бар-кода только для тех точек, которые достижимы с помощью градиентной оптимизации, что соответствует практическому подходу.

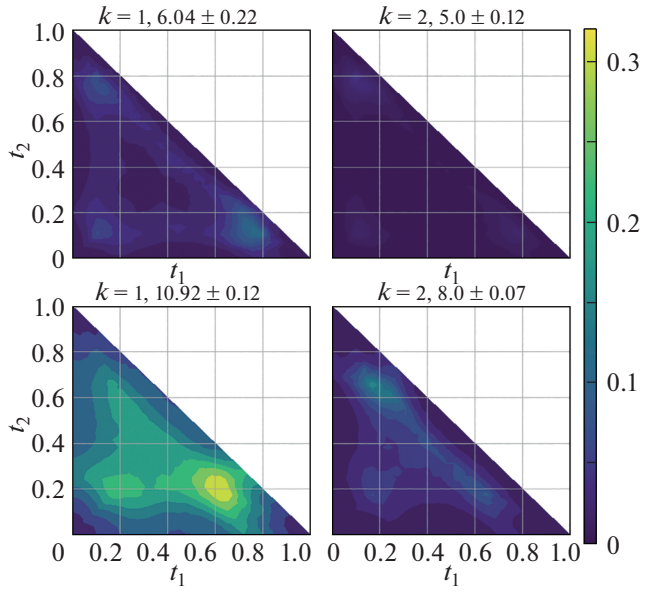

Рис. 3. Визуализация функции потерь внутри обученного треугольника для WideResNet-56-k и WideResNet-56-k-noshort с разной шириной. Заголовок каждого графика содержит спецификацию ширины k и тестовую ошибку.

Одним из возможных применений предложенного подхода является визуализация ландшафта функции потерь. Хотя работа [1] предлагает подход к визуализации через случайные направления в пространстве параметров, у него есть определенные недостатки. В то же время может быть более информативным понять, как процесс оптимизации проходит по поверхности функции потерь и находит ли он области с высокими барьерами. В этом случае предложенный подход более подходящий, так как он выявляет неустранимые препятствия на пути градиентного спуска.

Кроме того, используя предложенное расстояние $\mathbb{W}_\infty$ в пространстве бар-кодов, можно количественно оценить глобальную невыпуклость ландшафта потерь. Это может быть актуально, когда визуальная проверка невозможна или неоднозначна.

Более того, предложенный подход опирается как на пространство параметров, так и на процедуру оптимизации, следовательно, он исследует ландшафт более сложным способом. Этот взгляд является более практичным, так как позволяет получить представление о настройке гиперпараметров оптимизации и проектировании архитектуры.

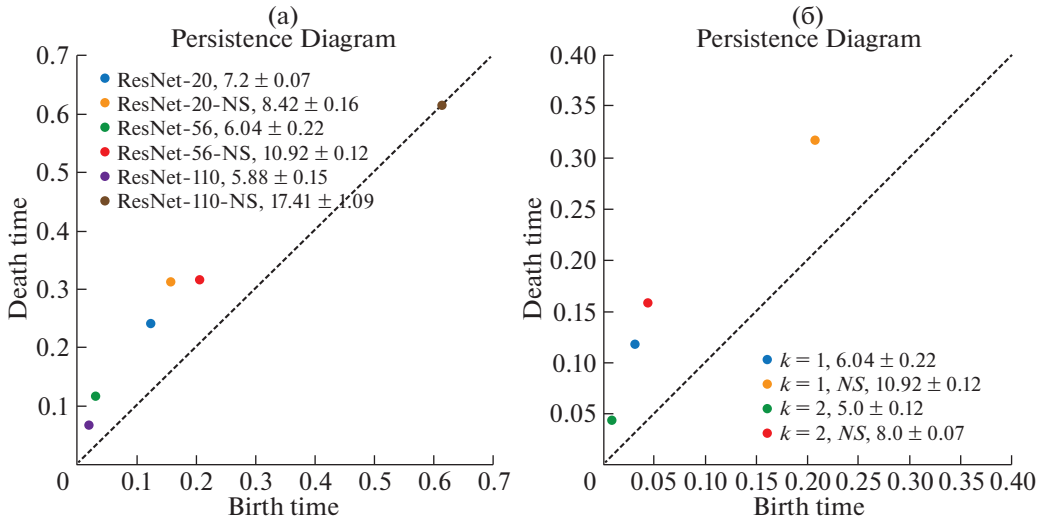

Рис. 4. Персистентные диаграммы. Подпись на каждой легенде содержит спецификацию модели и тестовую ошибку.

7. ЗАКЛЮЧЕНИЕ

Используя топологию, мы предоставляем новый подход к анализу ландшафта функции потерь, который основан на одномерных топологических инвариантах как мере невыпуклости функции потерь с точностью до перепараметризации. Мы предлагаем геометрическое описание этих особенностей и демонстрируем алгоритм их вычисления на практике. Также мы проверили предложенный подход на ряде глубоких нейронных сетей. Согласно нашим результатам, в большинстве случаев предложенный подход определяет топологию ландшафта потерь. Однако он может быть дополнительно усилен с помощью топологических инвариантов более низкой/высокой размерности. Мы обсуждаем несколько случаев использования, где наш подход может быть применим. Таким образом, мы ожидаем, что предложенный метод будет способствовать дальнейшему исследованию ландшафта функции потерь и продвижению топологического анализа функции потерь.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Данная работа была частично поддержана программой Next Generation Program (3rd Call for Proposals).

СПИСОК ЛИТЕРАТУРЫ

1. *Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, Tom Goldstein*. Visualizing the Loss Landscape of Neural Nets. NeurIPS 2018.
2. *Namuk Park, Songkuk Kim*. How Do Vision Transformers Work? ICLR 2022.
3. *Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, Ping Tak Peter Tang*. On Large-Batch Training For Deep Learning: Generalization Gap and Sharp Minima. ICLR 2017.
4. *Laurent Dinh, Razvan Pascanu, Samy Bengio, Yoshua Bengio*. Sharp Minima Can Generalize For Deep Nets. ICML 2017.
5. *Felix Draxler, Kambis Veschgini, Manfred Salmhofer, Fred A. Hamprecht*. Essentially No Barriers in Neural Network Energy Landscape. ICML 2018.
6. *Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, Andrew Gordon Wilson*. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. NeurIPS 2018.
7. *Gregory W. Benton, Wesley J. Maddox, Sanae Lotfi, Andrew Gordon Wilson*. Loss Surface Simplexes for Mode Connecting Volumes and Fast Ensembling. ICML 2021.
8. *Stanislav Fort, Stanislaw Jastrzebski*. Large Scale Structure of Neural Network Loss Landscapes. 2019.
9. *Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur*. The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks. ICLR 2022.
10. *Samuel K. Ainsworth, Jonathan Hayase, Siddhartha Srinivasa*. Git Re-Basin: Merging Models Modulo Permutation Symmetries. ICLR 2023.
11. *Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E. Gonzalez, Kannan Ramchandran, Michael W. Mahoney*. Taxonomizing Local Versus Global Structure in Neural Network Loss Landscapes. NeurIPS 2021.
12. *Ian J. Goodfellow, Oriol Vinyals, Andrew M. Saxe*. Qualitatively Characterizing Neural Network Optimization Problems. ICLR 2015.
13. *Leslie N. Smith, Nicholay Topin*. Exploring Loss Function Topology with Cyclical Learning Rates. ICLR 2017.
14. *Haowei He, Gao Huang, Yang Yuan*. Asymmetric Valleys: Beyond Sharp and Flat Local Minima. 2019.

15. *Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, Richard Socher.* Using Mode Connectivity for Loss Landscape Analysis. 2018.
16. *Ivan Skorokhodov, Mikhail Burtsev.* Loss Surface Sightseeing by Multi-Point Optimization. NeurIPS 2019.
17. *Wojciech Marian Czarnecki, Simon Osindero, Razvan Pascanu, Max Jaderberg.* A Deep Neural Network's Loss Surface Contains Every Low-dimensional Pattern. 2020.
18. *Serguei Barannikov, Alexander Korotin, Dmitry Oganesyan, Daniil Emtsev, Evgeny Burnaev.* Barcodes as summary of loss function's topology. 2020.
19. *Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Sanjeev Arora, Rong Ge.* Explaining Landscape Connectivity of Low-cost Solutions for Multilayer Nets. NeurIPS 2019.
20. *Chazal Frédéric and Michel Bertrand.* An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. Frontiers in artificial intelligence. 2021. V. 4. P. 108.
21. *Le Peutrec D. and Nier F. and Viterbo C.* Precise Arrhenius Law for p-forms: The Witten Laplacian and Morse–Barannikov Complex, Annales Henri Poincaré. Annales Henri Poincaré, 2013.
22. *Zomorodian Afra J.* Computing and comprehending topology: Persistence and hierarchical Morse complexes (Ph.D.Thesis). 2001.
23. *Serguei Barannikov.* Framed Morse complexes and its invariants. Advances in Soviet Mathematics. 1994. V. 21. P. 93–116.

1-DIMENSIONAL TOPOLOGICAL INVARIANTS TO ESTIMATE LOSS SURFACE NON-CONVEXITY

D. S. Voronkova[a], S. A. Barannikov[a,c], and E. V. Burnaev[a,b]

[a]*Skolkovo Institute of Science and Technology, Moscow, Russia*
[b]*Artificial Intelligence Research Institute, Moscow, Russia*
[c]*CNRS, IMJ, Paris Cité University, France*
Presented by Academician of the RAS A.L. Semenov

We utilize the framework of topological data analysis to examine the geometry of loss landscape. With the use of topology and Morse theory, we propose to analyse 1-dimensional topological invariants as a measure of loss function non-convexity up to arbitrary re-parametrization. The proposed approach uses optimization of 2-dimensional simplices in network weights space and allows to conduct both qualitative and quantitative evaluation of loss landscape to gain insights into behavior and optimization of neural networks. We provide geometrical interpretation of the topological invariants and describe the algorithm for their computation. We expect that the proposed approach can complement the existing tools for analysis of loss landscape and shed light on unresolved issues in the field of deep learning.