# Остаточная сеть с рекуррентными структурами

Л. Цзян [1] Ч. Цуй [2] Ц. Ван [3]

Мы вводим рекуррентную структуру (пространственно) на остаточных сетях, с целью улучшить производительность сети при сохранении параметров. Также, мы исследуем поведение рекуррентных структур в остаточных сетях на основе римановых многообразий, вводя кривизну в качестве метрики для нейронных сетей. Кроме того, мы экспериментально подтверждаем, что усиление за счет рекуррентной структуры связано с кривизной, и демонстрируем универсальность рекуррентной структуры как метода повышения производительности сети.

**Ключевые слова:** нейронные сети, риманова геометрия, рекуррентные структуры, многообразие, трансформеры.

## 1. Введение

Глубокие искусственные нейронные сети часто мощнее неглубоких сетей. Но по мере увеличения глубины, количество параметров в них также стремительно растёт. Существует ли способ повысить качество предсказаний сети без увеличения параметров? Вдохновленные рекуррентными связями в головном мозге, некоторые работы, например [1, 2], вводят рекуррентные структуры в неглубокие нейронные сети, тем самым повышая эффективность их работы. Однако, данное направление пока не получило широкого развития, а сам принцип работы рекуррентных структур все ещё остается малоизученным.

Известно, что простое применение рекуррентной структуры увеличивает риск взрыва или исчезновения градиента [3]. Поэтому существует необходимость определения и анализа архитектур, подходящих для

[1] *Цзян Лэй* — аспирант каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: kiwee@outlook.com

Jiang Lei — graduate student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems.

[2] *Чжэнюй Цуй* — аспирант каф. математической кибернетики ВМиК. ф-та МГУ, e-mail: ourobros1234@outlook.com

Cui Zhenyu — graduate student, Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Chair of Mathematical Cybernetics.

[3] *Ван Цзыци* — аспирант каф. теории чисел мех.-мат. ф-та МГУ, e-mail: wangziqi101@outlook.com

Wang Ziqi — graduate student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Number Theory.

включения в них рекуррентных структур. Другой вопрос — какие изменения в сети происходят при включении рекуррентных структур в сеть и каковы соответствующие результаты таких изменений?

Отвечая на первый вопрос, мы обнаружили, что архитектура остаточных сетей (ResNet) [4] подходит для рекуррентных структур больше, чем традиционные нейронные сети. При обратном распространении ошибки градиент в остаточных сетях намного стабильнее градиента, наблюдаемого в обычных сетях [5, 6, 7]. Это значительно снижает риск взрыва или исчезновения градиента из-за рекуррентной структуры.

В отношении второго вопроса мы продолжаем работу [8], вводя риманову геометрию для объяснения работы рекуррентной структуры: рассматривая набор данных как кривую в высоко-размерном пространстве, процесс обучения нейронной сети можно рассматривать как процесс обучения представлению координат для этой кривой. Исходя из этого, мы оцениваем конечное представление координат путем изучения внешней кривизны кривой в выходном пространстве сети и обнаруживаем, что введение рекуррентной структуры позволяет сети научиться лучшему представлению координат. Мы подтвердили нашу теорию экспериментально и обнаружили, что изменение внешней кривизны положительно коррелирует с точностью классификации сети. Согласно нашим сведениям, в нашей работе риманова геометрия впервые вводится в интерпретируемость рекуррентной структуры.

## 2. Методология

Остаточные нейронные сети — одна из важнейших архитектур для глубокого обучения. Типичный остаточный модуль (residual module) состоит из остаточного соединения (residual connection) и пропускных соединений (shortcut connection). Остаточный модуль содержит ряд обучаемых весов, а пропускное соединение добавляет входные данные модуля непосредственно к выходным данным. Формализовать процесс можно следующим образом:

$$F^l(x^l) = x^l + f(\mathbf{W}^l x^l + \mathbf{b}^l),$$

где $f(\mathbf{W}^l x^l + \mathbf{b}^l)$ указывает на остаточный модуль; $\mathbf{W}, \mathbf{b}$ — весовая матрица и смещение; $f(x)$ — нелинейная функция активации.

Процесс прямого распространения сигнала в остаточной сети выглядит следующим образом:

$$x^L = x_0 + \sum_{l=1}^{L} F^l(x^l).$$

## 2.1. В остаточной сети набор данных представлен как $\mathcal{C}^1$-многообразие

1) Предположим, что набор данных $D$ представляет собой топологическое многообразие (хаусдорфово топологическое пространство, которое имеет счетную базу и локальное сходство с евклидовым пространством).

2) Входное пространство нейронной сети обозначим через $Y^0$, а выходное пространство каждого слоя $L$-слойной нейронной сети обозначим через $Y^1, \cdots Y^L$, где $Y^L$ — последнее выходное пространство нейронной сети. Предположим, что все эти пространства имеют одинаковую размерность, обозначаемую $d$. Здесь мы также предполагаем, что размерность набора данных равна $d$.

3) Отображение из набора данных во входное пространство нейронной сети обозначим через $\varphi_0 : D \to Y^0$. Затем, определим отображение $\varphi_i : D \to Y^i$ из набора данных в каждое выходное пространство $Y^i, i = 1, 2, \cdots L$ нейронной сети следующим образом:

$$\varphi_i := F^{i-1} \circ \cdots \circ F^0 \circ \varphi_0, \qquad i = 1, 2, \cdots L,$$

где функция $F^i : Y^i \to Y^{i+1}, i = 0, 1, \cdots L$ является функцией преобразования координат. В остаточной сети такие преобразования принадлежат множеству непрерывно-дифференцируемых функций $\mathcal{C}^1$.

В остаточной сети набор данных обозначается как $\mathcal{C}^1$-многообразие, а это значит, с точки зрения непрофессионала, что выходные пространства различных слоев нейронной сети можно трактовать как набор координат для точки из набора данных, то есть все выходные пространства нейронной сети представляет собой атлас этого многообразия. А функция преобразования между слоями в остаточной сети — это функция преобразования координат между этими локальными системами координат.

Полагая, что процесс обучения нейронной сети — это нахождение набора данных в соответствующих локальных системах координат с помощью преобразований между этих системами, каждое такое преобразование задается соответствующими весами $W$ и смещениями $b$ слоя.

## 2.2. Процесс обучения нейронной сети с геометрической точки зрения

В задаче классификации нейронная сеть сближает удаленные точки во входных данных, принадлежащие к одному и тому же классу, и оттягивает соседние точки, принадлежащие к разным классам. Это явственно

отразится на выходных пространствах различных слоев нейронной сети, то есть, нейронная сеть может менять взаимное расположение точек в выходных пространствах различных слоев нейронной сети. Например, если две точки, находящиеся далеко друг от друга на прямой, принадлежат к одному и тому же классу, тогда нейронная сеть сблизит их с помощью изгиба этой прямой. Мы вводим понятие из римановой геометрии — кривизна, которая может измерять дисторсию окрестности точки в пространстве, благодаря чему мы можем исследовать поведение нейронных сетей, анализируя геометрические свойства входных данных в выходном пространстве различных слоев нейронной сети.

Из раздела 2.1 мы знаем, что набор данных можно понимать как $\mathcal{C}^1$-многообразие, а каждое выходное пространство различных слоев нейронной сети представляет собой локальную систему координат этого многообразия. В данной задаче мы рассматриваем одномерное многообразие $l^0(\theta)$ во входном пространстве, где $\theta$ — скалярные координаты на этом многообразии. Обозначим данное многообразие в разных локальных системах координат (в выходных пространствах различных слоев нейронной сети) через $l^i(\theta)$, где $i$ - индекс локальной системы координат (т.е. индекс выходного пространства различных слоев нейронной сети). Далее мы количественно анализируем геометрические свойства набора входных данных в выходном пространстве различных слоев нейронной сети, вычисляя внешнюю кривизну в данной точке на $l^i(\theta)$, и исследуем взаимосвязь между этим геометрическим свойством и точностью классификации нейронной сети. Теперь в каждой точке $\theta$ имеет соприкасающуюся окружность, которая касается прямой в данной точке и порядок которой не ниже 2. Эта окружность является наилучшим приближением заданной кривой в окрестности данной точки. Пусть радиус окружности в данной точке равен $R(\theta)$, тогда внешняя кривизна $k(\theta)$ определяется как $k(\theta) = \frac{1}{R(\theta)}$ и зависит от вложения в выходных пространствах различных слоев нейронной сети (ее образа в локальных системах координат) и инвариантна относительно конкретной параметризации $\theta$. Пусть $v^i(\theta) = \frac{\partial l^i(\theta)}{\partial \theta}$ и $a^i(\theta) = \frac{\partial v^i(\theta)}{\partial \theta}$, тогда точная формула для кривизны $k^i(\theta)$ в $i$-м выходном пространстве различных слоев нейронной сети следующая:

$$k^i(\theta) = \frac{\sqrt{(v^i(\theta), v^i(\theta))(a^i(\theta), a^i(\theta)) - (v^i(\theta), a^i(\theta))^2}}{(v^i(\theta), v^i(\theta))^{\frac{3}{2}}}, \qquad i = 0, \cdots L.$$

В своей задаче мы исследуем изменение кривизны $k^i(\theta)$ кривой в выходном пространстве каждого из слоев нейронной сети $Y^i, i \in [1, L]$ при обучении сети.

## 2.3. Эксперимент

Мы утверждаем, что рекуррентная структура увеличивает способность сети преобразовывать входные данные и, таким образом, лучше подстраиваться под них. Сначала мы проверили нашу гипотезу на небольшой сети, и для этого вычислили кривизну точки на кривой. Мы изучили изменение кривизны трех остаточных сетей (неглубокой, рекуррентной и глубокой) в рамках небольшой задачи классификации, которая имеет два набора точек на плоскости. Задача состоит в том, чтобы разделить точки на две группы с помощью гиперплоскости. Каждый слой этих трех сетей состоит всего из двух нейронов: неглубокая состоит из 4 слоев, количество параметров - 24, глубокая содержит 13 слоев, количество параметров - 78, тогда как в рекуррентной сети у нас есть 1 слой с 13-ю рекуррентными шагами, а количество параметров - **6**.

Изменение кривизны и оценка качества предсказаний сети показаны на рисунке 1 и в таблице 1 соответственно. Можно заметить, что внешняя кривизна увеличивается по мере увеличения количества слоев. Внешняя кривизна глубокой сети выше внешней кривизны однослойной рекуррентной сети, которая, в свою очередь, выше внешней кривизны неглубокой четырехслойной сети. Примечательно, что величина внешней кривизны положительно коррелирует с качеством классификации.

Координатное представление точек показано на рисунке 2. Мы видим, что, в сравнении с неглубокой сетью, рекуррентная преобразует координаты точек таким образом, чтобы их было легче разделить гиперплоскостями. В свою очередь это улучшает качество классификации точек.

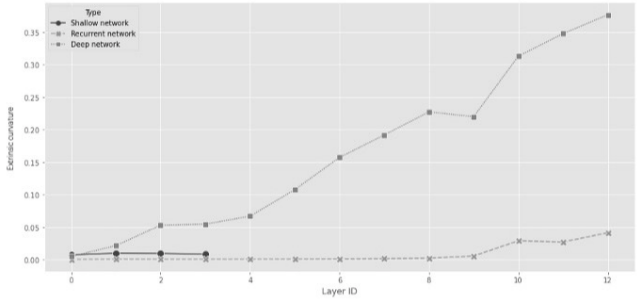


Рис. 1. Изменение внешней кривизны с увеличением количества слоев в обученной сети: глубокая сеть имеет более высокую внешнюю кривизну, нежели рекуррентная и неглубокая сети

| Модель | Неглубокая | Рекуррентная | Глубокая |
|---|---|---|---|
| Потери/Точность | 0.46/77% | **0.15/ 97%** | 0.05/99% |

Таблица 1. Потери и точность сети после обучения

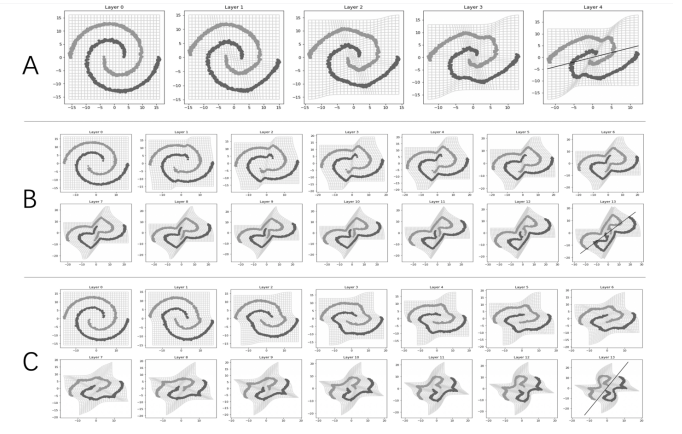


Рис. 2. Визуализация координатных представлений точек кривых, полученных для трех различных сетей: (A), (B), (C) — точки неглубокой, рекуррентной и глубокой сетей соответственно

## 3. Рекуррентные структуры на практике

Мы сравниваем изменение внешней кривизны в разных сетях на общем наборе данных. В качестве объекта исследования внешней кривизны мы выбрали Vision Transformer (ViT) [9, 10], натренированный на датасете ImageNet [11]. Далее мы используем четыре вариации ViT сети: (A) неглубокая ViT сеть, состоящая из трёх модулей трансформера; (B) рекуррентная ViT сеть, состоящая из трёх модулей трансформеров, с двух-, четырёх-, и снова двух-рекуррентными ступенями соответственно; (C) глубокая ViT, состоящая из семи трансформеров; (D) глубокая рекуррентная ViT, также состоящая из семи трансформеров, но с двумя рекуррентными ступенями в каждом трансформере. На рисунке 3 представлено изменение внешней кривизны трех сетей A, B, C. Подобные графики мы уже наблюдали в неглубоких сетях на рисунке 1. В таблице 2 представлено качество предсказаний всех четырех сетей ViT (A, B, C, D). Как можно заметить, для больших сетей внешняя кривизна также положительно коррелирует с оценкой качества сети.

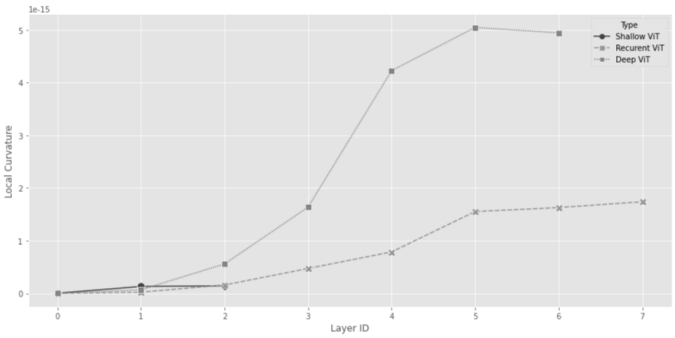


Рис. 3. Внешняя кривизна неглубокой, рекуррентной и глубокой сетей ViT

| Модель | (A) 3-без | (B) 3-с | (C) 7-без | (D) 7-с |
|---|---|---|---|---|
| Точность | 61.53 | **67.16** | 71.63 | **74.57** |

Таблица 2. Лучшая точность ViT (%) с рекуррентной структурой и без нее. Данные получены для датасета ImageNet

## 4. Заключение

Мы первыми вводим в рекуррентную структуру внешнюю кривизну как меру способности сети к преобразованию пространства данных, а также впервые вводим рекуррентную структуру в ViT, обнаруживая, что она может значительно улучшить точность классификации без дополнительных параметров. Мы демонстрируем, что с помощью рекуррентной структуры сети можно получить лучшие промежуточные представления, тем самым улучшая производительность сети. Данный факт подтверждается нами экспериментально.

## Список литературы

[1] Lan, Zhenzhong and Chen, Mingda and Goodman, Sebastian and Gimpel, Kevin and Sharma, Piyush and Soricut, Radu, “Albert: A lite bert for self-supervised learning of language representations”, 2019.

[2] Kubilius, Jonas and Schrimpf, Martin and Nayebi, Aran and Bear, Daniel and Yamins, Daniel LK and DiCarlo, James J, “CORnet: modeling the neural mechanisms of core object recognition”, 2018.

[3] Pascanu, Razvan and Mikolov, Tomas and Bengio, Yoshua, "On the difficulty of training recurrent neural networks", *International conference on machine learning*, 2013, 1310–1318.

[4] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, "Deep residual learning for image recognition", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–778.

[5] Zaeemzadeh, Alireza and Rahnavard, Nazanin and Shah, Mubarak, "Norm-preservation: Why residual networks can become extremely deep?", *IEEE transactions on pattern analysis and machine intelligence*, **43** (2020), 3980–3990.

[6] Tarnowski, Wojciech and Warcho, Piotr and Jastrzebski, Stanislaw and Tabor, Jacek and Nowak, Maciej, "Dynamical isometry is achieved in residual networks in a universal way for any activation function", *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, 2221–2230.

[7] Balduzzi, David and Frean, Marcus and Leary, Lennox and Lewis, JP and Ma, Kurt Wan-Duo and McWilliams, Brian, "The shattered gradients problem: If resnets are the answer, then what is the question?", *International Conference on Machine Learning*, 2017, 342–350.

[8] Hauser, Michael and Ray, Asok, "Principles of Riemannian geometry in neural networks", *Advances in neural information processing system*, **30** (2017).

[9] Dosovitskiy, Alexey and Beyer, Lucas and Kolesnikov, Alexander and Weissenborn, Dirk and Zhai, Xiaohua and Unterthiner, Thomas and Dehghani, Mostafa and Minderer, Matthias and Heigold, Georg and Gelly, Sylvain and others, "An image is worth 16x16 words: Transformers for image recognition at scale", 2020.

[10] Yuan, Li and Chen, Yunpeng and Wang, Tao and Yu, Weihao and Shi, Yujun and Jiang, Zi-Hang and Tay, Francis EH and Feng, Jiashi and Yan, Shuicheng, "Tokens-to-token vit: Training vision transformers from scratch on imagenet", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 558–567.

[11] Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Li, Kai and Fei-Fei, Li, "Imagenet: A large-scale hierarchical image database", *2009 IEEE conference on computer vision and pattern recognition*, 2009, 248–255.

**Residual network with recurrent structures**
**Jiang Lei, Cui Zhenyu, Wang Ziqi**

We introduce a recurrent structure (spatially) on residual networks, which can improve the performance of the network while saving parameters. We investigate the behaviour of recurrent structures in residual networks based on Riemannian manifolds, introducing curvature as a metric for neural networks. We also experimentally verify that the gain due to the recurrent structure is related to the curvature, and demonstrate the generality of the recurrent structure as a method to improve the performance of the network.

**Keywords:** Neural Netrowks, Riemannian geometry, Recurrent structures, Manifold, Transformers.

## Список литературы

[1] Lan, Zhenzhong and Chen, Mingda and Goodman, Sebastian and Gimpel, Kevin and Sharma, Piyush and Soricut, Radu, "Albert: A lite bert for self-supervised learning of language representations", 2019.

[2] Kubilius, Jonas and Schrimpf, Martin and Nayebi, Aran and Bear, Daniel and Yamins, Daniel LK and DiCarlo, James J, "CORnet: modeling the neural mechanisms of core object recognition", 2018.

[3] Pascanu, Razvan and Mikolov, Tomas and Bengio, Yoshua, "On the difficulty of training recurrent neural networks", *International conference on machine learning*, 2013, 1310–1318.

[4] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, "Deep residual learning for image recognition", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–778.

[5] Zaeemzadeh, Alireza and Rahnavard, Nazanin and Shah, Mubarak, "Norm-preservation: Why residual networks can become extremely deep?", *IEEE transactions on pattern analysis and machine intelligence*, **43** (2020), 3980–3990.

[6] Tarnowski, Wojciech and Warcho, Piotr and Jastrzebski, Stanislaw and Tabor, Jacek and Nowak, Maciej, "Dynamical isometry is achieved in residual networks in a universal way for any activation function", *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, 2221–2230.

[7] Balduzzi, David and Frean, Marcus and Leary, Lennox and Lewis, JP and Ma, Kurt Wan-Duo and McWilliams, Brian, "The shattered

gradients problem: If resnets are the answer, then what is the question?", *International Conference on Machine Learning*, 2017, 342–350.

[8] Hauser, Michael and Ray, Asok, "Principles of Riemannian geometry in neural networks", *Advances in neural information processing system*, **30** (2017).

[9] Dosovitskiy, Alexey and Beyer, Lucas and Kolesnikov, Alexander and Weissenborn, Dirk and Zhai, Xiaohua and Unterthiner, Thomas and Dehghani, Mostafa and Minderer, Matthias and Heigold, Georg and Gelly, Sylvain and others, "An image is worth 16x16 words: Transformers for image recognition at scale", 2020.

[10] Yuan, Li and Chen, Yunpeng and Wang, Tao and Yu, Weihao and Shi, Yujun and Jiang, Zi-Hang and Tay, Francis EH and Feng, Jiashi and Yan, Shuicheng, "Tokens-to-token vit: Training vision transformers from scratch on imagenet", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 558–567.

[11] Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Li, Kai and Fei-Fei, Li, "Imagenet: A large-scale hierarchical image database", *2009 IEEE conference on computer vision and pattern recognition*, 2009, 248–255.