Выбор оптимальной структуры модели глубокого обучения с контролем эксплуатационных характеристик

Фирсов Сергей

Московский физико-технический институт

2024

Цель исследования:

- **Neural Architecture Search:** Метод автоматизированного поиска оптимальной архитектуры нейросети.
- **Цель:** Получать архитектуры решающие поставленную ML задачу, при этом удовлетворяя вычислительным или ресурсным ограничениям.
- **Проблемы:**
  - Обширное пространство для поиска
  - Баланс точности и сложности
  - Получение семейства решений

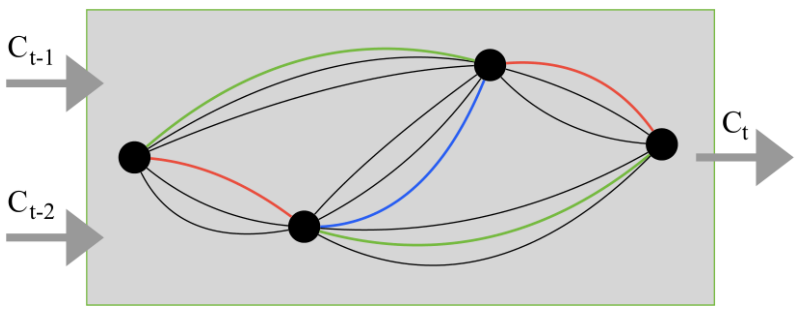


Иллюстрация клетки из метода DARTS

Постановка задачи

- Модель — отображение $\mathbb{X} \times \Gamma \times \mathbb{W} \to \mathbb{Y}$, где $\gamma \in \Gamma$ задает вычислительный граф модели.
- Будем задавать $\gamma \sim \mathbf{GS}(\boldsymbol{\alpha}(\boldsymbol{S}, \boldsymbol{a}), t)$, а архитектурные параметры $\boldsymbol{\alpha}$ представим как функцию от **векторного** параметра сложности $\boldsymbol{S}$ в пространство архитектур Γ:

$$\boldsymbol{\alpha} : (\boldsymbol{S}, \boldsymbol{a}) \to \Gamma,$$

  где $\boldsymbol{S} \in \Delta^{k-1}$ — $(k-1)$-размерный симплекс, k это количество операций, а $\boldsymbol{a} \in \mathbb{R}^n$ — параметры гиперсети.
- **Оптимизационная задача:** минимизация функционала

$$\mathbb{E}_{\boldsymbol{S} \sim \boldsymbol{U}(\Delta^{k-1})} \mathbb{E}_{\gamma \sim \mathbf{GS}(\boldsymbol{\alpha}(\boldsymbol{S}, \boldsymbol{a}), t)} \left[L(\boldsymbol{w}^*, \boldsymbol{\alpha}(\boldsymbol{S}, \boldsymbol{a})) + \kappa \cdot \mathrm{Lat}(\boldsymbol{\alpha}(\boldsymbol{S}, \boldsymbol{a})) \right] \to \min_{\boldsymbol{w}, \boldsymbol{a}},$$

где $\mathrm{Lat}(\boldsymbol{\alpha}(\boldsymbol{S}, \boldsymbol{a})))$ — регуляризационный член отвечающий за сложность модели и задержку выполнения операций

Архитектура решения

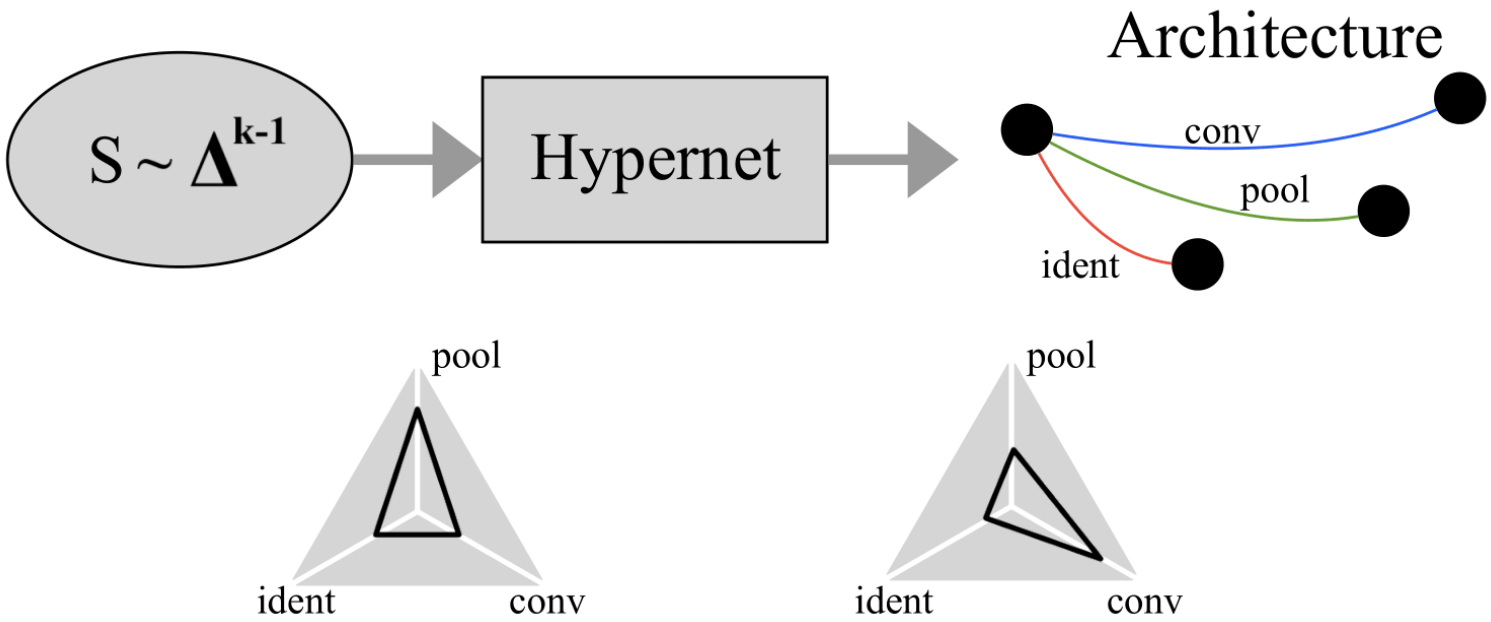


Рис. 2: Иллюстрация предлагаемого метода.

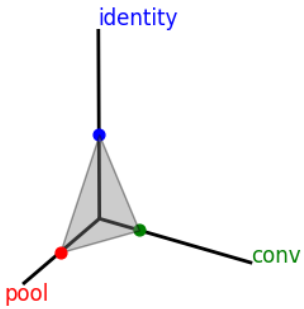
- Предлагается использовать векторный параметр сложности $\boldsymbol{S}$, компоненты которого — коэффициенты регуляризации по соответствующим операциям
- Гиперсеть на основе $\boldsymbol{S}$ генерирует архитектурные параметры для нейросети

Постановка эксперимента

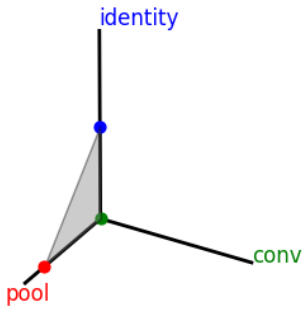
- Цель эксперимента – проверка работоспособности предлагаемого метода, качества получаемых моделей и возможности контроля эксплуатационных характеристик.
- Эксперимент проводится на выборке Fashion-MNIST.
- Модель состоит из трех ячеек, в каждой по 4 вершины.
- Во время обучения температура распределения гумбель-софтмакс понижалась от 1 до 0.2.

Результаты

1. Одинаковый штраф **0.33, 0.33, 0.33**



2. Увеличенный штраф за свёртки **0.15,0.15,0.70**



3. Увеличенный штраф за пулинг **0.70,0.15,0.15**

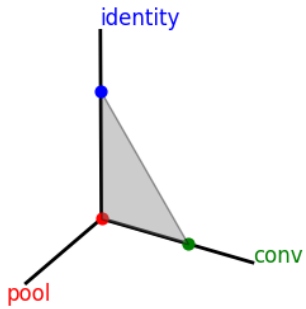


Рис 3. Иллюстрации распределения операций в моделях

Модели	1	2	3
Accuracy	82.5%	79.2%	85.0%
Количество параметров	38304	5120	143488
Количество pooling	12	17	**0**
Количество convolution	6	**0**	13
Количество identity	10	11	15

Результаты

- Предложен метод позволяющий получать семейство моделей с возможностью контроля сложности обучения и аппаратных ограничения.
- Метод обладает возможностью получать архитектуры моделей за счёт изменения вектора параметра сложности без необходимости дополнительного дообучения.
- Вычислительные эксперименты подтверждают работоспособность метода и демонстрируют заявленную гибкость получаемого решения.

References I

[1] Han Cai, Ligeng Zhu и Song Han. *ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware*. https://arxiv.org/abs/1812.00332. 2019. arXiv: 1812.00332 [cs.LG].

[2] Thomas Elsken, Jan Hendrik Metzen и Frank Hutter. "Neural Architecture Search: A Survey". В: *Journal of Machine Learning Research* 20.55 (2019), с. 1—21. URL: http://jmlr.org/papers/v20/18-598.html.

References II

[3] Olga Grebenkova, Oleg Bakhteev и Vadim Strijov. “Deep Learning Model Selection With Parametric Complexity Control”. В: *Proceedings of the 15th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, INSTICC. SciTePress, 2023, с. 65—74. ISBN: 978-989-758-623-1. DOI: 10.5220/0011626900003393. URL: https://www.scitepress.org/PublicationsDetail.aspx?ID=lRGAknY5+Ow=&t=1.

[4] Hanxiao Liu, Karen Simonyan и Yiming Yang. “DARTS: Differentiable architecture search”. В: *International Conference on Learning Representations*. 2019.

[5] Bichen Wu и др. “FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search”. В: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Июнь 2019.

References III

[6] Konstantin Yakovlev и др. “Neural architecture search with structure complexity control”. В: *Recent Trends in Analysis of Images, Social Networks and Texts*. 2022, с. 207—219. URL: https://link.springer.com/book/10.1007/978-3-031-15168-2.