

Автоматическое построение нейросети оптимальной сложности.

*Маркин В. О.¹, Забазнов А. Г.¹, Горян Н. А.¹, Губанов С. Е.¹, Таранов С. К.¹,
Криницкий К. Д.¹, Товкес А. А.¹, Улитин А. Ю.¹*

markin1198@mail.ru; antoniozabaznov@yandex.ru; goryan.na@phystech.edu;

sergey.gubanov@phystech.edu; taranov.sk@phystech.edu;

krinitskiy.kd@phystech.edu; tovkes.aa@phystech.edu; ulitin.ayu@phystech.edu

Московский физико-технический институт¹

В работе рассматривается задача построения оптимальной структуры нейронной сети и исследуется вопрос устойчивости построенной модели. Для оптимизации структурных параметров используется переход от выбора конкретной архитектуры к выбору комбинации различных архитектур сети и вариационный подход. Также исследуется влияние изменения данных на структуру сети. Для оценки качества и устойчивости моделей, построенных при помощи данного метода, проводятся эксперименты на выборке CIFAR10. Проводится сравнение предложенного алгоритма с другими методами поиска оптимальных моделей нейронной сети.

Ключевые слова: *нейронные сети, автоматическое построение нейронных сетей, оптимальная структура нейронной сети*

Введение

При использовании нейросетевых моделей в анализе данных часто встает вопрос о выборе архитектуры модели. Нейронная сеть имеет большое число гиперпараметров. Например, нейронная сеть, построенная по восьмислойной архитектуре AlexNet, имеет около 60 миллионов параметров. Долгое время для поиска оптимальной структуры нейросети и настройки её параметров использовались перебор и различные эвристические соображения [1]. Такие подходы вычислительно неэффективны и не дают гарантий оптимальности полученной модели.

Под оптимальной моделью понимается структура обучаемой сети и совокупность её гиперпараметров, которая даёт приемлемое качество классификации или регрессии. В данной работе в качестве критерия выбора модели предлагается сложность модели, то есть величина, учитывающая сложность описания совокупности выборки и модели. Под описанием выборки понимается приближенная оценка сложности модели, основанная на связи с её правдоподобием[2]

Существует несколько подходов выбора модели оптимальной сложности. В работе [3] представлен метод выбора модели с использованием прореживания нейронной сети, который заключается в обучении максимально большой сети, при последующем удалении части связей. Другой подход заключается в предсказании структуры модели другой нейросетью [4].

В данной работе для выбора оптимального набора гиперпараметров проводится процедура релаксации [5] — переход от дискретного множества возможных значений гиперпараметров к непрерывному множеству их комбинаций. Эта процедура позволяет параметризовать структуру модели некоторым действительным вектором. Такой подход дает возможность применять методы непрерывной оптимизации для нахождения наилучшего набора гиперпараметров. В основе разработанного метода лежит алгоритм DARTS,

предложенный в работе[6]. Оптимизация гиперпараметров проводится градиентными методами [7, 8, 9] либо с использованием Гауссовских процессов и Байесовской оптимизации.

Проверка и анализ метода проводится на выборке CIFAR-10[10]. В ходе экспериментов оценивается не только качество, которое дает полученная модель но и её вычислительная сложность и устойчивость. Проводится сравнение представленного метода с эвристическими алгоритмами выбора модели, а также с алгоритмом DARTS.

Постановка задачи

Пусть заданы обучающая и валидационная выборки:

$$\mathfrak{D}^{\text{train}} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \dots, m^{\text{train}},$$

$$\mathfrak{D}^{\text{valid}} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \dots, m^{\text{valid}},$$

состоящие из множеств пар объект-метка,

$$\mathbf{x}_i \in \mathbf{X} \subset \mathbb{R}^n, \quad y_i \in \mathbf{Y} \subset \mathbb{R}.$$

Метка y объекта \mathbf{x} принадлежит множеству $y \in \mathbf{Y} = \{1, \dots, Z\}$, где Z - количество классов.

Модель задаётся ориентированным графом $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, где для каждого ребра (i, j) заданы базовые функции $\mathbf{g}^{i,j}$, $|\mathbf{g}^{i,j}| = K^{i,j}$ и их веса $\gamma^{i,j}$. Требуется построить такую модель \mathbf{f} с параметрами $\mathbf{W} \in \mathbb{R}^n$:

$$\mathbf{f}(\mathbf{x}, \mathbf{W}) = \{\mathbf{f}_i(\mathbf{x}, \mathbf{w}_i)\}_{i=1}^{|\mathbf{V}|}$$

где $\mathbf{f}_i(\mathbf{x}, \mathbf{w}_i)$ - подмодель с параметрами \mathbf{w}_i задаётся как:

$$\mathbf{f}_i(\mathbf{x}, \mathbf{w}_i) = \sum_{j \in \text{adj}(i)} \langle \gamma^{i,j}, \mathbf{g}^{i,j} \rangle \mathbf{f}_j(\mathbf{x}, \mathbf{w}_j)$$

Тогда параметры модели определяются как конкатенация всех параметров каждой подмодели: $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_{|\mathbf{V}|}]$, а структура модели $\mathbf{\Gamma}$ задаётся вектором $\{\gamma^{i,j}\}_{\mathbf{E}}$.

Функция потерь на обучении L и функция потерь на валидации Q задаются как:

$$L(\mathbf{W}, \mathbf{A}, \mathbf{\Gamma}) = \log p(\mathbf{Y}^{\text{train}} | \mathbf{X}^{\text{train}}, \mathbf{W}, \mathbf{\Gamma}) + e^{\mathbf{A}} \|\mathbf{W}\|^2,$$

$$Q(\mathbf{W}, \mathbf{\Gamma}) = \log p(\mathbf{Y}^{\text{valid}} | \mathbf{X}^{\text{valid}}, \mathbf{W}, \mathbf{\Gamma}) + \lambda p(\mathbf{\Gamma}),$$

где \mathbf{A} и λ — регуляризационные слагаемые, $p(\mathbf{\Gamma})$ - произведение всех произведение вероятностей всех $\gamma^{i,j} \in \mathbf{\Gamma}$. Перед подсчётом значения функции потерь на валидации делается априорное предположение о распределении \mathbf{j} том, что вектор $\mathbf{\Gamma} = \{\gamma^{i,j}\}$ имеет распределение Дирихле[11].

Вектор $\{\gamma^{i,j}\}$ имеет распределение Дирихле с параметром α , если:

$$f(\gamma) = f(\gamma_1, \dots, \gamma_K) = \begin{cases} \frac{\mathbf{F}(K \times \alpha)}{\mathbf{F}(\alpha)^K} \prod_{i=1}^K \gamma_i, & \gamma \in \mathbf{S} \\ 0, & \gamma \notin \mathbf{S} \end{cases}$$

где \mathbf{F} - гамма-функция, \mathbf{S} - симплекс: $\{\gamma \in \mathbb{R}^K : \sum_{i=1}^K \gamma_i = 1, \gamma_i \geq 0\}$.

Требуется решить задачу двухуровневой оптимизации, оптимизируя параметры модели по обучающей выборке, а структуру модели по валидационной:

$$\mathbf{W}^*(\Gamma) = \arg \min_{\mathbf{W}} L(\mathbf{W}, \Gamma)$$

$$\Gamma^*, \mathbf{A}^* = \min_{\Gamma, \mathbf{A}} Q(\mathbf{W}^*(\Gamma), \Gamma)$$

Релаксация

Известно множество всех возможных операций $\mathbf{g}^{i,j} \in \mathbf{G}$. Для перехода к непрерывному пространству таких функций проводится релаксация каждой операции с использованием softmax:

$$\bar{\mathbf{g}}^{(i,j)}(x) = \sum_{\mathbf{g} \in \mathbb{K}} \frac{\exp(\gamma_{\mathbf{g}}^{(i,j)})}{\sum_{\bar{\mathbf{g}} \in \mathbb{K}} \exp(\gamma_{\bar{\mathbf{g}}}^{(i,j)})} \mathbf{g}(x),$$

где $\gamma^{i,j}$ — вектор, параметризующий комбинацию базовых функций. Таким образом, путём подбора непрерывных параметров γ осуществляется переход к задаче поиска базовой функции. В конце поиска, каждая комбинация базовых функций $\bar{\mathbf{g}}^{(i,j)}(x)$ меняется на $\mathbf{g}^{(i,j)} = \arg \max_{\mathbf{g} \in \mathbb{K}} \gamma_{\mathbf{g}}^{(i,j)}$.

Эксперимент

Для проверки влияния регуляризации структуры на итоговую модель был проведен вычислительный эксперимент. В качестве выборки использовалась выборка изображений CIFAR-10 [10]. Рассматривалась задача классификации. Основным критерием качества выступал

$$Accuracy = \frac{1}{m^{\text{valid}}} \sum_{\mathbf{x}, y \in \mathcal{D}^{\text{valid}}} I(\mathbf{f}(\mathbf{x}), y)$$

Эксперимент проводился в следующих режимах:

1) Алгоритм выбора структуры со слабой регуляризацией.

В качестве регуляризации структуры выступало распределение Дирихле с параметром $\alpha = 1$.

2) Алгоритм выбора структуры с меняющейся регуляризацией.

В качестве регуляризации структуры выступало распределение Дирихле с параметром α , меняющимся на каждой эпохи в интервале $(10^{-30}; 1)$.

3) Алгоритм выбора структуры с сильной регуляризацией.

В качестве регуляризации структуры выступало распределение Дирихле с параметром $\alpha = 10^{-30}$.

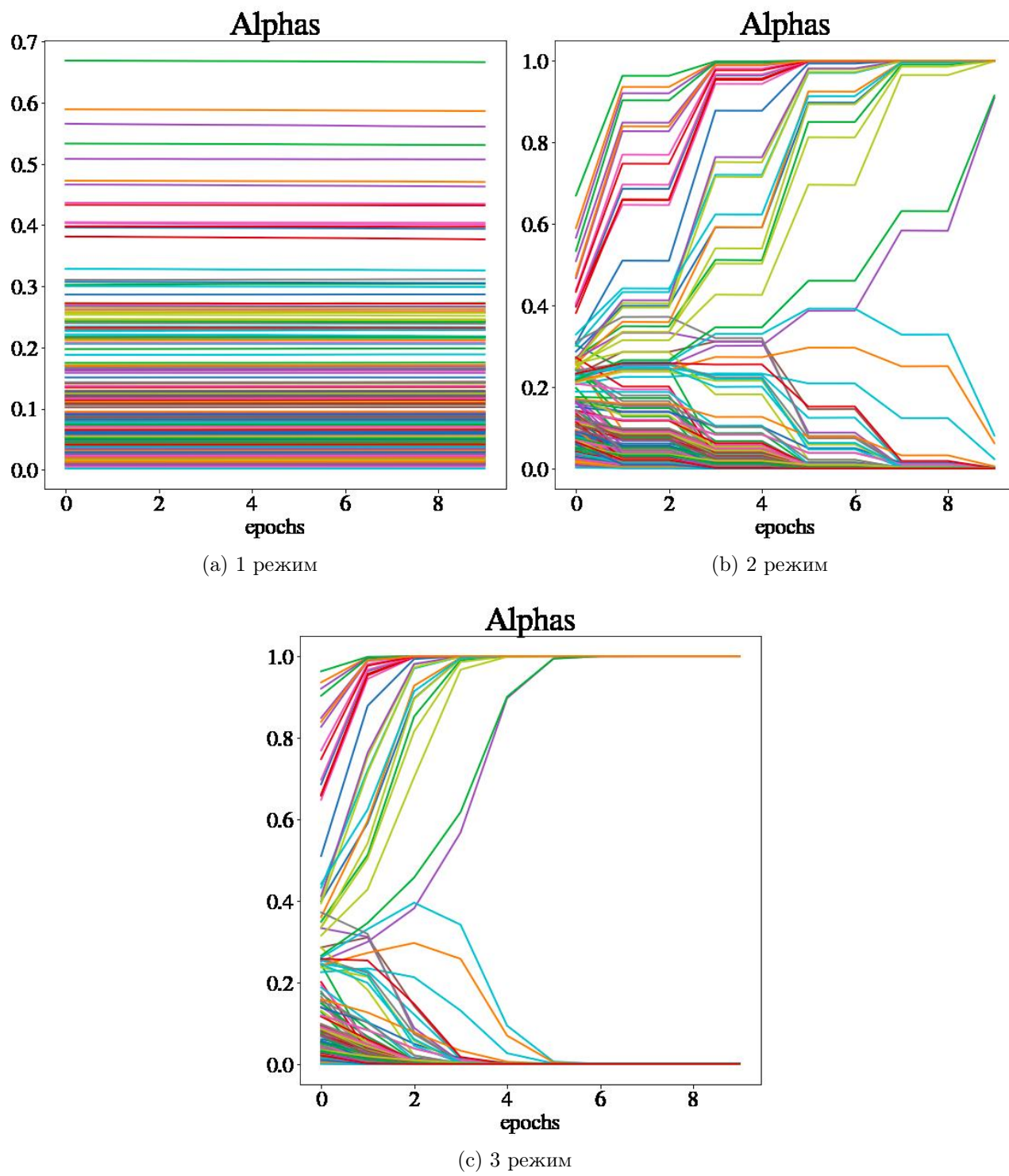


Рис. 1.

На рис. 1 изображена общая динамика изменения структурных параметров в процессе обучения.

На рис. 2 для 3 запусков в каждом режиме соответствие между top-1 accuracy на тестовых данных и количеством параметров в сети, больших порога 0.1.

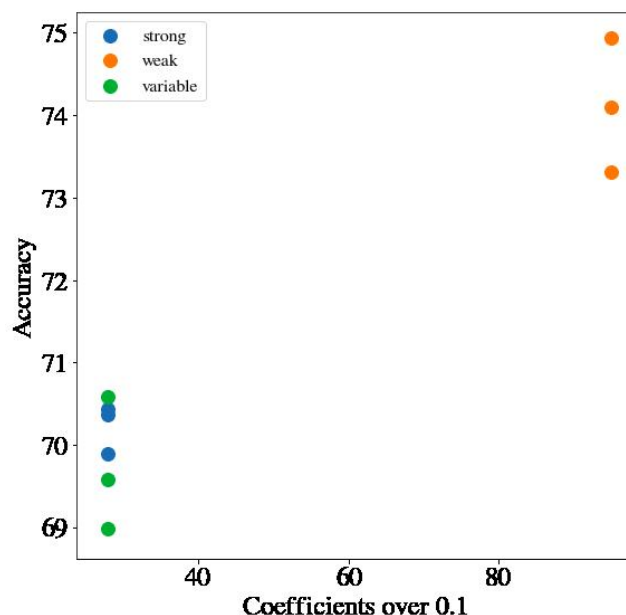


Рис. 2.

На рис. 3 показана таблица с усредненным по 3 запускам в каждом режиме top-1 и top-5 accuracy.

	experiment	top1 accuracy	top5 accuracy
0	strong	70.233330	98.109996
1	weak	74.113330	98.629997
2	variable	69.719999	97.986664

Рис. 3.

Рисунок 4 отражает поведение в сети в зависимости от присутствия в данных случайного шума, используется нормальное распределение с нулевым математическим ожиданием. График показывает что сильная и меняющаяся регуляризации оказываются более устойчивыми к зашумлению данных

На рис. 5 также изображено поведение модели в условиях наложение случайного шума на параметры модели.

Литература

- [1] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29,*

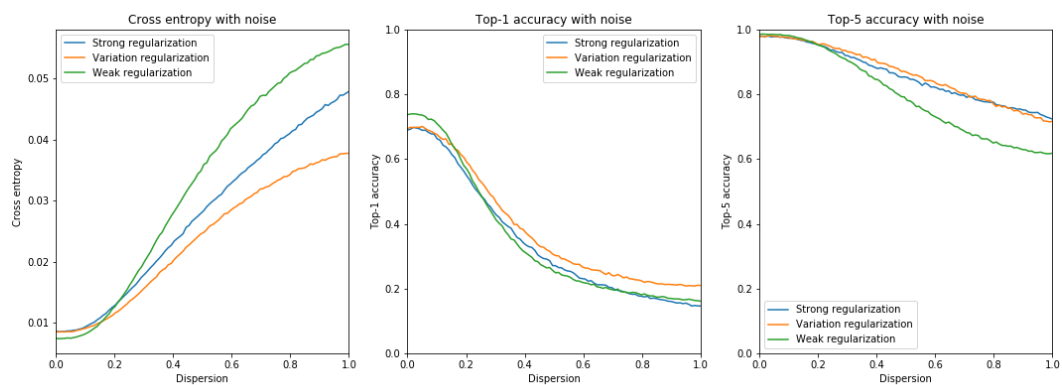


Рис. 4.

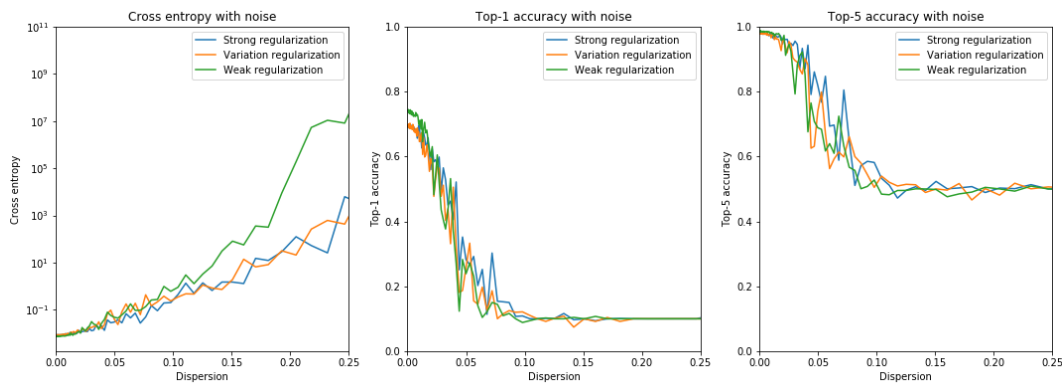


Рис. 5.

- 2014, Doha, Qatar, *A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751, 2014.
- [2] Grunwald P. A tutorial introduction to the minimum description length principle. 2005.
 - [3] Yann Le Cun, John S. Denker, and Sara A. Solla. *Advances in neural information processing systems 2. chapter Optimal Brain Damage*, pages 598–605. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.
 - [4] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, Cambridge, MA, USA, 2014. MIT Press.
 - [5] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *CoRR*, abs/1806.09055, 2018.
 - [6] Yang .Y Hanxiao L., Simonyan K. Darts: Differentiable architecture search. 2018.
 - [7] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2113–2122, Lille, France, 07–09 Jul 2015. PMLR.
 - [8] Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1165–1173, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.
 - [9] Fabian Pedregosa. Hyperparameter optimization with approximate gradient. pages 737–746, 2016.
 - [10] G. Hinton A. Krizhevsky, V. Nair. The cifar-10 dataset. 2009.
 - [11] Tommi S. Jaakkol Harald Steck. On the dirichlet prior and bayesian regularization.