

Автоматическое построение нейросети оптимальной сложности

Маркин Валерий, Забазнов Антон, Горян Николай,
Сергей Губанов, Сергей Таранов, Товкес Артём, Улитин
Александр, Криницкий Константин

Московский физико-технический институт

10 декабря, 2018г.

Иследуется

Задача выбора структуры нейронной сети.

Требуется

Найти нейросеть оптимальной сложности.

Проблемы

- Большое количество параметров,
- Высокая вычислительная сложность оптимизации,
- Невозможность использования эвристических и переборных алгоритмов выбора структуры модели

- *Yang .Y Hanxiao L., Simonyan K.*
Darts: Differentiable architecture search. 2018.
- *Dougal Maclaurin, David Duvenaud, Ryan P. 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.
- *Tommi S. Jaakkol Harald Steck* On the dirichlet prior and bayesian regularization.

Постановка задачи

Рассматриваем задачу классификации.

Пусть заданы обучающая и валидационная выборки:

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

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

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

$y \in \mathbf{Y} = \{1, \dots, Z\}$, где Z - количество классов.

Модель задаётся ориентированным графом $\mathbf{G} = (\mathbf{V}, \mathbf{E})$

$\mathbf{g}^{i,j}$ — базовые функции ребра (i, j) с весами $\gamma^{i,j}$

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:

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

Постановка задачи

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

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

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

где \mathbf{A} и λ — регуляризационные слагаемые, $p(\Gamma)$ — произведение всех произведение вероятностей всех $\gamma^{i,j} \in \Gamma$.

Требуется построить модель классификации \mathbf{f} с параметрами \mathbf{W} , доставляющую минимум функции потерь на валидации Q .

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

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

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 $\mathbf{g}^{(i,j)} = \arg \max_{\mathbf{g} \in \mathbb{K}} \gamma_{\mathbf{g}}^{(i,j)}$.

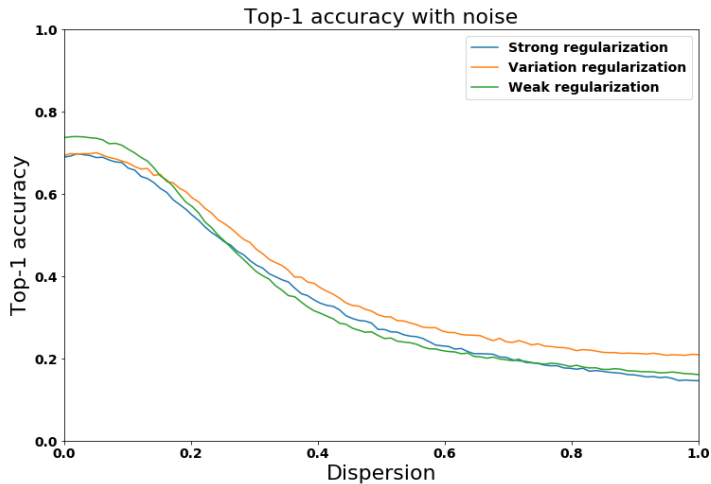
В качестве выборки использовалась выборка изображений CIFAR-10. Основным критерием качества выступал

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

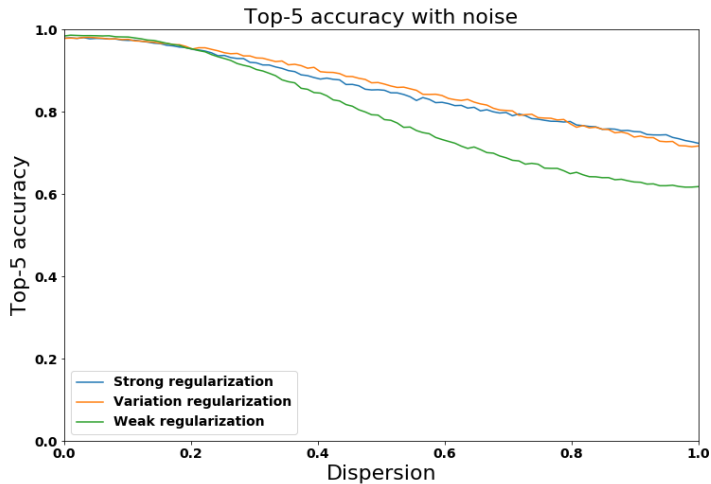
В качестве регуляризации использовались:

- 1 Слабая регуляризация (Дирихле с параметром $\alpha = 1$).
- 2 Меняющейся регуляризация (Дирихле с параметром α , меняющимся на каждой эпохи в интервале $(10^{-30}; 1)$).
- 3 Сильной регуляризацией (распределение Дирихле с параметром $\alpha = 10^{-30}$).

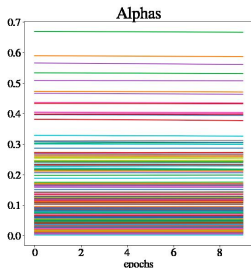
Устойчивость модели к шуму в данных



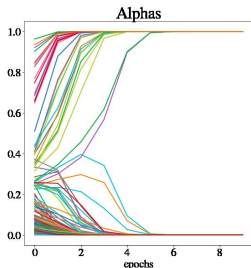
Устойчивость модели к шуму в данных



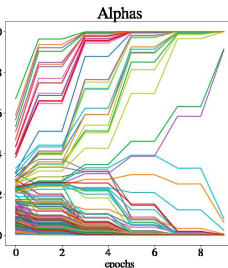
Динамика изменения структурных параметров в процессе обучения



слабая
регуляризация



сильная
регуляризация



вариационная
регуляризация

- Проведен эксперимент по регуляризации структуры нейронной сети на выборке CIFAR-10
- Эксперименты показали, что структура нейронной сети вырождается в дискретную при использовании регуляризации
- Эксперимент показал, что оптимизация подвержена застреванию в стационарных точках пространства структурных параметров

Планируется

- Предложить метод регуляризации, позволяющий избежать застревания в стационарных точках