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.

Цель работы

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 струкутры модели

Литература

- *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.
- *LeCun Y., Denker J. , Solla S.*
  Optimal Brain Damage // Advances in Neural Information Processing Systems, 1989. Vol. 2. P. 598–605.

Постановка задачи

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

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

$y \in \mathbf{Y} = \{1, \ldots, Z\}$, 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 $\mathbf{W} \in \mathbb{R}^{\mathbf{n}}$:

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

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

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

.

Постановка задачи

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

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

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

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

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

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

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

Идея для базовых алгоритмов

Рассматриваем задачу классификации.
Пусть заданы обучающая и вылидационная выборки:

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

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

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 $\gamma^{i,j}$.
Модель $\mathbf{f}$ с параметрами $\mathbf{W} \in \mathbb{R}^{\mathbf{n}}$:

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

где $\mathbf{f_i}(\mathbf{x}, \mathbf{w_i})$ - подмодель:

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

Идея для базовых алгоритмов

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

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

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

Регуляризация структуры

Для регуляризация структуры модели к фунции потерь прибавляется специальное слагаемого $\lambda P(\Gamma)$, где $P(\Gamma)$ есть произведение всех вероятностей возникновения веса $\gamma_{i,j,k}$. Таким образом функция потерь принимает вид:

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

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

Таким образом наша задача принимает такой вид:

$$\mathbf{\Gamma}, \mathbf{A} = argmin_{\mathbf{\Gamma},\mathbf{A}} Q(\mathfrak{D}^{valid}, \mathbf{W}^*(\mathbf{\Gamma}, \mathbf{A}), \mathbf{\Gamma}, \mathbf{A}),$$

$$\mathbf{W}^* = argmin_W L(\mathfrak{D}^{train}, \mathbf{W}, \mathbf{\Gamma}, \mathbf{A})$$

Описание эксперимента

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

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

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 $\alpha = 10^{-30}$).

Вывод

- Проведен эксперимент по регялиразции 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 избегать застревания в стационарных точках