Регуляризация траектории параметров модели глубокого обучения на основе дистилляции знаний

М. Горпинич, О. Ю. Бахтеев, В. В. Стрижов
gorpinich4@gmail.com; bakhteev@phystech.edu; strijov@ccas.ru

Исследуется задача оптимизации параметров модели глубокого обучения. Во время оптимизации также учитывается информация, содержащаяся в модели с более сложной структурой, то есть применяется дистилляция. Предлагается обобщение методов дистилляции, заключающееся в градиентной оптимизации метапараметров. Под метапараметрами модели понимаются параметры оптимизационной задачи дистилляции, а именно, коэффициенты перед слагаемыми в функции ошибки и температура. Функция ошибки состоит из двух слагаемых: правдоподобия исходной выборки и правдоподобия выборки дистилляции. Температурой является коэффициент, на который домножаются логиты моделей при применении функции softmax. Исследуются свойства оптимизационной задачи и методы предсказания траектории оптимизации метапараметров модели. Данное обобщение дистиллирует модель с лучшими эксплуатационными характеристиками и за меньшее число итераций оптимизации. Проиллюстрирован данный подход с помощью вычислительного эксперимента на выборке CIFAR-10 и на синтетической выборке.

Ключевые слова: *машинное обучение; дистилляция знаний; оптимизация метапараметров*

DOI:

1 Введение

В работе исследуется проблема оптимизации моделей глубоких нейросетей. Данная оптимизация требует значительных вычислительных мощностей и является затратной по времени. В данной работе предлагается метод оптимизации, позволяющий улучшить точность предсказаний модели, а также ускорить сходимость траектории оптимизации параметров к точке оптимума.

Предлагается обобщение метода оптимизации на основе дистилляции знаний. Назовем *дистилляцией знаний* задачу оптимизации параметров модели прогнозирования, при которой учитывается не только информация, содержащаяся в выборке, но также и информация, содержащаяся в сторонней модели (модели-учителе). Рассматривается *модель-учитель* более сложной структуры, которая была обучена на выборке. Модель более простой структуры предлагается оптимизировать путем переноса знаний модели учителя на более простую модель, называемую *моделью-учеником*. При этом ее качество будет выше по сравнению с качеством, полученным после оптимизации на той же выборке. Данный подход описан в [1]. В [2] предложен подход к дистилляции знаний, переносящий знания на модель с архитектурой, значительно отличающейся от архитектуры модели-учителя.

Предлагается формулировка задачи в виде двухуровневой оптимизации. На первом уровне оптимизируются параметры модели, на втором уровне — ее метапараметры. Данный подход описан в [3–5]. В [3] рассматривается жадный градиентный метод оптимизации метапараметров. В [4] сравниваются различные градиентные методы оптимизации метапараметров, а также метод случайного поиска.

В работе рассматривается подход к прогнозированию значений метапараметров, полученных методом градиентной оптимизации. Под метапараметрами понимаются параметры задачи оптимизации. Сложность градиентной оптимизации для метапараметров является квадратичной по числу параметров, и потому вычислительно затратна. Предлагается аппроксимация траектории оптимизации метапараметров на основе приближения траектории линейной моделью. Вычислительный эксперимент проводится на выборке изображений CIFAR-10 [6], а также синтетической выборке.

2 Постановка задачи

Решается задача классификации вида:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m,\ \mathbf{x}_i \in \mathbb{R}^n,\ y_i \in \mathbb{Y} = \{1, \dots, K\}, \tag{1}$$

где y_i — это класс объекта, также будем обозначать $\mathbf{y}_i$ вектором вероятности для класса y_i.

Разобьем выборку следующим образом:

$$\mathfrak{D} = \mathfrak{D}_{\text{train}} \sqcup \mathfrak{D}_{\text{val}}. \tag{2}$$

Подвыборку $\mathfrak{D}_{\text{train}}$ будем использовать для оптимизации параметров модели, а подвыборку $\mathfrak{D}_{\text{val}}$ — для оптимизации метапараметров.

Внешним критерием качества назначена доля правильных ответов:

$$\text{accuracy} = \frac{1}{m}\sum_{i=1}^{m}[\mathbf{g}(\mathbf{x}_i, \mathbf{w}) = y_i], \tag{3}$$

где $\mathbf{g}$ — параметрическая модель классификации с параметрами $\mathbf{w}$.

Определение 1. Назовем *дистилляцией знаний* задачу оптимизации параметров модели прогнозирования, при которой учитывается не только информация, содержащаяся в выборке, но также и информация, содержащаяся в сторонней модели (модели-учителе).

Зафиксирована модель учителя $\mathbf{f}$. Функция потерь $\mathcal{L}_{\text{train}}$, в которой учитывается перенос информации от модели учителя $\mathbf{f}$ к модели ученика $\mathbf{g}$, имеет вид:

$$\mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}) = -\lambda_1 \sum_{(\mathbf{x},y)\in\mathfrak{D}_{\text{train}}} \underbrace{\sum_{k=1}^{K} y^k \log \mathbf{g}(\mathbf{x}, \mathbf{w})|_{T=1}}_{\text{исходная функция потерь}} - \lambda_2 \sum_{(\mathbf{x},y)\in\mathfrak{D}_{\text{train}}} \underbrace{\sum_{k=1}^{K} \mathbf{f}(\mathbf{x})|_{T=T_0} \log \mathbf{g}(\mathbf{x}, \mathbf{w})|_{T=T_0}}_{\text{слагаемое дистилляции}}, \tag{4}$$

где T — параметр температуры. Параметр температуры T имеет следующие свойства:

1) при $T \to 0$ получаем вектор, в котором один из классов имеет единичную вероятность;
2) при $T \to \infty$ получаем равновероятные классы.

Выражение $\cdot|_{T=t}$ обозначает, что параметр температуры T в предыдущей функции равняется t.

Зададим множество метапараметров $\boldsymbol{\lambda}$ как вектор, состоящий из температуры и коэффициента перед слагаемым дистилляции:

$$\boldsymbol{\lambda} = [\lambda_1, \lambda_2, T].$$

Итоговая оптимизационная задача:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^3} \mathcal{L}_{\text{val}}(\hat{\mathbf{w}}, \boldsymbol{\lambda}), \tag{5}$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^s} \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}), \tag{6}$$

где функция $\mathcal{L}_{\text{val}}$ определяется как:

$$\mathcal{L}_{\text{val}}(\mathbf{w}, \boldsymbol{\lambda}) = \sum_{(\mathbf{x},y)\in\mathfrak{D}_{\text{val}}} \sum_{k=1}^{K} y^k \log \mathbf{g}(\mathbf{x}, \mathbf{w})|_{T=1}. \tag{7}$$

Определение 2. Назовем *оператором оптимизации* алгоритм U выбора вектора параметров $\mathbf{w}'$ по параметрам предыдущего шага $\mathbf{w}$:

$$\mathbf{w}' = U(\mathbf{w}).$$

Оптимизируем параметры $\mathbf{w}$ при помощи η шагов оптимизации:

$$\hat{\mathbf{w}} = U \circ U \circ \cdots \circ U(\mathbf{w}_0, \boldsymbol{\lambda}) = U^\eta(\mathbf{w}_0, \boldsymbol{\lambda}), \tag{8}$$

где $\mathbf{w}_0$ — начальное значение вектора параметров $\mathbf{w}$, $\boldsymbol{\lambda}$ — совокупность метапараметров модели.

Переопределим задачу минимизации согласно определению оператора U:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^3} \mathcal{L}_{\text{val}}\big(U^\eta(\mathbf{w}_0, \boldsymbol{\lambda})\big). \tag{9}$$

Схема оптимизации метапараметров:

1. Для каждого $i = \overline{0, l}$, где l — число итераций, используемых для оптимизации метапараметров.
2. Решим задачу (9) и получим новое значение метапараметров $\boldsymbol{\lambda}'$.
3. Положим $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$.

3 Градиентные методы оптимизации

Оптимизационную задачу (5) и (6) решает оператор градиентного спуска:

$$U(\mathbf{w}, \boldsymbol{\lambda}) = \mathbf{w} - \gamma\nabla\mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}), \tag{10}$$

где γ — длина шага градиентного спуска.

Используем метод градиентного спуска, который зависит только от значений параметров $\mathbf{w}$ на предыдущем шаге. На каждой итерации получим следующее значение метапараметров:

$$\boldsymbol{\lambda}' = \boldsymbol{\lambda} - \gamma_{\boldsymbol{\lambda}}\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\text{val}}(U(\mathbf{w}, \boldsymbol{\lambda}), \boldsymbol{\lambda}) = \boldsymbol{\lambda} - \gamma_{\boldsymbol{\lambda}}\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\text{val}}(\mathbf{w} - \gamma\nabla\mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}), \boldsymbol{\lambda}). \tag{11}$$

Градиентная оптимизация является вычислительно затратной, поэтому предлагается аппроксимировать траекторию оптимизации модели.

Предлагается предсказывать траекторию изменения метапараметров модели (а конкретно, их градиенты) с помощью линейных сплайнов через определенное число итераций, а в остальное время использовать градиентные методы.

4 Вычислительный эксперимент

Целью эксперимента является проверка работоспособности предложенного метода дистилляции моделей, а также анализ полученных моделей и их метапараметров. Эксперимент проводится на двух выборках: синтетической модели и выборке CIFAR-10. Результаты данной работы и исходный код эксперимента опубликованы в [7] и могут быть проверены или использованы в дальнейшей работе.

4.1 Эксперимент на синтетической выборке

В эксперименте используется синтетическая выборка:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m},\ x_{ij} \in (0,1),\ j = 1, 2, x_{i3} = [\mathrm{sign}(x_{i1}) + \mathrm{sign}(x_{i2}) > 0]$$

$$y_i = \mathrm{sign}(x_{i1} * x_{i2} + \delta) \in \mathbb{Y},$$

где δ — это шум. При этом размер выборки модели-ученика намного меньше размера выборки модели-учителя.

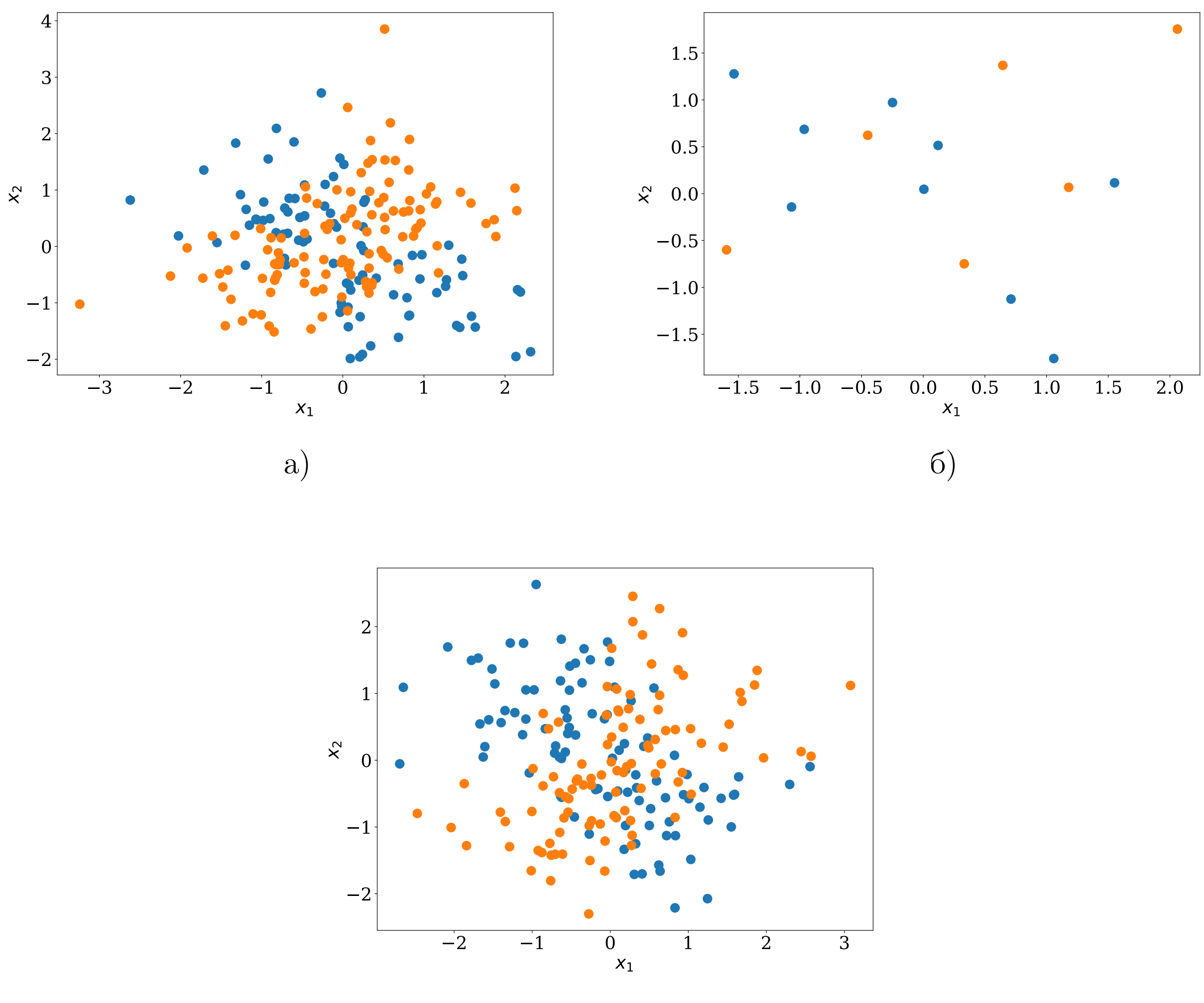

Рис. 1 Визуализация выборки а) для обучения учителя; б) для обучения ученика; в) тестовой выборки

Обучение модели-ученика проводилось несколькими методами: с использованием дистилляции и оптимизации метапараметров градиентными методами, дистилляции с предсказанием траектории оптимизации модели, дистилляции со случайными метапараметрами. При этом для обучения модели с использованием сплайнов дополнительно проводились серии экспериментов для определения наилучшего размера эпохи и наилучшего числа эпох между предсказаниями траектории с помощью сплайнов.

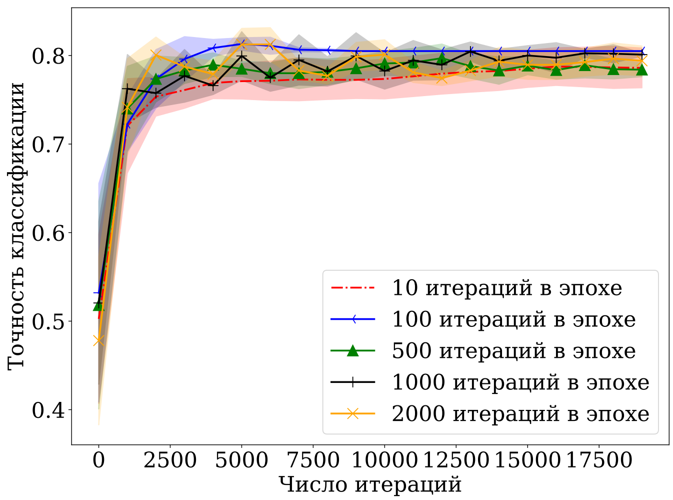

Рис. 2 График зависимости точности классификации от номера итерации при различных значениях размера эпохи

На рис. 2 показан график зависимости точности от номера итерации при различных размерах эпохи. Согласно данному графику размер эпохи был выбран равным 100.

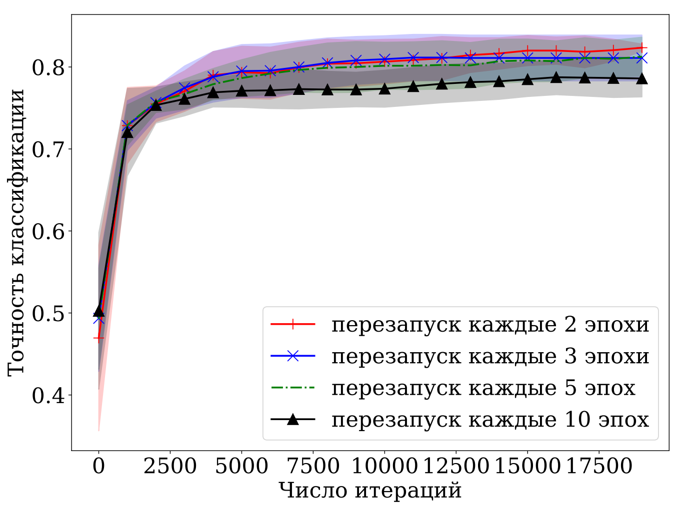

Рис. 3 График зависимости точности классификации от номера итерации при различных n

Пусть n — число эпох между использованием сплайнов. На рис. 3 показан график зависимости точности от номера итерации различных n. Наилучшие результаты достигнуты при $n = 2$.

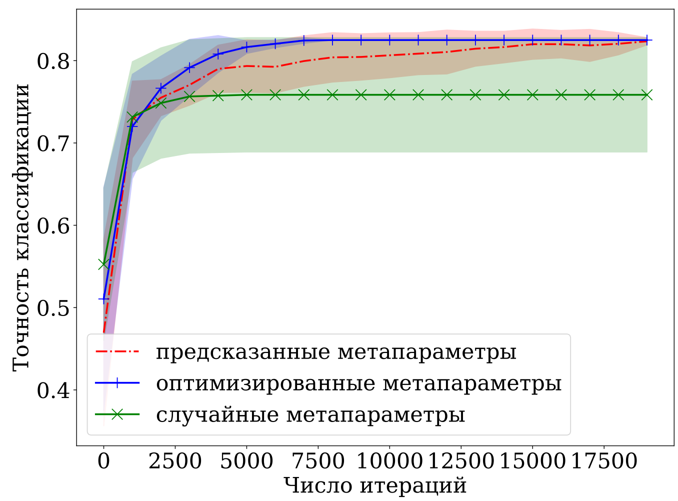

Рис. 4 График зависимости точности классификации от номера итерации

На рис. 9 показан график зависимости точности от номера итерации при различных подходах к обучению модели. Наилучшие результаты достигнуты при использовании оптимизированных метапараметров, но предсказание траектории с помощью сплайнов показало результат не намного хуже предыдущего, причем с увеличением числа итераций точность этих двух методов становилась одинаковой.

4.2 Эксперимент на выборке CIFAR-10

В эксперименте используется выборка CIFAR-10, которая состоит из 60000 цветных изображений размера 32×32 пикселя, разделенных на 10 непересекающихся классов. К каждому классу относится 6000 изображений. Выборка делится на обучающую (50000 изображений) и тестовую (10000 изображений) подвыборки. В тестовой выборке содержится 1000 изображений каждого класса.

Внешним критерием качества модели является точность (3). В качестве моделей-учителей рассматриваются модели из [2], а именно, ResNet-18 и сверточная нейросеть с тремя сверточными слоями и двумя слоями полносвязной нейросети.

Проведено сравнение среднего качества обучения модели-ученика без дистилляции после 5 запусков, с дистилляцией с моделью-учителем ResNet и сверточной нейросетью после 20 запусков. Значение коэффициента λ лежит в пределах от 0 до 1, значение температуры — от 0.1 до 10.

На рис. 6 изображена зависимость точности от величины коэффициента λ. Различные точки отвечают за точность модели без дистилляции, с дистилляцией ResNet и CNN. Можно заметить, что с уменьшением значения коэффициента λ значение точности увеличивается.

На рис. 7 изображена зависимость точности от T. Для изображения значений температуры используется логарифмическая шкала. По графику видно, что значение температу-

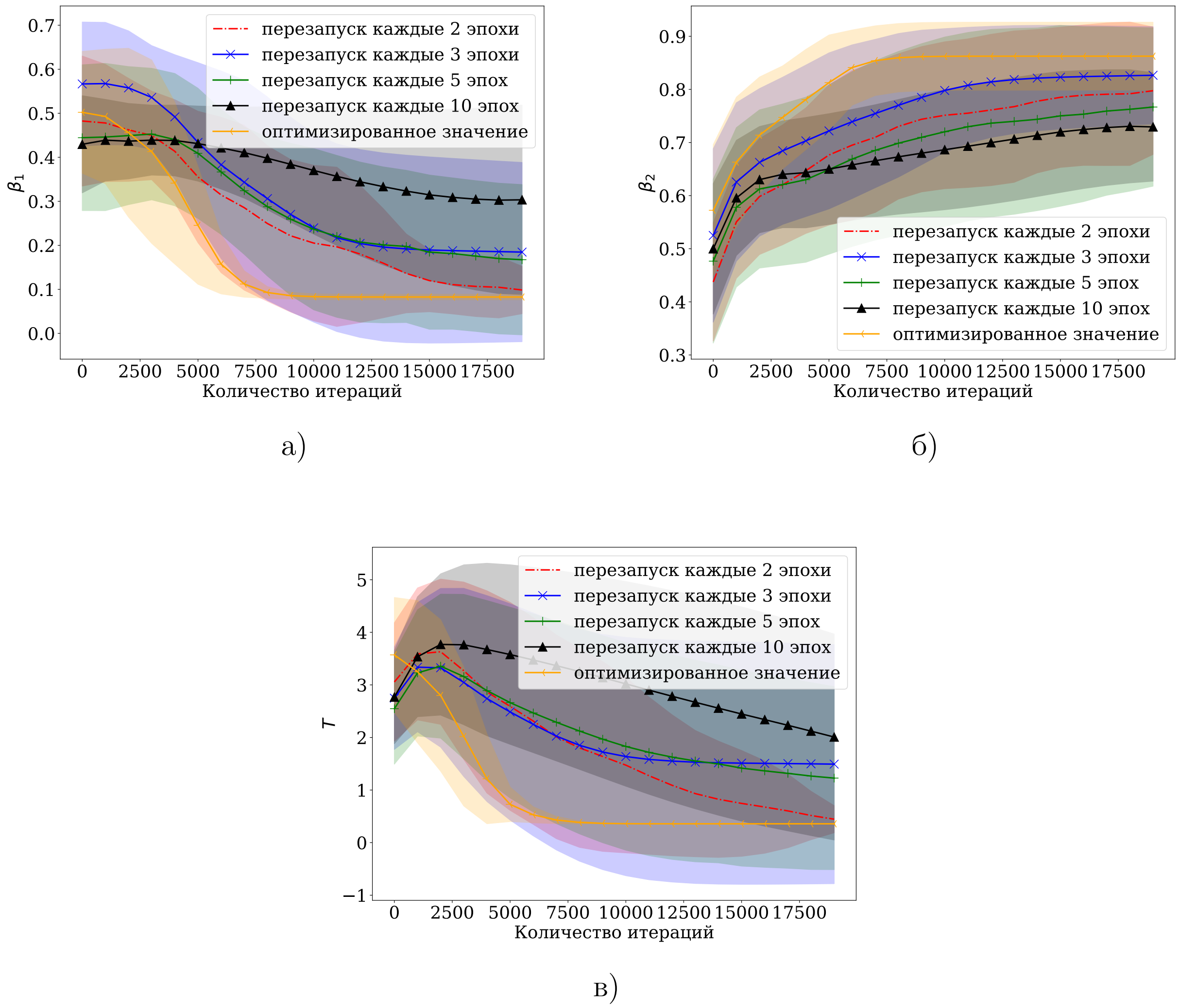

Рис. 5 График зависимости а) λ_1; б) λ_2; в) температуры от номера итерации

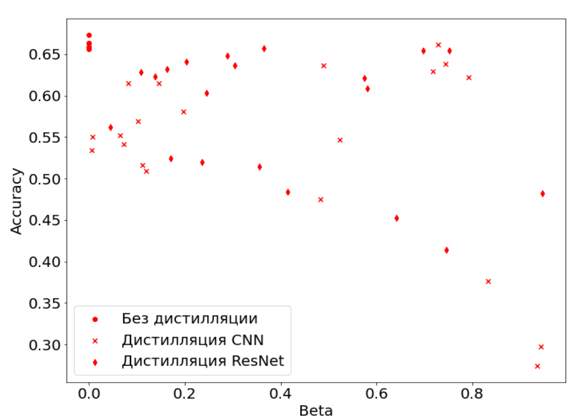

Рис. 6 График зависимости точности от λ

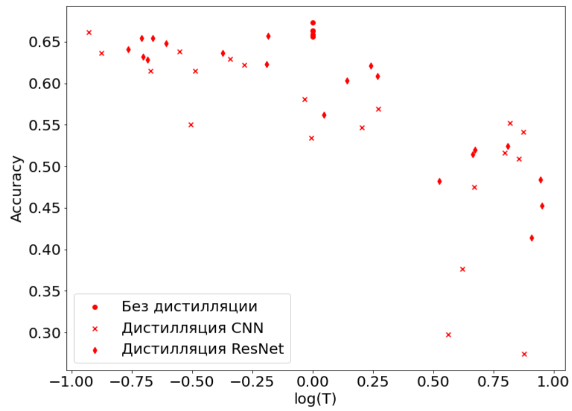

Рис. 7 График зависимости точности от температуры

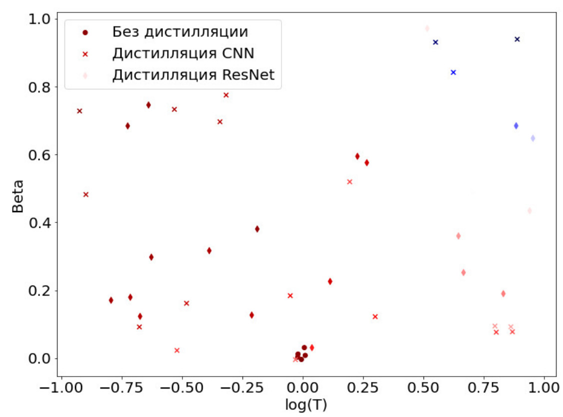

Рис. 8 График зависимости λ от температуры с выделенной цветом accuracy

ры уменьшается при увеличении логарифма температуры, но при значениях логарифма от 0.5 до 1 наблюдается резкое уменьшение точности.

На рис. 8 изображена зависимость λ от величины T с выделенной цветом accuracy. Заметим, что точки с большим значением точности в основном расположены в правом нижнем углу графика, а именно, при значениях λ от 0 до 0.5 и значениях $\log(T)$ от -1 до 0. Наоборот, точки с низким значением точности, расположены в правом верхнем углу графика.

На рис. 9 изображена зависимость точности от числа эпох для обучения модели-ученика без дистилляции, обучения с дистилляцией и случайными метапараметрами, обучения с дистилляцией и оптимизацией метапараметров, а также обучения с дистилляцией и оптимальными метапараметрами, полученными в ходе их оптимизации. Можно заметить, что точность обучения с дистилляцией гораздо выше, чем без дистилляции. Также наибольшая точность достигается при обучении с дистилляцией и оптимальными метапараметрами.

В таблице 1 приведены результаты эксперимента.

На рис. 10 изображена зависимость λ от числа итераций дистилляции.

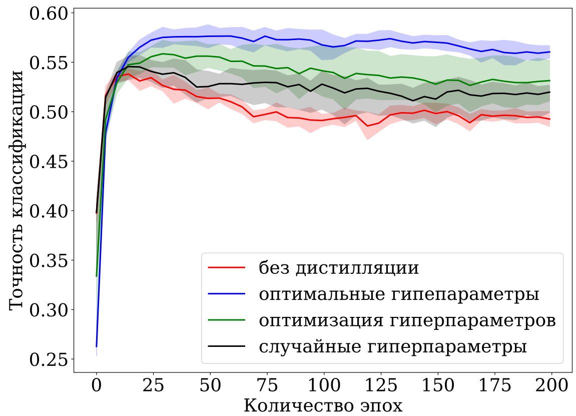

Рис. 9 График зависимости точности от числа эпох

Таблица 1 Результаты эксперимента

Рис. 10 График зависимости λ от числа итераций дистилляции

Рис. 11 График зависимости температуры от числа итераций дистилляции

На рис. 11 изображена зависимость T от числа итераций дистилляции.

5 Заключение

Была исследована задача оптимизации параметров модели глубокого обучения. Было предложено обобщение методов дистилляции, заключающееся в градиентной оптимизации метапараметров. На первом уровне оптимизируются параметры модели, на втором — метапараметры, задающие вид оптимизационной задачи. Были исследованы свойства оптимизационной задачи и методы предсказания траектории оптимизации метапараметров модели. Под метапараметрами модели понимаются параметры оптимизационной задачи дистилляции. Предложенное обобщение позволило производить дистилляцию модели с лучшими эксплуатационными характеристиками и за меньшее число итераций оптимизации. Комбинация данных подходов была проиллюстрирована с помощью вычислительного эксперимента на выборке CIFAR-10 и на синтетической выборке. Вычислительный эксперимент показал эффективность градиентной оптимизации для задачи выбора метарапараметров дистилляционной функции потерь. Проанализирована возможность аппроксимировать траекторию оптимизации метапараметров локально-линейной моделью. Планируется дальнейшее исследование оптимизационной задачи и анализ качества аппроксимации траектории оптимизации метапараметров более сложными прогностическими моделями.

Литература

[1] *Hinton Geoffrey E., Vinyals Oriol, Dean Jeffrey.* Distilling the knowledge in a neural network // CoRR, 2015. Vol. abs/1503.02531. URL: `http://arxiv.org/abs/1503.02531`.

[2] *Passalis Nikolaos, Tzelepi Maria, Tefas Anastasios.* Heterogeneous knowledge distillation using information flow modeling // CVPR. — 2020. P. 2336–2345. URL: `https://ieeexplore.ieee.org/xpl/conhome/9142308/proceeding`.

[3] *Luketina Jelena, Berglund Mathias, Greff Klaus, Raiko Tapani.* Scalable gradient-based tuning of continuous regularization hyperparameters // CoRR, 2015. Vol. abs/1511.06727. URL: `http://arxiv.org/abs/1511.06727`.

[4] *Bakhteev Oleg Yu., Strijov Vadim V.* Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Ann. Oper. Res, 2020. Vol. 289. No. 1. P. 51–65.

[5] *Maclaurin Dougal, Duvenaud David, Adams Ryan P.* Gradient-based hyperparameter optimization through reversible learning // CoRR, 2015. Vol. abs/1502.03492. URL: `http://arxiv.org/abs/1502.03492`.

[6] *Krizhevsky Alex et al.* Learning multiple layers of features from tiny images, 2009.

[7] URL: `https://github.com/Intelligent-Systems-Phystech/2021-Project-84`.

Received

Regularizing optimization trajectory of deep learning model parameters with knowledge distillation

M. Gorpinich, O. Yu. Bakhteev, V. V. Strijov

`gorpinich4@gmail.com; bakhteev@phystech.edu; strijov@ccas.ru`

The paper investigates parameter optimization problem for deep learning neural networks. The knowledge of a cumbersome model is considered during optimization, i.e. the knowledge distillation is used. The paper proposes generalization of knowledge distillation method to optimize meta-parameters by gradient descent. Meta-parameters are the parameters of knowledge distillation optimization problem, namely, the coefficients before terms in error function and the temperature factor. The error function is a sum of likelihood of the initial dataset and the one of distillation dataset. Temperature is a factor of logits of models in softmax function. The authors investigate the properties of optimization problem and methods to predict the optimization path of meta-parameters. Generalized method produces models with higher performance and uses less number of iterations. The algorithm is evaluated on CIFAR-10 dataset and synthetic data.

Keywords: *machine learning; knowledge distillation; metaparameter optimization*

DOI:

References

[1] *Hinton Geoffrey E., Vinyals Oriol, Dean Jeffrey.* Distilling the knowledge in a neural network // CoRR, 2015. Vol. abs/1503.02531. URL: `http://arxiv.org/abs/1503.02531`.

[2] *Passalis Nikolaos, Tzelepi Maria, Tefas Anastasios.* Heterogeneous knowledge distillation using information flow modeling // CVPR. — 2020. P. 2336–2345. URL: `https://ieeexplore.ieee.org/xpl/conhome/9142308/proceeding`.

[3] *Luketina Jelena, Berglund Mathias, Greff Klaus, Raiko Tapani.* Scalable gradient-based tuning of continuous regularization hyperparameters // CoRR, 2015. Vol. abs/1511.06727. URL: `http://arxiv.org/abs/1511.06727`.

[4] *Bakhteev Oleg Yu., Strijov Vadim V.* Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Ann. Oper. Res, 2020. Vol. 289. No. 1. P. 51–65.

[5] *Maclaurin Dougal, Duvenaud David, Adams Ryan P.* Gradient-based hyperparameter optimization through reversible learning // CoRR, 2015. Vol. abs/1502.03492. URL: http://arxiv.org/abs/1502.03492.

[6] *Krizhevsky Alex et al.* Learning multiple layers of features from tiny images, 2009.

[7] URL: https://github.com/Intelligent-Systems-Phystech/2021-Project-84.

Received