

# Регуляризация траектории параметров модели глубокого обучения на основе дистилляции

Мария Горпинич

Московский физико-технический институт

*Курс: Автоматизация научных исследований  
(практика, В. В. Стрижов)/Группа 874*

*Эксперт: В. В. Стрижов*

*Консультант: О. Ю. Бахтеев*

2021

# Задача дистилляции знаний

## Цель

Предложить метод оптимизации метапараметров в задаче обучения с применением дистилляции знаний.

## Исследуемая проблема

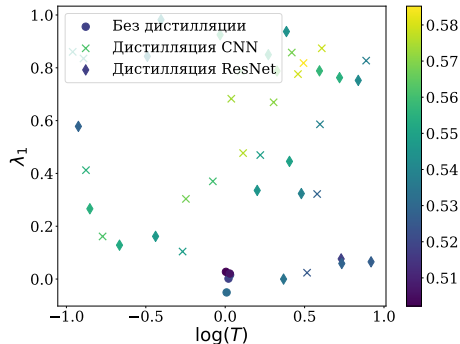
Назначение метапараметров задачи дистилляции является вычислительно сложной задачей.

## Решение

Предлагается рассмотреть задачу как двухуровневую задачу оптимизации. Решение задачи оптимизации метапараметров производится градиентными методами. Для ускорения вычислительно затратной процедуры оптимизации метапараметров производится прогнозирование локально-линейными моделями.




# Оптимизация параметров модели на основе дистилляции знаний

Назовем **дистилляцией знаний** задачу оптимизации параметров модели, при которой учитывается информация, содержащаяся в выборке и в сторонней модели (модели-учителе).



$T$  — температура,  $\lambda_1$  — доля правдоподобия исходной выборки в функции потерь

## Основная литература

-  Geoffrey E. Hinton, Oriol Vinyals и Jeffrey Dean. “Distilling the Knowledge in a Neural Network”. в: *CoRR* abs/1503.02531 (2015). URL: <http://arxiv.org/abs/1503.02531>.
-  Jelena Luketina и др. “Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters”. в: *CoRR* abs/1511.06727 (2015). URL: <http://arxiv.org/abs/1511.06727>.
-  Oleg Yu. Bakhteev и Vadim V. Strijov. “Comprehensive analysis of gradient-based hyperparameter optimization algorithms”. в: *Ann. Oper. Res* 289.1 (2020), с. 51—65.
-  Marcin Andrychowicz и др. “Learning to learn by gradient descent by gradient descent”. в: *CoRR* abs/1606.04474 (2016). URL: <http://arxiv.org/abs/1606.04474>.

# Постановка задачи дистилляции

## Задана выборка

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \quad \mathbf{x}_i \in \mathbb{R}^n, \quad y_i \in \mathbb{Y} = \{1, \dots, K\}, \quad \mathcal{D} = \mathcal{D}_{\text{train}} \sqcup \mathcal{D}_{\text{val}}.$$

$\mathbf{f}$  — зафиксированная модель учителя,  $\mathbf{g}$  — модель ученика.

## Функция ошибки дистилляции

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

# Алгоритм решения оптимизационной задачи

Множество метапараметров:

$$\lambda = [\lambda_1, \lambda_2, T].$$

Оптимизационная задача:

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

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

Оптимизационную задачу решает оператор градиентного спуска:

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

Обновим метапараметры последовательно согласно правилу:

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

**Гипотеза:** траектория градиентной оптимизации аппроксимируется локально-линейной моделью:

$$\lambda' = \lambda + \mathbf{c}^{\top} \begin{pmatrix} z \\ 1 \end{pmatrix},$$

где  $\sigma$  — сигмоида,  $z$  — номер итерации по модулю периодичности обучения линейной модели,  $\mathbf{c}$  — коэффициенты линейного многочлена.

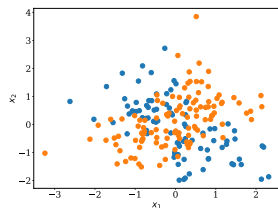
# Эксперимент на синтетической выборке

Целью вычислительного эксперимента является анализ градиентной оптимизации и проверка гипотезы об аппроксимации траектории оптимизации локально-линейной моделью.

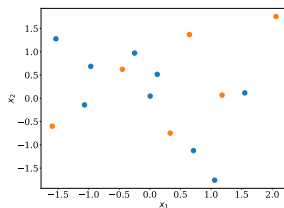
## Выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \quad x_{ij} \in \mathcal{N}(0, 1), \quad j = 1, 2, \quad x_{i3} = [\text{sign}(x_{i1}) + \text{sign}(x_{i2}) > 0], \\ y_i = \text{sign}(x_{i1} \cdot x_{i2} + \delta) \in \mathbb{Y}.$$

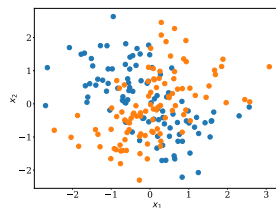
Размер выборки модели-ученика намного меньше размера выборки модели-учителя.



а)



б)

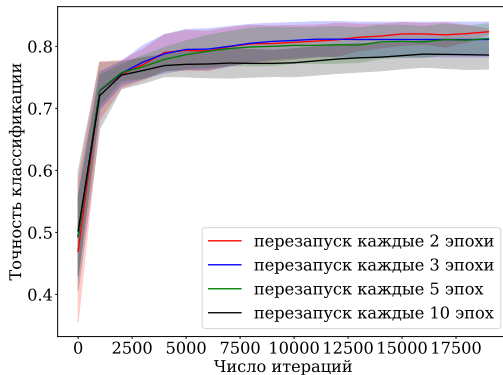


в)

Визуализация выборки а) учителя; б) ученика; в) тестовой

## Выбор периодичности перезапусков

График зависимости точности классификации от номера итерации при различном количестве перезапусков

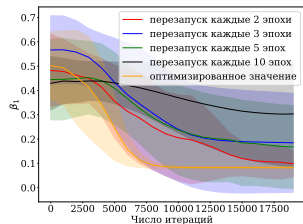


Наилучшее качество достигается при периодичности равной 2.

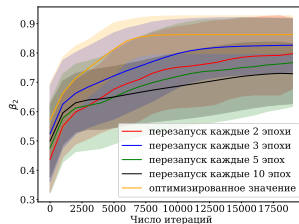


# Зависимость метапараметров от количества итераций

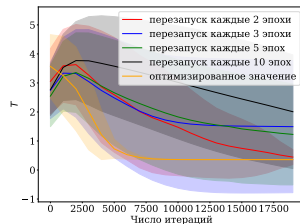
Графики зависимости значений метапараметров от номера итерации: а)  $\lambda_1$ ; б)  $\lambda_2$ ; в) температуры



а)



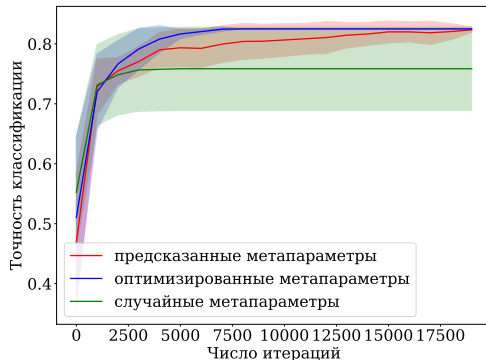
б)



в)

# Сравнение подходов к оптимизации

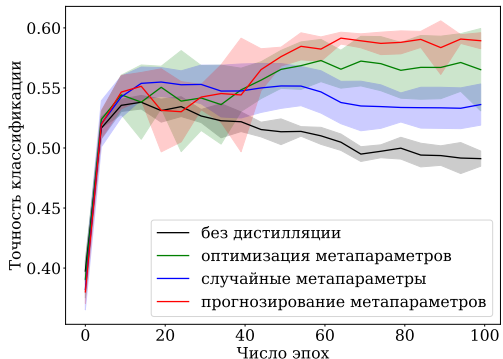
График зависимости точности классификации от номера итерации



Точность модели с предсказанными метапараметрами близка к точности модели с оптимизированными метапараметрами.

# Результаты эксперимента на выборке CIFAR-10

График зависимости точности классификации от номера итерации



Точность модели с предсказанными метапараметрами выше точности модели с оптимизированными метапараметрами.

# Заключение

- ▶ Исследовано применение градиентных методов оптимизации для метапараметров задачи дистилляции.
- ▶ Предложена и проверена гипотеза по аппроксимации траектории оптимизации метапараметров.
- ▶ Вычислительный эксперимент показал, что оптимизация метапараметров применима к задаче дистилляции.
- ▶ Подтверждена возможность аппроксимации метапараметров локально-линейными моделями.
- ▶ Планируется дальнейшее исследование оптимизационной задачи и анализ качества аппроксимации траектории оптимизации метапараметров более сложными прогностическими моделями.