Вариационная оптимизация модели глубокого обучения с контролем сложности модели

О. С. Гребенькова, О. Ю. Бахтеев, В. В. Стрижов
grebenkova.os@phystech.edu; bakhteev@phystech.edu; strijov@ccas.ru

В работе исследуется задача построения модели глубокого обучения с возможностью задания ее сложности. Под сложностью модели понимается минимальная длина описания, минимальное количество информации, которое требуется для передачи информации о модели и о выборке. Предлагается метод оптимизации модели, основанный на представлении модели глубокого обучения в виде гиперсети с использованием байесовского подхода, где под гиперсетью понимается модель, которая генерирует параметры оптимальной модели. Вводятся вероятностные предположения о параметрах модели глубокого обучения. Предлагается алгоритм, максимизирующий нижнюю вариационную оценку байесовской обоснованности модели. Вариационная оценка рассматривается как условная величина, зависящая от требуемой сложности модели. Для анализа качества предлагаемого алгоритма проводятся эксперименты на выборке MNIST.

Ключевые слова: *вариационная оптимизация модели; гиперсети; глубокое обучение; нейронные сети; байесовский подход; заданная сложность модели*

1 Введение

В работе рассматривается задача оптимизации модели глубокого обучения с заранее заданной сложностью модели. Под сложностью модели понимается обоснованность модели. Под моделью глубокого обучения понимается суперпозиция дифференцируемых по парметру функций. Построение модели заданной сложности является одной из фундаментальных проблем глубокого обучения, так как по построению данное семейство моделей имеет избытычное число параметров [1].

Предлагаемый метод заключается в представлении модели глубокого обучения в виде гиперсети, с использованием байесовского подхода. Гиперсеть — модель, которая задаёт параметры модели. На вход такой модели подается информация о структуре модели, а в результате работы порождается вектор параметров для слоев входной модели. В работе [2] рассмотрены статистические и динамические гиперсети для генерации весов сверточных и рекурентных моделей соответственно.

Вводятся вероятностные предположения о параметрах моделей глубокого обучения. В работе [1] предлагается использовать в качестве функции ошибки минимальную длину описания, т.е. минимальное количество информации, которое требуется для передачи информации о модели и о выборке, для оптимизации параметров модели глубокого обучения. Также в работе [1] получены виды функций потерь ошибки и потерь сложности для апроксимации апостериорной вероятности в случае нормального и дельта распределений.

Альтернативным подходом к построению модели заданной сложности выступает задача порождения и выбора оптимальной структуры моделей глубокого обучения. В работе [3] рассматривается возможность порождения широкого класса свёрточных моделей как подмоделей обощенной модели, которая называется «фабрикой»(англ. fabric). Данные структуры позволяют обойти процесс оптимизации параметров и проверки качества одиночных моделей. В работах [4, 5] представлены подходы для решения задачи выбора структуры нейросети с использованием дифференцируемых алгоритмов — стохастическая (англ. Stochastic Neural Architecture Search — SNAS) и дифференцируемая (англ.

Differentiable Neural Architecture Search — DNAS) нейронные архитектуры. Особенностью работы [5] является решение задачи выбора архитектуры модели, удовлетворяющей эксплутационным требованиям: быстродействию на различных типах процессоров.

В данной работе исследуется поведение обобщенной функции обоснованности модели, исследуется влияние априорного распределения на сложность модели. Для контроля сложности предлагается рассматривать задачу оптимизации параметров гиперсети, позволяющей порождать модели наперед заданной сложности с меньшими вычислительными затратами, чем в случае оптимизации модели, получаемой напрямую. Работа схожа с работой [6], где также исследовалась возможность получения гиперсети для предсказания наилучших гиперпараметров оптимизации модели. Вычислительный эксперимент проводится на выборке рукописных цифр MNIST [7].

2 Постановка задачи

Задана выборка:

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

где $\mathbf{x}_i \in \mathbb{R}^m$, $\mathbf{y}_i \in \{1, \ldots, Y\}$, Y — число классов. Рассмотрим модель

$$\mathbf{f}(\mathbf{x}, \mathbf{w}) : \mathbb{R}^m \times \mathbb{R}^n \longrightarrow \mathbb{R}^Y,$$

где $\mathbf{w} \in \mathbb{R}^n$ — пространство параметров модели.

Пусть задано априорное распределение вектора параметров в пространстве $\mathbb{W}$:

$$p(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}),$$

где $\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}$ — вектор средних и матрица ковариации априорного распределения. Заметим, что $\mathbb{W}$ является носителем априорного распределения. Тогда

$$p(\mathbf{w}|\mathfrak{D}) = \frac{p(\mathfrak{D}|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D})}$$

— апостериорное распределение вектора параметров $\mathbf{w}$ при заданной выборке $\mathfrak{D}$.

Пусть также задано распределение:

$$q(\mathbf{w}) \sim \mathcal{N}(\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}).$$

Здесь $\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}$ — вектор средних и матрица ковариации, аппроксимирующеее неизвестное апостериорное распределение $p(\mathbf{w}|\mathfrak{D})$.

Для модели $\mathbf{f}$ и соответствующего ей вектора параметров $\mathbf{w}$ определим логарифмическую функцию правдобподобия выборки:

$$\mathcal{L}_{\mathfrak{D}}(\mathbf{w}|\mathfrak{D}) \propto \log p(\mathfrak{D}|\mathbf{w}). \tag{1}$$

Оптимальные значения $\mathbf{w}$ находятся из максимизации $\mathcal{L}(\mathfrak{D})$ — логарифма обоснованности модели:

$$\log p(\mathfrak{D}) = \log \int\limits_{\mathbf{w} \in \mathbb{W}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{2}$$

Так как вычисление интеграла (2) является вычислительно сложной задачей, рассмотрим вариационный подход к решению задачи.

Оценим интеграл (2):

$$\mathcal{L}(\mathfrak{D}) = \log p(\mathfrak{D}) = \int\limits_{\mathbf{w}\in\mathbb{W}} q(\mathbf{w}) \log \frac{p(\mathfrak{D},\mathbf{w})}{q(\mathbf{w})} d\mathbf{w} - \int\limits_{\mathbf{w}\in\mathbb{W}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathfrak{D})}{q(\mathbf{w})} d\mathbf{w} \geqslant$$

$$\geqslant \int\limits_{\mathbf{w}\in\mathbb{W}} q(\mathbf{w}) \log \frac{p(\mathfrak{D},\mathbf{w})}{q(\mathbf{w})} d\mathbf{w} = \int\limits_{\mathbf{w}\in\mathbb{W}} q(\mathbf{w}) \log \frac{p(\mathbf{w})}{q(\mathbf{w})} d\mathbf{w} + \int\limits_{\mathbf{w}\in\mathbb{W}} q(\mathbf{w}) \log p(\mathfrak{D}|\mathbf{w}) d\mathbf{w} =$$

$$= \mathcal{L}_{\mathbf{w}}(\mathfrak{D},\mathbf{w}) + \mathcal{L}_E(\mathfrak{D}). \tag{3}$$

Первое слагаемое формулы (3) — это сложность модели. Оно определяется расстоянием Кульбака – Лейблера, то есть расстоянием между распределением $q(\mathbf{w})$ и априорном распределением $p(\mathbf{w})$:

$$\mathcal{L}_{\mathbf{w}}(\mathfrak{D},\mathbf{w}) = -D_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w})).$$

Второе слагаемое формулы (3) представляет собой математическое ожидание правдоподобия выборки $\mathcal{L}_{\mathfrak{D}}(\mathbf{w}|\mathfrak{D})$:

$$\mathcal{L}_E = \mathbb{E}_{q(\mathbf{w})}\mathcal{L}_{\mathfrak{D}}(\mathbf{w}|\mathfrak{D})$$

Обоснованность — это один из показателей сложности модели [1]. Рассматривается задача получения модели по обобщенной функции обоснованности:

$$\lambda\mathcal{L}_{\mathbf{w}}(\mathfrak{D},\mathbf{w}) - \mathcal{L}_E(\mathfrak{D}),$$

где λ контролирует влияние априорного распределения на итоговую модель.

Введём множество допустимых значений параметра сложности $\mathit{\Lambda} \subset \mathbb{R}^+$. Требуется найти такое отображение $\mathfrak{G} : \mathit{\Lambda} \longrightarrow \mathbb{W}$, чтобы для произвольного значения параметра сложности $\lambda \in \mathit{\Lambda}$ параметры доставляли бы максимум следующему функционалу:

$$\mathfrak{G}(\lambda) = \arg\max_{w\in\mathbb{W}}(\log p(\mathfrak{D}|\mathbf{w}) - \lambda D_{KL}(q(\mathbf{w})||p(\mathbf{w}))). \tag{4}$$

3 Построение гиперсети для контроля сложности модели

Решение задачи оптимизации (4) для произвольного значения $\lambda \in \mathit{\Lambda}$ является вычислительно сложной задачей. В данной работе для её решения предлагается использование гиперсетей.

Пусть задано множество параметров, контролирующих сложность модели Λ. Гиперсеть — это параметрическое отображение из множества Λ во множество параметров модели $\mathbb{W}$:

$$\mathbf{G} : \Lambda \times \mathbb{U} \to \mathbb{W},$$

где $\mathbb{U}$ — множество допустимых параметров гиперсети.

В работе были рассмотрены следующие виды гиперсетей: с отображением во множество матриц низкого ранга и линейной аппроксимацией.

$$\mathbf{G}_{\mathrm{lowrank}}(\lambda) = (\mathbf{f}(\lambda)\mathbf{U}_1)(\mathbf{f}(\lambda)\mathbf{U}_2)^\top + \mathbf{b}_1, \tag{5}$$

где λ — случайное число, сэмплируемое для каждого батча при обучении, $\mathbf{f}$ — функция, переводящаяся λ в скрытый слой; $\mathbf{U}_1, \mathbf{U}_2$ — матрицы, переводящие из скрытого слоя в

нужную размерность, их конкатенация принадлежит пространству параметров гиперсети: $[\mathbf{U}_1, \mathbf{U}_2] = \mathbf{U} \in \mathbb{U}$; $\mathbf{b_1}$ — константа, не зависящая от λ.

$$\mathbf{G}_{\text{linear}}(\lambda) = \lambda \mathbf{b}_2 + \mathbf{b}_3, \tag{6}$$

где $\mathbf{b}_2, \mathbf{b}_3$ — константы, не зависящие от λ.

Для аппроксимации оптимизационной задачи (4) предлагается оптимизировать параметры гиперсети $\mathbf{U} \in \mathbb{U}$ по случайно порожденным значениям параметра сложности $\lambda \in \Lambda$:

$$\mathsf{E}_{\lambda \sim \mathcal{U}(\Lambda)}(\log p(\mathfrak{D}|\mathbf{w}) - \lambda D_{KL}(q(\mathbf{w})||p(\mathbf{w})) \to \max_{\mathbf{U} \in \mathbb{U}},$$

где $\mathcal{U}$ — равномерное распределение.

4 Вычислительный эксперимент

Для анализа свойств обобщенной задачи оптимизации и предложенного метода построения гиперсети был проведен вычислительный эксперимент на выборке рукописных цифр MNIST [7]. Производилось сравнение метода построения модели напрямую без использования гиперсети (4) и с использованием гиперсетей (5), (6). В качестве критерия качества модели использовалась точность классификации:

$$\text{Accuracy} = 1 - \frac{1}{m} \sum_{i=1}^{m} [f(x_i, w_i) \neq y_i],$$

где m — длина тестовой выборки. Для каждой модели производилось прореживание параметров с применением подхода, описанного в [1]. Критерием удаления параметров выступала относительная плотность модели:

$$\rho(w_i) \approx \exp \frac{\mu_i^2}{2\sigma_i^2}$$

.

Была рассмотрена нейросеть состоящая из двух слоёв: 100 и 10 нейронов соответственно, где второй слой отвечает за softmax-функцию.

$$\mathbf{f}(\mathbf{x}, \mathbf{w}) = \mathbf{softmax}(\mathbf{x}_1^\top \mathbf{w}_1 + \mathbf{b}_1)\mathbf{w}_2 + \mathbf{b}_2,$$

где $\mathbf{w}_1, \mathbf{b}_1$ — параметры первого слоя нейросети, $\mathbf{w_2}, \mathbf{b}_2$ — параметры второго слоя нейросети,

$$\mathbf{softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{k} \exp(x_j)} \quad i = 1, \ldots, k$$

Нейросеть запускалась для разных значений параметра сложности $\lambda \in \mathit{\Lambda}$. На рис. 1 показано как меняется точность классификации при удалении параметров указанным методом.

Из графика видно, что вариационный метод позволяет удалить $\approx 80\%$ параметров при всех λ, кроме значений 0 и 0.1, без значительной потери точности классификации. При дальнейшем удалении качество для всех значений снижается.

При больших значениях λ, $\lambda \in [100, 1000]$, получается переупрощенная модель, которая зависит от малого числа параметров. Таким образом, удаление параметров нейросети при данных значении λ слабо влияет на точность классификации, однако изначальная точность невысока или держится на уровне случайного угадывания ($\lambda = 1000$).

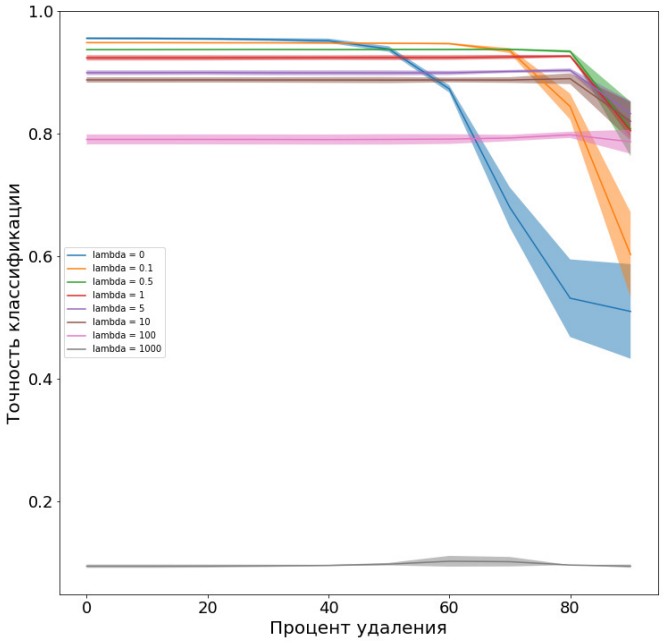

Рис. 1 График зависимости точности классификации от процента удалённых параметров

Далее была рассмотрена гиперсеть в двух реализациях: с отображением во множество матриц низкого ранга. и линейной аппроксимацией. Для обеих моделей использовался оптимизатор ADAM. Обучение проводилось на протяжении 50 эпох. В качестве параметра априорной дисперсии было выбрано значени 0.1.

Для первой модели (5) был рассмотрен случай с 50 нейронами в скрытом слое и с функций активации ReLU:

$$\mathrm{ReLU}(x) = \max(0, x).$$

При обучении второй модели (6) каждый батч проходил процесс оптимизации с 5 разными значениями сэмплируемой λ.

Прореживание нейросетей запускалось для разных значений параметра сложности $\lambda \in \Lambda$.

На рис. 2 показано как меняется точность классификации при удалении параметров указанным методом для модели с низкоранговой аппроксимацией. Как видно из графика, средняя точность классификации относительно базового экперимента понизилась на 20%. Также сильно увеличилось отклонение от среденего. При этом для всех значений $\lambda \in \Lambda$ получили более стабильную модель - точность классификации меньше зависит от удаления параметров. Можно наблюдать небольшую потерю точности при удалении более 80% . При больших значениях λ, $\lambda \in [100, 1000]$, получили результаты аналогичные базовому эксперименту. При значении $\lambda = 100$ модель показала результат схожий с небольшими значениями, $\lambda \in [0, 50]$.

На рис. 3 показано как меняется точность классификации при удалении параметров указанным методом для модели с линейной аппроксимацией. Линейная модель показала ещё более стабильные результаты относительно предыдущих экспериментов. При этом точность классификации остается постоянной и равной $\approx 85\%$. Отклонения от среднего

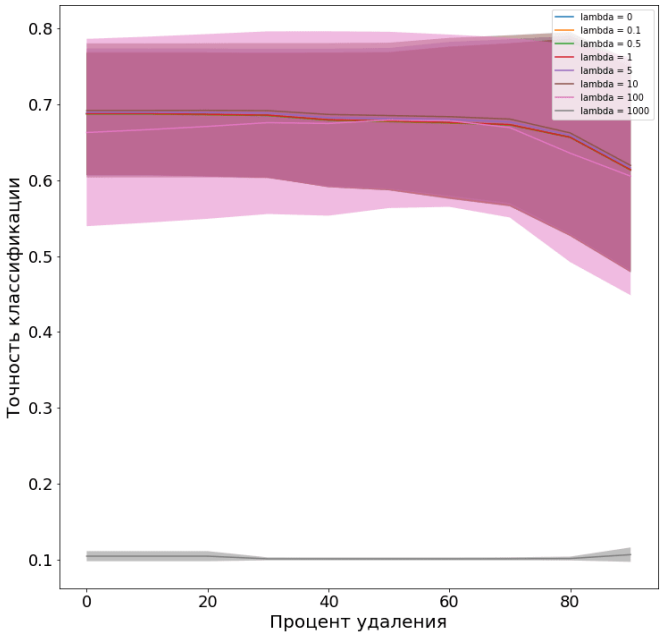

Рис. 2 График зависимости точности классификации от процента удалённых параметров для модели с низкоранговой аппроксимацией

незначительные для небольших значений λ, $\lambda \in [0, 50]$. При больших значениях λ, $\lambda \in [100, 1000]$, получили результаты аналогичные базовому эксперименту.

Далее данные модели были дообучены независимо от гиперсети в течении одной эпохи и эксперимент с прореживанием был запущен еще раз.

На рис. 4 показано как меняется точность у дообученной модели с низкоранговой аппроксимацией при удалении параметров. Как видно из графика, после обучения точность классификации увеличилась, уменьшилось отклонение от среднего. Однако понизилась стабильность модели и для небольших значений λ, $\lambda \in [0, 50]$, точность значительно падает при удалении более 60% параметров. Для значения $\lambda = 100$ модель показала значительные улучшения в точности классификации и большую стабильность относительно предыдущей версии модели. При больших значениях λ, $\lambda \in [100, 1000]$, λ получили результаты аналогичные базовому эксперименту.

На рис. 5 показано как меняется точность у дообученной модели с линейной аппроксимацией при удалении параметров. Как видно из графика, после обучения точность классификации увеличилась на $\approx 5\%$, незначительно увеличилось отклонение от среднего при удалении более чем 60% параметров. Понизилась стабильность модели, но она по-прежнему выше, чем в экспериментах с полноранговой моделью и моделью с низкоранговой аппроксимацией. Для значения $\lambda = 100$ модель показала улучшения в точности классификации. При больших значениях λ, $\lambda \in [100, 1000]$, λ получили результаты аналогичные базовому эксперименту.

Несмотря на потерю в качестве, гиперсеть позволяет получить схожие результаты, что и обычные модели при меньших вычислительных затратах. Более того, по графикам видно, что модель сохраняет схожие свойства при прореживании.

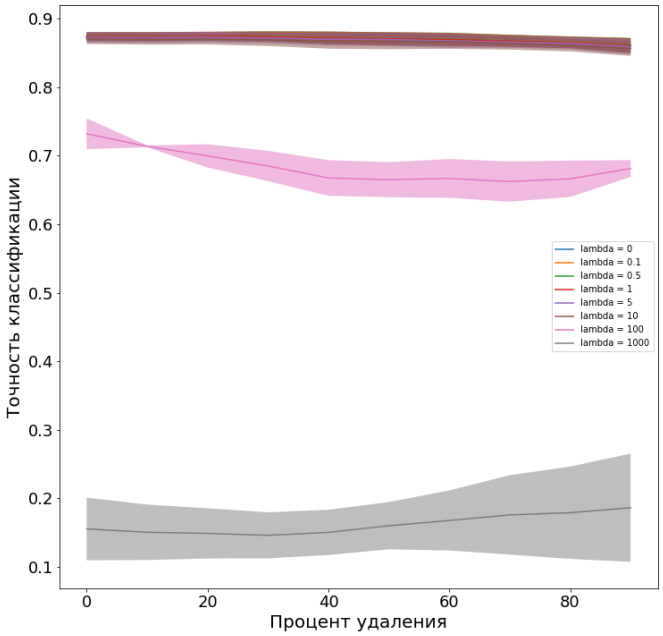

Рис. 3 График зависимости точности классификации от процента удалённых параметров для модели с линейной аппроксимацией

Литература

[1] *Graves Alex.* Practical variational inference for neural networks // Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain / Eds. John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, Kilian Q. Weinberger. — 2011. P. 2348–2356. URL: http://papers.nips.cc/book/advances-in-neural-information-processing-systems-24-2011.

[2] *Ha David, Dai Andrew M., Le Quoc V.* Hypernetworks // CoRR, 2016. Vol. abs/1609.09106. URL: http://arxiv.org/abs/1609.09106.

[3] *Saxena Shreyas, Verbeek Jakob.* Convolutional neural fabrics // CoRR, 2016. Vol. abs/1606.02492. URL: http://arxiv.org/abs/1606.02492.

[4] *Xie Sirui, Zheng Hehui, Liu Chunxiao, Lin Liang.* Snas: Stochastic neural architecture search // CoRR, 2018. Vol. abs/1812.09926. URL: http://arxiv.org/abs/1812.09926.

[5] *Wu Bichen, Dai Xiaoliang, Zhang Peizhao, Wang Yanghan, Sun Fei et al.* Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search // CoRR, 2018. Vol. abs/1812.03443. URL: http://arxiv.org/abs/1812.03443.

[6] *Lorraine Jonathan, Duvenaud David.* Stochastic hyperparameter optimization through hypernetworks // CoRR, 2018. Vol. abs/1802.09419. URL: http://arxiv.org/abs/1802.09419.

[7] *LeCun Yann, Cortes Corinna.* MNIST handwritten digit database, 2010. URL: http://yann.lecun.com/exdb/mnist/.

Received February 25, 2020

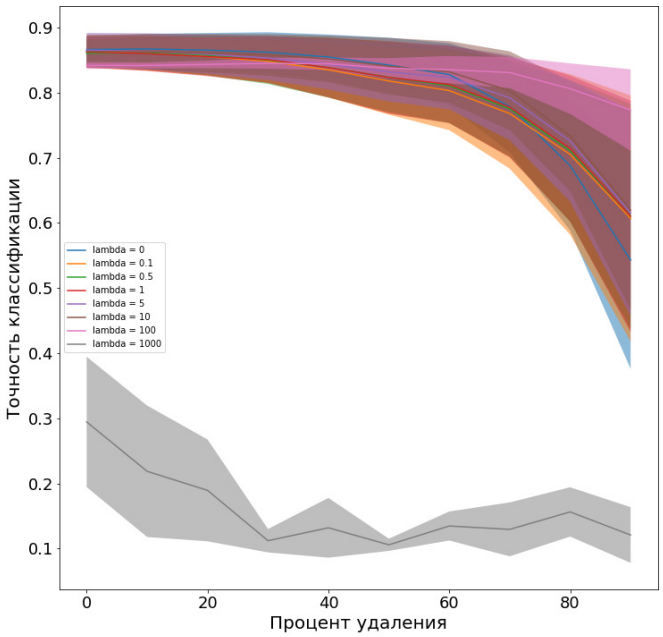

Рис. 4 График зависимости точности классификации от процента удалённых параметров для модели с низкоранговой аппроксимацией с дообучением

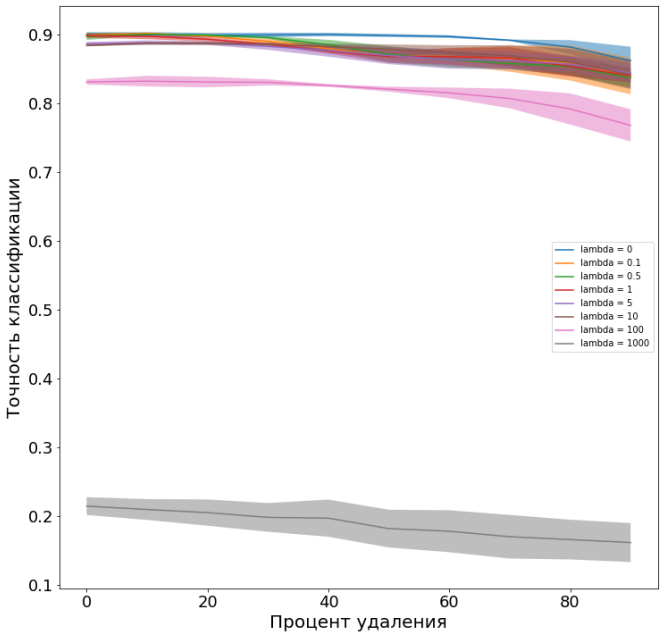

Рис. 5 График зависимости точности классификации от процента удалённых параметров для модели с линейной аппроксимацией с дообучением