Порождение моделей заданной сложности с использованием байесовских гиперсетей

О. С. Гребенькова

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем
Научный руководитель: к.ф.-м.н. Бахтеев Олег Юрьевич

Весна 2021 г.

Задача построения модели глубокого обучения

Цель

Предложить метод оптимизации модели глубокого обучения с контролем сложности модели.

Исследуемая проблема

По построению семейство моделей глубокого обучения имеет избытычное число параметров. Поэтому оптимизация и выбор модели с наперед заданной сложностью является вычислительно сложной задачей.

Метод решения

Предлагаемый метод заключается в представлении модели глубокого обучения в виде гиперсети, с использованием байесовского подхода.
Гиперсеть — сеть, которая порождает параметры для оптимальной модели.

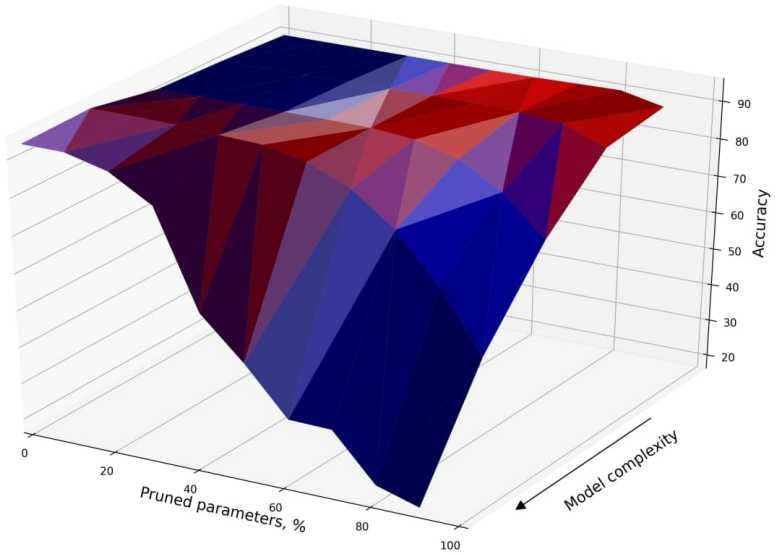


График зависимости точности классификации от процента удалённых параметров и параметра сложности модели

1. выборка

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \ldots, m, \quad \mathbf{x}_i \in \mathbb{R}^m, \quad y_i \in \{1, \ldots, Y\},$$

2. 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;

3. априорное распределение вектора параметров в пространстве $\mathbb{R}^n$:

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

где $\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}$ — вектор средних и матрица ковариации априорного распределения;

4. распределение, аппроксимирующее неизвестное апостериорное распределение $p(\mathbf{w}|\mathfrak{D})$:

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

Здесь $\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}$ — вектор средних и матрица ковариации. Предполагается, что:

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

Логарифмическая функция правдобподобия выборки:

$$\log p(\mathfrak{D}|\mathbf{w}).$$

Логарифм обоснованности модели:

$$\log p(\mathfrak{D}) = \log \int\limits_{\mathbf{w}\in\mathbb{R}^n} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}.$$

При оценке интеграла получаем:

$$\log p(\mathfrak{D}) \geq \int\limits_{\mathbf{w}\in\mathbb{R}^n} q(\mathbf{w}) \log \frac{p(\mathbf{w})}{q(\mathbf{w})} d\mathbf{w} + \int\limits_{\mathbf{w}\in\mathbb{R}^n} q(\mathbf{w}) \log p(\mathfrak{D}|\mathbf{w})d\mathbf{w} =$$

$$= -D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w})\big) + \mathbb{E}_{q(\mathbf{w})} \log p(\mathfrak{D}|\mathbf{w}).$$

Обобщенная функция обоснованности:

$$\mathfrak{L}_1(\lambda) = -\lambda D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w})\big) + \mathbb{E}_{q(\mathbf{w})} \log p(\mathfrak{D}|\mathbf{w}); \tag{1}$$

$$\mathfrak{L}_2(\lambda) = -D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\lambda)\big) + \mathbb{E}_{q(\mathbf{w})} \log p(\mathfrak{D}|\mathbf{w}); \tag{2}$$

$$\mathfrak{L}_3(\lambda) = \lambda\|\mathbf{w}\|^2 + \mathbb{E}_{q(\mathbf{w})} \log p(\mathfrak{D}|\mathbf{w}); \tag{3}$$

где под $p(\mathbf{w}|\lambda)$ понимаем распределение следующего вида $\sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{\lambda}\mathbf{A}_{\mathrm{pr}}^{-1})$.

Максимизация функционала

$$\mathfrak{G}_1(\lambda) = \arg\max_{\mathbf{w}\in\mathbb{R}^n}(\log p(\mathfrak{D}|\mathbf{w}) - \lambda D_{KL}(q(\mathbf{w})||p(\mathbf{w}))), \tag{4}$$

$$\mathfrak{G}_2(\lambda) = \arg\max_{\mathbf{w}\in\mathbb{R}^n}(\log p(\mathfrak{D}|\mathbf{w}) - D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\lambda))), \tag{5}$$

$$\mathfrak{G}_3(\lambda) = \arg\max_{\mathbf{w}\in\mathbb{R}^n}(\log p(\mathfrak{D}|\mathbf{w}) + \lambda\|\mathbf{w}\|^2). \tag{6}$$

Построение гиперсети для контроля сложности модели

Гиперсеть

Параметрическое отображение из множества Λ во множество параметров модели $\mathbb{R}^n$:

$$\mathbf{G} : \Lambda \times \mathbb{R}^u \to \mathbb{R}^n,$$

где $\mathbb{R}^u$ — множество допустимых параметров гиперсети, Λ — множество параметров, контролирующих сложность модели.

Реализация с линейной аппроксимацией

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

Реализация с кусочно-линейной аппроксимацией

$$\mathbf{G}_{\text{piecewise}}(\lambda) = \sum_{i=0}^{N} \mathbf{F}(t_i, t_{i+1}, \lambda),$$

$$\mathbf{F}(t_i, t_{i+1}, \lambda) = \begin{cases} \mathbf{b}(t_i) + \frac{\mathbf{b}(t_{i+1}) - \mathbf{b}(t_i)}{t_{i+1} - t_i}(\lambda - t_i), & t_i \leq \lambda \leq t_{i+1} \\ 0, \text{иначе} \end{cases}.$$

Критерий удаления параметров — относительная плотность модели:

$$\mathbf{g}(w_i) \propto \exp \frac{\mu_i^2}{2\sigma_i^2},$$

$$\mathbf{g}(w_i) \propto \exp\left(-w_i^2\right).$$

Теорема

Пусть выполнены следующие условия:

1. существует компакт $\mathbb{U} \in \mathbb{R}^n$, который содержит единственный минимум $\mathbf{w}^*(\lambda)$ для каждого $\lambda \in \Lambda$;
2. существует последовательность моделей $\mathbf{w}_n(\lambda) \in \mathbb{U}$ такая, что $\mathbb{E}\mathfrak{L}(\mathbf{w}_n(\lambda)) \underset{n\to\infty}{\to} \max$.

Тогда $\mathbf{g}(\mathbf{w}_n(\lambda)) \xrightarrow{p} \mathbf{g}(\mathbf{w}^*(\lambda))$, где $\mathbf{g}$ — это непрерывный критерий для удаления параметров.

Цель

Исследовать поведение обобщенной функции обоснованности модели. Сравнить методы построения разных моделей. Сравнить с теоретическими результатами.

Проведено сравнение следующих моделей:

(а) вариационная сеть;

(б) сеть с репараметризацией;

(в) базовая нейросеть с регуляризатором;

(г) вариационная сеть с линейной гиперсетью;

(д) сеть с репараметризацией и линейной гиперсетью ;

(е) базовая нейросеть с регуляризатором и линейной гиперсетью;

(ё) вариационная сеть с кусочно-линейной гиперсетью;

(ж) сеть с репараметризацией и кусочно-линейной гиперсетью ;

(з) базовая нейросеть с регуляризатором и кусочно-линейной гиперсетью.

Вид используемой нейросети для эксперимента на MNIST:

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

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

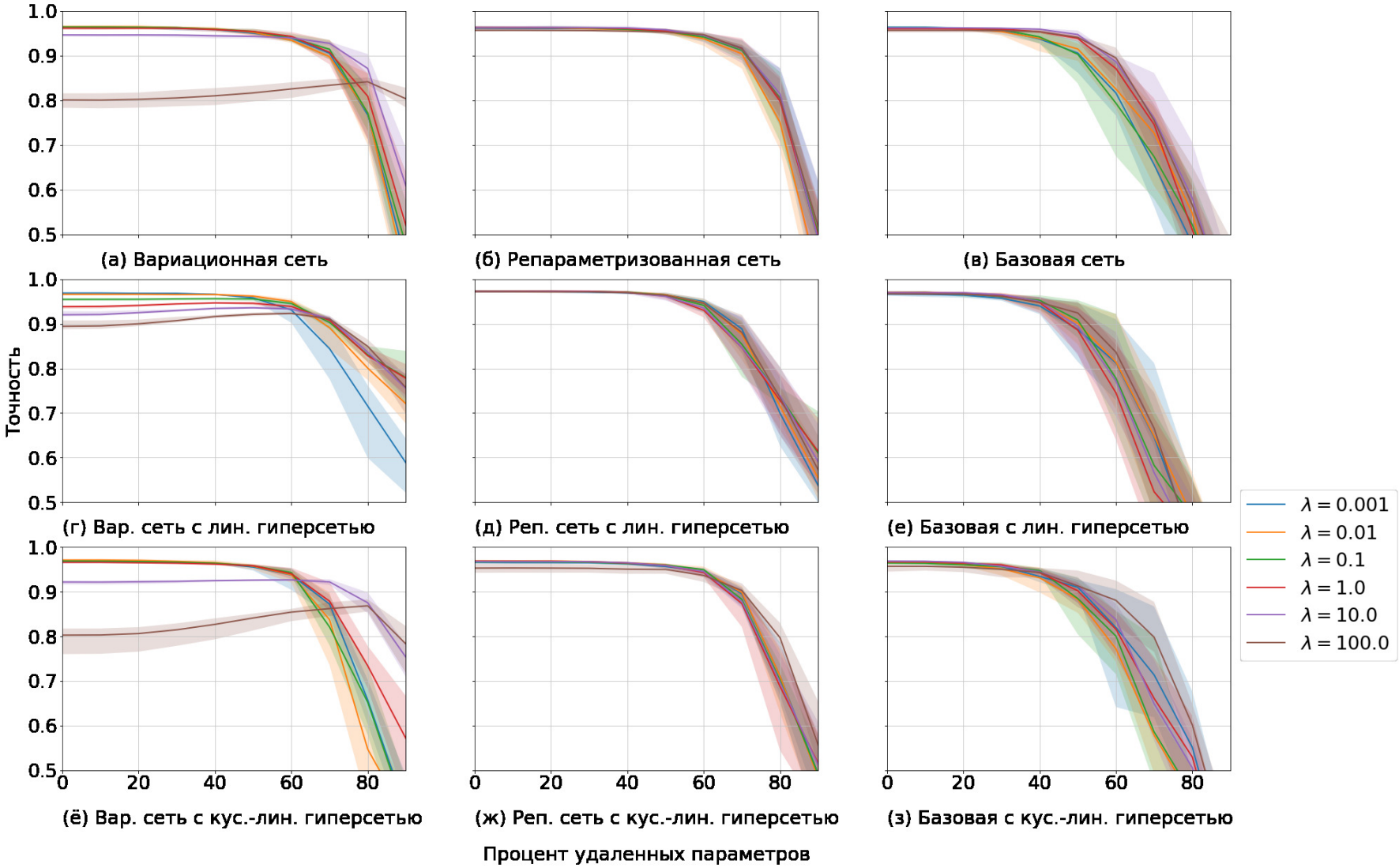
$$\mathbf{softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{k} \exp(x_j)} \quad i = 1, \ldots, k,$$

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

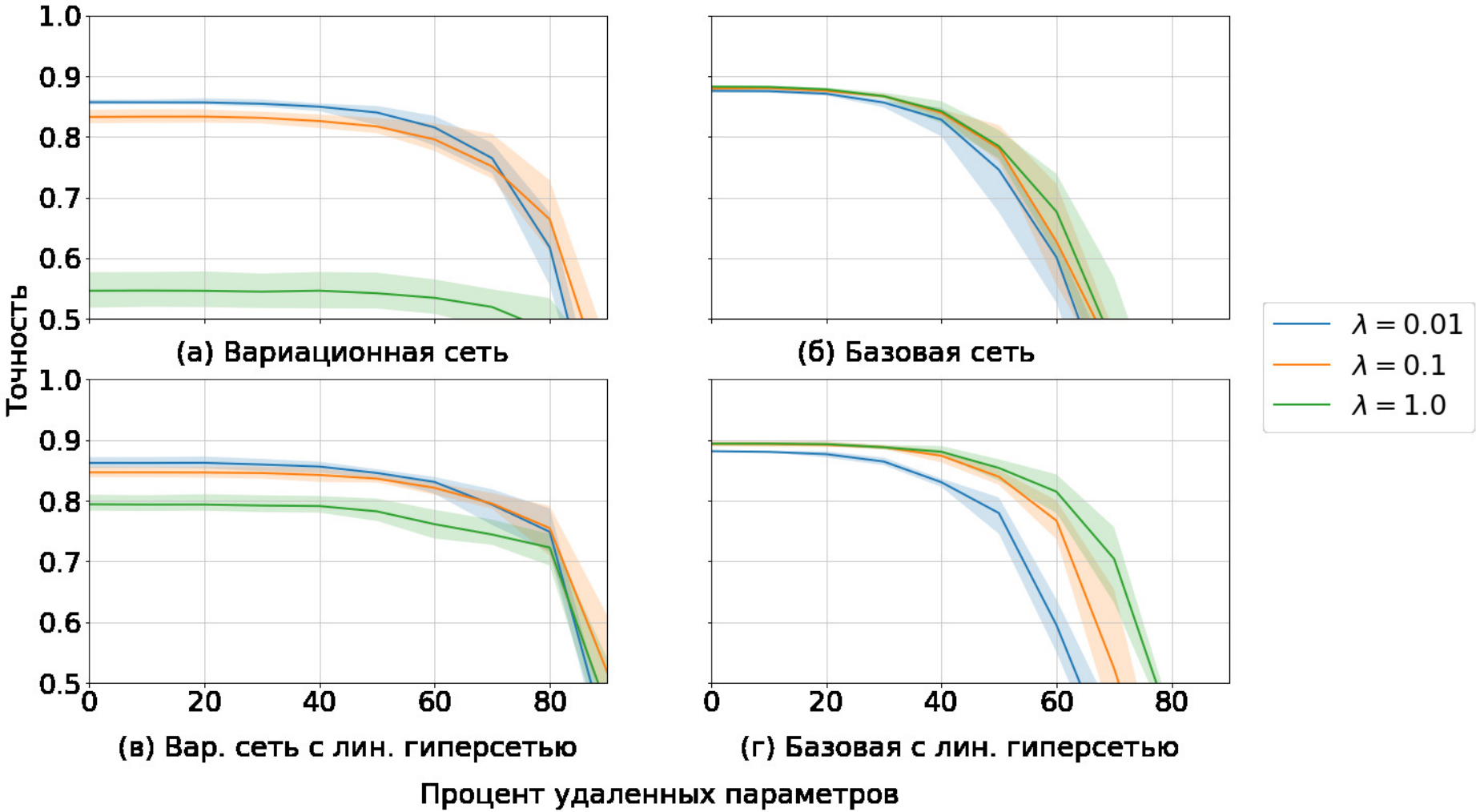
Критерий качества модели — точность классификации

$$\text{Accuracy} = 1 - \frac{1}{m} \sum_{i=1}^{m} [f(\mathbf{x}_i, \mathbf{w}) \neq y_i].$$

Сравнение моделей для выборки MNIST



Сравнение моделей для выборки CIFAR



Выносится на защиту

1. Исследована возможность прореживания модели глубокого обучения с помощью вариационных методов.
2. Показано, что несмотря на потерю в качестве, гиперсеть в разных реализациях получает схожие результаты, что и обычные модели, при меньших вычислительных затратах.
3. Экспериментально и теоретически доказано, что модели, полученные с помощью гиперсетей, сохраняют схожие свойства (к примеру точность классификации) при прореживании.
4. Проведены эксперименты для различных моделей и выборок, подтверждающие работоспособюность предложенного метода.