

# 1 Цели работы

Основная цель работы — предложить метод выбора оптимальной структуры модели на основе байесовского вывода. Предполагается, что предложенный метод будет работать не хуже существующих невероятностных методов (в плане ассигасы наиболее вероятных параметров). При этом такая модель будет более устойчива к возмущениям параметров/объектов, а также будет позволять прореживать блоки параметров на основании вероятностных оценок, полученных при поиске структуры.

Основные требования к методу:

1. Метод должен быть сопоставим с методами, предложенными Zoph et al.
2. Метод должен позволять проводить поиск не только CNN, но и полносвязных сетей.
3. Метод должен быть эффективен по ресурсами.

## 2 Обозначения

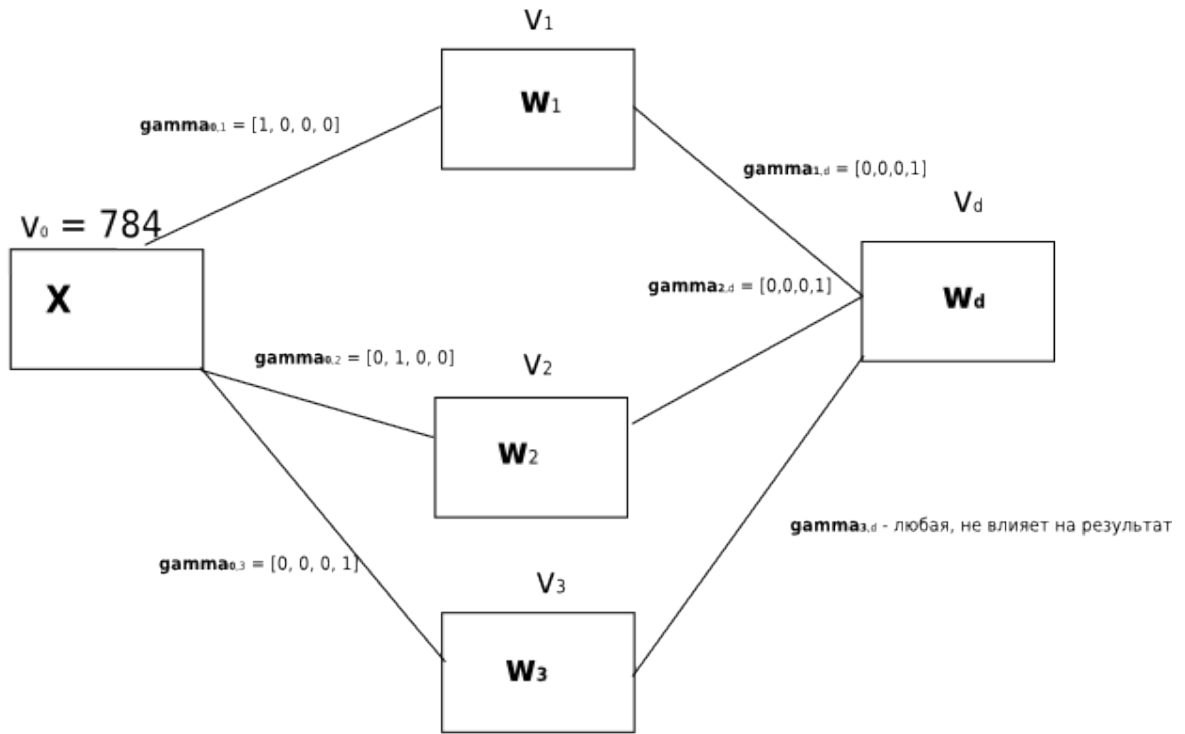
Выборка и пр.				
Что	Обозначение	Простр.	Что делает	С чем связано
Выборка	$\mathfrak{D}$	-	-	$\mathbf{X}, \mathbf{y}$
Объекты	$\mathbf{X}$	$\mathbb{R}^{m \times n}$	-	$\mathfrak{D}$
Метки	$\mathbf{y}$	$\mathbb{R}$ или $Z$	-	$Z, \mathfrak{D}$
Количество классов	$Z$	$\mathbb{Z}$	-	$\mathbf{y}$
Модель (как граф)				
Нелинейная функция	$\mathbf{o}$	-	Нелинейно отображает объекты внутри сети	-
Множество нелинейных функций	$\mathcal{O}$	-	Набор возможных нелинейных функций для каждого элемента сети	$\mathbf{o}, \mathbf{f}$
Вершина графа	$v$	$\mathbb{N}$	Определяет операция над выборкой. Число определяет размерность объекта после прохождения операции	$V$
Множество вершин	$V$	-	Множество возможных операций над выборкой. Каждая операция определяется своей матрицей, размерность которой соответствует числу $v \in V$	$v, d, \mathbf{f}$
Число вершин	$d$ -	-	$V$	
Ребро графа	$e$	$2^{d \times d - 1}$	Определяет наличие нелинейной функции, идущей от одной вершины к другой	$V$
Модель	$\mathbf{f}$	$E \rightarrow \mathcal{O}$	Отображение из ребра графа в нелинейную операцию	$E, \mathcal{O}$
Множество моделей	$\mathfrak{F}$	-	Множество моделей для заданного графа	$\mathbf{f}, V, E$

Модель (как вероятностная модель)				
Что	Обозначение	Простр.	Что делает	С чем связано
Параметры для одной вершины	$\mathbf{w}$	$\mathbb{R}^v$		$v, \mathbf{W}$
Вектор параметров модели	$\mathbf{W}$	-	Все параметры модели	$\mathbf{w}, \mathbf{f}$
Структура для вершины, бинарный вектор с одной единицей	$\gamma$	$2^{ O }$	Определяет веса нелинейных операций, входящих в вершину	$\mathbf{G}, v$
Структура для модели	$\mathbf{G}$	-	Определяет все веса нелинейных операций в модели	$\gamma$
Ковариация параметров	$\mathbf{A}^{-1}$	$\mathbb{R}^n$	Определяет prior для параметров	$\mathbf{W}$
Концентрация	$\alpha$	$\mathbb{R}$	Определяет prior для структуры	$\mathbf{G}$

Вариационный вывод				
Что	Обозначение	Простр.	Что делает	С чем связано
Вариационное распределение параметров модели	$q_{\mathbf{w}}$	-	Аппроксимирует апостериорное распределение	$\mathbf{A}^Q, \mu$
Среднее распределения параметров	$\mu$	-	-	$\mathbf{A}^Q, q_{\mathbf{w}}$
Вариационное распределение параметров структуры	$q_{\gamma}$	-	Аппроксимирует апостериорное распределение	$\mathbf{g}, \tau$
Параметры распределения гумбель-софтмакс	$\mathbf{g}$	-	-	$q_{\gamma}$
Параметр температуры гумбель-софтмакс	$\tau$	-	-	$q_{\gamma}$

Оптимизация				
Функция потерь	$L$	-	Оптимизируемая функция	$Q, \theta, T$
Функция валидации	$Q$	-	Оптимизируемая по гиперпараметрам функция	$L, \mathbf{h}, T$
Оптимизируемые параметры	$\theta$	-	-	$L, T$
Оптимизируемые гиперпараметры	$\mathbf{h}$	-	-	$Q, T$
Оператор оптимизации	$T$	-	-	$L, Q$
Количество итераций	$\eta$	-	-	$T$

$f(\mathbf{X}) = \mathbf{w}_d(\text{sigmoid}(\mathbf{w}_1\mathbf{X}) + \tanh(\mathbf{w}_2\mathbf{X})), n = 784, Z = 10$   
 $O = [\text{sigmoid}, \tanh, \text{identity}, \text{None}]$



$\mathbf{W} = [\mathbf{w}_0, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_d], p(\mathbf{W}) \sim N(\mathbf{0}, \mathbf{A}^{-1})$

$\mathbf{\Gamma} = [\mathbf{gamma}_{i,j}], \mathbf{gamma}_{i,j} \sim \text{Dir}(\alpha)$

### 3 Постановка задачи

Задана выборка

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

состоящая из множества пар «объект-метка»

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

Метка  $y$  объекта  $\mathbf{x}$  принадлежит либо множеству:  $y \in \mathbb{Y} = \{1, \dots, Z\}$  в случае задачи классификации, где  $Z$  — число классов, либо некоторому подмножеству вещественных чисел  $y \in \mathbb{Y} \subseteq \mathbb{R}$  в случае задачи регрессии. Определим множество архитектур моделей глубокого обучения для дальнейшего выбора оптимальной.

**Определение** Функция  $\mathbf{o}$  называется  $n_1, n_2$  - допустимой, если для вектора  $\mathbf{x} \in \mathbb{R}^{n_1}$  функция  $\mathbf{o}$  определена и  $\mathbf{o}(\mathbf{x}) \in \mathbb{R}^{n_2}$ .

**Определение** Пусть задан вектор натуральных чисел  $V = [v_0, v_1, \dots, v_d]$  и множество (возможно пустых) ребер  $E$ . Пусть также задано множество нелинейных функций  $O$ , такое что нулевая функция лежит в этом множестве. Соответствие  $\mathbf{f} : E \rightarrow O$  будем называть моделью, если оно удовлетворяет условиям:

1.  $\deg^+(v_0) = 0$
2.  $\deg^-(v_d) = 0$ .
3. Для каждого ребра  $e$  между вершинами  $v_i, v_j$  определена  $v_i, v_j$ -допустимая операция  $\mathbf{o} \in O$ .

**Определение** Параметрами модели (носителем?)  $\mathbf{f}$ , соответствующей графу  $\langle V, E \rangle$  с множеством операций  $O$  будем называть вектор матриц  $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_d]$  и бинарным множеством векторов, соответствующих ребрам графа  $\langle V, E \rangle$   $\mathbf{\Gamma} = \{\gamma_{i,j} : (v_i, v_j) \in E\}$ ,  $\gamma_{i,j} \in \mathbb{R}^{|O|}$ , т.ч.  $\forall \gamma \in \mathbf{\Gamma} : \|\gamma\|_1 \leq 1$ .

Результатом функции  $\mathbf{f}(\mathbf{x})$  будем полагать рекуррентное выполнение операций над  $\mathbf{x}$  с выполнением обхода от вершины  $v_0$ :

$$f_0(\mathbf{x}) = \mathbf{x},$$

$$f_i(\mathbf{x}) = \sum_{j:(j,i) \in E} \sum_{k=1}^{|O|} \gamma_{i,j}^k \mathbf{o}_k(\mathbf{w}_i f_j(\mathbf{x})).$$

Пусть для модели определено правдоподобие:  $p(\mathbf{y}|\mathbf{W}, \mathbf{\Gamma}, \mathbf{X})$ . Пусть также задано априорное распределение параметров модели  $\mathbf{W}$ :

$$\mathbf{W} \sim \mathcal{N}(0, \mathbf{A}^{-1}), \quad \mathbf{\Gamma} \sim \text{Dir}(\alpha),$$

где  $\mathbf{A}^{-1}$  — диагональная матрица гиперпараметров,  $\alpha$  — гиперпараметр концентрации весов параметров.

**Определение** Сложностью модели  $\mathbf{f}$  назовем правдоподобие модели:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \alpha) = \int_{\mathbf{W}, \mathbf{\Gamma}} p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma})p(\mathbf{w}|\mathbf{A})p(\mathbf{\Gamma}|\alpha)d\mathbf{W}d\mathbf{\Gamma}. \quad (2)$$

**Определение** Пусть заданы множество вершин  $V$ , множество операций  $O$ . Обозначим за  $\mathfrak{F}(V, E, O)$  множество всех моделей для заданных  $V, O$ . Модель  $\mathbf{f}$  назовем оптимальной, если достигается максимум интеграла:

$$\mathbf{f} = \arg \max_{\mathbf{f}' \in \mathfrak{F}(V, E, O)} \max_{\mathbf{A}} p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \alpha).$$

## 4 Вариационный вывод

Введем два аппроксимирующих параметрических распределения:

$$q_{\mathbf{w}} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_q^{-1}) \approx p(\mathbf{W}|\mathbf{X}, \mathbf{y}, \mathbf{A}, \boldsymbol{\Gamma}),$$

$$q_{\gamma} \sim \text{Gumbel-Softmax}(\mathbf{g}) \approx p(\boldsymbol{\Gamma}|\mathbf{X}, \mathbf{y}, \mathbf{A}, \alpha).$$

Пусть  $L$  — вариационная оценка правдоподобия:

$$L = \log p(\mathbf{y}|\hat{\mathbf{W}}, (\hat{\boldsymbol{\Gamma}}) - \text{KLD}(q_{\gamma}||p(\boldsymbol{\Gamma})) - \text{KLD}(q_{\mathbf{w}}||p(\mathbf{W})).$$

Пусть  $\boldsymbol{\theta}$  — параметры распределений  $q_{\mathbf{w}}$  и  $q_{\gamma}$ :

$$\boldsymbol{\theta} = [\boldsymbol{\mu}, \mathbf{A}_q^{-1}, \mathbf{g}].$$

Пусть  $\mathbf{h}$  — параметры априорных распределений, а также гиперпараметры распределений  $q_{\mathbf{w}}$  и  $q_{\gamma}$ :

$$\mathbf{h} = [\mathbf{A}, \alpha, \tau],$$

где  $\tau$  — температура Гумбель-софтмакс распределения.

Сформулируем задачу поиска оптимальной модели как двухуровневую задачу.

$$\hat{\mathbf{h}} = \arg \max_{\mathbf{h} \in \mathbb{R}^h} Q(T^{\eta}(\boldsymbol{\theta}_0, \mathbf{h})), \quad (3)$$

где  $T$  — оператор оптимизации, решающий задачу оптимизации:

$$L(T^{\eta}(\boldsymbol{\theta}_0, \mathbf{h})) \rightarrow \min.$$

### 4.1 Функция валидации Q

Предварительно общий вид:

$$Q_{\eta} = c_1 \mathbb{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Gamma}) + c_2 D_{\text{KL}}(q_{\mathbf{W}}^{\eta} | p(\mathbf{W}, \mathbf{A}^{\eta})) + c_3 D_{\text{KL}}(q_{\boldsymbol{\Gamma}}^{\eta} | p(\boldsymbol{\Gamma}, \alpha^{\eta})) + c_4 \sum_{t=1}^{\eta} D_{\text{KL}}(q_{\boldsymbol{\Gamma}}^t | p(\boldsymbol{\Gamma}, \alpha^{\tau}))$$

Идея последнего члена: <https://arxiv.org/pdf/1704.03003.pdf> Для адекватности оптимизации требуется, чтобы  $\alpha$  или  $\tau$  вырождали структуру при  $T \rightarrow \inf$ .

**Утверждение (предварительно)**

1. При  $c_1 = 1, c_2 = -1, c_3 = -1, c_4 = 0$  получается оптимизация Evidence. Оптимизацию можно проводить одним уровнем.
2. При  $c_1 = c_2 = c_3 = 0, c_4 > 0$  оптимизация максимизирует количество информации при изменении структуры, близко к полному перебору.

3. При  $c_1 > 0, c_2 = c_3 = c_4 = 0$  и вырождении априорных распределений получается обычная оптимизация без кросс-валидации.
4. При  $c_1 = 1, c_2 = -1, c_3 = -1, c_4 = 0$  получается ridge-отбор признаков.
5. При  $c_1 = 1, c_2, c_3 = -1$  и использовании оператора вида  $T = \nabla Q + \epsilon$  оптимизация будет проходить схожим образом с динамикой Ланжевена. Можно доказать, что оптимизация сходится к апостериорным значениям параметров. Связь с генетикой?
6. При постоянном варьировании  $c_2, c_3$  можно добиться оптимизации типа Add-Del.

## 5 Нерешенные вопросы

1. В работе Zoph et al. структуры оптимизируются с ограничением, что сеть является цепочкой повторяемых подмоделей. Таким образом, уменьшается количество возможных вариантов оптимизации и затраты по ресурсам. В данной версии постановки задачи ограничений нет.