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 accuracy 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.

Основныые 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 Постановка задачи

Задана выборка

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

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: $y \in \mathbb{Y} = \{1, \ldots, Z\}$ 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 $V = [v_0, v1, \ldots, v_d]$ и множество (возможно пустых) ребер E. Пусть также задано множество нелинейных функций O, такое что нулевая функция лежит в этом множестве. Соответствие $\mathbf{f} : E \to O$ 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 $< V, E >$ с множеством операций O будем называть вектор матриц $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_d]$ и бинарным множеством векторов, соответствующих ребрам графа $< V, E > \mathbf{\Gamma} = \{\boldsymbol{\gamma}_{i,j} : (v_i, v_j) \in E\}, \boldsymbol{\gamma}_{i,j} \in \mathbb{R}^{|O|}$, т.ч. $\forall \gamma \in \mathbf{\Gamma} : ||\gamma||_1 = 1$.

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

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

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

Далее будем рассматривать следующую параметризацию распределения Дирихле:

$$p(x_1, \ldots, x_K) = \frac{1}{B(\frac{\mathbf{m}}{c})} \prod_{i=1}^{K} x_i^{\frac{m_i}{c}-1},$$

c — параметр точности (precision), $\mathbf{m}$ — средний вектор.

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

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

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

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

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

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

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

3 Вариационный вывод

Введем два аппроксимирующих параметрических распределения:

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

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

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 вариационных распределений:

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

Пусть $\mathbf{h}$ — параметры априорных распределений:

$$\mathbf{h} = [\mathbf{A}, \mathbf{m}].$$

Сформулируем задачу поиска оптимальной модели как двухуровневую задачу.

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

где T — оператор оптимизации, решающий задачу оптимизации:

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

3.1 Функция валидации Q

Пусть η — общее число итераций оптимизации. Разобьем оптимизацию модели на π эпох, каждая по η' итераций. Функция валидации имеет вид:

$$Q = c_1 \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Gamma}) + c_2 D_{\text{KL}}(q_{\mathbf{W}}^{\eta}|p(\mathbf{W}|\mathbf{A}^{\eta})) + c_3 D_{\text{KL}}(q_{\boldsymbol{\Gamma}}^{\eta}|p(\boldsymbol{\Gamma}, \mathbf{m}^{\eta})) + c_4 \sum_{t=1}^{[\frac{T}{\eta'}]} D_{\text{KL}}(p(\boldsymbol{\Gamma}|\mathbf{m}^{\eta})|p(\boldsymbol{\Gamma}|\mathbf{m}^{t\eta'})).$$

Идея последнего слагаемого: https://arxiv.org/pdf/1704.03003.pdf

Рассмотрим подробнее режимы оптимизации, получаемые из данной функции.

Тривиальные режимы Пусть $c \to_t 0$. При отказе от оптимизации гиперпараметров и при фиксированных $\mathbf{A}^q \to 0$ задача сводится оптимизации правдоподобия выборки.

При отказе от оптимизации гиперпараметров и сэмплировании среднего из $q_{\mathbf{W}}$ задача сводится к гребневой регрессии.

Оптимизация правдоподобия модели Пусть $c \to_t 0$. Пусть $c_1 = 1$, $c_2 = c_3 = -1$. Тогда оптимизация эквивалентна оптимизации Evidence. TODO: рассмотреть, когда оптимизация эквивалентна одновременной оптимизации L, Q.

Полный перебор Пусть $c >> 0$ для каждой итерации начала эпохи и $c \to_t 0$ для каждой итерации конца эпохи. Пусть $c_4 > 0, c_1, c_3 = 0$. Тогда задача сводится к оптимизации правдоподобия модели (по параметрам модели) и перебору (по структуре). TODO: доказать.

Add-Del по параметрам Пусть $c >> 0$ для каждой итерации начала эпохи и $c \to_t 0$ для каждой итерации конца эпохи. Пусть для каждой четной эпохи $c_2 = 1$, для нечетной: $c_2 < 1$.

TODO: формализация Add-Del.

Add-Del по структуре Пусть $c >> 0$ для каждой итерации начала эпохи и $c \to_t 0$ для каждой итерации конца эпохи. Пусть для каждой четной эпохи $c_3 = 1$, для нечетной: $c_3 < 1$.

TODO: формализация Add-Del.

Эволюционный алгоритм TODO: формализация

4 Обозначения

Что	Обозначение	Простр.	Что делает	С чем связано
Выборка и пр.				
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	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	$\mathbf{f}$	$E \to O$	Отображение из ребра графа в нелинейную операцию	E, 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	γ	$2^{\lvert O\rvert}$	Определяет веса нелинейных операций, входящих в вершину	$\mathbf{\Gamma}, v$
Структура для модели	$\mathbf{\Gamma}$	-	Определяет все веса нелинейных операций в модели	γ
Ковариация параметров	$\mathbf{A}^{-1}$	$\mathbb{R}^n$	Определяет prior для параметров	$\mathbf{W}$
Концентрация	α	$\mathbb{R}$	Определяет prior для структуры	$\mathbf{\Gamma}$

Вариационный вывод

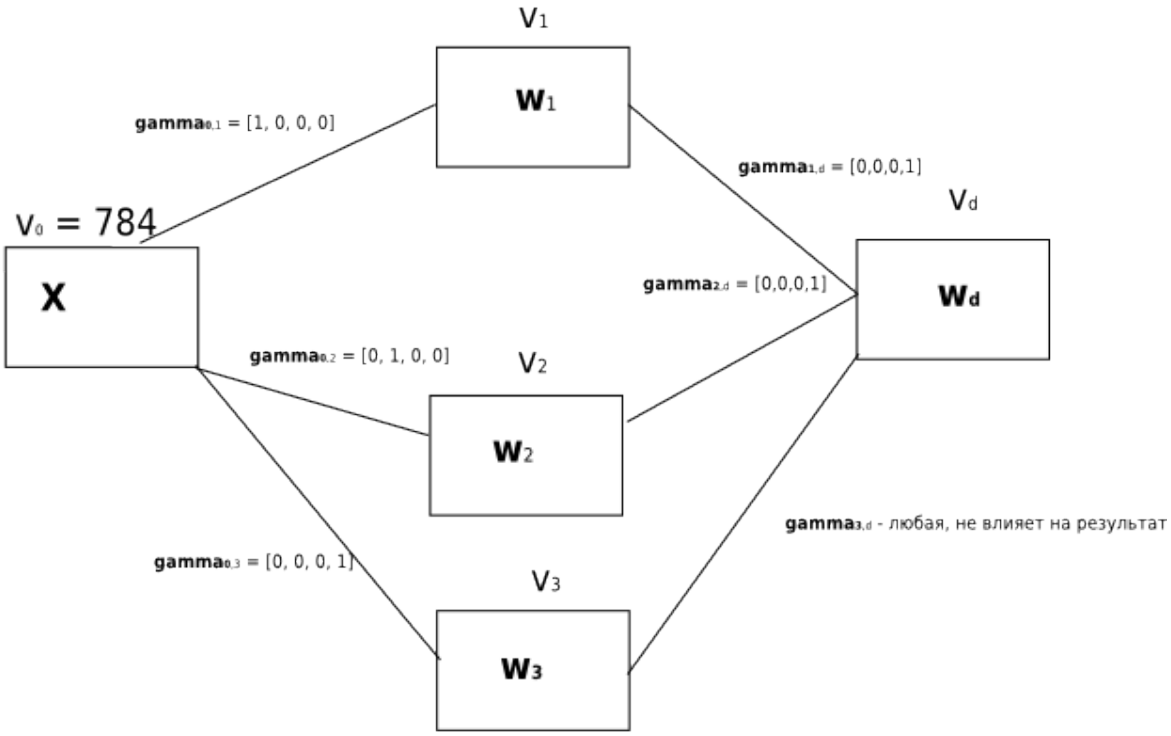
Что	**Обозначение**	**Простр.**	**Что делает**	**С чем связано**
Вариационное распределение параметров модели	$q_{\mathbf{w}}$	-	Аппроксимирует апостериорное распределение	$\mathbf{A}^Q, \boldsymbol{\mu}$
Среднее распределения параметров	$\boldsymbol{\mu}$	-	-	$\mathbf{A}^Q, q_{\mathbf{w}}$
Вариационное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

Оптимизация

Функция потерь	L	-	Оптимизируемая функция	$Q, \boldsymbol{\theta}, T$
Функция валидации	Q	-	Оптимизируемая по гиперпараметрам функция	$L, \mathbf{h}, T$
Оптимизируемые параметры	$\boldsymbol{\theta}$	-	-	L, T
Оптимизируемые гиперпараметры	$\mathbf{h}$	-	-	Q, T
Оператор оптимизации	T	-	-	L, Q
Количество итераций	η	-	-	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 = [sigmoid, tanh, identity, None]



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

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