Байесовский выбор архитектуры нейросетевой модели

Сотников А. Д., группа М05-004б
Научный руководитель: к. ф-м н. Бахтеев О.Ю.

Московский Физико-Технический институт
Кафедра интеллектуальных систем

21 июня, 2022

Введение

Задача

Поиск архитектуры нейросетевой модели (ПАНМ) — метод автоматического проектирования архитектуры нейронной сети на заданных задаче и наборе данных.

Мотивация

Методы ПАНМ не обладают достаточной устойчивостью, то есть уязвимы к внешним воздействиям.

Гипотеза

Проведение байесовского вывода повышает устойчивость базового метода.

Предложение

Реализовать модификацию базового метода поиска архитектуры с помощью байесовского вывода распределений параметров и структуры модели.
Предоставить теоретическую интерпретацию предлагаемого метода.

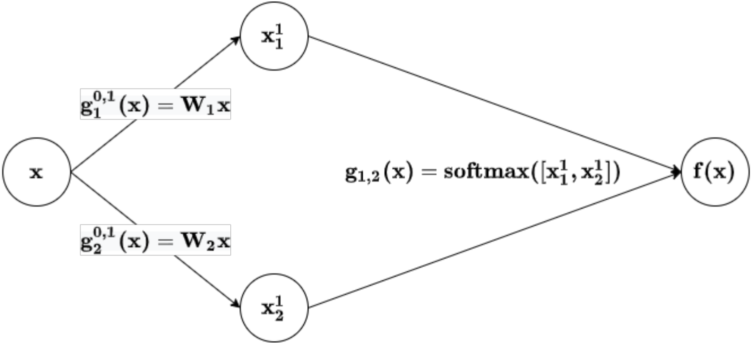
Основные определения

Определение 1

Моделью называется дифференцируемая по параметрам функция

$$\mathbf{f}(\mathbf{w}, \mathbf{x}) = y : \mathbb{W} \times \mathbb{X} \to \mathbb{Y},$$

где $\mathbf{w} \in \mathbb{W}$ - параметры модели, $\mathbf{x} \in \mathbb{X}$ - признаковое описание входного объекта, $y \in \mathbb{Y}$ - метка входного объекта.



Определение 2

Пусть на ребре (j, k) задан вектор операций $\mathbf{g}^{j,k}$, $|\mathbf{g}^{j,k}| = N^{j,k}$. *Структурными параметрами* назовём вектор $\gamma^{j,k} = [0, 1]^{N^{j,k}}$.

Структурой модели называется конкатенация её структурных параметров $\mathbf{\Gamma} = \{\gamma^{j,k} | (j, k) \in E\}$.

Определение 3

Архитектурой модели называется совокупность её параметров и структуры.

Определение 4

Гиперапараметрами $\mathbf{h} \in \mathbb{H}$ модели назовём параметры распределения $\mathbf{p}(\mathbf{w}, \mathbf{\Gamma} \mid \mathbf{h})$.

Существующие методы

Методы ПАНМ

- Методы, основанные на обучении с подкреплением
- Градиентные методы
- Вероятностный подход

Методы, повышающие устойчивость ПАНМ

- "Отравление" пространства поиска
- Оптимизация мер устойчивости
- Регуляризация структурных параметров

Постановка задачи

- Набор данных $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, $\mathbf{x_i} \in \mathbb{R}^m$, $y_i \in \mathbb{N}$
- Параметры $\mathbf{w} \sim p(\mathbf{w} \,|\, \mathbf{h})$ и структура $\mathbf{\Gamma} \sim p(\mathbf{\Gamma} \,|\, \mathbf{h})$ модели являются независимыми случайными величинами
- Вероятностная модель:

$$p(\mathbf{w}, \mathbf{\Gamma} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h}) = p(\mathbf{w} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h}) \cdot p(\mathbf{\Gamma} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h})$$

- Вводится вариационное распределение

$$q(\mathbf{w}, \mathbf{\Gamma} \,|\, \boldsymbol{\theta}) = q_{\mathbf{w}}(\mathbf{w} \,|\, \boldsymbol{\theta}_{\mathbf{w}}) \cdot q_{\mathbf{\Gamma}}(\mathbf{\Gamma} \,|\, \boldsymbol{\theta}_{\mathbf{\Gamma}})$$

- Функция ошибки:

$$\mathcal{L}(\mathbf{w}, \mathbf{\Gamma}) = \mathcal{D}_{KL}\left(q(\mathbf{w}, \mathbf{\Gamma} \,|\, \boldsymbol{\theta}) \,||\, p(\mathbf{w}, \mathbf{\Gamma} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h})\right)$$

- Ставится двухуровневая оптимизационная задача:

$$\min_{\mathbf{\Gamma}} \; \mathcal{L}_{val}(\hat{\mathbf{w}}, \mathbf{\Gamma})$$
$$s.t. \;\; \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathcal{L}_{train}(\mathbf{w}, \mathbf{\Gamma})$$

Вариационная нижняя оценка обоснованности

По теореме Байеса

$$p(\mathbf{w}, \mathbf{\Gamma} \mid \mathbf{X}, \mathbf{y}, \mathbf{h}) = \frac{p(\mathbf{y} \mid \mathbf{w}, \mathbf{\Gamma}, \mathbf{X}) \cdot p(\mathbf{w}, \mathbf{\Gamma} \mid \mathbf{X}, \mathbf{h})}{p(\mathbf{y} \mid \mathbf{X}, \mathbf{h})}$$

Выразим обоснованность через вариационное распределение

$$\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{h}) = \iint_{\mathbf{w}, \mathbf{\Gamma}} q(\mathbf{w}, \mathbf{\Gamma} \mid \boldsymbol{\theta}) \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{h}) d\mathbf{\Gamma} d\mathbf{w} =$$

$$\mathbb{E}_{q(\mathbf{w}, \mathbf{\Gamma} \mid \boldsymbol{\theta})} \left[\log \frac{p(\mathbf{y}, \mathbf{w}, \mathbf{\Gamma} \mid \mathbf{X}, \mathbf{h})}{q(\mathbf{w}, \mathbf{\Gamma} \mid \boldsymbol{\theta})} \right] + \mathcal{D}_{KL} \left(q(\mathbf{w}, \mathbf{\Gamma} \mid \boldsymbol{\theta}) \mid\mid p(\mathbf{w}, \mathbf{\Gamma} \mid \mathbf{X}, \mathbf{y}, \mathbf{h}) \right)$$

Утверждение 1

$$\min_{\boldsymbol{\theta}} \mathcal{D}_{KL} \left(q(\mathbf{w}, \mathbf{\Gamma} \mid \boldsymbol{\theta}) \mid\mid p(\mathbf{w}, \mathbf{\Gamma} \mid \mathbf{X}, \mathbf{y}, \mathbf{h}) \right) \iff \max_{\boldsymbol{\theta}} \mathbb{E}_{q(\mathbf{w}, \mathbf{\Gamma} \mid \boldsymbol{\theta})} \left[\log \frac{p(\mathbf{y}, \mathbf{w}, \mathbf{\Gamma} \mid \mathbf{X}, \mathbf{h})}{q(\mathbf{w}, \mathbf{\Gamma} \mid \boldsymbol{\theta})} \right]$$

- Матожидание можно представить в виде

$$\mathbb{E}_{q(\mathbf{w},\boldsymbol{\Gamma} \mid \boldsymbol{\theta})}\left[\log \frac{p(\mathbf{y},\mathbf{w},\boldsymbol{\Gamma} \mid \mathbf{X},\mathbf{h})}{q(\mathbf{w},\boldsymbol{\Gamma} \mid \boldsymbol{\theta})}\right] =$$

$$= \mathbb{E}_{q(\mathbf{w},\boldsymbol{\Gamma} \mid \boldsymbol{\theta})}\Big[\log p(\mathbf{y} \mid \mathbf{w},\boldsymbol{\Gamma},\mathbf{X})\Big] - \mathcal{D}_{KL}\big(q(\mathbf{w},\boldsymbol{\Gamma} \mid \boldsymbol{\theta}) \mid\mid p(\mathbf{w},\boldsymbol{\Gamma} \mid \mathbf{X},\mathbf{h})\big)$$

- KL-дивергенция раскладывается на

$$\mathcal{D}_{KL}\big(q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma} \mid \boldsymbol{\theta}_{\boldsymbol{\Gamma}}) \mid\mid p(\boldsymbol{\Gamma} \mid \mathbf{h})\big) + \mathcal{D}_{KL}\big(q_{\mathbf{w}}(\mathbf{w} \mid \boldsymbol{\theta}_{\mathbf{w}}) \mid\mid p(\mathbf{w} \mid \mathbf{h})\big)$$

Окончательно, функция ошибки принимает вид

$$\mathcal{L}(\mathbf{w},\boldsymbol{\Gamma}) = -\mathbb{E}_{q(\mathbf{w},\boldsymbol{\Gamma} \mid \boldsymbol{\theta})} \log p(\mathbf{y} \mid \mathbf{w},\boldsymbol{\Gamma},\mathbf{X}) + \\ + \mathcal{D}_{KL}\big(q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma} \mid \boldsymbol{\theta}_{\boldsymbol{\Gamma}}) \mid\mid p(\boldsymbol{\Gamma} \mid \mathbf{h})\big) + \mathcal{D}_{KL}\big(q_{\mathbf{w}}(\mathbf{w} \mid \boldsymbol{\theta}_{\mathbf{w}}) \mid\mid p(\mathbf{w} \mid \mathbf{h})\big)$$

Вариационный вывод ПАНМ

Теорема 1 (Сотников, 2022)

Пусть заданы две структуры $\hat{\mathbf{\Gamma}}_1$ и $\hat{\mathbf{\Gamma}}_2$, функция ошибки на которых принимает одинаковое значение, то есть $\mathcal{L}(\mathbf{w}, \hat{\mathbf{\Gamma}}_1) = \mathcal{L}(\mathbf{w}, \hat{\mathbf{\Gamma}}_2)$. Пусть также $||\nabla^2_{\mathbf{\Gamma}}\mathcal{L}(\mathbf{w}, \hat{\mathbf{\Gamma}}_1)|| < ||\nabla^2_{\mathbf{\Gamma}}\mathcal{L}(\mathbf{w}, \hat{\mathbf{\Gamma}}_2)||$. Тогда справедливо

$$||\hat{\mathbf{\Gamma}}_1 - \mathbf{\Gamma}^*|| > ||\hat{\mathbf{\Gamma}}_2 - \mathbf{\Gamma}^*||,$$

где $\mathbf{\Gamma}^* = \arg\min\limits_{\mathbf{\Gamma}} \mathcal{L}_{val}(\mathbf{w}, \mathbf{\Gamma})$ – оптимальная структура.

Теорема 2 (Сотников, 2022)

Пусть задана функция $G(\mathbf{w}, \mathbf{\Gamma})$. Тогда для любого распределения $q_{\mathbf{\Gamma}}(\mathbf{\Gamma} \mid \delta)$ такого, что компоненты $\mathbf{\Gamma}$ – независимые случайные величины, справедливо

$$\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma} \mid \delta)}\left[G(\mathbf{w}, \mathbf{\Gamma})\right] \approx G(\mathbf{w}, \boldsymbol{\mu}) + \frac{\sigma^2}{2}\mathrm{Tr}\left(\nabla^2_{\mathbf{\Gamma}} G(\mathbf{w}, \boldsymbol{\mu})\right),$$

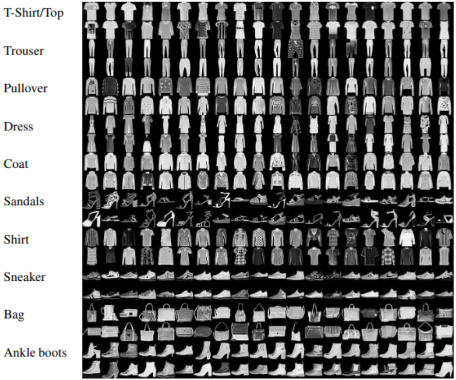
где $\boldsymbol{\mu} = \mathbb{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma} \mid \delta)}\left[\mathbf{\Gamma}\right]$ и $\boldsymbol{\sigma}^2 = \mathbb{D}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma} \mid \delta)}\left[\mathbf{\Gamma}\right]$.

Параметры эксперимента

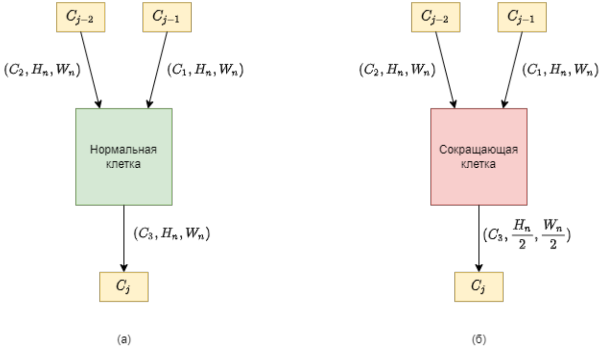
Размер изображения	Число классов	Размер $\mathfrak{D}_{train}$	Размер $\mathfrak{D}_{val}$
$1 \times 28 \times 28$	10	60000	10000

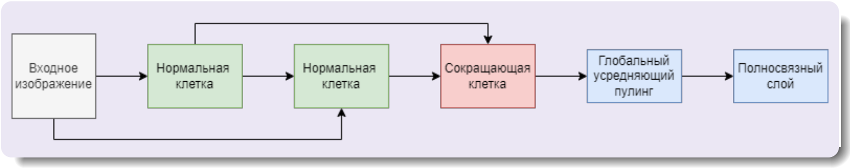
Априорные распределения

- $q_{\mathbf{w}}(\mathbf{w} \,|\, \boldsymbol{\theta}_{\mathbf{w}}) \sim \mathcal{N}(\mathbf{m}_{\mathbf{w}}, \mathbf{A}_{\mathbf{w}}^{-1})$
- $q_{\Gamma}(\boldsymbol{\Gamma} \,|\, \boldsymbol{\theta}_{\Gamma}) \sim$ Gumbel-Softmax$(\alpha_1, \dots \alpha_N)$

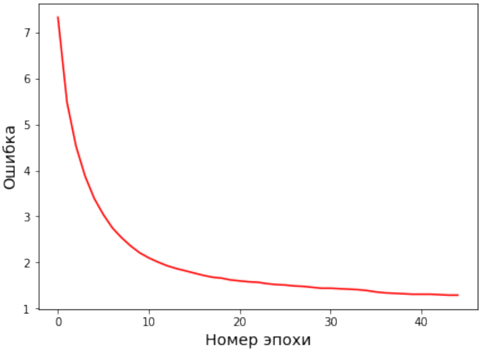


Архитектура поиска



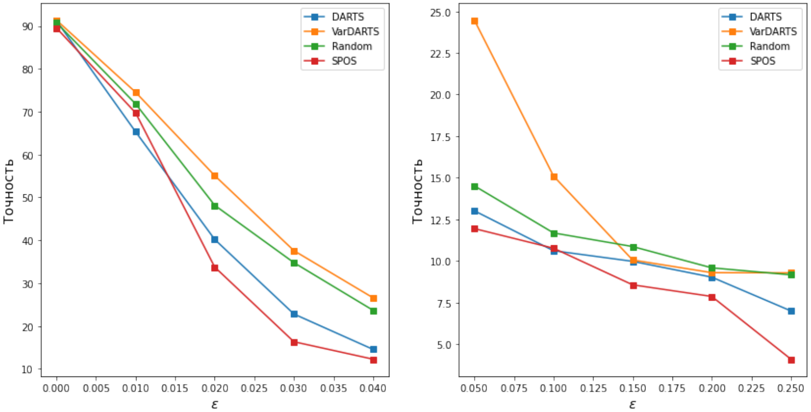


Сравнение с существующими методами

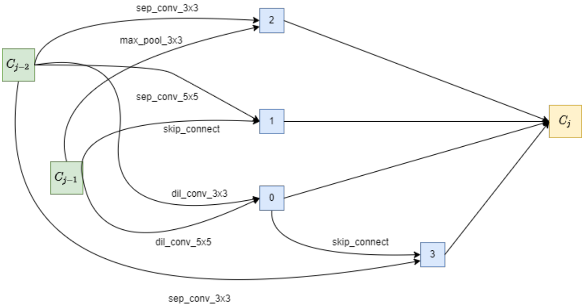


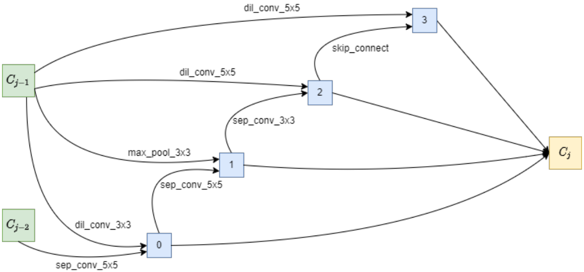
Метод	Accuracy top-1, %	Время обучения, ч
DARTS	91.22	2
Random	90.79	1.5
SPOS	89.54	**0.5**
VarDARTS	**91.36**	5

Сравнение с существующими методами



Обученные клетки





Выносится на защиту

Полученные результаты

- Предложен метод, повышающий устойчивость градиентного ПАНМ, с помощью байесовского вывода
- Предложена теоретическая интерпретация реализованного метода

Дальнейшие исследования

- добавить зависимость распределения параметров модели от её структуры;
- добавить распределения на гиперпараметры модели;
- заменить априорные предположения на другие распределения.