Поиск согласованных нейросетевых моделей в задаче мультидоменного обучения

К. Д. Яковлев[1] О. Ю. Бахтеев[1,2] В. В. Стрижов[1,2]

{iakovlev.kd, bakhteev, strijov}@phystech.edu

[1]Москва, Московский физико-технический институт

[2]Москва, Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН

2022

Цель исследования

Цель

Предложить метод поиска архитектуры модели глубокого обучения в задаче мультидоменного обучения.

Проблема

Модели, не учитывающие разделение выборки на домены, имеют низкую обобщающую способность.

Метод решения

Предлагаемый метод основан на построении мультимодели. Для каждого домена оптимизируется отдельная структура. Также предлагаются два метода регуляризации: структурная и регуляризация пространства скрытых представлений модели.

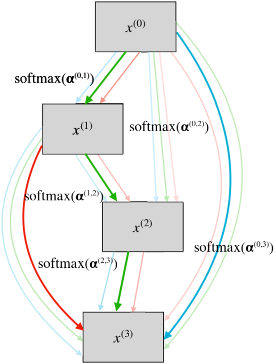
Основная литература

- Hanxiao Liu and Karen Simonyan and Yiming Yang. *DARTS: Differentiable Architecture Search*. CoRR, 2018.
- Wang, Q., Ke, J., Greaves, J., Chu, G., Bender, G., Sbaiz, L., Go, A., Howard, A., Yang, M., Gilbert, J. & Others *Multi-path neural networks for on-device multi-domain visual classification*. CoRR, 2021.
- Yakovlev, K., Grebenkova, O., Bakhteev, O. & Strijov, V. *Neural Architecture Search with Structure Complexity Control*. CoRR, 2022.

Постановка задачи поиска архитектуры

- Архитектура модели представляет собой ориентированный ациклический граф. Каждому ребру ставится в соответствие отображение $\boldsymbol{g}^{(i,j)}$, причем
$$\boldsymbol{x}^{(j)} = \sum_{i<j} \boldsymbol{g}^{(i,j)}(\boldsymbol{x}^{(i)}).$$
- Пусть вектор $\vec{\boldsymbol{g}}^{(i,j)}$ – вектор, составленный из доступных для ребра (i,j) отображений. Пусть вектор $\boldsymbol{\alpha}^{(i,j)}$ – вектор структурных параметров. Смешанная операция
$$\hat{\boldsymbol{g}}^{(i,j)}(\boldsymbol{x}^{(i)}) = \langle \boldsymbol{softmax}(\boldsymbol{\alpha}^{(i,j)}), \vec{\boldsymbol{g}}^{(i,j)}(\boldsymbol{x}^{(i)})\rangle.$$
- Задана выборка $\mathfrak{D} = \mathfrak{D}_{\text{train}} \cup \mathfrak{D}_{\text{val}}$. Задана функция потерь $\mathcal{L}_{\text{train}}$, $\mathcal{L}_{\text{val}}$. Пусть $\boldsymbol{\alpha} = [\boldsymbol{\alpha}^{(i,j)}]$. Пусть $\boldsymbol{w}$ – параметры модели. Двухуровневая задача оптимизации
$$\min_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\boldsymbol{w}^*, \boldsymbol{\alpha}),$$
$$\text{s.t.} \quad \boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \mathcal{L}_{\text{train}}(\boldsymbol{w}, \boldsymbol{\alpha})$$

Архитектура модели



- Смешанная операция:

$$\hat{\boldsymbol{g}}^{(i,j)} = \mathrm{softmax}(\alpha^{(i,j)})_1 \boldsymbol{g}_1^{(i,j)}(\boldsymbol{x}^{(i)}) + \mathrm{softmax}(\alpha^{(i,j)})_2 \boldsymbol{g}_2^{(i,j)}(\boldsymbol{x}^{(i)}) + \mathrm{softmax}(\alpha^{(i,j)})_3 \boldsymbol{g}_3^{(i,j)}(\boldsymbol{x}^{(i)})$$

Постановка задачи поиска архитектуры на мультидоменных данных

- Задано D доменов, $\{\mathfrak{D}_d\}_{d=1}^{D}$, $\mathfrak{D}_d = (\mathbf{X}_d, \mathbf{y}_d)$, $\mathbf{y}_d \subset \mathcal{Y}$. Пусть заданы функции потерь $\mathcal{L}_{\text{train}}^{(d)}$ и $\mathcal{L}_{\text{val}}^{(d)}$. Модель задается разделяемыми между доменами параметрами $\mathbf{w}$. Пусть $\boldsymbol{\alpha}_d$ – вектор структурных параметров домена d, $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_d]_{d=1}^{D}$.
- Задано распределение на доменах $d \sim \mathrm{Categorical}(\mathbf{p})$. Двухуровневая задача оптимизации

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \sum_{d=1}^{D} p_d \mathcal{L}_{\text{valid}}^{(d)}(\mathbf{w}^*, \boldsymbol{\alpha}_d),$$

$$\text{s.t.} \quad \mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{d=1}^{D} p_d \mathcal{L}_{\text{train}}^{(d)}(\mathbf{w}, \boldsymbol{\alpha}_d).$$

- Получение итоговой архитектуры для каждого из доменов:

$$\mathbf{g}^{(i,j)}(.) = \vec{\mathbf{g}}_{k^*}^{(i,j)}, \quad k^* = \arg\max_{k} (\alpha_d^{(i,j)})_k.$$

Предлагаемые регуляризаторы

- Заданы распределения на ребрах графа $P_d^{(i,j)} = \mathrm{Categorical}(\mathbf{softmax}(\boldsymbol{\alpha}_d^{(i,j)}))$. Структурный регуляризатор:

$$\mathcal{L}_{\text{struct}}(\boldsymbol{\alpha}) = \frac{1}{D(D-1)} \sum_{(i,j)} \sum_{d=1}^{D} \sum_{d'=1,\ d'\neq d}^{D} \mathrm{JS}(P_d^{(i,j)} || P_{d'}^{(i,j)}),$$

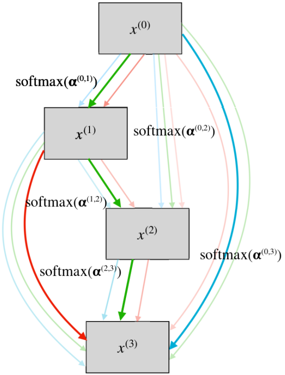
где $\mathrm{JS}(.||.)$ – дивергенция Йенсена—Шеннона.

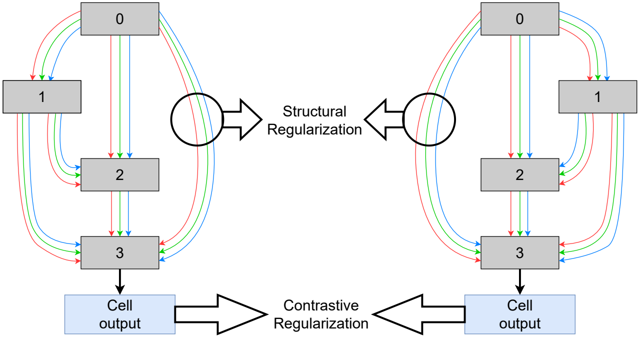
- Заданы наборы объектов для каждого из доменов $\{\mathcal{S}_b^{(d)}\}_{d=1}^{D}$ Регуляризатор пространства скрытых представлений модели:

$$\mathcal{L}_{\text{contr}}(\mathbf{w}, \boldsymbol{\alpha}) = \mathsf{E}_{d,d' \sim \mathrm{Categorical}(\mathbf{p})} \mathsf{E}_{\mathcal{S}_b^{(d)}, \mathcal{S}_b^{(d')}} \mathcal{L}_{\text{triplet}}(\mathbf{w}, \boldsymbol{\alpha}, \mathcal{S}_b^{(d)}, \mathcal{S}_b^{(d')}),$$

где $\mathcal{L}_{\text{triplet}}$ – триплетная функция потерь для скрытых представлений $\mathcal{S}_b^{(d)}, \mathcal{S}_b^{(d')}$.

Построение мультимодели





Задача оптимизации

- Заданы коэффициенты регуляризации $\beta_{\text{trip}} \geq 0,\ \beta_{\text{struct}} \geq 0$. Оптимальный вектор структурных параметров в задаче выбора архитектуры на мультидоменных данных находится из следующей задачи оптимизации:

$$\begin{aligned} \boldsymbol{\alpha}^* &= \arg\min_{\boldsymbol{\alpha}} \sum_{d=1}^{D} p_d \mathcal{L}_{\text{val}}^{(d)}(\mathbf{w}^*, \boldsymbol{\alpha}) + \beta_{\text{trip}} \mathcal{L}_{\text{contr}}(\mathbf{w}, \boldsymbol{\alpha}) + \beta_{\text{struct}} \mathcal{L}_{\text{struct}}(\boldsymbol{\alpha}), \\ \text{s.t.} \quad \mathbf{w}^* &= \arg\min_{\mathbf{w}} \sum_{d=1}^{D} p_d \mathcal{L}_{\text{train}}^{(d)}(\mathbf{w}, \boldsymbol{\alpha}) + \beta_{\text{trip}} \mathcal{L}_{\text{contr}}(\mathbf{w}, \boldsymbol{\alpha}). \end{aligned}$$

Постановка вычислительного эксперимента

- Цель – получение зависимости качества работы мультимодели и количества ее параметров в зависимости от используемого регуляризатора.
- Эксперимент проводится на подвыборке MNIST. В качестве доменов рассматриваются изображения, повернутые на угол, кратный $\pi/2$. Число доменов меняется от 1 до 4. Сравниваются следующие модели: мультимодель со структурной регуляризацией, мультимодель с регуляризацией скрытых представлений, а также модель, не учитывающая разбиение выборки на домены.
- Оценивается средняя точность (accuracy) на тестовой выборке для каждого из доменов. Также приводится количество параметров для каждой модели.

Результаты вычислительного эксперимента

model	accuracy	num. of params
1 domain		
single	60.59	5029
2 domains		
single, union	66.95	6560
multimodel, struct	62.86	**5248**
multimodel, contr	**69.64**	9328
3 domains		
single, union	63.02	5685
multimodel, struct	64.85	**6826**
multimodel, contr	**65.01**	12096
4 domains		
single, union	67.16	6560
multimodel, struct	63.15	**7872**
multimodel, contr	**67.98**	13685

Заключение

- Рассмотрена задача поиска архитектуры модели глубокого обучения на мультидоменных данных. Задача рассматривалась как задача мультимоделирования.
- Предложены два метода регуляризации: регуляризация структуры и регуляризация пространства скрытых представлений модели.
- Продемонстрирована работоспособность предлагаемого решения. При использовании первого регуляризатора мультимодель имеет меньшее число параметров. При использовании второго регуляризатора модель имеет лучшую точность классификации.
- В дальнейшем планируется провести вычислительный эксперимент на задаче мультиязычного языкового моделирования.