Дифференцируемый алгоритм поиска архитектуры с контролем сложности

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

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

[1]Москва, Московский физико-технический институт

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

2021

Цель исследования

Цель

Предложить метод поиска архитектуры модели глубокого обучения с контролем сложности модели.

Проблема

Модели глубокого обучения имеет избыточное число параметров. Поиск архитектуры на дискретном множестве является вычислительно сложной задачей.

Метод решения

Предлагаемый метод основан на непрерывной релаксации. Структурные параметры задаются гиперсетью, зависящей от коэффициента, задающего сложность архитектуры. Под гиперсетью понимается модель, порождающая параметры оптимизируемой модели.

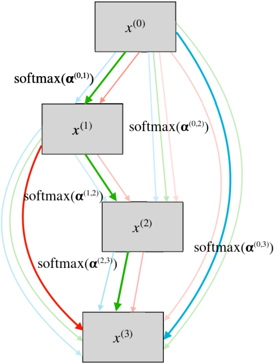
Основная литература

- Hanxiao Liu and Karen Simonyan and Yiming Yang. *DARTS: Differentiable Architecture Search*. CoRR, 2018.
- David Ha and Andrew M. Dai and Quoc V. Le. *HyperNetworks*. CoRR, 2016.
- Grebenkova, O., Bakhteev, O.Y., Strijov, V. *Variational deep learning model optimization with complexity control* 2021
- Jang, E., Gu, S., Poole, B. *Categorical reparameterization with gumbel-softmax*. CoRR, 2016.

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

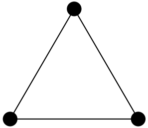


- Смешанная операция:

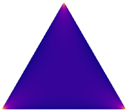
$$\hat{\boldsymbol{g}}^{(i,j)} = \text{softmax}(\alpha^{(i,j)})_1 \boldsymbol{g}_1^{(i,j)}(\boldsymbol{x}^{(i)}) + \text{softmax}(\alpha^{(i,j)})_2 \boldsymbol{g}_2^{(i,j)}(\boldsymbol{x}^{(i)}) + \text{softmax}(\alpha^{(i,j)})_3 \boldsymbol{g}_3^{(i,j)}(\boldsymbol{x}^{(i)})$$

Распределение гумбель-софтмакс

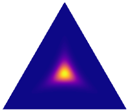
Распределение гумбель-софтмакс определено на симплексе. Пусть $\boldsymbol{X} \sim \mathcal{GS}(\boldsymbol{\alpha}, t)$, где $\boldsymbol{\alpha} \in \mathbb{R}^n_{++}$, $t > 0$.



$t \to 0$



$t = 0.995$



$t = 5.0$

Контроль сложности с помощью гиперсети

- Смешанная операция

$$\hat{\boldsymbol{g}}^{(i,j)}(\boldsymbol{x}^{(i)}) = \langle \gamma^{(i,j)}, \vec{\boldsymbol{g}}^{(i,j)} \rangle, \quad \gamma^{(i,j)} \sim \mathcal{GS}(\boldsymbol{exp}(\boldsymbol{\alpha}^{(i,j)}), t).$$

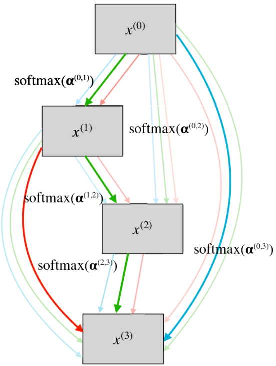
- Пусть $\Lambda \subset \mathbb{R}$ – множество параметров, задающих сложность. Гиперсеть – это параметрическое отображение из множества Λ во множество структурных параметров модели

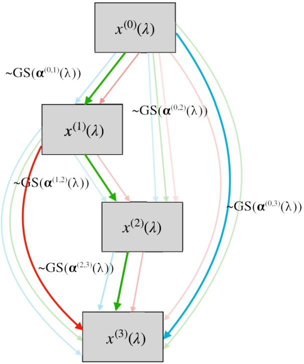
$$\boldsymbol{\alpha}^{(i,j)} = \boldsymbol{\alpha}^{(i,j)}(\lambda, \boldsymbol{a}^{(i,j)}), \quad \lambda \in \Lambda.$$

- В работе используется кусочно-линейная гиперсеть

$$\boldsymbol{\alpha}^{(i,j)}(\lambda, \boldsymbol{a}^{(i,j)}) = \sum_{k=0}^{N-1} \left(\frac{\lambda - t_k}{t_{k+1} - t_k} \boldsymbol{a}_k^{(i,j)} + \left(1 - \frac{\lambda - t_k}{t_{k+1} - t_k} \right) \boldsymbol{a}_{k+1}^{(i,j)} \right) I[\lambda \in [t_k, t_{k+1}]]$$

DARTS с использованием гиперсети





Структурные параметры порождаются гиперсетью, зависящей от коэффициента, задающего сложность архитектуры. Структурные параметры подчинены распределению Gumbel-Softmax.

Задача оптимизации

- Пусть вектор $\boldsymbol{n}(\vec{\boldsymbol{g}}^{(i,j)})$ хранит количество параметров каждого отображения. Регуляризатор, контролирующий сложность

$$\lambda \sum_{(i,j)} \langle \textbf{softmax}\left(\alpha^{(i,j)}(\lambda, \boldsymbol{a}^{(i,j)})\right), \boldsymbol{n}(\vec{\boldsymbol{g}}^{(i,j)})\rangle.$$

- Пусть задано распределение $p(\lambda)$ на Λ. Пусть $\boldsymbol{\gamma} = [\gamma^{(i,j)}]$. Параметры $\boldsymbol{a} = [\boldsymbol{a}^{(i,j)}]$ гиперсети находятся из задачи оптимизации

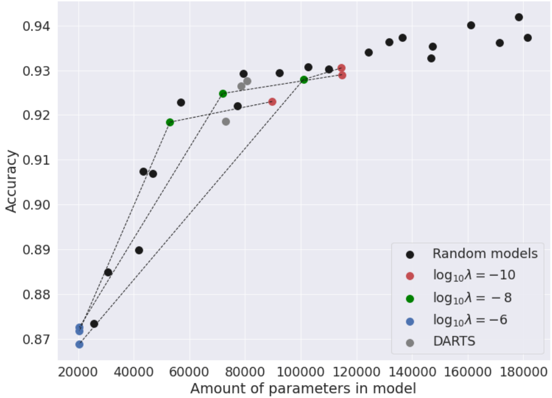
$$\min_{\boldsymbol{a}} \mathsf{E}_{\lambda \sim p(\lambda)} \left(\mathsf{E}_{\gamma} \mathcal{L}_{\text{val}}(\boldsymbol{w}^*, \gamma) + \lambda \sum_{(i,j)} \langle \textbf{softmax}\left(\alpha^{(i,j)}(\lambda, \boldsymbol{a}^{(i,j)})\right), \boldsymbol{n}(\vec{\boldsymbol{g}}^{(i,j)})\rangle \right),$$

$$\text{s.t.} \quad \boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \mathsf{E}_{\lambda \sim p(\lambda)} \mathsf{E}_{\gamma} \mathcal{L}_{\text{train}}(\boldsymbol{w}, \gamma).$$

Постановка вычислительного эксперимента

- Цель эксперимента – получение зависимости обобщающей способности модели от количества её параметров.
- Вычислительный эксперимент проводится на выборке Fashion-MNIST. Сравниваются архитектуры, полученные с помощью DARTS, предлагаемого метода и случайные архитектуры.
- Модель состоит из трех ячеек. Коэффициент $\lambda \sim \mathcal{U}[10^{-10}, 10^{-6}]$. Во время обучения температура распределения гумбель-софтмакс понижалась от 1 до 0.2.

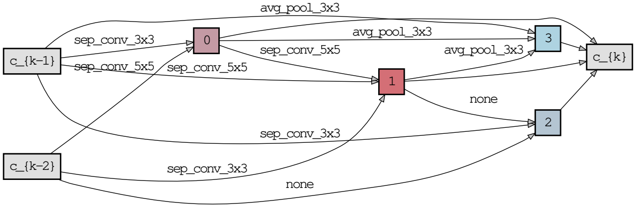
Результаты вычислительного эксперимента



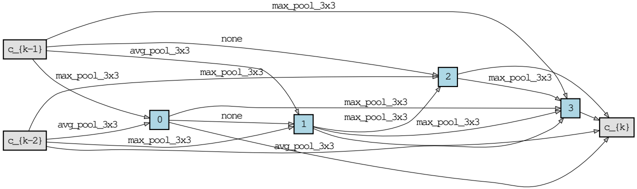
Зависимость качества классификации от количества параматров модели.

Предложенный метод позволяет контролировать сложность архитектуры, изменяя коэффициент регуляризации λ.

Полученные архитектуры



(a) Архитектура, полученная при $\lambda = 10^{-10}$.



(b) Архитектура, полученная при $\lambda = 10^{-6}$.

Чем больше коэффициент регуляризации λ, тем проще получаемая архитектура.

Заключение

- Предложен метод, позволяющий контролировать сложность модели в процессе поиска архитектуры.
- Метод обладает тем свойством, что изменение сложности итоговой модели происходит изменением коэффициента, задающего сложность архитектуры, без дополнительного обучения.
- Также результаты показывают, что данный метод сопоставим по качеству с DARTS.