

Дифференцируемый алгоритм поиска архитектуры модели с контролем её сложности

Константин Дмитриевич Яковлев

Московский физико-технический институт

Курс: Автоматизация научных исследований
(практика, В. В. Стрижов)/Группа 874

Эксперт: В. В. Стрижов

Консультант: О. С. Гребенькова, О. Ю. Бахтеев

2021

Цель исследования

Цель

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

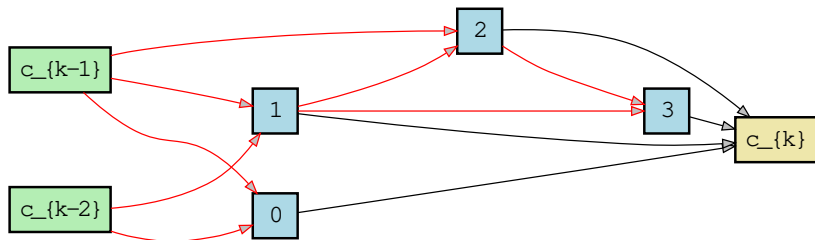
Проблема

Семейство моделей глубокого обучения имеет избыточное число параметров. Использование моделей, работающих с дискретной архитектурой, является вычислительно сложной задачей.

Метод решения

В основе метода лежит дифференцируемый алгоритм поиска архитектуры (DARTS). Гиперсеть выступает в качестве функции релаксации. Гиперсеть – это модель, генерирующая параметры другой модели.

DARTS с использованием линейной гиперсети



- ▶ Смешанная операция:

$$\hat{o}^{(i,j)}(x) = \sum_{o \in \mathcal{O}} \alpha_o^{(i,j)} o(x)$$

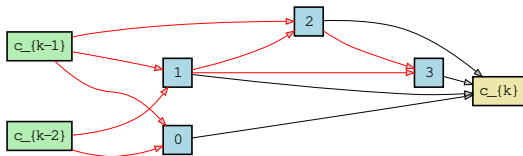
- ▶ Линейная гиперсеть задает параметры архитектуры:

$$\alpha = \lambda b_1 + b_2,$$

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.
-  Xiangxiang Chu, Tianbao Zhou, Bo Zhang, and Jixiang Li. *Fair DARTS: Eliminating Unfair Advantages in Differentiable Architecture Search*. CoRR, 2019.

Постановка задачи поиска архитектуры модели



► **Узлы:**

$\{x^{(i)}\}_{i=1}^N$ – узлы ориентированного ациклического графа.

► **Значение в текущем узле:**

$$x^{(j)} = \sum_{i < j} o^{(i,j)}(x^{(i)}) \quad o^{(i,j)} \in \mathcal{O}$$

► **Смешанная операция:**

$$\hat{o}^{(i,j)}(x) = \sum_{o \in \mathcal{O}} \alpha_o^{(i,j)} o(x)$$

► **Вектор параметров архитектуры:**

$$\alpha = [\alpha^{(i,j)}]$$

Линейная гиперсеть задает параметры архитектуры

Гиперсеть

$$G : \Lambda \times \mathbb{U} \rightarrow \mathbb{A}.$$

\mathbb{A} – пространство параметров архитектуры, \mathbb{U} – множество параметров гиперсети. Λ – множество параметров, контролирующие сложность модели.

- ▶ Вектор параметров архитектуры определяется гиперсетью:

$$\alpha = \lambda b_1 + b_2, \quad [b_1, b_2]^T \in \mathbb{U}, \quad \lambda \in \Lambda = \mathbb{R}, \quad \alpha \in \mathbb{A}.$$

- ▶ Задача оптимизации:

$$\begin{aligned} & \min_{\alpha} \mathcal{L}_{\text{val}}(w^*, \alpha), \\ & \text{s.t.} \quad w^* = \arg \min_w \mathcal{L}_{\text{train}}(w, \alpha). \end{aligned}$$

$\mathcal{L}_{\text{train}}(w, \alpha)$, $\mathcal{L}_{\text{val}}(w, \alpha)$ – кросс-энтропия.

Алгоритм DARTS

-
- 1: Для каждого узла создадим смешанную операцию $\hat{o}^{(i,j)}$, параметризованную $\alpha^{(i,j)}$
 - 2: **while** алгоритм не сошелся **do**
 - 3: обновить α : $\nabla_{\alpha} \mathcal{L}_{\text{val}}(w - \xi \nabla_w \mathcal{L}_{\text{train}}(w, \alpha), \alpha)$
 - 4: обновить веса w : $\nabla_w \mathcal{L}_{\text{train}}(w, \alpha)$
 - 5: **end while**
 - 6: получить окончательную архитектуру из полученного α
-

Получение дискретной архитектуры:

$$o^{(i,j)} = \arg \max_{o \in \mathcal{O}} \alpha_o^{(i,j)}.$$

Вычислительный эксперимент

Цель

Получение зависимости качества работы предложенного метода от параметра гиперсети $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Эксперимент проводится на выборке MNIST.

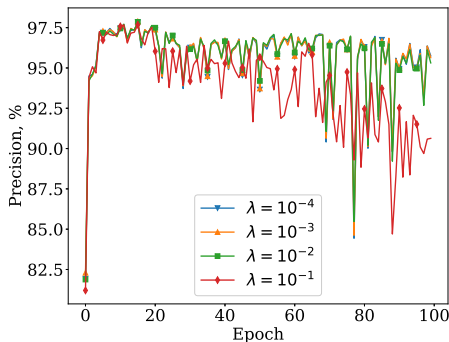
Критерии качества

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

В качестве $\mathcal{L}_{\text{train}}(w, \alpha)$, $\mathcal{L}_{\text{val}}(w, \alpha)$ рассматривается кросс-энтропия.

Качество DARTS с использованием гиперсети

Зависимость качества модели от числа прошедших эпох для разных параметров λ гиперсети.



Для $\lambda = 10^{-1}$ качество модели заметно хуже, чем для других значениях λ . Также для каждой эпохи и для каждого $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ качество модели практически не меняется.

Сравнение предложенного метода и DARTS с L_1 регуляризатором

Значения качества моделей на валидации.

Модель	Precision, %			
	эпоха 30	эпоха 50	эпоха 70	эпоха 100
Гиперсеть, $\lambda = 10^{-1}$	96,3500	95,4800	94,1200	90,6333
Гиперсеть, $\lambda = 10^{-2}$	95,8900	96,7167	91,0633	95,3133
Гиперсеть, $\lambda = 10^{-3}$	95,9133	96,6200	90,6400	95,6400
Гиперсеть, $\lambda = 10^{-4}$	95,9733	96,5367	90,4067	95,7533
DARTS+ L_1 , $\lambda = 10^{-1}$	86,6000	87,3967	89,3433	89,5067
DARTS+ L_1 , $\lambda = 10^{-2}$	84,1333	90,6333	91,9100	92,7333
DARTS+ L_1 , $\lambda = 10^{-3}$	97,6533	97,5800	98,0833	97,9467
DARTS+ L_1 , $\lambda = 10^{-4}$	98,5800	98,8867	98,9467	99,2400

Для модели DARTS с использованием гиперсети качество на валидации уменьшается с ростом номера эпохи.

- ▶ Предложен метод, позволяющий контролировать сложность модели в процессе поиска архитектуры.
- ▶ Метод обладает тем свойством, что изменение сложности итоговой модели происходит заменой параметра λ гиперсети без дополнительного обучения.
- ▶ Также результаты показывают, что данный метод сопоставим по качеству на валидационной выборке с DARTS.