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

18 января, 2022

Мотивация

- Многие современные архитектуры нейросетевых моделей, созданные экспертами вручную, не демонстрируют наилучшее качество на разных наборах данных.
- Поиск архитектуры нейронной сети (англ. Neural Architecture Search, NAS) — это процесс автоматизации проектирования архитектуры нейронной сети. Система NAS получает на вход набор данных и тип задачи (классификация, регрессия и т.д.), и на выходе дает архитектуру модели.
- Предлагается реализовать процедуру автоматической генерации структуры нейронной сети, которая обобщала бы наилучшим образом конкретный набор данных (имела наилучшее качество).

Базовый подход

Пусть дано зафиксированное пространство поиска $\mathcal{O}$. В работе [**DBLP:journals/corr/abs-1806-09055**] представлен алгоритм DARTS, использующий идею релаксации дискретного пространства поиска в непрерывное с помощью операции softmax:

$$o^{(i,j)}(x) = \sum_{\wr \in \mathcal{O}} \frac{exp(\alpha_o^{i,j})}{\sum_{\prime\prime \in \mathcal{O}} exp(\alpha_{o\prime}^{i,j})} \cdot o(x)$$

Задачей NAS в таком случае становится выучивание параметров $\alpha^{i,j}$. В конце, для получения итоговой архитектуры, на каждом ребре архитектуры выбирается операция, удовлетворяющая условию

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

Проблемы

- Пространство поиска является дискретным набором заранее заданных операций (пулинги, свертки заданных размеров и т.п.). Процедура поиска оптимальной структуры сети на дискретном пространстве является очень долгой и затратной по вычислительным ресурсам [**DBLP:journals/corr/ZophVSL17**].
- Существующие подходы выведены в условии независимости распределений структуры и параметров модели, что в общем случае неверно.
- Градиентные подходы NAS страдают от застревания в локальных минимумах, из-за чего моделью предпочитается неоптимальная операция в рассматриваемом ребре.

Постановка задачи

- Задан набор данных $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ где каждому входу $\mathbf{x}_i \in \mathbf{X}$ соответствует целевая переменная $y_i \in \mathbf{Y}$. Элементы $(\mathbf{x}_i, y_i)$ являются случайными величинами, взятыми из совместного распределения $\mathbf{p}(\mathbf{x}, y)$. Назовём через Γ - суперграф архитектуры, $\mathcal{A} \subset \Gamma$ - архитектура, $\mathbf{w}_\mathcal{A} \sim p(\mathbf{w}_\mathcal{A})$ - её структурные параметры. Через $\mathbf{w} \sim p(\mathbf{w})$ обозначим параметры модели.
- Вероятностная модель задается следующим образом

$$p(\mathbf{w}, \mathbf{w}_\mathcal{A}|\mathbf{X}, \mathbf{y}, \theta) = p(\mathbf{w}|\mathbf{w}_\mathcal{A}, \mathbf{X}, \mathbf{y}, \theta) \cdot p(\mathbf{w}_\mathcal{A}|\mathbf{X}, \mathbf{y}, \theta).$$

- В качестве оптимальных параметров $\mathbf{w}^*, \mathbf{w}^*_\mathcal{A}$ предлагается использовать те, которые максимизируют их совместное условное распределение.
- Таким образом ставится следующая оптимизационная задача:

$$\mathbf{w}^*, \mathbf{w}^*_\mathcal{A} = \arg max p(\mathbf{w}, \mathbf{w}_\mathcal{A}|\mathbf{X}, \mathbf{y}, \theta^*),$$

$$p(\theta|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{X}, \theta) \cdot p(\theta).$$

Вариационный вывод распределений структур и параметров модели.

Основной целью является вывод апостериорного распределения параметров модели при помощи теоремы Байеса. Главной проблемой является интеграл в знаменателе теоремы, который крайне сложно адекватно вычислить в силу высокой размерности пространства параметров модели. В связи с этим предлагается оценить его с помощью вариационного вывода.

Правдоподобие модели:

$$\log P(\mathfrak{D}|\mathcal{A}) = \int_{\mathbf{w}\in\mathcal{W}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathcal{A})d\mathbf{w}$$

Вариационная оценка:

$$\log P(\mathfrak{D}|\mathcal{A}) = \int_{\mathbf{w}\in\mathcal{W}} q(\mathbf{w})\frac{p(\mathfrak{D},\mathbf{w}|\mathcal{A})}{q(\mathbf{w})}d\mathbf{w} - \int_{\mathbf{w}\in\mathcal{W}} q(\mathbf{w})\frac{p(\mathbf{w}|\mathfrak{D},\mathcal{A})}{q(\mathbf{w})}d\mathbf{w} \approx$$

$$\int_{\mathbf{w}\in\mathcal{W}} q(\mathbf{w})\frac{p(\mathfrak{D},\mathbf{w}|\mathcal{A})}{q(\mathbf{w})}d\mathbf{w} =$$

$$\int_{\mathbf{w}\in\mathcal{W}} q(\mathbf{w})\frac{\log p(\mathbf{w}|\mathcal{A})}{q(\mathbf{w})}d\mathbf{w} + \int_{\mathbf{w}\in\mathcal{W}} q(\mathbf{w})\log p(\mathfrak{D}|\mathcal{A},\mathbf{w})d\mathbf{w} =$$

$$\mathcal{L}_{\mathbf{w}}(\mathfrak{D},\mathcal{A},\mathbf{w}) + \mathcal{L}_E(\mathfrak{D},\mathcal{A}).$$

Первое слагаемое - дивергенция Кульбака-Лейблера, второе - матожидание правдоподобия выборки. Минимизируется выведенная величина

Текущее состояние

- На текущий момент продолжается вывод теоретических результатов.
- Ставятся первые эксперименты на наборе данных CIFAR-10.

Технические детали проводящихся экспериментов

- Реализация программной части проходит на языке Python с помощью библиотеки для поиска нейросетевых архитектур nni.retiarii, pytorch.
- Для проведения экспериментов и формирования оптимальной структуры (назовем ее ячейкой) используется набор данных CIFAR-10. В дальнейшем предполагается попробовать обучить архитектуру, являющейся композицией нескольких таких ячеек, на наборе данных ImageNet и сравнить полученные метрики качества с существующими SOTA моделями.
- Попробовать в качестве априорного распределения параметров модели распределение Дирихле.

Эксперименты

to be continued...

Дальнейшая работа

- Провести вычислительные эксперименты, используя текущие подходы к аппроксимации распределений. Сравнить полученные показатели качества, а также сравнить робастность генерируемых моделей относительно adversarial атак. Провести сравнительный анализ с существующими подходами NAS.
- Уточнить теоретический вывод апостериорных вероятностных распределений для весов и структуры моделей.