Регуляризация траектории оптимизации параметров модели глубокого обучения на основе дистилляции знаний

М. Горпинич, О. Ю. Бахтеев, В. В. Стрижов
gorpinich4@gmail.com; bakhteev@phystech.edu; strijov@ccas.ru

Исследуется задача оптимизации параметров модели глубокого обучения. Предлагается обобщение методов дистилляции, заключающееся в градиентной оптимизации гиперпараметров. На первом уровне оптимизируются параметры модели, на втором — гиперпараметры, задающие вид оптимизационной задачи. Исследуются свойства оптимизационной задачи и различные виды оператора оптимизации. Предложенное обобщение оптимизации позволяет производить дистилляцию модели с лучшими эксплуатационными характеристиками и за меньшее количество итераций оптимизации. Иллюстрировать применение комбинации данных подходов предлагается с помощью вычислительного эксперимента на выборке CIFAR-10.

Ключевые слова:

DOI:

1 Введение

В работе рассматривается процесс оптимизации глубоких нейросетей. Данная задача требует больших вычислительных мощностей и является затратной по времени. В данной работе предлагается метод оптимизации, позволяющий улучшить эксплуатационные характеристики модели, а также ускорить ее сходимость к точке оптимума.

Предлагается обобщение метода оптимизации на основе дистилляции знаний. Рассматривается *модель-учитель* более сложной структуры, которая была обучена на выборке. Модель более простой структуры предлагается оптимизировать путем переноса знаний модели учителя на более простую модель, называемую *моделью-учеником*, при этом ее качество будет выше по сравнению с качеством, полученным после оптимизации на той же выборке. Примером применения данного подхода является [1]. В работе [2] предложен подход к дистилляции знаний, позволяющий переносить знания на модель с архитектурой, значительно отличающейся от архитектуры модели-учителя.

Предлагается представление задачи в виде двухуровневой оптимизации. На первом уровне оптимизации происходит оптимизация параметров модели, на втором уровне — ее гиперпараметров. Данный подход описан в работах [3–5]. В работе [3] рассматривается жадный градиентный метод оптимизации гиперпараметров, в работе [4] сравниваются различные градиентные методы оптимизации гиперпараметров, а также метод случайного поиска.

В работе рассматривается вид задачи оптимизации, а также различные виды оператора оптимизации. Данный подход с использованием нейросети LSTM описан в работе [6].

Вычислительный эксперимент проводится на выборке изображений CIFAR-10.

2 Название раздела

Данный документ демонстрирует оформление статьи, подаваемой в электронную систему подачи статей `http://jmlda.org/papers` для публикации в журнале «Машинное обучение и анализ данных». Более подробные инструкции по стилевому файлу `jmlda.sty` и использованию издательской системы LaTeX 2$_\varepsilon$ находятся в документе

authors-guide.pdf. Работу над статьёй удобно начинать с правки TEX-файла данного документа.

Обращаем внимание, что данный документ должен быть сохранен в кодировке UTF-8 without BOM. Для смены кодировки рекомендуется пользоваться текстовыми редакторами Sublime Text или Notepad++.

2.1 Название параграфа

Разделы и параграфы, за исключением списков литературы, нумеруются.

3 Заключение

Желательно, чтобы этот раздел был, причём он не должен дословно повторять аннотацию. Обычно здесь отмечают, каких результатов удалось добиться, какие проблемы остались открытыми.

Литература

[1] *Hinton Geoffrey E., Vinyals Oriol, Dean Jeffrey.* Distilling the knowledge in a neural network // CoRR, 2015. Vol. abs/1503.02531. URL: http://arxiv.org/abs/1503.02531.

[2] *Passalis Nikolaos, Tzelepi Maria, Tefas Anastasios.* Heterogeneous knowledge distillation using information flow modeling // CVPR. — 2020. P. 2336–2345.

[3] *Luketina Jelena, Berglund Mathias, Greff Klaus, Raiko Tapani.* Scalable gradient-based tuning of continuous regularization hyperparameters // CoRR, 2015. Vol. abs/1511.06727. URL: http://arxiv.org/abs/1511.06727.

[4] *Bakhteev Oleg Yu., Strijov Vadim V.* Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Ann. Oper. Res, 2020. Vol. 289. No. 1. P. 51–65.

[5] *Maclaurin Dougal, Duvenaud David, Adams Ryan P.* Gradient-based hyperparameter optimization through reversible learning // CoRR, 2015. Vol. abs/1502.03492. URL: http://arxiv.org/abs/1502.03492.

[6] *Andrychowicz Marcin, Denil Misha, Colmenarejo Sergio Gomez, Hoffman Matthew W., Pfau David et al.* Learning to learn by gradient descent by gradient descent // CoRR, 2016. Vol. abs/1606.04474. URL: http://arxiv.org/abs/1606.04474.

Received