Анализ методов оптимизации глубоких нейронных сетей с использованием распределенных вычислений

Тони Кастильо Мартин

студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: cmtony4@gmail.com

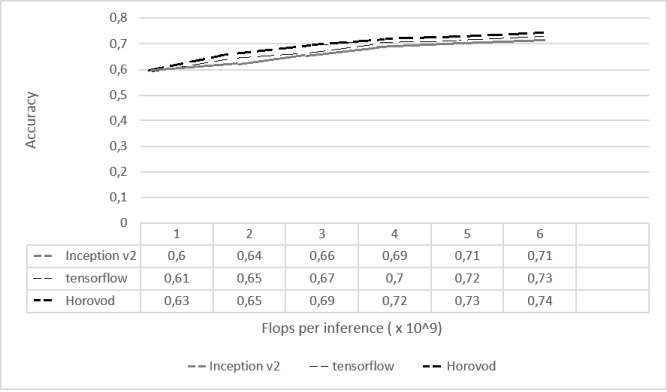
Научный руководитель *— к.ф.-м.н., Доцент Попова Нина Николаевна*

За последние несколько лет достижения в области глубокого обучения привели к огромному прогрессу в обработке изображений, распознавании речи и прогнозировании. Однако, алгоритмы машинного обучения являются дорогостоящими (с точки зрения времени и ресурсов) для проведения настройки модели с нуля для конкретных приложений. Некоторые автоматизированные подходы пытаются ускорить этот процесс путем поиска подходящих существующих моделей. Примерами таких подходов являются Neural Architecture Search и AdaNet[1], использующих машинное обучение для поиска пространства проектирования, чтобы найти улучшенные архитектуры. В качестве альтернативы можно использовать существующую архитектуру для решения аналогичной проблемы и в один прием оптимизировать ее для решения поставленной задачи.

В докладе рассматривается метод оптимизации модели глубокой нейросети, реализованный на базе пакета Morp-net[2], и проводится сравнение этого метода с распределенной реализацией метода оптимизации, выполненной с использованием фреймворков Tensorflow и Horovod, принимая существующую нейронную сеть в качестве входных данных и создавая новую нейронную сеть, которая меньше, быстрее и предлагает лучшую производительность, адаптированную к новой проблеме. Оптимизация осуществляется путем сокращения и расширения сети, выявляя неэффективные нейроны и удаляя их из сети, применяя разреживающего регуляризатор [3], так что функция полной потери сети включает стоимость каждого нейрона. В процессе обучения оптимизатор оценивает стоимость ресурсов при расчете градиентов и таким образом узнает, какие нейроны являются ресурсоэффективными, а какие могут быть удалены.

Экспериментальное исследование рассмотренных методов проводится на примере сети Inception v2 [4] в качестве модели нейронной

входной сети и наборов данных ImageNet и JFT. В докладе обсуждаются основные полученные результаты. Во- первых, было достигнуто лучшее использование имеющихся ресурсов в процессе обучения модели в предлагаемой реализации с использованием Horovod в качестве основы. Было увеличено на 10% количество обрабатываемых изображений в секунду, а также снижено на 8% (Tensorflow) и 12% (Horovod) количество используемых вещественных операций без ущерба для точности обучения. Реализация распределенных вычислений с использованием Horovod также показала улучшение точности на 2,5% и 1,3% по отношению к наборам данных JTF и ImageNet соответственно.



	1	2	3	4	5	6
Inception v2	0,6	0,64	0,66	0,69	0,71	0,71
tensorflow	0,61	0,65	0,67	0,7	0,72	0,73
Horovod	0,63	0,65	0,69	0,72	0,73	0,74

Рис. 1.Поведение достигаемой точности модели в зависимости от числа выполненных операций.

На рисунке 1 представлено поведение достигаемой точности в зависимости от числа выполненных операций для различных вариантов построения модели: заданного базового варианта Inception v2; модели, построенной с использованием morph-net в Tensorflow и модели, полученной с использованием распределенных вычислений с помощью Horovod. Применение morph-net снижает число используемых операций (FLOP) на 8% (TF) и 12% (HRV) соответственно по отношению к базовому варианту в дополнение к увеличению точности на 2,5% при тех же затратах.

Литература

1. Barret Zoph, Quoc V. Le., Neural Architecture Search with Reinforcement Learning, International Conference on Learning Representations, 2017.
2. Ariel Gordon, Elad Eban, Ofir Nachum, Bo Chen, Hao Wu, Tien-Ju Yang,Edward Choi MorphNet: Fast & Simple Resource-Constrained Structure Learning of Deep Networks, 2018.
3. Rosasco, Lorenzo; Poggio, Tomasso. "A Regularization Tour of Machine Learning". 2014, MIT-9.520 Lectures Notes.
4. Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna.Rethinking Z. The inception architecture for computer vision.InProceedings of the IEEE Conference on Computer Visionand Pattern Recognition, 2016, P. 2818—2826,