

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

О Т Ч Е Т
по учебной практике
на факультете компьютерных наук НИУ ВШЭ

Выполнил:

Студент группы БПМИ205

Подпись

А.А.Макогон

И.О.Фамилия

02.02.2021

Дата

Руководитель практики

Колесниченко Елена Юрьевна

(ФИО руководителя практики)

департамент больших данных и информационного поиска ФКН ВШЭ, доцент

Подразделение ФКН, должность

Дата — 2022

Подпись

Москва 2022

Содержание

1	Аннотация	3
2	Введение	3
3	Современные подходы	3
4	Large Scale Distributed Deep Networks	3
4.1	Downpour SGD	4
4.2	Sandblaster L-BFGS	4
4.3	Достигнутые результаты и проверка эффективности	4
	Список литературы	6

1 Аннотация

Тут будет аннотация

2 Введение

Графические процессоры способны быстро обрабатывать большие объемы данных, но имеют меньшую точность по сравнению с центральными процессорами, но тем не менее точности, достигаемой с помощью GPU, хватает для решения многих задач, в том числе для большинства задач машинного обучения. В ML очень важна скорость обучения моделей, особенно она критична для громоздких моделей с большим количеством параметров, например, для нейронных сетей, поэтому возникает резонное желание ускорить процесс тренировки моделей за счет распределённого обучения и использования графических процессоров.

3 Современные подходы

Вот так работает цитирование [1]

4 Large Scale Distributed Deep Networks

Начнем изучение темы распределённого обучения моделей со статьи “Large Scale Distributed Deep Networks” [2], в которой делается большой шаг в сторону распределенного обучения нейронных сетей.

Нейронная сеть состоит из нескольких слоев нейронов и синапсов, связей между нейронами. Каждый нейрон принимает на вход какие-то небольшие данные, производит над ними работу и передает полученный результат по синапсу следующему нейрону. У каждого синапса есть вес, который влияет на то, какой вклад внесет результат вычислений одного нейрона на результат вычислений следующего.

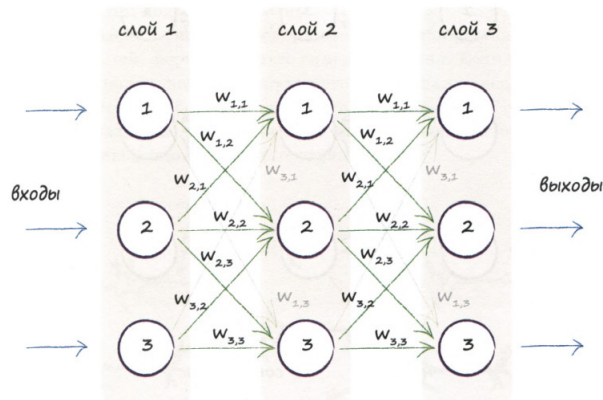


Рис. 1: Визуализация нейронной сети

Веса синапсов - это одни из параметров нейронной сети и в задачу обучения сети входит подбор весов синапсов так, чтобы минимизировать ошибку модели. Таким образом, подбор весов нейронной сети сводится к задаче минимизации функции ошибки.

Основной вклад, который внесла статья в развитие распределённого обучения моделей, заключается в том, что авторы разработали два распределенных метода оптимизации и провели серии экспериментов, показывающих эффективность этих моделей по сравнению с нераспределёнными, работающими на CPU или на сооптимизированном GPU.

В статье авторы описывают два распределенных метода оптимизации, которые можно применять для поиска весов нейронной сети: Downpour SGD и Sandblaster L-BFGS. Оба метода были разработаны для фреймворка DistBelief, поддерживающего параллелизм моделей, когда модель делится на части и каждая часть обучается на одном и том же датасете, и параллелизм данных, когда копия модели обучается на разных поднаборах датасета.

4.1 Downpour SGD

Downpour SGD - асинхронная вариация градиентного спуска SGD. В основе распределенности этой модели лежит параллелизм данных. Несколько копий модели запускаются на разных репликах, каждой реплике отведен поднабор параметров, который она будет менять во время обучения. Модели обмениваются обновленными параметрами с помощью центрального сервера параметров, который хранит текущее состояние всех параметров, которые используются на разных репликах.

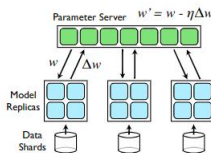


Рис. 2: Визуализация Downpour SGD

Преимуществом Downpour SGD по сравнению с синхронной распределенной вариацией SGD является его устойчивость к отказам машин, поскольку каждая реплика индивидуально выполняет свою работу и по завершении не ждет остальных, чтобы обновить значения параметров. Основной проблемой асинхронного подхода является невозможность поддержания консистентности значений параметров для каждой реплики. Поскольку каждая реплика действует независимо, то нет никакой гарантии, что в любой момент времени к параметрам на разных машинах было применено одинаковое количество обновлений или что обновления были применены в одном и том же порядке.

4.2 Sandblaster L-BFGS

Авторы рассматривают распределенное обучение модели L-BFGS, заключающееся в параллелизме данных.

Ключевые идеи распределенности для алгоритма L-BFGS, которые предлагают авторы статьи:

- распределенное хранение параметров
- выделение процесса координатора, который не имеет доступа к параметрам модели и выступает в роли оркестратора всех реплик, выполняющих обучение

Координатор назначает каждой из реплик модели небольшую часть работы, которую она может выполнить независимо, и каждый раз, когда какая-то реплика заканчивает свою работу, координатор назначает ей новую часть. Такой подход позволяет хорошо использовать быстрые машины и исключает ожидания самой медленной реплики, чтобы продолжить процесс. Результаты работы, а также кэшируемая информация сохраняются локально на каждой реплике и не отсылаются в центральный сервер, что позволяет запускать обучение достаточно больших моделей без оверхеда на отправку и сбор всех параметров.

4.3 Достигнутые результаты и проверка эффективности

Авторы статьи применяют описанные методы оптимизации для двух задач глубинного обучения: распознавание объектов на изображениях и акустическая обработка для распознавания речи. Ниже приведены графики

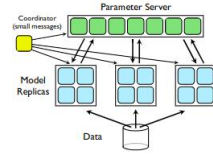


Рис. 3: Визуализация Sandblaster L-BFGS

для задачи обработки речи. Авторы сравнивают три оптимизационных модели, которые тренируются распределённо: Downpour SGD, Downpour SGD w Adagrad, Sandblaster L-BFGS с двумя моделями, которые были обучены на центральном процессоре одной реплики (черная линия) и на оптимизированном графическом процессоре (розовая пунктирная линия).

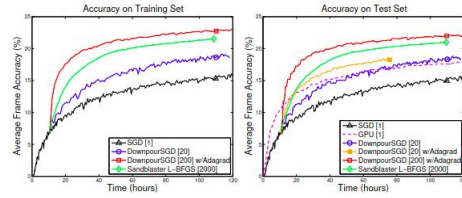


Рис. 4: Зависимость точности модели от времени обучения

Весьма значительным результатом является то, что авторы статьи применили разработанный метод оптимизации к модели с 1.7 миллиардами параметров и получили точность классификации 15% на кросс валидации, что является относительным улучшением более чем на 60% по сравнению с лучшей производительностью в этой задаче классификации.

Список литературы

1. A Unified Architecture for Accelerating Distributed DNN Training in Heterogeneous GPU/CPU Clusters / Y. Jiang [и др.] // 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20). — USENIX Association, 11.2020. — С. 463—479. — ISBN 978-1-939133-19-9. — URL: <https://www.usenix.org/conference/osdi20/presentation/jiang>.
2. Large Scale Distributed Deep Networks / J. Dean [и др.] // Advances in Neural Information Processing Systems 25 (NIPS 2012). — NeurIPS Proceedings, 11.2012. — URL: <https://papers.nips.cc/paper/2012/hash/6aca97005c68f1206823815f66102863-Abstract.html>.