Stochastic Training is Not Necessary for Generalization

Jonas Geiping, Micah Goldblum, Phillip E. Pope, Michael Moeller, Tom Goldstein

Доклад подготовили:

Полина Гусева
Максим Першин
Николай Карташев
Александра Сендерович

БПМИ181

Докладчик

Полина Гусева

Обобщающая способность SGD

SGD с малым размером батча:

- Дает шумную оценку градиента
- Сходится к точке с лучшей обобщающей способностью, чем SGD с большим размером батча

Помогают ли шумные градиенты обобщающей способности?

Обобщающая способность SGD

Из теории у SGD два основных преимущества над GD:

- Стабильность и скорость сходимости
- Шум добавляет смещение, ведущее к плоским минимумам

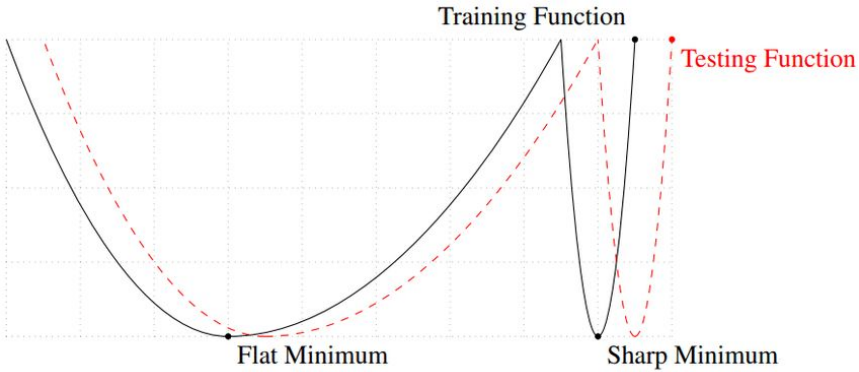
При этом смещение можно вывести явно!

Гипотеза:

GD со специальной регуляризацией и без шума может достичь тех же результатов, что и SGD.

Широкие и острые минимумы

- Широкие минимумы ведут к лучшей обобщающей способности.
- Меньше батч — шире минимум[1]
- Шум в градиенте помогает не застрять в остром минимуме
- Гауссовский шум в градиенте[2] =>



$$\mathrm{d}\theta_t = -\nabla\left(L(\theta_t) + \frac{\tau}{4}||\nabla L(\theta)||^2\right)\mathrm{d}t + \sqrt{\tau\Sigma_t}\,\mathrm{d}W_t,$$

L — функция потерь

τ — длина шага

W_t — винеровский случайный процесс

Σ_t — ковариационная матрица шума

1. Keskar et al. Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. ICLR 2016.
2. Li et al. Stochastic Modified Equations and Adaptive Stochastic Gradient Algorithms. ICML 2017.

Явная регуляризация в SGD

- Неявная регуляризация SGD может быть описана явно[1]

$$L(\theta) + \frac{\tau}{4|\mathcal{B}|} \sum_{B \in \mathcal{B}} \left\| \frac{1}{|B|} \sum_{x \in B} \nabla \mathcal{L}(x, \theta) \right\|^2, \qquad \mathcal{B} \text{ — множество батчей}$$

- Можно использовать, чтобы увеличить размер батча, но при этом сохранить свойства малых батчей.
- Вероятно, может сломаться при слишком больших батчах[2].

1. Barrett and Dherin. Implicit Gradient Regularization. ICLR 2020.
2. Smith et al. On the Origin of Implicit Regularization in Stochastic Gradient Descent. ICLR 2020.

Регуляризация для GD

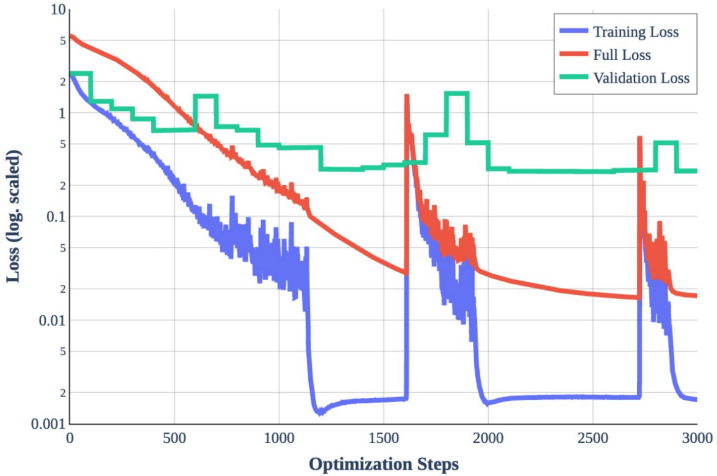
- Сильно упрощаем прошлую формулу для GD

$$L(\theta) + \frac{\tau}{4} ||\nabla L(\theta)||^2,$$

- Гессиан в градиенте регуляризатора => приближаем конечными разностями
- **Смысл:** минимизация градиентов на малых подмножествах

Стабилизация обучения

- GD => edge of stability[1], то есть нестабильное обучение
- Рекомендуется малая длина шага => теряем регуляризатор.
- **Решение:** медленный разогрев + агрессивный клиппинг градиентов



1. Cohen et al. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. ICLR 2021.

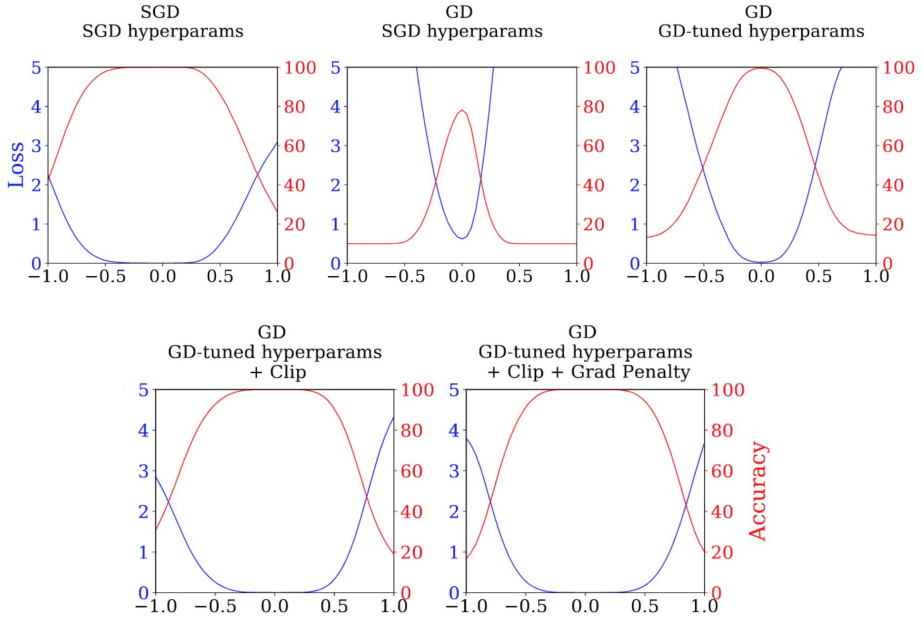
Результаты экспериментов. Аугментации

- Данные - CIFAR-10, модель ResNet-18
- SGD бейзлайн использует моментум и weight decay. Объекты не перемешиваются. Размер батча 128.
- Аугментации horizontal flip и random crop

Experiment	Mini-batching	Epochs	Steps	Modifications	Val. Accuracy %
Baseline SGD	✓	300	117,000	-	95.70(±0.11)
Baseline FB	✗	300	300	-	75.42(±0.13)
FB train longer	✗	3000	3000	-	87.36(±1.23)
FB clipped	✗	3000	3000	clip	93.85(±0.10)
FB regularized	✗	3000	3000	clip+reg	95.36(±0.07)
FB strong reg.	✗	3000	3000	clip+reg+bs32	95.67(±0.08)
FB in practice	✗	3000	3000	clip+reg+bs32+shuffle	95.91(±0.14)

Поверхность функции потерь

- Параметры меняются вдоль случайного направления



Результаты экспериментов. Другие модели

- Проверили на других сверточных сетях, результат стабильный
- Подобранные гиперпараметры можно переиспользовать
- В таблице доля верных ответов на валидации

Experiment	ResNet-18	ResNet-50	Resnet-152	DenseNet-121
Baseline SGD	95.70	95.83	95.98	95.84
Baseline FB	75.42	54.32	58.62	76.87
FB train longer	87.36	83.31	91.02	82.06
FB clipped	93.85	94.15	91.41	93.44
FB regularized	95.36	95.51	95.82	95.47
FB strong reg.	95.67	96.05	96.01	95.81
FB in practice	95.91	96.56	96.76	95.86

Результаты экспериментов. Без аугментаций

- Аугментации добавляются до обучения — из каждой картинки генерируется 10 новых.
- Утверждается, что аугментации не повлияли на обобщающую способность.

Experiment	Fixed Dataset	Mini-batching	Steps	Modifications	Val. Accuracy
Baseline SGD	CIFAR-10	✓	117,000	-	84.32(±1.12)
Baseline SGD*	CIFAR-10	✓	117,000	-	90.07(±0.48)
FB strong reg.	CIFAR-10	✗	3000	clip+reg+bs32	89.17(±0.24)
Baseline SGD	10× CIFAR-10	✓	117,000	-	95.20(±0.09)
FB	10× CIFAR-10	✗	3000	-	88.44(−)
FB strong reg.	10× CIFAR-10	✗	3000	clip+reg+bs32	95.11(−)

Итоги

- Результаты GD могут быть сравнимы с SGD, но надо постараться.
- Стохастический шум не критичен для обобщающей способности

Рецензент

Максим Першин

Практик

Николай Карташев

История публикаций

- Статья подана на ICLR 2022, решение будет принято 24 января
- Баллы от рецензентов: 6, 10, 8, 5, 6 (где 6 - слабое принятие, 5 - слабое отклонение)
- На момент доклада доступны две версии, в них нет принципиальных отличий, были слегка изменены формулировки, расширен обзор литературы, добавлено несколько фраз про практическую неэффективность полного градиентного спуска, добавлен апендикс со значениями функций потерь на тренировочном и валидационном наборе данных для разных аугментаций

Авторы

- Jonas Geiping - University of Siegen в статье, University of Maryland, College Park на Google Scholar
- Micah Goldblum - University of Maryland, College Park
- Phillip E. Pope - University of Maryland, College Park
- Michael Moeller - University of Siegen
- Tom Goldstein - University of Maryland, College Park

Практически все они занимаются в основном отравлением данных (Data Poisoning), а также распределенным обучением, адверсальными примерами и приватностью данных.

Основные источники

- Train longer, generalize better: Closing the generalization gap in large batch training of neural networks, by Elad Hoffer, Itay Hubara, and Daniel Soudry. Advances in Neural Information Processing Systems, 30, 2017. - Побочные эффекты увеличения размера батча можно побороть лучшим подбором гиперпараметров
- Measuring the Effects of Data Parallelism on Neural Network Training, by Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. in Journal of Machine Learning Research, 20(112):1–49, 2019. ISSN 1533-7928. - О необходимости неинтуитивного подбора размера шага и коэффициента моментума для больших батчей
- Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability by Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. In International Conference on Learning Representations, September 2020. - О феномене немонотонной не эффективной оптимизации при полном спуске

Цитирования

- В некотором смысле статья-конкурент “Never Go Full Batch (in Stochastic Convex Optimization)”

  В статье доказывают, что если делать *честный* градиентный спуск, то скорость сходимости асимптотически не менее четвертой степени, когда для SGD скорость не более чем квадратична

- 5 статей цитируют статью более касательно - статьи посвящены новым методам оптимизации и теоретическим рассуждениям об оптимизации нейросетей

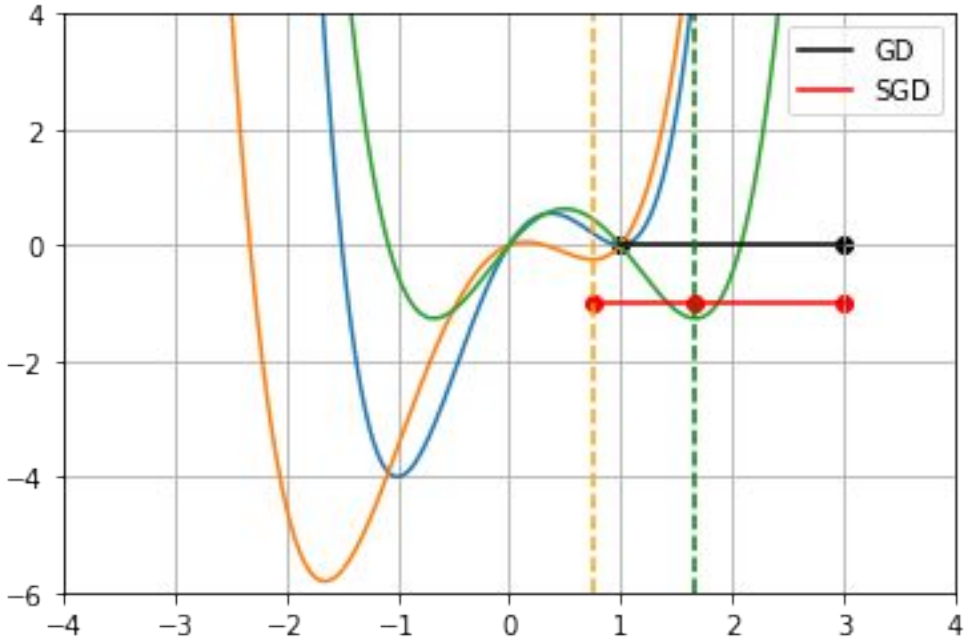
Будущие направления исследований

- Современные методы подобраны под эффективное обучение с помощью SGD. Было бы интересно посмотреть на архитектуры, созданные с Full-Batch GD в уме, и на то насколько хуже для них работает стохастический спуск
- Проведенные на наборах данных для компьютерного зрения эксперименты дают серьезные основания для выдвинутой гипотезы, но набор экспериментов на более широком наборе областей может дать больше информации об особенностях полного градиентного спуска - хотелось бы увидеть эксперименты на текстовых данных, особенно учитывая сложности с аугментацией и существующие проблемы с тем как метод максимизации правдоподобия работает на задачах обработки последовательностей

Хакер

Александра Сендерович

GD в двумерном случае



Эксперименты

- Батч вместе с сетью помещается в память видеокарты:
  - Сеть – 1 свёрточный и 1 линейный слой
  - MNIST:
    - Увеличение числа эпох в 20 раз (с 5 до 100) даёт нужный результат
  - FashionMNIST:
    - Увеличение числа эпох в 60 раз (с 5 до 300) даёт нужный результат
  - Клиппинг с подобранными параметрами – улучшение на 0.5 процента
  - GP ничего не даёт
- Cifar10:
  - Сеть – 2 свёртки с пулингом + 2 линейных слоя
  - Делаем backward на каждом батче, а step – после эпохи
  - На небольшом числе эпох клиппинг и GP ничего не дают
- Код авторов понятный и работающий (проверено на MNIST)