Stochastic Training is Not Necessary for Generalization

Практик-исследователь Николай Карташев

1) Статья была подана на ICLR 2022, официальное решение о принятии еще не принято, но баллы выглядят многообещающе (6, 10, 8, 5, 6).

   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.
2) Авторы:
   a) Jonas Geiping - University of Siegen в статье, University of Maryland, College Park на Google Scholar
   b) Micah Goldblum - University of Maryland, College Park
   c) Phillip E. Pope - University of Maryland, College Park
   d) Michael Moeller - University of Siegen
   e) 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.
3) Я бы сказал что тут нет непосредственно сильного влияния, но выделил бы важные промежуточные шаги - статьи о обучении с большими батчами
   a) 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
   b) 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
   c) 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
4)
   a) 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

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

5) Я бы не сказал что статья в прошлом пункте прямой конкурент - скорее там дается ограничение GD с другой стороны: если использовать только методы первого порядка и не использовать батчевые статистики, градиентный спуск не будет так же эффективно сходится
6) Попробовать подобрать модели, которые лучше работают для GD чем для SGD, так как прошлые 10 лет архитектуры в глубинном обучении подбирались для максимизации эффективности SGD. Было бы интересно посмотреть архитектуры, подобранные под полнобатчевый GD.
7) Его нет, статья почти полностью теоретическая.