Stochastic Training is Not Necessary for Generalization

В статье описана роль SGD и GD в аспекте обобщающей способности модели. В частности, авторы сравнивают SGD с full-batch GD вкупе с применением явной регуляризации к последнему. На примере нескольких архитектур и CIFAR-10 авторы показывают, что full-batch подход может достичь результатов SGD, если применить к нему ряд улучшений (регуляризация, клиппинг, увеличенный LR и так далее).В результате авторы подкрепляют свои предположения экспериментами, где можно наблюдать, что GD с правильной регуляризацией действительно доходит до качества SGD.

Положительные стороны:

- Статья хорошо написана, в аппендиксе много подробных экспериментов.
- В целом, авторы учли все замечания после публикации и опубликовали вторую версию статьи – сделали более понятный абстракт, добавили еще экспериментов, а также обновили заключение, с оговоркой, что обновление на полном батче все еще сильно хуже по производительности: “Nonetheless, our training routine is highly inefficient compared to SGD (taking far longer run time), and stochastic optimization remains a great practical choice for practitioners in most settings.”
- Большая ценность, хоть и больше эвристик, чем теоретических обоснований: авторы показывают, что в целом обобщающая способность модели с обычным GD и явной НЕстохастической регуляризацией примерно сравнима с SGD, где эта регуляризация неявная

Отрицательные стороны:

- В первой версии казалось, будто авторы считают, что GD в данном случае предпочтительнее SGD, в новой версии это поправили, с оговоркой про скорость при использовании GD, но не указали, насколько это все же дольше.
- Не очень понятно, как сложится картина на других (например, больших) наборах данных – в статье это не описано.
- Также из статьи не очень ясно, можно ли добавить какую-нибудь явную регуляризацию поверх SGD, чтобы улучшить модель.
- Не понятно, почему увеличение learning rate в два раза дает результат лучше, нет какого-то явного обоснования

Воспроизводимость:

- Просто с точки зрения кода, в целом авторы статьи достаточно хорошо расписали подробности экспериментов, и выглядит, как будто их нетрудно реализовать, но

- Это все тяжело с точки зрения необходимости иметь достаточные вычислительные мощности

Итог:

- Оценка – 7
- Уверенность – 4