Рецензия на статью:
“GradInit: Learning to Initialize Neural Networks for Stable and Efficient Training”
Пахалко Илья, БПМИ181

Авторы статьи предлагают новый алгоритм инициализации нейронных сетей. Метод не зависит от конкретной нейросетевой архитектуры, поскольку любую модель рассматривает абстрактно - в виде блоков параметров, соединённых между собой. Принцип работы метода заключается в умножении каждого из таких блоков на некоторый скалярный вес; веса подбираются градиентным спуском по целевым данным. Исследователи утверждают, что предложенный алгоритм обладает рядом преимуществ относительно уже существующих: от меньшего расходования вычислительных ресурсов до качественного улучшения результатов работы сети.

Достоинства:
- Первые три секции статьи хорошо читаемы. Словесное описание метода в целом понятно (за исключением пары моментов); математическое описание метода не вызвает вопросов: обозначения определены чётко и явно, алгоритмическое описание полное и ясное.
- Ясно описаны потенциальные преимущества по сравнению с конкурирующими методиками: метод не зависит от конкретной архитектуры; заявлено, что предложенный метод требует меньше вычислительных мощностей, чем MetaInit, а также за меньшее число итераций даёт лучшее качество по сравнению с warmup.
- Непосредственный замер итогового качества (accuracy score для изображений, BLEU для машинного перевода) позволяет сделать вывод, что метод работает как минимум не хуже всех остальных из рассмотренных.
- Достаточно обширные эксперименты для изображений: рассмотрены популярные архитектуры адекватного размера (в основном, семейства ResNet), с различными модификациями (с/без BatchNorm, с/без Skip Connections). Эксперименты в основном проводились на CIFAR-10, однако также есть и запуски на Imagenet.
- Методология описана достаточно подробно: с воспроизведением графиков и качественных результатов не должно возникнуть проблем.

Недостатки:
- Четвёртая секция статьи, посвящённая постановке экспериментов, вызывает у меня много вопросов. Во-первых, результаты тяжело сравнивать между собой: для некоторых экспериментов выходит, что описание результатов представлено в основной части статьи, а сравниваемые между собой графики находятся в приложении, к тому же разделены между собой несколькими страницами. Также иногда авторы сравнивают конкретные показатели между собой, но они не отмечены на графиках явно (например, измеряется отношение значений в начале и в конце кривой, при этом сам график изображён в логарифмической шкале). Во-вторых, многие эксперименты, на мой взгляд, описаны сумбурно: не очень ясно, что доказывает тот или иной построенный график. Возможно, это следствие визуальной несостоятельности, но в любом случае, хотелось бы ясности.
- Эксперименты с текстовыми данными проведены лишь на одном датасете и с оригинальной архитектурой Transformer. Пожалуй, для современной статьи хотелось бы запуск и более современной архитектуры, близкой к SOTA - тем более, если заявлена широкая применимость метода (отказ от warmup).

- Не вполне ясен выбор конфигурации использованного оборудования: почти все эксперименты были проделаны на 1x GTX 2080Ti; кроме части с Imagenet, для которой использовалась связка 4x 2080Ti. В таком случае вопрос: если всё же был доступ к четырём видеокартам, почему на них не был произведён запуск более тяжеловесной текстовой модели?
- Утверждение о превосходстве в расходовании вычислительных ресурсов по сравнению с MetaInit оставлены без численного подтверждения.

Общее впечатление от статьи смешанное. С одной стороны, предложен новый метод инициализации, который неплохо себя показывает с точки зрения итогового качества по сравнению с ближайшими конкурентами; применимость алгоритма также обоснована: метод не зависит от архитектуры, позволяет работать с данными разного рода. С другой стороны - презентация результатов большинства экспериментов, призванных объяснить внутренний принцип работы метода, оставляет желать лучшего. Также хотелось бы видеть более подробные эксперименты с текстовыми данными - другие архитектуры, датасеты. Моя оценка: 6 (marginally above acceptance threshold), уверенность - 3 из 5. Готов был бы повысить оценку, если бы эксперименты были описаны яснее.

Рецензенты на OpenReview присвоили статье общую оценку 7 (оценки 6, 6, 7, 7). В числе достоинств отмечают широкую область применения метода, новизну. Многие замечания авторам удалось устранить: например, были включены кривые качества целиком (вместо сравнения первой и последней эпох). Ревьюеры также отмечают зависимость описания экспериментов от графиков, представленных в приложении, однако в финальной версии статьи этот недостаток так и не был исправлен, по всей видимости. Также из замечаний ревьюеров: не перечислен диапазон перебора гиперпараметра tau; превосходство по качеству относительно других методов инициализации не превышает одного процента accuracy.