Sharpness-aware minimization for efficiently improving generalization

Анализ контекста

Кириллов Дмитрий

Работа [2] была представлена в качестве спотлайта на ICLR 2021. Оценки рецензентов были однозначно положительными. Большая часть замечаний была связана с концептуальной схожестью представленного подхода и ряда других публикаций. После первой редакции авторы заметно увеличили количество описанных экспериментов и добавили сравнения с некоторыми аналогами.

Все авторы аффилированы с программой исследований Google.
Pierre Foret — резидент Google AI, получил степень магистра в Беркли. Вместе с Behnam Neyshabur и Hossein Mobahi являлся организатором соревнования по предсказанию обобщающей способности нейросетей в рамках NIPS 2020.
У Ariel Kleiner до значительного перерыва можно отметить работы, связанные с высокопроизводительным машинным обучением и большими данными.
Среди авторов у Behnam Neyshabur наибольшее число публикаций связанных с изучением обобщающей способности и регуляризации нейросетей.
В том числе совместная с Hossein Mobahi работа [3], в которой авторы исследовали различные подходы к измерению генерализации моделей. В их числе были метрики оперирующие и PAC-Bayesian оценкой, и понятием "остроты" локального минимума. Поэтому разумным видится, что текущая работа является логичным продолжением исследований последних двух авторов с упором на эмпирическое исследование применимости предложенного метода в различных задачах.

Важной частью работы является вывод теоремы, которая выражает разрыв между популяционной функцией ошибки и ошибкой на обучающей выборке через "остроты" оптимума. Авторы явно указывают, что вдохновлялись ранней публикацией [1] при ее доказательстве. Общая идея подхода всех методов, связанных с "остротой" оптимума функции потерь состоит в том, что ошибка на обучающей выборке отличается от популяционной ошибки. И как следствие, если в окрестности оптимума функция потерь значительно меняется, то даже при маленьком изменении ландшафта при переходе к тестовой ошибке, выбранная точка может оказаться заметно субоптимальной. Для теоретического обоснования обычно прибегают к `PAC-Bayesian` оценкам и их следствиям. Рассматриваемая работы выделяется масштабным исследованием применимости предложенного метода оптимизации при относительно скромном объеме формальных выводов.

Существует несколько подходов к улучшению обобщающей способности моделей. Один из рецензентов отметил сходство предложенного алгоритма с известным [4] еще с 1976 года "экстраградиентным" методом. В другой работе [7] из иных предпосылок авторы также пришли к похожему шагу оптимизатора. Более того модели, сошедшиеся к острому минимуму функции потерь, могут

быть более подвержены специфичным состязательным атакам [6]. Что в некотором смысле можно рассматривать как плохую обобщающую способность.

На ICML 2021 в качестве постера было опубликовано прямое продолжение [5] работы, предлагающее модификацию метода, инвариантную к шкалированию весов модели. 60 других статей ссылаются на SAM. Подавляющее большинство авторов не использует предложенный метод оптимизации для собственных экспериментов, но приводят сравнение собственных результатов с опубликованными в рассматриваемой статье.

В рамках собственного исследования авторы реализовали предложенный метод и провели эксперименты с помощью фреймворка `JAX` на языке `Python`. Однако заметно более популярной стала реализация для библиотеки `PyTorch` (которая также поддерживает ASAM [5] модификацию). Существует репозиторий позволяющий применять метод и в рамках библиотеки `TensorFlow`. Усилиями сообщества метод SAM был реализован для всех наиболее популярных библиотек, используемых для глубинного обучения. Более того, технически метод может быть использован при оптимизации очень разнообразных моделей. Поэтому уже сейчас любой желающий относительно легко может использовать предложенный метод оптимизации в собственных проектах.

Список литературы

[1] Niladri S Chatterji, Behnam Neyshabur и Hanie Sedghi. “The intriguing role of module criticality in the generalization of deep networks”. в: *arXiv preprint arXiv:1912.00528* (2019).

[2] Pierre Foret и др. “Sharpness-Aware Minimization for Efficiently Improving Generalization”. в: *CoRR* abs/2010.01412 (2020). arXiv: 2010.01412. URL: https://arxiv.org/abs/2010.01412.

[3] Yiding Jiang и др. “Fantastic generalization measures and where to find them”. в: *arXiv preprint arXiv:1912.02178* (2019).

[4] G. M. Korpelevich. “The extragradient method for finding saddle points and other problems”. в: 1976.

[5] Jungmin Kwon и др. “ASAM: Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks”. в: *CoRR* abs/2102.11600 (2021). arXiv: 2102.11600. URL: https://arxiv.org/abs/2102.11600.

[6] Xu Sun и др. “Exploring the vulnerability of deep neural networks: A study of parameter corruption”. в: *arXiv preprint arXiv:2006.05620* (2020).

[7] Colin Wei и Tengyu Ma. “Improved Sample Complexities for Deep Networks and Robust Classification via an All-Layer Margin”. в: *CoRR* abs/1910.04284 (2019). arXiv: 1910.04284. URL: http://arxiv.org/abs/1910.04284.