Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time

Mitchell Wortsman, Gabriel Ilharco et al.
Рецензия Дениса Козлова, 191

Суть работы

Авторы предлагают новый способ файн-тьюна больших предобученных моделей, основанный на усреднении весов нескольких дообученных моделей с разными гиперпараметрами. Таким подходом авторы достигают SOTA точности на ImageNet не увеличив затраты на применение модели.

Контекст

Статья появилась на arXiv в марте 2022 года и участвовала постером на ICML 2022. Авторы статьи — студенты и сотрудники Вашингтонского, Колумбийского и Тель-Авивского университетов, а также сотрудники Google Brain и Meta AI Research. Часть авторов пишет таким составом не первый раз: очень похожий состав встречался в более ранней статье «Robust fine-tuning of zero-shot models» (CVPR 2022). Авторы уже несколько лет работают и интересуются большими моделями, в частности занимаются разработкой OpenCLIP. Вероятно, в своих предыдущих работах они получили большой опыт ландшафтами потерь и смогли сделать выводы в «Robust fine-tuning…» и в этой работе.

Повлиявшие статьи

В основном идеи статьи не новы, новизну представляет их смешение. Пожалуй, самые важные источники это:

- «What is being transferred in transfer learning?» (Neyshabur et al.): В этой работе авторы приходят к выводу что, если файн-тьюнить предобученную модель с разными параметрами (даже с разными наборами данных!), то веса модели сходятся в одну «область» (basin). Они демонстрируют это интерполируя веса модели из одного результата в другой и обнаруживают, что заметного падения качества не наблюдается — в отличие от интерполяции весов моделей, обученных с нуля, у которых веса сходятся в разные «области», и поэтому при интерполяции заметно сильное падение качества модели.
  В «Model soups…» авторы пользуются этим свойством файн-тьюна предобученных моделей и исследуют, как смешивать веса моделей, чтобы достичь наилучших результатов.
- «No One Representation to Rule Them All: Overlapping Features of Training Methods» (Gontijo-Lopes et al.): В этой работе авторы демонстрируют, что модели, обученные с разными гиперпараметрами, находят разные обобщения в данных — и поэтому ошибаются в разных местах. Делая ансамбль из всего двух таких по-разному обученных моделей, авторы получают большой прирост в точности относительно каждой индивидуальной модели.

В «Model soups…» исследователи тоже перебирают большой набор гиперпараметров, чтобы «супам» было проще достичь высокой точности.

Цитирования статьи

Статья достаточно свежая, поэтому ее не успели сильно доработать и процитировать. Большинство цитирований приходят из-за того, что в «Model soups…» достигнут рекорд точности на ImageNet, поэтому это сильный бейзлайн, с которым многие статьи сравнивают свои результаты. Тем не менее, статьи по теме тоже начинают появляться. Пожалуй, самое заметное продолжение — статья «Patching open-vocabulary models by interpolating weights» от этих же авторов, в которой продолжают свои исследования вокруг весов после файн-тьюна моделей и предлагают способ как интерполяцией весов достичь хороших результатов на сложных задачах для zero-shot моделей при этом не теряя в качестве на уже хорошо выполняемых задачах.

Конкуренты статьи

Файн-тьюн больших предобученных моделей — крайне актуальная тема в последнее время, поэтому неудивительно, что уделяется много внимания различным способам их дообучения. К конкурентам я могу отнести работы от других авторов, которые тоже занимаются смешиванием весов дообученных моделей для достижения тех или иных целей.

- В статье «Merging Models with Fisher-Weighted Averaging» исследуются альтернативный способ смешения весов: не среднее значение, а зависящий от информации Фишера. Такой способ требует больше вычислений, но имеет шанс достичь результатов лучше, чем просто линейная комбинация весов нескольких моделей.
- Авторы «Fusing finetuned models for better pretraining» предлагают способ переиспользовать существующие файн-тьюны моделей, чтобы из них составить новую базовую модель, файн-тьюн которой, судя по экспериментам, имеет шанс достичь лучших результатов, чем файн-тьюн просто базовой модели. В статье авторы не предлагают способа как автоматически выбирать конкретные модели для комбинации, но демонстрируют, что в этом направлении могут быть обнаружены интересные результаты.

И вообще идея переиспользования весов (а не получения своих весов каждый раз, как это происходит сейчас) сейчас активно развивается. Хочу отметить профессора университета Северной Калифорнии Colin Raffel, который объявил свои намерения и область интереса лаборатории в блоге «A Call to Build Models Like We Build Open-Source Software».

Сильные стороны работы

Работа получилась отличная! Авторы опытные, прошли с ней на ICML, достигли нового SOTA. Авторами предоставлен и код, и веса, что, к сожалению, сейчас далеко не всегда происходит. Кода немного, но он хорошо читается, а на все вопросы к нему авторы исправно и быстро отвечают.

Авторы уже давно работают над схожими задачами и продолжили ими заниматься даже после выхода «Model soups…», что показывает их заинтересованность и подчеркивает знания и опыт.

Большим удивлением оказались теоретические доводы в статье, в которых более формально объясняются причины наблюдаемого поведения. Математические формулы и выводы подкреплены экспериментами.
Статья помогает решить проблему, что очень часто дообученные модели выбрасываются или не имеют никакого применения кроме одной крайне специфичной задачи, и что для достижения хороших результатов иногда приходится прибегать к дорогостоящим ансамблям больших моделей.

Слабые стороны работы

Пожалуй неправильно называть «слабой стороной» — но это точно не «сильная сторона» — что статья по сути не представляет особой научной новизны, а просто элегантно и просто соединяет уже полученные ранее выводы. Хотя, с другой стороны, наверное, именно такая концентрированность позволила углубиться в эксперименты, перебить SOTA, написать математическое обоснование результатов.
Мне (и многим комментаторам в интернете) показалось недостаточным количество экспериментов про смешивание супов. Почти вся статья разбирает только uniform soup и greedy soup — два крайне простых метода — и хочется увидеть побольше внимания другим методам.

Предлагаемые улучшения

Мне кажется, будет интересным улучшением как раз провести побольше экспериментов и исследований по поводу других стратегий смешивания моделей. Посмотреть на разные параметры Learned soup, ввести новые стратегии, может быть для более простых моделей попробовать провести полный перебор жадных супов, чтобы получить оценки максимально достижимой точности, иметь точку для сравнения.
Также вероятно будет интересным посмотреть на поведение сильно меньших моделей. Если поведение там будет сильно отличаться, это может оказаться интересной темой для исследования.
Последнее предложение: посмотреть на то, что будет если постараться дообучить уже какой-то суп. А если сделать несколько супов и их всех немного дообучить и попробовать смешать еще раз — получится «суп второго порядка», который может быть имеет шанс оказаться еще лучше!

Практическое применение

Мне кажется, что эта статья может пригодиться в компаниях или лабораториях. Когда авторы файн-тьюнят большие модели для своих внутренних целей, у них может оставаться немало обученных, но не идеальных моделей: у которых был немного не тот набор данных, где были чуть-чуть другие гиперпараметры… Да и чекпоинты финальной модели тоже доступны. Все эти модели являются файн-тьюнами поверх одних и тех же предобученных весов, поэтому идея статьи применима. Вполне есть шанс, что результаты получится улучшить, что может отразиться в бизнес-показателях.