Рецензия на статью
Robust fine-tuning of zero-shot models

Шишков Алексей, БПМИ192

Для начала небольшой повторение того, что было в статье. Что было до неё? Раньше была возможность дообучить модель на датасет, приобретя в качестве, но потеряв в обобщающей способности модели. Как статья устраняет эту проблему? Последообучения предлагается брать не последний чекпоинт, а комбинацию fine-tuned и не fine-tuned результатов, получая лучшую модель, как с точки зрения устойчивости и обобщающей способности, так и с точки зрения качества на целевом наборе данных.

Немного общих фактов про статью и её авторов. Во-первых, можно сказать, что статья недавняя – первая версия вышла в сентябре 2021 года (последняя – меньше полугода назад). Последняя версия статьи была представлена на конференции CVPR – конференция о компьютерном зрении и распознавании паттернов. На этой конференции она была номинирована на звание лучшей статьи, но не победила в этой номинации.

Авторы статьи – Mitchell Worstman и Gabriel Illharco из унивеситета Вашингтона. У них было несколько работ, связанных со статьёй Robust fine-tuning of zero-shot models.

Во-первых, у них есть общая работа над Open-source реализацией Clip – OpenClip. В обозреваемой статье использовалась некоторая реализация Clip – с его помощью авторы и проводили эксперименты. В целом это логично – если такая библиотека ими реализована, значит авторы в ней хорошо разбираются, что значительно упрощает эксперименты.

Также у Митчела были статьи про некоторое "глубинное понимание" нейронных сетей, такие, как Learning Network Subspaces и What's Hidden in a Randomy Weighted Neural Network. Я считаю, что такое полное теоретическое изучение внутреннего устройства нейронных сетей позволяет делать более корректные и обоснованные предположения для экспериментов, а значит может повлиять и на конечный вид статьи.

Среди других статей, которые повлияли на обозреваемую статью, нельзя не отметить статью Averaging Weights Leads to Wider Optima and Better Generalization, которую мы разбирали ранее. В ней также говорится про улучшение генерализации модели путём усреднения весов нескольких чекпоинтов, и в обозреваемой статье присутствует множество её упоминаний. Также много отсылок к другой статье, описывающей улучшение устойчивости модели к разным наборам данных – Learning transferable visual models from natural language supervision. Там используются текстовые данные для предобучения моделей зрения. Также эта статья используется в качестве примера, показывающего, что дообучение на конкретный набор данных сильно уменьшает обобщающую способность.

Немного про сильные стороны статьи. Во-первых, проведено действительно много экспериментов, чуть ли не две трети объёма статьи занимают различные графики, описания экспериментов, визуализации, доказательства, псевдокод и прочее. Во-вторых, существует официальный репозиторий, в котором реализованы идеи из статьи, он понятно написан, его приятно читать. И, в-третьих – и это скорее обобщение всех плюсов статьи – несмотря на то, что статья содержит одну небольшую идею, написана она достаточно полно, чтобы можно было понять, откуда эта идея пришла, в чём она состоит и как это использовать.

Не уверен, что я в праве говорить о серьёзных минусах, обсуждая статьи такого уровня, однако из того, что я бы хотел в статье улучшить, можно отметить следующее. Во-первых, из-за того, что статья по сути небольшая, возникает вопрос про новизну идеи – казалось бы, все идеи уже были придуманы, основная идея – в не самом сложном их комбинировании. Также вопрос, который у меня возник при чтении статьи, состоит в том, что идея статьи похожа на идею обучения на двух датасетах одновременно – в этом случае тоже повышается обобщающая способность и тоже увеличивается качество, возможно даже на двух наборах данных. Да, конечно, есть отличие в том, что нам не надо ещё раз переучивать целую модельку на большом наборе данных, но всё же хотя бы сравниться с таким подходом стоило бы. Третьим вопросом, который возник у меня при прочтении, был про выбор θ – коэффициента усрденения модели. В статье бралось $\theta = 0.5$, либо делалось утверждение "возьмите оптимальное θ" для вашей задачи. Однако не совсем понятно, что такое оптимальное θ – нам же хочется не только увеличить качество на целевом наборе данных, но и сохранить устойчивость. Как искать баланс между устойчивостью и качеством я не совсем понимаю.

В качестве выводов хочется сказать следующее. Из этой статьи мы вынесли несколько фактов. Во-первых, – технический – описан способ дообучать модели, не понижая устойчивости. Во-вторых, авторы описали ещё одно применение усреднению весов моделей, что является довольно интересным фактом, как мне кажется. В-третьих – и это уже можно использовать всем нам, как авторам статей, – даже статьи с небольшой главной

идеей, но хорошем оформлением, могут быть полезными и иметь успех в научном сообществе. А мир станет немного лучше, если чуть больше статей будут оформлены грамотно.