Название статьи: When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations
Авторы статьи: Xiangning Chen, Cho-Jui Hsieh, Boqing Gong
Автор текста: Дарья Виноградова

Довольно свежая работа, первая версия ее была загружена на arxiv 3 июня 2021 года. Этой осенью она была подана на ICLR-2022 и принята в spotlight формате.
Авторов трое. Xiangning Chen - PhD из университета Калифорнии. Cho-Jui Hsieh - его научный руководитель. К их интересам в основном относятся эффективный DL, оптимизации нейросетей. Boqing Gong - исследователь из гугла, специализирующийся на компьютерном зрении, в частности анализе внутреннего устройства CV-нейросетей. В его команде не раз стажировался первый автор. Становится вполне ясным, как эти люди нашли друг друга.
Интересы Xiangning Chen весьма разнообразны - рекомендательные системы, детекция объектов,поиск оптимальных нейросетевых архитектур. Большая часть его работ относится именно с третьей теме, за что он даже получил outstanding paper award на ICLR в прошлом году. Это как раз-таки написанные вместе с Cho-Jui Hsieh статьи. С Boqing Gong работа также продолжительная, две предыдущие статьи были про улучшение процесса обучения детекции (CV+NAS опять же) и про определение визуальных отношений детектируемых объектов (чистый CV). Нельзя назвать обсуждаемую работу продолжением какой-то из имеющихся, разве что вспомнить, что Boqing Gong уже издавал статьи про генерализацию моделей. Это именно статья про новый метод, закономерно следующая из интересов авторов.
Из связанных работ считаю наиболее важной An Empirical Study of Training Self-Supervised Vision Transformers, принадлежащую все тому же Xiangning Chen. В ней не предлагается никаких новых методов, но детально исследуется предобучение трансформера с помощью Contrastive Learning, подробно разбирается эффект от ablation, сравнивается, как разные конфигурации оптимизатоов влияют на процесс обучения. Словом, это способ объяснить, почему трансформер устроен именно так. Именно в этой работе было сделано замечание про возникновение острых локальных минимумов при обучении, что, судя по всему, послужило причиной возникновения исследования из статьи. В разделе про предшествующие работы в основном упоминаются статьи про базовые архитектуры, вокруг которых потом строится обсуждение.
Эту статью цитируют в исследованиях различных надстроек для улучшения работы трансформеров. Прямым продолжением является аналогичное исследования влияния SAM на текстовые модели (там оно тоже принесло улучшения). Xiangning Chen и Cho-Jui Hsieh сами продолжают активно прорабатывать эту область: они написали еще одну статью про оптимизацию SAM для обучения ViT, которая заметно ускоряет процесс.
Описанная идея в принципе довольно простая (возьмем новый оптимизатор, применим к разным моделям, сравним), и ее можно экстраполировать на другие области (в NLP такую попытку уже предприняли). Из более узких идей: можно было бы попробовать провести аналогичную с ViT оптимизацию для других CV моделей и ускорить их обучение.