Название статьи (авторы статьи): **When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations** (Xiangning Chen, Cho-Jui Hsieh, Boqing Gong)
Автор исследования: Александра Сендерович

Публикация

Первая версия статьи выложена на arxiv 3 июня 2021 года, вторая – 11 октября 2021 года. Рецензенты поставили оценки 56888, оценка 5 была поставлена из-за того, что статья по сути представляет из себя чисто экспериментальное исследование предложенного ранее метода, SAM. После рецензий авторы добавили несколько экспериментов, но идейно ничего не поменялось. Статья принята на ICLR 2022 как Spotlight.

Авторы

1) Xiangning Chen – стажировка в Google Research, AutoML Team + PhD в University of California, Los Angeles, ранее – закончил бакалавриат в Цинхуа. Все статьи с 2020 года – в соавторстве со вторым автором.
    a) Кроме этой статьи, подавал на ICLR2022 ещё 2 статьи про ViTs: “Can Vision Transformers Perform Convolution?” и “Sharpness-Aware Minimization in Large-Batch Training: Training Vision Transformer In Minutes”, но они были отозваны.
    b) В 2020 году опубликовал статью, в которой улучшал один из алгоритмов NAS (Neural Architecture Search, по сути AutoML) путём сглаживания поверхности функции потерь. Похожая вещь происходит и в обсуждаемой статье.
    c) В этом году 3 принятых на ICLR2022 статьи: обсуждаемая, распределённое adversarial обучение, автоматический подбор шага оптимизатора. В 2021 году опубликовал 3 статьи по NAS – Neural Architecture Search (по сути AutoML) + одну по автоматической аугментации для улучшения устойчивости к атакам в задачах компьютерного зрения (в соавторстве с обоими авторами текущей статьи).
2) Cho-Jui Hsieh – его научный руководитель из UCLA, доцент. Глава UCLA Computational Machine Learning Group. Группа занимается adversarial robustness и model compression. У его группы много публикаций в 2021, он везде последний или предпоследний автор.
    a) Оптимизатор LAMB для одного из сетапов берется из его статьи 2020 года.
3) Boqing Gong – научный руководитель из Google. Занимается компьютерным зрением и adversarial robustness. Тоже за прошлый год много статей, где он последний или предпоследний автор.

Получается, что в основном первый автор занимается NAS в соавторстве со вторым автором; кроме того, в прошлом году у него была 1 статья на пересечении AutoML и компьютерного зрения в соавторстве с третьим автором. Из авторов текущей статьи только третий занимается в основном компьютерным зрением. Недавно первый автор начал работать над Vision Transformers и в прошлом году участвовал в написании 3 статей на эту тему, но 2 другие были отозваны из-за низких оценок. При этом в 2 из 3 прошлогодних статей применялся метод SAM. Ранее в другой области автор уже применял похожую на SAM глобальную идею: гладкий минимум ведёт к лучшей обобщающей способности модели.

Ссылки и цитирования

Работа опирается на 4 главные статьи: авторы анализируют ViT (Dosovitsky, 2020) и MLP Mixer (Tolstikhin, 2021), применяют к ним SAM (Foret, 2021) и сравнивают с ResNet (He, 2015). В качестве одной из нестандартных метрик кривизны функции потерь используют предложенное в (Xiao et al., 2020) NTK condition number.

По данным SemanticScholar у статьи 26 цитирований, из них 17 – в обзоре литературы. В основном это статьи, которые по-своему оптимизируют обучение Vision трансформеров и сравниваются с результатами этой (например, “Bootstrapping ViTs: Towards Liberating Vision Transformers from Pre-training”). Единственная работа, которая может претендовать на звание прямого продолжения текущей – это “Sharpness-Aware Minimization Improves Language Model Generalization”, в которой SAM применяется для обычных текстовых трансформеров. Авторы этой статьи пишут во введении, что начали работу, “вдохновившись успехом SAM для трансформеров в зрении”.

Возможные последующие исследования и приложения

1) Авторы статьи говорят о том, что с использованием SAM ViTs (особенно их первые слои) стали довольно разреженными. Это говорит о том, что можно попробовать найти сжатую параметризацию трансформеров.
2) В статье про mixup проводились эксперименты с corrupted labels. Если авторы говорят, что SAM полностью заменяет аугментации, то такие эксперименты тоже было бы интересно провести для SAM.
3) В статье утверждается, что свёртки можно заменить трансформерами. Возможно, это означает, что получится сделать трансформер, решающий одновременно и задачи из CV, и задачи из NLP, то есть одну многофункциональную модель.