Robust fine-tuning of zero-shot models

1) В статье предлагается метод, как можно получать модель, которая дает высокое качество на целевом распределении и при этом имеет хорошую обобщающую способность. Идея — усреднять веса zero-shot моделей (например CLIP) и их fine-tuned версию на целевом распределении. У такой модели остается хорошая устойчивость к сдвигам распределения данных, но также для целевых данных достигается качество, сравнимое с fine-tuned моделями.
2) Статья вышла в сентябре 2021 года. Опубликована на 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Авторы: Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, Ludwig Schmidt.

Mitchell Wortsman — PhD студент Вашингтонского университета под консультацией профессора Ali Farhadi и доцента Ludwig Schmidt.
Gabriel Ilharco — также PhD студент, консультируемый Ali Farhadi и доцентом Hannaneh Hajishirzi.
Jong Wook Kim - член технического персонала OpenAI, также работал над моделью CLIP.
Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes — научные сотрудники Google.

Предшествующие работы авторов:

- Learning Neural Network Subspaces - февраль 2021, авторы - Mitchell Wortsman и Ali Farhadi и др., изучается геометрическая связь между весами моделей с высокой точностью, нахождение пути между этими весами, показывается, что модель с весами посередине дает высокую точность и лучшую устойчивость.
- OpenCLIP: An open source implementation of CLIP — 2021, авторы - Mitchell Wortsman, Gabriel Ilharco, Hannaneh Hajishirzi, Ali Farhadi, Ludwig Schmidt
- Measuring Robustness to Natural Distribution Shifts in Image Classification — 2020, авторы - Ludwig Schmidt и др., также он участвовал в написании других статей, посвященных исследованию чего-то со сдвигом в распределении данных.
- The Evolution of Out-of-Distribution Robustness Throughout Fine-Tuning — июнь 2021, авторы - Rebecca Roelofs и др. - о том, как датасет, на котором дообучается модель, влияет на обобщающую способность.

Работа не является последовательным улучшением других работ. Скорее возникла идея попробовать работать с пространством весов, также посмотреть на модели на пути от одних до других для существующей проблемы с неустойчивым к сдвигом данных дообучением.

3) У работы очень много ссылок — 111. Я бы выделила следующие статьи из них как базовые:

    - Measuring Robustness to Natural Distribution Shifts in Image Classification — 2020, Rohan Taori, et al. - о том, что текущие модели, обученные на imageNet, не устойчивы к естественным сдвигам данных.
    - What is being transferred in transfer learning? - 2020, Behnam Neyshabur et al. - о том, что между двумя дообученными моделями есть линейный путь в пространстве весов с низкой ошибкой.
    - Learning transferable visual models from natural language supervision — 2021, Radford et al. - исследование устойчивости при дообучении, CLIP, исследование пяти сдвигов ImageNet, на которых проводились эксперименты в статье.

4) Последующие связанные работы:
    - Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time - март 2022 года, авторы - Mitchell Wortsman, Gabriel Ilharco, Ali Farhadi, Hongseok Namkoong, Ludwig Schmidt и др.
    - Patching open-vocabulary models by interpolating weights — август 2022, авторы - Gabriel Ilharco, Mitchell Wortsman, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, Ludwig Schmidt и др.
    - Context-Aware Robust Fine-Tuning — ноябрь 2022, Xiaofeng Mao, et al. (от других авторов)

- Exploring The Landscape of Distributional Robustness for Question Answering Models — октябрь 2022, авторы - Mitchell Wortsman, Gabriel Ilharco, Hannaneh Hajishirzi, Ludwig Schmidt и др.

5) Примерно в то же время вышла статья:
   - Amortized Prompt: Lightweight Fine-Tuning for CLIP in Domain Generalization — ноябрь 2021, Xin Zhang, et al. В статье предлагается новая архитектура, использующая подсказки домена данных для улучшения качества предсказаний CLIP. Сама модель не дообучается.
6) Сильные стороны: подробный разбор области с большим числом источников, понятно, откуда взялась идея, актуальность работы, большое число экспериментов и много обученных моделей, красивые и понятные картиночки, исследование гиперпараметров, оформление результатов и анализ, работающий код на github с кратким описанием и инструкцией
7) Слабые стороны: уже исследованную идею применили к уже имеющейся проблеме, небольшое разнообразие данных.
8) В статье прекрасно реализована и оформлена идея, которая, как показали авторы, также хорошо работает. Для убедительности можно сделать эксперименты на других датасетах, не только исследуя сдвиги ImageNet.