Адверсальное дообучение моделей интерактивной сегментации изображений для улучшения устойчивости к возмущениям ограничивающих прямоугольников

Голубева Алёна Алексеевна

студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: al.gol.02@inbox.ru

Научный руководитель *— Шахуро Владислав Игоревич*

Модели компьютерного зрения, такие как Segment Anything Model (SAM) [1], Mobile SAM [3] и Segment Anything Model 2 (SAM2) [2] показывают впечатляющие результаты при сегментации объектов по идеальным ограничивающим прямоугольникам (tight bbox). Однако на практике bbox, полученные от детекторов или пользователей, часто содержат неточности (сдвиги, масштабирование), что приводит к резкому падению качества сегментации. Данная работа посвящена повышению робастности SAM к таким неидеальностям на входе.

В работе предложен алгоритм адверсального DoRA дообучения [4], который имитирует «плохие» bbox в процессе тренировки. Алгоритм имеет два режима обработки батча:

- Обучение на «хороших» данных: Модель обучается предсказывать маску по tight bbox (стандартный подход).

- Адверсальная атака и дообучение: Процесс итеративно «ухудшает» tight bbox, двигая его в направлении градиента функции потерь (Dice loss), чтобы максимизировать ошибку сегментации. После $N = 3$ шагов атаки модель дообучается на полученных «плохих» bbox, при этом целевой маской служит предсказание модели для исходного tight bbox. Это позволяет сгладить ландшафт функции потерь вокруг оптимального входа.

В качестве бейзлайна также реализован метод случайного сглаживания, усредняющий предсказания для нескольких случайных вариаций одного bbox.

Сравнение проводилось по метрике IoU на тестовой выборке с синтезированными неточностями в bbox. Сравнивались:

1. Базовая модель SAM;
2. Модель, дообученная без атакованных батчей;
3. Модель, дообученная с атакованными батчами (Adversarial DoRA);
4. Базовая модель с постобработкой случайным сглаживанием.

Эксперименты демонстрируют, что предложенный метод Adversarial DoRA дообучения эффективно повышает устойчивость модели к искажениям ограничивающих прямоугольников, превосходя как стандартное дообучение, так и техники случайного сглаживания (Таблица 1).

Таблица 1: Сравнение производительности различных методов на примере Mobile SAM

Dataset	Исходная	Случ. сглаж.	DoRA	Адверс. DoRA
COCO-Stuff 10K	0.5408	0.5313	0.5139	**0.5665**
GrabCut	0.8249	0.8271	0.8015	**0.8451**
Berkeley	0.7995	0.8093	0.7625	0.8055

Литература

1. Kirillov A., Mintun E., Ravi N., Mao H., Rolland C., Gustafson L., Xiao T., Whitehead S., Berg A. C., Lo W.-Y., Dollár P., Girshick R. Segment Anything // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023. P. 4015–4026.
2. Ravi N., Gabeur V., Hu Y.-T., Hu R., Ryali C., Ma T., Khedr H., Rädle R., Rolland C., Gustafson L., Mintun E., Pan J., Alwala K. V., Carion N., Wu C.-Y., Girshick R., Dollár P., Feichtenhofer C. SAM 2: Segment Anything in Images and Videos // ICLR 2025 (Oral), Singapore, Singapore, 2025.
3. Zhang C., Han D., Qiao Y., Kim J. U., Bae S. H., Lee S., Hong C. S. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications // arXiv preprint. 2023. № 2306.14289.
4. Liu S., Wang C.-Y., Yin H., Molchanov P., Wang Y.-C. F., Cheng K.-T., Chen M.-H. DoRA: Weight-Decomposed Low-Rank Adaptation // Proceedings of the 41st International Conference on Machine Learning (ICML Oral), Vienna, Austria, 2024.