DOI 10.37882/2223-2966.2024.8.15

ИСПОЛЬЗОВАНИЕ ТЕХНИК РЕПРЕЗЕНТАТИВНОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ПРИКЛАДНЫХ И ИНДУСТРИАЛЬНЫХ ЗАДАЧ В ОБЛАСТИ КОМПЬЮТЕРНОГО ЗРЕНИЯ

UTILIZING REPRESENTATION LEARNING TECHNIQUES FOR SOLVING APPLIED AND INDUSTRIAL PROBLEMS IN COMPUTER VISION

A. Gurov
E. Kamilov

Summary. The paper investigates methods for solving applied and industrial problems in the field of computer vision, with a particular focus on granulometry. Despite advancements in machine learning, such tasks lack effective solutions due to limited annotated data. The paper reviews existing approaches, emphasizing representation learning methods, including those from related fields. The findings highlight the potential and limitations of current approaches, indicating the need for further research to effectively address such challenges.

Keywords: computer vision, segmentation, representation learning, foundational models.

Гуров Андрей Владимирович
Аспирант, Федеральное государственное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»
avgurov@itmo.ru
Камилов Эркин Махмуджанович
Аспирант, Федеральное государственное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»
emkamilov@itmo.ru

Аннотация. В статье исследуются методы решения прикладных и индустриальных задач в области компьютерного зрения, гранулометрии в частности. Несмотря на достижения машинного обучения, подобные задачи не имеют эффективного решения из-за ограниченности аннотированных данных. В статье рассматриваются существующие подходы с акцентом на методы репрезентативного обучения, в том числе из смежных областей. Выводы подчеркивают потенциал и ограничения текущих подходов, указывая на необходимость дальнейших исследований для эффективного решения подобных задач.

Ключевые слова: компьютерное зрение, сегментация, репрезентативное обучение, фундаментальные модели.

Введение

Методы машинного обучения продемонстрировали многообещающие результаты в прикладных и индустриальных задачах компьютерного зрения, таких как верификация лиц, управление беспилотными автомобилями, выявление заболеваний на рентгеновских снимках и детекция дефектов на производстве. Однако многие индустриальные задачи, как сегментация мелких, повторяющихся зернистых объектов остается сложной задачей из-за ограниченности аннотированных данных. Это приводит к дефициту эффективных решений для таких задач, как оптическая гранулометрия. Исследователи используют различные методы, включая передобучение существующих моделей и адаптацию больших моделей через оптимизацию промптов.

В данной работе рассматриваются методы, основанные на техниках репрезентативного обучения, демонстрирующие отличные результаты в задачах few-shot и zero-shot learning, с акцентом на сегментацию объектов зернистой природы, но не ограничивается ими.

Цель данной работы — определить достоинства и недостатки существующих на сегодняшний день решений в рассматриваемой области, а также выявить направления будущих исследований, направленных на улучшение этих методов.

Для достижение этой цели, ставятся следующие задачи:

- Обзор существующих решений индустриальных задач на примере задачи сегментации объектов зернистой природы/гранулометрии, их достоинств и недостатков;
- Исследование методов и подходов, основанных на адаптации больших моделей к новым задачам и данным;
- Исследование методов, использующих техники адаптации, для решения задач сегментации данных специфического домена, оценка их результатов и перспектив использования.

Обзор существующих решений

Для начала стоит определить, какие объекты мы имеем в виду, говоря о зернистой структуре. В данном раз-

деле к объектам с зернистой структурой относятся следующие материалы:

- Материалы, имеющие зерно в качестве фундаментальной иерархической единицей организации (например, некоторые металлы) [1, 2, 3];
- Сыпучие материалы гранулированной формы (зерно, камни) [4];
- Другие материалы и вещества, состоящие из почти однородных по форме, размеру и цвету плотно расположенных друг к другу объектов (пузыри пены, листва на дереве и т. д.) [5, 6].

Чаще всего такие структурные особенности материалов встречаются в горнодобывающей промышленности, обработке металлов, сельском хозяйстве и строительстве. Долгое время для сегментации объектов такой природы использовались классические подходы компьютерного зрения, такие как: настраивание порогового значения [7], методы на основе градиентов изображения [7], HED методы [8]. Несмотря на вычислительную эффективность и простоту использования, эти методы имели ряд серьезных недостатков: необходимость ручной настройки параметров/гиперпараметров, плохая обобщаемость и низкое качество.

Были и другие попытки адаптировать классические подходы компьютерного зрения к данной задачи. Сегментация изображений с помощью суперпикселей и по сей день остается популярным подходом, в основании которого лежит идея разделения всего изображения на отдельные однородные участки (суперпиксели/супервоксели [9, 10, 11]) с последующим слиянием с помощью распределений Гаусса [12] и NCut-методами [13], что выражается в их большой вычислительной сложности.

В большинстве своём при использовании машинного обучения при сегментации объектов зернистой структуры исследователи пользуются хорошо зарекомендовавшими себя семействами моделей, такими как: U-Net [14], YOLO [15] или Mask R-CNN [16] и их модификациями. Так, например, модификация Mask R-CNN и YOLO хорошо показала себя при сегментации пшеницы и других сельскохозяйственных культур с плотной зернистой структурой колоса [17, 18].

Другие работы были направлены на разработку новых семейств нейронных сетей, специализирующихся именно на решении поставленной задачи, а не адаптацию существующих моделей [19].

Хотя использование таких инструментов как нейронные сети и дает существенный прирост в качестве алгоритмов, оно усугубляет проблему, связанную с кропотливой и долгой разметкой данных для обучения.

Преодолеть это ограничения попытались с помощью комбинирования нейронных сетей с классическими методами. При таком подходе нейронные сети обучаются не на размеченных вручную данных, а на синтетических данных или данных, размеченных с помощью классических подходов компьютерного зрения [20, 21, 22]. При таких подходах также значительно увеличить уже имеющуюся обучающую выборку путем обогащения ее различными сегментирующими масками, немного отличными друг от друга, пытаясь решить таким образом проблему недостатка данных и недостатка аннотаций к данным.

Были также попытки исследовать саму природу таких данных, не используя при это разметку, выучивая репрезентативные представления. Такие методы используют подходы self-supervised learning и показывают выдающиеся результаты во многих областях машинного обучения, в том числе и в сегментации изображений [23, 24, 25, 26].

По причине недавнего успеха self-supervised подходов многие исследователи стали адаптировать их для задачи сегментации зернистых объектов. В работе [27] авторы попытались расширить существующие фреймворки репрезентативного обучения, чтобы они были способны «обращать внимание» на гранулированные и зернистые структуры, что является одной из первых попыток обобщить модели не на конкретный домен данных, а на объекты определенной природы.

Несмотря на все возможные перспективы, связанные с использованием техник репрезентативного обучения или адаптаций больших моделей, в научном сообществе пока не сформировался тренд к использованию этих методик к узкоспециализированным областям и доменам данных, как сегментация зернистых объектов.

Большие модели

В набирающей обороты тенденции к изучению больших нейросетевых моделей, способных решать широкий круг задач (большие модели), и исследованию их возможностей одним и важнейших направлений стало сближение достижений в области обработки естественного языка (NLP) и компьютерного зрения (CV).

Основополагающим вкладом в эту траекторию является разработка и дальнейшее улучшение GPT (Generative Pre-trained Transformer) моделей, в частности GPT-3, обширной языковой модели, которая имела на тот момент беспрецедентное количество параметров в 175 миллиардов [28]. GPT-3 не только демонстрировала исключительную эффективность в NLP задачах, но также и предзнаменовала парадигматический сдвиг, демонстрируя выдающиеся способности к обучению с помощью few-shot learning, показывая при этом впечатляющие результаты. Многогранный успех модели

в машинном переводе, в задачах ответа на вопросы, в оперативном рассуждении и адаптации предметной области подчеркнул потенциал для развития универсальных, общих языковых систем [28].

Одновременно с этим в сфере обработки естественного языка наметилась тенденция к поиску других методов адаптации больших моделей к новым задачам и новым данным помимо обучения с нуля и переобучения (fine-tuning). Один из таких методов вообще не требовал никаких действий с параметрами уже обученных моделей в отличие от обучения с учителем, используя при этом оптимизацию промптов (prompt), последовательностей входных слов, для прямого воздействия на поведение модели [29]. Такая схема "обучения" с помощью промптов продемонстрировала высокое качество в zero/few-shot задачах без обилия аннотированных данных. Новый подход к адаптации больших языковых моделей, воплощенный в правильной стратегии оптимизации входных промптов, не только послужил в качестве новой парадигмы в сфере NLP, но и выступил инструментом для повышения доступности в этой области. Авторы также утверждают, что такие эффективные и простые в использовании методы, как оптимизация промптов, являются катализаторами будущих научных достижений и ключом для исследования возможностей больших языковых моделей [29].

Все эти достижения и успешные практики в сфере NLP нашли применения в других областях, в том числе в компьютерном зрении. Так, например, была предложена методология по улучшению моделей на базе Vision Transformer посредством включения в её архитектура специальных токенов памяти (memory tokens) [30]. Это токены, встроенные в каждый слой первоначальной модели, служат дополнительной контекстной информацией для конкретных наборов данных, способствуя эффективной адаптации уже обученной модели. Идея дополнительных контекстных токенов была заимствована из сферы NLP, где к тому моменту техники оптимизации входных дополнительных промптов были широко распространены. Концептуальное отличие от первоначальной идеи лишь в том, что в случае с Vision Transformer эти промпты представлены в виде обучаемых параметров (токенов памяти). В работе также были использованы и другие нововведения, такие как маскированное внимание (masked attention), упрощающие адаптацию модели для последующих новых задач и новых данных. Пример использования подобных техник в задачах компьютерного зрения представляет убедительный аргумент в пользу эффективности стратегий обобщения больших моделей для новых задач и новых данных в данной сфере, закладывая основу для будущих научных исследований.

Другой попыткой адаптировать уже обученные модели к новым данным была работа, направленная, в отличии от предыдущей, на изучения одного универсального токена, не зависящего от входных данных, который при применении к уже обученной модели, такой как CLP [31], обеспечит эффективное выполнения новых задач [32]. В статье также исследуется неожиданная эффективность такого подхода по сравнению со многими другими решениями, а также устойчивость к сдвигу распределений входных данных. Эта работа только укрепила позиции новой парадигмы в методах адаптации моделей глубокого обучения к новым данным/задачам и побудила к дальнейшим исследованиям в данном направлении, направленных на понимание условий и контекстов, при которых визуальные промпты оказывают влияние на способность моделей компьютерного зрения к обобщению.

Другим подтверждением того, что новые методы адаптации на основе визуальных промптов зачастую показывают лучший результат по сравнению с fine-tuning методами является метод настройки визуальных промптов (visual prompts tuning), предложенный в [32]. Но, в отличии от предыдущей работы, в нем оптимизируется не только входной промпт для определенной задачи, но и дополнительный легковесный слой адаптации. Остальные веса уже обученной модели остаются замороженными.

Все эти примеры моделей и подходов, рассмотренных в данном разделе, демонстрируют нам перспективу использования больших моделей в совокупности с методами эффективной адаптации в качестве потенциальных решений узконаправленных задач на специфичных доменах данных. За счет большого количества параметров и предобучения в self-supervised манере большие модели имеют выдающиеся способности к обобщению на новые классы задач и на новые распределения данных, ранее не видимых для нее.

Адаптация больших сегментирующих моделей

Для исследования применимости вышеописанных техник в сфере решения прикладных задач, была выбрана модель SAM в качестве большой модели для дальнейшей адаптации. За счет self-supervised подходов, использованных перед тренировкой модели, эта модель обладает выдающимися способностями к few-shot и zero-shot сегментации, что говорит о ее хорошей обобщенности на разных доменах данных [34].

Последние достижения в области эффективных методик адаптации больших моделей, особенно в контексте Segment Anything модели, подчеркивают растущую необходимость в предметно-направленных улучшениях для преодоления ограничений в специализированных задачах. Одним из таких подходов является Conv-LoRA [35], что был разработан как общий фреймворк для улуч-

шения производительности SAM в задачах семантической сегментации. Метод успешно преодолевает ограничения SAM в некоторых специализированных областях, демонстрируя превосходные возможности адаптации.

Также были попытки, направленные не на получения широко обобщенной вариации SAM на всевозможных доменах, а на адаптацию к одному классу данных — медицинским снимкам. Учитывая врожденные ограничения SAM, сегментации определенных классов данных не достигает, порой, даже средних по качеству результатов, так как ни данные такого рода, ни данные из смежных отраслей или схожей природы не использовались в процессе тренировки модели. К таким данным относятся медицинские КТ и МРТ снимки как в 2D, так и в 3D формате [36, 37]. Оба метода в этих работах используют техники PEFT (parameter efficient fine-tuning) для адаптации модели к новым данным, в том числе и к 3D снимкам. Medical SAM Adapter [36] использует обычный MLP-прослойки в качестве адаптеров, что позволяет снизить почти до минимума количество обучаемых параметров. Несмотря на это, метод достигает SOTA результатов на 17 наборах данных в задачах сегментации изображений, подчеркивая свою эффективность в решении проблемы недостаточной производительности SAM на медицинских снимках. Авторы SAM-Adapter [37] представляет более универсальный подход, используя помимо адаптеров-прослоек дополнительную информацию, путем интегрирования предметно-специфической информации в форме дополнительных визуальных данных. SAM-Adapter значительно улучшает производительность SAM в сложных задачах, превосходя модели сетей, специфичных для задач, и достигая SOTA результатов как на медицинских данных, так и на других доменах. Эта адаптивность открывает новые возможности для применения SAM в различных областях, включая обработку медицинских изображений, сельское хозяйство и дистанционное зондирование.

Расширяя область задач сегментации переднего плана, Explicit Visual Prompting (EVP) [38] представляет универсальную структуру для различных задач сегментации переднего плана в компьютерном зрении, таких как выделение существенных объектов, обнаружение подделок, обнаружение нефокусированных размытий, обнаружение теней и обнаружение камуфлированных объектов. Предложенная структура, Explicit Visual Prompting (EVP), вдохновлена протоколами предварительного обучения и настройки запросов в обработке естественного языка (NLP). EVP фокусируется на явном визуальном содержании в отдельных изображениях, используя замороженные вложения патчей и компоненты высокой частоты. Метод превосходит полную настройку и другие эффективные методы точной настройки параметров в различных задачах сегментации переднего плана, продемонстрировав свою масштабируемость по различным наборам данных, архитектурам и предварительно обученным весам.

В заключение, текущее состояние дел в моделях семантической сегментации подчеркивает изменение парадигмы в сторону эффективных методов адаптации для преодоления ограничений в конкретных областях. Успех Conv-LoRA, Med-SA, SAM-Adapter и EVP подчеркивает потенциал универсальных, предметно-специфичных улучшений, подчеркивая необходимость целенаправленных адаптаций для решения задач в реальных приложениях. Текущий тренд в исследованиях предполагает нюансированный подход, настраивая модели сегментации под конкретные задачи с использованием эффективных методов адаптации, тем самым повышая применимость и устойчивость этих моделей в различных областях.

ЛИТЕРАТУРА

1. Yun Bai, Grady Wagner, and Christopher B Williams. Effect of particle size distribution on powder packing and sintering in binder jetting additive manufacturing of metals. Journal of Manufacturing Science and Engineering, 139(8):081019, 2017.
2. Choong Do Lee. Effect of grain size on the tensile properties of magnesium alloy. Materials Science and Engineering: A, 459(1-2):355–360, 2007.
3. Philipp Schempp, CE Cross, Ralf Häcker, Andreas Pittner, and Michael Rethmeier. Influence of grain size on mechanical properties of aluminium gta weld metal. Welding in the World, 57:293–304, 2013.
4. Heinrich M Jaeger, Sidney R Nagel, and Robert P Behringer. The physics of granular materials. Physics today, 49(4):32–38, 1996.
5. Wei-Jian Hu, Jie Fan, Yong-Xing Du, Bao-Shan Li, Naixue Xiong, and Ernst Bekkering. Mdfc–resnet: an agricultural iot system to accurately recognize crop diseases. IEEE Access, 8:115287–115298, 2020.
6. Jakob D Redlinger-Pohn, Matthias Grabner, Philipp Zauner, and Stefan Radl. Separation of cellulose fibres from pulp suspension by froth flotation fractionation. Separation and Purification Technology, 169:304–313, 2016.
7. John Canny. A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence, (6):679–698, 1986.
8. Saining Xie and Zhuowen Tu. Holistically nested edge detection. In Proceedings of the IEEE international conference on computer vision, pages 1395–1403, 2015.
9. David Bryant Mumford and Jayant Shah. Optimal approximations by piecewise smooth functions and associated variational problems. Communications on pure and applied mathematics, 1989.
10. Ren and Malik. Learning a classification model for segmentation. In Proceedings ninth IEEE international conference on computer vision, pages 10–17. IEEE, 2003.
11. David Stutz, Alexander Hermans, and Bastian Leibe. Superpixels: An evaluation of the state-of-the-art. Computer Vision and Image Understanding, 166:1–27, 2018.
12. Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels. Technical report, 2010.

13. Ming-Yu Liu, Oncel Tuzel, Srikumar Ramalingam, and Rama Chellappa. Entropy rate superpixel segmentation. In CVPR 2011, pages 2097–2104. IEEE, 2011.
14. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, pages 234–241. Springer, 2015.
15. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.
16. Kaiming He, Georgia Gkioxari, Piotr Doll á r, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.
17. Keyhan Najafian, Alireza Ghanbari, Mahdi Sabet Kish, Mark Eramian, Gholam Hassan Shirdel, Ian Stavness, Lingling Jin, and Farhad Maleki. Semi-self-supervised learning for semantic segmentation in images with dense patterns. Plant Phenomics, 5:0025, 2023.
18. Xin Xu, Qing Geng, Feng Gao, Du Xiong, Hongbo Qiao, and Xinming Ma. Segmentation and counting of wheat spike grains based on deep learning and textural features. Plant Methods, 19(1):77, 2023.
19. Javad Manashti, Pouyan Pirnia, Alireza Manashty, Sahar Ujan, Matthew Toews, and Fran çois Duhaime. Psdnet: Determination of particle size distributions using synthetic soil images and convolutional neural networks. arXiv preprint arXiv:2303.04269, 2023.
20. Jun Long, Yuxi Yang, Liujie Hua, and Yiqi Ou. Self-supervised augmented patches segmentation for anomaly detection. In Proceedings of the Asian Conference on Computer Vision, pages 1926–1941, 2022.
21. Peter Warren, Nandhini Raju, Abhilash Prasad, Shajahan Hossain, Ramesh Subramanian, Jayanta Kapat, Navin Manjooran, and Ranajay Ghosh. Grain and grain boundary segmentation using machine learning with real and generated datasets. arXiv preprint arXiv:2307.05911, 2023.
22. Philipp Schempp, CE Cross, Ralf Häcker, Andreas Pittner, and Michael Rethmeier. Influence of grain size on mechanical properties of aluminium gta weld metal. Welding in the World, 57:293–304, 2013.
23. Mathilde Caron, Hugo Touvron, Ishan Misra, Herv é Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660, 2021.
24. Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33:21271–21284, 2020.
25. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan Yen Lo, et al. Segment anything. arXiv preprint arXiv:2304.02643, 2023.
26. Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499, 2023.
27. Pan Zhou, Yichen Zhou, Chenyang Si, Weihao Yu, Teck Khim Ng, and Shuicheng Yan. Mugs: A multi-granular self-supervised learning framework. arXiv preprint arXiv:2203.14415, 2022.
28. Brown T. et al. Language models are few-shot learners. Advances in neural information processing systems. — 2020. — T. 33. — C. 1877–1901.
29. Liu P. et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys. — 2023. — T. 55. — №. 9. — C. 1–35.
30. Sandler M. et al. Fine-tuning image transformers using learnable memory. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2022. — C. 12155–12164.
31. Radford A. et al. Learning transferable visual models from natural language supervision. International conference on machine learning. — PMLR, 2021. — C. 8748–8763.
32. Bahng H. et al. Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274. — 2022.
33. Jia M. et al. Visual prompt tuning. European Conference on Computer Vision. — Cham: Springer Nature Switzerland, 2022. — C. 709–727.
34. Kirillov A. et al. Segment anything. arXiv preprint arXiv:2304.02643. — 2023.
35. preprint, article currently of review.
36. Wu J. et al. Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620. — 2023.
37. Chen T. et al. SAM Fails to Segment Anything? — SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, and More. arXiv preprint arXiv:2304.09148. — 2023.
38. Liu W. et al. Explicit Visual Prompting for Universal Foreground Segmentations. arXiv preprint arXiv:2305.18476. — 2023.

Журнал «Современная наука: актуальные проблемы теории и практики»