Multi-modal Data Processing for Foundation Models: Practical Guidances and Use Cases

Daoyuan Chen
Alibaba Group, Hangzhou, China
daoyuanchen.cdy@alibaba-inc.com

Yaliang Li
Alibaba Group, Bellevue, WA USA
yaliang.li@alibaba-inc.com

Bolin Ding
Alibaba Group, Bellevue, WA USA
bolin.ding@alibaba-inc.com

ABSTRACT

In the foundation models era, efficiently processing multi-modal data is crucial. This tutorial covers key techniques for multi-modal data processing and introduces the open-source Data-Juicer system, designed to tackle the complexities of data variety, quality, and scale. Participants will learn how to use Data-Juicer's operators and tools for formatting, mapping, filtering, deduplicating, and selecting multi-modal data efficiently and effectively. They will also be familiar with the Data-Juicer Sandbox Lab, where users can easily experiment with diverse data recipes that represent methodical sequences of operators and streamline the creation of scalable data processing pipelines. This experience solidifies the concepts discussed, as well as provides a space for innovation and exploration, highlighting how data recipes can be optimized and deployed in high-performance distributed environments.

By the end of this tutorial, attendees will be equipped with the practical knowledge and skills to navigate the multi-modal data processing for foundation models. They will leave with actionable knowledge with an industrial open-source system and an enriched perspective on the importance of high-quality data in AI, poised to implement sustainable and scalable solutions in their projects. The system and related materials are available at https://github.com/modelscope/data-juicer.

ACM Reference Format:
Daoyuan Chen, Yaliang Li, and Bolin Ding. 2024. Multi-modal Data Processing for Foundation Models: Practical Guidances and Use Cases . In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3637528.3671441

TUTORIAL OUTLINE

1. Introduction and Overview: Multi-modal Data Processing and the Data-Juicer System
We will begin with an overview highlighting the importance of heterogeneous, high-quality datasets for advancing large generative models [3, 8, 13, 17], shifting from traditional unimodal data to multi-modal cases. We will explore the challenges of multi-modal data processing, such as improving data quality, handling large volumes, and integrating cross-modal information.

Then, in the context of foundation models, we will discuss the evolution of related data processing framework [1, 5, 10, 18, 19], and key techniques for managing data quality, quantity, diversity, privacy, and processing challenges [2, 6, 7, 9, 11, 12, 14–16, 20]. Next, we will introduce the open-source data processing system, Data-Juicer [4], designed for foundation models. An outline of Data-Juicer's architecture, target users, practical applications, and key features will set the preliminaries for its use in subsequent sections.

2. Building Blocks of Data Processing: Data-Juicer's Operators
We'll introduce Data-Juicer's operators, the key to streamlining multi-modal data processing for foundation models. These atomic units—equipped with models or algorithms—address various data scenarios, offering workflow simplicity, flexibility, and scalability.

Specifically, we'll demonstrate five types of operators: Formatter, for standardizing raw data; Mapper, for editing and augmenting; Filter, for assessing targeted metrics and discarding unqualified data; Deduplicator, for removing duplicates; and Selector, for targeted sampling. Operators allow customization of pipelines to fit various tasks, adapting to the scale and complexity required. Participants will learn to apply these operators effectively, with practical examples and industry tips.

3. Composing Atomic Capabilities: Data-Juicer's Data Recipes
We will delve into Data-Juicer's Data Recipes, which are structured operator sequences establishing versatile, trackable data workflows. This part will cover their purpose, structure, and merits, showing how to craft workflow definitions and manage data mixtures.

Attendees will gain hands-on experience with Data-Juicer's configuration system, learning to utilize and adapt built-in recipes. Code snippets and tips on customizing recipes will be provided, emphasizing practical reuse and customization.

4. Exploring Data Recipes: The Data-Juicer Sandbox Lab
The Data-Juicer Sandbox Lab is a key innovative part for fostering experimental learning, and developing effective data recipes. Attendees will experiment in an integrated ecosystem of open-source models, datasets, and benchmarks. The lab encourages cost-efficient data-model co-design with a focus on using compact datasets and models for rapid prototyping. Enhanced by user-friendly co-development tools, visualizations, tracers, auto-evaluators, and automatic hyper-parameter tuning, the lab accelerates the data science exploration process.

5. From Exploration to Production: High-Performance Data Factory
Transitioning from the exploration of promising data recipes to their application in large-scale production scenarios is a critical step in the data processing lifecycle. Our tutorial will address taking data recipes from trial to large-scale use, focusing on efficient practices like Data-Juicer's operator fusion and concurrent optimizations such as multiprocessing and GPU acceleration. We will also discuss

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671441

distributing processing using modern infrastructures like Ray [1], and Alibaba Cloud's PAI-DLC [2] to boost scalability and performance.

6. Use Cases: From Text to Video Data Processing
Our exploration will guide participants from pure text data to multi-modal (text-to-X) data for foundation models. This journey emphasizes the evolution from basic textual characteristics to complex aspects such as time dependencies, spatio-temporal synchronization, and integrating cross-modal information. Throughout this section, we will present multiple case studies and delve into data processing strategies, covering the entire spectrum from pre-training to fine-tuning, and drawing upon diverse data sources such as web content, books, code repositories, image-text pairs, scientific charts, and openly available videos.

Finally, we will showcase that by utilizing Data-Juicer recipes, which are designed to improve the quality, diversity, and alignment of data, various state-of-the-art foundation models can achieve notable enhancements in performance and training efficiency.

8. Conclusion and Resources
In concluding this tutorial, we will recapitulate the benefits that Data-Juicer provides to the realm of data processing. Furthermore, we will outline a collection of open problems and potential research directions, informed by a synthesis of industrial experiences and academic insights.

Attendees will be furnished with a collection of resources, including online Jupyter Notebook tutorials, comprehensive Data-Juicer documentation, curated "awesome lists", and a guide to pertinent data processing competitions for foundation models. This not only marks the culmination of our tutorial but also the beginning of a new chapter in attendees' journey toward mastering multi-modal data processing.

TUTORS BIOGRAPHY

Daoyuan Chen, a research scientist at Alibaba Tongyi Lab, earned his Master's in Computer Application Tech from Peking University in 2019. His expertise lies in foundational models, efficient machine learning (ML) systems and applications, with over 30 publications in top-tier conferences like KDD, SIGMOD, ICML, ICLR and NeurIPS, many as first author. He has also organized several LLM competitions and actively contributed as a core developer of many open-source projects that benefit both academia and industry.

Yaliang Li, a research scientist and director at Alibaba Tongyi Lab, earned his Ph.D. in Computer Science from SUNY Buffalo in 2017. Before Alibaba, he worked as a researcher at Baidu Research and Tencent Medical AI Lab. With over 100 publications in elite venues like KDD, NeurIPS, ICML, and VLDB, his research interests span machine learning and data mining with a focus on data integration, AutoML, Federated Learning, and recently Multi-Agent. He's held roles as a (senior) area chair for NeurIPS, ICML, ACL, and AAAI, and co-chaired three workshops for IJCAI-TUSION and NeurIPS.

Bolin Ding is a research scientist and senior director at Alibaba Tongyi Lab. He completed his Ph.D. in Computer Science at UIUC and worked as a researcher in Microsoft Research. His recent research focuses on making systems, including physical computing systems such as databases and ML systems, and social systems, intelligent and efficient with ML and optimization techniques. He holds more than 30 patents on data privacy, databases, and ML, some of which have been deployed into important products in the industry. His research results are published in top conferences on databases, data mining, algorithms, and ML, including KDD, SIGMOD, VLDB, ICDE, ICML, NeurIPS, ICLR, and SODA.

[1] https://github.com/ray-project/ray
[2] https://www.alibabacloud.com/product/machine-learning

REFERENCES

[1] Stephen H. Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M. Saiful Bari, Thibault Févry, et al. Promptsource: An integrated development environment and repository for natural language prompts. In *ACL (demo)*, pages 93–104, 2022.

[2] Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous language tasks and client resources. *arXiv:2402.11505*, 2024.

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv:2311.15127*, 2023.

[4] Daoyuan Chen, Yilun Huang, Zhijian Ma, Hesen Chen, Xuchen Pan, Ce Ge, Dawei Gao, Yuexiang Xie, Zhaoyang Liu, Jinyang Gao, Yaliang Li, Bolin Ding, and Jingren Zhou. Data-juicer: A one-stop data processing system for large language models. In *SIGMOD*, 2024.

[5] Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, April 2023.

[6] Ce Ge, Zhijian Ma, Daoyuan Chen, Yaliang Li, and Bolin Ding. Data mixing made efficient: A bivariate scaling law for language model pretraining. *arXiv:2405.14908*, 2024.

[7] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *EMNLP (Findings)*, pages 3356–3369, 2020.

[8] Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. Enhancing multimodal large language models with vision detection models: An empirical study. *arXiv:2401.17981*, 2024.

[9] Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. In *KDD*, 2024.

[10] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. The bigscience ROOTS corpus: A 1.6tb composite multilingual dataset. In *NeurIPS*, 2022.

[11] Yin Lin, Bolin Ding, H. V. Jagadish, and Jingren Zhou. Smartfeat: Efficient feature construction through feature-level foundation model interactions. *arXiv:2309.07856*, 2023.

[12] Zhenqing Ling, Daoyuan Chen, Liuyi Yao, Yaliang Li, and Ying Shen. On the convergence of zeroth-order federated tuning for large language models. In *KDD*, 2024.

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023.

[14] Mengsha Liu, Daoyuan Chen, Yaliang Li, Guian Fang, and Ying Shen. Chartthinker: A contextual chain-of-thought approach to optimized chart summarization. In *COLING*, 2024.

[15] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *CoRR*, abs/2305.13169, 2023.

[16] Yingqian Min, Kun Zhou, Dawei Gao, Wayne Xin Zhao, He Hu, and Yaliang Li. Data-cube: Data curriculum for instruction-based sentence representation learning. In *ACL (Findings)*, 2024.

[17] OpenAI. Gpt-4 technical report. *CoRR*, abs/2303.08774, 2023.

[18] Yichen Qian, Yongyi He, Rong Zhu, Jintao Huang, Zhijian Ma, Haibin Wang, Yaohua Wang, Xiuyu Sun, Defu Lian, Bolin Ding, and Jingren Zhou. Unidm: A unified framework for data manipulation with large language models. In *MLSys*, volume 6, pages 465–482, 2024.

[19] Soldaini, Luca and Lo, Kyle and Kinney, Rodney and Naik, Aakanksha and Ravichander, Abhilasha and Bhagia, Akshita and Groeneveld, Dirk and Schwenk, Dustin and Magnusson, Ian and Chandu, Khyathi. The Dolma Toolkit, 2023.

[20] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.