Deep Multimodal Learning for Information Retrieval

Wei Ji
National University of Singapore
weiji0523@gmail.com

Yinwei Wei
Monash University
weiyinwei@hotmail.com

Zhedong Zheng
National University of Singapore
zdzheng@nus.edu.sg

Hao Fei
National University of Singapore
haofei37@nus.edu.sg

Tat-seng Chua
National University of Singapore
dcscts@nus.edu.sg

ABSTRACT

Information retrieval (IR) is a fundamental technique that aims to acquire information from a collection of documents, web pages, or other sources. While traditional text-based IR has achieved great success, the under-utilization of varied data sources in different modalities (*i.e.*, text, images, audio, and video) would hinder IR techniques from giving its full advancement and thus limits the applications in the real world. Within recent years, the rapid development of deep multimodal learning paves the way for advancing IR with multi-modality. Benefiting from a variety of data types and modalities, some latest prevailing techniques are invented to show great facilitation in multi-modal and IR learning, such as CLIP, ChatGPT, GPT4, *etc.* In the context of IR, deep multi-modal learning has shown the prominent potential to improve the performance of retrieval systems, by enabling them to better understand and process the diverse types of data that they encounter. Given the great potential shown by multimodal-empowered IR, there can be still unsolved challenges and open questions in the related directions. With this workshop, we aim to provide a platform for discussion about multi-modal IR among scholars, practitioners, and other interested parties.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**.

KEYWORDS

Information retrieval, Multi-modal, CLIP

ACM Reference Format:
Wei Ji, Yinwei Wei, Zhedong Zheng, Hao Fei, and Tat-seng Chua. 2023. Deep Multimodal Learning for Information Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3581783.3610949

1 BACKGROUND AND MOTIVATION

The emergence of multimodal learning offers a feasible way for multimodal IR. Within recent decades with the rapid development of deep learning techniques, the triumph of multimodal learning has been witnessed [12, 14, 17]. Deep multimodal learning has been defined as to use of deep neural techniques to model and learn from multiple sources of data or modalities among others. In the context of IR, deep multimodal learning has shown great potential to improve the performance and application scope of retrieval systems, *i.e.*, by enabling better understanding and processing of the diverse types of data [1–11, 13, 15, 16, 18–22].

The ACM Multimedia main conference covers a broader range of general topics of multimodal applications, while the discussion on multimodal learning-based IR could be scarce. Our workshop can be a good complementarity to place the major focus on multimodal IR. This workshop sets the goal to extend existing work in this direction, by bringing together and facilitating the community of researchers and practitioners. And meanwhile, we aim to encourage an exchange of perspectives and solutions between industry and academia to bridge the gap between academic design guidelines and the best practices in the industry regarding multimodal IR.

2 TOPICS AND THEMES

List of topics covered in this workshop (but not limited to) is shown as follow:

- Image-text Multimodal Learning and Retrieval
  - Image-text Compositional Retrieval
  - Vision-language Alignment Analysis
  - Multimodal Fusion and Embeddings
  - Vision-language Pre-training
  - Structured Vision-language Learning
  - Visually Grounded Interaction of Language Modeling
  - Commonsense-aware Vision-language Learning
  - Visually Grounded Language Parsing
  - Semantic-aware Vision-language Discovery
- Video-text Understanding and Retrieval
  - Video-text Retrieval
  - Video (Corpus) Moment Retrieval
  - Video Relation Detection
  - Video Scene Graph Generation
  - Video Question Answering
  - Video Dialogue
- Dialogue Multimodal Retrieval
  - Multimedia Pre-training in Dialogue
  - Multimedia Search and Recommendation
  - Multimodal Response Generation
  - User-centered Dialogue Retrieval
  - New Applications on ChatGPT & Visual-GPT and Beyond

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3610949

- Reliable Multimodal Retrieval
  - Explainable Multimodal Retrieval
  - Typical Failures of ChatGPT and other Large Models
  - Adversarial Attack and Defense
  - New Evaluation Metrics
- Multimedia Retrieval Applications
  - Multimodal-based Reasoning
  - Unapired Image Captioning
  - Cross-modal Scene Graph Understanding
  - Multimodal Information Extraction
  - Opinion-oriented Multimodal Analysis for IR
  - Multimodal Translation
  - Multimodal Learning for Social Good

3 ACTIVITIES AND INVITED KEYNOTES

We plan to hold a hybrid format of workshop, *i.e.*, both onsite and online. For the onsite one at least two organizers will attend in person to host the workshop. The workshop will include two major activities, the invited keynotes, and the paper presentations. We will invite keynote presentations for a half-day workshop, following by accepted workshop presentations. The speakers are experts on the relevant community from different organizations globally.

4 PAPER SUBMISSION AND REVIEWING

In this workshop, we welcome three types of submissions, all of which should relate to the topics and themes as listed in Section 2:

1. Position or perspective papers (up to 4 pages in length, plus unlimited pages for references): original ideas, perspectives, research vision, and open challenges in the area of evaluation approaches for explainable recommender systems;

2. Featured papers (title and abstract of the paper, plus the original paper): already published papers or papers summarizing existing publications in leading conferences and high-impact journals that are relevant for the topic of the workshop;

3. Demonstration papers (up to 2 pages in length, plus unlimited pages for references): original or already published prototypes and operational evaluation approaches in the area of explainable recommender systems.

Page limits include diagrams and appendices. Submissions should be single-blind, written in English, and formatted according to the current ACM two-column conference format. Suitable LaTeX, Word, and Overleaf templates are available from the ACM Website (use "sigconf" proceedings template for LaTeX and the Interim Template for Word).

5 ORGANIZER INFORMATION

Wei Ji (https://jiwei0523.github.io) is a Research Fellow in the School of Computing at National University of Singapore. He received the Ph.D. degree in computer science from the Zhejiang University in 2020. He has published several papers in top conferences such as CVPR, ECCV, SIGIR, AAAI, ACM MM, and journals including TPAMI, TIP and TCYB. His current research interests include multi-modal learning, vision and language, and cross-modal retrieval.

Yinwei Wei (https://weiyinwei.github.io) is a research fellow, Faculty of Information Technology, Monash University. His research interests include multimodal computing, information retrieval, and recommender system, particular in multimedia personalized recommendation. He has published several papers in top conferences such as SIGIR, ACM MM, WSDM, and journals including IEEE TKDE, TIP and TMM. Moreover, he has served as the PC member for ACM MM, AAAI, WSDM and IJCAI conference, and the invited reviewer for prestigious journals including IEEE TPAMI, TIP, TKDE, and TMM.

Zhedong Zheng (https://zdzheng.xyz) is a research fellow at NExT++, School of Computing, National University of Singapore. He received the Ph.D. degree from the University of Technology Sydney, Australia, in 2021 and the B.S. degree from Fudan University, China, in 2016. His research interests include robust learning for multimedia retrieval, generative learning for data augmentation, and unsupervised domain adaptation. He has published 33 papers in highly selective venues such as CVPR, ICCV, ACM MM, TMM, IJCV and TPAMI with a citation of 7,000+ times in Google Scholar. Six of the research papers are elected as ESI highly-cited papers. He received the IEEE Circuits and Systems Society Outstanding Young Author Award of 2021. He has served as the reviewer and program committee (PC) member for multiple conferences and journals, including TPAMI, TMM, IJCV, CVPR, ICCV, ECCV, IJCAI, AAAI and ACM Multimedia, and organized a special session on reliable retrieval at ICME 2022. Besides, he is also invited as a keynote speaker at CVPR 2020, and 2021, and a tutorial speaker at ACM Multimedia 2022.

Hao Fei (https://scofield7419.github.io) is currently a postdoctoral research fellow at NExT++, School of Computing, National University of Singapore. He previously received the Ph.D. degree from Wuhan University, China, in 2021. His research interests cover natural language processing, text mining and multimodal learning, especially with the angle of the structural learning. His research has been published at top-tier relevant conferences, *e.g.*, ICML, NeurIPS, ACL, AAAI, SIGIR, IJCAI, WWW, EMNLP *etc*, and journals, *e.g.*, TOIS, TNNLS, TASLP *etc*. He served as Area Chair or Senior Program Committee in EMNLP 2022, WSDM 2022, IJCAI 2023 and ACL 2023, and General Chair of NSSDM 2023, Volunteer Chair of WSDM 2023.

Tat-seng Chua (https://www.chuatatseng.com/) received the Ph.D. degree from the University of Leeds, U.K. He is the KITHCT Chair Professor with the School of Computing, National University of Singapore, where he was the Acting and Founding Dean of the School from 1998 to 2000. His main research interests include multimedia information retrieval and social media analytics. In particular, his research focuses on the extraction, retrieval, and question-answering (QA) of text and rich media arising from the Web and multiple social networks. He is the Co-Director of NExT, a joint center between NUS and Tsinghua University, to develop technologies for live social media search. He is the 2015 winner of the prestigious ACM SIGMM Award for Outstanding Technical Contributions to Multimedia Computing, Communications, and Applications. He is the Chair of Steering Committee of the ACM International Conference on Multimedia Retrieval (ICMR) and Multimedia Modeling (MMM) conference series. He is also the General Co-Chair of ACM Multimedia 2005, ACM CIVR (now ACM ICMR) 2005, ACM SIGIR 2008, and ACM Web Science 2015. He serves on the editorial boards of

four international journals. He is the Co-Founder of two technology startup companies in Singapore.

ACKNOWLEDGMENTS

This workshop and related research were supported by Sea-NExT++ Joint Lab.

REFERENCES

[1] Hui Cui, Lei Zhu, Jingjing Li, Yang Yang, and Liqiang Nie. 2019. Scalable deep hashing for large-scale social image retrieval. *IEEE Transactions on image processing* 29 (2019), 1271–1284.

[2] Yali Du, Yinwei Wei, Wei Ji, Fan Liu, Xin Luo, and Liqiang Nie. 2023. Multi-queue Momentum Contrast for Microvideo-Product Retrieval. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 1003–1011.

[3] Wei Ji, Long Chen, Yinwei Wei, Yiming Wu, and Tat-Seng Chua. 2022. MRTNet: Multi-Resolution Temporal Network for Video Sentence Grounding. *arXiv preprint arXiv:2212.13163* (2022).

[4] Wei Ji, Xi Li, Fei Wu, Zhijie Pan, and Yueting Zhuang. 2019. Human-centric clothing segmentation via deformable semantic locality-preserving network. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 12 (2019), 4837–4848.

[5] Wei Ji, Yicong Li, Meng Wei, Xindi Shang, Junbin Xiao, Tongwei Ren, and Tat-Seng Chua. 2021. VidVRD 2021: The Third Grand Challenge on Video Relation Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4779–4783.

[6] Wei Ji, Renjie Liang, Zhedong Zheng, Wenqiao Zhang, Shengyu Zhang, Juncheng Li, Mengze Li, and Tat-seng Chua. 2023. Are Binary Annotations Sufficient? Video Moment Retrieval via Hierarchical Uncertainty-based Active Learning. (2023).

[7] Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 2020. What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation. In *Proceedings of the 28th ACM International conference on Multimedia*. 3487–3495.

[8] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4654–4662.

[9] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2928–2937.

[10] Yaxin Liu, Jianlong Wu, Leigang Qu, Tian Gan, Jianhua Yin, and Liqiang Nie. 2022. Self-supervised Correlation Learning for Cross-Modal Retrieval. *IEEE Transactions on Multimedia* (2022).

[11] Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. *AAAI* (2023).

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*. 8748–8763.

[13] Xindi Shang, Yicong Li, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2021. Video visual relation detection via iterative inference. In *Proceedings of the 29th ACM international conference on Multimedia*. 3654–3663.

[14] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5100–5111.

[15] Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Tat-Seng Chua. 2022. Rethinking the two-stage framework for grounded situation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2651–2658.

[16] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022. Video as Conditional Graph Hierarchy for Multi-Granular Question Answering. AAAI.

[17] Shuyu Yang, Yinan Zhou, Yaxiong Wang, Yujiao Wu, Li Zhu, and Zhedong Zheng. 2023. Towards Unified Text-based Person Retrieval: A Large-scale Multi-Attribute and Language Search Benchmark. In *Proceedings of the 2023 ACM on Multimedia Conference*.

[18] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–10.

[19] Chenchen Ye, Lizi Liao, Fuli Feng, Wei Ji, and Tat-Seng Chua. 2022. Structured and natural responses co-generation for conversational search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 155–164.

[20] Xuzheng Yu, Tian Gan, Yinwei Wei, Zhiyong Cheng, and Liqiang Nie. 2020. Personalized item recommendation for second-hand trading platform. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3478–3486.

[21] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–23.

[22] Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: datasets, algorithms and challenges. *EMNLP* (2022).