International Journal of Multimedia Information Retrieval (2022) 11:523–524
https://doi.org/10.1007/s13735-022-00265-2

EDITORIAL

Special issue on cross-modal retrieval and analysis

Jianlong Wu[1,2] · Richang Hong[3] · Qi Tian[4]

Published online: 3 December 2022

With the development of the Internet and social media, a large amount of multimedia data are generated and uploaded every day. Although these multimedia data might have different modalities, such as texts, images, videos, and audio, there is a semantic correlation among them. Effective cross-modal and multi-modal learning imposes great opportunities for many practical applications, such as cross-modal retrieval, matching, recommendation, and classification, which play important roles in public security, social media, entertainment, healthcare, etc. However, due to the natural heterogeneous property of cross-modal data, it is very challenging to investigate the correlation among data of different modalities to deal with practical tasks.

This special issue aims to assemble recent advances in cross-modal retrieval and analysis to handle these existing problems and benefit relevant researchers. It is a joint special issue that cooperates with the China Multimedia Conference 2022. We received 36 submissions, and seven papers are selected for publication after at least double peer-review process. We are pleased to present them in the following.

In order to investigate the precise inter-modality relationship for cross-modal retrieval tasks, the paper, "Prototype Local-Global Alignment Network for Image-Text Retrieval" by L. Meng, F. Zhang, X. Zhang and C. Xu, presents a novel framework to jointly perform the fine-grained local alignment and high-level global alignment. On the one hand, prototype-based local alignment divides the region-word alignment into the region-prototype and word-prototype alignment, which can well bridge the modality gap and avoid cumbersome calculation. On the other hand, multi-scale global alignment can help to perceive hierarchical global information. Through joint learning, these two alignments can complement and boost performance for each other.

Cross-modal retrieval has many practical and downstream applications. For example, in the paper, "Who is Gambling? Finding Cryptocurrency Gamblers Using Multimodal Retrieval Methods" by Z. Huang, Z. Liu, J. Chen, Q. He, S. Wu, L. Zhu and M. Wang, the authors propose a new framework to tackle the problem of detecting gambling contracts and addresses from multi-modal data. It consists of three modules, including the EVM disassembler, opcode feature extractor, and contract classifier, where the memory-bank mechanism can well boost the classification results. It also releases a large-scale benchmark for evaluation and achieves good performance.

Multi-modal analysis and fusion lead to more remarkable results than the unimodal on various applications, since it can aggregate more information. For the task of micro-expressions recognition, the paper, "Your Heart Rate Betrays You: Multimodal Learning with Spatio-temporal Fusion Networks for Micro-expression Recognition" by R. Zhang, N. He, S. Liu, Y. Wu, K. Yan, Y. He and K. Lu, proposes a dual-stream multi-modal learning method to combine the heart rate and spatio-temporal features. An adaptive balancing method is also presented to handle the imbalance issue. Experiments on several datasets demonstrate the superiority of the proposed multi-modal fusion strategy.

For the task of visual dialog, the paper, "Multi-aware Coreference Relation Network for Visual Dialog" by Z. Zhang, T. Jiang, C. Liu and Y. Ji, introduces a novel network to handle the problem of coreference resolution in visual dialog from both textual and visual views. The model contains three modules, where the first one concentrates on the textual relations for semantic reasoning, the second one stores adaptive-visual relationships under the influence of semantic knowledge for information selection, and the third one generates relation-aware monolithic representation by fusing cross-modal features. Experimental results on VisDial show that the proposed dialog agent obtains reliable visual

✉ Richang Hong
hongrc.hfut@gmail.com

Jianlong Wu
jlwu1992@pku.edu.cn

Qi Tian
tian.qi1@gmail.com

[1] Harbin Institute of Technology (Shenzhen), Shenzhen, China

[2] Shandong University, Jinan, China

[3] Hefei University of Technology, Hefei, China

[4] Huawei Cloud & AI, Shenzhen, China

and textual coreferential comprehension as well as obvious improvement.

Towards agricultural obstacle detection, the paper, "Video Deblurring and Flow-Guided Feature Aggregation for Obstacle Detection in Agricultural Videos" by K. Cheng, X. Zhu, Y. Zhan and Y. Pei, combines video deblurring and object detection tasks for joint optimization based on RNN and flow-guided feature aggregation. They also propose a region-shared strategy to improve efficiency. Extensive experiments on the FieldSAFE and GOPRO datasets show that the proposed method achieves much better detection performance with less computational cost.

Contrastive learning attracts much attention recently since it can learn more discriminative feature representations without manual annotations. In the paper, "TCKGE: Transformers with Contrastive Learning for Knowledge Graph Embedding" by X. Zhang, Q. Fang, J. Hu, S. Qian and C. Xu, the authors incorporate it with transformers to learn complex semantics in multi-relational knowledge graphs. Specifically, a transformer-based deep hierarchical architecture is first designed to dynamically learn the embeddings of entities and relations. Then, they present a contrastive learning scheme to facilitate optimization by exploring the effectiveness of several different data augmentation strategies. Evaluation on two benchmark datasets shows their superiority over state-of-the-art methods.

The architecture of convolutional neural networks plays an important role in deep learning-based cross-modal learning. For this issue, the paper, "FDAM: Full-Dimension Attention Module for Deep Convolutional Neural Networks" by S. Cai, C. Wang, J. Ding, J. Yu and J. Fan, comes up with a lightweight full-dimensional attention module, where the generated 3-D attention maps have both channel-wise and spatial information interaction. They also present a generalized Elo rating algorithm to make use of the historical channel-wise information. More importantly, the proposed module can be seamlessly integrated into the end-to-end training of the CNN framework.

We would like to thank all the authors for their high-quality submissions and the expert reviewers for their insightful comments to this special issue. We also deeply appreciate the editorial office of IJMIR, especially Prof. Michael S. Lew, for their support.