ISSN: 2424-8460(O)
2251-2608(P)

Original Research Article

A review of multimodal representation learning

Xiangyu Chen[1], Qinglin Wang[2], Zhen Wang[1*]
[1] *College of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo, Zhejiang, 315100, China*
[2] *School of Statistics, University of International Business and Economics, Beijing, 100029, China*

Abstract: This paper provides an overview of multimodal representation learning, using modal combination types as the core classification framework. It systematically reviews representation learning methods, association mechanisms, and fusion strategies under different modal pairs. Unlike previous reviews based on model architectures or task categories, this paper adopts a bottom-up perspective on modal interactions, revealing how the collaborative characteristics of different modal pairs influence representation modelling and summarizing the technical evolution of cross-modal learning. Additionally, this paper summarizes the general paradigms and practical challenges in multi- modal architecture design, offering new perspectives for model optimization in complex scenarios.
Keywords: multimodal; representational learning; deep learning; modal combination

1. Introduction

Multimodal representation learning, which seeks to bridge the heterogeneity gap among modalities to exploit their correlations and specificities[1-3], has become a cornerstone for applications in autonomous systems, medical imaging, and so on. Prevailing surveys typically classify methods via two orthogonal views: One centered on architectural paradigms, which offer general templates but often overlook modality-specific constraints; the other organized by application domains, which prioritizes task adaptability but may sacrifice generality. Departing from these conventions, this survey introduces a systematic taxonomy structured explicitly by modality combinations. We comprehensively analyze five prevalent categories (1) Vision-language, (2) Audio-language, (3) Image-Audio, (4) Multi-Order Fusion, and (5) Special Sensing modalities—Thereby directly addressing the unique representational challenges and technical constraints inherent to each data pairing. For each category, we further dissect representative methods along architectural lines, offering a unified yet fine-grained perspective.

2. Vision-language modalities

In early studies, Kernel Correlation Component Analysis (KCCA)[4] was employed to map image and text features onto a shared latent space for multimodal representation learning. However, KCCA primarily relies on kernel expansion methods based on linear correlations, making it difficult to fully capture the complex nonlinear and higher-order semantic relation- ships between images and text. To overcome this limitation, researchers subsequently introduced Deep Boltzmann Machines (DBM)[5]. As a deep generative model, DBM can model the joint distribution of images and text, learning cross-modal shared latent representations through a multi-layer latent variable structure.

With the rapid advancement of deep learning in computer vision and natural language processing, researchers have progressively moved away from traditional shallow models toward deep neural networks capable of automatically extracting image and text features, enabling multimodal alignment within a unified embedding space. For the image modality, Convolutional Neural Networks (CNNs) are commonly employed for feature extraction (e.g.,VGGNet[6], ResNet[7] and GoogLeNet[8]). For text modalities with sequential characteristics, recurrent neural networks (RNNs) and their variants are employed to capture context dependencies.

After, transformers have emerged as the mainstream framework for constructing multimodal representation

ISSN: 2424-8460(O)
2251-2608(P)

learning models. Based on different multimodal interaction mechanisms, Transformer fusion strategies can be primarily categorized into the following two types: (1) Dual-stream cross-modal fusion mechanism: This mechanism feeds different modality data into independent Transformer branches, employing cross-attention at intermediate layers, such as ViLBERT[9,10]. (2) Single-stream cross-modal fusion mechanism concatenates images and text into a unified sequence at the token level, feeding it into the Transformer encoder, such as VisualBERT [11], VL-BERT [12], UNITER[13], Unicoder-VL[14], and Multitrans[15] in clinical settings. Beyond single-stream and dual-stream attention mechanisms, researchers have explored hybrid architectures combining Transformers with other fusion strategies to enhance multimodal representation capabilities[16,17]. This approach fuses text vectors extracted by BERT with image vectors extracted by ViT through tensor product fusion.

Faced with challenges in training scale and model generalization capabilities, researchers have turned to joint pre-training on large-scale image-text pairs and introduced efficient strategies such as contrastive learning. The core idea of contrastive learning is to construct positive and negative sample pairs, minimizing semantic relevance between related samples and maximizing irrelevance between unrelated samples within a unified embed- ding space, thereby achieving cross-modal fine-grained alignment. Such as CLIP[18]. Building upon CLIP, BLIP[19] introduces three complementary tasks: ITC (Image-Text Contrastive), ITM (Image-Text Matching), and LM (Language Modeling). It designs a visual-text hybrid encoder that balances alignment, understanding, and generation, further enhancing performance in downstream retrieval, question-answering, and caption generation tasks.

3. Audio-language modalities

Compared to vision-language multimodal representation learning, audio- language modalities face greater challenges in alignment methods. Audio is a continuous, high-frequency signal that can contain, for example, up to 16,000 frames per second at a sampling rate of 16kHz, whereas text is a discrete, low-frequency sequence, with each sentence usually consisting of only a few words.

To address the lack of robustness of early fusion methods in the face of modal noise, attention mechanisms have been introduced in recent years to achieve dynamic inter-modal fusion. The attention mechanism in the Transformer architecture can adaptively focus on the more reliable and informative parts of the audio and text modalities, which can effectively mitigate the interference of noise on the learning of multimodal representations. For example, the InterVoxNet model proposed by Ding et al. in[20] constructs a bimodal fusion architecture that fuses Transformer with LSTAFN (Long-Short-Term Attention Fusion Network) for the depression detection task. Ultimately, the joint temporal information is further integrated by a second- stage bidirectional LSTM to obtain a multimodal representation that is synergistically enhanced by semantic and temporal information. In addition, Luong et al. proposed a new framework for audio-text cross-modal retrieval task, mini-batch Learning to Match (m-LTM), in[21].

Inspired by the image-text multimodal contrast learning model CLIP, Elizalde et al. proposed the audio-text contrast learning architecture CLAP (Contrastive Language-Audio Pretraining)[22]. By introducing a contrast loss function, CLAP achieves the learning of joint representations between auditory and linguistic modalities, enabling the model to obtain more discriminative cross-modal embedding representations on large-scale unsupervised audio-text pairing data, thus demonstrating superior performance in tasks such as audio retrieval, audio classification, and cross-modal understanding.

4. Conclusion and future directions

Based on the classification perspective of topology, combined with typical models and methods, the similarities and differences of each modality in terms of representation, fusion mechanism and application strategy are discussed in depth, which reveals the intrinsic connection and evolutionary vein of the current mainstream technologies. In addition, this paper combs through the development history of multimodal representation learning, which is divided into four major stages: starting from the shallow representation method represented by linear dimensionality reduction, and gradually transitioning to the complex cross-modal modelling system based on deep neural networks, which reflects the technological leap from the primary unimodal

ISSN: 2424-8460(O)
2251-2608(P)

modelling to the deep multimodal semantic understanding in this field. Overall, with the deepening knowledge of the topological relationship between multimodalities and the continuous improvement of model structure and learning ability, multimodal representation learning is gradually becoming one of the core technologies to support intelligent systems to achieve semantic understanding, task generalization and cross-domain migration. In the future, this direction will still show broad application prospects and far-reaching research value in key fields such as multimodal intelligences, human-computer interaction, automatic driving, and medical diagnosis.

Fundings

A project supported by scientific research fund of Zhejiang provincial education department (Number: Y202455079). Provincial undergraduate training program on innovation and entrepreneurship (Number: S202410876066).

About the author

Xiangyu Chen（2003.10.11), male, Han, native place: Zhangzhou, Fujian, bachelor's degree, research field: Medical imaging.

Zhen Wang (2005.11.05), male, Han, native place: Daqing Heilongjiang, bachelor's degree, research field: Medical imaging.

Author contribution

All authors contributed to the conception and design of the study. Literature search and analysis were performed by Xiangyu Chen. The manuscript was drafted by Zhen Wang and critically revised by Qinglin Wang. All authors read and approved the final version of the manuscript.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[2] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, Zheng, X. Xia, et al., Glm-130b: An open bilingual pre-trained model, arXiv preprint arXiv:2210.02414 (2022).

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[4] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, Journal of Artificial Intelligence Research 47 (2013) 853–899.

[5] N. Srivastava, R. R. Salakhutdinov, Multimodal learning with deep boltzmann machines, Advances in neural information processing sys- tems 25 (2012).

[6] K. Simonyan, A. Zisserman, Very deep convolutional networks for large- scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[7] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Er- han, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[9] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Advances in neural information processing systems 32 (2019).

[10] X. Zhan, Y. Wu, X. Dong, Y. Wei, M. Lu, Y. Zhang, H. Xu, X. Liang, Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining, in: Proceedings of the IEEE/CVF interna- tional conference on computer vision, 2021, pp. 11782–11791.

[11] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, arXiv preprint arXiv:1908.03557 (2019).

ISSN: 2424-8460(O)
2251-2608(P)

[12] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, Vl-bert: Pre-training of generic visual-linguistic representations, arXiv preprint arXiv:1908.08530 (2019).

[13] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng,J. Liu, Uniter: Universal image-text representation learning, in: Euro- pean conference on computer vision, Springer, 2020, pp. 104–120.

[14] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training, in: Pro- ceedings of the AAAI conference on artificial intelligence, Vol. 34, 2020, pp. 11336–11344.

[15] D. Ma, M. Wang, A. Xiang, Z. Qi, Q. Yang, Transformer-based classi- fication outcome prediction for multimodal stroke treatment, in: 2024 IEEE 2nd International Conference on Sensors, Electronics and Com- puter Engineering (ICSECE), IEEE, 2024, pp. 383–386.

[16] F. Alzamzami, A. El Saddik, Transformer-based feature fusion approach for multimodal visual sentiment recognition using tweets in the wild, IEEE Access 11 (2023) 47070–47079.

[17] A. Xiang, Z. Qi, H. Wang, Q. Yang, D. Ma, A multimodal fusion net- work for student emotion recognition based on transformer and tensor product, in: 2024 IEEE 2nd International Conference on Sensors, Elec- tronics and Computer Engineering (ICSECE), IEEE, 2024, pp. 1–4.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transfer- able visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[19] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre- training for unified vision-language understanding and generation, in: International conference on machine learning, PMLR, 2022, pp. 12888– 12900.

[20] H. Ding, Z. Du, Z. Wang, J. Xue, Z. Wei, K. Yang, S. Jin, Z. Zhang, J. Wang, Intervoxnet: a novel dual-modal audio-text fusion network for automatic and efficient depression detection from interviews, Frontiers in Physics 12 (2024) 1430035.

[21] M. Luong, K. Nguyen, N. Ho, R. Haf, D. Phung, L. Qu, Revisiting deep audio-text retrieval through the lens of transportation, arXiv preprint arXiv:2405.10084 (2024).

[22] B. Elizalde, S. Deshmukh, M. Al Ismail, H. Wang, Clap learning audio concepts from natural language supervision, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Pro- cessing (ICASSP), IEEE, 2023, pp. 1–5.

[23] K. Sung-Bin, A. Senocak, H. Ha, A. Owens, T.-H. Oh, Sound to visual scene generation by audio-to-visual latent alignment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni- tion, 2023, pp. 6430–6440.

[24] T. Feng, Y. Xie, X. Guan, J. Song, Z. Liu, F. Ma, F. Yu, Unisync: A unified framework for audio-visual synchronization, arXiv preprint arXiv:2503.16357 (2025).