(IJRST) 2022, Vol. No. 12, Issue No. III, Jul-Sep **e-ISSN: 2249-0604, p-ISSN: 2454-180X**

An In-Depth Analysis of the Multimodal Representation Learning with Respect to the Applications and Linked Challenges in Multiple Sectors[1]

Arnav Goenka

Vellore Institute of Technology, Vellore, Tamil Nadu, India

DOI:10.37648/ijrst.v12i03.009

Received: 31 August 2022; Accepted: 27 September 2022; Published: 03 October 2022

ABSTRACT

Representation learning is a machine learning type wherein a system automatically uses deep models to extract features from raw data. It is essential for tasks like classifications, regression, and identification. Multimodal representation learning is a subset of representation learning that focuses on feature extraction from several heterogeneous, interconnected modalities. Although these modalities are frequently heterogeneous, they show correlations and relationships. These modalities include text, images, audio, or videos. Several difficulties arise from this intrinsic complexity, including combining multimodal data from various sources by precisely characterizing the relationships and correlations between modalities and jointly deriving features from multimodal data. Researchers are becoming increasingly interested in these problems, particularly as deep learning gains momentum. In recent years, many deep multimodal learning techniques have been developed. We present an overview of deep multimodal learning in this study, focusing on techniques that have been proposed in the past decade. We aim to provide readers with valuable insights for researchers, especially those working on multimodal deep machine learning, by educating them on the latest developments, trends, and difficulties in this field.

INTRODUCTION

To facilitate the extraction of important information or features for the construction of learning models, such as classifiers, regressors, and recognizers, representation learning aims to learn efficient representations of raw data automatically[1]. Representation learning is an essential machine learning data preparation technique that has a direct impact on learning model performance. In the big data era, information of the same event or phenomena is frequently originating from several sources and modalities, including text, video, audio, and more. As a result, representation learning for multimodal data has emerged as an intriguing field that poses novel and important difficulties because of the variety of the data.

Fusion is the general processing approach for multimodal data[2,3]. Either the feature space or the data space can experience multimodal data fusion. The former incorporates raw data straight from sensors and is referred to as raw data fusion or early data fusion. Determining the proper sampling rates between various sensors and synchronizing data from multiple sources are two of the issues this technology tackles. Raw data fusion is deemed inferior to intermediate data fusion, which takes place in the feature space. Because of its many uses, it has garnered a lot of interest recently and is currently a popular study topic in deep learning. The two main types of methods for fusing intermediate data are joint representation approaches and coordinated representation techniques.

A few review articles exist on multimodal machine learning; nevertheless, they mainly deal with

[1] *How to cite the article:*
Goenka A, An In-Depth Analysis of the Multimodal Representation Learning with Respect to the Applications and Linked Challenges in Multiple Sectors, IJRST, Jul-Sep 2022, Vol 12, Issue 3, 50-57, DOI: http://doi.org/10.37648/ijrst.v12i03.009

multimodal data fusion. Our goal in this study is to offer an extensive overview of multimodal representation learning, including techniques and applications (refer to Figure 1). In addition to summarizing current developments, we also discuss the problems and tendencies in this area [2,4]. This publication seeks to provide insightful information to scholars working on relevant projects.

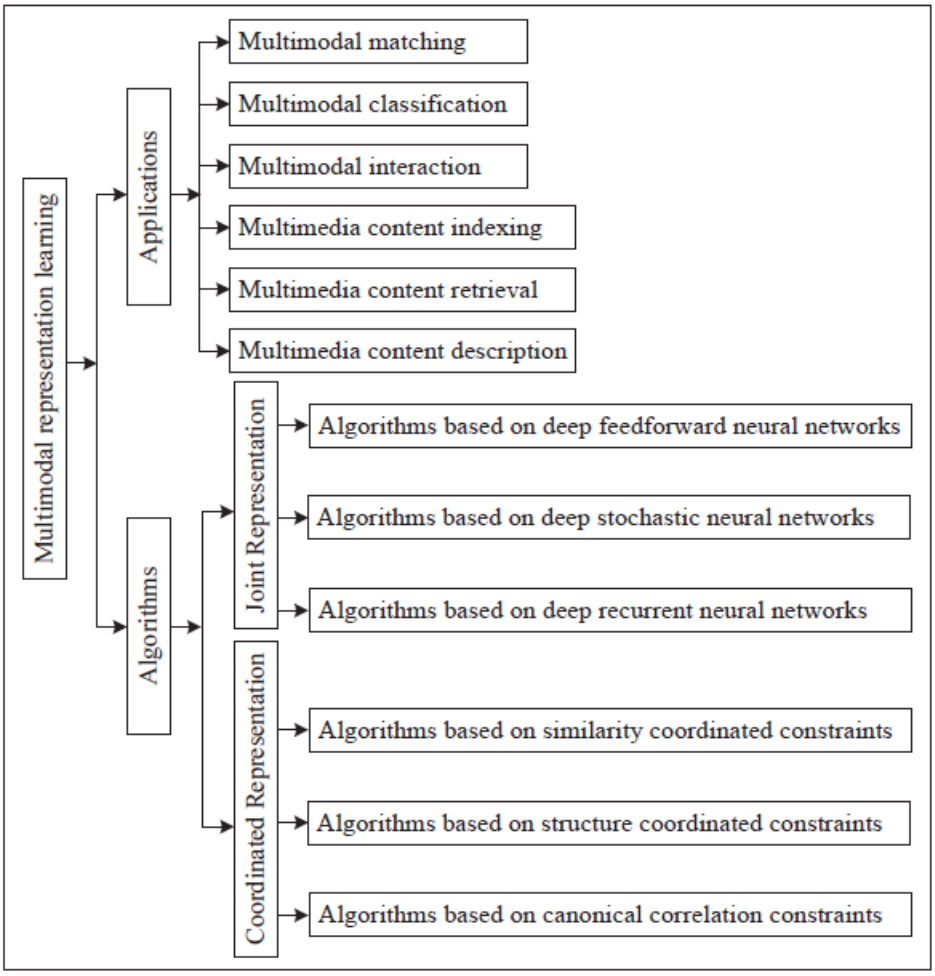

FIGURE 1. Applications and algorithms of multimodal representation learning

APPLICATION-DRIVEN ALGORITHM DESIGN

Applications drive the design of many different algorithms for multimodal representation learning. The main applications include indexing, retrieval, description, matching and classification of multimodal contexts and multimodal interactions. (refer figure1)

Multiple-Modal Pairing: The aim is to establish links between various modalities such as text and visual. Similarly, Zhao et al. [5] presented a technique called Multimodality Robust Line Segment (MRLS). This method creates MRLS descriptors using very similar corners and line segments from two multimodal images. Image matching is then done by measuring resemblance between these descriptors. Zhang et al. [6] introduced a multimodal matching approach that integrates feature extraction and matching into a single convolutional neural network (CNN) framework. This CNN considers every modal dataset as a separate class and is trained to minimize multiclassification loss in cross-modal matching studies where Pitts et al. [7] found that even slight differences can have large effects on match values and intra-individual match variability.

Multimodal Classification: Choi and Lee [8] proposed a deep learning-based multimodal fusion architecture for classification problems. This method has two main

advantages: it can work with any type of learning model; it effectively handles modal loss. Bahrampour et al. [9] put forward a task-driven dictionary learning algorithm for multi-modal data representation with joint sparsity constraint which makes representations from different modality sources share same sparse pattern thus enabling complementary processing between them instead of treating each modality separately. A significant benefit of this algorithm is its efficiency in jointly learning multi-modal dictionaries together with their respective classifiers. Gomez-Chova et al.[10] reviewed various approaches to remote sensing image multi-classification across seven challenging remote sensing applications.

According to Turk people naturally interact with the world through multiple senses [11]. With advances in hardware and software brought about by powerful mobile devices such as smartphones, tablets, wearables and other sensors, multimodal interaction has gained prominence. Multimodal interaction aims at enabling users to communicate with computers using text, audio, video among other modalities. Vidakis et al [12]. introduced a multimodal framework that supports the deployment of different modalities in blended learning environment.Mi et al. proposed a deep CNN-based architecture for learning human-centered object affordance which led to a multimodal fusion framework for effective object grasping [13].

Multimodal retrieval and multimodal indexing are closely related because indexing is used to accelerate retrieval [14]. In order for cross-modal retrieval to be possible, Hu et al. [14] proposed a Multi Adversarial Network (MAN) which has several modality-specific generators as well as a discriminator; additionally it also contains a multi-modal discriminant analysis loss function which allows it to project data from different modalities into one space so that similarity between them can be directly computed using just one distance measure. Shang et al. [15] combined GANs [16] with dictionary learning in their work on cross-modal retrieval called DLA-CMR (Dictionary Learning Algorithm for Cross-Modal Retrieval); here adversarial learning extracts statistical characteristics for every modality while at the same time dictionary learning reconstructs discriminative features. Cao et al.'s hybrid representation learning approach involves three steps: learning hybrid representations by means of joint autoencoder and feedforward neural network, deep Restricted Boltzmann Machines (RBMs) being used to extract modality-specific features, stacked bimodal autoencoders being employed to derive final shared representations for each modality [17]. A unified framework for multimodal retrieval was proposed by D. Rafailidis et al. [18]. Different data modalities were integrated by this framework aiming at improving the accuracy and efficiency of retrieval systems. Park et al.'s work involved two social network tasks: customized picture captioning problem solving through post creation by generating genuine text caption for the post; hashtag prediction where relevant hashtags were predicted and suggested [19]. Niu et al.'s study on picture annotation introduced a large-scale image annotation method based on multi-scale deep learning which focuses on multimodal feature representation and determination of optimal number of class labels needed to ensure accurate and meaningful annotations are made [20]. An image captioning method was developed by Zhao et al., which is based on multimodal fusion [21]. To make the video captioning process efficient and accurate, Wu et al. proposed hierarchical attention-based multimodal fusion to videos [22]. Chou et al. created technique for multimodal video-to-near-scene annotation [23].

Apart from the mentioned applications, there are other areas where multimodal representation learning can be useful such as multisensory systems, healthcare and medical image processing.

ALGORITHMS ADDRESSING APPLICATION PROBLEMS

Researchers have created a wide range of multimodal representations learning algorithms in the last decade. Joint representation algorithms and Coordinated representation algorithms are the two basic categories into which these algorithms can be widely divided. Three distinct types can be found within each of these groups (refer to Figure 1).

i)Combined Representation Techniques: These algorithms project the unimodal data onto a shared space, so combining the data into a single representation space. Figure 2 depicts the idea of collaborative representation learning.

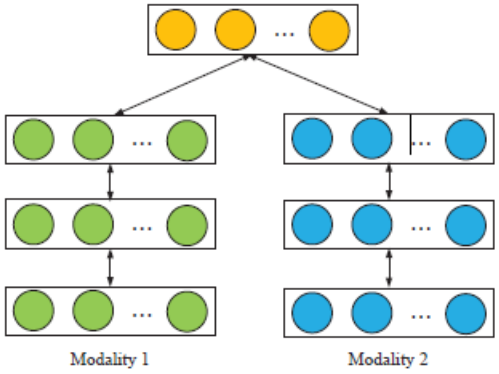

FIGURE 2. The diagram of joint representation learning

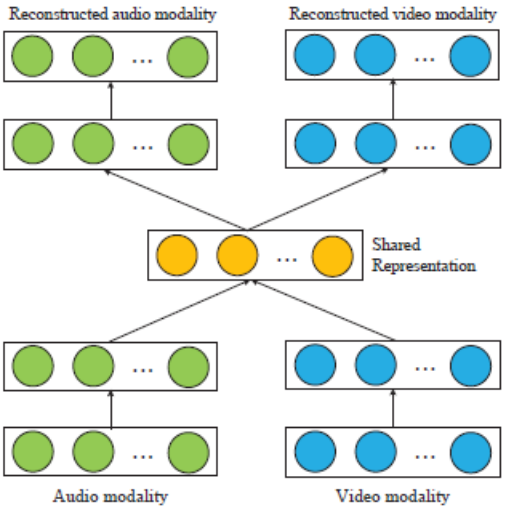

FIGURE 3. the idea of multimodal deep learning in [27]

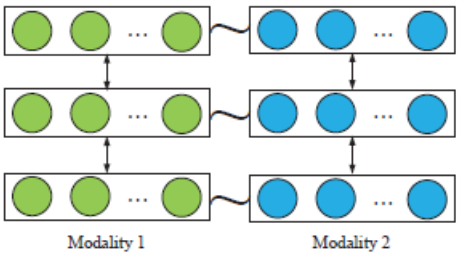

FIGURE 4. The diagram of coordinated representation learning

a) Deep Feedforward Neural Networks (DFNN)

Ngiam et al. [24] carried out the groundbreaking work in the framework of multimodal representation learning based on deep feedforward neural networks (DFNN). In their work, they automatically learned features from multimodal data using a deep stacked autoencoder. The method used in [24] is simple and is illustrated in Figure 3. Shekhar et al. [25] developed a combined sparse representation technique for the recognition of multimodal biometric data, building on this technological basis. Gu et al. [26] created a technique for learning joint multimodal representations that is based on multi-fusion deep neural networks.

b) Deep Stochastic Neural Networks (DSNN)

Srivastava and Salakhutdinov [27] presented the ground-breaking approach within the context of multimodal representation learning based on deep stochastic neural networks (DSNN). Deep Boltzmann machines were employed by their DSNN to extract useful characteristics from multimodal data. Extending this methodology, Sohn et al. [28] presented an enhanced algorithm that integrates information variation, while Amer et al. [29] suggested a hybrid technique for deep multimodal data fusion, augmenting the potential of DSNNs in managing intricate multimodal datasets.

c) Deep Recurrent Neural Networks (DRNN)

A number of noteworthy research have been produced under the paradigm of multimodal representation learning utilizing deep recurrent neural networks (DRNN). Long short-term memory (LSTM) networks were extended to multimodal representation learning by Rajagopalan et al. [30]. Multimodal recurrent neural networks were used by Feng et al. [31] for audio-visual speech recognition. For the purpose of labelling interior sceneries, Abdulnabi et al. [32] suggested a multimodal recurrent neural network with information transfer layers.

ii) Coordinated Representation Learning

Algorithms for Coordination Representation The material that is presented does not go into detail about these algorithms, but generally speaking, they entail coordinating and aligning features from several modalities without necessarily combining them into a single shared space. This makes it possible for each modality to have its distinctive qualities while still being incorporated into a well-rounded model.

Coordinated Representation Learning is a technique in machine learning where multiple models or agents learn representations simultaneously, ensuring their learned features are aligned or complementary. This coordination enhances the overall performance by leveraging diverse perspectives or data sources, promoting robustness and generalization [33]. Typically employed in multi-view learning, multi-task learning, and federated learning scenarios, the method involves sharing information between models through techniques like co-training, consensus regularization, or shared latent spaces. Coordinated representation learning is particularly effective in complex tasks requiring integration of heterogeneous data, leading to more comprehensive and insightful models.

a) Similarity-Coordinated Constraints

Similarity-Coordinated Constraints (SCC) are a technique in optimization and machine learning used to maintain coherence and consistency within data clusters or groups. By enforcing constraints that ensure similar data points or entities are treated alike, SCC improves the quality of clustering, classification, and other data analysis processes. This method balances flexibility and structure, allowing models to adapt to underlying patterns while preserving essential relationships [34]. SCC is particularly useful in applications like image recognition, natural language processing, and recommendation systems, where maintaining the integrity of similar items is crucial for accurate and reliable outcomes[35,36].

b) Structure-Coordinated Constraints

The framework was pioneered by Bronstein et al. [37], who enforced structure coordinate constraints using similarity-sensitive hashing. Their approach was predicated on similarity-sensitive functions, in contrast to conventional hashing. A hashing function for cross-view similarity search was created by Kumar and Udupa [38] and later extended to deep learning settings by Jiang and Li [39].

c) Canonical-Coordinated Constraints

Canonical Correlation Analysis (CCA) is used by methods in this category to extract important and pertinent information for later tasks. A feature extraction approach called FaRoC, which combines CCA and rough sets, was first presented by Mandal and Maji [40]. A category-based deep CCA for fine-grained venue discovery from multimodal data was proposed by Yu et al. [41]. In order to recognise human actions, Elmadany et al. [42] created a multimodal feature learning method based on bimodal/multimodal hybrid centroid CCA.

CONCLUSIONS

This paper presented a comprehensive survey on multimodal representation learning, including its applications and the underlying algorithms. Applications are what drives this field, and in order to overcome certain practical challenges a variety of algorithms have been created. The author have identified five trends or difficulties within multimodal representation learning.

1.Handling Modality Miss using Deep Generative Models: This refers to any situation where some modalities are missing, one approach is by employing

generative adversarial networks (GANs) and variational autoencoders (VAEs).

2.Multimodal Similarity Preserving Hashing via Deep Representation Learning: It means coming up with efficient representations that support similarity across various modalities.

3.Measuring Correlation Between Different Modalities: Finding robust methods for estimating how much different modalities correlate with each other still poses major challenges.

4.Learning Joint Distributions over Different Modalities How do different modalities interact? Models need to be designed which can capture these joint distributions effectively.

5.Learning Conditional Distributions Given One Modality: In tasks where we use information from one modality to predict or infer another, it becomes important to know the conditional distribution of one modality given another.

These points highlight where researchers should concentrate their efforts next while also indicating just how complicated things can get when working on multimodal representation learning systems.

REFERENCES

[1] Y. Bengio, A. Courville, P. Vincent. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(8):1798-1828.

[2] D. Lahat, T. Adali, C. Jutten. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. Proceedings of the IEEE, 2015, 103(9):1449-1477.

[3] Y. Zheng. Methodologies for Cross-Domain Data Fusion: An Overview. IEEE Transactions on Big Data, 2015, 1(1):1-14.

[4] D. Ramachandram, G. W. Taylo. Deep Multimodal learning: a survey on recent advances and trends. IEEE Signal Processing Magazine, 2017, 34(6):96-108.

[5] C. Zhao, H. Zhao, J. Lv, et al. Multimodal image matching based on Multimodality Robust Line Segment Descriptor. Neurocomputing, 2016, 177:290-303.

[6] Y. Zhang, Y. Gu, X. Gu. Two-Stream Convolutional Neural Network for Multimodal Matching. International Conference on Artificial Neural Networks (ICANN18), 2018, Pages:14-21.

[7] B. Pitts, S. L. Riggs, N. Sarter. Crossmodal Matching: A Critical but Neglected Step in Multimodal Research, IEEE Transactions on Human-Machine Systems, 2016, 46(3):445-450.

[8] J. H. Choi, J. S. Lee. EmbraceNet: A robust deep learning architecture for multimodal classification. Information Fusion, 2019, 51:259-270.

[9] S. Bahrampour, N. M. Nasrabadi, A. Ray, et al. Multimodal Task-Driven Dictionary Learning for Image Classification. IEEE Transactions on Image Processing, 2015, 25(1):24-38.

[10] L. Gomez-Chova, D. Tuia, G. Moser, et al. Multimodal Classification of Remote Sensing Images: A Review and Future Directions. Proceedings of the IEEE, 2015, 103(9):1-25.

[11] M. Turk. Multimodal interaction: A review. Pattern Recognition Letters, 2014, 36:189-195.

[12] N. Vidakis, K. Konstantinos, G. Triantafyllidis. A Multimodal Interaction Framework for Blended Learning. International Conference on Interactivity, Game Creation, Design, Learning, and Innovation, Pages:2016, 205-211.

[13] J. Mi, S. Tang, Z Deng, et al. Object affordance based multimodal fusion for natural Human-Robot interaction. Cognitive Systems Research, 2019, 54:128-137.

[14] P. Hu, D Peng, X. Wang, et al. Multimodal adversarial network for cross-modal retrieval. Knowledge-Based Systems, online first, 2019, https://doi.org/10.1016/j.knosys.2019.05.017.

[15] F. Shang, H. Zhang, L. Zhu, et al. Adversarial cross-modal retrieval based on dictionary learning. Neurocomputing, 2019, online first, https://doi.org/10.1016/j.neucom.2019.04.041.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. Generative adversarial nets. In: Proceedings of the 2014 Conference on Advances in Neural Information Processing Systems 27. Montreal, Canada: Curran Associates, Inc., 2014. 2672-2680.

[17] W. Cao, Q. Lin, Z. He, et al. Hybrid representation learning for cross-modal retrieval. Neurocomputing, 2019, 345:45-57.

[18] D. Rafailidis, S. Manolopoulou, P. Daras. A unified framework for multimodal retrieval. Pattern Recognition, 2013, 46:3358-3370.

[19] C. C. Park, B. Kim, G. Kim. Towards Personalized Image Captioning via Multimodal Memory Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(4):999-1012.

[20] Y. Niu, Z. Lu, J. R. Wen, et al. Multi-Modal Multi-Scale Deep Learning for Large-Scale Image Annotation. IEEE Transactions on Image Processing, 2019, 28(4):1720-1731.

[21] D. Zhao, Z. Chang, S. Guo. A multimodal fusion approach for image captioning. Neurocomputing, 2019, 329:476-485.

[22] C. Wu, Y. Wei, X. Chu, et al. Hierarchical attention-based multimodal fusion for video captioning. Neurocomputing, 2018, 315:362-370.

[23] C. L. Chou, H. T. Chen, S. Y. Lee. Multimodal Video-to-Near-Scene Annotation. IEEE Transactions on Multimedia, 2017, 19(2):354-366.

[24] J. Ngiam, A. Khosla, M. Kim, et al. Multimodal deep learning. in Proc. 28th Int. Conf. Machine Learning (ICML-11), 2011, pp. 689-696.

[25] S. Shekhar, V. M. Patel, N. M. Nasrabadi, et al. Joint Sparse Representation for Robust Multimodal Biometrics Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(1):113-126.

[26] Z. Gu, B. Lang, T. yue, et al. Learning Joint Multimodal Representation Based on Multi-fusion Deep Neural Networks. International Conference on Neural Information Processing, 2017, pp.276-285.

[27] N. Srivastava, R. R. Salakhutdinov. Multimodal learning with deep Boltzmann machines. in Proc. Advances in Neural Inform. Processing Syst., 2012, pp. 2222-2230.

[28] K. Sohn, W. Shang, H. Lee. Improved multimodal deep learning with variation of information. in Proc. Advances in Neural Information Processing Systems., 2014, pp. 2141-2149.

[29] M. R. Amer, T. Shields, B. Siddiquie, et al. Deep Multimodal Fusion: A Hybrid Approach. International Journal of Computer Vision, 2018, 126(2-4):440-456.

[30] S. S. Rajagopalan, L. P. Morency, T. Baltrusaitis, et al. Extending long short-term memory for multi-view structured learning. in Proc. Eur. Conf. Comput. Vis., 2016, pp. 338-353.

[31] W. Feng, N. Guan, Y. Li, et al. Audio visual speech recognition with multimodal recurrent neural networks. 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14-19 May 2017, 681-688.

[32] A. H. Abdulnabi, B. Shuai, Z. Zuo, et al. Multimodal Recurrent Neural Networks with Information Transfer Layers for Indoor Scene Labeling. IEEE Transactions on Multimedia, 2018, 20(7):1656-1671.

[33] J. Weston, S. Bengio, N. Usunier. WSABIE: Scaling up to large vocabulary image annotation. In Proc. 7th Int. Joint Conf. Artif. Intell., 2011, pp. 2764-2770.

[34] A. Frome, G. Corrado, J. Shlens. DeViSE: A deep visualsemantic embedding model. in Proc. 28th Int. Conf. Neural Inf. Process. Syst., 2013, pp. 2121-2129.

[35] R. Kiros, R. Salakhutdinov, R. S. Zemel. Unifying visualsemantic embeddings with multimodal neural language models. Trans. Assoc. Comput. Linguistics, 2015, pp. 1-13.

[36] Y. Pan, T. Mei, T. Yao, et al. Jointly modeling embedding and translation to bridge video and language. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4594-4602.

[37] M. M. Bronstein, A. M. Bronstein, F. Michel, et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2010, pp. 3594-3601.

[38] S. Kumar, R. Udupa. Learning hash functions for cross-view similarity search. in Proc. 7th Int. Joint Conf. Artif. Intell., 2011, pp. 1360-1365.

[39] Q. Y. Jiang, W. J. Li. Deep cross-modal hashing. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 3270-3278.

[40] A. Mandal, P. Maji. FaRoC: Fast and Robust Supervised Canonical Correlation Analysis for Multimodal Omics Data. IEEE Transactions on Cybernetics, 2018, 48(4):1229-1241.

[41] Y. Yu, S. Tang, K. Aizawa, et al. Category-Based Deep CCA for Fine-Grained Venue Discovery from Multimodal Data. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(4):1250-1258.

[42] N. E. D. Elmadany, Y. He, L. Guan. Multimodal Learning for Human Action Recognition Via Bimodal/Multimodal Hybrid Centroid Canonical Correlation Analysis. IEEE Transactions on Multimedia, 2019, 21(5):1317-1331.