An Overview of Multimodal Fusion Learning

Fan Yang, Bo Ning(✉), and Huaiqing Li

Dalian Maritime University, Dalian 116000, Liaoning, China
ningbo@dlmu.edu.cn

Abstract. With the rapid development of modern science and technology, information sources have become more widely available and in more diverse forms, resulting in widespread interest in multimodal learning. With the various types of information captured by humans in understanding the world and perceiving objects, a single modality cannot provide all of the information about a specific object or phenomenon. Multimodal fusion learning opens up new avenues for tasks in deep learning, making them more scientific and human in their approach to solving many real-world problems. An important challenge confronting multimodal learning today is how to efficiently facilitate the fusion of multimodal features while retaining the integrity of modal information to reduce information loss. This paper summarizes the definition and development process of multimodality, analyzes and discusses briefly the main approaches to multimodal fusion, common models, and current specific applications, and finally discusses future development trends and research directions in the context of existing technologies.

Keywords: Multimodal learning · Multimodal fusion · Deep learning

1 Introduction

The goal of multimodal learning is to learn and understand a variety of different types of information. With the rapid development of deep learning in recent years, multimodal fusion has become a popular topic. Philosophers and artists used the term "multimodality" to define forms of expression and rhetorical methods that fused different contents as early as the fourth-century BC [1,2]. Since the twentieth century, the widespread use of the web and mobile devices has made multimodal data the primary form of data resource in recent times, and research into multimodal learning is critical for computers to understand heterogeneous data from multiple sources. Multimodal learning is currently used in a variety of applications, including face recognition [3], visual question answering [4], image captioning [5], sentiment analysis [6], and multimodal retrieval [7].

Y. Chenggang et al. (Eds.): MobiMedia 2022, LNICST 451, pp. 259–268, 2022.
https://doi.org/10.1007/978-3-031-23902-1_20

2 Definition and Development Process of Multimodal Learning

2.1 Definition of Multimodal Learning

The term "modality" encompasses a wide range of concepts. Humans gather information about a thing through sight, hearing, touch, and smell. A modality is a term used to describe any method of obtaining information. Likewise, sounds, images, and text obtained in various ways can be considered modalities. A typical form of multimodal information is depicted in Fig. 1.

While unimodal representation learning aims to convert information into numerical or feature vectors for further computer processing, multimodal learning improves the ability to understand and learn multimodal information by leveraging complementarity and filtering redundancy. Recent examples usually involve multimodal learning with images, text, and sounds.

Fig. 1. Multimodal data for a "Autumn Streets" scene (image, sound and text)

2.2 Development of Multimodal Learning

Multimodal learning methods have made significant progress in several areas of intelligent information processing since the 1980s. McGurk et al. proposed the effect of vision on speech perception in 1976, which was used in Audio-Visual Speech Recognition (AVSR) technology [8] and served as a prototype for the multimodal concept. Many computer scientists, influenced by the McGurk effect, have worked on developing multimodal speech recognition systems based on vision and hearing, such as lip-sound speech recognition systems [9], which can effectively improve recognition accuracy when compared to audio alone. Atrey et al. classified existing multimodal fusion methods into fusion methods and fusion levels in 2010 [10]. Wang proposed the Deep Multimodal Hashing with Orthogonal Regularization Constraints (DMHOR) method [11] in 2015 as a method for reducing information redundancy in multimodal representations. Zhang et al. [12] and Wang et al. [13] have since made significant contributions to cross-modal information matching and retrieval; Liu et al. have analyzed and studied visual and haptic data and applied it to integrated robotic perception scenarios [14], and Fu et al. have made significant advances in the field of semantic annotation of images [15].

3 Multimodal Fusion Methods

Multimodal fusion is a key component of multimodal learning and is broadly classified into two types: model-independent fusion methods and model-based fusion methods. Model-independent methods do not directly rely on deep learning methods, which are simple but less practical in terms of information loss during fusion; Model-based methods, which are complex but more accurate, include the Multiple Kernel Learning (MKL) methods, Graphical Model (GM) methods, and the Neural Network (NN) methods.

3.1 Model-Independent Fusion Methods

Model-independent fusion methods are classed as early fusion, late fusion, and hybrid fusion depending on when the fusion happens. Early fusion incorporates features just after they're extracted, late fusion does so after each model's output, and hybrid fusion combine the advantages of the first two.

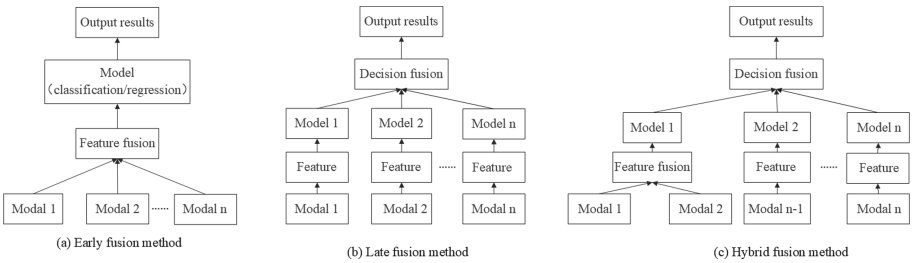

Fig. 2. Model-independent fusion methods

Early Fusion Methods. Early fusion is the technique of doing the fusion on both the feature and data levels immediately after feature extraction, most typically by using a simple join operation on the features. Early fusion methods can take advantage of the correlation and interactions between low-level elements of each modality when the modalities are highly linked. However, at the feature and data levels, this correlation is more difficult to extract, and Hinton et al. argue that the information contained in data from different modalities can only be correlated at a higher level [16]. According to Martinez et al., early fusion of multimodal data may not fully exploit the complementarity between the modalities and may even result in redundant vector inputs [17]. As a result, early fusion is not the best fusion procedure.

Furthermore, the issue of time synchronization between multimodal features must be considered during early fusion. Convolution and pool fusion, which can combine discrete event sequences with continuous signals, was proposed by ROBIN et al. to solve the time synchronization problem [18]. The structure of early fusion methods is depicted in Fig. 2(a), in which extracted features are fused directly first, then features from different modalities are integrated for model training, and finally the final output results.

Late Fusion Methods. Late fusion, also known as decision level fusion, entails first training a separate model for each modality and then fusing the outputs of several models. Late fusion methods, such as Bayesian rule fusion, maximum fusion, mean fusion, and other rule fusion methods, currently rely on rules to decide the combination of distinct model outputs [19].

When compared to previous fusion approaches, this sort of fusion can manage simple data asynchrony while providing greater flexibility in selecting the best-suited method for each modality to be analyzed, such as Hidden Markov Models (HMM) for audio and Support Vector Machines (SVM) for images. However, it ignores the low-level interaction of several modalities, making fusion harder. Figure 2(b) depicts the structure of the late fusion methods, which involves training each modal data individually in the first stage and fusing the output of several models in the second stage using a decision-making methodology.

Hybrid Fusion Methods. While hybrid fusion methods incorporate the benefits of early and late fusion, they also result in a more complicated model structure and higher training difficulty. Because the structure of deep learning models is flexible and diverse, hybrid fusion methods are ideal for this purpose, and so hybrid fusion methods have been widely applied in domains such as visual question answering and multimedia. Ni et al. [20] suggested a hybrid fusion strategy for multimedia analysis, for example, by presenting an image fusion method based on several BP (Back Propagation) networks. The combined elements of the video and sound signals are then sent into the audiovisual depth neural network model to generate model predictions, and the final results are generated by combining the predictions of each model [21]. The issue of the rationality of the hybrid fusion method's combination approach is a critical aspect in increasing the model's performance. Figure 2(c) shows the hybrid fusion method's structure, which is a blend of early and late fusion.

Each of the three methods has its own set of benefits and drawbacks. Early fusion can better capture the relationship between different features, but it is prone to overfitting; late fusion can solve the overfitting problem, but it does not allow the classifier to train all of the data at once; hybrid fusion methods can combine the benefits of the first two, but they must choose a suitable fusion method based on a combination of practical application problems.

3.2 Model-Based Fusion Methods

Model-based fusion methods are used to handle the problem of fusing disparate modalities by implementing technical and model viewpoints, and they have a broader variety of applications than model-independent methods. Multiple Kernel Learning (MKL) methods, Graphical Model (GM) methods, Neural Network (NN) methods, and so on are some of the most often used methodologies.

Multi-kernel Learning (MKL) Methods. Multi-kernel learning methods are a type of machine learning algorithm that extends the kernel Support Vector

Machine (SVM) method by replacing a single kernel with a collection of basic kernels, where different kernels correspond to different views of the data and are combined into a unified kernel after learning, as shown in Fig. 3. Multi-kernel learning methods are more flexible and capable of fusing heterogeneous data, and they're widely used in applications like multimodal sentiment recognition [22] and multimodal sentiment analysis, where different kernels are used for semantic, video, and text features to achieve better analysis results than single kernel modal fusion. McFee et al. employed MKL to rank the similarity of music artists based on auditory, semantic, and sociological triads to combine diverse data into a reasonable similarity space [23].

Flexible kernel selection, convex loss functions, and the ability to train models using common optimization packages and globally optimal solutions are advantages of the multi-kernel learning method, so a good MKL algorithm can improve accuracy while reducing complexity and training time. Its downside is that while testing, it is dependent on the training function and can consume a lot of memory.

The most basic way to construct a Multi-kernel model is to consider a convex combination of multiple elementary kernel functions:

$$K(x, z) = \sum_{i=1}^{M} \beta_i K_i(x, z) \tag{1}$$

$$\sum_{i=1}^{M} \beta_i = 1 \tag{2}$$

$$\beta_i \geq 0 \tag{3}$$

where K(x,z) is the basic kernel function, M is the total number of basic kernels, and β is the combination factor.

Graphical Model (GM) Methods. The Graphical Model (GM) Method is a popular fusion method that focuses on picture segmentation, stitching, and prediction to fuse shallow or deep layers to achieve the ultimate fusion result [24]. Graphical models are broadly classified as either joint probabilistic generative models or conditional probabilistic discriminative models. Initially, generative models were primarily used, particularly in statistical natural language processing. Examples include the Hidden Markov Model [25], Dynamic Bayesian Networks, and others. These models make extensive use of joint probabilities in their modeling. Subsequent discriminative models, such as Conditional Random Fields (CRF) techniques, are simpler and easier to learn than generative models and have been widely employed and demonstrated good results in multimedia classification tasks, multimodal session segmentation [26], and other applications.

Generative methods seek the best classification surface across distinct categories and adapt to differences in heterogeneous data, whereas discriminative

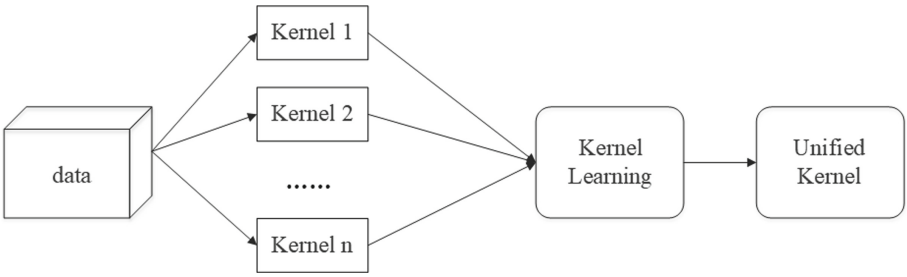

Fig. 3. Procedures multi-kernel learning

models seek to model posterior probabilities and describe data distribution statistically. Generic models require more data than discriminative models, but they are less accurate.

The advantages of Graphical Models include that they can easily uncover and exploit the geographical and temporal structure in the data, that they are suited for modeling time-series data, and that they enhance the model's interpretability by incorporating expert knowledge within the model. The downsides are that the models have limited generalization capabilities and that the feature relationships are complex.

Neural Network (NN) Methods. Neural networks have generated excellent results in unimodal feature extraction, particularly when processing data such as images, sound, and text. In multimodal tasks such as visual question answering [27], image captioning [28], and so on, neural networks are increasingly being used. Each modal data is initially transmitted through several different neural network layers, followed by several hidden layers to map the modalities to the joint space, resulting in the joint features, when utilizing neural networks to build multimodal feature representations.

Vo et al. proposed the Text Image Residual Gating (TIRG) method [29], where the gating module is used to preserve the spatial structure features of image modalities and the residual module obtains the modified features, which are added together to obtain the final combined features for image retrieval.

$$f_{gate}(\phi_x, \phi_t) = \sigma(W_{g2} * RELU(W_{g1} * [\phi_x, \phi_t]) \odot \phi_x) \tag{4}$$

$$f_{res}(\phi_x, \phi_t) = W_{r2} * RELU(W_{r1} * ([\phi_x, \phi_t])) \tag{5}$$

$$\phi_{xt}^{rg} = w_g f_{gate}(\phi_x, \phi_t) + w_r f_{res}(\phi_x, \phi_t) \tag{6}$$

where f_{gate},f_{res} are the gating and the residual features.W_g,W_r are learnable weights to balance them.

In 2019, Xu et al. proposed the Multi-Interactive MemoryNetwork, which uses the Aspect-guided attention mechanism to guide the model to generate

Attention vectors for text and images, while also using the Multi-interactive attention mechanism to capture interaction information between multimodalities and within a single modality [30]. This means that textual and visual information are fused via Attention over several hops. Zhang improved the algorithm for the multimodal machine translation task by deforming the Transformer to input the image representation at the Decoder side as well, implying that it should be coded as Q for the sentence and then K and V for the image features, i.e. finding semantically similar parts in the image for Attention fusion, and finally sending them together to the Decoder side for translation [31]. How to uncover the similarity and distinctiveness between multiple modalities is a critical topic in multimodal fusion. The information from several modalities is easily distinguishable and, at the same time, complementing. It is critical to understand how to locate these two representations, and Lu presented a new cross-modal shared feature transfer algorithm (cm-SSFT) to address this issue [32].

The neural network method has the advantage of being able to learn from large amounts of data and having high scalability. The downside is that it requires a significant amount of data to train, is difficult to achieve convergence, and gets less interpretable as the number of modalities grows.

4 Research Challenges in Multimodal Fusion Technology

By fully exploiting complementary information between modalities, the multimodal fusion technique provides for a more complete and accurate representation of features. When one of the modalities is lacking, the entire system can still function, such as when a person is unable to talk but can still gather and interpret emotions using visual data. Obtaining varying degrees of reinforcement for distinct features while ensuring the model's efficacy can aid in improving the performance of deep learning models. However, several issues, such as the heterogeneity gap and the semantic gap [33], remain unaddressed.

Different neural network structures characterize different data modalities, such as hierarchical networks for image features and sequential networks for text features. The various structures of neural networks result in various abstractions of data representation. As a result, they are not directly comparable, resulting in a heterogeneity gap.

Another issue in multimodal learning is maintaining semantic similarity, as it is difficult to record the intricate relationships between distinct modal inputs. Because neural networks' purpose is to normalize the properties of different modalities into a common space, the semantic similarity between modalities is critical. More effective semantic embedding approaches must be investigated to address these challenges and enable more effective communication and interoperability amongst modal information. There is also a need to develop a more universal evaluation standard to assess the value of feature fusion.

5 Summary and Outlooks

This study outlines multimodal data fusion approaches and the current state of research, as well as summarizes and analyzes the existing difficulties. Model-independent fusion methods and model-based fusion methods are the two forms of multimodal fusion methods. Model-independent methods are further classified as early, late, and hybrid fusion; model-based methods include multiple kernel learning, graphical models, and neural networks. Each of these strategies has benefits and drawbacks, and they all play a role in different domain applications. The current task is to bridge the heterogeneity and semantic gaps.

Multimodal learning is expected to be fully developed in the future because it is more closely aligned with human behavior in perceiving things than unimodal information, and it is more in line with real-world applications. In-depth research on issues such as the semantic gap of modalities and feature fusion evaluation metrics will be conducted in the future to promote the application and development of multimodal fusion technology in the emerging field of machine learning.

References

1. Pedwell, R.K., Hardy, J.A., Rowland, S.L.: Effective visual design and communication practices for research posters: exemplars based on the theory and practice of multimedia learning and rhetoric. Biochem. Mol. Biol. Educ. **45**(3), 249–261 (2017)
2. Welch, K.E., Thompson, G.: Electric rhetoric: classical rhetoric, oralism, and a new literacy. Coll. Compos. Commun. **52**(1), 153 (2000)
3. Bilge, Y.C., Yucel, M.K., Cinbis, R.G., Ikizler-Cinbis, N., Duygulu, P.: Red carpet to fight club: partially-supervised domain transfer for face recognition in violent videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3358–3369 (2021). https://doi.org/10.1109/WACV48630.2021.00340
4. Chen, L., Yan, X.: Counterfactual samples synthesizing for robust visual question answering. IEEE (2020)
5. Alikhani, M., Sharma, P., Li, S.: Cross-modal coherence modeling for caption generation. The Association for Computational Linguistics (2020)
6. Mao, Y., Sun, Q., Liu, G.: DialogueTRM: exploring the intra- and inter-modal emotional behaviors in the conversation (2020)
7. Anwaar, M.U., Labintcev, E., Kleinsteuber, M.: Compositional learning of image-text query for image retrieval. WACV, pp. 1139–1148 (2021). https://doi.org/10.1109/WACV48630.2021.00118
8. Mcgurk, H., Macdonald, J.: Hearing lips and seeing voices. Nature **264**(5588), 746–748 (1976)
9. Petajan, E.D.: Automatic lip-reading to enhance speech recognition (1985)
10. Atrey, P.K., Hossain, M.A., El Saddik, A., et al.: Multimodal fusion for multimedia analysis: a survey. Multimed. Syst. **16**(6), 345–379 (2010). https://doi.org/10.1007/s00530-010-0182-0
11. Wang, D., Cui, P., Ou, M.: Deep multimodal hashing with orthogonal regularization. AAAI Press (2015)

12. Zhang, L., Zhao, Y., Zhu, Z.: Multi-view missing data completion. IEEE Trans. Knowl. Data Eng. **30**(7), 1296–1309 (2018)
13. Wang, L., Sun, W., Zhao, Z.: Modeling intra- and inter-pair correlation via heterogeneous high-order preserving for cross-modal retrieval. Signal Process. **131**, 249–260 (2017)
14. Liu, H., Li, F., Xu, X.: Multi-modal local receptive field extreme learning machine for object recognition. Neurocomputing **277**, 4–11 (2017)
15. Fu, K., Jin, J., Cui, R.: Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2321–2334 (2017)
16. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
17. Martínez, H.P., Yannakakis, G.N.: Deep multimodal fusion. In: The 16th International Conference (2014)
18. Murphy, R.R.: Computer vision and machine learning in science fiction. Sci. Robot. **4**(30), eaax7421 (2019)
19. Kahou, S.E., Pal, C., Bouthillier, X.: Combining modality specific deep neural networks for emotion recognition in video. In: ACM on International Conference on Multimodal Interaction, pp. 543–550 (2013). https://doi.org/10.1145/2522848.2531745
20. Ni, J., Ma, X., Xu, L.: An image recognition method based on multiple BP neural networks fusion. In: International Conference on Information Acquistition, pp. 323–326 (2004)
21. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. J. Mach. Learn. Res. **12**, 2211–2268 (2011)
22. Jaques, N., Taylor, S.: Multi-task, multi-kernel learning for estimating individual wellbeing
23. Mcfee, B., Lanckriet, G.: Learning multi-modal similarity (2010)
24. He, J., Zhang, C.Q.: Survey of research on multimodal fusion technology for deep learning. Comput. Eng. **46**(5), 1–11 (2020)
25. Friedman, N.: Learning the structure of dynamic probabilistic networks. Comput. Sci. 139–147 (2010)
26. Reiter, S., Schuller, B., Rigoll, G.: Hidden conditional random fields for meeting segmentation. In: IEEE International Conference on Multimedia and Expo (ICME 2007), pp. 639–642 (2007)
27. Khademi, M.: Multimodal neural graph memory networks for visual question answering. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7177–7188 (2020)
28. Chen, S., Jin, Q., Wang, P., Wu, Q.: Say as you wish: fine-grained control of image caption generation with abstract scene graphs. In: IEEE, pp. 9962–9971 (2020)
29. Vo, N., Lu, J., Chen, S.: Composing text and image for image retrieval - an empirical odyssey. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6432–6441 (2019). https://doi.org/10.1109/CVPR.2019.00660
30. Xu, N., Mao, W., Chen, G.: Multi-interactive memory network for aspect based multimodal sentiment analysis. In: 33rd AAAI Conference on Artificial Intelligence, pp. 371–378 (2019)
31. Zhang, Z., Chen, K., Wang, R.: Neural machine translation with universal visual representation. In: ICLR 2020: Eighth International Conference on Learning Representations (2020)

32. Lu, Y., Wu, Y., Liu, B.: Cross-modality person re-identification with shared-specific feature transfer. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
33. Wei, C.A.: New ideas and trends in deep multimodal content understanding: a review. Neurocomputing (2020)