·前沿聚焦：优化技术与应用专栏·　　DOI：10. 13471/j. cnki. acta. snus. 2023A022

A survey of multi-modal learning theory*

HUANG Yu, *HUANG Longbo*✉

Institute for Interdisciplinary Information Sciences, *Tsinghua University*, *Beijing 100084*, *China*

Abstract: Deep multi-modal learning, a rapidly growing field with a wide range of practical applications, aims to effectively utilize and integrate information from multiple sources, known as modalities. Despite its impressive empirical performance, the theoretical foundations of deep multi-modal learning have yet to be fully explored. In this paper, we will undertake a comprehensive survey of recent developments in multi-modal learning theories, focusing on the fundamental properties that govern this field. Our goal is to provide a thorough collection of current theoretical tools for analyzing multi-modal learning, to clarify their implications for practitioners, and to suggest future directions for the establishment of a solid theoretical foundation for deep multi-modal learning.

Key words: multi-modal learning, machine learning theory, optimization, generalization

CLC number: TP18　　**Document code:** A　　**Article ID:** 2097－0137（2023）05－0038－12

1 Introduction

Our perception of the world is based on different modalities, e. g. sight, sound, movement, touch, and even smell(Smith et al. , 2005). Inspired from the success of deep learning(Krizhevsky et al. , 2017; He et al. , 2016), deep multi-modal research is also activated, which covers fields like audio-visual learning(Chen et al. , 2020a; Wang et al. , 2020), RGB-D semantic segmentation(Seichter et al. , 2020; Hazirbas et al. , 2017) and Visual Question Answering(Goyal et al. , 2017; Anderson et al. , 2018). Deep multi-modal learning has undoubtedly made a significant impact on a plethora of fields, resulting in remarkable performance gains. However, despite its practical success, the machine learning community still lacks a comprehensive understanding of the underlying theoretical principles. Despite the many advancements that have been made, there remains a significant amount of work to be done in order to gain a deeper understanding of why deep multi-modal learning methods work and when they may fail. This includes further theoretical research into the various components that make up these models, such as the neural networks, the multi-modal data inputs, and the optimization techniques used to train them. Additionally, it is important to explore the different types of data that these models can handle, as well as the different types of tasks they can be applied to. Ultimately, by gaining a deeper understanding of these models and their underlying principles, we can continue to improve upon their performance and make them more widely applicable to a variety of fields.

Typically, when developing a theory for deep multi-modal learning, we must address two fundamental problems that are crucial to understanding machine learning:

* **Received**: 2023－03－28　　**Accepted**: 2023－04－11　　**Published online**: 2023－09－04

Supported by Technology and Innovation Major Project of the Ministry of Science and Technology of China（2020AAA0108400，2020AAA0108403）；Tsinghua Precision Medicine Foundation（10001020109）

✉**Corresponding author**: HUANG Longbo（*longbohuang@tsinghua.edu.cn*）
HUANG Yu（*y-huang20@mails.tsinghua.edu.cn*）

1) Generalization, which ensures that the difference between the training error and the test error is minimal. This means that the model is able to accurately predict new data, even if it has not seen before;

2) Optimization, which refers to the effectiveness of algorithms during the training process in solving various learning tasks. This involves finding the optimal solution for the problem at hand, so that the model can perform well on different types of data.

Consequently, we have conducted a thorough review of the literature on multi-modal learning theory, primarily utilizing the aforementioned taxonomy as a guide. A significant portion of multi-modal learning algorithms that are supported by theoretical guarantees are derived from a generalization perspective. Based on the types of learning tasks, we have classified the existing theoretical work on the generalized properties of multi-modal learning into three primary categories: (i) supervised learning, (ii) semi-supervised learning, and (iii) self-supervised learning.

On the other hand, from an optimization perspective, in addition to traditional convergence analysis based on convex objectives, the theory of deep multi-modal learning must contend with the non-convex landscape of neural networks, which is a challenging task that lacks proper theoretical tools. In light of recent developments in deep learning theory, we have also reviewed the optimization of multi-modal learning theory from the perspectives of convergence guarantees and feature learning. Furthermore, we have also examined theoretical work on multi-modal learning that cannot be explicitly categorized within the aforementioned frameworks. The basic organization of this survey is summarized in Fig. 1.

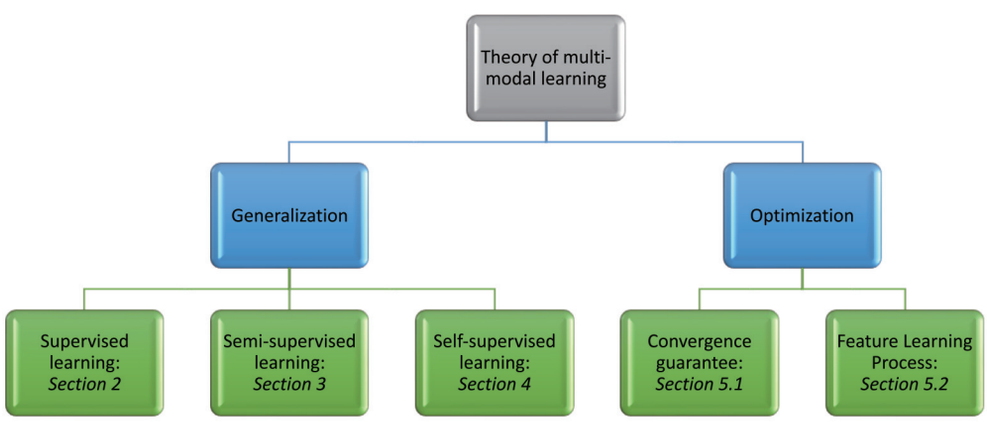

Fig. 1 The basic organization of this survey

The objective of this survey is to re-examine the theoretical advancements in multi-modal learning and to succinctly summarize the significant obstacles encountered in current analysis. Additionally, we aim to present our perspectives on potential future directions within this field. Our expectation is that this survey will furnish researchers with robust tools for the theoretical examination of the performance of multi-modal learning methods, thereby enabling them to gain a deeper understanding of this area and ultimately promote its advancement.

1. 1 Compared with existing surveys

There are several related studies, which reviewing the development of multi-modal learning from different perspectives (Sun, 2013; Xu et al. , 2013; Ramachandram et al. , 2017; Baltrušaitis et al. , 2019; Li et al. , 2019; Guo et al. , 2019; Gao et al. , 2020; Liang et al. , 2022). However, most of these literature surveys just focused on the applications of the representative multi-modal approaches or simply discussed the underlying theoretical foundations. Among them, the close efforts to our survey are Sun(2013) and Liang et al. (2022). Below, we elaborate the above key differences between these two surveys and our work.

Sun(2013) provided a comprehensive review of theories on multi-view learning and classified them into four categories: CCA, effectiveness of co-training, generalization error analysis for co-training, and generalization er-

ror analysis for other multi-view learning approaches. They mainly investigated the multi-view supervised/semi-supervised methods and their analyses are largely relied on conventional statistical learning and convex optimization wisdom. However, the research progress in multi-modal learning has grow rapidly over the last decade, especially in the area of self-supervised learning since Sun(2013) has been published. Moreover, the traditional wisdom cannot directly apply to the deep multi-modal setting. In comparison, our survey considers the recent advances of self-supervised multi-modal methods and also takes the up-to-date theoretical techniques in deep learning into consideration.

Recently, Liang et al. (2022) conducted an in-depth review of multi-modal learning by defining two fundamental principles of modality and presenting a categorization of six crucial technical difficulties. Furthermore, they delved into some theoretical advancements by utilizing this framework as a lens. In contrast to the comprehensive taxonomy provided by Liang et al. (2022), which covers all aspects of multi-modal learning, our survey focused specifically on the theoretical domain of multi-modal learning. The categories we formulated were derived from a technical perspective in theory, offering novel perspectives for multi-modal theoretical research. We hope that our survey will serve as a catalyst for further related research in this field.

2 Generalization: Supervised setting

Supervised multi-modal learning considers integrating information from multiple modalities supervised by a common signal. Conventional statistical theory provides various approaches to tackle the generalization error, which include the PAC-Bayes analysis(McAllester, 1999), different complexity measures-based bounds(e. g. VC dimension (Blumer et al. , 1989) Rademacher complexity(Koltchinskii et al. , 2000)) and uniform stable(Bousquet et al. , 2002). Researchers have utilized such traditional statistical wisdom to explore the theoretical underpinnings of supervised multi-modal learning.

2. 1 PAC-Bayes bounds

PAC-Bayes bounds (McAllester, 1999) are one of prevailing tools for studying the generalization properties of learning algorithms by providing probably approximately correct(PAC) guarantees. The adavantage of PAC-Bayes analysis is that one can obtain tight bounds since it provides data-dependent guarantees. Sun et al. (2017) adopted a novel PAC-Bayes analysis encoding the relationship between the two views through the data distribution dependent priors, and provided various PAC-Bayes bounds for supervised multi-view learning.

2. 2 Rademacher complexity based error bounds

Another approach for analyzing generalization is Rademacher complexity(Koltchinskii et al. , 2000), a distribution-dependent complexity measure for real-valued functions. For multi-modal learning, Farquhar et al. (2005) proposed a novel single optimisation method termed SVM-2K, by combinging the classic kernel Canonical Correlation Analysis(KCCA) and Support Vector Machine(SVM), and provided performance guarantees in term of empirical rademacher complexity. Amini et al. (2009) derived a rademacher complexity based generalization bound for classification with multiple artificially generated views, which identify a a trade-off between the number of views, the size of the training set and the quality of the view generating functions.

The above multi-view analysis typically assumes that each view alone is sufficient to predict the target accurately, which may not hold in current multi-modal setting. For instance, it is difficult to build a classifier just using a weak modality with limited labeled data, e. g. , depth modality in RGB-D images for object detection task (Gupta et al. , 2016). Based on view insufficiency assumption, Xu et al. (2015) presented a learning algorithm Multi-view Intact Space Learning (MISL), to explore the latent intact representation of the data, and they also proposed a novel concept of *multi-view stability*, which is together with the rademacher complexity to bound the generalization error of MISL. Recently, Huang et al. (2021) proved that learning with multiple modalities achieves a smaller population risk than only using its subset of modalities utilizing the rademacher complexity analysis un-

der a most popular multi-modal fusion framework. To the best of our knowledge, this is the first theoretical treatment to explain the superiority of multi-modal from the generalization standpoint. Besides, there is a by-product implication in Zhang et al. (2019) that multi-view representation can recover the same performance as only using the single-view observation by constructing the versatility with strict assumptions on the relationship across different modalities. However, their result is derived by simply analyzing the optimal solution of certain loss function, which neither belongs to typical generalization nor optimization analysis.

2. 3 Uniform stablity

Uniform stablity(Bousquet et al. , 2002) is also a classic notion used to derive high probability generalization error bounds. Zantedeschi et al. (2019) theoretically analyzed a landmark-based SVMs in a multi-view classification setting by applying uniform stability framework. More recently, Sun et al. (2022) introduced a view-consistency regularization and utilize the analyzed technique from uniform stability to deduce a tighter stability-based PAC-Bayes bound for multi-view algorithms.

3 Generalization: Semi-supervised setting

Supervised multi-modal learning approaches, which utilize both multi-modal and label information, have achieved satisfactory performance, however, they are faced with a significant challenge. The collection of large-scale, well-annotated multi-modal training data is both prohibitively expensive and time-consuming. To address the issue of limited labeled data in multi-modal learning, semi-supervised (Guillaumin et al. , 2010; Cheng et al. , 2016) and self-supervised(Lee et al. , 2019; Alayrac et al. , 2020) methods that make less use of label information and rely on exploiting correlation between modalities are proposed. Semi-supervised multi-modal learning considers the setting that each modality of data may only contain a small number of labeled data and a large number of unlabeled data. The majority semi-supervised multi-modal learning methods can be divided into two branches: co-training (Blum et al. , 1998) style algorithms and co-regularize (Brefeld et al. , 2006) style algorithms. Therefore, in this section, we will primarily discuss existing theories and techniques for these two branches of semi-supervised multi-modal learning methods.

3. 1 Co-training style algorithms

Co-training is one of the most representative semi-supervised multi-modal approaches, which utilizes initial small set of labeled two-modal data to learn a weak predictor on each modality and then enables them to label confident instances for each other for further training.

Under the PAC-learning(Valiant, 1984) framework, Blum et al. (1998) showed that for any initial weak predictors, co-training can boost their performances to arbitrarily high accuracy by only using unlabeled data samples. Dasgupta et al. (2001) proved that the generalization error of a classifier from each modality is upper bounded by the disagreement rate of the classifiers from the two modalities. For a special case of co-training that the hypothesis class is the class of linear separators, Balcan et al. (2005) proved that with polynomially many unlabeled examples and a single labeled example, there exists a polynomial-time algorithm to efficiently learn a linear separator under proper assumptions.

However, the above analysis made strong assumption that the true classifiers may not correlate to make predictions, i. e. the two modality is conditional independent given the class label. Abney(2002) showed that weak dependence can also guarantee the success of co-training by generalizing the error bound in Dasgupta et al. (2001) with weaker restrictions that classifiers from different modalities are weakly dependent and nontrivial. Later, Balcan et al. (2004) relaxed the conditional independence to a much weaker expansion assumption, which is proven to be sufficient for iterative co-training to succeed given appropriately strong PAC-learning algorithms on each modality. They also showed that such expansion assumption is necessary to some extent.

It is worthy noting that all above theoretical analyses on co-training rely on a crucial assumption that each of

the modality alone is sufficient to correctly predict the label, which cannot meet in many real applications. Wang et al. (2013) proved that large diversity between two modalities can also lead to good performance of co-training when neither modality is sufficient. Wang et al. (2017) further summarized the key issues for co-training and disagreement-base methods and provided theoretical analyses to tackle such issues, serving as a theoretical foundation of co-training and disagreement-based approaches.

3. 2 Co-regularize style algorithms

The main idea of co-regularize style approaches is to directly minimize the disagreements over different modal predictions. Rosenberg et al. (2007) provided the empirical Rademacher complexity for the Sindhwani et al. (2008) function class of co-regularized least squares and then derived the generalization bound. Later, Sindhwani et al. (2008) constructed a novel Reproducing Kernel Hilbert Spaces(RKHSs), where the reproducing kernel for this RKHS can be utilized to simplify the proof for Rademacher complexity results in . Furthermore, they gave more refined generalization bounds by such techniques based on localized Rademacher complexity. Rosenberg et al. (2009) extended such analysis to the setting that more than two modalities are considered. Since information theory provides the natural language to illustrate the relationship between different modalities, Sridharan et al. (2008) attempted to theoretically understand co-regularization from the information theoretical perspective. They showed that the e excess error between the output hypothesis of co-regularization and the optimal classifier is bounded by the term $\sqrt{\varepsilon_{\text{info}}}$, where $\varepsilon_{\text{info}} < 1$ measures the different information provided by the two modalities.

Canonical Correlation Analysis(CCA)(Hotelling, 1992) based algorithms are also seen as co-regularization method, which aim to compute a shared representation of both sets of variables through maximizing the correlations among the variables among these sets. CCA and its variants, such as Sparse CCA(Witten et al. , 2009), Kernel CCA(Akaho, 2006), Cluster CCA(Chaudhuri et al. , 2009) have been widely used in multi-modal learning. Therefore, existing theoretical analyses for CCA(Bach et al. , 2003; Kuss et al. , 2003) can provide theoretical justifications for the performance of multi-modal CCA-based algorithms. Recently, a survey of multi-modal CCA (Guo et al. , 2019) provided an overview of many representative CCA-based multi-modal learning approaches and also discussed their theoretical foundations.

It is worthy mentioning that all above theoretical analyses on co-regularization are based on the assumption that different modalities can provide almost the same predictions. Unfortunately, such assumption is hardly to meet in practice, since typically there exist divergences among different modalities.

3. 3 Other semi-supervised multi-modal learning algorithms

Szedmak et al. (2007) characterized the generalization performance of semi-supervised version of SVM-2K, which has been discussed in Section 2. 1. Sun et al. (2010) proposed a sparse semi-supervised learning algorithm named sparse multi-view SVMs, where they provided the generalization error analysis. Sun et al. (2011) further considered manifold regularization and introduced multi-view Laplacian SVMs. They also derived the generalization error bound and empirical Rademacher complexity for the proposed method. Recently, Sun et al. (2020) proposed a novel information-theoretic method, called Total Correlation Gain Maximization (TCGM) by maximizing the total correlation Total Correlation Gain(TCG) over classifiers of all modalities. They showed that the the optimal classifiers for such a TCG are equivalent to the Bayesian posterior classifiers given each modality under some permutation function, which theoretically guarantee the success of TCGM.

4 Generalization: Self-supervised setting

Self-supervised learning adopts self-defined signals to learn representations from unlabeled data. A large amount of prevailing self-supervised approaches naturally leverage the property of multi-view data, where the input (e. g. the original image) and the self-supervised signal (e. g. image with data augmentation) can be treated as two views of the same data. In practice, incorporating the view from different modalities as supervision into

such methods leads to remarkable successes(Zhang et al. , 2020; Desai et al. , 2021; Radford et al. , 2021) in multi-modal self-supervised learning. Therefore, seminal works that established theoretical foundations of self-supervised learning in general multi-view setting are conducive to our understanding of self-supervised multi-modal learning. The methods and techniques in this literature may be adapted or extended to more realistic multi-modal setting.

Arora et al. (2019) provided a theoretical guarantee of contrastive learning which aims to minimize $L(\phi) = \mathbb{E}\left[\ell\left(\phi(X)^{\mathrm{T}}\left(\phi(X_+) - \phi(X_-)\right)\right)\right]$, where X_+ and X_- are corresponding positive and negative views for data X. Under the class conditional independence(CI) assumption, i. e. $X_+ \perp X_-$, they show that contrastive learning yields good representations ϕ for downstream tasks. Lee et al. (2021) considered reconstruction-based self-supervised methods, where two views (X_1, X_2) are available for each data point, and the learning objective is to minimize the reconstruction error of X_2 based on a function of X_1: $L(\phi) = \mathbb{E}\left\| X_2 - \phi(X_1)\right\|^2$. They established a good performance guarantee for the learned ϕ on downstream tasks under a relaxed approximately conditional independence (ACI) assumption for each view on the label.

In work concurrent with Lee et al. (2021), Tosh et al. (2021b) shows guarantees for contrastive object with a multi-view redundancy assumption, which is analogous to ACI. Tosh et al. (2021a) also studied the problem of contrastive learning similar to Tosh et al. (2021b). Specifically, they considered the topic modeling setting and showed that when the two views of the data point correspond to random partitions of a document, contrastive learning recovers information related to the underlying topics that generated the document. Arora et al. (2019) and Lee et al. (2021) both assume strong independence between the views conditioning on the downstream tasks. Tsai et al. (2020) and Tosh et al. (2021b) considered a weak version of such assumption where they only assume independence between the downstream task and one view conditioning on the other view, while Tsai et al. (2020) mainly focuses on the mutual-information perspective. Meanwhile, motivated by the information bottleneck principle(Tishby et al. , 2000), Federici et al. (2020) proposed an unsupervised multi-modal method, where they provided a rigorous theoretical analysis of its application, but they also relied on the strong assumption that each modality provides the same task-relevant information.

5 Optimization

5. 1 Convergence guarantee

Xu et al. (2015) theoretically showed that the proposed Iteratively Reweight Residuals(IRR) technique for their multi-view intact space algorithm will lead to a local-minimized solution under mild assumptions. Arora et al. (2016) developed stochastic approximation approaches for Partial Least Squares(PLS) problem with two views in un-supervised setting and they provided iteration complexity bounds for the proposed methods. Mou et al. (2017) formulated a novel multi-view based PLS framework and proposed algorithms to optimize the objective function. They provided a rough convergence analysis of proposed method. Seminal work(Fukumizu et al. , 2007; Hardoon et al. , 2009; Cai et al. , 2011) provided the convergence analyses for kernel CCA.

5. 2 Understanding the feature learning process

In recent years, great efforts have been made to explore the learning dynamic of neural networks. Among this, a line of work studied how multi-view or multi-modal features can be learned by different algorithms and architectures in deep learning. Allen-Zhu et al. (2020) developed a theory to explain how ensemble and knowledge distillation work when the data has "multi-view" structures. Wen et al. (2021) studied how contrastive learning with two view generated by data augmentations learns the feature representations, and Wen et al. (2022) showed the mechanism of how non-contrastive self-supervised learning such as SimSiam(Chen et al. , 2021) can still learn competitive representations. Specific to the multi-modal setting, Huang et al. (2022) aim to demystify the

performance drop in naive multi-modal joint training that the best uni-modal network outperforms the jointly trained multi-modal network by understanding the feature learning process. Remarkably Huang et al. (2022) took the training process of neural networks into consideration, which is the first theoretical treatment towards the degenerating aspect of multi-modal learning in neural networks. Du et al. (2021) also studied such negative aspect of multi-modal learning, where they proved that with more modalities, some hard-to-learn features cannot be learned. While they attempted to analyze the feature learning process, they did not fully address the optimization challenges encountered during this process.

6 Challenges in theoretical research for multi-modal learning

As per our comprehensive review above, it is evident that within the current framework of multi-modal learning theory, there exist two crucial assumptions that play a vital role in shaping the overall understanding: (i) each modality of itself, is capable of providing sufficient information for specific tasks, i. e. the target functions from all modalities, as well as the combined modalities, maintain label consistency on every example. (ii) Each modality is conditionally independent given the class label.

However, in practice, different modalities are of various importance under specific circumstance (Ngiam et al. , 2011; Liu et al. , 2018; Gat et al. , 2020). It is common that information from one single modality may be incomplete to build a good classifier(Yang et al. , 2015; Gupta et al. , 2016; Liu et al. , 2018). For instance, it is difficult to build a classifier just using a weak modality with limited labeled data, e. g. , depth modality in RGB-D images for object detection task. Moreover, in real world application, different modalities (e. g. text, image, audio) are often closely related and contain information that can be used to infer the other modalities (Reed et al. , 2016; Wu et al. , 2018). Specifically, in image-text tasks, the text description can provide information about the objects and actions present in an image, while the image can provide information about the context and background of the scene.

Consequently, to continue to progress the theoretical foundations of deep multi-modal learning, it is necessary to re-examine the current assumptions underlying the modeling of relationships between modalities and tasks, as well as the capture of dependencies among different modalities. This would involve relaxing the assumptions of sufficiency and conditional independence, in order to more accurately reflect the intricacies of real-world multi-modal data.

7 Future directions

We are now discussing several important future directions which could lead to new and exciting developments in the field of multi-modal learning theory.

7. 1 Optimization perspective

Despite some theoretical exploration of multi-modal learning from various perspectives, there has been relatively limited theoretical analysis of optimization-related aspects specifically. Theoretically understanding the optimization dynamic of multi-modal learning in modern deep learning can be quiet challenging due to the non-convex optimization landscapes and complexity of multi-modal data. Moreover, it is important to understand the role of the training dynamics in terms of generalization in deep multi-modal learning, which is beyond the scope of convex optimization and conventional statistical learning. Recent developments in analysis techniques for deep learning theory, such as those proposed by Allen-Zhu et al. (2020); Li et al. (2021); Wen et al. (2021), have aimed to address optimization dynamics and non-convex landscapes. Building on this research, it would be worthwhile to further investigate multi-modal learning from an optimization perspective.

7. 2 Self-supervised learning

Among the recent advances of deep multi-modal learning, self-supervised approaches such as contrastive

learning have gained significant attention and success(Radford et al. , 2021; Jia et al. , 2021). These methods have shown to be effective in learning representations for multiple modalities in a unified manner by leveraging the inherent relationships of them, which may be difficult or impossible to annotate. However, the underlying mechanisms behind how they learn to explore the relationships between different modalities remains unclear from a theoretical perspective. Future theoretical studies for multi-modal learning on this topic are in high demand, which have the potential to provide practical insights to improve performance in various multi-modal tasks.

7. 3 Foundation models

Foundation models(Bommasani et al. , 2021), pre-trained on a large and diverse pool of data and able to transfer the learned knowledge to a wide range of tasks, have revolutionized the way models are built in various areas(Devlin et al. , 2018; Brown et al. , 2020; Chen et al. , 2020b), including multi-modal learning(Radford et al. , 2021; Ramesh et al. , 2022). Theoretical understanding and mathematical tools for foundation models are generally lacking and some recent attempts towards this direction mostly focus on extracting features from one module(Arora et al. , 2019; HaoChen et al. , 2021; Wei et al. , 2021; Kumar et al. , 2022). Can existing analysis be extended beyond the single module? How do foundation models learn useful representations from different modules? As the multi-modal learning is becoming an increasingly popular paradigm in modern machine learning, we believe the community will significantly benefit from more theoretical insights towards addressing these questions.

8 Conclusions

In this survey, we have endeavored to comprehensively review the theoretical analysis of multi-modal learning from the generalization and optimization perspectives. We organized the literature into technical categories to provide a systematic reference of mathematical tools for future theory research. Additionally, we discussed the challenges faced in existing theoretical work and highlighted several promising areas for future research. The goal of this survey is to provide a comprehensive understanding of the theoretical foundation of multi-modal learning and to inspire further exploration in this field. With the increasing popularity of multi-modal learning, it is vital for the community to continue to deepen their understanding of the theoretical foundations in order to unlock the full potential of this powerful approach. In conclusion, we believe that the multi-modal learning has a lot of potential to improve the performance of various machine learning tasks, and that future theoretical studies will be essential to fully realize this potential.

References:

ABNEY S，2002. Bootstrapping［C］//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics：360−367.

AKAHO S，2006. A kernel method for canonical correlation analysis［EB/OL］. arXiv：cs/0609071. https：//arxiv.org/abs/cs/0609071.

ALAYRAC J B，RECASENS A，SCHNEIDER R，et al，2020. Self-supervised multimodal versatile networks［C］//Proceedings of the 34th International Conference on Neural Information Processing Systems：25−37.

ALLEN-ZHU Z，LI Y Z，2020. Towards understanding ensemble，knowledge distillation and self-distillation in deep learning［EB/OL］. arXiv ：2012.09816. https：//arxiv.org/abs/2012.09816.

AMINI M R，USUNIER N，GOUTTE C，2009. Learning from multiple partially observed views：an application to multilingual text categorization［C］//Proceedings of the 23rd International Conference on Neural Information Processing Systems：28−36.

ANDERSON P，WU Q，TENEY D，et al，2018. Vision-and-language navigation：Interpreting visually-grounded navigation instructions in real environments［C］//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition：3674−3683.

ARORA R，MIANJY P，MARINOV T V，2016. Stochastic optimization for multiview representation learning using partial least squares［C］//Proceedings of the 33rd International Conference on International Conference on Machine Learning ：2859−2867.

ARORA S，KHANDEPARKAR H，KHODAK M，et al，2019. A theoretical analysis of contrastive unsupervised representation learning[EB/OL]. arXiv：1902.09229. https://arxiv.org/abs/1902.09229.

BACH F R，JORDAN M I，2003. Kernel independent component analysis[C]//2003 IEEE International Conference on Acoustics，Speech，and Signal Processing：IV-876.

BALCAN M F，BLUM A，2005. A PAC-style model for learning from labeled and unlabeled data[M]// AUER P，ed. Learning Theory. Berlin：Springer：111-126.

BALCAN M F，BLUM A，YANG K，2004. Co-training and expansion：Towards bridging theory and practice[C]// Proceedings of the 17th International Conference on Neural Information Processing Systems：89-96.

BALTRUŠAITIS T，AHUJA C，MORENCY L P，2019. Multimodal machine learning：A survey and taxonomy[J]. IEEE Trans Pattern Anal Mach Intell，41(2)：423-443.

BLUM A，MITCHELL T，1998. Combining labeled and unlabeled data with co-training[C]//Proceedings of the eleventh annual conference on Computational learning theory：92-100.

BLUMER A，EHRENFEUCHT A，HAUSSLER D，et al，1989. Learnability and the vapnik-chervonenkis dimension[J]. J ACM，36(4)：929-965.

BOMMASANI R，HUDSON D A，ADELI E，et al，2021. On the opportunities and risks of foundation models[EB/OL]. arXiv：2108.07258. https://arxiv.org/abs/2108.07258.

BOUSQUET O，ELISSEEFF A，2002. Stability and generalization[J]. J Mach Learn Res，2：499-526.

BREFELD U，GÄRTNER T，SCHEFFER T，et al，2006. Efficient co-regularised least squares regression[C]//Proceedings of the 23rd international conference on Machine learning：137-144.

BROWN T B，MANN B，RYDER N，et al，2020. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems：1877-1901.

CAI J，SUN H，2011. Convergence rate of kernel canonical correlation analysis[J]. Sci China Math，54(10)：2161-2170.

CHAUDHURI K，KAKADE S M，LIVESCU K，et al，2009. Multi-view clustering via canonical correlation analysis[C]//Proceedings of the 26th Annual International Conference on Machine Learning：129-136.

CHEN H，XIE W，VEDALDI A，et al，2020a. Vggsound：A large-scale audio-visual dataset[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics，Speech and Signal Processing（ICASSP）：721-725.

CHEN T，KORNBLITH S，NOROUZI M，et al，2020b. A simple framework for contrastive learning of visual representations[C]// Proceedings of the 37th International Conference on Machine Learning：1597-1607.

CHEN X，HE K，2021. Exploring simple Siamese representation learning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）：15745-15753.

CHENG Y H，ZHAO X，CAI R，et al，2016. Semi-supervised multimodal deep learning for RGB-D object recognition[C]//：Proceedings of the 25th International Joint Conference on Artificial Intelligence：3345-3351.

DASGUPTA S，LITTMAN M L，McALLESTER D，2001. PAC generalization bounds for co-training[C]//Proceedings of the 14th International Conference on Neural Information Processing Systems：Natural and Synthetic：375-382.

DESAI K，JOHNSON J，2021. VirTex：learning visual representations from textual annotations[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）：11157-11168.

DEVLIN J，CHANG M W，LEE K，et al，2018. BERT：Pre-training of deep bidirectional transformers for language understanding [EB/OL]. arXiv：1810.04805. https://arxiv.org/abs/1810.04805.

DU C，TENG J，LI T，et al，2021. Modality laziness：Everybody's business is nobody's business[EB/OL]. https://openreview.net/pdf? id=1eGFH6yYAJn.

FARQUHAR J D R，HARDOON D R，MENG H Y，et al，2005. Two view learning：SVM-2K，theory and practice[C]//Proceedings of the 18th International Conference on Neural Information Processing Systems：355-362.

FEDERICI M，DUTTA A，FORRÉ P，et al，2020. Learning robust representations via multi-view information bottleneck[EB/OL]. arXiv：2002.07017. https://arxiv.org/abs/2002.07017.

FUKUMIZU K，BACH F R，GRETTON A，2007. Statistical consistency of kernel canonical correlation analysis[J]. J Mach Learn Res，8：361-383.

GAO J，LI P，CHEN Z，et al，2020. A survey on deep learning for multimodal data fusion[J]. Neural Comput，32(5)：829-864.

GAT I，SCHWARTZ I，SCHWING A，et al，2020. Removing bias in multi-modal classifiers：Regularization by maximizing functional entropies[EB/OL]. arXiv：2010.10802. https://arxiv.org/abs/2010.10802.

GOYAL Y，KHOT T，SUMMERS-STAY D，et al，2017. Making the V in VQA matter：Elevating the role of image understanding in visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition：6904-6913.

GUILLAUMIN M，VERBEEK J，SCHMID C，2010. Multimodal semi-supervised learning for image classification[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition：902-909.

GUO C F，WU D R，2019. Canonical correlation analysis（CCA）based multi-view learning：An overview[EB/OL]. arXiv：1907.01693. https://arxiv.org/abs/1907.01693.

GUO W，WANG J，WANG S，2019. Deep multimodal representation learning：A survey[J]. IEEE Access，7：63373-63394.

GUPTA S，HOFFMAN J，MALIK J，2016. Cross modal distillation for supervision transfer[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition（CVPR）：2827-2836.

HAOCHEN J Z，WEI C，GAIDON A，et al，2021. Provable guarantees for self-supervised deep learning with spectral contrastive loss[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems：5000-5011.

HARDOON D R，SHAWE-TAYLOR J，2009. Convergence analysis of kernel canonical correlation analysis：Theory and practice[J].Mach Learn，74(1)：23-38.

HAZIRBAS C，MA L，DOMOKOS C，et al，2017. FuseNet：Incorporating depth into semantic segmentation via fusion-based CNN architecture[C]//Asian Conference on Computer Vision：213-228.

HE K，ZHANG X，REN S，et al，2016. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition（CVPR）：770-778.

HOTELLING H，1992. Relations between two sets of variates[M]//KOTZ S，ed. Breakthroughs in Statistics. New York：Springer：162-190.

HUANG Y，DU C Z，XUE Z H，et al，2021. What makes multi-modal learning better than single（provably）[EB/OL]. arXiv：2106.04538. https://arxiv.org/abs/2106.04538.

HUANG Y，LIN J Y，ZHOU C，et al，2022. Modality competition：What makes joint training of multi-modal network fail in deep learning？（provably）[EB/OL]. arXiv：2203.12221. https://arxiv.org/abs/2203.12221.

JIA C，YANG Y F，XIA Y，et al，2021. Scaling up visual and vision-language representation learning with noisy text supervision[C]// Proceedings of the 38th International Conference on Machine Learning：4904-4916.

KOLTCHINSKII V，PANCHENKO D，2000. Rademacher processes and bounding the risk of function learning[M]// GINÉ E，ed. High Dimensional Probability II. Boston：Birkhäuser：443-457.

KRIZHEVSKY A，SUTSKEVER I，HINTON G E，2017. ImageNet classification with deep convolutional neural networks[J]. Commun ACM，60(6)：84-90.

KUMAR A，RAGHUNATHAN A，JONES R，et al，2022. Fine-tuning can distort pretrained features and underperform out-of-distribution[EB/OL]. arXiv：2202.10054. https://arxiv.org/abs/2202.10054.

KUSS M，GRAEPEL T，2003. The geometry of kernel canonical correlation analysis[R]. Tübingen：Max Planck Institute for Biological Cybernetics.

LEE J D，LEI Q，SAUNSHI N，et al，2021. Predicting what you already know helps：Provable self-supervised learning[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems：309-323.

LEE M A，ZHU Y，SRINIVASAN K，et al，2019. Making sense of vision and touch：Self-supervised learning of multimodal representations for contact-rich tasks[C]//2019 International Conference on Robotics and Automation（ICRA）：8943-8950.

LI Y，YANG M，ZHANG Z，2019. A survey of multi-view representation learning[J]. IEEE Trans Knowl Data Eng，31(10)：1863-1883.

LI Z Y，WANG T H，ARORA S，2021. What happens after SGD reaches zero loss？：A mathematical framework[EB/OL]. arXiv：2110.06914. https://arxiv.org/abs/2110.06914.

LIANG P P，ZADEH A，MORENCY L P，2022. Foundations and trends in multimodal machine learning：Principles，challenges，and open questions[EB/OL]. arXiv：2209.03430. https://arxiv.org/abs/2209.03430.

LIU K，LI Y，XU N，et al，2018. Learn to combine modalities in multimodal deep learning[EB/OL]. arXiv：1805.11730. https://

arxiv.org/abs/1805.11730.

McALLESTER D A，1999. Some PAC-Bayesian theorems[J]. Mach Learn，37：355−363.

MOU Y，ZHOU L，YOU X，et al，2017. Multiview partial least squares[J]. Chemom Intell Lab Syst，160：13−21.

NGIAM J，KHOSLA A，KIM M，et al，2011. Multimodal deep learning[C]//Proceedings of the 28th International Conference on International Conference on Machine Learning：689−696.

RADFORD A，KIM J W，HALLACY C，et al，2021. Learning transferable visual models from natural language supervision[EB/OL]. arXiv：2103.00020. https://arxiv.org/abs/2103.00020.

RAMACHANDRAM D，TAYLOR G W，2017. Deep multimodal learning：A survey on recent advances and trends[J]. IEEE Signal Process Mag，34(6)：96−108.

RAMESH A，DHARIWAL P，NICHOL A，et al，2022. Hierarchical text-conditional image generation with CLIP latents[EB/OL]. arXiv：2204.06125. https://arxiv.org/abs/2204.06125.

REED S，AKATA Z，YAN X C，et al，2016. Generative adversarial text to image synthesis[C]// Proceedings of the 33rd International Conference on International Conference on Machine Learning：1060−1069.

ROSENBERG D S，BARTLETT P L，2007. The rademacher complexity of co−regularized kernel classes[C]// Proceedings of the 11th International Conference on Artificial Intelligence and Statistics：396−403.

ROSENBERG D S，SINDHWANI V，BARTLETT P L，et al，2009. Multiview point cloud kernels for semisupervised learning [Lecture Notes][J]. IEEE Signal Process Mag，26(5)：145−150.

SEICHTER D，KÖHLER M，LEWANDOWSKI B，et al，2020. Efficient RGB-D semantic segmentation for indoor scene analysis [EB/OL]. arXiv：2011.06961. https://arxiv.org/abs/2011.06961.

SINDHWANI V，ROSENBERG D S，2008. An RKHS for multi-view learning and manifold co-regularization[C]//Proceedings of the 25th international conference on Machine learning：976−983.

SMITH L，GASSER M，2005. The development of embodied cognition：Six lessons from babies[J]. Artif Life，11(1/2)：13−29.

SRIDHARAN K，KAKADE S M，2008. An information theoretic framework for multi-view learning[C]// Conference on Learning Theory ：403−414.

SUN S，2013. A survey of multi-view machine learning[J].Neural Comput Appl，23(7/8)：2031−2038.

SUN S，JIN F，2011. Robust co-training[J]. Int J Patt Recogn Artif Intell，25(7)：1113−1126.

SUN S，SHAWE-TAYLOR J，2010. Sparse semi-supervised learning using conjugate functions[J]. J Mach Learn Res，11：2423−2455.

SUN S，SHAWE-TAYLOR J，MAO L，2017. PAC-Bayes analysis of multi-view learning[J]. Inf Fusion，35：117−131.

SUN S，YU M，SHAWE-TAYLOR J，et al，2022. Stability-based PAC-Bayes analysis for multi−view learning algorithms[J]. Inf Fusion，86/87：76−92.

SUN X，XU Y，CAO P，et al，2020. TCGM：an information-theoretic framework for semi-supervised multi-modality learning [M]//VEDALDI A，ed. Computer Vision：ECCV 2020，Cham：Springer ：171−188.

SZEDMAK S，SHAWE-TAYLOR J，2007. Synthesis of maximum margin and multiview learning using unlabeled data[J]. Neurocomputing，70(7/8/9)：1254−1264.

TISHBY N，PEREIRA F，BIALEK W，2000. The information bottleneck method[EB/OL]. arXiv：physics/0004057. https://arxiv.org/abs/physics/0004057.

TOSH C，KRISHNAMURTHY A，HSU D J，2021a. Contrastive estimation reveals topic posterior information to linear models [J]. J Mach Learn Res，22(1)：12883−12913.

TOSH C，KRISHNAMURTHY A，HSU D，2021b. Contrastive learning，multi-view redundancy，and linear models[C]//Proceedings of the 32nd International Conference on Algorithmic Learning Theory：1179−1206.

TSAI Y H H，WU Y，SALAKHUTDINOV R，et al，2020. Self-supervised learning from a multi-view perspective[EB/OL]. arXiv：2006.05576. https://arxiv.org/abs/2006.05576.

VALIANT L G，1984. A theory of the learnable[J]. Commun ACM，27(11)：1134−1142.

WANG W，TRAN D，FEISZLI M，2020. What makes training multi-modal classification networks hard?[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）：12692−12702.

WANG W，ZHOU Z H，2013. Co-training with insufficient views[C]//Proceedings of the 5th Asian Conference on Machine Learn-

ing：467−482.

WANG W，ZHOU Z H，2017. Theoretical foundation of co-training and disagreement-based algorithms［EB/OL］. arXiv：1708.04403. https://arxiv.org/abs/1708.04403.

WEI C，XIE S M，MA T，2021. Why do pretrained language models help in downstream tasks？ an analysis of head and prompt tuning［C］//Proceedings of the 35th International Conference on Neural Information Processing Systems：16158−16170.

WEN Z X，LI Y Z，2021. Toward understanding the feature learning process of self-supervised contrastive learning［EB/OL］. arXiv：2105.15134. https://arxiv.org/abs/2105.15134.

WEN Z X，LI Y Z，2022. The mechanism of prediction head in non−contrastive self-supervised learning［EB/OL］. arXiv：2205.06226. https://arxiv.org/abs/2205.06226.

WITTEN D M，TIBSHIRANI R J，2009. Extensions of sparse canonical correlation analysis with applications to genomic data［J］. Stat Appl Genet Mol Biol，8（1）：1−27.

WU M，GOODMAN N，2018. Multimodal generative models for scalable weakly-supervised learning［EB/OL］. arXiv：1802.05335. https://arxiv.org/abs/1802.05335.

XU C，TAO D C，XU C，2013. A survey on multi-view learning［EB/OL］. arXiv：1304.5634. https://arxiv.org/abs/1304.5634.

XU C，TAO D C，XU C，2015. Multi-view intact space learning［J］. IEEE Trans Pattern Anal Mach Intell，37（12）：2531−2544.

YANG Y，YE H J，ZHAN D C，et al，2015. Auxiliary information regularized machine for multiple modality feature learning［C］// Proceedings of the 24th International Joint Conference on Artificial Intelligence：1033−1039.

ZANTEDESCHI V，EMONET R，SEBBAN M，2019. Fast and provably effective multi-view classification with landmark-based SVM［C］//Joint European Conference on Machine Learning and Knowledge Discovery in Databases：193−208.

ZHANG C Q，HAN Z B，CUI Y J，et al，2019. CPM-nets：Cross partial multi-view networks［C］// Proceedings of the 33rd International Conference on Neural Information Processing Systems：559−569.

ZHANG Y H，JIANG H，MIURA Y，et al，2020. Contrastive learning of medical visual representations from paired images and text［EB/OL］. arXiv：2010.00747. https://arxiv.org/abs/2010.00747.

（责任编辑　冯兆永）