IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 16, 2019, accepted May 4, 2019, date of publication May 15, 2019, date of current version May 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2916887

Deep Multimodal Representation Learning: A Survey

WENZHONG GUO[1,2], (Member,IEEE), JIANWEN WANG[1,2,3,4], AND SHIPING WANG[1,2], (Member, IEEE)

[1]College of Mathematics and Computer Sciences, Fuzhou University, Fuzhou 350116, China
[2]Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University, Fuzhou 350116, China
[3]College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350117, China
[4]Fujian Provincial Engineering Technology Research Center for Public Service Big Data Mining and Application, Fujian Normal University, Fuzhou 350117, China

Corresponding author: Shiping Wang (shipingwangphd@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61502104 and Grant 61672159, in part by the Fujian Collaborative Innovation Center for Big Data Application in Governments, and in part by the Technology Innovation Platform Project of Fujian Province under Grant 2014H2005.

ABSTRACT Multimodal representation learning, which aims to narrow the heterogeneity gap among different modalities, plays an indispensable role in the utilization of ubiquitous multimodal data. Due to the powerful representation ability with multiple levels of abstraction, deep learning-based multimodal representation learning has attracted much attention in recent years. In this paper, we provided a comprehensive survey on deep multimodal representation learning which has never been concentrated entirely. To facilitate the discussion on how the heterogeneity gap is narrowed, according to the underlying structures in which different modalities are integrated, we category deep multimodal representation learning methods into three frameworks: joint representation, coordinated representation, and encoder-decoder. Additionally, we review some typical models in this area ranging from conventional models to newly developed technologies. This paper highlights on the key issues of newly developed technologies, such as encoder-decoder model, generative adversarial networks, and attention mechanism in a multimodal representation learning perspective, which, to the best of our knowledge, have never been reviewed previously, even though they have become the major focuses of much contemporary research. For each framework or model, we discuss its basic structure, learning objective, application scenes, key issues, advantages, and disadvantages, such that both novel and experienced researchers can benefit from this survey. Finally, we suggest some important directions for future work.

INDEX TERMS Multimodal representation learning, multimodal deep learning, deep multimodal fusion, multimodal translation, multimodal adversarial learning.

I. INTRODUCTION

To convey the comprehensive information about objects in the world, various cognitive signals describing different aspects of the same object are recorded in different kinds of media such as text, image, video, sound, and graph. In the representation learning area, the word ‘‘modality’’ refers to a particular way or mechanism of encoding information. Hence, different types of media listed above also refer to modalities, and the representation learning tasks involving several modalities will be characterized as multimodal.

Since multimodal data depict an object from different viewpoints, usually complementary or supplementary in contents, they are more informative than unimodal data. For example, early research on speech recognition showed that the visual modality provides information on lip motions and articulations of the mouth including open and close, thus can help to improve the speech recognition performance. Therefore, it is valuable to exploit the comprehensive semantics provided by several modalities.

However, although it is easy for human beings to perceive the world through comprehensive information from multiple sensory organs [3], how to endow machines with analogous cognitive capabilities is still an open question. One of the challenges we are confronted with is the heterogeneity gap

The associate editor coordinating the review of this manuscript and approving it for publication was Canbing Li.

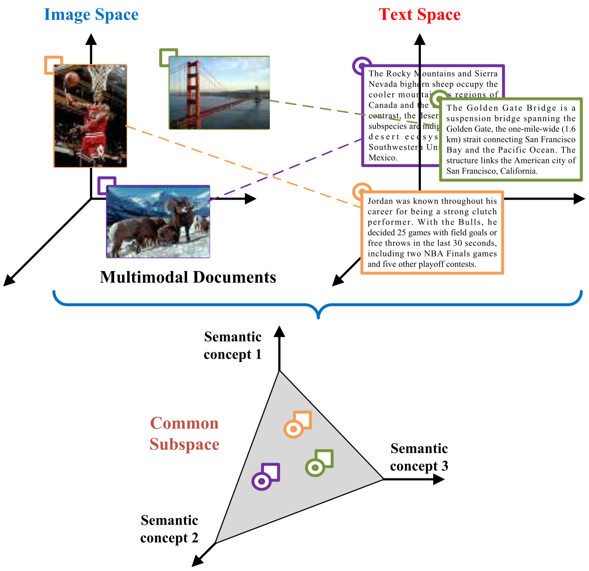
Image Space
Text Space
The Rocky Mountains and Sierra Nevada bighorn sheep occupy the cooler mountai s regions of Canada and the contrast, the deser subspecies are indig desert ecosys Southwestern Uni Mexico.
The Golden Gate Bridge is a suspension bridge spanning the Golden Gate, the one-mile-wide (1.6 km) strait connecting San Francisco Bay and the Pacific Ocean. The structure links the American city of San Francisco, California.
Jordan was known throughout his career for being a strong clutch performer. With the Bulls, he decided 25 games with field goals or free throws in the last 30 seconds, including two NBA Finals games and five other playoff contests.
Multimodal Documents
Semantic concept 1
Common Subspace
Semantic concept 3
Semantic concept 2

FIGURE 1. Schematic of the common subspace learning (adapted from [5]), which aims to project the heterogeneous data of different modalities into a common subspace, where the multimodal data with similar semantics will be represented by similar vectors.

in multimodal data. As Fig. 1 shows, since the feature vectors from different modalities originally located in unequal subspaces, the vector representations associated with similar semantics would be completely different. Here, this phenomenon is referred to as heterogeneity gap, which would hinder the multimodal data from being comprehensively utilized by the subsequent machine learning modules [4]. A popular method for addressing this problem is projecting the heterogeneous features into a common subspace, where the multimodal data with similar semantics will be represented by similar vectors [5]. Thus, the primary objective of multimodal representation learning is narrowing the distribution gap in a joint semantic subspace while keeping modality specific semantics intact.

To narrow the heterogeneity gap, numerous researches with various approaches have been conducted in the past decades. As a result, the advancement of multimodal representation learning has benefited plenty of applications. For example, by the utilization of fused features from multimodalities, improved performance can be achieved in cross-media analysis tasks, such as video classification [6], event detection [7], [8], and sentiment analysis [9], [10]. Further, via the exploitation of cross-modal similarity or cross-modal correlation, it becomes possible for us to retrieve images using a sentence as input or vice versa, which is a task known as cross-modal retrieval [11]. Most recently, a novel type of multimodal application, cross-modal translation [12], has drawn great attention in the computer vision community. As the name suggests, it strives to translate one modality into another. Exemplary applications within this category include image caption [13], video description [14], and text-to-image synthesis [15].

In recent years, due to the powerful representation ability with multiple levels of abstraction, deep learning has demonstrated outstanding results in various applications involving computer vision, natural language processing, and speech recognition [16]. Additionally, another key advantage of deep learning is that a hierarchical representation can be learned directly using a general-purpose learning procedure, without requiring a design or selection process of handcrafted features. Motivated by this success, deep multimodal representation learning, which is a natural extension of its unimodal version, has recently attracted tremendous research attention.

The goal of this article is to provide a comprehensive survey on deep multimodal representation learning and suggest the future direction in this active field. Generally, the machine learning tasks based on multimodal data include three necessary steps: modality-specific features extracting, multimodal representation learning which aims to integrate diverse features from different modalities in a common subspace, and a reasoning step such as classification or clustering. This paper mainly focuses on the second step, multimodal representation learning in deep learning scenarios, and will also make a brief reference to the other two steps but not go into the details.

The focus of this paper is the key issues on how to narrow the heterogeneity gap while keeping modality specific semantics intact in different multimodal application scenes. To facilitate the discussion, according to the underlying structures in which different modalities are integrated, shown as Fig. 2, we category these methods into three types of frameworks: joint representation, coordinated representation, and encoder-decoder. Each framework has its distinct architecture and approach of integrating multimodal features. Additionally, we review some typical models including probabilistic graphical models (PGM), multimodal autoencoders, deep canonical correlation analysis (DCCA), generative adversarial networks (GAN), and attention mechanism, which have either proven to be effective or shown promising results.

The connection between the typical models and the three frameworks can be seen in Table 1. Each of the typical models described here can be categorized into one or more of the frameworks or can be integrated with them. For each type of framework or model, we will discuss its basic structure, learning objective, application scenes, key issues, advantages, and disadvantages, such that both novel and experienced researchers will benefit from this survey. The key issues relevant to different frameworks and models will be marked in bold and summarized in Section IV (Table 3).

Most recently, several surveys [17]–[20] related to the topic of multimodal learning have been published. Comparing to previous reviews, the focus of our paper is distinctive in that we seek to survey the literature from a cross-perspective of multimodal representation learning and deep learning, which has never been concentrated fully. For example, the review proposed by Zhao *et al.* [17] mainly focuses on conventional methods. The work proposed by Baltrušaitis *et al.* [18] focuses on the challenges of multimodal machine learning, as one of the five challenges they

TABLE 1. The relationship between typical models and three types of deep multimodal representation learning frameworks. Each of the typical models may belong to (denoted by ✓) or can be integrated with (denoted by △) the relevant framework.

	Joint representation	Coordinated representation	Encoder-decoder
PGM	✓	✓	
Multimodal autoencoders	✓	✓	
DCCA		✓	
GAN	△	△	△
Attention mechanism	△	△	△

defined, representation learning is only a small part of their concern. From the perspective of multimodal representation learning, the closest work to ours is that of Li *et al.* [19] which concentrates on multi-view representation learning, including shallow and deep methods, while, by contrast, ours highlight the latter which have gained more attention in recent years. From the perspective of multimodal deep learning, the closest effort to ours is [20] which mainly reviews the models and applications relying on multimodal features fusion (categorized as joint representation in ours). Comparing to [20], in this paper, more types of integration frameworks and models will be discussed.

In contrast to previous surveys, another difference of ours is that we highlight on the key issues of newly developed technologies such as encoder-decoder model, generative adversarial networks (GAN), and attention mechanism in a multimodal representation learning perspective, which, to our best knowledge, have never been reviewed previously, even though they have become the major focuses of much contemporary research. For example, previously, the encoder-decoder models were mainly introduced as one of the implementation ways used for cross-modal translation task, while, for the first time, they are discussed further from the representation learning perspective in this paper.

The rest of this paper is organized as follows: In Section II, we discuss the key issues on how to narrow the heterogeneity gap in three types of frameworks. In Section III, we review the typical models listed in Table 1. In Section IV, we finish with a conclusion and suggest some future directions in this active field.

II. DEEP MULTIMODAL REPRESENTATION LEARNING FRAMEWORKS

To facilitate the discussion on how to narrow the heterogeneity gap and inspired by the definitions in [18], according to the underlying structures illustrated in Fig. 2, we category deep multimodal representation methods into three types of frameworks: (i) joint representation, which aims to project unimodal representations together into a shared semantic subspace such that the multimodal features can be fused; (ii) coordinated representation including cross-modal similarity models and canonical correlation analysis, which seeks to learn separated but constrained representations for each modality in a coordinated subspace; (iii) encoder-decoder models, which endeavors to learn an intermediate representation used for mapping one modality into another. Each framework has its way of integrating several modalities and is shared by some applications. To obtain a general impression of their possible applications, in table 2, a summary of typical applications and the relevant modalities involved in these frameworks has been given.

TABLE 2. A summary of typical applications of three frameworks. Each application may include some of the modalities such as audio, video, image, and text which are denoted by their first letter. Here, different integration ways are denoted by + (fusion), ∼ (coordination) and → (translation).

Frameworks	Applications	Modalities	References
Joint representation	Video classification	A + V	[21]
		A + V + T	[6]
	Event detection	A + V	[8]
		A + V + T	[7]
	Sentiment analysis	A + V + T	[9] [10] [22]
	Visual question answering	I + T	[23] [24]
	Emotion recognition	A + V	[25]
		A + V + T	[26]
	Speech recognition	A + V	[1] [27]
Coordinated representation	Cross-modal retrieval	I ∼ T	[11] [28] [29] [30] [31]
	Image caption	I ∼ T	[32]
	Cross-modal embeddings	V ∼ T	[33]
		I ∼ T	[34] [35] [36]
	Transfer learning	I ∼ T	[37] [38]
Encoder-decoder	Image caption	I → T	[13] [39] [40]
	Video description	V → T	[41] [42] [43]
	Text to image synthesis	T → I	[15] [44]

As Fig. 2 shows, before multimodal representation learning can be applied, modality-specific features should be extracted via appropriate methods. Thus, in this section, we will first introduce unimodal representation methods which may significantly impact the performance, and then start our discussion on three types of frameworks.

A. MODALITY-SPECIFIC REPRESENTATIONS

Although a variety of different multimodal representation learning models may share similar architectures, the essential components used for extracting modality-specific features could be quite different from each other. Here, we will introduce some of the most popular components appropriate for different modalities, without going into technical details.

The most popular models used for image feature learning are convolutional neural networks (CNN) such as LeNet [45], AlexNet [46], GoogleNet [47], VGGNet [48],

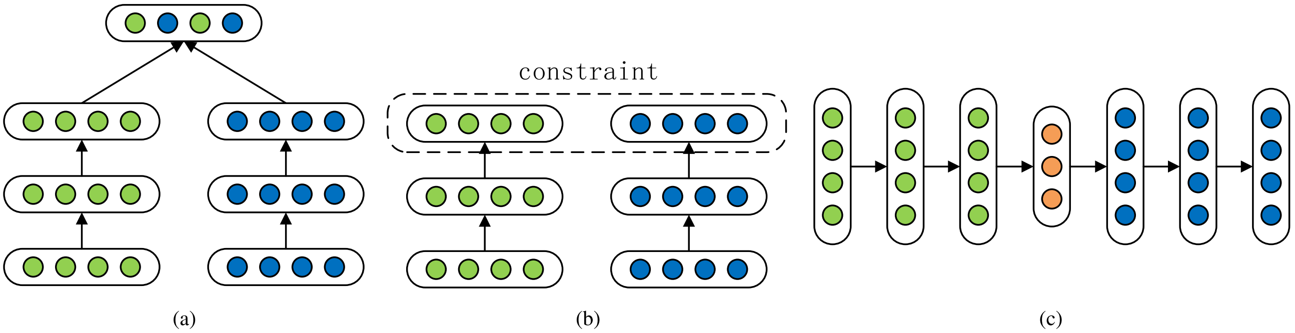

FIGURE 2. **Three types of frameworks about deep multimodal representation. (a) Joint representation aims to learn a shared semantic subspace. (b) Coordinated representation framework learns separated but coordinated representations for each modality under some constraints. (c) Encoder-decoder framework translates one modality into another and keep their semantics consistent.**

and ResNet [49]. They can be integrated into multimodal learning models and trained together with other components. However, considering the requirement for sufficient training data and computation resources, the pre-trained version of CNN may be a better choice for multimodal representation learning.

The fundamental works for neural language processing involve representing words and encoding sentences. A popular way to represent words is word embedding such as word2vec [50] or Glove [51] which maps words into a distributional vector space, where the similarity between words can be measured. In NLP tasks, a common issue that should be considered is the unknown word problem, also known as out-of-vocabulary (OOV) words, that can potentially affect the performance of many systems. To deal with unknown word issue, character embeddings [52], [53] is a viable option for representing language inputs. For example, Kim *et al.* [52] trained a convolution neural network to yield word representations based on character-level embeddings. Bojanowski *et al.* [53] proposed to learn the vector representations of character n-grams, then, by treating each word as a bag of character n-grams, the embedding of a word can be obtained by the sum of these vector representations. Experiments [54], [55] showed that handling OOV issue properly would improve the performance of NLP systems considerably.

Recurrent neural networks (RNN) [56] is a powerful tool for dealing with varying length sequences such as sentences, videos, and audios. Since the activation of the current hidden state at time t depends on that of all the previous time steps, it can be seen as a summarization of the sequence up to step t. However, vanilla RNNs is difficult to capture long-term dependencies because of the gradient vanishing problem [57]. In practice, a better choice is long short-term memory (LSTM) [58], [59] networks or gated recurrent unit (GRU) [60] networks, which has a better performance in capturing long-term dependencies [61], [62]. Further, bidirectional recurrent neural networks (BRNN) [63] and the bidirectional edition of LSTM [64] or GRU [65] are also widely used for capturing the semantics. In addition to RNN, CNN is another widely used model for extracting salient n-gram features from sentences. Experiments showed that CNN based models perform remarkably well in sentence-level classification [66] and sentiment analysis tasks [67].

As to video modality, since the input of each time step is an image, its feature can be extracted via the techniques used for handling images. In addition to deep features, handcrafted features are still widely used in video and audio modalities [10], [68]. Further, some toolkits have been developed to extract handcrafted features. For example, OpenFace [69] can be used to extract facial features such as facial landmark, head pose, and eye gaze. Another tool is Opensmile [70] which can be used to extract acoustic features including Mel-frequency cepstral coefficients (MFCC), voice intensity, pitch, and their statistics. After the frames of videos and audios have been encoded, CNN or RNN networks aforementioned can be used to summarize the sequences into individual vector representations.

B. JOINT REPRESENTATION

The strategy of integrating different types of features to improve the performance of machine learning methods has long been used by researches. A natural extension of this strategy in a multimodal setting is the utilization of fused heterogeneous features. Following this strategy, promising results have been shown in many multimodal classification or clustering tasks, such as video classification [6], [21], event detection [7], [8], sentiment analysis [9], [10], and visual question answering [23].

To bridge the heterogeneity gap of different modalities, joint representation aims to project unimodal representations into a shared semantic subspace, where the multimodal features can be fused [18]. As Fig. 2(a) showed, after each modality is encoded via an individual neural network, both of them will be mapped into a shared subspace, where the conceptions shared by modalities will be extracted and fused into a single vector.

The simplest way for fusing multimodal features is to concatenate them directly. However, mostly this subspace is implemented by a distinct hidden layer, in which the transformed modality specific vectors will be added, and thus the semantics from different modalities will be combined. This property can be seen from (1), where z is the activation of output nodes in the shared layer, v is the output of modality-specific encoding network, w is the weights connecting between modality specific encoding layer to the shared layer and the subscript index denotes different modalities.

$$z = f(w_1^T v_1 + w_2^T v_2) \tag{1}$$

Other than the fusion process in a distinct hidden layer, usually called as an additive approach, a multiplicative method is also adopted in some literature. In a sentiment analysis task, Zadeh *et al.* [10] proposed to fuse language, video, and audio modalities in a tensor, which is constructed from the out product of all the modality-specific feature vectors. By this way, the author intends to exploit either intra-modality or inter-modality dynamics. The definition of the fused tensor can be formulated as follows:

$$\mathbf{z}^m = \begin{bmatrix} \mathbf{z}^l \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}^v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}^a \\ 1 \end{bmatrix} \tag{2}$$

where z^m denotes the fused tensor, z^l, z^v, z^a denotes different modalities respectively, and $\otimes$ indicates the outer product operator. However, since the outer product is cost expensive, in a more efficient way, Fukui *et al.* [23] alternatively propose to utilize Multimodal Compact Bilinear pooling (MCB) to fuse language and image modalities. Formulated as (3), given vectors x and q, the proposed method seeks to reduce the dimension of the outer product $x \otimes q$ by Count Sketch projection function Ψ. Particularly, the count sketch of the outer product can be decomposed into a convolution of separated count sketches [71], which means that the computation of an outer product can be avoided. Further, the authors use Fast Fourier Transform (FFT) to accelerate the computation.

$$\begin{aligned} \Phi &= \Psi(x \otimes q) \\ &= \Psi(x) * \Psi(q) \\ &= \mathrm{FFT}^{-1}\left(\mathrm{FFT}\left(\Psi(x)\right) \odot \mathrm{FFT}\left(\Psi(q)\right)\right) \end{aligned} \tag{3}$$

Although the model shown in Fig. 2(a) is designed for the setting in which parallel data are available during training and inference steps, the ability to deal with partial data missing problem in some modalities is also desired, such that more training data can be exploited or the performance of downstream tasks is influenced only slightly in the case of data missing from one or more modalities. To this end, a widely used method is training the model via the data including only some modalities, excluding a modality in different training epochs [1], [72].

Interestingly, the training trick used for tackling data missing is also helpful for **obtaining modality-invariant property**, which means that the difference of the statistical distribution between modalities is minimized, or, in other words, the feature vectors contains minimum modality-specific characteristics. The work proposed by Aytar *et al.* [73] shows that constrained by a statistical regularization which encourages activations in the intermediate hidden layers to have similar statistics distribution across modalities, the modality-invariant property can be strengthened. Their model encourages different modalities to be aligned with each other automatically in the representation layer, even when the training data is unaligned.

To be more expressive, the learned vector is expected to **fuse complementary semantics** form different modalities. The property, complementary, cannot be guaranteed automatically since joint representation tends to preserve shared semantics across modalities while ignoring modality-specific information. A solution is adding extra regularization terms to the optimization objectives [74]. For example, the reconstruction loss used in multimodal autoencoders [1] can be considered as a regularization term playing as a role to preserve modality independence. Another example is the approach proposed by Jiang *et al.* [21], which impose a trace norm regularization over the network weights to reveal the hidden correlations and diversity of the multimodal features. Intuitively, if a pair of features are highly correlated, the weights used for fusing them should be similar such that their contributions to the fused representation will be roughly equal. Thus, the goal of trace norm regularization is to discover the relationship between modalities and adjust the weights of the fusion layer accordingly. Their experiments in video classification tasks showed that this regularization term is helpful for improving performance.

Comparing to other frameworks, one of the advantages of joint representation is that it is convenient to fuse several modalities since there is no need to coordinate modalities explicitly. Another advantage is that the shared common subspace tends to be modality-invariant, which is helpful for transferring knowledge from one modality to another [1], [73]. While one of the disadvantages of this framework is that it cannot be used to infer the separated representations for each modality.

C. COORDINATED REPRESENTATION

Another type of methods popular in multimodal learning is coordinated representation. As Fig. 2(b) showed, instead of learning representations in a joint subspace, coordinated representation framework learns separated but coordinated representations for each modality under some constraints [18]. Since the information contained in different modalities is unequal, learning separated representations is beneficial for persevering the exclusive and useful modality-specific characteristics [31]. Typically, condition on the constraint types, coordinated representation methods can be categorized into two groups, cross-modal similarity based and cross-modal correlation based. Cross-modal similarity based methods aim to learn a common subspace where the distance of vectors from different modalities can be measured directly [75], while cross-modal correlation based methods aim to learn a shared subspace such that the correlation of the representation sets from different modalities is maximized [5]. In this section, we will review the former and leave the latter in Section III-C.

Cross-modal similarity methods learn coordinated representations under constraints of similarity measurement. The learning objective of this model is to **preserve inter-modality and intra-modality similarity structure**, which expects the cross-modal similarity distance associated with the same semantics or object to be as minimum as possible, while expects the distance with dissimilar semantics to be as maximum as possible.

A widely used constraint is *cross-modal ranking*. Take visual-text embedding for example, ignoring the regularization terms and denoting the matched embedding vectors of visual and text as $(v, t) \in D$, the optimization objective can be expressed as a loss function in (4), where α is the margin, S is the similarity measurement function, t^- is the embedding vectors unmatched to v and v^- is the embedding vectors unmatched to t. Commonly, t^- and v^- are known as negative samples which are selected randomly from the dataset D, and (4) is known as margin rank loss [36].

$$rankLoss = \sum_{v}\sum_{t^-}\max(0, \alpha - S(v,t) + S(v,t^-)) + \sum_{t}\sum_{v^-}\max(0, \alpha - S(t,v) + S(t,v^-)) \quad (4)$$

Based on the cross-modal ranking constraint, a variety of cross-modal applications have been developed. For example, Frome *et al.* [34] used a combination of dot-product similarity and margin rank loss to learn a visual-semantic embedding model (DeViSE) for visual recognition. DeViSE firstly pre-trains a pair of deep networks to map images and their correlated labels into embedding vectors v and t then leverages the cross-modal similarity model to learn a shared semantic embedding space for both modalities. Following the notations in (4), the loss function for each training sample can be defined as follows:

$$loss(v, t) = \sum_{t^-}\max(0, \alpha - tMv + t^-Mv) \quad (5)$$

where M is a linear transformation matrix used for transforming v into the shared semantic embedding space, and the dot-product between t and Mv is the similarity measurement used for both training and testing. Under the constraint in (5), the model is expected to produce a higher dot-product similarity between matched vectors than between unmatched ones and subsequently endows images embedding with rich semantic information which is transferred from language modality. This idea is also shared by the work proposed by Lazaridou and Baroni [35], which aims to integrate and propagate visual information into word embeddings. Their experimental results implied that the transferred visual knowledge is helpful for representing abstract concepts.

Inspired by the success of DeViSE, Kiros *et al.* [36] extended this model to learn a joint image-sentence embedding used for image captioning. They pre-trained a CNN network to obtain image features v and trained an LSTM network to encode its relevant sentences into t, then mapped both encodings into a coordinated embedding space where the similarity between them can be exploited by a cross-modal similarity model similar to [34]. Their model adopted the same similarity measurement used in DeViSE but employed a bi-directional rank loss formulated in (4) such that much richer cross-modal relationships can be discovered. This model is also employed in the work proposed by Socher *et al.* [32], which aims to map sentences and images into a common space for cross-modal retrieval. They introduced dependency trees based recursive neural network (DTRNN) to encode language modality and argued that the proposed DTRNN is robust to surface changes such as word order.

Further, Karpathy and Fei-Fei [76] extended this framework to learn a fine-grained cross-model alignment between words and image regions for generating region-level descriptions of images. Unfortunately, this task suffers from lacking of necessary supervision information. Given images and their correlated sentences, the one-to-one correspondence between a word and the region it described is not yet known. To address this problem, they selected to infer the alignment between segments of sentences and the regions of the image in a cross-modal embedding space. The key idea is to formulate the image-sentence score as a function of the individual region-word similarity. Let v_i denotes the image regions and s_t denotes the words in a sentence, the score between image k and sentence l is defined as follows:

$$S_{kl} = \sum_{t \in g_l} max_{i \in g_k} v_i^T s_t \quad (6)$$

where, g_k is the set of fragments in image k, g_l is the set of snippets in sentence l and each word s_t aligns to a unique best image region. Additionally, assuming that $k = l$ denotes a matched image-sentence pair, the cross-modal ranking constraint can be defined as a loss function in (7), which encourages aligned image-sentences pairs to have a higher score than misaligned pairs.

$$rankLoss = \sum_{k}\sum_{l} max\,(0, 1 - S_{kk} + S_{kl}) + \sum_{k}\sum_{l} max\,(0, 1 - S_{kk} + S_{lk}) \quad (7)$$

The strategy to measure image-sentence similarity based on individual region-word scores is also adopted by Peng *et al.* [31], who aim to preserve the modality-specific characteristics by utilizing the fine-grained information within each modality during the cross-modal correlation learning. The authors argued that different modalities have imbalanced and complementary relationships, thus, instead of measuring the similarity in a common space, they construct an independent semantic space for each modality and measure the cross-modal similarity in both spaces simultaneously. After that, the modality-specific similarity scores will be combined into a final measurement used for cross-modal retrieval.

In addition to cross-modal ranking, another widely used constraint is *Euclid distance*. The mainstream approach in this category is to minimize the distance of paired samples [33], [77], [78]. An example is a model proposed by Pan *et al.* [33], which aims to learn a visual-semantic embedding used for generating video descriptions. The model projects both visual and language representations into a low-dimensional embedding space, where the distances between paired samples are minimized such that the semantics of visual embeddings will be consistent with their relevant sentences. This constraint can be expressed as a loss term:

$$distanceLoss = \sum_{(v,s)\in D} \|T_v v - T_s s\|_2^2 \tag{8}$$

where T_v and T_s are transform matrices for video v and sentence s, v and s are paired samples form dataset D.

Another example is the model for cross-modal matching proposed by Liong *et al.* [78], which aims to reduce the modality gap of paired data by minimizing the difference of hidden representations over all layers. Suppose that visual modality v and text modality t are encoded via homogeneous feed-forward neural networks, the loss can be formulated as follows:

$$distanceLoss = \sum\nolimits_{l=1}^{L-1} \sum\nolimits_{i=1}^{N} \left\| h_{it}^{l} - h_{iv}^{l} \right\|_2^2 \tag{9}$$

where l indicates a layer of both modality-specific networks, i indicates a pair of instances of training data and h denotes the hidden representations. Further, the authors also imposed a large margin criterion to the distance of unpaired data which aims to minimize the intra-class distance and maximize the inter-class distance, such that more discriminative information can be exploited. This criterion is defined as follows:

$$\begin{cases} \left\| t_i - v_j \right\|_2^2 \le \theta_1, & \text{if } l_{t_i,v_j} = 1 \\ \left\| t_i - v_j \right\|_2^2 \ge \theta_2, & \text{if } l_{t_i,v_j} = -1 \end{cases} \tag{10}$$

where t_i denotes the sentence i, v_j denotes image j, and θ_1, θ_2 are the small and large thresholds respectively. The condition $l_{t_i,v_j} = 1$ means that t_i and v_j belong to the same class, otherwise, belong to the different class.

Except for learning inter-modality similarity measurement, another key issue of cross-modal applications is to preserve the intra-modality similarity structure. A widely used strategy is classifying the category of learned features such that they are also discriminative within each modality [30], [79]. Additionally, another method is to keep the neighborhood structure within each view. The constraint in (10) is one of the implementations in this group. Another example is the work from Wang *et al.* [80], which proposed to learn image-text embeddings via coordinated representation model which combines cross-view ranking constraints with within-view neighborhood structure preservation constraints in the loss function. Let $N(v_i)$ denotes the neighborhood of image v_i and $N(t_i)$ denotes the neighborhood of sentence t_i, the within-view neighborhood structure preservation constraints can be formulated as follows:

$$\begin{cases} d\left(v_i, v_j\right) + m < d\left(v_i, v_k\right) & \forall v_j \in N\left(v_i\right), \ \ \forall v_k \notin N\left(v_i\right) \\ d\left(t_i, t_j\right) + m < d\left(t_i, t_k\right) & \forall t_j \in N\left(t_i\right), \ \ \forall t_k \notin N\left(t_i\right) \end{cases} \tag{11}$$

In addition to the applications characterized as finding one modality from another such as cross-modal retrieval [75], [77], [80] and retrieval-based visual description [32], another type of application of coordinated representation is transfer knowledge across modalities, which may enhance the semantic description capability of the embeddings in target modality. The basic idea is minimized the cross-modal similarity of paired multimodal data in a common subspace during training, such that the embeddings can capture their shared semantics, which means that the knowledge has been transferred. Several pieces of literature mentioned above [33]–[36] can be considered as representative examples of this idea. Furthermore, coordinated representation can also be used for cross-domain transfer learning which would partially reduce the need for labeled data. For example, in order to transfer knowledge from a large-scale cross-media dataset to small-scale one, the works from Huang *et al.* [37], [38] proposed to train a pair of networks, each for one of the domains, and coordinate them via minimizing the maximum mean discrepancy (MMD) [81].

Comparing to other frameworks, coordinated representation tends to persevere the exclusive and useful modality-specific characteristics within each modality [31]. Since different modalities are encoded in separated networks, one of the advantages is that each modality can be inferred individually. This property is also beneficial for cross-modal transfer learning which aims to transfer knowledge across different modalities or domains. A disadvantage of this framework is that, mostly, it is hard to learn representations with more than two modalities.

D. ENCODER-DECODER

Recently, Encoder-decoder framework has been widely used for multimodal translation tasks which map one modality into another, such as image caption [13], [39], video description [14], [41], and image synthesis [15], [82]. Typically, as shown in Fig. 2(c), the encoder-decoder framework is mainly composed of two components, an encoder, and a decoder. The encoder maps source modality into a latent vector v, and then, based on the vector v, the decoder will generate a novel sample of target modality.

Although most of the encoder-decoder models contain only an encoder and a decoder, some of the variants can also be composed of several encoders or decoders. For example, Mor *et al.* [83] proposed a model to translate music across musical instruments, where a single encoder and several decoders are involved. The shared encoder is responsible for extracting domain-independent music semantics, and each decoder will reproduce a piece of music in the target domain. An example including two encoders is the image-to-image translation model proposed by Huang *et al.* [84]. It consists of a content encoder and a style encoder, each is responsible for part of the duty.

The generalized learning objective of encoder-decoder models, take visual description as an example [41], can be expressed as follows:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{(V,S)} \log p(S|V;\theta) \tag{12}$$

which maximizes the log likelihood of the sentence S given the corresponding visual content V and the model parameters θ. Further, assuming that each word in the sequence is produced in order, the log probability of the sentence can be expressed as:

$$\log p(S|V;\theta) = \sum_{t=0}^{N} \log p\left(S_{w_t}|V, S_{w_1}, \ldots, S_{w_{t-1}}\right) \tag{13}$$

where S_{w_i} represents the ith word in the sentence and N is the total number of words.

Superficially, the latent vector learned by the encoder-decoder model seems to relate only to the source mode, but in fact, it closely relates to both source and target modalities. Since the flowing direction of the error correction signal is from the decoder to the encoder, the encoder is guided by the decoder during training. Subsequently, the generated representation tends to capture the shared semantics from both modalities.

To **capture shared semantics** more effectively, a popular solution is keeping the semantic consistency among modalities via some regularization terms. It depends on the coordination between the encoder and the decoder. Both the correct understanding of semantics in source modality and the pertinent generating of novel samples in target modality are important for success. Take image caption [85] as an example, the description generated by the decoder may cover multiple visual aspects of an image including objects, attributes such as color and size, backgrounds, scenes and spatial relationships, hence, the encoder has to detect and encode necessary information correctly, and further, the decoder will be responsible for reasoning high-level semantics and generating grammatically well-formed sentences.

An example of explicitly considering the semantic consistency between modalities is the model proposed by Gao *et al.* [42], which aims to translate videos into sentences. To tackle this problem, on the one hand, they maximized the likelihood formulated in (13) such that sentences can be generated correctly, on the other hand, they minimized the representation difference in a common subspace such that their semantics are correlated with each other. Suppose that v denotes the visual features, s denotes the sentence embeddings, and R denotes a matrix used for linearly projecting s into the subspace where v located, the consistency constraint can be written as loss term in (14). Another example is the work proposed by Reed *et al.* [15], which endeavors to translate characters into pixels via Generative adversarial network (GAN) [82]. In their model, within each class, the similarity between the source and target encodings is maximized such that the semantics in both modalities will keep consistent. Since the models of image synthesis are mostly implemented by GAN, more example of this task will be left to Section III-D which concentrates on generative adversarial learning.

$$loss = \|v - Rv\|_F^2 \tag{14}$$

On condition that the semantic consistency between modalities has been modeled explicitly, this framework can be used to learn cross-modal semantic embedding. For example, based on the encoder-decoder framework, Gu *et al.* [86] proposed to learn cross-modal embeddings used for retrieval. Their model translates each of the modality into another via distinct encoder-decoder networks and expects that the generated images or sentences are similar to their sources. In this model, the similarity between the generated sentence and its corresponding reference sentences is measured by a standard evaluation metric like BLEU [87], and the similarity between images is measured by a discriminator which is responsible for distinguishing whether an image comes from the generator or not.

In early works [88], [89], the representation of visual modality is usually a fixed visual semantic list such as objects and their relationship which is detected explicitly by the encoder. Then based on n-gram language models or sentence templates, a sentence is generated by the decoder. In this way, the problem is simplified. However, it is difficult for these models to deal with large vocabulary or to model complex sentence structure [41].

Recently, a more accessible way of representing source modality is encoding essential information into a single vectorial representation [14]. Comparing to traditional methods, it is more convenient for neural networks to encode information and generate samples. However, using the single vector as a bridge, it is challenging for both encoder and decoder to translate semantics between modalities. A problem for the encoder is that the high-level vectorial representation distilled from the source may lose some information which is useful for generating target modality [13]. Also, another problem will arise in decoder once RNN models are used for generating a long sequence. The information contained in the original representation vector will be diminished during its delivery through time steps.

Attention mechanism has become a popular solution for both aforementioned problems. Rather than merely using a single vector resulting from the last step of the encoder, attention mechanism allows utilizing the intermediate representations which distribute among time steps in an RNN network [90] or localized regions in a CNN network [91]. For the encoder, this mechanism relieves the requirement that the full information should be integrated into a single vector, and thus gives more flexibility to the design of encoder.

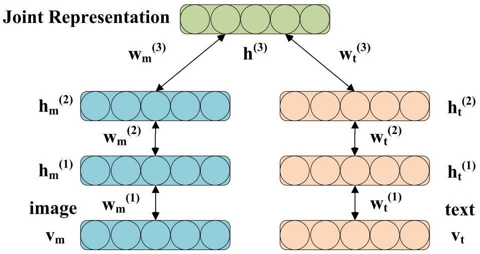

FIGURE 3. The model of deep multimodal RBM (adapted from [96]), which models the joint distribution over image and text inputs.

On the other hand, for the decoder, this mechanism provides an ability to concentrate on the part of the scene selectively and dynamically during the prediction process. Due to its ability to select the prominent features, attention mechanism has been successfully used in a variety of neural networks and has demonstrated its unique power in improving performance in many applications [90]–[92]. Considering its significance on multimodal representation learning, we will take a more detailed look at its impact in Section III-E.

To address the encoding and decoding problems of multimodal sequence, deep reinforcement learning (DRL) is another promising solution, in which either encoding or decoding of a sequence can be treated as sequential decision-making problem. For example, via deep reinforcement learning, Chen *et al.* [93] proposed to train a feature selection module used for determining whether an input at time step t should be included or not during encoding, such that salient features can be involved while noise will be excluded. Conversely, an exemplary application of deep reinforcement learning during decoding is image captioning [94], [95].

Comparing to other frameworks, one of the advantages of the encoder-decoder framework is its being able to generate novel samples of target modality condition on the representations of source modality. On the contrary, the disadvantage of this framework is that each encoder-decoder can only encode one of the modalities. Further, the complexity in designing the generator should be taken into consideration, since the technique for generating plausible target is still on its development.

III. TYPICAL MODELS

In this section, some typical models in deep multimodal representation learning will be summarized. They range from conventional models, including probabilistic graphical models, multimodal autoencoders, and deep canonical correlation analysis, to newly developed technologies, including generative adversarial networks and attention mechanism. The typical models described here can be categorized into one or more of the frameworks above introduced or can be integrated with them.

A. PROBABILISTIC GRAPHICAL MODELS

In the deep representation learning area, probabilistic graphical models include deep belief networks (DBN) [97] and deep Boltzmann machines (DBM) [98]. Although both of them are trained from stacked restricted Boltzmann machines (RBM) [99] layer wisely, their structures are different. The former is a partially directed model which consists of a directed belief network and an RBM layer, while the latter is a fully undirected model.

An example of probabilistic graphical models is multimodal DBN proposed by Srivastava and Salakhutdinov [72]. By adding a shared RBM hidden layer on top of modality-specific DBN, it learns a joint representation across modalities. Another model also from Srivastava and Salakhutdinov [96] is multimodal deep Boltzmann machines which alternatively using DBMs as the basic units for processing data from each modality. As a fully undirected model, the states of hidden units will influence each other across modalities. Hence, the modality fusion process is distributed across all hidden units of all layers.

The learning objective of multimodal probabilistic graphical models is to **maximize the joint distribution** over modalities. Take multimodal DBM as an example which is illustrated in Fig. 3, suppose that each modality is encoded via a DBM with two hidden layers, the joint distribution can be written as:

$$\begin{aligned} P(v_m, v_t, \theta) = \sum_{h_m^{(2)}, h_t^{(2)}, h^{(3)}} & P(h_m^{(2)}, h_t^{(2)}, h^{(3)}) \\ & (\sum_{h_m^{(1)}} P(v_m, h_m^{(1)}, h_m^{(2)})) \\ & (\sum_{h_t^{(1)}} P(v_t, h_t^{(1)}, h_t^{(2)})) \end{aligned} \tag{15}$$

where v_m, v_t denote image and text input respectively, θ denotes the parameters, $h_m = \{h_m^{(1)}, h_m^{(2)}\}$, $h_t = \{h_t^{(1)}, h_t^{(2)}\}$ denotes the hidden layers in each modality and $h^{(3)}$ denotes the shared representation layer.

Unlike the strategy which connects different modalities via a shared representation layer, Feng *et al.* [28] tended to maximize the correspondence between modalities layer wisely. At each equivalent hidden layer, two RBMs from different modalities are connected respectively by a correlation loss function. In this way, the essential cross-model correlation for cross-modal retrieval is captured.

By fusing modalities together in a unified latent space, probabilistic graphical models can be used to learn the essential cross-modal correlations. Based on multimodal deep belief networks, several applications such as audio-visual emotion recognition [25], audio-visual speech recognition [27], and information trustworthiness estimation [100] have been reported. Also, based on multimodal deep Boltzmann machines, several solutions used for human pose estimation [101] and video emotion prediction [26] have been proposed.

One of the advantages of probabilistic graphical models is that they can be trained in an unsupervised fashion, allowing the use of unlabeled data. Another advantage comes from

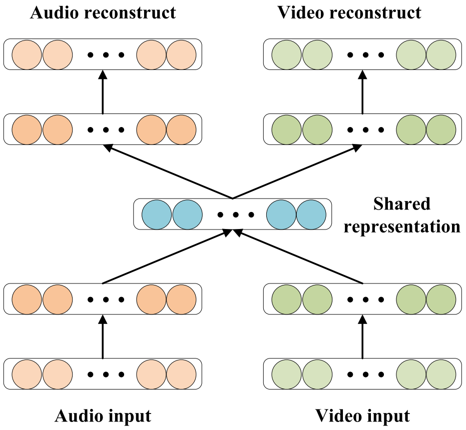

FIGURE 4. The model of bimodal autoencoder (adapted from [1]), which aims to learn a shared representation across audio and video modalities.

their generative nature, which makes it possible to generate the missing modality condition on the other ones [96]. However, due to the expensive approximate inference algorithm, a crucial disadvantage of multimodal deep Boltzmann machines is its considerably high computational cost [102].

B. MULTIMODAL AUTOENCODERS

Autoencoders is popular for its ability to learn representations in an unsupervised manner, no labels are needed [103]. The basic structure of autoencoders includes two components, one is an encoder and the other is a decoder. The encoder converts the input into a compressed hidden vector, also known as latent representation, while the decoder endeavors to reconstruct the input based on this latent representation such that the reconstruction loss is minimized.

Inspired by denoising autoencoders [104], Ngiam *et al.* [1] extended the autoencoders to a multimodal setting. They trained a bimodal deep autoencoder to learn a shared representation across audio and video modalities. Showed as Fig. 4, in this model, two separated autoencoders are combined in the common latent representation layer while keeping their encoders and decoders independent. To capture the cross-modal correlations robustly, depended on the shared representation, each modality can be reconstructed, even when one of the modalities is absent. Let (x_i, y_i) denotes a pair of inputs and $(\hat{x}_i, \hat{y}_i)$ denotes their reconstructed outputs, the basic optimization objective of this model is to **minimize the reconstruction loss** of both modalities formulated as follows:

$$Loss = \sum_{i=1}^{N} \left(\left\| x_i - \hat{x}_i \right\|_2^2 + \left\| y_i - \hat{y}_i \right\|_2^2 \right) \qquad (16)$$

Similar to the work from Ngiam, Silberer and Lapata [105] proposed a variant to learn semantic representations from textual and visual input. In addition to reconstruction loss, a classification loss is also optimized simultaneously to ensure the ability that different objects can be discriminated based on the learned latent representations. Another variant is the model proposed by Wang *et al.* [106] which imposed orthogonal regularization on the weights to reduce the redundancy in the learned representation.

Other than learning representation in a common subspace, Feng *et al.* [11] proposed to learn a couple of independent while correlated representation for each modality. In their model, each modality is encoded via individual autoencoders. In addition to the reconstruction loss of both modalities, the model minimizes the similarity loss between modalities such that the correlation between them can be captured. The author implied that a balance between both losses is vital for higher performance. This idea is also adopted by Wang *et al.* [107] who assigned separated weights to reconstruction loss of different modalities.

Besides the above-mentioned models, autoencoders are also used for extracting intermediate features. Generally, this type of models can be characterized as two stages of learning strategy. In the first step, based on unsupervised learning, the modality-specific features are extracted via separated autoencoders. Then, in the next step, a particular supervised learning procedure will be imposed to capture the cross-modal correlations. For example, based on autoencoders, Liu *et al.* [6] extracted modality-specific features separately, then fused them in a neural network via supervised learning. Another instance is the work of Hong *et al.* [108], which learns a mapping from one modality to another based on the features learned from autoencoders.

The first advantage of autoencoders is that the learned latent representation can preserve the dominant semantic information of input data. In the view of the generative model, since the input can be reconstructed from this latent representation, it is believable that the critical factors for generating the input have been encoded. The second advantage is that it can be trained by unsupervised manner, without labels required. However, since this model is mainly designed for general purpose, in order to improve its performance in specific tasks, additional constraints or supervised learning process should be involved.

C. DEEP CANONICAL CORRELATION ANALYSIS

Canonical correlation analysis (CCA) [109] is a method originally used for measuring the correlation between a pair of sets. In the multimodal representation learning scenario, given two sets of data $x[x_1, x_2, \cdots, x_n] \in R^{n \times d_x}$ and $y[y_1, y_2, \cdots, y_n] \in R^{n \times d_y}$, where each pair (x_i, y_i) is a data sample including two modalities, CCA aims to find two sets of basis vectors w_x and w_y used for mapping multimodal data into a shared d dimensional subspace, such that the correlation between the projected representations, $P_x = w_x^T x$ and $P_y = w_y^T y$, is maximized [5], [110]. In the case each set x and y has a zero mean, the objective function can be written as (17), where ρ denotes the correlation coefficient, and C denotes the covariance matrix.

$$\begin{aligned} \rho &= \max_{w_x, w_y} corr(w_x^T x, w_y^T y) \\ &= \max_{w_x, w_y} \frac{w_x^T C_{xy} w_y^T}{\sqrt{(w_x^T C_{xx} w_x)(w_y^T C_{yy} w_y)}} \end{aligned} \qquad (17)$$

Since ρ is invariant to the scale of w_x or w_y, the optimization objective can be further reformed as a constrained optimization problem as follows:

$$\max_{w_x, w_y} w_x^T C_{xy} w_y^T \; s.t. \; w_x^T C_{xx} w_x = 1, \quad w_y^T C_{yy} w_y = 1 \tag{18}$$

The basic CCA is limited to modeling linear relationship, regardless of the truth of probability distribution in different data views. To address this problem, many extensions have been proposed. One of the non-linear extensions is kernel CCA [111] which transforms the data into a higher dimensional Hilbert space before applying the CCA method. However, KCCA suffers from poor scalability [112], in that its closed form solution requires computation of high time complexity and memory consumption. Alternatively, some approximation methods such as Nyström method [113], incomplete Cholesky decomposition [114], partial Gram Schmidt orthogonalization [115], and block incremental SVD [116] can be used to speed up this model. Another drawback of KCCA is its poor efficiency, which results from its requirement of accessing to all training sets when transforming an unseen instance [117].

A new extension of CCA is deep CCA [117], which aims to learn a pair of more complex non-linear transformation for different modalities. The basic structure of this model can be illustrated by Fig. 2(b), where each modality is encoded by a deep neural network, then in a common subspace, the canonical correlation between modalities is maximized. Let $H_x = f_x(x, \theta_x)$ and $H_y = f_y(y, \theta_y)$ are non-linear transformation functions implemented by neural network which mapped x and y into a shared subspace, the optimization objective is to **maximize the cross-modality correlation** between H_x and H_y formulated as follows:

$$\max corr(H_x, H_y) = \max_{\theta_x, \theta_y} corr(f_x(x, \theta_x), f_y(y, \theta_y)) \tag{19}$$

Comparing to a particular kernel function used in KCCA, the non-linear function learned from the neural network is far more general. Hence, DCCA exhibits better performance in adaptability and flexibility. Meanwhile, as a parametric method, DCCA scales better with data size and does not require to reference to train data during testing.

Commonly, a maximized correlation objective focuses on learning the shared semantic information but tends to ignore modality specific knowledge. To address this problem, extra regularization terms should be considered. For example, Wang *et al.* [118] proposed a variant of DACC name deep canonically correlated autoencoders (DCCAE). In addition to maximize the correlation between views, this model also minimizes the reconstruction error of each view via autoencoders architecture. The role of additional autoencoders can be interpreted as a regularization item which aims to raise the lower bound of mutual information between views.

So far, most DCCA based applications can be characterized as predicting one modality given another, while DCCA can also be used to generate novel samples. Based on the probabilistic interpretation of CCA [119], Wang *et al.* [120] proposed an extension named deep variational canonical correlation analysis (VCCA). As a generative model, VCCA enables us to obtain unseen samples of each view. The basic probabilistic interpretation of CCA assumes that two views of observed variable x and y are generated according to conditional probabilities p($x|z$) and $p(y|z)$, where z is a latent variable shared by both views. Other than a linear assumption between x, y and z, implemented via DNN network, VCCA aims to model a non-linear relationship among them, which potentially has a stronger representation power. Specifically, the optimization objective of VCCA is a variational lower bound of the likelihood which can be expressed as a sum over data samples. Hence, the model can be trained via stochastic gradient descent method conveniently.

A challenge for DCCA is its relatively poor scalability. Directly inherited from basic CCA, the standard correlation function couples all training samples together and cannot be expressed as a sum of all data samples. Thus, Andrew *et al.* [117] choose a batch-based algorithm (L-BFGS) to optimize the network. However, it computes gradients over entire data samples and requires high memory volume which is infeasible for large datasets. In order to improve the scalability of DCCA, some efforts have been made. Wang *et al.* [121], [122] adopted a stochastic optimization method with large mini-batch to approximate the gradients. As a result, the problem of memory consumption is relieved.

Recently, a more efficient optimization solution named Soft CCA, which requires lower computation complexity, has been proposed by Chang *et al.* [123]. Unlike to traditional CCA which constrains the correlation matrix over the training batch to be an identity matrix, Soft CCA relaxes this constraint to a loss in (20), which minimizes the L_1 loss of off-diagonal element in constraint matrix. By expressing CCA objective as a loss function, Soft CCA avoids some computationally expensive operations such as matrix inversion and singular value decomposition (SVD). Thus, Soft CCA is effective and more scalable in computation.

$$L_{SDL} = \sum\nolimits_{i=1}^{k} \sum\nolimits_{j \neq i}^{k} |\phi_{ij}| \tag{20}$$

Comparing to another type of model in the coordinated framework, cross-modal similarity method, one of the advantages of DCCA is that it can be trained in an unsupervised manner. Due to these advantages, DCCA has been widely used for various multi-view and multimodal learning tasks including word embedding in a multilingual context [124], [125], acoustic features representation [121], matching images and text [29], music retrieval [126], and speech recognition [127], [128]. On the contrary, the drawback of DCCA is the higher computation complexity which may limit its scalability in data size.

D. GENERATIVE ADVERSARIAL NETWORK

Generative adversarial network (GAN) is an emerging deep learning technique. As an unsupervised learning method,

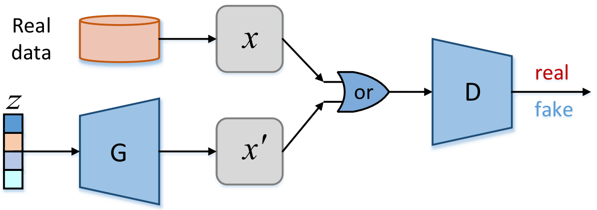

FIGURE 5. The conceptual structure of basic generative adversarial networks.

it can be used for learning data representation without involving labels, which will significantly lower the dependence on manual annotations. Also, as a generative method, it can be used for generating high-quality novel samples according to the distribution of training data. Since 2014, after being proposed by Goodfellow *et al.* [82], the generative adversarial learning strategy has been successfully used for various unimodal applications. One of the best-known applications is image synthesis [82], [129], [130], which generates high-quality images according to a random input drawn from a normal distribution. The other successful examples including image-to-image translation [131] and image super-resolution [132]. Most recently, generative adversarial learning strategy is further extended to multimodal cases such as text-to-image synthesis [15], [44], visual captioning [40], [43], cross-modal retrieval [30], multimodal features fusion [4], and multimodal storytelling [133]. In this section, we will briefly introduce the fundamental concepts of GAN and discuss its role in multimodal representation learning.

Generally, a generative adversarial network is composed of two components, a generative network G playing as a generator and a discriminative network D playing as a discriminator, contesting with each other. The network G is responsible for generating new samples according to the learned data distribution. While the network D aims to discriminate the difference between an instance generated by network G and an item sampled from the training set. Commonly, both components, G and D, are implemented via deep neural networks.

The generator G can be considered as a function mapping a vector in latent space, z, into a sample in data space, and this mapping can be formulated as $G(z; \theta_g) \rightarrow x$, where θ_g is the parameters of G. Similarly, the discriminator D can be formulated as $D(x, \theta_d) \rightarrow p$, mapping a matrix or a vector into a scalar probability value predicting whether a sample is drawn from training data or not, where θ_d is the parameters of D and $p \in (0, 1)$. Although G generates novel samples from distribution $P_g(x)$, it endeavors to capture the ground truth $P_{data}(x)$. Once the distribution P_g estimates P_{data} well enough, the discriminator D will be confused, and its prediction accuracy will be lowered. Theoretically, Goodfellow *et al.* [82] shows that the global optimum can be achieved on condition that $P_g = P_{data}$. In such a case, the discriminator is unable to distinguish the difference between them, and the predicted probability p will be 0.5 for all inputs.

$$\min_G \max_D V(G, D) \tag{21}$$

$$V(G, D) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}(1 - \log D(G(z))) \tag{22}$$

The optimization objective of GANs is a solution of (21), where function $V(G, D)$ is the cross-entropy loss of discriminator D which formulated in (22). During the training process, G and D will be updated in an iterative paradigm. While one of the components is updated, the parameters of another one will keep fixed. In step one, given samples from either generator or training dataset, the discriminator is trained to tell them apart. This objective is achieved by maximizing function V. On the other hand, in step two, the generator is trained to produce samples sufficient to confuse the discriminator. This objective is achieved by minimizing the function V. In such an adversarial manner, both subnets evolve alternately.

Comparing to classic representation learning methods, a visible difference for GANs is that the learning process of data representation is not straightforward. It is rather in an implicit paradigm. Unlike traditional unsupervised representation methods, such as autoencoders, which learns a mapping from data to latent variables directly, GANs learns a reverse mapping from latent variables to the data samples. Specifically, the generator maps a random vector into a distinctive sample. Thus, this random signal is a representation corresponding to generated data. On condition that P_g fits P_{data} well, this random signal is a good enough representation for realistic training data.

However, despite the success of GANs in image synthesis, a disadvantage of basic GANs is that the latent representation is hard to be interpreted since such a random representation has no connection with meaningful semantics. To improve the interpretability of this latent representation, Chen *et al.* [134] introduced a semantically meaningful method name Info-GAN which separates the random noise vector into several groups, z and $c = (c_1, \ldots, c_L)$. By maximizing the mutual information between latent variable c and generator distribution $G(z, c)$, the model encourages the different c_i to represent uncoupled salient attributes. As a result, a modification on the value of c_i will lead to a change of its relevant data attributes such as shape or style.

Another disadvantage of basic GANs is its lacking of a direct mapping from data to latent space which is critical for representation learning in traditional tasks such as retrieval and classification. To address this problem, some techniques equipped with an additional inference network have been proposed [135], [136]. Other typical models which can translate representations between data space and latent space bi-directionally include Adversarially Learned Inference model (ALI) [137] and Bidirectional Generative Adversarial Networks (BiGANs) [138]. In these models, the generator comprises a pair of parallel networks: a decoder used for mapping a latent vector z into a novel sample $\hat{x}$,

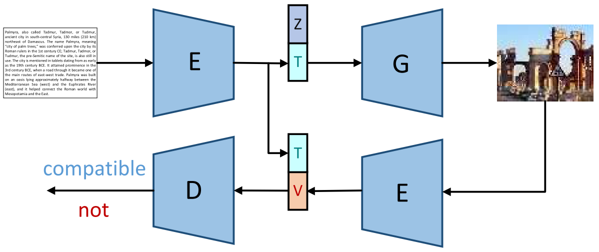
Palmyra, also called Tadmur, Tadmor, or Tudmur, ancient city in south-central Syria, 130 miles (210 km) northeast of Damascus. The name Palmyra, meaning "city of palm trees," was conferred upon the city by its Roman rulers in the 1st century CE; Tadmur, Tadmor, or Tudmur, the pre-Semitic name of the site, is also still in use. The city is mentioned in tablets dating from as early as the 19th century BCE. It attained prominence in the 3rd century BCE, when a road through it became one of the main routes of east-west trade. Palmyra was built on an oasis lying approximately halfway between the Mediterranean Sea (west) and the Euphrates River (east), and it helped connect the Roman world with Mesopotamia and the East.

FIGURE 6. **The conceptual structure of text-to-image generative adversarial networks.**

and an encoder which is responsible for inferring $\hat{z}$ from x. The decoder and the encoder will be optimized jointly such that the tuples $\left(\hat{x}, z\right)$ and $\left(x, \hat{z}\right)$ are similar enough to fool the discriminator.

Most recently, generative adversarial learning strategy has been extended to multimodal representation cases, mainly including cross-modal translation and retrieval. Although in both applications, the core role of adversarial learning is **narrowing the distribution difference** between modalities, their focuses are slightly different. Specifically, in cross-modal translation applications, GANs will help the encoder to capture shared semantic concepts among modalities, while, in cross-modal retrieval, given paired multimodal inputs, they will help the coupled encoders to yield paired representations that are similar enough in common subspace.

In cross-modal translation area, take text-to-image synthesis as an example, one of the key challenges is to properly encode visual concepts such as object categories, colors and location from text descriptions into a vector such that another modality can be generated correctly according to this intermediate representation. To address this problem, based on conditional generative adversarial nets (CGAN) [139], Reed *et al.* [15] proposed an end-to-end architecture to train the text encoder. As Fig. 6 illustrated, in this model, the text input acted as the condition is encoded into vector T, then T along with a noise vector Z are translated into an image, after that, the discriminator will tell whether T and the image encoding V is compatible or not. To gain a visually-discriminative vector representation of text descriptions, the optimization objective is a structured loss [140] formulated as follows:

$$\frac{1}{N}\sum\nolimits_{n=1}^{N} \Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n)) \tag{23}$$

where $\{(v_n, t_n, y_n), n = 1, \ldots, N\}$ is the training set, Δ is the 0-1 loss, v_n are the images, t_n are the text descriptions, and y_n are the class labels. Classifiers f_v and f_t are defined as follows:

$$f_v(v) = \underset{y\in\mathcal{Y}}{\arg\max}\mathbb{E}_{t\sim\mathcal{T}(y)}\left[\phi(v)^T\varphi(t)\right)] \tag{24}$$

$$f_t(t) = \underset{y\in\mathcal{Y}}{\arg\max}\mathbb{E}_{v\sim\mathcal{V}(y)}\left[\phi(v)^T\varphi(t)\right)] \tag{25}$$

where φ denotes the text encoder, ϕ denotes the image encoder, $\mathcal{T}(y)$ denotes the text set belongs to class y and likewise $\mathcal{V}(y)$ for images. Via optimizing loss function (23), the adversarial process between G and D will not only guide the generator to align images with the text descriptions but also help the text encoder to capture shared visual semantic concepts among modalities.

To improve the performance of text-to-image synthesis, several models [44], [141], [142] which share the same basic structure illustrated in Fig. 6 have been proposed. In different ways, they improved the text encoder such that visual information from text descriptions can be encoded more explicitly. Zhang *et al.* [44] adopted a sketch-refinement process to generate photo-realistic images. Conditioned on text descriptions, their model firstly sketches a low-resolution image and then generates a high-resolution image in the refinement stage. Also, in this model, in order to improve the diversity of the synthesized images, they introduce a Conditioning Augmentation technique to encourage the text encoding to be smooth in the latent conditioning space. In [141], Reed *et al.* combined object location information, which is provided by bounding boxes or key points, with the text descriptions to describe what content to draw in which location. Other than using a sentence as the condition, Johnson *et al.* [142], instead, proposed to use a scene graph as the input of the translation network. To process the scene graphs, in the proposed model, a graph convolution network is designed to encode the nodes and edges information into representation vectors. Comparing to unstructured text, the structured scene graphs which describe objects and their relationships explicitly will help for generating complex images.

In the cross-modal retrieval area, the main role of GANs is to help the coupled encoders to yield paired representations that are similar enough in common subspace. The key idea is mapping paired inputs into a common subspace such that the discriminator cannot distinguish which modality a feature comes from. According to the input contents of the discriminator, the typical structures of cross-modal adversarial models can be generalized into two categories. In the first category, which is illustrated in Fig. 7(a), the inputs of the modality discriminator are features generated by encoders. While in the second category illustrated by Fig. 7(b), the inputs are data samples. The rest of this section, we will describe both types of learning strategies.

As Fig. 7(a) showed, the cross-modal adversarial model of the first category is composed of two generators and a discriminator. Each generator is a feature encoder used for mapping either text or images into a common latent subspace, where features from different modalities can be compared directly. The desired goal is to narrow the distribution gap of different modalities, which means that the data with a similar semantic from different modalities may be mapped into the adjacent points in common space. During training, the generators seek to yield modality-invariant representations; on the contrary, a modality classifier, also the discriminator of GANs, is used for discriminating where a feature comes from. Once the discriminator cannot distinguish the source

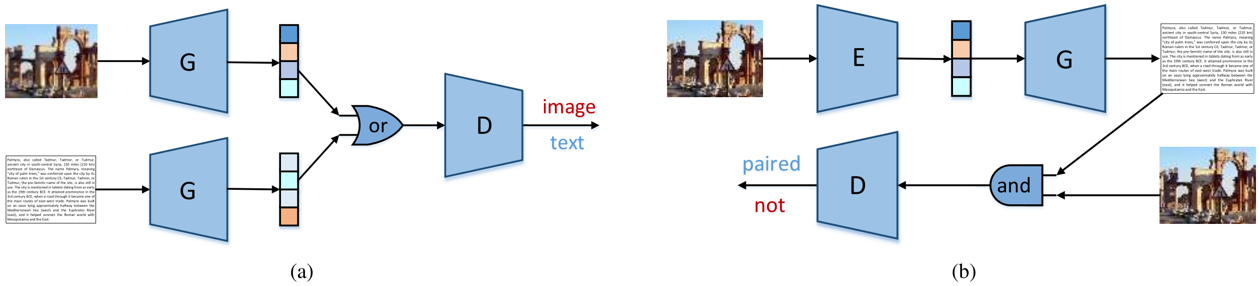

FIGURE 7. Two methods used for improving modality-invariant property via adversarial learning. The key idea is mapping paired inputs into a common subspace such that the discriminator cannot distinguish which modality a feature comes from. (a) Discriminate which modality a feature comes from. (b) Discriminate whether the input is a pair or not.

of feature vectors, the distribution gap of different modalities will be minimized accordingly.

Based on the learning strategies of the first category, several models used for cross-modal retrieval have been proposed [4], [30], [143]. In these models, the adversarial process is served to enforce the distributions of projected representations from different modalities to be closer to each other. The main difference between them is the way how to preserve the intra-modality and inter-modality similarities simultaneously. For example, Wang *et al.* [30] proposed to learn presentations that are modality-invariant and discriminative. In addition to the modality classifier, a label predictor is also integrated into this model to keep the learned features discriminative within each modality. Further, a triplet margin rank constraint is added to the label classifier such that inter-modality similarity can be preserved.

Peng *et al.* [4] proposed to learn discriminative common representation for bridging the heterogeneity gap. In their model, the generator is formed by a cross-modal autoencoder with weight-sharing constraint, and the discriminator is composed of two kinds of discriminative modules: intra-modality and inter-modality discriminators. The generator seeks to project multimodal inputs into common subspace with two useful properties, keeping semantic consistency within each modality and distribution consistency among modalities, on the contrary, the discriminators tries to detect the inconsistency. Specifically, the intra-modality discriminator aims to distinguish generated reconstruction feature from the original input, while the inter-modality discriminator endeavors to tell which modality a feature comes from.

The model proposed by Xu *et al.* [143] aims to learn cross-modal representations which are maximally correlated and statistically indistinguishable in the common subspace. They decompose the whole problem into three loss terms: an adversarial loss which is utilized to minimize the statistical difference between distributions of different modalities, a feature discrimination loss which ensures the representations to be discriminative within each modality, and a cross-modal correlation loss which is responsible for keeping cross-modal similarity structure. Specifically, the cross-modal correlation loss is measured by the square distance between pairs of samples come from different modalities. If a pair come from the same category, its distance will be minimized. Otherwise, it is maximized.

As Fig. 7(b) showed, the cross-modal adversarial model of the second category contains an encoder-decoder network, which translates one modality into another. For example, given a pair of input (v, t), the encoder maps t into a representation vector, then the decoder, playing as the generator, maps this vector into a reproduced sample $\hat{v}$. The generated sample $\hat{v}$ is expected to sufficiently similar to v, such that the reproduced pair $(\hat{v}, t)$ is considered as a real pair by the discriminator. On condition that the learned representation can be translated into another modality soundly, it is believable that the cross-modal invariant property has been preserved. An example in this category is the model proposed by Gu *et al.* [86] which integrated a generative adversarial network in their model to train a text encoder. In the following, more examples will be shown to demonstrate how this model can be used in practice.

Zhang *et al.* [144] adopted GANs to model cross-modal hashing in an unsupervised fashion. In addition to preserving inter-modality and intra-modality correlations in the common hash space, the property preserving manifold structure across different modalities is also desired in their model. Given a sample from a modality, the generator is trained to select a sample from another modality located in the same manifold. Then, the discriminator will determine whether the generated pair of samples belonging to the same manifold structure or not. Here, the hash codes play a key role for both generator and discriminator. Specifically, the generator selects samples conditioned on hash codes; also, the discriminator judges their correlation between modalities based on hash codes. The adversarial learning process is used for enhancing the property of preserving cross-modal manifold structure in a common hash space.

Wu *et al.* [145] extended CycleGAN [146] to learn cross-modal hash functions on the condition without paired training samples are available. CycleGAN can be seen as a special case of the second category, which includes a pair of encoder-decoder, each of them is designed to translate one modality into another. For example, given an input v,

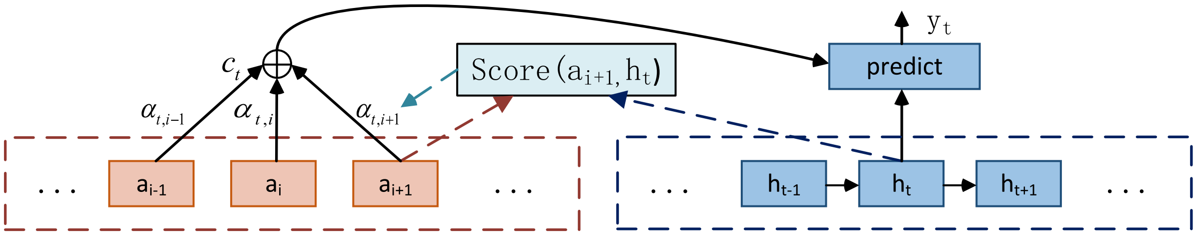

FIGURE 8. The typical structure of key-based attention mechanism. The attention module uses the current state (h_t) as a key to search salient elements in the source ($\{a_i\}$).

the model translates v into t, and then t is reversely translated back to $\hat{v}$, it is expected that $v \approx \hat{v}$. Similarly, given an input t, a reconstructed $\hat{t}$ is expected to roughly equal with t. Based on the cycle-consistent constraint in both modalities, the model can be trained in the absence of paired training samples.

One of the advantages of GAN is that it can be trained by unsupervised learning which will significantly lower the dependence on manual annotations. Another advantage is its powerful ability to generate high-quality novel samples according to the distribution of training data. However, though a unique global optimum is existent theoretically, it is challenging to train a GAN system which may suffer from training instability, either "collapsing" or failing to converge [147]. Although several improvements have been proposed [147]–[150], the way for stabilizing the training of GANs remains an open problem.

E. ATTENTION MECHANISM

Attention mechanism allows a model to focus on specific regions of a feature map or specific time steps of a feature sequence. Via attention mechanism, not only an improved performance can be achieved, but also better interpretability of feature representations can be seen. This mechanism mimics the human ability to extract the most discriminative information for recognition. Rather than using all of the information at once, the attention decision process prefers to concentrate on the part of the scene selectively which is needed [151]. Recently, this method has demonstrated its unique power in improving performance in many applications such as visual classification [152]–[154], neural machine translation [155], [156], speech recognition [92], image captioning [13], [91], video description [42], [90], visual question-answering [24], [157], cross-modal retrieval [31], [158], and sentiment analysis [22].

According to whether a key is used during selecting part of the features, attention mechanism can be categorized into two groups: key-based attention, and keyless attention. Key-based attention used a key to search for salient localized features. Take image caption as an example [13], its typical structure can be illustrated as Fig. 8, where a CNN network encodes the image into a feature set $\{a_i\}$, and then an RNN network decodes the input into hidden states $\{h_t\}$. In time step t, the output y_t is predicted based on h_t and c_t, where c_t is the salient feature summarized from $\{a_i\}$. During the process of extracting the salient feature c_t, the current state h_t in decoder plays as a key and the encoder states $\{a_i\}$ play as a source to be searched [159]. The computation method of attention mechanism [13], [156] can be defined as (26) to (28), and the compatibility scores between the key and the sources can be evaluated via one of the three different functions listed in (29).

$$e_{ti} = score(a_i, h_t) \tag{26}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{i=1}^{L} \exp(e_{ti})} \tag{27}$$

$$c_t = \sum\nolimits_{i=1}^{L} \alpha_{ti} a_i \tag{28}$$

$$score(a_i, h_t) = \begin{cases} h_t^T a_i \\ h_t^T W_a a_i \\ v_a^T \tanh(W_a[h_t; a_i]) \end{cases} \tag{29}$$

Key-based attention is widespread in visual description applications [13], [90], [160], where an encoder-decoder network is commonly used. It brings us an approach to **evaluate the importance of the features** within a modality or among modalities. On the one hand, attention mechanism can be used to select the most salient features within a modality, on the other hand, it can be used to balance the contribution among modalities during fusing several modalities.

In order to recognize and describe objects contained in the visual modality, a set of localized region features, which potentially encode different objects distinctly, would be more helpful than a single feature vector. By selecting the most salient regions in an image or time steps of a video sequence dynamically, both system performance and noise tolerance can be improved. For example, Xu *et al.* [13] adopted attention mechanism to detect salient objects in an image and fused them with text features in a decoder unit for captioning. In such a case, guided by current text generated in time step t, the attention module will be used to search local regions appropriate for predicting next word.

For locating local features more accurately, several attention models have been proposed. Yang *et al.* [157] proposed a stacked attention network for searching image regions. They suggested that multiple steps of search or reasoning are helpful to locate to fine-grained regions. In the beginning, the model locates one or more local regions in the image by attention using language features as a key and then combines the attended visual and language features into a vector, which also plays as a key used for next iterator. After K steps, not only the appropriate local regions are located, but both features are fused. Zhu *et al.* [161] proposed a structured attention model to capture the semantic structure among image regions, and their experiments showed that this model is capable of inferring spatial relations and attending to the right region. Chen *et al.* [162] proposed to incorporate spatial and channel wise attentions in a CNN network. In their model, not only local regions but also channels of CNN features are filtered simultaneously.

So far, attention models are mostly trained using indirect cues because of lacking explicit attention annotations. Alternatively, Gan *et al.* [163] trained the attention module using direct supervision. They collected link information between visual segments and words from several datasets and then utilized the link information to guide the training of attention module explicitly. The experiments showed that improved performance could be achieved.

Balancing the contribution of different modalities is a key issue that should be considered during fusing multimodal features. By contrast to concatenation or fixed weights fusion methods, an attention-based method can adaptively balance the contributions of different modalities. Several pieces of research [90], [91], [164] have been reported that dynamically assigning weights to modality-specific features condition on a context is helpful to improve application performance.

Hori *et al.* [90] proposed to tackle multimodal fusion based on attention for video description. In addition to attending on specific regions and time steps, the proposed method highlights attending on modality-specific information. After modality-specific features have been extracted, the attention module produces appropriate weights to combine features from different modalities based on the context. In a cross-modal retrieval task, Chen *et al.* [164] adopted a similar strategy to adaptively fuse modalities and filter out unrelated information within each modality according to search keys.

Lu *et al.* [91] introduced an adaptive attention frame to determine whether including a visual feature or not during generating the caption. They argued that some words such as ''the'' are not related to any visual object. Therefore, no visual feature is needed in this case. Suppose that the visual feature is excluded, the decoder would just depend on the language features to predict a word.

Keyless attention is mostly used for classification or regression task. In such an application scene, since the result is generated in a single step, it is hard to define a key to guide the attention module. Alternatively, the attention is applied directly on the localized features without any key involved. The computation functions can be illustrated as flow:

$$e_i = score(a_i) \tag{30}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=1}^{L} \exp(e_i)} \tag{31}$$

$$c_i = \sum_{i=1}^{L} \alpha_i a_i \tag{32}$$

$$score(a_i) = \begin{cases} v^T a_i \\ v^T \tanh(W a_i) \end{cases} \tag{33}$$

Due to the nature to select prominent cues from raw input, keyless attention mechanism is suitable for a multimodal feature fusion task which suffers from issues such as semantic confliction, duplication, and noise. Through the attention mechanism, it provides us an approach to evaluate the relationship between parts of modalities, which may be complementary or supplementary. By **selecting complementary features** from different modalities and fusing them into a single representation, the semantic ambiguity could be eased.

The advantage of attention mechanism in multimodal fusion has been proven in many applications. For example, Long *et al.* [165] compared four multimodal fusion methods and demonstrated that attention based method is the most effective one for addressing the video classification problem. They performed experiments in different setups: early fusion, middle-level fusion, attention-based fusion, and late fusion, which corresponding to different fusion points. The experimental result also shows that attention based fusion method is robust across various datasets. Some other researches also demonstrated the promising perspective of attention based methods for multimodal features fusion [166], [167].

A special issue on multimodal feature fusion is fusing features from several variable length sequences such as videos, audios, sentences or a set of localized features. A simple way to tackle this problem is fusing each sequence independently via the attention mechanism. After each sequence has been combined into a weighted representation with a fixed length, they will be concatenated or fused into a single vector. This way is beneficial for fusing several sequences, even in the case that their lengths are different, which is commonly seen in a multimodal dataset. However, such a simplified method does not explicitly consider the interaction between modalities, and thus may ignore the fine-grained cross-modal relationships.

A solution to model the interactions between attention modules is constructing a shared context as an extra condition for the computation of modality-specific attention modules. For example, Lu *et al.* [24] proposed to construct a global context by calculating the similarity between visual and text features. Nam *et al.* [158] used an iterative strategy to update the shared context and modality-specific attention distribution. Firstly, modality-specific features will be summarized based on attention modules, then they are fused into a context used for next iterator.

Recently, a novel learning strategy named multi-attention mechanism, which utilizes several attention modules to extract different types of features from the same input data, has been exploited. Generally, each type of feature locates in a distinct subspace and reflects different semantics. Hence, the multi-attention mechanism is helpful in discovering different inter-modal dynamics. For example, Zadeh *et al.* [22] proposed to discovery diverse interactions between modalities using multi-attention mechanism. At each time step t,

TABLE 3. A summary of the key issues, advantages and disadvantages for each framework or typical model described in this paper. One thing should be mentioned is that both cross-modal similarity model and deep canonical correlation analysis (DCCA) are belonged to coordinated representation framework.

Frameworks and models	Key issues	Advantages	Disadvantages
Joint representation	obtain modality-invariant property	fuse several modalities	cannot infer individual modality
	fuse complementary semantics		
Coordinated representation	maximize the cross-modal similarity or correlation	infer each modality individually	hard to coordinate modalities more than two
Cross-modal similarity	preserve inter-modality and intra-modality similarity	measure the cross-modal similarity	
DCCA	maximize cross-modal correlation	unsupervised learning	
Encoder-decoder	capture shared semantics	generate novel samples	encode only one of the modalities
PGM	maximize the joint distribution	generate the missing modality	high computational cost
		unsupervised learning	
Multimodal autoencoders	minimize the reconstruction loss	persevere modality-specific characteristics	designed for general purpose
		unsupervised learning	
GAN	narrow the distribution difference	generate high quality novel samples	suffer from training instability
		unsupervised learning	
Attention mechanism	evaluate the importance of features	select salient localized features	no obvious drawback reported
	select complementary features	filter noise	

the hidden states h_t^m from all modalities were concatenated into vector h_t, then multi-attentions will be applied on h_t to extract K different weighted vectors which reflect distinctive cross-modal relationships. After that, all the K vectors are fused into a single vector which represents the shared hidden state across modalities at time t.

Another example is the model form Zhou *et al.* [167], which fused heterogeneous features of user behavior based on multi-attention mechanism. Here, a user behavior type can be seen as a distinctive modal, because different types of behaviors have distinctive attributes. The author supposed that the semantics of user behavior can be affected by the context. Hence, the semantic intensity of that behavior also depends on the context. Firstly, the model project all types of behaviors into a concatenated vector denoted as S, which is a global feature and plays as the context in the attention module. Then, S is projected into K latent semantic sub-spaces to represent different semantics. After that, the model fuses K sub-spaces through attention module.

One of the advantages of attention mechanism is its capability to select salient and discriminative localized features, which can not only improve the performance of multimodal representations but also lead to better interpretability. Additionally, by selecting prominent cues, this technique can also help to tackle issues such as noise and help to fuse complementary semantics into multimodal representations.

IV. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we provided a comprehensive survey on deep multimodal representation learning. According to the underlying structures in which different modalities are integrated, we category deep multimodal representation learning methods into three groups of frameworks: joint representation, coordinated representation, and encoder-decoder. Additionally, we summarize some typical models in this area, which range from conventional models to newly developed technologies, including probabilistic graphical models, multimodal autoencoders, deep canonical correlation analysis, generative adversarial networks, and attention mechanism. For each framework or model, we describe its basic structure, learning objective, and application scenes. Additionally, we also discuss their key issues, advantages, and disadvantages which have been briefly summarized in Table 3.

When coming into the learning objectives and key issues in all kinds of learning frames or typical models, we can clearly see that the primary objective of multimodal representation learning is to narrow the distribution gap in a joint semantic subspace while keeping modality specific semantic intact. They achieve this objective in different ways: joint representation framework maps all modalities into a global common subspace; coordinated representation framework maximizes the similarity or correlation between modalities while keeping each modality independent; encoder-decoder framework maximizes the condition distribution among modalities and keep their semantics consistent; probabilistic graphical models maximize the joint probability distribution across modalities; multimodal autoencoders endeavor to keep modality specific distribution intact by minimizing the reconstruction errors; generative adversarial networks aims to narrow the distribution difference between modalities by an adversarial process; attention mechanism selects salient features from modalities, such that they are similar in local manifolds or such that they are complementary with each other.

With the rapid development of deep multimodal representation learning methods, the need for much more training data is growing. However, the volume of the current multimodal datasets is limited because of the high cost of manual labeling. The acquirement of high-quality labeled datasets is extremely labor-consuming. A popular solution to address this problem is transfer learning, transferring general knowledge from the source domain with a large-scale dataset to target domain with insufficient data [168]. Transfer learning has been widely used in the multimodal representation learning area and has been shown to be effective in improving performance in many multimodal tasks. One of the examples is the reuse of pre-trained CNN network such as VGGNet [48], ResNet [49], which can be used for extracting image features in a multimodal system. The second example is word embeddings such as word2vec [50], Glove [51]. Although these representations of words are trained only on general-purpose language corpora, they can be transferred to other datasets directly even without fine-tuning.

In contrast to the widespread use of convenient and effective knowledge transfer strategy in image and language modality, similar methods are not yet available within audio or video modality. Hence, the deep networks used for extracting audio or video features would more easily suffer from overfitting due to the limited training instances. As a result, in many applications such as sentiment analysis and emotion recognition which based on fused multimodal features, it is relatively hard to improve the performance when only audio and video data are available. Alternatively, most works have to increasingly rely on a stronger language model. Although some efforts have been made to transfer cross-domain knowledge to audio and video modalities, in the multimodal representation learning area, more convenient and effective methods are still required.

In Addition to the knowledge transferring within the same modality, cross-modal transfer learning which aims to transfer knowledge from one modality to another is also a significant research direction. For example, recent studies show that the knowledge transferred from images can help to improve the performance of video analysis tasks [169]. Besides, an alternative but the more challenging approach is the transfer learning between multimodal datasets. The advantage of this method is that the correlation information among different modalities in the source domain can also be exploited, while the weakness is its complexity, both modality difference and domain discrepancy should be tackled simultaneously.

Another feasible future direction to tackle the problem of relying on large scale labeled datasets is unsupervised or weakly supervised learning, which can be trained using the ubiquitous multimodal data generated by internet users. Unsupervised learning has been widely used for dimensionality reduction and feature extraction on unlabeled datasets. That is why conventional unsupervised learning methods such as multimodal autoencoders are still active today, although comparing to CNN or RNN features their performance are not so good. Due to a similar reason, generative adversarial nets has recently attracted much attention in the multimodal learning area.

Most recently, weakly supervised learning has demonstrated its potential in exploiting useful knowledge hidden behind the multimodal data. For example, given an image and its description, it is highly possible that a segment can be described by some words in the sentence. Although the one-to-one correspondences between them are fully unknown, the work proposed by Karpathy and Fei-Fei [76] shows that these hidden relationships can be discovered via weakly supervised learning. Potentially, a more promising application of these type of weak supervision based methods is video analysis, where different modalities such as actions, audios, languages have been roughly aligned in the timeline.

For a long time, multimodal representation learning suffers from issues such as semantic confliction, duplication, and noise. Although attention mechanism can be used to address these problems partially, they work implicitly and cannot be controlled actively. A more promising method for this problem is integrating reasoning ability into multimodal representation learning networks. Via the reasoning mechanism, a system would have the capability to select evidence actively which is sorely needed and could play an important role in mitigating the impact of these troubling issues. We believe that the close combination of representation learning and their reasoning mechanism will endow machines with intelligent cognitive capabilities.

REFERENCES

[1] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[2] S. Wang and W. Guo, "Sparse multigraph embedding for multimodal feature representation," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1454–1466, Jul. 2017.

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, p. 746, 1976.

[4] Y. Peng, J. Qi, and Y. Yuan. (2017). "CM-GANs: Cross-modal generative adversarial networks for common representation learning." [Online]. Available: https://arxiv.org/abs/1710.05106

[5] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[6] Y. Liu, X. Feng, and Z. Zhou, "Multimodal video classification with stacked contractive autoencoders," *Signal Process.*, vol. 120, pp. 761–766, Mar. 2016.

[7] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan, "Zero-shot event detection using multi-modal fusion of weakly supervised concepts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2665–2672.

[8] A. Habibian, T. Mensink, and C. G. M. Snoek, "Video2vec embeddings recognize events when examples are scarce," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 2089–2103, Oct. 2017.

[9] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, Jan. 2016.

[10] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[11] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 7–16.

[12] J. Qi and Y. Peng, "Cross-modal bidirectional translation via reinforcement learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 2630–2636.

[13] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[14] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2625–2634.

[15] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[17] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Inf. Fusion*, vol. 38, pp. 43–54, Nov. 2017.

[18] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[19] Y. Li, M. Yang, and Z. Zhang. (2016). "A survey of multi-view representation learning." [Online]. Available: https://arxiv.org/abs/1610.01206

[20] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[21] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 352–364, Feb. 2018.

[22] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, P. Vij, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–35.

[23] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 457–468.

[24] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 289–297.

[25] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3687–3691

[26] L. Pang and C.-W. Ngo, "Mutlimodal learning with deep Boltzmann machine for emotion prediction in user generated videos," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 619–622.

[27] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7596–7599.

[28] F. Feng, R. Li, and X. Wang, "Deep correspondence restricted Boltzmann machine for cross-modal retrieval," *Neurocomputing*, vol. 154, pp. 50–60, Apr. 2015.

[29] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3441–3450.

[30] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 154–162.

[31] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5585–5599, Nov. 2018.

[32] R. Socher, Q. V. L. A. Karpathy, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Trans. Assoc. Comput. Linguistics*, vol. 2, no. 1, pp. 207–218, 2014.

[33] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4594–4602.

[34] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov, "DeViSE: A deep visual-semantic embedding model," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, vol. 2, 2013, pp. 2121–2129.

[35] A. Lazaridou and M. Baroni, "Combining language and vision with a multimodal skip-gram model," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 153–163.

[36] R. Kiros, R. Salakhutdinov, and R. S. Zemel. (2014). "Unifying visual-semantic embeddings with multimodal neural language models." [Online]. Available: https://arxiv.org/abs/1411.2539

[37] X. Huang, Y. Peng, and M. Yuan, "Cross-modal common representation learning by hybrid transfer network," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 1893–1900.

[38] X. Huang and Y. Peng, "Deep cross-media knowledge transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8837–8846.

[39] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3156–3164.

[40] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition GAN for visual paragraph generation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 3362–3371.

[41] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 1494–1504.

[42] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[43] Y. Yang *et al.*, "Video captioning by adversarial LSTM," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5600–5611, Nov. 2018.

[44] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 5907–5915.

[45] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[47] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[50] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (2013). "Efficient estimation of word representations in vector space." [Online]. Available: https://arxiv.org/abs/1301.3781

[51] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[52] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2741–2749.

[53] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[54] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2016, pp. 1715–1725.

[55] H. Peng, E. Cambria, and X. Zou, "Radical-based hierarchical embeddings for Chinese sentiment analysis at sentence level," in *Proc. 13th Int. Flairs Conf.*, 2017, pp. 1–6.

[56] J. L. Elman, "Finding structure in time," *Cognit. Sci.*, vol. 14, no. 2, pp. 179–211, Mar. 1990.

[57] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[58] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[59] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, 2000.

[60] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[61] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2342–2350.

[62] Y. Dai, W. Guo, X. Chen, and Z. Zhang, "Relation classification via LSTMs based on sequence and tree structure," *IEEE Access*, to be published.

[63] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[64] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5, pp. 602–610, 2005.

[65] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling." [Online]. Available: https://arxiv.org/abs/1412.3555

[66] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751.

[67] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 655–665.

[68] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 873–883.

[69] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2016, pp. 1–10.

[70] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[71] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 239–247.

[72] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Proc. Int. Conf. Mach. Learn. Workshop*, vol. 79, 2012, pp. 1–8.

[73] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba, "Cross-modal scene networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2303–2314, Oct. 2018.

[74] S. Wang, H. Zhang, and H. Wang, "Object co-segmentation via weakly supervised data fusion," *Comput. Vis. Image Understand.*, vol. 155, pp. 43–54, Feb. 2017.

[75] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, Jul. 2016.

[76] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3128–3137.

[77] R. Xu, C. Xiong, W. Chen, and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2346–2352.

[78] V. E. Liong, J. Lu, Y. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1234–1244, Jun. 2017.

[79] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 405–420, Feb. 2017.

[80] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5005–5013.

[81] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[82] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[83] N. Mor, L. Wolf, A. Polyak, and Y. Taigman. (2018). "A universal music translation network." [Online]. Available: https://arxiv.org/abs/1805.07848

[84] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.

[85] R. Bernardi *et al.*, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *J. Artif. Intell. Res.*, vol. 55, pp. 409–442, Jan. 2016.

[86] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7181–7189.

[87] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[88] G. Kulkarni *et al.*, "Baby talk: Understanding and generating simple image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1601–1608.

[89] S. Guadarrama *et al.*, "YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2712–2719.

[90] C. Hori *et al.*, "Attention-based multimodal fusion for video description," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 4203–4212.

[91] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 6, Jun. 2017, pp. 375–383.

[92] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 4945–4949.

[93] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 163–171.

[94] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1151–1159.

[95] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 7008–7024.

[96] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.

[97] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[98] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *Proc. 29th Int. Conf. Artif. Intell. Statist.*, 2009, pp. 448–455.

[99] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.

[100] L. Ge, J. Gao, X. Li, and A. Zhang, "Multi-source deep learning for information trustworthiness estimation," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 766–774.

[101] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2329–2336.

[102] R. Salakhutdinov and H. Larochelle, "Efficient learning of deep Boltzmann machines," in *Proc. 30th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 693–700.

[103] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and Helmholtz free energy," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 3–10.

[104] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[105] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2014, pp. 721–732.

[106] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in *Proc. 24th Int. Conf. Artif. Intell.*, 2015, pp. 2291–2297.

[107] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," *VLDB Endowment*, vol. 7, no. 8, pp. 649–660, 2014.

[108] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang, "Multimodal deep autoencoder for human pose recovery," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5659–5670, Dec. 2015.

[109] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.

[110] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[111] S. Akaho. (2006). "A kernel method for canonical correlation analysis." [Online]. Available: https://arxiv.org/abs/cs/0609071

[112] N. Mallinar and C. Rosset. (2018). "Deep canonically correlated LSTMs." [Online]. Available: https://arxiv.org/abs/1801.05407

[113] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 682–688.

[114] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, Jan. 2002.

[115] N. Cristianini, J. Shawe-Taylor, and H. Lodhi, "Latent semantic kernels," *J. Intell. Inf. Syst.*, vol. 18, nos. 2–3, pp. 127–152, 2002.

[116] R. Arora and K. Livescu, "Kernel CCA for multi-view learning of acoustic features using articulatory measurements," in *Proc. Symp. Mach. Learn. Speech Lang. Process.*, 2012, pp. 1–4.

[117] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[118] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 1083–1092.

[119] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Dept. Statist., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. 688, 2005.

[120] W. Wang, X. Yan, H. Lee, and K. Livescu. (2016). "Deep variational canonical correlation analysis." [Online]. Available: https://arxiv.org/abs/1610.03454

[121] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 4590–4594.

[122] W. Wang, R. Arora, K. Livescu, and N. Srebro, "Stochastic optimization for deep CCA via nonlinear orthogonal iterations," in *Proc. Allerton Conf. Commun., Control Comput.*, Sep./Oct. 2015, pp. 688–695.

[123] X. Chang, T. Xiang, and T. M. Hospedales, "Scalable and effective deep CCA via soft decorrelation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1488–1497.

[124] A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu, "Deep multilingual correlation for improved word embeddings," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 250–256.

[125] G. Rotman, I. Vulić, and R. Reichart, "Bridging languages through images with deep partial canonical correlation analysis," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 910–921.

[126] Y. Yu, S. Tang, F. Raposo, and L. Chen. (2017). "Deep cross-modal correlation learning for audio and lyrics in music retrieval." [Online]. Available: https://arxiv.org/abs/1711.08976

[127] Y. Takashima, T. Takiguchi, Y. Ariki, and K. Omori, "Audio-visual speech recognition for a person with severe hearing loss using deep canonical correlation analysis," in *Proc. 1st Int. Workshop Challenges Hearing Assistive Technol.*, 2017, pp. 77–81.

[128] Q. Tang, W. Wang, and K. Livescu, "Acoustic feature learning via deep variational canonical correlation analysis," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1656–1660.

[129] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[130] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–16.

[131] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 5967–5976.

[132] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 4681–4690.

[133] Z. Chen, X. Zhang, A. P. Boedihardjo, J. Dai, and C.-T. Lu, "Multimodal storytelling via generative adversarial imitation learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 3967–3973.

[134] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[135] A. Creswell and A. A. Bharath. (2016). "Inverting the generator of a generative adversarial network." [Online]. Available: https://arxiv.org/abs/1611.05644

[136] Z. C. Lipton and S. Tripathi. (2017). "Precise recovery of latent vectors from generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1702.04782

[137] V. Dumoulin *et al.*, "Adversarially learned inference," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–18.

[138] J. Donahue, P. Krähenbuhl, and T. Darrell, "Adversarial feature learning," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–18.

[139] M. Mirza and S. Osindero. (2014). "Conditional generative adversarial nets." [Online]. Available: https://arxiv.org/abs/1411.1784

[140] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 49–58.

[141] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 217–225.

[142] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1219–1228.

[143] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, Mar. 2019.

[144] J. Zhang, Y. Peng, and M. Yuan, "Unsupervised generative adversarial cross-modal hashing," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[145] L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1602–1612, Apr. 2019.

[146] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 2223–2232.

[147] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[148] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–25.

[149] M. Arjovsky, S. Chintala, and L. Bottou. (2017). "Wasserstein GAN." [Online]. Available: https://arxiv.org/abs/1701.07875

[150] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[151] R. A. Rensink, "The dynamic representation of scenes," *Vis. Cognit.*, vol. 7, nos. 1–3, pp. 17–42, 2000.

[152] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[153] W. Pei, T. Baltrušaitis, D. M. Tax, and L.-P. Morency, "Temporal attention-gated model for robust sequence classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 820–829.

[154] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 3156–3164.

[155] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[156] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[157] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 21–29.

[158] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 299–307.

[159] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[160] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 451–466.

[161] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma, "Structured attentions for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 1291–1300.

[162] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 6298–6306.

[163] C. Gan, Y. Li, H. Li, C. Sun, and B. Gong, "VQS: Linking segmentations to questions and answers for supervised attention in VQA and question-focused semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 1811–1820.

[164] K. Chen, T. Bui, C. Fang, Z. Wang, and R. Nevatia, "AMC: Attention guided multi-modal correlation learning for image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 6203–6211.

[165] X. Long *et al.*, "Multimodal keyless attention fusion for video classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[166] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[167] C. Zhou *et al.*, "ATRank: An attention-based user behavior modeling framework for recommendation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[168] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[169] J. Zhang, Y. Han, J. Tang, Q. Hu, and J. Jiang, "Semi-supervised image-to-video adaptation for video action recognition," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 960–973, Apr. 2017.

WENZHONG GUO received the Ph.D. degree from the Department of Physics and Information Engineering, Fuzhou University, Fuzhou, China, in 2010. He was a Postdoctoral Research Scholar with the Department of Computer Science, National University of Defense and Technology, Changsha, China, from 2011 to 2014, a Visiting Professor with the Faculty of Engineering, Information and System, University of Tsukuba, Japan, in 2013, and a Visiting Professor with the Department of Computer Science and Engineering, State University of New York at Buffalo, USA, in 2016. He is currently a Professor and the Director of the Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing. His research interests include the fields of data mining, machine learning, and artificial intelligence.

JIANWEN WANG is currently pursuing the Ph.D. degree with the College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China. He is currently a Lecturer with the College of Mathematics and Informatics, Fujian Normal University, Fuzhou. His research interests include multimodal machine learning and computer vision.

SHIPING WANG received the Ph.D. degree from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2014. From 2015 to 2016, he was a Research Fellow with Nanyang Technological University, Singapore. He is currently a Qishan Scholar with the College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China. His research interests include machine learning, computer vision, and granular computing.