New Ideas and Trends in Deep Multimodal Content Understanding: A Review

Wei Chen [a], Weiping Wang [b], Li Liu [b,c], Michael S. Lew [a,*]

[a] LIACS, Leiden University, Leiden, 2333 CA, The Netherlands
[b] College of Systems Engineering, NUDT, Changsha, 410073, China
[c] Center for Machine Vision and Signal Analysis, University of Oulu, Finland

ARTICLE INFO

ABSTRACT

The focus of this survey is on the analysis of two modalities of multimodal deep learning: image and text. Unlike classic reviews of deep learning where monomodal image classifiers such as VGG, ResNet and Inception module are central topics, this paper will examine recent multimodal deep models and structures, including auto-encoders, generative adversarial nets and their variants. These models go beyond the simple image classifiers in which they can do uni-directional (*e.g.* image captioning, image generation) and bi-directional (*e.g.* cross-modal retrieval, visual question answering) multimodal tasks. Besides, we analyze two aspects of the challenge in terms of better content understanding in deep multimodal applications. We then introduce current ideas and trends in deep multimodal feature learning, such as feature embedding approaches and objective function design, which are crucial in overcoming the aforementioned challenges. Finally, we include several promising directions for future research.

1. Introduction

Semantic information that helps us illustrate the world usually comes from different sensory modalities in which the event is processed or is experienced (*i.e.* auditory, tactile, or visual). Thus, the same concept or scene can be presented in different ways. If we consider a scene where "*a large yellow dog leaps into the air to catch a frisbee*", then one could select audio or video or an image, which also indicates the multimodal aspect of the problem. To perform multimodal tasks well, first, it is necessary to understand the content of multiple modalities. Multimodal content understanding aims at recognizing and localizing objects, determining the attributes of objects, characterizing the relationships between objects, and finally, describing the common semantic content among different modalities. In the information era, rapidly developing technology makes it more convenient than ever to access a sea of multimedia data such as text, image, video, and audio. As a result, exploring semantic correlation to understand content for diverse multimedia data has been attracting much attention as a long-standing research field in the computer vision community.

Recently, the topics range from speech-video to image-text applications. Considering the wide array of topics, we restrict the scope of this survey to image and text sepcifically in the multimodal research community, including tasks at the intersection of image and text (also called cross-modal). According to the available modality during testing stage, multimodal applications include bi-directional tasks (*e.g.* image-sentence search [1,2], visual question answering (VQA) [3,4]) and uni-directional tasks (*e.g.* image captioning [5,6], image generation [7,8]), both of them will be introduced in the following sections.

With the powerful capabilities of deep neural networks, data from visual and textual modality can be represented as individual features using domain-specific neural networks. Complementary information from these unimodal features is appealing for multimodal content understanding. For example, the individual features can be further projected into a common space by using another neural network for a prediction task. For clarity, we illustrate the flowchart of neural networks for multimodal research in Fig. 1. On the one hand, the neural networks are comprised by successive linear layers and non-linear activation functions, the image or text data is represented in a high abstraction way, which leads to the "semantic gap" [9]. On the other hand, different modalities are characterized by different statistical properties. Image is 3-channel RGB array while text is often symbolic. When represented by different neural networks, their features have unique distributions and differences, which leads to the "heterogeneity gap" [10]. That is to say, to understand multimodal content, deep neural

---

* Corresponding author.
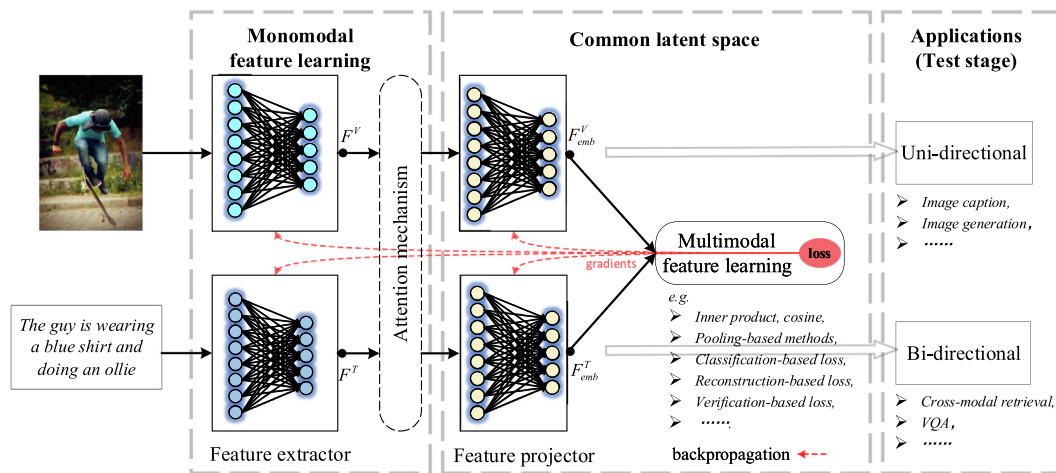 *E-mail address:* m.s.k.lew@liacs.leidenuniv.nl (M.S. Lew).

Fig. 1. A general flowchart of deep multimodal feature learning. Each modality starts with using an individual neural network to process the data (*e.g.* CNN for images and RNN for text), which implements monomodal feature learning. The attention module is an optional module for aligning two monomodal features. The extracted features F^V and F^T are not directly comparable and are distributed inconsistently due to the process of individual domain-specific neural networks. To understand the multimodal content, these monomodal features are embedded into a common latent space with the help of mapping functions (*e.g.* MLP). According to the taxonomy [20], feature embedding in the common space can be categorized into a joint and a coordinated representation. Afterwards, the optimized multimodal features F^T_{emb} and F^V_{emb} are comparable for specific applications. Common latent feature learning is a crucial module in deep multimodal learning, which aims at learning semantically discriminative features. The current ideas and trends for improving performance are the central topic of this paper.

networks should be able to reduce the difference between high-level semantic concepts and low-level features in intra-modality representations, as well as construct a common latent space to capture semantic correlations in inter-modality representations.

Much effort has gone into mitigating these two challenges to improve content understanding. Some works involve deep multimodal structures such as cycle-consistent reconstruction [11–13], while others focus on feature extraction nets such as graph convolutional networks [14–16]. In some algorithms, reinforcement learning is combined with deep multimodal feature learning [17–19]. These recent ideas are the scope of this survey. In a previous review [20], the authors analyze intrinsic issues for multimodal research but mainly focus on machine learning. Some recent advances in deep multimodal feature learning are introduced [21], but it mainly discusses feature fusion structures and regularization strategies.

In this paper, we focus on two specific modalities, image and text, by examining recent related ideas. First, we focus on the structures of deep multimodal models, including auto-encoders and generative adversarial networks [22] and their variants. These models, which perform uni-directional or bi-directional tasks, go beyond simple image classifiers (*e.g.* ResNet). Second, we analyze recent methods of multimodal feature extraction which aim at getting semantically related features to minimize the heterogeneity gap. Third, we focus on current popular algorithms for common latent feature learning, which are beneficial for network training to preserve semantic correlations between modalities. In conclusion, the newly applied ideas mitigate the "heterogeneity gap" and the "semantic gap" between visual and textual modalities.

The rest of this paper is organized as follows: Section 2 introduces image-text related applications, followed by corresponding challenges and intrinsic issues for these image-text applications in Section 3. Regarding these challenges, we analyze the current ideas and trends in deep multimodal learning in Section 4. Finally, we conclude with several promising directions in Section 5.

2. Multimodal applications

This section aims to summarize various multimodal applications where image and text data are involved. These applications

have gained a lot of attention lately and show a natural division into uni-directional and bi-directional groups. The difference is that for uni-directional scenarios only one modality is available at the test stage, whereas in bi-directional scenarios, two modalities are required.

2.1. Uni-directional multimodal applications

An important concern in deep multimodal research is to map (translate) one modality to another. For example, given an entity in visual (or textual) space, the task is to generate a description of this entity in textual (or visual) space according to the content. For some tasks, these mapping processes are uni-directional, *i.e.* either from image to text or from text to image.

2.1.1. Image-to-text tasks

Image captioning is a task that generates a sentence description for an image and requires recognizing important objects and their attributes, then inferring their correlations within the image [23]. After capturing these correlations, the captioner yields a syntactically correct and semantically relevant sentence. To understand the visual content, images are fed into convolutional neural networks to learn hierarchical features, which constitutes the feature encoding process. The produced hierarchical features are transformed into sequential models (*e.g.* RNN, LSTM) to generate the corresponding descriptions. Subsequently, the evaluation module produces description difference as the feedback signals to update the performance of each block. Deep neural networks are commonly used in image captioning. For other methods, including retrieval- and template-based methods, we recommend the existing surveys [23–25]. In the following sections, we will examine the methods widely used to improve image captioning performance, including evolutionary algorithm [26], generative adversarial networks [22,27,28], reinforcement learning [17–19], memory networks [29–31], and attention mechanisms [32–35].

Image captioning is an open-ended research question. It is still difficult to evaluate the performance of captioning, which should be diverse, creative, and human-like [36]. Currently, the metrics for evaluating the performance of image captioning include *BLEU* (Bilingual Evaluation Understanding), *ROUGE* (Recall-Oriented

Understudy for Gisting Evaluation), *METEOR* (Metric for Evaluation of Translation with Explicit ORdering), *CIDEr* (Consensus-based Image Description Evaluation), and *SPICE* (Semantic Propositional Image Captioning Evaluation). It is hard to access the generated captions by linguists. Thus it is necessary to further study an evaluation indicator which is more in line with human judgments and is flexible to new pathological cases [36]. Furthermore, image captioning systems might suffer from dataset bias issue. The trained captioner overfits to the common objects in seen context (*e.g.* book and desk), but it would be challenging for the captioner to generalize to the same objects in unseen contexts (*e.g.* book and tree).

According to captioning principles, researchers focus on specific caption generation tasks, such as image tagging [37], visual region captioning [38], and object captioning [39]. Analogously, these tasks are also highly dependent on the regional image patch and sentences/phrases organization. The specific correlations between the features of objects (or regions) in one image and the word-level (or phrase-level) embeddings are explored instead of global dependence of the holistic visual and textual features.

2.1.2. Text-to-image tasks

Compared to generating a sentence for a given image, generating a realistic and plausible image from a sentence is even more challenging. Namely, it is difficult to capture semantic cues from a highly abstract text, especially when the text is used to describe complex scenarios as found in the MS-COCO dataset [40,41]. Text-to-image generation is such a kind of task which maps from textual modality to visual modality.

Text-to-image generation requires synthesized images to be photo-realistic and semantically consistent (*i.e.* preserving specific object sketches and semantic textures described in text data). Generally, this requirement is closely related to the following two aspects: the heterogeneity gap [10] and the semantic gap [9,42]. The first addresses the gap between the high-level concepts of text descriptions and the pixel-level values of an image, while the second exists between synthetic images and real images.

The above issues in text-to-image application are exactly what generative models attempt to address, through methods such as Variational Auto-Encoders (VAE) [43], auto-regressive models [44] and Generative Adversarial Networks (GANs) [8,22]. Recently, various new ideas and network architectures have been proposed to improve image generation performance. One example is to generate a semantic layout as intermediate information from text data to bridge the heterogeneity gap in image and text [45–47]. Some works focus on the network structure design for feature learning. For image synthesis, novel derivative architectures from GANs [48] have been explored in hierarchically nested adversarial networks [49], perceptual pyramid adversarial networks [50], iterative stacked networks [51,52], attentional generative networks [53,54], cycle-consistent adversarial networks [11,13] and symmetrical distillation networks [42].

Image generation is a promising multimodal application and has many applicable scenarios such as photo editing or multimedia data creation. Thereby, this task has attracted lots of attention. However, there are two main limitations to be explored further. Similar with image captioning, the first limit is regarding the evaluation metrics. Currently, Inception Score (IS) [51,52,55], Fréchet Inception Distance (FID) [55], Multi-scale Structural Similarity Index Metrics (MS-SSIM) [56,57], and Visual-semantic Similarity (VS) [49] are used to evaluate generation quality. These metrics pay attention to generated image resolution and image diversity. However, performance is still far from human perception. Another limitation is that, while generation models work well and achieve promising results on single category object datasets like Caltech-UCSD CUB [58] and Oxford-102 Flower [58], existing methods are still far from promising on complex dataset like MS-COCO

where one image contains more objects and is described by a complex sentence.

To compensate for these limitations, word-level attention [53], hierarchical text-to-image mapping [46] and memory networks [59] have been explored. In the future, one direction may be to make use of the Capsule idea proposed by Hinton [60] since capsules are designed to capture the concepts of objects [48].

2.2. Bi-directional multimodal applications

As for bi-directional applications, features from visual modality are translated to textual modality and vice versa. Representative bi-directional applications are cross-modal retrieval and visual question answering (VQA) where image and text are projected into a common space to explore their semantic correlations.

2.2.1. Cross-modal retrieval

Single-modal and cross-modal retrieval have been researched for decades [61]. Different from single-modal retrieval, cross-modal retrieval returns the most relevant image (text) when given a query text (image). As for performance evaluation, there are two important aspects: retrieval accuracy and retrieval efficiency.

For the first, it is desirable to explore semantic correlations across an image and text features. To meet this requirement, the aforementioned heterogeneity gap and the semantic gap are the challenges to deal with. Some novel techniques that have been proposed are as follows: attention mechanisms and memory networks are employed to align relevant features between image and text [62–65]; Bi-directional sequential models (*e.g.* Bi-LSTM [66]) are used to explore spatial-semantic correlations [1,62]; Graph-based embedding and graph regularization are utilized to keep semantic order in text feature extraction process [67,68]; Information theory is applied to reduce the heterogeneity gap in cross-modal hasing [219]; Adversarial learning strategies and GANs are used to estimate common feature distributions in cross-modal retrieval [69–71]; Metric learning strategies are explored, which consider inter-modality semantic similarity and intra-modality neighborhood constraints [72–74].

For the second, recent hashing methods have been explored [2,75–84] owing to the computation and storage advantages of binary code. Essentially, methods such as attention mechanisms [78] and adversarial learning [81,82,85] are applied for learning compact hash codes with different lengths. However, the problems should be considered when one employs hashing methods for cross-modal retrieval are feature quantization and non-differential binary code optimization. Some methods, such as self-supervised learning [82] and continuation [85], are explored to address these two issues. Recently, Yao et al. [84] introduce an efficient discrete optimal scheme for binary code learning in which a hash codes matrix is construct. Focusing on the feature quantization, Wang et al. [83] introduce a hashing code learning algorithm in which the binary codes are generated without relaxation so that the large quantization and non-differential problems are avoided. Analogously, a straightforward discrete hashing code optimization strategy is proposed, more importantly, in an unsupervised way [86].

Although much attention has been paid to cross-modal retrieval, there still exists room for performance improvement (see Fig. 8–9). For example, to employ graph-based methods to construct semantic information within two modalities, more context information such as objects link relationships are adopted for more effective semantic graph construction [61].

2.2.2. Visual question answering

Visual question answering (VQA) is a challenging task in which an image and a question are given, then a correct answer is
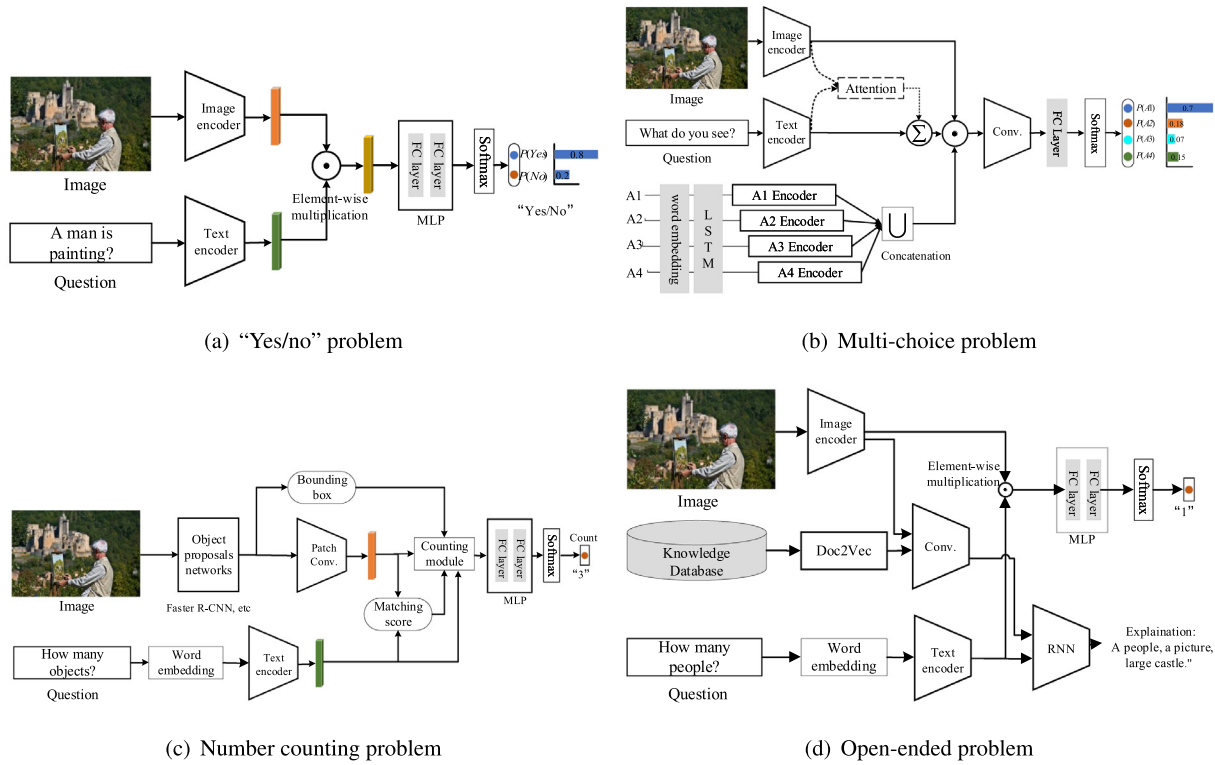
(a) "Yes/no" problem

(b) Multi-choice problem

(c) Number counting problem

(d) Open-ended problem

Fig. 2. Common types of visual question answering. "Yes/No" problem and multi-choice problem can be regarded as a classification problem, while number counting problem and open-ended problem can be viewed as a caption generation problem.

inferred according to visual content and syntactic principle. We summarize four types of VQA [87] in Fig. 2. VQA can be categorized into image question answering and video question answering. In this paper, we target recent advances in image question answering. Since VQA was proposed, it has received increasing attention in recent years. For example, there are some training datasets [88] built for this task, and some network training tips and tricks are presented in work [89].

To infer correct answers, VQA systems need to understand the semantics and intent of the questions completely, and also should be able to locate and link the relevant image regions with the linguistic information in the questions. VQA applications present twofold difficulties: feature fusion and reasoning rationality. Thus, VQA more closely reflects the difficulty of multimodal content understanding, which makes VQA applications more difficult than other multimodal applications. Compared to other applications, VQA has various and unknown questions as inputs. Specific details (*e.g.* activity of a person) in the image should be identified along with the undetermined questions. Moreover, the rationality of question answering is based on high-level knowledge and advanced reasoning capability of deep models. As for performance assessment, answers on open-ended questions are difficult to evaluate compared to the other three types in Fig. 2 where the answer typically is selected from specific options, or the answer contains only a few words [88].

As summarized in Fig. 2, the research on VQA includes: free-form open-ended questions [90], where the answer could be words, phrases, and even complete sentences; object counting questions [91] where the answer is counting the number of objects in one image; multi-choice questions [32] and Yes/No binary problems [92]. In principle, the type of multi-choice and Yes/No can be viewed as classification problems, where deep models infer the candidate with maximum probability as the correct answer. These

two types are associated with different answer vocabularies and are solved by training a multi-class classifier. In contrast, object counting and free-form open-ended questions can be viewed as generation problems [88] because the answers are not fixed, only the ones related to visual content and question details.

Compared to other three mentioned multimodal applications, VQA is more complex and more open-ended. Although much attention has been paid to visual question answering research, there still exist several challenges in this field. One is related to accuracy. Some keywords in question might have been neglected and some visual content might remain unrecognized or misclassified. Because of this, a VQA system might give inaccurate even wrong answers. Another is related to diversity and completeness of the predicted answer, which is especially crucial for free-form open-ended problems, as the output answers should be as complete as possible to explain the given question, and not limited to a specific domain or restricted language forms [88]. The third one is the VQA datasets, which should have been less biased. For the existing available datasets, questions that require the use of the image content are often relatively easy to answer. However, harder questions, such as those beginning with "Why", are comparatively rare and difficult to answer since it needs to more reasoning [93]. Therefore, the biased question type impairs the evaluation for VQA algorithms. For recommendations, a larger but less biased VQA dataset is necessary.

3. Challenges for deep multimodal learning

Typically, domain-specific neural networks process different modalities to obtain individual monomodal representations and are further embedded or aggregated as multimodal features. Importantly, it is still difficult to fully understand how multimodal

features are used to perform the aforementioned tasks well. Taking text-to-image generation as an example, we can imagine two questions: First, how can we organize two types of data into a unified framework to extract their features? Second, how can we make sure that the generated image has the same content as the sentence described?.

These two kinds of questions are highly relevant to the heterogeneity gap and the semantic gap in deep multimodal learning. We illustrate the heterogeneity gap and the semantic gap in Fig. 3. Recently, much effort has gone into addressing these two challenges. These efforts are categorized into two directions: towards minimizing the heterogeneity gap and towards preserving semantic correlation.

3.1. Heterogeneity gap minimization

On the one hand, although complementary information from multiple modalities is beneficial for multimodal content understanding, their very different statistical properties can impair the learning of this complementary. For example, an image comprises 3-channel RGB pixel values, whereas the symbolic text consists of words with different lengths. Meanwhile, image and text data contain different ways of conveying semantic information. Usually, the text has more abstract semantics than image while the content of the image is easier and more straightforward to understand than text.

On the other hand, neural networks are comprised by successive linear layers and non-linear activation functions. The neurons in each layer have different receptive fields so that the systems have various learning capacities. Usually, the last layer of neural networks are used as a way for representing data. Due to the diverse structures of neural networks, the data representations are in various abstractions. Usually, image features are extracted from hierarchical networks and text features are from sequential networks. Naturally, these features are distributed inconsistently so that they are not directly comparable, which leads to the heterogeneity gap. Both the modality data and the network itself contribute to this discrepancy.

Therefore, to correlate features among different modalities, it is necessary to construct a common space for these multimodal features to narrow the heterogeneity gap. In general, there are two strategies to narrow the heterogeneity gap. One direction is from the viewpoint of deep multimodal structures and another is from the viewpoint of feature learning algorithms.

Auto-encoders and generative adversarial networks (GANs) [22] are two important structures for representing the multimodal structure. We will introduce both of them in the following sections. Generative adversarial networks learn features to bridge image data and text data. For example, GANs are commonly applied to generate images according to their descriptive sentences. This idea is then developed into several variants, such as StackGAN [51], HDGAN [49], and AttnGAN [53]. Auto-encoders are used to correlate multimodal features based on feature encoding and feature reconstruction. For example, Gu et al. [94,95] use cross-reconstruction method to preserve multimodal semantic similarity where image (text) features are reconstructed to text (image) features.

In addition, much effort has gone into minimizing the gaps in uni-modal representations. For instance, sequential neural networks (*e.g.* RNNs) are employed to extract multi-granularity including character-level, word-level, phrase-level and sentence-level text features [58,96–98]. Graph-based approaches have been introduced to explore the semantic relationship in text feature learning [68,99].

Regarding the goal of reducing the heterogeneity gap, uni-modal representations are projected into a common latent space under joint or coordinated constraints. Joint representations combine uni-modal features into the same space, while coordinated representations process uni-modal features separately but with certain similarity and structure constraints [20].

3.2. Semantic correlation preserving

Preserving semantic similarity is challenging. On the one hand, the differences between high-level semantic concepts (*i.e.* features) and low-level values (*e.g.* image pixel) result in a semantic gap among intra-modality embeddings. On the other hand, uni-modal visual and textual representations make it difficult to capture complex correlations across modalities in multimodal learning.

As images and text are used to describe the same content, they should share similar patterns to some extent. Therefore, using several mapping functions, uni-modal representations are projected into the common latent space using individual neural networks (see Fig. 1). However, these embedded multimodal features cannot reflect the complex correlations of different modalities because the
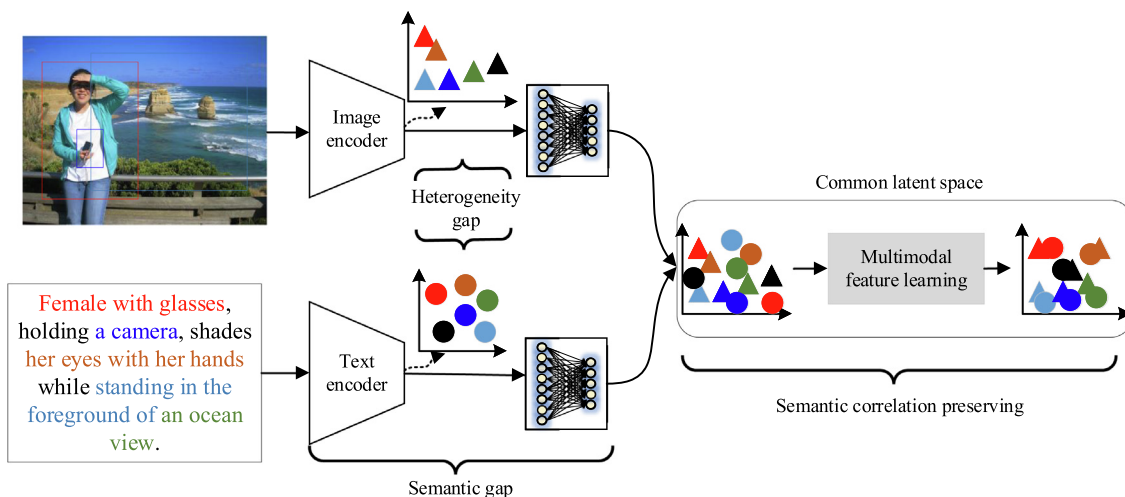

Fig. 3. A conceptual illustration of two challenges. We use different shapes to denote different modalities; the circle represents text feature distributions, and the triangle represents image feature distributions. Different shapes with the same color mean that they are semantically similar in content. Apart from the "semantic gap" [9] which is commonly mitigated in monomodal deep visual tasks such as image classification, the key for deep multimodal content understanding also lies in mitigating the "heterogeneity gap" [10], *i.e.* reducing the inter-modality gap and exploring the semant.ic correlations.

individual networks are not pre-trained and the feature projection are untargeted. The neural networks mainly act as a way of normalizing the monomodal features into a common latent space. Therefore, preserving the semantic correlations between projected features of similar image-text pairs is another challenge in deep multimodal research. More specifically, if an image and text are similar in content, the semantic similarity of features in common latent space should be preserved, otherwise, the similarity should be minimized.

To preserve the semantic correlations, one must measure the similarity between multimodal features where joint representation learning and coordinated representation learning can be adopted. Joint presentation learning is more suitable for the scenarios where all modalities are available during testing stage, such as in visual question answering. For other situations where only one modality is available for testing, such as cross-modal retrieval and image captioning, coordinated representation learning is a better option.

Generally, feature vectors from two modalities can be concatenated directly in joint representation learning; Then the concatenated features are used to make classification or are fed into a neural network (*e.g.* RNN) for prediction (*e.g.* producing an answer). Simple feature concatenation is a linear operation and less effective, advanced pooling-based methods such as compact bilinear pooling [100,101] are introduced to connect the semantically relevant multimodal features. Neural networks are also alternative for exploring more corrections on the joint representations. For example, Wang et al. [102] introduce a multimodal transformer for disentangling the contextual and spatial information so that a unified common latent space for image and text is construct. Similarly, auto-encoders, as unsupervised structures, are used in several multimodal tasks like cross-modal retrieval [74] and image captioning [103]. The learning capacity of the encoder and decoder are enhanced by improving on the structure of the sub-networks by stacking attention [104,105], paralleling LSTM [106,103], and ensembling CNN [107]. Different sub-networks have their own parameters. Thereby, auto-encoders would have more chances to learn comprehensive features.

The key point for coordinated representation learning is to design optimal constraint functions. For example, computing inner product or cosine similarity between two cross-modal features is a simple way to constrain dot-to-dot correlations. Canonical correlation analysis [71,108–110] is commonly used to maximize semantic correlations between vectors; For better performance and stability improvement, metric learning methods such as bidirectional objective functions [111,112,72] are utilized. However, mining useful samples and selecting appropriate margin settings remain empirical in metric learning [73]. Regarding these limits in metric learning, some new methods, such as adversarial learning [70,82,69] and KL-divergence [73], are introduced for these demerits. Instead of selecting three-tuple samples, these alternative methods consider the whole feature distributions in a common latent space. In addition, attention mechanisms [33,34,113] and reinforcement learning [114,115] are popularly employed to align relevant features between modalities.

To address the above-mentioned challenges, several new ideas, including methods for feature extraction, structures of deep networks, and approaches for multimodal feature learning, have been proposed in recent years. The advances from these ideas are introduced in the following sections.

4. Recent advances in deep multimodal feature learning

Regarding the aforementioned challenges, exploring content understanding between image and text has attracted sustained

attention and lots of remarkable progresses have been made. In general, these advances are mainly from a viewpoint of network structure and a viewpoint of feature extraction/enhancement. To this end, combining the natural process pipeline of multimodal research (see Fig. 1), we categorize these research ideas into three groups: deep multimodal structures presented in Section 4.1, multimodal feature extraction approaches introduced in Section 4.2, and common latent space learning described in Section 4.3. Deep multimodal structures indicate the basic framework in community; Multimodal feature extraction is the prerequisite which supports the following similarity exploring; Common latent space learning is the last but a critical procedure to make the multimodal features comparable. For a general overview for these aspects in multimodal applications, we make a chart for the representative methods in Fig. 4.

4.1. Deep multimodal structures

Deep multimodal structures are the fundamental frameworks to support different deep networks for exploring visual-textual semantics. These frameworks, to some extent, have critical influences for the following feature learning steps (*i.e.* feature extraction or common latent space learning). To understand the semantics between images and text, deep multimodal structures usually involve computer vision and natural language processing (NLP) field [173]. For instance, raw images are processed by hierarchical networks such as CNNs and raw input text can be encoded by sequential networks such as RNN, LSTM [98], and GRU [173]. During the past years, a variety of related methods have blossomed and accelerated the performance of multimodal learning directly in multimodal applications, as shown in Fig. 4.

Deep multimodal structures include generative models, discriminative models. Generative models implicitly or explicitly represent data distributions measured by a joint probability $P(X, Y)$, where both raw data X and ground-truth labels Y are available in supervised scenarios. Discriminative models learn classification boundaries between two different distributions indicated by conditional probability $P(Y|X)$. Recent representative network structures for multimodal feature learning are auto-encoders and generative adversarial networks. There are some novel works to improve the performance of multimodal research based on these two basic structures (see Fig. 4).

4.1.1. Auto-encoders

The main idea of auto-encoder for multimodal learning is to first encode data from a source modality as hidden representations and then to use a decoder to generate features (or data) for the target modality. Thus, it is commonly used in dimensionality reduction for bi-directional applications where two modalities are available at the test stage. For this structure, reconstruction loss is the constraint for training encoder and decoder to well capture the semantic correlations between image and text features. For clarity, we identify three ways for correlation learning using auto-encoders in Fig. 5. For instance, as shown in Fig. 5a, the input images and text are processed separately with non-shared encoder and decoder, after which these hidden representations from the encoder are coordinated through a constraint such as Euclidean distance [116]. The coordinated methods can be replaced by joint methods in Fig. 5b where image and text features are projected into a common space with a shared multilayer perceptron (MLP). Subsequently, the joint representation is reconstructed back to the original raw data [124]. Alternatively, feature correlations are captured by cross reconstruction with similarity constraints (*e.g.* a similarity matrix is used as supervisory information [126,95]) between hidden features. The idea of constraining sample similarity is also incorporated with GANs into a cycle-consistent forma-
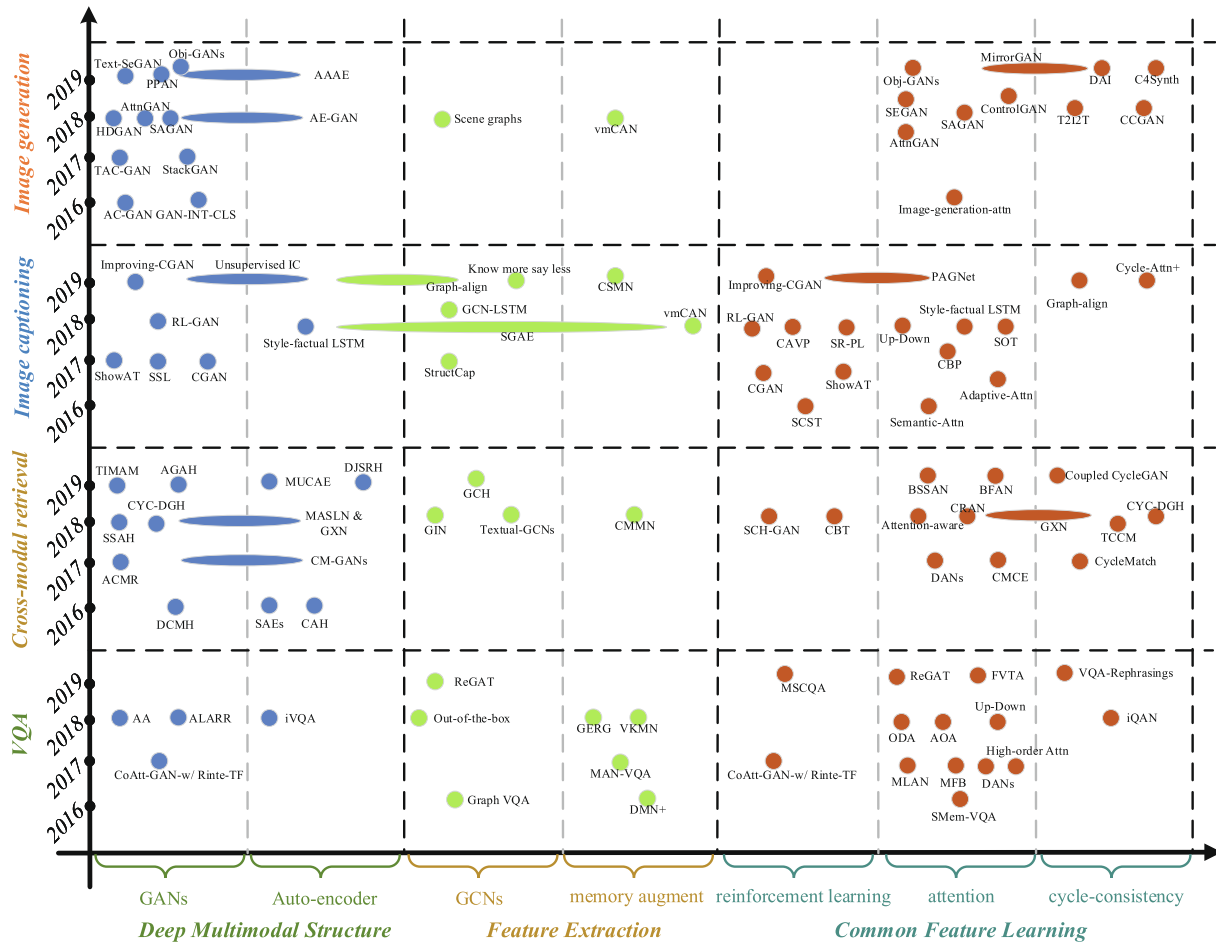
Fig. 4. Representative approaches for multi-modal content understanding. We categorize these new ideas and trends in the perspective of deep multimodal structure, feature extraction and common feature learning, which are applied into different applications: Text-SeGAN [7], PPAN [50], MUCAE [116], AAAE [117], HDGAN [49], AttnGAN [53], SAGAN [55], AE-GAN [118], TAC-GAN [119], StackGAN [51], AC-GAN [57], GAN-INT-CLS [8], Unsupervised IC [120], Improving-CGAN [121], RL-GAN [122], ShowAT [123], SSL [28], CGAN [27], SSAH [82], CYC-DGH [12], MASLN [124], GXN [94], ACMR [70], CM-GANs [125], DCMH [126], TIMAM [127], AGAH [128], DJSRH [129], SAEs [130], CAH [95], AA [131], ALARR [132], iVQA [133], CoAtt-GAN-w/Rinte-TF [134], Scene graphs [45], vmCAN [135], Graph-align [136], Know more say less [137], GCN-LSTM [15], SGAE [16], StructCap [138], GCH [139], GIN [140], Textual-GCNs [141], CSMN [31], CMMN [64], ReGAT [142], Out-of-the-box [143], Graph VQA [144], GERG [145], VKMN [146], MAN-VQA [147], DMN+ [148], MSCQA [115], SCH-GAN [81], CBT [149], SCST [17], CAVP [18], SR-PL [19], SMem-VQA [150], ODA [151], AOA [152], Up–Down [32], Attention-aware [78], BSSAN [62], CRAN [149], CBP [36], SOT [5], PAGNet [34], MirrorGAN [153], DAI [154], T2I2T [11], CCGAN [13], C4Synth [155], Cycle-Attn+ [156], Coupled CycleGAN [157], TCCM [158], CycleMatch [159], VQA-Rephrasings [160], iQAN [161], MLAN [162], MFB [163], DANs [164], High-order Attn [165], FVTA [3], CMCE [166], BFAN [167], Semantic-Attn [168], Adaptive-Attn [169], Obj-GANs [170], Image-generation-attn [43], ControlGAN [171], SEGAN [172].

tion for cross-modal retrieval like GXN [94] and CYC-DGH [12] (see Fig. 4).

The neural networks contain in the encoder-decoder framework can be modality specific. For image data, the commonly used neural networks are CNN while sequential networks like LSTM are most often used for text data. When applied for multimodal learning, the decoder (*e.g.* LSTM) constructs hidden representations of one modality in another modality. The goal is not to reduce reconstruction error but to minimize the output likelihood estimation. Therefore, most works focus on the decoding since it is a process to project the less meaningful vectorial representations to meaningful outputs in target modality. Under this idea, several extensions have been introduced. The main difference among these algorithms lies in the structure of the decoder. For example, "stack and parallel LSTM" [106,103] is to parallelize more parameters of LSTMs to capture more context information. Similar ideas can be found in "CNN ensemble learning" [107]. Instead of grabbing more information by stacking and paralleling, "Attention-LSTM" [106,174] combines attention technique into LSTM to highlight most relevant correlations, which would be more targeted. An adversarial training strategy is employed into the decoder to make

all the representations discriminative for semantics but indiscriminative for modalities so that intra-modal semantic consistency is effectively enhanced [124]. Considering the fixed structure in the decoder like RNN might limit the performance, Wang et al. [26] introduce evolutionary algorithm to adaptively generate neural network structures in the decoder.

4.1.2. Generative adversarial networks

As depicted in Fig. 4, adversarial learning from generative adversarial networks [22] has been employed into applications including image captioning [28,121,123], cross-modal retrieval [70,82,69,124,81,78] and image generation [7,49,51,52,50,53], but has been less popular in VQA tasks. GANs combine generative sub-models and discriminative sub-models into a unified framework in which two components are trained in an adversary manner.

Different from auto-encoders, GANs can cope with the scenarios where there are some missing data. To accurately explore the correlations between two modalities, multimodal research works involving GANs have been focusing on the whole network structure and its two components: *generator* and *discriminator*.
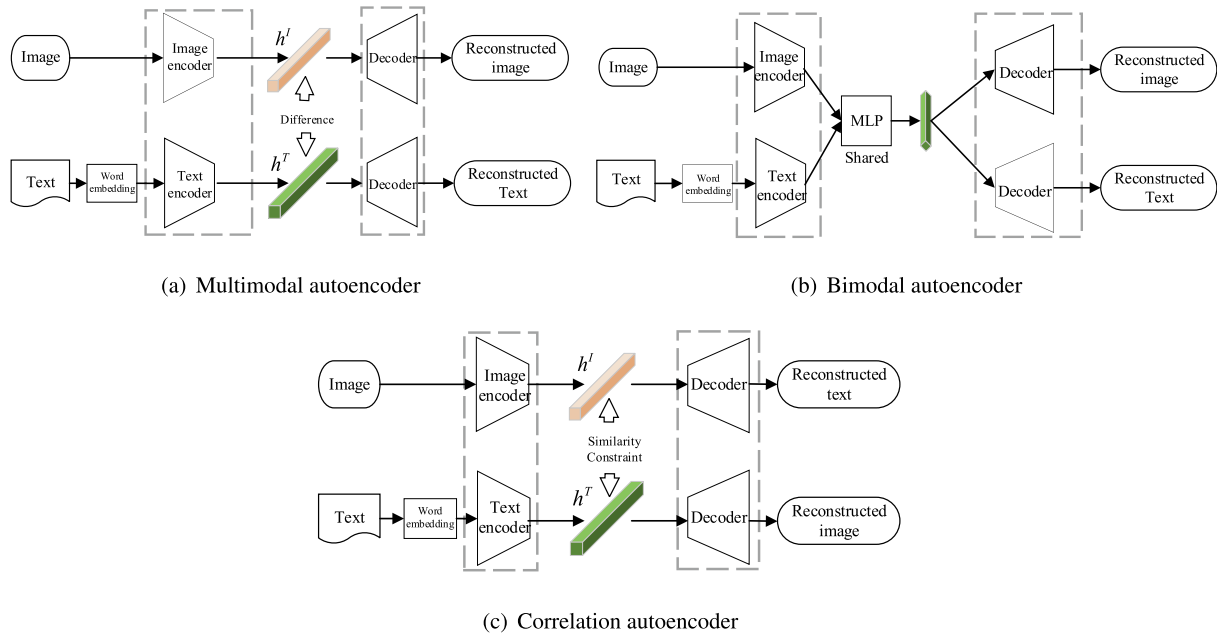
(a) Multimodal autoencoder

(b) Bimodal autoencoder

(c) Correlation autoencoder

Fig. 5. Convolutional autoencoder used for deep multimodal learning. The branch for image feature learning can adopt hierarchical networks such as CNNs; the branch for text feature learning can capture the dependency relations in a sentence by sequential models such as RNN and LSTM. Usually, a reconstruction loss function is used to optimize network training.

For the generator which also can be viewed as an encoder, an attention mechanism is often used to capture the important key points and align cross-modal features such as AttnGAN [53] and Attention-aware methods [78]. Sometimes, Gaussian noise is concatenated with the generator's input vector to improve the diversity of generated samples and avoid model collapse, such as the conditioning augmentation block in StackGAN [51]. To improve its capacity for learning hierarchical features, a generator can be organized into different nested structures such as hierarchical-nested [49] and hierarchical-pyramid [50], both of them can capture multi-level semantics.

The discriminator, which usually performs binary classification, attempts to discriminate the ground-truth labels from the outputs of the generator. Some recent ideas are proposed to improve the discrimination of GANs. Originally, discriminator in the first work [22] just needs to classify different distributions into "*True*" or "*False*" [8]. However, discriminator can also make a class label classification where a label classifier is added on the top of discriminator [57,119]. Apart from the label classification, a semantic classifier is designed to further predict semantic relevances between a synthesized image and a ground-truth image for text-to-image generation [7]. Only focusing on the paired samples leads to relatively-weak robustness. Therefore, the unmatched image-text samples can be fed into a discriminator (*e.g.* GAN-INT-CLS [8] and AACR [71]) so that the discriminator would have a more powerful discriminative capability.

According to previous work [48], the whole structure of GANs in multimodal research are categorized into direct methods [8,119,57], hierarchical methods [49,50] and iterative methods [51–53]. Contrary to direct methods, hierarchical methods divide raw data in one modality (*e.g.* image) into different parts such as a "style" and "structure" stage, thereby, each part is learned separately. Alternatively, iterative methods separate the training into a "coarse–fine" process where details of the results from a previous generator are refined. Besides, cycle-consistency from cycleGAN [175] is introduced for unsupervised image translation where a self-consistency (reconstruction) loss tries to retain the patterns

of input data after a cycle of feature transformation. This network structure is then applied into tasks like image generation [13,11] and cross-modal retrieval [94,12] to learn semantic correlation in an unsupervised way.

Preserving semantic correlations between two modalities is to reduce the difference of inconsistently distributed features from each modality. Adversarial learning keeps pace with this goal. In recent years, adversarial learning is widely used to design algorithms for deep multimodal learning [69,70,78,82,81,124]. For these algorithms, there are no classifiers for binary classification. Instead, two sub-networks are trained with the constraints of competitive loss functions.

As the dominant popularity of adversarial learning, some works are performed by combining auto-encoders and GANs in which the encoder in auto-encoders and the generator in GANs share the same sub-network [117,118,120,124,94,125] (see Fig. 4). For example, in the first work about unsupervised image captioning [120], the core idea of GANs is used to generate meaningful text features from scratch of text corpus and cross-reconstruction is performed between synthesized text features and true image features.

4.2. Multimodal feature extraction

Deep multimodal structures support the following learning process. Thereby, feature extraction is closer for exploring visual-textual content relations, which is the prerequisite to discriminate the complementarity and redundancy of multiple modalities. It is well-known that image features and text features from different deep models have distinct distributions although they convey the same semantic concept, which results in a heterogeneity gap. In this section, we introduce several effective multimodal feature extraction methods for addressing the heterogeneity gap. In general, these methods focus on (1) learning the structural dependency information to reasoning capability of deep neural networks and (2) storing more information for semantic correlation learning during model execution. Moreover, (3) feature align-

ment schemes using attention mechanism are also widely explored for preserving semantic correlations.

4.2.1. Graph embeddings with graph convolutional networks

Words in a sentence or objects within an image have some dependency relationships, and graph-based visual relationship modelling is beneficial for the characteristic [35]. Graph Convolutional Networks (GCNs) are alternative neural networks designed to capture this dependency information. Compared to standard neural networks such as CNNs and RNNs, GCNs would build a graph structure which models a set of objects (nodes) and their dependency relationships (edges) in an image or sentence, embed this graph into a vectorial representation, which is subsequently integrated seamlessly into the follow-up steps for processing. Graph representations reflect the complexity of sentence structure and are applied to natural language processing such as text classification [176]. For deep multimodal learning, GCNs receive increasing attention and have achieved breakthrough performance on several applications, including cross-modal retrieval [14], image captioning [15,16,35,138], and VQA [144,143,142]. Recent reviews [177,178] have reported comprehensive introductions to GCNs. However, we focus on recent ideas and processes in deep multimodal learning.

Graph convolutional networks in multimodal learning can be employed in text feature extraction [14,144].

[35,65] and image feature extraction [15,16,138]. Among these methods, GCNs capture semantic relevances of intra-modality according to the neighborhood structure. GCNs also capture correlations between two modalities according to supervisory information. Note that vector representations from graph convolutional networks are fed into subsequent networks (*e.g.* "encoder-decoder" framework) for further learning.

GCNs aim at determining the attributes of objects and subsequently characterize their relationships. On the one hand, GCNs can be applied in a singular modality to reduce the intra-modality gap. For instance, Yu et al. [14] introduce a "GCN + CNN" architecture for text feature learning and cross-modal semantic correlation modeling. In their work, Word2Vec and *k*-nearest neighbor algorithm are utilized to construct semantic graphs on text features. GCNs are also explored for image feature extractions, such as in image captioning [15,16,138]. In previous work [138], a tree structure embedding scheme is proposed for semantic graph construction. Specifically, input images are parsed into several key entities and their relations are organized into a visual parsing tree (VP-Tree). This process can be regarded as an encoder. The VP-Tree is transformed into an attention module to participate in each state of LSTM-based decoder. VP-Tree based graph construction is somewhat in a unified way. Alternative methods are introduced to construct more fine-graded semantic graphs [15,16]. Specifically, object detectors (*e.g.* Faster-RCNN [179]) and visual relationship detectors (*e.g.* MOTIFS [180]) are used to get image regions and spatial relations, semantic graphs and spatial graphs are constructed based on the detected regions and relations, respectively. Afterwards, GCNs extract visual representations based on the built semantic graphs and spatial graphs.

Graph convolutional networks are also introduced to mitigate the inter-modality gap between image and text [144,143]. Take the work [143] for VQA as an example, an image is parsed into different objects, scenes, and actions. Also, a corresponding question is parsed and processed to obtain its question embeddings and entity embeddings. These embedded vectors of image and question are concatenated into node embeddings then fed into graph convolutional networks for semantic correlation learning. Finally, the output activations from graph convolutional networks are fed into sequential networks to predict answers.

As an alternative method, graph convolutional networks are worthy more exploration for correlations between two modalities. Moreover, there exist two limitations in graph convolutional networks. On the one hand, graph construction process is overall time- and space-consuming; On the other hand, the accuracy of output activations from graph convolutional networks mostly relies on supervisory information to construct an adjacency matrix by training, which are more suitable for structured data, so flexible graph embeddings for image and/or text remains an open problem.

4.2.2. Memory-augmented networks

To enable deep networks to understand multimodal content and have better reasoning capability for various tasks, a solution may be the mentioned GCNs. Moreover, another solution that has gained attention recently is memory-augmented networks. Directly, when much information in mini-batch even the whole dataset is stored in a memory bank, such networks have greater capacity to memorize correlations.

In conventional neural networks like RNNs for sequential data learning, the dependency relations between samples are captured by the internal memory of recurrent operations. However, these recurrent operations might be inefficient in understanding and reasoning overextended contexts or complex images. For instance, most captioning models are equipped with RNN-based encoders, which predict a word at every time step based only on the current input and hidden states used as implicit summaries of previous histories. However, RNNs and their variants often fail to capture long-term dependencies [31]. For this limitation, memory networks [30] are introduced to augment the memory primarily used for text question–answering [87]. Memory networks improve understanding of both image and text, and then "remember" temporally distant information.

Memory-augmented networks can be regarded as recurrent neural networks with explicit attention methods that select certain parts of the information to store in their memory slots. As reported in Fig. 4, memory-augmented networks are used in cross-modal retrieval [64], image captioning [31,98,181,6] and VQA [146–148,145,150]. We illustrate memory-augmented networks for multimodal learning in Fig. 6. A memory block, which acts as a compressor, encodes the input sequence into its memory slots. The memory slots are a kind of external memory to support learning; the row vectors in each slot are accessed and updated at each time-step. During training, a network such as LSTM or GRU, which acts as a memory controller, refers to these memory slots to compute reading weights (see Fig. 6). According to the weights, the essential information is obtained to predict the output sequence. Meanwhile, the controller computes writing weights to update values in memory slots for the next time-step of the training [182].

Memory networks can expand the "memory" of networks thus store more information. The network performance relates to the memory slots' initialization strategy and the stored information. For this aspect, memory networks have been combined with other techniques like attention mechanisms [105] to further improve its feature learning capability. For example, Xiong et al. [148] explore the impact of different initialization strategies to demonstrate that initializations from the outputs of pre-trained networks have better performance. This was verified in works [150] where output features from image patches are stored into memory slots of spatial memory networks for VQA. Thereby, generated answers are updated based on gathering evidence from the accessed regions in memory slots. Similarly, Ma et al. [147] adopt LSTM to obtain text features of each sentence and store into memory slots. Then memory-augmented networks are utilized to determine the importance of concatenated visual and text features over the whole training data. Further considering both two modalities, a visual knowledge memory networks is introduced in which
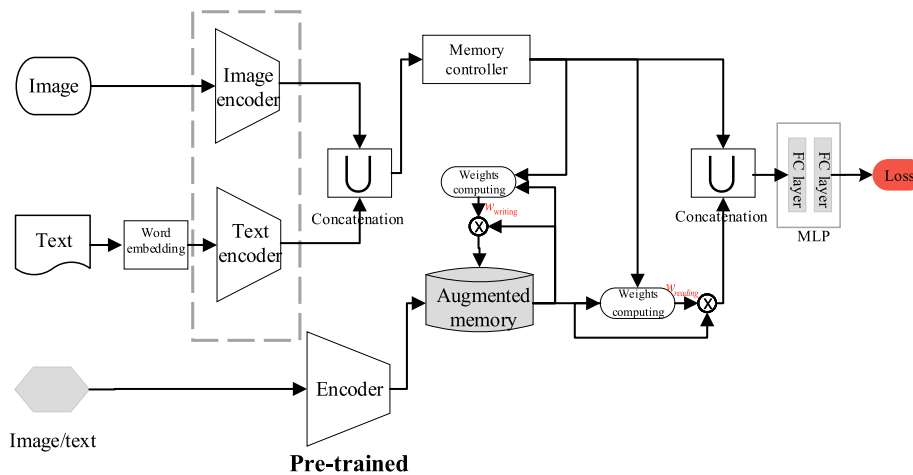
Fig. 6. The paradigm of memory-augmented networks in multimodal learning. Augmented memory stores the features of all long-term sequences processed by a pre-trained classifier. The stored vectors can be selected (read) by the memory controller according to the reading weights. Meanwhile, these stored vectors are updated according to the writing weights at each time-step.

memory slots store key-value vectors computed from images, query questions and a knowledge base [146]; Instead of storing the actual output features, Song et al. [64] adopt memory slots to store a prototype concept representation from pre-trained concept classifiers, which is inspired from the process of human memory.

Memory-augmented networks improve the performance of deep multimodal content understanding by offering more information to select. However, this technique is less popular in image generation, image captioning and cross-modal retrieval than VQA (see in Fig. 4). A possible reason for this is that in cross-modal retrieval, memory-augmented networks might require extra time when memory controllers determine when to write or read from the external memory blocks. It will hurt overall retrieval efficiency.

4.2.3. Attention mechanism

As mentioned in Section 3.2, one challenge for deep multimodal learning is to preserve semantic correlations among multimodal features. Regarding this challenge for content understanding, feature alignment plays a crucial role. Image and text features are first processed by deep neural networks under a certain structure like auto-encoders. Naturally, the final output global features include some irrelevant or noisy background information, which is not optimal for performing multimodal tasks.

Attention are commonly used mechanisms to tackle this issue and have been widely incorporated into various multimodal tasks, such as visual question answering [151,152,183–188].

[98], image captioning [5,34,35,15,189,190], and cross-modal retrieval [62,78,65]. In principle, the attention mechanisms compute different weights (or importances) according to relevances between two global (or local) multimodal features and assign different importances to these features. Thereby, the networks will be more targeted at the sub-components of the source modality–regions of an image or words of a sentence. To further explore the relevances between two modalities, the attention mechanisms are adopted on multi-level feature vectors [150,189], employed in a hierarchical scheme [188,191], and incorporated with graph networks for modelling semantic relationships [35].

To elaborate on the current ideas and trends of attention algorithms, we categorize this popular mechanism into different types. According to objective computing vectors, we categorize the current attention algorithms into four types: visual attention, textual attention, co-attention, and self-attention. Their diagrams are introduced in Fig. 7. We further categorize the attention algorithms

into single-hop and multiple-hop (i.e. stacked attention) according to the iterations of importance calculation.

4.2.3.1. Visual attention. As shown in Fig. 7a, visual attention schemes are used in scenarios where text features (e.g. from a query question) are used as context to compute their co-relevance with image features, and then the relationships are used to construct a normalized weight matrix. Subsequently, this matrix is applied to original image features to derive text-guided image features using element-wise multiplication operation (linear operation). The weighted image features have been aligned by the correlation information between image and text. Finally, these aligned multimodal features are utilized for prediction or classification. This idea is common in multimodal feature learning [5,32] and has been incorporated to get different text-guided features. For example, Anderson et al. [32] employ embedded question features to highlight the most relevant image region features in visual question answering. The predicted answers are more accurately related to the question type and image content. Visual attention is widely used to learn features from two modalities.

4.2.3.2. Textual attention. Compared to visual attention, the textual attention approach is relatively less adopted. As shown in Fig. 7b, it has an opposite computing direction [149,192,193]. The computed weights are based on text features to obtain relevances for different image regions or objects. According to the work [87], the reason why textual attention is necessary is that text features from the multimodal models often lack detailed information for a given image. Meanwhile, the application of textual attention is less dominant as it is harder to capture semantic relevances between abstract text data and image data. Moreover, image data has always contained more irrelevant content for similar text. In other words, the text might describe only some parts within an image.

4.2.3.3. Co-attention. As shown in Fig. 7c, co-attention algorithm is viewed as a combination of visual attention and textual attention, which is an option to explore the inter-modality correlations [62,186,188].

[105,164,190,194,166,65]. Co-attention is a particular case of joint feature embedding in which image and text features are usually treated symmetrically. Co-attention in a bi-directional way is beneficial for spatial-semantic learning. As an example, Nguyen et al. [186] introduce a dense symmetric co-attention method to
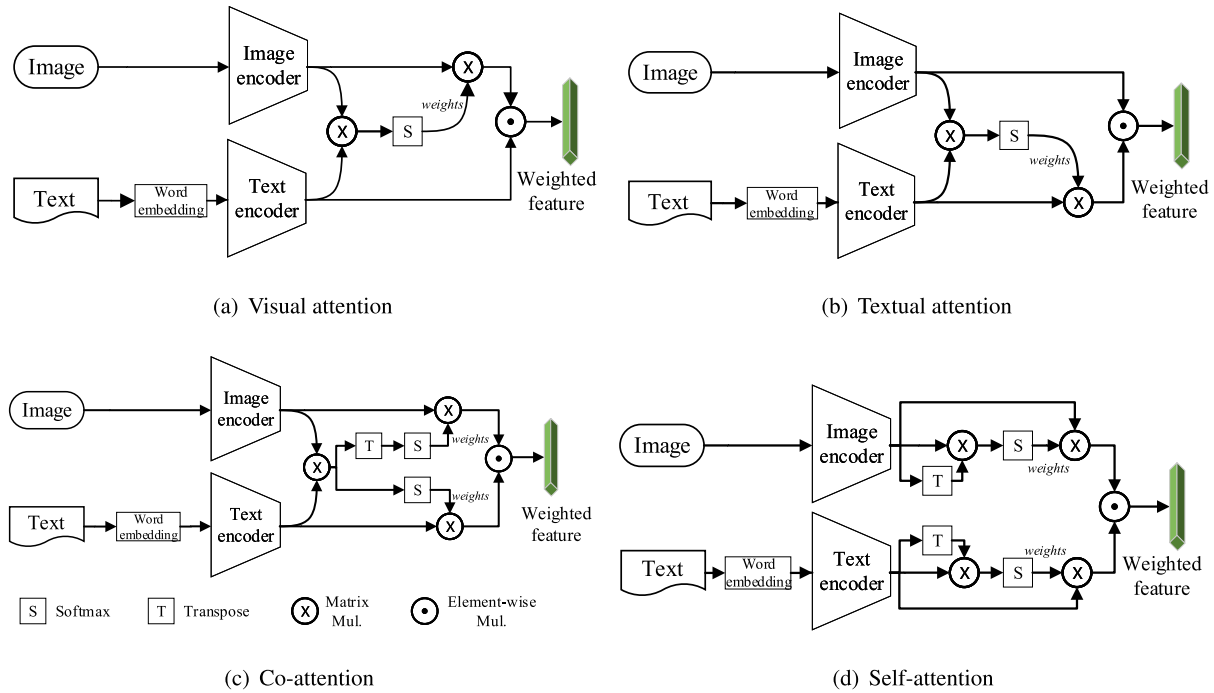
(a) Visual attention

(b) Textual attention

(c) Co-attention

(d) Self-attention

Fig. 7. Diagram for different types of attention mechanisms used in deep multimodal learning.

improve the fusion performance of image and text representations for VQA. In their method, features are sampled densely to fully consider each interaction between any word in question and any image region. Similarly, Huang et al. [62] also employ this idea to capture underlying fine-granularity correlations for image-text matching, and Ding et al. [195] capture similar fine-granularity by two types of visual-attention for image captioning. Meanwhile, several other works explore different formations of co-attention. For instance, Lu et al. [188] explore co-attention learning in a hierarchical fashion where parallel co-attention and alternating co-attention are employed in VQA. The method aims at modelling the question hierarchically at three levels to capture information from different granularities. Integrating image feature with hierarchical text features may vary dramatically so that the complex correlations are not fully captured [163]. Therefore, Yu et al. [163,184] develop the co-attention mechanism into a generalized Multi-modal Factorized High-order pooling (MFH) block in an asymmetrical way. Thereby, higher-order correlations of multi-modal feature achieve more discriminative image-question representation and further result in significant improvement on the VQA performance.

4.2.3.4. Self-attention. Compared to the co-attention algorithm, self-attention, which considers the intra-modality relations, is less popular in deep multimodal learning. As intra-modality relation is complementary to the inter-modality relation, its exploration is considered improving the feature learning capability of deep networks. For example, in the VQA task, the correct answers are not only based on their associated words/phrases but can also be inferred from related regions or objects in an image. Based on this observation, a self-attention algorithm is proposed for multimodal learning to enhance the complementary between intra-modality relations and the inter-modality relations [184,78,194,102]. Self-attention has been used in different ways. For example, Gao et al. [194] combine the attentive vectors form self-attention with co-attention using element-wise product. The linear modelling method for inter- and intra-modality information flow lead to less effectiveness since the complex correlations cannot be fully learned. Therefore, more effective strategies are introduced. Yu

et al. [184] integrate text features from the self-attention block by Multimodal Factorized Bilinear (MFB) pooling approach rather than linear method to produce joint features. Differently, Zhang et al. [78] propose a learnable combination scheme in which they employ a self-attention algorithm to extract the image and text features separately. Then these attended and unattended features are trained in an adversary manner.

It is important to note that when these four types of attention mechanisms are applied, they can be used to highlight the relevances between different image region features and word-level, phrase-level or sentence-level text features. These different cases just need region/object proposal networks and sentence parsers. When multi-level attended features are concatenated, the final features are more beneficial for content understanding in multimodal learning.

As for single-hop and multiple-hop (stacked) attention, the difference lies in whether the attention "layer" will be used one or more times. The four mentioned attention algorithms can be applied in a single-hop manner where the relevance weights between image and text features are computed once only. However, for multiple-hop scenarios, the attention algorithm is adopted hierarchically to perform coarse-to-fine feature learning, that is, in a stacked way [65,188,164,150,104,152,105]. For example, Xu et al. [150] introduce two–hop spatial attention learning for VQA. The first hop focuses on the whole and the second one focuses on individual words and produces word-level features. Yang et al. [104] also explore multiple attention layers in VQA thereby the sharper and higher-level attention distributions will contribute refined query features for predicting more relevant answers. Singh et al. [152] achieve marginal improvements using "attention on attention" framework in which the attention module is stacked in parallel and for image and text feature learning. Nevertheless, a stacked architecture has tendency for gradient vanishing [105]. Regarding this, Fan et al. [105] propose stacked latent attention for VQA. Particularly, all spatial configuration information contained in the intermediate reasoning process is retained in a pathway of convolutional layers so that the vanishing gradient problem is tackled.

In summary, to better understand the content in visual and textual modality, attention mechanisms provide a pathway for aligning the multimodal semantic correlations. With different multimodal applications, attention mechanisms (single-hop or multiple-hop) can have different benefits. To this end, we briefly make a comparison for single-hop and multiple-hop with respect to their advantages, disadvantages, and the applicable scenarios in Table 1.

4.3. Common latent space learning

As illustrated in Fig. 1, feature extractors (*e.g.* GCNs) would yield modality-specific presentations. In other words, these features distribute inconsistently and are not directly comparable. To this end, it is necessary to further map these monomodal features into a common latent space with the help of an embedding networks (*e.g.* MLP). Therefore, common latent feature learning has been a critical procedure for exploiting multimodal correlations. In the past years, various constraint and regularization methods have been introduced into multimodal applications (see Fig. 4). In this section, we will include these ideas, such as attention mechanisms, which aim to retain similarities between monomodal image and text features.

According to the taxonomy methods [20], multimodal feature learning algorithms include joint and coordinated methods. The joint feature embedding is formulated as:

$$J = \mathcal{J}(x_1, \ldots, x_n, y_1, \ldots, y_n) \tag{1}$$

while coordinated feature embeddings are represented as:

$$F = \mathcal{F}(x_1, \ldots, x_n) \sim \mathcal{G}(y_1, \ldots, y_n) = G \tag{2}$$

where J refers to the jointly embedded features, F and G denote the coordinated features. $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are n-dimension monomodal feature representations from two modalities (*i.e.* image and text). The mapping functions $\mathcal{J}(\cdot), \mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)$ denote the deep networks to be learned, "$\sim$" indicates that the two monomodal features are separated but are related by some similarity constraints (*e.g.* DCCA [196]).

Table 1
Brief comparisons of two attention categories.

Hop(s)	Advantages	Disadvantages	Applicable scenarios
Single	More straightforward and training effective since the visual-textual interaction occurs a single time	Less focused on complex relations between words. Insufficient to locate words or features on complicated sentences	No explicit constraints for visual attention. Suitable for capturing relations in short sentences as tends to be paid much to the most frequently words.
Multiple	More sophisticated and accurate, especially for complicated sentences. Each iteration provides newly relevant information to discover more fine-grained correlations between image and text.	Less training effective due to re-assigning attention weights multiple times. Sharing structures and parameters leads to attention bias (similar attention weights in all hops). Might suffer from the gradient vanishing problem [105].	Beneficial for multimodal learning involved long sentences. More suitable for sentence embedding in text classification or machine translation tasks. Beneficial for combining with memory networks due to the repeatedly or iteratively information extraction process.

4.3.1. Joint feature embedding

In deep multimodal learning, joint feature embedding is a straightforward way in which monomodal features are combined into the same presentation. The fused features are used to make a classification in cross-modal retrieval [63]. It also can be used for performing sentence generation in VQA [32].

In early studies, some basic methods are employed for joint feature embedding such as feature summation, feature concatenation [51–53], and element-wise inner product [148,186], the resultant features are then fed into a multi-layer perceptron to predict similarity scores. These approaches construct a common latent space for features from different modalities but cannot preserve their similarities while fully understanding the multimodal content. Alternatively, more complicated bilinear pooling methods [100] are introduced into multimodal research. For instance, Multimodal Compact Bilinear (MCB) pooling is introduced [197] to perform visual question answering and visual grounding. However, the performance of MCB is based on a higher-dimensional space. Regarding this demerit, Multimodal Low-rank Bilinear pooling [198,187] and Multimodal Factorized Bilinear pooling [163] are proposed to overcome the high computational complexity when learning joint feature. Moreover, Hedi et al. [199] introduce a tensor-based Tucker decomposition strategy, MUTAN, to efficiently parameterized bilinear interactions between visual and textual representations so that the model complexity is controlled and the model size is tractable. In general, to train an optimal model to understand semantic correlations, classification-based objective functions [119,57] and regression-based objective functions [51,53] are commonly adopted.

Bilinear pooling methods are based on outer products to explore correlations of multimodal features. Alternatively, neural networks are used for jointly embedding features since its learnable ability for modelling the complicated interactions between image and text. For instance, auto-encoder methods, as shown in Fig. 5b, are used to project image and text features with a shared multi-layer perceptron (MLP). The similar multimodal transformer introduced in [102] constructs a unified joint space for image and text. In addition, sequential networks are also adopted for the latent space construction. Take visual question answering as an example, based on the widely-used "encoder-decoder" framework, image features extracted from the encoder are fed into the decoder (*i.e.* RNNs [106]), and finally combined with text features to predict correct answers [88,103,121,147]. There are several ways to combine features. Image features can be viewed as the first "word" and concatenate all real word embeddings from the sentences. Alternatively, image features can be concatenated with each word embedding then fed them into RNNs for likelihood estimation. Considering the gradient vanishing in RNNs, CNNs are used to explore complicated relations between features [112,200]. For example, convolutional kernels are initialized under the guidance of text features. Then, these text-guided kernels operate on extracted image features to maintain semantic correlations [200].

The attention mechanisms in Section 4.2.3 can also be regarded as a kind of joint feature alignment method and are widely used for common latent space learning (see Fig. 4). Theoretically, these feature alignment schemes aim at finding relationships and correspondences between instances from visual and textual modalities [20,87]. In particular, the mentioned co-attention mechanism is a case of joint feature embedding in which image and text features are usually treated symmetrically [63]. The attended multimodal features are beneficial for understanding the inter-modality correlations. Attention mechanisms for common latent space learning can be applied in different formations, including bi-directional [186,62], hierarchical [188,163,184], and stacked [188,164,150,104]. More importantly, the metrics for measuring similarity are crucial in attentive importance estimation. For

example, the importance estimation bu simple linear operation [188] may fail to capture the complex correlations between visual and textual modality while the Multi-modal Factorized High-order pooling (MFH) method can learn higher-order semantic correlations and achieve marginal performance.

To sum up, joint feature embedding methods are basic and straightforward ways to allow learning interactions and perform inference over multimodal features. Thus, joint feature embedding methods are more suitable for situations where image and text raw data are available during inference, and joint feature embedding methods can be expanded into situations when more than two modalities are present. However, for content understanding among inconsistently distributed features, as reported in previous work [88], there is potential for improvement in the embedding space.

4.3.2. Coordinated feature embedding

Instead of embedding features jointly into a common space, an alternative method is to embed them separately but with some constraints on features according to their similarity (*i.e.* coordinated embedding). For example, the above-noted reconstruction loss in auto-encoders can be used to constraining multimodal feature learning in the common space [68,74]. Using traditional canonical correlation analysis [108], as an alternative, the correlations between two kinds of features can be measured and then maintained [110,196]. To explore semantic correlation in a coordinated way, generally, there are two commonly used categories: classification-based methods and verification-based methods.

For classification-based methods when class label information is available, these projected image and text features in the common latent space are used for label prediction [70,72,82,194]. Cross-entropy loss between the inference labels and the ground-truth labels is computed to optimize the deep networks, see Fig. 1, via the back-propagation algorithm. For classification-based methods, class labels or instance labels are needed. They map each image feature and text feature into a common space and guarantee the semantic correlations between two types of features. Classification-based methods mainly concern the image-text pair with the same class label. For the image and unmatched text (vice versa), classification-based methods have less constraints.

Different from classification-based methods, the commonly used verification-based methods can constrain both the matched image-text pairs (similar or have the same class labels) and unmatched pairs (dissimilar or have the different class labels). Verification-based methods are based on metric learning among multimodal features. Given similar/dissimilar supervisory information between image and text, these projected multimodal features should be mapped based on their corresponding similar/dissimilar information. In principle, the goal of the deep networks is to make similar image-text features close to each other while mapped dissimilar image-text features further away from each other. Verification-based methods include pair-wise constraint and triplet constraint, both of which form different objective functions.

For pair-wise constraint, the key point lies in constructing an inference function to infer similarity of features. For example, Cao et al. [95] use matrix multiplication to compute the pair-wise similarity. In other examples, Cao et al. [80,82,79] construct a Bayesian network, rather than a simple linear operation, to preserve the similarity relationship of image-text pairs. In addition, triplet constraint is also widely used for building the common latent space. Typically, bi-directional triplet loss function is applied to learn feature relevances between two modalities [63,19,77,72,78,149,70]. Inter-modality correlations are learned well when triplet samples interchange within image and text. However, a complete deep multimodal model should also be able to capture intra-modality similarity, which is a complementary

part for inter-modality correlation. Therefore, several works consider combining intra-modal triplet loss in feature learning in which all triplet samples are from the same modality (*i.e.* image or text data) [63,77,183,194].

These classification-based and verification-based approaches are widely used for deep multimodal learning. Although the verification-based methods overcome some limits of classification-based methods, they still face some disadvantages such as the negative samples and margin selection, which inherit from metric learning [73]. Recently, new ideas on coordinated feature embedding methods have combined adversarial learning, reinforcement learning, cycle-consistent constraints to pursue high performance. Several representative approaches are shown in Fig. 4.

4.3.2.1. Combined with adversarial learning.
Classification- and verification-based methods focus on the semantic relevance between similar/dissimilar pairs. Adversarial learning focuses on the overall distributions of two different modalities instead of just focusing on each pair. The primary idea in GANs is to determine whether the input image-text pairs are matched [45,46,48,71].

In new ideas of adversarial learning for multimodal learning, an implicit generator and a discriminator are designed with competitively goals (*i.e.* the generator enforces similar image-text features be close while the discriminator separates them into two clusters). Therefore, the aim of adversarial learning is not to make a binary classification ("True/False"), but to train two groups of objective functions adversarially, it will enable the deep networks with powerful ability and focus on holistic features. For example, in recent works [70,82,124,81,201], a modality classifier is constructed to distinguish the visual modality and textual modality according to the input multimodal features. This classifier is trained adversarially with other sub-networks which constrain similar image-text feature to be close. Furthermore, adversarial learning is also combined with a self-attention mechanism to obtain attended regions and unattended regions. This idea is imposed on the formation of a bi-directional triplet loss to perform cross-modal retrieval [78].

4.3.2.2. Combined with reinforcement learning.
Reinforcement learning has been incorporated into deep network structures (*e.g.* encoder-decoder framework) for image captioning [17–19,114,103,202,203], visual question answering [115,134] and cross-modal retrieval [81]. Because reinforcement learning avoids exposure bias [19,18] and non-differentiable metric issue [17,19]. It is adopted to promote multimodal correlation modeling. To incorporate reinforcement learning, its basic components are defined (*i.e.* "agent", "environment", "action", "state" and "reward"). Usually, the deep models such as CNNs or RNNs are viewed as the "agent", which interacts with an external "environment" (*i.e.* text features and image features), while the "action" is the prediction probabilities or words of the deep models, which influence the internal "state" of the deep models (*i.e.* the weights and bias). The "agent" observes a "reward" to motivate the training process. The "reward" is an evaluation value through measuring the difference between the predictive distribution and ground-truth distribution. For example, the "reward" in image captioning is computed from the CIDEr (Consensus-based Image Description Evaluation) score of a generated sentence and a true descriptive sentence. The "reward" plays an important role for adjusting the goal of predictive distribution towards the ground-truth distribution.

Reinforcement learning is commonly used in generative models in which image patch features or word-level features are regarded as sequential inputs. When incorporating reinforcement learning into deep multimodal learning, it is important to define an algorithm to compute the expected gradients and the "reward" as a reasonable optimization goal.

For the first term, the expected gradients, REINFORCE algorithm [204] is widely used as a policy gradient method to compute gradients, then to update these "states" via back-propagation algorithms [17–19]. [81,94,134,202,203,220]. For the second term, there are several different alternatives. For example, the difference, evaluated by the popular metric CIDEr, between the generated captions and true description sentences in image captioning is used as a "reward" [18,202,17,103,203]. Instead of measuring the difference, sample similarity is more straightforward to track. As an example, visual-textual similarity is used as "reward" after deep networks are trained under the ranking loss (*e.g.* a triplet loss) [19,114,81]. Note that, the design of triplet ranking loss function is diverse, such as in a bi-directional manner [114] or based on inter-modal triplet sampling [19]. For example, Wang et al. [115] devise a three-stage "reward" for three different training scenarios according to similarity scores and binary indicators.

4.3.2.3. Combined with cycle-consistent constraint. Class label information or relevance information between image and text is crucial for understanding semantic content. However, this supervisory information sometimes is not available for training deep networks. In this case, a cycle-consistent constraint is employed for unpaired image-text inputs. The basic idea of a cycle-consistent constraint is dual learning in which a closed translation loop is used to regularize the training process. This self-consistency constraint allows a predictive distribution to retain most of the correlations of the original distribution to improve the stability of network training. In principle, a cycle-consistent constraint includes a forward cycle and backward cycle. The former relies on the loss function $F(G(X)) \approx X$, while the latter relies on another loss function $G(F(Y)) \approx Y$. In these two functions, $F(\cdot)$ is a mapping process from Y to X and $G(\cdot)$ is a reversed process from X to Y. Cycle-consistency has been used on several tasks such as cross-modal retrieval [157,94,12,149,124], image generation [11,13,94] and visual question answering [161].

Cycle-consistency is an unsupervised learning method for exploring semantic correlation in the common latent space. To ensure predictive distribution and retain as many correlations as possible, the aforementioned forward and backward cycle-consistent objective functions are necessary. The feature reconstruction loss function acts as the cycle-consistency objective function. For example, Gorti et al. [11] utilize the cross-entropy loss between generated words and the actual words as cycle-consistency loss values to optimize the process text-to-image-to-text translation. For cross-modal retrieval tasks, Li et al. [157] adopt Euclidean distance between predictive features and reconstructed features as the cycle-consistency loss where the two cycle loss functions interact in a coupled manner to produce reliable codes.

Currently, the application of cycle-consistent constraints for deep multimodal learning can be categorized as structure-oriented and task-oriented. The former group focuses on making several components in a whole network into a close loop in which output of each component is used as the input for another component. Differently, task-oriented group concerns to exploit the complementary relations between tasks. Thus, there are two independent tasks (*e.g.* VQA and VQG) in the close loop.

For structure-oriented groups, the cycle-consistent idea is combined with some popular deep networks, such as GANs, to make some specific combinations. In these methods, image features are projected as "text features" and then reconstructed back to itself. Currently, the combination with GANs is a popular option since paired correspondence of modalities can be learned in the absence of a certain modality (*i.e.* via generation). For example, Wu et al. [12] plug a cycle-consistent constraint into feature projection

between image and text. The inversed feature-learning process is constrained using the least absolute deviation. The whole process is just to learn a couple of generative hash functions through the cycle-consistent adversarial learning. Regarding this limit, Li et al. [157] devise an outer-cycle (for feature representation) and an inner-cycle (for hash code learning) constraint to combine GANs for cross-modal retrieval. Thereby, the objects for which the cycle-consistency loss constrains have increased. Moreover, in their method, the discriminator should distinguish if the input feature is original (viewed as *True*) or generated (viewed as *False*).

For task-oriented groups, cycle-consistency is adopted into dual tasks. In cycle-consistency, we use an inverse process (*task A* to *task B* to *task A*) to improve the results. When a whole network performs both tasks well, it indicates that the learned features between the tasks have captured the semantic correlations of two modalities. For example, Li et al. [161] combine visual question answering (VQA) and visual question generation (VQG), in which the predicted answer is more accurate through combining image content to predict the question. In the end, the complementary relations between questions and answers lead to performance gains. For text-image translation, a captioning network is used to produce a caption which corresponds to a generated image from a sentence using GANs [11]. The distances between the ground truth sentences and the generated captions are exploited to improve the network further. The inverse translation is beneficial for understanding text context and the synthesized images. To sum up, there are still some questions to be explored in task-oriented ideas, such as the model parameter sharing scheme, and these implicit problems make the model more difficult to train and might encounter gradient vanishing problems, the task-oriented cycle-consistent constraint is applied to unify multi-task applications into a whole framework and attracts more research attention.

5. Conclusion and future directions

In this survey, we have conducted a review of recent ideas and trends in deep multimodal learning (image and text) including popular structures and algorithms. We analyzed two major challenges in deep multimodal learning for which these popular structures and algorithms target. Specifically, popular structures including auto-encoders, generative adversarial nets and their variants perform uni-directional and bi-directional multimodal tasks. Based on these popular structures, we introduced current ideas about multimodal feature extraction and common latent feature learning which plays crucial roles for better content understanding within a visual and textual modality. For multimodal feature extraction, we introduced graph convolutional networks and memory-augmented networks. For common latent feature learning, we presented the joint and coordinated feature embedding methods including the recently proposed objective functions.

The aforementioned ideas have made some progress on various multimodal tasks. For example, for cross-modal retrieval, we presented the achieved progress and state-of-the-art of recent methods on the Flickr30K [205] and the MS-COCO [41] datasets in Fig. 8. For hashing retrieval methods, we presented the achievement on the MIRFlickr25k [206] and the NUS-WIDE [207] datasets in Fig. 9. Meanwhile, other tasks, (*i.e.* image generation, image captioning and VQA) are listed in Tables in the Appendix. As we can see from these statistics, the progress is notable in recall rate (*i.e.* the fraction of queries for which the top K nearest neighbors are retrieved correctly) and mAP (*i.e.* the mean of the average precision scores for each query) in cross-modal retrieval. However, there is still room for improvement in the current limitations of multimodal content understanding.
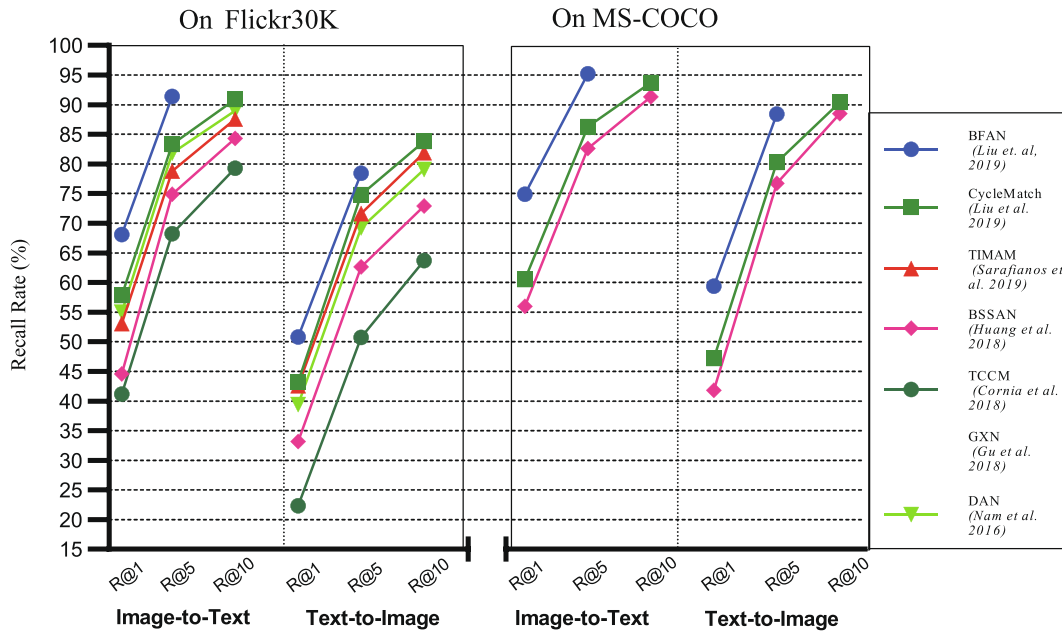
Fig. 8. The achieved progress of cross-modal retrieval on the Flickr30K [205] and the MS-COCO [41] datasets.
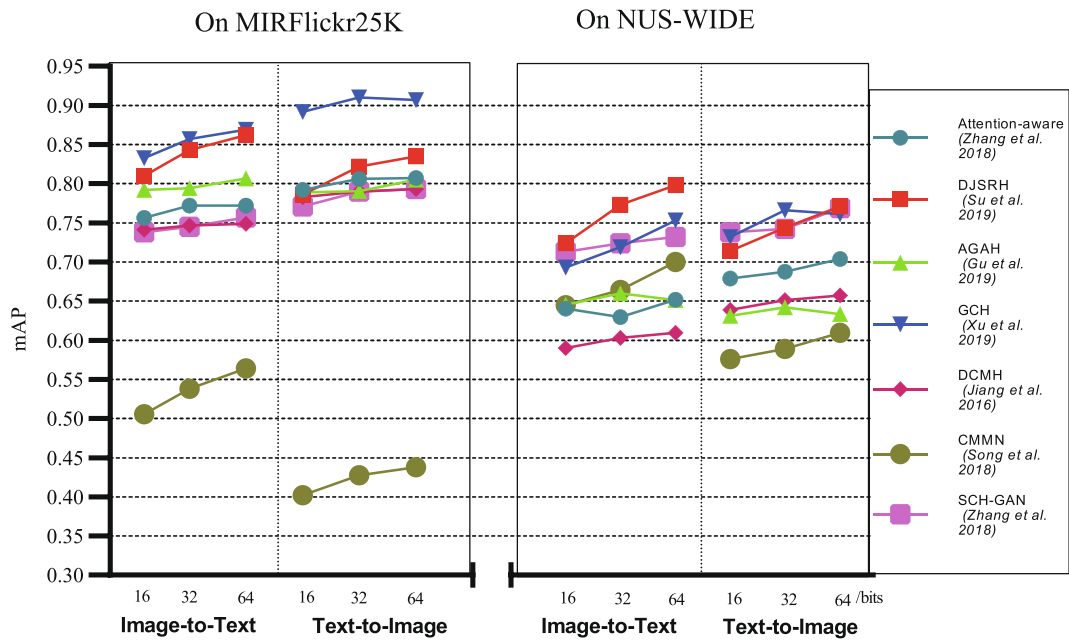

Fig. 9. The achieved progress of cross-modal hash retrieval on the MIRFlickr25k [206] and the NUS-WIDE [207] datasets. Hashing methods have higher retrieval efficiency using the binary hash codes.

For future directions, multi-task integrated networks might be helpful and complementary for content understanding as different applications capture semantic correlations from different perspectives. Effort has been made on integrating image captioning and cross-modal retrieval tasks, image captioning and visual question answering, image generation and image retrieval. Nevertheless, these combined applications are only based on two modalities. Considering the complementary characteristic among modalities (conveying the same concept), it might be promising to fuse more than two modalities to enable machines to understand their

semantic correlations. Undoubtedly, it will be more challenging for aligning these diverse data. There are some explorations in this direction. Aytar et al. [208] present a deep cross-modal convolutional network to learn a representation that is aligned across three modalities: sound, image, and text. The network is only trained with "image + text" and "image + sound" pairs. He et al. [209] construct a new benchmark for cross-media retrieval in which image, text, video, and audio are included. It is the first benchmark with 4 media types for fine-grained cross-media retrieval. However, this direction is still far from satisfactory.

Deep neural networks, including convolutional neural networks and recurrent neural networks, have made the monomodal feature extraction and multimodal feature learning end-to-end trainable. The representations from multimodal data can be automatically learned effectively, without the need of requiring expert knowledge in a certain field, which makes the process of understanding of multimodal content more intelligent. However, the disadvantages of deep networks for multimodal learning are obvious. It is well-known that the deep networks depend on a massive of multiple-modality data to train, but the less biased datasets are not so common. More importantly, deep networks for multimodal learning lacks of interpretability to some extent. Although joint embedding or coordinated embeddings methods can be utilized, it still needs to figure out which modality (or its features) plays more important role for the final content understanding.

From a technical viewpoint, graph-based networks are an important direction for future research. Currently, graph representation is constructed within intra-modality to present the semantic relations, which can be further explored in the future. Meanwhile, the exploration of graph-based networks can be deepened by examining scalability and heterogeneity [178]. Finally, generation-based tasks such as image generation and image captioning are effective for unsupervised learning, since numerous labeled training data can be generated from the deep networks. Combined with reinforcement learning, the image generation process is more controllable. For example, some fine-grained attributes including texture, shape and color can be specified during deep network training. Once it understands the content between modalities, the deep network, like an agent, will synthesize photo-realistic images, which can be used in other applications.

CRediT authorship contribution statement

Wei Chen: Investigation, Methodology, Resources, Writing - original draft. **Weiping Wang:** Formal analysis, Writing - review & editing. **Li Liu:** Formal analysis, Writing - review & editing. **Michael S. Lew:** Funding acquisition, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Appendix A

The progress of other multimodal application in recent years including image captioning in Table 2, image generation in Table 3, and visual question answering in Table 4. We list the topic-relevant representative works in several datasets. The evaluation metric are these common-used in each task. These approaches make sense for improving performance.

Table 2
Performance of image captioning on the MS-COCO dataset [41].

Methods	Year	CIDEr		ROUGE-L		METEOR		BLEU1		BLEU2		BLEU3		BLEU4		KeyNotes
		c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	
StructCap [138]	2016	94.3	95.8	53.5	68.2	25.0	33.5	73.1	90.0	56.5	81.5	42.4	70.9	31.6	59.9	Attention + VP-tree for visual features
Semantic-Attn [168]	2016	95.3	94.8	53.4	68.4	25.1	34.0	72.4	90.7	55.8	82.2	42.3	71.7	32.0	60.7	Visual attention
CGAN [27]	2017	102.0	–	52.7	–	24.8	–	–	–	–	–	39.3	–	29.9	–	CGAN + Reinforcement learning
Adaptive-Attn [169]	2017	104.2	105.9	55.0	70.5	26.4	35.9	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	Adaptive visual attention
SCST [17]	2017	114.7	116.7	56.3	70.7	27.0	35.5	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	Reinforcement learning
RL-GAN [122]	2018	–	–	–	–	24.3	–	71.6	–	51.8	–	37.1	–	26.5	–	GAN + Reinforcement learning
SOT [5]	2018	106.1	108.7	55.5	69.9	25.9	34.2	78.7	93.5	61.5	85.5	46.5	74.8	34.5	63.3	Visual attention
SR-PL [19]	2018	117.1	–	57.0	–	27.4	–	80.1	–	63.1	–	48.0	–	35.8	–	Reinforcement learning
Up–Down [17]	2018	117.9	120.5	57.1	72.4	27.6	36.7	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	Top-down and bottom-up attention
CAVP [18]	2018	121.6	123.8	58.2	73.1	28.1	37.0	80.1	94.9	64.7	88.8	50.0	79.7	37.9	69.0	Reinforcement learning
RFNet [107]	2018	122.9	125.1	58.2	73.1	28.2	37.2	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	Visual attention
iVQA [133]	2018	168.2	–	46.6	–	20.1	–	42.1	–	32.0	–	25.3	–	20.5	–	Reinforcement learning
UnsupervisedIC [120]	2019	54.9	–	43.1	–	17.9	–	58.9	–	40.3	–	27.0	–	19.6	–	VAE + GAN (unsupervised)
Graph-align [136]	2019	69.5	–	–	–	20.9	–	67.1	–	47.8	–	32.3	–	21.5	–	VAE + Graph embed + cycle (unpaired)
Self-critical [202]	2019	112.6	115.3	56.1	70.4	26.9	35.4	77.6	93.1	61.3	86.1	46.5	76.0	34.8	64.6	Reinforcement learning
PAGNet [34]	2019	118.6	–	58.6	–	30.4	–	83.2	–	62.8	–	46.3	–	40.8	–	Attention + Reinforcement learning
RL-CGAN [121]	2019	123.1	124.3	59.0	74.4	28.7	38.2	81.9	95.6	66.3	90.1	51.7	81.7	39.6	71.5	GAN + Reinforcement learning
SGAE [16]	2019	123.8	126.5	58.6	73.6	28.2	37.2	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	VAE + graph embedding

Table 3
Performance of image generation.

Methods	Year	Caltech-UCSD Birds200		Oxford Flowers102		MS-COCO		KeyNotes
		IS	FID	IS	FID	IS	FID	
GAN-INT-CLS [8]	2016	2.88 ± .04	–	2.66 ± .03	–	7.88 ± 0.07	–	Vanilla GAN for image generation
TAC-GAN [119]	2017	–	–	3.45 ± .05	–	–	–	Discriminator learns class information
GAWWN [54]	2017	3.60 ± .07	–	–	–	–	–	Conditional objection location is learned
StackGAN [51]	2017	3.70 ± .04	51.89	3.20 ± .01	55.28	8.45 ± .03	74.05	GAN in a stacked structure
StackGAN++ [52]	2018	4.04 ± .05	15.3	3.26 ± .01	48.68	8.30 ± .10	81.59	GAN in a tree-like structure
HDGAN [49]	2018	4.15 ± .05	–	3.45 ± .07	–	11.86 ± 0.18	–	GAN in a hierarchically-nested structure
AttnGAN [53]	2018	4.36 ± .03	–	–	–	25.89 ± .47	–	Attentional generative network
Scene graphs [45]	2019	–	–	–	–	7.3 ± 0.1	–	Graph convolution for graphs from text
Obj-GANs [170]	2019	–	–	–	–	27.37 ± 0.22	25.85	Attentive generator and object-wise discriminator
vmCAN [135]	2019	–	–	–	–	10.36 ± 0.17	–	Visual-memory method in GAN
AAAE [117]	2019	–	–	–	103.46	–	–	Auto-encoders + GAN for adversarial approximation
Text-SeGAN [7]	2019	–	–	4.03 ± 0.07	–	–	–	Semantic relevance matching in GAN
DAI [154]	2019	3.58 ± 0.05	18.41 ± 1.07	2.90 ± 0.03	37.94 ± 0.39	8.94 ± 0.2	27.07 ± 2.55	Dual inference mechanism disentangled variables
C4Synth [155]	2019	4.07 ± 0.13	–	3.52 ± 0.15	–	–	–	Image generation using multiple captions
PPAN [50]	2019	4.35 ± .05	–	3.53 ± .02	–	–	–	GAN in perceptual pyramid structure
MirrorGAN [153]	2019	4.56 ± 0.05	–	–	–	26.47 ± 0.41	–	Task-oriented cycle consistency + attention
ControlGAN [171]	2019	4.58 ± 0.09	–	–	–	24.06 ± 0.6	–	Attention + region-wise attribute generation
SD-GAN [210]	2019	4.67 ± 0.09	–	–	–	35.69 ± 0.5	–	Disentangling high-/low-level semantics in GAN

‡ To evaluate the identification and diverse of generated image, Inception Score (IS) and Fréchett Inception Distance (FID) are commonly used. For Inception Score, higher is better. For Fréchet Inception Distance, lower is better.

Table 4
Performance of visual question answering on VQA 1.0 dataset [211].

Methods	Year	Open-ended test-std	Open-ended test-dev	MC test-std	MC test-dev	KeyNotes
Smem-VQA [150]	2016	58.24	57.99	–	–	Spatial memory network stores image region features
DMN+ [148]	2016	60.4	60.3	–	–	Improved dynamic memory network for VQA
MLB [198]	2016	65.07	64.89	68.89	–	Low-rank bilinear pooling for similarity learning
MCB [197]	2016	66.5	66.7	70.1	70.2	Multimodal compact bilinear pooling for similarity learning
High-order Attn [165]	2017	–	–	69.3	69.4	Attention mechanisms learn high-order feature correlations
DAN [164]	2017	64.2	64.3	69	69.1	Co-attention networks for multimodal feature learning
MLAN [162]	2017	65.3	65.2	70	70	Multi-level co-attention for feature alignment
SVA [212]	2017	66.1	66	–	–	Visual attention on grid-structured image region feature learning
MFB [163]	2017	66.6	66.9	71.4	71.3	Multi-modal factorized bilinear pooling for similarity learning
MUTAN [199]	2017	67.36	67.42	–	–	Multimodal tucker fusion for similarity learning
Graph VQA [144]	2017	70.42	–	74.37	–	Graph representation for scene and question feature learning
MAN-VQA [147]	2018	64.1	63.8	69.4	69.5	Memory-augmented network for feature learning and matching
QGHC [200]	2018	65.9	65.89	–	–	Question-guided convolution for visual-textual correlations learning
Dual-MFA [213]	2018	66.09	66.01	69.97	70.04	Co-attention for visual-textual feature learning
VKMN [146]	2018	66.1	66	69.1	69.1	Visual knowledge memory network for feature learning
CVA [214]	2018	66.2	65.92	70.41	70.3	Cubic visual attention for object-region feature learning
DCN [186]	2018	67.02	66.89	–	–	Dense co-attention for feature fusion
DRAU [215]	2018	67.16	66.86	–	–	Recurrent co-attention for feature learning
ODA [151]	2018	67.97	67.83	72.23	72.28	Object-difference visual attention to fuse features
ALARR [132]	2018	68.43	68.61	71.28	68.43	Adversarial learning for pair-wise feature discrimination
DF [216]	2019	68.48	68.62	73.05	73.31	Differential network for visual-question feature learning
Relational Encoding [217]	2019	69.3	69.1	–	–	Textual attention for question feature encoding
DCAF [218]	2019	70.0	69.9	–	–	Dense co-attention for feature fusion

References

[1] C.C. Park, Y. Kim, G. Kim, Retrieval of sentence sequences for an image stream via coherence recurrent convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 945–957.

[2] D. Mandal, K.N. Chaudhury, S. Biswas, Generalized semantic preserving hashing for cross-modal retrieval, IEEE Trans. Image Process. 28 (1) (2019) 102–112.

[3] J. Liang, L. Jiang, L. Cao, Y. Kalantidis, L.-J. Li, A.G. Hauptmann, Focal visual-text attention for memex question answering, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2019) 1893–1908.

[4] P. Wang, Q. Wu, C. Shen, A. Dick, A. van den Hengel, Fvqa: Fact-based visual question answering, IEEE Trans. Pattern Anal. Mach. Intell. 40 (10) (2018) 2413–2427.

[5] H. Chen, G. Ding, Z. Lin, S. Zhao, J. Han, Show, observe and tell: Attribute-driven attention model for image captioning., in: IJCAI, 2018, pp. 606–612.

[6] Q. Wu, C. Shen, P. Wang, A. Dick, A. van den Hengel, Image captioning and visual question answering based on attributes and external knowledge, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1367–1381.

[7] M. Cha, Y.L. Gwon, H. Kung, Adversarial learning of semantic relevance in text to image synthesis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3272–3279.

[8] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: 33rd International Conference on Machine Learning, 2016, pp. 1060–1069.

[9] X. Li, T. Uricchio, L. Ballan, M. Bertini, C.G. Snoek, A.D. Bimbo, Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval, ACM Computing Surveys (CSUR) 49 (1) (2016) 14.

[10] K. Wang, Q. Yin, W. Wang, S. Wu, L. Wang, A comprehensive survey on cross-modal retrieval, arXiv preprint arXiv:1607.06215.

[11] S.K. Gorti, J. Ma, Text-to-image-to-text translation using cycle consistent adversarial networks, arXiv preprint arXiv:1808.04538.

[12] L. Wu, Y. Wang, L. Shao, Cycle-consistent deep generative hashing for cross-modal retrieval, IEEE Trans. Image Process. 28 (4) (2018) 1602–1612.

[13] X. Liu, G. Meng, S. Xiang, C. Pan, Semantic image synthesis via conditional cycle-generative adversarial networks, 2018, pp. 988–993.

[14] J. Yu, C. Yang, Z. Qin, Z. Yang, Y. Hu, Y. Liu, Textual relationship modeling for cross-modal information retrieval, CoRR abs/1810.13151.

[15] T. Yao, Y. Pan, Y. Li, T. Mei, Exploring visual relationship for image captioning, in: ECCV, 2018.

[16] X. Yang, K. Tang, H. Zhang, J. Cai, Auto-encoding scene graphs for image captioning, CoRR abs/1812.02378.

[17] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7008–7024.

[18] D. Liu, Z.-J. Zha, H. Zhang, Y. Zhang, F. Wu, Context-aware visual policy network for sequence-level image captioning, arXiv preprint arXiv:1808.05864.

[19] X. Liu, H. Li, J. Shao, D. Chen, X. Wang, Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 338–354.

[20] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443.

[21] D. Ramachandram, G.W. Taylor, Deep multimodal learning: A survey on recent advances and trends, IEEE Signal Process. Mag. 34 (6) (2017) 96–108.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[23] M. Hossain, F. Sohel, M.F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, ACM Computing Surveys (CSUR) 51 (6) (2019) 118.

[24] X. He, D. Li, Deep learning for image-to-text generation: A technical overview, IEEE Signal Process. Mag. 34 (6) (2017) 109–116.

[25] S. Bai, S. An, A survey on automatic image caption generation, Neurocomputing 311 (2018) 291–304.

[26] H. Wang, H. Wang, K. Xu, Evolutionary recurrent neural network for image captioning, Neurocomputing. 401 (2020) 249–256.

[27] B. Dai, S. Fidler, R. Urtasun, D. Lin, Towards diverse and natural image descriptions via a conditional gan, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2970–2979.

[28] R. Shetty, M. Rohrbach, L. Anne Hendricks, M. Fritz, B. Schiele, Speaking the same language: Matching machine to human captions by adversarial training, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4135–4144.

[29] J. Weston, S. Chopra, A. Bordes, Memory networks, arXiv preprint arXiv:1410.3916.

[30] S. Sukhbaatar, J. Weston, R. Fergus, et al., End-to-end memory networks, in: Advances in neural information processing systems, 2015, pp. 2440–2448.

[31] C.C. Park, B. Kim, G. Kim, Towards personalized image captioning via multimodal memory networks, IEEE transactions on pattern analysis and machine intelligence.

[32] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.

[33] W. Wang, Z. Chen, H. Hu, Hierarchical attention network for image captioning, in: Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

[34] L. Song, J. Liu, Q. Buyue, Y. Chen, Connecting language to images: A progressive attention-?-guided network for simultaneous image captioning and language grounding, in: Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

[35] J. Wang, W. Wang, L. Wang, Z. Wang, D.D. Feng, T. Tan, Learning visual relationship and context-aware attention for image captioning, Pattern Recogn. 98 (2020) 107075.

[36] Y. Cui, G. Yang, A. Veit, X. Huang, S. Belongie, Learning to evaluate image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5804–5812.

[37] J. Jin, H. Nakayama, Annotation order matters: Recurrent image annotator for arbitrary length image tagging, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 2452–2457.

[38] Y. Li, W. Ouyang, B. Zhou, K. Wang, X. Wang, Scene graph generation from objects, phrases and region captions, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1261–1270.

[39] P. Anderson, S. Gould, M. Johnson, Partially-supervised image captioning, Adv. Neural Inform. Process. Syst. (2018) 1879–1890.

[40] O.B. El, O. Licht, N. Yosephian, Gilt: Generating images from long text, arXiv preprint arXiv:1901.02404.

[41] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[42] M. Yuan, Y. Peng, Text-to-image synthesis via symmetrical distillation networks, in: 2018 ACM international conference on Multimedia, ACM, 2018, pp. 1407–1415.

[43] E. Mansimov, E. Parisotto, J.L. Ba, R. Salakhutdinov, Generating images from captions with attention, In ICLR.

[44] S. Reed, A. van den Oord, N. Kalchbrenner, S.G. Colmenarejo, Z. Wang, Y. Chen, D. Belov, N. de Freitas, Parallel multiscale autoregressive density estimation, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2912–2921.

[45] J. Johnson, A. Gupta, F.F. Li, Image generation from scene graphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[46] S. Hong, D. Yang, J. Choi, H. Lee, Inferring semantic layout for hierarchical text-to-image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[47] F. Tan, S. Feng, V. Ordonez, Text2scene: Generating compositional scenes from textual descriptions, arxiv.org/abs/1809.01110.

[48] H. He, P.S. Yu, C. Wang, An introduction to image synthesis with generative adversarial nets, arXiv preprint arXiv:1803.04469 (2018).

[49] Z. Zhang, Y. Xie, Y. Lin, Photographic text-to-image synthesis with a hierarchically-nested adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[50] L. Gao, D. Chen, J. Song, X. Xu, D. Zhang, H. Shen, Perceptual pyramid adversarial networks for text-to-image synthesis, in: Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

[51] Z. Han, X. Tao, H. Li, Stackgan, Text to photo-realistic image synthesis with stacked generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5907–5915.

[52] Z. Han, X. Tao, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, Stackgan++: Realistic image synthesis with stacked generative adversarial networks, IEEE Transactions on Pattern Analysis and Machine Intelligence PP (99) (2018) 1–1.

[53] Q.H.H.Z.Z.G.X.H.X.H. Tao Xu, Pengchuan Zhang, Attngan: Fine-grained text to image generation with attentional generative adversarial networks.

[54] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, H. Lee, Learning what and where to draw.

[55] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: International Conference on Machine Learning, 2019, pp. 7354–7363.

[56] J. Snell, K. Ridgeway, R. Liao, B.D. Roads, M.C. Mozer, R.S. Zemel, Learning to generate images with perceptual similarity metrics, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 4277–4281.

[57] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier gans, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2642–2651.

[58] S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 49–58.

[59] S. Zhang, H. Dong, W. Hu, Y. Guo, C. Wu, D. Xie, F. Wu, Text-to-Image Synthesis via Visual-Memory Creative Adversarial Network: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part III, 2018, pp. 417–427. doi:10.1007/978-3-030-00764-5_38.

[60] S. Sabour, N. Frosst, G.E. Hinton, Dynamic routing between capsules, in: NIPS, 2017.

[61] Y. Peng, X. Huang, Y. Zhao, An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2018) 2372–2385.

[62] F. Huang, X. Zhang, Z. Zhao, Z. Li, Bi-directional spatial-semantic attention networks for image-text matching, IEEE Trans. Image Process. 28 (4) (2019) 2008–2020.

[63] S. Wang, Y. Chen, J. Zhuo, Q. Huang, Q. Tian, Joint global and co-attentive representation learning for image-sentence retrieval, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1398–1406.

[64] G. Song, D. Wang, X. Tan, Deep memory network for cross-modal retrieval, IEEE Transactions on Multimedia.

[65] J. Yu, Y. Lu, W. Zhang, Z. Qin, Y. Liu, Y. Hu, Learning cross-modal correlations by exploring inter-word semantics and stacked co-attention, Pattern Recogn. Lett. 130 (2019) 189–198.

[66] A. Graves, S. Fernández, J. Schmidhuber, Bidirectional lstm networks for improved phoneme classification and recognition, in: International Conference on Artificial Neural Networks, Springer, 2005, pp. 799–804.

[67] M. Zhang, H. Zhang, J. Li, L. Wang, Y. Fang, J. Sun, Supervised graph regularization based cross media retrieval with intra and inter-class correlation, J. Vis. Commun. Image Represent. 58 (2019) 1–11.

[68] Y. Wu, S. Wang, Q. Huang, Learning semantic structure-preserved embeddings for cross-modal retrieval, in: ACM Multimedia Conference on Multimedia Conference, ACM, 2018, pp. 825–833.

[69] Y. Peng, J. Qi, Cm-gans: Cross-modal generative adversarial networks for common representation learning, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15 (1) (2019) 22.

[70] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: Proceedings of the 25th ACM international conference on Multimedia, ACM, 2017, pp. 154–162.

[71] Y. Wu, S. Wang, G. Song, Q. Huang, Augmented adversarial training for cross-modal retrieval, IEEE Transactions on Multimedia.

[72] L. Wang, Y. Li, J. Huang, S. Lazebnik, Learning two-branch neural networks for image-text matching tasks, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 394–407.

[73] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 686–701.

[74] Y. Zhan, J. Yu, Z. Yu, R. Zhang, D. Tao, Q. Tian, Comprehensive distance-preserving autoencoders for cross-modal retrieval, in: 2018 ACM Multimedia Conference on Multimedia Conference, ACM, 2018, pp. 1137–1145.

[75] L. Jin, J. Tang, Z. Li, G.-J. Qi, F. Xiao, Deep semantic multimodal hashing network for scalable multimedia retrieval, arXiv preprint arXiv:1901.02662.

[76] X. Zhang, G. Dong, Y. Du, C. Wu, Z. Luo, C. Yang, Collaborative subspace graph hashing for cross-modal retrieval, in: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ACM, 2018, pp. 213–221.

[77] C. Deng, Z. Chen, X. Liu, X. Gao, D. Tao, Triplet-based deep hashing network for cross-modal retrieval, IEEE Trans. Image Process. 27 (8) (2018) 3893–3903.

[78] X. Zhang, H. Lai, J. Feng, Attention-aware deep adversarial hashing for cross-modal retrieval, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 591–606.

[79] Q.-Y. Jiang, W.-J. Li, Deep cross-modal hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3232–3240.

[80] Y. Cao, B. Liu, M. Long, J. Wang, Cross-modal hamming hashing, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 202–218.

[81] J. Zhang, Y. Peng, M. Yuan, Sch-gan: Semi-supervised cross-modal hashing by generative adversarial network, IEEE transactions on cybernetics.

[82] C. Li, C. Deng, N. Li, W. Liu, X. Gao, D. Tao, Self-supervised adversarial hashing networks for cross-modal retrieval, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4242–4251.

[83] Y. Wang, X. Luo, L. Nie, J. Song, W. Zhang, X.-S. Xu, Batch: A scalable asymmetric discrete cross-modal hashing, IEEE Transactions on Knowledge and Data Engineering.

[84] T. Yao, Y. Han, R. Wang, X. Kong, L. Yan, H. Fu, Q. Tian, Efficient discrete supervised hashing for large-scale cross-modal retrieval, Neurocomputing 385 (2020) 358–367.

[85] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, H.T. Shen, Binary generative adversarial networks for image retrieval, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[86] T. Wang, L. Zhu, Z. Cheng, J. Li, Z. Gao, Unsupervised deep cross-modal hashing with virtual label regression, Neurocomputing. 386 (2020) 84–96.

[87] D. Zhang, R. Cao, S. Wu, Information fusion in visual question answering: A survey, Information Fusion.

[88] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, A. van den Hengel, Visual question answering: A survey of methods and datasets, Comput. Vis. Image Underst. 163 (2017) 21–40.

[89] D. Teney, P. Anderson, X. He, A. van den Hengel, Tips and tricks for visual question answering: Learnings from the 2017 challenge, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4223–4232.

[90] Q. Li, Q. Tao, S.R. Joty, J. Cai, J. Luo, Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions, in: ECCV, 2018.

[91] Y. Zhang, J.S. Hare, A. Prügel-Bennett, Learning to count objects in natural images for visual question answering, International Conference on Learning Representations (ICLR) abs/1802.05766.

[92] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, D. Parikh, Yin and yang: Balancing and answering binary visual questions, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5014–5022.

[93] K. Kafle, C. Kanan, Visual question answering: Datasets, algorithms, and future challenges, Comput. Vis. Image Underst. 163 (2017) 3–20.

[94] J. Gu, J. Cai, S.R. Joty, L. Niu, G. Wang, Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7181–7189.

[95] Y. Cao, M. Long, J. Wang, H. Zhu, Correlation autoencoder hashing for supervised cross-modal search, in: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ACM, 2016, pp. 197–204.

[96] H. Li, J. Zhu, C. Ma, J. Zhang, C. Zong, Read, watch, listen and summarize: Multi-modal summarization for asynchronous text, image, audio and video, IEEE Transactions on Knowledge and Data Engineering.

[97] Q. You, Z. Zhang, J. Luo, End-to-end convolutional semantic embeddings, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5735–5744.

[98] T. Chen, Z. Zhang, Q. You, C. Fang, Z. Wang, H. Jin, J. Luo, "factual"or"emotional", Stylized image captioning with adaptive learning and attention, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 519–535.

[99] L. Jin, K. Li, H. Hu, G.-J. Qi, J. Tang, Semantic neighbor graph hashing for multimodal retrieval, IEEE Trans. Image Process. 27 (3) (2018) 1405–1417.

[100] Y. Gao, O. Beijbom, N. Zhang, T. Darrell, Compact bilinear pooling, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 317–326.

[101] Y. Liu, L. Liu, Y. Guo, M.S. Lew, Learning visual and textual representations for multimodal matching and classification, Pattern Recogn. 84 (2018) 51–67.

[102] W. Wang, P. Liu, S. Yang, W. Zhang, Dynamic interaction networks for image-text multimodal learning, Neurocomputing 379 (2020) 262–272.

[103] W. Zhao, B. Wang, J. Ye, M. Yang, Z. Zhao, R. Luo, Y. Qiao, A multi-task learning approach for image captioning., in: IJCAI, 2018, pp. 1205–1211.

[104] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 21–29.

[105] H. Fan, J. Zhou, Stacked latent attention for multimodal reasoning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1072–1080.

[106] X. Zhu, L. Li, J. Liu, Z. Li, H. Peng, X. Niu, Image captioning with triple-attention and stack parallel lstm, Neurocomputing 319 (2018) 55–65.

[107] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, T. Zhang, Recurrent fusion network for image captioning, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 499–515.

[108] H. Hotelling, Relations between two sets of variates, in: Breakthroughs in statistics, Springer, 1992, pp. 162–190.

[109] Y. Hoshen, L. Wolf, Unsupervised correlation analysis, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 3319–3328.

[110] T. Tommasi, A. Mallya, B. Plummer, S. Lazebnik, A.C. Berg, T.L. Berg, Combining multiple cues for visual madlibs question answering, Int. J. Comput. Vision 127 (1) (2019) 38–60.

[111] Y. Liu, Y. Guo, E. Bakker, M.S. Lew, Learning a recurrent residual fusion network for multimodal matching, in: ICCV 2017, 2017.

[112] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, Y.-D. Shen, Dual-path convolutional image-text embedding with instance loss, arXiv preprint arXiv:1711.05535.

[113] Y. Zhou, R. Ji, J. Su, X. Sun, W. Chen, Dynamic capsule attention for visual question answering, in: Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

[114] Z. Ren, X. Wang, N. Zhang, X. Lv, L.-J. Li, Deep reinforcement learning-based image captioning with embedding reward, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 290–298.

[115] Y. Wang, H. Jin, A deep reinforcement learning based multi-step coarse to fine question answering (mscqa) system, in: Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

[116] X. Liu, M. Wang, Z.-J. Zha, R. Hong, Cross-modality feature learning via convolutional autoencoder, ACM Trans. Multimedia Comput., Commun., Appl. (TOMM) (1s) 15 (2019) 7.

[117] W. Xu, S. Keshmiri, G.R. Wang, Adversarially approximated autoencoder for image generation and manipulation, IEEE Transactions on Multimedia.

[118] J. Wang, W. Zhou, J. Tang, Z. Fu, Q. Tian, H. Li, Unregularized auto-encoder with generative adversarial networks for image generation, in: 2018 ACM Multimedia Conference on Multimedia Conference, ACM, 2018, pp. 709–717.

[119] A. Dash, J.C.B. Gamboa, S. Ahmed, M. Liwicki, M.Z. Afzal, Tac-gan-text conditioned auxiliary classifier generative adversarial network, arXiv preprint arXiv:1703.06412.

[120] Y. Feng, L. Ma, W. Liu, J. Luo, Unsupervised image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4125–4134.

[121] C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu, Q. Ju, Improving image captioning with conditional generative adversarial nets, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8142–8150.

[122] S. Yan, F. Wu, J.S. Smith, W. Lu, B. Zhang, Image captioning using adversarial networks and reinforcement learning, in: 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 248–253.

[123] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, M. Sun, Show, adapt and tell: Adversarial training of cross-domain image captioner, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 521–530.

[124] X. Xu, J. Song, H. Lu, Y. Yang, F. Shen, Z. Huang, Modal-adversarial semantic learning network for extendable cross-modal retrieval, in: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ACM, 2018, pp. 46–54.

[125] Y. Peng, J. Qi, Y. Yuan, Cm-gans: Cross-modal generative adversarial networks for common representation learning, CoRR abs/1710.05106.

[126] Q.-Y. Jiang, W.-J. Li, Deep cross-modal hashing, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 3270–3278.

[127] N. Sarafianos, X. Xu, I.A. Kakadiaris, Adversarial representation learning for text-to-image matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5814–5824.

[128] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, W. Wang, Adversary guided asymmetric hashing for cross-modal retrieval, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, ACM, 2019, pp. 159–167.

[129] S. Su, Z. Zhong, C. Zhang, Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3027–3035.

[130] W. Wang, X. Yang, B.C. Ooi, D. Zhang, Y. Zhuang, Effective deep learning-based multi-modal retrieval, The VLDB Journal?, Int J. Very Large Data Bases 25 (1) (2016) 79–101.

[131] V. Sharma, A. Vaibhav, S. Chaudhary, L. Patel, L. Morency, Attend and attack: Attention guided adversarial attacks on visual question answering models.

[132] Y. Liu, X. Zhang, F. Huang, Z. Li, Adversarial learning of answer-related representation for visual question answering, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, ACM, 2018, pp. 1013–1022.

[133] F. Liu, T. Xiang, T.M. Hospedales, W. Yang, C. Sun, Inverse visual question answering: A new benchmark and vqa diagnosis tool, IEEE Trans. Pattern Anal. Mach. Intell.

[134] Q. Wu, P. Wang, C. Shen, I. Reid, A. van den Hengel, Are you talking to me? reasoned visual dialog generation through adversarial learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6106–6115.

[135] S. Zhang, H. Dong, W. Hu, Y. Guo, C. Wu, D. Xie, F. Wu, Text-to-image synthesis via visual-memory creative adversarial network, in: Pacific Rim Conference on Multimedia, Springer, 2018, pp. 417–427.

[136] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, G. Wang, Unpaired image captioning via scene graph alignments, arXiv preprint arXiv:1903.10658.

[137] X. Li, S. Jiang, Know more say less: Image captioning based on scene graphs, IEEE Trans. Multimedia.

[138] F. Chen, R. Ji, J. Su, Y. Wu, Y. Wu, Structcap: Structured semantic embedding for image captioning, ACM Multimedia (2017).

[139] R. Xu, C. Li, J. Yan, C. Deng, X. Liu, Graph convolutional network hashing for cross-modal retrieval, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 982–988.

[140] J. Yu, Y. Lu, Z. Qin, W. Zhang, Y. Liu, J. Tan, L. Guo, Modeling text with graph convolutional network for cross-modal information retrieval, in: Pacific Rim Conference on Multimedia, Springer, 2018, pp. 223–234.

[141] J. Yu, C. Yang, Z. Qin, Z. Yang, Y. Hu, Z. Shi, Semantic modeling of textual relationships in cross-modal retrieval, in: International Conference on Knowledge Science, Engineering and Management, Springer, 2019, pp. 24–32.

[142] L. Li, Z. Gan, Y. Cheng, J. Liu, Relation-aware graph attention network for visual question answering, CoRR abs/1903.12314.

[143] M. Narasimhan, S. Lazebnik, A. Schwing, Out of the box: Reasoning with graph convolution nets for factual visual question answering, Adv. Neural Inform. Proces. Syst. (2018) 2654–2665.

[144] D. Teney, L. Liu, A. van den Hengel, Graph-structured representations for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1–9.

[145] J. Li, H. Su, J. Zhu, S. Wang, B. Zhang, Textbook question answering under instructor guidance with memory networks, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 3655–3663.

[146] Z. Su, C. Zhu, Y. Dong, D. Cai, Y. Chen, J. Li, Learning visual knowledge memory networks for visual question answering, 2018 IEEE Conference on Computer Vision and Pattern Recognition (2018) 7736–7745.

[147] C. Ma, C. Shen, A.R. Dick, Q. Wu, P. Wang, A. van den Hengel, I.D. Reid, Visual question answering with memory-augmented networks, 2018 IEEE Conference on Computer Vision and Pattern Recognition (2018) 6975–6984.

[148] C. Xiong, S. Merity, R. Socher, Dynamic memory networks for visual and textual question answering, in: ICML, 2016.

[149] J. Qi, Y. Peng, Cross-modal bidirectional translation via reinforcement learning, IJCAI (2018) 2630–2636.

[150] H. Xu, K. Saenko, Ask, attend and answer: Exploring question-guided spatial attention for visual question answering, in: ECCV, 2016.

[151] C. Wu, J. Liu, X. Wang, X. Dong, Object-difference attention: A simple relational attention for visual question answering, in: 2018 ACM Multimedia Conference on Multimedia Conference, ACM, 2018, pp. 519–527.

[152] J. Singh, V. Ying, A. Nutkiewicz, Attention on attention: Architectures for visual question answering (vqa), arXiv preprint arXiv:1803.07724.

[153] T. Qiao, J. Zhang, D. Xu, D. Tao, Mirrorgan: Learning text-to-image generation by redescription, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1505–1514.

[154] Q. Lao, M. Havaei, A. Pesaranghader, F. Dutil, L.D. Jorio, T. Fevens, Dual adversarial inference for text-to-image synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7567–7576.

[155] K. Joseph, A. Pal, S. Rajanala, V.N. Balasubramanian, C4synth: Cross-caption cycle-consistent text-to-image synthesis, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 358–366.

[156] Y. Wu, S. Zhao, J. Chen, Y. Zhang, X. Yuan, Z. Su, Improving captioning for low-resource languages by cycle consistency, in: IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2019, pp. 362–367.

[157] C. Li, C. Deng, L. Wang, D. Xie, X. Liu, Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 176–183.

[158] M. Cornia, L. Baraldi, H.R. Tavakoli, R. Cucchiara, Towards cycle-consistent models for text and image retrieval, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[159] Y. Liu, Y. Guo, L. Liu, E.M. Bakker, M.S. Lew, Cyclematch: A cycle-consistent embedding network for image-text matching, Pattern Recogn. 93 (2019) 365–379.

[160] M. Shah, X. Chen, M. Rohrbach, D. Parikh, Cycle-consistency for robust visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6649–6658.

[161] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, X. Wang, M. Zhou, Visual question generation as dual task of visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6116–6124.

[162] D. Yu, J. Fu, T. Mei, Y. Rui, Multi-level attention networks for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4709–4717.

[163] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1821–1830.

[164] H. Nam, J.-W. Ha, J. Kim, Dual attention networks for multimodal reasoning and matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 299–307.

[165] I. Schwartz, A. Schwing, T. Hazan, High-order attention models for visual question answering, Adv. Neural Inform. Process. Syst. (2017) 3664–3674.

[166] S. Li, T. Xiao, H. Li, W. Yang, X. Wang, Identity-aware textual-visual matching with latent co-attention, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1890–1899.

[167] C. Liu, Z. Mao, A.-A. Liu, T. Zhang, B. Wang, Y. Zhang, Focus your attention: A bidirectional focal attention network for image-text matching, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, 2019, pp. 3–11.

[168] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4651–4659.

[169] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 375–383.

[170] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, J. Gao, Object-driven text-to-image synthesis via adversarial training, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12174–12182.

[171] B. Li, X. Qi, T. Lukasiewicz, P.H. Torr, Controllable text-to-image generation, arXiv preprint arXiv:1909.07083.

[172] H. Tan, X. Liu, X. Li, Y. Zhang, B. Yin, Semantics-enhanced adversarial nets for text-to-image synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10501–10510.

[173] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, IEEE Comput. Intell. Maga. 13 (3) (2018) 55–75.

[174] W. Jiang, L. Ma, X. Chen, H. Zhang, W. Liu, Learning to guide decoding for image captioning, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[175] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[176] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907.

[177] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, M. Sun, Graph neural networks: A review of methods and applications, CoRR abs/1812.08434.

[178] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P.S. Yu, A comprehensive survey on graph neural networks, CoRR abs/1901.00596.

[179] S. Ren, K. He, R.B. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2015) 1137–1149.

[180] R. Zellers, M. Yatskar, S. Thomson, Y. Choi, Neural motifs: Scene graph parsing with global context, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 5831–5840.

[181] C.C. Park, B. Kim, G. Kim, Attend to you: Personalized image captioning with context sequence memory networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6432–6440.

[182] D.A. Hudson, C.D. Manning, Compositional attention networks for machine reasoning, arXiv preprint arXiv:1803.03067.

[183] B. Patro, V.P. Namboodiri, Differential attention for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7680–7688.

[184] Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering, IEEE Trans. Neural Networks and learning systems (99) (2018) 1–13.

[185] Y. Bai, J. Fu, T. Zhao, T. Mei, Deep attention neural tensor network for visual question answering, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 20–35.

[186] D.-K. Nguyen, T. Okatani, Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6087–6096.

[187] T. Qiao, J. Dong, D. Xu, Exploring human-like attention supervision in visual question answering, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[188] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, Adv. Neural Inform. Process. Syst. (2016) 289–297.

[189] Q. Wang, A.B. Chan, Gated hierarchical attention for image captioning, arXiv preprint arXiv:1810.12535.

[190] Z. Gan, Y. Cheng, A.E. Kholy, L. Li, J. Liu, J. Gao, Multi-step reasoning via recurrent dual attention for visual dialog, arXiv preprint arXiv:1902.00579.

[191] A. Agrawal, D. Batra, D. Parikh, A. Kembhavi, Don't just assume; look and answer: Overcoming priors for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4971–4980.

[192] R. Li, J. Jia, Visual question answering with question representation update (qru), Advances in Neural Information Processing Systems (2016) 4655–4663.

[193] A. Santoro, D. Raposo, D.G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, in: Advances in neural information processing systems, 2017, pp. 4967–4976.

[194] G. Peng, H. Li, H. You, Z. Jiang, P. Lu, S. Hoi, X. Wang, Dynamic fusion with intra-and inter-modality attention flow for visual question answering, arXiv preprint arXiv:1812.05252.

[195] S. Ding, S. Qu, Y. Xi, S. Wan, Stimulus-driven and concept-driven analysis for image caption generation, Neurocomputing 398 (2020) 520–530.

[196] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: International conference on machine learning, 2013, pp. 1247–1255.

[197] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 457–468.

[198] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, B.-T. Zhang, Hadamard product for low-rank bilinear pooling, arXiv preprint arXiv:1610.04325.

[199] H. Ben-Younes, R. Cadene, M. Cord, N. Thome, Mutan: Multimodal tucker fusion for visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2612–2620.

[200] P. Gao, H. Li, S. Li, P. Lu, Y. Li, S.C. Hoi, X. Wang, Question-guided hybrid convolution for visual question answering, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 469–485.

[201] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, H.T. Shen, Unsupervised cross-modal retrieval through adversarial learning, in: 2017 IEEE International Conference on Multimedia and Expo (ICME) IEEE, 2017, pp. 1153–1158.

[202] J. Gao, S. Wang, S. Wang, S. Ma, W. Gao, Self-critical n-step training for image captioning, arXiv preprint arXiv:1904.06861.

[203] N.Y.S.G. Siqi Liu, Zhenhai Zhu, K. Murphy, Improved image captioning via policy gradient optimization of spider, in: ICCV 2017, 2017.

[204] R.J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Mach. Learn. 8 (3–4) (1992) 229–256.

[205] B.A. Plummer, L. Wang, C.M. Cervantes, J.C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2641–2649.

[206] M.J. Huiskes, M.S. Lew, The mir flickr retrieval evaluation, in: Proceedings of the 1st ACM international conference on Multimedia information retrieval, ACM, 2008, pp. 39–43.

[207] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: Proceedings of the ACM international conference on image and video retrieval, ACM, 2009, p. 48.

[208] Y. Aytar, C. Vondrick, A. Torralba, See, hear, and read: Deep aligned representations, arXiv preprint arXiv:1706.00932.

[209] X. He, Y. Peng, L. Xie, A new benchmark and approach for fine-grained cross-media retrieval, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1740–1748.

[210] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, J. Shao, Semantics disentangling for text-to-image generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2327–2336.

[211] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C.L. Zitnick, D. Parikh, D. Batra, Vqa: Visual question answering, Int. J. Comput. Vision 123 (1) (2017) 4–31.

[212] C. Zhu, Y. Zhao, S. Huang, K. Tu, Y. Ma, Structured attentions for visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1291–1300.

[213] P. Lu, H. Li, W. Zhang, J. Wang, X. Wang, Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[214] J. Song, P. Zeng, L. Gao, H.T. Shen, From pixels to objects: Cubic visual attention for visual question answering., in: IJCAI, 2018, pp. 906–912.

[215] A. Osman, W. Samek, Dual recurrent attention units for visual question answering, arXiv preprint arXiv:1802.00209.

[216] C. Wu, J. Liu, X. Wang, R. Li, Differential networks for visual question answering, In AAAI.

[217] F. Liu, J. Liu, Z. Fang, H. Lu, Language and visual relations encoding for visual question answering, in: IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 3307–3311.

[218] F. Liu, J. Liu, Z. Fang, R. Hong, H. Lu, Densely connected attention flow for visual question answering, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 869–875.

[219] W. Chen, N. Pu, Y. Liu, E. Bakker, M. Lew, Domain Uncertainty Based On Information Theory for Cross-Modal Hash Retrieval, 2019 IEEE International Conference on Multimedia and Expo (ICME),IEEE, 2019, pp. 43–48.

Wei Chen is a doctoral candidate in Leiden Institute of Advanced Computer Science at Leiden University. His research interest focuses on cross-modal retrieval with deep learning methods. Before starting with PhD study in Leiden University, he received his Master degree from the National University of Defense Technology (NUDT), China, in 2016. He has published papers in international conferences and journal including ICPR, ICME and TMM.

Weiping Wang received the Ph.D. degree in systems engineering from the National University of Defense Technology (NUDT), Changsha, China, where he is currently a Professor. His research interests include systems engineering and simulation. He has more than 200 papers published on journals and conferences including IEEE Transactions on Vehicular Technology, Simulation, Simulation Modeling Practice and Theory, Software and Systems Modeling.

Li Liu received the BSc degree in communication engineering, the MSc degree in photogrammetry and remote sensing and the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), China, in 2003, 2005 and 2012, respectively. She joined the faculty at NUDT in 2012, where she is currently an Associate Professor with the College of System Engineering. During her PhD study, she spent more than two years as a Visiting Student at the University of Waterloo, Canada, from 2008 to 2010. From 2015 to 2016, she spent ten months visiting the Multimedia Laboratory at the Chinese University of Hong Kong. From 2016.12 to 2018.11, she worked as a senior researcher at the Machine Vision Group at the University of Oulu, Finland. She was a co-chair of nine International Workshops at CVPR, ICCV, and ECCV. She was a guest editor of special issues for IEEE TPAMI and IJCV. She serves as Area Chair for ACCV 2020 and ICME 2020. She currently serves as Associate Editor of the Visual Computer Journal and Pattern Recognition Letter. Her current research interests include computer vision, pattern recognition and machine learning. Her papers have currently over 2500+ citations in Google Scholar.

Michael S. Lew is head of Deep Learning and Computer Vision Research Group Director, LIACS Media Lab. He received his doctorate from University of Illinois at Urbana-Champaign and then became a postdoctoral researcher at Leiden University. One year later, he became the first Leiden University Fellow which was a pilot program for tenure track professors. In 2003, he became a tenured associate professor at Leiden University and was invited to serve as a chair full professor in computer science at Tsinghua University (the MIT of China). He has published over 100 peer reviewed papers with three best paper citations in the areas of computer vision, content-based retrieval, and machine learning. Currently (September 2014), he has the most cited paper in the history of the ACM Transactions on Multimedia. In addition, he has the most cited paper from the ACM International Conference on Multimedia Information Retrieval (MIR) 2008 and also from ACM MIR 2010. He has served on the organizing committees for over a dozen ACM and IEEE conferences. He served as the founding the chair of the ACM ICMR steering committee and had served as chair for both the ACM MIR and ACM CIVR steering committees. In addition he is the Editor-in-Chief of the International Journal of Multimedia Information Retrieval (Springer) and a member of the ACM SIGMM Executive Board which is the highest and most influential committee of the SIGMM.