Chapter 8

New Ideas and Trends in Deep Multimodal Content Understanding

In previous chapters, we focused the research on image retrieval and cross-modal retrieval in the context of non-incremental or incremental learning. In the past years, deep learning has also been explored for the field of multimodal learning.

In this Chapter, we present the recent new ideas and trends in multimodal content understanding filed, focusing on the analysis of two modalities: image and text. These new methods can be further used for intelligent image retrieval to seek performance improvement. Unlike classic reviews of deep learning where unimodal image classifiers such as VGG, ResNet, and Inception module are central topics, this chapter examines 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.

Keywords

Multimodal deep learning, Ideas and trends, Content understanding

This chapter is based on the following publication [39]:

- Chen, W., Wang, W., Liu, L., and Lew, M.S., “New Ideas and Trends in Deep Multimodal Content Understanding: A Review.” Neurocomputing, 2020, pp. 195-215.

8.1 Introduction

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 data specifically 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 [264], visual question answering (VQA) [265]) and uni-directional tasks (*e.g.* image captioning [266], image generation [22, 267]), both of them will be introduced in the following sections.

Data from visual and textual modalities are represented as unimodal features using domain-specific networks. Complementary information from these unimodal features is appealing for multimodal content understanding. For example, the unimodal features can be further projected into a common space by using another neural network for an vision task. For clarity, we illustrate the flowchart of deep multimodal research in Figure 8.1. 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 is helpful for reducing the "semantic gap" [10], as defined in Chapter 3. 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" [176]. To understand multimodal content, deep neural 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 associate 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 [268, 269], while others focus on feature extraction nets such as graph convolutional networks [270, 271]. In some algorithms, reinforcement learning is combined with deep multimodal feature learning [272, 273]. These recent ideas are the scope of this chapter.

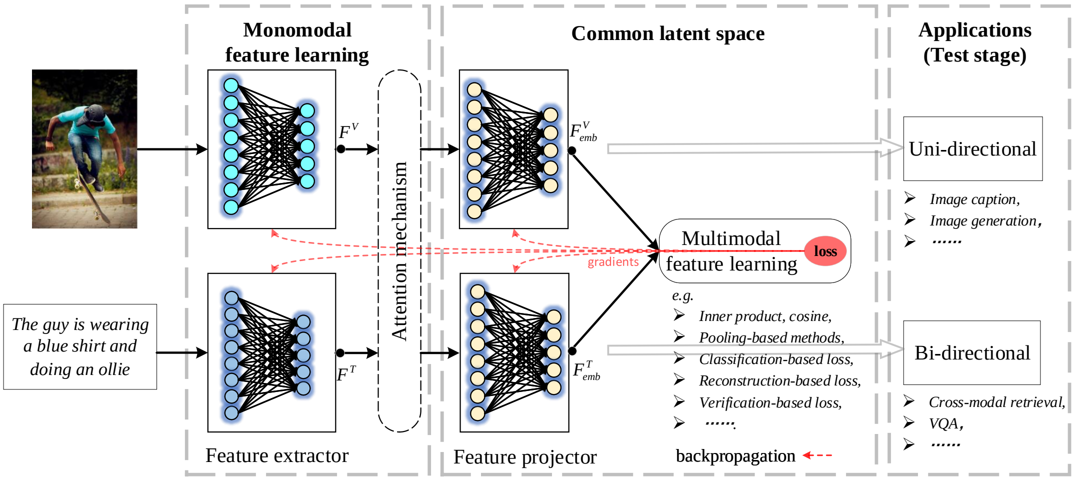


Figure 8.1: A general flowchart of deep multimodal feature learning.

8.2 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.

8.2.1 Uni-directional applications

a. 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 [274]. 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. In the following sections, we will examine the methods widely used to improve image captioning performance, including evolutionary algorithm [275], generative adversarial networks [180, 276], reinforcement learning [272, 273], memory networks [277, 278], and attention mechanisms [279, 280].

According to captioning principles, researchers focus on specific caption generation tasks, such as image tagging [281], visual region captioning [282], and object captioning [283]. 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.

b. 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 [192, 284]. 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) through architectures such as Variational Auto-Encoders (VAE) [285], auto-regressive models [286] and Generative Adversarial Networks (GANs) [22, 180]. One example is to generate a semantic layout as intermediate information from text data to bridge the heterogeneity gap in image and text [287, 288]. Some works focus on the network structure design for feature learning. For image synthesis, novel derivative architectures from GANs [180] have been explored in hierarchically nested adversarial networks [289], perceptual pyramid adversarial networks [290], iterative stacked networks [23, 291], attentional generative networks [292, 293], cycle-consistent adversarial networks [268], and symmetrical distillation networks [294].

One of limitations of image generation is that, while generation models work well and achieve promising results on single category object datasets like Caltech-UCSD CUB [295] and Oxford-102 Flower [295], 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 this limitation, word-level attention [292], hierarchical text-to-image mapping [288] and memory networks [296] have been explored. In the future, one direction may be to make use of the Capsule idea proposed by Hinton [297] since capsules are designed to capture the concepts of objects.

8.2.2 Bi-directional applications

a. Cross-modal retrieval

Cross-modal retrieval has been researched for decades. The aim is to return the most relevant image (text) when given a query text (image). There are two aspects should be considered: 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 [298]; Bi-directional sequential models (*e.g.* Bi-LSTM [188]) are used to explore spatial-semantic correlations [264]; Graph-based embedding and graph regularization are utilized to keep semantic order in text feature extraction process [299]; Information theory is applied to reduce the heterogeneity gap in cross-modal hashing [34]; Adversarial learning strategies and GANs are used to estimate common feature distributions [177, 300].

For the second, recent hashing methods have been explored owing to the computation and storage advantages of binary code [178]. Essentially, methods such as attention mechanisms and adversarial learning [178] 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. Focusing on the feature quantization, Wang *et al.* [301] 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.

There still exists room for performance improvement (see Figure 8.4-8.5). 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.

b. 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 inferred according to visual content and syntactic principle. Since VQA was proposed, it has received increasing attention in recent years. For example, there are some training datasets [302] built for this task, and some network training tips and tricks are presented in work [303].

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 two-fold 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.

The research on VQA includes: free-form open-ended questions [304], where the answer could be words, phrases, and even complete sentences; object counting questions [305] where the answer is counting the number of objects in one image; multi-choice questions [279] and Yes/No binary problems [306]. 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 [302] because the answers are not fixed, only the ones related to visual content and question details.

8.3 Recent Advances in Content Understanding

Lots of remarkable progresses about exploring content understanding between image and text 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 Figure 8.1), we categorize these research ideas into three groups: deep multimodal structures presented in Section 8.3.1, multimodal feature extraction approaches introduced in Section 8.3.2, and common latent space learning described in Section 8.3.3.

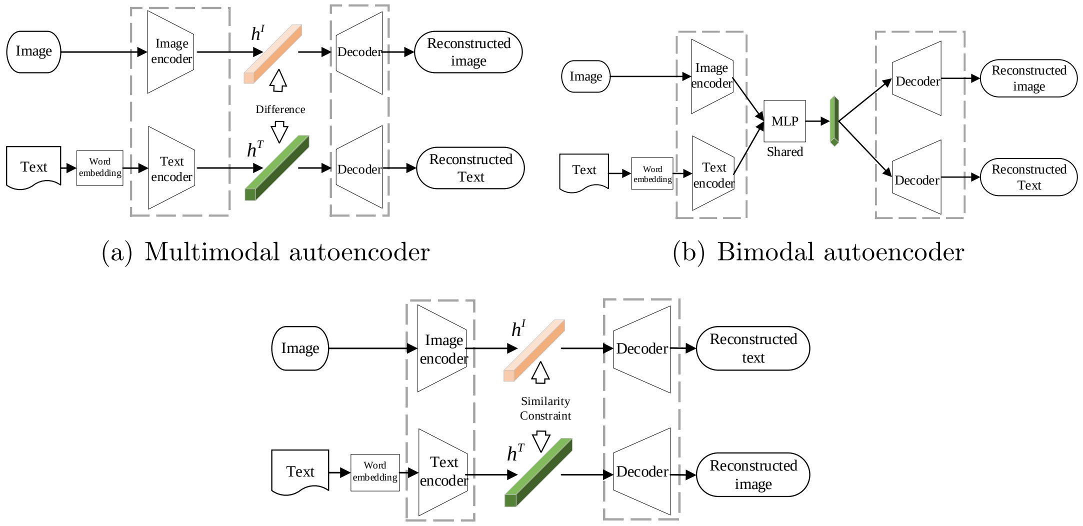
8.3.1 Deep multimodal structures

Deep multimodal structures are the fundamental frameworks to support different deep networks for exploring visual-textual semantics. These frameworks have critical influences for the following feature extraction and 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 [307]. During the past years, a variety of related methods have blossomed and accelerated the performance of multimodal learning directly in multimodal applications.

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.

a. AutoEncoders

The main idea of auto-encoder 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 for bi-directional applications where



(a) Multimodal autoencoder

(b) Bimodal autoencoder

(c) Correlation autoencoder

Figure 8.2: 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.

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 Figure 8.2. For instance, as shown in Figure 8.2(a), 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 [308]. The coordinated methods can be replaced by joint methods in Figure 8.2(b) 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 [309]. Alternatively, feature correlations are captured by cross reconstruction with similarity constraints between hidden features. The idea of constraining sample similarity is also incorporated with GANs into a cycle-consistent formation for cross-modal retrieval such as CYC-DGH [269].

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” [310] is to parallelize more parameters of LSTMs to capture more context information. Similar ideas can be found in “CNN ensemble learning” [311]. Instead of grabbing more information by stacking and paralleling, “Attention-LSTM” [310] combines attention technique into LSTM to highlight most relevant correlations. 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 [309]. Considering the fixed structure in the decoder like RNN might limit the performance, Wang *et al.* [275] introduce evolutionary algorithm to adaptively generate neural network structures in the decoder.

b. Generative adversarial networks

Adversarial learning from generative adversarial networks [180] has been employed into applications including image captioning [312], cross-modal retrieval [309] and image generation [23, 289, 291], 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.

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*.

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 Attention-aware methods [292]. 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 [23]. To improve its capacity for learning hierarchical features, a generator can be organized into different nested structures to capture multi-level semantics such as hierarchical-nested [289] and hierarchical-pyramid [290].

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 [180] just needs to classify different distributions into “ *True*” or “*False*” [22]. However, discriminator can also make a class label classification where a label classifier is added on the top of discriminator [313]. 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 [267]. 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 [22] and AACR [300]) so that the discriminator would have a more powerful discriminative capability.

The application of GANs in multimodal research are categorized into direct methods [22, 313], hierarchical methods [289, 290], and iterative methods [23, 291, 292]. 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-to-fine" process where details of the results from a previous generator are refined. Besides, cycle-consistency [314] 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 idea is then applied into tasks like image generation [268] and cross-modal retrieval [269] to learn semantic correlation in an unsupervised way.

In recent years, adversarial learning is widely used to design algorithms for deep multimodal learning [177, 178, 309]. 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 [309, 315, 316]. For example, in the first work about unsupervised image captioning [316], 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.

8.3.2 Multimodal feature extraction

Feature extraction is closer for exploring visual-textual content relations, which is the prerequisite to discriminate the complementarity and redundancy of multiple modalities. 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 alignment schemes using attention mechanism are also widely explored for preserving semantic correlations.

a. 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. 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 [154]. For deep multimodal learning, GCNs receive increasing attention and have achieved breakthrough performance on several applications, including cross-modal retrieval [317], image captioning [270, 271, 318], and VQA [319].

Graph convolutional networks in multimodal learning can be employed in text feature extraction [317, 319] and image feature extraction [270, 271, 318]. 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 are introduced to determine the attributes and subsequently characterize the relationships between image and text [319]. To use GCNs, 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 GCNs are fed into sequential networks to predict answers.

As an alternative method, GCNs are worthy more exploration for correlations between two modalities. Moreover, there exist two limitations in GCNs. On the one hand, graph construction process is overall time- and space-consuming; On the other hand, the accuracy of output activations from GCNs 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.

b. Memory-augmented networks

To enable deep networks to understand multimodal content and have better reasoning capability for various tasks, 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. 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 [278]. For this limitation, memory networks [277] are introduced to augment the memory primarily used for text question-answering [320]. Memory networks improve understanding of both image and text, and then "remember" temporally distant information.

Memory-augmented networks are used in cross-modal retrieval [298], image captioning [278], and VQA [321]. 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. The memory slots are a kind of external memory to support learning. During training, a network such as LSTM or GRU, which acts as a memory controller, refers to these memory slots to compute reading weights. 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 [322].

The performance of memory networks 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 [323] to further improve its feature learning capability. For example, Xiong *et al.* [324] 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 [325] 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.* [326] 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 memory slots store key-value vectors computed from images, query questions and a knowledge base [321]; Instead of storing the actual output features, Song *et al.* [298] adopt memory slots to store a prototype concept representation from pre-trained concept classifiers, which is inspired from the process of human memory.

c. Attention mechanism for deep multimodal learning

Attention mechanisms are widely used to tackle this issue in various multimodal tasks, such as VQA [327, 328, 329] and image captioning [266, 270, 280]. In principle, the attention mechanisms compute different importances according to relevances between two global (or local) multimodal features and assign different importances to these features. Thereby, the networks are 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

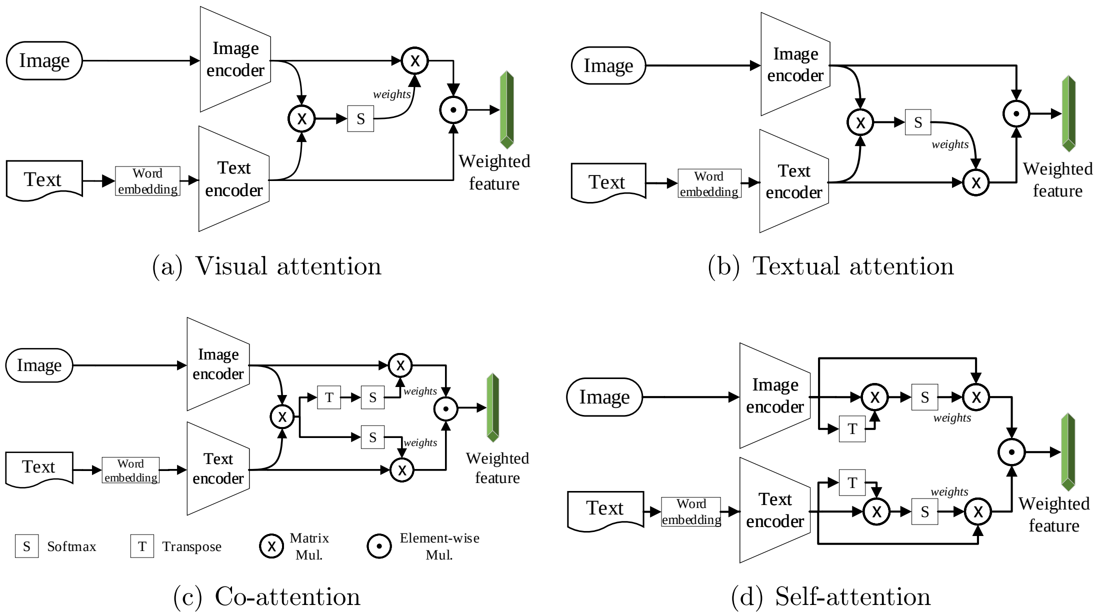


Figure 8.3: Diagram for different types of attention mechanisms used in deep multimodal learning.

on multi-level feature vectors [325], employed in a hierarchical scheme [330], and incorporated with graph networks for modelling semantic relationships.

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, as illustrated in Figure 8.3. We further categorize the attention algorithms into single-hop and multiple-hop (*i.e.* stacked attention) according to the iterations of importance calculation.

Visual attention. As shown in Figure 8.3(a), 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 [266, 270, 279, 280, 325, 327, 328] and has been incorporated to get different text-guided features. For example, Anderson *et al.* [279] 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.

Textual attention. Compared to visual attention, the textual attention approach is relatively less adopted. As shown in Figure 8.3(b), it has an opposite computing direction [105]. The computed weights are based on text features to obtain relevances for different image regions or objects. According to [320], 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.

Co-attention. As shown in Figure 8.3(c), co-attention algorithm is viewed as a combination of visual attention and textual attention, which is an option to explore the inter-modality correlations [202, 204, 323, 329]. 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.* [329] introduce a dense symmetric co-attention method to 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. Meanwhile, several other works explore different formations of co-attention. Integrating image feature with hierarchical text features may vary dramatically so that the complex correlations are not fully captured. For this, Yu *et al.* [331] 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.

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 [332]. Self-attention has been used in different ways. For example, Gao *et al.* [333] combine the attentive vectors form self-attention with co-attention using element-wise product. Thereby the inter- and intra-modality information flow are modeled by the linear method.

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 [202, 323, 325, 328]. For example, Xu *et al.* [325] 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. Singh *et al.* [328] 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 [323]. Regarding this, Fan *et al.* [323] 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 8.1.

Table 8.1: 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 [323].	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.

8.3.3 Common latent space learning

As illustrated in Figure 8.1, unimodal features distribute inconsistently and are not directly comparable. It is necessary to further map these unimodal features into a common latent space with the help of an embedding networks (*e.g.* MLP). Due to this, 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. In this section, we include these ideas, such as attention mechanisms, which aim to retain similarities between unimodal image and text features.

According to [334], multimodal feature learning methods include joint and coordinated methods. The joint feature embeddings are defined as:

$$J = \mathcal{J}(x_1, ..., x_n, y_1, ..., y_n) \tag{8.1}$$

while coordinated feature embeddings are represented as:

$$F = \mathcal{F}(x_1, ..., x_n) \sim \mathcal{G}(y_1, ..., y_n) = G \tag{8.2}$$

where J refers to the jointly embedded features, F and G denote the coordinated features. $x_1, ..., x_n$ and $y_1, ..., y_n$ are n-dimension unimodal feature representations from two modalities. 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 unimodal features are separated but are related by some similarity constraints.

a. Joint feature embedding

In deep multimodal learning, joint feature embedding is a straightforward way in which unimodal features are combined into the same presentation. The fused features are used to make a classification in cross-modal retrieval. It also can be used for performing sentence generation in VQA [279, 304].

In early studies, some basic methods are employed for joint feature embedding such as feature summation, feature concatenation [23, 291, 292], and element-wise inner product [324, 329], 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 such as Multimodal Compact Bilinear (MCB) pooling [335]. However, the performance of MCB is based on a higher-dimensional space. Regarding this demerit, Multimodal Low-rank Bilinear pooling [336] and Multimodal Factorized Bilinear pooling [331] are proposed to overcome the high computational complexity when learning joint feature. Moreover, Hedi *et al.* [337] 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 [313] and regression-based objective functions [23, 292] 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 Figure 8.2(b), are used to project image and text features with a shared multi-layer perceptron (MLP). The similar multimodal transformer introduced in [332] 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 [310]), and finally combined with text features to predict correct answers [302, 312, 326]. 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 [203, 338]. 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 [338].

The attention mechanisms in Section 8.3.2 can also be regarded as a kind of joint feature alignment method and are widely used for common latent space learning. Theoretically, these feature alignment schemes aim at finding relationships and correspondences between instances from visual and textual modalities [320, 334]. In particular, the mentioned co-attention mechanism is a case of joint feature embedding in which image and text features are usually treated symmetrically. 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 [329], hierarchical [331], and stacked [202, 325]. More importantly, the metrics for measuring similarity are crucial in attentive importance estimation. For example, the importance estimation by simple linear operation 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 [302], there is potential for improvement in the embedding space.

b. 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. Using traditional canonical correlation analysis, as an alternative, the correlations between two kinds of features can be measured and then maintained. 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 [177, 178]. Cross-entropy loss between the inference labels and the ground-truth labels is computed to optimize the deep networks, see Figure 8.1, 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 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, Li *et al.* [178] 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 [177, 339]. 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).

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 [186]. Recently, new ideas on coordinated feature embedding methods have combined adversarial learning, reinforcement learning, cycle-consistent constraints to pursue high performance.

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 [287, 288, 300].

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 [177, 178, 309], 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.

Combined with reinforcement learning. Reinforcement learning has been incorporated into deep network structures (*e.g.* encoder-decoder framework) for image captioning [272, 273], VQA [340] and cross-modal retrieval. Because reinforcement learning avoids exposure bias [273, 339] and non-differentiable metric issue [272, 339]. 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 [341] is widely used as a policy gradient method to compute gradients, then to update these "states" via back-propagation algorithms [340, 342]. 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" [272, 342]. 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) [339]. The design of triplet ranking loss function is diverse, such as in a bi-directional manner or based on inter-modal triplet sampling [339].

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 [269], image generation [268], and VQA [343].

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.* [268] 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.* [344] 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.* [269] 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. For this limit, Li *et al.* [344] 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.* [343] 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 [268]. 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.

8.4 Results and Discussions

The aforementioned ideas have made some progress on various multimodal tasks. For example, for cross-modal retrieval, we presented the achieved progress and state-

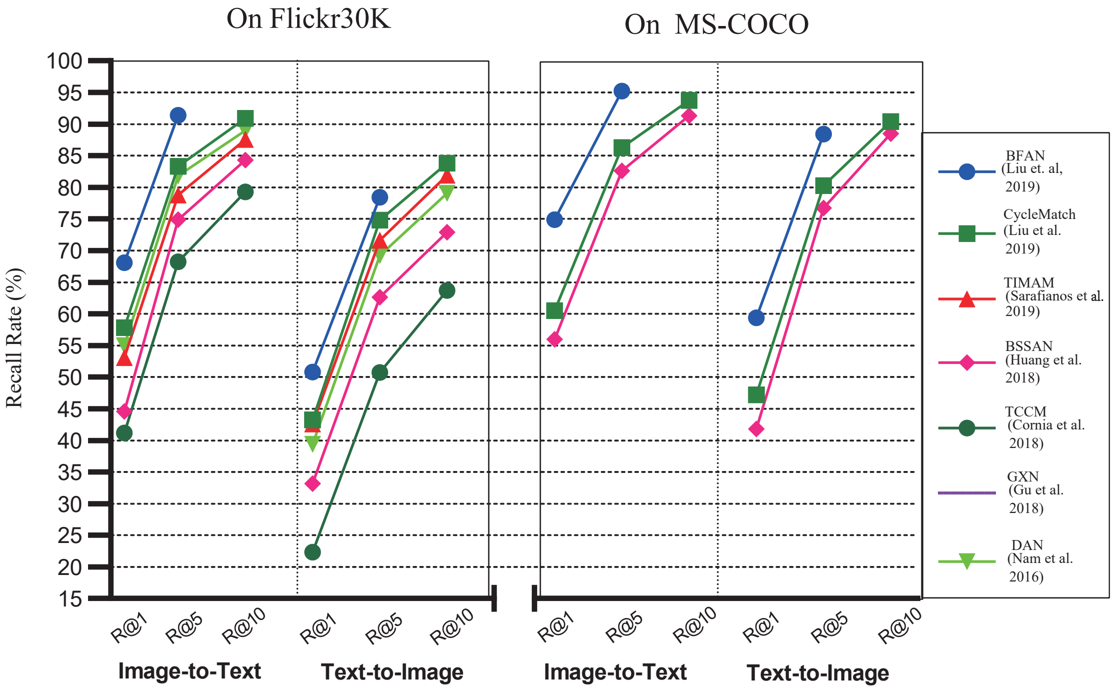


Figure 8.4: The achieved progress of cross-modal retrieval on the Flickr30K [345] and the MS-COCO [192] datasets.

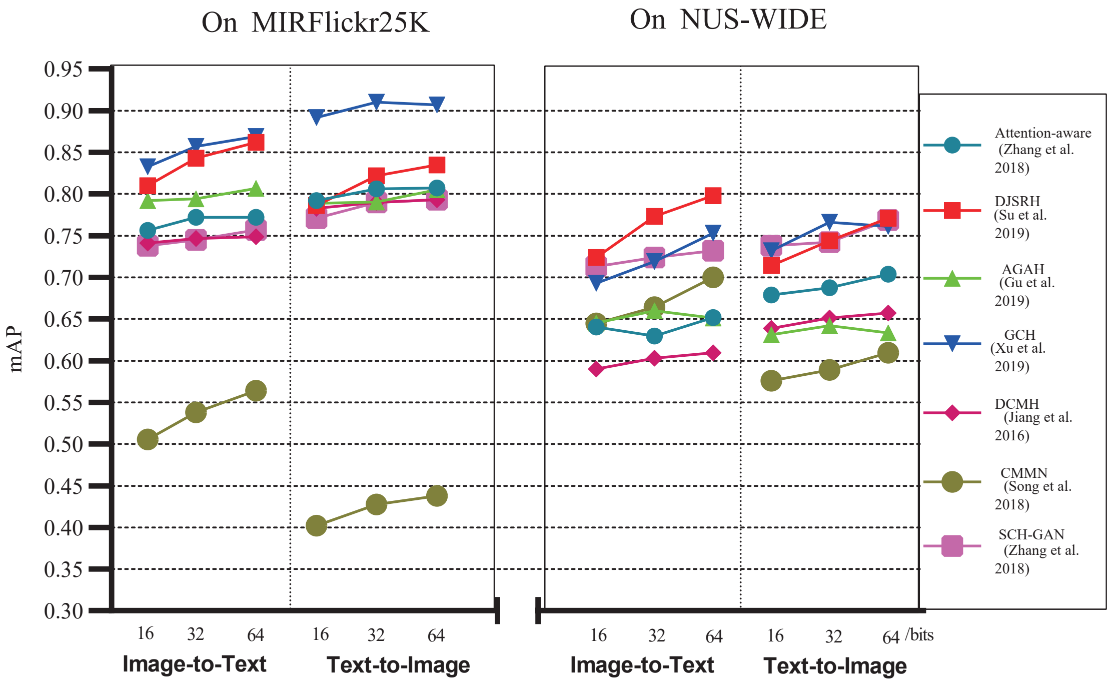


Figure 8.5: The achieved progress of cross-modal hash retrieval on the MIRFlickr25k [182] and the NUS-WIDE [183] datasets. Hashing methods have higher retrieval efficiency using the binary hash codes.

of-the-art of recent methods on the Flickr30K [345] and the MS-COCO [192] datasets in Figure 8.4. For hashing retrieval methods, we presented the achievement on the MIRFlickr25k [182] and the NUS-WIDE [183] datasets in Figure 8.5. 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. As can be seen from the results, there is still room for improvement in the current limitations of multimodal content understanding. In terms of other three tasks, (*i.e.* image generation, image captioning and VQA), the achieved performance in recent years are reported in Tables 8.2, 8.3, and 8.4, respectively.

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.* [367] 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.* [368] 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 unimodal 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. Finally, generation-based tasks such as image generation and image captioning are effective for unsupervised learning, since numerous labeled

Table 8.2: Performance of image captioning on the MS-COCO dataset [192]

Methods	**CIDEr**		**ROUGE-L**		**METEOR**		**BLEU1**		**BLEU2**		**BLEU3**		**BLEU4**		**KeyNotes**
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	
StructCap [318]	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 [346]	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 [276]	102.0	-	52.7	-	24.8	-	-	-	-	-	39.3	-	29.9	-	CGAN + Reinforcement learning
Adaptive-Attn [347]	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 [272]	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 [348]	-	-	-	-	24.3	-	71.6	-	51.8	-	37.1	-	26.5	-	GAN + Reinforcement learning
SOT [266]	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 [339]	117.1	-	57.0	-	27.4	-	80.1	-	63.1	-	48.0	-	35.8	-	Reinforcement learning
Up-Down [272]	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 [273]	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 [311]	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 [349]	168.2	-	46.6	-	20.1	-	42.1	-	32.0	-	25.3	-	20.5	-	Reinforcement learning
UnsupervisedIC [316]	54.9	-	43.1	-	17.9	-	58.9	-	40.3	-	27.0	-	19.6	-	VAE + GAN (unsupervised)
Graph-align [314]	69.5	-	-	-	20.9	-	67.1	-	47.8	-	32.3	-	21.5	-	VAE + Graph embed + cycle (unpaired)
Self-critical [342]	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 [280]	118.6	-	58.6	-	30.4	-	83.2	-	62.8	-	46.3	-	40.8	-	Attention + Reinforcement learning
RL-CGAN [312]	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 [271]	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 8.3: Performance of image generation

Methods	Caltech-UCSD Birds200		Oxford Flowers102		MS-COCO		KeyNotes
	IS	FID	IS	FID	IS	FID	
GAN-INT-CLS [22]	2.88±.04	-	2.66±.03	-	7.88±0.07	-	Vanilla GAN for image generation
TAC-GAN [313]	-	-	3.45±.05	-	-	-	Discriminator learns class information
GAWWN [293]	3.60±.07	-	-	-	-	-	Conditional objection location is learned
StackGAN [23]	3.70±.04	51.89	3.20±.01	55.28	8.45±.03	74.05	GAN in a stacked structure
StackGAN++ [291]	4.04±.05	15.3	3.26±.01	48.68	8.30±.10	81.59	GAN in a tree-like structure
HDGAN [289]	4.15±.05	-	3.45±.07	-	11.86±0.18	-	GAN in a hierarchically-nested structure
AttnGAN [292]	4.36±.03	-	-	-	25.89±.47	-	Attentional generative network
Scene graphs [287]	-	-	-	-	7.3±0.1	-	Graph convolution for graphs from text
Obj-GANs [350]	-	-	-	-	27.37±0.22	25.85	Attentive generator and object-wise discriminator
vmCAN [296]	-	-	-	-	10.36±0.17	-	Visual-memory method in GAN
AAAE [315]	-	-	-	103.46	-	-	Auto-encoders + GAN for adversarial approximation
Text-SeGAN [267]	-	-	4.03±0.07	-	-	-	Semantic relevance matching in GAN
DAI [351]	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 [352]	4.07±0.13	-	3.52±0.15	-	-	-	Image generation using multiple captions
PPAN [290]	4.35±.05	-	3.53±.02	-	-	-	GAN in perceptual pyramid structure
MirrorGAN [353]	4.56±0.05	-	-	-	26.47±0.41	-	Task-oriented cycle consistency + attention
ControlGAN [354]	4.58±0.09	-	-	-	24.06±0.6	-	Attention + region-wise attribute generation
SD-GAN [355]	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 8.4: Performance of visual question answering on VQA 1.0 dataset [356]

Methods	**Open-ended test-std**	**Open-ended test-dev**	**MC test-std**	**MC test-dev**	**KeyNotes**
Smem-VQA [325]	58.24	57.99	-	-	Spatial memory network stores image region features
DMN+ [324]	60.4	60.3	-	-	Improved dynamic memory network for VQA
MLB [336]	65.07	64.89	68.89	-	Low-rank bilinear pooling for similarity learning
MCB [335]	66.5	66.7	70.1	70.2	Multimodal compact bilinear pooling for similarity learning
High-order Attn [357]	-	-	69.3	69.4	Attention mechanisms learn high-order feature correlations
DAN [202]	64.2	64.3	69	69.1	Co-attention networks for multimodal feature learning
MLAN [358]	65.3	65.2	70	70	Multi-level co-attention for feature alignment
SVA [359]	66.1	66	-	-	Visual attention on grid-structured image region feature learning
MFB [331]	66.6	66.9	71.4	71.3	Multi-modal factorized bilinear pooling for similarity learning
MUTAN [337]	67.36	67.42	-	-	Multimodal tucker fusion for similarity learning
Graph VQA [319]	70.42	-	74.37	-	Graph representation for scene and question feature learning
MAN-VQA [326]	64.1	63.8	69.4	69.5	Memory-augmented network for feature learning and matching
QGHC [338]	65.9	65.89	-	-	Question-guided convolution for visual-textual correlations learning
Dual-MFA [360]	66.09	66.01	69.97	70.04	Co-attention for visual-textual feature learning
VKMN [321]	66.1	66	69.1	69.1	Visual knowledge memory network for feature learning
CVA [361]	66.2	65.92	70.41	70.3	Cubic visual attention for object-region feature learning
DCN [329]	67.02	66.89	-	-	Dense co-attention for feature fusion
DRAU [362]	67.16	66.86	-	-	Recurrent co-attention for feature learning
ODA [327]	67.97	67.83	72.23	72.28	Object-difference visual attention to fuse features
ALARR [363]	68.43	68.61	71.28	68.43	Adversarial learning for pair-wise feature discrimination
DF [364]	68.48	68.62	73.05	73.31	Differential network for visual-question feature learning
Relational Encoding [365]	69.3	69.1	-	-	Textual attention for question feature encoding
DCAF [366]	70.0	69.9	-	-	Dense co-attention for feature fusion

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.

8.5 Chapter Conclusions

In this chapter, 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.