https://doi.org/10.26599/AIR.2022.9150008 CAAI Artificial Intelligence Research

A Survey of Vision and Language Related Multi-Modal Task

Lanxiao Wang[1], Wenzhe Hu[1], Heqian Qiu[1], Chao Shang[1], Taijin Zhao[1], Benliu Qiu[1], King Ngi Ngan[2],
and Hongliang Li[1] ✉

ABSTRACT

With the significant breakthrough in the research of single-modal related deep learning tasks, more and more works begin to focus on multi-modal tasks. Multi-modal tasks usually involve more than one different modalities, and a modality represents a type of behavior or state. Common multi-modal information includes vision, hearing, language, touch, and smell. Vision and language are two of the most common modalities in human daily life, and many typical multi-modal tasks focus on these two modalities, such as visual captioning and visual grounding. In this paper, we conduct in-depth research on typical tasks of vision and language from the perspectives of generation, analysis, and reasoning. First, the analysis and summary with the typical tasks and some pretty classical methods are introduced, which will be generalized from the aspects of different algorithmic concerns, and be further discussed frequently used datasets and metrics. Then, some other variant tasks and cutting-edge tasks are briefly summarized to build a more comprehensive vision and language related multi-modal tasks framework. Finally, we further discuss the development of pre-training related research and make an outlook for future research. We hope this survey can help relevant researchers to understand the latest progress, existing problems, and exploration directions of vision and language multi-modal related tasks, and provide guidance for future research.

KEYWORDS

deep learning; vision and language; multi-modal generation; multi-modal analysis; multi-modal reasoning; pre-training

In the early days of the development of deep learning, some works paid more attention to single-modal related research, especially in the field of Computer Vision (CV) and Natural Language Processing (NLP). With the continuous development of single-modal related tasks such as classification tasks, detection tasks, and segmentation tasks, major breakthroughs have been made in the single-modal related research field. However, the practical applications of the real world usually involve information over multiple modalities. For example, online shopping usually requires image retrieval based on text descriptions. In autonomous driving, in addition to voice interaction, it is also necessary to analyze road conditions based on the collected videos or images, so as to give the driver language feedback. Therefore, researchers have gradually realized that just making research on the single-modal is far from enough, and further multi-modal related research is needed on feature fusion and transformation between different modalities.

In recent years, more and more studies have begun to focus on multi-modal related tasks. Multi-modal related tasks usually involve two or more modalities information. Common multi-modal information includes vision, hearing, and language, among which the most common modalities are vision and language. In this paper, we conduct an in-depth study on typical vision and language related multi-modal tasks. Figure 1a shows common vision and language related multi-modal tasks. According to the function of tasks, these tasks can be divided into (i) generation related tasks, (ii) analysis related tasks, and (iii) reasoning related tasks.

The most classic generation related task is visual captioning, including image captioning and video captioning. These tasks aim to realize the mapping between different modalities, which requires the network to understand and transform the information between the two different modalities. The analysis related tasks mainly come from the variants of traditional CV tasks including detection and segmentation, which requires the network to understand and fuse the information of the two input modalities, and generate predictions for the corresponding tasks. The representative analysis related tasks are visual grounding and referring segmentation. According to different location targets, visual grounding could be further subdivided into referring expression comprehension and phrase localization. Reasoning related tasks require the network to perform one or more times of knowledge reasoning between different modalities to obtain answers, which proposes a higher challenge to semantic analysis within modalities and semantic mapping and fusion between different modalities. The typical reasoning related task is visual question answering.

In order to introduce vision and language related multi-modal tasks more comprehensively and systematically, some typical tasks and classic methods are introduced in detail first from the above three perspectives. We summarize recent classic methods based on their characteristics, and further discuss common datasets and metrics for corresponding tasks. Then, some other variant tasks and cutting-edge tasks are briefly summarize to build a more comprehensive vision and language related multi-modal task structure.

With the continuous expansion of the training data scale, some works begin to pay attention to the vision and language pre-

1 Department of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
2 The Chinese University of Hong Kong, Hong Kong 999077, China
Address correspondence to Hongliang Li, hlli@uestc.edu.cn

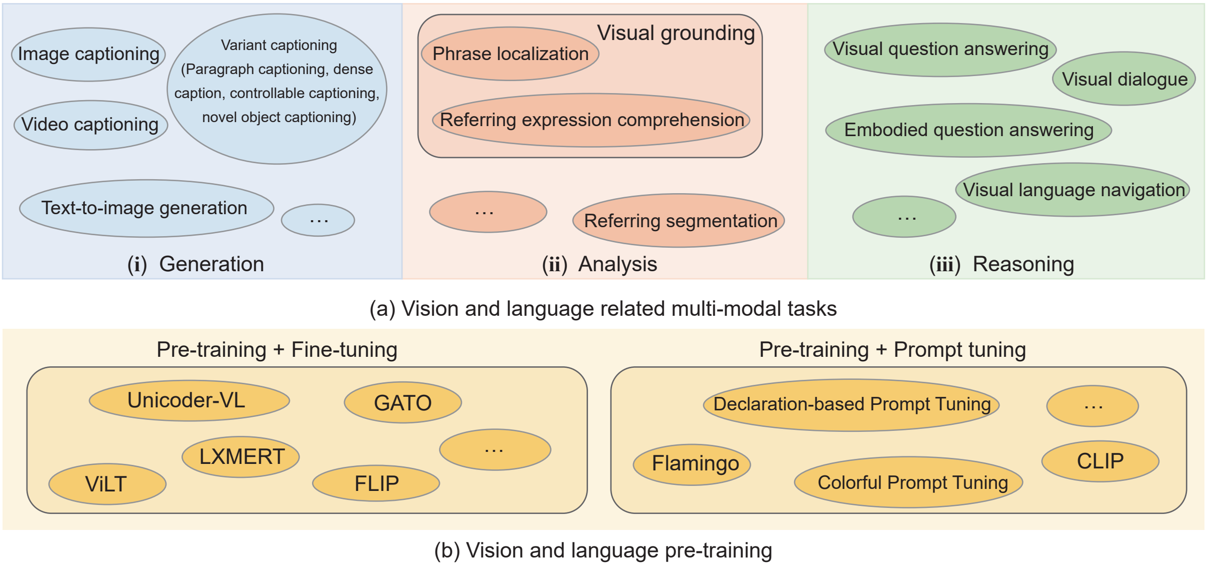


Image captioning
Variant captioning (Paragraph captioning, dense caption, controllable captioning, novel object captioning)
Video captioning
Text-to-image generation
…
(i) Generation
Visual grounding
Phrase localization
Referring expression comprehension
…
Referring segmentation
(ii) Analysis
Visual question answering
Visual dialogue
Embodied question answering
Visual language navigation
…
(iii) Reasoning
Pre-training + Fine-tuning
Unicoder-VL
GATO
LXMERT
…
ViLT
FLIP
Pre-training + Prompt tuning
Declaration-based Prompt Tuning
…
Flamingo
CLIP
Colorful Prompt Tuning

(a) Vision and language related multi-modal tasks

(b) Vision and language pre-training

Fig. 1　Main framework of the survey about vision and language related multi-modal task.

training tasks, hoping to learn the semantic relationships between different modalities through the pre-training of large-scale datasets, which can be used to get more outstanding performance on the multi-modal downstream tasks. Therefore, as shown in Fig. 1b, the development of vision and language pre-training research is discussed from the two perspectives of “pre-training + fine-tuning” and “pre-training + prompt tuning”. Finally, we analyze the challenges faced by the vision and language related multi-modal field and the potential development directions in the future, hoping to provide guidance for relevant research.

1　Generation Related Tasks

1.1　Image captioning

1.1.1　Task definition

Image captioning task aims to describe the content or event of the given image in natural language. This task requires the network to understand visual information especially the objects and their relationships and transform visual modal into linguistic modal. In recent years, with the development of deep neural networks, many researchers devote themselves to designing effective and efficient methods for image captioning based on the structure of encoder-decoder, shown in Fig. 2. Image captioning, as a classic vision and language related task, faces enormous challenges because it requires the model not only to recognize the objects and understand the events in the images but also to generate a reasonable natural linguistic description. Moreover, the descriptions of the same image may be diverse, making it more difficult to judge whether a description is appropriate.

1.1.2　Methods

In the beginning, early approaches are based on conventional image processing and machine learning algorithms. Some researchers converted image captioning to a description retrieval process[1–3], linking the most similar sentence to an image. Some other methods[4,5] entail template to restrain the sentence structures, and employ language models to guide description generation. Though these methods obtain relatively low performance, their thoughts are still worthy of further exploration. As deep learning neural networks come into being, some deep learning-based methods play paramount and prevalent roles in image captioning, such as Convolutional Neural Networks (CNN)-Recurrent Neural Networks (RNN) structure, attention mechanism, Transformer-based structure, and graph structure.

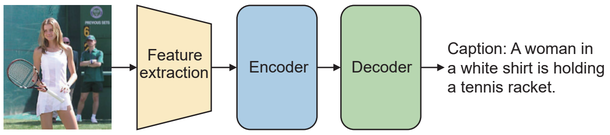

Feature extraction
Encoder
Decoder
Caption: A woman in a white shirt is holding a tennis racket.

Fig. 2　Framework of image captioning.

CNN-RNN

With the development of CNN and deep learning networks, the feature map, usually the output of the last convolutional layer, is employed to represent high-level visual contents. Thus, some methods make use of CNN-based networks as a visual feature extractor to acquire proper representation and feed them into a language model to generate descriptions. As the boost of image recognition networks, many approaches leverage them as powerful feature extractors. Generally, the extractor is pre-trained on a large-scale image dataset, such as ImageNet[6], and then extracts visual features on target datasets. For the language model, those methods adopt Recurrent Neural Networks for generating word sequences.

Karpathy and Li[7] got features from AlexNet[8] and RCNN[9], while Vinyals et al.[10] used the more powerful GoogleNet[11] for feature extraction. Afterwards, more approaches[12–16] take CNN features as the input of recurrent model to generate captions. As for the structure of encoder-decoder, researchers usually use the CNN network to extract features which is regarded as the encoder, and the RNN model, usually LSTM, to generate captions regarded as the decoder. Such an encoder-decoder paradigm becomes prevalent in the afterward methods.

Attention mechanism

As deep learning research continues to deepen, researchers realize that the network should focus on more important places. Thus, the attention mechanism is widespread in computer vision and natural language processing applications. It can make the model focus on a certain part of the input rather than all of them.

Xu et al.[17] first incorporated the attention mechanism into the image captioning model. They propose two kinds of attention: stochastic hard attention and deterministic soft attention. Though hard attention achieves better performance, it is less practical because the gradient cannot be back-propagated. The soft attention is to compute the similarity between the hidden state and visual features, then predict a probability distribution indicating the importance of feature vectors at each time step. Such attention mechanism is widely used and improved by further studies[18,19]. SCA-CNN[20] combines spatial and channel-wise attention to grasp both the visual and semantic information of input features. Lu et al.[21] designed a visual sentinel to adaptively attend to visual content and language prior when predicting visual and non-visual words. Some methods[22–24] leverage the saliency map, i.e., the human attention, to guide the attention module.

The aforementioned methods employ CNN feature maps. Whereas in 2018, Anderson et al.[25] proposed the bottom-up top-down attention model, which firstly adopts the region-level visual features. It adopts the Faster R-CNN model[26] for object detection to generate a set of proposals, and mean-pool the feature vectors of each proposal. In this way, the feature map is transferred into a batch of region features. The number of proposals is set to 36 in that paper. Those extracted features are cached and then fed into the language model. Notably, subsequent approaches mainly follow this structure and directly process the cached offline features[27–29]. For the decoder, it designs a two-layer LSTM network with an attention mechanism, where one LSTM is for attention and the other for language. Specifically, the first LSTM takes the previous word, the mean-pooled feature, and the previous hidden state as inputs. Then the attention module predicts a distribution over region features based on the current hidden state and image features. The weighted sum of features is fed into the second LSTM, and the current word is predicted.

Transformer-based structure

The powerful Transformer model is proposed by Vaswani et al.[30] for machine translation. This architecture is completely based on the attention mechanism, discarding recurrent and convolution models. The proposed multi-head self-attention mechanism computes the cosine similarity of the query Q and the key K and assigns weights on the value V. The Transformer model and multi-head self-attention inspire further development in computer vision, including image captioning.

Some methods adapt the Transformer architecture for captioning. Li et al.[31] designed two sub-encoder, both following the Transformer model, to respectively capture visual and linguistic information. Herdade et al.[32] and Guo et al.[33] successively incorporated geometry embedding to encode object spatial relationships.

Some other approaches employ the self-attention mechanism to explore the latent relationships between visual features. The AoA model[34] adds another attention module after Transformer to determine the relevance between attended features and queries. In the decoder, the AoA module refines the visual features, while in the decoder it computes the context vector used for word prediction. Cornia et al.[35] proposed Meshed-Memory Transformer, aiming at integrating a priori information and learning multi-level features. It adds memory slots to the encoder to encode a priori information, called memory-augmented attention. The decoder devises meshed attention. In each decoder layer, the response is computed based on all the outputs of encoder layers, thus it can explore multi-level relationships. The X-LAN model[36] intends to capture the second-order, even higher-order interaction between the objects. It utilizes bilinear pooling techniques to model the second-order cross attention and further employs ELUs to capture infinite order interactions.

Despite the regional features showing outstanding performances, Jiang et al.[37] turned to prove that the CNN features, also called grid features, can be a substitute if extracted by proper backbone networks. It revives the grid features. RSTNet[38] makes use of the grid features and leverages the Transformer model, incorporating geometry embedding for grids and adaptive attention for visual and non-visual words. Note that it designs a BERT-based[39] language model to learn linguistic representations. Likewise, DLCT[40] uses both grid and region features, while building an alignment graph to link them. Moreover, it integrates multiple techniques such as absolute and relative positional encoding, and designs several attention modules to capture object intrinsic properties and latent spatial and semantic relationships.

Graph structure

Exploring relationships in images is a critical process in visual understanding, while the graph can be a logical and reasonable representation. Thus, some researchers consider using graphs to model the latent relationships.

Yao et al.[41] utilized Graph Convolutional Networks (GCN) to encode the object interactions, both spatial and semantic. A pre-trained classifier is applied for predicting the relation of the given object pair, which is used to encode semantic relationships. The spatial relations are determined by the geometry properties of proposals, including IoU, distance, and angle. After that, Yao et al.[42] augmented the encoder by parsing image contents into tree structures, varying from region-level to instance-level. The root node is the whole image, and the region-level and instance-level nodes are acquired from Faster R-CNN[26] and Mask R-CNN[43]. The region-level nodes follow the coarse-to-fine paradigm. Finer small objects are the children of the coarse object.

Scene graph[44,45] represents the relationships between objects, from object pairs to word triplets (*subject-predicate-object*). Yang et al.[46] made use of scene graphs for visual encoding. It parses dependency trees from sentences and builds scene graphs based on objects, attributes, and relationships. In addition, the model learns a dictionary to help reconstruct sentences, incorporating human language inductive bias. Nguyen et al.[47] proposed SG2Caps utilizing only scene graph labels for captioning. In order to bridge the gap between visual and textual scene graphs, it adopts Human-Object Interaction (HOI) inference to augment partial graphs involving the "person" category, imitating human language inductive bias.

Other methods

There are some other attempts to improve the performance. To make the generated caption more human-like and natural, the Generative Adversarial Network (GAN) is employed in some models[48–50]. Usually, these methods leverage a general captioning model at the first stage, and use GAN to discriminate whether the captions are human-like and appropriate. Some approaches extract part-of-speech information to restrict and guide the generation step[49,51,52]. Some methods are trained in a multi-task learning manner, integrating other computer vision tasks that can help the model describe images[53,54]. For example, the model is trained to classify objects in parallel, thus it can be able to learn category-aware representations.

1.1.3 Datasets

Figure 3 shows some typical datasets about image captioning task. Initially, the image-caption pairs are acquired from Flickr website, for example, Flickr8k[55] and SBU[3]. SBU collects over 1 million images with captions from users of Flickr, where one image is only linked to one description. Flickr8k collects 5 captions for each image, after which Flickr30k[56] extends the volume. Flickr30k has been one of the standard datasets of image captioning. Afterwards, Flickr30k Entities[57] links entity words in captions

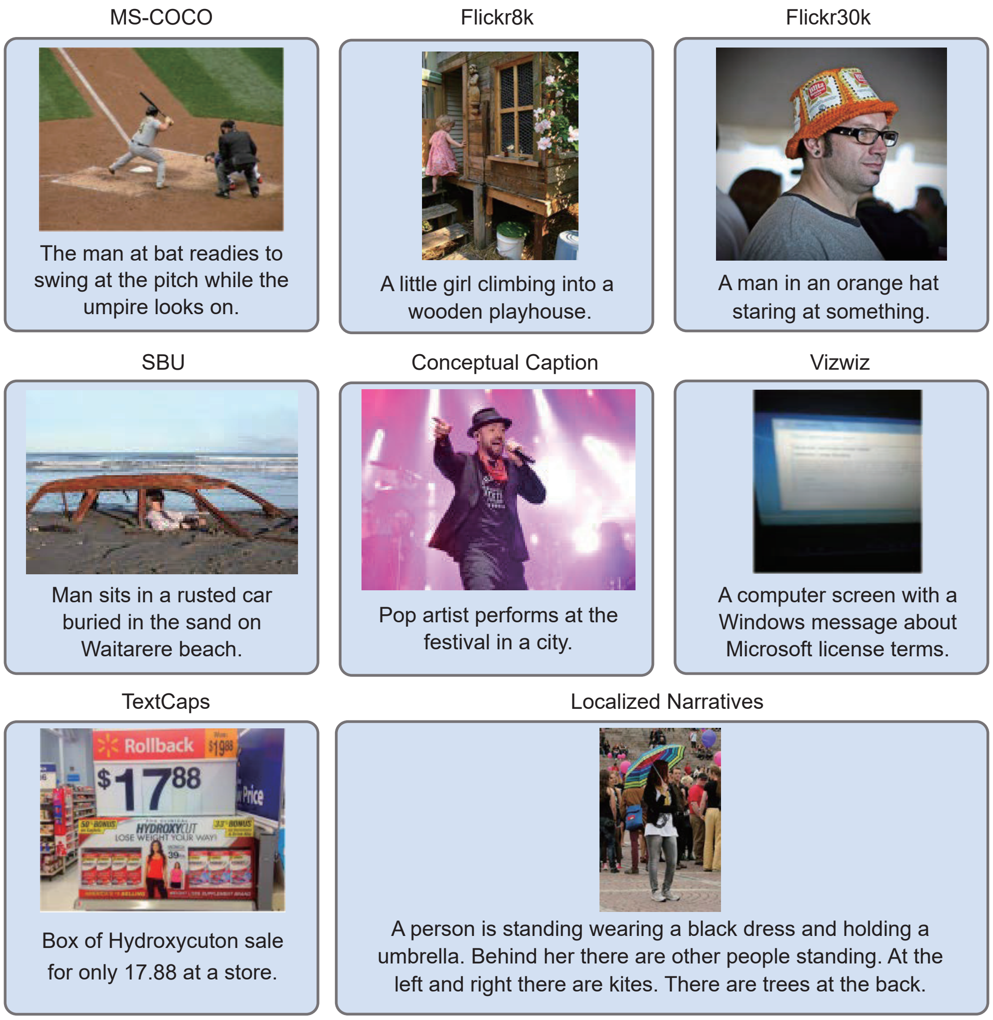

MS-COCO
The man at bat readies to swing at the pitch while the umpire looks on.
Flickr8k
A little girl climbing into a wooden playhouse.
Flickr30k
A man in an orange hat staring at something.
SBU
Man sits in a rusted car buried in the sand on Waitarere beach.
Conceptual Caption
Pop artist performs at the festival in a city.
Vizwiz
A computer screen with a Windows message about Microsoft license terms.
TextCaps
Box of Hydroxycuton sale for only 17.88 at a store.
Localized Narratives
A person is standing wearing a black dress and holding a umbrella. Behind her there are other people standing. At the left and right there are kites. There are trees at the back.

Fig. 3 Some examples in common datasets of image captioning.

with objects in images. Specifically, it has about 244 000 coreference chains associating the same entities in sentences and bounding boxes in images. Flickr30k provides not only the captions but also localization information, thus can be used for some other tasks related to localization, such as visual grounding.

Another standard dataset, Microsoft COCO[58], usually dubbed MS-COCO, is a large-scale dataset containing plenty of scenes and objects in people's daily life. This dataset is built for many computer vision tasks, including object detection and segmentation, and Chen et al.[59] collected descriptions of images and utilized them for image captioning. MS-COCO consists of 123 287 images, each with 5 captions. For reasonable evaluation, Karpathy and Li[7] re-splitted the dataset, assigning 5 000 images for validation, another 5 000 for testing, and the rest 113 287 images for training. Note that MS-COCO also offers an interface[※] for official online testing. Later, Conceptual Caption (CC)[60, 61] gathers descriptions corresponding to an image from websites with a filtering procedure, which is usually used in vision-and-language pre-training.

Except for these event-specific datasets, there are also some stylish datasets with certain specific use. Vizwiz[62] collects images and captions from people who are blind or visual-impaired, aiming to assist disabled people. The images are acquired from blind people, which are closer to real life. TextCaps[63] collects images with text and contains the text in captions. Localized Narratives[64] links image and captions with mouse traces, thus it can ground the words to specific regions in images.

1.1.4 Metrics

Assessing the quality of descriptions is a complex and subjective question, requiring evaluating the fluency and logic of the sentences and judging whether a description is appropriate. Generally, the evaluation metrics are based on calculating the similarity between the generated captions and groundtruths. The prevalent metrics are introduced as follows. It should be noted that some metrics, such as BLEU[65], METEOR[66], ROUGE[67] and SPICE[68], are originally proposed for the evaluation of natural language processing.

BLEU

BLEU[65] is designed for machine translation, focusing on *n-gram* precision in the predictions, where the value of n ranges from 1 to 4. *n-gram* means the n continual words in a sentence. For example, the score of BLEU-1 indicates how many words in the groundtruth are predicted in the generated sentence. BLEU-1 can be regarded as the word-level precision, while BLEU-4 can measure the fluency of captions.

METEOR

METEOR[66] is also a metric for machine translation. Different from BLEU matching the exact words, METEOR considers both precision and recall, and the recall of unigrams plays a more

[※] https://competitions.codalab.org/competitions/3 221 important role when calculating the geometry average. METEOR also uses some extra knowledge sources to expand the synonym wordset, and takes into account the synonyms and words with the same stem.

ROUGE

ROUGE[67] aims to evaluate the quality of summaries, which attend much more to recall. Specially, ROUGE-L is commonly used in visual captioning tasks, and it is designed based on the Longest Common Subsequence (LCS) to measure the similarity between the generated captions and the groundtruth, whose main idea is that two sentences with larger longest common subsequence are more similar.

CIDEr

CIDEr[69] is a special metric for visual captioning. CIDEr computes the cosine similarity between the generated caption and the groundtruth on the sentence level by Term Frequency-Inverse Document Frequency (TF-IDF). The matched *n*-gram is weighted by their frequency, rather than treating all words equally, so some high-frequency phrases may not contribute much to the score. Researchers usually use CIDEr-D as the metric because it restrains the length of sentences and punishes long but verbose captions.

SPICE

SPICE[68] is also designed for visual captioning tasks. Different from the above metrics, SPICE is based on semantic structures rather than *n*-grams. By extracting objects, attributes and relationships in the generated captions, SPICE builds syntactic dependencies trees, and further generates scene graphs. Finally, SPICE scores the similarity of tuples in scene graphs to evaluate the quality of the generated caption. Therefore, SPICE focuses more on semantic contents instead of fluency.

1.1.5 Training strategy

There are two common training strategies for visual captioning tasks: cross-entropy loss and reinforcement learning. The cross-entropy loss supervises the word distribution at each time step. Given target words $y1:T$, the cross-entropy loss is formulated as

$$L_{\mathrm{XE}}(\theta) = -\sum_{t=1}^{T}\log P(y_i|y1:t-1) \qquad (1)$$

However, only using cross-entropy loss to supervise caption model might exist following questions: existing evaluation metrics, such as BLEU, ROUGE-L, METEOR, CIDEr and SPICE, evaluate the sentences from different perspectives, but the cross-entropy loss aims to fairly treat all words, and does not in line with the evaluation metrics. To this end, Rennie et al.[70] proposed Self-Critical Sequence Training (SCST), an optimization approach based on reinforcement learning. The SCST strategy employs the reward mechanism to optimize CIDEr score, hoping that the caption model can generate more fluent and vivid sentences.

1.2 Video captioning

1.2.1 Task definition

Video captioning task aims to generate a description based on the given video, which can describe the details and content in video. Compared with image captioning task, video captioning task needs extra attention to temporal information, which makes it more difficult and challenging. As shown in Fig. 4, video captioning method firstly extracts the frames in video. Then, feature extraction is performed on the serialized image to obtain 2D, 3D, and detection features. Finally, researchers use the structure of encoder-decoder to achieve the prediction of captions.

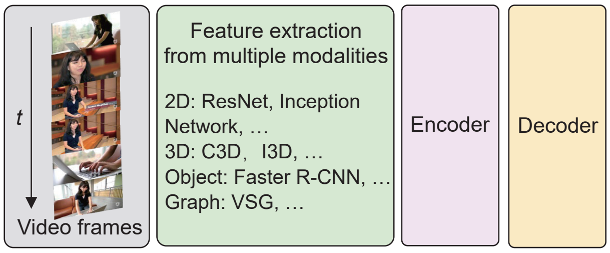


Fig. 4 Framework of video captioning.

1.2.2 Methods

Similar to image captioning task, the methods of video captioning task have also undergone two stages of development. In the early days, researchers design fixed-structure sentence templates to generate descriptions by filling in the form of words. With the emergence of the encoder-decoder structure in recent years, researchers usually use the idea of attention mechanism, scene graph, part-of-speech, etc., for network design in the encoder-decoder structure. At present, the Transformer structure has also begun to gradually develop, and it has also begun to be applied to solve video captioning task. In this section, we will introduce the development of methods for video captioning task in detail from the perspective of method design.

Two stages of development

In the early days of video captioning task, deep learning has not yet emerged, and some traditional methods[71–75] focus on simple sentence generation and with a fixed structure. Some researchers set some sentence templates according to grammatical structures and description habits. By predicting the words to fill the blanks in the sentence template, the descriptions can be generated. In 2012, Barbu et al.[71] proposed the first video captioning system based on the idea: who do what to whom, and where and how they do it, which is a groundbreaking attempt based on the fixed sentence templates. Although the description can be predicted through the fixed sentence templates, the generated sentence is very single and inflexible, which is limited by the structure of the fixed sentence template. In 2013, Das et al.[72] combined the ideas of bottom-up and top-down in image captioning methods to build caption model to capture more important semantic information in videos. Rohrbach et al.[74] modelled the natural language generation problem in video captioning task as a machine translation problem, using visual semantic information as the source language and captions as the target language. In 2014, Rohrbach et al.[75] followed a two-step approach to learn semantic representation of video and generate suitable multi-sentence descriptions. Since 2015, deep learning has flourished. Traditional methods have gradually faded out of researchers' vision.

With the rise of CNN and RNN, the mainstream methods for video captioning task begin to adopt the structure of encoder-decoder, which is separated from the fixed sentence template and makes the captions more flexible and diverse. Venugopalan et al.[76] began to use CNN as encoder to extract visual features of video frames and then use RNN as decoder for caption prediction for the first time. Further, Venugopalan et al.[77] tried to use the idea of sequence-to-sequence in machine-translation task to achieve caption prediction. Since then, the structure of encoder-decoder has been widely adopted.

Attention mechanism

In real life, the information is very rich and complex, but not all the information is beneficial. Therefore, humans usually selectively focus on the effective part of the information, while ignoring the information that brings noise. Researchers call this idea as attention mechanism. Research on attention mechanism is very common in the field of video captioning task. Researchers[78–85] usually use attention mechanisms in the temporal and spatial (regional) dimensions. Yao et al.[86] tried to fuse different visual information in different frames based on the temporal attention module. Some methods[80,83] use a spatial attention mechanism to enhance the important parts within each frame. And some other works[81,82,84,87] achieve both temporal and spatial attention, capturing not only the importance between different frames in time, but also the relationship between different regions within a frame. MGSA[83] proposes an optical flow graph-guided spatial attention mechanism, which uses the optical flow graph to describe the dynamics of behavior to better capture the changes of objects in the video. MARN[88] builds an attention-based RNN decoder with a memory structure to store descriptive information for each word, which can establish correspondence between each word and its relevant visual semantic information. SHAN[89] builds a hierarchical content attention module and a syntactic attention module to adaptively integrate visual features in different frames. SGN[90] utilizes a semantic attention mechanism to achieve alignment between video frames and phrases, which can predict decoded words for different semantic groups.

Scene graph

In order to generate more vivid and detailed captions, exploring the relationship between different objects in videos attracts the attention of researchers. Researchers[91–95] tried to generate finer captions by building a comprehensive graph network structure. OA-BTG[91] builds a bidirectional temporal graph along temporal order to capture the temporal trajectory of each salient object, and then proposes an object-aware aggregation module to learn discriminative representations for different objects. Using object interactions in space and time, Pan et al.[92] built a spatio-temporal scene graph with interpretable links, and further exploit the idea of knowledge distillation to use local object information to obtain higher-quality global scene features. ORG-TRL[93] builds an object relational graph to improve visual representations by enhancing interaction information. At the same time, it takes full advantage of successful external language models to expand language knowledge and generate higher quality caption. D-LSG[94] creatively proposes a conditional graph to add spatio-temporal information into the visual representation and dynamically obtains visual words with higher semantic information. Similarly, Hua et al.[95] designed an object-scene graph model based on object detectors and scene segmenters, which can explore association information in visual representations to capture more comprehensive visual information. These methods of enhancing visual representations with graph structures illustrate that more advanced visual representations can generate better quality captions.

Sentence structure

In addition to rich visual semantic information, sentence structure is also very important to generate sentences that conform to human description habits. Therefore, many researchers[87,89,96–99] conducted extensive research on the Part-of-Speech (POS). Wang et al.[96] constructed a novel gated fusion network and POS sequence generator to control the grammar of generated caption. Hou et al.[97] used the learnable POS tags to address caption bias caused by word imbalance in a multi-tasking manner. SAAT[98] focuses more on changes in video frames, especially verbs, which emphasizes the prediction of interactions between objects. RMN[99] uses three specially designed spatio-temporal reasoning modules to predict POS, thereby realizing more diverse and complex visual feature reasoning fusion. SHAN[89] uses hierarchical content attention and syntax attention to obtain semantic and syntax cues for better visual and sentence-context features fusion. PDA[87] designs an adaptive POS feature extractor, and dynamically adjusts the feature mapping methods of different POS to achieve more accurate captions. Different from the previous method focusing on the POS of a single word, SGN[90] adopts the idea of classifying phrases and predicts the semantic information of the phrases separately.

Other methods

Researchers have also studied video captioning task from other perspectives. Some methods[100,101] employ the idea of Reinforcement Learning (RL) to generate longer and higher quality descriptions. To this day, this reinforcement learning idea is still widely used in visual caption tasks. In addition, the Transformer structure has excellent performance on vision tasks. Lin et al.[102] groundbreakingly proposed an end-to-end framework based on transofmer architecture, which can directly predict the caption of video without using any 2D/3D feature extractors. PAC[103] network aims to generate multiple captions for different perspectives in the video, which puts forward higher requirements for the understanding of video.

1.2.3 Datasets

As shown in Fig. 5, the datasets related to the video captioning task usually contain continuous video frames and corresponding descriptions. In the early days of video captioning task, related datasets are usually relatively simple. In 2011, the first comprehensive dataset MSVD[73] is proposed, which contains 1970 videos on different topics. In 2013, a single-topic dataset YouCook[72] is proposed, which mainly focuses on the cooking process and consists of only 88 videos downloaded from YouTube website. Since the collection and annotation process of video captioning datasets are very complicated, the datasets at this stage are relatively small and simple.

With the continuous development of technology, since 2015, more and more large-scale video captioning datasets have appeared. In 2015, two themed datasets about movie appeare. MPII-MD[104] contains 68 337 clips from 94 Hollywood movies, and M-VAD[105] contains 48 986 video clips from 92 different movies. Since 2015, the scale of the dataset, such as the number of videos and the richness of words, has made significant breakthroughs. In 2016, a new comprehensive video captioning dataset MSR-VTT[106] is proposed. Compared with MSVD, in addition to the number of videos up to 20 000, MSR-VTT also annotates 20 language descriptions and 1 category label for each video. Until today, MSR-VTT is still one of the most challenging datasets for video captioning task.

As more and more researchers focus on video captioning task, researchers no longer only label descriptions for videos, but also label a lot of additional information for video, such as grounded labels of time information, various language annotations. For example, ActivityNet Captions[107] annotates the time grounded labels of each sentence corresponding to the video. VaTEX[108] collects 41 250 videos, including more than 600 human activities and different video contents. In addition to annotating 10 english captions for each video, VaTEX also annotates 10 chinese captions for each video.

1.2.4 Metrics

Similar to image captioning, there are also five common evaluation metrics in video captioning task: BLEU[65], ROUGE-L[67], METEOR[66], CIDEr[69], SPICE[68]. Among them, the first four are the most commonly used in video captioning task. More details

YouCook

The woman has all the ingredients ready for making muffins. She shows all the ingredients like flour, chocolate chips essence eggs etc.

MSVD

A man lights matches and yells.

MPII-MD

They rush out onto the street. A man is trapped under a cart. Valjean is crouched down beside him.

M-VAD

Grasping her hand, SOMEONE <Jack> helps SOMEONE <Rose> onto the bow rail platform.

MSR-VTT

A black and white horse runs around.

VATEX

A person in a comic bear suits falls and rolls around in a moonbounce.

ActivityNet Captions

- An elderly man is playing the piano in front of a crowd.
- A woman walks to the piano and briefly talks to the elderly man.
- Another man starts dancing to the music, gathering attention from the crowd.

Fig. 5 Some examples in common datasets of video captioning.

can be seen in Section 1.1.4.

1.2.5 Traning strategy

As for video captioning task, similar to image captioning, there are also two common training strategies: cross-entropy loss and reinforcement learning. More details can be seen in Section 1.1.5.

1.3 Other generation related tasks

Based on traditional image captioning task, many researchers have proposed multiple similar vision-language tasks. These tasks maintain the prime purpose of image captioning, while adding some other unique objectives and properties.

Dense captioning task[109] aims to combine localization and description of salient regions in images, which is proposed by Johnson et al. in 2016. Note that the descriptions are mainly short phrases rather than long sentences. They propose an end-to-end fully convolutional localization network, incorporating object detection techniques with LSTM networks to generate bounding boxes and short descriptions. Yin et al.[110] argued that regional descriptions generated by RoI features lack contextual coherence with surroundings. Thus, they design an end-to-end framework, leveraging graph structures for feature interaction between RoIs, and then pass refined features to LSTMs for caption generation. Kim et al.[111] took a step further and introduce dense relational captioning task, aiming at generating object relations in images. They propose a Multi-Task Triple-Stream Network (MTTSNet), using part-of-speech tags as a prior to guide word generation. Chen et al.[112] extended this task to 3D scenes and propose 3D dense captioning task. To accomplish this task, their model adopts a relational graph module to extract object relation features. Yuan et al.[113] improved this by exploring Transformer structure and knowledge distillation techniques. It transfers 2D cross-modal knowledge to their 3D student model.

Image paragraph captioning task aims to generate paragraphs with rich semantics and coherent content for the input image. In 2017, Krause et al.[114] contended that a single sentence can only describe images at a coarse level, while dense captioning, which generates separate phrase descriptions, is unable to tell a coherent and unified story. Therefore, they propose image paragraph captioning task, build a new dataset of image-paragraph pairs, and develop a model to generate relatively long descriptions. Their model employs a hierarchical RNN, which is composed of a word

RNN and a sentence RNN. The sentence RNN decides the topic of each sentence, and the word RNN generates word sequences as traditional captioning. The main challenge of this task is to organize sentences in a natural order. Wang et al.[115] proposed DAM model to explore the depth of images. DAM model determines the order of images based on spatial locations so that getting rid of verbose on the same object. Wang et al.[116] designed Convolutional Auto-Encoding (CAE) networks to model the topics on region-level features, and further feed these topic vectors into a two-layer LSTM network. Liu et al.[117] proposed DuelRel model to capture both spatial and semantic relationships, where spatial relations are acquired from a geometry pattern and semantic relations are modeled in a weakly supervised manner.

Grounded image captioning[118] attempt to ground the objects in videos when generating corresponding words. Specifically, when a model generates object words, such as “man” or “car”, it needs to output a bounding box containing the man or the car. Based on regional features, the attention module assigns weights, i.e., a distribution, to these regions, and the region with the highest weight is regarded as the grounding box. Zhou et al.[52] employed knowledge distillation techniques to transfer knowledge from an image-text retrieval model to captioning model, and further adopt the reinforcement learning strategy to improve performance. Ma et al.[119] designed a cyclical learning regimen to refine bounding boxes and captions, resulting in higher performance without grounding supervision. Chen et al.[120] proposed distributed attention network to merge proposals in attention module in order to alleviate the partial grounding problem.

Controllable image captioning task begins to be polpular these years, which hopes to enhance the interpretability of deep learning network through various control signals. In 2019, Cornia et al.[121] proposed controllable image captioning, aiming to improve higher controllability of image captioning models. Given a set or sequence of image regions, which they called the control signal, the model can generate captions following the provided constraints. After that, some researchers change the form of control signals to achieve better controllability and performance. Deng et al.[122] used the length of sentences as the control signal, as longer sentences usually contain more information and details. They adopt the Transformer model and add a length-level vector for captioning. Chen et al.[123] represented human intention via abstract scene graphs, which only contain abstract concepts such as object, attribute, and relationship, to control sentence structure and generate captions as expected. Pont-Tuset et al.[64] built the Localized Narratives dataset with image captions and synchronized mouse tracks, and meanwhile propose to control captioning with the provided mouse traces, resulting that the model can describe image regions in a specified order. Chen et al.[124] extracted semantic roles by a language model, and control the captions to be more human-like and logical via the verb-specific semantic roles.

Novel object captioning task hopes to describe new objects that do not exist in the training set. In 2016, Hendricks et al.[125] pointed out the problem that previous models rely largely on paired image-caption data, thus being unable to describe novel objects in test data. They add object recognition dataset and external text corpora into training data, and train a lexical classifier to recognize novel objects. They re-split the MSCOCO dataset to create a novel object captioning setting, named held-out COCO. After that, Wu et al.[126] decoupled this task as object recognition and image captioning, using placeholders to replace object words in sentences, thus employing a blank-filling paradigm to construct the whole descriptions. In 2019, Agrawal et al.[127] built a new dataset, particularly for novel object captioning, named *nocaps*. They also design metrics to evaluate performances. The whole work is the first large-scale benchmark for novel object captioning. Hu et al.[128] explored the Transformer structure to conduct large-scale pretraining. The VIVO model[128] is trained on a large amount of image-tag pair data and builds a visual vocabulary. Noticeably, this model even surpasses human performance. Unlike previous models relying on object detection models, Vo et al.[129] proposed a Transformer-based end-to-end model, named NOC-REK. They make use of a dictionary and acquire word vectors based on their explanations. In this way, the model can perform vocabulary retrieval based on the similarity of visual features and explanations.

Text-to-image generation[130] has recently attracted ubiquitous attention, which is the reverse of image captioning. This task refers to automatic synthesis of realistic images from text. Models to tackle this task must read a sentence as input and output an image of which the content is described by the input sentence. Reed et al.[131] are the first to implement text to image synthesis based on generative adversarial networks. Zhang et al.[132] proposed a vector quantized diffusion model to remove the unconditional bias and avoid accumulated prediction errors. Utilizing diffusion models[133] in text-to-image generation is a promising avenue to explore.

2 Analysis Related Tasks

2.1 Visual grounding

2.1.1 Task definition

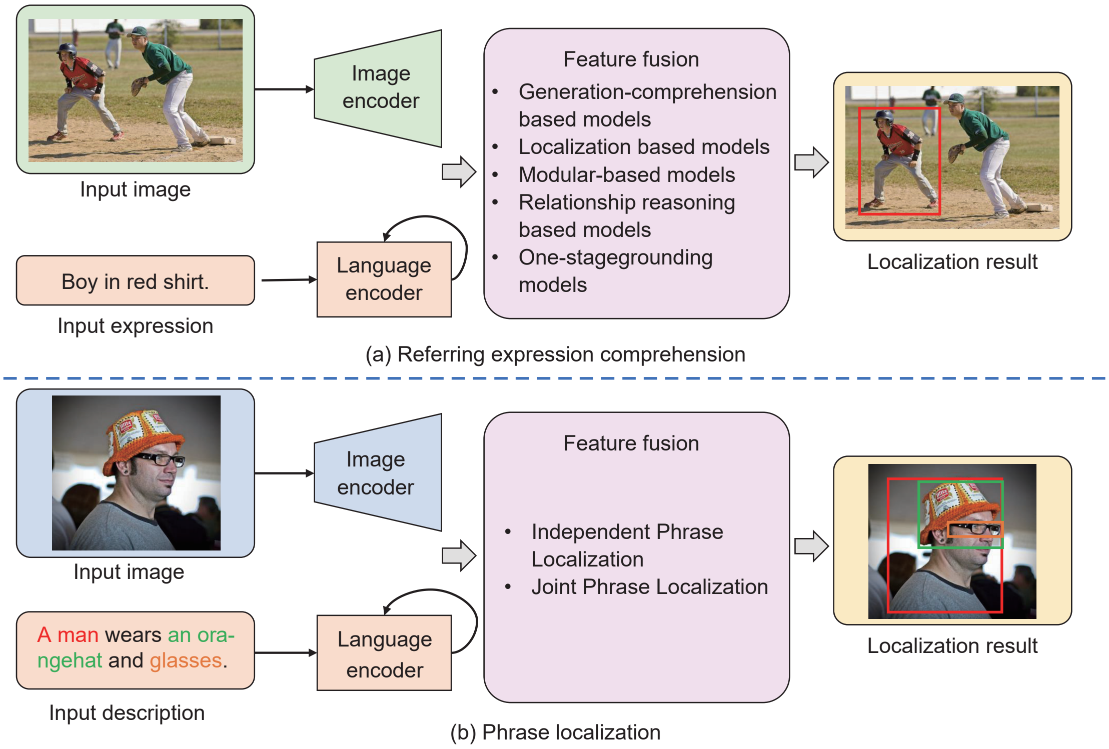
Visual grounding studies how to localize objects in images based on language descriptions. In more detail, visual grounding could be subdivided into Referring Expression Comprehension (REC) and phrase localization. Given an image-expression pair, referring expression comprehension expects to locate the unique target object in the image according to the expression with natural language, as shown in Fig. 6. Similar but different from referring expression comprehension, phrase localization aims at localizing objects in an image corresponding to all noun phrases from a sentence. Visual grounding requires jointly understanding rich image content and complex natural language, which is crucial and challenging for bridging the communication between humans and machines.

2.1.2 Methods

In recent years, a large number of studies have been proposed to move this task forward. Based on whether to locate each phrase in the description independently or in conjunction with the entire sentence/ contextual phrases in the sentence, solutions of phrase localization could be classified into independent localization (I-PL) methods and joint localization methods (J-PL). Localizing each phrase independently does not use information from the whole sentence, so phrase localization in this case is essentially equivalent to REC. For clear organization, we first introduce the methods for REC/I-PL, and then we introduce J-PL methods. Solutions for REC/I-PL could further be roughly divided into five categories: generation-comprehension based models, localization based models, modular based models, relationship reasoning based models and one-stage grounding models.

Generation-comprehension-based models

Benefitting from the ability of CNN and RNN to extract image and text features respectively, early-stage solutions for REC jointly



Input image
Image encoder
Feature fusion
• Generation-comprehension based models
• Localization based models
• Modular-based models
• Relationship reasoning based models
• One-stagegrounding models
Localization result
Boy in red shirt.
Input expression
Language encoder
(a) Referring expression comprehension
Input image
Image encoder
Feature fusion
• Independent Phrase Localization
• Joint Phrase Localization
Localization result
A man wears an ora-ngehat and glasses.
Input description
Language encoder
(b) Phrase localization

Fig. 6 Framework of visual grounding, including referring expression comprehension and phrase localization.

model REC and Referring Expression Generation (REG) within one system, which are called generation-comprehension based method. For REG, given an image I and a region R in the image, the system extracts the image region feature via a CNN and sends the feature into an RNN to generate a sentence S. The system could be represented by $p(S|R,I)$. For REC, given I and a region collect C for I, the system selects the R in the C that maximizes $p(S|R,I)$.

Following the pipeline, Mao et al.[134] proposed a Maximum Mutual Information (MMI) training strategy to constrain the differences of sentences generated by different regions of the image for more discriminative sentences. Similarly, Yu et al.[135] and Liu et al.[136] also focused on the reduction of ambiguity in the generated sentences. But the difference is that Yu et al. design a visual comparison module named Visdif to generate more discriminative visual object features. Liu et al. proposed an explored the usage of the attribute (attr). They define a set of attributes and connect visual features and language features via attributes. MMI, Visdif, and attr can be adopted together to obtain better REC and REG performances.

Considering the usage of visual context, Hu et al.[137] proposed a Spatial Context Recurrent Convnet (SCRC) model which handles the language query, the whole image, the spatial feature, and the visual region feature via two LSTM networks to score candidate regions.

Localization-based models

The same as generation-comprehension based methods, localization based methods also uses CNN and RNN to extract visual features and language features respectively. The difference is that localization based methods directly compute $p(R|S,I)$ for each region R in the C, R that maximizes $P(R|S,I)$ will be chosen as the result.

For the utilization of visual context, Nagaraja et al.[138] proposed to find the context regions via multiple-instance learning (MIL) and maps a language expression to each region and its context region. Considering that MIL based methods could only model one context region, Zhang et al.[139] proposed to use a variational Bayesian framework to learn contexts.

Zhuang et al.[140] proposed a Parallel AttentioN (PLAN) network which contains two attention branches that attends an image and proposals of the image via a language expression, respectively. The attended image feature and region features are fused to predict the final matching result. A similar attention mechanism is also adopted by Deng et al.[141]. They proposed to jointly attend a language query, an image, and proposals by an Accumulated Attention (A-ATT) mechanism.

Modular-based models

The modular-based methods are deepening of localization based methods. Previous works deal with the language expression as a whole, while modular-based methods decompose an expression into different parts and compute matching scores with visual region representations in different ways. The sum of the matching scores of different components is the final matching score between an expression and a visual region.

Compositional Modular Networks (CMNs)[142] contains a language representation module, a localization module, and a relation module. The language representation module divides each expression into the subject, relationship and object. The localization module matches the subject and object language representations with visual region representations, and the relationship module matches relationship language representations with two visual region representations. CMNs mainly focuses on fixed relation modeling through different part of a language expression, though these expressions might have various forms. In view of this, Yu et al.[143] proposed a more general modular based method named Modular Attention Network (MAttNet) for adaptive modeling the input expression by language based attention and visual attention. Based on MattNet, Liu et al.[144] designed an erasing approach named Cross-Modal Attention-Guided Erasing (CM-Att-Erase) for better textual–visual correspondences.

Relationship reasoning based models

In order to disambiguate and precisely describe a target object, referring expressions normally describe the properties of the target itself as well as its relationship to other objects and complex linguistic structure. Thus, only focusing on the target itself is not enough to achieve correct target localization. But it is necessary to carefully model and reason the relationships between objects under the guidance of referring expression. GroundNet[145] is the first approach to utilize syntactic parse of the input expressions to dynamically construct relationships between multiple objects. Based on a parse tree of the referring expression with Stanford Parser, GroundNet dynamically creates a computation graph into a neural architecture that maps the syntactic constituents and relationships in the tree to aid the target localization. NMTREE[146] transforms a Dependency Parsing Trees (DPT)[147] into a neural module tree network for composite reasoning, which assembles three module networks (i.e., Single module for independent objects, Sum and Comp modules for relation reasoning) and accumulates their grounding scores along the tree in a bottom-up fashion.

In addition to tree-based structure, graph-based structure is also appealing for building high-semantic relationship reasoning. LGRAN[148] proposes a language-guided graph attention network, which is composed of a node attention component for attending relevant object regions and an edge attention for capturing intra-class and inter-class relationships. By summarizing these attended informative regions, it enriches the object representation for accurately localizing the described target. Furthermore, DGA[149] designs a differential analyzer to decompose the expression as a sequence of constituent expressions, and performs a multi-step dynamic visual reasoning on cross-modal graph to gradually update the compound object representation at each node. To fully exploit the linguistic structure for subsequent reasoning, SGMN[150] parses the complex expression by a scene graph parser[151], which has a consistent representation with image semantic graph. Jing et al.[152] additionally introduced a permutation loss and a semantic cycle-consistency loss in a self-supervised manner to address one-to-one graph matching between language graph and image graph, which helps to find every node correspondence. To some extent, these methods enable the grounding process to be visualizable and explainable rather than a black box.

One-stage grounding models

The above methods usually regard the referring expression comprehension task as a retrieval problem, which mainly follows a two-stage paradigm. In the first stage, they employ a pre-trained object detector[26, 43, 153] to extract a series of region proposals in an image. In the second stage, they calculate the matching score of each expression-region pair and select the best match proposal region as the final localization result. However, the performance of two-stage methods is limited to the quality of pre-trained object detectors. Once the target object is incorrectly detected or missed, it will not be successfully matched in the second stage. In addition, these methods spend a lot of computation cost to extract plenty of object proposals while only one proposal is finally selected, which is not conducive for real-time referring expression comprehension.

To address these limitations, recent research attention has begun to gravitate toward end-to-end one-stage grounding models that get rid of expensive region proposal extraction. FAOA[154] simply fuses the expression embedding and spatial image features into popular YOLOv3 object detector[155] and directly regresses the target object localization, which shows great potential of one-stage grounding model in terms of both accuracy and speed. On this basis, ReSC[156] proposes a recursive sub-query construction network to deal with long and complex expression queries. The sub-queries are recursively constructed to refine the text-conditional visual features by multiple rounds for reducing the referring ambiguity and reasoning the referred object. LBYL-Net[157] further takes into account the relationship modeling in one-stage grounding network and achieves competitive results over previous two-stage methods, which designs a landmark feature convolution module to encode contextual information and model spatial relations between objects from different directions.

Unlike previous methods based on YOLOv3 detector, RCCF[158] introduces the cross-modality correlation filtering into an anchor-free CenterNet[159] object detector. It regards the encoded language features as filter kernels, and performs correlation filter on the image feature maps to generate a correlation heatmap. The peak value of heatmap means the center point of the predicted target, which combines the regressed object size and coordinates to form the final target localization. Instead of previous rectangular object representation, HFRN[160] proposes a hierarchical fine-grained representation network based on Reppoints[161], which adaptively samples a set of key points based on the image and language to capture more fine-grained object information (e.g., object shape and pose) at local word level and global sentence level. Inspired by powerful Transformer, TransVG[162] establishes intra-modality and inter-modality context information by simply stacking Transformer encoders with self-attention mechanism. In addition, Luo et al.[163] and Li et al.[164] proposed multi-task one-stage network to jointly learn referring expression comprehension and segmentation, which significantly improves the performance by mutual assistance between these two high-related tasks.

Joint phrase localization methods

For phrase localization, unlike I-PL methods, J-PL methods not only use the phrase itself to localize an object, but also consider other phrases in the description or information of the whole description sentence. Richer contextual information is utilized in these approaches, leading to more accurate localization.

Wang et al.[165] proposed a structured matching method for phrase localization. The key idea is that the semantic relation between phrases and visual regions should have consistency. The semantic relation is constructed by extracting visual and language embeddings and computing cosine similarities between them. Chen et al.[166] proposed a strategy using reinforcement learning to exploit contextual information in the descriptions. They proposed QRC-Net[166], penalizing the results predicted by the network corresponding to contextual phrases of the description. SeqGROUND[167] adopts three sets of LSTM sequences to encode image regions, all phrases in sentences, and previously grounded phrase-box pairs, respectively. Abundant context information is extracted for grounding phrase referred objects sequentially. To encode the context, LCMCG[168] constructs two graph networks to generate scene graphs for phrases and visual objects. Two scene graphs are matched via a graph similarity network for seeking the relationship between visual regions and language phrases.

2.1.3 Datasets

In this section, we will introduce popularly used datasets for visual grounding, and some examples of datasets are shown in Fig. 7. In 2014, Kazemzadeh et al.[169] introduced the first large-scale real-world REC dataset ReferItGame (also known as RefCLEF) which contains 19 894 images from ImageCLEF IAPR[170]. 130 525 expressions are annotated to 96 654 objects in these images.

With the emergence and popularity of the MS-COCO[58]

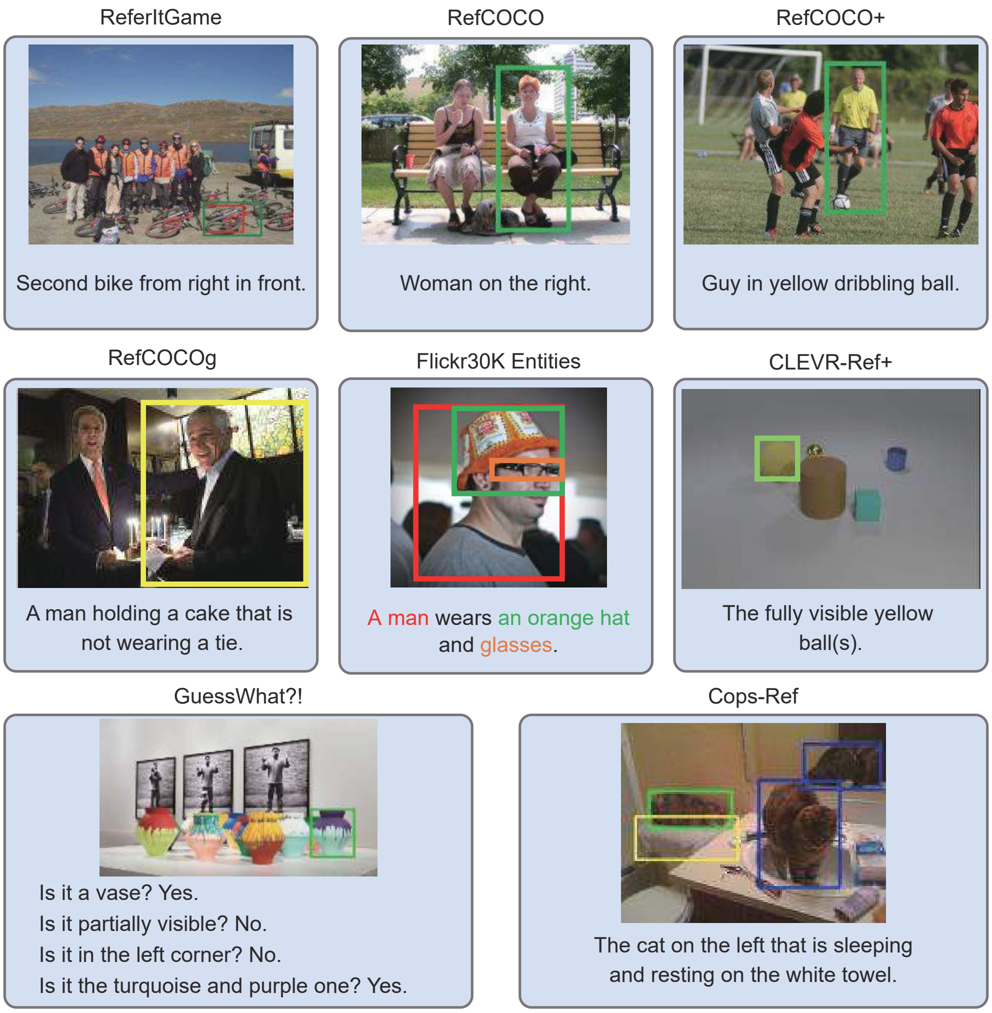

ReferItGame
Second bike from right in front.
RefCOCO
Woman on the right.
RefCOCO+
Guy in yellow dribbling ball.
RefCOCOg
A man holding a cake that is not wearing a tie.
Flickr30K Entities
A man wears an orange hat and glasses.
CLEVR-Ref+
The fully visible yellow ball(s).
GuessWhat?!
Is it a vase? Yes.
Is it partially visible? No.
Is it in the left corner? No.
Is it the turquoise and purple one? Yes.
Cops-Ref
The cat on the left that is sleeping and resting on the white towel.

Fig. 7 Some examples in common datasets of visual grounding.

dataset, researchers constructed RefCOCO[135], RefCOCO+[135], and RefCOCOg[134] based on COCO images. RefCOCO contains 142 210 expressions for 50 000 objects from 19 994 images while RefCOCO+ contains 141 564 expressions for 49 856 objects from 19 992 images. Location words are not allowed to use in RefCOCO+. RefCOCOg contains 104 560 expressions for 54 822 objects from 26 711 images. Expressions in RefCOCOg are generally longer than expressions in RefCOCO/RefCOCO+.

What's more, Flickr30K Entities[57] is generated by abstracting sentencesbasedontheoriginalFlickr30K[56].Thedatasetcontains31 783 images, 158 915 descriptions, and 456 107 expression-targets pairs. It is worth mentioning that Flickr30K could be used for both referring expression comprehension and phrase localization. Compared with locating an object via one expression, Guesswhat?![171] provides a new paradigm that locates a goal by asking a series of questions. This dataset contains 66 537 images and 155 280 dialogues corresponding to 134 073 objects. Besides real-world REC datasets, CLEVR-Ref+[172] is a synthesized image dataset that contains 100 000 images based on CLEVR[173], and each image is corresponded to 10 expressions. Cops-Ref[174] focuses on the process of logical reasoning, with 148 712 expressions and 1 307 885 proposals in 75 299 images.

2.1.4 Metrics

For each prediction of the target object in the image, if the IoU (intersection over union) between the generated bounding box and the groundtruth bounding box is larger than 0.5, the prediction will be seen as a True Positive (TP) one. Otherwise, the prediction will be seen as a False Positive (FP) one. For the model, Top-1 Accuracy (Top-1 Acc) is used to evaluate the performance of a visual grounding model:

$$\text{Top-1Acc} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \tag{2}$$

where N_{TP} and N_{FP} represent the number of TP and FP predictions, respectively.

2.2 Referring segmentation

2.2.1 Task definition

As shown in Fig. 8, the aim of referring segmentation task is to analyze about the objects referred in the linguistic information according to the given natural language expression, and segment the target referent at the pixel level. This task not only requires segmentation methods to parse the relationship between different objects in vision, but also a comprehensive understanding of the semantics expressed in language. Compared with the visual grounding task, which only uses the rectangular detection boxes to roughly locate the referent, the referring segmentation task needs more fine-grained mask prediction for the referent, so it is more challenging and becomes one of the most fundamental and influential cross-modal tasks, and has a wide application prospect in the field of human-robot interaction.

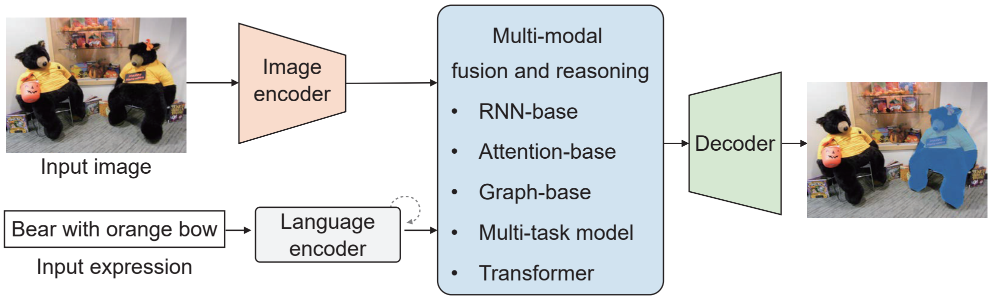


Fig. 8 Framework of referring segmentation.

2.2.2 Methods

For the referring segmentation, how to effectively obtain the semantic information expressed in the language and match it with the corresponding object in the image is the key to this task. Hu et al.[175] first proposed this task and apply a straightforward idea to address this problem, which extracts visual features of images through a pre-trained convolutional network, and extracts global features of text using LSTM, then concatenates the two modal features and generates the segmentation predictions through a full convolution network. With the continuous progress of deep neural network, more and more advanced technologies are applied to referring segmentation task. Next, we roughly divide the existing methods into five categories and introduce them with more details.

RNN-based

Since language expressions are often presented in the form of sequences, and recurrent neural networks (RNN) have achieved great success in natural language processing methods, it becomes a natural idea to apply the form of RNN to referring segmentation task. Different from the original works[175], which fuses global language information with image features, RMI[176] concatenates each word feature with image features and then serves as the input of LSTM, so as to obtain better performance. To obtain more detailed image information, RRN et al.[177] fuses the language features with the top-level image features, and uses ConvLSTM to continuously fuse more low-level visual features recurrently. The recurrent neural network uses fixed parameters to process different inputs, and the text content is often various, so the DMN[178] maps the word features to the parameters of the recurrent network to improve the flexibility and capacity of the network. To obtain more visual cues from the visual-textual co-embedding, STEP[179] designs the Bottom-up network to generate the heat map of the referent, and uses the ConvRNN to recurrently correct the segmentation map through the Top-down method.

Attention mechanism

In the early works, the fusion style of the features from different modalities is just a simple concatenation. However, the visual areas corresponding to different words are often different, and concatenation is difficult to make the network focus more on the object of interest and more important words, so the attention mechanism is widely used in referring segmentation.

To obtain the corresponding relationship between different words and different visual regions, KWA[180] uses cross-modal attention to match keywords for different visual regions and also interactively matches the corresponding visual contexts for different words. CMSANet[181] firstly fuses each word with image features, and uses the self-attention mechanism to fully capture the long-distance dependencies between the two modalities, thereby enhancing multi-modal features. And then it uses gates to achieve multi-level fusion. Similar to KWA, to fuse the features of the two modalities sufficiently, BRINet[182] proposes bidirectional cross-modal attention to achieve visual-guided language and language-guided visual attention, and uses a gate mechanism for bidirectional feature fusion. In terms of presentation form, the text of the language is coarse-grained, while the pixels of the image are more fine-grained. To map modal information in different forms into the same space, EFN[183] proposes an asymmetric co-attention model to obtain cross-modal dependencies, while using a spatial Transformer networks to achieve boundary enhancement to further refine the segmentation results. CGAN[184] uses the attention mechanism to perform multi-modal feature inference for multiple groups and multiple times, and supervises attention at each step to improve the alignment between different modalities.

Graph structure

Neither word features of text nor pixel features of images can be analyzed independently, but more complete semantics must be captured according to the relationship between the current feature and the context. For example, different words have different parts of speech in different sentences, and different visual areas also have complex spatial relationships. It is difficult to capture global information simply by performing pixel-by-word attention matching, so some methods apply graph models to cross-modal global understanding. To match words with different parts of speech and image content, CMPC[185] divides the language content into three parts: entity, attribute and relationship, combines the corresponding visual information to build a graph model, and highlights the target referent through graph analysis. In language text, the importance of different words is different. In order to emphasize more important keywords, LSMC[186] firstly builds a word graph and uses a Dependency Parsing Tree to suppress irrelevant word relationships to obtain a sparser word graph. There are often multiple objects of the same category in the same image. To separate objects from similar objects more accurately, BUSNet[187] builds a language relationship graph and gradually disambiguates similar regions during the analysis process.

Multi-task mechanism

To obtain more accurate segmentation results, it is an important prerequisite to obtain accurate referent location. Therefore, some methods combine the referring expression comprehension task with referring segmentation to improve the performance of segmentation. MattNet[143] utilizes a pre-trained object detector to obtain location information and visual features of different objects in the image, and generates subject, location and relationship features from textual information, and then matches each RoI with text feature to find the most relevant visual region. To bridge the gap between the two tasks of detection and segmentation, MCN[163] proposes a consistency energy maximization module to simultaneously achieve alignment between vision and language and more accurate localization, and then suppresses irrelevant regions for more accurate segmentation results.

Transformer-based structure

With the development of more robust Transformer models, more recent methods are based on Transformers to achieve better performance. To obtain more accurate location, LTS first[188] predicts the location of the referent after inputting language features and image features into Transformer, and then generates a more accurate segmentation mask based on the location information. Due to the diversity of language expressions, it is necessary to understand language information from multiple aspects. Therefore, VLT[189] generates different visual features according to the information of different combinations of language, which are used as the query input of Transformer, and obtains the most matching visual features through Transformer's analysis.

2.2.3 Datasets

Figure 9 shows some common datasets in the field of referring segmentation. The earliest work[175], which proposes the referring segmentation task first, uses the ReferitGame[169] dataset for training and testing the performance of the method. ReferitGame collects 19 894 images from IAPR TC-12[190] and designs a two-player game to obtain 130 525 referring expressions that refer 96 654 object regions. The length of language expressions is relatively short, and the content includes not only people and things, but also stuff, such as sky, grass, forest, etc.

With the popularity of the MS-COCO[58] dataset, some new MS-COCO-based datasets have emerged. Yu et al.[135] leveraged a two-player interactive game[169] to build RefCOCO dataset. The language expressions of RefCOCO consist of an average of 3.5 words, each image contains more than two objects in the same category, and the expressions contain more descriptions of object location information, so it is more difficult than ReferitGame. Similar to RefCOCO, RefCOCO+[135] is also obtained from the MS-COCO dataset. For the content expressed in language, RefCOCO+ avoids the description of the object location, and prefers to describe the target attributes and contextual information around the object.

To obtain more complex language descriptions, different from interactive games, RefCOCOg[134,138] collects 104 560 language descriptions for 54 822 objects based on Amazon Mechanical Turk, and collects 26 711 images from MS-COCO. RefCOCOg contains two data partitioning modes: Google[134] and UMD[138]. The expressions have a longer average length of 8.4 words and contain more complex semantics, making them more challenging.

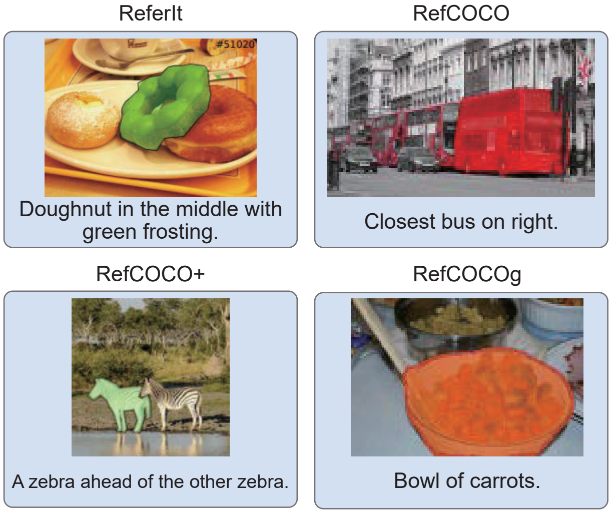

ReferIt
#S1020
Doughnut in the middle with green frosting.
RefCOCO
Closest bus on right.
RefCOCO+
A zebra ahead of the other zebra.
RefCOCOg
Bowl of carrots.

Fig. 9 Some examples in common datasets of referring segmentation.

2.2.4 Metrics

For referring segmentation evaluation, two metrics are commonly used to evaluate the accuracy of the mask predictions: Overall Intersection-over-Union (IoU) and Precision@X (Prec@X). Overall IoU is used to measure the overall performance of the segmentation performance of the method by calculating the total intersection regions over total union regions between the predicted masks and groundtruth. Prec@X is used to analyze the overlap between segmentation results and groundtruth by calculating the percentage of mask preditions with IoU higher than the threshold X, where $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

3 Reasoning Related Task

3.1 Visual question answering

3.1.1 Task definition

Generally speaking, Visual Question Answering (VQA) requires machines to have a human capacity of reading a question and answering it after seeing the given image related to the question. For VQA task, it is not enough for a machine to understand both vision and language modalities in two individual ways. This task requires machines can be at least endowed with treating two totally-discrepant modalities in one common semantic space, which makes this task extremely challenging. Many researchers think VQA is one of the AI-complete problems, which means this problem is equivalent to the core problem of artificial intelligence. As shown in Fig. 10, in present, mainstream methods adopt a double stream architecture. A vision encoder is applied to extract visual feature of an image and this visual information is incorporated in a multi-modal fusion module with the linguistic information mined by a language encoder. The final answer is produced from the fused information by an answer predictor.

3.1.2 Methods

Similar to the development of visual captioning tasks, the research focus of visual question answering tasks also focuses on attention mechanism, graph structure, Transformer structure, and so on.

Attention mechanism

Attention mechanism is the most classic type of VQA methods and methods of remaining types apply attention mechanisms more or less. Attention-based methods focus on devising exquisite attention modules which is used to interact the input image and question by addition, multiplication, etc.

Up-Down[25] is a classical method in visual captioning task and VQA task. At that time, visual attention mechanisms are top-down and operated on CNN grid features. Inspired by human visual system, Up-Down utilizes a modern two-staged object detector Faster-RCNN[26] to extract object-level features and then calculates bottom-up attention. Therefore, Up-Down uses not only top-down attention on grid features but also bottom-up attention on object-level features.

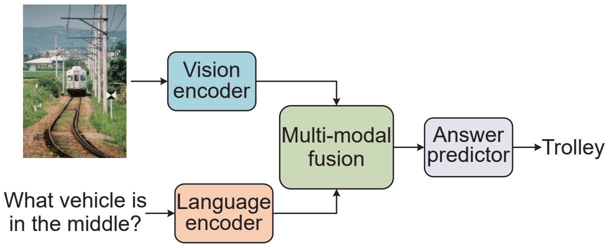


Fig. 10 The framework of visual question answering.

In 2022, Peng et al.[191] built MRA-Net to model object-level relations by attention module. Beyond exploring simple relations between objects, MRA-Net (short for multi-modal relation attention network) captures further the word relations, the binary and the trinary visual relations. The relations of important words in a question can provide information about semantic relational knowledge for generating answers. Although information of pairwise objects’ relation can be contained in binary relations, trinary visual relations are indispensable for capturing more complicated relations involving three objects.

Graph structure

Graph is generally thought of possessing the ability of modeling complicated relations between data. And a tree can be treated as a special case of a graph, which is still with highly expressive ability. Therefore, some methods endeavor to involve graph or tree structures into VQA models.

LCGN[192] focuses on learning relations among objects with the help of graph networks. Contrary to existing methods modeling relations until then, the key novelty of Language-Conditioned Graph Networks (LCGN) is that graph networks are conditioned on language to replace the local appearance-based visual features with the context-aware ones. LCGN is naturally suitable for tasks with an image and a sentence as input, including visual question answering and referring expressions.

What's more, Cao et al.[193] proposed Parse-Tree-Guided Reasoning Network (PTGRN), which possesses the capacity to reason globally on a dependency parsing tree from the question. Benefited from the dependency parsing tree, PTGRN owns an improved interpretability over other pure attention-based methods. The whole architecture consists of attention modules, gated residual composition modules and edge modules, and it is established according to the structure of a dependency tree parsed by an off-the-shelf parser and pruned non-noun leaf nodes. Every noun node is replaced by a PTGRN module, which contains an attention module and two gated residual composition modules, and every two adjacent nodes are connected by an edge module.

Neural-symbolic methods

This is a brand new type of VQA methods, since connectionism represented by neural networks and symbolism represented by logic programs have enormous arguments and conflicts in the past half century. However, both have begun to absorb advantages from one another recently.

In 2018, Neural-Symbolic VQA (NS-VQA)[194] lies at an intersection of deep representation learning and symbolic programs. Deep representation learning does well in perception while symbolic programs are devised for reasoning. Putting their own bright spots together, NS-VQA achieves a near-perfect accuracy of 99.8% in CLEVR. Specifically, NS-VQA firstly parses the scene in an image by Mask R-CNN[43] into a structural scene representation which includes object identifier, size, shape, material, color, and position coordinates. In the meantime, a question parser is used to generate a program. Finally, a programs executor reasons on the structural scene representation and the program to obtain final answer.

Since the performances of neural-symbolic methods in the synthetic dataset CLEVR have reached the upperbound, this work[195] endeavors to adapt this type of methods into a real-world dataset GQA. For real-world images, it is difficult for a neural model to achieve a perfect perception, and faults from the perception will mislead symbolic programs' learning, which results in the model's degradation of performances. To deal with this issue, DFOL-VQA[195] (short for differential first-order logic VQA) proposes a framework to separate the perception from reasoning, and evaluations for the model's reasoning can be accomplished independently from the assessment on the perception. In addition, DFOL-VQA devises a novel top-down calibration technique wherein a differentiable first-order logic is formalized to explicitly disentangle question reasoning from visual perception, which enables the model to answer questions with an imperfect perception.

Causality-based methods

In last two years, causality theory has delved into deep learning, computer vision, multi-modal, etc. Some researchers have utilized causality to mitigate language bias in VQA and achieved a great deal of success.

CounterFactual VQA (CF-VQA)[196] is one of the representative work incorporating causality theory into VQA. CF-VQA establishes a novel counterfactual framework for VQA, on the foundation of which image's and question's causal effects on answers are analyzed and main cause of language bias can be described as a direct effect of the question to the answer in the causal graph. Therefore, CF-VQA mitigates language bias by subtracting the direct effect from the total effect of the question to the answer. Chen et al.[197] proposed CSS to consider causality mechanisms in the sample-level and propose a counterfactual sample synthesis method in order to alleviate language biases. Instead of relying on GANs, Counterfactual Samples Synthesis (CSS) generates enormous samples that are masked key objects in images or important words in questions, and assigns modified groundtruth answers. These two operations on images and questions equip the model with characteristics of visual-explanability and question-sensitivity, respectively.

Transformer-based structure

Following the tremendous success of Transformer-based structure in NLP, the trend has been spreading over computer vision and vision-and-language communities. The structure of Transformer can be widely used for VQA task. In 2019, ViLBERT[198] follows an outstanding work in NLP called BERT, which outperformed the whole states of the art in 11 NLP tasks at that time. This method is composed of a visual stream and a linguistic stream in parallel and interacts information from vision and language modalities using novel co-attentional Transformer layers. ViLBERT focuses on masked multi-modal modelling task and multi-modal alignment prediction task. The former masks around 15% of both words and image region and then requires the model to reconstruct them. The latter presents the model an image-text pair and makes the model predict whether the text depicted the image.

Multi-modal End-to-end TransformER (METER)[199] is a typical fully Transformer-based model, which means that the image encoder is also Transformer-based like text encoder and multi-modal fusion module. This work empirically studies transfomer-based models' design in five aspects: vision encoders, text encoders, multi-modal fusion module, architectural design and pre-training objectives. Moreover, METER is constructed with the study results as guidelines, which surpasses state of the art in VQA2.0 including the previous best region-feature-based model VinVL[200] and best ViT-based model ALBEF[201].

3.1.3 Datasets

Visual question answering is established on top of a rich diversity of datasets. As shown in Fig. 11, the related datasets contain rich scenarios and involve various questions and answers. The overall evolving trend of VQA datasets is from small to large scale, from imbalanced to balanced, from synthetic to real-world, from commonsense to reasoning.

The first mature dataset VQA[202] is proposed in 2015. It

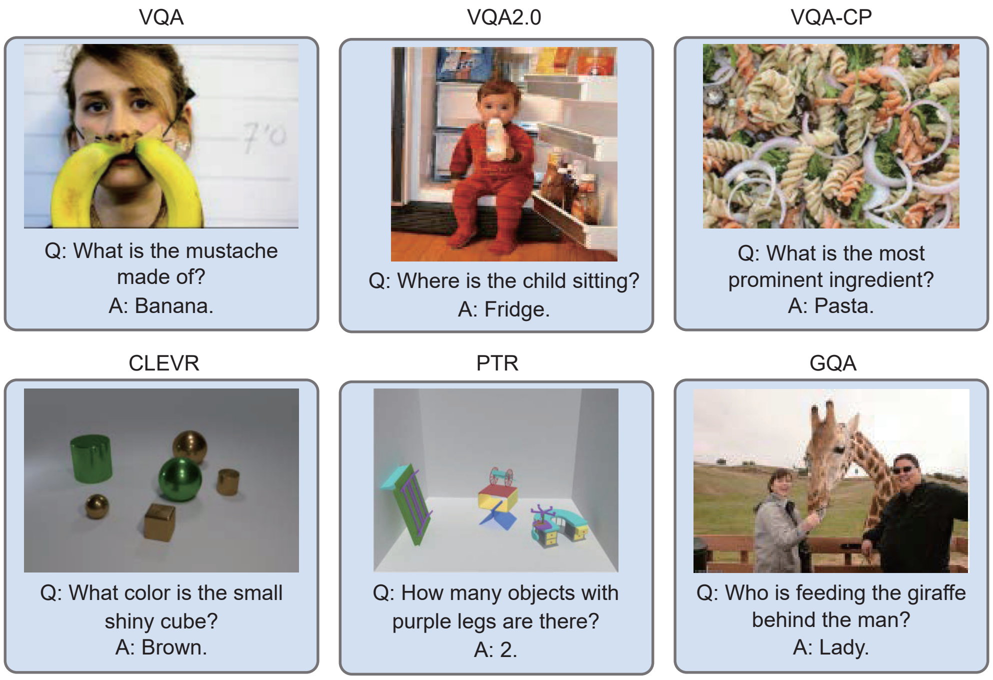

VQA
Q: What is the mustache made of?
A: Banana.
VQA2.0
Q: Where is the child sitting?
A: Fridge.
VQA-CP
Q: What is the most prominent ingredient?
A: Pasta.
CLEVR
Q: What color is the small shiny cube?
A: Brown.
PTR
Q: How many objects with purple legs are there?
A: 2.
GQA
Q: Who is feeding the giraffe behind the man?
A: Lady.

Fig. 11 Some examples in common datasets of visual question answering.

includes 614 163 questions and 7 984 119 answers for 204 721 real images from MS-COCO[58], and 150 000 questions and 1 950 000 answers for 50 000 abstract scenes from a newly-created image set[203] at that time. Many questions in VQA are answered with commonsense knowledge and visual understanding to the scenes. Although VQA dataset has pushed the process of visual question answering, more and more researchers discover that plenty of language biases exist in this dataset, which are much easier for models to catch and make models ignore visual information. VQA2.0[204] is proposed to tackle language biases by balancing VQA dataset. For every binary question, a complementary image is collected in VQA2.0, making a pair of similar images that can result two distinct answers associated with a question.

In order to evaluate models' visual reasoning abilities, including attribute identification, counting, comparison, spatial relationships and logical operations, some researchers build CLEVR[173] in 2017. Additionally, it also contains scene graph annotations and functional program representations for all train and validation images. In 2018, based on VQA and VQA2.0, VQA-CP v1 and VQA-CP v2[205] are proposed. The full name of VQA-CP dataset is Visual Question Answering under Changing Priors dataset. With the popularity of large-scale dataset, GQA[206] is established on top of Visual Genome[207] in 2019. GQA contains 113 018 images (each annotated with a dense scene graph) and 22 669 678 questions (each associated with a functional program). Each question in GQA is categorized into two types: structural and semantic. The structural type includes specifically verify, query, choose, logical reason and compare, while the semantic type contains object, attribute, category, relation and global view.

In recent years, aiming at tackling the problem that existing datasets could not reflect the human-like ability of parsing visual scenes into object parts rather than only individual objects, PTR[208] is proposed recently, which is composed of 70 000 synthetic images and 700 000 questions. Each image is annotated with object categories and part-level semantic instance segmentations as groundtruth.

3.1.4 Metrics

Compared with captioning tasks, a highlighted advantage of VQA is that the evaluation of models' performances is simple and direct. In a long time, VQA Accuracy is almost the only measurement for VQA models.

Question categories of visual question answering include two types: open-ended questions and multi-choice questions. The simple accuracy is suitable for multi-choice questions, because the answer is specific for each question. However, applying the simple accuracy directly to open-ended questions will cause some problems. For example, when the groundtruth for a question is "man", answer "person" is more correct than answer "woman", but in the process of calculating the simple accuracy, the "person" and the "woman" both are classified into wrong answers, which leads evaluations to deflect away true reflections of models' performances.

To relieve the problem of simple accuracy, VQA accuracy is proposed along with the introduction to VQA[202] dataset. Each question accompanies 10 human-annotated answers in VQA dataset. If more than 3 annotators give the same answer, then it is deemed the groundtruth. The VQA accuracy (Acc) is defined as follows:

$$\mathrm{Acc} = \min\left(\frac{n}{3}, 1\right) \tag{3}$$

3.2 Other reasoning related tasks

Visual question answering has many more challenging variations, including Visual Language Navigation (VLN), Embodied Question Answering (EQA), and Visual Dialog (VD).

Visual language navigation[209] requires a robot to understand human instructions, perceive its surrounding environment, and reach the target. Several methods[210,211] explore to solve VLN in the perspective of pre-training. Qi et al.[212] proposed to extract intra-modal and inter-modal high-level semantic features for the sake of performance improvement. Many other works resort to learning graph representation[213], establishing memory-augmented models[214], or utilizing auxiliary tasks[215] so as to boost performance on VLN.

Embodied question answering[216] firstly puts an agent in a random position of a 3D environment and then asks the agent a question that needs to be answered by exploring the environment through an egocentric vision. The crucial characteristic of EQA is that it allows an agent to actively choose which direction it should go and thus what images it will see. Therefore, agents in EQA must have the ability to take actions according to their perception and the given natural language question. Das et al.[217] proposed a hierarchical neural modular controller that contains a master policy to generate subgoals and a few sub-policies to achieve these subgoals. Ilinykh et al.[218] examined the influence of vision perturbation in EQA and find that models answer correctly even though observing perturbed images.

In visual dialog task[219], an agent must have a conversation with a human in natural language, which is specifically required to answer a series of questions about the visual content. Qi et al.[220] proposed two model-agnostic causal principles to boost the performance of many existing methods. To resolve the visual co-reference problem for visual dialog, Niu et al.[221] proposed a novel attention mechanism, dubbed as recursive visual attention. To handle commonsense-required questions, Zhang et al.[222] utilized external knowledge and build a model to reason with multi-structure commonsense knowledge.

4 Vision and Language Pre-training

In multi-modal correlation research, different tasks need to use different datasets for training. These kinds of training strategies need to spend a lot of time on retraining for different vision and language related multi-modal tasks. In recent years, the pre-training research on learning general multi-modal representation from large-scale datasets and applying it to downstream tasks has received extensive attention. At present, the mainstream method adopts the pre-training with fine-tuning paradigm. First, these methods perform pre-training based on a large-scale general vision and language datasets, and then they further use proprietary datasets for fine-tuning to gain task specific knowledge for downstream tasks. With the development of prompt tuning strategy in the field of NLP research, some new multi-modal pre-training related works begin to apply the training mechanism of prompt tuning to downstream tasks, and propose a new pre-training with prompt tuning paradigm.

4.1 Pre-training with fine-tuning

Vision and language pre-training tasks[223–225] usually perform large-scale data training based on vision-text pairs by the tasks of Mask Language Modeling (MLM)[223–225], Mask Region Modeling (MRM)[223, 224], Image-Text Matching (ITM)[223–225], and so on. It aims to extract and fuse information of different modalities to learn the general knowledge representation. Then, these works need to fine-tune downstream tasks based on specific datasets. Pre-training with fine-tuning breaks the barriers between different downstream tasks to a large extent, makes full use of a large number of multi-modal datasets, and greatly improves the performance on downstream tasks while avoiding waste of time and data resources.

Specifically, MLM is a text prediction task, which randomly masks some words in the input text and aims the model to predict the masked words based on context information and visual information. MRM task can be seen as the generalized MLM task into visual modality. Some visual regions are masked, and MRM task aims to predict these masked regions using vision and language information. Different from the MLM, and MRM tasks, ITM task is to judge whether the text is the description of the corresponding image, which randomly replaces the image or text in the image-text pairs with the image or text in other samples when pre-training.

In 2019, Tan et al.[223] thought the understanding of visual concepts and language semantics is of vital importance for vision-and-language reasoning. Thus, they proposed LXMERT with an object relationship encoder and a language encoder to learn the context information in the modalities, and build a cross-modality encoder to learn cross-modality relationships. LXMERT achieves pre-training based on the MLM, MRM and ITM tasks. Unlike LXMERT[223], which uses a dual stream structure, Unicoder-VL[224] also bases on the MLM, MRM and ITM tasks, but uses a single stream structure to achieve vision and language pre-training. It feeds both image and text contents into the same multi-layer Transformer for vision and language cross-modal pre-training. In 2021, ViLT[225] cuts the input image into an image patch sequence, and converts it into feature embedding through linear projection rather than convolution operation. ViLT can solve the problem of low efficiency due to heavily rely on the complex image feature extraction processes, which greatly improves speed and efficiency with better performance in downstream tasks. In 2022, Li et al.[226] proposed the Fast Language-Image Pre-training (FLIP) based on the structure of the CLIP model. It adds a simple mask module to randomly mask some image regions, and then only encode visible image regions, which not only increases the speed by 3.7 times, but also improves the performance in downstream tasks. In order to achieve more downstream tasks by using a single sequence model, Reed et al.[227] built a generalist agent GATO, which has the characteristics of multi-modal, multi-task, multi-embodiment. GATO[227] serializes all modalities data into a flat token sequence to realize different downstream tasks based on the sequence prediction, instead of training for each downstream task separately.

4.2 Pre-training with prompt tuning

As models become larger and larger, the cost of fine-tuning becomes higher and higher. Some researchers hope to avoid introducing additional parameters by adding extra templates and proposing a new prompt tuning paradigm based on pre-training models. With the great breakthrough of GPT-3[228] and PET[229] in the field of NLP, some latest works[230–232] begin to apply prompt tuning to vision and language pre-training.

In 2021, CLIP[233] uses the idea of contrastive learning to build a visual language similarity matrix. It only needs the marks of positive and negative samples to capture features with strong semantics information. Furthermore, based on the constructed prompt template, CLIP achieves outstanding performance on the downstream task zero shot. What's more, CPT[230] uses color-based co-referential markers in image and text as sub-prompt templates to reformulate visual grounding task as a fill-in-the-blank task, so as to minimize the objective form between pre-training and downstream tasks. In 2022, Flamingo[231] regards the single-modal pre-training models of CV and NLP as components and freezes relevant parameters to reduce training costs. It creatively introduces a new fusion module as the prompt templates, and only fine-tune the new fusion module when used on the downstream tasks, which shows good generalization on many downstream tasks such as few shot and zero shot. Liu et al.[232] also aim to solve the problem of inconsistent objective forms of pre-training and fine-tuning tasks, which not only severely limits the generalization of pre-training model to downstream tasks, but also introduces additional parameters and a large number of labeled data for fine-tuning. Liu et al.[232] put forward a DPT model for visual question answering, which converts the given question into a declarative sentence as a prompt template, and then reformulates the answer prediction into MLM and ITM tasks.

5 Future Direction

The field of multi-modal tasks has grown rapidly over the past few years. And more and more researchers have begun to pay attention to the study of vision and language related tasks. Table 1 summarizes the vision and language related typical tasks, datasets, metrics and mainstream methods in recent years from the perspectives of generation, analysis and reasoning. Although some methods of the past few years have achieved a significant performance improvements on standard datasets, some new sub-tasks and how to further enhance the fusion and translation of models across different modalities deserve further study. In this section, we will further discuss the future development trends in vision and language related tasks.

5.1 Analysis of published papers of some representative conference

To more intuitively analyze the development of multi-modal tasks related to vision and language, we further count the published papers from some highly representative conferences, including CVPR†, ACM MM‡, ECCV§, and ICCV¶, since 2015. The statistics are shown in Fig. 12.

As shown in Fig. 12, visual captioning task (image captioning and video captioning), visual question answering task, and referring task (visual grounding and referring segmentation) show a positive development trend in recent years overall, especially the referring task. However, in 2019 and 2020, visual captioning and visual question answering experience a small trough. We believe that these two reasons lead to the decrease. First, the existing research on traditional tasks encounter a bottleneck, and the performance on public datasets reach very high scores, which make some researchers in related fields begin to explore new subtasks. Second, some researchers on multi-modal tasks turn to focus on pre-training of vision and language, resulting in a decrease in downstream tasks. In 2021, vision and language related tasks are in turn boosted with the emergence of variants of traditional tasks and new related datasets.

5.2 Image captioning

Image captioning is a crucial task of visual interpretation and understanding, while complex and tricky integrating computer vision and natural language processing. Despite the increasing development of methods and performances, the generated captions are still far from ideal ones. The main problem is that the model is a black box, which lacks controllability and interpretability. For one, most models tend to generate homogeneous and simple descriptions, with no diversity. However, it is hard for people to control the generation step. For another, people cannot see the internal operations of the model intuitively. Though the attention mechanism can show some clues, the reason that the model predicts some words such as "riding" or "dirty" remains unclear. To these ends, some sub-tasks are proposed for better handling the image captioning task.

Controllable image captioning is first proposed by Cornia et al.[121], which aims at increasing the controllability of the black-box model. Specifically, given a control signal, the model can generate corresponding captions with both high quality and diversity. It is obvious that a proper control signal can make the model able to generate diverse descriptions. However, the design of the control signal is a tricky problem. A control signal with much human knowledge can boost the quality of captions while decreasing the generalization performance. Therefore, how to find a proper control signal is an intriguing problem.

Grounded Image Captioning (GIC) is a task incorporating image captioning with visual grounding, requiring the model to ground the corresponding object in the image when describing it. This task is first proposed by Zhou et al.[118] for grounded video captioning, aiming to alleviate the black-box problem. Notably, GIC only picks the region the model most attends to when generating an object word and outputs the bounding box of the cached feature. There is no bounding box regression process in GIC. Many state-of-the-art approaches[119, 120, 241, 242] attempt to

Table 1 The datasets, metrics and mainstream methods of vision and language related tasks

Category	Task	Dataset	Metric	Mainstream method
Generation	Image captioning	MS-COCO[58], Flickr8k[55] Flickr30k[56], SBU[3] Conceptual Caption[60, 61] Vizwiz[62], TextCaps[63] Localized Narratives[64]	BLEU[65], ROUGE-L[67] METEOR[66], CIDEr[69] SPICE[68]	Karpathy et al.[7], Show and Tell[17] Show Attend and Tell[10] Bottom-up Top-down[25] GCN-LSTM[41], AoA[34] X-LAN[36], $\mathcal{M}^2$ Transformer[35] RSTNet[38], DLCT[40]
	Video captioning	YouCook[104], MSVD[73] MPII-MD[104], M-VAD[105] MSR-VTT[106], VATEX[108] ActivityNet Captions[107]	BLEU[65], ROUGE-L[67] METEOR[66], CIDEr[69] SPICE[68]	HMM[234],Picknet[235], Recnet[236] HTM[237], D-LSTM[238], MARN[88] OA-BTG[91], SibNet[239], POS+CG[96] POS+VCT[97], ORG-TRL[93] SAAT[98], RMN[99], SHAN[89] PDA[87], HRNAT[240]
Analysis	Visual grounding	ReferitGame[169], RefCOCO[135] RefCOCO+[135], RefCOCOg[134, 138] Flickr30K Entities[57], Guesswhat?![171] CLEVR-Ref+[172], Cops-Ref[174]	Top-1 Accuracy	MMI[134], Visdif[135], attr[136] SCRC[137], Neg Bag[138], VC[139] PLAN[140], A-ATT[141], CMNs[142] MAttNet[143], CM-Att-Erase[144] GroundNet[145], NMTREE[146] LGRAN[148], DGA[149], SGMN[150] FAOA[154], ReSC[156], LBYL-Net[157] RCCF[158], HFRN[160], TransVG[162] MCN[163], RefTR[164], Structured Matching[165], QRC Net[166], SeqGROUND[167], LCMCG[168]
	Referring segmentation	ReferitGame[169], RefCOCO[135] RefCOCO+[135], RefCOCOg[134, 138]	Overall IoU Prec@X	CNN-LSTM[175], RMI[176] MattNet[143], KWA[180] CMPC[185], VLT[189]
Reasoning	Visual question answering	VQA[202], VQA2.0[204] VQA-CP[205], CLEVR[173] PTR[208], GQA[206]	Simple Accuracy VQA Accuracy[202]	Up-Down[25], MRA-Net[191], LCGN[192] PTGRN[193], NS-VQA[194] DFOL-VQA[195], CF-VQA[196] CSS[197], ViLBERT[198], METER[199]

†Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
‡Proceedings of the ACM International Conference on Multimedia
§Proceedings of the European Conference on Computer Vision
¶Proceedings of the International Conference on Computer Vision

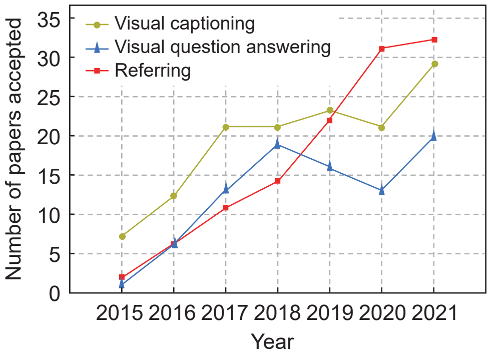


Fig. 12 Number of published papers related to vision and language related tasks of some representative conference in recent years, including CVPR, ACM MM, ECCV, and ICCV

address this problem using attention mechanisms and graph structures, however, it still does not always generate human-like results. The bounding boxes are sometimes deviated, partial, or enlarged. Though there are some efforts to alleviate such problems, it is still worthy of further study to guide and rectify the attention mechanism and help the model generate more grounded image captions.

5.3 Video captioning

Although great breakthroughs have been made in the current video captioning task, there are still some problems and hotspots that need to be further researched. First, most of the current video captioning methods are two-stage, that is, the first stage is to extract visual features for video frames, and the second stage is to predict captions through the encoder-decoder network. This two-stage method requires a lot of time for network pre-training and feature extraction, and requires additional feature storage space. What's more, it also limits the flexibility of the network. Therefore, with the continuous development of technology, it is necessary to explore simpler and more efficient end-to-end video captioning models in the future.

Second, the current evaluation metrics are pretty diverse, and most of them are based on machine translation task. Different evaluation metrics often have different tendencies. It is meaningful to explore the caption evaluation standard, which is more comprehensive, more targeted, and more in line with human language habits.

In addition, many diverse video captioning tasks have emerged in recent years, such as ground video captioning, dense video captioning, multi-perspective video captioning, which require the models to achieve segmentation understanding of videos or localization of key objects, and generate more detailed and diverse paragraph descriptions. These tasks place higher demands and challenges for caption models to achieve understanding and translation between visual and textual modalities.

5.4 Visual grounding

Although the visual grounding task has been made remarkable achievements over the past decade, there is still much development room in the future. Several potential research directions are deserved to be discussed.

(1) Larger and more challenging visual grounding datasets. As reported in Ref. [243], current visual grounding datasets usually exhibit strong dataset bias problem. This indicates the model achieves surprisingly high performance even if there is no input expression or complex linguistic structure, which doubts the reasoning and understanding ability of existing visual grounding models.

In addition, an image only contains limited distracting information (e.g., a few objects for the same category) in these datasets, which is unsatisfactory for real-world complex scenes.

(2) Open-vocabulary visual grounding task. Labeling object expressions and bounding boxes usually require expensive time and resource costs, but the annotations of existing datasets are limited to the real world. Thus, it is necessary to explore the grounding of unseen categories or vocabularies without supervision.

5.5 Referring segmentation

With the gradual development of multi-modal tasks, referring segmentation has achieved significant improvements, but there are still many contents worthy of further research. The referring segmentation task involves knowledge in multiple domains such as computer vision, natural language processing, and semantic reasoning, so the global understanding of information from different modalities is crucial. With the rise of Transformer, more and more cross-modal tasks have made great progress based on Transformer. In future research, we can continue to explore how to better use the emerging Transformer structure to solve the information conversion between different modalities in the referring segmentation task, and achieve greater performance breakthroughs.

In addition, supervised task learning requires a large amount of data annotations. However, the segmentation map annotation is very complex, and it is necessary to finely label the outline of the object, which usually requires expensive time and resource costs. Therefore, the research of few-shot learning and unsupervised learning in the field of referring segmentation task are also worth exploring.

5.6 Visual question answering

Visual question answering has developed rapidly in these years, and has achieved great performance improvement on regular datasets. Nowadays, Transformer-based models have shown dominant performances in vision and language community and this trend will last for several years. However, the VQA task still faces the black-box problem. As the core problem of artificial intelligence, how to judge whether the network really realizes the understanding of images and problems is very important. At present, researchers tend to output attention feature maps to show the focus of the network. However, this is far from enough. In the future, the steps in the network’s answer generation process can be explored by exploiting additional intermediate annotations in the dataset to understand the network’s reasoning process.

In addition, current mainstream methods heavily rely on additional knowledge bases. Pre-training a big vision and language transform-based model consumes tremendous electricity and an enormous amount of time. It's not practicable to retrain such a big model when some new tasks are encountered in applications. On the one hand, researchers can design a memory storage module that can memory external knowledge. On the other hand, finetuning is one choice to avoid retraining from scratch, but sometimes it might cause catastrophic forgetting[244]. Therefore, continual learning is indispensable for big models. Continual VQA might be a promising choice as a future direction on VQA.

5.7 Multi-modal tasks

Multiple multi-modal tasks (e.g., image captioning, visual question answering, visual grounding) have been proposed in recent years, but these tasks are developed independently in their respective field. Each task is specifically designed and tuned by a long time of effort. Therefore, it is very necessary to build a unified multi-modal multi-task framework. It will be greatly advance the development of the multi-modal field by constructing a joint framework to simultaneously handle multiple tasks, which is a critical step towards general intelligence.

6 Conclusion

Our survey investigates the latest development of vision and language related research. Related research fields are summarized based on three perspectives, including generation, analysis and reasoning, so that researchers can have a systematic understanding of vision and language related research fields. In order to introduce vision and language related research fields in more detail, this paper summarizes typical tasks, common datasets, evaluation metrics and the latest research progress from different perspectives, including image captioning, video captioning, visual grounding, and visual question answering. Finally, we further discuss the development of pre-training related research, and summarize the challenges of vision and language related research fields and the future research possibility, which will provide potential directions for future research.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 61831005).

Article History

Received: 4 July 2022; Revised: 15 December 2022; Accepted: 27 December 2022

References

[1] J. Y. Pan, H. J. Yang, P. Duygulu, and C. Faloutsos, Automatic image captioning, in *Proc. 2004 IEEE Int. Conf. Multimedia and Expo*, Taipei, China, 2004, pp. 1987–1990.

[2] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, Every picture tells a story: Generating sentences from images, in *Proc. 11th European Conf. Computer Vision*, Heraklion, Greece, 2010, pp. 15–29.

[3] V. Ordonez, G. Kulkarni, and T. L. Berg, Im2Text: Describing images using 1 million captioned photographs, in *Proc. 24th Int. Conf. Neural Information Processing Systems*, Granada, Spain, 2011, pp. 1143–1151.

[4] Y. Yang, C. Teo, H. Daumé, and Y. Aloimonos, Corpus-guided sentence generation of natural images, in *Proc. Conf. Empirical Methods in Natural Language Processing*, Edinburgh, UK, 2011, pp. 444–454.

[5] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, BabyTalk: Understanding and generating simple image descriptions, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, 2013.

[6] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, ImageNet: A large-scale hierarchical image database, in *Proc. 2009 IEEE Conf. Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248–255.

[7] A. Karpathy and F. F. Li, Deep visual-semantic alignments for generating image descriptions, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 3128–3137.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Proc. 25th Int. Conf. Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 580–587.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, Show and tell: A neural image caption generator, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 3156–3164.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 1–9.

[12] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, Deep captioning with multimodal recurrent neural networks (m-RNN), in *Proc. 3rd Int. Conf. Learning Representations*, San Diego, CA, USA, 2015, pp. 1–17.

[13] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, Long-term recurrent convolutional networks for visual recognition and description, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 2625–2634.

[14] X. Chen and C. L. Zitnick, Mind' s eye: A recurrent visual representation for image caption generation, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 2422–2431.

[15] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, Guiding the long-short term memory model for image caption generation, in *Proc. 2015 Int. Conf. Computer Vision*, Santiago, Chile, 2015, pp. 2407–2415.

[16] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, Image captioning with semantic attention, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 4651–4659.

[17] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. S. Zemel, and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in *Proc. 32nd Int. Conf. Machine Learning*, Lille, France, 2015, pp. 2048–2057.

[18] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, Skeleton key: Image captioning by skeleton-attribute decomposition, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 7378–7387.

[19] W. Jiang, L. Ma, Y. G. Jiang, W. Liu, and T. Zhang, Recurrent fusion network for image captioning, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 510–526.

[20] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. S. Chua, SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 6298–6306.

[21] J. Lu, C. Xiong, D. Parikh, and R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 3242–3250.

[22] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, Top-down visual saliency guided by captions, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 3135–3144.

[23] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, Paying more attention to saliency: Image captioning with saliency and context attention, *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 2, p. 48, 2018.

[24] S. Chen and Q. Zhao, Boosted attention: Leveraging human attention for image captioning, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 72–88.

[25] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, Bottom-up and top-down attention for image

captioning and visual question answering, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6077–6086.

[26] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[27] D. Liu, Z. J. Zha, H. Zhang, Y. Zhang, and F. Wu, Context-aware visual policy network for sequence-level image captioning, in *Proc. 26th ACM Int. Conf. Multimedia*, Seoul, Republic of Korea, 2018, pp. 1416–1424.

[28] L. Ke, W. Pei, R. Li, X. Shen, and Y. W. Tai, Reflective decoding network for image captioning, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 8887–8896.

[29] Y. Qin, J. Du, Y. Zhang, and H. Lu, Look back and predict forward in image captioning, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 8359–8367.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[31] G. Li, L. Zhu, P. Liu, and Y. Yang, Entangled transformer for image captioning, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 8927–8936.

[32] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, Image captioning: Transforming objects into words, in *Proc. 33rd Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2019, pp. 11137–11147.

[33] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, Normalized and geometry-aware self-attention network for image captioning, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10324–10333.

[34] L. Huang, W. Wang, J. Chen, and X. Y. Wei, Attention on attention for image captioning, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 4633–4642.

[35] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, Meshed-memory transformer for image captioning, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10575–10584.

[36] Y. Pan, T. Yao, Y. Li, and T. Mei, X-linear attention networks for image captioning, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10968–10977.

[37] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, In defense of grid features for visual question answering, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10264–10273.

[38] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, and R. Ji, RSTNet: Captioning with adaptive attention on visual and non-visual words, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 15460–15469.

[39] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805, 2019.

[40] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C. W. Lin, and R. Ji, Dual-level collaborative transformer for image captioning, *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, pp. 2286–2293, 2021.

[41] T. Yao, Y. Pan, Y. Li, and T. Mei, Exploring visual relationship for image captioning, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 711–727.

[42] T. Yao, Y. Pan, Y. Li, and T. Mei, Hierarchy parsing for image captioning, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 2621–2629.

[43] K. He, G. Gkioxari, P. Dollár, and R. Girshick, Mask R-CNN, in *Proc.* 2017 *Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 2980–2988.

[44] C. Lu, R. Krishna, M. Bernstein, and F. F. Li, Visual relationship detection with language priors, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, Netherlands, 2016, pp. 852–869.

[45] D. Xu, Y. Zhu, C. B. Choy, and F. F. Li, Scene graph generation by iterative message passing, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 3097–3106.

[46] X. Yang, K. Tang, H. Zhang, and J. Cai, Auto-encoding scene graphs for image captioning, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 10677–10686.

[47] K. Nguyen, S. Tripathi, B. Du, T. Guha, and T. Q. Nguyen, In defense of scene graphs for image captioning, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision*, Montreal, Canada, 2021, pp. 1387–1396.

[48] B. Dai, S. Fidler, R. Urtasun, and D. Lin, Towards diverse and natural image descriptions via a conditional GAN, in *Proc. 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 2989–2998.

[49] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. Forsyth, Fast, diverse and accurate image captioning guided by part-of-speech, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 10687–10696.

[50] P. Dognin, I. Melnyk, Y. Mroueh, J. Ross, and T. Sercu, Adversarial semantic alignment for improved image captions, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 10455–10463.

[51] J. Zhang, K. Mei, Y. Zheng, and J. Fan, Integrating part of speech guidance for image captioning, *IEEE Trans. Multimedia*, vol. 23, pp. 92–104, 2021.

[52] Y. Zhou, M. Wang, D. Liu, Z. Hu, and H. Zhang, More grounded image captioning by distilling image-text matching model, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 4776–4785.

[53] C. Wang, H. Yang, and C. Meinel, Image captioning with deep bidirectional LSTMs and multi-task learning, *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 2s, p. 40, 2018.

[54] W. Zhao, B. Wang, J. Ye, M. Yang, Z. Zhao, R. Luo, and Y. Qiao, A multi-task learning approach for image captioning, in *Proc. 27th Int. Joint Conf. Artificial Intelligence*, Stockholm, Sweden, 2018, pp. 1205–1211.

[55] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, Collecting image annotations using Amazon’S Mechanical Turk, in *Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, CA, USA, 2010, pp. 139–147.

[56] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, *Trans. Assoc. Comput. Linguist.*, vol. 2, pp. 67–78, 2014.

[57] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, in *Proc. 2015 IEEE Int. Conf. Computer Visio*n, Santiago, Chile, 2015, pp. 2641–2649.

[58] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, Microsoft COCO: Common objects in context, in *Proc. 13th European Conf. Computer Vision*, Zurich, Switzerland, 2014, pp. 740–755.

[59] X. Chen, H. Fang, T. Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, Microsoft COCO captions: Data collection and evaluation server, arXiv preprint arXiv: 1504.00325, 2015.

[60] P. Sharma, N. Ding, S. Goodman, and R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in *Proc. 56th Annu. Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 2018, pp. 2556–2565.

[61] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, Conceptual

12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 3557–3567.

[62] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, Captioning images taken by people who are blind, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 417–434.

[63] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, TextCaps: A dataset for image captioning with reading comprehension, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 742–758.

[64] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, Connecting vision and language with localized narratives, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 647–664.

[65] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, Bleu: A method for automatic evaluation of machine translation, in *Proc. 40th Annu. Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA, 2002, pp. 311–318.

[66] S. Banerjee and A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA 2005, pp. 65–72.

[67] C. Y. Lin, ROUGE: A package for automatic evaluation of summaries, in *Proc. Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74–81.

[68] P. Anderson, B. Fernando, M. Johnson, and S. Gould, SPICE: Semantic propositional image caption evaluation, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, Netherlands, 2016, pp. 382–398.

[69] R. Vedantam, C. L. Zitnick, and D. Parikh, CIDEr: Consensus-based image description evaluation, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 4566–4575.

[70] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, Self-critical sequence training for image captioning, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 1179–1195.

[71] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, et al. , Video in sentences out, in *Proc. 28th Conf. Uncertainty in Artificial Intelligence*, Catalina Island, CA, USA, 2012, pp. 102–112.

[72] P. Das, C. Xu, R. F. Doell, and J. J. Corso, A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching, in *Proc. 2013 IEEE Conf. Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013, pp. 2634–2641.

[73] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko, YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition, in *Proc. 2013 IEEE Int. Conf. Computer Vision*, Sydney, Australia, 2013, pp. 2712–2719.

[74] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, Translating video content to natural language descriptions, in *Proc. 2013 IEEE Int. Conf. Computer Vision*, Sydney, Australia, 2013, pp. 433–440.

[75] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele, Coherent multi-sentence video description with variable level of detail, in *Proc. 36th German Conf. Pattern Recognition*, Münster, Germany, 2014, pp. 184–195.

[76] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, Translating videos to natural language using deep recurrent neural networks, in *Proc. 2015 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, CO, USA, 2015, pp. 1494–1504.

[77] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, Sequence to sequence-video to text, in *Proc. 2015 IEEE Int. Conf. Computer Vision*, Santiago, Chile, 2015, pp. 4534–4542.

[78] J. Xu, T. Yao, Y. Zhang, and T. Mei, Learning multimodal attention LSTM networks for video captioning, in *Proc. 25th ACM Int. Conf. Multimedia*, Mountain View, CA, USA, 2017, pp. 537–545.

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

[80] X. Li, B. Zhao, and X. Lu, MAM-RNN: Multi-level attention model based RNN for video captioning, in *Proc. 26th Int. Joint Conf. Artificial Intelligence*, Melbourne, Australia, 2017, pp. 2208–2214.

[81] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, STAT: Spatial-temporal attention mechanism for video captioning, *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 229–241, 2020.

[82] B. Zhao, X. Li, and X. Lu, Cam-RNN: Co-attention model based RNN for video captioning, *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5552–5565, 2019.

[83] S. Chen and Y. G. Jiang, Motion guided spatial attention for video captioning, *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 8191–8198, 2019.

[84] L. Gao, X. Wang, J. Song, and Y. Liu, Fused GRU with semantic-temporal attention for video captioning, *Neurocomputing*, vol. 395, pp. 222–228, 2020.

[85] B. Shi, L. Ji, Z. Niu, N. Duan, M, Zhou, and X. Chen, Learning semantic concepts and temporal alignment for narrated video procedural captioning, in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 4355–4363.

[86] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, Describing videos by exploiting temporal structure, in *Proc. 2015 IEEE Int. Conf. Computer Vision*, Santiago, Chile, 2015, pp. 4507–4515.

[87] L. Wang, H. Li, H. Qiu, Q. Wu, F. Meng, and K. N. Ngan, POS-trends dynamic-aware model for video caption, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4751–4764, 2022.

[88] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y. W. Tai, Memory-attended recurrent network for video captioning, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 8339–8348.

[89] J. Deng, L. Li, B. Zhang, S. Wang, Z. Zha, and Q. Huang, Syntax-guided hierarchical attention network for video captioning, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 880–892, 2022.

[90] H. Ryu, S. Kang, H. Kang, and C. D. Yoo, Semantic grouping network for video captioning, *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, pp. 2514–2522, 2021.

[91] J. Zhang and Y. Peng, Object-aware aggregation with bidirectional temporal graph for video captioning, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 8319–8328.

[92] B. Pan, H. Cai, D. A. Huang, K. H. Lee, A. Gaidon, E. Adeli, and J. C. Niebles, Spatio-temporal graph for video captioning with knowledge distillation, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10867–10876.

[93] Z. Zhang, Y. Shi, C. Yuan, B. Li, P. Wang, W. Hu, and Z. J. Zha, Object relational graph with teacher-recommended learning for video captioning, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 13275–13285.

[94] Y. Bai, J. Wang, Y. Long, B. Hu, Y. Song, M. Pagnucco, and Y. Guan, Discriminative latent semantic graph for video captioning, in *Proc. 29th ACM Int. Conf. Multimedia*, China, 2021, pp. 3556–3564.

[95] X. Hua, X. Wang, T. Rui, F. Shao, and D. Wang, Adversarial

reinforcement learning with object-scene relational graph for video captioning, *IEEE Trans. Image Process.*, vol. 31, pp. 2004–2016, 2022.

[96] B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, and W. Liu, Controllable video captioning with POS sequence guidance based on gated fusion network, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 2641–2650.

[97] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, Joint syntax representation learning and visual cue translation for video captioning, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 8917–8926.

[98] Q. Zheng, C. Wang, and D. Tao, Syntax-aware action targeting for video captioning, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 13093–13102.

[99] G. Tan, D. Liu, M. Wang, and Z. J. Zha, Learning to discretely compose reasoning module networks for video captioning, in *Proc. 29th Int. Joint Conf. Artificial Intelligence*, Yokohama, Japan, 2021, pp. 745–752.

[100] X. Wang, W. Chen, J. Wu, Y. F. Wang, and W. Y. Wang, Video captioning via hierarchical reinforcement learning, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 4213–4222.

[101] L. Li and B. Gong, End-to-end video captioning with multitask reinforcement learning, in *Proc. 2019 IEEE Winter Conf. Applications of Computer Vision*, Waikoloa, HI, USA, 2019, pp. 339–348.

[102] K. Lin, L. Li, C. C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, SwinBERT: End-to-end transformers with sparse attention for video captioning, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 17928–17937.

[103] Y. Bin, X. Shang, B. Peng, Y. Ding, and T. S. Chua, Multi-perspective video captioning, in *Proc. 29th ACM Int. Conf. Multimedia*, China, 2021, pp. 5110–5118.

[104] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, A dataset for movie description, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 3202–3212.

[105] A. Torabi, C. Pal, H. Larochelle, and A. Courville, Using descriptive video services to create a large data source for video annotation research, arXiv preprint arXiv: 1503.01070, 2015.

[106] J. Xu, T. Mei, T. Yao, and Y. Rui, MSR-VTT: A large video description dataset for bridging video and language, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 5288–5296.

[107] B. Ghanem, J. C. Niebles, C. Snoek, F. C. Heilbron, H. Alwassel, R. Khrisna, V. Escorcia, K. Hata, and S. Buch, ActivityNet challenge 2017 summary, arXiv preprint arXiv: 1710.08011, 2017.

[108] X. Wang, J. Wu, J. Chen, L. Li, Y. F. Wang, and W. Y. Wang, VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 4580–4590.

[109] J. Johnson, A. Karpathy, and F. F. Li, DenseCap: Fully convolutional localization networks for dense captioning, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 4565–4574.

[110] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, Context and attribute grounded dense captioning, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 6234–6243.

[111] D. J. Kim, J. Choi, T. H. Oh, and I. S. Kweon, Dense relational captioning: Triple-stream networks for relationship-based captioning, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 6264–6273.

[112] Z. Chen, A. Gholami, M. Nießner, and A. X. Chang, Scan2Cap: Context-aware dense captioning in RGB-D scans, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 3192–3202.

[113] Z. Yuan, X. Yan, Y. Liao, Y. Guo, G. Li, S. Cui, and Z. Li, X-Trans2Cap: Cross-modal knowledge transfer using transformer for 3D dense captioning, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 8553–8563.

[114] J. Krause, J. Johnson, R. Krishna, and F. F. Li, A hierarchical approach for generating descriptive image paragraphs, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 3337–3345.

[115] Z. Wang, Y. Luo, Y. Li, Z. Huang, and H. Yin, Look deeper see richer: Depth-aware image paragraph captioning, in *Proc. 26th ACM Int. Conf. Multimedia*, Seoul, Republic of Korea, 2018, pp. 672–680.

[116] J. Wang, Y. Pan, T. Yao, J. Tang, and T. Mei, Convolutional auto-encoding of sentence topics for image paragraph generation, in *Proc. 28th Int. Joint Conf. Artificial Intelligence*, Macao, China, 2019, pp. 940–946.

[117] Y. Liu, Y. Shi, F. Feng, R. Li, Z. Ma, and X. Wang, Improving image paragraph captioning with dual relations, in *Proc. 2022 IEEE Int. Conf. Multimedia and Expo*, Taipei, China, 2022, pp. 1–6.

[118] L. Zhou, Y. Kalantidis, X. Chen, J. J. Corso, and M. Rohrbach, Grounded video description, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 6571–6580.

[119] C. Y. Ma, Y. Kalantidis, G. AlRegib, P. Vajda, M. Rohrbach, and Z. Kira, Learning to generate grounded visual captions without localization supervision, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 353–370.

[120] N. Chen, X. Pan, R. Chen, L. Yang, Z. Lin, Y. Ren, H. Yuan, X. Guo, F. Huang, and W. Wang, Distributed attention for grounded image captioning, in *Proc. 29th ACM Int. Conf. Multimedia*, China, 2021, pp. 1966–1975.

[121] M. Cornia, L. Baraldi, and R. Cucchiara, Show, control and tell: A framework for generating controllable and grounded captions, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 8299–8308.

[122] C. Deng, N. Ding, M. Tan, and Q. Wu, Length-controllable image captioning, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 712–729.

[123] S. Chen, Q. Jin, P. Wang, and Q. Wu, Say as you wish: Fine-grained control of image caption generation with abstract scene graphs, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 9959–9968.

[124] L. Chen, Z. Jiang, J. Xiao, and W. Liu, Human-like controllable image captioning with verb-specific semantic roles, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 16841–16851.

[125] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, Deep compositional captioning: Describing novel object categories without paired training data, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 1–10.

[126] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, Decoupled novel object captioner, in *Proc. 26th ACM Int. Conf. Multimedia*, Seoul, Republic of Korea, 2018, pp. 1029–1037.

[127] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, nocaps: Novel object captioning at scale, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 8947–8956.

[128] X. Hu, X. Yin, K. Lin, L. Zhang, J. Gao, L. Wang, and Z. Liu, VIVO: Visual vocabulary pre-training for novel object captioning, *Proc. AAAI Conf. Artificial Intelligence*, vol. 35, no. 2, pp. 1575–1583, 2021.

[129] D. M. Vo, H. Chen, A. Sugimoto, and H. Nakayama, NOC-REK: Novel object captioning with retrieved vocabulary from external knowledge, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp.

17979–17987.

[130] S. Frolov, T. Hinz, F. Raue, J. Hees, and A. Dengel, Adversarial text-to-image synthesis: A review, *Neural Netw.*, vol. 144, pp. 187–209, 2021.

[131] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, Generative adversarial text to image synthesis, in *Proc. 33rd Int. Conf. Machine Learning*, New York, NY, USA, 2016, pp. 1060–1069.

[132] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, Vector quantized diffusion model for text-to-image synthesis, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 10686–10696.

[133] F. A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, Diffusion models in vision: A survey, arXiv preprint arXiv: 2209.04747, 2022.

[134] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy, Generation and comprehension of unambiguous object descriptions, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 11–20.

[135] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, Modeling context in referring expressions, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, Netherlands, 2016, pp. 69–85.

[136] J. Liu, L. Wang, and M. H. Yang, Referring expression generation and comprehension via attributes, in *Proc. 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 4866–4874.

[137] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, Natural language object retrieval, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 4555–4564.

[138] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, Modeling context between objects for referring expression understanding, in *Proc. 14th Eur. Conf. Computer Vision*, Amsterdam, Netherlands, 2016, pp. 792–807.

[139] H. Zhang, Y. Niu, and S. Fu. Chang, Grounding referring expressions in images by variational context, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 4158–4166.

[140] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. Van Den Hengel, Parallel attention: A unified framework for visual object discovery through dialogs and queries, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 4252–4261.

[141] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, and M. Tan, Visual grounding via accumulated attention, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7746–7755.

[142] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, Modeling relationships in referential expressions with compositional modular networks, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 4418–4427.

[143] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, MAttNet: Modular attention network for referring expression comprehension, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 1307–1315.

[144] X. Liu, Z. Wang, J. Shao, X. Wang, and H. Li, Improving referring expression grounding with cross-modal attention-guided erasing, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 1950–1959.

[145] V. Cirik, T. Berg-Kirkpatrick, and L. P. Morency, Using syntax to ground referring expressions in natural images, in *Proc. 32nd AAAI Conf. Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conf. and 8th AAAI Symp. Educational Advances in Artificial Intelligence*, New Orleans, LA, USA, 2018, pp. 6756–6764.

[146] D. Liu, H. Zhang, Z. J. Zha, and F. Wu, Learning to assemble neural module tree networks for visual grounding, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 4672–4681.

[147] D. Chen and C. Manning, A fast and accurate dependency parser using neural networks, in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 740–750.

[148] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. van den Hengel, Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 1960–1968.

[149] S. Yang, G. Li, and Y. Yu, Dynamic graph attention for referring expression comprehension, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 4643–4652.

[150] S. Yang, G. Li, and Y. Yu, Graph-structured referring expression reasoning in the wild, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 9949–9958.

[151] S. Schuster, R. Krishna, A. Chang, F. F. Li, and C. D. Manning, Generating semantically precise scene graphs from textual descriptions for improved image retrieval, in *Proc. 4th Workshop on Vision and Language*, Lisbon, Portugal, 2015, pp. 70–80.

[152] C. Jing, Y. Wu, M. Pei, Y. Hu, Y. Jia, and Q. Wu, Visual-semantic graph matching for visual grounding, in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 4041–4050.

[153] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, Feature pyramid networks for object detection, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 936–944.

[154] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, A fast and accurate one-stage approach to visual grounding, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 4682–4692.

[155] J. Redmon and A. Farhadi, YOLOv3: An incremental improvement, arXiv preprint arXiv: 1804.02767, 2018.

[156] Z. Yang, T. Chen, L. Wang, and J. Luo, Improving one-stage visual grounding by recursive sub-query construction, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 387–404.

[157] B. Huang, D. Lian, W. Luo, and S. Gao, Look before you leap: Learning landmark features for one-stage visual grounding, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 16883–16892.

[158] Y. Liao, S. Liu, G. Li, F. Wang, Y. Chen, C. Qian, and B. Li, A real-time cross-modality correlation filtering method for referring expression comprehension, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10877–10886.

[159] X. Zhou, D. Wang, and P. Krähenbühl, Objects as points, arXiv preprint arXiv: 1904.07850, 2019.

[160] H. Qiu, H. Li, Q. Wu, F. Meng, H. Shi, T. Zhao, and K. N. Ngan, Language-aware fine-grained object representation for referring expression comprehension, in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 4171–4180.

[161] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, RepPoints: Point set representation for object detection, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 9656–9665.

[162] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, TransVG: End-to-end visual grounding with transformers, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision*, Montreal, Canada, 2021, pp. 1749–1759.

[163] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, and R. Ji, Multi-task collaborative network for joint referring expression comprehension and segmentation, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10031–10040.

[164] M. Li and L. Sigal, Referring transformer: A one-step approach to

multi-task visual grounding, in *Proc. 35th Conf. Neural Information Processing Systems*, Vancouver, Canada, 2021, pp. 19652–19664.

[165] M. Wang, M. Azab, N. Kojima, R. Mihalcea, and J. Deng, Structured matching for phrase localization, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, Netherlands, 2016, pp. 696–711.

[166] K. Chen, R. Kovvuri, and R. Nevatia, Query-guided regression network with context policy for phrase grounding, in *Proc. 2017 IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 824–832.

[167] P. Dogan, L. Sigal, and M. Gross, Neural sequential phrase grounding (SeqGROUND), in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 4170–4179.

[168] Y. Liu, B. Wan, X. Zhu, and X. He, Learning cross-modal context graph for visual grounding, *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 11645–11652, 2020.

[169] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, ReferItGame: Referring to objects in photographs of natural scenes, in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 787–798.

[170] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, The IAPR TC-12 benchmark: A new evaluation resource for visual information systems, In International workshop ontoImage,http://thomas.deselaers.de/publications/papers/grubinger_lrec06.pdf.

[171] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, GuessWhat?! Visual object discovery through multi-modal dialogue, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 4466–4475.

[172] R. Liu, C. Liu, Y. Bai, and A. L. Yuille, CLEVR-Ref+: Diagnosing visual reasoning with referring expressions, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 4180–4189.

[173] J. Johnson, B. Hariharan, L. Van Der Maaten, F. F. Li, C. L. Zitnick, and R. Girshick, CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 1988–1997.

[174] Z. Chen, P. Wang, L. Ma, K. Y. K. Wong, and Q. Wu, Cops-ref: A new dataset and task on compositional referring expression comprehension, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10083–10092.

[175] R. Hu, M. Rohrbach, and T. Darrell, Segmentation from natural language expressions, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, Netherlands, 2016, pp. 108–124.

[176] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, Recurrent multimodal interaction for referring image segmentation, in *Proc. 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 1280–1289.

[177] R. Li, K. Li, Y. C. Kuo, M. Shu, X. Qi, X. Shen, and J. Jia, Referring image segmentation via recurrent refinement networks, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 5745–5753.

[178] E. Margffoy-Tuay, J. C. Pérez, E. Botero, and P. Arbeláez, Dynamic multimodal instance segmentation guided by natural language queries, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 656–672.

[179] D. J. Chen, S. Jia, Y. C. Lo, H. T. Chen, and T. L. Liu, See-through-text grouping for referring image segmentation, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 7453–7462.

[180] H. Shi, H. Li, F. Meng, and Q. Wu, Key-word-aware network for referring expression image segmentation, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 38–54.

[181] L. Ye, M. Rochan, Z. Liu, and Y. Wang, Cross-modal self-attention network for referring image segmentation, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 10494–10503.

[182] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, Bi-directional relationship inferring network for referring image segmentation, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 4423–4432.

[183] G. Feng, Z. Hu, L. Zhang, and H. Lu, Encoder fusion network with co-attention embedding for referring image segmentation, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 15501–15510.

[184] G. Luo, Y. Zhou, R. Ji, X. Sun, J. Su, C. W. Lin, and Q. Tian, Cascade grouped attention network for referring expression segmentation, in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 1274–1282.

[185] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, and B. Li, Referring image segmentation via cross-modal progressive comprehension, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10485–10494.

[186] T. Hui, S. Liu, S. Huang, G. Li, S. Yu, F. Zhang, and J. Han, Linguistic structure guided context modeling for referring image segmentation, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 59–75.

[187] S. Yang, M. Xia, G. Li, H. Y. Zhou, and Y. Yu, Bottom-up shift and reasoning for referring image segmentation, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 11261–11270.

[188] Y. Jing, T. Kong, W. Wang, L. Wang, L. Li, and T. Tan, Locate then segment: A strong pipeline for referring image segmentation, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 9853–9862.

[189] H. Ding, C. Liu, S. Wang, and X. Jiang, Vision-language transformer and query generation for referring segmentation, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision*, Montreal, Canada, 2021, pp. 16301–16310.

[190] H. J. Escalante, C. A. Hernández, J. A. Gonzalez, A. López-López, M. Montes, E. F. Morales, L. E. Sucar, L. Villaseñor, and M. Grubinger, The segmented and annotated IAPR TC-12 benchmark, *Comput. Vis. Image Underst.*, vol. 114, no. 4, pp. 419–428, 2010.

[191] L. Peng, Y. Yang, Z. Wang, Z. Huang, and H. T. Shen, MRA-net: Improving VQA via multi-modal relation attention network, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 318–329, 2022.

[192] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko, Language-conditioned graph networks for relational reasoning, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 10293–10302.

[193] Q. Cao, X. Liang, B. Li, and L. Lin, Interpretable visual question answering by reasoning on dependency trees, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 887–901, 2021.

[194] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. B. Tenenbaum, Neural-symbolic VQA: Disentangling reasoning from vision and language understanding, in *Proc. 32nd Int. Conf. Neural Information Processing Systems*, Montréal, Canada, 2018, pp. 1039–1050.

[195] S. Amizadeh, H. Palangi, A. Polozov, Y. Huang, and K. Koishida, Neuro-symbolic visual reasoning: Disentangling “ visual ” from “ reasoning ”, in *Proc. 37th Int. Conf. Machine Learning*, 2020, pp. 279–290.

[196] Y. Niu, K. Tang, H. Zhang, Z. Lu, X. S. Hua, and J. R. Wen, Counterfactual VQA: A cause-effect look at language bias, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 12695–12705.

[197] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, Counterfactual samples synthesizing for robust visual question answering, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10797–10806.

[198] J. Lu, D. Batra, D. Parikh, and S. Lee, ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language

tasks, in *Proc. 33rd Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2019, pp. 13–23.

[199] Z. Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng, et al. , An empirical study of training end-to-end vision-and-language transformers, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 18145–18155.

[200] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, VinVL: Revisiting visual representations in vision-language models, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 5575–5584.

[201] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, arXiv preprint arXiv: 2107.07651, 2021.

[202] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, VQA: Visual question answering, in *Proc. 2015 IEEE Int. Conf. Computer Vision*, Santiago, Chile, 2015, pp. 2425–2433.

[203] C. L. Zitnick and D. Parikh, Bringing semantics into focus using visual abstraction, in *Proc. 2013 IEEE Conf. Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013, pp. 3009–3016.

[204] Y. Goyal, T. Khot, A. Agrawal, D. Summers-Stay, D. Batra, and D. Parikh, Making the V in VQA matter: Elevating the role of image understanding in visual question answering, *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 398–414, 2019.

[205] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, Don' t just assume; look and answer: Overcoming priors for visual question answering, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 4971–4980.

[206] D. A. Hudson and C. D. Manning, GQA: A new dataset for real-world visual reasoning and compositional question answering, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 6693–6702.

[207] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[208] Y. Hong, L. Yi, J. B. Tenenbaum, A. Torralba, and C. Gan, PTR: A benchmark for part-based conceptual, relational, and physical reasoning, in *Proc. 35th Neural Information Processing Systems*, 2021, pp. 17427–17440.

[209] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 3674–3683.

[210] A. Majumdar, A. Shrivastava, S. Lee, P. Anderson, D. Parikh, and D. Batra, Improving vision-and-language navigation with image-text pairs from the web, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 259–274.

[211] W. Hao, C. Li, X. Li, L. Carin, and J. Gao, Towards learning a generic agent for vision-and-language navigation via pre-training, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 13134–13143.

[212] Y. Qi, Z. Pan, S. Zhang, A. van den Hengel, and Q. Wu, Object-and-action aware model for visual language navigation, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 303–317.

[213] K. Chen, J. K. Chen, J. Chuang, M. Vazquez, and S. Savarese, Topological planning with transformers for vision-and-language navigation, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 11271–11281.

[214] Y. Zhu, F. Zhu, Z. Zhan, B. Lin, J. Jiao, X. Chang, and X. Liang, Vision-dialog navigation by exploring cross-modal memory, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10727–10736.

[215] F. Zhu, Y. Zhu, X. Chang, and X. Liang, Vision-language navigation with self-supervised auxiliary reasoning tasks, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10009–10019.

[216] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, Embodied question answering, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 1–10.

[217] A. Das, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, Neural modular control for embodied question answering, in *Proc. 2nd Conf. Robot Learning*, Zürich, Switzerland, 2018, pp. 53–62.

[218] N. Ilinykh, Y. Emampoor, and S. Dobnik, Look and answer the question: On the role of vision in embodied question answering, in *Proc. 15th Int. Conf. Natural Language Generation*, Waterville, ME, USA, 2022, pp. 236–245.

[219] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, Visual dialog, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 1080–1089.

[220] J. Qi, Y. Niu, J. Huang, and H. Zhang, Two causal principles for improving visual dialog, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10857–10866.

[221] Y. Niu, H. Zhang, M. Zhang, J. Zhang, Z. Lu, and J. R. Wen, Recursive visual attention in visual dialog, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 6672–6681.

[222] S. Zhang, X. Jiang, Z. Yang, T. Wan, and Z. Qin, Reasoning with multi-structure commonsense knowledge in visual dialog, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, New Orleans, LA, USA, 2022, pp. 4599–4608.

[223] H. Tan and M. Bansal, LXMERT: Learning cross-modality encoder representations from transformers, in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, China, 2019, pp. 5100–5111.

[224] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training, arXiv preprint arXiv: 1908.06066, 2019.

[225] W. Kim, B. Son, and I. Kim, ViLT: Vision-and-language transformer without convolution or region supervision, in *Proc. 38th Int. Conf. Machine Learning*, 2021, pp. 5583–5594.

[226] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, Scaling language-image pre-training via masking, arXiv preprint arXiv: 2212.00794, 2022.

[227] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, et al. , A generalist agent, arXiv preprint arXiv: 2205.06175, 2022.

[228] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. , Language models are few-shot learners, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 1877–1901.

[229] T. Schick and H. Schütze, Exploiting cloze-questions for few-shot text classification and natural language inference, in *Proc. 16th Conf. European Chapter of the Association for Computational Linguistics*, 2021, pp. 255–269.

[230] Y. Yao, A. Zhang, Z. Zhang, Z. Liu, T. S. Chua, and M. Sun, CPT: Colorful prompt tuning for pre-trained vision-language models, arXiv preprint arXiv: 2109.11797, 2022.

[231] J. B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., Flamingo: A visual language model for few-shot learning, arXiv preprint arXiv: 2204.14198, 2022.

[232] Y. Liu, W. Wei, D. Peng, and F. Zhu, Declaration-based prompt tuning for visual question answering, in *Proc. 31st Int. Joint Conf. Artificial Intelligence*, Vienna, Austria, 2022, pp. 3264–3270.

[233] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. , Learning transferable visual models from natural language supervision, in *Proc. 38th Int. Conf. Machine Learning*, 2021, pp. 8748–8763.

[234] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan, Hierarchical memory modelling for video captioning, in *Proc. 26th ACM Int. Conf. Multimedia*, Seoul, Republic of Korea, 2018, pp. 63–71.

[235] Y. Chen, S. Wang, W. Zhang, and Q. Huang, Less is more: Picking informative frames for video captioning, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 367–384.

[236] B. Wang, L. Ma, W. Zhang, and W. Liu, Reconstruction network for video captioning, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7622–7631.

[237] Y. Hu, Z. Chen, Z. J. Zha, and F. Wu, Hierarchical global-local temporal modeling for video captioning, in *Proc. 27th ACM Int. Conf. Multimedia*, Nice, France, 2019, pp. 774–783.

[238] Y. Zhu and S. Jiang, Attention-based densely connected LSTM for video captioning, In *Proc. 27th ACM Int. Conf. Multimedia*, Nice, France, 2019, pp. 802–810.

[239] S. Liu, Z. Ren, and J. Yuan, SibNet: Sibling convolutional encoder for video captioning, in *Proc. 26th ACM Int. Conf. Multimedia*, Seoul, Republic of Korea, 2018, pp. 1425–1434.

[240] Y. Lei, Z. He, P. Zeng, J. Song, and L. Gao, Hierarchical representation network with auxiliary tasks for video captioning, in *Proc. 2021 IEEE Int. Conf. Multimedia and Expo*, Shenzhen, China, 2021, pp. 1–6.

[241] F. Liu, X. Ren, X. Wu, S. Ge, W. Fan, Y. Zou, and X. Sun, Prophet attention: Predicting attention with future attention, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 1865–1876.

[242] W. Zhang, H. Shi, S. Tang, J. Xiao, Q. Yu, and Y. Zhuang, Consensus graph representation learning for better grounded image captioning, in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, pp. 3394–3402, 2021.

[243] V. Cirik, L. P. Morency, and T. Berg-Kirkpatrick, Visual referring expression recognition: What do systems actually learn? arXiv preprint arXiv: 1805.11818, 2018.

[244] R. M. French, Catastrophic forgetting in connectionist networks, *Trends Cogn. Sci.*, vol. 3, no. 4, pp. 128–135, 1999.