

# Visual to Text: Survey of Image and Video Captioning

Sheng Li , *Member, IEEE*, Zhiqiang Tao , Kang Li, and Yun Fu , *Fellow, IEEE*

**Abstract**—Visual data such as images and videos are easily accessible nowadays, and they play critical roles in many real-world applications like surveillance. This raises a series of technological demands for automatic visual understanding and content summarization, which has guided the research community to move towards a better achievement of such capabilities. Meanwhile, it presents the big challenge of semantic understanding of video content and automatically translating them into human language. When developing such automatic translation systems, one critical issue is how to bridge the gap between low level features and high level semantic information. Furthermore, as a large amount of videos are captured under unconstrained conditions by nonprofessional users, this issue becomes even more serious. Therefore, brand new sets of technologies are required to address these difficulties and narrow the semantic gap effectively. These thoughts drive us to survey the complete state-of-the-art techniques in the visual to text topic. Existing methods, popular datasets, technical difficulties, and promising future directions are discussed systematically. In particular, we classify existing methods by their mechanism to link visual information (including both images and videos) and text descriptions, and emphasize the latest advances on deep learning based approaches. The quantitative evaluations of representative approaches on benchmark dataset are also presented and discussed. Finally, we provide with the promising research directions on this topic.

**Index Terms**—Visual to text, image and video captioning, content understanding, text description, summarization.

## I. INTRODUCTION

**N**OWADAYS visual data, such as images and videos, can be collected quickly and cheaply, which bring plentiful information for addressing real-world problems in many domains such as health care and public surveillance. For instance, consumer-grade video is becoming abundant online, and it's much easier than before to search and obtain any type of visual contents. This raises urgent technological demands for automatic visual understanding and content summarization,

which has guided the research community to move towards a better achievement of such capabilities. The availability of vast amounts of text gave a huge boost to the Natural Language Processing (NLP) research community, which was critical in order to organize the amount of information that had suddenly become available. The above-mentioned visual material is set to do the same for intelligent analysis, and we argue that techniques that can transform visual contents into accurate and concise textual descriptions will be a major goal in organizing these information. We call techniques under this topic “Visual to Text”.

The long-standing problem in computer vision and artificial intelligence is the semantic gap between low-level visual data and high-level abstract knowledge. Visual to Text is a major technique that could bridge such a semantic gap in many real-world applications, such as video surveillance systems, visual assistive systems, etc. Although many classical computer vision approaches for classification or detection have shown promising results in some of the aforementioned applications, they usually generate partial and unstructured outputs, such as bounding boxes and object labels in a video frame. These methods provide us with semantic primitives, and they can be considered as the basic steps in “visual to text”. On the other hand, the natural language generation based end-to-end visual to text techniques directly produce sentences for describing the visual observations, which is much easier for understanding.

### A. Challenges for Visual to Text

Humans can easily categorize and describe a visual scene in natural language. However it is still a difficult task to teach machine to do the same thing. Machines are able to recognize the human activities in videos to a certain extent [1], but the automatic description of visual scenes has remained unsolved. Moreover, while action recognition is a well-studied problem in the computer vision community, automatic understanding of activities, especially for the complex and long-term human activities [2], is still challenging. From a linguistic perspective, activity recognition is about extracting semantic similarity between human actions represented by verb phrases, and transforming visual information to semantic text is analogous to grounding words in perception and action [3].

There is one key difference between visual and textual modalities: abstraction level. The data form of the visual content is an image or video that contains a specific theme, or more specifically an object, a scene, an event, an activity, etc. In contrast, the basic data form of the textual contents are words, which are strings of characters. Although a word may also describe a specific object or activity, it only provides a high level label

Manuscript received June 28, 2018; revised December 3, 2018; accepted December 16, 2018. Date of publication January 28, 2019; date of current version July 22, 2019. This work was supported in part by the National Science Foundation, Information and Intelligent Systems Award 1651902 and in part by the U.S. Army Research Office Award W911NF-17-1-0367. (Sheng Li and Zhiqiang Tao contributed equally to this work.) (Corresponding author: Sheng Li.)

S. Li is with the Department of Computer Science, University of Georgia, Athens, GA 30602 USA (e-mail: sheng.li@uga.edu).

Z. Tao is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: zqtao@ece.neu.edu).

K. Li is with the Northeastern University, Boston, MA 02115 USA (e-mail: kongkong115@gmail.com).

Y. Fu is with the Department of Electrical and Computer Engineering and the Khoury College of Computer and Information Sciences, Northeastern University, Boston, MA 02115 USA (e-mail: yunfu@ece.neu.edu).

Digital Object Identifier 10.1109/TETCI.2019.2892755

of a concept objects, activities, abstracting from the real world, which may imply many aspects of information from different modalities of human perception, such as five sense of sight, smell, hearing, taste and touch. Visual information formed by sight is just one of them. Although named entities would provide more details about a specific object or event, an image or a video containing that object or event is still more vivid than the pure label or concept. These differences, vivid visual representation versus highly abstracted text representation, is what defines the major challenges in machine vision and is also a key topic in natural language processing. In computer vision we are interested in constructing and learning models which can characterize images or videos by recognizing their categories or other high level features. In natural language processing, we usually encounter the inverse challenges to parse a language description by identifying connotation and denotation of the sequence of words. These challenges arise because languages are directly related concepts rather than the lossless recording of objects or activities in the real world.

The extraction of labels for objects and activities have been extensively studied in the past decade, which form the foundations of visual-to-text techniques. The focus of this survey is one step past classification problem, or category-level of textual description. We want to focus on more attractive tasks, such as learning how to generate detailed scripts for images and video contents automatically.

The major challenges in “visual to text” include:

- **Fine grained natural descriptions:** Recognizing the fine details of visual contents as well as the interactions of objects is a challenging task. The biggest challenge here is the subtleness of the action units. Sometimes they are not visible, or hard for vision techniques to detect. For instance, unclear unit boundaries and occlusions of interactive objects presents other difficulties to accurately decode the intention of the human activities in a video.
- **Intermediate representation learning:** Learning mid-level representations between visual domain and natural language domain is a key problem in visual to text techniques.
- **Recounting of visual contents:** Even if we were to somehow recognize many semantic elements that appear throughout the visual data, it is hard to rank the importance of them with accordance to theme of the image or video that may guide us to generate more relevant textual descriptions. Also, we need to find out how much detail we are looking to recount and what type of language complexity is to be applied.
- **Benchmark datasets with rich text:** To automatically evaluate generated language descriptions for visual contents, we need standard datasets for evaluating new methods and algorithms. Sentence-level annotations that are aligned to the image and video are a basic requirement.

This survey aims to present a comprehensive overview of both traditional natural language generation models and recent deep learning based techniques, targeting visual to text.

## B. Organization

The rest of the paper is organized as follows. In Section II, we present the foundations of visual to text techniques, including visual features, image annotation, and video recounting/summarization/abstraction. In Section III, we discuss the natural language generation methods. Section IV and V present the recent advances on image captioning and video captioning, respectively. Section VI discusses the empirical performance of the existing visual to text techniques on benchmark datasets. Finally, Section VII presents the conclusion and future perspectives.

## II. FOUNDATIONS OF VISUAL TO TEXT

Though very limited works have been done on translating visual contents, researchers have done a lot of excellent work on other level of intelligence for visual content analysis.

### A. Visual Features

There are a large variety of features, ranging from low-level features directly computed from the input signals, to high-level features that capture spatiotemporal relationships exhibited by the lower-level features. Popular features include a large number of visual descriptors such as histogram of gradients [4], SIFT [5], SURF [6], Gabor textures [7], etc., as well as object and interest point detectors and trackers. The features will be computed on objects and interest points as well as on the whole scene, salient scene segments, and super-pixels [8]. This allows the features to capture both overall scene descriptions and more object-level descriptions. All of these features are indexed to allow rapid retrieval of the relevant segments of the archive based on similarity or range queries.

A visual concept detection module can detect human, objects, actionlets, and scenes from input images or videos. This module is mainly grounded on the cutting-edge techniques for the recognition of visual semantics. It will include following sub-modules: human detection [4], [9], object detection [10], [11], scene detection [12], [13], and action detection [14], [15].

In real-world applications, some problem-dependent cues can also provide complementary information to enrich visual features. For example, social context [16] is able to enhance the face recognition, and thus further facilitate the following caption generation. For another example, a concept words detector [17] is developed to provide high-level information for a wide range of visual to language tasks, such as captioning, retrieval and question answering.

### B. Image Annotation

Image annotation is a fundamental research problem in computer vision. It can be considered as a basic task of translating image to text. The major challenge in translating visual information to text is the so-called semantic gap [18]. From the perspective of artificial intelligence, bridging the semantic gap is equivalent to addressing a visual symbol ground problem [19]. As suggested by S. Harnad [20], the symbol grounding problem

can be decomposed into two levels, categorical representations and symbolic representations. It is worth noting that, symbol grounding can also be linked to situation model theory [3], in order to create so-called “Grounded Situation Models”, which are representations that can be useful for situated embodied agents, such as Robots, when they have to fluidly communicate using Natural Language regarding current, past, and imaginary states of their environment and body.

Image annotation belongs to the level of categorical representations, which has been extensively investigated in the last decades [21], [22]. Some research directions have been investigated to advance the image annotation techniques. For example, to deal with the inconsistency between objects and scenes, several approaches have been proposed to use contextual information for image understanding [23]. Moreover, several projects on collecting large-scale image datasets greatly promote the research progress on image understanding, such as the ImageNet [24]. Besides, there are also some interesting works that aim to investigate the human language development [25], enhance the image labeling via gamification [26] and link the computing resource to the physical world via human-robot cloud [27]. These research efforts not only provide alternative ways for data augmentation, but also explore the connection between vision and text in multiple views.

### C. Video Recounting/Abstraction/Summarization

Recently, the Multimedia Event Recounting (MER) evaluation, as a part of the TREC Vid Evaluation, has been introduced by NIST. MER aims at evaluating video recounting systems that could summarize the key information of the detected events as human-understandable textual descriptions. Some representative works on video content recounting include [28], [29]. Kojima *et al.* developed a system that first detects head and body movements in videos and then generates sentences from the detected case frames [28]. Tan *et al.* detected audio-visual concepts from the Internet videos, and then generated textual descriptions using the rule-based grammar [29].

Video abstraction is also a closely related research topic. Truong *et al.* [30] discussed the video abstraction techniques that target videos from various domains, including documentaries, movies, home recordings, etc. The current video abstraction techniques can be mainly divided into two subcategories, key-frame extraction and skim generation.

Video summarization is usually converted to the problem of key frame extraction [31], where the appearance based features such as colors play a major role in change detection [32]. Some methods also try to employ the overall camera motions [33], [34]. However, these methods simply employed the fixed rules for the entire set of videos. The aforementioned methods are visual summarization that generates a “teaser” for the video. In addition, researchers try to translate video content into human language directly. In this case, bridging the gap between low-level visual features and high-level semantic information is the major challenge. Moreover, beyond the visual to text, many research efforts have been devoted to parsing, generating and translating among humans’ representations (e.g., *the*

*spoken, motor, vision language and etc*), such as the “Poeticon” project [2].

## III. CLASSICAL MODELS: NATURAL LANGUAGE GENERATION FROM VISUAL CONTENTS

Introducing natural language models to visual understanding systems has been a popular research paradigm in the past decade. We can roughly divide the natural language generation based “visual to text” techniques into four different categories, including: (1) language rules or templates based scripts generation, (2) borrowing available descriptions from other visual contents, (3) complementing existing descriptions by visual recognition, and (4) creating a language model to generate descriptions. Table I lists some representative works in each of the four categories.

### A. Language Rules or Templates based Generation

Many works on video recounting utilize manually specified templates or rules to obtain descriptions from a middle-level semantic representation. Nagel [62] presented the Naos system [63] to fill case grammar frames of street surveillance videos by detecting bounding boxes and tracking their relationships. Case frames grammar represents the key elements in a sentence, including location, object, predicate, agent, etc. Further, Nagel and Zimmermann introduced a dialog system in which visual scenes were depicted [62]. Case frames were also used later by Kojima *et al.* who constructed an action concept hierarchy represented by a hierarchical case frame [28].

The following works turned their focuses to real world videos [40]. Arens *et al.* utilized a situation graph for expression of traffic scenes and placed them into templates to produce language descriptions [64]. The DARPA Mind’s eye project incorporates a video corpus annotated with 48 different verbs. Based on a predefined template, they generated text from their detected semantic elements. Similarly, Khan *et al.* recounted video events on the TREC Video summarization competition by designing a language template [39]. Khan and Gotoh proposed a method that can recognize a limited set of six humans activities as well as gender, age, and emotions based on facial features [65]. Then a template filling approach is proposed for sentence generation, which is based on the context free grammar (CFG). Yang *et al.* proposed to use external information like text corpus to enhance the language generation [66]. On the UIUC Pascal Sentence dataset they recognized objects and scenes [67]. Then they added prepositions and activities based on a language model learned from the newswire Gigaword corpus. Given the language model and detected objects, a Hidden Markov Model (HMM) can be used to generate a template based sentence. This category of approaches allow precise generation of language and is effective for a lot of applications in a *limited* domain, especially when the visual content sets, objects or activities, are small and the expected language description are simple.

### B. Exploiting Auxiliary Descriptions

The second category of methods reduces the generation process by borrowing available descriptions from other visual



TABLE I  
CATEGORIZATION OF NATURAL LANGUAGE MODELS FOR VISUAL TO TEXT

Category	Representative Papers
Language rules or templates based scripts generation	IJCV 2002 [28], CVPR 2009 [35], CVPR 2011 [36], ACM MM 2011 [29], ACL 2011 [37], UAI 2012 [38], ICCV-W 2011 [39], ECCV-W 2012 [40], CVPR 2013 [41], ICCV 2013 [42]
Borrowing available descriptions from other visual contents	ECCV 2010 [43], NIPS 2011 [44], CoNLL 2011 [45], CVPR 2014 [46], ACL-W 2014 [47], AAAI 2014 [48]
Complementing existing descriptions by visual recognition	ACL 2010 [49], ICCV 2013 [50], TACL 2014 [51], CVPR 2014 [52]
Creating a language model to generate descriptions	IEEE Proceedings 2010 [53], ECCV 2012 [54], ICCV 2013 [55], IEEE TPAMI [56], IEEE Multimedia [57], COLING 2014 [58], CVPR 2014 [59], AAAI 2015 [60], ACL 2015 [61]

contents locally [44] or globally [43]. By using a Markov Random Field (MRF), intermediate semantic representations of visual elements can be constructed, including scenes, objects, actions, etc [43]. These representations can then be mapped to the sentence set. They borrowed the most relevant sentence from training dataset, by evaluating the similarity to the predicted language description by employing a sentence-semantic representation mapping trained on the UIUC Pascal Sentence dataset. In order to match language description to out-of-vocabulary words, the algorithm utilizes the semantic distance in the WordNet according to the Lin's measure [68], especially for objects and scenes. Similarly, Ordonez *et al.* searched for the most relevant captions in the Captioned Photo Dataset as the description for a test image [44]. By applying a greedy search strategy, they first selected the some most similar images, and then run more advanced detectors for objects, scenes, human, and actions. A linear regression was used to learn optimized weight on the training set [69]. Mason *et al.* presented a data-driven framework for image caption generation [70], which incorporates visual and textual features with varying degrees of spatial structures. This category of approaches would achieve good performance when the additional sources are aligned well with the targeted visual contents. However, the unreliable, low-quality sources may easily degrade the model performance.

### C. Complementing Descriptions by Visual Recognition

In some cases, there are available textual description associated with the image or video at the *testing* phase. The third category of methods can incorporate text labels or descriptions into visual to text models after visual recognition. In order to create descriptions of the tagged locations or buildings through web searching, Aker and Gaizauskas generated image descriptions by using multi-document summarization techniques [49]. Not only rely on textual data, Feng and Lapata proposed to generate captions by representing text and images jointly using the bag-of-words model [71]. They demonstrated that a mixed topic model (i.e., with both textual and visual words) can significantly improve text only model. The captions are constructed by retrieving sentences from news articles. Moreover, the most similar text descriptions based on visual features are also integrated. Finally, attachment probabilities are introduced to model long distance dependencies and grammar structure among different phrases. [37] focused on improving the precision of

automatically generated image captions. They compressed the descriptions by considering both linguistic fluency and consistency of the language and visual elements. The visual elements are obtained by an object recognition model learned on ImageNet [24] and the optimization is implemented by beam search and dynamic programming. On the SBU dataset that contains about one million images, their improved captions are much more relevant and accurate than the original ones. Socher *et al.* [51] designed a recursive neural network model based on dependency trees, which learns latent representations for sentences and images. The key idea is to embed images and sentences into a latent space, and then find the matched counterpart for a given query. For example, given a query image, the proposed model could find a sentence to describe it. The performance of this category of approaches heavily depend on the visual recognition model.

### D. Creating a Language Model to Generate Descriptions

The fourth category of work goes beyond borrowing available descriptions via creating a language model to generate concise textual descriptions [72]. Yao *et al.* utilized a large-scale image dataset with informative annotations, and employed a natural language generation (NLG) method that converts the image parsing results to textual descriptions [53]. In particular, their method consists of four major components: (1) an image parsing engine; (2) an and-or graph based visual knowledge representation; (3) a semantic web; (4) a text generation engine. The image parsing engine converts images or videos into parse graphs. Then the and-or graph based knowledge representation is employed to provide top-down hypotheses during image parsing. The general knowledge embedded in the semantic web is then adopted to enrich the semantic representations. Given semantic representations, the text generation engine can output query-able and human readable descriptions. Their work shows that, a successful image to text system heavily relies on the four components. Fernandez *et al.* inferred the human activities from some predefined conceptual primitives, and then incorporated a natural language model for text generation [73]. In particular, their system extracts geometrical information from videos, and then converts these information into predicates in fuzzy logic formalism. The discourse representation structures are then utilized to facilitate the generation of natural language texts. [74] defines the Topic-Oriented Multimedia Summarization (TOMS)

task based on natural language generation. In particular, given a set of features that are automatically extracted a video (e.g., ASR transcripts and visual concepts), a TOMS system aims to produce a paragraph to summarize the key information contained in the video associated to a particular topic, and also offer proper explanations on the retrieval results. However, their approach can only describe a few pieces of contents which are relevant to a predefined event topic. Also, they do whole-video-level matching by human judges as evaluation metric, which is hard to measure the general quality of content translation, since a single key evidence may directly lead to a good match.

Some other works focus on training a language generation model through an aligned corpus of images with descriptions, in which the advanced language generation methods are incorporated. Kuznetsova *et al.* searched candidate words or phrase according to the scene and object recognition results on the SBU dataset [37]. They generate the relevant and high-quality scripts by applying an optimization algorithm for content planning and language realization. Mitchell *et al.* applied the visual recognition system to predicts ordered noun-phrases on the same dataset [75]. Then, by adding necessary prepositions, predicates, the proposed approach can automatically form novel phrases. Most recently, deep learning methods such as recurrent neural networks have been introduced to train language models for image caption generation [76]–[78].

The key in this category of approaches is how to effectively adapt the language models to the tasks of visual to text. Most of existing works explore the language models when fixing the visual learning model. However, it would be more effective to jointly train the visual and language models.

#### IV. RECENT ADVANCES: FROM IMAGE TO TEXT

The recent advances in training deep neural networks (i.e., also termed as *deep learning*) have significantly promoted the development of image captioning, and pushed towards a tight link between the fields of computer vision (CV) and natural language processing (NLP). Specifically, deep convolutional neural networks (CNNs) [80]–[83], pre-trained on the large-scale image datasets, provide hierarchical and rich feature representations to parse visual world into higher-level semantics; on the other hand, recurrent neural network (RNN) has made great progress for machine translation [84]–[86], which enables to generate readable target sentence conditioned on the semantic information extracted from source sentence. Analogous to machine translation, recent image captioning algorithms mainly adopt such a framework as *image encoder*  $\rightarrow$  *text decoder*, which amounts to the nature of from image to text, i.e., translating image to sentence.

The key point for such above architecture is how to encode representative vision cues from still images, and decode image representations as meaningful sentences. Along this line, a great deal of research efforts have been made to develop various schemes for image encoding and text decoding process.

##### A. Visual-Text Embedding

A very pioneering work that introduces neural network to image captioning was proposed by Kiros *et al.* [72], where

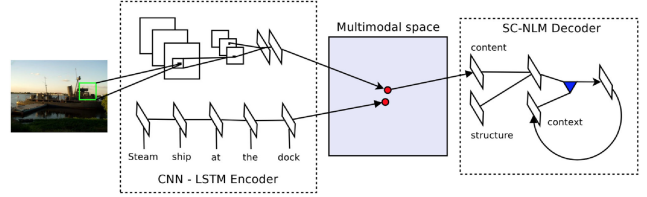


Fig. 1. An example of multimodal embedding with language model decoder. Figure adapted from [79].

they devised a multimodal log-bilinear (LBL) model to work as a feed-forward neural network for predicting words. In this work, the image feature encoded by CNN is formulated as the bias within a LBL model, to provide discriminative visual clue during the caption generation process. Shortly after that, Kiros *et al.* [79] proposed the “encoder-decoder” paradigm (see Fig. 1) to mimic machine translation for the caption generation task, and devised a multimodal neural language model (MNLM) based on their previous work. In MNLM, the encoder seeks for a joint multimodal embedding space that could associate the image representation to its corresponding description; while the decoder is also formulated by a LBL based language model conditioned on the embedding space. The AlexNet [80] and LSTM are used to extract image feature and learn sentence representation, respectively, and a pairwise ranking loss is eventually employed to achieve jointly embedding (among image and sentence).

Similar to [79], Karpathy and Li [88] proposed a deep visual-semantic (DeepVS) alignment model to learn a multimodal embedding space. However, rather than aligning the whole image and its description, they considered the alignment in a more fine-grained level, i.e., they aimed to generate dense captions for image regions. Specifically, the Region-CNN (R-CNN) [89] is leveraged to extract a set of object regions from the image, and thus feature representation for each region is obtained; while one bidirectional RNN (BRNN) model is adopted to characterize the descriptions corresponding to each region. The embedding matrices for both image and text are learned with a max-margin ranking loss and image-sentences pairs. Unlike [79], DeepVS develops language model by using RNNs, which predict next word in a sequence by considering the current word and hidden states from previous time steps. By conditioning the image content on the initial state of the RNN model, it finally delivers a multimodal RNN to tackle with the caption generation.

These multimodal (i.e., *visual-text*) embedding based methods share the similar idea with some previous works, which first find a common embedding space for image and text and then generate human-readable sentence with the visual embeddings. Nevertheless, the difference lies at the way for caption generation. Recent methods mainly develop a language model with powerful neural networks, which enable to generate novel descriptions for image. Besides captioning, this embedding approach could also benefit the tasks of image annotation and image retrieval. However, the joint embedding space is highly depended on the ranking function, which may demand carefully tuning.

Instead of encoding image with visual-text embedding, a more natural way for image captioning is to directly decode the image representation as sentences.

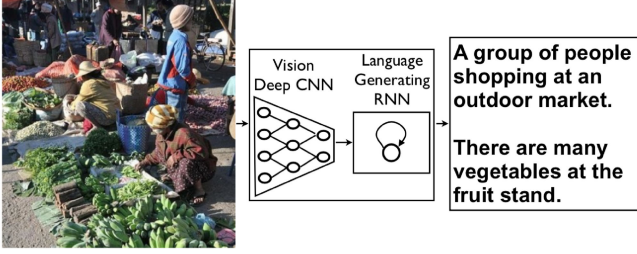


Fig. 2. Overview of encoder-decoder framework for image captioning. Figure adapted from [87].

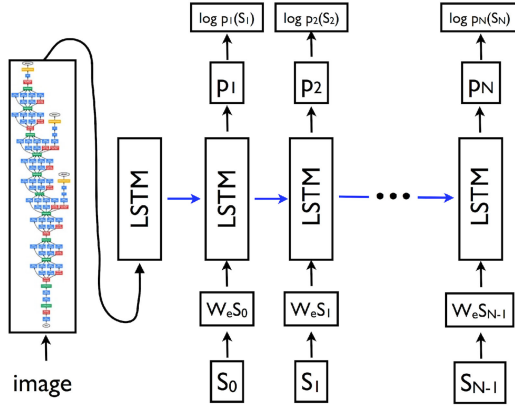


Fig. 3. Illustration of the Google NIC model. Figure extracted from [90].

### B. CNN Encoder to RNN Decoder

As shown in Fig. 2, one of the most representative encoder-decoder models is based on a CNN image encoder and a RNN text decoder, where CNN extracts various vision cues from one still image as a single real-valued feature representation, and RNN generates caption for that image conditioned on its representation at the very beginning.

More specifically, we take the neural image caption (NIC) [87], [90] model as an example to elaborate this popular scheme. Given an image  $I$  and its description  $S = \{S_0, \dots, S_N\}$ , NIC targets to maximize the conditional probability of  $S$  given  $I$  as  $\max_{\theta} \sum_{(I, S)} \log p(S|I; \theta) = \sum_{t=0}^N \log p(S_t|I, S_0, \dots, S_{t-1}; \theta)$ , where  $\theta$  parameterizes the NIC model. To tackle with the likelihood maximization, NIC employs a pre-trained CNN to encode the image, and then feed the encoded representation to a RNN (e.g., commonly a one-layer LSTM) to decode the sentences. One obvious benefit of using RNN is its ability to handle variable length of sentence. It models the time sequence data with hidden state or memory (i.e., a fixed-length vector), conditioning on the current input and the state in last time step. As shown by Fig. 3, NIC unrolls the procedure of caption generation by

$$x_{-1} = W_I \text{CNN}(I), \quad (1)$$

$$x_t = W_e S_t, \quad (2)$$

$$h_{t+1} = \text{LSTM}(x_t, h_t) \quad (3)$$

$$p_{t+1} = \text{softmax}(h_{t+1}), \quad (4)$$

where  $W_I$  ( $W_e$ ) is image (word) embedding matrix that needs to be learned,  $h_t$  denotes the hidden state of the LSTM layer at time  $t$ ,  $t \in \{0, \dots, N-1\}$ , and  $p_{t+1}$  is the probability distribution over all the words, which is given by  $h_{t+1}$  followed by a softmax layer. It is worth noting that, NIC only shows the image content to the RNN decoder once (at time  $t = -1$ ), that is, to initialize the hidden state for LSTM.

Many other same period works employ the similar architecture of CNN encoder and RNN decoder, yet with subtle differences. Mao *et al.* proposed a multimodal RNN (m-RNN) model [76] to cope with the caption generation, where the image feature is also extracted by pre-train CNN and the words are represented with two-layer embedding and modeled by vanilla RNN. Different from [90], the image content is visible to RNN at each time step, and the visual as well as text information are incorporated with an additional multimodal layer (followed by softmax) to predict the words. The multimodal layer in essence consists of three MLPs, which project the word embedding, hidden state of RNN and image feature into the same space, and fuse them into one single vector via element-wise add.

Similar to m-RNN [76], the long-term recurrent convolutional networks (LRCNs) proposed by Donahue [77] also feed the image content to the RNN decoder at each time step. In LRCN, a stacked two-layer LSTM (refer to the factored one in their paper) is formulated as decoder, where the bottom LSTM is only fed with the previous word embedding, and the top LSTM takes as input the image feature and the output given the bottom LSTM. The advantage for this stacked architecture is that the bottom LSTM can focus on modeling the text data; and the top one fuses the context and image information to predict the word distribution, which is similar to the multimodal layer in the m-RNN [76].

Another interesting work is proposed by Chen *et al.* [78], which introduces a recurrent visual feature to assist the long-term memory for the RNN decoder. This model not only maximizes the sentence likelihood given image and previous words, but also considers the likelihood w.r.t visual feature conditioned on the previous words. By this means, they develop an addition hidden layer within the RNN decoder to model the visual memory, and thus help the word prediction.

### C. Attention Mechanism

All the methods above mainly encode image with the top layer of pre-trained CNNs, and keep the image content fixed during the decoding process for generating natural language sentence. However, it is not an easy task to distill all the necessary information into one single vector, considering the cluttered background and multiple objects, as well as the complex relationship between objects. Thus, it will be helpful for caption generation by looking at different image regions according to the context. In light of this, attention mechanism has been widely used for image captioning, which generally learns *where and what* the RNN decoder should attend to. In the next, we will review two common attention mechanisms, including spatial attention and semantic attention.

1) *Spatial Attention*: Xu *et al.* [91] first introduced spatial attention (see Fig. 4) for the task of image captioning. In their



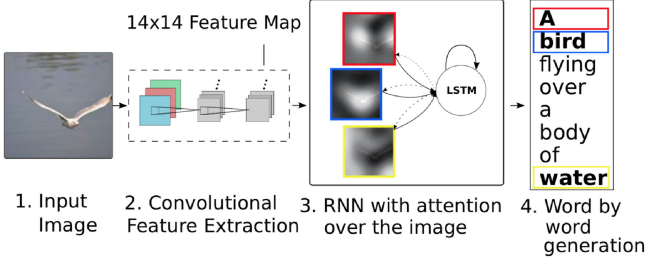


Fig. 4. Illustration of the spatial attention mechanism. Figure reproduced from [91].

model, the last convolutional layer of pre-trained CNNs is employed for the image encoder, instead of using a fully connected layer. By this means, the visual information is vectorized as a set of representations, e.g.,  $a = \{a_1, \dots, a_L\}$ ,  $a_i \in \mathbb{R}^D$ , which are corresponding to different (i.e.,  $L$ ) regions of the given image, and hence allow the RNN decoder attending to different spatial image regions under the attention mechanism. Like previous works [90], the RNN decoder is also formulated as a one-layer LSTM. However, instead of keeping the visual content fixed, they introduce a key concept of *context vector* to compute the hidden state of LSTM at each time step as the following:

$$h_t = \text{LSTM}(x_{t-1}, h_{t-1}, z_t), \quad (5)$$

where  $z_t$  represents the context vector at time  $t$ , and explicitly considers the relevance between image regions and the generated words at each time step. It is defined by

$$z_t = \phi(\{a_i\}, \{\alpha_i\}). \quad (6)$$

where  $\alpha_i$  represents the weight of each image region, and  $\phi(\cdot)$  works as a fusion function that computes a single vector representation upon image regions and their corresponding weights. The weight  $\alpha_i$  of each image region is obtained by

$$e_{ti} = f_{att}(a_i, h_{t-1}), \quad (7)$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^L \exp(e_{tk})}, \quad (8)$$

where  $f_{att}$  represents the attention function formulated by a multilayer perceptron network and conditioned on the previous hidden state. In [91], the authors provide two ways to implement the function  $\phi(\cdot)$ , termed as hard attention and soft attention, where the former selects the attended position based on Monte Carlo sampling; while the later models the relative importance among all the regions through blending them as  $z_t = \sum \alpha_i a_i$ . The spatial attention mechanism [91] formulates another landmark for the image captioning task, followed by a wide range of variants. In the next, we will briefly introduce two interesting works that mainly adopt the spatial attention.

Yang *et al.* [97] proposed to extend context vectors as thought vectors, by considering global information with a ReviewNet. They adopt both RNN and CNN networks as visual encoders to capture the whole view of image content and also formulate it as sequence. The ReviewNet formulated by LSTM layer is employed to capture the visual sequence with spatial attention and generate the thought vectors by using temporal attention.

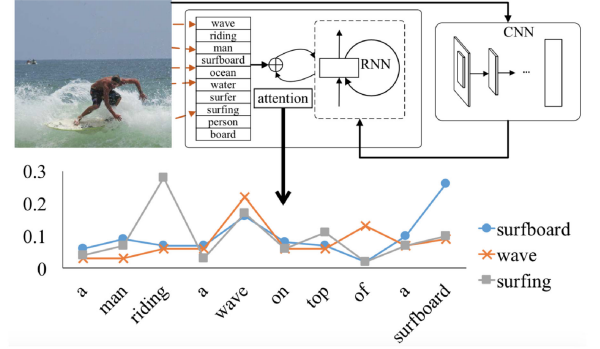


Fig. 5. Illustration of the semantic attention mechanism. Figure reproduced from [95].

This reviewing process is repeated several times to formulate a data sequence. The text decoder in [97] also adopts a LSTM layer with putting attention mechanism on the input thought vectors. Different from [91], which leverages spatial attention with image features during the caption generation process, the ReviewNet “reviews” where to attend to in advance and expects to explore more global properties than directly using the encoded visual features.

On the other hand, Lu *et al.* [99] not only target to enable the decoder knowing where to look, but also focus on the problem of when to look. In their model, they first modify the spatial attention as being conditioned on the current hidden state to obtain context vector with image features, and then introduce a visual sentinel to determine whether the decoder should attend image for predicting the next word. The visual sentinel is in essence a latent representation that computed with memory state from the LSTM layer, and controlled by current input and previous hidden state. It is combined with the context vector through a weight parameter to be learned, resulting in an adaptive attention scheme to balance the visual content and language memory.

2) *Semantic Attention*: High-level semantic information have shown to be useful for assisting the architecture of visual encoder to text decoder [100], where the semantics are usually formulated as visual attributes [95], [98], [100]. These methods generally feed to the text decoder both the encoded image content and detected visual attributes, to achieve complimentary information for the generation process.

You *et al.* [95] transferred the attention mechanism from spatial image regions to the “visual words” (see Fig. 5). In their model, they only show image features at the beginning to give an overview visual clue to the RNN decoder, and then design two attention models at each time step to take good care of input and output for the decoding process. In details, the input attention model takes as input a set of visual attributes (i.e., semantic information represented by word embedding) detected from image, and learns a semantic context vector with being conditioned on the previous word. On the other hand, the output attention model adopts a similar attention mechanism to the input one, except being conditioned on the current hidden state. By recursively applying these two models during the generation, the RNN decoder could consistently update the hidden state with

the attended attributes. One benefit of semantic attention is to provide a natural way to bridge the gap between visual and semantic spaces, which also puts more emphasis on language modeling with recurrent attention. An extensive investigation of how to incorporate high-level semantics into the RNN decoder is provided by Yao *et al.* [98]. Though this work does not employ attention mechanism, they explicitly consider the relationship among different attributes.

In this survey work, we focus on the captioning task of generic visual data, which is not limited to still images. More discussions on image captioning could be referred to Refs [110]–[113]. In the next, we will present recent works related to video captioning.

## V. RECENT ADVANCES: VIDEO CAPTIONING

Similar to image captioning, video captioning aims to translate a sequence of video frames into a human readable sentence, which, in recent years, also mainly adopts an encoder-decoder architecture. To our best knowledge, Venugopalan *et al.* [101] first introduced such a framework for the video captioning task, where the encoder employs the AlexNet [80] to extract visual feature for each frame and simply utilizes average pooling to encode the whole video as a single vector, while the decoder leverages the widely-used LSTM layer as a language model to decode the encoded visual feature as a sequence of words. Though this method does not consider the temporal information of video, it inspires a wide range of the following works, which pay attention to the visual encoding process by modeling the temporal information [86] and exploring the hierarchical structure [107], [108]; or focus on the text decoding part with exploiting the attention mechanism [103], considering the hierarchy between sentence and paragraph [106], as well as using the semantic information [107], [109]. Table III gives an overview of the recent video captioning methods based on the encoder-decoder framework, which could be roughly divided into two categories by the visual encoding way, i.e., either based on LSTM or spatial-temporal CNN.

### A. LSTM Based Encoding

The LSTM based encoding methods generally treat video as a sequence of visual features, and try to encapsulate the context information of sequence data into the hidden state of LSTM layer at the final step. To this end, the pre-trained deep CNN networks are usually used to extract visual feature from video frames, such as using VGG [81] on RGB frames and AlexNet [80] on optical flow images in [102], employing GoogleNet [82] in [107], and using ResNet [83] with C3D [105] in [108]. On the other hand, to explore the temporal information of video, various networks have been designed upon LSTM.

Venugopalan *et al.* [102] successfully transferred the popular *sequence to sequence* [86] model in neural machine translation to the video captioning task, and thus proposed a novel model termed as S2VT. As following [77], a two-layer stacked LSTMs network is used to encode the video sequence, which, in contrast to [77], is also used as a decoder to unroll the visual content as a text sequence. In S2VT, the first-layer LSTM

models the temporal information between video features during the encoding stage, while the second-layer LSTM takes as input the hidden state of the first LSTM and the previous word, and only focuses on the language modeling part. However, since LSTM may not capture the very long-range dependency and stacking LSTMs inevitably burdens the computation, some works [107], [108] attempt to divide the long video into several short clips by utilizing the hierarchical structure underlying in video, while keep modeling the relationship between different clips in sequence.

To achieve above goals, Pan *et al.* [107] proposed a Hierarchical Recurrent Neural Encoder (HRNE) model by using two-layer LSTMs. HRNE employs the first LSTM as a temporal filter to explore the local temporal structure of each subsequence and thus delivers a hidden representation, which is analogous to applying convolution filter to capture the local structure of image. Meanwhile, it uses the second LSTM to maintain the context information among all the subsequences. A hierarchical video representation is eventually obtained as the output of the second LSTM. Moreover, the temporal attention mechanism [85] is also applied in their encoding process. For the decoding part, one-layer LSTM is used as the language model.

Following HRNE [107], Baraldi *et al.* [108] further proposed to adaptively detect the video clip by providing a novel Boundary-Aware LSTM cell (LSTM-BA), rather than simply segmenting the video in a hand-crafted way as [107]. A time boundary detection unit is incorporated into the LSTM cell, which learns to decide whether to transfer the hidden state and memory content to the next step. It enables LSTM-BA resetting the hidden state and memory cell when a new video boundary is detected, and interrupts the continuous update for sequence modeling. One benefit of LSTM-BA is to avoid blending the hidden states of different subsequences. A higher-layer LSTM is also employed to composite all the subsequences into a compact representation. In the decoding process, one-layer Gated Recurrent Unit (GRU) [84] is used to decode meaningful sentences.

### B. Spatial-Temporal Based CNN Encoding

Different from LSTM based encoding, the methods in this track mainly rely on the 3-D convolution network (C3D) [104], [105] to capture the temporal information in the encoding process, and usually pay more attention to the decoding part, such as incorporating the attention mechanism [103], [106] or involving high-level semantics [107], [109].

1) *Attention Mechanism*: Yao *et al.* [103] first introduced the C3D based video encoder, and employed the temporal attention [85] during the decoding stage. One-layer LSTM plays the role as caption generator in their model, which conditions on the previous hidden state and takes as input the previous word as well as a temporal context vector. The temporal context vector is given by the attention function, which computes a weight for each sequence data with its visual feature and previous hidden state at each time step. By this means, it is able to dynamically model the video's global temporal structure. Following [103], Yu *et al.* [106] devised a hierarchical RNN decoder (h-RNN) to exploit both spatial and temporal attention mechanisms, and



explicitly consider the hierarchy structure between sentence and paragraph. Specifically, two-layer GRUs [84] are formulated as sentence and paragraph generator, respectively. In their model, each video frame is divided into several patches, which makes it possible to consider the spatial relationship between visual feature within each frame. In the decoding stage, the sentence generator of h-RNN first computes a spatial attention score for each frame, and then employs the temporal attention to model the whole video sequence. The sentence is generated by the first-layer GRU conditioning on the previous hidden state with spatial-temporal context vector and previous word. On the other hand, h-RNN utilizes the second-layer GRU to explore the relationship between generated sentences, and finally delivers a meaningful paragraph.

2) *High-level Semantics*: Pan *et al.* [107] tackled the video captioning task by jointly learning a visual-semantic co-embedding space. In details, they formulated video as a single vector by average pooling the appearance features given by CNNs [80], [81] and the temporal features captured by C3D [105]. Instead of directly using the video representation, they fed the visual-semantic embedding to the LSTM decoder to bridge the gap between visual and semantic space, which facilitates the caption generation. To this end, they trained the model by jointly considering a co-embedding loss. Different from seeking a co-embedding representation, Pan *et al.* [109] detected visual attributes from frames/videos with a multiple instance learning (MIL) [114] model, and fed these attributes as high-level semantics to a two-layer stacked LSTM decoder, where the first LSTM is initialized with video representation and takes care of word sequence; and the second one generates meaningful words from visual attributes and the output of the first LSTM. To explore the temporal structure with visual attributes, they designed a transfer unit to control the visual attributes with context information, and incorporated it into the stacked LSTMs. In details, the transfer unit integrates the information from input words, visual attributes (including frames and video) and the previous hidden state into a semantic representation, which is provided as input to the second LSTM time-wisely.

## VI. EVALUATION

### A. Datasets & Validation Criteria

1) *Image Datasets*: The major datasets for evaluating the image caption generation performance include the BBC News dataset [71], the UIUC Pascal Dataset [43], the SBU Captioned Photo dataset [44], the Flickr30k Images dataset [115], and the Microsoft COCO (MS-COCO) dataset [116]. The BBC News dataset creates captions from news documents, the SBU dataset adopts user-generated captions, while the other three datasets use crowd-sourced captions. Please refer to [117] for a recent survey that discusses the datasets for vision and language research. Here, we mainly introduce the Flickr30k and MS-COCO image datasets, as they are widely used by the recent image captioning methods.

The Flickr30k image dataset [115] collects 31,783 images from Flickr, which covers a wide range of human activities.

Each image is described by five crowd-sourced captions. Generally, we may split 1000 images for validation and testing, respectively, and keep the remainder as the training set. On the other hand, the MS-COCO image dataset [116] is a more challenge one, as images in this dataset may contain multiple objects, cluttered background, and complex semantic relationship. It includes 82,783, 40,504 and 40,775 images for training, validation and testing, respectively. Each image in MS-COCO is corresponding to 5 human annotated captions. Following [87], [91], [95], one general data split setting is 82,783/5000/5000 images of training/validation/test set.

2) *Video Datasets*: For evaluating video to text techniques, a very popular benchmark is the Saarbrücken Corpus of Textually Annotated Cooking Scenes (TACoS), which contains a set of video descriptions (in natural language) and timestamp-based alignment with the videos [118].

In addition, Microsoft Research Video Description Corpus (MSVD) [119], Montreal Video Annotation Dataset (M-VAD) [120] and MPII Movie Description Corpus (MPII-MD) [121] are three common benchmark datasets for the video captioning task, each of which is briefly introduced as follows.

- MSVD [119] collects 1,970 video snippets from YouTube, where each video roughly has 40 available English descriptions. As described in [101], one general setting for this dataset is 1,200/100/670 videos for training/validation/testing set.
- M-VAD [120] is a large-scale movie description corpora consisting of 49,000 DVD movie snippets extracted from 92 DVD movies, each of which is accompanied with single sentence from semi-automatically transcribed descriptive video service (DVS) narrations. The standard split of this dataset is to set training/validation/test sets as 39,000/5,000/5,000 video clips.
- MPII-MD [121] is another large-scale movie snippets collection, which contains around 68,000 video clips with corresponding sentences from 94 Hollywood movies. MPII-MD is built in a similar way to M-VAD [120], while its alignment between video clips and descriptions is manually proofed. It is generally split as 56,861/4,930/6,584 training/validation/test samples.

3) *Validation Criteria*: The most popular evaluation metrics include BLEU [69], ROUGE-L [123], METEOR [124], CIDEr [125], SPICE [126], etc. Specifically, (1) BLEU measures the effective overlap between a reference sentence  $X$  and a candidate sentence  $Y$ . It is defined as a multiplication of the geometric mean of the  $n$ -gram precision scores (i.e.,  $\text{BLEU}@n$ ), by the brevity penalty factor BP in order to penalize the short translations. In addition, the smoothed BLEU metric can also be used to perform sentence-level analysis [127]. (2) ROUGE-L measures the longest common subsequence of tokens between a reference  $X$  and a candidate  $Y$ . BLEU and ROUGE-L are the most widely-used evaluation metrics, but recent studies have shown that these two metrics are weakly correlated with human judgement. (3) METEOR is the harmonic mean of unigram precision and recall, which is suitable for exact and paraphrase matchings between the reference  $X$  and candidate  $Y$ . (4) CIDEr is short for the Consensus-based Image Description Evaluation,

TABLE II  
OVERVIEW OF ENCODER-DECODER ARCHITECTURE FOR IMAGE CAPTIONING

Methods	Visual Encoder			Text Decoder		
	Image feature	Finetune	Timestep-wise	Architecture	Word representation (finetune)	Attention mechanism
MNLM [79]	AlexNet [80]/VGG [81]	✗	✓	LBL [95]	word2vector [96] (✓)	✗
DeepVS [89]	VGG [81]	✓	✗	multimodal RNN	word2vector [96] (✗)	✗
Google NIC [91]	GoogleNet-BN [97]	✗	✗	one-layer LSTM	word embedding	✗
m-RNN [92]	AlexNet [80]/VGG [81]	✗	✓	one-layer RNN	word embedding	✗
LRCN [77]	AlexNet [80]/VGG [81]	✓	✓	two-layer LSTM	word embedding	✗
Hard-Attention [94]	VGG [81]	✗	✓	one-layer LSTM	word embedding	✓
ATT-FCN [98]	GoogleNet [82]	✗	✗	one-layer LSTM	Glove word rep. [99] (✗)	✓
ReviewNet [100]	VGG [81]	✗	✓	one-layer LSTM	word embedding	✓
MSM [101]	GoogleNet [82]	✗	✗/✓	one-layer LSTM	word embedding	✗
VisualSentinel [102]	ResNet [83]	✓	✓	one-layer LSTM	word embedding	✓

TABLE III  
OVERVIEW OF ENCODER-DECODER ARCHITECTURE FOR VIDEO CAPTIONING

Methods	Visual sequence encoding	Text sequence Decoding
LSTM-YT [104]	AlexNet [80] + average pooling	two-layer stacked LSTMs
S2VT [105]	VGG [81] & AlexNet [80] + two-layer stacked LSTMs	two-layer stacked LSTMs
SA [106]	GoogleNet [82] & C3D [107], [108]	one-layer LSTM + temporal attention [85]
h-RNN [109]	VGG [81] & C3D [107], [108]	two GRUs [84] + spatial & temporal [85] attention
HRNE [110]	GoogleNet [82] + two LSTMs + temporal attention [85]	one-layer LSTM
LSTM-BA [111]	ResNet [83] & C3D [108] + one LSTM-BA + one-layer LSTM	one-layer GRU [84]
LSTM-E [112]	AlexNet [80]/VGG [81] & C3D [108] + average pooling	one-layer LSTM with co-embedding loss
LSTM-TSA [113]	VGG [81] & C3D [108] + image/video attributes	two-layer stacked LSTMs + transfer unit

which measures the similarity of a generated sentence against a set of ground truth sentences. Both METEOR and CIDEr have shown better correlation with human judgements. Moreover, as suggested by [125], METEOR exhibits more accurate evaluation than other metrics, especially when the number of references is small. Please refer to [125], [128] for detailed comparisons of these metrics. Besides, Semantic Propositional Image Caption Evaluation (SPICE) [126] is a recently proposed metric specifically designed for image captioning, which computes the caption similarity based on the consensus of scene-graph tuples of the candidate sentence and all the reference ones. It also shows promising evaluation performance for the image captioning task. Most recently, some learning based metrics [129], [130] are also proposed for captioning evaluation, which mainly implement deep neural networks to distinguish between generated captions and ground truth. In the next, we mainly adopt BLEU@n, METEOR, CIDEr and SPICE, and may denote BLEU@n as B-n for simplicity.

### B. Evaluation on Image Captioning

Table IV summarizes the performance of recent encoder-decoder based image captioning methods on the Flickr30k and MS-COCO image dataset in terms of BLEU, METEOR and CIDEr, respectively. We select eight representative methods from Table II, which mainly adopt the architecture of CNN-encoder and LSTM-decoder. As shown by Table IV, several conclusion could be made: 1) Attention based methods

generally outperform others, which validates the effectiveness of attention mechanism to caption generation. 2) High-level semantics show great potential to the image captioning task, as it provides complementary information and helps bridge the gap between visual and text space. 3) Ensemble models could provide higher and more robust results. We also collect the results of recent methods given by MS-COCO testing server in Table V, where c5 and c40 indicate the testing set that describe image with 5 and 40 sentences, respectively. As can be seen, it shares similar observations with Table IV.

### C. Evaluation on Video Captioning

We present the evaluation of video to text work on the TACoS Corpus [118]. Detailed results of different methods are provided in Table VI. When retrieving the most relevant sentence from the training raw video with aligned corpus, as shown in row one of Table VI, it achieves BLEU@4 of 6.0%. Instead of using the raw features, high-level semantic representations could improve the results to 12.0% and 13.0%. We notice that most advanced methods can improve those baseline approaches, up to 18.9% BLEU@4. On the same dataset, but particularly for a different subset, [65] received 14.9% and 56.7%, respectively, which indicates that the results on the different subsets of the dataset can be comparable.

Moreover, we summarize the performance of several recent video captioning methods on three benchmark datasets in Table VII, where LSTM-YT [101] serves as a strong baseline, and the others could be roughly divided into two groups: (1) the LSTM-based encoding methods including S2VT [102], HRNE [107] and LSTM-BA [108]; (2) the spatial-temporal CNN based encoding methods including SA [103], h-RNN [106], LSTM-E [107] and LSTM-TSA [109]. Interestingly, though the LSTM based encoder is expected to capture the temporal information from video sequence better, the spatial-temporal CNN based methods overall perform slightly better than LSTM encoding methods in practice. This is possibly due to the fact that the CNN-encoder pays more attention to the decoding process, which is more likely the key factor to the final performance. Moreover, high-level semantics (e.g., LSTM-E [107] and LSTM-TSA [109]) also exhibit a good performance, similar to the image captioning task, which again shows the benefit of introducing semantic information to caption generation. Some example videos with detailed results are shown in Fig. 6.

TABLE IV  
PERFORMANCE SUMMARIZATION OF RECENT ENCODER-DECODER IMAGE CAPTIONING METHOD ON THE FLICKR30K AND MS-COCO IMAGE DATASETS, WHERE  $\dagger$  INDICATES ENSEMBLE MODELS

Methods	Flickr30k						MS-COCO					
	B-1	B-2	B-3	B-4	METEOR	CIDEr	B-1	B-2	B-3	B-4	METEOR	CIDEr
DeepVS [89]	0.573	0.369	0.240	0.157	0.157	0.247	0.625	0.450	0.321	0.230	0.195	0.660
Google NIC [91]	0.663	0.423	0.277	0.183	-	-	0.666	0.461	0.329	0.246	-	-
m-RNN [92]	0.600	0.410	0.280	0.190	-	-	0.670	0.490	0.350	0.250	-	-
LRCN [77]	0.587	0.390	0.250	0.165	-	-	0.628	0.442	0.304	0.210	-	-
Hard-Attention [94]	0.669	0.439	0.296	0.199	0.185	-	0.718	0.504	0.357	0.250	0.230	-
ATT-FCN $^\dagger$ [98]	0.647	0.460	0.324	0.230	0.189	-	0.709	0.537	0.402	0.304	0.243	-
MSM $^\dagger$ [101]	-	-	-	-	-	-	0.730	0.565	0.429	0.325	0.251	0.986
VisualSentinel [102]	0.677	0.494	0.354	0.251	0.204	0.531	0.742	0.580	0.439	0.332	0.266	1.085

TABLE V  
PERFORMANCE SUMMARIZATION OF RECENT ENCODER-DECODER IMAGE CAPTIONING METHOD ON THE MS-COCO TESTING SERVER

Methods	B-1		B-2		B-3		B-4		METEOR		ROUGE-L		CIDEr		SPICE	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
MS Captivator [126]	0.715	0.907	0.543	0.819	0.407	0.710	0.308	0.601	0.248	0.339	0.526	0.680	0.931	0.937	0.180	0.609
Google NIC [91]	0.713	0.895	0.542	0.802	0.407	0.694	0.309	0.587	0.254	0.346	0.530	0.682	0.943	0.946	0.182	0.636
m-RNN [92]	0.716	0.890	0.545	0.798	0.404	0.687	0.299	0.575	0.242	0.325	0.521	0.666	0.917	0.935	0.174	0.600
LRCN [77]	0.718	0.895	0.548	0.804	0.409	0.695	0.306	0.585	0.247	0.335	0.528	0.678	0.921	0.934	0.177	0.599
Hard-Attention [94]	0.705	0.881	0.528	0.779	0.383	0.658	0.277	0.537	0.241	0.322	0.516	0.654	0.865	0.893	0.172	0.598
ATT-FCN [98]	0.731	0.900	0.565	0.815	0.424	0.709	0.316	0.599	0.250	0.335	0.535	0.682	0.943	0.958	0.182	0.631
ReviewNet [100]	0.720	0.900	0.550	0.812	0.414	0.705	0.313	0.597	0.256	0.347	0.533	0.686	0.965	0.969	0.185	0.649
MSM [101]	0.751	0.926	0.588	0.851	0.449	0.751	0.343	0.646	0.266	0.361	0.552	0.709	1.049	1.053	0.197	0.669
VisualSentinel [102]	0.748	0.920	0.584	0.845	0.444	0.744	0.336	0.637	0.264	0.359	0.550	0.705	1.042	1.059	0.197	0.673

TABLE VI  
EVALUATION OF GENERATED DESCRIPTIONS ON TACoS VIDEO-DESCRIPTION CORPUS. HUMAN JUDGMENTS FROM 1–5, WHERE 5 IS BEST

Approach	BLEU%		Human Judgments		
	BLEU@4	BLEU@1	Grammar	Correctness	Relevance
Sentence retrieval (raw video features) [23]	6.0	32.3	NA	NA	NA
Sentence retrieval (attributes classifiers) [135]	12.0	39.9	4.6	2.3 (3.1/2.0/2.7)	2.1
Sentence retrieval (CRF predictions) [136]	13.0	40.0	4.6	2.8 (3.7/2.5/3.0)	2.6
CRF + N-gram generation [55]	16.0	56.2	4.7	2.9 (3.9/2.6/2.7)	2.6
CRF + annotations (All) [137]	11.2	38.5	NA	NA	NA
CRF + annotations (Last) [137]	16.9	44.5	NA	NA	NA
CRF + annotations (Semantic overlap) [137]	18.9	48.1	4.6	2.9 (3.7/2.6/3.2)	2.6
CRF+ sentence level predictions [137]	6.0	32.3	4.6	3.1 (3.9/2.9/3.3)	2.8
Human descriptions [55]	36	66.9	4.6	4.6 (4.6/4.7/3.7)	4.3

TABLE VII  
PERFORMANCE (%) SUMMARIZATION OF RECENT ENCODER-DECODER VIDEO CAPTIONING METHOD ON THREE DATASETS

Methods	M-VAD	MPII-MD	MSVD					
	METEOR	METEOR	METEOR	CIDEr	BLEU@1	BLEU@2	BLEU@3	BLEU@4
LSTM-YT [104]	6.1	6.7	29.1	-	-	-	-	33.3
S2VT [105]	6.7	7.1	29.8	-	-	-	-	-
HRNE [110]	6.8	-	33.1	-	79.2	66.3	55.1	43.8
LSTM-BA [111]	7.3	7.0	32.4	63.5	-	-	-	42.5
SA [106]	5.7	-	29.6	51.7	80.0	64.7	52.6	41.9
h-RNN [109]	-	-	32.6	65.8	81.5	70.4	60.4	49.9
LSTM-E [112]	6.7	7.3	31.0	-	78.8	66.0	55.4	45.3
LSTM-TSA [113]	7.2	8.0	33.5	74.0	82.8	72.0	62.8	52.8




	<b>Attributes from videos:</b> person: 0.962 doing: 0.732 man: 0.675 room: 0.633 boy: 0.564 cleaning: 0.398 machine: 0.382 his: 0.368 someone: 0.333 riding: 0.258	<b>Attributes from images:</b> young: 0.420 girl: 0.319 holding: 0.308 child: 0.210 little: 0.200 floor: 0.186 pair: 0.185 it: 0.176 woman: 0.168 playing: 0.166	<b>GT:</b> a baby is cleaning <b>LSTM:</b> a boy is playing with a toy <b>LSTM-TSA<sub>N</sub>:</b> a boy is cleaning the floor
	<b>Attributes from videos:</b> riding: 0.710 man: 0.707 two: 0.503 each: 0.455 other: 0.453 together: 0.445 going: 0.404 bike: 0.401 talk: 0.400 motor: 0.399	<b>Attributes from images:</b> man: 0.543 woman: 0.409 sitting: 0.391 two: 0.342 wearing: 0.341 riding: 0.311 smiling: 0.281 young: 0.233 people: 0.210 motorcycle: 0.202	<b>GT:</b> a man and woman is riding a motorcycle <b>LSTM:</b> a woman is riding a horse <b>LSTM-TSA<sub>N</sub>:</b> a man and woman are riding a motorcycle
	<b>Attributes from videos:</b> animals: 0.806 ground: 0.756 something: 0.743 black: 0.636 man: 0.611 animal: 0.603 baby: 0.506 forest: 0.453 searching: 0.434 walking: 0.416	<b>Attributes from images:</b> bear: 0.521 forest: 0.460 walking: 0.369 woods: 0.362 some: 0.335 area: 0.242 standing: 0.220 two: 0.212 grass: 0.188 rocks: 0.186	<b>GT:</b> bear eats dirt <b>LSTM:</b> a badger is walking <b>LSTM-TSA<sub>N</sub>:</b> a bear is walking in the forest

Fig. 6. Attributes and sentences generation results on MSVD dataset. GT means ground truth. Figure reproduced from [109].



## VII. CONCLUSION AND FUTURE PERSPECTIVES

The visual to text techniques aim at accurately describing visual contents using natural language descriptions. One well-known challenge is the long-standing semantic gap between computable low-level features and semantic information that they encode. In this paper, we gave a comprehensive survey of relevant work on this topic. We can clearly notice that significant progress has been achieved in coupling visual recognition and computational linguistics recently. In this section, we discuss the future work and promising directions.

### A. Generating Natural and Diverse Descriptions

One important direction of work will be generating more natural and diverse descriptions, rather than the predefined templates. As discussed in [134], meaningful captions shall have three properties, *fidelity*, *naturalness*, and *diversity*, where the last two are essential properties of human language. However, most of existing image or video captioning works mainly focus on the fidelity of the generated descriptions. Most recently, some works try to achieve natural and diverse captions by the means of contrastive learning [135], conditional GAN [134] and variational auto-encoder [136].

Especially, this task will be more challenging for video captioning. First, many existing techniques on video captioning couldn't fully exploit the temporal structure of video, which is extremely important when describing a long video with multiple sentences. This problem has been approached through sliding window [137], fine grained segmentation plus recognition [41], and spatial-temporal based CNN encoding [103]. But these approaches are still not satisfactory in real applications. Second, with a few exceptions [106], many existing approaches are focusing on short duration videos or video clips which can be described in a couple of sentences [29], [39], [41]. In the future, researchers may pay more attention to generating paragraphs for more complex, long videos. When it comes to multiple sentences or paragraphs, generating coherent sentences is the key.

Also, the visual to text techniques should be able to leverage more flexible semantic units. When constructing mapping functions between visual and linguistic units, current approaches typically are restricted themselves on single semantic relationships, such as between action and verbs, objects and nouns. However, human language has so many types of combination of verbs, nouns, and other language units, which requires that visual translation methods should have similar flexibility in terms of constructing bigger semantic units. For example, an action with an interactive objects can be mapped to a long phrases as transitive verbs taking an object. Similar ideas has been approached in [138], where visual phrases of objects are used. Though modeling small semantic units can lead to better occurrence statistics, we believe that, in order to achieve a more accurate understanding, it is critical to form bigger semantic units by translating several elements jointly.

Moreover, it would be very important by extending current techniques to more unconstrained situations, where longer videos with multiple sentence descriptions are preferred.

### B. Deep Reinforcement Learning for Visual to Text

As an alternative paradigm to the encoder-decoder framework described in Section IV and V, deep reinforcement learning has been applied to the image and video captioning tasks recently. Several pioneering works on this topic include the decision-making framework based on a policy network and a value network [139], the self-critical sequence training model [140], reinforced video using captioning entailment-enhanced reward [141], etc. On one hand, designing reward functions plays a key role in these methods, and it still requires a lot of research efforts in the future. On the other hand, modeling "visual to text" tasks as reinforcement learning problems makes it easier to integrate with other innovative machine learning strategies, such as bringing a human in the loop [142].

### C. Unified Framework for Visual to Text

Most of current methods have approached semantic understanding of visual contents and language generation separately, or loosely combined the two. Computer vision techniques mainly focus on constructing learning models which can characterize images or videos by recognizing their characteristics. On the other hand, NLP is trying to parse a language description by identifying connotation and denotations. However, it could be better to formulate the problem as a unified framework which can automatically find the balance or a good meeting point between the two big areas of AI. For researchers in both sides, the big challenge here is to dive into fields they might not be familiar with, and leverage the advances from both sides.

### D. Visual Understanding and Reasoning

Existing visual to text techniques mainly focus on the visual description problem. It would be more interesting to think one step further, and develop visual understanding systems, such as visual question answering and visual reasoning. By leveraging the visual to text techniques, the high-level visual understanding techniques are expected to achieve better performance in the near future.

### E. Large-scale Benchmarks and Evaluations

Like many other machine learning driven applications, visual to text technique highly depends on the training data in terms of both quality and quantity. The next big achievement would be using most advanced supervised machine learning techniques on large-scale visual-text aligned data sets. How to build large-scale image and video datasets with accurate and diverse text descriptions in an effective manner will be another major challenge in the future work.

## REFERENCES

- [1] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 273–280.
- [2] C. Wallraven, M. Schultze, B. Mohler, A. Vataakis, and K. Pastra, "The Poeticon enacted scenario corpus—A tool for human and computational experiments on action understanding," in *Proc. 9th IEEE Conf. Autom. Face Gesture Recognit.*, 2011, pp. 484–491.

- [3] N. Mavridis, "Grounded situation models for situated conversational assistants," Ph.D. dissertation, Dept. Archit., Massachusetts Inst. Technol., Cambridge, MA, USA, 2007.
- [4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.
- [5] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. 17th IEEE Int. Conf. Comput. Vis.*, 1999, vol. 2, pp. 1150–1157.
- [6] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.
- [7] M. Turner, "Texture discrimination by Gabor functions," *Biol. Cybern.*, vol. 55, no. 2, pp. 71–82, 1986.
- [8] J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille, "Efficient multilevel brain tumor segmentation with integrated Bayesian model classification," *IEEE Trans. Med. Imag.*, vol. 27, no. 5, pp. 629–640, May 2008.
- [9] Y. Hou and G. Pang, "People counting and human detection in a challenging situation," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 41, no. 1, pp. 24–33, Jan. 2011.
- [10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. I–511.
- [11] O. Barinova, V. Lempitsky, and P. Kholi, "On detection of multiple object instances using hough transforms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1773–1784, Sep. 2012.
- [12] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
- [13] L. Baraldi, C. Grana, and R. Cucchiara, "A deep siamese network for scene detection in broadcast videos," in *Proc. 23rd Annu. ACM Conf. Multimedia*, 2015, pp. 1199–1202.
- [14] D. Zhang and S. Chang, "Event detection in baseball video using superimposed caption recognition," in *Proc. 10th ACM Int. Conf. Multimedia*, 2002, pp. 315–318.
- [15] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1234–1241.
- [16] N. Mavridis, W. Kazmi, and P. Toulis, *Friends With Faces: How Social Networks Can Enhance Face Recognition and Vice Versa*. London, U.K.: Springer, 2010, pp. 453–482.
- [17] Y. Yu, H. Ko, J. Choi, and G. Kim, "End-to-end concept word detection for video captioning, retrieval, and question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3261–3269.
- [18] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.
- [19] C. Town, "Ontological inference for image and video analysis," *Mach. Vis. Appl.*, vol. 17, no. 2, pp. 94–115, 2006.
- [20] S. Harnad, "The symbol grounding problem," *Physica D*, vol. 42, no. 1–3, pp. 335–346, 1990.
- [21] J. Li and J. Wang, "Real-time computerized annotation of pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 985–1002, Jun. 2008.
- [22] D. Xu and S. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1985–1997, Nov. 2008.
- [23] D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective," *Int. J. Comput. Vis.*, vol. 80, pp. 3–15, 2008.
- [24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 248–255.
- [25] D. Roy *et al.*, "The human speechome project," in *Proc. Third Int. Conf. Emergence Evol. Linguistic Commun.: Symbol Grounding Beyond*, 2006, pp. 192–196.
- [26] L. von Ahn and L. Dabbish, "Designing games with a purpose," *Commun. ACM*, vol. 51, no. 8, pp. 58–67, 2008.
- [27] N. Mavridis, T. Bourlai, and D. Ognibene, "The human-robot cloud: Situated collective intelligence on demand," in *Proc. IEEE Int. Conf. Cyber Technol. Automat., Control Intell. Syst.*, 2012, pp. 360–365.
- [28] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 171–184, 2002.
- [29] C. Tan, Y. Jiang, and C. Ngo, "Towards textually describing complex video contents with audio-visual concept classifiers," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 655–658.
- [30] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 3, no. 1, 2007, Art. no. 3.
- [31] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *J. Vis. Comun. Image Representation*, vol. 19, no. 2, pp. 121–143, 2008.
- [32] H. L. Wang and L. F. Cheong, "Taxonomy of directing semantics for film shot classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 10, pp. 1529–1542, Oct. 2009.
- [33] X. Zhu, A. K. Elmagarmid, X. Xue, L. Wu, and A. C. Catlin, "Insightvideo: Towards hierarchical video content organization for efficient browsing, summarization and retrieval," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 648–666, Aug. 2005.
- [34] G. Abdollahian, C. Taskiran, Z. Pizlo, and E. J. Delp, "Camera motion-based analysis of user generated video," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 28–41, Jan. 2010.
- [35] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2012–2019.
- [36] G. Kulkarni *et al.*, "Baby talk: Understanding and generating simple image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1601–1608.
- [37] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, 2012, vol. 1, pp. 359–368.
- [38] A. Barbu *et al.*, "Video in sentences out," in *Proc. 28th Conf. Uncertainty Artificial Intell.*, 2012, pp. 102–112.
- [39] M. U. G. Khan, L. Zhang, and Y. Gotoh, "Human focused video description," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 1480–1487.
- [40] P. Hanckmann, K. Schutte, and G. J. Burghouts, "Automated textual descriptions for a wide range of video events with 48 human actions," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 372–380.
- [41] 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. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2634–2641.
- [42] S. Guadarrama *et al.*, "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. 14th Int. Conf. Comput. Vis.*, 2013, pp. 2712–2719.
- [43] A. Farhadi *et al.*, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 15–29.
- [44] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.
- [45] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proc. 15th Conf. Comput. Natural Lang. Learn.*, 2011, pp. 220–228.
- [46] D. Lin, S. Fidler, C. Kong, and R. Urtasun, "Visual semantic search: Retrieving videos via complex textual queries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2657–2664.
- [47] J. Malmoud, E. J. Wagner, N. Chang, and K. Murphy, "Cooking with semantics," in *Proc. ACL Workshop Semantic Parsing*, 2014, pp. 33–38.
- [48] I. Naim, Y. C. Song, Q. Liu, H. Kautz, J. Luo, and D. Gildea, "Unsupervised alignment of natural language instructions with video segments," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1558–1564.
- [49] A. Aker and R. Gaizauskas, "Generating image descriptions using dependency relational patterns," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics*, 2010, pp. 1250–1258.
- [50] V. Ramanathan, P. Liang, and L. Fei-Fei, "Video event understanding using natural language descriptions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 905–912.
- [51] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 207–218, 2014.
- [52] R. N. Chen Sun, "Discover: Discovering important segments for classification of video events and recounting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2569–2576.
- [53] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, "I2T: Image parsing to text description," *Proc. IEEE*, vol. 98, no. 8, pp. 1485–1508, Aug. 2010.

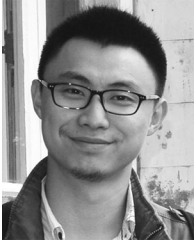
- [54] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script data for attribute-based recognition of composite activities," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 144–157.
- [55] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, "Translating video content to natural language descriptions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 433–440.
- [56] Y. Feng and M. Lapata, "Automatic caption generation for news images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 797–812, Apr. 2013.
- [57] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu, "Joint video and text parsing for understanding events and answering queries," *IEEE MultiMedia*, vol. 21, no. 2, pp. 42–70, Apr.–Jun. 2014.
- [58] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney, "Integrating language and vision to generate natural language descriptions of videos in the wild," in *Proc. 25th Int. Conf. Comput. Linguistics*, 2014, pp. 1218–1227.
- [59] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler, "What are you talking about? text-to-image coreference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3558–3565.
- [60] R. Xu, C. Xiong, W. Chen, and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2346–2352.
- [61] J. Devlin *et al.*, "Language models for image captioning: The quirks and what works," *Proc. 53rd Annual Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Language Proc.*, vol. 2, pp. 100–105, 2015.
- [62] H.-H. Nagel, "From image sequences towards conceptual descriptions," *Image Vis. Comput.*, vol. 6, no. 2, pp. 59–74, 1988.
- [63] B. Neumann and H.-J. Novak, "Event models for recognition and natural language description of events in real-world image sequences," *Comput. Compacts*, vol. 1, pp. 724–726, 1983.
- [64] M. Arens, R. Gerber, and H. Nagel, "Conceptual representations between video signals and natural language descriptions," *Image Vis. Comput.*, vol. 26, no. 1, pp. 53–66, 2008.
- [65] M. Khan and Y. Gotoh, "Describing video contents in natural language," in *Proc. Workshop Innovative Hybrid Approaches Process. Textual Data*, 2012, pp. 27–35.
- [66] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 444–454.
- [67] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
- [68] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Mach. Learn.*, vol. 68, no. 3, pp. 267–276, 2007.
- [69] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
- [70] R. Mason and E. Charniak, "Domain-specific image captioning," in *Proc. 18th Conf. Comput. Lang. Learn.*, 2014, vol. 1, pp. 11–20.
- [71] Y. Feng and M. Lapata, "How many words is a picture worth? automatic caption generation for news images," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics*, 2010, pp. 1239–1249.
- [72] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 595–603.
- [73] C. Fernández Tena, P. Baiget, X. Roca, and J. González, "Natural language descriptions of human behavior from video sequences," in *Proc. Annu. Conf. Artif. Intell.*, 2007, pp. 279–292.
- [74] D. Ding *et al.*, "Beyond audio and video retrieval: Towards multimedia summarization," in *Proc. 2nd ACM Int. Conf. Multimedia Retrieval*, 2012, p. 2.
- [75] M. Mitchell *et al.*, "Midge: Generating image descriptions from computer vision detections," in *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2012, pp. 747–756.
- [76] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (M-RNN)," *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.
- [77] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.
- [78] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2422–2431.
- [79] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *Trans. Assoc. Comput. Linguistics*, 2015.
- [80] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [81] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.
- [82] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
- [83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [84] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.
- [85] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.
- [86] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 3104–3112, 2014.
- [87] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017.
- [88] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2015, pp. 3128–3137.
- [89] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.
- [90] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Boston, MA, USA, Jun. 7–12, 2015, pp. 3156–3164.
- [91] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
- [92] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 641–648.
- [93] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Proc. Int. Conf. Learning Representations (ICLR)*, 2013.
- [94] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, vol. 37, pp. 448–456.
- [95] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4651–4659.
- [96] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, vol. 14, pp. 1532–1543.
- [97] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," *Adv. Neural Inf. Process. Syst.*, vol. 29, pp. 2361–2369, 2016.
- [98] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4904–4912.
- [99] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3242–3250.
- [100] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 203–212.
- [101] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2015, pp. 1494–1504.



- [102] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence - video to text," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4534–4542.
- [103] L. Yao *et al.*, "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4507–4515.
- [104] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1725–1732.
- [105] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.
- [106] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4584–4593.
- [107] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1029–1038.
- [108] L. Baraldi, C. Grana, and R. Cucchiara, "Hierarchical boundary-aware neural encoder for video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3185–3194.
- [109] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 984–992.
- [110] R. Bernardi *et al.*, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *J. Artif. Intell. Res.*, vol. 55, no. 1, pp. 409–442, 2016.
- [111] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, 2018.
- [112] M. Z. Hossain, F. Soheli, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, 2018.
- [113] X. Liu, Q. Xu, and N. Wang, "A survey on deep neural network-based image captioning," *Vis. Comput.*, pp. 1–26, Jun. 2018.
- [114] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1417–1424.
- [115] 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. Linguistics*, vol. 2, pp. 67–78, 2014.
- [116] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [117] F. Ferraro *et al.*, "A survey of current datasets for vision and language research," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 207–213.
- [118] M. Regneri, M. Rohrbach, D. Wetzell, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Trans. Assoc. Comput. Linguistics*, vol. 1, pp. 25–36, 2013.
- [119] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Human Lang. Technol.*, 2011, pp. 190–200.
- [120] A. Torabi, C. J. Pal, H. Larochelle, and A. C. Courville, "Using descriptive video services to create a large data source for video annotation research," 2015, arXiv:1503.01070.
- [121] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3202–3212.
- [122] H. Fang *et al.*, "From captions to visual concepts and back," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1473–1482.
- [123] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, Barcelona, Spain, 2004.
- [124] D. Elliott and F. Keller, "Image description using visual dependency representations," in *Proc. Empirical Methods Natural Lang. Process.*, 2013, vol. 13, pp. 1292–1302.
- [125] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDER: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.
- [126] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 382–398.
- [127] J. H. Clark, C. Dyer, A. Lavie, and N. A. Smith, "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability," in *Proc. 49th Annu. Meeting. Assoc. Comput. Linguistics, Human Lang. Technol.*, 2011, pp. 176–181.
- [128] D. Elliott and F. Keller, "Comparing automatic evaluation measures for image description," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Short Papers*, 2014, pp. 452–457.
- [129] N. Sharif, L. White, M. Bennamoun, and S. A. A. Shah, "Learning-based composite metrics for improved caption evaluation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 14–20.
- [130] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, "Learning to evaluate image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5804–5812.
- [131] N. Inoue, Y. Kamishima, T. Wada, K. Shinoda, and S. Sato, "Tokyotech+ canon at trecvid 2011," in *Proc. NIST TRECVID Workshop*, 2011.
- [132] T. Starner, "Visual recognition of American sign language using hidden Markov models," in *Proc. Int. Workshop Autom. Face Gesture Recognit.*, Massachusetts Inst. Technol., Cambridge, MA, USA, DTIC Document Rep., 1995.
- [133] D. Vail, M. Veloso, and J. Lafferty, "Conditional random fields for activity recognition," in *Proc. 6th Int. Joint Conf. Auton. Agents Multiagent Syst.*, 2007, pp. 1–8.
- [134] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional GAN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2989–2998.
- [135] B. Dai and D. Lin, "Contrastive learning for image captioning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 898–907.
- [136] L. Wang, A. Schwing, and S. Lazebnik, "Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5758–5768.
- [137] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1194–1201.
- [138] M. Christel, "Automated metadata in multimedia information systems: Creation, refinement, use in surrogates, and evaluation," *Synthesis Lectures Inf. Concepts, Retrieval, Serv.*, vol. 1, no. 1, pp. 1–74, 2009.
- [139] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1151–1159.
- [140] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1179–1195.
- [141] R. Pasunuru and M. Bansal, "Reinforced video captioning with entailment rewards," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 979–985.
- [142] H. Ling and S. Fidler, "Teaching machines to describe images via natural language feedback," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5068–5078.



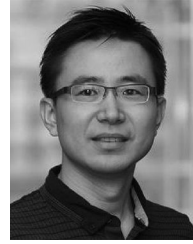
**Sheng Li** (S'11–M'17) received the B.Eng. degree in computer science and engineering and the M.Eng. degree in information security from Nanjing University of Posts and Telecommunications, Nanjing, China, and the Ph.D. degree in computer engineering from Northeastern University, Boston, MA, USA, in 2010, 2012, and 2017, respectively. Since 2018, he has been a Tenure-Track Assistant Professor with the Department of Computer Science, University of Georgia, Athens, GA, USA. He was a Research Scientist with Adobe Research from 2017 to 2018. He has authored or coauthored more than 70 papers at leading conferences and journals. His research interests include robust machine learning, dictionary learning, visual intelligence, and behavior modeling. He was the recipient of the best paper awards (or nominations) at SDM 2014, IEEE ICME 2014, and IEEE FG 2013. He is currently an Associate Editor for the IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE, *Neurocomputing*, *IET Image Processing*, and *SPIE Journal of Electronic Imaging*, and is also an Editorial Board Member for the *Neural Computing and Applications*. He was a reviewer for several IEEE Transactions, and program committee member for NIPS, ICML, IJCAI, AAAI, CVPR, and KDD.



**Zhiqiang Tao** received the B.E. degree in software engineering from the School of Computer Software, and the M.S. degree in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2012 and 2015, respectively. He is currently working toward the Ph.D. degree with Northeastern University, Boston, MA, USA. His research interests include subspace learning, ensemble clustering and representation learning.



**Kang Li** received the B.S. degree in information and computational science and the M.S. degree in expert system and intelligent control from Northwestern Polytechnical University, Xi'an, China, in 2004 and 2007, respectively, the M.S. degree in computer science and engineering from the State University of New York at Buffalo, Buffalo, NY, USA, in 2011, and the Ph.D. degree in computer engineering from Northeastern University, Boston, MA, USA, in 2014. His research interests include computer vision, applied machine learning, and data mining.



**Yun Fu** (S'07–M'08–SM'11–F'19) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xian Jiaotong University, Xi'an, China, respectively, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, respectively. Since 2012, he has been an Interdisciplinary Faculty Member affiliated with College of Engineering and the Khoury College of Computer and Information Sciences, Northeastern University, Boston, MA, USA. His research interests include machine learning, computational intelligence, big data mining, computer vision, pattern recognition, and cyber-physical systems. He has extensive publications in leading journals, books/book chapters, and international conferences/workshops. He is currently as Associate Editor, Chairs, PC member, and reviewer of many top journals and international conferences/workshops. He was the recipient of seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; nine Best Paper Awards from IEEE, IAPR, SPIE, SIAM; and many major Industrial Research Awards from Google, Samsung, and Adobe, etc. He is currently an Associate Editor for the IEEE Transactions on Neural Networks and Learning Systems and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is fellow of the IAPR, OSA, and SPIE, a Lifetime Distinguished Member of the ACM, a Lifetime Member of AAAI, and Institute of Mathematical Statistics, a member of ACM Future of Computing Academy, Global Young Academy, AAAS, INNS, and a Beckman Graduate Fellow during 2007–2008.

His research interests include machine learning, computational intelligence, big data mining, computer vision, pattern recognition, and cyber-physical systems. He has extensive publications in leading journals, books/book chapters, and international conferences/workshops. He is currently as Associate Editor, Chairs, PC member, and reviewer of many top journals and international conferences/workshops. He was the recipient of seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; nine Best Paper Awards from IEEE, IAPR, SPIE, SIAM; and many major Industrial Research Awards from Google, Samsung, and Adobe, etc. He is currently an Associate Editor for the IEEE Transactions on Neural Networks and Learning Systems and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is fellow of the IAPR, OSA, and SPIE, a Lifetime Distinguished Member of the ACM, a Lifetime Member of AAAI, and Institute of Mathematical Statistics, a member of ACM Future of Computing Academy, Global Young Academy, AAAS, INNS, and a Beckman Graduate Fellow during 2007–2008.