A Comprehensive Survey of Deep Learning for Image Captioning

MD. ZAKIR HOSSAIN, FERDOUS SOHEL, MOHD FAIRUZ SHIRATUDDIN, and HAMID LAGA, Murdoch University, Australia

Generating a description of an image is called image captioning. Image captioning requires recognizing the important objects, their attributes, and their relationships in an image. It also needs to generate syntactically and semantically correct sentences. Deep-learning-based techniques are capable of handling the complexities and challenges of image captioning. In this survey article, we aim to present a comprehensive review of existing deep-learning-based image captioning techniques. We discuss the foundation of the techniques to analyze their performances, strengths, and limitations. We also discuss the datasets and the evaluation metrics popularly used in deep-learning-based automatic image captioning.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Supervised learning**; **Unsupervised learning**; **Reinforcement learning**; **Neural networks**;

Additional Key Words and Phrases: Image captioning, deep learning, computer vision, natural language processing, CNN, LSTM

ACM Reference format:
Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2018. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.* 51, 6, Article 118 (February 2019), 36 pages.
https://doi.org/10.1145/3295748

1 INTRODUCTION

Every day, we encounter a large number of images from various sources such as the internet, news articles, document diagrams, and advertisements. These sources contain images that viewers have to interpret themselves. Most images do not have a description, but humans can largely understand them without their detailed captions. However, machines need to interpret some form of image captions if humans need automatic image captions from it.

Image captioning is important for many reasons. For example, it can be used for automatic image indexing. Image indexing is important for content-based image retrieval (CBIR), and therefore, it can be applied to many areas, including biomedicine, commerce, the military, education, digital libraries, and web searching. Social media platforms such as Facebook and Twitter can directly

This work was partially supported by an Australian Research Council grant DE120102960.
Authors' addresses: Md. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, Murdoch University, School of Engineering and Information Technology, Perth, Western Australia, 6150, Australia; emails: {MdZakir.Hossain, F.Sohel, f.shiratuddin, H.Laga}@murdoch.edu.au.

0360-0300/2018/02-ART118 $15.00
https://doi.org/10.1145/3295748

generate descriptions from images. The descriptions can include where we are (e.g., beach, cafe), what we are wearing, and, importantly, what we are doing there.

Image captioning is a popular research area of artificial intelligence (AI) that deals with image understanding and a language description for that image. Image understanding entails detecting and recognizing objects, as well as understanding scene type or location, object properties, and their interactions. Generating well-formed sentences requires both syntactic and semantic understanding of the language [143].

Understanding an image largely depends on obtaining image features. The techniques used for this purpose can be broadly divided into two categories: (1) traditional machine-learning-based techniques and (2) deep machine-learning-based techniques.

In traditional machine learning, handcrafted features such as local binary patterns (LBPs) [107], scale-invariant feature transform (SIFT) [87], the histogram of oriented gradients (HOG) [27], and a combination of such features are widely used. In these techniques, features are extracted from input data. They are then passed to a classifier such as support vector machines (SVMs) [17] in order to classify an object. Since handcrafted features are task specific, extracting features from a large and diverse set of data is not feasible. Moreover, real-world data such as images and video are complex and have different semantic interpretations.

On the other hand, in deep machine-learning-based techniques, features are learned automatically from training data and they can handle a large and diverse set of images and videos. For example, convolutional neural networks (CNNs) [79] are widely used for feature learning, and a classifier such as Softmax is used for classification. CNN is generally followed by recurrent neural networks (RNNs) in order to generate captions.

In the last 5 years, a large number of articles have been published on image captioning, with deep machine learning being popularly used. Deep learning algorithms can handle complexities and challenges of image captioning quite well. So far, only three survey papers [8, 13, 75] have been published on this research topic. Although the papers have presented a good literature survey of image captioning, they could only cover a few papers on deep learning because the bulk of them were published after the survey papers. These survey papers mainly discussed template-based, retrieval-based, and a very few deep-learning-based novel image caption generating models. However, a large number of works have been done on deep-learning-based image captioning. Moreover, the availability of large and new datasets has made learning-based image captioning an interesting research area. To provide an abridged version of the literature, we present a survey mainly focusing on the deep-learning-based papers on image captioning.

The main aim of this article is to provide a comprehensive survey of deep learning for image captioning. First, we group the existing image captioning articles into three main categories: (1) template-based image captioning, (2) retrieval-based image captioning, and (3) novel image caption generation. The categories are discussed briefly in Section 2. Most deep-learning-based image captioning methods fall into the category of novel caption generation. Therefore, we focus only on novel caption generation with deep learning. Second, we group the deep-learning-based image captioning methods into different categories, namely, (1) visual space based, (2) multimodal space based, (3) supervised learning, (4) other deep learning, (5) dense captioning, (6) whole scene based, (7) encoder-decoder architecture based, (8) compositional architecture based, (9) LSTM (Long Short-Term Memory) [54] language model based, (10) other language model based, (11) attention based, (12) semantic concept based, (13) stylized captions, and (12) novel-object-based image captioning. We discuss all the categories in Section 3. We provide an overview of the datasets and commonly used evaluation metrics for measuring the quality of image captions in Section 4. We also discuss and compare the results of different methods in Section 5. Finally, we give a brief discussion and future research directions in Section 6 and then a conclusion in Section 7.

2 IMAGE CAPTIONING METHODS

In this section, we review and describe the main categories of existing image captioning methods, including template-based image captioning, retrieval-based image captioning, and novel caption generation. Template-based approaches have fixed templates with a number of blank slots to generate captions. In these approaches, different objects, attributes, and actions are detected first and then the blank spaces in the templates are filled. For example, Farhadi et al. [34] use a triplet of scene elements to fill the template slots for generating image captions. Li et al. [80] extract the phrases related to detected objects, attributes, and their relationships for this purpose. A conditional random field (CRF) is adopted by Kulkarni et al. [74] to infer the objects, attributes, and prepositions before filling in the gaps. Template-based methods can generate grammatically correct captions. However, templates are predefined and cannot generate variable-length captions. Moreover, later on, parsing-based language models have been introduced in image captioning [2, 32, 76, 77, 101], which are more powerful than fixed template-based methods. Therefore, in this article, we do not focus on these template-based methods.

Captions can be retrieved from visual space and multimodal space. In retrieval-based approaches, captions are retrieved from a set of existing captions. Retrieval-based methods first find the visually similar images with their captions from the training dataset. These captions are called candidate captions. The captions for the query image are selected from this captions pool [47, 55, 108, 130]. These methods produce general and syntactically correct captions. However, they cannot generate image-specific and semantically correct captions.

Novel captions can be generated from both visual space and multimodal space. A general approach of this category is to analyze the visual content of the image first and then generate image captions from the visual content using a language model [70, 152, 155, 156]. These methods can generate new captions for each image that are semantically more accurate than previous approaches. Most novel caption generation methods use deep machine-learning-based techniques. Therefore, deep-learning-based novel image-caption-generating methods are our main focus in this literature.

An overall taxonomy of deep-learning-based image captioning methods is depicted in Figure 1. The figure illustrates the comparisons of different categories of image captioning methods. Novel generation-based image caption methods mostly use visual space and deep machine-learning-based techniques. Captions can also be generated from multimodal space. Deep-learning-based image captioning methods can also be categorized by learning techniques: supervised learning, reinforcement learning, and unsupervised learning. We group the reinforcement learning and unsupervised learning into "other deep learning." Usually captions are generated for a whole scene in the image. However, captions can also be generated for different regions of an image (dense captioning). Image captioning methods can use either simple encoder-decoder architecture or compositional architecture. There are methods that use attention mechanisms, semantic concepts, and different styles in image descriptions. Some methods can also generate descriptions for unseen objects. We group them into one category as "Others." Most of the image captioning methods use LSTM as the language model. However, there are a number of methods that use other language models such as CNN and RNN. Therefore, we include a language model-based category as "LSTM vs. Others."

3 DEEP-LEARNING-BASED IMAGE CAPTIONING METHODS

We draw an overall taxonomy in Figure 1 for deep-learning-based image captioning methods. We discuss their similarities and dissimilarities by grouping them into visual space versus multimodal space, dense captioning versus captions for the whole scene, supervised learning versus other

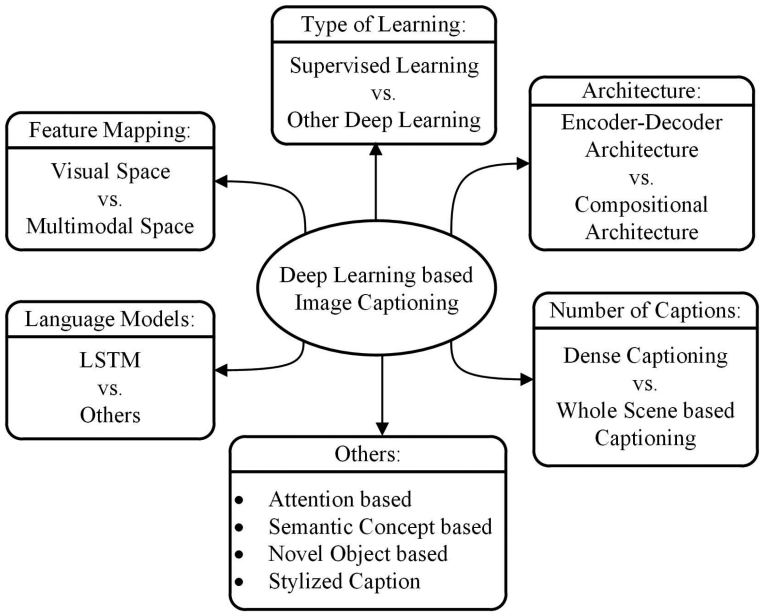


Fig. 1. An overall taxonomy of deep-learning-based image captioning.

deep learning, and encoder-decoder architecture versus compositional architecture, and one "others" group that contains attention-based, semantic-concept-based, stylized, and novel-object-based captioning. We also create a category named LSTM versus Others.

A brief overview of the deep-learning-based image captioning methods is shown in Table 1. Table 1 contains the name of the image captioning methods, the type of deep neural networks used to encode image information, and the language models used in describing the information. In the final column, we give a category label to each captioning technique based on the taxonomy in Figure 1.

3.1 Visual Space versus Multimodal Space

Deep-learning-based image captioning methods can generate captions from both visual space and multimodal space. Understandably, image captioning datasets have the corresponding captions as text. In the visual space-based methods, the image features and the corresponding captions are independently passed to the language decoder. In contrast, in a multimodal space case, a shared multimodal space is learned from the images and the corresponding caption text. This multimodal representation is then passed to the language decoder.

3.1.1 Visual Space. The bulk of the image captioning methods use visual space for generating captions. These methods are discussed in Section 3.2 to Section 3.5.

3.1.2 Multimodal Space. The architecture of a typical multimodal space-based method contains a language encoder part, a vision part, a multimodal space part, and a language decoder part. A general diagram of multimodal space-based image captioning methods is shown in Figure 2. The vision part uses a deep convolutional neural network as a feature extractor to extract the image features. The language encoder part extracts the word features and learns a dense feature embedding for each word. It then forwards the semantic temporal context to the recurrent layers. The multimodal space part maps the image features into a common space with the word features.

Table 1. An Overview of the Deep-Learning-Based Approaches for Image Captioning

Reference	Image Encoder	Language Model	Category
Kiros et al. 2014 [69]	AlexNet	LBL	MS, SL, WS, EDA
Kiros et al. 2014 [70]	AlexNet, VGGNet	1. LSTM 2. SC-NLM	MS, SL, WS, EDA
Mao et al. 2014 [95]	AlexNet	RNN	MS, SL, WS
Karpathy et al. 2014 [66]	AlexNet	DTR	MS, SL, WS, EDA
Mao et al. 2015 [94]	AlexNet, VGGNet	RNN	MS, SL, WS
Chen et al. 2015 [23]	VGGNet	RNN	VS, SL, WS, EDA
Fang et al. 2015 [33]	AlexNet, VGGNet	MELM	VS, SL, WS, CA
Jia et al. 2015 [59]	VGGNet	LSTM	VS, SL, WS, EDA
Karpathy et al. 2015 [65]	VGGNet	RNN	MS, SL, WS, EDA
Vinyals et al. 2015 [142]	GoogLeNet	LSTM	VS, SL, WS, EDA
Xu et al. 2015 [152]	AlexNet	LSTM	VS, SL, WS, EDA, AB
Jin et al. 2015 [61]	VGGNet	LSTM	VS, SL, WS, EDA, AB
Wu et al. 2016 [151]	VGGNet	LSTM	VS, SL, WS, EDA, AB
Sugano et at. 2016 [129]	VGGNet	LSTM	VS, SL, WS, EDA, AB
Mathews et al. 2016 [97]	GoogLeNet	LSTM	VS, SL, WS, EDA, SC
Wang et al. 2016 [144]	AlexNet, VGGNet	LSTM	VS, SL, WS, EDA
Johnson et al. 2016 [62]	VGGNet	LSTM	VS, SL, DC, EDA
Mao et al. 2016 [92]	VGGNet	LSTM	VS, SL, WS, EDA
Wang et al. 2016 [146]	VGGNet	LSTM	VS, SL, WS, CA
Tran et al. 2016 [135]	ResNet	MELM	VS, SL, WS, CA
Ma et al. 2016 [90]	AlexNet	LSTM	VS, SL, WS, CA
You et al. 2016 [156]	GoogLeNet	RNN	VS, SL, WS, EDA, SCB
Yang et al. 2016 [153]	VGGNet	LSTM	VS, SL, DC, EDA
Anne et al. 2016 [6]	VGGNet	LSTM	VS, SL, WS, CA, NOB
Yao et al. 2017 [155]	GoogLeNet	LSTM	VS, SL, WS, EDA, SCB
Lu et al. 2017 [88]	ResNet	LSTM	VS, SL, WS, EDA, AB
Chen et al. 2017 [21]	VGGNet, ResNet	LSTM	VS, SL, WS, EDA, AB
Gan et al. 2017 [41]	ResNet	LSTM	VS, SL, WS, CA, SCB
Pedersoli et al. 2017 [112]	VGGNet	RNN	VS, SL, WS, EDA, AB
Ren et al. 2017 [119]	VGGNet	LSTM	VS, ODL, WS, EDA
Park et al. 2017 [111]	ResNet	LSTM	VS, SL, WS, EDA, AB
Wang et al. 2017 [148]	ResNet	LSTM	VS, SL, WS, EDA
Tavakoli et al. 2017 [134]	VGGNet	LSTM	VS, SL, WS, EDA, AB
Liu et al. 2017 [84]	VGGNet	LSTM	VS, SL, WS, EDA, AB
Gan et al. 2017 [39]	ResNet	LSTM	VS, SL, WS, EDA, SC
Dai et al. 2017 [26]	VGGNet	LSTM	VS, ODL, WS, EDA
Shetty et al. 2017 [126]	GoogLeNet	LSTM	VS, ODL, WS, EDA
Liu et al. 2017 [85]	Inception-V3	LSTM	VS, ODL, WS, EDA
Gu et al. 2017 [51]	VGGNet	1. Language CNN 2. LSTM	VS, SL, WS, EDA
Yao et al. 2017 [154]	VGGNet	LSTM	VS, SL, WS, CA, NOB

(Continued)

Table 1. Continued

Reference	Image Encoder	Language Model	Category
Rennie et al. 2017 [120]	ResNet	LSTM	VS, ODL, WS, EDA
Vsub et al. 2017 [140]	VGGNet	LSTM	VS, SL, WS, CA, NOB
Zhang et al. 2017 [161]	Inception-V3	LSTM	VS, ODL, WS, EDA
Wu et al. 2018 [150]	VGGNet	LSTM	VS, SL, WS, EDA, SCB
Aneja et al. 2018 [5]	VGGNet	Language CNN	VS, SL, WS, EDA
Wang et al. 2018 [147]	VGGNet	Language CNN	VS, SL, WS, EDA

VS = Visual Space, MS = Multimodal Space, SL = Supervised Learning, ODL = Other Deep Learning, DC = Dense Captioning, WS = Whole Scene, EDA = Encoder-Decoder Architecture, CA = Compositional Architecture, AB = Attention Based, SCB = Semantic Concept Based, NOB = Novel Object Based, SC = Stylized Caption.

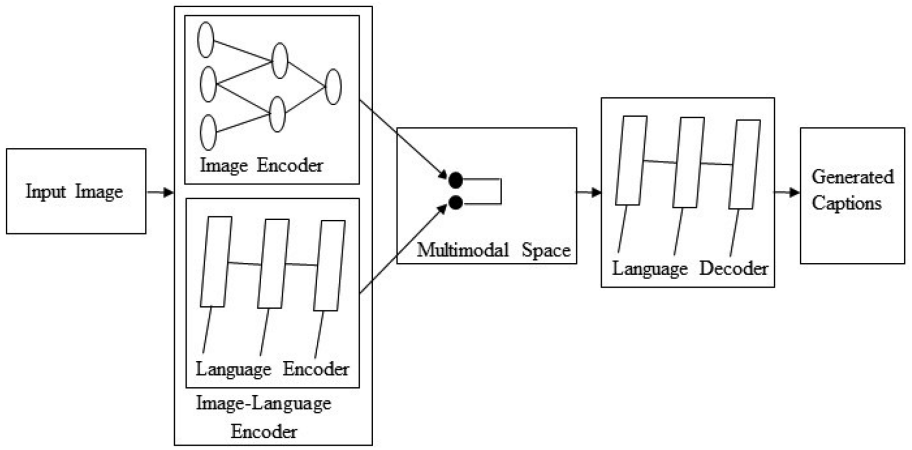


Fig. 2. A block diagram of multimodal space-based image captioning.

The resulting map is then passed to the language decoder, which generates captions by decoding the map.

The methods in this category follow the following steps:

(1) Deep neural networks and the multimodal neural language model are used to learn both image and text jointly in a multimodal space.
(2) The language generation part generates captions using the information from Step 1.

An initial work in this area was proposed by Kiros et al. [69]. The method applies a CNN for extracting image features in generating image captions. It uses a multimodal space that represents both image and text jointly for multimodal representation learning and image caption generation. It also introduces the multimodal neural language models such as the Modality-Biased Log-Bilinear Model (MLBL-B) and the Factored Three-Way Log-Bilinear Model (MLBL-F) of [104] followed by AlexNet [73]. Unlike most previous approaches, this method does not rely on any additional templates, structures, or constraints. Instead, it depends on the high-level image features and word representations learned from deep neural networks and multimodal neural language models, respectively. The neural language models have limitations in handling a large amount of data and are inefficient to work with long-term memory [64].

Kiros et al. [69] extended their work in [70] to learn a joint image sentence embedding where LSTM is used for sentence encoding and a new neural language model called the Structure-Content

Neural Language Model (SC-NLM) is used for image caption generations. The SC-NLM has an advantage over existing methods in that it can extricate the structure of the sentence to its content produced by the encoder. It also helps to achieve significant improvements in generating realistic image captions over the approach proposed by [69].

Karpathy et al. [66] proposed a deep, multimodal model embedding of image and natural language data for the task of bidirectional images and sentences retrieval. The previous multimodal-based methods used a common embedding space that directly maps images and sentences. However, this method works at a finer level and embeds fragments of images and fragments of sentences. This method breaks down the images into a number of objects and sentences into a dependency tree relation (DTR) [28] and reasons about their latent intermodal alignment. It shows that the method achieves significant improvements in the retrieval task compared to other previous methods. This method has a few limitations as well. In terms of modeling, the dependency tree can model relations easily, but they are not always appropriate. For example, a single visual entity might be described by a single complex phrase that can be split into multiple sentence fragments. The phrase "black and white dog" can be formed into two relations: (CONJ, black, white) and (AMOD, white, dog). Again, for many dependency relations we do not find any clear mapping in the image (e.g., "each other" cannot be mapped to any object).

Mao et al. [94] proposed a multimodal recurrent neural network (m-RNN) method for generating novel image captions. This method has two subnetworks: a deep recurrent neural network for sentences and a deep convolutional network for images. These two subnetworks interact with each other in a multimodal layer to form the whole m-RNN model. Both the image and fragments of sentences are given as input in this method. It calculates the probability distribution to generate the next word of the caption. There are five more layers in this model: two word embedding layers, a recurrent layer, a multimodal layer, and a SoftMax layer. Kiros et al. [69] proposed a method that is built on a Log-Bilinear Model and used AlexNet to extract visual features. This multimodal recurrent neural network method is closely related to the method of Kiros et al. [69]. Kiros et al. use a fixed-length context (i.e., five words), whereas in this method, the temporal context is stored in a recurrent architecture, which allows an arbitrary context length. The two word-embedding layers use one hot vector to generate a dense word representation. It encodes both the syntactic and semantic meaning of the words. The semantically relevant words can be found by calculating the Euclidean distance between two dense word vectors in embedding layers. Most sentence-image multimodal methods [38, 66, 70, 128] use precomputed word embedding vectors to initialize their model. In contrast, this method randomly initializes word embedding layers and learns them from the training data. This helps them to generate better image captions than the previous methods. Many image captioning methods [66, 69, 95] are built on recurrent neural networks at the contemporary times. They use a recurrent layer for storing visual information. However, m-RNN uses both image representations and sentence fragments to generate captions. It utilizes the capacity of the recurrent layer more efficiently, which helps to achieve better performance using a relatively small dimensional recurrent layer.

Chen et al. [23] proposed another multimodal space-based image captioning method. The method can generate novel captions from images and restore visual features from the given description. It also can describe a bidirectional mapping between images and their captions. Many of the existing methods [55, 66, 128] use joint embedding to generate image captions. However, they do not use reverse projection that can generate visual features from captions. On the other hand, this method dynamically updates the visual representations of the image from the generated words. It has an additional recurrent visual hidden layer with RNN that makes a reverse projection.

3.2 Supervised Learning versus Other Deep Learning

In supervised learning, training data come with desired output called a *label*. Unsupervised learning, on the other hand, deals with unlabeled data. Generative adversarial networks (GANs) [48] are a type of unsupervised learning techniques. Reinforcement learning is another type of machine-learning approach where the aims of an agent are to discover data and/or labels through exploration and a reward signal. A number of image captioning methods use reinforcement learning and GAN-based approaches. These methods sit in the category of "other deep learning."

3.2.1 Supervised-Learning-Based Image Captioning. Supervised-learning-based networks have successfully been used for many years in image classification [53, 73, 127, 133], object detection [44, 45, 116], and attribute learning [40]. This progress makes researchers interested in using them in automatic image captioning [23, 65, 94, 142]. In this article, we have identified a large number of supervised-learning-based image captioning methods. We classify them into different categories: (1) encoder-decoder architecture, (2) compositional architecture, (3) attention based, (4) semantic concept based, (5) stylized captions, (6) novel object based, and (7) dense image captioning.

3.2.2 Other Deep-Learning-Based Image Captioning. In our day-to-day lives, unlabeled data are increasing because it is often impractical to accurately annotate data. Therefore, recently, researchers have been focusing more on reinforcement learning and unsupervised-learning-based techniques for image captioning.

A reinforcement learning approach is designed by a number of parameters such as agent, state, action, reward function, policy, and value. The agent chooses an action, receives reward values, and moves to a new state. Policies are defined by actions and values are defined by reward functions. The agent attempts to select the action with the expectation of having a maximum long-term reward. It needs continuous state and action information to provide the guarantees of a reward function. Traditional reinforcement learning approaches face a number of limitations such as the lack of guarantees of a reward function and uncertain state-action information. Policy gradient methods [132] are a type of reinforcement learning that can choose a specific policy for a specific action using gradient descent and optimization techniques. The policy can incorporate domain knowledge for the action that guarantees convergence. Thus, policy gradient methods require fewer parameters than reward-function-based approaches.

Existing deep-learning-based image captioning methods use variants of image encoders to extract image features. The features are then fed into the neural-network-based language decoders to generate captions. The methods have two main issues: (1) They are trained using maximum likelihood estimation and backpropagation [114] approaches. In this case, the next word is predicted given the image and all the previously generated ground-truth words. Therefore, the generated captions look like ground-truth captions. This phenomenon is called the exposure bias [10] problem. (2) Evaluation metrics at test time are nondifferentiable. Ideally sequence models for image captioning should be trained to avoid exposure bias and directly optimize metrics for the test time. A typical architecture of reinforcement-learning-based image captioning method has two networks: (1) the policy network and (2) the value network. Sometimes they are referred to as actor and critic, respectively. The critic (value network) can be used in estimating the expected future reward to train the actor (captioning policy network). Reinforcement-learning-based image captioning methods sample the next token from the model based on the rewards they receive in each state. Policy gradient methods in reinforcement learning can optimize the gradient in order to predict the cumulative long-term rewards. Therefore, it can solve the nondifferentiable problem of evaluation metrics.

The methods in this category follow the following steps:

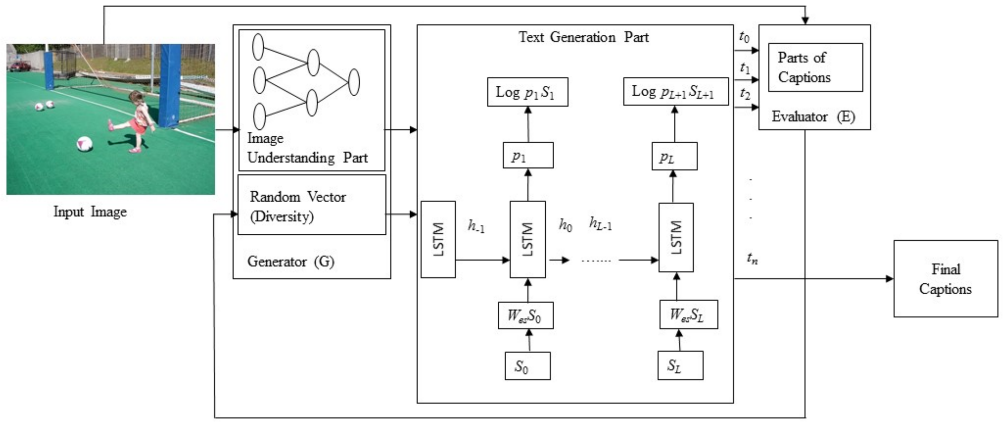


Fig. 3. A block diagram of other deep-learning-based captioning.

(1) A CNN- and RNN-based combined network generates captions.
(2) Another CNN-RNN-based network evaluates the captions and sends feedback to the first network to generate high-quality captions.

A block diagram of a typical method of this category is shown in Figure 3.

Ren et al. [119] introduced a novel reinforcement-learning-based image captioning method. The architecture of this method has two networks that jointly compute the next best word at each time step. The "policy network" works as local guidance and helps to predict the next word based on the current state. The "value network"' works as global guidance and evaluates the reward value considering all the possible extensions of the current state. This mechanism is able to adjust the networks in predicting the correct words. Therefore, it can generate good captions similar to ground-truth captions at the end. It uses an actor-critic reinforcement learning model [71] to train the whole network. Visual semantic embedding [117, 118] is used to compute the actual reward value in predicting the correct word. It also helps to measure the similarity between images and sentences that can evaluate the correctness of generated captions.

Rennie et al. [120] proposed another reinforcement-learning-based image captioning method. The method utilizes the test-time inference algorithm to normalize the reward rather than estimating the reward signal and normalization in training time. It shows that this test-time decoding is highly effective for generating quality image captions.

Zhang et al. [161] proposed an actor-critic reinforcement-learning-based image captioning method. The method can directly optimize nondifferentiable problems of the existing evaluation metrics. The architecture of the actor-critic method consists of a policy network (actor) and a value network (critic). The actor treats the job as a sequential decision problem and can predict the next token of the sequence. In each state of the sequence, the network will receive a task-specific reward (in this case, the evaluation metrics score). The job of the critic is to predict the reward. If it can predict the expected reward, the actor will continue to sample outputs according to its probability distribution.

GAN-based methods can learn deep features from unlabeled data. They achieve this representation applying a competitive process between a pair of networks: the generator and the discriminator. GANs have already been used successfully in a variety of applications, including image captioning [26, 126], image-to-image translation [56], text-to-image synthesis [15, 115], and text generation [36, 145].

There are two issues with GAN. First, GAN can work well in generating natural images from real images because GANs are proposed for real-valued data. However, text processing is based on discrete numbers. Therefore, such operations are nondifferentiable, making it difficult to apply backpropagation directly. Policy gradients apply a parametric function to allow gradients to be backpropagated. Second, the evaluator faces problems in vanishing gradients and error propagation for sequence generation. It needs a probable future reward value for every partial description. Monte Carlo rollouts [157] are used to compute this future reward value.

GAN-based image captioning methods can generate a diverse set of image captions in contrast to the model based on conventional deep convolutional networks and deep recurrent networks. Dai et al. [26] also proposed a GAN-based image captioning method. However, they do not consider multiple captions for a single image. Shetty et al. [126] introduced a new GAN-based image captioning method. This method can generate multiple captions for a single image and showed impressive improvements in generating diverse captions. GANs have limitations in backpropagating the discrete data. The Gumbel sampler [58, 91] is used to overcome the discrete data problem. The two main parts of this adversarial network are the generator and the discriminator. During training, the generator learns the loss value provided by the discriminator instead of learning it from explicit sources. The discriminator has true data distribution and can discriminate between generator-generated samples and true data samples. This allows the network to learn diverse data distribution. Moreover, the network classifies the generated caption sets, either real or fake. Thus, it can generate captions similar to human-generated ones.

3.3 Dense Captioning versus Captions for the Whole Scene

In dense captioning, captions are generated for each region of the scene. Other methods generate captions for the whole scene.

3.3.1 Dense Captioning. The previous image captioning methods can generate only one caption for the whole image. They use different regions of the image to obtain information of various objects. However, these methods do not generate region-wise captions.

Johnson et al. [62] proposed an image captioning method called DenseCap. This method localizes all the salient regions of an image and then generates descriptions for those regions.

A typical method of this category has the following steps:

(1) Region proposals are generated for the different regions of the given image.
(2) CNN is used to obtain the region-based image features.
(3) The outputs of Step 2 are used by a language model to generate captions for every region.

A block diagram of a typical dense captioning method is given in Figure 4.

Dense captioning [62] proposes a fully convolutional localization network architecture, which is composed of a convolutional network, a dense localization layer, and an LSTM [54] language model. The dense localization layer processes an image with a single, efficient forward pass, which implicitly predicts a set of regions of interest in the image. Thereby, it requires no external region proposals unlike Fast R-CNN or a full network (i.e., RPN (Region Proposal Network [44])) of Faster R-CNN. The working principle of the localization layer is related to the work of Faster R-CNN [116]. However, Johnson et al. [62] use a differential, spatial soft attention mechanism [49, 57] and bilinear interpolation [57] instead of an ROI pooling mechanism [44]. This modification helps the method to backpropagate through the network and smoothly select the active regions. It uses the Visual Genome [72] dataset for the experiments in generating region-level image captions.

One description of the entire visual scene is quite subjective and is not enough to bring out the complete understanding. Region-based descriptions are more objective and detailed than global

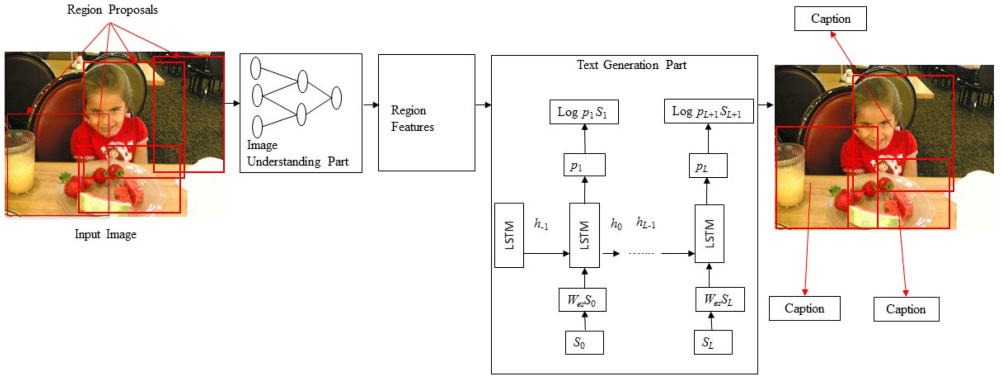


Fig. 4. A block diagram of dense captioning.

image description. The region-based description is known as dense captioning. There are some challenges in dense captioning. As regions are dense, one object may have multiple overlapping regions of interest. Moreover, it is very difficult to recognize each target region for all the visual concepts. Yang et al. [153] proposed another dense captioning method that can tackle these challenges. First, it addresses an inference mechanism that jointly depends on the visual features of the region and the predicted captions for that region. This allows the model to find an appropriate position of the bounding box. Second, they apply a context fusion that can combine context features with the visual features of respective regions to provide a rich semantic description.

3.3.2 Captions for the Whole Scene. Encoder-decoder architecture, compositional architecture, attention-based, semantic-concept-based, stylized captions, novel object-based image captioning, and other deep learning network-based image captioning methods generate single or multiple captions for the whole scene.

3.4 Encoder-Decoder Architecture versus Compositional Architecture

Some methods use just simple vanilla encoders and decoders to generate captions. However, other methods use multiple networks for it.

3.4.1 Encoder-Decoder Architecture-Based Image Captioning. The neural-network-based image captioning methods work in a simple end-to-end manner. These methods are very similar to the encoder-decoder framework-based neural machine translation [131]. In this network, global image features are extracted from the hidden activations of CNN and then fed into an LSTM to generate a sequence of words.

A typical method of this category has the following general steps:

(1) A vanilla CNN is used to obtain the scene type and to detect the objects and their relationships.
(2) The output of Step 1 is used by a language model to convert them into words and combined phrases that produce an image caption.

A simple block diagram of this category is given in Figure 5.

Vinyals et al. [142] proposed a method called Neural Image Caption Generator (NIC). The method uses a CNN for image representations and an LSTM for generating image captions. This special CNN uses a novel method for batch normalization, and the output of the last hidden layer of CNN is used as an input to the LSTM decoder. This LSTM is capable of keeping track of the

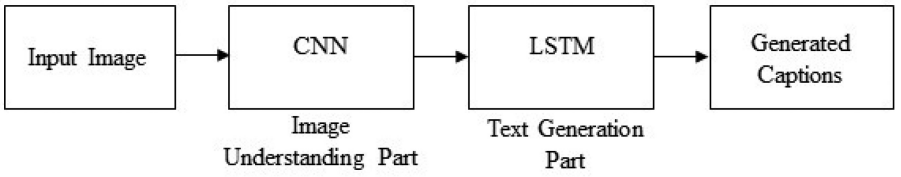


Fig. 5. A block diagram of simple encoder-decoder architecture-based image captioning.

objects that already have been described using text. NIC is trained based on maximum likelihood estimation.

In generating image captions, image information is included in the initial state of an LSTM. The next words are generated based on the current time step and the previous hidden state. This process continues until it gets the end token of the sentence. Since image information is fed only at the beginning of the process, it may face vanishing gradient problems. The role of the words generated at the beginning is also becoming weaker and weaker. Therefore, LSTM is still facing challenges in generating long-length sentences [7, 24]. Therefore, Jia et al. [59] proposed an extension of LSTM called guided LSTM (gLSTM). This gLSTM can generate long sentences. In this architecture, it adds global semantic information to each gate and cell state of LSTM. It also considers different length normalization strategies to control the length of captions. Semantic information is extracted in different ways. First, it uses a cross-modal retrieval task for retrieving image captions, and then semantic information is extracted from these captions. The semantic-based information can also be extracted using a multimodal embedding space.

Mao et al. [92] proposed a special type of text generation method for images. This method can generate a description for a specific object or region that is called a referring expression [37, 46, 68, 102, 103, 136, 141]. Using this expression, it can then infer the object or region that is being described. Therefore, a generated description or expression is quite unambiguous. In order to address the referring expression, this method uses a new dataset called the ReferIt dataset [68] based on the popular MS COCO dataset.

Previous CNN-RNN-based image captioning methods use LSTMs that are unidirectional and relatively shallow in depth. In unidirectional language generation techniques, the next word is predicted based on visual context and all the previous textual contexts. Unidirectional LSTM cannot generate contextually well-formed captions. Moreover, recent object detection and classification methods [73, 127] show that deep, hierarchical methods are better at learning than shallower ones. Wang et al. [144] proposed a deep bidirectional LSTM-based method for image captioning. This method is capable of generating contextually and semantically rich image captions. The proposed architecture consists of a CNN and two separate LSTM networks. It can utilize both past and future context information to learn long-term visual language interactions.

3.4.2 Compositional Architecture-Based Image Captioning. Compositional architecture-based methods are composed of several independent functional building blocks: First, a CNN is used to extract the semantic concepts from the image. Then a language model is used to generate a set of candidate captions. In generating the final caption, these candidate captions are reranked using a deep multimodal similarity model.

A typical method of this category maintains the following steps:

(1) Image features are obtained using a CNN.
(2) Visual concepts (e.g., attributes) are obtained from visual features.

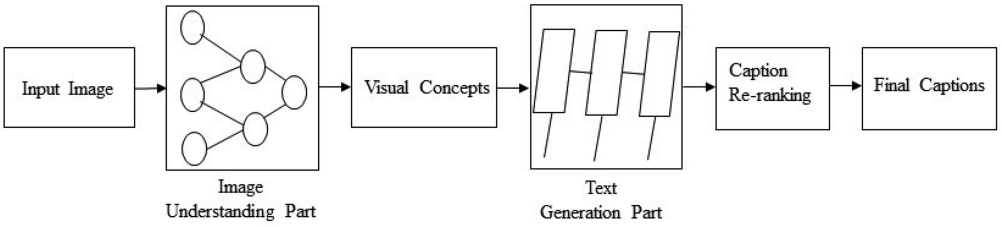


Fig. 6. A block diagram of a compositional network-based captioning.

(3) Multiple captions are generated by a language model using the information of Step 1 and Step 2.
(4) The generated captions are reranked using a deep multimodal similarity model to select high-quality image captions.

A common block diagram of compositional network-based image captioning methods is given in Figure 6.

Fang et al. [33] introduced generation-based image captioning. It uses visual detectors, a language model, and a multimodal similarity model to train the model on an image captioning dataset. Image captions can contain nouns, verbs, and adjectives. A vocabulary is formed using the 1,000 most common words from the training captions. The system works with the image subregions rather that the full image. Convolutional neural networks (both AlexNet [73] and VGG16Net) are used for extracting features for the subregions of an image. The features of subregions are mapped with the words of the vocabulary that are likely to be contained in the image captions. Multiple instance learning (MIL) [96] is used to train the model for learning discriminative visual signatures of each word. A maximum entropy (ME) [12] language model is used for generating image captions from these words. Generated captions are ranked by a linear weighting of sentence features. Minimum error rate training (MERT) [106] is used to learn these weights. Similarities between image and sentence can be easily measured using a common vector representation. Image and sentence fragments are mapped with the common vector representation by a deep multimodal similarity model (DMSM). It achieves a significant improvement in choosing high-quality image captions.

Until now, a significant number of methods have achieved satisfactory progress in generating image captions. The methods use training and testing samples from the same domain. Therefore, there is no certainty that these methods can perform well in open-domain images. Moreover, they are only good at recognizing generic visual content. There are certain key entities such as celebrities and landmarks that are out of their scope. The generated captions of these methods are evaluated on automatic metrics such as BLEU [110], METEOR [1], and CIDEr [139]. These evaluation metrics have already shown good results on these methods. However, in terms of performance, there exists a large gap between the evaluation of the metrics and human judgment of evaluation [20, 30, 74]. If it is considered real-life entity information, the performance could be weaker. However, Tran et al. [135] introduced a different image captioning method. This method is capable of generating image captions even for open-domain images. It can detect a diverse set of visual concepts and generate captions for celebrities and landmarks. It uses an external knowledge base, Freebase [16], in recognizing a broad range of entities such as celebrities and landmarks. A series of human judgments are applied for evaluating the performances of generated captions. In experiments, it uses three datasets: MS COCO, Adobe-MIT FiveK [19], and images from Instagram. The images of the MS COCO dataset were collected from the same domain, but the images of other datasets were chosen from an open domain. The method achieves notable performances especially on the challenging Instagram dataset.

Ma et al. [90] proposed another compositional network-based image captioning method. This method uses structural words <object, attribute, activity, scene> to generate semantically meaningful descriptions. It also uses a multitask method similar to the multiple instance learning method [33] and multilayer optimization method [52] to generate structural words. An LSTM encoder-decoder-based machine translation method [131] is then used to translate the structural words into image captions.

Wang et al. [146] proposed a parallel-fusion RNN-LSTM architecture for image caption generation. The architecture of the method divides the hidden units of RNN and LSTM into a number of same-size parts. The parts work in parallel with corresponding ratios to generate image captions.

3.5 Others

Attention-based, semantic-concept-based, and novel object-based methods and stylized captions are put together into the "others" group because these categories are independent of other methods.

3.5.1 Attention-Based Image Captioning. Neural encoder-decoder-based approaches were mainly used in machine translation [131]. Following these trends, they have also been used for the task of image captioning and found very effective. In image captioning, a CNN is used as an encoder to extract the visual features from the input image and an RNN is used as a decoder to convert this representation word by word into a natural language description of the image. However, these methods are unable to analyze the image over time while they generate the descriptions for the image. In addition to this, the methods do not consider the spatial aspects of the image that are relevant to the parts of the image captions. Instead, they generate captions considering the scene as a whole. Attention-based mechanisms are becoming increasingly popular in deep learning because they can address these limitations. They can dynamically focus on the various parts of the input image while the output sequences are being produced.

A typical method of this category adopts the following steps:

(1) Image information is obtained based on the whole scene by a CNN.
(2) The language generation phase generates words or phrases based on the output of Step 1.
(3) Salient regions of the given image are focused in each time step of the language generation model based on generated words or phrases.
(4) Captions are updated dynamically until the end state of the language generation model.

A block diagram of the attention-based image captioning method is shown in Figure 7.

Xu et al. [152] were the first to introduce an attention-based image captioning method. The method describes the salient contents of an image automatically. The main difference between the attention-based methods and other methods is that they can concentrate on the salient parts of the image and generate the corresponding words at the same time. This method applies two different techniques: stochastic hard attention and deterministic soft attention to generate attention. Most CNN-based approaches use the top layer of ConvNet for extracting information of the salient objects from the image. A drawback of these techniques is that they may lose certain information that is useful to generate detailed captions. In order to preserve the information, the attention method uses features from the lower convolutional layer instead of the fully connected layer.

Jin et al. [61] proposed another attention-based image captioning method. This method is capable of extracting the flow of abstract meaning based on the semantic relationship between visual information and textual information. It can also obtain higher-level semantic information by proposing a scene-specific context. The main difference between this method and other attention-based methods is that it introduces multiple visual regions of an image at multiple scales. This

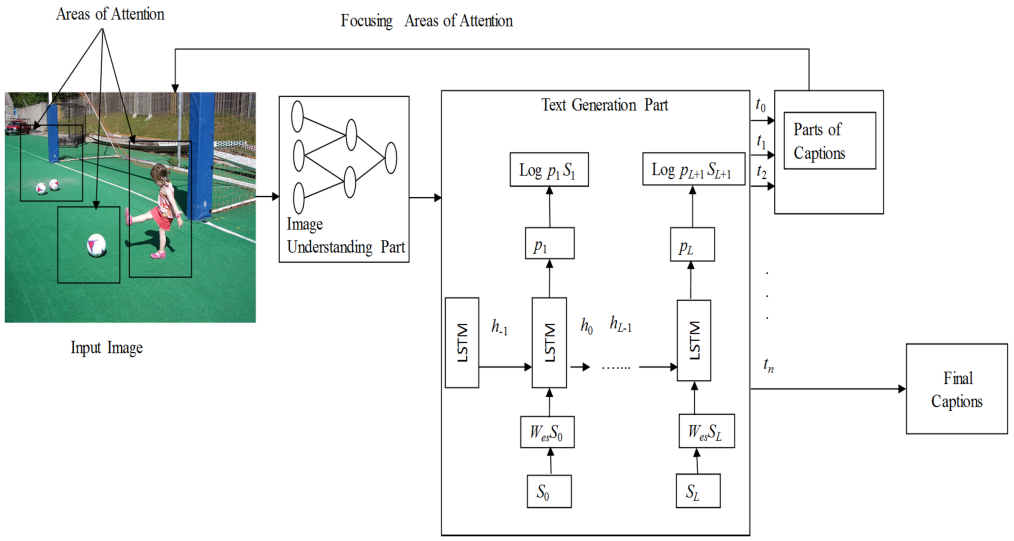


Fig. 7. A block diagram of a typical attention-based image captioning technique.

technique can extract proper visual information of a particular object. For extracting a scene-specific context, it first uses the latent Dirichlet allocation (LDA) [14] for generating a dictionary from all the captions of the dataset. Then a multilayer perceptron is used to predict a topic vector for every image. A scene-factored LSTM that has two stacked layers is used to generate a description for the overall context of the image.

Wu et al. [151] proposed a review-based attention method for image captioning. It introduces a review model that can perform multiple review steps with attention on CNN hidden states. The output of the CNN is a number of fact vectors that can obtain the global facts of the image. The vectors are given as input to the attention mechanism of the LSTM. For example, a reviewer module can first review: what are the objects in the image? Then it can review the relative positions of the objects and another review can extract the information of the overall context of the image. This information is passed to the decoder to generate image captions.

Pedersoli et al. [112] proposed an area-based attention mechanism for image captioning. Previous attention-based methods map image regions only to the state of the RNN language model. However, this approach associates image regions with caption words given the RNN state. It can predict the next caption word and corresponding image region in each time step of RNN. It is capable of predicting the next word as well as corresponding image regions in each time step of RNN for generating image captions. In order to find the areas of attention, previous attention-based image caption methods used either the position of the CNN activation grid or object proposals. In contrast, this method uses an end-to-end trainable convolutional spatial transformer along with a CNN activation gird and object proposal methods. A combination of these techniques helps this method to compute image-adaptive areas of attention. In experiments, the method shows that this new attention mechanism together with the spatial transformer network can produce high-quality image captions.

Lu et al. [88] proposed another attention-based image captioning method. The method is based on an adaptive attention model with a visual sentinel. Current attention-based image captioning methods focus on the image in every time step of RNN. However, there are some words or phrases (e.g., "a," "of") that do not need to attend visual signals. Moreover, these unnecessary visual signals

could affect the caption generation process and degrade the overall performance. Therefore, their proposed method can determine when it will focus on an image region and when it will just focus on a language generation model. Once it determines to focus on the image, then it must have to choose the spatial location of the image. The first contribution of this method is to introduce a novel spatial attention method that can compute spatial features from the image. Then, in their adaptive attention method, they introduced a new LSTM extension. Generally, an LSTM works as a decoder that can produce a hidden state at every time step. However, this extension is capable of producing an additional visual sentinel that provides a fall-back option to the decoder. It also has a sentinel gate that can control how much information the decoder will get from the image.

While attention-based methods look to find the different areas of the image at the time of generating words or phrases for image captions, the attention maps generated by these methods cannot always correspond to the proper region of the image. It can affect the performance of image caption generation. Liu et al. [84] proposed a method for neural image captioning. This method can evaluate and correct the attention map at time steps. Correctness means making a consistent map between image regions and generated words. In order to achieve these goals, this method introduced a quantitative evaluation metric to compute the attention maps. It uses the Flickr30K entity dataset [113] and the MS COCO [83] dataset for measuring both a ground-truth attention map and semantic labelings of image regions. In order to learn a better attention function, it proposed a supervised attention model. Two types of supervised attention models are used here: strong supervision with alignment annotation and weak supervision with semantic labeling. In the strong supervision with alignment annotation model, it can directly map ground-truth words to a region. However, ground-truth alignment is not always possible because collecting and annotating data is often very expensive. Weak supervision is performed to use bounding box or segmentation masks on the MS COCO dataset. In experiments, the method shows that the supervised attention model performs better in mapping attention as well as image captioning.

Chen et al. [21] proposed another attention-based image captioning method. This method considers both spatial and channel-wise attentions to compute an attention map. The existing attention-based image captioning methods only consider spatial information for generating an attention map. A common drawback of these spatial attention methods is that they compute weighted pooling only on attentive feature maps. As a result, these methods lose the spatial information gradually. Moreover, they use the spatial information only from the last conv-layer of the CNN. The receptive field regions of this layer are quite large, which limits the gap between the regions. Therefore, they do not get significant spatial attention for an image. However, in this method, CNN features are extracted not only from spatial locations but also from different channels and multiple layers. Therefore, it gets significant spatial attention. In addition to this, in this method, each filter of a convolutional layer acts as a semantic detector [159], while other methods use external sources for obtaining semantic information.

In order to reduce the gap between the human-generated description and machine-generated description, Tavakoli et al. [134] introduced an attention-based image captioning method. This is a bottom-up saliency-based attention model that can take advantage of comparisons with other attention-based image captioning methods. It found that humans first describe the more important objects before less important ones. It also shows that the method performs better on unseen data.

Most previous image captioning methods applied a top-down approach for constructing a visual attention map. These mechanisms typically focused on some selective regions obtained from the output of one or two layers of a CNN. The input regions are of the same size and have the same shape of receptive field. This approach has little consideration for the content of the image. However, the method of Anderson et al. [4] applied both top-down and bottom-up approaches. The bottom-up attention mechanism uses Faster R-CNN [116] for region proposals that can select

salient regions of an image. Therefore, this method can attend to both object-level regions and other salient image regions.

Park et al. [111] introduced a different type of attention-based image captioning method. This method can generate image captions addressing personal issues of an image. It mainly considers two tasks: hashtag prediction and post generation. This method uses a context sequence memory network (CSMN) to obtain the context information from the image. Descriptions of an image from a personalized view have a lot of applications in social media networks. For example, everyday people share a lot of images as posts on Facebook, Instagram, or other social media. Photo taking or uploading is a very easy task. However, describing the images is not easy because it requires a theme, sentiment, and context of the image. Therefore, the method considers past knowledge about the user's vocabularies or writing styles from the prior documents for generating image descriptions. In order to work with this new type of image captioning, the CSMN method has three contributions: First, the memory of this network can work as a repository and retain multiple types of context information. Second, the memory is designed in such a way that it can store all the previously generated words sequentially. As a result, it does not suffer from the vanishing gradient problem. Third, the proposed CNN can correlate with multiple memory slots, which is helpful for understanding contextual concepts.

Attention-based methods have already shown good performance and efficiency in image captioning as well as other computer vision tasks. However, attention maps generated by these attention-based methods are only machine dependent. They do not consider any supervision from human attention. This creates the necessity to think about the gaze information and whether it can improve the performance of these attention methods in image captioning. Gaze indicates the cognition and perception of humans about a scene. Human gaze can identify the important locations of objects in an image. Thus, gaze mechanisms have already shown their potential performances in eye-based user modeling [18, 35, 109, 122, 124], object localization [100] or recognition [67], and holistic scene understanding [158, 160]. However, Sugano et al. [129] claimed that gaze information has not yet been integrated in image captioning methods. This method introduced human gaze with the attention mechanism of deep neural networks in generating image captions. The method incorporates human gaze information into an attention-based LSTM model [152]. For experiments, it uses the SALICON dataset [60] and achieves good results.

3.5.2 Semantic-Concept-Based Image Captioning. Semantic-concept-based methods selectively attend to a set of semantic concept proposals extracted from the image. These concepts are then combined into hidden states and the outputs of recurrent neural networks.

The methods in this category follow the following steps:

(1) The CNN-based encoder is used to encode the image features and semantic concepts.
(2) Image features are fed into the input of the language generation model.
(3) Semantic concepts are added to the different hidden states of the language model.
(4) The language generation part produces captions with semantic concepts.

A typical block diagram of this category is shown in Figure 8.

Karpathy et al. extended their method [66] in [65]. The later method can generate natural language descriptions for both images and their regions. This method employs a novel combination of CNN over the image regions, bidirectional recurrent neural networks over sentences, and a common multimodal embedding that associates the two modalities. It also demonstrates a multimodal recurrent neural network architecture that utilizes the resultant alignments to train the model for generating novel descriptions of image regions. In this method, dependency tree relations (DTRs) are used to train to map the sentence segments with the image regions that have a fixed window

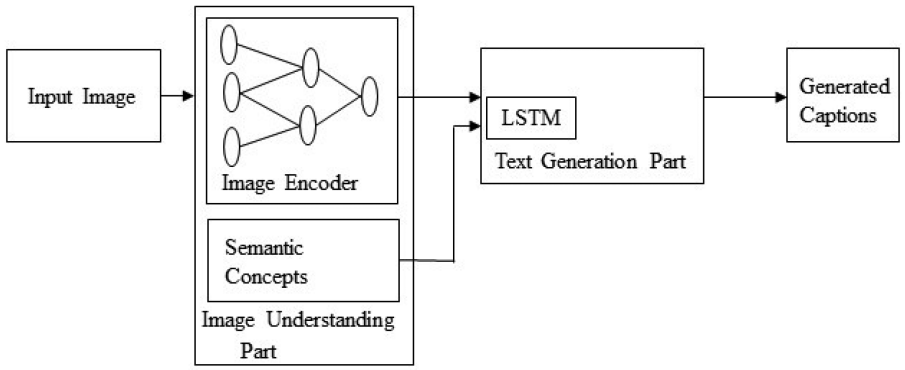


Fig. 8. A block diagram of a semantic-concept-based image captioning.

context. In contrast to their previous method, this method uses a bidirectional neural network to obtain word representations in the sentence. It considers contiguous fragments of sentences to align in the embedding space, which is more meaningful, interpretable, and not fixed in length. Generally an RNN considers the current word and the contexts from all the previously generated words for estimating a probability distribution of the next word in a sequence. However, this method extends it for considering the generative process on the content of an input image. This addition is simple, but it makes it very effective for generating novel image captions.

Attributes of an image are considered as rich semantic cues. The method of Yao et al. [155] has different architectures to incorporate attributes with image representations. Mainly, two types of architectural representations are introduced here. In the first group, it inserts only attributes to the LSTM or image representations to the LSTM first and then attributes, and vice versa. In the second group, it can control the time step of the LSTM. It decides whether image representation and attributes will be input once or every time step. These variants of architectures are tested on the MS COCO dataset and common evaluation metrics.

You et al. [156] proposed a semantic-attention-based image captioning method. The method provides a detailed, coherent description of semantically important objects. The top-down paradigms [23, 31, 65, 93, 94, 142, 152] are used for extracting visual features first and then convert them into words. In bottom-up approaches, [32, 34, 74, 76, 78, 80], visual concepts (e.g., regions, objects, and attributes) are extracted first from various aspects of an image and then combined. Fine details of an image are often very important for generating a description of an image. Top-down approaches have limitations in obtaining fine details of the image. Bottom-up approaches are capable of operating on any image resolution and therefore they can do work on fine details of the image. However, they have problems in formulating an end-to-end process. Therefore, semantic-based attention models applied both top-down and bottom-up approaches for generating image captions. In top-down approaches, the image features are obtained using the last 1,024-dimensional convolutional layer of the GoogleNet [133] CNN model. The visual concepts are collected using different nonparametric and parametric methods. The nearest neighbor image retrieval technique is used for computing nonparametric visual concepts. The fully convolutional network (FCN) [86] is used to learn attributes from local patches for parametric attribute prediction. Although Xu et al. [152] considered attention-based captioning, it works on fixed and predefined spatial location. However, this semantic-attention-based method can work on any resolution and any location of the image. Moreover, this method also considers a feedback process that accelerates to generate better image captions.

Previous image captioning methods do not include high-level semantic concepts explicitly. However, Wu et al. [150] proposed a high-level semantic-concept-based image captioning. It uses an intermediate attribute prediction layer in a neural-network-based CNN-LSTM framework. First, attributes are extracted by a CNN-based classifier from training image captions. Then these attributes are used as high-level semantic concepts in generating semantically rich image captions.

Recent semantic-concept-based image captioning methods [150, 156] applied the semantic concept detection process [40] to obtain explicit semantic concepts. They use these high-level semantic concepts in the CNN-LSTM-based encoder-decoder and achieve significant improvements in image captioning. However, they have problems in generating semantically sound captions. They cannot distribute semantic concepts evenly in the whole sentence. For example, Wu et al. [150] consider the initial state of the LSTM to add semantic concepts. Moreover, it encodes a visual feature vector or an inferred scene vector from the CNN and then feeds it to the LSTM for generating captions. However, Gan et al. [41] introduced a semantic compositional network (SCN) for image captioning. In this method, a semantic concept vector is constructed from all the probable concepts (called tags here) found in the image. This semantic vector has more potential than the visual feature vector and scene vector and can generate captions covering the overall meaning of the image. This is called a compositional network because it can compose most semantic concepts.

Existing LSTM-based image captioning methods have limitations in generating a diverse set of captions because they have to predict the next word in a predefined word-by-word format. However, a combination of attributes and subjects and their relationship in a sentence irrespective of their location can generate a broad range of image captions. Wang et al. [148] proposed a method that locates the objects and their interactions first and then identifies and extracts the relevant attributes to generate image captions. The main aim of this method is to decompose the ground-truth image captions into two parts: skeleton sentences and attribute phrases. The method is also called Skeleton Key. The architecture of this method has ResNet [53] and two LSTMs called Skel-LSTM and Attr-LSTM. During training, skeleton sentences are trained by the Skel-LSTM network and attribute phrases are trained by the Attr-LSTM network. In the testing phase, skeleton sentences are generated first that contain the words for the main objects of the image and their relationships. Then these objects look back through the image again to obtain the relevant attributes. It is tested on the MS COCO dataset and a new Stock3M dataset and can generate more accurate and novel captions.

3.5.3 Novel-Object-Based Image Captioning. Despite recent deep-learning-based image captioning methods having achieved promising results, they largely depend on the paired image and sentence caption datasets. These types of methods can only generate a description of the objects within the context. Therefore, the methods require a large set of training image-sentence pairs. Novel-object-based image captioning methods can generate descriptions of novel objects that are not present in paired image-caption datasets.

The methods of this category have the following general steps:

(1) A separate lexical classifier and a language model are trained on unpaired image data and unpaired text data.
(2) A deep caption model is trained on paired image caption data.
(3) Finally, both models are combined together to train jointly, which can generate captions for novel object.

A simple block diagram of a novel-object-based image captioning method is given in Figure 9.

Current image captioning methods are trained on image-caption paired datasets. As a result, if there are unseen objects in the test images, they cannot present them in their generated captions.

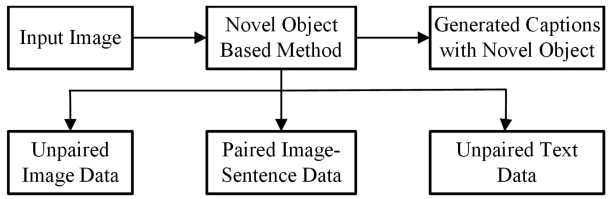


Fig. 9. A block diagram of a typical novel-object-based image captioning.

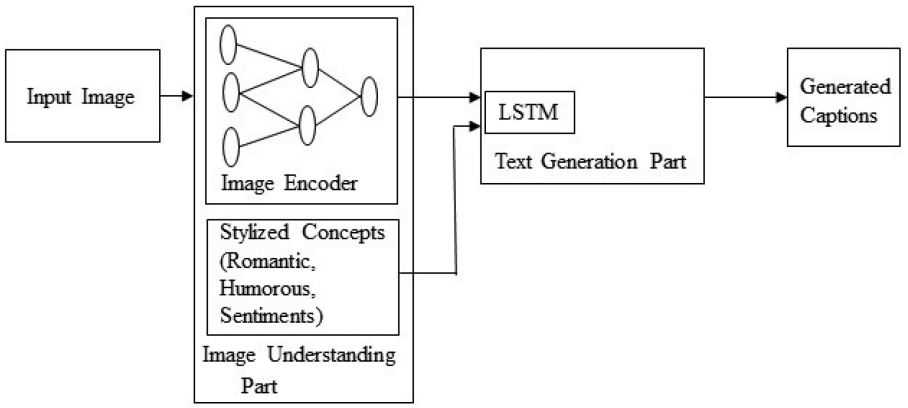


Fig. 10. A block diagram of image captioning based on different styles.

Anne et al. [6] proposed a deep compositional captioner (DCC) that can represent the unseen objects in generated captions.

Yao et al. [154] proposed a copying mechanism to generate a description for novel objects. This method uses a separate object recognition dataset to develop classifiers for novel objects. It integrates the appropriate words in the output captions by a decoder RNN with a copying mechanism. The architecture of the method adds a new network to recognize the unseen objects from unpaired images and incorporate them with the LSTM to generate captions.

Generating captions for the unseen images is a challenging research problem. Venugopalan et al. [140] introduced a novel object captioner (NOC) for generating captions for unseen objects in the image. They used external sources for recognizing unseen objects and learning semantic knowledge.

3.5.4 Stylized Caption. Existing image captioning systems generate captions based on only the image content, which can also be called factual descriptions. They do not consider the stylized part of the text separately from other linguistic patterns. However, the stylized captions can be more expressive and attractive than only the flat description of an image.

The methods of this category have the following general steps:

(1) The CNN-based image encoder is used to obtain the image information.
(2) A separate text corpus is prepared to extract various stylized concepts (e.g., romantic, humorous) from training data.
(3) The language generation part can generate stylized and attractive captions using the information of Step 1 and Step 2.

A simple block diagram of stylized image captioning is given in Figure 10.

Such captions have become popular because they are particularly valuable for many real-world applications. For example, everyday people are uploading a lot of photos in different social media. The photos need stylized and attractive descriptions. Gan et al. [39] proposed a novel image captioning system called StyleNet. This method can generate attractive captions, adding various styles. The architecture of this method consists of a CNN and a factored LSTM that can separate factual and style factors from the captions. It uses multitask sequence-to-sequence training [89] for identifying the style factors and then adds these factors at runtime to generate attractive captions. More interestingly, it uses an external monolingual stylized language corpus for training instead of paired images. However, it uses a new stylized image caption dataset called FlickrStyle10K and can generate captions with different styles.

Existing image captioning methods consider the factual description about the objects and scene and their interactions in an image in generating image captions. In our day-to-day conversations, communications, interpersonal relationships, and decision making, we use various stylized and nonfactual expressions such as emotions, pride, and shame. However, Mathews et al. [97] claimed that automatic image descriptions are missing these nonfactual aspects. Therefore, they proposed a method called SentiCap. This method can generate image descriptions with positive or negative sentiments. It introduces a novel switching RNN model that combines two CNN+RNNs running in parallel. In each time step, this switching model generates the probability of switching between two RNNs. One generates captions considering the factual words and the other considers the words with sentiments. It then takes inputs from the hidden states of both RNNs for generating captions. This method can generate captions successfully given the appropriate sentiments.

3.6 LSTM versus Others

Image captioning intersects computer vision and natural language processing (NLP) research. NLP tasks, in general, can be formulated as a sequence-to-sequence learning. Several neural language models such as neural probabilistic language models [11], log-bilinear models [105], skip-gram models [98], and RNNs [99] have been proposed for learning sequence-to-sequence tasks. RNNs have widely been used in various sequence learning tasks. However, traditional RNNs suffer from vanishing and exploding gradient problems and cannot adequately handle long-term temporal dependencies.

LSTM [54] networks are a type of RNN that have special units in addition to standard units. LSTM units use a memory cell that can maintain information in memory for long periods of time. In recent years, LSTM-based models have dominantly been used in sequence-to-sequence learning tasks. Another network, the gated recurrent unit (GRU) [25], has a similar structure to LSTM but it does not use separate memory cells and uses fewer gates to control the flow of information.

However, LSTMs ignore the underlying hierarchical structure of a sentence. They also require significant storage due to long-term dependencies through a memory cell. In contrast, CNNs can learn the internal hierarchical structure of the sentences and are faster in processing than LSTMs. Therefore, recently, convolutional architectures have been used in other sequence-to-sequence tasks, e.g., conditional image generation [137] and machine translation [42, 43, 138].

Inspired by the above success of CNNs in sequence learning tasks, Gu et al. [51] proposed a CNN language model-based image captioning method. This method uses a language CNN for statistical language modeling. However, the method cannot model the dynamic temporal behavior of the language model only using a language CNN. It combines a recurrent network with the language CNN to model the temporal dependencies properly. Aneja et al. [5] proposed a convolutional architecture for the task of image captioning. They use a feed-forward network without any recurrent function. The architecture of the method has four components: (1) input embedding layer, (2) image embedding layer, (3) convolutional module, and (4) output embedding layer. It also uses

Ground Truth Caption: Two brown bears playing in a field together.

Generated Caption: Two brown bears playing on top of a lush green field.

Ground Truth Caption: A plate of breakfast food with a silver tea pot.

Generated Caption: A close up of a plate of food with a folk and a knife on a table.

Fig. 11. Captions generated by Wu et al. [149] on some sample images from the MS COCO dataset.

an attention mechanism to leverage spatial image features. They evaluate their architecture on the challenging MS COCO dataset and show comparable performance to an LSTM-based method on standard metrics.

Wang et al. [147] proposed another CNN+CNN-based image captioning method. It is similar to the method of Aneja et al. except that it uses a hierarchical attention module to connect the vision CNN with the language CNN. The authors of this method also investigate the use of various hyperparameters, including the number of layers and the kernel width of the language CNN. They show that the influence of the hyperparameters can improve the performance of the method in image captioning.

4 DATASETS AND EVALUATION METRICS

A number of datasets are used for training, testing, and evaluating the image captioning methods. The datasets differ in various perspectives, such as the number of images, the number of captions per image, format of the captions, and image size. Three datasets, Flickr8K [55], Flickr30K [113], and MS COCO [83], are popularly used. These datasets, together with others, are described in Section 4.1. In this section, we show sample images with their captions generated by image captioning methods on the MS COCO, Flickr30K, and Flickr8K datasets. A number of evaluation metrics are used to measure the quality of the generated captions compared to the ground truth. Each metric applies its own technique for computation and has distinct advantages. The commonly used evaluation metrics are discussed in Section 4.2. A summary of deep-learning-based image captioning methods with their datasets and evaluation metrics are listed in Table 2.

4.1 Datasets

4.1.1 MS COCO Dataset. The Microsoft COCO Dataset [83] is a very large dataset for image recognition, segmentation, and captioning. There are various features of the MS COCO dataset such as object segmentation, recognition in context, multiple objects per class, more than 300,000 images, more than 2 million instances, 80 object categories, and five captions per image. Many image captioning methods [26, 39, 61, 112, 119, 126, 135, 144, 149, 151, 156] use the dataset in their experiments. For example, Wu et al. [149] use the MS COCO dataset in their method, and the generated captions of two sample images are shown in Figure 11.

4.1.2 Flickr30K Dataset. Flickr30K [113] is a dataset for automatic image description and grounded language understanding. It contains 30K images collected from Flickr with 158K captions provided by human annotators. It does not provide any fixed split of images for training,

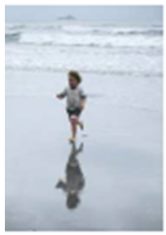
Table 2. An Overview of Methods, Datasets, and Evaluation Metrics

Reference	Datasets	Evaluation Metrics
Kiros et al. 2014 [69]	IAPR TC-12,SBU	BLEU, PPLX
Kiros et al. 2014 [70]	Flickr 8K, Flickr 30K	R@K, mrank
Mao et al. 2014 [95]	IAPR TC-12, Flickr 8K/30K	BLEU, R@K, mrank
Karpathy et al. 2014 [66]	PASCAL1K, Flickr 8K/30K	R@K, mrank
Mao et al. 2015 [94]	IAPR TC-12, Flickr 8K/30K, MS COCO	BLEU, R@K, mrank
Chen et al. 2015 [23]	PASCAL, Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Fang et al. 2015 [33]	PASCAL, MS COCO	BLEU, METEOR, PPLX
Jia et al. 2015 [59]	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Karpathy et al. 2015 [65]	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Vinyals et al. 2015 [142]	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Xu et al. 2015 [152]	Flickr 8K/30K, MS COCO	BLEU, METEOR
Jin et al. 2015 [61]	Flickr 8K/30K, MS COCO	BLEU, METEOR, ROUGE, CIDEr
Wu et al. 2016 [151]	MS COCO	BLEU, METEOR, CIDEr
Sugano et at. 2016 [129]	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Mathews et al. 2016 [97]	MS COCO, SentiCap	BLEU, METEOR, ROUGE, CIDEr
Wang et al. 2016 [144]	Flickr 8K/30K, MS COCO	BLEU, R@K
Johnson et al. 2016 [62]	Visual Genome	METEOR, AP, IoU
Mao et al. 2016 [92]	ReferIt	BLEU, METEOR, CIDEr
Wang et al. 2016 [146]	Flickr 8K	BLEU, PPL, METEOR
Tran et al. 2016 [135]	MS COCO, Adobe-MIT, Instagram	Human Evaluation
Ma et al. 2016 [90]	Flickr 8k, UIUC	BLEU, R@K
You et al. 2016 [156]	Flickr 30K, MS COCO	BLEU, METEOR, ROUGE, CIDEr
Yang et al. 2016 [153]	Visual Genome	METEOR, AP, IoU
Anne et al. 2016 [6]	MS COCO, ImageNet	BLEU, METEOR
Yao et al. 2017 [155]	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Lu et al. 2017 [88]	Flickr 30K, MS COCO	BLEU, METEOR, CIDEr
Chen et al. 2017 [21]	Flickr 8K/30K, MS COCO	BLEU, METEOR, ROUGE, CIDEr
Gan et al. 2017 [41]	Flickr 30K, MS COCO	BLEU, METEOR, CIDEr
Pedersoli et al. 2017 [112]	MS COCO	BLEU, METEOR, CIDEr
Ren et al. 2017 [119]	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Park et al. 2017 [111]	Instagram	BLEU, METEOR, ROUGE, CIDEr
Wang et al. 2017 [148]	MS COCO, Stock3M	SPICE, METEOR, ROUGE, CIDEr
Tavakoli et al. 2017 [134]	MS COCO, PASCAL 50S	BLEU, METEOR, ROUGE, CIDEr
Liu et al. 2017 [84]	Flickr 30K, MS COCO	BLEU, METEOR
Gan et al. 2017 [39]	FlickrStyle10K	BLEU, METEOR, ROUGE, CIDEr
Dai et al. 2017 [26]	Flickr 30K, MS COCO	E-NGAN, E-GAN, SPICE, CIDEr
Shetty et al. 2017 [126]	MS COCO	Human Evaluation, SPICE, METEOR
Liu et al. 2017 [85]	MS COCO	SPIDEr, Human Evaluation
Gu et al. 2017 [51]	Flickr 30K, MS COCO	BLEU, METEOR, CIDEr, SPICE
Yao et al. 2017 [154]	MS COCO, ImageNet	METEOR
Rennie et al. 2017 [120]	MS COCO	BLEU, METEOR, CIDEr, ROUGE
Vsub et al. 2017 [140]	MS COCO, ImageNet	METEOR
Zhang et al. 2017 [161]	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Wu et al. 2018 [150]	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Aneja et al. 2018 [5]	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Wang et al. 2018 [147]	MS COCO	BLEU, METEOR, ROUGE, CIDEr

Generated Caption: A young baseball player is sliding into a base.

Generated Caption: A young boy playing with a soccer ball in a field.

Fig. 12. Captions generated by Chen et al. [22] on some sample images from the Flickr30K dataset.



Ground Truth Caption: A little boy runs away from the approaching waves of the ocean.

Generated Caption: A young boy is running on the beach.

Ground Truth Caption: A brunette girl wearing sunglasses and a yellow shirt.

Generated Caption: A woman in a black shirt and sunglasses smiles.

Fig. 13. Captions generated by Jia et al. [59] on some sample images from the Flickr8K dataset.

testing, and validation. Researchers can choose their own choice of numbers for training, testing, and validation. The dataset also contains detectors for common objects, a color classifier, and a bias toward selecting larger objects. Image captioning methods such as [22, 65, 142, 144, 150] use this dataset for their experiments. The generated captions by Chen et al. [22] of two sample images of the dataset are shown in Figure 12.

4.1.3 Flickr8K Dataset. Flickr8K [55] is a popular dataset and has 8,000 images collected from Flickr. The training data consists of 6,000 images and the test and development data, and each consists of 1,000 images. Each image in the dataset has five reference captions annotated by humans. A number of image captioning methods [21, 59, 61, 144, 150, 152] have performed experiments using the dataset. Two sample results by Jia et al. [59] on this dataset are shown in Figure 13.

4.1.4 Visual Genome Dataset. The Visual Genome dataset [72] is another dataset for image captioning. Image captioning requires not only recognizing the objects of an image but also reasoning about their interactions and attributes. Unlike the first three datasets where a caption is given to the whole scene, the Visual Genome dataset has separate captions for multiple regions in an image. The dataset has seven main parts: region descriptions, objects, attributes, relationships, region graphs, scene graphs, and question-answer pairs. The dataset has more than 108K images. Each image contains an average of 35 objects, 26 attributes, and 21 pairwise relationships between objects.

4.1.5 Instagram Dataset. Tran et al. [135] and Park et al. [111] created two datasets using images from Instagram, which is a photo-sharing social networking service. The dataset of Tran et al. has about 10K images, which are mostly from celebrities. However, Park et al. used their dataset for hashtag prediction and postgeneration tasks in social media networks. This dataset contains 1.1M posts on a wide range of topics and a long hashtag list from 6.3K users.

4.1.6 IAPR TC-12 Dataset. The IAPR TC-12 dataset [50] has 20K images. The images are collected from various sources such as sports, photographs of people, animals, landscapes, and many other locations around the world. The images of this dataset have captions in multiple languages. Images have multiple objects as well.

4.1.7 Stock3M Dataset. The Stock3M dataset has 3,217,654 images uploaded by users and it is 26 times larger than the MS COCO dataset. The images of this dataset have a diversity of content.

4.1.8 MIT-Adobe FiveK Dataset. The MIT-Adobe FiveK [19] dataset consists of 5,000 images. These images contain a diverse set of scenes, subjects, and lighting conditions and they are mainly about people, nature, and manmade objects.

4.1.9 FlickrStyle10K Dataset. The FlickrStyle10K dataset has 10,000 Flickr images with stylized captions. The training data consists of 7,000 images. The validation and test data consist of 2,000 and 1,000 images, respectively. Each image contains romantic, humorous, and factual captions.

4.2 Evaluation Metrics

4.2.1 BLEU. BLEU (Bilingual evaluation understudy) [110] is a metric that is used to measure the quality of machine-generated text. Individual text segments are compared with a set of reference texts and scores are computed for each of them. In estimating the overall quality of the generated text, the computed scores are averaged. However, syntactical correctness is not considered here. The performance of the BLEU metric is varied depending on the number of reference translations and the size of the generated text. Subsequently, Papineni et al. introduced a modified precision metric. This metric uses n-grams. BLEU is popular because it is a pioneer in automatic evaluation of machine-translated text and has a reasonable correlation with human judgments of quality [20, 29]. However, it has a few limitations, such as that BLEU scores are good only if the generated text is short [20]. There are some cases where an increase in BLEU score does not mean that the quality of the generated text is good [82].

4.2.2 ROUGE. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [81] is a set of metrics that are used for measuring the quality of text summaries. It compares word sequences, word pairs, and n-grams with a set of reference summaries created by humans. Different types of ROUGE such as ROUGE-1, 2, ROUGE-W, and ROUGE-SU4 are used for different tasks. For example, ROUGE-1 and ROUGE-W are appropriate for single-document evaluation, whereas ROUGE-2 and ROUGE-SU4 have good performance in short summaries. However, ROUGE has problems in evaluating multidocument text summaries.

4.2.3 METEOR. METEOR (Metric for Evaluation of Translation with Explicit ORdering) [9] is another metric used to evaluate the machine-translated language. Standard word segments are compared with the reference texts. In addition to this, stems of a sentence and synonyms of words are also considered for matching. METEOR can make a better correlation at the sentence or the segment level.

4.2.4 CIDEr. CIDEr (Consensus-based Image Description Evaluation) [139] is an automatic consensus metric for evaluating image descriptions. Most existing datasets have only five captions

per image. Previous evaluation metrics work with this small number of sentences and are not enough to measure the consensus between generated captions and human judgment. However, CIDEr achieves human consensus using term frequency-inverse document frequency (TF-IDF) [121].

4.2.5 SPICE. SPICE (Semantic Propositional Image Caption Evaluation) [3] is a new caption evaluation metric based on semantic concepts. It is based on a graph-based semantic representation called a scene graph [63, 123]. This graph can extract the information of different objects and attributes and their relationships from the image descriptions.

Existing image captioning methods compute log-likelihood scores to evaluate their generated captions. They use BLEU, METEOR, ROUGE, SPICE, and CIDEr as evaluation metrics. However, BLEU, METEOR, and ROUGE are not well correlated with human assessments of quality. SPICE and CIDEr have better correlation, but they are hard to optimize. Liu et al. [85] introduced a new caption evaluation metric that is a good choice by human raters. It is developed through a combination of SPICE and CIDEr, and termed as SPIDEr. It uses a policy gradient method to optimize the metrics.

The quality of image captioning depends on the assessment of two main aspects: adequacy and fluency. An evaluation metric needs to focus on a diverse set of linguistic features to achieve these aspects. However, commonly used evaluation metrics consider only some specific features (e.g., lexical or semantic) of languages. Sharif et al. [125] proposed learning-based composite metrics for evaluation of image captions. The composite metric incorporates a set of linguistic features to achieve the two main aspects of assessment and shows improved performances.

5 COMPARISON ON BENCHMARK DATASETS AND COMMON EVALUATION METRICS

While formal experimental evaluation was left out of the scope of this article, we present a brief analysis of the experimental results and the performance of various techniques as reported. We cover three sets of results:

(1) We find that a number of methods use the first three datasets listed in Section 4.1 and a number of commonly used evaluation metrics to present the results. These results are shown in Table 3.
(2) A few methods fall into the following groups: attention-based and other deep-learning-based (reinforcement learning and GAN-based methods) image captioning. The results of such methods are shown in Tables 4 and 5, respectively.
(3) We also list the methods that provide the top two results scored on each evaluation metric on the MS COCO dataset. These results are shown in Table 6.

As shown in Table 3, on Flickr8K, Mao et al. [94] achieved 0.565, 0.386, 0.256, and 0.170 on BLEU-1, BLEU-2, BLEU-3, and BLEU-4, respectively. For the Flickr30K dataset, the scores are 0.600, 0.410, 0.280, and 0.190, respectively, which are higher than the Flickr8K scores. The highest scores were achieved on the MS COCO dataset. The higher results on a larger dataset follow the fact that a large dataset has more data and a comprehensive representation of various scenes, complexities, and their own natural context. The results of Jia et al. [59] are similar for Flickr8K and Flickr30K datasets but higher on the MS COCO dataset. The method uses visual space for mapping image features and text features. Mao et al. use multimodal space for the mapping of image features and text features. On the other hand, Jia et al. use visual space for the mapping. Moreover, the method uses an encoder-decoder architecture, where it can guide the decoder part dynamically. Consequently, this method performs better than that of Mao et al. [94].

Table 3. Performance of Different Image Captioning Methods on Three Benchmark Datasets and Commonly Used Evaluation Metrics

Dataset	Method	Category	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8K	Mao et al. 2015 [94]	MS, SL, WS	0.565	0.386	0.256	0.170	–
	Jia et al. 2015 [59]	VS, SL, WS, EDA	0.647	0.459	0.318	0.216	0.201
	Xu et al. 2015 [152]	VS, SL, WS, EDA, AB	0.670	0.457	0.314	0.213	0.203
	Wu et al. 2018 [150]	VS, SL, WS, EDA, SCB	0.740	0.540	0.380	0.270	–
Flickr30K	Mao et al. 2015 [94]	MS, SL, WS	0.600	0.410	0.280	0.190	–
	Jia et al. 2015 [59]	VS, SL, WS, EDA	0.646	0.466	0.305	0.206	0.179
	Xu et al. 2015 [152]	VS, SL, WS, EDA, AB	0.669	0.439	0.296	0.199	0.184
	Wu et al. 2018 [150]	VS, SL, WS, EDA, SCB	0.730	0.550	0.400	0.280	–
MS COCO	Mao et al. 2015 [94]	MS, SL, WS	0.670	0.490	0.350	0.250	–
	Jia et al. 2015 [59]	VS, SL, WS, EDA	0.670	0.491	0.358	0.264	0.227
	Xu et al. 2015 [152]	VS, SL, WS, EDA, AB	0.718	0.504	0.357	0.250	0.230
	Wu et al. 2018 [150]	VS, SL, WS, EDA, SCB	0.740	0.560	0.420	0.310	0.260

A dash (–) in the table indicates results are unavailable.

Table 4. Performance of Attention-Based Image Captioning Methods on MS COCO Dataset and Commonly Used Evaluation Metrics

Method	Category	MS COCO						
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Xu et al. 2015 [152], soft	VS, SL, WS, EDA, VC	0.707	0.492	0.344	0.243	0.239	–	–
Xu et al. 2015 [152], hard	VS, SL, WS, EDA, VC	0.718	0.504	0.357	0.250	0.230	–	–
Jin et al. 2015 [61]	VS, SL, WS, EDA, VC	0.697	0.519	0.381	0.282	0.235	0.509	0.838
Wu et al. 2016 [151]	VS, SL, WS, EDA, VC	–	–	–	0.290	0.237	–	0.886
Pedersoli et al. 2017 [112]	VS, SL, WS, EDA, VC	–	–	–	0.307	0.245	–	0.938

A dash (–) in the table indicates results are unavailable.

Table 5. Performance of Other Deep-Learning-Based Image Captioning Methods on MS COCO Dataset and Commonly Used Evaluation Metrics

Method	Category	MS COCO							
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Shetty et al. 2017_{GAN} [126]	VS, ODL, WS, EDA	–	–	–	–	0.239	–	–	0.167
Ren et al. 2017_{RL} [119]	VS, ODL, WS, EDA	0.713	0.539	0.403	0.304	0.251	0.525	0.937	–
Zhang et al. 2017_{RL} [161]	VS, ODL, WS, EDA	–	–	–	0.344	0.267	0.558	1.162	–

A dash (–) in the table indicates results are unavailable.

Xu et al. [152] also show better performance on the MS COCO dataset. This method outperformed that of both Mao et al. [94] and Jia et al. [59]. The main reason behind this is that it uses an attention mechanism, which focuses only on relevant objects of the image. The semantic-concept-based methods can generate semantically rich captions. Wu et al. [151] proposed a semantic-concept-based image captioning method. This method first predicts the attributes of different objects from the image and then adds these attributes with the captions that are semantically meaningful. In terms of performance, the method is superior to all the methods mentioned in Table 3.

Table 4 shows the results of attention-based based methods on the MS COCO dataset. Xu et al.'s stochastic hard attention produced better results than deterministic soft attention. However, these

Table 6. Top Two Methods Based on Different Evaluation Metrics and MS COCO Dataset (Bold and Italic Indicates the Best Result; Bold Indicates the Second-Best Result)

Method	Category	MSCOCO							
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Lu et al. 2017 [88]	VS, SL, WS, EDA, AB	***0.742***	***0.580***	**0.439**	0.332	**0.266**	–	**1.085**	–
Gan et al. 2017 [41]	VS, SL, WS, CA, SCB	**0.741**	**0.578**	***0.444***	**0.341**	0.261	–	1.041	–
Zhang et al. 2017 [161]	VS, ODL, WS, EDA	–	–	–	***0.344***	***0.267***	***0.558***	***1.162***	–
Rennie et al. 2017 [120]	VS, ODL, WS, EDA	–	–	–	.319	0.255	**0.543**	1.06	–
Yao et al. 2017 [155]	VS, SL, WS, EDA, SCB	0.734	0.567	0.430	0.326	0.254	0.540	1.00	***0.186***
Gu et al. 2017 [51]	VS, SL, WS, EDA	0.720	0.550	0.410	0.300	0.240	–	0.960	**0.176**

A dash (–) in the table indicates results are unavailable.

results were outperformed by the method of Jin et al. [61], which can update its attention based on the scene-specific context.

Wu et al. [151] and Pedersoli et al. [112] only show BLEU-4 and METEOR scores, which are higher than the aforementioned methods. The method of Wu et al. uses an attention mechanism with a review process. The review process checks the focused attention in every time step and updates it if necessary. This mechanism helps to achieve better results than the prior attention-based methods. Pedersoli et al. propose a different attention mechanism that maps the focused image regions directly with the caption words instead of the LSTM state. The behavior of this method drives it to achieve top performances among the mentioned attention-based methods in Table 4.

Reinforcement-learning (RL)-based and GAN-based methods are becoming increasingly popular. We name them as "other deep-learning-based image captioning." The results of the methods of this group are shown in Table 5. The methods do not have results on commonly used evaluation metrics. However, they have their own potential to generate the descriptions for the image.

Shetty et al. [126] employed adversarial training in their image captioning method. This method is capable of generating diverse captions. The captions are less biased with the ground-truth captions compared to the methods using maximum likelihood estimation. To take the advantage of RL, Ren et al. proposed a method that can predict all possible next words for the current word in the current time step. This mechanism helps them to generate contextually more accurate captions. Actor-critics of RL are similar to the generator and the discriminator of GAN. However, at the beginning of the training, both actor and critic do not have any knowledge about data. Zhang et al. [161] proposed an actor-critic-based image captioning method. This method is capable of predicting the ultimate captions at its early stage and can generate more accurate captions than other reinforcement-learning-based methods.

We found that the performance of a technique can vary across different metrics. Table 6 shows the methods based on the top two scores on every individual evaluation metric. For example, Lu et al. [88], Gan et al. [41], and Zhang et al. [161] are within the top two methods based on the scores achieved on the BLEU-n and METEOR metrics. The BLEU-n metrics use variable-length phrases of generated captions to match against ground-truth captions. METEOR [9] considers the precision, recall, and alignments of the matched tokens. Therefore, the generated captions by these methods have good precision and recall accuracy, as well as good similarity in word level. ROUGE-L evaluates the adequacy and fluency of generated captions, whereas CIDEr focuses on grammaticality and saliency. SPICE can analyze the semantics of the generated captions. Zhang et al. [161], Rennie et al. [120], and Lu et al. [88] can generate captions, which have adequacy, fluency, and saliency and are more grammatically correct than other methods in Table 6. Gu et al. [51] and Yao et al. [155] perform well in generating semantically correct captions.

6 DISCUSSION AND FUTURE RESEARCH DIRECTIONS

Many deep-learning-based methods have been proposed for generating automatic image captions in recent years. Supervised learning, reinforcement learning, and GAN-based methods are commonly used in generating image captions. Both visual space and multimodal space can be used in supervised-learning-based methods. The main difference between visual space and multimodal space occurs in mapping. Visual-space-based methods perform explicit mapping from images to descriptions. In contrast, multimodal-space-based methods incorporate implicit vision and language models. Supervised-learning-based methods are further categorized into encoder-decoder architecture based, compositional architecture based, attention based, semantic concept based, stylized captions, dense image captioning, and novel-object-based image captioning.

Encoder-decoder architecture-based methods use a simple CNN and a text generator for generating image captions. Attention-based image captioning methods focus on different salient parts of the image and achieve better performance than encoder-decoder architecture-based methods. Semantic-concept-based image captioning methods selectively focus on different parts of the image and can generate semantically rich captions. Dense image captioning methods can generate region-based image captions. Stylized image captions express various emotions such as romance, pride, and shame. GAN- and RL-based image captioning methods can generate diverse and multiple captions.

MS COCO, Flickr30K, and Flickr8K datasets are common and popular datasets used for image captioning. The MS COCO dataset is a very large dataset and all the images in these datasets have multiple captions. The Visual Genome dataset is mainly used for region-based image captioning. Different evaluation metrics are used for measuring the performances of image captions. The BLEU metric is good for small-sentence evaluation. ROUGE has different types and they can be used for evaluating different types of texts. METEOR can perform an evaluation on various segments of a caption. SPICE is better in understanding the semantic details of captions compared to other evaluation metrics.

Although success has been achieved in recent years, there is still room for improvement. Generation-based methods can generate novel captions for every image. However, these methods fail to detect prominent objects and attributes and their relationships to some extent in generating accurate and multiple captions. In addition to this, the accuracy of the generated captions largely depends on syntactically correct and diverse captions, which in turn rely on powerful and sophisticated language generation models. Existing methods show their performances on the datasets where images are collected from the same domain. Therefore, working on an open-domain dataset will be an interesting avenue for research in this area. Image-based factual descriptions are not enough to generate high-quality captions. External knowledge can be added in order to generate attractive image captions. Supervised learning needs a large amount of labeled data for training. Therefore, unsupervised learning and reinforcement learning will be more popular in the future in image captioning.

7 CONCLUSIONS

In this article, we have reviewed deep-learning-based image captioning methods. We have given a taxonomy of image captioning techniques, shown generic block diagrams of the major groups, and highlighted their pros and cons. We discussed different evaluation metrics and datasets with their strengths and weaknesses. A brief summary of experimental results was also given. We briefly outlined potential research directions in this area. Although deep-learning-based image captioning methods have achieved remarkable progress in recent years, a robust image captioning method that is able to generate high-quality captions for nearly all images is yet to be achieved. With the

advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time.

REFERENCES

[1] Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 115–118.

[2] Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1250–1258.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*. Springer, 382–398.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and VQA. *arXiv preprint arXiv:1707.07998* (2017).

[5] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. 2018. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5561–5570.

[6] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR'15). arXiv: https://arxiv.org/abs/1409.0473*.

[8] Shuang Bai and Shan An. 2018. A survey on automatic image caption generation. *Neurocomputing*.

[9] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Vol. 29. 65–72.

[10] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*. 1171–1179.

[11] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3 (Feb. 2003), 1137–1155.

[12] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22, 1 (1996), 39–71.

[13] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research (JAIR)* 55 (2016), 409–442.

[14] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, (Jan. 2003), 993–1022.

[15] Cristian Bodnar. 2018. Text to image synthesis using generative adversarial networks. *arXiv preprint arXiv:1805.00676*.

[16] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. ACM, 1247–1250.

[17] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*. ACM, 144–152.

[18] Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Troster. 2011. Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 4 (2011), 741–753.

[19] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. IEEE, 97–104.

[20] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *Proceedings of European Chapter of the Association for Computational Linguistics (EACL)*, Vol. 6. 249–256.

[21] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 6298–6306.

[22] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *IEEE International Conference on Computer Vision (ICCV'17)*, Vol. 2.

[23] Xinlei Chen and C. Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2422–2431.

[24] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Association for Computational Linguistics*. 103–111.

[25] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

[26] Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. 2017. Towards diverse and natural image descriptions via a conditional GAN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 2989–2998.

[27] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR'05)*, Vol. 1. IEEE, 886–893.

[28] Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, Vol. 6. 449–454.

[29] Etienne Denoual and Yves Lepage. 2005. BLEU in characters: Towards automatic MT evaluation in languages without word delimiters. In *Companion Volume to the Proceedings of the 2nd International Joint Conference on Natural Language Processing*. 81–86.

[30] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*.

[31] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2625–2634.

[32] Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1292–1302.

[33] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, and John C. Platt. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1473–1482.

[34] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*. Springer, 15–29.

[35] Alireza Fathi, Yin Li, and James M. Rehg. 2012. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*. Springer, 314–327.

[36] William Fedus, Ian Goodfellow, and Andrew M. Dai. 2018. Maskgan: Better text generation via filling in the _. *arXiv preprint arXiv:1801.07736*.

[37] Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. 2013. Learning distributions over logical forms for referring expression generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1914–1925.

[38] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*. 2121–2129.

[39] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3137–3146.

[40] Chuang Gan, Tianbao Yang, and Boqing Gong. 2016. Learning attributes equals multi-source domain generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 87–97.

[41] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 1141–1150.

[42] Jonas Gehring, Michael Auli, David Grangier, and Yann N. Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.

[43] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

[44] Ross Girshick. 2015. Fast r-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*. 1440–1448.

[45] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 580–587.

[46] Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 410–419.

[47] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*. Springer, 529–545.

[48] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[49] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. 2015. DRAW: A recurrent neural network for image generation. In *Proceedings of Machine Learning Research*. 1462–1471.

[50] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop Ontoimage*, Vol. 5, 10.

[51] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. 2017. An empirical study of language CNN for image captioning. In *Proceedings of the International Conference on Computer Vision (ICCV'17)*. 1231–1240.

[52] Yahong Han and Guang Li. 2015. Describing images with hierarchical concepts and object class localization. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 251–258.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[54] Sepp Hochreiter and JÃijrgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[55] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[56] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR'17)*. 5967–5976.

[57] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial transformer networks. In *Advances in Neural Information Processing Systems*. 2017–2025.

[58] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations (ICLR'17)*.

[59] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*. 2407–2415.

[60] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1072–1080.

[61] Junqi Jin, Kun Fu, Runpeng Cui, Fei Sha, and Changshui Zhang. 2015. Aligning where to see and what to tell: Image caption with region-based attention and scene factorization. *arXiv preprint arXiv:1506.06272*.

[62] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4565–4574.

[63] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3668–3678.

[64] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

[65] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

[66] Andrej Karpathy, Armand Joulin, and Fei Fei F. Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems*. 1889–1897.

[67] S. Karthikeyan, Vignesh Jagadeesh, Renuka Shenoy, Miguel Eckstein, and B. S. Manjunath. 2013. From where and how to what we see. In *Proceedings of the IEEE International Conference on Computer Vision*. 625–632.

[68] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*. 787–798.

[69] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*. 595–603.

[70] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. In *Workshop on Neural Information Processing Systems (NIPS'14)*.

[71] Vijay R. Konda and John N. Tsitsiklis. 2000. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*. 1008–1014.

[72] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting

language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[73] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[74] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer.

[75] Akshi Kumar and Shivali Goel. 2017. A survey of evolution of image captioning techniques. *International Journal of Hybrid Intelligent Systems* Preprint 14, 3 (2017), 123–139.

[76] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Vol. 1*. Association for Computational Linguistics, 359–368.

[77] Polina Kuznetsova, Vicente Ordonez, Tamara L. Berg, and Yejin Choi. 2014. TREETALK: Composition and compression of trees for image descriptions. *TACL* 2, 10 (2014), 351–362.

[78] RÃl'mi Lebret, Pedro O. Pinheiro, and Ronan Collobert. 2015. Simple image description generator via a linear phrase-based approach. In *Workshop on International Conference on Learning Representations (ICLR'15)*.

[79] Yann LeCun, LÃl';on Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, 11 (1998), 2278–2324.

[80] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the 15th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 220–228.

[81] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, Vol. 8.

[82] Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 605.

[83] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft Coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 740–755.

[84] Chenxi Liu, Junhua Mao, Fei Sha, and Alan L. Yuille. 2017. Attention correctness in neural image captioning. In *AAAI*. 4176–4182.

[85] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*, Vol. 3. 873–881.

[86] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[87] David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 2 (2004), 91–110.

[88] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 3242–3250.

[89] Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *International Conference on Learning Representations (ICLR'16)*.

[90] Shubo Ma and Yahong Han. 2016. Describing images by feeding LSTM with structural words. In *2016 IEEE International Conference on Multimedia and Expo (ICME'16)*. IEEE, 1–6.

[91] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations (ICLR'17)*.

[92] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11–20.

[93] Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L. Yuille. 2015. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *Proceedings of the IEEE International Conference on Computer Vision*. 2533–2541.

[94] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-RNN). In *International Conference on Learning Representations (ICLR'15)*.

[95] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

[96] Oded Maron and Tomás Lozano-Pérez. 1998. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*. 570–576.

[97] Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2016. SentiCap: Generating image descriptions with sentiments. In *AAAI*. 3574–3580.

[98] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[99] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *11th Annual Conference of the International Speech Communication Association*.

[100] Ajay K. Mishra, Yiannis Aloimonos, Loong Fah Cheong, and Ashraf Kassim. 2012. Active visual segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 4 (2012), 639–653.

[101] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 747–756.

[102] Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics, 95–104.

[103] Margaret Mitchell, Kees Van Deemter, and Ehud Reiter. 2013. Generating expressions that refer to visible objects. In *HLT-NAACL*. 1174–1184.

[104] Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th International Conference on Machine Learning*. ACM, 641–648.

[105] Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th International Conference on Machine Learning*. ACM, 641–648.

[106] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Vol. 1*. Association for Computational Linguistics, 160–167.

[107] Timo Ojala, Matti Pietikäinen, and Topi Mäenpä. 2000. Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision*. Springer, 404–420.

[108] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*. 1143–1151.

[109] Dim P. Papadopoulos, Alasdair D. F. Clarke, Frank Keller, and Vittorio Ferrari. 2014. Training object class detectors from eye tracking data. In *European Conference on Computer Vision*. Springer, 361–376.

[110] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 311–318.

[111] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 6432–6440.

[112] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. Areas of attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*. 1251–1259.

[113] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*. 2641–2649.

[114] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations (ICLR'16)*.

[115] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *Proceedings of Machine Learning Research*, Vol. 48. 1060–1069.

[116] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. 91–99.

[117] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. 2015. Multi-instance visual-semantic embedding. *arXiv preprint arXiv:1512.06963*.

[118] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. 2016. Joint image-text representation by gaussian visual-semantic embedding. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 207–211.

[119] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 1151–1159.

[120] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 1179–1195.

[121] Stephen Robertson. 2004. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation* 60, 5 (2004), 503–520.
[122] Hosnieh Sattar, Sabine Muller, Mario Fritz, and Andreas Bulling. 2015. Prediction of search targets from fixations in open-world settings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 981–990.
[123] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the 4th Workshop on Vision and Language*, Vol. 2.
[124] Karthikeyan Shanmuga Vadivel, Thuyen Ngo, Miguel Eckstein, and B. S. Manjunath. 2015. Eye tracking assisted extraction of attentionally important objects from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3241–3250.
[125] Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah. 2018. Learning-based composite metrics for improved caption evaluation. In *Proceedings of ACL 2018, Student Research Workshop*. 14–20.
[126] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *IEEE International Conference on Computer Vision (ICCV'17)*. 4155–4164.
[127] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR'15)*.
[128] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* 2 (2014), 207–218.
[129] Yusuke Sugano and Andreas Bulling. 2016. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*.
[130] Chen Sun, Chuang Gan, and Ram Nevatia. 2015. Automatic concept discovery from parallel text and visual corpora. In *Proceedings of the IEEE International Conference on Computer Vision*. 2596–2604.
[131] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*. 3104–3112.
[132] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*. 1057–1063.
[133] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.
[134] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2487–2496.
[135] Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. 2016. Rich image captioning in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 49–56.
[136] Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Natural Language Generation Conference*. Association for Computational Linguistics, 130–132.
[137] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. 2016. Conditional image generation with pixelCNN decoders. In *Advances in Neural Information Processing Systems*. 4790–4798.
[138] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.
[139] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4566–4575.
[140] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1170–1178.
[141] Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the 5th International Natural Language Generation Conference*. Association for Computational Linguistics, 59–67.
[142] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.
[143] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2017), 652–663.

[144] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional LSTMs. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 988–997.

[145] Heng Wang, Zengchang Qin, and Tao Wan. 2018. Text generation based on generative adversarial nets with latent variables. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 92–103.

[146] Minsi Wang, Li Song, Xiaokang Yang, and Chuanfei Luo. 2016. A parallel-fusion RNN-LSTM architecture for image caption generation. In *2016 IEEE International Conference on Image Processing (ICIP'16)*. IEEE, 4448–4452.

[147] Qingzhong Wang and Antoni B. Chan. 2018. CNN+ CNN: Convolutional decoders for image captioning. *arXiv preprint arXiv:1805.09019*.

[148] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W. Cottrell. 2017. Skeleton key: Image captioning by skeleton-attribute decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 7378–7387.

[149] Qi Wu, Chunhua Shen, Anton van den Hengel, Lingqiao Liu, and Anthony Dick. 2015. Image captioning with an intermediate attributes layer. *arXiv preprint arXiv:1506.01144*.

[150] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2018. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1367–1381.

[151] Zhilin Yang, Ye Yuan, Yuexin Wu, Ruslan Salakhutdinov, and William W. Cohen. 2016. Encode, review, and decode: Reviewer module for caption generation. In *30th Conference on Neural Image Processing System (NIPS'16)*.

[152] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. 2048–2057.

[153] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. 2016. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 1978–1987.

[154] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017. Incorporating copying mechanism in image captioning for learning novel objects. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. IEEE, 5263–5271.

[155] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision (ICCV'17)*. 4904–4912.

[156] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4651–4659.

[157] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu Seqgan. 2017. Sequence generative adversarial nets with policy gradient. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

[158] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg. 2013. Exploring the role of gaze behavior and object detection in scene understanding. *Frontiers in Psychology* Article 917 (2013), 1–14.

[159] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*. Springer, 818–833.

[160] Gregory J. Zelinsky. 2013. Understanding scene understanding. *Frontiers in Psychology* Article 954 (2013), 1–3.

[161] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M. Hospedales. 2017. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*.

Received April 2018; revised October 2018; accepted October 2018