IJIREEICE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

NCIET- 2020

National Conference on Innovations in Engineering & Technology
Govt., Polytechnic College, Palakkad, Kerala

Vol. 8, Special Issue 1, February 2020

A Survey on Image Captioning Techniques using Deep Learning

Hamdha Sherief[1], Bincy K[2], Arathi M[3], Surabhi C[4], Rafeeque P C[5]

Computer Science and Engineering, Government College of Engineering, Kannur, Kerala, India

Abstract: Image captioning is an emerging technology in the field of computer vision and machine learning. It has been an important research topic in the recent years as it involves understanding images and language modeling. Image captioning requires recognizing the important objects, their attributes and their interactions in an image. It also needs to generate syntactically correct and semantically accurate sentences. Traditional machine learning based methods and deep machine learning based methods can be used to achieve this. Deep learning-based techniques are better capable of handling the complexities and challenges of image captioning over traditional methods. In this survey paper, we aim to present a review of existing image captioning techniques focusing primarily on deep-learning based methods. We discuss the foundation of the techniques, their strengths, limitations and analyze their performances.

Keywords: Image captioning, novel concept, deep learning, computer vision.

I. INTRODUCTION

Image captioning is important and it is used in many fields. For example, it can be used for automatic image indexing which is relevant in Content-Based Image Retrieval (CBIR) and therefore, it can be applied to many areas such as web searching, education, the military, biomedicine, commerce, digital libraries, etc. Social media platforms such as Facebook, Twitter and Instagram are able to directly generate descriptions from images. The descriptions can tell where we are (e.g., mall, beach), what we are wearing and even what we are doing there.

Image captioning is a very popular research area of Artificial Intelligence (AI) that deals with image understanding. In image understanding objects need to be detected and recognized along with the scene type or location, object features and their interactions. Generating well-structured sentences requires syntactic and semantic grasp of the language [58]. Understanding an image mainly depends on extracting image features. The methods used for this purpose can be divided into two categories: (i) Traditional machine learning based methods and (ii) Deep machine learning based methods. In traditional machine learning, Local Binary Patterns (LBP) [42], Histogram of Oriented Gradients (HOG) [9], Scale-Invariant Feature Transform (SIFT) [35] and a combination of such features are majorly used. In deep machine learning based techniques, features are learned by the machine on its own from training data and these techniques are able to handle a large and diverse set of images and videos. Hence, they are preferred over traditional methods. In the last few years, numerous articles have been published on image captioning with deep learning. To provide an abridged version of the literature, this paper presents a survey majorly focusing on the deep learning-based papers on image captioning. Initially, we organize the existing image captioning methods into three main categories: (i) Template-based Image captioning, (ii) Retrieval-based image captioning, and (iii) Novel image caption generation. The categories are briefly discussed in Section II.

Since most of the deep learning-based image captioning methods fall into the novel caption generation category, we focus on novel caption generation with deep learning. We further classify the deep learning-based 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. We discuss all the categories in Section III. Finally, we conclude in Section IV.

II. DIFFERENT METHODS OF IMAGE CAPTIONING

In this section, we discuss the three main categories of existing image captioning methods: template-based image captioning, retrieval-based image captioning, and novel caption generation. Template-based techniques have fixed templates with blank slots to generate captions. In these systems, the different objects, actions and attributes are first identified and then the gaps in the templates are filled. For example, Farhadi et al. [11] use three different elements of a scene to fill the template slots for generating image captions. A Conditional Random Field (CRF) is leveraged by Kulkarni et al. [33] to detect the objects, attributes, and prepositions before filling in the blanks. Template-based approaches are able to generate grammatically correct captions, but since the templates are predefined, it cannot generate variable-length captions.

In retrieval-based methods, captions are retrieved from a pool of existing captions. Retrieval based methods initially find images that are visually similar images to the given image along with their captions from the training data set. These captions are called candidate captions. The captions for the given image are selected from this caption set [17], [41]. These types of systems produce general and grammatically correct captions. However, they cannot generate more descriptive and semantically correct captions.

Novel captions can be generated from visual and multimodal spaces. In these types of systems, the visual content of the image is first analyzed and then captions are generated from the visual content using a language model [29], [60], [63], [64]. These approaches can generate new, more semantically accurate captions for each image. Most novel caption generation techniques employ deep machine learning. Therefore, in this paper we focus primarily on deep learning based novel image caption generating methods.

Deep learning-based image captioning methods can also be classified based on learning techniques: Supervised learning, Reinforcement learning, and Unsupervised learning. We clump reinforcement learning and unsupervised learning into Other Deep Learning. Captions are most often generated for a whole scene in the image. However, captions can also be generated for different areas of an image such as in Dense captioning. Image captioning methods can either use simple Encoder- Decoder architecture or Compositional architecture.

III. DEEP LEARNING BASED IMAGE CAPTIONING METHODS

An overall classification for deep learning-based image captioning methods is done. We compare them by grouping them into dense captioning vs. captions for the whole scene, Encoder-Decoder architecture vs. Compositional architecture, Supervised learning vs. Other deep learning, visual space vs. multimodal space.

A. Visual Space vs. Multimodal Space

Deep learning-based image captioning methods can generate captions from both visual and multimodal spaces. Image captioning datasets have the corresponding captions stored as text. In the visual space-based methods, the image features and the corresponding captions are given as separate inputs to the language decoder. Whereas, in multimodal space-based approaches, common multimodal spaces are learned from the images and corresponding caption-text, then these multimodal representations are passed to the language decoder.

1) *Visual Space:* Most of the image captioning methods use visual space for generating captions. These methods are discussed in Section B to Section D.

2) Multimodal Space: A typical multimodal space-based system consists of a language Encoder section, a vision section, a multimodal space section, and a language decoder section. The vision section employs a deep convolutional neural network as a feature extractor to extract the image features. The language encoder obtains the word features and learns a dense feature embedding for each word. Then the semantic temporal context is forwarded to the recurring layers. The multimodal space section matches the image features with the word features in a common space. The result is passed to a language decoder which generates captions.

The methods in this category follow the steps below:

- Both image & text are jointly learned in a multimodal space using Deep neural networks & multimodal neural language model.
- The language generation part generates captions.

Kiros et al. [28] were one of the first to propose a work in this area. The proposed method employs a CNN for extracting image features and uses a multimodal space that represents both image and text together for multimodal representation learning and image caption generation. In contrast to most of the previous approaches, this method does not depend upon any additional templates, structures, or constraints. Instead it relies on the high-level image features and word representations learned from deep neural networks and multimodal neural language nets respectively. The neural language models cannot efficiently handle a large amount of data and therefore works poorly with long term memory [24].

Kiros et al. [28] extended their work in [29] where LSTM is employed for sentence encoding and introduced a novel neural language model called the structure-content neural language model (SC-NLM) for image captions generations, to learn a joint image sentence embedding. The SC-NLM has one advantage over existing techniques i.e., it can extract the structure of the sentence to its content produced by the encoder. It also generates more realistic image captions than the approach proposed by [28].

Mao et al. [38] put forth a multimodal Recurrent Neural Network (m-RNN) method for generating novel image captions. This method has two sub-networks: a deep recurrent neural network (RNN) for sentences and a deep convolutional network (CNN) for images. These two sub-networks exchange information with one other in a multimodal layer to form the whole m-RNN model. Both image and parts of sentences are given as input in this method. It calculates the probability distribution to generate the next word. There are five more layers in this model: Two-word embedding layers, a multimodal layer, a recurrent layer and a SoftMax layer. This multimodal recurrent neural network method is similar to the method of Kiros et al. [28]. Kiros et al. use a fixed length context, but in this method, the temporal context is stored in a recurrent architecture that allows an arbitrary context length. Only one hot vector is used by the two word embedding layers to generate a dense word representation. It encodes both the syntax and semantics of the words.

The semantically relevant words is found by calculating the Euclidean distance between two dense word vectors in embedding layers. Unlike most of the sentence-image multimodal methods which [13], [27], [29], [51] use pre-determined word embedding vectors to initialize their model, this method arbitrarily initializes word embedding layers and learn them from the training set. This enables them to generate better image captions than previous methods. Nowadays, many image captioning techniques [39], [28], [39] are built on recurrent neural networks. They make use of a recurrent layer for storing visual information. However, (m-RNN) leverage both image representations and sentence bits to generate captions. It achieves a better performance with the help of a relatively small dimensional recurrent layer.

Shuang Ma et al. [36] introduced a multimodal image captioning technique in which they tackled the problem of translating instances from one modality to another without the help of paired data by leveraging an intermediate modality that was common to the two other modalities. In the paper, Shuang Ma et al. [36] chose to translate images to speech and leveraged disjoint datasets with one shared modality, i.e., image-text pairs and text-speech pairs, with text as the shared modality. Since the shared modality is skipped during the generation process, they called this problem" skip-modal generation". Shuang et al. [36] proposed a multimodal information bottleneck approach to tackle the problem. The model showed qualitative results on image-to-speech synthesis and also improves performance on traditional cross-modal generation.

B. Supervised Learning vs. Other Deep Learning

In supervised learning, training data is labelled with desired output. Unsupervised learning, on the other hand, handles unlabeled data. A type of unsupervised learning technique uses Generative Adversarial Networks (GANs) [18]. Reinforcement learning is another type where the aim of a model is to discover data and/or labels through exploration and a reward signal. Several image captioning methods employ reinforcement learning and GAN based approaches. These methods belong to “Other Deep Learning” category.

1) *Supervised Learning-Based Image Captioning:* Supervised learning-based networks have been used for many years in image classification [19], [32], [49], [54], object detection [15], [16], [44], and attribute learning [14]. This paper identifies a large number of supervised learning-based image captioning methods. They are classified into different categories: (i)Encoder-Decoder Architecture, (ii) Compositional Architecture, (iii) Dense image captioning.

2) Other Deep Learning-Based Image Captioning: It is often difficult to accurately label data. Therefore, recent focus has been more on reinforcement learning and unsupervised learning-based techniques for image captioning. A reinforcement learning model selects an action, receives a reward signal, and proceeds to a new state. The model endeavors to select the action with the maximum long-term reward. It requires continuous state and action information, to provide the expectations of a value function. State-of-the-art reinforcement learning techniques face a lot of limitations such as the lack of guarantees of a value function and non-specific state-action information.

Policy gradient methods [53] are a type of reinforcement learning that can select a specific policy for a specific action with the help of gradient descent and optimization techniques. Policy gradient methods require fewer parameters than value-function based approaches. Existing deep learning-based image captioning methods use different image encoders to extract image features. The features are then given as input to the neural network-based language decoders to generate captions. The methods have two main disadvantages:

- They are trained using back- propagation [43] approaches and maximum likelihood estimation. In this method, the next word is predicted using the given image and all the previously generated ground-truth words. Therefore, the generated captions resemble ground-truth captions. This is called exposure bias [3] problem.
- Evaluation metrics at test time are non-differentiable.

Ideally sequence models should be trained to prevent exposure bias and directly optimise evaluation metrics. In actor-critic-based reinforcement learning algorithm, critic can be employed in calculating the expected future reward to train the actor. Reinforcement learning (RI) based image captioning methods sample the next token from the model depending upon the rewards they receive in each state. Policy gradient methods in RI can optimize the gradient in order to predict the amassed long-term reward value. Therefore, it is able to solve the non-differentiable issue of evaluation metrics.

The methods in this category follow the following steps:

- A combined CNN and RNN based network generate captions.
- A second CNN-RNN based network evaluates the captions and sends feedback to the first network in order to generate high quality captions.

Ren et al. 2017 [47] proposed a novel reinforcement learning based image captioning method. The architecture of this method consists of two networks that predict the next best word at each time step, together. The job of the policy network is to act as a local supervisor and assist in predicting the next word depending upon the current state. The value network acts as a global supervisor and estimates the reward value by taking all the possible extensions of the current state into account. This mechanism can adjust the networks in predicting the correct words and can therefore, generate good captions that are like the ground truth captions at the end. It also uses an actor-critic reinforcement learning model [30] to train the entire network. Visual semantic embedding [45], [46] is employed to compute the actual reward value in predicting the correct word and also to measure the likeness between images and sentences to evaluate the correctness of generated captions.

Most of the existing RL-based image captioning methods rely primarily on a single policy network and reward function which is a limped approach to the multi-level (word and sentence) and multi-modal (vision and language) nature of the task. To solve this problem, Xu et al. [61] proposed a novel multi-level policy and reward RL framework for image captioning that can be easily integrated with RNN-based captioning models, language metrics, or visual-semantic functions for optimization. The multi-level policy network jointly updates the word and sentence-level policies for word generation. The multi-level reward function leverages both a vision-language reward and a language-language reward together to guide the policy. Furthermore, Xu et al.[61] proposed a guidance term to bridge the policy and the reward for RL optimization. Experiments were conducted on the MSCOCO and Flickr30k datasets and the analyses show that the proposed framework achieves significant performances on a variety of evaluation metrics. In addition, ablation studies were performed on multiple variants of the proposed framework and several representative image captioning models and metrics for the word-level policy network and the language- language reward function were explored to measure the generalization ability of the proposed framework.

GAN based methods are able to learn deep features from unlabeled data. They achieve these representations with the help of a pair of networks: The Generator and the Discriminator. GANs have already been employed successfully in a number of applications such as image captioning [4], image to image translation [21], text to image synthesis [5], and text generation [12], [59]. There are two main issues with GAN. (i) GAN can efficiently generate natural images from real images as GANs are proposed for real-valued data, but since, text processing is depending upon discrete numbers, such operations are non-differentiable, thus making it difficult to apply back-propagation directly. Policy gradients make use of a parametric function to enable gradients to be back-propagated. (ii) The evaluator faces difficulties with vanishing gradients and error propagation during sequence generation. It is in need of a probable future reward value for every partial description. Monte Carlo rollouts [65] is used to calculate this future reward value.
GAN based image captioning methods holds an advantage over conventional deep convolutional network and deep recurrent network-based models in that it is able to generate a diverse set of image captions. Shetty et al. [48] proposed a new GAN based image captioning method. This method can generate multiple captions for a single image and can generate more diverse captions compared to previous methods. GANs faces limitations in back- propagating discrete data and so Gumbel sampler [23], [37] is used to overcome the problem. The two main parts of this network are the generator and the discriminator. During training, the generator learns the loss value from the discriminator instead of learning it from explicit sources. The discriminator has true data distribution and can differentiate between generator-generated samples and true data samples. This helps the network to learn diverse data distribution. Moreover, the network can classify the generated caption sets as either real or fake. Thus, it can generate captions similar to human generated one.

C. Dense Captioning vs. Captions for the whole scene

In dense captioning, each region of the scene is captioned. In other methods, captions are generated for the whole scene.

1) Dense Captioning:

Previous image captioning methods are able to generate only one caption for the entire image. They make use of different regions of the image to extract information of various objects. However, these methods do not generate captions region wise. Johnson et al. [25] introduced an image captioning technique called DenseCap in which all the salient regions of an image is localized and then descriptions generated for those regions. A usual method of this category has the following steps: (i) Region suggestions are generated for the various regions of the given image. (ii) CNN is leveraged to obtain region-based image features. (iii) The outputs of Step 2 are made use of by a language model to generate captions for every region.

In Dense captioning [25], there is a fully convolutional localization network architecture, which constitutes a convolutional network, a dense localization layer, and an LSTM [20] language model. The dense localization layer takes an image and processes it with a single, efficient forward pass, which implicitly predicts a collection of regions of interest in the image. Therefore, it requires no external region proposals unlike Fast R-CNN or a full network of Faster R-CNN. The working principle of the localization layer is similar to the work of Faster R-CNN [44].

However, Johnson et al. [25] use a differential, spatial soft attention method [26], [22] and bilinear interpolation [22] instead of ROI pooling mechanism [15]. This modification enables the method to backpropagate through the network to smoothly select the active regions. It employs the Visual Genome [31] dataset for the experiments to generate image captions for different regions. Region-based descriptions are more detailed and objective when compared to global image descriptions. Region-based description is known as dense captioning.

There are a few challenges to dense captioning. As regions are dense, one object could have multiple overlapping regions of interest. Moreover, it is quite difficult to recognize each target region for every visual concept. Yang et al. [62] proposed another dense captioning method that can tackle these obstacles. First, it addresses an inference mechanism that depends on the visual features of the region as well as the predicted captions for that region. This lets the model find one appropriate position for the bounding box. Second, in order to to provide a rich semantic description, they apply a context fusion that can join context features with the visual features of the respective region.

2) Captions for the whole scene:

The following methods all generate single or multiple captions for the whole scene: Encoder-Decoder architecture and Compositional architecture.

D. Encoder-Decoder Architecture vs Compositional Architecture

Some methods use simple standard encoders and decoders to generate captions. However, other methods make use of multiple networks for the same.

1) Encoder-Decoder Architecture-Based Image captioning:

Neural network-based image captioning methods work in a simple end to end manner. These methods are largely similar to the encoder-decoder framework-based neural machine translation [52]. In this net, global image features are extricated from the hidden layers of CNN and then fed into an LSTM to generate a sequence of words.

A classic method of this category has the following steps: (1) A typical CNN is used to obtain the scene type, to detect the objects and their relationships. (2) The output of Step 1 is leveraged by a language model to convert them into words, combined phrases that produce image captions.

Vinyals et al. [57] introduced a method called Neural Image Caption Generator (NIC). This method takes advantage of a CNN for image representations and an LSTM for generating image captions. This special CNN employs a new method for batch normalization and the output of the last hidden activations of CNN is given as an input to the LSTM decoder. This LSTM is able to keep track of the objects that have previously been described using text. NIC is trained depending upon maximum likelihood estimation. Image information is included in the initial state of an LSTM. The next words are generated according to the current time step and the previous hidden state. This process continues till it reaches the end fragment of the sentence. It may face vanishing gradient problems since image data is fed only at the start of the process. The role of the words generated at the beginning also becomes less significant. Therefore, LSTM is still facing issues in generating long length sentences [2], [8].

Previous CNN-RNN based image captioning methods employ unidirectional and relatively shallow LSTMs. In unidirectional techniques, the next word is predicted based on both visual context and all the previous textual contexts and it is unable to generate contextually well formed captions. Also, recent object detection and classification methods [32], [50] prove that deep, hierarchical techniques are better at learning.

Wang et al. [7] proposed a deep bidirectional LSTM- based method capable of generating contextually and semantically rich image captions. The proposed architecture includes a CNN and two separate LSTM networks. It utilizes both past and future context information to learn long term visual-language interactions.

2) Compositional Architecture-Based Image Captioning:

Compositional architecture-based methods are composed of several independent functional building blocks: First, a CNN is employed to extract semantic concepts from the image. Then a language model is used to generate a group of candidate captions. Then these candidate captions are re-ranked using a deep multimodal similarity model to generate the final caption. A typical method of this class follows the following steps: (i) Image features are extracted using a CNN. (ii) Visual concepts (e.g. attributes) are extricated from visual features. (iii) Multiple captions are generated by a language model using the information from Step i and Step ii. (iv) The generated captions are re-ranked with the help of a deep multimodal similarity model which selects high quality image captions.

Fang et al. [10] introduced generation-based image captioning. To train the model on an image captioning dataset, it uses visual detectors, a language model, and a multimodal similarity model. Convolutional neural networks, AlexNet [32] and VGG16Net, are used for extracting features of the sub-regions of an image. These features are mapped to the words that are likely to be contained in the image captions.

Multiple instance learning (MIL) [40] is employed to train to learn discriminative visual attributes of each word. A maximum entropy (ME) [1] language model is used to generate image captions from these words. Generated captions are then ranked with the help of linear weighting of sentence features. Minimum Error rate training (MERT) [106] is used in order to learn these weights.

A Deep Multimodal Similarity Model (DMSM)is employed to map image and sentence fragments with their common vector representations. It chooses significantly high-quality image captions. Most methods use training and testing samples from the same domain. Therefore, there is no guarantee that these methods can perform well in open-domain images. Moreover, they are only good at identifying generic visual content. They are unable to recognize entities such as celebrities and landmarks. The generated captions of these methods are evaluated on automatic evaluation metrics. These methods have already shown good results on these metrics. However, if it is considered real life entity information, the performance could be weaker. Tran et al. [56] introduced a different image captioning method which is capable of generating image captions for open domain images as well. It can note a diverse set of visual concepts and generate captions for celebrities and landmarks. It leverages an external knowledge base Freebase [34] to recognize a broad range of entities. Human judgments are used to evaluate the performance. In experiments, it uses three datasets: MS COCO, Adobe-MIT FiveK [6], and images from Instagram; where only images of MS COCO dataset were collected from the same domain. The method achieves significant performances, most notably on the Instagram dataset.

The main issue with sequential models is that they usually result in overgeneralized expressions that lack details that may be present in an input image. To overcome this problem Tian et al. [55] introduced a hierarchical framework for image captioning that explores both compositionality and sequentiality of natural language by selectively attending to different modules corresponding to unique attributes of each object detected in an input image in order to include specific descriptions such as counts and color. Experiments carried out on the MSCOCO dataset showed that the proposed model achieves significant improvement in performance over state-of-the art models across multiple evaluation metrics, more importantly, presenting visually interpretable results.

IV. CONCLUSION

In this paper, we have discussed different deep learning- based image captioning methods and reviewed their pros and cons. Although deep learning-based image captioning methods has achieved remarkable success in recent years, there is still a huge room for improvement. While generation-based methods are able to generate novel captions for every image, these methods fail to detect relevant objects and attributes and their interactions to a certain extent. Also, the accuracy of the generated captions mainly rely on syntactically correct and unique captions which in turn depend upon efficient and powerful language generation models. Factual descriptions based on images alone are not enough to generate high quality captions. External knowledge can be used in order to generate more pleasing captions. Supervised learning requires a huge amount of labelled data for training and therefore, unsupervised learning and reinforcement learning will gain more relevance in the future in image captioning.

ACKNOWLEDGMENT

We give all honour and praise to the GOD who gave us wisdom and blessings during this survey. First and foremost we would like to express our sincere thanks to our guide and the H.o.D, **Dr. Rafeeque P. C.** for providing his time, inspiration, encouragement, help, valuable guidance, constructive criticism and constant interest. He took personal interest in spite of numerous commitments and busy schedule to help us. We would like to extend our sincere gratitude to the Principal, **Dr. Rejoin V. O.** for rendering us all facilities throughout the survey. Lastly, we would like to thank all friends for providing helping hands in times of need.

REFERENCES

[1]. Vincent J Della Pietra, Adam L Berger, and Stephen A Della Pietra. *A maximum entropy approach to natural language processing.* Computational linguistics, 22(1):39–71, 1996.
[2]. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural machine translation by jointly learning to align and translate.* International Conference on Learning Representations (ICLR), pages 3156–3164, 2015.
[3]. Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. *Scheduled sampling for sequence prediction with recurrent neural networks.* Advances in Neural Information Processing Systems, pages 1171–1179, 2015.
[4]. Raquel Urtasun Bo Dai, Dahua Lin and Sanja Fidler. *Towards diverse and natural image descriptions via a conditional gan.* Proceeding of the IEEE conference on computer vision and pattern recognition (CVPR), pages 2989–2998, 2017.
[5]. Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frdo Durand. *Learning photographic global tonal adjustment with a database of input/output image pairs.* Computer Vision and Pattern Recognition (CVPR),IEEE Conference, pages 97–104, 2011.
[6]. Christian Bartz Cheng Wang, Haojin Yang and Christoph Meinel. *Image captioning with deep bidirectional lstms.* Proceedings of the 2016 ACM on Multimedia Conference. ACM, pages 988–997, 2016.
[7]. Kyunghyun Cho, Bart Van Merrinboer, Dzmitry Bahdanau, and Yoshua Bengio. *On the properties of neural machine translation: Encoder-decoder approaches.* Association for Computational Lin- guistics, pages 103–111, 2014.
[8]. Navneet Dalal and Bill Triggs. *Histograms of oriented gradients for human detection.* IEEE Computer Vision and Pattern Recognition, 1:886–893, 2005.
[9]. Hao Fang, Saurabh Gupta, Rupesh K Srivastava Forrest Iandola, Li Deng, Piotr Dollr, Xiaodong He Jianfeng Gao, Margaret Mitchell, and John C Platt. *From captions to visual concepts and back.* Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1473–1482, 2015.
[10]. Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. *Every picture tells a story: Generating sentences from images*. European conference on computer vision,Springer, pages 15–29, 2010.
[11]. William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: *Better text generation.* arXiv preprint arXiv:1801.07736, 47, 2018.
[12]. Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, and Tomas Mikolov. *Devise: A deep visual-semantic embed- ding model.* Advances in neural information processing systems,
pages 2121–2129, 2013.

[13]. Chuang Gan, Tianbao Yang, and Boqing Gong. *Learning attributes equals multi-source domain generalization*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 87– 97, 2016.
[14]. Ross Girshick. *Fast r-cnn*. Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015.
[15]. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation*. Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.
[17]. Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. *Improving image-sentence embeddings us- ing large weakly annotated photo collections*. European Conference on Computer Vision. Springer, pages 529–545, 2014.
[18]. Ian Goodfellow, Jean Pouget-Abadie, Bing Xu Mehdi Mirza, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. *Generative adversarial nets*. Advances in neural information processing systems, pages 2672–2680, 2014.
[19]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep residual learning for image recognition*. Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770– 778, 2016..
[21]. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. *Image-to-image translation with conditional adversarial networks*. Proceedings of the IEEE International Conference on Computer Vision (CVPR), pages 5967–5976, 2017.
[22]. Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. *Spatial transformer networks*. Advances in Neural Information Processing Systems, pages 2017–2025, 2015.
[23]. Eric Jang, Shixiang Gu, and Ben Poole. *Categorical reparameteri- zation with gumbel-softmax*. International Conference on Learning Representations (ICLR), 2017.
[24]. Rafal, ,MikeSchuster,NoamShazeer,and Yonghui Wu. *Exploring the limits of language modeling*. arXiv preprint arXiv:1602.02410, 2016.
[25]. Andrej Karpathy Justin Johnson and Li Fei-Fei. Densecap: *Fully convolutional localization networks for dense captioning*. Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4565–4574, 2016.
[26]. Alex Graves Danilo Jimenez Rezende Karol Gregor, Ivo Danihelka and Daan Wierstra. *Draw: A recurrent neural network for image generation*. Proceedings of Machine Learning Research, 9(8):1462– 1471, 2015.
[27]. Andrej Karpathy, Armand Joulin, and Fei Fei F Li. *Deep fragment embeddings for bidirectional image sentence mapping*. Advances in neural information processing systems, pages 1889–1897, 2014.
[28]. Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. *Multimodal neural language models*. Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 595–603, 2014.
[29]. Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. *Unifying visual-semantic embeddings with multimodal neural language models*. Workshop on Neural Information Processing Systems (NIPS), 2014.
[30]. Vijay R Konda and John N Tsitsiklis. *Actor-critic algorithms*. Advances in neural information processing systems, pages 1008– 1014, 2000.
[31]. Ranjay Krishna ,Yuke Zhu , Oliver Groth , Justin Johnson ,Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, and et al David A Shamma. *Visual genome: Connecting language and vision using crowdsourced dense image annotations*. International Journal of Computer Vision, 123(1):32–73, 2017.
[32]. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. *Imagenet classification with deep convolutional neural networks*. Advances in neural information processing systems, pages 1097–1105, 2012.
[33]. Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: *Understanding and generating image descriptions*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35:2891–2903, June 2013.
[34]. Praveen Paritosh Tim Sturge Kurt Bollacker, Colin Evans and Jamie Taylor. *Freebase: a collaboratively created graph database for structuring human knowledge*. Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, pages 1247–1250, 2008.
[35]. David G Lowe. *Distinctive image features from scale-invariant keypoints*. International journal of computer vision, 60(2):91110, 2004.
[36]. Shuang Ma, Daniel McDuff, and Yale Song. *Unpaired image-to- speech synthesis with multimodal information bottleneck*. The IEEE International Conference on Computer Vision (ICCV), pages 7598– 7607, 2019.
[37]. Chris J Maddison, Andriy Mnih, and Yee Whye Teh. *The concrete distribution: A continuous relaxation of discrete random variables*. International Conference on Learning Representations (ICLR), 2017.
[38]. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. *Deep captioning with multimodal recurrent neural networks (m-rnn)*. International Conference on Learning Repre- sentations (ICLR), 2015.
[39]. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. *Explain images with multimodal recurrent neural networks*. arXiv preprint arXiv:1410.1090, 2014.
[40]. Oded Maron and Toms Lozano-Prez. *A framework for multiple- instance learning*. Advances in neural information processing systems, pages 570–576, 1998.
[41]. Peter Young Micah Hodosh and Julia Hockenmaier. *Framing image description as a ranking task: Data, models and evaluation metrics*. Journal of Artificial Intelligence Research, 47:853–899, 2013.
[42]. Timo Ojala, Matti Pietikinen, and Topi Menp. *Gray scale and rotation invariant texture classification with local binary patterns*. European Conference on Computer Vision. Springer, pages 404– 420, 2000.
[43]. MarcAurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. *Sequence level training with recurrent neural networks*. International Conference on learning Representations (ICLR), 2016.
[44]. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. *Faster r-cnn: Towards real-time object detection with region proposal networks*. Advances in neural information processing systems, pages 91–99, 2015.
[45]. Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. *Multi-instance visual-semantic embedding*. arXiv preprint arXiv:1512.06963(2015), 2015.
[46]. Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. *Joint image-text representation by gaussian visual-semantic embedding*. Proceedings of the 2016 ACM on Multimedia Conference. ACM, pages 207–211, 2016.
[47]. Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. *Deep reinforcement learning-based image captioning with embed- ding reward*. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pages 1151–1159, 2017.
[48]. Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. *Speaking the same language: Matching machine to human captions by adversarial training*. IEEE International Conference on Computer Vision (ICCV), pages 4155–4164., 2016.
[49]. Karen Simonyan and Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition*. International Conference on Learning Representations (ICLR), 2015.
[50]. Karen Simonyan and Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition*. International Conference on Learning Representations (ICLR), 2015.
[51]. Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, , and Andrew Y Ng. *Grounded compositional semantics for finding and describing images with sentences*. Transactions of the Association for Computational Linguistics 2, pages 207–218, 2014.
[52]. Ilya Sutskever, Oriol Vinyals, , and Quoc V Le. *Sequence to sequence learning with neural networks*. Advances in neural information processing systems, pages 3104–3112, 2014.

[53]. Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. *Policy gradient methods for reinforcement learning with function approximation*. Advances in neural information processing systems, pages 1057–1063, 2000.
[54]. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. *Going deeper with convolutions.* Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.
[55]. Junjiao Tian and Jean Oh. *Image captioning with compositional neural module networks*. Proceedings of the Twenty-Eighth Interna- tional Joint Conference on Artificial Intelligence (IJCAI-19), pages 3576–3584, 2019.
[56]. Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Cara- pcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. *Rich image captioning in the wild.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 49– 56, 2016.
[57]. Oriol Vinyals ,Alexander Toshev, ,Samy Bengio ,and Dimitru Erhan. *Show and tell: A neural image caption generator.* Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3156–3164, 2015.
[58]. Oriol Vinyals ,Alexander Toshev, ,Samy Bengio ,and Dimitru Erhan. *Show and tell: Lessons learned from the 2015 mscoco image captioning challenge.* IEEE transactions on pattern analysis and machine intelligence, 39(4):652–663, 2017.
[59]. Heng Wang, Zengchang Qin, and Tao Wan. *Text generation based on generative adversarial nets with latent variables.* Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pages 92–103, 2018.
[60]. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. *Show, attend and tell: Neural image caption generation with visual attention*. International Conference on Machine Learning, 2048- 2057, 2015.
[61]. N. Xu, H. Zhang, A. Liu, W. Nie, Y. Su, J. Nie, and Y. Zhang. *Multi-level policy and reward-based deep reinforcement learning framework for image captioning*. IEEE Transactions on Multimedia, pages 1–1, October 2019.
[62]. Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. *Dense captioning with joint inference and visual context.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1978–1987, 2016.
[63]. Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. *Boosting image captioning with attributes*. IEEE International Conference on Computer Vision (ICCV), pages 4904–4912, 2017.
[64]. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. *Image captioning with semantic attention.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4651–4659, 2016.
[65]. Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu Seqgan. *Sequence generative adversarial nets with policy gradien*. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017.