e-ISJN: A4372-3114
p-ISJN: A4372-3115

ISSN: 2321-7782 (Online)
ISSN: 2347-1778 (Print)
Impact Factor: 7.327

Volume 8, Issue 2, February 2020

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Deep Learning Based Image Captioning: The State of The Art

Yugandhara A. Thakare[1]
Asst. Professor,
Computer Science & Engineering Department
SIPNA COET,
Amravati, MS, India

Dr. K. H. Walse[2]
Professor & Head,
Computer Science & Engineering Department
Anuradha Engineering College,
Chikhli, MS, India

Dr. V. M. Thakare[3]
Professor & Head,
P.G. Department of CSE, SGBAU,
Amravati, MS, India

Abstract: ***In Artificial Intelligence, Image captioning is a well-known research area that deals with understanding and generation of a language description for that image. Detection and recognition of objects is of necessity in Image understanding. It also needs to interpret scene type, object properties and their interactions. Generation of attractive sentences requires a sentence to be both syntactically and semantically proper. A first look is enough for human being to recognize and explain what the image is all about. Automatically producing textual explanation from an artificial system is the job of image captioning. The job is straightforward – a single sentence should be generated as a output which describes what is in fact presented in the image – the things existing, the activities being performed, the correlation amongst the things and their properties etc.***

Keywords: Image Captioning, RNN, CNN, LSTM.

I. INTRODUCTION

Image captioning is nothing but generating textual explanation of an image. The job Image captioning is to identify the key objects, objects relationships and attributes in an image. The produced sentence should be semantically and syntactically correct. Techniques that are based on Deep learning are able to handle challenges and complexity of image captioning. The job of automatic image explanation comprises of taking an image, examining its visual content and producing a textual description which is usually a sentence. Generation of textual description requires the use of Computer Vision and Natural Language Processing (NLP) techniques too. From a Computer Vision point of sight, the explanation could cover any visual side of the image: it can talk about things/objects, objects attributes and features of the scene like outdoor or indoor, or express the interaction of the persons and things in the scene. Additional interesting, it can reference the things that are not displayed and provide background information that cannot be resulted straight from the image. In other words, image understanding is essential, but undoubtedly not enough for generating a good description. A good description should be comprehensive but concise, while it consists of grammatically well-formed sentences. In this survey, we follow [1] and assume that for this study the descriptions that are of interest are the ones which express visual and conceptual facts in an image.

There are many reasons which stated the importance of Image captioning. There are different applications of Image captioning. For example it can be used in video surveillance, Google image Retrieval, automatic image indexing. In many areas Image indexing can be applied, including the military field, education field, bio medicine, commerce field, and in web searching. A public network platform like Twitter along with Facebook is able to produce information from images. The

descriptions can cover at what place we are, what clothing we are wearing and the most important part is what activity we performed there.

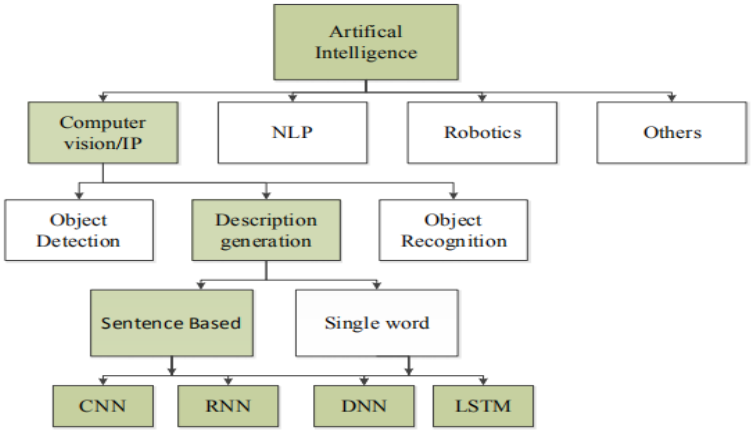

Fig.1. Description generation Taxonomy [28]

Image captioning is a well-known fact finding of Artificial Intelligence that concern about understanding image means to detect and recognize objects and generating explanation for the given input image. In image captioning it is essential to interpret background area of image, object characteristics and their coordination. A proper sentence can be generated by using both semantic and syntactic understanding of the language [2].This paper presents a brief survey of methods for description-generation of images and some technical aspects. An extensive relevant literature has appeared over the last five years. The aim behind this review is to provide an across-the-board outline of this literature, covering state-of-the-art models, evaluation metrics and datasets that have been developed or adopted in this research area.

Current image captioning methods are categories in following types: (1) Image captioning using Template (2) Image captioning by Retrieval method (3) Image captioning by Novel-based method. Novel caption generation category uses the deep learning for image caption generation. Here we discussed the review of novel caption generation methods which uses the deep learning.

II. Methodologies in Deep Learning for Image Caption Generation

In recent years, several methodology's have been implemented that produce automatic textual description of image. Such method includes supervised learning, GAN based methods and reinforcement learning is usually used to generate image captions. In supervised learning-based methods both visual and multimodal space are widely used. The dissimilarity among visual and multimodal space occurs in mapping. Main task of visual space-based methods is to achieve explicit mapping from images to explanations. On other side, the method which uses multimodal space integrates language models and implicit vision.

Methods based on supervised learning are classified as Encoder Decoder based, Compositional based, Attention-based, Semantic based, Stylized, Dense image captioning, and Novel object image captioning.

A. Encoder Decoder architecture

Text generator and CNN are used for generating image captions in Encoder Decoder architecture.

This method has the following steps:

(1)To identify the objects and their relationship and to gain the scene type vanilla CNN is used.

(2)The output of Step 1 is used as input to language model that uses the output from step1 and translate them into words, united phrases that generate an image captions.

The method proposed by [21] called Neural Image Caption Generator (NIC). For image illustrations this method makes use of LSTM and CNN for creating image captions. Convolutional neural network uses a unique technique for normalization of batch and for LSTM decoder the input is given as the result of last layer of CNN. The job LSTM is to keep track of the objects which are previously described with the help of text. Fig. 2 shows the image captioning methods which uses encoder decoder architecture. This outline is initially aimed to convert one language sentence into another. Image captioning is articulated as a translation issue, where input and output is an image and sentence respectively [28]. Under this framework, a neural network encoder firstly encodes an image to an intermediate representation, and it then inputted to recurrent neural network decoder and at last generates a textual sentence or description word by word.

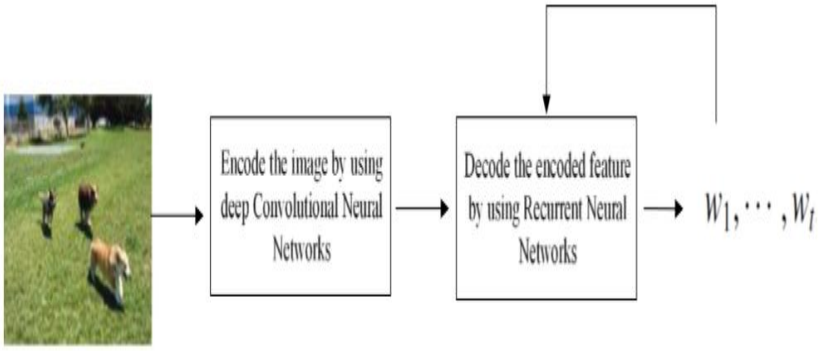

Fig. 2 General structure of encoder–decoder based image captioning methods.

B. Compositional architecture

Compositional architecture-based methods consist of numerous independent functional blocks: First, semantic concepts are extracted from the image by using CNN. A set of candidate captions is generated using Language model. Final caption is generated by re-ranking candidate captions using a deep multimodal similarity model.

This method has the following general steps:

(1) CNN is used to gain Image features.

(2) Visual features are used to generate visual attributes

(3) Language model generates multiple captions using the

(4) Superior quality image captions can be generated by using the concept of re-ranking which uses deep multimodal similarity model.

[29] Introduced generation-based image captioning to train the model on dataset of image captioning. Visual indicators, a language generator, and multimodal model are used. Image captions may include verbs, nouns, adjectives. A dictionary made up of thousand mutual words out of the captions from the images given for the training. Rather than the complete image, the system functions with the image sub-regions. For extracting features of sub-regions of an image CNN have been used. Those features which are extracted from sub-regions are depicted with dictionary words that probable to be enclosed in the image captions.

C. Attention-based

Compare to the performance of encoder-decoder based architecture method; Attention-based image captioning methods gives the attention on various salient parts of the image. For machine translation, neural encoder-decoder based methods were primarily used [30]. By considering these fashions, they have been used for image captioning effectively. To extract the visual features from the given image CNN is used as an encoder and as decoder RNN is used, which convert word-by-word in natural language explanation of the image. Moreover, these methods generate only the captions in the view of scene as a whole of

image and fail to consider spatial features of the image which are the essential portions of the image captions. Rather, they produce captions taking in to account the scene as a whole. In deep learning Attention based mechanisms are becoming increasingly popular as they are able to report these limitations. This mechanism dynamically emphases on the several portions of the image which is given as an input and finally output as sequence of sentences are being generated.

This method has the following steps:

(1) CNN is used to obtain Image facts based on the entire scene.

(2) Language generation phase generates words or phrase in language generation phase by taking the output of CNN.

(3) Depending on produced words or phrases, salient area of given image is concentrated in every phase of language generator.

(4) Till the end of language generation model, dynamically captions are updated.

A structural view of the classic attention-based image captioning method is shown in Figure 3.

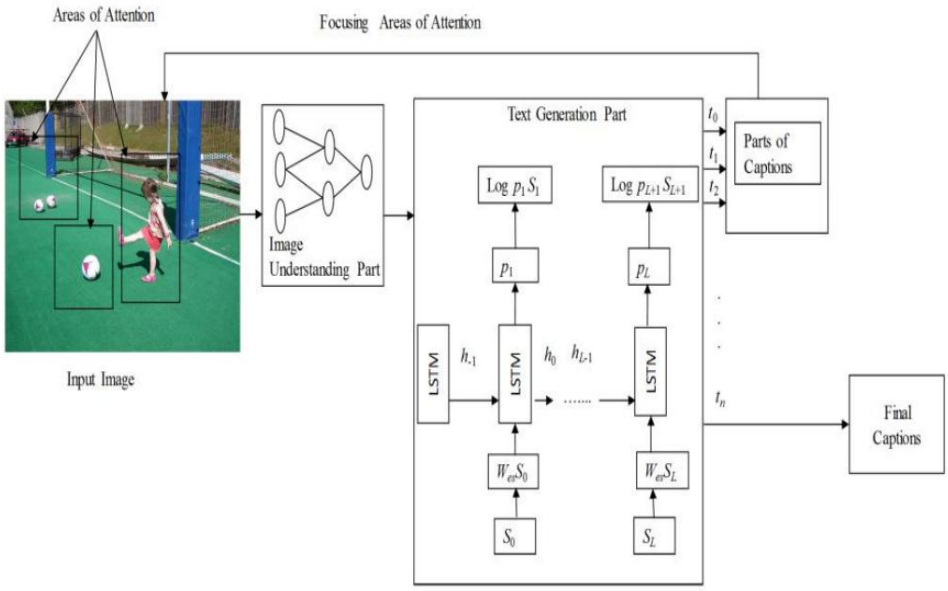

Fig. 3 A structural view of the classic attention-based image captioning method. [24]

D. Semantic concept-based

Image captioning method that are based on semantic concept mainly emphasis on various area of an image and can be able to produce semantically correct captions. Set of semantic concept proposals are obtained from the image to which this method selectively attend. These concepts are then combined into states which are hidden and the outputs of RNN.

This method has the below steps:

(1) Image features and semantic concepts are encrypted by using CNN encoder.

(2) Image features are given to the input of language generation model.

(3) Semantic concepts are further added to the different out of sight states of the language model.

(4) Caption is generated with semantic meaning by using language generation portion.

Karpathy et al. extended their method [23] in [22]. This advanced method is capable of generating caption for image and for their region too. This method make use of a new grouping of CNN on to the image areas, bidirectional RNN on to sentences, and a common multimodal embedding that associates the two modalities. It also presents multimodal recurrent neural network

framework. As compare to their earlier method, bidirectional neural network is used by this method to acquire word illustrations in the sentence.

E. Stylized captions

Stylized image captions express different emotions like love, pride, and embarrassment. Present image captioning systems are capable of generating captions based on only the image content. In that method, stylized part of the text is not considered separately from other linguistic patterns. Though, the captions with style are more informative and smart compare to only plane explanation of an image.

This method has the below steps:

(1) Image details are obtained by CNN image encoder.

(2) Different stylized thoughts such as loving, funny are extracted from training data for which a separate text corpus is arranged.

(3) By using the information from above two steps the language generation model can produce stylized and attractive captions.

StyleNet technique for novel image captioning system is proposed by [4]. This technique is capable of generation eye-catching caption adding several styles. The structural design of this technique comprises of a factorized LSTM and Convolutional neural network which is able to isolate factual and stylish elements from captions. Style factors are being identified by multitask sequence to sequence training [21] which is used to identify the style factors and then these factors are added at run time for generating attractive captions. More interesting is that, instead of paired images for training purpose it makes the use of an external monolingual stylish language corpus. FlickrStyle10k is a dataset that consist of stylish image captions which is used to generate captions with different styles.

F. Dense image captioning

Region based image captions are generated by using Dense image captioning method. Only one caption can be generated by the previous image captioning methods by considering image as a whole. Different area of the image is used to acquire facts of various objects. Though, this method doesn't produce area wise captions. [27] Come up with a method based on image captioning called DenseCap. By identifying the salient region from image, it generated descriptions for that salient region.

This method has the following general steps:

(1) For the various regions of the given image, Region proposals are generated.

(2) CNN is used to get the region-based image features.

(3) Language model used the outputs of CNN to generate captions for every region.

Dense captioning [27] recommend architecture of a totally convolutional localized grid that consists of a convolutional network with a dense localized layer, and language model as LSTM.

G. Novel object-based image captioning

Image captioning methods, GAN and RL can generate diverse and multiple captions. In spite of achieving most favorable outcomes with deep learning image captioning methods, they mostly dependent on the sentence caption datasets and paired image. This kind of technique can generate only explanation of things inside the particular area. So that, that kind of method needs a huge set of training image-sentence pairs. This method generated explanation for novel objects that are not present in paired image-captions datasets.

This method has the below steps:

(1) On unpaired image and text data, lexical classifier and a separate language model are trained.

(2) A deep caption model is trained based on data of paired image caption.

(3) Lastly, both models are joined together to train them jointly for generating captions for novel object.

Image captions paired datasets are used for training of currently used image captioning method. As an outcome, if they detect any unnoticed things in test images, such method cannot present that object in their produced captions. [25] Projected a Deep Compositional Captioner (DCC) which denotes objects that are unseen in generated captions. Yao et al. [26] proposed a framework called copying mechanism which generates explanation of novel items. Such technique makes the use of a discrete item identification dataset to build classifiers for novel items. This contains suited words in the generated captions with copying mechanism and by a decoder RNN. Novel grid to identify the unobserved items from unpaired images is added by architecture of this method and integrated them with LSTM for generating captions.

III. DATASETS

For image captioning, most commonly used and popular datasets are MSCOCO, Flickr30k and Flickr8k. A very large dataset having all the images multiple captions is MSCOCO dataset.

A. MS COCO Dataset:

For image segmentation, image recognition, and captioning a huge dataset is used called MS COCO dataset. MS COCO dataset has features such as object segmentation, context recognition, and multiple objects per class. MSCOCO dataset is having more than 300,000 images, 80 object categories, more than 2 million instances, and 5 captions per image. Various image captioning methods [3, 4, 5, 6, 7, 8, 9] use the dataset in their experiments.

B. Flickr30K Dataset :

Flickr30K is a dataset used for automatically generating image explanation with ground truth. Flicker30K consist of 30000 images gathered from Flickr having 158000 captions were given by human annotators. For training, testing, and validation it doesn't make available any fixed split of images. Researcher is able to select their preferred number of count for training purpose, testing purpose and validation of images. The dataset includes indicators for mutual items, a color classifier, and a unfairness in the direction of selecting more items. Image captioning techniques [10, 11, 7, and 12] use this dataset for their experiments.

C. Flickr8K Dataset:

Flickr8k consist of 8000 images collected from Flickr. Flicker8K is a popular dataset in which training data comprises of six thousand images, the development plus test data, each comprises of one thousand images. All datasets images have five referential captions which is annotated by humans. Different methods of image captioning [13, 7, and 12] have accomplished experimentations with the same dataset.

D. Conceptual Captions Dataset:

The Conceptual Captions dataset comprises of about 3.3M images for training, for validating 28000 and for test set 22.5. Image-captioning dataset is having about 3.3M examples which is larger in magnitude than that of MSCOCO. It comprises of a widespread range of images, which includes natural images, professional photos, cartoons, drawings, etc. Its captions are based on descriptions taken from original Alt-text attributes, automatically transformed to achieve a balance between cleanliness, in formativeness and learnability [14].s

IV. Evaluation Metrics

For evaluating or measuring the performance of image captions various evaluation metrics are used. Small sentence can be evaluated by using BLEU metric. ROUGE is having different types. ROUGE is used evaluating different types of texts. Evaluation of several segments of a caption can be performed by METEOR. Compared to other evaluation metrics SPICE metrics is superior in understanding semantic particulars of captions.

A. BLEU:

[15] Is a metric which is used to estimate the superiority of text produced by machine. By comparing separate text parts with a set of referential texts, scores are calculated for all. Approximate total superiority of the produced text is evaluated based on averaged computed scores. Though, syntactical correctness of generated text is not considered here. Depending on the size and number of reference translations of the generated text performance of BLEU metrics may vary.

B. ROUGE:

[16] The quality of text summary can be measured by using ROUGE metrics. It processes by comparing word sequences, n-grams and word pairs with a set of reference summaries created by humans. There are different types of ROUGE such as ROUGE-1, 2, ROUGE-SU4, ROUGE-W.

C. METEOR:

[17] To evaluate the machine translated language METEOR metric is used. In METEOR standard word parts are matched with the referential texts. Further, for matching purpose synonyms and words stems of a sentence are too granted for matching.

D. CIDEr :

[18] It is a automatic consensus metric for evaluating description of image. Maximum existing datasets is having only 5 captions for each image. With the help of earlier evaluation metrics we can work with these small numbers of sentences but these metrics are not sufficient to measure the consensus between human judgment and generated captions. Though,

Human consensus is achieved by using term frequency-inverse document frequency (TF-IDF) in CIDEr.

E. SPICE:

[3] Based on semantic concept a new caption evaluation metric comes in to picture called SPICE which is built on a graph-based semantic illustration called scene-graph [20, 19], which is able to find out the data or information of various objects, their characteristics and relations from the image explanations.

V. Comparative Analysis

Sr. No.	Reference No.	Classifier	Metrics	Datasets
1	[2]	Convolution neural network,LSTM	BLEU	MS COCO, Flickr30k
2	[3]	Long Short Term Memory	ENGAN, E-GAN, SPICE,CIDEr	Flickr 30K MS COCO
3	[4]	Long Short Term Memory	BLEU, METEOR, ROUGE, CIDEr	FlickrStyle10K
4	[5]	Recurrent Neural Network	BLEU, METEOR, CIDEr	MS COCO
5	[6]	Long Short Term Memory	BLEU, METEOR, ROUGE, CIDEr	MS COCO
6	[7]	Long Short Term Memory	BLEU, R@K	Flickr 8K/30K, MS COCO
7	[8]	Long Short Term Memory	BLEU, METEOR, CIDEr	MS COCO
8	[9]	Recurrent Neural Network	BLEU, METEOR, ROUGE, CIDEr	Flickr 30K, MS COCO

9	[11]	Long Short Term Memory	BLEU, METEOR, CIDEr	Flickr 8K/30K, MS COCO
10	[12]	Long Short Term Memory	BLEU, METEOR, CIDEr	Flickr 8K/30K, MS COCO
11	[13]	Long Short Term Memory	BLEU, METEOR, ROUGE, CIDEr	Flickr 8K/30K, MS COCO
12	[14]	RNN, Long Short Term Memory	CIDEr, ROUGE-L, and METEOR	Conceptual Captions,COCO
13	[22]	Recurrent Neural Network	BLEU, METEOR, CIDEr	Flickr 8K/30K, MS COCO
14	[23]	Dependency tree relations	R@K, mrank	PASCAL1K, Flickr 8K/30K
15	[25]	Long Short Term Memory	ImageNet BLEU, METEOR	MS COCO,
16	[26]	Long Short Term Memory	ImageNet METEOR	MS COCO
17	[27]	Long Short Term Memory	METEOR, AP, IoU	Visual Genome

Table1. Deep learning-based approaches for image captioning

VI. Conclusion

Image captioning is a boom topic of image understanding. Image captioning is composed of two parts look and language expression which correspond to the two most important fields of artificial intelligence i.e. machine vision and natural language processing". Image captioning involves providing a brief and concise description of an image in natural language and is currently accomplished by techniques that use a combination of computer vision, natural language processing (NLP), and machine learning methods. In this, we have done review of image captioning methods based on deep learning. We discussed and compared different evaluation metrics and datasets.

VII. Future Scope

Though there is huge success achieved in recent years in image captioning but still there is a big room for enhancement. Supervised learning required huge set of labelled data for training purpose. Hence, in image captioning, unsupervised and reinforcement learning will be more popular in near future.

References

1. Micah Hodosh, PeterYoung, and Julia Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. Journal of Artificial Intelligence Research, 47:853–899, 2013.
2. 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.
3. 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). 2989–2998.
4. 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.
5. 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.
6. 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). 1151–1159.
7. 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.
8. Zhilin Yang Ye Yuan Yuexin Wu and Ruslan Salakhutdinov William W Cohen. 2016. Encode, Review, and Decode: Reviewer Module for Caption Generation. In 30th Conference on Neural Image Processing System(NIPS).
9. 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.
10. 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 The IEEE International Conference on Computer Vision (ICCV), Vol. 2.
11. 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.

12. 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, 1367–1381.
13. 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). 6298–6306.
14. Piyush Sharma, Nan Ding, Sebastian Goodman, Radu Soricut 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. Google AI Venice, CA 90291.In proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
15. 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.
16. 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. Barcelona, Spain.
17. 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.
18. 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.
19. 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 fourth workshop on vision and language, Vol. 2.
20. 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.
21. 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).
22. 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.
23. 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.
24. MD. ZAKIR HOSSAIN, FERDOUS SOHEL, MOHD FAIRUZ SHIRATUDDIN, HAMID LAGA. 2018.A Comprehensive Survey of Deep Learning for Image Captioning *ACM Comput. Surv.* 0, 0, Article 0
25. Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
26. 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)*. IEEE, 5263–5271.
27. 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.
28. R. Kiros, R. Salakhutdinov, R. Zemel, Unifying visual-semantic embeddings with multimodal neural language models, arXiv:1411.2539(2018).
29. Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr DollÃąr, 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.
30. 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.

