International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-9, July 2020

Analysis of Different Neural Network Techniques Used for Image Caption Generation

Samiksha Lambat, S. S. Sonawane

Abstract*: Artificial intelligence has open doors to new opportunities for research and development. And the recent development in machine learning and deep learning has paved a way to deal with complex problem easily. Now a day's every aspect of human life can now be thought of as a problem statement that can be implemented and is useful in one way or the other. One such aspect is human ability to understand and describe the surrounding to which they interact and take decision accordingly. This ability can also be used in machines or bots to make human and machine interaction easier. Generating captions for images is same we need to describe images based on what you see. This task can be considered as a combination of computer vision and natural language processing. In this paper we performs a survey of various methods and techniques that can be useful in understanding how this task can be done. The survey mainly focuses on neural network techniques, because they give state of the art results.

Keywords:* Image Captioning, deep learning, image annotation, caption generation.

I. INTRODUCTION

Machines are developing today at a rapid speed, they are getting more and more intelligent. They are able to make decisions and tasks easier for humans. So now a day's every human action is under observation, as to decide which operations can be replicated with machines. So thinking in this direction, humans can understand their surrounding environment and act accordingly by taking proper decisions. By having a look at any given object, by one's own imagination they can share the same knowledge with others. Humans can act, react, write and show emotions accordingly. Image captioning is the same process done by a machine or application that will generate descriptions of images up to human understanding automatically. Image captioning can be categorized under computer vision, natural language processing domain. Because not only we need to consider the objects in the image but also the context of the image. Considering it with computer vision alone earlier it was a difficult task but with the tremendous development in machine learning and deep learning algorithms and methods, this task has become easy.

Revised Manuscript Received on July 30, 2020.
***** Correspondence Author
Samiksha Lambat*, department of Computer Engineering, Pune Institute of Computer Technology. E-mail: lambatsami@gmail.com
Dr. S. S. Sonawane, department of Computer Engineering, Pune Institute of Computer Technology. E-mail: sssonawane@pict.edu

A. Motivation

It is a good opportunity to learn more about images and language processing models, and how they work. Image caption generation is also known as image annotation which is multi-label classification and multi-label ranking tasks that are popular in the machine learning community, which gathers the attention of many. Also its various applications such as an aid to blind people, CCTV cameras, self-driving cars, etc.

B. Challenges

- Compositionality and Naturalness: For compositionality we need to take into consideration the context of objects in visual scene (Image) and same should be reflected in natural language processing. Naturalness deals with irrelevant semantics that may appear in generated captions.
- Generalization: Some objects are common and may appear in different situations based on its requirements so while training the system may get confused so as which context is correct, so for this we need to have generalization.
- Evaluation: The generated captions are evaluated based on metric but this metrics only consider the sentences to check whether it is correct or not. It does not consider image while using evaluation**.**

II. IMAGE CAPTION GENERATION

We can widely categorize this task into two type of approach such as traditional and deep learning-based methods. The traditional category can include retrieval and template-based methods. This are the early approach that were used to generate captions, but due to deep learning result are more relevant.

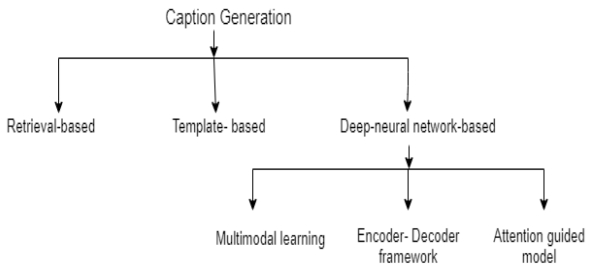

Fig.1, Categorization of approaches

- **Retrieval based image captioning**
  Consider we have a pool of already defined or specified sentences or captions, and we have a given query image, then the retrieval-based method will try to search or generate a caption or sentences from this available pool. This can be a new sentence or an existing one.

- **Template-based image captioning**
  The template-based method deals with the generation of captions as a syntactical and semantic process. Firstly we need to detect the visual concepts of an image and then connect it through sentence templates such as the grammar rules or optimized algorithm used to generate a sentence.
- **Deep learning-based methods**
  Deep neural networks use various different methods to make the task easier than traditional methods. The most commonly used method in this is encoder-decoder methods, where first the image is converted into an intermediate representation, then the decoder RNN take this as input and generates a sentence word by word.

III. CLASSIFICATION METHODS

Through this survey one can identify that every system that is present in this paper or not, have somewhat similar methods or architectures used, or a certain set of methods in common. So commonly used famous architectures are RNN, CNN and LSTM [1][2][3][4][5][6][9][10][12].

A. CNN (Convolution Neural Network)

A Convolution Neural Network (CNN) is a Deep Learning algorithm which is mostly used for visualizing images. CNN can be thought of as a multilayer perceptron that is a fully connected network. CNN uses a mathematical operation called convolution instead of general matrix multiplication. For images, CNN especially assigns importance (weights and biases) to various aspects in the image and be able to differentiate one from the other.

B. RNN (Recurrent Neural Network)

Recurrent Neural Network is a type of artificial neural network, which is also a feedforward network that has internal memory. RNN is recurrent in nature. The output of the present input depends on past output computation. After producing the output, it's copied and sent back to the recurrent network. for taking a decision, In other neural networks, all the inputs are independent whereas, in RNN, all the inputs are associated with one another. The aim is to employ sequential information.

C. LSTM (Long Short Term Memory)

Long Short-Term Memory (LSTM) networks are similar to recurrent neural networks, only they use a different function to compute the hidden state. They have memory cells that decide what to remember and not. LSTM is used to remember information for long periods of time. The vanishing gradient problem of RNN is resolved with LSTM.

IV. LITERATURE SURVEY

Jyoti Aneja et.al.[1] proposed a CNN(Convolution Neural Network) based image captioning method against the available LSTM based method. The CNN based model is used along with the attention method. The main aim of authors is to deal with the vanishing gradient problem. The model used in the paper was the Convolution machine translation model. For image embedding, the authors used the fc7 layer of the VGG16 network which is already pretrained on the imagenet dataset.

The authors used the Gated Linear Unit(GLU) activation for the convolution layer. According to the author, CNN with attention gives improved performance rather than LSTM along with attention. Dataset used is MSCOCO. Min Yang et.al.[2] designed and developed MLADIC(Multitask Learning Algorithm for cross-Domain Image Captioning). It is a multitasking system that performs the dual-task which is image captioning and image synthesis. The authors first tried to generate the caption for available and then from available text tries to generate images. For image captioning authors used the encoder-decoder model i.e CNN and LSTM. For image synthesis, they used the C-GAN (Conditional-Generative Adversarial Network). According to the authors, the earlier study uses the In-domain image captioning to generate the captions where the training and test images are from the same data distribution but authors tried to improve it by using images from different domains i.e. from the different data distribution. So the source domain used is MSCOCO and the target domain uses the Flickr and Oxford-102 dataset. MLADIC is a first model that performs the dual-task and to improve the performance they added a top-down attention mechanism to CNN which can generate plausible images. The decoder used by authors consists of two LSTM networks. The performance is evaluated against various baseline methods such as DCC, SAtT, Bottom-up, and top-down, attention model, SAdT, DCC dual, SAtT dual, SAdT models. Datasets used are MSCOCO, Flickr30k, and Oxford102. Oriol Vinyals et.al.[3] developed NIC(Neural Image Caption) which is an end-to-end neural network system used to generate the caption for images in the English language. NIC uses a generative model that is based on deep CNN that uses the neural and probabilistic framework for generating descriptions. The neural net is trained using stochastic gradient descent. CNN is used as an encoder that performs the pre-train task and then uses the last hidden layer as an input to RNN decoder that will generate captions. The LSTM model is used for the prediction of words in the sentences. The authors used the Beam Search for selecting the best sentence as a caption for the image. The authors also proposed an encoder-decoder system that is trained to maximize the log-likelihood of the target image descriptions. For evaluation purposes, they used the Amazon Mechanical Turk experiment to rate the caption manually by users. Jiuxiang Gu et.al.[4] proposed a system that can help to solve the problem of vanishing gradient of RNN. For this, the authors proposed to use the language CNN that is used to capture the long-range dependencies in sequences with RNN, which is then can be used in place of the LSTM network. Because the LSTM network has a memory cell that holds the information for some time steps only. The authors' combined language CNN with RNN. Pooling operations were used in earlier systems but authors removed this and instead applied a stack of convolution layers on top of each other. They added the multimodal fusion layer along with the language CNN which mixes the representation of the word with image features. The authors used Adam, a stochastic gradient descent method to train all models. The models used by authors are Simple RNN, language CNN, and language CNN with RNN. Dataset used is MSCOCO and Flickr30k. Xinpeng Chen et.al.

[5] proposed a new architecture called ARNeT(Auto-Reconstructor Network). This framework aims to improve the performance of the available traditional encoder-decoder method by reconstructing the previous hidden state with the present one. This, in turn, reduces the discrepancy between training and the inference process of caption generation. Also, the authors have used ARNeT along with its other variants to improve the performance on both image captioning and source code captioning. Authors used the Inception-V4 to encode any given Image along with LSTM which acts as an encoder for handling input sequence data. According to the authors, ARNeT aims at exploiting the relationship between the neighboring hidden states and also its reconstruction strategy behaves similarly to that of zoneout regularizer. The algorithm used was Adam. Dataset used was MSCOCO and Habeas Corpus for code captioning. The proposed framework was tested against available models such as NIC, Soft attention model, etc. Further to again check the regularizing ability of ARNet, the authors tested it on the MNIST to classify digits. Ilya Sutskerver et.al[6] proposes a multilayer LSTM, where the first LSTM is used for mapping input to fixed-size vectors. And another LSTM is used for decoding from the first vector. The whole idea of the authors was to learn data from a long-range of available sequences. The conclusion was that LSTM works well on long sequences. The algorithm used was a beam search algorithm. The dataset used was WMT 14 English to French. For evaluation purposes, the BLEU score was 34.8.Nal Kalchbrenner et.al.[7] proposed a Recurrent Continuous Translation model(RCTM) and its variants to perform 4 experiments for improving multilingual translations. RCTM provides a generalized way that can map source language to target language. Two variants used by authors were RCTM1 and RCTM2. RCTM1 used a convolution sentence model(CSM) that uses the n-gram representation, CSM here is a hierarchical structure. RCTM2 defines the length of the target sentence, then constructs the target with the help of n-gram. Evaluation is done by using rescoring and BLEU. For rescoring and performance purposes “cdee” system, and BLEU were used respectively. Dataset used was the WMT news commentary section which is a bilingual corpus, the source is English and the target is French. Yezhou Yang et.al[8] proposed a system that makes use of the semantic and syntactic concepts of sentence formation. For more relevant captions the context of the image is taken into consideration while caption generation. The important points to include while generating caption is a noun, verb, scene, and prepositions. The authors conducted three experiments to test and observe the best result. The authors used the SVM classifier for object detection and Pascal VOC 2008. Dataset used was English Gigaword, UIVC Pascal Sentences. For evaluation purposes, the precision score was used to determine the effectiveness of generated text, also the performance measure used was ROGUE 1.Yoshitaka Ushiku et. al.[9] developed an online platform for sentence generation from images. This system mainly focuses on the “multi-keyphrase” problem. In this the authors make use of semantic knowledge such as grammar semantics along with multi-keyphrase to generate sentences. The online learning algorithm called PAAL was designed to perform multilabel classification along with a random selection of labels. PAAL was implemented using MATLAB. Dataset used were Corel 5K, ESP game, IAPR-TC12. For evaluation purposes, precision, recall, F-measure and BLEU(4-gram) and NIST(5-gram) was used.Chuang Gan et.al[10] provided a new way to improve available caption generation. According to author captions generated by previous systems lacks some factors that make it easier to understand what the image is about. So the author considered style which is either humorous or romantic, which results in the generation of stylized captions that can convey the context of the image more effectively. The same encoder-decoder method of the neural network was used. They replaced the traditional LSTM with the factored LSTM. For authors this style generation was a semi-supervised and unsupervised captioning problem that uses the Adam algorithm. Dataset used was created by authors themselves which includes the FlickrStyle10K based on Flickr30K, for video captioning they used the Youtube2text dataset. Ali Farhadi et.al[11] give a simplified model for sentence generation. The authors used a pool of sentences for searching a sentence that matches the image descriptions. The match made is based on some factors such as image potentials and sentence potentials. According to authors the concept of potentials is to match the image and available respective sentences based on some meaning. Meaning is an intermediate understanding in the form of a triplet (object, action, scene). Here also the authors have created their customized dataset based on PASCAL2008. For matching sentences they used the LIN similarity measure for objects and scenes, action co-occurrences, node and edge potentials. The performance measures used were Tree-F1, BLEU. Since they have their own dataset they used Amazon Mechanical Turk to generate available sentences.

Table-I: The following table compares the result of performance measure given for caption generations. The BLEU (N=1,2,3,4) measure is used as B1, B2, B3, B4. M stands for METEOR, R stands for ROUGE and C stands for CIDEr metrics.

Methods	B1	B2	B3	B4	M	R	C
LSTM[1]	.710	.537	.399	.299	.246	.523	.904

CNN+Attention[1]	.715	.545	.408	.304	.246	.525	.910
MLADIC[2]	79.4	63.1	48.2	36.1	28.1	57.5	119.6
NIC[3]	---	---	---	27.7	23.7	---	85.5
Encoder-Decoder + ArNET[5]	0.196	0.107	0.075	0.058	0.089	0.215	---
Attention Encoder-Decoder + ArNET[5]	0.255	0.173	0.139	0.120	0.123	0.289	---
StyleNet(F)[10]	41.2	21.4	12.1	7.7	0.135	0.36	0.24
StyleNet(Romantic)[10]	46.1	24.8	15.2	10.4	0.154	0.38	0.31
StyleNet(F)[10]	42.9	22.3	12.9	7.7	0.135	7.7	0.23
StyleNet(Humorous)[10]	48.7	25.4	14.6	10.1	0.152	10.1	0.27

Table-II: The following table provides the dataset distribution for different dataset used in the literature survey of this paper.

Paper	Dataset	Training Images	Validation Images	Testing Images
[1]	MSCCOCO	113287	5000	5000
[2]	MSCOCO	2783	40504	40775
	Flickr30k	29000	1000	1000
	Oxford-102	6189	1000	---
[3]	PASCAL VOC 2008	---	---	1000
	Flickr8k	6000	1000	1000
	Flickr30k	28000	1000	1000
	MSCOCO	82783	40504	40775
	SUB	1M	---	---
[4]	MSCOCO	---	5000	5000
	Flickr30k	29000	1000	1000
[10]	Youtube2Text	1200	100	670

V. DATASET

A. PASCAL (Pattern Analysis, Statistical Modeling, and Computational Learning)

Standardized image data sets for object class recognition. Also, it provides a common set of tools for accessing the data sets and annotations. It is a dataset available from PASCAL Visual Object Classes Challenge.

Table-III: Following table gives the details of the number of classes used while considering an image, and number of total images used for that particular year dataset.

Year	Classes	Images	Images With Annotated objects
2005	4	1578	2209
2006	10	2618	4754
2007	20	9963	24640
2008	20	4340	10363
2009	20	70504	17218
2010	20	10103	23374
2011	20	11530	27450
2012	20	11530	27450

B. MSCOCO (Microsoft Common Objects in Context)

Microsoft COCO is image recognition, segmentation, and captioning dataset. Images available in this dataset do not focus on iconic images. MSCOCO[21] dataset was released in two parts, one in 2014 and others in 2015 respectively. In 2014, the size of the dataset was 82,783 images for training, 40504 for validation and 40775 for testing. In 2015, the size of the dataset was 165482 images for training, 81,208 for validation and 81434 for testing. This dataset covers mostly 80 object categories and gives 5 captions per image.

C. FLICKR

It is a recognized benchmark collection where we have images paired with five different captions that provide clear descriptions of salient entities and events. The images presented were chosen from six different Flickr groups and tend not to contain any well-known person. Images in this dataset do not contain any famous person or place so that the entire image can be learned based on all the different objects in the image.

VI. PERFORMANCE MEASURES

Performance measures are something that is a quantifiable entity or a result of activities that are needed to make sure that whatever output we are getting is up to the mark and can be trusted. To make sure of this, we need some kind of metrics or measures that can evaluate the generated output. For the purpose of caption generation also, we also need to use some following performance measures:

A. BLEU (Bilingual Evaluation Understudy)

It is originally a score that is used for comparing a candidate translation of text to at least one or more reference translations. And also it can often be used for evaluating text that is generated for natural language processing tasks[17]. The counting of matching n-grams is modified to make sure that it takes the occurrence of the words within the reference text under consideration, not rewarding a candidate translation that generates an abundance of reasonable words. This is often referred as modified n-gram precision. For BLEU[17] we'd like to supply a score for the entire corpus using the modified precision scores for the segments are combined using the geometric mean multiplied by a brevity penalty to stop very short candidates from receiving too high a score. Let *r* be the total length of the reference corpus, and *c* the total length of the translation corpus.

BP = {1 if c>r } or
BP = { $e^{(1-r/c)}$ if c ≤ r } Then,
BLEU = BP * exp ($\sum_{n=1}^{N} w_n log\ p_n$)
The ranking behavior is more immediately apparent within the log domain,

$$\log BLEU = \min\left(1-\frac{r}{c},0\right)+\sum_{n=1}^{N} w_n log\ p_n$$

B. METEOR (Metric for Evaluation of Translation with Explicit ORdering)

METEOR [18], is a metric for machine translation evaluation that is based on unigram matching between the machine generated translation and human-produced reference. These unigrams considers, stem, synonym, and paraphrase that matches between words and phrases. Also, METEOR uses the harmonic mean based on precision and recall. In this alignment is done based on mapping.

C. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

It is used for evaluating summarization of texts as well as machine translation. [19]It works by comparing an n-gram recall between a candidate summary and a set of reference summaries. There are various version of ROUGE available, ROUGE-1 refers to unigram, ROUGE-2 refers to bigram, ROUGE-L refers to Longest Common Subsequence (LCS), ROUGE-W refers to Weighted LCS-based statistic, ROUGE-S refers to Skip-bigrams based on co-occurrence statistics and ROUGE-SU refers to Skip-bigram plus unigram-based co-occurrence statistics

D. CIDEr (Consensus-based Image Description Evaluation)

This metric consider the generated text to human-evaluated texts to determine how close the results are. It checks how many n-grams are present in the candidate as well as reference sentences. All the words are first mapped to their root words. This metric for evaluating image captioning is based on consensus, where consensus-based evaluation means that how close the generated text is to human's ability of sentence generation.

VII. CONCLUSION

In this paper, we have done a survey on how to generate captions for images. There are various methods available that can be used but we particularly explored the neural network based techniques available, and how they work. This survey also explains how different combination of neural network techniques generates the acceptable results. Also we highlighted some of the dataset that are commonly used for this purpose.

This paper not only does survey of caption generation images but also does survey of sentence generation techniques that can be used along with it. Then the focus was shifted on performance measurement, where different metrics that are commonly used for evaluation purpose was surveyed.

REFERENCES

1. J. Aneja, A. Deshpande, A. G. Schwing, "Convolutional Image Captioning", IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
2. M. Yang, W. Zhao, W. Xu, Y. Feng, Z. Zhao, X. Chen, and K. Lei, "Multitask Learning for Cross-Domain Image Captioning" IEEE Transaction on Multimedia, Vol.21, NO. 4, April 2019.
3. O. Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and Tell: A Neural Image Caption Generator", arx1411.4555v2, 20 April 2015.
4. J. Gu, G. Wang, J. Cai, T. Chen, "An Empirical Study of Language CNN for Image Captioning", IEEE International Conference of Computer Vision, 2017.
5. X. Chen, L. Ma, W. Jiang, J. Yao, and W. Liu, "Regularizing RNNs for Caption Generation by Reconstructing the Past with the Present", IEEE/CVF Conference on Computer Vision and Pattern Recognition,2018.
6. I. Sutskever, O. Vinyals, Q. V. Le, "Sequence to sequence learning with neural network", Proceedings of the Advances in Neural Information Processing Systems, 2014.
7. N. Kalchbrenner, P. Blunso, "Recurrent Translation Models", Proceedings of the Conference on Empirical Methods in Natural Language Processing 2013.
8. Y. Yang, C. L. Teo, H. Daume, Y. Aloimono- Corpus, "Guided sentence generation of natural images", Proceedings of the Conference on Empirical Methods in Natural Language Processing 2011, pp. 444-454.
9. Y. Ushiku, T. Harada, Y. Kuniyoshi, "Efficient image annotation for automatic sentence generation", Proceedings of the 20th ACM International Conference on Multimedia, 2012.
10. C. Gan, Z. Gan, X. He, J. Gao, L. Deng, "StyleNet: Generating Attractive Visual Captions with Styles", IEEE Conference on Computer Vision and Pattern Recognition, 2017.
11. A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchain, J. Hockenmaier, M. D. Forsyth, "Every picture tells a story: Generating sentences from images", Proceedings of the European Conference on Computer Vision, 2010, pp.15-29.
12. M. Tanti, A. Gatt, and K. P. Camilleri, "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator", Proceedings of The 10th International Natural Language Generation conference, pages 51–60, September 4-7 2017.
13. R. C. Luo, Y.T. Hsu, Y.C. Wen and H. J. Ye, "Visual Image Caption Generation for Service Robotics and Industrial Applications", IEEE, 2019
14. M. Konno, K. Suzuki and M. Sakamoto, "Sentence Generation System Using Affective Image", 2018 joint 10th International Conference on Soft Computing and Intelligent Systems and 19th International Symposium on Advanced Intelligent Systems.
15. S. Venugopalan, L. A. Hendricks, and M. Rohrbach, "Captioning Images with Diverse Objects", arXiv: 1606.07770 [cs.CV] 20 Jul 2017.
16. H. Fang, S. Gupta, F. Iandola and R. K. Srivastava, "From Captions to Visual Concepts and Back", arXiv: 1411.4952v3 [cs.CV] 14 Apr 2015.
17. K. Papineni , S. Roukos , T. Ward, and W. J Zhu, " BLEU: a Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.
18. S. Banerjee, A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, June 2005.
19. Chin-Yew Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", Text Summarization Branches Out, July 2004.
20. R. Vedantam, C. Lawrence Zitnick, D. Parikh, "CIDEr: Consensus-based Image Description Evaluation", arXiv: 1411.5726v2 [cs.CV] 3 June 2015.
21. T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollar, "Microsoft COCO: Common Objects in Context", arXiv: 1405.0312v3 [cs.CV] 21 Feb 2015.

AUTHORS PROFILE

Samiksha A. Lambat, B.Tech in Information Technology from Government College of Engineering, Amravat, Maharashtra, India. And is currently pursuing Masters in Computer Engineering, from Pune Institute of Computer Technology, Pune, India.

Dr. S. S. Sonawane, PhD in Computer Engineering from College of Engineering Pune (COEP). And she is currently a Associate Professor in Computer Engineering department at Pune Institute of Computer Technology, Pune, India.