IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Conceptual Review of Deep Learning Methods for Automatic Image Caption Generation

S. H. Patel[1*], N.M. Patel[2], D.G. Thakore[3]

[1,2,3]Department of Computer Engineering, Birla Vishvakarma Mahavidhyala, Vallabh Vidhyanagar, India

[*]*Corresponding Author: smit.academics@gmail.com, Tel.: +91-94093-72273*

DOI: https://doi.org/10.26438/ijcse/v7i3.987991 | Available online at: www.ijcseonline.org

Accepted: 07/Mar/2019, Published: 31/Mar/2019

Abstract— Automatic generation of caption for given images is a complex AI task. It is a problem of generating textual description for a given input image. This involves both image understanding and natural language generation. This is a very dynamic field. A lot of work has been done and currently ongoing in this domain. The recent frontiers of the fields are based on deep learning based methods. The purpose of this article is to provide overview of deep learning based image captioning methods to readers. The readers will first get basic concepts which are used to in development of various methods. Then basic information on datasets is given. Then three existing work are discussed followed by very brief discussion on other works. Concisely, this article presents classification of existing approaches, popular datasets and some of existing models followed by brief discussion of other works. Initially, the topic is introduced and then broader classification of deep learning based methods is discussed. At last, brief discussions on some methods are done.

Keywords— Image Caption Generation, Deep Learning, Computer Vision

I. Introduction

Image caption generation is a task of generating natural language description (typically single line) for a given image input. It can also be viewed as image understanding and machine translation problem.

Image caption generation is considered as one of the complex AI problems as it involves both image understanding and natural language generation. So, it has large research community from both computer vision and natural language generation fields. Despite many years of its origin, it's in still very hot research area.

Image captions have very wide scope of applications. It can be used to index images in image retrieval. It can also be used to assist blind people. Also, it can be used by social media to understand images uploaded by users.

Image captioning problem was initially approached by template based methods and retrieval based methods. After rise of deep learning methods, the problem is approached by pipeline based methods and end-to-end methods. Here, deep learning methods have shown state of the art results recently. In this paper, we will briefly summarize some of deep learning based methods in more detail and get overview of some other notable work.

After generating captions using various methods, they should be compared with the given captions from dataset to measure goodness of generated captions. This process of evaluation should be automated as it is difficult for human to assess all generated captions of large datasets in specific time. There are several matrices that can give measure of goodness for generated captions. The evaluation of generated captions are done using matrices like BLEU [1], METOR [2], CIDEr [3] etc. Typically, these matrices give score in range (0, 1) or (0, 100). The larger score indicate better matching of generated captions with reference captions. Although, these are not discussed in current work, readers are highly encouraged to get basic understanding of these from sources like cited references and internet.

In this paper, our contribution is to present the overall introduction of the field and present important concepts in relatively simple manner.

The rest of the paper is organized as follows. First section is introduction. The second section contains classification of methods. The third section provides overview of datasets. The fourth section contains discussion of existing works in the field. The fifth section contains summary of discussion. And, the sixth section contains references.

II. Classification of Methods

Image captioning techniques are classified in various categories. P. Shah, V. Bakrola, and S. Pati [4] classified them into retrieval based techniques and generation based techniques. Retrieval based techniques operate in two phases, image matching and caption generation. Generation based techniques are further classified as pipeline based techniques and end to end techniques. Pipeline based techniques perform language modeling task and image recognition task separately and then combines them. Whereas encoder decoder based techniques learn both the parts at the same time.

M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga [5] classified image captioning techniques into traditional machine learning based and deep learning based techniques. Traditional machine learning based methods used methods like LBP(Local binary patterns), HOG(Histogram of Oriented Gradients), SIFT(Scale-invariant feature transform) etc. whereas, Deep learning based techniques used combination of models like CNNs and RNNs for caption generation. Where, CNNs are used to extract image features and RNNs are used for language modeling. [5]

M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga [5] claimed that most image captioning based articles are classified into template based, retrieval based and novel caption generation. They further claimed that most existing review at time of his writing were mainly covering template based and retrieval methods in greater details and very few deep learning based methods were discussed in those reviews and hence he mainly focused on survey of deep learning based methods for caption generation. This paper will also focus more on deep learning based methods.

A detailed discussion on classification of deep learning based methods for image captioning is done by M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga [5] in their paper. We will briefly discuss the categories of image captioning methods based on discussion and methodology adopted by M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga [5].

A. Visual Space vs. Multimodal Space

In visual space based methods, the image features and their respective captions are handled separately and are independently passed to the language decoder. Whereas in multimodal space based techniques, a combined and shared multimodal representation is passed to language model [5].

B. Supervised Learning vs. Other Deep Learning

In Supervised methods, Images and their corresponding labels are used to learn the task. Whereas, other techniques like generative adversarial networks (GANs), Reinforcement learning etc were also applied to this task. These methods are termed as other deep learning based methods. Generative adversarial networks contain generator models and discriminator models and they are collectively trained by adversarial training which is inspired from game theory to train generator. Reinforcement learning techniques train agent through exploration and reward [5].

C. Dense Captioning vs. Captions for the whole scene

In dense captioning, captions are generated for each region whereas in whole scene based approach, the caption is generated based on entire scene and not region vise [5].

D. Encoder-Decoder Architecture vs. Compositional Architecture

Encoder-decoder framework provides simple and end to end model by using CNN as image encoder and RNN as language decoder. These are very similar to the encoder-decoder networks used in machine translation. Whereas the compositional architectures are made-up of several independent functioning blocks [5]. M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga [5] described it as "CNN is used to extract the semantic concepts from the image. Then a language model is used to generate a set of candidate captions. In generating the final caption, these candidate captions are re-ranked using a deep multimodal similarity model."

E. Other architectures

Attention-based, Semantic concept based, novel object based and stylized captions are put under other architecture group by M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga [5]. Attention-based approaches focus on salient features of image on each time step of language model. Semantic concept based methods selectively attend to a set of semantic concept proposals extracted from the image and then combine them into hidden states and the outputs of RNN. Novel object based methods can generate descriptions of novel objects which are not present in paired image-captions dataset. Stylized caption based methods can generates captions that can be more expressive and attractive than just only flat descriptions which are generated by other methods [5].

III. DATASETS

Image captioning datasets contains image and their respective reference caption. S. Bai and S. An [6] stated that some of the widely used datasets for image captioning are Flicker8k [7], Flicker30k [8], MSCOCO [9].

A. Flicker8k[7]

Flicker8k [7] contains 8000 images extracted from Flicker. It mainly contains natural images of human and animals. Each image is annotated by five sentences based on crowd sourcing service from Amazon Mechanical Turk.

B. Flicker30k[8]

It is extension of Flicker30k [8]. It contains approximately 30000 images along with their respective captions.

C. MSCOCO[9]

MSCOCO [9] is one of the huge dataset available for image captioning and object annotation. It contains 328,000 images and 5 sentences for each image.

IV. EXISTING WORK IN IMAGE CAPTION GENERATION

Following text will give brief overview of some existing

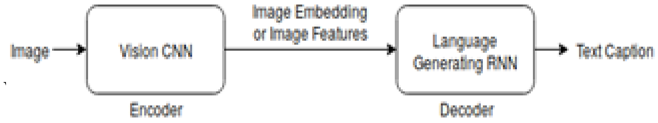

Figure 1: The encoder decoder based image captioning model [10]

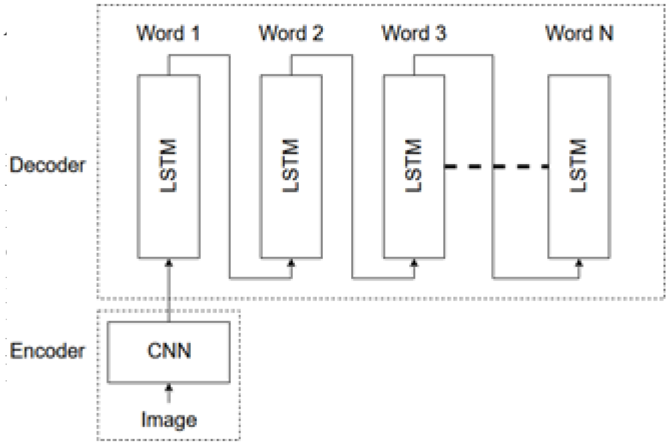

Figure 2: The expanding structure of Show and tell NIC[11]

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan [10] showed BLEU-1 score of 66 and 63 on Flicker30k [8], Flicker8k [7] respectively. They also reported BLEU-4, METEOR and CIDER scores of 27, 7, 23.7, 85.5 respectively on MSCOCO [9] dataset.

B. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[12]

Xu et al [12] proposed new image captioning model with attention mechanism. It can focus on salient regions of image

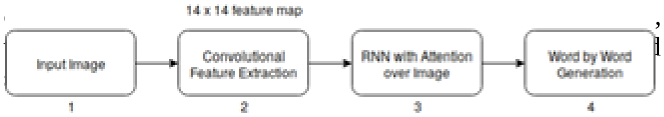
14 x 14 feature map
Input Image
Convolutional Feature Extraction
RNN with Attention over Image
Word by Word Generation
1
2
3
4

Figure 3: Architecture of Show Attend Tell model [12]

Major contributions in his work [12] are reported as "(1) Attention based two models under common framework, 'soft' deterministic attention mechanism which is trainable by standard back-propagation methods and 'hard' stochastic attention mechanism which is trainable by maximizing an approximate variational lower bound. (2) How we can gain insight and interpretations of this framework by visualizing 'where' and 'what' the attention focused on. (3) Quantitative validation of attention mechanism in caption generation on 3 benchmark datasets Flicker8k [7], Flicker30k [8], MSCOCO [9] is done".

Xu et al.[12] reported score of 67, 45.7, 31.4 and 21.3 of BLEU-1, BLEU-2, BLEU-3 and BLEU-4 respectively on flicker8k[7] dataset, score of 66.9, 43.9, 29.6 and 19.9 of BLEU-1, BLEU-2, BLEU-3 and BLEU-4 respectively on Flicker30k[8] dataset and score of 71.8, 50.4, 35.7, 25.0 and 23.04 of BLEU-1, BLEU-2, BLEU-3 and BLEU-4 respectively on MSCOCO[9] dataset.

C. Deep Captioning with multimodal recurrent neural networks (M-RNN)[13]

Novel multimodal recurrent neural network architecture is proposed by Mao et al [13] it consists of five layers of two word embedding layers, recurrent layer, multimodal layer, and softmax layer. [13]

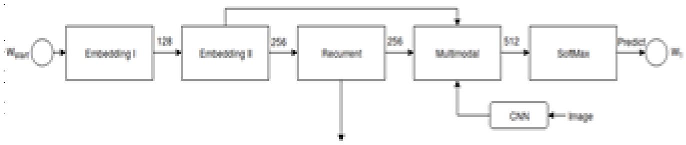

Figure 4: Architecture of Multimodal RNN proposed and described by Mao et al [13]

Mao et al [13] reported score of 60, 41, 28 and 19 of BLEU-1, BLEU-2, BLEU-3 and BLEU-4 respectively on Flicker30k [8] dataset and score of 67, 49, 35 and 25 of BLEU-1, BLEU-2, BLEU-3 and BLEU-4 respectively on MSCOCO [9] dataset.

D. Other works in brief

D.J. Kim, D. Yoo, B. Sim, and I. S. Kweon [14] proposed novel method to perform transfer learning on sentences. The image features are extracted with Deep Fisher Kernel. Modification in LSTM called gLSTM(guided LSTM) is employed which prevents dilution of CNN feature information as soon as subsequent words are given. Some performance improvements were reported in this method.

Anderson et al [15] proposed model to combine top-down and bottom-up attention mechanism in single model for image captioning and visual question answering.

J. Lu, C. Xiong, D. Parikh, and R. Socher [16] proposed adaptive attention mechanism which learns when and where to attend. The authors argued that not all words require visual attentions as some of the words can be reliably predicted from language model. For example the word "phone" can be predicted at the end of "talking on a cell" and words like "is" and "the" are purely dependent on language model rather than any visual object. The proposed model used visual sentinel to address this and at each step, model decides whether to attend image region or visual sentinel. The model improved performance and visualization of attention.

M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek [17] proposed novel attention based model "Areas of Attention". They claimed that his approach models the dependencies between image regions, caption words and state of RNN language model. They also proposed and compared 3 other ways to generate attention area: CNN activation grids, object proposals, spatial transformers net applied in a convolutional fashion. Reported results are good and comparable with the best results.

K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang [18] proposed image captioning system that exploits parallel structure between images and captions. They showed that the model aligns the process of generation of caption and shifting of attention among visual regions.

A. Poghosyan and H. Sarukhanyan [19] introduced modified LSTM cell with additional gate responsible for image features. They claimed that this modification results in improved caption generation.

V. Mullachery and V. Motwani [20]. Discussed results from their experiments. They incorporated transfer learning, hidden layers in RNN and ResNet in the lieu of VGGNet as experimental modifications. They had also shown a toy application of the captioning system on video and shows the challenges encountered.

V. SUMMARY

The image caption generation is still very fast growing research field. Every day, new works are published. Our effort in this work is intended to give brief conceptual overview of the field. We have summarized some important terminology for classification of existing work. And, we have also taken basic overview of some popular datasets and some noteworthy contributions. This will hopefully help readers to get overview of the field and basic understanding of existing work.

REFERENCES

[1] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, “Bleu: a method for automatic evaluation of machine translation,” in proc. Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 311–318, 2002.

[2] M. Denkowski and A. Lavie, “Meteor universal: Language specific translation evaluation for any target language,” in proc. EACL 2014 Workshop on Statistical Machine Translation, 2014, Baltimore, USA.

[3] R. Vedantam, C. L. Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,”, in proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4566-4575

[4] P. Shah, V. Bakrola, and S. Pati, “Image captioning using deep neural architectures,” in proc. International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), IEEE, Piscataway, NJ, mar 2017.

[5] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, “A comprehensive survey of deep learning for image captioning,” CoRR, vol. abs/1810.04020, 2018.

[6] S. Bai and S. An, “A survey on automatic image caption generation,” Neurocomputing, vol. 311, pp. 291–304, 2018.

[7] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” Journal of Artificial Intelligence Research, vol. 47, pp. 853–899, aug 2013.

[8] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, IEEE, dec 2015, pp. 2641-2649.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Computer Vision – ECCV 2014, pp. 740–755, Springer International Publishing, 2014.

[10]O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, jun 2015, pp. 3156-3164.

[11] X. Liu, Q. Xu, and N. Wang, “A survey on deep neural network-based image captioning,” The Visual Computer, jun 2018, pp. 1–26.

[12]K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in proc. 32Nd International Conference on Machine Learning - Volume 37, ICML’15, JMLR.org, 2015, pp. 2048–2057.

[13]J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, “Deep captioning with multimodal recurrent neural networks (m-rnn),” eprint arXiv:1412.6632 [cs.CV], Jun 2015.

[14]D.-J. Kim, D. Yoo, B. Sim, and I. S. Kweon, “Sentence learning on deep convolutional networks for image caption generation,” in proc. 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), IEEE, Xi'an, China, aug 2016.

[15]P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and VQA,” in proc. IEEE Conference on Computer Vision and Pattern Recognition, (CVPR) 2018, Salt Lake City, UT, USA, Jun, 2018, pp. 6077-6086.

[16]J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via a visual sentinel for image captioning,” in proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, USA, jul 2017..

[17]M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, “Areas of attention for image captioning,” in proc. IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, Italy, oct 2017.

[18]K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, “Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, pp. 2321–2334, dec 2017.

[19]A. Poghosyan and H. Sarukhanyan, “Short-term memory with read-only unit in neural image caption generator,” in proc. Computer Science and Information Technologies (CSIT), IEEE, Yerevan, Armenia, sep 2017.

[20]V. Mullachery and V. Motwani, “Image captioning,” arXiv:1805.09137 [cs.CV], may 2018.

Authors Profile

Mr. S H Patel pursed Bachelor of Engineering in Computer Engineering from Gujarat Technological University, Gujarat, India in 2017. He is currently pursuing Master of Technology in Software Engineering

Dr Narendra M Patel received his BE degree in Electronics Engineering from M S University, Baroda in 1993 and ME degree from M S University, Baroda in 1997. He received PhD degree from SVNIT, surat in 2012. He is currently working as Associate Professor in Computer Engineering Department, BVM Engineering College, V V Nagar, and Gujarat. He has more than 24 years of academic experience. His research interests include Digital Image Processing, Real time operating systems, Distributed systems and Computer Graphics. He authored more than 60 papers which are published in Prominent international and conference proceedings. He has guided more than 55 ME dissertations in Computer engineering. He has rendered his service as expert in various STTPs, conferences and workshops.

Dr.D.G. Thakore pursued his bachelor of engineering in computer engineering from Sardar Patel University (SPU) and master of technology in computer engineering from IIT-Delhi. He pursued his PhD. in computer engineering from M.S University, Vadodara, Gujarat, India and currently working as Professor and Head of Computer Department in Birla Vishvakarma Mahavidyalaya, Gujarat Technological University (GTU) since year 1991 and 2014 respectively. He has professional membership of ISTE, India. He successfully presented and published multiple papers in national journals and international journal. His main research work focuses on Digital Image Processing. He has 25 years of teaching experience along with research experience and 4 months of industry experience.