iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IV **Month of publication:** April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33270

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

A Survey on Visual Story Generation for Images in Sequence

Aneri Patel[1], Prof. Pranay Patel[2], Dr. Hemant Vasava[3]
[1, 2, 3]*Department of Computer Engineering Birla Vishvakarma Mahavidyalaya Vallabh Vidyanagar, Gujarat, India*

Abstract: ***Stories are considered as a fundamental tool of humans for communication. We can enable this understanding to computers also to generate a story from the given input image sequence i.e. image understanding. This emphasizes a classic captioning problem of computer vision. Generating a story from images in sequence is truly interesting than generating a single line captions for an image. Story generation from image sequence not only involves current image but also its previous image vector to generate caption at that stage and this process is carried out until the end of the sequence. This is needed as we need to generate the story in continuation. Common methods for this purpose, according to our literature survey are Advanced Neural Network like CNN, RNN and LSTM. But as CNN lacks remembering capacity and as we have to consider the previously generated caption till the end of sequence for continuation of story, CNN proves a perfect misfit for the said purpose. Hence RNN along with its LSTM is used to get more better results. Also feature vector generation, pre-processing, NLP and evaluation jointly contributes to generate interesting stories. This paper provides a brief survey of various research carried out so far in this field. It also discusses the existing data sets, feature Generation methodologies and evaluation metrics used in this system.***

Index Terms: Image Understanding, Pre-Processing, Feature Generation, Sequence to Sequence Model, Recurrent Neural Network – Long Short Term Memory, Natural Language Processing

I. INTRODUCTION

Story generation is a challenging Computer Vision problem where a story based description must be generated from a sequence of images [1]. It requires both, methods from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. Recently, deep learning methods have achieved state-of- the-art results on examples of this problem[1]. This problem is different from the basic image captioning problem where a single image is given and a caption is generated out of that as an output. But visual story generation takes sequence of images as an input and generates story for this sequence in continuation by understanding the image and by considering both previous and current images for generating story.

Initially feature vectors for those images are required to be generated using any of the Neural Network models. These feature vectors serves as an input to RNN. All necessary Pre-Processings to reduce the corpus also is needed along with NLP, and at last Evaluation Metrics helps to check the relevance of the story.

Here in the figure 1, we can see that sequence of 5 images is given as an input to the system and captions are generated out of that sequence. Starting from the first image, once the caption for the first image is generated, model uses that caption as the partial caption for the next image in the sequence. Now the next image vector along with the previous partial caption generates the caption for the current image which is embedded in the partial caption. This process of embedding the current results with partial caption is carried out until last image in sequence and at the end fully flourished story at the end of the sequence is generated.

Fig. 1. Visual Story Generation[2]

In the figure-1, system is trained using RNN-LSTM to generate the story in continuation. Many datasets have been introduced in the past few years to solve the Visual story generation problem. The most commonly used is VIST [2], but other datasets like MSCOCO [15], Flickr[14] etc. are also prominent. More research work has been published in this area in the last few years.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

II. LITERATURE REVIEW

Computer Vision have gained a lot of attention in recent years and development in field has got huge recognition in today's AI era. With the growing interest in the field of Computer vision, there are increasing research works also carried out. Visual Story Generation is a combination of two research areas: Image Understanding and NLP. Image understanding focuses on capturing areas of interest from the image. This content is converted to a specific form and presented in textual representation in a correct order, emphasizing the language construct.

In the work by Xin Wang et al. [43], They proposed an Adversarial Reward Learning (AREL) framework[40] and applied it to boost visual story generation. They also evaluated an approach on the Visual Storytelling (VIST) dataset[12] and achieved the state-of-the-art results on automatic metrics. They also designed and performed a comprehensive human evalu- ation via Amazon Mechanical Turk[45], which demonstrated the superiority of the generated stories of their method on relevance, expressiveness, and concreteness. They not only introduced a novel adversarial reward learning algorithm to generate more human-like stories given image sequences, but also empirically analyzed the limitations of the automatic metrics for story evaluation.

Taehyeong Kim et al. [44] proposed a deep learning network model, GLAC Net, that generates visual stories by combining global-local (glocal) attention and context cascading mecha- nisms. Their model incorporates two levels of attention, i.e overall encoding level and image feature level, to construct image-dependent sentences. While standard attention config- uration needed a large number of parameters, the GLAC Net implements them in a very simple way via hard connections from the outputs of encoders or image features onto the sentence generators. The coherency of the generated story is further improved by conveying (cascading) the information of the previous sentence to the next sentence serially. They also evaluated the performance of the GLAC Net on the visual storytelling dataset (VIST)[12] and achieved very competitive results compared to the state- of-the-art techniques.

Chih-Chia Li et al. [45] introduced KG-Story, a three- stage frame-work that allows the story generation model to take advantage of external Knowledge Graphs to produce interesting stories. KG-Story distills a set of representative words from the input prompts, enriches the word set by using external knowledge graphs, and finally generates stories based on the enriched word set. This distill-enrich-generate framework allowed the use of external resources not only for the enrichment phase, but also for the distillation and generation phases. In this paper, they showed the superiority of KGStory for visual storytelling[12], where the input prompt was a sequence of five photos and the output was a short story. Per the human ranking evaluation, stories generated by KG-Story were on average ranked better than that of the state-of-the art systems. In the work by Oriol Vinyals et al. [4] they presented an idea of CNN[21] as an image "encoder", by first pre-training it for an image classification task and using the last hidden layer as an input to the RNN [22] decoder that generates sentences . They called this model the Neural Image Caption, or NIC[23]. Their model is quite accurate, and they verified it both qualitatively and quantitatively. For instance, while the BLEU-1 score[24] on the Pascal dataset[25] was 25, their approach yielded to 59, to be comparable with human performance around 69. They also evaluated this among other datasets too and got BLEU-1 score improvements on Flickr30k[14], from 56 to 66, and on SBU[26], from 19 to 28. They also experimented this on newly released COCO dataset[15], and achieve a BLEU-4 of 27.7. So there evaluation showed accurate results and hence their methodology worked well.

Marko Smilevski et al. [3] proposed the approach wherein the deep architecture that they suggested uses a pre-trained CNN such as AlexNet[27] or VGG[28] to extract the features from the image that is captioned. The activation layer from the pre-trained network was then used as an input feature in the caption generator. The caption generator used a language model to model the captions in the form of a vanilla recurrent neural network[29] a long-short term memory recurrent neural network[30].

Ilija Lalkovski et al. [3] also covered the problems that are related to generating a description about a sequence of images or a video. A common solution they provided was a sequence to sequence modelling, where a pre-trained CNN[21] is used for feature extraction from the images (that are part of the sequence) and an RNN[22] is used to model the temporal behaviour of the sequence of image features.

In the work by Diana Gonzalez-Rico et al. [6] , their model extended the image description model which consists of an encoder-decoder architecture[32]. The encoder as an Convolutional Neural Network (CNN)[21] and the decoder as an Long Short-Term Memory (LSTM)[22] network . Here The image was passed through the encoder generating the image representation that was further used by the decoder to know the content of the image and generate the description word by word. They obtained competitive METEOR[33] scores in both the public and hidden test sets and performed well in the human evaluation also. Even their idea of bi-directional LSTM[34] was amazing. If Bi-directional LSTM[34] is used then results can be obtained in shorter amount of time as prediction coming from front end and back end and meeting in middle ease the prediction process in LSTM[34] . And hence it causes the faster and accurate results regardless of greedy or beam approach.

MD. Zakir Hossain et al. [7] specified Dense[35] which is a fully convolutional localization network architecture, which is composed of a convolutional network, a dense localization layer, and an LSTM language model[30]. The dense localiza- tion layer processes an image with a single, efficient forward pass, which implicitly predicts a set of region of interest in the image. Thereby, it requires no external region proposals unlike to Fast R-CNN[36] or a full network of Faster R-CNN[37]. The working principle of the localization layer they specified was related to the work of Faster R- CNN[37] .Descriptions of the entire visual scene were quite subjective The region-based description is known as dense captioning[35]. But there are some challenges in that, as regions are dense, one object may have multiple overlapping regions of interest. Moreover, it is very difficult to recognize each target region for all the visual concepts.

Licheng Yu a et al. [8] proposed an approach that makes use of the Visual Storytelling dataset and a model composed of three hierarchically-attentive Recurrent Neural Nets (RNNs)[22] to encode the album photos, select representative (summary) photos, and compose the story. Automatic and human evaluations showed that their model achieved better performance on selection, generation, and retrieval than base- lines.

Simao Herdade et al. [9] specified the way in which algorithms uses feature vectors extracted from the region pro- posals obtained from an object detector[38]. In this work they introduced the Object Relation Transformer[38], which ex- plicitly incorporates information about the spatial relationship between input detected objects through geometric attention. Quantitative and qualitative results demonstrate the importance of such geometric attention for image captioning, leading to improvements on all common captioning metrics on the MS- COCO dataset[16].

They proposed Transformer that encodes 2D position and size relationships between detected objects in images, building upon the bottom-up and top down image cap- tioning approach. Their results on the MS-COCO dataset[16] demonstrated that the Transformer does indeed benefit from incorporating spatial relationship information when comparing the relevant sub-metrics of the SPICE captioning metric[39].

They have also presented qualitative examples of how incorporating this information can yield captioning results demon- strating better spatial awareness. Currently their model only takes into account geometric information in the encoder phase. They also aimed to do this by explicitly decoding words with object bounding boxes. This leads to additional performance gains and improved interpretability of the model.

In the work by Shuang Liu et al. [10], they mainly de- scribed three image captioning methods using the deep neural networks: CNN-RNN based[21,22], CNN-CNN based[21] and Reinforcement-based framework[40]. They also specified the degree of matching between the caption sentence and the reference sentence to evaluate the generation results. The commonly used Evaluation metrics methods include BLEU [24], METEOR[33], ROUGE[41] ,CIDEr[42] and SPICE[39] measurement indicators. Among them, BLEU[24] and ME- TEOR[33] are derived from machine translation, ROUGE[41] is derived from text abstraction, and CIDEr[42] and SPICE[39] are specific indicators based on image captioning.

They checked the performance of all 5 indicators and the best results of the above three methods for five evaluation metrics were that the both the CNN-RNN[21,22] based and the Rein- forcement based methods can get the better performance than the CNN-CNN[21] based framework, which greatly improves the training speed without seriously affecting the accuracy. Besides all of them, the reinforcement framework gives best performance.

Several other authors have focused on performance and checking evaluation of all the methods and finding the best one. Some authors have also discussed their Future work in Bi-Directional LSTM[34] where in they supported Bi- Directional LSTM[34] for accurate and faster results but in implementation Bi-Directional LSTM[34] has not got much attention and there are hardly implementations done on these Bi-Directional LSTM[34] in the field of Captioning. Nvidia ® Corporation is also researching in this field.[46]

This Research area got an insight in several fields like: Self driving cars — Automatic driving is one of the biggest challenges and if we can properly caption the scene around the car, it can give a boost to the self driving system. Aid to the blind — We can create a product for the blind which will guide them travelling on the roads without the support of anyone else. We can do this by first converting the scene into text and then the text to voice. Both are now famous applications of Deep Learning. CCTV cameras are everywhere today, but along with viewing the world, if we can also generate relevant captions, then we can raise alarms as soon as there is some malicious activity going on somewhere. This could probably help reduce crimes and/or accidents. Automatic Captioning can help, make Google Image Search as good as Google Search, as then every image could be first converted into a caption and then search can be performed based on the caption.[11]

III. DATASETS AND EVALUATION METRICS

A number of datasets have been proposed for Visual Story Generation. These datasets also contain some additional an- notations associated with the image. The most common one for images in sequence is VIST i.e Visual story telling dataset also known as SIND i.e Sequential vision-to-language [11]. The first release of this dataset, SIND1 v.1, includes 81,743 unique photos in 20,211 sequences, aligned to both descriptive (caption) and story language. They have establish several strong baselines for the storytelling task, and motivate an automatic metric to benchmark progress. This Dataset strongly supports the potential to move artificial intelligence from basic understandings of typical visual scenes towards more and more human-like understanding of grounded event structure and subjective expression[12].

			
DII	A group of people that are sitting next to each other.	Adult male wearing sunglasses lying down on black pavement.	The sun is setting over the ocean and mountains.
SIS	Having a good time bonding and talking.	[M] got exhausted by the heat.	Sky illuminated with a brilliance of gold and orange hues.

Fig. 2. Example language difference between descriptions for images in isolation (DII) vs. stories for images in sequence (SIS)[12].

Figure-2 below clearly shows that how captions are affected when generated in isolation. It limits its narrative ability in single image caption generation . It can be viewed as a concept of narration for an example, “sitting next to each other” versus “having a good time”, or “sun is setting” versus “sky illuminated with a brilliance...”. The first descriptions capture image content that is literal and concrete; the second requires further inference about what a good time may look like, or what is special and worth sharing about a particular sunset[12]. This differs in sequential image caption generation because they are inter-related and also previous and current state decides outcome rather than only current state.

Flickr dataset[14]: The Flickr30k dataset has become a standard benchmark for sentence-based image description. This paper presents Flickr30k Entities, which augments the 158k captions from Flickr30k with 244k coreference chains, linking mentions of the same entities across different captions for the same image, and associating them with 276k manually annotated bounding boxes. Such annotations are essential for continued progress in automatic image description and grounded language understanding[14].

There are other versions of Flickr dataset like Flickr 8k , Flickr 10K etc many other human edited versions also . This dataset has a huge range of images and is widely used in captioning [14].

MS-COCO[15]: COCO is a large-scale object detection, seg- mentation, and captioning dataset[15].They have pre- sented a new dataset by gathering images of complex everyday scenes containing common objects in their natural context. Objects are labeled using per-instance segmentations to aid in precise object localization. Their dataset contains photos of 91 objects types that would be easily recognizable by a 4 year old. With a total of 2.5 million labeled instances in 328k images, the creation of their dataset drew upon extensive crowd worker involvement. They have also described how they annotated images. Because labelling 2.5 million objects is a very tedious and time consuming process. So there annotation task involved (a) labeling the categories present in the image, (b) locating and marking all instances of the labeled categories and (c) segmenting each object instance[16].

ImageNet[17]: The ImageNet Large Scale Visual Recogni- tion Challenge is a benchmark in object category classification and detection on hundreds of object categories and millions of images. The challenge has been run annually from 2010 to present, attracting participation from more than fifty institu- tions [17,18].

Fig. 3. Annotations[18]

This dataset is deeply annotated as seen in above figure-3. They have classified a bird also in flamingo, cock, ruffed grouse, quail, patridge etc. This deep annotations can help in creating interesting stories as more and more entities are involved and training is done through them so more deeper training . This would definitely affect in good evaluation if tested on test data. More entities also emphasizes describingspecialized form of any real world entity rather than leaving it under generalized form.

IV. EVALUATION

Evaluation can be performed on number of metrics like BLEU[19], METEOR[20], SPICE[39] etc but most prominent ones are BLUE and METEOR. BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text
[19] This metrics take hypothesis and reference as parameter for evaluation. Here the reference would be our generated output story and the hypothesis can be both the previously trained samples all together.
METEOR (Metric for Evaluation of Translation with Ex- plicit ORdering) is another metric for the evaluation[20].
Meteor metrics produce good correlation with human judge- ment at the sentence or segment level. This differs from the BLEU metric in that BLEU seeks correlation at the corpus level.[19,20].
This means that when hypothesis is kept as human judge- ment it scores good in meteor and when hypothesis is kept as corpus i.e our vocabulary of trained samples it scores good in BLUE metrics. The scores in BLEU metrics and METEOR metrics both differ due to selection of their hypothesis even though if reference is kept same.
Its calculation is also very easy . For an example ' The cat is going' is our reference and 'The cat is going to garden' is our Hypothesis, Here our reference i.e our output on which evaluation needs to be done is only matching with one trained sample from entire corpus that too not in entirety, hence according to score – Hypothesis of 6 words and reference of 4 words matches each other which gives out 4/6 = 0.67 Here score would always range between [0,1] . If it is 0 then extremely poor, if it is 1 then perfect, if it crosses 0.5 then above average and if it is below 0.5 then below average. But these metrics are highly dependent on training images i.e if we have trained images over cats and dogs and if our testing images are also of cats and dogs then scores would be good but if testing done in human based images then evaluation will be very poor . So evaluation metrics highly focuses on hypothesis and reference.

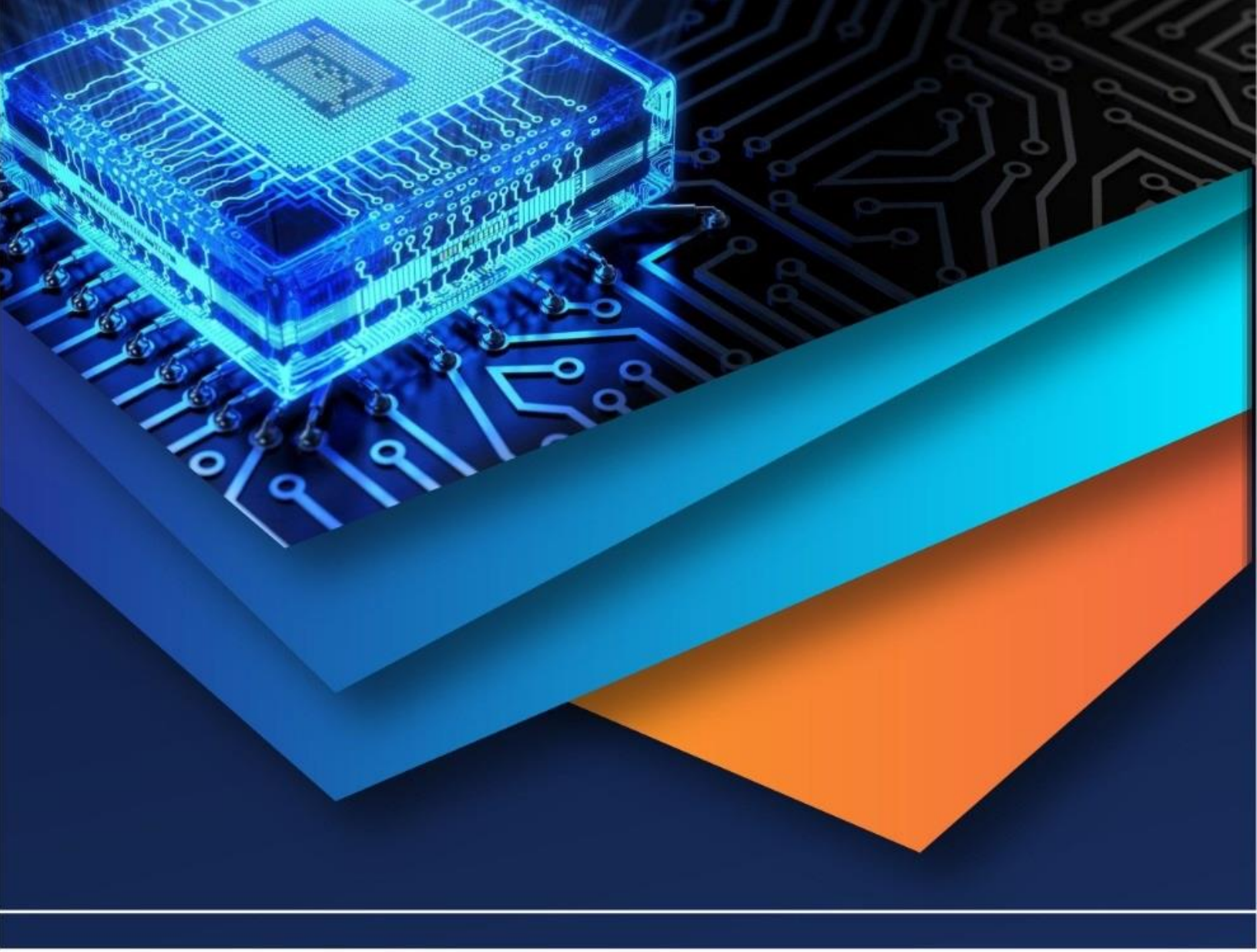
V. CONCLUSION

Visual Story Generation for images in sequence is a tech- nique that emphasizes the image understanding in textual form using NLP. This paper presents a comprehensive review on the ongoing research works done in the field of visual Story generation. The number of works in this fields are constantly increasing day by day due to the interesting features of this research area. Most of the work in this field has been done on Image Captioning. There are various new methodologies such as Bi-Directional LSTM, Using Generator Function to reduce load on LSTM etc. This paper also provides a comprehensive survey of various datasets that are utilized so far in the realization of Visual Story Generation. Various evaluation met- rics methods have also been discussed and differentiated. We believe that the ongoing and future work on these particular points will benefit the specific task of Visual Story Generation as well as the general objective of Image understanding in computer vision.

REFERENCES

[1] Jason Browniee. (2019). 'How to develop a deep learning photo caption generator from scratch', develop-a-deep-learning-caption-generation-model-in-python, 27 June. Available at:https://machinelearningmastery.com/develop-a-deep-learning-caption- generation-model-in-python/.
[2] IST dataset available at URL:https://visionandlanguage.net/VIST/.
[3] Marko Smilevski , Ilija Lalkovski and Gjorgji Madjarov ," Stories for Images-in-Sequence by using Visual and Narrative Components", arxiv:1805.05622.
[4] Oriol Vinyals , Alexander Toshev , Samy Bengio , Dumitru Erhan, "Show and Tell: A Neural Image Caption Generator", arxiv: 1411.4555.
[5] Jing Wang, Jianlong Fu, Jinhui Tang , Zechao Li ,Tao Mei (2018), "Show, Reward and Tell: Automatic Generation of Narrative Paragraph from Photo Stream by Adversarial Training", In Proceedings of the AAAI Conference on Artificial Intelligence .
[6] Kelvin Xu ,Jimmy Lei Ba ,Ryan Kiros ,Kyunghyun Cho ,Aaron Courville ,Ruslan Salakhutdinov ,Richard S. Zemel ,Yoshua Bengio ,"Show,Attend and Tell: Neural Image Caption Generation with Visual Attention",arxiv:1502.03044.
[7] MD. ZAKIR HOSSAIN, FERDOUS SOHEL, FAIRUZ SHIRATUD-DIN, HAMID LAGA, "A Comprehensive Survey of Deep Learning for Image Captioning", arxiv:1810.04020.
[8] Licheng Yu and Mohit Bansal and Tamara L. Berg, "Hierarchically- Attentive RNN for Album Summarization and Storytelling" ,arxiv:1708.02977.
[9] Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares , "Image Captioning: Transforming Objects into Words", arxiv:1906.05963.
[10] Liu, Shuang Bai, Liang Hu, Yanli Wang, Haoran. (2018). Image Cap- tioning Based on Deep Neural Networks. MATEC Web of Conferences. 232. 01052. 10.1051/matecconf/201823201052.
[11] Harshall Lamba. (2018). 'Image captioning with keras', image- captioning-with-keras-teaching-computers-to-describe-pictures, 4 November. Available at: https://towardsdatascience.com/image- captioning-with-keras-teaching-computers-to-describe-pictures- c88a46a311b8.
[12] IST dataset at availableat URL: https://visionandlanguage.net/VIST/dataset.html.
[13] Ting-Hao (Kenneth) Huang, , Francis Ferraro, Nasrin Mostafazadeh , Ishan Misra , Aishwarya Agrawal , Jacob Devlin , Ross Girshick , Xiaodong He , Pushmeet Kohli , Dhruv Batra , C. Lawrence Zitnick, Devi Parikh , Lucy Vanderwende , Michel Galley , Margaret Mitchell, "Visual Storytelling", arxiv:1604.03968..
[14] Flickr Dataset available at URL: https://www.kaggle.com/hsankesara/flickr- image-dataset.
[15] COCO dataset available at URL: https://cocodataset.org/home. [16]Tsung-Yi Lin , Michael Maire , Serge Belongie , Lubomir Bourdev, Ross Girshick , James Hays Pietro Perona , Deva Ramanan , C. Lawrence Zitnick , Piotr Dollar, "Microsoft COCO: Common Objects in Context", arXiv:1405.0312v3

[16] Imagenet Dataset available at URL:http://www.image-net.org/. [18]Olga Russakovsky, Jia Deng, Hao Su , Jonathan Krause , Sanjeev Satheesh , Sean Ma , Zhiheng Huang, Andrej Karpathy ,Aditya Khosla, Michael Bernstein , Alexander C. Berg , Li Fei-Fei." ImageNet Large Scale Visual Recognition Challenge". arXiv:1409.0575v3.
[17] BLEU metrics Information availableat URL:https://en.wikipedia.org/wiki/BLEU.
[18] METEOR metrics Information availableat URL:https://en.wikipedia.org/wiki/Meteoroid.
[19] Bengio, Y. Lecun, Yann. (1997). Convolutional Networks for Images, Speech, and Time-Series.
[20] Work by David Rumelhart Information available at various sites. [23]Qingzhong Wang and Antoni B. Chan , "Convolution decoders for image captioning ", arXiv:1805.09019v1.
[21] Papineni, Kishore Roukos, Salim Ward, Todd Zhu, Wei Jing. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. 10.3115/1073083.1073135.
[22] Kaggle based dataset Pascal available at URL: https://www.kaggle.com/huanghanchina/pascal-voc-2012.
[23] Large-scale Training of Shadow Detectors with Noisily-Annotated Shadow Examples, Vicente, T.F.Y., Hou, L., Yu, C.-P., Hoai, M., Samaras, D., Proceedings of European Conference on Computer Vision (ECCV), 2016.
[24] Krizhevsky, Alex Sutskever, Ilya Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Infor- mation Processing Systems. 25. 10.1145/3065386.
[25] Simonyan, Karen Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.
[26] F. ROSENBLATT , "THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN" , Psychological Review Vol. 65, No. 6, 19S8. [30]Hochreiter, Sepp Schmidhuber, J ürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735. [31]Junyoung Chung , Caglar Gulcehre , KyungHyun Cho , Yoshua Bengio ," Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling", arXiv:1412.3555v1.
[27] Tong Wang† Ping Chen† Kevin Amaral† Jipeng Qiang, "An Experimen- tal Study of LSTM Encoder-Decoder Model for Text Simplification", arXiv:1609.03663v1.
[28] Satanjeev Banerjee Alon Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" ,available at URL: https://www.aclweb.org/anthology/W05-0909.
[29] Schuster, Mike Paliwal, Kuldip. (1997). Bidirectional recurrent neural networks. Signal Processing, IEEE Transactions on. 45. 2673 - 2681. 10.1109/78.650093.
[30] Gelenbe, Erol Yin, Yonghua. (2017). Deep Learning with Dense Random Neural Networks. 10.1007/978-3-319-67792-71.
[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun," Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", arXiv:1506.01497v3.
[32] Ren, Shaoqing He, Kaiming Girshick, Ross Sun, Jian. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelli- gence. 39. 10.1109/TPAMI.2016.2577031.
[33] He, Sen Liao, Wentong Rezazadegan Tavakoli, Hamed Yang, Michael Ying Rosenhahn, Bodo Pugeault, Nicolas. (2020). Image Captioning through Image Transformer.
[34] Peter Anderson , Basura Fernando , Mark Johnson, Stephen Gould, " SPICE: Semantic Propositional Image Caption Evaluation" arXiv:1607.08822v1.
[35] Reinforcement learning information available athttp://www.cs.cmu.edu/ mmv/papers/03TR-advRL.pdf.
[36] Lin, Chin-Yew. (2004). ROUGE: A Package for Automatic Evaluation of summaries. Proceedings of the ACL Workshop: Text Summarization Braches Out 2004. 10.
[37] Ramakrishna Vedantam , C. Lawrence Zitnick , Devi Parikh, " CIDEr: Consensus-based Image Description Evaluation " , arXiv:1411.5726v2.
[38] Xin Wang , Wenhu Chen , Yuan-Fang Wang , William Yang Wang, "No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling" , arXiv:1804.09160v2.
[39] Taehyeong Kim, Min-Oh Heo , Seonil Son , Kyoung-Wha Park , Byoung-Tak Zhang, "GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation" , arXiv:1805.10973v3.
[40] Chao-Chun Hsu1, Zi-Yuan Chen2, Chi-Yang Hsu, Chih-Chia Li , Tzu- Yuan Lin, Ting-Hao (Kenneth) Huang3 Lun-Wei Ku," Knowledge- Enriched Visual Storytelling", arXiv:1912.01496v1.
[41] "Horus by NVIDIA – A life-changer for the blind " YouTube, uploaded by GiGadgets, 06 november 2016, https://www.youtube.com/watch?v=rLyF4XQLwr0.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY