International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 01 | Jan 2021 www.irjet.net p-ISSN: 2395-0072

# Image Caption Generator using Machine Learning

**Annagiri Sai Bhargav[1], Chandrashekara K V[2], Kartheek R[3], Sathyanarayana M[4], Suraj Kumar B P[5]**

*[1,2,3,4]Dept. of Computer Science, Sir M Visvesvaraya Institute of Technology, Bengaluru, India*
*[5]Assistant Professor, Dept. of Computer Science, Sir M Visvesvaraya Institute of Technology, Bengaluru, India*

---

**Abstract –** *As of late, with the fast improvement of innovation, the strategies that a machine can do is step by step expanding. One of these strategies is perceiving an image and giving us an outline of what's going on in that image in a language that is reasonable by people. This strategy joins the information on computer vision and natural language processing. Since artificial intelligence is the base information in acquiring the image captioning, numerous specialists are pulled in to it and has become an intriguing and challenging assignment. This paper presents a far reaching investigation of the cutting edge models and progression in exploration that has occurred in image captioning.*

***Key Words***: **Computer Vision, Natural Language Processing, Artificial Intelligence, Image Captioning, Technology.**

## 1. INTRODUCTION

Image caption generator is a mainstream research territory of computerized reasoning that manages picture understanding and a language depiction for that picture. To create an all-around framed sentence, we need the information on both syntactic and semantic comprehension of the language. Having the option to portray the substance of a picture utilizing precisely framed sentences is a difficult errand, yet have numerous significant applications like, it can help outwardly hindered individuals better comprehend the photos, in web-based media, picture ordering and so forth This errand is altogether harder in contrast with the picture characterization or article acknowledgment assignments that have been well-informed. The main undertaking is unquestionably having the option to make portrayal that can depict the items in the picture as well as the connection among at that point and the scene going on in that picture. This should be possible by understanding the climate of that image and the work done by the item.

Allow us to comprehend this momentarily utilizing a dataset model where two felines are dozing on a seat.

**Fig -1**: Dataset Example

At the point when a human investigate this image, the individual can say 'Two felines are dozing on a seat' or 'Two felines are dozing'. It appears to be simple for us people to take a gander at a picture and depict it fittingly. In any case, a machine or a PC doesn't have that capacity. So we need to give that capacity to machines utilizing a technique either by utilizing convolutional neural organizations or by utilizing profound learning and any dataset like flicker8k.

In this paper, we have investigated different works done in the space of picture inscribing. We have recorded down our discoveries in a consecutive request under the Writing study segment.

## 2. LITERATURE SURVEY

(Oriol Vinyals, Alexander Toshev, Samy Bengio and Dumitru Erhan [1]) have introduced in their paper about a generative model dependent on a profound repetitive engineering that joins ongoing advances in PC vision and machine interpretation. This can be utilized to produce sentences depicting a picture. This model is prepared to boost the probability of the objective portrayal sentence given the preparation picture. This model gains exclusively from picture depictions. The present status of the workmanship BLEU score on the Pascal dataset is 25, this methodology gets 59, and the human presentation is around 69.

(Philip Kinghorn, Li Zhang, and Ling Shao [2]) have referenced in their paper that comprehensive strategies may lose subtleties identifying with significant angles in a scene. They proposed novel locale based profound

learning design for picture portrayal age. This model uses territorial item identifier, encoder – decoder language generator inserted with two repetitive neural organizations (RNN), to deliver refined and itemized depictions of the given picture. Above all the proposed framework utilizes nearby based methodology that can additionally improve the current comprehensive techniques. This model was assessed with the IAPR TC-12 dataset.

(Parth Shah, Vishvajit Bakrola and Supriya Pati [3]) this paper examine about different accessible models for picture inscribing task. They have investigated about how the progression in the errand of machine interpretation and item acknowledgment has enormously improved the presentation of picture inscribing model as of late. Notwithstanding this they likewise talked about how this model can be executed and assessed the exhibition of model utilizing assessment lattices. They inferred that we can consolidate late progression in picture naming and programmed machine interpretation into a start to finish half and half neural organization framework so the framework is proficient to freely see a picture and produce a depiction in regular language with better precision and effortlessness.

(Recovery Alahmadi, Chung Hyuk park and James Hahn [4]) has examined in this paper about how the picture subtitling is standing out enough to be noticed from the man-made reasoning local area. Numerous current models utilizes the encoder-decoder machine interpretation to naturally create inscriptions for pictures. For this the vast majority of the works utilized convolutional neural organization (CNN) as a picture encoder and intermittent neural organization (RNN) as a decoder to create subtitle. They proposed a model which utilizes RNN as a picture encoder that utilizes encoder-decoder machine interpretation model so the contribution to the model is a bunch of pictures that speaks to the items in the given picture. The request for the items depends on the request in the subtitles. They utilized flickr30k dataset.

(Ying Hua Tan and Chee Seng Chan [5]) states that the most work done in the field of picture inscribing as unadulterated consecutive information. Regular language, anyway have a fleeting progressive system structure with complex conditions between every aftereffect. In this paper they proposed an expression based picture subtitling model that utilizes progressive long transient memory (phi-LSTM) design to produce the picture portrayal. Not at all like any convolutional arrangements which utilizes unadulterated consecutive way to produce inscription, phi-LSTM unravels picture subtitle from expression to sentence. It comprises an expression decoder that unravels the thing expression of a variable length and a condensed sentence decoder which interprets the truncated type of the picture portrayal. Complete picture subtitle is created by joining the produced phrases with sentence during the induction step. This model shows better outcomes on flickr8k, flickr30k and ms-coco datasets.

(Eric Ke Wang, Xun Zhang, Fan Wang, Tsu-Yang Wu and Chien-Ming Chen [6]) proposed a paper that utilizes multilayer thick consideration model for picture subtitle. To remove picture highlights as the coding layer they utilized a quicker repetitive convolutional neural organization and long transient memory to unravel the multilayer thick consideration model. Inclination streamlining in fortification learning is utilized to enhance the model boundaries. By utilizing thick consideration measures in the encoding layer we can adequately stay away from the impedance of non-notable data and can yield the portrayal specifically for the deciphering part.

(Ansar Hani, Najiba Tagougui and Monji Kherallah [7]) The errand of picture inscribing includes creating an applicable portrayal of a picture in typical language and can be performed utilizing PC vision and normal language handling and AI techniques. They create a model that can play out the picture subtitling utilizing blend of convolution neural organizations and repetitive neural organizations which can perform extraction of highlights and to produce text from these highlights individually. They joined the consideration component while producing subtitles and assessed utilizing MSCOCO information base.

(Daouda Sow, Zengchang Qin, Mouhamed Niasse and Tao Wan [8]) There have been such countless advancements in PC vision and regular language handling that have made unpredictable and troublesome assignments in understanding semantics, for example, portrayal age from common pictures. In this errand, utilizing convolutional neural organization and intermittent neural organization as picture encoder and decoder separately, we had the option to accomplish better execution. In this new model, they manage the decoder utilizing a successive directing organization and we furnish the entire model itself with extra direction utilizing long transient memory. The model can be prepared in a start to finish way by utilizing picture/enlightening sets. They approve their exploration by leading investigations on ms coco inscriptions, a benchmark dataset. The proposed new model accomplished some huge improvement over probably the freshest profound learning models.

(Adele Puscasiu, Alexandra Fanca, Dan-credit Gota, Honoriu Valean [9]) This paper presents a composite model comprising of profound convolutional network and repetitive neural organization. The profound convolutional network is utilized for include extraction that utilizes move learning and repetitive neural organization is utilized for building the depiction. The examination on understanding this model depends on AI calculations and systems for both common language

handling and picture preparing. The convenience is utilizing existing bundles in actualizing AI calculations. The execution of the calculation is it takes a picture and gives a wide portrayal about the picture. In the wake of assessing the necessities, an investigation was led for acclimation with the area and potential models. Then, research was made for any accessible innovations that can help construct the application was made and afterward the plan, usage and approval on the proposed model started. For backend usage, they utilized keras and tensorflow. Finally a prepared model that portrays a picture utilizing regular language is acquired.

(Himanshu Sharma, Manmohan Agrahari, Sujeet Kumar Singh, Mohd Firoj and Ravi Kumar Mishra [10]) They expressed that the cycle of picture inscribing needs to distinguish objects in picture, activities their relationship and some striking component that might be absent in the picture. The subsequent stage after ID is to create a generally typical and brief portrayal for the picture which ought to be grammatically and semantically right. It is hard for a machine to copy human mind capacity however analysts in this field demonstrated an incredible improvement. To deal with such issues profound learning procedures like CNN and long transient memory is sufficient. This model can be utilized in numerous smart control frameworks and Web of Things (IOT) based gadgets. They introduced an alternate way to deal with picture inscribing, for example, recovery based, profound learning based and layout based strategies.

(Chetan Amritkar and Vaishali Jabade [11]) The substance of a picture are created consequently which included Characteristic Language Handling and PC vision in artificial intelligence. The neural model which relies upon PC vision and machine interpretation that is regenerative is made. This model creates a characteristic sentence which can portray a picture. It comprises of Convolutional Neural Organization and Intermittent neural organizations. This model works in a manner that if picture is given as an information, we get a characteristic language subtitle that portrays the picture. The model is tried on various datasets for its precision, perfection and order of language that model gains from the picture depictions.

(I.Hrga, M.Ivasic-kos [12]) They gave a diagram of issues and ongoing picture inscribing research with an exceptional interest on models which utilizes profound encode-decoder design. They additionally expressed that the picture subtitling is expanding its methodology progressively applications and furthermore that it has seen fast advancement from starting layout based models to the ones that depend on profound neural organizations. They had examined the favorable circumstances and weaknesses of various methodologies alongside evaluating probably the most regularly utilized datasets.

**3. End**

Huge endeavors have been made to create cutting edge models for picture inscribing. Various calculations have arisen in the field of picture preparing, AI and profound learning procedures to upgrade the subtitling systems. The capacity of convolutional neural organizations and repetitive neural organizations made it conceivable to acquire better outcomes in producing portrayals and highlight extractions. Many component extraction models are utilized in numerous applications like aiding outwardly debilitated individuals and so on In this overview paper we have introduced a thorough investigation of best in class models and progression in examination that has occurred in utilization of picture subtitling. There are many intriguing difficulties like identifying scenes in a video, following item in a picture or a video keeps this area open for broad exploration.

**REFERENCES**

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio and Dumitru Erhan, ”Show and tell: A neural image caption generator”, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, doi : 10.1109/CVPR.2015.7298935

[2] Philip Kinghorn, Li Zhang, and Ling Shao, “A region-based image caption generator with refined descriptions”, Neurocomputing, Volume 272,Pages 416-424, doi: 10.1016/j.neucom.2017.07.014

[3] Parth Shah, Vishvajit Bakrola and Supriya Pati,” Image Captioning using Deep Neural Architectures”, 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India, doi : 10.1109/ICIIECS.2017.8276124

[4] Rehab Alahmadi, Chung Hyuk park and James Hahn, ” Sequence to sequence Image Caption Generator”, Eleventh International Conference on Machine Vision (ICMV 2018), 2018, Munich, Germany, doi: 10.1117/12.2523174

[5] Ying Hua Tan and Chee Seng Chan, “Phrase based image caption generator with hierarchical LSTM network”, Neurocomputing, Volume 333, Pages 86-100, doi : 10.1016/j.neucom.2018.12.026

[6] Eric Ke Wang, Xun Zhang, Fan Wang, Tsu-Yang Wu and Chien-Ming Chen, “Multilayer Dense Attention Model for Image Caption”, IEEE Access ( Volume: 7), Page(s): 66358 – 66368, doi: 10.1109/ACCESS.2019.2917771

[7] Ansar Hani, Najiba Tagougui and Monji Kherallah, “Image Caption Generation Using A Deep

Architecture", 2019 International Arab Conference on Information Technology (ACIT), Al Ain, United Arab Emirates, doi: 10.1109/ACIT47987.2019.8990998

[8] Daouda Sow, Zengchang Qin, Mouhamed Niasse and Tao Wan, "A Sequential Guiding Network with Attention for Image Captioning", ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, doi: 10.1109/ICASSP.2019.8682505

[9] Adele Puscasiu, Alexandra Fanca, Dan-loan Gota, Honoriu Valean, "Automated image captioning", 2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania, doi: 10.1109/AQTR49680.2020.9129930

[10] Himanshu Sharma, Manmohan Agrahari, Sujeet Kumar Singh, Mohd Firoj and Ravi Kumar Mishra, "Image Captioning: A Comprehensive Survey", 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), Mathura, Uttar Pradesh, India, doi: 10.1109/PARC49193.2020.236619

[11] Chetan Amritkar and Vaishali Jabade, "Image Caption Generation Using Deep Learning Technique", 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, doi: 10.1109/ICCUBEA.2018.8697360

[12] I.Hrga, M.Ivasic-kos, "Deep Image Captioning: An Overview", 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, doi: 10.23919/MIPRO.2019.8756821