



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJIRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.569



9940 572 462



6381 907 438



ijirset@gmail.com



www.ijirset.com

Dynamic Image Description Generation using Hybrid Deep Learning Framework

Gitanjali S.Patil

PG Student, Dept of Computer Engineering, SSVPS's B.S. Deore College of Engineering, Dhule Maharashtra, India

ABSTRACT: With the fast growth of artificial intelligence in recent years, imagetag has progressively drawn the attention of many artificial intelligence researchers and has become a fascinating and difficult problem. Scene comprehension, which combines the expertise of computer vision with natural language processing, includes picture tag, which automatically generates natural language descriptions based on the information seen in an image. The use of picture tags is widespread and important, as shown by the realization of human-computer interaction. The attention mechanism, which plays an essential role in computer vision and has lately been extensively utilized in picture tag generation tasks, is the subject of this article, which outlines the relevant techniques and concentrates on it. Additionally, the benefits and drawbacks of these techniques are addressed, as well as the most frequently utilized datasets and assessment criteria in this area. Traditional approaches cannot handle the complexity and difficulties of image tagging as well as deep learning-based approaches. In this paper we proposed deeplearning framework for tag generation on complex objects. First, we go through several current image tagging approaches, with an emphasis on deep-learning-based approaches and how they are used for pre-processing. Then CNN module implementation for feature extraction and tag generation in dense layer. The experimental analysis we describes how proposed system provides better results than existing systems.

KEYWORDS: Convolutional Neural Network, Long short term memory, image tag generation, Natural Language processing

INTRODUCTION

Convolutional neural-nets i.e., CNN are usually pre trained on a particular picture dataset, which uproots visual attributes from those pictures. This helps the defined model to learn on new data in a much swift way. Elaborating an image in a text format with genuine grammatical syntax is still a demanding problem in the field of computer vision and natural language processing. Comprehending a scene, which combines the expertise of computer vision with NLP, involves image tag, which automatically generates ethical English language descriptions based on the information seen in a frames. Tagging images is a long process if done by people physically, it also contributes errors generated by humans which is very significant if the generated description is used in sensitive applications like medical or space research to name few. From computer vision point of view the content within the images is very much valuable. They help a machine understand and perform accordingly. Image tagging has various applications such as recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications. Through various ways of deep learning solution to the process of image annotation has been achieved at a greater level and has contributed to various fields. Through deep learning people have come up with various innovative ideas to tackle this application. From those results it has been demonstrated that deep learning models are able to achieve optimum results in the field of tag generation problems. To generate quality level image descriptions, it is important to understand not only what the objects within the image are but also the relationship that exists between them. Today the encoder – decoder models are considered one of the state of the art models for image tagging [4].

In an object, a lot of information is saved. On social networking sites, including celestial objects, enormous image data can be generated each day, but this is an up with the fast thing. It takes time to annotate photographs of humanity, and the likelihood of error is greater. Deep learning models are used to accurately compile such images, thereby eliminating the manual corrections done.[16] This would significantly reduce human failure and also the efforts by removing any need for

human involvement. The development of image annotations has numerous real benefits, varying from supporting the mentally challenged to helping the automated, cost-saving marking of images shared online every day, guidelines for processing software, useful for smart devices, image encoding, visually disabled people, social networking sites, and so many other natural literature.

Models like Convolutional Neural Network have played an important role in this picture. Here, we try to show and highlight different methods for image feature extraction and how it will be used for tag creation. There has been a lot of study in the field of data science for advancing image tag generation model. Natural language processing plays an important role for creating a description, which is to the mark and has semantics to go with. Recurrent Neural Network (RNN) is the preferred network in description creation as RNN is basically used for sequence generation. Researchers have put in lot of effort in creating enormous magnitude datasets. Some of the famous datasets are the MSCOCO which is been provided by Microsoft. Other well-known and benchmark datasets are the Flickr 8K, Flickr 30K, PASCAL and some others [5]. Through this paper we have highlighted successful approaches that focus on deep learning to generate image descriptions. They have proved to be successful and have evolved with time. Evaluation of the overall model with the description generated is evaluated using evaluation metrics like the CIDEr, BLEU, METEOR and other metrics.

II. LITERATURE SURVEY

Retrieval-based image tagging is one of the most popular techniques used in early work. Retrieval-based techniques generate a tag for a query image by obtaining one or more sentences from a pre-specified sentence pool. The produced tag may be a statement that already exists or one that is made out of the recovered sentences. Let's start by looking at the line of study that utilises recovered phrases as image tags.

Semantics implementation [6] is supposed to design a clean way of injecting sentiment into the current image tagging system. Novel objects must be included for the expansion of scenarios. There have been several insights on why this is still an open issue. First of all, usually models are built on very specific datasets, which do not cover all possible scenarios and are not applicable in describing diverse environment. The same with vocabulary as it has a limited number of words and their combinations. Second, models are usually thought to perform on one specific task, while humans are able to work on many tasks simultaneously

LSTM (long-short-term memory) was developed from RNN, with the intention to work with sequential data. It is now considered as the most popular method for image tagging due to its effectiveness in memorizing long term dependencies through a memory cell. Undoubtedly this requires a lot of storage and is complex to build and maintain. There have been intentions to replace it with CNN [7], but as we can see from the number of times this method is used in most of the articles found during this SLR (68 of 78), scientists always come back to LSTM. LSTM works by generating a tag by making one word at every time step conditioned on a context vector, together with the previous hidden state and the earlier generated words.

Computational speed not only depends on the feature detection model, but also on the size of the vocabulary—each new word added consumes more time. Just recently [8] scientists have tried to solve the image tagging task by resizing the vocabulary dictionary. Usually the vocabulary size might vary from 10,000 to 40,000 words, while their model relies on 258 words. The decrease is quite sharp—reduced by 39 times if compared to 10,000, but the results are high, with some space for improvements.

According to [89] recently confirmed that the better approach is still top-down attention mechanisms as the results from experiments with humans and with machines showed similar results. In the top down model, the process starts from a given image as input and then converts it into words. Moreover, a new multi-modal dataset is created with the highest number of new instances from human fixations and scene descriptions.

Most of the works are evaluated on Flickr30k [9] and MSCOCO [10] datasets. Both datasets are rich in the number of images and each image has five tags assigned which makes it very suitable to train and test the models. It is of course necessary to continuously compare models with the same datasets in order to check the performance, however, they are very limited in the object classes and scenarios presented. The need of new datasets has always been an open question in image tagging.

III. PROPOSED METHODOLOGY

Image Tag Methodologies

The encoder part receives the image as input and generates a feature vector of high dimensions. The image goes through the layers of the Encoder model that converts the image to feature vector. The system does not understand anything but 0's and 1's and thus the image is converted to 0s and 1s. The image vector created is the form of binary format. The encoder is a group of convolutions that uses filters and uses pooling to help the system to focus on certain regions of the image. The encoded features from the encoder are given as input to the decoder. The decoder takes this high dimension features and generates a semantic segmentation mask. It is a process of recurrent units where it predicts an output at each stage. It accepts a hidden state and produces an output and its own hidden state. Various activation functions are used in the encoder-decoder model as per the application of the user.

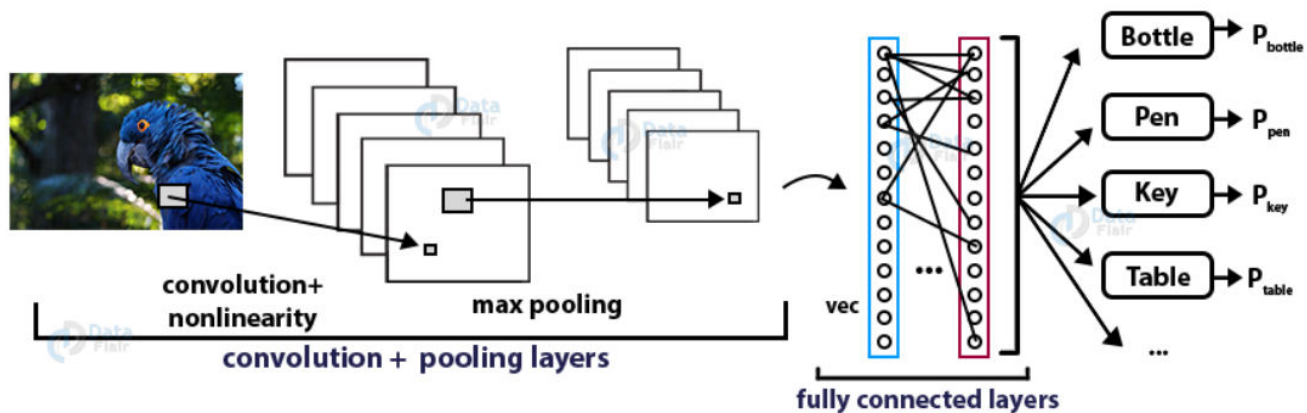


Figure 1: basic framework of image tag generation

Encoders:

Convolutional and Recurrent networks are the head and tail of the encoder decoder model. The convolution neural network works on the two dimensional image data and can be used in other dimensional data too. The model is used in Feature extraction for images. It has two versions viz. the VGG16 and VGG19, each of these models have 16 and 19 layers. These layers are stacked on top of each other. The architecture mainly consists of the convolution which then passes through the relu activation and then the pooling, at last flatten the vector value which is passed to the decoders. VGG faces issues with gradient decent as you go deeper in the model. ResNets are like successors of VGG. They are more deeper and better in their applications. ResNets have majorly these versions viz. ResNet 34,50,101,152. ResNets have proven to be useful and overcome the issue of vanishing gradient. The technique of skip connection is convenient it allows an alternate path for the gradient that helps the network from the gradient gradually approaching a very small value. Skip connections in deep architectures, in the computer program, skip a level and feed the performance about one layer also as layer known. The framework of Inception is planned in such a manner that it functions well, even when under strict machine and storage time restraints. It uses a significantly lower set of variables and gives our text recognition tasks a recommended efficiency as seen in the outcomes [5]

Decoders:

Recurrent neural networks like the Long Short Term Memory (LSTM) and GRU (Gated Recurrent Unit) are used to decode the vectors from the encoder. Pattern generation is implemented through the LSTM and GRU models. They have been widely used in speech recognition, NLP, and other areas. From the point of generating cinematic intervals, the Long Short Term Memory model is developed, consisting of an internal mechanism called gates that regulates the information flow. The gates decide what data to hold and to discard, which provides the benefit of passing the necessary data to predict the sequence chain. This processes information that passes on data as it transmits forward. It comprised of three windows, namely the gateway of input, gate of output and gateway of forget. The GRU is the latest Recurrent Neural Networks technology and is very close to the LSTM. The Gated Recurrent Device has got rid of both the cell state and is transmitting

information using the secret state. There are also actually two ports, an include some and an upgrade entrance. These gates control the flow of information within the model. These gates determine what data to maintain which information to discard. Attention models are focused on two areas viz, soft attention and the hard attention. Both these models have their pros. In case of soft attention, the entire image is considered but certain areas are more focused or attended to more. Whereas the hard attention covers a subset of the image and focuses that particular part.[2] This model requires comparatively less memory and computation power as it covers a subset of the image, whereas the soft attention model requires more memory and computational power. As the decoder will use the previous hidden state, previously generated word and a context vector. The attention model uses the context vector which is formulated by the convolution neural network. The weights calculated using the attention function can be done using the soft or hard attention mechanism. Attention can be broadly interpreted as a vector of importance weights. The visualization of the attention weights clearly demonstrates which regions of the image the model is paying attention to so as to output a certain word. It eliminates the vanishing gradient problem, as they provide direct connections between the encoder states and the decoder.

Generative Adversarial Networks:

These networks mainly known as GANs have given a new vision in neural networks world, GANs being originally found in 2014 have come a long way with wide usage in many applications. It consists majorly of two major components the Generator and the Discriminator. The Generator's work is to generate a synthetic data with some noise added in such way that it resembles the original distributed data. The noise is data from a random distribution. The job of Discriminator model is to identify if the generated output from the Generator is real or fake i.e., applying binary classification. In image tag the sentences generated must be more human like thus using it to define the image more human like. The GANs are provided with a truth value which the discriminator uses to classify the results thus implying the quality of the output. Its usage is also important in applications where we need more data to train the deeper models i.e., the data augmentation.

In probability theory, the core concept of GAN arises from the Statistical distribution [1]. Two game players are assumed: one generator and one clustering algorithm. The purpose of the generator is to learn how real data is transmitted, while the discriminator attempts to correctly decide if the input data is from the actual data or from the generator. To win the game, the two players need to constantly refine themselves to enhance the skill of the community and the ability to discriminate, respectively. The goal of the process of optimization is to find a Nash compromise between the various individuals.

Figure 2 displays the GAN computing procedure or composition. As the generator as well as the discriminator, any separable method can be used. Here, to describe the loss function and the generator, we use distinguishable functions D and G and their contributions are actual data x and probability distributions z , respectively. $G(z)$ describes the sample created by G and obeys the real data distribution p_{data} . If the discriminator D 's input is from the actual data x , D should identify it as valid and mark it as 1. If the output is from $G(z)$, it should be categorized as false by D and labeled as 0. The purpose of D is to ensure that the data source is properly classified, whereas the purpose of G is to ensure that the output of the generated data $G(z)$ on D (i.e. $D(G(z))$) is inconsistent with the growth of real data x on D (i.e. $D(x)$). The method of adversarial optimization steadily improves the efficiency of D and G . Eventually, when D 's capacity to differentiate has been enhanced to a high degree but cannot accurately discriminate against the data source, it is assumed that the G creator has produced the representation of real data.

The GANs have evolved since with different approaches being used the paper describes some of them.

Conditional GAN: CGANs perform better as the generated instance is labelled by the user and would allow to specify to which class it belongs. CGANs can build a model that may generate man or woman image that are imaginary and belong to one of these classes. From computer vision point it is one of the major advances, it uses self-attention layer in both generator and discriminator. To control the gradient, spectral normalization is applied on both the generator and discriminator models. Most image tagging methods include picture feature extraction, object recognition, and tag synthesis components, as described in the introduction. These models boost performance by analyzing the whole picture or selected picture areas. Those characteristics, on the other hand, are irrelevant for verbs, which have no significant relationship with visual characteristics. If a model is trained to produce the word "walk" from an image of a person walking with open legs, it will likewise produce the word "walk" from an image of a person swimming with similarly open legs. Despite the fact that the proper answer is the word "swim," this is the case. Another example is that a model may find it challenging to produce the words "standing" or "performing gymnastics" from a picture of a person standing inside and performing gymnastics with their hands rose. As a result, in this situation, a standard tagging algorithm is unable to create a verb and instead produces

the tag "person in a room." As a result, a model [4] with a motion estimation component has been suggested. When learning the relationship between picture characteristics and verbs is challenging, this model can learn the relationship between motion features and verbs. However, as mentioned in the introduction, background motion elements are used. This flaw may result in two significant problems. To begin, tags mostly explain the contents of the image's objects. As a result, backdrop characteristics aren't as important as they formerly were. When more characteristics in the backdrop are estimated than in the object regions, the tagging model includes more background characteristics, which may have a detrimental effect on tag creation. Second, the motion estimate accuracy is not always great. There are some problems in several of the motion features. As a result, utilising all of the motion characteristics may lead to additional chances to utilise the incorrect features, lowering tag production accuracy. To address this problem, we only utilise motion characteristics in the image's object area. In the first instance, the tag model may analyse key picture areas for picture tagging explicitly. In the latter instance, the likelihood of the tag model using the incorrect characteristics is reduced.

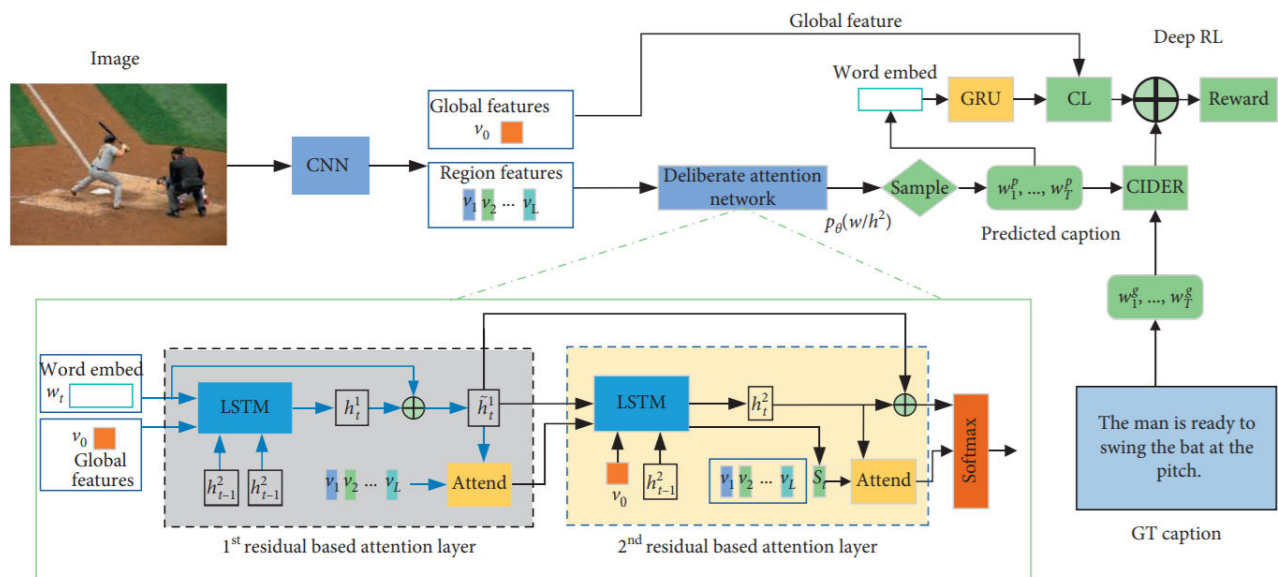


Figure 2: Proposed system architecture

The above Figure 3 depicts the proposed system architecture for tag generation using deep learning model. In below section we describe an each phase of system to generate the final outcome of validation set.

Preprocessing and Normalization: The data collection has done from synthetic repository called MSCOCO. The data might be imbalance with dimension thus it needs to balance before proceeding to CNN. In preprocessing we define fix set of dimensions of image (300*300) and eliminate those objects which are not readable for misclassified. The cross validation has also done in this similar phase for training and testing. (e.g. 5 fold, 10 fold, 15 fold etc.). In the dataset each image having tag as ground truth tag, and that used for analysis. Due to the large size of the dataset, data processing was done on the host computer and the results were saved as pickle files before being sent to the cloud for model training.

CNN: The module has used for CNN, to extract the features and selection as well. The below code snippet demonstrates the building of CNN with . The global image feature vectors for the frames are retrieved from Keras applications using the final fully connected layer of the CNN. Before extraction, the pictures are compressed to 300x300 arrays and the pixels are normalized to a scale of 0 to 255 to suit the model input. A pretrained model is utilized for the picture scene features. To match the input shape of the model, the images are compressed to 224x224 and normalized on a scale of 0 to 255. Both networks output feature vectors are saved as pickle files for simple uploading for validation testing.

GRU: attention model basically work for selection of features that received from CNN. The selection features eliminate the redundant and non-essential features. The attention model has divide into four different phases these are describes in below, A recurrent GRU can extract a lot of information from a picture, both temporally and spatially. However, semantic information is required to enhance the features obtained from the encoder in order to produce semantically correct sentences. The dense layer classification method is utilized in our approach to generate semantic vectors, and a semantic compositional network is used as the decoder. Furthermore, a dual learning method is employed to save the semantic vector's information throughout training, allowing the semantic information in forward flow to be effectively used. In GRU we applied the greedy approach for tag generation in both training testing. In both module generated tag has evaluated with ground truth sentence for validation and generate the 4 BELU scores and cosine similarity weight for confusion matrix. In result section we depicts the various ground text features extraction methods has used for generate the BELU score and similarity weight.

IV. RESULTS AND DISCUSSIONS

The datasets used are benchmark datasets that are used to assess the system's performance. Around 1000 images are used for testing and validation in both Flickr datasets. Transfer learning can be used to implement these models. Models are trained on high-configuration CPUs that can be found online as well as those on the system. The training period varies depending on the CPU; for such applications, systems with larger Gigabyte CPUs, such as 4 GB, are desirable, and software such as Python with IDEs such as Jupyter Notebook and other python libraries is necessary.

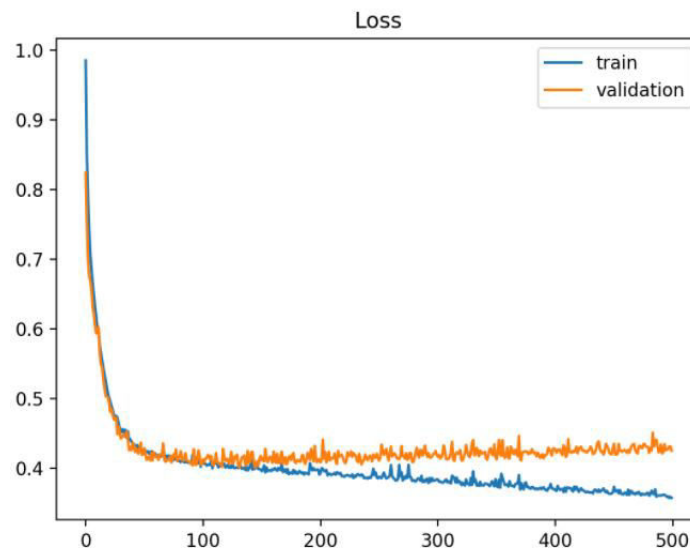


Figure3: Model performance on training

The above Figure 3 describes training and testing model performance with proposed on MSCOCO dataset. Almost 9000 images validated in training set and 300 for testing. When system deals with high epoch values it reduces the overall loss in training while it slightly increases or constant for validation.

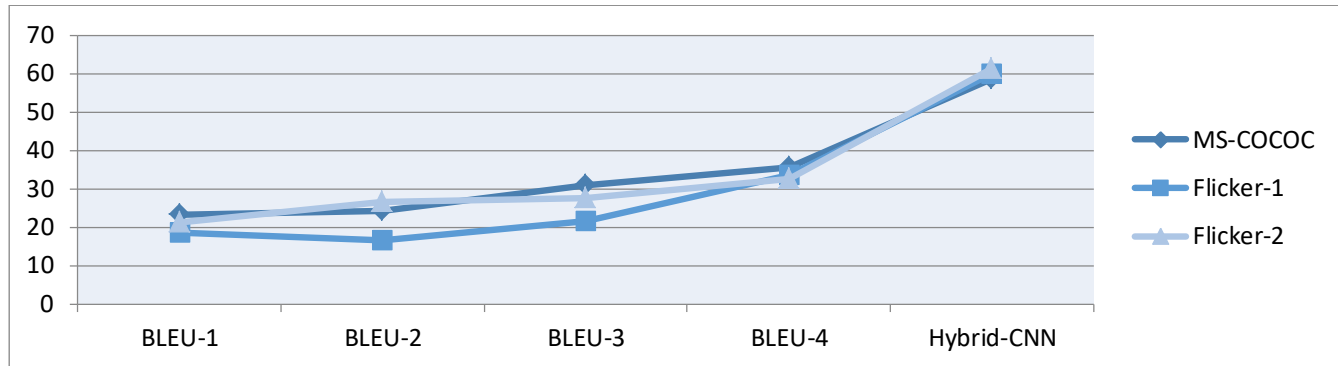


Figure 4 : Model performance testing with various feature selection techniques

The Figure 4 describes the tag generation in validation phase for two different images. The generated tag has evaluated with both ground truth sentence and evaluated with 4 BELU functions and Bigrams methods with cosine similarity algorithm. According to both results we conclude the default cosine similarity produces higher accuracy over the other methods.

V.CONCLUSION

In this research, we have compiled all aspects of the image tag generation task, discussed the model framework proposed in recent years to solve the description task, focused on the algorithmic essence of different attention mechanisms, and summarized how the attention mechanism is applied. We summarize the large datasets and evaluation criteria commonly used in practice. We show how deep learning techniques are used to generate image tags in this study. We employed CNN models with GRU to obtain an accurate description of the image. The ideas of the generalized Encoder – Decoder model have been discussed. Its application, as well as the attention model and how it was applied in various techniques. We also discussed several datasets that can be used to evaluate a system by using them as benchmark datasets. Finally, we discussed assessment metrics and how to use them to evaluate our model. This research focused on the image tagging task, which is a fundamental problem in artificial intelligence. For extraction, the CNN module was employed, and GRU was used to create tags for the images. To obtain the blue score for the complete testing dataset and demonstrate the performance of the proposed system, several cosine similarity and N-Gram feature extraction algorithms were applied.

REFERENCES

- [1] Vinyals O., Toshev A., Bengio S., et al.: Show and tell: A neural image tag generator. Proceedings of the International Conference on Computer Vision and Pattern Recognition.
- [2] Jinfei Zhou, Yaping Zhu, Hong Pan.: Image tag based on Visual Attention Mechanism, Proceedings of the 2019 International Conference on Image, Video and Signal Processing
- [3] Kishore papineni, SalimRoukos, Todd Ward, Wei-Jing Zhu.: BLEU: a Method for Automatic Evaluation of Machine Translation Kishore, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics
- [4] ViktorAtliha ,DmitrijSe'sok.: Comparison of VGG and ResNet used as Encoders for Image Tagging , 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)
- [5] ShreyasiCharu, S.P.Mishra, Tapan Gandhi.: Vision to Language: Tagging Images using Deep Learning, 2020 International Conference on Artificial Intelligence and Signal Processing (AISP).
- [6] You, Q.; Jin, H.; Luo, J. Image tagging at will: A versatile scheme for effectively injecting sentiments into image descriptions. arXiv 2018, arXiv:1801.10121.
- [7] Aneja, J.; Deshpande, A.; Schwing, A. Convolutional image tagging . In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5561–5570.
- [8] Mishra, A.; Liwicki, M. Using deep object features for image descriptions. arXiv 2019, arXiv:1902.09969



- [9] Plummer, B.A.; Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2641–2649.
- [10] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755



**Impact Factor:
7.569**



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 **9940 572 462**  **6381 907 438**  **ijirset@gmail.com**



www.ijirset.com

Scan to save the contact details