Image Description using Encoder and Decoder LSTM Methods: Some Issues

Nirmala, Gopalkrishna Joshi, P S Hiremath

***Abstract:** Description of images has an important role in image mining. The description of images provides an insight into the location, its surroundings and other information related to it. Different procedures of describing the images exist in literature. However, a well trained description of images is still a tedious task to achieve. Several researchers have come up with solutions to this problem using various techniques. Herein, the concept of LSTM is used in generating a trained description of images. The said process is achieved through encoders and decoders. Encoders use techniques of maxpooling and convolution, while the decoders use the concept of recurrent neural networks. The combined architecture of encoders and decoders result in trained classifiers, which enable reliable description of images. The working has been implemented by considering a sample image. It has been found that slight variations with regard to accuracy, naturalness, missing concepts, deficiency of sufficient semantics and incomplete description of image still exist. Hence, it can be inferred that, with reasonable amount of enhancement in the technique and using the techniques of natural language processing, more accuracy in image descriptions could be achieved.*

***Keywords:** Convolution Neural Network, Data Processing, Decoder, Encoder, LSTM, Maxpooling*

I. INTRODUCTION

Description of images can be done in several ways and captioning is one such simple technique. Here, the contents of image are described using natural language. The current work deals with the same by making use of specific architectures of encoders and decoders [1]. Architecture of encoders involve extraction of images using max pooling techniques and convolution layers, using which, meaningful features can be extracted. Architecture of decoders describes the recurrent neural network, which can retain long-term memory. Such an architecture, known as long-short-term-memory (LSTM) takes input from the encoder and produces short captions as output from decoder. The combined architecture of both encoder and decoder is used herein for generating the captions. The major steps involved in developing the said architecture of encoder and decoder are: (i) Data collection (ii) Vocabulary building (iii) Image processing and (iv) Decoding using neural network.

(i) Data Collection: The performance of the application is completely dependent on the quality of the dataset. The quality of the dataset can be measured using six parameters as depicted in Figure 1.

Revised Manuscript Received on September 05, 2020.

Mrs Nirmala, Department of Computer Science & Engineering, Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, India.

Dr,Gopalkrishna Joshi, Dean Director, Centre for Engineering Education Research B. V. Bhoomaraddi College of Engg. & Technology, Hubli, Karnataka, India.

Dr P S Hiremath, Professor, Department of Computer Science, BVB College of Engineering & Technology, Hubli Karnataka, India.

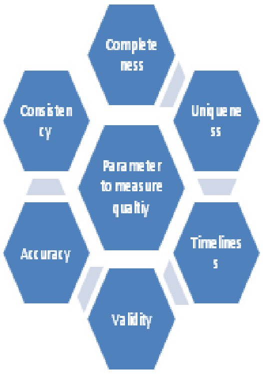


Figure 1: Parameter of quality measurement

Completeness: It is the proportion of stored data against the potential of "100% complete". Completeness is achieved through validity and accuracy.

Uniqueness: Uniquness implies that nothing will be recorded more than once. It is measured against all records within a single dataset.

Timeliness: Timeliness refers to the degree to which data represents reality from the required point in a specified time.

Validity: Data is said to be valid if it conforms to the syntax (format, type, range) of its definition. It includes comparison between the data nd the metadata for any data item.

Accuracy: It is the degree to which the data mirrors the characteristics of the real world object or objects it represents. Validity is a related dimension because, in order to be accurate, values must be valid, the right value and in the correct representation.

Consistency: Consistency refers to the absence of difference, when comparing two or more representations. Consistency can be achieved even in the absence of validity or accuracy.

By considering the above parameters, the datasets are reduced to Coco dataset and Flicker30k dataset. COCO is a large-scale object detection, segmentation, and captioning dataset and has the following features:

- Object segmentation
- Recognition in context
- Super pixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with key points

Herein, coco dataset and its python API have been used for preprocessing the vocabulary.

(ii)Vocabulary Building

After the collection of quality datasets, the dataset of vocabulary is created for the captions which represent the image dataset. The dataset of vocabulary is called dictionary. In vocabulary dictionary, each word is indexed according to the frequency of its occurrence. If the frequency of occurrence of the word is more than a pre specified threshold value, the words are retained. Otherwise, they are not considered for further analysis. Different stages of creating the dictionary of words are discussed below:

Tokenization of Captions: Tokenization is the process of splitting the text into smaller parts called tokens. The argument of this function is the text that needs to be tokenized.

The complete sentence of the caption is taken as input and vector of words is produced as output. Here, each word is considered as a token. Figure 2 depicts the same.

Captions as input
Vector or words of captions

Figure 2: Tokenization

Tokenization depends on the type of language. Languages such as English and French are referred to as space-delimited, as most of the words are separated from each other by white spaces. Languages such as Chinese and Thai are referred to as unsegmented, as words do not have clear boundaries. Tokenizing unsegmented language sentences requires additional lexical and morphological information. Tokenization is also affected by writing system and the typographical structure of the words. Structures of languages can be grouped into three categories:

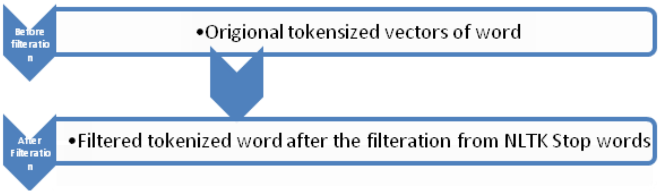
Isolating: Words that do not divide into smaller units. Example: Mandarin Chinese

Agglutinative: Words that divide into smaller units. Example: Japanese, Tamil

Inflectional: Boundaries between morphemes are not clear and are hence ambiguous in terms of grammatical meaning. Example: Latin.

Stop Words And Threshold Vectors [2]: Text may contain stop words like 'the', 'is' and 'are'. Such stop words can be filtered from the text to be processed.

Stop words can be filtered from the sentences by using open source tools. Python NLTK stop words is one such tool, which consists of a dataset of stop words, that can be used to filter the word. The steps are indicated in Figure 3.



•Origional tokensized vectors of word

•Filtered tokenized word after the filteration from NLTK Stop words

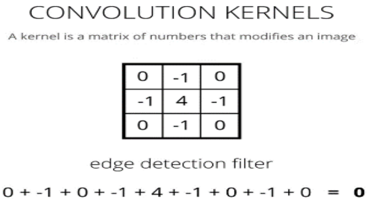
Figure 3: Stop words filtering

After processing vectors by removing stop words, threshold values are determined either through experimentation on the data or by using statistical techniques. Words above threshold value only are then selected for further considerations.

(iii) Image Processing: The images are processed as a series of image resizing, transforming, normalization and scaling, leading to extraction of features. Other procedures like random cropping [3] and random flipping are also done. The processed image is further enhanced through image visualization using Hue Saturation Intensity (HIS).

Feature Extraction From Images [4,5]: The extraction of features leads to two layers, called convolution layer and maxpooling layer. The color image comes with three different layers .i.e Red, Green, Blue (RGB). Since working with three layers is tedious, sometimes images are converted to grey scale for further analysis.

Convolution Layer [6]: The convolution step involves transformation using convolution kernel. Convolution kernel is nothing but a matrix which helps in extraction pixel by pixel. The dimensions of the convolution matrix must be passed as a parameter with the condition that, the dimension of convolution kernel must be less than the dimension of the actual image and sum of elements of the kernel matrix should add up to the value zero. Figure 4 shows Kernel matrix.

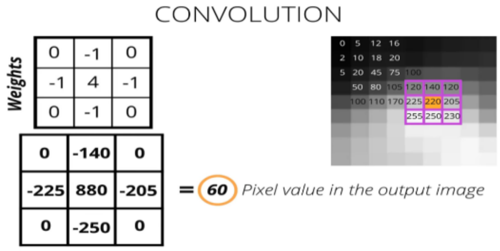


0	-1	0
-1	4	-1
0	-1	0

0 + -1 + 0 + -1 + 4 + -1 + 0 + -1 + 0 = 0

Figure 2 Kernel matrix

Using the kernel matrix, the convolution layer is generated, by super-imposing the matrix to image and multiplying the sectional area with the kernel matrix and replacing the result with the calculated value in the position of the previous centered pixel as demonstrated below in Figure 5.



0	-1	0
-1	4	-1
0	-1	0

0	-140	0
-225	880	-205
0	-250	0

Figure 5 convolution

The output of the convolution layer is taken as input to the maxpooling layer.

Maxpooling Layer: The output of the convolution layer is taken as input to the maxpooling layer. Maxpooling layer has structure as shown in Figure 6.

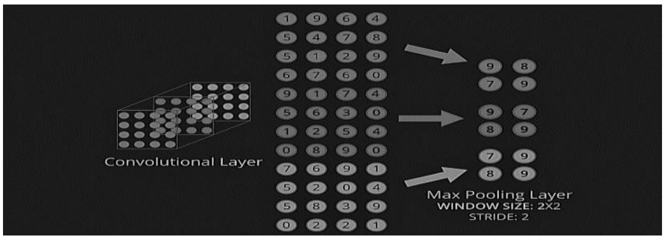


Figure 6 Max pooling layer

The order of convolution and maxpooling layer may vary from different architectures. The maxpooling layer takes 2 parameters, viz, window size and stride. The window size refers to the matrix size that covers particular portion of image matrix and stride is the size which jumps after a certain action either horizontally or vertically. The window size helps to select the maximum value available in the image size which is imposed by the window matrix and runs for next window by jumping to the next stride size and the process continues. At the end of this step, the new filtered layer that is obtained is known as maxpooling layer. Both convolution and maxpooling layers together constitute CNN Architecture. Resnet-152 Architecture is used herein for feature extraction. ResNet is a LSVRC2012 classification [6]. The core idea behind the ResNet is to introduce "Identity Shortcut Connection" that skips one or more layers as shown in Figure 7.

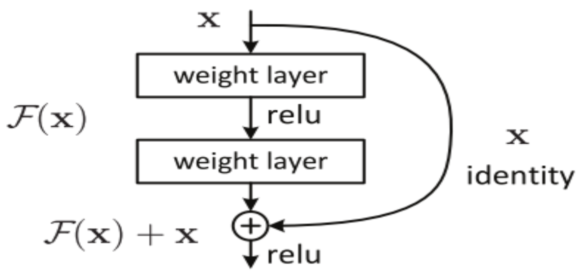

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 7 A residual Block

(iv) Decoding using neural network: The extracted feautures of the image[1] are fed to the decoder which helps to predict the words sequences. A special kind of approach known as "Transferred Learning" has been used here. This approach helps to provide pre-trained weights of the architecture. The convolution and maxpooling layers are connected to the modified functional layer. While training, only the weights of functional layer is changed and optimized and the rest of the architecture is already having the most optimised weights. The decoder takes the extracted feature vector as input and gives output to the LSTM (Long-Short-Term-Memory) architecture, which is a modification of RNN architecture. This architecture contains four logic gates viz, "Learn gate", "Forget gate", "Remember Gate" and "Use gate" [5]. The process maps the feature vectors to the word which is present in the vocabulary file, so that it can predict only the words present in the vocabulary file. The decoder takes extracted feature vector as input and gives output to LSTM architecture. The process maps the feature vectors to the word which is present in the vocabulary file, so that it can predict only the words present in the vocabulary file. The features from encoder are used in decoder (LSTM) and act as input to predict the captions in human readable from with the help of vocabulary that is already generated.

II. LSTM ARCHITECTURE

The proposed architecture has the following four logic gates, Learn gate, Forget gate, Remember gate and Use gate.[5][8].

The Learn Gate: This gate takes old STM memory and performs mathematical operations with the current event to produce a new memory The mathematical operation of this gate is a simple Tanh operation as explained below.

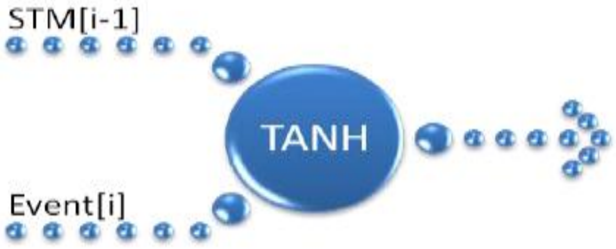


Figure 8: Learn Gate

N[i] = $Tanh(W[i](STM[i-1], Event[i]$

Here, tanh is a hyperbolic tangent operation on weights matrix formed by STM[I-1] and Event[i] and sum up to the bias(B). Sometimes it is necessary to ignore saved memory from unnecessary contents. To do that multiply the N[i] to ignore factor. The ignore factor is calculated by performing sigmoid operation on current event with the previous STM[i-1] memory.

$I = \sigma(W[i](STM[I-1], Event[i])$

*N[i] - > N[i] * I*

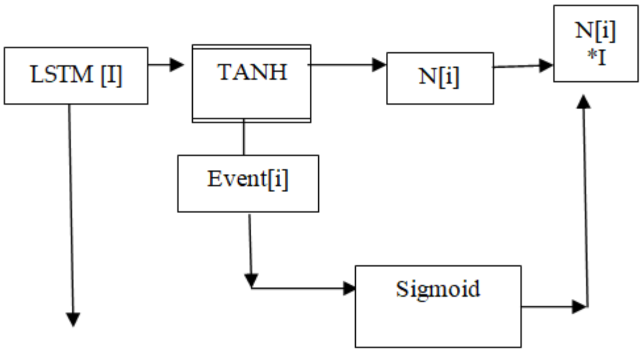
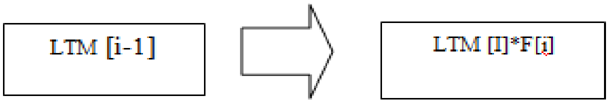


Figure 9: Learn gate with detailed representation

Forget Gate: The forget is similar to the Learn gate. It simply multiplies the long term memory with the forget factor



The forget factor is calculated by the sigmoid operation with the short term memory and the current event. Mathematically it is calculated as:

F[i] = $\sigma(W[i](STM[I-1], Event[i])$

The forget gate is represented as in Figure 10.

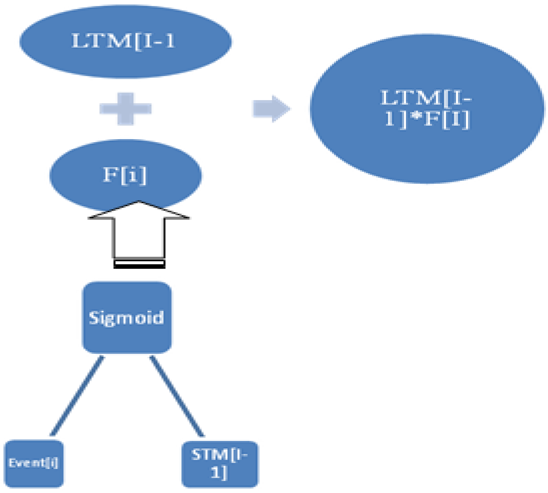


Figure 10 Forget gate

Remember Gate: The remember gate takes output of the Learn gate Forget gate as inputs and adds them up. Logically the Gate looks as shown in Figure 11.

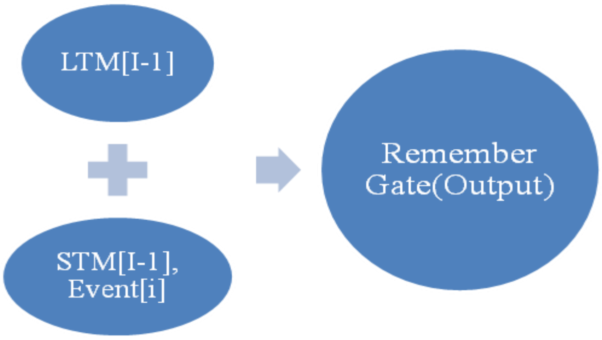


Figure 11 Remember Gate

Use Gate: The use gate is considered to be the most difficult gate, as it performs multiple operations to get the result. The first output is calculated by taking input of LTM [I-1] to the forget gate and the result is passed to the Tanh function. The second output is calculated by doing sigmoid operation with STM [I-1], Event[i], as shown in Figure 12.

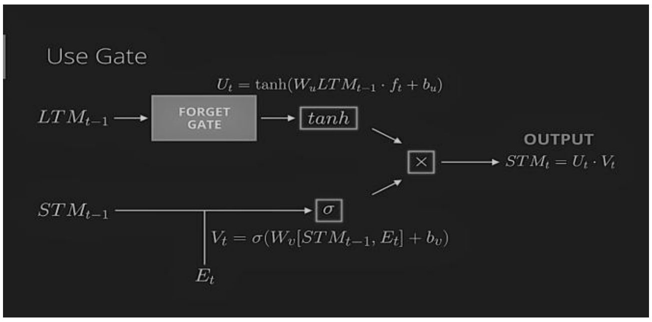


Figure 12 Use Gate

Finally, the decoder model is built as shown in Figure 13.

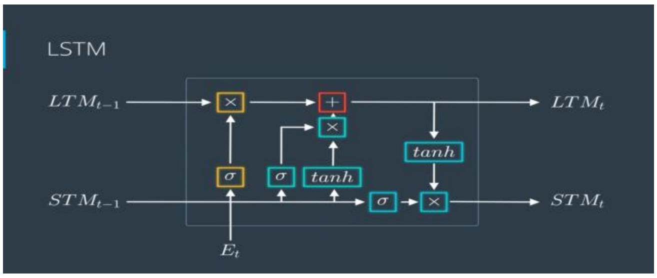
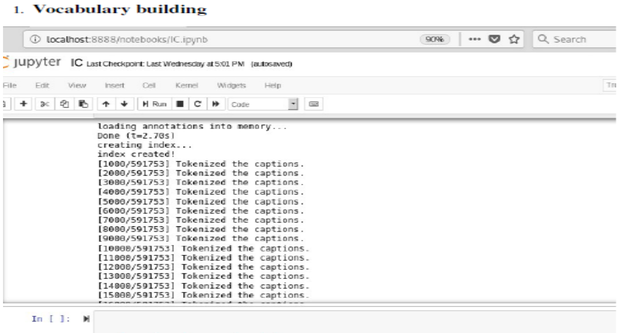


Figure 3 LSTM

III. RESULTS AND DISCUSSION

The accuracy in the mentioned architecture and model design is difficult to find. For Example, "A Man sitting in a Diner and Eating" Or "A Man in a restaurant is eating with forks" may convey the same message, using the model. But, the model generates reasonable caption to only those category which are present in the training set. For Example, if a training set consists of only images of flower and the model is being tested on animal, from the caption generated, it is difficult to judge between flowers and animals because it has never seen animals during training. The snapshot in Figure 14 is the working of application with user experience .i.e. to select an image from directory and display the result.



1. **Vocabulary building**

```
loading annotations into memory...
Done (t=2.70s)
creating index...
index created!
[1000/591753] Tokenized the captions.
[2000/591753] Tokenized the captions.
[3000/591753] Tokenized the captions.
[4000/591753] Tokenized the captions.
[5000/591753] Tokenized the captions.
[6000/591753] Tokenized the captions.
[7000/591753] Tokenized the captions.
[8000/591753] Tokenized the captions.
[9000/591753] Tokenized the captions.
[10000/591753] Tokenized the captions.
[11000/591753] Tokenized the captions.
[12000/591753] Tokenized the captions.
[13000/591753] Tokenized the captions.
[14000/591753] Tokenized the captions.
[15000/591753] Tokenized the captions.
```

localhost:8888/notebooks/IC.ipynb
90%
Search
jupyter IC Last Checkpoint: Last Wednesday at 5:01 PM (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code
In [ ]:

Figure 14: Snapshot

Example: A sample input image is shown in Figure 15.

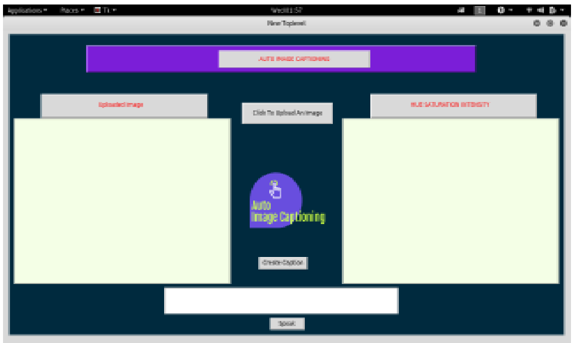
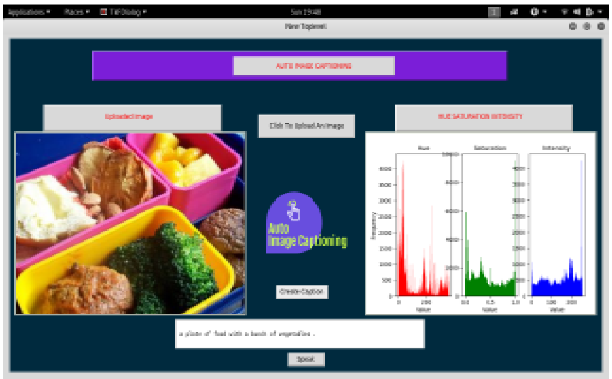


Figure 15: Input image

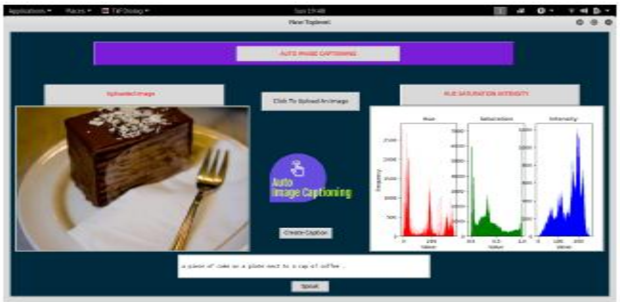
The out images giving various descriptions of the input image are depicted in Figures 16, 17, 18 and 19.



a plate of food with a bunch of vegetables .

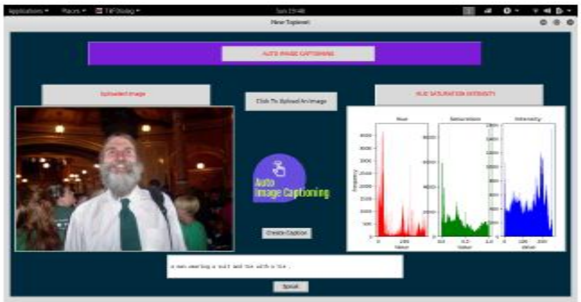
Figure 16: Image Description1

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication



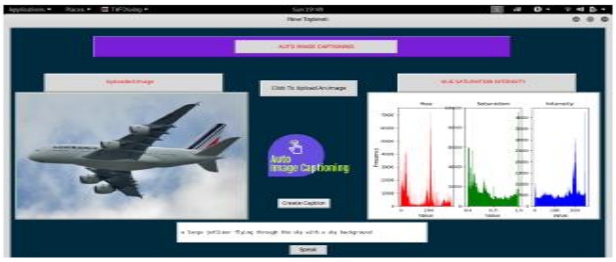
a piece of cake on a plate next to a cup of coffee .

Figure 17: Image Description2



a man wearing a suit and tie with a tie .

Figure 18: Image Description3



a large jetliner flying through the sky with a sky background

Figure 19: Image Description4

IV. PERFORMANCE EVALUATION USING BLEU METRIC

For the data MS COCO Dataset which contains 333,000 images, more than 2 million instances, 80 object categories, and 5 captions per imageTest data consists of 2010 images respectively. Each image factual captionsEvaluation Metric considered is BLEU. Metric that is used to measure the quality of machine generated text.Individual text segments are compared with a set of reference texts and scores are c. Observations from performances : The image descriptions have lack of accuracy and naturalness,imbalance and missing concepts and deficiency of semantics relevance.Thus, the observed drawback is incomplete descriptions of the input imageomputed for each of them

Method	BLEU-1 gram	BLEU-2 gram	BLEU-3 gram	BLEU-4 gram
LSTM	0.585	0.378	0.256	0.178

V. CONCLUSION AND FUTURE SCOPE

It can be observed from the output images that all the four descriptions obtained are different. All these four descriptions have lack of accuracy and naturalness, imbalance and missing concepts and deficiency of sufficient semantics. Thus, all of them provide incomplete descriptions of the input image. One way to overcome this drawback could be to use Natural Processing Language (NLP) Techniques, using which it could be possible to generate only one object decription of the entire image. Such descriptions might identify objects as well as the relationship between objects and the activities they are involved in

REFERENCES

1. Micah Hodosh, Peter Young, Julia Hockenmaier, “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics”, Journal of Artificial Intelligence Research, Vol. **47**, Issue. **1**, pp. 4188-4192, **2013**.
2. Parth Shah, Vishvajit Bakrola , Supriya Pati, “Image Captioning using Deep Neural Architectures”, In the Proceedings of the 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, **India, 2017.**
3. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition” , Computer Vision Foundation, pp.770-778, **2015.**
4. Quanzeng You , HailinJin , Zhaowen Wang , Chen Fang and Jiebo Luo, “ Image Captioning with Semantic Attention”, Computer Vision Foundation, pp.4651-4659, **2016.**
5. Mingxing Zhang , Yang Yang , Hanwang Zhang , Yanli Ji , Heng Tao Shen, Tat-Seng Chua, “More is Better:Precise and Detailed Image Captioning using Online Positive Recall and Missing Concepts Mining”, IEEE Transactions on Image Processing, Vol. 28, Issue. 1, pp. 32-44, **2019.**
6. Sepp Hochreiter, Jurgen Schmidhuber, “LONG SHORT-TERM MEMORY”, Neural Computation, Vol. 9, Issue 8, pp. 1735-1780,**1997.**
7. Rad, R., Jamzad, M.: Automatic image annotation by a loosely joint non-negative matrix factorisation. IET Comput. Vis. 9(6), 806–813 (2015)
8. hang, J., Gao, Y., Feng, S., Yuan, Y. and Lee, C.-H.: Automatic image region annotation through segmentation based visual semantic analysis and discriminative classification. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1956–1960 (2016)
9. Ivasic-Kos, M., Pobar, M., Ribaric, S.: Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme. Pattern Recogn. 52, 287–305 (2016)
10. Bhaskari, D.L., Harini, D.N.: Automatic Image Annotation: Towards a Fusion of Segmentation, Retrieval and Embedding. Scholar’s Press, Saarbrücken (2017)
11. Verma, Y., Jawahar, C.V.: Image annotation by propagating labels from semantic neighbourhoods. Int. J. Comput. Vis. 121(1), 126–148 (2017) Automatic Image Annotation: A Review of Recent Advances … 281
12. Tian, F., Shen, X., Shang, F.: Automatic image annotation with real-world community contributed data set. Multimed. Syst. (2017
13. Hao, Z., Ge, H. and Gu, T.: Automatic image annotation based on particle swarm optimization and support vector clustering. Math. Probl. Eng. (2017). Article ID 8493267
14. Yun, G., Xue, H., Yang, J.: Cross-modal saliency correlation for image annotation. Neural Process. Lett. 45(3), 777–789 (2017) Rad, R., Jamzad, M.: Automatic image annotation by a loosely joint non-negative matrix factorisation. IET Comput. Vis. 9(6), 806–813 (2015)
15. hang, J., Gao, Y., Feng, S., Yuan, Y. and Lee, C.-H.: Automatic image region annotation through segmentation based visual semantic analysis and discriminative classification. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1956–1960 (2016)
16. Ivasic-Kos, M., Pobar, M., Ribaric, S.: Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme. Pattern Recogn. 52, 287–305 (2016)
17. Bhaskari, D.L., Harini, D.N.: Automatic Image Annotation: Towards a Fusion of Segmentation, Retrieval and Embedding. Scholar’s Press, Saarbrücken (2017)
18. Verma, Y., Jawahar, C.V.: Image annotation by propagating labels from semantic neighbourhoods. Int. J. Comput. Vis. 121(1), 126–148 (2017) Automatic Image Annotation: A Review of Recent Advances … 281
19. Tian, F., Shen, X., Shang, F.: Automatic image annotation with real-world community contributed data set. Multimed. Syst. (2017

20. Hao, Z., Ge, H. and Gu, T.: Automatic image annotation based on particle swarm optimization and support vector clustering. Math. Probl. Eng. (2017). Article ID 8493267
21. Yun, G., Xue, H., Yang, J.: Cross-modal saliency correlation for image annotation. Neural Process. Lett. 45(3), 777–789 (2017)

AUTHOR PROFILE

Dr Gopalkrishna Joshi Professor of Computer Science, Dean(Curriculum Innovation), Director, Centre for Engineering Education Research in B. V. Bhoomaraddi College of Engg. & Tech., Hubli . pursued Bachelor of Science from Karnataka University Dharwad and Master from NIT Warangal and Phd from JNTU Hydrabad . published 40 papers in reputed international journals including UGC & Scopus Index and conferences including IEEE with 24 years of experience in teaching

Dr Prakash Hiremath Professor of Computer Science, MCA, Gulbarga University, BVB College of Engg & Tech, Hubli having 50 years of experience in teaching . His work focus on Image processing Pattern Recognition web mining manets and published 300 papers in reputed international journals including UGC & Scopus Index and conferences including IEEE and it is also available online having citations 3229 with h index and i10 index

Mrs. Nirmala pursued Bachelor of Science from Karnataka University Dharwad 1999 and Master of Science from Visvesvaraya Technological University in year 2009. Currently pursuing Ph.D. and currently working as Assistant Professor in the Department of Computer Science at Nitte Meenakshi Institute of Technology, Bangalore. She has published 12 papers in reputed international journals including UGC & Scopus Index and conferences including IEEE and it is also available online. Her main research work focuses on Web Image Mining, Data Mining and Computational Intelligence based education. She has 16 years of teaching experience and 4 years of Research Experience.