TEST Engineering & Management

November-December 2019
ISSN: 0193-4120 Page No. 2732 - 2742

Image Captioning with SEBL Net: Squeeze and Excitation Block combined with Bi-Long-short term memory Network

[1]Vijayarani.A, [2]*Lakshmi Priya G. G.
School of Information Technology and Engineering,
Vellore Institute of Technology, Vellore, Tamil Nadu, India.
vijayarani.a@vit.ac.in

2*School of Information Technology and Engineering,
Vellore Institute of Technology, Vellore, Tamil Nadu, India.
lakshmipriya.gg@vit.ac.in

Article Info
Volume 81
Page Number: 2732 - 2742
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *14 December 2019*

Abstract:

Rapid developments in the advancing Deep Learning (DL) made significant progress in the methodologies for Automated Captioning. Automatic captioning for digital images or videos is a great challenge in Artificial intelligence. Though most algorithms used Convolution Neural networks (CNN), this work emphasize the use of Squeeze and Excitation (SE) technique with the Long Short-Term Memory (LSTM). This combination works well to generate the caption from a sequence of words based on the learning. This proposed work bridges the gap between visual and language system by combining the two vital methodologies for image caption.

Keywords: *Deep Learning, Image Captioning, Long-Short Term Memory Block, Squeeze and Excitation, Convolution Process*

I. Introduction

Automated Captioning is the process of generating natural language descriptions for an image or video. This is a difficult and challenging task. A model that generates such a combination of visual and language system is worth and has its applications in making intelligent robots and very helpful for the visually disabled people. Commonly, generation of automatic captioning uses Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN). CNN obtains the features whereas RNN generates the captions. These two architectures proved to be dominant but performance and complexity are still endless to argue.

Computer vision from image to text model is classified into several categories based on various options. LSTM is the most employed model to generate text because it performs well. In, hierarchical LSTM model [3, 4, 15, 17, 26, 28, 29] blocks are arranged in top-down order to produce better text information. That extracted text formatof an image will be tags or phrases or sentences. Phrases are generated [3, 4, 28], sentences are extracted in [8, 47] whereas remaining works generate the caption.

LSTM is controlled by gates while generating texts for given temporal. Additional guideline given externally to each gate for temporal control [24, 31]. Some models [7, 15, 18, 32, 33, 40-42, 44] added weights to the visual feature according to its spatial location rather than using average CNN feature.

In dense captioning model [14, 18, 20, 29, 33, 40], synthesis fully localize information and

regions(salient or region of interest) for feature extraction in deep. It trains the model in end-to-end basis so it avoids the re-computation in feature extraction process. Though the deep extracted feature is better, each region sent separately to the network which repeats the process. This makes the process inefficiency.

Jie Hu et al [25] proposed squeeze and excitation model which accounts inter channel dependencies while calibrating features and proved that avoids redundant computation and feature is deep. Most of the existing works used dense or attention model to extract the spatial feature instead of the convention method like convolution. Dense and attention model are more complexed and expensive. Embedding SE block with convolution gave highlighted features at no cost and complexity is less.

So, the proposed work presents a novel architecture that combines the Squeeze and Excitation technique along with LSTM architecture for the captioning of attributes for images. Compared with images, captioning videos is very difficult as the sets of objects, scenes, attributes are diverse. The LSTM is used to overcome the vanishing gradients problem by enabling the network to learn when to forget previous hidden states and when to update hidden states by integrating memory units. CNNs are able to produce image representations that capture hierarchical patterns and attain global theoretical receptive field [25]. Squeeze and Excitation (SE) block, with the goal of improving quality of representations produced by a network by explicitly modelling the interdependencies between the channels of its convolutional features [2].

II. Related Works

Earlier, RNN used to predict the text information but it couldn't retain the information for long time which leads to gradient descendent problem. LSTM overcomes that by keeping the information for long time. So it becomesa powerful model to predict the text information of the image. But this LSTM alone is not enough to transform the visual information into textual form. Images are represented as complex in the visual form. This should be reduced by diminishing the background information.

The convolution process is good enough to extract the highlighted features. So most existing work employs the Convolution Neural Network(CNN) to down sample the images and the extracted features are passed into LSTM model to predict the textual information. Based on this the captioning model classified into several categories.

Encoder-decoder model [1, 5, 6, 11, 12, 16, 19 and 22] which is a convention model in image captioning. Generally this model, encodes the images into its feature vectors and this transmitted to the LSTM network where decoder applied and predicted the corresponding textual information for the given image.Cross domain image captioning was proposed in[12] which optimized two coupled objectives via a dual learning mechanism: image captioning and text-to-image synthesis in parallel.

Also this encoder-decoder mechanism performs well in handling the improper dataset[19] when the ground truth is not following the standards. The downside of this model is, encoded features are not strong to predict the textual information for the dominant objects of the image.

Dense model [10, 14, 18, 20, and 33] was introduced where emphasised regions of images fused with the local information and then features are extracted. This provided better result than the conventional model.Next word predicted in [10] based on the content of the image rather than the previous word. It is achieved by LSTM with dense concept. Dense achieved as a channel based feature recalibration [14] instead of dynamic recalibration. Attributes are penetrated deeply with image features and text information in each layer of training. Though the encoded features are well, extraction is more expensive.

The attention model proposed in [7, 15, 21, 32, 34, 42] where the visual information weights are added along with the spatial information.

November-December 2019
ISSN: 0193-4120 Page No. 2732 - 2742

This made precise feature extraction. The features are calibrated with the channel information and learnt weight was added for better performance.

Subsequent layer of the captioning model is, predicting the textual information for the received features. LSTM is a great network to predict text. Even though the single LSTM block predicts the text, many types of LSTM blocks are introduced for better performances.

Some models LSTM blocks are arranged in stacking manner for the perfect captioning which is called as Hierarchical LSTM(HiLSTM).HiLSTM decodes image caption from phrase to sentences in contrast to conventional solutions[4]. It generated more novel captions richer in word contents. HiLSTMwith Adaptive Attention approach applied [28] for both image and video utilized spatial and temporal attention for selecting the specific region or frames for prediction of words.

The problem of spatial attention and the feature channels and semantic concepts are addressed[11] in the Attentive Linear Transformation. During captioning, external reset is required to control the HiLSTM, Gers F et al[27] designed a self-reset control for HiLSTM. High level context information and low level visual features are handled simultaneously to generate captioning in [15, 17, 26].

Sometimes the output of the LSTM block is given as input to itself, is called as Bidirectional LSTM(BiLSTM). Cheng Wang et.al proposed an end-to-end trainable deep BiLSTM [33] to address the problem of captioning. Deep CNN is combined with two LSTMs is used to learn long-term visual language interactions by using history and future contexts. Xiao F et al[42] developed a attention based model with dual biLSTM for image captioning.

Generated captions of few models are not in natural language format. It might be continuous words or phrases from the ground truth. Another type of LSTM called as “phrase based LSTM(phiLSTM)” handles this well. It consists of a phrase decoder at the bottom hierarchy to decode noun phrases of variable length, and an abbreviated sentence decoder at the upper hierarchy to decode an abbreviated form of the image description.phiLSTM used in [11, 28] to generate caption.

Another form of LSTM is called as gridLSTM(gLSTM) where the visual content and the features are fused together. Wu L et al[30] proposed a model where visual features are extracted along with its latent information, this leads the decoder to receive information dynamically without any surplus parameters. The corresponding image contents are recalled while generating text.

Though many models other than the convention have discussed in several works, the computation cost is expensive. The feature recalibration with channel information discussed in [13, 14, 18, 37] having less complexity and high in performance.

III. Proposed Model

A model was proposed with convolution, feature recalibration with channel information and a BiLSTM block. The design and development of new CNN architecturesis a difficult engineering task, typically requiring the selection of many new hyperparameters and layer configurations. By contrast, the structure of the SE block is simple andcan be used directly in existing state-of-the-art architecturesby replacing components with their SE counterparts, wherethe performance can be effectively enhanced. SE blocks arealso computationally lightweight and impose only a slightincrease in model complexity and computational burden [25].

The ultimate goal of our model is to improve the performance and thereby the efficiency. In this section, introduced the proposedSEBLNet architecture. The images are covoluted by the CNN architecture and then passed on to the SE block. The SE block improves the channel interdependencies at no cost.

November-December 2019
ISSN: 0193-4120 Page No. 2732 - 2742

In this section, the proposed model for captioning that constructs in two stages: the first stage represents the feature vector input from convoluted layer to SE module while the second trains the feature with source sentence vector and generating captions of the image. The data set is loaded and convolution takes place that establishes non-linear connection between the channels and max pooling is obtained.

1. Extraction of Features by CNN

In this model each input image is passed into the CNN layer. CNNs use convolutional filters for extracting the hierarchical information from the images. The lower layers finds the edges, ridges and the upper layers detects the objects and other shapes. All the important information that is essential for solving the problem is extracted from the images or videos. Spatial features and channels combine to perform this task.

Convolution preserves the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation that takes two inputs such as image matrix and a filter or kernel [25].

2. Activation function- ReLu

A neural network is comprised of layers of nodes. Each node's weight is multiplied with the weight of the node and summed together. It is then transformed by an activation function which transforms and specifies the output. Figure-1 shows the proposed architecture diagram.

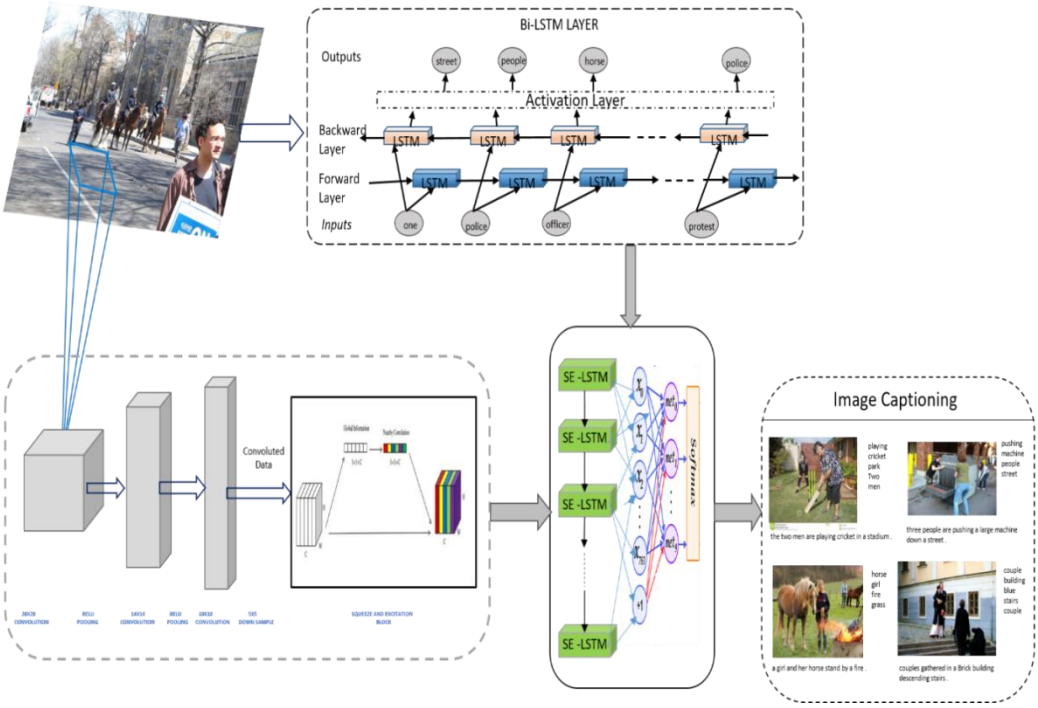


Figure-1 Proposed Architecture SEBLNet

Earlier sigmoid and hyperbolic tangent functions were used if the nodes need more learning, these were widely used.

Convolution of an image with different filters can perform operations such as edgedetection, blur and sharpen by applying filters. The filters include stride, padding, Non-linearity called ReLU, pooling layer – spatial, max, average and sum. All these are based on the inputs and how perfectly they fit. Rectified Linear Unit (ReLu) is an activation function that is used when there is need for application of transformation.

The input image is seen as a set of matrices i.e pixels. Later, with the help of the filters, it is shifted to the matrix. If filters do not fit in the image, padding can be done. ReLU is needed for the non-linearity in the image. Pooling reduces the number of parameters if they are more in number. Max pooling is done here. After pooling, the fully connected layer converts the

November-December 2019
ISSN: 0193-4120 Page No. 2732 - 2742

feature map into *n* vectors. Any number of layers can be added until the task is completed.

Here the images as vectors are being passed into the CNN and the weights are calculated. The activation function ReLu sums the weights of the nodes. Then the image vectors are fed to the SEI block.

3.Squeeze and Excitation Network

This technique maps each feature vector to a single value. *n*vectors obtained from CNN are used as weights for the feature maps. The computational cost for this mapping is very less. This can be added to any model. The Squeeze-and-Excitation block is a computational unit which can be constructed for any given transformation of input image X, where $X \in R^{H' \times W' \times C'}$ into feature U, where $U \in R^{H \times W \times C}$. H' is the height, W' is the width with C' number of channels of the input image. Let V = [v_1, v_2, . . . ,v_C] denote the learned set of filter kernels, where $v_c$refers to the parameters of the c-th filter. The outputs of Ftr can be written as

Ftr : X → U, X ∈ R H′×W′×C ′ ,
U ∈ R H×W×C
U = [u1, u2, . . . ,uC], where

$$u_c = v_c * X = CX' \quad (1)$$

Here ∗ denotes convolution,
$v_c = [v_{1c}, v_{2c}, \ldots, v_{C'c}]$

and $\tilde{X} = [x1, x2, \ldots xC']$ (2)

forward and backward sequence respectively. F_w is forward sequence weight and B_w is backward sequence weight learned from network. To get intense feature, the textual information and the spatial information are combined andthis is computed as:

$$f_h^2 = B\left(f_h^1 X_t; fw_b\right) \quad (6)$$

$$b_h^2 = B\left(b_h^1 X_t; bw_b\right) \quad (7)$$Here,

fw_b and bw_b are the learned weights of forward sequence and backward sequence, B is the Bi-LSTM, it predicts the relation between tag and Equations 1 and 2 mathematically define the SE block. The highlighted features are extracted after the SE process.

4. Long-Short Term Memory

The LSTM approach pushed back the old RNN approaches and brings an effective model by vanishing gradient descendent problem. This LSTM blocks are classified as Hierarchical LSTM (Hi-LSTM), Multi-model LSTM(MM-LSTM) and Bi-directional LSTM(Bi-LSTM). In the proposed model Bi-LSTM used where the output of current cell is forwarded to next cell and input to the same cell. It comprisesof memory cells where each cell$c_t$is controlled by three gates like input gate I_t, forget gate f_t, and output gate o_t over the time period t.Cell in LSTM gets input fromany source orpreceding cell c_{t-1} for the predicted time t. The input gate decides to retain the input or not. The content of the previous cell will be forgotten when the forget gate is on.The proposed Bi-LSTM computes as follows:

$$X_t = X\Theta Co_w \quad (3)$$

$$f_h^1 = \tau\left(F_m F_t; F_w\right) \quad (4)$$

$$b_h^1 = \tau\left(B_m B_t; B_w\right) \quad (5)$$

Here, X_t indicates the image for the time t, Cw is the weight learned from the convolution network and Θ is the convolution process. f_h^1 is the forward hidden calculated at the time t, τ represents the text based LSTM process, F_m and B_m are forward matrix and backward matrix derived from network. F_t and B_t are tags of visual information at various time stamps.This made the cells as a Long-Short Term Memory cells.

5. Squeeze and Excitation with BiLong-short term Network(SEBLNet)

The proposed model combines SE and Bi-LSTM to extract feature and train the dataset. The flexibility of the SE block means that it can be directly applied to transformationsbeyond standard convolutions. To illustrate this point, the SEBLNetdeveloped by integrating SE

November-December 2019
ISSN: 0193-4120 Page No. 2732 - 2742

blocks into modern architectures like Bi-LSTM with sophisticated designs.

Initially the data is convoluted through convolution layers and it inputted to SE block to extract the highlighted features along with inter channel dependency details. SE block squeezes the highlighted features and weaken others. The excitation process recalibrate the highlighted features.

In parallel, the proposed model passes the data into Bi-LSTM network to encode the data for specific temporal values. When the sequence is long, carrying information from the previous layer to next layers, RNNs suffer vanishing gradient problem by leaving important information that has to be given knowledge transfer to the new ones. During the process, time plays a vital role and at one point of time, gradient shrinks. As the gradients are the values that giveweights to the network layer, they stop learning over a period of time. Then the textual information to be accounted for captioning.

Every image is provided with the ground truth information as sentences. The given source sentences of the image X_t will be converted into a special vector $V_s$which has sequence of words. It is represented as

$$V_s = \{S_0, S_1, ...S_n\} \quad (8)$$

Where S_0 represents the token "START" added in the pre-processing. The model aims to improve the sum of likelihood $\log P_l$ of the related words.

$$P_l = Max\sum_{t=1}^{n} \log_p (St_t \mid X_t, St_0, St_1, ...St_{t-1}; \Phi) \quad (9)$$

Φ is a parameter which has to be learnt from network. This vector sent to SE-BiLSTM layer for word prediction as follows

$$\log_p (St_t \mid X_t, St_0, St_1, ..., St_{t-1}) = \mathrm{Re}lu(h_t) \quad (10)$$

Here $\mathrm{Re}lu$ is a non-linear function that predicts the probability of St_t Then Softmax applied to predict the related word.

These are passed into SE-Bi-LSTM for further predictions. This block computes as follows:

$$T_X = \left(\tilde{X}_i \oplus f_h^2 b_h^2; P_i\right) \quad (11)$$

In training phase of the network, 15 layers of SE is concatenated $(\oplus)$ with 5 layers of LSTM.

IV. Experimental Evaluation

Performance of a model is assessed by the experiments done. The data sets and the metrics are the key points in the evaluation.

1. Data set

Three benchmark datasets Flickr8K, Flickr30K and MSCOCO are used to verify the performance of the proposed model. Flickr8K has 8000 images, Flickr30K has 31784 and MSCOCO has118287 images. These images are annotated with minimum 5 sentences manually. Most of the image in both databases depicts human involvement in an activity.Table-1 gives the statistical details of dataset.

Table-1 Dataset Description

Dataset	Train	Test	Validation
Flickr8K	6000	1000	1000
Flickr30K	29000	1784	1000
MSCOCO	82783	40504	5000

2. Pre processing

Every image is provided with minimum of five sentences annotated by human. These given source sentences of the image will be converted into a special vector which has sequence of unique keywords and every vector has the first sentence as "START". Images are convoluted through the convolution process and then passed to the SE blocks.

3.Evaluation metrics

The main challenges faced by the evaluation for image captioning are the blind spots [50] found in the rule-based metrics and the correlation with human judgements. Bilingual Evaluation Understudy (BLEU), ROUGE, METEOR is some of the metrics that are commonly used.

Flickr and MSCOCO datasets being used these are very accurate in prediction. BLEU [49] is commonly used to measure the similarity between the two sentences. It is defined as the geometric mean of n-gram precision scores

November-December 2019
ISSN: 0193-4120 Page No. 2732 - 2742

multiplied by a brevity penalty for short sentences.

$$BLEU_n(as,bs)=\frac{\sum_{wt_n\in as}\min\left(c_{as}(Wt_n),\max_{j=1,\ldots bs}c_{bs_j}(Wt_n)\right)}{\sum_{Wt_n\in as}c_{as}(Wt_n)} \quad (12)$$

Where as is candidate sentence and bs is reference sentence, wt_n is n-gram and $c_x(y_n)$ is count of n-gram y_n in sentence x

This is a simple and quick to understand and inexpensive too. It correlates with human judgement. Recall Oriented Understudy of Gisting Evaluation (ROUGE) compares overlapping n-grams, word sequences and word pairs.

$$ROUGE_{(as,bs)}=\frac{\sum_{j=1}^{bs}\sum_{wt_n\in bs_j}\min\left(c_{as}(wt_n),c_{bs_j}(wt_n)\right)}{\sum_{j=1}^{|bs|}\sum_{wt_n\in bs_j}c_{bs_j}(wt_n)} \quad \mathbf{(13)}$$

METEOR is the harmonic mean of the precision and recall of unigram matches of sentences. It is computed as

$$METEOR=\max_{j=1}^{bs}\left(\frac{10PR}{R+9P}\right)\left(1-\frac{1}{2}\left(\frac{nc}{mu_n}\right)^3\right) \quad (14)$$

Here P is unigram precision, R is unigram recall, and nc is set of unigrams adjacent in as and bs_j.

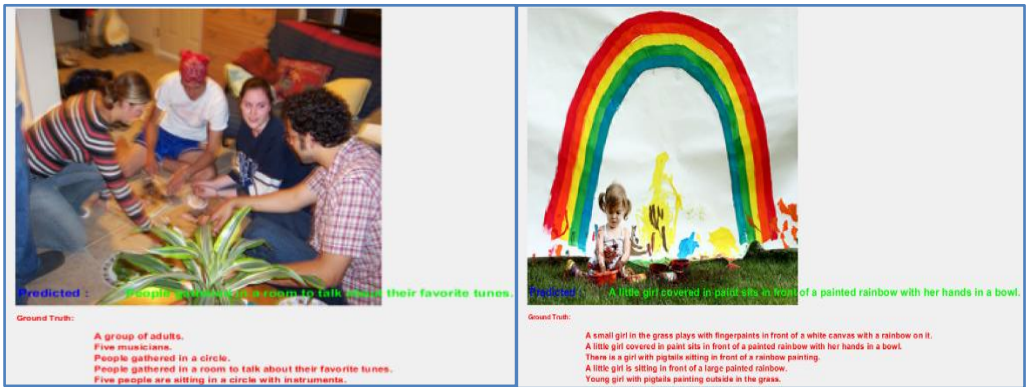
4. Implementation Details

In the proposed model, a convolution with 32 filters followed by Relu and maxpooling are used for the first three layers. Images are down sampled with a standard size of 32x32.

Then, two fully connected layers applied with Relu and softmax classifier. These two are applied with a weight learned from the network.

SqueezeNet used to encode visual features of the images, in the squeezing process, 1x1 filter applied for squeezing and 3x3 filter for excitation process. Fifteen layers of SE used in our model with the stochastic gradient descent with momentum optimizer.

Textual information of the image is given as minimum of five sentences. This converted as image vector where important key terms are stored in the vectors. SE-BiLSTM predicts the closest key terms out of the sentence by accounting the previous and future contents. The key terms of the sentences are compared and predicted to get the closest ones of the image. Figure-2 shows some sample image captioning of our model.



November-December 2019
ISSN: 0193-4120 Page No. 2732 - 2742

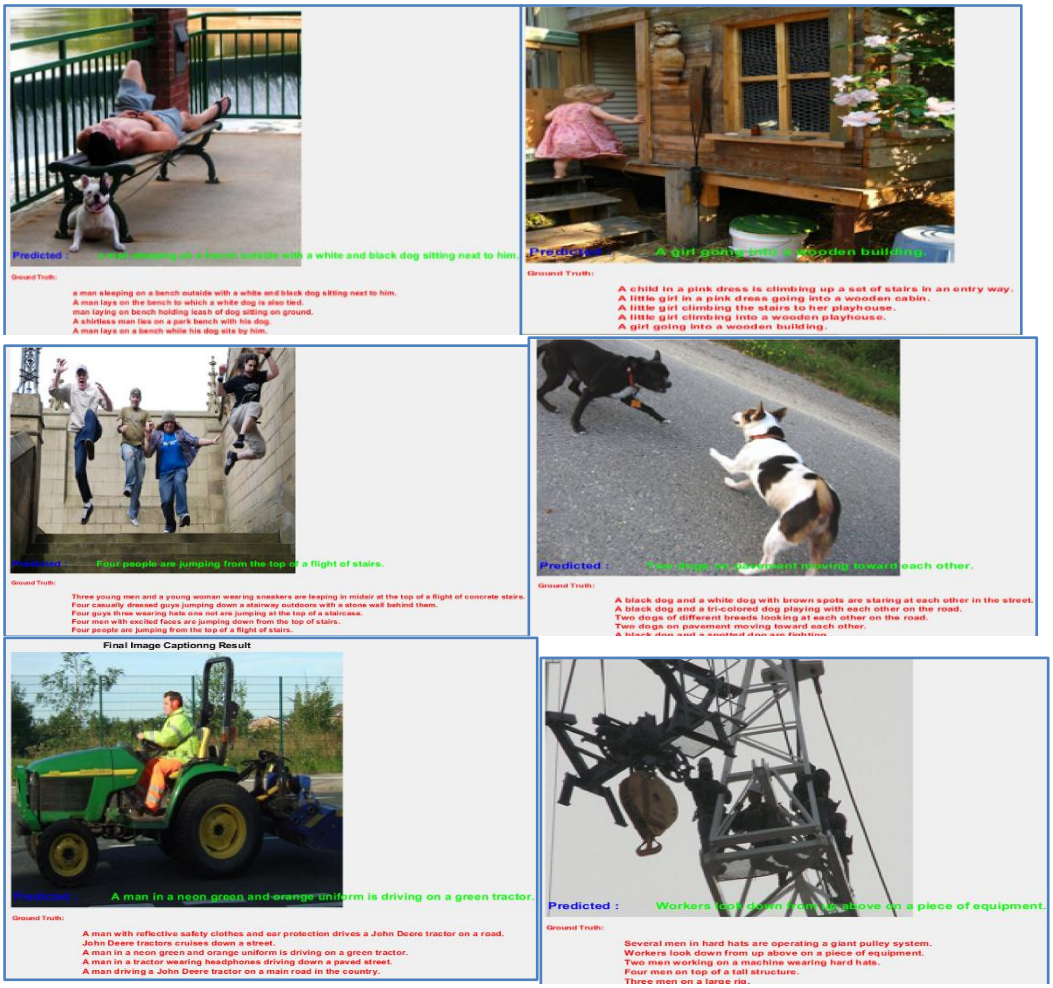


Figure-2 Sample images with predicted captioning of the proposed model. Red color indicates the ground truth generated by human. Green color is the predicted captioning of the SEBLNet.

Table-2 Comparison of proposed metrics along with existing models Highlighted number indicates that it is the higher value in that category. '-' indicates data not available.

Model Name	Flickr 8K			Flickr 30			MSCOCO		
	BLEU	ROUGE	METEOR	BLEU	ROUGE	METEOR	BLEU	ROUGE	METEOR
Recall Network[20]	-	-	-	32.23	53.9	25.92	-	-	-
ALT-ALTM[11]	-	-	-	27	48	21.2	35.5	55.9	27.4
Dense Att[18]	33.4	46.9	23.1	**39.1**	51	26	34	**56.5**	24.8
Attention Model[43]	21.5	-	-	21.1	-	-	23.5	50.5	23.5
DAA[42]	-	-	-	26.6	48.3	21.5	34.6	55.9	27.1
DTM-SBA[39]	-	-	-	25.5	-	20.9	33.8	-	26.9
SLNN[37]	18.7	40.6	19.2	21.2	41.9	19.1	30.3	52.3	21.6
SEBLNet[Proposed Model]	**34.6**	**50.4**	**24.2**	36.8	**56.7**	**27.6**	37.1	54.6	**29.7**

Proposed model compared with existing models and it proved gives better result in overall. The comparison result given in Table-2.Performance of the proposed work among various dataset for various measure is given in the Figure-3.

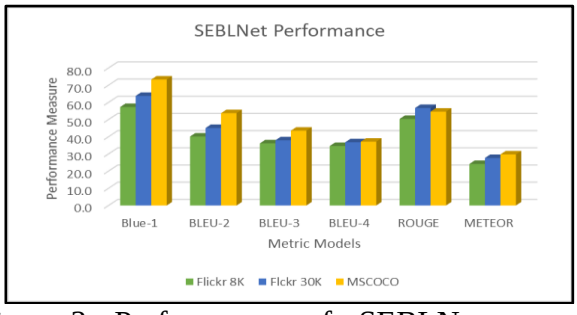


Figure-3 Performance of SEBLNet among various dataset and different metric models.

November-December 2019
ISSN: 0193-4120 Page No. 2732 - 2742

V. Conclusion

In the proposed work, image caption has predicted using SEBLNet. Convolution process applied to down sample the image and it passed to the SE block used to extracted highlighted and important features of images. First level of Bi-LSTM applied to train the network to predict the related key-terms. Then these passed to the proposed SE-BiLSTM blocks for the final caption prediction. It is observed thatthe proposed work results are better than the existing models. In future, this will be combined with Natural Language Processing to generate the captioning for videos.

References

1. Yang, Z., Yuan, Y., Wu, Y., Salakhutdinov, R., & Cohen, W. W. Encode, Review, and Decode: Reviewer Module for Caption Generation. arXiv 2016. *arXiv preprint arXiv:1605.07912*.
2. Wang, C., Yang, H., Bartz, C., &Meinel, C. (2016, October). Image captioning with deep bidirectional LSTMs. In *Proceedings of the 24th ACM international conference on Multimedia* (pp. 988-997). ACM.
3. Tan, Y. H., & Chan, C. S. (2017). Phrase-based Image Captioning with Hierarchical LSTM Model. *arXiv preprint arXiv:1711.05557*.
4. Tan, Y. H., & Chan, C. S. (2016, November). phi-LSTM: a phrase-based hierarchical LSTM model for image captioning. In *Asian Conference on Computer Vision* (pp. 101-117). Springer, Cham.
5. Yang, Z., Yuan, Y., Wu, Y., Cohen, W. W., &Salakhutdinov, R. R. (2016). Review networks for caption generation. In *Advances in Neural Information Processing Systems* (pp. 2361-2369).
6. Wang, M., Song, L., Yang, X., & Luo, C. (2016, September). A parallel-fusion RNN-LSTM architecture for image caption generation. In *2016 IEEE International Conference on Image Processing (ICIP)* (pp. 4448-4452). IEEE.
7. Fu, K., Jin, J., Cui, R., Sha, F., & Zhang, C. (2016). Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. *IEEE transactions on pattern analysis and machine intelligence*, *39*(12), 2321-2334.
8. Park, C. C., Kim, Y., & Kim, G. (2017). Retrieval of sentence sequences for an image stream via coherence recurrent convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 945-957.
9. Wu, Q., Shen, C., Wang, P., Dick, A., & van den Hengel, A. (2017). Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, *40*(6), 1367-1381.
10. Wu, Q., Shen, C., Wang, P., Dick, A., & van den Hengel, A. (2017). Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, *40*(6), 1367-1381.
11. Ye, S., Han, J., & Liu, N. (2018). Attentive Linear Transformation for Image Captioning. *IEEE Transactions on Image Processing*, *27*(11), 5514-5524.
12. Yang, M., Zhao, W., Xu, W., Feng, Y., Zhao, Z., Chen, X., & Lei, K. (2018). Multitask learning for cross-domain image captioning. *IEEE Transactions on Multimedia*, *21*(4), 1047-1061.
13. Dimou, A., Ataloglou, D., Dimitropoulos, K., Alvarez, F., &Daras, P. (2018). LDS-Inspired Residual Networks. *IEEE Transactions on Circuits and Systems for Video Technology*.
14. Zhang, K., Guo, Y., Wang, X., Yuan, J., & Ding, Q. (2019). Multiple Feature Reweight DenseNet for Image Classification. *IEEE Access*, *7*, 9872-9880.
15. Gao, L., Li, X., Song, J., & Shen, H. T. (2019). Hierarchical LSTMs with adaptive attention for visual captioning. *IEEE transactions on pattern analysis and machine intelligence*.
16. Weng, Y., Zhou, T., Liu, L., & Xia, C. (2019). Automatic Convolutional Neural Architecture Search for Image Classification Under Different Scenes. IEEE Access, 7, 38495-38506.
17. Xiao, X., Wang, L., Ding, K., Xiang, S., & Pan, C. (2019). Deep Hierarchical Encoder-Decoder Network for Image Captioning. IEEE Transactions on Multimedia.
18. Wang, E. K., Zhang, X., Wang, F., Wu, T. Y., & Chen, C. M. (2019). Multilayer Dense Attention Model for Image Caption. IEEE Access.
19. Xian, Y., & Tian, Y. (2019). Self-Guiding Multimodal LSTM-when we do not have a perfect training dataset for image captioning. IEEE Transactions on Image Processing.
20. Wu, L., Xu, M., Wang, J., & Perry, S. (2019). Recall What You See Continually Using GridLSTM in Image Captioning. *IEEE Transactions on Multimedia*.

November-December 2019
ISSN: 0193-4120 Page No. 2732 - 2742

21. Zhang, J., Wang, J., Sun, Q., Li, C., Liu, B., Zhang, Q., & Wei, X. (2019). Second-Order Response Transform Attention Network for Image Classification. *IEEE Access*, *7*, 117517-117526.
22. Park, D., Hoshi, Y., & Kemp, C. C. (2018). A multimodal anomaly detector for robot-assisted feeding using an lstm-based variationalautoencoder. *IEEE Robotics and Automation Letters*, *3*(3), 1544-1551.
23. Wu, Q., Shen, C., van den Hengel, A., Liu, L., & Dick, A. (2015). Image captioning with an intermediate attributes layer. *arXiv preprint arXiv:1506.01144*.
24. Jia, X., Gavves, E., Fernando, B., &Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE international conference on computer vision* (pp. 2407-2415).
25. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).
26. Krause, J., Johnson, J., Krishna, R., &Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 317-325).
27. Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM.
28. Tan, Y. H., & Chan, C. S. (2019). Phrase-based image caption generator with hierarchical LSTM network. *Neurocomputing*, *333*, 86-100.
29. Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G. (2017). Hierarchical multimodal lstm for dense visual-semantic embedding. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1881-1889).
30. Wu, L., Xu, M., Wang, J., & Perry, S. (2019). Recall What You See Continually Using GridLSTM in Image Captioning. *IEEE Transactions on Multimedia*.
31. Vinyals, O., Toshev, A., Bengio, S., &Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156-3164).
32. You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4651-4659).
33. Xiao, X., Wang, L., Ding, K., Xiang, S., & Pan, C. (2019). Dense semantic embedding network for image captioning. *Pattern Recognition*, *90*, 285-296.
34. Ding, S., Qu, S., Xi, Y., Sangaiah, A. K., & Wan, S. (2019). Image caption generation with high-level image features. *Pattern Recognition Letters*, *123*, 89-95.
35. Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., &Rigoll, G. (2013). LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, *31*(2), 153-163.
36. Tariq, A., &Foroosh, H. (2018). Designing a symmetric classifier for image annotation using multi-layer sparse coding. *Image and Vision Computing*, *69*, 33-43.
37. Peng, Y., Liu, X., Wang, W., Zhao, X., & Wei, M. (2019). Image caption model of double LSTM with scene factors. *Image and Vision Computing*, *86*, 38-44.
38. Wu, C., Wei, Y., Chu, X., Su, F., & Wang, L. (2018). Modeling visual and word-conditional semantic attention for image captioning. *Signal Processing: Image Communication*, *67*, 100-107.
39. Yang, J., Sun, Y., Liang, J., Ren, B., & Lai, S. H. (2019). Image captioning by incorporating affective concepts learned from both visual and textual components. *Neurocomputing*, *328*, 56-68.
40. He, X., Yang, Y., Shi, B., & Bai, X. (2019). VD-SAN: Visual-densely semantic attention network for image caption generation. *Neurocomputing*, *328*, 48-55.
41. Zhu, X., Li, L., Liu, J., Li, Z., Peng, H., &Niu, X. (2018). Image captioning with triple-attention and stack parallel LSTM. *Neurocomputing*, *319*, 55-65.
42. Xiao, F., Gong, X., Zhang, Y., Shen, Y., Li, J., &Gao, X. (2019). DAA: Dual LSTMs with adaptive attention for image captioning. *Neurocomputing*, *364*, 322-329.
43. Ding, S., Qu, S., Xi, Y., & Wan, S. (2019). Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing*.
44. Li, Y., Zhu, Z., Kong, D., Han, H., & Zhao, Y. (2019). EA-LSTM: Evolutionary attention-based LSTM for time series prediction. *Knowledge-Based Systems*.
45. Xu, N., Liu, A. A., Liu, J., Nie, W., & Su, Y. (2019). Scene graph captioner: Image captioning based on structural visual representation. *Journal of Visual*

November-December 2019
ISSN: 0193-4120 Page No. 2732 - 2742

Communication and Image Representation, *58*, 477-485.

46. Stafylakis, T., Khan, M. H., &Tzimiropoulos, G. (2018). Pushing the boundaries of audiovisual word recognition using Residual Networks and LSTMs. *Computer Vision and Image Understanding*, *176*, 22-32.
47. Islam, M. S., Mousumi, S. S. S., Abujar, S., & Hossain, S. A. (2019). Sequence-to-sequence Bangla Sentence Generation with LSTM Recurrent Neural Networks. *Procedia Computer Science*, *152*, 51-58.
48. Wang, C., Yang, H., &Meinel, C. (2018). Image captioning with deep bidirectional lstms and multi-task learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *14*(2s), 40.
49. Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., &Erdem, E. (2016). Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600*.
50. Cui, Y., Yang, G., Veit, A., Huang, X., &Belongie, S. (2018). Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5804-5812).
51. Karpathy, A., &Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3128-3137).
52. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., ...& Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2891-2903.
53. Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010, September). Every picture tells a story: Generating sentences from images. In *European conference on computer vision* (pp. 15-29). Springer, Berlin, Heidelberg.