



# CSL Net: Convoluted SE and LSTM Blocks Based Network for Automatic Image Annotation

Vijayarani. A, Lakshmi Priya G. G.

**Abstract:** Due to advancement of multimedia technology, availability and usage of image and video data is enormous. For indexing and retrieving those data, there is a need for an efficient technique. Now, Automatic keyword generation for images is a focussed research which has lot of attractions. In general, conventional auto annotation methods having lesser performance over deep learning methods. The annotation is transformed as captioning in deep learning models. In this paper, we propose a new model CSL Net (CSLN) as a combination of convoluted squeeze and excitation block with Bi-LSTM blocks to predict tags for images. The proposed model is evaluated using the various benchmark datasets like CIFAR10, Corel5K, ESPGame and IAPRTC12. It is observed that, the proposed work yields better results compared to that of the existing methods in term of precision, recall and accuracy.

**Keywords:** Automatic image annotation, Image captioning, Deep learning, Convolution, Squeeze and Excitation Block, Long – short term memory block.

## I. INTRODUCTION

Automatic Image annotation has much attention in the computer vision research domain because image search needs the content to be described. In this smart phone era, digital images are growing rapidly in social medias, blogger sites, CCTV footages, etc. These digital images need to be annotated automatically instead of manual which is expensive. Digital images are represented as multi-formats like text information and visual content. Mostly there is a discrepancy between the visual content and text information. In early stage, many conventional annotation models are available like Generative, Discriminative, Nearest Neighbour and Tag Completion models. These models used handcrafted feature extraction like Local Binary Pattern (LBP), Scale-Invariant Feature Transform (SIFT), Speed-Up Robust Features (SURF), Histogram of Oriented Gradients (HOG), etc. or combination of these, for image annotation. As images are made up of complex data and obtaining handcrafted feature extraction is tedious, so it results lesser performance in annotation task. Later deep learning model has been introduced and it is proven in various research papers [42, 43, 44] as better model because the network learns itself. Recently, deep learning methods are rocking in this research area and its slowly transforms the annotation into captioning. In deep learning, annotation models can be classified as annotating with tags [10,11,14,20,29,29,34], finding sequence of words [5,9,13,15,22,24,25,27], captioning [17,18,19,26,21] and

classification [6,7,14,16]. In the first model, assigning tags to images based on extracted features or objects in the image. Related tags are identified and make it as a sequence of tags whereas the third model combined related tags with Natural Language processing as meaningful sentence called as captioning. In deep learning annotation models, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) perform well to encode features of image and decode the features into the natural language representation [2]. Later Long-Short-Term-Memory (LSTM) has been introduced to conserve the dependency for future reference and good in natural language generating [13] which RNN can't. Memory cell of LSTM has controlled by three gates like input gate which receives data to process, history gate to calculate/retain the information and output gate to produce the output to the next cell or not. LSTM performs well to generate long sequences which can be applied for captioning. Various forms of LSTM have been applied to convert the computer vision information into natural language. Bi-directional LSTM (Bi-LSTM), Hierarchical LSTM (Hi-LSTM) and Multimodal LSTM (Mm-LSTM) are various models of LSTM used to decode the sentences or tags. Controlling In CNN, feature extraction has been performed as fusing the channel and spatial information. Recent research shows that the convolutional networks are strengthened by adding interdependencies between channels. So, the channel information are squeezed and then expands the highlighted features. Squeeze and Excitation (SE) is an intermediate block created by establishing the contingency between each channels during feature extraction. Heaping multiple blocks [30] of this kind has identified as Squeeze and Excitation Network (SEN) which enhances the quality of CNN feature extraction. Generally, deep learning image annotation applies any convolution models for feature extraction and use classifier to classify images. The proposed work combines convolution layer, SE block and Bi-LSTM as **CSL Net (CSLN)**. Here, SE block used to extract the highlighted features from the convoluted input and Bi-LSTM used to train the network and annotates images. Overview of this paper is listed below:

- Proposed CSL Net convolutes image features in phase-1. Then the convoluted features inputted to a SEN to extract the highlighted features.
- In Phase-2, dataset passed to Bi-LSTM network to learn the text information.
- In phase-3, those extracted information are combined and are considered as features for training to generate tags and train the model to generate tags.

Revised Manuscript Received on December 30, 2019.

\* Correspondence Author

**Vijayarani. A**, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.

**Lakshmi Priya G.G.**, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](http://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

## II. RELATED WORK

In general, images are represented as complex and huge values, so that the extracted features are not good enough for further training. Convolution is a process which encodes image features with channel information as vectors and it will feed into networks to train models better. CNN used to extract discriminative image features in [1,3,6,21,22,25]. CNN applied to extract words from a sentence [9]. RNN applied to classify the detected shapes [7,8,16,17,20, 21,25] and find discriminative objects [9]. Ultimate process of CNN is the convolution process among channel-based and spatial information in the local receptive fields of each layer. Even though convolution process enhances the feature extraction well, dependencies between spatial information and channels are not accounted. This diminishes the potency of feature extraction. Several works proved that the CNN extraction feature was fervent while embed with SE block in the CNN. Jie Su et al [30] introduced SE blocks to enhance the convoluted features by giving explicit attention to the interdependency among each channels. Descriptor of each channel is extracted and computed the global pooling which is identified as *squeezing*. Excitation has obtained through learning non-linear relationship and interaction among channels. Then, channel wise amalgamation applied between the learned scalar and feature maps. Features are shorten and applied to a block then expanded the highlighted features and diminish others. Non-linearity in Residual Networks (RN) made it as ineffective with redundant and diversity. Supplemented the SE blocks [32, 33] in RN to overcome the mentioned problem and re-imagined a feature map which is used to exhibit the dependencies among channels. SE blocks are embedding with bi-linear CNN [36] to address fine-grained problem of the Croatian and QUT fish data set. Also transfer-learning used to classify fine-grained images. Super-resolution refurbishment applied for data augmentation along with pre-train and high quality images. Abhijit Guha Roy et al [37] incorporate the SE blocks with fully connected CNN to recalibrate the feature extraction. Modified the SE block and introduced three variations; spatial squeeze and channel excitation which emphasises important channels, channel squeeze and space excitation which highlights the spatial information and combination of concurrent spatial, channel squeeze and channel excitation which recalibrates spatial information and channel. So that, the SE can be applied explicitly either to channel information or spatial information. Attention branch network has created on the heap of SE block [40] and are used for visual understanding like fine-grained recognition, image classification and multiple facial attribute. SE block integrate with Long-Term-Recurrent Convolutional Network [41] to understand human activities in a temporal sequence data. Residual network is implemented for recalibration of spatial information and enhance the dependency of the channels. These kind of rich feature extractions from SE motivates to embed SE in the proposed model for efficient feature extraction. LSTM is a feed backward network which retains history for the future. Various models of LSTM applied for natural language prediction in the computer vision. Bi-LSTM feeds the output to the current cell to retain the information in addition to the input [1,3,10]. Multilevel or tree structure of LSTMs are connected to make Hi-LSTM which is used to

produce sentences from phrase [2,27]. Mm-LSTM [6, 9,15,19,22,24,27] applied to generate sentences for the image in most cases. Various forms of LSTM have applied to convert the computer vision information into natural language. In Bi-LSTM, history and future contents are accounted to produce informative sentences for the image and that model has proposed by combining image and sentences which transfers the visual information into captioning or continuous sentences or multiple tags of images. Class attention model [3] used to classify an aerial view image which has only one label, but in reality it is related with multi labels. Also LSTM employed in Human Activity Recognition (HAR)[7,16] and future prediction[6]. In HAR, video information and sensor data are fed into RNN to encode the visual data and LSTM decodes the corresponding text and classifies the images to predict the lip movement during voice activity. Xingjian Shi Et al [6] proposed the rain fall prediction based on spatial and temporal data. Abbreviated sentence decoder and Phrase decoder are combined [2, 27] in a Hi-LSTM to produce sentence of an image from the phrase by correlating image to sentence and phrase to regions. Also decides what to be the next like phrase or word at every temporal. Marco Lippi et al [13] obtained the similarity between the LSTM generated text and real work. Few famous novels are taken as sample and trained LSTM to generate a long sequence of sentence which is similar writing of the novel. Amir Vatani et al [11] proposed image annotation with low-level and high-level feature extraction, tag generation and annotation which optimizes the common issue in image annotation called as semantic gap. Jacobian matrix creation in visual control systems is a challenging one which leads to give opportunity for estimation error, observation error and filter error. This is minimized while training the LSTM with Robust Kalman Filter algorithm. Salient objects in the image along with position are considered for image-text correlation [10,15] as bundled object model context. Context objects are fed sequentially into LSTM and relevant information are aggregated and encode the features of object as context based compact vectors. Fine grained was addressed [10, 28] when attributes have lack in natural language sentence or more attributes or image sizes are very small. LSTMs are trained with raw data for better performances. Multiple LSTMs are connected fully or in nested form or in parallel to make Mm-LSTM. In the paper [9] authors proposed Mm-LSTM as inner and outer LSTM which are used to encode visual features of image and predicts the corresponding text information for the image. Sometime there is no strong correlation between image and sentence [19] which requires supplement textual features. This is handled by Mm-LSTM well. In traditional image annotation models, hand crafted features are extracted and it was fed into classifier for annotation. It couldn't retain the information on the temporal basis which is an essential information to generate tag. It was proved that various LSTM models are outstanding to generate text for the extracted features. So that, Bi-LSTM has used in the proposed model to generate tags.

III. PROPOSEDWORK (CSL NET)

In this paper, a new CSL Net is proposed. Images are convoluted in the first stage of the proposed model. Squeeze and Excitation blocks used to get rich features. LSTM used to find the textual information for the images. Proposed network combines Convolution layer, Squeeze and excitation block and LSTM, then it is named as CSL Net.

In the early stage of deep learning models, RNN is used to transform visual information into textual information. For huge data,RNN couldn't handle the vanishing gradient where LSTM can. Generally,LSTM based deep learning models apply CNN to extricate image features. But its spatial element has to be intensified to get rich features. So that, SE Net is used in the proposed model to retrieve rich features from the convoluted data. We proposed CSL Net to overcome the mentioned challenges. It embeds convolution

layer for meaningful dimension reduction and SE layer applied over the convoluted data for feature extraction. SE-LSTM block has created by combining Bi-LSTM blocks with SE blocks and this could generate tags related to the visual content of the image. Block diagram of the proposed work is shown in figure-1 and description of each block is discussed in the subsections.

A. Convolution Layer

CNN learns a weight from network based on the filter information and channels. This weight applied over the local information to diminish the complexity of the network[43]. Proposed model's convolution layer built with three layers of embedded convolution, Relu and maxpooling as the first level. Then one fully connected layer followed by Relu and another fully connected layer followed by classification layer.

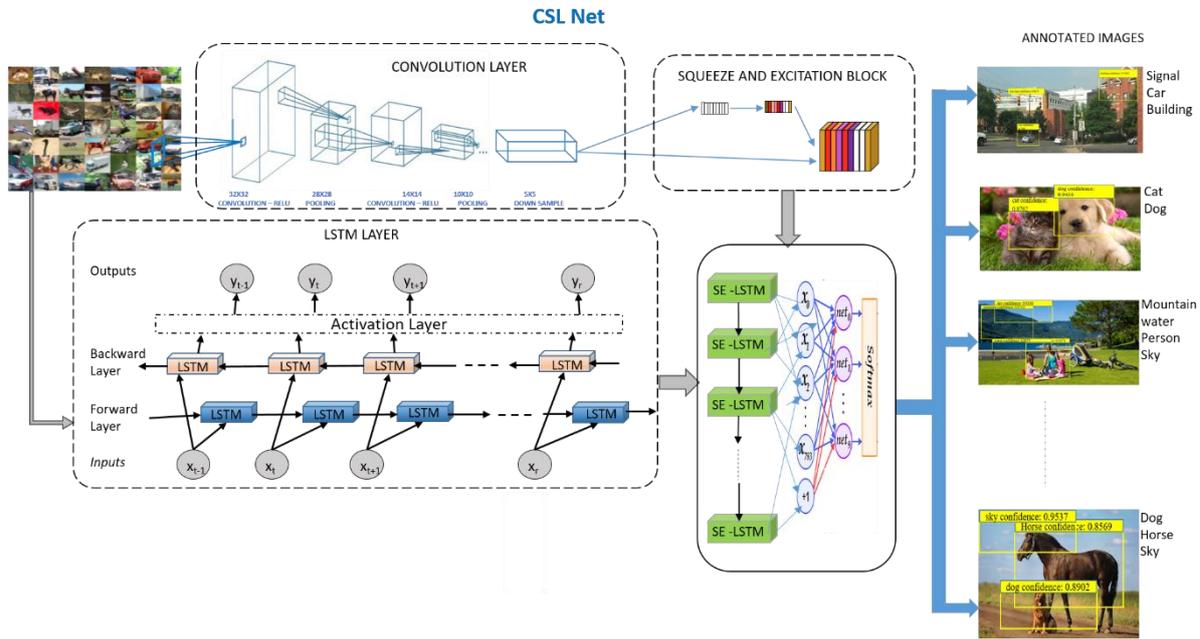


Fig.1. Proposed CSL Net

B.Squeeze and Excitation (SE)

SE block recalibrates the spatial information to find interdependencies between each channel. Squeeze process has been achieved by applying the global pooling over spatial information. Gating mechanism followed by non-linearity function applied to the squeezed information, this produces excited information. In our method, SE block adapted as follows. Applying feature transformation over the input image  $X \in R^{H \times W \times C}$ , encoder/decoder will be obtained which will be called as feature map  $U \in R^{H \times W \times C}$ .  $H$  is the height,  $w$  is the width with  $C$  is the number of channels of the input image  $X$ . The feature map of the  $i$ th image,  $u_i$  is calculated as,

$$u_i = v_i \Theta X = \sum_{s=1}^{C'} v_i^s \Theta x^s \quad (1)$$

where  $V$  is the parameter of learned filter kernels from the network,  $v_i$  is filter kernels of  $i$ th image,  $\Theta$  is the convolution of filter kernels over each channel of the  $i$ th

input image.  $v_i$ 's spatial kernel is represented by  $v_i^s$ . Squeezing ( $S_q$ ) of  $i$ th image  $ze_i$  will be obtained as,

$$ze_i = S_q(u_i) = \frac{1}{H \times W} \sum_{j=1}^H \sum_{k=1}^W u_i(j, k) \quad (2)$$

Excitation( $E_x$ ) will be extracted by applying a gating mechanism over squeezed spatial information ( $S_q$ ) with few fully connected layers. It will be computed as:

$$re = E_x(ze, Wt) = \sigma(g(ze, Wt)) = \sigma(Wt_2 \delta(Wt_1 ze)) \quad (3)$$

Here  $\delta$ -refers non-linearity,  $Wt_1 \in R^{C_r \times C}$  is known as squeezing the spatial information and recalibration achieved through  $Wt_2 \in R^{C \times C_r}$ . Finally the scaling will be figured as:

$$\tilde{X}_i = Sc(u_i, re_i) = re_i u_i \quad (4)$$

### C. Long-Short Term Memory(LSTM)

Second layer of the proposed model built with Bi-LSTM. It is a progressed model over RNN with the added advantage of feedback information to the memory cell. LSTM memory cell  $c_t$  is surrounded by few controlled gates and it decides the amount of data to be retained or cleared or assessed. Cell received input  $x_t$  from any source for a particular time(t) and accumulated, if the input gate( $I_t$ ) is activated. If the forget gate( $f_t$ ) is on for the time period t, the previous cell( $c_{t-1}$ ) status would be forgotten.

Output of the cell at the time(t) will be carried forward if the output gate( $o_t$ ) is activated. It decides the amount of data transfer from the current active cell to hidden state. Previous hidden unit is denoted as  $h_{t-1}$ . This can be processed as follows:

$$i_t = \sigma(Wt_{xi}x_t + Wt_{hi}h_{t-1}) \quad (5)$$

$$f_t = \sigma(Wt_{xf}x_t + Wt_{hf}h_{t-1}) \quad (6)$$

$$o_t = \sigma(Wt_{xo}x_t + Wt_{ho}h_{t-1}) \quad (7)$$

$$g_t = \phi(Wt_{xc} + Wt_{hc}h_{t-1}) \quad (8)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \quad (9)$$

$$h_t = o_t \otimes \phi(c_t) \quad (10)$$

$\sigma$ , denotes an activation non-linearity function,  $\otimes$  denotes multiplication yields with gate value,  $\phi$  is the hyperbolic tangent and  $Wt$  is the weight learned from network. LSTM maintains long-term temporal behaviours through selective inputs or avoiding previous states and eliminates the gradient vanishing problem. LSTM hidden output( $h_{t_0}$ ) computed as follows:

$$h_{t_0} = \{h_{ts}\}_{s=1}^S \quad (11)$$

Here,  $h_{t_0} \in \mathfrak{R}^S$ ,  $\mathfrak{R}^S$  is the dimension of  $h_{t_0}$  and  $S$  is the size of vocabulary. Softmax is applied to find the next word for the image  $X_j$ , with the learned weight  $Wt_s$  from Softmax. It is computed as:

$$\Gamma(p_{ij}; Wt_s) = \frac{\exp(Wt_s h_{ij})}{\sum_{k=1}^S \exp(Wt_s h_{ik})} \quad (12)$$

Here  $p_{ij}$  represents the probability distribution of the predicted keyword and  $Wt_s$  is the weight matrix learned from Softmax. To boost the performance of LSTM, output of each cell back propagated itself, which is identified as Bi-LSTM. Several works adapted LSTM network with Bi-LSTM cells and proved that it outperforms. Forward LSTM layer and backward LSTM layer are combined to make Bi-

LSTM in the proposed model. Starting time of both are differs forward LSTM initiates at the time t when it has the value whereas the backward LSTM starts at the time t when it has the value  $B_t$ . forward hidden sequence denoted as  $F_h$  and backward hidden sequence denoted as  $B_h$ . For the raw input image  $X$ , forward tag is identified as  $F_t$  and backward tag is identified as  $B_t$ . Our proposed network computes the encoding as follows,

$$X_t = X \Theta C o_w \quad (13)$$

$$f_h^1 = \Gamma(p_{ft}; Wt_s) \quad (14)$$

$$b_h^1 = \Gamma(p_{bt}; Wt_s) \quad (15)$$

$X_t$ , indicates the image for the time t,  $C o_w$  is the weight learned from the convolution network and  $\Theta$  is the convolution process.  $f_h^1$  is the forward hidden calculated at the time t,  $\Gamma$  represents Softmax function, probability distribution  $p_{ft}$  and  $p_{bt}$  represent forward sequence and backward sequence respectively. Then the network fuses the spatial and textual information to extract rich information, it is computed as:

$$f_h^2 = B(f_h^1 X_t; f w_b) \quad (16)$$

$$b_h^2 = B(b_h^1 X_t; b w_b) \quad (17)$$

Here,  $f w_b$  and  $b w_b$  are the learned weights of forward sequence and backward sequence,  $B$  is the Bi-LSTM, it predicts the relation between tag and visual information at various time stamps.

### D. Squeeze and Excitation-Long Short Term Memory(SE-LSTM)

Here a network is created by combining SE block with Bi-LSTM cell which will be called as SE-LSTM cell. SE squeeze and recalibrate the pre-processed data received from the previous layers. Many research works combine LSTM with convolution or embed SE in CNN to convert the computer vision into human understandable format like tag, phrases, captioning, etc. In our method, convoluted data passed to SE, raw data input into Bi-LSTM and then SE-LSTM block created by fusing this SE with Bi-LSTM. SE-LSTM layers receive extracted features from SE and keywords from LSTM. Now the network is activated with hidden units, non-linear function Relu and pooling. Fully-connected established as the end layer. Finally Softmax function applied to predict the tag  $T_X$  for the image  $X_j$ , is computed as

$$T_{X_j} = (\tilde{X}_i \oplus f_h^2 b_h^2) \quad (18)$$

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



It performs well and this work has compared with other works of conventional tag generation and deep learning tag generation.

#### IV EXPERIMENTAL RESULTS

Experiment is a key term in every research to know the performance of the proposed model and data set is a backbone to conduct the experiment. Experiments conducted in various perceptions and its performance discussed in this subsections.

##### A. Datasets

Three benchmark dataset which are used widely are selected for evaluating the proposed work.

Corel5K [43], Espgame[47], and Iaprtc12 [45]. Corel5K contains 5000 images which classify 4000 images as train image and 1000 images as test images. On an average 3.5 tags assigned to each image which are annotated manually. The overall distinct keyword is 260. Corel5k images depict humans, animals and natural scenes. Espgame dataset contains 20K images which are divided into 18689 images as training and 2081 images for testing. It has various types like logos, game screen and personal photo, all these taken from ESP collaborative image labelling tasks. The average keyword for an image is 4.6, maximum of 15 keywords for an image and overall keywords is 268. Iaprtc12 has more than 19K images which are classified into huge varieties like social life, flora, fauna, natural images, humanities, social and modern urban life. In that, 17665 images are used for training purpose and 1962 images are used for testing. The average keyword for an image is 5.7 and the overall available keyword is 291. The table-1 summarizes these benchmark dataset.

Table-I: Dataset used

Dataset	Vocabulary size	Training Images	Testing Images	Average Key words / Image	Total Keywords
Corel5K	5000	4000	1000	3.5	260
Espgame	20770	18689	2081	4.6	268
Iaprtc12	19627	17665	1962	5.7	291
CIFAR10	60000	50000	10000	-	-

##### B. Evaluation Measures

Models are assessed with standard measures like precision(P), recall(R) and f-measure(F). In this, the first step is, images are annotated with the available keywords of dataset. Precision and recall values are calculated for each keyword. Precision is the ratio between correctly predicted images for each tag and predicted images for each tag. Whereas recall is the ratio between correctly predicted images for each tag and ground truth tags. Precision ( $P_i$ ) for the keyword ( $k_i$ ) is computed as:

$$P_i = \frac{CPT(k_i)}{PT(k_i)} \quad (19)$$

Recall ( $R_i$ ) for the keyword ( $k_i$ ) is computed as:

$$R_i = \frac{CPT(k_i)}{GT(k_i)} \quad (20)$$

Here,  $CPT(k_i)$  is the correctly predicted images for the tag  $k_i$ , where  $i \in (1, \dots, S)$ ,  $S$  is the vocabulary size of dataset.

$PT(k_i)$  is predicted images for the tag  $k_i$ .  $GT(k_i)$  is the ground truth. Mean of the accuracy (P) and recall (R) of each keyword is computed as,

$$P = \frac{1}{S} \sum_{i=1}^S P_i \quad (21)$$

$$R = \frac{1}{S} \sum_{i=1}^S R_i \quad (22)$$

Finally the F measure has calculated, it is the balanced value between precision and recall.

$$F = \frac{2 \times P \times R}{P + R} \quad (23)$$

##### C. 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. To encode the text information, Bi-LSTM used with a kind of ResNet model [32]. Here five layers are constructed with 100 hidden units. Highlighted textual information of the raw data was updated with the aid of trained data from Bi-LSTM. Adaptive moment estimation optimizer have used in this model. Ultimately network trained with the proposed model by fusing the visual space and textual information. Implementation was run in CPU based system with 16GB RAM and Windows operating System. The network has trained as multiple batches.

##### D. Discussion

The proposed model CSL Net tested over four benchmark datasets like Corel5K, Espgame, IAPRTC-12 and CIFAR10. The performance of the proposed model compared with previous models for each dataset and it given in Table-2. Comparison of self-performance over various data sets is depicted in the Table 2.. It gives better performance for all dataset when compare with the existing models.

In the self-performance, CIFAR10 dataset result is the

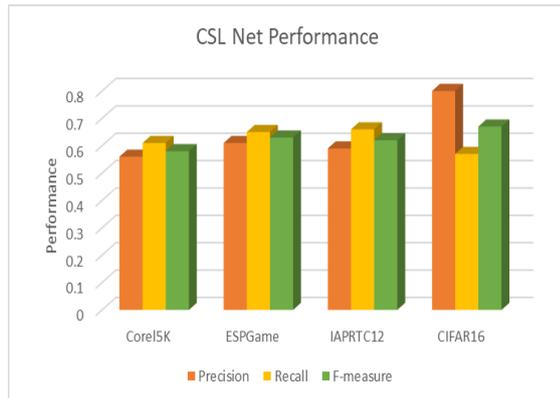
# CSL Net: Convoluted SE and LSTM Blocks Based Network for Automatic Image Annotation

best. CIFAR10 is a huge dataset of the selected data and CSL Net is a kind of deep learning model. As deep learning models are good for huge dataset, the CIFAR10 result is the

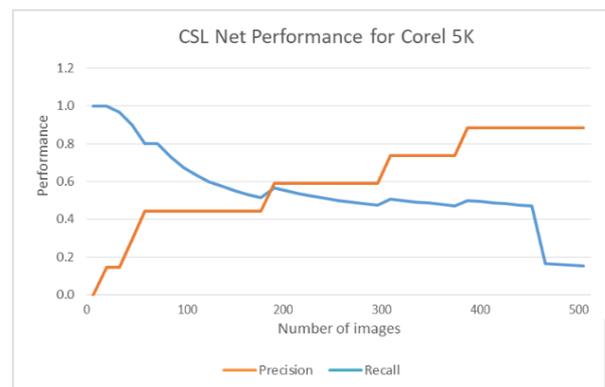
best. The Corel5K performance is low because it has lesser data.

**Table -II: Compares proposed model of CSLN with exist works.**

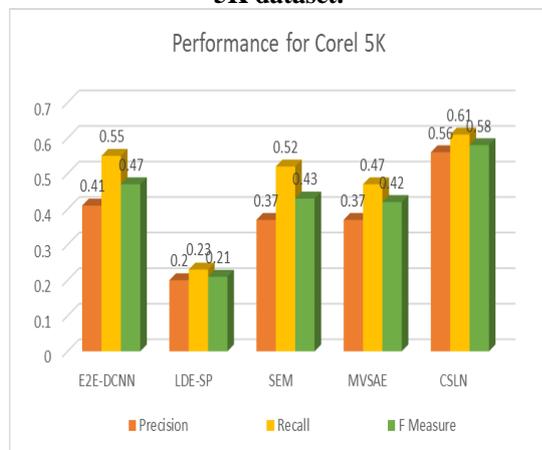
Model	Corel 5K			ESP Game			laprtc12		
	Precision	Recall	F Measure	Precision	Recall	F Measure	Precision	Recall	F Measure
E2E-DCNN	0.41	0.55	0.47	0.48	0.39	0.43	0.48	0.43	0.45
LDE-SP	0.2	0.23	0.21	0.23	0.18	0.2	0.3	0.2	0.24
SEM	0.37	0.52	0.43	0.38	0.42	0.4	0.41	0.39	0.4
MVSAE	0.37	0.47	0.42	0.47	0.28	0.34	0.43	0.38	0.4
CSLN	0.56	0.61	<b>0.58</b>	0.61	0.65	<b>0.63</b>	0.59	0.66	<b>0.62</b>



**Fig.2. CSL Net performance among various datasets**  
In this section, performance of CSLN and existing models are compared for every dataset. Fig.2 shows performance of CSLN and existing models for the Corel 5K dataset.

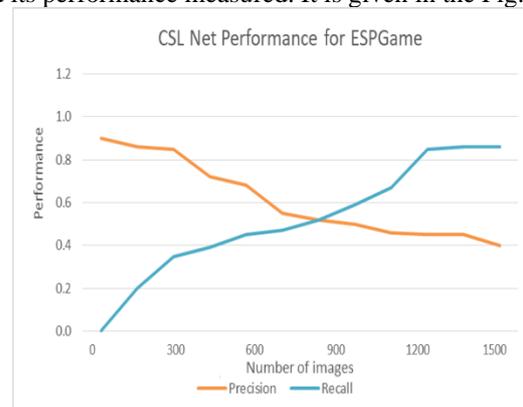


**Fig.4. CSL Net performance of Corel5K**  
Randomly picks 1500 test images from ESPGame data set and its performance measured. It is given in the Fig. 5.



**Fig.3. Performance of Corel 5**

Randomly 500 test images are taken to test the Corel5K dataset with the proposed model. The performance is given in the Fig.3. CSL Net performance of Corel5K is depicted in Fig.4.



**Fig.5. CSL Net performance of ESPGame**

Dataset	Sample Image	Predicted Tags	Ground Truth
Corel 5K		Plane Sky Cloud Jet White	Plane Sky Cloud
		Cloud Water Beach Sand Mountain	Sand Sky Cloud Beach

		Flowers Tree Sky Grass Building	Flower Tree Sky
ESPGames		Girl Doll Hair Shoe Red	Doll Blue Wall
		Woman Megaphone Fence Hat Snow	Tribal MegaPhone People
		Shadow Mountain Sky Peak portrait	Mountain Snow Sunset Sky
IAPRTC12		Building People Tree Car Cloud	People Building Car Tree
		Girl Dog Rock Grass Hat	Girl Road Dog
		Building White Car Sky Staircase	Building Car Road

Fig.6. Predicted keywords of CSLN and ground truth of sample images.

Some sample images are provided in Fig.6. along with the predicted tags of our model and ground truth. From IAPRTC12 and CIFAR10 data set, 1500 and 3000 test images are chosen at random to test the performance of CSL Net. Fig. 7 and 8 show that.

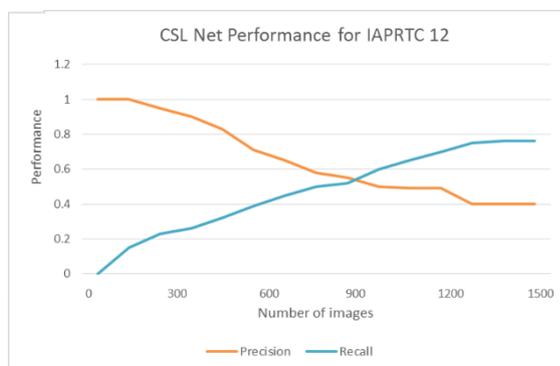


Fig.7. CSL Net Performance of IAPRTC12

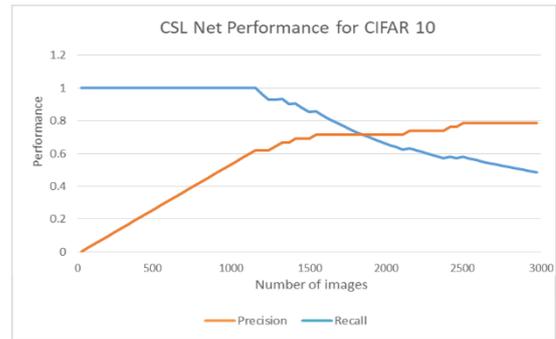


Fig.8. CSL Net Performance of CIFAR10

## V CONCLUSION

In this paper, a new CSL Net model is proposed which convolute images and fed into SE block. It squeezes the highlighted features of image and excites the squeezed feature. This extracted feature is richer feature than the convoluted features. In addition to this, Bi-LSTM block is used to extract the features which in turn used for annotating keywords for each images. Then these data from SE block and Bi-LSTM blocks are combined which yields SE-LSTM features for training and predicting keywords. The proposed model produces better quality in automatic image annotation than the conventional methods which are compared with few existing works. In future works, this model can be applied to generate captioning for images.

## REFERENCES

- 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.
- Tan, Y. H., & Chan, C. S. (2017). Phrase-based Image Captioning with Hierarchical LSTM Model. arXiv preprint arXiv:1711.05557.
- Hua, Y., Mou, L., & Zhu, X. X. (2019). Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification. ISPRS journal of photogrammetry and remote sensing, 149, 188-199.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
- Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In Advances in neural information processing systems (pp. 3294-3302).
- Xingjian, S. H. I., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Advances in neural information processing systems (pp. 802-810).
- Aung, Z. H., & Ritthipravit, P. (2015, November). Robust visual voice activity detection using Long Short-Term Memory recurrent neural network. In Image and Video Technology (pp. 380-391). Springer, Cham.
- Yan, G., Wang, Y., & Liao, Z. (2016). LSTM for Image Annotation with Relative Visual Importance. In BMVC.
- Song, J., Tang, S., Xiao, J., Wu, F., & Zhang, Z. M. (2016). LSTM-in-LSTM for generating long descriptions of images. Computational Visual Media, 2(4), 379-388.
- Huang, F., Zhang, X., Zhao, Z., & Li, Z. (2018). Bi-directional spatial-semantic attention networks for image-text matching. IEEE Transactions on Image Processing, 28(4), 2008-2020.
- Vatani, A., Ahvanooy, M. T., & Rahimi, M. (2018). An Effective Automatic Image Annotation Model Via Attention Model and Data Equilibrium. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 9(3), 269-277.
- Zhou, Z., Zhang, R., & Zhu, Z. (2019). Robust Kalman filtering with long short-term memory for image-based visual servo control. Multimedia Tools and Applications, 1-31.

13. Lippi, M., Montemurro, M. A., Degli Esposti, M., & Cristadoro, G. (2019). Natural Language Statistical Features of LSTM-Generated Texts. *IEEE Transactions on Neural Networks and Learning Systems*.
14. Yan, F., Huang, X., Yao, Y., Lu, M., & Li, M. (2019). Combining LSTM and DenseNet for Automatic Annotation and Classification of Chest X-Ray Images. *IEEE Access*, 7, 74181-74189.
15. Li, X., & Jiang, S. (2018). Bundled Object Context for Referring Expressions. *IEEE Transactions on Multimedia*, 20(10), 2749-2760.
16. Sarma, N., Chakraborty, S., & Banerjee, D. S. (2019, January). Activity Recognition through Feature Learning and Annotations using LSTM. In 2019 11th International Conference on Communication Systems & Networks (COMSNETS) (pp. 444-447). IEEE.
17. Qu, S., Xi, Y., & Ding, S. (2017, May). Visual attention based on long-short term memory model for image caption generation. In 2017 29th Chinese Control And Decision Conference (CCDC) (pp. 4789-4794). IEEE.
18. 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.
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. Kinghorn, P., Zhang, L., & Shao, L. (2018). A region-based image caption generator with refined descriptions. *Neurocomputing*, 272, 416-424.
21. Kinghorn, P., Zhang, L., & Shao, L. (2017). A hierarchical and regional deep learning architecture for image description generation. *Pattern Recognition Letters*.
22. Balderas, D., Ponce, P., & Molina, A. (2019). Convolutional long short term memory deep neural networks for image sequence prediction. *Expert Systems with Applications*, 122, 152-162.
23. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).
24. Huang, Y., Wang, W., & Wang, L. (2017). Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2310-2318).
25. 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).
26. 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).
27. 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).
28. Reed, S., Akata, Z., Lee, H., & Schiele, B. (2016). Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 49-58).
29. Khaing, P.P. (2019). A Survey in Deep Learning Model for Image Annotation.
30. 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).
31. Linsley, D., Shiebler, D., Eberhardt, S., & Serre, T. (2018). Learning what and where to attend.
32. Hu, Y., Wen, G., Luo, M., Dai, D., Ma, J., & Yu, Z. (2018). Competitive inner-imaging squeeze and excitation for residual network. *arXiv preprint arXiv:1807.08920*.
33. Singh, P., Mazumder, P., & Nambodiri, V. P. (2019). Accuracy Booster: Performance Boosting using Feature Map Re-calibration. *arXiv preprint arXiv:1903.04407*.
34. Lu, J., Li, R., Zhang, Y., Zhao, T., & Lu, Z. (2010). Image annotation techniques based on feature selection for class-pairs. *Knowledge and information systems*, 24(2), 325-337.
35. Vallet, A., & Sakamoto, H. (2015). A multi-label convolutional neural network for automatic image annotation. *Journal of information processing*, 23(6), 767-775.
36. Qiu, C., Zhang, S., Wang, C., Yu, Z., Zheng, H., & Zheng, B. (2018). Improving transfer learning and squeeze-and-excitation networks for small-scale fine-grained fish image classification. *IEEE Access*, 6, 78503-78512.
37. Roy, A. G., Navab, N., & Wachinger, C. (2018). Recalibrating Fully Convolutional Networks With Spatial and Channel "Squeeze and Excitation" Blocks. *IEEE transactions on medical imaging*, 38(2), 540-549.
38. Hu, J., Shen, L., Albanie, S., Sun, G., & Vedaldi, A. (2018). Gather-excite: Exploiting feature context in convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 9401-9411).
39. Wang, R., Xie, Y., Yang, J., Xue, L., Hu, M., & Zhang, Q. (2017). Large scale automatic image annotation based on convolutional neural network. *Journal of Visual Communication and Image Representation*, 49, 213-224.
40. Fukui, H., Hirakawa, T., Yamashita, T., & Fujiyoshi, H. (2019). Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 10705-10714).
41. An, G., Zhou, W., Wu, Y., Zheng, Z., & Liu, Y. (2018, August). Squeeze-and-Excitation on Spatial and Temporal Deep Feature Space for Action Recognition. In 2018 14th IEEE International Conference on Signal Processing (ICSP) (pp. 648-653). IEEE.
42. Wu, X., Zhang, L., Li, F., & Wang, B. (2018). A Novel Model for Multi-label Image Annotation. 2018 24th International Conference on Pattern Recognition (ICPR), 1953-1958.
43. Ma, Y., Liu, Y., Xie, Q., & Li, L. (2019). CNN-feature based automatic image annotation method. *Multimedia Tools and Applications*, 78(3), 3767-3780.
44. Ke, X., Zhou, M., Niu, Y., & Guo, W. (2017). Data equilibrium based automatic image annotation by fusing deep model and semantic propagation. *Pattern Recognition*, 71, 60-77.
45. Ke, X., Zou, J., & Niu, Y. (2019). End-to-End Automatic Image Annotation Based on Deep CNN and Multi-Label Data Augmentation. *IEEE Transactions on Multimedia*.
46. Yang, Y., Zhang, W., & Xie, Y. (2015). Image automatic annotation via multi-view deep representation. *Journal of Visual Communication and Image Representation*, 33, 368-377.
47. Jin, C., & Jin, S. W. (2016). Image distance metric learning based on neighborhood sets for automatic image annotation. *Journal of Visual Communication and Image Representation*, 34, 167-175.

## AUTHORS PROFILE



**Vijayarani A** received the M.Tech degree in Computer Science and Engineering from VIT University, India. She is an Assistant Professor in the Department of Information Technology, VIT University, India. Her research interests include image processing, data mining, computer vision and multimedia, Semantic Indexing, and Machine learning.



**Lakshmi Priya G. G.** is currently with the School of Information Technology and Engineering, VIT University, Vellore, India. She received the M.C.A. and M.E. degrees in 2004 and 2007, respectively, and the Ph.D. degree from the National Institute of Technology at Tiruchirappalli, Tiruchirappalli, India, in 2014. Her research interests include temporal video segmentation, content-based video retrieval, and video analysis