D-CNN: A New model for Generating Image Captions with Text Extraction Using Deep Learning for Visually Challenged Individuals

Madhuri Bhalekar
School of Computer Engineering and Technology
MIT World Peace University
Pune, India
madhuri.bhalekar@mitwpu.edu.in

Mangesh Bedekar
School of Computer Engineering and Technology
MIT World Peace University
Pune, India
mangesh.bedekar@mitwpu.edu.in

Received: 24 January 2022 | Revised: 14 February 2022 | Accepted: 22 February 2022

Abstract-Automatically describing the information of an image using properly constructed sentences is a tricky task in any language. However, it has the potential to have a significant effect by enabling visually challenged individuals to better understand their surroundings. This paper proposes an image captioning system that generates detailed captions and extracts text from an image, if any, and uses it as a part of the caption to provide a more precise description of the image. To extract the image features, the proposed model uses Convolutional Neural Networks (CNNs) followed by Long Short-Term Memory (LSTM) that generates corresponding sentences based on the learned image features. Further, using the text extraction module, the extracted text (if any) is included in the image description and the captions are presented in audio form. Publicly available benchmark datasets for image captioning like MS COCO, Flickr-8k, Flickr-30k have a variety of images, but they hardly have images that contain textual information. These datasets are not sufficient for the proposed model and this has resulted in the creation of a new image caption dataset that contains images with textual content. With the newly created dataset, comparative analysis of the experimental results is performed on the proposed model and the existing pre-trained model. The obtained experimental results show that the proposed model is equally effective as the existing one in subtitle image captioning models and provides more insights about the image by performing text extraction.

Keywords-image captioning; text extraction; convolutional model; long short-term memory; deep learning

I. INTRODUCTION

Image captioning is one of the numerous computer vision tasks that have been aided by deep learning. Image captioning is helpful to annotate objects which might be overlooked by a human observer. It can be useful when images are extremely cluttered and it is difficult to recognize objects in them. Currently, the majority of research is done with image description, which can be extended on scene comprehension through images and videos. With domain-specific image captioning datasets (such as sports, night vision images, forest animal images, etc.) we can further improve the accuracy of a captioning system. In the existing available research, the textual part present in an image is not considered while generating its caption. Considering this research gap, the objective of the proposed system is to generate a new image captioning dataset and to perform more prescribed image captioning with text extraction from the image and including the extracted text in generated captions. The proposed system can be deployed as a mobile application where the user takes a photograph as an input image and the system generates the description of the image. This will be helpful as a smart assistant for children and especially for visually challenged people. The proposed system can also be deployed on robots to help them understand their surroundings. It can be used even in an application to identify famous objects for a tourist who goes without a tour guide. The user can capture the monument or artifact in his mobile phone camera and the application displays relevant information in the form of text.

II. RELATED WORK

Many different architectures have been proposed for generating captions for a given image. Authors in [1] presented an image captioning system that takes advantage of image and sentence parallel structures. It also introduces scene-specific contexts, resulting in a novel model. For an input image, it finds visual regions and generates words for these regions, and performs alignment between visual regions and words. The attention model aligns words and regions very well. Authors in [2] proposed a two-stage image captioning module which consists of detectors. A collection of attribute classifiers is applied to each candidate object region and the sentences are constructed using a Conditional Random Field (CRF) to predict labelling for an input image. For experimentation they used the UIUC PASCAL dataset. This approach can be further extended to accommodate more general image contents, such as actions and scenes, and can also be used to describe video content. In [3], the authors proposed a system in which region-wise natural language descriptions of the images are generated. The suggested model is a novel combination of CNNs applied on images and Bidirectional Recurrent Neural Networks (BRNNs)

Corresponding author: Madhuri Bhalekar

applied on text and the alignment of these two models is described as the Multimodal Recurrent Neural Network architecture. A Region Convolutional Neural Network (RCNN) is used to recognize items in each image. This approach is made up of two different models. One approach is to utilize a multimodal embedding to match word samples with the visual regions they describe, and then use the correlation as training data for a Multimodal Recurrent Neural Network model, to learn to produce the samples. The system can be further improved by creating a single model which can be trained with the image annotation dataset to region-level annotations.

Authors in [4] presented an architecture with a linguistic model and a multimodal part. A Deep CNN extracts the image features. A one-layer representation connects the language model and the deep CNN in the multimodal component. A perplexity-based cost function is used to learn the m-RNN model. The errors are propagated backwards through the three sections of the m-RNN model, updating the model parameters. The m-RNN architecture performs three tasks: i) sentence creation, ii) sentence retrieval (finding the most appropriate sentences for a given image), and iii) image retrieval. This technique can also be expanded to include more complicated image features and a more complex language model. In [5], a generative model based on the deep recurrent architecture is proposed. This research combines Vision Deep CNN with Language Generating RNN to generate natural sentences that describe an image. Given the training image, the model is trained to maximize the probability of the desired description sentence. This model is referred to as the Google NIC (Neural Image Caption) Generator. To produce descriptions from images, the authors propose a neural and probabilistic framework. The proposed model is quite accurate. However, the problem of how to use unsupervised data to improve the image description is not solved and is discussed in the proposed method. In [6], the bi-directional mapping between images and their sentence-based descriptions is proposed. An RNN is the key component of the approach, as it attempts to dynamically construct a visual representation of the scene. On its own, the representation learns to remember long-term visual notions. The model has the ability of producing unique captions and reconstructing visual features from an image description. For learning, the Back Propagation Through Time (BPTT) algorithm is used. The RNN model works in both directions. As a result, it is possible to derive image features from sentences as well as sentences from image features. Long-term interactions can be learned by the model. It would be interesting to replace the proposed RNNs with Long Short-Term Memory (LSTM) models to learn a bi-directional model as the authors had not examined the usage of LSTM models that have proven the ability to learn long-term notions.

The proposed method in [7] employs an attention-based model which is capable to describe the image content. It has an Encoder-Decoder form, in which a CNN is used as an encoder to extract a set of feature vectors, and the LSTM network is used as a decoder. The authors presented two different processes for the attention model: stochastic and deterministic attention. Both models can pay attention to important parts of an image while producing a caption, and the output is more interpretable as a result of the attention mechanism. Authors in [8] used pre-trained CNN models VGG16 and ResNet-101 and the experimental results on Flickr8K dataset were tested and compared. In [9], for the classification task a small portion of the training images were labeled. The purpose is to predict the class label of test images which are not having tags. When paired with visual characteristics, tags provide a suitable feature that further improves the performance of classification. It uses the Multiple Kernel Learning (MKL) framework which provides a semi-supervised learning strategy for using the information contained in the tags linked with unlabeled images.

For classification of leaf diseases, the hybrid deep learning approach gives better performance [10]. In [11], authors presented a deep learning approach for the early detection of banana diseases, which can be deployed on a mobile system. Along with this, authors in [12, 13] presented their work on object detection methods. In [12], the authors used a binary mask which is created around the object to identify the borders of the image. The binary mask is applied in both vertical and horizontal fashions to identify the borders of the image. After that, a refinement process is applied and repeated until the classification is as precise as possible. As a result, the basic strategy is to use the entire image as an input and perform localization using a Deep Neural Network (DNN). The DNN localizer is used on a small set of big windows for precision localization. This approach has some computational cost during training because each object type and mask type must be trained separately. Authors in [13] propose a special formulation for Multiple Instance learning (MIL) for object detection. The MIL optimization problem is modified into a convex problem and is addressed using Stochastic Gradient Descent (SGD). Relaxed Multiple Instance (RMI) SVM can be used to solve various recognition problems, such as visual tracking, image classification, and object detection. The experimental results show that RMI-SVM outperforms other known MIL benchmark methods.

III. THE PROPOSED SYSTEM MODEL

After analyzing the existing systems, we have identified some of the research gaps. The current research work is done mostly with generic image datasets, with domain-specific image captioning datasets we can further improve the accuracy of a captioning system. In the existing available research, attention on textual part present in image is missing while generating the captions from an image. Considering these research gaps, the main objective of the proposed system is to generate a new image captioning dataset and further to perform more prescribed image captioning with text extraction from the image and include the extracted text in the generated captions. The overall system flow of the proposed model is presented in in Figure 1.

A. Dataset Creation (Phase I)

In the first phase, we have created a new image captioning dataset, in which images with textual information are included. It builds on [14] to understand the dataset creation model for image captioning task in which the authors have created and presented the Microsoft COCO dataset. Considering this benchmark dataset for image captioning, we have created the image dataset of MIT World Peace University campus vicinity

consisting of 1500 images which are labeled. These labels talk about the significant objects present in the image. For extracting the text information, we have considered images which contained textual information. Validation on images is done while creating the new dataset. The images should provide a good representation of the classes which can further help the classification process, diverse images must be collected to have proper training on the dataset in order to avoid the problem of over fitting or under fitting during training. Along with the image files, the new dataset also contains relevant caption files (in CSV and JSON format) which are further used in generating the natural language description of the image. Further, the model was trained and validated with the CNN architecture. This allows the system to predict multiple significant objects present in the input image and will generate an appropriate description using natural language processing. This dataset is used in Phases II and III as shown in Figure 2 which represents the phase wise working of the proposed model for image captioning with text extraction.

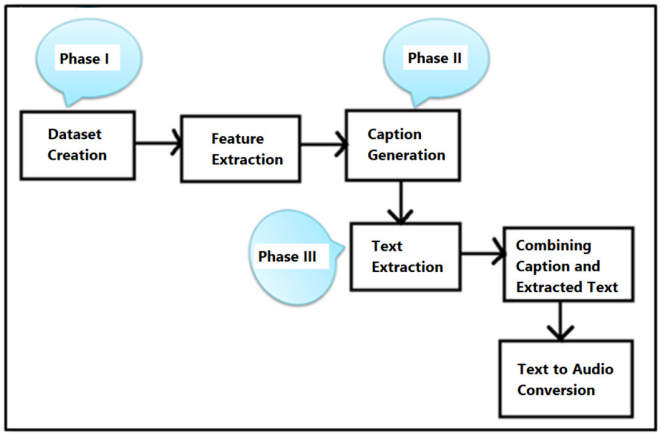


Fig. 1. Overall system flow of the proposed architecture.

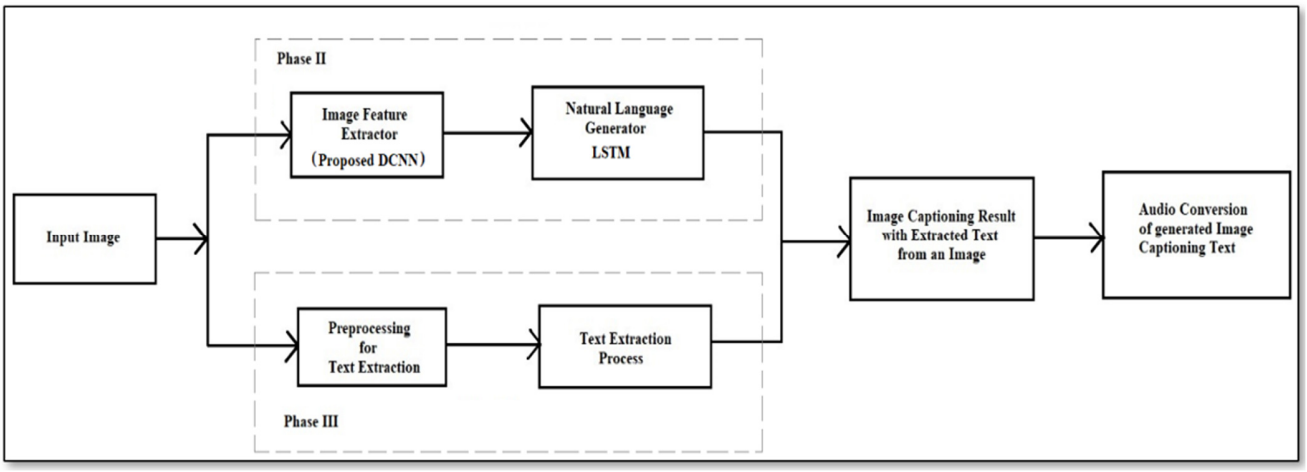


Fig. 2. Overview of the proposed model.

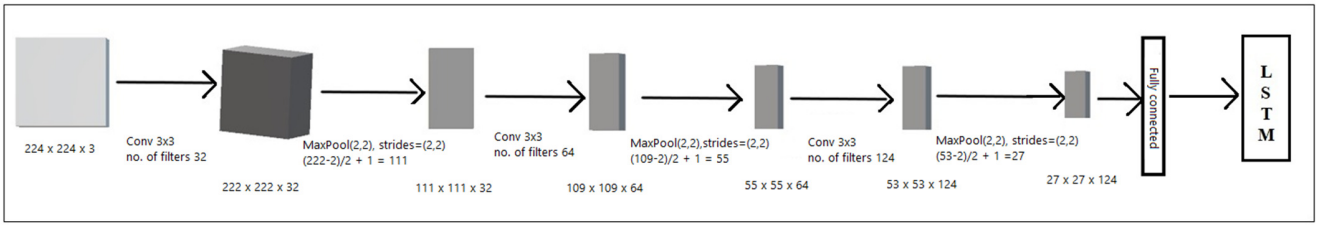


Fig. 3. Proposed D-CNN model followed by LSTM for image feature extraction.

B. Image Feature Extractor with the Proposed CNN Model (Part of Phase II)

CNN works fantastically in computer vision tasks such as recognizing different objects present in an image. It consists of multiple layers of a CNN architecture [24]. The ReLU layer serves as an activation function, maintaining non-linearity which passes through each layer of the network. The fully connected layer allows performing the classification task.

In the proposed model we used three convolution layers which allow seeing specific features using filters. The convolutional layers consist of 32, 64, and 124 filters followed by the max pooling layer which reduces the sample size of the extracted feature map which makes processing faster as it reduces the number of parameters from processing. The pooling layer provides the pooled feature map. We propose the enhanced Deep CNN (D-CNN) model for image feature extraction followed with LSTM as shown in Figure 3. While creating the proposed optimized enhanced D-CNN model we performed optimization and took into consideration different parameters like the number of convolutional layers, flatten layers, and dense layers, the number and size of filters,

optimizer type, learning rate, number of epochs, batch size, etc. Based on this experimentation, the different hyper parameters used in model training are presented in Table I. We tuned the model with different values of batch size, number of epochs, and different types of optimizers such as Stochastic Gradient Descent (SGD), Adam optimizer, and Root Means Square Propagation (RMSprop) optimizer. After analyzing the obtained results, we selected the SGD optimizer which updates the weight and bias parameters to get the minimum loss/error while training the CNN as shown in (1) and (2) where w represents the weight and b represents the bias value. From the obtained loss value, we calculate the partial derivative of loss and then adjust the weight to reduce the error.

$$w = w - LearningRate * (\partial Loss / \partial w) \quad (1)$$

$$\because \partial Loss / \partial w = 1/n \sum_{i=1}^{n} (actual - predicted\, OutputValue)$$

$$b = b - LearningRate * (\partial Loss / \partial b) \quad (2)$$

$$\because \partial Loss / \partial b = 1/n \sum_{i=1}^{n} (actual - predicted\, Output\, Value)$$

SGD gradient measures how much the output of a function changes if there is a change in input value. Gradient is the slope of function in which a higher value indicates that the model learns faster and a lower value with almost zero slope indicates that the model stopped learning. After analyzing the obtained results, the hyper parameters for the proposed model were selected: batch size 64, number of epochs 60, and SGD optimizer. The results of hyper-parameter tuning for the proposed model are shown in Table I.

TABLE I. HYPER PARAMETER TUNING FOR THE PROPOSED MODEL

Batch size	No. of epochs used	Optimizer's used		
		SGD	Adam	RMSprop
32	20	78.12	75.30	75.90
	40	79.33	78.08	77.53
	60	79.86	78.66	79.19
	100	79.81	78.93	79.27
64	20	79.36	78.77	79.85
	40	79.28	78.90	79.49
	60	79.96	79.85	74.40
	100	79.72	77.29	77.38
Selected hyper parameters for the proposed model: Batch size 64, 60 epochs, SGD optimizer.				

The obtained training and validation accuracy was 86.30% and 74.56%, due to the over fitting problem. To avoid the over fitting problem, initially we used the dropout method which is a regularization technique. Using the dropout normalization, the system achieved 85.28% training accuracy and 83.02% validation accuracy. To improve the performance of the proposed model, the model was trained with dropout and batch normalization and 87.43% training accuracy and 86.30% validation accuracy were achieved. The results of cross entropy loss and classification accuracy with dropout and batch normalization are shown in Figure 4. The training loss and validation loss, after training the model and tuning the hyper parameters, are shown in Figure 5. Figure 6 shows the accuracy during training and validation. The optimized parameters for the proposed D-CNN model are shown in Table II.

TABLE II. OPTIMIZED PARAMETERS OF THE PROPOSED MODEL

Parameters	Learning rate	Regularization method	Training time/ epoch	Training accuracy	Validation accuracy
6,237,936	0.01	-	8.345 sec	86.30%	74.56%
6,237,936	0.01	Dropout	8.4532 sec	85.28%	83.02%
6,328,816	0.01	Dropout and batch normalization	9.4787 sec	87.43%	86.30%
Selected hyper parameters for the proposed model: Batch size 64, 60 epochs, SGD optimizer.					

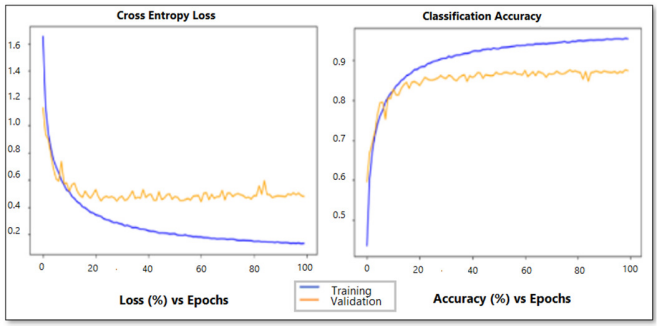


Fig. 4. Cross entropy loss and classification accuracy with dropout and batch normalization.

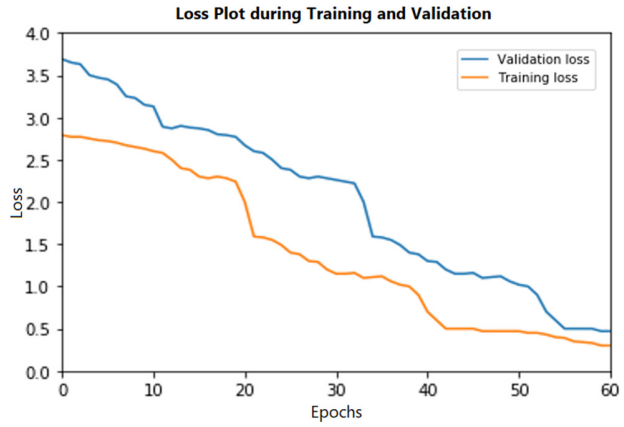


Fig. 5. Loss plot during training and validation.

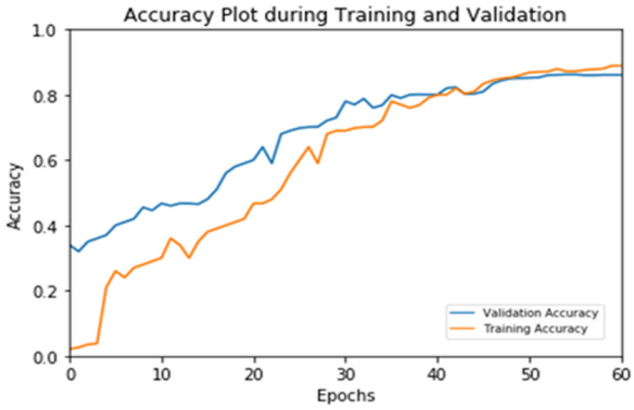


Fig. 6. Accuracy plot during training and validation.

C. Caption Generation Using the LSTM Model ((Phase II)

After obtaining the extracted image features, the next step is to use Vector to sequence representation with the LSTM model [15, 16] to generate the caption for the image as shown in Figure 7 and the LSTM structure as shown in Figure 8. LSTM consists of forget gate ft, input gate it, output gate ot with one cell state Co with each gate having different weight vectors Wc, Wi, Wo, Wf. Sigmoid activation function is used in LSTM. The computations are presented in the following equations [15, 16]:

$$\text{Forget Gate: } ft = \sigma(Wf\, St\text{-}1 + Wf\, Xt) \quad (3)$$

$$\text{Input Gate: } it = \sigma(Wi\, St\text{-}1 + Wi\, Xt) \quad (4)$$

$$\text{Output Gate: } ot = \sigma(Wo\, St\text{-}1 + Wo\, Xt) \quad (5)$$

The intermediate cell state is:

$$Ct = \tanh(Wc\, St\text{-}1 + Wc\, Xt) \quad (6)$$

The value for the input gate is obtained using the sigmoid activation function. After that, the intermediate cell Ct is obtained using the tanh activation as shown in (6). Similarly, the calculation of forget gate is done and its multiplication with old state Co is performed and by adding these we get the new cell state C1. Finally the new state S1 is generated by using output gate and the cell state.

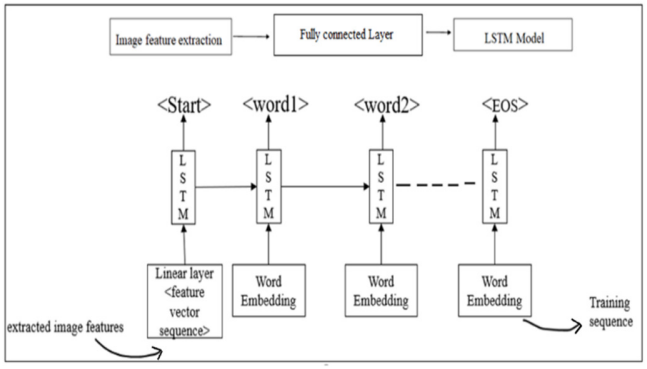


Fig. 7. Image vector to sequence the representation using the LSTM model.

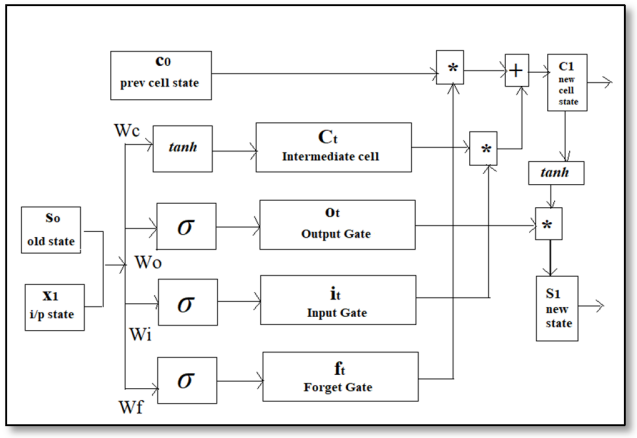


Fig. 8. Structure of LSTM.

Using the proposed model with LSTM, the generated captions for an input image given from the new dataset are shown in Figure 9. For evaluating the results of caption generation of the proposed model, comparative analysis of the existing approach was conducted, i.e. VGG with LSTM and Inception V3 model on Flickr-8K dataset and the new dataset.

D. Caption Generation with Text Extraction (Phase III)

Along with identifying different objects in the image, textual information present in the image is also important. This textual information can give more information about the image, like image background. Extracting the text from an image and including it in the generated image caption gives more details about the image. This approach is missing in the existing systems. In the proposed model, we present image caption generation along with text extraction. The pre-processing on the input image is carried out, the original image is converted to grayscale, and then Otsu's threshold method is used which produces a single intensity threshold that divides pixels into foreground and background. Then, dilation is applied on the threshold image, which is a mathematical morphological operation to find the contours as shown in Figure 10.

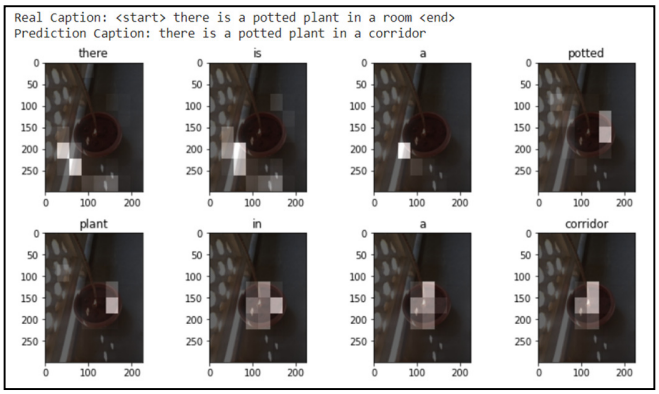


Fig. 9. Result of image captioning using the proposed model.

Fig. 10. Text extraction.

After looping through the contours, the rectangular part is cropped and passed to the Optical Character Recognition (OCR) system [17] for text extraction. With the deep learning approach, text is extracted with better accuracy [25]. Figure 10 shows some sample results of text extraction. The performance evaluation of text extraction is measured with precision and recall metrics [18]. A sample of precision and recall calculation follows.

System summary: There is plotted plant

Reference summary: There is potted plant in garden

Recall = number of overlapping words/total words in reference summary: (4/6=0.6)

Precision = number of overlapping words/total words in system summary: (4/4=1)

The results of precision and recall computations are shown in Figure 11 and the average result is summarized in Table III.

Image ID	Referance Text	System text	Recall	Precision
IMG_556	MAEER'S maharshtra institute of technology Department of computer engineering	MAEER'S maharshtra institute of technology Department of computer engineeri ng	0.89	0.80
IMG_557	vision of department to build a value based academic center of excellence in computer engineering	vision of department to build a value bas academic center of excellence in computer engin eering	0.87	0.81
IMG_558	Mission of department	Mission department	0.67	1.00
IMG_559	maharshtra institute of technology N block	technology N block inst	0.50	0.75
IMG_560	maharshtra institute of technology A block	maharshtra institute of technoloy A block	1.00	1.00
IMG_561	Third floor department of E & TC engineering	Third floor department of E TC eng	0.75	0.86
IMG_562	MAEER's maharshtra institute of technology, pune innovation center	maharshtra institut of technology, pune innovation cent	0.63	0.86
IMG_563	There is potted plant in garden	There is plotted plant	0.60	1.00

Fig. 11. Results of precision-recall computations for text extraction.

TABLE III. TEXT EXTRACTION EVALUATION

Metric used	Recall	Precision
Obtained result	0.75	0.86

IV. RESULTS

For evaluating the quality of the text of generated captions of the proposed system the utilized performance metrics are BLUE score [19], METEOR, CIDEr [20] which measures how much the machine generated output is closer to the human generated output, i.e. how similar is the candidate text to the reference text. It normally ranges between 0 and 1.

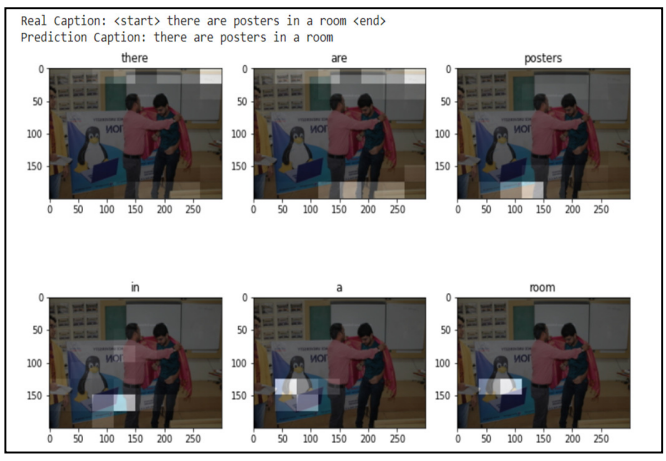


Fig. 12. Result of image captioning without text extraction.

Figures 12-13 show some results of image captioning without the text extraction module and Table IV shows the results of image captioning with the text extraction module using the proposed system, which is able to generate a more detailed description of the image. In order to check the performance of the proposed image captioning model, it was trained on Flickr-8k [21] dataset. Table V shows the performance of the proposed model for image captioning (without text extraction) on Flickr-8k. The obtained results exhibit a satisfactory performance. Comparative analysis of existing architectures like VGG-16 [22], Inception V3 [23], and the proposed model for image captioning (without the text extraction module) on the new dataset is shown in Table VI. It is seen that the proposed system outperforms the existing methods with improved BLEU score and METEOR score and also has an accuracy that reaches 87% for caption generation without text.

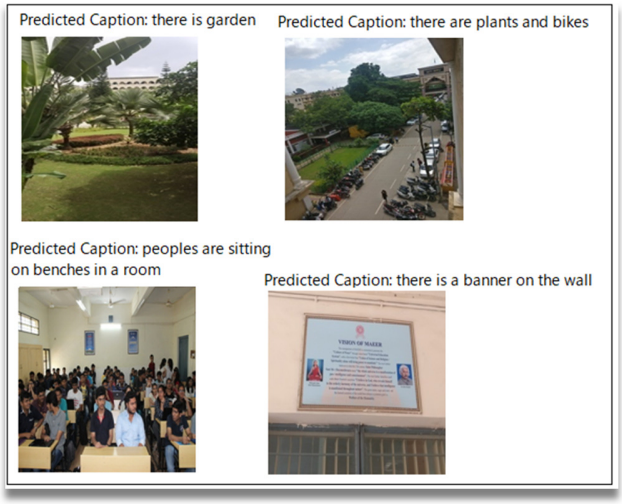


Fig. 13. Results of image captioning without text extraction.

TABLE IV. IMAGE CAPTIONING WITH TEXT EXTRACTION RESULTS

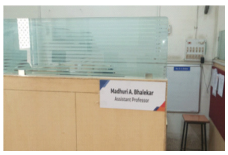
Input image	**Validation of the proposed model**
	Predicted caption is: there is a poster on the wall Extracted text: Madhuri A. Bhalekar Assistant Professor Generated captions is: there is a poster on the wall and the extracted text is Madhuri A. Bhalekar Assistant Professor
	Predicted Caption is: there is a poster Extracted text: MAEER'S MAHARASHRTA INSTITUTE OF TECHNOLOGY, PUNE 17/01/20008 Generated captions is: there is a poster and the text extracted is MAEER'S MAHARASHRTA INSTITUTE OF TECHNOLOGY, PUNE 17/01/20008

TABLE V. PERFORMANCE COMPARISON OF VGG-LSTM AND THE PROPOSED MODEL FOR IMAGE CAPTIONING WITHOUT TEXT EXTRACTION

Dataset	**Architecture**	**Evaluation metrics**				
		BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**
[21]	VGG with LSTM [22]	0.647	0.459	0.318	0.216	0.201
	Proposed model	0.495	0.352	0.276	0.201	0.425

The result of performance evaluation for caption generation with text extraction model is shown in Table VII. We see that the proposed model was an accuracy that reaches 83%, thus providing more information about the image. Further, the generated captions can be presented in audio form. Also, the model can be deployed on smart glass which will be useful for visually challenged individuals and can be used as a navigation system.

TABLE VI. COMPARATIVE ANALYSIS OF EXISTING ARCHITECTURES AND OF THE PROPOSED MODEL FOR IMAGE CAPTIONING WITHOUT TEXT EXTRACTION

Dataset	Architecture	Evaluation metrics				
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Our new dataset	VGG with LSTM [22]	0.644	0.522	0.441	0.374	0.644
	Inception V3 model [23]	0.534	0.452	0.343	0.253	0.534
	Proposed	0.879	0.752	0.743	0.678	0.653

TABLE VII. EVALUATION OF THE PROPOSED MODEL FOR IMAGE CAPTIONING WITH TEXT EXTRACTION

Dataset	Architecture	Evaluation metrics				
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Our new dataset	Proposed model with LSTM	0.832	0.645	0.672	0.611	0.589

V. CONCLUSION AND FUTURE WORK

This paper reviewed various methodologies used for image captioning from the domain of machine learning and neural networks. Various methodologies were analyzed for feature extraction, classification, and caption generation. A new dataset was implemented which trained further the proposed Deep CNN model. The paper also analyzed text extraction from the image and included the extracted text as a part of the generated image captions. Extracting the text from an image and including it in the image caption gives more details about the image. The proposed system outperforms the existing methods with the curated new dataset and also gives satisfactory performance on the benchmark Flickr8k dataset. Considering the recent advances in image captioning systems, generating more accurate image captioning and image description is still an open problem. Automatic image captioning will continue to be a major research topic for a while, due to the advancement of deep learning-based network architectures. Further, fine tuning the hyper-parameters of the proposed model can give better results. There is a huge impact of image captioning in many social relevance applications such as video surveillance, navigation for visually challenged individuals, online education systems, medical assistance, and many more which add significant value in our daily lives.

ACKNOWLEDGMENT

The authors would like to thank the authorities of MIT World Peace University for permitting and becoming a vital part of the new dataset and for supporting this research.

REFERENCES

[1] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "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*, vol. 39, no. 12, pp. 2321–2334, Dec. 2017, https://doi.org/10.1109/TPAMI.2016.2642953.

[2] G. Kulkarni *et al.*, "Baby talk: Understanding and generating simple image descriptions," in *CVPR 2011*, Colorado Springs, CO, USA, Jun. 2011, pp. 1601–1608, https://doi.org/10.1109/CVPR.2011.5995466.

[3] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664–676, Dec. 2017, https://doi.org/10.1109/TPAMI.2016.2598339.

[4] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain Images with Multimodal Recurrent Neural Networks," *arXiv:1410.1090 [cs]*, Oct. 2014, Accessed: Feb. 23, 2022. [Online]. Available: http://arxiv.org/abs/1410.1090.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3156–3164, https://doi.org/10.1109/CVPR.2015.7298935.

[6] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 2422–2431, https://doi.org/10.1109/CVPR.2015.7298856.

[7] K. Xu *et al.*, "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, Lille, France, Apr. 2015, pp. 2048–2057.

[8] M. Bhalekar, S. Sureka, S. Joshi, and M. Bedekar, "Generation of Image Captions Using VGG and ResNet CNN Models Cascaded with RNN Approach," in *Machine Intelligence and Signal Processing*, Singapore, 2020, pp. 27–42, https://doi.org/10.1007/978-981-15-1366-4_3.

[9] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, Jun. 2010, pp. 902–909, https://doi.org/10.1109/CVPR.2010.5540120.

[10] S. Nuanmeesri, "A Hybrid Deep Learning and Optimized Machine Learning Approach for Rose Leaf Disease Classification," *Engineering, Technology & Applied Science Research*, vol. 11, no. 5, pp. 7678–7683, Oct. 2021, https://doi.org/10.48084/etasr.4455.

[11] S. L. Sanga, D. Machuve, and K. Jomanga, "Mobile-based Deep Learning Models for Banana Disease Detection," *Engineering, Technology & Applied Science Research*, vol. 10, no. 3, pp. 5674–5677, Jun. 2020, https://doi.org/10.48084/etasr.3452.

[12] C. Szegedy, A. Toshev, and D. Erhan, "Deep Neural Networks for Object Detection," in *Advances in Neural Information Processing Systems*, 2013, vol. 26.

[13] X. Wang, Z. Zhu, C. Yao, and X. Bai, "Relaxed Multiple-Instance SVM with Application to Object Discovery," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Sep. 2015, pp. 1224–1232, https://doi.org/10.1109/ICCV.2015.145.

[14] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, 2014, pp. 740–755, https://doi.org/10.1007/978-3-319-10602-1_48.

[15] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Aug. 1997, https://doi.org/10.1162/neco.1997.9.8.1735.

[16] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, Jul. 2017, https://doi.org/10.1109/TNNLS.2016.2582924.

[17] G. A. Robby, A. Tandra, I. Susanto, J. Harefa, and A. Chowanda, "Implementation of Optical Character Recognition using Tesseract with the Javanese Script Target in Android Application," *Procedia Computer*

Science, vol. 157, pp. 499–505, Jan. 2019, https://doi.org/10.1016/j.procs.2019.09.006.

[18] F. Alotaibi, M. T. Abdullah, R. B. H. Abdullah, R. W. B. O. K. Rahmat, I. A. T. Hashem, and A. K. Sangaiah, "Optical Character Recognition for Quranic Image Similarity Matching," *IEEE Access*, vol. 6, pp. 554–562, 2018, https://doi.org/10.1109/ACCESS.2017.2771621.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, USA, Apr. 2002, pp. 311–318, https://doi.org/10.3115/1073083.1073135.

[20] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4566–4575, https://doi.org/10.1109/CVPR.2015.7299087.

[21] M. Hodosh, P. Young, and J. Hockenmaier, "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, Aug. 2013, https://doi.org/10.1613/jair.3994.

[22] C. Alippi, S. Disabato, and M. Roveri, "Moving Convolutional Neural Networks to Embedded Systems: The AlexNet and VGG-16 Case," in *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, Porto, Portugal, Apr. 2018, pp. 212–223, https://doi.org/10.1109/IPSN.2018.00049.

[23] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdu, China, Jun. 2017, pp. 783–787, https://doi.org/10.1109/ICIVC.2017.7984661.

[24] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, Nov. 2021, Art. no. 53, https://doi.org/10.1186/s40537-021-00444-8.

[25] B. Ahmed, G. Ali, A. Hussain, A. Baseer, and J. Ahmed, "Analysis of Text Feature Extractors using Deep Learning on Fake News," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 7001–7005, Apr. 2021, https://doi.org/10.48084/etasr.4069.

AUTHORS PROFILE

Madhuri Bhalekar is a Ph.D. research scholar in the School of Computer Engineering & Technology, MIT World Peace University, Pune, India. She is currently working as an Assistant Professor in the School of Computer Engineering & Technology, MIT World Peace University, Pune, India. Her research interests include image processing, computer vision, and natural language processing. ORCID ID: 0000-0003-1863-5293.

Mangesh Bedekar is currently working as a Professor in the School of Computer Engineering and Technology at MIT World Peace University, Pune, India. He received his M.Sc. and Ph.D. from BITS Palani, Rajasthan, India. He is a member of CSI, ISTE, IET, and ACM. His primary research interests include web data mining, web personalization, user interface design, user interface improvements, browser customization, affective computing, information cisualization, computer vision, and natural language processing. ORCID ID: 0000-0003-4461-9641