Image Captioning Using Improved YOLO V5 Model and Xception V3 Model

M. SAROJA
Sarah Tucker College
Ani Brown Mary (✉ anibrownvimalraj@gmail.com)
Sarah Tucker College https://orcid.org/0000-0002-6029-4472

Research Article

Keywords: Image captioning, object detection, YOLO V5, Xception V3, NLP

Posted Date: August 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2856192/v1

IMAGE CAPTIONING USING IMPROVED YOLO V5 MODEL AND XCEPTION V3 MODEL

M. SAROJA[1] and Dr N. ANI BROWN MARY[2]

[1]II M.Sc. Computer Science, Sarah Tucker College, Tirunelveli- 21spcssaroja@sarahtuckercollege.edu.in

[2]Assistant Professor, Computer Science, Sarah Tucker College, Tirunelveli- anibrownmarycs@sarahtuckercollege.edu.in

Corresponding Author : Dr N. ANI BROWN MARY

Funding Info

No financial support for this research, authorship, and/or publication of this article.

Conflicts of interest/Competing interests

No Conflict of Interest

Ethics approval

All applicable international, national, and/or institutional guidelines for the care and use of animals were followed. This manuscript does not contain any studies with human participants or animals performed by any of the two authors.

Consent to Participate (Ethics)

I understand that my contribution will be confidential and that there will be no personal identification in the data that I agree to allow to be used in the study.

Consent to Publish (Ethics)

Authors should make sure to also seek consent from individuals to publish their data prior to submitting their manuscript to this journal.

Authors' contributions

Conceived and designed the analysis - M. SAROJA
Collected the data - Dr. N. Ani Brown Mary
Contributed data or analysis tools - Dr. N. Ani Brown Mary
Performed the analysis - M. SAROJA
Wrote the paper - M. SAROJA

Availability of data and material

NO

IMAGE CAPTIONING USING IMPROVED YOLO V5 MODEL AND XCEPTION V3 MODEL

M. SAROJA[1] and Dr N. ANI BROWN MARY[2]

[1]II M.Sc. Computer Science, Sarah Tucker College, Tirunelveli- 21spcssaroja@sarahtuckercollege.edu.in

[2]Assistant Professor, Computer Science, Sarah Tucker College, Tirunelveli- anibrownmarycs@sarahtuckercollege.edu.in

ABSTRACT

Image captioning provides the process of describing the content from an image The task of generating image captions considers object detection for single-line descriptions. To improve the quality of the generated caption, object detection features are applied. In this proposed work, features are extracted from improved YOLO V5 model. This improved YOLO V5 model enhances the performance of the object detection process. Xception V3 model is applied to generate the sequence of the word from predicted object feature. Finally the caption generated from Xception V3 is used to hear in voice and text with any selected language. Flickr 8k, Flicr30k and MSCOCO data sets are used for this proposed method. Natural Language Processing (NLP) is a technique used to understand the description of an image. This proposed method is very much used for visually impaired people. The results show that the proposed method provides 99.5% Accuracy, 99.1% Precision , 99.3% Recall, 99.4% F1 score on MS COCO data set using improved YOLO V5 model and Xception V3 model. Compared to the existing techniques, this proposed method shows 11% to 15% improved accuracy.

Key words : Image captioning, object detection, YOLO V5, Xception V3, NLP

1. INTRODUCTION

1.1 IMAGE CAPTIONING

Image captioning provides the process of automatically describe the content from an image. The creation of textual descriptions for images is known as image captioning. The automatic generation of image captions is a more challenging task than object recognition and image classification. The captions are generated using both computer vision and natural language processing. In this proposed work, features are extracted from improved YOLO V5 model.

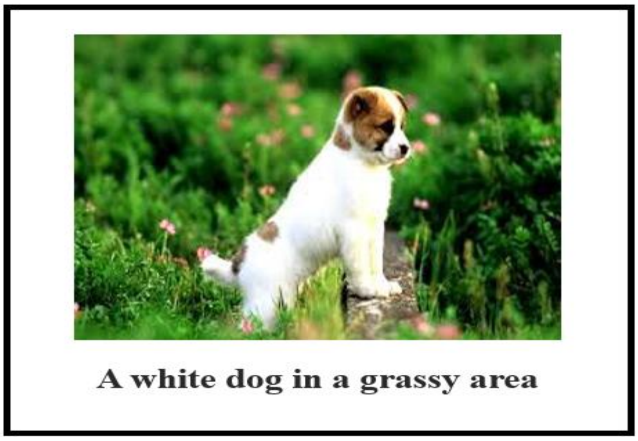


Figure 1 Example of image captioning

This improved YOLO V5 model enhances the performance of the object detection process. Xception V3 model is applied to generate the sequence of the word from predicted object feature. Finally the caption generated from Xception V3 model is used to hear in voice and text with any selected language. The combination of these two models used for more accurate and descriptive captions for images. Xception V3 is a modified version of the Inception V3 model that has fewer parameters and a higher accuracy. Flickr 8k, Flicr30k and MSCOCO data sets are used for this proposed method. These data sets are large collection of images with accompanying captions, which can be used to train a deep learning model to generate captions for new images. The deep learning model is trained using a combination of supervised and unsupervised learning techniques. The supervised learning involves training the model on a set of image-caption pairs, while the unsupervised learning involves training the model to predict the next word in a sequence of words. Natural Language Processing (NLP) is a technique used to understand the description of an image as shown in the Figure 1.

1.2 NATURAL LANGUAGE PROCESSING (NLP)

The computer only understands the language of binary values (0 and 1), it does not understand human languages like english or any other languages. The use of Natural Language Processing (NLP), the computer system can understand English or any other languages. The two main approaches in Natural Language Processing (NLP) are syntax and semantic analysis. Semantic analysis examines the grammar of sentences, including the placement of words, phrases, and clauses, to identify the connections between different items in a given context. When compared to accurate grammar rules, syntax analysis evaluates whether the content is meaningful. The main benefit of NLP, is that improves the way humans and computers communicate with each other. This proposed method is very much used for visually impaired people.

2. LITERATURE SURVEY

Researchers from various angles have proposed a large number of algorithms and strategies.

Vinyals et al. [1] have presented the neural image caption generator model based on a recurrent architecture incorporates the most latest events in object recognition with machine translation and can be utilized to produce natural language that describe images. Mathur et al. [2] have proposed a real-time image captioning generator using deep learning based on computer vision and machine translation. Here, describe a simplified encoder- and decoder based implementation of image captioning and its conventional methods, allowing us to execute these models on low-end hardware found in portable devices. Shuang Liu et al. [17] have proposed a multimodal Recurrent Neural Network (m-RNN) model that creatively combines the CNN and RNN models to resolve the captioning problem. LSTM model is an important type of structure of the RNN model that can solve the problem. Because the gradient disappearance and limited memory problem of conventional RNN. Three additional control units (cells), input, output, and forget gates, are added. The cells in the model will evaluate the information as it enters. Nonconforming material will be lost, while information that meets with the requirements will be preserved. The long and complicated sequential dependency issue in the neural network can be overcome using these method. Other works such as image captioning model [3,18,28] have presented the machine translation and image captioning model using CNN and RNN. Given its success in computer vision and the related studies that have been performed on machine translation, CNN has recently become a popular technique. These enhancements to the convolutional model can also be used for image captioning[21-30].

The decoder architectures that plays the image repeatedly to create the caption with the aid of machine translation and trained data stored use LSTM (Long Short Term Memory) or its updated version GRU - Gated Recurrent unit, which provides the developed caption as an output [4]. Akash Verma et al. [6] have proposed intelligence embedded image caption generator using LSTM based RNN model. Here the process of developing an LSTM-based RNN model for image captioning allows information to be scanned, extracted, and converted into a single-line sentence in english using natural language. Overfitting of data is considered to be difficult to prevent most of the time, but we are grateful to have solved that challenge. M.M. Ali Baig et al. [7] have proposed image caption generator with novel object injection. Here that utilized a pre-trained caption generator, image captioning with novel words injection attempts to inject objects into the caption that are not present in the dataset. Specifically, BLEU, CIDEr, and ROUGE-L are the usual metrics we use to evaluate the model. The outcomes exceed the underlying model both qualitatively and numerically. The basic CNN-RNN method by adding in image and word information at each timestep in their proposed Long-term Recurrent Convolutional Network (LRCN) [19]. N. Komal Kumar et al. [8] have proposed detection and recognition of objects in image caption generator system is demonstrated through experiments implementing the python programming language on the Flickr 8k data set.

S.H. Han et al. [9] have developed explainable image caption generator using attention and bayesian inference. Here, the explanation generator that creates a loss in the image-sentence relevance and affects how the training in the generation module is done. The explanation module also creates a weight matrix that shows the relationships between the regions taken from the provided image and the words in the created caption. Andrej Karpathy et al. [20] have proposed deep visual-semantic alignments for generating image descriptions model developed a paradigm for deep visual-semantic matching that produced descriptions of regions or images. This method begins by utilising an object detection technique to compute the values for locations before training a generative model with a multi-modal Recurrent Neural Network (m-RNN) using picture caption data and previously calculated scores. For an input region, a sentence is produced using the trained model. Image caption generation using deep learning technique model is developed to produces natural sentences that ultimately explain the image. Convolutional Neural Networks (CNN) and Recurrent Neural Networks make this model (RNN). CNN is used to extract picture features, and the RNN is used to generate sentences. The model have been developed to generate captions that, given an input image, almost correctly describe the image [10,11,12].

3. PROPOSED METHODOLOGY

In this proposed work correctly identifies the specific features and the relevant caption as shown in the Figure 2. In this proposed work, features are extracted from improved YOLO V5 model. The Prediction part of the YOLO V5 model consists of three convolution layers that predict the bounding boxes, the object scores and the class labels. In improved YOLO V5 model, one more prediction convolution layer is added. Convolution layer increases the accuracy of the model. It improves the performance of the model to detect object features. Then, it will feed into the Xception V3 model. Xception V3 model is used to generate a sequence of words, that correctly describe the image. Each word of the description will automatically aligned to different objects of the input image. After the description is processed, it will convert into voice and text with any selected language. The proposed system architecture is described as shown in the Figure 2.

3.1 ARCHITECTURE DESIGN

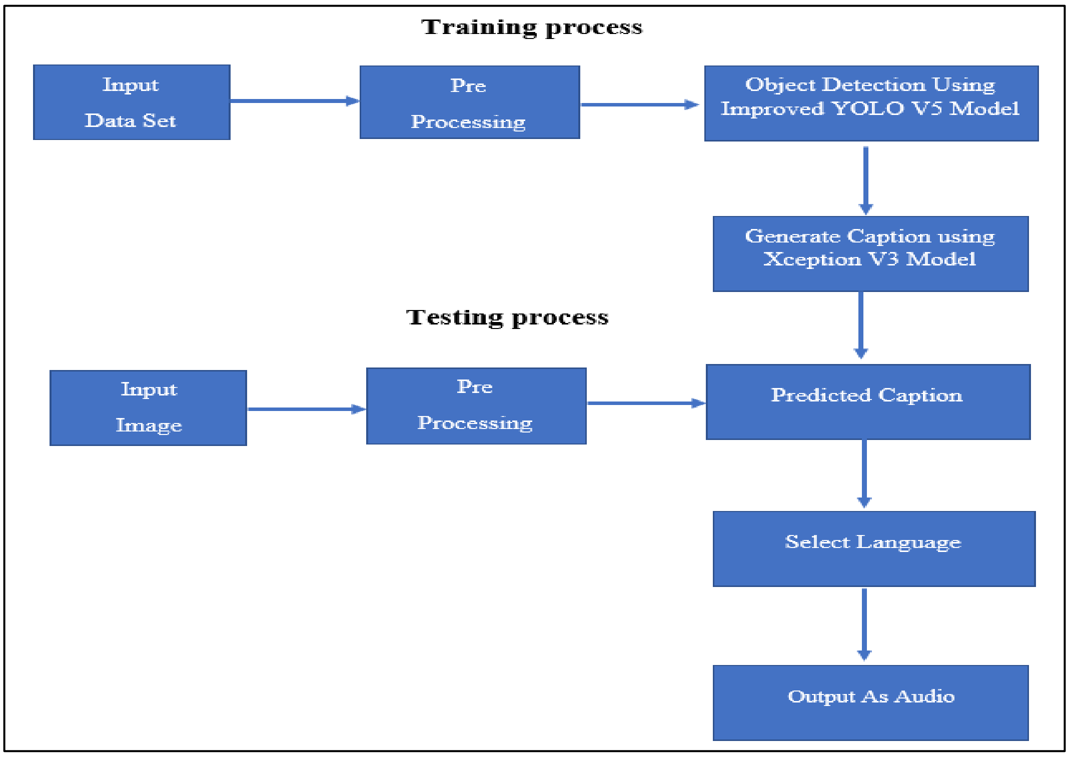


Figure 2 The proposed system architecture

3.3 PRE PROCESSING

A crucial stage for the pre-processing is image resizing. Image interpolation technique used for resize image from one pixel grid to another.

3.4 OBJECT DETECTION USING IMPROVED YOLO V5 MODEL

The YOLO V5 model is used to detect object features. The YOLO V5 model is a single-stage object detector. This model consists of three components : Backbone, Neck, Prediction.

The Backbone is a pre-trained network. It is used to extract rich feature representation for images. CSP-Darknet53 as a backbone. This helps reducing the spatial resolution of the image and increasing its feature (channel) resolution.

Spatial Pyramid Pooling (SPP) and Path Aggregation Network are used to extract feature pyramids from the model neck. This improves the model's good generalization to objects of various sizes and scales.

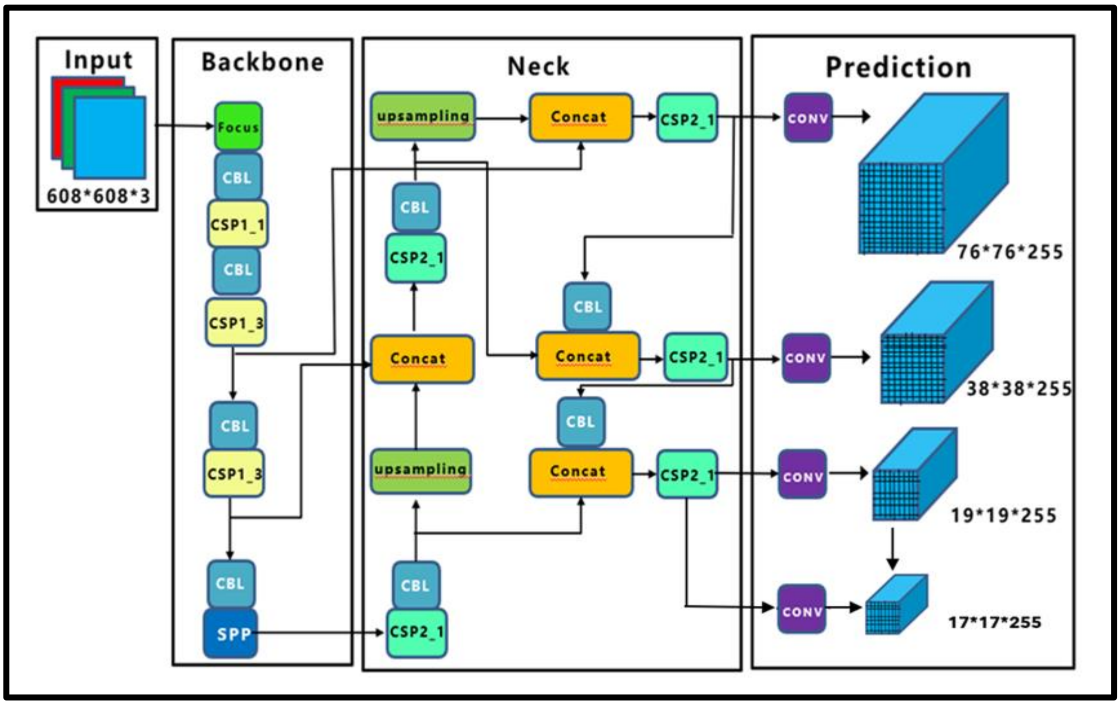


Figure 3 Improved YOLO V5 model Architecture

The prediction part it composed from three convolution layers that predicts the location of the bounding boxes, objects scores and objects classes. Improved YOLO V5 model one more prediction convolution layer is added. Convolution layer it increase the accuracy of the model. It improve the performance of the model to detect object features. Improved YOLO V5 model architecture as shown in the Figure 3.

3.5 GENERATE CAPTION USING XCEPTION V3 MODEL AND OUTPUT

Predicted object features is fed into the Xception V3 model. Predicted object features will generate a sequence of words that correctly describe the image. Xception V3 uses depth wise separable convolutions extensively throughout the network, which allows it to achieve state-of-the-art accuracy on image classification benchmarks such as ImageNet. Xception V3 also includes other advanced techniques, such as residual connections and batch normalization, which further improve its performance.

Xception V3 has been used in a wide range of applications beyond image classification, including object detection, semantic segmentation, and image captioning. The architecture has also inspired the development of other models, such as MobileNet and EfficientNet, which use similar depth wise separable convolution techniques to achieve high performance with fewer parameters and computations. Xception V3 model contains four parts. Convolution layer, global average pooling layer, fully connected layer, prediction.

Convolution Layer : The Xception V3 architecture uses convolutional layers to extract features from the input. The feature extraction process involves applying a series of convolutional filters to the input to extract features at different levels of abstraction.

Global Average Pooling : Once the features have been extracted, the Xception V3 architecture performs a global average pooling operation to summarize the features and reduce the dimensionality of the data. This involves taking the average of the feature maps across all spatial locations.

Fully Connected Layers : The output of the global average pooling layer is fed into a fully connected layer, which applies a linear transformation to the features and maps them to a set of intermediate representations.

Prediction: Finally, the output of the fully connected layer is passed through a SoftMax activation function to generate a probability distribution over the possible output classes. The class with the highest probability is chosen as the predicted output. Finally the caption is generated from Xception V3 model. It is used to hear in voice and text with any selected language.

4.   RESULT AND DISCUSSION

4.1 DATA SET DETAILS

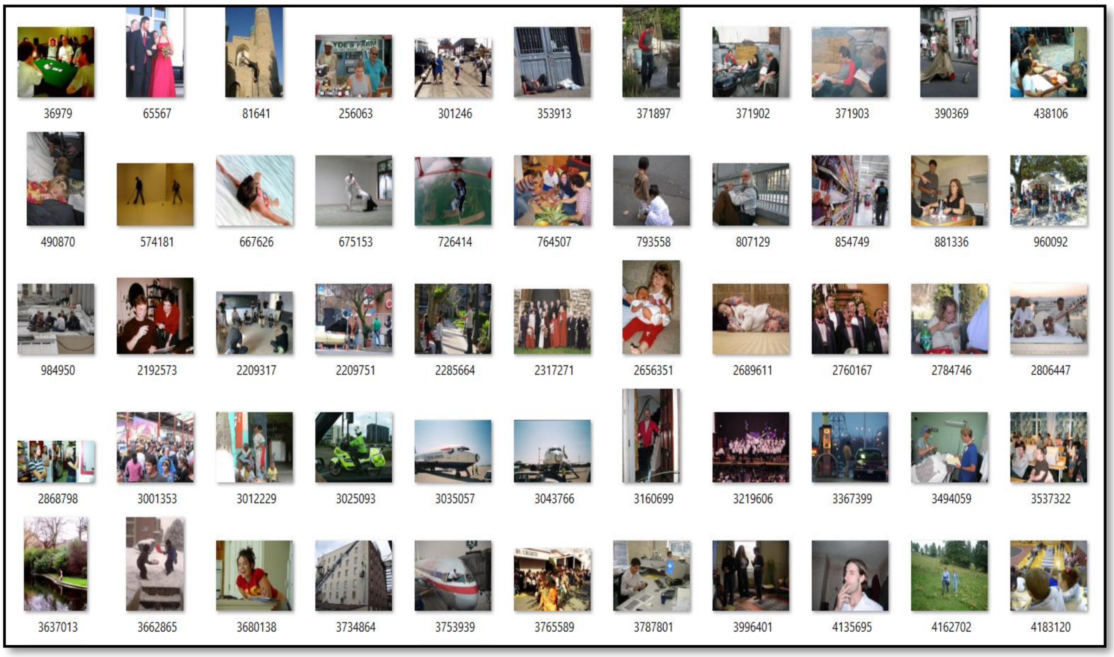


Figure 4 Data sets details

Flickr 8k Data Set, Flickr 30K Data Set, MSCOCO Data Sets are used for this work. Flickr8k Data Set contains 8091 images with five english captions per image. This Data Set is available on Kaggle website and have a size on 1GB. This data set has over 31,000 images. Each image in data set has five reference sentences provided by human annotators. MS COCO data collection, which stands for "Microsoft Common Objects in Context" is a large-scale object detection, segmentation, key-point detection, and captioning data set. This data set consists of 328K images. These data sets details as shown in Figure the 4.

4.2 PERFORMANCE ANALYSIS

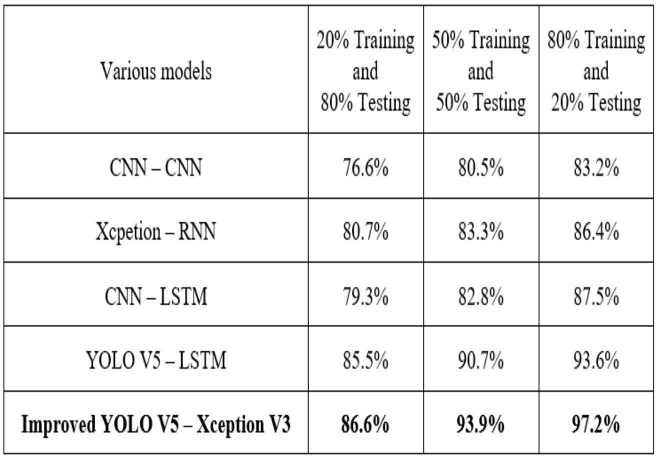
In this project, various deep learning algorithms like Improved YOLO V5 and Xception V3 algorithms are used to generate the captions from images. For evaluation metrics like Accuracy, Precision, Recall, F1 Score are considered [31-50].

4.3 RESULT

4.3.1 COMPARISON OF ACCURACY

Accuracy is compared in three forms of training and testing, namely 20% of training and 80% of testing. 50% of training and 50% of testing, 80% of training and 20% of testing and these training models are compared on Flickr8k data set, Flickr30k data set and MS COCO data set.

Table 1 Comparison of accuracy on Flickr8k data set



Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	76.6%	80.5%	83.2%
Xcpetion – RNN	80.7%	83.3%	86.4%
CNN – LSTM	79.3%	82.8%	87.5%
YOLO V5 – LSTM	85.5%	90.7%	93.6%
Improved YOLO V5 – Xception V3	**86.6%**	**93.9%**	**97.2%**

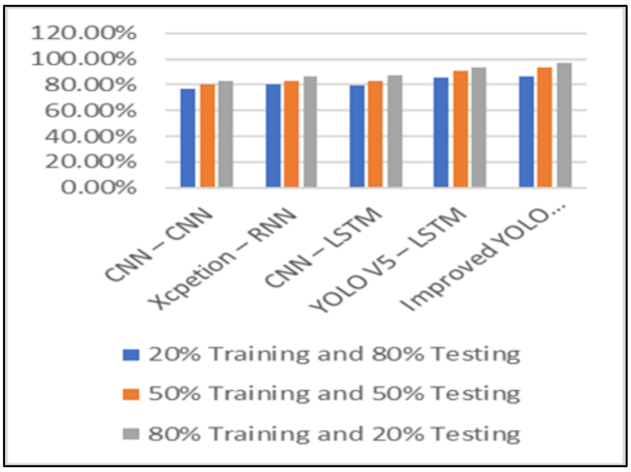
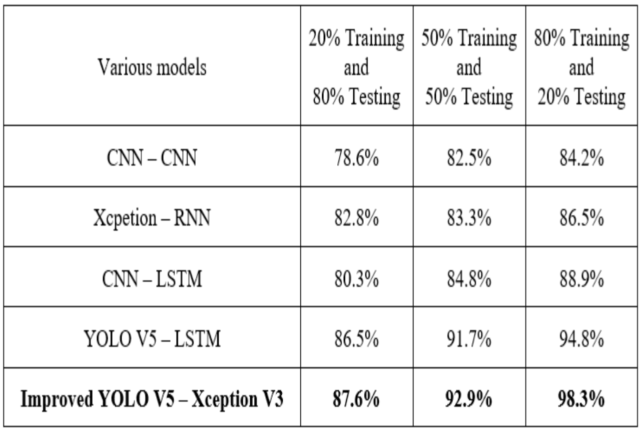


Figure 5 Comparison of accuracy on Flickr8k data set

The comparison of accuracy on flickr8k data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 97.2% accuracy is a best training model compared to existing models like CNN – CNN with 83.2%, Xception – RNN with 86.4%, CNN - LSTM with 87.5%, YOLO V5 – LSTM with 93.6% on flickr8k data set. The comparison of accuracy on flickr8k data set as shown in the Figure 5 and as shown in the Table 1.

Table 2 Comparison of accuracy on Flickr30k data set



Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	78.6%	82.5%	84.2%
Xcpetion – RNN	82.8%	83.3%	86.5%
CNN – LSTM	80.3%	84.8%	88.9%
YOLO V5 – LSTM	86.5%	91.7%	94.8%
Improved YOLO V5 – Xception V3	**87.6%**	**92.9%**	**98.3%**

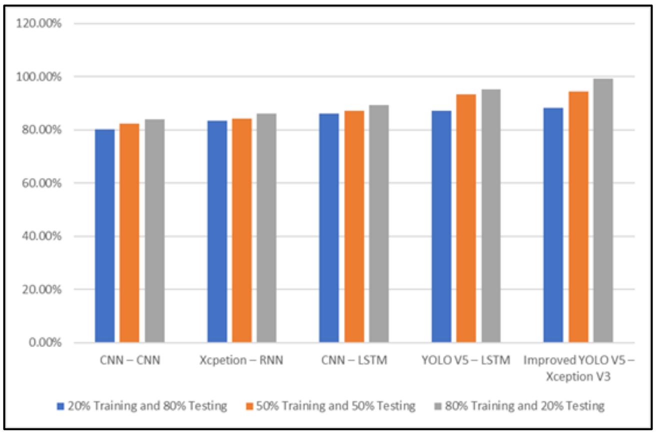


Figure 6 Comparison of accuracy on Flickr30k data set

The comparison of accuracy on flickr30k data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 98.3% accuracy is a best training model compared to existing models like CNN – CNN with 84.2%, Xception – RNN with 86.5%, CNN - LSTM with 88.9%, YOLO V5 – LSTM with 94.5% on flickr30k data set. The comparison of accuracy on flickr30k data set as shown in the Figure 2 and as shown in the Table 6.

Table 3 Comparison of accuracy on MS COCO data set

Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	80.6%	82.5%	84.2%
Xcpetion – RNN	83.8%	84.9%	86.5%
CNN – LSTM	86.3%	87.8%	89.9%
YOLO V5 – LSTM	87.5%	93.7%	95.8%
Improved YOLO V5 – Xception V3	**88.6%**	**94.9%**	**99.5%**

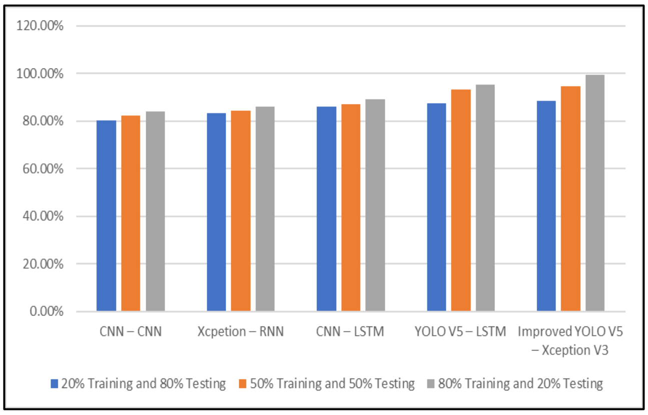


Figure 7 Comparison of accuracy on MS COCO data set

The comparison of accuracy on MS COCO data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 99.5% accuracy is a best training model compared to existing models like CNN – CNN with 84.2%, Xception – RNN with 86.5%, CNN - LSTM with 89.9%, YOLO V5 – LSTM with 95.8% on MS COCO data set. The comparison of accuracy on MS COCO data set as shown in the Figure 7 and as shown in the Table 3.

The overall comparison of accuracy results shows that the 80% of training and 20% of testing with 99.5% accuracy on MS COCO data set is best training model compared to others.

4.3.2 COMPARISON OF PRECISION

Precision score is compared in three forms of training and testing, namely 20% of training and 80% of testing. 50% of training and 50% of testing, 80% of training and 20% of testing and these training models are compared on Flickr8k data set, Flickr30k data set and MS COCO data set.

Table 4 Comparison of precision score on Flickr8k data set

Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	74.6%	80.5%	81.2%
Xcpetion – RNN	80.7%	82.3%	85.4%
CNN – LSTM	78.3%	82.3%	86.5%
YOLO V5 – LSTM	84.5%	90.1%	92.6%
Improved YOLO V5 – Xception V3	**86.2%**	**93.2%**	**96.9%**

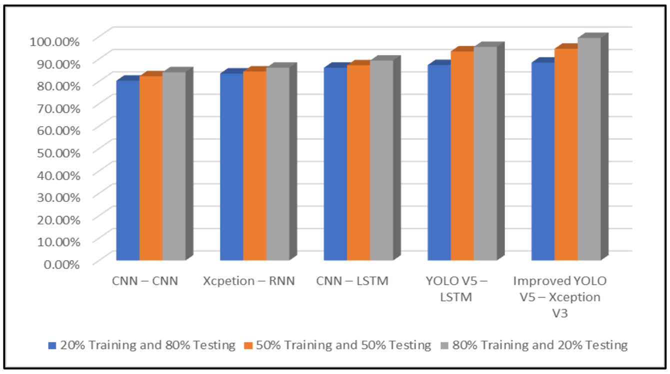
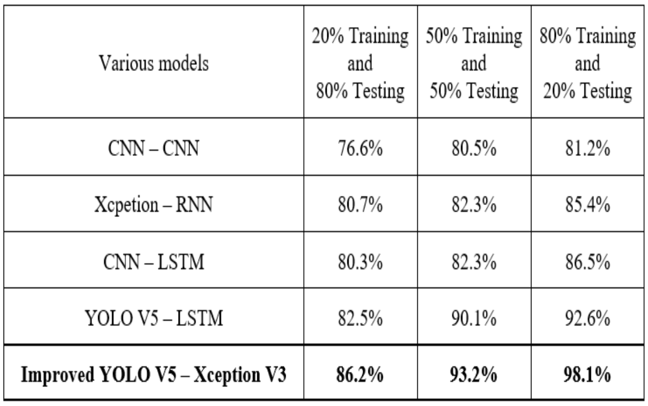


Figure 8 Comparison of precision score on Flickr8k data set

The comparison of precision score on flickr8k data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 96.9% precision score is a best training model compared to existing models like CNN – CNN with 81.2%, Xception – RNN with 85.4%, CNN - LSTM with 86.5%, YOLO V5 – LSTM with 92.6% on flickr8k data set. The comparison of precision score on flickr8k data set as shown in the Figure 4 and as shown in the Table 8.

Table 5 Comparison of Precision score on Flickr30k data set



Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	76.6%	80.5%	81.2%
Xcpetion – RNN	80.7%	82.3%	85.4%
CNN – LSTM	80.3%	82.3%	86.5%
YOLO V5 – LSTM	82.5%	90.1%	92.6%
Improved YOLO V5 – Xception V3	**86.2%**	**93.2%**	**98.1%**

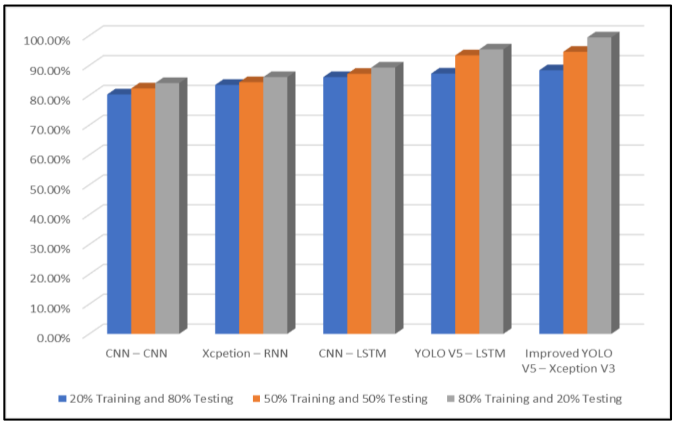


Figure 9 Comparison of Precision score on Flickr30k data set

The comparison of precision score on flickr30k data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 98.1% precision score is a best training model compared to existing models like CNN – CNN with 81.2%, Xception – RNN with 85.4%, CNN - LSTM with 86.5%, YOLO V5 – LSTM with 92.6% on flickr8k data set. The comparison of precision score on flickr8k data set as shown in the Figure 9 and as shown in the Table 5.

Table 6 Comparison of Precision score on MS COCO data set

Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	80.6%	82.5%	84.2%
Xcpetion – RNN	83.8%	84.9%	86.5%
CNN – LSTM	86.3%	87.8%	89.9%
YOLO V5 – LSTM	87.5%	93.7%	95.8%
Improved YOLO V5 – Xception V3	**88.6%**	**94.9%**	**99.1%**

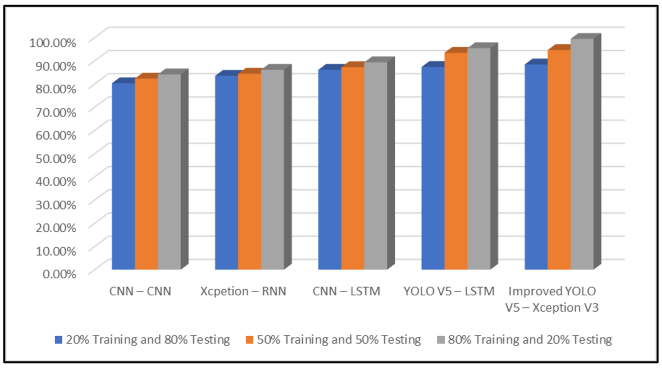


Figure 10 Comparison of Precision score on MS COCO data set

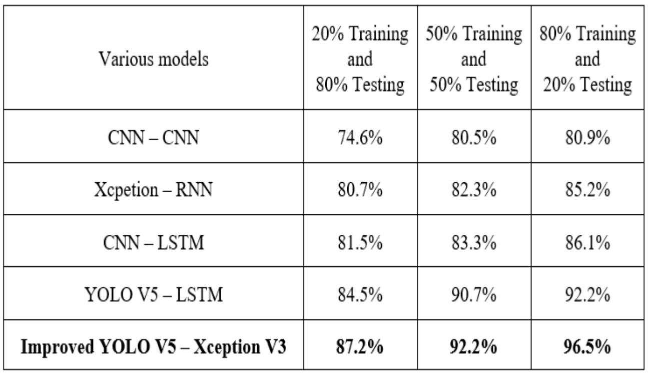
The comparison of precision score on MS COCO data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 99.1% precision score is a best training model compared to existing models like CNN – CNN with 84.2%, Xception – RNN with 86.5%, CNN - LSTM with 89.9%, YOLO V5 – LSTM with 95.8% on MS COCO data set. The comparison of precision score on MS COCO data set as shown in the Figure 10 and as shown in the Table 6.

The overall comparison of precision score results shows that the 80% of training and 20% of testing with 99.1% precision score on MS COCOC data set is best training model compared to others.

4.3.3 COMPARISON OF RECALL

Recall score is compared in three forms of training and testing, namely 20% of training and 80% of testing. 50% of training and 50% of testing, 80% of training and 20% of testing and these training models are compared on Flickr8k data set, Flickr30k data set and MS COCO data set.

Table 7 Comparison of recall score on Flickr8k data set



Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	74.6%	80.5%	80.9%
Xcpetion – RNN	80.7%	82.3%	85.2%
CNN – LSTM	81.5%	83.3%	86.1%
YOLO V5 – LSTM	84.5%	90.7%	92.2%
Improved YOLO V5 – Xception V3	**87.2%**	**92.2%**	**96.5%**

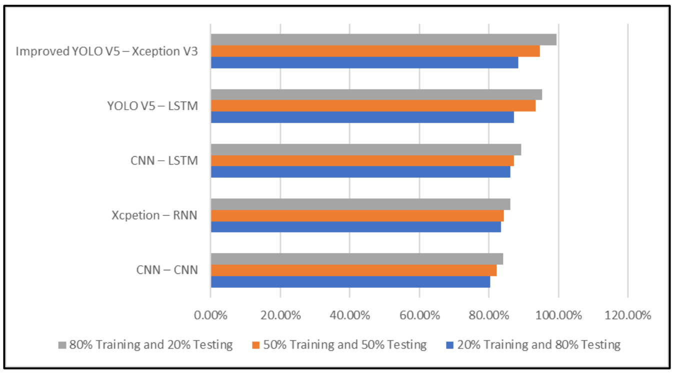
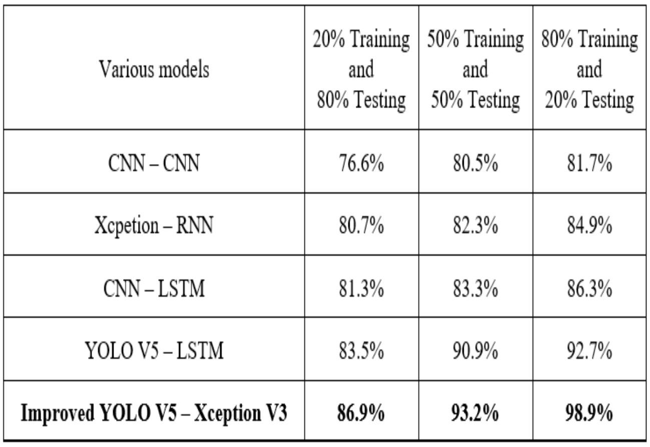


Figure 11 Comparison of recall score on Flickr8k data set

The comparison of recall score on flickr8k data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 96.5% recall score is a best training model compared to existing models like CNN – CNN with 80.9%, Xception – RNN with 85.2%, CNN - LSTM with 86.1%, YOLO V5 – LSTM with 92.2% on flickr8k data set. The comparison of recall score on flick8k data set as shown in the Figure 11 and as shown in the Table 7.

Table 8 Comparison of recall score on Flickr30k data set



Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	76.6%	80.5%	81.7%
Xcpetion – RNN	80.7%	82.3%	84.9%
CNN – LSTM	81.3%	83.3%	86.3%
YOLO V5 – LSTM	83.5%	90.9%	92.7%
Improved YOLO V5 – Xception V3	**86.9%**	**93.2%**	**98.9%**

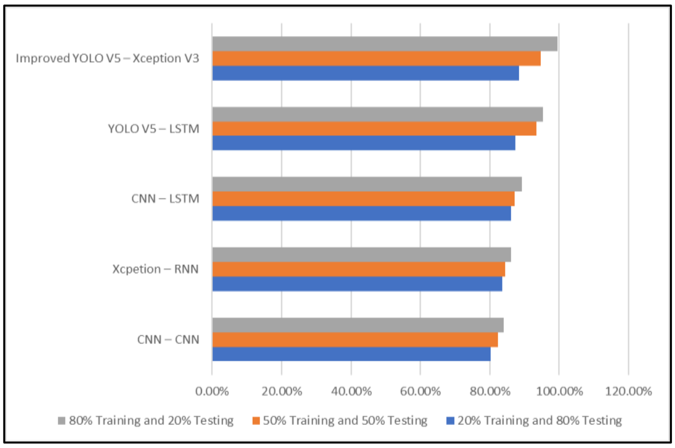


Figure 12 Comparison of recall score on Flickr30k data set

The comparison of recall score on flickr30k data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 98.9% recall score is a best training model compared to existing models like CNN – CNN with 81.7%, Xception – RNN with 84.9%, CNN - LSTM with 86.3%, YOLO V5 – LSTM with 92.7% on flickr30k data set. The comparison of recall score on flick30k data set as shown in the Figure 12 and as shown in the Table 8.

Table 9 Comparison of recall score on MS COCO data set

Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	81.6%	82.5%	85.2%
Xcpetion – RNN	83.8%	84.9%	87.5%
CNN – LSTM	85.3%	87.8%	88.9%
YOLO V5 – LSTM	87.5%	91.7%	94.8%
Improved YOLO V5 – Xception V3	**88.6%**	**95.9%**	**99.3%**

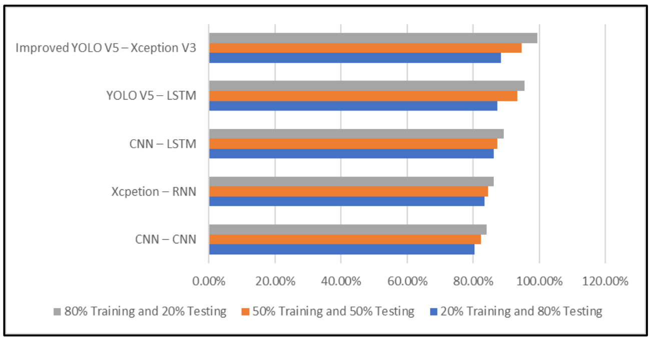


Figure 13 Comparison of recall score on MS COCO data set

The comparison of recall score on MS COCO data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 99.3% recall score is a best training model compared to existing models like CNN – CNN with 85.2%, Xception – RNN with 87.5%, CNN - LSTM with 88.9%, YOLO V5 – LSTM with 94.8% on MS COCO data set. The comparison of recall score on MS COCO data set as shown in the Figure 13 and as shown in the Table 9.

The overall comparison of recall score results shows that the 80% of training and 20% of testing with 99.3% recall score on MS COCO data set is best training model compared to others.

4.3.4 COMPARISON OF F1 SCORE

F1 Score is compared in three forms of training and testing, namely 20% of training and 80% of testing. 50% of training and 50% of testing, 80% of training and 20% of testing and these training models are compared on Flickr8k data set, Flickr30k data set and MS COCO data set.

Table 10 Comparison of F1 Score on Flickr8k data set

Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	76.6%	79.5%	83.2%
Xcpetion – RNN	80.7%	81.8%	86.4%
CNN – LSTM	81.3%	82.8%	87.5%
YOLO V5 – LSTM	85.5%	93.7%	93.6%
Improved YOLO V5 – Xception V3	**86.6%**	**93.5%**	**97.8%**

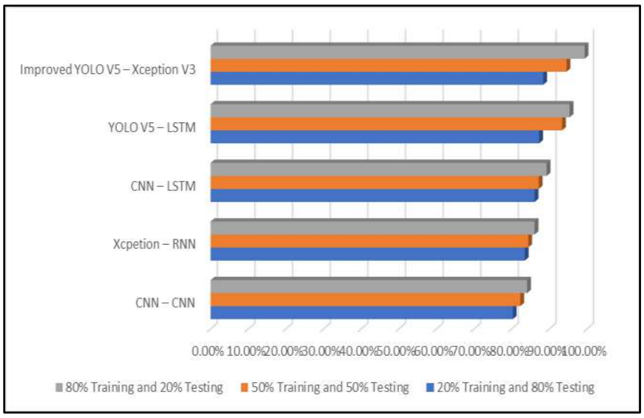


Figure 14 Comparison of F1 Score on Flickr8k data set

The comparison of F1 score on flickr8k data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 99.3% F1 score is a best training model compared to existing models like CNN – CNN with 83.2%, Xception – RNN with 86.4%, CNN - LSTM with 87.5%, YOLO V5 – LSTM with 93.6% on flickr8k data set. The comparison of F1 score on flickr8k data set as shown in the Figure 14 and as shown in the Table 10.

Table 11 Comparison of F1 score on Flickr30k data set

Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	78.6%	82.4%	83.9%
Xcpetion – RNN	82.8%	82.9%	86.2%
CNN – LSTM	83.3%	84.8%	88.6%
YOLO V5 – LSTM	86.9%	92.7%	94.2%
Improved YOLO V5 – Xception V3	**87.6%**	**92.9%**	**98.2%**

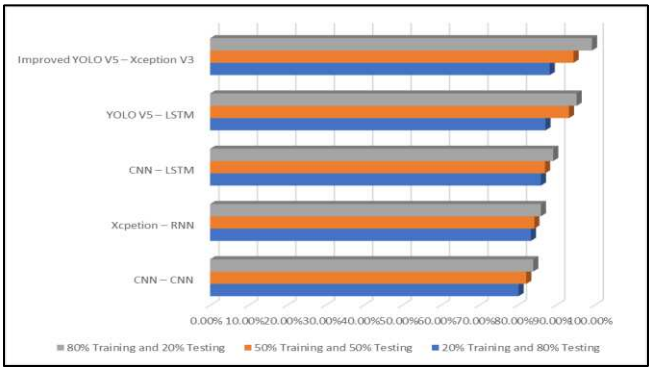


Figure 15 Comparison of F1 score on Flickr30k data set

The comparison of F1 score on flickr30k data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 98.2% F1 score is a best training model compared to existing models like CNN – CNN with 83.9%, Xception – RNN with 86.2%, CNN - LSTM with 88.6%, YOLO V5 – LSTM with 94.2% on flickr30k data set. The comparison of F1 score on flickr30k data set as shown in the Figure 15 and as shown in the Table 11.

Table 12 Comparison of F1 Score on MSCOCO data set

Various models	20% Training and 80% Testing	50% Training and 50% Testing	80% Training and 20% Testing
CNN – CNN	80.3%	82.3%	84.1%
Xcpetion – RNN	83.5%	84.4%	86.1%
CNN – LSTM	86.1%	87.2%	89.3%
YOLO V5 – LSTM	87.3%	93.4%	95.4%
Improved YOLO V5 – Xception V3	**88.4%**	**94.6%**	**99.4%**

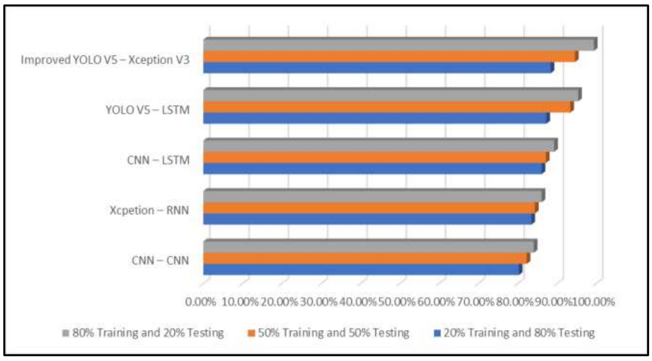


Figure 16 Comparison of F1 Score on MSCOCO data set

The comparison of F1 score on MS COCO data set results shows that the 80% of training and 20% of testing using Improved YOLO V5 – Xception V3 with 99.4% F1 score is a best training model compared to existing models like CNN – CNN with 84.1%, Xception – RNN with 86.1%, CNN - LSTM with 89.3%, YOLO V5 – LSTM with 95.4% on flickr30k data set. The comparison of F1 score on flickr30k data set as shown in the Figure 16 and as shown in the Table 12.

The overall comparison of F1 score results shows that the 80% of training and 20% of testing with 99.4% F1 score on MS COCO data set is best training model compared to others.

5. CONCLUSION

In conclusion, the combination of Improved YOLO V5 and Xception V3 powerful approaches is used for image captioning. Improved YOLO V5 model used for object detection and Xception V3 model used for generate caption, it is more accurate and descriptive representation of the image. However, it is important to note that building an image captioning system is a complex task that requires expertise in both computer vision and natural language processing. It involves not only designing and training the neural networks but also preprocessing the image and text data, fine-tuning the hyperparameters, and evaluating the performance of the system. Therefore, it is recommended to use existing libraries and frameworks, such as TensorFlow, PyTorch, or Keras, that provide pre-trained models and tools used for building image captioning systems. The results show that the proposed method provides 99.5% Accuracy, 99.1% Precision, 99.3% Recall, 99.4% F1 score on MS COCO data set using improved YOLO V5 model and Xception V3 model.

6. FUTURE ENHANCEMENT

In future, this work can be enhanced by developing a Zero-Shot Captioning. Zero-Shot Captioning is the ability to generate captions for images of novel objects or scenes that were not present in the training data. Developing techniques for zero-shot captioning could help to improve the generalization ability of image captioning models and make them more versatile.

7. REFERENCES

[1] Vinyals, Oriol, et al. "Show and Tell: A Neural Image Caption Generator." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 3156–64. DOI.org (Crossref), https://doi.org/10.1109/CVPR.2015.7298935.

[2] Mathur, Pranay, et al. "Camera2Caption: A Real-Time Image Caption Generator." 2017 International Conference on Computational Intelligence in Data Science(ICCIDS), IEEE, 2017, pp. 1–6. DOI.org (Crossref), https://doi.org/10.1109/ICCIDS.2017.8272660.

[3] Liu, Shuang, et al. "Image Captioning Based on Deep Neural Networks." MATEC Web of Conferences, edited by Yansong Wang, vol. 232, 2018, p. 01052. DOI.org (Crossref), https://doi.org/10.1051/matecconf/201823201052.

[4] Geetha, G., et al. "Image Captioning Using Deep Convolutional Neural Networks (CNNs)." Journal of Physics: Conference Series, vol. 1712, no. 1, Dec. 2020, p. 012015. DOI.org (Crossref), https://doi.org/10.1088/1742-6596/1712/1/012015.

[5] R, Mohana Priya, et al. "Building A Voice Based Image Caption Generator with Deep Learning." 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2021, pp. 943–48. DOI.org (Crossref), https://doi.org/10.1109/ICICCS51141.2021.9432091.

[6] Verma, Akash, et al. "Intelligence Embedded Image Caption Generator Using LSTM Based RNN Model." 2021 6th International Conference on Communication and Electronics Systems (ICCES), IEEE, 2021, pp. 963–67. DOI.org (Crossref), https://doi.org/10.1109/ICCES51350.2021.9489253.

[7] Baig, Mirza Muhammad Ali, et al. "Image Caption Generator with Novel Object Injection." 2018 Digital Image Computing: Techniques and Applications (DICTA), IEEE, 2018, pp. 1–8. DOI.org (Crossref), https://doi.org/10.1109/DICTA.2018.8615810.

[8] Kumar, N. Komal, et al. "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach." 2019 5th International Conference on Advanced Computing &

Communication Systems (ICACCS), IEEE, 2019, pp. 107–09. DOI.org (Crossref), https://doi.org/10.1109/ICACCS.2019.8728516.

[9] Han, Seung-Ho, and Ho-Jin Choi. “Explainable Image Caption Generator Using Attention and Bayesian Inference.” 2018 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, 2018, pp. 478–81. DOI.org (Crossref), https://doi.org/10.1109/CSCI46756.2018.00098.

[10] Chetan Amrikar and Vaishali Jabade. “Image caption using deep learning techniques”. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), IEEE 2018.

[11] Wang, Haoran, et al. “An Overview of Image Caption Generation Methods.” Computational Intelligence and Neuroscience, vol. 2020, Jan. 2020, pp. 1–13. DOI.org (Crossref), https://doi.org/10.1155/2020/3062706.

[12] Kabra, Palak, et al. “Image Caption Generator Using Deep Learning.” International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 10, Oct. 2022, pp. 621–26. DOI.org (Crossref), https://doi.org/10.22214/ijraset.2022.47058.

[13] Lakshminarasimha Srinivasan et al. “Image Captioning – A Deep Learning Approach”. 2018 International Journal of Applied Engineering Research ISSN 0973-4562 vol 13, no 9, pp.7239-7242.

[14] Masotti, Caterina, et al. “Deep Learning for Automatic Image Captioning in Poor Training Conditions.” Italian Journal of Computational Linguistics, vol. 4, no. 1, June 2018, pp. 43–55. DOI.org (Crossref), https://doi.org/10.4000/ijcol.538.

[15] Aishwarya Maroju et al. “Image Caption Generating Deep Learning Model”. 2021 International Journal of Engineering Research & Technology (IJERT). ISSN : 2278-0181, vol 10, issue 09, sep 2021.

[16] Al-Malla, Muhammad Abdelhadie, et al. “Image Captioning Model Using Attention and Object Features to Mimic Human Image Understanding.” Journal of Big Data, vol. 9, no. 1, Dec. 2022, p. 20. DOI.org (Crossref), https://doi.org/10.1186/s40537-022-00571-w.

[17] Weivu et al. “Explain Image with Multimodal Recurrent Neural Networks”.2014.

[18] Aneja, Jyoti, et al. Convolutional Image Captioning. 2017. DOI.org (Datacite), https://doi.org/10.48550/ARXIV.1711.09151.

[19] Jeff Donahue et al. “Long-term Recurrent Convolutional Network for Visual Recognition and Description”. 2015 IEEE.

[20] Andrej Karpathy and LiFei-Fei. “Deep Visual – sematic Alignments for Generating Image Descriptions.

[21] R. Kiros, R. Zemel, and R. Salakhutdinov. Multimodal neural language models. In ICML, 2014.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neuralnetworks. In NIPS, pages 1097–1105, 2012.

[23] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In CVPR, 2011.

[24] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Muller. Efficient backprop. In ¨ Neural networks: Tricks of the trade, pages 9–48. Springer, 2012.

[25] C.-Y. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In ACL, page 605, 2004.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[27] T. Mikolov, M. Karafiat, L. Burget, J. Cernock ´ y, and S. Khudanpur. Recurrent neural network based language model. In INTERSPEECH, pages 1045–1048, 2010.

[28] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur. Extensions of recurrent neural network language model. In ICASSP, pages 5528–5531, 2011.

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, pages 3111–3119, 2013.

[30] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daume III. Midge: Generating image descriptions from computer vision detections. In ´ EACL, 2012.

[31] N.Ani Brown Mary, "Profit Maximization For Saas Using SLA Based Spot Pricing in Cloud Computing", published in the Proceedings of the International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013).

[32] N.Ani Brown Mary, "Profit Maximization for Service Providers using Hybrid Pricing in Cloud Computing" published in the Proceedings of the International Journal of Computer Applications Technology and Research Volume 2, Issue 3, 218 - 223, 2013.

[33] N.Ani Brown Mary et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 1-5.

[34] Ani Brown Mary, N.: Profit maximization for SAAS using SLA based SPOT PRICING in CLOUD COMPUTING. Int. J. Emerg. Technol. Adv. Eng. **3**(1), 19–25 (2013).

[35] Ani Brown Mary, N., Saravanan, K.: Performance factors of CLOUD COMPUTING data centers using [(M/G/1):(/GDMODEL)] queuing systems. Int. J. Grid Comput. Appl. **4**(1), 1–9 (2013).

[36] Ani Brown Mary, N.: Profit maximization for service providers using hybrid pricing in cloud computing. Int. J. Comput. Appl. Technol. Res. **2**(3), 218–223 (2013).

[37] Ani Brown Mary, N., Jayapriya, K.: An extensive survey on QoS in cloud computing. Int. J. Comput. Sci. Inf. Technol. **5**(1), 1–5 (2014).

[38] Ani Brown Mary N, Dejey D (2017) Classification of coral reef submarine images and videos using a novel Z with tilted Z local binary pattern (Z⊕TZLBP). Wirel Pers Commun. https://doi.org/10.1007/s11277-017-4981-x.

[39] Ani Brown Mary N, Dharma D (2017) Coral reef image classification employing Improved LDP for feature extraction. Elsevier J Vis Commun Image Represent 49:225–242.

[40] Ani Brown Mary N, Dharma D, 2018, Coral reef image/video classification employing novel octa-angled pattern for triangular sub region and pulse coupled convolutional neural network (PCCNN). Multimed Tools Appl. https://doi.org/10.1007/s11042-018-6148-5, pp 1–35. Print-ISSN: 13807501, E-ISSN: 14321882.

[41] Mary AB, Dharma D (2018) Classification of coral reef submarine images and videos using a novel z with tilted z local binary pattern. Springer Wireless Personnel Communications 98(3):2427–2459

[42] Mary AB, Dharma D (2019) A novel framework for real-time diseased coral reef imageclassification. Springer, Multimedia Tools and Applications 78:11387–11425.

[43] Ani Brown Mary , A. Sherly, et al. “Classification of Membrane Protein Using Tetra Peptide Pattern.” Analytical Biochemistry, vol. 606, Oct. 2020, p. 113845. DOI.org (Crossref), https://doi.org/10.1016/j.ab.2020.113845.

[44] V.S, Abbiramy, and Ani Brown Mary. Comparison of Statistical Methods for Classification of Human Semen Data. preprint, In Review, 14 Mar. 2023. DOI.org (Crossref), https://doi.org/10.21203/rs.3.rs-2322083/v1.

[45] N Ani Brown Mary, Mrs N Adline Rajasenah Merryton, D Sheefa Ruby Grace (2021) “BANANA LEAVES DISEASES CLASSIFICATION using DPVP and PCCNN”, Cape Comorin Trust, India, Pages 62.

[46] N Ani Brown Mary, A Robert Singh, Suganya Athisayamani (2021) “Classification of banana leaf diseases using enhanced gabor feature descriptor”, Inventive Communication and Computational Technologies, Pages 229-242, Publisher Springer, Singapore.

[47] K Jayapriya, I Jeena Jacob, N Mary (2020) “Person re-identification using prioritized chromatic texture (PCT) with deep learning”, Multimedia Tools and Applications, Volume 79, Issue 39, Pages 29399-29410, Publisher Springer US.

[48] N Ani Brown Mary, Gomathi, S., et al. Employing a Novel Tri-Code Embedding Vector with LSTM and SoftMax Layer for Membrane Protein Classification. preprint, In Review, 14 Mar. 2023. DOI.org (Crossref), https://doi.org/10.21203/rs.3.rs-1873422/v1.

[49] N Ani Brown Mary et al., “Classification of Diseases in Banana Leaves using Diagonal Path Value Pattern” International Journal of Scientific Development and Research (IJSDR) , vol 7, issue 10, 2022.

[50] N Ani Brown Mary et al., “AN EXTENSIVE SURVEY ON SUBMARINE IMAGE ENHANCEMENT TECHNIQUES” Strad Research Journal,vol 8, issue 12, 2021.

[51] Flick, Carlos. "ROUGE: A Package for Automatic Evaluation of summaries." The Workshop on Text Summarization Branches Out2004:10. (2014).

[52] Vedantam, Ramakrishna, C. L. Zitnick, and D. Parikh. "CIDEr: Consensus-based Image Description Evaluation." Computer Science ,4566-4575. (2014) 20. Anderson, Peter, et al. "SPICE: Semantic Propositional Image Caption Evaluation." Adaptive Behavior 11.4 382-398. (2016).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Biography.docx