Vol.5 No.1 Jan. 2024

Multimedia Research

Received 20 August, Revised 27 October, Accepted 30 November

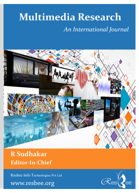

Image Caption Generation Framework using Hybrid Grey Wolf Optimization and Crow Search Algorithm

Saketh Kamatham
Masters in Computer Science,San Jose State University, CA, USA

Abstract: The image captioning process is used to generate an image textual description. To generate caption, image captioning uses both natural language processing as well as computer vision. Nevertheless, most image captioning systems present indistinct depictions about objects such as “Man”, “woman”, “group of people”, “building” and so on. Therefore, an intelligent based image captioning technique is developed in this work. The proposed technique consists of some steps such as the formation of the sentence, generation of words, and generation of the caption. At first, the input image is fed to a Deep learning classifier named Convolutional Neural Network (CNN). Here, the main aim of the classifier is used to train the appropriate words which are associated with the image, and can simply classify related words of a given image. Furthermore, the Long-Short Term Memory (LSTM) technique is exploited to form a set of sentences with generated words. Subsequently, the Maximum Like hood (ML) function is exploited to calculate formed sentences likelihood, as well as sentences with maximum probability is exploited that is furthermore exploited to generate a visual illustration of the scene regarding the image caption. The major objective of this work is to improve CNN performance by optimally tuning its activation function and weight. Therefore, for this optimal selection, this work adopts a novel improved optimization approach named Grey Wolf Optimization (GWO) algorithm and Crow Search Algorithm (CSA), which is termed as the Hybrid GWO-CSA algorithm. At last, the adopted captioning technique performance is evaluated with other existing techniques regarding the statistical analysis.

Keywords: Image Captioning, Deep Learning Classifier, Lstm, Optimization Algorithm, Sentence.

Nomenclature

Abbreviations	Descriptions
CNN	Convolutional Neural Network
GWO	Grey Wolf Optimization Algorithm
RNNs	Recurrent Neural Networks
LSTM	Long-Short Term Memory
DNN	Deep Neural Networks
CSA	Crow Search Algorithm
FCN	Fully Convolutional Network
avtmNet	adaptive visual-text merging network
CNN	Convolutional Neural Network

1. Introduction

The usage of intelligent mobile phone as well as photoshop tools have performed increasingly images appear on social network platforms with speedy advancement of Internet as well as information technology [1]. The generation of image caption, a crossing domain of natural language processing as well as computer vision attempts to produce the textual caption for given image. The image caption generation task turn out to be more valuable present with the remarkable raise of the multimedia social media. Additionally, the image resources efficient, the generation of image caption has extensive application forecast in associated fields namely content-based image retrieval, automatic labeling, human-computer interaction, as well as dialog system. In the meantime, the generation of image caption is facing numerous confronts [2].

Generation of caption is a motivating artificial intelligence issue wherein an eloquent sentence is produced for a given image [3]. It includes the double models from computer vision to recognize the image content as well as a language model from the natural language processing field to turn the considerate of the image into words in the right order. Image captioning has several applications namely recommendations in editing applications, tradition in virtual assistants, for visually impaired persons, for image indexing, for social media, and numerous other natural language processing applications. In recent times, deep learning techniques have attained conventional outcomes on instances of this issue. It has been shown that deep learning techniques are capable to attain optimum outcomes in the field of caption generation field issues. To predict a caption a single end-to-end technique is explained, describe a photo instead of necessitate complex preparation of data else a specifically designed pipeline models [4].

In computer vision and scene understanding models, the main issue is the automatically generating a portrayal for an image. This task is considered as a challenging task because training model is not only exploited to identify objects in an image, but also it should indicate object properties as well as their relationships in natural language [5]. In recent times, various studies have conducted regarding the visual recognition techniques on the basis of the CNNs as well as RNNs which considerably enhanced the generated captions quality. However, a great deal remains to be done in an intelligent system to match the accurateness of human image descriptions. The main significant element aspects of the visual recognition system are integrating the contextual information like object presence, scene context for instance a farm, and object co-occurrence to explain a scene. For instance, if we need to predict a caption for an image significant only which it comprises a horse, a farm, as well as a human, then almost certainly generate a sentence like a man is next to the horse else boy is riding the horse. The additional vital ability of the visual system is visual-spatial attention that directs consideration to a particular position in a scene or image. The visual-spatial attention is for the most part significant while the scene comprises background clutter; humans do not explain everything in a scene, other than as an alternative look at the significant regions and objects [2].

The main objective of this work is to propose an image captioning model which includes the steps such as the formation of the sentence, generation of words, and generation of the caption. Initially, the input image is fed to the CNN. As the appropriate words are formerly trained using the classifier, and it is able to classify the related words of the presented image simply. Additionally, using the LSTM technique, a set of sentences is produced with generated words. By exploiting, the ML function, the created sentences likelihood is calculated as well as the sentences with higher probability is used that are subsequently deployed to generate the visual illustration of the scene regarding the image caption. Furthermore, this work tries to improve CNN performance by tuning its activation function, as well as weight in an optimal way for this Hybrid GWO-CSA technique, is used. At last, the developed technique performance is evaluated with existing techniques to validate adopted model efficiency.

2. Literature Survey

In 2020, Xianrui Li et al [1], proposed joint attribute recognition as well as a visual attention model for clothes image captioning. Particularly, a pre-trained CNN was initially used to learn features that can characterize more information regarding the attribute of clothing. An encoder/decoder model was adopted for the learned feature, initially; the clothes feature was encoded, subsequently, it was input to a language LSTM technique to decode clothes descriptions. The technique highly improves clothes image captioning performance and minimizes the deceptive consideration.

In 2020, Maofu Liu et al [2], worked on visual attention to deepen indulgence of the image, integrating image labels produced by FCN into image caption generation. Moreover, the adopted technique uses textual attention to augment the reliability of the information. At last, label generation, attached to textual attention model, as well as generation image caption, were concatenated to create an end-to-end trainable modelIn 2020, Heng Song et al [3], introduced a visual attention approach was initially to produce a text attention technique as well as the visual information to create the text information correspondingly and an avtmNet . This concatenating network can competently concatenate the text information and visual information, and mechanically ascertain the proportion of both text information as well as visual information to produce subsequent caption words.In 2020, Huan Liu et al [4], developed an automated technique to manifest construction activity scenes using image captioning – a technique rooted in natural language as well as computer vision generation. At first, to manifest the scenes a linguistic description model was adopted and two different dedicated image captioning datasets were formed for technique examination. Subsequently, a fundamental technique framework of image captioning was modeled by integrating an encoder/decoder model with DNN, pursued by three investigational tests including model learning schemes chosen and performance estimate measures.

In 2021, Imad Afyouni et al [5], presented AraCap, a hybrid object-based, attention-augmented image captioning framework, with a focal point on the Arabic language. There were three techniques were exhibited all of them were validated and trained on Flickr30k as well as COCO datasets, and subsequently examined by constructing an Arabic version of a COCO dataset subset. An object-based captioner was an initial technique that can handle one or multiple detected objects. Next, an integrated pipeline was adopted which employs both attention-based captioning as well as to object detector.

3. System Model

In this paper, to generate the descriptions from images, a probabilistic as well as neural model is introduced. Present advancements in the transformation of arithmetic machines have shown that with a superior sequence technique it is possible to obtain the existing results by raising the probability of precise translation, while an input sentence is presented in an "end-to-end" way for both pieces of training as well as inference.

These techniques use an RNN model which encodes changing input lengths to a dimensional vector of predetermined as well as it furthermore uses this exhibition into a decode it to precise outcome sentence. Hence, it is adequate to use the same technique that is while an image is subjected, translating rule it into its description can be used. Hence, a technique is described to directly raise the exact description of image probability using eq. (1). Here, T indicates exact transcription, θ indicates constraint of the given model, Im indicates image (edge image, center image, and complete image). Since T representing any sentence, its length is unbounded.

$$\theta^* = \arg\max_{\theta} \sum_{(Im,T)} \log p(T|Im;\theta) \tag{1}$$

Hence, to model joint probability with $T_0,...T_N$ it is basic to use the chain rule, where N denotes the length of the specific exhibited as stated in Eq. (2), the dependency on θ is unused for expediency.

$$\log p(T|Im;\theta) = \sum_{s=0}^{N} \log p(T_s|Im,T_0,...T_{s-1}) \tag{2}$$

(T,Im) indicates a training instance pair, summation of log probabilities are optimized as stated in Eq. (2) over complete training set using the stochastic gradient descent during the training. It is general to model $p(T_s|Im,T_0,...T_{s-1})$ with an RNN, in that changing number of words up to $s-1$ is identified using a memory g_s or predetermined length hidden state. Furthermore, memory g_s is updated subsequent to observing a novel input y_s using a non-linear function h as stated in Eq. (3).

$$g_{s+1} = h(g_s, y_s) \tag{3}$$

To make the aforesaid RNN much sensible, 2 significant modeling decisions need to be devised: "what is the exact form of h and how are words" as well as images subjected as inputs y_s. An LSTM net is exploited for h, which has shown existing performance on series tasks namely translation. In this paper, an optimized CNN model is exploited to represent the image. Here, the adopted optimized CNN is represented as a novel technique that batches the normalization as well as obtains the present optimal performance based upon the ILSVRC 2014 classification competition [6]. Additionally, they also simplify the other tasks like scene classification by exploiting the "transfer learning" [7] and by exploiting the embedding technique the words are represented.

3.1 Sentence Generator using LSTM

In Eq. (3), h selection is controlled by its ability to handle exploding as well as eradicating gradients [8] that is a significant problem for RNNs designing as well as training. To deal with this problem, an exact kind of recurrent net, called LSTM, was produced [8] used for sequence translation and generation with great attainment [9] [10] [11]. LSTM model contains "memory cell" b that is regulated using 3 gates. The cell update description, gates, as well as output, are indicated in Eq. (4)-Eq. (9), in that Θ denotes the "product with a gate value and the diverse M matrices" [8]. Sigmoid $\sigma(.)$ represents nonlinearities and $g(.)$ represents hyperbolic tangent. m_s is exploited to feed the Softmax which generates a probability distribution p_s for the whole words in eq. (9).

$$i_s = \sigma(M_{iyu} yu_s + M_{im} m_{s-1}) \tag{4}$$

$$h_s = \sigma(M_{hyu} yu_s + M_{hm} m_{s-1}) \tag{5}$$

$$o_s = \sigma(M_{oyu} yu_s + M_{om} m_{s-1}) \tag{6}$$

$$b_s = h_s \Theta b_{s-1} + i_s \Theta g(M_{byu} yu_s + M_{bm} m_{s-1}) \tag{7}$$

$$m_s = o_s \Theta b_s \tag{8}$$

$$p_{s+1} = \mathrm{Soft\,max}(m_s) \tag{9}$$

Training: The most important element of the LSTM technique is a "memory cell" b which encodes knowledge at every time step concerning inputs that are seen in this step. The LSTM model is fed to the training which forecasts every word of sentence subsequent it having seen the image and whole previous words is exhibited as $p(T_s|Im,T_0,...T_{s-1})$. Therefore, it is improved to believe in LSTM in unrolled format – "a copy of the LSTM memory is formed for every sentence word and the image such that the entire LSTMs shares the similar constraints". Furthermore, the output of LSTM's m_s-1 at a time $s-1$ is subjected to LSTM at time s. The whole recurrent links are altered to feed-forward connects in the unrolled edition. In brief, if the input image is stated as Im as well as $T=(T_0,...T_N)$ indicated as a true sentence of the image, the unrolling process is read as exhibited as in Eq. (10) to (12).

$$yu_{-1}=CNN(Im) \tag{10}$$

$$yu_s=M_eT_s,\quad s\in\{0.....N-1\} \tag{11}$$

$$p_{s+1}=lstm(y_s),\quad s\in\{0.....N-1\} \tag{12}$$

Therefore, T_0 signifies a particular begin word and T_N signifies the particular final word which denotes begin and final of the sentence. Particularly, the LSTM signals would be generated which represents a complete sentence by neglecting the stop word. Both the image as well as the words are plotted to the same space. Moreover, by exploiting the CNN image are plotted and by exploiting the word embedder M_e the words are plotted. The Im is subjected as input only once, at $s=-1$, to inform LSTM concerning image contents. Additionally, it is tested and verified that feeding image at each time step as a supplementary input obtains low-grade results, as the network can employ image noise overtly additionally it ensembles more purely. Eq. (13) states loss is a summary of the "negative log-likelihood" of the precise word at every step. The loss gets minimized regarding all the LSTM parameters, M_e as well as the CNN top layer.

$$L(Im,T)=-\sum_{s=1}^{N}\log p_s(T_s) \tag{13}$$

3.2 Classification using CNN Model

The spatial information is used by the CNN between image pixels, thus they are based upon discrete convolution [14].

The CNN primary elements are exhibited in [12]. In eq. (14), a grayscale image is supposed to be represented by a function is stated thus the image Im is recognized by an array size of $a_1\times a_2$. Let filter as $D\in\Re^{2ge_1+1\times 2ge_2+1}$ for image and it is stated in eq. (15), the discrete convolution with filter D in that filter D is stated using eq. (16). Eq. (17) indicates a basically exploited for the smoothing purpose which is the discrete Gaussian filter $D_{H(\sigma)}$ [27] σ states the standard deviation of Gaussian distribution. Let layer S as a convolution layer, the input comprises $n_1^{(S-1)}$ feature maps from the previous layer. If s is considered as "1", input is represented as a single image Im that includes one or more channels. Thus, a CNN takes into consideration of "unprocessed images as input directly".

The layer s output consists of $n_1^{(S)}$ feature maps with size $n_2^{(S)}\times n_3^{(S)}$. Eq. (18) indicates i^{th} feature map in the layer S, stated as $Xe_i^{(S)}$ is stated as, here, $C_i^{(S)}$ represents bias matrix and $D_{i,j}^{(S)}$ represents filter size of $2ge_1^{(S)}+1\times 2ge_2^{(S)}+1$ which connects the i^{th} feature map in layer S with j^{th} feature map in the layer $(S-1)$.

$$Im:\{1,...a_1\}\times\{1,...a_2\}\to P\subseteq\Re,(i,j)\mapsto Im_{i,j} \tag{14}$$

$$(Im*D)_{l,r}:=\sum_{v=-ge_1}^{ge_1}\sum_{u=-ge_2}^{ge_2}D_{v,u}\,Im_{l+v,r+u} \tag{15}$$

$$D=\begin{pmatrix}D_{-ge_1,-ge_2} & ... & D_{-ge_1,-ge_2}\\ \vdots & D_{0,0} & \vdots\\ D_{ge_1,-ge_2} & .... & D_{ge_1,ge_2}\end{pmatrix} \tag{16}$$

$$\left(D_{H(\sigma)}\right)_{l,r} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{l^2 + r^2}{2\sigma^2}\right) \tag{17}$$

$$Xe_i^{(S)} = C_i^{(S)} + \sum_{j=1}^{n_1^{s-1}} D_{i,j}^{(S)} * Xe_j^{(S-1)} \tag{18}$$

As aforementioned, border effects influences $n_2^{(S)}$ as well as $n_3^{(S)}$ by exploiting discrete convolution in the applicable segment of input feature maps which is for pixels in that sum of Eq. (15) is shown exactly, output feature maps comprise a size as stated in Eq. (19). Eq. (20) is used to connect the convolutional layer and its function to multilayer perceptron, this formulation is reformulated. Moreover, Eq. (21) exhibits the output, attained accordingly by the evaluation using the unit at location (l,r).

$$n_2^{(S)} = n_2^{(S-1)} - 2ge_1^{(S)} \text{ and } n_3^{(S)} = n_3^{(S-1)} - 2ge_2^{(S)} \tag{19}$$

$$\left(Xe_i^{(S)}\right)_{l,r} = \left(C_i^{(S)}\right)_{l,r} + \sum_{j=1}^{n_1^{(S-1)}} \left(D_{i,j}^{(S)} * Xe_j^{(S-1)}\right)_{l,r} \tag{20}$$

$$\left(Xe_i^{(S)}\right)_{l,r} = \left(C_i^{(S)}\right)_{l,r} + \sum_{j=1}^{n_1^{(S-1)}} \sum_{v=-ge_1^S}^{ge_1^S} \sum_{u=-ge_2^S}^{ge_2^S} \left(L_{i,j}^{(S)}\right)_{v,u} \left(D_{i,j}^{(S)}\right)_{l,r} \left(Xe_j^{(S-1)}\right)_{l+v,r+u} \tag{21}$$

In the network, the trainable weights can be ascertained in the bias matrices $C_i^{(S)}$ and filters $D_{i,j}^{(S)}$. The input of S is indicated as $n_1^{(S)}$ feature maps its output consists of $n_1^{(S)} = n_1^{(S-1)}$ feature maps if S is a non-linearity layer as represented as Eq. (22), where f indicates activation function.

If S and $S-1$ is a fully connected layer, Eq. (23) is exploited, where $y^{(S-1)}$, $we^{(S)}$ and $z^{(S)}$ denotes own vectors as well as matrix indications of outputs f weights, $we_{i,k}^{(S)}$ and actual inputs $z_i^{(S)}$ respectively. Otherwise, the layer S expects $n_1^{(S-1)}$ size feature maps $n_2^{(S-1)} \times n_3^{(S-1)}$ as input as well as i^{th} input in layer S estimates Eq. (24). Here, $we_{i,j,l,r}^{(S)}$ indicates weights connecting unit at (l,r) $_a$ location in the layer $(S-1)$ as well as i^{th} unit in the layer S.

$$Xe_i^{(S)} = f\left(Xe_i^{(S-1)}\right) \tag{22}$$

$$z_i^{(S)} = \sum_{k=0}^{n^{(S-1)}} we_{i,k}^{(S)} y_k^{S-1} \quad \text{or} \quad z^{(S)} = we^{(S)} y^{(S-1)} \tag{23}$$

$$y_i^{(S)} = f\left(z_i^{(S)}\right) \text{ with } z_i^{(S)} = \sum_{j=1}^{n_1^{(S-1)}} \sum_{l=1}^{n_2^{(S-1)}} \sum_{r=1}^{n_3^{(S-1)}} we_{i,j,l,r}^{(S)} \left(X_k^{S-1}\right)_{l,r} \tag{24}$$

In this paper, to construct the captioning process more precisely, it is intended to tune a few of parameters optimally. Significantly, activation function (f) and weight (we) of CNN are optimally tuned. Nevertheless, to solve these optimization problems, tuning with optimal values is not at all a simple task; this work aspires to develop a novel approach named Hybrid CSA-GWO.

4. Proposed Hybrid GWO-CSA model

The GWO possesses better exploitation capability however poor exploration ability hence in the adopted optimization approach [13], a superior value of $f1$ is used to exploit the CSA superior exploration quality as represented in eq. (25).

In the adopted technique, rather than updating from α, β and δ, as stated in Eq. (25), a search agent is allowed for updating its position.

$$Y(t+1) = Y + f1 \times \text{rand} \times \left(\left(Y_1 - Y\right) + \left(Y_2 - Y\right)\right) / 2 \tag{25}$$

To maintain population diversity, not all individuals are updated by α, as well as β updating direction in the population, but by α merely in the adopted model. This performs as a shrinking scheme that set up adopted mode to evade from local optimum.

$Y(t+1) = Y + f1 \times \text{rand} \times (Y_1 - Y)$ (26)

A fixed balance probability amid Eq. (25) and Eq. (26) is not good to attain the required ratio of exploitation/exploration. Therefore, a probability of adaptive balance is adopted that permits the adopted model to attain acceleration all through previous steps of the optimization process whereas afterward steps of optimization capable solutions will have a maximum probability to be used in this paper. The probability of adaptive balance p is calculated as follows:

$$p = 1 - \left(1.01 \times t^3 / Max\_iter^3\right) \tag{27}$$

Wherein Max_iter signifies a maximum number of iterations and t denotes the current iteration. Eq. (28) is used to generate the control parameter values a at the time of the optimization process. This model allows the adopted model to effectually look at search space compared to existing GWO.

$$a = 2 - \left(\cos\left(ran = (\ )\right) \times 1 / Max\_iter\right) \tag{28}$$

5. Result and Discussion

In this section, the adopted Hybrid GWO-CSA technique experimentation analysis regarding the image caption generation was demonstrated. Here, the proposed Hybrid GWO-CSA model was compared with the CNN, PSO-CNN [17], GWO-CNN [15], and GA-CNN [16].

The adopted dataset was "**Flickr8k_Dataset** that comprises a total of 8092 images in JPEG format with different shapes and sizes, of that 6000 images are exploited for training, 1000 images for testing, and 1000 images for validation". Here, cosine similarity, as well as Jaccard similarity, was used as standard metrics to compute generated image caption quality. In the proposed model, statistical evaluation was performed by deviating the weights, as well as outcomes, was achieved.

Fig 1 demonstrates the analysis of the adopted technique with existing models regarding Jaccard Similarity. Here, the proposed model is 16% better than the CNN, 18% better than the PSO-CNN, 14% better than the GWO-CNN and 19% better than the GA-CNN models for best. Fig 2 demonstrates the analysis of the adopted technique with existing models regarding cosine Similarity. Here, the proposed model is 18% better than the CNN, 15% better than the PSO-CNN, 12% better than the GWO-CNN and 16% better than the GA-CNN models for best.

Fig 3 reveals the analysis of the adopted technique with existing models regarding precision. Here, the proposed model is 28% better than the CNN, 26% better than the PSO-CNN, 23% better than the GWO-CNN and 11% better than the GA-CNN models for best. Fig 4 reveals the analysis of the adopted technique with existing techniques regarding the recall. In this fig it is clearly shown that the proposed model is 26% better than the CNN, 29% better than the PSO-CNN, 21% better than the GWO-CNN and 29% better than the GA-CNN models for best. Moreover, the adopted technique is executed 5 times, results are taken under certain metrics such as best, worst, median, mean, as well as standard deviation, correspondingly. From evaluation, superior results are analyzed to be obtained using the proposed model than the conventional models.

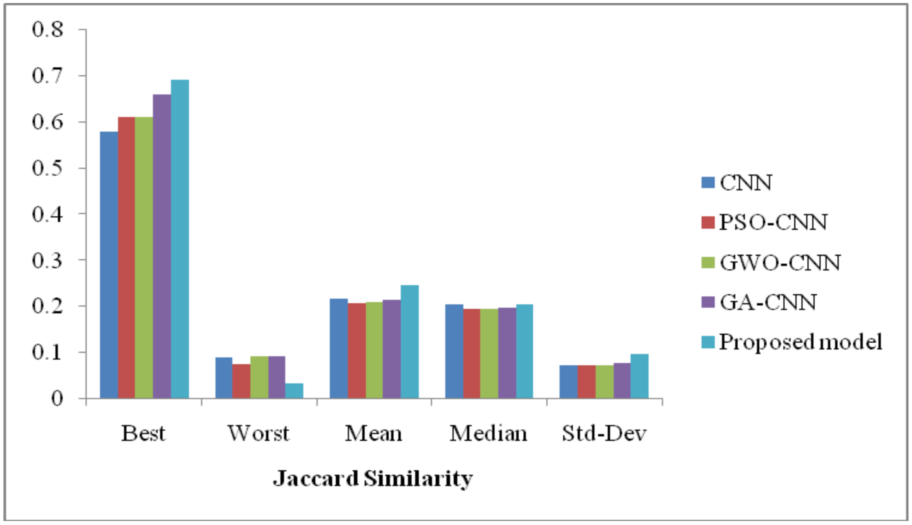

Fig.1. *Analysis of the adopted technique with existing techniques concerning Jaccard Similarity*

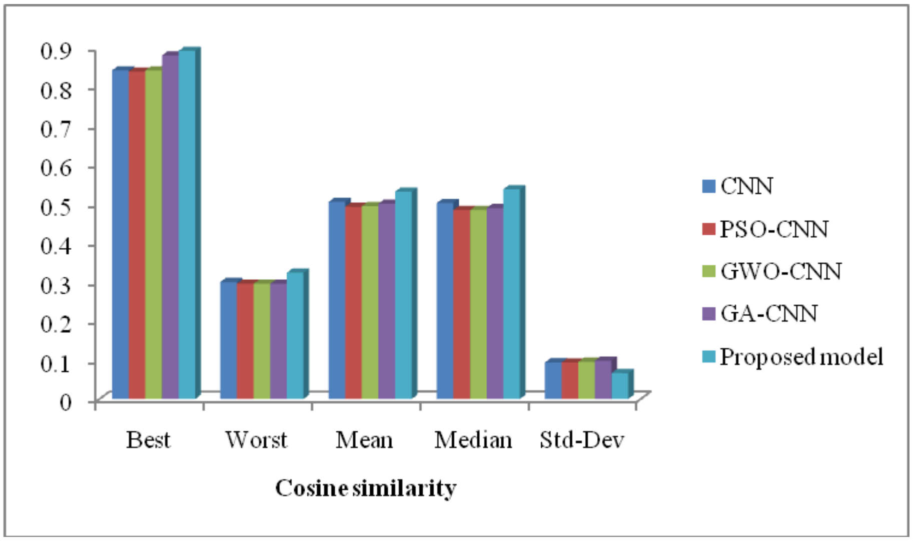

Fig.2. *Analysis of the adopted technique with existing techniques concerning Cosine Similarity*

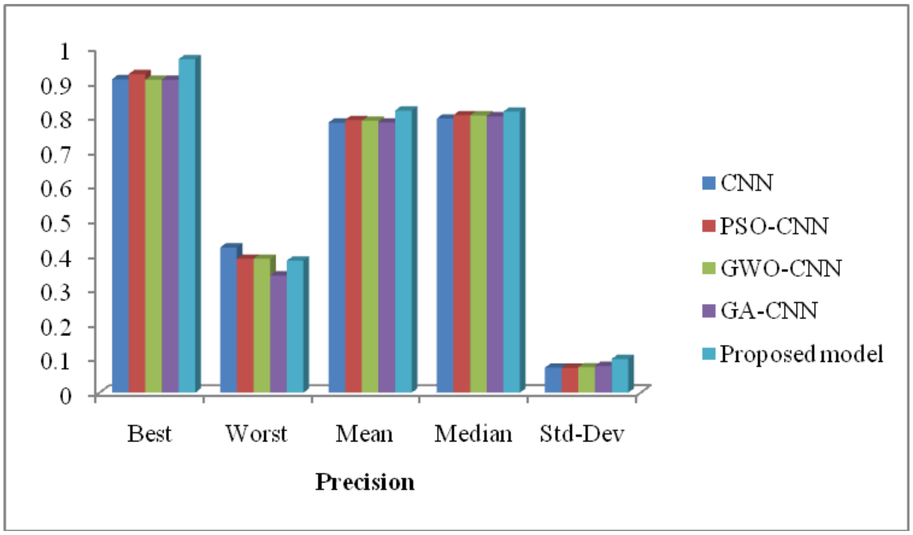

Fig.3. *Analysis of the adopted technique with existing techniques concerning Precision*

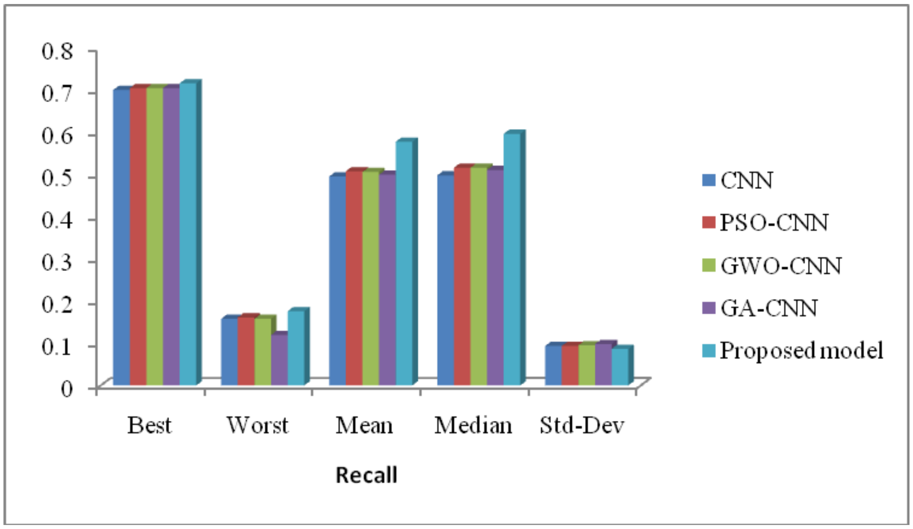

Fig.4. *Analysis of the adopted technique with existing techniques concerning Recall*

6. Conclusion

The major contribution of this research was to present an image captioning approach, which consists of steps such as the formation of the sentence, generation of words, and generation of the caption. At first, the input image was fed to CNN. Here, the classifier was exploited to train the appropriate words that can able to easily classify related words of the presented image. Moreover, a set of sentences was formed with generated words exploiting the LSTM technique. By exploiting the ML function, the likelihood of the created sentences was calculated. To generate a visual indication of the scene regarding the image caption, the sentences with higher probability were exploited. Furthermore, this work tries to enhance CNN performance by tuning its activation function, as well as its weight in an optimal way for this RR-BOU approach, was used. At last, the proposed model performance was evaluated with the conventional models to verify the efficiency of the adopted technique. The overall analysis states that the superiority of the adopted technique with existing techniques.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

Reference

[1] Xianrui LiZhiling YeMingbo Zhao,"Clothes image caption generation with attribute detection and visual attention model", Pattern Recognition Letters, 10 December 2020.

[2] Maofu LiuLingjun LiJing Tian,"Image caption generation with dual attention mechanism", Information Processing & Management, 12 December 2019.

[3] Heng SongJunwu ZhuYi Jiang,"avtmNet:Adaptive Visual-Text Merging Network for Image Captioning", Computers & Electrical Engineering, 15 April 2020.

[4] Huan LiuGuangbin WangXiaochun Luo," Manifesting construction activity scenes via image captioning", Automation in Construction, 6 July 2020.

[5] Imad AfyouniImtinan AzharAshraf Elnagar,"AraCap: A hybrid deep learning architecture for Arabic Image Captioning", Procedia Computer Science, 14 July 2021.

[6] Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, In arXiv:1502.03167, 2015.

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang,E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition", In ICML, 2014.

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8), 1997.

[9] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation, In EMNLP, 2014.

[10] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequencelearning with neural networks, In NIPS, 2014.

[11] A. Graves. Generating sequences with recurrent neural networks. arXiv:1308.0850, 2013.

[12] Y. LeCun, K. Kavukvuoglu, and C. Farabet, “Convolutional networks and applications in vision”, In Circuits and Systems, International Symposium on, pp. 253–256, 2010.

[13] S. Arora, H. Singh, M. Sharma, S. Sharma and P. Anand, "A New Hybrid Algorithm Based on Grey Wolf Optimization and Crow Search Algorithm for Unconstrained Function Optimization and Feature Selection," IEEE Access, vol. 7, pp. 26343-26361, 2019.

[14] Chaitrali Prasanna Chaudhari, Satish Devane, "Improved Framework using Rider Optimization Algorithm for Precise Image Caption Generation", International Journal of Image and Graphics, 2021.

[15] Amolkumar Narayan Jadhav,Gomathi N,"DIGWO: Hybridization of Dragonfly Algorithm with Improved Grey Wolf Optimization Algorithm for Data Clustering", Multimedia Research, vol. 2, no. 3, July 2019.

[16] Raviraj Vishwambhar Darekar,Ashwinikumar Panjabrao Dhande,"Emotion Recognition from Speech Signals Using DCNN with Hybrid GA-GWO Algorithm",Multimedia Research, vol. 2, no. 4, October 2019.

[17] Yogesh R kulkarni,Senthil Murugan T,"Hybrid Weed-Particle Swarm Optimization Algorithm and C-Mixture for Data Publishing",Multimedia Research, vol. 2, no. 3, July 2019.