





Efficient Evolving Deep Ensemble Medical Image Captioning Network

Dilbag Singh , Senior Member, IEEE, Manjit Kaur , Senior Member, IEEE, Jazem Mutared Alanazi, Ahmad Ali AlZubi , and Heung-No Lee , Senior Member, IEEE

Abstract—With the advancement in artificial intelligence (AI) based E-healthcare applications, the role of automated diagnosis of various diseases has increased at a rapid rate. However, most of the existing diagnosis models provide results in a binary fashion such as whether the patient is infected with a specific disease or not. But there are many cases where it is required to provide suitable explanatory information such as the patient being infected from a particular disease along with the infection rate. Therefore, in this paper, to provide explanatory information to the doctors and patients, an efficient deep ensemble medical image captioning network (DCNet) is proposed. DCNet ensembles three well-known pre-trained models such as VGG16, ResNet152V2, and DenseNet201. Ensembling of these models achieves better results by preventing an over-fitting problem. However, DCNet is sensitive to its control parameters. Thus, to tune the control parameters, an evolving DCNet (EDC-Net) was proposed. Evolution process is achieved using the self-adaptive parameter control-based differential evolution (SAPCDE). Experimental results show that EDC-Net can efficiently extract the potential features of biomedical images. Comparative analysis shows that on the Open-i dataset, EDC-Net outperforms the existing models in terms of BLUE-1, BLUE-2, BLUE-3, BLUE-4, and kappa statistics (KS) by 1.258%, 1.185%, 1.289%, 1.098%, and 1.548%, respectively.

Index Terms—Medical, diagnosis, pre-trained models, explanatory information.

I. INTRODUCTION

WITH the advancement in artificial intelligence (AI) based E-healthcare applications, the role of automated

Manuscript received 4 February 2022; revised 22 August 2022; accepted 30 September 2022. Date of publication 18 November 2022; date of current version 6 February 2023. This work was supported in part by the Ministry of Science and ICT, Korea, through the Information Technology Research Center Support Program under Grant IITP-2021-0-01835 supervised by the Institute for Information and Communications Technology Planning and Evaluation, in part by the National Research Foundation of Korea funded by the Korean Government, MSIP, under Grant NRF-2021R1A2B5B03002118, and in part by the Researchers Supporting Project, King Saud University, Riyadh, Saudi Arabia, under Grant RSP-2021/395. (Corresponding author: Heung-No Lee.)

Dilbag Singh, Manjit Kaur, and Heung-No Lee are with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea (e-mail: dg-gill2@gmail.com; manjitkaur@gist.ac.kr; heungno@gist.ac.kr).

Jazem Mutared Alanazi and Ahmad Ali AlZubi are with the Computer Science Department, Community College, King Saud University, Riyadh 11451, Saudi Arabia (e-mail: ajazem@ksu.edu.sa; aalzubi@ksu.edu.sa).

Digital Object Identifier 10.1109/JBHI.2022.3223181

diagnosis of various diseases has increased at a rapid rate. Deep learning models have recently been used by many researchers to classify patients suffering from a particular disease. But these models generally provide results in a binary fashion such as whether the patient is infected from a specific disease or not. However, there are many cases where it is required to provide suitable information to the patients such as the patient being infected with some disease with this much infection rate. Therefore, these days many researchers have started utilizing image captioning techniques to provide explanatory information to doctors and patients.

Recently, semantic concepts were used to detect the captions from the image. Image-caption pairs were used to train the concept detector. However, it suffers from vocabulary discrepancy and deficiency of required information [1]. To increase the accuracy of pathological information in diagnostic reports, Semantic fusion networks (SFNet) were utilized [2]. Attention-based models were also used in image captioning. The attention masks were queried using hidden states of LSTM from image features. It provided better image information for training deep sequential networks. However, these models did not ensure that layers significantly focused on regions of interest due to indirectly supervised learning [3]. These problems were overcome by utilizing differentiable neural networks to obtain the captions from the images [4]. To describe the disease information in ultrasound images, LSTM was used to decode the encoding vectors [5].

Most image captioning techniques are based on visual information and rarely rely on semantic content. Therefore, it is also needed to express the emotions of text descriptions for better captioning. Yang et al. [6] made sentences using both emotional and visual information by combining latent codes. The quality of AI-generated diagnostic reports is equally important as the development of models. However, [6] performs poorly for images with poor visibility.

Babar et al. [7] proposed a new measure to evaluate the quality of generated diagnostic reports. Convolutional neural network (CNN)-based image captioning models can retain only a few features of the original image. The image captioning DCNet based on a recurrent neural network (RNN) suffers from a gradient vanishing problem. The multimodal fusion method solved these issues by using CNN and RNN into the same DCNet for image captioning [8]. A semi-supervised deep generative DCNet generated the description more naturally from images [9]. An adaptive multimodal attention network was also designed to produce better quality captioned images [10].

Kavur et al. [11] utilized various ensemble models such as majority voting, average combiner, product combiner, and min/max combiner. To prevent overfitting problems with deep learning models, U-Net, deepmedic, V-Net, and dense V-networks were ensemble to enhance the liver segmentation from CT images.

However, it is found that the hyper-parameters selection of the existing models was achieved using the trial-and-error basis. Additionally, the designed model suffers from the over-fitting and gradient vanishing problem. To overcome these problems, an efficient EDC-Net model is designed.

The main contributions of this paper are as follows:

- 1) To provide explanatory information to the doctors and patients, an efficient deep ensemble medical image captioning network (DCNet) is proposed.
- 2) The proposed DCNet ensembles three well-known pre-trained models such as VGG16, ResNet152V2, and DenseNet201.
- 3) To tune the control parameters, evolving DCNet (EDC-Net) was proposed. The evolution process is achieved using the self-adaptive parameter control-based differential evolution (SAPCDE)
- 4) By considering benchmark datasets, extensive experiments are conducted.

The remainder of the paper is organized as follows: Section II describes related work. Section III provides a mathematical formulation of EDC-Net. Section IV presents the comparative analysis. Section V concludes the paper.

II. RELATED WORK

Yu et al. [12] used order embedding to caption the topic-oriented image. A convolutional neural network (CNN)-based classifier is used to select the topics for images from candidates. Wang et al. [13] captioned the images using a recurrent memory network (RMN). In training, topic words were recorded from a topic repository. After that, the image was tested using the retrieval method for the generation of a topic word. Finally, the sentence was generated through a recurrent memory network by incorporating the retrieved topic words. Zhang et al. [1] proposed an image captioning DCNet using missing concepts mining and online positive recall. The suitable captions were automatically generated using the element-wise selection method. Zhang et al. [3] captioned the images using the visual aligning attention DCNet (VAA). The visual aligning loss was designed to optimize the attention layer in the training stage. The non-visual words were filtered out using visual vocab to minimize their effect on the attention layer. Zhou et al. [14] enhanced single-phase image captioning using a saliency-enhanced re-captioning model. In this, two-phase learning was applied to get better results. In the first phase, semantic and visual saliency cues were extracted from the model. In the second phase, cues were fused for the self-boosting of the model. Zhao et al. [15] captioned the images using DCNet adaption and cross-modal retrieval in cross-domain. Firstly, the source domain was used to pre-train the cross-modal and then utilized in the target domain to extract

the initial image-sentence pairs. These pairs were further reefered using a retrieval model. Yang et al. [16] proposed cross-domain image captioning using Multitask learning (ML). Textual explanations of the images were generated using CNN-LSTM. Images were synthesized using a Conditional generative adversarial network (C-GAN) based on generated text descriptions. Hoxha et al. [17] captioned the images using CNN and Recurrent neural network (RNN). Firstly, visual features were extracted and translated into a textual explanation. Secondly, embedding techniques were used to convert the textual explanation into feature vectors. Finally, similar images were retrieved by calculating the similarity between vectors of textual explanation and archive images. Huang et al. [18] utilized a Multimodal attribute detector (MAD) and Subsequent attribute predictor (SAP) to improve the performance of image captioning. MAD used image features as well as word embedding to improve attribute detection accuracy. A concise attribute was predicted with SAP every time to reduce the diversity of image attributes. Yang et al. [19] proposed a Dual generator generative adversarial network (DGGAN) based image captioning model. This technique ensembled the generation-based and retrieval-based image captioning models. Yuan et al. [20] captioned the images using multi-label attribute graph convolution and multi-level attention. The attention module focused on both specific spatial and scale features. For image captioning, attribute features were learned using the attribute graph convolution module. Huang et al. [21] utilized the Denoising-based multi-scale feature fusion (DMSFF) technique to caption the images. Monay et al. [22] captioned the images using the probabilistic latent semantic analysis (PLSA) model. The textual and visual modalities are assumed equally by an expectation-maximization algorithm. Yu et al. [23] proposed an image captioning DCNet using Multimodal Transformer (MT) model. Multi-view visual features were also introduced to improve the performance. Xian et al. [24] used multimodal LSTM to caption the images. Amirian et al. [25] proposed a deep learning-based image captioning model. Li et al. [26] designed an efficient image captioning model that can be used to extract potential information from digital images. It has shown effective performance over competitive models. But this model suffers from the over-fitting issue.

Park et al. [27] designed a multi-difference average pooling LSTM (mDiAP-LSTM)-based medical image captioning model. The most appropriate method for capturing the differences was determined through a study of feature representation methods. Hou et al. proposed an efficient Full-adversarial reinforcement learning (Full-ARL) model for medical image captioning. Li et al. [28] designed an efficient knowledge-driven encode, retrieve, paraphrase (KERP) model to obtain better medical image captions. It decomposes captions into explicit disease-related abnormality graphs, which were then analyzed using natural language modeling. Hou et al. [29]. An overall score based on accuracy and fluency was provided by two additional discriminators as the evaluator. A report generator was implemented, which produces discrete sequences of words based on decision probabilities, as opposed to generative adversarial networks (GANs) used in image generation. Furthermore, it prevented

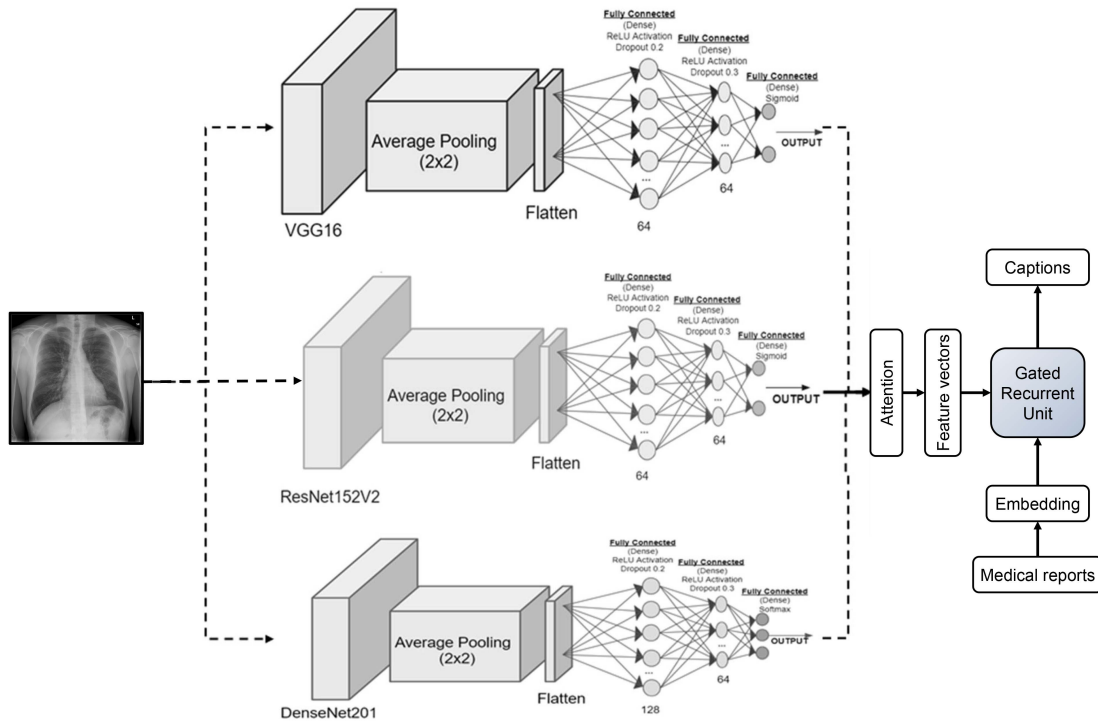


Fig. 1. Proposed ensemble deep transfer network with gated recurrent unit.

the gradients from being transmitted between discriminators and generators.

Wang et al. [30] designed an efficient relational-topic retrieval and generation framework (R-paraNet). Reports incorporated semantically consistent medical terms and encourage the generation of sentences for rare abnormal descriptions. Li et al. [31] designed a hybrid retrieval-generation reinforced agent (HRGR-Agent) that integrated retrieval-based strategies with sophisticated learning techniques to provide robust image captions. Hierarchical decision-making was employed by HRGR-Agent. To obtain a caption, a low-level generation module invoked either a high-level retrieval policy or a high-level retrieval policy module. Reinforcement learning guided the HRGR-Agent's updates via rewards at the word and sentence levels. Jing et al. [32] designed a multitask learning model that predicted tags and paragraphs simultaneously. It utilized co-attention to locate abnormal regions and generate narrations for them. The hierarchical LSTM (HLSTM) was proposed for generating long paragraphs.

It is observed that the existing models suffer from hyper-parameters tuning problems. Moreover, the designed model exhibits over-fitting and gradient vanishing problems.

III. PROPOSED MODEL

This section discusses the proposed medical image captioning model.

A. Ensemble Deep Transfer Model

An ensemble deep transfer network (DCNet) is proposed. VGG16 [33], ResNet152V2 [34], and DenseNet201 [35] models

have been used to develop the proposed model. Fig. 1 shows the EDC-Net with a gated recurrent unit. Ensembling of models provides better results by preventing over-fitting [36], [37]. It also enhances the extraction of features and performance of supervised models [38], [39]. Fig. 1 demonstrates the proposed ensemble deep transfer model. 128 neurons have been utilized for input dense layer [40]. VGG16 [33], ResNet152V2 [34], and DenseNet201 [35] models are used to obtain potential features. These models have been trained using 20 epochs with a batch size of 8. Fully connected layers [41] having size 128 neurons along with dropouts of 0.2 and 0.25, respectively have been utilized to prevent memorization problems with the competitive models. $l_r = 0.001$ has been utilized as learning rate.

B. Gated Recurrent Unit

Cho et al. [42] designed gated recurrent unit (GRU). GRU allows every recurrent unit to dynamically obtain dependencies of different time scales. Similar to LSTM, GRU has gating units that modulate the knowledge flow within the unit, but, without utilizing any additional memory cells.

Activation (α_n^k) of GRU at time t is a linear interpolation among candidate activation ($\tilde{\alpha}_n^k$) and succeeding activation (α_{t-1}^k). α_n^k can be computed as:

$$\alpha_n^k = (1 - \beta_n^k)\alpha_{t-1}^k + \beta_n^k\tilde{\alpha}_n^k, \quad (1)$$

Here, update gate (β_n^k) monitors and controls the activation. An update gate can be evaluated as:

$$\beta_n^k = \sigma(W_z\gamma_n + U_z\alpha_{t-1}^k). \quad (2)$$

Here, $W_r \gamma_n$ represents the weighted matrix. GRU is unable to limit the degree to which its state is exposed. However, it can expose the whole state during every iteration.

A candidate activation ($\tilde{\alpha}_n^k$) can be computed [43] as:

$$\tilde{\alpha}_n^k = \tanh(W_r \gamma_n + U(\mathbf{r}_n \odot \alpha_{t-1}))^k \quad (3)$$

Here, \odot shows an element-wise multiplication. \mathbf{r}_n contains a group of reset gates.

When r_n^k approaches 0, the reset gate can efficiently develop a unit act as if it is utilizing the first symbol of the input sequence by forgetting the earlier evaluated state.

Reset gate (r_n^k) can be evaluated according to the update gate as:

$$r_n^k = \sigma(W_r \gamma_n + U_r \alpha_{t-1})^k. \quad (4)$$

C. Adaptive Differential Evolution

The proposed DCNet suffers from the hyper-parameters tuning issue. Therefore, in this paper, a differential evolution variant is used to evolve the proposed model. Bilingual Evaluation Understudy (BLUE) [44] is used as an objective function. Differential evolution (DE) is used in various fields due to its advantages such as strong robustness, good performance, and simple structure to solve optimization problems. The performance of DE mainly depends on the selection of control parameters (crossover rate CR , scale factor F , and population size NZ) and trial vector generation strategy (crossover and mutation). According to the problem nature, these parameters should be selected for better optimization results. The setting of these parameters is a challenging task for any problem. The solution to this problem is given by Cui et al. [45] by proposing self-adaptive parameter control-based DE (SAPCDE). It is based on the idea that good parameters should be propagated from generation to generation and the bad parameters should learn from the good parameters. In SAPCDE, two populations are utilized i.e., the solution population and the parameter population. Every solution has its control parameters. Parameter population is also evolved with each generation. SAPCDE is a combination of basic DE and parameter self-adaptation control methods.

Lets suppose that initial parameter population for DCNet is represented as $C^0 = \{C_1^0, C_2^0, \dots, C_{NZ}^0\}$ with $C_i^0 = \{F_{i,1}^0, CR_{i,2}^0\}$, where NZ denotes the population size. The initial solution population is represented as $S^0 = \{S_1^0, S_2^0, \dots, S_{NZ}^0\}$ with $S_i^0 = \{s_{i,1}^0, s_{i,2}^0, \dots, s_{i,D}^0\}$, where D represents the number of variables. G_n represents the number of generations. The parameter population evolved in the same way as the solution population in DE. Initially, parameter population is generated uniformly and randomly between $[0, 0]$ and $[1, 1]$. Thereafter, for each individual $C_i^{G_n}$, mutation parameter ($VC_i^{G_n}$) is generated using mutation operator such as

$$VC_i^{G_n} = C_{r1}^{G_n} + CF \cdot (C_{r2}^{G_n} - C_{r3}^{G_n}) \quad (5)$$

$$VC_i^{G_n} = C_{r1}^{G_n} + CF \cdot (SC_k^{G_n} - C_{r2}^{G_n}) \quad (6)$$

Here, C_{r1} , C_{r2} , and C_{r3} are selected from parameter population randomly. $SC_k^{G_n}$ denotes a good parameter that is selected

randomly. Next, the trailing parameter ($TC_i^{G_n}$) is generated using crossover operator such as

$$TC_i^{G_n} = \begin{cases} VC_{i,j}^{G_n}, & \text{if } (rand_{i,j} \leq CCR \text{ or } j == j_{rand}) \\ C_{i,j}^{G_n}, & \text{Otherwise} \end{cases} \quad (7)$$

where $j = \{1, 2\}$ and $i = \{1, 2, \dots, NZ\}$. $rand_{i,j} \in [0, 1]$ represents the uniform random number. $CCR \in [0, 1]$ denotes the new CR. Lastly, the selection operator is applied to select the good parameter for the next generation. In SAPCDE, good parameter individual $C_i^{G_n}$ is one which helps the $S_i^{G_n}$ to produce better offspring $TS_i^{G_n}$. Otherwise, $C_i^{G_n}$ is considered as a bad control parameter. The selection operator for a good parameter is defined as

if $C_i^{G_n}$ is a good parameter

$$C_i^{G_{n+1}} = \begin{cases} C_i^{G_n}, & \text{if } rand(0, 1) < \lambda_1 \\ TC_i^{G_{n+1}}, & \text{Otherwise} \end{cases} \quad (8)$$

else

$$C_i^{G_{n+1}} = \begin{cases} TC_i^{G_{n+1}}, & \text{if } rand(0, 1) < \lambda_2 \\ C_i^{G_n}, & \text{Otherwise} \end{cases} \quad (9)$$

where λ_1 and λ_2 control the values of parameters to explore the new values and keep the previous values. The basic working of SAPCDE is illustrated in Algorithm 1. Initially, solution and parameter populations are generated in line 1 and line 2, respectively. In line 3, number of generation (G_n) is initialized to 1. F_{eval} represents the number of function evaluations. Line 5 represents the termination condition of the algorithm when F_{eval} reaches F_{max} . GP stores the good parameter individuals (line 6). It is initialized with 0. The solution population is evolved through lines 7 to 16. In line 7, the mutation operator is applied to generate a mutant vector ($SV_i^{G_n}$) using parameter individual $C_i^{G_n}$ such as

$$SV_i^{G_n} = S_{r1}^{G_n} + F_{i,1}^{G_n} \cdot (S_{r2}^{G_n} - S_{r3}^{G_n}) \quad (10)$$

where S_{r1} , S_{r2} , and S_{r3} are randomly selected from solution population. $F_{i,1}^{G_n} \in C_i^{G_n}$ represents the scale factor. In line 8, crossover operator is applied to obtain a trial vector ($TS_i^{G_n}$) using $C_i^{G_n}$ such as

$$TS_i^{G_n} = \begin{cases} SV_{i,j}^{G_n}, & \text{if } (rand_{i,j} \leq CR_{i,2} \text{ or } j == j_{rand}) \\ S_{i,j}^{G_n}, & \text{Otherwise} \end{cases} \quad (11)$$

where $CR_{i,2} \in C_i^{G_n}$ and $i = \{1, 2, \dots, NZ\}$. A selection operator is applied to select the best solution (lines 10–15). If fitness of $TS_i^{G_n}$ is better than $S_i^{G_n}$, the corresponding $C_i^{G_n}$ is considered as a good parameter individual and flagged as 1 (i.e., $val = 1$) (line 11). It is also added in the GP in line 12. If $C_i^{G_n}$ is a bad parameter, then it is flagged as 0 (line 14). Next, the parameter population is evolved using mutation, crossover, and selection operators (lines 17–37). If $GP = 0$ at any G_n , then bad parameters are initialized randomly (line 31) to explore the new values.

Algorithm 1: Self-Adaptive Parameter Control-Based Differential Evolution for Evolving Ensemble Model.

```

1 Generate initial solution ( $S^0$ ) and parameter population ( $C^0$ )
2 Set  $G_n = 1$  and  $F_{eval} = 0$ 
3 while  $F_{eval} < F_{max}$  do
4   Set  $GP = 0$ 
5   for  $i = 0$  to  $NZ$  do
6     Mutant vector  $SV_i^{G_n}$  is obtained using Eq. 10
       and  $C_i^{G_n}$ 
7     Trial vector  $TS_i^{G_n}$  is obtained using Eq. 11 and
        $C_i^{G_n}$ 
8     if  $f(TS_i^{G_n}) \geq f(S_i^{G_n})$  then
9        $S_i^{G_{n+1}} = TS_i^{G_n}$ ,  $val(i) = 1$ 
10      Put  $C_i^{G_n}$  into  $GP$ 
11     else
12        $S_i^{G_{n+1}} = S_i^{G_n}$ ,  $val(i) = 0$ 
13     end
14   end
15   for  $i = 1$  to  $NZ$  do
16     if  $val(i) == 1$  then
17       if  $rand(0, 1) < \lambda_1$  then
18          $C_i^{G_{n+1}} = C_i^{G_n}$ 
19       else
20         Generate  $TC_i^{G_n}$  using Eq. (5) and Eq.
           (7)
21          $C_i^{G_{n+1}} = TC_i^{G_n}$ 
22       end
23     else
24       if  $rand(0, 1) < \lambda_2$  then
25         if  $GP \neq 0$  then
26           Generate  $TC_i^{G_n}$  using Eqs. (6) and
             (7)
27            $C_i^{G_{n+1}} = TC_i^{G_n}$ 
28         else
29           initialize  $C_i^{G_{n+1}}$  randomly
30         end
31       else
32          $C_i^{G_{n+1}} = C_i^{G_n}$ 
33       end
34     end
35   end
36    $G_n = G_{n+1}$ 
37 end

```

IV. PERFORMANCE ANALYSIS

In this paper, a dataset is obtained from Open-i [46]. It contains chest X-ray images and radiology text reports. Each image is paired with respective captions. 7,471 chest X-ray images are available with frontal and lateral views of given patients. Additionally, VGG16 [33], ResNet152V2 [34], and DenseNet201 [35] models are also ensembled using majority voting (MV), min, and max combiner. These models are

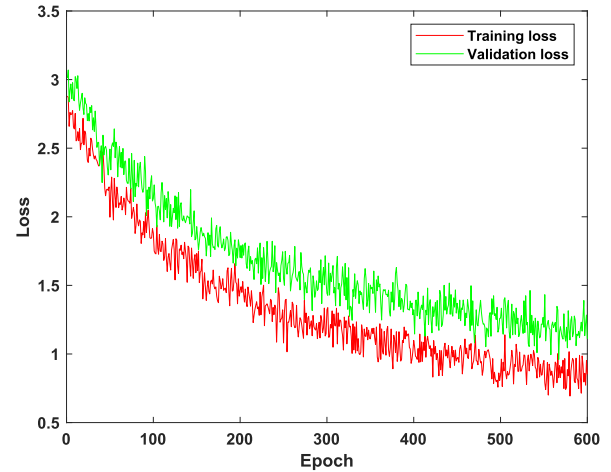


Fig. 2. Training and validation loss analysis of VGG16.

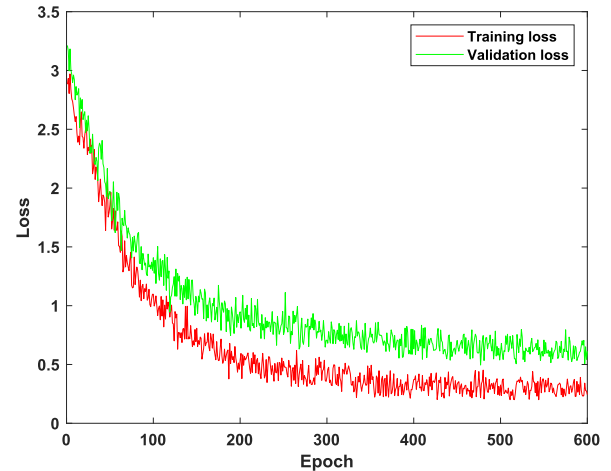


Fig. 3. Training and validation loss analysis of ResNet152V2.

renamed as MVE_1, Min_4, and Max_5, respectively. Also, inspired from [47], during the evolution phase, validation data is used to evaluate the performance of EDC-Net. The final trained EDC-Net is then tested on the training dataset.

A. Training and Validation Loss Analysis

This section discusses the training and validation analysis using the loss curves. Fig. 2 shows the training and validation loss analysis of VGG16 with respect to a number of epochs. It is found that the VGG16 achieves the best training and validation loss values as 0.7387 and 1.3481, respectively. It shows that there is an impact of over-fitting too.

Fig. 3 demonstrates the training and validation loss analysis of ResNet152V2. It is found that RESNET152 V achieves the best training and validation loss values of 0.6297 and 1.1726, respectively. It shows that there is an impact of over-fitting. But it shows better performance than VGG-16. Fig. 4 shows the training and validation loss analysis of DenseNet201. It is found that the best training and validation loss values are 0.3207 and 0.5218, respectively), thus DenseNet201 is least affected by the

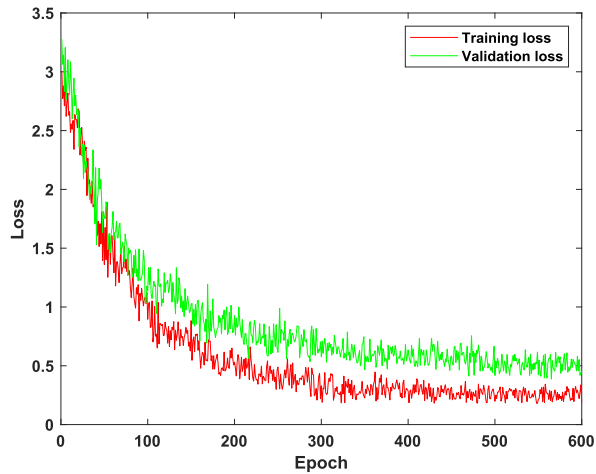


Fig. 4. Training and validation loss analysis of DenseNet201.

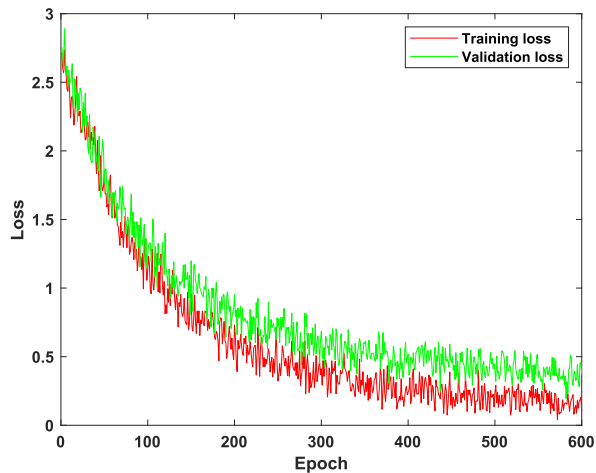


Fig. 5. Training and validation loss analysis of the proposed EDC-Net.

over-fitting problem. But still, there is room for improvement in the convergence curve. Fig. 5 demonstrates the training and validation loss analysis of EDC-Net. It is found that EDC-Net approaches toward minimum possible training loss. Also, there is a lesser difference between best training and validation loss values, (i.e., 0.2189 and 0.4368, respectively), thus EDC-Net is least affected by the over-fitting problem.

B. Visual Analysis

Fig. 6 shows the correctly obtained captions. It is found that the predicted captions are identical to the actual captions. So, for such captions, we received a maximum BLUE score. Since we have considered it as a classification problem, therefore, it represents the truly predicted class too. It clearly shows that EDC-Net can effectively provide captions for medical images.

Fig. 7 demonstrates the incorrectly obtained captions. It is found that the predicted captions are far away from the actual captions. So, for such captions, we received a minimum BLUE score. Since we have considered it as a classification problem, therefore, it represents the falsely predicted class

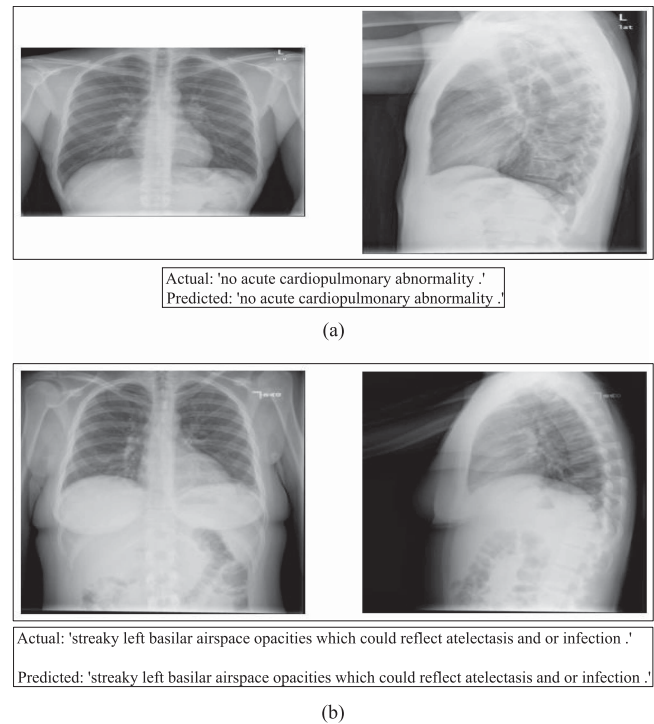


Fig. 6. Correctly classified captions.

TABLE I
BLUE SCORE AND KAPPA STATISTICS (KS) ANALYSIS OF THE EDC-NET

Model	BLUE-1	BLUE-2	BLUE-3	BLUE-4	KS
RMN	0.563	0.342	0.229	0.125	0.919
VGG16	0.568	0.364	0.249	0.126	0.923
CNN-RNN	0.511	0.351	0.253	0.123	0.912
DMSFF	0.539	0.353	0.244	0.126	0.924
DGGAN	0.563	0.365	0.238	0.127	0.932
ResNet152V2	0.571	0.309	0.256	0.119	0.896
DenseNet201	0.574	0.327	0.227	0.128	0.938
MVE_1	0.573	0.363	0.252	0.127	0.934
Min_1	0.571	0.361	0.250	0.124	0.915
Max_1	0.569	0.359	0.249	0.121	0.908
EDC-Net	0.577	0.367	0.258	0.129	0.952

too. It clearly shows that in certain cases when the visibility of the images is poor then EDC-Net fails to provide correct captions.

C. Quantitative Analysis

This section discusses the performance analyses using the confusion matrix. Fig. 8 shows the confusion matrix obtained from the VGG16 on the medical image captioning dataset. It is found that the VGG16 achieves 93.5% accuracy. Fig. 9 shows the confusion matrix obtained from the ResNet152V2 on the medical image captioning dataset. It is found that the ResNet152V2 achieves 94.8% accuracy. Fig. 10 shows the confusion matrix obtained from the DenseNet201 on the medical image captioning dataset. It is found that the DenseNet201 achieves 95.4% accuracy. Fig. 11 shows the confusion matrix obtained from EDC-Net on the medical image captioning dataset. It is found

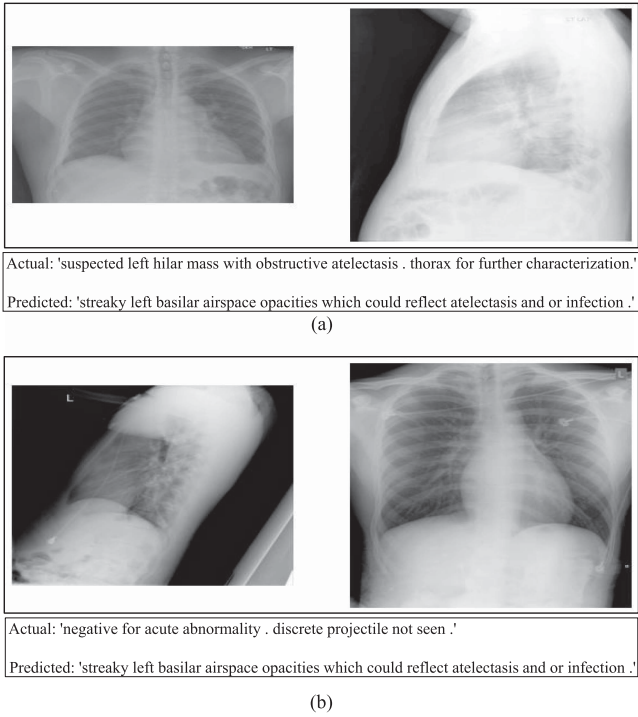


Fig. 7. Incorrectly classified captions.

Output Class	Class 1	Class 2	Class 3	Class 4	
Class 1	22 14.4%	4 2.6%	0 0.0%	1 0.7%	81.5% 18.5%
Class 2	5 3.3%	31 20.3%	0 0.0%	0 0.0%	86.1% 13.9%
Class 3	0 0.0%	0 0.0%	41 26.8%	0 0.0%	100% 0.0%
Class 4	0 0.0%	0 0.0%	0 0.0%	49 32.0%	100% 0.0%
	81.5% 18.5%	88.6% 11.4%	100% 0.0%	98.0% 2.0%	93.5% 6.5%
	Class 1	Class 2	Class 3	Class 4	

Fig. 8. Confusion matrix analysis of VGG16 on medical image captioning dataset.

that EDC-Net achieves 97.4% accuracy. Thus, EDC-Net outperforms the competitive models with an average improvement of 2.0%.

D. Discussion

Table I shows the BLUE score [44] analysis of EDC-Net. Comparisons are drawn between the proposed and the

Output Class	Class 1	Class 2	Class 3	Class 4	
Class 1	26 17.0%	1 0.7%	0 0.0%	1 0.7%	92.9% 7.1%
Class 2	1 0.7%	34 22.2%	0 0.0%	3 2.0%	89.5% 10.5%
Class 3	0 0.0%	0 0.0%	41 26.8%	2 1.3%	95.3% 4.7%
Class 4	0 0.0%	0 0.0%	0 0.0%	44 28.8%	100% 0.0%
	96.3% 3.7%	97.1% 2.9%	100% 0.0%	88.0% 12.0%	94.8% 5.2%
	Class 1	Class 2	Class 3	Class 4	

Fig. 9. Confusion matrix analysis of ResNet152V2 on medical image captioning dataset.

Output Class	Class 1	Class 2	Class 3	Class 4	
Class 1	27 17.6%	4 2.6%	0 0.0%	0 0.0%	87.1% 12.9%
Class 2	0 0.0%	28 18.3%	0 0.0%	0 0.0%	100% 0.0%
Class 3	0 0.0%	0 0.0%	41 26.8%	0 0.0%	100% 0.0%
Class 4	0 0.0%	3 2.0%	0 0.0%	50 32.7%	94.3% 5.7%
	100% 0.0%	80.0% 20.0%	100% 0.0%	100% 0.0%	95.4% 4.6%
	Class 1	Class 2	Class 3	Class 4	

Fig. 10. Confusion matrix analysis of DenseNet201 on medical image captioning dataset.

competitive models such as RMN [13], VGG16 [33], CNN-RNN [7], DMSFF 9057472, DGGAN [19], ResNet152V2 [34], and DenseNet201 [35] using BLUE scores [44] and kappa statistics (KS) [48], [49]. It is found that DenseNet201 [35] achieved the highest BLUE-1 as 0.574 as compared to the other competitive models. DGGAN [19] and VGG16 [33] models have achieved highest BLUE-2 values as 0.365 and 0.364, respectively. ResNet152V2 [34] achieved highest BLUE-3 value as 0.56 than the existing models. DenseNet201 [35]. Ensemble-based models, i.e., MVE_1, Min_4, and Max_5 achieved better results than most of the existing models. Out of these models,

Output Class	Class 1	Class 2	Class 3	Class 4	
Class 1	24 15.7%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Class 2	3 2.0%	35 22.9%	1 0.7%	0 0.0%	89.7% 10.3%
Class 3	0 0.0%	0 0.0%	40 26.1%	0 0.0%	100% 0.0%
Class 4	0 0.0%	0 0.0%	0 0.0%	50 32.7%	100% 0.0%
	88.9% 11.1%	100% 0.0%	97.6% 2.4%	100% 0.0%	97.4% 2.6%
	Class 1	Class 2	Class 3	Class 4	

Fig. 11. Confusion matrix analysis of EDC-Net on medical image captioning dataset.

Source	SS	df	MS	F	Prob>F
Columns	2.267	10	0.2267	37.7833	2.016e-04
Error	0.294	49	0.0060		
Total	2.561	59			

Fig. 12. ANOVA table of BLUE-1.

Source	SS	df	MS	F	Prob>F
Columns	2.836	10	0.2836	32.9767	0.0156
Error	0.423	49	0.0086		
Total	3.259	59			

Fig. 13. ANOVA table of BLUE-2.

Source	SS	df	MS	F	Prob>F
Columns	0.53724	10	0.3301	47.8405	0.0023
Error	0.44408	20	0.0069		
Total	0.98132	24			

Fig. 14. ANOVA table of BLUE-3.

MVE_1 outperformed the Min_4 and Max_5. In terms of KS, DGGAN achieved remarkable performance over the existing models. From Table I, it is found that EDC-Net outperforms the existing models by achieving higher values of BLUE scores. Bold values indicate higher performance. It is found that the proposed EDC-Net outperforms the existing models in terms

Source	SS	df	MS	F	Prob>F
Columns	3.198	10	0.3198	53.3000	0.0003
Error	0.297	49	0.0060		
Total	3.495	59			

Fig. 15. ANOVA table of BLUE-4.

Source	SS	df	MS	F	Prob>F
Columns	1.748	10	0.1748	22.4102	1.435e-03
Error	0.385	49	0.0078		
Total	2.133	59			

Fig. 16. ANOVA table of kappa statistics (KS).

TABLE II
COMPARATIVE ANALYSIS OF EDC-NET WITH THE STATE-OF-THE-ART MODELS

Model	BLUE-1	BLUE-2	BLUE-3	BLUE-4
mDiAP-LSTM [27]	0.373	0.226	0.147	0.101
Full-ARL [29]	-	-	-	0.125
R-paraNet (VGG-19)[30]	0.505	0.329	0.230	0.168
R-paraNet (DneseNet-19)[30]	0.503	0.333	0.236	0.175
HRGR-Agent [31]	0.438	0.298	0.208	0.151
HLSTM [32]	0.455	0.288	0.205	0.154
KERP [28]	0.216	0.214	0.087	0.066
EDC-Net	0.516	0.348	0.238	0.178

of BLUE-1, BLUE-2, BLUE-3, BLUE-4, and KS by 1.258%, 1.185%, 1.289%, 1.098%, and 1.548%, respectively. Also, additional ANOVA analyses are performed on each performance metric which has shown that the EDC-Net significantly outperforms the competitive models.

For statistical analysis, ANOVA is used. Each performance metric has the following hypotheses:

$$\begin{cases} H_0 & \mu M_1 = \mu M_2 = \dots = \mu M_{11}, \\ H_A & \text{Means are not equal.} \end{cases} \quad (12)$$

Here, Null and alternate hypotheses are defined by H_0 and H_A , respectively. μM_i represents EDC-Net and the existing image captioning models. M_{11} depicts EDC-Net. The ANOVA table contains degrees of freedom, a sum of squares, the mean sum of squares, F-statistics (F_V), and P-value (p). If p value of F_V falls below the significance level, we will reject H_0 and state that there is a significant difference between the models. For all the metrics analysis presented in Table I, H_A is accepted (see Figs. 12–16). It shows that there is a significant difference in performance between different models.

Table II shows the comparative analysis between the EDC-Net and state-of-the-art models on IU X-RAY [50]. It is found that EDC-Net outperforms the state-of-the-art models in terms of BLUE-1, BLUE-2, BLUE-3, and BLUE-4 by 1.357%, 1.249%, 1.115%, and 1.031%, respectively.

V. CONCLUSION

To provide explanatory information to the doctors and patients, an efficient deep ensemble medical image captioning network (DCNet) was proposed. DCNet has ensembled three well-known pre-trained models such as VGG16, ResNet152V2, and DenseNet201. Ensembling of these models has shown better results by preventing over-fitting. DCNet can efficiently extract the potential features of biomedical images, but it is sensitive to its control parameters. Therefore, to tune the control parameters, evolving DCNet (EDC-Net) was proposed. Evolution was achieved using the self-adaptive parameter control-based differential evolution. Comparative analysis has shown that EDC-Net achieves higher performance than the existing models to provide explanatory information for biomedical images. Comparative analysis has shown that EDC-Net achieves 97.4% accuracy. According to the results obtained using Open-i dataset, the proposed EDC-Net outperformed the existing models in terms of BLUE-1, BLUE-2, BLUE-3, BLUE-4, and KS by 1.258%, 1.185%, 1.289%, 1.098%, and 1.548%, respectively. Additionally, the EDC-Net model outperformed the state-of-the-art models on IU X-RAY dataset in terms of BLUE-1, BLUE-2, BLUE-3, and BLUE-4 by 1.357%, 1.249%, 1.115%, and 1.031%, respectively.

In the near future, to handle the poor visibility, noise, poor registration, etc. kind of problems with medical images, we will integrate various preprocessing techniques such as image filtering, image registration, visibility restoration, etc. with the proposed EDC-Net. Since preprocessing techniques may reduce the performance of EDC-Net, therefore, a novel selective preprocessing model will be designed which will decide whether preprocessing is required or not. If preprocessing is required then which operation(s) should be applied on the given image(s).

Data Availability Statement: The data collected during the data collection phase are available from the corresponding authors upon request.

Conflicts of Interest: The authors would like to confirm there are no conflicts of interest regarding the study.

REFERENCES

- [1] M. Zhang, Y. Yang, H. Zhang, Y. Ji, H. T. Shen, and T.-S. Chua, "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 32–44, Jan. 2019.
- [2] X. Zeng, L. Wen, Y. Xu, and C. Ji, "Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models," *Comput. Methods Programs Biomed.*, vol. 197, 2020, Art. no. 105700.
- [3] Z. Zhang, W. Zhang, W. Diao, M. Yan, X. Gao, and X. Sun, "VAA: Visual aligning attention model for remote sensing image captioning," *IEEE Access*, vol. 7, pp. 137355–137364, 2019.
- [4] R. Sharma, A. Kumar, D. Meena, and S. Pushp, "Employing differentiable neural computers for image captioning and neural machine translation," in *Proc. Int. Conf. Smart Sustain. Intell. Comput. Appl. Procedia Comput. Sci.*, 2020, vol. 173, pp. 234–244.
- [5] X. Zeng, L. Wen, B. Liu, and X. Qi, "Deep learning for ultrasound image caption generation based on object detection," *Neurocomputing*, vol. 392, pp. 132–141, 2020.
- [6] J. Yang, Y. Sun, J. Liang, B. Ren, and S.-H. Lai, "Image captioning by incorporating affective concepts learned from both visual and textual components," *Neurocomputing*, vol. 328, pp. 56–68, 2019.
- [7] Z. Babar, T. van Laarhoven, F. M. Zanzotto, and E. Marchiori, "Evaluating diagnostic content of AI-generated radiology reports of chest x-rays," *Artif. Intell. Med.*, vol. 116, 2021, Art. no. 102075.
- [8] D. Zhao, Z. Chang, and S. Guo, "A multimodal fusion approach for image captioning," *Neurocomputing*, vol. 329, pp. 476–485, 2019.
- [9] N. Zakharov, H. Su, J. Zhu, and J. Gläscher, "Towards controllable image descriptions with semi-supervised vae," *J. Vis. Commun. Image Representation*, vol. 63, 2019, Art. no. 102574.
- [10] S. Yang, J. Niu, J. Wu, Y. Wang, X. Liu, and Q. Li, "Automatic ultrasound image report generation with adaptive multimodal attention mechanism," *Neurocomputing*, vol. 427, pp. 40–49, 2021.
- [11] A. E. Kavur, L. I. Kuncheva, and M. A. Selver, "Basic ensembles of vanilla-style deep learning models improve liver segmentation from ct images," 2020, *arXiv:2001.09647*.
- [12] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic-oriented image captioning based on order-embedding," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2743–2754, Jun. 2019.
- [13] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 256–270, 2020.
- [14] L. Zhou, Y. Zhang, Y.-G. Jiang, T. Zhang, and W. Fan, "Re-caption: Saliency-enhanced image captioning through two-phase learning," *IEEE Trans. Image Process.*, vol. 29, pp. 694–709, 2020.
- [15] W. Zhao, X. Wu, and J. Luo, "Cross-domain image captioning via cross-modal retrieval and model adaptation," *IEEE Trans. Image Process.*, vol. 30, pp. 1180–1192, 2021.
- [16] M. Yang et al., "Multitask learning for cross-domain image captioning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1047–1061, Apr. 2019.
- [17] G. Hoxha, F. Melgani, and B. Demir, "Toward remote sensing image retrieval under a deep image captioning perspective," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4462–4475, 2020.
- [18] Y. Huang, J. Chen, W. Ouyang, W. Wan, and Y. Xue, "Image captioning with end-to-end attribute detection and subsequent attributes prediction," *IEEE Trans. Image Process.*, vol. 29, pp. 4013–4026, 2020.
- [19] M. Yang et al., "An ensemble of generation- and retrieval-based image captioning with dual generator adversarial network," *IEEE Trans. Image Process.*, vol. 29, pp. 9627–9640, 2020.
- [20] Z. Yuan, X. Li, and Q. Wang, "Exploring multi-level attention and semantic relationship for remote sensing image captioning," *IEEE Access*, vol. 8, pp. 2608–2620, 2020.
- [21] W. Huang, Q. Wang, and X. Li, "Denoising-based multiscale feature fusion for remote sensing image captioning," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 436–440, Mar. 2021.
- [22] F. Monay and D. Gatica-Perez, "Modeling semantic aspects for cross-media image indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1802–1817, Oct. 2007.
- [23] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4467–4480, Dec. 2020.
- [24] Y. Xian and Y. Tian, "Self-guiding multimodal LSTM—when we do not have a perfect training dataset for image captioning," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5241–5252, Nov. 2019.
- [25] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, "Automatic image and video caption generation with deep learning: A concise review and algorithmic overlap," *IEEE Access*, vol. 8, pp. 218386–218400, 2020.
- [26] X. Li et al., "COCO-CN for cross-lingual image tagging, captioning, and retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2347–2360, Sep. 2019.
- [27] H. Park, K. Kim, S. Park, and J.-W. Choi, "Medical image captioning model to convey more details: Methodological comparison of feature difference generation," *IEEE Access*, vol. 9, pp. 150560–150568, 2021.
- [28] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 6666–6673.
- [29] D. Hou, Z. Zhao, Y. Liu, F. Chang, and S. Hu, "Automatic report generation for chest x-ray images via adversarial reinforcement learning," *IEEE Access*, vol. 9, pp. 21236–21250, 2021.
- [30] F. Wang, X. Liang, L. Xu, and L. Lin, "Unifying relational sentence generation and retrieval for medical image report composition," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5015–5025, Jun. 2022.
- [31] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

- [32] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2577–2586.
- [33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
- [36] K. Siwek, S. Osowski, and R. Szupiluk, "Ensemble neural network approach for accurate load forecasting in a power system," *Int. J. Appl. Math. Comput. Sci.*, vol. 19, no. 2, pp. 303–315, 2009.
- [37] J. Islam and Y. Zhang, "Brain MRI analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain Informat.*, vol. 5, no. 2, 2018, Art. no. 2.
- [38] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Ensembles of deep learning models and transfer learning for ear recognition," *Sensors*, vol. 19, no. 19, 2019, Art. no. 4139.
- [39] D. Singh, V. Kumar, M. Kaur, M. Y. Jabarulla, and H.-N. Lee, "Screening of covid-19 suspected subjects using multi-crossover genetic algorithm based dense convolutional neural network," *IEEE Access*, vol. 9, pp. 142566–142580, 2021.
- [40] D. Singh, V. Kumar, V. Yadav, and M. Kaur, "Deep neural network-based screening model for covid-19-infected patients using chest x-ray images," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 35, no. 03, 2021, Art. no. 2151004.
- [41] D. Singh, V. Kumar, and M. Kaur, "Densely connected convolutional networks-based COVID-19 screening model," *Appl. Intell.*, vol. 51, no. 5, pp. 3044–3051, 2021.
- [42] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*.
- [43] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
- [44] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
- [45] L. Cui, G. Li, Z. Zhu, Z. Wen, N. Lu, and J. Lu, "A novel differential evolution algorithm with a self-adaptation parameter control method by differential evolution," *Soft Comput.*, vol. 22, no. 18, pp. 6171–6190, 2018.
- [46] U. L. of Medicine, "Open-i service of the national library of medicine," 2022. [Online]. Available: <https://openi.nlm.nih.gov>
- [47] A. E. Kavur et al., "Chaos challenge - combined (CT-MR) healthy abdominal organ segmentation," *Med. Image Anal.*, vol. 69, 2021, Art. no. 101950.
- [48] T. Toprak, B. Belenlioglu, B. Aydın, C. Guzelis, and M. A. Selver, "Conditional weighted ensemble of transferred models for camera based onboard pedestrian detection in railway driver support systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5041–5054, May 2020.
- [49] M. A. Selver, "A robotic system for warped stitching based compressive strength prediction of marbles," *IEEE Trans. Ind. Informat.*, vol. 16, no. 11, pp. 6796–6805, Nov. 2020.
- [50] H. Park, K. Kim, J. Yoon, S. Park, and J. Choi, "Feature difference makes sense: A medical image captioning model exploiting feature difference and tag information," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics: Student Res. Workshop*, 2020, pp. 95–102.