Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2022, Article ID 2756396, 13 pages
https://doi.org/10.1155/2022/2756396

Research Article

Caption Generation Based on Emotions Using CSPDenseNet and BiLSTM with Self-Attention

Kavi Priya S,[1] **Pon Karthika K**,[1] **Jayakumar Kaliappan**,[2] **Senthil Kumaran Selvaraj**,[3] **Nagalakshmi R**,[4] **and Baye Molla**[5]

[1]*Department of Computer Science and Engineering, Mepco Schlenk Engineering College (Autonomous), Sivakasi 626005, Tamil Nadu, India*
[2]*Department of Analytics, School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Vellore 632014, Tamil Nadu, India*
[3]*Department of Manufacturing Engineering, School of Mechanical Engineering (SMEC), Vellore Institute of Technology (VIT), Vellore 632014, Tamil Nadu, India*
[4]*Department of Computer Science and Engineering, Faculty of Engineering and Technology, Kalinga University, Raipur, Chhattisgarh, India*
[5]*School of Mechanical Engineering, Engineering and Technology College, Dilla University, P.O.Box. 419, Dilla, Ethiopia*

Correspondence should be addressed to Senthil Kumaran Selvaraj; sskumaranvit@gmail.com and Baye Molla; bayem@du.edu.et

Received 2 February 2022; Accepted 22 August 2022; Published 17 September 2022

Academic Editor: Dimitrios A. Karras

Automatic image caption generation is an intricate task of describing an image in natural language by gaining insights present in an image. Featuring facial expressions in the conventional image captioning system brings out new prospects to generate pertinent descriptions, revealing the emotional aspects of the image. The proposed work encapsulates the facial emotional features to produce more expressive captions similar to human-annotated ones with the help of Cross Stage Partial Dense Network (CSPDenseNet) and Self-attentive Bidirectional Long Short-Term Memory (BiLSTM) network. The encoding unit captures the facial expressions and dense image features using a Facial Expression Recognition (FER) model and CSPDense neural network, respectively. Further, the word embedding vectors of the ground truth image captions are created and learned using the Word2Vec embedding technique. Then, the extracted image feature vectors and word vectors are fused to form an encoding vector representing the rich image content. The decoding unit employs a self-attention mechanism encompassed with BiLSTM to create more descriptive and relevant captions in natural language. The Flickr11k dataset, a subset of the Flickr30k dataset is used to train, test, and evaluate the present model based on five benchmark image captioning metrics. They are BiLingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit Ordering (METEOR), Recall-Oriented Understudy for Gisting Evaluation (ROGUE), Consensus-based Image Description Evaluation (CIDEr), and Semantic Propositional Image Caption Evaluation (SPICE). The experimental analysis indicates that the proposed model enhances the quality of captions with 0.6012(BLEU-1), 0.3992(BLEU-2), 0.2703(BLEU-3), 0.1921(BLEU-4), 0.1932(METEOR), 0.2617(CIDEr), 0.4793(ROUGE-L), and 0.1260(SPICE) scores, respectively, using additive emotional characteristics and behavioral components of the objects present in the image.

1. Introduction

Artificial intelligence (AI) becomes the driving force behind evolving technologies such as IoT (Internet of Things) [1], sensor networks [2], robotics, and big data. Automatic image captioning is also an emerging interdisciplinary field of research in AI, related to NLP (natural language processing) and computer vision. The idea of an automatic caption generation system is to create a relevant and meaningful description for an image in an appropriate natural language format. The automatic caption generation system has an extensive array of applications such as intelligent human-

machine interactions, supporting visually impaired people, and image indexing. Image caption generation is a tedious process of identifying various objects present in the image and the association between those objects, attributes, and behavior. The advancement of deep learning techniques in various applications[3–6] has a wider scope in caption generation models. The typical architecture of an image captioning system using deep learning techniques follows the encoder-decoder framework. In recent works, convolutional neural network (CNN) is applied to encode the image by creating the semantic feature vector representation, and recurrent neural network (RNN) is applied to decode the feature vectors for generating the captions in natural language. Although these deep learning neural network models provided promising results, still it is an open research problem to identify the suitable and efficient CNN and RNN models for image captioning applications.

Recently developed deep learning-based caption generation techniques [7–9] considers only the factual and semantic content of the image. But, the emotional perspective of the image plays a vital role in generating high-quality descriptions more intelligently. To capture the emotional aspects of the image, the Face-Cap model is introduced by Nezami et al. [10]. The system can detect and derive the facial expression features and apply these features to generate captions for the image. The system extracted the features of facial expression from the image using a FER technique. Additionally, the system considers the features from the ImageNet dataset trained using the Oxford Visual Geometry Group network (VGGnet) [11] with a weighted attention mechanism. Finally, to generate the caption, the system utilized a long short-term memory (LSTM) network. Inspired by this work, the proposed system embeds the emotional analysis in the image caption generation model to automatically generate descriptions based on both the emotional features and salient image features. The proposed caption generation model employs CSPDenseNet [12] and the FER model [10] to extract image features and emotion features and introduced a self-attentive BiLSTM (bidirectional long short-term memory) network to describe the image more effectively. The contributions of the proposed work are as follows:

(1) Extracts more extensive and salient image features using CSPDenseNet instead of the standard CNN model and generates an image representation vector.

(2) Self-attentive BiLSTM model is introduced to generate captions. The BiLSTM model process the textual knowledge from both the forward direction and the backward direction. The self-attention mechanism is implemented for improving the quality of caption generation by focusing on the important text features as well as the contextual features.

(3) The experimental outcomes exhibit that the newly designed caption generation system is capable of describing an image with better quality in comparison to the existing image captioning models.

The remainder of the article is organized in the following sections. Section 2 examines the related works. The detailed description of the proposed model is explored in section 3. The experimental settings, results, and discussion are illustrated in section 4. Finally, section 5 concludes the present article with a summary.

2. Literature Survey

The existing works about image caption generation and emotion identification methods are reviewed in the following subsections.

2.1. Image Captioning Methodologies. Based on the recently proposed image caption generation models, the methodologies are categorized into three distinct groupings: template methods, retrieval methods, and novel image captioning methods. The template method generates captions based on fixed or specified templates containing numerous empty spaces. The empty slots in the templates are filled by identifying the different objects, object features, and behaviors present in the input image. A triplet of object, action, and scene spaces is proposed in [13]. This method fills the blank slots by predicting the triplet from the image based on the multilabel Markov random field. On the contrary, [14] detected multiple objects, classifiers, and their prepositional relationships to fill the empty template spaces using conditional random fields (CRF). Though the templates are predefined and the number of slots is fixed, they can generate grammatically correct captions.

The retrieval method aims to generate captions by retrieving semantically similar captions from the existing set of captions or the captions are simply retrieved from visually similar images. The model presented in [15] generated a description for a query image using the global representation of the image to retrieve the captions related to the query image. Then, the relevant caption is produced using the direct estimate of image contents. The clustering-based retrieval method is used in [16] to cluster similar images and associated phrases together using the scores of visual and semantic similarity contents. Then, the caption for the target image is retrieved from the cluster with similar images. This method can generate syntactically correct descriptions, but it cannot generate semantically appropriate image-specific captions.

The contemporary caption generation method typically follows either machine learning or deep learning models to generate the descriptions. This approach first examines the visual image content and produces descriptions using a language model by feeding the extracted visual image content. This method typically follows the encoder-decoder architecture to generate the description. In [17], the authors proposed a complete neural network-based model, in which a CNN network generated the image representation and LSTM generated the sequence of texts. The authors of [18] embedded visual attention mechanisms into the previous model with two variants: soft deterministic attention and hard stochastic attention. RNN embedded with the attention

technique significantly improved the model performance. Inspired by the effect of the attention mechanism in caption generation, the bidirectional self-attention mechanism for caption generation is mentioned in [19]. In the prior model, the attention is computed on both forward direction and backward direction to particularly focus on the relevant information.

ResNet50 architecture with CNN and soft attention-based LSTM is used in [20]. The soft attention mechanism is adopted to concentrate on the specific image regions for predicting the next word in the caption. The framework introduced in [21] exploited DenseNet-based CNN for global image feature representation. Also, the LSTM is applied as the language decoder an adaptive attention strategy is introduced to determine whether to consider the features of the image in sentence prediction. The novel image captioning approach can describe the image with semantically more precise captions when compared to previously discussed approaches. Machine learning-based caption generation models are demonstrated in [22]. The authors experimented by fusing several distinct pretrained models. To compare the various outputs produced by the model, a variety of pretrained CNNs and embedded matrices are employed. Although the solution to the problem of automated image description generation yields acceptable results, there is still space for improvement because certain photos are not well described. Metric-oriented focal mechanism (MFM) [23] is introduced to assist the caption generation model in focusing more on the complex examples during training. To gauge the difficulty of examples, MFM used the procedural metrics of captioning, and during training, it can increase the weights of challenging examples. The survey presented by [24] provides an in-depth analysis of picture captioning methods, including image encoding and text generation, as well as training methodologies, datasets, and assessment measures. The study concludes that there are still many unresolved issues including accuracy, robustness, and generalization outcomes. Also, novel captions can be produced from the multimodal visual space. Even though these recent models are showing great progress in image captioning research, most of the existing works are aimed at generating a subjective description of the image instead of exhibiting the emotional content present in the image.

2.2. Image Captioning with Emotions. This subsection provides a short review of some of the works that have integrated the emotional aspects into the caption generation model. SentiCap [11] designed a system to generate descriptions with emotions based on positive and negative sentiments. The model combined two CNN and RNNs running in parallel, whereas one combination is used to capture the factual content generation, and another one is used to generate the sentimental content. For training this model, sentiment vocabulary and rewriting of original captions are required including the sentiment-related terms and phrases.

In contrast, Face-Cap [10] captured human facial expressions from the image and incorporated the emotions to generate the captions. There are four variants of the Face-Cap model: Face-Step model, Face-Init model, Face-Cap$_F$ model, and Face-Cap$_L$ model. For all time steps including the initial one, the Face-Cap$_F$ model fed the facial features and the LSTM used these features to generate each word. In the Face-Cap$_L$ model, the memory cell vectors and hidden state vectors are initialized with the facial feature vectors. Face-Step is a variant of the Face-Cap$_F$ model without computing the face loss function. Face-Init is the variant of the Face- Cap$_L$ model, in which the facial features are applied at the initial step. All of these models used the standard CNN network for extracting image features and the standard LSTM network for creating captions. Extending this work, the Face-Attend model with attention mechanism is introduced in [25]. The model used two linked LSTMs: one for capturing the features of facial expressions and the other for general visual features using the Up-Down-Captioner. Even though the existing systems achieved greater results, still there are ways to produce more relevant and comprehensive captions by extracting a rich set of features.

3. The Proposed Caption Generation Model

This section investigates the novel automatic caption generation model based on human emotions using CSPDenseNet and BiLSTM with the self-attention mechanism. The main motive of the proposed model is to generate a fine-grained caption revealing the emotional content present in the image. The conceptual framework depicted in Figure 1 represents the novel end-to-end caption generation prototype. The two units, namely, encoder unit and decoder unit are the two core components of the entire image captioning system. Especially, the encoder architecture incorporates CSPDenseNet and FER models to extract the visual features and facial emotion features, respectively. The decoder architecture uses BiLSTM for language decoding to generate a description for the given input image. To enhance the prediction of the next word in the sequence, the self-attention mechanism is embedded in the decoder to concentrate on the semantic content based on its importance in the particular context. The complete interpretation of the proposed model is depicted in the following subsections.

3.1. Image Encoder

3.1.1. Emotional Feature Extraction. The system adapts the Facial Expression Recognition(FER) model proposed in [10] to obtain the emotional content present in the input image. The FER model is learned on the FER-2013 dataset to obtain the emotional attributes from the images in image captioning datasets. Before training the model, face preprocessing is applied to make the samples uniform in both datasets. The process of face preprocessing involves the following steps:

(1) Identify the faces in the image by applying a face detection algorithm using the CNN model.
(2) Crop the identified faces from every instance.
(3) Transform the cropped faces to grayscale.

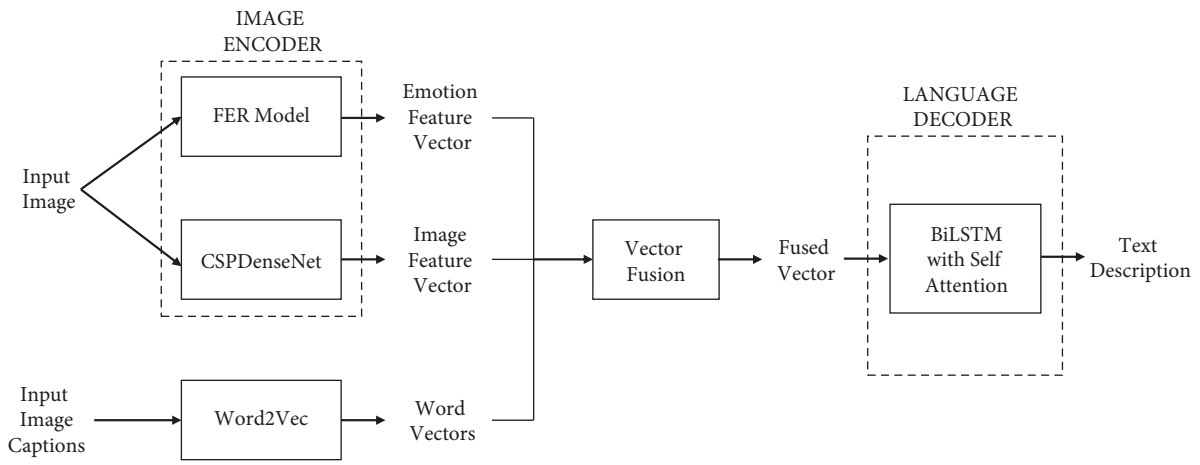


Figure 1: Caption Generation framework using CSPDenseNet-BiLSTM-Self-Attention network.

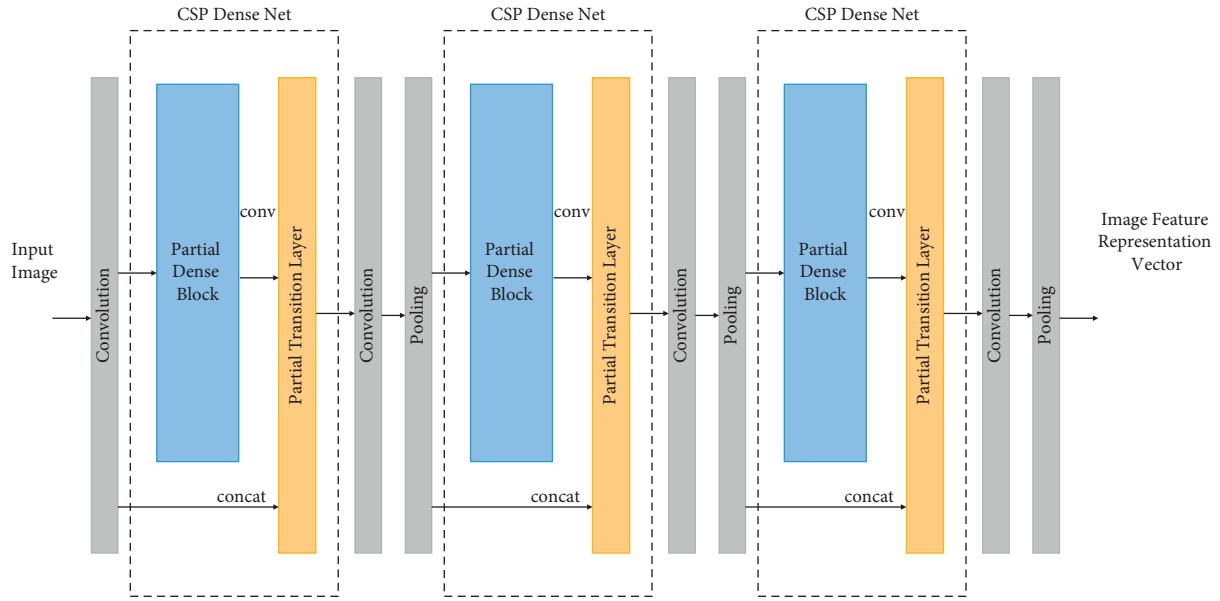


Figure 2: CSPDenseNet architecture with partial dense blocks and partial transition layers.

(4) Resize it to 48-by-48 pixels to match the size of the images in the FER-2013 dataset.

The VGGnet model [26] is trained on the FER-2013 data to identify facial emotions. There are seven emotional classes present in the FER-2013 dataset comprising happy, sad, surprise, fear, disgust, anger, and neutral. To represent the class distribution probabilities of these emotions, the output layer of the model contains seven neurons. The probabilities are referred to as the vector $p = (p_1, p_2, \ldots, p_7)$. Then, the trained model is used to extract the probabilities of emotions in the preprocessed image dataset. Finally, an aggregated one-hot encoding vector of emotion features $e = (e_1, e_2, \ldots, e_7)$ is constructed for each image as given in equation (1).

$$e_k = \begin{cases} 1, & \text{for } k = \arg\max \sum_{1 \le i \le n} p_i, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

In equation (1), the variable n represents the total face count in the image.

3.1.2. Extraction of Image Features. To extract the factual content of the image, the system uses the CNN-based object detector called *Cross Stage Partial Dense Network (CSPDenseNet)* introduced in [12]. CSPNet is a cornerstone in enhancing the learning ability of CNN. The rationale behind this model development is two-fold: (i) to attain a strong gradient combination and (ii) to reduce the computation overhead. To achieve these two aspects simultaneously, the base layer feature maps are divided into two subvectors for processing. Then, the processed subvectors are integrated using the cross-stage hierarchy. Figure 2 illustrates the architecture of CSPDenseNet. A single phase of CSPDenseNet contains two components: (1) a partial dense unit/block and (2) a partial transition layer. To balance the computation of each layer, the partial dense block is constructed in such a way that increases the gradient path and

minimizes the memory traffic. On the other hand, to amplify the variance of gradient combination, the partial transition layer is added to the construction. The working methodology of partial dense block and partial transitional layer is mentioned in the following steps.

(i) Partial Dense Block

(1) The base layer feature vectors(x_0) are partitioned into two subvectors via a medium, $x_0 = [x_0', x_0'']$.
(2) The subvector x_0'' is directly connected to the end of the phase.
(3) Another subvector x_0' will undergo the dense block.

(ii) Partial Transition Layer

(1) The input fed into a transition layer is the output obtained from dense layers $[x_0'', x_1, x_k]$.
(2) The additional input fed into the transition layer is the output obtained from the previous transition layer (x_T) concatenated with x_0'.
(3) Finally, the transition layer generates the output x_U, which will undergo further processing of the next CSPDenseNet unit after passing through convolution and pooling layers.

Then, the following equation (2) is used to update the feed-forward pass where $*$ denotes the convolution operation, the square brackets([ ]) denote the concatenation of the input vectors, and x_i and w_i denote the outputs and weights of the dense layer(i) respectively.

$$\begin{aligned} x_k &= w_k * [x_0'', x_1, \ldots, x_{k-1}], \\ x_T &= w_T * [x_0'', x_1, \ldots, x_k], \\ x_U &= w_U * [x_0', x_T], \end{aligned} \tag{2}$$

$$\begin{aligned} w_k' &= f(w_k, g_0'', g_1, g_2, \ldots, g_{k-1}), \\ w_T' &= f(w_T, g_0'', g_1, g_2, \ldots, g_k), \\ w_U' &= f(w_U, g_0', g_T). \end{aligned} \tag{3}$$

Equation (3) is used for weight updating where f is the weight updating function, g_i denotes the gradient passed to the dense layer(i). The gradients arrived through the dense layers are combined independently as given in the equation. Also, the integration of feature vectors that did not pass through the dense layers (x_0) is done independently and the weight update does not contain any redundant gradient information. In the proposed work using CSPDenseNet, the output image feature vectors are directly obtained from the pooling operation. The image features are extracted for each region of the image represented by a D-dimensional vector. Finally, it produces an image feature representation vector for the whole image in the form of L number of vectors, $r = \{r_1, r_2, \ldots, r_L\}, r_i \in \mathfrak{R}^D$ where L is the total count of regions in the image.

3.2. Vector Fusion. The language decoder takes the input as the fusion of three vectors: the facial emotion feature vector e, the image feature vector r, and the word vector obtained by training the training captions using word2vec model V . Word2Vec is a technique in NLP to learn word embeddings for the given text using a shallow neural network. For each time step t, the vectors are fused to obtain $x_t = \{e_t, r_t, V_t\}$, and this vector is fed to the language model.

3.3. Language Decoder. In the proposed caption generation system, the fusion of BiLSTM and Self-attention strategy is applied to create the description for the input image. Figure 3 depicts the pro-posed language decoder architecture embedded with self-attention. The three prominent layers present in the proposed decoder model are (i) BiLSTM, (ii) self-attention, and (iii) softmax activation layer. The working methodology of these three layers is discussed in the subsequent subsections. In the caption generation process, the words are generated for the current time step t using the emotion features associated with weights, image features associated with weights, the words induced at the previous time step $t-$ 1, and the actual reference image captions.

3.3.1. BiLSTM Layer. The BiLSTMs in the BiLSTM layer are composed of two individual LSTMs for processing the information of a text sequence from two directions. One of the LSTMs is meant to process the information from the forward direction and another one is meant to process the information from the backward direction of a sequence and then combine both the information gained to feed it to the next level. LSTM is a unique variant of RNN, specially designed to retain the long-term dependencies to resolve the gradient vanishing problem in traditional RNN networks. The LSTM neural network consists of a cell memory gate to keep track of values over changing time intervals and three inherent gates, namely, input, output and forget gate to control the information flow. Similar to $Face - Cap_L$ in [9], the proposed model computes the input gate (i_t), output gate (o_t), forget gate (f_t), hidden state (h_t), input modulation gate (g_t), and cell memory state (c_t) for current time step using the word embedding vector (x_{t-1}) from precedent time step, hidden state vector (h_{t-1}) from precedent time step and the factual features of the image (r_t) as given in equation (4).

$$\begin{aligned} i_t &= \sigma(W_i{}^* x_{t-1} + H_i{}^* h_{t-1} + R_i{}^* r_t + b_i), \\ f_t &= \sigma\big(W_f{}^* x_{t-1} + H_f{}^* h_{t-1} + R_f{}^* r_t + b_f\big), \\ o_t &= \sigma(W_o{}^* x_{t-1} + H_o{}^* h_{t-1} + R_o{}^* r_t + b_o), \\ g_t &= \sigma(W_c{}^* x_{t-1} + H_c{}^* h_{t-1} + R_c{}^* r_t + b_c), \\ c_t &= f_t{}^* c_{t-1} + i_t{}^* g_t, \\ h_t &= o_t{}^* \tanh(c_t). \end{aligned} \tag{4}$$

In the above equation, t denotes the particular time step, W,H, and R are learned weights, b is the bias and the logistic sigmoid function is represented by the variable σ. The LSTM learns to generate the next word automatically based on the emotion features fixed at all time steps as mentioned in equation (4). Initially, the facial emotion features are fed into two independent multilayer perceptrons for initializing the

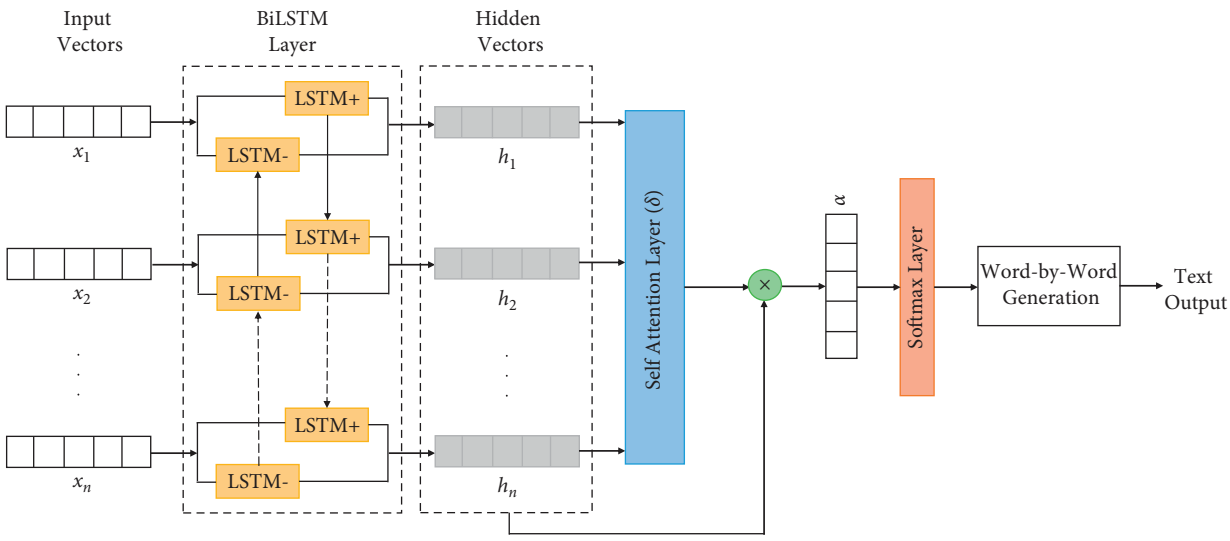


Figure 3: The architecture of BiLSTM with self-attention mechanism.

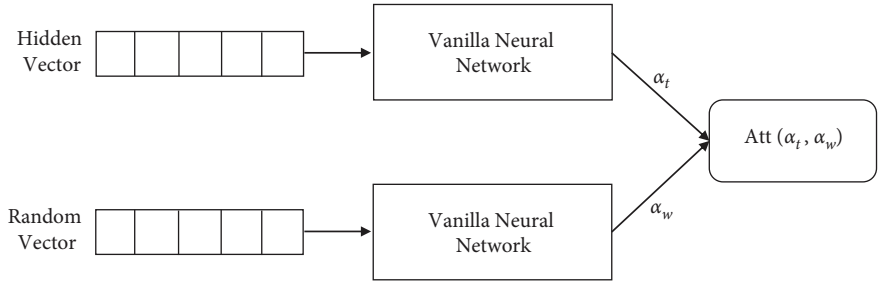


Figure 4: Basic flow diagram of self-attention mechanism.

values of hidden state (h_0) and cell state (c_0) as given in equation (5).

$$c_0 = \tanh_{init,c}(e),$$
$$h_0 = \tanh_{init,h}(e). \tag{5}$$

In BiLSTM, the forward LSTM evaluates the hidden state vector hf_t using x_t (input embedding vector) and hf_{t-1} (previous hidden state vector). Similarly, the backward LSTM evaluates the hidden state vector hb_t using x_t and hb_{t-1} from the reverse direction. Then the terminal hidden state vector of the BiLSTM is obtained by combining the forward and backward hidden vector (hf_t and hb_t) as shown in equation (6).

$$h_t = [hf_t, hb_t]. \tag{6}$$

3.3.2. Self-Attention Layer. The typical BiLSTM cannot identify the salient features required to generate the captions based on the particular context. To capture text features that are more related to the contextual information, the mechanism called self-attention is added to the proposed decoder model. The self-attention mechanism allows the input to interact with each other and extracts the important features by assigning more weight to indicate the significance of the feature. The hidden vector h_t obtained from the BiLSTM layer is fed as input to the layer of Self-attention. Figure 4 depicts the fundamental flow diagram of the self-attention mechanism, where the vanilla neural network is a simple Multilayer Perceptron neural network and *Att* represents the function used to compute the similarity between the new hidden vector and context vector. The equations of the self-attention mechanism are mentioned in (7).

$$\alpha_t = \tanh(W_w * h_t + b_w),$$
$$\delta_t = \frac{\exp(\alpha_t * \alpha_w)}{\sum_t \exp(\alpha_t * \alpha_w)}, \tag{7}$$
$$a = \sum_t \delta_t * h_t,$$

where W_w and b_w denote the weight and bias, respectively, which is used to find the new representation of hidden vector(α_t). α_w represents the high dimensional context vector used to predict the significance of different features in the text sequence, which is initialized randomly during the training phase and learned jointly.

3.3.3. Softmax Layer. The softmax layer in the decoder is used to generate a word for the next timestep based on the resulting output from the self-attention layer. At each time step t, it selects the words with high probability and sends

them to the next time step for generating the complete text caption. The functionality of softmax activation used in the proposed model is defined in equation (8).

$$y_t = \text{soft}\max(W_a * a + b_a), \tag{8}$$

where y_t,W_a, and b_a represent the weighted matrix and the bias, respectively.

4. Experimental Results

4.1. Datasets. For the experimentation purpose, two benchmark datasets are utilized: (1) the FER-2013 (Facial Expression Recognition) dataset to train the emotion identification model and (2) the FlickrFace11k dataset to train the caption generation model.

The FER-2013 dataset comprises 35,887 face images in a grayscale of size 48×48 pixels. Among these, 25,109 images are selected for model learning, and the remaining 3,589 images are utilized for testing and performance comparison with other recent caption generation models. The images are annotated based on the facial emotions classified as *Happy, Sad, Surprise, Disgust, Fear, Angry and Neutral.*

The FlickFace11k dataset is the subset of the Flickr30k dataset created by [9]. It is composed of 11,696 samples including human faces. The split up of the dataset is 8696 for learning and 3000 for validation and testing.

4.2. Experimental Settings. The hyperparameters of the neural network models play a vital role in achieving great results. The image encoder uses the CSPNet applied on DenseNet121 with three number of Dense blocks and three number of transition layers. The input image vectors are fed to the convolution layer with 2000 filters, each of size 7×7 and the stride value of 2. Then, a maximum pooling layer with a stride value of 2 and window size of 3×3 is placed to evaluate the largest patch value of the feature vector. The output feature vector from the max pooling layer is given as input to the CSPDenseBlock1. An average pooling layer of size 7×7 is placed after three consequent dense blocks and transition layers to obtain the final image feature vectors. The learning rate for CSPDenseNet is $1e - 5$. In the BiLSTM language decoder, the number of hidden layers is 50, and the dimension of the cell memory and the hidden state is 512. The word embedding vector size is set to 300. To avoid model overfitting, the dropout rate is fixed at 0.1 and the learning rate is fixed at 0.001. The momentum value is assigned as 0.8, the weight value is assigned as 0.999, and the batch size is initialized to 20. The attention dimension is assigned to 100 and the model is learned for 20 epochs and Adam optimizer is used to speed up the training process with a gradual decrease in learning rate.

4.3. Model Training

4.3.1. Face Loss Function. During the training phase, the hidden vector h_t is used to compute the negative log-likelihood of e in every time step t. This is known as the face loss function adapted from [9] and the loss value is calculated and averaged over all time steps as given in (9). Based on this method, the record of e, x_{t-1}, and r_t can be logged at every time step.

$$\text{Loss}(h_t, e) = -\sum_{1 \le i \le 7} 1_{(i=e)} \log\left(p(i|h_t)\right). \tag{9}$$

4.3.2. Caption Analysis. In the model training phase, the back propagation method is used to trace back the error of BiLSTM in two directions. (1) From current time t, the statistical error value at every time step is computed. (2) The error is estimated by moving the statistical error to one layer in advance. The weight gradient value is computed for every weight using the respective error value. At last, a stochastic gradient descent algorithm is applied for updating the parameters. The objective loss function is derived using the cross-entropy error as mentioned in equation (10).

$$E = -\sum_{1 \le i \le n} (x_i)\log(p(x_i)). \tag{10}$$

In equation (10), n denotes the distinct verbal count generated by the model, E represents the entropy score, and $p(x_i)$ represents the probability of every distinctive verb x_i.

4.4. Evaluation Metrics. The proposed image captioning model is evaluated using five well-known standard metrics such as BLEU [27], METEOR [28], ROUGE-L [29], CIDEr [30], and SPICE [31]. BLEU is language independent measure used to evaluate a machine-generated text to the original reference text. It calculates the precision of an n-gram(n =1, 2, 3, 4) among the reference captions and generated caption. The scores can be computed using the reference caption length, generated caption, modified precisions of n-gram, and uniform weights.

METEOR is one of the standard evaluation metrics meant for machine translation. The score can be computed using the harmonic mean of precision and recall for words, where the weightage of recall is higher than precision.

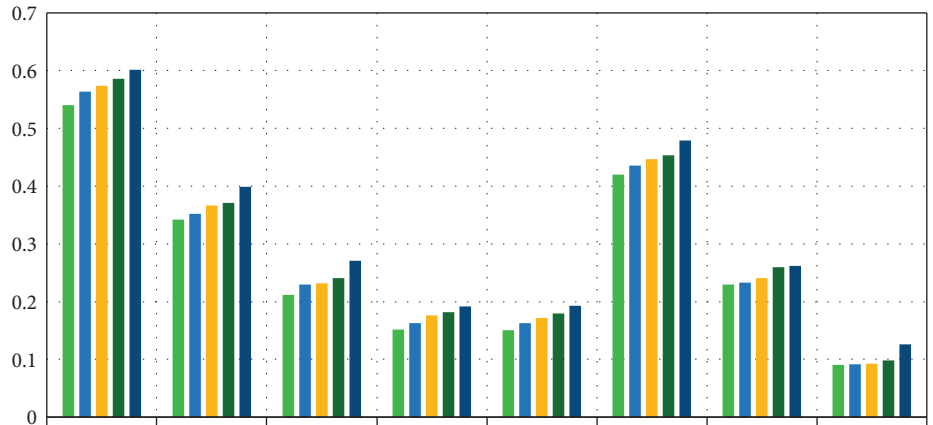
ROUGE-L is a scoring algorithm that computes the similarity between the actual caption and the set of reference captions using the Longest Common Subsequence. It determines the quality of generated caption using the similarity of the longest cooccurring n-gram sequence.

CIDEr calculates the similarity among the actual reference captions and the machine-generated captions. CIDEr measure considers grammaticality, accuracy, and saliency.

SPICE evaluates the quality of the caption by constructing the scene graphs for both the original and generated caption. The graph indicates the semantic representation of the captions.

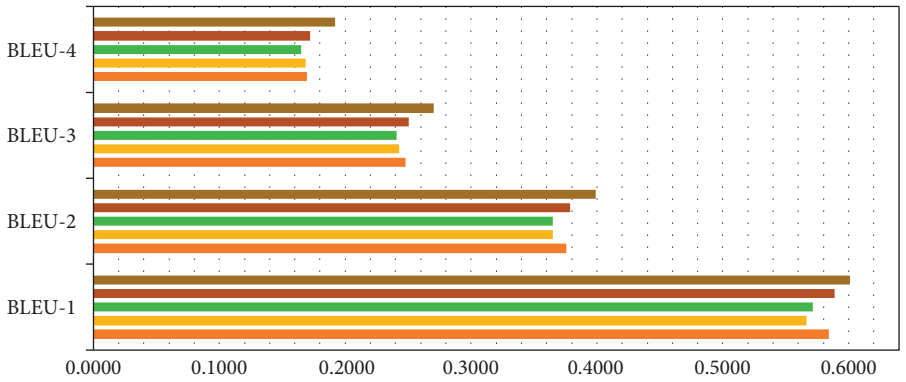
For all the evaluation metrics, the score value spans from 0 to 1. The value of 1.0 denotes the perfect match and 0.0 indicates the perfect mismatch between the generated caption and the reference caption.

4.5. Analysis of Proposed Model. The proposed model is evaluated based on five different combinations: CNN-LSTM (extracting features of an image using standard CNN and generating captions using conventional LSTM decoder),



	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN-LSTM	0.5403	0.3421	0.2112	0.1512	0.1501	0.4201	0.229	0.0899
CNN-BiLSTM	0.5631	0.3519	0.2293	0.1631	0.1624	0.4354	0.2332	0.0919
CSPDN-LSTM	0.5733	0.3661	0.2319	0.176	0.1712	0.4462	0.2411	0.0922
CSPDN-BiLSTM	0.5862	0.3712	0.2402	0.1812	0.1799	0.4539	0.2591	0.0981
CSPDN-BiLSTM-SelfAtt	0.6012	0.3992	0.2703	0.1921	0.1932	0.4793	0.2617	0.1260

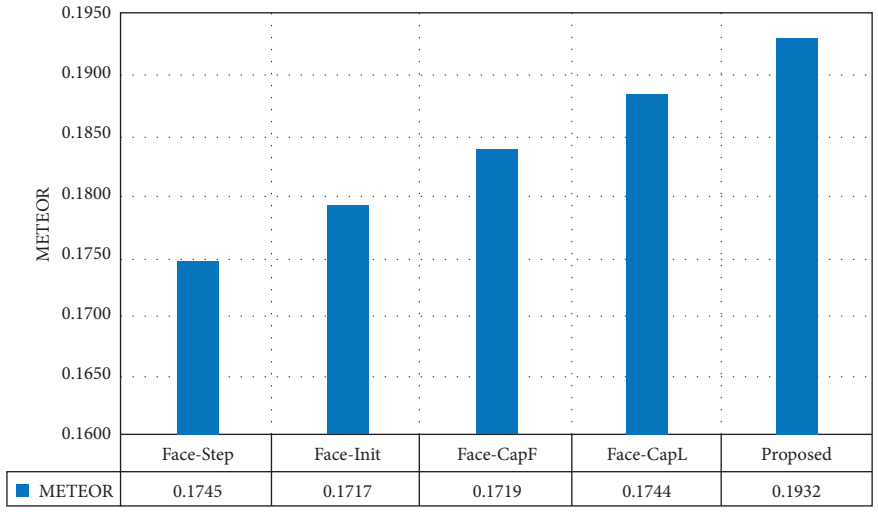
Figure 5: Performance comparison of image captioning models on FlickrFace11k dataset.



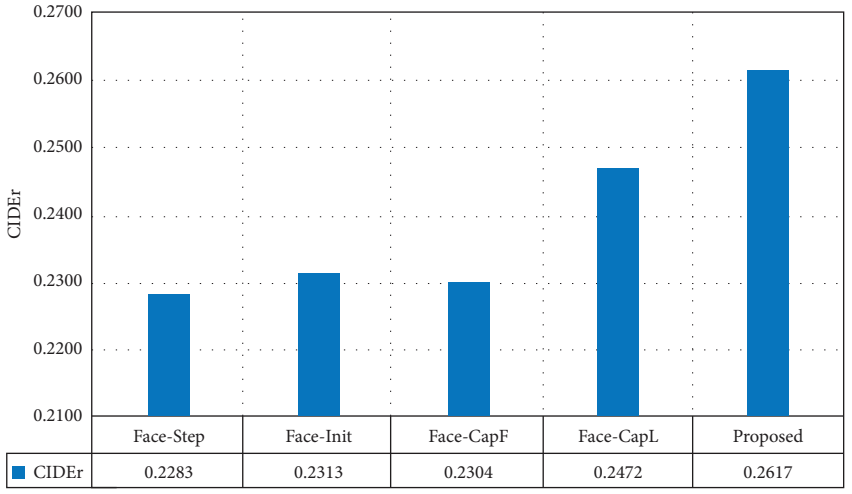
	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Proposed	0.6012	0.3992	0.2703	0.1921
Face-CapL	0.5890	0.3789	0.2507	0.1719
Face-CapF	0.5713	0.3651	0.2407	0.1652
Face-Init	0.5663	0.3649	0.2430	0.1686
Face-Step	0.5843	0.3756	0.2478	0.1696

Figure 6: Performance comparison based on BLEU-N scores for N = 1, 2, 3, 4.

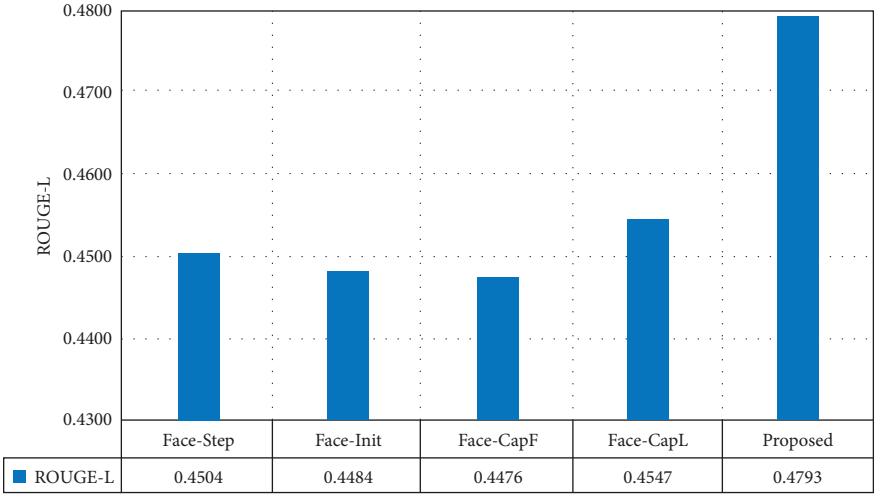
CNN-BiLSTM (extracting features of an image using standard CNN and generating captions using conventional BiLSTM decoder), CSPDN-LSTM (extracting image features using CSPDenseNet and generate captions using LSTM decoder), CSPDN-BiLSTM (extracting image features using CSPDenseNet and generate captions using BiLSTM decoder), CSPDN-BiLSTM-SelfAtt (extracting image features using CSPDenseNet and generate captions using BiLSTM decoder with self-attention mechanism). Figure 5 illustrates the performance comparison of the above-discussed different models on the FlickrFace11k dataset. The results show that the performances of the models increase continuously for each evaluation metric. The performance of the basic LSTM model is low since the traditional LSTM model cannot capture semantic information effectively from the sentence. The BiLSTM model comprises a forward LSTM and a backward LSTM, which can collect more information from two directions. Hence, it can capture more semantic information needed to describe the image. The BiLSTM model increases the accuracy of generated captions when compared to the basic LSTM model. But still, the traditional BiLSTM alone cannot capture the information of a particular context. To capture the context information within the sentence, the self-attention mechanism is added to the



(a)

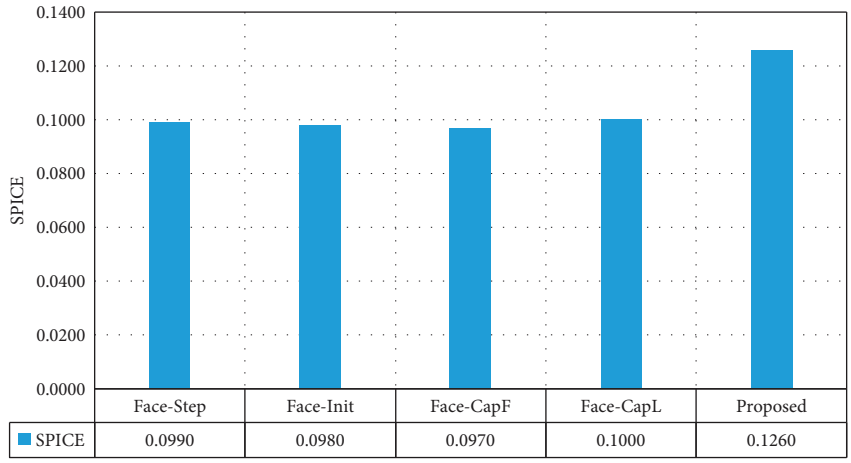


(b)



(c)

Figure 7: Continued.



(d)

Figure 7: Comparison graphs of METEOR, CIDEr, ROUGE-L, and SPICE scores on the Flickr dataset. (a) METEOR. (b) CIDEr. (c) ROUGE-L. (d) SPICE.

traditional BiLSTM model. And, it is observed that CSPDN-BiLSTM-SelfAtt shows better performance, which implies the significance of the attention mechanism in enhancing the accuracy of caption generation.

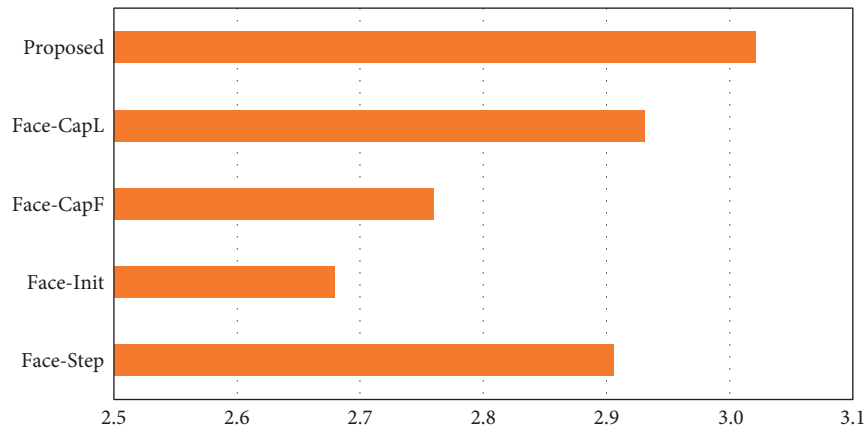
4.6. Comparison with the Baseline Models. This section examines the performance comparison of the presented image captioning model with the four variants of the baseline face captioning model (Face-Cap) discussed earlier. Figure 6 illustrates the performance comparison based on BLUE scores. Figure 7 depicts the comparison results based on METEOR, CIDEr, ROUGE-L, and SPICE scores. The comparison of entropy values of all the verbs generated using the different models is displayed in Figure 8. For computing the BLEU-N scores of the different models, they are tested with four N values as $N=1$ represents unigrams, $N=2$ represents bigrams, $N=3$ represents trigrams and $N=4$ represents four-grams. The score is evaluated based on the refined precision metric for n-grams which considers the two important aspects of translation: fluency and adequacy. The matching of longer n-grams between the reference caption and generated caption determines the fluency. The adequacy is decided based on the count of words present in the generated caption that is the same as that of the reference caption. From Figure 6, it is noticed that the BLEU scores of the present model are superior for all four N-grams in comparison with the other baseline models. And, the image captions generated by the newly introduced model are more adequate and fluent.

Figure 7(a) presents the METEOR score comparison of the presented model with other models. The METEOR score is computed using the explicit word-to-word matching of the reference captions and the machine-generated caption. It first identifies all the word (unigram) matches between both the captions. For the identified word matches, the score is evaluated by combining the word precision, word recall, and a fragmentation measure. The fragmentation measure is used to capture the explicit ordering of the matched unigrams in the generated and reference caption. From the comparison scores, it is clear that the current model generates captions with more precision and recall in comparison with the baseline emotion models. And, the words generated by the proposed model follow the explicit word-to-word occurrence equivalent to the reference captions.

ROUGE-L measure uses the LCS (Longest Common Subsequence) based F-score to evaluate the relatedness between computer-generated captions and ideal reference captions. It considers only the in-sequence matches. The ROUGE-L comparison graph of the different models including the currently presented model is depicted in Figure 7(c). The presented model has a better LCS-based ROUGE score when compared to the other models. The LCS present in the computer-generated caption matches with the LCS in the original reference captions.

CIDEr evaluates the image descriptions using a consensus-based evaluation strategy, which makes use of human consensus. The similarity score for the generated description is computed based on the ground truth human descriptions of the image. It also captures the three main aspects of sentence similarity including grammaticality, importance, saliency, and accuracy. From Figure 7(b), it is observed that the proposed model generates image descriptions satisfying the above aspects of sentence similarity. The CIDEr score of the proposed model is better when compared to the baseline models. And, it is recognized that the generated descriptions are more similar to most of the ground truth descriptions composed by humans.

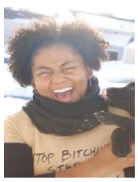
SPICE metric incorporates the semantic propositional component for evaluating the image descriptions. To exploit the semantic structure and meaning, the metric constructs a scene graph to encode the semantic propositional content of the generated and reference captions. F-score is used to calculate the similarity between the scene graphs of the machine-generated caption and the reference captions. The SPICE scores of the proposed and other caption generation



	Face-Step	Face-Init	Face-CapF	Face-CapL	Proposed
Entropy	2.9059	2.6792	2.7592	2.9306	3.0214

Figure 8: Entropy comparison for all the verbs generated using different models.

Face-Step : A young boy in red shirt is running with a dog
Face-Init : Adog is chasing a young boy
Face-CapL : A young boy in red shirt is playing with a dog
Face-CapF : A young boy in red shirt is running with a dog and playing
Proposed : A young boy in red shirt is playing Frisbee with a dog in the grass



Face-Step : A woman with yellow shirt is laughing with a dog
Face-Init : A woman with yellow shirt is laughing and holding a dog
Face-CapL : A woman with yellow shirt is laughing with a dog in her hand
Face-CapF : A woman with yellow shirt is laughing and holding a dog
Proposed : A woman with a black scarf is laughing as a black dog try to lick her

Face-Step : A group of people are smiling
Face-Init : A group of people are smiling at a camera
Face-CapL : A group of people are smiling
Face-CapF : A group of people are smiling at a camera
Proposed : A group of people holding bottles are smiling at a camera

Face-Step : A young girl with pink shirt is running
Face-Init : A young girl with pink shirt is playing
Face-CapL : A young girl with pink shirt is playing in grass
Face-CapF : A young girl with pink shirt is running
Proposed : A young girl with pink shirt and green pants is running through a flower field

Face-Step A young man in red shirt is jumping in water
Face-Init A young man in red shirt is playing
Face-CapL A young man in red shirt is playing in water
Face-CapF A young man in red shirt is surfing
Proposed A young man in red life vest is smiling and riding wakeboard

Figure 9: Sample captions generated using different models for the images in the Flickr dataset [Source: Kaggle, https://www.kaggle.com/hsankesara/flickr-image-dataset].

models are shown in Figure 7(d). From the graph, it is noticed that the SPICE score of the current model is better in comparison with the scores gained by other models. Hence, the proposed model generates captions that can effectively capture the objects and their characteristics present in the image to exhibit the relation between them. The comparison based on entropy values of all verbs generated using different models is depicted in Figure 8. The highest entropy value shows that the proposed system generates verbs with a more diverse distribution when compared to other models. Figure 9 demonstrates the captions generated by the rival methods and the proposed system for the images in the Flickr dataset. The proposed model first identified the emotional features from the images using the FER model. Then, more comprehensive and fine-tuned image features are obtained using the CSPDense neural network. The


Face-Step : A group of people are sitting at night
Face-Init : A group of people are sitting with light
Face-Capı : A group of people are sitting in fire
Face-Capı : A group of people are sitting in fire
Proposed : A group of people are singing in night fire


Face-Step : A young boy is smiling
Face-Init : A young boy is laughing
Face-Capı : A young boy is laughing
Face-Capı : A young boy is laughing
Proposed : A little boy with blue shirt and green glass is laughing

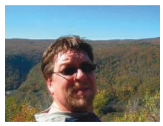

Face-Step : A man is looking at a camera
Face-Init : A man is with black cap is smiling
Face-Capı : A man is smiling at a camera
Face-Capı : A man is with black cap is smiling
Proposed : A man is smiling and standing in a mountain

Figure 10: Inappropriate captions generated using different models for the images in the Flickr dataset [Source: Kaggle, https://www.kaggle.com/hsankesara/flickr-image-dataset].

extracted emotion and image features along with the word embedding vector(from training captions) are sent to the language decoder for generating the description for the input image. There exists a complex relationship between objects and scenes in the images presented in Figure 9. Self-attention plays a vital role in selecting salient features and their correlations with each other. By analyzing the descriptions generated by the proposed model, it is observed that the captions exhibit the emotional content of the images such as *smiling* and *laughing*. The model also generates *playing* since it is an indication of positive emotion. Hence, the proposed model generates the captions which signify the emotional state of the people, the link between an object and the scene, and the object and object features along with their behavior. It also generates captions with appropriate grammar in simple language. Figure 10 depicts the inappropriate captions generated by the proposed methods and other comparative models. From the figure, it is inferred that the proposed model can identify the objects, relationships, and scenes better when compared to the other models but the facial emotions are not recognized clearly using the basic FER model. Irrelevant phrases and verbs such as smiling, singing, and laughing are encountered irrespective of the facial features in the sample images.

5. Conclusion

Caption generation with emotions paved the way to explore a wider scope of image captioning applications. In this work, a novel encoder-decoder architecture for the automatic caption generation process is introduced by incorporating human emotions extracted from facial features. Human facial emotions are captured and emotion feature vectors are created using a FER model trained on the FER-2013 dataset. CSPDenseNet, a new variant of CNN is adopted for encoding the image to capture the more intense feature vectors. The cross-stage feature fusion strategy is employed in CSPDenseNet which makes it lightweight to run on CPUs. A Word2vec model is developed and trained using the human-annotated captions to extract the word feature vectors. Finally, the extracted emotion, image, and word feature vectors are fused and fed into the language decoder. The language decoder employed BiLSTM with self-attention to produce the captions. The BiLSTM network is implemented to extract the semantic knowledge and self-attention is incorporated to focus on the salient contextual features in the text. The experimental outcomes demonstrated that the model proposed in this article generates image captions more effectively when compared to the state-of-art models in terms of BLEU, METEOR, CIDEr, ROUGE-L, and SPICE. The future work aims to develop novel facial expression recognition models that can capture a wider range of emotions. Then the extracted intense and relevant facial feature vectors can be infused to generate image descriptions efficiently.

Data Availability

All data used to support the findings of the study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to extend their gratitude to the Management and Principal of Mepco Schlenk Engineering College(Autonomous), Sivakasi, for providing ample facilities and assistance for this research project.

References

[1] S. K. Priya, G. Shenbagalakshmi, and T. Revathi, "IoT based automation of real time in-pipe contamination detection system in drinking water," in *Proceedings of the International Conference on Communication and Signal Processing*, pp. 1014–1018, Chennai, India, 2018.

[2] S. K. Priya, T. Revathi, K. Muneeswaran, and K. Vijayalakshmi, "Heuristic routing with bandwidth and energy constraints in sensor networks," *Applied Soft Computing*, vol. 29, pp. 12–25, 2015.

[3] H. Naeem and A. A. Bin-Salem, "A CNN-LSTM network with multi-level feature extraction- based approach for automated detection of coronavirus from CT scan and X-ray images," *Applied Soft Computing*, vol. 113, Article ID 107918, 2021.

[4] S. R. Sahoo and B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," *Applied Soft Computing*, vol. 100, Article ID 106983, 2021.

[5] D. Jain, A. Kumar, and G. Garg, "Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN," *Applied Soft Computing*, vol. 91, Article ID 106198, 2020.

[6] S. Barzut, M. Milosavljević, S. Adamović, M. Saračević, N. Maček, and M. Gnjatović, "A novel fingerprint biometric cryptosystem based on convolutional neural networks," *Mathematics*, vol. 9, no. 7, p. 730, 2021.

[7] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–36, 2019.

[8] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, 2018.

[9] H. Wang, Y. Zhang, X. Yu, and F. Solari, "An overview of image caption generation methods," *Computational Intelligence and Neuroscience*, vol. 2020, Article ID 3062706, 13 pages, 2020.

[10] O. Nezami, M. Dras, P. Anderson, and L. Hamey, "Face-cap: image captioning using facial expression analysis," in *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science*, M. Berlingerio, F. Bonchi, T. Ga¨rtner, N. Hurley, and G. Ifrim, Eds., Springer, Berlin, Germany, Article ID 11051, 2019.

[11] A. Mathews, L. Xie, and X. He, "SentiCap: generating image descriptions with sentiments," 2016, https://arxiv.org/abs/1510.01431.

[12] C. Wang, H. Liao, I. Yeh, Y. Wu, P. Chen, and J. Hseih, *CSPNet: A New Backbone that Can Enhance Learning Capability of CNN*, 2019, https://arxiv.org/abs/1911.11929.

[13] A. Farhadi, M. Hejrati, M. Sadeghi et al., "Every picture tells a story: generating sentences from images," in *European Conference on Computer Vision*, pp. 15–29, Springer, Berlin, Germany, 2010.

[14] G. Kulkarni, V. Premraj, S. Dhar et al., "Baby talk: understanding and generating image descriptions," in *Proceedings of the CVPR Means IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2891–2903, Colorado Springs, CO, USA, 2011.

[15] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: describing images using 1 million captioned photographs," *Advances in Neural Information Processing Systems*, vol. 24, pp. 1143–1151, 2011.

[16] C. Sun, C. Gan, and R. Nevatia, "Automatic concept discovery from parallel text and visual corpora," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2596–2604, Santiago, Chile, 2015.

[17] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: a neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, Boston, MA, USA, 2015.

[18] K. Xu, J. Ba, R. Kiros et al., "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, Lille, France, 2015.

[19] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, and M. Bennamoun, "Bi-San-cap: bi-directional self-attention for image captioning," in *Proceedings of the 2019 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–7, Perth, WA, Australia, 2019.

[20] Y. Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang, "Automatic image captioning based on ResNet50 and LSTM with soft attention," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8909458, 7 pages, 2020.

[21] Z. Deng, Z. Jiang, R. Lan, W. Huang, and X. Luo, "Image captioning using DenseNet network and adaptive attention," *Signal Processing: Image Communication*, vol. 85, Article ID 115836, 2020.

[22] B. Smolka, B. Predić, D. Manić, M. Saračević, D. Karabašević, and D. Stanujkić, "Automatic image caption generation based on some machine learning algorithms," *Mathematical Problems in Engineering*, vol. 2022, Article ID 4001460, 11 pages, 2022.

[23] J. Ji, Y. Ma, X. Sun, Y. Zhou, Y. Wu, and R. Ji, "Knowing what to learn: a metric-oriented focal mechanism for image captioning," *IEEE Transactions on Image Processing*, vol. 31, pp. 4321–4335, 2022.

[24] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: a survey on deep learning-based image captioning," 2022, https://arxiv.org/abs/2107.06912.

[25] O. Mohamad Nezami, M. Dras, S. Wan, and C. Paris, "Image captioning using facial expression and attention," *Journal of Artificial Intelligence Research*, vol. 68, pp. 661–689, 2020.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[27] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA, 2002.

[28] M. Denkowski and A. Lavie, "Meteor universal: language specific translation evaluation for any target language," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380, Baltimore, MD, USA, 2014.

[29] C. Lin, "ROUGE: a package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, 2004.

[30] R. Vedantam, C. Zitnick, and D. Parikh, "CIDEr: consensus-based image description evaluation," 2015, https://arxiv.org/abs/1411.5726.

[31] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," 2016, https://arxiv.org/abs/1607.08822.