Paper

High-level and Low-level Feature Set for Image Caption Generation with Optimized Convolutional Neural Network

Roshni Padate, Amit Jain, Mukesh Kalla, and Arvind Sharma

Department of Computer Science and Engineering, Sir Padampat Singhania University, India

https://doi.org/10.26636/jtit.2022.164222

Abstract — Automatic creation of image descriptions, i.e. captioning of images, is an important topic in artificial intelligence (AI) that bridges the gap between computer vision (CV) and natural language processing (NLP). Currently, neural networks are becoming increasingly popular in captioning images and researchers are looking for more efficient models for CV and sequence-sequence systems. This study focuses on a new image caption generation model that is divided into two stages. Initially, low-level features, such as contrast, sharpness, color and their high-level counterparts, such as motion and facial impact score, are extracted. Then, an optimized convolutional neural network (CNN) is harnessed to generate the captions from images. To enhance the accuracy of the process, the weights of CNN are optimally tuned via spider monkey optimization with sine chaotic map evaluation (SMO-SCME). The development of the proposed method is evaluated with a diversity of metrics.

Keywords* — *CNN, image caption, proposed contrast, sharpness, SMO-SCME algorithm

Tab. 1. List of terms and abbreviations.

AACR	Algebraic amalgamation-based composed representation
BI-LSTM	Bidirectional LSTM
BI-GRU	Bidirectional gated recurrent units
Bleu_1	Bilingual evaluation understudy
CNN	Convolutional neural network
CV	Computer vision
CMBO	Cat mouse-based optimization
DG-GAN	Dual generator GAN
DL	Deep learning
DBN	Deep belief network
FCA	Face centrality attribute
GLP	Global leader phase
GLL	Global leader learning phase
GLD	Global leader decision phase
gLSTM	Guiding LSTM
GAN	Generative adversarial networks
IGGAN	Interactions guided GAN
LLD	Local leader decision phase
LSTM	Long short-term memory
LLL	Local leader learning phase
LP	Learning percentage
LLP	Local leader phase
MK-KDES	Multiple kernel-kernel descriptors
MFO	Moth flame optimization
NLD	Natural language description
NLP	Natural language processing
NLG	Natural language generation
NN	Neural network
RMCNet	ResNet with multi-scale module and adaptive channel attention
SSA	Salp swarm algorithm
SSO	Shark smell optimization
SMO-SCME	SMO with sine chaotic map evaluation
sLSTM	Stacked LSTM
SMO	Spider monkey optimization
TR	Tag refinement
WHO	Wild horse optimizer

1. Introduction

Automated image captioning entails image capturing, examining its content, and generating a linguistic description [1]–[3]. Automated portrayal of image contents, with an appropriate level of expressiveness and accuracy, is a demanding task. In order for the captioning of an image to be effective, three fundamental requirements need to be satisfied, namely identifying attributes, objects, and their mutual associations [4]–[7]. However, identification of objects and their associations with images is insufficient for producing informative textual descriptions. Therefore, in order to recognize the content of a given image and to generate a text, an automated tool is necessary which would be capable of create various inferences from these identified objects and their associations [8]–[10]. An effective image captioning model aids in multimodal machine conversion and word sensing disambiguation by reducing the level of uncertainty in words [7], [11]–[13].

Three key architectures of DL-oriented image captioning schemes may be distinguished, namely end-to-end, attention-based, and compositional. The schemes deployed for captioning images are classified as those relying on the template-based matching and retrieval-based approaches. In the former, all attributes, actions, and objects visible in the image are derived and packed into a stiff sentence pattern [14], [15]. The yield of the template-oriented matching model is not always sufficient and clear. In the latter approach, a visually identical image is recovered from a larger database and the recovered image's captions are matched with the query image [16], [17]. This model is characterized by a low level of flexibility in terms of adjusting the recovered captions and lacks in expressiveness and fluency [7], [18].

The contribution of the presented work is as follows. The authors proposed a new model for extracting both low-level and high-level data from image captions. The said data is then used for caption generation with enhanced CNN and SMO-SCME model weight adjustments.

The paper is arranged as follows. Section 2 presents related works. Section 3 introduces the model concerned and Section 4 explains the extraction of the proposed features. Section 5 describes the optimal CNN with SMO-SCME-aided optimization. Sections 6 and 7 illustrate the outcome and offer conclusions, respectively. Table 1 summarizes the abbreviations used.

2. Literature Review

Singh *et al.* [7] proposed an encoder-decoder oriented framework in which CNN was deployed to encode visual features of an image and sLSTM was relied upon, in combination with bi-directional LSTM and unidirectional LSTM, to generate captions in Hindi. To encode the ocular features of the image, sLSTM and "V GG19" designs were deployed in the process of generating captions at the decoder side. The results have shown that the adopted method achieved better results.

Ye *et al.* [19] proposed an image caption generation scheme depending on the optimized BI-LSTM model. A variant of MFO was developed for optimization purposes. The performance of the adopted approach was proven using varied datasets, such as Flicker 8k, Flicker 30k, VizWik and Coco datasets, using renowned metrics, such as Cider, Bleu, Spice and Rough. Performance simulations have proven that B-LSTM achieved better results over the remaining schemes. Zhang *et al.* [20] presented a system for creating a novel feature, known as MK-KDES-1, with the extraction of 3 KDES features and MKL scheme fusing. MK-KDES-1 features contribute to enhancing the Bleu score of captions. The subsequent issue was resolved by a new, effective two-layer tag refinement (TR) approach incorporated into the NLG scheme. Analyses have proven the system to be suitable for creating image captions.

Sur *et al.* [21] defined AACR as a simplifying, language structuring and modeling linguistic attribute (associated with grammar and language) that improved the grammatical structure and corrected the sentences. AACR allowed structure and represent feature spaces in a more accurate and unique manner. Shan *et al.* [22] developed the IGGAN approach for captioning images in an unsupervised mode. This technique combined object-to-object communications with multi-scale feature representation. Images were encrypted using RMC Net to gain a robust representation of features. The created sentence and the image were deployed to reconstruct one another in IGGAN. The approach produced sentences with no manual labeled image caption pairs. Yiwei *et al.* [23] presented a multi-attention method by deploying both non-local and local evidence for more effective image caption analysis. This scheme, known as Magan, includes a discriminator and a generator. The adopted generator aided in generating more exact sentences. At the same time, the discriminator was utilized to determine if the sentences created were machine-generated or case-specific.

Yang *et al.* [24] presented a capable approach relying on the EnsCaption model. The solution intended to improve an ensemble of generation-oriented and retrieval-oriented image captioning schemes via an innovative DG-GAN network. By relying on the adversary training procedure, caption re-ranking and caption generation, it offered better retrieved captions with superior scores and allowed lower ranking scores to be assigned to the retrieved and created image captions, thus offering improved effectiveness over other schemes.

Zhao *et al.* [25] developed a multimodal fusion scheme to produce descriptions portraying image contents. The developed technique involved four networks: CNN for feature extraction, image attribute extraction, sentence modeling, and recurrent network. Unlike extant models that anticipate the specific words based upon hidden states, the presented approach deployed the attributes of the image and designed recorded words.

Ding *et al.* [41] deployed high-level visual characteristics to serve as a foundation of a novel image captioning model. To identify areas of an image that need attention, low-level data, such as image value, was blended high-level elements. Tests relying on MSCOCO and Flickr data sets proved that the model performed well. Table 2 shows the comparison of different image captioning schemes.

3. Proposed Image Caption Generation Model

The proposed image captioning approach comprises three stages. Initially, feature extraction is performed, deriving low-level (contrast, sharpness, color) and high-level (motion and facial impact score) features. The said features are then fed to CNN that generates captions based on images. The CNN weights are optimized via the SMO-SCME model, resulting in precise image captioning. The SMO-SCME model is depicted in Fig. 1.

Tab. 2. Comparison of conventional image captioning systems.

Reference	Deployed schemes	Features	Notes
[7]	CNN	• high efficiency • improved Bleu score	Not capable of captivating the alphanumerical image content
[19]	B-LSTM	• high Bleu score • enhanced Cider score	Occlusion may occur
[20]	MK-KDES-1	• improved Bleu score • creates more coherent features	Needs attention when dealing with visual image contents
[21]	AACR	• high Meteor score • high Cider score	Other constructive elements should be included
[22]	IGGAN	• robust representation of features • generates realistic sentences	Suffers from insufficient common sensical reasoning
[23]	Magan	• high Cider score • high Meteor score	Additional discriminators increase the memory usage
[24]	EnsCaption model	• high Rouge score • high Cider score	Ranking procedure needs to be improved
[25]	CNN	• chooses optimal captions • enhanced accuracy	No deliberation on denser image captions
[41]	High- and low-level features	• efficiency is high • accuracy is high	It is still difficult to describe visuals that have several objects in them

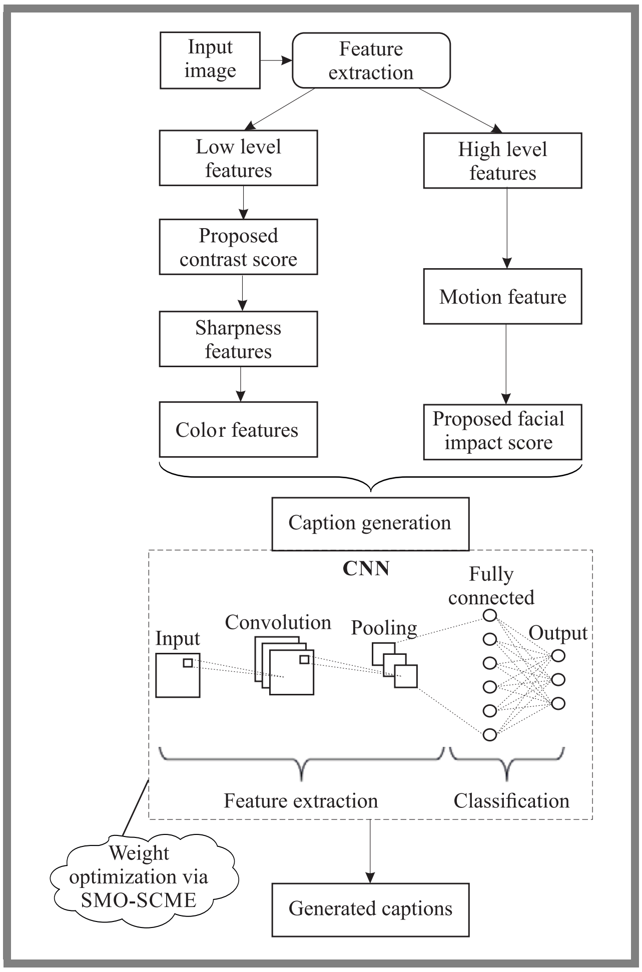


Fig. 1. Proposed image caption generation scheme.

4. Extraction of High-level and Low-level Features

Contrast [26] is defined as the ratio between the black and white parts of the image, representing the gradual change from black to white. The greater the contrast ratio, the more gradients there are, and the richer color in a given area. Initially, a gamma-corrected image is generated from the input image as:

$$R = CX^{\delta}, \tag{1}$$

where C is a constant positive constraint for controlling brightness, X refers to the original image (its values should be 0 and 1) and δ refers to a constant positive parameter which indicates the gamma value.

The formula for calculating is:

$$C = \sum_{y} \gamma(i,j)^2 P_{\gamma}(i,j), \tag{2}$$

where $P_{\gamma}(i,j)$ is probability of pixel distribution of gradation variation of γ between neighboring pixels.

The $\gamma(i,j)$ parameter in Eq. (2) is calculated:

$$\gamma(i,j) = |i-j|. \tag{3}$$

However, for the proposed contrast score $\gamma(i,j)$ is:

$$\gamma(i,j) = \frac{|\vec{\mu}-i|}{|\vec{\mu}-j|}, \tag{4}$$

which represents the difference between neighboring pixels. $\vec{\mu}$ implies the average of γ for neighboring pixels.

The candidate areas in the image for computing contrast score C are initially transformed to luminance, and are then filtered and re-sampled for obtaining width and height [26].

Color is a significant feature for representing an image. It is invariant, regardless of the image's rotation, translation, and scaling. Color spacing, quantification, and similarity computations are the key elements of extracting color features [27].

Image sharpness is an important parameter for evaluating image quality. Sharpness quality directly impacts the subjective feelings of humans [26]. In an image, sharpness refers to its overall clarity in terms of both focus and contrast [28]. The derived low-level features (sharpness, color, and contrast) are together represented as FT_{low}.

These features offer temporal data and are generally attained using optical flow for detection the inappropriate motion at the backdrop [29]. A motion processing analysis may be performed for detecting motion, i.e. finding those locations where anything is moving within the image [30].

Facial information plays a significant role in expressing image semantics [26]. The presence of a face in every candidate region is determined by the f score. Initially, a face denoted by f_s is standardized to a frame by:

$$FS = \frac{Wi_f^2}{Ht_{im} \times Wi_m}, \tag{5}$$

in which Wi_f^2 refers to the width of the face bound box in pixels, Ht_{im} refers to the height of the image and Wi_m refers to the width of the image.

In the next step, FCA is formulated based on the Gaussian formula as:

$$\text{FCA} = \frac{1}{2\pi\sigma_y\sigma_x}\text{e}^{-\frac{\left|\frac{\Delta y^2}{\sigma_y^2}+\frac{\Delta y^2}{\sigma_y^2}\right|}{2}}, \tag{6}$$

where $\text{Cen}Y_f$ and $\text{Cen}X_f$ are the centroid column and the centroid row of the facial area, respectively, and:

$$\sigma_x = \frac{2Wi_m}{3}, \tag{7}$$

$$\sigma_y = \frac{Hy_{im}}{2}, \tag{8}$$

$$\Delta y = \left|\text{Cen}Y_f - \frac{wi_m}{2}\right|, \tag{9}$$

$$\Delta x = \left|\text{Cen}X_f - \frac{Ht_{im}}{5}\right|. \tag{10}$$

The facial impact score is:

$$f = \alpha FS \times \beta\,\text{FCA}, \tag{11}$$

where α and β refer to scalars evaluated using logistic maps.

The derived high-level features (motion and proposed facial impact score) are indicated as FT_{high}.

5. CNN with SMO-SCME for Caption Generation.

The derived low-level and high-level features are together sent to CNN for caption generation. In CNN [26], every neuron gets connected to nearby neurons located in the preceding layer. At position in the l-th layer of the linked w-th feature map, the features are computed as:

$$B^l_{r,t,w} = W_w^{l^T} PI^l_{r,t} + D^l_w, \tag{12}$$

where W^l_w is the weight, D^l_w stands for the bias of the w-th filter linked to the l-th layer. At the center location (r, t) of the l-th layer, the patch input is indicated as $PI^l_{r,t}$. The activation value $\text{act}^l_{r,t,w}$ related with convolutional features $B^l_{r,t,w}$ is:

$$\text{act}^l_{r,t,w} = \text{act}(B^l_{r,t,w}). \tag{13}$$

In the pooling layer the value of $C^l_{r,t,w}$ is:

$$C^l_{r,t,w} = \text{pool}(\text{act}^l_{m,h,w}), \quad \forall(m_h) \in NN_{r,t}, \tag{14}$$

where $NN_{r,t}$ is the neighbor's nearer position (r, t).

The prediction result factor occurs at output layer of CNN. The CNN loss is:

$$\text{Loss} = \frac{1}{wn}\sum_{h=1}^{wn} l\left[\theta; C^{(h)}, F^{(h)}\right]. \tag{15}$$

The general constraint related with W^l_w and D^l_w is denoted as θ, the counts of output-input relation is $PI^{(h)}, C^{(h)}; h \in [1, \ldots, wn]$. The h-th input, the label and output are denoted as $PI^{(h)}, C^{(h)}$ and $F^{(h)}$, respectively.

For solution encoding, CNN weights W are optimally elected via the SMO-SCME scheme. Representation of the solution is shown in Fig. 2, where symbolizes the entire count of CNN weights. The objective Obj of this research is:

$$\text{Obj} = \text{Min}(Er), \tag{16}$$

where Er implies an error.

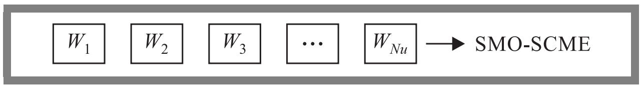


Fig. 2. Solution encoding scheme.

5.1. Proposed SMO-SCME Algorithm

The existing SMO model [31] offers a low level of accuracy. Hence, to overcome this disadvantage, self-improvements are proposed [32]–[39]. The proposed SMO-SCME model includes four steps:

1) The monkey distance to the food calculation as kind of fitness.
2) The positions are updated and fitness is re-computed for each individual.
3) The local leader updates the position if the group is distributed as per perturbation rate.
4) When the group exceeds the bound of a maximum group, a parent group is formed.

The initialization of all i-th spider monkeys SP_i takes place by:

$$SP_{ij} = SP_{\min j} + D_u(0,1) \times (SP_{\max j} - SP_{\min j}), \tag{17}$$

where $SP_{\max j}$ and $SP_{\min j}$ are the upper and lower limits of SP_i in the j-th dimension, and $D_u(0, 1)$ is a random number.

In the LLP phase, the i-th SP of the k-th subgroup is specified using:

$$\begin{aligned} SP_{\text{new}\,ij} = {} & SP_{\min j} + D_U(0,1) \times (ll_{kj} - SP_{ij}) + \\ & D_u(-1,1) \times (SP_{rj} - SP_{ij}), \end{aligned} \tag{18}$$

where SP_{ij} is the i-th SP in the j-th dimension, ll_{kj} is the j-th dimension of the k-th local group leader location, gl is the group leader and SP_{rj} is the r-th SP chosen from a random k-th group in which $r \neq i$. In the proposed SMO-SCME, the update occurs as:

$$\begin{aligned} SP_{\text{new}\, ij} = {} & SP_{\min j} + D_u(0,1) \times (ll_{kj} - SP_{ij}) + \\ & D_u(-1,1) \times (SP_{rj} - SP_{ij}) \times \\ & \left(w_{\max} - \frac{t}{t_{\max}} (w_{\max} - w_{\min}) \right), \end{aligned} \tag{19}$$

where implies the weight factor from 0 to 9.

During the GLP phase, the spider monkey position gets updated as:

$$\begin{aligned} SP_{\text{new}\, ij} = {} & SP_{\min j} + D_u(0,1) \times (gl_j - SP_{ij}) + \\ & D_u(-1,1) \times (SP_{rj} - SP_{ij}), \end{aligned} \tag{20}$$

where gl_j is GL position and $j \in 1, 2, \ldots, I$, V is a random index. In the proposed SMO-SCME, the update occurs as:

$$\begin{aligned} SP_{\text{new}\, ij} = {} & SP_{\min j} + D_u(0,1) - \text{prob}_i \times (gl_j - SP_{ij}) + \\ & D_u(-1,1) - \text{prob}_i \times (SP_{rj} - SP_{ij}), \end{aligned} \tag{21}$$

where, prob_i is a random probability factor created using the logistic map, $D_u(0,1)$ is computed based on a chaotic sine map.

The pro_i factor is:

$$\text{pro}_i = y \times \frac{\text{fit}_i}{\max - \text{fit}} + x, \tag{22}$$

where fit_i is fitness of the i-th monkey. The best outcomes are gained when $y = 0.9$ and $x = 0.1$.

The GL position is updated by deploying a greedy selection procedure and SP with the best fitness value is elected to determine the new position of GL. If GL position is similar to the old one, the global limit count is increased by 1.

Similarly, the ll position of entire groups is updated using a greedy selection procedure, and then SP with the best fitness is selected. If the ll position is similar, then the local limit count is increased by 1. If the position of ll is not updated, i.e. the local leader limit is met, then the first step is repeated using:

$$\begin{aligned} SP_{\text{new}\, ij} = {} & SP_{ij} + D_u(0,1) \times (gl_j - SP_{ij}) + D_u(0,1) \times \\ & (SP_{ij} - ll_{kj}). \end{aligned} \tag{23}$$

The decision is taken based on the gl position. The population is separated into subgroups if the gl position is not updated. The groups are separated until the group count reaches the highest allowed number of groups mg and, next, they are fused to form a group.

6. Simulation Results

The proposed CNN + SMO-SCME method for generating image captions was implemented in Python. The analysis was performed using the "Flickr Dataset" [40] and the performance of the CNN + SMO-SCME approach was evaluated in comparison with the faster R-CNN [40], GLSTM [34], LSTM, DBN, BI-GRU, CNN + CMBO, CNN + SSA, CNN + WHO, and CNN + SSO, with a wide variety of metrics benchmarked. Accordingly, the analysis was carried out by varying the LPs and different scores to represent the effectiveness of CNN + SMO-SCME. The results are shown in Fig. 3.

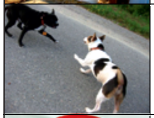
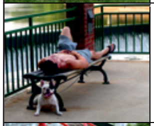
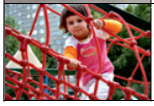
	Pink child in a dress is climbing up of setting a way stairs	A child in a pink dress is climbing up a set of stairs on a way
	The spotted dog and a black dog are a fighting	A black dog and a spotted dog are fighting
	A hands paint covered in sits in front of a rainbow painted with her hands in a bowl	A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl
	The man lays on a sits bench while his dog is by him	A man lays on a bench while his dog sits by him
	Playing net with a child on a rope	A child was playing on a rope net

Fig. 3. Sample image representation, DBN, and CNN + SMO-SCME results.

6.1. Analysis of Varied Scores

Meteor, Cider, and Rouge scores achieved by the CNN + SMO-SCME scheme are revealed in Fig. 4. Rouge provides a set of metrics used to assess the performance of machine translation and automatic summarization software in natural language processing. The metrics contrast is an automatically generated summary, a translation with a summary reference, translation or a collection of references. The Meteor measure has demonstrated stronger connection with human participants in terms of assessing the overall quality of the description. The remaining metrics rely on caption rankings and are unable to assess novel content.

It has been found that the correlation between these measurements and human judgement is rather poor. Our metric compares a machine-created sentence with a group of sentences that were written by humans for reference. The concepts of grammaticality, saliency, relevance, and accuracy (precision and recall) are essentially captured by our metric using sentence similarity.

The presented model was compared with Faster R-CNN [40], gLSTM [34], LSTM, DBN, BI-GRU, CNN + CMBO, CNN + SSA, CNN + WHO, and CNN + SSO schemes. The Meteor score attained by the CNN + SMO-SCME scheme, in comparison with Faster R-CNN [40], gLSTM [34], LSTM, DBN, BI-GRU, CNN + CMBO, CNN + SSA, CNN + WHO, LSTM [41] and CNN + SSO, regarding diverse LPs, is shown in Fig. 4a. The proposed model achieves a higher Meteor score than the schemes it is compared with. The CNN + SMO-SCME model offers enhanced values at the 90th LP, compared with the Meteor score outputs at other LPs. Similarly, the Cider score of CNN + SMO-SCME method reached a higher value of 1.609 at the 90-th LP, whereas the conventional models achieved lower values.

Figure 4c shows that the CNN + SMO-SCME technique achieved a better Rouge score (0.70) than other models at the 90th LP.

In the next step, a CNN that has been optimized to produce captions based on an image is used to create the captions. The weights of CNN are tuned using SMO with SCME to increase caption generating accuracy.

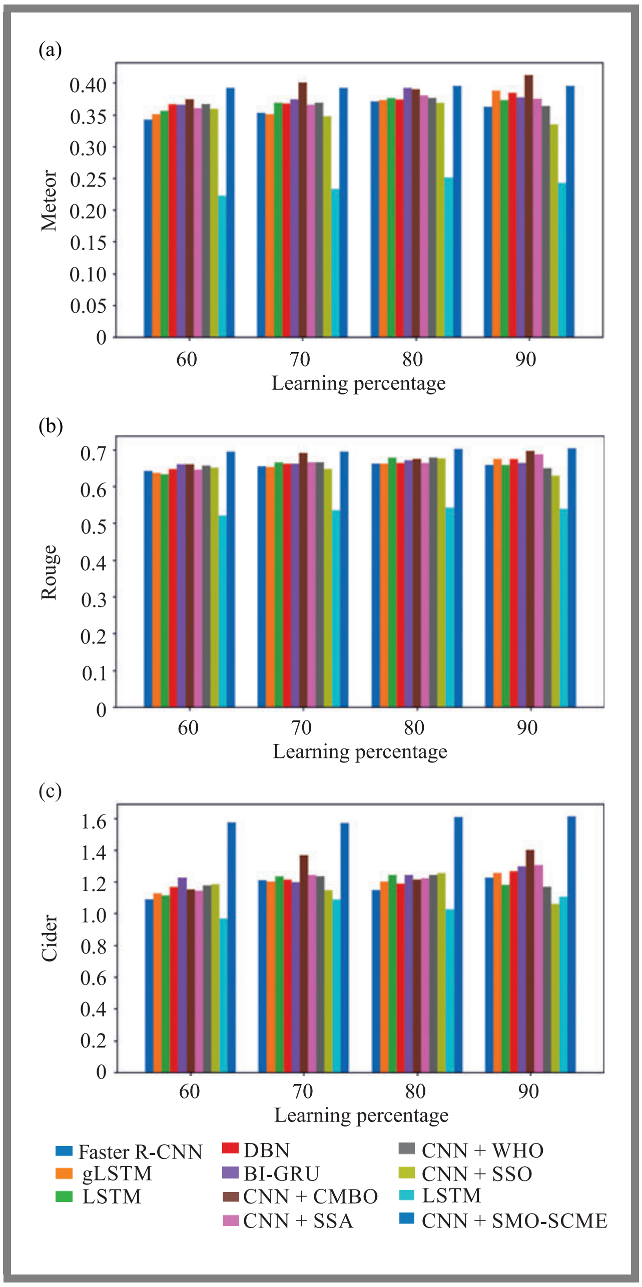


Fig. 4. Comparison of the proposed approach with other schemes for: (a) Meteor, (b) Cider, and (c) Rouge scores.

6.2. Analysis Using Bleu Score

Bleu is a bilingual evaluation understudy algorithm which assesses accuracy of text in such a way that quality is defined as the similarity between machine-generated output and a professional human translation. Today, Bleu still remains one of the most widely used automated metrics. The Bleu score of the proposed CNN + SMO-SCME scheme is evaluated in comparison with Faster R-CNN [40], gLSTM [34], LSTM [41], DBN, BI-GRU, CNN + CMBO, CNN + SSA, CNN + WHO solutions, as well as with the CNN + SSO scheme for various LPs. The evaluation was performed with datasets from [40], and the relevant results are summarized in Table 3. As shown in the table, the CNN + SMO-SCME scheme provides better outputs at the 90th LP than other LPS. In the Blue 1 case, the DBN model achieved the worst results. Thus, the advantage of the developed model is proven.

6.3. Convergence Analysis

The cost function of the SMO-SCME scheme, compared with CMBO, SSO, SSA, WHO, and SMO, is illustrated in Fig. 5. The SMO-SCME has generated lower costs in the 5–25 iteration range and higher cost values at lower iteration values. However, the cost factor still remains better than that of the compared schemes.

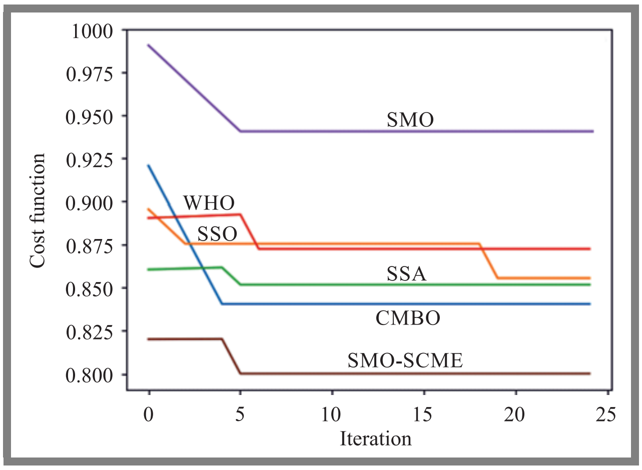


Fig. 5. Convergence analysis: SMO-SCME scheme vs. other models.

6.4. Time Analysis

Table 4 presents the comparison of computational time typical of the SMO-SCME method and other methods (CMBO, SSO, SSA, WHO, and SMO) and allows to evaluate performance of the system. The SMO-SCME system achieved lower values than CMBO, SSA, WHO, and SMO methods. The SSO is characterized by the lowest value among all techniques assessed. However, since the remaining scores attained by the SMO-SCME are better than those of SSO, this difference can be disregarded. Next to SSO, the WHO model has gained the time over CMBO, SSA, and WHO. Thus, the computational time improvement of the SMO-SCME scheme is proven.

Tab. 3. Comparison of CNN + SMO-SCME model with other schemes using the Bleu score.

LP		Faster R – CNN [40]	gLSTM [34]	LSTM [41]	DBN	BI-GRU	CNN + CMBO	CNN + SSA	CNN + WHO	CNN + SSO	CNN + SMO-SCME
60	Bleu-1	0.78	0.76	0.76	0.77	0.79	0.77	0.75	0.79	0.78	0.80
60	Bleu-2	0.61	0.60	0.59	0.61	0.63	0.59	0.58	0.59	0.58	0.63
60	Bleu-3	0.47	0.46	0.45	0.47	0.49	0.46	0.44	0.48	0.48	0.49
60	Bleu-4	0.24	0.27	0.27	0.29	0.30	0.27	0.24	0.30	0.30	0.36
70	Bleu-1	0.81	0.80	0.79	0.79	0.79	0.81	0.79	0.80	0.79	0.80
70	Bleu-2	0.64	0.64	0.63	0.61	0.62	0.65	0.62	0.63	0.61	0.63
70	Bleu-3	0.49	0.50	0.49	0.48	0.48	0.52	0.48	0.49	0.46	0.49
70	Bleu-4	0.27	0.30	0.27	0.24	0.28	0.33	0.26	0.26	0.22	0.36
80	Bleu-1	0.79	0.77	0.79	0.79	0.80	0.78	0.78	0.78	0.78	0.80
80	Bleu-2	0.62	0.61	0.62	0.62	0.64	0.63	0.62	0.62	0.62	0.64
80	Bleu-3	0.47	0.47	0.49	0.49	0.49	0.48	0.48	0.49	0.49	0.50
80	Bleu-4	0.28	0.32	0.34	0.32	0.30	0.24	0.28	0.34	0.38	0.36
90	Bleu-1	0.79	0.80	0.78	0.77	0.79	0.81	0.79	0.78	0.76	0.80
90	Bleu-2	0.62	0.64	0.62	0.61	0.63	0.65	0.63	0.61	0.59	0.64
90	Bleu-3	0.48	0.50	0.49	0.48	0.50	0.51	0.49	0.48	0.46	0.50
90	Bleu-4	0.31	0.35	0.34	0.33	0.32	0.36	0.28	0.33	0.29	0.37

Tab. 4. Time analysis (in seconds) using SMO-SCME vs. conventional models.

LP	CMBO	SSO	SSA	WHO	SMO	SMO-SCME
60	186	63	435	178	116	99
70	185	93	415	180	116	99
80	195	93	415	180	116	100
90	195	93	415	180	116	100

6.5. Feature Analysis

Table 5 presents the result of analysis of the developed CNN + SMO-SCME scheme and its comparison with the SMO-SCME model without lower-level features, SMO-SCME without higher-level features, and the proposed model without optimization. The study covers such metrics as Bleu, Meteor, Cider, and Rouge scores. The CNN + SMO-SCME achieved better values than SMO-SCME without lower-level features, SMO-SCME without higher-level features, and the proposed model without optimization. Moreover, the SMO-SCME model without lower-level features was characterized by relatively lower Bleu, Meteor, Cider, and Rouge scores than those of the SMO-SCME scheme without higher-level features and those of the proposed model without optimization.

Table 6 summarizes the performance of the former EC + SI-EFO scheme and the proposed CNN + SMO-SCME model. From the results, one may notice that the CNN + SMO-SCME model achieved better results than EC + SI-EFO. This is due to enhancements in the proposed methodology.

Tab. 5. Feature analysis and summary.

Metrics	CNN + SMO-SCME	SMO-SCME without lower features	SMO-SCME without higher features	Proposed without optimization
Bleu-1	0.80	0.62	0.76	0.77
Bleu-2	0.63	0.55	0.50	0.60
Bleu-3	0.49	0.36	0.36	0.46
Bleu-4	0.36	0.14	0.17	0.27
Cider	1.57	1.09	1.11	1.15
Rouge	0.69	0.64	0.54	0.66
Meteor	0.39	0.34	0.25	0.37

Tab. 6. Analysis of the EC + SI-EFO model and the proposed CNN + SMO-SCME scheme.

Metrics	CNN + SMO-SCME	EC + SI-EFO
Bleu-1	0.80	0.77
Bleu-2	0.63	0.58
Bleu-3	0.49	0.44
Bleu-4	0.36	0.26
Cider	1.57	1.43
Rouge	0.69	0.63
Meteor	0.39	0.37

7. Conclusion

The proposed image caption-generating model relying on low-level and high-level features provides improved caption generation accuracy with the CNN weights tuned via SMO-SCME. The better performance of the offered scheme was proven in comparison with other methods. Specifically, the CNN + SMO-SCME scheme achieved better estimation results for all Bleu scores at the 90th LP. In the case of Bleu 1, the DBN model was characterized by outputs that were worse than those of other schemes, e.g. Faster R-CNN, gLSTM, LSTM, BI-GRU, CNN + CMBO, CNN + SSA, CNN + WHO, and CNN + SSO at the 90th LP. The SMO-SCME model achieves the minimum calculation cost value of 99.24 at the the 70th LP, which is negligible compared to CMBO, SSA, WHO, and SMO approaches.

References

[1] 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, 2020 (DOI: 10.1016/j.image.2020.115836).

[2] J. Su, J. Tang, Z. Lu, X. Han, and H. Zhang, "A neural image captioning model with caption-to-images semantic constructor", *Neurocomputing*, vol. 367, 2019, pp. 144–151 (DOI: 10.1016/j.neucom.2019.08.012).

[3] S. Bang and H. Kim, "Context-based information generation for managing UAV-acquired data using image captioning", *Automation in Construction*, vol. 112, 2020 (DOI: 10.1016/j.autcon.2020.103116).

[4] H. Wang, H. Wang, and K. Xu, "Evolutionary recurrent neural network for image captioning", *Neurocomputing*, vol. 401, pp. 249–256, 2020 (DOI: 10.1016/j.neucom.2020.03.087).

[5] R. Li, H. Liang, Y. Shi, F. Feng, and X. Wang, "Dual-CNN: A convolutional language decoder for paragraph image captioning", *Neurocomputing*, vol. 396, pp. 92–101, 2020 (DOI: 10.1016/j.neucom.2020.02.041).

[6] J. Guan and E. Wang, "Repeated review based image captioning for image evidence review", *Signal Processing: Image Communication*, vol. 63, pp. 141–148, 2018 (DOI: 10.1016/j.image.2018.02.005).

[7] A. Singh, T.D. Singh, and S. Bandyopadhyay, "An encoder-decoder based framework for hindi image caption generation", *Multimed. Tools Appl 80*, pp. 35721–35740, 2021 (DOI: 10.1007/s11042-021-11106-5).

[8] Ph. Kinghorn, L. Zhang, and L. Shao, "A region-based image caption generator with refined descriptions", *Neurocomputing*, vol. 272, pp. 416–424, 2018 (DOI: 10.1016/j.neucom.2017.07.014).

[9] Q. Liu, Y. Chen, J. Wang, and S. Zhang, "Multi-view pedestrian captioning with an attention topic CNN model", *Computers in Industry*, vol. 97, pp. 47–53, 2018 (DOI: 10.1016/j.compind.2018.01.015).

[10] G. Christie, A. Laddha, A. Agrawal, S. Antol, and D. Batra, "Resolving vision and language ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes", *Computer Vision and Image Understanding*, vol. 163, pp. 101–112, 2017 (DOI: 10.1016/j.cviu.2017.09.001).

[11] F. Xiao, X. Gong, Y. Zhang, Y. Shen, and X. Gao, "DAA: Dual LSTMs with adaptive attention for image captioning", *Neurocomputing*, vol. 364, pp. 322–329, 2019 (DOI: 10.1016/j.neucom.2019.06.085).

[12] G. Huang and H. Hu, "c-RNN: A Fine-Grained Language Model for Image Captioning", *Neural Process Lett*, 2018 (DOI: 10.1007/s11063-018-9836-2).

[13] C. Wu, Y. Wei, X. Chu, F. Su, and L. Wang, "Modeling visual and word-conditional semantic attention for image captioning", *Signal Processing: Image Communication*, vol. 67, pp. 100–107, 2018 (DOI: 10.1016/j.image.2018.06.002).

[14] J. Yang, Y. Sun, J. Liang, B. Ren, and S. Lai, "Image captioning by incorporating affective concepts learned from both visual and textual components", *Neurocomputing*, 2018 (DOI: 10.1016/j.neucom.2018.03.078).

[15] T. Yinghua and C.S. Chee, "Phrase-based Image Caption Generator with Hierarchical LSTM Network", *Neurocomputing*, 2018 (DOI: 10.1016/j.neucom.2018.12.026).

[16] A. Yuan, X. Li, and X. Lu, "3G structure for image caption generation", *Neurocomputing*, 2018 (DOI: 10.1016/j.neucom.2018.10.059).

[17] Ch. Fan, Z. Zhang, and D.J. Crandall, "Deepdiary: Lifelogging image captioning and summarization", *Journal of Visual Communication and Image Representation*, vol. 55, pp. 40–55, 2018 (DOI: 10.1016/j.jvcir.2018.05.008).

[18] X. Chen, M. Zhang, Z. Wang, L. Zuo, and Y. Yang, "Leveraging Unpaired Out-of-Domain Data for Image Captioning", *Pattern Recognition Letters*, In press, accepted manuscript, 2018 (DOI: 10.1016/j.patrec.2018.12.018).

[19] Z. Ye, *et al.*, "A novel automatic image caption generation using bidirectional long-short term memory framework", *Multimed Tools Appl 80*, pp. 25557–25582, 2021 (DOI: 10.1007/s11042-021-10632-6).

[20] H. Zhang *et al.*, "Novel model to integrate word embeddings and syntactic trees for automatic caption generation from images", *Soft Comput 24*, pp. 1377–1397, 2020 (DOI: 10.1007/s00500-019-03973-w).

[21] C. Sur, "AACR: Feature Fusion Effects of Algebraic Amalgamation Composed Representation on (De)Compositional Network for Caption Generation for Images", *SN Comput. Sci. 1, 229*, 2020 (DOI: 10.1007/s42979-020-00238-4).

[22] C. Shan, A. Gaoyun, Z. Zhenxing, and R. Qiuqi, "Interactions guided generative adversarial network for unsupervised image captioning", *Neurocomputing*, vol. 417, pp. 419–431, 2020 (DOI: 10.1016/j.neucom.2020.08.019).

[23] Y. Wei, L. Wang, and C. Wu, "Multi-Attention Generative Adversarial Network for image captioning", *Neurocomputing*, vol. 387, pp. 91–99, 2019 (DOI: 10.1016/j.neucom.2019.12.073).

[24] M. Yang *et al.*, "An Ensemble of Generation- and Retrieval-Based Image Captioning With Dual Generator Generative Adversarial Network", *IEEE Transactions on Image Processing*, vol. 29, pp. 9627–9640, 2020 (DOI: 10.1109/TIP.2020.3028651).

[25] D. Zhao, Z. Chang, and S. Guo, "A multimodal fusion approach for image captioning", *Neurocomputing*, vol. 329, pp. 476–485, 2019 (DOI: 10.1016/j.neucom.2018.11.004).

[26] S. Ding, S. Qu, and S. Wan, "Image caption generation with high-level image features", *Pattern Recognition Letters*, vol. 123, pp. 89–95, 2019 (DOI: 10.1016/j.patrec.2019.03.021).

[27] S.R. Kodituwakku, "Comparison of Color Features for Image Retrieval", *Indian Journal of Computer Science and Engineering*, vol. 1, no. 3, pp. 207–211 (`http://www.ijcse.com/docs/IJCSE10-01-03-06.pdf`).

[28] –, `https://photography.tutsplus.com/tutorials/what-is-image-sharpening--cms-26627`.

[29] T. Bouwmans, C. Silva, C. Marghes, M.S. Zitouni, H. Bhaskar, and C. Frelicot, "On the role and the importance of features for background modeling and foreground detection", *Computer Science Review*, vol. 28, pp. 26–91, 2018 (ISSN 15740137, DOI: 10.1016/j.cosrev.2018.01.004).

[30] –, `https://en.wikipedia.org/wiki/Motion_analysis`.

[31] S. Harish, G. Hazrati, and J.C. Bansal, "Spider Monkey Optimization Algorithm", 2019 (DOI: 10.1007/978-3-319-91341-4_4).

[32] B.R. Rajakumar, "Impact of Static and Adaptive Mutation Techniques on Genetic Algorithm", *International Journal of Hybrid Intelligent Systems*, vol. 10, no. 1, pp. 11–22, 2013 (DOI: 10.3233/HIS-120161).

[33] B.R. Rajakumar, "Static and Adaptive Mutation Techniques for Genetic algorithm: A Systematic Comparative Analysis", *International Journal of Computational Science and Engineering*, vol. 8, no. 2, pp. 180–193, 2013 (DOI: 10.1504/IJCSE.2013.053087).

[34] S.M. Swamy, B.R. Rajakumar and I.R. Valarmathi, "Design of Hybrid Wind and Photovoltaic Power System using Opposition-based Genetic Algorithm with Cauchy Mutation", *IET Chennai Fourth International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2013)*, 2013 (DOI: 10.1049/ic.2013.0361).

[35] A. George and B.R. Rajakumar, "APOGA: An Adaptive Population Pool Size based Genetic Algorithm", *AASRI Procedia – 2013 AASRI Conference on Intelligent Systems and Control (ISC 2013)*, vol. 4, pp. 288–296, 2013 (DOI: 10.1016/j.aasri.2013.10.043).

[36] B.R. Rajakumar and A. George, "A New Adaptive Mutation Tech-

nique for Genetic Algorithm", *In proceedings of IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1–7, 2012 (DOI: 10.1109/ICCIC.2012.6510293).

[37] M.B. Wagh and N. Gomathi, "Improved GWO-CS Algorithm-Based Optimal Routing Strategy in VANET", *Journal of Networking and Communication Systems*, vol. 2, no. 1, pp. 34–42, 2019 (DOI: 10.46253/jnacs.v2i1.a4).

[38] S. Halbhavi, S.F. Kodad, S.K. Ambekar, and D. Manjunath, "Enhanced Invasive Weed Optimization Algorithm with Chaos Theory for Weightage based Combined Economic Emission Dispatch", *Journal of Computational Mechanics*, Power System and Control, vol. 2, no. 3, pp. 19–27, 2019 (DOI: 10.46253/jcmps.v2i3.a3).

[39] A.N. Jadhav and N. Gomathi, "DIGWO: Hybridization of Dragonfly Algorithm with Improved Grey Wolf Optimization Algorithm for Data Clustering", *Multimedia Research*, vol. 2, no. 3, pp. 1–11, 2019 (DOI: 10.46253/j.mr.v2i3.a1).

[40] –, https://www.kaggle.com/ming666/flicker8k-dataset.

[41] D. Songtao, *et al.*, "Image caption generation with high-level image features", *Pattern Recognition Letters 123*, pp. 89–95, 2019 (DOI: 10.1016/j.patrec.2019.03.021).

Mukesh Kalla is serving as Head & Assistant Professor in the Department of Computer Science and Engineering from the Sir Padampat Singhania University, Udaipur, India. He has received his Doctoral Degree in Computer Engineering in year 2017. He is having teaching and administrative experience of 19 years. He has about 24 research publications in international journals and conferences.

E-mail: mukesh.kalla@spsu.ac.in

Department of Computer Science and Engineering, Sir Padampat Singhania University, India

Amit Jain is currently working as Assistant Professor in Computer Science and Engineering Department, Sir Padampat Singhania University, Udaipur, India. He has completed his Doctoral Degree in Computer Engineering in 2016. He is having 25 years teaching and administrative experience and about 33 research publications in international journals and conferences.

E-mail: amit.jain@spsu.ac.in

Department of Computer Science and Engineering, Sir Padampat Singhania University, India

Arvind Sharma is presently working as Assistant Professor in Mathematics Department, Sir Padampat Singhania University, Udaipur, India. He has completed his Doctoral Degree in Mathematics in year 2017. He is having teaching and administrative experience of 17 years. He has about 10 research publications in international journals and conferences.

E-mail: sharma.arvind@spsu.ac.in

Department of Computer Science and Engineering, Sir Padampat Singhania University, India

Roshni Padate is Ph.D. scholar in Sir Padampat Singhania University, Udaipur, India and an Assistant Professor in the Department of Computer Engineering, Fr. Conceicao Rodrigues college of Engineering, Mumbai University. She has completed M.E. (Computer Engineering) from Mumbai University, India in 2010 and B.E. (Computer Science and Engineering) from Shri Sant Gajanan Maharaj College of Engineering Shegaon, Amravati University, India in 2000. Her area of research is image processing, artificial intelligence, machine learning, deep learning, and data mining.

E-mail: Roshni.padate@spsu.ac.in

Department of Computer Science and Engineering, Sir Padampat Singhania University, India