Multimedia Tools and Applications (2023) 82:3801–3830
https://doi.org/10.1007/s11042-022-13443-5

An accurate generation of image captions for blind people using extended convolutional atom neural network

Tejal Tiwary[1] · Rajendra Prasad Mahapatra[2]

Received: 5 June 2021 / Revised: 15 February 2022 / Accepted: 2 July 2022 /
Published online: 15 July 2022

Abstract
Recently, the progress on image understanding and AIC (Automatic Image Captioning) has attracted lots of researchers to make use of AI (Artificial Intelligence) models to assist the blind people. AIC integrates the principle of both computer vision and NLP (Natural Language Processing) to generate automatic language descriptions in relation to the image observed. This work presents a new assistive technology based on deep learning which helps the blind people to distinguish the food items in online grocery shopping. The proposed AIC model involves the following steps such as Data Collection, Non-captioned image selection, Extraction of appearance, texture features and Generation of automatic image captions. Initially, the data is collected from two public sources and the selection of non-captioned images are done using the ARO (Adaptive Rain Optimization). Next, the appearance feature is extracted using SDM (Spatial Derivative and Multi-scale) approach and WPLBP (Weighted Patch Local Binary Pattern) is used in the extraction of texture features. Finally, the captions are automatically generated using ECANN (Extended Convolutional Atom Neural Network). ECANN model combines the CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory) architectures to perform the caption reusable system to select the most accurate caption. The loss in the ECANN architecture is minimized using AAS (Adaptive Atom Search) Optimization algorithm. The implementation tool used is PYTHON and the dataset used for the analysis are Grocery datasets (Freiburg Groceries and Grocery Store Dataset). The proposed ECANN model acquired accuracy (99.46%) on Grocery Store Dataset and (99.32%) accuracy on Freiburg Groceries dataset. Thus, the performance of the proposed ECANN model is compared with other existing models to verify the supremacy of the proposed work over the other existing works.

✉ Tejal Tiwary
ttiwary92@gmail.com

[1] Department of Computer Science and Engineering, SRMIST, NCR Campus, Ghaziabad, India

[2] Department of CSE, SRM Institute of Science & Technology, Delhi, NCR Campus, Ghaziabad, India

Keywords Image captioning · Blind people · Alternative text · Natural language processing · Deep learning · Extended convolutional atom neural network (ECANN)

1 Introduction

Image captioning is an interesting research area due to its various applications like supporting blind people, helpful for image indexing and other NLP applications [35]. Web accessing through image captions is a major part in the everyday life of blind people. At the same time, identifying the images on the web is a very challenging task for the blind people [4]. The idea of retrieving web data and accomplishing the everyday jobs like banking, grocery shopping is difficult for the blind person [15]. Web is an important source for the blind people and giving most important autonomy to the blind people [7]. Hence, the web accessibility practice is utilized to describe the images with the alternative text (alt text). This alt text gives small captions substitutes to the image that expresses general meaning of the image [31].

The caption of the images permits the blind people to contribute on the social activities and getting more information from online and helps in purchasing the products. Generation of the captions automatically enable the blind people to get more information about the images [18]. Image captioning is the process of automatically creating a caption for the image. In artificial intelligence, creation of caption in images providing more attention and is becoming more significant [3]. Recently researchers are focused on the enhancements in the accessibility of web with different techniques. These methods are categorized into 3 classes are, crowdsourcing, machine based and hybridized technologies.

In the first crowdsourcing process, the captions are produced in the images with the human annotators [1, 20]. In the machine-based techniques, the image is identified with the learning methodologies and provides the corresponding captions [9, 23]. The hybridized techniques provide the image captions automatically. It lessens the time consumption as well as the expense of the process [8]. Numerous methodologies are presented in the previous image captioning works for blind people. They are, ZFNet, VGGNet, GoogLeNet, AlexNet, Nearest Neighbor (NN) [19], LSTM [23]. Deep learning based techniques are broadly utilized in the image captioning and attained better results in the image captioning [2, 13, 22]. Deep learning (DL) is the advanced subtype of ML (Machine Learning) which is very useful in interpreting huge amounts of data and make the process easier and faster. Recent studies have shown that, identifying diseases using deep learning architectures can reduce the workload of radiologists compared to other diagnostic methods. In the field of medical processing, deep learning can act as an automatic screening and detection of COVID-19 disease [24] and is very convenient and fast compared to ML techniques. Deep learning is the most popular and convenient method for providing fast detection in various applications such as image classification, object detection, entertainment, healthcare [21, 25–27], fake news detection, crop yield prediction, robotics etc.

Number of researchers have been hardly working to enhance the accessibility of online images via a huge range of techniques. Deep learning based captioning models produce captions generated by machine by using the trained CV (computer vision) models that identify the relationships and objects in an image and produce suitable captions. In this work, a new DL framework named ECANN is presented to generate multiple image captions and make use of reverse search strategy to select the most appropriate caption for the image input. The proposed ECANN model progresses the image captions accessibility by means of the fully-automated principle and explores the feasibility of images that are being shared online.

1.1 Motivation and problem statement

Recently, computer vision-based assistive technologies have been designed for assisting the blind people. In particular, web accessing through image captions is a major part in the everyday life of blind people. Generation of image captioning techniques are supports them for easy identification of images with captions in various applications such as grocery shopping, education and so on. Thus, it gained significant attention and deliberated as one of the most famous topics in the field of computer vision. But accurate generation of captions in image is difficult task in this field. The process of caption generation faces the major problems such as: to create the entire NL (Natural language) sentences/captions similar to humans, to make the caption generated and its semantics to be accurate, correct and being consistent with the input image. With the advancements of the image caption generation system, blind people can see the world like normal humans. The problems originate in accordance with the naturalness and the compositionality. The conventional system on image captioning suffers because of the lack of naturalness and NL compositional nature in which the captions are generated in a sequential way that creates the language structure to be semantically inappropriate. Another significant challenge is the impact of dataset bias, which makes the trained models to over fit on common/similar objects, which struggles to generalize the suitable captions. To overcome these challenges, a new DL based automatic caption generation strategy is introduced in this work, which bridges the semantic gap among the vision points and language to incorporate the need of accurate scene understanding. The presented approach automatically generates the alternative caption of images on the websites that are not captioned and make web access easier for blind people.

1.2 Contributions

The main contributions of ECANN based image caption generation are stated as:

- To develop a deep learning based neural network model named ECANN to generate the accurate image captions. The proposed ECANN model is designed based on understanding the difficulties faced by the blind people in shopping online and the error occurrence in the generation of image caption is minimized using AAS algorithm.
- The proposed ECANN model work on the pre-defined captions generated and choose the most appropriate caption with reversible search enabling quick browsing.
- The ECANN framework satisfy the necessities of blind people in alternative image caption generation with higher accuracy compared to other baseline classifiers.
- The selection of non-captioned images from the dataset is performed using ARO, which is the hybrid combination of FCM (Fuzzy C Means) clustering, optimized by the ARO algorithm. Next, the appearance and texture features are extracted using SDM and WPLBP.
- Finally, the ECANN allows the automatic generation of image captions and effectively helps the visually impaired people in online Grocery shopping.

The outline of this paper is structured as: Section 1 presents the introduction and highlights the motivation and contributions. Section 2 discuss about the recent related works on generation of image captions. Section 3 clearly describes about the proposed methodology. Section 4 provides the detailed description about Implementation and the results obtained. Section 5 concludes the presented work followed by references.

2 Related work

Some of the recent literature works related to image caption generation are discussed as follows:

Heng Song et al. [29] presented an innovative visual text merging (VTM) framework for providing the image captions. Initially, an attention model was developed for the visual data to get the accurate image captions. In avtmNet (adaptive VTM network) the merging of visual, text data was done correspondingly. The datasets used were COCO2014 and Flickr30K. The presented merging network merges the text and visual data accurately. The developed adaptive VTM methodology produces the resulting image captions based on the attained text and visual data. The performance obtained was evaluated for the dual datasets based on different scores such as BLEU, CIDEr, ROUGE-L, SPICE and METEOR. The evaluation results for Flickr30 were BLEU (0.248), ROUGE-L (0.494), METEOR (0.208), CIDEr (0.598), SPICE (0.157). While for COCO2014 the outcomes were BLEU (0.3317), ROUGE-L (0.567), METEOR (0.273) CIDEr (1.126) and SPICE (0.201) illustrates the efficacy of merging network. Deng et al. [5] developed an image caption framework called DenseNet+LSTM. In the encoding step, DenseNet extricates the features from the image. In the parallel, sentinel gate in the framework selects the feature data for the creation of caption. In the decoding step LSTM was utilized to enhance the image quality in the creation of text in the images. It was very useful for the blind people for better guidance. The datasets used for execution were Flickr30K and COCO2014. The performance obtained with Flickr30K was based on BLEU (0.667), METEOR (0.214) scores. Also, the COCO2014 obtained the scores with BLEU (0.739) and METEOR (0.270).

Yuchen Wei et al. [30] presented the review of various deep learning procedures and its challenges in identifying the marketing products for blind people. VI people deal with number of challenges in everyday life like shopping. The number of datasets used were GroZi-120, RPC, D2S, and Groci-3.2 K. The DL framework was utilized for finding the correct products through the captions which should be very accurate in helping VI people. The lack of accuracy in captions creates difficulty in retail store shopping. The different frameworks on DL provides image captions and recognizes the products effectively. Min Yang et al. [34] developed an innovative retrieval dependent caption creation framework called Ensemble caption (EnsCaption). It was the combined technique of both caption creation and caption retrieving. The model fuses the personalised texts for the query image. The re-ranking procedure recovers the accurate caption for the image from the dataset of created captions. The adversarial network finds the variations among the created and the retrieved captions for the image. The introduced technique provides more accurateness in the caption generation. The datasets used were Flickr-30 K and MSCOCO. The outcomes obtained with flickr-30 K dataset was BLEU (76.2) and CIDEr (69.3) and for MSCOCO was BLEU (81.7) and CIDEr (125.5).

Niange Yu et al. [36] introduced a CNN based multiple labelling classifier for the image caption. Here, the captions were provided for the images based on their topics by utilizing the presented classifier. In the developed framework, input was the image and their topic, and the output was the image caption. The hierarchical framework was conserved with an embedding technique. The retrieving the image captions were attained in the work was utilizing the bi-directional caption image retrieving procedure. The developed technique gives the better quality in the generation of image captions. Fen Xiao et al. [32] developed an image captioning framework with dual LSTM for enhancing accessibility of blind people. In that, two separate LSTM frameworks were integrated with the adaptive semantic

attention framework. The first LSTM was adapted after the attention layer which was utilized for attaining the visual sentinel information. Next LSTM was utilized for creating the text for the images as captions. Finally attains the accurate text sequence for the given input images. The dataset used was MSCOCO, Flickr30K. The performance obtained with Flickr30K was BLEU (68.6), METEOR (21.5) whereas for MSCOCO the score obtained was BLEU (75.8), METEOR (27.1).

Loganathan et al. [17] presented an automatic captions generation for the images with the combined CNN and LSTM frameworks. The framework provides the accurate captions for the images with the learning process. Here, the effective learning procedure results the better image caption. These combined machine learning procedures gives the accurate image captioning with reduced complexity. The automatic caption generation results the easy access to the blind people for identifying the images without any difficulties. The dataset used was Flicker8K dataset. Singh, A et al. [28] introduced the encoder-decoder based framework for image captioning. Here, the CNN model was used as the encoder for image visual features. Then, the captions were generated for images by the stacked LSTM, which was the integration of bi-directional LSTM and unidirectional LSTM. The VGG19 based CNN model was utilized for encoding the visual features. Hindi genome dataset was used for the validation of this framework. The evaluation metrics were RIBES (0.17) and BLEU (3.28). The presented framework did not capture the alphanumeric content of the images.

Iwamura et al. [10] presented a trainable end-to-end approach for generating the image caption with three datasets namely several copyright-free images, MSCOCO and MSR-VTT2016-image. In this framework, the four phases were performed such as feature extraction, motion estimation, object detection and caption generation. The motion-CNN model was developed here for the automatic motion feature extraction. CNN extracted the features from the input images and passed them on to the object detection phase for detecting the object. Then, the attention model was used in caption generation component for the computation of attention features. LSTM get the correlated attention features for caption generation. The performance obtained on dataset MSR-VTT2016-Image was BLEU (49.9), METEOR (16.1)] whereas for MSCOCO dataset was BLEU (75.9), METEOR (26.7). Khurram et al. [12] developed a Dense-CaptionNet deep learning model for the image captioning. This deep learning model was region based architecture to describe the image semantics. In this framework, three modules were available. The dataset used were MSCOCO, Visual Genome and IAPR TC-12. In the first module, the object relationship and the region description were generated. The object attributes available in the scene was generated in the second module. The textual descriptions attained from first two module were provided as input to third module for the caption generation. This approach provided the detailed description of the images. Table 1 illustrates the review on different existing models.

2.1 Research questions (RQs)

This section describes the set of research questions that clearly focus and pinpoints the major objectives of the proposed framework. The following are the proposed RQs that are clearly stated as:

> RQ 1: How the proposed deep learning based image caption generation model creates accurate image captions for VI people?

Table 1 Review on various existing methods

Author Name	Dataset	Technique	Pre-processing	Performance (%)
Heng Song et al. [29]	COCO2014 and Flickr30K	avtmNet	Image Resizing	Flickr30K: BLEU (0.248), ROUGE-L (0.494), METEOR (0.208), CIDEr (0.598), SPICE (0.157) COCO2014: BLUE (0.3317), ROUGE (0.567), METEOR (0.273), CIDEr (1.126), SPICE (0.201)
Zhenrong Deng et al. [5]	COCO 2014 and Flickr30K	DenseNet+LSTM	Image Resizing	Flickr30K: BLEU (0.667), METEOR (0.214). COCO2014: BLEU (0.739) and METEOR (0.270).
Yuchen Wei et al. [30]	RPC, D2S, Grozi-120, Groci-3.2 K	Deep learning models (Priming Network, RetinaNet, DNN, One-shot learning)	Noise elimination and removing redundant data	RPC: Priming Network-mAP (97.91%) D2S: RetinaNet -mAP (89.6%) Groci-120: DNN-Precision (45.20%), Recall (52.70%) Groci-3.2 K: One-shot Learning-Precision (92.19%), Recall (87.89%)
Fen Xiao et al. [32]	MS COCO and Flickr30K	Dual LSTM	–	Flickr30K: BLEU (68.6), METEOR (21.5) MS COCO: BLEU B-1(75.8), METEOR (27.1)
Singh, A et al. [28]	Hindi genome dataset	Encoder (CNN)and Decoder (LSTM) based model	IndicNLP Tokenizer	BLEU (3.57), RIBES (0.08)
Iwamura et al. [10]	MSR-VTT2016-Image, MSCOCO	CNN_LSTM	Image Resizing	MSR-VTT2016-Image: BLEU (49.9), METEOR (16.1) MSCOCO: BLEU (75.9), METEOR (26.7)

RQ 2: How the metaheuristic optimization is used with neural network for hyperparameter optimization?
RQ 3: What are the various performance metrics used for the evaluating the performance?
RQ 4: What are the possible grocery datasets used for online shopping and the major future directions?

3 Proposed methodology

Recently, Caption Generation (CG) in computer vision is expected to have much attention due to its extensive applications such as virtual assistants, image understanding, image retrieval or

indexing and helping blind people. Automatic caption generation (ACG) system is a challenging task which helps the blind people by providing better understandings about what is happening around them. Humans naturally have the ability to identify or recognize the images at a quick glance. But the lifestyle of blind people differs from normal people because they make use of other senses (touch, hearing) as assistance to know the objects placed nearby. This work aims to design a new image CG task with the deep learning model named ECANN (Extended Convolutional Atom Neural Network) which make use of computers to imitate the ability of humans to better understand the visual world. Figure 1 signifies the schematic model of the proposed method.

Automatic image captioning (AIC) offers simpler image captions to the images that are non-captioned. The proposed AIC with the EACNN model is processed with the combination of subsequent stages such as Data Collection, Non-captioned image selection, Extraction of appearance, texture features and Generation of automatic image captions. The initial stage is the selection of non-captioned images from the database by using ARO algorithm. After the selection of non-captioned images, appearance and texture features are extracted using spatial derivative & multi scale (SDM) feature and weighted patch based local binary pattern (WPLBP). Moreover, the extracted features are utilized for the accurate differentiation of the images. Finally, caption is accurately produced for corresponding images with the help of ECANN architecture. The ECANN based alternate image captioning process introduced a caption reusable system based on AI (Artificial Intelligence) with a reverse image search to reuse pre-existing captions for the target image. However, this proposed ECANN model is used to generate alternate captions to images which should be semantically genuine to the original image. In this framework, error occurrence in the image captioning is reduced using (AAS) algorithm.

3.1 Data collection

The process of collecting or gathering the information from any online public source which enables us to evaluate the results is named as data collection. It can be viewed as a

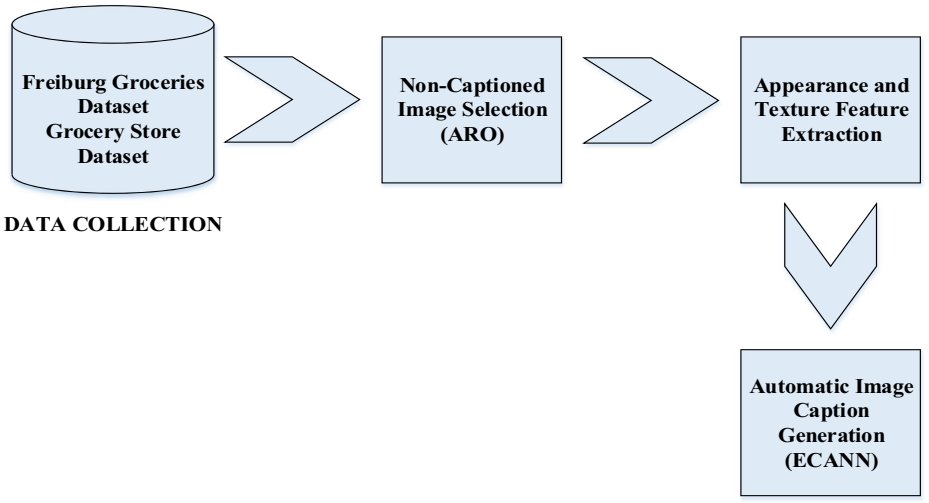

Fig. 1 Schematic model of proposed method

systematic framework to collect data from different sources to acquire an accurate and complete outcome on the particular area of interest. Precise data collection is very much essential to maintain the integrity of the research and guaranteeing quality assurance. In this work, the dual datasets namely Freiburg Groceries Dataset and Grocery Store Dataset are collected from dual online sources and used for the analysis and implementation purpose. These online groceries dataset includes multiple-classes which comprises of different types of objects and products.

3.2 Non-captioned image selection

The raw image input from the dataset consists of both captioned and non-captioned images. The main objective of this work is to generate automatic captions to the non-captioned images. Here, initially the adaptive rain optimization (ARO) algorithm is used to select only the non-captioned images. This ARO algorithm is the combination of Fuzzy C Means (FCM) and Rain Optimization (RO) algorithm. The selection of non-captioned images using ARO solves the problem of imbalanced dataset learning and can be extensively applied in various research fields namely engineering, image processing etc. The basic idea behind ARO is clustering which groups the images into dual groups such as captioned and non-captioned images. The ARO algorithm follows the unsupervised clustering strategy and allows each data to be associated with more number of clusters.

FCM algorithm [33] optimized with RO algorithm is named as ARO which assigns lesser weights to the samples which are equivalent to C clusters. Let the input data points can be given as $Z = \{z_1, z_2, \ldots, z_n\}$ which is divided into C clusters as $C = \{C_1, C_2, \ldots, C_k\}$ where n indicates the number of elements and k signifies the number of pre-defined clusters. The minimization of objective function can be expressed as,

$$J_{\min} = \sum_{i=1}^{C} \sum_{j=1}^{n} U_{ij}^{g} D_{ij}^{2} \tag{1}$$

where, the fuzzification element is g, the membership degree is U_{ij} and the Euclidean distance of j^{th} instance to i^{th} cluster centroid is specified as D_{ij}. Equation (1) is minimized with reference to the following conditions:

$$1 \leq j \leq n : \sum_{i=1}^{C} U_{ij} = 1 \tag{2}$$

$$1 \leq i \leq C : \sum_{j=1}^{n} U_{ij} > 0 \tag{3}$$

Here, the Lagrange function is used to minimize $J_{\min}$ by means of setting the derivative with reference to zero value for both U_{ij} and C_j parameters given as,

$$J_{\min} = \sum_{i=1}^{C} \sum_{j=1}^{n} U_{ij}^{g} D_{ij}^{2} + \sum_{j=1}^{n} \lambda_k \left(\sum_{i=1}^{n} U_{ij} - 1 \right) \tag{4}$$

where, the Lagrange function is represented as $\lambda = [\lambda_1, \lambda_2, \ldots\lambda_n]^T$ and the cluster centroid is represented as,

$$C_i = \frac{\sum_{j=1}^{n} U_{ij}^{g} z_j}{\sum_{j=1}^{n} U_{ij}^{w}}, i \in [1, C] \tag{5}$$

The degree of membership is specified as,

$$U_{ij} = \frac{D_{ij}^{\frac{2}{1-g}}}{\sum_{k=1}^{C} D_{kj}^{\frac{2}{1-g}}}, i \in [1, C], j \in [1, n] \tag{6}$$

The fuzzification element value is $g = 2$. FCM model is generalized using the fuzzy set and the initial outcomes are produced with proper competency and simplicity in fuzzy clustering. In FCM model, the Lagrange function λ_k is optimized with RO algorithm which overcomes the issues related to higher time computation, sensitive to noisy data and initialization. RO algorithm simulate the behaviour of rain drops. Each solution is considered a raindrop and the initial population is generated. Radius is the main function in RO and the population is initialized. The rain droplets R_1, R_2 connects with R radius is described as,

$$R = \left(R_1^m + R_2^m\right)^{1/m} \tag{7}$$

where, m indicates the variables of each droplet count. Thus, by increasing the number of total iterations the cluster group (captioned and non-captioned image group) is formed in minimal time with noise elimination.

3.3 Extraction of appearance and texture features

The process of capturing the image visual content for indexing and retrieval is termed as feature extraction. Texture and appearance features play a vital role in the feature extraction process which extracts relevant image information. Feature extraction begins with the initial set of data and constructs features that can be non-redundant, informative and facilitates successive learning which leads to better human understandings.

3.3.1 Appearance features

The appearance of an image is extracted using SDM feature model. Differential feature is defined as a vector which is linked with a point in an image and it can be utilized to extract the appearance based features of an image. The differential feature is evaluated from the image spatial derivatives. For a given image I, with w point the lower order derivatives can be utilized as a feature which can be expressed as,

$$H = \left\langle \frac{\partial I}{\partial x}(w), \frac{\partial I}{\partial y}(w), \frac{\partial^2 I}{\partial x^2}(w), \frac{\partial^2 I}{\partial xy}(w), \frac{\partial^2 I}{\partial y^2}(w) \right\rangle \tag{8}$$

From the image, the valuable statistical information can be captured using the derivatives. The first-order derivatives indicate the intensity edgeness or gradient whereas the second-order derivatives are used to illustrate bars. The derivative features are known as appearance features

which can be denoted as vector H. However, these features approximate the intensity surface local shape using the Taylor Series (TS) expansion. The TS specifies that the pixel derivatives are required to estimate the intensity value in a neighbourhood around it.

3.3.2 Texture features

Texture is usually a feature which can be used to divide the images into regions. This texture feature offers information by means of the spatial arrangement of intensities or colours in an image. Texture defines the surface characteristics in terms of shape, size, arrangement, density etc. The image textures can be rough or smooth, hard or soft, glossy or mat etc. and characterized using the spatial distribution of intensity levels in a neighbourhood. In this work, the extraction of texture features can be processed using the approach called weighted patch based local binary pattern (WPLBP).

The LBP approach extracts the local information by matching the pixel differences from each minor region in image. Then the local information extracted is encoded into a shorter string of bits and this string represents the DG (Directional Gradient) information by means of bit '0' or '1'. The texture feature extraction using LBP on every central z_{cn} pixel depends on the neighbouring z_p ($p = 0, 1, \ldots P - 1$) pixel with radius R can be expressed as:

$$LBP = \sum_{p=0}^{P-1} f\left(z_p - z_{cn}\right) 2^p \tag{9}$$

$$f(z) = \begin{cases} 1, z \geq 0; \\ 0, o.w \end{cases} \tag{10}$$

where, P signifies the number of neighbouring pixels (p = 8 or 16) and the threshold function is denoted as $f(z)$. The extended or improved form of LBP is WPLBP which make use of a pyramidal structure. The WPLBP uses any of the kernel $s_k^p \in S_k$ which can be summed with the patches of $z_k^p \in Z_k$ given as,

$$WPLBP = \sum_{k=1}^{K} f\left(\sum_{p=0}^{P-1} z_{k-1}^p s_k^p\right) 2^{(k-1)} \tag{11}$$

where, K signifies the number of levels in the adopted pyramid model. The above Eq. (11) can be expressed in convolution form as,

$$Z_{k+1} = S_k * Z_k \tag{12}$$

where, the 2D convolution operation is signified as $*$, the weighted kernel matrix is denoted as S_k and the original image is Z_0. The feature maps absorbed from the original image can be given as, $Z_1, Z_2, \ldots, Z_k$whereas Z_1, Z_k represents the lowest and highest levels of abstraction. In WPLBP, a proper training set is required for the purpose of weight matrix learning.

$$N\text{-}GD = \left(\sum_{t=1}^{T} \overline{Z}_t^{i,j}\right)^{1/2}, \forall i, j \in \left(1, 2, \ldots, \sqrt{M}\right) \tag{13}$$

The training sets are generated using the N-dimensional GD (Gradient Descriptor). In WPLBP, the gradients are calculated by using the Sobel operator on number of patches $\overline{Z_t}$, $(t = 1, 2, \ldots T)$ that can be sampled randomly from the scaled image. Using the WPLBP approach, the image is resized to gather the patches as training set and the texture feature is finally extracted.

3.4 Generation of automatic image captions

The procedure of generating automatic image captions is known as image captioning. A single image can hold large amount of information as everyday massive image data is generated. After the extraction of valuable information from image, the next step is the automatic generation of image captions. Here, with the help of ECANN the automatic image captions are generated for the non-captioned images. This ECANN concept works under automatic caption generation with reverse image search to reuse pre-existing captions for the target image. In this framework, error occurrence in the image captioning is reduced using AAS algorithm. ECANN is a deep learning model which signifies the combination of CNN and LSTM architecture. ECANN can be used to annotate the images automatically by reusing the pre-existing captions generated. Thus, the deep learning based caption generation process can significantly minimize the human errors and make use of AI models in generating the most suitable captions which helps the visually impaired people.

Figure 2 signifies the hybrid combination of CNN and LSTM which represents the ECANN architecture. The most commonly used deep learning architectures are the CNN and LSTM. Here, the CNN models are used to generate the captions and LSTM models follows the revere image search strategy to select the most probable caption from the pre-existing captions for the target image. The proposed ECANN model utilizes the capability of all the layers to learn the internal time-series representation of data and use the LSTM network utilize the training set attributes to categorize the short as well as long-term dependencies. CNN is otherwise named as ConvNet which is a form of ANN model and used for the purpose of image analysis. CNN models allows the user to provide an image input which further assigns learnable weights and biases to generate pre-trained image captions. The layers in CNN model are as follows: Input layer (IL), Convolutional layer (CL), Pooling Layer (PL), Fully Connected Layer (FCL) and Output Layer (OL). The architecture of CNN is illustrated in Fig. 3.

The first layer of CNN is the IL which is responsible for passing the input to the ECANN model. The size of the image input is 3-dimensiional and it can either be black and white or coloured. This IL layer is different from the other layers which may not contain the weighted inputs. It holds neurons of artificial input which are passive in nature. The next CL is the main layer where maximum number of computations occur. The main function of this CL is to extract the minute details from the image with the use of multiple hidden layers. The number of CLs is more than one and it includes filters for detecting the objects and patterns. The PL is used to divide the features into patches and ignores the useless image details. In between the CLs the PLs are used. The two forms of pooling are Max_pooling and Average_pooling. Max_pooling suppress the image noise and reduce the image dimensions. Average_pooling is also used to minimize the dimensions but the performance is lower when compared to Max_pooling. The FCL is otherwise called as feed forward (FF) layer which is present next

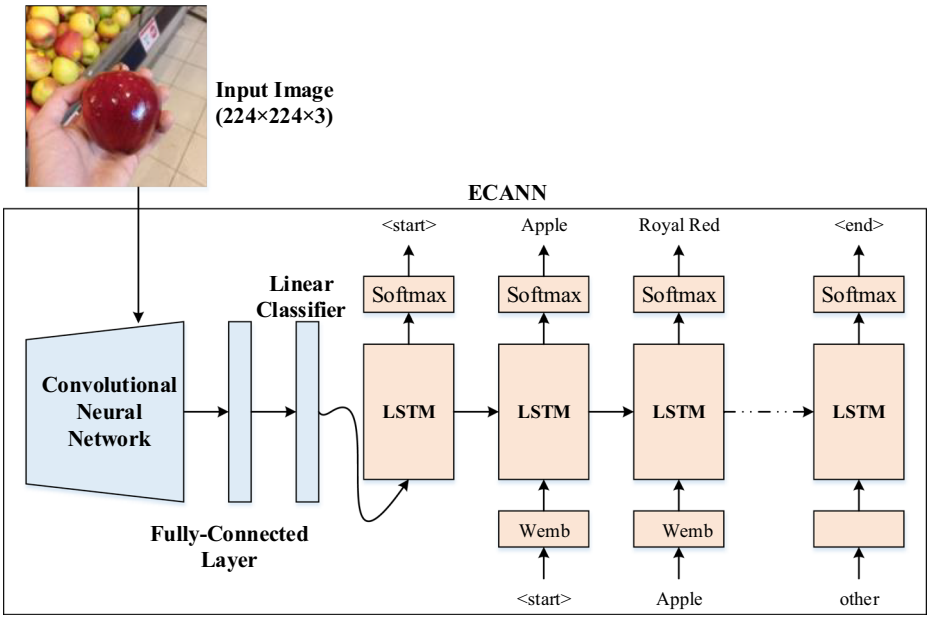

Fig. 2 Schematic representation of ECANN model

to pooling layer. The input to FCL is the output received from the last PL. This FC layer is very dense and each node is being connected to each and every other node which is existing in the previous layer. The last layer is the OL or Softmax layer where the pre-trained captions are generated based on the highest probability acquired by the FC layer.

The final output of CNN is the generation of pre-trained captions on identifying the objects in the input image. Therefore, the hybrid ECANN model that exploits the advantages of both deep learning techniques and enhance the prediction accuracy. LSTM models are more-suited for processing, classifying and making accurate predictions. LSTM is a kind of RNN (Recurrent Neural Network) which can be used to overcome the problems of long term dependency. LSTM network is used to execute the machine translation task and can be used as a language model to create the most appropriate captions based on the given input vector. Figure 4 signifies the diagrammatic representation of LSTM memory block.

LSTM networks are the broadly used type of RNN structural design. The architecture of LSTM is better than traditional RNNs for capturing long-term dependencies. The difference among the LSTM model and regular RNN is that each traditional node in the hidden layer of

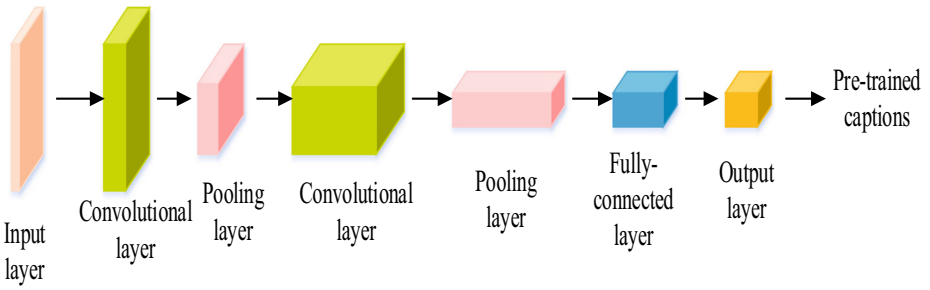

Fig. 3 Architecture of CNN

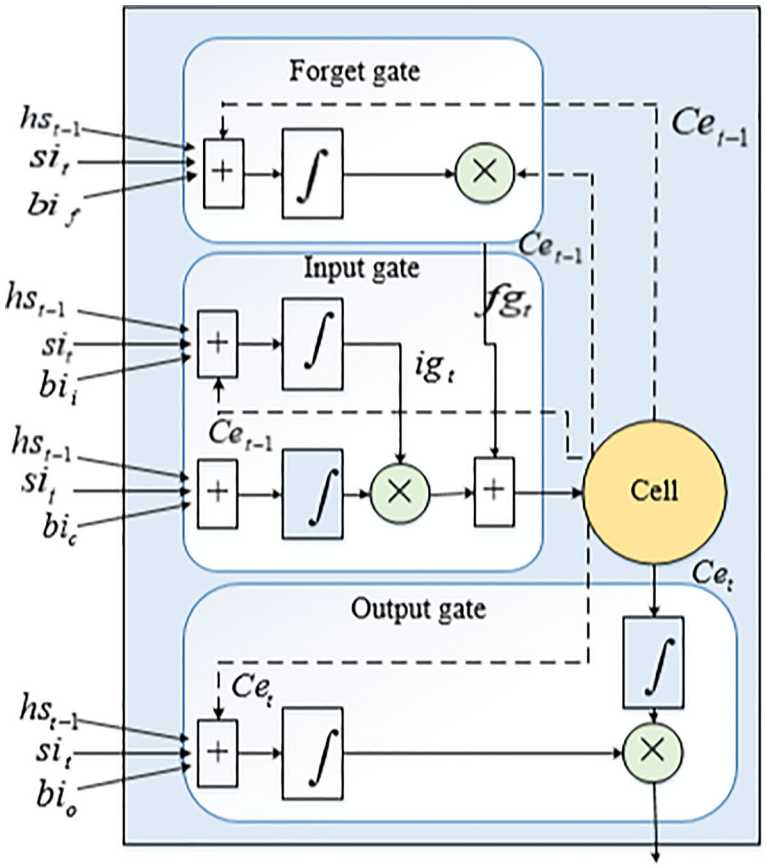

Fig. 4 Memory Block of LSTM

LSTM is interchanged with memory cells (MCs). The architecture of LSTM includes memory cells, gate units, and memory blocks. The main component of LSTM is the cell state which indicates a horizontal straight line that connects the network structure. The gate units in LSTM are utilized to control the flow of information. The tri-gate units are namely input, output, and forget gate. With the combination of forget, input gate the information is disappeared or stored from the MC. Also, the multiplicative input gate units are utilized in order to eliminate the harmful effects originated from the irrelevant inputs. However, the input gate controls the input flow towards the MC while the output of hidden state (hs_t) is controlled by means of output gate. The sigmoid (σ) activation is used for the execution which specifies the range 0 to 1.

The forget gate is controlled using single-layered NN (Neural Network), in LSTM memory block. The sigmoid function is expressed as:

$$fg_t = \sigma\left(We_f si_t + We_f hs_{t-1} + bi_f\right) \tag{14}$$

Where, fg_t denotes forget gate, ***We*** indicate weight vectors, input series is si_t, hidden state of previous block (hs_{t-1}) and bias is denoted as bi_f.

After the input squashing, the input gate ranges from 0 to 1 and newer memory is created by the simple NN which has the activation function as well as the previous memory block. The expressions for the input gate and input squashing function are:

$$ig_t = \sigma(We_i si_t + We_i hs_{t-1} + bi_i) \tag{15}$$

$$si_t = \tanh(We_s si_t + We_s hs_{t-1} + bi_s) \tag{16}$$

The MC activation function processing at time t is given as,

$$ce_t = ig_t \otimes si_t + fg_t \otimes ce_{t-1} \tag{17}$$

The output gate is denoted in Eq. (18) and the Eq. (19) illustrates the hidden vector

$$og_t = \sigma(We_o si_t + We_o hs_{t-1} + bi_o) \tag{18}$$

$$hs_t = og_t \otimes \tanh(ce_t) \tag{19}$$

Where, the element-wise multiplication is $\otimes$. Thus, the LSTM includes number of layers where each one is linked with each other as a form of repeating chain modules. The captions generated can be given as input to each of the LSTM layers. One caption is inputted to each layer and this LSTM model learn from these captions and gets optimized by itself. Each one of the LSTM layer predicts the most suitable caption for the input image. The last Softmax layer (SL) of LSTM is iteratively trained to reduce the loss present in the network. The loss function in SL is named as cross-entropy (CE) loss which is expressed in Eq. (20). However, this loss is optimized by means of adaptive atom search (AAS) algorithm.

$$CE\left(\widehat{z}\right) = -\log\left(\widehat{z}\right) \tag{20}$$

The AAS algorithm is a meta-heuristic algorithm based on the population criterion and inspired by the behavior of molecular dynamics. It mimics the atomic motion being controlled by the constraint forces and interaction in order to design the model which is very effective to solve the problems related to global optimization. The atomic motion is expressed as,

$$F_i + C_i = ac_i \times ma_i \tag{21}$$

Where, the resulting force of interaction on atom I is F_i, the constraining force on atom I is C_i, acceleration is ac and mass is signified as ma. However, the AAS algorithm signifies a feasible solution in the algorithm search space. The significance of the algorithm model is represented using the atom of the feasible solution. All the atoms in the structure repel or attract to each other depending on the distance characteristics which makes the lighter atoms to get attracted towards the heavier atoms. The fitness function can be evaluated by considering the mass $ma_i t$ at t-th iteration of atom ® represented as,

$$ma_i(t) = \frac{S_i(t)}{\sum_{j=1}^{N} S_j(t)} \tag{22}$$

$$S_i(t) = e^{-(F_i(t)-F_b(t)/F_w(t)-F_b(t))} \tag{23}$$

where, N signifies the number of total atoms, the fitness function of atom ® at *t*-th iteration can be represented as $F_i(t)$ the fitness of best and worst atoms can be given as$F_w(t)$, $F_b(t)$. The normal AS (Atom Search) optimization algorithm make use of the random initialization of solutions whereas the AAS algorithm procedure use the chaotic map based initialization for the evaluation of the fitness value. Choas is regarded as a non-linear phenomenon used in the SI (Swarm Intelligence) algorithm initialization by avoiding the algorithm to fall under the strategy of local optimum. The circular chaotic map is used for the atomic population initialization. The updated condition on velocity, position at (t + 1) iteration are expressed as:

$$V_i^{di}(t+1) = rand_i^{di} V_i^{di}(t) + ac_i^{di}(t) \tag{24}$$

$$X_i^{di}(t+1) = X_i^{di}(t) + V_i^{di}(t+1) \tag{25}$$

where, *di* represents the dimension, V_i denotes the velocity and X_i^{di} represents the updated position. By using the immune detection (ID) operator the accuracy as well the convergence of the AAS is improved which is illustrated based on Eq. (26),

$$X_i(t+1) = \begin{cases} X_{N_i}(t), F_i(t). > NF_i(t) \\ X_i(t), F_i(t) \leq NF_i(t) \end{cases} \tag{26}$$

In ECANN, the loss function is minimized and the accuracy is improved using AAS which further optimize and update the network weights. Also, this ECANN model can manage the process of training intended for the automatic generation of image captions. Thus, the entire ECANN structure learns to create the most suitable image captions and supports the task of labelling or captioning the images in an automatic manner. Table 2 illustrates the hyperparameters of ECANN model.

4 Implementation results

This section reveals the details about the implementation procedure and the analysis on the evaluation metrics. The proposed ECANN model is implemented in the PYTHON platform.

Table 2 Hyperparameters of ECANN model

Sl. No	Hyperparameters	ECANN
1.	Learning algorithm	AAS
2.	Initial learning rate	0.001
3.	Activation Function	sigmoid
4.	Loss Function	cross-entropy
5.	Mini batch size	30
6.	Size of word vectors	100
7.	Maximum length of sequence	1000
8.	No of neurons in hidden layer	100
9.	Hidden layer	10
10.	Max epochs	100

The proposed deep learning based AIC generation is used to achieve better outcomes. By evaluating the overall model on various metrics the efficiency of the proposed technique is better analyzed. This will lead to generate the image captions precisely and effectively. The performance results on proposed ECANN approach is better when compared with other similar methods.

4.1 Description about dataset

The performance of proposed ECANN method is tested on two publicly available datasets such as Freiburg Groceries and Grocery Store Datasets. Figure 5 shows the sample images of Freiburg Groceries Dataset. However, the first Freiburg dataset includes total 4947 images with 25 image categories. The images were collected from apartments, offices and stores in Germany with phone cameras. In this work, the total number of images used for processing are 4947 with training (3462 images) and testing (1485 images). Here the training and testing data is considered in the ratio as 70:30.

Figure 6 shows the sample images of Grocery Store Datasets The second Grocery Store dataset involves natural and ionic images. This dataset is used in the classification of natural images and assists the blind people. The images are recorded using the phone camera. The total image in this dataset is 5125 having 80-classes. Here, we consider 22 classes with 862 images. From the grocery store website, the iconic images are taken which represents the product information like nutrient values, weight, and origin country. In this work, the total number of images used for processing are 862 which includes 603 training data and 259 testing data. Here the training and testing data is considered in the ratio as 70:30.

Fig. 5 Sample images of freiburg groceries dataset

Fig. 6 Sample images of grocery store dataset

4.2 Evaluation metrics

The performance of the proposed method is evaluated using the metrics such as accuracy, recall, precision, F-score. The outcome shows that the proposed ECANN method provides high performance than the other existing methods. The performance metrics are explained as follows:

Precision (P): It can be defined as the ratio of number of positive samples that is classified from the total number of samples. It is based on percentage of cases that are wrongly categorised.

$$P = \frac{TP}{TP + FP} \times 100\% \tag{27}$$

Recall (R): It can be defined as the ratio of number of positive samples classified as positive to total number of positive samples. It is based on percentage of cases that are accurately categorised.

$$R = \frac{TP}{TP + FP} \times 100\% \tag{28}$$

F-score: This metric also called as F1 score and F-measure which is utilized for testing the weighted consonant mean of recall and precision.

$$F\text{-}score = \frac{2*P*R}{P+R} \tag{29}$$

Accuracy (A): It is defined as the proportion of correctly identified samples to that of the total number of samples. The value close to the true value is defined as the accuracy.

$$A = \frac{TP + TN}{TP + TN + FN + FP} \times 100\% \tag{30}$$

Where, TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative).

4.3 Performance analysis

The outcomes of the proposed ECANN is analysed using the evaluation metrics accuracy, recall, F-score and precision on the groceries dataset (Freiburg and Grocery Store). Also, the supremacy of the proposed ECANN method is compared with other DL models to illustrate the effectiveness of the proposed approach.

Table 3 signifies the results of proposed ECANN method. The outcomes obtained on testing with the dual groceries dataset signifies the effective nature of the proposed ECANN based automatic image caption generation. For both the dataset, the classification accuracies are obviously higher with the softmax classifier, and the fine-tuning process offers enhanced results consistently. Thus, the accuracy obtained with Grocery Store Dataset is higher (99.46%) whereas the Freiburg Groceries Dataset obtained equivalent performance of (99.32%) accuracy.

Table 4 illustrates the comparison of precision and recall on various datasets. The various DL based methods can be used to compare with the proposed ECANN architecture. The metrics used for comparing the evaluated results are the precision and recall. The proposed ECANN method is evaluated on the 2 different groceries datasets whereas the existing DL models were assessed on other groceries datasets which identify the objects or food items in certain constrained environments. Hence, the results prove that the proposed ECANN method have acquired higher outcomes in terms of recall as well precision with the other DL models.

Table 5 represents the accuracy metric comparison on Grocery Dataset. For the evaluation of the automatic generation of image captions, the accuracy metric is used more precisely which illustrates the ratio of predictions the proposed ECANN model has acquired right. Hence, the highest score indicates the accurateness of the captions generated. Here, the proposed ECANN model gained higher outcomes compared to the other DL models. Table 6 signifies the accuracy metric comparison on Freiburg Dataset. Here, the existing DenseNet-169, VGG16 and AlexNet models are implemented in this work for Freiburg Dataset. The higher accuracy value is obtained using ECANN is (99.32%) and the existing CaffeNet [11],

Table 3 Outcomes of ECANN based automatic image caption generation

Method	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
ECANN (Proposed)	Freiburg Groceries Dataset	99.32	99.73	98.94	99.33
	Grocery Store Dataset	99.46	99.35	99.57	99.46

Table 4 Comparison on precision and recall [6]

Metrics/Method	Dataset	Precision (%)	Recall (%)
ECANN (Proposed)	Freiburg Groceries Dataset	**99.73**	**98.94**
	Grocery Store Dataset	**99.35**	**99.57**
VGG16	Grozi-3.2 K	61.73	43.22
	Grozi-120	50.44	30.69
	GP-20	90.95	92.82
	GP-180	89.92	87.63
VGG16+AT(BRISK)	Grozi-3.2 K	64.78	46.22
	Grozi-120	46.32	29.50
	GP-20	94.33	93.85
	GP-180	85.55	80.74
VGG16+AT(SIFT)	Grozi-3.2 K	65.83	45.52
	Grozi-120	49.05	29.37
	GP-20	93.85	93.85
	GP-180	92.19	87.89
ResNet-18 [16]	Grocery Store dataset	89.97	88.95
ResNet-101 [16]	Grocery Store Dataset	93.68	93.55
DenseNet-169 [16]	Grocery Store Dataset	93.22	92.55

Table 5 Accuracy comparison on grocery store dataset [14]

Methods	Accuracy (%)
ECANN (Proposed)	99.46
DenseNet-169	84.0
VGG16	73.8
AlexNet	69.3

DenseNet-169, VGG16 and AlexNet models acquired lower accuracy values as 78.9%, 82.51%, 70.86% and 67.43%.

Table 7 denotes the comparison on accuracy considering with and without feature extraction using Grocery store dataset. The ECANN approach used the softmax classifier for better classification whereas the other DL approaches used the SVM (Support Vector Machine) based classification. The accuracy acquired with feature extraction provided higher results than without processing the feature extraction phase.

Table 8 shows the accuracy comparison of proposed and existing models with and without feature extraction using Freiburg dataset. Here also, the similar existing approaches are considered for comparison on with and without feature extraction. From Table 7, it is clearly proved that the proposed ECANN model with feature extraction obtained better accuracy than

Table 6 Accuracy comparison on Freiburg dataset

Methods	Accuracy (%)
ECANN (Proposed)	99.32
CaffeNet [11]	78.9
DenseNet-169	82.51
VGG16	70.86
AlexNet	67.43

Table 7 Accuracy comparison on with and without feature extraction (Grocery Store Dataset) [14]

Method/ Classifier	ECANN (Proposed)		$VGG16_6$		$VGG16_7$		$AlexNet_6$		Densenet-169	
	Softmax	Softmax-ft	SVM	SVM-ft	SVM	SVM-ft	SVM	SVM-ft	SVM	SVM-ft
Accuracy (%)	90.14%	99.46%	62.1%	73.3%	57.3%	71.7%	69.2%	72.6%	72.5%	85.0%

existing DL models. Tables 7 and 8 illustrates the accuracy obtained by the proposed ECANN model using the softmax classifier and existing models using SVM classifier with and without feature extraction.

Table 9 depicts the accuracy achieved by the proposed ECANN and existing $VGG16_6$, $VGG16_7$, $AlexNet_6$, and Densenet-169 using softmax classifier for both datasets. When comparing the obtained accuracy of proposed ECANN for two datasets, the accuracy obtained by proposed ECANN for Grocery store dataset is slightly higher than the Freiburg dataset. From comparative analysis, the accuracy achieved by the proposed ECANN is higher than the existing DL models. Table 10 illustrates the comparison of proposed model with basic models for the two datasets using softmax classification.

Figure 7 specifies the evaluation results in terms of Graphical analysis. With the Freiburg dataset, the outcomes obtained can be accuracy (99.32%), recall (98.94%), precision (99.73%) and F-score (99.33%). The outcomes acquired from Grocery datasets can be accuracy (99.46%), precision (99.35%), F-score (99.46%) and recall (99.57%). The proposed ECANN model on both datasets gained equivalent results on the automatic image caption generation.

Figure 8 illustrates the graphical assessment on accuracy comparison. The proposed ECANN model evaluated on the Grocery Store Dataset acquired higher accuracy (99.46%) when compared to other DL architectures such as DenseNet-169 (84.0%), VGG16 (73.8%) and AlexNet (69.3%). The proposed ECANN approach gained improved results due to the ARO based selection of non-captioned images, feature extraction and DL based automatic image caption generation.

Figure 9 illustrates the accuracy comparison of proposed ECANN and existing CaffeNet, DenseNet-169, VGG16 and AlexNet models for Freiburg dataset. Table 5 and Fig. 9 shows that the proposed ECANN model achieved higher accuracy than the existing models. Thus, the proposed ECANN model is fit for automatic generation of image captions for visually impaired people.

Figure 10 represents the comparison on accuracy based on with and without feature extraction for Grocery store dataset. The proposed ECANN approach using Softmax classification gained higher accuracy (99.46%) with feature extraction. Whereas, without performing feature extraction using Softmax classifier the proposed ECANN model obtained lower accuracy (90.14%). The existing DL models such as $VGG16_6$, $VGG16_7$, $AlexNet_6$ and

Table 8 Accuracy comparison on with and without feature extraction (Freiburg Dataset)

Method/ Classifier	ECANN (Proposed)		$VGG16_6$		$VGG16_7$		$AlexNet_6$		Densenet-169	
	Softmax	Softmax-ft	SVM	SVM-ft	SVM	SVM-ft	SVM	SVM-ft	SVM	SVM-ft
Accuracy (%)	90.03%	99.32%	61.58%	72.7%	55.12%	70.9%	68.49%	71.6%	72%	84.79%

Table 9 Accuracy comparison Proposed and existing models using softmax classifier with and without feature extraction (Grocery store and Freiburg dataset)

Method/Classifier	ECANN (Proposed)		$VGG16_6$		$VGG16_7$		$AlexNet_6$		Densenet-169	
	Softmax	Softmax-ft	Softmax	Softmax-ft	Softmax	Softmax-ft	Softmax	Softmax-ft	Softmax	Softmax-ft
Accuracy (Grocery store dataset)	90.14%	99.46%	66.9%	78.76%	61.72%	74.58%	72.49%	76.93%	77.3%	87.62%
Accuracy (Freiburg dataset)	90.03%	99.32%	64.56%	77.12%	60.92%	72.82%	70.03%	74.13%	75.63%	86%

Table 10 Comparative assessment of proposed with other basic models

Method	Accuracy (%)		Precision (%)		Recall (%)		F-score (%)	
	Freiburg Groceries Dataset	Grocery Store Dataset	Freiburg Groceries Dataset	Grocery Store Dataset	Freiburg Groceries Dataset	Grocery Store Dataset	Freiburg Groceries Dataset	Grocery Store Dataset
ECANN	**99.32**	**99.46**	**99.73**	**99.35**	**98.94**	**99.57**	**99.33**	**99.46**
CNN	97.01	97.56	97.64	97.22	96.35	97.21	97.16	97.53
RNN	96.26	96.74	96.57	96.63	95.47	96.43	96.32	96.76
DNN	96.87	96.93	96.93	96.87	95.39	96.92	96.88	96.92
DBN	95.29	95.37	95.76	95.54	94.62	95.67	95.31	95.42

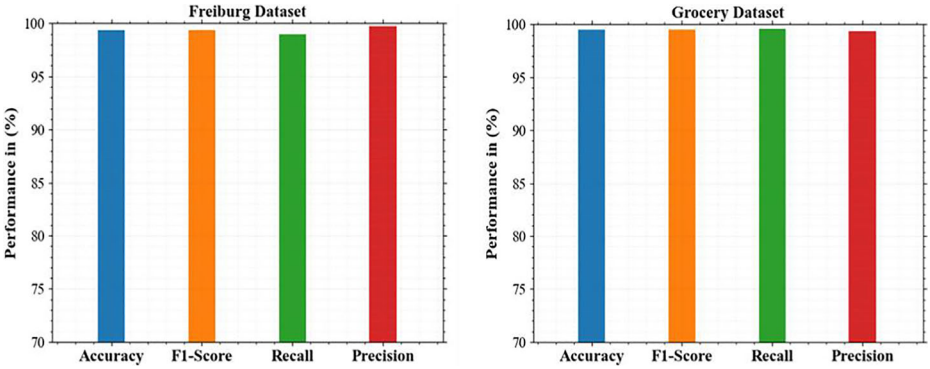

Fig. 7 Graphical assessment of ECANN based image caption generation

DenseNet-169 using the SVM classification achieved lower results on with and without feature extraction.

Figure 11 depicts the comparative analysis of proposed and existing models with and without feature extraction for Freiburg dataset. Similar to Fig. 10, here also the accuracy of the

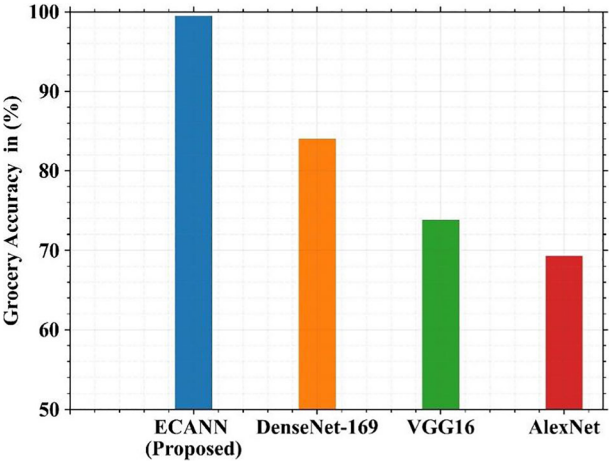

Fig. 8 Accuracy comparison on various DL models for grocery store dataset

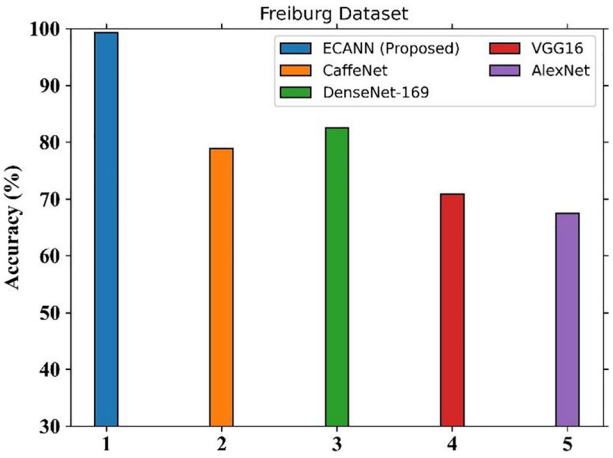

Fig. 9 Accuracy comparison on various DL models for Freiburg dataset

proposed ECANN model with feature extraction obtained higher accuracy (99.32%) than without feature extraction accuracy (90.03%). Likewise, when comparing the accuracy of proposed model with existing models, which obtained lower accuracy on with and without feature extraction.

Figure 12a and b represents the accuracy evaluation of proposed ECANN and existing DL models using softmax classifier with and without feature extraction on grocery store and Freiburg datasets. In the above all result analysis, the existing models used SVM classifier for the automatic image caption generation. Here, the accuracy is computed for proposed ECANN and existing approaches namely $VGG16_6$, $VGG16_7$, $AlexNet_6$ and DenseNet169 using softmax classifier. From the comparative analysis, it is cleared that the deep learning models using softmax classifier provided better accuracy than SVM classifier.

Figure 13 specifies the ROC curve for ECANN approach. This ROC (Receiver Operating Characteristics) curve is used to execute the quantitative analysis. It defines the plot between

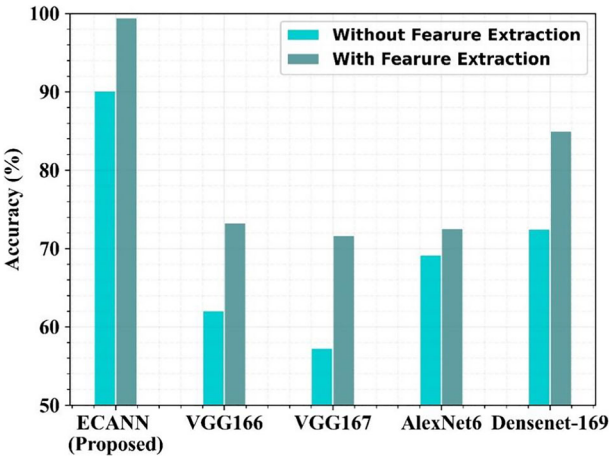

Fig. 10 Accuracy comparison on different models with and without feature extraction

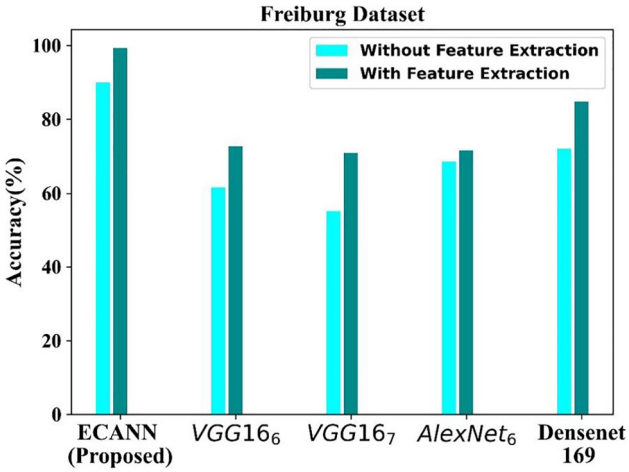

Fig. 11 Accuracy comparison of different models on with and without feature extraction

TPR (True Positive Rate) and FPR (False Positive Rate). The AUC of the ROC curve space is 0.994 and 0.993 for dataset Grocery and Freiburg respectively.

Figure 14 represents the comparison of precision, recall on various datasets. The proposed ECANN obtained higher values of precision (9.73%), recall (98.94%) on Freiburg Groceries dataset and the Grocery Store dataset gained higher values precision (99.35%) and recall (99.57%). The existing methods such as VGG16, VGG16 + AT (BRISK), VGG16 + AT (SIFT), ResNet-18, ResNet-101 and DenseNet169 acquired lower values when tested with the different Groceries dataset.

Figure 15 shows the training and testing metrics for loss and accuracy are considered for the proposed ECANN classification model. For evaluating the learning performance, the curves like accuracy and loss curves are determined. The training and testing phase for accuracy and loss are carried out by varying the size of epoch from 1 to 100 sequentially. The enhancement in accuracy and minimization in loss take place because of increasing epoch size. The ECANN model attains a training accuracy of 89%, 96% and 97% respectively when the epoch size is

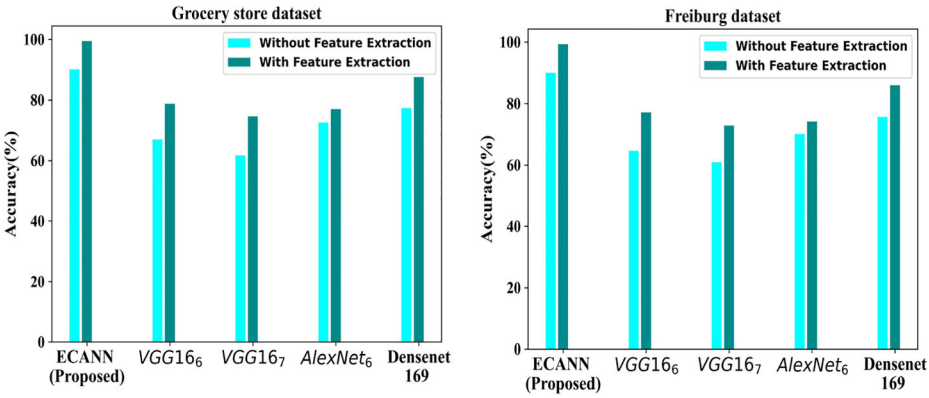

Fig. 12 **a** and **b** Accuracy comparison of various DL models using softmax classifier with and without feature extraction

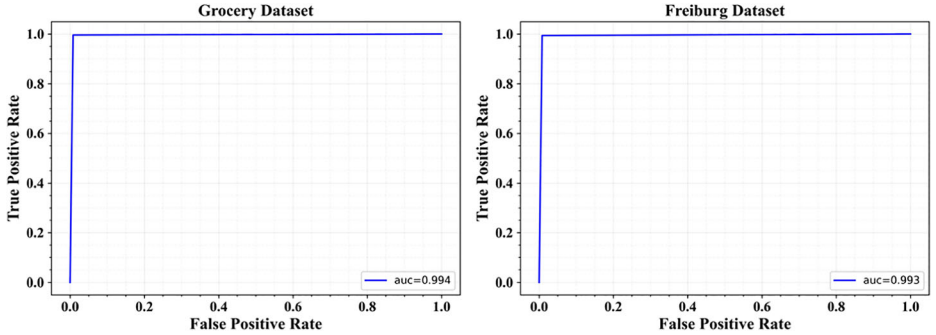

Fig. 13 ROC curve

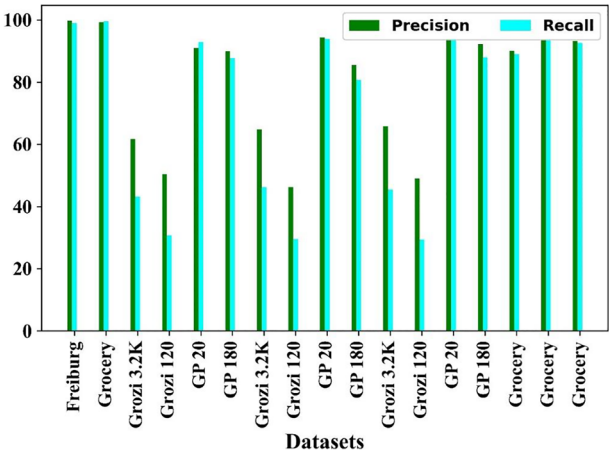

Fig. 14 Precision and recall comparison on different datasets

20, 40 and 60. From the Fig. 15a, it is noted that the proposed model obtains maximum accuracy values, and it is same for both cases (testing and training). For performing the training phase, the better values are arranged and when the training stage is exit when the best values

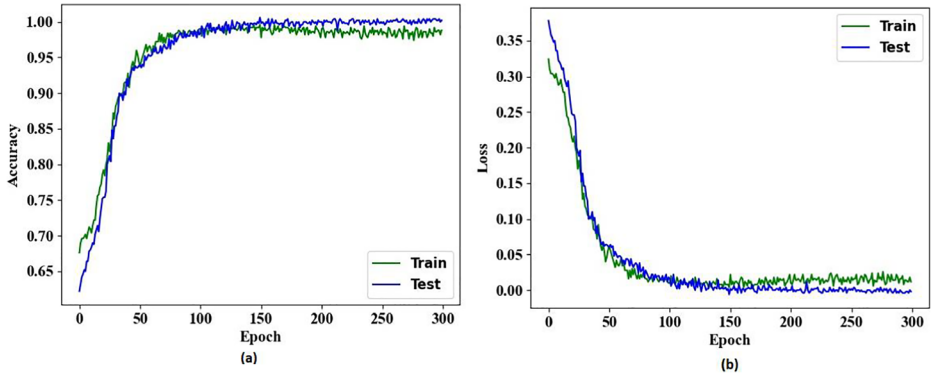

Fig. 15 Training and testing performance of the proposed ECANN model (**a**) Accuracy (**b**) Loss

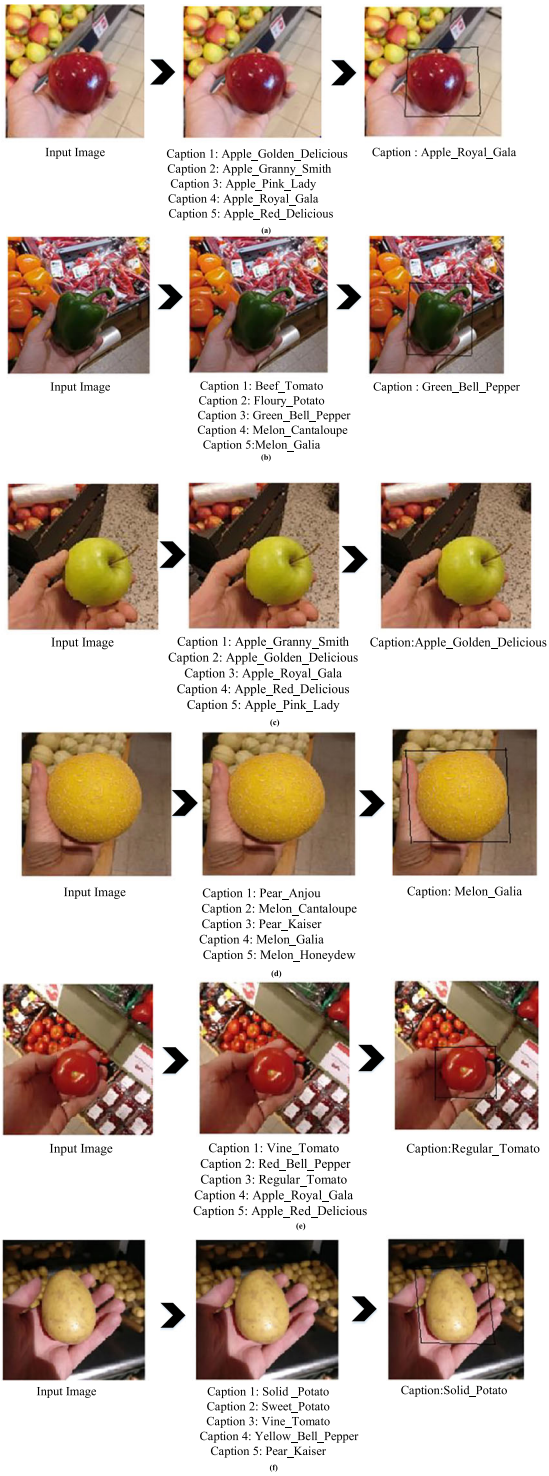

Fig. 16 Outcomes of image captions

are evaluated. Figure 15b depicts the losses of both training and validation cases. In a loss curve, when the size of epoch are20, 40 and 60, it has a training loss of 2.2, 0.2 and 0.3 respectively.

Figure 16 signifies the outcomes of automatically generated image captions. The proposed ECANN builds a model by generating pre-trained image captions with CNN and make use of LSTM network to process on these captions and selects the most relevant image caption and removes the non-relevant image caption.

4.4 Discussion

This section discusses the results of the proposed ECANN and its comparative analysis. The novelty of the proposed work is that automatic generation of image captions with the ECANN deep learning model that uses computers to come up with the solution of solving the inconveniences faced by VI people in shopping grocery items. The proposed ECANN model includes the combination of dual deep networks namely CNN and LSTM architectures to perform a caption reusable system based on AI (Artificial Intelligence) with a reverse image search to reuse pre-existing captions for the target image and selects the most accurate image captions. However, this proposed ECANN model is used to generate alternate captions to images which should be semantically genuine to the original image. Also, the ECANN network is trained using the optimization algorithm AAS. Optimization algorithms are responsible for reducing the losses and provide the most accurate results. Moreover, the optimization algorithms are used to change the neural network attributes such as learning rate and weights in order to reduce the losses. In the proposed work, the AAS algorithm is used for optimizing the network attributes and it is a very competitive optimization algorithm and has been applied in various research fields such as flow scheduling problems, economic load dispatch problems, detection, and classification. When comparing with other optimization algorithms this AAS algorithm benefits from avoiding high local optima, which leads to avoidance of overlapping features during classification and helps to optimize the hyperparameters in the network to obtain the most accurate results. In addition, AAS optimization is integrated with ECANN which finds the best acceptable solution for a given problem. In this framework, error occurrence in the image captioning system is reduced using the AAS algorithm.

The efficiency of the proposed automatic image caption generation approach is computed in terms of precision, recall and accuracy. The precision, accuracy and recall attained by the proposed method is compared with different existing models in the above subsection. From the comparative analysis, the proposed ECANN shows higher efficiency than the existing models in automatic generation of image caption for easy web access of blind people. Two publicly available datasets such as Freiburg and Grocery store database are utilized for the performance evaluation. Moreover, the state-of-the-art deep learning models are considered for the performance comparison namely VGG16, AlexNet, DenseNet-169 [14] and CaffeNet [11] models. These existing pre-trained classifiers has shown lower performance because of the accumulation of error generation and inappropriate text context. These classifier models shown severe complexity during the network training since each caption can be treated similarly without mentioning the importance of diverse words. Moreover, during the generation of captions the scenes or semantic objects can be wrongly recognized. It is very challenging to describe the image content automatically using the accurately formed English language sentences and greatly impact in

assisting the VI people in everyday life. Due to these drawbacks, the existing classifiers shown lower performance in image caption generation. Hence in proposed work, the recall, precision and accuracy of the ECANN is evaluated separately for both the datasets and compared with various deep learning models. The ROC curve depicts the accurate outcomes of image captions with less error. The accuracy of the proposed model using softmax classifier with feature extraction is higher than without feature extraction. The feature extraction plays a vital role in the accurate image caption generation process. It is proved from the comparative analysis with and without feature extraction. From the overall comparative analysis, it is clearly showed that the proposed ECANN is appropriate for automatic generation of image captions, and it make web accessing easier for blind people.

5 Conclusion

This work presents a new deep learning based human-computer interaction for automatic generation of image captions which benefits the blind people. The proposed ECANN model acts as an effectual tool in image caption generation which use the reverse search strategy to select the most suitable captions for the input image. The presented EACNN based automatic image caption generator model assists the blind people by incrementing their spatial awareness level and building the Internet to be more accessible. The ECANN approach collects the image data from dual online sources which includes the packaged items (juice products, vegetables, and fruits). Initially, the captioned and non-captioned images are clustered using the ARO model which saves the processing time. Next, the appearance and texture features are extracted from the images which signifies more visual information about these images. The ECANN model is the hybrid of CNN and LSTM networks in which CNN generates the pre-trained captions and the LSTM executes the reverse search strategy to select the most suitable caption for the image input and the loss in the network is minimized using the optimization named AAS. Moreover, this ECANN generates automatic captions, and the images are better described to the people with low vision or blind. Thus, the proposed ECANN model achieved (99.46%) accuracy on Grocery store Dataset and (99.32%) accuracy on Freiburg Groceries dataset which better supports the blind people in distinguishing the food items. The major limitation of the proposed AIC generation is that no hardware platforms namely smart glasses or smart phones are integrated. In future, we will extend this work by analyzing the performance on bigger datasets like MSCOCO or Flickr30k in order to develop precise subtitles. Further, we try to embed NLP and smart phone application with speaker options may help the VI people to understand the surrounding events which turns the life of VI people an enjoyable experience. Also, we plan to use new attention based deep learning model to generate image captions that deal with multiple languages.

Data availability statement Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest Tejal Tiwary and Rajendra Prasad Mahapatra declared that they have no conflict of Interest.

References

1. Al-Muzaini HA, Al-Yahya TN, Benhidour H (2018) Automatic Arabic image captioning using RNN-LST M-based language model and CNN. Int J Adv Comput Sci Appl 9(6):67–73
2. Amritkar C, Jabade V (2018) Image caption generation using deep learning technique. In 2018 fourth international conference on computing communication control and automation (ICCUBEA). IEEE, Pune, pp 1–4
3. Bai S, An S (2018) A survey on automatic image caption generation. Neurocomputing 311:291–304
4. Bigham JP, Lin I, Savage S (2017) The effects of not knowing what You Don't know on web accessibility for blind web users. In proceedings of the 19th international ACM SIGACCESS conference on computers and accessibility, 101-109
5. Deng Z, Jiang Z, Lan R, Huang W, Luo X (2020) Image captioning using dense net network and adaptive attention. Signal Process Image Commun 85:1–9
6. Geng, W, Han F, Lin J, Zhu L, Bai J, Wang S, He L, Xiao Q, Lai Z (2018) Fine-grained grocery product recognition by one-shot learning. In Proceedings of the 26th ACM international conference on Multimedia, pp 1706–1714
7. Giraud S, Thérouanne P, Steiner DD (2018) Web accessibility: filtering redundant and irrelevant information improves website usability for blind users. International Journal of Human-Computer Studies 111:23–35
8. Guinness D, Cutrell E, Morris MR (2018) Caption crawler: enabling reusable alternative text descriptions using reverse image search. In proceedings of the 2018 CHI conference on human factors in computing systems, Montréal, QC, Canada, pp 1–11
9. Hossain MDZ, Sohel F, Shiratuddin MF, Laga H (2019) A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR) 51(6):1–36
10. Iwamura K, Kasahara JYL, Moro A, Yamashita A, Asama H (2021) Image captioning using motion-CNN with object detection. Sensors 21(4):1–13
11. Jund P, Abdo N, Eitel A, Burgard W (2016) The freiburg groceries dataset. arXiv preprint arXiv: 1611.05799
12. Khurram I, Fraz MM, Shahzad M, Rajpoot NM (2021) Dense-captionnet: a sentence generation architecture for fine-grained description of image semantics. Cogn Comput 13(3):595–611
13. Kim D-J, Choi J, Oh T-H, Kweon IS (2019) Image captioning with very scarce supervised data: adversarial semi-supervised learning approach arXiv preprint arXiv:1909.02201
14. Klasson M, Zhang C, Kjellström H (2019) A hierarchical grocery store image dataset with visual and semantic labels. In 2019 IEEE winter conference on applications of computer vision (WACV), 491-500
15. Kuber R, Yu W, Strain P, Murphy E, McAllister G (2020) Assistive multimodal interfaces for improving web accessibility. UMBC Information Systems Department Collection
16. Leo M, Carcagnì P, Distante C (2021) A systematic investigation on end-to-end deep recognition of grocery products in the wild. In 2020 25th international conference on pattern recognition (ICPR), IEEE, 7234-7241
17. Loganathan K, Kumar RS, Nagaraj V, John TJ (2020) CNN & LSTM using python for automatic image captioning. Materials Today: Proceedings, CNN & LSTM using python for automatic image captioning, pp 1–5
18. MacLeod H, Bennett CL, Morris MR, Cutrell E (2017) Understanding blind people's experiences with computer-generated captions of social media images. In proceedings of the 2017 CHI conference on human factors in computing systems, 5988-5999
19. Makav B, Kılıç V (2019) A new image captioning approach for visually impaired people. In 2019 11th international conference on electrical and electronics engineering (ELECO), IEEE, 945-949
20. Melas-Kyriazi L, Rush AM, Han G (2018) Training for diversity in image paragraph captioning. In proceedings of the 2018 conference on empirical methods in natural language processing, 757-761
21. Sadeghi D, Shoeibi A, Ghassemi N, Moridian P, Khadem A, Alizadehsani R, Teshnehlab M, Gorriz JM, Nahavandi S (2021) An overview on artificial intelligence techniques for diagnosis of schizophrenia based on magnetic resonance imaging modalities: methods, challenges, and future works. arXiv preprint arXiv: 2103.03081
22. Sehgal S, Sharma J, Chaudhary N (2020) Generating image captions based on deep learning and natural language processing. In 2020 8th international conference on reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE, 165–169
23. Sharma G, Kalena P, Malde N, Nair A, Parkar S (2019) Visual image caption generator using deep learning. In 2nd international conference on advances in Science & Technology (ICAST)
24. Shoeibi A, Khodatars M, Alizadehsani R, Ghassemi N, Jafari M, Moridian P, Khadem A et al (2020) Automated detection and forecasting of covid-19 using deep learning techniques: a review. arXiv preprint arXiv:2007.10785:1–20

25. Shoeibi A, Khodatars M, Jafari M, Moridian P, Rezaei M, Alizadehsani R, Khozeimeh F, Gorriz JM, Heras J, Panahiazar M, Nahavandi S, Acharya UR (2021) Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: a review. Comput Biol Med 136:104697
26. Shoeibi A, Sadeghi D, Moridian P, Ghassemi N, Heras J, Alizadehsani R, Khadem A, Kong Y., Nahavandi S., Zhang Y.D., Gorriz J.M. (2021) Automatic diagnosis of schizophrenia in EEG signals using CNN-LSTM models. Frontiers in Neuroinformatics 15
27. Shoeibi A, Ghassemi N, Khodatars M, Moridian P, Alizadehsani R, Zare A, Khosravi A, Subasi A, Acharya UR, Gorriz JM (2022) Detection of epileptic seizures on EEG signals using ANFIS classifier, autoencoders and fuzzy entropies. Biomedical Signal Processing and Control 73:103417
28. Singh A, Singh TD, Bandyopadhyay S (2021) An encoder-decoder based framework for hindi image caption generation. Multimedia tools and applications, 1-20
29. Song H, Zhu J, Jiang Y (2020) avtmNet: adaptive visual-text merging network for image captioning. Comput Electr Eng 84:1–12
30. Wei Y, Tran S, Xu S, Kang B, Springer M (2020) Deep learning for retail product recognition: challenges and techniques. Comput Intell Neurosci 1–23
31. Wu S, Wieland J, Farivar O, Schiller J (2017) Automatic alt-text: computer-generated image descriptions for blind users on a social network service. In proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, 1180–1192
32. Xiao F, Gong X, Zhang Y, Shen Y, Li J, Gao X (2019) DAA: dual LSTMs with adaptive attention for image captioning. Neurocomputing 364:322–329
33. Yang M-S, Nataliani Y (2017) Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters. Pattern Recogn 71:45–59
34. Yang M, Liu J, Shen Y, Zhao Z, Chen X, Wu Q, Li C (2020) An Ensemble of Generation-and Retrieval-Based Image Captioning with dual generator generative adversarial network. IEEE Trans Image Process 29: 9627–9640
35. You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention. In proceedings of the IEEE conference on computer vision and pattern recognition, 4651–4659
36. Yu N, Hu X, Song B, Yang J, Zhang J (2018) Topic-oriented image captioning based on order-embedding. IEEE Trans Image Process 28(6):2743–2754