Multimedia Tools and Applications (2023) 82:20283–20302
https://doi.org/10.1007/s11042-023-14434-w

CNN deep learning-based image to vector depiction

Safa Riyadh Waheed[1,2] · Mohd Shafry Mohd Rahim[1] · Norhaida Mohd Suaib[1] · A.A. Salim[3]

Received: 30 November 2021 / Revised: 22 November 2022 / Accepted: 21 January 2023 /
Published online: 31 January 2023

Abstract

In the computational science and engineering domains, the depiction of picture information remains an intricate problem. Such a description needs an accurate recognition of various objects and individuals together with their attributes, correlations, and panorama information. Based on this fact, we depict the image contents in the natural language or image description generation methods using the convolutional neural networks (CNNs)-assisted deep learning (CNN-DL) approach, wherein the images are transformed to vectors. The DL and study attributes via the machine-learned data were used to construct the complete pictures from the real world. Two sections were considered based on image classification for CNN's improvement method to develop a classification model and the good results of the classification via a novel method for describing an image to the vector of each object in the image. The learning and relationship activity included all the essential categorizing and classifying entities. In addition, the developed system was extended to handle the open detection and hazards classification. The performance evaluation (using the CIFAR dataset) of the newly developed system revealed its better strength and flexibility in managing the test images from a new-fangled and isolated field than the reported techniques.

Keywords Deep learning · Classification · Image description · CNN · Image vector

✉ A.A. Salim
asalim@utm.my

Safa Riyadh Waheed
safa_albdeary@hotmail.com

Norhaida Mohd Suaib
haida@utm.my

[1] Faculty of Engineering, School of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia

[2] Computer Techniques Engineering Department, Faculty of information Technology, Imam Jaafar Al-sadiq University, Najaf, Iraq

[3] Laser Center and Physics Department, Faculty of Science, Universiti Teknologi Malaysia, Johor, Malaysia

1 Introduction

Over a prolonged period, the common believe was that the machines could easily accomplish the human intelligence level for understanding the visual world. However, extensive research enabled the resolution of such mysteries [11, 12, 17, 21, 45]. Now, the researchers can produce a very small error rate using exceptionally deep convolutional neural networks (CNNs) and large-scale image classification [17]. During the training procedure, each image is first annotated with a label from a predefined collection of categories in order to predict the category in each image. Thus, properly supervised training enabled the computer to learn and classify an image. The image content categorization includes a predominantly classified object, which are generally the easy tasks [20]. Conversely, the scenario can be more intricate if the computers are required to understand the complicated scenes, wherein one such task is subtitling the images. The difficulties are due to two reasons [1, 50]. To begin, the system must detect significant semantic concepts in the image and understand their connectivity in order to consistently describe the entire picture content. This in turn produces a meaningful and syntactically fluent caption, including language and common-sense knowledge, without incorporating object recognition. In addition, due to the complexity of the image scenes, it is intricate to describe all nuanced and finer changes using the simple categorical characteristics. A complete description of any image contents in natural languages or image description generator methods is often inaccurate, thus lacking a thin alignment between its sub-regions and description terms in the training supervision of the model image [51]. Furthermore, different practical techniques have been developed for describing the contents of images in contrast to the image taxonomy [31]. In fact, it is instantly possible to recognize the legitimacy of the classification results via the comparison with the ground reality [50].

The determination of the correctness of the created title is extremely difficult. Generally, for the practical evaluation of an image title human is used [51]. However, assessment of human is not only expensive but also time consuming. To overcome this limitation, several automated approaches have been developed that act as a proxy for speeding up the development cycle [51]. Early techniques can categorize pictures into two groups. The matching template approaches are first recognizes the objects, actions, scenes, and qualities before describing them in a manual and rigid structure based on the photographs [29–31]. These approaches do not have fluid and easy-to-read subtleties. The other approaches are built on the descriptive methods, which select a collection of visually comparable photos from a huge database and then transfer the images' captions that are retrieved into the query image [22, 40, 46]. The words based on the query image contents are inflexible because they rely directly on the training image captions and are free to generate new captions. Deep CNNs with fluent and expressive subtlety can solve these two problems, generalising beyond the training. Particularly, pictures classified using the NNs [20, 45, 46] and object detection [55] generated renewed interest in the application of neural visual underlying networks.

Recently, visual attributes and associated descriptions for images have been incorporated into the DL-based approach [24, 55]. *Esteva* et al. advocated using a variational autoencoder for image captioning and dense image descriptions that were created for each feature [15]. *Chen* et al. used REINFORCE algorithms as a technique for self-critical succession training [9]. *Piasco* et al. aimed to optimise an assessment meter that is undetected by normal gradient algorithms. Using value and policy networks, image descriptions can be generated inside a critical actor's framework, maximising a visual semantic prize that assesses the similarity of image-derived descriptions [41]. *Yu* et al. presented the generative adversarial networks (GANs)-based models

for producing the text that can be used to generate image captions [59]. The generator was also modelled using SeqGAN as a stochastic policy for enhanced learning in the discrete output, such as text. In addition, *Lin* et al. provided a range of discriminator losses utilising RankGAN, which meticulously assessed the generated text quality, leading to an excellent generator [33]. All of these achievements encouraged the researchers to design learning enhancement strategies for direct optimization of various models to acquire further advantages [10].

In this paper, we offer a fundamentally new strategy for the image description method called Image to Vector (IV). First, build an enhanced model-based CNN to categorise images appropriately. Second, each object is described using the classification model, which shows the IV. This approach is entirely based on CNNs, which have been trained to create visual descriptions. It has an obvious advantage when it comes to supervised training on huge datasets. The system learns common discriminative characteristics for classification and description tasks. The deep network models the image descriptors' separate representations and dependencies. This method produced higher accuracy while requiring less complexity. The feature extraction in the CNN algorithm was advantaged as a precept essential by analyzing the CNN structure. It is given the possibility to reuse hidden connected layers to apply learning representation and accurately describe visual features. Through the unique design of the proposed method, we were able to obtain a new image description that outperformed the previous methods with higher accuracy and reliability and less complexity compared to the previous reports, which used an additional method to get feature extraction. To overcome the dataset's quality and size effects through training, the CNN algorithm was developed to guarantee the hidden layer weights, which are accurate and proportional to the number of objects included in the images. The rest of this study will be organised as follows: In the next sections, we discuss the importance of using deep learning in computer vision and its applications and illustrate the structure of the original CNN that we will use to build the purposed model of image classification based on the common objects in context (COCO) dataset for training and the model evaluation using the CIFAR 10 dataset as a test.

2 The criteria

Several datasets have been generated to enable picture captioning research. The collection of data using the PASCAL sentencing [43] and Flickr [58] datasets was generated to enable various picture captions. Lately, Microsoft introduced the largest image subscription dataset in the public domain, called COCO [32]. In recent years, substantial progress on picture subtitling has been made due to the availability of wide-ranging datasets. The COCO submission challenge was attended by about 15 organizations in 2015, wherein the challenge entries were assessed by a person [16]. Five human judgment metrics are shown in Table 1. The findings of measure 1 (C1) and criterion 2 were used for evaluating the competition (C2). The other measures were utilized to diagnose and interpret the results. Each job was evaluated using human judgment, with an image and two captions, one of which was generated automatically and the other by a human. The judge was requested to provide a better description of the image for M1, or the same choice if it is of similar quality. For C2, it demanded the judge, who was produced by a person. The judge was deemed to have passed the Turing test when he picked an automatically produced title or chose the "can't say" option.

Table 2 presents the results for the 15 submissions to the 2015 COCO captioning challenge. Among them was the entry for Microsoft Research (MSR), which returns the highest in the

Table 1 Measurements for human evaluation by 2015 COCO

Criterion	The meaning
C1	Assessed captions' percentages better or equivalent to human captions.
C2	Captions' percentages of passing the Turing test.
C3	Average accuracy (incorrect–correct) of captions on a scale of (1–5).
C4	The average quantity of details on 1–5 scales (deficiency in details–very comprehensive).
C5	The percentage of captions comparable to description by humans.

Turing measure, while the Google team outperformed other human subtitles in terms of their percentages. Consequently, both were jointly awarded the first prize in the COCO picture subtitling contest in 2015, resulting in the evolution of new systems since this event. Now, it is important to describe the results for human and random systems. Earlier, human assessment was never performed because of its exorbitant expense. In fact, COCO benchmarked the organizers by installing an automated evaluation server. In this process, the server received the new system-generated captions, assessed them, and automatically submitted the results of a blind test. Table 3 displays 40 references per picture for the top 24 of them (in 2015), including the SPICE human system [19]. All these 24 systems outperformed the human system. A significant disparity was observed between the finest systems and a human being because of human judgment (Table 2).

3 Need of DL

Deep learning (DL) is a component of the machine-learning (ML) method that incorporates data processing via deep networks. McClulloch and Pitts in 1943 termed DL "cybernetics" [37]. Gradually, DL garnered interest amongst researchers due to its capability and unique characteristics to imitate the manner in which the brain processes information before making decisions. Furthermore, DL has been designed to process information either through supervised or unsupervised approaches, in which learning is conducted on representations and multi-layered

Table 2 Human rating by 2015 COCO [18]

Entry	C5	C4	C3	C2	C1	Date
Human	0.352	3.428	4.836	0.675	0.638	2015
Google	0.233	2.742	4.107	0.317	0.273	2015
MSR	0.234	2.662	4.137	0.322	0.268	2015
Montreal/Toronto	0.197	2.832	3.932	0.272	0.262	2015
MSR Captivator	0.233	2.565	4.149	0.301	0.25	2015
Berkeley LRCN m-	0.204	2.786	3.924	0.268	0.246	2015
RNN	0.202	2.595	3.897	0.252	0.223	2015
Nearest Neighbor	0.196	2.716	3.801	0.255	0.216	2015
PicSOM	0.182	2.552	3.965	0.25	0.202	2015
Brno University m-	0.154	3.482	3.079	0.213	0.194	2015
RNN (Baidu/UCLA)	0.195	2.548	3.831	0.241	0.19	2015
MIL	0.159	2.915	3.349	0.197	0.168	2015
MLBL	0.156	2.42	3.659	0.196	0.167	2015
NeuralTalk	0.147	2.742	3.436	0.192	0.166	2015
ACVT	0.155	2.599	3.516	0.19	0.154	2015
Tsinghua Bigeye	0.116	2.163	3.51	0.146	0.1	2015
Random	0.013	3.247	1.084	0.02	0.007	2015

Table 3 Obtained automated measures by different image captioning systems (2016) [4, 18]

Entry	SPICE (×10)	BLEU-4	METEOR	CIDEr-D	Date
Watson Multimodal	0.204	0.344	0.268	1.123	2016
DONOT_FAIL_AGAIN	0.199	0.32	0.262	1.01	2016
Human	0.198	0.217	0.252	0.854	2015
MSM@MSRA	0.197	0.343	0.266	1.049	2016
MetaMind/VT_GT	0.197	0.336	0.264	1.042	2016
ATT-IMG (MSM@MSRA)	0.193	0.34	0.262	1.023	2016
G-RMI(PG-SPIDEr-TAG)	0.192	0.331	0.255	1.042	2016
DLTC@MSR	0.19	0.331	0.257	1.003	2016
Postech_CV	0.19	0.321	0.255	0.96	2016
G-RMI (PG-BCMR)	0.187	0.332	0.257	1.013	2016
feng	0.187	0.323	0.255	0.986	2016
THU_MIG	0.186	0.323	0.251	0.969	2016
MSR	0.186	0.291	0.247	0.912	2015
reviewnet	0.185	0.313	0.256	0.965	2016
Dalab_Master Thesis	0.183	0.316	0.253	0.96	2016
ChalLS	0.183	0.309	0.252	0.955	2016
ATT_VC_REG	0.182	0.317	0.254	0.964	2016
AugmentCNNwithDe	0.182	0.315	0.251	0.956	2016
AT	0.182	0.316	0.25	0.943	2015
Google	0.182	0.309	0.254	0.943	2015
TsinghuaBigeye	0.181	0.314	0.248	0.939	2016

features. Several breakthroughs associated with DL have been reported in terms of enhancing solutions and solving problems with the help of highly advanced computation models. Since DL can perform learning on multi-layered representations, it is considered superior at deriving outcomes for sophisticated problems. In this respect, DL-based methods for data processing and abstraction in multiple layers can be regarded as the most refined technique. These features make DL an ideal method for the investigation and analysis of data on gene expression. Ontop, DL can learn multi-layered representations, imparting flexibility to achieve correct results in a rapid way. The multi-layered representation component forms a part of the overall architecture of DL [6]. The performance of both DL and ML depends on the amount of data. DL is unable to perform on a low-dimensional learning dataset because it needs high-dimensional data for complete learning [52].

4 Deep of image classification

Image classification, localization, image segmentation, and object identification are examples of major challenges in computer vision. Among these difficulties, picture classification is the most fundamental. It serves as the foundation for various computer vision challenges. Image classification algorithms are used in a wide range of applications, including diagnostic imaging, object recognition in satellite images, traffic management systems, brake light detection, machine vision, and many more. Image classification is a fundamental activity that aims to interpret a whole image in its entirety. The purpose is to categorize the image by providing it with a label. Image classification usually involves one-object pictures. Object detection, on the other hand, includes both classification and localization operations and is used to investigate more realistic circumstances in which many items may be present in an image. Image categorization is the process of

extracting information classes from a multiband raster image. Thematic maps can be created using the image categorization raster. Based on the interaction between the analyzer and the machine during classification, there are two types of classification: supervision and unsupervision. The classification technique is a multi-step process, and image classification was created to provide an integrated environment for classifications. Convolutional neural networks (CNNs) are a type of deep learning neural network. CNN represents a significant advancement in image recognition. They are most usually employed to examine visual imagery and are extensively utilized in picture classification. They are used in anything from photo tagging to self-driving cars. It is hard at work behind the scenes in fields ranging from healthcare to security. A pixel-based image is analyzed by a computer. It accomplishes this by treating the image as an array of matrices, the size of which is determined by the image resolution. Simply put, picture classification is the processing of statistical data by a computer utilizing algorithms. Image classification in digital image processing is accomplished by automatically grouping pixels into predefined groups, referred to as "classes." The algorithms divide the image into a succession of its most noticeable elements, reducing the workload on the final classifier. These features inform the classifier about what the image represents and which class it may belong to. The characteristic extraction method is the most crucial stage in categorizing an image because it is the foundation for the remainder of the steps. Image classification, especially supervised classification, is heavily dependent on the data provided to the algorithm. A well-optimized classification dataset outperforms a bad dataset with data imbalance based on class and low image and annotation quality. "Supervised classification" and "unsupervised classification' are two of the most common methods for classifying the whole image using training data.

4.1 Supervised images classification

Supervised classification uses spectral signatures acquired from training samples to classify images. It can rapidly build training samples that represent the classes it needs to extract. It may also quickly construct a signature file from the training samples, which is then used by the multivariate classification tools to classify the image.

4.2 Unsupervised categorization

Without the intervention of an analyst, unsupervised classification discovers spectral classes (or clusters) in a multiband image. Unsupervised classification can provide access to tools for creating clusters, the ability to examine cluster quality, and references to classification tools.

4.3 CNN constructions

The construction design of the CNN consists of three layers: the entry, hidden (latent), and output. Hidden or secret layers have been referred to as the pooling, completely connected, or conveyor layers. Figure 1 depicts the fundamental CNN architecture [4]. The next sub-section provides a brief summary of this layer.

4.3.1 The convolutional layer

The convolution method is being used iteratively to perform these functions to generate a change in the output function [37]. This convolutional layer is made up of a number of

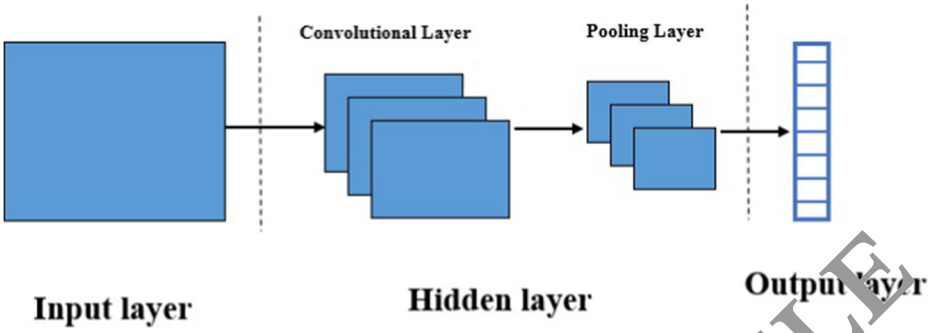

Fig. 1 The diagrams of the basic CNN [4]

neuronal maps that are either referred to as "filter maps" or "characteristic maps." A quantification of the discrete convolution of receivers may interpret the neural reactivity. The quantification process entails computing the overall neural weights of the input as well as the activation function assignments. Figure 2 depicts the structure of a typical discrete convolutional layer.

4.3.2 The max pooling layer

The max pooling layer generates a large number of meshes from the output of the segmented convolutional layer. The maximum grid value is used to create matrices in sequence [4]. The operators are used to get the average or maximum value for each matrix. Figure 3 depicts the building of the greatest pooling layer.

4.3.3 The full connection layer

This layer refers to the full CNN, which contains 90% of the overall structural components. This layer enables the input to be transmitted over the networks with the preconfigured vector length [6]. The data is transformed by a layer in this network before it is graded. The convolutionary layer has also been transformed to preserve the integrity of the information.

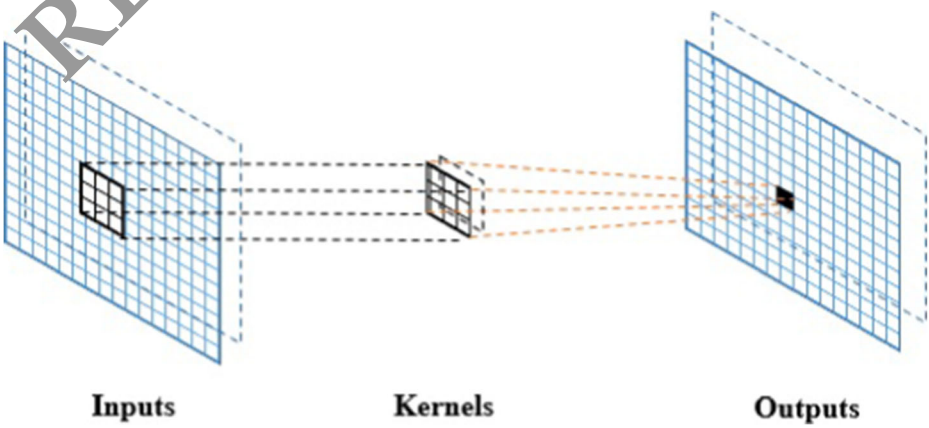

Fig. 2 The convolutional layer [37]

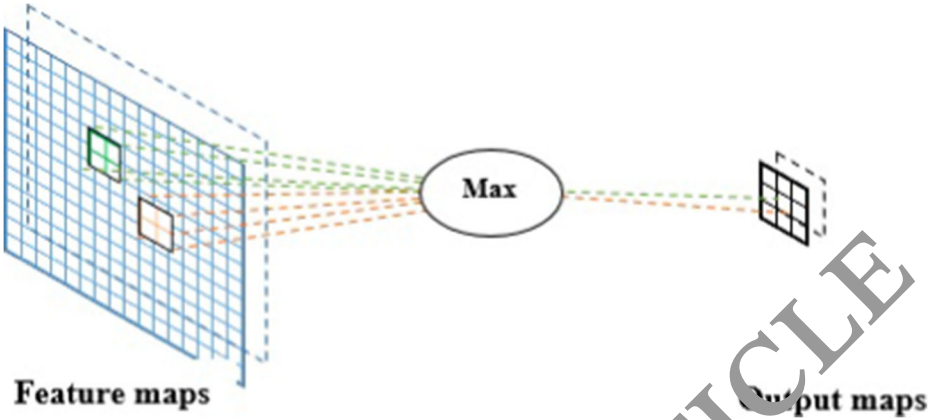

Fig. 3 The max pooling layer [37]

For complete connection layers, the neurons of every previous layer are used. These fully connected layers serve as the network's last layer and are categorized. Figure 4 depicts the whole connection layer configuration. Figure 5 depicts a typical complete CNN with all three layers. It should be mentioned that the conventional CNN design described here may not be the ideal choice for solving the CV problem because it was developed for object recognition. To optimize performance, a bespoke network structure must be created to adapt to the issue area. However, the experimental findings suggest that the developed CNN is capable of achieving the needed performance.

5 Paradigm of proposed method for image description

The proposed method for image description is based on transforming the image into a vector. Figure 6 displays the CNN framework for various image captions. Recent successes in machine translation learning of visual descriptions introduced the global visual characteristic. First, the vector encodes the same raw image and displays the whole semantic image data

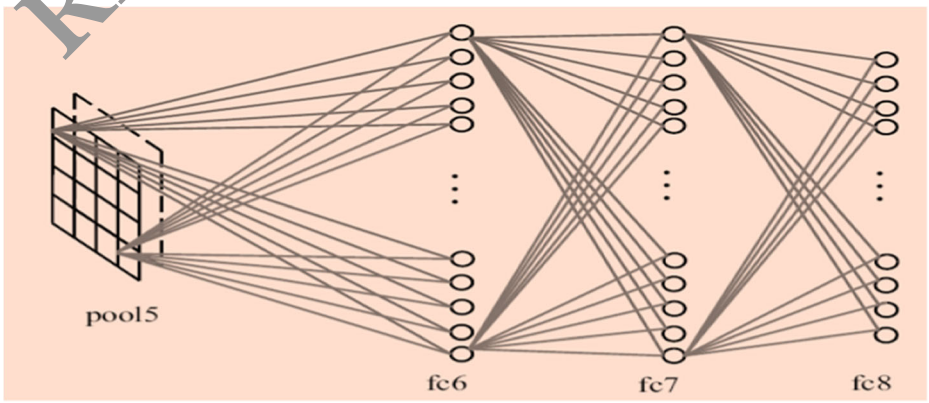

Fig. 4 The full connection layers [6]

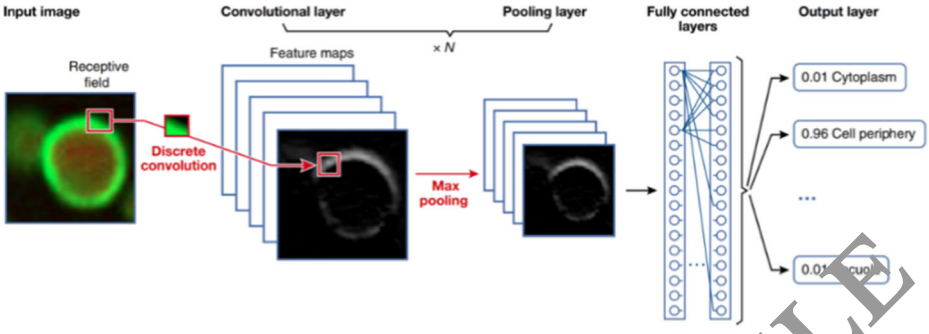

Fig. 5 The architecture for a complete CNN

using deep CNNs. The CNN is fully connected and contains many convolutionary, maximal bundling, normalized responses (Fig. 7). The design was highly successful for large-scale imaging classifications [24], and the know-how has been transferred to a wide range of vision tasks [57]. Usually, the activation values in the latter, fully connected layer, as the overall visual function vector, are retrieved in raw images.

Several studies have been conducted using CNN models based on linguistic architectures [8, 13, 25, 27, 36, 48, 49, 56]. Recently, a study was conducted to understand the mechanism of image captioning [54]. Figure 8 shows the attention architecture wherein CNN used a series of visual vectors for the sub-region images and a global visual vector. The CNN is able to eliminate these vectors from a lower convolutional layer [34]. The CNN refers to those sub-region vectors at every step of the language development process and determines the possibility of every sub-region's relevancy to the existing word production states. Finally, the attention architecture creates a context vector by combining the sub-region and relevant vectors for decoding the following CNN words. In another work, a module was added to improve the attention mechanism, and a method

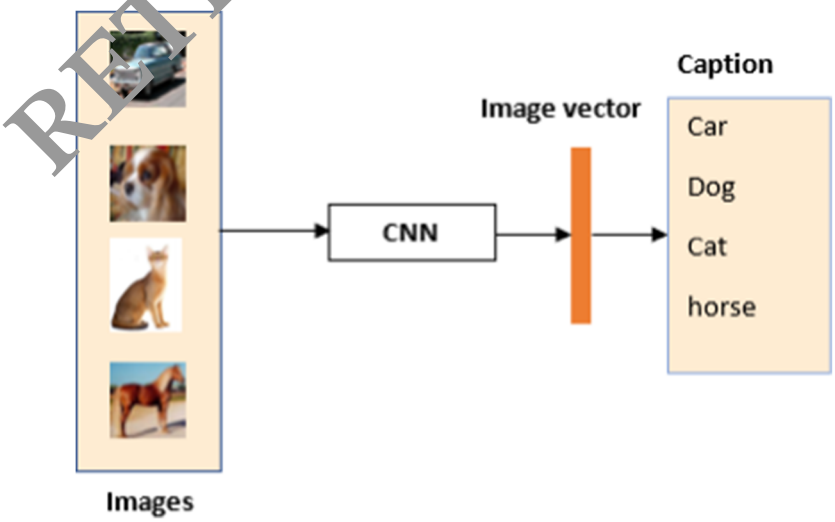

Fig. 6 CNN framework for image captions

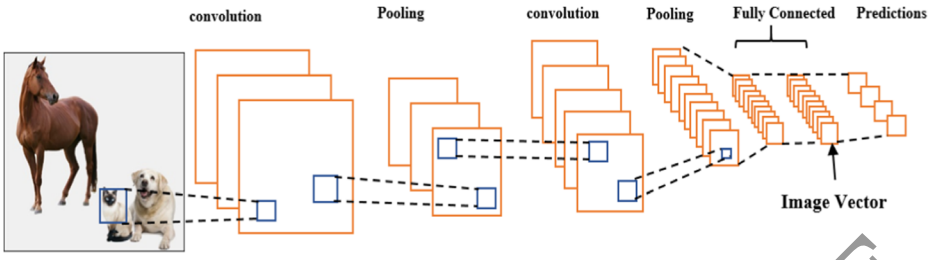

Fig. 7 Deep CNN structure showing overall visual feature vector, with second last dense layer representing the semantic information content for the entire image

was proposed to enhance the accuracy of seeing [35, 56]. In addition, a bottom-up attention model was introduced that showed the most advanced photographic subtitling performance based on object recognition [2]. The end-to-end formats could include the back-to-back names and every parameter of the CNN model.

6 Proposed CNN architecture for learning representation

Based on the abovementioned facts, various CNN-deep learning (CNN-DL) model architectures were examined to see their accuracy in image captioning. The CNN model was configured after the data collection and feature extraction. Convolutional architectures with totally connected layers were considered the default structural design. These architectural designs were appropriate for dealing with the image datasets in high- and multi-dimensional formats like 2D images or genomic data. In order to assess the improvements caused by the increased CNN depth, Krizhevsky principles were used to design the proposed CNN layer configurations [28]. Representation learning consisted of learning representative data characteristics that simplified the extraction of valuable information for future learning tasks [7, 39]. The remarkable achievement of DL has led to immense improvements in the representations learned by deep neural networks (DNNs) over the hand-made functions used in most of the learning tasks [20, 53].

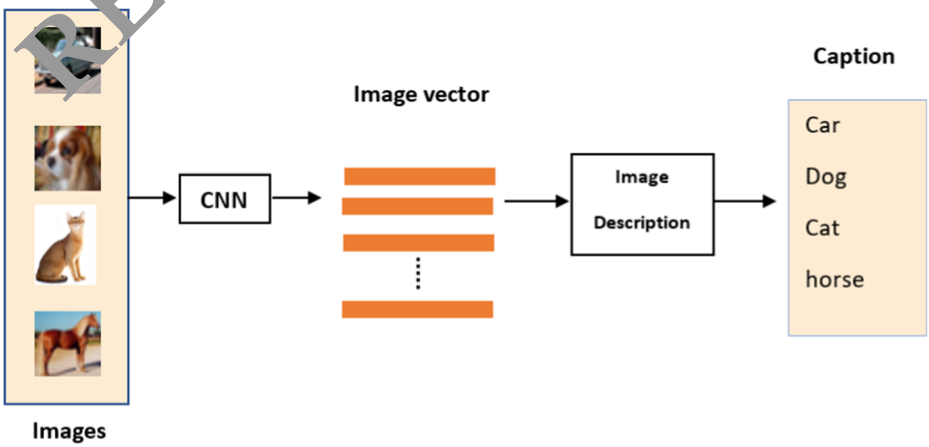

Fig. 8 Representation of the method for generating image description

Layered architectures that learn various functionalities at different levels are profound learning models. These hierarchically layered feature representations can eventually be linked to the end layer during the classification (generally a completely connected layer) to produce the final results. For example, the use of the Krizhevsky CNN configurational model in the absence of its last categorization layer enables the conversion of the object into a novel task area hidden in the state-based n-dimensional vectors (nodes in the last hidden layer) [18]. It is the most commonly used method for learning transmission across DNNs. Figure 9 displays the new learning method using Krizhevsky CNN as the feature extractor.

Multiple parameters across several layers of CNN's coding were fine-tuned during training. The fully connected layer's convolution filters, decision tree nodes, and hidden neurons were constantly adjusted to the data. Figures 10 and 11 depict the proposed CNN parameters and structure configuration, respectively.

7 Performance evaluation

The evaluation criteria are the primary components for determining the robustness of any classification method. It serves as a guide for developing and improving the classification models. Table 4 shows all the measurements that are derived from four factor values, such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [4]. The most common classification measurements are the true positive rate (TPR, which includes recall, detection, and sensitivity), the correct classification rate (CCR), the false positive rate (FPR), the false negative rate (FNR), and the true negative rate (TNR), which includes specificity and precision. For the performance evaluation of the proposed CNN-DL-based image-to-vector depiction model, we used measures like CCR, FPR, TPR, Precision, and FP or 1-Precision.

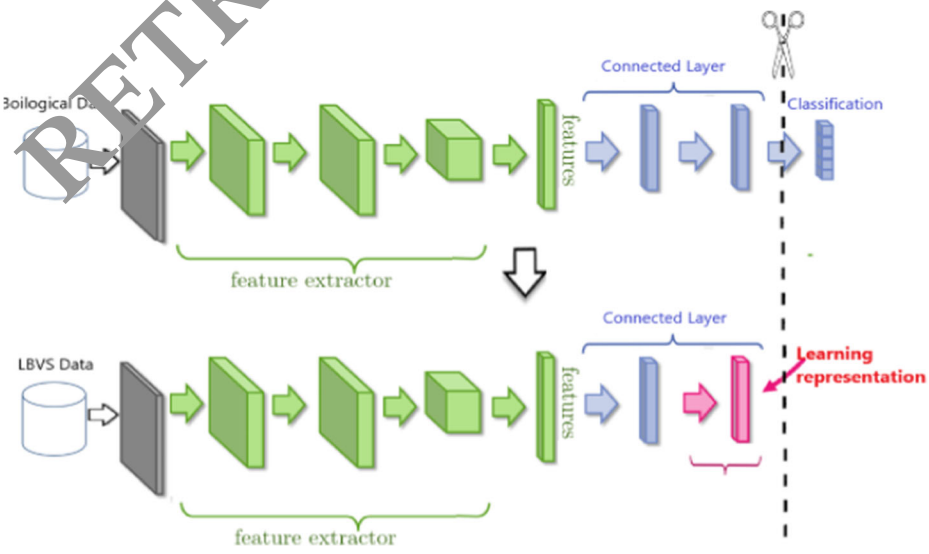

Fig. 9 Learning representation of the proposed CNN model

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 254, 254, 64)	640
conv2d_2 (Conv2D)	(None, 252, 252, 32)	18464
max_pooling2d_1 (MaxPooling2	(None, 126, 126, 32)	0
dropout_1 (Dropout)	(None, 126, 126, 32)	0
flatten_1 (Flatten)	(None, 508032)	0
dense_1 (Dense)	(None, 128)	65[illegible]
dropout_2 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 10)	[illegible]90

Total params: 65,048,618
Trainable params: 65,048,618
Non-trainable params: 0

Fig. 10 The summary of the proposed CNN parameters

The expression of CCR (defined as the percentage of patterns correctly classified) is given by:

$$CCR = \frac{T_P + T_N}{\text{Total number of patterns}}$$

The expression TPR also called the Detection Rate, Recall, or Sensitivity (defined as the percentage of positive pattern correctly classified as belonging to the positive class) is can be written as:

$$TPR = \frac{T_P}{T_P + F_N}$$

The expression for FPR (defined as the percentage of negative patterns identified wrongly as positive) yields:

$$FPR = \frac{F_P}{F_P + T_N}$$

The expression for TNR (defined as the proportion of negatives properly identified as negative classes) is given by:

$$TNR = \frac{T_N}{T_N + F_P}$$

The expression for FNR (defined as percentage of positive patterns incorrectly classified as belonging to the negative class) can be written as:

$$FNR = \frac{F_N}{F_N + T_P}$$

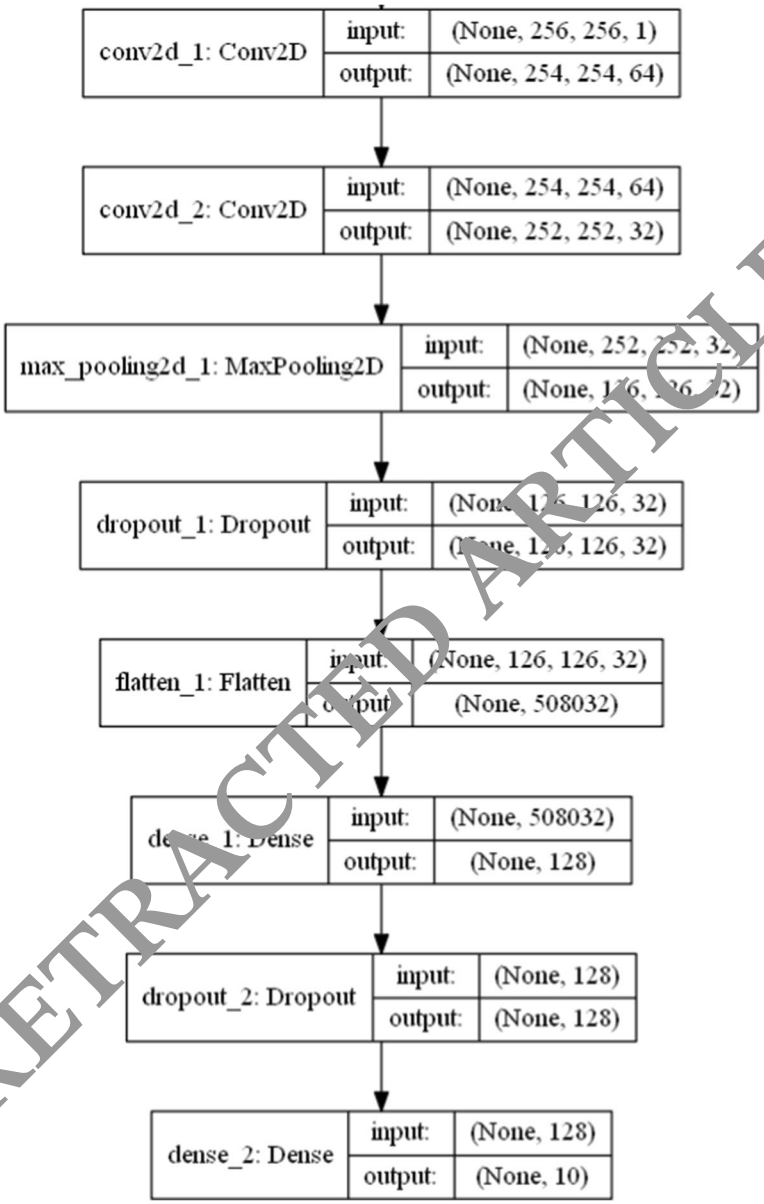

Fig. 11 The proposed CNN Model structure configuration

Table 4 Definition of the measuring parameters

Parameter	Definition
T_P	Pattern correctly classified as positive
F_N	Pattern incorrectly classified as negative
F_P	Pattern incorrectly classified as positive
T_N	Pattern correctly classified as negative

The expression for Precision (defined as the ratio of the number of properly categorized positive instances to the total number of positive instances) is written as:

$$\text{Precision} = \frac{T_P}{T_P + F_P}$$

The expression for Recall (that measures the number of positive class forecasts from all positive data instances) can be written as:

$$\text{Recall} = \frac{T_P}{T_P + F_N}$$

The F-Measure (F1, defined as a single score that balances the precision concerns and recalls them in a single number) is given by:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The Matthews Correlation Coefficient (MCC) is a contingency matrix method of determining Pearson product-moment correlation between actual and anticipated values, which is an alternative measure that is not influenced by imbalanced datasets. The equation of MCC is as follows:

$$MCC = \frac{(T_P \times T_N) - (F_P \times F_N)}{\sqrt{(T_P + F_P) \times (T_P + F_N) \times (T_N + F_p) \times (T_N + F_N)}}$$

8 The results and comparison

The success of the representation education was tracked and discussed in terms of the theoretical advantages of the distributed and profound representations, concluding with the broader idea of the underlying assumptions regarding the data generation process and causes of the observed data. Depending on the representation of the information, many data processing jobs might be either easy or very hard. It is a broad concept that applies to daily life in general and computer science in particular. One may consider advanced networks that are taught as representation learning via supervised learning. Particularly, a linear classifier, being usually the last layer on the network, can be represented by the remainder of the network. Training under a supervised criterion naturally results in the display of characteristics that simplify the classification job on any hidden layer (but closer to the top hidden layer). For example, the last hidden layer that is not linearly detached from the input characteristics becomes linear. In fact, in principle, the last layer can be another type of model (like the closest classification of neighbor), as shown in Table 5. Functions in the pre-last layer should learn different characteristics based on the last layer type, as depicted in Fig. 12.

To demonstrate the validity of the proposed method of image description, the image classification was used to calculate the image description's accuracy based on the proposed model's findings. By building a new model based on an improved CNN algorithm, image descriptions begin when the last connected layer is gathered before the classification layer to represent the image vector, as shown in Fig. 12. Table 5 displayed some of the image vector

Table 5 Images description for each object of the dataset and difference amid the descriptions

	0	0	0	0	83	52	71	0	0	0	111	0	0	0	80	0	0	113	0	0	69	0	102	0	0	63	51	112	0	99
	0	0	0	0	25	14	17	0	0	0	34	0	0	0	20	0	0	29	0	0	17	0	30	0	0	20	20	29	0	22
	0	0	0	0	79	46	71	0	0	0	116	0	0	0	84	0	0	99	0	0	69	0	104	0	0	71	49	125	0	91
	0	12	14	18	12	0	0	0	0	8	0	16	21	18	0	0	16	0	0	12	0	0	0	0	14	10	0	13	11	12
	0	21	14	20	12	0	0	0	0	14	0	23	26	20	0	0	17	0	0	13	0	0	0	0	18	8	0	15	16	19
	0	24	20	24	16	0	0	0	0	18	0	28	33	27	0	0	23	0	0	19	0	0	0	0	21	13	0	18	24	0
	0	11	16	19	0	12	0	20	0	10	24	0	0	0	0	0	0	0	0	13	19	0	19	25	0	30	19	24	19	0
	0	18	16	23	0	16	0	20	0	9	27	0	0	0	0	0	0	0	0	10	19	0	24	28	0	34	23	27	22	20	0
	0	18	21	22	0	17	0	26	0	15	33	0	0	0	0	0	0	0	0	17	26	0	23	0	0	27	24	29	25	0
	910	19	0	14	0	0	0	0	31	0	16	0	17	0	38	19	17	14	20	10	23	23	19	0	0	16	0	0	26	
	912	19	0	13	0	0	0	0	23	0	9	0	16	0	28	15	17	14	18	13	12	12	26	0	0	11	0	0	24	
	924	16	0	10	0	0	0	0	19	0	14	0	10	0	25	17	15	9	14	8	19	15	22	20	0	0	8	0	0	17
	342	13	11	0	13	0	12	0	18	13	0	0	13	15	21	9	0	0	4	0	0	0	11	0	5	19	9	12	0	0
	367	13	9	0	17	0	15	0	19	13	0	0	14	18	25	9	0	0	7	10	0	0	10	0	8	21	14	12	0	0
	370	13	11	0	16	0	15	0	23	15	0	0	14	18	23	8	0	0	7	12	0	0	10	0	8	25	15	16	0	0

outcomes. Thereafter, the evaluation of the testing results is based on several measurement performances in recent studies [38]. Table 6 showed the calculation of the accuracy, precision, recall, and F-score percentage that were obtained through testing the classification model on the CIFAR10 dataset.

For the performance evaluation, the proposed CNN-DL model configuration used the R, G, and B color channels as input features representing objects selected from the CIFAR-10 dataset. The profound learning models, as previously stated, were layered architectures that learned various functionalities at different levels. Finally, in the absence of its final taxonomy layer, such layers were linked to the final layer to produce the eventual outcomes. This in turn allowed the transformation of the object into an innovative task domain, the hidden-states based n-dimensional vector. Consequently, the features were extracted from an object

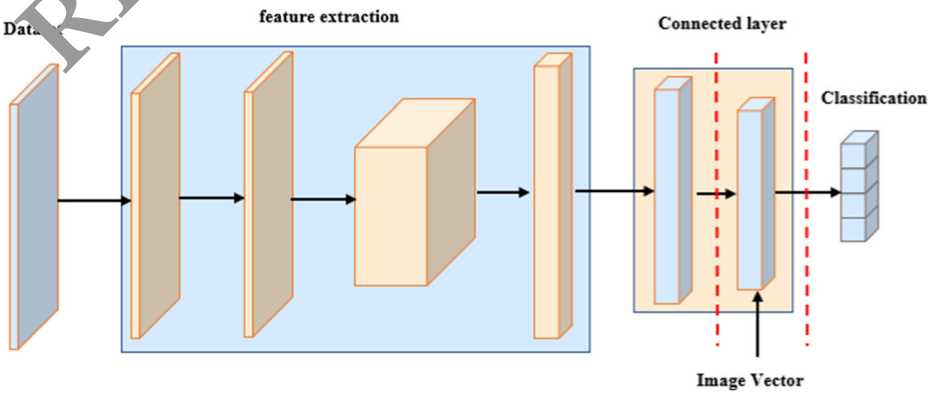

Fig. 12 Image description extraction from the penultimate layer that contains all the important information of the object

Table 6 Classification evaluation of CIFAR 10 dataset

Classes	Test Dataset	TPR	FPR	TNR	FNR	CCR	Precision	Recall	F1
Plane	202	98.02	2.97	97.03	1.98	97.52	97.06	98.02	97.54
Car	212	97.30	3.96	96.04	2.70	96.70	96.43	97.30	96.86
Bird	199	98.97	2.94	97.06	1.03	97.99	96.97	98.97	97.96
Cat	192	98.89	2.94	97.06	1.11	97.92	96.74	98.89	97.80
Deer	199	95.96	4.00	96.00	4.04	95.98	95.96	95.96	95.96
Dog	185	97.65	2.00	98.00	2.35	97.84	97.65	97.65	97.65
Frog	207	94.29	7.84	92.16	5.71	93.24	92.52	94.2[illegible]	93.40
Horse	202	98.02	2.97	97.03	1.98	97.52	97.06	98.02	97.54
Boat	199	93.27	2.11	97.89	6.73	95.48	97.98	[illegible].27	95.57
Truck	203	99.02	1.98	98.02	0.98	98.52	98.06	99.[illegible]	98.54

classification assignment that utilized the information from the object detection task. Table 6 shows the evaluation result for the final layer as a classification. We have built our model depending on the proposed method.

In addition to improving accuracy, the proposed method reduced methodological complexity. We could use feature extraction as a fundamental of the CNN algorithm by examining the structure of the CNN. It can access hidden, connected layers if it wants to apply a learning representation and precisely describe visual features [45]. The proposed method's novel architecture allowed us to produce a new image description that is both more accurate and reliable than its predecessors and simpler to implement. Previous research employed a different strategy to obtain feature extraction, leading to less precise results and more work [5, 14, 23, 44]. Using computer vision and CNN to classify images has limits. These techniques may not work in other circumstances if the dataset is unavailable. This study aims to develop a CNN

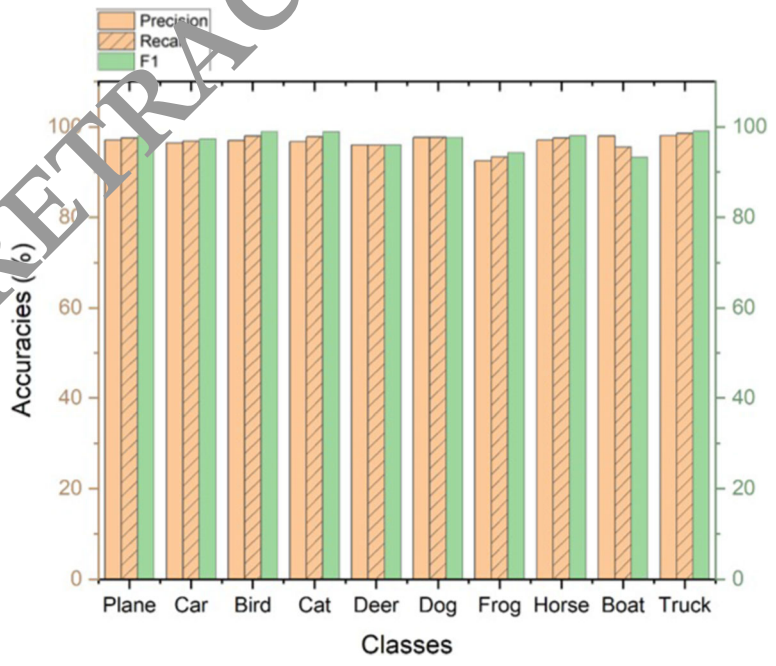

Fig. 13 Demonstrates the main finding of our proposed method

Table 7 Comparative evaluation of the proposed method

References	Algorithm	Accuracy	Precision	Recall	F1
[47]	CNN-GA	96.78	–	–	–
[42]	CNN	96.81			
[3]	CNN	–	96.77	–	92.3
	Proposed method	96.871	96.643	97.139	96.882

model for image classification and IV to improve prior work. Our research focuses on high-accuracy image classification and description generation with less complexity. However, our proposed method's limitation was the MCC, which our suggested technique reached at 93.87%.

The first step in describing an image using the new model based on the enhanced CNN algorithm is to collect the last connected layer before the classification layer, which represents the image vector. Next, some of the image vector outcomes are shown in Table 5. After that, several measurement performances from recent studies are used to evaluate the test results [38]. Finally, the calculated accuracy, precision, recall, and F-score percentage from testing the classification model on the CIFAR10 dataset are displayed in Fig. 13. Furthermore, based on the main findings, we compared the proposed method with the state-of-the-art techniques. Table 7 compares the results of the proposed method to those of state-of-the-art techniques.

9 Conclusion

This paper proposes an enhanced CNN-DL approach to describe the image contents in the natural language in which the images were transformed to vectors, recognizing the cross-disciplinary value of precise images to text production in computer vision and the processing of natural languages. First, it focused on picture classification and provided a new strategy for using CNN to create a classification model. Second, based on the classification's success, we propose a new description method—an image to victor—to characterise each object in the image. The performance of the developed model was trained on the COCO dataset and evaluated using CIFAR10. Besides, it provided the technical basis for other significant applications. In addition, a convolutional neural network (CNN) and a deep learning (CNN-DL technique were implemented to convert images into vectors and describe their contents in plain English. The major advancements in DL research and industrial deployment by the community and their impacts were examined. Image sub-sections being a critical area for the multimodal intelligence image-natural language, a new strategy for training the CNN architecture that could remove the locally matched visual descriptors was proposed, with NNs-based profound training playing a significant role. As a result, using the CIFAR dataset, a newly built system performed better than reported methods for processing test images from a distinct and isolated field. The experimental results demonstrated more detailed descriptions of the image contents using the principles that have been introduced in the field of distance metrics, which were stimulated through the training with positive and negative constraints simultaneously. The empirical outcomes of the model with the cross-domain picture datasets reaffirmed its high flexibility, reliability, and stability when compared with other state-of-the-art techniques reported in the literature. It was established that the present approach may contribute to the future development of multimodal intelligence related to AI capabilities.

Acknowledgments Authors are extremely thankful to Universiti Teknologi Malaysia (UTM), Ministry of Higher Education Malaysia (MOHE), and RMC for research grant FRGS Q.J130000.2509.21H11, and UTM RA ICONIC GRANT Q.J130000.4354.09G60, FRGS-04E86 and UTMFR 21H78. Authors are also grateful to Research Management Centre-Universiti Teknologi Malaysia (RMC-UTM) for supporting under Postdoctoral fellowship scheme.

Declarations

Conflict of interest Please check the following as appropriate:
All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.
The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

References

1. Adnan MM, Rahim MSM, Rehman A, Mehmood Z, Saba T, Naqvi RA (2021) Automatic image annotation based on deep learning models: a systematic review and future challenges. IEEE Access 9:50253–50264
2. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086)
3. Ayadi W, Elhamzi W, Charfi I, Atri M (2021) Deep CNN for brain tumor classification. Neural Process Lett 53(1):671–700
4. Banerjee S, Lavie A (2005) METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65-72)
5. Benyahia S, Meftah B, Lézoray O (2022) Multi-features extraction based on deep learning for skin lesion classification. Tissue Cell 74:101701
6. Bianchini M, Scarselli F (2014) On the complexity of neural network classifiers: A comparison between shallow and deep architectures. IEEE TransacNeural Netw Learn Syst 25(8):1553–1565
7. Bullins B, Hazan E, Kalai A, Livni R (2019) Generalize across tasks: efficient algorithms for linear representation learning. In *algorithmic learning theory* (pp. 235-246). PMLR
8. Chen X, Lawrence Zitnick C (2015) Mind's eye: A recurrent visual representation for image caption generation. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2422-2431)
9. Chen Y, Liu L, Tao J, Chen X, Xia R, Zhang Q, Xie J (2021) The image annotation algorithm using convolutional features from intermediate layer of deep learning. Multimed Tools Appl 80(3):4237–4261
10. Chun PJ, Yamane T, Maemura Y (2022) A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage. Comput-Aided Civil Infrastruc Eng 37(11):1387–1401
11. Dahl GE, Yu D, Deng L, Acero A (2011) Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang Process 20(1):30–42
12. Deng L, Yu D (2014) Deep learning: methods and applications. Foundations Trends® Sig Proc 7(3–4):197–387
13. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2625-2634)
14. El-Komy A, Shahin OR, Abd El-Aziz RM, Taloba AI (2022) Integration of computer vision and natural language processing in multimedia robotics application. Inform Sci Lett 11(3):9
15. Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, Socher R (2021) Deep learning-enabled medical computer vision. NPJ Digital Med 4(1):1–9
16. Gan Z, Gan C, He X, Pu Y, Tran K, Gao J, Deng L (2017) Semantic compositional networks for visual captioning. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5630-5639)

17. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT press
18. He X, Deng L (2017) Deep learning for image-to-text generation: A technical overview. IEEE Signal Process Mag 34(6):109–116
19. He X, Deng L (2018) Deep learning in natural language generation from images. In deep learning in natural language processing (pp. 289–307). Springer, Singapore
20. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778)
21. Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Proc Mag 29(6):82–97
22. Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: data, models and evaluation metrics. J Artif Intell Res 47:853–899
23. Idicula SM (2019) Dense model for automatic image description generation with game theoretic optimization. Information 10(11):354
24. Jena B, Saxena S, Nayak GK, Saba L, Sharma N, Suri JS (2021) Artificial intelligence-based hybrid deep learning models for image classification: the first narrative review. Comput Biol Med 137:104803
25. Jia X, Gavves E, Fernando B, Tuytelaars T (2015) Guiding the long-short term memory model for image caption generation. In proceedings of the IEEE international conference on computer vision (pp. 2407-2415)
26. Kadhim KA, Adnan MM, Waheed SR, Alkhayyat A (2021) Automated high-security license plate recognition system. Materials Today: Proceedings, WITHDRAWN: Automated high-security license plate recognition system
27. Kiros R, Salakhutdinov R, Zemel R (2014) Multimodal neural language models. In international conference on machine learning (pp. 595-603). PMLR
28. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
29. Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y, Berg TL (2013) Babytalk: understanding and generating simple image descriptions. IEEE Trans Pattern Anal Mach Intell 35(12):2891–2903
30. Li S, Kulkarni G, Berg T, Berg A, Choi Y (2011) Composing simple image descriptions using web-scale n-grams. In proceedings of the fifteenth conference on computational natural language learning (pp. 220-228)
31. Li S, Xiao T, Li H, Zhou B, Yue D, Wang X (2017) Person search with natural language description. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1970-1979)
32. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Zitnick CL (2014) Microsoft coco: common objects in context. In European conference on computer vision (pp. 740–755). Springer, Cham
33. Lin K, Li D, He X, Zhang Z, Sun MT (2017) Adversarial ranking for language generation. Adv Neural Inf Proces Syst 30
34. Liu Y, An X (2017) A classification model for the prostate cancer based on deep learning. In 2017 10th international congress on image and signal processing, BioMedical engineering and informatics (CISP-BMEI) (pp. 1-6). IEEE
35. Liu C, Mao J, Sha F, Yuille A (2017) Attention correctness in neural image captioning. In Thirty-first AAAI conference on artificial intelligence, Attention Correctness in Neural Image Captioning
36. Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A (2014) Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632
37. McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5(4):115–133
38. Najar FH, Al-Jawahry HM, Al-Khaffaf MS, Al-Hasani AT (2021) A novel hybrid feature extraction method using LTP, TFCM, and GLCM. In *journal of physics: conference series* (Vol. 1892, no. 1, p. 012018). IOP publishing
39. O'Connor P, Neil D, Liu SC, Delbruck T, Pfeiffer M (2013) Real-time classification and sensor fusion with a spiking deep belief network. Front Neurosci 7:178
40. Ordonez V, Kulkarni G, Berg T (2011) Im2text: describing images using 1 million captioned photographs. Advances in neural information processing systems, 24
41. Piasco N, Sidibé D, Gouet-Brunet V, Demonceaux C (2021) Improving image description with auxiliary modality for visual localization in challenging conditions. Int J Comput Vis 129(1):185–202
42. Qin J, Pan W, Xiang X, Tan Y, Hou G (2020) A biological image classification method based on improved CNN. Ecolog Inform 58:101093
43. Rashtchian C, Young P, Hodosh M, Hockenmaier J (2010) Collecting image annotations using amazon's mechanical turk. In proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical Turk (pp. 139-147)
44. Shao H, Lin J, Zhang L, Galar D, Kumar U (2021) A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance. Inform Fusion 74:65–76

45. Sharma H, Jalal AS (2022) Image captioning improved visual question answering. Multimed Tools Appl 81(24):34775–34796
46. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
47. Sun Y, Xue B, Zhang M, Yen GG, Lv J (2020) Automatically designing CNN architectures using the genetic algorithm for image classification. IEEE Transac Cybernet 50(9):3840–3854
48. Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T, Saenko K (2015) Sequence to sequence-video to text. In proceedings of the IEEE international conference on computer vision (pp. 4534-4542)
49. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164)
50. Waheed SR, Alkawaz MH, Rehman A, Almazyad AS, Saba T (2016) Multifocus watermarking approach based on discrete cosine transform. Microsc Res Tech 79(5):431–437
51. Waheed SR, Suaib NM, Rahim MSM, Adnan MM, Salim AA (2021) Deep learning algorithms-based object detection and localization revisited. In journal of physics: conference series (Vol. 1892, no. 1, p. 012001). IOP publishing
52. Wang H, Meghawat A, Morency LP, Xing EP (2016) Select-additive learning: improving cross-individual generalization in multimodal sentiment analysis. arXiv preprint arXiv:1609.05244
53. Wu FX, Li M (2019) Deep learning for biological/clinical data. Neurocomputing 324:1–2
54. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In international conference on machine learning (pp. 2048-2057). PMLR
55. Xu S, Wang J, Shou W, Ngo T, Sadick AM, Wang X (2021) Computer vision techniques in construction: a critical review. Arch Computa Meth Eng 28(5):3383–3397
56. Yang Z, Yuan Y, Wu Y, Cohen WW, Salakhutdinov RR (2016) Review networks for caption generation. Adv Neural Inf Proces Syst 29
57. Yao K, Peng B, Zhang Y, Yu D, Zweig G, Shi Y (2014) Spoken language understanding using long short-term memory neural networks. In 2014 IEEE spoken language technology workshop (SLT) (pp. 189-194). IEEE
58. Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Transac Assoc Comput Linguist 2:67–78
59. Yu L, Zhang W, Wang J, Yu Y (2017) Seqgan: sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence, 31 (1)