IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 29, 2020, accepted March 12, 2021, date of publication March 24, 2021, date of current version April 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068897

Automatic Image Annotation Based on Deep Learning Models: A Systematic Review and Future Challenges

MYASAR MUNDHER ADNAN[1,2], MOHD SHAFRY MOHD RAHIM[3], AMJAD REHMAN[4], (Senior Member, IEEE), ZAHID MEHMOOD[5], TANZILA SABA[4], (Senior Member, IEEE), AND RIZWAN ALI NAQVI[6], (Member, IEEE)

[1]Faculty of Engineering, School of Computing, University Teknologi Malaysia, Johor Bahru 81310, Malaysia
[2]Islamic University, Najaf 202001, Iraq
[3]School of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[4]Artificial Intelligence and Data Analytics Laboratory, College of Computer and Information Sciences (CCIS), Prince Sultan University, Riyadh 11586, Saudi Arabia
[5]Department of Computer Engineering, University of Engineering and Technology, Taxila, Taxila 47050, Pakistan
[6]Department of Unmanned Vehicle Engineering, Sejong University, Seoul 05006, South Korea

Corresponding authors: Zahid Mehmood (zahid.mehmood@uettaxila.edu.pk) and Rizwan Ali Naqvi (rizwanali@sejong.ac.kr)

ABSTRACT Recently, much attention has been given to image annotation due to the massive increase in image data volume. One of the image retrieval methods which guarantees the retrieval of images in the same way as texts are automatic image annotation (AIA). Consequently, numerous studies have been conducted on AIA, particularly on the classification-based and probabilistic modeling techniques. Several image annotation techniques that performed reasonably on standard datasets have been developed over the last decade. In this paper, a review of the image annotation method was conducted, focusing more on deep learning models. Automatic image annotation (AIA) methods were also classified into five categories, including i) Convolutional Neural Network (CNN) based on AIA, ii) Recurrent Neural Network (RNN) based on AIA, iii) Deep Neural Networks (DNN) based on AIA, iv) Long-Short-Term Memory (LSTM) based on AIA, and v) Stacked auto-encoder (SAE) based on AIA. An assessment of the five varieties of AIA methods was also offered based on their principal notion, feature mining technique, explanation precision, computational density, and examined aggregated data. Moreover, the evaluation metrics used to evaluate AIA methods were reviewed and discussed. The need for careful consideration of methods throughout the improvement of novel procedures and datasets for image annotation assignment was highly demanded. From the analysis of the achievements so far, it is certain that more attention should be paid to automatic image annotation.

INDEX TERMS Automatic image annotation (AIA), deep learning, feature's extraction, digital learning.

I. INTRODUCTION

The progressively cumulative volume of ordinal images and the need to meet the users' requirements for gigantic data volumes have necessitated an accurate and efficient image retrieval technology. One of the image retrieval methods which guarantees the retrieval of images in the same way as texts are automatic image annotation (AIA). According to Barnard *et al.* [1], AIA is an important problem in computer vision. As images often contain complex and different kinds of content information, query, retrieve, and organize image information quickly and effectively becomes a crucial issue. AIA can be applied in various fields, including online/offline data exploration, image manipulation, and annotation application used in mobile gadgets [2]–[4]. In a typical image annotation system, two things are significant; (i) a semantic appreciation of ordinal images and (ii) a natural language processing (NLP) unit which will interpret the images' semantic data into an output that a human can read. Various methods have recently been proposed on AIA systems, giving rise to several AIA algorithms. These methods contain the practice of texture resemblance, Support Vector Machines, Bayesian,

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.

and Instance-based methods [1]–[6]. However, deep learning techniques have, over the last decade, performed excellently in image processing. Visual attention has also been successfully deployed with deep neural networks in many NLP and computer revelation methods. Its usage for image annotation has also been reported in several studies [7]–[10]. Despite the prevailing deep learning-based methods to improve AIA frameworks' enactment, AIA is still prone to numerous key challenges. Among these challenges is its requirement of a huge data volume to perform an accurate prediction. The other two major challenges of AIA are the management of imbalanced keywords distribution, as well as the selection of appropriate features. Previous works on AIA have developed several deep learning procedures, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Deep Neural Networks (DNN) to address these issues. However, inspired by the cranial nerve theory, DNN has started to become widely used in the arena of computer vision, NLP, and so on. In 1943 [11], Warren McCulloch and Pitts proposed and presented the Artificial Neural Network (ANN) concept and the mathematical model of artificial neurons, which is considered the foundation for the theory of neurons in biology and physiology. The milestone in ANN research is the invention of the Backpropagation algorithm (BP) [12]. The ANN is closer to the human brain in structure, principle, and function. It can adapt to the environment itself, summarize laws, perform some operations, identification, or process control. It was not until 2012 that the ANN became popular due to Deep Convolutional Neural Network's realization in image classification [13], [14]. Although deep learning methods can efficiently handle huge data, their efficiency usually decreases with increases in the model's complexity and scope. Additionally, for outsized-scale datasets, many systems do not ruminate the unique labels of the datasets. This work focuses on five categories of deep learning techniques based on AIA; these categories are CNN based on AIA, RNN based on AIA, DNN based on AIA, Long-Short-Term Memory (LSTM) based on AIA, and Stacked Auto-Encoder (SAE) built on AIA. Furthermore, an analysis and comparison of these AIA methods were performed based on their basic concepts, the main contribution, the annotation accuracy, and the computational complexity. The remaining part of this work is organized; thus: Section 2 discussed image segmentation and feature extraction, while section 3 described various deep learning-based AIA techniques. Section 4 presented the summary and conclusion of the review.

II. IDEEP LEARNING FOR IMAGE ANNOTATION

Deep learning is responsible for the dramatic advancements in state-of-the-art AI-based research such as speech recognition, entity recognition, and machine translation. Deep learning could be used to solve numerous complicated AI tasks due to its deep architecture [15]. Consequently, deep learning is currently extended to several modern tasks and domains; this is in addition to conventional errands such as surface acknowledgment, etymological prototypes, or object discovery. For instance, the study by [16] reported the use of a recurrent neural network (RNN) to de-noise speech signals. In contrast, the discovery of gene expression and clustering patterns using SAE has been reported [17]. Another study by [17] generates images with different styles using a neural model, while [18] depended on deep learning to permit simultaneous sentiment analysis from numerous modalities. This era will experience a boost in deep learning-related studies. Deep learning does implement better than other machine learning procedures as the pragmatic outcomes recommend. Some have proposed that it is for the reason that deep learning can roughly impersonate the brain's purposes (numerous deposits of neural networks arranged one after an additional like the conventional brain prototypical). Nevertheless, there is no vigorous speculative context for deep learning [5]. Usually, deep learning technologies execute superior to the predictable ML implementations due to their training on the feature extraction part. With deep learning methods, feature hierarchies are learned so that features from the higher hierarchy are formed by compiling features from the subordinate hierarchy. The automatic learning of features at manifold abstraction levels will tolerate a method to acquire intricate functions; it will help the system directly map the input to the output without depending on the human-crafted features [5]. For instance, during image recognition, the normal system is to remove/fetch and feed the programmed features to SVM. In the deep learning schemes, the extracted features are also optimized, and this is why deep learning methods perform better Figure 1 describes the reasoning for the use of deep learning. Deep learning differed from traditional ML in its performance as the volume of data increases. With small data, deep learning performance is not nice because it necessitates enormous information to achieve perfect learning. Contrarily, traditional ML algorithms work better with small data owing to their handcrafted rules. This fact is summarized in the image below [18].

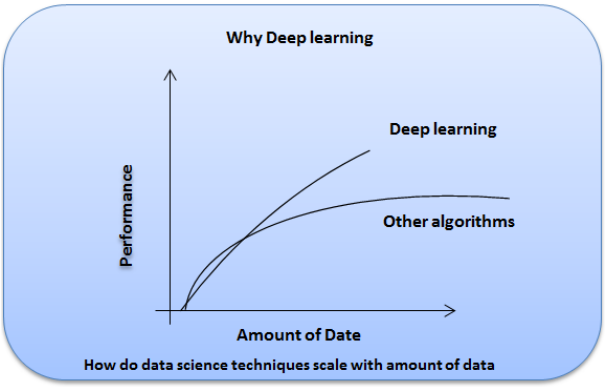

FIGURE 1. Why deep learning ?.

III. FEATURE EXTRACTION

This section discussed feature extraction (FE), an important step in an AIA model to convert raw images into features. There are two categories of image features; these are low-level and high-level image features. The low-level

image features such as shape, color, and texture are extracted via image processing, while the high-level features represent the words or concepts from an image. Furthermore, image features used in the existing AIA techniques could also be classified into the region and global image features; the region-based features require image segmentation while global image features are calculated from all the images. There are two representations of global image features; these are gist and color histograms [19]. Three color spaces (RGB, LAB, and HSV) are involved in the extraction of color histograms and these color spaces are the most utilized in computer vision. The local features can capture more semantic image contents compared to the comprehensive features. The *scale-invariant feature transform (SIFT)* and a vigorous type descriptor are normally implemented as two confined features. Both features were considered in this study to ensure appropriate image representation. A description of the structure of both features is presented as:

A. *Low-level image features:* These are a combination of features or autonomous entities in an image [9]. They provide a specific description of the images' components, such as the background, color, texture, or shape by concentrating on the basic micro details of images [15].
B. *High-level features:* These features are important image representation attributes as they represent the image from a global perspective and refers to the concept or definition of an image [9], [20]. These features can mimic the human perceptual system efficiently.

IV. IMAGE SEGMENTATION

In most studies, segmentation methods that depend mainly on the color space of an image are utilized. These methods are mainly used for the efficient local or global extraction of image visual features via image segmentation. For the global methods, a single set of features is computed from the entire image. In contrast, the local methods work by partitioning the images into blocks or regions before computing a set of features for each block. Thus, images can be represented with features at the object level and still provide spatial image information. However, the unsupervised segmentation associated with region features may affect their accuracy since segmentation performance is normally dependent on the applications. Among the popular algorithms for image, segmentation is grid-based techniques, clustering-based techniques, contour-based techniques, region growing-based techniques, and statistical model-based techniques [6]. The variance intra-cluster maximization method is one of the efficient image segmentation methods because, in grey-level images, it ensures the selection of a global threshold value by maximizing the separability of the classes [24].

V. AUTOMATIC IMAGE ANNOTATION METHODS USING DEEP LEARNING

In this section, a brief review of the deep learning methods for AIA was conducted. These methods are classified into A) Convolution neural network (CNN) based on AIA, B) Recurrent Neural Network (RNN) based on AIA, C) Deep neural network (DNN) based on AIA, D) Long-Short-Term Memory (LSTM) based on AIA, E) Stacked Auto-Encoder (SAE) based on AIA.

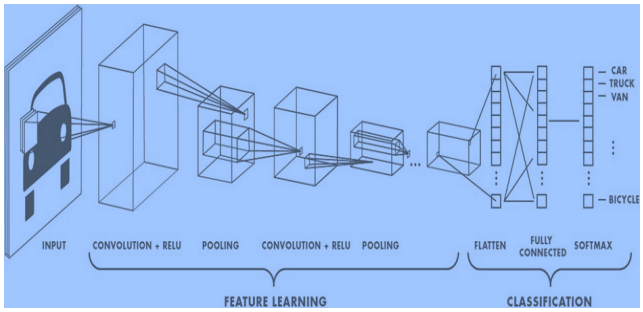

FIGURE 2. The general CNN configuration.

A. CONVOLUTION NEURAL NETWORK (CNN)

CNN or ConvNet represents a class of deep feed-forward ANNs that are mostly used in visual image analysis. It is a well-known DL architecture that got stimulated by the expected visual sensitivity tool of active things. Several CNN architecture variants exist in the literature, such as LeNet-5, Alexnet, VGG, GoogleNet, and Deep Residual Learning [16]. These variants of CNN are, however, similar in their basic components. For instance, the famous LeNet-5 comprises three basic layers (convolutional, pooling, & fully-connected) depicted in Figure 2. It describes the input feature representation learned by the convolutional layer. This layer consists of several convolution kernels that help in dissimilar characteristics maps computation. For individual neurons, its feature map is directly linked to a region of nearby neurons in the preceding layer (a region known as the neuron's receptive field in the preceding layer). The input will first be convolved with a trained kernel before smearing a component-wise nonlinear triggering function on the convolved outcomes to obtain the new feature map. It should be noted that before generating each feature map, all the inputs' spatial locations must share the same kernel. Different kernels are required to obtain the complete feature maps [15].

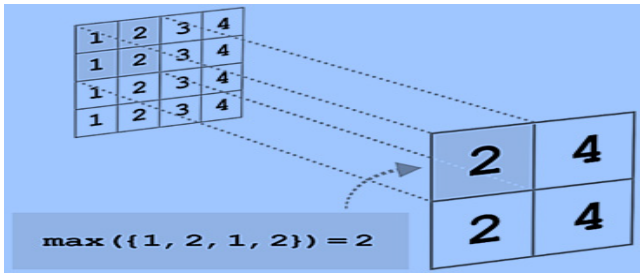

FIGURE 3. The max-pooling operation using 2 × 2 filters.

As presented in Figure 3, Max Pooling Layer commonly involves the periodic insertion of a pooling layer in-between

successive conventional layers in a ConvNet architecture. This layer's function ensures a reduced level of computation and the expanse of considerations in the network by progressively minimizing the representation's spatial size, thereby controlling overfitting. The pooling layer's operation is not dependent on the response's depth slice as it spatially resizes it via MAX operation. The commonest form of pooling layer is a pooling layer with 2×2 filters, which is applied with the progress of two down samples per complexity portion in the response by two along both width and height, leaving 75% of the activations [25].

The performance of numerous computer vision tasks has been improved by CNNs owing to their associated modeling complexity when learning from a huge volume of supervised data. Numerous models have been suggested for CNN-based AIA and retrieval; some of these models are discussed as follows. The combination of the CNN feature of an image with a semantic extension model (SEM) using the well-known CNN model-AlexNet has been proposed by Ma *et al.* [4]. The study extracted the CNN features by stripping its final layer; this proved to be a useful SEM model approach. The performance of the SEM was evaluated on corel5k [5], espGame [26], and Iaprtc12 [8], which are 3 publicly available standard image annotations datasets. However, this model is disadvantaged by the inadequacy in its precision due to the non-uniform tags distribution in the large dataset, making the model enlarge the prediction tags when searching for the neighbor group of an image, thereby causing a decrease in the precision. Another study by Wang *et al.* [5] modified the CNN model of CaffeNet to build a large-scale model called MVAIACNN for image annotation. In the proposed MVAIACNN, the layers are shallow. Each category is directly regarded as a label; it performs large-scale image annotation using raw images as inputs and depends on CNN for large-scale image annotation. The performance of the model was evaluated on the MIRFlickr25K and NUSWIDE datasets. To address a fixed number of labels appearing during the multilabel image annotation process and label annotation according to the ranking function. In [31] the application of a CNN-THOP for image annotation. First, a CNN model is used to predict the probability for each class of labels. Due to the VGG16 network architecture's merits, we improved the CNN structure in this study based on VGG16. A BN added within the CNN significantly accelerates the convergence speed, and the network structure and parameters are adjusted to make them more suitable for the datasets. In another approach, Luo *et al.* [27] suggested a novel CNN-based technique of annotating power grid objects' images. The images' attribute list is first obtained before building the image database for power grid objects. This database is comprised of a huge number of images in multilabel networks. Then, the image is annotated using an attribute-specific segmentation model. the accuracy of the proposed method was evaluated and found to be 94.83 %. Lin *et al.* [28] focused on the feature combination technique for image annotation and retrieval. This method utilizes low-level color features of the original images, while the extracted features are learned from CNN's. The progression of the projected technique is as follows: i) extract the color feature from the original images, build a visual lexicon, and use a bigram to present a co-occurrence relation, ii) input images into the CNN through five convolutional layers and let the pooling layer attain a visual feature, and iii) combine the two features at the first hidden layer of the DNN system. Mahmood *et al.* [29] suggested a DL and computer vision-based framework for automatic unlabelled coral images' annotation. This proposed framework depends on a novel coral classification framework that exploits the robust image representation capability of CNNs. Owing to the lack of basic truth labels for numerous coral reef images, the loop incorporates a human expert to validate the new method's accuracy. The trained coral classifier was used to analyze the coral reefs of the Abrolhos Islands, which is considered one of the unique marine areas in Western Australia. The unlabelled coral mosaics of 3 sites of this coral reef (covering two-year new method's accuracy increase in the method's performance was observed when Abrolhos Islands' performing all years. This indicates the challenges faced during the training phase; the test set images were collected over several years, which could affect the performance. This is mainly caused by the time-related changes occurring in the coral reefs. However, the major contributors to misclassification are the uncertain corals-non-corals boundaries. A study by Sarangi *et al.* [30] suggested a technique that performs image annotation using convex DL models such as Tensor Deep Stacking Networks (T-DSN) and Kernel Deep Convex Network (DCN). The study also proposed the use of features extracted with DCN as the input to the convex models. Observably, the convex models with DCN-extracted features as input provided the best performance. The extraction of the features becomes easier after training the convolutional network on a large image set. The convex networks' training time is short; this makes them ideal for image annotation tasks. Considering the K-CDN and T-DSN models, it was observed that, in each hidden layer, it is not beneficial to have different numbers of nodes. In this approach, no criterion for selecting a proTheer nodes networks' training times ended; however, it will be useful to determine the appropriate number of nodes and global optimum parameters for T-DSN and DCN, respectively. Simple and effective image annotation models that depend on CNN-extracted image categories and expression set in vectors to capture their allied labels were presented by Venkatesh *et al.* [11]. the versatility of the mid-level deep learning model's visual model, a method of extracting the mid-level convolutional characteristics, is developed and studied. Based on this, an image annotation method based on positive examples is proposed. The deep learning mid-level convolution feature extraction method used in the paper does scale dataset training model [37] model's visual meaning that the deep feature data volume and hardware costs are reduced. The first set of models that model the visual features and textual features of the data were based

on the Canonical Correlation Analysis (CCA) framework. The last layer of CaffeNet in the CNN-based model was substituted with a projection layer (for regression tasks); the subsequent network was then trained for semantic mapping of images' evocative word embedding vectors. There are two advantages of this type of modeling; i) numerous handcrafted features are not required; hence, metric learning or how to effectively combine those features is not important; ii) it is a relatively easier approach to formulate than the other discriminative or generative models. Additionally, when used in the earlier models, it improves the effectiveness of CNN-extracted features. An AIA approach that depends on DL models has been presented by Kashani *et al.* [32]. At first, the approach uses CNN models for feature extraction from the feedback image. The mined feature vectors are against all the training images; the most relevant tags are allocated to the participant image. This approach is a search-based annotation method that leverages deep architectures (i.e., CNN) as feature extractors. In this way, the proposed approach takes advantage of the search-based method and deep models simultaneously. CNN features are extracted from images using pre-trained models such as Caffe, VGG-16, VGG-19, and ResNet networks. Zhang *et al.* [10] presented a technique for feature extraction based on DL for annotating skin biopsy histopathological images. They also used CNN as a feature training model. The model depended on both nonlinear transformations of the original features and multiple-layer weighted combinations to learn the abstract features. For the generation of the annotation results, a supervised MIML learning model was placed on top of the DL model. However, some problems remain to be solved; the first one is the model's capability to execute only region-based administered learning. The uncertainty of the level at multiple-instances makes it impossible to propagate the model output loss through the network, leading to the inability to conduct a supervised fine-tuning of the network weights. The second problem relates to the scheme of the multiple-instance data sample CNN. This problem probes whether it is possible to build a CNN model that can scramble a manifold-instance model rather than just an occurrence. Rajchl *et al.* [34] suggested combining a NN model with an iterative graphical optimization technique. This combination aims to develop a model that can recover pixel-wise object segmentations with the associated bounding box annotations from an image database. This concept was based on the popular GrabCut [35] approach, which involves an iterative fitting of an intensity appearance model to a segment and later regularizes it to achieve segmentation. Similarly, the suggested DeepCut approach iteratively updates the training targets using a CNN model and regularizes the segmentation using a fully connected conditional random field (CRF). A generic form of this approach was formulated; thus, it can be easily applied to any image or object modality [40]. This article proposed an end-to-end deep learning framework for multi-label annotation of RS images that exploits dual-level semantic concepts. The framework includes a shared CNN for visual feature learning, a classification branch for multilabel annotation, and an embedding branch for maintaining the similarity relationships among the triplet images grouped by scene-level semantic concepts. An attention mechanism is introduced in the classification branch for salient object detection, while the skip connection is incorporated to combine information from multiple layers. The proposed method's main drawback is that it fails to consider the label dependences at the object level and the label relationships between the scene and the object level. Thus, we plan to adopt the RNN to model the label relationships between the intralevel and interlevel semantic concepts of RS images in future work.

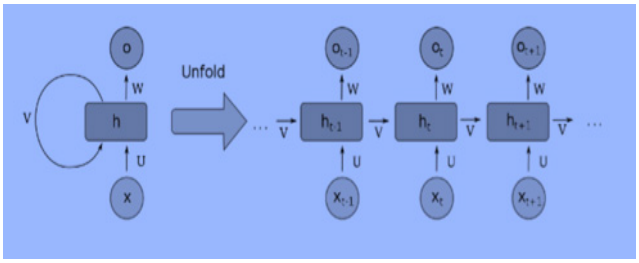

FIGURE 4. A simple RNN model; x_i: inputs, h_i: hidden states (memory), o_i: outputs, U: input weights, V: hidden weights (recurrent connection), W: output weights.

B. RECURRENT NEURAL NETWORK (RNN)

The RNN, also called the Elman network, has three layers in each period [36]. The design of RNN is illustrated in Figure 4. The three layers in each of its periods are the feedback word layer w, the recurrent layer r, and the output layer y, and the activation of these layers at time t is represented as w(t), r(t), and y(t), respectively. W(t) is the existing word trajectory that can take the form of a simple 1-of-N coding representation h(t) (i.e. the one-hot representation; it is binary, and its dimension is the same as that of the dictionary size with just one non-zero component). The calculation of Y (t) can proceed thus (Mikolov *et al.* 2010):

$$x(t) = [w(t)\ r\ (t-1)];\ r(t) = f1(U.x(t));\ y(t) = g1(V.\ r(t)); \tag{1}$$

here, x (t) represents the concatenation vector for w (t) & r (t-1), while f1 (:) is the element-wise sigmoid & g1 (:) is the softmax function. U & V are the learnable weights. A simple RNN model is typified in Figure 4. In the RNN, the size of the network is a function of the input sequence length. The recurrent layers connect the sub-networks at different time frames. Hence, when executing backpropagation, there is a need to ensure the inaccuracy is disseminated back in time via recurrent connections [12]. The study by Tsochatzidis *et al.* [20] presented a method of facilitating object instances annotation. In the approach, a polygon that outlines an object is predicted using Polygon-RNN, while the corrections are easily incorporated from an annotator in the loop. The system's analysis showed that it achieved an improved annotation speed of up

to factor 4.74 and its annotation agreement was the same as that performed by human annotators. This approach is advantageous because it offers structurally plausible object annotations and permits predefined annotation accuracy by the annotator in just a few clicks.

A multimodal RNN (m-RNN) framework was presented by Mao *et al.* [39]. For three tasks sentence generation, this platform works efficiently as the method adopted, sentence retrieval based on the probe image, and image reclamation based on the query sentence. The model components include a deep RNN and a deep CNN, which interact in a multimodal layer. The proposed m-RNN is strong in connecting sentences and images; more sophisticated language models and complex image representations could also be easily introduced into the system. The study by Joonmyun *et al.* [22] presented an approach for automatic extraction of subject-allied keywords from users' natural language observations on their media files. Here, 'theme' refers to the concepts that described media files' content, such as natural sites, pets, places, and palaces. In this approach, RNN was employed; RNN is good at implicit pattern recognition in sequential data. A semantically regularized CNN-RNN model was proposed by Feng *et al.* [23] for image annotation. With this semantic regularization, the CNN-RNN interface becomes semantically significant. It ensures the distribution of the correlation tasks and label prediction between both models and makes the full models' training more efficient and stable. The approach was evaluated on NUS-WIDE & MSCOCO and it showed efficacy on both image captioning and multilabel classification. A model for image regions' natural language descriptions generation was projected by Karpathy *et al.* [41]. This model is based on weak labels in image and sentence datasets and with few hardcoded assumptions. In this approach, there is a well-worn classification model for aligning parts of language and visual modalities. This model was shown to provide good image-sentence ranking performance. Also described was an m-RNN ar *REASON* architecture that generates visual data descriptions whose performance was evaluated on both region-level and full-frame experiments. In both cases, them-RNN performed better than other retrieval models. Most of the prevailing CNN-RNN-based techniques are prone to misprediction and object missing due to their dependence on global representation at the image-level. Hence, Linghui *et al.* [9] addressed this issue by proposing the global-local attention (GLA) method for image caption. The GLA method was believed to selectively focus on the semantically important image regions while maintaining the global context information via attention mechanisms to combine the native features (at object-level) and the global features (at image-level). When evaluated on Microsoft COCO caption datasets, the proposed GLA showed good performance using different evaluation metrics. There are two aspects of the advantages of RNN in AIA; one is its ability to generate outputs of varying lengths. The other is that it can predict the current time step output by recalling the previous inputs. However, the RNN-based image annotation method's disadvantage is the inconsistency between the value R and P and the estimation of imprecise semantic classes of a word to the precise connotation of the word in the text due to the noise in the attributes used to sequence the data.

C. DEEP NEURAL NETWORK (DNN)-BASED AIA

A DNN is a network characterized by a definite convolution level; it is a neural network consisting of more than two layers. The DNNs depend on sophisticated mathematical modeling for data processing in complex ways. Chengjian *et al.* [42] developed a new framework of a multimodal deep learning network to learn transitional depictions and deliver a decent network initialization. Then, the distance metric functions on each modality were optimized using backpropagation; finally, an optimization of the combinational weights of different modalities was performed by applying the exponentiated gradient online learning algorithm. Advance deep learning research, which will emphasize the number of feature proportions for achieving a satisfactory system performance for a given neural network framework, is necessary. Another characteristic to consider is the tool to improve a specified deep learning architecture and progress its strength. Yang *et al.* [43] proposed a new MVSAE model for a joint establishment of the correlations between high-level semantic keywords and low-level image features for automatic image annotation. First, the SAE was modified by using an iteration algorithm and a sigmoid function predictor. Then image keywords were solved with an imbalanced distribution. The influence of the imbalance learning method at different levels of keyword frequency varies. The F1 score decreases slightly towards high-frequency keywords because of a low frequency to cause a high-frequency keyword's misclassification. Contrarily, the low-level frequency keywords present a better performance compared to the original SAE. A multi-view stacked auto-encoder (MVSAE) framework has been proposed by Yang *et al.* [43] for finding the correlations between high-level semantic information and low-level visual features. Experiments on three popular datasets proved the proposed framework's effectiveness in achieving a favorable image annotation performance. The DNNs with multiple nonlinear hidden layers can learn complex input-output relationships; however, the network can be exposed to local optima and convergence difficulty due to the nonlinear mapping between the outputs and the inputs when using the BP algorithm [36].

D. LONG-SHORT-TERM MEMORY (LSTM) BASED ON AIA

When modeling temporal dynamics in sequences, RNN is a good choice even though traditional RNN finds it difficult to study long-term dynamics due to the issue of exploding and vanishing gradients to address these issues, the LSTM network was proposed [14] and the core of its architecture (refer to Figure 5) is a memory cell for storing the state of the cells over time; there is also the gates for controlling how and when the states of the cells can be updated.

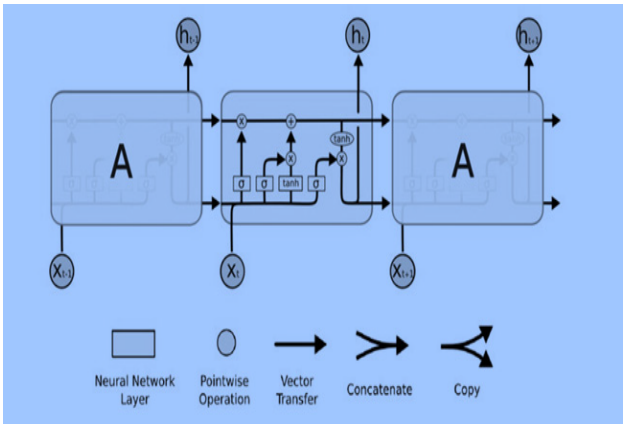

FIGURE 5. The architecture of LSTM.

Between the gates and the memory cells exists several variants with different connections. Word generation in the LSTM model depends mainly on the word's embedment at the current time-step and on the preceding hidden state (this includes the pre-information of the image) [50]. This process is systematically sustained until the end token of a sentence is encountered. Meanwhile, the progression of this procedure weakens the relevance of the image data served at the beginning. The generated words at the start of a sequence are also prone to the same issue; hence, the generation may be performed almost ''blindly'' for a long sentence towards the end of the sentence. Despite the capability of LSTM to maintain long-term memory to a certain level, it is still a problem for sentence generation [4, 1]. The study by Jia *et al.* [28] suggested a modification to the LSTM model by introducing semantic data as additional feedback to each LSTM block unit. With this modification, the model can better describe an image's content without diving into the common but unrelated phrases. This modification also made it possible to search various length regularization types for beam exploration, thereby preventing short sentences' preference and making the results better. Another effort by Zhang *et al.* [45] presented an automatic natural language description generation model for videos. This model is dependent on an LSTM based sentence generator and a 2-stream video representation-learning model. It also has a novel model for parallel video representation, which merges both motion boundary history frames and RGB frames; both frames are laden with complementary information from temporal motions and visual appearance. Notably, the suggested framework could learn the simultaneous combination of several feature streams effectively and perform end-to-end preparation of the whole model. A comparison of the model with the existing models for video description was done on 3 different datasets and the outcome showed the proposed model to perform better than the others. The study by Sarangi *et al.* [30] presented a generative AIA model that exploits both fronts' recent improvements. It uses a deep-CNN for image region detection. The experiments showed the model to achieve better training and accuracy when coupling image illustration from a discovery model with the embedment of the feedback word; it was also observed that a good portion of the information contained in the last layer of the detection model disappeared when fed to LSTM decoder as a vector. This observation could represent the class probabilities and the bounding box of the detected objects by the information contained in the last fully connected YOLO models' layer. This information is not sufficiently rich.

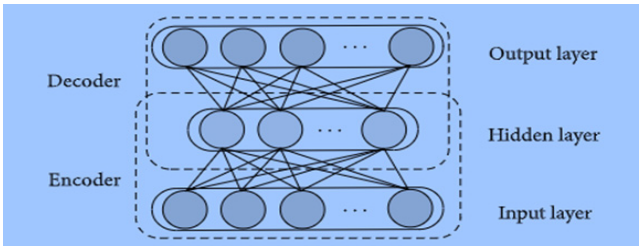

FIGURE 6. Stacked auto-encoders.

E. STACKED AUTO-ENCODER (SAE) BASED ON AIA

As a form of unsupervised learning structure, an auto-encoder is also comprised of 3 layers -input, hidden, and output layers (refer to Figure 6). There are two aspects of auto-encoder training: the encoder and the interpreter; the encoder is involved in transforming the participation information into concealed illustration, while the decoder is for the input data reconstruction from the hidden representation. The SAE is an NN with several auto-encoder layers commonly used as a DL method for dimensionality reduction [18] and feature learning [17, 22, 4, and 19].

An effective multi-model retrieval system was presented by Luo *et al.* [27] based on SAE. This system maps the extracted features (high-dimensional) from dissimilar media information types into a single low-dimensional space for metric knowledge. When using DNN to solve image annotation problems, image features are normally used as inputs to the model, while keywords serve as the model object. However, modeling the complex relationship between tags and features requires the application of several hidden layers. Having trained the DNN, it can then generate the appropriate keywords for new images. Because the performance of DNN is a function of its initial parameters, its optimization becomes a challenge. Zhou *et al.* [51] suggested two strategies for addressing data imbalance in image annotation. The study also proposed a robust, balanced and algorithm (RB-SAE) for improving the annotation effect of low-frequency tags to enhance the training of low-frequency tags. This model was also proposed to increase the annotation stability through enhanced training by a group in sub-B-SAE models. With this approach, the ability of the model to address the issue of data imbalance was ensured. Regarding the annotation process, the hypothesis of the native symmetry dataset of the unknown image was made by taking the unknown image as the starting

point. Simultaneously, there was discrimination between the high and low-frequency image attributes to establish different annotation processes. The low-frequency images were annotated using the local semantic propagation algorithm (SP), while the high-frequency images were annotated using the RB-SAE algorithm. The attribute discrimination annotation (ADA) framework was also formed to improve the overall image annotation effect. SAE is one of the commonly used DL techniques; it has received much consideration in fault diagnosis and has been studied as a common aspect of DNN [17]. The study by Jia *et al.* [7] suggested a DNN-based SAE diagnose faults in planetary gearbox and roller bearing; the input for this model was the frequency spectra after undergoing Fourier transform. Another study by Guo *et al.* [49] utilized the multidimensional statistical attributes of raw vibration signals as input for SAE; this can be considered a feature combination. Liu *et al.* [8] used STFT-created normalized spectrograms as input to SAEs to diagnose faults in a rolling bearing. Another study by [42] presented a multi-view SAE (MVSAE) model with a sigmoid predictor for annotating images. Here, the features of images are utilized as inputs to the model, while keywords are the model objects. However, the modeling of complex feature-tags relationships requires setting several hidden layers. Since the performance of DNN is a function of its initial parameters, the pre-trained constraints were adopted for MVSAE prototypical optimization. Specifically, the SAE was first trained using the chromatic feature I as the input x to aid the preliminary likelihood dissemination D 1 of the keyword. Then, the SAE is retrained using I and D 1 as the new inputs x to aid the generation of the final probability distribution D 2 of the keyword. The last process is the derivation of the image keywords $\wedge$ T from D 2.

VI. SUMMARY

The five types of DL based on AIA methods were discussed in the previous sections based on their concepts, models, algorithms, and associated problems. Some of the advantages of DL-based AIA methods were also pointed out to include the mass data, the ability to learn complex relationships, generate strong features, and the need for no manual selection. The other advantages include the derivation of sufficient side information and calculation of the alternative number of labels. However, some of the related problems of DL-based AIA methods include local optimum entrapment, the need for numerous training images, and the inability to control the training process. Conclusively, DL-based AIA methods are associated with both opportunities and challenges for AIA. The advancements in DL have brought large-scale improvements to the AIA routine on image datasets on the one hand, and on the other hand, certain challenges are still encountered in DL-based AIA methods. Among the challenges is the low efficiency of DL-based AIA procedures with increases in both the depth and breadth of DNNs [51]. Despite the capability of DNNs to learn complex input-output relationships, they still suffer from local optimum entrapment and may not converge when using the BP algorithm. Finally, irrespective of a combined RNN-CNN network's ability to solve label dependencies and label quantity prediction for large-scale image annotation, there is still a need to better rank label orders; RNN requires an ordered sequential list as input. The major challenge of the AIA techniques based on deep learning is related to high-dimensional feature analysis. All the existing features currently have the problem of not sufficiently describing the images. Another problem is the simultaneous performance of annotation and ranking in the existing methods, which is not ideal for efficient image retrieval. There is also an image ranking in each of the resulting categories from a single labeling approach to achieve better retrieval accuracy; this problem still opens new issues.

VII. DATASETS

A. COREL-5K [41]

This dataset consists of 4500 preparation images and 499 testing images, with respective images marked with up to five labels (approximately 3.4 labels per image). The Corel-5K is one of the oldest datasets for image annotation.

B. ESP GAME [23]

The ESP Game contains 18689 training images and 2081 testing images and each image is annotated with up to 15 labels (approximately 4.7 labels per image). The dataset was formed from an online game where 2 players are meant to assign labels to a given image, with a point scored for each common label. Several participants are involved in the manual annotation task; thereby, making it a challenging dataset.

C. IAPR TC-12 [42]

This dataset contains 17665 preparation images and 1962 analysis images with respective images marked with up to twenty-three labels (approximately 5.7 tags per image). Each image has a lengthy depiction of several languages. Makadia *et al.* [38], [39] used the English language to extract nouns from the image descriptions and considered them as observations. Since then, it has been a widely used method for the evaluation of image explanation methods.

D. NUS-WIDE [43]

The NUS-WIDE is the prevalent freely accessible image annotation dataset. It comprises 269648 images, which were downloaded from Flickr and with 81 labels in the vocabulary. Each image in the dataset is annotated with up to 3 labels (approximately 2.40 labels per image). Based on earlier reports [10], [34], images with labels were discarded in this report, leaving a net of 209,347 images split into 125000 preparation images and 80000 images for analysis using the split method proposed by the authors of the dataset.

E. MS-COCO [44]

The MS-COCO is next to NUS-WIDE in size and a popular image annotation dataset. It comprises 82783

TABLE 1. Performance comparison of different AIA-based methods on the Corel, ESPGame, and IAPR TC-12 datasets.

MODELS	Corel dataset				ESPGame dataset				IAPR TC-12 dataset			
	P	R	F1-score	N+	P	R	F1-score	N+	P	R	F1- score	N+
KCCA + SVM [2]	-	-	-	-	-	-	-	-	42.7	47.7	50.2	-
KCCA +2PKNN [2]	-	-	-	-	-	-	-	-	48.9	47.3	51.4	-
SEM [4]	0.37	0.52	0.43	-	0.38	0.42	0.4	-	-	-	-	-
MVAIACNN [5]	-	-	-	-	-	-	-	-	0.2035	0.4760	0.2851	81
CNN-RNN [8]	-	-	-	-	-	-	-	-	55.65	50.17	52.77	-
MVNMF-SK [21]	40.3	45.8	42.9	182	41.5	29.2	34.3	248	-	-	-	-
MVNMF-DK [21]	44	47.5	45.6	197	43.7	31.4	36.7	254	-	-	-	-
FF-CNN [29]	0.41	0.37	0.39	-	-	-	-	-	-	-	-	-
KDCN [31]	-	-	0.34	-	-	-	-	-	-	-	-	-
CCN-R [32]	32	41.3	37.2	166	44.5	28.5	34.7	248	49	31	37.9	272
NN-CNN [33]	0.43	0.51	0.47	201	0.47	0.34	0.4	256	0.51	0.36	0.42	279
MVSAE [47]	0.37	0.477	0.42	175	0.47	0.28	0.34	246	0.43	0.38	0.40	283
CCA-KNN [32]	42	52	46	201	46	36	41	260	45	38	41	278
2PKNN [2]	46.4	70.9	66.5	-	-	-	-	-	39.7	52.2	48.0	-
MV+KNN [5]	-	-	-	-	-	-	-	-	0.2468	0.3816	0.2997	81
MBRM FROM [2]	-	-	-	-	18	19	18.4	209	24	19	21.2	223
MIDDLE LAYER OF DEEP LEARNING MODEL [56]	26.93	41.43	32.64	161	43.74	33.08	37.67	258	46.15	32.80	38.35	258
PRM DEEP (2018) [55]	0.453	0.5173	0.483	201	-	-	-	-	0.4918	0.4023	0.4426	281
L-ADA (2017) [44]	0.31	0.38	0.34	164	0.35	0.19	0.25	247	0.42	0.26	0.32	273
MULTIMODULE DEEP TRANSFER LEARNING [46]	0.203	0.527	0.212	79	-	-	-	-	-	-	-	-
CNN-THOP [32]	0.527	0.583	0.553	-	-	-	-	-	-	-	-	-

preparation images with eighty tags, with the respective image being annotated with an average of 2.9 labels. Although it is not available publicly, it is used for image recognition.

VIII. EVALUATION METRICS

The dissimilar types of AIA methods' performance are evaluated using several evaluation metrics such as recall, precision, F1-score, and N+ [52].

- ***Recall and Precision*:** Given any keyword, let the number of images in the assessment dataset annotated with the label be m1. At the same time, m2 represents the number of appropriately annotated images with the label. Also, let m3 be the number of annotated images using the ground-truth data. In this case, the recall will be given as m2 m3, while the precision will be m2 m1. The recall represents the relevant information retrieval capability, while precision measures uncorrelated information refusal capability. AIA models' performance is usually evaluated using a combination of recall and precision; however, the evaluation of AIA models' performance using only recall and precision is difficult because both metrics conflict with each other. Observe that AIA methods usually perform forced annotation of test images with k (generally 5) labels even when the images are associated with more or fewer labels in the ground accuracy. Thus, the values of recall and precision may be biased even when all the ground truth labels are predicted by the model [48].
- ***F1-score*:** This is calculated thus: F 1 = 2∗P∗R P+ R. being that the performance of AIA models cannot be adequately evaluated using either the recall or the precision. There is a need to integrate them into one assessment catalog. Additionally, the F1-score can measure AIA methods' robustness, where a larger F1-score is suggestive of a more robust model.
- *N+:* This metric measures the number of correctly assigned keywords to at least one test image. It also portrays the number of keywords whose recall values are positive. Good performing AIA models usually present high N+ values.

IX. COMPARISON RESULTS

Table 1 compares 22 algorithms on three datasets Corel, ESPGame, and IAPR TC-12 are introduced comprehensively. From this chart, We find that the precision, R, F1 score, and N + of 2PKNN [2] are 46.4%, 70.9%, and 66.5%, respectively, which are the best results on this dataset Corel and

the best F1 score for all databases. Besides, the results did not reach the desired level on other databases than the rest of the algorithms. As we can see in Table 1, CNN-RNN [8] 55.65% 50.17% 52.77%, respectively, the performance of CNN-RNN [8] has significant improvement with the best precision Compared to remaining algorithms.

X. PERFORMANCE COMPARISON OF ANNOTATION METHODS

This section presents a comparison of the performance of some typical models, such as CNN, RNN, LSTM, & SAE, whose details are mentioned in Table 2.

XI. DISCUSSION AND CONCLUSION

This paper presents a review of various methods to AIA based on deep learning. The reviewed approaches are Convolution neural network (CNN) based on AIA, Recurrent Neural Network (RNN) based on AIA, Deep neural network (DNN) based AIA, Long-Short-Term Memory (LSTM) based on AIA, and Stacked Auto-Encoder (SAE) based on AIA. An assessment of the five types of AIA methods was offered based on their original idea, feature mining technique, annotation correctness, computational complexity, and datasets. Moreover, the assessment metrics used to assess AIA methods were reviewed and the advantages and issues associated with each technique were explained. The major challenge of the AIA techniques based on deep learning is related to high-dimensional feature analysis. All the existing features currently have the problem of not sufficiently describing the images. There is no existing feature that is significant enough to represent the high variation between images efficiently. The AIA technique's challenge is to reduce the semantic gap between low-level visual image features captured by machines and high-level semantic concepts perceived by humans. Many studies have been conducted on mining the image-image, image-label, and label-label correlation. Open issues, such as class-imbalance and weak-labeling of the training dataset [52]. Another problem is the simultaneous performance of annotation and ranking in the existing methods, which is not ideal for efficient image retrieval. There is also an image ranking in each of the resulting categories from a single labeling approach to achieve better retrieval accuracy [54]. Finally, some of the research on image Annotation systems is inclined to accomplish high accuracy and squat recall; the strength of image annotation is to ensure a balance between precision and recall by making sure the recall scores improve without maintaining precision. On the other hand, some image annotation models take a long time and computational complexity in the training phase, making them computationally intensive when faced with large training datasets. It is necessary to highlight the need for careful consideration of these aspects when building new image annotation techniques and datasets for future tasks [53].

TABLE 2. Performance comparison of annotation-based methods.

ANNOTATION METHOD	ADVANTAGES	DISADVANTAGE
CNN	First, CNN does not necessitate programmed feature abstraction; parameters called a kernel, an optimizable feature extractor, are functional at respective image spots, which brands CNN's exceedingly effectual for image dispensation. Subsequently, a feature may arise anyplace in the image.	Since the current CNNs are becoming deeper and deeper, they necessitate large-scale datasets and gigantic computing power for preparation. Manually accumulating a tagged dataset entails massive sums of human energy. CNN is far more information-starved because of its heaps of learnable constraints to evaluate. Thus, it is more computationally costly, demanding graphical processing units (GPUs) for prototypical preparation.
RNN	The benefits of RNN in AIA lie in four characteristics. On the one hand, RNN can produce yield with diverse lengths. On the other hand, RNN can denote the aforementioned inputs in forecasting the contemporary time step output. The third benefit of RNN over ANN is that RNN can use prototypical data arrangement (i.e., time series) so that each model can be expected to be reliant on preceding ones, and last one Recurrent neural network are exercised with convolutional layers to spread the operative pixel region.	The disadvantage of the RNN-based image annotation method in five aspects the firstly, There is an inconsistency between the value R and P; secondly, even in case the characteristics adopted for the preparation data has some noise inaccurate semantic classes of a word concerning the word's precise denotation in text, problematic for customary RNN to study long-term underlying forces because of the issue of disappearing and explosion gradients, Preparation an RNN is very problematic. It cannot practice very lengthy classifications if using tanh or relu as an initiation function.
LSTM	To amend vanishing or exploding Gradient that The basic RNN models suffer from it, can artificially study to recall and overlook data using precise gates to regulate the data flow.	the disadvantage of LSTM uncertain tags by aggregating the contrast among pixels (the restraint of RGB input) the conduct of LSTM networks instigates these errors
SAE	SAE robust capability to pact with the unstable data. To increase the annotation stability through enhanced training, (SAEs) have concerned substantial courtesy in fault diagnosis	The SAE might be more sensitive to input errors different from those in the training set or changes in underlying relationships that a human would notice. The autoencoder algorithm requires an objective function for evaluating the accuracy of encoded/decoded input data [50].

ACKNOWLEDGMENT

(Myasar Mundher Adnan and Rizwan Ali Naqvi are co-first authors.)

COMPETING INTEREST

All the authors declare no competing interest.

REFERENCES

[1] K. Barnard, P. Duygulu, and D. Forsyth, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, no. 2, pp. 1107–1135, 2003.

[2] T. Uricchio, L. Ballan, L. Seidenari, and A. Del Bimbo, "Automatic image annotation via label transfer in the semantic space," *Pattern Recognit.*, vol. 71, pp. 144–157, Nov. 2017, doi: 10.1016/j.patcog.2017.05.019.

[3] C.-W. Shih, H.-C. Chu, Y.-M. Chen, and C.-C. Wen, "The effectiveness of image features based on fractal image coding for image annotation," *Expert Syst. Appl.*, vol. 39, no. 17, pp. 12897–12904, Dec. 2012.

[4] Y. Ma, Y. Liu, Q. Xie, and L. Li, "CNN-feature based automatic image annotation method," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3767–3780, Feb. 2019.

[5] R. Wang, Y. Xie, J. Yang, L. Xue, M. Hu, and Q. Zhang, "Large scale automatic image annotation based on convolutional neural network," *J. Vis. Commun. Image Represent.*, vol. 49, pp. 213–224, Nov. 2017.

[6] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. Eur. Conf. Comput. Vis.*, May 2002, pp. 97–112.

[7] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding long-short term memory for image caption generation," 2015, *arXiv:1509.04942*. [Online]. Available: http://arxiv.org/abs/1509.04942

[8] H. Wang and B. Raj, "On the origin of deep learning," 2017, *arXiv:1702.07800*. [Online]. Available: http://arxiv.org/abs/1702.07800

[9] J. I. Katuka, D. Mohamad, and T. Saba, "An analysis of object appearance information and context based classification," *3D Res.*, vol. 5, p. 24, Sep. 2014, doi: 10.1007/s13319-014-0024-5.

[10] G. Zhang, C.-H.-R. Hsu, H. Lai, and X. Zheng, "Deep learning based feature representation for automated skin histopathological image annotation," *Multimedia Tools Appl.*, vol. 77, no. 8, pp. 9849–9869, Apr. 2018.

[11] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas bimmanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 33–115, 1943.

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 60, no. 2, pp. 1097–1105, 2012.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[15] Y. Bengio, "Learning deep architectures for AI," *Found. Trend Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[16] K. Osako, R. Singh, and B. Raj, "Complex recurrent neural networks for denoising speech signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2015, pp. 1–5.

[17] A. Gupta, H. Wang, and M. Ganapathiraju, "Learning structure in gene expression data using deep architectures, with an application to gene clustering," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2015, pp. 1328–1335.

[18] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1995–2003.

[19] M. H. Ali, K. Al-Jawaheri, M. M. Adnan, A. Aasi, and A. H. Radie, "Improved intrusion detection accuracy based on optimization fast learning network model," in *Proc. 3rd Int. Conf. Eng. Technol. Appl. (IICETA)*, Sep. 2020, pp. 198–202.

[20] L. Tsochatzidis, K. Zagoris, N. Arikidis, A. Karahaliou, L. Costaridou, and I. Pratikakis, "Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach," *Pattern Recognit.*, vol. 71, pp. 106–117, Nov. 2017.

[21] R. Rad and M. Jamzad, "Image annotation using multi-view non-negative matrix factorization with different number of basis vectors," *J. Vis. Commun. Image Represent.*, vol. 46, pp. 1–12, Jul. 2017.

[22] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[23] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in *Proc. 31sst AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 4133–4139.

[24] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, and T. Liu, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[25] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2004, pp. 319–326.

[26] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The iapr TC-12 benchmark: A new evaluation resource for visual information systems," in *Proc. LREC Workshop Image Lang. Resour.*, 2016, pp. 13–23.

[27] W. Luo, M. Feng, Q. Fan, Q. Peng, G. Li, X. Hao, L. Yu, and Y. Xia, "Image annotation of power grid objects based on convolutional neural networks," in *Proc. 8th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2016, pp. 1–4.

[28] Y. Lin and H. Zhang, "Automatic image annotation via combining low-level colour feature with features learned from convolutional neural networks," *NeuroQuantology*, vol. 16, no. 6, pp. 679–685, Jun. 2018.

[29] A. Mahmood, M. Bennamoun, S. An, F. Sohel, F. Boussaid, R. Hovey, G. Kendrick, and R. B. Fisher, "Automatic annotation of coral reefs using deep learning," in *Proc. OCEANS MTS/IEEE Monterey*, Sep. 2016, pp. 1–5.

[30] N. Sarangi and C. Chandra Sekhar, "Automatic image annotation using convex deep learning models," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2015, pp. 92–99.

[31] J. Cao, A. Zhao, and Z. Zhang, "Automatic image annotation method based on a convolutional neural network with threshold optimization," *PLoS ONE*, vol. 15, no. 9, Sep. 2020, Art. no. e0238956.

[32] M. M. Kashani and S. H. Amiri, "Leveraging deep learning representation for search-based image annotation," in *Proc. Artif. Intell. Signal Process. Conf. (AISP)*, Oct. 2017, pp. 156–161.

[33] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[34] M. Rajchl, M. C. H. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, and D. Rueckert, "DeepCut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 674–683, Feb. 2017.

[35] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[36] S. Waheed, M. Adnan, N. Suaib, and M. S. Rahim, "Fuzzy logic controller for classroom air conditioner," *J. Phys., Conf. Ser.*, vol. 1484, no. 1, 2020, Art. no. 012018.

[37] Y. Chen, L. Liu, J. Tao, X. Chen, R. Xia, Q. Zhang, J. Xiong, K. Yang, and J. Xie, "The image annotation algorithm using convolutional features from intermediate layer of deep learning," *Multimedia Tools Appl.*, vol. 80, no. 3, pp. 4237–4261, Jan. 2021.

[38] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-RNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5230–5238.

[39] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," 2014, *arXiv:1412.6632*. [Online]. Available: http://arxiv.org/abs/1412.6632

[40] P. Zhu, Y. Tan, L. Zhang, Y. Wang, and J. Mei, "Deep learning for multilabel remote sensing image annotation with dual-level semantic concepts," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4047–4060, Jan. 2020.

[41] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.

[42] S. Chengjian, S. Zhu, and Z. Shi, "Image annotation via deep neural network," in *Proc. 14th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, May 2015, pp. 518–521.

[43] Y. Yang, W. Zhang, and Y. Xie, "Image automatic annotation via multi-view deep representation," *J. Vis. Commun. Image Represent.*, vol. 33, pp. 368–377, Nov. 2015.

[44] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.
[45] C. Zhang and Y. Tian, "Automatic video description generation via LSTM with joint two-stream encoding," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2924–2929.
[46] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[47] M. M. Adnan, M. S. M. Rahim, S. M. Khaleel, and K. Al-Jawaheri, "A survey automatic image annotation based on machine learning models," *J. Eng. Appl. Sci.*, vol. 14, no. 20, pp. 7627–7635, Oct. 2019.
[48] I. Supriana and Y. Pratama, "Face recognition new approach based on gradation contour of face color," *Int. J. Elect. Eng. Inform.* vol. 9, no. 1, p. 125, 2017.
[49] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, "Semantic regularisation for recurrent image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2017, pp. 2872–2880, doi: 10.1109/CVPR.2017.443.
[50] A. Benba, J. Abdelilah, and A. Hammouch, "Detecting patients with Parkinson's disease using Mel frequency cepstral coefficients and support vector machiness," *Int. J. Elect. Eng. Inform.* vol. 7, no. 2, p. 297, 2015.
[51] M. Satone and G. Kharate, "Feature selection using genetic algorithm for face recognition based on PCA, wavelet and SVM," *Int. J. Elect. Eng. Inform.*, vol. 6, no. 1, p. 39, 2014.
[52] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, "A survey and analysis on automatic image annotation," *Pattern Recognit.*, vol. 79, pp. 242–259, Jul. 2018.
[53] P. K. Bhagat and P. Choudhary, "Image annotation: Then and now," *Image Vis. Comput.*, vol. 80, pp. 1–23, Dec. 2018.
[54] M. M. Adnan, M. S. Mohd Rahim, K. Al-Jawaheri, M. H. Ali, S. R. Waheed, and A. H. Radie, "A survey and analysis on image annotation," in *Proc. 3rd Int. Conf. Eng. Technol. Appl. (IICETA)*, Sep. 2020, pp. 203–208.

MYASAR MUNDHER ADNAN received the B.E. degree in computer science from Alkufa University, Iraq, in 2011, and the M.S. degree in computer science from UTM, in 2014, where he is currently pursuing the Ph.D. degree. His current research interests include deep learning, image processing, and machine learning.

MOHD SHAFRY MOHD RAHIM received the B.Sc. degree (Hons.) in computer science and the M.Sc. degree in computer science from Universiti Teknologi Malaysia, in 1999 and 2002, respectively, and the Ph.D. degree in spatial modelling from Universiti Putra Malaysia, in 2008. He is currently a Professor of Image Processing with the School of Computing, Universiti Teknologi Malaysia. He also had appointed to be Deputy Director of the Centre for Joint Programme, UTMSPACE. He current focused his research together with his research group, UTM ViCube Lab under the Faculty of Computing, UTM. He is also an Expert in the research area of computer graphics and image processing.

AMJAD REHMAN (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Computing, Universiti Teknologi Malaysia, with a specialization in forensic documents analysis and security, in 2010. He is currently a Senior Researcher with the Artificial Intelligence and Data Analytics Laboratory, Prince Sultan University, Riyadh, Saudi Arabia. He is also a PI in several funded projects and also completed projects funded from MOHE Malaysia, Saudi Arabia. He is the author of more than 200 ISI journal papers, conferences. His H-index is 40 with 4000 citations. His research interests include data mining, health informatics, and pattern recognition. He received a Rector Award for the 2010 Best Student from Prince Sultan University.

ZAHID MEHMOOD received the B.S. degree (Hons) in computer engineering from the COMSATS University of Sciences and Technology–Wah, Pakistan, in 2009, the M.S. degree in electronic engineering with a specialization in signal and image processing from International Islamic University (IIU), Islamabad, Pakistan, in 2012, and the Ph.D. degree in computer engineering with a specialization in content-based image retrieval (CBIR) from the University of Engineering and Technology (UET), Taxila, Pakistan, in 2016. He has published more than 70 publications in impact factor journals (ISI indexed) and international conferences. He is also a Team-Lead of the Forensic Analysis, Machine Learning, and Information Retrieval (FAMLIR) research group. He is a reviewer of international journals and conferences, such as IEEE Access, *Pattern Recognition*, *Information Fusion*, *Soft Computing*, *Pattern Recognition Letter*, *Neural Computing and Applications*, *Neurocomputing*, *Journal of Electronic Imaging*, *Journal of Information Science*, *Computer & Electrical Engineering*, PAMI, and CVPR. His research interests include content-based image retrieval (CBIR), medical imaging, deep learning, image forensic, computer vision, and machine learning.

TANZILA SABA (Senior Member, IEEE) received the Ph.D. degree in document information security and management from the Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia, in 2012. She is currently serving as an Associate Chair for the Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia. Her research interests include medical imaging, pattern recognition, data mining, MRI analysis, and soft-computing. She has above 100 publications that have around 4370 citations with H-index 42. Her mostly publications are in biomedical research published in ISI/SCIE indexed. Due to her excellent research achievement, she is included in Marquis Who's Who (S&T) 2012. She is also an editor and a reviewer of reputed journals and on the panel of TPC of international conferences. She has full command of a variety of subjects and taught several courses at the graduate and postgraduate levels. On the accreditation side, she is a skilled lady with ABET and NCAAA quality assurance. She is also the Leader of the Artificial Intelligence and Data Analytics Research Laboratory, PSU, and the Active Professional Member of ACM, AIS, and IAENG organizations. She is also the PSU WiDS (Women in Data Science) Ambassador with Stanford University and Global WomenTech Conference. She won the Best Student Award in the Faculty of Computing, UTM, in 2012. She received the Best Researcher Award from PSU for consecutive four years. She has been nominated as a Research Professor with PSU since September 2019.

RIZWAN ALI NAQVI (Member, IEEE) received the B.S. degree in computer engineering from COMSATS University Islamabad, Pakistan, in 2008, the M.S. degree in electrical engineering from Karlstad University, Sweden, in 2011, and the Ph.D. degree in electronics and electrical engineering from Dongguk University, South Korea, in 2018. From 2011 to 2012, he was a Lecturer with the Computer Science Department, Sharif College of Engineering and Technology, Pakistan. He joined the Faculty of Engineering and Technology, The Superior College, Pakistan, as a Senior Lecturer, in 2012. After his Ph.D. degree, he worked as a Postdoctoral Researcher with Gachon University, South Korea, from 2018 to 2019. He is currently working as an Assistant Professor with Sejong University, South Korea. His research interests include gaze tracking, biometrics, computer vision, artificial intelligence, machine learning, deep learning, and medical imaging analysis.