ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 11, No. 2, April 2023, pp.176-185

Automatic image annotation system using deep learning method to analyse ambiguous images

Ali Abbas Al-Shammary [1], Nizar Zaghden [2], Med Salim Bouhlel [3]
[1] National School of Electronics and Telecommunications of Sfax, University of Sfax.
[2] Higher School of business, university of sfax
[3] ISBS, university of sfax
[1.2.3] Research Lab: Smart systems for Engineering & E-health based on Technologies of Image & Telecommunications. (SETIT), ISBS. University of Sfax, Tunisia

ABSTRACT

Image annotation has gotten a lot of attention recently because of how quickly picture data has expanded. Together with image analysis and interpretation, image annotation, which may semantically describe images, has a variety of uses in allied industries including urban planning engineering. Even without big data and image identification technologies, it is challenging to manually analyze a diverse variety of photos. The improvements to the Automated Image Annotation (AIA) label system have been the subject of several scholarly research. The authors will discuss how to use image databases and the AIA system in this essay. The proposed method extracts image features from photos using an improved VGG-19, and then uses nearby features to automatically forecast picture labels. The proposed study accounts for both correlations between labels and images as well as correlations within images. The number of labels is also estimated using a label quantity prediction (LQP) model, which improves label prediction precision. The suggested method addresses automatic annotation methodologies for pixel-level images of unusual things while incorporating supervisory information via interactive spherical skins. The genuine things that were converted into metadata and identified as being connected to pre-existing categories were categorized by the authors using a deep learning approach called a conventional neural network (CNN) - supervised. Certain object monitoring systems strive for a high item detection rate (true-positive), followed by a low availability rate (false-positive). The authors created a KD-tree based on k-nearest neighbors (KNN) to speed up annotating. In order to take into account for the collected image backdrop. The proposed method transforms the conventional two-class object detection problem into a multi-class classification problem, breaking the separated and identical distribution estimations on machine learning methodologies. It is also simple to use because it only requires pixel information and ignores any other supporting elements from various color schemes. The following factors are taken into consideration while comparing the five different AIA approaches: main idea, significant contribution, computational framework, computing speed, and annotation accuracy. A set of publicly accessible photos that serve as standards for assessing AIA methods is also provided, along with a brief description of the four common assessment signs.

Keywords: VGG19 algorithm, Annotated Image, Image Analysis, AIA model

Corresponding Author:

Ali Abbas Al-Shammary
National School of Electronics and Telecommunications of Sfax,
University of Sfax.
E-mail: alidyala1987@gmail.com

1. Introduction

Raw images are becoming increasingly common as technology develops in various activities and places. Millions of images are frequently stored without any form of detailed labelling. As a result, finding desired photographs has grown difficult and time-consuming. [1] Giving one or more labels (tags) that characterize the contents of a picture is the image labelling strategy, sometimes referred to as image annotation. The automatic

© The Author 2023. This work is licensed under a Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/) that allows others to share and adapt the material for any purpose (even commercially), in any medium with an acknowledgement of the work's authorship and initial publication in this journal.

extraction of language from photographs [4] and the automatic classification of images for social media [2] are only a few of the uses for this technology. These issues led to the development of automatic image annotation (AIA), which has since grown into a crucial and significant academic field. AIA models concepts with easily accessible pre-annotated photo sources. The trained model will then be used to complete incompletely labelled photographs or to classify unidentified photos [5]. Aerial photography is one of the subfields of remote sensing. the most popular and commonly used platforms from which aircraft and satellites conduct remote sensing observations. Aerial photography, a branch of remote sensing, is used for a variety of tasks. Balloons and kites were used in the early days of aerial photography [6] [7]. There are six main characteristics of every image, including the following: (1) Synoptic view: Seeing through the eyes of a person is akin to social media photographs or space recordings taken across a sizable area. These technologies make it possible to recognize small details and the spatial relationships between them. (2) They are thought of as real recordings of the circumstances that existed on the Earth's surface at a certain epoch and may be utilized as historical records. (3) The ability to stop motion: By showing a stop-action image of the dynamic situation, researchers were able to investigate dynamic / changeable phenomena including flooding, migrating animals, transportation, oil spills, forest fires, and changing dynamics in natural phenomena. (4) Three-dimensional perspective Measurements in both the horizontal and vertical planes are possible because to the stereoscopic portrayal of the Earth's surface in aerial photographs. (5) Resolution: Images can be affected by electromagnetic waves with wavelengths which are beyond the range of the human eye's sensitivity. (6) Availability: Images may be captured at user-specified intervals and in a variety of sizes, providing a permanent record of any location. All of these variables were covered in this study in order for the researchers to quickly choose the finest ones. The majority of these attributes are stored in multiple databases, including (Kaggle, Image Net). The capturing of images typically contains multiple aberrations then image manipulation is a rapidly enhancing technical method [8]. For example, image procuration, it's possible that any type of noise will be present. In this paper [9], the authers investigated how image problems impact CNN-based image classifiers. There are various kinds of image blur, including motion blur, average blur, Gaussian blur, out-of-focus blur, and blur caused by air turbulence. [10] [11]. In addition to severe distortions such as image noise / independent noise / spike noise). Poisson noise (shot noise), and Poisson noise (shot noise). The effect of two types of picture distortions—Gaussian blur and Salt & Pepper noise—on CNN-based image classifiers has been explored in this study. Before being used, CNNs are trained. The study is structured into four sections: a review of related literature, the function of deep learning, a practical component containing the datasets obtained, and conclusions and recommendations for future research.

2. Literature survey

This section's literature review covers the in-depth study that provides a window into the work done using deep models. Table 1 wraps up the study of earlier work for categorizing and identifying objects sources.

1. Qimin Cheng, et. al, 2018 [12]: The authors synthesized 138 publications from the previous 20 years to give their detailed evaluation of contemporary AIA methodologies. They divided AIA techniques into five groups: 1) Image annotation based on generative models, 2) Image annotation based on nearest neighbors, 3) Image annotation based on discriminative models, 4) Image annotation based on tag completion, and 5) Image annotation based on deep learning. Based on the fundamental concept, primary contribution, model framework, computational complexity, computation time, and annotation accuracy, comparisons of the five different AIA approaches are made. They also provided an overview of four typical assessment measures and five publicly accessible picture datasets that are used as benchmarks for assessing AIA approaches. Then, using a benchmark dataset and accepted assessment measures, the effectiveness of a few typical or well-behaved models is evaluated.
2. Nizar Zaghden et. al 2013 [13]: The authors presented a hybrid system based on a local approach (SIFT descriptor) and a global approach (fractal dimension). Since it is rotation invariant and quite resilient to shifting light, the Scale Invariant Feature Transform looks to work well with their application. In the first stage, the computation of the fractal dimension is used to filter out images that contain attributes that are out of alignment with the specifications of the image request. The SIFT is then used to display which photos best fit the request. However, as compared to "fractal dimension" and "SIFT descriptor" methodologies applied separately, the hybrid approach's average matching time is faster. The key benefit of utilizing the SIFT descriptor as the second phase is the time saved over the authors' original method of matching one image with 1,000 other photos in their database. The problem of picture matching can be solved more effectively and efficiently by combining fractal dimensions with keypoint descriptors. Several experiments

were used, and they demonstrated the system's reliability for things like text graphics and figures from old texts.

3. Allan Hanbury 2008 [14]: In order to evaluate automated image annotation and object identification algorithms, a set of images with suitably annotated descriptions of each shot is required. This study covers the free text annotation, keyword annotation, and ontology-based annotation methods for picture annotation. The use of photo annotation is then considered. In this article, they talked about developing keyword vocabularies for automatic picture annotation assessment. They also explore alternative approaches to increase the efficiency of ground truth generation because direct hand annotation of photos involves a lot of time and effort. A list of annotated photo datasets for computer vision research is available.
4. Marina Ivasic-Kos, Ivo Ipsic et. al, 2010 [15]: The authors worked with a semantic picture and proposed a framework for fuzzy petri net-based content analysis. In the context of the researched outdoor domain, knowledge scheme is utilized to establish more general and sophisticated semantic ideas and their relationships. There is a formal explanation of the spatial and hierarchical relationships between ideas from the outdoor image domain. The fuzzy recognition algorithm-based automated image annotation process, which associates high-level semantics with images, is given. A solution that can improve picture management and retrieval has emerged: automatic image annotation. The goal is to annotate images with higher-level ideas that match to keywords that consumers naturally use while looking for images online. It is difficult to infer high-level semantics from an image's attributes since it is crucial to study all picture components and their interactions as well as include knowledge needed for the overall image's semantic interpretation. For knowledge representation, this work employs the KRFPN formalism, which is based on the Fuzzy Petri Net theory. A well-defined semantics and straightforward graphical notation are used in this representation to make the model simple to understand. At different semantic levels of abstraction, the well-defined inference techniques may be employed for picture annotations. The paper briefly describes a paradigm that associates feature descriptors with domain classes. In this approach, there are two different types of weighting that are used: weighing the descriptor values and weighting the descriptors' effects on classification performance. Additionally, a portion of the knowledge base that focuses on generalization, spatial, and partial links among ideas is offered. Although the research is restricted to a single field of application, we think our strategy will work well with bigger datasets of comparable photos.
5. Jian Kang, et. al 2021 [16]: The segmentation of images or any object from high-resolution remote sensing (RS) images is critical for detection, and calculation. CNNs have been used to create these ojects, an approach that approved its usefulness. Yet, extensive, pixel-level annotations are needed to fully use CNNs' predictive capabilities. There have been few attempts to divide building footprints with partial annotations, since most cutting-edge CNN-based algorithms focus on creating network topologies for improved predictions of building footprints with complete annotations. In this paper, they introduced a unique semi-supervised learning technique for segmenting building footprints that can precisely forecast building footprints using a network trained with sparse annotations (for example, only 0:0324km2 out of 2:25km2 area is annotated).
6. Sirine Ammar, Nizar Zaghden, et.al., 2020 [17] A model for segmenting and classifying moving objects in video footage was developed using generative adversarial networks and the unsupervised anomaly detection framework DeepSphere (GANs). The outcomes show that the proposed method outperforms cutting-edge techniques in the segmentation and classification of moving objects from video surveillance. The suggested Deep learning approach locates, distinguishes, and classifies moving objects in video sequences by utilizing the Deep Sphere architecture. The Deep Sphere framework is initially used to segment moving objects. To get binary segmentation labels and morphological filtering, threshold the network outputs. All BGS Library algorithms have been shown to underperform Deep Sphere-based object segmentation. The proposed approach first extracts deep features from extracted pictures and then classifies the extracted images using the efficient data classification capabilities of the GAN discriminator.
7. Ferentinos et. al. 2018 [18] Using a publicly accessible database of 25 distinct plant species, CNN models AlexNet, AlexNetOWTBn, GoogLeNet, Overfeat, and VGG were used, and accuracy of 99.53% was obtained. A deep learning model called DLHPQE was proposed by P. Mahajan et al. [19] for evaluating the impact of environmental factors, such as haze, on forecasting As compared to two-stage tiny CNN architectures like MobileNet, SqueezeNet, and NasNet Mobile, large scale architectures InceptionV3 and VGG16 for detecting and diagnosing rice illnesses have been presented. Eight distinct kinds of data were gathered in a real-world setting from the paddy fields of the Bangladesh Rice Research Institute (BRRI). Using the UC Merced Land Use aerial dataset, the effectiveness of the AlexNet and GoogleNet designs was

assessed under the impact of Gaussian Blur. The study's final findings looked at CNN's model's resistance to blur. Compared to AlexNet, GoogleNet is more adaptable to different degrees of Gaussian blurring.

The corpus of research demonstrated many ways that VGG-19 may be studied in dynamic contexts. This study's goal is to assess how well VGG-19 performs in terms of visual distortions. The dataset is improved using salt-pepper noise and Gaussian blurring. This study contributes by outlining the impacts of noise and blurring on SVM classification performance [24].

3. Datasets

ADE20K data [31] are tested more than 27K photos from the SUN and Places databases make up ADE20K. Over 3K item categories are thoroughly tagged into the images. Numerous photos also include item components and portions of other parts. For a modal segmentation, we additionally offer the original annotated polygons and object instances. Additionally, pictures are identified, masking faces and car plates.

The dataset, the authors looked for a big collection of scenes in-depth annotations of all the visual notions. The three types of visual conceptions are as follows: A discrete object is one that has a distinct shape, like a car or a human. A backdrop object, like grass or the sky, comprises amorphous parts. An object portion, or component of an existing object instance that performs a specific function, such as a head or a leg, is a third choice. Images from the LabelMe [27], SUN datasets [29], and Locations [32] The 1900 types of scenes listed in the ADE20K database were covered by databases. One knowledgeable employee tagged the photographs using the LabelMe interface. Figure 1 displays screenshots of the annotation interface as well as a fully segmented image. The authors provided three main sorts of annotations: named object and stuff segments, object components, and object attributes. As each instance of an object has a unique segmentation, the dataset may be used to train and evaluate segmentation or detection algorithms. To identify the dataset's segments, polygons are employed. There are several places where the polygons from various objects overlap since every item in the collection is completely marked, even in the areas where they are occluded. To construct a segmentation mask from the annotated polygons, the depth layers of each picture are sorted. Background classes are positioned as the bottom layer, such as the sky or the wall. When an object from the inner polygon is completely ringed by another polygon, a closer depth layer is assigned to it, and the remaining object depths are determined as follows. When two items only partially overlap, we examine the location of their polygonal intersections and award the nearest object to the polygonal junction with the greatest number of points. When the objects are sorted, the segmentation mask is made by continually passing through the items with a decreasing depth to make sure that no object is ever covered by its components or by the pieces of another object [33].

Image	Basis object	Forecast
	Sky Building Tree Arch, Window, Roof	Building Tree Windows Roof, Arch, Sky
	Sky control tower mountain dome, airplane, building, car, box, runway	Airplane control tower building car box, runway, dome
	Sky Road Person building bench	Building Road Person Bench Sky
	Boats Sea Airplane Sun, Buildings, Ocean buoy	Boats Sea Airplane Sun, Buildings

Figure 1. shows a snapshot annotation interface as well as one completely segmented picture

4. The role of deep learning

To determine the right labels and semantic objects, this research aims to detect any image's missing semantic objects. Before recommendations concerning the selection of network design, Further research is critically needed to train deep neural networks and improve computer efficiency. It is crucial to consider the research topic of verbal description of concepts. Yet unlike discrete key words, which are typically employed to identify images, sentences are more substantial, concise, and sensitive in their representations of information. Furthermore, it is expected that the subsequent step of phrase synthesis would be straightforward with a strong intermediate space connecting pictures and tags. Additionally, a difficult issue with numerous applications is the use of a human-like approach, such as stochastic Petri-nets (SPN) graphs, for a more thorough comprehension of the natural language text phrase. Robotics, medicine, and a variety of other industries have all seen a change because of too deep learning. One deep learning model that can automatically acquire spatial feature hierarchies is the convolutional neural network, which analyzes data with grid patterns like photographs. This section investigates the suggested technique for our research [25]. Figure (2) shows the architecture of CNN which frequently has pooling layers, completely connected layers, and many convolutional layers. N stands for inputs No, h and w for the height and breadth, and c for the input's depth (RGB image). This four-dimensional tensor is the input to a CNN. The convolution layer rolls the input k for size filter. [23] The convolution layer concatenates the input. ($k_h \times k_w \times k_c$), where k_h is less than h, k_w less than w, and k_c less than or equal to c, produces k of feature images to the next layer. The foundation of local connections is formed by these kernels, which have identical properties. The kernel and a section of the layer's input matrix are dot products at each convolution cycle (typically, the Frobenius inner product), and an activation function—typically, the ReLU function—is then applied. To produce a feature map, the kernel repeatedly iterates through the input matrix. The feature maps are subsampled by the pooling layers, which reduces the number of training parameters required. Depending on where the neurons are positioned in the feature map, the pooling technique frequently employs the maximum (max pooling), average (average pooling), or global (global pooling) value of the neurons. These techniques make CNN a better deep learning system by drastically lowering the number of parameters and speeding up convergence. Since they completely use the input image data's two-dimensional structure, CNNs are preferred in computer vision applications. [26].

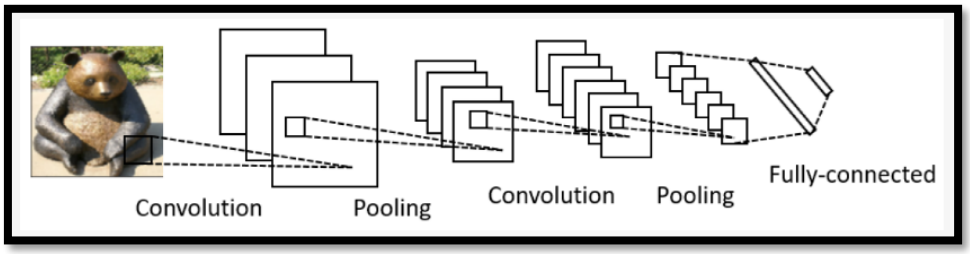


Figure 2. CNN Architecture [27]

The convolutional network is indicated as f(.) for a collection of pictures x, where the photos are filtered using dense linked layers and convolutional layers. A scoring function made up of a vector of activations at the point x of data that is the output of the function f(.). For training, it is assigned that n is no. of photos and c no of labels. The function of lose WARP (Weighted Approximate Ranking) minimizes the following formula: [28]

$$J = \sum_{i=1}^{r} \sum_{j=1}^{c+} \sum_{k=1}^{c-} \mathrm{L}(r_j) \max\left(0, 1 - \mathrm{fj(xi)} + \mathrm{fk(xi)}\right) \quad \ldots\ldots (1)$$

where L(.) is an used by for various ranks, and for picture I r_j gives the rank for the j^{th} class. As seen in (2), A weight function L(.) is as follows:

$$L(.) = \sum_{j=1}^{r} a_j \qquad \ldots\ldots (2)$$

Where the weights determined by L(.) control the top-k of the optimization and a_j is defined as 1/j. If a positive label is given top priority in the label list, L(.) will just give the loss a little weight and won't overpay for it. L(.) will give the loss a much larger weight if a positive label is not placed first, pushing the positive label to the top. Additionally, given c classes and s sample trials, equation (3) calculates the rank rj.

$$r_j = \left| \frac{c-1}{s} \right| \qquad \ldots\ldots (3)$$

5. Methodology

Figure.2 shows the fundamental framework for using the VGG19 model to determine if any image is available, regardless of how clear it is. Before adding distortions such noise, secondary data was gathered in the first stage to assess their impact on the availability of object detection.

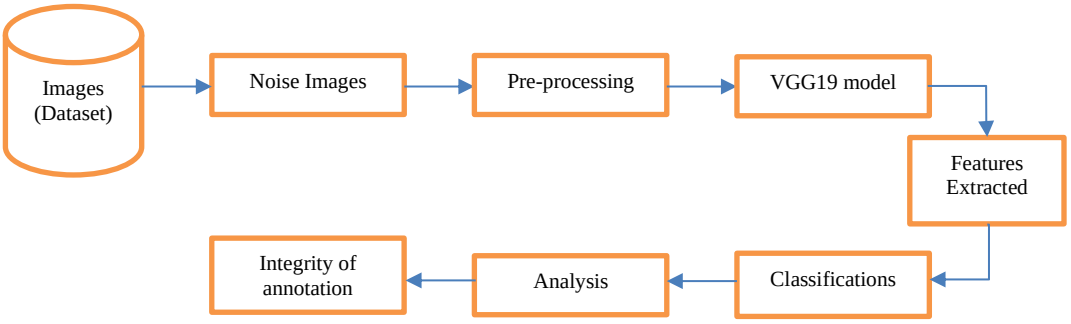


Figure 2. Proposed Methodology

The pre-processing is done by applying Gaussian function given by the equation specified below. It amends the pixel's value conferring to averaging its neighbouring pixels [20]. The equation is as follows:

$$G(x,y) = \frac{1}{2\pi\sigma^2}\, e^{-(x^2+y^2)/2\sigma^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Here σ, the blur factor is denoted by, the Euler number is denoted by e, and the horizontal and vertical distances from the centre pixel are denoted by (x, y).
Similarly, noise is introduced into dataset by applying Salt and pepper noise expressed as [29]:

$$n(s) = \begin{cases} N_a, & s = a \\ N_b, & s = b \\ 0, & otherwise \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

where s represents pixel's intensity values of a noisy image. Also, a and b represents noise impulses, for b>a, intensity b appears to be light point while a pitch up as dark point onto image.

Table 1: *Unique Features of The Vgg19 Model Implemented In The Proposed Work*

Models	Years	Total Parameters	Top1 Accuracy	Top5 Accuracy
ResNet50 [29]	2012	25,636,712	0.75	0.92
VGG16 [30]	2014	138,357,544	0.71	0.90
VGG19 [30]	2014	143,667,240	0.72	0.90
InceptionV3 [30]	2015	2,385,178,4	0.78	0.94
EfficientNetB0 [27]	2019	4.049.564	0.77	0.93

6. The results

The outcomes of the experiment that is now underway have been studied and generalized. The main goal of the experiment is to analyse the ramifications of these distortion-related problems, particularly how the number of classes used to train the network affects these concerns. The test setup includes an Intel(R) Core(TM) 3227U CPU operating at 1.90GHz and 8 GB of RAM. On a Windows 10 computer, CNNs are implemented using Python. Precision, recall, and F1-score values have all been calculated as performance measures for TP (true positive), TN (true negative), FP (false positive), and FN (false negative).
Precision is measured as the ratio of properly projected positive observations to all correctly expected positive observations.

$$Precision = \frac{TP}{TP + FP} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

By dividing all observations that are part of the actual class by the percentage of observations that were correctly predicted to be positive, recall—also known as sensitivity—is calculated. [34].

$$Recall = \frac{TP}{TP + FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

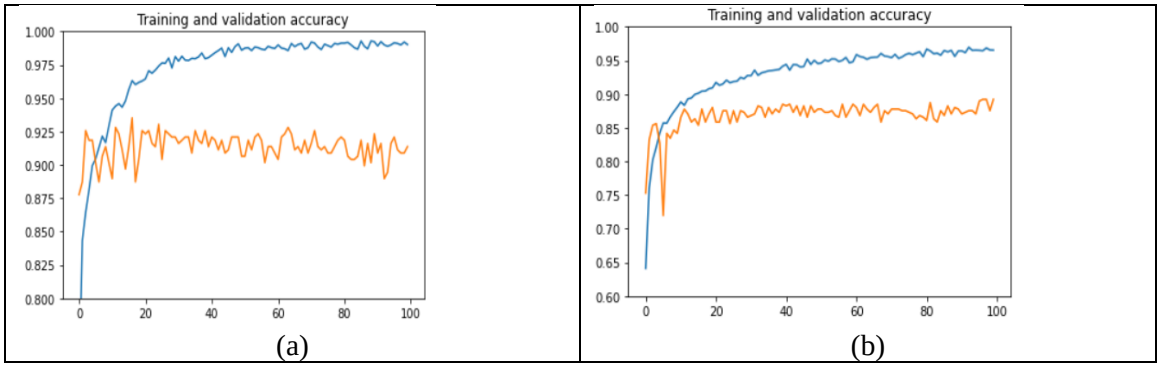
The F1 Score is the weighted average of Precision and Sensitivity. False outcomes are taken into account, both good and negative.

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Table 2. Dataset for classification of ade20ks' images

Class	ADE20K (Not annotated)	ADE20K (annotated)	Other Dataset
Train Images	**200**	**205**	**872**
Test Images	**214**	**213**	**347**
Total No. of Images	**414**	**418**	**1219**

To reach the highest level of accuracy, all of the aforementioned models are applied to data with 100 epochs and either tagged photos are present or not. Several training-validation plot curves were created for each model. Figure 3 shows the plotted history of the VGG19 model's training and validation accuracy both with and without image distortions (Gaussian Blur and Salt & Pepper Noise) [35]. Similarly, all of the aforementioned CNN models have been used, with EfficientNetB0 having the highest validation accuracy (almost 90%). Table 3 provides a summary of the variations in validation accuracy brought on by image distortions.



(a) (b)

Figure. 3 shows the history of the VGG19 Model's training and validation accuracy on original (a) and noise-filled (b) pictures, respectively.

However, Precision, Recall, and F1-score calculations evaluated the classifications' quantitative and qualitative performance to confirm that the VGG19 model achieved the study's objectives. To assess the severity of the poor image quality, the three different sorts of photos must be carefully examined.

Table 3. Performance metrics of different CNN architectures.

Datasets		Models						
		Proposed model	**ResNet50 [9]**	**EfficientNetB0 [10]**	**VGG19 [23]**	**InceptionV3 [25]**	**TMC**	**CNN+ WARP**
ADE20K	P	**89.15**	**78.19**	**92.91**	**88.53**	**89.95**	**-**	**53**
	R	**89.54**	**62.12**	**90.35**	**89.28**	**89.53**	**-**	**60**
	F1	**88.60**	**60.11**	**91.27**	**88.51**	**89.72**	**-**	**56**
Noise Images	P	**91.05**	**34.20**	**91.47**	**90.81**	**87.13**	**-**	**52**
	R	**86.78**	**45.00**	**89.50**	**87.17**	**87.37**	**-**	**58**
	F1	**87.97**	**35.83**	**90.26**	**88.27**	**87.18**	**-**	**55**

The results are confirmed, and confusion matrices were constructed, providing us counts of TP, TN, FP, and FN, respectively, in accordance with the performance of the VGG19 model and the two various types of images (Original, Noise). Regarding pre-trained implemented ResNet, EfficientNet0, and our suggested architecture for VGG19, Table 3 displays the confusion matrix that was developed. The values produced by the confusion matrices were then used to construct the precision, recall, and F1-score performance measures. Table 3 summarizes the performance characteristics of the deployed CNN architectures for identifying actual objects, taking into consideration the performance of well-behaved AIA models as well as the influence of distortions.
The ROC (Receiver Operating Characteristics) Curve is a graphical representation of the false positive rate (FPR) against the true positive rate for a variety of classification models with threshold values between 0.0 and 1.0. (TPR). It offers a probability curve that gauges the effectiveness of classification models. If a classification model's ROC is larger, it is better. For each image, tags determined by AIA (right panel) and hand annotation (left panel) are provided so that they may be compared. Nearly all generative model-based AIA approaches are inferior to other types of AIA methods because generative models are unable to capture the intricate relationships between image features and labels. Despite the numerous benefits of tag completion-based AIA approaches (such as their robustness to noise and lack of a training phase), their performance for annotating pictures is not sufficient. Tag completion-based AIA, however, still offers a lot of potential as a new area of study. By combining DML and graph learning, the closest neighbor model-based AIA techniques outperform other methods in terms of annotation performance. In comparison to other approaches, the deep learning-based AIA algorithms perform the best since they can gather substantial characteristics. Additionally, the AIA approaches based on deep learning often annotate a picture with a predetermined amount of class labels. The paper calculated the real values of ROC and AUC to find the following results:

1. InceptionV3 has shown exceptional performance, with P values of 88.53 and 90.81 for original, noisy images.
2. ResNet50 deprived low values, i.e. 78.19, 34.20, for original, noisy, and blurred pictures. With models with AUC values between 34.20 and 1, there is a good chance that the classifier will really want to separate the positive class values from the negative class values. ResNet50 underperforms, as seen by various graphs and performance statistics, with an accuracy of 63% and weak classification metrics in contrast to others.
3. The EfficientNetB0 and InceptionV3 models significantly outperformed ResNet in accuracy while requiring less parameters, however ResNet50 is unable to correctly categorize distorted pictures.
4. It correctly classified distorted photos with enhanced performance metrics, proving that the model's classification procedure is unaffected.

7. Conclusion & future works

In this work, pertained Deep Learning models were used to the task of subject detection. With the original and deformed pictures, several CNN architectures were explored to determine the effects of distortions like blur and noise. The ADE20K dataset, which includes examples of objects, things, and parts that cover a wide variety of visual ideas in scenes, was presented in this work. The dataset was painstakingly annotated by a single annotator to ensure precise item boundaries inside the picture and consistent object names throughout the photographs. Many models, including ResNet50, VGG16, VGG19, InceptionV3, and EfficientNetB0, were trained using the dataset, which is split into two groups. The ResNet50 model has low performance from distortions in comparison. However, EfficientNetB0 outperforms the installed CNN. In the future, the labelled approach will display a sentence rather than a single word and provide additional information.

Declaration of competing interest

The authors declare that they have no known financial or non-financial competing interests in any material discussed in this paper.

Funding information

No funding was received from any financial organization to conduct this research.

Acknowledgements

The authors are grateful to the Iraqi Ministry of Higher Education and Scientific Research (MOHESR) for technically supporting the current research.

References

[1] B. K. A. A. R. M. Nizar Zaghden, “Text Recognition in both ancient and cartographic documents,” vol. 1308, no. 6309, 2013..

[2] J. Chen, P. Ying, X. Fu, X. Luo, H. Guan and K. Wei, “Automatic tagging by leveraging visual and annotated features in social media,” IEEE, vol. 9210, pp. 1-12, 2021.

[3] A. Stangl, M. Morris and D. P. S. Gurari, “, Tree. Is the Person Naked? What People with Vision Impairments Want in Image Descriptions.,” in In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 2020.

[4] H. Ben, Y. Pan, Y. Li, T. Yao, R. Hong, M. Wang and T. Mei, “Unpaired Image Captioning with Semantic-Constrained Self-Learning,” IEEE Trans. Multimed. , 2021.

[5] N. Zaghden, S. Ben Moussa and A. M Alimi, "Categorizing ancient documents", International Journal of computer science, issues 2, 2013

[6] C.Ambrosi and T. Strozzi, “Study of landslides in Ticino: Photointerpretation and analysis of deformations with satellite radar interferometry,” Bull. Of the Ticino Society of Sci. Nat., pp. 19-27, 2008.

[7] F. a. W. J. El-Baz, Remote Sensing In Archaeology, New York: Springer Books, 2007.

[8] L. L. J. N. J. G. I. a. J. L. E. Kurimo, “The Effect of Motion Blur and Signal Noise on Image Quality in Low Light Imaging,” in in Proc. Scandinavian Conference on Image Analysis,, 2009.

[9] M. HAMOUDA and M. S. BOUHLEL, “Modified Convolutional Neural Networks Architecture for Hyperspectral Image Classification (Extra-Convolutional Neural Networks),” IET Image Processing, vol. DOI: 10.1049/ipr2., no. 12169, pp. 1-8, 2021.

[10] D. S. a. S. Kulkarni, “Different types of Noises in Images and Noise Removing Technique,” International Journal of Advanced Technology in Engineering and Science, vol. 03, no. 01, pp. 50-62, 2015.

[11] S. J. a. S. Goswami, “A Comparative Study of Various Image Restoration Techniques with Different Types of Blur,” International Journal of Research in Computer Applications and Robotics, vol. 3, no. 11, pp. 54-60, 2015.

[12] Q. Cheng, Q. Zhang, P. Fu, C. Tu and S. Li, “A survey and analysis on automatic image annotation,” Pattern Recognition, vol. 79, pp. 242-259, 2018.

[13] R. M. A. M. Nizar ZAGHDEN, “A proposition of a robust system for historical document images indexation,” International Journal of Computer Applications, vol. 11, pp. 224-235, 2010.

[14] A. Hanbury, “A survey of methods for image annotation,” Elsevier, vol. 19, p. 617–627, 2008.

[15] I. I. Marina Ivasic-Kos, “A Lightweight Network for Building Extraction from Remote Sensing Images,” IEEE, Vols. 0196-2892, pp. 99-112, 2021.

[16] Jian Kang, “PiCoCo: Pixelwise Contrast and Consistency Learning for Semisupervised Building Footprint Segmentation,” IEEE, vol. 14, pp. 10548 - 10559, 2021.

[17] S. Ammar, T. Bouwmans, Nizar ZAGHDEN and M. Neji, “Deep detector classifier (DeepDC) for moving objects,” IET Image Processing, vol. 14, no. 1212, pp. 1490-1501, 2020.

[18] K. P. Ferentinos, “Deep learning models for plant disease detection and diagnosis,” Computers and Electronics in Agriculture, vol. 145, pp. 311-318., 2018.

[19] V. J. P. A. P. L. B. N. S. a. S. C. G. P. Mahajan, “Perceptual Quality Evaluation of Hazy Natural Images,” IEEE Transactions on Industrial Informatics, vol. 10.1109/TII.2021.3065439, 2021.

[20] E. S. ,. T. D. Jonathan Long, “Fully Convolutional Networks for Semantic Segmentation,” IEEE, pp. 3431-3440, 2015.

[21] I. S. a. G. E. H. A. Krizhevsky, “Imagenet classification with deep convolutional neural networks,” NIPS, vol. 3, no. 5, pp. 1-2, 2012.

[22] K. S. a. A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, abs/1409.1556, vol. 4, no. 5, pp. 1-3, 2014.

[23] W. L. Y. J. P. S. S. R. D. A. D. E. V. V. a. A. R. C. Szegedy, “Going deeper with convolutions.,” CoRR, vol. 1409, no. .4842,, pp. 1-3, 2014.

[24] Nizar ZAGHDEN, M. B. MS Jasim, “Identified of Collision Alert in Vehicle Ad hoc based on Machine learning,” 2021.

[25] J. S. E. D. T. Long, “ Fully convolutional networks for semantic segmentation,” proceedings of IEEE conference on Computer Vision and pattern recognition, pp. 3431-3440, 2015.

[26] A. Krizhevsky, I. Sutskever and G. Hinton, “Imagenet classification with deep convolutional neural networks communication,” ACM, vol. 60, p. 84–90, 2017.

[27] M. T. a. Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in International Conference on Machine Learning, 2019.

[28] Q. Chenga, Q. Zhang, P. Fu, C. Tu and S. Li, “A survey and analysis on automatic image annotation,” ELSEVIER, vol. 79, pp. 242-259, 2018.

[29] J. J.-K. P. K. M. G. a. M. B. A. Brodzicki, Pre-Trained Deep Convolutional Neural Network for Clostridioides Difficile Bacteria Cytotoxicity Classification Based on Fluorescence Images, sensors, 2020.

[30] M. T. a. D. Hanbay, “Plant disease and pest detection using deep learning-based features,” Turkish Journal of Electrical Engineering & Computer Sciences, pp. 1636-1651., 2019.

[31] X. Guo, “Comparison and evaluation of annual NDVI time series in china,” NOAA AVHRR LTDR and terra MODIS mod13c1 products., 17 7 2017. [Online].

[32] Q. W. D. C. Y. Z. ,. W. L. ,. G. D. X. L. Wei Wei, “Automatic image annotation based on an improved nearest neighbor technique with tag semantic extension model,” Elsever, Vols. 1877-0509, pp. 616-623, 2021.

[33] X. Yang, “Pixel-level automatic annotation for forest fire image,” Elsevier, pp. 1-14, 2021.

[34] N. W. D. R. C. J. W. M. M. a. C. W. Calum R. Wilson, “Receiver Operating Characteristic curve analysis determines association of individual potato foliage volatiles with onion thrips preference,” cultivar and plant age, pp. 5-9, 2019.

[35] F. P. J. B. N. B. M. B. N Jaouedi, “Prediction of Human Activities Based on a New Structure of Skeleton Features and Deep Learning Model,” Sensors, vol. 17, no. 4944, pp. 1-15, 2020.