International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.7532
***Article Received**: 26 April 2023 **Revised:** 18 June 2023 **Accepted:** 30 June 2023*

Deep Learning Based Fine Grained Image Classification

Priti P. Vaidya[1], Dr. S. M. Kamalapur[2]
[1]Research Scholar, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, Savitribai Phule Pune University, Pune, Maharashtra, India, ppvaidya@kkwagh.edu.in
[2] Professor and Research Guide, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, Savitribai Phule Pune University, Pune, Maharashtra, India, smkamalapur@kkwagh.edu.in

Abstract : Image classification, specifically object classification is the focused research area in the computer vision and machine learning field in the past decade. In image classification a label or category is assigned to an input image based on its content. With breakthroughs in deep learning-based approaches, performance of image classification models' has improved significantly, particularly fine-grained image classification, which includes discriminating between items of the same category with slight changes. The object classification can be categorised as coarse grained object classification, which identifies highly diverse object categories, such as an elephant and a bus. One example of this type of object classification is a bus and an elephant. On the other hand, fine-grained image categorization seeks to recognise photos as belonging to distinct species of animals, birds, or plants, as well as distinct models of automobiles, versions of aircraft, and so on. The purpose of this study is to evaluate previously published research that investigates deep learning techniques for the classification of fine-grained images and to compare the effectiveness of these techniques using datasets that are open to the public.

Keywords: Computer Vision, Machine Learning, Fine-grained image classification, Coarse-grained image classification, Object classification.

I. INTRODUCTION

In the field of Computer Vision, image classification plays an essential role in a wide variety of applications, including image retrieval, object detection and identification, scene comprehension etc. The purpose of image classification is to do an automated analysis of the visual content of a picture in order to identify any recognizable patterns or objects that may be included within it.

Traditional approaches based on machine learning have used the handmade features, such as SIFT or HOG, to describe the picture content. These features are used by a classifier to predict the image labels. To get high performance with these approaches requires professional knowledge and substantial feature engineering. Moreover, it is possible that these methods are not resistant to fluctuations in illumination, size, and orientation.

Image classification, specifically object classification is the field of the computer science that deals with classifying the objects. It may either classify an object in a picture or make an educated guess as to what kind of object it is based on its appearance. For instance, an image that depicts a Siberian husky is categorised as belonging to the canine category. Numerous apps that are used in the real world that are based on object identification have been developed for the aim of automatically tagging images, captioning images, and analysing user interests.

The ability of coarse-grained object classification to comprehend image content on a deeper level is limited. Using a coarse-grained object classification system, it is simple to determine the likelihood that an image contains a dog; however, it is more difficult to determine which breed of dog is depicted in the image. It takes a significant amount of domain expertise to construct a classifier, which is required for detecting and classifying a specific pattern that exists among multiple visually similar dog breeds.

On the other hand, there has been an increasing interest in the research of sub-category classification, also known as fine-grained classification. As a sub-field of object classification, extremely fine-grained classification is a relatively young discipline in the field of computer vision and pattern recognition. Its primary function is to categorise objects at a finer level. In contrast to coarse-grained object classification, which seeks to identify the most appropriate general category, such as a bird, dog, or plant, fine-grained picture classification seeks to categorise the specific subcategory, as seen in figure 1.

Distinguishing subcategories is a challenging problem and its solution is applicable to all similar fine-grained classification problems. Examples of fine grained problems are to identify breeds of dogs/cats and horses as well as species of birds and plants, and/or models of cars, variants of aircrafts etc.

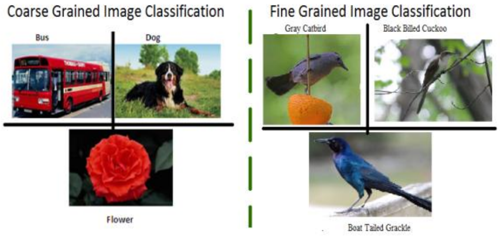
Coarse Grained Image Classification
Bus
Dog
Flower
Fine Grained Image Classification
Gray Catbird
Black Billed Cuckoo
Boat Tailed Grackle

Figure 1: Comparision of Coarse grained image classification against the Fine-grained image classification

In recent years, however has been a great deal of innovation seen in the methods that depend on deep learning for the categorization of images. These advancements include the establishment of novel network topologies, training procedures, and data augmenting techniques. These advancements have made it possible to construct models that can attain human-level performance on a variety of picture classification tasks. Also, they have opened up new potential for the use of computer vision in fields such as healthcare, autonomous driving, and robotics.

The basic objective of fine-grained item classification systems is to recognise the sub-categories that are contained within the confines of the same meta-category. Obtaining an accurate fine-grained classification for object sub-categories is a task that is difficult to do. There are two issues that contribute to the difficulty of the fine-grained picture classification problem that computer vision faces. The first significant problem originates in the fact that several classes that are part of the same category, such a bird or a dog, visually appear extremely similar. This phenomenon is referred to as low inter-class variance. The vast amount of variety in attitude, illumination, and lighting that exists within the same sub-species is referred to as intra-class variation. This presents a second issue. In order to address these issues, it has been suggested that significant efforts be made to discover local locations and characteristics that can identify minor variations in appearance between species.

Section II focuses on fine grained image classification techniques using deep learning. Section III focuses on datasets used for fine grained classification. Section IV represenrts performance analysis on benchmark datasets followed by results and its analysis.

II. RELATED WORK

The approaches based on deep learning uses features from the unprocessed pixel values of a picture directly by utilizing deep neural networks. These methods have improved the performance of image classification tasks, particularly fine-grained image classification, which requires distinguishing between items that belong to the same category but have slight variances in appearance.

Coarse grained image classification

Target objects in generic image classification belong to the coarse-grained meta-categories, and as a result, they have a very distinct appearance from one another. Machine learning strategies are heavily included into the many methodologies that are used for coarse-grained picture classification. The performance of classification can be improved by amassing more datasets, gaining a deeper understanding of more robust models, and using more effective strategies to avoid overfitting.

The improved performance on the ImageNet dataset (2012) was achieved using the AlexNet architecture [6]. Particularly, the AlexNet architecture was able to categories photos into tens of thousands of distinct categories. The architecture of AlexNet is made up of five convolutional layers, which are followed by three fully connected layers. All of these layers work together to prevent overfitting by utilising ReLU activation and dropout regularisation in various ways across the network. When utilising a method that considers the median of the projections of two classifiers that were trained on Fisher Vectors (FVs) obtained from two different kinds of densely-sampled features, the best results that have been published were 45.7% and 25.7%, respectively. This strategy was used to get these findings.

Because of the depth of the neural networks used, the VGG architecture [7] achieves even higher accuracy than AlexNet on the ImageNet dataset. This is accomplished by combining smaller convolutional filters with up to 19 weight layers of deep neural networks. Stochastic gradient descent with weight decay is the method that is used to train VGG networks. These networks can have as many as 19 layers and frequently use 3x3 filters. It was discovered that the representation depth is advantageous for the accuracy of the classification, and it is possible to attain a better performance on the ImageNet dataset as a result.

Since 2014, convolutional networks with deep connections have demonstrated significant improvement across a variety of benchmarks. Increases in model size and the associated computing costs result in quick quality gains for the majority of jobs, at least up until sufficient labelled data is made available for training. The following table provides an overview of the results obtained by applying deep neural networks to the ImageNet and PASCAL VOC benchmark datasets, both of which are utilised for coarse-grained image categorization.

***Article Received**: 26 April 2023 **Revised:** 18 June 2023 **Accepted:** 30 June 2023*

When it comes to establishing how accurate an image categorization is, one of the most important factors is the extraction of features. The local features, such as Scale invariant feature transform (SIFT) [2], histogram of oriented gradient (HOG) [3], are extracted from the image by traditional classification algorithms. These algorithms then use Vector of locally collected descriptors (VLAD) [4] or Fisher vector[5] code model to perform feature encoding in order to get the final required feature representation. On the other hand, because of our restricted ability to characterise the characteristics, the categorization effect is frequently unsatisfactory.

Table.1 Coarse grained Image Classification on ImageNet dataset

Year	Models with highest map	Classification accuracy
2012	AlexNet	63.30%
2014	VGG	74.40%
2015	ResNet-152	78.57%
2016	Inception ResNet V2	80.10%
2016	SimpleNet V1	81.24%
2017	NasNet –A(6)	82.70%
2017	PNASNet – 5	82.90%
2018	AmoebaNet-A	83.90%
2019	FixResNeXt-101	86.40%
2020	FixEfficientNet-L2	88.50%
2021	ViT-G/14	90.88%
2022	Coca (finetuned)	91.00%

Table.2 Object detection on PASCAL VOC dataset

Year	Models with highest map	Classification accuracy
2013	R-CNN	58.50%
2015	Fast R-CNN	70.00%
2015	Faster R-CNN	73.20%
2016	COCO	81.60%
2017	CoupleNet	82.70%
2019	VGG16	83.00%
2020	NAS-FPN	89.30%
2021	YOLOv4	81.80%
2022	InternImage-H	94.00%

Fine grained image classification

In recent years, the exploration of Fine-grained Image Classification, additionally referred to as FGIC, has been an important topic of research, notably in the subfields of machine learning and recognition of patterns. Approaches developed by FGIC can be utilised in a diverse array of domains, including person/vehicle re-identification, intelligent retail, autonomous biodiversity monitoring, and intelligent transportation, amongst others. These approaches have also resulted in a positive impact in a number of areas, including face recognition, ecosystem conservation (recognising biological species), and e-commerce.

The previous work in fine-grained classification can, to a large extent, be split in two different directions. The initial step is to identify the parts of the image that are discriminative of the objects in order to adjust for fluctuations that are a nuisance, such as attitude. For the classification of birds, cars, aircrafts, and dogs, numerous approaches based on component parts and including geometric limitations have been developed. Some of the efforts directly use parts annotations from the dataset to train a highly supervised parts detector, with the goal of minimising the effect that position and viewpoint have on the results. These methods, on the other hand, frequently call for ground-truth bounding boxes of the bird's location in addition to annotations that provide the location of particular fascination regions.

The second strategy is to derive features that are both robust and discriminatory. The conventional method hand-crafted feature descriptors, which include the Scale Invariant Feature Transform (SIFT), the Histogram of Oriented Gradients (HoG), and the Colour Histogram, are being successfully adopted for application in fine-grained classification, where they make use of colour, texture, and edge information. This is due to the fact that these feature descriptors are able to take advantage of the information presented by the image. Other methods, such as the Part-based One-vs-One Features (POOFs), concentrate on modelling the activation of matching components and have been specifically developed for fine-grained categorization.

Fine-grained object categorization has reached state-of-the-art performance thanks to deep convolutional neural network (DCNN) techniques. These approaches can train very resilient image features. The majority of the techniques, including AlexNet, VGG, ResNet, DenseNet, GoogleNet, and others, utilised DCNN in order to directly classify images.

The fine-grained classification approaches (Figure 2) are organized into two paradigms-

1. Strongly Supervised Learning
2. Weakly Supervised Learning

***Article Received**: 26 April 2023 **Revised:** 18 June 2023 **Accepted:** 30 June 2023*

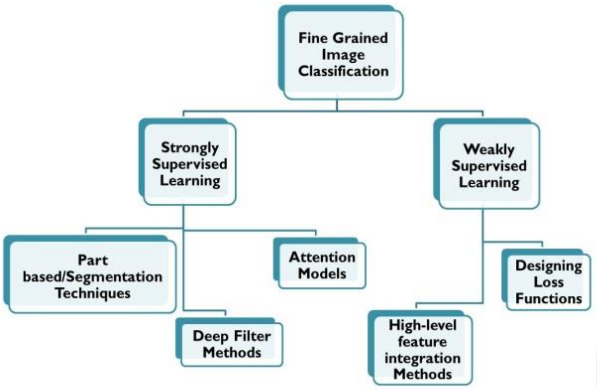
Fine Grained Image Classification
Strongly Supervised Learning
Weakly Supervised Learning
Part based/Segmentation Techniques
Attention Models
Deep Filter Methods
High-level feature integration Methods
Designing Loss Functions

Figure 2: Fine Grained Image Classification approaches

Figure 2 shows that both of the fine-grained image techniques make use of the annotation information, which includes picture labels, bounding boxes, part annotations, and so on.

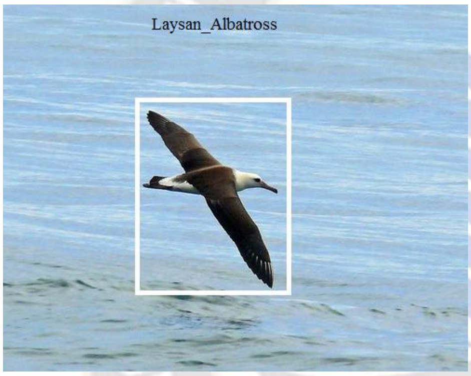

Figure 3. Bounding Boxes

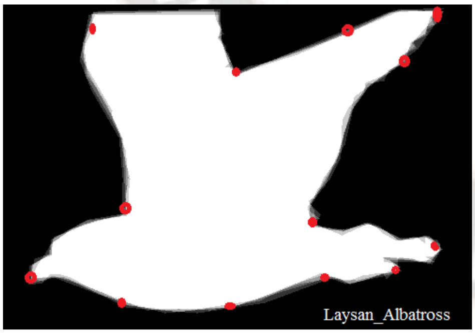

Figure 4. Part annotations

1. Strongly Supervised Learning

Strongly Supervised Learning makes use of the annotation information that is included in the datasets in order to offer the most accurate description of the object feasible. The two types of annotation information that are utilised the most frequently are object bounding boxes and part annotations. Here, bounding boxes represent discriminative picture patches (Figure 3), and part annotations detect the parts of an object that contain finer features (Figure 4). Figure 3: Bounding boxes show discriminative image patches. Methods for strongly supervised learning can be broken down into three categories: i. techniques that are part-based or segmentation-based; ii. methods that use deep filters; and iii. methods that use attention models.

i. Part based or Segmentation Techniques

The part-based or segmentation approaches [8, 9], [10] are used to zero in on significant and localised picture regions of a fine-grained object, such as bird heads and tails, automobile lights, dog ears and torsos, and so on. These techniques are used to focus on crucial image parts of the item. In order to localise and segment the objects, these methods have utilised large capacity CNN architectures to bottom-up component proposals. Because of this, it has the potential to improve the model's capacity for learning, which will ultimately result in a significant increase in performance.

An architecture for fine-grained recognition was developed by Branson et al. [11], and it computes an approximation of the pose. This approximation of the pose is then utilised to figure out local level picture features, which are then used for classification. The pose is responsible for the localization and standardisation of the calculated characteristics.

Zhang et al. [12] first learns the entirety of the object, and then makes use of the part-based annotations that are provided in the datasets. After that, the features are extracted from the local semantic portions in order to create a pose-normalized representation, and after that, a classifier is trained in order to accomplish the ultimate object recognition.

Further, Di et al. [13] suggested a Valve Linkage Function (VLF) that makes use of a sub-network that connects localization and classification modules and refines local level information based on the findings of part alignment. A CNN architecture that incorporates semantic part detection and abstraction was developed by SPDA-CNN [14]. One of this network's sub-networks is dedicated to detection, and the other is dedicated to recognition. In order to anticipate part localisation, faster R-CNN is utilised.

Stacked-CNN Components The fully convolutional network was utilised in [15] in order to locate various object pieces. Additionally, a two-stream classification network that simultaneously encodes object-level and part-level characteristics was constructed. This network was adopted. Mask-CNN [16] made use of segmentation models, which concentrate on the representation of the more minute parts of the image.

In recent times, several algorithms have done various feature fusion procedures, such as LSTMs [17], graphs [18], or knowledge distillation [19]. These strategies produce recognition accuracy that is much greater than average.

Table 3: Part based and/or segmentation techniques on Benchmark Datasets (Birds (CUB200-2011 [37]), Dogs (Stanford Dogs [38]), Cars (Stanford Cars [39]), and Aircrafts (FGVC Aircraft [40])

Method	Image Size	Datasets
Part Based –RCNN using AlexNet (2014)	224*224	CUB200_2011
Part Stacked – CNN using CaffeNet (2016)	227*227	CUB200_2011
Mask – CNN using VGG 16 (2018)	448*448	CUB200_2011
HSNet using GoogLeNet (2017)	224*224	CUB200_2011, Stanford Cars
Graph-propagation based Correlation Learning using ResNet-50 (2020)	448*448	CUB200-2011, Stanford Cars, FGVC Aircraft
Filtration & Distillation using ResNet-50 (2020)	448*448	CUB200-2011, Stanford Dogs, Stanford Cars, FGVC Aircraft

ii. Deep Filter Methods

The computer is given the ability to learn from visual samples and derive internal representations through the use of a deep convolutional neural network [20]. of the field of computer vision, visual descriptors are descriptions of the visual aspects of the contents of images that provide such descriptions. Image descriptors are a subset of visual descriptors.

Deep neural networks can be utilised in fine-grained classification tasks without the need for any part level annotations for the purpose of representing the object. Instead, filter outputs are utilised in order to act as part detectors [21], [22], [23], [24], [25], [26], [27], [28].

A method for automatic fine-grained recognition called Picking Deep Filter Responses [29] was proposed. This method does not make use of any part annotation. PDFS investigates a unified architecture that is built on two processes of picking deep filter responses. In the first stage, they uncover certain patterns and develop a set of part detectors by iteratively alternating between the mining of new positive samples and the retraining of part models. In the second stage, deep filter responses are incorporated into the final representation. This step also collects these replies. This technique identifies regions of the image that are distinct and consistent based on their portion.

Table 4: Deep filter methods on Benchmark Datasets (Birds (CUB200-2011 [37]), Dogs (Stanford Dogs [38]), Cars (Stanford Cars [39]), and Aircrafts (FGVC Aircraft [40])

Method	Image Size	Datasets
Two-level attention model in DCNN using VGG-16 (2015)	Not given	CUB200-2011
Picking Deep Filter Responses using VGG -16 (2016)	Not given	CUB200-2011, Stanford Dogs,
Discriminative Filter bank within CNN using VGG – 16 (2018)	448*448	CUB200-2011, Stanford Cars, FGVC Aircraft
Selective Sparse Sampling using ResNet – 50 (2019)	448*448	CUB200-2011, Stanford Cars, FGVC Aircraft

iii.Using attention Models

The work that has been done in the past on fine-grained approaches has demonstrated robust classification performance; nevertheless, the primary limitation of these methods is that they require correctly described object pieces. Utilising attention models is yet another method that can be used to locate pieces. Because of this, CNNs are able to attend fine-grained objects in regions that are only vaguely defined. [30] The human eye is able to more accurately capture the visual structure of an object. Iteratively generating area attention maps on the basis of past forecasts is the goal of RA-CNN [31].

A multi-attention convolutional neural network, or CNN, was used in the part learning strategy that was proposed in MACNN [32]. The classification network assigns each component of an image to the appropriate category and generates additional distinguishing characteristics. Multi-level attention models have been proposed by Peng et al. [33] and Zheng et al. [34] in order to acquire information about hierarchical attention levels. X. He and colleagues [35] used multilevel attention to locate numerous discriminative regions simultaneously for each image through the use of an n-pathway, which in turn produced information that was diverse and complementary.

Table 5: Using attention models on Benchmark Datasets (Birds (CUB200-2011 [37]), Dogs (Stanford Dogs [38]), Cars (Stanford Cars [39]), and Aircrafts (FGVC Aircraft [40])

Method	Image Size	Datasets
Recurrent Attention CNN using VGG -19 (2017)	448*448	CUB200-2011, Stanford Dogs, Stanford Cars
Object Part Attention Model using VGG – 16	Not Given	CUB200-2011, Stanford Cars

(2018)		
Trilinear Attention Sampling Network using ResNet-50 (2019)	224*224	CUB200-2011, Stanford Cars
Progressive Attention Networks using VGG -19 (2020)	448*448	CUB200-2011, Stanford Cars, FGVC Aircraft

Zheng et al. [36] developed a trilinear attentiveness sampling network. This network of neurons learns finer details from a number of part level suggestions and places the learnt features into a single CNN. In the case of small scale data, it has a tendency to over-fit.

2. Weakly Supervised Learning

When classifying fine-grained images, Weakly Supervised Learning merely makes use of the category labels provided by the training set. At neither the training nor the testing stage do we make use of any manual annotations. The many techniques to Weakly Supervised Learning can be broken down into the following categories: i) High-level features integration methods. ii) Creating loss functions through design.

i. High-level feature integration methods

When representing an image, earlier methods of deep learning relied on the characteristics of layers that were fully connected. After some time had passed, the feature maps of deeper convolutional layers of data were formed [41]. These feature maps contain information on both the mid-level and the high-level. Furthermore, some encoding techniques were used which resulted in substantial improvements [42], [43], [44].

The high-order feature integration methods uses covariance matrix as a region descriptor [45], [46]. The covariance matrix-based format with deep descriptors has shown promise accuracy in fine-grained recognition during the past few years.

Table 6: High-level feature integration methods on Benchmark Datasets (Birds (CUB200-2011 [37]), Dogs (Stanford Dogs [38]), Cars (Stanford Cars [39]), and Aircrafts (FGVC Aircraft [40])

Method	Image Size	Datasets
Bilinear CNN using VGG-16 + VGG-M (2015)	448*448	CUB200-2011, Stanford Cars, FGVC Aircraft
Compact Bilinear pooling using VGG-16 (2016)	448*448	CUB200-2011, Stanford Cars, FGVC Aircraft
Low-Rank using VGG-16 (2017)	224*224	CUB200-2011, Stanford Cars, FGVC Aircraft
Hierarchical Bilinear pooling using VGG-16 (2018)	448*448	CUB200-2011, Stanford Cars, FGVC Aircraft
Deep Bilinear Transformation using VGG-16 (2019)	448*448	CUB200-2011, Stanford Cars, FGVC Aircraft
Multi-Objective Matrix Normalization using VGG-16 (2020)	448*448	CUB200-2011, Stanford Cars, FGVC Aircraft

Using Bilinear Convolutional Neural Networks [47, 48], an image is fed into two CNNs, and their outputs at each location are merged using the matrix outer product to obtain the bilinear feature representation. The average of the pooled outputs is then used to create the bilinear representation. In order to obtain class predictions, this is first processed by a linear layer, and then by a softmax layer. However, the outside product operation leads to exceptionally high dimensional characteristics, which might lead to over-fitting and make the product unusable for applications that are based in the actual world.

Gao et al.'s [49] research provides a solution to this issue. They proposed two compact bilinear representation having the same discriminative capacity as the full bilinear representations but with only a few thousand dimensions [50]. These representations are compact since they only have a few thousand dimensions. Kong et al. [51] implemented a low-rank bilinear classifier and presented the features in the form of a covariance matrix. The classifier that was created as a consequence was tested, and the results showed that not only was the amount of computation time cut significantly, but also the number of parameters that needed to be learned.

ii. Designing loss function

The accuracy of the neural network's representation of the training data is evaluated using a loss function, which does a comparison of the target and predicted output values. It is essential to design loss functions for fine-grained image recognition since the items that are subjected to fine-grained classification appear to be highly similar to one another.

Following this, Dubey et al. [52], [53] tackled the over fitting problem using the Pairwise Confusion optimisation function. They accomplished this by bringing closer together the various class-dependent probability distributions. This confused the deep network, which ultimately led to improved generalisation performance.

Sun et al. [54] came up with the idea for a one-of-a-kind attention-based convolutional neural network that controls different object parts over a wide range of input images.

A discriminality element and an array of characteristics component [55] make up mutual-channel loss. These two channel-specific components are what make up mutual-channel loss. Both of these components are unique to their respective channels. At the conclusion of the procedure, a collection of characteristics will be generated that reflect multiple geographically discriminative patches for a certain class.

Table 7: Loss functions methods on Benchmark Datasets (Birds (CUB200-2011 [37]), Dogs (Stanford Dogs [38]), Cars (Stanford Cars [39]), and Aircrafts (FGVC Aircraft [40])

Method	Image Size	Datasets
Maximum Entropy using Bilinear CNN (2018)	224*224	CUB200-2011, Stanford Dogs, Stanford Cars, FGVC Aircraft
Pairwise Confusion using Bilinear CNN (2018)	224*224	CUB200-2011, Stanford Dogs, Stanford Cars, FGVC Aircraft
Channel Interaction Network using ResNet-101 (2020)	448*448	CUB200-2011, Stanford Cars, FGVC Aircraft
Mutual Channel-Loss using Bilinear CNN (2020)	448*448	CUB200-2011, Stanford Cars, FGVC Aircraft

III. FINE GRAINED IMAGE CLASSIFICATION DATASETS

In recent years, members of the vision community have made available a large number of fine-grained benchmark datasets spanning a wide variety of subject matter, including but not limited to animals, canines, automobiles, aircraft, flowers, vegetables, fruits, meals, clothing, and retail goods, amongst other things. In addition, it is important to note that even the most well-known large-scale picture classification dataset, known as ImageNet, has fine-grained classifications that span a significant number of distinct sub-categories pertaining to canines and avians.

Table.8 Publically available Fine-Grained Image Classification Datasets

Dataset	Classes	Total Images	Description
CUB-200-2011	200	11,788	A dataset of birds with fine-grained categories and bounding boxes for object localization.
Stanford Dogs	120	20,580	A dataset of dogs with fine-grained breed categories.
Oxford Flowers	102	8,189	A dataset of flowers with fine-grained categories.
FGVC-Aircraft	100	10,000+	A dataset of aircraft with fine-grained categories and varying viewpoints.
Food-101	101	101,000+	A dataset of food items with fine-grained categories.
iNaturalist	8,142	437,513	A dataset of natural images with fine-grained categories and bounding boxes for object localization.
NABirds	555	48,562	A dataset of North American birds with fine-grained categories and bounding boxes for object localization.
Stanford Cars	196	16,185	A dataset of cars with fine-grained make and model categories.
LeafSnap	185	30866	A dataset of tree leaves from the Northeastern United States.
Flavia	32	1907	A dataset of highly constrained leaf images.

DATASET SPECIFICATION AND COLLECTION

Caltech-UCSD Birds-200-2011 Dataset

The classification of bird species is a challenging subject that stretches the visual capabilities of both people and computers to their absolute limits. Even though all bird species are composed of the same fundamental anatomical components, the shapes and external appearances of different bird species can be extremely dissimilar to one another (compare pelicans to sparrows, for example). At the same time, there are some pairs of bird species that are practically impossible to tell apart visually, even for experienced bird watchers (for example, many species of sparrows have very similar appearances). There is a large amount of variety within the class itself as a result of differences in lighting and background, as well as a significant amount of variation in position (for example, there are birds that are soaring, birds that are swimming, and perched birds that are partially hidden by branches).

DOI: https://doi.org/10.17762/ijritcc.v11i7s.7532
***Article Received**: 26 April 2023 **Revised:** 18 June 2023 **Accepted:** 30 June 2023*

A hard dataset consisting of 200 different species of birds, CUB-200-2011 is an expanded version of the original CUB-200. The number of photos included in each category has approximately been increased by a factor of two in the extended version, and new annotations pertaining to part localisation have also been included. All of the photos have bounding boxes, part positions, and attribute label annotations added to them. Multiple users on Mechanical Turk screened the images and annotations to remove unwanted content.

Bird Species: The dataset include 11,788 photos representing 200 different kinds of birds. Each species has its own article on Wikipedia, which is ordered according to its position in the scientific categorization system (order, family, genus, and species). The names of the species were gathered with the assistance of an online field guide. The images were obtained through the use of the Flickr picture search, and then they were filtered by displaying each image to a number of participants on Mechanical Turk. Each image has labels indicating the bounding box, the location of the parts, and the properties of those parts. Figure 5 provides some examples of the dataset's accompanying images.

Bounding Boxes: It is a box that surrounds the object in the image and is referred to as the bounding box. The box holds not only information regarding the object but also coordinates, which carry information regarding the location of the object inside the image.

Attributes: An online tool for the identification of bird species served as the basis for the selection of a lexicon consisting of 28 attribute groupings and 312 binary characteristics (for example, the attribute group belly colour comprises 15 different colour options). All characteristics are visual in origin, with the majority of traits referring to a colour, pattern, or shape of a certain component.

Figure 5: Sample images from CUB-200-2011

Part Locations: There were a total of 15 sections that were annotated, and their visibility in each image was based on the pixel location. The "ground truth" part positions were determined by taking the median of the five different user locations for each image that was submitted to Mechanical Turk.

Stanford Dogs Dataset

One of the standard data sets that is utilised in the process of fine-grained picture categorization is the Stanford Dogs dataset. Images of 120 different dog breeds from around the world are included in the dataset. There are a total of 20,580 photos of various canine breeds available here. There is information provided in the form of annotations, such as class labels and bounding boxes. For the purpose of fine-grained picture categorization, this dataset has been constructed with the assistance of images and annotations taken from ImageNet. Initially, it was gathered for the purpose of fine-grained image categorization, which is a difficult challenge to solve due to the fact that several dog breeds share nearly similar characteristics or differ in colour and age. Figure 6 provides some examples of the dataset's accompanying images.

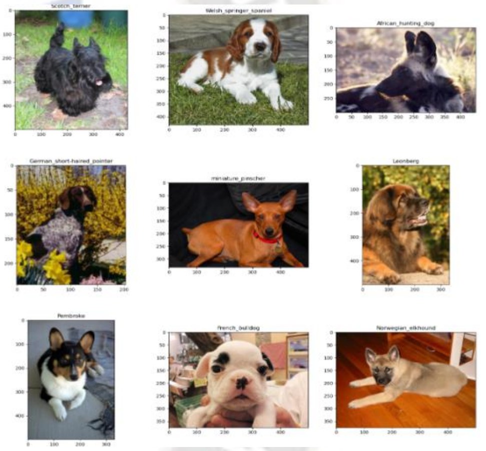

Figure 6: Sample images from Stanford Dogs dataset

BRCars Dataset

The high class imbalance necessitated the creation of two sets in order to do an evaluation of the FGVC problem. These sets are referred to as BRCars-196 and BRCars-427. The datasets for BRCars can be accessed at this location: https://github.com/danimtk/brcars-dataset.

A. BRCars-196 set

Each of the 196 classes in BRCars-196 corresponds to a certain make and model of automobile. To begin the process of putting together this set, we started by choosing just the models that contain at least 200 different instances of automobiles. Following that, 200 different cases were chosen at random for each car model. In the end, each of the 200 groups of photographs that were selected at random from the total of 200 sets was put together to form the images of each model. The most popular car models were chosen to be

DOI: https://doi.org/10.17762/ijritcc.v11i7s.7532
***Article Received**: 26 April 2023 **Revised:** 18 June 2023 **Accepted:** 30 June 2023*

included in this collection as a consequence of the selection method that was just described. There is a little variance in the total number of photographs that correspond to each model due to the fact that the total number of images produced by each car instance is different. After the removal of photos that contained noise, BRCars-196 has a total of 212,609 images, of which 170,151 are designed for training and 42,458 are designed for testing.

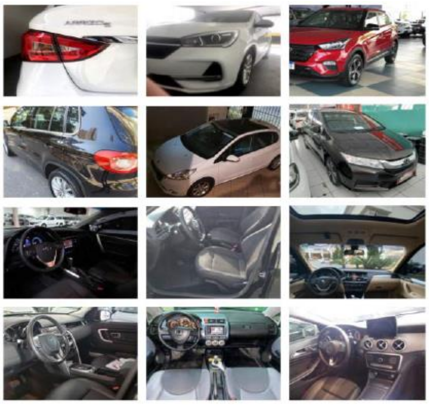

Figure 7: Samples of the images from the BRCars dataset

B. BRCars-427 set

BRCars-427 is comprised of BRCars-196 and an additional 231 classes, all of which refer to models with an instance count of fewer than 200 cars. We rejected models that had fewer than 20 instances in order to get rid of classes that had an abnormally low number of occurrences. This allowed us to get rid of classes that were significantly underrepresented. These additional 231 classes have a wide range of pictures across their entirety. The addition of classes that have fewer occurrences has been done with the intention of simulating the difficulty of dealing with rarer models. Following the completion of the noise removal process, BRCars-427 is made up of a total of 300,325 photos, of which 239,668 are designed for training purposes and 60,657 are intended for testing purposes. Figure 7 provides some examples of the dataset's accompanying images.

FGVC Aircraft

The FGVC-Aircraft database has 10,000 photos of different aircraft, each of which has been annotated with the model and bounding box of the predominant aircraft in the image. Only the third, fourth, and fifth levels of the hierarchy that organises aircraft models are relevant here. This hierarchy contains four levels.

Model: This is the class label that is the most exact, for example, Boeing 737-76J. This level is not regarded useful for FGVC since the variations between models might not be visually apparent, at least not provided an image of the outside of the aeroplane. since of this, this level is not considered relevant for FGVC.

Variant: Model variations are a finer differentiation level that may be visually detected. Model variants were generated by combining visually indistinguishable models. For instance, the variant Boeing 737-700 includes 87 models such as 737- 7H4, 737-76N, and 737-7K2, amongst others. The dataset has 100 variants.

Family: The versions of the model that differ from one another in oblique ways, which magnify the discrepancies across families. The creation of a categorization activity with an intermediate level of challenge is the objective of this level. For instance, the family Boeing 737 includes the types 737-200, 737-300, and so on up to 737-900; there are a total of 70 families in the dataset.

Manufacturer: A gathering of families that are all produced by the same company is referred to as a manufacturer. For instance, Boeing is home to the 707, 727, and 737 families of aeroplanes. There are thirty distinct manufacturers represented in this dataset's collection of aeroplanes.

Figure 8 provides a list of model variants together with the sample photos that correspond to them. FGVC-Aircraft contains one hundred example images for each of the one hundred model variants it supports. The image resolution is somewhere between 1 and 2 mega pixels. The quality of the photographs varies due to the fact that they were taken over a range of decades, but they are typically extremely nice. Because the dominant aircraft is typically well centred, this makes it easier to concentrate on fine-grained discrimination as opposed to object identification. The images are then randomly distributed among three subsets: one for training, one for validation, and one for testing. Each subset has either 33 or 34 images for each variant. In order to prevent algorithms from being overfit, they should be built using the training and validation subsets before being tested merely once using the test subset. The information contained in the bounding boxes may be utilised for the purpose of training aircraft classifiers; however, this information must never be utilised during testing.

The dataset can be accessed by the general public at http://www.robots.ox.ac.uk/ vgg/data/fgvc-aircraft/. However, its use is restricted to academic study.

Figure 8: Sample images of FGVC aircraft dataset

ImageNet Dataset

WordNet serves as the structural foundation for ImageNet, which is a massive image ontology project known as ImageNet. WordNet contains approximately 80,000 synsets, and ImageNet's goal is to populate the bulk of these synsets with an average of 500 to 1000 high-quality images per. Because of this, tens of millions of annotated photos will be produced, all of which will be arranged according to the semantic hierarchy of WordNet.

ImageNet is constructed using the hierarchical framework offered by WordNet as a foundation. When it is finished, ImageNet will hopefully include somewhere in the neighbourhood of 50 million photos that have been neatly annotated and have a full resolution (between 500 and 1000 per synset).

Scale: ImageNet's mission is to give coverage of the visual world that is both the most extensive and varied possible. The currently active 12 sub trees include a total of 3.2 million photos that have been neatly labelled and are distributed throughout 5247 categories (Fig. 9). Over 600 photos are collected for each synset, on average. This is already the clean picture dataset that the community of vision researchers has access to that is the largest in terms of the total number of photos, the number of images contained inside each category, and the number of categories themselves.

Hierarchy: ImageNet creates a densely packed semantic hierarchy that is then used to organise the many categories of photos. The semantic structure of WordNet, often known as its ontology of concepts, is the database's most valuable feature. ImageNet's synsets of images, much like WordNet's, are connected to one another by a variety of relations, with the "IS-A" relation serving as the most complete and beneficial of the bunch.

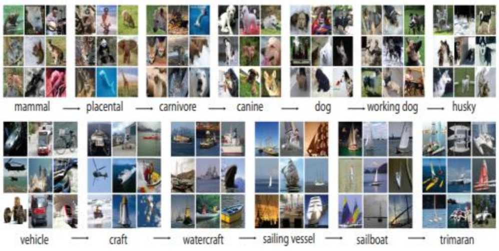

Figure 9: Sample images from ImageNet dataset

Diversity: ImageNet was built with the intention that the objects in the photographs would have a variety of different appearances, positions, views, and poses, in addition to various levels of background clutter and occlusions. In order to solve the challenging issue of quantifying image diversity, the average image of each synset was computed, and the size of the lossless JPG file was measured. The size of the file is a measure of the amount of information contained in each image. A synset that contains a variety of photos will produce an image that is more blurry on average, with a grey image serving as the most extreme example; in contrast, a synset that contains few different images will produce an image that is more structured and clearer on average.

IV. PERFORMANCE ANALYSIS

The performance parameters used for image classification algorithms are discussed in this chapter. Metrics help determine the performance of all the models that are trained.

Accuracy: Accuracy is the most commonly used metric for classification algorithms due to its simplicity. Accuracy refers to the total number of correct predictions made, divided by the total number of all predictions.

Accuracy=Correct Prediction / Total Prediction

True Positive (TP): Number of positive class samples the model predicted correctly.

True Negative (TN): Number of negative class samples the model predicted correctly.

False Positive (FP): Number of negative class samples the model predicted incorrectly. In statistical terminology, it's known as a Type-I error.

False Negative (FN): Number of positive class samples the model predicted incorrectly. In statistical terminology, it's known as a Type-II error.

Accuracy = (TP+TN) / (TP+FP+TN+FN)

Precision: Precision refers to the ratio of true positive samples predicted versus the total number of positive samples predicted.

Precision=TP / (TP+FP)

DOI: https://doi.org/10.17762/ijritcc.v11i7s.7532
***Article Received**: 26 April 2023 **Revised:** 18 June 2023 **Accepted:** 30 June 2023*

Recall: Recall refers to the ratio of true positive samples predicted against all the available positive samples. It's also known as sensitivity or hit rate.

Recall=TP / (TP+FN)

F1-Score: The F1-score metric is used to get the best of both worlds since the formula represents the harmonic mean of recall and precision.

F1-score=2*(precision * recall) / (precision + recall)

V. RESULTS AND DISCUSSION

Comparative results of fine grained recognition methods under weakly supervised and strongly supervised categories are given in Table 9, Table 10, Table 11, Table 12, and Table 13.

The classification accuracy perceived by detection and segmentation methods is given in Table 9. The major work has been carried out on CUB200-2011 dataset of bird species with accuracy reported in the range of 76% to 88%. Table 10 shows classification accuracy perceived by deep filter methods. The major work has been carried out on CUB200-2011 dataset of bird species with accuracy reported in the range of 77% to 88%. Table 11 shows classification accuracy perceived by attention models. The major work has been carried out on CUB200-2011 dataset of bird species and Cars dataset with make and models of different cars. The classification accuracy reported on the birds dataset is in the range of 77% to 88% and 92% to 93% on cars dataset.

The classification accuracy perceived by high-level feature integration methods is shown in Table 12. The work has been carried out on CUB200-2011, Stanford cars, and FGVC aircraft datasets. The classification accuracy reported on the birds dataset is in the range of 84% to 87%, 91% to 93% on cars dataset and 84% to 90% on aircrafts dataset. The classification accuracy perceived by loss function methods is shown in Table 13. The work has been carried out on all four benchmark datasets CUB200-2011, Stanford dogs, Stanford cars, and FGVC aircraft datasets.

Table 9: Comparative Fine-Grained Recognition Results of part based / segmentation techniques

Method	**Classification Accuracy**			
	Birds	**Dogs**	**Cars**	**Aircrafts**
Part Based –RCNN using AlexNet (2014)	76.4%	-	-	-
Part Stacked – CNN using CaffeNet (2016)	76.6%	-	-	-
Mask – CNN using VGG 16 (2018)	85.7%	-	-	-
HSNet using GoogLeNet (2017)	87.5%	-	93.9%	-
Graph-propagation based Correlation Learning using ResNet-50 (2020)	88.3%	-	94.0%	93.2%
Filtration & Distillation using ResNet-50 (2020)	88.6%	85.0%	94.3%	93.4%

Table 10: Comparative Fine-Grained Recognition Results of deep filter methods

Method	**Classification Accuracy**			
	Birds	**Dogs**	**Cars**	**Aircrafts**
Two-level attention model in DCNN using VGG-16 (2015)	77.9%	-	-	-
Picking Deep Filter Responses using VGG -16 (2016)	84.5%	72.0%	-	-
Discriminative Filter bank within CNN using VGG - 16 (2018)	86.7%	-	93.8%	92.0%
Selective Sparse Sampling using ResNet – 50 (2019)	88.5%	-	94.7%	92.8%

Table 11: Comparative Fine-Grained Recognition Results of using attention models

Method	**Classification Accuracy**			
	Birds	**Dogs**	**Cars**	**Aircrafts**
Recurrent Attention CNN using VGG - 19 (2017)	85.3%	87.3%	92.5%	-
Object Part Attention Model using VGG – 16 (2018)	85.8%	-	92.2%	-
Trilinear Attention Sampling Network using ResNet-50 (2019)	87.9%	-	93.8%	-
Progressive Attention Networks using VGG -19 (2020)	87.8%	-	93.3%	91.0%

Table 12: Comparative Fine-Grained Recognition Results of High-level feature integration methods

Method	**Classification Accuracy**			
	Birds	**Dogs**	**Cars**	**Aircrafts**
Bilinear CNN using VGG-16 + VGG-M (2015)	84.1%	-	91.3%	84.1%
Compact Bilinear pooling using VGG-16 (2016)	84.3%	-	91.2%	84.1%
Low-Rank using VGG-16 (2017)	84.2%	-	90.0%	87.3%
Hierarchical Bilinear pooling using VGG-16 (2018)	87.1%	-	93.7%	90.3%
Deep Bilinear Transformation using VGG-16 (2019)	87.5%	-	94.1%	91.2%
Multi-Objective Matrix Normalization using VGG-16 (2020)	87.3%	-	92.8%	90.4%

Table 13: Comparative Fine-Grained Recognition Results of loss functions methods

Method	**Classification Accuracy**			
	Birds	**Dogs**	**Cars**	**Aircrafts**
Maximum Entropy using Bilinear CNN(2018)	85.3%	83.2%	92.8%	86.1%
Pairwise Confusion using Bilinear CNN (2018)	85.6%	83.0%	92.4%	85.7%
Channel Interaction Network using ResNet-101(2020)	88.1%	-	94.5%	92.8%
Mutual Channel-Loss using Bilinear CNN (2020)	86.4%	-	94.4%	92.9%

VI. CONCLUSION

This study presents a thorough analysis of recent developments in the field of fine-grained image categorization and includes a summary of those developments. The fine-grained picture categorization is difficult to do because it requires the model to collect adequate features as well as minute differences in objects that appear to be the same. These particulars may consist of the form, surface, and colour of the thing being described. Specifically, a list of the shortcomings of the already available research was presented, which demonstrated that the issue of FGIC is still a long way from being resolved.

The fluctuation in the look of the item, which inhibits the performance of categorization, makes it difficult for fine-grained characteristics to accurately represent the thing they are trying to describe. The recognition of non-rigid fine-grained items becomes more difficult than the recognition of rigid fine-grained ones. Therefore, in order to establish whether or not fine-grained image classification algorithms are successful, it is necessary to locate and properly utilise the local area information of an item.

The CUB200-2011 benchmark, the Stanford Dogs dataset, the Stanford Cars dataset, and the FGVC Aircraft dataset are among the most influential datasets in the field of fine-grained recognition. The results obtained using the fine-grained approaches are summarised in Tables 9 through Table 13. A classification accuracy ranging from roughly 85% to 89% has been recorded for the CUB200-2011 dataset, which is the fine-grained dataset that is utilised the most frequently for measuring the effectiveness of the system.

REFERENCES

[1] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, Serge Belongie, “Fine-Grained Image Analysis with Deep Learning: A Survey”, 10.1109/TPAMI.2021.3126648, 2021 IEEE Transactions on Pattern Analysis and Machine Intelligence.

[2] Lowe D G. “Object recognition from local scale-invariant features”, Proceedings of the 7th IEEE International Conference on Computer Vision. Kerkyra. Greece: IEEE, 1099. pp. 1150-1157.

[3] Dalal N, Triggs B. “Histograms of oriented gradients for human detection”. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego. USA: IEEE. 2005. pp. 886-893.

[4] Jegou H. Douze M, Schmid C. Perez P. “Aggregating local descriptors into a compact image representation” Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco. USA: IEEE, 2010.

[5] Sanchez Rerronnin E Mensink I, Verbeek “Image classification with the Fisher vector: theory and practice” International journal of Computer Vision. 2013, 105(3): pp.222-245.

[6] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 1097-1105.

[7] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[8] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Conf. Neural Inf. Process. Syst., 2015, pp. 91–99.

[9] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3431–3440.

[10] R. Girshick, J. Donahue, T. Darrell, and J.Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 580–587.

[11] S. Branson, G. Van Horn , S. Belongie, and P. Perona, “Bird species categorization using pose normalized deep convolutional nets,” in Proc. Brit, Mach. Vis. Conf., 2014, pp. 1–14.

[12] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, “Part-based R-CNNs for fine-grained category detection,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 834–849.

[13] D. Lin, X. Shen, C. Lu, and J. Jia, “Deep LAC: Deep localization, alignment and classification for fine-grained recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1666–1674.

[14] H. Zhang et al., “SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1143–1152.

[15] S. Huang, Z. Xu, D. Tao, and Y. Zhang, “Part-stacked CNN for fine-grained visual categorization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1173–1182.

[16] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, “Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization,” Pattern Recognit., vol. 76, pp. 704–714, 2018.

[17] M. Lam, B. Mahasseni, and S. Todorovic, “Fine-grained recognition as HSnet search for informative image parts,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2520–2529.

[18] Z. Wang, S. Wang, H. Li, Z. Dou, and J. Li, “Graph-propagation based correlation learning for weakly supervised fine-grained image classification,” in Proc. AAAI Int. Conf. Artif. Intell., 2020, pp. 122 89–122 96.

[19] C. Liu, H. Xie, Z.-J. Zha, L. Ma, L. Yu, and Y. Zhang, “Filtration and distillation: Enhancing region attention for

fine-grained visual categorization," in Proc. AAAI Int. Conf. Artif. Intell., 2020, pp. 11 555–11 562.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[21] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 818–833.

[22] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 842–850.

[23] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 4749–4757.

[24] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1143–1151.

[25] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1134–1142.

[26] Aaziz Fadhil , R. ., & Haddi Hassan, Z. A. . (2023). A Hybrid Honey-Badger Intelligence Algorithm with Nelder-Mead Method and Its Application for Reliability Optimization. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 136–145. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2580

[27] Y.Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4148–4157.

[28] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective sparse sampling for fine-grained image recognition," in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 6599–6608.

[29] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 8662–8672.

[30] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1134–1142.

[31] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in Proc. Conf. Neural Inf. Process. Syst., 2010, pp. 1243–1251.

[32] García, A., Petrović, M., Ivanov, G., Smith, J., & Cohen, D. Enhancing Medical Diagnosis with Machine Learning and Image Processing. Kuwait Journal of Machine Learning, 1(4). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/143

[33] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4438–4446.

[34] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 5209–5217.

[35] Y. Peng, X. He, and J. Zhao, "Object-part attention model for finegrained image classification," IEEE Trans. Image Process., vol. 27, no. 3, pp. 1487–1500, Mar. 2018.

[36] H. Zheng, J. Fu, Z.-J. Zha, J. Luo, and T. Mei, "Learning rich part hierarchies with progressive attention networks for fine-grained image recognition," IEEE Trans. Image Process., vol. 29, pp. 476–488, Jun. 2020.

[37] X. He, Y. Peng, and J. Zhao, "Fast fine-grained image classification via weakly supervised discriminative localization," IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 5, pp. 1394–1407, May 2019.

[38] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5012–5021.

[39] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200–2011 dataset," Univ. California, Los Angeles, Ca, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[40] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei , "Novel dataset for fine-grained image categorization," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop Fine-Grained Vis. Categorization, 2011, pp. 806–813.

[41] J. Krause, M. Stark, J. Deng, and L. Fei-Fei , "3D object representations for fine-grained categorization," in Proc. IEEE Int. Conf. Comput. Vis. Workshop 3D Representation Recognit., 2013, pp. 554–561.

[42] Prof. Parvaneh Basaligheh. (2020). Mining Of Deep Web Interfaces Using Multi Stage Web Crawler. International Journal of New Practices in Management and Engineering, 9(04), 11 - 16. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/94

[43] Khetani, V. ., Gandhi, Y. ., Bhattacharya, S. ., Ajani, S. N. ., & Limkar, S. . (2023). Cross-Domain Analysis of ML and DL: Evaluating their Impact in Diverse Domains. International Journal of Intelligent Systems and Applications in Engineering, 11(7s), 253–262.

[44] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in Proc. IEEE Int. Conf. Comput. Vis., 2011, pp. 2018–2025.

[45] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in Proc. IEEE Conf.Comput. Vis. Pattern Recognit., 2015, pp. 1798–1807.

[46] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3828–3836.

***Article Received**: 26 April 2023 **Revised:** 18 June 2023 **Accepted:** 30 June 2023*

[47] B.-B. Gao, X.-S. Wei, J. Wu, and W. Lin, “Deep spatial pyramid: The devil is once again in the details,” 2015, arXiv:1504.05277.

[48] L. Wang, J. Zhang, L. Zhou, C. Tang, and W. Li, “Beyond covariance: Feature representation with nonlinear kernel matrices,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 4570–4578.

[49] Q. Wang, J. Xie, W. Zuo, L. Zhang, and P. Li, “Deep CNNs meet global covariance pooling: Better representation and generalization,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 8, pp. 2582–2597, Aug. 2021.

[50] T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear CNN models for fine-grained visual recognition,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1449–1457.

[51] T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear convolutional neural networks for fine-grained visual recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1309–1322, Jun. 2018.

[52] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, “Compact bilinear pooling,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016,pp. 317–326.

[53] N. Pham and R. Pagh, “Fast and scalable polynomial kernels via explicit feature maps,” in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2013, pp. 239–247.

[54] S. Kong and C. Fowlkes, “Low-rank bilinear pooling for finegrained classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 365–374.

[55] A. Dubey, O. Gupta, R. Raskar, and N. Naik, “Maximum entropy fine-grained classification,” in Proc. Conf. Neural Inf. Process. Syst., 2018, pp. 637–647.

[56] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, “Pairwise confusion for fine-grained visual

[57] G. Sun,H.Cholakkal, S.Khan, F. S.Khan, and L. Shao, “Fine-grained recognition: Accounting for subtle differences between similar classes,” in Proc. AAAI Int. Conf. Artif. Intell., 2020, pp. 12047–12054.

[58] D. Chang et al., “The devil is in the channels: Mutual-channel loss for fine-grained image classification,” IEEE Trans. Image Process., vol. 29, pp. 4683–4695, Feb. 2020.