sciendo

VOLUME 18, N° 2 2024
Journal of Automation, Mobile Robotics and Intelligent Systems

Comparative Analysis of CNN-Based Smart Pre-Trained Models for Object Detection on DOTA

Submitted: 24th April 2023; accepted: 20th September 2023

Hina Hashmi, Rakesh Kumar Dwivedi, Anil Kumar

DOI: 10.14313/JAMRIS/2-2024/11

Abstract:
In this paper, we proposed a comparative research project on the classification of various objects in satellite images using some pre-trained models of CNN (VGG-19, ResNet-50, Inception-V3, EfficientNet-B7) and R-CNN. In this research work, we have used the DOTA dataset, which combines data from 14 classes. We have implemented above-mentioned pre-trained models of CNN and R-CNN to achieve optimal results for accuracy as well as productivity in detection of various objects such as ships, tennis courts, swimming pools, vehicles, and harbors from remotely accessed images. In this study, a convolutional neural network (CNN) is used as the base model. For complex computations and for speeding up results, transfer learning is used. With the help of experimental analysis, we have discovered that R-CNN and Inception-V3 performed best out of the five pre-trained models.

Keywords: *remote sensing images, CNN, R-CNN, transfer learning, object detection*

1. Introduction

Artificial Intelligence (AI) can be considered as the future for the world by seeing the ongoing progress of AI in various fields. Many industries are currently growing dependent on AI based applications. As time passes, advancements and improvisations are continuously progressing in the field of AI along with IT [1]. In the recent past it was not thought that AI would play this major role in our day-to-day lives, as it was considered something out of the field of science, only used for the implementation of robots, and that kind of thing. But AI has proven itself a required element of many disciplines, from basic requirements like facial recognition to complex calculations like automation, and so on. The digital industries gained control of various basic challenges through the digital transformation of AI technologies. The work abilities of AI systems have forced AI to sit at the core of various developmental industries [2]. The integration of AI features with many other applications has reduced the overburden of developers in maintaining and improving productivity, efficiency, and quality. From the fields of manufacturing and research to smart healthcare and finance systems, AI has brought many things onto a single platform in a short period of time.

AI has made possible the development of IT systems at a broad level.

AI and machine learning (ML) are the fields that correspond to each other as an integral part of computer science. Most commonly, AI and ML are used as synonyms, but they are different from each other, as AI can be described as the simulation of human through machine for developing many capabilities such as thinking and decision making. ML is a part of AI, the foundation of the process of training the computers with the help of datasets to accomplish AI tasks. ML is used to build smart systems by implementing various learning algorithms. With the help of these ML algorithms, the computers can perform various tasks such as object detection, recognition, localization, and classification [3] in images as well as the applications such as fraud detection in live videos as well [4]. These automated systems are developed by ML algorithms to learn from training images and the model is smart enough to keep itself updated while producing outputs without any human interaction.

1.1. New Paradigms of AI-based Computing – the ML to DL

There has been an outperformance of expectations in terms of the performance of deep learning (DL) models, and the results achieved are state of the art. Deep learning techniques are a subset of ML but improve the process of training and learning. DL achieves powerful results by learning from real-world data in a nested and hierarchical form. The biggest and advantage of DL over ML is that DL learns high-level features from training data in the form of images and videos as an incremental process. DL over ML solves any problem in an end-to-end approach whereas ML follows the approach of breaking down a problem in several parts to solve the problem as a whole. For example, for multiple-objects detection, DL technologies like YOLO take the input in the form of image and video and produce the output in the form of object location with name of object whereas in ML algorithms like SVM, other algorithms like bounding-box object detection is required initially for identification of all possible objects. DL outshines this, as there is no limitation and worries about feature engineering in various applications such as image classification [5], anomaly detection, video surveillance, instance segmentation [6], and object detection.

One of the most demanding approaches of DL is object detection, which is one of the most popular areas today. However, object detection has evolved from about 20 years ago. Before 2014, object detection was performed by several traditional approaches: Viola-Jones Detector (2001), HOG Detector (2006), DPM (2008). Roughly from 2014, DL techniques came into existence, with one- and two-stage object detection algorithms. Various two-stage object detectors, such as SPPNet and RCNN (2014), Fast R-CNN and Faster R-CNN (2015), Mask RCNN (2017), GRCNN (2021), FPN (2017), and one-stage detectors such as SSD (2016), YOLO (2016), RetinaNet (2017), YOLOv3 (2018), YOLOv4 (2020), YOLOR (2021) are performing by achieving state-of-the-art results in the field of object detection.

1.2. Facets of Advanced DL

Deep Representation Learning or deep learning (DL) has improved and dramatically achieved success performing various learning tasks in past few years. Traditional neural network or multilayer perception is a form of neural network that represents the framework of connected layers. It is used in various classification and regression tasks. Artificial Neural Network (ANN) generally involves multilayer perceptrons, Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN), and more. Deep learning techniques are serving in various dimensions: object detection, classification, object recognition, speech recognition, natural language processing, bio-informatics, visual image recognition, drug discovery, visual art processing, and so on. Deep learning follows multiple layers to extract the highest and deepest features of data. Each layer produces more abstract and significant features. From raw data in the form of images and videos, feature levels at each successive level are extracted.

1.3. Convolutional Neural Network

In the near past, CNN attracted researchers, and it has achieved remarkable fame and success because of its working nature. It uses kernel methods along with weight-sharing processes. CNN has set its benchmarks in various domains, most specifically in computer vision. CNN is an advanced form of traditional neural networks, processing complex data at initial stages of computation and pre-processing. CNN is an advanced form of deep learning that accepts images or videos as input, assigns weights to various objects present in video or images, and ably detects and differentiates various objects from each other. Above various other deep-learning-based neural network models, CNN is achieving excellent results in multiple domains.

Convolutional neural networks have several layers, including the input layer, the convolutional layer, the pooling layer, and the fully connected layer. In order to obtain attributes from the input image, the convolutional layer applies filters. In order to reduce computation, the pooling layer down-samples the image before the fully connected layer makes the final prediction. With the help of gradient descent and back-propagation, the network learns the best filters.

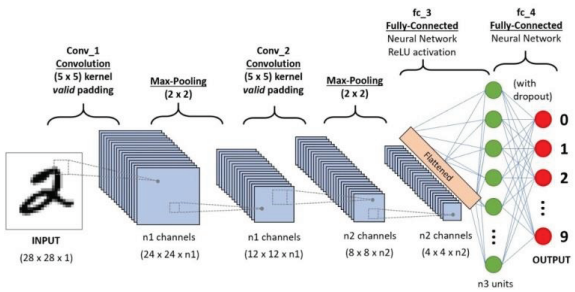


Figure 1. Work flow Diagram of CNN

Each successive convolution layer is followed by a pooling layer using nonlinear functions to reduce the size of the input layer. In addition, they facilitate the reduction of the amount of computation and parametric information in the network. Casteluccio et al., 2015, indicate that it also helps control over-fitting. A [2 x 2] filter is used in Figure 1 to represent the pooling operation. Pooling operations include the following:

- Layers of convolution consist of learnable filters (or kernels) that have the same width, height, and depth as the input volume (3 if it's an image input layer).
- As an example, let's take a 34x34x3 image and run convolution. It is possible to create filters as large as axax3, where a can be anything from 3, 5, or 7, but smaller than the size of the image.
- Forward pass filters are calculated in steps by sliding each filter across the input volume as a dot product of kernel weights and patches.
- The stride of each filter is determined by the number of steps and the height of each filter is the height of the input volume.
- The output volume of the output volume will consist of a depth equal to the number of filters. As we slide our filters, the output volume will have a 2-D output. All filters will be learned by the network.

It is also known as covnets when it is a complete architecture based on Convolution Neural Networks. Every layer of a covnet is a differentiable function that transforms one volume into another. Let's take an image of 32 x 32 x 3 dimensions and run a network on it.

Input Layers: These are layers where we provide input to our models. Images or sequences of images are commonly used as inputs in CNN. Raw input for this layer is 32x32 pixels, 32x32 pixels, and 3x3 pixels.

Convolutional Layers: An input dataset is processed with this layer, which extracts the feature from it. Image input is processed using a set of learnable filters called kernels. A filter/kernel is typically a 2x2, 3x3, or 5x5 matrix. This function calculates a dot product between kernel weights and corresponding input image patches based on input image data. This layer produces feature maps as its output. We'll get a 32 x 32 x 12 output volume if we use 12 filters for this layer.

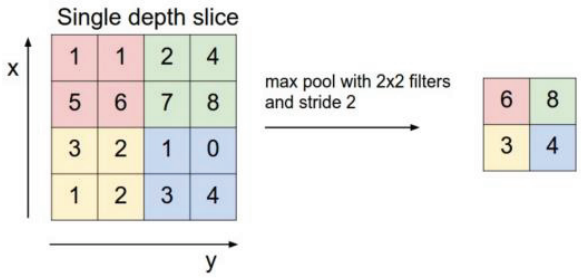


Figure 2. Max pooling

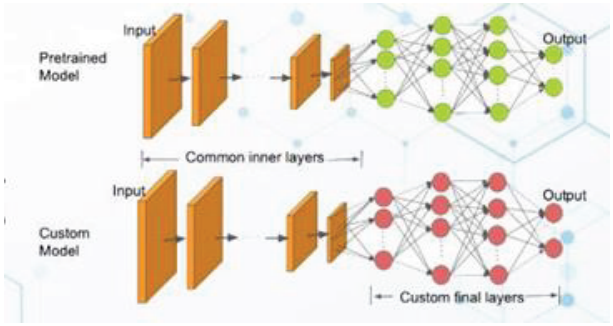


Figure 3. Workflow of transfer learning

Activation Layers: An activation layer adds non-linearity to a network by adding an activation function to its output. A convolution layer's output will be activated element-by-element. In addition to RELU: max(0, x), Tanh, and Leaky RELU, some common activation functions are RELU: max(0, x), and Leaky RELU. Therefore, the output volume will have dimensions 32 x 32 x 12 due to the volume remaining unchanged.

Covnet pooling layer: This layer is periodically inserted into the covnets and is used to reduce the volume of the computation, which in turn reduces memory usage. This layer also prevents overfitting. A maximum pooling layer and an average pooling layer are two common types of pooling layers. With a maximum pool and 2x2 filters, we will have a 16x16x12 volume.

Flattening: After convolution and pooling, the feature maps are flattened into one-dimensional vectors, which can be used for categorization or regression.

Fully Connected Layers: This layer calculates the final classification or regression task based on the input from the previous layer.

Output Layer: As each class's output is converted into its probability score by the logistic function, the output is then fed into a sigmoid or softmax function for classification purposes.

1.4. The Advancement Using Transfer Learning

In the current scenario, deep learning has achieved the remarkable strengths in training deep neural networks to achieve accuracy in predictions and decision making. The network is capable enough to get trained through labels, sentences, images, and predictions as well. Transfer learning is the advanced attraction and charm in the field of AI as it allows reusing an existing labeled data network trained on some specific task or domain. This pre-trained model can be used for some other problem of interest having the same nature. The pre-trained model is used as the initial step for a different model being developed for some other task that involves computer-vision or Artificial Intelligence. Neural networks basically perform the edges detection at the first layer and form it to the middle layer; later on it performs the problem-specific features at the final layers. When we develop a model using smart and advanced transfer learning technology, the task involved in initial and middle layers gets reduced, only we have to perform the retraining process at the latter layers.

As Figure 3 is showing the pre-trained models are used as the initial step in building or customizing any existing model. In the transfer learning, the customized model uses pre-trained model to train its neural networks on a new dataset [9]. By the growing advancements of transfer learning it is offering a number of benefits to the researchers that majorly include reducing training time, improving performance, etc. It involves the use of various pre-trained models as a starting point that is trained on a problem; this reduces the development efforts and time. It is flexible approach also it allows custom model building. A number of top rated models are provided by Keras trained on ImageNet to perform image processing tasks like VGG, ResNet, and Inceptionv3 [10].

1.5. Contributions

This research primarily contributes the following:

1.5.1. Demonstrating how deep learning techniques can be used for object detection using transfer learning techniques

1.5.2. Identify and compare accuracy and precision of different CNN-based pre-trained models.

1.5.3. For achieving state-of-the-art performance, these models are evaluated on the DOTA dataset.

1.6. Organization of Paper

Section 1 contains the fundamental aspects of New Paradigms of AI-based Computing along with facets of deep learning and the advancements of transfer learning in today's AI world. Section 2 gives a brief discussion of object detection in satellite images. Section 3 highlights the survey about various existing tools and techniques for object detection in satellite images using deep learning approaches. Section 4 gives in-depth study of high level and architectural view of transfer learning paradigms. Section 5 defines the comparative analysis of various CNN based pre-trained models for object detection. Section 6 is dedicated to the details of dataset used for the experimentation.

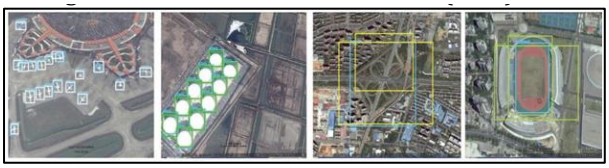


Figure 4. Demonstration of class-based Object Detection in remote sensing images

Section 7 moves towards model implementations in which model training, processing and results are discussed. Section 8 shows the summary and future direction for the research.

2. Object Detection in Remote Sensed Imagery

Since this is the age of Artificial Intelligence, computer vision technology is used and applied to interpret and analyze the images to extract important and useful information from it. Among all those computer vision image processing techniques, object detection plays a vital role to detect various objects availability in an image or video. With the improved versions of object detection, object identification and localization can also be done on the images. Whenever it is required to count or track the precise locations of the objects in an image or video, object detection also helps in labeling the objects in an image.

Object detection in satellite images is a challenging issue as these types of images have different kind of resolutions and dimensions as compared to normal images. Still a lot of research has been done in the field of satellite and aerial imagery to get desired information for Earth observation, change detection, agricultural areas, etc. Over recent years the spatio-temporal earth observation data is increasing drastically. This data is useful to extract important and hidden information to monitor, control or analyze land-surface dynamics in a huge range. Some application areas of satellite image processing involve distortion of settlements, urban growth, vegetation cover areas, water availability etc [11]. Most commonly this kind of study involves various classes like buildings, trees, grassland, shorelines, aircrafts, overpass, oil-tank, playgrounds, etc (Figure 4 is showing). Most commonly, extraction of various deep internal features from remote imagery is done through convolutional neural network (CNN).

3. Various Tools & Techniques

In 1983, Chittineni et al. [12] proposed a framework for line and edge detection in multi-dimensional images. They have represented edges as the direct jump to functions and the weighted sum to expand grey-tone surface in multi-dimensional images. In their work, the noise is assumed as Gaussian. Computational efficiency is achieved by recursive relations. The experiments are performed on Landsat satellite data.

During the years 1993 and 1994, Ionescu et al. [25] focused their research on determining the features of SAR images, since these images have a higher level of complexity due to noise and quality issues associated with them. Feature extraction from SAR images of large-scale objects, such as rivers, lakes, and highways, was accomplished in this paper using an automated algorithm. To detect homogenous areas, the watershed algorithm was employed. In the above areas, possible objects are detected based on similarities between neighboring regions and differences from background areas. For the experiments, images of the Ottawa area are used through SAR imagery.

The detection of forest cover changes from Landsat TM data was first described by Coppin et al. [ext 6] in 1994. The sensor calibration offsets are minimized by using multiple years of imagery. The water reflectance value is obtained with a correlation mechanism between bi-temporal band pairs typically ranging from 0.9884 to 0.9998 combined with dark object subtraction. We applied two different change detection algorithms to bi-temporal vegetation index pairs for intervals of two, four, and six years. Radiometrically defined change classes are investigated. This study demonstrates how reflective TM data can be used to stratify forest cover change in a forest cover change assessment phase before more detailed analysis.

The Indian Remote Sensing (IRS) satellite imagery was used by Mandal et al. [15] to detect man-made objects like roads, bridges, airports, and industrial areas in 1996. For the initial classification, the image pixels are classified into six types of land cover using a multi-valued recognition system. Some heuristic rules about spatial knowledge and their inter-relationships are applied on clustered images to identify certain targets. By using multiple classifications, the detection process became more effective.

In 1996, Rong et al. [13] presented a model for target detection for not only man-made things, but also for natural backgrounds. They have used a kind of self-organizing model for background learning, and then applied some reinforcement learning on that model through contextual information. Experimental results show that they have achieved optimal results in their work.

In 1999, Ng et al. [14] tried to detect human faces. With this model, faces can be detected in multiple views and poses can be estimated at a near-frame rate. Their work extends SVMs to model the 2D appearance of human faces that undergo nonlinear changes over a view sphere.

As part of their development of a system for image analysis, Garnesson et al. [15] proposed the Multi Expert System for Scene Interpretation and Evaluation (MESSIE) based on geometry, context, and radiometry. The system basically operates on geometrical modelling. A study is undertaken to ascertain the class of an object by examining its general structure. The developed model identifies objects of interest initially. Additionally, by examining the characteristics

of gathered salient objects, one can conclude further search for new objects in the scene. A study identifies roads and buildings using suburban images.

An automated method for detecting and classifying hidden targets in hyper-spectral images was proposed by Ren et al. [16] in 1998, which is capable of identifying targets without previous knowledge. A benchmark is achieved in three phases. First, we will choose a band. Second, we will apply a band rationing approach, and, finally, we will use ATDCA (Automatic Target Detection and Classification Algorithm). By analyzing image scenes from the Hyper-spectral Digital Imagery Collection Experiment (HYDICE), we evaluate the effectiveness of the CADCM. These results demonstrate that this model is capable of detecting targets hidden by natural background shades, man-made objects, or shade effects.

Shufelt et al. [17] examined the performance evaluation of four monocular building extraction methods by using image space and object space matrices on 83 images of 18 buildings in 1999. In this analysis, they examine how image obliquity effects system performance, as well as object complexity. Additionally, they analyzed the impact of edge fragmentation on the system. To extract buildings, we used photogrammetric primitive representations and strict object space modeling.

According to Shanks et al. [18], they developed a system that detects cloudiness and aerosol pollution on remote sensing signals to minimize their deleterious effects. They have reviewed existing cloud impacts reduction techniques in this paper.

From airborne light and range (LIDAR) imagery, Haithcoat et al. developed an automated method for extracting the footprint of buildings and reconstructing their 3D shape in 2001 [19]. The objects higher than the ground surface are first extracted from a digital surface model (DSM). The size, height, and shape of a building distinguish it from other objects. In order to improve the quality of the extracted building footprints, an orthogonal algorithm is applied. Ridgelines and slopes are used to identify roofs. In the final step of accuracy assessment, the results are compared with manually digitized building reference data.

Building features can be extracted from images in three steps by Chen et al. [20] in 2002. Anartificial neural network–supervised algorithm is used to categorize roofs using RGB color bands and image textures. A hybrid approach of edge and region segmentation is then applied to extract useful spatial information about objects. The results are then refined using the spatial information. Tests are conducted on AUSIMAGE/spl trade digital imagery.

The model developed by Secord et al. [21] used LIDAR aerial images and range data to approach trees. In this paper, they proposed a two-step process for detecting trees, including segmentation and classification. Here, weighted features are used to segment using the region-growing algorithm. Weights are determined using the random walk learning method. This approach allows for control of the rate of misclassification through weighted support vector machines. Experiments demonstrate its effectiveness.

Chaudhuri et al. [22] used multispectral imagery in 2008 to discover bridges over water bodies. Multispectral images are classified into eight types of land cover using the multi-seed supervised classification technique. Tri-level images are created by identifying water, concrete, and background information from classification. Furthermore, the study uses a knowledge-based approach to reveal the spatial arrangement of the bridge and its surroundings. A recursive scanning method is used to extract the river features. Bridge pixels are identified using neighborhood operators. A spatial resolution of 23.52 m is used for testing on the IRS-1C/1-D satellite.

Using TerraSAR-X ScanSAR images (19-m resolution) as a data source in 2010, Paes et al. [14] proposed concepts for ship detection. TSX images are compared with the K-distribution by means of the Kolmogorov-Smirnov test to determine the goodness of fit. This isdone in a develop-and-verify target detection algorithm.

Using a neighborhood model based on loose spatial contingency, Grant et al. [23] presented a method for detecting amorphous objects in 2012 that uses a maximum probability to tell whether a pixel surrounded by the object of interest contains it as well. The evaluation is done on hypothesis imagery.

Using panchromatic satellite imagery, Elbakary et al. [24] developed a method for detecting shadows in 2014. This method uses a geometric active contour model. On the image, shadows and dark areas are segmented after detection. To distinguish the real shadows from other shadows, they proposed selecting the best threshold and boundary complexity metric. Performance is also validated through experiments.

In 2016, the automatic content-based analysis [25] presented by Sevo et al. allows for arbitrary objects to be detected in aerial images. This paper implements a two-stage training model and then verifies it against remote imagery using convolutional neural networks. UCMerced's data set is used to test the model's accuracy, which is 98.6%.

A feature fusion method was proposed by Yu et al. [27] in 2019 that utilized multiple layers of remotely sensed images to extract some of the fine-grained features, since the images have a number of similarities and differences between classes, along with multidirectional objects. Initially, ResNet50 is applied to extract the features from multiple layers, and then channel attention is applied to enhance them. Fusion is performed using multilayer bilinear pooling and feature connections. The training model is built using PyTorch, a deep learning framework.

The paper by Yu et al. in 2020 [26] outlined an approach to detect vehicles from remote imaging that relies on convolutional capsule networks. In the beginning, an image is segmented into super pixels, and patches are generated without redundant information. Finally, repetitive detection is eliminated using non-maximum suppression.

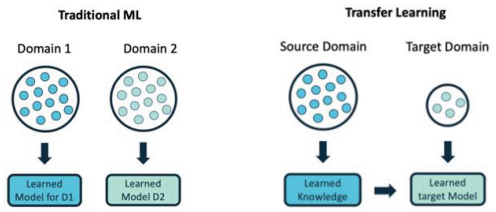
Table 1. Classification based on tools used by various researchers

Algorithm/Model Methodology	DataSet	Findings
MSCNN (Yao et al. 2021)	VHR satellite images for geospatial applications. Challenging data set for the NWPU VHR-10. The images range from 0.5 m to 2.0 m in resolution.	Filters should be replaced with smaller ones
DCL-based object detection method Xiwen (Yao et al., 2021)	NWPU-VHR-10.v2 data set	A convolution based on depth followed by a convolution based on points
SE-MGMM (Xue et al. 2021)	Synthetic aperture radar images	Kernels with small sizes are used
Convolutional capsule Network (Yu et al., 2020)	Open access Remote imagery	Skip connections are used to identify mappings
PTAN (A patch-based three-stage aggregation network) (BingSui1, et al. 2020)	1. DOTA 2. NWPU VHR-10	Optimizes accuracy as well as floating-point operations with multi-objective neural architecture search.
Compatibility loss clustering method (CLCM) Yongsai Han et al., 2020	1. DOTA 2. UCAS-AOD 3. NWPU VHR-10 4. RSOD-Dataset	Fine grained features from multiple layers
Object relationship reasoning CNN (ORRCNN) (Li et al., 2020)	Aerial Image. data set (AID) [16], UC Merced Land-Use data set&WHU-RS19 data set	Accuracy for multiband data
AASM (Femin et al., 2020)	Open access satellite images	Ability and Efficiency are considered
R-CNN algorithm with dialed convolution (Wei et al., 2020)	HRSC2016 dataset	Accuracy with respect to its feature extraction
Active contour excluding edges models (Rai et al., 2020)	Synthetic Aperture Radar (SAR) images	Working ability on dense dataset
Deep learning algorithms on NVIDIA DGX-1 supercomputer (Larionov et al., 2020)	Pre-trained dataset of SpaceNetfine-tuned on planet database.	Time Complexity & Efficiency are performed
Selective Search and Edge Boxes (Farooq et al., 2017)	NWPU VHR-10	Resulting in high recall rates
An enhanced deep CNN based (Deng et al., 2017)	VHR-10 data set	Substantial number of densely packed objects
Two-stage training model using convolutional neural network (Sevo et al., 2016)	UCMerced Dataset	Arbitrary objects are detected
Neighborhood model (Grant et al., 2012)	hypothesis imagery and DIRSIG	Amorphously shaped objects
Kolmogorov-Smirnov (Paes et al., 2010)	TerraSAR-X (TSX) ScanSAR images (19-m resolutoin)	
A model to detect bridges over water bodies (Chaudhuri et al., 2008)	IRS-1C/1-D satellite images of 23.5 23.5m.	Multispectral imagery are processed

Experimental results indicate that the proposed method achieves its benchmarks better than the traditional methods in terms of completeness, correctness, and quality.

In this study, the author seeks to visualize the interpretations of deep convolutional neural networks for aerial images and to comprehend how these interpretations vary across datasets or under different network weight conditions. Their visualization findings shed light on the robustness and generalizability of well-known networks like VGG19, ResNet50, and DenseNet121. AID and UCM datasets are used to show how common classification methods like convolutional networks have evolved to include object and texture detectors [28].

Tan and Le [7] created the final family of CNNs for image recognition covered in this study, the EfficientNets, using the same search space in late 2019. The baseline model EfficientNet-B0 performed similarly to MnasNet because the search space remained unchanged and the optimization technique was same as MnasNet. The scaling strategy, however, is what makes EfficientNet successful. Tan and Le [7] presented compound scaling, balancing scaling in the network's depth, width, and resolution. In comparison to AlexNet seven years earlier, as a result EfficientNet-B7 achieved 97.1% acc@5 at 66M parameters.



(a) (b)

Figure 5. (a) Not utilizing knowledge from other domains. (b) Transfer learning utilizes prior knowledge from another domain

3.1. Classification Based on Tools Used

Table 1 shows the comparison of models developed by many researchers for various parameters on the different datasets. The prime objective of this classification is to categorize various deep learning algorithms for various parameters like accuracy, time complexity, change detection, speed, and high recall rate.

4. Pre-Trained Modeling Based on CNN

4.1. High Level View of Transfer Learning Paradigm

Learning from one domain or task to another is easy for us as humans. We do not have to start over when we encounter a new task. Then, we can learn and adapt faster and more accurately to the new task from our previous experiences [29]. We have witnessed astounding leaps in the application of artificial intelligence in recent years, thanks to advances in supervised and unsupervised machine learning. Our technology has reached a point where we can develop automatic vehicles, robots with artificial intelligence, and disease detection systems that are human-level or superhuman.

Machine learning models are not capable of generalizing beyond the circumstances encountered to increase their performance [30]. They were inspired by the human capacity to transfer knowledge, which has led them to focus during training [31], which hampers their ability to on transfer learning to resolve these issues. When compared to the traditional machine learning paradigm, where learning occurs in isolation, not utilizing knowledge from other domains (Figure 5(a)), and transfer learning utilizes prior knowledge from another domain (source) to learn about a new domain (target) (Figure 5(b)).

4.2. Architectural View of Transfer Learning Paradigm

As a powerful deep learning technique in computer vision, Transfer Learning (TL) is a powerful tool for constructing high-performance models. Knowledge can be re-used across different areas using TL—the knowledge-reusability concept. You don't need to reinvent the wheel with each new situation or model. You can leverage previous experience. By applying previous knowledge to new tasks, you can perform them more efficiently. An example of transfer learning is when a model developed for one task is used as the basis of a model developed for another.

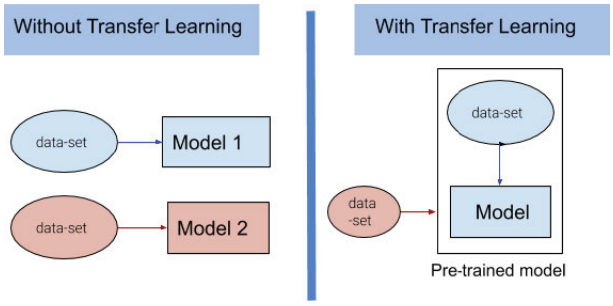


Figure 6. Transfer learning vs traditional learning

This is a popular method of developing neural networks as the starting point for computer vision and natural language processing tasks due to the enormous computer resources and time required for developing neural networks on these problems, as well as the huge leaps in performance they provide when dealing with related problems. In simpler terms, transfer learning entails re-training a model for a second but related task after it has been trained to perform the first task. The Figure 7 illustrates transfer learning overtraditional learning.

Looking at the diagram above (Fig. 6), we see that both models for different tasks are trained from scratch in the traditional approach that is without transfer learning. A transfer learning approach, however, uses our data set to train a pre-trained model that can perform a different task. When you pay attention, you can see that the data set in red (the second one) is smaller than the first. Transfer learning consists of the following steps:

1) Determine the weights of the network.
2) Train on your new images after unfreezing the "head" layers that are fully connected.
3) Training with the weights from the previous training and unfreezing the latest convolutional layers. If we do not do #2, we will trigger large gradient updates.

5. Comparison of Pre-trained CNN Models for Object Detection

As discussed in Section 4, in transfer learning, we can use an already-trained model for new tasks by using a large dataset. Due to the fact that the datasets have been vetted, we can be sure that the quality of the datasets is high. This reduces the cost of training deep-learning models. In satellite object detection; across research and industry, some datasets are highly popular. Following (Table 2) are some of the prominent ones:

Table 2 shows various datasets for satellite images like DOTA, Google Earth, and others. The categories of objects in datasets are also mentioned. DOTA has 14 categories for various objects like harbor, storage tank, ship, baseball diamond, ground track field, etc. Number of instances and number of images shows the count for various object classes and the number of pictures in the dataset, respectively.

Table 2. Datasets for Satellite Imagery [32]

Main categories	Source	Dataset	Number of instances	Annotation way	Bands of image	Number of images
14 categories: Harbor, storage tank, ship, baseball diamond, ground track field, soccer ball field, tennis court, Airplane, large vehicle, roundabout, helicopter, swimming pool, bridge, basketball court	Google Earth	DOTA (Xia et al., 2017)	178,282	Oriented BB	RGB	2816
Two categories: Vehicle, Airplane	Google Earth	UCAS-AOD (Zhu et al., 2015)	13,596	Oriented BB	RGB	1610
Ten categories:, ship, baseball diamond, storage tank, tennis court , ground track field, basketball court, harbor, bridge, airplane, and vehicle	Google Earth and Vaihingen	NWPU VHR-10 (Cheng et al., 2014, 2016b; Cheng and Han, 2016a)	3675	Horizontal BB	RGB	850
Four categories: storage tank, airplane, overpass playground,	Google Earth and Tianditu	RSOD (Xiao et al., 2015; Long et al., 2017)	6850	Horizontal BB	RGB	986
One category: oil well	Google Earth	Oil well dataset (Wang et al., 2021) Zhang et al. (2021)	1092	Horizontal BB	RGB	442

In this paper, we have covered five pre-trained models of CNN (VGG-19, ResNet50, Inception-V3, and EfficientNet-B7) and R-CNN for object detection from satellite data. These models are prominent in their fields and are widely used in the industry as well. By the successful execution of these models on DOTA we have made a comparative analysis for finding the best model with optimal results in terms of accuracy and precision. Here is the in-depth discussion about each.

5.1. VGG-19

A convolutional neural network with a depth of 19 layers is called the VGG. In their article "Very Deep Convolutional Networks for Large-Scale Image Recognition," K. Simonyan and A. Zisserman from the University of Oxford put forth the CNN model known as VGG-19. In the top five tests, CNN performs at 92.7% accuracy on the ImageNet dataset, which consists of over 14 million images divided into 1000 classes [33]. In ILSVRC-2014, this model was among the well-known ones. By sequentially substituting several 33 kernel-sized filters for AlexNet's big kernel-sized filters (11 and 5, respectively, in the first and second convolutional layers), it improves upon AlexNet. Up to 1000 items can be categorized by this network due to its pre-training. The network was trained with 224x224-pixel colored pictures, which indicates that the matrix had the shape (224,224,3). Here is some quick information on its size and capabilities.

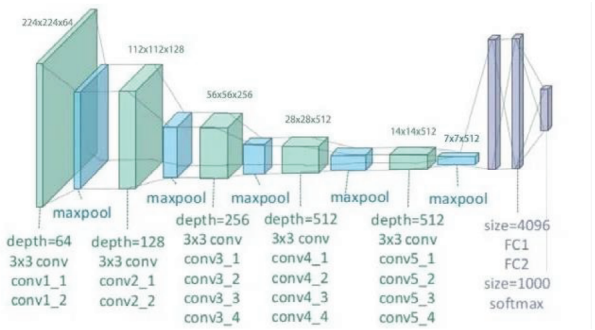


Figure 7. Architecture of VGG-19

According to Figure 7, the network received an RGB image with a fixed size of (224 * 224), indicating that the matrix had the shape of (224, 224,3). One preprocessing step was applied to each pixel, which calculated the average RGB value [34]. Using kernels with a size of (3x3) and a stride size of 1 pixel, they were able to cover the entire image. Spatial padding was applied to the image to maintain its spatial resolution. A 2 x 2 pixel window was used for max pooling with stride 2. To enhance classification accuracy and computation time, a Rectified Linear Unit (ReLu) was added afterward. The performance of this model was significantly better than previous models that used tanh or sigmoid functions. A third layer was implemented using a softmax function that used a 1000-way ILSVRC as part of the classification, while the first two layers had a size of 4096.

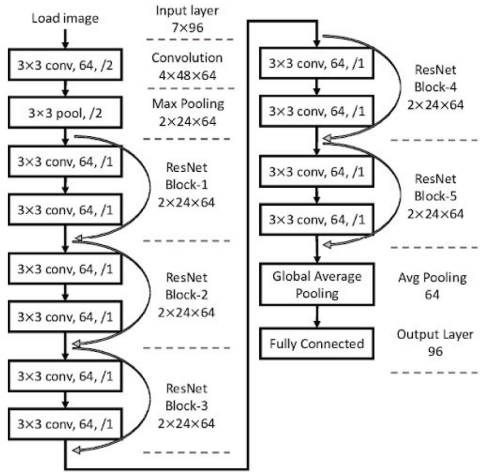


Figure 8. Architecture of ResNet-12, a basic model

5.2. ResNet50

A convolutional neural network with 50 layers in depth is called ResNet50. You may find the model performance results on Microsoft's article titled Deep Residual Learning for Image Recognition, which was developed and trained in 2015. To train this model, hundreds of thousands of photos were taken from the ImageNet database. The network, like VGG-19, can categorize up to 1000 objects based on 224x224 pixel colored images [35].

As can be observed, there are 4 comparable layers utilizing merely different filter sizes after beginning with a single convolutional layer and Max Pooling. All of these levels use the 3 * 3 convolution process. Additionally, we are omitting or skipping the layer between every two convolutions. Identity shortcut connections are what is referred to as residual blocks and are what have missed connections. Simply put, the ResNet authors demonstrate that applying a residual mapping is significantly simpler than applying the actual mapping and that this should be done for all layers. It's also noteworthy to note that the ResNet creators claim that performance shouldn't degrade as we add more layers to the model [36]. Contrary to what we observed in Inception, this is nearly identical to VGG-19 in that it just involves stacking ResNet layers on top of one another while altering the underlying mapping.

5.3. Inception-V3

Szegedy first described the Inception micro-architecture in their work "Going deeper with convolution" in 2014. As can be seen in Figure 9, the Inception Module merely applies convolutions to the input using various filter sizes, applies Max Pooling, and then concatenates the outcome for the following Inception module [37]. Compared to its predecessors, Inception v3 features 42 layers and a lower error rate.

5.4. EfficientNet-B7

EfficientNetB7 is a state-of-the-art convolutional neural network that was trained and released to the public by Google with the paper "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" in 2019.

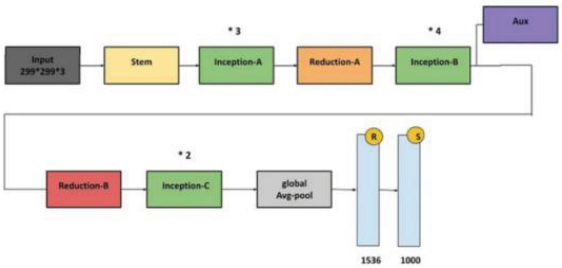


Figure 9. Inception-V3 architecture

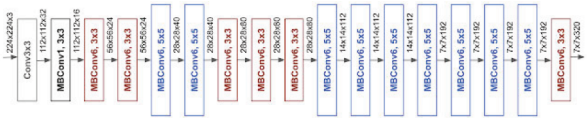


Figure 10. EfficientNet-B7 architecture

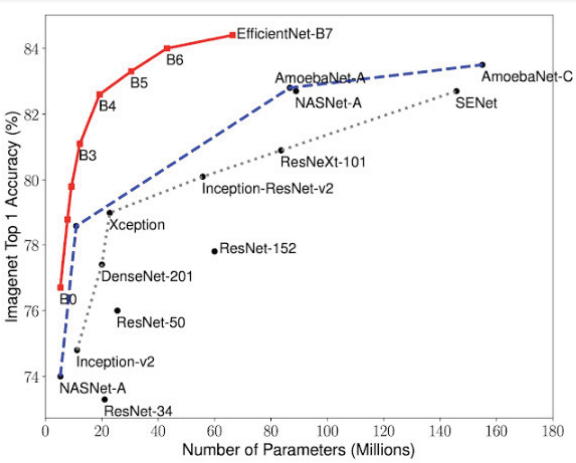


Figure 11. Comparative performances of popular models incl. EfficientNet family

The EfficientNet implementations range from B0 to B7, and even the simplest implementation, EfficientNetB0, is outstanding. In its Top-1 accuracy performance, it achieved 77.1% with 5.3 million parameters [38].

Mobile inverted bottleneck convolution is often known as MBConv (and isrelated to MobileNetv2). The following scaling coefficients are included in their compound scaling formula as well.

- Depth = 1.20
- Width = 1.10
- Resolution = 1.15

A new family of EfficientNets, EfficientNetB0 to EfficientNetB7, is constructed using this algorithm [39]. The performance of this family in comparison to other well-liked models is depicted in the straightforward graph that follows (Fig. 11).

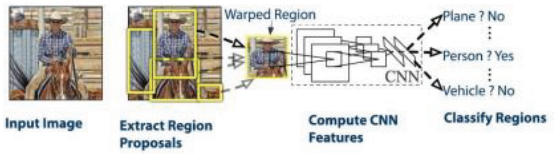


Figure 12. R-CNN: Regions with CNN features

	Baseball Diamond	Basketball Court	Bridge	Ground Track Field	Harbor	Large Vehicle	Plane	Roundabout	Ship	Small Vehicle	Soccerer Ball Field	Storage Tank	Swimming
Training	185	345	238	148	2495	9865	2014	161	5999	21862	201	2944	486
Test	55	105	95	89	929	2903	1030	38	1408	5502	105	775	99
Training : 49053 Test:13772 Total: 62825 images													

Figure 13. Numerical data of objects used

5.5. R-CNN

One of the earliest investigations into deep learning–based object detection is R-CNN. VGG Net and ResNet methods were used to train the network, and it was during this process that the number of classes + background for the classifier layer was calculated. The range of 0 to 1 is used to calculate the similarity ratio. The similarity ratio of the search object to the image is represented by this number. Regardless of the number of classes, the R-CNN test procedure offers 2000 different areas [40]. Four of the closest places from among these 2000 different regions are suggested for each class, as seen in Figure 12.

6. Dataset

In this paper, the DOTA (Xia et al., 2018) dataset issued for the experimentation purpose. 14 kinds of objects were identified in the dataset. For training, 888 photos of various sizes were used. High quality photographs in the size range of 432x559 to 5193x6054 are referred to as training images. There are 888 photos with 49053 objects divided into 14 classes. The test involved 277 photos. These pictures are available in sizes ranging from 448x511 to 6313x6400. The images consist of both nighttime black-and-white and colored photographs. The DOTA dataset's object coordinates are reorganized into rectangular dimensions. The varying sizes of the high resolution samples in the dataset are thought to be a crucial factor in determining how well deep learning models perform generally. Therefore, it was intended to demonstrate the numerous aspects in which the crisis will manifest itself. There are 13,772 items in all throughout the photos utilized for the test. Figure 13 depicts the training and testing classrooms together with the amount of items in each one.

7. Model Implementation: Training, Processing, and Results

The deep learning modeling tool Pytorch was used to implement all of the calculation's code. As the computing system, we have used a workstation with an RTX 2070 SUPER graphics card, an AMD Ryzen7 3700X CPU, and 32-GB of RAM.

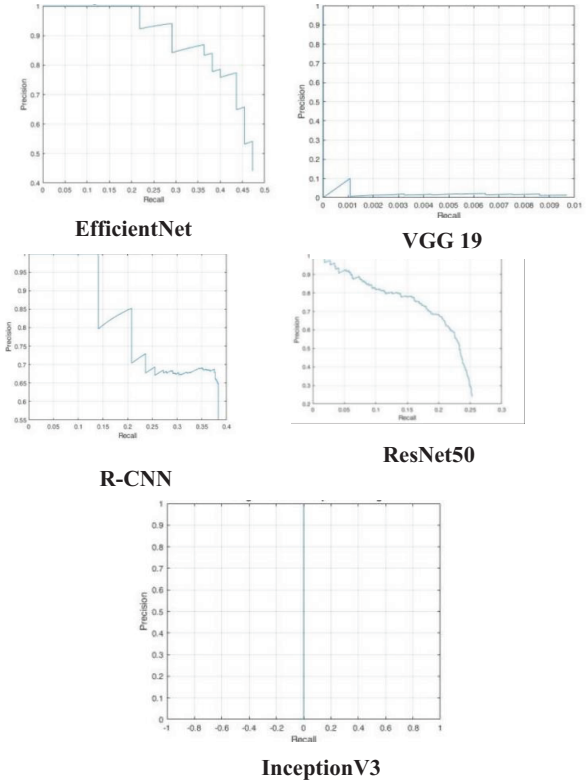


Figure 14. Test results of implemented deep learning models

888 photos that were divided into 14 classes are used for training purposes. Because of the extensive training time, the number of epochs was limited to a minimum.

Deep learning models VGG-19, EfficientNet-B7, ResNet, InceptionV3, and R-CNN were trained.

The testing phase was initiated following the completion of each algorithm's training operations. According to IoU>0.5, precision and recall values were computed. As a result, the rates of each algorithm's detection of objects from 14 classes were made. Figure 14 displays some graphics showing the precision and recall levels attained for this.

The graphs show precision and recall values for studied models and object classes. The recall value starts at 0 and grows towards 1, while the graphic precision value starts at 1 and moves towards 0 in determinations with a high performance rate. This condition makes it clear that the graphics with the highest performance rates achieve good results.

7.1. Results & Discussions

Figure 15 displays the experimentation results by studied CNN models. The results of each model are displayed separately. It is clear that images with high spatial resolution give better results in comparison to low spatial resolution images. These findings lead to the outcomes of 14 objects being detected; 5 deep algorithmic values for each object are shown in bold in the table of learning models Table 3.

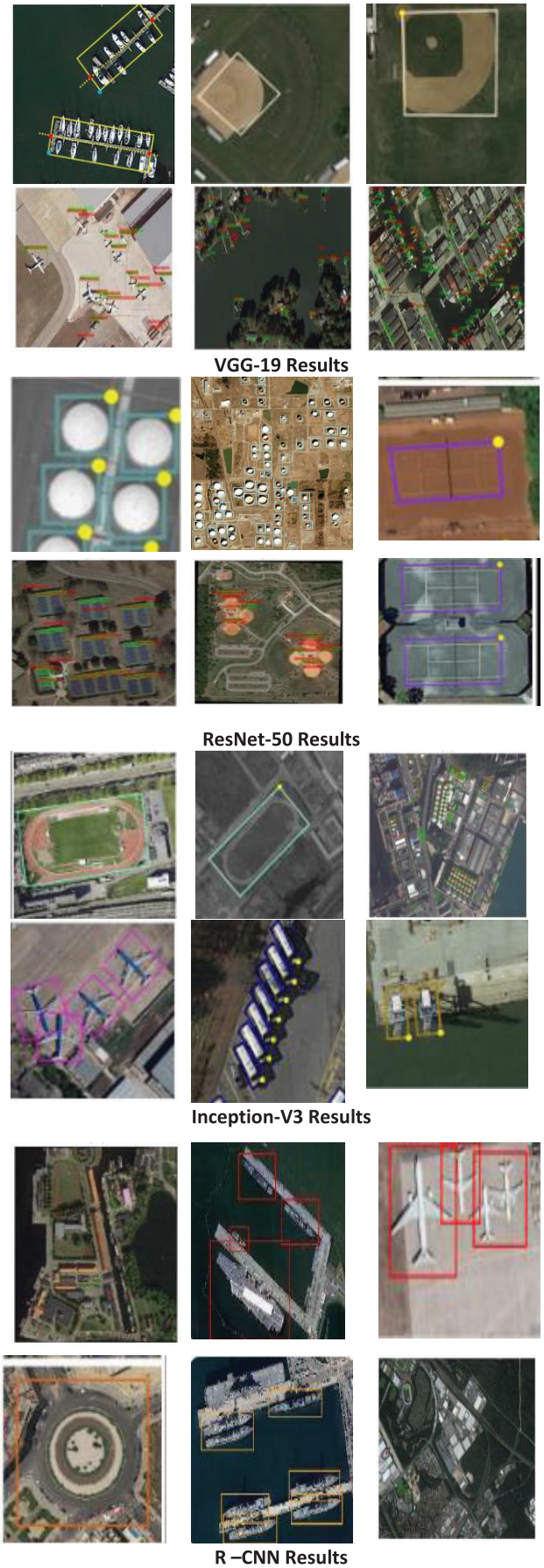


Figure 15. Continued

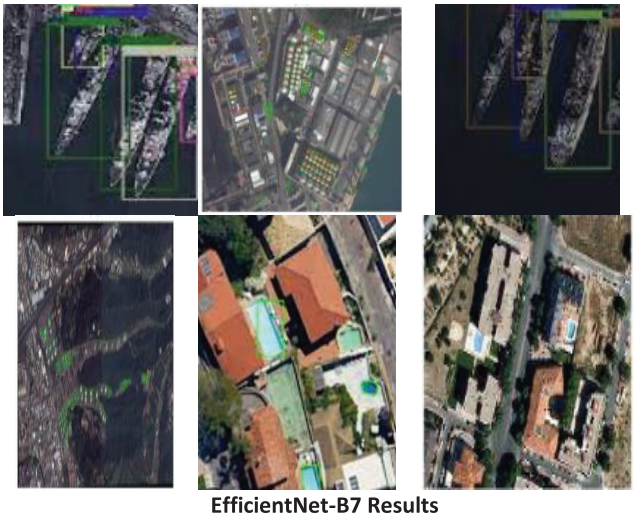


Figure 15. Results of object detection of various deep learning models

Table 3 lists the object detection performance findings for 5 deep learning techniques over 14 classes. The last row of the table also includes the average performance rates for each model.

According to the results, the InceptionV3 and R-CNN models had the highest average performance rates (18.78 and 41.78 respectively). R-CNN outperformed other models in detecting several classes with state-of-the-art results. VGG-19 had the lowest average performance, scoring 19.5, 32.64, and 15.42, 38.07, respectively. The tennis court has the highest availability rate of any object, at 52, 34, and it was acquired using ResNet-50. Roundabout and bridge had the lowest finding rate, at 9.09. In all models, it was discovered at a very low rate. The limited quantity of training and testing samples is believed to be one of the causes of the low availability rate of roundabout and bridge lessons. However, it is believed that the models are unable to define objects that fall under the roundabout and bridge object class in a way that is accurate. These two classes are regarded as lacking some training. The samples in the dataset are close to the ground surface and are simpler for the model to identify in terms of shape after examination of the tennis court picture samples, the object class with the highest availability rate. Models with fewer layers tend to perform better than models with more layers, according to research. This is interpreted as evidence that when detecting many objects, deep learning models shouldn't favor those with a high number of layers. The two most successful results in object classes are shown in Table 4 in light of the results from the models.

Feature comparison results of experimented pre-trained models are shown in Table 4. The accuracy of these models in top-1 and top-5 predictions is also determined when they were used for classification and a comparison of these accuracies is shown in the table below (Table 5).

Table 3. Results of object detection of various deep learning models

Mean Average Precision	VGG-19	ResNet-50	Inception-V3	R-CNN	EfficientNet-B7
Large Vehicle	12, 45	28, 54	**31, 56**	**36, 64**	23, 74
Basketball Court	9, 09	9,0 9	9,0 9	9,0 9	9,0 9
Baseball Diamond	**22, 77**	9,0 9	16, 23	**16, 84**	9,0 9
Ground Track Field	9,0 9	9,0 9	**15, 38**	14, 55	**15, 38**
Bridge	9,0 9	9,0 9	9,0 9	9,0 9	9,0 9
Harbor	**17, 59**	15, 03	16, 88	**20, 25**	17, 00
Plane	12, 16	23, 24	**27, 49**	**31, 42**	27, 39
Roundabout	9,0 9	9,0 9	9,0 9	9,0 9	9,0 9
Ship	23, 55	22, 78	24, 07	**25, 41**	**27, 37**
Small Vehicle	23, 55	14, 86	**15, 45**	**15, 47**	13, 91
Soccer Ball Field	9,0 9	9,0 9	**15, 51**	**12, 73**	10, 54
Storage Tank	9,0 9	**17, 55**	14, 63	17, 15	**17, 79**
Swimming Pool	24, 85	17, 44	24, 09	**25, 56**	**28, 60**
Tennis Court	29, 87	**52, 34**	**49, 01**	25, 56	45, 11
Average Score Rate	15.42, 38.07	17.28, 32.53	19.5, 32.64	18.78, 41.78	18.42, 37.07

Table 4. Class-wise object classes detection result

Models	Object Classes Detected	Top Two Object Classes
VGG-19	Baseball Diamond, Harbor	Baseball Diamond, Harbor
ResNet	Storage Tank, Tennis Court	Tennis Court , Storage Tank
InceptionV3	Large Vehicle, Ground Track Field, Plane, Small Vehicle, Soccer Ball Field, Tennis Court	Tennis Court, Large Vehicle,
R-CNN	Swimming Pool, Large Vehicle, Baseball Diamond, Harbor, Plane, Small Vehicle, Ship, Soccer Ball Field,	Large Vehicle, Plane
EfficientNet-B7	Ground Track Field, Ship, Storage Tank, Swimming Pool	Swimming Pool, Ship

Table 5. Feature comparison result of studied transfer learning models

Module	Complexity	No. of Parameters	Speed
InceptionV3	Low	23.62 million	High
R-CNN	Low	22.85 million	High
VGG-19	High	138 million	Low
ResNet50	Low	23 million	High
EfficientNetB7	Low	5.3 million -66 million	Low

Table 6. Comparison of Models with respect to classification accuracy

Module	Top-1 Accuracy	Top-5 Accuracy
InceptionV3	0.782	0.941
R-CNN	**0.790**	**0.945**
VGG-19	0.715	0.901
ResNet50	0.770	0.933
EfficientNetB7	0.710	0.931

In this paper, several object detection techniques like CNN (VGG-19, ResNet-50, Inception-V3, and EfficientNet-B7) and R-CNN etc. are discussed and compared. From the discussions, it was found that R-CNN is improved more than other CNN-based pre-trained models in terms of accuracy and precision while working with the DOTA dataset. But Inceptionv3 also acquires next to top position in detecting multiple classes from satellite images after R-CNN.

The results make it clear that as soon as the number of classes for multiple object detection rises, performance rate falls. It is also noted that object detection becomes more challenging as the size of the images produced by high-resolution remote sensing rises. It has been shown that all algorithms struggle to detect objects in photographs that are taken above the ground and cover substantially bigger areas. The number of detected objects and their performance has been found to grow as the field of view narrows. According to the classifications, random selection was used to choose the photos in the dataset that would be used for training and testing.

As per the future direction, it can be stated that for research to be done in this domain, picking the appropriate samples is crucial for testing structures. A better evaluation of the outcomes will depend on the number of samples to be utilized in the training exceeding a particular threshold. By eliminating these issues while doing training and testing will produce better results. Researchers who are interested in this area should take extra care to consider these issues. We might advise that multispectral LIDAR data be used to test these models.

AUTHORS

Hina Hashmi* – College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, 244001, India, e-mail: hinahashmi170@gmail.com.

Rakesh Kumar Dwivedi – College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, 244001, India, e-mail: dwivedi.rakesh02@gmail.com.
Anil Kumar – Indian Institute of Remote Sensing (IIRS), Indian Space Research Organisation (ISRO), Dehradun, 248001, India, e-mail: aniliirsisro@gmail.com.
*Corresponding author

ACKNOWLEDGEMENTS

First and foremost, we extend our heartfelt appreciation to our advisors, for their invaluable guidance, expertise, and continuous support throughout this research project. Their insights and constructive feedback have been instrumental in shaping this work. We are also thankful to our colleagues and fellow researchers who provided valuable input and suggestions during the course of our study. Their collaboration and discussions greatly enriched our understanding of the subject matter. We extend our gratitude to the creators of the DOTA dataset and the authors of the pre-trained models we utilized in our research. Their efforts in data collection and model development have made this work possible.

Additionally, we appreciate the support and encouragement from our friends and family throughout this journey. Last but not least, we would like to express our appreciation to the reviewers and editors for their valuable feedback and suggestions, which helped enhance the quality of this paper.

This research would not have been possible without the collective efforts and support of all these individuals and organizations. Thank you for your contributions to our work.

References

[1] M. Sharp, R. Ak, and T. Hedberg. "A survey of the advancing use and development of machine learning in smart manufacturing." *Journal of Manufacturing Systems*, 48, 2018, 170–179. doi: 10.1016/j.jmsy.2018.02.004.

[2] A. D. Preez, G. A. Oosthuizen. "Machine learning in cutting processes as enabler for smart sustainable manufacturing." *Procedia Manufacturing*, 33, 2019, 810–817. doi: 10.1016/j.promfg.2019.04.102.

[3] H. Hashmi, R. K. Dwivedi, A. Kumar, "Identification of Objects using AI & ML Approaches: State-of-the-Art," 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), 2021, pp. 1–5, doi: 10.1109/SMART52563.2021.9676273.

[4] H. Kumar, S. A. Hashmi, Khan and S. Kazim Naqvi, "SSE: A Smart Framework for Live Video Streaming based Alerting System," 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), 2021, pp. 193–197, doi: 10.1109/SMART52563.2021. 9675306.

[5] K. He, X. Zhang, S. Ren, J. Sun, "Deepresidual learning for image recognition," in: *CVPR*, 2016.

[6] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in: *CVPR*, 2014.

[7] M. Tan, Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." *Proc. Mach. Learn. Res.* 97, 2019, 6105–6114.

[8] Pan, S. J., and Yang, Q. "A Survey on Transfer Learning." *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, 2010, 1345–1359. doi: 10.1109/tkde.2009.191.

[9] X. Sun et al., "Multi-type Microbial Relation Extraction by Transfer Learning," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 2021, pp. 266–269, doi: 10.1109/BIBM52615. 2021.9669738.

[10] X. Wang, S. Liu and C. Zhou, "Classification of Knee Osteoarthritis Based on Transfer Learning Model and Magnetic Resonance Images," 2022 International Conference on Machine Learning, Control, and Robotics (MLCR), Suzhou, China, 2022, pp. 67–71, doi: 10.1109/MLCR57210.2022.00021.

[11] Z. Xia, J. Liu, X. Chen, X. Li, and P. Chen, "Airplane Object Detection in Satellite Images Based on Attention Mechanism and Multi-scale Feature Fusion," 2022 4th International Conference on Robotics and Computer Vision (ICRCV), Wuhan, China, 2022, pp. 142–147, doi: 10.1109/ICRCV55858.2022.9953228.

[12] C. B. Chittineni, "Edge and Line Detection in Multidimensional Noisy Imagery Data," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-21, no. 2, pp. 163–174, April 1983, doi: 10.1109/TGRS.1983.350485.

[13] S. Rong and B. Bhanu, "Modeling clutter and context for target detection in infrared images," Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1996, pp. 106–113, doi: 10.1109/CVPR.1996.517061.

[14] J. Ng and Shaogang Gong, "Multi-view face detection and pose estimation using a composite support vector machine across the view sphere," Proceedings International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV'99 (Cat. No.PR00378), 1999, pp. 14–21, doi: 10.1109/RATFG.1999.799218.

[15] P. Garnesson, G. Giraudon, and P. Montesinos, "An image analysis, application for aerial imagery interpretation," [1990] Proceedings. 10th International Conference on Pattern Recognition, Atlantic City, NJ, USA, 1990, pp.

210–212, vol. 1, doi: 10.1109/ICPR.1990.118094.

[16] Hsuan Ren and Chein-I Chang, "A computer-aided detection and classification method for concealed targets in hyperspectral imagery," IGARSS '98. Sensing and Managing the Environment. 1998 IEEE International Geoscience and Remote Sensing. Symposium Proceedings. (Cat. No.98CH36174), 1998, pp. 1016–1018, vol. 2, doi: 10.1109/IGARSS.1998.699658.

[17] J. A. Shufelt, "Performance evaluation and analysis of monocular building extraction from aerial imagery," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, pp. 311–326, April 1999, doi: 10.1109/34.761262.

[18] J. G. Shanks and B. V. Shetler, "Confronting clouds: detection, remediation and simulation approaches for hyperspectral remote sensing systems," Proceedings 29th Applied Imagery Pattern Recognition Workshop, 2000, pp. 25–31, doi: 10.1109/AIPRW.2000.953599.

[19] T. L. Haithcoat, W. Song, and J. D. Hipple, "Building footprint extraction and 3-D reconstruction from LIDAR data," IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (Cat. No.01EX482), 2001, pp. 74–78, doi: 10.1109/DFUA.2001.985730.

[20] Keping Chen and R. Blong, "Extracting building features from high resolution aerial imagery for natural hazards risk assessment," IEEE International Geoscience and Remote Sensing Symposium, 2002, pp. 2039–2041, vol. 4, doi: 10.1109/IGARSS.2002.1026437.

[21] J. Secord and A. Zakhor, "Tree Detection in Urban Regions Using Aerial Lidar and Image Data," in *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 2, April 2007, pp. 196–200, doi: 10.1109/LGRS.2006.888107.

[22] D. Chaudhuri and A. Samal, "An Automatic Bridge Detection Technique for Multispectral Images," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 9, Sept. 2008, pp. 2720–2727, doi: 10.1109/TGRS.2008.923631.

[23] C. S. Grant, T. K. Moon, J. H. Gunther, M. R. Stites and G. P. Williams, "Detection of Amorphously Shaped Objects Using Spatial Information Detection Enhancement (SIDE)," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, April 2012, pp. 478–487, doi: 10.1109/JSTARS.2012.2186284.

[24] M. I. Elbakary and K. M. Iftekharuddin, "Shadow Detection of Man-Made Buildings in High-Resolution Panchromatic Satellite Images," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 9, Sept. 2014, pp. 5374–5386, doi: 10.1109/TGRS.2013.2288500.

[25] Sevo and A. Avramović, "Convolutional Neural Network Based Automatic Object Detection on Aerial Images," in *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 5, May 2016, pp. 740–744, doi: 10.1109/LGRS.2016.2542358.

[26] D. Yu, H. Guo, Q. Xu, J. Lu, C. Zhao, and Y. Lin, "Hierarchical Attention and Bilinear Fusion for Remote Sensing Image Scene Classification," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, 2020, pp. 6372–6383, doi: 10.1109/JSTARS.2020.3030257.

[27] Y. Yu, T. Gu, H. Guan, D. Li and S. Jin, "Vehicle Detection from High-Resolution Remote Sensing Imagery Using Convolutional Capsule Networks," in *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 12, Dec. 2019, pp. 1894–1898, doi: 10.1109/LGRS.2019.2912582.

[28] B. Vasu and A. Savakis, "Resilience and Plasticity of Deep Network Interpretations for Aerial Imagery," in *IEEE Access*, vol. 8, 2020, pp. 127491–127506, doi: 10.1109/ACCESS.2020.3008323.

[29] X. Sun et al., "Multi-type Microbial Relation Extraction by Transfer Learning," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 2021, pp. 266–269, doi: 10.1109/BIBM52615.2021.9669738.

[30] B. Huang, X. Chen, Y. Sun, and W. He, "Multi-agent cooperative strategy learning method based on transfer Learning," 2022 13th Asian Control Conference (ASCC), Jeju, Korea, Republic of, 2022, pp. 1095–1100, doi: 10.23919/ASCC56756.2022.9828357.

[31] Zou and Q. Zhang, "eyeSay: Make Eyes Speak for ALS Patients with Deep Transfer Learning-Empowered Wearable," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 377–381, doi: 10.1109/EMBC46164.2021.9629874.

[32] M. Thoreau and F. Wilson, "SaRNet: A Dataset for Deep Learning Assisted Search and Rescue with Satellite Imagery," 2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA), Zagreb, Croatia, 2021, pp. 204–208, doi: 10.1109/ISPA52656.2021.9552103.

[33] N. Dey, Y. D. Zhang, V. Rajinikanth, R. Pugalenthi, and N. S. M. Raja, "Customized VGG19 architecture for pneumonia detection in chest X-rays." *Pattern Recognition Letters,* vol. 143, 2021, 67–74.

[34] A. Bagaskara, M. Suryanegara, "Evaluation of VGG-16 and VGG-19 Deep Learning Architecture for Classifying Dementia People." In 2021 4th International Conference of Computer and Informatics Engineering (IC2IE) (pp. 1–4). IEEE.

[35] S. Mascarenhas, M. Agarwal, "A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification."

In 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON) (vol. 1, pp. 96–99). IEEE.

[36] B. Koonce, B. Koonce, B. "ResNet 50. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization," 2021, pp. 63–72.

[37] M. G. D. Dionson, P. B. El Jireh, "Inception-V3 architecture in dermatoglyphics-based temperament classification." *Philippine Social Science Journal*, vol. 3, no. 2, 2020, pp. 173–174.

[38] L. P. Kothala, L. P., and Guntur, S. R. (2022, December). Segmentation of Intracranial Hemorrhage through an EfficientNetB7-based UNET model. In 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) (pp. 1–5). IEEE.

[39] M. K. Islam, C. Kaushal, M. A. Amin, "Smart Home-Healthcare For Skin Lesions Classification With Iot Based Data Collection Device." Kushtia, Bangladesh: Islamic University, 2021.

[40] Y. W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, R. Sukthankar, R. (2018). "Rethinking the faster r-cnn architecture for temporal action localization." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. pp. 1130–1139.

[41] P. Bharati, A. Pramanik, "Deep learning techniques—R-CNN to mask R-CNN: a survey." *Computational Intelligence in Pattern Recognition: Proceedings of CIPR,*2020, pp. 657–668.