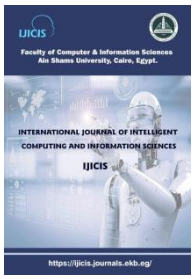
IJICIS, Vol.22, No.2, 63-73
DOI: 10.21608/ijicis.2022.118834.1163

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



RepConv: A novel architecture for image scene classification on Intel scenes dataset

Mohamed Soudy*

Bioinformatics Program, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
M.soudy@cis.asu.edu.eg

Yasmine M. Afify

Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
yasmine.afify@cis.asu.edu.eg

Nagwa Badr

Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
nagwabadr@cis.asu.edu.eg

Received 2022-01-28; Revised 2022-03-24; Accepted 2022-03-25

Abstract

Image understanding and scene classification are keystone tasks in computer vision. The advancement of technology and the abundance of available datasets in the field of image classification and recognition study provide plenty of attempts for advancement. In the scene classification problem, transfer learning is commonly utilized as a branch of machine learning. Despite existing machine learning models' superior performance in image interpretation and scene classification, there are still challenges to overcome. The weights and current models aren't suitable in most circumstances. Instead of using the weights of data-dependent models, in this work, a novel machine learning model for the scene classification task is provided that converges rapidly. The proposed model has been tested on the Intel scenes dataset for a comprehensive evaluation of our model. The proposed model RepConv over-performed four existing benchmark models in a low number of epochs and training parameters, and it achieved 93.55 ± 0.11, 75.54 ± 0.14 accuracies for training and validation data respectively. Furthermore, re-categorization of the data set is performed for a new classification problem that is not previously reported in the literature (natural scenes; real scenes). The accuracy of the proposed model on the binary model was 98.08 ± 0.05 on training data and 92.70 ± 0.08 on validation data which is not reported previously in any other publication.

Keywords

Image scene classification; Intel scene classification; Machine learning; Deep learning;

***Corresponding Author**: Mohamed Soudy

Bioinformatics Program, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

Email address: M.soudy@cis.asu.edu.eg

Introduction

Image recognition and Scene Classification (SC) are keystone tasks in computer vision. With the proliferation of picture and video collections, effective software solutions for data retrieval and processing are essential [1]. Human brains can differentiate multiple objects in real-time, while software tools and algorithms aim to mimic this ability in various ways. Even though various attempts were made to understand images and extract descriptive features, there is still wide room for enhancement.

Object Detection (OD) and Classification (OC) are widely used in scene classification. Researchers use objects in the scene to build a quiet knowledge about the scene structure. They assume that more efforts and research in image understanding will impact the research of image classification as the two tasks are interconnected. The optimization in the image understanding era will facilitate the image classification task. Researchers use one or more object detectors to optimize and enhance classification accuracy. Multiple attempts and approaches were made to optimize the performance of object detection models. Models such as You Only Look Once (YOLO), and Single-shot multi-box detector are the most used models in the last decade, achieving optimal accuracy in the OC task [2]. The initial optimal performance of YOLO lead to the implementation of its derivatives and it becomes widely employed in real-time face and object detection from images and videos.

Furthermore, other researchers made several efforts in the Scene Recognition (SR) challenge using Low-level image data, with the goal of understanding and classifying scenes using low-level color features, orientation, global multi-scale orientation, local dominating orientation, and texture. The theory behind this algorithm aims to classify scenes without identifying the said objects using the features and methods that describe the base features and, then use these descriptors for the scene classification. The approach tries to avoid the propagation of error in the algorithms that use the object to classify scenes as the wrong object assignment will lead to the wrong assignment of the scene. Moreover, the researchers that use this approach take into account the resource management of time and memory [3].

Scenes are usually split into four categories: natural scenes, cultured scenes, indoor scenes, and activity scenes. Because the constituent features of various scenes differ greatly, diverse outcomes may be achieved when the same detection algorithm is applied to multiple scene datasets, particularly for outdoor and interior scenes. The current state of scene categorization research is fraught with difficulties. For starters, sceneries are complicated and diverse, and scene images might differ greatly even when recorded in the same sort of scenario. Second, external elements may cause

interference during the photography process. Different filming angles in the same scenario might result in visual changes across scene images.

Machine Learning (ML) and its extension Deep learning (DL) show an optimal performance in object detection and classification task. Machine learning methods are used as extensions for shallow learning methods such as Histogram of Gradient (HOG). HOG counts the number of times a gradient orientation appears in a certain area of an image and speeds up robust features that use an integer approximation of the determinant of Hessian blob detector, which can be computed with three integer operations using a precomputed integral image. Then researchers use machine learning algorithms such as Support Vector Machines (SVMs) or Decision Trees (DT) for the classification task based on the linearity of the data [4] [5].

Researchers use shallow learning for feature extraction and then machine learning for classification, this approach shows good results on the clean data sets, while the complex images that consist of multiple objects and have less homogeneity within groups lead to lower performance. Using machine learning in feature extraction optimizes the results while consuming the training resources.

Most of the deep learning approaches for image processing rely on the architecture of Conventional Neural Networks (CNNs). It's an extension of the Neural Networks (NNs) but with modification in the connections between layers to ignore less important features and include the vital features that will be used in the classification task. CNN is the most used algorithm in deep learning that extracts the vital features from the image using the Conventional layer, which applies a filter to an image to produce a feature map that describes the detection of features in the input. Pooling layers take the output of the conventional and down-sample the feature map by sliding a two-dimensional filter over each channel of the feature map and summing the features inside the filter's region. The last layer is the fully connected layer which is the feed-forward layer that compiles the final output as a probability assigned to each class of the input data representing the weights [1].

Transfer learning, a branch of machine learning, is widely used on two benchmark scene datasets (Sun397; Places) showing significant results [6] [7]. Models such as VGG16 [8], VGG19 [8], Xception [9], ResNet50 [10], and InceptionV3 [11] were used on scene recognition achieving acceptable performance. The benchmark models are usually tested and validated regarding their efficiency, time, and memory management on the ImageNet dataset and Places datasets, these datasets are benchmarked for object and scene classification. Moreover, with the attempts performed on the aforementioned datasets, there is still a wide room for improvements.

Transfer learning aims to use the weights of the pre-trained model and fine-tune these weights for the classification of the new task. Transfer learning involves initially training a base network on a baseline dataset and task, and then repurposing or transferring the acquired features to a new target

network to be trained on a target dataset and task. This method is more likely to succeed if the characteristics are universal, that is, valid to both the base and target tasks, rather than particular to the base job. While transfer learning aims to reduce the cost of model development using previously trained weights on a semi-generic dataset that contains a large number of classes, the approach is still facing many challenges. Selecting the best model among other models is crucial for obtaining optimal results. Despite the optimal performance of the transfer learning techniques, there are some limitations to overcome as the weights of the pre-trained models are data-dependent according to the training data which is not suitable in most of the classification problems.

Derived by these findings, the contribution of this work is to propose a novel machine learning model for scene classification tasks that rapidly converges while maintaining the accuracy of the classification. To summarize, this work's contribution is two-fold:

- The novel architecture of the machine learning model is composed of 5 Conventional layers followed by batch normalization and Relu layers.
- Re-categorization of the data for the binary classification problem (natural scenes vs. real scenes), with natural scenes being forest, glacier, mountain, and sea, and real scenes including building and streets. The six categories are included in the multi-classification task.

To comprehensively assess the proposed model, evaluate and compare its performance, an experiment is performed against benchmark models: ResNet 50, ResNet-50 (Places architecture), ResNet 101, and SE ResNet 101. The performance evaluation is based on accuracy.

The rest of this paper is structured as follows. Section 2 presents literature work. Section 3 presents materials and methods. Section 4 presents the achieved results. Section 5 presents the conclusion and discussion.

Related work

The work of scene classification entails categorizing scenes from pictures. To attain optimal accuracy, this job frequently employs objects or visual descriptors. Things are identified in scene recognition based on their structure within the picture as well as the surrounding backdrop, as opposed to object classification, which concentrates on classifying influential objects in the foreground.

Scene identification may be accomplished in a variety of ways. At a high level, the techniques may be classified into two categories: those that use low-level characteristics and those that use object recognition. Many more techniques such as probabilistic and/or fuzzy techniques are merged into each of these methodologies in order to cope with the uncertainty that frequently accompanies the outcome of picture interpretation. When comparing low-level feature approaches to object

identification approaches, the purpose of picture interpretation must be considered. When low-level characteristics are applied, scene recognition works better [1].

SC may be further analyzed by examining its close relationships with associated computer vision such as item classification and texture categorization. These challenges also include feature representation and categorization, as is typical of pattern recognition challenges. Scene pictures, on the other hand, are more involved than object or texture categorization. Moreover, it is necessary to further investigate the content of the scene, what the semantic pieces are, how they are structured together, and what their semantic links with each other are. Despite decades of progress in scene categorization, most approaches are still incapable of performing at a level appropriate for varied real-world settings. The inherent challenge stems from the nature of intricacy and the wide range of sceneries [2].

Transfer learning targets a well-trained model on a large dataset and use the previously trained weights for the classification of a new task in which data is similar or close to the data that the model was trained on. Models such as VGG16 [8], and VGG19 [8] arose from the need to reduce the number of parameters in the CONV layers while improving training time [11]. Moreover, Xception and InceptionV3 are developed based on the theory of going wider rather than deeper in terms of the number of layers. Within the same layer, multiple kernels of varying sizes are implemented [12]. Subsequently, ResNet architecture that is developed since neural networks are notorious for failing to find a simpler mapping when one exists, two types of shortcut connections are introduced: identity shortcut and projection shortcut. [13].

The idea behind the ResNet is the input layers of this network are made up of many residual blocks, and the operating idea is to optimize a residual function. It has been discovered that improving CNN performance is still mostly accomplished by adjusting the feature extractor structure. Extraction of the output of CNN's convolutional layers and fully connected layers yields a significant number of visual attributes. Scene-RecNet deeper models are built on a larger number of ResNet, and the ResNet feature extraction module has been enhanced. To change the features extracted by the feature extractor, a feature adjustment module with one convolutional layer and one fully connected layer is added. For simpler storage, the completely linked layer reduces the feature dimension.

These models require a large amount of data for the training process and if the data isn't sufficient enough; the model overfits. Moreover, these models require a lot of time for training and extracting the hidden patterns or crucial features from the input data. In this work, we decipher the limitations and challenges of deep neural network models by proposing a novel architecture for scene classification that regards the training time, memory, and accuracy of the final results.

Materials and Methods

In this work, an optimization of the performance of machine learning is performed by building a machine learning model that optimizes the training set size, training time, and accuracy. The model is trained and validated on data obtained from Kaggle (https://www.kaggle.com/puneet6060/intel-imageclassification)

Intel scenes dataset

The data contains around 25k images of size 150x150 distributed under 6 categories (building; forest; glacier; mountain; sea; street). Data is available on Kaggle at https://datahack.analyticsvidhya.com by Intel to host an image classification challenge.

Re-categorization of the data for the binary classification problem (natural scenes vs. real scenes), with natural scenes being forest, glacier, mountain, and sea, and real scenes including building and streets is performed. Moreover, the six categories are included in the multi-classification task.

This new categorization of the dataset will lead to a new classification problem in the scene classification filed by the repurposing of this benchmark dataset into a complex problem that includes both multi and binary classification. Moreover, the homogeneity within the binary classes will lead to more popularity and usage for such data in the scene classification. The binary classification opens a wide room for research by providing a new challenge using the existing data set in another classification task.

Benchmark approaches

ResNet

ResNet is firstly proposed a convolutional neural network to overcome previous limitations in base machine learning models. Deep convolutional networks inherently incorporate data from multiple levels, and deeper features may be recovered by further deepening the network's structure. As a result, the bigger the number of layers in a convolutional network, the more features may be recovered using this network. When employing deeper networks, however, gradient vanishing and explosion difficulties develop. This difficulty is substantially overcome by typical initialization and regularization layers, which ensure that networks with hundreds of layers may converge, but the gradient disappearance or explosion problem persists as the number of layers increases. Another issue is network degradation, which converts the initial network's layers into a residual block [10].

ResNet50

ResNet50 is a fifty-layer deep convolutional neural network built by researchers from Microsoft's research group. The architecture is designed to overcome the theory of vanishing gradient phenomena in models such as AlexNet and VGG. The input layers of this network are made up of many residual blocks, and the operating idea is to optimize a residual function [13]. The network's

pre-trained implementation was trained over a million photos from the ImageNet dataset. The pre-trained network can identify photos into thousands of item categories, such as keyboards, rats, pencils, and numerous animals. As a consequence, the network has learned detailed feature representations for a diverse set of pictures.

ResNet 50 (Places architecture)

The architecture is designed to overcome the theory of vanishing gradient phenomena in models such as AlexNet and VGG. The input layers of this network are made up of many residual blocks, and the operating idea is to optimize a residual function. ResNet-50 (Places architecture) is the same architecture as ResNet-50 but is trained on the places dataset [19].

ResNet 101

ResNet101 is a one hundred-layer deep convolutional neural network built by researchers from Microsoft's research group. The architecture is designed to overcome the theory of vanishing gradient phenomena in models such as AlexNet and VGG. The input layers of this network are made up of many residual blocks, and the operating idea is to optimize a residual function

Squeeze-and-Excitation (SE) ResNext 101

Squeeze-and-Excitation Networks (SENets) is a CNN building block that improves channel interconnections at nearly no computational burden. They were employed in this year's ImageNet competition and helped to enhance the outcome by 25% over previous years. Aside from providing a significant speed improvement, they are also simple to include in current systems.

Proposed model

In this paper, a machine learning model RepConv is built that consists of five residual layers followed by flattening and dense layers with 64 filters and 2 strides in the Conv layers. To address prior constraints in base machine learning models, the RepConv architecture is compared to the ResNet architecture. Deep convolutional networks absorb data from numerous layers by default, and deeper characteristics may be retrieved by deepening the network's structure even more. As a result, the more layers in a convolutional network there are, the more features may be retrieved using this network. However, when using deeper networks, gradient vanishing and explosion issues arise. Typical initialization and regularization layers significantly alleviate this obstacle, ensuring that networks with hundreds of layers may converge, but the gradient disappearance or explosion problem continues as the number of layers rises. The residual blocks overcome the problem of vanishing gradient in the other models by creating some shortcut connections across the model architecture. The model aims to extract the optimal parameters from the image features to consider the training time.

The data set is split into training data and validation data with ratios of 80%, and 20% respectively. The hyperparameters govern the rate of learning or the pace with which the model learns. It controls how much-assigned error is used to update the model's weights each time they are updated, such as at the conclusion of each batch of training examples. Because of the fewer changes, smaller learning rates need more training epochs. Larger learning rates, on the other hand, result in quicker changes. As a result, a learning rate of 0.01 is used in conjunction with stochastic gradient descent, an optimization approach that estimates the error gradient for the present state of the model using examples from the training dataset and then utilizes backpropagation to update the model's weights.

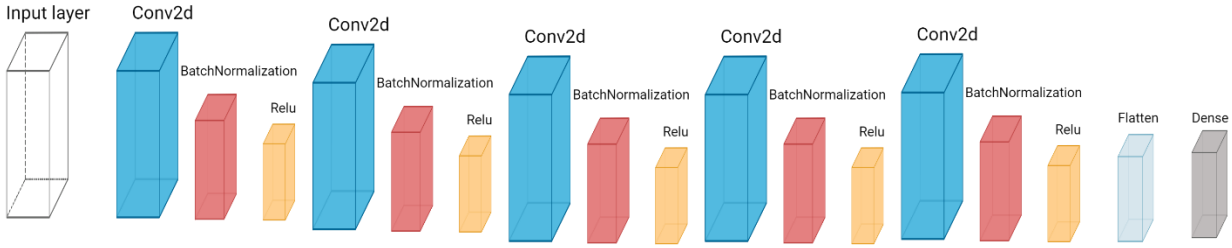


Figure 1: The architecture of the proposed RepConv Model.

To contribute to the advance of scene recognition research, our work is made available at:

Challenge link: https://www.kaggle.com/puneet6060/intel-imageclassification

Dataset link: https://datahack.analyticsvidhya.com

Data-preprocessing and models: https://github.com/MohmedSoudy/Intel-SceneClassification

Evaluation criteria

Accuracy and loss of the trained models are calculated using the following formulas: Equation 1

$$Accuracy = 100\% - Error\ Rate$$

Equation 2

$$Error\ Rate = \frac{|Observed\ - Actual\ value|}{Actual\ value} * 100$$

Results

Extensive tests were carried out to evaluate the correctness of the presented model RepConv. The proposed model is built and evaluated on the Intel scenes dataset and compared to the benchmark models.

The benchmark models are selected based on their performance on the same data selecting the best performance among the multiple attempts made on the dataset.

Accuracies of 95.33%, 94.63%, 94.72%, 94.36%, and 93.55% were achieved for ResNet 50 places, ResNet 50, ResNet 101, SE ResNext 101, and RepConv respectively. The results for all models were comparable with a variance up to 0.75 between highest and lowest accuracies which is a negligible

variance. Compared to benchmark approaches, our model achieved an optimal accuracy taking into consideration the advantage of having the lowest training parameters, architecture depth, and training time (# of epochs) as illustrated in Table 1.

Table1: Accuracy on Intel scene dataset

Backbone Model	Parameters	Accuracy
ResNet 50 (Places architecture)	25,636,712	**95.33**
ResNet 50	25,636,712	94.63
ResNet 101	44,707,176	94.72
SE ResNext 101	44,177,704	94.36
RepConv	160,390	**94.58**

Compared to the benchmark approaches, the proposed model parameters are reduced by 159.83 times when compared to the ResNet Places architecture and the ResNet 50 models, and by 278.74 times when compared to ResNet 101 and 275.43 times when compared to SE ResNet 101. The # of epochs for the proposed model is only 10% of the number of epochs compared to the aforementioned models. Moreover, if the number of epochs is increased, our model achieves optimal accuracy as shown in Figure 2.

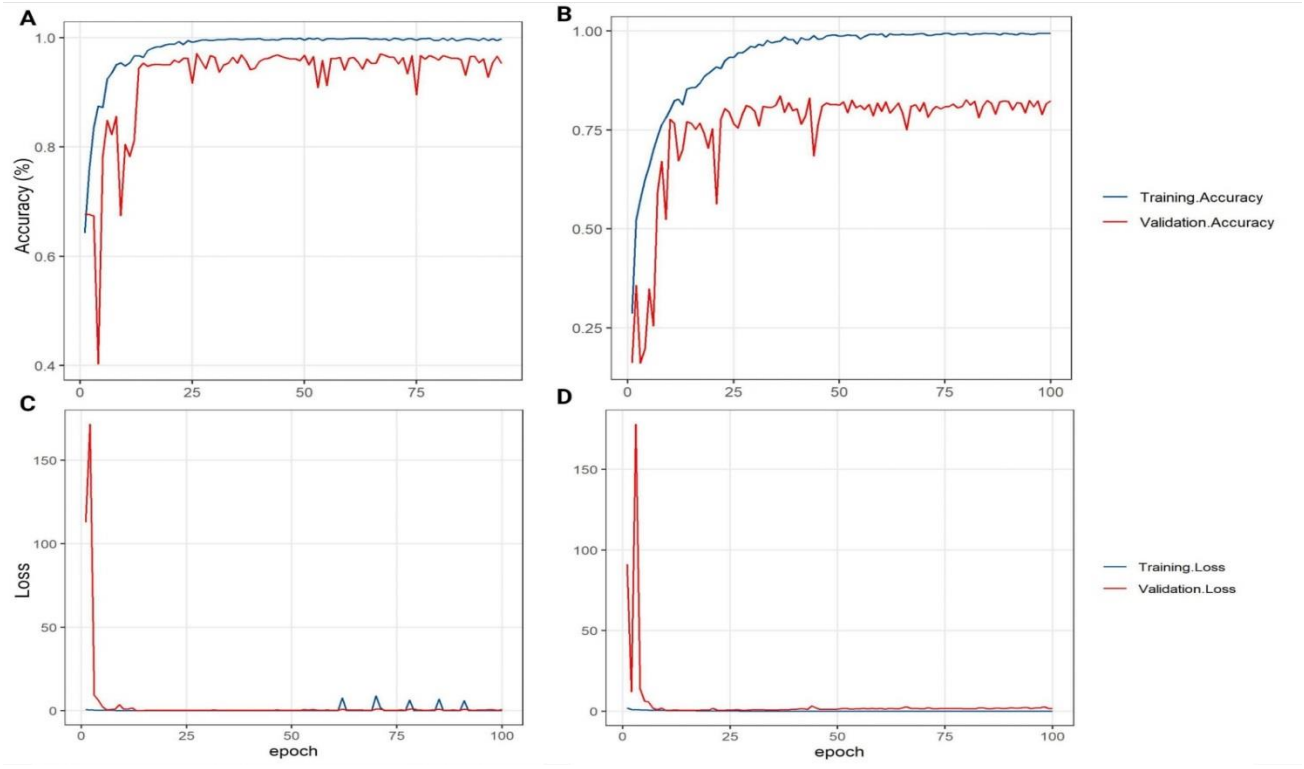


Figure2: Training and validation accuracy of the RepConv model. A: Accuracy on natural scenes vs real scenes. B: Accuracy of 6 classes classification (building; forest; glacier; mountain; sea; street). C: Loss of natural scenes vs real scenes. D: Loss of 6 classes classification.

The model converges and learns the patterns from the training data using the stochastic gradient descent as an optimizer and an initial learning rate of 0.01.

The model performance converges faster on the training data compared to the validation dataset showing that the homogeneity in the training data affected the model performance. Moreover, normalizing the training images impacted the accuracy significantly. The normalization is performed on the whole batch by setting each input mean to zero and dividing the image by the standard deviation of the dataset [14]. Furthermore, the accuracy on the binary model was 98.08 ± 0.05 on training data and 92.70 ± 0.08 on validation data which is not reported previously in any other publication, while the multi-classification model achieved 93.55 ± 0.11, 75.54 ± 0.14 accuracies for training and validation data respectively.

Conclusion

Image understanding and scene classification are two of the most significant challenges in computer vision. The evolution of technology, as well as the number of available datasets in the field of picture classification and recognition research, provide ample opportunities for progress. Scene classification research is currently fraught with difficulties. For starters, scenery is difficult and diversified, and scene shots might vary substantially even when taken in the same type of situation. Second, outside factors may interfere with the photographing process. Different shooting angles in the same scene may result in aesthetic differences across scene pictures.

In this work, re-categorization of the data for the binary classification problem (natural scenes vs. real scenes) is performed, with natural scenes being forest, glacier, mountain, and sea, and real scenes including building and streets. The six categories are included in the multi-classification task. Moreover, RepConv is introduced as a novel machine learning model for scene classification on the Intel scene dataset. The proposed model shows a comparable result to existing benchmark models and achieved 93.55 ± 0.11, 75.54 ± 0.14 accuracies for training and validation data on multi-classification, while it achieved 98.08 ± 0.05 on training data and 92.70 ± 0.08 on validation data on the binary classification problem. These promising results are achieved with the advantage that the proposed model has the lowest depth and parameters compared to the other models. As future work, a standard platform could be provided for scene categorization as a web application to ease user engagement and to assist academics in building and testing their models with a few clicks.

References

1. Singh, V., Girish, D., & Ralescu, A. (2017). Image Understanding-a Brief Review of Scene Classification and Recognition. In MAICS (pp. 85-91).
2. Huang, Rachel, Jonathan Pedoeem, and Cuixian Chen. "YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.
3. Wang, Ning, Yuanyuan Wang, and Meng Joo Er. "Review on deep learning techniques for marine object recognition: Architectures and algorithms." Control Engineering Practice (2020): 104458.
4. Noble, William S. "What is a support vector machine?." Nature Biotechnology 24.12 (2006): 15651567.
5. Quinlan, J. Ross. "Induction of decision trees." Machine learning 1.1 (1986): 81-106
6. Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010, June). Sun database: Large-scale scene recognition from abbey to zoo. In 2010 IEEE computer society conference on computer vision and pattern recognition (pp. 3485-3492). IEEE.
7. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence, 40(6), 1452-1464.
8. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
9. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).
10. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
11. Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
12. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.
13. Sobti, P., Nayyar, A., & Nagrath, P. (2021). EnsemV3X: a novel ensembled deep learning architecture for multi-label scene classification. PeerJ Computer Science, 7, e557.
14. Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.