Enhanced Zero shot learning: Frameworks and Techniques

Second Progress Seminar
Report

Submitted in partial fulfillment of the requirements

of the degree of

Doctor of Philosophy (Science & Technology)
Electronics and Telecommunication Engineering

by

Ansari Shaista Khanam
Reg No: 32

Supervisor:

Dr. Poonam N. Sonar

Department of Electronics and Telecommunication Engineering
Rajiv Gandhi Institute of Technology
Andheri (W), Mumbai-400053
University of Mumbai

June 2023

CERTIFICATE

This is to certify that the **SECOND** progress seminar report entitled **"Enhanced Zero shot learning: Frameworks and Techniques"** is a bonafide work of "**Ansari Shaista Khanam"** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of Doctor of Philosophy in Science and Technology in the subject of Electronics and Telecommunication Engineering of University of Mumbai.

Dr. Poonam N. Sonar
Supervisor/Guide

Dr. S. D. Deshmukh
Head of Department

Dr. Sanjay U. Bokade
Principal

Certificate of Approval of Examiners

This **SECOND** progress seminar report entitled ***Enhanced Zero shot learning: Frameworks and Techniques*** by ***Ansari Shaista khanam*** is approved for the partial fulfillment of requirement for the degree of Doctor of Philosophy in Science and Technology in the subject of Electronics and Telecommunication Engineering of Universityof Mumbai.

Examiners

1._____

2._____

Supervisors

1._____

2._____

Chairman

_____

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature)

Ansari Shaista Khanam

Date:

Abstract

Computer vision has advanced with lot of development in visual recognition systems, which poses restrictions to expand for huge numbers of image classes. This restriction is due to huge image classes with unlabeled images which cannot be used in training the machine learning algorithm. Traditional machine learning methods of classification are based on the classification of categories which are available at the time of training. Technique of Zero-shot learning (ZSL) recognizes categories of test sets which do not appear while training the model. The enhanced ZSL using deep neural network ResNet 50 technique proposed gives better accuracy and reduced model loss than CNN model. The Hierarchical approach is proposed based on two levels of classification, primary level, and secondary level. Primary classification is using simple CNN based classifier which classifies the input image in subclass-1 or subclass-2. The secondary level of classifier is based on ResNet 50 based Zero shot learning classifier which gives top 5 predictions of Labels of input image. Zero shot learning classifier is using ResNet 50 as visual model and fasText as language model. The proposed Hierarchical approach is evaluated using accuracy, computational and time complexity on two standard dataset SUN and AWA2. The proposed approach gives improvement in accuracy from 10% to 15% and reduction in computational complexity from 23% to 47%.

List of figures

List of tables

Acronym

ZSL	Zero shot Learning
GZSL	Generalized Zero shot learning
FS	Feature Space
VSE	Visual Semantic Embedding
ALE	Attribute label embedding
SSE	Semantic Similarity Embedding
GAN	Generative adversarial network
VAE	Variational Autoencoder
RestNet	Residual netwrok
EZSL ResNet 50	Enhanced zero shot learning using deep neural network ResNet 50

Index

Chapter1: Introduction

Conventional Classification approaches have accomplished extensive achievement in many areas. Though, there are some constraints for these methods under this learning model. Sufficient labeled training examples are required for each class. The trained model can only categorize the examples belonging to categories covered by the training data, and unable to recognize with unseen category. Nevertheless, in practical scenario, it is not always possible to have labeled categories. Lack of labeled data is due to a large data set which is expensive to annotate and due to rare data class.

Zero-shot learning (ZSL) deals with recognition of unlabeled data without training it. The model is trained with some classes, called seen classes and classes employed at testing time are called unseen classes. As compared to conventional learning method, this method is testing the classes which are not available at the time of training. To counteract information of about unseen classes, List of attributes, a set of words or sentences in natural language are used to illustrate each class semantically. The key concept of ZSL is thus to understand some intermediary features from training data, which can be applied during the test to map the seen classes with the unseen classes. These intermediary features can indicate the colors or textures (fur, hill, water, sand...) or even some part of objects (paws, claws, eyes, ears, trunk, tail.), eating habits (plants, animals), Since such features are likely to be present in both seen and unseen categories, and this information is used in categorizing unseen classes based on the semantic features [1]. These Semantic features and extracted visual features are mapped during training. During the prediction phase model has access to unlabeled image and its semantic features.in this phase model predict the semantic features corresponding to input image and corresponding class is predicted based on the similarity of predicted features and features of all classes using Euclidean distance.

The enhanced ZSL using the deep visual semantic model is developed by building a visual model using CNN and ResNet 50 and the language model is developed using FastText. The implemented model can classify unknown image categories. The proposed model is tested on the standard datasets SUN [2] and AWA2 [3]. Once a model is tested for unknown categories (zero-shot classes), the performance of the designed Model is evaluated by calculating per class average accuracy.

The Hierarchical approach for zero-shot learning with improved accuracy and reduced time complexity is proposed. This approach is a hierarchical kind of approach with a primary and secondary level of classification. This approach consists of three models, one model in primary classification and two models in secondary classification. The primary classification uses a

simple CNN-based classifier model trained on the reduced number of classes. The secondary classification is using ResNet 50-based ZSL classifiers. ResNet 50-based ZSL classifiers-1 is trained on subclass-1 and ResNet 50-based ZSL classifiers-2 is trained on subclass-2. The performance of the proposed Hierarchical approach is evaluated with per-class accuracy and time complexity metrics. This approach is tested for zero-shot classes on the standard datasets SUN [2] and AWA2 [3].

Chapter 2: Literature review

Image classification is one of the most demanding applications of complexityr vision. Zero shot learning is a kind of classification of images which model has not seen. Lot of research has been made in study of Zero shot learning.

In 2009 Lampert [4] et al. proposed attribute-based classification for unseen classes. Attributes are manually made features for groups of classes. Attributes like feather type, body structure, animal habitat etc., are used as auxiliary information for unseen categories that are not available in training. Attribute-based zero-shot image classification was introduced in [5]. This method uses direct attribute prediction (DAP) and indirect attribute prediction (IAP), which are probabilistic classifiers. Attribute label embedding [ALE] suggested in [6] works better than DAP.

Frome [7] has developed deep visual semantic embedding (DeViSE), which extracts visual features with a convolutional neural network and semantic features using a skip-gram language model. The trained model is checked for its prediction using the Hinge loss function. This model can be used for larger dictionaries and trained on massively bigger text corpora can improve the quality of prediction.

The Deep Weighted Attribute Prediction method [8] uses a deep neural network for feature extraction and class prediction. And it does not require hand-crafted specific features. Models perform better with more accurate attribute prediction. SRC (Sparse representation coefficient) uses weighted attributes for prediction, improving classification accuracy.

The author in [9],[10],[11] have developed generative based methods. The main building block of these methods are Generative adversarial network, Conditional Variational autoencoder which makes the system relatively complicated. These methods yield better accuracy leading to better ZSL classification.

The hybrid Feature model [12] uses a conditional autoencoder conditioned on semantic space. This method uses two autoencoders; one encoder uses visual and semantic information, and another autoencoder is provided with only visual information.

The proposed methodology of Hierarchical approach of zero shot learning emphasizes the computational complexity of models used for Zero shot learning. The literature on computational and time complexity of models is as follows.

RICH LEE [13] has shown various factors in CNN image classification which affects the time complexities. It depends on different layers of CNN, depth of CNN, number of epochs and model optimizer. Various models are evaluated for accuracy, loss, training time, training and validation accuracy.

Pedro J. [14] has provided a systematic approach for evaluating and contrasting the computational complexity of neural network layers in digital signal processing. This paper explains how to calculate four metrics of computational complexity for feedforward and recurrent neural networks. The four metrics are Number of Real Multiplications, Number of Bit Operations, Number of Shift and Add Operations and Number of Hardware Logic Gates. Number of shift and Add operations is a newly introduced metric which describes the bit width along with quantization used in the arithmetical operations.

The main idea of [15] is to evaluate the time complexity of CNN models. This work has identified the factors that affect the model performance, time taken by each layer to run. Time complexity analysis has been done on eight different models by changing the parameters such as size of filters, depth of CNN model, number of filters, number of fully connected layers and kernel size.

The literature for zero shot learning shows that the implemented approach with good accuracy is quite complicated. The accuracy of the exiting method can be further improved with reduced complexity. The proposed Hierarchical approach of zero shot learning is based on deep visual semantic embedding [5] method and is evaluated for accuracy, computational and time complexity.

Chapter 3: Aim of study:

- To develop an efficient approach with optimized design architecture for zero shot learning image classification which effectively makes use of the training data and the supporting information to understand the classification model for the testing classes.

- To identify the challenges and issues in classification with zero shot learning method.

- To enhance performance and reduce complexity.

Chapter 4: Objective of study

- To design and develop an approach for zero shot Image classification with enhanced performance.
- To propose an Architecture to enhance the performance of the model with reduced model complexity.
- To compare the proposed Hierarchical approach with Enhanced Zero shot learning using deep neural network ResNet 50 and evaluate the performance of the proposed architecture with existing methods using standard dataset.

Chapter 5: Methodology

Zero shot learning has received growing interest which is Motivated by humans' skill to identify a new class without ever seen visually. In ZSL method labeled dataset is used for training the model with seen classes, the main aim of ZSL to identify new class that has never seen by the model before. ZSL reduces the annotation cost to a great extent. The common phases in zero shot learning are visual feature extraction, semantic representation and Visual semantic mapping.

5.1 Visual features Extraction:

Extracting visual features is very important in complex vision. Effective methods should be used for extracting visual features. Visual features are color, texture, shape etc. Traditional feature extraction methods such as Hue, Saturation Value, Histogram of oriented Gradient (HOG), gray-level cooccurrence matrix etc. are used. Deep learning models have shown great improvement over traditional methods of feature extraction. Currently VGG, Google Net and Rest net etc models are used.

5.2 Semantic Representation:

ZSL is a method of recognizing unseen data based on knowledge gained from seen data. ZSL is trained with labelled seen data and tested on unseen classes which are not available at the time of training. The traditional method finds it difficult to transfer knowledge which is learned in training for the identification of unseen classes. To overcome this problem some secondary information is used to relate seen classes with unseen classes. This secondary information is called semantic feature space. The semantic features work as a link between seen and unseen categories which are attribute-based, or word vector based.

5.2.1 Attribute Space:

In attribute space, a list of human understandable characteristics illustrating various properties of the classes are defined as attributes. Attributes are words or sentences describing one property of the classes. For example, animals' attributes are their properties such as their body color, their habitat, and their visual properties etc. Like zebra can be described by stripe, forest animals etc are attributes of zebra. These attributes are called semantic feature features which are common in some animals and hence can be used to identify unseen classes.

5.2.2 Word Vector Based:

Natural language processing techniques are using word embedding vectors which are generated using Word2vec. Word2Vec is one of the most widely used methods to understand word embeddings using shallow neural networks. Two basic architectures named continuous bag-of-words (CBOW) and skip-gram are used to generate word2vec vectors [16].

It can be obtained using two methods:

5.2.2.1. Common Bag of Words (CBOW):

In this method, context of the sentence predicts the word in the middle of sentence. CBOW is quite faster to train than the skip-gram and it gives better accuracy for the frequent words.

5.2.2.2. Skip Gram Model:

In this method, the input word is used to predict the context. This method represents rare words well and gives good accuracy with a small amount of data.

5.3 Visual Semantic Mapping:

Visual features are mapped with semantic features. After developing visual semantic mapping in visual space, semantic space or intermediate space unseen classes are predicted. There are three ways of Visual semantic mapping based on it ZSL methods are classified. The three mapping frameworks are shown in fig. 5.1.

Forward Mapping. Forward mapping is a type of mapping in which visual features are associated with semantic features, and the identification of unknown classes are achieved in semantic space.

Common Mapping. Image features and semantic features are associated to common space. Identification of unseen class is accomplished in common space.

Reverse mapping. Reverse mapping is a type of mapping in which semantic features are mapped to visual space. Identification of unseen class is accomplished in visual space.

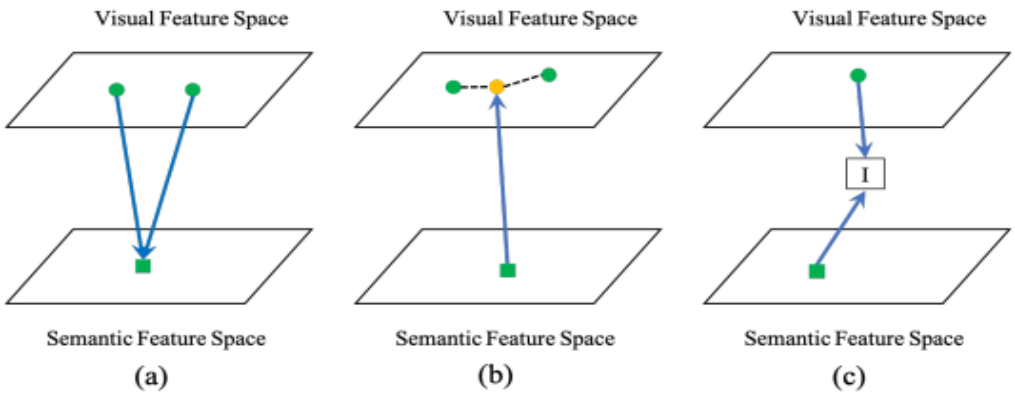


Figure 5.1: Three learning frameworks of zero-shot learning: (a) Forward Mapping (b) Reverse Mapping (c) Common mapping [17]

5.4 Enhanced Zero shot learning using Deep neural network ResNet 50:

Enhanced Zero shot learning using Deep neural network ResNet 50 (EZSL ResNet50) is for Image classification of unseen data i:e zero shot classes merges two models, visual model and language model. This method is using forward mapping approach of visual semantic mapping where visual features are mapped to semantic space.

5.4.1 Visual model:

Visual models are used to extract visual features using deep Convolutional neural network. There rae various models available in literature such as Alex Net, VGG19, Google Net, RestNet50 etc. In this method CNN basic model and Resnet 50 models are used to extract the visual features from the images. Both the models are pretrained on ImageNet data set.

5.4.1.1 Convolutional Neural Network:

It is a very extensively utilized deep neural network. The basic operation used in convolution is a linear operation of matrices. CNN is always composed of the first layer as a convolutional layer which is used to extract various features of an image based on filters. The dimensionality of the network is lowered by the pooling layer. The fully connected layer is used after pooling layer which connects all neurons of the prior layer. Various applications of CNN exist in computer vision for image classification, object detection etc. CNN also finds wide application in natural language processing [18].

5.4.1.2 ResNet 50:

ResNet 50 is an architecture of Deep Convolutional neural networks which are used to extract low, mid, and high-level features from the images. A deep convolutional network is stacked with layers to get better accuracy, but it also leads to challenges of vanishing gradient and exploding gradient. This model converges slowly or starts oscillating. One of the solutions to reduce the challenge of vanishing gradient and exploding gradient is to use skip connection which is used in the Resnet 50 model.

ResNet 50 Architecture as shown in figure.5.2 consists of several convolutional layers, a pooling layer, and a fully connected layer fusing to a total of 50 layers Models consist of several convolutional layers, max pooling etc to extract deep features of an image. The special feature of ResNet 50 architecture is Skip Connection. Skip connection skips some of the layers in the ResNet 50 architecture and provides the output of one layer as the input to the next layers. Skip connections overcome the problem of vanishing gradient to a great extent.

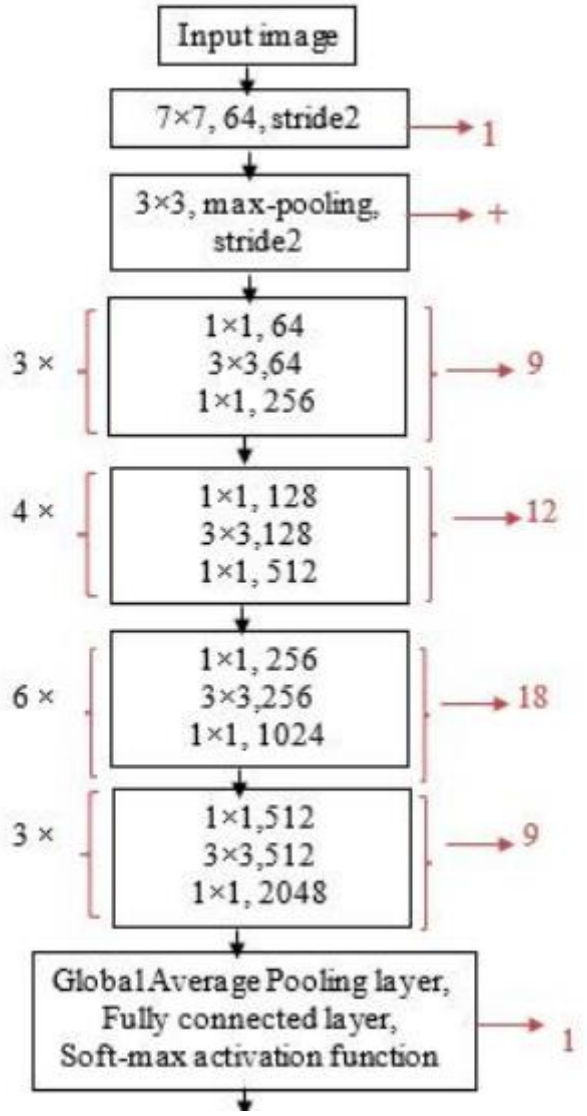


Figure 5.2: ResNet 50 Architecture [29]

5.4.2: Language model:

Language models are used to convert words into word vectors. The Wor2vec method is used for language modeling. The word2vec algorithm uses a deep neural network model to understand word similarity based on a large amount of text. Once a model is trained it can predict semantically similar words. This model represents each label by fixed length vectors called embedding vector. As synonyms tend to occur in similar context, this leads models to learn similar embedding vectors for semantically related words. The language model is implemented using Fast text an open-source library which is used to convert words into embeddings based on text data. Fast Text is created by Facebook's AI Research (FAIR) lab.

Labels of Classes are converted into embedding vectors of 300 Dimensional. The model can generate feature vectors of 50, 100, 200 and 300 dimensions. A more dimension word vector gives more information hence for this experiment 300 dimension is selected.

5.4.3: Deep visual semantic model

Deep visual semantic model [1] combines visual model with language model. Models are trained with training images along with their embedding vector of label generated by Fast Text library. Visual features and language models are mapped with 300-dimensional representation. At the test time, when a new image from zero shot class (unseen class), first model complexity visual feature vector using visual model, then search for nearest labels in embedding vector space using cosine similarity. The top 5 nearest embedding vectors are displayed as result.

The proposed Enhanced zero shot learning using deep neural network using CNN model is shown in Figure.5.3 which consists of two convolutional layers followed by a Max pooling layer, two dense layers and a Relu activation function in the convolution layer. A drop out layer is used to avoid overfitting of model. CNN visual model gives a feature vector of 300 dimensional.

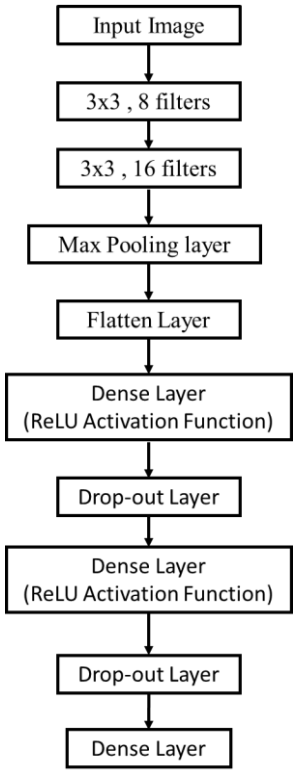


Figure. 5.3. Enhanced zero shot learning using deep neural network CNN Model

The proposed Enhanced zero shot learning using deep neural network using ResNet 50 model is shown in figure 5.4 and is pre-trained on the ImageNet dataset which is followed by two dense layers and drop out layer. The visual model with RestNet 50 also gives a feature vector of 300 dimensions.

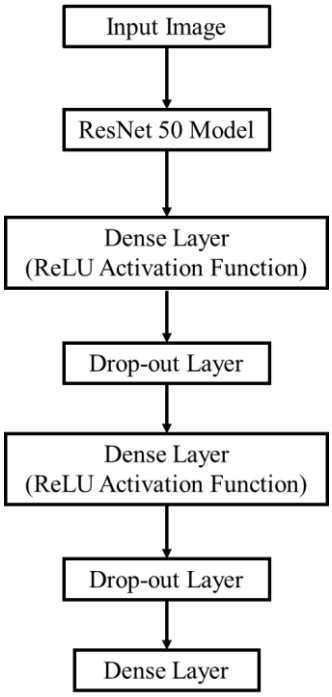


Fig. 5.4. Enhanced zero shot learning using deep neural network ResNet-50

5.5: Algorithm

- **Training:**
    - Input image (seen classes) given to visual model.
    - Visual features (300 dimensional) extracted using CNN based model.
    - Class Labels (seen classes) given to language model.
    - Embedding vectors (300 dimensional) generated using Fast Text model.
    - Visual features and embedding vectors are mapped.
    - Model trained and model loss calculated**.**
- **Testing:**
    - Unseen image given to trained model.
    - Embedding vector is predicted.
    - Top 5 most similar embedding vectors labels based on cosine similarity are displayed.

5.6 Hierarchical Approach for Zero shot learning

Enhanced deep visual semantic embedding method of zero-shot learning uses ResNet 50-based single model for classification which is trained on the whole training dataset. The hierarchical approach is using three models for classification as shown in figure 5.5. The primary classification is based on the CNN model trained for classifying images into two sub-classes i:

e class-1, and class-2. The CNN model used in the first level of classification is trained on a few classes of datasets which is helpful for the classification of Zero shot classification. This CNN model broadly classifies the images into subclasses class-1 or class-2. The secondary classification uses two models, one for each subclass -1 and class 2, which is trained for zero-shot classification. Subclass -1 and class-2 are trained on subcategories of images with ResNet 50-based model.

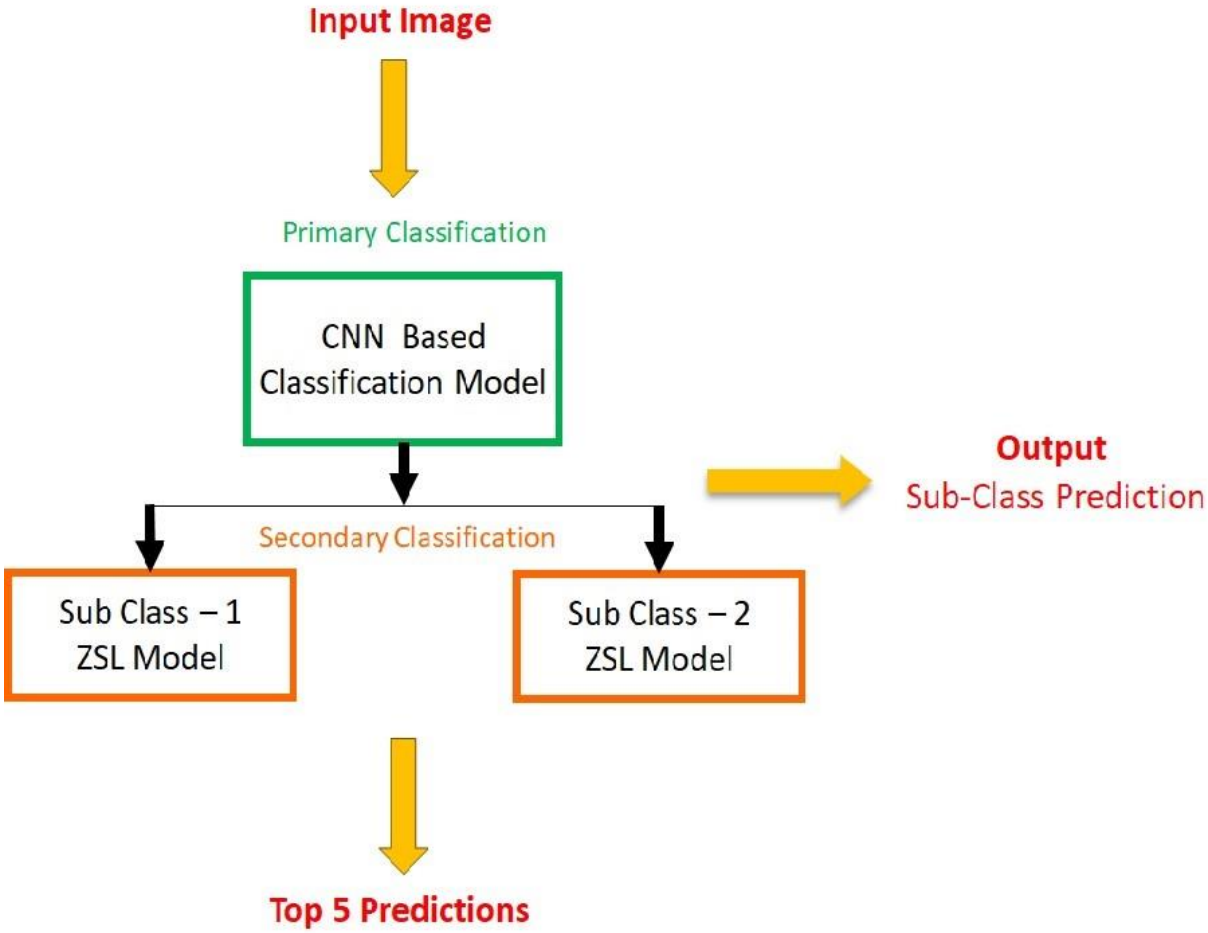


Figure 5.5 – Architecture of Proposed Hierarchical Approach for Zero shot learning.

The zero-shot classification (ZSL) model uses visual model and language model as shown in figure 5.6. Visual model is using ResNet 50 pretrained model trained on images. Language model is also pretrained model for image labels.

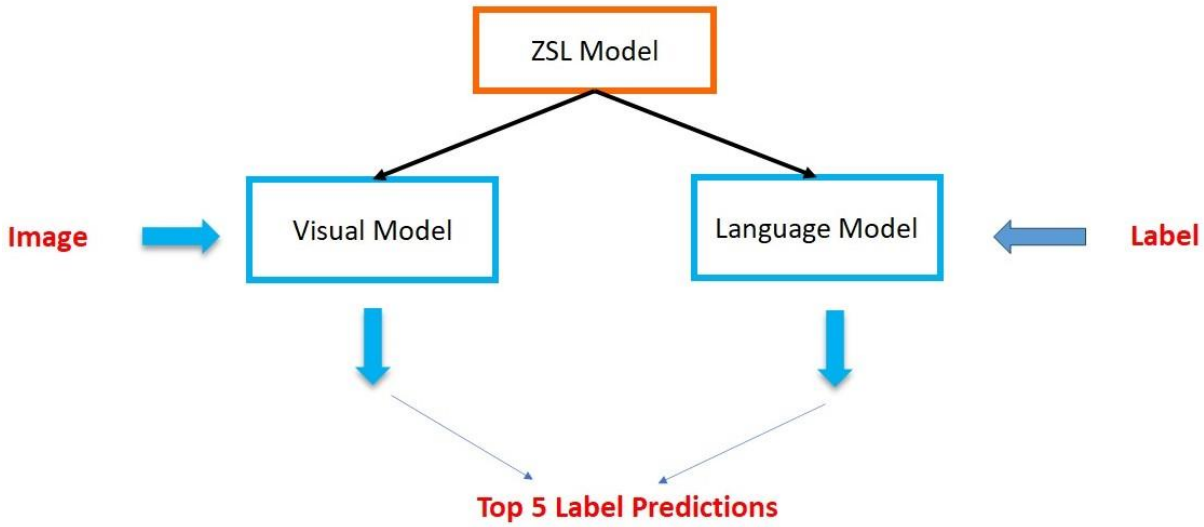


Figure 5.6 – ZSL Model of Proposed Hierarchical Approach for Zero shot learning.

In this approach, the dataset is divided into two categories based on visual similarity and structure of the dataset. AWA2 dataset is divided into two categories, the first category is consisting of monkeys, gorillas, bears, polar bears, and cows, ox whereas the second category is consisting of 4 leg animals, water animals, furry animals such as lions, tigers, rhinoceros, walrus, whale, dolphin, skunk, squirrel, rabbit, mouse, and cat. SUN dataset is divided into two categories, first categories consist of images of nature Mountain, Greenery, snow, and waterbodies and the second category is based on man-made structure such as building, house, and market. Figure 5.7 shows Dataset of Proposed Hierarchical Approach for Zero shot learning.

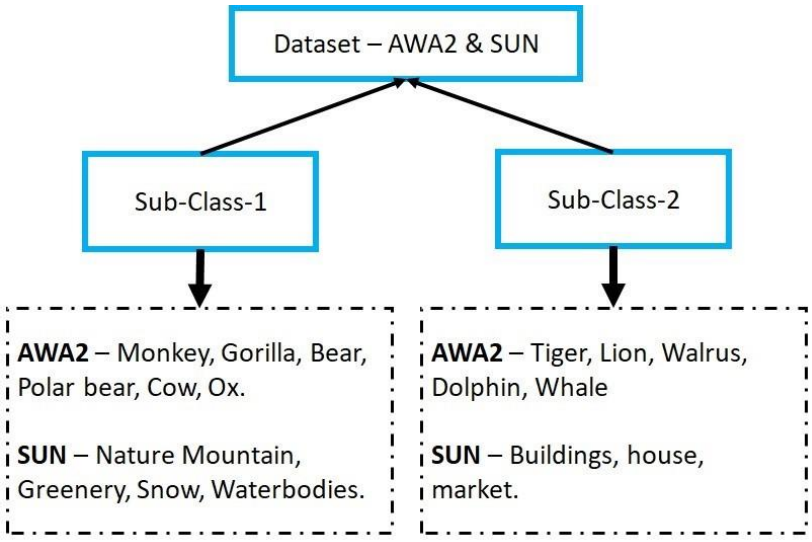


Figure 5.7 – Dataset of Proposed Hierarchical Approach for Zero shot learning.

The primary classifier is a simple CNN-based classifier for broad classification into two subclasses, Subclass 1 and Subclass 2. The CNN-based classifier for the SUN dataset is shown in figure 5.8, consisting of a convolutional layer, batch normalization layer, max pooling layer, and dense layer. Various combinations of layers with different kernel sizes, different combinations of layers have been tested, and the following model has been finalised. This CNN classifier gives training accuracy and validation accuracy of 85.86% and 87.26%, respectively. The CNN classifier for the SUN dataset uses six numbers of convolutional layers, whereas the AWA2 dataset uses five numbers of convolutional layers. The CNN classifier for AWA2 shown in figure 5.9 gives training accuracy and validation accuracy of 84.72% and 84.28%, respectively. The accuracy plots for the CNN classifier for the SUN and AWA2 datasets are shown in figures 5.10 and 5.11.

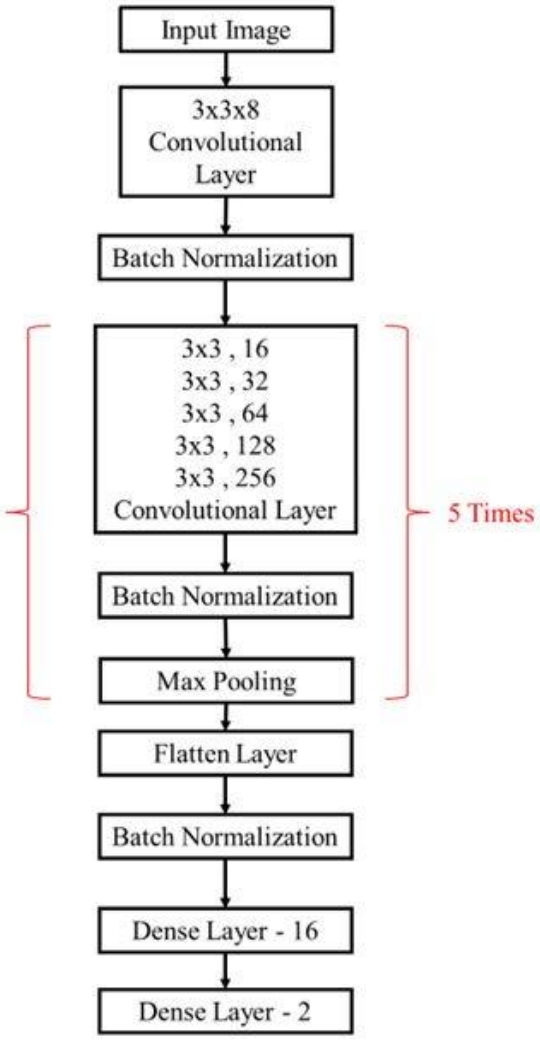


Figure 5.8 – Proposed CNN based classifier for primary classification – SUN dataset.

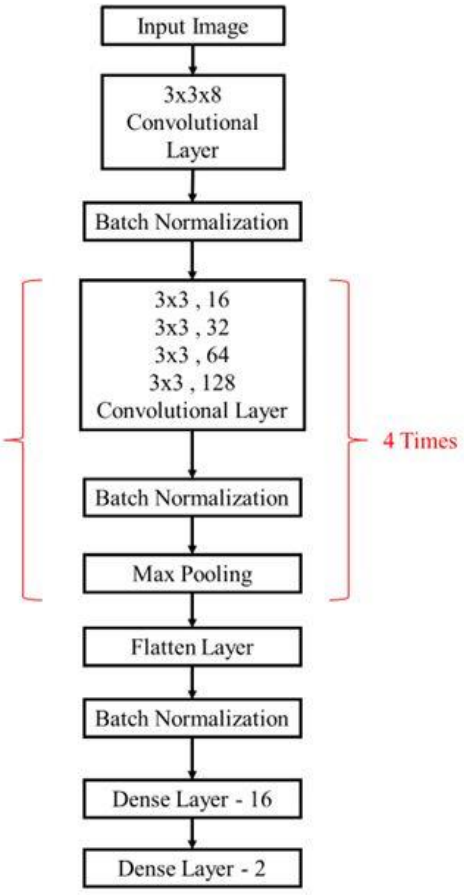


Figure 5.9 – Proposed CNN based classifier for primary classification – AWA2 dataset.

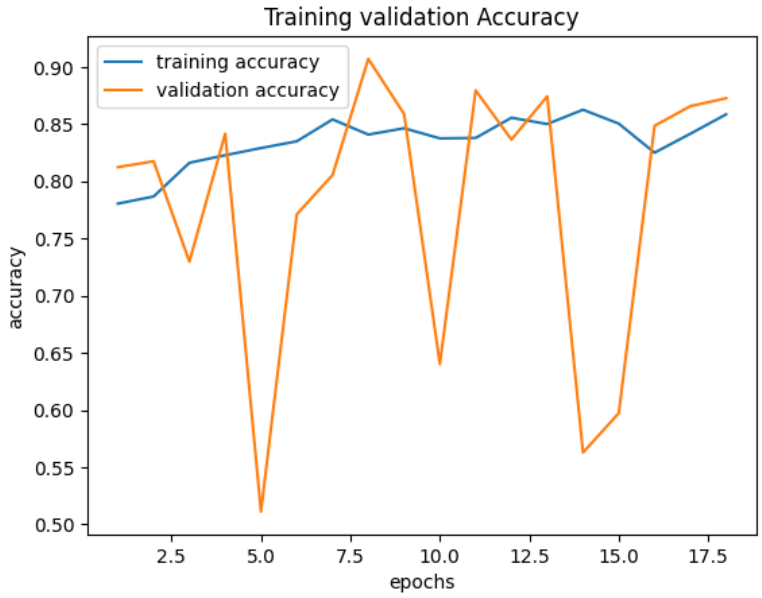


Figure 5.10 – Accuracy plot of Proposed CNN based classifier for primary classification – SUN dataset.

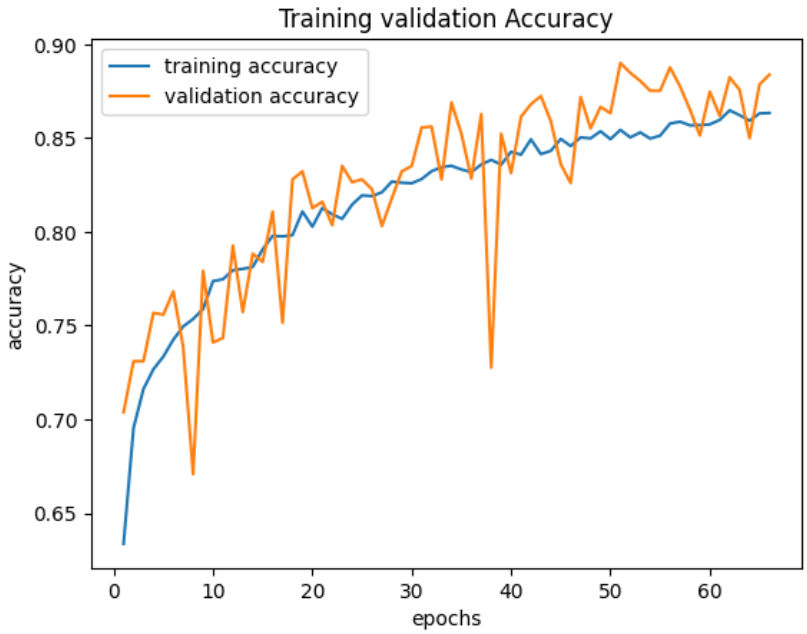


Figure 5.11 – Accuracy plot of Proposed CNN based classifier for primary classification – AWA2 dataset.

5.6.1 Metrics of Computational Complexity

The proposed Hierarchical Approach is evaluated on various metrics of computational complexity.

5.6.1.1 FLOPs: Floating point operations term is an important metric to give complexity of model. This could be an addition, subtraction, division, multiplication, or any other operation that involves a floating-point value. Different layers of Convolutional neural network perform different number of computational operations based on type of layer. FLOPs are used in calculating inference time which is time to take forward propagation. Following are FLOPs calculations for different layer of CNN.

5.6.1.1.1 Convolutional layer: The convolutional layer is the main structure block of a CNN, and it is where most of the computation occurs. It needs input image, kernel. Input image is of height, width, and depth whereas kernel also called filter is a feature detector. Filter is applied to an input image; dot product is calculated between input image and filter by using equation - 1 is called convolution.

$$y_i^f = \sum_{n=1}^{n_i} \sum_{j=1}^{n_k} x_{i+j-1,n}^{in} . k_{j,n}^f + b^f \quad (5.1)$$

where y_i^f i denotes the output, known as a feature map, of a convolutional layer built by the filter f in the i-th input element, n_k is the kernel size, n_i is the size of the input vector, x^{in}

represents the raw input data, k_j^f j denotes the j-th trainable convolution kernel of the filter f and b^f is the bias of the filter f. FLOPs for convolutional layer is calculated using equation-5.2 [19].

FLOPs for convolutional layer = $2 \times n \times$ kernel shape $\times$ output shape (5.2)

n = number of kernels

Kernel shape $= \mathrm{W} \times H$ (5.3)

W and H are width and height of the kernel.

Output shape of convolutional layer = $(M - W + 1)(N - H + 1)$ (5.4)

Where M and N are height and width of the input image.

For first convolutional layer of Model for AWA2 dataset is using 8 kernels of size 3 × 3 and input image is of 224 × 224 ×3.

M = 224, N=224, W = 3, H = 3

Output shape of convolutional layer = $(224 - 3 + 1)(224 - 3 + 1)$

$= 222 \times 222$

FLOPs for convolutional layer = $2 \times 8 \times 3 \times 3 \times 222 \times 222$

= 7,096,869

5.6.1.1.2 Fully connected layer: In the fully connected layer, each node in the output layer connects directly to a node in the previous layer. FLOPs for fully connected layer is calculated using equation-5.5.

Flops for fully connected layer= $2 \times number\ of\ nodes\ in\ input\ layer \times number\ of\ nodes\ in\ output\ layer$ (5.5)

The fully connected layer of Model for AWA2 dataset is classified into 2 classes so number of nodes in output layer is 2. Input to layer is 16 neurons.

Flops for fully connected layer = $2 \times 16 \times 2$

= 64

5.6.1.1.3 Pooling layer: Pooling layer is used to reduce dimensionality reduction. Similar to the convolutional layer, the pooling operation sweeps a filter across the entire input, but the difference is that this filter does not have any weight. Instead, the kernel applies an aggregation function to the values within the receptive field. FLOPs for pooling layer is calculated using the equation-5.6.

FLOPs for pooling layer = $Height \times width \times Depth\ of\ an\ input\ image\ to\ pooling\ layer$ (5.6)

Last Pooling layer of Model for AWA2 dataset Height and width is 12 × 12 and depth is 128.

FLOPs for pooling layer = $12 \times 12 \times 128$

= 18,432

5.6.1.1.4 Batch normalization layer: It is used to normalize the output of the previous layers. The activations scale the input layer in normalization. Using batch normalization learning becomes efficient also it can be used as regularization to avoid overfitting of the model.

FLOPs for batch normalization is given by equation -5.7.

FLOPs for batch normalization = $C \times 4 \times Output\ shape$ (5.7)

C is channels in the convolution layer's output.

FLOPs for first batch normalization layer which consists of 8 channels and output shape of 222 × 222.

FLOPs for batch normalization = $8 \times 4 \times 222 \times 222$

= 1,577,088

FLOPs for EZSL ResNet 50, CNN based classifier for SUN dataset and CNN based classifier for AWA2 dataset are calculated using the equation 5.2 to 5.7. Details of the layers and calculated FLOPs are shown in Table-5.1, Table-5.2 and Table-5.3.

Table-5.1 Details of the layers and calculated FLOPs of EZSL ResNet 50 Model

Sr No.	**Layers of EZSL ResNet 50 Model**	**FLOPs**
1.	ResNet 50 Pretrained Model [20]	3.8×10^9
2.	Dense Layer - 1	18,35,008
3.	Dense Layer - 2	3,44,064
4.	Dense Layer - 3	2,30,400
Total Number of FLOPs EZSL ResNet 50 Model		$\mathbf{3.802409 \times 10^9}$

Table-5.2 Details of the layers and calculated FLOPs of CNN based classifier model for SUN dataset.

Sr No.	**Layers of CNN based classifier model for SUN dataset**	**FLOPs**
1.	Convolution layer – 1	7,096,896
2.	Batch Normalization Layer -1	1,577,088
3.	Convolution layer – 2	111,503,600
4.	Batch Normalization Layer - 2	3,097,600
5.	Max Pooling Layer -1	193,600
6.	Convolution layer – 3	107,495,424
7.	Batch Normalization Layer – 3	746,496
8.	Max Pooling Layer -2	93,312

9.	Convolution layer – 4	99,680,256
10.	Batch Normalization Layer – 4	692,224
11.	Max Pooling Layer - 3	43,264
12.	Convolution layer – 5	84,934,656
13.	Batch Normalization Layer – 5	294,912
14.	Max Pooling Layer - 4	18,432
15.	Convolution layer – 6	58,982,400
16.	Batch Normalization Layer – 6	10,240
17.	Max Pooling Layer - 5	6400
18.	Flatten layer	0
19.	Batch Normalization Layer – 7	1024
20.	Dense Layer - 1	204,800
21.	Dense Layer - 2	64
Total Number of FLOPs CNN based classifier model for SUN dataset.		**476.682×10^6**

Table-5.3 Details of the layers and calculated FLOPs of CNN based classifier model for SUN dataset.

Sr No.	**Layers of CNN based classifier model for AWA2 dataset**	**FLOPs**
1.	Convolution layer – 1	7,096,896
2.	Batch Normalization Layer -1	1,577,088
3.	Convolution layer – 2	111,503,600
4.	Batch Normalization Layer - 2	3,097,600
5.	Max Pooling Layer -1	193,600
6.	Convolution layer – 3	107,495,424
7.	Batch Normalization Layer – 3	746,496
8.	Max Pooling Layer -2	93,312
9.	Convolution layer – 4	99,680,256
10.	Batch Normalization Layer – 4	692,224
11.	Max Pooling Layer - 3	43,264
12.	Convolution layer – 5	84,934,656

13.	Batch Normalization Layer – 5	294,912
14.	Max Pooling Layer - 4	18,432
15.	Flatten layer	0
16.	Drop out layer	0
17.	Dense Layer - 1	58,982,400
18.	Batch Normalization Layer – 6	1024
19.	Drop out layer	0
20.	Dense Layer - 2	64
Total Number of FLOPs CNN based classifier model for AWA2 dataset.		**417.774× 10^6**

Total number of of FLPOs for Hierarchical Approach based model for SUN dataset =
FLOPs of CNN based classifier model for SUN dataset + EZSL ResNet 50 Model for subclass -1 + EZSL ResNet 50 Model for subclass -2
= 476.682 × 10^6 + 3.802409 × 10^9 + 3.802409 × 10^9
= 8.0815*10^9

Total number of of FLPOs for Hierarchical Approach based model for AWA2 dataset =
FLOPs of CNN based classifier model for AWA2 dataset + EZSL ResNet 50 Model for subclass -1 + EZSL ResNet 50 Model for subclass -2
= 417.774 × 10^6 + 3.802409 × 10^9 + 3.802409 × 10^9
= 8.022*10^9

5.6.1.2 Training complexity

The training complexity of model depends on number of FLOPs, number of epochs, number of input image for training.

Training complexity = $Number\ of\ FLOPS \times Number\ of\ epohcs \times number\ of\ input$ (5.8)

Training complexity of model depends on dataset on which it is trained. If Size of dataset increases the training complexity of model increases. At the same time for large dataset, it takes a smaller number of epochs to fit the model then models training complexity can be reduced.

Training complexity of EZSL ResNet 50 = $Number\ of\ FLOPS \times Number\ of\ epohcs \times number\ of\ input$

$Number\ of\ FLOPS$ EZSL ResNet 50 $= 3.8024 \times 10^9$

$Number\ of\ epohcs = 19$

$number\ of\ input = 11{,}735$

Training complexity of EZSL ResNet 50 = $3.8024 \times 10^9 \times 19 \times 11{,}735$

$= 84.78 \times 10^{13}$

Similar to above calculations Training complexity of Enhanced ZSL Model, Hierarchical Approach based model for SUN dataset, Hierarchical Approach based model for AWA2 dataset are evaluated and shown in Table 5.4.

Table – 5.4 Training complexities of evaluation of various models.

Sr No.	**Model**		**Number of FLOPs**	**Number of epochs**	**Number of input image**	**Training Complexity**
1.	Enhanced ZSL Model - SUN		3.802409×10^9	19	11735	**84.78×10^{13}**
2.	Enhanced ZSL Model - AWA2		3.802409×10^9	15	24270	**1.38426×10^{15}**
3.	Hierarchical Approach based model for SUN dataset	primary classifier Model	476.682×10^6	18	2715	2.3295×10^{13}
		Sub class-1 model	3.802409×10^9	25	991	9.4204×10^{13}
		Sub class-2 model	3.802409×10^9	23	454	3.9704×10^{13}
		Total Training Complexity				**19.69×10^{13}**
4.	Hierarchical Approach based model for AWA2 dataset	primary classifier Model	417.774×10^6	66	9767	2.6931×10^{14}
		Sub class-1 model	3.802409×10^9	11	3278	1.37×10^{13}
		Sub class-2 model	3.802409×10^9	10	6408	2.4365×10^{13}
		Total Training Complexity				**6.499×10^{14}**

5.6.2 Algorithm:

- **Training:**
    - Training dataset is prepared with a visual resemblance with test classes.
    - Training dataset is divided into two subclasses class-1, and class-2 based on visual similarity and structure.
    - Primary classification is using CNN based model, trained on an explicitly. prepared training dataset.
    - Primary classification model accuracy and loss are evaluated.
    - Secondary ZSL classification models 1 and 2 are trained on subclass-1 and subclass-2 datasets respectively.
    - ZSL classification models are composed of a visual model and a language model.
    - Input image (seen classes) given to the visual model.
    - Visual features (300 dimensional) extracted using ResNet 50-based model.
    - Class Labels (seen classes) given to language model.
    - Embedding vectors (300 dimensional) generated using the Fast Text model.
    - Visual features and embedding vectors are mapped.
    - ZSL classification Models are trained, and model loss is evaluated.
    - Time and computational complexity are evaluated for each model as well on the complete model.
- **Testing:**
    - Unseen image from the zero-shot class is given to the trained model.
    - Image is classified into either subclass-1 or subclass-2 using primary classification.
    - Embedding vectors are predicted using the Class -1 – ZSL model or Class -2 – ZSL model based on primary classification.
    - Top 5 most similar embedding vector labels based on cosine similarity are displayed.

Chapter 6: Result

With the increase in size of digital data, unlabeled data increases with huge quantity. Conventional image classification method works good with labeled data, raises demand for classification of unlabeled data. Zero shot learning is a solution for this which classify unseen data. Deep visual semantic embedding is one of the methods of zero shot learning which is used to classify unseen classes. The hierarchical approach of zero shot learning is implemented using CNN based classifier for broad classification of unseen class into subclass-1 or subclass-2. Model of Subclass-1 or subclass-2 is doing zero shot learning classification. This method is compared with Enhanced zero shot learning using ResNet 50.

Results are tested on the following datasets:

- **SUN dataset:**

  The SUN [2] dataset comprises 14,340 images of scenes images such as rivers, hills, churches, beaches etc. It consists of a split of 645 seen categories for training and 72 unseen classes for testing is used for zero-shot learning.

- **AWA2 dataset:**

  The AwA-2 dataset [3] comprises 50 categories of 37322 images of animals. In this experiment for the AWA2 dataset, 40 categories are utilized for training and 10 categories are utilized for testing.

Results are evaluated using following parameters:

- **Per class Average Accuracy** = $\frac{1}{N}\sum_{i=1}^{N}[\frac{Y_{correct\ class}^{class\_i}}{Y_{Total}^{class\_i}}]$ (6.1)

  N is total number of classes in dataset

- **Model loss** = $-\sum_{i=1}^{N}[l^2norm(Y-true) * l^2norm(Y-pred)]$ (6.2)

  Cosine similarity model loss is used. Its value is a number between -1 and 1. When it is a negative number between -1 and 0, 0 indicates orthogonality and values closer to -1 indicate greater similarity. The values closer to 1 indicate greater dissimilarity.

- **Flops** – Number of floating-point operations of model
- **Training Time** – Time to train the model.
- **Testing Time** – Time to test one image.
- **Training complexity** – Training complexity depends on number of FLOPs, number of epochs and number of input image.

6.1 Enhanced Zero shot learning using deep neural network ResNet50 – Results:

The enhanced ZSL using the deep visual semantic model is developed by building a visual model using CNN and ResNet 50 and the language model is developed using FastText. The implemented model can classify unknown image categories. The proposed model is tested on the standard datasets SUN [2] and AWA2 [3]. The unseen image when given to a model, it extracts the visual features and predicts the embedding vectors. Top 5 most similar Embedding vectors labels are predicted. Top 5 prediction results for SUN and AWA2 datasets are shown in figure 6.1 and 6.2 respectively.

Actual Label: Valley
Top 5 Prediction for an image: mountain, **valley**, hillside, hill, tree-line

Actual Label: forest
Top 5 Prediction for an image: garden, back-garden, house, farm, gardens

Figure 6.1 Results for SUN dataset - Enhanced ZSL ResNet 50.

Actual Label: chimpanzee
Top 5 Prediction for an image: gorilla, gorillas, chimp, chimpanzee, orangutan

Actual Label: tiger
Top 5 Prediction for an image: tiger, bobcat, tigers, **leopard**, tiger-striped

Figure 6.2 Results for AWA2 dataset - Enhanced ZSL ResNet 50.

The model is trained for 20 epochs in call back mode. In call back command early stopping optimizing technique is used to avoid overfitting. Adam optimizer is used. Model loss for SUN and AWA2 datasets are shown in figures 6.3 and 6.4 respectively. Model loss reduces with training of model with increase in epochs.

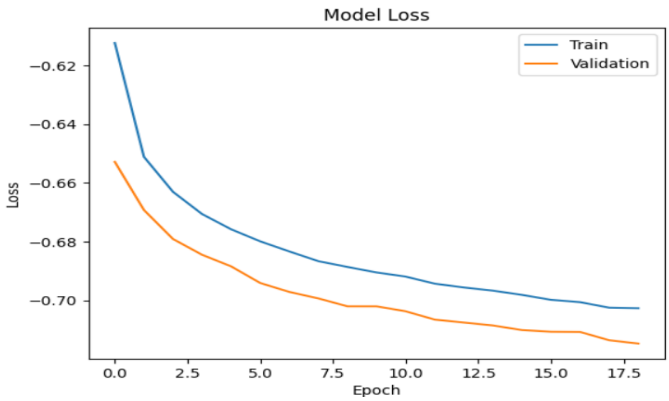


Figure.6.3 Model loss – SUN dataset

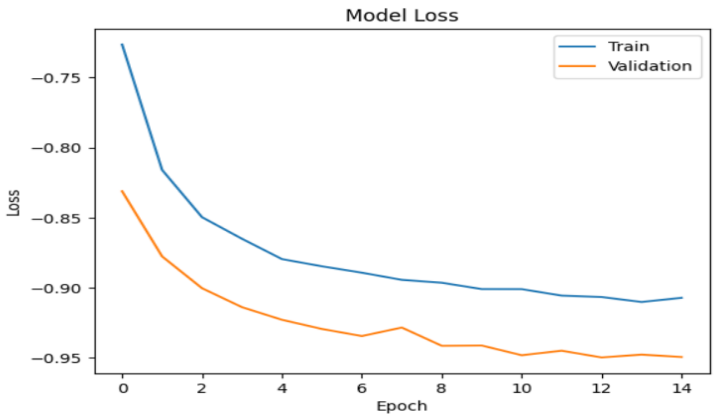


Figure.6.4 Model loss - AWA2 dataset

6.2 Hierarchical Approach for Zero shot learning – Results:

Hierarchical Approach is built using primary and secondary level of classification. The primary level of classification is accomplished using CNN based classifier which broadly classifies image in two subclasses. Secondary level of classification is accomplished by ResNet 50 based model. For Zero shot class image CNN based classifier classify image into sub-Class-1 or sub class-2 based visual similarity. ZSL classifier for sub class-1 or sub class-2 extracts visual features using ResNet 50 and predicts the embedding vectors. The top 5 most similar embedding vector labels are predicted as a result. Tops 5 prediction results for SUN and AWA2 datasets are shown in figure 6.5 and 6.6 respectively.

Actual Label: Valley

Top 5 Prediction for an image: mountain, hillside, meadow, hill, **valley**

Actual Label: forest
Top 5 Prediction for an image: forest, meadow, tree-line, foresty, woodlot

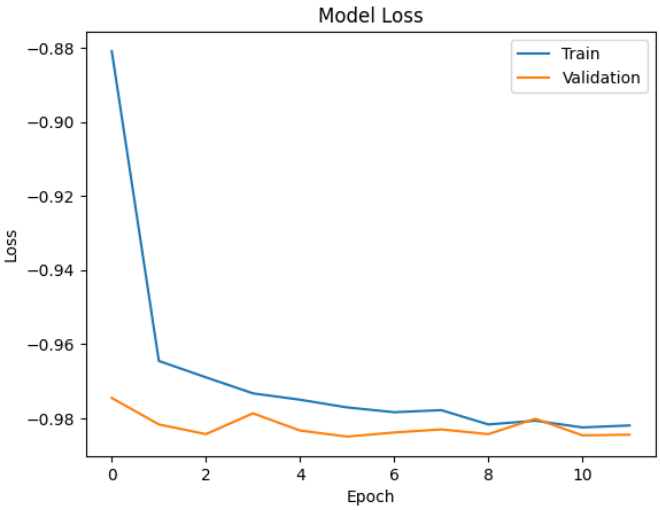
Figure 6.5 Results for SUN dataset – Hierarchical Approach

Actual Label: chimpanzee
Top 5 Prediction for an image: gorilla, gorillas, chimp**, chimpanzee,** orangutan

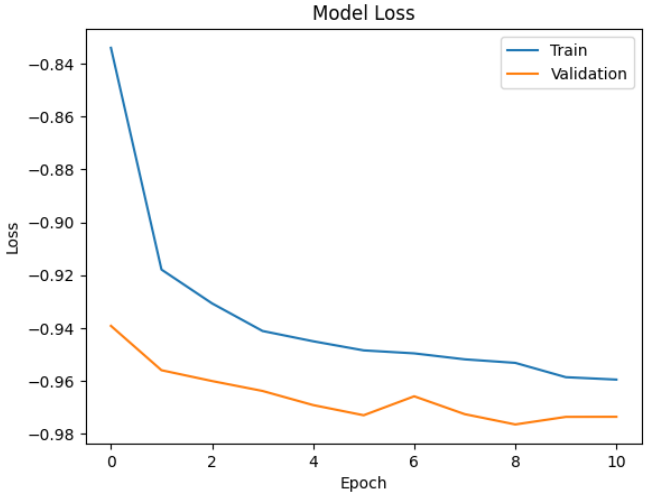
Actual Label: tiger
Top 5 Prediction for an image: tiger, bobcat, tigers, **leopard**, tigress

Figure 6.6 Results for AWA2 dataset – Hierarchical Approach

The model is trained for 20 epochs in call back mode. In call back command early stopping optimizing technique is used to avoid overfitting. Adam optimizer is used. Model loss for SUN and AWA2 datasets are shown in figures 6.7 and 6.8 respectively. Model loss reduces as model is trained with increase in epochs.



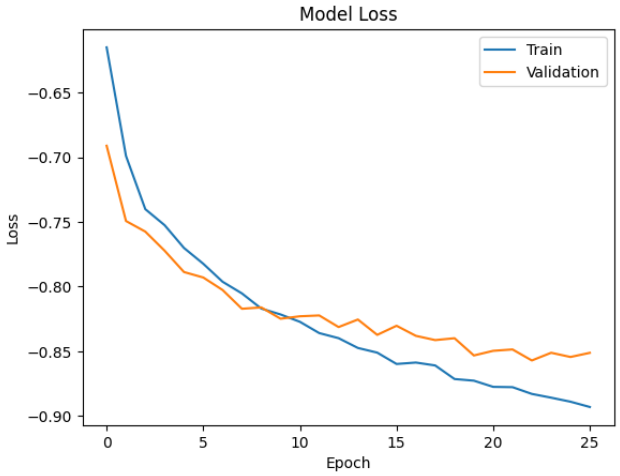
(a) Model loss – ZSL model -1



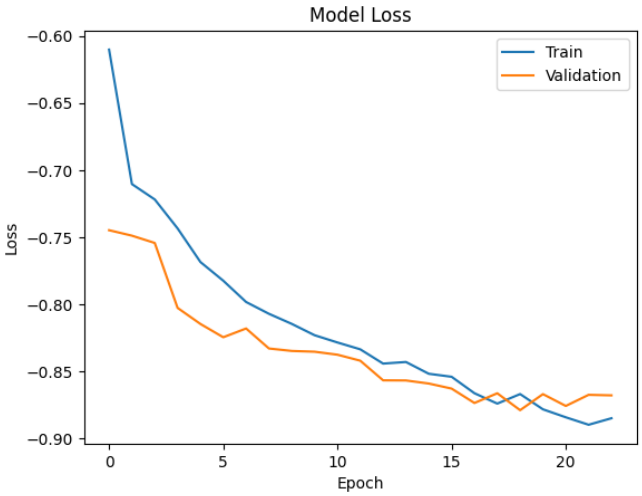
(b) Model loss – ZSL model -2

(c) Model loss - Primary model

Figure 6.7 Model loss –AWA2 dataset - Hierarchical Approach



(a) Model loss – ZSL model -1



(b) Model loss – ZSL model -2

(c) Model loss - Primary model

Figure 6.8 Model loss –SUN dataset - Hierarchical Approach

6.3 Comparative analysis of Quality, Computational and time Complexity:

Once the model is trained using Enhanced Zero shot learning method and Hierarchical Approach, model is tested for unseen classes (zero shot classes), accuracy for both the approaches are shown in Table 6.1. The hierarchical Approach shows improvement in accuracy as compared to the enhanced zero shot learning approach. Improvement in accuracy is approximately 10% to 15%. Table 6.2 shows model computational complexity parameters. Computational Complexity of model is given by FLOPs. FLOPs are increasing in hierarchical approach, but training complexity is reduced. Hyperparameter number of input image affects the training complexity of model. A hierarchical approach is trained on smaller datasets which improves training complexity. Training complexity improves approximately from 23% to 47%. Training and testing time of hierarchical approach shows enhancement as compared to Enhanced Zero shot learning method. Hence hierarchical approach gives better classification accuracy with reduced timing complexity.

TABLE-6.1 Quantitative analysis of quality

Sr.No	Method	Dataset	Accuracy
1.	**Enhanced Zero shot learning using deep neural network ResNet50.**	SUN	39.5%
		AwA2	44.786%
2.	**Hierarchical Approach for Zero Shot learning**	SUN	**50.65%**
		AWA2	**60.438%**

TABLE-6.2 Quantitative analysis of Computational and time complexity

Sr.No	Method	Dataset	Training Time (sec)	Testing Time (sec)	FLOPS	Training Complexity
1.	**Enhanced Zero shot learning using deep neural network ResNet50**	SUN	809.9183	0.2201	3.802409*10^9	84.78*10^13
		AwA2	817.5256	0.3912	3.802409*10^9	13.8426*10^14
2.	**Hierarchical Approach for Zero Shot learning**	SUN	656.3926	0.264	8.0815*10^9	**19.69*10^13**
		AWA2	4,859.7809	**0.297**	8.022*10^9	**6.499*10^14**

Chapter:7 Conclusion

Zero shot learning is a machine learning model used to classify the classes that are not included in the training set. Zero shot learning consists of visual feature extraction, semantic feature representation, visual semantic mapping, and the classification of unseen classes. The proposed Enhanced zero shot learning using deep neural network ResNet 50, which gives better accuracy as compared to the CNN based visual model. A hierarchical approach for zero-shot learning is proposed and implemented on two standard datasets, SUN and AWA2. A comparative analysis of enhanced zero shot learning using the deep neural network ResNet50 with a hierarchical approach has been done based on quality and computational complexity. The accuracy of the proposed hierarchical approach has been improved by 10% to 15% as compared to EZSL ResNet50 approach. Training complexity of the proposed hierarchical approach improves training complexity from 23% to 47%. The hierarchical approach uses three models for ZSL classification, which increases FLOPs and training time, along with improved training complexity and testing time.

Chapter 8: Bibliography

1. Palatucci M, Pomerleau D, Hinton G, Mitchell T (2009) "Zero-shot learning with semantic output codes", Adv Neural Inf Proces Syst 1:1410–1418
2. Sun Attribute Database: Discovering, annotating, and recognizing scene attributes (no date) SUN Attribute Dataset. Available at: https://cs.brown.edu/~gmpatter/sunattributes.html (Accessed: December 20, 2022).
3. Available at: https://cvml.ist.ac.at/AwA2/ (Accessed: December 20, 2022).
4. Lampert, C.H., Nickisch, H. and Harmeling, S. (2009) 'Learning to detect unseen object classes by between-class attribute transfer', *2009 IEEE Conference on Computer Vision and Pattern Recognition* [Preprint]. doi:10.1109/cvpr.2009.5206594.
5. Lampert, C.H., Nickisch, H. and Harmeling, S. (2014) 'Attribute-based classification for zero-shot visual object categorization', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3), pp. 453–465. doi:10.1109/tpami.2013.140.
6. Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. "Label-embedding for image classification". *IEEE TPAMI*, 38(7):1425–1438, 2015.
7. Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *In Advances in neural information processing systems*, pages 2121–2129, 2013.
8. Xuesong Wang, Chen Chen, Yuhu Cheng , Xun Chen and Yu Liu "Zero-Shot Learning Based on Deep Weighted Attribute Prediction" IEEE Transactions on Systems, Man, and Cybernetics: Systems (Volume: 50, Issue: 8, Aug. 2020)
9. Yongqin Xian, Tobias Lorenz, Bernt Schiele, Zeynep Akata "Feature Generating Networks for Zero-Shot Learning", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5542-5551
10. Gao, R., Hou, X., Qin, J., Liu, L., Zhu, F., Zhang, Z. (2019). A Joint Generative Model for Zero-Shot Learning. In: Leal-Taixé, L., Roth, S. (eds) Computer Vision – ECCV 2018 Workshops. ECCV 2018. Lecture Notes in Computer Science (), vol 11132. Springer, Cham. https://doi.org/10.1007/978-3-030-11018-5_50
11. Varun Khare, Divyat Mahajan, Homanga Bharadhwaj, Vinay Verma, Piyush Rai, "Generative Framework for ZSLwith Adversarial Domain Adaptation", in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV).
12. Al Machot, F.; Ullah, M.; Ullah, H. HFM: A Hybrid Feature Model Based on Conditional Auto Encoders for Zero-Shot Learning. J. Imaging 2022, 8, 171. https://doi.org/10.3390/jimaging8060171

13. LEE, R. and CHEN, I.-Y. (2020) ‘The Time Complexity Analysis of neural network model configurations’, *2020 International Conference on Mathematics and Complexity in Science and Engineering (MACISE)* [Preprint]. doi:10.1109/macise49704.2020.00039.
14. Freire, P. J., Srivallapanondh, S., Napoli, A., Prilepsky, J. E., & Turitsyn, S. K. (2022) ‘Computational Complexity Evaluation of Neural Network Applications in Signal Processing’, *ArXiv*. /abs/2206.12191
15. Shah, B. and Bhavsar, H. (2022b) ‘Time complexity in deep learning models’, Procedia Computer Science, 215, pp. 202–210. doi: 10.1016/j.procs.2022.12.023.
16. Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 427–431, 2017.
17. Socher, R.; Ganjoo, M.; Manning, C.D.; Ng, A. Zero-shot learning through cross-modal transfer. Adv. Neural Inf. Process. Syst. 2013, 26, 1–10.
18. S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.
19. *Lecture 41: Space and Computational Complexity in DNN*. nptel. Available at: https://www.youtube.com/watch?v=hGu2VIaEbHE.
20. He, K. *et al.* (2016) ‘Deep residual learning for image recognition’, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* [Preprint]. doi:10.1109/cvpr.2016.90. – ResNet FLOPs

Publication:

Jain College of Engineering
Belagavi, India

IEEE BANGALORE SECTION

4th International Conference of Emerging Technology
(INCET 2023)

26th – 28th May 2023

Certificate

This is to certify that Dr./Prof./Mr./Ms. Ansari Shaista Khanam has presented paper entitled Enhanced Zero shot learning using deep neural network ResNet50 in 4th International Conference of Emerging Technology (INCET 2023) during 26th to 28th May 2023.

Dr. Krupa Rasane
Convener - INCET 2023

Dr. J. Shivakumar
General Chair - INCET 2023