6th Semester Mini-Project

Image Classification with Active Zero Shot Learning

Indian Institute of Information Technology,Allahabad

Submitted by :

Roll No	Name
IIT2015016	Kaustubh Rakesh
IIT2015038	Vipul Kumar
IIT2015040	Tushar Minj
IIT2015043	Vikash Kr Choudhary
IIT2015068	Rishabh

Supervised by
Dr. K.P.Singh

Candidate's Declaration

This is to certify that the project report entitled

Image Classification with Active ZSL

submitted to the Dept. of Information Technology, Indian Institute of Information Technology, Allahabad in partial fulfilment of the 6th Semester Mini-Project work, is a record bonafide work carried out by:

IIT2015016	Kaustubh Rakesh
IIT2015038	Vipul Kumar
IIT2015040	Tushar Minj
IIT2015043	Vikash Kr Choudhary
IIT2015068	Rishabh

This project is our original work, and it has not been presented anywhere else for any purpose.

Supervisor's Certficate

This is to certify that the project report entitled

Image Classification with Active ZSL

submitted to the Dept. of Information Technology, Indian Institute of Information Technology, Allahabad in partial fulfilment of the 6th Semester Mini-Project work, is a record bonafide work carried out by:

IIT2015016	Kaustubh Rakesh
IIT2015038	Vipul Kumar
IIT2015040	Tushar Minj
IIT2015043	Vikash Kr Choudhary
IIT2015068	Rishabh

under my supervision and guidance.

No part of this project has been submitted elsewhere for any purpose

Dr. K.P. Singh
Dept.of I.T.
IIIT Allahabad

Contents

Abstract

Zero Shot Learning (ZSL) tries to classify novel class samples that do not have any labeled instances in training set. ZSL has gained popularity recently because of its ability to classify in domains where abundance of unlabeled data is present, but labeling is expensive. ZSL is typically achieved by transferring knowledge from seen training classes to unseen testing classes via some side information about the domain. To explore side information, we map visual features of an image to an embedding space spanned by class semantic information. Class semantic information is learnt from text corpus in an unsupervised manner. For obtaining the embedding function from visual space to semantic space, we use Mainfold regularized Cross-Modal Embedding(MCME) approach which preseves the intrinsic geometry of visual features and also aligns pairwise consistency. Further, we also propose an active class selection strategy to optimize the performance of unseen class prediction. Experimental results on two benchmark datasets AwA2 and CUB show promising results with respect to the state-of-the-art methods.

1. Introduction

We live in era of Information Technology. Almost unlimited number of images are available to us via various image sharing or social media platforms. Traditional classification methods that use Supervised Learning approach fail to classify when they encounter any novel class in testing phase. However, labeling of all the images is very expensive and impractical task. So, to solve these challenges, Zero Shot Learning has gained popularity in recent years.
Zero Shot Learning is a transfer learning approach that aims to recognize objects of the unseen classes, from which no examples are available at the training stage. Two main challenges with ZSL are:
1. Collect auxiliary information for each class, for both seen and unseen classes.
2. Find a relationship between visual samples and class auxiliary information.
For first component, auxiliary information can be achieved with: attribute based approaches, and word vector based approaches. Attribute based approaches provide intermediate-level descriptions that shares characteristics among different classes. For eg- for animal dataset, classes can be annotated by attributes like fur, paw, food , habitat etc. Word vector based approaches extract semantic information in unsupervised manner from a linguistic corpus with neural language models, like Word2Vec[4].
For second component, different cross-modal methods can be applied, which vary in loss functions. We construct a semantic embedding space to find the relationship. By doing this, we associate each class with a vector in this embedding space, so that knowledge can be transferred from visual space to class label. We use word vectors as our embedding space, and each class label is projected onto it. Furthermore, instead of passsively learning from labeled

data collected for fixed subset of classes, we actively decide which classes are most useful, for selecting labeled data to train our model for predicting remaining classes. In short, we show that unseen class predictions can be benefited from wisely selecting seen classes.

2. Motivation

The ability to learn object categories from few examples, and at a rapid pace, has been demonstrated in humans, and it is estimated that a child has learned almost all of the 10 to 30 thousand object categories in the world by the age of six. This is due not only to the human mind's computational power, but also to its ability to synthesize and learn new object classes from existing information about different, previously learned classes. Given two examples from two different object classes: one, an unknown object composed of familiar shapes, the second, an unknown, amorphous shape; it is much easier for humans to recognize the former than the latter, suggesting that humans make use of existing knowledge of previously learned classes when learning new ones.

The key motivation for the zero-shot learning technique is that systems, like humans, can use prior knowledge about object categories to classify new objects. Zero Shot learning[1] enables us to solve classification problems, when not much annotated data is available for all the classes.

3. Problem Statement

To classify images into classes whose instances are not present in training set. For auxillary information, word vector embeddings are to be used. Let,
$\mathbf{X} \in \mathbb{R}^p$ denote the visual space, and
$\mathbf{Z_s} \in \mathbb{R}^q$ where, $\mathcal{Z}_S = \{z_1, \ldots, z_K\}$ denote K discrete classes.
Then, training set
$S = \{(\mathbf{x}_i, l_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Z}_S$, where
x_i= p-dimensional visual feature extracted from image
l_i= q-dimensional word vector representation of class labels.
There are n tuples in training set.
Let, $U = \{\mathbf{x}_i\}_{i=1}^m$ be the testing set.
This set contains m samples.Further, training and testing classes are disjoint.
Our aim is to learn an embedding function :

$$f : \mathbb{R}^p \to \mathbb{R}^q$$

with training set S.
Subsequently, ZSL[1] is performed by projecting visual feature of testing image to the word vector space with learned embedding function ,f, and then matching the embedding vectors with the vector representation of unseen classes.

4. Literature Review

4.1 Introduction to Zero Shot Learning[1]

The ability to classify instances of an unseen visual class is called zero-shot learning.ZSL aims at learning classification models for the novel classes with no labeled data for training. It learns to predict classes indirectly via their semantic attributes. The two key issues should be considered for ZSL:

- Collect auxiliary information for each class,including seen classes and unseen classes.
- Connect the visual samples and the class auxiliary information this connection is often achieved by constructing a semantic embedding space.

4.2 Approach to Zero Shot Learning[1]

From the perspective of embedding space, ZSL approaches fall into the following three categories:

- Attribute-based approaches
- Word-vector-based approaches
- Approaches using both attributes and word vectors

4.2.1 Attribute based approaches

These approaches build an attribute space for the seen and unseen classes, enabling learning unseen classes only with their descriptions. Since it requires human label, its applicability for large-scale Zero Shot Learning is limited.

One of the typical approaches is Directed Attribute Prediction (DAP) ,in which the attributes serve as an intermediate space between the visual features and the labels. In this case, a probabilistic classifier is learned for each attribute at the training stage, and the unseen classes are then inferred with the learned estimators.

4.2.2 Word-vector based approaches[3]

Attribute-based approaches are giving way to word-vector-based approaches with the rapid progress of computational linguistics techniques.
Word Vector is a distributed representation of a word .Each word is represented by a distribution of weights across the elements corresponding to a vector with some given dimension (say 1000).So instead of a one-to-one mapping between an element in the vector and a word, the representation of a word is spread across all of the elements in the vector, and each element in the vector contributes to the definition of many words.

We find that the learned word representations in fact capture meaningful syntactic and semantic regularities in a very simple way. Specifically, the

regularities are observed as constant vector offsets between pairs of words sharing a particular relationship.

4.2.3 Attribute and Word-vector based approaches

Considering that the attribute and word vector may provide complementary information, some recent work pays attention to exploiting both of them to enhance the performance of Zero Shot Learning.

5. Proposed methods

5.1 *Our Approach to ZSL[1]*

In our approach, the sample x_i can be projected into the class embedding space by a linear model, i.e.,

$$f(x_i) = W^T x_i$$

, where $\mathbf{W} \in \mathbb{R}^{p \times q}$ denotes the transformation matrix. Ideally, W can be estimated by maximizing the correlation between $f(x_i)$ and the corresponding prototype y_{li}, where l_i is the class label of sample x_i in S. The correlation between $f(x_i)$ and the class prototypes is defined as follows:

$$\rho\left(f\left(\mathbf{x}_i\right), \mathbf{y}_{l_i}\right) = \frac{\mathbf{x}_i^{\mathrm{T}} \mathbf{W} \mathbf{y}_{l_i}}{\sqrt{\left(\mathbf{W}^{\mathrm{T}} \mathbf{x}_i\right)^{\mathrm{T}} \left(\mathbf{W}^{\mathrm{T}} \mathbf{x}_i\right)} \sqrt{\mathbf{y}_{l_i}^{\mathrm{T}} \mathbf{y}_{l_i}}} \tag{5.1}$$

Since the prototype for each class has been obtained with neural language models ${y^T}_{li}\, y_{li}$ is a constant. Thus, maximizing(5.1) is turned to the following optimization problem:

$$arg \max_{\mathbf{W}} \mathbf{x}_i^{\mathrm{T}} \mathbf{W} \mathbf{y}_{l_i} \quad s.t.\ \mathbf{x}_i^{\mathrm{T}} \mathbf{W} \mathbf{W}^{\mathrm{T}} \mathbf{x}_i = 1 \tag{5.2}$$

Equation(5.2) reflects the potential maximal correlation between a visual sample x_i and its corresponding prototype y_{li}. Furthermore, this correlation should be larger than those with the other class prototypes y_j in the training set, such that the objective function is further defined as follows:

$$\begin{gathered} arg \max_{\mathbf{W}} \sum_{i=1}^{n} \sum_{j=1}^{K} (\mathbf{x}_i^{\mathrm{T}} \mathbf{W} \mathbf{y}_{l_i} - \mathbf{x}_i^{\mathrm{T}} \mathbf{W} \mathbf{y}_j) - \tfrac{\lambda_1}{2} \parallel \mathbf{W} \parallel_F^2 \\ s.t.\ \mathbf{X}_S^{\mathrm{T}} \mathbf{W} \mathbf{W}^{\mathrm{T}} \mathbf{X}_S = 1, \end{gathered} \tag{5.3}$$

where $\mathbf{X}_S = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ denotes all the training data, 1 is a positive regularization parameter, l_F indicates the Frobenius Norm, and $\parallel \mathbf{W} \parallel_F^2$ is used to avoid the overfitting. From the perspective of space transformation, the objective function of Eq(5.3) forces the training inputs close to their corresponding semantic embedding vectors while far away from the other class semantic vectors.This pairwise objective function learns a discriminative

function to map the visual sample to the class label space.

Further, considering that structure preserving is useful in embedding methods , we add a manifold term to preserve the locally visual structure:

$$arg\min_{\mathbf{W}} \sum_{i,j=1}^{n} S_{ij} \parallel \mathbf{W}^{\mathrm{T}} (\mathbf{x}_i - \mathbf{x}_j) \parallel_2, \tag{5.4}$$

where S_{ij} denotes the similarity between x_i and x_j. Specifically, we construct a nearest neighbor graph model, where each vertex represents a sample. More precisely, if x_i and x_j belong to the same class, and if x_i is among the knearest neighbors of x_j or x_j is among the knearest neighbors of x_i, we define the cosine distance to measure the relationship between two samples, i.e., $S_{ij} = \cos(x_i, x_j)$; otherwise $S_{ij} = 0$. This reflects the manifold assumption that visually similar images are more likely embedding closer in the class semantic space, i.e., having a small distance measured by $\parallel \mathbf{W}^{\mathrm{T}} (\mathbf{x}_i - \mathbf{x}_j) \parallel_2$ Eq(5.4) can be deduced as:

$$arg\min_{\mathbf{W}} \sum_{i,j=1}^{n} S_{ij} \parallel \mathbf{W}^{\mathrm{T}} (\mathbf{x}_i - \mathbf{x}_j) \parallel_2 = \mathbf{W}^{\mathrm{T}} \mathbf{X}_S \mathbf{L} \mathbf{X}_S^{\mathrm{T}} \mathbf{W}, \tag{5.5}$$

where the Laplacian matrix L is defined as L = D-S, and $\mathbf{D}_{ii} = \sum_j S_{ij}$ is the degree of the i^{th} data. Finally, simultaneously taking Eq(5.3) and Eq(5.5) together leads to the final model:

$$\begin{aligned} arg\min_{\mathbf{W}} \sum_{i=1}^{n} \sum_{j=1}^{K} & -\left(\mathbf{x}_i^{\mathrm{T}} \mathbf{W} \mathbf{y}_{z_i} - \mathbf{x}_i^{\mathrm{T}} \mathbf{W} \mathbf{y}_j\right) + \tfrac{\lambda_1}{2} \parallel \mathbf{W} \parallel_F^2 + \tfrac{\lambda_2}{2} \mathbf{W}^{\mathrm{T}} \mathbf{X}_S \mathbf{L} \mathbf{X}_S^{\mathrm{T}} \mathbf{W} \\ & s.t. \qquad \mathbf{X}_S^{\mathrm{T}} \mathbf{W} \mathbf{W}^{\mathrm{T}} \mathbf{X}_S^{\mathrm{T}} = 1, \end{aligned} \tag{5.6}$$

where λ_2 is a positive parameter to reconcile the two losses in Eq(5.3) and Eq(5.5). Specifically, the first term ensures that labeled samples are projected closer to their corresponding class prototypes than to any other class prototypes. The middle term is introduced to avoid the overfitting. The last one holds the locally visual structure. It can be seen that Eq(5.6) is typically a Regularized Least Square (RLS) problem that can be solved efficiently.

The optimal W for Eq(5.6) is denoted as W^* and can be derived as follows with the method of Lagrange multipliers:

$$\mathbf{W}^* = \left(\mathbf{X}_S\mathbf{X}_S^{\mathrm{T}} + \lambda_1\mathbf{I} + \lambda_2\mathbf{X}_S\mathbf{L}\mathbf{X}_S^{\mathrm{T}}\right)^{-1}\left(\mathbf{X}_S\mathbf{Y}_S^{\mathrm{T}} - \mathbf{X}_S\hat{\mathbf{Y}}_S^{\mathrm{T}}\right), \tag{5.7}$$

where I is a unit matrix, and $\mathbf{Y}_S = [\mathbf{y}_{l_1}, \ldots, \mathbf{y}_{l_n}] \in \mathbb{R}^{q\times n}$ is the average of all the class prototypes of seen classes, i.e., $\hat{\mathbf{y}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{y}_k$

5.2 *Tweaking Active ZSL[2]*

We propose to iteratively add from the pool of unseen classes more labels that are informative about the remaining unseen classes. The connectivities between the classes can be indicators of information about one class carried by others. Specifically, the connectivity between the ith unseen class and the other unseen classes can be measured by various centrality metrics of the corresponding ith node on the subgraph of $\mathcal{G}$ consisting of all unseen classes. For example, the degree centrality of the ith unseen class can be calculated as $\sum_{j=1}^{k} K_{ij}^{\mathcal{U}\mathcal{U}}$, where k is the current number of unseen classes. We have used the strategy max-deg-uu which selects the unseen class with the maximal degree. This selection strategy does not consider the distribution of the class similarities between class i and others: class i can be strongly connected to only a few unseen classes with high weights, but barely so to the remaining majority classes. Such a class can still have a high degree, but does not add much information about the remaining unseen classes.

Instead, we use entropy to characterize how the connectivities $K_{ij}^{\mathcal{U}\mathcal{U}}, j = 1, \ldots, k$ distribute. First, the similarities in $K^{\mathcal{U}\mathcal{U}}$ are normalized to a probability distribution:

$$P^{\mathcal{U}\mathcal{U}} = \mathrm{diag}(1^\top K^{\mathcal{U}\mathcal{U}})^{-1}K^{\mathcal{U}\mathcal{U}}, \tag{5.8}$$

where diag(v) denotes the diagonal matrix with diagonal elements being the entries of the vector v , and is the all-one vector. Then, we calculate the

entropy of $K_{ij}^{\mathcal{UU}}$ for the ith unseen class:

$$H(i) = -\sum_{j=1}^{k} P_{ij}^{\mathcal{UU}} \log P_{ij}^{\mathcal{UU}}, \quad i = 1, \ldots, k. \tag{5.9}$$

We select the top c classes that have the highest entropies and move them from U to S . Thus, by applying above modifications, the training set gets enriched.

6. Software Requirements

Software Used

- OS used - Ubuntu 16.04 or Windows 10
- Python 3.5.2
- Pycharm Community Edition 2017

Libraries Used

- Tensor Flow
- Keras
- NumPy
- SciPy
- Sklearn

System Specification

- CPU Intel i5 5th gen
- GPU Nvidia Geforce 920M
- RAM 12 GB

7. Experiments

Datasets used

We have used the following two datasets:

- **AwA2** (Animals with Attributes 2)
  50 classes of animal images (30,475 images)
  40 Training classes and 10 testing classes.

- **CUB** (Caltech-UCSD Birds)
  Contains 11,788 images of 183 bird species . 123 Training classes and 60 Testing classes.

For each dataset,training and testing sets belong to mutually exclusive classes. For obtaining visual features, we use VGG19[5] features, which are 4096-dimension features from the 'fc2' layer of very deep19-layer CNN pre-trained on ILSVRC2014.
No task specific preprocessing was done.
Word2Vec trained on Google News was used to generate 300-dimension word vectors of labels of classes.

8. Results and Discussion

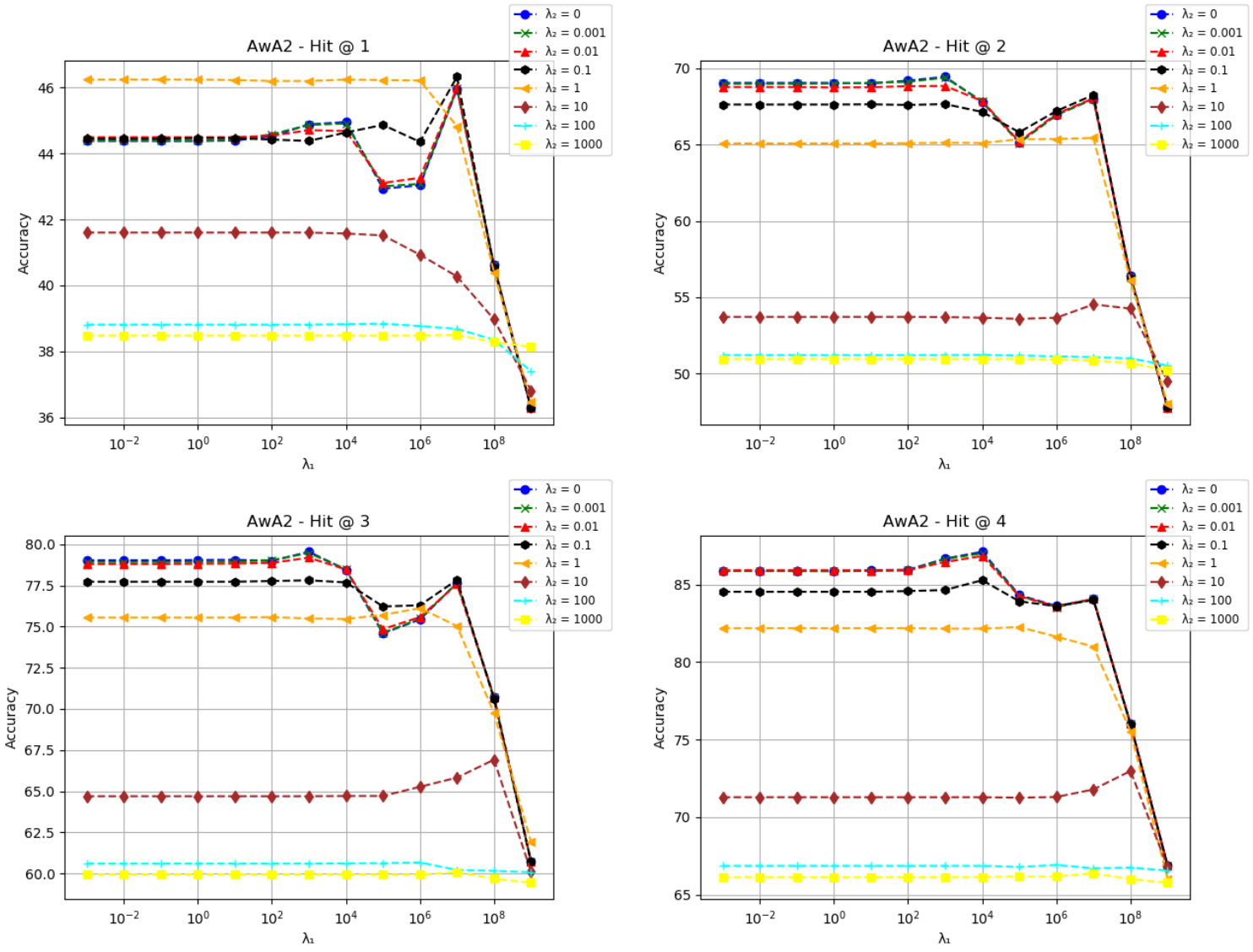
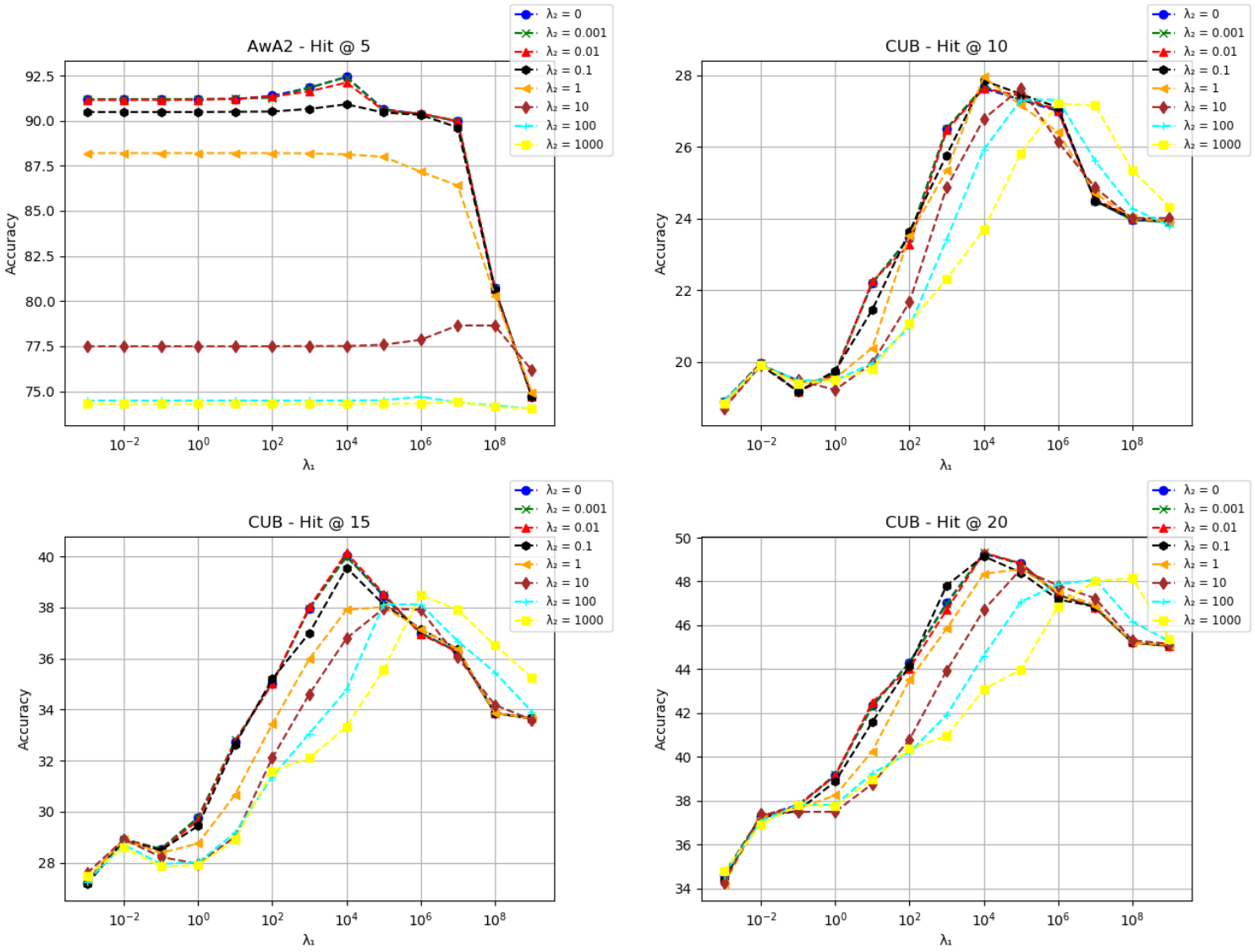


Figure 8.1: The impacts of the parameters λ_1 and λ_2 on AwA2 dataset.

In the above pictures the variables are λ_1 and λ_2.While calculating optimal $\mathbf{W}$ i.e. $\mathbf{W}^*$,we have used λ_1 = (10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9) and various values of λ_2 = (0, 0.001, 0.01, 0.1, 1, 10, 100, 1000). Furthermore, during calculation of Laplacian matrix there was notion of calculating k nearest neighbors for which values of k = (50, 100, 150, 200) were used . Above graphs are obtained using various values of λ_1

and λ_2 (with k=100). Another parameter being varied is the "hit" which takes values 1 to 5 (hit@1, hit@2, hit@3, hit@4, hit@5) .

We performed the same for CUB datasets with a slight variation of Hit 10, 15 and 20 (hit@10, hit@15, hit@20). The images are depicted below.



Iteration	Without Active ZSL	With Active ZSL
1	81.63	89.84
2	81.63	80.67
3	81.63	72.98
4	81.63	91.57
5	81.63	81.25

Table 8.1: Comparing Accuracy with Active ZSL on AwA2

We performed Active ZSL for division of testing and training classes for both AWA2 and CUB .

After actively selecting testing and training classes we apply our previous method on so called enriched training and testing classes . This was done various times (5 iterations). Results are summarized for AWA2 above and CUB datasets below.

Iteration	Without Active ZSL	With Active ZSL
1	47.55	71.08
2	47.55	73.02
3	47.55	68.83
4	47.55	71.62
5	47.55	74.71

Table 8.2: Comparing Accuracy with Active ZSL on CUB

9. Conclusion

A novel manifold regularized cross-modal embedding approach for word-vector-based ZSL is presented in this paper. Furthermore, instead of passively using predefined training and testing classes, we employ active class selection strategy to improve our results. Extensive experiments on the benchmark datasets, AWA2 and CUB shows that Actice ZSL improves the accuracy of traditional ZSL.

For future work,we can do the following :

- While calculating the similarity between two vectors, the distance measure being used is cosine distance.Instead of that,other distance measures can also be used,such as Dot Product.
- While actively selecting our training classes,we can use min-deg-us and uncertainty instead of max-deg-uu[2].

we can change the similarity measure used in calculating Lagrangian matrix to inner product similarity metrices.

Bibliography

[1] Z. Ji, Y. Yu, Y. Pang, J. Guo, and Z. Zhang. Manifold regularized cross-modal embedding for zero-shot learning. Information Sciences, 378:48 58, 2017

[2] Sihong Xie and S Yu Philip. Active zeroshot learning: a novel approach to extreme multi-labeled classification. International Journal of Data Science and Analytics

[3] https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

[4] http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

[5] https://keras.io/applications/#vgg19