Semi-Supervised Learning via Triplet Network Based Active Learning (Student Abstract)

Divyanshu Sundriyal,[1] **Soumyadeep Ghosh,**[1] **Mayank Vatsa,**[2] **Richa Singh**[2]
[1]IIIT Delhi,
[2]IIT Jodhpur
{divyanshu18096, soumyadeepg}@iiitd.ac.in, {mvatsa, richa}@iitj.ac.in

Abstract

In recent years deep learning models have pushed state-of-the-art accuracies for several machine learning tasks. However, such models require a large amount of data for training. Active learning techniques help us in utilizing unlabelled data which may result in an improved classification model. In this research, we present an active learning algorithm which can help in increasing performance of deep learning models by using large amount of available unlabelled data. A novel active learning algorithm (Triplet AL) is proposed which uses a triplet network to select samples from an unlabelled data set. Previous active learning methods rely on classification model's final prediction scores as a measure of confidence for an unlabelled sample. We propose a more reliable confidence measure called Top-Two-Margin which is given by Triplet Network. The proposed algorithm shows improved performance compared to other active learning approaches.

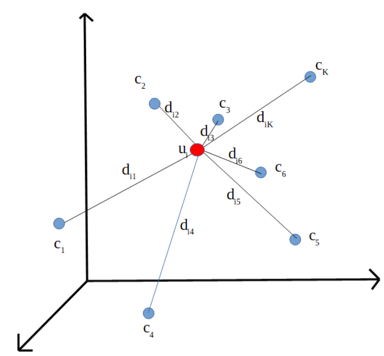

Figure 1: Distance of unlabelled data point u_i to K class centers. Embedding space of size N is illustrated using an embedding space of size 3 for representational purposes.

Introduction

Semi-supervised learning involves utilization of a small labelled set of data along with a large unlabelled set which can be used to increase model performance. There are many cases in the real world where the number of labelled samples available is not adequate to train a good classification model. For example, most object recognition databases (eg: the CIFAR-10 dataset) contains several thousand labelled images. However in internet there are a huge amount of such images present but these are unlabelled. The proposed algorithm can be used to select samples from the unlabelled set in iterative manner and pseudo label the selected samples such that these pseudo labelled samples when added to current initial labelled set would allow us to train an improved supervised classification model.

Proposed Algorithm

Conventional active learning methods consists of a classification model M_1 whose purpose is to learn a function $g_\theta : \mathbb{R}^Z \longrightarrow \mathbb{R}^K$, where Z is the dimensionality of input manifold and K is the number of classes. In proposed novel **Triplet AL** method, along with classification model M_1, a Triplet model M_2 is also used. Purpose of model M_2 is to learn a function $f_\phi : \mathbb{R}^Z \longrightarrow \mathbb{R}^N$ using triplet loss (Schroff, Kalenichenko, and Philbin 2015), where Z is the dimensionality of the input manifold and N is the size of embedding space. The proposed algorithm consists of the following steps:

- Train classification model M_1 and embedding model M_2 using initial labelled set L. Find class centers C_1,C_2,C_3...C_K in the embedding space. To find center of a class, find embeddings of all samples belonging to that class and take mean of the embeddings.
- For each unlabelled sample p_i, convert it to its embedding e_i using embedding model M_2. Find its distance to all the K centers $[d_{i1}, d_{i2}, d_{i3}, ...d_{ik}]$ in the embedding space as shown in Figure 1. Sort distances in ascending order $[D_{i1}, D_{i2}, D_{i3}, ..., D_{iK}]$. Distance D_{i1} is the distance of unlabelled sample p_i from it's nearest class center and distance D_{i2} is the distance of unlabelled sample p_i from it's second nearest class center and so on.
- We propose a metric **Top-Two-Margin** as a difference between distance from nearest center and distance from second nearest center. Thus Top-Two-Margin of an unlabelled sample p_i is calculated as, $Top - Two - Margin_i = D_{i2} - D_{i1}$.
- Top-Two-Margin indicates confidence of Model M_2 on

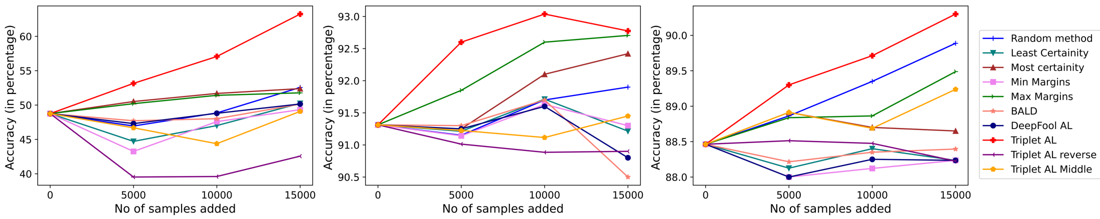

Figure 2: Results on the STL-10 Dataset showing changes in accuracy of model M_1 when samples are added from unlabelled dataset for 3 different architectures namely, M_1 - LeNet-5, ResNet-152 and DenseNet-121.

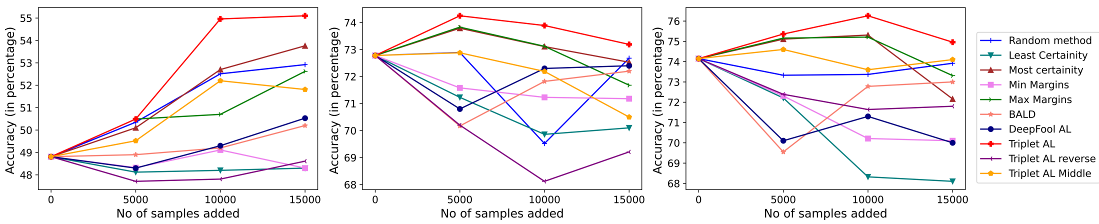

Figure 3: Results on the CIFAR-10 Dataset showing changes in the accuracy of model M_1 when samples are added from unlabelled dataset for 3 different architectures of M_1 - LeNet-5, ResNet-152 and DenseNet-121.

class of unlabelled sample p_i. If Top-Two-Margin is large, then model M_2 is sure that the sample belongs to the class of its nearest class center. If Top-Two-Margin has a lower value, then model M_2 is very less certain that it belongs to class of its nearest class center. In each iteration of active learning, we select S samples whose Top-Two-Margin is largest i.e. model M_2 is most confident about their class because model M_2 is used as Oracle.

- For pseudo labelling a selected sample s_i, we convert it to its embedding e_i using M_2. We calculate the distance of e_i to all class centers. We then assign a class C to the sample s_i whose center is closest to the sample in embedding space. The pseudo labelled points are then added to the labelled set and are removed from unlabelled set, so that they are not chosen again. Models M_1 and M_2 are retrained using new labelled set.

Results

To evaluate the proposed algorithm, we used the STL-10 dataset (Coates, Ng, and Lee 2011) and CIFAR-10 dataset (Krizhevsky and Hinton 2009). For STL-10 dataset, we have 5000 initial labelled samples and 100,000 unlabelled samples. For CIFAR-10 dataset, we have partitioned the training set of 50000 samples into 5000 labelled and 45,000 unlabelled samples. To demonstrate architectural neutrality, we used 3 different architectures for classification model M_1 in our experiments- LeNet-5, ResNet-152 and DenseNet-121. We compared proposed Triplet AL algorithm with other active learning methods like Random sampling, Least Certainty, Most Certainty, Minimum margin, Maximum margin and two recent active learning algorithms BALD (Gal, Islam, and Ghahramani 2017) and Deep Fool active learning (Ducoffe and Precioso 2018). We performed three active learning iterations and in each iteration we selected 5000 samples from the unlabelled set.

The proposed Triplet AL method outperformed other active learning approaches used in this work. Figures 2 and 3 show that increase in accuracy is higher when samples are selected from unlabelled dataset using proposed algorithm. For STL-10 dataset, proposed Triplet AL method is able to increase accuracy (in percentage) of classification model M_1 by 14.55, 1.72 and 1.84 when classification model used is LeNet-5, ResNet-152 and DenseNet-121 respectively. For CIFAR-10 dataset, proposed Triplet AL method is able to increase accuracy (in percentage) of classification model M_1 by 6.29, 1.47 and 2.11 when classification model used is LeNet-5, ResNet-152 and DenseNet-121 respectively.

Conclusion and Future Work

We have exhibited that using an unlabelled set, the proposed Triplet AL algorithm is able to improve the performance of a classification model on unseen test data. The proposed algorithm outperforms other popular active learning algorithms. In future, we plan to use a multi task learning model instead of separate classification and triplet models.

References

Coates, A.; Ng, A.; and Lee, H. 2011. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *AISTATS*.

Ducoffe, M.; and Precioso, F. 2018. Adversarial Active Learning for Deep Networks: A Margin Based Approach. *arXiv preprint arXiv:1802.09841* .

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep Bayesian Active Learning with Image Data. In *ICML*, 1183–1192.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A Unified Embedding for Face Recognition and Clustering. In *IEEE CVPR*, 815–823.