Active Learning on a Minimal Budget

Tala Al-Sharif

Introduction

Deep learning architecture had a huge impact on machine learning success, however it requires to be trained on a large quantity of data. Thus, Active learning framework gives the opportunity to achieve data efficiency in deep learning.

Active learning targets a real-world situation where labeled data are scarce but unlabeled data are abundant and labeling a large amount of data is very difficult, costly and time consuming. Thus, it is very attractive to propose a proper labeling scheme to reduce the number of labels required in order to train a classifier.

Data

The dataset used is the MNIST dataset for handwritten digits.

http://yann.lecun.com/exdb/mnist/

MNIST dataset has a 55,000 labeled images in the training set and 10,000 images in the test set. However, we will assume only few images are labeled to implement active learning techniques.

Predictive Model

The model used to build an image classifier is the Conventional Neural Network (CNN) using the machine learning library TensorFlow.

Initially, the model was trained on the entire MNIST dataset to make sure that our CNN model can generalize from the training data to unseen data by measuring the performance of the learned model on the test set that is separate from the training set.

Active Learning

In this project, the focus is on a pool-based active learning. Pool-based active learning consists of two main engines, a learning engine and sampling engine.

Initially, the model is introduced and trained on a small set of labelled data to build a predictive model. Then the learning algorithm will decide which instances from the unlabeled pool to label. The decision is based on evaluating the informativeness and the value of instances. However, there is in fact no exact measure of the value, but we can estimate the value by implementing the uncertainty measures which assumes that instances with higher classification uncertainty are most critical to the label. Then, those instances are added into the labelled set and let the learning engine constructs a new classifier based on the updated set. After each iteration, the performance of the model is improving because of the increase in number of the labelled data. This process is run repeatedly, and the active learner iteratively selects informative query instances until we run out of budget.

This technique is computationally intensive and requires evaluating the entire data set at each iteration. However, it can be applied to a wide variety of applications from text classification, image classification, speech recognition to cancer diagnosis.

Query Selection Strategies

Random Sampling

The active learner chooses instances at random to query their actual labels. Then add those instances into the labelled set to retrain the classifier.

This method doesn’t pick instances that are representative of the underlying distribution. However, as seen in the next graph the learning climbs up substantially till it reaches 94% accuracy score on the test set.

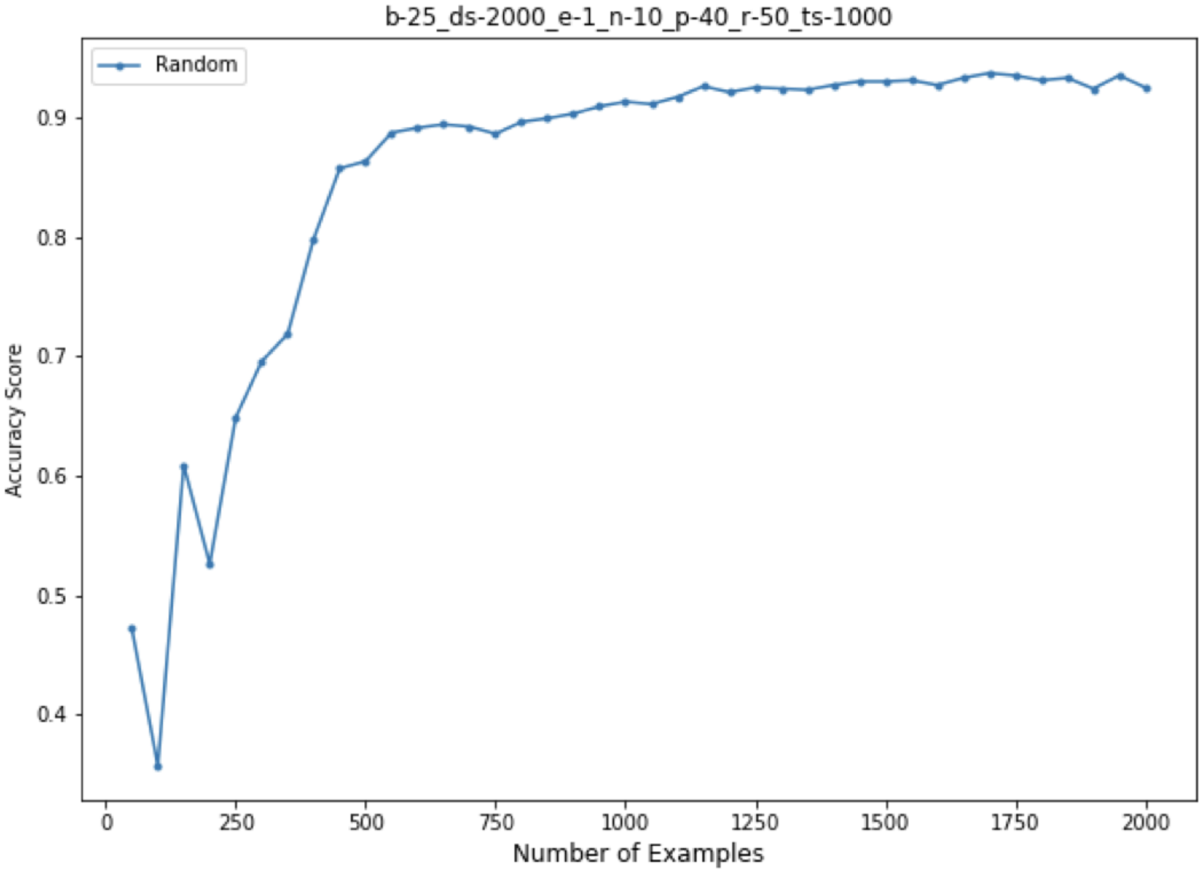

Figure 1: Learning Curve of Random Sampling

Most Confident

In this method we will query instances which the active learner is most certain about. In the introduction I mentioned that the decision for query selection will be based on uncertainty measure but out of curiosity wanted to see how the model will perform when choosing instances with high probability prediction.

In figure 2, we can notice that random sampling acted better throughout the whole learning curve and that denoted to the fact that most confident approach is querying irrelevant or redundant instances that add no value to the existing labelled set.

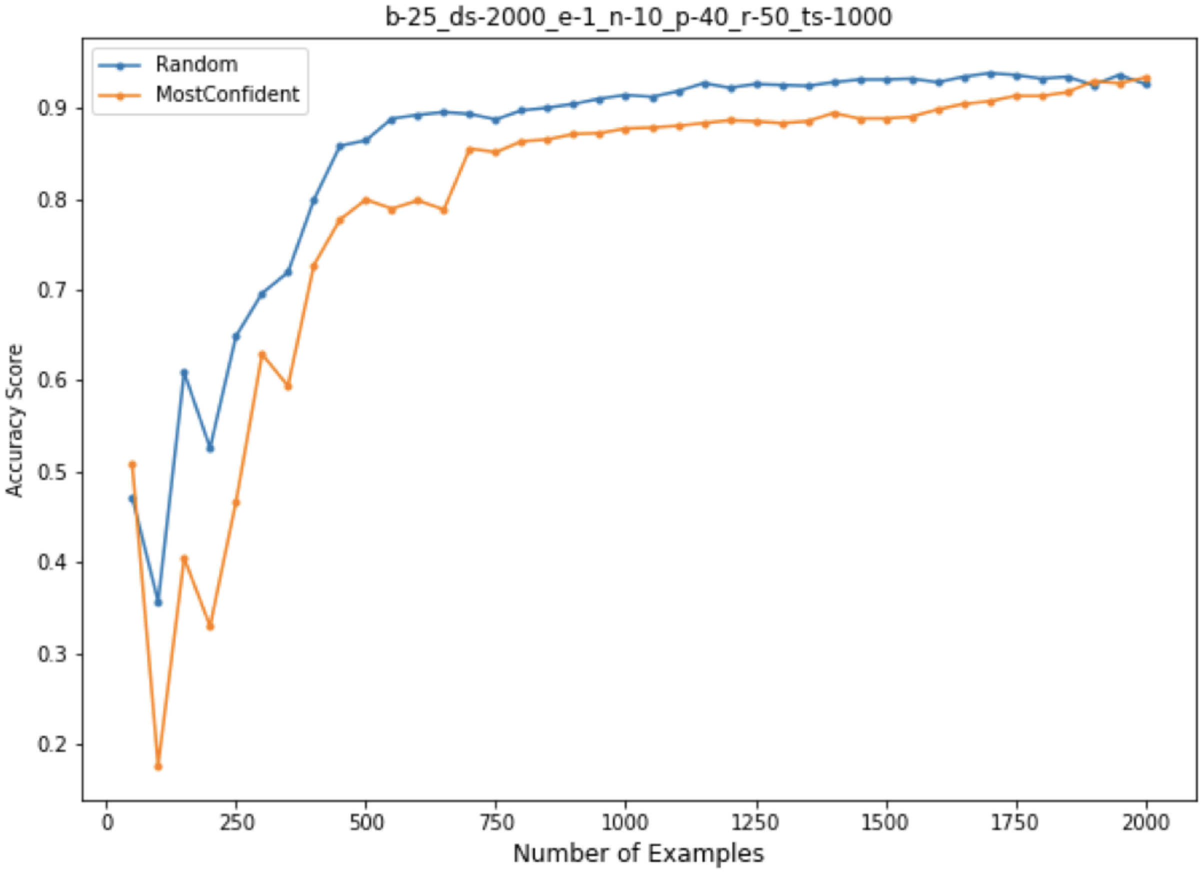

Figure 2: Learning Curve of Random Sampling and Most Confident

Uncertainty Sampling

The active learner queries instances for which it has least confidence in its most likely label, so the focus on instances closest to its decision boundary, assuming it can adequately explain those in other parts of the input space of the unlabeled ones. As a result, it avoids requesting labels for redundant or irrelevant instances, and achieves a higher accuracy than random sampling.

Several ways to measure uncertainty, label probability is our chosen technique which can be divided into least confident and maximum entropy.

Least confident queries an instances whose predictions are the least confident. This method considers information about the most probable label and discards information about the remaining label distribution.

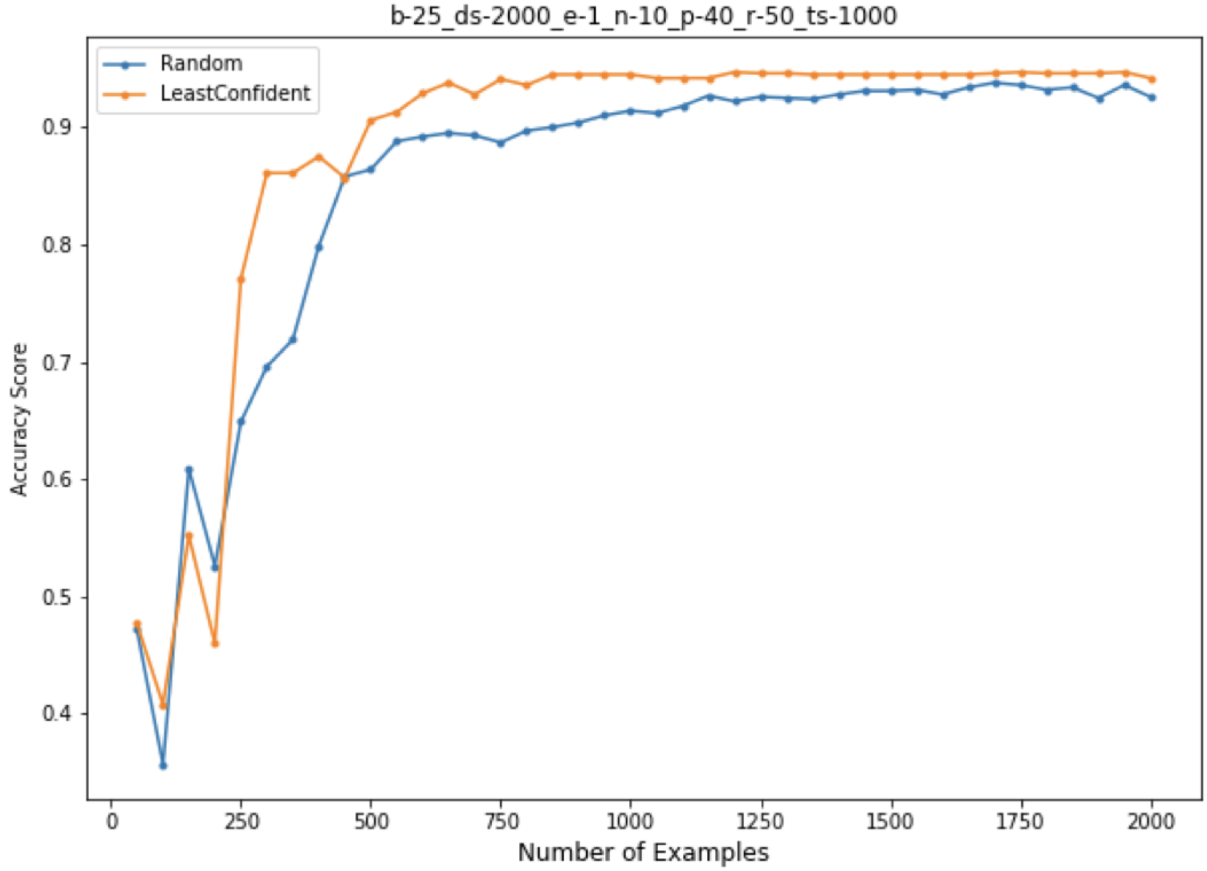

Figure 3: Learning Curve of Random Sampling and Least Confident

Maximum entropy queries instances whose label entropies is maximum. The entropy measure is highly influenced by the probability values of the unimportant classes. In other words, when computing entropies, the small probability values of unimportant classes will contribute to a higher entropy score even though the classifier is much confident about the classification example.

Entropy is a measure of uncertainty. Therefore; if an unlabeled point in the pool set has a distribution with a higher entropy then the classifier is more uncertain about its class membership.

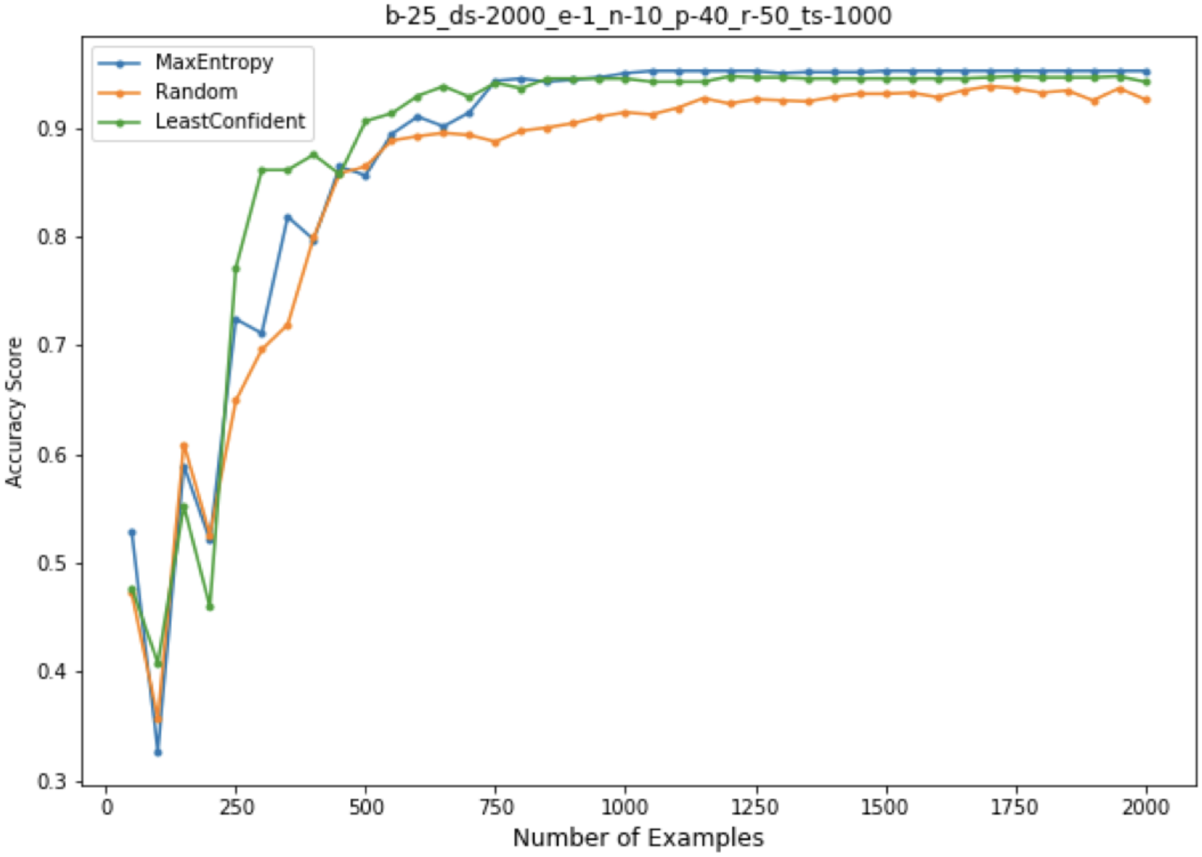

Figure 4: Learning Curve of Random and Uncertainty Sampling

Uncertainty sampling outperformed random sampling and reached the non-change state in the accuracy with fewer labelled instances with only 800 examples and 97% accuracy score.

Graph-Based Approach using Pairwise distance

In the graph-based method, instances are represented by vertices in a graph with edges encode distance between nodes, distance serves as a notion of similarity (or dissimilarity) between instances. We captured this notion of similarity using pairwise distances. Distances are computed for the predictions on the entire dataset then query to label the instance that is furthest from the labeled instances then update the labelled set with the new instance and choose the next instance that is furthest from the updated labeled set. This process is repeated till we reach the size selection needed, then update the classifier with the new labeled data to leverage its knowledge to choose which instances to query next.

This method handles high-dimensional data, and therefore; it outperformed other active learners throughout the whole learning curve. In addition, label uncertainty on instances can be reduced because the label of the unlabelled instances is predicted by nearby labelled instances.

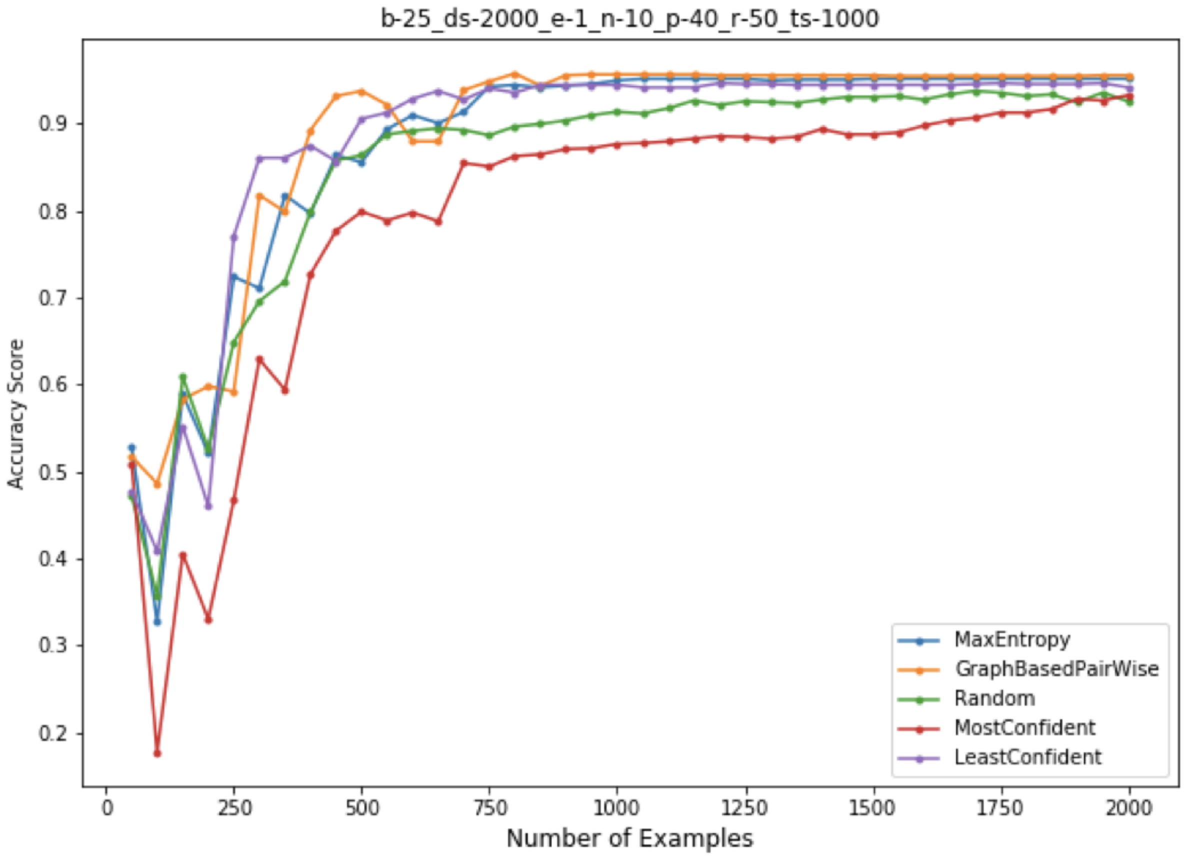

Figure 5: Active Learners Curves

I use T-SNE library to visualize the images. The green color represents the unlabeled instances which is in our case, 2000 examples. The red color represent the labelled instances with only 50 labelled examples and the blue color represents the instance the active learner will query its actual label.

The below graph was captured in the very first iteration where are querying 50 instance one at a time. It is noteworthy that the location of picking instances is moving further away from the labelled instances after each update of the labelled set.

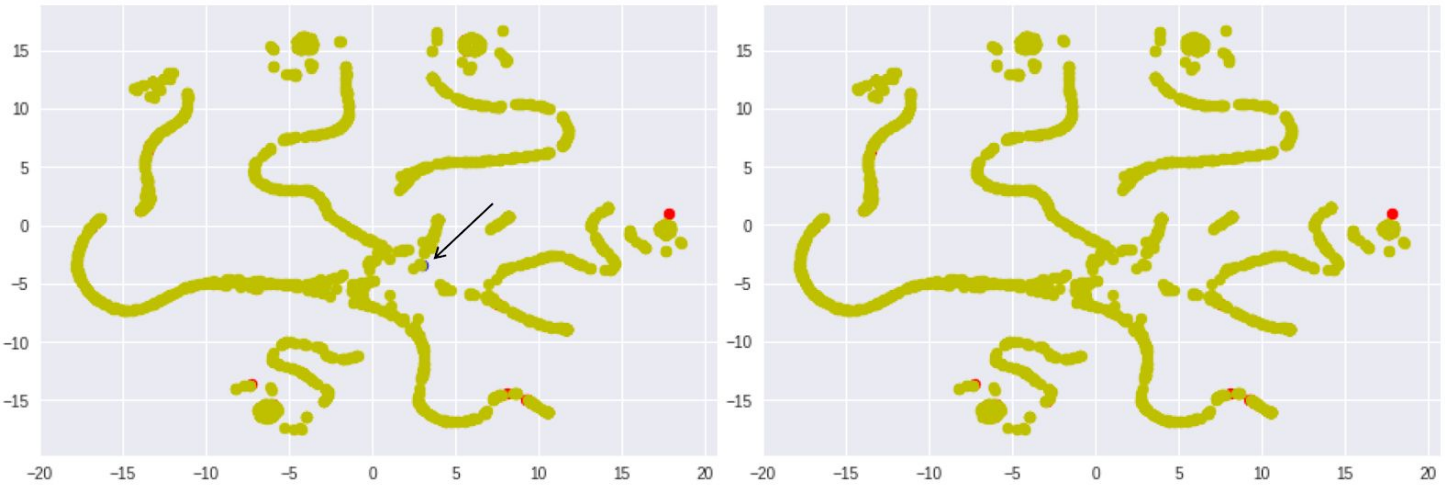

Figure 6: Visualization of instances for the first and second instance

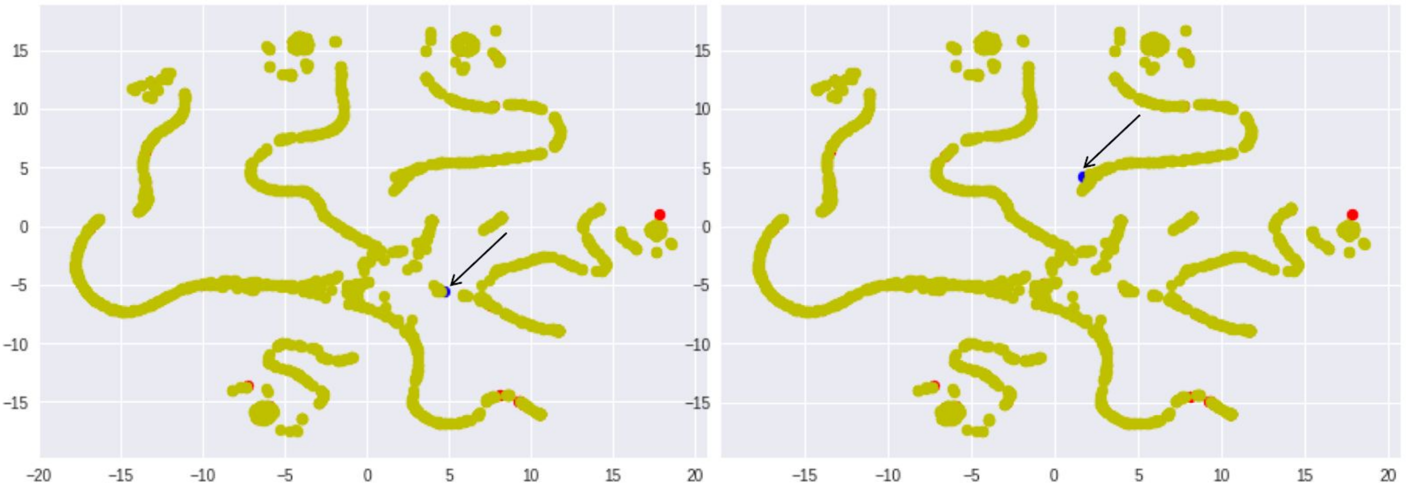

Figure 7: Visualization of instances for the fifth and sixth instance

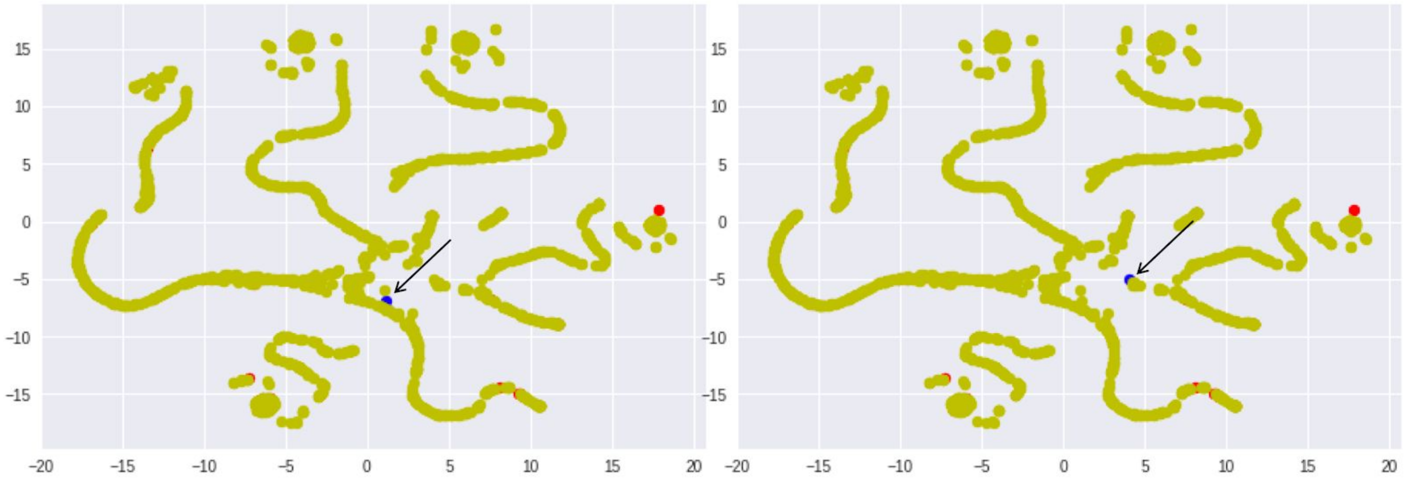

Figure 8: Visualization of instances for the seventh and twenty-first instance

Exploratory Data Analysis

Stopping Criteria

It doesn’t mean if we are getting a high accuracy score on the test set to stop labeling. You can let the active learner querying for more labels after the performance does not improve significantly. In other words, after a certain threshold where the accuracy reaches a non-change state.

In figure 9, we can stop requesting for more labels with approximately 800 examples since the accuracy reached the non change state. In figure 10, will take a close up to the first 500 examples to demonstrate the stopping criteria with different data size. We notice that none of the active learners reached the steady state and random sampling performed better than maximum entropy compared to figure 9 where 2000 examples where included in the labelled set.

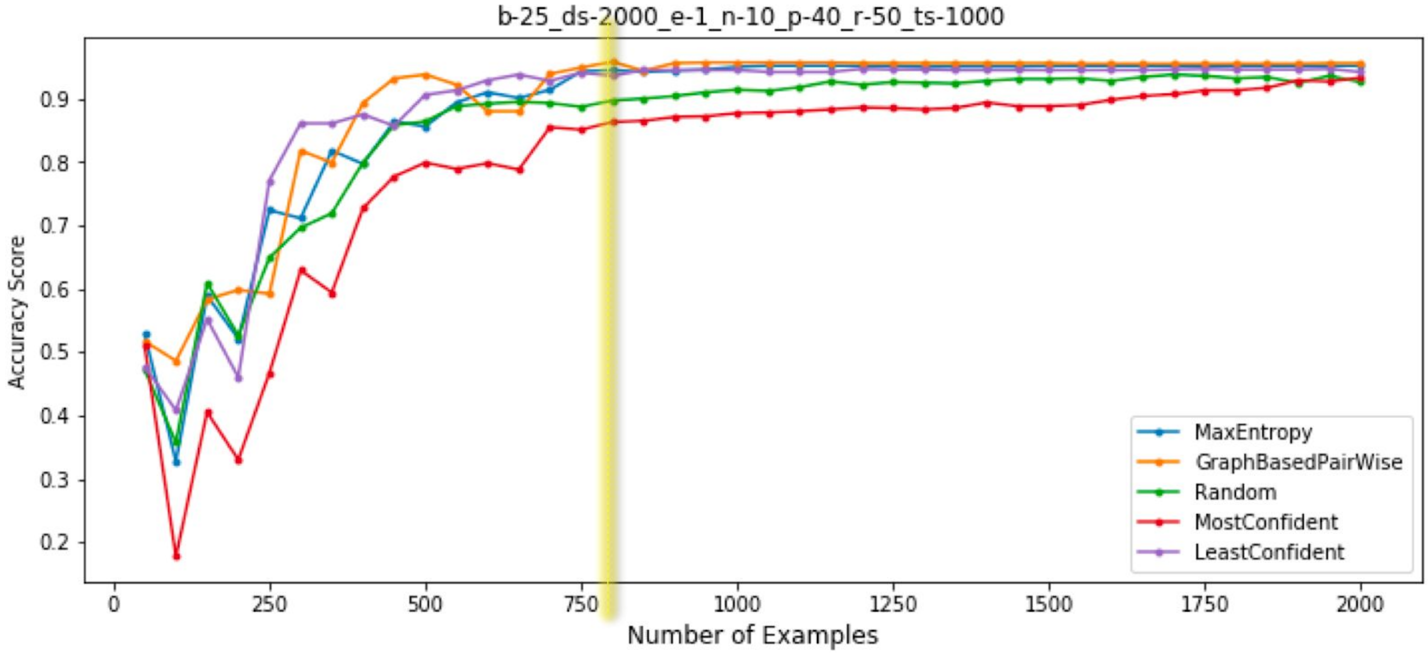

Figure 9: Learning Curves with Threshold at 800

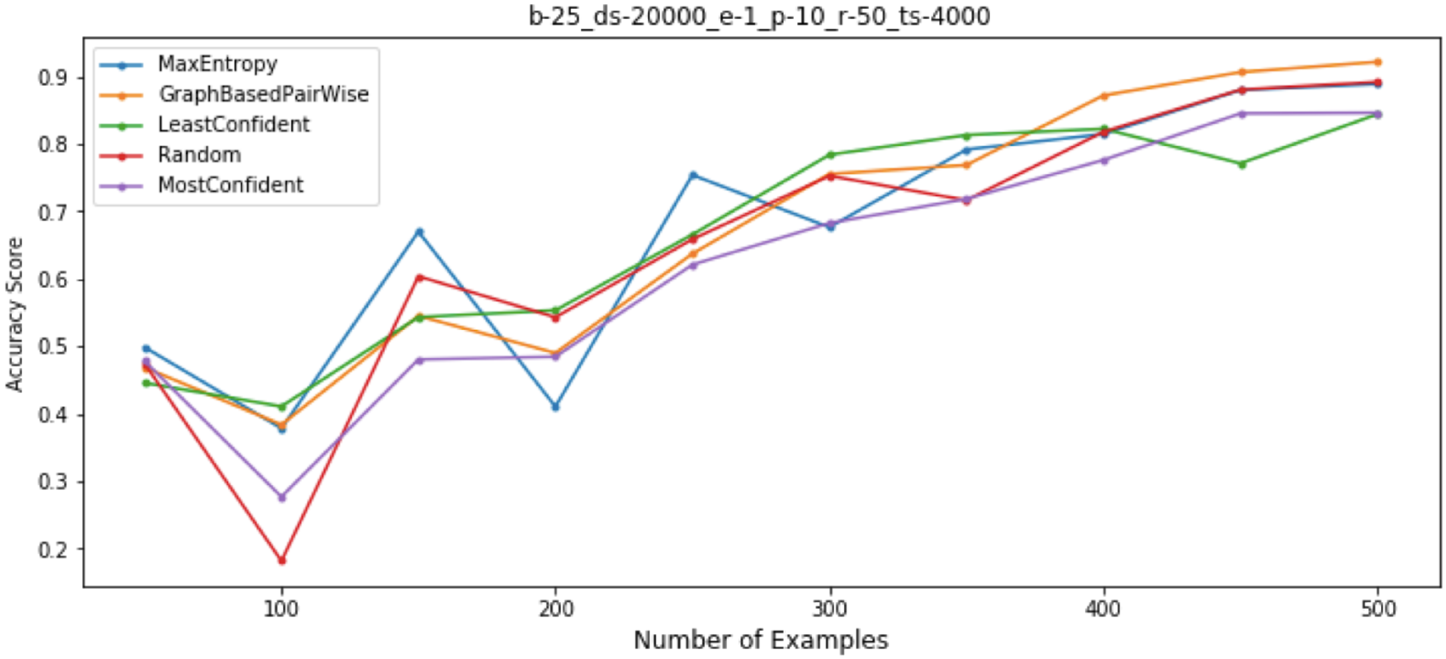

Figure 10: Learning Curves with 500 Labelled Examples

Data Efficiency

I look into the significance of querying few data points and demonstrate how the model will perform with even fewer instances.

In figure 11, there are three learning curves with 10,25,50 queries per step and the one with 50 queries had the highest accuracy score after the threshold point. However, It is noteworthy that at the learning curve with 10 queries per rating had the highest score at the beginning of the learning curve , which reinforces the concept of the informativeness of the examples being selected with few data points.

Computation time is affected by the number of queries selected, therefore, it is wise to understand the significance of choosing the optimal number in terms of computation time and accuracy score.

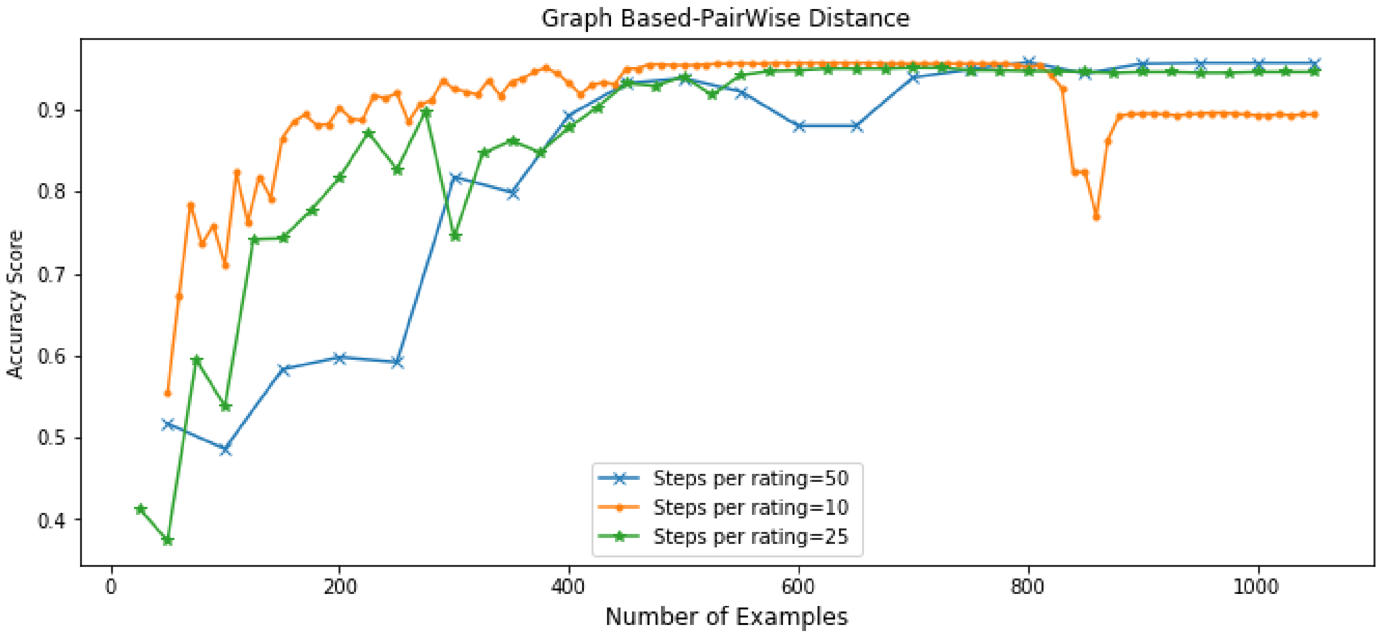

Figure 11: Learning Curve with Different Steps per Rating

Supervised Learning

Supervised learning in deep learning framework must be trained on a large amount of labelled data. The results in figure 12, is based on 2000 labelled examples. The learning curve with 2000 examples chosen randomly from the labelled set to is fluctuating which confirms that more labeled data should be fed into the model to build ro build robust classifier.

Additionally, If assuming that only 50 examples are labeled and train the model with those ones as we did in active learning we will get around 45% accuracy score. So we can conclude that if we let the learning algorithm choose the data it wants to learn from, it can perform better than supervised learning.

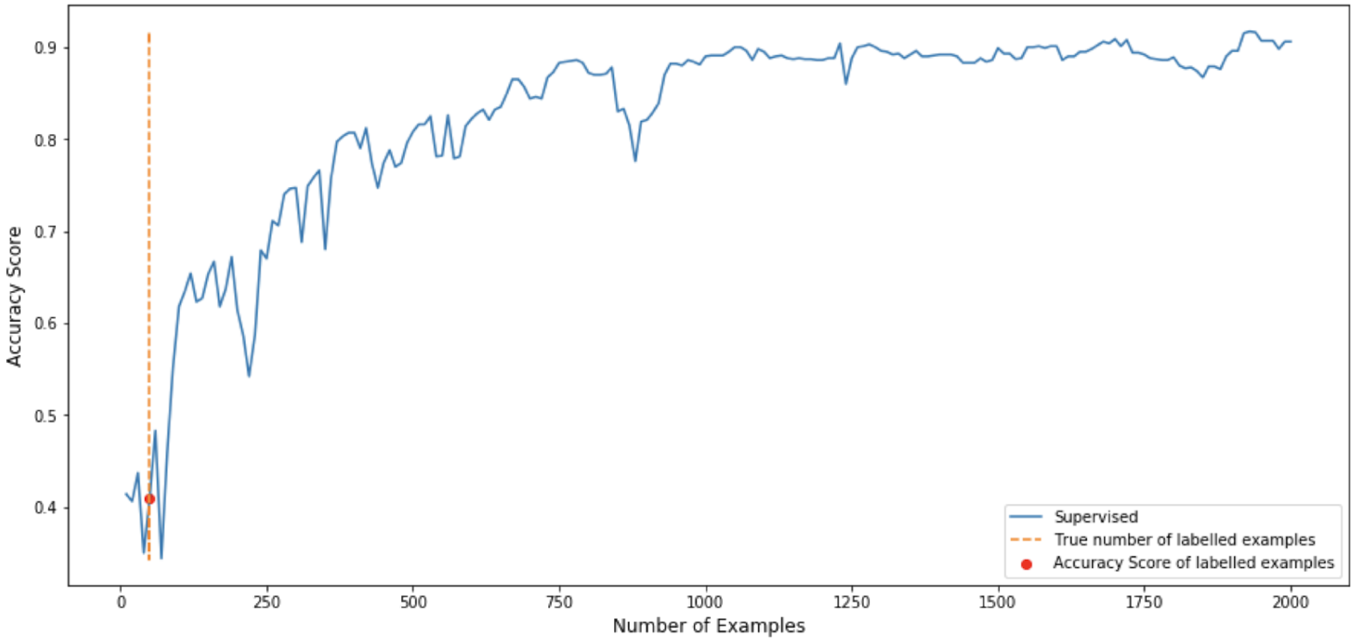

Figure 12: Supervised learning Curve

Conclusion

Graph-based approach achieved higher accuracy than other strategies at the beginning of all tests because it tends to sample more informative instances than other active learning methods. Active learning can overcome the labeling bottleneck and reduce the number of labeled data required to train a classifier.

In real-world application, different examples have different labeling costs. Economic factor is the main stopping criteria for labeling and data efficiency can be achieved in deep learning through active learning.