Active Learning Based Classification of Cell Organelles Images

Akash Ramachandran
Computational Biology
akashr1@andrew.cmu.edu

1 Theory

For this project, a neural network base learner was coupled with pool-based uncertainty sampling. The uncertainty measures commonly used in the literature [1], [2] for multi-class image classification are BVSB(Best-Versus Second Best) approach and the Entropy Approach.

1.0.1 BVSB

The data point that is selected in every iteration of active learning is the one which has the smallest difference between the top-2 class-probabilities calculated for it by the base learner. Greater this difference, smaller the uncertainty of the point.

1.0.2 Entropy

Entropy of a datapoint X is defined below:

$$H(X) = -\sum_{k=1}^{K} p_k \log p_k$$

Here K is the number of classes. Intuitively, a point which is given equal probability of belonging to each of the K classes has the highest uncertainty. Such a point will have highest $H(X)$. So in this project, in very active learning iteration, the point which has the highest entropy will be added to the pool of labeled instances L that will be used to update the hypothesis space (in this project, the weights and biases of the neural network).

2 Methods

2.1 Software

Google's Tensorflow [3] was used to implement the Neural Network.

2.2 Algorithm

The steps used in this project are listed below:
The algorithm is the same as the generic pool-based uncertainty sampling covered in class. Initially there are no points in the labeled set L. Every iteration, all data points are passed through the network and the probability values of each class for each example is obtained (using the softmax function in the neural net).

In the BVSB method, the point with the smallest difference between its top-2 softmax probabilities is added to L and the cumulative L set is used to update the weights of the network. In the entropy method, the point with the highest entropy is added to L and then used to update the weights of the network. The random learner selects a random point to add to L every iteration.

3 Results

3.0.1 Easy dataset

The neural network architecture that achieved highest performance on the test set is given below:

1. Learning Rate : 0.05
2. Dropout : 0.5
3. Number of Layers = 2
4. Number of hidden units in layer 1: 60
5. Number of hidden units in layer 2: 30
6. Weight Decay: 0.001
7. random seed = 2017

Table 1: Performance on the Easy Test Set

Learner	Test Accuracy Achieved(%)
BVSB	92.5926
Entropy	91.8919
Random	89.6897

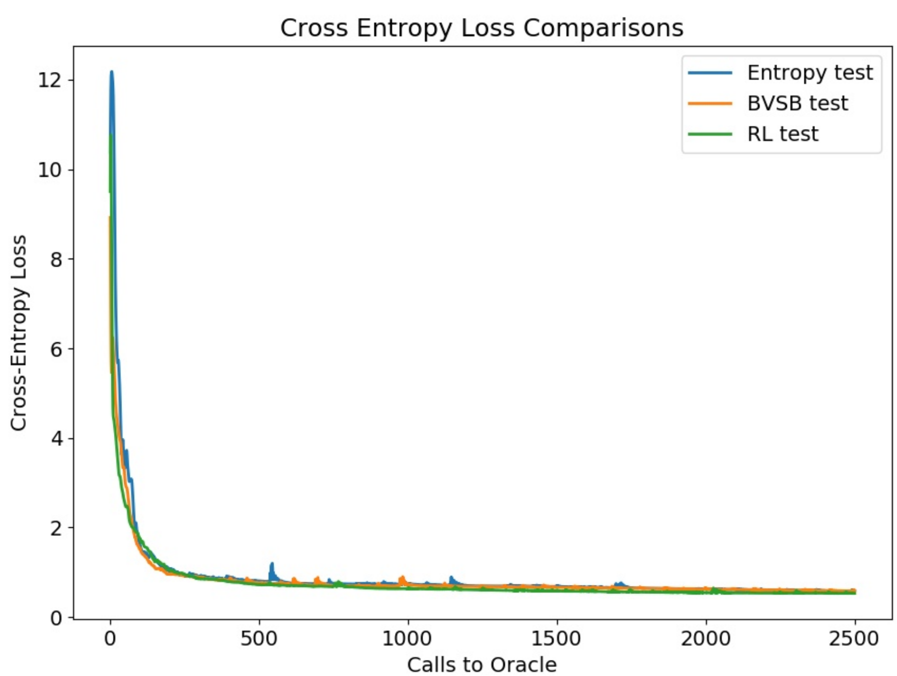

Figure 1: Medium Dataset - Cross Entropy Curves

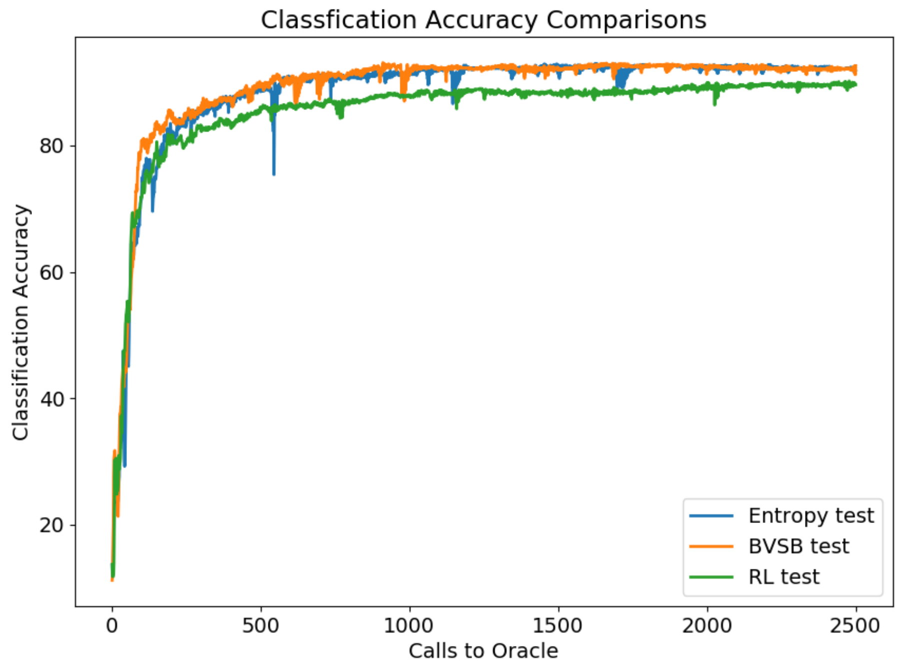

Figure 2: Medium Dataset - Classification Accuracy Curves

3.0.2 Medium dataset

The neural network architecture that achieved highest performance on the test set is given below:

1. Learning Rate : 0.05
2. Dropout : 0.5
3. Number of Layers = 2
4. Number of hidden units in layer 1: 80
5. Number of hidden units in layer 2: 40
6. Weight Decay: 0.001
7. random seed = 2017

Table 2: Performance on the Medium Test Set

Learner	Test Accuracy Achieved(%)
BVSB	85.8859
Entropy	85.7858
Random	82.5826

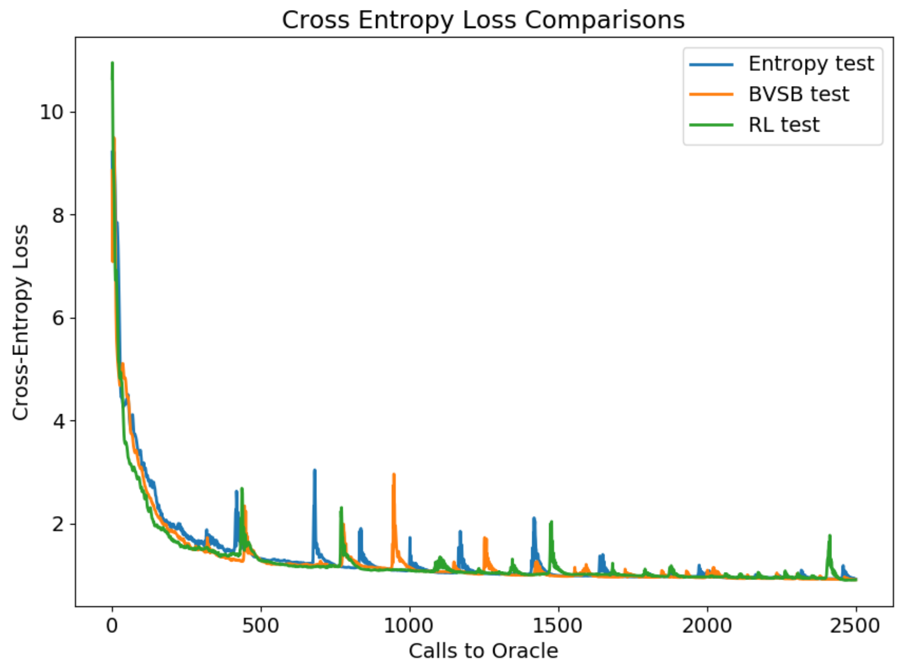

Figure 3: Medium Dataset - Loss Curves

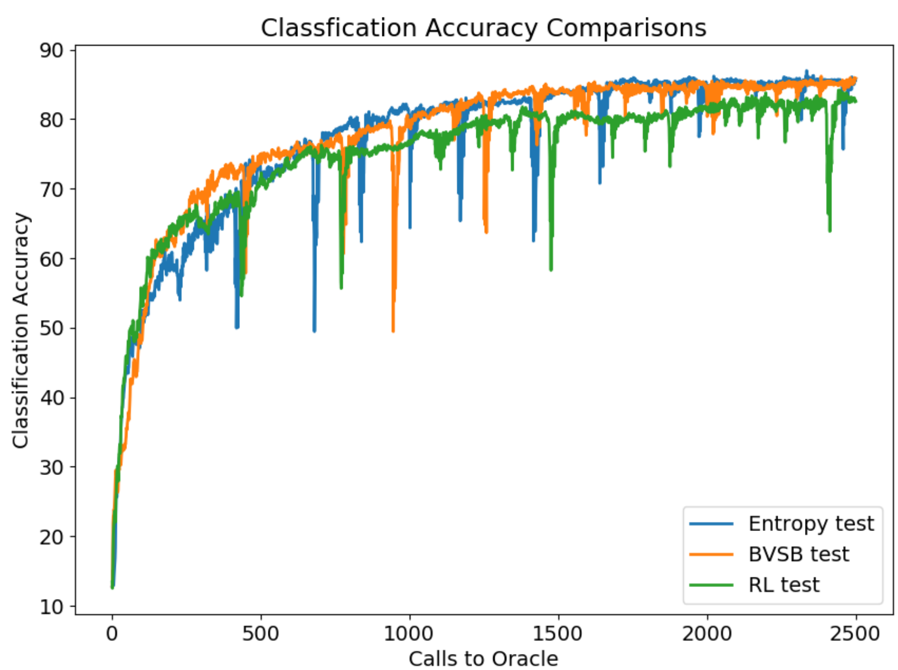

Figure 4: Medium Dataset - Classification Accuracy Curves

3.0.3 Difficult dataset

The neural network architecture that achieved highest performance on the test set is given below:

1. Learning Rate : 0.05
2. Dropout : 0.5
3. Number of Layers = 2
4. Number of hidden units in layer 1: 65
5. Number of hidden units in layer 2: 35
6. Weight Decay: 0.001
7. random seed = 2017

Table 3: Performance on the Difficult Test Set

Learner	Test Accuracy Achieved(%)
BVSB	79.98
Entropy	79.4795
Random	72.2723

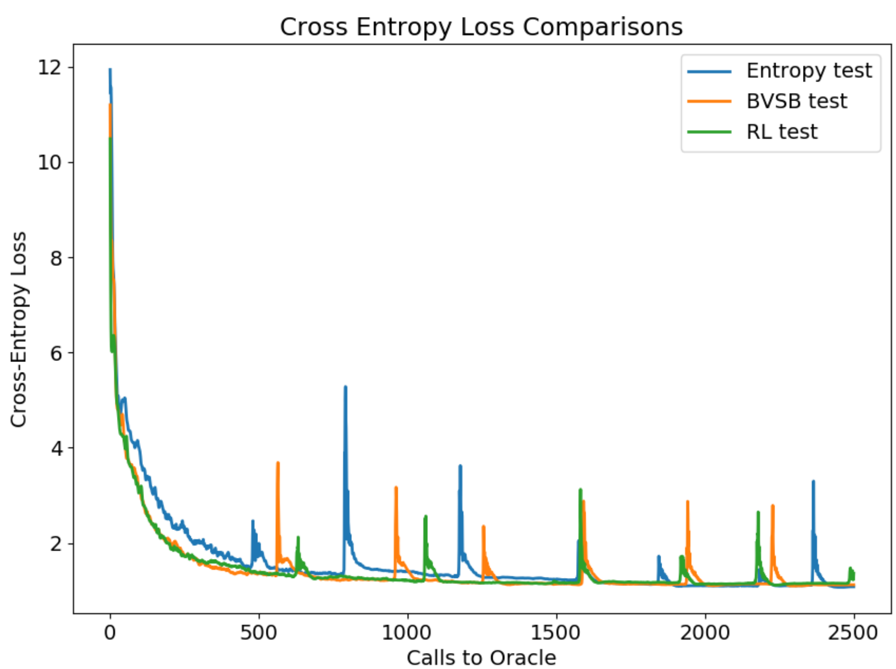

Figure 5: Difficult Dataset - Loss Curves

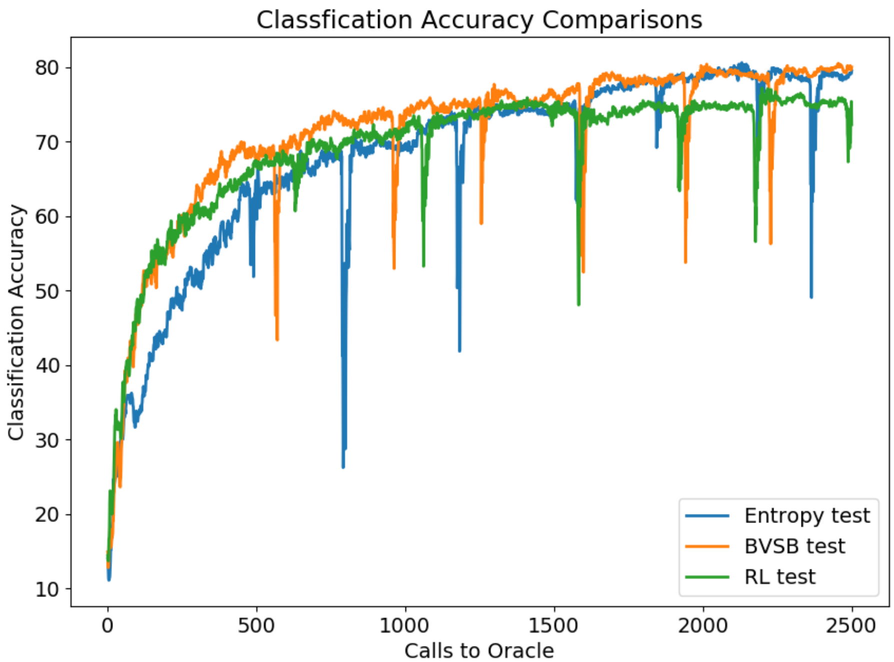

Figure 6: Difficut Dataset - Classification Accuracy Curves

4 Observations

1. In easy, medium and hard datasets, the active learning approaches of BVSB and entropy achieve better test accuracies than the random learner as expected.
2. For some reason, this pattern is not observed in the cross-entropy losses of the 3 methods.
3. BSVB performs better than the Entropy method on test accuracies.
4. As the difficulty increases, there are larger fluctuations in the accuracy plots. One reason for this could be because the data is so noisy / diverse that the network's hypothesis space has to change significantly after each oracle call.

References

[1] A. J. Joshi, F. Porikli and N. Papanikolopoulos, "Multi-class active learning for image classification," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 2372-2379. doi: 10.1109/CVPR.2009.5206627

[2] Y. Gu, Z. Jin and S. C. Chiu, "Active learning combining uncertainty and diversity for multi-class image classification," in IET Computer Vision, vol. 9, no. 3, pp. 400-407, 6 2015. doi: 10.1049/iet-cvi.2014.0140

[3] Abadi et.al, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.