Min Entropy Sampling Might Lead to Better Generalization in Deep Active Text Classification

Nimrah Shakeel

Abstract

We investigate the usefulness of entropy based sampling in deep active learning for text classification in a set of specially designed experiments.

Motivation: Uncertainty sampling sometimes performs worse than random sapling in active learning, with neural networks.

Inspiration: Lowell et al. report that there is no advantage in training a successor model with an actively acquired dataset.

Method: Train one classifier and use it to select training data points for another classifier.

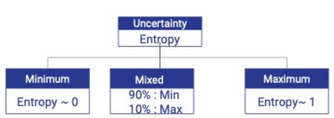
AL in Pool Based Batch Mode: Steps

- Train a Classifier trained on a small subset of data(randomly selected)
- Subset of unlabelled pool selected according to some strategy
  - Classifier predicted labels for uncertainty sampling
- Classifier trained on labelled data
- Repeat until criteria met
  - Budget exhausted, accuracy reached

Sampling Strategy: Uncertainty

- Uncertainty: select points about which classifier is most uncertain
- Simple Deep Uncertainty: Entropy, Least Confidence, Margin based
- Entropy of classifier prediction over all output classes
  - E.g. a movie review with 0.53 predicted positive sentiment, 0.47 negative ~ High Entropy
  - Another with 0.9 positive, 0.1 negative~ Low Entropy

Uncertainty: Strategies Used



Deep Learning with Few Labels

- Acquiring labels is expensive
- Aim: train a classifier on a few data points, get max accuracy
- Active learning: classifier selects points, asks for labels
- Pool based batch-mode AL:
- All points available as a pool (not arriving one by one)
- Labels acquired for a batch (not getting labels one by one)
- AL strategies sometimes work badly with neural networks

Experimental Setting: Difference from AL

- Models : CNN-1 and CNN-2 (similar CNNs)
- Data : D_train, D_val and D_test
- Train CNN-1 on D_train, validate on D_val
- Estimate uncertainty of CNN-1on D_test
- For each sampling strategy
  - Select a batch of k top points
  - Train one instance of CNN-2 with this batch
  - Note accuracy of CNN-2 on D_val
- Repeat for various values of k

Summary

- Our experimental setting is different from AL setting
  - Uncertainty estimates from a fully trained classifier(CNN-1)
  - Used to train another classifier(CNN-2)
- Results somewhat counterintuitive
  - Might shed light on the behaviour of AL strategies with deep models
  - Poor behaviour of max-entropy sampling reported in literature

Results on 3 Datasets (IMDB, Yelp, Amazon)

Conclusions

- Experiments on 3 binary text classification datasets(IMDB, Yelp, Amazon)
- Max-entropy : often worse than random
- Min and mixed entropy : often better than random
- Min-entropy samples seem to be helping representation learning

Table 2: Classification accuracy achieved by CNN-2 on D_{val} after training on D'_{test} against $|D'_{test}|$. CNN-2 uses word2vec embeddings.

$\lvert D'_{test}\rvert$	100	250	500	1000	2500	5000	10000	15000	20000	25000
IMDB										
MIN-ENT	49.44	54.16	**76.83**	76.92	82.77	**84.39**	85.2	85.82	86.31	86.76
MIXED-ENT	49.44	**63.05**	71.4	**79.98**	**83.06**	83.99	84.99	85.51	**86.95**	86.85
MAX-ENT	50.63	50.82	51.44	52.07	52.31	59.38	80.35	85.69	86.47	**86.86**
RANDOM	**50.96**	55.18	57.82	70.43	82.32	84.23	**85.42**	**85.83**	86.2	86.7
AMAZON										
MIN-ENT	52.53	55.39	**69.71**	**74.02**	78.37	**80.76**	**82.92**	83.4	84.81	**85.48**
MIXED-ENT	52.53	**60.34**	68.63	73.99	**78.54**	80.72	82.43	**83.99**	85.01	85.35
MAX-ENT	47.5	47.48	47.87	49.35	50.55	56.65	74.05	82.72	**85.29**	85.34
RANDOM	**54.56**	55.74	64.61	70.09	78.07	80.58	82.64	**83.99**	84.94	85.32
YELP										
MIN-ENT	46.73	56.99	66.23	77.52	82.14	84.54	86.85	88.28	89.38	90.56
MIXED-ENT	47.81	60.38	**75.3**	**81.8**	84.39	86.38	87.53	88.36	89.36	**90.57**
MAX-ENT	53.77	53.77	53.77	53.77	54.89	66.86	87.74	**90.34**	**90.45**	90.55
RANDOM	**56.76**	**61.64**	72.38	80.94	**85.94**	**87.46**	**88.9**	89.47	90.41	90.56

Table 3: Classification accuracy achieved by CNN-2 on D_{val} after training on D'_{test} against $|D'_{test}|$. CNN-2 uses Glove embeddings.

$\lvert D'_{test}\rvert$	250	500	1000	2500	5000	10000
IMDB						
MIN-ENT	**64.38**	**71.08**	**73.18**	**74.97**	**76.1**	**77.76**
MAX-ENT	51.2	51.29	51.34	52.42	53.65	65.18
RANDOM	54.88	61.84	65.26	69.45	74.2	77
AMAZON						
MIN-ENT	**64.41**	**69.03**	**70.01**	**72.66**	**74.05**	**76.02**
MAX-ENT	53.24	52.89	52.28	52.95	54.39	61.72
RANDOM	57.23	61.02	64.79	69.23	72.32	75.47
YELP						
MIN-ENT	57.88	65.33	69.04	74.74	**78.33**	**80.59**
MAX-ENT	52.65	51.35	53.11	54.14	57.08	73.62
RANDOM	**59.67**	**66.59**	**70.79**	**75.01**	76.85	79.59