Recursive Ensembles Active Learning for Text Classification

Anonymous COLING submission

Abstract

Supervised classification of texts relies on availability of reliable class labels for the training data. However, the process of collecting data labels can be complex and costly. Therefore, more common task is semi-supervised learning where the number of texts is often much higher than the number of labels, but the labels can be obtained by querying an annotator. Active learning is a process of selection of unlabeled data records for which knowledge of the label would bring the highest discriminability of the dataset. Various methods for active learning have been proposed in many different fields that use supervised learning models. In this project, we study suitability of Bayesian approaches and propose a recursive modification of the deep ensemble active learning approach. We compare existing state-of-the-art active learning classifiers with the proposed approach and study uncertainty calibration for both existing and new approaches.

1 Introduction

Active learning strategy is a techniuqe of semi-supervised machine learning that actively searches which unlabeled sample should be given to annotator and extend the training set. If the active search is well optimized, the number of queries for corrcet classification can be greatly reduced. Our goal is to compare performace of Bayesian active learning methods on the text classification problem. People have already been solving the active learning problem for anomaly detection (Das et al., 2018), image processing (Gal et al., 2017) (Sener and Savarese, 2017), named entity recognition (Shen et al., 2017), (Lowell et al., 2019), (Burkhardt et al., 2018), and others.

A typical approach to labeling huge amount of unlabeled data, is often not optimal. People are randomly choosing unlabeled text data. These data are annotated by the subject matter experts, and used for training and testing of the models. If the model performance is weak after the training, more text documents are selected and annotated. This approach is costly because nobody knows how many text documents must be selected in order to have good model scores. Active learning strategy has potential to greatly reduce this effort. While it was introduced almost two decades ago, recent improvements in deep learning motivate our attempt to revisit the topic. For example, SVM-based active learning approaches for text classification date back to 2001 (Tong and Koller, 2001), where superiority of active learning over random sampling are demonstrated. Since deep recurrent and convolutional neural networks achieve better classification result, Bayesian active learning methods for deep network gained popularity especially in image classification (Gal et al., 2017; Lowell et al., 2019). The Bayesian approach is concerned with querying data where the classified is most uncertain. While acquisition functions provide similar results, representation of uncertainty is often more important. The most popular approach using Dropout MC (Gal et al., 2017) has been tested on text classification (An et al., 2018), however other techniques such as Langevin dynamics (Welling and Teh, 2011) and ensembles (Lakshminarayanan et al., 2017) are available. Deep ensembles often achieve better performance (Snoek et al., 2019) but require higher computational cost since they train an ensemble of networks after each extension of the data set.

This problem has been recently addressed in (Ulrych and Smidl, 2020), where it was proposed that the ensemble should not be trained from scratch but initialized randomly around the position of ensemble network from the previous iteration. In this contribution, we test this approach and compare it with the dropout MC and Langevin dynamics representations.

2 Methods

Transformer models such as BERT (Devlin et al., 2018) or other different modifications showed their strength in context understanding. However, we assume that Fast Text (Mikolov et al., 2018) text

encoding is efficient enough for our text classification purposes. Representation of a text document $\mathbf{x}$ is calculated as a mean value from all words embeddings in the text

$$\mathbf{x}_i = \frac{1}{|\mathcal{D}_i|} \sum_{j \in \mathcal{D}_i} f_{\text{Fast text}}(\mathbf{c}^{(j)}),$$

where $\mathcal{D}_i$ is the set of indices of all words in the i-th document in the common vocabulary, $\mathbf{c}^{(j)}$ is j-th one hot encoded word vector and $f_{\text{Fast text}}$ is a function that creates Fast Text embeddings with respect to given one-hot encoded word.

For supervised classification, each document $\boldsymbol{x}_i$ should have a label y_i. We are concerned with binary classification for simplicity, however, extension to multiclass is straightforward. We assume that an initial set of few labels is available, splitting the set of text documents to the labeled, $X^{(l)}$, and unlabeled parts. We design a selection procedure for the annotator, by maximizing the entropy of the predictive probability density function

$$i^{(l)} = \arg\max_{j \in \mathcal{J}} \mathsf{E}\left(\log(p(y_j|\boldsymbol{x}_1 \dots \boldsymbol{x}_l, \boldsymbol{x}_j))\right),$$

where $\mathcal{J}$ is the set of indexes of unlabeled texts, E is the expectation operator over the uncertainty of model parameters. When the selected text is annotated, the text is added with its label to the labeled data set $X^{(l+1)} = [X^{(l)}, x_{i^{(l)}}]$. And the procedure is repeated.

Key component is the method of representation of uncertainty. We will compare the following methods: i) SGLD: Stochastic Gradient with Langevin dynamics (Welling and Teh, 2011), which adds additional noise to the gradient in stochastic gradient descent, ii) Dropout MC: samples binary mask disabling selected paths through the network (Gal et al., 2017). and iii) Deep ensembles: consist of N networks trained in parallel from different initial conditions (Lakshminarayanan et al., 2017). This approach is the current state-of-the-art in active learning (Beluch et al., 2018).

While many of these has been tested in active learning, the authors always assumed that after each step of active learning, the network training starts from the initial conditions. This is clearly suboptimal, since the information from previous training is lost. A simple solution was presented in (Ulrych and Smidl, 2020), where it was argued that estimated results from the previous step can be used as centroids around which the new initial point is sampled. Since this is a form of warm start, we also test warm-start strategies for Dropout. In summary, we test the following algorithms:

DEnFI: a deep ensemble method with warm start using weights of ensemble members as initial conditions for new ensemble. The weights are perturbed by Gaussian nose of variance q which is a hyperparameter. The ensemble is trained run 2000 epochs with additional 700 epochs after each extension of the learning data set.

Dropout MC: in two versions, pure hot-start and hot-start with weights perturbed by additive noise of variance q. The network is trained run 3000 epochs, dropout rate is 0.5, and number of epochs per sample is 50.

SGLD: variance of the noise added to the gradient descent is $\sqrt{\epsilon}$ where ϵ is the learning rate with initial value of 0.01 and which is calculated in $n+1$ iteration as $\epsilon_{n+1} = \frac{\epsilon_n}{n-3000} + 0.05$. The noise is added to a gradient only after 3000 of initial training epochs.

3 Experiments

The active learning simulation compares different algorithms with respect to two strategies: i) random, and ii) entropy based selection of text documents. In this paper we decided to use positive negative tweets from Tweets Dataset (Go A., 2009) and 5 pairs of categories form News Category Dataset (Misra, 2018). Names of the categories are shown in table 1 and in figure 1. The categories were chosen with respect to different classification complexity. We randomly choose initial training set that has 10 samples. Described 10 random samples are chosen from 1000 text documents (500 text documents per category). We reduce the size of all above mentioned datasets to 1000 documents. Each dataset is split on testing and training data.

Next we initialize two runs. First run is random selection of the text documents and second run is entropy based selection of the documents. Both runs start with the same training dataset, and then they chose additional training documents based on their strategies. We consider continuous new data selection from 1000 documents dataset and imitating annotators labeling process. All in all, we repeat text samples selection 200 times ($U = 200$). We select 10 random samples in the beginning, train our model, and make

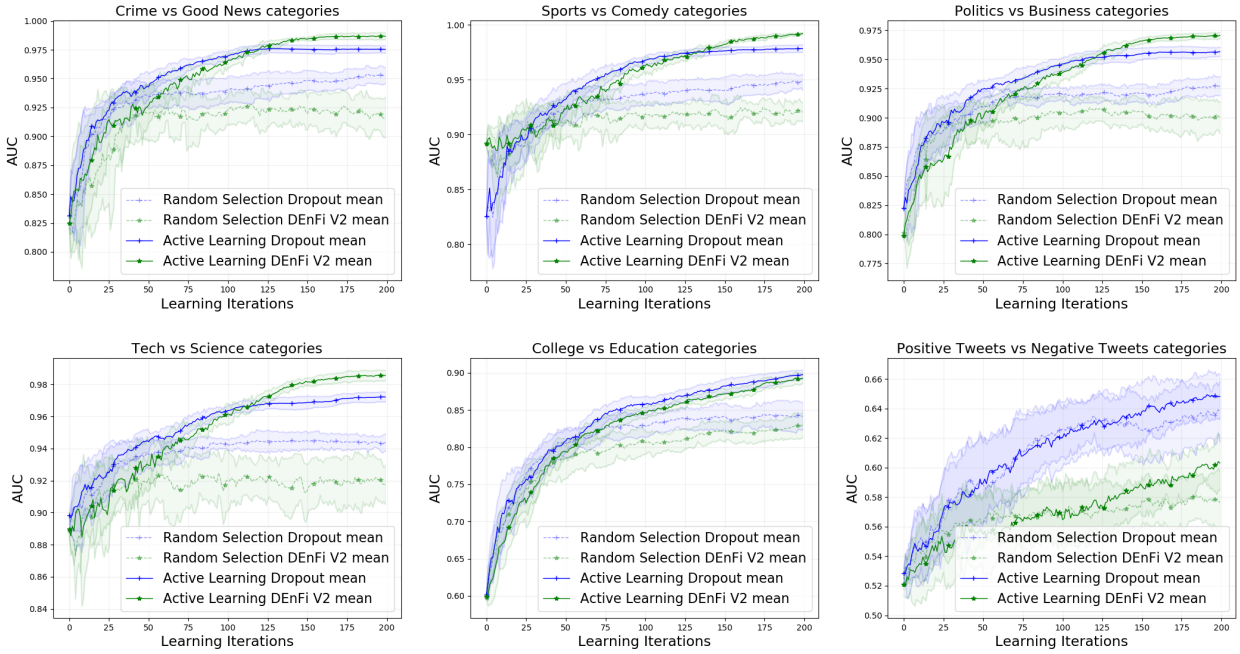

Figure 1: AUC mean evolution with respect to learning iterations for DEnFi, Dropout cold start algorithms and six pairs of categories. The uncertainty bounds are illustrated as one standard deviation from the mean value. Both DEnFi and Dropout were initially trained on 10 labeled text documents before sequential learning strategies were initialized

a prediction on the complement to the training dataset (990 text documents). As a further step, we select 1 new sample from the set on which the prediction was done. New text sample for labeling is chosen with respect to acquisition function or random choice. Before we extend our training dataset with a new labeled sample, we calculate the area under ROC curve (AUC) metrics on the complement to the dataset (990 text documents). This process is repeated 200 times. Thus, by the end of the simulation, our training set will have 210 text documents and testing set (complement to a training set) will have 790 data samples. In order to make our results statistically valid, we repeat the described simulation loop 10 times.

3.1 Neural Network Approaches Comparison

In table 1 is shown comparison of different neural networks based active learning algorithms with respect to AUC mean. The results in table 1 are shown for 200 active learning iterations and entropy acquisition function. We belief that DEnFi and Dropout cold start algorithms show the best performance with respect to shown results.

In figure 1 is illustrated whole evolution of AUC mean over 10 runs with uncertainty bounds for six pairs of categories. It is seen that in four out of six plots DEnFi shows better active learning results. It is also seen that DenFi mimic behavior of Dropout curve for College vs Education categories. In comparison to the categories where DEnFi showed its superiority over Dropout we can see that AUC values were higher than 95%, whereas maximal AUC value for College vs Education is around 90%. It can be said that DEnFi did not have enough active learning iterations in order to explore the dataset better. It is seen that in the final steps of a simulation, DEnFi curve is converging to Dropout curve. The behavior that was seen for College vs Education categories can be also observed for Tweets Dataset. The AUC scores are very low and it is seen that both DEnFi and Dropout are copying random selection strategy evolution. However, we can observe that in the right part of the plot, DEnFi active learning strategy starts to show better scores than random selection strategy. In this case we can also conclude that DEnFi did not have enough training samples for overcoming Dropout.

3.2 Hot Start Noise Calibration

Both DEnFi and Dropout with hot start are quite sensitive with respect to added noise. DEnFi and Dropout hot start mean AUC results with respect to different gaussian noise variance configuration for

Entropy based acquisition function	Crime vs Good News	Sports vs Comedy	Politics vs Business	Tech vs Science	Education vs College	Positive vs Negative Tweets
SGLD	**0.989**	0.968	0.944	0.984	0.881	0.621
DEnFi	0.987	**0.992***	**0.971***	**0.986**	**0.893**	0.603
Dropout cold start	0.975	0.978z	0.957	0.972	**0.898***	**0.648**
Dropout hot start	0.978	0.979	0.954	0.973	0.877	**0.657***
Dropout hot start w noise	0.978	0.951	0.944	**0.989***	0.824	0.561

$(*)$ maximal value per column

Table 1: Binary classification mean AUC results over 10 runs for 6 different algorithms with respect to 200 active learning iterations and six different pairs of categories. Bold values represent intersection of a mean value with respect to one standard deviation interval from a maximal value $(*)$

Noise variance	**Active learning iterations** 0	50	100	150	200
0.1	0.867	**0.945***	**0.968***	0.974	0.976
0.2	**0.893***	0.932	0.964	0.976	0.978
0.3	0.890	0.930	0.961	**0.982***	0.986
0.4	0.886	0.909	0.948	0.976	**0.990***
0.6	0.846	0.874	0.921	0.952	0.979
1	0.777	0.805	0.871	0.906	0.941

$(*)$ maximal value per column

(a) DEnFi

Noise variance	**Active learning iterations** 0	50	100	150	200
0.1	**0.902***	**0.936**	**0.966***	0.972	0.976
0.2	**0.901**	**0.938***	**0.966***	**0.981***	0.983
0.3	0.899	0.920	0.956	**0.980**	**0.989***
0.4	0.900	0.917	0.955	0.976	**0.988**
0.6	0.898	0.894	0.948	0.972	**0.986**
1	0.866	0.859	0.914	0.941	0.970

$(*)$ maximal value per column

(b) Dropout hot start with noise

Table 2: DEnFi and Dropout hot start with noise binary classification mean AUC results over 10 runs for Tech vs Science categories with respect to different noise variance configuration and number of active learning iterations. Bold values represent intersection of a mean value with respect to one standard deviation interval from a maximal value $(*)$

Tech vs Science categories are shown in table 2. It is clearly seen that for little amount of training data strategy with less noise works better. However, when the amount of training data is increasing, the strategy with higher amount of noise is more efficient. Described trend is observed both for DEnFi and Dropout algorithms and can be seen as a linear curve formed from bold values in table 2. As a result, it can be assumed that adaptive noise modification can lead to the significantly better results.

4 Conclusion

In this work it was illustrated and verified that an ensembles exploration power can overcome a dropout based active learning algorithm if a querying process lasts long enough. Described behavior was perfectly seen both in the plots for the News Category Dataset where the classification scores were high and for the Twitter Dataset where the classification scores were low. Another important aspect of the work are hot start methods and a noise calibration. The DEnFi method showed great results by using the knowledge from previous iterations. The dropout hot start method has also showed good results. However, the hot start results were not good enough in comparison to the cold start method. We are concerned, that the trend of increasing noise and better classification for higher amount of training documents may let us to reach higher scores with adaptive noise addition. Combination of the facts mentioned above can be a strong basis for the future research.

References

Bang An, Wenjun Wu, and Huimin Han. 2018. Deep active learning for text classification. In *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, pages 1–6.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. 2018. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377.

Sophie Burkhardt, Julia Siekiera, and Stefan Kramer. 2018. Semisupervised bayesian active learning for text classification. In *Bayesian Deep Learning Workshop at NeurIPS*.

Shubhomoy Das, Md Rakibul Islam, Nitthilan Kannappan Jayakodi, and Janardhan Rao Doppa. 2018. Active anomaly detection via ensembles. *arXiv preprint arXiv:1809.06477*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org.

Huang L. Go A., Bhayani R. 2009. Twitter sentiment classification using distant supervision.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413.

David Lowell, Zachary C Lipton, and Byron C Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 21–30.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Rishabh Misra. 2018. News category dataset, 06.

Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.

Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980.

Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

Lukas Ulrych and Vaclav Smidl. 2020. Deep ensemble filter for active learning. Technical Report 2383, Institute of Information Theory and Automation.

Max Welling and Yee W Teh. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688.