Active Learning for Text Classification using Deep Ensemble Filtering

Anonymous COLING submission

Abstract

Supervised classification of texts relies on the availability of reliable class labels for the training data. However, the process of collecting data labels can be complex and costly. A common procedure is to add labels sequentially by querying an annotator until reaching satisfactory performance. Active learning is a process of selection of unlabeled data records for which the knowledge of the label would bring the highest discriminability of the dataset. Bayesian methods are frequently used due to their ability to represent the uncertainty of the classification procedure. In this project, we apply deep ensemble filter, i.e. warm-start modification of the deep ensemble representation, for the task of active text classification. We show that while the conventional dropout Monte Carlo approach provides good results for a low number of requests, the warm-start ensembles improve with a growing number of requests, outperforming the dropout in the long run.

1 Introduction

The development of a text classifier on a new problem requires the availability of the training data and their labels. Labeling involves human annotators and a common practice is to label as many text documents as possible, train a classifier and search for new data and labels if the performance is unsatisfactory. Random choice of the documents for the data set extension can be costly because the new documents may not bring new information for the classification. Active learning strategy aims to select among available unlabeled documents those that the classifier is most uncertain about and query an annotator for their labels. Therefore, it has the potential to greatly reduce the effort needed for the development of a new system. 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 the superiority of active learning over random sampling is demonstrated. Since deep recurrent and convolutional neural networks achieve better classification results, 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 label for such data for which the classifier predicts the greatest uncertainty. The uncertainty is quantified using the so-called acquisition function, such as predictive variance or predictive entropy. While different acquisition functions often provide similar results, different representations of predictive distribution yield much more diverse results. The most popular approach using Dropout MC (Gal et al., 2017) has been tested on text classification (An et al., 2018) and named entity recognition (Shen et al., 2017; Lowell et al., 2019), however other techniques such as Langevin dynamics (Welling and Teh, 2011) and deep ensembles (Lakshminarayanan et al., 2017) are available. Deep ensembles often achieve better performance (Beluch et al., 2018; Snoek et al., 2019) but require higher computational cost since they train an ensemble of networks after each extension of the data set. One potential solution of this problem has been recently proposed in (Ulrych and Smidl, 2020), where the ensemble is not trained from a fresh random initialization after each query but initialized randomly around the position of the ensembles from the previous iteration. In this contribution, we test this approach and compare it with the dropout MC and Langevin dynamics representations. We also provide sensitivity study for the choice of the hyperparameters.

Active learning for fake news detection has been considered in (Bhattacharjee et al., 2017) using uncertainty based on probability of classification. It was later extended to context aware approach (Bhattacharjee et al., 2019). An entropy based approach has been presented in (Hasan et al., 2020) using an ensemble of three different kinds of network.

2 Methods

Throughout the paper, we will use four different encoding algorithms such as Fast Text (Mikolov et al., 2018), LASER , BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). BERT and RoBERTa are transformer models based on multi-head attention layers (Vaswani et al., 2017). Transformers models provide state of the art results in context understanding and it is interesting to compare behavior of active learning algorithms with respect to different encoding techniques. Representation of the i-th text document $\mathbf{x}_i$ is calculated as the mean value from sentence embeddings of all sentences in the text

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

where $\mathcal{D}_i$ is the set of vectors where each each vector represents a sentence as a set of word indices in the i-th document in the common vocabulary, $|\mathcal{D}_i|$ is a cardinality of $\mathcal{D}_i$, $\mathbf{C}^{(j)}$ is a matrix of j-th sentence where words are encoded with one hot or byte pair encoding (Shibata et al., 1999) technique and $f_{\text{Sentencet embed}}$ is a function that creates sentence embeddings with respect to the given one-hot or byte pair encoded word. Fast Text encoding is made with respect to a one-hot encoded words. LASER and transformer based models take as an input byte per encoded words. For every algorithm (Fast Text, Laser, BERT, RoBERTa), sentence embeddings are calculated as a mean value through embedded words.

For supervised classification, each document $\boldsymbol{x}_i$ should have a label y_i. We are concerned with binary classification for simplicity, however, an extension to multiclass is straightforward. We assume that for the full corpus of documents text documents $X = [\boldsymbol{x}_1, \dots \boldsymbol{x}_n]$, only an initial set of $l_0 \ll n$ labels is available, $Y^{(0)} = [y_1, \dots y_{l_0}]$, splitting the full set X to the labeled, $X^{(0)} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_{l_0}]$, and unlabeled parts, $X \backslash X^{(0)}$. Active learning is defined as a sequential extension of the training data set. In each iteration, $l = 1, \dots L$, the algorithm computes entropy of predictive probability distribution for each document in the unlabeled dataset and selects the index of the document with the highest entropy (entropy acquisition function), formally:

$$k_l = \arg\max_{k \in \mathcal{K}} \mathsf{E}_{p(\theta|X^{(l-1)})} \left(-\log(p(y_k|\theta, \boldsymbol{x}_1 \dots \boldsymbol{x}_{l_0+l}, \boldsymbol{x}_k))\right) \tag{1}$$

where $\mathcal{K}$ is the set of indexes of all unlabeled documents, E is the expectation operator over the posterior probability of the classifier parameters θ trained on all labeled data $p(\theta|X^{(l-1)}, Y^{(l-1)})$. When the selected text is annotated, the text is added with its label to the labeled data set $X^{(l)} = [X^{(l-1)}, \boldsymbol{x}_{k_l}], Y^{(l)} = [Y^{(l-1)}, y_{k_l}]$. The procedure is repeated L times.

The key component of the method is a representation of the posterior distribution of the parameter θ. Due to the complexity of the neural networks it is always represented by samples, with a different method of their generation. 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 have 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 a warm-start, we also test warm-start strategies for Dropout. The methods for representation of parametric uncertainty are:

DEnFI: a deep ensemble method with 10 neural networks in the ensembles and warm-start using weights of the ensemble members as initial conditions for the new ensemble. The weights are perturbed by a Gaussian noise of variance q which is a hyperparameter. The ensemble is trained to run 2000 epochs on the initial data with additional 700 epochs after each extension of the learning data set.

Dropout MC: in three versions: i) cold-start with 3000 epochs after each request, ii) hot-start, with 50 epochs, and iii) warm-start with weights from the previous iteration perturbed by an additive noise of variance q with 700 epochs. Dropout rate is 0.5.

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. Then we draw 50 samples with 100 epochs between consecutive samples to avoid correlation.

Noise variance	AUC after # of iterations 50	100	150	200
0.1	**0.945***	**0.968***	0.974	0.976
0.2	0.932	0.964	0.976	0.978
0.3	0.930	0.961	**0.982***	0.986
0.4	0.909	0.948	0.976	**0.990***
0.6	0.874	0.921	0.952	0.979
1	0.805	0.871	0.906	0.941

(a) **Fast Text** DEnFi

Noise variance	AUC after # of iterations 50	100	150	200
0.1	**0.936**	**0.966***	0.972	0.976
0.2	**0.938***	**0.966***	**0.981***	0.983
0.3	0.920	0.956	**0.980**	**0.989***
0.4	0.917	0.955	0.976	**0.988**
0.6	0.894	0.948	0.972	**0.986**
1	0.859	0.914	0.941	0.970

(b) **Fast Text** Dropout warm-start

Noise variance	AUC after # of iterations 50	100	150	200
0.1	**0.935**	**0.970***	**0.982**	0.988
0.2	**0.940***	0.954	**0.983***	**0.990***
0.3	0.903	0.930	0.967	0.987
0.4	0.883	0.911	0.945	0.976
0.6	0.799	0.853	0.885	0.945
1	0.661	0.750	0.818	0.875

(c) **RoBERTa** DEnFi

Noise variance	AUC after # of iterations 50	100	150	200
0.1	**0.930***	**0.970**	0.982	0.986
0.2	**0.923**	**0.971***	**0.986***	**0.990***
0.3	0.917	0.959	0.982	**0.990***
0.4	0.878	0.949	0.974	**0.990***
0.6	0.822	0.908	0.952	0.980
1	0.643	0.741	0.862	0.921

(d) **RoBERTa** Dropout warm-start

Table 1: **Fast Text** and **RoBERTa** encoding based AUC of text classification after selected number of requests of the active learning using DEnFi and Dropout warm-start for various selection of the perturbation noise q. Average over 10 runs on the Tech vs Science categories. The best result is denoted by a star, results within one standard deviation of the winner are displayed in bold font.

3 Experiments

The methods were compared on the positive/negative tweets from the Tweets Dataset (Go et al., 2009), 5 pairs of categories from the News Category Dataset (Misra, 2018) and two types of the news datasets . The names of the tested categories are shown in table 2 , figure 1 and in figure . The categories were chosen to represent different classification complexity and are displayed in the order of increasing complexity from the left to the right side. We have randomly chosen the initial training set that has $l_0 = 10$ samples from 1000 text documents (500 text documents per category), which were the initial 1000 documents of the datasets. Documents from the News Category dataset were downloaded using links provided in the dataset.

The active learning strategy is initialized from a dataset of 10 samples. For each strategy $L = 200$ requests are simulated. The element selected by active learning is accepted with probability $\epsilon = \frac{\exp(l-40)}{\exp(l-40)+1}$, i.e. the ϵ-greedy approach (Watkins, 1989), otherwise a random document is selected for labeling. After each request, the classification accuracy is evaluated on the remaining part of the selected dataset (i.e. on the 990 text documents in the first evaluation) using the area under the ROC curve (AUC) metrics (Fawcett, 2006). In order to make the results statistically valid, we repeat the described simulation loop 10 times.

3.1 Hyperparameter tuning

The classification network was designed as feed-forward NN with one dense layer of 100 neurons with sigmoid activation functions, and softmax output layer. Hyperparameter tuning was performed by a grid search for both Fast Text and RoBERTa text encoding techniques. Since the main focus of the paper is on the effect of the warm-start strategy, the results of the effect of the noise variance q for DEnFi and dropout MC is displayed in Table 1 for active learning strategy on the Tech vs Science task.

Note that the variance of the perturbation noise of the best result is increasing with the number of requests. We conjecture that the variance has the role of a selection of the exploration/exploitation tradeoff. Low variance favors exploitation and improves quickly, higher variance implies less accurate guesses in the initial iterations but better performance in the long run. Since calibration of the variance for all methods and all datasets would be too computationally expensive, we run all remaining experiments with $q = 0.3$. However, tuning of this hyperparameter for an application scenario or its adaptive strategy offers clearly a potential for further improvement.

3.2 Influence of uncertainty representation

A comparison of AUC of the active learning strategy after 200 requests for all tested algorithms and Fast Text encoding is reported in table 2. While all methods achieved best results on some datasets, the most consistent

method	Crime/ Good News	Sports/ Comedy	Politics/ Business	Tech / Science	Education/ College	Pos./Neg. Tweets
SGLD	**0.989**	0.968	0.944	0.984	0.881	0.621
DEnFi, $q = 0.3$	0.987	**0.992***	**0.971***	**0.986**	**0.893**	0.603
Dropout cold-start	0.975	0.978	0.957	0.972	**0.898***	**0.648**
Dropout hot-start	0.978	0.979	0.954	0.973	0.877	**0.657***
Dropout warm-start, $q = 0.3$	0.978	0.951	0.944	**0.989***	0.824	0.561
DEnFi 1 Ensemble, $q = 0.0$	0.953	0.939	0.901	0.938	0.800	0.609

Table 2: Average AUC over 10 runs for five different algorithms after 200 iterations of active learning and six different datasets. The best result is denoted by a star, results within one standard deviation of the winner are displayed in bold font.

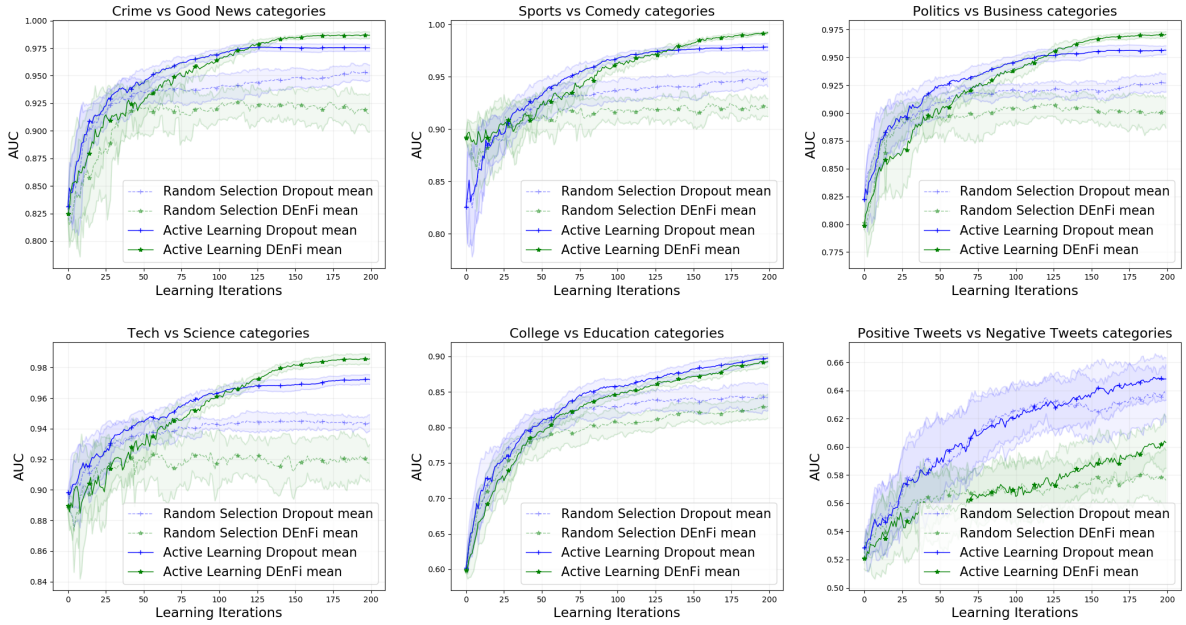

Figure 1: AUC mean evolution with respect to learning iterations for DEnFi, Dropout cold-start algorithms and six pairs of categories. The uncertainty bounds are illustrated as one standard deviation from the mean value. Both DEnFi and Dropout were initially trained on 10 labeled text documents before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

results were provided by the DEnFi and Dropout methods. Detailed analysis of the DEnFi and Dropout strategies is provided in figure 1 for all tested datasets. For better insight, we also display the performance of the passive learning strategy where training data are extended using randomly sampled documents. Note that for passive learning strategy, dropout outperforms consistently DEnFi on all tasks. We conjecture that this is due to the robustness of the dropout regularization. However, the power of DEnFi becomes apparent with the increasing number of requests. It is improving slower than dropout at the beginning, but improves faster, thus outperforming dropout in the long run. We conjecture that this is due to better exploration capability of the DEnFi while dropout excels at exploitation. The speed of improvement depends on the complexity of the learning task. For simpler tasks (such as crime vs. Good News), AUC over 0.98 is achieved quickly. However, for more complex tasks, such as Positive vs Negative Tweets, the number of data needed for improvement is much higher. The active learning is on par with the passive strategy up to 125 requests and even after 200 requests, the AUC is below 0.7 indicating poor performance. Note that the active learning strategy of DEnFi starts improving over the passive one sooner than dropout with a sharper slope which indicates a high probability of obtaining the same profile as the other datasets in the long run.

3.3 Fake News classification and uncertainty representation

Following experiment is made with respect to the same initial conditions as in Section 3.2 but with a wider choice of encoding techniques. Based on positive results from Section 3.2 we show a comparison of DEnFi and Dropout warm-start with respect to four different types of encodings and two fake news datasets.

Noise variance	AUC after # of iterations 50	100	150	200
0.2	**0.794**	**0.886***	**0.910**	**0.942***
0.3	**0.806***	**0.877**	**0.911***	**0.932**
0.4	**0.806***	0.866	**0.901**	0.911

(a) **Fast Text** DEnFi

Noise variance	AUC after # of iterations 50	100	150	200
0.2	**0.806**	**0.879***	**0.916***	**0.949***
0.3	**0.804**	**0.884***	0.914	0.939
0.4	**0.809***	**0.878**	**0.912**	**0.944**

(b) **Fast Text** Dropout warm-start

Noise variance	AUC after # of iterations 50	100	150	200
0.1	**0.929***	**0.986***	**0.992***	0.991
0.2	**0.923**	0.975	0.990	**0.998***
0.3	**0.891**	0.935	0.955	0.975

(c) **RoBERTa** DEnFi

Noise variance	AUC after # of iterations 50	100	150	200
0.1				
0.2	0.917	0.974	0.992	0.995
0.3	0.917	0.959	0.982	**0.990***

(d) **RoBERTa** Dropout warm-start

Table 3: **Fast Text** and **RoBERTa** encoding based AUC of text classification after selected number of requests of the active learning using DEnFi and Dropout warm-start for various selection of the perturbation noise q. Average over 5 runs on the Fake and True news categories (Fake News Detection dataset). The best result is denoted by a star, results within one standard deviation of the winner are displayed in bold font.

4 Conclusion

We have studied the suitability of various uncertainty representations for the task of active text classification. The established dropout methodology was compared against deep ensembles. To reduce the computational cost, we studied warm-start strategies for both ensembles (called DEnFi) and dropout. The resulting methods exhibit a different tradeoff between exploration and exploitation. While dropout has been found to be more reliable in passive learning and improving faster at the beginning of the training, the DEnFi was found to prefer exploration sacrificing performance at the beginning but outperforming dropout in the long run. The same has been observed for tuning of the variance of the noise used in the warm-start where higher variance implied shift towards exploration in the learning process. Both methods achieved significantly faster learning than the passive approach on all tested datasets of various complexity.

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.

Sreyasee Das Bhattacharjee, Ashit Talukder, and Bala Venkatram Balantrapu. 2017. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 556–565. IEEE.

Sreyasee Das Bhattacharjee, William J Tolone, and Ved Suhas Paranjape. 2019. Identifying malicious social media contents using multi-view context-aware active learning. *Future Generation Computer Systems*, 100:365–379.

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*.

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

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.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford. `http://help.sentiment140.com/for-students/` accessed on 26 june 2020.

Md Saqib Hasan, Rukshar Alam, and Muhammad Abdullah Adnan. 2020. Truth or lie: Pre-emptive detection of fake news in different languages through entropy-based active learning and multi-model neural ensemble.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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. `https://www.kaggle.com/rmisra/news-category-dataset`, accessed on 26 june 2020.

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*.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching. Technical report, Citeseer.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Christopher John Cornish Hellaby Watkins. 1989. Learning from delayed rewards.

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.