

Deep Active Learning for Text Classification with Diverse Interpretations

Qiang Liu^{1,2}, Yanqiao Zhu^{1,2}, Zhaocheng Liu³, Yufeng Zhang¹, and Shu Wu^{1,2,*}

¹Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences

²School of Artificial Intelligence, University of Chinese Academy of Sciences ³RealAI

{qiang.liu, shu.wu}@nlpr.ia.ac.cn, {yanqiao.zhu, yufeng.zhang}@cripac.ia.ac.cn, lio.h.zen@gmail.com

ABSTRACT

Recently, Deep Neural Networks (DNNs) have made remarkable progress for text classification, which, however, still require a large number of labeled data. To train high-performing models with the minimal annotation cost, active learning is proposed to select and label the most informative samples, yet it is still challenging to measure informativeness of samples used in DNNs. In this paper, inspired by piece-wise linear interpretability of DNNs, we propose a novel Active Learning with DivErse iNterpretations (ALDEN) approach. With local interpretations in DNNs, ALDEN identifies linearly separable regions of samples. Then, it selects samples according to their diversity of local interpretations and queries their labels. To tackle the text classification problem, we choose the word with the most diverse interpretations to represent the whole sentence. Extensive experiments demonstrate that ALDEN consistently outperforms several state-of-the-art deep active learning methods.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**;
Active learning settings; *Neural networks*.

KEYWORDS

Active learning; text classification; diverse interpretations

ACM Reference Format:

Qiang Liu, Yanqiao Zhu, Zhaocheng Liu, Yufeng Zhang, and Shu Wu. 2021. Deep Active Learning for Text Classification with Diverse Interpretations. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21)*, November 1–5, 2021, Virtual Event, QLD, Australia. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3459637.3482080>

1 INTRODUCTION

In recent years, Deep Neural Networks (DNNs) have achieved the state-of-the-art supervised performance in numerous research tasks. Among them, a typical task in natural language processing is text classification, where deep models such as Convolutional Neural

Networks (CNNs) [14] and Recurrent Neural Networks (RNNs) [28] are often adopted. However, such deep models require a large number of labeled samples, which are expensive and labor-consuming to obtain in real-world applications. Fortunately, active learning, which aims to identify and label the most informative samples from a pool of unlabeled data to train deep models with limited labels, is a promising approach to relieve this problem [1, 3, 29, 33].

Existing works on active learning mainly select samples based on uncertainty and diversity. Taking Expected Gradient Length (EGL) [12] as an example, it computes the sample uncertainty as the norms of gradients of losses with respect to the model parameters. Following EGL, EGL-Word [33] selects the word with the largest EGL among all samples to query its label so as to maximize the model performance for text classification. In addition, Bayesian Active Learning by Disagreement (BALD) [6] measures the uncertainty according to the probabilistic distribution of the model output via Bayesian inference, where an approximation by dropout is usually incorporated [5]. On the other hand, to measure the diversity of samples, some works define the active learning task as a CORESET problem [24] and uses the embedding of the last layer in deep models as the representation of samples. There are also attempts to trade off between uncertainty and diversity [13, 29]. For example, Batch Active learning by Diverse Gradient Embeddings (BADGE) [1] can be viewed as a combination of EGL and CORESET. Meanwhile, there are empirical experiments to evaluate above approaches on text classification [3, 21, 26, 30].

Recently, the interpretability of DNNs has received increasingly attention, among which most works focus on local piece-wise interpretability [2, 22]. To be specific, previous works [2, 10, 17] investigate the local interpretability of DNNs and show that a deep model with piece-wise linear activations, e.g., Maxout [8] and the family of ReLU [7, 18], can be regarded as a set of numerous local linear classifiers. The linear separable regions corresponding to these linear classifiers can be determined by the local piece-wise interpretations in DNNs that are calculated via gradient backpropagation [15, 23, 27, 32] or feature perturbation [4, 9]. In other words, samples used in a DNN could be divided into numerous linearly separable regions according to their local interpretations and samples in the same linearly separable region are classified by the same local linear classifier [2]. Therefore, fitting a DNN model is roughly equivalent to fitting all the linear classifiers in different linearly separable regions. Inspired by this, we propose to actively select samples in different linearly separable regions with the maximally diverse local interpretations, so that linear classifiers in different linearly separable regions can be all well trained.

In this paper, we propose a novel Active Learning with DivErse iNterpretations (ALDEN) approach for text classification. In our

*To whom correspondence should be addressed.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8446-9/21/11...\$15.00
<https://doi.org/10.1145/3459637.3482080>

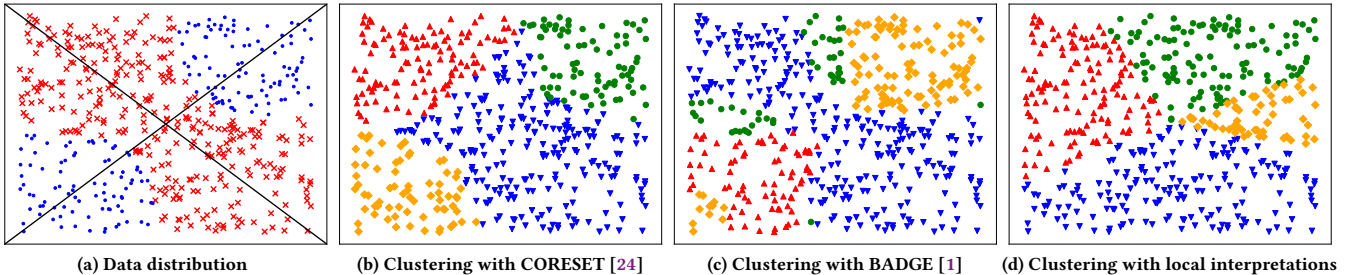


Figure 1: Illustrating local interpretations in DNNs. We artificially generate a series of data samples that could be roughly divided into four linearly separable regions (shown in four triangle areas). We perform k -Means clustering on the example data, where the representations of samples are from CORESET [24] and BADGE [1], as well as the local interpretations in DNNs computed using Eq. (1). The clusters are shown in four different colors. It is seen that only with interpretations we are able to correctly identify the four linearly separable regions.

proposed approach, we first calculate the local interpretations in DNN for each sample as the gradient backpropagated from the final predictions to the input features [15, 23]. Then, we use the most diverse interpretation of words in a sample to measure its diverseness. Accordingly, we select unlabeled samples with the maximally diverse interpretations for labeling and retrain the model with these labeled samples. We conduct experiments on two text classification datasets, with two representative deep classifiers: CNN [14] and Bi-directional Long Short-Term Memory (BiLSTM). Extensive experimental results show that ALDEN can constantly outperform state-of-the-art deep active learning approaches.

2 LOCAL INTERPRETATIONS IN DEEP NEURAL NETWORKS

Recently, extensive works have been conducted to study local piece-wise interpretability of DNNs, which can be computed using the gradient backpropagation from the predictions to the input features [15, 16, 23, 27, 32]. To be specific, we first train a deep model and obtain the prediction \hat{y}_i given input features x_i of a specific sample. Then, we can calculate local interpretations as

$$I_i = \frac{\partial \hat{y}_i}{\partial x_i}. \quad (1)$$

As in Li et al. [15], local interpretations could be formulated by

$$\hat{y}_i \approx I_i x_i^\top + b, \quad (2)$$

where b is the bias term. As mentioned in previous works [2, 17, 22], a DNN model with piece-wise linear activation functions (such as Maxout and ReLU [7, 8, 18]) can be regarded as a combination of numbers of local linear classifiers, which are introduced by the local interpretations in DNN. That is to say, local interpretations I_i of sample x_i as calculated in Eq. (1) can be partitioned into several clusters and each of them corresponds to a specific local linear classifier. With the local piece-wise interpretations in DNNs, samples can be divided into numerous linearly separable regions and samples in the same linearly separable region are classified by the same local linear classifier [2]. Therefore, fitting a DNN model means fitting all the linear classifiers in different linearly separable regions. Accordingly, if we select samples according to diverse local interpretations,

different linear classifiers in different linearly separable regions can be optimized in a more balanced way, so that the corresponding DNN model can be better trained. Thus, we argue that adopting local interpretations in DNN could potentially benefit deep active learning.

To demonstrate that the local Interpretations in DNNs can help promote deep active learning, we present a concrete example as shown in Figure , where example data are drawn from a probability distribution $p(y_i = 1 | x_i) = \sigma(x_{i,1} \cdot x_{i,2})$, where $x_{i,1}$ and $x_{i,2}$ are uniformly drawn from $[-5.0, 5.0]$, and $\sigma(\cdot)$ is the sigmoid function. The distribution of these artificially generated samples is shown in Figure 1a, which exhibits clear nonlinear characteristics. In addition, it is seen there are roughly four linearly separable regions, corresponding to the four triangle areas. For these samples, we run k -Means clustering on the representations generated by CORESET [24] and BADGE [1], as well as local interpretations in a Multi-Layer Perception (MLP) model, all trained on the example data. We set the number of clusters in k -Means to 4 and present the results in Figures 1b, 1c, and 1d respectively. We can observe that CORESET focuses on the original feature distribution and different classes, while BADGE pays more attention to the decision boundaries. Clearly, we can only use local interpretations to distinguish the four linearly separable regions. Therefore, with the help of local interpretations in DNNs, we are able to identify samples in different linearly separable regions. Inspired by this observation, we propose a deep active learning strategy to better fit all the linear classifiers corresponding to the DNN model.

3 THE PROPOSED ALDEN APPROACH

In this section, we introduce the ALDEN approach for text classification in detail.

3.1 Problem Formulation

In this work, we apply pool-based active learning in the batch mode [3, 25, 31, 33]. Specifically, we have a small set of labeled samples \mathcal{L} and a large set of unlabeled samples \mathcal{U} . Sample $x_i \in \mathcal{L}$ is associated with label y_i , while sample $x_i \in \mathcal{U}$ has no labels. The feature vector x_i is denoted as $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,|x_i|})$, where $x_{i,j}$ is a word in the

sample. With the labeled samples in \mathcal{L} , we can train a text classifier $f(x|\theta): \mathcal{X} \rightarrow \mathcal{Y}$. We need to develop an active learning strategy to select samples from \mathcal{U} and add them to \mathcal{L} for further training the classifier. We set the label budget to K samples per iteration of sample selection and train the model for a total of N iterations.

3.2 Approach Details

Regarding active learning for text classification, similar to Eq. (1), for a word in a specific sample used in a deep text classifier, we can compute its local interpretation as

$$I_{i,j} = \frac{\partial \hat{y}_i}{\partial x_{i,j}}, \quad (3)$$

where \hat{y}_i is the prediction of sample x_i . Equivalently, we can also calculate Eq. (3) using word embedding $e_{i,j}$ of word $x_{i,j}$:

$$I_{i,j} = \frac{\partial \hat{y}_i}{\partial e_{i,j}} e_{i,j}^\top. \quad (4)$$

Recall that the local interpretation of a word indicates its contribution to the final prediction; similar to Eq. (2), the prediction can be approximated [15] as

$$\hat{y}_i \approx \sum_{1 \leq j \leq |x_i|} \frac{\partial \hat{y}_i}{\partial e_{i,j}} e_{i,j}^\top + b. \quad (5)$$

Consider that local interpretations (i.e. the contribution to the model predictions) of the same word may be different among different samples, due to the complex nonlinear feature interactions modeled by deep models [2, 17, 22]. As discussed in Section 2, we need to select samples with the diverse local interpretations, so that linear classifiers in different linearly separable regions can be well optimized. Meanwhile, since diverse interpretations indicate different decision regions in the deep model, samples with the maximally diverse interpretations can provide the most comprehensive information to learn diverse decision logics in the deep model. For the task of text classification, as different samples consist of various numbers of words, we need to start with analyzing local interpretations of words in each sample. In particular, we calculate the interpretation diversity of a word $x_{i,j}$ compared to the same word appeared in labeled samples

$$D(\mathcal{L}, x_{i,j}) = \min_{\substack{x_m \in \mathcal{L} \\ 1 \leq w \leq |x_m| \\ x_{m,w} = x_{i,j}}} \|I_{i,j} - I_{m,w}\|, \quad (6)$$

which is similar to the distance calculation in the greedy k -Center algorithm [24]. However, some words may not appear in the labeled samples, which makes it infeasible to directly calculate Eq. (6). As a remedy, we search for the most similar embedding of the word appearing in the labeled samples as the neighbor, which is formulated as

$$N(\mathcal{L}, x_{i,j}) = \operatorname{argmin}_{\substack{x_m \in \mathcal{L} \\ 1 \leq w \leq |x_m|}} \|e_{i,j} - e_{m,w}\|. \quad (7)$$

Then, we can rewrite Eq. (6) as

$$D(\mathcal{L}, x_{i,j}) = \min_{\substack{x_m \in \mathcal{L} \\ 1 \leq w \leq |x_m| \\ x_{m,w} = N(\mathcal{L}, x_{i,j})}} \|I_{i,j} - I_{m,w}\|. \quad (8)$$

Considering that there are various numbers of words in sentences, which brings difficulties in directly using the local interpretations of all words in a sample. Therefore, we adopt a pooling strategy for active learning. Recall that in EGL-Word [33], the word with the largest EGL is used to represent the whole sentence. Aligning with EGL-Word, we also use the word with the maximally diverse interpretation to represent the whole sample for active learning. Formally, for a sample $x_i \in \mathcal{U}$, we have

$$D(\mathcal{L}, x_i) = \max_{1 \leq j \leq |x_i|} D(\mathcal{L}, x_{i,j}). \quad (9)$$

Based on the metric calculated using Eq. (9), we can select the unlabeled sample that has the maximally diverse interpretation for labeling:

$$x = \operatorname{argmax}_{x_i \in \mathcal{U}} D(\mathcal{L}, x_i). \quad (10)$$

Since we are given a budget of K in each iteration, we repeat the above process for K times to select and label K samples. Algorithm 1 summarizes the training procedure of the ALDEN approach.

Algorithm 1: The ALDEN approach

Data: Labeled samples \mathcal{L} , unlabeled samples \mathcal{U} , budget K in each iteration, and the number of iterations N .

- 1 Train an initial model $f(x | \theta_0)$ on \mathcal{L}
 - 2 **for** $n = 1, 2, \dots, N$ **do**
 - 3 **for** $x_i \in \mathcal{L} \cup \mathcal{U}$ **do**
 - 4 Calculate prediction $\hat{y}_i = f(x_i | \theta_{n-1})$
 - 5 **for** $1 \leq j \leq |x_i|$ **do**
 - 6 Calculate the local interpretation $I_{i,j}$ according to Eq. (4)
 - 7 **for** $k = 1, 2, \dots, K$ **do**
 - 8 **for** $x_i \in \mathcal{U}$ **do**
 - 9 **for** $1 \leq j \leq |x_i|$ **do**
 - 10 Find the neighbor $N(\mathcal{L}, x_{i,j})$ of word $x_{i,j}$ according to Eq. (7)
 - 11 Compute the diversity $D(\mathcal{L}, x_{i,j})$ of local interpretations of $x_{i,j}$ according to Eq. (8)
 - 12 Compute the diversity $D(\mathcal{L}, x_i)$ of the local interpretation of x_i according to Eq. (9)
 - 13 Select and label the sample x having the most diverse local interpretations
 - 14 $\mathcal{L} = \mathcal{L} \cup \{x\}$
 - 15 $\mathcal{U} = \mathcal{U} \setminus x$
 - 16 Train a new model $f(x | \theta_n)$ on \mathcal{L}
 - 17 **return** The final model $f(x | \theta_N)$
-

4 EXPERIMENTS

In this section, we empirically evaluate our proposed ALDEN approach on the task of text classification.

4.1 Baseline Approaches

To evaluate the effectiveness of ALDEN, we compare it with the following approaches:

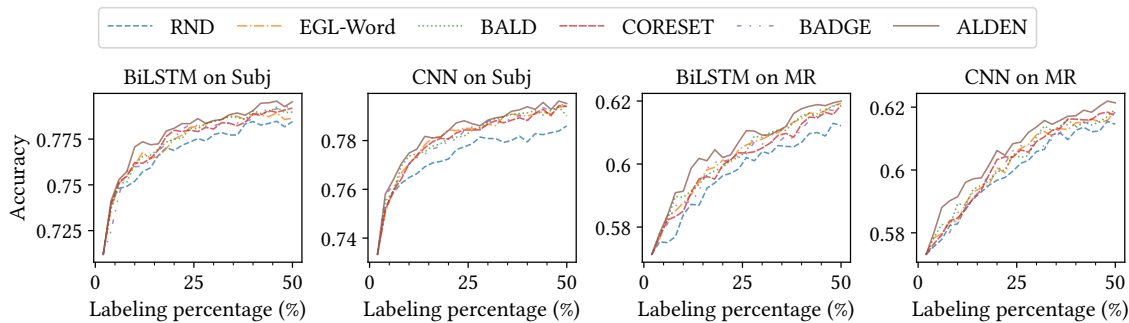


Figure 2: Learning curves in terms of accuracy of compared approaches with various labeling rates of training samples.

Table 1: Normalized area under curve scores of learning curves. The larger the values, the better the performances.

Model	Subj		MR	
	BiLSTM	CNN	BiLSTM	CNN
RND	0.688	0.658	0.531	0.594
EGL-Word	0.750	0.775	0.644	0.650
BALD	0.757	0.773	0.645	0.658
CORESET	0.752	0.764	0.612	0.659
BADGE	0.744	0.767	0.619	0.641
ALDEN	0.803	0.814	0.700	0.746

- **RND** is a simple baseline which randomly selects samples in each iteration.
- **EGL-Word** [33] is an extension of EGL [12], which utilizes norms of gradients to measure uncertainty for the task of text classification.
- **BALD** [11] is an uncertainty-based approach based on Bayesian inference. We apply dropout approximation [5, 6] in our experiments, where the dropout rate is set to 0.5.
- **CORESET** [24] uses the representations of the last layer in DNN as the representations.
- **BADGE** [1] can be viewed as a combination of EGL and CORESET.

4.2 Experimental Settings

To evaluate the performance of ALDEN, we use two sentence classification datasets¹: **Subj** [19] and **MR** [20], which contain 5000, 5331 positive samples and 5000, 5331 negative samples respectively. In our experiments, we use accuracy as the evaluation metric. We run each approach 10 times and report the median of results. We randomly select 60%, 20%, and 20% samples in each dataset for training, validation, and testing respectively. We train a word2vec² model on each dataset to initialize the word embeddings and set the hidden dimensionality to 100. We use two deep models: BiLSTM and CNN for comprehensive evaluation. For the implementation of BiLSTM, we use a single bidirectional LSTM layer with 100 hidden units. For

the implementation of CNN, we set the filter size to (3, 4, 5) and set the hidden dimension to 100 as well. In both BiLSTM and CNN, we apply the ReLU activation and the dropout rate is set to 0.5. We use 2% samples in the training set as the initial seed labeled set. Furthermore, we label 2% samples in the training set during each iteration until 50% samples in the training set have been labeled. In other words, we set N to 24 and K to 2% of training samples for each dataset.

4.3 Results and Analysis

We present the learning curve of the performance with different ratios of labeled samples in Figure 2. It is seen from the figure that in most cases, active learning approaches outperform random selection, which demonstrates the necessity of deep active learning. EGL-Word and BALD perform similarly and they both slightly outperform CORESET and BADGE. Meanwhile, it is clear that ALDEN constantly outperform other compared approaches, which is demonstrated especially in the middle parts of the learning curves.

Additionally in Table 1, we calculate the normalized area under curve scores of learning curves in Figure 2. This metric evaluates the global performance of each compared approach and it is evident that ALDEN achieves the best performance. In summary, these results strongly demonstrate the advantages of our proposed ALDEN approach.

5 CONCLUSION

In this paper, inspired by the local piece-wise interpretability of DNNs, we introduce the linearly separable regions of samples to the problem of deep active learning. For the task of text classification, we propose a novel ALDEN approach, which selects and labels samples according to the diverse interpretations of unlabeled sample. Specifically, we use the most diverse interpretation of words in a sample to measure the sample diversity. Experimental results on two text classification datasets with CNN and BiLSTM as classifiers show that the ALDEN approach is able to consistently outperform state-of-the-art deep active learning approaches.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (U19B2038, 61772528) and Shandong Provincial Key Research and Development Program (2019JZZY010119).

¹<http://www.cs.cornell.edu/people/pabo/movie-review-data/>

²<https://code.google.com/archive/p/word2vec/>

REFERENCES

- [1] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *ICLR*.
- [2] Lingyang Chu, Xia Hu, Juhua Hu, Lanjun Wang, and Jian Pei. 2018. Exact and Consistent Interpretation for Piecewise Linear Neural Networks: A Closed Form Solution. In *KDD*. 1244–1253.
- [3] Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *EMNLP*. 7949–7962.
- [4] Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *ICCV*. 3449–3457.
- [5] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *ICML*. 1050–1059.
- [6] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian Active Learning with Image Data. In *ICML*. 1183–1192.
- [7] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. In *AISTATS*. 315–323.
- [8] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron C. Courville, and Yoshua Bengio. 2013. Maxout Networks. In *ICML*. 1319–1327.
- [9] Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a Deep and Unified Understanding of Deep Neural Models in NLP. In *ICML*. 2454–2463.
- [10] Nick Harvey, Christopher Liaw, and Abbas Mehrabian. 2017. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *COLT*. 1064–1068.
- [11] Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian Active Learning for Classification and Preference Learning. *arXiv.org* (2011). [arXiv:1112.5745](https://arxiv.org/abs/1112.5745)
- [12] Jiayi Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. 2016. Active Learning for Speech Recognition: the Power of Gradients. *arXiv.org* (2016). [arXiv:1612.03226](https://arxiv.org/abs/1612.03226)
- [13] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. 2014. Active Learning by Querying Informative and Representative Examples. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 10 (2014), 1936–1949.
- [14] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*. 1746–1751.
- [15] Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky. 2016. Visualizing and Understanding Neural Models in NLP. In *HLT-NAACL*. 681–691.
- [16] Qiang Liu, Zhaocheng Liu, Haoli Zhang, Yuntian Chen, and Jun Zhu. 2021. Mining Cross Features for Financial Credit Risk Assessment. In *CIKM*.
- [17] Guido F. Montúfar, Razvan Pascanu, KyungHyun Cho, and Yoshua Bengio. 2014. On the Number of Linear Regions of Deep Neural Networks. In *NeurIPS*. 2924–2932.
- [18] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *ICML*. 807–814.
- [19] Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *ACL*. 271–278.
- [20] Bo Pang and Lillian Lee. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *ACL*. 115–124.
- [21] Ameeya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling Bias in Deep Active Classification: An Empirical Study. In *EMNLP/IJCNLP*. 4056–4066.
- [22] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *KDD*. 1135–1144.
- [23] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* 128, 2 (2020), 336–359.
- [24] Ozan Sener and Silvio Savarese. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *ICLR*.
- [25] Burr Settles. 2009. *Active Learning Literature Survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.
- [26] Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study. In *EMNLP*. 2904–2909.
- [27] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. *arXiv.org* (2017). [arXiv:1706.03825](https://arxiv.org/abs/1706.03825)
- [28] Ruishuang Wang, Zhao Li, Jian Cao, Tong Chen, and Lei Wang. 2019. Convolutional Recurrent Neural Networks for Text Classification. In *IJCNN*. 1–6.
- [29] Zheng Wang and Jieping Ye. 2015. Querying Discriminative and Representative Samples for Batch Mode Active Learning. *ACM Trans. Knowl. Discov. Data* 9, 3 (2015), 17:1–17:23.
- [30] Yifan Yan, Sheng-Jun Huang, Shaoyi Chen, Meng Liao, and Jin Xu. 2020. Active Learning with Query Generation for Cost-Effective Text Classification. In *AAAI*. 6583–6590.
- [31] Kai Yang, Jie Ren, Yanqiao Zhu, and Weiye Zhang. 2018. Active Learning for Wireless IoT Intrusion Detection. *IEEE Wireless Communications* 25, 6 (Dec. 2018), 19–25.
- [32] Hao Yuan, Yongjun Chen, Xia Hu, and Shuiwang Ji. 2019. Interpreting Deep Models for Text Analysis via Optimization and Regularization Methods. In *AAAI*. 5717–5724.
- [33] Ye Zhang, Matthew Lease, and Byron C. Wallace. 2017. Active Discriminative Text Representation Learning. In *AAAI*. 3386–3392.