Deep Active Learning by Model Interpretability

Qiang Liu[1,2], Zhaocheng Liu[1], Xiaofang Zhu[1], Yeliang Xiu[3], Jun Zhu[1,2]

[1]RealAI, [2]Tsinghua University, [3]Sun Yat-sen University

{qiang.liu,zhaocheng.liu,xiaofang.zhu}@realai.ai, 1448560127@qq.com, dcszj@mail.tsinghua.edu.cn

Abstract

Recent successes of Deep Neural Networks (DNNs) in a variety of research tasks, however, heavily rely on the large amounts of labeled samples. This may require considerable annotation cost in real-world applications. Fortunately, active learning is a promising methodology to train high-performing model with minimal annotation cost. In the deep learning context, the critical question of active learning is how to precisely identify the informativeness of samples for DNN. In this paper, inspired by piece-wise linear interpretability in DNN, we first introduce the linear separable regions of samples to the problem of active learning, and propose a novel Deep Active learning approach by Model Interpretability (DAMI). To keep the maximal representativeness of the entire unlabeled data, DAMI tries to select and label samples on different linear separable regions introduced by the piece-wise linear interpretability in DNN. We focus on two scenarios: 1) Multi-Layer Perception (MLP) for modeling tabular data; 2) language models for modeling textual data. On tabular data, we use the local piece-wise interpretation in DNN as the representation of each sample, and directly run K-mediods clustering to select and label the central sample in each cluster. On textual data, we propose a novel aggregator to find the most informative word in each sentence, and use its local piece-wise interpretation as the representation of the sentence. To be noted, this whole process of DAMI does not require any hyper-parameters to tune manually. To verify the effectiveness of our approach, extensive experiments have been conducted on both tabular datasets and textual datasets. The experimental results demonstrate that DAMI constantly outperforms several state-of-the-art compared methods.

Introduction

Over the past decades, Deep Neural Networks (DNNs) have represented the state-of-the-art supervised learning models and shown unprecedented success in numerous research tasks. However, these successes heavily rely on large amount of labeled training samples. A promising approach to address this problem is active learning, which aims to find effective ways to identify and label the maximally informative samples from a pool of unlabeled data (Wang and Ye 2015; Ash et al. 2020).

Previous works on active learning mainly quantify samples from uncertainty and representative. Expected Gradient Length (EGL) (Huang et al. 2016; Zhang, Lease, and Wallace 2017) is a typical uncertainty-based method, which regards the norms of gradients of losses with respect to the model parameters as the uncertainty evaluation. Bayesian Active Learning by Disagreement (BALD) (Houlsby et al. 2011; Gal, Islam, and Ghahramani 2017; Siddhant and Lipton 2018) measures uncertainty according to the probabilistic distribution of model outputs via Bayesian inference (Zhu et al. 2017), where an approximation by dropout are usually incorporated (Gal and Ghahramani 2016). Among representative-based approaches, in the deep learning context, some works define the active learning task as a CORE-SET problem (Sener and Savarese 2018), which uses the representations of the last layer in DNN as representations of samples. Besides, there are several approaches trade off between uncertainty and representative (Wang and Ye 2015; Ash et al. 2020). For the active learning task in deep learning, Batch Active learning by Diverse Gradient Embeddings (BADGE) (Ash et al. 2020) utilizes gradients of losses with respect to the representations of the last layer in DNN as representations of samples, on which clustering is conducted for capturing both uncertainty and representative.

Recently, the interpretability of DNN has been widely studied, among which most works focus on local piece-wise interpretability (Ribeiro, Singh, and Guestrin 2016; Chu et al. 2018). Specifically, the local piece-wise interpretations of DNN can be calculated via gradient backpropagation (Li et al. 2015; Selvaraju et al. 2017) or feature perturbation (Fong and Vedaldi 2017; Guan et al. 2019). Some previous works (Montufar et al. 2014; Harvey, Liaw, and Mehrabian 2017; Chu et al. 2018) deeply investigate the local interpretability of DNN, and shows that DNN with piece-wise linear activations, e.g., Maxout (Goodfellow et al. 2013) and the family of ReLU (Nair and Hinton 2010; Glorot, Bordes, and Bengio 2011), can be regraded as a set of numerous local linear classifiers. That is to say, with DNN, samples are divided into numerous linear separable regions. As we know, we usually need the same numbers of samples for fitting different linear classifiers on different linear separable regions. For example, each linear classifier on two-dimensional space requires two samples for fitting. Thus, to select samples for optimally training DNN, differ-

ent linear separable regions should be considered in a balance way. From this perspective, with the help of local interpretability of DNN, we can identify different linear separable regions of samples, and potentially promote the effectiveness of deep active learning.

Accordingly, in this paper, we introduce the linear separable regions of samples to the problem of active learning for DNN, and propose a novel Deep Active learning approach by Model Interpretability (DAMI). Specifically, we calculate the local interpretations in DNN via the gradient backpropagation from the final predictions to the input features (Li et al. 2015; Selvaraju et al. 2017). In this paper, we focus on two scenarios: 1) Multi-Layer Perception (MLP), which has fixed number of input features, for classification on tabular data; 2) language models, which have variant numbers of input features, for classification on textual data. On tabular data, we use local interpretations in DNN as the representations of samples, and directly run K-mediods clustering (Park and Jun 2009) to select and label the central sample in each cluster. On textual data, we propose a novel aggregator to find the most informative word in each sentence, and use its local piece-wise interpretation as the representation of the sentence, on which K-mediods clustering is also performed. We have conducted extensive experiments on four tabular datasets and two textual datasets. The experimental results show that DAMI can constantly outperform state-of-the-art active learning approaches.

Related Works

In this section, we briefly review some related works on active learning, as well as interpretability of DNN.

Active Learning

Based on a certain sampling strategy, active learning approaches actively samples a small batch of informative instances from the unlabeled data for labeling. Roughly speaking, there exist two major types of strategies: representative-based sampling and uncertainty-based sampling.

Representative-based sampling aims to select unlabeled samples that are representative according to the data distribution. In the deep learning context, this is usually done based on CORESET construction (Sener and Savarese 2018), in which the representations of the last layer in DNN are used as representations of samples. Adversarial learning can also be considered to select most indistinguishable samples (Ducoffe and Precioso 2018).

Uncertainty-based sampling aims to select samples that can maximally reduce the uncertainty of the classifier. Such approaches are widely applied in the deep learning context. EGL (Huang et al. 2016) measures uncertainty based on the norms of gradients of losses with respect to the model parameters. For the task of sentence classification, EGL-word (Zhang, Lease, and Wallace 2017) seeks to find the word with largest norm of gradients in a sentence, and uses the corresponding norm as the uncertainty measurement. BALD (Houlsby et al. 2011) measures uncertainty according to the probabilistic distribution of model outputs via Bayesian inference (Zhu et al. 2017). Inspired by the finding that Bayesian inference can be approximated by dropout
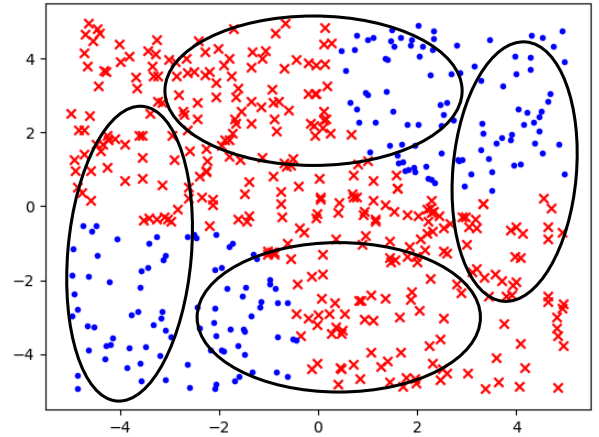

Figure 1: Distribution of example data in the Sigmoid dataset, which is formulated in Eq. (2). Four linear separable regions are shown in the circles.

in deep models (Gal and Ghahramani 2016), the dropout approximation is usually applied in deep active learning to perform BALD (Gal, Islam, and Ghahramani 2017). And the BALD approach is successfully applied in the task of sentence classification (Siddhant and Lipton 2018). Meanwhile, the uncertainty-based approaches have been empirically studied and evaluated for deep active learning on textual data (Prabhu, Dognin, and Singh 2019).

Furthermore, some works consider uncertainty and representative at the same time, and make trade-off between them (Huang, Jin, and Zhou 2010; Wang and Ye 2015; Hsu and Lin 2015). For example, such trade-off is considered for text classification (Yan et al. 2020). In the context of deep learning, BADGE (Ash et al. 2020) is proposed to take use of gradients of losses with respect to the representations of the last layer in DNN as representations of samples, in which both uncertainty and representative can be preserved to some extend.

Interpretability of DNN

Recently, the interpretability of DNN has drawn great attention in academia, and research works mostly focus on local piece-wise interpretability, which means assigning a piece of local interpretation for each sample (Guidotti et al. 2018). Some unified approaches are proposed to fit a linear classifier in each local space of input samples (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017). Some works investigate the gradients from the final predictions to the input features in deep models, which can be applied in the visualization of deep vision models (Zhou et al. 2016; Selvaraju et al. 2017; Smilkov et al. 2017; Alvarez-Melis and Jaakkola 2018), as well as the interpretation of language models (Li et al. 2015; Yuan et al. 2019). Perturbation on input features is also utilized to find local interpretations of both vision models (Fong and Vedaldi 2017) and language models (Guan et al. 2019). Meanwhile, via adversarial diagnosis of neural networks, adversarial ex-
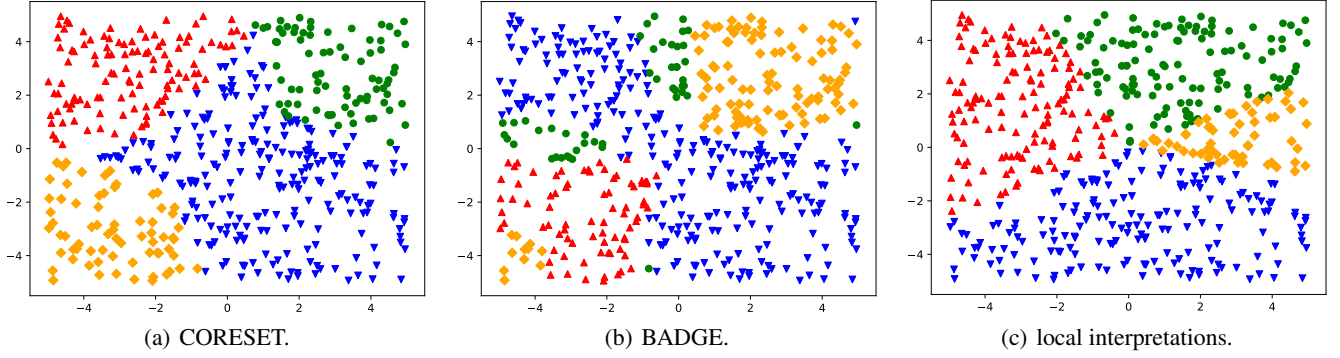
| (a) CORESET. | (b) BADGE. | (c) local interpretations. |

Figure 2: K-mediods clustering on samples in the Sigmoid dataset, which is formulated in Eq. (2) and illustrated in Fig. 1. The representations of samples are from CORESET (Sener and Savarese 2018) and BADGE (Ash et al. 2020), as well as the local interpretations in DNN via the calculation in Eq. (1). Clearly, only the local interpretations can find the four linear separable regions in the Sigmoid dataset.

amples can also be introduced for local interpretation of DNN (Koh and Liang 2017; Dong et al. 2018). In some views, attention in deep models can also be regarded as local interpretations (Ghaeini, Fern, and Tadepalli 2018; Wang et al. 2019).

As discussed in some previous works (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017), the nonlinear DNN model can be regarded as a combination of numbers of linear classifiers, and the upper bound of the number in DNN with piece-wise linear activations has been given (Montufar et al. 2014). Moreover, piece-wise linear DNN has been exactly and consistently interpreted as a set of linear classifiers (Chu et al. 2018). In a word, with the local piece-wise interpretations of DNN, we can define the linear separable regions of input samples.

The DAMI Approach

In this section, we introduce the proposed DAMI approach on two scenarios: 1) the MLP model for the classification task on tabular data; 2) language models for the classification task on textual data.

Notations

In this work, we consider the pool-based AL case (Tong and Koller 2001; Settles and Craven 2008; Zhang, Lease, and Wallace 2017), in which we have a small set of labeled samples $\mathcal{L}$, and a large set of unlabeled samples $\mathcal{U}$.

For sample $s_i \in \mathcal{L}$, we have $s_i = (x_i, y_i)$, where x_i and $y_i \in \{0, 1\}$ are the corresponding features and label. For sample $s_i \in \mathcal{U}$, we have $s_i = (x_i,)$, where the label is unknown. With the labeled samples $\mathcal{L}$, we can train a deep learning-based classifier $f(x|\theta) \colon \mathcal{X} \to \mathcal{Y}$, which maps the features to the labels. Then, we can develop a AL strategy to select most informative samples from $\mathcal{U}$, and further optimize the classifier.

In this paper, we focus on two types of data, i.e., tabular data and textual data. For sample s_i in tabular data, x_i is a fixed-size feature vector, and denoted as $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,M})$, where M is the width of the tabular

data. For sample s_i in textual data, x_i is a variant-size feature sentence, and denoted as $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,|x_i|})$, where $x_{i,j}$ is a word in the sentence.

Learning from the Interpretations of DNN

Recently, extensive works have been conducted to study local piece-wise interpretability of DNN. And the calculation of local interpretations can be done via the gradient backpropagation from the predictions to the input features (Selvaraju et al. 2017; Li et al. 2015; Smilkov et al. 2017; Yuan et al. 2019). We need to first train a deep model, and obtain the predicted label $\hat{y}_i$ for sample $s_i \in \mathcal{U}$. Then, we can calculate local interpretations of sample s_i as

$$I_i = \frac{\partial \hat{y}_i}{\partial x_i}. \tag{1}$$

As mentioned in some works (Montufar et al. 2014; Ribeiro, Singh, and Guestrin 2016; Chu et al. 2018), a DNN model can be regarded as a combination of numbers of linear classifiers. That is to say, samples can be divided into numerous linear separable regions. To demonstrate the Interpretations of DNN can help to promote deep active learning, we draw some example data from the following probability distribution:

$$p(y_i = 1 | x_i) = \text{sigmoid}(x_{i,1} \cdot x_{i,2}), \tag{2}$$

where $x_{i,1}$ and $x_{i,2}$ are uniformly sampled from $[-5.0, 5.0]$. For simplicity, these toy samples are named the Sigmoid dataset, whose data distribution is shown in Fig. 1. This data is clearly nonlinear, and there are mainly four linear separable regions, which are illustrated in the four circles. Usually, we need the same numbers of samples for fitting different linear classifiers on different linear separable regions. For example, a linear classifier on t-dimensional space requires t samples for fitting. With the help of the interpretations of DNN, we are able to find different linear separable regions of the unlabeled samples, and propose a better deep active learning approach. To illustrate this, for samples in the Sigmoid dataset, we run K-mediods clustering (Park and Jun

Algorithm 1 DAMI on Tabular Data.
___
Require: Labeled samples $\mathcal{L}$, unlabeled samples $\mathcal{U}$, number of iterations M, budget k in each iteration of sample selection.
1: Train an initial MLP model $f(x|\theta_0)$ on $\mathcal{L}$;
2: **for** $m = 1, 2, ..., M$ **do**
3:    **for** $s_i \in \mathcal{U}$ **do**
4:       Make prediction $\hat{y}_i = f(x_i | \theta_{m-1})$;
5:       Compute the local interpretation I_i as in Eq. (1);
6:    **end for**
7:    Run K-mediods on $\{I_i \,|\, s_i \in \mathcal{U}\}$ to partition $\mathcal{U}$ into k clusters, and assign cluster centers as $\mathcal{L}_m$;
8:    Label samples $s_i \in \mathcal{L}_m$;
9:    $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{L}_m$;
10:   $\mathcal{U} \leftarrow \mathcal{U} \backslash \mathcal{L}_m$;
11:   Train a new MLP model $f(x|\theta_m)$ on $\mathcal{L}$;
12: **end for**
13: **return** The final model $f(x|\theta_M)$.
___

2009) on the representations generated by CORESET (Sener and Savarese 2018) and BADGE (Ash et al. 2020), as well as the local interpretations in a MLP model trained on the Sigmoid dataset. We set the number of clusters in K-mediods as 4, and results are shown in Fig. 2. We can observe that, CORESET focuses on the original feature distribution and different classes, while BADGE pays more attention to the decision boundaries. And clearly, we can only use local interpretations to find the four linear separable regions.

DAMI on Tabular Data

Inspired by the local interpretability of DNN, we can introduce piece-wise linear separable regions to the problem of deep active learning. Specifically, in this subsection, we detail the DAMI approach on tabular data.

In tabular data, there are fixed number of input features, and MLP is usually applied for modeling. To find most informative unlabeled samples, according to the calculation in Eq. (1), we can directly utilize the local interpretations as the representations of samples. K-center has been used for finding informative samples based on their representations (Sener and Savarese 2018; Ash et al. 2020). Recently, K-mediods (Park and Jun 2009) has been proven to obtain a more favorable loss bound than K-Center, and is more computational efficient (Wu et al. 2019). Accordingly, we adopt K-mediods in our approach. Specially, with budget k in each iteration of sample selection, we run K-mediods clustering on $\{I_i \,|\, s_i \in \mathcal{U}\}$ to partition unlabeled samples into k clusters, and then label the corresponding cluster centers. Detailed process can be found in Alg. 1.

DAMI on Textual Data

For modeling textual data, various language models, e.g., Convolutional Neural Networks (CNNs) (Kim 2014) and Bi-directional Long Short-Term Memory (BiLSTM), can be utilized. Different from tabular data, textual data usually has variant numbers of input features. That is to say, for different samples in textual data, the corresponding local interpretations are usually with different sizes. Thus, local interpretations in language models can not be directly utilized for deep active learning on textual data.

Algorithm 2 DAMI on Textual Data.
___
Require: Labeled samples $\mathcal{L}$, unlabeled samples $\mathcal{U}$, number of iterations M, budget k in each iteration of sample selection.
1: Train an initial CNN or BiLSTM model $f(x|\theta_0)$ on $\mathcal{L}$;
2: **for** $m = 1, 2, ..., M$ **do**
3:    **for** $s_i \in \mathcal{U}$ **do**
4:       Make prediction $\hat{y}_i = f(x_i | \theta_{m-1})$;
5:       **for** $1 \le j \le |x_i|$ **do**
6:          Compute the local interpretation $I_{i,j}$ as in Eq. (3);
7:       **end for**
8:    **end for**
9:    Compute the global average interpretation $\bar{I}$ as in Eq. (4);
10:   **for** $s_i \in \mathcal{U}$ **do**
11:      Compute the most informative word d_i as in Eq. (5);
12:      Obtain the representation r_i as in Eq. (6);
13:   **end for**
14:   Run K-mediods on $\{r_i \,|\, s_i \in \mathcal{U}\}$ to partition $\mathcal{U}$ into k clusters, and assign cluster centers as $\mathcal{L}_m$;
15:   Label samples $s_i \in \mathcal{L}_m$;
16:   $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{L}_m$;
17:   $\mathcal{U} \leftarrow \mathcal{U} \backslash \mathcal{L}_m$;
18:   Train a new CNN or BiLSTM model $f(x|\theta_m)$ on $\mathcal{L}$;
19: **end for**
20: **return** The final model $f(x|\theta_M)$.
___

EGL-word (Zhang, Lease, and Wallace 2017) proposes to find the word with largest norm of gradients in a sentence. Inspired by this, we plan to find the most informative word in a sentence, and use its local interpretation as the representation of the sentence. An informative word tends to have discriminative local interpretation compared to other words. To find such words, we need first to calculate the local interpretation of each word, as well as the global average interpretation:

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

$$\bar{I} = \frac{1}{\mathcal{U}} \sum_{s_i \in \mathcal{U}} \frac{1}{|x_i|} \sum_{1 \le j \le |x_i|} I_{i,j}. \tag{4}$$

Then, we can find the word with most discriminative interpretation in each sentence, and obtain the corresponding representation of each sample:

$$d_i = \underset{1 \le j \le |x_i|}{\operatorname{argmax}} \left(I_{i,j} - \bar{I}\right)^2, \tag{5}$$

$$r_i = I_{i,d_i}. \tag{6}$$

Similar as on tabular data, we run K-mediods clustering on $\{r_i \,|\, s_i \in \mathcal{U}\}$ to find samples to be labeled. Detailed process can be found in Alg. 2.

Experiments

In this section, we empirically evaluate our proposed DAMI approach. Extensive evaluations are conducted to answer the following research questions:

- **RQ1** What kind of samples are selected by DAMI?
- **RQ2** How are the performances of DAMI on tabular data?
- **RQ3** How are the performances of DAMI on textual data?

| (a) Original data distribution. | (b) EGL. | (c) BALD. |

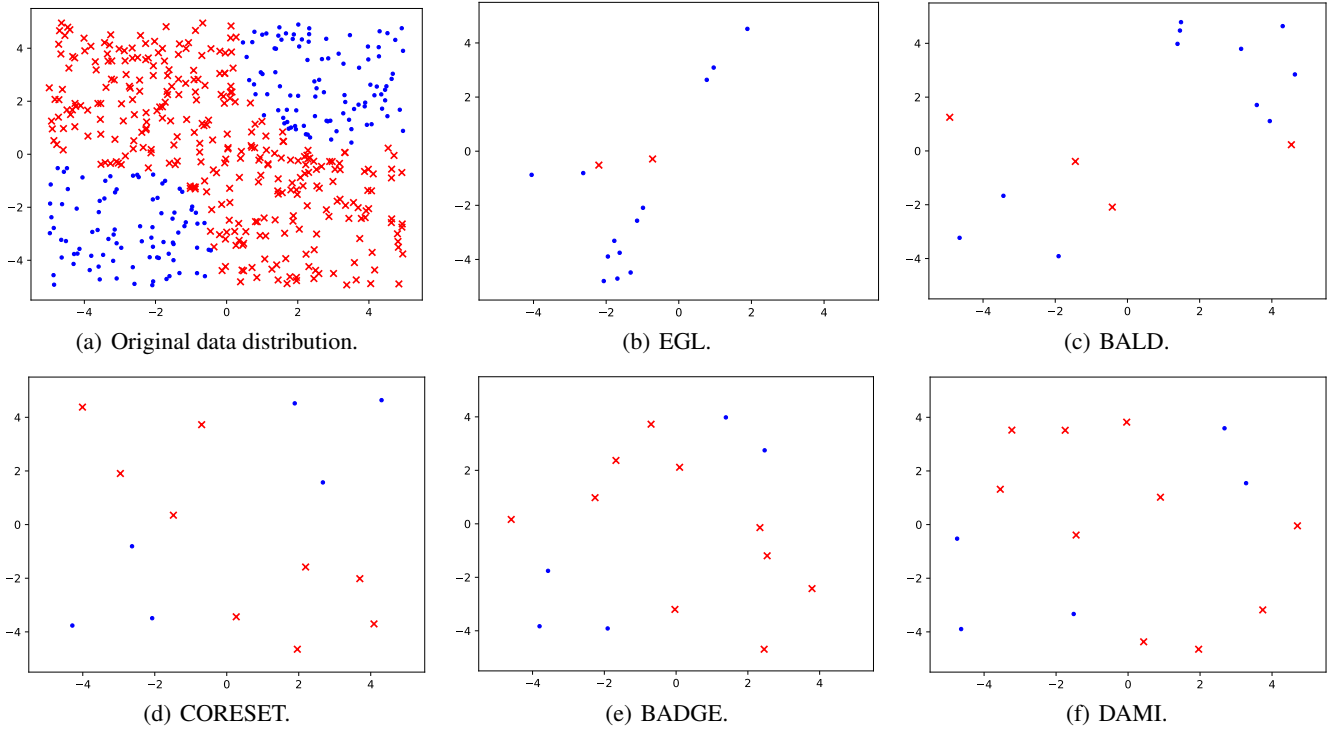| (d) CORESET. | (e) BADGE. | (f) DAMI. |

Figure 3: 15 unlabeled samples selected by EGL (Huang et al. 2016), BALD (Gal, Islam, and Ghahramani 2017), CORESET (Sener and Savarese 2018), BADGE (Ash et al. 2020) and DAMI (ours) on the Sigmoid dataset. The original data distribution is also illustrated. Compared with other state-of-the-art approaches, DAMI can best capture different linear separable regions.

Experiments on Toy Dataset

To find out what kind of samples are selected by DAMI, we conduct experiments on a toy dataset, i.e., the Sigmoid dataset. The samples in the Sigmoid Dataset are formulated in Eq. (2). We show the original data distribution of samples in the Sigmoid Dataset in Fig. 3(a). To show the difference among DAMI and other state-of-the-art deep active learning approaches, we perform following approaches:

- **EGL** (Huang et al. 2016) is a typical uncertainty-based approach, which utilizes norms of gradients.

- **BALD** (Houlsby et al. 2011) is a another uncertainty-based approach based on Bayesian inference. We apply dropout approximation (Gal and Ghahramani 2016; Gal, Islam, and Ghahramani 2017) in our experiments. And the the dropout rate is set as 0.5.

- **CORESET** (Sener and Savarese 2018) uses the representations of the last layer in DNN as the representations.

- **BADGE** (Ash et al. 2020) can be viewed as a combination of EGL and CORESET.

- **DAMI** is proposed in this paper, which conduct deep active learning based the local interpretability in DNN.

We run 3 layers of MLP on samples in the Sigmoid dataset, where then hidden units are set as $(16, 8)$. We randomly select 100 samples from the Sigmoid dataset as labeled samples, for the initial training of the MLP model.

Then, we select 15 unlabeled samples accordingly to the five compared deep active learning approaches.

As shown in Fig. 3(b) and 3(c), uncertainty-based approaches focuses on samples with large uncertainty, most of which are near to the classification boundaries, but fail to capture the data distribution. CORESET in Fig. 3(d) nearly uniformly selects samples according to the data distribution, but can not well capture the classification boundaries in the lower left corner and the upper right corner. Compared to CORESET, samples selected by BADGE in Fig. 3(e) are more concentrated. It can capture the classification boundaries in the lower left corner, but fails in the upper right corner. Meanwhile, according to Fig. 3(f), DAMI can best capture different linear separable regions and different classification boundaries.

Experiments on Tabular Datasets

To evaluate the performances of DAMI on tabular data, we use four tabular datasets: **Employee**[1], **Telescope**[2], **Default**[3] and **NewsPopularity**[4]. There are 32771, 19020, 30000 and 39797 samples respectively in these datasets, and 9, 11, 24 and 61 input features respectively in these datasets. We use

---

[1] www.kaggle.com/c/amazon-employee-access-challenge/data

[2] archive.ics.uci.edu/ml/datasets/magic+gamma+telescope

[3] archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

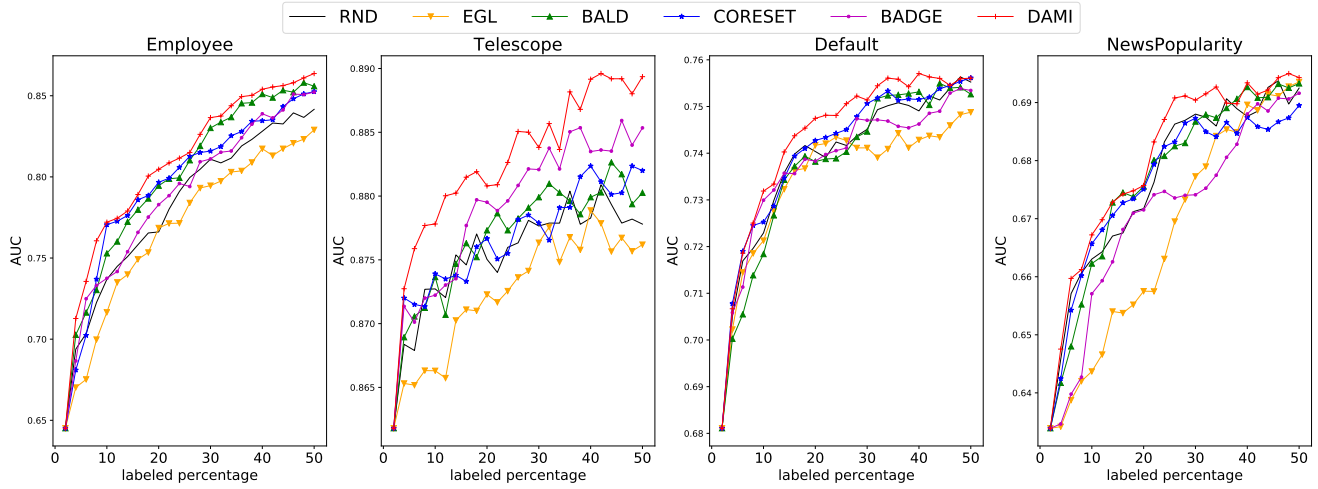[4] archive.ics.uci.edu/ml/datasets/online+news+popularity

Figure 4: Performance comparison with different ratios of labeled samples on tabular datasets.
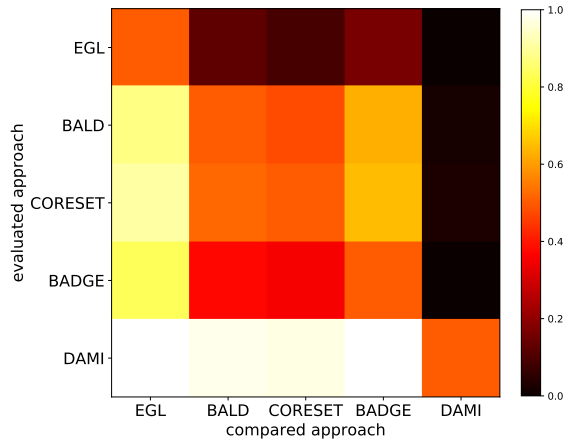

Figure 5: Pairwise comparison matrix over all experiments on tabular datasets. Each square represents the ratio that the corresponding evaluated approach outperforms the corresponding compared approach.

70%, 10% and 20% samples in each dataset for training, validation and testing respectively. Same compared approaches are involved as in the toy experiments, i.e., **EGL** (Huang et al. 2016), **BALD** (Gal, Islam, and Ghahramani 2017), **CORESET** (Sener and Savarese 2018), **BADGE** (Ash et al. 2020) and **DAMI**. We also involve **RND**, which randomly selects samples in each iteration.

We run 3 layers of MLP on samples in each dataset, where then hidden units are set as $(16, 8)$. Considering these datasets are class-imbalanced, we use AUC (Area Under Curve) as the evaluation metric. We use 2% samples in the training set as initial labeled samples. Then, we label 2% samples in the training set during each iteration of sample selection, until 50% samples in the training set are covered. We run each approach 10 times, and report the average experimental results.

Fig. 4 shows the performance comparison among RND, EGL, BALD, CORESET, BADGE and DAMI with different ratios of labeled samples. In most cases, active learning approaches can outperform the random selection, which demonstrates the necessity of deep active learning. We can observe that, EGL performs poor, and is even outperformed by RND. This may indicate that, the uncertainty evaluation based the norms of gradients is not stable. On the Employee, Telescope and Default datasets, BALD, CORESET and BADGE have close performances, and each of them achieves the best performance among the five baseline methods on different datasets. Meanwhile, BADGE performs poor on the NewsPopularity dataset. Moreover, it is clear that, DAMI has best performances on the four tabular datasets, and can constantly outperform other baseline approaches. Specifically, in the middle parts of the curves, i.e., labeled percentage in the range of $[15\%, 35\%]$, DAMI usually has great advantages.

We also illustrate the pairwise comparison matrix over all experiments on tabular datasets in Fig. 5. Each square in the figure indicates the ratio that the corresponding evaluated approach outperforms the corresponding compared approach. Brighter the square, large the advantage of the evaluated approach on the compared approach. It is clear that, EGL has the worst performances. BALD and CORESET are the two best baseline approaches on tabular datasets. Moreover, DAMI can clearly outperform other approaches. These results strongly demonstrate the effectiveness of our proposed DAMI approach on tabular datasets.

Experiments on Textual Datasets

To evaluate the performances of DAMI on textual data, we use two sentence classification datasets: **subj** (Pang and Lee 2004) and **MR** (Pang and Lee 2005). In subj, there are 5000 positive samples and 5000 negative samples. In MR, there are 5331 positive samples and 5331 negative samples. We use 70%, 10% and 20% samples in each dataset for training, validation and testing respectively. As in the tabular ex-
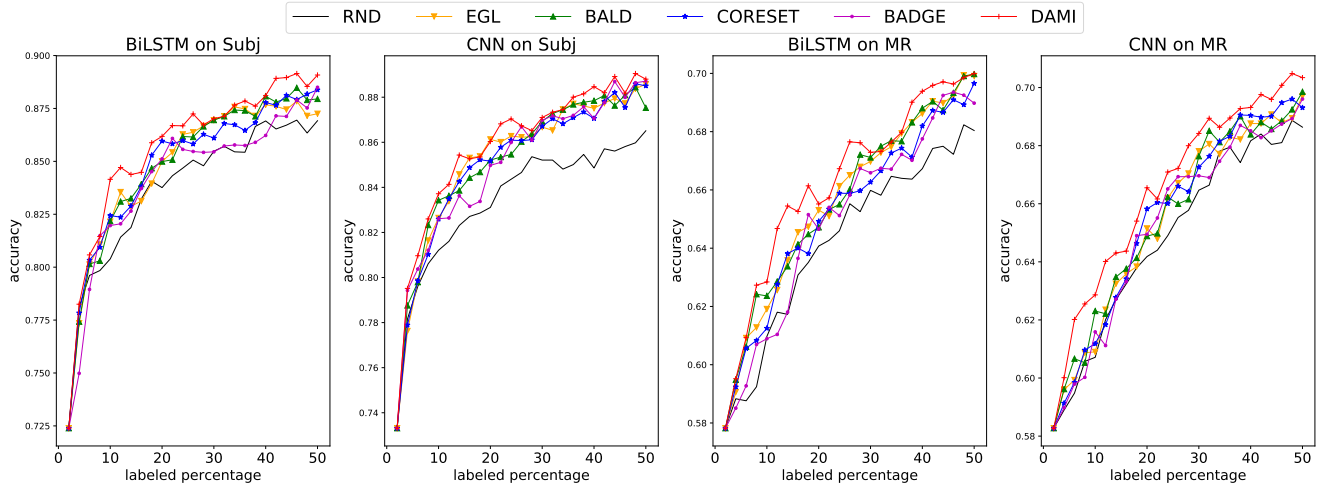
Figure 6: Performance comparison with different ratios of labeled samples on textual datasets.
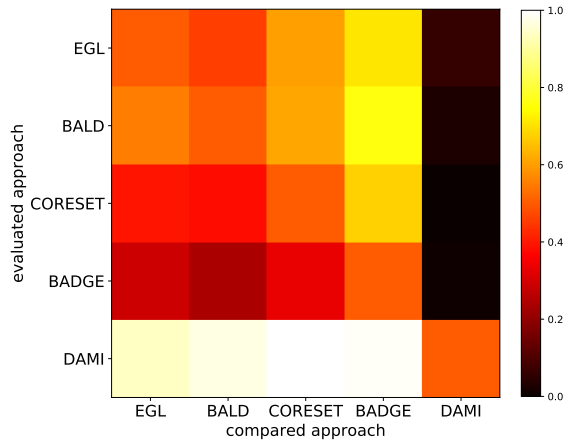

Figure 7: Pairwise comparison matrix over all experiments on textual datasets. Each square represents the ratio that the corresponding evaluated approach outperforms the corresponding compared approach.

periments, we have the same compared approaches, except we replace the conventional EGL (Huang et al. 2016) with **EGL-word** (Zhang, Lease, and Wallace 2017). For simplification, we still name EGL-word as EGL in our results.

We involve two language models: BiLSTM and CNN. We train word2vec[5] on each dataset for the initialization of word embeddings, whose dimensionality is 100. For the implementation of BiLSTM, we have a single layer and 100 hidden units. For the implementation of CNN, we have filter sizes $(3, 4, 5)$ and 100 feature maps. Considering subj and MR are class-balanced, we use accuracy as the evaluation metric. We use 2% samples in the training set as initial labeled samples. Then, we label 2% samples in the training set during each iteration of sample selection, until 50% samples

in the training set are covered. We run each approach 10 times, and report the average experimental results.

As in the tabular experiments, we illustrate the performance comparison and pairwise comparison matrix on textual datasets in Fig. 6 and 7 respectively. This time, EGL performs well, and is one of the two best baseline approaches, where the other one is BALD. This indicates that, EGL-word, which selects the most informative word in a sentence, is clearly a better active learning approach for modeling textual data compared to the conventional EGL. And uncertainty-based approaches are shown to be effective for modeling textual data. Meanwhile, BADGE performs relatively poor. And clearly, these experimental results show that, DAMI is the best deep active learning approach for the sentence classification task.

Conclusion

In this paper, inspired by the local piece-wise interpretability of DNN, we introduce the linear separable regions of samples to the problem of active learning. Accordingly, we propose a novel DAMI approach, which selects and labels samples on different linear separable regions for optimally training DNN. We mainly focus on two scenarios: 1) MLP for classification on tabular data; 2) language models for classification on textual data. For tabular data, we use the local piece-wise interpretation in DNN as the representation of each sample, and directly run K-mediods clustering to select and label the central sample in each cluster. For textual data, we propose a novel aggregator to find the most informative word in each sentence, and use its local piece-wise interpretation as the representation of the sentence. Extensive experiments on both tabular and textual datasets demonstrate the effectiveness of our proposed DAMI approach.

References

Alvarez-Melis, D., and Jaakkola, T. S. 2018. Towards robust interpretability with self-explaining neural networks. In *NeurIPS*.

---

[5]code.google.com/archive/p/word2vec/

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*.

Chu, L.; Hu, X.; Hu, J.; Wang, L.; and Pei, J. 2018. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution. In *KDD*.

Dong, Y.; Su, H.; Zhu, J.; and Bao, F. 2018. Towards interpretable deep neural networks by leveraging adversarial examples. In *CVPR*.

Ducoffe, M., and Precioso, F. 2018. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*.

Fong, R. C., and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *CVPR*.

Gal, Y., and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *ICML*.

Ghaeini, R.; Fern, X. Z.; and Tadepalli, P. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *EMNLP*.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep sparse rectifier neural networks. In *AISTATS*.

Goodfellow, I. J.; Warde-Farley, D.; Mirza, M.; Courville, A.; and Bengio, Y. 2013. Maxout networks. *arXiv preprint arXiv:1302.4389*.

Guan, C.; Wang, X.; Zhang, Q.; Chen, R.; He, D.; and Xie, X. 2019. Towards a deep and unified understanding of deep neural models in nlp. In *ICML*.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51(5).

Harvey, N.; Liaw, C.; and Mehrabian, A. 2017. Nearly-tight vc-dimension bounds for piecewise linear neural networks. In *COLT*.

Houlsby, N.; Huszár, F.; Ghahramani, Z.; and Lengyel, M. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Hsu, W.-N., and Lin, H.-T. 2015. Active learning by learning. In *AAAI*.

Huang, J.; Child, R.; Rao, V.; Liu, H.; Satheesh, S.; and Coates, A. 2016. Active learning for speech recognition: the power of gradients. *arXiv preprint arXiv:1612.03226*.

Huang, S.-J.; Jin, R.; and Zhou, Z.-H. 2010. Active learning by querying informative and representative examples. In *NIPS*.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. In *ICML*.

Li, J.; Chen, X.; Hovy, E.; and Jurafsky, D. 2015. Visualizing and understanding neural models in nlp. In *ACL*.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *NIPS*.

Montufar, G. F.; Pascanu, R.; Cho, K.; and Bengio, Y. 2014. On the number of linear regions of deep neural networks. In *NIPS*.

Nair, V., and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*.

Pang, B., and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*.

Park, H.-S., and Jun, C.-H. 2009. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications* 36(2).

Prabhu, A.; Dognin, C.; and Singh, M. 2019. Sampling bias in deep active classification: An empirical study. In *EMNLP*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "why should i trust you?" explaining the predictions of any classifier. In *KDD*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*.

Sener, O., and Savarese, S. 2018. Active learning for convolutional neural networks: A core-set approach. In *ICLR*.

Settles, B., and Craven, M. 2008. An analysis of active learning strategies for sequence labeling tasks. In *EMNLP*.

Siddhant, A., and Lipton, Z. C. 2018. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. In *EMNLP*.

Smilkov, D.; Thorat, N.; Kim, B.; Vigas, F.; and Wattenberg, M. 2017. Smoothgrad: removing noise by adding noise. In *ICML*.

Tong, S., and Koller, D. 2001. Support vector machine active learning with applications to text classification. *JMLR* 2(Nov).

Wang, Z., and Ye, J. 2015. Querying discriminative and representative samples for batch mode active learning. *ACM TKDD* 9(3).

Wang, J.; Liu, Q.; Liu, Z.; and Wu, S. 2019. Towards accurate and interpretable sequential prediction: A cnn & attention-based feature extractor. In *CIKM*.

Wu, Y.; Xu, Y.; Singh, A.; Yang, Y.; and Dubrawski, A. 2019. Active learning for graph neural networks via node feature propagation. *arXiv preprint arXiv:1910.07567*.

Yan, Y.-F.; Huang, S.-J.; Chen, S.; Liao, M.; and Xu, J. 2020. Active learning with query generation for cost-effective text classification. In *AAAI*.

Yuan, H.; Chen, Y.; Hu, X.; and Ji, S. 2019. Interpreting deep models for text analysis via optimization and regularization methods. In *AAAI*.

Zhang, Y.; Lease, M.; and Wallace, B. C. 2017. Active discriminative text representation learning. In *AAAI*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*.

Zhu, J.; Chen, J.; Hu, W.; and Zhang, B. 2017. Big learning with bayesian methods. *National Science Review* 4(4).