Multifaceted Uncertainty Estimation for Label-Efficient Deep Learning

Weishi Shi[1] **Xujiang Zhao**[2] **Feng Chen**[2] **Qi Yu**[1*]
Rochester Institute of Technology[1] University of Texas at Dallas[2]
{ws7586, qi.yu}@rit.edu[1] {xujiang.zhao, feng.chen}@utdallas.edu[2]

Abstract

We present a novel multi-source uncertainty prediction approach that enables deep learning (DL) models to be actively trained with much less labeled data. By leveraging the second-order uncertainty representation provided by subjective logic (SL), we conduct evidence-based theoretical analysis and formally decompose the predicted entropy over multiple classes into two distinct sources of uncertainty: vacuity and dissonance, caused by lack of evidence and conflict of strong evidence, respectively. The evidence based entropy decomposition provides deeper insights on the nature of uncertainty, which can help effectively explore a large and high-dimensional unlabeled data space. We develop a novel loss function that augments DL based evidence prediction with uncertainty anchor sample identification. The accurately estimated multiple sources of uncertainty are systematically integrated and dynamically balanced using a data sampling function for label-efficient active deep learning (ADL). Experiments conducted over both synthetic and real data and comparison with competitive AL methods demonstrate the effectiveness of the proposed ADL model.

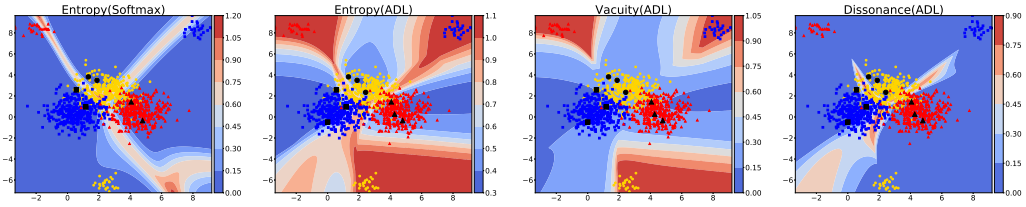
1 Introduction

Deep learning (DL) models establish dominating status among other types of supervised learning models by achieving the state-of-the-art performance in various application domains. However, such an advantage only emerges when a huge amount of labeled training data is available. This limitation slows down the pace of DL, especially when being applied to knowledge-rich domains, such as medicine and biology, where large-scale labeled samples are too expensive to obtain from well-trained experts. Meanwhile, active learning (AL) has demonstrated great success by showing that for many supervised models, training samples are not equally important in terms of improving the model performance [1]. As a result, a carefully selected smaller training set can achieve a model equally well or even better than a randomly selected large training set.

An interesting question arises, which is whether DL models can be actively trained using much less labeled data. Recent efforts show promising results in this direction through Bayesian modeling [2] and batch model sampling [3]. However, as DL models are commonly applied to high dimensional data such as images and videos, a fundamental challenge still remains, which is how to most effectively explore the exponentially growing feature space to select the most useful data samples for active model training. Existing AL models usually leverage the model provided information, such as estimated decision boundaries or predicted entropy for data sampling. However, the deep structure and the large number of parameters of DL models make model overfitting almost inevitable especially in the early stage of AL when only very limited training data is available. As a result, the model may provide misleading information that makes data sampling from a high-dimensional search space even more difficult. Besides a high dimensionality, complex data may contain a large number of classes

*Corresponding author

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.



(a) (b) (c) (d)

Figure 1: A dataset consists of three mixtures of Gaussian's (shown in red, blue, and yellow), each of which has a large and small clusters of data samples. (a) Softmax predicted entropy; (b) ADL predicted entropy; (c) ADL predicted vacuity; (d) ADL predicted dissonance.

and data samples from certain classes may be completely missing. Such situations are quite common for domains, such as scientific discovery (e.g., gene function prediction) and anomaly detection. AL models should be able to effectively discover these out of distribution (OOD) samples for labeling in order to achieve an overall good prediction performance.

Uncertainty sampling has been one of the most commonly used pool-based AL models. In particular, a model chooses the data sample that it is least certain about. Thus, once the sample is labeled, model uncertainty can be significantly reduced. As an information-theoretic measure, entropy provides a general criterion for uncertainty sampling. Some commonly used sampling methods, including least confident and margin based strategies, are equivalent to entropy-based sampling in binary classification [1]. It is also straightforward to generalize to multi-class problems.

A key challenge of entropy-based sampling for AL is that the predicted entropy may be highly inaccurate, especially in the early state of the AL. Such an issue may become more severe when training a neural network (NN)/DL active learner due to model overfitting as described above. Figure 1(a) shows the predicted entropy by an NN active learner trained using nine labeled data samples, which are in black color and evenly distributed in three classes. The standard *softmax* layer is used in the output layer to generate class probabilities over three classes, each of which is a mixture of two Gaussian's. It turns out that all the data samples in the three small clusters located in the top left, top right, and bottom center, are wrongly predicted with high confidence, as indicated by the low entropy. As a result, data samples from these three clusters are less likely to be selected for labeling. In contrast, the data samples that are close to the center of the three major clusters are more likely to be selected. However, labeling these samples will have the effect of fine-tuning a wrongly predicted decision boundary, leading to a much higher (but less effective) labeling cost.

Figure 1(b) shows the result from the proposed active deep learning (ADL) model. While the samples from the small clusters are still wrongly predicted due to lack of training data, they are predicted with a much lower confidence as indicated by the high entropy. However, even with a more accurately predicted entropy, the active learner may still sample from the center of the three major clusters as it is also assigned a high entropy along with the areas that cover the three smaller clusters. By leveraging the evidenced based probabilistic constructs developed under the subjective logic (SL) framework [4], we formally *decompose entropy into two distinct sources of uncertainty: vacuity and dissonance*, which are caused by lack of evidence and conflict of strong evidence, respectively. When putting the vacuity and dissonance as shown in Figures 1(c) and (d) together, it is interesting to see that we recover the entropy as shown in Figure 1(b), which empirically verifies our theoretical results. Entropy decomposition provides further insights on the sources on uncertainty, which is instrumental to guide the data sampling process. Intuitively, given the dataset in Figure 1, an effective sampling strategy will first choose samples from the three small clusters according to vacuity in the early stage of AL to properly establish the shape of the decision boundary. It can then fine-tune the decision boundary by sampling according to dissonance. Such an uncertainty-aware sampling strategy will be critical for a high-dimensional space with multiple competing classes where data samples are scarcely distributed and the decision boundary becomes more complicated.

Our major contribution is threefold: (1) theoretical decomposition of entropy into evidence-based second-order uncertainties, including belief vacuity and belief dissonance; (2) a multi-source uncertainty prediction model that accurately quantifies different sources of uncertainty; (3) an active deep learning model that systematically integrates different types of uncertainty for effective data sampling in a high-dimensional space. Extensive experiments are conducted over both synthetic and real-world data to demonstrate the effectiveness of the proposed ADL model.

2 Related Work

Uncertainty Quantification in Belief/Evidence Theory: In belief/evidence theory, uncertainty reasoning has been substantially explored through Fuzzy Logic [5], Dempster-Shafer Theory (DST) [6], and Subjective Logic (SL) [4]. Unlike the efforts made in ML/DL, belief theorists focused on reasoning of inherent uncertainty in information resulting from unreliable, incomplete, deceptive, and/or conflicting evidence. SL considered uncertainty in subjective opinions in terms of *vacuity* (i.e., lack of evidence) and *vagueness* (i.e., failure of discriminating a belief state) [4]. Vacuity has been used as an effective vehicle to detect OOD queries through evidence learning, achieved under the typical DL setting with ample training samples [7]. Recently, other dimensions of uncertainty have been studied, such as *dissonance* (due to conflicting evidence) and *consonance* (due to evidence about composite subsets of state values) [8].

Uncertainty in Deep Learning: In DL, *aleatoric* uncertainty (AU) and *epistemic* uncertainty (EU) have been studies using Bayesian Neural Networks (BNNs) for computer vision. AU consists of homoscedastic uncertainty (i.e., constant errors for different inputs) and heteroscedastic uncertainty (i.e., different errors for different inputs) [9]. A Bayesian DL (BDL) framework was presented to estimate both AU and EU simultaneously in regression (e.g., depth regression) and classification settings (e.g., semantic segmentation) [10]. A new type of uncertainty, called *distributional uncertainty*, is defined based on distributional mismatch between the test and training data distributions [11]. Other than exploring the new sources of uncertainty, recent works also focus on better estimating the well-known first-order uncertainty, predictive entropy, in DL models through calibration [12] or ensemble [13] methods. Even though the recent efforts offer abundant uncertainty measurements for DL, how to leverage these uncertainty information for better active sampling remains sparse. For example, while distributional uncertainty can be used for data sampling in AL, the prior network in [11] needs to be properly trained as its parameter must encapsulate knowledge of both in-domain distribution and the decision boundary, making it not very suitable for AL. This is also evidenced by our experimental results on real-world data. The Noise-Contrastive Priors can also be used to support better exploration in data sampling as it encourages high uncertainty near the boundary of the training data [14]. However, in the initial phase of AL when the training data is very limited, this measure can be insufficient to explore data samples faraway from the training data.

Active Learning in Deep Learning: The common AL methods other than DL-based ones are surveyed in [1]. There are limited efforts on actively training DL models for high-dimensional data with a few exceptions. In [15], an AL model was developed for DL using three metrics for data sampling: least confidence, margin sampling, and entropy. A new approach combines recent advances in BDL into the AL framework to achieve label-efficient DL training [2]. Another approach advances the AL development by introducing a cost-effective strategy to automatically select and annotate the high-confidence samples, which improves the traditional sample selection strategies [16]. Data sampling in DL has also been approached as core-set selection problem [3] or disparate and high magnitude selection in a hallucinated gradient space [17], which require a large batch to work well.

Different from all existing works, the proposed ADL model decomposes the accurately estimated uncertainty into two second-order uncertainties: vacuity and dissonance, and dynamically balances multi-source uncertainty to achieve active training of DL models with much less labeled data.

3 Evidence-Aware Entropy Decomposition

As discussed earlier, a high entropy may be contributed by difference sources of uncertainty with distinct characteristics. In this section, we conduct a fine-grained theoretical analysis of different types of uncertainty that arise in the context of multi-class problems. The decomposition is conducted under the SL framework, which provides key building blocks for our theoretical analysis.

3.1 Theory of Subjective Logic

Overview. SL is an uncertain probabilistic logic that is built upon probabilistic logic (PL) [18] and belief theory (BT) [19] while making two unique extensions. First, SL explicitly represents uncertainty by introducing vacuity of evidence (or uncertainty mass) in its opinion representation, which addresses the limitation of using PL to model lack of confidence in probabilities. Second, SL extends the traditional belief function of BT by incorporating base rates, which serve as the prior probability in Bayesian theory. The Bayesian nature of SL allows it to use second-order uncertainty to express and reason the uncertainty mass, where second-order uncertainty is represented in terms of

a probability density function (PDF) over first-order probabilities [4]. In particular, for multi-class problems, we use a multinomial distribution (first-order uncertainty) to model class probabilities and use a Dirichlet PDF (second-order uncertainty) to model the distribution of class probabilities. Second-order uncertainty enriches uncertainty representation with evidence information, which plays a central role in entropy decomposition as detailed later.

Subjective opinions (or opinions) are the arguments in SL. In the multi-class setting, the subjective opinion of a multinomial random variable y in domain $\mathbb{Y} = \{1, ..., K\}$ is given by a triplet

$$\omega = (\boldsymbol{b}, u, \boldsymbol{a}), \text{ with } \sum_{k=1}^{K} b_k + u = 1 \tag{1}$$

where $\boldsymbol{b} = (b_1, ..., b_K)^T$, u, and $\boldsymbol{a} = (a_1, ..., a_K)^T$ denote the belief mass over $\mathbb{Y}$, uncertainty mass representing vacuity of evidence, and base rate distribution over $\mathbb{Y}$, respectively, and $\forall k, a_k \geq 0, b_k \geq 0, u \geq 0$. The probability that y is assigned to class k is given by

$$P(\mathrm{y} = k) = b_k + a_k u, \quad \forall k \in \mathbb{Y} \tag{2}$$

which combines the belief mass with the uncertain mass using the base rates. In the multi-class setting, a_k can be regarded as the prior preference over class k. When no preference is given, we assign all the base rates as $1/K$.

Transition between first- and second-order uncertainties. In existing SL literature, there *lacks a clear transition* between the first-order uncertainty given in equation 2 and the second-order uncertainty expressed as a Dirichlet PDF. Here, we make this transition *explicit* by introducing a set of random variables $\mathbf{p} = (p_1, ..., p_K)^T$, distributed on a $(K-1)$-dimensional simplex. We introduce a conditional distribution $P(\mathrm{y} = k|\mathrm{p}_k) = \mathrm{p}_k$, which allows us to represent the marginal distribution in equation 2 by $P(\mathrm{y}) = \int P(\mathrm{y}|\mathbf{p})p(\mathbf{p})d\mathbf{p}$. We define $p(\mathbf{p})$ as a Dirichlet PDF over $\mathbf{p}$: $\mathrm{Dir}(\mathbf{p}|\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)^T$ is K-dimensional strength vector, with $\alpha_k \geq 0$ denoting the effective number of observations of class k. SL explicitly introduces the uncertainty evidence through a non-informative weight W and redefine the strength parameter as

$$\alpha_k = r_k + a_k W, \text{ with } r_k \geq 0, \forall k \in \mathbb{Y} \tag{3}$$

where r_k is the amount of evidence (or the number of observations) to support the k-th class and W is usually set to K, i.e., the number of classes. Given the new definition of the strength parameter, the expectation of the class probabilities $\mathbf{p} = (p_1, ..., p_K)^T$ is given by

$$\mathbb{E}[p_k] = \frac{\alpha_k}{\sum_{j=1}^{K} \alpha_j} = \frac{r_k + a_k W}{\sum_{j=1}^{K} r_j + W} \tag{4}$$

where $a_k = 1/K$. Marginalizing $\mathbf{p}$ leads to an evidence-based expression of belief mass and uncertainty mass (existing SL literature [4] arrives at the following definitions using a set of mapping rules that are not strictly defined using probability terms as in our approach):

$$b_k = \frac{r_k}{S} \quad \forall k \in \mathbb{Y}, \quad u = \frac{W}{S}, \text{ with } S = \sum_{k=1}^{K} \alpha_k \tag{5}$$

Second-order (evidence-based) uncertainties. SL categorizes uncertainty into two primary sources [4, 8]: (1) basic belief uncertainty that results from specific aspects of belief mass in isolation and (2) intra-belief uncertainty that results from the relationships between belief masses and uncertainty mass. Since we focus on the multi-class setting, no composite values (i.e., simultaneously assigned to multiple classes) are allowed. As a result, these two sources of evidence-based uncertainty boil down to *vacuity* and *dissonance*, respectively, that correspond to vacuous belief and contradicting beliefs. In particular, vacuity of an opinion $vac(\omega)$ is captured by uncertainty mass u, which is defined in equation 5 and dissonance of an opinion is defined as

$$diss(\omega) = \sum_{k=1}^{K} \left(\frac{b_k \sum_{j \neq k} b_j \mathrm{Bal}(b_j, b_k)}{\sum_{j \neq k} b_j} \right), \mathrm{Bal}(b_j, b_k) = \begin{cases} 1 - \frac{|b_j - b_k|}{b_j + b_k} & \text{if } b_i b_j \neq 0 \\ 0 & \text{if } \min(b_i, b_j) = 0 \end{cases} \tag{6}$$

where $\mathrm{Bal}(b_j, b_k)$ is the relative mass balance function between two belief masses. The belief dissonance of an opinion is measured based on how much belief supports individual classes. Consider a binary classification example with a binomial opinion given by $(b_1, b_2, u, \boldsymbol{a}) = (0.49, 0.49, 0.02, \boldsymbol{a})$. Based on equation 6, it has a dissonance value of 0.98. In this case, although the vacuity is close to zero, a high dissonance indicates that one cannot make a clear decision because both two classes have the same amount of supporting evidence and hence strongly conflict with each other.

3.2 Evidence-Based Entropy Decomposition

By leveraging the second-order uncertainty representation, we formally show that the entropy of a predicted class distribution $P(\mathrm{y})$ can be decomposed into vacuity and dissonance. Our major theoretical results indicate that the uncertainty of a high-entropy data sample may be caused by either lack of evidence (i.e., high vacuity) or conflict of strong evidence (i.e., high dissonance) but not both. By clearly identifying the sources of uncertainty instead of using them in a combined form as in entropy, the evidence based decomposition of entropy provides deeper insights on the nature of uncertainty, which provides important guidance for an AL model to more effectively explore a large and high-dimensional search space for efficient data sampling.

Lemma 1. Dissonance maximization. *Given a total Dirichlet strength $S = CK$, where $C \geq 1$ and K is the number of classes, for any opinion ω on a multinomial random variable* y*, we have*

$$\max diss(\omega) = 1 - \frac{1}{C} \tag{7}$$

Corollary 1. *The dissonance $diss(\omega)$ is approaching (but not reaching) 1 when all the evidence r_k's are set to equal and $S \to \infty$; it reaches 0 when $S = K$:*

$$\begin{cases} \lim_{S\to\infty} diss(\omega) = 1 & \text{if } r_1 = ...r_k... = r_K \\ diss(\omega) = 0 & \text{if } S = K \end{cases} \tag{8}$$

Lemma 2. Vacuity maximization. *For any opinion ω on a multinomial random variable* y*, we have $0 \leq vac(\omega) \leq 1$ and the maximum vacuity is achieved when $\sum_{k=1}^{K} r_k = 0$.*

Theorem 1. *Let* y *denote a multinomial random variable, ω_{y} denote its opinion, S denote its total Dirichlet strength, and $H[\mathrm{y}]$ be the entropy of* y*. $H[\mathrm{y}]$ can be maximized under two different and non-overlapping conditions: (**1**) $\mathrm{y}^* = \arg\max H[\mathrm{y}] \Leftrightarrow \mathrm{y}^* = \arg\max vac(\omega_{\mathrm{y}})$ if $S = K$ and $a_k = 1/K, \forall k$; (**2**) $\mathrm{y}^* = \arg\max H[\mathrm{y}] \Leftrightarrow \mathrm{y}^* = \arg\max diss(\omega_{\mathrm{y}})$, if $C \to \infty$.*

Details proofs are provided in Appendix A. A more intuitive interpretation of the main results in Theorem 1 is as following. A high-entropy data sample supported by a strong evidence (i.e., $S \gg K$) is caused by a high dissonance (i.e., conflict of evidence); a high-entropy data sample supported by little evidence (i.e., $S \approx K$) is caused by a high vacuity (i.e., lack of evidence). Through the second-order uncertainty representation, we offer an evidence based interpretation of entropy that allows us to identify two different sources of uncertainty that both cause a high entropy. The multi-source uncertainty will provide important information to design a fine-grained sampling function for AL, which will be detailed in next section.

4 Multi-source Uncertainty Aware Active Deep Learning

In order to best use the uncertainty information, the ADL model should first be able to provide an accurate uncertainty estimation based on very limited training data. This, coupled with the large number of parameters of the DL model, poses a fundamental challenge due to a higher risk of model overfitting. As shown earlier, inaccurate uncertainty estimation will cause the model to miss labeling important data samples that can help accurately detect the decision boundary if labeled.

In addition, since both vacuity and dissonance are derived from second-order uncertainty, solely predicting the class label or its distribution does not provide sufficient information for multi-source uncertainty prediction. Instead, the proposed ADL model directly estimates the supporting evidence (i.e., r_k's) for each class, which is a central element that can be used to quantify belief mass and uncertainty mass according to equation 5. To better address overfitting, we develop a novel loss function that augments DL based evidence prediction with uncertainty anchor sample identification. These anchor samples are unlabeled data that inform the ADL which areas of the data space are important but less explored. Optimizing this loss function will ensure that ADL predicts high vacuity over these areas. Finally, we introduce our novel sampling function that systematically integrates accurately estimated multi-source uncertainty for active deep learning.

4.1 Uncertainty Anchor Sample Identification

The desired anchor samples should meet two important criteria: (i) indicative of the less explored areas in the data space, and (ii) representative of a group of unlabeled samples instead of being an isolated outlier. These properties ensure ADL to effectively prioritize the sampling order

over these areas and iteratively visit them based on data density. It is also important that anchor samples can be identified efficiently to support fast data sampling for real-time AL. To this end, we define *relative similarity* for efficient anchor sample selection: $\mathbb{A}^* = \{\boldsymbol{x}|\text{RS}(\boldsymbol{x}) \geq \lambda\}$, with $\text{RS}(\boldsymbol{x}) = \sum_{\boldsymbol{x}_n \in \mathbb{X}_u} \text{sim}(f(\boldsymbol{x}), f(\boldsymbol{x}_n)) / \sum_{\boldsymbol{x}_n \in \mathbb{X}_t} \text{sim}(f(\boldsymbol{x}), f(\boldsymbol{x}_n))$, where $\mathbb{X}_u$ and $\mathbb{X}_t$ denote the sets of unlabeled candidate and training samples, respectively; f is an embedding function that maps a high-dimensional feature vector $\boldsymbol{x}$ to a low-dimensional embedding, which can be achieved through a neural network (e.g., a convolutional autoencoder for image data or word embedding for text data); and $\text{sim}(\cdot, \cdot)$ is a similarity function defined in the low-dimensional embedding space (a RBF kernel is used in our experiments). Intuitively, the numerator ensures that the selected area has abundant candidate unlabeled data points to sample so that it has lower risk of containing isolated noise. The denominator makes sure the selected region is located OOD with respect to the current training data.

4.2 Multi-Source Uncertainty Prediction

The set of uncertainty anchor samples $\mathbb{A}^*$ represents areas in the data space that are cohesively distributed far away from the current training data. As these data are essentially OOD with respect to the current training data, their predicted vacuity should be high, which implies low predicted evidence due to Lemma 2. We leverage this information by constructing an evidence strength loss, $\mathcal{L}_{Evi}^{(u)}$, which forces the model to predict low evidence for $\boldsymbol{x}_u \in \mathbb{A}^*$:

$$\mathcal{L}_{Evi}^{(u)}(\mathbb{A}^*, \Theta) = \mathbf{1}_{(\mathrm{x_u} \in \mathbb{A}^*)}^T f(\boldsymbol{x}_i|\Theta) \tag{9}$$

where $\mathbf{1}_{(\mathrm{C})} = \mathbf{1}$ if C is true and 0 otherwise; $\boldsymbol{r}_i = f(\boldsymbol{x}_u|\Theta)$ is the output of the DL model, representing the predicted supporting evidence of $\boldsymbol{x}_u$, and Θ is the set of DL model parameters. Since we require $r_k \geq 0$, an activation layer (i.e., ReLu) is used to replace the softmax layer as commonly used in other NN classifiers. The evidence strength loss is the key component of our proposed overall loss function. Samples in $\mathbb{A}^*$ act as anchors to provide the model a preview of certain areas that out of its current knowledge. The model is guided to put less belief mass on those areas, leading to more accurate uncertainty estimation and eventually benefit the multi-source uncertainty based data sampling. Furthermore, since the activation layer is used for model output, equation 9 essentially performs l_1 regularization to last hidden layer's weight matrix and bias vector. We want to emphasize that our approach demands no additional labeling cost. The anchor samples are dynamically detected according to the current training and put into use without their actual label being known.

We proceed to define our overall loss function. For training sample $\boldsymbol{x}_i$, let $\boldsymbol{y}_i$ encode the ground-true class label k by setting $y_{ik} = 1$ and $y_{ij} = 0, \forall j \neq k$. Let $\text{Cat}(\hat{\mathrm{y}}_i = k|\mathbf{p}_i(\Theta))$ be the likelihoood, where $\mathbf{p}_i(\Theta) \sim \text{Dir}(\mathbf{p}_i|\boldsymbol{\alpha}_i(\Theta))$ and $\boldsymbol{\alpha}_i(\Theta) = f(\boldsymbol{x}_i|\Theta) + W\boldsymbol{a}_i$. We set the non-informative weight $W = K$ and base rates $a_{ik} = 1/K, \forall k$. The expected sum of squares loss is defined as

$$\mathcal{L}^{(i)}(\Theta) = \mathbb{E}_{\mathbf{p}_i \sim \text{Dir}(\mathbf{p}_i(\Theta)|\boldsymbol{\alpha}_i(\Theta))}||\boldsymbol{y}_i - \mathbf{p}_i||_2^2 = \sum_{j=1}^{K}(y_{ij}^2 - 2y_{ij}\mathbb{E}[\mathrm{p}_{ij}(\Theta)] + \mathbb{E}[\mathrm{p}_{ij}(\Theta)^2]) \tag{10}$$

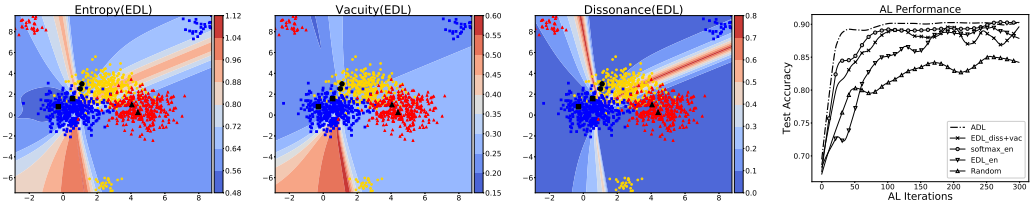
Minimizing $\mathcal{L}^{(i)}(\Theta)$ has the effect of jointly minimizing the prediction error and the variance of $\mathbf{p}_i$ [7], hence reducing the uncertainty. This can be seen using identity $\mathbb{E}[\mathrm{p}_{ij}(\Theta)^2] = \mathbb{E}[\mathrm{p}_{ij}(\Theta)]^2 + \text{Var}(\mathrm{p}_{ij}(\Theta))$ and rearranging the terms on the r.h.s. of equation 10. Our overall loss function is:

$$\sum_{\boldsymbol{x}_i \in \mathbb{X}_t} \mathcal{L}^{(i)}(\Theta) + \lambda_1 \sum_{\boldsymbol{x}_u \in \mathbb{X}_u} \mathcal{L}_{Evi}^{(u)} + \lambda_2 \mathcal{L}_2(\Theta) \tag{11}$$

where $\mathcal{L}_2(\Theta)$ is the L_2 regularizer of network parameters.

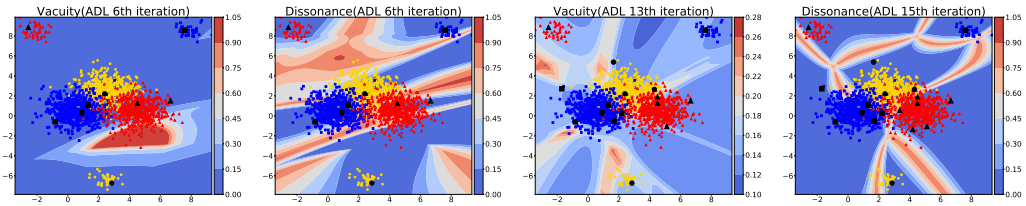
4.3 Data Sampling for Active Deep Learning

According to Lemma 2, a data sample's vacuity is maximized when the model assigns zero evidence to all K classes. This indicates the model has never seen a similar data sample from training. Annotating samples with a large vacuity can help the ADL gain most new knowledge of the data space. It has the effect of guiding the model to explore the most important areas, which is especially critical for a high-dimensional data space. In AL, the true decision boundary can be easily skewed due to limited initial training. The vacuity-aware search helps the model fast converge to the true decision boundary without excessively sampling around the wrong one. Moreover, it is also effective to discover new types of classes whose instances have never been exposed to the model, as shown in our experiments. According to Lemma 1, a data sample's dissonance is maximized when the model assigns equally



(a) (b) (c) (d)

Figure 2: Uncertainty prediction results from EDL: (a) Entropy, (b) Vacuity, and (c) Dissonance; (d) AL performance comparison



(a) (b) (c) (d)

Figure 3: (a) Vacuity and (b) Dissonance of iteration 6 when ADL first discovers all hidden OOD areas; (c) Vacuity of iteration 13 when ADL starts penalizing vacuity in the sampling function; (d) Dissonance after two iterations of penalizing vacuity.

high (close to infinity) evidence to all K classes. These strong conflicting evidence received from different classes indicate that the data sample is located near the decision boundary where multiple classes are heavily overlapped. Annotating samples with high dissonance helps the model further fine-tune the decision boundary, leading to better discriminative power.

We design a sample function that best leverages these two important and complementary sources of uncertainty to most effectively guide ADL. Intuitively, we would like ADL to rely more on vacuity in the early phase of AL, which can most effectively shape the decision boundary and avoid fine-tuning the wrong decision areas. As AL goes, dissonance should gradually gain a higher weight, which allows ADL to further fine-tune the decision boundary that has the right shape but is less accurate, aiming to maximize the discriminate power of the model. The sample function is given:

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}\in\mathbb{X}_u} [diss(\omega(\boldsymbol{x})) + \beta vac(\omega(\boldsymbol{x})] \tag{12}$$

where β is an annealing coefficient to gradually balance between vacuity and dissonance. Specifically, the importance of vacuity reduces as there are less "vacuous" areas in the data space w.r.t. the current training data. This implies that the training data can well approximate the entire data space. To achieve this, we define $1/\beta = \min_{\mathbf{x}_u\in\mathbb{X}_u} \max_{\mathbf{x}_t\in\mathbb{X}_t} \text{sim}(f(\boldsymbol{x}_t), f(\boldsymbol{x}_u))$ and set $\beta = \max\{0, 1 - dT\}$ if it does not change within the past few AL iterations, where T denotes the current iteration of AL and d is a fixed decay rate (set to $1/100K$ in our experiments).

5 Experiments

We report our experimental results on both synthetic and real-world data. The former aims to verify the key theoretical properties of ADL, including entropy decomposition and multi-source uncertainty prediction, and how these properties contribute to AL. The real-data experiment aims to compare ADL and its competitors. We focus on testing in classical AL environment, where the initial training set only includes limited samples from some classes with samples from other classes completely missing. In each AL iteration, we sample one data instance. This is fundamentally different than some recent DL based AL methods, such as [3, 17], which perform batch-mode sampling with a large batch size (larger than our entire labeled samples). Thus, these models are not applicable when only limited label budgets are available which is true for many special domains where labeling is very costly. All models uses the same DL architecture. For synthetic data, we adopt a 3-layer MLP with tanh for activation. For real data, we use LeNet with Relu for activation.

5.1 Synthetic data

The synthetic experiment is designed to show: (1) whether ADL accurately captures different sources of uncertainty, and (2) whether accurately estimated uncertainty leads to better AL behavior. To

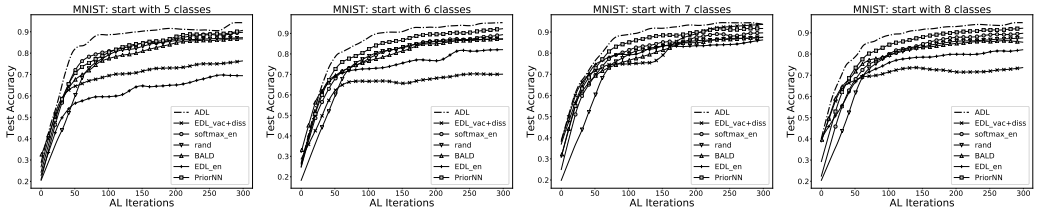


Figure 4: AL performance on MNIST (start with 5-8 classes)

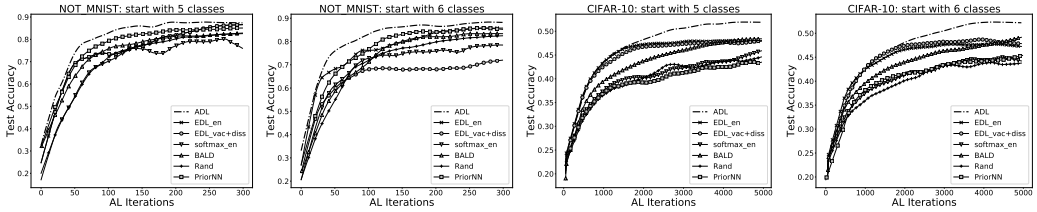


Figure 5: AL performance on notMNIST and CIFAR-10 (start with 5, 6 classes; 7,8 in Appendix)

mimic the existence of OOD, we generate three mixtures of Gaussian's. Each mixture consists of a major and a smaller (i.e., OOD) clusters with 750 and 50 samples, respectively. We center the major Gaussian components from each class and put their corresponding OOD components away from them. Figure 1 shows that a classical DL model with a softmax layer provides very inaccurate uncertainty estimation. In contrast, the proposed ADL model not only provides accurate entropy prediction but also successfully decomposes it into vacuity and dissonance. Figure 2 shows the uncertain prediction result from EDL [6], which can also provide evidence prediction but requires ample training data. Suffering from insufficient training, EDL is inaccurate in its entropy prediction, especially for the OOD clusters. While EDL does not provide entropy decomposition, we use its predicted evidence to compute vacuity and dissonance as shown in Figure 2. However, neither of them is accurately predicted as low vacuity is predicted for the three OOD clusters where there is no training data and high dissonance is predicted in areas with no nearby training data to show conflicting evidence.

Figure 3 shows the first time when ADL selects at least one data sample from each OOD area, high vacuity is assigned to an area with no training data but many unlabeled data. Meanwhile, high dissonance indicates that refining the decision boundary may be more instrumental to improve the model performance. A few iterations later, ADL starts to penalize dissonance. While vacuity is still accurately estimated (high in vacuous regions), it becomes less useful for sampling (since very few unlabeled data is located nearby). Two iterations later, penalizing on dissonance helps to choose data samples that significantly refine the decision boundary. The superior AL performance of ADL as shown in Figure 2 further confirms the effectiveness of ADL's key properties as demonstrated above.

5.2 Real data

The real-world experiment is conducted on three datasets, MNIST, notMNIST, and CIFAR-10, all of which have ten classes. To mimic the real-world AL scenario, we leave 2-5 classes out for initial training. A good AL model is expected to discover samples of unknown classes in an early stage to effectively improve model accuracy. We compare the proposed model with **EDL** [7] (entropy, vacuity+dissonance), **BALD** [2] (epistemic), **PriorNN** [11] (distributional uncertainty), and **softmax** (entropy, random), where in the parenthesis are the uncertainty measurements used for sampling. Figures 4 and 5 show that ADL consistently outperforms other models on all three datasets. The advantages of ADL are twofold. First, entropy decomposition gives ADL flexibility to meet distinct sampling need at different AL phases. In an early stage, the fast accuracy improvement is achieved by the vacuity guided sampling where the most *representative* samples are labeled with high priority. Gradually, ADL switches to dissonance guided sampling to refine the decision boundary by labeling the most *informative* samples to improve its discriminative power. In contrast, sampling methods utilizing a unified uncertainty (e.g., epistemic, distributional uncertainty and entropy) lack such flexibility to adjust the sampling behavior, leading to either slow convergence or lower model accuracy. Second, compared with EDL, which can also perform evidence prediction, ADL is superior due to accurate uncertainty estimation using the effective loss function. For both synthetic and real

data, we observe that ADL identifies samples from missing classes at least around 20% faster than using EDL and other models.

We also conducted a thorough ablation study to justify the effectiveness of each major technical component, including sampling function, vacuity/dissonance balance, anchor sample identification, and related model parameters. Additional comparisons with the state-of-the-art batch-mode AL models [3, 17] are also performed. ADL shows clear advantage in single batches and also performs better in small batches of 2-4. Since ADL is not designed for batch sampling, it becomes less effective for larger batches with a size greater than 6. Detailed results are reported in Appendix C and D.

5.3 Discussions

The subjective logic framework provides a strong theoretical underpinning to perform a principled, fine-grained analysis between the 1st-order uncertainty (i.e., predictive entropy as the total uncertainty) and 2nd-order uncertainty (vacuity + dissonance), where evidence plays a central role to unveil the underlying (dynamic) relationship among different uncertainties. Understanding this dynamic relationship is essential to derive a theoretically sound data sampling process in AL. In particular, Theorem 1 shows the total uncertainty dynamically shifts between high vacuity and high dissonance as more evidence is collected. Guided by this theory, *AL can be regarded as an evidence collection process*. The evidence-based uncertainties (i.e., vacuity + dissonance) derived under SL, offer a natural way to determine the sources of uncertainty during AL, which starts by focusing on vacuity in early stage when evidence is limited and then gradually shifts to dissonance. Since vacuity and dissonance both depend on evidence, using evidence provides a principled way to trace the dynamic shift between these two sources of uncertainty to best guide data sampling in AL. This is the key advantage over other types of uncertainty, such as epistemic, aleatoric, and distributional uncertainty, which only focus a certain aspect of uncertainty, and their (dynamic) relationship is hard to be precisely quantified as in vacuity and dissonance. Thus, using these uncertainty measures lacks the capability to dynamically adjust the sampling process as the nature and focus of uncertainty change when more data samples are labeled. In sum, while there are various forms of uncertainty measures, the evidence-based (2nd-order) uncertainty, i.e., vacuity + dissonance, offers the most suitable way for active sampling, as justified by our theory and empirical evaluation. Our experimental results clearly show that ADL converges much faster than other uncertainty based sampling functions, which empirically confirms its effectiveness in AL.

6 Conclusion

We present a novel active deep learning model that systematically leverages two distinct sources of uncertainty, vacuity and dissonance, to effectively explore a large and high-dimensional data space for label-efficient training of DL models. The proposed ADL model benefits from the evidence-based entropy decomposition that follows from our theoretical analysis of belief vacuity and belief dissonance under the SL framework. The multi-source uncertainty can be accurately estimated through a novel loss function that augments DL based evidence prediction with vacuity-aware regularization of the model parameters. Through dynamically balancing the importance of vacuity and dissonance, a sampling function is designed to first explore the critical areas of the data space and then fine-tune the decision boundary to maximize its discriminate power. Extensive experiments conducted over both synthetic and real data help verify the theoretical properties and empirical performance of the proposed ADL model.

Broader Impact

The ability to train high-quality supervised learning models opens the gate to effectively leveraging machine intelligence in many critical domains, where domain expertise is scarce and data labeling is costly. The proposed research suggests a paradigm shift to train statistical models that is fundamentally different from existing data-intensive statistical analysis tools, including most deep learning models, where massive amounts of training data are required to ensure model performance. Furthermore, the research will also contribute novel (both theoretical and empirical) methodologies to conduct uncertainty analysis of neural network/deep learning models. The importance of uncertainty analysis plays an essential role to establish human-machine trust to support complex human-machine collaborative decision-making in critical missions and to ensure AI safety as AI systems have been more broadly adopted in society.

Acknowledgement

Weishi Shi and Qi Yu are supported in part by an ONR award N00014-18-1-2875 and an NSF IIS award IIS-1814450. Xujiang Zhao and Feng Chen are supported by the NSF under Grant No #1815696 and #1750911. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

References

[1] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

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

[3] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[4] Audun Jøsang. *Subjective logic*. Springer, 2016.

[5] Clarence W De Silva. *Intelligent control: fuzzy logic applications*. CRC press, 2018.

[6] Kari Sentz, Scott Ferson, et al. *Combination of evidence in Dempster-Shafer theory*, volume 4015. Citeseer, 2002.

[7] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pages 3179–3189, 2018.

[8] Audun Jøsang, Jin-Hee Cho, and Feng Chen. Uncertainty characteristics of subjective opinions. In *Fusion*, pages 1998–2005. IEEE, 2018.

[9] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 2016.

[10] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, pages 5574–5584, 2017.

[11] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *arXiv preprint arXiv:1802.10501*, 2018.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

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

[14] Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. Reliable uncertainty estimates in deep neural networks using noise contrastive priors. 2018.

[15] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pages 112–119. IEEE, 2014.

[16] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

[17] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2020.

[18] Nils J Nilsson. Probabilistic logic. *Artificial intelligence*, 28(1):71–87, 1986.

[19] Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

[20] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[21] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.