Inconsistency-Based Data-Centric Active Open-Set Annotation

Ruiyu Mao, Ouyang Xu, Yunhui Guo

University of Texas at Dallas
{ruiyu.mao, oxu, yunhui.guo}@utdallas.edu

Abstract

Active learning, a method to reduce labeling effort for training deep neural networks, is often limited by the assumption that all unlabeled data belong to known classes. This closed-world assumption fails in practical scenarios with unknown classes in the data, leading to active open-set annotation challenges. Existing methods struggle with this uncertainty. We introduce NEAT, a novel, computationally efficient, data-centric active learning approach for open-set data. NEAT differentiates and labels known classes from a mix of known and unknown classes, using a clusterability criterion and a consistency measure that detects inconsistencies between model predictions and feature distribution. In contrast to recent learning-centric solutions, NEAT shows superior performance in active open-set annotation, as our experiments confirm. Additional details on the further evaluation metrics, implementation, and architecture of our method can be found in the public document at https://arxiv.org/pdf/2401.04923.pdf.

Introduction

The remarkable performance of modern deep neural networks owes much to the availability of large-scale datasets such as ImageNet (Deng et al. 2009). However, creating such datasets is a challenging task that requires a significant amount of effort to annotate data points (Sorokin and Forsyth 2008; Snow et al. 2008; Raykar et al. 2010; Welinder et al. 2010). Fortunately, active learning offers a solution by enabling us to label only the most *significant* samples (Settles 2009; Cohn, Ghahramani, and Jordan 1996; Settles 2011; Beygelzimer, Dasgupta, and Langford 2009; Gal, Islam, and Ghahramani 2017). In active learning, a small set of labeled samples from known classes is combined with unlabeled samples, with the goal of selecting specific samples to label from the pool to enhance model training. Typical active learning methods for training deep neural networks involve selecting samples that have high levels of uncertainty (Settles 2009; Cohn, Atlas, and Ladner 1994; Balcan, Beygelzimer, and Langford 2006), are in close proximity to the classification boundary (Ducoffe and Precioso 2018), or reveal cluster structures from the data (Sener and Savarese 2017; Ash et al. 2019).

Despite the considerable body of research on active learning, its practical implementation in an *open-world* context remains relatively unexplored (Ning et al. 2022). Unlike in closed-world active learning settings, where the unlabeled data pool is assumed to consist only of known classes, the open world introduces unknown classes into the unlabeled data pool, making active open-set annotations a challenge. For example, when collecting data to train a classifier to distinguish between different dog breeds as known classes, a large number of unlabeled images collected from online sources may include images from unknown classes such as wolves and coyotes. The task is to identify images with known classes and select informative samples for labeling, while avoiding samples from unknown classes. Existing active learning methods face a significant obstacle, as they tend to select samples from unknown classes due to their high uncertainty (Ning et al. 2022), which is undesirable.

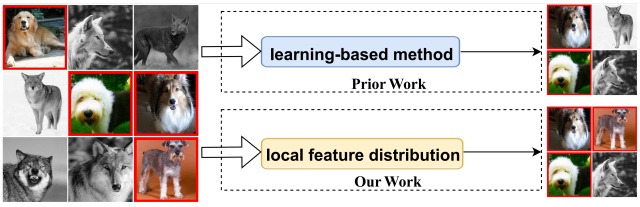

Figure 1: Dataset consists of color images as known dog class and gray-scale images as unknown wolf class. The objective is to find and label examples from the known classes in unlabeled data pool. Prior work using learning-based approach may identify some ambiguous unknown classes samples as known classes samples. Our work focusing on local feature distribution can find known classes samples more accurately.

Addressing the active open-set annotation problem requires specialized methods, and one such approach, called LFOSA was recently proposed by Ning et al. (2022). LFOSA is a learning-based active open-set annotation as it involves training a detector network with an additional output for unknown classes. The predictions of the detector network on unlabeled data are used for identifying known classes. While this approach can achieve impressive performance, it has two limitations: 1) training the additional detector network is costly, 2) and it is difficult to identify informative samples from the known classes as there is a contradiction in that while excluding unknown classes is necessary, it is also easy

to exclude informative samples from the known classes.

In this paper, we propose a novel inconsistency-based data-centric active learning method to actively annotate informative samples in an open world, which not only reduces the computational cost but also improves the performance of active open set annotation (Figure 1). Rather than using a learning-based approach to differentiate known and unknown classes, we suggest a data-centric perspective that naturally separates them by label clusterability, eliminating the need for an additional detector network. In addition, our method involves selecting informative samples from known classes by estimating the inconsistency between the model prediction and the local feature distribution. For example, suppose the model predicts an unlabeled sample as a wolf, but the majority of nearby samples are actually dogs. In that case, we would choose to label this unlabeled sample. The proposed inconsistency-based active learning approach shares a similar spirit to the version-space-based approach (Cohn, Atlas, and Ladner 1994; Balcan, Beygelzimer, and Langford 2006; Dasgupta 2005). However, there are two key differences. Firstly, our hypothesis class consists of deep neural networks. Secondly, the version-space-based approach identifies uncertain examples by analyzing the consistencies among multiple models, our approach leverages a fixed model and estimates the consistencies between the model prediction and the local feature distribution.

The contributions of this paper are summarized below,

- We introduce NEAT, a novel and efficient inconsistency-based data-centric active learning method. This method is designed to select informative samples from known classes within a pool containing both known and unknown classes. To the best of our knowledge, our work stands as the pioneering effort in utilizing large language models for active open-set annotation.
- Compared with the learning-based method, the proposed data-centric active learning method is more computationally efficient and can effectively identify informative samples from the known classes.
- Extensive experiments show that NEAT achieves much better results compared to standard active learning methods and the method specifically designed for active open set annotation. In particular, NEAT achieves an average accuracy improvement of 9% on **CIFAR10**, **CIFAR100** and **Tiny-ImageNet** compared with existing active open-set annotation method given the same labeling budget.

Related Work

Active learning, a topic which is extensively studied in machine learning (Lewis and Gale 2001; Settles 2009; Cohn, Ghahramani, and Jordan 1996; Settles 2011; Beygelzimer, Dasgupta, and Langford 2009), focuses on selecting the most uncertain samples for labeling. Uncertainty-based active learning methods (Lewis and Gale 2001) employ measures such as entropy (Shannon 1948), least confident (Settles 2009), and margin sampling (Scheffer, Decomain, and Wrobel 2001). Another widely used approach is the version-space-based method (Cohn, Atlas, and Ladner 1994; Balcan, Beygelzimer, and Langford 2006; Dasgupta 2005, 2011) which maintains multiple models consistent with labeled samples. If these models make inconsistent predictions on an unlabeled sample, a label query is prompted.

For deep neural networks, specific active learning methods have been developed. For instance, Core-Set (Sener and Savarese 2017) advocates for an active learning method that selects samples which are representative of the whole data distribution. BADGE (Ash et al. 2019) selects samples based on predictive uncertainty and sample diversity. In particular, BADGE leverages k-MEANS++ to select a set of samples which have diverse gradient magnitudes. The approach BGADL (Tran et al. 2019) integrates active learning and data augmentation. It leverages Bayesian inference in the generative model to create new training samples from selected existing samples. These generated samples are then utilized to enhance the classification accuracy of deep neural networks. CEAL (Wang et al. 2016) combines pseudo-labeling of clearly classified samples and actively labeled informative samples to train deep neural networks. For deep object detection, active learning is introduced in (Brust, Käding, and Denzler 2018), utilizing prediction margin to identify valuable instances for labeling. DFAL (Ducoffe and Precioso 2018) is an adversarial active learning method for deep neural networks that selects samples based on their distance to the decision boundary, approximated using adversarial examples. The samples closest to the boundary are chosen for classifier training. Recently, LFOSA (Ning et al. 2022) is proposed as the first active learning method for active open-set annotation. LFOSA trains a detector network to identify known classes and selects samples based on model confidence. MQNET (Park et al. 2022) is an open-set active learning method that addresses the Purity-Informativeness Dilemma and dynamically solves it using a Meta-learning MLP (Multilayer Perceptron) network, based on the samples' purity and informativeness scores.

Preliminary

Active Open-Set Annotation

Assume the sample-label pair (X, Y) has the joint distribution P_{XY}. For a given sample $\mathbf{x}$, the label y can be determined via the conditional expectation,

$$\eta(\mathbf{x}) = \mathbb{E}[Y|X = \mathbf{x}] \tag{1}$$

We consider a pool-based active learning setup. We begin by randomly sampling a small labeled set denoted as $L = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$ from the joint distribution P_{XY}. We also have access to a large set of unlabeled samples, denoted as U. Given a deep neural network f_θ parameterized by θ, the expected risk of the network is computed as $R = \mathbb{E}[\ell(y, f_\theta(\mathbf{x}))] = \int l(y, f_\theta(\mathbf{x}))dP_{XY}$, where $\ell(y, f_\theta(\mathbf{x}))$ represents the classification error. We conduct T query rounds, where in each round t, we are allotted a fixed labeling budget B to identify B informative samples from U, denoted as U_B. The queried samples U_B are given labels and added to the initial labeled set, with the aim of minimizing the sum of cross-entropy losses $\sum_{(\mathbf{x},y)\in L \bigcup U_B} \ell_{\text{CE}}(f_\theta(\mathbf{x}), y)$ and reducing the expected risk of the model f_θ (Settles 2009). The term $\ell_{\text{CE}}(f_\theta(\mathbf{x}), y)$ represents the cross-entropy loss between the predicted and true labels for a sample $(\mathbf{x}, y)$ (Bishop 2006).

In the context of active open-set annotation (Ning et al. 2022), the set of unlabeled samples U comprises data from both known classes and unknown classes, denoted as S_{known} and $S_{unknown}$, respectively. Importantly, S_{known} represents the classes observed during the initial labeling process, and it is ensured that $S_{known} \cap S_{unknown} = \emptyset$. In each active query round, we label the samples from known classes with their true labels and label the samples from unknown classes as *invalid*. Active open-set annotation presents a greater challenge than standard active learning, as it requires the active learning method **1)** to distinguish between known and unknown classes among the unlabeled samples **2)** and to select informative samples exclusively from the known classes.

Learning-Based Active Open-Set Annotation

The study of active open-set annotation is currently limited. A recent paper (Ning et al. 2022) proposed a new learning-based method called LFOSA which is specially designed for active open-set annotation. This approach combines a model, f_θ, trained on known classes with a detector network that identifies unknown class samples. Suppose there are C known classes, the detector network has $C+1$ outputs, similar to OPENMAX(Bendale and Boult 2016), which allow for the classification of unlabeled data into known and unknown classes. During an active query, LFOSA focuses only on unlabeled data classified as known classes, and then uses a Gaussian Mixture Model to cluster the activation values of detected known classes data into two clusters. The data closer to the larger cluster mean is selected for labeling. This learning-based approach has shown significant improvement over traditional active learning methods for active open-set annotation. However, there are still challenges that need to be addressed, such as the additional computation cost required to train the detector network and the difficulty of identifying informative samples from the known classes.

NEAT

Algorithm 1 describes the detailed algorithm of NEAT. Initially, NEAT begins with a randomly drawn labeled set L from known classes. In each query round, NEAT performs two main steps. First, it identifies unlabeled samples whose neighbors in the labeled set belong to known classes. Secondly, NEAT selects a batch of unlabeled samples for labeling by estimating the inconsistency between the model's prediction and the local feature distribution. Unlike the learning-based method discussed in (Ning et al. 2022), NEAT is more computationally efficient. Furthermore, it effectively decouples the detection of known classes from the identification of informative samples, enabling it to identify informative samples even from the known classes, which is a challenge for existing active learning methods.

Data-Centric Known Class Detection

In the learning-based method (Ning et al. 2022), an additional detector network is trained to differentiate known and unknown classes. However, the training cost is high. In contrast, NEAT adopts a data-centric perspective for finding known classes samples based on feature similarity. In particular,

Algorithm 1: NEAT: Inconsistency-Based Data-Centric Active Open-Set annotation.

Require: A deep neural network f_θ, initial labeled set L, a set of known class Y_{Known} , unlabeled labeled set U, number of query rounds T, number of examples in each query batch B, a pre-trained model M, number of neighbors K.

1: Use the pre-trained model M for extracting features on L and U.
2: **for** $t \leftarrow 1$ to T **do**
3: Train the model f_θ on L by minimizing $\sum_{(x,y)\in L} \ell_{\text{CE}}(f_\theta(x), y)$
4: $S \leftarrow \{\}$.
5: For each sample $x \in U$:
6: Compute the output of the softmax function as P_x.
7: Find the K-nearest neighbors $\{N_i(\mathbf{x})\}_{i=1}^K$ of $\mathbf{x}$ in L based on the extracted features.
8: If all the labels from $\{N_i(\mathbf{x})\}_{i=1}^K$ belong to known classes Y_{Known}, then $S \leftarrow S \bigcup \{\mathbf{x}\}$.
9: Compute the score $I(x)$ using Eq. 5 for each sample $\mathbf{x} \in S$.
10: Rank the samples based on $I(\mathbf{x})$ and denote the B samples which have the largest scores as U_B.
11: Query the labels of each sample in U_B.
12: $U \leftarrow U \setminus U_B$, $L \leftarrow L \bigcup U_B$.
13: **end for**

NEAT relies on label clusterability to identify known classes from the unlabeled pool.

Label clusterability. We leverage the intuition that samples with similar features should belong to the same class (Gao, Yang, and Zhou 2016; Zhu, Song, and Liu 2021; Zhu, Dong, and Liu 2022). This can be formally defined as follows,

Definition 0.1. ((K, σ_K) label clusterability). A dataset D satisfies (K, σ_K) label clusterability if for all $\mathbf{x} \in D$, the sample $\mathbf{x}$ and its K-Nearest-Neighbors (K-NN) $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K\}$ belong to the same label with probability at least $1 - \sigma_K$.

When σ_K = 0, it is called K-NN label clusterability (Zhu, Song, and Liu 2021). Recent methods for noisy label learning (Zhu, Song, and Liu 2021; Zhu, Dong, and Liu 2022) leverage label clusterability for detecting examples with noisy labels.

Feature extraction. To utilize the clusterability of labels for identifying known classes, we need to extract features from unlabeled inputs that group semantically similar samples in the feature space. Additionally, the quality of these features will inevitably impact the clusterability of the labels. Instead of developing a separate detector, as demonstrated in Ning et al. (2022), we suggest leveraging pre-trained large language models for feature extraction, which have been demonstrated to possess exceptional zero-shot learning ability (Radford et al. 2021). In particular, we leverage CLIP (Radford et al. 2021) to extract features for both the labeled and unlabeled data, providing high-quality features for calculating feature similarity.

Known class detection. By utilizing the features ex-

tracted by CLIP from both the labeled set L and the unlabeled set U, we can identify the K-nearest neighbors $\{N_1(\mathbf{x}), N_2(\mathbf{x}), ..., N_K(\mathbf{x})\}$ in the labeled set L for each sample $\mathbf{x} \in U$, using cosine distance. Each $N_k(\mathbf{x}) \in L$ represents the k-th closest samples in L to the unlabeled sample $\mathbf{x}$. Subsequently, we compute the count of neighbors with known and unknown classes for each unlabeled sample $\mathbf{x}$, assuming label clusterability. If an unlabeled sample is in proximity to numerous known class samples, it is more likely to belong to a known class. Therefore, we classify an unlabeled sample as belonging to a known class if all of its neighboring samples are from known classes.

Theoretical analysis. Here we give an upper bound on the detection error of the proposed known class detection. The first step is to understand in what condition our method will make a mistake. Given the sample $\mathbf{x}$, although all of its K-nearest neighbors $\{N_1(\mathbf{x}), N_2(\mathbf{x}), ..., N_K(\mathbf{x})\}$ belong to known classes, it is still possible that the sample is from the unknown classes. This can caused by the randomness in the labeling process and the quality of the features. Instead of using the definition 0.1 which is too coarse for our analysis, we introduce the following assumptions,

Assumption 0.1. There exists a constant $C > 0$ and $0 < \alpha < 1$ such that for any $\mathbf{x}$ and $\mathbf{x}'$,

$$\mathbb{P}(y_{true}(\mathbf{x}) \neq y_{true}(\mathbf{x}')) \leq C\rho(\mathbf{x}, \mathbf{x}')^{\alpha} \tag{2}$$

Here, $\rho(\mathbf{x}, \mathbf{x}')$ represents the distance between samples $\mathbf{x}$ and $\mathbf{x}'$, while $y_{true}(\mathbf{x})$ and $y_{true}(\mathbf{x}')$ correspond to the true labels of $\mathbf{x}$ and $\mathbf{x}'$, respectively. We use $r_K(\mathbf{x})$ to denote the radius of the ball centered at $\mathbf{x}$ such that $\forall\, \mathbf{x}'$ in the K-nearest neighbors of $\mathbf{x}$, $\rho(\mathbf{x}, \mathbf{x}') \leq r_K(\mathbf{x})$.

Assumption 0.2. For any $\mathbf{x}$, the labeling error is upper-bounded by a small constant e,

$$\mathbb{P}(y(\mathbf{x}) \neq y_{true}(\mathbf{x})) \leq e \tag{3}$$

Then, we can establish an upper bound for the detection error as follows,

Theorem 0.1. Given the assumption 0.1. and 0.2. and the number of neighbors K, the probability of making a detection error is upper-bounded as,

$$\mathbb{P}(\text{Error}|K) \leq \sum_{k=\lceil \frac{K-1}{2} \rceil + 1}^{K} \binom{K}{k} e^k (1-e)^{K-k} C^k r_K(\mathbf{x})^{\alpha k} + \sum_{k=0}^{\lfloor \frac{K+1}{2} \rfloor - 1} \binom{K}{k} e^k (1-e)^{K-k} C^{(K-k)} r_K(\mathbf{x})^{\alpha(K-k)} \tag{4}$$

The proof of the theorem will be given in the Appendix. From the theorem, we can conclude that a good feature will decrease the detection error as we expected.

Inconsistency-Based Active Learning

To improve accuracy with a fixed labeling budget, it is crucial to select informative samples for labeling. One active learning strategy, motivated by theory, is the version-space-based approach (Cohn, Atlas, and Ladner 1994; Balcan, Beygelzimer, and Langford 2006; Dasgupta 2005, 2011). This approach involves maintaining a set of models that are consistent with the current labeled data, and an unlabeled sample is selected for labeling if two models produce different predictions. However, implementing this approach for deep neural networks is challenging due to the computational cost of training multiple models (Ash et al. 2019). To address this issue, we propose an inconsistency-based active learning method that does not require training multiple models and naturally leverages features produced by CLIP.

Given an unlabeled sample $\mathbf{x} \in U$, the model f_θ's prediction for $\mathbf{x}$, denoted as $P_\mathbf{x} \in \mathbb{R}^C$, represents a probability vector where $P_\mathbf{x}[c]$ is the model's confidence that $\mathbf{x}$ belongs to class c. Since we lack ground-truth labels, measuring prediction accuracy is impossible. To address this, we propose evaluating the importance of the sample for improving model training by assessing whether the model prediction aligns with local feature similarity. For example, if the model predicts an unlabeled sample as a dog with low probability but the majority of the sample's neighbors belong to the dog class, then either the model's prediction is incorrect or the sample is near the decision boundary between the dog class and the true class. In either case, the sample can be labeled to improve model training.

Given the K-nearest neighbors $\{N_i(\mathbf{x})\}_{i=1}^K$ of the example $\mathbf{x}$, we first construct a vector $V_\mathbf{x} \in \mathbb{R}^C$ with $V_\mathbf{x}[c] = \sum_k \mathbf{1}(Y_k(\mathbf{x}) = c)$, where $Y_k(\mathbf{x})$ is the label of k-th nearest neighbor of $\mathbf{x}$ and $\mathbf{1}(\cdot)$ is an indicator function. Then the vector $V_\mathbf{x}$ is normalized via the softmax function to be a probabilistic vector $\tilde{V}_\mathbf{x}$. The inconsistency between the model prediction and local feature similarity is computed using cross-entropy as,

$$I(\mathbf{x}) = -\sum_{c=1}^{C} P_\mathbf{x}[c] \log \tilde{V}_\mathbf{x}[c]. \tag{5}$$

A large $I(\mathbf{x})$ indicates that the model prediction is inconsistent with local feature distribution. Similar to version-space-based approach, the unlabeled sample is selected for labeling. In each query round, we rank all the identified known classes samples in the first stage using $I(\mathbf{x})$ and select the top B samples for labeling.

Experiments

Experimental Settings and Evaluation Protocol

Datasets and models. We consider **CIFAR10** (Krizhevsky and Hinton 2009), **CIFAR100** (Krizhevsky and Hinton 2009), and **Tiny-Imagenet** (Le and Yang 2015), to evaluate the performance of our proposed method. Similar to the existing methods for active open-set annotations (Ning et al. 2022), we leverage a ResNet-18 (He et al. 2016) architecture to train the classifier for the known classes. For the proposed method, we leverage CLIP (Radford et al. 2021) to extract the features for both known and unknown classes.

Active open-set annotation. In accordance with (Ning et al. 2022), the experiment randomly selects 40 classes, 20 classes and 2 classes from **Tiny-Imagenet**, **CIFAR100**, and **CIFAR10**, respectively, as known classes, while the remaining

classes are treated as unknown classes. To begin with, following (Ning et al. 2022), 8% of the data from the known classes in **Tiny-ImageNet** and **CIFAR100**, and 1% of the data from the known classes in **CIFAR10** are randomly selected to form an initial labeled set. The rest of the known class data and the unknown class data are combined to form the unlabeled data pool.

Baseline methods. We consider the following active learning methods as baselines,

1. RANDOM: The naive baseline which randomly selects samples for annotation;
2. UNCERTAINTY (Settles 2009): A commonly used active learning method which selects samples with the highest degree of uncertainty as measured by entropy;
3. CERTAINTY (Settles 2009): A common baseline for active learning which selects samples with the highest degree of certainty as measured by entropy;
4. CORESET (Sener and Savarese 2017): This method identifies a compact, representative subset of training data for annotation;
5. BGADL (Tran et al. 2019): A Bayesian active learning method which leverages generative to select informative samples;
6. OPENMAX (Bendale and Boult 2016): A representative open-set classification method which can differentiate between known classes and unknown classes;
7. BADGE (Ash et al. 2019): An active learning method designed for deep neural networks which select a batch of samples with diverse gradient magnitudes;
8. MQNET (Park et al. 2022): An open-set active learning method employs a Meta-learning MLP network to dynamically select samples based on their purity and informativeness scores.
9. LFOSA (Ning et al. 2022): A learning-based active open-set annotation method which selects samples based on maximum activation value (MAV) modeled by a Gaussian Mixture Model;
10. NEAT (Passive): We also consider a baseline that is the passive version of NEAT. In NEAT (Passive), we do not leverage the proposed inconsistency-based active learning methods for selecting samples for labeling but instead randomly sample from the identified known classes samples.

Metrics. We evaluate various active learning methods based on three key metrics: accuracy, precision, and recall. Accuracy reflects how accurately the model makes predictions on the test set. To measure recall, we use the ratio of selected known class samples to the total number of known class samples present in the unlabeled pool, denoted as N^t_{known} and $N^{\text{total}}_{\text{known}}$, respectively, in the query round t. Precision, on the other hand, measures the proportion of true known class samples among the selected samples in each query round,

$$\text{precision} = \frac{N^t_{\text{known}}}{B} \qquad \text{recall} = \frac{\sum_{t=1}^{T} N^t_{\text{known}}}{N^{\text{total}}_{\text{known}}} \tag{6}$$

Implementations details. The classification model, specifically ResNet-18, was trained for 100 epochs using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01. The learning rate decayed by 0.5 every 20 epochs, and the training batch size is 128. There were 9 query rounds with a query batch size of 400. All experiments were repeated three times with different random seeds and the average results and standard deviations were reported. The experiments were conducted on four A5000 NVIDIA GPUs.

Results

NEAT VS. Baselines. We evaluate the performance of NEAT and other methods by plotting curves as the number of queries increases (Figure 2). It is evident that regardless of the datasets, NEAT consistently surpasses other methods in all cases. In particular, NEAT achieves much higher selection recall and precision compared to existing active learning methods which demonstrates the effectiveness of the data-centric known class detection. **1)** In terms of recall, NEAT consistently outperforms other active learning methods by a significant margin. Notably, on **CIFAR10**, **CIFAR100**, and **Tiny-ImageNet**, NEAT achieves improvements of 4%, 12%, and 7%, respectively, compared to LFOSA (Ning et al. 2022), which is a learning-based active open-set annotation method. **2)** In terms of precision, NEAT consistently maintains a higher selection precision than other baselines, with a noticeable gap. Importantly, NEAT's ability to differentiate between known and unknown classes is significantly improved by adding examples from unknown classes in the first query round. **3)** In terms of accuracy, NEAT consistently outperforms LFOSA (Ning et al. 2022) across all datasets, including **CIFAR10**, **CIFAR100**, and **Tiny-ImageNet**, with improvements of 6%, 11%, and 10%, respectively. These results indicate that the proposed NEAT method effectively addresses the open-set annotation (OSA) problem. It is worth noting that NEAT (Passive) achieves slightly higher recall and precision than NEAT. The possible reason is that by selecting informative samples from known classes, it is also more likely to include samples from unknown classes with high uncertainty. Despite having a slightly lower recall, NEAT achieves a higher accuracy in all data sets compared to NEAT (Passive). This demonstrates that NEAT can effectively identify informative samples from known classes to train the model.

The Use of Pre-Trained Models for Active Open-Set Annotation. Unlike the majority of existing active learning methods, NEAT incorporates a pre-trained model as an component of the method. Although there was some initial works that utilized large language models for natural language processing (NLP) tasks (Bai et al. 2020; Seo et al. 2022), to our knowledge, our work represents the first instance of leveraging a large language model for active open-set annotation.

The utilization of a pre-trained model inevitably introduces prior knowledge. Although this pre-trained model has not been directly trained on the target dataset, it is essential to examine its potential impact on other active learning methods. To this end, we conducted experiments involving the pre-trained model module across various baseline methods. In particular, we adopt a hybrid approach that combines *data-centric known class detection* to initially identify potential known classes, followed by the application of different active learning methods using the data from these detected known classes. The empirical results indeed show the influence of the pre-trained model on classification accuracy. However, our

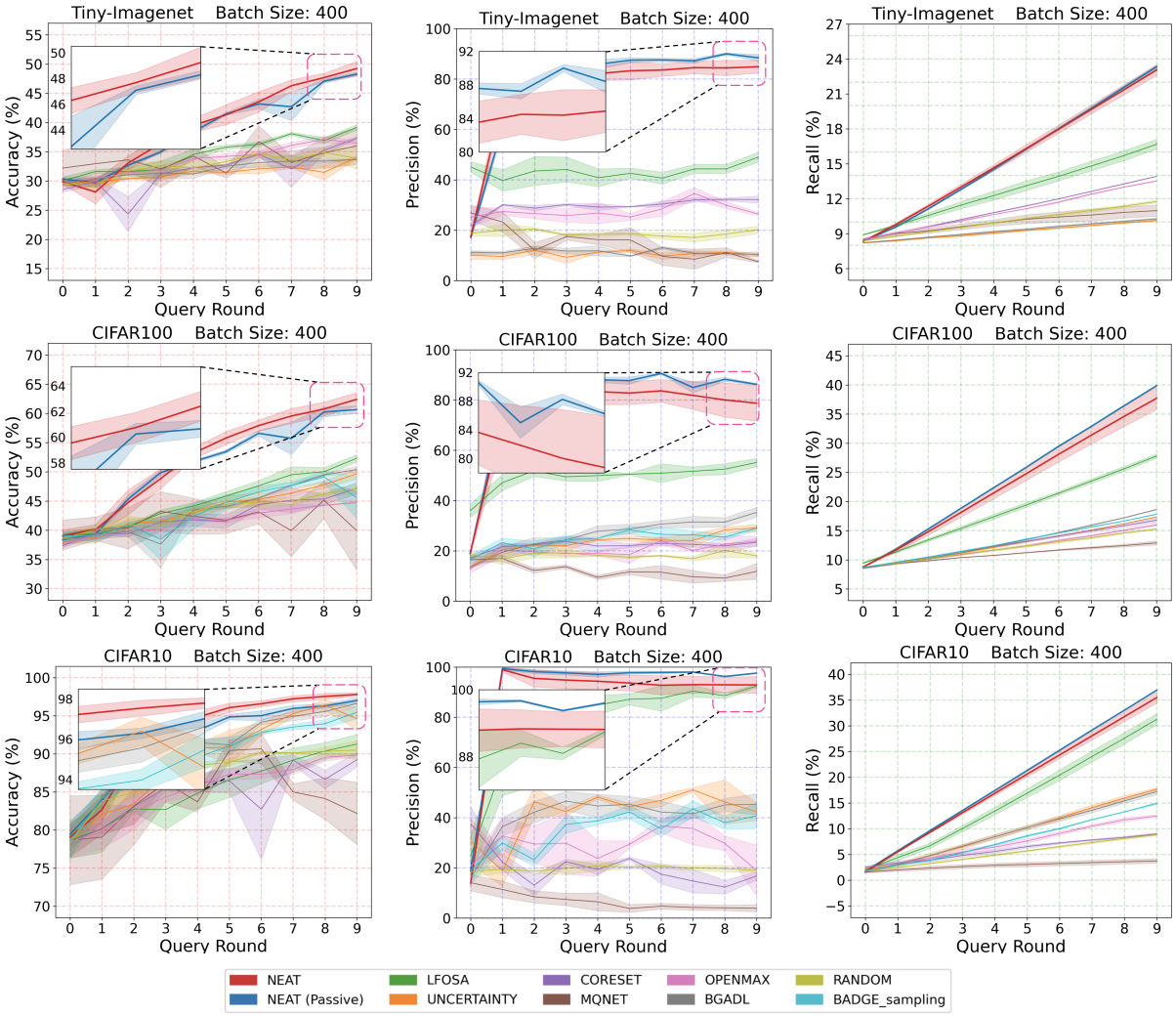

Figure 2: NEAT achieves higher precision, recall and accuracy compared with existing active learning methods for active open-set annotation. We evaluated NEAT and the baseline active learning methods on CIFAR10, CIFAR100 and Tiny-ImageNet based on accuracy, precision and recall.

proposed NEAT method still surpasses other active learning techniques in terms of accuracy on the **Tiny-ImageNet**, as shown in Figure 3.

Computational Efficiency. Our approach is more computationally efficient compared to learning-based open-set active learning methods, such as LFOSA. This efficiency is attributed to the fact that learning-based methods like LFOSA require the additional training of a detection network to discern known and unknown classes. The training of the detection network requires a significant amount of additional time. Consequently, the training time for our method on the **Tiny-ImageNet** is 88 minutes on an NVIDIA A5000 GPU, while the LFOSA method requires 156 minutes. For the **CIFAR100**, our method takes 21 minutes, compared to LFOSA's 59 minutes, and for the **CIFAR10**, our method requires 17 minutes, while LFOSA needs 25 minutes. Overall, we can observe that NEAT is much faster than LFOSA.

Ablation Studies

Impact of feature quality. We investigate the impact of different pre-trained models for feature extraction. The quality of the features is a crucial factor that may affect the label clusterability of the dataset. Specifically, we consider pre-trained ResNet-18, ResNet-34, and ResNet-50 on ImageNet, in addition to CLIP. Table 1 demonstrates that NEAT achieves significantly better accuracy than the baseline active learning methods, regardless of the pre-trained models used. This highlights the robustness of NEAT in detecting known classes. It's worth noting that while CLIP features achieve better accuracy on **CIFAR10** and **CIFAR100** than ResNets, the accuracy on **Tiny-ImageNet** is lower. This could be because the ResNets are pre-trained on ImageNet, and their features are better suited for Tiny-ImageNet. However, large language models like CLIP can generally provide high-quality features that are useful across different datasets.

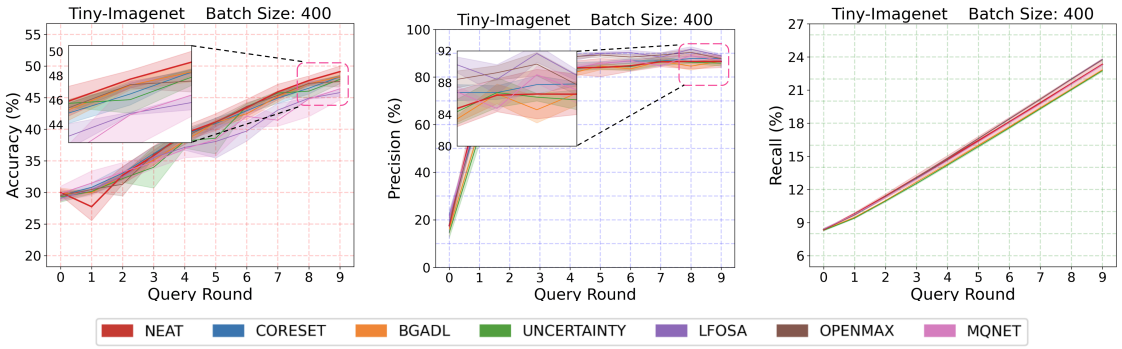

Figure 3: NEAT is effective compared with other active learning methods for deep neural networks.

Dataset	Model	Accuracy (avg ± std)
Tiny-Imagenet	ResNet-50	50.53 ± 1.15
	ResNet-34	50.17 ± 0.20
	ResNet-18	48.98 ± 0.42
	CLIP	49.01 ± 1.60
CIFAR100	ResNet-50	62.00 ± 1.26
	ResNet-34	62.73 ± 0.45
	ResNet-18	60.75 ± 0.58
	CLIP	63.68 ± 0.67
CIFAR10	ResNet-50	97.53 ± 0.24
	ResNet-34	97.48 ± 0.19
	ResNet-18	97.48 ± 0.26
	CLIP	98.15 ± 0.14

Table 1: We leverage ResNet-{18, 34, 50} and CLIP for feature extraction. The results show that NEAT is robust to the choices of pre-trained models.

Dataset	# of Neighbors	Accuracy (avg ± std)
Tiny-Imagenet	5	47.23 ± 1.21
	10	49.01 ± 1.60
	15	49.02 ± 0.61
	20	50.73 ± 0.29
CIFAR100	5	62.63 ± 0.66
	10	63.68 ± 0.67
	15	62.77 ± 0.84
	20	60.85 ± 1.70
CIFAR10	5	97.95 ± 0.27
	10	98.15 ± 0.14
	15	97.75 ± 0.14
	20	97.85 ± 0.21

Table 2: The impact of number of neighbors on the final performance of NEAT.

Influence of number of neighbors. We further investigate how the number of neighbors impacts the detection of data-centric known classes. We consider different values of K (5, 10, 15, 20) and present the results in Table 2. We observe that although a smaller value of K (e.g., $K = 5$) leads to slightly worse performance, other choices of K yield similar results. This suggests that a smaller K only captures the local feature distribution of the target sample, which not provide a good characterization of the underlying feature space. Ablation studies of query batch size and the impact of different classification models are included in the Appendix.

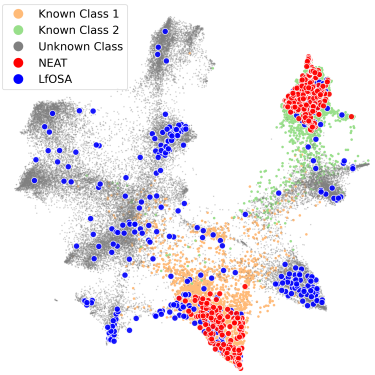

Figure 4: NEAT can accurately identify known classes from the unlabeled pool.

Visualization. We used t-SNE (Van der Maaten and Hinton 2008) to visualize the features produced by CLIP on **CIFAR10**, focusing on the samples selected by NEAT and LFOSA (Figure 4). We randomly select a query round for plotting. Our results show LFOSA selects many samples from unknown classes, explaining its low precision. Alternatively, nearly all samples NEAT selects are from known classes, demonstrating its effectiveness in the active open-set annotation problem.

Conclusion

In this paper, we propose a solution to the practical challenge of maintaining high recall when identifying examples of known classes for target model training from a massive unlabeled open-set. To address this challenge, we introduce a data-centric active learning method called NEAT, which utilizes existing large language models to identify known classes in the unlabeled pool. Our proposed method offers several advantages over traditional active learning methods. Specifically, NEAT uses the CLIP model to extract features, which eliminates the need to train a separate detector. Additionally, our approach achieves improved results with low query numbers, resulting in cost savings on labeling.

Acknowledgments

This work was supported by a startup funding by the University of Texas at Dallas.

References

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Bai, G.; He, S.; Liu, K.; Zhao, J.; and Nie, Z. 2020. Pretrained language model based active learning for sentence matching. *arXiv preprint arXiv:2010.05522*.

Balcan, M.-F.; Beygelzimer, A.; and Langford, J. 2006. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, 65–72.

Bendale, A.; and Boult, T. E. 2016. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1563–1572.

Beygelzimer, A.; Dasgupta, S.; and Langford, J. 2009. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, 49–56.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*, volume 4 of *Information science and statistics*. Springer. ISBN 9780387310732.

Brust, C.-A.; Käding, C.; and Denzler, J. 2018. Active learning for deep object detection. *arXiv preprint arXiv:1809.09875*.

Cohn, D.; Atlas, L.; and Ladner, R. 1994. Improving generalization with active learning. *Machine learning*, 15: 201–221.

Cohn, D. A.; Ghahramani, Z.; and Jordan, M. I. 1996. Active learning with statistical models. *Journal of artificial intelligence research*, 4: 129–145.

Dasgupta, S. 2005. Coarse sample complexity bounds for active learning. *Advances in neural information processing systems*, 18.

Dasgupta, S. 2011. Two faces of active learning. *Theoretical computer science*, 412(19): 1767–1781.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

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

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, 1183–1192. PMLR.

Gao, W.; Yang, B.-B.; and Zhou, Z.-H. 2016. On the resistance of nearest neighbor to random noisy labels. *arXiv preprint arXiv:1607.07526*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Lewis, D.; and Gale, W. 2001. A Sequential Algorithm for Training Text Classifiers. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in Information Retrieval*, 29.

Ning, K.-P.; Zhao, X.; Li, Y.; and Huang, S.-J. 2022. Active learning for open-set annotation. In *CVPR*, 41–49.

Park, D.; Shin, Y.; Bang, J.; Lee, Y.; Song, H.; and Lee, J.-G. 2022. Meta-Query-Net: Resolving Purity-Informativeness Dilemma in Open-set Active Learning. In *Proceedings of the 35th Neural Information Processing Systems (NeurIPS) Conference*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *Journal of machine learning research*, 11(4).

Scheffer, T.; Decomain, C.; and Wrobel, S. 2001. Active hidden markov models for information extraction. In *Advances in Intelligent Data Analysis: 4th International Conference, IDA 2001 Cascais, Portugal, September 13–15, 2001 Proceedings 4*, 309–318. Springer.

Sener, O.; and Savarese, S. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Seo, S.; Kim, D.; Ahn, Y.; and Lee, K.-H. 2022. Active learning on pre-trained language model with task-independent triplet loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11276–11284.

Settles, B. 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Settles, B. 2011. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, 1–18. JMLR Workshop and Conference Proceedings.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Snow, R.; O'connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 254–263.

Sorokin, A.; and Forsyth, D. 2008. Utility data annotation with amazon mechanical turk. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, 1–8. IEEE.

Tran, T.; Do, T.-T.; Reid, I.; and Carneiro, G. 2019. Bayesian generative active deep learning. In *International Conference on Machine Learning*, 6295–6304. PMLR.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, K.; Zhang, D.; Li, Y.; Zhang, R.; and Lin, L. 2016. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12): 2591–2600.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. 2010. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23.

Zhu, Z.; Dong, Z.; and Liu, Y. 2022. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning*, 27412–27427. PMLR.

Zhu, Z.; Song, Y.; and Liu, Y. 2021. Clusterability as an alternative to anchor points when learning with noisy labels. In *International Conference on Machine Learning*, 12912–12923. PMLR.