Inconsistency-Based Data-Centric Active Open-Set Annotation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Active learning is a commonly used approach that reduces the labeling effort required for training deep neural networks. However, the effectiveness of current active learning methods is limited by their closed-world assumptions, which assume that all data in the unlabelled pool comes from a set of predefined known classes. This assumption is often not valid in practical situations, as there may be unknown classes in the unlabeled data, leading to the active open-set annotation problem. The presence of unknown classes in the data can significantly impact the performance of existing active learning methods due to the uncertainty they introduce. To address this issue, we propose a novel data-centric active learning method called NEAT that actively annotates open-set data. NEAT is designed to label known classes data from a pool of both known and unknown classes unlabeled data. It utilizes the clusterability of labels to identify the known classes from the unlabeled pool and selects informative samples from those classes based on a consistency criterion that measures inconsistencies between model predictions and local feature distribution. Unlike the recently proposed learning-centric method for the same problem, NEAT is much more computationally efficient and is a data-centric active open-set annotation method. Our experiments demonstrate that NEAT achieves significantly better performance than state-of-the-art active learning methods for active open-set annotation.

1 Introduction

The remarkable performance of modern deep neural networks owes much to the availability of large-scale datasets such as ImageNet [1]. However, creating such datasets is a challenging task that requires a significant amount of effort to annotate data points [2, 3, 4, 5]. Fortunately, active learning offers a solution by enabling us to label only the most *significant* samples [6, 7, 8, 9, 10]. In active learning, a small set of labeled samples from known classes is combined with a pool of unlabeled samples, and the objective is to identify which samples to label from the pool for better training the model. Typical active learning methods for training deep neural networks involve selecting samples that have high levels of uncertainty [6, 11, 12], are in close proximity to the classification boundary [13], or unveil cluster structures from the data [14, 15].

Despite the considerable body of research on active learning, its practical implementation in an *open-world* context remains relatively unexplored [16]. 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 instance, 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

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

active learning methods face a significant obstacle in this task, as they may tend to select samples from unknown classes for labeling due to their high uncertainty [16], which is undesirable.

Addressing the active open-set annotation problem requires specialized methods, and one such approach called LFOSA is proposed recently by [16]. 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 computation cost but also improves the performance of active open-set annotation. 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's prediction and 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 [11, 12, 17]. 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.

In summary, the contributions are the paper are as follows,

- We propose a novel and efficient inconsistency-based data-centric active learning method, called NEAT, for selecting informative known classes samples from a pool of both known and unknown classes.
- Compared with the learning-based method, the proposed data-centric active learning method is computationally efficient and can effectively identify informative samples from the known classes.
- Extensive experiments show that NEAT achieves much better results compared with the standard active learning methods and the method which is 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 with the same labelling budget.

2 Related Work

Active Learning, extensively studied in machine learning [18, 6, 7, 8, 9], focuses on selecting the most uncertain samples for labeling. Uncertainty-based active learning methods [18] employ measures such as entropy [19], least confident [6], and margin sampling [20]. Another widely used approach is the version-space-based method [11, 12, 17, 21], where multiple models consistent with labeled samples are maintained, and inconsistent predictions prompt label querying.

For deep neural networks, specific active learning methods have been developed. For instance, Core-Set [14] advocates for an active learning method that selects samples which are representative of the whole data distribution. Badge [15] 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 [22] 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 [23] 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 [24], utilizing prediction margin to identify valuable instances for labeling. DFAL [13] 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 [16] 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.

3 Preliminary

3.1 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}, where $\mathbf{x}_i \in X$ and $y_i \in Y$. 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 [6]. The term $\ell_{\text{CE}}(f_\theta(\mathbf{x}), y)$ represents the cross-entropy loss between the predicted and true labels for a sample $(\mathbf{x}, y)$ [25].

In the context of active open-set annotation [16], the unlabeled samples in U include both data from known classes Y and unknown classes $\tilde{Y}$. In particular, $Y \cap \tilde{Y} = \emptyset$. 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 from the unlabeled samples 2) and to select informative samples exclusively from the known classes.

3.2 Learning-Based Active Open-Set Annotation

The study of active open-set annotation is currently limited. A recent paper [16] 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[26], 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 means 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.

4 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. Firstly, 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 prediction and local feature distribution. Unlike the learning-based method discussed in [16], NEAT is computationally efficient. Moreover, 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.

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.

Use the pre-trained model M for extracting features on L and U.
for $t \leftarrow 1$ to T **do**
    Train the model f_θ on L by minimizing $\sum_{(x,y)\in L} \ell_{\text{CE}}(f_\theta(x), y)$
    $S \leftarrow \{\}$
    For each sample $x \in U$:
        Compute the output of the softmax function as P_x
        Find the K-nearest neighbors $\{N_1(\mathbf{x}), N_2(\mathbf{x}), ..., N_K(\mathbf{x})\}$ of $\mathbf{x}$ in L based on the extracted features
        If all the labels of $\{N_1(\mathbf{x}), N_2(\mathbf{x}), ..., N_K(\mathbf{x})\}$ belong to known classes Y_{Known}, then $S \leftarrow S \bigcup \{\mathbf{x}\}$
    Compute the score $I(x)$ using Eq. 3 for each sample $\mathbf{x} \in S$
    Rank the samples based on $I(\mathbf{x})$ and denote the B samples which have the largest scores as U_B.
    Query the labels of each sample in U_B.
    $U \leftarrow U \setminus U_B$, $L \leftarrow L \bigcup U_B$
end for

4.1 Data-Centric Known Class Detection

In the learning-based method [16], an additional detector network is trained to differentiate known classes and unknown classes. However, the training cost is high. Instead, NEAT takes a data-centric perspective for finding known classes samples based on feature similarity. In particular, NEAT relies on label clusterability to find known classes from the unlabeled pool.

Label clusterability. We leverage the intuition that samples with similar features should belong to the same class [27, 28, 29]. This intuition can be formally defined as follows,

Definition 4.1. ((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, then it is called K-NN label clusterability [28]. Recent methods for noisy label learning [28, 29] leverage label clusterability for detecting examples with noisy labels. The label clusterability of the dataset is closely related to the smoothness condition introduced in [30] for analyzing nearest neighbor classifier [30] and non-parametric active learning [31],

Definition 4.2. ((α, L)-smooth) [30]. The conditional expectation η is (α, L)-smooth if for all $\mathbf{x}$ and $\mathbf{x}' \in D$,

$$|\eta(\mathbf{x}) - \eta(\mathbf{x}')| \leq L\mu(B(\mathbf{x}, \rho(\mathbf{x}, \mathbf{x}')))^\alpha \tag{2}$$

where $B(\mathbf{x}, \rho(\mathbf{x}, \mathbf{x}')$ is an open-ball centred at $\mathbf{x}$ with radius $\rho(\mathbf{x}, \mathbf{x}')$, μ is a measure on the input space.

When a conditional expectation is (α, L)-smooth, it intuitively implies that similar samples are highly likely to have the same label. Therefore, if the features accurately capture the semantic similarity between samples, the similarity in features reflects the clusterability of labels.

Feature extraction. To utilize the clusterability of labels for identifying known classes, we need to extract features from unlabelled 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 [16], we suggest taking advantage of pre-trained large language models for feature extraction, which have been demonstrated to possess exceptional zero-shot learning ability [32]. In particular, we leverage CLIP [32] 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 extracted 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 $\mathbf{x}$. Afterward, we compute the count of neighbors with known and unknown classes for each unlabeled sample $\mathbf{x}$, assuming label clusterability. If there are many unknown-class samples close to an unlabeled sample, it is more likely to belong to a known class. Finally, we determine the unlabeled samples with all neighbors belonging to known classes as potential known-class samples.

4.2 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 [11, 12, 17, 21]. 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 [15]. 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 $\{N_1(\mathbf{x}), N_2(\mathbf{x}), ..., N_K(\mathbf{x})\}$ 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,

$$I(\mathbf{x}) = -\sum_{c=1}^{C} P_\mathbf{x}[c] \log \tilde{V}_\mathbf{x}[c]. \tag{3}$$

A large $I(\mathbf{x})$ indicates that the model prediction is inconsistent with local feature similarity, 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.

5 Experiments

5.1 Experimental Settings and Evaluation Protocol

Datasets and models. We consider **CIFAR10** [33], **CIFAR100** [33], and **Tiny-Imagenet** [34], to evaluate the performance of our proposed method. Similar to existing methods for active open-set annotations [16], we leverage a ResNet-18 [35] architecture to train the classifier for the known classes. For the proposed method, we leverage CLIP [32] to extract the features for both the known classes and unknown classes.

Active open-set annotation. In accordance with [16], the experiment randomly selects 40 classes, 20 classes, and 2 classes from **Tiny-Imagenet**, **CIFAR100**, and **CIFAR10**, respectively, to be known classes, while the remaining classes are treated as unknown classes. To begin with, following [16], 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;

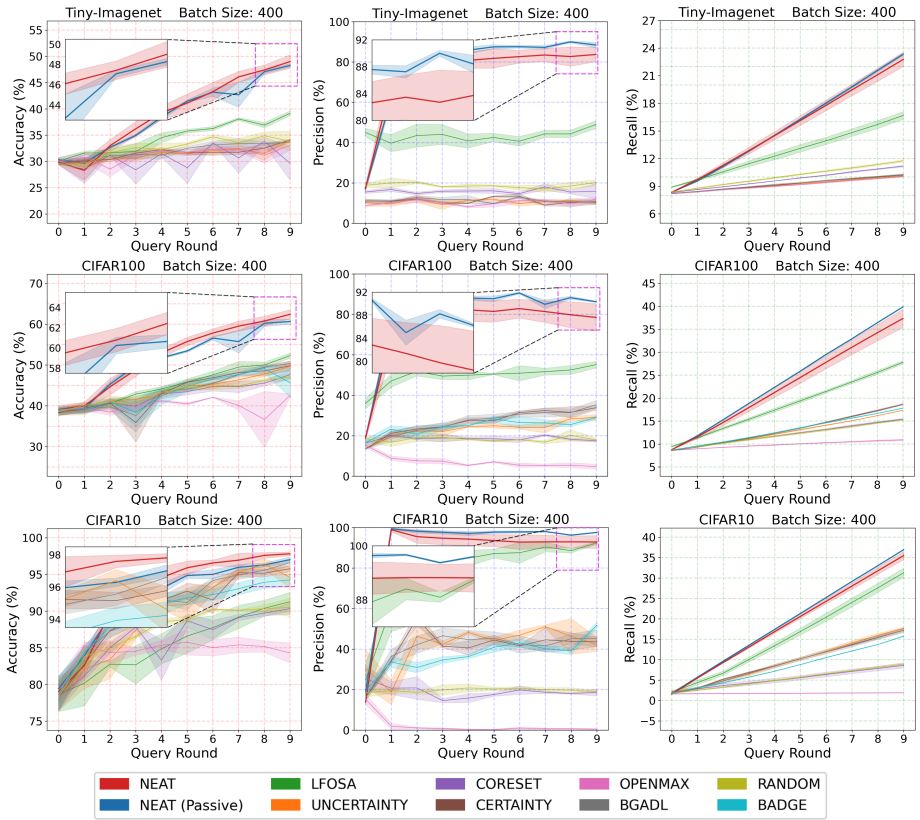

Figure 1: NEAT achieves higher precision, recall and accuracy compared with existing active learning methods for active open-set annotation. We evaluate NEAT and the baseline active learning methods on **CIFAR10**, **CIFAR100** and **Tiny-ImageNet** based on accuracy, precision and recall.

2. UNCERTAINTY [6]: A commonly used active learning method which selects samples with the highest degree of uncertainty as measured by entropy;
3. CERTAINTY [6]: A common baseline for active learning which selects samples with the highest degree of certainty as measured by entropy;
4. CORESET [14]: This method identifies a compact, representative subset of training data for annotation;
5. BGADL [22]: A Bayesian active learning method which leverages generative to select informative samples;
6. OPENMAX [26]: A representative open-set classification method which can differentiate between known classes and unknown classes;
7. BADGE [15]: A active learning method designed for deep neural networks which select a batch of samples with diverse gradient magnitudes;
8. LFOSA [16]: A learning-based active open-set annotation method which selects samples based on maximum activation value (MAV) modeled by a Gaussian Mixture Model;
9. NEAT (Passive): We also consider a baseline which is the passive version of NEAT, that is, we do not leverage the proposed inconsistency-based active learning methods for selecting samples for labeling and just 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. 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

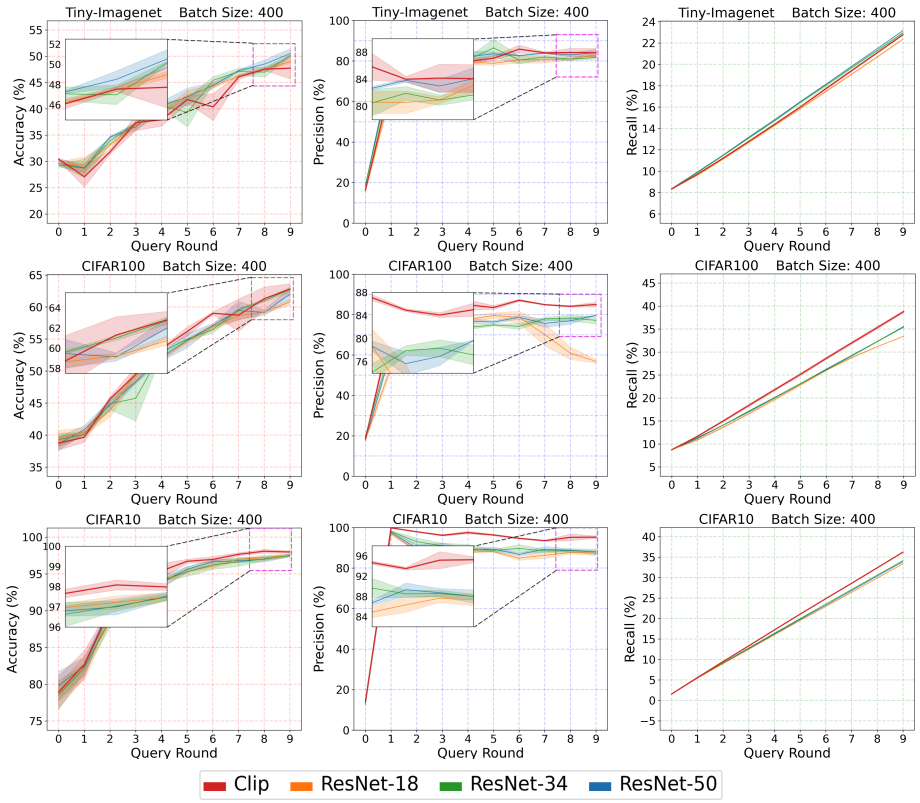

Figure 2: NEAT is robust to the choices of pre-trained models. We show the results of using different pre-trained models for feature extraction.

query round,

$$\text{precision} = \frac{N^t_{\text{known}}}{B} \qquad \text{recall} = \frac{\sum_{t=1}^{T} N^t_{\text{known}}}{N^{\text{total}}_{\text{known}}} \tag{4}$$

Implementations details. The classification model was trained for 100 epochs using SGD with an initial learning rate of 0.01. The learning rate decayed by 0.5 every 20 epochs, and the training batch size was set to 128. In total, there were 9 query rounds with a query batch size of 400. To ensure robustness, all experiments were repeated three times with different random seeds, and both the average results and standard deviations were reported. The experiments were conducted on four A5000 NVIDIA GPUs.

5.2 Results

5.2.1 NEAT VS. Baselines

We assess the performance of the proposed NEAT and compared methods by plotting curves with the number of queries increasing (Figure 1). 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 with 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 [16], 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 [16] across all datasets, including **CIFAR10**, **CIFAR100**, and **Tiny-ImageNet**, with improvements of 6%, 11%, and 10%, respectively. These results indicate that

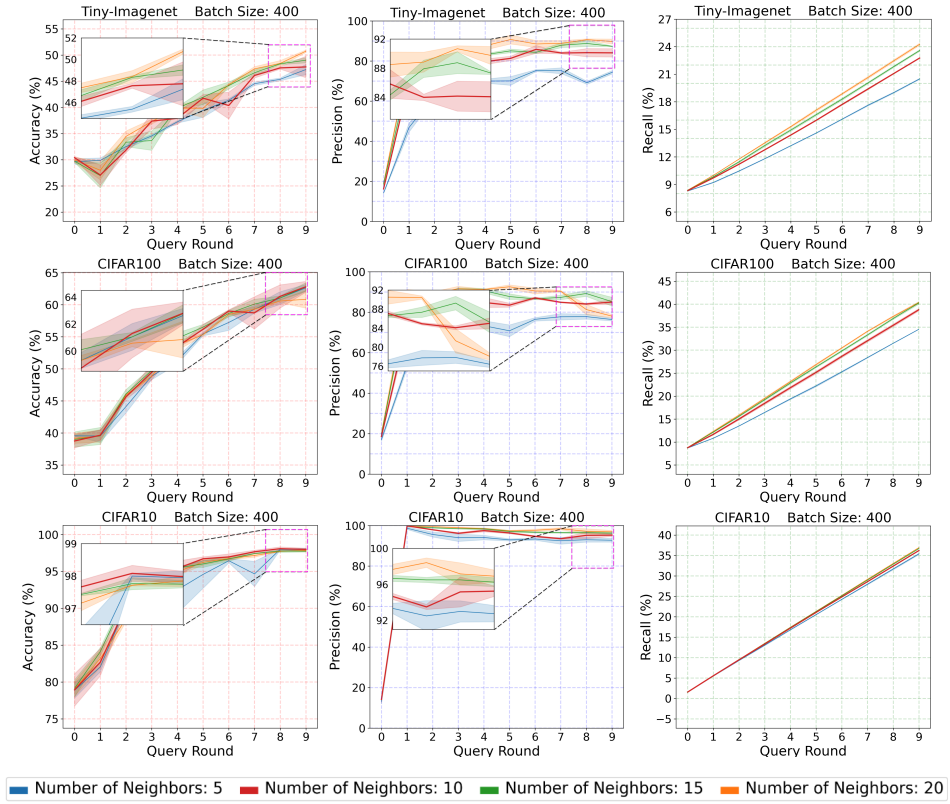

Figure 3: NEAT exhibits robustness to the choice of the number of neighbors, although a smaller value of K generally results in slightly lower performance.

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. However, despite a slightly lower recall, NEAT achieves higher accuracy across all datasets compared to NEAT (Passive). This demonstrates that NEAT can effectively identify informative samples from known classes to train the model.

5.2.2 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. Figure 2 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 **CIFAR-10** and **CIFAR-100** 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.

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 Figure 3. 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 may not provide a good characterization of the underlying feature space.

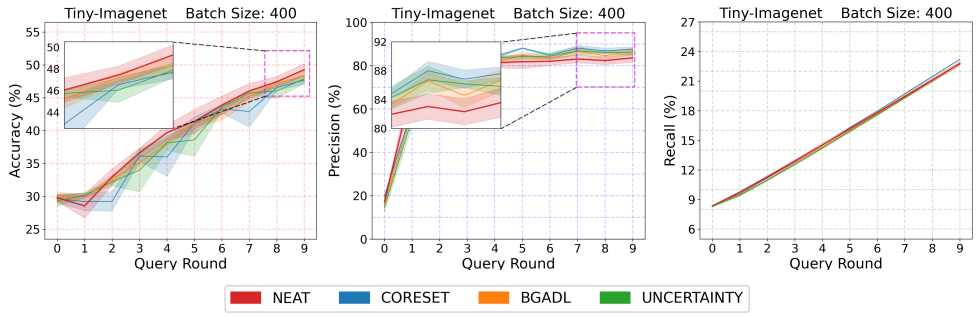

Figure 5: NEAT is effective compared with other active learning methods for deep neural networks.

Effectiveness of inconsistency-based active learning. We conducted an ablation study to evaluate the effectiveness of the inconsistency-based active learning step in comparison to other active learning methods for deep neural networks. 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 results on **Tiny-ImageNet** are presented in Figure 5. The findings demonstrate that our proposed inconsistency-based active learning method achieves higher accuracy when compared to other active learning methods. These results indicate that the proposed inconsistency-based active learning approach is capable of selecting more informative samples compared to alternative methods.

Ablation studies of query batch size and the impact of different classification models are included in the Appendix.

Visualization. We used t-SNE [36] to visualize the features produced by CLIP on CIFAR10, focusing on the samples selected by NEAT and LFOSA [16]. We randomly select a query round for plotting. Our results show that LFOSA selects a large number of samples from unknown classes, which explains its low precision. On the other hand, almost all of the samples selected by NEAT belong to the known classes, demonstrating its effectiveness in addressing the active open-set annotation problem.

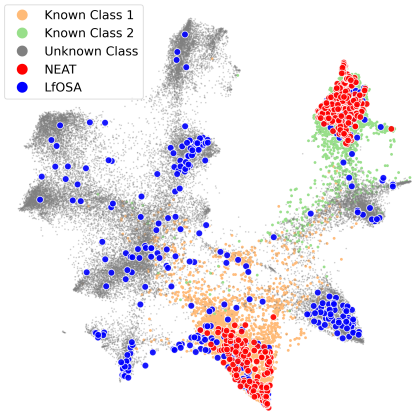

Figure 4: NEAT can accurately identify known classes from the unlabeled pool.

6 Summary

In this paper, we present 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 for identifying known classes from 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 for training a separate detector. Additionally, our approach achieves improved results with low query numbers, resulting in labeling cost savings. Furthermore, NEAT offers a plug-and-play solution with high adaptability to various datasets.

NEAT has the potential to create a positive societal impact by reducing the labeling cost for learning in an open world, which is crucial for practical applications such as autonomous driving and medical imaging. We believe the proposed NEAT should not raise any ethical considerations. However, a limitation of the proposed method is that for real-world application, maintaining pre-trained models like CLIP can result in significant memory costs. Therefore, it may be necessary to quantize and compress the model for more efficient deployment.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 1–8. IEEE, 2008.

[3] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 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*, pages 254–263, 2008.

[4] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11(4), 2010.

[5] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23, 2010.

[6] Burr Settles. Active learning literature survey. 2009.

[7] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

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

[9] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 49–56, 2009.

[10] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

[11] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15:201–221, 1994.

[12] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 65–72, 2006.

[13] Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

[14] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[15] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[16] Kun-Peng Ning, Xun Zhao, Yu Li, and Sheng-Jun Huang. Active learning for open-set annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–49, 2022.

[17] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. *Advances in neural information processing systems*, 18, 2005.

[18] David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.

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

[20] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 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*, pages 309–318. Springer, 2001.

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

[22] Toan Tran, Thanh-Toan Do, Ian Reid, and Gustavo Carneiro. Bayesian generative active deep learning. In *International Conference on Machine Learning*, pages 6295–6304. PMLR, 2019.

[23] 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.

[24] Clemens-Alexander Brust, Christoph Käding, and Joachim Denzler. Active learning for deep object detection. *arXiv preprint arXiv:1809.09875*, 2018.

[25] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[26] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

[27] Wei Gao, Bin-Bin Yang, and Zhi-Hua Zhou. On the resistance of nearest neighbor to random noisy labels. *arXiv preprint arXiv:1607.07526*, 2016.

[28] Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In *International Conference on Machine Learning*, pages 12912–12923. PMLR, 2021.

[29] Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning*, pages 27412–27427. PMLR, 2022.

[30] Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. *Advances in Neural Information Processing Systems*, 27, 2014.

[31] Nick Rittler and Kamalika Chaudhuri. A two-stage active learning algorithm for k-nearest neighbors. *arXiv preprint arXiv:2211.10773*, 2022.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[34] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.