Task-Relevant Active Learning without Prior Knowledge using Vision-Language Models

Usman Irshad Bhatti
Concordia University
Montreal, QC, Canada
2019mc7@student.uet.edu.pk

Ali Ayub
Concordia University
Montreal, QC, Canada
ali.ayub@concordia.ca

***Abstract*—Traditional active learning methods rely on uncertainty-based selection, which assumes the presence of initial labeled data to guide sample selection. These methods also assume that all the unknown data is relevant to the task to be performed by the robot, which is untrue in real world settings. This problem becomes more challenging when robots only know the names of relevant classes for their tasks but have no labeled data, rendering uncertainty sampling and other informative selection strategies infeasible. The only practical method in these situations is random sampling, which results in an ineffective use of labeling resources. As a solution to this problem, we suggest a task-relevant active learning method in which we pre-select samples that are most probably from relevant classes by using embeddings generated by vision-language models and calculating similarity scores between written task descriptions and object images. This minimizes annotation waste by ensuring that the labeled dataset begins with a high percentage of pertinent samples. According to experimental findings, our approach outperforms random sampling and traditional active learning in terms of model performance and relevant data selection, which makes it a workable option for robotic learning systems that have no initial labels and limited data to learn from in the real world.**

I. Introduction and Related Work

Active learning (AL) is a widely used approach to reduce annotation costs by selecting the most informative samples for labeling. Traditional AL strategies, such as uncertainty sampling [1], [2] and query-by-committee, require a small set of initially labeled data to estimate informativeness. Additionally, traditional AL methods assume that all the data in the environment is relevant to the task to be performed by the robot. This assumption, however, is violated in real-world settings. For example, a domestic robot might be tasked with setting up a table for cereal breakfast, and unknown objects, such as a toothbrush, would be irrelevant. Querying users to learn about irrelevant objects wastes labeling budget, increases human teaching load, and reduces model performance.

Recent works in open-set learning [3], [4] have attempted to address the problem of identifying relevant classes in the absence of fully labeled datasets. These methods differentiate between known and unknown categories, hence giving a better chance of the selected sample to be from the relevant classes but they rely on a portion of initially labeled data of all relevant object classes. However, in human-robot interaction (HRI) scenarios, initially robots might know the name of the task and the names of relevant classes but not have labeled data. This absence of initial annotations (known as the *cold start* problem) makes conventional AL and open-set AL methods ineffective, as uncertainty-based or task-relevant selection cannot be performed without prior labeled data. This forces robot to rely on random sampling, leading to inefficient annotation efforts [5], [6].

Several studies have addressed the cold-start problem in AL. One such method is the use of transfer learning [7], where models pretrained on large datasets do the initial sample selection. Another method is self-supervised learning [8], which uses unlabeled data to extract useful representations. However, these methods still require a certain degree of pretraining or domain adaptation, which may not always be feasible in robotic environments.

In this paper, we address the cold-start problem in open-set AL, and propose a novel method that focuses on a filtering mechanism based on vision-language model (e.g., CLIP [9]) embeddings to select the most relevant samples before any labeling occurs. CLIP (Contrastive Language-Image Pretraining) has demonstrated its effectiveness in learning transferable visual representations from natural language supervision. Leveraging CLIP, we compute similarity scores between images and textual descriptions of relevant classes, enabling the robot to pre-select images that are likely to belong to task-relevant classes. By filtering out irrelevant samples before annotation, our method maximizes the utility of the initial labeling budget, ensuring that most labeled samples contribute meaningfully to the learning process.

II. Methodology

Task-relevant AL (TRAL) extends traditional AL by introducing the challenge of distinguishing between *relevant* and *irrelevant* classes within an unlabeled dataset. Given an unlabeled pool $\mathcal{X}_U$, samples may belong to either a predefined closed set of *relevant classes* $\mathcal{Y}_r$ or an open set of *irrelevant classes* $\mathcal{Y}_i$, which contains unknown and potentially novel categories. The goal of TRAL is to train a model $f : \mathcal{X} \rightarrow \mathcal{Y}_r \cup \mathcal{Y}_i$ that can effectively differentiate between these two subsets while learning from the most informative relevant samples.

A fundamental challenge in TRAL is that, unlike standard AL where all queried labels improve the model's understanding of known classes, the presence of unknowns introduces

uncertainty regarding how to define a decision boundary between relevant and irrelevant samples. Ideally, a model should be trained to reject instances from $\mathcal{Y}_i$ while refining its classification on $\mathcal{Y}_r$. However, this raises an important question: **how can a model learn to identify irrelevant classes when it has no labeled examples of what is irrelevant?** This challenge complicates both the selection of informative samples for labeling and the design of effective query strategies in an open-set setting.

A. CLIP-Sim

To address the challenge of identifying relevant samples in an open-set setting without any initial labeled data, we propose a method based on CLIP-based similarity scores which we call CLIP-Sim. Given an unlabeled dataset $\mathcal{X}_U$ and a set of relevant class descriptions $\mathcal{T}_r$, we leverage CLIP embeddings to compute a similarity score for each sample. Let $\phi(x)$ denote the CLIP embedding of an image $x \in \mathcal{X}_U$, and let $\phi(t)$ represent the embedding of a textual description $t \in \mathcal{T}_r$. The relevance score for each image is defined as:

$$S(x) = \max_{t \in \mathcal{T}_r} \cos(\phi(x), \phi(t)) - \frac{1}{|\mathcal{T}_i|} \sum_{t' \in \mathcal{T}_i} \cos(\phi(x), \phi(t')) \quad (1)$$

where $\mathcal{T}_i$ represents a set of textual descriptions of irrelevant classes. The first term ensures that the sample is similar to at least one relevant class, while the second term penalizes similarity to irrelevant classes. Samples with the highest $S(x)$ values are selected for annotation. This process allows us to construct an informative labeled dataset while minimizing annotation costs, ensuring that the majority of labeled samples belong to relevant categories for training a classifier.

We note that the embedding $\phi(t)$ is not just the text embedding of a class label. Simply using the class labels forces the model to select homogeneous set of images that are closely related to a single text embedding for a class. This reduces the diversity and the informativeness of the samples. Instead, we generated a variety of textual descriptions for each object to improve the diversity and informativeness of the relevant samples. For example, an object class label might be "Cat". For this class we generate the descriptions: "A photo of a cat walking", "A photo of a cat laying down", "A cat is sleeping", etc. Additionally, the model does not have access to any irrelevant class labels or data in the beginning, therefore, the second term in eq. (1) is empty. To avoid that, we choose N (a hyperparameter) number of irrelevant class labels randomly at the beginning from a large open set of classes. However, the model encounters irrelevant classes in subsequent AL rounds because CLIP-Sim does not lead to 100% relevant samples selected in a round. This helps CLIP-Sim to use the irrelevant classes in the actual samples to improve relevant data selection.

III. Experiments

A. Dataset and Textual Representations

We implement our method on the commonly used CIFAR-10 dataset [10] consisting of 10 classes. The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. Among the 10 classes, 5 randomly chosen classes are selected as relevant and other 5 as irrelevant, however, the model doesn't know about these irrelevant classes yet as it assumes that it is an open set. The 5 relevant class labels are then used to generate different text descriptions. Each class has 15 different text representations in our experiments. We use $N = 3$ for the initial irrelevant classes in our experiments

B. Model Training and Evaluation

A ResNet-18 model is trained using the selected dataset in each AL round. The training set images are used in the sample selection using CLIP-Sim and then the samples labeled as relevant are used to train the ResNet-18. The ResNet-18 is used as a classifier between the 5 relevant classes. Performance is compared against the same ResNet-18 trained on random samples as well as sample selected using uncertainty sampling [1].For the uncertainty sampling method, it is necessary to have a model trained on some initial examples. Therefore, we randomly choose images that make up half of the budget, have them labeled, train the model on these labeled images, and then use uncertainty sampling to select the remaining half of the budget images.

Number of images that belong to relevant classes after being labeled and the classification accuracy on the relevant classes of the CIFAR-10 test set are used to evaluate the effectiveness of the selection strategies.

IV. Results and Discussion

Figure 1 shows the relationship between the number of samples selected and the training budget for three selection methods: random sampling, CLIP-Sim, and uncertainty sampling. Our results indicate that CLIP-Sim consistently selects more relevant samples compared to other methods, demonstrating the advantage of using CLIP-based embeddings for sample filtering.

Fig. 1. Selected Samples vs Training Budget. CLIP-Sim significantly outperforms other methods in terms of identifying relevant samples.

Figure 2 shows the relationship between the test set accuracy and the training budget for three selection methods. Our

results indicate that CLIP-Sim performs significantly better than uncertainty sampling and random sampling in the start but random sampling becomes competitive in larger budgets. This is due to less diversification at larger budgets where 15 textual representations for each class are not enough and the learning also plateaus at that point. Uncertainty samplings seems to get better with larger budgets but still lags behind. These results indicate that CLIP-based similarity scores provide a practical alternative to traditional AL approaches, especially in scenarios where no prior labeled data exists. Particularly, CLIP-Sim deals with the cold-start problem which provides a significant advantage over random and uncertainty sampling in the initial AL rounds. These results also show that in the TRAL setting, traditional AL methods perform worse than random sampling. Therefore, methods for selecting relevant as well as informative samples must be developed for real-world robotics applications.

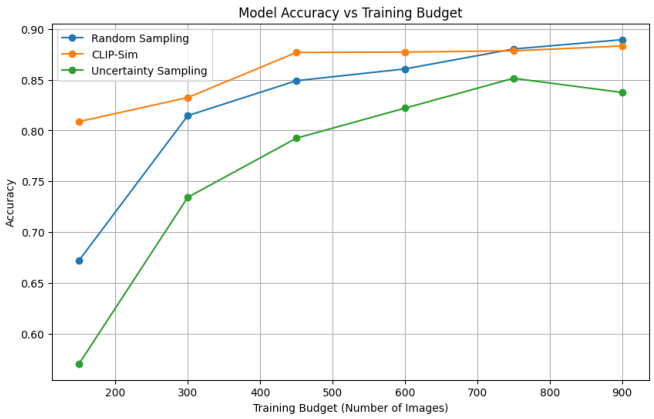

Fig. 2. Model Accuracy vs Training Budget. CLIP-Sim significantly outperforms other methods in the start at low budgets.

V. Conclusion

We introduce a task-relevant active learning method that uses CLIP embeddings for efficient sample selection in human-robot interaction tasks. Unlike other AL methods, which require initial labeled data for distinguishing between relevant and irrelevant classes, our method filters samples solely based on similarity scores, ensuring a higher proportion of relevant labeled samples. This reduces wasted annotations and improves the model performance in constrained labeling scenarios, particularly when robots must learn from limited interactions with human users. Future work could explore adaptive budget allocation and integration with uncertainty-based methods for further refinements. Multi-Modal models that can generate descriptions directly using the image itself can also be used for more accurate descriptions instead of randomly generating the textual representations of each class. We also only focused on the assumption that relevant labels are already known. Future works can ignore this assumption and work from a single task description rather than relevant object classes for a task.

References

[1] B. Settles, "Active learning literature survey," *University of Wisconsin-Madison*, 2009.

[2] A. Ayub and C. Fendley, "Few-shot continual active learning by a robot," *Advances in Neural Information Processing Systems*, vol. 35, pp. 30 612–30 624, 2022.

[3] D. Park, Y. Shin, J. Bang, Y. Lee, H. Song, and J.-G. Lee, "Meta-query-net: Resolving purity-informativeness dilemma in open-set active learning," in *NeurIPS '21*, 2021.

[4] K.-P. Ning, X. Zhao, Y. Li, and S.-J. Huang, "Active learning for open-set annotation," in *AAAI '22*, 2022.

[5] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *SIGIR '94*, 1994, pp. 3–12.

[6] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," in *ICML '01*, 2001, pp. 999–1006.

[7] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[8] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *ICCV '19*, 2019, pp. 1476–1485.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML '21*, 2021.

[10] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009, technical report, University of Toronto.