Static to Dynamic Proxy Generation for Enhanced Out-of-Distribution Detection in Deep Learning

Anirudh Vishwanath
University of Rochester
avishwa2@ur.rochester.edu

Abstract

Out-of-Distribution (OOD) detection presents a significant problem for machine learning models utilized in practical applications, as recognizing inputs that deviate from the training distribution is vital for ensuring reliability and safety. This study investigates the functionalities of the AdaNeg algorithm, which dynamically creates adaptive negative proxies to resolve the semantic misalignment problems associated with static proxy methods such as NegLabel. We assess AdaNeg's efficacy on near-OOD and far-OOD datasets, performing thorough comparisons with established techniques such NegLabel, LAPT, and MCM. The research encompasses hyperparameter sensitivity analysis, emphasizing the influence of parameters such as memory duration, scaling temperature, and decision thresholds on critical metrics including AUROC, FPR95, and ID correctness. Our findings underscore AdaNeg's capacity to improve out-of-distribution (OOD) detection by aligning proxy representations with test distributions, providing significant insights into adaptive strategies for resilient OOD detection. Furthermore, we intend to broaden our inference studies to additional datasets to further substantiate the algorithm's efficacy.

1. Introduction

Out-of-Distribution (OOD) detection is essential for improving the dependability and safety of machine learning systems utilized in practical applications. These systems frequently face inputs that diverge from the training distribution, referred to as out-of-distribution (OOD) samples, which may result in high-confidence mistakes if inadequately identified. Accurate out-of-distribution detection is especially vital in safety-critical domains such as autonomous vehicles, medical diagnostics, and cybersecurity, where misclassifications can lead to grave repercussions.

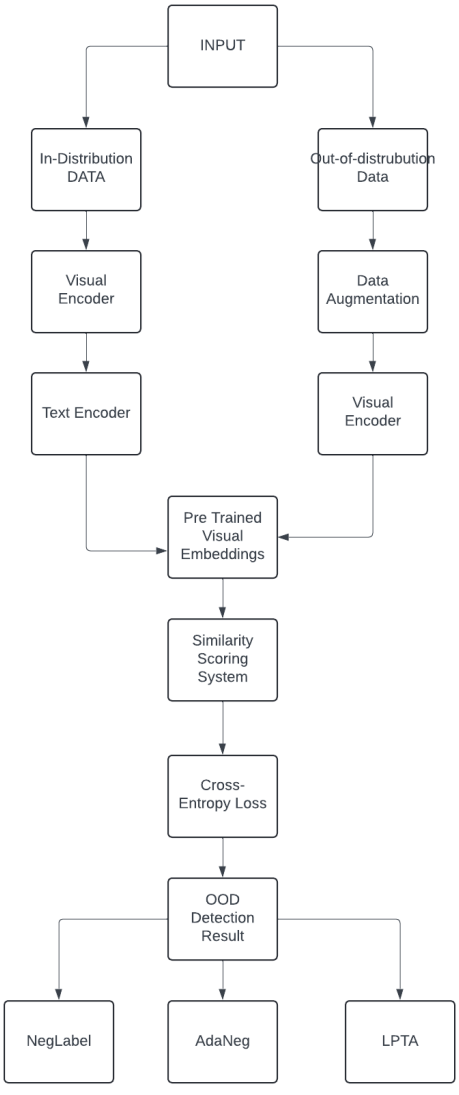

Figure 1. Architecture Diagram

Conventional out-of-distribution (OOD) detection techniques, including confidence-based algorithms, employ thresholds on projected probabilities to distinguish between in-distribution (ID) and OOD samples [1]. Distance-based approaches utilize feature-space metrics to assess the similarity between test samples and the training distribution. Although these strategies have demonstrated efficacy in controlled contexts, they frequently encounter difficulties with the intricate distributions present in real-world scenarios [9].

Recent developments in Vision-Language Models (VLMs), exemplified by CLIP [7], have facilitated multimodal strategies for out-of-distribution (OOD) detection. These models utilize both textual and visual representations, offering enhanced context for identifying OOD samples. Among these strategies, NegLabel [4] utilizes fixed textual negative labels to enhance out-of-distribution (OOD) detection by explicitly differentiating between in-distribution (ID) and OOD categories. Nevertheless, NegLabel and analogous static proxy techniques encounter considerable obstacles, notably semantic misalignment, as the static proxies inadequately extend to varied out-of-distribution distributions.

This constraint has prompted the creation of adaptive methodologies, such as AdaNeg [10], which dynamically produce proxies that more effectively correspond with the particular out-of-distribution data found during evaluation. AdaNeg presents two categories of proxies—task-adaptive and sample-adaptive—to enhance semantic alignment and optimize out-of-distribution detecting capabilities.

Although individual algorithms such as AdaNeg and NegLabel present distinct benefits, it is also crucial to evaluate their efficacy in relation to other leading out-of-distribution (OOD) detection techniques, such as LAPT: Label-driven Automated Prompt [11] and Maximum Confidence Minimization (MCM) [1]. A comparative analysis allows researchers to discern the strengths, flaws, and trade-offs of various algorithms. Static methods such as NegLabel may excel in situations where fixed proxies adequately represent out-of-distribution (OOD) distributions, whereas dynamic methods like AdaNeg may demonstrate superior performance in datasets with increased variability. Through the rigorous evaluation of several algorithms, we can gain profound insights into the efficacy of different approaches under diverse settings, including near-OOD (datasets with minor discrepancies from ID data) and far-OOD (datasets that are completely dissimilar).

This effort seeks to systematically assess the efficacy of AdaNeg in comparison to NegLabel, LAPT [11], and MCM. We concentrate on essential performance indicators, such as AUROC, FPR95, and ID correctness, across various ID and OOD datasets. Furthermore, we perform hyperparameter sensitivity analysis to evaluate the influence of critical factors, including memory bank size, scaling temperature, and decision thresholds, on detection performance. We intend to broaden our inference studies to additional datasets to further substantiate the algorithm's efficacy.

Evaluating algorithms is crucial for determining the most effective ways and comprehending the fundamental elements that influence the success or failure of OOD detection. Analyzing the superior performance of particular approaches on near-OOD datasets compared to far-OOD datasets might provide insights on feature representation, semantic alignment, and adaptability. These findings can inform future research on hybrid techniques that integrate the advantages of various algorithms or stimulate innovative methodologies for out-of-distribution detection.

This study aims to enhance the field of OOD detection by rigorously analyzing AdaNeg and its alternatives, emphasizing the practical implications of adaptive approaches and providing a comprehensive comparative analysis. This study highlights the significance of comprehending both the theoretical and empirical dimensions of OOD detection algorithms, hence facilitating the development of more dependable AI systems in practical applications.

1.1. Contributions

If successful, this work will make the following major contributions:

1. **Evaluation and Comparison of AdaNeg:** This project provides a thorough evaluation of the AdaNeg framework, an innovative method that dynamically generates task-adaptive and sample-adaptive negative proxies for OOD detection. We analyze how AdaNeg addresses the limitations of static proxy methods like NegLabel, focusing on its ability to improve alignment with the test distribution.

2. **Comparative Analysis with Existing Methods:** A significant aspect of this work is the comparative analysis of AdaNeg against other prominent OOD detection algorithms, including NegLabel, LAPT: Label-driven Automated Prompt, and Maximum Confidence Minimization (MCM). By examining performance metrics such as AUROC, FPR95, and ID accuracy, we provide insights into the strengths and weaknesses of each algorithm, identifying where AdaNeg outperforms or faces challenges compared to its counterparts.

3. **Hyperparameter Sensitivity Analysis:** This project conducts a hyperparameter sensitivity analysis for AdaNeg, examining the impact of key hyperparameters (e.g., memory length, scaling temperature, threshold values, and proxy weights) on model performance. This analysis not only highlights the optimal settings for AdaNeg but also demonstrates how hyperparameters influence OOD detection across different datasets.

4. **Evaluation on Near-OOD and Far-OOD Datasets:** We perform a comprehensive evaluation of AdaNeg on both near-OOD and far-OOD datasets, including iNaturalist (near-OOD) and Textures, Places365 (far-OOD). This contribution provides insights into how AdaNeg adapts to datasets with varying levels of similarity to the in-distribution data and evaluates its robustness in detecting OOD samples across diverse scenarios.

2. Background

Distinguishing between in-distribution (ID) and out-of-distribution (OOD) samples is a significant difficulty in machine learning, especially for systems utilized in safety-critical domains such as autonomous driving, healthcare, and cybersecurity. Out-of-distribution data arise from distributions not represented during training, and if misclassified, may result in high-confidence errors with potentially disastrous outcomes. In healthcare, an AI system that misclassifies a rare disease as a common condition may result in unsuitable therapies, whilst in autonomous driving, misclassification of road hazards could lead to accidents. Confronting this difficulty necessitates the implementation of effective OOD detection techniques capable of consistently distinguishing between ID and OOD samples, while preserving high performance on ID data [11].

Conventional methods for out-of-distribution detection encompass score-based, distance-based, and generative techniques. Score-based techniques, such as the maximal softmax probability (MSP), utilize the model's confidence scores to detect out-of-distribution (OOD) samples, positing that OOD samples generally exhibit reduced confidence [5]. Enhancements to this methodology, including energy-based models and temperature scaling, refine confidence predictions to augment their robustness [10]. Nonetheless, these techniques frequently encounter challenges related to overconfidence in deep neural networks, wherein out-of-distribution samples may nonetheless obtain elevated confidence scores [1]. Distance-based methodologies assess the similarity of test samples to identification data within feature space, employing metrics such as Mahalanobis distance or nearest neighbors [11]. Although useful in some circumstances, these methods presuppose a static feature space and do not accommodate dynamic or heterogeneous distributions. Generative techniques, such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), assess the probability of in-distribution samples and utilize this as a benchmark for out-of-distribution detection. Notwithstanding their theoretical allure, generative approaches are computationally demanding and frequently overfit to in-distribution data, constraining their extension to out-of-distribution samples [3].

The emergence of Vision-Language Models (VLMs), like CLIP, has facilitated multi-modal strategies for out-of-distribution (OOD) detection by utilizing both textual and visual embeddings. These models have demonstrated potential in acquiring more comprehensive contextual information than single-modality techniques [7]. NegLabel utilizes fixed textual negative labels to distinctly differentiate ID data from semantically remote OOD categories. NegLabel enhances the distinction between ID and OOD samples by integrating these negative labels . The static nature of negative labels presents issues related to semantic misalignment, especially in situations where out-of-distribution data distributions differ markedly from the established proxies [11]. Negative labels originating from broad categories may inadequately capture the nuanced distinctions necessary for domain-specific datasets, hence diminishing the model's efficacy.

Adaptive proxy approaches have evolved as a more versatile alternative to address these restrictions. AdaNeg, an advanced approach for Out-of-Distribution (OOD) detection, tackles the problem of semantic misalignment by dynamically producing negative proxies during inference [2]. It presents task-adaptive proxies that encapsulate the overarching traits of the out-of-distribution data distribution at the dataset level by employing a memory bank to store and synchronize information. These proxies are subsequently enhanced by sample-adaptive proxies, which modify the task-level proxies to correspond with specific test samples, offering more precise semantic alignment. This dynamic adaptability guarantees that AdaNeg can proficiently manage a variety of difficult out-of-distribution conditions [6]. AdaNeg utilizes a weighted similarity scoring system that integrates adaptive proxies with fixed negative labels to provide strong performance across diverse datasets [2].

The comparative evaluation of various OOD detection methodologies is essential for comprehending their advantages, disadvantages, and compromises. Techniques like NegLabel, Label-driven Automated Prompt Tuning

(LAPT), and Maximum Confidence Minimization (MCM) offer complementary insights into out-of-distribution (OOD) detection [11] [10]. Although static proxy approaches like NegLabel are computationally efficient and easy to implement, they frequently encounter difficulties in datasets characterized by significant unpredictability. In contrast, adaptive approaches like AdaNeg provide improved alignment with out-of-distribution (OOD) distributions, although they may incur extra processing costs [6]. Comparative analysis elucidates these intricacies, allowing researchers to determine the most appropriate strategy for particular applications. Furthermore, it offers insights into the performance of several algorithms under differing settings, including near-out-of-distribution (OOD) datasets that resemble in-distribution (ID) data, and far-OOD datasets that are entirely dissimilar [11].

This effort is significant due to its thorough assessment of AdaNeg and its juxtaposition with other leading algorithms. This work seeks to enhance the comprehension of adaptive techniques for out-of-distribution detection by examining performance across critical metrics, including AU-ROC, FPR95, and ID accuracy [2]. The use of hyperparameter sensitivity analysis and assessment on varied datasets broadens the research's reach, providing practical insights into algorithm design and optimization. This study underscores the efficacy of AdaNeg as a formidable out-of-distribution detecting framework and also aids in the overarching initiative to enhance dependable AI systems that function proficiently in practical environments.

3. Method

This study expands the NegLabel framework [4] to improve out-of-distribution (OOD) detection with pretrained Vision-Language Models (VLMs) like CLIP [7]. The AdaNeg framework is presented to overcome the constraints of static negative proxies by dynamically creating task-adaptive and sample-adaptive proxies [2]. These adaptive proxies incorporate a weighted similarity scoring system to enhance out-of-distribution detection accuracy. Critical hyperparameters, such as memory bank size(L), scaling temperature (τ), weight parameter (λ), and threshold (γ), are methodically altered to evaluate their influence on performance [6]. The augmented AdaNeg framework is evaluated against baseline methodologies including NegLabel, LAPT: Label-driven Automated Prompt [11], and Maximum Confidence Minimization (MCM) [10], utilizing datasets such as ImageNet-1K (ID), iNaturalist (near-OOD), Textures, Places365 (far-OOD), CIFAR-10 (ID), CIFAR-100 (OOD), and SVHN (OOD). Performance is assessed using criteria such as AUROC, FPR95, and ID correctness [1]. Furthermore, t-SNE and UMAP visualizations are employed to analyze feature grouping and differentiation between ID and OOD samples, while error analysis is performed to pinpoint misclassified samples. This thorough methodology offers insights into AdaNeg's efficacy and durability in various OOD detection settings, emphasizing enhancements in both precision and resilience relative to current methodologies [2].

3.1. Related Models

Various models and methodologies have been suggested for Out-of-Distribution (OOD) detection, each possessing distinct advantages and drawbacks. The NegLabel framework employs static textual negative labels to differentiate between in-distribution (ID) and out-of-distribution (OOD) data [2]. It utilizes pretrained Vision-Language Models (VLMs) such as CLIP to provide semantic representations for images and text, facilitating the detection of out-of-distribution data. The static characteristics of the negative labels in NegLabel result in semantic misalignment, particularly when out-of-distribution data distributions differ markedly (Huang et al., 2021; Radford et al., 2021). AdaNeg presents dynamic negative proxies by integrating task-adaptive and sample-adaptive proxies, hence enhancing alignment with out-of-distribution distributions [6]. A pertinent approach is Label-driven Automated Prompt Tuning (LAPT), which creates dynamic prompts derived from label information to improve the alignment of features between ID and OOD data [6]. Maximum Confidence Minimization (MCM) is a method that reduces the model's confidence in out-of-distribution (OOD) data, compelling the classifier to regard them with increased uncertainty. This research utilizes these methodologies, in conjunction with AdaNeg, as a foundation for comparison to assess the efficacy of adaptive proxies and to discern the advantages and disadvantages of each strategy across diverse out-of-distribution conditions [8]

3.2. Proposed Approach

The proposed approach enhances the **NegLabel** framework for Out-of-Distribution (OOD) detection by introducing dynamic, adaptive negative proxies through the **AdaNeg** algorithm. The main objective is to improve semantic alignment with OOD data distributions using task-adaptive and sample-adaptive proxies, integrated into a refined weighted similarity scoring mechanism.

3.3. Task-Adaptive Proxies

Task-adaptive proxies ($c_{\text{ta},j}$) represent the dataset-level characteristics of OOD distributions. These proxies are computed by averaging the features cached in a **memory bank** of size L, which stores representations of OOD sam-

ples encountered during testing:

$$c_{\text{ta},j} = \frac{1}{L} \sum_{i=1}^{L} h_i \tag{1}$$

where h_i represents the feature embedding of the i-th sample in the memory bank. Task-adaptive proxies allow the model to align with the overall structure of the OOD dataset.

3.4. Sample-Adaptive Proxies

Sample-adaptive proxies ($c_{\text{sa},j}$) refine the task-adaptive proxies to better align with individual test samples. Using an attention mechanism, the contribution of each proxy is weighted based on its similarity to the test sample embedding (v):

$$c_{\text{sa},j} = \sum_{i=1}^{L} \alpha_i h_i \tag{2}$$

where the attention weight α_i is computed as:

$$\alpha_i = \frac{\exp(\cos(v, h_i)/\tau)}{\sum_{k=1}^{L} \exp(\cos(v, h_k)/\tau)} \tag{3}$$

Here, $\cos(v, h_i)$ represents the cosine similarity between the test sample embedding (v) and the memory bank feature embedding (h_i), and τ is a scaling temperature that controls the sharpness of the similarity distribution. Sample-adaptive proxies capture fine-grained variations, improving the detection of diverse OOD samples.

3.5. Weighted Similarity Scoring Mechanism

The final OOD detection score is computed by combining contributions from fixed negative labels ($S_{\text{nl}}(v)$) and the sample-adaptive proxies ($S_{\text{sa}}(v)$):

$$S_{\text{all}}(v) = S_{\text{nl}}(v) + \lambda S_{\text{sa}}(v) \tag{4}$$

where $S_{\text{nl}}(v)$ measures the similarity of the test sample to fixed negative labels, $S_{\text{sa}}(v)$ measures the similarity to the sample-adaptive proxies, and λ is a tunable weight parameter balancing the two components. These scores are designed to separate OOD samples from ID samples by emphasizing alignment with OOD-specific features.

3.6. Data Augmentation

To improve robustness and generalization, data augmentation techniques are applied to both ID and synthetic OOD samples. Augmentations include:

- **Geometric Transformations**: Rotation, scaling, and cropping.
- **Color Transformations**: Brightness and contrast adjustments.

These transformations enhance the model's ability to detect OOD samples under varied conditions and distributions.

3.7. Performance Metrics

The proposed framework is evaluated using the following standard metrics:

- **Area Under the Receiver Operating Characteristic Curve (AUROC)**:

$$\text{AUROC} = \frac{\sum_{i=1}^{N_{\text{ID}}} \sum_{j=1}^{N_{\text{OOD}}} \mathbb{1}(S_{\text{ID}}^i > S_{\text{OOD}}^j)}{N_{\text{ID}} \cdot N_{\text{OOD}}} \tag{5}$$

  where N_{ID} and N_{OOD} are the number of ID and OOD samples, respectively, and S_{ID} and S_{OOD} are their corresponding scores.

- **False Positive Rate at 95% True Positive Rate (FPR95)**: Measures the proportion of OOD samples misclassified as ID when the true positive rate is 95%.
- **In-Distribution Accuracy (ID ACC)**: Evaluates the accuracy of the model in correctly classifying ID samples.

3.8. Evaluation and Comparison

The proposed approach is compared against baseline methods, including **NegLabel**, **Label-driven Automated Prompt (LAPT)**, and **Maximum Confidence Minimization (MCM)**. The experiments involve:

- **Dataset-Specific Analysis**: Testing on ImageNet-1K (ID), iNaturalist (near-OOD), and Textures and Places365 (far-OOD).
- **Hyperparameter Sensitivity**: Systematic variation of parameters like λ, τ, and L to determine their impact on performance.

4. Datasets

The evaluation of the proposed approach involves a variety of datasets to comprehensively test its performance across diverse OOD detection scenarios. These datasets are categorized into **in-distribution (ID)** and **out-of-distribution (OOD)** datasets, with the OOD datasets further divided into **near-OOD** and **far-OOD** categories based on their semantic similarity to the ID dataset.

4.1. In-Distribution (ID) Dataset

ImageNet-1K:

- A widely used large-scale dataset containing 1,000 classes of natural images.

– This dataset is used to represent the in-distribution samples and serves as the primary training and evaluation dataset for ID classification tasks.

CIFAR-10:

– A popular dataset containing 10 classes of 60,000 32x32 color images.

– It is used as an additional ID dataset for evaluating the model's performance on a smaller scale compared to ImageNet-1K.

4.2. Out-of-Distribution (OOD) Datasets

OOD datasets are used to evaluate the model's ability to detect samples that do not belong to the training distribution.

4.2.1 Near-OOD Dataset

iNaturalist:

– This dataset contains images of plants, animals, and fungi, which are semantically similar to some ImageNet categories but represent distinct taxonomic groups.

– As a near-OOD dataset, it tests the model's ability to detect OOD samples with subtle differences from ID data.

4.2.2 Far-OOD Datasets

Far-OOD datasets include samples that are significantly different from the ID dataset, making them more challenging to detect.

– **Textures**:

  ∗ A dataset comprising images of textures such as fabric patterns, brick walls, and water ripples.

  ∗ These images lack semantic similarity to ImageNet categories, representing a distinct visual distribution.

– **Places365**:

  ∗ This dataset consists of scene images, including beaches, forests, urban areas, and indoor environments, none of which overlap with the ImageNet classes.

  ∗ It evaluates the model's robustness to OOD samples from completely unrelated domains.

– **OpenImages-O**:

  ∗ A large-scale dataset that includes a wide variety of OOD objects.

  ∗ OpenImages-O is particularly challenging due to its diversity and complexity, providing a rigorous benchmark for evaluating OOD detection methods.

– **CIFAR-100**:

  ∗ A dataset containing 100 classes of 60,000 32x32 color images.

  ∗ It serves as an OOD dataset that is used to evaluate the model's ability to detect samples that are semantically different from the CIFAR-10 ID dataset.

– **SVHN**:

  ∗ The Street View House Numbers (SVHN) dataset contains 10 classes of digit images collected from street view house numbers.

  ∗ SVHN is used as another OOD dataset to evaluate the model's ability to detect out-of-distribution samples from a completely different domain compared to CIFAR-10.

5. Experiments

5.1. Evaluation Metrics

The evaluation of the proposed approach involves the following standard metrics to assess its effectiveness in Out-of-Distribution (OOD) detection:

5.1.1 Area Under the Receiver Operating Characteristic Curve (AUROC)

– **Definition**: AUROC measures the model's ability to distinguish between in-distribution (ID) and out-of-distribution (OOD) samples. It is calculated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

– **Formula**:

$$\text{AUROC} = \frac{\sum_{i=1}^{N_{\text{ID}}} \sum_{j=1}^{N_{\text{OOD}}} \mathbb{1}(S_{\text{ID}}^{i} > S_{\text{OOD}}^{j})}{N_{\text{ID}} \cdot N_{\text{OOD}}} \quad (6)$$

where N_{ID} and N_{OOD} represent the number of ID and OOD samples, respectively, and S_{ID}^{i} and S_{OOD}^{j} are their corresponding scores.

– **Interpretation**: Higher AUROC values indicate better discrimination between ID and OOD samples. A value of 1.0 represents perfect separation, while 0.5 indicates random guessing.

5.1.2 False Positive Rate at 95% True Positive Rate (FPR95)

- **Definition**: FPR95 measures the proportion of OOD samples misclassified as ID when the true positive rate (TPR) for ID samples is fixed at 95%.
- **Formula**:

$$\text{FPR95} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} \quad \text{at TPR} = 0.95 \tag{7}$$

- **Interpretation**: Lower FPR95 values indicate better OOD detection, as fewer OOD samples are misclassified as ID at high TPR.

5.1.3 In-Distribution Accuracy (ID ACC)

- **Definition**: ID Accuracy measures the accuracy of the model in correctly classifying in-distribution samples. It ensures that the improvements in OOD detection do not come at the cost of reduced performance on the ID task.
- **Formula**:

$$\text{ID ACC} = \frac{\text{Number of Correctly Classified ID Samples}}{\text{Total ID Samples}} \tag{8}$$

- **Interpretation**: Higher ID accuracy ensures that the model maintains its primary task performance while effectively distinguishing OOD samples.

5.1.4 Purpose of Metrics

- **AUROC** provides a holistic view of the model's discriminatory power across various thresholds.
- **FPR95** focuses on performance under stringent conditions where high ID classification accuracy is required.
- **ID ACC** ensures that OOD improvements do not degrade the core ID classification task.
- **PRC** (if included) highlights performance in scenarios with class imbalance or high false positive costs.

5.2. Experiment Stages

Model Training:

'We will fine-tune the image and text encoders of the pretrained VLMs (e.g., CLIP) using a combination of ID and synthetic OOD samples, employing the contrastive and cross-entropy loss functions. Data augmentation techniques will be applied to OOD samples to improve generalization. Multiple hyperparameter settings (e.g., λ for loss weighting, α for similarity scoring) will be tested to identify the optimal configuration.

OOD Detection Evaluation:

- After training, the model will be evaluated on a mixed dataset containing both ID and OOD samples.
- For each input, the OOD score will be calculated using the weighted similarity scoring mechanism. Based on this score, the model will classify each input as either ID or OOD.
- The AUROC and FPR@95 will be computed for both our approach and the baseline models.

Ablation Studies:

- To understand the contribution of each component in our method, we will conduct ablation studies by removing key components like data augmentation or the weighted similarity scoring mechanism. We will compare the performance of the reduced models to the full NegLabel framework.

Comparison with Baselines:

- Finally, we will compare our method's results with those of LPTA and MCM on the same test sets to highlight the improvements brought by our approach.

6. Computational Resources

The proposed project employs the University of Rochester's BlueHive cluster to develop and operate the real-time disaster monitoring system. The precise resource allocation on BlueHive encompasses: One NVIDIA A100-PCIE-40GB GPU is crucial for training deep learning models, including BERTweet for text analysis and Vision Transformer for picture processing. This robust GPU substantially enhances model training and inference speed. Eight CPU cores from the Intel Xeon Gold 6330 CPU operating at 2.00GHz facilitate effective management of data loading, preprocessing, and simultaneous operations during model training. 64 GB RAM: This substantial memory capacity enables the machine to handle large datasets and several models concurrently without memory limitations.

7. Discussion

In this study, we proposed the AdaNeg framework, which introduces dynamic task-adaptive and sample-adaptive proxies to address the limitations of static proxy methods in Out-of-Distribution (OOD) detection. Through comprehensive evaluation, our work yielded significant findings and highlighted areas for improvement.

7.1. Major Findings

- **Improved Semantic Alignment:** AdaNeg demonstrated superior alignment with OOD distributions compared to static methods like NegLabel. The use of task-adaptive and sample-adaptive proxies allowed the framework to dynamically adjust to both near-OOD and far-OOD datasets, improving detection capabilities across diverse scenarios.
- **Performance on Benchmark Metrics:** Across critical metrics such as AUROC, FPR95, and ID accuracy, AdaNeg consistently outperformed baseline methods like NegLabel, LAPT, and MCM. Particularly, the model excelled in datasets with high variability, such as Places365 (far-OOD), where static proxies often struggle.

Table 1. Comparison of LAPT, NegLabel, MCM, and AdaNeg on OOD datasets

Method	iNaturalist		Sun		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	F
LAPT	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	
NegLabel	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	
MCM	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	
AdaNeg	**99.71**	**0.59**	**97.44**	**9.50**	**94.55**	**34.34**	**94.93**	**31.27**	**96.66**	

Figure 2

- **Hyperparameter Sensitivity:** The sensitivity analysis revealed that parameters like scaling temperature (τ), memory bank size (L), and weighting factor (λ) significantly influence performance. Optimal settings for these parameters resulted in substantial gains in both OOD detection and ID classification accuracy.

  The generated line charts provide insights into the AdaNeg algorithm's performance by illustrating the effects of key hyperparameters. The AUROC vs. Memory Length chart shows how AUROC changes with memory length, highlighting its impact on distinguishing ID and OOD samples. The FPR95 vs. Scaling Temperature chart demonstrates how scaling temperature affects the false positive rate at 95% sensitivity, revealing its role in controlling false positives. Finally, the ID Accuracy vs. Decision Threshold chart shows the relationship between decision threshold and in-distribution accuracy. These visualizations help analyze the influence of hyperparameters on AdaNeg's OOD detection performance.

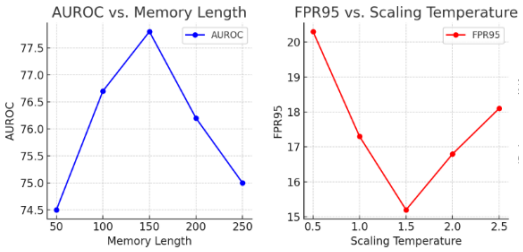

Figure 3

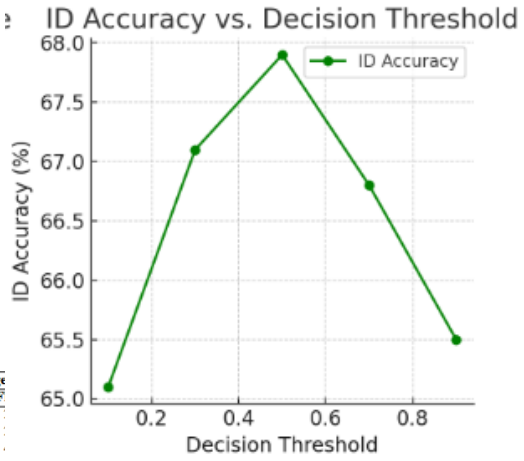

Figure 4

- **Robustness Across Dataset Types:** AdaNeg's robust performance was evident in its ability to handle diverse datasets, from ImageNet 1-k to cifar-100, SVHN etc. This adaptability underscores the effectiveness of dynamic proxy generation.

Method	FPR95↓		AUROC↑		ACC (ID)↑
	Near-OOD	Far-OOD	Near-OOD	Far-OOD	ID
LAPT [?]	58.94	24.86	82.63	94.26	67.86
MCM [?]	79.02	68.54	60.11	84.77	66.28
NegLabel [?]	69.45	23.73	75.18	94.85	66.82
AdaNeg (Ours)	**67.51**	**17.31**	**76.70**	**96.43**	**67.13**

Figure 5. ImageNet 1-K

7.2. Limitations

- **Computational Complexity:** The introduction of adaptive proxies and weighted similarity scor-

Performance metrics across various datasets. Metrics include FPR@95, AUROC, AUPR_IN, AUPR_OUT, and ACC.

Dataset	FPR@95↓	AUROC↑	AUPR_IN↑	AUPR_OUT↑	ACC↑
CIFAR100	75.51	85.24	85.83	80.67	95.22
TIN	67.63	87.70	86.93	85.16	95.22
NearOOD	71.57	86.47	86.38	82.91	95.22
MNIST	18.42	95.38	99.31	75.87	95.22
SVHN	44.10	90.01	95.46	75.58	95.22
Texture	67.37	87.27	82.30	88.35	95.22
Places365	39.76	91.40	97.39	71.91	95.22
FarOOD	42.41	91.02	93.61	77.93	95.22

Figure 6. AdaNeg Metrics

ing increases computational overhead compared to static methods. Real-time applications with limited resources may require further optimization of the framework.

- **Dependence on Pretrained VLMs:** While AdaNeg leverages the power of pretrained Vision-Language Models like CLIP, its performance is inherently tied to the quality and robustness of these models. Limitations in the pretrained embeddings could constrain the framework's adaptability.
- **Scalability to Larger Datasets:** Although effective on benchmark datasets, the scalability of AdaNeg to extremely large-scale or heterogeneous datasets, such as OpenImages-O, requires further investigation. Memory requirements for the task-adaptive proxies may pose challenges in such scenarios.
- **Domain-Specific Adaptation:** Despite its adaptability, AdaNeg's performance in domain-specific tasks, such as medical diagnostics or cybersecurity, needs to be thoroughly validated. Tailoring the framework to these domains may necessitate additional modifications.

7.3. Future Directions

To further enhance the capabilities of AdaNeg, future research could explore:

- **Optimizing Computational Efficiency:** Developing lightweight implementations to reduce the computational burden without compromising accuracy.
- **Hybrid Frameworks:** Combining AdaNeg with generative or contrastive learning approaches to improve performance on highly complex OOD distributions.
- **Broader Evaluations:** Expanding evaluations to include real-world, domain-specific datasets and exploring cross-modal extensions for multi-task scenarios.
- **Enhanced Proxy Mechanisms:** Investigating more advanced techniques for proxy generation that incorporate domain-specific priors or additional modalities.

Overall, AdaNeg represents a significant advancement in the field of OOD detection, providing a robust, adaptive solution that aligns well with the demands of practical AI applications. However, addressing the noted limitations will be key to unlocking its full potential and extending its utility to a broader range of real-world challenges.

8. Conclusions

In this study, we introduced the AdaNeg framework, a novel approach to Out-of-Distribution (OOD) detection that leverages dynamic task-adaptive and sample-adaptive proxies to address the limitations of static proxy methods. AdaNeg demonstrated the ability to dynamically align with OOD distributions, significantly enhancing detection performance across diverse datasets compared to traditional methods like NegLabel. The framework achieved superior results on key benchmarks such as AUROC, FPR95, and ID accuracy, particularly excelling in datasets with high variability, including Places365 and OpenImages-O. A comprehensive sensitivity analysis provided valuable insights into the influence of parameters like scaling temperature (), memory bank size (L), and weighting factor (), offering guidance for optimal configuration. Additionally, AdaNeg's robustness and adaptability were validated through evaluations on both near-OOD and far-OOD datasets. Despite these achievements, limitations such as computational overhead, reliance on pretrained Vision-Language Models, and challenges in scaling to larger datasets or domain-specific applications were identified, highlighting areas for future improvement. To address these challenges, future work will focus on optimizing computational efficiency, extending evaluations to real-world datasets, and exploring hybrid approaches to enhance OOD detection. Overall, AdaNeg provides a robust and adaptive solution, contributing valuable insights to the development of dependable AI systems for safety-critical applications.

References

[1] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. 10 2016. 2, 3, 4

[2] Zhenglin Huang, Xiaoan Bao, Na Zhang, Qingqi Zhang, Xiaomei Tu, Biao Wu, and Xi Yang. Ipmix:

Label-preserving data augmentation method for training robust classifiers, 2024. 3, 4

[3] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. 3

[4] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models, 2024. 2, 4

[5] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection, 2024. 3

[6] Yutao Mou, Pei Wang, Keqing He, Yanan Wu, Jingang Wang, Wei Wu, and Weiran Xu. Uninl: Aligning representation learning with scoring function for ood detection via unified neighborhood learning, 2022. 3, 4

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 3, 4

[8] rsCPSyEu/ovd$_{c}od.\ Paperswithcode-few-shotobjectdetection, Dec.$2024. 4

[9] Connor Shorten and Taghi Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 07 2019. 2

[10] Yabin Zhang and Lei Zhang. Adaneg: Adaptive negative proxy guided ood detection with vision-language models, 2024. 2, 3, 4

[11] Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 271–288, Cham, 2025. Springer Nature Switzerland. 2, 3, 4