Reproducing PT4AL: Challenges and Insights into Using Self-Supervised Pretext Tasks for Active Learning

Minseok Yoon, Sohee Bae, Seonyoung Yoon
Depratment of Data Science
Seoul National University of Science and Technology, Republic of Korea
pill0106@ds.seoultech.ac.kr, shbae2819@g.seoultech.ac.kr, adbstjsdud@seoultech.ac.kr

Reproducibility Summary

Scope of Reproducibility

This paper aims to reproduce the key claims presented in "PT4AL: Using Self-Supervised Pretext Tasks for Active Learning."[1] The main claims are: 1) The loss of the pretext task is correlated with the loss of the main task. 2) By utilizing pretext task loss and batch samplers, it is possible to consider both distribution data and uncertainty data. 3) PT4AL can effectively solve the cold start problem.

Methodology

Our approach closely follows the method described in the original paper, except for minor changes in the datasets. Since the provided code only included the rotation task for the CIFAR10 dataset, we implemented the code for other datasets and experiments based on the paper's descriptions.

Results

Our findings showed some differences from the authors' claims. We observed a very low correlation between the pretext task loss and the main task loss across all datasets tested (CIFAR10[2], Imbalanced CIFAR10[2], and Caltech101[3]), contrary to the authors' assumption. Additionally, while the PT4AL method showed higher accuracy than random sampling, we were unable to confirm its effectiveness in addressing the cold start problem.

What was easy

Following the overall framework and structure of the PT4AL method was relatively straightforward based on the paper's descriptions.

What was difficult

The official code provided in the paper often lacks most of the experiments, making it difficult to reproduce the results. The code is optimized for specific datasets and tasks, making it challenging to apply to other tasks or datasets while maintaining the framework.

Communication with original authors

We did not have any direct communication with the original authors during this reproduction study.

Submitted to ML Reproducibility Challenge 2020. Do not distribute.

1 Introduction

Recent successes in deep learning have led to significant advancements in computer vision tasks, primarily due to CNNs[4] and large-scale labeled datasets. However, building large-scale labeled datasets is costly. Active Learning[5] aims to select the most informative subsets that achieve the best performance within a fixed labeling budget. Existing AL approaches are divided into distribution-based methods and uncertainty-based methods. Distribution-based methods[6, 7] sample data that well represents the overall feature distribution but may fail to select data near the decision boundary. Uncertainty-based approaches[8] address this issue by sampling the most uncertain points.

Against this backdrop, PT4AL aims to enhance the efficiency of Active Learning by combining self-supervised pretext task losses with batch samplers to simultaneously sample representative and challenging data. PT4AL orders data based on loss values obtained from pretext tasks and then divides it into batches, from which the most uncertain data is sampled. By considering the data that is difficult for the model to predict based on loss values, and simultaneously utilizing batch sampling to effectively sample data that well represents the overall feature distribution, PT4AL improves the effectiveness of Active Learning.

2 Scope of reproducibility

This paper aims to reproduce some of the key claims presented in "PT4AL: Using Self-Supervised Pretext Tasks for Active Learning." The main claims of PT4AL are as follows:

- The loss of the pretext task is correlated with the loss of the main task.
- By utilizing pretext task loss and batch samplers, it is possible to consider both distribution data and uncertainty data.
- PT4AL can effectively solve the cold start problem.

In this paper, we design experiments to validate these claims and reproduce the results of the original paper to evaluate the effectiveness of PT4AL. Our goal is to verify whether the hypotheses and methodologies proposed by PT4AL consistently perform well. The structure of this paper is as follows: Section1 explains the claims and methodologies presented by the authors. Section 2 describes the scope of the reproduction. Section 3 describes the reproduction methods. Section 4 compares the results of experiments reproducing the authors' methods. Finally, Section 5 presents the conclusions.

3 Methodology

Our goal is to precisely replicate the method implemented in PT4AL. Therefore, except for minor changes in the datasets, our approach follows the method described in the original paper. This section describes the PT4AL method. Since the code provided by the authors only included the rotation task for the CIFAR10 dataset, we re-implemented the code for the other datasets and experiments based on the descriptions given in the paper.

3.1 Using Pretext Tasks for Active Learning

A key assumption of PT4AL is that the pretext task loss is highly correlated with the main task loss (e.g., image classification, segmentation). The authors propose and validate the following assumption:

- H1: The pretext task loss is correlated with the main task loss.

The authors of the paper hypothesize that if the pretext task is correlated with or representative of the main task, then images that are difficult for the pretext task (i.e., images with high loss values) will also be difficult for the main task. Figure1is extracted from PT4AL. It shows scatter plots of the pretext task loss and main task loss on three benchmark datasets. In the scatter plots, the x-axis represents the normalized rank of the main task loss, and the y-axis represents the normalized rank of the pretext task loss.

As shown in Figure1, there is a strong positive correlation between the pretext task loss and the main task loss. The authors of PT4AL observed high ρ values across all three datasets: CIFAR10 (ρ = 0.79), Caltech101 (ρ = 0.78), and

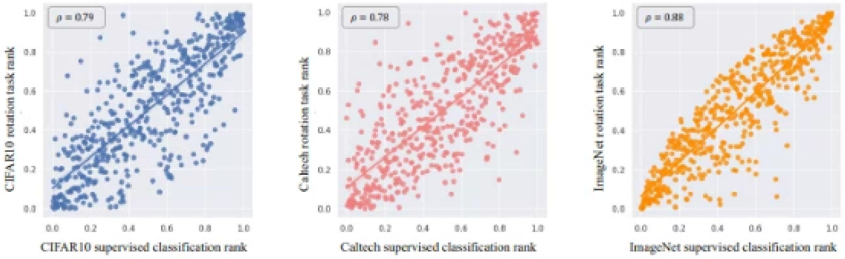


Figure 1: This figure is extracted from the paper. (From left to right) The loss rank correlation plots for the main task loss and the pretext task loss in CIFAR10, Caltech101 and ImageNet. The x and y axes represent the normalized rank of the two losses, respectively

ImageNet (ρ = 0.88). This validates the hypothesis of a correlation between the two losses and provides evidence that using pretext task loss in active learning is effective.

3.2 Pretext Task Learning for Batch Split

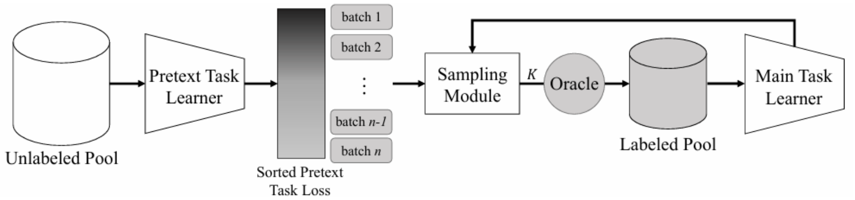


Figure 2: The overall framework of PT4AL. Unlabeled data are sorted by pretext task losses, split into batches, and sampled for training

Figure2 shows the overall structure of PT4AL extracted from the PT4AL paper. For batch splitting in PT4AL, the pretext task is first performed. Using rotation prediction task as an example, the model F_p is trained on the unlabeled dataset X_U using the pretext task loss, defined as the average cross-entropy loss for each of the four rotation directions.

$$\text{loss}(x_i; \theta_p) = \frac{1}{k} \sum_{y=1}^{k} \mathcal{L}_{CE}(F_p(g(x_i \mid y) \mid \theta_p), y) \tag{1}$$

The pretext task learning model F_p extracts the pretext task loss values from the unlabeled dataset X_U. The model weights θ_p are used for loss extraction. The extracted loss values are sorted in descending order, and the sorted data is divided into multiple batches. The number of batches is equal to the number of active learning iterations.

3.3 In-batch Sampling

After extracting high-uncertainty data using the pretext task, we employed active learning for image classification. The in-batch sampler selects K samples per iteration by computing the top-1 posterior probability within each batch using the prior main task learner F_m^{i-1}. K data points with the lowest confidence scores are chosen for annotation, ensuring the selection of both difficult and representative data. In the first iteration, K points are uniformly sampled from the initial batch, forming a labeled pool with the most informative data.

$$\phi(b_i, F_m^{i-1}) = \min_{\mathcal{K}} \left\{ \max \left(F_m^{i-1}(b_i \mid \theta_m) \right) \right\} \tag{2}$$

4 Replication Experiments

This section describes the replication experiments and results conducted to evaluate the PT4AL approach. First, we analyzed the correlation between pretext task loss and main task loss to validate the core assumption of PT4AL. Then, we conducted main task experiments to assess the effectiveness of the PT4AL method. In the PT4AL paper, various experiments using different datasets are mentioned, but the provided code is limited to the Rotation Prediction task on the CIFAR10 dataset. Therefore, we referred to this official code to implemented the code for the other datasets and experiments based on the descriptions given in the paper.

4.1 Experimental Setting

We replicated the experimental setup used in the PT4AL paper to verify the performance of PT4AL. The detailed experimental settings are as follows.

4.1.1 Dataset

- **CIFAR10**: Consists of 50,000 training images and 10,000 test images of size 32 × 32, categorized into 10 object classes.
- **Imbalanced CIFAR10**: An Imbalanced dataset created by adjusting the class proportions of the CIFAR10 dataset, with the same imbalance ratio as in the paper.
- **Caltech101**: Contains 9,144 images of approximately 300 × 200 pixels distributed across 101 classes, divided into 8,046 training images and 1,098 test images.

4.1.2 Pretext task

- **Rotation Prediction**: The input images were rotated by 0 degrees, 90 degrees, 180 degrees, and 270 degrees, and the network was trained to predict the rotated angle.
- **Colorization**: The network was trained to restore grayscale images to their original color images.

4.1.3 Hyperparameters

	Pretext task		Main
	Rotation	Colorization	Image classification
epoch	128	128	100
Batch size	256	256	
Optimizer	SGD	Adam	SGD
Learning rate	0.1	0.1	0.1

Upon running the official code for the Rotation Prediction task on the CIFAR10 dataset, we were unable to reproduce the performance claimed in the paper. Additionally, the official code was optimized only for the CIFAR10 dataset, limiting its reusability. To improve reproducibility, we re-implemented the Rotation Prediction task code for CIFAR10, making some modifications. During these modifications, we maintained the parameters and most of the structure closely aligned with the original framework of the Rotation Prediction task.

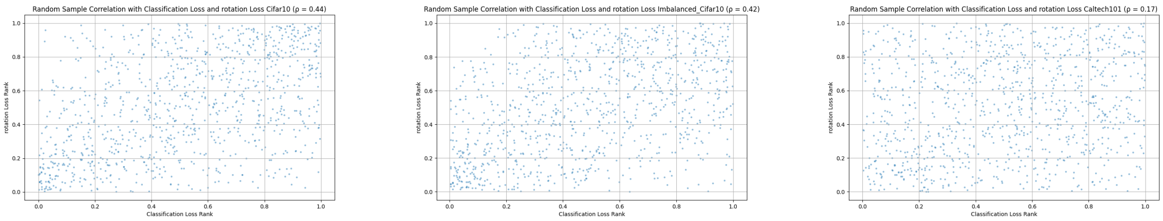
4.2 Results

The results of our replication experiments are detailed below. Each result is averaged over three repeated experiments. Our findings showed some differences from what the authors claimed.

4.2.1 Correlation Analysis: Pretext vs. Main Task Loss

We analyzed the correlation between pretext task loss and main task loss to validate the core assumption of PT4AL.The authors hypothesize that pretext task loss and main task loss are correlated, and that images with high loss values in the pretext task will also be difficult in the main task.

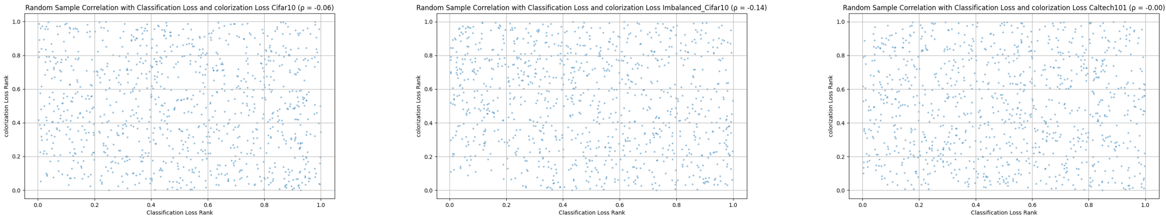
However, our study found that the core assumption of PT4AL is incorrect. Specifically, we observed that there is no high correlation between pretext task loss and main task loss. This implies that the method proposed by PT4AL may not perform as well as expected in selecting data. We first replicated the experiments performed in the paper to validate H1. We conducted correlation experiments using the CIFAR10, Imbalanced CIFAR10, and Caltech101 datasets, which were used in the original paper. For the pretext task, we performed rotation[1] and colorization[9] on each dataset. The results showed the following rotation correlation coefficients for each dataset.



(a) CIFAR10 Rotation **(b)** Imbalanced CIFAR10 Rotation **(c)** Caltech101 Rotation

Figure 3: The loss rank correlation plots for the main task loss (rotation task) and the pretext task loss in CIFAR10, Imbalanced CIFAR10, and Caltech101 datasets. The x and y axes represent the normalized rank of the two losses, respectively.

As shown in Figure3, the correlation between the pretext task loss and the main task loss was very low across all datasets tested: CIFAR10 ($\rho = 0.44$), Imbalanced Cifar10 ($\rho = 0.42$), and Caltech101 ($\rho = 0.17$). While the original paper reported correlation coefficients of $\rho = 0.79$ for CIFAR10 and $\rho = 0.78$ for Caltech101, our observations differed. Our results indicated that the correlation between the pretext task loss and the main task loss was very low across all datasets tested.



(a) CIFAR10 Colorization **(b)** Imbalanced CIFAR10 Colorization **(c)** Caltech101 Colorization

Figure 4: The loss rank correlation plots for the main task loss (colorization task) and the pretext task loss in CIFAR10, Imbalanced CIFAR10, and Caltech101 datasets.

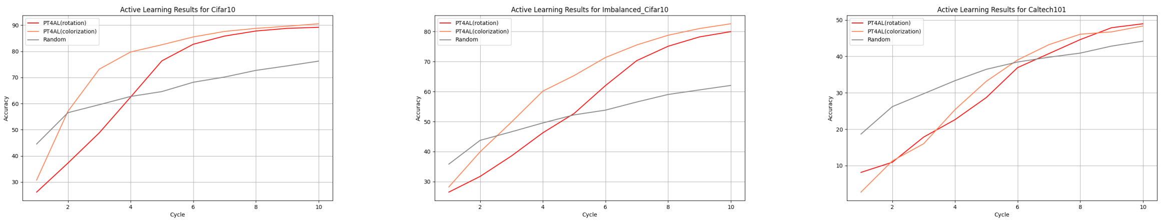
The authors only presented the correlation experiments for rotation, but we also examined the correlation for colorization, which was not included in the paper. Figure4 shows the results of the colorization experiments. The correlation results for colorization also showed no significant correlation: CIFAR10 ($\rho = -0.06$), Imbalanced CIFAR10 ($\rho = -0.14$), and Caltech101 ($\rho = -0.00$).

These results indicate that, contrary to the authors' assumptions, there is little correlation between the pretext task loss and the main task loss. The lack of significant correlation between pretext task loss and main task loss suggests that data with high pretext task loss does not necessarily indicate uncertain data. Since the core assumption of PT4AL is flawed, the proposed method may not be effective in improving performance on the main task.

4.2.2 Image Classification for Active learning

The experimental results for the main task, image classification, are presented below.

The experimental results for each dataset are presented in graphs. The results are obtained by performing the main task after using rotation and colorization as pretext tasks, with "random" referring to random sampling. Although the authors compared the performance with various methods including random sampling, our primary focus was to verify if the proposed method could surpass the performance of random sampling. Therefore, we conducted comparative experiments specifically with random sampling. For the CIFAR10, both PT4AL methods started with about 30% accuracy initially and reached approximately 90% accuracy by the 10th cycle. When comparing the two methods, colorization tended to perform better than rotation. In contrast, the random sampling method started with an initial



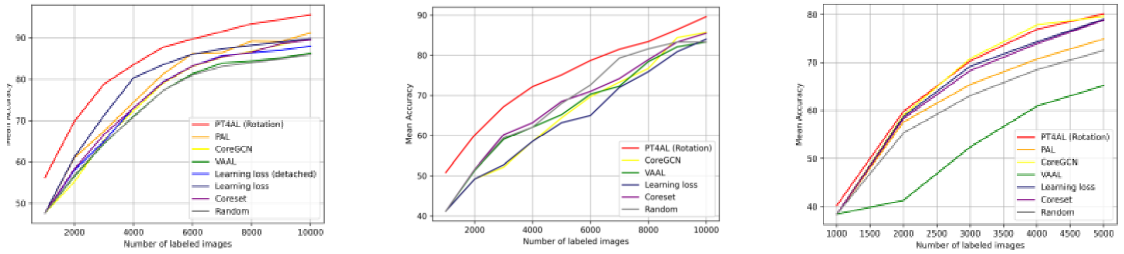
(a) CIFAR10 Colorization **(b)** Imbalanced CIFAR10 Colorization **(c)** Caltech101 Colorization

Figure 5: Comparison of image classification performance on CIFAR10, Imbalanced CIFAR10, and Caltech101.

accuracy of 30% and reached 70% accuracy by the 10th cycle, but it showed lower results compared to the two PT4AL methods.

For the Caltech101, both methods started with about 10% accuracy initially and reached approximately 90% accuracy by the 10th cycle, with no significant performance difference between the two methods. On the other hand, the random method started with about 20% accuracy initially but showed lower performance compared to the other two methods in the final cycle.

Lastly, for the Imbalanced CIFAR10, the performance started at about 30% initially and reached approximately 80% in the end. Again, colorization generally showed higher performance than rotation.



(a) CIFAR10 **(b)** Imbalanced CIFAR10 **(c)** Caltech101

Figure 6: This figure is extracted from the paper. Comparison of image classification performance on CIFAR10, Imbalanced CIFAR10, and Caltech101.

In all datasets, the PT4AL method showed higher accuracy than the random sampling method, but it was difficult to confirm the claimed advantage regarding the cold Start problem. The authors claim that, as shown in Figure 6, the initial performance in all datasets and methods is higher than that of random sampling, thereby resolving the cold Start problem. However, our experiments did not confirm the effectiveness of PT4AL in addressing the cold start problem. We believe that PT4AL did not perform as expected because we could not identify a correlation between the pretext task loss and the main task loss.

5 Conclusion

This report confirmed the main results reported in [1] through replication experiments. We used the publicly available code provided by the authors and implemented the code for other datasets and experiments besides the rotation task. The authors claimed a high correlation between the pretext task loss and the main task loss, suggesting that active learning utilizing this correlation would be effective. However, after conducting extensive experiments across various datasets (CIFAR10, imbalanced CIFAR10, Caltech101), we did not find a significant correlation between the pretext task loss and the main task loss. Additionally, contrary to the authors' claims, we were unable to resolve the cold start problem. These results suggest that the PT4AL method did not perform as well as the authors expected due to the lack of correlation between the pretext task loss and the main task loss. While we did not find a correlation between the pretext task and the main task in our experiments, we believe that applying this approach to manufacturing data and finding a more suitable correlation could make this pretext task a meaningful methodology.

References

[1] John Seon Keun Yi, Minseok Seo, Jongchan Park, and Dong-Geol Choi. Pt4al: Using self-supervised pretext tasks for active learning. In *European Conference on Computer Vision*, pages 596–612. Springer, 2022.

[2] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[3] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding*, 106(1):59–70, 2007.

[4] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

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

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

[7] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9583–9592, 2021.

[8] 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.

[9] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.