Synergistic Training: Harnessing Active Learning and Pseudo-Labeling for Enhanced Model Performance in Deep Learning

SEONGJIN OH[1], JONGPIL JEONG[1], CHAE-GYU LEE[1], JUYOUNG YOO[2], GYURI NAM[3]
[1]Department of Smart Factory Convergence
[2]Department of Electronics and Electrical Engineering
[3]Department of Chemical Engineering
Sungkyunkwan University
2066, Seobu-ro, Jangan-gu, Suwon-si, Gyeonggi-do, 16419
REPUBLIC OF KOREA

Abstract: - This research addresses the growing need for efficient data labeling methods by leveraging deep learning models. The proposed approach combines pre-training and active learning to automate the labeling process and reduce reliance on human annotators. In the pre-training phase, two deep learning models are trained using labeled data, adjusting the data ratio to ensure approximately 50% accuracy on the test set. In the active learning phase, the models generate pseudo labels for unlabeled data based on a confidence threshold, and the selected data is used to improve the models' performance through alternating epochs. The experimental results demonstrate the effectiveness of the approach, achieving significant improvements in accuracy compared to traditional methods. This research contributes to the trend of using deep learning for efficient data labeling and offers a promising solution for reducing the time and cost associated with manual annotation.

Key-Words: - Computer Vision, Deep Learning, Semi-supervised Learning, Active Learning, Pseudo Labeling, Human Annotator Replacement

Received: May 15, 2022. Revised: July 16, 2023. Accepted: August 27, 2023. Published: September 18, 2023.

1 Introduction

Recently, there has been a lot of research on how to leverage deep learning to automatically generate and efficiently label training data. This is mainly addressed in the fields of active learning and semi-supervised learning, which explore how to replace the role of human annotators with deep learning models. A recent trend in these studies is to develop efficient labeling techniques by combining methods such as active learning, self-training, co-training, and pseudo-labeling[1-6].

Due to the advancement of artificial intelligence technology, there is an increasing demand to utilize artificial intelligence models in various fields. However, sufficient training data is required to train artificial intelligence models based on deep learning. However, manually creating training data is a costly and time-consuming task. Therefore, there is a need for a method to efficiently label training data by replacing the role of human annotators with deep learning models. With this proposal as a background, the problem of studying how to replace the role of human annotators in active learning using deep learning models was raised[1].

Noting the recent advances in deep learning technology and the importance of training data, researchers have been working on various ideas to replace the work of human annotators. They explored how to use deep learning models to efficiently perform labeling and improve model performance without relying on human annotators. This idea generation process led to the proposed method. The proposed method consists of pre-learning and active learning phases and uses two deep learning models to efficiently label unlabeled data and improve performance. For this purpose, we use pre-trained models on a part of the dataset and utilize unlabeled data through Pseudo Labeling. Based on these ideas, the proposed technique is developed.

This study proposes a method for two deep learning models to complementarily label difficult data, rather than relying on a human annotator. The proposed method consists of pre-training and active learning phases, where two models are pre-trained with their respective labeled data, and then the models are trained by applying pseudo-labeling to unlabeled data. This improves the performance of the model and increases the efficiency of the training data. Experimental results show that the proposed method improves the performance of the model. The two pre-trained models showed 50% accuracy on the test data, and the performance of the models continued to improve as they learned unlabeled data through active learning. It is expected that the proposed method will contribute to increasing the efficiency of the training data generation and labeling process of deep learning models.

The paper is organized as follows. Section 2 describes related concepts and research. We describe the structure of our proposed technique in Section 3 and present experimental results in Section 4. Finally, we conclude in Section 5 with a discussion of the work and future work.

2 Related Work

2.1 Active Learning

Active learning is a form of machine learning, which refers to the process by which a model learns by selecting data on its own. Machine learning models are typically trained using labeled training data and then make predictions on new data. However, obtaining unlabeled data is costly and time-consuming, and Active Learning was developed to overcome these constraints[4,6].

The goal of Active Learning is to obtain the maximum performance gain by labeling as few samples as possible. To do this, it selects the most useful samples from the unlabeled dataset and offloads the labeling task to an oracle (e.g., a human annotator), with the goal of minimizing labeling costs while maintaining performance. Active Learning approaches can be categorized into the following scenarios: Membership Query Synthesis, Stream-based Selective Sampling, and Pool-based[1-3].

Membership query synthesis is an approach that aims to generate samples from the input space and query their labels. This method primarily leverages generative adversarial neural networks (GANs) for data generation, where the most informative samples can play an important role in improving model performance.

Stream-based approaches allow the model to request additional labels from data that arrive sequentially in the form of streams. When the input distribution is uniform, stream-based methods can behave similarly to membership query learning, but when the distribution is non-uniform and unknown, it makes sense to draw queries from the actual underlying distribution. This method is less studied in vision-related tasks compared to membership query synthesis and pool-based strategies but is effective for tasks where large amounts of data are generated in real time.

The pool-based active learning approach is used in situations where a small number of labeled data and many unlabeled data are available. This method involves selecting samples from a pool of data and querying them for labels and is often the most practical method because large amounts of unlabeled data can often be collected at once. Typical methods utilize entropy to measure uncertainty and select samples with higher uncertainty.

Active Learning has great potential for reducing the cost of labeling data and helping develop efficient models. The method can be used in real-world applications by reducing the time and money required for labeling. It can also be used effectively in areas where human annotators are required to minimize effort and domain expertise. Active Learning is one of the core principles of data-driven learning, and it is expected to show a lot of potential in real-world applications.

2.2 Sampling Startegy

In Active Learning, various sampling strategies have been developed to select the most informative data points to improve model performance. Uncertainty sampling is a strategy that selects data based on how uncertain the model is about its current predictions. It uses methods such as least confident, margin sampling, and entropy to calculate uncertainty and select the most uncertain data.

The least confident method selects the data with the lowest probability. Margin sampling selects the data with the smallest difference in probability between the most probable class and the next most probable class. The entropy method calculates the entropy and selects the data with the highest entropy. Uncertainty sampling, especially the entropy method, is the most widely used sampling strategy because it is simple and effective.

Other sampling strategies include Query-By-Committee, Expected Model Change, Variance Reduction, and Density-Weighted Methods.

Query-By-Committee is a method that uses multiple models or ensembles to select data. Each model makes predictions from a different perspective on the training data, estimates the uncertainty, and selects the most uncertain data. Expected Model Change measures the amount of information gain by predicting the change in the model when new data is added to training and selects data with the largest information gain.

Expected Error Reduction predicts whether adding new data to training will reduce the model's error and selects the data with the largest error reduction. Variance Reduction selects data to reduce the variance of the model's predictions. It reduces the uncertainty of the model by selecting data that has a large variance in the model's predictions. Density-Weighted Methods selects data by considering the distribution of data points. It selects data from low-density areas or border regions of the data distribution to help the model better explore the data space.

There are many other sampling strategies being researched. These sampling strategies utilize different principles and methods for efficient data selection in active learning to reduce labeling costs and improve model performance.

3 Proposed Method

Instead of asking a human annotator for labeling, the technique proposed in this paper induces complementary learning by allowing two deep learning models to annotate data that each has difficulty classifying. The overall structure is shown in Fig. 1.

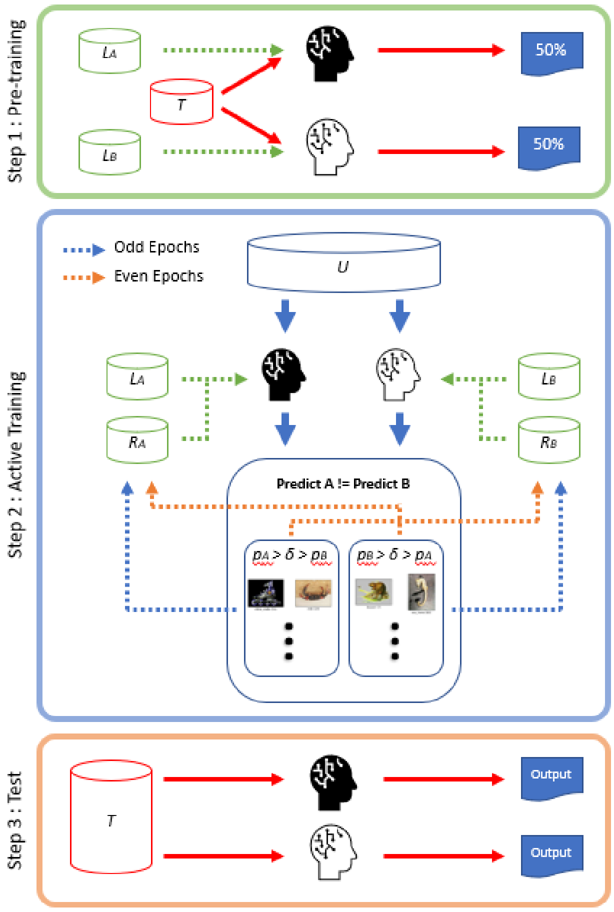

Fig. 1 A schematic of the proposed framework.

The operation procedure consists of 3 steps, and the data is divided into 4 groups for training as follows.

LA: Labeled data for pre-training the A model.
LB: Labeled data for pre-training the B model.
T: Labeled data for evaluating the accuracy.
U: Unlabeled data for performing active learning.

In the first step, the two deep learning models are pre-trained on their respective labeled data groups. In this case, the test data group **T** accounts for 20% of the total dataset, and the data groups **LA** and **LB** for pre-training are proportioned so that each model has 50% accuracy on **T** when trained to fit.

In the second step, Active Learning is performed, where each pre-trained model pseudo-labels the unlabeled data group **U** to create data to train each other. Pseudo-labeled data is created when the following conditions are satisfied when classifying the unlabeled data group **U**. The highest probability class result predicted by each model is different from each other, and compared to a predefined threshold, the prediction reliability of one model is higher than the threshold and the prediction reliability of the other model is lower than the threshold. Among the selected data, the data group with high prediction reliability of model A and the data group with high prediction reliability of model B are divided and used for model training. In Even Epochs, you use your own pseudo-labeled data as your training data, and in Odd Epochs, you use your opponent's labeled data as your training data. When learning, you learn the labeled data group **L** that you have pre-learned together. The unlabeled data group **U** is immutable, and **RA** and **RB** are newly created at each epoch.

Finally, in the third step, the accuracy is measured by classifying the test group **T** with each of the trained models. The user can select or recycle the better performing model among the models.

4 Experiment and Results

4.1 Datasets and Training Models

This section consists of experimental results of our proposed technique using the Caltech101 dataset and EfficientNet-B0.

The Caltech101 dataset is one of the widely used public datasets for image classification tasks. It consists of images belonging to 101 different categories.

The Caltech101 dataset mainly contains images of objects and animals. Each category contains at least 80 images, and often more. The dataset consists of a total of 9,144 images, split into training and test sets.

Each image has different sizes and resolutions, and may have variations in background, lighting, and rotation. These are included to mimic image classification tasks in the real world.

The Caltech101 dataset can be used for a variety of computer vision tasks, including performance evaluation of deep learning models, development of image classification algorithms, and transfer learning. The dataset is publicly available and heavily utilized by academic researchers and computer vision developers.

EfficientNet-B0 is the smallest model explored through Neural Architecture Search (NAS) and is a deep learning model with computational and parameter efficiency. EfficientNet-B0 is designed for image classification tasks and can be applied to a variety of computer vision tasks.

Compound Scaling: EfficientNet uses a concept called compound scaling to scale the network. This improves the performance and efficiency of the model by simultaneously adjusting the depth, width, and resolution of the network. EfficientNet-B0 keeps the model small while maintaining an efficient structure.

EfficientNet Architecture: EfficientNet uses an efficient network structure to maximize computational and parametric efficiency. It utilizes techniques such as Depthwise Separable Convolution, Inverted Residuals, Squeeze-and-Excitation, and more to achieve more expressiveness with fewer parameters.

Wide range of applications: EfficientNet-B0 can be utilized for a variety of computer vision tasks, including image classification, object detection, segmentation, transfer learning, and more. The model's small size and efficiency make it suitable for deployment on mobile devices or in lightweight environments.

4.2 Results

As a first step in our proposed technique, we adjust the proportion of training data so that the accuracy of each model on test group **T** is close to 50%. The pre-training data groups $\mathbf{L_A}$ and $\mathbf{L_B}$ each account for 20% of the total Caltech101 dataset and have different image data. With this data, models A and B were trained for 50 epochs to classify test group **T** with 99% accuracy on their respective data, and we found that the accuracy did not deviate significantly from 50% even after further pre-training.

Table 2 shows the accuracy of test group **T** for each additional training epoch. Fig. 2 is a graphical representation of Table 2.

Table 2 Accuracy of Each Model Except AL.

Epochs	1	2	3	4	5	6	7	8	9	10
EfficientNet A	48.71	51.10	49.36	51.65	50.07	50.55	52.40	51.04	52.26	51.38
EfficientNet B	49.95	50.67	49.63	49.35	51.27	51.18	50.06	50.73	50.94	51.94

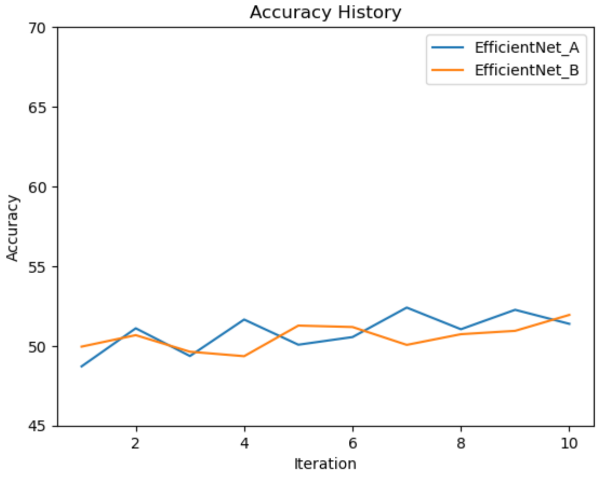

Fig. 2 Accuracy of Each Model Except AL.

The results of performing Active Learning are the result of the following two steps.

Condition 1: When you use the other party's labeled data for learning without exchanging pseudo-labeled data.
Condition 2: When you use your own labeled data for odd epochs and the other party's labeled data for even epochs.

Table 3 shows the Accuracy for Test Group T by Epoch when the other party's labeled data is used for learning without exchanging Pseudo Labeled Data in Condition 1. Fig. 3 is a graphical representation of Table 3.

Table 3 Accuracy of Each Model in Condition 1.

Epochs	2	4	6	8	10	12	14	16	18	20
EfficientNet A	56.56	61.63	62.37	61.58	62.72	61.89	62.34	62.43	62.60	62.16
EfficientNet B	59.38	60.14	60.32	61.75	62.31	62.34	61.76	62.03	61.93	61.95

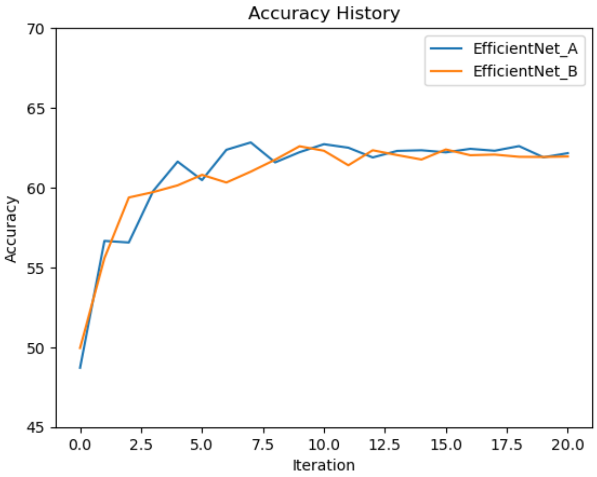

Fig. 3 Accuracy of Each Model in Condition 1.

We can see that after 7 epochs, both models converge to 62% accuracy.

Seeing that the two models no longer outperformed each other in Condition 1 because their predictions were identical, in Condition 2 they used their own labeled data for odd epochs and their opponent's labeled data for even epochs to increase the prediction bias of each model.

Table 4 shows the Accuracy for Test Group T per Epoch when trained with Condition 2. Fig. 4 is a graphical representation of Table 4.

Table 4 Accuracy of Each Model in Condition 2.

Epochs	2	4	6	8	10	12	14	16	18	20
EfficientNet A	61.42	62.34	63.14	63.37	64.85	65.41	64.49	65.69	65.44	65.82
EfficientNet B	60.04	62.23	62.89	63.82	63.84	65.25	65.85	65.68	66.20	66.19

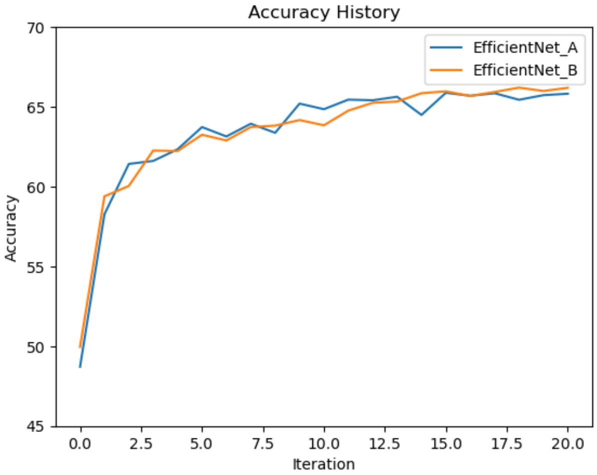

Fig. 4 Accuracy of Each Model in Condition 2.

As a result of exchanging pseudo-labeled data, the accuracy of Condition 1 was achieved at 4 epochs, and it continued to increase slightly thereafter. By 20 epochs, we achieved about 66% accuracy, which is about 16%p better than without active learning.

5 Conclusion

The technique proposed in this paper explores how to perform efficient Active Learning by replacing the role of a human annotator with a deep learning model. The proposed technique combines dictionary learning and pseudo-labeling to efficiently learn unlabeled data and improve the performance of the model. Experimental results show that the proposed technique improves performance through complementary learning between two deep learning models. By performing active learning, the performance was improved by 16 percentage points, and efficient utilization of training data was achieved.

The main contributions of this research are as follows: First, we propose a technique to efficiently generate training data by replacing the role of human annotators, which can save time and cost; Second, by combining pre-learning and pseudo-labeling, we effectively utilize unlabeled data to improve the performance of the model; Third, by using active learning, we maximize the efficiency of the labeling task by allowing the model to select and update training data on its own. Through these main contributions, the proposed technique succeeds in efficiently utilizing training data and improving performance at the same time.

Although the techniques proposed in this study have contributed to the efficient utilization of training data and improved model performance, there are some limitations. To overcome them, future research plans can include the following.

First, improving the quality of unlabeled data: Although the proposed technique utilizes unlabeled data through pseudo-labeling, the automatic labeling may cause errors in some data. Therefore, future research should investigate methods to further improve the quality of unlabeled data.

Second, scalability for small-scale datasets: In the current study, we utilized the Caltech101 dataset to conduct experiments, but in real-world applications, we need to deal with much larger and more diverse datasets. Future research should develop techniques that can scale from small-scale datasets to large-scale datasets to further expand their applicability in various applications.

Third, development of domain-specific Active Learning techniques: Although the current proposed techniques used a general Active Learning approach, effective Active Learning strategies may vary depending on domain characteristics. Future research should develop domain-specific Active Learning techniques to maximize the efficiency and performance improvement of labeling.

References:

[1] Khan Abbas. "PMAL: A Proxy Model Based Active Learning Approach for Image Classification." Master's thesis, Department of Engineering, School of Graduate Studies, Sejong University, Korea, 2022.

[2] Son Y ,and Lee J . "Active learning using transductive sparse Bayesian regression." Information Sciences-- (2016): 240-254.

[3] Bernard J ,Hutter M ,Zeppelzauer M ,Fellner D ,Sedlmair M . "Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study" IEEE transactions on visualization and computer graphics : 298-308.

[4] Song Liangchen,Xu Yonghao,Zhang Lefei,Du Bo,Zhang Qian,Wang Xinggang. "Learning From Synthetic Images via Active Pseudo-Labeling" IEEE transactions on image processing : 6452-6465.

[5] Kellenberger Benjamin. "Half a Percent of Labels is Enough: Efficient Animal Detection in UAV Imagery Using Deep CNNs and Active Learning" IEEE transactions on geoscience and remote sensing : 9524-9533.

[6] VUNUNU CALEB BRUCE NGANDU. "Unsupervised and Semi-supervised Learning Methods based on Deep Clustering and Explainable Active Learning." Domestic Doctoral Dissertation, Pukyong National University, 2021.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Sungjin Oh, led the project and carried out the simulation and the optimization.
Juyoung Yoo, researched the research data and designed the learning strategy.
Gyuri Nam, researched the research data and designed the sampling strategy.

Corresponding author: Prof. Jongpil Jeong and Prof. Chae-gyu Lee

Sources of funding for research presented in a scientific article or scientific article itself

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-2018-0-01417) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). Moreover, Following are results of a study on the “Leaders in INdustry-university Cooperation 3.0” Project, supported by the Ministry of Education and National Research Foundation of Korea.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0
https://creativecommons.org/licenses/by/4.0/deed.en_US