Is Meta-Learning Always Necessary?: A Practical ML Framework Solving Novel Tasks at Large-scale Car Sharing Platform

Hyunhee Chung*
SOCAR AI Research
Seoul, Republic of Korea
esther@socar.kr

Kyung Ho Park*
SOCAR AI Research
Seoul, Republic of Korea
kp@socar.kr

Abstract

While the deep neural networks achieved superior performance in various tasks under the supervised regime, the ML practitioners in the real world frequently encounter a novel task that cannot acquire the labeled dataset shortly. Even if they have become available in acquiring the target samples from the unlabeled dataset, conventional labeling procedures require the practitioners to invest in resource consumption. Pursuing an effective solution to these problems, our study proposes a practical ML framework that efficiently enables practitioners to solve novel tasks. Our ML framework consists of two solutions consisting of early and mature stages. First, the early stage solution lets the practitioners solve the novel task under the few-shot classification setting. Second, the mature stage solution enhances the labeling efficiency by retrieving samples that seem relevant to the target. Upon these solutions, acquiring a qualified representation power is the most important job. Under the public benchmark datasets and image recognition tasks in a large-scale car-sharing platform, we examined that the paradigm of supervised learning, surprisingly not meta-learning, produces the most beneficial representation power to solve novel tasks. We further scrutinized the supremacy of supervised representation derives from broader, nourished high-level representations in the neural networks. We highly expect our analyses can be a concrete benchmark to the ML practitioners who solve novel tasks in their domain.

1. Introduction

Deep neural networks have accomplished significant performances in various machine learning (ML) applications such as image recognition [10, 12] under the fully-supervised large-scale annotated dataset. Suppose that the ML practitioners successfully deployed a supervised model under the supervised regime. In this case, unfortunately, there comes a hardship regarding the following question: ”What if the ML practitioners should solve a *novel task* where sufficient samples are not accessible?” Note that we denote this problem as a *novel task* problem. In the real world, ML practitioners frequently encounter a novel task in which they cannot acquire sufficient annotated samples shortly. This problem frequently occurs when the ML practitioner shall wait for a particular time until sufficient samples are accumulated in the database. This insufficiency might be derived from the inherent difficulty of dataset acquisition (i.e., samples regarding the accident or disaster) or the nature of user-generated data (i.e., samples from the newly-launched business are rare). As the supervised paradigm requires a large annotated dataset, the ML practitioners cannot simply apply the conventional procedure due to the overfitting. One presumable approach is dividing the problem into two stages: early stage and mature stage. An early stage describes a circumstance where the ML practitioner cannot acquire sufficient samples in the database, but only have a few-labeled one. Note the *novel task* problem in the early stage has been actively studied as a few-shot classification [20, 15, 17]. The mature stage implies that the unlabeled database includes a particular number of novel samples in novel tasks, but the practitioners should invest particular resources into annotation procedures.

In the early stage, the ML practitioners can easily think of applying the meta-learning [8]. The meta-learning aims to adapt neural networks fastly to the classes not seen in the training set given only a few samples of each of these novel classes. Suppose the ML practitioners already have one labeled *base dataset* and a few samples at each label of the novel task. We can easily apply the meta-learning paradigm by regarding the *base dataset* as a meta-training dataset and few novel samples as a support set. The ML practitioners can simply train the neural networks with meta-training approaches on the base dataset and let the trained model solve the few-shot classification with the support samples. While this meta-learning approach might allow the practitioners to establish a promising baseline to solve the novel task, it bears several drawbacks. First, the meta-learning's representation power is not stable. As the performance of meta-learner deviates according to the chosen meta-training method [8, 18], the practitioners should experimentally validate the most effective one fit to their domain. Second, the meta-learning approach requires additional resource consumption of the ML practitioners. The ML practitioners should additionally train the meta-learner model from the given base dataset, and deploy it in the production environment. If the number of novel tasks increases, the resource consumption also increases excessively from training, deploying, and managing multiple meta-learners.

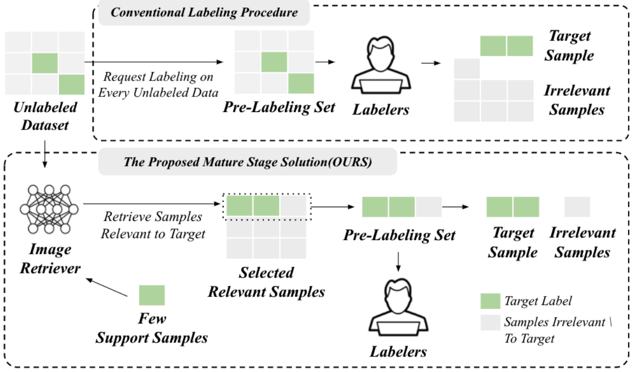
Conventional Labeling Procedure
Unlabeled Dataset
Request Labeling on Every Unlabeled Data
Pre-Labeling Set
Labelers
Target Sample
Irrelevant Samples
The Proposed Mature Stage Solution(OURS)
Image Retriever
Retrieve Samples Relevant to Target
Selected Relevant Samples
Pre-Labeling Set
Target Sample
Irrelevant Samples
Few Support Samples
Labelers
Target Label
Samples Irrelevant \ To Target

Figure 1. Comparison between conventional labeling procedure and the proposed labeling procedure

In the mature stage, acquiring novel labels from the unlabeled database is a straightforward method to solve the novel task. However, it requires excessive annotation efforts from the practitioners. Conventional labeling procedures consist of the following steps as shown in the upper diagram at Figure 1. As a foremost job, The ML practitioners prepare a few samples at each label to provide an annotation guideline for the labelers. The ML practitioners retrieve every unlabeled sample from the database to create a pre-labeling dataset and offer it to the labelers with the annotation guide. Given an unlabeled sample, the labeler annotates it as a target label if it fits the labeling guide and rejects it when it does not belong to any labels. Upon this conventional labeling procedure, we figured out several inefficiencies. First, the ML practitioners cannot know whether the pre-labeling dataset includes the target labels or not a priori. If the unlabeled pre-labeling dataset does not include target labels at all, unfortunately, the ML practitioners shall repeat the procedure with much resource consumption. Even the pre-labeling set includes target label samples, the practitioners would waste labeling costs as the pre-training set includes a particular amount of irrelevant samples.

Among recent studies on meta-learning, [18] examined that supervised representation power on the training set can significantly outperform existing meta-learners. Motivated by this recent work, we propose a practical ML framework to solve novel tasks with the supervised learner instead of meta-learners. Pursuing an efficient solution to this *novel task* problem, our study presents a case study in a large-scale car sharing platform illustrating how we solved the problem in the real world based on the proposed framework. We aim to introduce our approach to the novel task with novel labels problem and provide lessons learned from applying the proposed approach in the real world car sharing platform. We expect our study to be a concrete baseline for the candidate ML practitioners who solve similar problems. The key contributions of our study are as follows.

- Upon the public benchmark datasets in meta-learning studies, we experimentally discovered that [18]'s proposition is generally valid under both 1-shot and 5-shot settings, except for the case where the test set‘s distribution is shifted from the training set under the 5-shot setting.

- We unveiled the effective representation power of the supervised learner is located in the higher layers of the CNN. Especially, the comparative advantage of supervised representation is the capacity to describe the given image's various contextual information. We discovered the supervised learner could illustrate much broader, fruitful contextual representations regarding a given image; thus, it contributes to the supreme performance in various few-shot classification tasks.

- We set two real-world scenarios in a large-scale car sharing platform (where novel tasks are not excessively shifted from the training set), and validated that the supervised learner achieves a better few-shot classification performance than meta-learners.

- Given a set of the unlabeled dataset and a few target samples, we discovered that the supervised learner retrieved more relevant samples than the meta-learners. While the retrieval performance is not superficial, we at least examined that the supervised learner‘s representation is much more competitive in understanding novel labels in novel tasks.

2. Preliminaries

2.1. Problem Definition

First and foremost, we set a circumstance where the ML practitioners have already solved at least one image recognition task under the supervised regime, and we denote this task as a *base task*. This setting also presumes that the practitioners have one labeled dataset for the *base task*, and we denote this dataset as *base dataset*. Upon the aforementioned setting, we defined *few-shot novel task* problem as an image classification where its label space is distinct from the *base dataset*. Specifically, this *novel task* problem requires the practitioners to fulfill two solutions: early stage solution and mature stage solution. In the early stage of the *novel task* problem, the ML practitioners cannot acquire a sufficient number of labeled samples for *novel task*; thus, it requires the practitioners to solve a few-shot classification task. Conversely, in the mature stage, the practitioners can solve the novel task with a supervised classifier as the practitioners can acquire a particular amount of target samples from the unlabeled dataset. But, the obstacle lies in the labeling procedure because the practitioners have to invest much annotation efforts. Therefore, the mature stage solution requires the practitioners to resolve this labeling inefficiency. In a nutshell, the ML practitioners should solve tackle down two obstacles: 1) How can we solve the few-shot classification task in the early stage? and 2) How can we reduce the labeling inefficiency in the mature stage? Our study aims to develop a practical ML framework that resolves the aforementioned obstacles with the supervised learner.

2.2. Image Recognition Tasks

In this section, we illustrate image recognition tasks in a large-scale car sharing platform. `AnonymousInstitution` is the largest car sharing platform in the Republic of Korea, and it operates ten thousand fleets in every city (similar to ZipCar in the United States). The core business of the car sharing platform is letting the users borrow the car with a smartphone application. When users need to use a car, they reserve the car on the application and visit the designated parking station. Then, users can open the car on the application, use it, and return it to the parking station after their trips. `AnonymousInstitution` requires them to take pictures of the car (i.e., surfaces, seat, or cup holders) at particular events: before and after using the car, accident, car wash, and so on. These images taken by the users are uploaded to the database. To manage a large number of fleets without much human engagement, `AnonymousInstitution` has utilized various machine learning models in operational procedures. In pursuit of examining the effectiveness of the proposed ML framework, we presume the ML practitioners in `AnonymousInstitution` had solved one *base task*), and they were about to solve two novel image classification tasks (*novel tasks*). Given a *base dataset*, we aimed to solve *novel tasks* with a proposed practical ML framework. The image recognition tasks' detailed descriptions are elaborated in the following sections.

Base Task The *base task* in the car sharing platform is a car state recognition task. The car state recognition is a 10-class classification where each label illustrates a car‘s states, and every class is as follows: *Normal*, *Defect*, *Dirt*, *Bubble Wash*, *Cars inside of the Washing Machine*, *Dashboard*, *Cup Holder*, *Glovebox*, *Washer Fluid*, *Seat*. To solve this task in a supervised regime, we have acquired a labeled dataset (*base dataset*) denoted as **Car-Image**, and we visualized example images of the Car-Image and accessible URLs at supplementary material. Regarding the *novel task* problem, we utilized this Car-Image as a *base dataset*. It becomes a training set each learner described in the section 2.3.

Novel Task 1: Shape-Shifted Task The first target task is *shape-shifted* image recognition. The *shape-shifted* image recognition is a 3-class image classification between three labels: *Receipts*, *Documents*, and *Wheel*. This task had occurred when the company launched a new business with auto repair shops. When the company requests an auto repair shop to change the wheel, each shop resolves the request and sends images regarding repair operations. The auto repair shop sends *Document* images and *Receipts* images to prove a contract and transaction, respectively. Moreover, each shop sends *Wheel* images to prove that it completed a repair request of wheel change. These samples were accumulated in the database, and we aimed to establish an image recognition model that classifies a given sample‘s label. Note that we named this image recognition task as a *shape-shifted* task as these three labels are distinct from the *base dataset*, especially in shape. Each label of the *shape-shifted* task is illustrated in Figure **??**. (a) at supplementary material.

Novel Task 2: Texture-Shifted Task The *Texture-Shifted* image recognition is a 3-class image classification among three labels: *Black Cars*, *Gray Cars*, and *Cars with Snow*. While the *Shape-Shifted* task describes a circumstance where novel samples are distributionally shifted on their shapes, we synthetically established this *Texture-Shifted* samples that have different textures from the training set. Note that the *Normal*, *Defect*, *Dirt*, and *Bubble Wash* samples in the Car-Image only include white cars. We presume the samples in *Black Cars*, *Gray cars* experience a texture shift from the training set due to their colors. The *Cars with Snow* samples describe a car's surface with the snow on its surface; this, we presume the snow would cause a distribution shift in the perspective of texture. Each label of the *texture-shifted* task is described in Figure **??**. (b) supplementary material.

2.3. Experimental Settings

Research Questions Upon the *base dataset* and two *novel tasks*, we established several research questions to examine whether the supervised-learner can solve *novel task* problem in both the short and long run. Our research starts from validating the effectiveness of supervised-learner in the public benchmark setting. We then excavate underlying takeaways of the effectiveness and examine whether the supervised learner achieves significant performance in the real world setting. The detailed research questions are as follows.

- **RQ 1.** Does the meta-learner always outperform the supervised learner?
- **RQ 2.** What contributes to the qualified representation power for novel tasks?
- **RQ 3.** Does the supervised learner solve real world few-shot classification better?
- **RQ 4.** Does the supervised learner retrieve relevant samples from the unlabeled dataset well?

Baselines Our study employed several baselines to examine the effectiveness of supervised learners. For the baselines in the study, we aim to acquire many representation powers derived from various model training paradigms. As baselines of meta-learners, we utilized one metric-based approach [15], and one optimization-based approach (Model-Agonistic Meta-Learning; MAML) [4]. We selected these two methods among various meta-learners as they accomplished significant performance in a few-shot classification task under the public benchmark dataset. Furthermore, following the experiment setup in [18], we also employed the self-supervised learning paradigm as our study's baseline method following the reported effectiveness in [18]. Among various self-supervised learning methods, we employed a contrastive learning approach denoted as Bootstrap Your Own Latent (BYOL) [5] as it achieved promising performance in public benchmark datasets of the representation learning domain. Lastly, we additionally employed a neural network with a frozen weight pre-trained on the ImageNet [2]. As the ImageNet-weight has been widely utilized in various computer vision tasks due to its qualified representation power, we presume neural networks trained on the ImageNet would become a concrete baseline of our study.

Implementation Details Throughout the study, we employed five training paradigms in the journey of establishing a practical ML framework: *ImageNet*, *Supervised Learner*, *Self-Supervised Learner*, *Meta-Learner-Optimization*, and *Meta-Learner-Metric*. As implementation details, we commonly utilized the neural networks architecture of ResNet-50. For *ImageNet* option, we acquired the model's parameters trained on the ImageNet classification (which is easily accessible in machine learning frameworks such as PyTorch or Tensorflow). For the *Supervised Learner*, we utilized the model's parameters trained on the training set under the supervised regime, and we established *Self-Supervised Learner* by letting the model solve the contrastive learning tasks described in [5]. For *Meta-Learner-Optimization* and *Meta-Learner-Metric*, we followed the implementation procedures in [15] and [4], respectively. Unlike the meta-learners, *ImageNet*, *Supervised Learner*, and *Self-Supervised Learner* does not have a module to perform a target classification (novel task). Upon the experiment setups in [18], we froze the model's parameters at each learner, acquired the representations on the support samples and trained a simple logistic regression model. For the reproducibility of our implementations, we described our code in `anonymousURL`.

Table 1. 3-way 1-shot classification results on the public benchmark dataset settings

Method	**3-way 1-shot Classification**		
	CIFAR-FS & CIFAR-FS (Minimal Shift)	**CIFAR-FS & miniImageNet** (Particular Shift)	**CIFAR-FS & Double-MNIST** (Massive Shift)
ImageNet	39.75 ± 0.85	37.15 ± 0.69	33.99 ± 0.59
Supervised Learner	**71.22 ± 0.87**	**55.78 ± 0.86**	**41.48 ± 0.76**
Self-Supervised Learner	59.79 ± 0.43	39.29 ± 1.30	37.11 ± 0.61
Meta-Learner-Optimization	53.33 ± 0.31	40.45 ± 0.24	37.30 ± 0.38
Meta-Learner-Metric	49.18 ± 0.18	39.01 ± 0.34	36.98 ± 0.89

Table 2. 3-way 5-shot classification results on the public benchmark dataset settings

Method	**3-way 5-shot Classification**		
	CIFAR-FS & CIFAR-FS (Minimal Shift)	**CIFAR-FS & miniImageNet** (Particular Shift)	**CIFAR-FS & Double-MNIST** (Massive Shift)
ImageNet	38.64 ± 0.71	37.72 ± 0.65	28.93 ± 0.55
Supervised Learner	**84.46 ± 1.12**	65.65 ± 1.02	39.19 ± 0.69
Self-Supervised Learner	79.19 ± 0.79	**68.90 ± 0.83**	43.67 ± 0.18
Meta-Learner: Optimization	78.75 ± 0.95	60.27 ± 0.29	54.58 ± 0.40
Meta-Learner: Metric	68.13 ± 0.94	59.80 ± 0.58	**56.27 ± 1.59**

3. Does the Supervised Learner Outperform Meta-Learner under the Dataset Shift?

Setup As an answer to **RQ 1.**, we aim to examine whether the supervised learner bears more effective representation power than other baselines. Although the prior study [18] proved the supervised representation is much effective in several public benchmark datasets, we analyzed this takeaway shall be validated in more various dataset settings to let the ML practitioners apply it in the real world. An improvement avenue of [18] exists in the domain similarity between the training set and the test. Referring to the experiment settings in [18], the training set and the test set have been derived from the same source dataset. However, in the real world, there frequently exists a circumstance where the test set has a particular amount of domain shift [11] from the training set. As a starting research question, our study aims to check whether the supervised learner's representation power is still effective when the test set (*novel task* in our problem setting) experiences domain shift from the training set.

Given the training samples of CIFAR-FS [1] as a training set, we employed three test datasets: Test set at CIFAR-FS [1], Mini-ImageNet [3], and Double-MNIST [16] (denoting each test set as CIFAR-FS-Test, mini ImageNet-Test, and Double-MNIST-Test). The CIFAR-FS-Test implies a circumstance where the training and test sets share the same domain without any domain shifts (also known as distribution shifts or dataset shift) [11] (just as a conventional few-shot classification setting). The mini ImageNet-Test describes a particular dataset shift from the training set as samples in both CIFAR-FS-Training and mini ImageNet-Test are natural images (semantically similar), but other factors are different (i.e., size or resolution). The Double-MNIST-Test implies a massive dataset shift from the training set. To justify the aforementioned analysis with a quantified metric, we checked the out-of-distribution (OOD) scores on each test set in the perspective of the training set. If the trained model on the training set (CIFAR-FS-Training) understands a given sample properly, the OOD score would be low, and vice versa. Among various OOD detection approaches [21, 7], we employed the Maximum Softmax Probability (MSP) method [7, 19], which is a concrete baseline in OOD detection studies. The higher MSP implies high confidence in the given sample, resulting in a low OOD score (less likely to be out-of-distribution). Following the OOD score distribution shown in Figure 2, we figured out that both mini ImageNet-Test and Double MNIST-Test are distributionally shifted from the training set, and Double MNIST-Test is more shifted than the mini ImageNet-Test. In a nutshell, we justified our problem settings with the aforementioned three test sets are valid.

Upon the experiment setups described above, we acquired trained neural networks from different training paradigms elaborated on the section 2.3. We let each model solve the 3-way n-shot classification task with the acquired representation power, where the n is 1 and 5. Following the evaluation suite in the benchmark paper [18], in every test set, we measured a few-shot accuracy on 20000 episodes under the paradigm of episodic training [17]. The 3-way 1-shot and 3-way 5-shot classification results are shown in Table 1 and Table 2, respectively.

Analysis We figured out that adequate representation power differs along with each problem setting following the experiment results. As a foremost analogy, we discovered the [18]‘s proposition is valid on their experiment settings as the supervised learner accomplished the best performance on the minimally-shifted dataset at both 1-shot and 5-shot settings. We further examined that the supervised learner is effective in general circumstances except for the 5-shot classification under the massively-shifted test dataset. We analyzed that these results imply the effectiveness of supervised representation power over the others. Following the prior studies on the few-shot classification, there primarily exist two drivers of good representation in solving novel tasks: 1) the trained representation power and 2) the ability to fastly adapt its parameters. Under the 1-shot setting, there are not many samples that each learner can utilize to understand the novel labels. Thus, we expect the trained representation power would significantly influence the few-shot classification performance rather than the ability to adapt its parameters fastly. Upon the supreme performance of the supervised learner, we analyzed the supervised representation as to the one generally applicable in various problem settings even the test dataset is shifted from the training set.

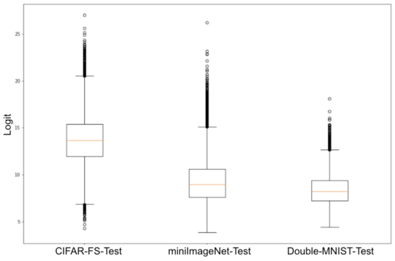

Figure 2. The OOD score distribution of each test data from the distribution of training set (CIFAR-FS-Training)

On the other hand, under the 5-shot setting, we expect both the trained representation power and the ability to adapt the model‘s parameters to the novel tasks are correlated to the experiment results. In the 5-shot setting, more support samples exist that the meta-learners can utilize to update their representation power. We expect the supervised learner accomplished the best performance under the minimal shift because the trained representation power (that shares the same domain with the test) was sufficient enough to analyze novel samples. However, under the 5-shot massive shift, we presume that an ability to fine-tune its parameters on novel samples fastly has become more influential than the trained representation power (as there are many support samples rather than a 1-shot setting). As meta-learners are intrinsically designed to adapt their parameters to the novel tasks effectively, we expect they accomplished better few-shot classification performance under the 5-shot massive shift. Note that we acknowledge these takeaways are presumable analogies regarding the experiment results (which leaves an improvement avenue to the future works). Consequently, we discovered that the supervised representation power generally understands novel samples without fine-tuning; thus, the candidate ML practitioners can utilize it in their tasks but should be cautious if their test dataset is massively shifted from the training set.

4. What Drives A Good Representation?

Setup As an answer to the **RQ 2.**, we further scrutinized what factor drives a supreme representation power of the supervised learner. Following the recent studies of interpreting the representation power [9], we aim to examine which layer of the deep neural networks contributes to the supervised learner‘s significant few-shot classification performance. We empirically assumed the ML practitioners would not frequently experience an excessively-shifted test dataset; thus, we narrowed the scope of analysis to the minimal shift. Under the conventional CNN architectures such as ResNet [6], the lower layers (which locates near the input image) are known to learn primitive features of the image, such as dots or lines (denoted as low-level features). On the other hand, representations at higher layers (which are located near the final softmax layer) illustrate contextual information of the image (denoted as high-level features) [13, 14, 13]. Upon the aforementioned notion, our study aims to examine what representation (low-level features or high-level features) significantly contributes to the effectiveness of supervised learner. Accordingly, our study estimated the representation similarity among multiple learners leveraging centered kernel alignment (CKA) [9], a useful tool to measure the similarity between two representations extracted from different neural networks. To discover the layer that contributes to the qualified representation power of the supervised learner, we measured the representation similarity between the supervised learner and other learners at both the lowest and highest layers. We fed samples in CIFAR-FS-Test to each learner to compare their representations, and every learners are trained under the CIFAR-FS-Training except for the *ImageNet*. The result is described in Table 3.

Table 3. Representation similarity between the Supervised Learner and other Learners

Supervised &	**Representation Similarity**	
	Lower Layer	**Hgher Layer**
Supervised Learner	1.0	1.0
ImageNet	0.8288	0.2926
Self-Supervised	0.7418	0.0928
Meta-Learner	0.7918	0.0067

Analysis Following Table 3, we resulted in that contextual representations at the higher layers of the neural network particularly contribute to the effective few-shot classification performance. Upon the lower layers, the low-level representation of the supervised learner is similar to those of the other learners. We interpret the low-level layers at each learner acquired similar knowledge; thus, we analyzed the low-level features do not seem to be significantly relevant to the supremacy of the supervised learner. On the other hand, we figured out that the supervised learner's high-level representations are distinct from those of the other learners. We presume contextual knowledge acquired under the supervised paradigm is advantageous to understand novel samples. Furthermore, we additionally visualized layer-wise representation similarities at a single model to compare an overall shape of the representation power. For example, given a ResNet-50 model trained under the supervised regime (supervised learner), we measured the similarities among every 50 residual blocks of the model, yielding a representation similarity matrix. As this matrix shows the correlations among every residual block of the model, we expect it to represent a unique characteristic of each learner's knowledge. The representation similarity matrices at each learner are visualized in Figure 3.

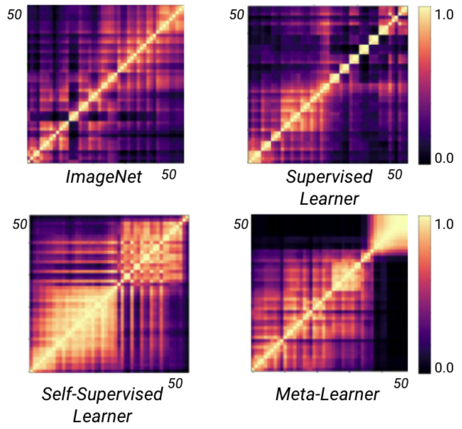

Figure 3. Layer-wise similarity matrix at each learner

Upon the Figure 3., we discovered the supervised learner‘s high-level representations are less correlated themselves while those of the other learners are highly correlated. On the supervised learner, representations at high-level layers have comparatively lower similarities to each other. We analyze these less-correlated high-level representations imply that each high-level layer illustrates different elements (characteristics) of the image, which means a larger capacity of describing an image. This larger capacity also implies a broader contextual knowledge of the image, and we presume this fruitful contextual representation of a given image contributes to better performance. On the other hand, high-level representations of the *ImageNet* and *Self-Supervised Learner* are particularly correlated themselves, and the high-level representations at the meta-learner are nearly the same. We expect the more similar high-level representations imply that the model learns a limited, narrow contextual knowledge on a given image. Then, it would not contribute to the effective understanding of the knowledge. In a nutshell, we figured out that an underlying reason behind the supervised learner‘s significant representation power resides in the high-level layers of the neural networks. Especially, we discovered that higher layers trained under the supervised regime could illustrate more nourished contextual patterns regarding given samples; thus, this large capacity of contextual illustration contributes to the significant few-shot classification performance of the supervised learner.

5. Early-Stage: Few-Shot Classification

Setup After we examined the supervised representation‘s effectiveness and where does this effectiveness comes from, we further applied the supervised learner in the real world. Pursuing an answer to the **RQ 3.**, we aim to vali-

Table 4. 3-way 1-shot results on the domain dataset settings, especially for shape-shifted and texture-shifted tasks.

Method	Shape-Shifted Task		Texture-Shifted Task	
	3-way 1-shot	**3-way 5-shot**	**3-way 1-shot**	**3-way 5-shot**
ImageNet	40.02 ± 1.11	57.63 ± 0.97	40.08 ± 0.71	59.72 ± 0.88
Supervised Learner	70.04 ± 1.24	**81.87 ± 1.22**	**65.99 ± 1.02**	**79.61 ± 0.96**
Self-Supervised Learner	**71.53 ± 1.25**	79.65 ± 1.23	64.19 ± 1.08	75.83 ± 0.91
Meta-Learner-Optimization	62.49 ± 1.49	71.39 ± 1.27	60.48 ± 1.20	75.41 ± 0.80
Meta-Learner-Metric	55.56 ± 1.67	66.10 ± 0.86	58.78 ± 0.98	67.82 ± 1.25

date whether the supervised learner can be an effective early stage solution to solve the *novel task* problem. As we elaborated on section 2.1, during the early stage of *novel task* problem, the ML practitioners cannot acquire sufficient labeled samples on the novel task; thus, they shall solve the novel task under the few-shot classification setting. Therefore, our study designed an experiment to check whether the supervised learner trained with the Car-Image (*base dataset*) can solve the two novel, real world tasks (*shape-shifted* task, *texture-shifted* task) under the few-shot classification problem setting. Suppose the supervised learner accomplishes effective few-shot classification performance on the two *novel tasks*. In that case, we can recommend the ML practitioners adapt our early-stage solution to solve tasks in their domain.

While the two *novel tasks* look semantically shifted from the Car-Image, we checked the OOD score distribution of each *novel dataset* at Figure 4. We expect this description regarding the distributional shift of the *novel dataset* would contribute to the clear understanding of our experiment to the candidate ML practitioners. Following the OOD score distribution in Figure 4., both tasks were shifted from the Car-Image-Training, and the *Shape-shifted* task experiences a larger dataset shift. Moreover, both *novel tasks* are not massively shifted compared to the domain shift between the *CIFAR-FS* & *Double-MNIST*; thus, we would like to highlight that the real world *novel tasks* in our study are not excessively shifted from the dataset. As detailed experiment setups, we measured the few-shot classification accuracy in 20000 episodes at each task under the 1-shot and 5-shot settings. Note that both novel tasks are *3-way n-shot* problem as each task includes three labels. The experiment results are shown in Table 4.

Analysis Following the experiment results shown in Table 4., we discovered the supervised learner outperformed other baselines in every problem settings except for 3-way 1-shot classification at *Shpe-Shifted Task*. Still, we evaluate our supervised learner accomplished competitive performance to the best accuracy of the *Self-Supervised Learner* as their performances have a minor gap. Regarding the experiment results at 3-way 1-shot classification on *Shape-Shifted Task*, we could not figure out why the *Self-Supervised Learner* achieves better performance rather than the supervised learner. Similar to our analogy in the prior section, we leave this question as an improvement avenue of our study. Consequentially, we examined the supervised learner’s effective few-shot classification performance on various problem settings; thus, we recommend the candidate ML practitioners refer to our early-stage solution on their tasks.

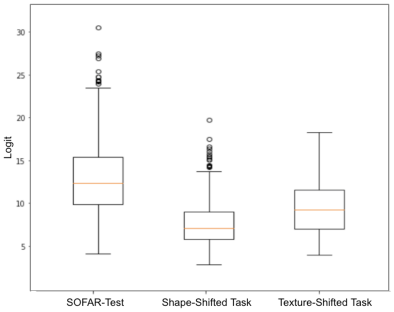

Figure 4. The OOD score distribution of each *novel datasets* from the distribution of training set (Car-Image)

6. Mature Stage: Zero-Shot Image Retrieval

Setup Towards the **RQ 4.**, our study aims to examine the validity of our mature stage solution to the *novel task* problem. We validate whether the supervised learner can retrieve the samples relevant to the target label from the unlabeled dataset, given few-labeled target samples. As the learner is not trained with samples of the target label, we denote this solution as a zero-shot image retrieval. We would like to highlight this zero-shot image retrieval is challenging as the unlabeled *Pool* is in an open-set setting (which has samples irrelevant to the task). Upon this setting, we presume the supervised learner can become an effective image retriever if its representation power is qualified enough. In the experiment, we set two target labels *Document* and *Cars with Snow*, which exist in *Shape-Shifted Task* and *Texture-Shifted Task*, respectively. We acquired 3000 random unlabeled samples from the live database as a target unlabeled dataset (denote as *Pool*). Note that samples in the *Pool* share the same label space with Car-Image, and it does not have any duplicated samples with Car-Image and datasets of *Shape-Shifted Task* and *Texture-Shifted Task.* As the *Pool* does not include samples of the target labels, we synthetically added target samples to it. We additionally retrieved a small-sized unlabeled dataset from the live database (without any duplicated samples with *Pool*), acquired 50 samples at each target label, and added to the *Pool*. Therefore, we could assure that the *Pool* includes 50 samples of the target labels. Given the *Pool* and five samples at each target task, we let the learner estimate an average similarity between the unlabeled sample and the given five samples. We employed two similarity metrics: Euclidean distance and cosine similarity. We iterate this similarity estimation on every sample in the *Pool* and select Top-50 samples with high similarity scores. As an evaluation metric, we measured the number of correctly-retrieved target samples in the retrieved 50 samples (Top-50 Accuracy). Note that the larger number implies a good performance of the learner as it retrieved many relevant samples from the *Pool*. We compared the performance of five learners on the aforementioned setting, and the experiment results are shown in Table 5.

Table 5. Zero-shot Retrieval experimental results

Method	Similarity Metric			
	Euclidean Distance		**Cosine Similarity**	
	Document	***Cars with Snow***	***Document***	***Cars with Snow***
ImageNet	2	8	2	6
Supervised Learner	**8**	**18**	**3**	**12**
Self-Supervised Learner	6	12	4	**12**
Meta-Learner-Optimization	0	2	2	9
Meta-Learner-Metric	0	2	4	10

Analysis Following Table 5., we figured out the supervised learner retrieved the largest number of relevant samples from the *Pool*. Regardless of the similarity estimation method and target labels, the supervised learner outperformed the image retrieval performance rather than other learners; thus, we validated that the supervised learner also can be applied in a zero-shot image retrieval as a mature stage solution to the *novel task* problem. Moreover, we discovered that the supervised learner achieved better retrieval performance on *Cars with Snow* rather than *Document*. We expect this result happens as the *Cars with Snow* samples are less distributionally-shifted from the training set compared to the *Documents*. Please check Figure 4 to check the domain shift of two target labels from the training set, Car-Image. Like the prior analogies in various few-shot classification tasks, we discovered again that the supervised representation becomes effective rather than other learners when the target label shifts less from the training set. Although we justified the effectiveness of supervised learner in zero-shot image retrieval, the absolute retrieval performances are not remarkable. We acknowledge that this becomes an improvement avenue of our study. In a nutshell, we recommend the candidate ML practitioners can apply the proposed mature stage solution on their task, but its effectiveness shall be improved to be actively applied in the real world.

7. Related Works

The meta-learning aims to empower a fast-adaptability toward the model to let it efficiently solve novel tasks. Especially, few-shot learning, which leverages few support samples on the novel task, is one prominent paradigm of meta-learning. Upon the few-shot learning studies, there exist two categories: optimization-based approach and metric-based approach. First, the optimization-based meta-learning trains fast-adapt parameters to converge with only a few support sets fastly. MAML [4], and its derived versions are representative methods of the optimization-based approach. A key takeaway of the MAML is utilizing gradient-steps to converge the model's parameters to adapt fastly to other novel tasks. Second, the metric-based meta-learning method embeds a given image into a fixed-shape representation vector. Metric-based approaches measure similarity between support samples and the validation sample. ProtoNet [15] and Relation Network [17] are highly popular ones under the metric-based meta-learning. In this study, we employed two meta-learning methods (MAML and ProtoNet) as baseline models representative approaches at the optimization-based and metric-based paradigm, respectively.

8. Discussions and Conclusion

In this study, we aim to conduct an evidential study on the practical ML framework that efficiently solves the novel task in the real world. We design our framework with two-stage solutions in the early and mature stages. The early stage solution aims to escalate the classification performance in novel tasks, given a few labeled samples. The mature stage solution contributes to reducing labeling inefficiencies through zero-shot image retrieval. To implement this framework, we conducted a series of experiments to examine whether the supervised representation can be utilized in both solutions upon the shared motivation with [18]. First and foremost, we discovered that the supervised learner is generally applicable in few-shot classification tasks under public benchmarks and real-world image recognitions. We further scrutinized this supremacy of the supervised learner derives from the nourished high-level representations within the trained neural networks. Lastly, we validated the supervised learner can be utilized as a zero-shot image retriever to escalate the labeling efficiency in the mature stage solution. Still, there exist several improvement avenues of our study. Future works shall validate a factor that drives different few-shot classification performances at different learners. Moreover, we shall improve the zero-shot image retrieval performance and excavate a key element of qualified representation power in this zero-shot problem setting. Lastly, the proposed ML framework and related takeaways had better be examined under various classification dataset settings, or various computer vision(i.e., object detection, semantic segmentation). As a closing remark, we highly expect our study can be a concrete benchmark to the candidate ML practitioners who solve the novel tasks in their domain.

9. Acknowledgement

This work is supported by the Korea Agency for Infrastructure Technology Advancement grant funded by the Ministry of Land, Infrastructure, and Transport (Grant RS-2022-0014579).

References

[1] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

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

[3] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*, 2019.

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[5] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

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

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[8] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[9] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[10] Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870, 2007.

[11] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[12] Siddhartha Sankar Nath, Girish Mishra, Jajnyaseni Kar, Sayan Chakraborty, and Nilanjan Dey. A survey of image classification methods and techniques. In *2014 International conference on control, instrumentation, communication and computational technologies (ICCICCT)*, pages 554–557. IEEE, 2014.

[13] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

[14] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

[15] Ori Sasson, Avishay Vaaknin, Hillel Fleischer, Elon Portugaly, Yonatan Bilu, Nathan Linial, and Michal Linial. Protonet: hierarchical classification of the protein space. *Nucleic acids research*, 31(1):348–352, 2003.

[16] Shao-Hua Sun. Multi-digit mnist for few-shot learning, 2019.

[17] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[18] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pages 266–282. Springer, 2020.

[19] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *arXiv preprint arXiv:2110.06207*, 2021.

[20] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[21] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.