Few-Shot One-Class Classification via Meta-Learning

Ahmed Frikha [1, 2, 4], **Denis Krompaß** [1, 2], **Hans-Georg Köpken** [3], **Volker Tresp** [2, 4]
[1]Siemens AI Lab
[2]Siemens Technology
[3]Siemens Digital Industries
[4]Ludwig Maximilian University of Munich
ahmed.frikha@siemens.com

Abstract

Although few-shot learning and one-class classification (OCC), i.e., learning a binary classifier with data from only one class, have been separately well studied, their intersection remains rather unexplored. Our work addresses the few-shot OCC problem and presents a method to modify the episodic data sampling strategy of the model-agnostic meta-learning (MAML) algorithm to learn a model initialization particularly suited for learning few-shot OCC tasks. This is done by explicitly optimizing for an initialization which only requires few gradient steps with one-class minibatches to yield a performance increase on class-balanced test data. We provide a theoretical analysis that explains why our approach works in the few-shot OCC scenario, while other meta-learning algorithms fail, including the unmodified MAML. Our experiments on eight datasets from the image and time-series domains show that our method leads to better results than classical OCC and few-shot classification approaches, and demonstrate the ability to learn unseen tasks from only few normal class samples. Moreover, we successfully train anomaly detectors for a real-world application on sensor readings recorded during industrial manufacturing of workpieces with a CNC milling machine, by using few normal examples. Finally, we empirically demonstrate that the proposed data sampling technique increases the performance of more recent meta-learning algorithms in few-shot OCC and yields state-of-the-art results in this problem setting.

Introduction

The anomaly detection (AD) task (Chandola, Banerjee, and Kumar 2009; Aggarwal 2015) consists in differentiating between normal and abnormal data samples. AD applications are common in various domains that involve different data types, including medical diagnosis (Prastawa et al. 2004), cybersecurity (Garcia-Teodoro et al. 2009) and quality control in industrial manufacturing (Scime and Beuth 2018). Due to the rarity of anomalies, the data underlying AD problems exhibits high class-imbalance. Therefore, AD problems are usually formulated as one-class classification (OCC) problems (Moya, Koch, and Hostetler 1993), where either only a few or no anomalous data samples are available for training the model (Khan and Madden 2014). While most of the developed approaches (Khan and Madden 2014) require a substantial amount of normal data to yield good generalization, in many real-world applications, e.g., in industrial manufacturing, only small datasets are available. Data scarcity can have many reasons: data collection itself might be expensive, e.g., in healthcare, or happens only gradually, such as in a cold-start situation, or the domain expertise required for annotation is scarce and expensive.

To enable learning from few examples, viable approaches (Lake et al. 2011; Ravi and Larochelle 2017; Finn, Abbeel, and Levine 2017) relying on meta-learning (Schmidhuber 1987) have been developed. However, they rely on having examples from each of the task's classes, which prevents their application to OCC tasks. While recent meta-learning approaches focused on the few-shot learning problem, i.e., learning to learn with few examples, we extend their use to the OCC problem, i.e., learning to learn with examples from only one class. To the best of our knowledge, the few-shot OCC (FS-OCC) problem has only been addressed in (Kozerawski and Turk 2018; Kruspe 2019) in the image domain.

Our contribution is fourfold: Firstly, we show that classical OCC approaches fail in the few-shot data regime. Secondly, we provide a theoretical analysis showing that classical gradient-based meta-learning algorithms do not yield parameter initializations suitable for OCC and that second-order derivatives are needed to optimize for such initializations. Thirdly, we propose a simple episode generation strategy to adapt any meta-learning algorithm that uses a bi-level optimization scheme to FS-OCC. Hereby, we first focus on modifying the model-agnostic meta-learning (MAML) algorithm (Finn, Abbeel, and Levine 2017) to learn initializations useful for the FS-OCC scenario. The resulting One-Class MAML (OC-MAML) maximizes the inner product of loss gradients computed on one-class and class-balanced minibatches, hence maximizing the cosine similarity between these gradients. Finally, we demonstrate that the proposed data sampling technique generalizes beyond MAML to other metalearning algorithms, e.g., MetaOptNet (Lee et al. 2019) and Meta-SGD (Li et al. 2017), by successfully adapting them to the understudied FS-OCC.

We empirically validate our approach on eight datasets from the image and time-series domains, and demonstrate its robustness and maturity for real-world applications by successfully testing it on a real-world dataset of sensor read-

ings recorded during manufacturing of metal workpieces with a CNC milling machine. Furthermore, we outperform the concurrent work One-Way ProtoNets (Kruspe 2019) and achieve state-of-the-art performance in FS-OCC.

Approach

The primary contribution of our work is to propose a way to adapt meta-learning algorithms designed for class-balanced FS learning to the underexplored FS-OCC problem. In this section, as a first demonstration that meta-learning is a viable approach to this challenging learning scenario, we focus on investigating it on the MAML algorithm. MAML was shown to be a universal learning algorithm approximator (**?**), i.e., it could approximate a learning algorithm tailored for FS-OCC. Later, we validate our methods on further meta-learning algorithms (Table 4).

Problem Statement

Our goal is to learn a one-class classification task using only a *few* examples. In the following, we first discuss the unique challenges of the few-shot one-class classification (FS-OCC) problem. Subsequently, we discuss the formulation of the FS-OCC problem as a meta-learning problem.

To perform one-class classification, i.e., differentiate between in-class and out-of-class examples using only in-class data, approximating a *generalized* decision boundary for the normal class is necessary. Learning such a class decision boundary in the few-shot regime can be especially challenging for the following reasons. On the one hand, if the model overfits to the few available datapoints, the class decision boundary would be too restrictive, which would prevent generalization to unseen examples. As a result, some normal samples would be predicted as anomalies. On the other hand, if the model overfits to the majority class, i.e., predicting almost everything as normal, the class decision boundary would overgeneralize, and out-of-class (anomalous) examples would not be detected.

In the FS classification context, N-way K-shot learning tasks are used to test the learning procedure yielded by the meta-learning algorithm. An N-way K-shot classification task includes K examples from *each* of the N classes that are used for learning this task, after which the trained classifier is tested on a disjoint set of data (Vinyals et al. 2016). When the target task is an OCC task, only examples from one class are available for training, which can be viewed as a 1-way K-shot classification task. To align with the anomaly detection problem, the available examples must belong to the normal (majority) class, which usually has a lower variance than the anomalous (minority) class. This problem formulation is a prototype for a practical use case where an application-specific anomaly detector is needed and only few normal examples are available.

Model-Agnostic Meta-Learning

MAML is a meta-learning algorithm that we focus on adapting to the FS-OCC problem before validating our approach on further meta-learning algorithms (Table 4). MAML learns a model initialization that enables quick adaptation to unseen tasks using only few data samples. For that, it trains a model explicitly for few-shot learning on tasks T_i coming from the same task distribution $p(T)$ as the unseen target task T_{test}. In order to assess the model's adaptation ability to *unseen* tasks, the available tasks are divided into mutually disjoint task sets: one for meta-training S^{tr}, one for meta-validation S^{val} and one for meta-testing S^{test}. Each task T_i is divided into two disjoint sets of data, each of which is used for a particular MAML operation: D^{tr} is used for adaptation and D^{val} is used for validation, i.e., evaluating the adaptation. The adaptation of a model f_θ to a task T_i consists in taking few gradient descent steps using *few* datapoints sampled from D^{tr} yielding θ_i'.

A good measure for the suitability of the initialization parameters θ for few-shot adaptation to a considered task T_i is the loss $L_{T_i}^{val}(f_{\theta_i'})$, which is computed on the validation set D_i^{val} using the task-specific adapted model $f_{\theta_i'}$. To optimize for few-shot learning, the model parameters θ are updated by minimizing the aforementioned loss across all meta-training tasks. This *meta-update*, can be expressed as:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_i \sim p(T)} L_{T_i}^{val}(f_{\theta_i'}). \tag{1}$$

Here β is the learning rate used for the meta-update. To avoid overfitting to the meta-training tasks, model selection is done via validation using tasks from S^{val}. At meta-test time, the FS adaptation to unseen tasks from S^{test} is evaluated. We note that, in the case of few-shot classification, K datapoints from *each* class are sampled from D^{tr} for the adaptation, during training and testing.

One-Class Model-Agnostic Meta-Learning

Algorithm. MAML learns a model initialization suitable for *class-balanced* (CB) FS classification. To adapt it to FS-OCC, we aim to find a model initialization from which taking few gradients steps with a few one-class (OC) examples yields the same effect as doing so with a CB minibatch. We achieve this by adequately modifying the objective of the inner loop updates of MAML. Concretely, this is done by modifying the data sampling technique during meta-training, so that the class-imbalance rate (CIR) of the inner loop minibatches matches the one of the test task.

MAML optimizes explicitly for FS adaptation by creating and using auxiliary tasks that have the same characteristic as the target tasks, in this case tasks that include only few datapoints for training. It does so by reducing the size of the batch used for the adaptation (via the hyperparameter K (**?**)). Analogously, OC-MAML trains explicitly for quick adaptation to OCC tasks by creating OCC auxiliary tasks for meta-training. OCC problems are binary classification scenarios where only few or no minority class samples are available. In order to address both of theses cases, we introduce a hyperparameter (c) which sets the CIR of the batch sampled for the inner updates. Hereby, c gives the percentage of the samples belonging to the minority (anomalous) class w.r.t. the total number of samples, e.g., setting $c = 0\%$ means only majority class samples are contained in the data batch. We focus on this extreme case, where no anomalous

Algorithm 1 Meta-training of OC-MAML

Require: S^{tr}: Set of meta-training tasks
Require: α, β: Learning rates
Require: K, Q: Batch size for the inner and outer updates
Require: c: CIR for the inner-updates
1: Randomly initialize θ
2: **while** not done **do**
3: Sample batch of tasks T_i from S^{tr}; $T_i = \{D^{tr}, D^{val}\}$
4: **for each** sampled T_i **do**
5: Sample K examples B from D^{tr} such that CIR$= c$
6: Initialize $\theta_i^{'} = \theta$
7: **for** number of adaptation steps **do**
8: Compute adapted parameters with gradient descent using B: $\theta_i^{'} = \theta_i^{'} - \alpha \nabla_{\theta_i^{'}} L_{T_i}^{tr}(f_{\theta_i^{'}})$
9: **end for**
10: Sample Q examples $B^{'}$ from D^{val} w/ CIR$= 50\%$
11: Compute outer loop loss $L_{T_i}^{val}(f_{\theta_i^{'}})$ using $B^{'}$
12: **end for**
13: Update θ: $\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{T_i} L_{T_i}^{val}(f_{\theta_i^{'}})$
14: **end while**
15: **return** meta-learned parameters θ

samples are available for learning. In order to evaluate the performance of the adapted model on both classes, we use a *class-balanced* validation batch $B^{'}$ for the meta-update. This way, we maximize the performance of the model in recognizing both classes after having *seen* examples from only one class during adaptation. The OC-MAML meta-training is described in Algorithm 1, and the cross-entropy loss was used for L. At test time, the adaptation to an unseen task is done by applying steps 5-9 in Algorithm 1, starting from the meta-learned initialization.

We note that the proposed episode sampling strategy, i.e., training on a one-class batch then using the loss computed on a class-balanced validation batch to update the meta-learning strategy (e.g., model initialization), is applicable to any meta-learning algorithm that incorporates a bi-level optimization scheme (examples in Table 4).

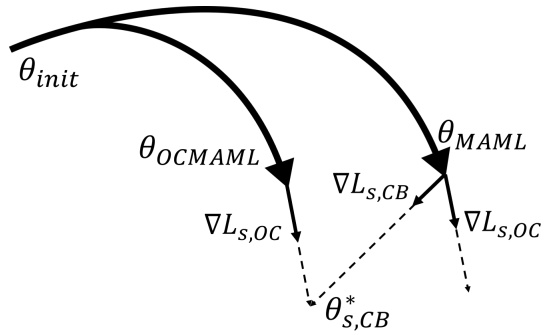

Figure 1: Adaptation to task T_s from the model initializations yielded by OC-MAML and MAML

Using OCC tasks for adaptation during meta-training favors model initializations that enable a quick adaptation to OCC tasks over those that require CB tasks. The schematic visualization in Figure 1 shows the difference between the model initializations meta-learned by MAML and OC-MAML. Hereby, we consider the adaptation to an unseen binary classification task T_s. $\theta^*_{s,CB}$ denotes a local optimum of T_s. The parameter initializations yielded by OC-MAML and MAML are denoted by θ_{OCMAML} and θ_{MAML} respectively. When starting from the OC-MAML parameter initialization, taking a gradient step using an OC support set $D_{s,OC}$ (gradient direction denoted by $\nabla L_{s,OC}$), yields a performance increase on T_s (by moving closer to the local optimum). In contrast, when starting from the parameter initialization reached by MAML, a class-balanced support set $D_{s,CB}$ (gradient direction denoted by $\nabla L_{s,CB}$) is required for a performance increase on T_s.

Theoretical Analysis: Why Does OC-MAML Work ? In this section we give a theoretical explanation of why OC-MAML works and why it is a more suitable approach than MAML for the FS-OCC setting. To address the latter problem, we aim to find a model parameter initialization, from which adaptation using few data examples from only *one* class yields a good performance on both classes, i.e., good generalization to the class-balanced task. We additionally demonstrate that adapting first-order meta-learning algorithms, e.g., First-Order MAML (FOMAML) (Finn, Abbeel, and Levine 2017) and Reptile (Nichol and Schulman 2018), to the OCC scenario as done in OC-MAML, does not yield initializations with the desired characteristics.

By using a Taylor series expansion the gradient used in the MAML update can be approximated to Equation 2 (Nichol and Schulman 2018), where the case with only 2 gradient-based updates is considered, i.e., one adaptation update on a minibatch (1), the support set including K examples from D^{tr}, and one meta-update on a minibatch (2), the query set including Q examples from D^{val}. We use the notation from (Nichol and Schulman 2018), where $\overline{g}_i$ and $\overline{H}_i$ denote the gradient and Hessian computed on the i^{th} minibatch at the initial parameter point ϕ_1, and α the learning rate. Here it is assumed that the same learning rate is used for the adaptation and meta-updates.

$$\begin{aligned} g_{MAML} &= \overline{g}_2 - \alpha \overline{H}_2 \overline{g}_1 - \alpha \overline{H}_1 \overline{g}_2 + O(\alpha^2) \\ &= \overline{g}_2 - \alpha \frac{\partial (\overline{g}_1 . \overline{g}_2)}{\partial \phi_1} + O(\alpha^2) \end{aligned} \tag{2}$$

Equation 2 shows that MAML maximizes the inner product of the gradients computed on different minibatches (Nichol and Schulman 2018). Under the assumption of local linearity of the loss function (which is the case around small optimization steps), and when gradients from different minibatches have a positive inner product, taking a gradient step using one minibatch yields a performance increase on the other (Nichol and Schulman 2018). Maximizing the inner product leads to a decrease in the angle between the gradient vectors and thus to an increase in their cosine similarity. Hence, MAML optimizes for an initialization where gradients computed on *small* minibatches have similar directions, which enables few-shot learning.

Equation 2 is independent of the data strategy adopted and

hence holds also for OC-MAML. However, in OC-MAML the minibatches 1 and 2 have different class-imbalance rates (CIRs), since the first minibatch includes examples from only one class and the second minibatch is class-balanced. So, it optimizes for increasing the inner product between a gradient computed on a one-class minibatch and a gradient computed on class-balanced data. Thus, OC-MAML optimizes for an initialization where gradients computed on one-class data have similar directions, i.e., a high inner product and therefore a high cosine similarity, to gradients computed on class-balanced data (Figure 1). Consequently, taking one (or few) gradient step(s) with one-class minibatch(es) from such a parameter initialization results in a performance increase on class-balanced data. This enables one-class classification. In contrast, MAML uses only class-balanced data during meta-training, which leads to a parameter initialization that requires class-balanced minibatches to yield the same effect. When adapting to OCC tasks, however, only examples from one class are available. We conclude, therefore, that the proposed data sampling technique modifies MAML to learn parameter initializations that are more suitable for adapting to OCC tasks.

A natural question is whether applying the same data sampling method to other gradient-based meta-learning algorithms would yield the same desired effect. We investigate this for First-Order MAML (FOMAML), a first-order approximation of MAML that ignores the second derivative terms and Reptile (Nichol and Schulman 2018), which is also a first-order meta-learning algorithm that learns an initialization that enables fast adaptation to test tasks using few examples from *each* class. We refer to the versions of these algorithms adapted to the FS-OCC setting as OC-FOMAML and OC-Reptile. We note that for OC-Reptile, the first $N-1$ batches contain examples from only one class and the last (N^{th}) batch is class-balanced. The approximated FOMAML and Reptile gradients are given by Equations 3 and 4 (Nichol and Schulman 2018), respectively.

$$g_{FOMAML} = \overline{g}_2 - \alpha \overline{H}_2 \overline{g}_1 + O(\alpha^2) \tag{3}$$

$$g_{Reptile} = \overline{g}_1 + \overline{g}_2 - \alpha \overline{H}_2 \overline{g}_1 + O(\alpha^2) \tag{4}$$

We note that these equations hold also for OC-FOMAML and OC-Reptile. By taking the expectation over minibatch sampling $\mathbb{E}_{\tau,1,2}$ for a task τ and two *class-balanced* minibatches 1 and 2, it is established that $\mathbb{E}_{\tau,1,2}[\overline{H}_1 \overline{g}_2] = \mathbb{E}_{\tau,1,2}[\overline{H}_2 \overline{g}_1]$ (Nichol and Schulman 2018). Averaging the two sides of the latter equation results in

$$\begin{aligned} \mathbb{E}_{\tau,1,2}[\overline{H}_2 \overline{g}_1] &= \frac{1}{2}\mathbb{E}_{\tau,1,2}[\overline{H}_1 \overline{g}_2 + \overline{H}_2 \overline{g}_1] \\ &= \frac{1}{2}\mathbb{E}_{\tau,1,2}[\frac{\partial(\overline{g}_1 . \overline{g}_2)}{\partial \phi_1}]. \end{aligned} \tag{5}$$

Equation 5 shows that, FOMAML and Reptile, like MAML, in expectation optimize for increasing the inner product of the gradients computed on different minibatches with the *same* CIR. However, when the minibatches 1 and 2 have different CIRs, which is the case for OC-FOMAML and OC-Reptile, $\mathbb{E}_{\tau,1,2}[\overline{H}_1 \overline{g}_2] \neq \mathbb{E}_{\tau,1,2}[\overline{H}_2 \overline{g}_1]$ and therefore $\mathbb{E}_{\tau,1,2}[\overline{H}_2 \overline{g}_1] \neq \frac{1}{2}\mathbb{E}_{\tau,1,2}[\frac{\partial(\overline{g}_1 . \overline{g}_2)}{\partial \phi_1}]$. Hence, despite using the same data sampling method as OC-MAML, OC-FOMAML and OC-Reptile do *not* explicitly optimize for increasing the inner product, and therefore the cosine similarity, between gradients computed on one-class and class-balanced minibatches. The second derivative term $\overline{H}_1 \overline{g}_2$ is, thus, necessary to optimize for an initialization from which performance increase on a class-balanced task is yielded by taking few gradient steps using one class data.

Related Works

Our proposed method addresses the FS-OCC problem, i.e., solving binary classification problems using only *few* datapoints from only *one* class. To the best of our knowledge, this problem was only addressed in (Kozerawski and Turk 2018) and (Kruspe 2019), and exclusively in the image data domain. In (Kozerawski and Turk 2018) a feed-forward neural network is trained on ILSVRC 2012 to learn a transformation from feature vectors, extracted by a CNN pre-trained on ILSVRC 2014 (Russakovsky et al. 2015), to SVM decision boundaries. At test time, an SVM boundary is inferred by using one image of one class from the test task which is then used to classify the test examples. This approach is specific to the image domain since it relies on the availability of very large, well annotated datasets and uses data augmentation techniques specific to the image domain, e.g., mirroring. Meta-learning algorithms offer a more general approach to FS-OCC since they are data-domain-agnostic, and do not require a pre-trained feature extraction model, which may not be available for some data domains, e.g., sensor readings.

The concurrent work One-Way ProtoNets (Kruspe 2019) adapts ProtoNets (Snell, Swersky, and Zemel 2017) to address FS-OCC by using 0 as a prototype for the *null* class, i.e., non-normal examples, since the embedding space is 0-centered due to using batch normalization (BN) (Ioffe and Szegedy 2015) as the last layer. Given the embedding of a query example, its distance to the normal-class prototype is compared to its norm. This method constraints the model architecture by requiring the usage of BN layers. We propose a model-architecture agnostic data sampling technique to adapt meta-learning algorithms to the FS-OCC problem. The resulting meta-learning algorithms substantially outperform One-Way ProtoNets (Kruspe 2019) (Table 4).

Class-Balanced Few-Shot Classification

Meta-learning approaches for FS classification approaches may be broadly categorized in 2 categories. Optimization-based approaches aim to learn an optimization algorithm (Ravi and Larochelle 2017) and/or a parameter initialization (Finn, Abbeel, and Levine 2017; Nichol and Schulman 2018), learning rates (Li et al. 2017), an embedding network (Lee et al. 2019) that are tailored for FS learning. Metric-based techniques learn a metric space where samples belonging to the same class are close together, which facilitates few-shot classification (**?**Vinyals et al. 2016; Snell, Swersky, and Zemel 2017; Sung et al. 2018; Oreshkin, López, and Lacoste 2018; Lee et al. 2019). Hybrid methods (**?**Lee and

Choi 2018) combine the advantages of both categories. Prior meta-learning approaches to FS classification addressed the *N*-way *K*-shot classification problem described in the problem statement section, i.e they require examples from *each* class of the test tasks. We propose a method to adapt meta-learning algorithm to the *1*-way *K*-shot scenario, where only few examples from *one* class are available.

One-Class Classification

Classical OCC approaches rely on SVMs (Schölkopf et al. 2001; Tax and Duin 2004) to distinguish between normal and abnormal samples. Hybrid approaches combining SVM-based techniques with feature extractors were developed to compress the input data in lower dimensional representations (Xu et al. 2015; Erfani et al. 2016; Andrews et al. 2016). Fully deep methods that jointly perform the feature extraction step and the OCC step have also been developed (Ruff et al. 2018). Another category of approaches to OCC uses the reconstruction error of antoencoders (Hinton and Salakhutdinov 2006) trained with only normal examples as an anomaly score (Hawkins et al. 2002; An and Cho 2015; Chen et al. 2017). Yet, determining a decision threshold for such an anomaly score requires labeled data from both classes. Other techniques rely on GANs (Goodfellow et al. 2014) to perform OCC (Schlegl et al. 2017; Ravanbakhsh et al. 2017; Sabokrou et al. 2018). The aforementioned hybrid and fully deep approaches require a considerable amount of data from the OCC task to train the typically highly parametrized feature extractors specific to the normal class, and hence fail in the scarce data regime (Table 1).

Experimental Evaluation

The conducted experiments [1] use some modules of the pyMeta library (Spigler 2019) and aim to address the following key questions: (a) How do meta-learning-based approaches using the proposed episode sampling technique perform compared to classical OCC approaches in the few-shot (FS) data regime? (b) Do the findings of our theoretical analysis about the differences between the MAML and OC-MAML initializations hold in practice? (c) Does the proposed episode sampling strategy to adapt MAML to the FS-OCC setting yield the expected performance increase and does this hold for further meta-learning algorithms?

Baselines and Datasets

We compare OC-MAML, with the classical OCC approaches One-Class SVM (OC-SVM) (Schölkopf et al. 2001) and Isolation Forest (IF) (Liu, Ting, and Zhou 2008) (Question (a)), which we fit to raw features and embeddings of the support set of the test task. Here, we explore two types of embedding networks which are trained on the meta-training tasks as follows: one is trained in a Multi-Task-Learning (MTL) (Caruana 1997) setting using one-class-vs-all tasks and the other trained using the "Finetune" baseline (FB) (Triantafillou et al. 2019). i.e., using multi-class classification on all classes available.

[1] Code available under https://github.com/AhmedFrikha/Few-Shot-One-Class-Classification-via-Meta-Learning

Moreover, we compare first-order (FOMAML and Reptile) and second-order (MAML) class-balanced meta-learning algorithms to their adapted versions to the OCC scenario, i.e., OC-FOMAML and OC-Reptile and OC-MAML (Question (b)). Finally, we compare MetaOptNet (Lee et al. 2019) and meta-SGD (Li et al. 2017) to their one-class counterparts that use our sampling strategy (Question (c)). We conducted a hyperparameter search for each baseline separately and used the best performing setting for our experiments. We evaluate our approach on 8 datasets from the image and time-series data domains, including two synthetic time-series (STS) datasets that we propose as a benchmark for FS-OCC on time-series, and a real-world sensor readings dataset of CNC Milling Machine Data (CNC-MMD). To adapt the image datasets to the OCC scenario, we create binary classification tasks, where the normal class is one class of the initial dataset and the anomalous class contains examples from *multiple* other classes.

Results and Discussion

In this section, we first discuss the performance of classical OCC approaches and the meta-learning algorithms in the FS-OCC problem setting, as well as the impact of the proposed data sampling strategy. Subsequently, we demonstrate the maturity of our approach on a real-world dataset. Thereafter, we further confirm our theoretical analysis with empirical results of cosine similarity between gradients. Finally, we show the generalizability of our sampling technique to further meta-learning algorithms beyond MAML, and compare the resulting algorithms to One-Way ProtoNets.

Table 1 shows the results averaged over 5 seeds of the classical OCC approaches (Top) and the meta-learning approaches, namely MAML, FOMAML, Reptile and their one-class versions (Bottom), on 3 image datasets and on the STS-Sawtooth dataset. For the meta-learning approaches, models were trained with and without BN layers and the results of the best architecture were reported for each dataset. The results of all the methods on the other 8 MT-MNIST task-combinations and on the STS-Sine dataset, are consistent with the results in Table 1.

While classical OCC methods yield chance performance in almost all settings, OC-MAML achieves very high results, consistently outperforming them across all datasets and on both support set sizes. Likewise, we observe that OC-MAML consistently outperforms the class-balanced and one-class versions of the meta-learning algorithms in all the settings, showing the benefits of our modification to MAML.

Moreover, OC-FOMAML and OC-Reptile yield poor results, especially without BN, confirming our theoretical findings that adapting first-order meta-learning algorithms to the OCC setting does not yield the desired effect. We found that using BN yields a substantial performance increase on the 3 image datasets and explain that by the gradient orthogonalizing effect of BN (Suteu and Guo 2019). In fact, gradient orthogonalization reduces interference between gradients computed on one-class and class-balanced batches. OC-MAML achieves high performance even without BN, as it reduces interference between these gradients by the means of its optimization objective (see theoretical analysis).

Adaptation set size	$K = 2$				$K = 10$			
Model \ Dataset	MIN	Omn	MNIST	Saw	MIN	Omn	MNIST	Saw
FB	50.0	50.6	56.5	50.0	50.0	51.2	50.3	50.0
MTL	50.0	50.0	49.7	50.0	50.2	50.0	45.3	50.0
OC-SVM	50.2	50.6	51.2	50.1	51.2	50.4	53.6	50.5
IF	50.0	50.0	50.0	50.0	50.7	50.0	50.9	49.9
FB + OCSVM	50.0	50.0	55.5	50.4	51.4	58.0	86.6	58.3
FB + IF	50.0	50.0	50.0	50.0	50.0	50.0	76.1	51.5
MTL + OCSVM	50.0	50.0	50.0	50.0	50.0	50.1	53.8	86.9
MTL + IF	50.0	50.0	50.0	50.0	50.0	55.7	84.2	64.0
Reptile	51.6	56.3	71.1	69.1	57.1	76.3	89.8	81.6
FOMAML	53.3	78.8	80.7	75.1	59.5	93.7	91.1	80.2
MAML	62.3	91.4	85.5	81.1	65.5	96.3	92.2	86
OC-Reptile	51.9	52.1	51.3	51.6	53.2	51	51.4	53.2
OC-FOMAML	55.7	74.7	79.1	58.6	66.1	87.5	91.8	73.2
OC-MAML (ours)	**69.1**	**96.6**	**88**	**96.6**	**76.2**	**97.6**	**95.1**	**95.7**

Table 1: Accuracies (in $\%$) computed on the class-balanced test sets of the test tasks of MiniImageNet (MIN), Omniglot (Omn), MT-MNIST with $T_{test} = T_0$ and STS-Sawtooth (Saw).

Several previous meta-learning approaches, e.g., MAML (Finn, Abbeel, and Levine 2017), were evaluated in a transductive setting, i.e., the model classifies the whole test set at once which enables sharing information between test examples via BN (Nichol and Schulman 2018). In anomaly detection applications, the CIR of the encountered test set batches, and therefore the statistics used in BN layers, can massively change depending on the system behavior (normal or anomalous). Hence, we evaluate all methods in a non-transductive setting: we compute the statistics of all BN layers using the few one-class adaptation examples and use them for predictions on test examples. This is equivalent to classifying each test example separately. We also use this method during meta-training. We note that the choice of the BN scheme heavily impacts the performance of several meta-learning algorithms (Bronskill et al. 2020).

Validation on the CNC-Milling Real-World Dataset. We validate OC-MAML on the industrial sensor readings dataset CNC-MDD and report the results in Table 2. We compute F1-scores for evaluation since the test sets are class-imbalanced. Depending on the type of the target milling operation (e.g., roughing), tasks created from *different* operations from the same type are used for meta-training. OC-MAML consistently achieves high F1-scores between 80% and 95.9% across the 6 milling processes. The high performance on the minority class, i.e., in detecting anomalous data samples, is reached by using only $K = 10$ non-anomalous examples ($c = 0\%$). These results show that OC-MAML yielded a parameter initialization suitable for learning OCC tasks in the time-series data domain and the maturity of this method for industrial real-world applications. Due to the low number of anomalies, it is not possible to apply MAML with the standard sampling, which would require K anomalous examples in the inner loop during meta-training. With OC-MAML, the few anomalies available are only used for the outer loop updates. We note that despite the high class-imbalance in the data of the meta-training processes, class-balanced query batches were sampled for the outer loop updates. This can be seen as an under-sampling of the majority class.

F_1	F_2	F_3	F_4	R_1	R_2
80.0%	89.6%	95.9%	93.6%	85.3%	82.6%

Table 2: OC-MAML F1-scores, averaged over 150 tasks sampled from the test operations, on finishing (F_i) and roughing (R_j) operations of the real-world CNC-MMD dataset, with only $K = 10$ normal examples ($c = 0\%$).

Model \ Dataset	MIN	Omn	MNIST	Saw
Reptile	0.05	0.02	0.16	0.02
FOMAML	0.13	0.14	0.31	-0.02
MAML	0.28	0.16	0.45	0.01
OC-Reptile	0.09	0.05	-0.09	0.03
OC-FOMAML	0.26	0.12	0.36	0.07
OC-MAML	**0.42**	**0.23**	**0.47**	**0.92**

Table 3: Cosine similarity between the gradients of one-class and class-balanced minibatches averaged over test tasks of MiniImageNet, Omniglot, MT-MNIST and STS-Sawtooth.

Cosine Similarity Analysis. We would like to directly verify that OC-MAML maximizes the inner product, and therefore the cosine similarity, between the gradients of one-class and class-balanced batches of data, while the other meta-learning baselines do not (see theoretical analysis). For this, we use the initialization meta-learned by each algorithm to compute the loss gradient of K normal examples and the loss gradient of a disjoint class-balanced batch. We use the best performing initialization for each meta-learning algorithm and compute the cosine similarities using on test tasks.

Support set size	$K = 2$			$K = 10$		
Model \ Dataset	MIN	CIFAR-FS	FC100	MIN	CIFAR-FS	FC100
MAML	62.3	62.1	55.1	65.5	69.1	61.6
OC-MAML (ours)	**69.1**	**70**	**59.9**	**76.2**	**79.1**	**65.5**
MetaOptNet	50	56	51.2	56.6	74.8	53.3
OC-MetaOptNet (ours)	**51.8**	**56.3**	**52.2**	**67.4**	**75.5**	**59.9**
MetaSGD	65	58.4	55	73.6	71.3	61.3
OC-MetaSGD (ours)	**69.6**	**71.4**	**60.3**	**75.8**	**77.8**	**64.3**
One-Way ProtoNets (Kruspe 2019)	67	70.9	56.9	74.4	76.7	62.1

Table 4: Test accuracies (in %) computed on the class-balanced test sets of the test tasks of MiniImageNet (MIN), CIFAR-FS and FC100 after using a one-class support set for task-specific adaptation

We report the mean cosine similarity on 3 image datasets and one time-series dataset in Table 3. The significant differences in the mean cosine similarity found between OC-MAML and the other meta-learning algorithms consolidate our theoretical findings.

Applicability to Further Meta-Learning Algorithms and Comparison to One-Way ProtoNets. To investigate whether the benefits of our sampling strategy generalize to further meta-learning algorithms beyond MAML, we apply it to MetaOptNet (Lee et al. 2019) and Meta-SGD (Li et al. 2017). Like MAML, these algorithms use a bi-level optimization scheme (inner and outer loop optimization) to perform few-shot learning. This enables the application of our proposed data strategy which requires two sets of data with different CIRs to be used. We refer to the OC versions of these algorithms as OC-MetaOptNet and OC-MetaSGD.

MetaOptNet trains a representation network to extract feature embeddings that generalize well in the FS regime when fed to linear classifiers, e.g., SVMs. For that, a differentiable quadratic programming (QP) solver (Amos and Kolter 2017) is used to fit the SVM (Lee et al. 2019) (inner loop optimization). The loss of the fitted SVM on a held-out validation set of the same task is used to update the representation network (outer loop optimization). Since solving a binary SVM requires examples from both classes and our sampling strategy provides one-class examples in the inner loop, we use an OC-SVM (Schölkopf et al. 2000) classifier instead. The embeddings extracted for few normal examples by the representation network are used to fit the OC-SVM, which is then used to classify the class-balanced validation set and to update the embedding network, analogously to the class-balanced scenario. To fit the OC-SVM, we solve its dual problem (Schölkopf et al. 2000) using the same differentiable quadratic programming (QP) solver (Amos and Kolter 2017) used to solve the multi-class SVM in (Lee et al. 2019). The ResNet-12 architecture is used for the embedding network. We use the meta-validation tasks to tune the OC-SVM hyperparameters.

Meta-SGD meta-learns an inner loop learning rate for each model parameter besides the initalization. Our episode sampling method is applied as done for MAML. Unlike the class-balanced MetaSGD, the meta-learning optimization assigns negative values to some parameter-specific learning rates to counteract overfitting to the majority class, which leads to performing gradient ascent on the adaptation loss. To prevent this, we clip the learning rates between 0 and 1.

Table 4 shows that applying the proposed sampling technique to MetaOptNet and Meta-SGD results in a significant accuracy increase in FS-OCC on the MiniImageNet, CIFAR-FS and FC100 datasets. Eventhough MetaOptNet substantially outperforms MAML and Meta-SGD in the class-balanced case (Lee et al. 2019), it fails to compete in the FS-OCC setting, suggesting that meta-learning a suitable initialization for the classifier is important in this scenario.

Finally, we compare to One-Way ProtoNets [2] and find that OC-MAML and OC-MetaSGD significantly outperform it on all three datasets. The poorer performance of One-Way ProtoNets and OC-MetaOptNet could be explained by the absence of a mechanism to adapt the feature extractor (the convolutional layers) to the unseen test tasks. OC-MAML and OC-MetaSGD finetune the parameters of the feature extractor by the means of gradient updates on the few normal examples from the test task. We conducted experiments using 5 different seeds and present the average in Table 4.

Conclusion

This work addressed the novel and challenging problem of few-shot one-class classification (FS-OCC). We proposed an episode sampling technique to adapt meta-learning algorithms designed for class-balanced FS classification to FS-OCC. Our experiments on 8 datasets from the image and time-series domains, including a real-world dataset of industrial sensor readings, showed that our approach yields substantial performance increase on three meta-learning algorithms, significantly outperforming classical OCC methods and FS classification algorithms using standard sampling. Moreover, we provided a theoretical analysis showing that class-balanced gradient-based meta-learning algorithms (e.g., MAML) do not yield model initializations suitable for OCC tasks and that second-order derivatives are needed to optimize for such initializations. Future works could investigate an unsupervised approach to FS-OCC, as done in the class-balanced scenario (Hsu, Levine, and Finn 2018).

[2] We re-implemented One-Way ProtoNets to conduct the experiments, since the code from the original paper was not made public.

References

Aggarwal, C. C. 2015. Outlier analysis. In *Data mining*, 237–263. Springer.

Amos, B.; and Kolter, J. Z. 2017. Optnet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 136–145. JMLR. org.

An, J.; and Cho, S. 2015. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE* 2: 1–18.

Andrews, J. T.; Tanay, T.; Morton, E. J.; and Griffin, L. D. 2016. Transfer representation-learning for anomaly detection. ICML.

Bronskill, J.; Gordon, J.; Requeima, J.; Nowozin, S.; and Turner, R. E. 2020. TaskNorm: Rethinking Batch Normalization for Meta-Learning. *arXiv preprint arXiv:2003.03284* .

Caruana, R. 1997. Multitask learning. *Machine learning* 28(1): 41–75.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41(3): 15.

Chen, J.; Sathe, S.; Aggarwal, C.; and Turaga, D. 2017. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, 90–98. SIAM.

Erfani, S. M.; Rajasegarar, S.; Karunasekera, S.; and Leckie, C. 2016. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition* 58: 121–134.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1126–1135. JMLR. org.

Garcia-Teodoro, P.; Diaz-Verdejo, J.; Maciá-Fernández, G.; and Vázquez, E. 2009. Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security* 28(1-2): 18–28.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Hawkins, S.; He, H.; Williams, G.; and Baxter, R. 2002. Outlier detection using replicator neural networks. In *International Conference on Data Warehousing and Knowledge Discovery*, 170–180. Springer.

Hinton, G. E.; and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *science* 313(5786): 504–507.

Hsu, K.; Levine, S.; and Finn, C. 2018. Unsupervised learning via meta-learning. *arXiv preprint arXiv:1810.02334* .

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* .

Khan, S. S.; and Madden, M. G. 2014. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review* 29(3): 345–374.

Kozerawski, J.; and Turk, M. 2018. CLEAR: Cumulative LEARning for One-Shot One-Class Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3446–3455.

Kruspe, A. 2019. One-Way Prototypical Networks. *arXiv preprint arXiv:1906.00820* .

Lake, B.; Salakhutdinov, R.; Gross, J.; and Tenenbaum, J. 2011. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10657–10665.

Lee, Y.; and Choi, S. 2018. Gradient-based meta-learning with learned layerwise metric and subspace. *arXiv preprint arXiv:1801.05558* .

Li, Z.; Zhou, F.; Chen, F.; and Li, H. 2017. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835* .

Liu, F. T.; Ting, K. M.; and Zhou, Z.-H. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, 413–422. IEEE.

Moya, M. M.; Koch, M. W.; and Hostetler, L. D. 1993. One-class classifier networks for target recognition applications. *NASA STI/Recon Technical Report N* 93.

Nichol, A.; and Schulman, J. 2018. Reptile: a Scalable Metalearning Algorithm. *arXiv preprint arXiv:1803.02999* .

Oreshkin, B.; López, P. R.; and Lacoste, A. 2018. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, 721–731.

Prastawa, M.; Bullitt, E.; Ho, S.; and Gerig, G. 2004. A brain tumor segmentation framework based on outlier detection. *Medical image analysis* 8(3): 275–283.

Ravanbakhsh, M.; Nabi, M.; Sangineto, E.; Marcenaro, L.; Regazzoni, C.; and Sebe, N. 2017. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, 1577–1581. IEEE.

Ravi, S.; and Larochelle, H. 2017. Optimization as a Model for Few-Shot Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL https://openreview.net/forum?id=rJY0-Kcll.

Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *International Conference on Machine Learning*, 4393–4402.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.;

Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3): 211–252. doi:10.1007/s11263-015-0816-y.

Sabokrou, M.; Khalooei, M.; Fathy, M.; and Adeli, E. 2018. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3379–3388.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Schmidt-Erfurth, U.; and Langs, G. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, 146–157. Springer.

Schmidhuber, J. 1987. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. Ph.D. thesis, Technische Universität München.

Schölkopf, B.; Platt, J. C.; Shawe-Taylor, J.; Smola, A. J.; and Williamson, R. C. 2001. Estimating the support of a high-dimensional distribution. *Neural computation* 13(7): 1443–1471.

Schölkopf, B.; Williamson, R. C.; Smola, A. J.; Shawe-Taylor, J.; and Platt, J. C. 2000. Support vector method for novelty detection. In *Advances in neural information processing systems*, 582–588.

Scime, L.; and Beuth, J. 2018. Anomaly detection and classification in a laser powder bed additive manufacturing process using a trained computer vision algorithm. *Additive Manufacturing* 19: 114–126.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 4077–4087.

Spigler, G. 2019. Meta-learnt priors slow down catastrophic forgetting in neural networks. *arXiv preprint arXiv:1909.04170* .

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Suteu, M.; and Guo, Y. 2019. Regularizing Deep Multi-Task Networks using Orthogonal Gradients. *arXiv preprint arXiv:1912.06844* .

Tax, D. M.; and Duin, R. P. 2004. Support vector data description. *Machine learning* 54(1): 45–66.

Triantafillou, E.; Zhu, T.; Dumoulin, V.; Lamblin, P.; Xu, K.; Goroshin, R.; Gelada, C.; Swersky, K.; Manzagol, P.; and Larochelle, H. 2019. Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. *CoRR* abs/1903.03096. URL http://arxiv.org/abs/1903.03096.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080* .

Xu, D.; Ricci, E.; Yan, Y.; Song, J.; and Sebe, N. 2015. Learning Deep Representations of Appearance and Motion for Anomalous Event Detection. *Procedings of the British Machine Vision Conference 2015* doi:10.5244/c.29.8. URL http://dx.doi.org/10.5244/C.29.8.