Learning by Small Loss Approach Multi-label to Deal with Noisy Labels

Vitor Sousa(✉), Amanda Lucas Pereira, Manoela Kohler, and Marco Pacheco

Pontifícia Universidade Católica Do Rio de Janeiro, Rio de Janeiro, RJ, Brazil
vitorbds@aluno.puc-rio.br
https://puc-rio.ai/

Abstract. Noisy data samples is a common problem for deep learning models applied to real-world applications. In this context, noisy samples refer to samples with incorrect labels, which can potentially degenerate the robustness of a model. Several works account for this issue in multi-class scenarios. However, despite a number of possible applications, multi-label noise remains an under-explored research field. In this work, two novel approaches to handle noise in this scenario are presented. First, we propose a new multi-label version of the Small Loss Approach (SLA), formerly multi-class, to handle multi-label noise. Second, we apply the multi-label SLA to a novel model, *Learning by SLA Multi-label*, based on Co-teaching. The proposed model achieves a performance gain of 15% in the benchmark UcMerced when compared to its baseline Co-teaching and a standard model (without any noise-handling technique). In addition, the model is also evaluated in a real-world scenario of underwater equipment imagery classification, yielding a relative improvement of 9% in F1-Score.

Keywords: Deep Learning · Noisy Samples · Noisy Labels

1 Introduction

Fully-supervised deep learning models have a large number of applications in Computer Vision, solving a variety of multi-label classification problems [1,5,7,34]. However, a large amount of carefully annotated samples is required to train such models, which can limit the performance obtained for some scenarios. Depending on the problem at hand, it may be difficult to obtain an equally representative number of labeled samples for each class. For real-world industry applications, labels can pose an even greater challenge, since it commonly demands a group of highly specialized personnel to be able to provide a set of accurate labels.

Common approaches to mitigate the problem of labels' availability are: the use of crowdsourcing [31], web queries [22] and the reliance on non-specialists in

Supported by organization Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023, LNCS 13956, pp. 385–403, 2023.
https://doi.org/10.1007/978-3-031-36805-9_26

the process of labeling the set. Such solutions, although they may help increase the number of available samples, can potentially yield a high presence of noisy labels [12,21]. In this context, *noisy labels* refer to the fact that some samples from the set are incorrectly labeled, which can degenerate the robustness of the deep learning model [12].

Several works have been developed in order to solve the problem of label noise in multi-class scenarios [12,15,29,35]. Benchmark datasets to compare the performance of models are available in the literature [32], and a variety of approaches to handle this type of noise can be found, each with its own advantages [12,15,21]. On the other hand, the development of models that are able to account for multi-label noise remains a less explored field [3], where both evaluation protocols and benchmark datasets are still at early stage, being investigated in works such as [3,14,28].

Most multi-class models supposedly can be directly adjusted to fit multi-label scenarios [3], as in [15,20,29]. However, we identified that the most common strategy from state-of-the-art (SOTA) methods for multi-class noisy labels – which relies mainly on Small Loss Approach (SLA) [12], used in several works [26,29,35,36] –, presents two major problems when directly applied to a multi-label scenario: (1) the possibility of noise being present only in some of the classes whilst others may be clean for the same sample; and (2) the SLA excludes samples with any level of noise from the training process.

Therefore, to alleviate the problem of label noise in multi-label scenarios, a novel multi-label version of the SLA is proposed (SLAM, as in SLA Multi-label), on which we mitigate the problems mentioned above. Furthermore, we also propose a new model, "Learning by SLA Multi-label", based on Co-teaching [12], where we replace the former SLA with its new multi-label version. The model is evaluted in a noisy version of the multi-label benchmark UcMerced [4, 33], where we synthetically injected 25% multi-label noise in mixed form [3]. We share the noisy version of the benchmark along with a clean test set in [https://github.com/ICA-PUC/Uc MercedNoiseDataset] for reproducibility and also for future works that wish to assess the results of new models with comparison to ours under fair conditions.

We compare the results obtained by our model to the model Co-teaching (adapted to multi-class), and also to a standard baseline, where no noise-handling method is used during the training procedure. Our model reduced the present noise from 25% to 5%, yielding an improvement in the F1-score measured over the test set by 15%. In addition, we evaluate our model on a real-world industry problem of underwater equipment imagery classification, where the model tries to identify event(s) of interest given a frame from an underwater inspection video. Specialists from an oil&gas company observed the presence of noise in the set, so the methods developed in this work were applied. We share the results obtained for a controlled setting, where the same amount and type of multi-label noise mentioned above is injected. We note that the latter is a private set so images cannot be disclosed, as requested by the data owners, therefore the figures presented as for this application are fictional.

The main contributions of our work are the following:

1. We propose a new multi-label version of the Small Loss Approach (SLA) (formerly multi-class). This modification can be either applied to the development of new models, or in the adaptation of multi-class models to multi-label scenarios.
2. We propose a novel model that handles multi-label noise
3. We disclose a new public noisy version of the multi-label benchmark Uc Merced for reproducibility and comparison for future works under fair conditions
4. We evaluate the proposed model in a real-world industry application

This paper is organized as follows: in Sect. 2, we outline the main existent multi-label and multi-class models to deal with noise labels, going into detail for the SLA and the Co-teaching model; then, in Sect. 3, we describe the several modifications made in SLA to be suitable to multi-label and present our proposed model *Learning by Small Loss Aproach Multi-label*. In Sect. 4, we provide the experimental setup and a description for each dataset. In Sect. 5, the results of our model *Learning by Small Loss Aproach Multi-label* are presented and discussed. Finally, in Sect. 6 we present the conclusion and outline future works.

2 Related Work

2.1 Deep Learning Models for Multi-class Noise

The use of deep learning models to deal with multi-class noisy labels is a well explored issue in the literature, with a variety of works addressing the problem [12,15,36]. One common approach to deal with noisy label is based on the Transition Matrix [21,24,27]. However, in some cases, where the number of classes is large, the estimation of the Transition Matrix may be a challenging task [12]. In addition, some authors propose using a loss function robust to noisy labels [11,20,30], such as the design of a regularization term to prevent the memorization of noisy labels [20].

Another popular approach focuses on the selection of *clean samples* from the group of available data points, discarding the remainder during training [17,23]. Clean samples refer to data points with correct labels, whilst corrupted labels present some level of noise. Current state-of-the-art (SOTA) methods such as Co-teaching [12], Co-teaching+ [36], JoCoR [29] and JoCoR-SLR [35], use the small loss strategy [12] – which we refer to as Small Loss Approach (SLA) – to perform the samples selection, outperforming standard models in benchmarks with noisy samples as in [21].

This strategy is based on the value of the loss function to determine the clean samples. Once the clean samples are elected, these methods differ in the strategy adopted to leverage information from the clean set in order to improve the model's performance.

2.2 Deep Learning Models for Multi-label Noise

In contrast to the multi-class scenario, where there are a variety of established methods and benchmark datasets [32], multi-label classification with noisy labels remains relatively unexplored in literature. Although there are papers with multi-label applications, the approaches and datasets used are mostly divergent and the topic deserves more attention [3].

Some papers focus on *multi-task learning* where the network tries to simultaneously learn to clean and correctly classify the samples [28]. There is also an intersection of the field with the study of models that handle noise by a *partial multi-label learning* approach, where the labels for each sample are considered as “candidates" rather than assuming they are correct [14]. The main difference is that partial multi-label learning does not assume that noise can occur for all labels, only the relevant ones.

On the recent work [3], the authors propose adapting the models [29], ElR [20] and SAT [15] from the multi-class noise scenario to fit a multi-label application which handles noisy labels. The main modifications comprise exchanging the former categorical cross entropy loss and softmax activation function at the final layer of the network for a binary cross entropy and a sigmoid function.

Other works as [13,16] use a strategy based on teacher-student network, where the teacher supervises the training process of the student. The supervision is based on training the teacher on a clean subset, so it can clean the noisy labels for the student.

2.3 Small Loss Approach

Small Loss Approach (SLA) [12] is a technique widely used by SOTA models for multi-class noise [29,35,36]. It consists of selecting clean (correct) samples from the dataset and discarding the ones with corrupted labels. Based on the assumption that models trained with fewer noisy samples perform better, only the clean set is used for training.

During the training process, the samples that present the lowest loss value over the batch of images are chosen for the clean set. Generally, this process is done at every iteration for each epoch. Once the clean instances are selected, training can be performed through different strategies aiming to fully exploit the clean set.

This technique is based on observation that Deep learning models tends to learn easy instances first, and gradually adapt to hard instance over the epochs [12]. Hence, it is expected that mislabeled samples to be memorized at more advanced time during training.

2.4 Co-teaching

On the paper that introduces the method Co-teaching [12], the authors propose using two networks trained simultaneously with SLA to deal with noisy labels.

Since the SLA method is based on the fact that networks memorize easier patterns – which may correspond to clean labels– earlier in the training process [2]. Therefore, each network should be able to identify different errors from the labels, making the learning process more robust to noise.

During training, each network receives as input the entire mini-batch containing both noisy and clean samples. Then, through SLA, each net selects which of the presented samples are clean and should be backpropagated through its peer network. After each net performs their selection of clean samples, the chosen ones are used for learning whilst the others are discarded for the given mini-batch.

The networks share architecture but do not share parameters. The reason to train both networks simultaneously is based on the fact that two networks can identify different types of errors from the noisy labels, since they have different learning abilities [29], i.e. different starting weights. Therefore, when the networks exchange the selected clean samples in each mini-batch, error should be mutually reduced by their peer.

3 Learning by Small Loss Approach Multi-label

The model Learning by Small Loss Approach Multi-label is based on SLA, which is used on models for Single Noise Labels. However, there are two major problems in directly adapting the SLA to the Multi-Label Noise scenario. Therefore, we make several changes in SLA in order to address such problems, and this modified version of SLA we refer to as *Small Loss Approach Multi-Label* (SLAM).

The first problem is that, for multi-label data, given one sample, some classes can present noise whilst others do not (Fig. 1). In Fig. 1(A), we see a version of the image with clean labels, while in Fig. 1(B) and Fig. 1(C), we see two versions of noise labels for the same sample image. In this example, the classes Baresoil, Buildings, Car and Pavement are correctly for both scenarios (Fig. 1(B) and Fig. 1(C)). On the other hand, we note there is noise for the class Grass, since it is absent on the labels in Fig. 1(B); and mislabeled as Airplane in Fig. 1(C). Therefore, a multi-labeled sample image can contain both correct and incorrect labels.

Based on the different manners that noise can be present for a multi-label sample as described above, the first proposed modification is to perform SLA analysis *by class* instead of *by sample*. Since some classes may be noisy while others remain clean (Fig. 1), we modify SLA to be applied separately for each class. Therefore, the SLA is performed in the following manner: calculate the cost for each class independently using the same loss function; select as *clean* the *small loss instances over all classes*. In this manner, it is possible to identify the noise by class for each image, allowing the model to keep the cleans classes of an image during training while discarding the noisy ones.

The second problem is that the former SLA excludes samples with noise from the training stage. This can be an issue due to the fact that the discarded samples may contain important features of a class to which it belongs. Hence, excluding such samples from training, deep Learning models can be deprived from learning important features that could potentially improve the generalization of the

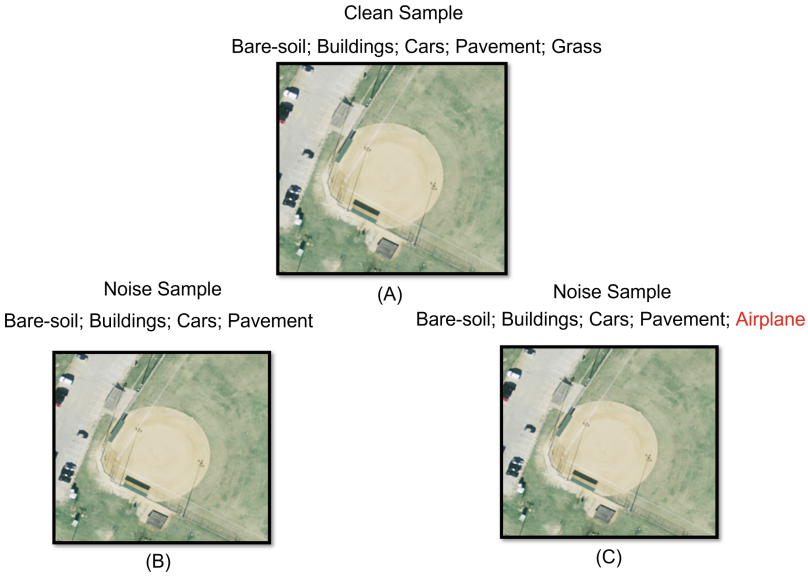

Fig. 1. (A) Example of clean sample of multi-label. (B) Example of noise sample where the class Grass was not annotated. (C) Example of noise sample where the class Grass was mislabel for Airplane class

model. To address this problem, we first observe that, for each one-hot encoded class, there are only two options: the class is either present in the image – **1** – or the class is not present in the image – **0**. So, once the class is identified as a noise label, we simply invert the label signal from **1** to **0** or from **0** to **1** to turn the noise label into a clean label.

3.1 Small Loss Approach Multi-Label(SLAM)

Given a dataset $D = (x_i, y_i)_{i=1}^{z}$ where x_i is the i-th sample of D and y_i is the multi-label class vector $y_i = [c_i^1, c_i^2, ..., c_i^M]$ with $c_i \in [0, 1]$ and **0** indicating the absence of a class and **1** indicating presence, for a number of classes M. Given $P_{model}(y|x_i; \theta)$ a model that produces a prediction over the classes labels vector y for a input sample x_i with parameters θ. Finally, the vector of the cross entropy loss ${CE_M}^i$ for sample x_i is given by:

$$CE_M(x_i) = -y_i \odot \log(P_{model}(y|x_i; \theta)) \tag{1}$$

where $\odot$ is an element-wise multiplier operator and log is also applied element-wise. So $CE_M(x_i)$ is the loss vector for all classes of instance x_i, being $CE_M(x_i) = [l_i^1, .., l_i^M]$, where l_i is loss for the i-th class. Then, CE_M for a batch of size B from the dataset D can be concatenated and be represented as matrix MC with B lines and M columns, MC_{BM}:

$$MC_{BM} = \begin{bmatrix} l_{11}\ l_{12}\ l_{13} \ldots l_{1M} \\ l_{21}\ l_{22}\ l_{23} \ldots l_{2M} \\ \ldots\ldots\ldots\ldots\ldots \\ l_{B1}\ l_{B2}\ l_{B3} \ldots l_{BM} \end{bmatrix} \tag{2}$$

In SLAM, we calculate the k maximum values for each column from MC_{BM}, where each column represents one class of y. k is an integer value, estimated by a function of the noise present in the specific class. This parameter can either be a fixed value or adjusted over the epochs. The vector of indexes $Y(m)$ of the k maximum values for the column m of MC_{BM} is given by:

$$Y(m) = arg\ max_k(MC_{1:B,m}) \tag{3}$$

Therefore, we have $Y(m) = [id_m^1, id_m^2, id_m^k]$, where id_m is the index of matrix MC_{BM} for class m. The operator $arg\ max_k$ return the index of the k max values of the input vector. Then, $MC_{1:B,m}$ is a vector equal to all lines of column m of MC_{BM}.

The indexes of $Y(m)$ are equivalent to the index of the samples x in batch B, where the class label m will be inverted. We calculate $Y(m)$ for all M classes. Finally, we assign the new labels for each class m for the samples from batch B where the indexes are those from $Y(m)$. To assign the new label for class m of sample x with index id_m from $Y(m)$, we simply invert the label, i.e. if the label is **1** it will be **0** or if it is **0** it will be **1**.

3.2 The Proposed Approach

In this section, we propose a novel deep learning model to deal with multi-noise labels, *Learning by SLA Multi-label*, built upon the model Co-teaching [12], adding the SLAM.

In Fig. 2, the training flow is illustrated. Following [12], we use peer networks that are trained simultaneously over the same mini-batch, where each network will identify the noise samples of the batch individually, assigning new labels by the method SLAM. Then, the mini-batch with corrected labels assigned by each network is used to train its peer. With the assigned labels, each network calculates the loss function considering these new labels, and proceeds to the back-propagation with the optimizer.

The model procedure is detailed in Algorithm 1. The SLAM procedure is applied after some epochs defined experimentally as a hyperparameter $start_{epoch}$ in step 5 of Algorithm 1. In order to define the k maximum values used in Eq. 5 for each epoch, we use the function $R(t)$ adapted from [12]:

$$R(t) = min[\frac{\tau t}{T_k}, \tau] \tag{4}$$

where the minimum value between $\frac{\tau t}{T_{max}}$ and τ is returned. T_k is an integer value, hyperparameter of model; τ is the estimated noise from the dataset; and t is the training epoch. The k value is given by multiplying the $R(t)$ by the batch size,

i.e. $k = R(t) * batchsize$. k is linearly increased until it reaches the value τ (the estimated noise).

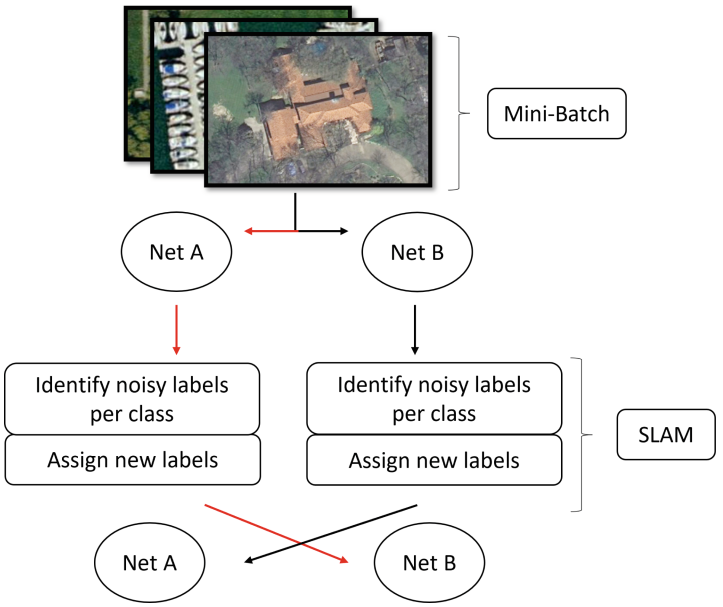

Fig. 2. Two networks, A and B, are trained simultaneously over the same mini-batch. The Network A will assign the new labels for network B, also B for A .

4 Experimental Details

In this section, the experiments are described. First, a thorough explanation of the method used to introduce noise for each dataset is provided. Then, the metric used to evaluate the performance of each model is introduced. Next, the benchmark and the real-world dataset of subsea equipment used to evaluate the model are presented. Finally, the experimental setup used for each set is explained.

4.1 Introduced Noise

To evaluate the robustness of our model, synthetically injected multi-label noise in mixed form is adopted, as in [3]. The noise was injected class-wise, and could be either additive or subtractive. Additive noise refers to adding a class that does not exist in the image, whereas subtractive noise excludes a class that appears in that sample.

Algorithm 1. Learning by SLA Multi-label Algorithm

Require: Network $w_f(\theta_1)$ and Network $w_g(\theta_2)$, learning rate η, fixed τ , epoch T_k, T_{max}, Noise Dataset $D = (x_i, y_i)_{i=1}^{z}$, number of class M, function $R(t)$,iteration I_{max},$batch_{size}$,$start_{epoch}$:

for $t = 1, 2..., T_{max}$ **do**
  Shuffle training set D
  for $n = 1, 2..., I_{max}$ **do**
    Fetch mini-batch D_n from D
    if $t > start_{epoch}$ **then**
      Create a copy $D_n^{w_g}$ from Dn
      Create a copy $D_n^{w_f}$ from Dn
      Calculate $CE_M(x) \forall x \in D_n$ with w_g
      Calculate $CE_M(x) \forall x \in D_n$ with w_f
      Obtain matrix $MC_{BM}^{w_g}$ from vectors $CE_M(x)$ of step 8
      Obtain matrix $MC_{BM}^{w_f}$ from vectors $CE_M(x)$ of step 9
      Calculate $Y(m)^{w_g}$ with Eq.5 $\forall\ m\ \in\ [1, .., M]$ using $MC_{BM}^{w_g}$ and k
      Calculate $Y(m)^{w_f}$ with Eq.5 $\forall\ m\ \in\ [1, .., M]$ using $MC_{BM}^{w_f}$ and k
      for $m = 1, 2..., M$ **do**
        Update Label $y_i[m]\ \forall\ i\ \in\ Y(m)^{w_g}$ of sample x_i in $D_n^{w_g}$
        Update Label $y_i[m]\ \forall\ i\ \in\ Y(m)^{w_f}$ of sample x_i in $D_n^{w_f}$
        ▷ Update Label is done by $logical_{not}$ operator.
        ▷ $logical_{not}$ $0 \rightarrow 1$; $1 \rightarrow 0$
      end for
    end if
    if $t > start_{epoch}$ **then**
      Update $\theta_1 = \theta_1 - \bigtriangledown L(w_g, D_n^{w_f})$
      Update $\theta_2 = \theta_2 - \bigtriangledown L(w_f, D_n^{w_g})$
    else
      Update $\theta_1 = \theta_1 - \bigtriangledown L(w_g, D_n)$
      Update $\theta_2 = \theta_2 - \bigtriangledown L(w_f, D_n)$
    end if
    Update: $R(t) = min[\frac{\tau t}{T_k}, \tau]$
    Update: $k = R(t) * batch_{size}$
  end for
end for

For both datasets, UcMerced and the Subsea Equipment, noise is added in the following manner: (1) given a class of interest, we randomly select 25% out of the total samples from the training set; (2) if the sample contains the class, we remove it (subtractive noise); if it does not, we add the class to its labels (additive noise).

We disclose the public noise version produced by adding the mixed noise to the dataset UcMerced, for future works that want compare the results in fair condition with our model.

The performance of the models are evaluated by comparison of the F1-Score measured over the test set per epoch as in the protocol used by [12,29,35,36]. The test set only contains clean samples, therefore the performance of the model is analyzed without the influence of noise.

4.2 Accuracy SLA Multi-label

The proposed method SLAM has its performance evaluated using the metric *Accuracy SLA Multi-Label*. This metric measures the accuracy of the labels assigned by SLAM for each class, at the end of every training epoch. The Accuracy SLA Multi-Label for class m is given by:

$$\text{Accuracy SLA Multi-Label(m)} = \frac{\#\text{ of correctly assigned labels}}{\#\text{ total of labels}} \tag{5}$$

It is important to highlight that the Accuracy SLA Multi-Label measures both class absence and class presence. So, the number of correctly assigned labels takes into account both presence and absence of corrected assignments performed by SLAM. The total of labels is equivalent to the number of samples from the set.

4.3 Dataset UCMerced Multi-label

We create a noisy version of the dataset UcMerced [4], which is a multi-label version of the former dataset [33]. This dataset contains 17 classes, depicting aerial satellite image scenarios of both urban and countryside regions. The classes include objects like cars, boats and airplanes, and also natural land-cover classes such as water, trees and buildings.

Following the protocol proposed by the authors from [12,29,35,36], first, the dataset is partitioned into train and test sets. For this dataset, this comprises of a training set of 1700 images and a test set of 217 images. Then, noise is introduced only in the training set while the test set remains clean. This version of the dataset is available at https://github.com/ICA9-PUC/UcMercedNoiseDataset.

4.4 Dataset Subsea Equipment

The Dataset Subsea Equipment is being elaborated in order to develop a deep learning model to assist specialists at underwater inspections of subsea equipment used for oil and gas extraction. It currently contains over 30,000 images with more than 39 potential classes per images, and it is a multi-label dataset. In this context, such potential classes refer to either structures (e.g. pipeline accessories) or events (e.g. a pipeline crossing).

The motivation behind testing our model in this set of data was built upon the notice of the presence of noise from underwater inspections specialists. Therefore, the experiments presented in this work were performed with a subset from this

set of data, part of the much larger set mentioned above, as a proof-of-concept, following the experimental setup described in Sect. 4.1.

The subset used for our experiments contain a total of 1327 images with 12 classes. The samples were all carefully revised in order to assure clean labels. Then, the set was partitioned into training and test sets, each with 664 and 664 images, respectively. As in the UCMerced Dataset (Sect. 4.3), noise is added only to training set whilst the test set is kept clean, following the protocol from [12,29,35,36].

The classes present at this set are the following: Anode, Anode Debris, Pipeline Buoy, Pipeline, Marine Life, Flange, Rope, Text, Chain, ROV, End Fitting, Pipeline Crossing. As the images are private, a fictional example queried from a search engine is presented in Fig. 3.

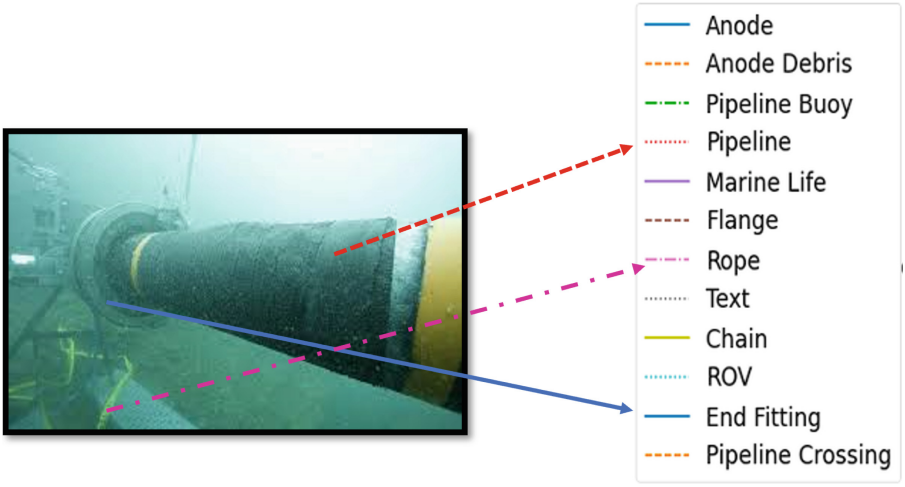

Fig. 3. Fictional example of an image from the dataset Subsea Equipment. This sample image depicts three different classes: Pipeline, Rope and End Fitting

4.5 Experimental Setup

We use peer networks $w_f(\theta_1)$ and $w_g(\theta_2)$ with VGG16 as the architecture [25], with pre-trained weights on ImageNet [9], and only the final 4 layers are optimized. To enable multiple predictions, we use a binary cross entropy loss, and exchange softmax by sigmoid layers. All input images to $(224, 224)$ and the default pre-processing of Tensorflow version 2.4 for VGG-16 is applied, using the function `vgg16.preprocess_input`. Model's performance is measured by f1-score [6] over test data, and the Accuracy SLA Multi-Label. The optimizer was the Adam [18] and the model was implemented using Tensorflow 2.4 and the experiments were carried out using the NVIDIA® Tesla® V100.

We compare our model with the standard VGG-16 network under same condition for multi-label. In addition, we compare our model with a Co-teaching model directly adapted for multi-label, i.e. using the conventional SLA for multi-class as pointed out in work [3]. The comparison of our model with this model is important to illustrate the performance gains obtained using SLAM instead of SLA, once this is the main difference between our model and Co-teaching model.

For the experiments with UcMerced, the learning rate η was set to 0.00025, and the network is trained for a total of epochs $T_{max} = 50$. The hyperparameter T_k was set to 40, the SLAM $start_{epoch}$ was 1, the estimated noise τ was 0.25, and a $batch_{size}$ 64. For UcMerced, the number of classes M was 17. For the experiments with Subsea Equipment, the learning rate η was 0.001, the total of epochs T_{max} was 200, $T_k = 30$, the SLAM $start_{epoch}$ was 5, the estimated noise τ was 0.25, the $batch_{size}$ 32. The number of classes M was 12.

5 Results and Discussion

5.1 Results Dataset UcMerced

In Fig. 4, the results obtained for the Accuracy SLA Multi-Label per epoch for all classes in dataset UcMerced is shown. The black solid line represents the initial clean percentage for all classes, which corresponds to a value of 75%, once an equal percentage of 25% of noise was introduced for each class. Every curve superior to this line indicates that the SLAM process is reducing the noise present for each class. On the other hand, a curve below that line reveals that there was an increase of noise for such class.

We note that the method SLAM was able to reduce the noise for most of the classes from UcMerced, achieving a percentage of 95% of clean samples for a few. Also, solely four classes present an increase of noisy samples during the process. The understanding of why such classes perform poorly remain an open question in this research. Intuitively, there may be some connection with the fact that the data is highly unbalanced. Nonetheless, this behaviour is not repeated on the results obtained for the Subsea Equipment dataset, so further experiments are required in order to reach a solid conclusion.

In Fig. 5, we compare our model (blue line) with a standard model (orange line) trained under the same conditions. Our model improved the F1-Score in 15%, showing that reducing noise per class-instance can significantly improve the results. As expected, this suggests that models trained on data with less noise tend to perform better.

We also compare our model with the model Co-teaching, directly adjusted from multi-class to the multi-label scenario (green line). First, we observe that the model Co-teaching can improve the performance when compared to the standard model (orange line). This illustrates that, indeed, the direct adaptation of SLA models to multi-class is possible as mentioned in [20]. However, the performance gains are only 2%, whereas when we replace the former SLA in Co-teaching by SLAM the performance improves significantly (blue line). It is important to highlight here that our model essentially replaces the SLA used in

Co-teaching with SLAM, so this result proves the benefit of replacing SLA by SLAM in Co-teaching model. We also show in Table 1, using ANOVA TEST [8], that the models belongs to different distributions. In the Table, STD refer to the standard model.

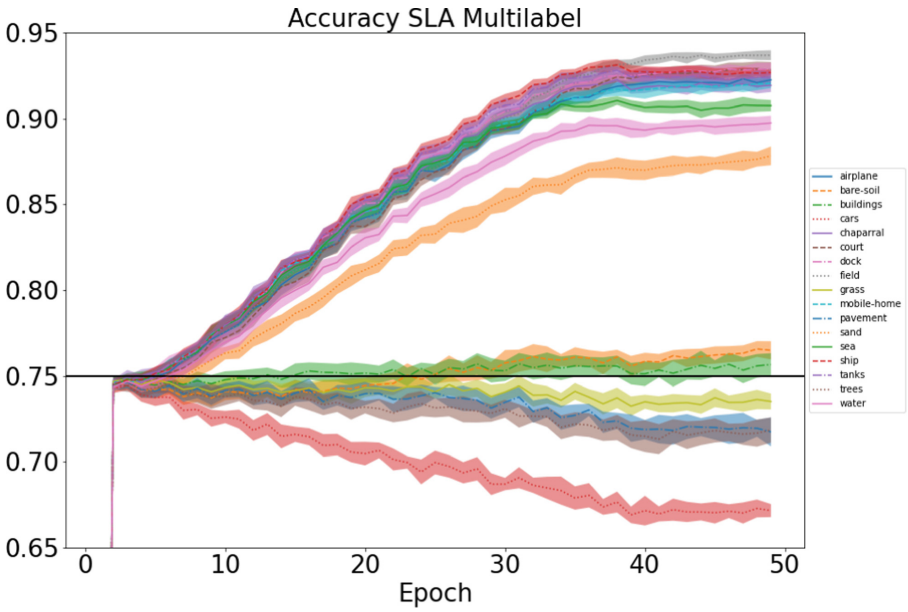

Fig. 4. Accuracy SLA Multi-Label for all classes in the dataset UcMerced. The curves over this line point a reduction in label noise, whilst below indicates an increase in noise during the SLAM process.

Table 1. ANOVA Test for the experiments. In the table, "STD" refer to the standard model.

Model 1	Model 2	Model 3	Dataset	Epoch	p-value	Distribution
SLAM	Co-teaching	STD	UcMerced	45	2^{-23}	Different
SLAM	Co-teaching	STD	Subsea	150	7^{-13}	Different

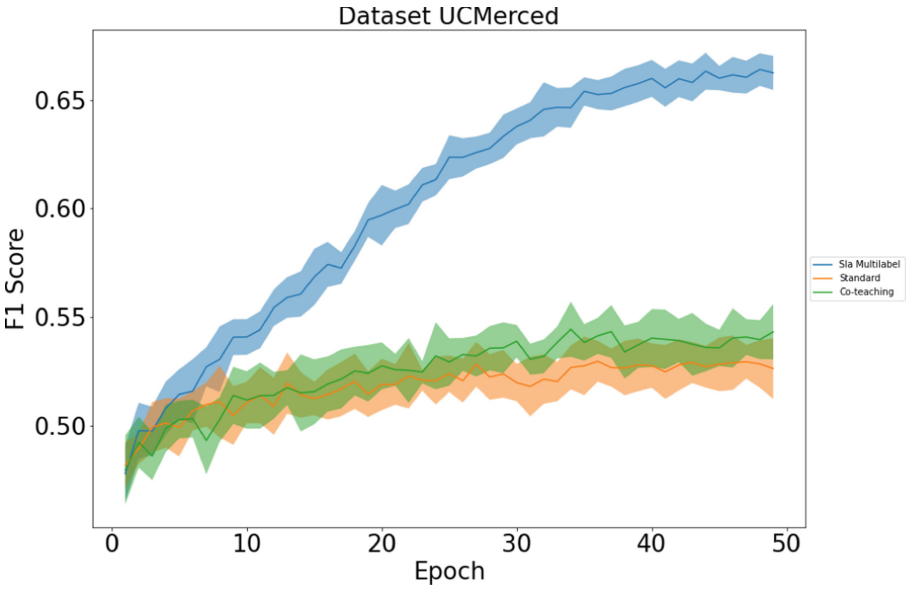

Fig. 5. Performance on the test set for the metric F1-Score

5.2 Results Dataset Subsea Equipment

In Fig. 6, the results obtained for the Accuracy SLA Multi-Label per epoch for all classes from the dataset Subsea Equipment is presented. The results point that SLAM was able to reduce the noise present at all classes, achieving a percentage of 90% clean samples for some classes. Contrary to the experiments from the SAR scenario, none of the classes had noise increase.

In epoch 25, we have an interesting observation, where we note a local peak performance to all classes in the Accuracy SLA Multi-Label. This local peak coincides with the performance peak observed for the F-1 score to our model (blue line in Fig. 7). Intuitively, fixing the labels assigned at this epoch would lead to an improvement in performance for the model. However, adding a hyperparameter to control this process would be extremely dataset-dependent, since such behavior was not observed in the experiments with UcMerced. Therefore, this hyperparameter may be hard to tune for real-world industry applications.

Despite that, an in-depth understanding of this phenomenon poses an interesting option for future works. Once in [12] we also observe a peak of Label Precision [12] around epoch 25 of SLA in single label for the datasets MNIST [10], CIFAR-10 and CIFAR-100 [19], the peak of performance in the Test accuracy [12] occurs at the same point. Understanding this process could enable an improvement at both the SLA and SLAM, thereby future models could be developed.

In Fig. 7, we compare our model with a standard model trained under same conditions. We note that our model improved the F1-Score in 8%, being a considerable improvement for a real-world industry application. We also present the comparison with the model Co-teaching, and as in the experiments with the UcMerced dataset, Co-teaching (green line) presented results that are superior to the ones from the standard model, but within a small margin of 2%. On the other hand, the performance gain that SLAM brings are solid, enabling the application of such method to real problems. We show in Table 1, using Analysis of Variance (ANOVA Test) [8], that the three models that were assessed (standard, Co-teaching and SLAM) belong to different distributions.

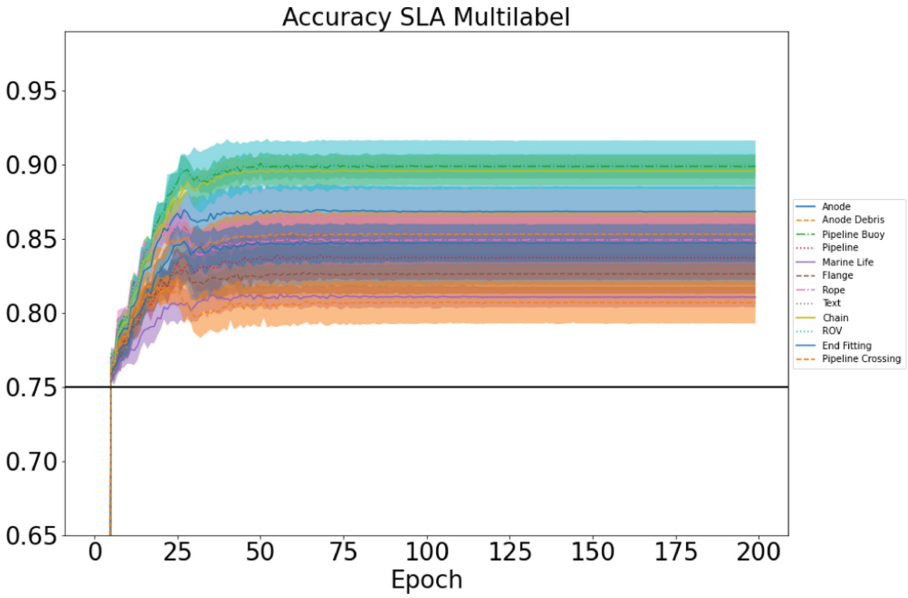

Fig. 6. Accuracy SLA Multi-Label for all classes in the dataset Subsea Equipment. The solid black line represents the initial amount of clean samples from the training set. The curves over this line point a reduction in label noise, whilst below indicates an increase in noise during the SLAM process.

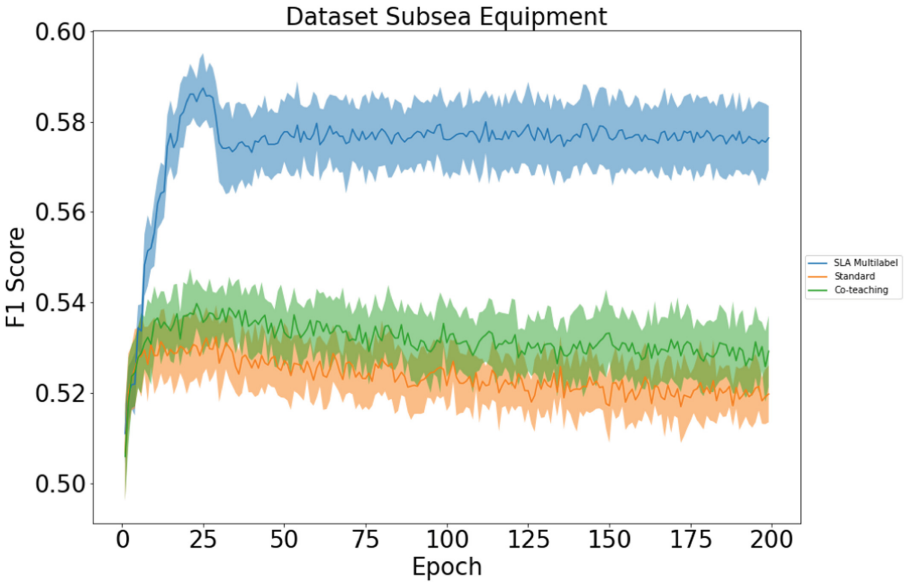

Fig. 7. Performance on the test set for the metric F1-Score

6 Conclusions and Future Work

In this work, we identified two major problems in directly adjusting the SLA (used by SOTA methods to deal with noisy label in multi-class problems) to the multi-label scenario. To address such problems, we propose several changes in the former SLA (for multi-class) in order to make it more suitable for multi-label applications. This new method we propose as Small Loss Approach Multi-label (SLAM). SLAM allows adapting formerly multi-class methods appropriately to multi-label scenarios, or it can be used to develop new methods and models. We also propose a new model, Learning by Small Loss Approach Multi-label, which is an adaptation of the model Co-teaching – used in multi-class noisy scenarios – to multi-label. The major difference of our model to the model Co-teaching is the replacement of SLA with the novel SLAM.

We evaluate our model in benchmark dataset UcMerced, and also provide the results obtained for a real-world industry application of Subsea Equipment, where we apply the model to underwater inspection videos. Our model significantly improved the results obtained for both datasets, with a performance gain in the F1-score of 15% for dataset Ucmerced and 9% for Subsea Equipment application in comparison to the standard model and also to the original Co-teaching (adapted to the multi-label scenario with its former SLA). The gain in performance illustrates the robustness of our model for dealing with noisy labels in multi-label scenarios, which emboldens its application in sets of data from either academia or industry problems.

Acknowledgements. The authors would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) and Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) for their financial support.

References

1. Ahmadi, Z., Kramer, S.: A label compression method for online multi-label classification. Pattern Recogn. Lett. **111**, 64–71 (2018)
2. Arpit, D., et al.: A closer look at memorization in deep networks. In: International conference on machine learning, pp. 233–242. PMLR (2017)
3. Burgert, T., Ravanbakhsh, M., Demir, B.: On the effects of different types of label noise in multi-label remote sensing image classification. IEEE Trans. Geosci. Remote Sens. **60**, 1–13 (2022)
4. Chaudhuri, B., Demir, B., Chaudhuri, S., Bruzzone, L.: Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method. IEEE Trans. Geosci. Remote Sens. **56**(2), 1144–1158 (2017)
5. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference On Computer Vision And Pattern Recognition, pp. 5177–5186 (2019)
6. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genom. **21**(1), 1–13 (2020)
7. Coulibaly, S., Kamsu-Foguem, B., Kamissoko, D., Traore, D.: Deep convolution neural network sharing for the multi-label images classification. Mach. Learn. Appl. **10**, 100422 (2022)
8. Cuevas, A., Febrero, M., Fraiman, R.: An anova test for functional data. Comput. Stat. Data Anal. **47**(1), 111–122 (2004)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference On Computer Vision And Pattern Recognition, pp. 248–255. IEEE (2009)
10. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process. Mag. **29**(6), 141–142 (2012)
11. Ghosh, A., Kumar, H., Sastry, P.S.: Robust loss functions under label noise for deep neural networks. In: Proceedings of the AAAI Conference On Artificial Intelligence, vol. 31 (2017)
12. Han, B., et al.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
13. Hu, M., Han, H., Shan, S., Chen, X.: Multi-label learning from noisy labels with non-linear feature transformation. In: Computer Vision-ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14. pp. 404–419. Springer (2019)
14. Hu, M., Han, H., Shan, S., Chen, X.: Weakly supervised image classification through noise regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11517–11525 (2019)
15. Huang, L., Zhang, C., Zhang, H.: Self-adaptive training: bridging the supervised and self-supervised learning. arXiv preprint arXiv:2101.08732 (2021)

16. Inoue, N., Simo-Serra, E., Yamasaki, T., Ishikawa, H.: Multi-label fashion image classification with minimal human supervision. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2261–2267 (2017)
17. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: International Conference On Machine Learning, pp. 2304–2313. PMLR (2018)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
20. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. Adv. Neural. Inf. Process. Syst. **33**, 20331–20342 (2020)
21. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Trans. Pattern Anal. Mach. Intell. **38**(3), 447–461 (2015)
22. Liu, W., Jiang, Y.G., Luo, J., Chang, S.F.: Noise resistant graph ranking for improved web image search. In: CVPR 2011, pp. 849–856. IEEE (2011)
23. Malach, E., Shalev-Shwartz, S.: Decoupling "when to update" from "how to update". In: Advances in Neural Information Processing Systems, vol. 30 (2017)
24. Sanderson, T., Scott, C.: Class proportion estimation with application to multiclass anomaly rejection. In: Artificial Intelligence and Statistics, pp. 850–858. PMLR (2014)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Song, H., Kim, M., Lee, J.G.: Selfie: Refurbishing unclean samples for robust deep learning. In: International Conference on Machine Learning, pp. 5907–5915. PMLR (2019)
27. Van Rooyen, B., Williamson, R.C.: A theory of learning with corrupted labels. J. Mach. Learn. Res. **18**(1), 8501–8550 (2017)
28. Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., Belongie, S.: Learning from noisy large-scale datasets with minimal supervision. In: Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition, pp. 839–847 (2017)
29. Wei, H., Feng, L., Chen, X., An, B.: Combating noisy labels by agreement: A joint training method with co-regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13726–13735 (2020)
30. Wei, T., Shi, J.X., Tu, W.W., Li, Y.F.: Robust long-tailed learning under label noise. arXiv preprint arXiv:2108.11569 (2021)
31. Welinder, P., Branson, S., Perona, P., Belongie, S.: The multidimensional wisdom of crowds. In: Advances in Neural Information Processing Systems, vol. 23 (2010)
32. Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: CVPR (2015)
33. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 270–279 (2010)
34. Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501 (2017)

35. Yao, Y., Sun, Z., Zhang, C., Shen, F., Wu, Q., Zhang, J., Tang, Z.: Jo-src: A contrastive approach for combating noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5192–5201 (2021)
36. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., Sugiyama, M.: How does disagreement help generalization against label corruption? In: International Conference on Machine Learning, pp. 7164–7173. PMLR (2019)