LEARNING TO UNDERSTAND EARTH OBSERVATION IMAGES WITH WEAK AND UNRELIABLE GROUND TRUTH

Rodrigo Caye Daudt[1,2], Adrien Chan-Hon-Tong[1], Bertrand Le Saux[1], Alexandre Boulch[1]

[1]DTIS, ONERA, Université Paris-Saclay, FR-91123 Palaiseau, France
[2]LTCI, Télécom ParisTech, FR-75013 Paris, France

ABSTRACT

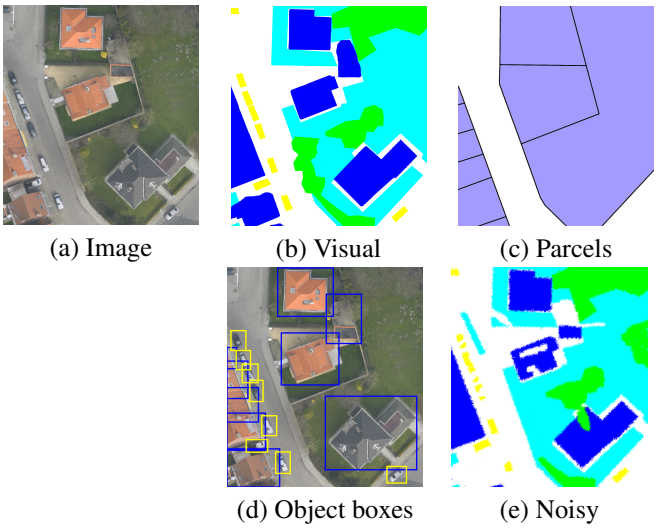
In this paper we discuss the issues of using inexact and inaccurate ground truth in the context of supervised learning. To leverage large amounts of Earth observation data for training algorithms, one often has to use ground truth which was not been carefully assessed. We address both the problems of training and evaluation. We first propose a weakly supervised approach for training change classifiers which is able to detect pixel-level changes in aerial images. We then propose a data poisoning approach to get a reliable estimate of the accuracy that can be expected from a classifier, even when the only ground-truth available does not match the reality. Both are assessed on practical land use and land cover applications.

***Index Terms*—** Weakly supervised learning, noisy data, Earth observation, change detection, data poisoning.

1. INTRODUCTION

Many of the recent advances in image understanding rely on machine learning algorithms which require large amounts of training data. More and more data are now online, but data which can be used for learning are scarce! Indeed, training data for learning a given task should be associated with information which indicates the desired output. For several Earth observation (EO) tasks, such as change detection or object recognition, very large labelled datasets which go beyond academic efforts are not easily available, mainly because manually annotating images for a task such as semantic segmentation can be very time consuming, and therefore costly.

One line of work that has gained traction recently is the field of weakly supervised learning, where the aim is to use labels that are easier to obtain to learn a more complex task [1]. Many strategies have thus been proposed to leverage large amounts of data at a cheaper cost. The first one is cross-referencing open datasets in order to create ground truth labels, such as mixing imagery with OpenStreetMap (crowd-sourced) or land lot information. It usually results in *inexact supervision*, because different sources imply different concepts: true land cover, administrative partition (parcels), user-oriented classes (see Fig. 1(c)). Second, another kind of inexact supervision is when one manually annotates datasets using



(a) Image (b) Visual (c) Parcels (d) Object boxes (e) Noisy

Fig. 1. Image (a) and various types of reference data: visual ground truth (b), parcels (c), objects detection boxes (d), and noisy visual annotation (e).

labels that are simpler to obtain than the final desired task, such as bounding boxes or image level tags for objects [2, 3, 4] (see Fig. 1(d)). *Inaccurate supervision* consists in using uncontrolled labels which might be noisy or false [5], for instance the results of an automatic process (see Fig. 1(e)). Actually, Earth observation questions the notion of defining a ground-truth: precise geo-localised cartography and images may differ due to incidence artefacts and ortho-rectification processes. How can a reference be defined?

These methods result in datasets that are very challenging in their usage for supervised learning. Automatically generated data and user provided data contain more noise than data annotated by trained analysts. Simpler types of annotations bring with them the challenge of training systems that understand the provided information and go further than what the given examples initially allowed.

In this paper we discuss these problems in the context of remote sensing. The first method, presented in Section 3, helps at training with inexact weak labels and is based on data cleaning. We assess the approach on a change detec-

tion problem with automatically generated labels. Then, we show in Section 4 how ideas from data poisoning can give us insights into the effects of label noise, allowing us to estimate upper and lower bounds on the performance of algorithms even when the only ground-truth available cannot be trusted.

2. RELATED WORK

The sensitivity of deep learning to label noise (*inaccurate supervision*) has been a topic of attention recently [5]. Lu et al. [4] proposed a method that performs weakly supervised learning for semantic segmentation that is robust to label noise by formulating the problem as a label noise reduction method based on L1 optimisation. Rolnick et al. [6] have shown that deep learning algorithms are quite robust to random label noise and proposed solutions to minimise the effect of label noise on the training process, such as increasing training batch size. Yet, this work has assumed unbiased label noise which is not always a valid assumption. Muoz-Gonzlez et al. [7] show that optimised noise can lead to large accuracy gaps.

Recent advances have been made in performing semantic segmentation from several types of weak labels (*inexact supervision*): points labels [3], image labels [4], and bounding box labels [2]. Khoreva et al. proposed in [2] recursive training schemes and showed that a naive recursive training scheme led to a decline in performance and obtained best results by applying problem heuristics to generate a training semantic segmentation dataset directly using an image classification network and class activation maps.

Supervised learning has a long history in remote sensing. It was proven efficient for land use and land cover classification [8]. Supervised learning techniques have been shown to perform change detection when it is treated as a semantic segmentation problem [9, 10]. In this context, open datasets are either small compared to other computer vision datasets [11, 9] or large in size but containing noisy labels [10].

3. THE PROBLEMS WITH TRAINING

The dataset presented in [10] is the first large scale dataset in the context of change detection. High resolution aerial image pairs were combined with openly available land cover vector data to generate pixel level labels for land cover maps and change maps. Three main sources contribute to label noise:

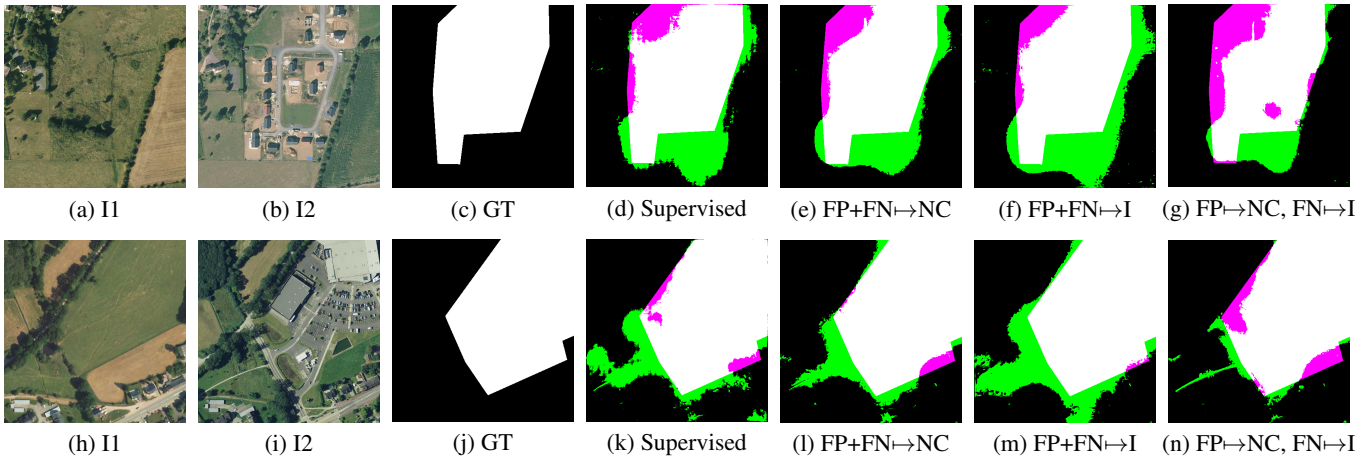
- The polygons used for generating the ground truth rasters are not true to the boundaries of the objects, they mark the land lots inside which a change has occurred (parcel case shown in Fig. 1(c)).
- There are mild discrepancies between the dates when the pictures were taken and when the vector maps were generated, and neither of these dates are available with the data.
- There are inaccuracies already present in the Urban Atlas data, as only 80-85% accuracy is guaranteed.

Using these data directly to train a change detection network works, but the output of the network consists of blobs around detected changes as it implicitly attempts to predict land lot information from the image [10]. With the aim of making more accurate change detections, we propose an iterative training scheme that builds upon the ideas proposed in [2]. Khoreva et al. have shown that simply using the output of the network as training data for itself results in a decrease in performance, but using problem specific heuristics between training iterations can lead to increases in performance. Many of the heuristics originally proposed in [2] are not applicable to the current problem as bounding box information is not available. Another difference is that in the case of this dataset a single polygon may contain several change objects while that is not the case for bounding boxes, which are assumed to contain a single object of a given class.

We propose an iterative training method that alternates between 1) training a fully convolutional neural network (FCNN) until convergence, and 2) using agreement between predictions and labels to clean the training data. The FCNN that was used here was the integrated terrain classification and change detection architecture referred to as Strategy 4.2 in [10]. For the sake of simplicity, only the change detection branch of the network is discussed here. For cleaning the data, three approaches were tested. If both the network prediction and the initial ground truth information agree on the pixel class, there is no reason to change it. If there is a disagreement, it is necessary to choose how to combine these sources of information.

1. **Mark all disagreements - false positive (FP) and false negatives (FN) - as no change (NC):** the motivation for this is the assumption that all disagreement come from overestimation of changes by either the ground truth or the current network (FP+FN$\mapsto$NC).
2. **Mark all disagreements as ignore (I):** it aims to discard all pixels where the network and the ground truth disagree in order to perform training using only pixels of whose labels we are confident (FP+FN$\mapsto$I).
3. **Mark false detections as no change and false negatives as ignore:** this hybrid strategy merges the two previous ideas into a single policy that assumes the network overestimates changes, but relies slightly more heavily on the original ground truth information (FP$\mapsto$NC, FN$\mapsto$I).

Figure 2 shows results obtained by the iterative training method for various scenarios. (a-b) and (h-i) are coregistered image pairs that are about six years apart. (c) and (j) contain



(a) I1 (b) I2 (c) GT (d) Supervised (e) FP+FN↦NC (f) FP+FN↦I (g) FP↦NC, FN↦I

(h) I1 (i) I2 (j) GT (k) Supervised (l) FP+FN↦NC (m) FP+FN↦I (n) FP↦NC, FN↦I

Fig. 2. Change detection results. (a)-(b) and (h)-(i): image pairs. (c) and (j): inaccurate ground truth labels. (d)-(g) and (k)-(n): predictions of FCNNs trained using the proposed iterated training strategies overlaid on original ground truth.

the initial ground truth labels obtained from Urban Atlas vector maps. Note that while changes actually occurred inside the marked areas, the boundaries are not precise and each region contains different types of changes, i.e. buildings, parking lots, roads, trees, etc.

Fig. 2 (d) and (k) display the change detection results from the network trained until convergence, i.e. 100 epochs, using only the initial ground truth labels. We then iterated the cleaning of the training dataset with further training of the network for 100 more epochs, repeating this process four times. To ensure an accurate comparison of methods, the same starting point was used for the FCNN weights on all tests. The results for methods 1, 2, and 3 for cleaning the training set are displayed in (e/l), (f/m), and (g/n) respectively. From these images we can see that FP↦NC, FN↦I, which combines the usage of an ignore class while prioritising ground truth data over network predictions, is able to learn through this iterative learning method to more accurately detect changes. Note in (g) how a patch of unchanged grass is no longer marked as change despite being surrounded by true changes, and in (n) how it was able to accurately mark the new pathway as a change without marking its surroundings as change. These results show clear improvements over the naive supervised training method.

4. THE EVALUATION PROBLEM

Another important question is how can we evaluate algorithms when only noisy data are available. For the evaluation of an already trained model, the issue is that if the available testing ground truth is not perfect, quantitative measures do not really reflect a given classifier's performance. For example, in the semantic segmentation context, producing a perfect visual prediction (e.g. Fig. 1(b)) will not lead to a 100% accuracy if ground truth is noised (e.g. Fig. 1(e)). More precisely, if a classifier has a true accuracy of $\mu\%$, and if $\rho\%$ of pixels have noisy labels in test set, then it follows that the measured accuracy is bounded in $[\mu - \rho, \mu + \rho]\%$. Of course, in some semantic segmentation datasets it is standard for edge labels to be discarded at testing time to calculate a fair metric, since edge pixels' labels are prone to being unreliable. Such *ad hoc* corrections can not be trivially designed for change detection based on a parcel ground truth (see Fig. 1(c)), and it may be dangerous to conclude the superiority of some methods over others using such quantitative metrics.

This problem is even more important when comparing different models. Let us consider two feature extractors: 1) a U-Net [12] cut before the last layer trained for semantic segmentation on a different dataset, and 2) a U-Net trained for auto encoding. In an attempt to know if one is better than the other to train an SVM considering any possible value of edge pixel labels, the minimal and maximal accuracies of each feature-SVM pipeline can be approximated by modifying edge pixel labels. We propose the application of data poisoning algorithms to estimate these bounds (see Fig. 3). The bound is obtained by:

- Considering the best/worse reachable classifier, e.g. by applying the training process to the testing data.
- Applying the best/worse classifier to the training data.
- Modifying training labels at the edges according to the best/worse classifier.
- Training from modified training data and measuring accuracy on testing data.

By modifying some training labels, we bend the energy landscape to make the training lead to a classifier closer to the best or worse one, eventually approximating a bound on accuracy of each feature-SVM pipeline according to edge pixel labels.

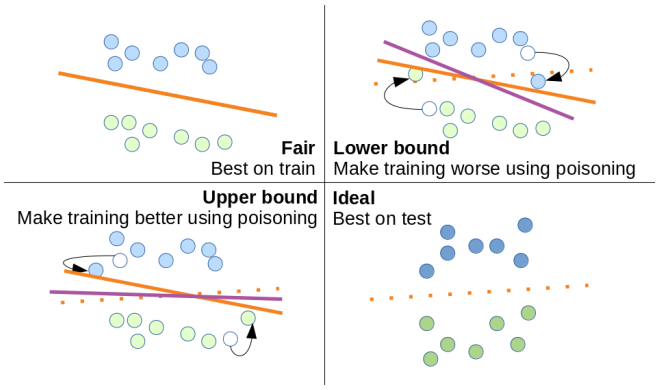


Fig. 3. Illustration of our procedure to bound accuracy gap due to the label noise. Poisoning is computed using test set and applied to training set, which leads to lower and upper bounds of test accuracy.

We show the feasibility of this algorithm using data from the Data Fusion Context 2015 [8]. We learn the last layer of the pretrained U-Net by stochastic gradient descent on image 315135_56865 where the labels of edge pixels are modified. Then, we evaluate the resulting U-Net on image 315140_56865. By modifying the pixels close to boundaries (4% of pixels), accuracy changes by as much as +4% to -6%, yielding an approximate bounds of the features accuracy.

A limitation of the proposed method is that the computation of the quality of the bound approximation is not trivial. Currently, state-of-the-art methods offer stronger ways to approximate such bounds only for convex and global classifiers. Indeed, the computation of our bounds would likely be better if the last layer were trained with state-of-the-art solvers like liblinear rather than with SGD. Unfortunately, encoding even only one these images leads to a 5 GB file intractable for these solvers. Also, the question of computing bounds for a complete deep network and not just for a feature-SVM pipeline seems theoretically feasible but intractable in practice [7].

Even if currently limited to few use cases, we highlight the usefulness of data poisoning tools to evaluate the impact of the noise at training time for semantic segmentation and the difficulty to quantify the performance of algorithms using ground truth with very different natures than visual ground truth. The calculated bounds contain not only an estimate of an algorithm's performance, but it also reflects our uncertainty regarding the calculated metric, which prevents us from being overconfident in the metrics when using them for comparisons or making decisions.

5. CONCLUSION

In this paper we discussed the problem of labelled data in the context of Earth observation image understanding. We analysed the sensitivity of supervised learning systems to noise in the data available for training and testing. We showed an example of how using problem specific heuristics to improve the training procedure can improve the obtained results. We also proposed a method that allows us to estimate the upper and lower bounds of the metrics used to evaluate and compare such algorithms. These ideas are essential to go beyond supervision and develop algorithms for EO data understanding.

6. REFERENCES

[1] O. Russakovsky et al., "Best of both worlds: human-machine collaboration for object annotation," in *CVPR*, 2015.

[2] A. Khoreva et al., "Simple does it: Weakly supervised instance and semantic segmentation," in *CVPR*, 2017.

[3] A. Chan-Hon-Tong and N. Audebert, "Object detection in remote sensing images with center only," in *IGARSS*, July 2018, pp. 7054–7057.

[4] Z. Lu et al., "Learning from weak and noisy labels for semantic segmentation," *IEEE TPAMI*, vol. 39, no. 3, pp. 486–500, 2017.

[5] B. Frénay and A. Kaban, "A comprehensive introduction to label noise," in *Proc. of ESANN*, 2014.

[6] D. Rolnick et al., "Deep learning is robust to massive label noise," *CoRR*, vol. abs/1705.10694, 2017.

[7] L. Muñoz-González et al., "Towards poisoning of deep learning algorithms with back-gradient optimization," in *Workshop AISec*. ACM, 2017, pp. 27–38.

[8] M. Campos-Taberner et al., "Processing of extremely high-resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS data fusion contest - Part A: 2-D contest," *IEEE JSTARS*, vol. 9, no. 12, 2016.

[9] R. Caye Daudt et al., "Fully convolutional siamese networks for change detection," in *ICIP*, October 2018, pp. 4063–4067.

[10] R. Caye Daudt et al., "High resolution semantic change detection," *arXiv preprint arXiv:1810.08452*, 2018.

[11] C. Benedek and T. Szirányi, "Change detection in optical aerial images by a multilayer conditional mixed markov model," *IEEE TGRS*, 2009.

[12] O. Ronneberger et al., "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.