POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Learning to Detect Illegal Landfills in Aerial Images with Scarce Labeling Data

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Authors: Corrado Fasana, Samuele Pasini

Advisor: Prof. Piero Fraternali

Co-advisor: Federico Milani

Academic year: 2021-2022

1. Introduction

Environmental protection is the practice related to the protection and conservation of natural resources with the aim of preserving all forms of life. Recently, the importance of these aspects guided nations to assess the environment's conditions and changes through environmental monitoring. Among the topics addressed by environmental monitoring, the control and careful treatment of waste cover a crucial role and require the consideration of Illegal Landfills (ILs), which are a serious source of hazards for the environment and society.

For this reason, detecting illegal disposal sites on time is essential. On-site inspection of potential ILs is still fundamental to assess the danger and potential impacts of outlawed activities but prevents keeping a wide territory under control. To speed up the process and reduce the inspected locations, Remote Sensing (RS) technologies that allow capturing aerial images can be exploited to check the presence or absence of ILs [7]. Furthermore, distinguishing among the different types of landfills can help prioritize the interventions but can be even more challenging. The advent of Computer Vision methods, and Deep Learning (DL), leads the way to the development of new automatic tools that can capture experts' knowledge and partially automate the process, reducing the number of on-site inspections. However, the lack of fine-grained annotated data, which is important to train deep architectures, leads to the development of methods able to solve the same tasks by reducing the need for huge quantities of annotated data, such as Weak and Self Supervision.

In this work, the ILs detection problem is approached as a multi-label classification problem, designing a model that can differentiate among the different types of landfills by exploiting Aerial Images. A multi-scale Convolutional Neural Network classifier and RGB Remote Sensing Images (RSIs) are used, extending the ILs binary classifier proposed by Torres et al. [7, 8], to a multi-label classifier. The proposed method is evaluated quantitatively and qualitatively, obtaining 56.43% average F1-score on the AerialWaste data set [8] for the multi-label classification task. The qualitative analysis, using Class Activation Maps (CAMs), highlights the strengths and weaknesses of the model in terms of localization capabilities. The tested self-supervised approaches obtain poor performances, demonstrating the need for specific pre-

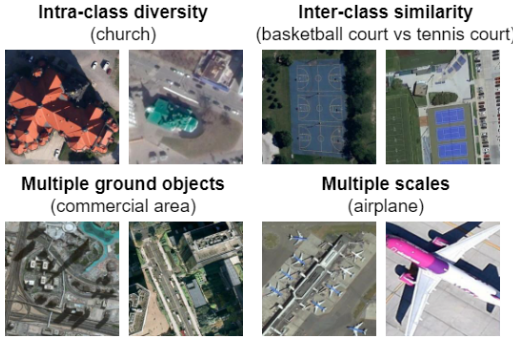

Figure 1: Examples of issues in RSIs.

text task adaptations.

2. Related Work

Some research fields require specific types of images that are more difficult to be retrieved and annotated with respect to natural images and may introduce new challenges that can affect the performance of generic DL models if not addressed. RSIs differ significantly from natural images given that object instances only occupy a small portion of large images, while in natural images, few big objects are usually present. RSIs are characterized by a large variation in objects' scale, appearance, orientation, and density, as well as cluttered backgrounds. Finally, the high intra-class diversity and inter-class similarity can strongly affect the generalization capabilities of DL models, making hard the transferability of knowledge from natural images to the RS domain. Figure 1 shows some examples of common issues in RSIs.

The increasing availability of RSIs allowed the application of DL methods to solve specific tasks, such as image classification and object detection. A large quantity of openly available RSIs can speed up tasks that are still mostly performed manually, e.g., ILs detection. The ILs detection use case presents the usual characteristics of RSIs due to the heterogeneity of the waste objects and confusing backgrounds. Over the years, different approaches have been developed to solve the ILs detection problem by exploiting aerial images. However, very few DL models have been proposed. Many ILs detection approaches are based on the availability of bounding boxes indicating the position of the landfill, but these fine-grained annotations are hard to obtain and usually require expert knowledge. Another major challenge in collecting ground truth samples is also given by the sensitivity of the domain.

Torres et al. [7] propose to solve this problem by considering ILs detection as a scene classification task that only needs image-level labels, indicating the presence or absence of an IL. This binary classification approach leads to promising results. However, the distinction of the different types of landfills can further boost the environmental monitoring process since it allows prioritizing the inspections based on the type of waste. Unfortunately, there is no previous work that addresses this task that requires multi-label classification. Torres et al. [8] propose a novel data set, AerialWaste, labeled for binary and multi-label (only for a small subset) ILs classification with annotations obtained by the Agency for Environmental Protection of Lombardy region (ARPA).

Given the data-hungry nature of DL architectures and the limited amount of available samples, other techniques can be taken into consideration to solve the task under analysis. One of the most widespread techniques is Transfer Learning (TL) which allows transferring knowledge from a source to a target task before fine-tuning it. More recently, Self-Supervised Learning (SSL) techniques have been proposed to learn more relevant features exploiting unlabeled data. The features are learned by training a so-called pretext task, i.e., a simpler problem, using supervision signals automatically generated from the data or maximizing the similarity between semantically identical inputs. The learned features are then transferred to a supervised downstream task (e.g., multi-label classification), just as in the case of TL. The effectiveness of the knowledge transfer is dictated by the difference between the two domains in the case of TL and by a well-designed pretext task in the case of SSL. Over the years, several approaches have been developed to perform SSL in natural images. For instance, the work by Noroozi et al. [6] learns relevant features by solving jigsaw puzzles, i.e. decomposing an image in tiles which are then mixed and letting the network learn how tiles were shuffled. Instead, Gidaris et al. [2] propose to learn features by predicting image rotations. If the network can solve the assigned pretext task, it should have learned specific features related to the image content.

While these methods have been proven to boost the performance of natural images, they are badly impacted by the characteristics of RSIs and need to be adapted to generalize well. For this reason, several methods have been studied on RSIs. Among them, the work by Jean et al. [4], Tile2Vec, exploits the idea that portions of an image that are near each other should be close in the feature space to learn proper features.
While SSL allows learning better feature representations, other techniques based on weak supervision allow for solving more complex tasks starting from coarse-grained labels. For instance, many approaches have been proposed to perform Weakly Supervised Instance Segmentation (WSIS) exploiting image-level labels such as IRNet [1] which generates pseudo-labels using inter-pixel relations and displacement fields. Unfortunately, no method is specifically designed for WSIS in RSIs.

3. Data Set and Methods

This work assesses the performance of multi-label fine-grained classification in the specific case of ILs detection. The aim is to build a model able to distinguish among different classes which is crucial if a subsequent localization task needs to be performed.
The AerialWaste data set [8] is used for the experiments. It consists of 10,434 RGB images (already split in train and test) of different sizes and resolutions. Binary labels are provided for each image to indicate whether an IL is present. Multi-class multi-label annotations are given for a subset of images based on the presence of specific waste objects and storage modes. Finally, a few test images are annotated with segmentation masks surrounding relevant waste objects.
The data set is highly imbalanced and considers 22 categories related to the waste type or the storage mode. Moreover, it is characterized by a high class co-occurrence, meaning that most of the time two or more classes appear together. The available classes are not easily recognizable since the images possess all the characteristics of RSIs, such as intra-class diversity and inter-class similarity. Figure 2 presents examples extracted from the AerialWaste data set, showing its characteristics.
To improve the multi-label classification performance by addressing the class imbalance and co-occurrence issues while working with the few available samples, several options can be considered. Particular attention must be taken into consideration to improve the discriminative power and localization capabilities of the network.

Figure 2: Examples of images from the AerialWaste data set [8] and their characteristics.

Since most images are around $1,000 \times 1,000$ pixels and the suspicious sites are often located in the center of the image, to increase the number of available samples, it is possible to crop each image in different patches and label them as the original image. The choice of crop size is critical since cropping images may result in cutting out landfills, thus generating mislabeled samples. Another major issue is related to class co-occurrence and imbalance. To deal with these problems, new samples can be generated and added to the original data set. This can be done via oversampling, i.e., creating multiple augmented copies of the available images. Before being fed to the network, each image is flipped and rotated so that the network will rarely see the same image twice. Alternatively, it is possible to generate new synthetic data, thanks to the availability of many negative samples in the AerialWaste data set. In this case, Synthetic Data Augmentation (SDA) can be performed by inserting patches of manually extracted ILs on images without landfills. In this way, a new image is obtained and labeled according to the type of landfill that is added. To increase the realism of the images, a simple idea is to blur the contours of the patch before placing it on the background, reducing the contrast between the two images.

Furthermore, to improve the possibility of obtaining more promising results given the complexity of the task, TL is used to initialize the weights of the network. Besides the widespread ImageNet pre-training, TL from Torres et al. [8] model (Torres-AerialWaste pre-training) is considered since, in this case, the source and target tasks are very similar, especially from the domain standpoint. Moreover, the knowledge obtained performing ILs binary classification can be important to perform fine-grained classification of illicit waste disposal sites. Indeed, the more specific fine-grained classification task requires implicit knowledge about ILs characteristics.
SSL is also tested to improve the learned features. Despite the techniques proposed by Noroozi et al. [6] (Jigsaw) and Gidaris et al. [2] (Rotation), the one proposed by Jean et al. [4] (Tile2Vec) is implemented. Tile2Vec requires the selection of three tiles, two of which should be similar to each other and both should differ from the third. The tiles selection cannot be performed randomly since landfills usually cover a small portion of the image and most of the time the selected tiles would not contain any waste disposal site. This would result in learning nothing about the different types of landfills which is instead crucial for successful classification. To solve this problem, a CAM-guided approach is proposed. The idea is to exploit the CAMs produced by an ILs binary classification model [7, 8] to discover where an illicit waste disposal site may be located and then use this information to select tiles that are highly likely to contain a landfill. Similar tiles will be extracted from areas where CAMs are highly focused while the different tiles will be selected from empty areas. Finally, once CAMs are produced (independently of the pre-training technique), it is possible to binarize them using a threshold or refine them using methods such as IRNet [1] to check the possibility of proceeding with a localization task.

4. Evaluation

The architecture used during the various experiments is the one proposed by Torres et al. [7, 8] which exploits a Residual network (ResNet50) [3] enhanced with a Feature Pyramid Network (FPN) [5] to account for the multi-scale nature of RSIs. When TL is performed, the first two stages of the network are frozen before fine-tuning.
Each experiment is evaluated from a quantitative and a qualitative point of view. Per-class and macro-average precision, recall, and F1-score are used to quantitatively assess the performance of each model, while the qualitative evaluation is performed through a visual inspection of CAMs to check the discriminative and localization capabilities of the network.
Initially, a set of preliminary experiments is performed to verify which AerialWaste classes are more distinguishable. For this reason, single-class experiments are conducted to analyze the influence of the crop size, data set configuration, and TL source task. The evaluation revealed that Torres-AerialWaste pre-training outperforms ImageNet pre-training since it allows to focus more on ILs. At the same time, a crop size of 650×650 pixels is the best choice to augment the training data set given that smaller crop sizes introduce too many mislabeled instances, while bigger crop sizes introduce too much context, causing a degradation of the performance. Furthermore, the importance of including in the data set negative samples containing other types of landfills is proven.
At the end of these experiments, five classes (*Rubble, Bulky items, Fire Wood, Scrap* and *Vehicles*) are selected. These classes correspond to the waste types with the most distinctive features that appear the most in the AerialWaste data set. Thus, a new data set is built keeping only a subset of the original data set, resulting in a training, validation, and test set containing 447, 111, and 201 samples. The training set is augmented cropping images using a crop size of 650 pixels, obtaining 1,482 samples.
The selected data set is used to perform multi-label classification using the previously described ResNet50+FPN architecture and the Torres-AerialWaste pre-training. A baseline model is trained on this data set, obtaining 57.15% average F1-score but a qualitative analysis reveals that the network is often unable to discriminate between classes. Since this behavior may be due to the characteristics of the data set, and especially to the very high class co-occurrence, other experiments are performed exploiting oversampling and SDA to mitigate this

issue.
The number of samples generated using oversampling is computed by exploiting a linear programming algorithm (Simplex algorithm) with constraints aimed at reducing the class co-occurrence and imbalance. Using oversampling, the performance slightly deteriorates (55.70% average F1-score) but an improvement in the discrimination between classes is shown, indicating that the idea of reducing the class co-occurrence is promising.
To further verify this idea, SDA is initially applied generating a large number of synthetic samples containing only a single class, which trivially reduces the class co-occurrence and imbalance. Using SDA, generating 750 samples per class with contours blurring, 56.72% average F1-score is obtained with a remarkable improvement in the discrimination between classes. Once again confirming the original hypothesis that class co-occurrence and imbalance are major issues to be tackled. Further experiments were conducted, generating multi-label synthetic data following the strategy indicated by the Simplex algorithm, reaching 61.95% average F1-score. The obtained model is able to distinguish classes better than the baseline but worse with respect to the previous approach. This can be due to the fact that the generated samples with multiple classes are more confusing, thus negatively affecting the network's discriminative power.
It is important to notice that, in this case, a quantitative improvement of the performance does not imply that the network is able to discriminate better between the five selected classes. From the previous experiments, if a model almost always predicts that three or more classes are present in an image, it can obtain a better average F1-score than a model that is less confident but has learned more accurate features. Still, this is an issue related to the high class co-occurrence.
Given the previous results, three different SSL algorithms are implemented to check the possibility of improving the feature representations by exploiting the wide amount of data annotated exclusively at binary level in the AerialWaste data set. More specifically, the prediction of image rotations, the resolution of jigsaw puzzles and CAM-guided Tile2Vec are considered as pretext tasks. After their training, the learned features are transferred to the downstream task (multi-label waste type classification) using the best data augmentation obtained so far (SDA with 750 samples and contours blurring). The three experiments resulted in 53.05%, 53.04%, and 46.06% average F1-score, revealing that using SSL in the ILs scenario is less effective than ImageNet pre-training (53.80%) and especially worse than Torres-AerialWaste pre-training (56.72%). This can be due to the fact that the ILs scenario requires the design of more ad-hoc pretext tasks to learn discriminative features of landfills.
Overall, the most promising model remains the one that exploits SDA with 750 single-class samples per category. This model and the baseline are selected and compared, for a final generalization capabilities assessment, on the test set. While the performance of the baseline drops to 44.47% average F1-score, the selected model is still able to reach 56.43% average F1-score, showing good generalization capabilities (Table 1). This reveals the importance of complementing the quantitative analysis with a qualitative one exploiting CAMs (Figure 3). While being better than the baseline in the large majority of the cases, the model still presents some limitations due to the confusion between classes and the inability to detect very large instances. However, this is coherent with the fact that most of the time small instances are present in the images and that many classes are co-occurring and similar to each other.

Table 1: Quantitative evaluation of the baseline and the selected model on the test set.

Experiment	**Metric**	**Macro avg.**
Baseline	*Precision*	51.41%
	Recall	52.73%
	F1-score	44.47%
Selected Model	*Precision*	**58.08%**
	Recall	**61.16%**
	F1-score	**56.43%**

5. Conclusions

In this work, the problem of ILs detection is addressed as a multi-label classification problem, paying particular emphasis on the discriminative

Figure 3: Examples of images of the test set and corresponding predictions performed by the selected model. In the first case, the model correctly predicts *Fire Wood* focusing on the right aspects but it also predicts *Rubble* due to inter-class similarity. In the second case, the model can correctly predict a small instance of *Fire Wood*.

and localization capabilities of the network. A multi-label data set (AerialWaste [8]), containing RSIs with ILs, is analyzed and evaluated.
A ResNet50 backbone augmented with an FPN is used to improve the feature extraction at different scales as proposed by Torres et al. [7, 8]. A subset of five classes, representing waste types, is selected from the original data set and is enlarged with new samples, to mitigate the effects of class co-occurrence and imbalance. Synthetic Data Augmentation is the method that provides the best results. To exploit the part of the data set annotated only at binary level, SSL approaches are analyzed and tested, revealing the need for more suitable pretext tasks for the scenario of ILs in RSIs.
A quantitative and qualitative evaluation of the models is performed on the validation set. The baseline and the best model are evaluated also on the test set, showing the superior generalization capabilities of the best model (56.43% F1-score) which obtains also satisfactory qualitative performance. At the moment, the obtained results prevent proceeding with a localization task. The experiments confirm the importance of tackling class co-occurrence and imbalance, as well as the need for more data. For these reasons, future work will concentrate on: 1) extending the data set, with a special focus on the previously describes issues, 2) exploiting hyperspectral data to reduce the effects of inter-class similarity and intra-class diversity, and 3) designing more suitable pretext tasks to efficiently exploit unlabeled or coarse-labeled RSIs.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2204–2213, 2019.

[2] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[4] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[6] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[7] Rocio Nahime Torres and Piero Fraternali. Learning to identify illegal landfills through scene classification in aerial images. *Remote Sensing*, 13(22):4520, 2021.

[8] Rocio Nahime Torres and Piero Fraternali. Aerialwaste: A dataset for illegal landfill discovery in aerial images, August 2022.