Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 185 (2021) 361–369

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference Theme: Big Data, IoT, and AI for a Smarter Future
Malvern, Pennsylvania, June 16-18, 2021

Deep Learning and Remote Sensing: Detection of Dumping Waste Using UAV

Ousmane Youme[1]* , Theophile Bayet[2] , Jean Marie Dembele[3], Christophe Cambier[4]

[1, 3] *Université Gaston Berger de Saint-Louis, Laboratoire d'Analyse Numérique et Informatique, BP.234 Saint Louis, Senegal*
[2,4] *SORBONNE UNIVERSITE, IRD, UCAD,UGB UMI UMMISCO, F-75006 Paris, France*

Abstract

An important success and use of Deep Learning in recent years has been in the field of image processing. Research on Deep Learning has shown that these architectures particularly convolution neuron network (CNN) can learn solutions with human-level capability for certain visual tasks. These techniques have been used in particular in remote sensing image analysis tasks, including object detection on images, image fusion, image recording, scene classification, segmentation, object-based image analysis, land use and land cover classification (LULC). In this paper we present an automatic solution for the detection of clandestine waste dumps using unmanned aerial vehicle (UAV) images in the Saint Louis area of Senegal, West Africa. This is a challenging task given the very high spatial resolution of UAV images (on the order of a few centimeters) and the extremely high level of detail, which require suitable automatic analysis methods. Our proposed method begins by 1) segmenting image into four (4) regions, which can be used as an input image 2) Reduce size of input images into 300x300x3 for the CNN entries 3) Labelling the image by determining region of interest. Next Single shot detector SSD is used to mine highly descriptive features from these datasets. The results show that the model recognizes well the areas concerned but presents difficulties on some areas lacking clear ground truths.

Keywords: Deep Learning; Remote Sensing; Convolutional Neuron Network; Single shot Detector

* Corresponding author. Tel.: +221 77 263 25 00.
E-mail address: youme.ousmane1@ugb.edu.sn

10.1016/j.procs.2021.05.037

1. Introduction

Artificial intelligence in general, Machine Learning in particular, is used in many fields such as medicine, science, environment, etc. in order to allow humans to better understand their ecosystem. Indeed, Machine Learning associated with some techniques is very useful to solve certain tasks. Among these techniques we have within the area of our study remote sensing (refers to all the methods that is used to study objects or phenomena on earth at distance). Remote sensing associated with Deep Learning is more and more used by researchers (200 papers) to solve certain problems. Deep learning algorithms have recently been introduced in the geosciences and remote sensing community for data analysis. Zhang et al [1] show that the use of deep learning has already been initiated at all levels of learning with remote sensing data: for traditional image pre-processing topics, for target classification and recognition, for recent recurrent tasks of extracting high-level semantic features and for understanding image scene. The methods applied in remote sensing, particularly in Geographic Information Systems (GIS), are similar to input-output data processing combined with various deep networks and parameters. Numerous experimental results confirm the good performance of algorithms based on Deep Learning in the analysis of remote sensing data. However, difficulties still exist. In particular, the main questions are:

- The amount of training data: there is generally a lack of high-quality training images because it is difficult to obtain labeled images. Under these circumstances, applying Deep Learning methods with a small amount of training data remains a great challenge. The complexity of remote sensing images contributes greatly to the difficulty of learning robust and discriminating representations from scenes and objects in Deep Learning.
- Model adaptation: the question is whether networks trained on a certain type of images will be effective for images acquired in a different context.
- The depth of the network: it can be expected that more the networks are deeper more the performance of the models is better. For supervised networks such as CNN, deeper layers may learn more complex distribution, but they may require many other parameters to be learned, and otherwise lead to an overlearning problem. Computation time is also an essential factor that must be taken into account in network design.

In the context of the NAOMI project, work concerns the semantic classification of vegetation maps from **hyperspectral** data using Deep Learning methods [2]. Y. Dou et al [3] propose a classification approach based on the DBN model for detailed urban mapping using **Pol-SAR** (Synthetic Aperture Radar) data. The classification is performed on polarimetric RADARSAT-2 data. Comparisons with SVM methods, conventional neural networks and probabilistic methods of the expectation-maximization type were conducted to evaluate the potential of the classification approach based on the DBN (Deep Belief Network). Experimental results show that the DBN-based method outperforms the other three approaches and produces homogeneous mapping results with preserved form details. In **LIDAR** (laser detection and ranging), work has been initiated in this direction with Deep Learning: Garcia-Gutierrez et al [4] present a preliminary comparison between the use of multiple linear regression with and without pre-processing by the auto-encoders on real LIDAR data from two Spanish areas, for biomass estimation. The results show that the auto-encoders statistically increased the quality of the multiple linear regression estimates by about 15-30%. The use of auto-encoders is made possible here because it is an unsupervised technique, which therefore does not require labelled data sets. In the field of **SAR** (synthetic aperture radar) processing, DL has been used to classify urban areas. The rotated urban target has different mechanisms and the network learns the parameters α and γ of the HH, VV and HV bands (H, horizontal; V, vertical polarization). J. Geng et al [5] used an eight-layer network with a convoluted layer to extract texture features from SAR imagery, a scale transformation layer to aggregate neighboring features, four-stack AE (SAE) layers for feature, optimization and classification, and a two-layer post-processor. The limitation to use Deep Learning methods for study with remote sensing is the acquisition of training datasets with sufficient size.

Deep Learning has been used on almost all parts of remote sensing with satisfactory results. In our case we use it in the context of an environmental problem, namely waste, which is a real and international challenge because it has a great impact on the economy, tourism, and the health of ecosystem's species. Most of this waste comes from clandestine dumping, industrial dumps, etc. It is composed of 80% plastic, but their destruction poses a problem because the degradation of the polymers and their fragmentation are transformed into microparticles. Therefore, it is essential to find techniques to detect and list the waste to facilitate its treatment. We choose a Deep Learning method CNN for the detection of clandestine waste deposits in the city of Saint Louis in Senegal using a UAV. Our objective is to train a detection model for finally set up a monitoring and planning tool that can help municipality to control the problem of clandestine waste dumps. We have to answer the questions previously asked above: data acquisition with

the UAV which presents high-resolution images, adaptation of the models taking into account the different study areas on the sand, along the river, the depth of the neural network and computation time.

Nomenclature

DL	Deep Learning
CNN	Convolution Neural Network
RS	Remote Sensing
DBN	Deep Belief Network
AE	Auto-encoder

2. Related Works

The Possibility of Monitoring of Waste Disposal Site Using Satellite Imagery has been shown by YONEZAWA et al. [6] using an operated earth observation satellite Japanese ALOS. The Japanese ALOS has two optical imagers, PRISM and AVNIR-2, and an L-band synthetic aperture radar (PALSAR) with ground resolution of 0.6 m for panchromatic imagery and 2.4 m for multispectral imagery. Pan-sharpen PRISM and AVNIR-2 images are useful for image interpretation by observation. Multispectral information greatly helped to distinguish between concrete and vegetation due to their differing spectral characteristics. It was difficult to identify scrap iron and plastic with a size of approximately 2 x 2 m lying on bare soil in either panchromatic or pan-sharpen images. Plastic, concrete, and bare soil generally exhibit high reflectance. Dabholkar et al. [7] propose to use deep learning approach GoogleNet and AlexNet to recognize various types of frequently dumped wastes. They explored various approaches and demonstrate the accuracy variance with regard to the number of classes, baseline models, and input image characteristics. In the first they built a naive approach with 102 images and 100 iteration for training; In Second approach with 407 images, more classes and 5000 iteration for training and in the third approach they cropped the images to contain only one target object(single object detection) thereby the deep learning model can more clearly distinguish different classes. Result show a variation in accuracy between classes that is unbalanced. Begur et al. [8] focuses on illegal dumping problems in the City of San Jose. It proposes an innovative smart mobile based service system that supports real-time illegal dumping detection, altering, monitoring, and management. For the detection it applies the faster R-CNN on the images taken from an edge-based sensors. Sensors (smartphone, Camera etc.) send the captured images to a server for processing. Result based on 9963 images and 24640 annotated objects shows 69.17% accuracy in illegal object detection.

3. Deep Learning methods

Deep neural networks have more complex connections between neurons than the conventional neural network, many more neurons to express the complexity of a problem, therefore require more hardware resources and machine power. Moreover, they were initially introduced to solve Big Data problems with successful applications in the field of voice recognition, automatic language processing, pattern recognition. Deep networks can be considered as an architecture divided into sub-problems and each of its sub-parts solve a simpler problem and pass the result to the next layer. Deep Learning is particularly used in image processing. In 2007, the Stanford Vision Lab set up ImageNet, a database of several million labeled images. In 2012, Deep Learning is back in the spotlight with the success of its performances at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [9]: an annual image recognition competition founded by Stanford University. The methods used in this competition using Deep Learning achieve more accuracy than any other. Since then, these methods have been deployed for many applications.
At present, we note four (4) Deep Learning algorithm architectures called: Convolutional Neural Network, Auto-encoder, Deep Belief Network (DBN) and Restricted Boltzmann Machine (RBM) [10].

- Convolutional Neural Network (CNN): Among the supervised approaches, convolutional neural networks CNN are neural networks dedicated to image processing, which have generated unequalled performance gains in the field of recognition [11]. The deep structure of the CNNs allows the model to learn increasingly abstract

layers of descriptors, resulting in representations that can clearly boost the performance of subsequent classifiers. The CNNs can be considered as a multilayer stack of perceptron, whose purpose to preprocess small amounts of information. They include convolution layers.

- Auto-encoding: An auto-encoder or AE is a symmetric neural network used to learn a representation (encoding) of a data set, usually with the aim of reducing the size of the data set [12]. The network is built by minimizing the reconstruction error between the input data on the encoding layer and its reconstruction at the output of the decoding layer.
- DBN and RBM: The Deep Belief Networks (DBN) model [13] is a widely studied and deployed deep learning architecture, one of the first historically developed. It combines unsupervised and supervised learning. It is a deep neural network, composed of multiple layers of latent variables ("hidden units"), with connections between layers but not between units within each layer. The neural layers are trained one by one in an unsupervised phase by Restricted Boltzmann Machines (RBMs). These RBMs are composed of a neuron layer, which receives the input, and a hidden neuron layer, with neurons of the same layer independent from each other. DBNs link supervised and unsupervised phases.

4. Methods and Data acquisition

4.1. Study area

Our place of study is located in the region of Saint Louis in the North of Senegal, West Africa. It is an area of the country that attracts many tourists because of its popular beaches, its warm sun, its atmosphere, its culture and its history. It is an ideally ecological zone with the Senegal River which crosses it from end to end, “the Tongue of Barbary”, "the island of birds" which sees thousands of migratory birds every year during the nesting season. However, some areas have waste deposits that can obstruct the view but above all cause environmental problems.
For our study, the areas we have selected are the Gaston Berger University of Saint Louis and along the Saint Louis River.

Fig. 1. Map of Saint-Louis, Senegal. Point of location the area of our study along the river

We flew the drone over an altitude of 5, 10 and 30 meters at a distance of 300 meters with a 90° angle of view. Based on the images taken, we can see that the 10-meter altitude satisfies more in terms of coverage, image resolution and visibility. We will take the images record at 10 meters and also at 30 meters to train our model. Our study area has special feature: wet at the edge of the shore with a darker background, sandy with sand dunes and forest with trees. This makes up an area specificity to train our model to be usable in various areas See Fig. 2.

4.2. Data acquisition

For the acquisition of the necessary data we use the DJI Mavic PRO Phantom UAV which has a high resolution with its L1D-20c RGB color camera taking images of resolution 5472 X 3648. The high resolution is essential for the performance of our model. We record a video while flying over our study area. The video is transformed into several screen images per frame of (2) two seconds. We obtain a database of images which are divided into a grid of (4) four images and then reduced to a size adapted for the inputs of the chosen algorithm: 300X300. The waste deposits being on a large surface this division will allow us to have the maximum of covered surface while preserving the quality of the image. We obtain our image shape to be labeled constituting an input to the algorithm for the model training. In addition, the UAV is easy to manipulate and contains metadata that can be read with exiftool software, which contains information about the images, such as shutter speed, angle, ISO and GPS coordinates (latitude, longitude and altitude).

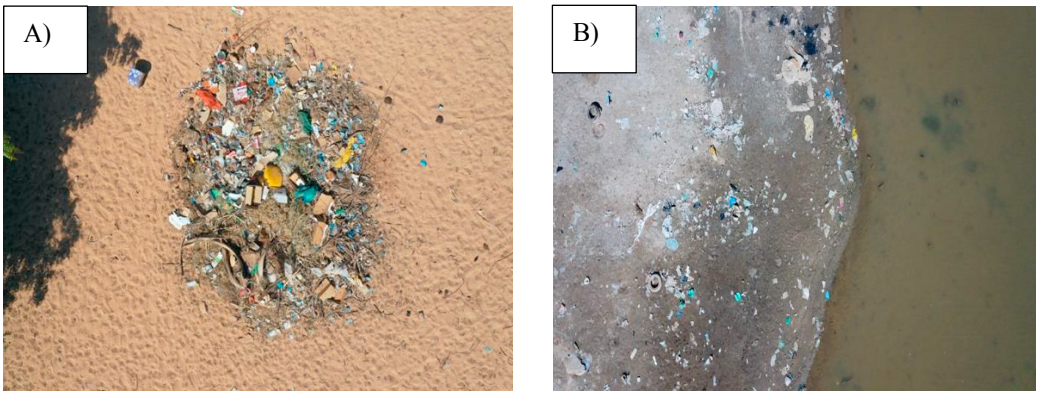

Fig. 2. Image Data acquisition. A): Dumping waste at the university with a sand background; near others under the trees. B): Waste along the river with a darker background

4.3. Algorithm

After having obtained the images reduced to the input size of the algorithm, we labelled the images by framing the parts containing waste. We choose the Single Shot Detector (SSD) as the detection algorithm. SSD [14] is a method for detecting objects in images using a single deep neural network. The SSD approach consists in discrediting the output space of selection boxes into a set of default boxes on different formats and scales with labels represented by the location of the object on the image. It is composed of Multi-scale feature maps by adding convolution layers of decreasing size to reduce dimensioning and predict multi-scale detections, then convolution filters called: Convolutional predictors for detection are imposed on each convolution layer. For a feature layer of size $m \times n$ with p channels, the basic element for predicting parameters of a potential detection is a $3 \times 3 \times p$ small kernel that produces either a score for a category, or a shape offset relative to the default box coordinates. At each of the $m \times n$ locations where the kernel is applied, it produces an output value. The bounding box offset output values are measured relative to a default box position relative to each feature map location (cf the architecture of YOLO [15] that uses an intermediate fully connected layer instead of a convolutional filter for this step). Finally, bounding box are default associated with each feature map cell. Each multiple feature map has offsets from the default box shapes in the cell and class scores that indicate the presence of a class instance in each of these boxes. This results in a total of $(c+4)k$ filters(c number of class object, 4 offset, k location) that are applied around each feature map location, giving $(c+4)kmn$ output for a $m \times n$ feature map.

Experimental results on the PASCAL VOC, COCO, and ILSVRC [14] datasets confirm that SSD has competitive accuracy to methods that utilize an additional object proposal step and is much faster, while providing a unified framework for both training and inference. For 300×300 input, SSD achieves 74.3 % mAP on VOC2007 test at 59 FPS on a Nvidia Titan X and for 512×512 input, SSD achieves 76.9 % mAP, outperforming a comparable state of the art Faster R-CNN [16] model. Compared to other single stage methods, SSD has much better accuracy even with a smaller input image size.

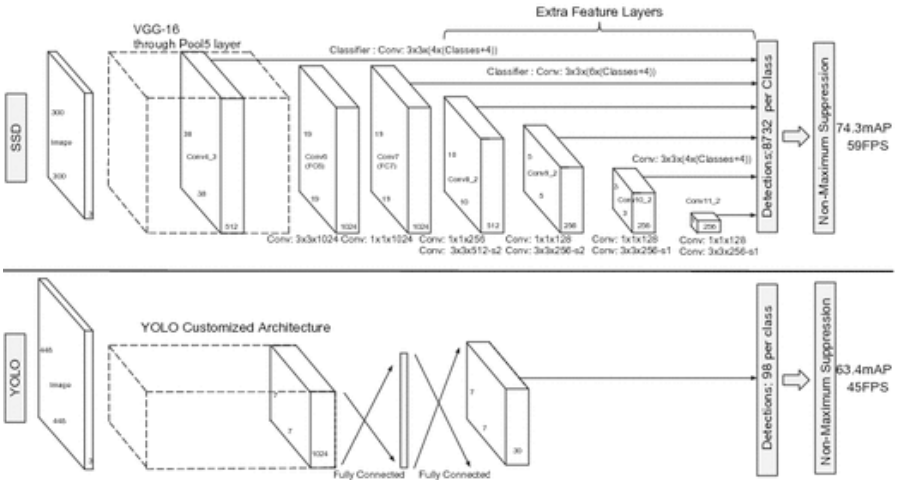

Fig. 3. A comparison between two single shot detection models: SSD and YOLO [12].

5. Experimental Results

We set hyperparameters according to the work done on SSD with default variables. These variables are chosen according to their positive effect on the data set after several tests.[14] We collected with the UAV 5000 images with an average of 5 items per image composing our database which we divided as follows: 90% for the training set and test set (data set) and 10% for the validation. Now we slice the dataset: 60% for the training set and 40% for the test set. When the dataset is run through completely once, it is called an epoch. We chose 10 the number of epochs and 32 batch size: number of items to be analyzed directly in one block in the GPU. With a GPU Intel Iris 1536 Mo, we note 730 minutes for training.

Training: The major difference between SSD and a detector with region proposal is the ground truth information is associated to specific outputs in the fixed set of detector outputs. During training, the bounding boxes generated according to their location, aspect ratio and scale are associated to the corresponding ground truth. The bounding boxes with a Jaccard overlap (Multibox [17] method) whose correspondence is greater than **threshold 0.5** are chosen. After this assignment the loss function and the backpropagation are applied from start to finish during training. The loss function (1) is calculated as Weight Sum of the localization loss e.g. Smooth L1 [16]) and confidence loss e.g. SoftMax (confidence: score for the presence of each object category c in each default box).

$$L(x, c, l, g) = {}^{1}/_{N} \left(L_{conf}(x, c) + \alpha L_{Loc}(x, l, g)\right) \quad (1)$$

N the number of matched default boxes; x_{ij}^{p} the indicate for matching the i-th default box to the j-th ground truth box of category p. C class confidences; l predicts box and g ground truth. We choice a training by cross validation $\alpha = 1$ and load the VGG16 weight which is pre-trained on the ILSVRC CLS-LOC dataset [18].

Here we have the curves of the loss function described above train loss and test loss (val loss). More N is small more the loss function value is larger. As we train the value decreases for both train loss and test loss, which shows that our model is on the good way. Our final weights will then be the weights of the epoch that had the lowest val loss.

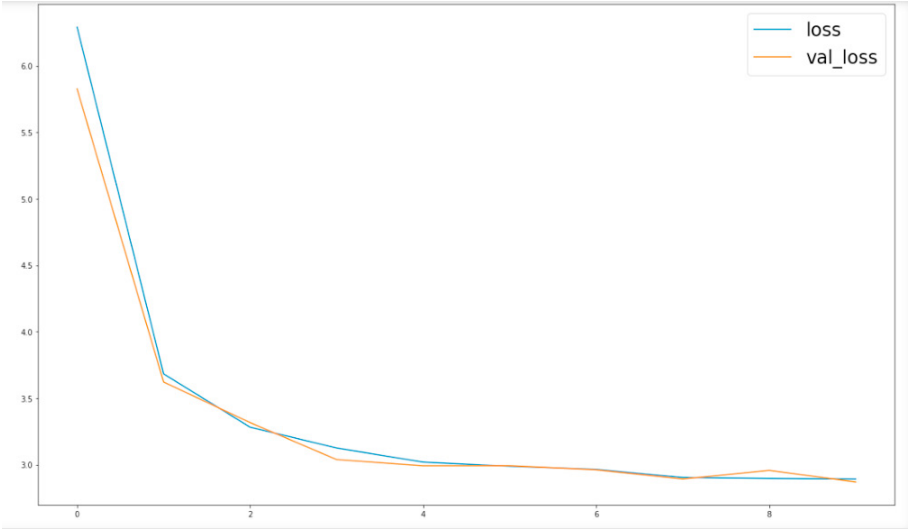

Fig. 4 The evolution of training and test loss (axes represent: the **value (y)** of function loss described **during (x)** training)

Inference: In this step we will pass the images of the validation set to the model that it can predict the areas with waste by framing them with a red rectangle and label “dechets” e.g. Waste. We give an input image for prediction in the expected format: the input images receive the same pre-processing as for training: grid division and size reduction. We set as **IoU Thresold 0.5** (Intersection Over Union, a metric to measure the intersection between the ground truth bounding box and the predict bounding boxes from our model). The confidence threshold is determined after testing several values on the validation set. We find for images with less ground truth (Visibility of object on the ground) a **confidence threshold of 0.45** and for images with more visible wastes a confidence threshold of **0.64** for the detection. We did not calculate the mean Average Precision (mAP) precision of the model knowing that the prediction results vary according to the zone. Its results can be explained by the diversity of land where the model is more efficient at the level of zones with more ground truth. In order to have better accuracy, the model must be improved by collecting more data.

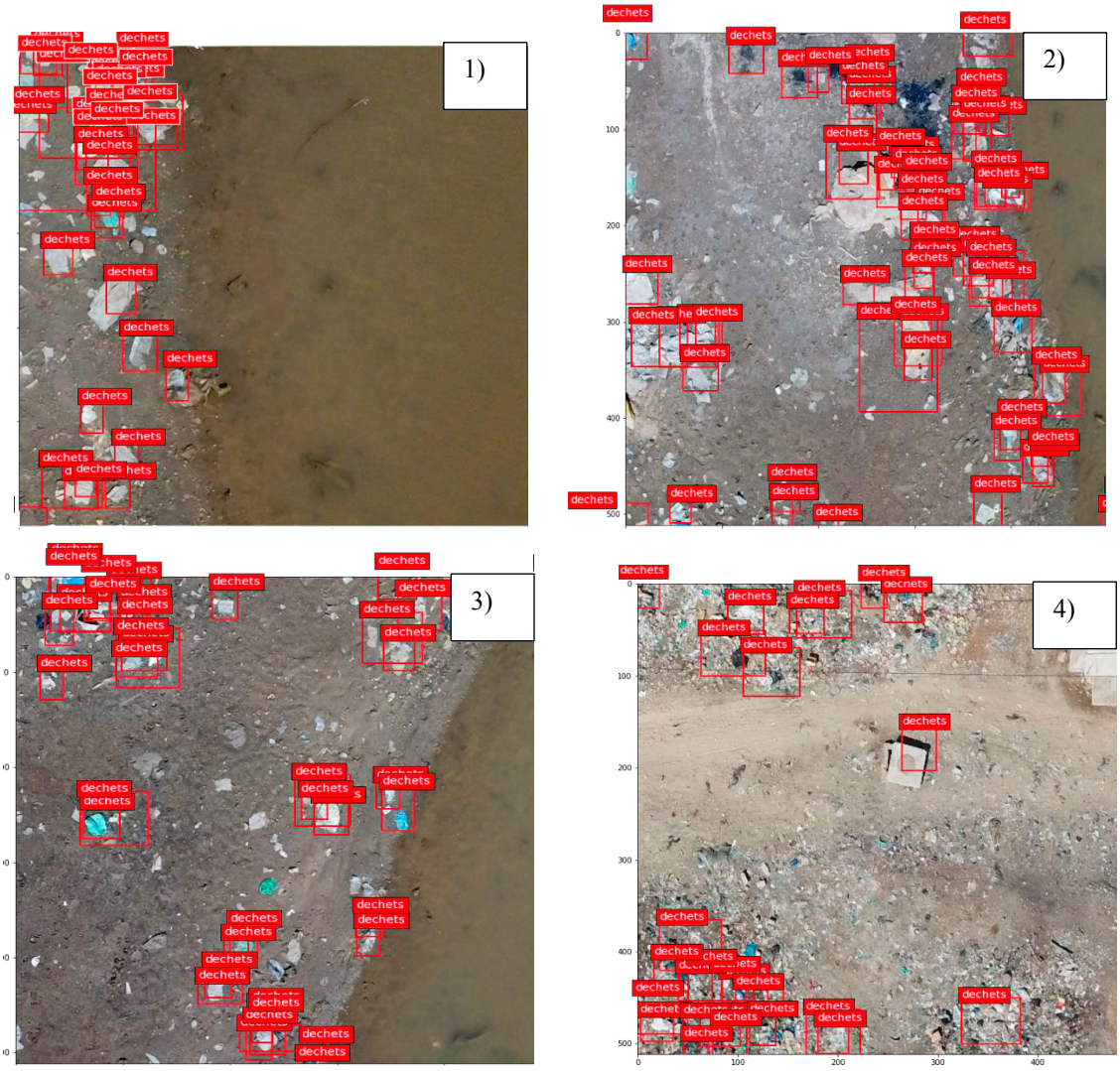

Fig. 5. Some result of prediction. 1, 2 and 3: examples predictions of images at river area who present more precision. 4: an example prediction at university area.

6. Discussion

In this work, we have built a model able of detecting dumping waste using remote sensing tools and Deep Learning (CNN). We obtained satisfactory results. However, we have tested the model on different area and we find that our model has detection problems when the image does not show ground-truth. For example, on the second site at the university level where there are many trees, the model presents inferences with many false positives. The model detects waste more over a large area and less over areas covered (by trees for example). Existing waste detection models usually detect only partially: plastic, bottles, which does not fit our problem because of the complexity of our ecosystem. Satellite observation satisfy large scale but does not give a lot of leeway, especially in certain circumstances it is difficult to detect the waste and even impossible to specify the type [6]. Mobile detection [8] can give us a better accuracy but limits us to a small scale. It is difficult to go around sites with dumping waste one by one for taking images or put camera there. Contrary to the UAV drone which satisfies both large scale and good resolution to the order of a few centimeters which will allow us to specify the type of waste. A general solution is to reconstruct the detection model with UAV for the no-covered areas and another detection model e.g. on smartphone for the

covered areas and group all results together after processing on a platform. To have a much more global method, we can switch to reinforcement learning, unsupervised phase where the agent launches an alert each time it detects waste at the level of the overflown area that it will classify as an anomaly.

7. Conclusion

Waste management is an important task for local governments in Senegal, which administer waste disposal sites and landfills to ensure that they operate appropriately, and also monitors illegal dumping.

In this work we built a model able of detecting waste deposits. The construction of the model was also possible by using several useful tools including remote sensing and Deep Learning technology good for object detection: CNN. Indeed, CNN is a good architecture in terms of image processing and object detection due to its abstraction and sophistication we recommend it for these effects.

From a general point of view, Deep Learning and remote sensing can be used to solve many problems related to the environment, floods, real estate etc. This use is all the more feasible with the possibility of acquiring data collected by oneself and which will be perfectly adapted to local environment or area. This has always been a problem for researchers who collected data from internet databases that were sometimes unfavorable to their area of treatment. In the end, these techniques make it possible to build sophisticated monitoring and planning tools customized to our needs. Other perspectives can be opened up for example instead of solving problems individually build a super model able to independently treat each given problem.

References

[1] Zhang L., and B. Du. (2016) "Deep learning for remote sensing data: A technical tutorial on the state of the art." *IEEE Geoscience and Remote Sensing Magazin*e, vol. 4, no 2, 22–40. 76

[2] Audebert, N., B. Le Saux, S. Lefevere, C. Taillandier, and D. Dubucq. (2017) "Deep learning on hyperspectral data for land use and vegetation."

[3 Lv, Q., Dou, Y., Niu, X., Xu, J., and Li, B. (2014) "Classification of land cover based on deep belief networks using polarimetric radarsat-2 data." *Geoscience and Remote Sensing Symposium* (IGARSS), IEEE International, IEEE, 4679–4682. 77

[4] García-Gutiérrez, J., González-Ferreiro, E., Mateos-García, D., and Riquelme-Santos, J. C. (2016) "A preliminary study of the suitability of deep learning to improve LiDAR-derived biomass estimation." *International Conference on Hybrid Artificial Intelligence Systems*, Springer, 588–596. 78

[5] Geng, Jie, Jianchao Fan, Hongyu Wang, Xiaorui Ma, Baoming Li, and Fuliang Chen. (2015) "High-resolution SAR image classification via deep convolutional autoencoders." *IEEE Trans. Geosci. Remote Sens. Lett*., **12(11)**, 2351–2355.

[6] Chinatsu, Yonezawa. (2009) "Possibility of monitoring of waste disposal site using satellite imagery." *JIFS*, vol. 6, 23-28.

[7] Dabholkar, Akshay, Bhushan Muthiyan, Shilpa Srinivasan, Swetha Ravi, Hyeran Jeon, and Jerry Gao. (2017) "Smart illegal dumping detection." *2017 IEEE Third International Conference on Big Data Computing Service and Applications* (BigDataService), IEEE, 255-260.

[8] Begur, Hema, Mithila Dhawade, Navit Gaur, Pulkit Dureja, Jerry Gao, Medhat Mahmoud, Jesse Huang, Sean Chen, and Xiaoming Ding. (2017) "An edge-based smart mobile service system for illegal dumping detection and monitoring in San Jose." *2017 IEEE SmartWorld*, (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), IEEE, 1-6.

[9] 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." IJCV.

[10] W.Liu, Z.Wang, X. Liu, N. Zeng, Y. Liu, and F.E. Alsaadi. (2017) "A survey of deep neural network architecture and their application." *Neurocomputing,* 234, 11-26.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. (2012) "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*, 25,1097–1105.

[12] G. E. Hinton and R. R. Salakhutdinov. (2006) "Reducing the dimensionality of data with neural networks." *Science*, vol. 313, no. 5786, 504–507.

[13] G. E. Hinton, S. Osindero, and Y. W. Teh. (2006) "A fast learning algorithm for deep belief nets." *Neural computation*, vol. 18, no. 7, 1527–1554.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. (2016) "SSD: Single Shot MultiBox Detector." *ECCV 2016: Computer Vision* – ECCV, 21-37.

[15] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016) "You only look once: unified, real-time object detection." CVPR.

[16] Girshick, R. (2015) "Fast R-CNN." ICCV.

[17] Erhan, D., Szegedy, C., Toshev, A., and Anguelov, D. (2014) "Scalable object detection using deep neural networks." CVPR.

[18] Simonyan, K., and Zisserman, A. (2015) "Very deep convolutional networks for large-scale image recognition." NIPS.