Automatic Pixel-Level Land-use Prediction Using Deep Convolutional Neural Networks

Mercedes Torres Torres[*1], Bertrand Perrat*[2], Mark Iliffe*[2], James Goulding*[2] and Michel Valstar[*1]

[1]School of Computer Science, University of Nottingham
[2]N-LAB, University of Nottingham
[3]Horizon Digital Economy Research, University of Nottingham

January 12, 2017

Summary

Land-use classification provides information on the types of human activity in an area. Its accurate mapping facilitates the study of the environmental impact that different planning decisions may have to the community. This is especially important in rapidly-developing cities with under-resourced local government, where centrally-organised mapping is difficult to maintain up to date. We show how deep convolutional neural networks can be combined with unnamed-aerial-vehicle (UAV) imagery to provide land-use pixel-level predictions. With data from Dar es Salaam we obtain a Jaccard Index of 78%, and show how in many cases predictions find clear annotation errors, outperforming human annotators.

KEYWORDS: Geo-demographics, Land-Use Classification, Big data, Deep Learning

1. Introduction

Land-use classification is a crucial environmental process that provides information on the types of human activity involved in an area. Its accurate assessment facilitates the study of the environmental impact that different planning decisions and activities may have to the community. However, while there is a large number of automatic approaches that have been developed using different types of imagery (i.e. satellite, hyper-spectral, aerial) and methods (i.e. NN, SVMs), the most accurate way of classifying land-use remains manual. This provides a particular challenge in developing economies, characterized as they are by under-resourced local government and rapid urbanization rendering central mapping endeavours rapidly out of date.

In this paper, we present a multi-disciplinary solution that applies state-of-the-art Deep Learning techniques to crowd-sourced labels to effectively attain pixel-level land-use predictions. We combine UAV imagery with *Fully Convolutional Neural Networks* to create, to our knowledge, the first pixel-level land-use system of its kind. We have compiled a dataset with UAV imagery of over 600 locations in Dar-es-Salaam (Tanzania) and we show that on a 9-class land-use problem, we obtain accuracies of 78%. More importantly, our results clearly show how in many cases the learned land-use segmentation actually outperforms human accuracy, by finding clear annotation errors.

2. Literature Review

There are two main categories of research for image based land use-classification t: manual and automatic. The most accurate way of mapping land-use classes in an area remains trained humans (Anderson, 1976). Manual annotators provide the gold standard against which all automatic approaches are compared. However, employing and training surveyors to produce accurate maps is expensive, labour intensive and time consuming. However, In under-funded and rapidly-changing areas, such as Sub-Saharan Africa, the prospect of employing human annotators is not manageable, since by the time mapping efforts have finished, the area has already changed dramatically.

* *firstname.lastname*@nottingham.ac.uk

Consequently a vast number of approaches that have been developed for automatic land-use classification using a variety of imagery and classifiers (e.g. Li et al, 2014). In recent years, deep learning has become the state-of-the-art technique in terms of image classification and segmentation (LeCun et al, 2015) and basic approaches have been applied to land-use classification (Castelluccio et al, 2015; Hu et al. 2015). However, these approaches rely in a per-image classification framework, which incorrectly assumes that all pixels within an image belong to the same class. On the other hand, state-of-the-art per-pixel classification approaches are few and in most cases (Längkvist et al. 2016; Volpi and Tuia, 2017) work with very limited number of classes (only four and/or five classes – building, ground, road, water and vegetation) and need a large amount of data (20,000 pixels of DSM plus multispectral orthophotoraphy imagery) and multiple CNNs to obtain its result.

3. Data

To perform this reseach we collected UAV imagery and demographic and environmental annotations, including land-use, from 641 regions in Dar Es Salaam (Tanzania). Annotations were collected as part of the *Dar Ramani Huria* project (Eichleay et al, 2015), with local community members creatng highly-accurate maps of their city and over 750k manual annotations. An example of the images and annotations present in our dataset is shown in Figure 1 below.

(a) example raw drone image

(b) 2 Land Use classes *(Purple: Buildings, Blue: No buildings)*

(c) 6 Land Use classes *(Purple: Building, Orange: Vegetation, Blue: Barren, Red: Roads)*

(d) 9 Land Use classes *(Orange: Vegetation, Blue: Barren Purple: Formal, Green: informal)*

(e) 10 Land Use classes *(Orange: vegetation, Blue: barren Purple: Formal, Green: informal: Red: Slum)*

Figure 1. Example of a drone image with different levels of results.

4. Methodology

Our system makes use of Fully Convolutional Networks (FCNs), developed by Long et al. (2015). FCNs were chosen because they learn dense per-pixels predictions, approaching the problem as a semantic segmentation task. Furthermore, they have produced state-of-the-art in some of the most popular computer vision problems, such as scene parsing in PASCAL VOC (Long et al., 2015). FCNs are similar to traditional Convolutional Neural Networks (CNNs), but diverge at the prediction level. As deep learning approaches, both CNNs and FCNs have similar architectures and can be divided into two stages: 1. automatic feature extraction (carried out by mostly convolutional layers) and 2. prediction generation, as shown in Figure 2. As such, both use the same set of building blocks, namely convolutional and subsampling layers. Convolutional layers, the core of deep neural networks, are used to compute the output of each of the neurons, while subsampling layers are used to progressively reduce the size of the features and reduce the amount of computation required by the network. It is also used to control and prevent overfitting.

However, while CNNs have one or more fully-connected layers to generate single predictions, FCNs substitute these by 1x1 convolutions accross the results from the previous layers. This change allows the network to output a heatmap, instead of single prediction. As a result, FCNs are able to 1) take input of arbitrary sizes and 2) produce per-pixel semantic segmentation, instead of per-image classification results. In other words, the FCNs classified and localised predictions.

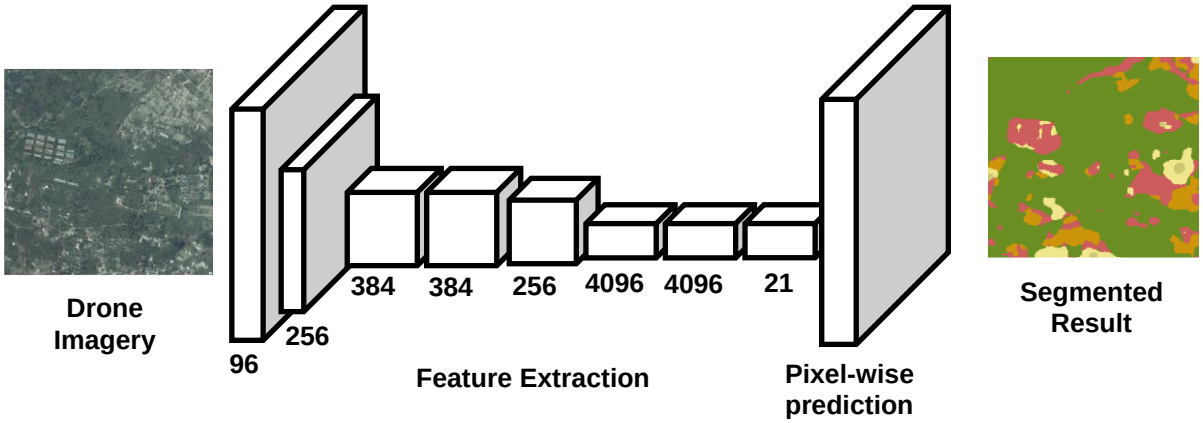


Figure 2. Fully Convolutional Neural Networks from Long et al. (2015)

5. Experimentation

In order to assess our approach we use the Jaccard Index (JI) to quantify the similarity between our predictions (*P*) and the ground-truth (*GT*), as shown in Equation (1). This index, widely used in segmentation tasks (Real and Vargas, 1996), measures the intersection over the union of two different sets. The JI ranges between 0 and 1, with 0 implying that both sets are completely different and 1 entails that both sets are equal:

$$\boldsymbol{J(GT, P)} = \frac{\left|\boldsymbol{GT} \cap \boldsymbol{P}\right|}{\left|\boldsymbol{GT} \cup \boldsymbol{P}\right|} \qquad (1)$$

5.1. Set up

For the purposes of this project, we extracted 5000 multi-class UAV images belonging to random sections of the city. Images were 440x440 pixels. Training was done with 80% of the data, randomly selected and testing was done with the remaining 20% of the data. Experiments were carried out on a machine using an NVIDIA's Titan X GPU. For each step, we ran each stage of the FCNs for 30,000 iterations (348 minutes) with a learning rate of 0.0001 and a step of 0.9.

In order to study the robustness of FCNs, we chose to test FCNs with four classifications schemes with different degrees of granularity: from a coarse 2-class scheme that only separated building from not buildings to a finer-grained 10-class scheme that included 5 different types of buildings (*urban, industrial, formal, informal,* and *slum).* Table A shows the four different classification schemes used in our system as well as the percentages of pixels for each class in the both the training and the test set.

Table A. *Different land-use classifications from fine-grained to coarse-grained. Pixel distribution for training and testing are shown in parenthesis.*

2-CLASS	6-CLASS	9-CLASS	10-CLASS
Building (34.74/35.10)	**Building** (34.74/35.1)	**Urban** (0.15/0.15)	**Urban** (0.15/0.15)
Non-Building (65.26/64.9)	**Road** (2.07/2.16)	**Industrial** (3.01/2.91)	**Industrial** (3.01/2.91)
	Vegetation (53.33/53.73)	**Formal Res.** (7.84/7.51)	**Formal Res.** (7.84/7.51)
	Barren (8.4/7.4)	**Informal Res** (23.75/24.53)	**Informal Res.** (18.87/19.39)
	Water (1.01/1.05)	**Road** (2.07/2.16)	**Slum** (4.88/5.14)
	Other (0.45/0.55)	**Vegetation** (53.33/53.73)	**Road** (2.07/2.16)
		Barren (8.4/7.4)	**Vegetation** (53.33/53.73)
		Water (1.01/1.05)	**Barren** (8.4/7.4)
		Other (0.45/0.55)	**Water** (1.01/1.05)
			Other (0.45/0.55)

5.2. Results

Table B summarises the results from our experiments all classification schemes, while Table C compares our results to the current state-of-the-art in terms of pixel-wise land-use classification (Volpi and Tuia, 2017).

Table B. *Mean and Median JI for our Dar es Salaam dataset with different classification schemes.*

NO. CLASSES	MEAN JI	MEDIAN JI
2 classes	0.55	0.77
6 classes	0.91	0.9
9 classes	0.89	0.89
10 classes	0.87	0.87

Table C. Comparison between our approach and current methods

APPROACH	RESULTS	NO. CLASSSES	DATASET
CNN (Volpi et al. 2017)	87.83	6	Vaihingen
	89.86	6	Postdam
FCNs	91	6	Dar es Salaam

5.3. Discussion

Our extensive testing showed that FCNs were both accurate and robust. With 5000 images, FCNs achieved highly accurate results, which obtaining a JI of over 87% for almost all cases and reaching

even 91% in the case of 6 classes. Additionally, and most remarkably, FCNs showed impressive robustness by being able to correctly classify images even when the annotations were incorrect. Firstly, they were are able to learn classes that were not present in the ground-truth. This is observable in Figures 3 and 4, were our system finds buildings and roads that were not annotated by humans. Secondly, this robustness also extended to not learning incorrect annotations. An example of this is shown in Figure X, were an incorrectly-annotated house is not learned by the network.

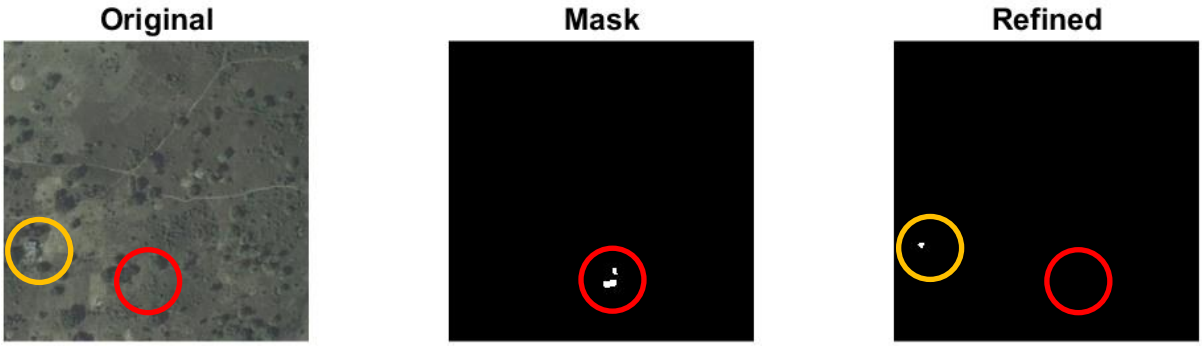
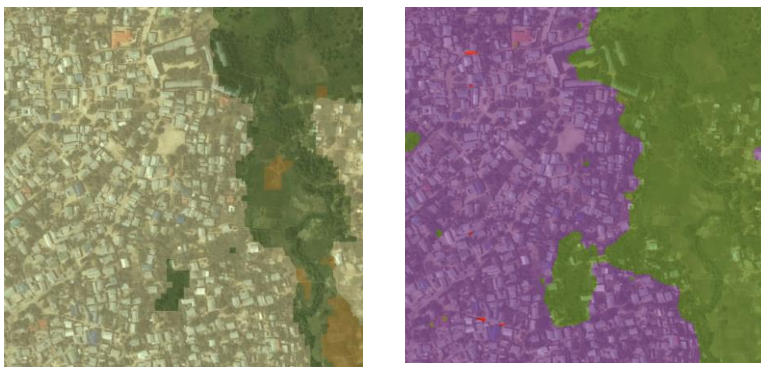


Figure 3. Here we show a 2-class segmentation, with the FCN in this case performing better than the annotators themselves. Annotations missed (yellow) are classified by the network and, at the same time, wrong annotations (red) are not.

(a) Ground Truth **(b)** Prediction

Figure 4. 6-class segmentation. FCNs are able to learn roads (red) that are not present in the ground-truth annotations, as well as finding patches of vegetation than were not annotated.



(a) Ground Truth **(b)** Prediction

Figure 5. 10-class segmentation. FCNs are able to learn more detailed patches of vegetation than were actually annotated.

6. Conclusions

Land-use classification is an essential environmental activity that helps assess the effect of planning decisions on an area. Manual classification remains the gold standard, but it is expensive, labour intensive and time consuming. On the other hand, accurate automatic land-use classification remains a challenge. These problems are exacerbated in areas where local government is under-funded. In this paper, we present the first automatic system for per-pixel land-use classification combining UAV imagery and deep learning. Results on our Dar es Salaam dataset show that our system improves accurate state-of-the-art results while, at the same time, producing pixel-wise predictions. Additionally, it is also shown that in many instances, FCNs outperform manual annotators.

7. Biography

Mercedes Torres Torres is a Transitional Assistant Professor School of Computer Science in the University of Nottingham. In 2015, she finished her PhD on *Automatic Phase 1 Classification Using Ground-taken Imagery* at the University of Nottingham. Her research interests include Image Processing and Machine Learning.

Bertrand Perrat is a Research Associate at the University of Nottingham, Horizon Digital Economy Research. He has an Engineering degree in Geospatial Sciences and is interested in application of novel Machine Learning and Prediction techniques to geospatial data.

James Goulding is Assistant Professor and Deputy Director of N-LAB, a new centre for International Analytics at Nottingham University Business School, specializing in novel forms of data science - time series analysis, summarization and visualization of mass data sets. His work focuses on how closed source, commercial dataset can be harnessed to promote of international development and social good.

Michel Valstar is an Associate Professor at the University of Nottingham, School of Computer Science, and a researcher in Automatic Visual Understanding of Human Behaviour, which encompasses Machine Learning, Computer Vision, and a good idea of how people behave. He is a member of the Computer Vision Lab, and Mixed Reality Lab.

References

Anderson, J.R., 1976. *A land use and land cover classification system for use with remote sensor data* (Vol. 964). US Government Printing Office.

Castelluccio, M., Poggi, G., Sansone, C. and Verdoliva, L., 2015. Land use classification in remote sensing images by convolutional neural networks. *arXiv preprint arXiv:1508.00092*.

Hu, F., Xia, G.S., Hu, J. and Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, *7*(11), pp.14680-14707.

Längkvist, M., Kiselev, A., Alirezaie, M. and Loutfi, A., 2016. Classification and Segmentation of Satellite Orthoimagery Using Convolutional Neural Networks. *Remote Sensing*, *8*(4), p.329.

Long, J., Shelhamer, E. and Darrell, T., (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431-3440).

Volpi, M. and Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(2), pp.881-893.