Automating the Classification of the Swiss Areal Statistics: Addressing the problem of Imbalance between the Classes

In Switzerland, the Federal Statistical Office (OFS) is in charge of the production of the land use and land cover (LCLU) statistics. The traditional method employed involves visual classification based on aerial images, which is particularly time consuming and costly. In the future, the OFS is heading towards integrating automatic classification methods into the process. To this end, the ADELE project showed that deep learning produces encouraging results for the partial automatization of the classification task. In this paper we study the particular problem of imbalanced label distribution, a common issue affecting the classification accuracy of categories with few samples. The results indicate that addressing the class imbalance problem leads to an increased accuracy on our test dataset for rare classes without damaging the accuracy of the frequent classes.

En Suisse, l'Office fédéral de la statistique (OFS) est chargée de la production de la statistique suisse de superficie. La classification de la couverture et de l'utilisation du sol par interprétation visuelle d'images aériennes est onéreuse et chronophage. A l'avenir, l'OFS se dirige vers des méthodes de classification automatique. Le projet ADELE a démontré que l'intelligence artificielle peut produire des résultats encourageants pour l'automatisation partielle de cette tâche. Ce projet aborde le problème de la distribution déséquilibrée des classes, un problème qui nuit à la classification des catégories avec peu d'exemples. Les résultats indiquent que l'utilisation de méthodes luttant contre le problème de déséquilibre des classes conduit à une amélioration de la prédiction des classes rares, pour nos données de test, sans nuire à celle des classes fréquentes.

In der Schweiz ist das Bundesamt für Statistik (BFS) für die Erstellung der Statistik der Bodennutzung und Bodenbedeckung zuständig. Die traditionelle Methode der Fotointerpretation ist besonders zeit- und kostenaufwendig. Für die Zukunft strebt das BFS deshalb den Einsatz von automatischen Klassifizierungsmethoden an. Das ADELE-Projekt hat gezeigt, dass Deep Learning erfolgversprechende Ergebnisse für die Teilautomatisierung der Klassifikationsaufgabe liefert. Das Projekt befasst sich mit dem Problem der unausgewogenen Verteilung von Landnutzungskategorien, einem häufigen Problem, das die Genauigkeit der Klassifizierung mit wenigen Proben behindert. Die Ergebnisse zeigen, dass eine adäquate Kompensation ungleich verteilter Klassen in unseren Testdaten zu einer erhöhten Genauigkeit für seltene Klassen führt, ohne die Genauigkeit der häufigen Klassen zu beeinträchtigen.

V. Zermatten, B. Kellenberger, D. Tuia

Introduction

The production of the Swiss land use statistics is led by the Federal Statistical Office (FSO) since 1980 and requires the classification of more than 4 million aerial images after each survey [1]. The classification task requires a high degree of reproducibility and reliability to ensure the compatibility of the labels from previous surveys. Up to now, this semantic labelling work has been performed manually through visual photointerpretation by a group of FSO experts. Nine to twelve years of work were needed in the previous surveys to update the statistics over the entire country. To respond to the rapid evolution of the landscape, the FSO aims at reducing the updating period to 6 years, a requirement that collides with the time needed to perform manual surveys.

Artificial intelligence offers promising solutions for the automatisation of this task. Deep learning[2], and in particular Convolutional Neural Networks (CNNs), allows images to be classified with excellent accuracy [3]. The FSO therefore launched the ADELE project with the support of the University of Neuchâtel, the FHNW (University of Applied Sciences and Arts Northwestern Switzerland) and the Zurich company Exolabs. ADELE [4] aims at developing semi-automatic classification methods and reduce the human workload for the production of land use and land cover (LCLU) statistics. Its algorithm manages to classify the images of the territory with a precision over 85% for certain land cover categories such as arable land, vineyard areas or water bodies. The determination of land use in residential and infrastructure areas is more complex and human intervention is still needed.

In this work we aim to go further in the task of automatic image classification for land use. We tackle the problem of imbalanced label category distribution, a common issue affecting the classification accuracy of categories with a relatively small number of samples. The Swiss land use statistics dataset potentially presents this issue since a low number of classes has a significantly higher number of samples than the rest (figure 1).

By default, deep learning models give equal importance to each sample regardless of their categories. As a result, classes with numerous images are better recognised since the model observes them more often. On the contrary, the model has difficulties to recognise a class with a small number of samples. This well-studied issue is known as the "class imbalanced problem". In this work, we aim at improving the recognition of rare classes by using specific methods addressing this issue.

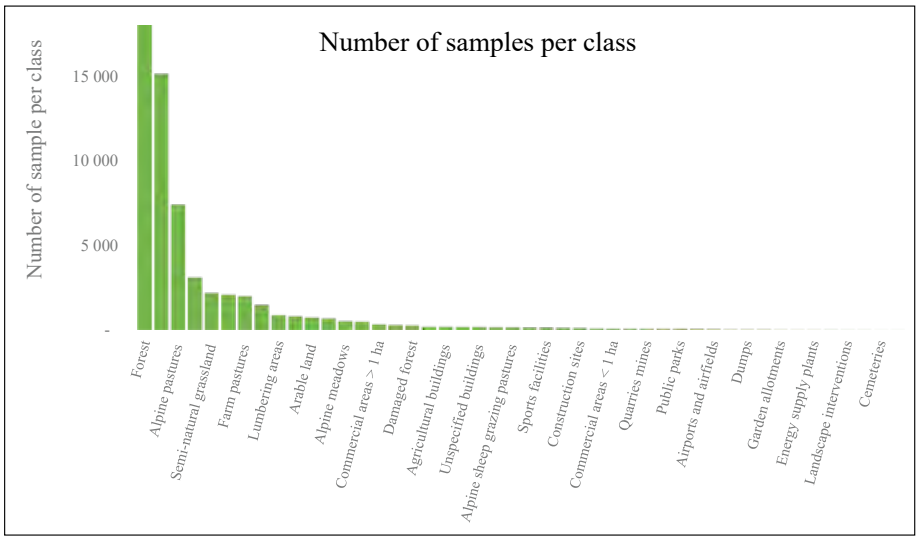

Fig. 1: The visualization of the number of samples for some of the 46 land use categories in our study area illustrates the imbalance between the classes.

Specific methods have been developed to give rare categories a more equal representation. Sampling methods (e.g. observing the samples of common classes less often during training) or specific loss functions (the functions assessing the accurarcy of the model) are commonly employed. In this study a specific loss functions known as the focal class balanced loss (fCBL)[5] is employed. The fCBL penalises the classifier more for making erroneous predictions on samples from rare classes, while being more permissive (i.e. penalising less) on the common ones. For more details about other methods, see [6].

Data

Our experiment is based on a test dataset in the greater region of Sion (VS) in the south-west of Switzerland that spreads over approximately 600 km^2. Aerial images were collected by Swisstopo in 2020 with four spectral bands: red, green, blue and near infrared (RGBI). The digital elevation model (DEM) from the Swisstopo product "SwissALTI3D" is also included as input.

The 46 land use labels are present on the area of interest. They were created for the 2013/2018 land use statistics based on aerial images from the 2013 aerial surveys for our area of interest. The study area presents a strongly skewed distribution towards some very frequent categories (figure 1): 10 land use categories make up 90% of the data, whereas the 10% remaining samples represent the 36 other classes. As illustrated in figure 1, categories such as *Forest* or *Alpine* pastures are much more frequent in the dataset than *Public parks* or *Sport facilities*.

Method

Data pre-processing

As input, we extract a surface of 50 × 50 m around each of the sample points for both the DEM and the RGBI images. Categories with less than 100 samples are removed due to their insuffcient number of samples to train a deep learinng model, leading to the suppression of a total of 18 classes.

After some preliminary experiments with 28 categories, we decided to apply manual data decontamination to our training set. This involved the relabelling or the removal of some selected examples. We removed two classes due to the time inconsistency between the label production based on aerial images from 2013 and the photography survey in 2020: the *Construction sites* category comprehends surface where temporary construction work is in progress. The 7 years gap lead to the end and the disappearance of the construction sites. Similarly, an important fraction of samples labelled as *Unexploited urban* areas seemed to have a new affectation due to rapid urban expansion.

In addition, some classes exhibit strong similarities or even overlaps, which makes them more diffcult to learn for the CNN. This is the case for three types of alpine pastures: *Alpine meadows in general, Alpine pastures in general* and *Alpine sheep grazing pastures,* and also for two types of grasslands: *Farm pastures in general* and *Semi-natural grassland in general.* Effectively, their visual similarity makes them practically undistinguishable without auxiliary information that was not accessible to us during the study. For this reason, we decided to merge categories of the same type together. Finally, the category *Unspecified buildings and surroundings* encompasses many different building types and requires information from the Federal register of buildings and dwellings (RBD) to be identified, which is not available as input to the CNN. Thus, it is also removed. The final number of categories equals 21.

Model architecture

As deep learning model, we opt for one specific CNN architecture for image classification: ResNet-50[7] pre-trained on the ImageNet[8] dataset. By pre-training, we mean that the ResNet-50 model has already been trained on a dataset of natural images (ImageNet), and is only further tuned on our dataset. This has the advantage of enjoying a model that can already make image recognition at a very accurate level and only needs to be adapted to the problem of landuse mapping.

To achieve such adaptation, the following modifications are performed to the original ResNet-50 model:

- The input layer of the CNN is modified to accept data with five channels composed of the RGBI images and the DEM, instead of the conventional three channels for natural RGB images.
- A 50% dropout layer[9] is added between the average pooling layer and the fully-connected layer to reduce overfitting.

Labels	Baseline				fCBL			Test set
Rare classes	F1-score	Precision	Recall		F1-score	Precision	Recall	size
Motorways	89.0%	88.0%	90.0%		82.0%	83.0%	81.0%	31
Alpine sports facilities	18.4%	56.0%	11.0%		24.2%	50.0%	16.0%	45
Sports facilities	34.4%	61.0%	24.0%		51.6%	85.0%	37.0%	46
Public buildings and surroundings	22.4%	44.0%	15.0%		25.4%	43.0%	18.0%	55
Agricultural buildings and surroundings	0.0%	0.0%	0.0%		16.7%	50.0%	10.0%	61
Golf courses	72.4%	63.0%	85.0%		83.0%	81.0%	85.0%	61
Parking areas	28.6%	45.0%	21.0%		39.1%	46.0%	34.0%	62
Damaged forest	66.9%	83.0%	56.0%		63.1%	74.0%	55.0%	87
Lakes	91.0%	93.0%	89.0%		88.8%	93.0%	85.0%	88
Industrial and commercial areas > 1 ha	71.5%	66.0%	78.0%		68.6%	58.0%	84.0%	101
Residential areas (blocks of flats)	52.5%	54.0%	51.0%		53.9%	52.0%	56.0%	152
Rivers streams	67.6%	83.0%	57.0%		68.6%	91.0%	55.0%	208
Arable land in general	54.4%	85.0%	40.0%		56.7%	74.0%	46.0%	225
Orchards	72.9%	76.0%	70.0%		72.0%	73.0%	71.0%	249
Average	**53.0%**	**64.1%**	**49.1%**		**56.7%**	**68.1%**	**52.4%**	
Frequent classes								
Roads	67.9%	71.0%	65.0%		66.5%	68.0%	65.0%	454
Residential areas (one and two-family houses)	75.2%	71.0%	80.0%		74.0%	74.0%	74.0%	630
Vineyards	93.0%	91.0%	95.0%		93.5%	93.0%	94.0%	932
Group of semi-natural grasslands	75.5%	75.0%	76.0%		74.9%	73.0%	77.0%	1262
Group of alpine pastures	83.0%	82.0%	84.0%		82.5%	81.0%	84.0%	2443
Unused	89.5%	91.0%	88.0%		89.5%	91.0%	88.0%	4552
Forest	93.4%	90.0%	97.0%		93.4%	91.0%	96.0%	5599
Average	**82.5%**	**81.6%**	**83.6%**		**82.0%**	**81.6%**	**82.6%**	

Tab. 1: Comparison of the classification metrics for the baseline model and the fCBL. The test set size indicates the number of samples used during the model testing to produce the accuracy metrics.

- The number of outputs in the fully-connected layer is set to 21 to match the number of classes.

Experimental procedure

Our experiments evaluate several techniques targeting the class-imbalanced distribution against a baseline model where no special measures are applied. The results for the focal class balanced loss (fCBL) and the baseline are reported here, other results are omitted for brevity reasons but can be consulted in [4].

Data are split into 60% training, 10% validation and 30% test sets (resp. about 35 000, 6000, 18 000 samples). Even if these numbers may seem important, the smallest class, Motorways, is present only 61 times in the training set due to imbalanced distribution. During the training phase, data augmentation methods such as random rotation, random flips and colour jittering are employed to slightly modify the images seen by the model at each iteration and hence increase diversity of the training samples. During the testing phase, the model performances are evaluated on the unseen test set data and accuracy metrics are computed.

Evaluation metrics

Several indicators are employed to evaluate the model performances.

- Precision indicates whether the samples classified into a class actually belong to that class and is calculated by dividing the number of correctly classified samples by the total number of samples predicted for the respective class.
- Recall is computed by dividing the correctly classified samples by the total number of samples pertaining to the class. In other words, recall measures the completeness of the predictions, assessing whether the classifier is under-predicting a given class.
- Precision and recall are best used together as the classifier is balancing between both: Increasing recall usually leads to reducing the precision and vice-versa. The F1 score combines these two metrics through a geometric mean and give a global measure of effectiveness of the classifier.

Results and Analysis

Results are presented in Table 1. The frequent classes obtain the best precision with more than 82% precision for both models. For rare classes, the baseline model receives on average lower accuracy metrics than the fCBL model. The fCBL method manages to compensate the class imbalance and increases the rare classes accuracy compared to the baseline without damaging the recognition abilities of the model of the frequent categories.

The scores for rare classes occupy a wide range of values. The classification does not only depend on the class frequency but also on the complexity of the class itself. The model trained with fCBL loss can produce accurate labels for classes even with a small number of samples, if

the class shows distinctive visual patterns that are repeated in many samples within the class but absent from other classes. The rare classes with distinctive features present on all samples such as Golf courses, Lakes and Motorways reach similar level of precision as the frequent classes, despite their small number of examples. Larger error rates seem to originate from classes with high variability between samples. The most erroneous categories among the rare classes are Alpine sport facilities, Sport facilities, Public buildings, Agricultural buildings and Parking areas. They show complex patterns and a high variability between samples from the same class. Without auxiliary data, both models have troubles to predict them. Interestingly, they are also the classes where the differences in performance between the baseline and the fCBL are the largest. The fCBL systematically outperforms the baseline with a positive difference in F1 score ranging from 3% to 16%.

Conclusion

This work addresses the class imbalance problem that is known to affect classification performance for rare categories. A loss function specialized in dealing with rare classes is compared with an unconstrained baseline deep learning model to observe their effectiveness. The category complexity appeared as a major driver of misclassification of images derived from minority classes in addition to the limited amount of data for a class. After a dataset cleaning the fCBL obtained an average reduction of 3% of the error rate on the rare classes compared to the baseline model. The use of methods addressing class imbalance offer sources for potential improvements in the classification of land use. In the future, such methods could help to further automatise the classification procedure and assist the operators in the labelling of difficult categories.

References:

[1] Office Fédérale de la Statistique, Statistique de la superficie selon nomenclature 2004, Office fédéral de la statistique. 2017.

[2] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[3] X. X. Zhu et al., 'Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources', IEEE Geoscience and Remote Sensing Magazine, vol. 5, no. 4, pp. 8–36, Dec. 2017, doi: 10.1109/MGRS.2017.2762307.

[4] D. Jordan, A. Meyer, N. Lack, M. Schonholzer, R. Leiter, and G. Milani, Bericht Prototyp KI Arealstatistik2020. Neuchâtel: Office fédéral de la statistique (BFS), 2019.

[5] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, 'Class-Balanced Loss Based on Effective Number of Samples', 2019, pp. 9268–9277.

[6] V. Zermatten, B. Kellenberger, and D. Tuia, 'Predicting Land Usage from Aerial Images with Deep Learning: A Case Study in the Valaisan Alps focusing on Class Imbalance', École Polytechnique Fédérale de Lausanne (EPFL), Sion, 2021.

[7] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', 2016, pp. 770–778.

[8] O. Russakovsky et al., 'ImageNet Large Scale Visual Recognition Challenge', Int J Comput Vis, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 'Dropout: a simple way to prevent neural networks from overfitting', J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

Valérie Zermatten
Benjamin Kellenberger
Devis Tuia
Environmental Computational Science and Earth Observation Laboratory
Ecole Polytechnique Fédérale de Lausanne
valerie.zermatten@alumni.epfl.ch
benjamin.kellenberger@epfl.ch
devis.tuia@epfl.ch