Scene classification for mining sites from satellite images using 3D covolutional neural networks

André E. C. Oliveira, Matheus C. Domingos, Valdivino A. de Santiago Júnior

Laboratório de Inteligência ARtificial para Aplicações AeroEspaciais e Ambientais (LIAREA)
Programa de Pós-Graduação em Computação Aplicada - CAP
Instituto Nacional de Pesquisas Espaciais - INPE

andre.estevam.unb@gmail.com matheuscorreadomingos@gmail.com
valdivino.santiago@inpe.br

Motivation

There has been extensive research regarding how to leverage temporal as well as spatial features for recognition tasks in Remote Sensing. Most approaches employ a pixel-by-pixel analysis when dealing with temporal data, which does not account for relationships between neighbouring pixels; while spatial features extraction already enjoys a number of well-stablished methods, but it does not take into account pixel value change through time.

A variety of methods have been proposed to handle spatiotemporal data, but most rely on adptations of Recurrent Neural Networks, which might not be the best as it handles spatiotemporal data as first spatial, and then temporal, effectively decoupling both dimensions. 3D convolutions can be an effective tool as it traverses the space and time dimensions at the same time.

Objective

This work evaluates the effectiveness of 3D-Convolutional Neural Networks (3DCNNs) for scene classification from high-resolution remote sensing images, specifically mining sites. The following actvities were carried out to achieve this objective:

- Create a dataset for scene classification of mining sites from Planet imagery and Mapbiomas classification;
- Train, validate and test a 3D-CNN on the newly created dataset;
- Train, validate and test a traditional CNN using a single ”slice” *t* of the dataset;
- Compare the 3D-CNN model’s performance with that of the (2D) CNN to evaluate the impact of incorporating temporal information, specifically for mining sites.

Methodology

A dataset containing 264,000 scenes from Planet imagery, covering the period between September 2020 and May 2024, was built. This imagery was labeled using Mapbiomas products [2] to create ground-truth data of mining sites. Two neural networks were developed: a (2D) CNN that processes a single image and a 3DCNN that captures spatial and temporal features from a datacube. Both models were evaluated using accuracy, precision, and recall.

The creation of the dataset involved developing a plugin for QGis, which generated the frames later used to clip imagery into scenes.

For a better understanding of the difference between a 2D and a 3D convolution, see Figure 1. While a 2D convolution (using sliding window) always results in a matrix, a 3D convolution preserves volume if we are dealing with a datacube, effectively preserving the temporal dimension.

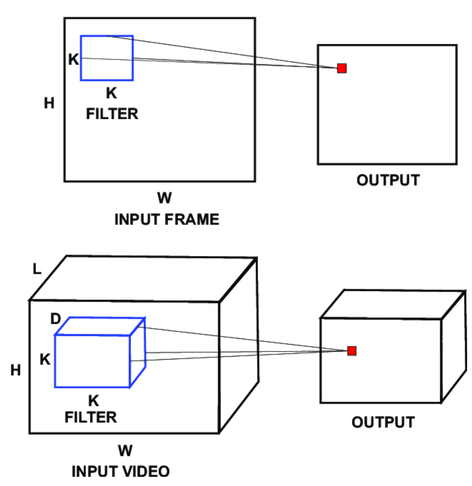

Figure 1: *Difference between a 2D and a 3D convolution operation [1].*

The overall architecture of both neural networks can be seen in 2. It follows the classic structure of features extraction and then classification. The classification step is a fully connected network of 3 layers with 4096, 4096 and 1 neurons, respectively, where the last neuron has a sigmoid activation function (for binary classification). The main difference between the (2D) CNN and the 3DCNN models rests in the features extraction phase: the kernel dimensions are 3x3 and 3x3x3, respectively. This way we asserted that any difference in performance could be attributed solemnly to the addition of the temporal dimension.

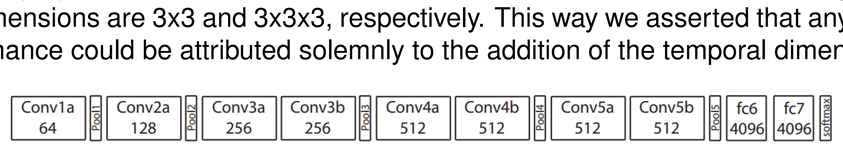

Figure 2: *Overall architecture for both (2D)CNN and 3DCNN [3].*

For training we did not utilized all samples due to technical limitations. We trained both models for 10 epochs each, using the Adam optimizer with learning rate of 0.001, beta values 0.9 and 0.999, epsilon 1e-08 and no weight decay.

We utilized the PyTorch framework for implementation of the model and training routine, using t4 GPUs available through Colab.

Results

Due to some technical limitation, we could not use every sample from the dataset nor train the model for many iteration, restricting it’s potential.

The experiments described in this study were executed using a set of 500 samples per class for training and 100 samples for test, randomly selected. The number of epochs was set to 10, and the threshold used was 0,9, i.e., outputs bigger than 0,9 were considered ”mining”.

As we can see in Figure 3, the training loss decreases smoothly until reaching a minimum in 10 epochs. This shows that the CNN model is learning well from the train set, being able to use it’s knowledge for inference in the test set as well.

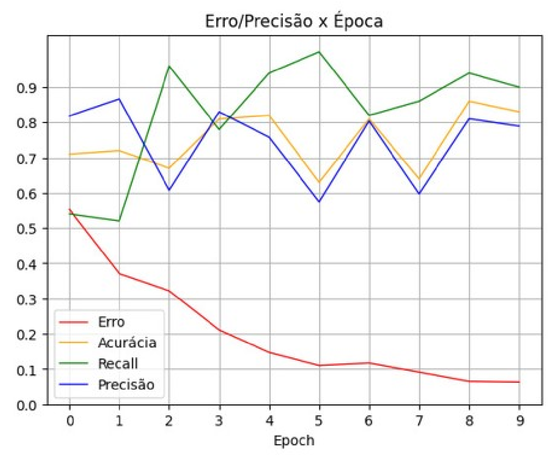

Figure 3: *Error and precision metrics through each training iteration for the CNN model.*

Curiously, the 3DCNN converged faster than the 2DCNN, as we can see in figure 4, but it seems to start overfitting as it reaches 10 plus epochs (the loss reaches a minimum while the performance metrics start to deviate).

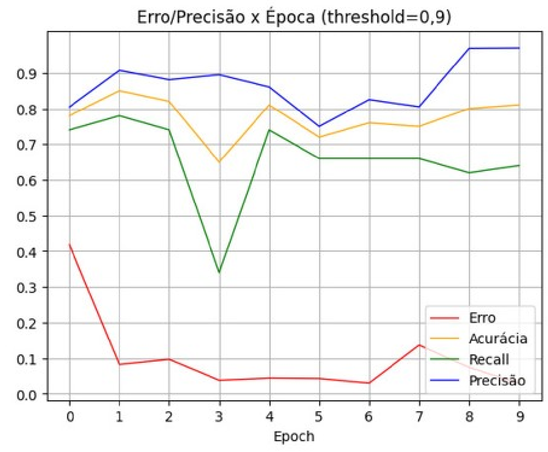

Figure 4: *Error and precision metrics through each training iteration for the 3DCNN model.*

The 2D CNN, which used only spatial information, achieved an accuracy of 86%, a recall of 94%, and a precision of 81%. The 3D-CNN, incorporating temporal data, achieved a slightly lower accuracy of 85% but a significantly higher precision of 90%, indicating that the model was more confident in its predictions. Although the 3D-CNN had a lower recall (74%), it demonstrated the advantage of temporal data for detecting subtle changes in mining activities.

Conclusions

According to the results for each model, the fact that the 3DCNN achieved better precision indicates that it’s inference can be more reliable than that of the (2D) CNN model, even if their overall accuracy were very similar. In contrast, the CNN’s recall outperformed it’s counterpart, which indicates that it mislabeled more but let fewer mining sites scenes pass undetected, which can be more desired for illegal mining sites prevention efforts, for example.

For future work, several directions promise to further improve the analysis and interpretation of remote sensing data. Compared to classical approaches such as pixel-by-pixel temporal analysis, there is potential for significant advances by integrating more sophisticated and detailed methods. Further analysis of the structured dataset can reveal additional patterns and insights, particularly in differentiating between mining and industrial mining, and in understanding the temporal progression of mining sites.

Acknowledgements

This work was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), INPE (National Institute for Space Research) and CAP (Department of Applied Computing).

References

[1] Wheidima Melo, Eric Granger, and Abdenour Hadid. Combining global and local convolutional 3d networks for detecting depression from facial expressions. In *14th IEEE International Conference on Automatic Face and Gesture Recognitio*, 05 2019.

[2] MapBiomas Project Team. Mapbiomas: Brazilian land cover and use mapping project, 2020.

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.