GeoLabels: Towards Efficient Ecosystem Monitoring using Data Programming on Geospatial Information

David Dao*
Department of Computer Science
ETH Zurich
8092 Zurich, Switzerland
daviddao@inf.ethz.ch

Johannes Rausch*
Department of Computer Science
ETH Zurich
8092 Zurich, Switzerland
johannes.rausch@inf.ethz.ch

Ce Zhang
Department of Computer Science
ETH Zurich
8092 Zurich, Switzerland
ce.zhang@inf.ethz.ch

Abstract

Monitoring, Reporting and Verification (MRV) systems for land use play a key role in the decision-making of climate investors, policymakers and conservationists. Remote sensing is commonly used for MRV but practical solutions are constrained by a lack of labels to train machine learning-based downstream tasks. GeoLabels is an automated MRV system that can rapidly adapt to novel applications by leveraging existing geospatial information and domain expertise to quickly create training sets through data programming. Moreover, GeoLabels uses dimensionality reduction interfaces, allowing non-technical users to create visual labeling functions.

1 Introduction

Humanity is in the midst of an unprecedented climate crisis, demanding urgent, decisive action in line with the Paris Agreement. Land use plays a crucial role in our climate, taking up about a quarter of annual anthropogenic carbon emissions. Careless usage of land includes a wide range of critical issues, including deforestation and forest degradation through agriculture, turning natural carbon sinks into carbon sources. Monitoring, Reporting and Verification (MRV) of land use is a key task of global forest conservation efforts [1], climate finance instruments, such as REDD+, Payment for Ecosystem Services and biodiversity schemes [2]. As such, MRV creates baselines and impact assessments for policy and decision makers. The promise that remote sensing holds for MRV is considerable. Earth observations from aircraft or satellite-based sensors combined with recent advances in computer vision have an enormous potential in automatically monitoring land that is vast and otherwise difficult to access. Although remote sensing provides a huge supply of data of detailed temporal and spatial resolutions, many practical solutions and actionable projects are constrained by a lack of task-specific labels. Current approaches to this problem require manual labeling (leveraging visual interpretation tools such as Collect Earth [3]) and an enormous amount of data cleaning, munging, and handling from domain experts. As such, annotating data for automated systems has become the biggest cost (and time) factor. The lack of available labeled data prevents the rapid adaptation/scaling of systems to new downstream tasks, or for new data sources.

We propose GeoLabels: An automated MRV land use monitoring system that addresses the high cost of manual annotation by leveraging existing geospatial information and weak supervision of

* Equal contribution.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

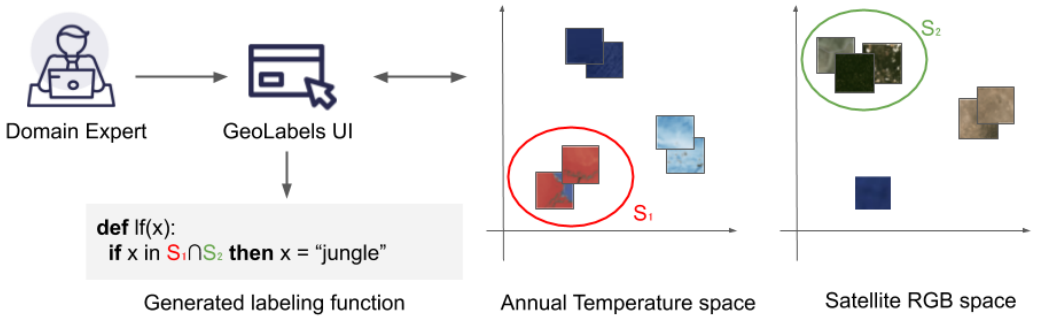

Figure 1: An overview of the labeling process of GeoLabels. Domain experts can select sets of tiles in different modalities (displayed using dimensionality reduction on tile2vec) and define visual labeling functions for data programming.

domain experts. GeoLabels provides a flexible approach that scales to different ecological and environmental downstream tasks for MRV that require non-technical domain knowledge such as tree species classification, land use segmentation or emission tracking.

2 Data Programming for Geospatial Maps

GeoLabels applies data programming [4], a method that is used to build massive, weakly labeled datasets by allowing users to define labeling functions instead of hand-labeling individual examples [5]. Based on overlaps of multiple labeling functions for given data points, a generative model is trained in order to predict probabilistic labels.

2.1 Extending Labeling Functions for Geospatial Data

Data programming is commonly used for textual data, where unsupervised methods such as Word2Vec [6] are used in order enrich words with high-dimensional representations. The lack of representation that incorporate rich contextual information for geospatial data has previously posed a challenge in writing labeling functions with high discriminate power. Tile2Vec [7], a technique for unsupervised representation learning on spatially distributed data has recently been introduced. Using this technique, map tiles can automatically be associated with vector embeddings that encode information such as visual similarity or spatial proximity. As vector embeddings can be generated for different modalities e.g. hyper-spectral satellite images, and spatial environmental and ecological maps, human annotators are able to define labeling functions that incorporate one or more modalities. For instance, environmental maps (e.g. annual temperature, annual precipitation) combined with ecological maps (e.g. nitrogen mineralization, soil PH) and raw satellite data (e.g. Sentinel, LANDSAT) can provide human annotators enough information to determine reforestation potential [8]. Combining overlapping labeling functions with the generated vector embeddings allows for the training of a generative model that automatically estimates the accuracy of the individual labeling functions.

2.2 Dimensionality Reduction as Labeling UI for Domain Experts

GeoLabels leverages domain expertise from non-technical users by allowing them to visually design labeling functions with dimensionality reduction (see Figure 1). Tile2Vec representations of different geospatial data sources are mapped to a lower dimension via PCA - allowing domain experts to visually interpret data. The domain experts can then select sets from different sources and define relationships between them. GeoLabels will automatically generate an internal labeling function from the user's input (e.g. distance from point x in Tile2Vec space).

3 Discussion

GeoLabels facilitates training of machine learning-based MRVs for various downstream tasks by leveraging domain expertise and combining it with a variety of available geospatial data sources.

Over the coming years, geospatial data is expected to grow rapidly, with more and more observation infrastructure in place (smaller satellites, more cost-efficient launching rockets). Better technology deployed as part of this growing infrastructure will deliver higher resolution images, as well as different non-image data streams. An interpretable and adaptable MRV system is therefore crucial to guide climate investors and policymakers in their decision-making.

Acknowledgments

The authors are thankful for the guidance and advise by the CONAF (Daniel Montaner, Cesar Mattar, Jose Antonio Prado). Part of this research has been developed as part of the OpenSurface platform and a real-world pilot in Chile, which was launched at the COP25 United Nation's Climate Summit. OpenSurface is funded by IDBLab and EIT Climate-KIC.

References

[1] David Dao, Catherine Cang, Clement Fung, Ming Zhang, Nick Pawlowski, Reuven Gonzales, Nick Beglinger, and Ce Zhang. GainForest: Scaling Climate Finance for Forest Conservation using Interpretable Machine Learning on Satellite Imagery. *ICML Climate Change AI workshop 2019*.

[2] Sandra Díaz, Sebsebe Demissew, Julia Carabias, Carlos Joly, Mark Lonsdale, Neville Ash, Anne Larigauderie, Jay Ram Adhikari, Salvatore Arico, András Báldi, et al. The IPBES Conceptual Framework—connecting nature and people. *Current Opinion in Environmental Sustainability*, 14:1–16, 2015.

[3] Adia Bey, Alfonso Sánchez-Paus Díaz, Danae Maniatis, Giulio Marchi, Danilo Mollicone, Stefano Ricci, Jean-François Bastin, Rebecca Moore, Sandro Federici, Marcelo Rezende, et al. Collect earth: Land use and land cover assessment through augmented visual interpretation. *Remote Sensing*, 8(10):807, 2016.

[4] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, pages 3567–3575, 2016.

[5] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proc. VLDB Endow.*, 11(3):269–282, November 2017.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[7] 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.

[8] Jean-Francois Bastin, Yelena Finegold, Claude Garcia, Danilo Mollicone, Marcelo Rezende, Devin Routh, Constantin M. Zohner, and Thomas W. Crowther. The global tree restoration potential. *Science*, 365(6448):76–79, 2019.