A Data-Centric AI Approach to Crop Detection with Deep Learning

Abstract/Rezumat

TEODORA SELEA

Supervisor: prof. dr. Dana Petcu

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Computer Science

West University of Timișoara
România

2024

Contents

Abstract

The exponential increase in the global population and the adverse effects of climate change pose substantial obstacles to agricultural methods. By 2050, food production needs to be increased by 60% to support the growing population demands. Simultaneously, climate change induces alterations in temperature and precipitation patterns, resulting in inconsistent agricultural circumstances and heightened occurrence of severe weather phenomena.

These issues necessitate inventive ways to improve crop yield and guarantee the long-term viability of farming methods. Scientists and agriculturalists are investigating novel approaches and advanced technologies to tackle these difficulties, including precision agriculture and sustainable farming practices. Current methodologies in crop detection using remote sensing and Artificial Intelligence face significant limitations due to the vast diversity and temporal dynamics of agricultural landscapes. The diversity of crop kinds, growth stages, and spatial patterns of farm landscapes presents a difficulty in creating precise and reliable models for crop detection. Moreover, the absence of extensive, specialised datasets for training artificial intelligence models impedes the capacity to extrapolate discoveries and construct dependable models.

These limitations underscore the need for a data-centric AI approach to handle agricultural landscapes' diversity and temporal dynamics and utilise large-scale, specific datasets for training and testing. A data-centric AI approach would enhance the accuracy and effectiveness of agricultural deep learning models. This thesis introduces **a large-scale, high quality, crop detection dataset, AgriSen-COG**, and the methodology to create and extend it.

First, during the dataset creation stage, we propose a novel parcel aggregation method to represent each crop parcel in the dataset accurately. This method considers individual crop parcels' spatial distribution and size, allowing for a more precise representation of the agricultural landscape. Second, we proposed a submodular evaluation framework to assess

the representativeness of the proposed aggregation method. The framework ensures that the aggregated time series captures the temporal dynamics of each crop parcel's growth and provides a robust basis for training and testing crop detection models.

Third, we employ an unsupervised anomaly detection process to improve the quality of the AgriSen-COG dataset. We performed a sequence of experiments using multiple deep-learning models to demonstrate the enhanced performance of the models when utilising a carefully selected dataset. Fourth, we propose to use another submodularity evaluation framework to assess the validity of the anomaly detection process analytically, disregarding the model bias.

In addition to this, we provide an extensive deep learning benchmark, highlighting how a data-centric approach can improve the deep learning models' performance. Finally, we analyse processing large amounts of geospatial data and provide further recommendations.

These contributions address the challenges of data diversity and temporal dynamics in agricultural landscapes. The development of the AgriSen-COG dataset, a large-scale crop detection dataset, has practical implications for both the AI and agricultural research communities. This dataset allows researchers to build and evaluate novel models and theories in crop detection and management thanks to its extensive size, diversity, and temporal resolution.

Key Words: data-centric AI, data quality, deep learning, distributed processing, crop detection, agriculture, sentinel-2, LPIS.

Rezumat

Creșterea exponențială a populației globale și efectele adverse ale schimbărilor climatice reprezintă provocări pentru domeniul agricol. Până în 2050, producția alimentară trebuie sa crească cu 60% pentru a sustine cerințele populației în creștere. În același timp, schimbările climatice induc modificări ale modelului temperaturilor și precipitațiilor, rezultând în circumstante agricole inconsistente si o incidență crescută a fenomenelor meteorologice severe.

Aceste probleme necesită metode pentru a îmbunătăți randamentul culturilor și pentru a garanta viabilitatea pe termen lung a metodelor de agricultură. Oamenii de știință și agricultorii investighează abordări noi și tehnologii avansate pentru a aborda aceste dificultăți, inclusiv agricultura de precizie și practici agricole sustenabile. Metodologiile curente în detectarea culturilor folosind teledetecția și Inteligența Artificială se confruntă cu limitări semnificative datorită diversității vaste și dinamicii temporale ale peisajelor agricole. Diversitatea tipurilor de culturi, stadiile de creștere și modelele spațiale ale peisajelor agricole reprezintă o dificultate în crearea unor modele precise și de încredere pentru detectarea culturilor. Mai mult, absența unor seturi de date specializate, extinse, pentru antrenarea modelelor de inteligență artificială împiedică capacitatea de a extrapola descoperirile și de a construi modele fiabile.

Aceste limitări subliniază necesitatea unei abordări AI centrate pe date pentru a gestiona diversitatea și dinamica temporală a peisajelor agricole și pentru a utiliza seturi de date specifice, la scară largă, pentru antrenament și testare. O abordare AI centrată pe date ar îmbunătăți acuratețea și eficacitatea modelelor de tip deep learning. Această teză introduce un set de date de detecție a culturilor la scară largă și de înaltă calitate, AgriSen-COG, și metodologia de creare și extindere a acestuia.

În primul rând, în etapa de creare a setului de date, propunem o metodă nouă de agregare a parcelelor pentru a reprezenta fiecare parcelă de cultură din setul de date în mod precis. Această metodă ia în considerare distribuția spațială și dimensiunea parcelelor

de cultură individuale, permițând o reprezentare mai precisă a peisajului agricol. În al doilea rând, am propus un cadru de evaluare submodular pentru a evalua reprezentativitatea metodei de agregare propuse. Cadrul asigură că seria temporală agregată captează dinamica temporală a creșterii fiecărei parcele de cultură și oferă o bază solidă pentru antrenarea și testarea modelelor de detectare a culturilor.

În al treilea rând, utilizăm un proces de detectare a anomaliilor nesupervizat pentru a îmbunătăți calitatea setului de date AgriSen-COG. Am efectuat o serie de experimente utilizând mai multe modele de tip deep learning pentru a demonstra performanța îmbunătățită a modelelor atunci când utilizează un set de date atent selectat. În al patrulea rând, propunem să utilizăm un alt cadru de evaluare a submodularității pentru a evalua analitic validitatea procesului de detectare a anomaliilor, ignorând prejudecățile modelului.

În plus, oferim un benchmark extensiv de învățare profundă, evidențiind cum o abordare centrată pe date poate îmbunătăți performanța modelelor de învățare profundă. În final, analizăm prelucrarea unor cantități mari de date geospațiale și oferim recomandări suplimentare.

Aceste contribuții abordează provocările diversității datelor și dinamicii temporale în peisajele agricole. Dezvoltarea setului de date AgriSen-COG, un set de date de detecție a culturilor la scară largă, are implicații practice atât pentru comunitățile de cercetare AI, cât și pentru cele agricole. Acest set de date permite cercetătorilor să construiască și să evalueze modele și teorii noi în detectarea și gestionarea culturilor datorită dimensiunii sale extinse, diversității și rezoluției temporale.

Cuvinte cheie: AI centrat pe date, calitatea datelor, învățare profundă, prelucrarea distribuită, detectarea culturilor, agricultură, Sentinel-2, LPIS.

Contributions

This thesis has made several contributions, which are summarized as follows:

(1) **Detailed Methodology for the Creation of a Large-Scale Crop Detection Dataset**: Addressed the challenges associated with constructing an extensive dataset for crop detection, emphasizing the steps and establishing processing benchmarks for future research.

(2) **Geospatial Processing Techniques**: Efficient file formats and the use of distributed processing to manage and process geospatial data efficiently. Innovative techniques for extracting time series data from large-scale data cubes, demonstrating how to leverage computing clusters for enhanced geospatial processing.

(3) **New parcel aggregation method**: The introduction of a new parcel aggregation method using the Dynamic Time Warping Barycenter Averaging (DBA) technique, which outperforms traditional methods by preserving phenological patterns within ecological datasets.

(4) **Development of AgriSen-COG Dataset**: Introduced a large-scale, comprehensive dataset for crop type mapping, covering five distinct regions and including 61 tiles with approximately 7 million fields. This dataset stands out for its scale and the inclusion of diverse geographical areas.

(5) **Anomaly Detection Method**: Proposed a novel anomaly detection methodology tailored for improving the quality of agricultural datasets, thereby enhancing the accuracy of crop type classification and mapping.

(6) **Submodular Evaluation Framework**: Developed an evaluation framework based on submodularity to assess the efficacy of the anomaly detection method. This framework offers a nuanced approach to evaluating methodological innovations in dataset curation.

(7) **Extensive Benchmarking Efforts**: Conducted thorough benchmarking for two crucial use cases—crop type classification and crop type mapping. Demonstrated

experimentally the improved performance of models when using the curated version of the AgriSen-COG dataset, thus underscoring the dataset's value.

(8) **Code & Data Repositories**: Provision of a comprehensive repository containing all intermediate steps and processing code, ensuring research transparency and reproducibility.

These contributions collectively advance the capabilities for agricultural monitoring based on AI techniques, offering new tools and methodologies for researchers and practitioners in the domain.

Thesis Outline

This thesis is divided in five chapters. The following is a summary of each of the chapters in this thesis:

Chapter 1 — Introduction: This chapter provides a brief context for the current research, focusing on a data-centric approach to crop detection. It also includes the identified research problems and the motivation to focus on data. Alongside, it presents the objectives of this thesis and the research contributions.

Chapter 2 — Background: The chapter presents a brief literature review of related work, providing an overview of current free satellite remote sensing products. It also includes a description of traditional machine learning methods, basic deep learning concepts, and recent deep learning-based semantic segmentation and time series classification developments.

Chapter 3 - Creating a Large Scale Crop Detection Dataset: This chapter presents the methodology and the Big Data processing involved in creating a large-scale dataset for crop detection. It also details several parcel aggregation techniques and the design of a submodular evaluation framework to evaluate the proposed parcel aggregation method.

Chapter 4 - Improving Deep Learning Performance: This chapter describes the proposed large-scale dataset, AgriSen-COG. It also details how we improved the dataset's quality by employing an anomaly detection method. It also presents the design of a submodular evaluation framework to validate the previously employed anomaly detection method.

Chapter 5 Conclusions and Future Work: The purpose of this chapter is to describe the most important aspects of this thesis by providing an overview of what was accomplished, focussing on the most important results, and analysing the contributions that this thesis has made. This chapter presents the conclusions that can be derived from the research findings and suggestions for further research are provided.

List of Publications

The techniques and experimental results developed in the context of this PhD thesis have been shared through several scientific publications listed below.

Journals

(1) Selea, T. (2023). AgriSen-COG, a Multicountry, Multitemporal Large-Scale Sentinel-2 Benchmark Dataset for Crop Mapping Using Deep Learning. Remote Sensing, 15(12), 2980. https://doi.org/10.3390/rs15122980
JCR - Q1 (Geosciences, Multidisciplinary), IF - 5.0 (Q1), AIS (Q2)
8 | 4 points, Citations: 0 (WoS/Scopus/Google Scholar)

Conferences

(1) Selea, T., & Pslaru, M. F. (2020, September). AgriSen-A Dataset for Crop Classification. In 2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp. 259-263). IEEE. https://doi.org/10.1109/SYNASC51798.2020.00049
Rank - D
1 points, Citations: 2 (WoS/Scopus/Google Scholar)

(2) Neagul, M., Panica, S., & Selea, T. (2019, September). Experiences in building a distributed Earth Observation Platform. In 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP) (pp. 545-550). IEEE. https://doi.org/10.1109/ICCP48234.2019.8959718
Rank - C
2 points, Citations: 0 (WoS/Scopus/Google Scholar)

(3) Selea, T., & Neagul, M. (2017, September). Using Deep Networks for Semantic Segmentation of Satellite Images. In 2017 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp.

409-415). IEEE. https://doi.org/10.1109/SYNASC.2017.00074

Rank - C

2 points, Citations: 4 (WoS/Scopus/Google Scholar)

Publications' score: 13 points

Total citations: 6 (WoS/Scopus/GoogleScholar)