Fachhochschule Nordwestschweiz
Hochschule für Technik

Studiengang Data Science

Konzept - Challenge X HS 2023

Nutzpflanzenerkennung

Challengenummer: cx

Challenge von:

Daniela Herzig
daniela.herzig@students.fhnw.ch

Thomas Mandelz
thomas.mandelz@students.fhnw.ch

Jan Zwicky
jan.zwicky@students.fhnw.ch

Data:

Zürich / Thurgau

Schweiz (ohne Label)

Fachbetreuer:

Daniel Perruchoud
daniel.perruchoud@fhnw.ch

Windisch, 1. August 2023

Inhaltsverzeichnis

1 Fragestellung

Jedes Jahr müssen Bauern, die Direktzahlungen beantragen, den Kantonen jeweils ihre Strukturdaten wie Flächen oder Anzahl Tiere sowie Daten zu der Lage ihrer Hofflächen melden. Einige Kantone haben dies mittlerweile digitalisiert mit dem sogenannten AgriPortal. Dort können die Landwirte ihre Daten digital erfassen und haben den Vorteil einer GIS-Anbindung. Die Eingabe erfolgt jedoch immer noch händisch.

Ziel der Challenge X ist es, mithilfe von Satellitendaten herauszufinden, was auf den Landwirtschaftsflächen angebaut wird. Hierzu müsste die Feldsegmentierung und Klassifizierung automatisiert werden und die Landwirte müssten nur Inhaber bzw. Pachtänderungen manuell erfassen. Des Weiteren könnte dies für viele andere Fragestellungen wie Vorhersage von Produktionsmengen, Entwicklung von Landwirtschaftsflächen bzw. deren Anbau aufgrund der Klimaveränderung genutzt werden.

Aufgrund der Komplexität der Fragestellung, werden wir uns insbesondere auf die Klassifizierung von bereits segmentierten Flächen konzentrieren.

2 Datenquellen

Als Grundlage dienen uns Satellitendaten. Da die Nutzpflanzen sich von der Aussaat bis zur Ernte stark verändern und dies auch nicht zeitgleich über alle Sorten, benötigen wir eine Sequenz von Satellitenbildern. Diese satellitenbilder müssen segmentiert (=Ackerflächen) und mit Nutzpflanzenarten gelabelt sein, damit sie weiterverwendet werden können für die Modellierung.

Wir möchten zwei unterschiedliche Datensätze nutzen. Beide Datensätze haben Polygonzüge im LV95-Koordinatensystem und weisen den Polygonen das entsprechende Label für die Nutzflächen zu. Diese Label weisen eine unterschiedliche Granularität auf, können jedoch mithilfe der sogenannten Kulturcodes harmonisiert werden (Kulturcodes / Labels).

2.1 Kanton Zürich und Thurgau

Der Datensatz beinhaltet 116'000 instanzierte Polygone mit einem entsprechenden Label für die Nutzfläche. Diese ground truth wurde vom Amt für Landwirtschaft bereitgestellt und durch die ETHZ aufgearbeitet[1]. Die 48 Label wurden in drei Hierarchien unterteilt (siehe Abbildung 1). Des Weiteren enthält der Datensatz 28'000 Sentinel-2 Bilder patches (24x24 pixels), wobei jedes Bild 71 mal beobachtet wurde im Zeitrahmen von Januar 2019 bis Dezember 2019. Dies deckt eine Fläche von 50 x 48 km ab. Der sogenannte ZueriCrop Datensatz ist Teil der torchgeo Bibliothek: Zürich/ Thurgau

[1] Crop mapping from image time series: deep learning with multi-scale label hierarchies, ETHZ, August 2021, Link

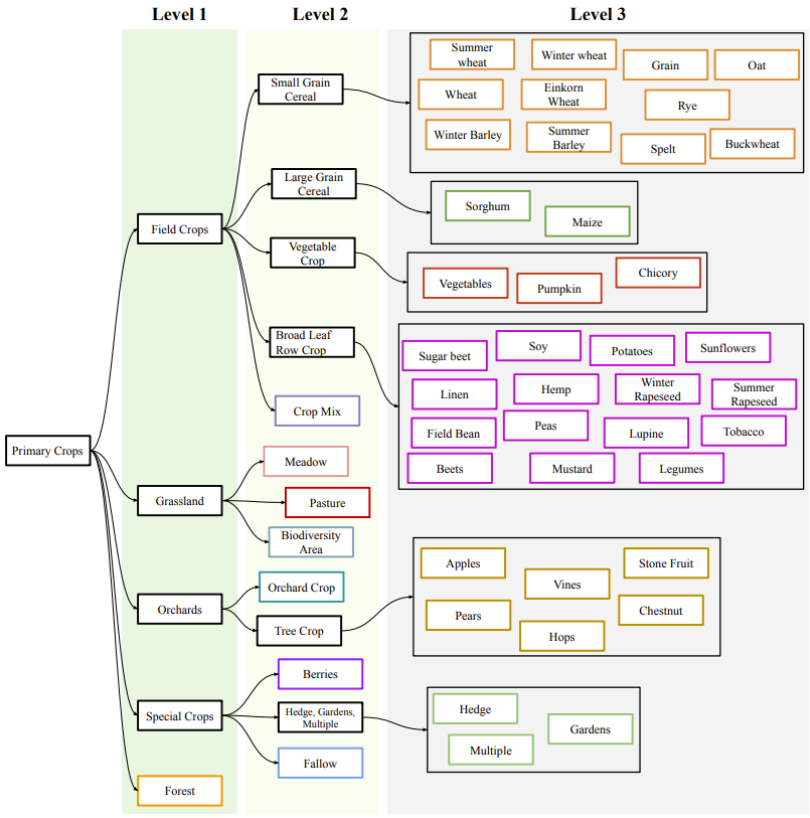

Abbildung 1: Hierarchie der Nutzpflanzen-Labels

2.2 Kanton Bern

Im Kanton Bern müssen die Daten aus unterschiedlichen Quellen aufbereitet werden. Die öffentlich zugängliche Ebene der Nutzungsflächen im Geoportal des Kanton Berns beinhalten die Informationen von Polygonen, die mit einem schweizweit harmonisierten Kulturcode attribuiert wurden. Diese Daten sind für das Jahr 2022 verfügbar. Die Granularität des Kulturcodes ist noch feiner als die 48 Label der ETH Zürich, dies gilt es in einem ersten Schritt zu harmonisieren.

Für die Satellitendaten von Copernicus verwenden wir Sentinel-2 Level 2a Daten, da sie für landwirtschaftliche Fragestellungen empfohlen werden und im Datensatz vom Kanton Zürich und Thurgau verwendet werden. Diese Daten beinhalten multispektrale, atmosphärische korrigierte Oberflächenreflexionsbilder, in denen bereits die Streuung von Reflexionen durch Aerosole oder Luftpartikel korrigiert wurden.

3 Vorgehen

Das Vorgehen für das Semester sieht wie folgt aus:

Grundlagen (SW 1-2):

- Administrativer Aufbau Gitlab / clean code / Kanban Board / detaillierte Definition Abgabeobjekte
- Literatur ETHZ Paper
- Literatur Segmentierung und Klassifizierung

Daten aufbereiten (SW 3-5):

- Umgang mit geo-Formaten
- Umgang mit Segmentierung mithilfe von Polygonzügen in Sentinel Daten Grids
- Datensatz Kanton Bern aufbereiten
- Allgemeine explorative Datenanalyse inkl. ersten Erkenntnissen
- Definition von Trainingsdaten (räumliche und zeitliche Dimension der "Grids")

- Train/Validate/Test Split Strategie und Umsetzung
- Messgrössen und qualitative Prüfung definieren

Baseline (SW 6-8):

- Pipeline aufbauen
- Baseline bauen
- Auswertung prüfen mit Baseline

ModelSelection (SW 9-12):

- Literaturrecherche weitere Architekturen
- Aufbau Modelle
- Model selection
- Hyperparameter tunen

Abschluss (SW 13-14):

- Bericht und Präsentation
- ggf. Demonstrationstool / Dashboard

3.1 Modellierungskonzepte

In einem ersten Schritt möchten wir eine Pipeline aufbauen, indem wir verschiedene Modellarchitekturen testen können. Als Baseline möchten wir einen einfachen Ansatz verwenden wie zum Beispiel RandomForest. In einem weiteren Schritt werden wir die Architektur des im Paper der ETHZ beschriebenen Netzwerk (ms-convSTAR, CNN + RNN kombination) verstehen und implementieren. Mithilfe eines Literaturstudiums werden wir sinnvolle andere Netz-Architekturen aufbauen. Dabei möchten wir mithilfe unserer Evaluierung erkennen, wo die Schwächen des ETH- Modells liegen und es dort verbessern. Wir möchten explizit nicht zufällig neue Architekturen ausprobieren und hoffen, dass diese besser sind.

3.2 Testkonzepte

Da wir die Modelle auf 48 Labels auswerten möchten, die alle noch unterschiedliche Grössen aufweisen, müssen wir uns Strategien überlegen, die diese Umstände berücksichtigen. Einerseits ist es möglich mit quantitativen Metriken wie Accuracy, Precision, Recall und F1 score die einzelnen Labels auszuwerten. Um jedoch eine globale Aussage über ein Modell machen zu können, sind gewichtete Metriken zu berücksichtigen.

Eine Confusionmatrix kann bei einer ersten qualitativen Beurteilung von Modellen helfen - so kann erkannt werden, ob bei gewissen Labels öfters eine Verwechslungsgefahr besteht oder Modelle Mühe haben, spezifische Labels überhaupt zu erkennen. Dies ersetzt jedoch nicht, dass wir uns einzelne Ergebnisse direkt anschauen um die Problematiken zu identifizieren.

Sollten wir erkennen, dass unsere Modelle Mühe haben mit Labels, die wenig vorkommen, sind andere Ansätze für das Trainingsset zu prüfen. Siehe hierzu auch das Kapitel Herausforderungen.

4 Abgabeobjekte

Dani, das ist uns noch nicht klar. Sind die Vorgaben die gleichen, wie bisher inkl. Pitch-Video etc.?

Abgabeobjekt	**Abgabedatum**
Exposé	28.08.2023
CHX Launch	18.09.2023
Arbeit / Portfolio	18.01.2024

Tabelle 1: Abgabeobjekte

5 Herausforderungen

Die Fragestellung beinhaltet sowohl eine Segmentierung von Bilddaten als auch die Klassifizierung der erkannten Segmente. Des weiteren haben wir im Team wenig Erfahrung mit sequentiellen Bilddaten. Daher ist es wichtig, dass wir uns beim Wissensaufbau und der Literaturrecherche gegenseitig austauschen und bei unserem CHX Coach Inputs holen, wenn wir nicht mehr weiterkommen. Die bereits durch die ETHZ beschriebenen Erkenntnisse können uns in der Erarbeitung ebenso unterstützen.

Die Daten vom Kanton Bern liegen uns in verschiedenen Geoformaten vor, die miteinander interagieren müssen. So kommen als Input Polygone mit LV-95 Koordinaten in die Pipeline und daraus sollen segmentierte Bilddaten von Sentinel-2 aufbereitet werden. Dieses Vorgehen muss genau definiert werden und die Interaktion mit weiteren Formaten müssen im Code klar ersichtlich sein.

Des Weiteren werden wir trotz der bereits prozessierten Sentinel Level2a Daten noch Rauschen vorfinden, wie zum Beispiel Gebäude, Wolken, Saharastaub oder ähnliches. Hier könnten Ansätze, wie das Vergrössern des Trainingsdatensatzes oder eine Auswahl nur von bestimmten Satellitenbildern mit einer bestimmten Qualität weiterhelfen (zum Beispiel nur eine geringe Wolkenüberdeckung).

Durch die nur geringe Anzahl von Daten, besteht die Gefahr von Overfitting. Dem versuchen wir mithilfe von Regularisierungstechniken wie Dropout, Data Augmentation oder Ansätzen wie Up- oder Downsampling zu begegnen.

Bei der Evaluierung ist darauf zu achten, dass die Labelgruppen nicht alle gleich gross sind. Dies ist mit den entsprechenden Evaluierungsmetriken zu berücksichtigen. Gewisse Label wie Biodiversitätsflächen, können sehr unterschiedliche Ausprägungen haben. Solche Flächen müssen entsprechend im Trainingsdatensatz genügend vorkommen.

Saisonale Abhängigkeiten von Nutzpflanzen erschweren die Aufgabe zusätzlich. So werden Sommer- und Wintergerste kaum in den gleichen Bilddaten zu erkennen sein. Allenfalls ist dem mit unterschiedlichen Modellen für unterschiedliche Saisons zu begegnen.

6 Projektrisiken

Wir sind uns noch nicht ganz im Klaren, ob die Aufgabenstellung für uns zu komplex ist aufgrund doch noch einiger fehlender technischer Grundlagen. Dies könnte dazu führen, dass wir uns in einigen Themen zu stark einlesen müssen oder uns in einem Thema verlieren.

Mit der Komplexität der Modelle steigen auch die Anforderungen an die Integration von bereits vortrainierten Architekturen wie z.B. dies der ETH Zürich. Des Weiteren müssen wir allenfalls auf externe Rechen-Ressourcen zugreifen können, um die Modelle in einem vertretbaren Zeitraum zu trainieren und evaluieren.

Es ist durchaus auch ein realistisches Szenario, dass wir am Ende des Semester kein besseres Modell finden, als die ETH Zürich. Trotz allem scheint es uns aber für den Aufbau von Kompetenzen im Team sinnvoll, diese Fragestellung zu bearbeiten.

7 Kompetenzen

Dani, hier ist uns nicht ganz klar, was erwartet wird. Vielleicht kannst du uns das erläutern.

Kompetenz	**Jan**	**Daniela**	**Thomas**
Domänenverständnis	1	1	1
Lösungsstrategien und -methoden	1	1	1
Engineering	1	1	1
Wissenschaftlichkeit und Reproduzierbarkeit	1	1	1
Projektmanagement	1	1	1
Transferkompetenz	1	1	1
Umgang mit Ideen und Problemen	1	1	1
Zusammenarbeitskompetenz	1	1	1
Reflexionsfähigkeit	1	1	1

Tabelle 2: Kompetenzeinteilung der einzelnen Personen