Formulating a Big Data Science Challenge for Land Imaging Time-Series Data

A Report of the National Geospatial Advisory Committee
Landsat Advisory Group
April 2021

Table of Contents

1.0 Introduction

In early 2020, the U.S. Geological Survey (USGS) requested that the Landsat Advisory Group (LAG), a subcommittee of the National Geospatial Advisory Committee (NGAC), provide input regarding the initiation of a Big Data Science Government Challenge to explore the benefits of computer vision and Machine Learning (ML), specifically Deep Neural Network (DNN)/ Convolutional Neural Network (CNN) methods for the purposes of exploiting Landsat Analysis Ready Data for time-series analysis and land change forecasting applications. These LAG high-level recommendations seek to provide support and encouragement for the development of an initial challenge to be conducted in 2021. Based on the outcomes of this initial challenge, the LAG encourages USGS to consider creating a series of follow-on challenges in 2022 and beyond in recognition of the 50th anniversary of the Landsat program.

2.0 Vision and Goals

Potential goals of a USGS Landsat government sponsored data challenge include the following:

- To expose the depth and breadth of the Landsat archive and encourage the development of novel and innovative new applications;
- To encourage the development of new ML/DNN-powered methods for the Land Change Monitoring, Assessment and Projection (LCMAP) program in order to operationalize them and improve program efficiency;
- To dramatically enhance the speed, accuracy, and capability of the USGS Land Imaging program;
- To support education in data science with ML and Earth Observation (EO) and to recruit new to talent to the EO profession;
- To engage data scientists globally;
- To support the growth of commercial markets for EO data;
- To bring citizen science to big data problems;
- To highlight cross-government cooperation in the fields of EO and ML; and
- To raise awareness of the utility of EO, and Landsat in particular, to provide actionable information for a wide range of government, commercial, and civil society users.

3.0 Background

Big data science challenges focused on spurring development of ML/DNN methods for image analysis have produced major advances in the fields of computer vision, artificial intelligence, and remote sensing, with great practical benefit to U.S. governmental and commercial stakeholders. Deep learning algorithms and technology have revolutionized automated image analysis, spurring billions of dollars of investment and producing significant technological advances for commercial applications and national security. However, training DNN algorithms

requires large amounts of labeled image samples, which has been a bottleneck in generalizing machine learning algorithms for the remote sensing field and broadening its application in the EO community. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) based on the ImageNet dataset of over 1 million human-labeled social media images (Deng, et al., 2009) was one of the early efforts that facilitated the rise of modern deep learning algorithms.

The SpaceNet challenges and datasets launched by In-Q-Tel and Maxar in August 2016, and now run by SpaceNet LLC, extended the ImageNet approach to high-resolution (<1m resolution) commercial satellite imagery. SpaceNet provides free, public access to over 67,000km2 of human-labeled imagery via Amazon AWS and has launched 7 major ML/DNN challenges, the most recent of which has focused on multi-temporal urban development (February 2021). Other satellite providers, including Planet, have adopted this approach to encourage algorithm development for their use cases, including a deforestation challenge hosted by Kaggle in 2016.

More recently, non-profit organizations have begun to create very large scale human-labeled datasets for medium-resolution public satellite imagery datasets focused on Copernicus Sentinel-2 10m imagery that can support creation of ML/DNN challenges. The International Institute for Applied Systems Analysis (IIASA) launched GeoWiki.org in 2009 to engage citizen scientists to produce geospatial training data for analysis of satellite imagery and scientific modeling. This has led to the creation of a 20 million human-labeled point dataset currently being used by the ESA World Cover consortium to produce a global 10m resolution land use land cover map from Copernicus Sentinel-2 imagery. Radiant Earth Foundation has produced a large dataset of 2000 tiles (13,000km^2) of human-labeled Sentinel-2 imagery called LandCoverNet (Alemohammad, et al., 2020). National Geographic Society/Impact Observatory has completed the creation of the Dynamic World 24,000 tile (624,000km^2) dataset of Sentinel-2 imagery (Brown, et al., 2021), containing over 5 billion human-labeled pixels specifically designed to support ML/DNN model creation. These Sentinel-2 based datasets with human labeling at 10m resolution are potentially adaptable to support development of ML/DNN algorithms for Landsat 30m resolution ARD.

4.0 LCMAP Initial Challenge, 2021: Problem Statement

The USGS charge to this working group is “to incentivize exploration into the utility and efficacy of ML/DNNs methods for purposes of exploiting Landsat ARD for time-series analysis and land change forecasting applications, and to augment those developed as part of the USGS LCMAP initiative.”

The LAG recommends that USGS craft a problem statement that directs competitors to mine the breadth and depth of the Landsat archive. With almost 5 decades of imagery available for analysis there is no better open repository of global coverage to analyze. This is a unique opportunity to attract and engage the world’s brightest data scientists. An appropriately scoped and promoted challenge could unlock new methods for analysis and most importantly identify new insights into our changing global landscape. The LAG recommends the challenge focus attention on key value propositions of the Landsat archive:

- **Moderate Resolution:** 15m pansharpened observations represent a significant spatial resolution improvement over standard 500m to 100m global monitoring systems (e.g., NASA MODIS, NOAA VIIRS, Copernicus CGLS-LC100). At this spatial resolution, ML/DNN model builders can explore a rich set of spatial-spectral features.
- **Long Time Series:** Extend back to the 1980s with Thematic Mapper sensor, and to the 1970s with the MSS. The world population has doubled since 1970, with enormous increases in demands on human infrastructure and food systems and impacts on the natural world.
- **Thermal Bands:** Landsat TIRS Bands 10 & 11 at 100m resolution provide a unique sensor modality used in the U.S. for monitoring water consumption and heat islands.
- **Spectral Calibration of Other Sensors for Persistent Observations:** The quality of Landsat observations enables the archive to provide a basis for spectral calibration and correction of other instruments, especially observations by newer, commercial small satellites. Consistent observations across satellite constellations enable global persistent monitoring of land use and land cover change processes.

The actual development of the recommended challenge statement is beyond the scope and capability of the LAG. However, but we believe that the Landsat Science Team and U.S. industry representatives could be well equipped to support further refinement of the problem statement. Additionally, several Federal employees have strong reputations for their work in designing challenges. The LAG identified the following individuals USGS might wish to consult as a challenge is designed:

- Jenn Gustetic, NASA Director of Early Stage Innovations and Partnerships
- Jarah Meador, Director of GSA's Challenge.gov
- Amy Kaminski, NASA's Prizes and Challenges Program Executive
- Lynn Buquo, Manager of NASA's Center of Excellence for Collaborative Innovation
- Kevin Murphy, NASA PE for Earth Science Data Systems
- Tsengdar J. Lee, NASA High Performance Computing
- Steven Babitch, AI leader for GSA's Technology Transformation Service

5.0 Challenge Management

The successful design and management of a robust Landsat data challenge will require a significant commitment of person-hours. The U.S. government has long encouraged the use of challenges by Federal agencies. Challenge.gov is a website with a wealth of information that USGS will find useful in designing and planning challenges. On the commercial side, Kaggle has run successful satellite imagery challenges for the space data industry and for governments, e.g., UK Defence Science and Technology Laboratory (Dstl) Kaggle competition (Kaggle, 2017).

The LAG recommends that USGS consider hiring an organization to participate in the design and to oversee the management of the challenge. There are numerous organizations – both commercial and not-for-profit – that specialize in designing and managing such challenges. Specifically, the management of the challenge will require planning for the following activities:

- Supporting ideation, problem statement identification, acquiring analysis, training and benchmark data, criteria, and performance measures;
- Managing timeline;
- Overseeing marketing and advertising of the challenge;
- Contracting with infrastructure hosting platform;
- Rule monitoring;
- Independent evaluation of code;
- Leader board management;
- Prize allocation; and
- Communicating and marketing the results of the challenge.

6.0 Multi-Agency and Commercial Partnerships

Given the growth in the availability of Earth Observation (EO) data and the rise of data science focused on the analysis of these data streams, the USGS could fashion this program as an ongoing partnership between many of the market segments of the remote sensing industry. Specifically, USGS could consider the inclusion of:

- NASA, NOAA, EPA, USDA/Foreign Agriculture Service, USAID, and other interested Federal agencies;
- International cooperation with entities such as the European Space Agency (ESA) and the Copernicus program, and other organizations including Japan's JAXA;
- Commercial U.S. space data suppliers such as Maxar, Planet, and Capella; and
- Training data partnerships with Radiant Earth Foundation, SpaceNet LLC, World Resources Institute, and National Geographic Society/Impact Observatory.

While the addition of these participants would add complexity to the process of design it would likely to lead to stronger results.

7.0 Challenge Data

In the government's instruction to the LAG there is a clear intention of including additional data streams in the challenge. The charge states "Methods are particularly encouraged to consider complementary Earth Observation (EO) data sets integrated with Landsat data", e.g., government and/or commercial multispectral (MSI), hyperspectral (HSI), synthetic aperture radar (SAR), thermal (LWIR/TIR), and lidar (e.g., NASA GEDI) remote sensing data collected from

air or space, non-image EO data, etc.). The government must identify the geographic area for which the challenge will be run and secure the data sets that will be made available for analysis in the challenge. The opportunity to open this challenge to a wide breadth of data sets both public and private, could provide an interesting opportunity to the USGS. Most if not all of the commercial EO data providers have opened data sets for similar challenges via SpaceNet and other open science data portals (e.g., Google Earth Engine). Once the problem statement and the geographic boundaries of the challenge have been defined USGS should reach out to the commercial companies to gauge their interest in participating by contributing data to the challenge.

Training and benchmark data are often the bottleneck in supporting robust data challenges. USGS recently released a large Landsat landcover training data set over the continental U.S. Additionally, Radiant Earth Foundation and National Geographic Society/Impact Observatory have landcover training data sets that could be fundamental to supporting USGS data challenges. Training data sets for the other data streams both open and commercial must be identified early in the competition design.

The LAG recommends that training datasets be constructed and made public that extend the state of the art, and should aim for a minimum of 100 million human-labeled pixels across at least 3000 locations. To demonstrate the full value of the Landsat archive, this training dataset should be distributed globally.

8.0 Target Participants and Teaming Rules

Based upon the desired goal(s) and problem statement, USGS will need to clearly define who are the target participants in this competition and conduct marketing and outreach to reach this audience to ensure broad participation. Additionally, USGS must determine if the competition is to be team-based or only based on individual effort. Given the global nature of EO and data science USGS should anticipate strong non-U.S. citizen participation. Given that USGS cannot award prize money to foreign participants, careful consideration should be given to how to manage this situation. One potential remedy is to require all teams to be led by U.S. individuals or organizations. However, as suggested earlier in this paper USGS may find it advantageous for many reasons to hire an organization to manage this competition. One advantage is that the managing entity may be able to raise and award prize money to the winners regardless of their citizenship. Additionally, a managing entity could also work to find philanthropic and/or corporate sponsorships and participation.

9.0 Rules and Evaluation Criteria

USGS must develop evaluation metrics that clearly define how each competitor's submission will be scored. These metrics must be quantitative and easily scored. Further, each high-ranking submission must be independently validated. As an example, SpaceNet has developed a number of metrics to support their individual challenges.

Rules for the competition may include topics such as:

- Multiple accounts per user, collaboration, or membership across multiple teams are not allowed.
- Code may not be shared privately. Any code that is shared must be made available to all competition participants through the platform.
- Solutions must use publicly available, open-source packages only, and all packages must be the most updated versions.
- Solutions must not infringe the rights of any third party and you must be legally entitled to assign ownership of all rights of open source of all winning solution code.
- Participants will be disqualified if they do not respond within the timeframe given in the request for code.

10.0 Follow-on Challenge, 2022

The LAG strongly encourages the USGS to consider creating a series of challenges to begin in 2022, the fiftieth anniversary of the Landsat program, based upon the results and lessons learned from the 2021 challenge. The LAG strongly encourages big thinking and swift action to envision a challenge program that that will focus on enhancing existing products as well as the development of new products and services generated from the analysis of Landsat and other observations from other government or commercial sources. We recommend a careful evaluation of the Copernicus Masters Challenge for a model of what to design and implement. We believe such a challenge program will enhance the utility and exposure of Landsat data to solve some of society's most persistent challenges such as climate change mitigation and adaptation as well as progress toward achieving the Sustainable Development Goals.

Other challenge themes that maybe of interest to USGS and the broader Landsat community include the following potential challenge topics:

10.1 Anomaly Detection

Data anomalies are values that deviate from long term mean or expected "normal" patterns. More specifically, anomalies in ARD can be defined by both spatial and temporal variability in reflectance data from long-term averages over a pixel in which the z-score (the number of standard deviations above or below the mean) outlier rejection test can be implemented for time-series products (Daszykowski et al., 2007). A fully data-driven (unsupervised) method to identify spatio-temporal anomalies in ARD data is an important first step to increase trustworthy of artificial intelligence (AI)-driven processing of Landsat data for various application. This challenge would encourage development of novel statistical machine learning approaches capable of capturing anomalies in data and identifying errors in source data (i.e., ARD) before they can be used for further analysis.

10.2 Automation and Augmentation of Landsat ARD Data (AAA)

Although AI/ML, particularly deep learning, has been overwhelmingly successful in many other fields, its use in remote sensing community has been limited due to the fact that deep learning

applications require large amounts of labeled training samples. For many remote sensing applications, collecting ground-truth data (labeled training samples) is often tedious, time consuming, and labor-intensive work, which is a current bottleneck for deep learning in remote sensing. Respondents to this challenge would be expected to develop automated procedures to create labeled ground truth data for remote sensing applications. The specific type of application would not be limited, and could include land-cover/land-use, crop type mapping, change detection, etc.

10.3 End-to-End Mapping (E2E)

Most of the of remote sensing studies that used deep learning have focused on feature engineering, meaning remote sensing imagery are converted to spectral and texture indices as input to deep learning architectures. There have been very few studies that used reflectance data as the direct input to a ML/DNN model. Participants in this challenge would be encouraged to develop E2E algorithms using time-series reflectance imagery as the direct input which produces the desired product in a fully automated fashion excluding manual process from the loop.

10.4 Data Fusion

Landsat ARD are best suited for medium resolution mapping, e.g., crop type mapping, extracting built up areas, or land-cover/land-use mapping. It is limited in its ability to map smaller man-made structures such as canals, trails and roads in mountainous areas and hydrological dams and buildings. With its long record of data, Landsat is an invaluable asset to document human geography across the world. This limitation can be overcome by fusing Landsat ARD data with other higher resolution imagery, e.g., WorldView/PlanetScope, as well as Synthetic Aperture Radar (SAR) imagery. Participants in this challenge would be expected to develop data fusion methods (pixel level, information fusion, or decision level fusion) and demonstrate the efficacy of the method in more than one application. Creative and challenging human geography applications like detecting trails and larger wildlife (e.g., elephants) in support of countering wildlife trafficking, and man-made construction with an estimated capacity (inhabitants) would be welcome.

11.0 Summary

In summary, the LAG strongly encourages the USGS conduct a LCMAP spectral and temporal resolution-focused data challenge in 2021 and to consult with the Landsat Science Team and Federal challenge experts to refine this focus. We suggest that USGS would be well served to hire an organization to facilitate the design and management of the challenge and further that USGS engage other Federal agencies to participate as sponsors of the activity.

Finally, based upon the results and the reception by the data science community to the first challenge, we encourage USGS to evaluate the development of a data challenge program to fully exploit the value of Landsat and other Earth observation data. We believe that USGS would find many productive partnerships with Federal agencies, the commercial sector, and

academia. We believe that the value of this effort can bring untold benefits to USGS as an organization as well as to society.

Acknowledgments

This paper was approved by the NGAC Landsat Advisory Group (LAG) on March 25, 2021 and was adopted by the NGAC as a whole on April 27, 2021. The LAG team developing this paper included Team Lead Anne Hale Miglarese (Saildrone), Frank Avila (National Geospatial-Intelligence Agency), Steven Brumby (Impact Observatory), Vasit Sagan (Saint Louis University), Robbie Schingler (Planet), and May Yuan (University of Texas – Dallas).

References

1. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.
2. Alemohammad S.H., Ballantyne A., Bromberg Gaber Y., Booth K., Nakanuku-Diggs L., & Miglarese A.H. (2020) "LandCoverNet: A Global Land Cover Classification Training Dataset", Version 1.0, Radiant MLHub. https://doi.org/10.34911/rdnt.d2ce8i
3. Christopher F. Brown, Steven P. Brumby, Brookie Guzder-Williams, et al., “Dynamic World, automated global 10m land use land cover maps”, Geo for Good, October 2020.
4. Kaggle, "Dstl Satellite Imagery Feature Detection", 2017. https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection.