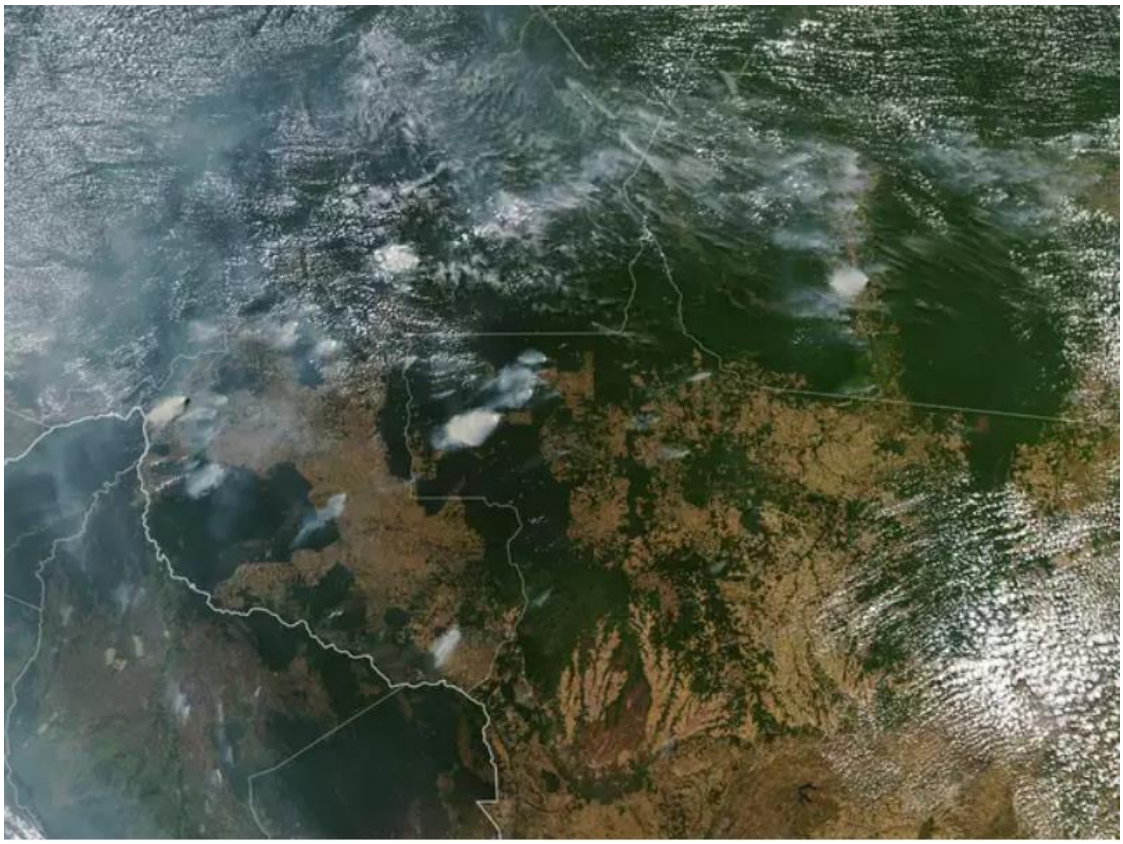
Springboard Capstone Project 2 – Project Proposal
Amazon Rainforest Satellite Image Classification



Connor McAnuff

October 9, 2019

1. Problem Description and Value to Client

1.1 Amazon Rainforest

The Amazon rainforest covers an area of 6,000,000 km^2 across multiple countries and is located in the largest river basin in the world. It contains several million species of insect, animal, plant, and tree life. Many areas of the Amazon also contain human civilization.

Despite efforts beginning in the 1990s by the Brazilian government and international bodies to protect the rainforest, human encroachment, exploitation, deforestation and other forms of destruction continue to harm the health of the Amazon Rainforest, causing reduced biodiversity, habitat loss, climate change, desertification, and soil erosion, among other issues [1][2].

1.2 Satellite Imagery

Due to the immense area of the Amazon Rainforest, monitoring deforestation is difficult, if not impossible to do so from the ground. Previously, research on tracking changes in forests has been performed using satellite imagery with a resolution of 30 m/pixel (Landsat) or even 250 m/pixel (MODIS). These resolutions are too coarse to identify small-scale deforestation/degradation or identify whether the cause is human or natural. The company Planet plans to collect daily satellite imagery of the Earth's surface at a resolution of 3-5 m/pixel, which would allow for manual (and perhaps machine-learning automated) classification of Amazon Rainforest satellite images, including the location and cause of small-scale deforestation/degradation [3].

1.3 Value to client

Automated classification of Amazon Rainforest satellite images would provide great value to governments and other organizations. The scale of the area and frequency of the images means that it would be extremely cumbersome and resource intensive for images to be manually classified.

Automated classification would ideally allow for the 'human footprint' in the Amazon to be accurately and frequently mapped to identify deforestation/degradation. The human footprint map can be used by governments or organizations to take action to prevent further harm. The data would also be valuable for other reasons – for example, it could be used for infrastructure, aid, census, and resource planning.

2. Dataset Description

2.1 Overview

The dataset to be used in for this project has been made available for a former Kaggle competition that occurred in 2017 [4]. The images were collected using Planet satellites and are given as chips (tiles). Each chip has been manually assigned a label (or multilabel) through crowd sourcing of labour. The client has stated that some mislabels are present, however this issue is common in satellite imagery classification.

The labels stem from three categories: atmospheric conditions, common land cover/use, and rare land cover/use. The labels are unbalanced, for example there are many more images of primary forest than there are images of selective logging. Images commonly have more than one label.

2.2 Dataset Format

Dataset available: https://www.kaggle.com/c/planet-understanding-the-amazon-from-space/data

1) 40,479 images (chips) of amazon rainforest satellite images
    - 256 x 256 pixels (947.2m x 947.2m on ground) in GeoTiff format (stripped of GeoTiff metadata)
    - 21.2 GB total
    - Four bands of data: red, green, blue, near infrared (16-bit)
    - 17 classifications (unbalanced)
        - More common:
            - Cloudy
            - Partly cloudy
            - Hazy
            - Primary rain forest
            - Water (rivers and lakes)
            - Habitation
            - Agriculture
            - Road
            - Cultivation
            - Bare ground
        - Less common:
            - Slash and burn
            - Selective logging
            - Blooming
            - Conventional mining
            - “Artisanal” mining
            - Blow down
2) List of image file names and their associated labels (train.csv)
    - CSV format

3. Proposed Solution Methodology

3.1 Data wrangling

- The CSV list of image file names and associated labels can be imported and stored as a Pandas DataFrame.
- Images will be imported for viewing and training as needed using imread.

3.2 Exploratory data analysis and data storytelling

- Frequency of each label.
- Co-occurrence of each label with other labels.
- Examples of images with each of the labels attached.
- Clustering of images.
- Outlier / erroneous image detection.
- Additional tasks to be determined as the data is explored.

3.3 Machine learning model overview

Target variable: Label/multilabel of image tiles.

Features: Image tiles.

Feature Engineering: Image processing strategies to be explored.

Model type: Multilabel classification.

Model Evaluation Criteria: F2 score.

The classification algorithm will very likely involve deep learning, as it is state-of-the-art for image classification.

4. Deliverables

- **Jupyter Notebooks:** Annotated code with figures and brief explanations for the data processing, exploratory data analysis, machine learning implementation, etc.
- **Technical Report:** In depth explanation and discussion of dataset, problem solving methodology, results, etc.
- **Slide deck:** High level overview of the project with voiceover.

5. References

[1] Encyclopaedia Britannica, "Amazon Rainforest," 27 08 2019. [Online]. Available: https://www.britannica.com/place/Amazon-Rainforest. [Accessed 2019].

[2] Pachamama Alliance, "Effects of Deforestation," 2019. [Online]. Available: https://www.pachamama.org/effects-of-deforestation. [Accessed 2019].

[3] Planet, "Understanding the Amazon from Space," 2017. [Online]. Available: https://www.kaggle.com/c/planet-understanding-the-amazon-from-space/overview/description. [Accessed 2019].

[4] Planet, "Understanding the Amazon from Space - Data," 2017. [Online]. Available: https://www.kaggle.com/c/planet-understanding-the-amazon-from-space/data. [Accessed 2019].