Clasificación de cobertura terrestre en la Amazonía usando imágenes satelitales

Cristian Daniel Muñoz Botero
Ana Isabel Patiño Osorio
Jonathan Andrés Granda Orrego
Universidad de Antioquia

1. Contexto de Aplicación

La selva amazónica es uno de los ecosistemas más biodiversos del planeta, pero también uno de los más amenazados debido a la deforestación, la expansión agrícola y el cambio climático.

Este proyecto se desarrolla en el marco de una competencia de Kaggle organizada por Planet Labs, una empresa que proporciona imágenes satelitales diarias de la Tierra [1]. El objetivo es utilizar datos satelitales para clasificar automáticamente la cobertura terrestre y las condiciones atmosféricas del Amazonas, asignando etiquetas a distintas clases y situaciones ambientales.

2. Objetivo de Machine Learning

El problema es de **clasificación multietiqueta**: dado un conjunto de imágenes satelitales RGB, el objetivo es predecir una o más etiquetas por imagen, que pueden incluir:

- *Cobertura terrestre:* `agriculture`, `artisinal_mine`, `bare_ground`, `blow_down`, `conventional_mine`, `cultivation`, `habitation`, `primary`, `road`, `selective_logging`, `slash_burn`, `water`
- *Condiciones atmosféricas:* `blooming`, `clear`, `cloudy`, `haze`, `partly_cloudy`

Cada imagen puede contener múltiples etiquetas simultáneamente, reflejando la complejidad del entorno amazónico y la coexistencia de diferentes coberturas y condiciones atmosféricas en una misma escena.

3. Dataset

Fuente: Planet Labs (conjunto de datos proporcionado originalmente en la competencia de Kaggle [1]). Para este trabajo se utiliza un dataset que fue publicado posteriormente como un dataset independiente [2], y que adecua al conjunto de datos original, dado que los archivos originales en formato `.torrent` resultaban demasiado pesados para la capacidad de cómputo disponible, y los protocolos de descarga generan conflictos en entornos en la nube.

Tipo de datos: Imágenes satelitales RGB de 256×256 píxeles.

Tamaño del dataset:

- 40,479 imágenes en el conjunto de entrenamiento, y 40,669 en el conjunto de prueba (estas últimas sin etiquetas).
- Tamaño en disco: aproximadamente 3 GB comprimido.

Formato: Imágenes en `.jpg` + archivo `.csv` con etiquetas.

Distribución de clases: La distribución de clases es desbalanceada. Algunas etiquetas como `primary` y `clear` son mucho más comunes que otras como `slash_burn` o `blow_down`, estas primeras representando alrededor del 80-85 % del conjunto total de imagenes.

4. Métricas de Desempeño

Métrica de ML: F2 Score (promedio macro).
La métrica principal utilizada para evaluar el modelo será el F2 Score, que asigna un mayor peso al recall que a la precisión. Esto resulta útil en contextos donde es más importante detectar todos los casos relevantes (minimizar falsos negativos) que evitar falsos positivos.

$$F_2 = (1 + \beta^2) \cdot \frac{(\text{Precision} \cdot \text{Recall})}{(\beta^2 \cdot \text{Precision}) + \text{Recall}}$$

$$\text{Precision (precisión)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall (sensibilidad)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

El promedio macro (macro-averaged F2) se calcula promediando el F2 score de cada clase, tratándolas de forma equitativa pese al desequilibrio de datos.

Métricas de negocio: Dado que el proyecto tiene un enfoque principalmente técnico, podríamos hablar en cuanto a métricas de negocio que son importantes para la aplicación de modelos, esta se relacionan con la eficiencia en el procesamiento de imágenes satélitales. Un modelo más preciso y rápido podría incrementar el **Throughput** del sistema, es decir, el número de imágenes procesadas automáticamente por unidad de tiempo, permitiendo una detección más temprana de eventos como deforestación, incendios o minería ilegal, por lo que seria bueno contemplar un modelo que tenga menos complejidad computacional en la emisión de predicciones.

$$\text{Throughput} = \frac{\text{Imágenes procesadas}}{\text{Unidad de tiempo}}$$

5. Estado del arte

Diversos estudios han abordado la clasificación multietiqueta en imágenes satelitales aplicadas al monitoreo ambiental y la cobertura terrestre. Singh y Shankar (2022) propusieron un enfoque basado en modelos preentrenados y técnicas de *data augmentation*, junto con funciones de pérdida adaptadas al desbalance de clases, logrando mejoras en tareas de clasificación sobre la Amazonía [4].

Por su parte, Ji et al. (2020) introdujeron una arquitectura híbrida que combina redes convolucionales (CNN) con redes recurrentes (LSTM) y mecanismos de atención para capturar dependencias semánticas entre etiquetas, mejorando la detección de clases raras en entornos naturales complejos [5].

Estas metodologías sirven como base conceptual para el presente trabajo, que busca aplicar principios similares de transferencia de aprendizaje y optimización de métricas sensibles al *recall*, como el F2 Score, en el contexto específico de la clasificación de cobertura terrestre amazónica.

6. Referencias

Referencias

[1] Planet Labs. (2017). *Planet: Understanding the Amazon from Space* [Competencia de Kaggle]. Recuperado de https://www.kaggle.com/competitions/planet-understanding-the-amazon-from-space/overview

[2] Rom, N. (2017). *Planets Dataset* [Conjunto de datos en Kaggle]. Recuperado de https://www.kaggle.com/datasets/nikitarom/planets-dataset/data

[3] Planet Labs Inc. (2017). *Planet imagery and data products*. Disponible en https://www.planet.com/

[4] Singh, A.K. & Shankar, B.U. (2022). *Multi-Label Classification on Remote-Sensing Images*. arXiv:2201.01971.

[5] Ji, J., Jing, W., Chen, G., Lin, J., & Song, H. (2020). *Multi-Label Remote Sensing Image Classification with Latent Semantic Dependencies*. Remote Sensing, 12(7), 1110.