Instituto Tecnológico y de Estudios Superiores de Monterrey

Avance 2. Ingeniería de características

Ingeniería de Características Visuales: De Máscaras a Texto con Segmentación Semántica

Juan Ricardo Albarracín Barbosa
Luis Ángel Oporto Añacato
David Alexis García Espinosa
Dr.Gerardo Jesús Camacho González

Proyecto Integrador MNA-V 2025

11 de mayo de 2025

1. Introducción

La ingeniería de características en el contexto de visión por computadora se refiere al proceso de transformar datos visuales, como imágenes o máscaras, en representaciones estructuradas que puedan ser utilizadas por modelos posteriores. En este proyecto, se trabaja con segmentaciones semánticas derivadas del dataset PASTIS24, generando descripciones en lenguaje natural a partir de las clases presentes en cada parche de imagen.

Este enfoque permite vincular los resultados visuales del modelo de segmentación con salidas comprensibles por humanos, utilizando técnicas de mapeo entre clases, conteo, distribución espacial y reglas heurísticas. De esta manera, se facilita la generación de texto que describe la composición agrícola de una región observada por satélite.

2. Objetivos

Objetivo general: Transformar máscaras de segmentación semántica en descripciones textuales mediante ingeniería de características visuales.

Objetivos específicos:

- Extraer características estadísticas y espaciales relevantes de las máscaras.
- Establecer una representación intermedia interpretable entre la máscara y el texto.
- Automatizar la generación de descripciones en lenguaje natural a partir de reglas basadas en dominio.

3. Transformación de Imágenes en Patch Embeddings para su Consumo por Transformers

Los modelos Transformer, originalmente diseñados para tareas de procesamiento de lenguaje natural (NLP), han demostrado un rendimiento sobresaliente en visión por computadora al ser adaptados mediante el concepto de *patch embeddings*, como se propone en los Vision Transformers (ViTs) [2]. Sin embargo, debido a que los Transformers operan sobre secuencias de vectores de igual longitud, es necesario transformar las imágenes, que son estructuras matriciales bidimensionales (o tridimensionales en el caso multiespectral), en una representación compatible con este tipo de arquitectura.

3.1. Proceso de Generación de Patch Embeddings

Dado un conjunto de imágenes satelitales de tamaño $H \times W \times C$ (alto, ancho y número de canales), el primer paso consiste en dividir cada imagen en parches (*patches*) no superpuestos de tamaño fijo $P \times P$ píxeles. Esta segmentación produce un total de:

$$N = \frac{H}{P} \times \frac{W}{P}$$

patches por imagen, cada uno de los cuales se aplana para formar un vector de dimensión $D = P \times P \times C$.

Posteriormente, cada vector se proyecta a un espacio de menor dimensión mediante una capa lineal:

$$\text{Embedding} = \text{Linear}(D \to d_{\text{model}})$$

donde d_{model} es la dimensión del espacio latente del Transformer (por ejemplo, 512 o 768). A cada embedding resultante se le suma una codificación posicional que representa la ubicación espacial del patch dentro de la imagen original. Esto es necesario porque los Transformers no tienen una noción inherente de orden o estructura espacial.

3.2. Extensión a Series de Tiempo Satelitales (SITS)

En aplicaciones como las series de tiempo satelitales, donde las imágenes están distribuidas a lo largo del tiempo, este proceso se extiende a una secuencia temporal. Las imágenes se organizan como un tensor de cuatro dimensiones ($T \times H \times W \times C$), donde T representa la dimensión temporal. Cada imagen se divide en patches y se embebe siguiendo el procedimiento anterior, resultando en una secuencia de embeddings de tamaño:

$$T \times N = T \times \left(\frac{H}{P} \times \frac{W}{P} \right)$$

Finalmente, esta secuencia de vectores se alimenta al Transformer, el cual procesa tanto la dimensión espacial como la temporal mediante mecanismos de atención, permitiendo modelar relaciones complejas entre regiones de una misma imagen y entre observaciones a lo largo del tiempo.

3.3. Importancia del Proceso de Patch Embedding

La transformación de imágenes en secuencias de patch embeddings es el paso esencial que permite adaptar los modelos Transformer, originalmente secuenciales, a tareas de visión computacional. Sin esta etapa, no sería posible representar las imágenes en un formato compatible con la arquitectura del Transformer, ya que sus entradas deben tener forma de secuencia con dimensión fija.

Además, este proceso tiene ventajas clave:

- Permite aprovechar el paralelismo de los Transformers en el dominio visual.
- Reduce la dimensionalidad de entrada, haciendo el entrenamiento más eficiente.
- Hace posible la reutilización de arquitecturas preentrenadas en NLP.
- Facilita la incorporación de señales temporales y espaciales de manera explícita, lo cual es crítico en tareas con imágenes multitemporales como las de SITS.

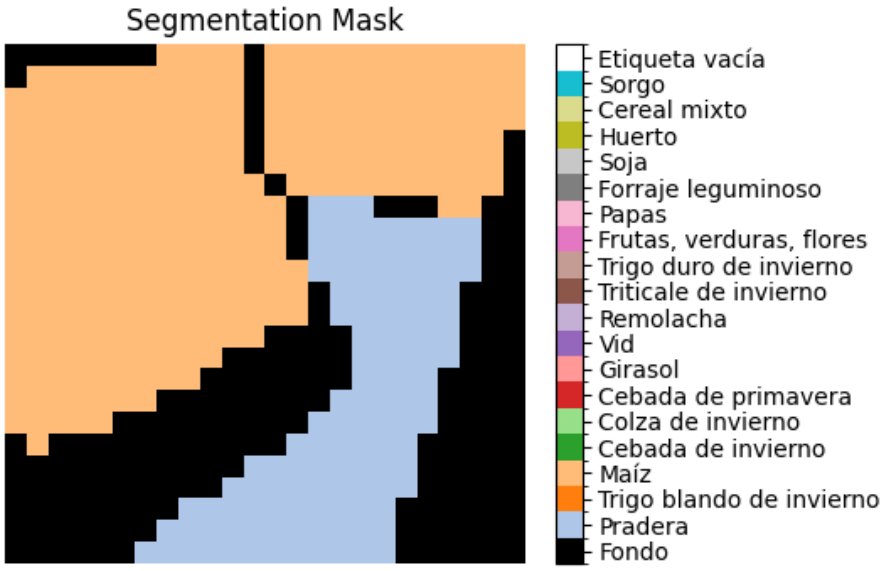

Figura 1. Image post procesamiento, notamos como se consigue la segmentación y se logra etiquetar a que segmento de cultivo pertenecen.

4. Ingeniería de Características Visuales

4.1. Extracción de características de máscaras

Se parte de una colección de máscaras de segmentación generadas a partir del dataset PASTIS24 [1]. Cada máscara contiene valores enteros representando clases agrícolas (por ejemplo, maíz, girasol, vid, etc.). A partir de estas, se aplicaron las siguientes transformaciones:

- Cálculo de frecuencia de clases por máscara (conteo de píxeles).
- Cálculo de proporción relativa por clase (porcentaje sobre el total del parche).
- Identificación de clases dominantes (por encima de un umbral fijo).

4.2. Transformaciones semánticas

A partir de las representaciones anteriores, se propuso una fase de transformación heurística que convierte información visual en texto comprensible. Esta fase incluye:

- Traducción de identificadores numéricos a etiquetas en español (e.g., 3 → Maíz).
- Reglas de redacción del tipo: “La parcela contiene principalmente X, seguida de Y y Z.”
- Priorización de cultivos con mayor superficie y omisión de etiquetas de fondo o ruido.

4.3. Codificación y normalización

Aunque no se trabaja con variables numéricas tradicionales, se aplica una codificación implícita al transformar las máscaras a vectores de frecuencias, normalizados en escala [0,1], para facilitar la representación semántica posterior.

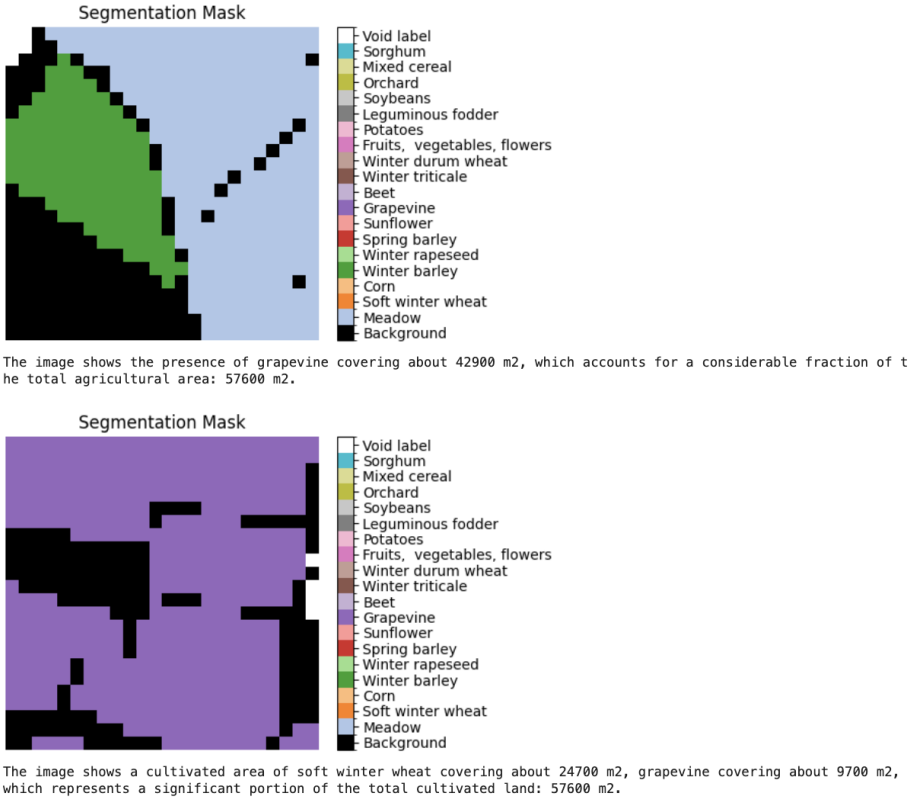

Figura 2. Ejemplo de resultados de la transformación de la imagen a un concepto.

4.4. Selección de características

Dado que no todas las clases están presentes en cada muestra, se filtraron aquellas categorías con baja frecuencia global. También se aplicaron umbrales mínimos de cobertura para considerar clases relevantes en la descripción textual, evitando ruido en las salidas.

5. Transformaciones espectrales en desarrollo

Como parte de los trabajos en curso, se está explorando la incorporación de transformaciones espectrales derivadas de las bandas multiespectrales disponibles en Sentinel-2, específicamente mediante el cálculo del **Índice de Vegetación de Diferencia Normalizada (NDVI)**. Esta métrica es ampliamente utilizada en agricultura para monitorear la salud de la vegetación.

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1}$$

En imágenes Sentinel-2, la banda del infrarrojo cercano (NIR) corresponde a la banda 10 y la banda roja (Red) a la banda 4. El índice se calcula para cada pixel y puede

agregarse como canal adicional o usarse como base para descripciones temporales de cobertura vegetal.

A continuación, se muestra un ejemplo de cómo evoluciona el NDVI a lo largo del tiempo para un parche específico:

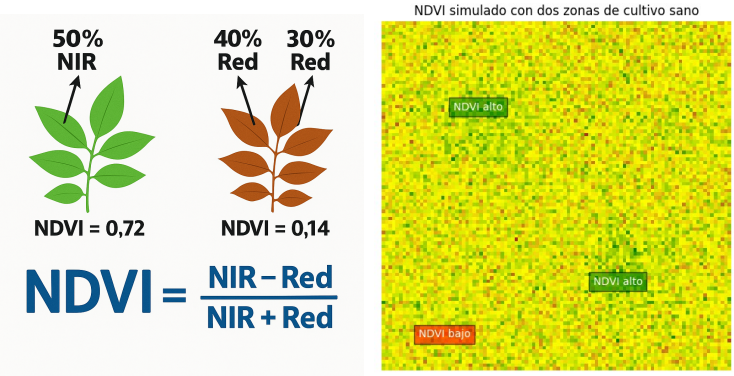

Figura 3. Evolución del NDVI en una ubicación específica.

Este tipo de representación permite capturar el comportamiento fenológico de los cultivos. A futuro se prevé integrar el NDVI como entrada adicional para enriquecer las descripciones generadas, o como parte de un conjunto de reglas basado en dinámicas de crecimiento.

5.1. Repositorio de Referencia y Experimentos Relacionados

Los conceptos descritos en esta sección se han comenzado a implementar y explorar en el repositorio de código disponible en GitHub. En particular, se han desarrollado dos líneas experimentales clave:

- **Conversión de máscaras a texto**: Se encuentra en desarrollo un método preliminar para interpretar regiones segmentadas en imágenes satelitales, generando descripciones textuales de las máscaras mediante técnicas de análisis semántico y estructural. Este enfoque puede ser consultado en el siguiente cuaderno: `https://github.com/juanrtato/crop-image-deepanalysis/blob/channels_analysis/notebooks/mask2text.ipynb`
- **Experimentación con NDVI**: En paralelo, se ha iniciado la evaluación del índice de vegetación diferencial normalizado (NDVI) como fuente para enriquecer las representaciones multiespectrales y mejorar la separación semántica en entornos agrícolas. Esta línea de trabajo se puede revisar en el cuaderno: `https://github.com/juanrtato/crop-image-deepanalysis/blob/channels_analysis/notebooks/channels_ndvi_experimentation.ipynb`

Estos experimentos constituyen los primeros pasos para integrar capacidades de extracción semántica y análisis espectral dentro de una arquitectura basada en Vision Transformers, con el objetivo de mejorar la interpretación automatizada de imágenes satelitales en escenarios agrícolas y de monitoreo ambiental.

6. Conclusión

En esta etapa del proyecto se ha establecido un puente entre datos visuales y texto, mediante un proceso estructurado de ingeniería de características centrado en máscaras de segmentación semántica. Las representaciones extraídas permiten generar descripciones que reflejan con precisión la distribución de cultivos en una parcela. Este trabajo se alinea con la fase de "Preparación de los datos" de la metodología CRISP-ML(Q) [3], al sentar las bases para futuros modelos generativos de lenguaje más sofisticados, como BLIP o LLaVA.

Referencias

[1] S. Garnot, B. Webb, N. Le Roux, L. Chevallier and F. Tardy, *Panoptic segmentation of satellite image time series with convolutional temporal attention networks*, in *CVPR Workshops*, 2021.

[2] Tarasiou, M., Chavez, E., & Zafeiriou, S. (2023). ViTs for SITS: Vision Transformers for Satellite Image Time Series. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10418–10428.

[3] L. Visengeriyeva, A. Kammer, I. Bär, A. Kniesz, M. Plöd, *CRISP-ML(Q): The ML Lifecycle Process*, INNOQ, 2023. Disponible en: `https://ml-ops.org/content/crisp-ml`