Instituto Tecnológico y de Estudios
Superiores de Monterrey

# *Propuesta de Proyecto*

**Fusión de Visión Computacional y Modelos de Lenguaje para la Transferencia de Conocimiento en Agricultura de Precisión**

**Juan Ricardo Albarracín Barbosa**
**Luis Ángel Oporto Añacato**
**David Alexis García Espinosa**

**Dra. Grettel Barceló Alonso**
**Dr. Luis Eduardo Falcón Morales**
**Mtra. Verónica Sandra Guzmán de Valle**

**Dr. Gerardo Jesús Camacho González**

Proyecto Integrador

27 de abril de 2025

# 1. Introducción

La inteligencia artificial ha impulsado una transformación en el monitoreo agrícola mediante el análisis de imágenes capturadas por satélites y drones. Sin embargo, el análisis de cultivos y la adaptabilidad de estos sistemas a diferentes entornos geográficos y climáticos sigue siendo un desafío, principalmente por la falta de datos homogéneos y las dificultades en transferir modelos preentrenados entre regiones.

Este contexto resalta la necesidad de desarrollar modelos capaces de generalizar mejor a variaciones espaciales y temporales en las condiciones de observación. Además, la naturaleza dinámica de los cultivos, combinada con la variabilidad en las adquisiciones de imágenes (debido a factores como cobertura de nubes, ángulos de adquisición o diferencias fenológicas), requiere métodos que no solo sean robustos, sino también capaces de capturar de manera explícita las dependencias temporales y espaciales en los datos.

En respuesta a estos desafíos, el presente proyecto propone una solución basada en un pipeline multimodal que traduce las salidas visuales de modelos de segmentación semántica, en entradas comprensibles para modelos de lenguaje de gran escala (LLMs). De esta manera, se busca realizar un análisis efectivo de los cultivos y facilitar la transferencia de conocimiento entre diversas regiones agrícolas.

# 2. Antecedentes

## 2.1. Contexto

El monitoreo de actividades humanas sobre la superficie terrestre, como la agricultura, es fundamental para diseñar intervenciones que aumenten el bienestar y la resiliencia de las sociedades. La observación del desarrollo de cultivos permite optimizar estrategias orientadas a mejorar los ingresos de los agricultores y fortalecer los sistemas de producción alimentaria [1], además de controlar prácticas ambientalmente responsables, como la adopción de técnicas amigables con el medio ambiente y la diversificación de cultivos, que son esenciales para garantizar una agricultura sostenible y respetuosa con el entorno.

En la actualidad, el aumento notable de la cantidad de datos de Observación de la Tierra (EO) recolectados desde el espacio [2], las herramientas

disponibles para su procesamiento y los avances en técnicas de visión por computadora y aprendizaje automático, representa una oportunidad clave para desarrollar soluciones que permitan monitorear de forma automática el crecimiento de los cultivos.

## 2.2. Estado del arte

Modelos de segmentación de imágenes, como U-Net [3] y DeepLabV3+ [4], han sido ampliamente utilizados en aplicaciones agrícolas, como la detección de estrés hídrico, la identificación de plagas y la delimitación de parcelas.

Por otro lado, modelos multimodales como CLIP [5] han demostrado la capacidad de mapear imágenes y texto en un espacio semántico común. Recientes investigaciones han explorado esta convergencia en el sector agrícola, como el trabajo de Lu et al. (2022) sobre CropFormer, y el modelo AgriBERT, que adapta representaciones lingüísticas al ámbito agroalimentario.

Además, los avances en el campo de Visual Question Answering (VQA) y el desarrollo de modelos como BLIP-2 [6] han abierto nuevas perspectivas metodológicas para integrar de manera efectiva visión y lenguaje.

Finalmente, Michail Tarasiou et al., en su artículo ViTs for SITS: Vision Transformers for Satellite Image Time Series [1], establecen una base sólida para la implementación de Vision Transformers en series temporales de imágenes satelitales, lo cual resulta fundamental para este proyecto.

# 3. Entendimiento del negocio

## 3.1. Formulación del problema

Dado que en la actualidad los modelos de segmentación semántica han alcanzado una alta precisión, además del creciente impulso en el uso de modelos generativos aplicados al lenguaje natural y visión artificial, surge la inquietud de aplicar estas nuevas tecnologías en el contexto de la producción agrícola.

La problemática abordada consiste en analizar, describir y cuantificar los cultivos sembrados a partir de una serie de imágenes satelitales. Para ello, las imágenes deben ser segmentadas según el tipo de cultivo, para luego generar una respuesta final en lenguaje natural.

## 3.2. Objetivos

**Objetivo general:** Diseñar e implementar un sistema multimodal basado en un pipeline que integre visión computacional y modelos de lenguaje, par realizar un análisis efectivo de los cultivos y facilitar la transferencia de conocimiento entre diversas regiones, contextos agrícolas y condiciones específicas.

**Objetivos específicos:**

- Desarrollar una arquitectura que combine modelos de segmentación de imágenes con LLMs, mediante embeddings y técnicas avanzadas de Deep Learning, para mejorar la comprensión semántica y la representación de información agrícola.
- Implementar una metodología de entrenamiento robusta y escalable que permita la generalización del conocimiento agrícola en distintos contextos regionales, considerando variaciones climáticas, geográficas y culturales que afectan el desarrollo de los cultivos.
- Evaluar el rendimiento del sistema en cuanto a precisión semántica, transferibilidad interregional y eficiencia computacional, utilizando métricas cuantitativas y cualitativas para validar su aplicabilidad y eficiencia en la monitorización y análisis agrícola.

## 3.3. Preguntas clave

- ¿Qué tipos de cultivos pueden ser diferenciados de manera efectiva mediante segmentación semántica, a partir de imágenes satelitales disponibles?
- ¿Qué resolución y calidad de imagen son necesarias para lograr una segmentación y clasificación efectiva?
- ¿Qué nivel de precisión es aceptable para las estimaciones de superficie y producción?
- ¿Cómo varían los calendarios agrícolas por región y tipo de cultivo, y cómo puede modelarse esta variabilidad?
- ¿Cómo se validará el modelo? ¿Qué métricas de desempeño se usarán?

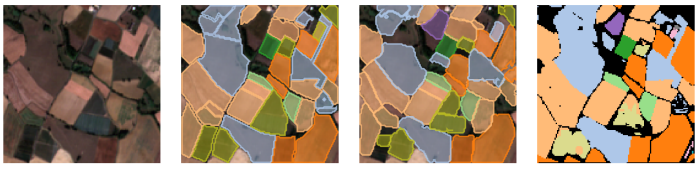
## 3.4. Involucrados

- Asesor Investigador
  - Definición de requerimientos.
  - Retroalimentación del proceso de modelamiento y validación de resultados.
- Equipo de Ciencia de Datos
  - Integración de fuentes de datos.
  - Limpieza, normalización, almacenamiento de datasets.
  - Preprocesamiento e ingeniería de datos.
- Equipo de modelamiento en Inteligencia Artificial/Machine Learning
  - Definición de la arquitectura multimodal.
  - Entrenamiento, ajuste y evaluación del modelo.

# 4. Entendimiento de los datos

## 4.1. Descripción de los datos

Los datos provienen de series temporales de imágenes satelitales (SITS) usadas para el monitoreo de la superficie terrestre, especialmente para la identificación de cultivos agrícolas. Concretamente, se plantea usar el dataset PASTIS (Patch-based Agricultural Semantic Time Series) [7], uno de los utilizados para el entrenamiento de la arquitectura TSViT propuesta en [1]. PASTIS contiene series temporales de imágenes satelitales Sentinel-2 enfocadas en regiones agrícolas en Francia. Cada muestra corresponde a un pequeño parche espacial (24×24 píxeles) acompañado de información de clases de cultivos.



(a) Image from the sequence. (b) Panoptic annotation. (c) Panoptic segmentation. (d) Semantic segmentation.

Figura 1. Segmentación en conjunto de datos PASTIS. [7]

Las imágenes capturan variaciones temporales a lo largo del ciclo de crecimiento de las plantas y contemplan distintas clases de cultivos agrícolas, así como clases de fondo y etiquetas de vacío. A continuación, se enumeran las clases consideradas:

- 0: Background (Fondo)
- 1: Meadow (Pradera)
- 2: Soft winter wheat (Trigo blando de invierno)
- 3: Corn (Maíz)
- 4: Winter barley (Cebada de invierno)
- 5: Winter rapeseed (Colza de invierno)
- 6: Spring barley (Cebada de primavera)
- 7: Sunflower (Girasol)
- 8: Grapevine (Vid)
- 9: Beet (Remolacha)
- 10: Winter triticale (Triticale de invierno)
- 11: Winter durum wheat (Trigo duro de invierno)
- 12: Fruits, vegetables, flowers (Frutas, verduras, flores)
- 13: Potatoes (Papas)
- 14: Leguminous fodder (Forraje de leguminosas)
- 15: Soybeans (Soya)
- 16: Orchard (Huerto frutal)
- 17: Mixed cereal (Cereal mixto)
- 18: Sorghum (Sorgo)
- 19: Void label (Etiqueta vacía)

Cada muestra está representada como un tensor con la siguiente estructura:

- Número de muestras.
- Número de adquisiciones a lo largo del tiempo.
- Dimensiones espaciales (píxeles por adquisición).
- Número de canales correspondientes a diferentes longitudes de onda captadas por el sensor satelital.

Estos datos capturan la variabilidad espectral, espacial y temporal necesaria para segmentar y caracterizar los cultivos en las imágenes.

La salida del modelo de segmentación es un mapa de predicción de clases, donde cada píxel recibe una etiqueta correspondiente a un tipo de cultivo y está representada con la siguiente estructura:

- Número de muestras.
- Número de clases posibles.
- Dimensiones espaciales de la predicción.

Cada predicción representa la distribución de probabilidades sobre las clases para cada píxel, de la cual se extrae la clase más probable. Esta representación sería utilizada como entrada para un LLM, aún por definir, con el objetivo de generar descripciones en lenguaje natural que resuman las características de cada parcela.

## 4.2. Arquitectura y Modelamiento

En la fase de modelamiento, se hará uso de la arquitectura TSViT (Temporo-Spatial Vision Transformer) para la segmentación semántica de las imágenes satelitales, basados en la arquitectura ViT (Vision Transformer). El modelo utilizará como datos de entrada series temporales de imágenes satelitales (SITS). Posteriormente, para la generación de descripciones en lenguaje natural, se contempla la incorporación de un modelo de lenguaje de gran escala

(LLM), el cual procesará los resultados obtenidos de la segmentación semántica.

Se considera la posibilidad de realizar un fine-tuning del LLM utilizando las representaciones de las predicciónes del modelo de segmentación y ejemplos específicos del dominio agrícola para mejorar la calidad y relevancia de las respuestas generadas.

**Nota:** Cabe resaltar que, dado el carácter exploratorio e investigativo del proyecto, las arquitecturas específicas, algoritmos y técnicas de aprendizaje automático a utilizar seguirán un enfoque profundo-supervisado. Sin embargo, podrián ajustarse o evolucionar durante el desarrollo del trabajo, en función de los hallazgos, necesidades específicas del problema y resultados obtenidos en etapas intermedias.

# Referencias


[1] M. Tarasiou, E. Chavez y S. Zafeiriou, «Vits for sits: Vision transformers for satellite image time series,» en *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, págs. 10 418-10 428.

[2] M. Tarasiou y S. Zafeiriou, «Deepsatdata: Building large scale datasets of satellite images for training machine learning models,» en *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2022, págs. 4070-4073.

[3] O. Ronneberger, P. Fischer y T. Brox, «U-Net: Convolutional Networks for Biomedical Image Segmentation,» *MICCAI*, 2015.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy y A. L. Yuille, «DeepLabv3+: A Unified Approach for Semantic Image Segmentation,» *CVPR*, 2018.

[5] A. Radford, J. W. Kim, C. Hallacy, A. El-Nouby y et al., «Learning Transferable Visual Models From Natural Language Supervision,» *ICML*, 2021.

[6] J. Li, X. He, L. Yu y et al., «BLIP-2: Bootstrapping Language-Image Pretraining with Frozen Image Encoders and Large Language Models,» *arXiv preprint arXiv:2301.12597*, 2023.

[7] V. Sainte Fare Garnot y L. Landrieu, «Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks,» *ICCV*, 2021.