Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Avance 3. Baseline

Análisis Comparativo de Modelos Multimodales para Interpretación de Máscaras con LLM

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

Proyecto Integrador MNA-V 2025

18 de mayo de 2025

Introducción

La interpretación automática de máscaras de segmentación semántica mediante modelos multimodales y grandes modelos de lenguaje (LLM) representa un avance significativo en la vinculación de datos visuales con representaciones textuales comprensibles. Esta capacidad es fundamental para transformar imágenes segmentadas, como las del dataset PASTIS24, en descripciones precisas y humanas sobre la composición de cultivos en zonas agrícolas.

Los modelos multimodales recientes ofrecen capacidades integradas para procesar texto e imágenes simultáneamente, facilitando el entendimiento contextual y la generación de texto explicativo. Este documento compara los principales modelos desarrollados entre 2023 y 2025, evaluando su idoneidad para responder a la tarea específica de convertir máscaras de segmentación en texto mediante LLM.

1. Comparativa de modelos multimodales recientes

2. Discusión orientada a la tarea de interpretación de máscaras con LLM

Para el objetivo específico de transformar máscaras de segmentación semántica en descripciones textuales, es fundamental que el modelo multimodal cuente con:

- **Capacidad para integrar información visual estructurada**, tal como las regiones segmentadas, con contexto semántico.
- **Fuerte comprensión del lenguaje natural** para generar textos claros, coherentes y específicos a partir de los datos visuales.
- **Facilidad para manejar datos multiespectrales o extendidos** (e.g., NDVI u otras bandas satelitales), que enriquecen la interpretación.
- **Flexibilidad para instrucciones y tareas específicas**, ya que la generación de texto depende de reglas heurísticas y datos de dominio agrícola.

En este sentido, modelos open source como *BLIP-2*, *LLaVA* y *InstructBLIP* destacan por su balance entre eficiencia, capacidad de razonamiento visual y naturalidad en generación textual. *MiniGPT-4* también presenta una arquitectura ligera útil para prototipos y despliegues con recursos limitados.

Modelos comerciales como *GPT-4*, *Kosmos-2* y *PaLM-E* ofrecen mayor robustez y desempeño, particularmente en tareas complejas o con múltiples modalidades, pero su acceso restringido limita la experimentación y ajuste fino para dominios especializados como la agricultura satelital.

La incorporación de técnicas como la extracción de *patch embeddings* para las máscaras y la integración de métricas espectrales como NDVI pueden complementar el input visual, mejorando la precisión semántica de la descripción generada por el LLM.

3. Conclusiones

Para el proyecto de ingeniería de características visuales orientado a la interpretación textual de máscaras semánticas satelitales, se recomienda priorizar modelos multimodales open source como *BLIP-2*, *LLaVA* e *InstructBLIP*, por su capacidad avanzada de razonamiento visual-textual y soporte para instrucciones. Estos modelos facilitan la integración de datos específicos del dominio y permiten generar descripciones precisas y ricas en contexto. Los modelos comerciales pueden ofrecer mejoras de rendimiento, pero es necesario evaluar cuidadosamente sus costos y limitaciones de acceso. La combinación de datos segmentados con indicadores espectrales como NDVI y el uso de reglas heurísticas continúan siendo elementos esenciales para lograr precisión y utilidad en el sistema final. Actualmente, se está explorando la conexión de embeddings derivados de las representaciones visuales con el modelo *GPT-2*, buscando validar su eficacia para interpretar y generar textos a partir de información visual estructurada, apuntando a emplear modelos más livianos en tareas específicas y optimizando recursos sin sacrificar calidad.

Se contempla evaluar *Missisipi* de H2O.ai, una solución enfocada en la integración y análisis avanzado de datos multimodales que podría ofrecer ventajas en escalabilidad y despliegue en entornos empresariales, facilitando la adopción práctica del sistema. La estrategia combina modelos open source con enfoques híbridos que integran datos espectrales y heurísticos, junto con experimentos en conexión con modelos de lenguaje más ligeros, abriendo una ruta flexible y eficiente para la interpretación textual de datos visuales satelitales.

Referencias

- BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models
- OpenFlamingo GitHub Repository
- MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models
- LLaVA: Large Language and Vision Assistant
- mPLUG-2: Modular Prompting and Large-Scale Vision-Language Pre-training
- InstructBLIP: Towards Universal Image-Language Understanding and Generation
- PaLM-E: An Embodied Multimodal Language Model
- OpenAI GPT-4 Official Announcement
- Microsoft Kosmos-2 Blog
- Microsoft Florence 2.0 Blog

Modelo	Año	Parámetros	Modalidades	Arquitectura / Base	Casos de uso	Open Source	Organización
GPT-2	2019	1.5B	Texto	Transformer autoregresivo	Generación texto	Sí	OpenAI
BLIP-2	2023	∼1B	Texto + Imagen	Transformer + ViT	Captioning, Visual QA	Sí	Salesforce Research
OpenFlamingo	2023	∼3B	Texto + Imagen	Flamingo + LLM	Few-shot multimodal, diálogo	Sí	OpenAI + Comunidad
MiniGPT-4	2023	Desconocido	Texto + Imagen	LLM + ViT	Chat multimodal	Sí	OpenAI + Comunidad
LLaVA	2023	13B	Texto + Imagen	LLaMA + ViT	Diálogo, análisis visual	Sí	UCLA + Comunidad
mPLUG-2	2023	>1B	Texto + Imagen	Transformer multimodal	Razonamiento visual	Sí	Microsoft Research
InstructBLIP	2023	Desconocido	Texto + Imagen	Derivado de BLIP-2	Tareas instruccionales	Sí	Salesforce Research
Falcon-40B	2023	40B	Texto	Transformer autoregresivo	NLP general	Sí	Technology Innovation Institute
Mistral-7B	2023	7B	Texto	Transformer autoregresivo	NLP general	Sí	Mistral AI
Stable Diffusion	2022	∼1B	Texto + Imagen	Latent Diffusion Model	Generación imágenes	Sí	Stability AI
CLIP	2021	400M	Texto + Imagen	Transformer	Emparejamiento texto-imagen	Sí	OpenAI
OpenCLIP	2021	Varía	Texto + Imagen	Open implementation of CLIP	Emparejamiento texto-imagen	Sí	OpenCLIP Community

Tabla 2. Comparativa horizontal de modelos multimodales open source recientes (2022–2025).

Modelo	**Año**	**Parámetros**	**Modalidades**	**Arquitectura / Base**	**Casos de uso**	**Costo Aproximado**	**Open Source**	**Organización**
GPT-4	2023	Desconocido	Texto + Imagen	Transformer multimodal	Chat, generación avanzada	$0.03 - $0.12 por 1K tokens	No	OpenAI
Claude	2023	Desconocido	Texto	Transformer	Diálogo seguro y ético	$0.02 - $0.06 por 1K tokens	No	Anthropic
DeepSeek	2024	Desconocido	Texto / Búsqueda	Modelo propietario	Búsqueda semántica avanzada	Variable, desde $1000/mes+	No	DeepSeek Inc.
PaLM	2022	540B+	Texto	Transformer grande	Multitarea, generación texto	No público (Google Cloud Pricing)	No	Google
Bard	2023	Desconocido	Texto	Basado en PaLM	Chatbot conversacional	Gratuito / Google API costo variable	No	Google
Flamingo	2022	Desconocido	Texto + Imagen	Transformer multimodal	Few-shot multimodal	No público	No	DeepMind
DALL · E 2	2022	Desconocido	Texto + Imagen	Transformer	Generación de imágenes	$0.016 por imagen 1024x1024	No	OpenAI
ChatGPT Enterprise	2023	Basado en GPT-4	Texto + Imagen	Transformer multimodal	Empresas, seguridad y soporte	Desde $20 / usuario/mes	No	OpenAI
Jasper AI	2021	Basado en GPT-3	Texto	NLP generación contenido	Marketing, contenido	Desde $29/mes	No	Jasper Labs

Tabla 3. Comparativa horizontal de modelos multimodales closed source con costos aproximados (2022–2025).