Image Analysis for AI-Powered Post Creation

Basic Image Analysis Techniques

Image analysis encompasses several key techniques:

1. **Image Classification**: Identifies what a whole image represents (e.g., "This is a restaurant interior")
2. **Object Detection**: Locates and identifies multiple objects within a single image (e.g., "There's a steak, a glass of wine, and a salad")
3. **Semantic Segmentation**: Classifies each pixel in an image to a specific class (e.g., precisely outlining the boundaries of all objects)
4. **Instance Segmentation**: Combines object detection and semantic segmentation to identify separate instances of objects (e.g., distinguishing between multiple steaks on a table)

For our project, **object detection** is indeed the most appropriate approach, as we need to identify multiple food items, restaurant decor elements, and other objects that would be relevant for caption generation.

Role of Image Analysis in Our Project

Image analysis serves a critical function in the AI-powered post creation workflow:

1. The system analyzes uploaded images to detect key objects
2. These detected objects are presented to the user
3. Users can select relevant items to focus on while removing irrelevant or incorrectly detected objects
4. The selected objects provide context for generating accurate, relevant captions that match the image content
5. This improves the quality of AI-generated captions by ensuring they reference what's actually in the image

This approach gives users control over what aspects of the image to emphasize in captions while leveraging AI to identify content automatically.

Implementation Options

Table 1: Implementation Options Comparison

Option	Approach	Pros	Cons	Feasibility

Option	Approach	Pros	Cons	Feasibility
1	Run pre-trained models (YOLO) locally on developer machine	• Zero hosting costs • Simplest initial setup • Complete privacy • No internet dependency • Good for proof-of-concept	• Only runs on developer's machine • Team members can't test it • Requires GPU for good performance • Not usable in production	Low - Limited to single developer testing
2	Deploy entire backend with pre-trained models (YOLO) on cloud servers (AWS EC2, Google Colab)	• Complete backend solution • Integrated with all other features • Production-ready environment • Centralized deployment	• Highest cost (full server) • Most complex setup • GPU costs even when not using ML • Requires server management • Overkill for small project	Low - Excessive overhead for image analysis alone
3	Self-host pre-trained models (YOLO) as API on cloud servers (AWS EC2, Google Colab)	• Dedicated ML microservice • Only pay for ML processing • Separates ML from main backend • Reusable across projects • Can optimize server for ML	• Still requires separate GPU server • Deployment complexity • API design and implementation needed • Time-consuming for a small feature	Medium - Requires significant initial setup

Option	Approach	Pros	Cons	Feasibility
4	Use free third-party image analysis APIs	• No setup required • Free to use • Simple API integration	• Poor detection quality (see Table 2)	Low - Inadequate quality (tested)
5	Use paid commercial vision APIs (AWS Rekognition, Google Vision API)	• Best accuracy for food items • No server setup or maintenance • Highly reliable services	• Usage costs (~$1-2 per 1,000 images) • API key management • Data privacy concerns	High - Best balance of effort vs quality

Table 2: Free API Test Results

Image	Expected Detection	Cloudmersive Results
	Steak, Red wine	cup
	Pasta, Shrimp, Garlic, Herbs, Clam, White wine	wine glass, chair
	Burger	sandwich

Image	Expected Detection	Cloudmersive Results
	Pancake, Berries, Whipped Cream	cake, bowl

Conclusions

Option 5 (Paid Commercial APIs) offers the best balance between implementation effort and quality outcomes. Given the scope of our project, the associated costs would be minimal while delivering the necessary food detection capabilities for generating high-quality captions.

Further research on AI Vision APIs

API Service	Pricing	Free Tier	Food Detection Capabilities	Other Benefits
AWS Rekognition	$1 per 1,000 images	5,000 images/month (first 12 months)	Restaurant-specific object detection with detailed labels	Robust food and ingredient recognition; Extensive documentation
Google Cloud Vision	$1.50 per 1,000 images	1,000 units/month	Strong food recognition with ingredient breakdown	Food label detection; Integration with other Google services
BLIP (Replicate)	~$0.73 per 1,000 images	~100 calls per day	Combined object detection and caption generation	Generates descriptive captions; Single API call for both detection and description

Next steps

1. Research and compare sample responses from AWS Rekognition, Google Vision API, and BLIP for restaurant food images
2. Set up developer accounts and obtain API keys for the selected service
3. Create a simple API wrapper in our backend to handle image upload and analysis
4. Implement the user interface for displaying and selecting detected objects before caption generation
5. Test the integration with actual restaurant food images to verify detection quality