More Than a Thousand Words

Teaching AI to See for the Visually Impaired

A PICTURE is worth a thousand words, but only if you can see it. For the 253 million people around the world living with visual impairments, the digital world, a world increasingly built on images, can feel like a book with most of its pages torn out.

Standard artificial intelligence can offer a caption, but it's often frustratingly literal. An image of a triumphant marathon finish line might be described as "a group of people on a street." A cherished family photo from a holiday becomes "a man, a woman, and a child." Technically correct, yes. But it misses the point entirely. It captures the objects but loses the story, the emotion, and the details that give an image its meaning.

For someone who is blind or has low vision, a vague caption isn't just unhelpful; it's an obstacle. It widens the digital divide, making it harder to participate in social media, understand a news article, or even shop online. This project, which is my master's research, began with a simple belief: we can do better. We don't just need AI that can label images; we need AI that can *describe* them with the richness and control that a sighted person takes for granted.

The Problem with a Single Description

The core challenge is that a single, one-size-fits-all caption is never enough. Sometimes, a user needs a quick summary of an image. Other times, they might want to know specific details: What is the expression on a person's face? What text is written on a sign in the background?

Current AI captioning systems typically look at an image and generate a single sentence. Our project takes a fundamentally different approach. We decided that to describe an image properly, an AI first needs to understand it like a detective – by breaking down the scene into its core components.

Building a Blueprint of the Scene

Before our system writes a single word, it first creates a "scene graph": a structured blueprint of everything happening in the image. It doesn't just see a "bird"; it identifies the bird, its attributes ("small," "blue"), and its relationship to other objects ("flying over the water"). This is a crucial step. By a converting the visual chaos of pixels into an organized map of objects and relationships, we give the AI a much deeper, more structured understanding of the scene.

We built on the foundation of powerful models like OpenAI's CLIP and GPT-2, but our innovation lies in how we fuse their capabilities with this detailed scene graph. The result is a system that doesn't just guess a plausible sentence; it constructs a description from a foundation of understood facts.

The Real Innovation: Captions on Demand

This deeper understanding unlocks the most important feature of our work: *hierarchical, controllable captions*. Instead of a single, static description, our system can generate layers of detail, giving the user complete control over how much information they receive.

Imagine an image of a kitchen scene. A user could navigate the description like this:

- **Level 1: The Gist.** "A woman is standing in a kitchen."
- **Level 2: Adding Detail.** "A smiling woman with a red apron is standing in a modern kitchen, holding a wooden spoon."
- **Level 3: The Full Picture.** "A smiling woman with a red apron is standing in a modern kitchen, holding a wooden spoon over a steaming pot on the stove. Sunlight is streaming through a window on the left."

This isn't just a longer caption; it's an interactive experience. The user decides how deep they want to go. This simple shift from a static label to a dynamic, controllable description is a profound change for assistive technology. It empowers the user, giving them the agency to explore a visual world on their own terms.

Looking Ahead: A More Accessible World

This project is a step toward a future where AI serves as a true visual interpreter. The goal is to create technology that offers not just access, but understanding. By moving beyond vague labels and toward rich, controllable descriptions, we can build AI that helps bridge the information gap for millions. The work continues, but the vision is clear: to harness the power of AI not just to see the world, but to share its stories with everyone.