

Motivation

In recent years, autonomous driving technology has made tremendous progress, yet interpreting complex urban driving scenes remains a major challenge. The dynamic interactions of vehicles, pedestrians, cyclists, and infrastructure, especially in dense cityscapes, complicate decision-making processes for autonomous systems. While current models excel at visual recognition, they often lack interpretability, which is crucial for ensuring safety and reliability in real-world driving scenarios. The integration of Natural Language Processing (NLP) with visual data presents an opportunity to improve the explainability of these systems. By generating descriptive captions for driving scenes, we aim to enhance model interpretability and enable better understanding of critical scenarios, ultimately improving safety in autonomous driving.

Method

We employed two datasets for this research: **Cityscapes** and **GTA5**. The **Cityscapes dataset** consists of 5,000 images of urban driving scenes captured from 50 cities across Germany, offering a variety of real-world traffic situations. The **GTA5 dataset** provides 24,966 synthetic images with per-pixel semantic annotations, simulating hazardous and rare driving environments.

For automatic caption generation, we utilized Google's **Gemini 1.5 Flash** model, a multimodal large language model (LLM) that interprets images and generates scene descriptions. Each image was paired with the prompt: *"Describe the image according to traffic data in 2-3 sentences,"* focusing on key elements like vehicle positions, traffic signals, and pedestrians.

To improve accuracy, the machine-generated captions were manually reviewed and corrected. We observed common errors in recognizing object positions, traffic signal colors, and distinguishing between left and right. Human corrections provided more contextually precise and interpretable descriptions, enhancing scene understanding and overall system performance.

Experiments

Caption Generation:

- Machine-generated captions described the number of vehicles, their positions, and traffic light colors.
- Captions were produced for both GTA5 and Cityscapes datasets.

Challenges with Machine-Generated Captions:

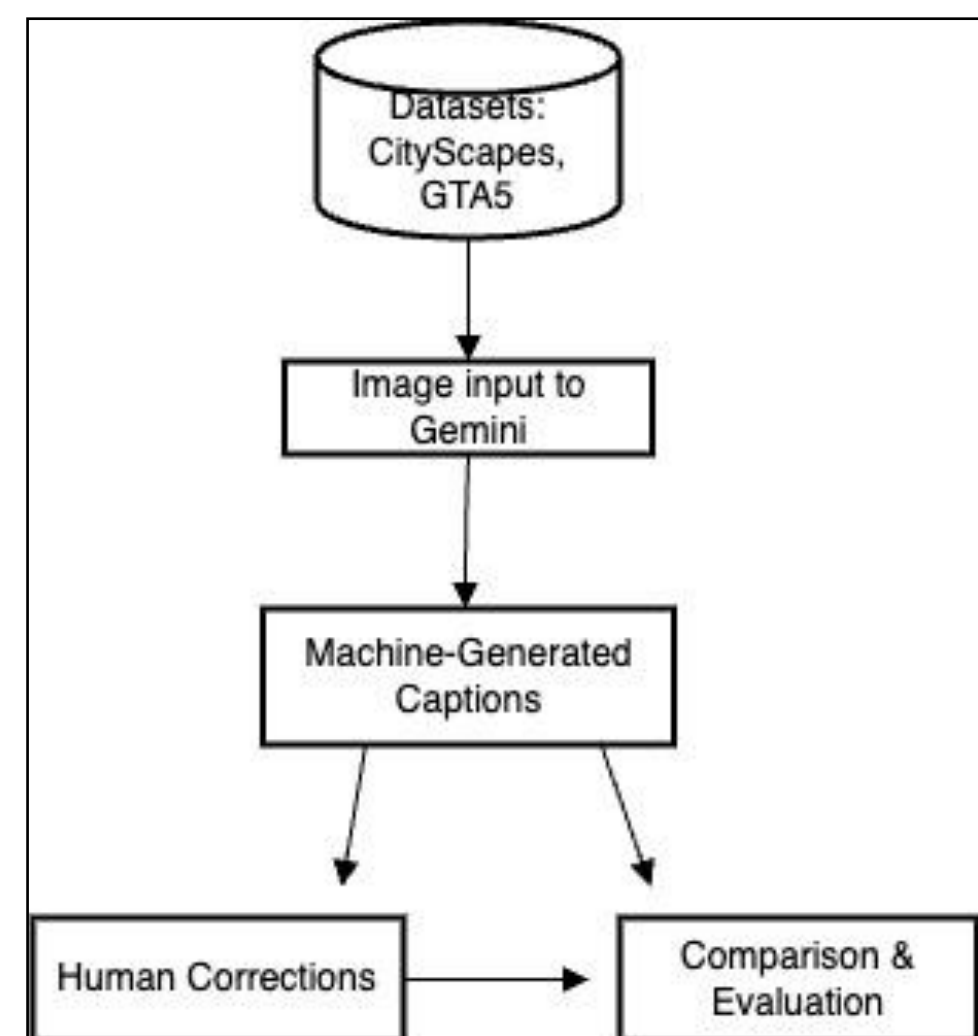
- Struggled with distinguishing left from right, leading to incorrect vehicle positions.
- Inconsistent recognition of traffic signal lights (e.g., green vs. red).
- Failed to correctly identify objects at the edges or distant parts of the image.

Human-Corrected Captions:

- Accurate descriptions of vehicle and pedestrian positions.
- Clearer sentence structures, leading to improved interpretability.
- More precise and contextually accurate, overall improvement of scenes.

Manual Review and Corrections:

- Machine-generated captions were manually inspected for accuracy.
- Corrections were made to fix errors in vehicle positioning, traffic flow direction, and color recognition.



Results

Question: Describe the image according to traffic data in 2-3 sentences.

Machine Generated Caption VS Human Corrected Caption

GTA



Machine Answer: A white car is driving down a street with a single lane in each direction. A black car is on the left side of the road. A red car is on the right, just past the lane markings, and is not visible in the photo. There is a line of parked cars on the right side of the road and a line of parked cars across the street from them. The white car is driving down the left lane.

Human Answer: A white car is driving in the right lane of a two-way street, with a single lane in each direction. Several cars are parked along both sides of the road, with three on the left side and two on the right. The road is otherwise clear of moving traffic, and there is a fire hydrant visible on the left side.

CityScapes



Machine Answer: The image shows a street with a bus driving down the center lane. There are cars parked on the right side of the street. The bus is driving towards the left side of the image, so the road is likely one way. The white lines indicate the lane divisions. The car in the foreground is also driving down the center lane.

Human Answer: The image shows a street with two buses driving down the right lane toward the stoplight. Seven cars are parked on the right side of the street, and one car is parked on the left side. The buses are driving towards the right side of the image, indicating that the road is likely a multi-lane road with clearly marked lanes for traffic in both directions. The white lines indicate the lane divisions, and the car in the foreground is also driving down the center lane.

Conclusions

- Machine-generated captions were prone to errors in interpreting complex urban driving scenes. Human corrections dramatically improved caption accuracy and scene interpretability.
- Integrating NLP with visual data improves the interpretation of complex urban driving scenes. While machine-generated captions struggled with accuracy, human corrections significantly enhanced scene comprehension. Combining real-world and synthetic data bolstered model robustness, particularly in rare or critical driving scenarios. This approach improves the safety and decision-making of autonomous vehicles in urban environments.

References

Richter, Stephan R., et al. "Playing for data: Ground truth from computer games." Cordts, Marius, et al. "The cityscapes dataset for semantic urban scene understanding."