Exploring Scene Understanding with Large Language Models (LLM)

Introduction

This report documents the comparison of different machine learning models that were designed for image description. The primary goal of this was to see how accurately and contextual these models can describe images of different environments from around the world. The models I have chosen are GIT Large COCO, BLIP, Kosmos-2, and ViT-GPT2. Each model's strengths and weaknesses are assessed based on its ability to describe the details of the scene, identify the environment and provide context on the image. This research has potential to improve the scene understanding of autonomous cars and make it better at recognizing things in their ways.

Methodology

Models Used

Here are the 4 image-to-text machine learning models that I have used for this research.

- GIT Large COCO: It is a transformers model that uses COCO dataset. It generates text descriptions from images. It is chosen because of it is training with the COCO dataset
- BLIP: A model optimized for image-text alignment tasks. It is chosen because it is strong on generating captions for straightforward scenarios.
- ViT-GPT2: A vision transformer-based model connected with GPT2 for text generation. It is selected because of its usage of GPT2.
- Kosmos-2: A multimodal AI system specializing in grounded scene analysis with detailed descriptions. It is chosen because of its multi model capabilities.

Dataset

In total 648 Google Street View images were used in this research, these images were chosen because each image is taken from different places and different conditions. These images had different types of variables inside them such as:

Daytime: Morning, afternoon, evening, and night scenes.

Weather: Clear skies, rain, snow, and fog.

Environment: Urban areas, highways, rural roads, and intersections.

Subjects: Cars, pedestrians, and other environmental elements.

These images were chosen to challenge models' ability to adapt and describe capabilities. All of the images were captioned either by me or Chat GPT since there were 648 images, Chat GPT was utilized.

Evaluation

BLEU and ROUGE were used to evaluate the models. BLEU compares the human written output to the machine output, scoring is based on how close the machine output is to the human written descriptions. It uses bigrams to calculate the score; for example, if one out of four bigrams matches, the score would be 0.25. ROUGE works similarly by comparing the machine output to the human output. It has three key metrics: ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 and ROUGE-2 evaluate the overlap of unigrams (single words) and bigrams (pairs of consecutive words), respectively, between the human and machine outputs, while ROUGE-L measures the Longest Common Subsequence (LCS), focusing on sequences of words that appear in the same order in both texts. The scoring for ROGUE-1, ROUGE-2, and ROUGE-L ranges from 0 to 1, with scores closer to 1 indicating a better overlap between the texts. Similarly, BLEU scores are calculated on a scale of 0 to 1, where higher values indicate better alignment between the machine and human outputs.values indicating better alignment between the machine and human outputs.

Experiment Setup

Hardware: Evaluations were conducted on a system equipped with a NVIDIA RTX 2070 GPU, ensuring efficient model inference.

Software: Python 3.12.3 with the following libraries:

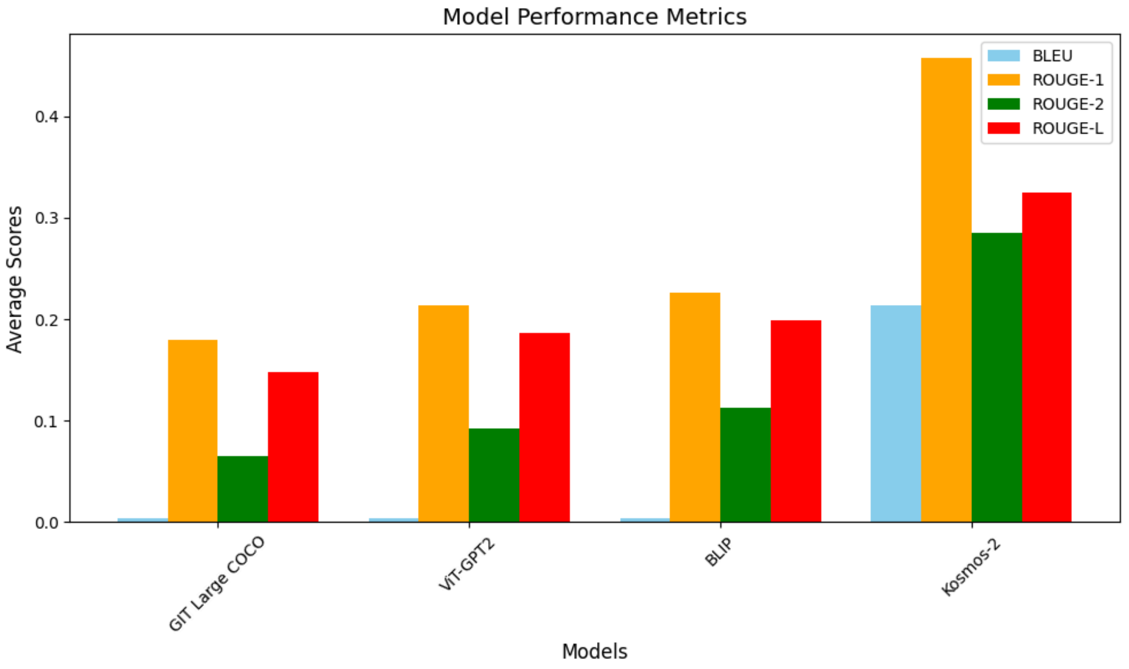
- ➔ transformers
- ➔ torch
- ➔ nltk
- ➔ Rouge_score
- ➔ Pillow

Procedure

Images were fed into the model and generated captions were compared to human written references in a json file. The comparison was done by BLEU and ROUGE scores. All of the generated captions were recorded with the scores next to each model in a text file.

Results

Model	BLEU AVERAGE SCORE	ROGUE-1 AVERAGE SCORE	ROGUE-2 AVERAGE SCORE	ROGUE-L AVERAGE SCORE
GIT Large COCO	0.0036	0.1799	0.0649	0.1478
VIT-GPT2	0.0035	0.2137	0.0927	0.1861
BLIP	0.0036	0.2259	0.1130	0.1986
Kosmos-2	0.2139	0.4581	0.2849	0.3247



FINDINGS

BLEU, which measures overlapping words between the generated output and the human output, showed very low performance, almost 0 at 0.00036 on average, between GIT Large COCO, ViT-GPT2, and BLIP. However Kosmos-2 achieved a higher BLEU score of 0.2139 outperforming other models by a lot. This tells us that Kosmos-2 generated captions that are more related with the human generated captions.

ROUGE-1 measures the similarity between unigrams on the human generated text and model generated text. The scores were way better compared to the BLEU test.
GIT LARGE-COCO scored the lowest at 0.1799, this shows that the least amount of words overlapped with each other. VIT-GPT2 scored 0.2137 and BLIP scored 0.2259 which is very close to each other which reflects that they had almost the same amount of overlapping words. On the other hand Kosmos-2 again scored the highest with 0.4581 more than the double compared to others.

ROUGE-2 measures bigram overlap, gives more insight into coherence of the captions. GIT Large COCO got the lowest score of 0.0649 this shows its difficulty in producing coherent phrases. ViT-GPT2 and BLIP were similar again with scores of 0.0927 and 0.1130, respectively. Kosmos again led the pack with a higher score of 0.2849.

ROUGE-L measures the sentence-level frequency between texts. GIT Large COCO still was the last with a score of 0.1478. VIT-GPT2 and BLIP scored 0.1861 and 0.1986 respectively; this shows moderate coherence between sentences. As always Kosmos-2 the highest score with a 0.3247 still outperforming other models.

Overall Observations

GIT Large COCO: It was the dead last on every evaluation even though if we look at some of the output it gave there were some that was pretty detailed such as “a street view of a building with a closed garage door in the middle of it” but most of the captions were in the lines of something like this “view of the house from the street” which lacked a lot of detail

VIT GPT2 and BLIP: These models almost performed the same they both gave sometime good details but most of the time missed a lot of stuff in the images

Kosmos-2: This model dominated across all metrics and received the highest scores compared to others. Which is fair since the descriptions it gave had a lot of details about the whole image rather than just one part of the image such as; The image shows a narrow street with houses on either side. There are two cars parked on the street, one on the left side and the other on the right side. In addition to the cars, there is a truck parked further down the street. The street appears to be located in a residential area, and there are a few people walking on the sidewalk. There is also a bench located near the center of the scene, providing a place for people to sit and enjoy the surroundings. This detailed explanation was never reached with the other models.

FUTURE DIRECTIONS

The results for the GIT Large COCO model was calculated differently since 3 out of 4 captions it generated was “this is a google street view image”. To solve this a filter has been applied so that only one valid caption is used for each scoring in the future a different set of images can be utilized so that GIT Large COCO can function better. Since many of the evaluations required a human typed caption to evaluate the images automating this process with an very accurate Image Captioner would save up a lot of time in the future

Since evaluations were conducted on a NVIDIA RTX 2070 GPU which made going through all the images very long. Utilizing Google Cloud or AWS could significantly lessen the time it takes to process all the images. Furthermore Advanced API’s and Model can be utilized such as Claude AI and Chat GPT to make the captioning and the generating captions through models better and more accurate to the images since these models get updated continuously.

In the future research an AI model can be trained just for captioning images from dash cam cameras or Google Street View.

CONCLUSION

This research analyzed the performances of four different image-to-text models, which were GIT Large COCO, ViT-GPT2, BLIP, and Kosmos-2, regarding their performance on tasks on scene understanding and image captioning. In the research, the mentioned models were compared by generating different captions for various images gathered from Google Street View and limits and performances of each model were compared. Kosmos-2 gave the best results among other models, with the highest scores in BLEU and ROUGE metrics. Its multimodal architecture showed much better capabilities in generating descriptive and detailed captions, making it more suitable for advanced scene understanding tasks. In contrast, other models, such as GIT Large COCO, suffered from repetitive or generic outputs, again underlining the importance of dataset alignment and fine-tuning.

This assessment, in the end, shows the huge potential of multimodal AI systems in those fields that really require precise image captioning, like autonomous cars. While Kosmos-2 evaluations were high, there is still a lot of room for improvement on different models that were used. Further research should be based on better datasets, with powerful computers and specialized model exploration. Coupled with these, AI-driven scene understanding will continue to improve and pave the way for safer and more effective technologies in the years to come.

CITATIONS

Wikipedia contributors. "ROUGE (metric)." Wikipedia, The Free Encyclopedia, 28 Nov. 2023. Wikipedia, The Free Encyclopedia. Web. 28 Nov. 2024.

Wikipedia contributors. "BLEU." Wikipedia, The Free Encyclopedia, 16 Sep. 2024. Wikipedia, The Free Encyclopedia. Web. 28 Nov. 2024.

Wang, Jianfeng, et al. "GIT: A Generative Image-to-text Transformer for Vision and Language." ArXiv preprint, arXiv:2205.14100, 2022. Web.

Li, Junnan, et al. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation." Proceedings of the International Conference on Machine Learning (ICML), 2022. Web.

Peng, Zhiliang, et al. "Kosmos-2: Grounding Multimodal Large Language Models to the World." ArXiv preprint, vol. abs/2306, 2023. Web.

Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." ArXiv preprint, 2023. Web.

ydshieh. "ViT-GPT2 Image Captioning Model." Hugging Face Models Repository, NLPConnect, 2024, https://huggingface.co/nlpconnect/vit-gpt2-image-captioning. Accessed 2 Dec. 2024.