DOI: 10.54254/2755-2721/109/2024.17892

Using LLM Model to Process Sensor-detected Images

Shunqi Wang

University of Toronto, Toronto, Ontario, M5R 0A3, Canada

wang2669003240@gmail.com

Abstract. Modern digital Image processing was developed within the 1960s at Massachusetts Institute Of Technology. Image processing is still a popular topic in recent technology development as countless industries need to process the image in an efficient and economical way. In this paper, this paper will first briefly introduce the features of the large language model (LLM). Then, a discussion between the LLM and the sensor will be set up, introducing the feasibility that LLM can process the image effectively. In the next section, this study conducts an experiment on a LLM-based image-recognizing program to indicate language models are indeed competent in describing what happens in the image. The whole experiment flow will be go through detailedly and some experiment data is going to be presented. Lastly, a conclusion to the experiment will be made, including the usages of the program in the experiment, and the possible improvement of the experiment and the program in future.

Keywords: Large language model, sensor-detected images, image processing.

1. Introduction

A multitude of experts contend that large language models (LLMs) have excelled in diverse applications, hence enhancing their popularity in both academia and industry. A credential poll regarding the LLM further underscores its significance. Both academia and the corporate sector have demonstrated significant interest in large language models (LLMs). Previous research indicates that LLMs exhibit exceptional performance, suggesting they may ultimately supplant artificial general intelligence (AGI) in contemporary contexts [1]. In recent years, individuals have exerted considerable effort to analyze LLMs from multiple perspectives. Although LLM is robust and extensive, it is not infallible. An essay that examines the present limitations of LLMs and contends that more complex duties pose significant challenges for LLMs to manage. The critical interpretation of information is a primary challenge faced by current models, especially with ever complex tasks or signals. This issue may result in biases and mistakes being intertwined with erroneous data, potentially leading users to reach incorrect conclusions—particularly if they rely excessively on these models (i.e., automation bias) [2]. Consequently, to enhance the performance of LLM, it is essential to focus on instances where LLM yields poor outputs and to consider how prompts influence these results.

Moreover, the selection of language models is also crucial. It is normal to observe two LLMs executing essentially identical tasks, however they may vary significantly in size. Although a larger model enhances performance, it does not inherently utilize time or resources efficiently. Recent advancements in Transformer-based large language models (LLMs) have resulted in substantial efficiency improvements across several occupations. These benefits necessitate a substantial

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

augmentation in model size, potentially leading to protracted and costly inference times. In reality, the generations produced by LLMs consist of varying levels of complexity. Some predictions significantly benefit from utilizing the full capabilities of all models, while other continuations are more straightforward and can be managed with reduced computational resources [3]. While it is alluring to utilize the more sophisticated LLM, the additional time required by the forward-looking LLM will detract from the efficiency and effectiveness of the program employing LLM.

2. LLM and sensor

2.1. Sensor detection via LLM

Humans perceive the world through their eyes and brain; however, how can robots or computers perceive it? Researchers are diligent in creating sensors that can identify human behavior. Over the past two decades, there has been an increased interest in Human Activity Recognition (HAR) utilizing sensor technology because to its potential applications in smart home settings, security surveillance, and healthcare. There is a growing demand for sensor technology, and many assert that LLM is proficient in this context. Researchers and experts have extensively endorsed the interpretation of sensor images through machine language. The integration of Internet of Things sensors with machine learning techniques can yield an ambient intelligence environment. Previous research has demonstrated that large language models (LLMs) can interpret complex patterns and trends in sequential data, such as in time-series analysis and as general pattern learners [5].

2.2. Encoder pairs bringing innovation to image processing

Indeed, the real-world image detected by the sensor contains countless pixels that are far more complicated than text or number inputs, which challenges the LLM to handle sequential and massive data. While traditional LLMs are trained with simple data sets and are only capable of text recognition, contemporary development in LLM makes image recognition feasible via encoders. A text encoder and an image encoder are paired in contrastive vision language models, or VLMs. To produce an embedding vector, the encoders process each input from the corresponding media. After that, researchers use descriptions and photos to train these models. As a result, the model is able to interpret visual concepts by comprehending the relationship between the text and image [6].

A pair of encoders brings the possibility to LLM for converting what is included in a visual image into concepts that are intelligible by computers or humans.

2.3. A more productive way to seek training data

Although the concept of encoder pairs is innovative, a significant drawback for the LLM's implementation is that encoders require an extensive volume of training data. This necessity is justified, given the multitude of substances present in the real world, as the training data are essential for associating each object with a corresponding text encoder, particularly for substances that are similar to one another. While acquiring training data is a crucial aspect of language machine learning, a recent study has proposed an innovative data-free method to obtain data from another extensive model, ChatGLM. In a study on Data-free Multi-label Image identification, researchers propose a framework for multi-label image identification that does not utilize any training data. A large language model (ChatGLM) can serve as a repository for encyclopedic knowledge. This is motivated by the discovery that, when provided with a linguistic description, individuals can identify an object in an image with considerable accuracy. Subsequently, employing the acquired knowledge, a pre-trained vision-language model (CLIP) is swiftly refined to enhance multi-label classification through the aligned visual and linguistic embedding space [7]. This novel data training methodology significantly enhances the creation of big language models, and its adoption is rapidly rising across numerous LLM platforms.

2.4. Solving the issue came from the far-distance in the image

Besides, those who use LLM to process 3D images might notice that the performance of LLM in image recognition is seriously affected by geometric changes such as the position and size of the object to be recognized in the image. For instance, the instance at a distant place of the image is more difficult for LLM to distinguish and output accurate results. In one research, a generative model that is effective at capturing the generating processes of observations is an SL-HMM. Using SL-HMMs, the suggested approach can extract features invariant to geometric alterations and use the recovered features to create an accurate classifier based on discriminative models [8]. As the invariance to geometric variations of the objects in the picture is extracted, applying the extracted information, LLM can distinguish the object located on the farther side more precisely.

3. An application of LLM in image detection

3.1. Functionality of the program in experiment

The image sensor is one of the most significant sensors among a diverse array of sensors. The primary function is to convert visual perceptions into signals comprehensible to computers, image sensors perform a similar role, transforming incoming light (photons) or visual signals into an electrical signal appropriate for storage, analysis, or display [9]. This paper will present a method utilizing LLM to replicate the conversion of information from a photograph into computer-readable data. The application can process the image regardless of the presence of text. The application will initially present the event depicted in the image, followed by any text contained inside the image, along with the text's placement.

3.2. Event recognition model and text recognition model

This paper uses two language machines. One language machine can briefly describe what happens in the image(event recognition model) and another language machine can tell what text occurred in the image(text recognition model). The event recognition mode is downloaded from model-scope, an open-source platform for large language models. The mPLUG task is a downstream effort that involves fine-tuning image description on the MS COCO Caption dataset for English image description. A multimodal fundamental paradigm for unified comprehension and generation is called the mPLUG model. The model suggests using skip-connections to create an effective cross-modal fusion framework [10]. Figure 1 explains how the mPLUG model works in detail.

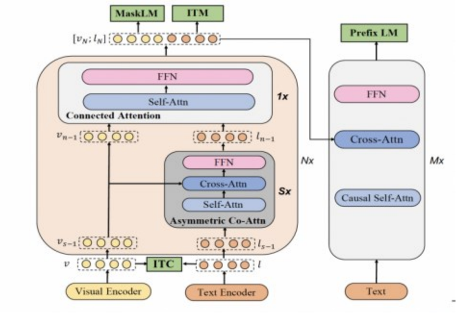


Figure 1. Model architecture of mPLUG [10]

3.3. Introduction of the experiment process

Initially, the sensor, such as a camera, captures an external image and transmits it to the computer. The computer collects photographs and organizes them into a designated folder for reading. The subsequent phase of the experimental program will establish the event recognition model, utilized to delineate occurrences or items present within the image. The program then stores the output of the event recognition model in a set. Utilize this output to ascertain the presence of text within the image. If no text is identified, the output will be transmitted directly to the user. Upon detection of text by the event recognition model, the software will initialize the text recognition model, which is employed to identify the text within an image, and the output from this model will be documented in a set. By amalgamating the outputs from two language models, we are highly likely to obtain an accurate description of the input image. Subsequently, present the consolidated result to the users and inform them that the image contains text. Finally, if the image is excessively abstract and the LLM cannot ascertain its content, an error notice will be dispatched to the user, indicating that the submitted image is not unsuitable.

3.4. Experiment data and a review on the data

The input image(Figure 2) includes text, in this situation the program will call text recognition model to read the text and image recognition model to describe event in the image. And the output image is Figure 3.

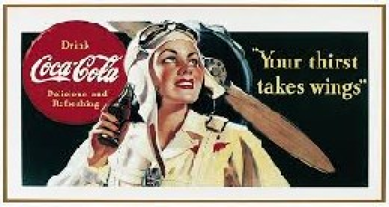
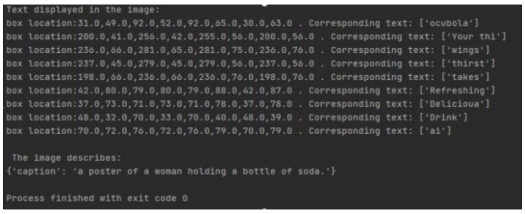


Figure 2. An advertisement, which is the experiment input image that contains text



```
Text displayed in the image:
box location:31.0,49.0,92.0,52.0,92.0,65.0,30.0,63.0 . Corresponding text: ['ocubola']
box location:200.0,41.0,256.0,42.0,255.0,56.0,200.0,56.0 . Corresponding text: ['Your thi']
box location:236.0,66.0,281.0,65.0,281.0,75.0,236.0,76.0 . Corresponding text: ['wings']
box location:237.0,45.0,279.0,45.0,279.0,56.0,237.0,56.0 . Corresponding text: ['thirst']
box location:198.0,66.0,236.0,66.0,236.0,76.0,198.0,76.0 . Corresponding text: ['takes']
box location:42.0,80.0,79.0,80.0,79.0,88.0,42.0,87.0 . Corresponding text: ['Refreshing']
box location:37.0,73.0,71.0,73.0,71.0,78.0,37.0,78.0 . Corresponding text: ['Delicioua']
box location:48.0,32.0,70.0,33.0,70.0,40.0,48.0,39.0 . Corresponding text: ['Drink']
box location:70.0,72.0,76.0,72.0,76.0,79.0,70.0,79.0 . Corresponding text: ['ai']

 The image describes:
{'caption': 'a poster of a woman holding a bottle of soda.'}

Process finished with exit code 0
```

Figure 3. The experiment outcome given the input image is figure2

The input image(Figure 4) excludes text, in this situation the program will only call image recognition model to describe event in the image. Output image is Figure 5.

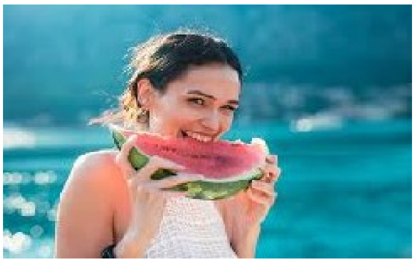
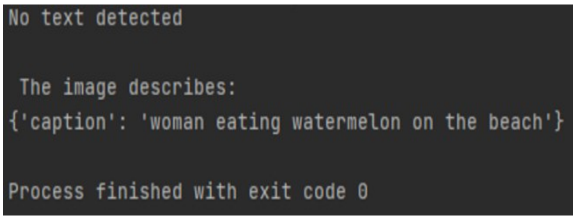


Figure 4. Image of a woman, which is the experiment input image without text



```
No text detected

 The image describes:
{'caption': 'woman eating watermelon on the beach'}

Process finished with exit code 0
```

Figure 5. The experiment outcome given the input image is figure4

The experiment's results indicate that the image containing text (Figure 2) corresponds to an output image (Figure 3) that delineates the positions of the text boxes. The x and y coordinates of the four corners of each box are provided, and the associated text within each box is matched and displayed in the console immediately following the box's location. Figure 3 illustrates that the characters present in Figure 2 have been identified by the text recognition model, hence validating the program's capacity to recognize text. The section in Figure 3 pertains to the image description, stating: 'A poster of a woman holding soda,' which accurately reflects the content of the image. Consequently, we can ascertain that the application functions effectively when the image contains text.

Upon examining the text-excluded image (Figure 4), the resultant output image (Figure 5) reveals that no text has been discovered. The subsequent section in Figure 4 contains the image description, stating: ‘woman eating watermelon on the beach,’ which accurately corresponds to the scenario depicted in the image (Figure 4). Consequently, it may be inferred that the program functions effectively in scenarios devoid of text within the image. Furthermore, as only a single LLM is required for the text-excluded image, the program's performance significantly improves.

4. Discussion

4.1. An overview of the experiment result and instructive advice

The experiment is successful; the author's program effectively describes the event depicted in the image and identifies any text there. The program's output is available for the robot or computer to document or examine. A remarkable transformation occurs from human-readable visuals to computer-readable data. Scholars are extensively discussing the growing focus on low-cost sensors. The swift expansion of the mobile phone industry in recent years has significantly propelled the trend toward the development of affordable, energy-efficient image sensors [11]. Employing the LLM for image processing in the sensor is far more economical and expedient than conventional methods. Utilizing my program in the experiment, the sensor can assess the events occurring in the detected image from the actual world with greater detail and efficiency. Consequently, the sensor's processing quality and responsiveness will be enhanced.

4.2. Future improvement and prospects

The experiment use the computer's camera to identify real-world images; nevertheless, this method of replicating sensor detection is sluggish and inflexible, necessitating the development of a more adaptable approach in the future. Although LLM is significantly more rapid than conventional image processing methods, it remains time-consuming during the loading phase, resulting in delays in sensor detection.

Although LLM is generally reliable, it exhibits inconsistencies in some instances. Therefore, regular monitoring is essential for optimal outcomes from the large language model. However, these repeated checks will prolong the duration of image processing, necessitating a significant trade-off for the software. Inappropriate for characterizing a complicated or multifaceted occurrence. Consequently, the software must decompose a complex image into simpler images that the LLM can process. The subsequent stage is to amalgamate the outcomes from basic photographs. A different LLM capable of semantic fusion is necessary in this instance.

5. Conclusion

This work has presented the contemporary advancements of large language models (LLM), their limitations, and the significance of choosing appropriate models. This essay provides an in-depth analysis of image sensors, facilitated by a language model specializing in image processing. The research posits that image sensors are crucial to various contemporary industries, including engineering and nursing, and highlights a growing demand for more efficient sensors. Despite skepticism regarding the viability of image processing with LLMs, including obstacles in locating training datasets and the susceptibility of geometric alterations in images to errors, emerging technology empowers LLMs to address these issues. The book includes a description of an experiment involving image processing through a program primarily composed of two large language models, one for text recognition and the other for image recognition. The experiment is deemed successful as the output from the program accurately represents the tested input image. The paper subsequently examines the prospective application of the experimental program in practice. Nevertheless, there are certain deficiencies in the experiment, and the author intends to address these flaws and explore this topic further in the future.

References

[1] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), 1–45. https://doi.org/10.1145/3641289

[2] van Diessen, E., van Amerongen, R. A., Zijlmans, M., & Otte, W. M. (2024). Potential merits and flaws of large language models in epilepsy care: A critical review. Epilepsia, 65(4), 873–886. https://doi.org/10.1111/epi.17907

[3] Schuster, T., Fisch, A., Gupta, J., Dehghani, M., Bahri, D., Tran, V., Tay, Y., & Metzler, D. (2022, December 6). Confident adaptive language modeling. Advances in Neural Information Processing Systems. https://proceedings.neurips.cc/paper_files/paper/2022/hash/6fac9e316a4ae75ea244ddcef1982c71-Abstract-Conference.html

[4] Chen, X., Cumin, J., Ramparany, F., & Vaufreydaz, D. (2024, June 25). Towards LLM-powered ambient sensor based multi-person human activity recognition. arXiv.org. https://arxiv.org/abs/2407.09529

[5] Ouyang, X., & Srivastava, M. (2024, March 28). LLMSense: Harnessing llms for high-level reasoning over spatiotemporal sensor traces. arXiv.org. https://arxiv.org/abs/2403.19857

[6] Esfandiarpoor, R. (2024, July 25). Vision language models: How llms boost image classification. Snorkel AI. https://snorkel.ai/improving-vision-language-models-two-studies-on-vlm-llm-cooperation/

[7] Yang1, S., & author., C. (n.d.). Data-free multi-label image recognition via LLM-powered prompt tuning. https://arxiv.org/html/2403.01209v1

[8] Y. Tsuzuki, K. Sawada, K. Hashimoto, Y. Nankaku and K. Tokuda, "Image recognition based on discriminative models using features generated from separable lattice HMMS," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 2607-2611, doi: 10.1109/ICASSP.2017.7952628.

[9] Understanding the digital image sensor - lucid vision labs. LUCID Vision Labs - Modern Machine Vision Cameras. (2024, June 4). https://thinklucid.com/tech-briefs/understanding-digital-image-sensors/

[10] ModelScope. (n.d.-a). MPLUGImage description model-Chinese-base. Moda Community. https://www.modelscope.cn/models/damo/mplug_image-captioning_coco_base_zh/

[11] Leili, B. (n.d.). Development of a mote for wireless image sensor networks. Development of a Mote for Wireless Image Sensor Networks. https://geometry.stanford.edu/papers/dra_cogis-06/dra_cogis-06.pdf