SMART INDIA HACKATHON 2024

TITLE PAGE

- **Problem Statement ID –** 1604
- **Problem Statement Title-** Conversational Image recognition Chatbot
- **Theme-** Smart Automation
- **PS Category-** Software
- **Team ID-**
- **Team Name -** Dragons of the Realm

Dragons of the Realm

Conversational Image Recognition Chatbot

Solution we offer

- Chatbot harnessing the power of **Vision Language Model** (VLM) & **Zero-shot object detection Model**
- The user can upload an image, detect the objects in it and start the chat session.
- **Enhanced spatial understanding** of objects in the images. It happens due to **inter-communication** between both the models.
- Image question answering chatbot with feature of object detection.
- Chat bot History and detection output **interaction** of the system.
- We used **9 representative VLMs** on 10 Benchmarks in **Open compass multimodal leaderboard.**
- **Best performing and most used** object detection models like OWLv2 model, Grounding DINO model.

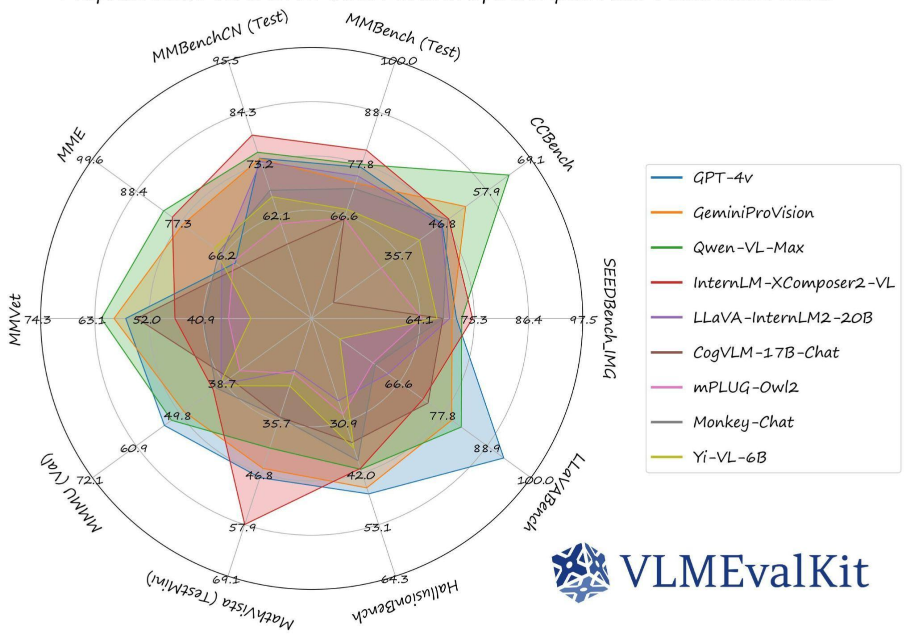
How it addresses the problem

- **Fulfils all the aspects** of the problem statement.
- Both the models help in addressing the problem since they were **pretrained on household** datasets.
- Provide unparalleled results and i**nference speed.**
- Our approach is **easy and can be implemented** with minimal efforts.

Unique value propositions

- Correct responses lexically and grammatically.
- Usage of **state of the art** and novel research work in field of image understanding. Everything we have used is **open source**work.
- **Flexibility** to use different models with reference documentation.
- **Working and hosted demo application**

9 Representative VLMs on 10 Benchmarks in OpenCompass Multi-Modal Leaderboard.



Dragons of the Realm

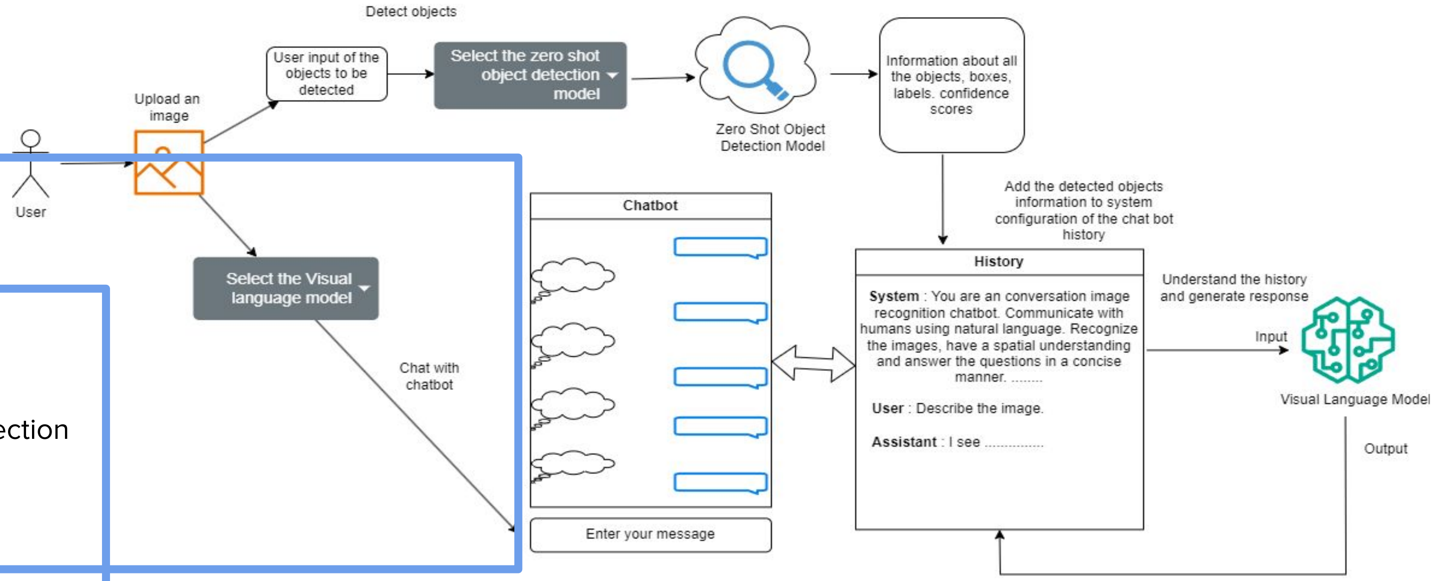
TECHNICAL APPROACH

SMART INDIA HACKATHON 2024

Technologies used

Programming languages : Python
Libraries : Transformers, Pytorch, Image libraries like PIL
Hardware : Nvidia T4 medium 8vCPU 30GB RAM
Platforms : Hugging face, arXiv, other research articles
Deployment tools : Gradio, Streamlit, Hugging face spaces hardware

Process Flow Diagram


Detect objects
User input of the objects to be detected
Select the zero shot object detection model
Zero Shot Object Detection Model
Information about all the objects, boxes, labels. confidence scores
Upload an image
User
Add the detected objects information to system configuration of the chat bot history
Select the Visual language model
Chat with chatbot
Chatbot
Enter your message
History
Understand the history and generate response
Input
Visual Language Model
Output

System : You are an conversation image recognition chatbot. Communicate with humans using natural language. Recognize the images, have a spatial understanding and answer the questions in a concise manner. ....

User : Describe the image.

Assistant : I see ....

Product status

- Built a **working demo and deployed** on hugging face spaces.
- Used google/owlv2-base-patch16-ensemble as zero shot object detection model and Qwen/Qwen2-VL-2B-Instruct as VLM
- Gradio framework and transformers library for development

Demo Link - Hugging Face Space Google Colab demo

HF space has 2vCPU-16GB RAM and **no GPU** deployed in the free tier. So the inference speed of our chatbot is very slow.
To see the demo it is highly recommended to use the Google colab demo we have provided. Get started with the demo with minimal efforts. The **inference speed increases drastically on google colab** with T4 GPU runtime.

Dragons of the Realm

FEASIBILITY AND VIABILITY

SMART INDIA HACKATHON 2024

	Technical	Financial	Market	Operational
Potential challenges	● **Computation** of large models is expensive. ● We require **powerful GPUs** like Nvidia T4 medium 8vCPU 30GB RAM, Nvidia 1xL4 8vCPU 30GB RAM etc.	● The starting cost these GPUs cost 0.60$ per hour In the **extensive usage,** using the app may be **expensive**.	● This chatbot is a very simple and **basic use case as per Problem statement.** ● But for **specific use cases** the models should be **fine tuned on respective data**	● The challenge for us to proceed is to get **a GPU with high RAM** for **deployment purpose**. We are using multiple models to give **flexibility**.
Strategies for overcoming	● As the **use case** is most basic, we have selected the best performing **models** which do not require fine tuning and give **state of the art results**. ● **9 Representative VLMs on 10 Benchmarks**	● Optimized inference pipeline ● Reduced waste for every model ● But we **need GPUs** for better **performance.**	If we really want to finetune model, then we have got a solution ● **No need to finetune all parameters.** ● We can use **PEFT** library and **adaptor** fine tuning techniques. ● This also **preserves the performance.**	● The only way to reduce the operational cost is to **reduce the number of models to be used.** ● But there will be no flexibility.

Dragons of the Realm

IMPACT AND BENEFITS

SMART INDIA HACKATHON 2024

Use cases

- Reduced **customer support** costs for businesses.
- **Anomaly**/ hazard **detection** in images/ scenes.
- **Feedback** and reviews of products using only images.
- Integration with **Autonomous** vehicle and weapon system.
- Guided tours and **information of artifacts** in museums.
- A **powerful educational tool** for anyone who interacts with images
- The solution opens avenues for **business growth and innovation** in areas like e-commerce, education, and tech support.

User experience

- Any user can interact with application using **natural language** making it easy for non-technical users.
- User can also provide ongoing **feedback and suggestions,** improving user satisfaction and experience.
- Use a model of your choice. **Flexibility**

Social and economic benefits

- Supports diverse user needs and equal access to **information to everyone i**n organization
- Assistance in navigating complex financial processes, such as **filling out applications** or understanding banking terms.
- Analyze **images of prescription drugs** to identify the drug name, composition, and expiration date, aiding in **patient safety** and medication management.
- **Monitoring patient's health** using **medical images** and data

Dragons of the Realm

RESEARCH AND REFERENCES

SMART INDIA HACKATHON 2024

Platforms

For **development** and experimentation : Kaggle, Google collab

Loading **models** and many other uses : Hugging face, Vertex AI, Open AI,

Version control system : Github

Engineering Designs : Draw.io

Articles and other resources

Tasks page :
https://huggingface.co/tasks/image-text-to-text
https://huggingface.co/tasks/visual-question-answering
https://huggingface.co/tasks/zero-shot-object-detection

Open VLM Leaderboard :
https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

Citations of the research work used in demo

Zero Shot object detection model
https://huggingface.co/google/owlv2-base-patch16-ensemble

arXiv:2306.09683 [cs.CV]

Visual Language Model
https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct

The paper is not yet published for this model