LLM Project Proposal

Title

VisionPulse: AI-Enhanced Learning and Interaction Assistant for Visually Impaired people at IIIT Delhi

Abstract

"VisionPulse" - our revolutionary idea in the educational domain to cater to the needs of visually impaired students at IIIT Delhi by leveraging multiple facets of Computer Vision and Natural Language Processing techniques to build a Large Multi-Model (LMM) that will mitigate the challenges faced by them & bridging the gap in their educational realm by providing them with highly tailored, context-aware assistance system tailored to work in the college campus setting. This will enable them to interact seamlessly with their peers and other objects within the campus environment. Our model will be powered with facial recognition techniques, deciphering complex scenarios in a localized environment and delivering precise, actionable knowledge about the surroundings to turn daily campus navigation and social encounters into a vivid experience. VisionPulse uplifts the barrier of passive participation in classroom environments by leveraging various techniques to extract audio from lecture recordings and to create them into structured, accessible notes, which will further facilitate dynamic, content-specific Q&A sessions using Retrieval Augmented Generation (RAG). The benefits of this technology ensure that no student feels alienated due to specific barriers in their education by enhancing their class participation through information retention techniques. Besides this, our project aims to provide the user with more features like real-time obstacle detection, spatial location detection, aiding in reading emails, reading, understanding, and conveying text-based instructions in front of the user, enhancing the feasibility and easing the individual's life. VisionPulse will be able to perform actions based on voice commands and be capable of generating audio-based responses as well. Our contribution of VisionPulse to the IIIT Delhi community can act as a gateway in the realm of education for visually impaired students by creating a more inclusive, empowering, and independent educational experience for them. Our project has

the potential to become one of the breakthrough contributions in the academic landscape which leverages existing tools and technology to a significant effect, which will help bridge the gaps in education for differently-abled students and create a healthy, inclusive environment where every student can thrive & unlock their full potential to scale great heights & achieve all their dreams.

Objectives

1. Leverage Large Multi-modal Models for answering questions based on knowledge extracted from visual data.
2. The objective of the pipeline will be to address key challenges faced by Visually Impaired Persons at IIIT Delhi. A few examples are listed below:
    a. Identify objects in a localized environment.
    b. Provide detailed and descriptive answers to user queries.
    c. Read, understand, and answer questions based on text found in images, such as checking expiry dates on food items, reading documents, or identifying items on a menu.
    d. Identify locations by reading and recognizing banners or signs (e.g., identifying ‘MIDAS Labs’ or ‘Cross Caps Lab’).
    e. Inform the user whether they can move forward and provide real-time alerts, such as beeping, when obstacles are detected.
    f. Pipeline to annotate visual feed using audio feed for knowledge base creation.
    g. Since the visually impaired person may not be able to access documents like IIIT Delhi policies, etc., the model should have knowledge related to IIIT Delhi built within it, which can be accessed and queried by the user at ease.
    h. Parallel Pipeline to process the video and audio feed to text with “smart storage” for future QA based on past context using RAG.
    i. Detect the user's current location and provide guidance on routes to reach their desired destination.
    j. Recognize and identify faces relevant to the user.

k. Integrate the student's mailbox by developing an email agent integrated with SMTP to enable users to read and write emails using audio prompts eg: "write a formal mail to admin B.Tech inquiring about the last date of fee payment".

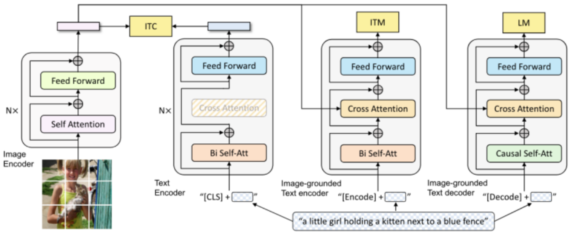
l. Allow users to control and navigate the system through voice commands. We aim to tailor each feature of this system to enhance its usability on the IIIT Delhi campus.

Literature review

1. **Bootstrapping Language Image Pre-training (BLIP)**

The BLIP model architecture proposed by Li et al. is a new Vision-Language Pretraining framework that achieves state-of-the-art performance on various vision-language tasks by addressing the limitations of existing methods. BLIP utilizes a new dataset bootstrapping technique called CapFit, which generates synthetic captions and filters out noisy captions to improve the quality of the dataset. The proposed framework introduces a multimodal mixture of encoder-decoder (MED) model architecture and leverages pre-training objectives such as image-text contrastive learning, image-text matching, and image-conditioned language modeling to achieve flexible transfer learning and effective multi-task pre-training.

From this paper, we will be using the BLIP model for Visual Question Answering and its various evaluation metrics discussed. The relevant metrics have been mentioned and explained in the evaluation criteria section.



2. **Pathways Language and Image (PALI)**

PaLI (Pathways Language and Image Model) builds on recent advancements in large language models (LLMs) and vision models by jointly modeling vision and language tasks. This approach addresses the growing demand for models that can handle multimodal inputs, where both visual and textual data are processed together. Previous research has demonstrated the effectiveness of scaling models for specific tasks, particularly in language, using architectures like Transformers and Vision Transformers (ViTs) for vision. The model achieves state-of-the-art performance in several tasks, such as image captioning, visual question-answering, and scene-text understanding. We will be using the PaLI-VQA model for our question-answering tasks.

2. **LLama 3**

The paper introduces a new series of foundational models named Llama 3. These models are a collection of language models designed to inherently handle multilingual tasks, coding, reasoning, and tool utilization. The most advanced model in this series is a dense Transformer with 405 billion parameters and a context window capable of accommodating up to 128,000 tokens. The paper includes a thorough empirical assessment of Llama 3, revealing that it achieves performance comparable to top language models like GPT-4 across a wide range of tasks. Llama 3 is being made available to the public, with both pre-trained and fine-tuned versions of the 405 billion parameter model, as well as the Llama Guard 3 model, for enhanced input and output safety. Additionally, the paper details experiments where image, video, and speech capabilities were incorporated into Llama 3 using a compositional method, which demonstrates competitive performance with current state-of-the-art techniques in these areas. We will use this model based on various language-related functionalities provided by VisionPulse and will also use different versions of fine-tuned Llama VQA models, such as **MiniCPM-Llama3-V-2_5.**

4. Wu-Palmer Similarity Score (WUPS)

The paper by Malinowski et al. introduces a performance measure called the WUPS score for evaluating the quality of system-generated answers. It draws inspiration from the Fuzzy Sets theory and utilizes the Wu-Palmer Similarity (WUPS)
score to account for semantic fuzziness between classes. WUPS score penalizes both underestimation and overestimation of answers. The formula considers the intersection of system and ground-truth answers, employing a soft membership measure. Empirical findings suggest a WUP score of approximately 0.9 for precise

answers, prompting down-weighting for scores below a threshold. A curve over thresholds illustrates the trade-off between precision and forgiveness, with WUPS at 0 being the most lenient measure and WUPS at 1.0 equating to standard accuracy. Further details about the evaluation metric will be discussed in the subsequent sections.

Approach (Methodologies and Techniques)

1. **Data collection and processing:**

   A. Permanent Context Knowledge base

   A permanent context knowledge base means knowledge that the model needs to have at all times and which, once fully developed, doesn’t require frequent updates due to the static nature of this knowledge. Creating an organization-wide knowledge base. For example, the images of various places in IIIT Delhi are a permanent context knowledge base since they are static in nature and will have minimal to no change in the future.

   Image Feeds: Click images that will contain different faces of people and places that provide personal context around the user's life in the organization, i.e. IIIT Delhi.

   Image Key-frame Extraction: Process video feed to collect key-frames, enhancing the dataset with more custom-clicked photographs and videos (if required) to enhance the model's performance in the setting of IIIT Delhi.

   Create a custom dataset of images of places and faces from the feed, clean the dataset and try to augment the dataset / click more images/videos of the places occurring in the video to make the dataset more comprehensive and generalized to work well from any angle the user clicks a photograph in during usage and minimize errors.

   Annotation: Annotate people and places from a visual feed and use an audio interface to annotate, i.e. provide knowledge of the Identity of that place.

   B. Rolling Context Knowledge base

   Rolling context knowledge base are those data which will keep increasing as the model is used. For example, the audio-based data collected for note making of lecture keeps increasing with each lecture, and hence effective storage methods need to be employed to keep only the relevant data / preserve as much data as possible. This could include summarizing the existing data to reduce it in token size or dropping data that is older than a certain cutoff.

CC Audio: Audio-to-Text Conversion of Audio feed.

***"CC"* Key Frames:** Generating descriptions of the important key-frames using visual-based models. Detailed Captioning of video feed to text from the key-frames.

OCR Key Frames: Text extraction from extracted key-frames. Using Optical character recognition, get the numbers and words in a visual feed for processing.
Used for reading text-heavy objects/environments like documents, menus, street signs, etc., for better context and aiding Visual QA.

Summarized CC: This will be a summary of the text-based feeds generated by LLM. To store only themes of old conversations.

***"Smart Storage"*:** The continuous visual and audio feeds will generate a large volume of data. Therefore, smart management is required which shifts the feed to a lower storage requirement type of feed (see note below) and deletes the data from modality as per disk and compute constraints.

Creating a resource regime to manage the computing and storage. Managing storage of these continuous feeds.
Note: Storage Requirements for various input raw feeds and generated feeds are the following:-

Video feed > Key-Frame feed >~ Audio feed > CC Key-Frame >~ CC Audio > OCR Key-Frame > Summarized CC

Regime will be like storing video feed for the last few minutes, audio feed for the last few hours, closed captions of audio and joined key frame description and OCR for last few days and summarization of all the CC containing themes. In general using this kind of calculation to find the balance between storage capacity and data quality.

2. **Implementation plan for each of the objectives:**

    a. Visual Question Answeri

We aim to cover many facets and features described in the objectives above by answering the in-context visual question. For visual QA, we are planning to use the Large Language Model known as BLIP (Bootstrapping Language-Image Pre-training), which is a Vision-Language Pre-training (VLP) model built to perform many vision-language tasks like visual question answering and can answer questions based on the image it has in front of it. We would use its image recognition and visual question-answering capabilities with the LLaMA 3.1 Instruct model for a chat bot-like feel to the user using audio.

Like any other model, BLIP is fine-tuned for visual question answering on a general dataset, and we aim to fine-tune it to specifics related to IIIT Delhi. To do this, we will employ the various techniques of parameter-efficient fine-tuning discussed in class, such as LoRA (Low-rank adaptation), Adapter-based tuning, prompt tuning, and prefix tuning. We will also fine-tune it to provide context-based summaries of images. For example, a general BLIP output of a park would be: “This is a park” However, we will aim our model to give mode context-aware outputs like: “This is the park of IIIT Delhi, the one nearby RnD block”. We will adapt the model to be able to answer questions like “Where am I?” by integrating landmarks of the college and enabling it to recognize where the image has been clicked, e.g., combining it with RAG-based pipelines of a database of pictures of various places in IIIT Delhi and use image matching to answer.

Other than this, we aim to make the model capable enough to answer questions related to person recognition to be able to answer questions like:

A. Tell me, who is this person?

B. What is this in front of me?

C. Where am I?

To achieve this, we would be doing fine-tuning, implementing RAG pipelines or using rolling adaptations by using techniques like modular adapters within the same model to save space and make the processes optimized without the use of much storage of having a separate model for every feature and just having small DNN's to get a customized model on the go.

We will build a RAG Pipeline, which will have annotated data like the image of a person and their identity, the image of a place, and a corresponding location to help recognize any specific person or place.

We will also be integrating the model with OCR reading capabilities from the image so that the model can extract relevant information like reading expiry dates from the back of a food packet, reading what's on the menu in the canteen, etc.

b. Real-time obstacle detection

Obstacle detection is something that needs to be done at all times since it needs to be almost in real-time with minimal latency, and any kind of delays in providing an alert to the user will deem this feature useless, hence using a huge model such as an LLM for image processing would not be helpful. Therefore, we plan to train a model based on the classification of various depths and train it in a supervised manner by labeling the images as an "alert" or "not alert" The model will be processing real-time video feed from the camera and in real-time detect the obstacles and beep for alerting when required. Hence, we are willing to look for relatively lighter models from hugging faces for this purpose and fine-tune them for obstacles in IIIT Delhi or train our light-supervised model with manual labeling for the task depending on what works better and gives better results on testing. This feature is something which, even though it may not exactly be related to LLMs due to the usage of a lighter model however, we feel that this is a necessary feature to have in such a system.

c. RAG-based system for IIIT Delhi's knowledge base

Since the visually impaired user may be unable to access text-based documents from the various websites, we would collect all the relevant policies and documents available on the website and vector-store them using a RAG-based system so that the user can easily question and answer them using audio-based interaction, ask for summaries for documents, etc.

d.. Agent-based system for reading and sending emails

We aim to implement an email agent using LLaMA 3.1 that integrates with the mailbox of the user using Simple Mail Transfer Protocol (SMTP) and is capable of reading emails, remembering emails, and also be able to send emails based on audio-based user commands, the system automatically drafts an email based on the audio prompt of user based on prompts such as “Write me a formal email to ask for fee details to admin B.tech”, The existing emails could be integrated with RAG based pipelines for future question answering.

e. Control and navigate the system through voice commands

A visually impaired person at any place can click the image of the place and say “Where am I?”, now this speech input would be converted into text and the model will take this text+image as input and identify the place using the RAG pipeline, Model will generate the output in text format which will then again be converted into the speech and ultimately the person will get the answer to his/her query.

f. Building app-based agents (if time permits)

We will be making LLaMA 3.1 a central piece of understanding the natural language request of the user (speech-to-text). the LLaMA model would be integrated with various tools with access to applications like Scheduling meetings, writing messages, calling numbers, etc. The LLaMA 3.1 would read and comprehend the query and direct it to the appropriate agent which would then be doing the needful by keeping its goal in mind.

Considerations

Privacy Concerns: The continuous feed of visual and audio around the subject will be quite a large privacy concern for users and other people.

Workarounds on the potential harms:

1. Prediction of sensitive input coming to and automatic suggestion and stopping input feeds selectively or/and all feeds.
2. Active deletion of visual feed with sensitive data, as soon as the scope of usage demises.
3. Storing blurred visual fields of people with only the name of the person once processed.

4. Audio command integration to stop visual feed storage as a soft stop of input.
5. Physical off button to stop all input or any selective feed for hard stop of input.

Timeline

1. Data Collection & Creation (1-2 weeks)
    a. Capturing high-quality images & audio clips.
    b. Extracting image frames from the video feed to diversify the image pool.
2. Data Preprocessing & Annotating, Extraction of text from audio (2-4 weeks)
    a. Annotating the images.
    b. Speech-to-text conversion for model training for the task of summarization.
    c. Storing the data in (Image and Corresponding text description) format.
    d. Optimizing Storage of Data for VisualPulse's different set of tasks (summarisation, Visual Q&A, etc.).
3. Integrating Different Model Functionalities
    a. Empowering the model with facial recognition by incorporating image & text data.
    b. Integrating mailbox reading & summarization.
    c. Empowering the model to identify the spatial position of a person to guide them on a path to their preferred destination.
    d. Incorporating the functionality of obstacle detection & alarming the person to avoid it & take a separate route.
4. BLIP Model inference and finetuning on the custom dataset. (4-6 weeks)
    a. Generating inference from generic BLIP Model.
    b. Applying RAG (Retrieval Augmented Generation) for Visual QnA.
    c. Fine-tuning BLIP on the custom dataset
5. Testing of Model through RAG pipeline and Generic Model (6-8 weeks)

   a. Generating inference from the RAG-based model.

   b. Improving upon the results of the RAG-based model by giving human feedback.

5. Comparison and Analysis of Results and Documentation. (8-10 weeks)

   a. Comparison of model results from the state-of-the-art publications.

   b. Visualizing the evaluation metrics using plots & graphs for better insights into model performance.

   c. Analysis of model results & identifying points of differences (unique findings of our evaluation) from the state-of-the-art publications.

   d. Documenting the results in the form of a table & generation of plots to be put together in a report/paper for potential publication.

Evaluation Criteria

1. **VQAScore**: This metric will used to evaluate the performance of models on the Visual Question Answering (VQA) task. The VQA task requires a model to answer a question posed in natural language based on a given image.

   This is computed as follows:

   a. Multiple human annotators (typically 10) answer each question in the dataset. This allows for variations in possible correct answers.

   b. A model's predicted answer is compared to these human-provided answers. The score for the model's prediction is calculated based on how many annotators provided the same answer. The scoring formula is:

      VQAScore = min(#human answers matching prediction/3, 1)

2. **Wu-Palmer Similarity Score (WUPS):** The WUPS calculates the similarity between two words based on their longest common subsequence in the taxonomy tree. If the similarity between two words is less than a threshold then a score of zero will be given to the candidate answer. We will be using the Wordnet database to measure the similarity between the words. The WUPS Score is calculated using the below formula

$$\text{WUPS}(A,T) = \frac{1}{N}\sum_{i=1}^{N} \min\{ \prod_{a\in A^i} \max_{t\in T^i} \mu(a,t), \\ \prod_{t\in T^i} \max_{a\in A^i} \mu(a,t)\}$$

We will be providing the VQAScore and WUP Score for different Visual Question Answering models for comparison.

3. **Bilingual Evaluation Understudy (BLEU) Score**: This metric will be used to evaluate the performance of our VQA model and also the LLM model which will be used for context-based question answering tasks. The metric is a geometric average of the modified n-gram precisions, pn, using n-grams up to length N and positive weights wn summing to one. Let c be the length of the predicted sentence and r be the ground truth sentence length.

   a. Brevity Penalty BP

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

   b. BLEU score

$$BLEU = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

   We utilize BLEU-1, BLEU-2, BLEU-3, and BLEU-4 by respectively adjusting N and applying uniform weights wn = 1/N.

4. **BERT Score:** BERTScore leverages the pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity. It has been shown to correlate with human judgment on sentence-level and system-level evaluation. Moreover, BERTScore computes precision, recall, and F1 score, which can be useful for evaluating different language generation tasks.

Team Details

Name	Roll number
Akash Kushwaha	2021514
Shreyas Kabra	2021563
Utkarsh Venaik	2021570
Manav Mittal	2021538
Lakshay Kumar	2021061
Tanmay Singh	2021569

References

1. Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In *International conference on machine learning*, pp. 12888-12900. PMLR, 2022.

2. Chen, Xi, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman et al. "Pali: A jointly-scaled multilingual language-image model." *arXiv preprint arXiv:2209.06794* (2022).

3. Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur et al. "The llama 3 herd of models." *arXiv preprint arXiv:2407.21783* (2024).

4. Malinowski, Mateusz, and Mario Fritz. "A multi-world approach to question answering about real-world scenes based on uncertain input." *Advances in neural information processing systems* 27 (2014).