AudiEyes

Ayoub Maimmadi

# 1. Project Inception

## 1.1. Framing the Business Idea as an ML Problem

* Business case description

Audieyes is a groundbreaking project aimed at enhancing the independence and safety of visually impaired individuals through advanced machine learning technologies. The service utilizes image and video captioning to provide real-time, detailed descriptions of environments and objects, improving navigation and accessibility. This addresses a critical need within the underserved assistive technology market, offering both significant social impact and commercial potential. By focusing on inclusivity and technological innovation, Audieyes not only opens new markets for businesses but also profoundly enhances the quality of life for its users. For more details click [here](https://github.com/AyoubMaimmadi/Audieyes?tab=readme-ov-file#business-case).

* Business value of using ML

Utilizing machine learning in the Audieyes project enables precise and adaptive recognition of visual data, greatly enhancing the service's capability to assist visually impaired users. ML algorithms optimize the accuracy and speed of image and video captioning, ensuring real-time feedback that is critical for navigation and interaction. Moreover, continuous learning from user interactions and feedback refines the system's effectiveness, ensuring that the technology evolves to meet diverse user needs. This not only fosters user dependency and satisfaction but also positions Audieyes as an innovative leader in assistive technologies, potentially increasing market share and generating sustainable revenue streams. For more details click [here](https://github.com/AyoubMaimmadi/Audieyes?tab=readme-ov-file#business-value-of-using-ml).

* Data overview

Audieyes utilizes a rich dataset comprising image-text pairs from diverse sources such as COCO, flickr30k, vqa, and nlvr. These datasets provide a wide variety of visual scenarios and associated descriptions, essential for training the ML models to recognize and articulate the content of images accurately. The data includes everyday objects, people, scenes, and activities, ensuring comprehensive coverage and relevance to real-world situations faced by visually impaired users. This extensive data foundation enables the ML models to deliver precise and contextually appropriate captions, critical for the functionality of Audieyes. For More details click [here](https://github.com/AyoubMaimmadi/Audieyes?tab=readme-ov-file#data-sources).

|  |  |  |
| --- | --- | --- |
| Model checkpoints | Dataset trained on | Pretrained model |
| model\_base.pth | /export/share/junnan-li/VL\_pretrain/annotation/coco\_karpathy\_train.json | N/A |
| model\_base\_nlvr.pth | http://export/share/datasets/vision/NLVR2/ | https://storage.googleapis.com/sfr-vision-language-research/BLIP/models/model\_base\_nlvr.pth |
| model\_base\_retrieval\_coco.pth | Coco: http://export/share/datasets/vision/coco/images/ | https://storage.googleapis.com/sfr-vision-language-research/BLIP/models/model\_base\_retrieval\_coco.pth |
| model\_vqa.pth | http://export/share/datasets/vision/VQA/Images/mscoco/ | https://storage.googleapis.com/sfr-vision-language-research/BLIP/models/model\_base\_vqa\_capfilt\_large.pth |
| model\_base\_retrieval\_flickr.pth | Flickr: http://export/share/datasets/vision/flickr30k/ | https://storage.googleapis.com/sfr-vision-language-research/BLIP/models/model\_base\_retrieval\_flickr.pth |

A screenshot of a computer program

Description automatically generated

* Project archetype

Audieyes represents a hybrid of autonomous real-time systems and human-in-the-loop architectures. It combines the speed and scalability of automated machine learning for instant image and video captioning with the precision of human oversight to enhance accuracy and contextual relevance. This model ensures that the system can operate independently while still benefiting from periodic human intervention to refine outputs, making it highly effective for applications in dynamic and varied environments. The design caters specifically to the needs of the visually impaired, promoting independence through technology while maintaining the flexibility to adapt and improve with user feedback. For more details click [here](https://github.com/AyoubMaimmadi/Audieyes?tab=readme-ov-file#architectural-archetypes-for-vision-captioning).

## 1.2. Feasibility analysis

* + literature review

The development of "Audieyes" is underpinned by significant advancements in machine learning, particularly in the fields of image and video processing that cater to enhancing accessibility for visually impaired individuals. This literature review examines the relevant technologies and models that form the basis of the Audieyes project, focusing on their applicability and effectiveness in real-world scenarios.

1. **. Machine Learning Models for Image and Video Captioning**
   1. **. Salesforce's BLIP Model:**

The BLIP (Bootstrapped Language Image Pretraining) model from Salesforce is a cornerstone in the image captioning domain, known for its efficacy in generating accurate and contextually relevant descriptions from images. BLIP’s architecture is designed to leverage large-scale datasets combining images with textual annotations, which is ideal for the diverse inputs expected in Audieyes. This model's proven capabilities in visual question answering and image captioning provide a reliable foundation for developing Audieyes’ core functionalities.

* 1. **. Moondream:**

Moondream is a compact, open-source computer vision model characterized by its portability and efficiency. Its design allows it to run on a wide range of devices, making it particularly suitable for real-time applications in accessibility technologies. The model’s robust performance in various settings demonstrates its potential to be integrated into Audieyes, ensuring that the application is both scalable and adaptable across different hardware environments.

* 1. **. Hugging Face’s Vit-GPT2 Image Captioning:**

Hosted on the Hugging Face platform, the Vit-GPT2 model combines vision-transformer (ViT) and GPT-2 technologies to create a potent image-to-text system. This model excels in translating visual data into descriptive text, a key feature for Audieyes. The use of transformer architectures allows for nuanced understanding and generation of captions, which can significantly enhance the user experience for the visually impaired by providing detailed and accurate descriptions of their surroundings.

* + Model choice/ specification of a baseline

For the Audieyes project, we have chosen the BLIP (Bootstrapping Language-Image Pre-training) model as our baseline due to its comprehensive abilities in both vision-language understanding and generation, which are essential for real-time image and video captioning applications aimed at assisting visually impaired users.

**Key Features and Specifications of BLIP:**

**Architecture:** BLIP employs a Multimodal Mixture of Encoder-Decoder (MED) architecture, which is versatile in functioning as a unimodal encoder, an image-grounded text encoder, or an image-grounded text decoder. This flexibility allows it to adaptively handle various tasks, including image captioning and text retrieval, within the same framework.

**Pre-training Methodology:** BLIP uses a novel dataset bootstrapping technique, CapFilt, which involves generating synthetic captions for web images and filtering out noisy ones. This method improves the quality of training data, especially considering the prevalent noise in web-crawled datasets.

**Performance Metrics:**

**Image-text retrieval:** Improved recall@1 by +2.7% on average.

**Image captioning:** Enhanced CIDEr scores by +2.8%.

**Visual Question Answering (VQA):** Increased VQA score by +1.6%.

**Generalization:** BLIP demonstrates strong zero-shot generalization to video-language tasks, which could be beneficial for extending Audieyes to video captioning in the future.

**Advantages for Audieyes:**

**High Relevance and Accuracy:** BLIP's capability to generate contextually relevant captions directly supports the core functionality of Audieyes, enhancing the user's understanding of their surroundings.

**Scalability and Flexibility:** The model’s architecture and the CapFilt training approach allow for easy scalability and adaptation to various data environments, crucial for deployment across different geographic locations.

* + Metrics for business goal evaluation

To effectively measure the success and impact of the Audieyes project, it is essential to establish clear, quantifiable metrics that align with the business objectives and operational goals. The following metrics have been identified to evaluate the performance and business value of Audieyes:

**1. User Engagement Metrics:**

**. Daily Active Users (DAU) and Monthly Active Users (MAU):** These metrics will help track the usage frequency and retention of the application, providing insights into its acceptance and value to users.

**. Session Length:** Measures the average duration users interact with the app per session, indicating the application's utility and user reliance.

**2. Performance Metrics:**

**. Caption Accuracy:** Accuracy of the image and video captions generated by the BLIP model, assessed through human validation or comparison with benchmark datasets.

**. System Latency:** Time taken from image/video input to caption output, crucial for real-time performance. A lower latency ensures a seamless experience for users.

**3. Customer Satisfaction Metrics:**

**. Net Promoter Score (NPS):** This metric gauges user satisfaction and likelihood to recommend Audieyes to others, which is pivotal for organic growth in the consumer base.

**. User Feedback and Reviews:** Qualitative assessments from users, providing insights into the app's impact on their daily lives and areas needing improvement.

**4. Accessibility Impact Metrics:**

**. Tasks Completed:** The number of tasks users successfully complete with the aid of Audieyes, such as navigation in new environments or identification of objects.

**. Incidence of Accessibility Issues:** Tracking issues reported by users related to accessibility, helping to refine and enhance the app’s functionality.

**5. Economic Metrics:**

**. Customer Acquisition Cost (CAC):** The cost associated with acquiring a new customer, essential for evaluating the efficiency of marketing strategies.

**. Lifetime Value (LTV):** The total revenue expected from a typical customer over the life of their relationship with the app, indicating the long-term viability of the project.

# 2. ML Pipeline Development - From a Monolith to a Pipeline

## 2.1. Ensuring ML Pipeline Reproducibility (milestone 2, 15%)

* Project structure definition and modularity

The Audieyes project is structured into a modular architecture to enhance scalability, ease maintenance, and facilitate collaboration among development teams. This architecture is divided into several key layers: the Data Layer, responsible for data ingestion, preprocessing, and storage, ensuring that data remains consistent and accessible across the system; the Model Layer, which houses the machine learning models including the baseline BLIP model and any additional models for specific tasks; the Service Layer, which interfaces the model outputs with the application layer, handling API requests and responses efficiently; and the Application Layer, which consists of both the user-facing front-end interface and the back-end application logic that allows for user interactions and data visualization.

Each of these layers is designed as an independent module with well-defined interfaces for interactions with other modules. This modularity allows for individual components of the system to be updated or replaced without disrupting the entire system's functionality, which is crucial for implementing continuous integration and deployment practices effectively.

Furthermore, the modular design supports parallel development where different teams can work on separate modules without interference, thereby speeding up the development process and enhancing efficiency. It also simplifies testing and maintenance since modules can be individually tested and maintained, thus reducing complexity and minimizing the risk of introducing system-wide failures. Lastly, the approach promotes reusability, where common functionalities are abstracted into shared libraries or services that can be reused across different parts of the project or even in future projects.

Overall, this structured and modular approach not only ensures the robustness and reproducibility of the ML pipeline for Audieyes but also prepares the system for future expansions and adaptations, aligning with the project’s long-term vision and objectives.

* Code versioning

For the Audieyes project, we utilize GitHub and Docker Hub to ensure robust code versioning and management. GitHub serves as the central repository for all project code, including the application logic, machine learning models, and system configurations. This platform facilitates version control, allowing the team to track changes, revert to previous states, and manage branches for feature development and bug fixes efficiently. Collaborative features such as pull requests and code reviews help maintain code quality and consistency across the development lifecycle.

Additionally, Docker Hub is employed to manage the Docker images of the project. This integrates seamlessly with our development pipeline, enabling consistent environments from development through to production. By storing and versioning Docker images on Docker Hub, we ensure that any member of the team or the CI/CD pipeline can pull the exact versions of the environment needed to run the application, enhancing reproducibility and reducing "works on my machine" issues.

* Data versioning

For the Audieyes project, data versioning is managed through Data Version Control (DVC), which plays a crucial role in maintaining the integrity and reproducibility of the machine learning pipeline. DVC allows the team to track versions of data sets and models, ensuring that every experiment can be reproduced and every model training session can be traced back to its data source. This tool integrates seamlessly with existing version control systems like Git, but it manages large data files and machine learning models that Git cannot handle efficiently.

Using DVC, changes to datasets are tracked in a manner similar to source code, which enables the team to pinpoint the exact data version used for specific training runs. This capability is vital for debugging and improving models, as well as for team collaborations, where consistency in data usage needs to be maintained across different members and potentially remote environments. DVC also supports data storage in remote locations, providing flexibility in how and where data is accessed and stored, crucial for handling large-scale datasets typical in machine learning projects like Audieyes.

* Experiment tracking and model versioning

critical components in managing the machine learning lifecycle effectively. MLflow offers a centralized platform to monitor experiments, including tracking of parameters, metrics, and model artifacts across various stages of the ML pipeline. This functionality allows our team to compare different model versions objectively, facilitating the identification of the most effective configurations based on empirical data.

Model versioning in MLflow provides a systematic way to version and store models, ensuring that every iteration is cataloged and retrievable. This capability is crucial for rolling back to previous versions if needed and for auditing purposes, allowing us to maintain a detailed history of model development and changes over time.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

By integrating MLflow, the Audieyes project benefits from enhanced reproducibility and accountability in model development, enabling seamless transitions between model iterations and ensuring consistent results in deployment environments. This setup not only streamlines the development process but also supports robust testing and deployment strategies, crucial for delivering a reliable assistive technology solution.

* Setting up a meta store for metadata

In the Audieyes project, we've integrated MLflow to set up a meta store for efficiently managing metadata. MLflow's meta store centralizes and streamlines the storage of essential metadata such as model parameters, configurations, and experiment results. This setup not only enhances the reproducibility of our ML pipeline but also facilitates seamless access and analysis of metadata across the project's lifecycle. By using MLflow's meta store, we ensure robust management of metadata, enabling effective tracking of experiments and model versions, which is crucial for maintaining consistency and accelerating development cycles in our Audieyes system. For more details click [here](https://github.com/AyoubMaimmadi/Audieyes?tab=readme-ov-file#experiment--metadata-tracking-with-mlflow).

* Setting up the machine learning pipeline under an MlOps platform

In the Audieyes project, we have set up our machine learning pipeline using ZenML, an open-source MLOps framework that enhances the entire lifecycle of machine learning from data ingestion to model deployment. ZenML supports our goal of developing and testing the ML pipeline with a structured and scalable approach, emphasizing reproducibility, collaboration, and automation. Our pipeline includes key components such as data validation to ensure the integrity of input data, data ingestion from diverse JSON sources, and data splitting to segregate data into training and validation sets effectively. Additionally, we use ZenML for training our models using the BLIP architecture for image captioning and rigorously evaluate model performance on the validation set to pinpoint areas for improvement. To ensure our models perform reliably and meet quality standards, we integrate GitHub Actions for running behavioral model tests, which automatically validate changes in the codebase and maintain high standards of functionality and performance. This comprehensive setup under ZenML not only streamlines our ML pipeline but also embeds best practices of MLOps, ensuring that the Audieyes project achieves its technological and business objectives efficiently.

View all the details of the pipline [here](https://github.com/AyoubMaimmadi/Audieyes?tab=readme-ov-file#milestone-4-ml-pipeline-development-and-testing).

## 2.2. Pipeline Components (Milestone 3 and 4, 20%)

* + 1. Setup of data pipeline within the larger ML pipeline/ MLOps Platform
  + (5 pts) Data Validation and Verification
  + (5 pts) Preprocessing and Feature Engineering
    1. (5 pts) Integration of model training and offline evaluation into the ML pipeline / MLOps Platform
    2. (5pts) Development of model behavioral tests
    3. Energy efficiency measurement (optional, 2 pts bonus)

# 3. Model Deployment (Milestones 5-6, 35%)

## 3.1 ML System Architecture

* (5 pts) Drawing with architecture highlights

## 3.2 Application development

* (5 pts) Model service development
* (5 pts) Front-end client development

## 3.3 Integration and Deployment

* (3 pts) Packaging and containerization
* (5 pts) Integration with a CI/CD Pipeline
* (3 pts) Hosting the application

## 3.4. Model Serving and online testing

* (3 pts) Model serving runtime
* (3 pts) Serving mode (batch, on demand to a human, on demand to a machine)
* (3 pts) Online testing (A/B Testing, Bandit)

# Monitoring and Continual Learning (milestone 7, 25%)

## 4.1. (2 pts) Resource Monitoring

## 4.2. (10 pts) Model Performance Monitoring or data distribution drift monitoring

## 4.3. (10 pts) Continual Learning: CT/CD pipeline

## 4.4. (3 pts) Pipeline orchestration

# 5. Responsible AI (milestone 8-optional, for later, 15% bonus)

## 5.1 Evaluation Beyond Accuracy

* (7.5 pts) Audit Model for Bias
* (7.5 pts) Model Explainability and Interpretability

# 6. Conclusion

* Summary of Achievements
* Lessons Learned
* Future Directions

# References

# Appendices (if necessary)