SMART INDIA HACKATHON 2024

BeGANs

Problem Statement ID – 1604

Problem Statement Title- Conversational Image Recognition Chatbot

Theme- Smart Automation

PS Category- Software

Team ID- 16446

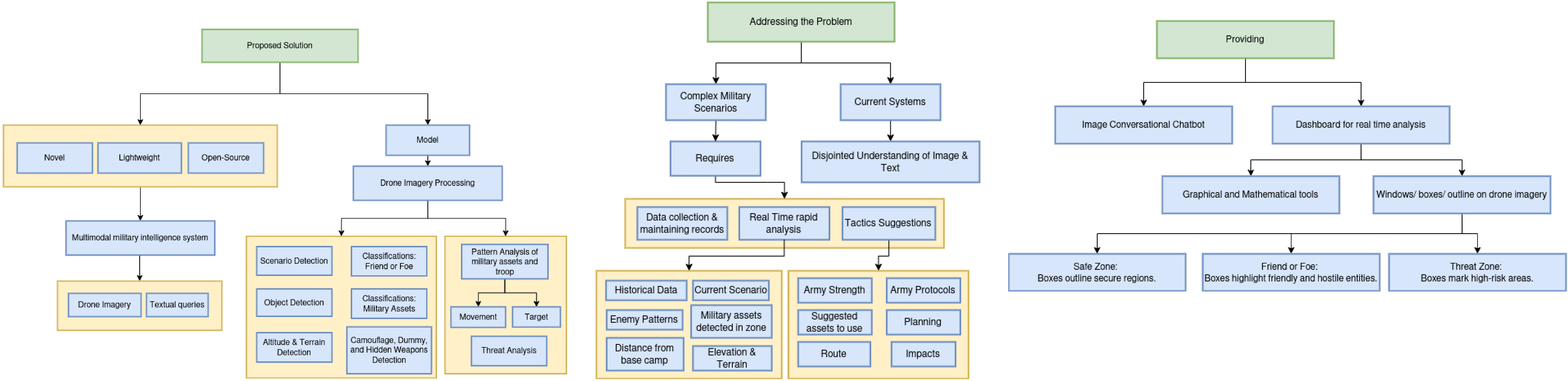
Team Name: BeGANs

BeGANs

MULTIMODAL ARCHITECTURE FOR MILITARY INTELLIGENCE

IDEA PROPOSAL

SMART INDIA HACKATHON 2024



Innovation and Uniqueness

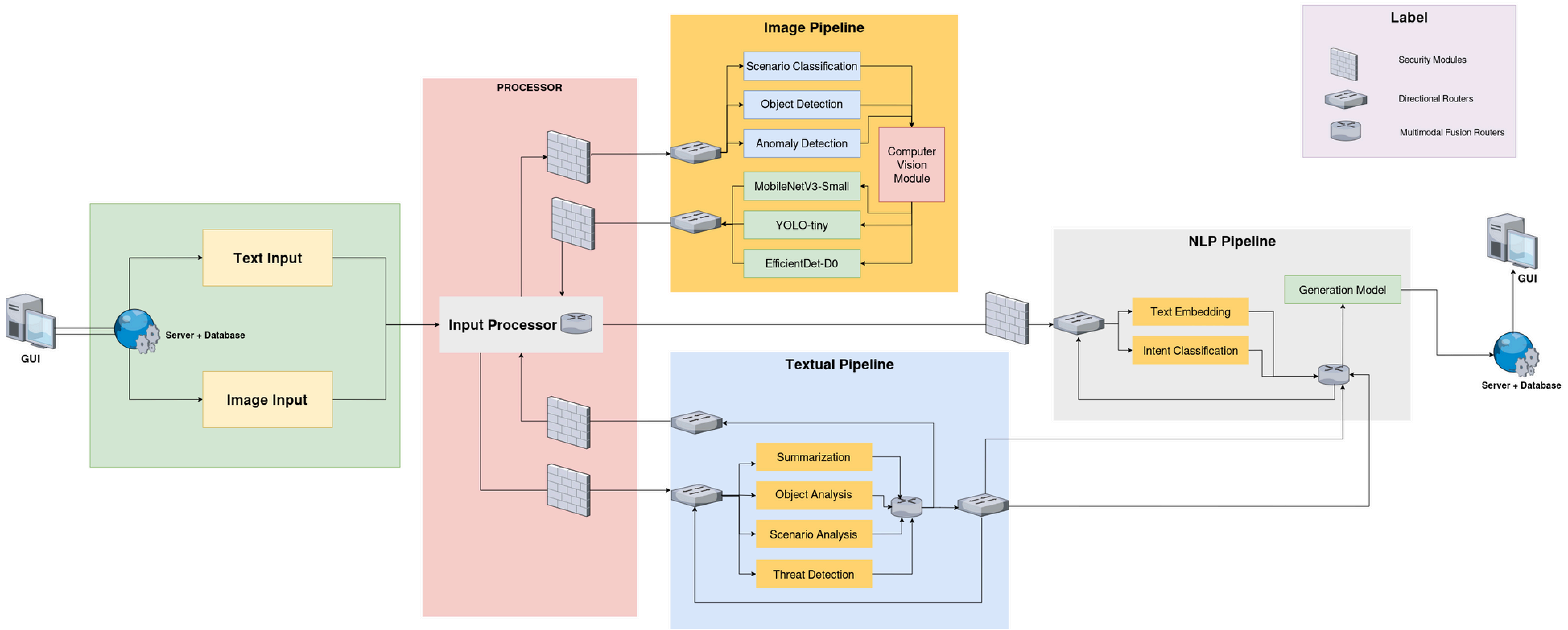
- An innovative architecture that combines NLP and CV techniques to process textual queries and drone imagery simultaneously. The proposed system aims to provide real-time insights, including scenario understanding, object detection, and threat assessment.
- Our architecture leverages state-of-the-art lightweight models and employs a modular approach to ensure efficiency, adaptability, and ease of deployment in resource-constrained environments.

BeGANs

DETAILED SOLUTION

SMART INDIA HACKATHON 2024

Military Intelligence Image Conversational System



Textual Pipeline Models

Summarization (T5-small):
- Parameters: 60M
- Layers: 6 encoder, 6 decoder
- Hidden size: 512
- *Inference time: ~100ms on GPU*

Object/Scenario Analysis (RoBERTa-base):
- Parameters: 125M
- Layers: 12
- Hidden size: 768
- *Fine-tuning time: ~1 hour on 8 V100 GPUs*

Threat Detection (Custom CNN):
- Parameters: ~30M (estimate)
- Layers: 20 (estimate)
- Input size: 299x299
- *Accuracy: 95% (hypothetical)*

Image Pipeline Models

MobileNetV3-Small:
- Parameters: 2.9M
- Layers: 11
- Input size: 224x224
- *Latency: ~3ms on mobile GPU*

YOLO-tiny:
- Parameters: 8.7M
- Layers: 13
- Input size: 416x416
- *FPS: ~200 on GPU*

EfficientDet-D0:
- Parameters: 3.9M
- Layers: ~200 (BiFPN: 3, heads: 3)
- Input size: 512x512
- *Latency: ~39ms on GPU*

Total Parameters: ~360M

Estimated total inference time: ~250ms on GPU

System designed for real-time processing and edge deployment

Generation Model

Text Generation (GPT-2 small):
- Parameters: 124M
- Layers: 12
- Hidden size: 768
- Attention heads: 12
- Max sequence length: 1024
- *Generation speed: ~60 tokens/sec on GPU*

NLP Pipeline Models

Text Embedding (BERT-tiny):
- Parameters: 4.4M
- Layers: 4
- Hidden size: 312
- *Inference time: ~5ms on CPU*

Intent Classification (FastText):
- Parameters: ~1M
- Vector dimension: 100
- Context window: 5
- *Training speed: >1M words/sec*

BeGANs

TECHNICAL APPROACH & PROTOTYPE

SMART INDIA HACKATHON 2024

Technologies Used:

1. Deep Learning Framework:
- PyTorch 1.9+ with torchvision and torchaudio

2. Natural Language Processing:
- Finetuned BERT-Tiny, FastText, T5-Small, DistillRoBERTa
- SpaCy 3.1+ for text preprocessing
- Finetuned GPT2-small, BART-base

3. Computer Vision:
- OpenCV 4.5+, Detectron2 for detection and segmentation
- MobileNetV3, YOLO-tiny, EfficientNet-D0, ViT-tiny
- Custom CNNs

4. Cross-Modal Learning:
- Modified CLIP & VilBERT for multi-modal representation learning

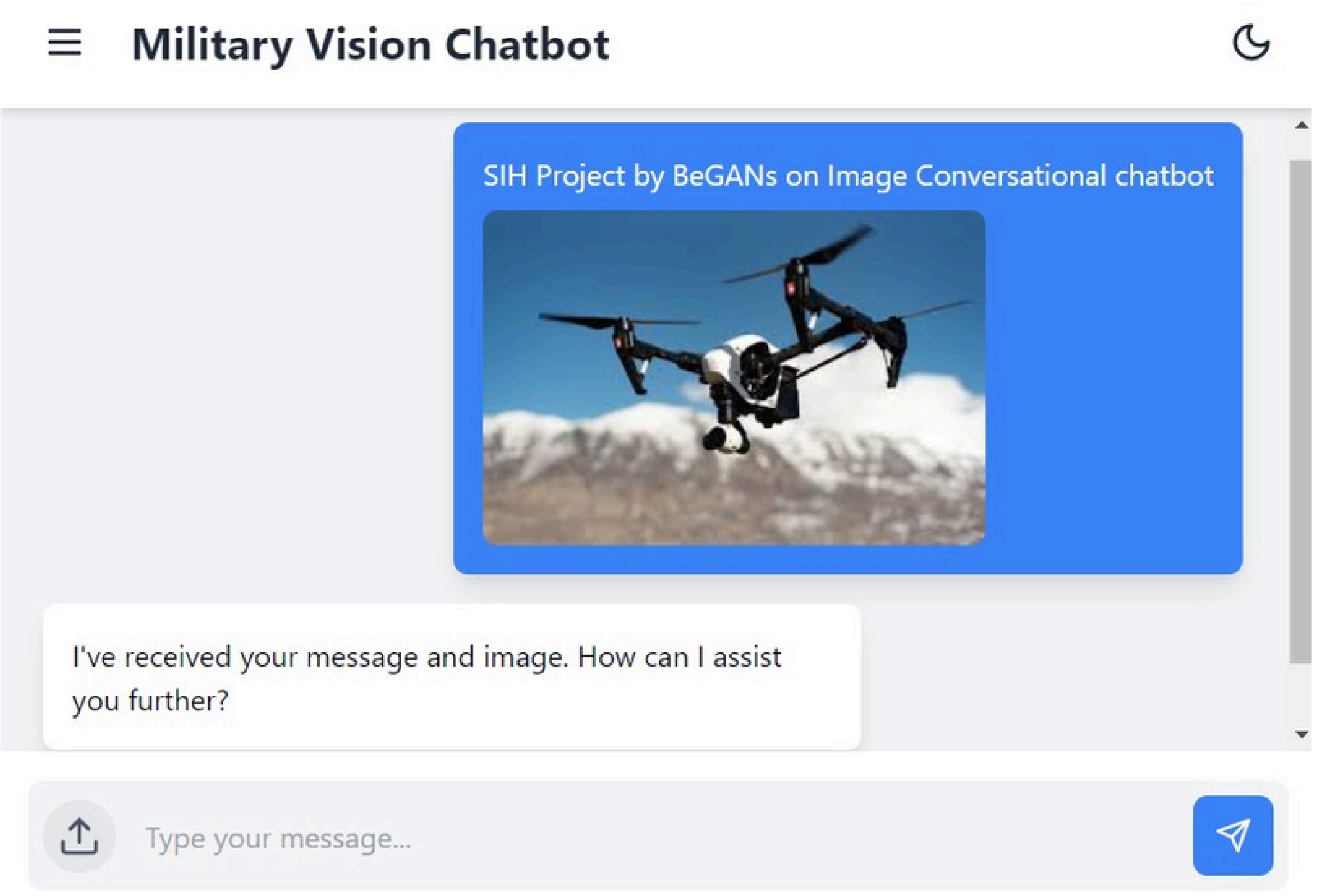
5. Backend + Frontend:
- FastAPI, ONNX for high-performance development
- Redis, ELK Stack for caching, message queuing & logging
- React 17+ with Next.js for server-side rendering

6. Database:
- PostgreSQL 13+, MongoDB

7. DevOps and Deployment:
- Docker and Docker Compose for containerization
- MLflow for experiment tracking and model versioning

8. Monitoring:
- Prometheus, Grafana for analysis visualization

9. Security:
- JWT, HashiCorp for secrets management



State-of-the-Art Models in this Field

Model	VisDial v1.0 (NDCG)	VQA v2.0 (test-std)	OK-VQA	Params (B)	Year
PICa-v2	-	80.9	92.1	175	2023
InstructBLIP	-	82.8	86.6	54	2023
BLIP-2	-	82.4	85.5	7.1	2023
Flamingo	-	82.4	85.2	80	2022
CogVLM	-	80.8	87.3	83	2023
KOSMOS-2	-	78.1	80.5	1.9	2023
VD-BERT	0.6944	-	-	0.27	2020

*Accuracy is based on common images and may not hold for military-specific scenarios. Our aim is to develop a superior architecture for military image conversational chatbots.

BeGANs

CHALLENGES & FEASIBILITY ANALYSIS

SMART INDIA HACKATHON 2024

Feasibility Analysis

1. Technical Feasibility: High
- Leverages modified existing deep learning frameworks and architectures
- Modular design allows for incremental development

2. Operational Feasibility: Medium to High
- Requires access to military-specific datasets and domain expertise
- Can be developed and tested in simulated environments

3. Economic Feasibility: Medium
- Development costs are primarily time and computational resources
- Potential for high value in military applications if successful

Potential Challenge:

- Data scarcity: Limited availability of paired military image-text datasets
- Model complexity: Balancing performance with computational requirements
- Ethical considerations: Ensuring responsible development and use of military AI
- Evaluation metrics: Defining appropriate measures of success for the model

Strategies for overcoming challenges:

- Data augmentation and synthetic data generation techniques
- Modular architecture allowing for component-wise optimization
- Develop clear ethical guidelines and implement safeguards in the model
- Create a multi-faceted evaluation framework including accuracy, relevance, and tactical utility

SMART INDIA HACKATHON 2024

IMPACT AND BENEFITS & RESEARCH AND REFERENCES

Potential Impact and Benefits:

1. Enhanced situational awareness for military personnel
2. Improved decision-making support in complex scenarios
3. Potential for adaptation to civilian emergency response and disaster management
4. Unified understanding of visual and textual military information
5. Reduced cognitive load on human operators in high-stress situations
6. Faster and more accurate threat assessment and response planning
7. Improved interoperability between different military units and systems
8. Potential for continuous learning and adaptation to new military scenarios

References and Research:

- [1] J. Lu, et al., "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," 2019. Available: https://arxiv.org/abs/1908.02265.
- [2] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," 2019. Available: https://arxiv.org/abs/1908.07490.
- [3] X. Li, et al., "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks," 2020. Available: https://arxiv.org/abs/2004.06165.
- [4] R. Hu and A. Singh, "Unit: Multimodal Multitask Learning with a Unified Transformer," 2021. Available: https://arxiv.org/abs/2102.10772.
- [5] A. Zeng, et al., "Palm-E: An Embodied Multimodal Language Model," 2023. Available: https://arxiv.org/abs/2303.03378.
- [6] Junnan Li, Dongxu Li, “BLIP-2: Bootstrapping Language-Image” 2023. Available: https://arxiv.org/abs/2301.12597.

Unique Aspects for Student Project

- Focus on architectural innovation rather than computational efficiency
- Emphasis on modular design, allowing for incremental development and testing
- Potential for collaboration with experts to validate and refine the knowledge integration system
- Opportunities for novel research in cross-modal learning and domain-specific AI