PROJECT PROPOSAL FOR

FAER - Integra SCHOLAR PROGRAM: 2023-24

ON

Integrating Deep Learning for Safety Management in Construction: Semantic Image-Rule Matching for Unsafe Behavior Detection Using YOLO-ESCA

Submitted to

Foundation for Advancement of Education and Research

Submitted by

Abishek S, IV – Year

Mohamed Aslam K, IV – Year

Under the Guidance of

Dr. P.Swathika

Assistant Professor Senior Grade

Department of Artificial Intelligence and Data Science

MEPCO SCHLENK ENGINEERING COLLEGE

SIVAKASI - 626 005

October 2024

Foundation for Advancement of Education and Research

G5, Swiss Complex, 33, Race Course Road, Bangalore- 560001
E-mail: office@faer.ac.in, Website: www.faer.ac.in

FAER SCHOLAR PROGRAM: 2024-2025

Integrating Deep Learning for Safety Management in Construction: Semantic Image-Rule Matching for Unsafe Behavior Detection Using YOLO-ESCA.

Name of the Students	: Abishek S, Mohamed Aslam K
Degree registered	: Bachelor of Technology
Branch	: Artificial Intelligence and Data science
Year	: 4th Year
Address (residential)	: 143, Jeevanagar 1st street Jaihindpuram Madurai-11
E-mail address	: alifnisha89295_bai25@mepcoeng.ac.in
Phone / Cell no	: 7695932602
Name and Address of the College (including pin code)	: Mepco Schlenk Engineering College Sivakasi,Virudhunagar-626005
Web site of the college	: www.mepcoeng.ac.in
Phone No	: 04562-235109
Name of the Supervisor / Project Guide with E-mail address and mobile Number	: Dr.P.Swathika, PhD., swathikap@mepcoeng.ac.in +91 90474 28267

1. Whether the students belong to SC/ST : No
2. Title of the Project : **Integrating Deep Learning for Safety Management in Construction: Semantic Image-Rule Matching for Unsafe Behavior Detection Using YOLO-ESCA.**
3. Area : Artificial & Hybrid Intelligence, collaborative working of humans with robots.

   Relevance to the area of the contest : The integration of AI-driven safety systems like deep learning-based unsafe behaviour detection enhances overall construction site safety by continuously monitoring workers and their environment. These systems use computer vision to identify hazards, unsafe actions, and rule violations, preventing accidents in real-time. In Hybrid Intelligence settings, AI automates safety surveillance, ensuring that both human workers and robots can operate safely together. By detecting risks early, AI enables timely interventions from human supervisors, improving decision-making and accident prevention. In collaborative human-robot tasks, AI ensures that human behaviour follows safety protocols, reducing risks while optimizing robot efficiency. This creates a safer, smarter construction site were humans and AI work together seamlessly.

   Type of Project: Concepts / Experimental / Application / Technology Development / Computer based Product

   Objective of the Project:
   - To enhance detection accuracy by integrating cross-attention between image features and safety rule embeddings.
   - To utilize YOLO-ESCA for extracting image features for safety monitoring.

- To integrate a GRU network for embedding safety rules effectively.
- To leverage multimodal learning for accurate interpretation of safety compliance.
- To integrate human-in-the-loop for critical decision-making.

Brief Description : The AI-powered safety management system for construction sites based on the YOLO-ESCA architecture aims to revolutionize safety practices in high-risk environments. Construction sites are inherently hazardous, with workers, machinery, and equipment constantly in motion, making it critical to monitor the site for potential safety violations. The proposed system utilizes advanced computer vision, deep learning techniques, and natural language processing to ensure real-time safety monitoring, significantly improving compliance with safety regulations and reducing the likelihood of accidents. By leveraging a specialized YOLO-ESCA model designed for real-time object detection and embedding safety rule data, this system provides a comprehensive solution for detecting unsafe behaviours and environmental hazards.

At the core of this system lies a YOLO-ESCA model, a modified version of the popular YOLO (You Only Look Once) object detection framework, which is optimized for high-speed, accurate object detection in dynamic environments such as construction sites. The backbone network of the YOLO-ESCA model extracts multi-scale image features, capturing key aspects of the environment such as workers, machinery, tools, and potential hazards. The model's ability to process features at multiple scales ensures that it can detect small objects, like safety helmets or protective gear, as well as larger objects, such as vehicles or cranes, making it highly versatile and capable of adapting to various construction scenarios.

The neck network within YOLO-ESCA further refines these features through a series of up sample layers and convolutions, enabling the detection of objects across different resolutions. The multi-scale feature extraction mechanism ensures that the system remains sensitive to details critical for safety enforcement, such as whether workers are wearing the appropriate protective equipment or whether machinery is operating within defined safety zones. The model generates predictions in the form of bounding boxes and object classes, which are essential for identifying the location and nature of each detected object on the construction site.

One of the distinguishing aspects of this system is its integration of text-based safety rules as embedding vectors, enabling it to understand the context of detected objects in relation to site-specific safety regulations. These safety rules are embedded using Glove embeddings, a technique in natural language processing that represents textual data in vector form, allowing the system to interpret written rules as part of its decision-making process. This text data provides a crucial layer of understanding, allowing the system not only to detect objects but also to evaluate whether the detected scenario complies with safety standards.

To achieve this, the embedding vectors from the YOLO-ESCA model are concatenated with the safety rule embeddings, resulting in a combined representation that incorporates both visual and textual data. This fusion of visual features and contextual safety knowledge allows the system to make more informed predictions about safety violations. For instance, if the system detects a worker operating heavy machinery, it can cross-reference this with embedded safety rules to determine if the worker is adhering to proper protocols, such as maintaining a safe distance from others or wearing protective gear.

Once the visual and safety rule data have been combined, they pass through a series of fully connected layers that further refine the predictions. These layers output classifications for both object classes (such as workers, machinery, or tools) and attribute classes (such as worker attire or safety compliance). The system then generates alerts based on these classifications, drawing attention to any detected safety violations. For example, if a worker is identified without a hard hat or if machinery is operating in a restricted zone, the system immediately flags these conditions, allowing site managers to intervene and address the potential hazard before an accident occurs.

The system is trained on a custom dataset of construction site images, ensuring that it is tailored to the unique conditions and challenges of construction environments. By focusing on specific site attributes, such as machinery types, worker behaviour, and environmental conditions, the system becomes highly adept at

recognizing critical safety-related features. This customized training enables the AI system to handle diverse construction site layouts and operational workflows, ensuring scalability and adaptability across various types of projects.

Another key feature of the system is its ability to operate in real-time, providing continuous monitoring of site conditions and worker behaviour. Construction sites are fast-paced environments where conditions can change rapidly, and the ability to detect safety issues as they occur is essential. The real-time feedback offered by the YOLO-ESCA model allows construction managers to receive immediate notifications of any safety concerns, enabling quick decision-making and action. This proactive approach to safety management helps prevent accidents and fosters a culture of safety awareness on-site.

In addition to monitoring current conditions, the system also provides data-driven insights that can inform future safety practices. By analysing patterns of detected violations over time, site managers can identify areas where safety protocols may need to be strengthened or where additional worker training may be required. This not only improves safety outcomes but also enhances overall site efficiency by reducing downtime caused by accidents or non-compliance.

Implementation Plan :

Existing Approaches :

Literature Survey:

Chorowski J., Bahdanau D., Serdyuk D., Cho K., & Bengio Y. (2015) - Attention-Based Recurrent Neural Network (RNN) Model

- Merits:
  - The introduction of a location-aware attention mechanism enhances the model's ability to focus on relevant parts of the input, which improves the accuracy in speech recognition tasks.
  - Tested on the TIMIT Phoneme Recognition Dataset, a well-established dataset for benchmarking, showing its robustness in phoneme recognition.
  - The model significantly reduces the Phoneme Error Rate (PER), a critical metric in speech processing.
- Demerits:
  - The model might suffer from overfitting if the attention mechanism focuses too much on specific parts of the data, making it less generalizable to new datasets.
  - Requires substantial computational resources due to the complexity of the attention mechanism, limiting its scalability in low-resource environments.

Fang Q., Li H., Luo X., Ding L., Luo H., Rose T. M., & An W. (2018) - Faster R-CNN Model

- Merits:
  - Achieves high precision (90.4) and recall (91.5), demonstrating its effectiveness in detecting objects from large surveillance datasets (over 100,000 frames), making it highly applicable in construction site surveillance.
  - The model's ability to process large datasets ensures applicability in real-time safety monitoring scenarios.
- Demerits:
  - Faster R-CNN is computationally expensive, especially when applied to large datasets such as construction surveillance footage, which could hinder real-time performance.
  - The model may struggle with identifying small or partially obscured objects in cluttered environments, reducing its effectiveness in highly dynamic and complex scenes like construction sites.

Cho K., van Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., & Bengio Y. (2014) - RNN Encoder-Decoder Architecture

- Merits:
  - Introduction of a novel hidden unit improves the model's performance in sequence-to-sequence tasks, particularly in language translation, as evidenced by the BLEU score of 34.64 on the WMT'14 translation dataset.
  - Outperforms the baseline models, demonstrating significant improvements in handling long-range dependencies in translation tasks.
- Demerits:
  - Like many RNN-based models, the architecture struggles with processing long sequences effectively, leading to potential degradation in performance for very long inputs.
  - The reliance on the WMT'14 dataset for evaluation may limit the generalizability of the model across different languages or contexts outside translation.

Ben-Alon L., & Sacks R. (2017) - Agent-Based Simulation (ABS) combined with Building Information Modeling (BIM)

- Merits:
  - By integrating economic utility models, this paper effectively combines financial considerations with safety in high-rise residential construction, offering a comprehensive simulation framework for project planning.
  - Achieves a high F1-Score of 84.3, which suggests a good balance between precision and recall, making it suitable for simulating safety behavior in construction environments.
- Demerits:
  - The model might be too complex and computationally intensive for smaller projects, limiting its usability in smaller construction environments or less resource-rich scenarios.
  - The approach may face difficulties in capturing real-world unpredictability, such as sudden accidents or unmodeled behaviors, which are common in construction sites.

Waehrer G. M., Dong X. S., Miller T., Haile E., & Men Y. (2007) - Comprehensive Cost Model

- Merits:
  - Provides a thorough analysis of quality-of-life costs related to workplace safety, making it a valuable resource for policymakers and stakeholders aiming to improve safety measures in construction.
  - Uses a large dataset from the Bureau of Labor Statistics (BLS), which enhances the credibility and reliability of the findings.
- Demerits:
  - The focus on economic costs might overshadow other critical aspects of workplace safety, such as psychological impacts or long-term health outcomes, making the model somewhat limited in scope.
  - The accuracy of 80.5% may be insufficient for complex decision-making processes, especially in high-risk industries like construction, where more precise predictions are needed.

The paper introduces a semantic image-text matching approach using a Stacked Cross Attention Network (SCAN), which pairs images of unsafe behaviours with safety rules for improved precision. This method goes beyond models like Faster R-CNN by incorporating bidirectional GRU to efficiently embed safety rules and detect multiple unsafe behaviours in a single image. Unlike earlier models limited by smaller datasets, it uses pretrained models like MS-COCO to enhance generalizability and performance. The SCAN attention mechanism ensures reliable detection by focusing on both relevant image regions and safety rules.

1. Originally:
   - Existing Approaches:
     - Previous methods like Faster R-CNN (Fang et al., 2018) were primarily focused on detecting unsafe behaviours using handcrafted rule-based techniques or object detection models. While they effectively identified individual unsafe actions (e.g., lack of helmet), they didn't establish a connection between the detected behaviour and the underlying safety rule.
     - Models such as attention-based RNNs (Chorowski et al., 2015) or RNN Encoder-Decoder architectures (Cho et al., 2014) were used for tasks like phoneme recognition or translation but didn't address the challenge of combining visual and textual data in construction safety contexts.
   - Proposed Approach:
     - This proposal is original in that it uses a Stacked Cross Attention Network (SCAN) to match unsafe behaviours detected in images with semantic safety rules. The model aligns visual features from images with the text from safety rules, allowing for accurate detection of multiple behaviours and their related safety violations in one shot.
     - It employs a bidirectional GRU to embed safety rules, ensuring that the context of the rule is captured in both directions, enhancing the model's ability to match behaviours to complex safety rules.

2. Performance:
   - Existing Approaches:
     - Faster R-CNN and other similar models achieved high precision and recall in detecting objects but faced challenges in identifying multiple unsafe behaviours within the same image. They struggled in cluttered, real-world environments like construction sites, where detecting small, partially obscured objects is crucial.
     - Hand-crafted rule-based systems were rigid and lacked scalability, requiring extensive manual updates when new behaviours or rules were introduced. This limited their adaptability in dynamic, real-time monitoring.
   - Proposed Approach:
     - The proposed model shows a significant performance improvement by using cross-attention between image regions and safety rules. This enables it to detect multiple unsafe behaviours in real-time, addressing the major limitation of previous approaches.
     - Experimental results demonstrate that the model achieves up to 97% precision and 90% recall, particularly excelling in scenarios that involve detecting more than one unsafe action simultaneously. These improvements are due to the stacked attention mechanism, which focuses on both visual features and their alignment with safety rule text.
     - The model's ability to learn from large pretrained datasets (MS-COCO) also enhances its robustness, ensuring better generalization across various construction sites and environments, unlike previous models that were more dataset-dependent.

3. Costs/Benefits:
   - Existing Approaches:
     - Models like Faster R-CNN and handcrafted rule-based systems were computationally expensive, requiring significant processing power for real-time detection, particularly when applied to large datasets like surveillance footage from construction sites. The constant manual updates to rules and reprogramming made these systems labour-intensive and costly to maintain over time.
     - Systems using handcrafted rules were not scalable, as each new safety violation or environment required re-training or creating new detection models, which led to higher operational costs.
   - Proposed Approach:
     - The cost-benefit of this proposal is substantially better due to its scalability and automation. Once trained, the model can detect a range of unsafe behaviours without needing continuous manual updates or retraining. This drastically reduces labour costs associated with safety monitoring.
     - The use of pretrained models and efficient GRU-based text processing also lowers computational costs. The system can process safety rules and behaviours simultaneously, leading to faster detection and improved accuracy with reduced hardware requirements.
     - Long-term benefits include improved real-time detection, reducing the number of accidents and improving safety compliance, which lowers the financial and human costs associated with workplace accidents.

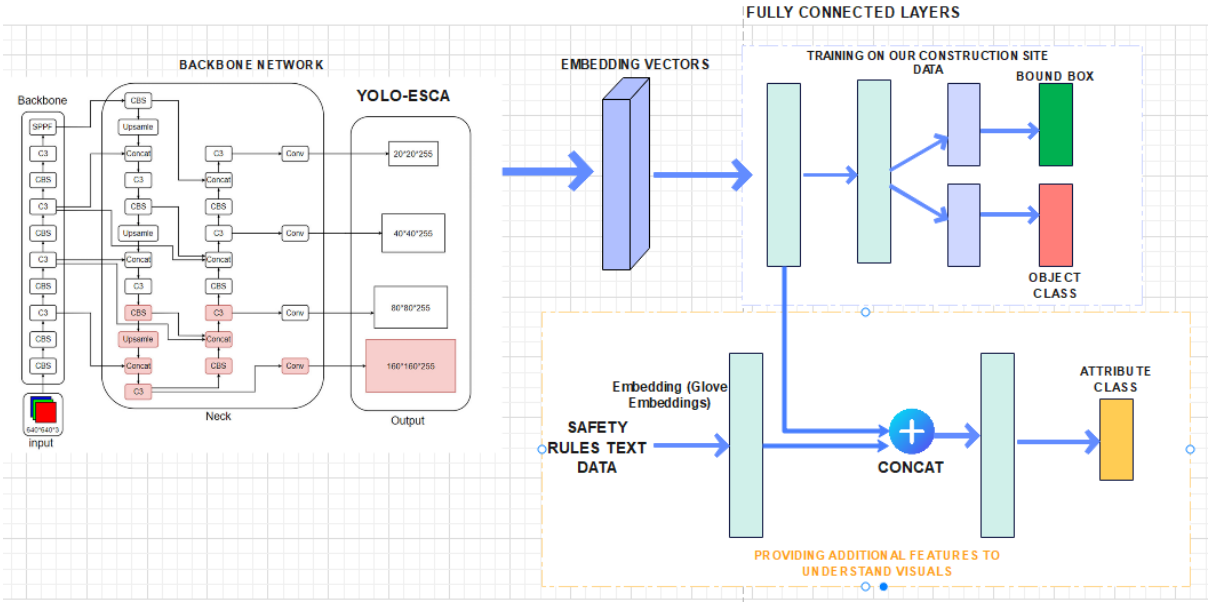
References:

Reference	Model	Dataset	Significant Factor	Evaluation Measures
Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015).	Attention-Based Recurrent Neural Network(RNN) model	TIMIT Phoneme Recognition Dataset	Location-aware attention mechanism	Phoneme Error Rate (PER)
Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T. M., & An, W. (2018)	Faster R-CNN (Region-Convolutional Neural Network)	More than 100,000 image frames from surveillance videos of construction sites	Location-aware attention mechanism	Precision: 90.4 Recall : 91.5
Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014)	RNN Encoder-Decoder architecture	WMT'14 translation Dataset	Novel hidden unit	BLEU scores of 34.64 on the test set compared to 33.30 for the baseline model
Ben-Alon, L., & Sacks, R. (2017)	Agent-Based Simulation (ABS) combined with Building Information Modeling (BIM)	High-rise residential construction dataset.	Economic utility model	F1-Score:84.3
Waehrer, G. M., Dong, X. S., Miller, T., Haile, E., & Men, Y. (2007).	Comprehensive cost model	the Bureau of Labor Statistics (BLS) dataset.	Quality of life costs	Accuracy:80.5

Approach to Solving the Problem:

1. Analysis:
   - The problem is focused on improving safety monitoring in construction environments through the automatic detection of unsafe behaviours (e.g., improper helmet usage) using computer vision and deep learning.
   - The challenge includes detecting multiple safety violations within a single image and associating those behaviours with textual safety rules.
   - From the second paper, we extend the analysis to include semantic matching between images and text-based safety rules.

2. Design:
   - The design integrates the YOLO-ESCA model for feature extraction from images (used for helmet detection) and a Stacked Cross Attention Network (SCAN) for matching detected behaviours with corresponding safety rules.
   - YOLO-ESCA uses EIoU loss, Soft-NMS, and CBAM attention to improve detection accuracy and target occlusion handling.
   - The GRU network is used for embedding and processing safety rules, improving context understanding.
   - The overall system follows an image-text matching pipeline, linking image regions to semantic safety rules using cross-attention mechanisms.



3. Experimental Setup:

- A supervised learning model is employed, with the YOLO-ESCA model pretrained on helmet datasets and GRU on safety rule texts.
- The system will be trained on image-text matching datasets, pairing images of unsafe behaviours with textual safety rules.
- Experiments will run on a server equipped with Nvidia A100 GPUs to handle the large dataset and the complexity of the cross-attention model.
- Evaluation metrics include mAP (mean average precision) and precision-recall for behaviour detection and rule matching.

4. Data Preparation:

- Image Dataset: High-resolution images of construction sites, focusing on helmet usage and general safety, with annotations for unsafe behaviours (e.g., No helmet, Helmet).
- Textual Dataset: Safety rules from official standards (e.g., Usage of machine to transport workers") for matching with image data.
- Annotation: Images are labelled with corresponding safety rules, and unsafe behaviours are categorized (helmet, no helmet).

5. Sample Sizes:

- Images: 193K images from the Multiple sources of MSCOCO, Peoples and ladders, Safety equipment's.
- Safety Rules: A dataset of eight safety rules (e.g Approaching dangerous zone without the usage of edge protection (i.e., roof, scaffold, holes, stairwell, elevator, and foundation pit).) paired with image data for rule matching and anomaly detection.

6. Consultation Details:

- Safety experts from construction industries are consulted to validate the unsafe behaviours and safety rules.
- Deep learning specialists provide feedback on model optimization and cross-attention network tuning.

7. Experimental Parameters:

Hardware Requirements:

- GPU: Nvidia RTX 3060 with 12GB memory for handling image and text data processing.
- Storage: 1TB External Hard Drive for storing datasets of annotated images and rule texts.

Training Parameters:

- Learning Rate: Optimized at 0.001, reducing over time using a scheduler.
- Batch Size: 64 images per batch to balance memory use and performance.
- Epochs: 50-100 epochs for convergence.

Evaluation Metrics:

- Precision, Recall, mAP for image behavior detection.
- Cosine similarity and attention weights to evaluate image-text rule matching accuracy.

Likely Problems That May Be Encountered

1. Data Quality:
    - Issue: Limited availability of high-quality images for training the model may lead to inaccurate predictions.
    - Impact: The model may not perform well in real-world scenarios.
2. Computational Resources:
    - Issue: Access to high-performance computers or GPUs may be limited, affecting the training time.
    - Impact: Longer training periods can delay project completion.
3. Integration Challenges:
    - Issue: Difficulty in integrating the new detection system with existing safety protocols on construction sites.
    - Impact: Resistance from users or stakeholders may lead to underutilization of the system.
4. Real-time Detection:
    - Issue: Achieving accurate real-time detection of unsafe behaviours might be technically challenging.
    - Impact: Delays in detection could undermine the effectiveness of the solution.

Reasons Why This Proposal Should Be Considered for Selection

1. Relevance to Safety:
    - This proposal addresses a critical issue in construction safety by providing an automated solution for detecting unsafe behaviours.
2. Innovative Technology:
    - The use of deep learning and image-text matching introduces a modern approach to safety monitoring, which is highly relevant in today's tech-driven environment.
3. Scalability:
    - The proposed system can be adapted to different construction sites, making it versatile and applicable in various contexts.
4. Practical Impact:
    - By improving safety compliance, this project can potentially reduce accidents, benefiting the workforce and the industry.

Budget Estimates:

1. GPU for Training:
    - Nvidia RTX 3060 or equivalent: ₹25,000 - ₹27,000
        - A capable GPU for training deep learning models, balancing performance and cost.
2. External Storage:
    - External Hard Drive (1TB): ₹5,000
        - For storing datasets, backups, and model outputs.

Declaration – 1

We will take up this project if selected, for FAER Scholar Project only and will not submit this to any other contest / institution. We will comply with FAER's requests on submission of action plans, summary, monthly progress reports. We will regularly discuss with mentors and follow his advice. Mentors are there to help us. We will follow the rules and regulations of FAER.

Name and
Signature of Students

Signature of Project Advisor / Project Guide

Date:

Signature of Principal with seal of the institution

Declaration – 2

We declare that the project work is not a copy or a purchased activity. We will do the project by ourselves with support from our faculty and use the facilities of the college. Wewill not involve in plagiarism. We will be sincere and take interest in executing and completing the project as proposed.

All statements stated are true to our knowledge.

Name and
Signature of Students

Signature of Project Advisor / Project Guide

Date:

Signature of Principal with seal of the institution