Machine learning for industrial safety culture in the developing world

Abstract

A disproportionate number of deaths and injuries in the workplace are concentrated in the developing world due to inadequate enforcement capabilities in such areas, a lack of capital equipment and a paucity of expertise. This requires solutions that enable automated detection of harmful violations of safety standards and sending appropriate warnings at real-time. We propose an automated method for highlighting such violations of safety measures by foregoing the usage of safety helmets, hard hats and similar safety equipment in the factory setting. A pipeline is proposed wherein camera feed at real-time is processed to detect the safety critical objects and their placement using a set of deep learning based architectures tailored for detection and localization under compute-constrained environments, with the helmet compliance sub-task treated as an exemplar of our approach.

1 Introduction

A key aspect of fostering socioeconomic development in the developing world has been the incentivization of industrial growth, with a particular focus on enabling revenue and employment generation through small and medium scale manufacturing. As with most activity centred on human-machine interaction, the production line is germane for mishaps arising due to faulty equipment, human error, unprecedented floor issues or a combination of similar factors. This implies a definite possibility of accidents arising in the factory floor and a consequent possibility of personnel being adversely affected in terms of their health and well-being. Thus, it is essential that a certain safety culture is continually maintained and extant rules and regulation concerning health and safety be consistently adhered to. This requires a continual process of observation and course correction of any breaches during operation, and can be done by qualified staff members on the shop floor and other operating areas. This is a difficult task requiring a significant investment in terms of employing qualified personnel who need to be vigilant. Often, an oversight over multiple areas are required and as such, multiple trained personnel need to be deployed. This creates a serious budgetary and operational constraint for factories and manufacturing units, particularly in the developing world. Industrial safety in the developing world has also been compromised due to insufficient regulatory heft, lack of expertise and poor equipment availability. At the same time, a sub-optimal safety culture in such settings is a leading cause of injury, impairment and long term damage to health and safety at the workplace in developing countries. This problem is particularly acute in small and medium scale manufacturing establishments and in safety critical factory segments where statutory regulations are often not followed due to a lack of enforcement. A possible solution to these challenges can arise from addressing the issue of inadequate manpower and augmenting the ability of supervisors to effectively enforce safety protocols such as wearing helmets at all times, wearing requisite safety gear like masks, vests and so on are adhered to at real time. This can be accomplished by continuously processing the video feed obtained from strategically located closed circuit cameras and with state-of-the-art object detection methods. In recent years, machine learning methods have revolutionized computer vision application including object detection and localization, along with corresponding applications in different settings in the wild [1]. Building over traditional methods of feature based object detection, deep learning methods were employed beginning with [2], R-CNN [3] which employed grid regions

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

at the level of input images, and each such region is classified for a given class using a pretrained classifier. This was followed by Fatser RCNN with a region proposal network [4], and a Single Shot Detector (SSD) [5] and YOLO with its various versions such as the YOLO v2[10], and the YOLO v3 [6] which were able to perform a single stage detection and localization.

Given the proliferation of closed circuit cameras, it is pertinent that such infarctions of safety be monitored at real time and suitable corrective measures immediately deployed. This is possible by integrating object detection based frameworks on the edge. We propose an automated method for highlighting such violations of safety measures by foregoing the usage of safety helmets and hard hats, masks, vests and similar safety equipment in the factory setting. A pipeline is proposed where camera feed at real-time is processed to detect the safety critical objects and their placement using a low memory footprint architecture with a YOLO based object detection framework. The key priorities considered were 1) accuracy of classification of safety aspect, and localization 2) compact architectures for deployment with low memory and compute requirements 3) validation over multiple factory environments for robustness to variable scene dynamics

2 Methodology

To start with, we explore the feasibility of performing identification and localization of safety related aspects in real-time detection on a video streams. We consider different classes of safety adherence pertaining to wearing safety helmets, where we define three classes of 'persons wearing helmets', 'persons not wearing helmets' and 'helmets' themselves, so that our detection and localization models can identify such situations from the video frames and draw a bounding box around the area of interest, with an additional feature that activates an audible alarm in case the class 'person without helmet' is detected in a camera feed from such a zone where wearing helmets is of critical importance. To look into model generalization, the evaluation was performed at different settings from a factory environment, with scenes from a shop floor, storage and logistics zones, aisles, maintenance activities, gate meetings and emergency evacuation. Data is curated from open source stock images relevant to each of the classes under study, and validation is performed on curated videos from different factory scenes sourced from across different industrial location in Low and Middle Income Countries (LMIC).

Safety aspect detection. To ensure the implementation of our detection and localization pipeline at a low memory budget and under compute-constrained environments for real-time inference, we rely on a backbone inspired by the SqueezeNet, MobileNet and MobileNet v2 architectures, with alpha = 0.25 for MobileNet and alpha = 0.35 set as a design choice based on optimal values of mAP and classification accuracies observed in a holdout set from curated video frames.These models are representative of the main types of compact architectures — squeeze-excite convolution blocks [7], group convolutions [8] and depthwise separable convolutions in groups [9]. Most other compact models proposed in computer vision literature derive from these basic architectures. The convolutional backbones are feature extractors adapted to work with a YOLO v2 [10] object detector and classifier. The benefit of using an YOLO based approach is that it is a single stage detector. In our case, the model requires to load a particular frame once and then a grid of regions in the frame produce five bounding boxes and a class confidence score for a given class. The overall confidence score for a given region is a combined score of the class confidence and the bounding box.

Implementation. The dataset consisted of a curated set of images from the three classes considered for the helmet compliance task aggregated from stock images with a total of 1262 frames consisting of 570 frames of persons wearing helmets, 410 of persons without helmets, and 282 frames of different variants of safety and crash helmets in various settings, orientations and illumination conditions. Data augmentation was applied with vertical and horizontal flips, and random cropping. Annotation of the frames was performed with class labels ('persons wearing helmets', 'persons not wearing helmets' and 'helmets'), and bounding boxes. The training for initial stages is performed over batches of 100 frames in 2000 epochs, with a learning rate of 0.001 and a logistic regression objective for bounding box regression along with a binary cross entropy loss term for the classification part (for determining whether or not a particular class is evident in a proposed region). The validation was performed on a validation set of images (from a 80:20 split), followed by further tests on the set of videos accummulalted from various scenes of industrial units. The ability to perform inference on a standard workstations points to significant savings in terms of operational and capital costs, as the only capital expenditure would then be to install a network of suitably placed closed circuit cameras, a local area

Table 1: Evolution of model performance over the first safety aspect task of helmet compliance. The average percentage classification accuracy for individual classes are presented

	Helmets	***Persons with helmets***	***Persons without helmets***
SqueezeNet + YOLO)	0.761	0.712	0.642
MobileNet + YOLO	0.792	0.751	0.720
MobileNetv2 + YOLO	0.827	0.804	0.825

Figure 1: (Detection of persons wearing and not wearing helmets in a large group in factory settings.

network (with a possibility of using a wireless network or a bluetooth based system instead) and the inference is possible without requiring any special modification. This is contrast to hiring and retaining specially qualified personnel and subsequent costs in retraining and allied aspects.

An Intersection over Union metric for the validation examples is used with the obtained bounding boxes versus the ground truth data. The mean IoU across classes is 0.61 for SqueezeNet backbone, 0.67 for the MobileNet backbone and 0.71 for the MobileNet v2 architectures on the localization task for the helmet compliance task. Classwise classification accuracies are shown in table 1. It is noted that the MobileNet v2 backbone enables better classification across classes along with more efficient localization. The implementation for the MobileNet v2 was found to be 6x faster than the SqueezeNet backbone detector and 3.5x faster than the one using MobileNet, demonstrating the efficiency gains achieved through group convolution based model optimization.

2.1 Future work

The proposed pipeline is a proof-of-concept towards using lightweight deep learning models for ensuring automated safety compliance ad alarm generation in the manufacturing and industrial sector of developing countries, by addressing key shortages in manpower and equipment with a real-time detection based approach for identifying aspects relevant to various safety compliance tasks using the helmet usage task as an exemplar of approach. Future work would look at expanding the technique to other aspects of safety compliance in an incremental learning setting and performing action recognition for directly categorising actions in the industrial workplace. It is hoped that the generalizability demonstrated in this approach is replicable in a variety of diverse situations with regards to health and safety monitoring in the workplace without constraints of significant capital and operational expenditure, which in turn would help reduce on the key causes behind a diminished quality of life that tends to afflict working age populations in many developing societies.

3 References

1. Zou, Zhengxia, et al. "Object Detection in 20 Years: A Survey." arXiv preprint arXiv:1905.05055 (2019).

2. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

3. Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

Figure 2: (Detection of persons wearing and not wearing helmets different work environments.

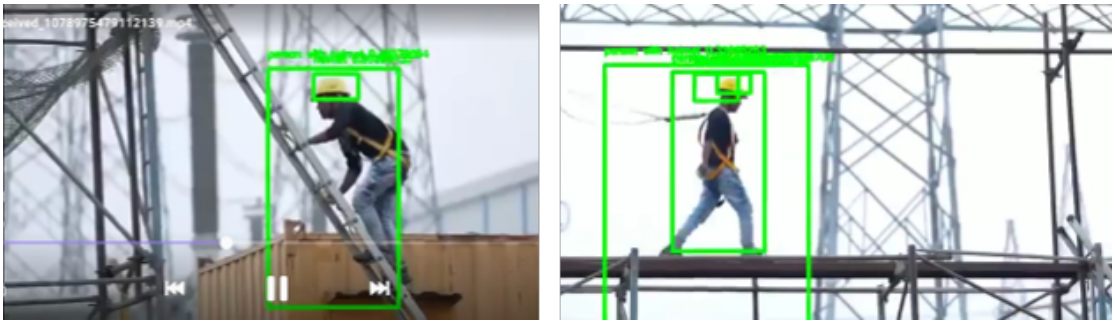

Figure 3: (Detection in outdoor gangways and ladders.

4. Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99)..

5. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

6. Redmon, J., Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767..

7. Iandola, Forrest N., et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size." arXiv preprint arXiv:1602.07360 (2016).

8. Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

9. Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

10. Redmon, J., Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).

4 Appendix

Qualitative examples for different industrial scenes are presented here.

Figure 4: (Detection in training and rescue.

Figure 5: (Detection for evacuation and firefighting scenes.