Detection and classification of threatening objects in baggage screening images using deep learning

Gabriel Ganzer[1]*; Gabrielle Maria Romeiro Lombardi[2]

[1] Computer Engineer. Rua José Vilagelim Neto – Taquaral; 13076-280 Campinas, São Paulo, Brazil
[2] PECEGE. PhD in Genetics and Plant Development. Rua Alexandre Herculano, 120 – Vila Monteiro; 13418-446 Piracicaba, São Paulo, Brazil
*corresponding author: gabriel.ganzer@icloud.com

Detection and classification of threatening objects in baggage screening images using deep learning

Summary

Security control is a mandatory protocol in today's airports. Passengers must undergo a rigorous baggage screening process to detect and eliminate prohibited items from their belongings before entering the boarding area. This critical operation involves X-ray machines closely monitored by qualified personnel. However, due to this task's labor-intensive and repetitive nature, there is an inherent risk of human error. This study focuses on training the state-of-the-art YOLOv4 model for precisely detecting and classifying threatening objects in baggage screening images. This work also explores two distinct activation functions that leverage the model's performance: the Leaky ReLU and Mish. This investigation aims to refine the training process by identifying potential improvements and trade-offs associated with each activation function. The results demonstrate the YOLOv4 model's exceptional performance, showcasing its ability to accurately detect and classify threatening objects in a dataset of baggage X-ray images with a precision exceeding 91% for both activation functions. Notably, the training process proved to be 1.5 times faster when employing the Leaky ReLU activation function compared to Mish, highlighting potential trade-offs with this approach.

Key-words: computer vision; object detection; baggage screening; convolutional neural network.

Introduction

A single event has changed the way passengers travel by air internationally. The terrorist attacks of September 11, 2001, on both the World Trade Center in New York City and the Pentagon in Washington DC, led the federal government of the United States to emit the Aviation and Transportation Security Act (ATSA), which required screeners to inspect airline passengers, all transported luggage as well as air cargo (Pekoske, 2021). Soon enough, security procedures became mandatory in all airports of several countries, for both domestic and international flights, as currently, almost all passengers must undergo additional control before entering a "secured area," which includes a hands-off screening of carry-on baggage using X-ray or Computed Tomography (CT) scanners, walk-through metal detectors to screen passengers, and pat-down screening when necessary (TSA, 2023).

Although international regulatory organizations enforced these restrictions after September 11, terrorist attacks were not rare throughout the world. In Brazil, an incident in 1988 has marked the national aviation history. Raimundo Nonato Alves da Conceição boarded flight 375 from the now-extinct company Vasp flying from Porto Velho to Rio de Janeiro. When descending to its final destination, Raimundo hijacked the airplane and diverted its course to the federal capital of Brasília. He had planned to crash the aircraft to the government office, the Palácio do Planalto. His motive was to kill José Sarney, the Brazilian president at that time. Raimundo had lost all his savings due to hyperinflation, as he was also discontent with the unemployment rates, blaming the presidential acts. Without fuel and visibility to land in Brasília,

the pilots received instructions to land the airplane in Goiânia, a nearby city. As the flight landed, Raimundo required another aircraft, but the Federal Police immediately arrested him during this transfer. The flight was transporting 105 passengers, and Raimundo was able to enter the airplane carrying a gun in his backpack as there were no security measures to identify the threatening object in his luggage at that time, which could have prevented the incident (Brant, 2018).

The baggage screening process involves monitoring X-ray machines by operators trained to identify prohibited objects among the passengers' belongings in images projected on a screen using a color pattern based on different materials' density and atomic heaviness (Seyfi, 2023). The International Civil Aviation Organization (ICAO) states that the human factor is the leading cause of inefficiency in this monotonous and labor-intensive task. Operators generally lose their sense of vigilance after 20-30 minutes due to distraction or fatigue (ICAO, 2018). Furthermore, the operator's experience is critical in image classification, and finding skilled personnel to perform this role could also be quite challenging.

The Dangerous Goods Panel Working Group of the Whole, which met in 2006 in China, established the categories of the most dangerous items for aviation security: firearms, guns, and weapons; explosive and flammable substances; pointed/edged weapons and sharp objects; and blunt instruments (DGP, 2006). A quick assessment of these items is crucial to guarantee the safety of passengers, crew, and staff.

Advances in machine learning algorithms permitted the introduction of innovative solutions to aid X-ray machine operators in their tasks, particularly those in deep learning using a Convolutional Neural Network (CNN) to identify and classify objects in images (Akcay, 2022). A CNN is a supervised method inspired by the human learning process and consists of an artificial neural network that deals with images as inputs (Sakib, 2019). Its architecture differs from other approaches as it uses properties from the 2D structure of an image to extract its characteristics. Akçay et al. (2016) was one of the first studies that examined the use of a CNN on classifying images, where the model predicted a global image label (gun vs. no-gun). In the same year, Redmond et al. (2016) published the You Only Look Once (YOLO) network, which performed a multi-label classification of objects directly on an entire image with high performance. Since then, researchers highly recommend the YOLO model for weapon detection (Narejo, 2021), smart surveillance (Oguine, 2022), and recognition of objects in X-ray images (Wang, 2022).

This study aims to train and validate the YOLO model to detect and classify objects in x-ray images of carry-on baggage screening into five classes: gun, knife, pliers, scissors, and wrench.

Material and Methods

You Only Look Once (YOLO)

The neural networks in object detection often use a convolutional network to divide the image into regions and propose potential bounding boxes to feed general-purpose classifiers that perform the object inference (Girshick et al., 2014). The YOLO model differs from this conventional approach by using a single convolutional network that simultaneously predicts multiple bounding boxes and class probabilities for those boxes, as seen in Figure 1. The network looks for features in the entire image to generate the bounding boxes for all classes of objects simultaneously, i.e., it learns to generalize representations of objects (Redmond et al., 2016). Because of that, the YOLO model is faster and smaller than models implementing other approaches, with the drawback of misleading detection of small objects and localization errors.

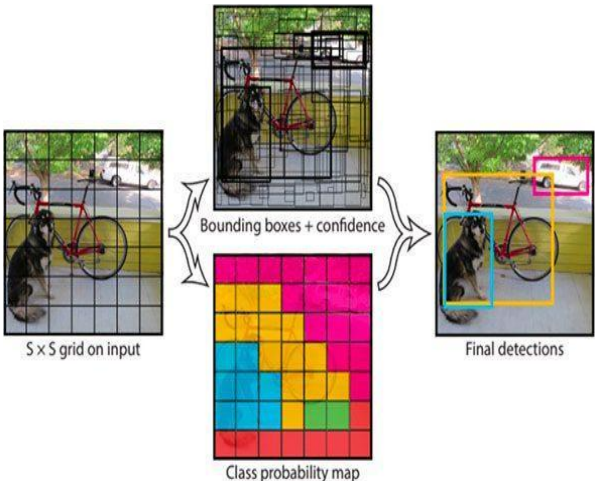

Figure 1. Object detection in YOLO
Source: You only look once: Unified, real-time object detection Redmond et al. (2016)

The architecture of the YOLO network consists of 24 convolutional layers followed by two fully connected layers and alternating 1 × 1 convolutional layers followed by 3 × 3 convolutional layers to reduce the feature space from preceding ones, as depicted in Figure 2. The final layer predicts both class probabilities and bounding box coordinates. The output is a 7 × 7 × 30 tensor of predictions. These layers were pre-trained with the ImageNet 1000-class competition dataset developed by Russakovsky et al. (2015).

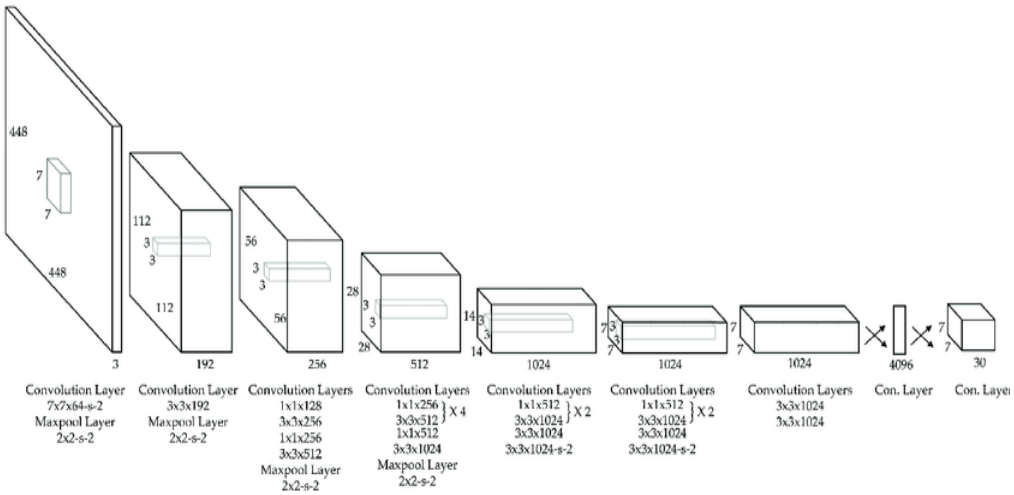

Figure 2. YOLO architecture

Source: You only look once: Unified, real-time object detection Redmond et al. (2016)

The authors Wang et al. (2021) took a step further and scaled the YOLO model to perform real-time object detection in general GPUs, introducing the YOLOv4 version. This version is faster than the previous ones. It presents higher accuracy mainly due to incorporating cross-stage partial connections, which reduced the number of computations, and the introduction of the Leaky Rectified Linear Unit (Leaky ReLU) and Mish activation functions. The authors published YOLOv4 under an open-source license; this is the version used throughout the experiments in this research.

Activation function

Computer vision tasks such as object detection have a high computational cost and often require a Graphics Processing Unit (GPU) or a Tensor Processing Unit (TPU) as hardware accelerators to increase the speed. Furthermore, according to Doherty et al. (2022), in a neural network like YOLO, the activation functions are crucial to balancing the computational cost and the model's accuracy.

The Leaky ReLU activation function is a modified version of the ReLU activation function that adds a slight slope from negative values instead of a flat one, as demonstrated in eq. (1) and eq. (2). This subtle difference allows the neural network to reach convergence faster than the regular ReLU with the counter effect of moderately lower accuracy (Maas et al., 2013).

$$f(x)_{ReLU} = max\,(0, x) \tag{1}$$

$$f(x)_{LeakyReLU} = \{0.01xxforx < 0forx \geq 0 \tag{2}$$

Misra (2019) introduced the Mish, a novel self-regularized non-monotonic activation function. Like Leaky ReLU, it also enables small negative values but adds non-monotonicity,

leading to a smooth profile, as shown in eq. (3) . This provides a better generalization of the activation function, which increases accuracy and lowers computational costs.

$$f(x)_{Mish} = x * tanh\left(ln(1 + ex)\right) \quad (3)$$

Figure 3 compares the curves from the Leaky ReLU and Mish activation functions. The study explores the impact of these functions on training the YOLOv4 model through two distinct experiments.

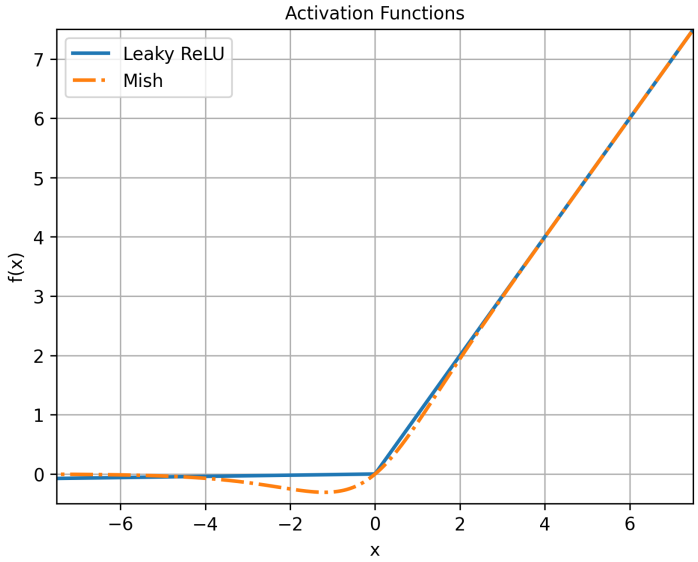

Figure 3. Leaky ReLU vs. Mish
Source: Original Research Results (2024)

Dataset

The dataset often dictates the success of a model in image recognition, playing a fundamental role in the final accuracy (Paullada, 2021). Therefore, this study used the SIXray dataset presented by Miao et al. (2019), which contains 1,059,231 X-ray images of baggage screening captured from surveillance equipment. The dataset was preprocessed by selecting only the images displaying at least one object belonging to one of the five classes of the experiment. This custom dataset contained 61,663 images, from which 55,497, or approximately 90% of the dataset, were randomly picked to train the model. The remaining 6,166, or about 10% of the dataset, was kept for validation. The images were subsequently annotated to the YOLOv4 format with the aid of labelImg (Vostrikov, 2019). This process consisted of drawing bounding boxes around the target object and annotating it with their respective label, as shown in Figure 4.

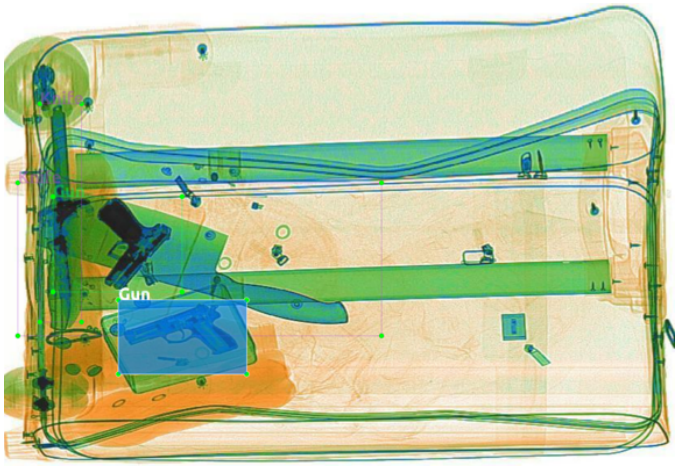

Figure 4. Sample of an annotated image from the custom dataset. The image depicts two guns and two knives with their respective bounding boxes and annotations
Source: SIXRay dataset Miao et al. (2019)

Training and validation

The advantage of using pre-trained weights to fine-tune a model to a custom dataset is that it utilizes a transfer learning approach to adapt the neural network from a large dataset to a specific goal, thus, offering a speed up to the training process Gupta, Neeraj (2021). The weights provided by the authors Wang et al. (2021) were used in the experiments as a starting base for training. These weights achieved an average precision of 57.9% on the COCO test-dev dataset Lin et al. (2014). The neural network configuration was modified for the custom dataset, i.e., the number of filters and the final layer were adapted for the output of five classes.

The model's training was performed on a machine with CUDA capability for speed up and equipped with an NVIDIA GeForce® RTX 3090 GPU. Redmond et al. (2016) recommend training the model for the number of classes * 2000 iterations, but not less than the number of images and not less than 6000 iterations. Given the five classes and the quantity of images, the training ran for 10000 iterations. The training script performs a validation of the weights after every 1000 iterations and it computes the following metrics (Wang et al., 2021):

1. The mean Average Precision (mAP) that takes into account the true positives (TP) and true negatives (TN) at different levels of confidence, with higher values indicating better object detection performance;

2. The Intersection over Union (IoU) measures the overlap between the objects detected by the model and the coordinates in the annotations, with 0.5 as the goal value;

3. The recall and precision are also calculated at this stage and combined in an F1-Score metric as shown in eq. (4). These values range from 0 to 1, with 0 indicating poor performance and 1 indicating overfitting;

$$F1_{score} = 2\frac{precision*recall}{precision+recall} \quad (4)$$

4. False positives (FP) and false negatives (FN).

The validation function plots the mAP and the average loss achieved at each stage on a chart. The training script saves the weights in a backup folder for traceability at every 1000 iterations. The final weights are those that performed the best mAP throughout the entire training process.

Results and Discussion

The fine-tuned YOLOv4 model effectively detected and classified objects within the designated custom classes for both activation functions. Figures 5 and 6 illustrate a consistent improvement in the mean average precision after every 1000 iterations in both scenarios, culminating in peak performance towards the conclusion of training. As described in the work of Maas et al. (2013), the training employing the Leaky ReLU activation function yielded slightly lower precision compared to the Mish activation function, although converging more rapidly.

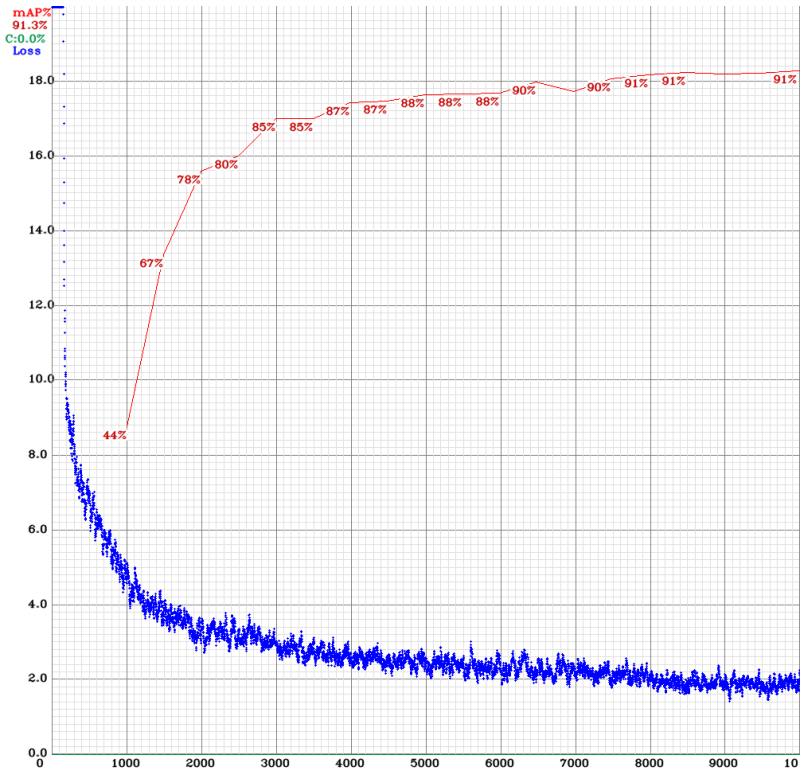

Figure 5. Validation chart for custom YOLO model trained with the Leaky ReLU activation function.
Source: Original Research Results (2024)

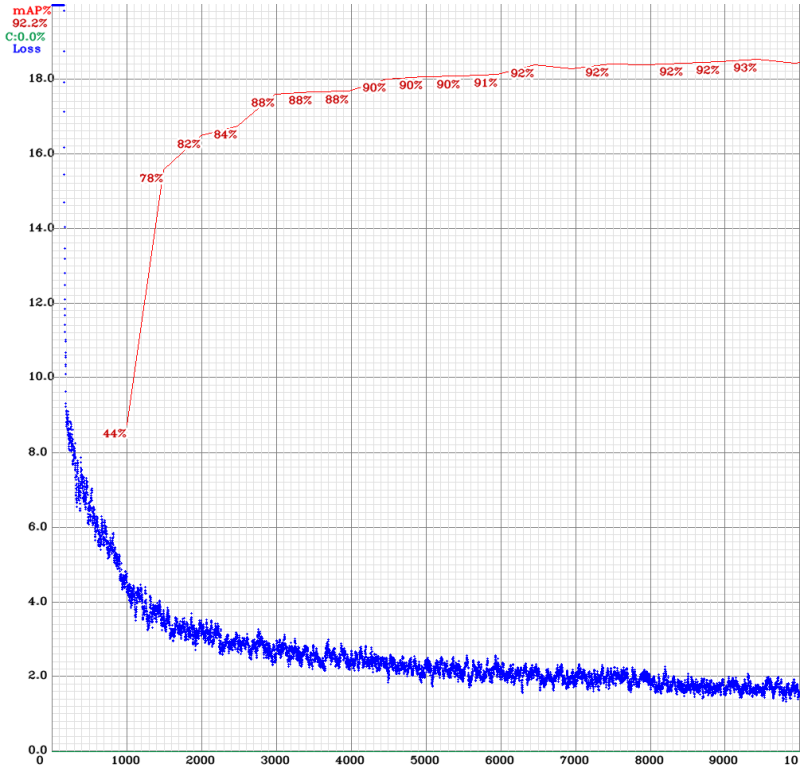

Figure 6. Validation chart for custom YOLO model trained with the Mish activation function. Source: Original Research Results (2024)

The model exhibited consistent detection performance across both experiments, effectively identifying each of the five objects. Russakovsky et al. (2015) acknowledge YOLOv4's limitations in detecting small objects clustered together or densely distributed within a scene. As shown in Table 1, the class "gun" achieved the highest average precision. The distinct shape and metallic density of guns often ensure a color contrast in images, facilitating accurate identification. On the contrary, lower precision results can be observed for "pliers", "scissors", and "wrench", items frequently packed in groups by passengers. Additionally, as highlighted by Diwan et al. (2023), the imbalance in instances among different object classes, notably for "knife", "scissors", and "wrench", can impact the model performance.

Table 1. Average precision on detection of each class of object for both activation functions

Class	Leaky ReLU average precision (%)	Mish average precision (%)	No. of instances
Gun	98.27	98.26	4392
Knife	84.18	85.50	2735
Pliers	92.25	94.06	4826
Scissors	90.66	94.44	1042
Wrench	91.07	90.00	2779

Source: Original Research Results (2024)

Figure 7 showcases the output of the YOLOv4 model, trained with the Mish activation function, directly applied to the sample images used as inputs. The model accurately detected and classified all proposed object classes with a high confidence. It's noteworthy that the sample images have different resolutions and portray various objects beyond those included in the model's training classes, yet the model disregards these extraneous objects in its output despite Wei et al. (2014) and Diwan et al. (2023) observations that image detectors generally underperform for inputs having different scales and with objects located at various positions and poses.

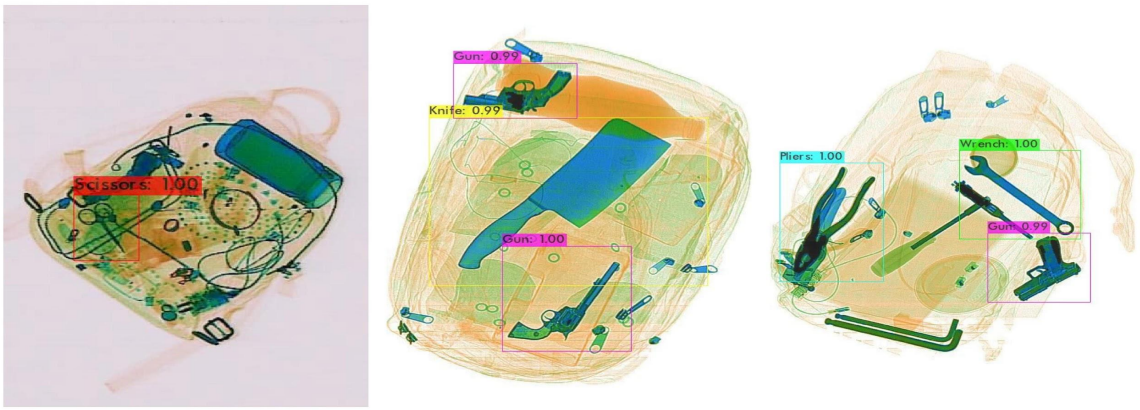

Figure 7. An output image of the YOLOv4 model trained with the Mish activation function and used in detection mode. Objects from all five classes are highlighted with their respective levels of confidence.
Source: Original Research Results (2024)

Table 2 presents a comparative analysis of the final metrics obtained from both training experiments. The mean average precision values are computed across various IoU thresholds, ranging from 0.5 to 0.95, providing an evaluation of the model's performance across different confidence levels.

Table 2. Overall results achieved with the custom YOLO model in both activation functions

Metric	Leaky ReLU	Mish
mAP(IoU=0.5)	91.29%	92.57%
Average IoU	75.87%	74.87%
Precision	0.92	0.91
Recall	0.88	0.90
F1-Score	0.90	0.90
Time training	5.52 hours	8.38 hours

Source: Original Research Results (2024)

In the study conducted by Wang et al. (2022), a YOLOv4 model was trained specifically to detect a single class, "gun," in X-ray images, achieving a mAP of 90.62% at an IoU threshold of 0.5. Addressing the complexity of multi-label image classification, Wei et al. (2014) highlight its challenges compared to single-label image classification, as the latter is often applied to aligned images with a clear depiction of objects in the foreground. As Table 2 illustrates, the

outcomes of training the YOLOv4 model for multi-label X-ray image classification with the Leaky ReLU activation function yielded a slightly higher mAP of 91.29% at an IoU threshold of 0.5, while the Mish activation function showcased even superior performance with a mAP of 92.57%. These results validate the efficacy of both models in addressing the specified task.

The average IoU scores suggest that in both experiments the model effectively overlapped its bounding boxes with those provided in the annotations across the majority of images from the dataset. As noted by Diwan et al. (2023), localization errors often arise from background pixel occupancy within bounding boxes or detecting objects with similar characteristics, a challenge pertinent to the object categories in this study. This observation is further reinforced by the F1-Score of 0.9 in both experiments, affirming the model's effectiveness without overfitting.

Finally, one notable advantage of employing the Leaky ReLU activation function becomes evident in the training time, which was faster compared to Mish's activation and without any significant drawbacks to the overall performance of the model.

Final Considerations

The training of YOLOv4 on the specified classes of objects has yielded successful results. The model achieved outstanding precision with both activation functions. The Mish activation function exhibited slightly superior performance. At the same time, the version applying Leaky ReLU finished the training process 1.5 times faster with minimal counter effects in the accuracy, a favorable trade-off given that this difference in precision was less than 1.28%. Overall, the findings of this research confirm that deep learning techniques are a suitable approach to effectively detecting and classifying potentially hazardous objects during baggage screening in airports.

References

Akçay, S. and Breckon, T. 2022. Towards automatic threat detection: A survey of advances of deep learning within X-ray security imaging. Pattern Recognition, 122, p.108245.

Akçay, S., Kundegorski, M.E., Devereux, M. and Breckon, T.P. 2016, September. Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery. In 2016 IEEE International Conference on Image Processing (ICIP) (pp. 1057-1061). IEEE.

Brant, Ana Clara. 2018. A história do Boeing que seria jogado no Palácio do Planalto. Estado de Minas. Política. Available at: <https://www.em.com.br/app/noticia/politica/2018/09/28/interna_politica,992438/o-dia-emque-um-aviao-seria-jogado-sobre-o-palacio-do-planalto.shtml>. Acessed on: Oct 01 2023.

DGP. 2006. Report of the meeting of the working group of the whole. Dangerous Goods Panel (DGP) meeting of the Working Group of the Whole. Beijing, China.

Diwan, T., Anirudh, G. and Tembhurne, J.V., 2023. Object detection using YOLO: Challenges, architectural successors, datasets and applications. multimedia Tools and Applications, 82(6), pp.9243-9275.

Doherty, J., Gardiner, B., Kerr, E., Siddique, N. and Manvi, S.S., 2022. Comparative Study of Activation Functions and Their Impact on the YOLOv5 Object Detection Model. In International Conference on Pattern Recognition and Artificial Intelligence (pp. 40-52). Cham: Springer International Publishing.

Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

Gupta, Neeraj. 2021. A Pre-Trained Vs. Fine-Tuning Methodology in Transfer Learning. In Journal of Physics: Conference Series (Vol. 1947, No. 1, p. 012028). IOP Publishing.

ICAO. 2018. Innovative Solutions applied for automation of X-ray machine operator's state and performance monitoring, and Enhancement of their job functions. Agenda Item 3: Global Aviation Security Plan. Montreal, QC, Canada.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.

Maas, A.L., Hannun, A.Y. and Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In Proc. icml (Vol. 30, No. 1, p. 3).

Miao, C., Xie, L., Wan, F., Su, C., Liu, H., Jiao, J. and Ye, Q. 2019. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2119-2128).

Misra, D., 2019. Mish: A self regularized non-monotonic activation function. arXiv preprint arXiv:1908.08681.

Narejo, S., Pandey, B., Esenarro Vargas, D., Rodriguez, C. and Anjum, M.R. 2021. Weapon detection using YOLO V3 for smart surveillance system. Mathematical Problems in Engineering, 2021, pp.1-9.

Oguine, K.J., Oguine, O.C. and Bisallah, H.I. 2022, November. YOLO v3: Visual and Real-Time Object Detection Model for Smart Surveillance Systems (3s). In 2022 5th Information Technology for Education and Development (ITED) (pp. 1-8). IEEE.

Paullada, A., Raji, I.D., Bender, E.M., Denton, E. and Hanna, A. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns, 2(11).

Pekoske, David P. 2021. Security20 Years After 9/11: The State of the Transportation Security Administration. House Committee on Homeland. Washington, DC, United States of America. Available at: < https://www.tsa.gov/news/press/testimony/2021/09/29/20-years-after-911-state-transportation-security-administration>. Accessed on: October 01 2023.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C. 2015. Imagenet large scale visual recognition challenge. International journal of computer vision, 115, pp.211-252.

Sakib, S., Ahmed, N., Kabir, A.J. and Ahmed, H. 2019. An overview of convolutional neural network: Its architecture and applications. Preprints 2018, 2018110546

Seyfi, G., Esme, E., Yilmaz, M. and Kiran, M.S., 2023. A literature review on deep learning algorithms for analysis of X-ray images. International Journal of Machine Learning and Cybernetics, pp.1-17.

TSA. 2023. Security Screening. Transportation Security Administration. Springfield, VA, United States of America. Available at: < https://www.tsa.gov/travel/security-screening>. Accessed on: October 01 2023.

Vostrikov, A. and Chernyshev, S. 2019, June. Training sample generation software. In Intelligent Decision Technologies 2019: Proceedings of the 11th KES International Conference on Intelligent Decision Technologies (KES-IDT 2019), Volume 2 (pp. 145-151). Singapore: Springer Singapore

Wang, C.Y., Bochkovskiy, A. and Liao, H.Y.M., 2021. Scaled-yolov4: Scaling cross stage partial network. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition (pp. 13029-13038).

Wang, M., Yang, B., Wang, X., Yang, C., Xu, J., Mu, B., Xiong, K. and Li, Y. 2022. YOLO-T: multitarget intelligent recognition method for x-ray images based on the YOLO and transformer models. Applied Sciences, 12(22), p.11848.

Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y. and Yan, S., 2014. CNN: Single-label to multi-label. arXiv preprint arXiv:1406.5726.