BNL-207942-2018-COPA

Using Deep Machine Learning to Conduct Object-Based Identification and Motion Detection on Safeguards Video Surveillance

Y. Cui,

Submitted to the Symposium on International Safeguards: Building Future Safeguards Capabilities Conference
to be held at Vienna, Austria
November 05 - 08, 2018

Nonproliferation and National Security Department

Brookhaven National Laboratory

U.S. Department of Energy

USDOE National Nuclear Security Administration (NNSA), Office of Nonproliferation and Verification Research and Development (NA-22)

DISCLAIMER

USING DEEP MACHINE LEARNING TO CONDUCT OBJECT-BASED IDENTIFICATION AND MOTION DETECTION ON SAFEGUARDS VIDEO SURVEILLANCE

Y. CUI
Brookhaven National Laboratory, Upton, USA
Email: ycui@bnl.gov

Z.N. GASTELUM[2], R. REN[1], M.R. SMITH[2], Y. LIN[1], M.A. THOMAS[2], S. YOO[1], W. STERN[1]
[1]Brookhaven National Laboratory, Upton, USA
[2]Sandia National Laboratories, Albuquerque, USA

Abstract

Video surveillance is one of the core monitoring technologies used by the International Atomic Energy Agency (IAEA) Department of Safeguards at safeguarded nuclear facilities worldwide. Current IAEA image-review software has functions for scene-change detection, black image detection and missing scene analysis, but their capabilities are not optimum. The current workflow for the detection of safeguards relevant events heavily depends on inspectors' laborious visual examination of surveillance videos, which is a time-consuming process and prone to errors. To improve the accuracy of the process and reduce inspectors' burden, the paper proposes using deep machine learning to detect objects of interest in video streams and to conduct object-based motion detection. The hypothesis of this work is that deep machine learning will reduce the burden on inspectors and reduce errors by automatically locating and identifying objects and activities of interest in video streams. Objects of interest include casks and fuel assemblies that are typically monitored by inspectors. The algorithm being developed in this work is based on a computationally efficient deep machine learning algorithm – You Only Look Once (YOLO) – but is further devised to address specific challenges related to the operation of nuclear facilities. The developed model (which is called YOLO-SG – YOLO for Nuclear Safeguards) is evaluated with data sets collected at the test facilities at Brookhaven National Laboratory (BNL) and Sandia National Laboratories (SNL). The initial focus of the research is for application at safeguarded nuclear reactors, such as pressurized heavy-water reactors, where video surveillance is broadly deployed, but can be extended to other use cases of nuclear safeguards. The detailed structure of YOLO-SG model is introduced and the test results are reported in the paper.

1. INTRODUCTION

Containment and surveillance (C/S) are important measures in the implementation of the International Atomic Energy Agency (IAEA) safeguards. Video surveillance systems monitor and record activities at nuclear facilities and provides visual evidence of events to support the drawing of safeguards conclusions [1]. During review of surveillance video data, inspectors review all the footage with specific focuses, e.g. general checking, looking for and examining specific events, and investigating anomalous events recorded by other safeguards sensors or equipment. However, the video surveillance data covers not only safeguards-relevant activities, but also daily operation of safeguarded facilities. The latter doesn't involve movement of objects of interest, but is a large portion of the video surveillance data set, which makes the video-review process time consuming especially when the review software cannot help filter out these irrelevant video clips. Although the IAEA has been upgrading its video surveillance system hardware and the Next Generation Surveillance System (NGSS) has incorporated the newly designed digital surveillance cameras, surveillance video review still depends on the General Advanced Review Station Software (GARS). Despite the recent improvements in GARS to enable the review of records produced by the NGSS, the software has limited functions to automate the review process. To improve the efficiency of surveillance video review, advanced features like object-based scene change detection are required.

The paper proposes using machine learning algorithms to improve the efficiency of the surveillance video review process. Specifically, the focus of this research is to develop a machine learning algorithm for object detection and object-based motion detection. Currently, machine learning algorithms are used in many domains with large amounts of data where it is not feasible for a human to analyse all of the data. In computer vision, deep machine learning algorithms have reached better than human performance in detecting and locating objects in images. By applying these machine learning techniques to surveillance video review, an inspector can more easily

review the video with respect to the objects of interest and, hence, the efficiency of video review process can be improved significantly.

Although object detection algorithms have been developed in general, applying the technique to video surveillance in an IAEA safeguards environment is still challenging. First, the setup of nuclear facilities is complex and varies from one facility to another. Although the objects of interest may be similar, the background of videos/images can be very different. Additionally, the objects themselves may also vary in size, colour and shape. Second, in general, performance of machine learning algorithms depends on the quality of the training data set from which it learns. A more representative data set containing a larger number of images with different views of object of interest results in better trained models. However, the number and quality of images of objects of interest at a nuclear facility may be limited due to time constraint, physical access or the operator's proprietary information concerns. Hence, the available training set is small. Third, ideally the algorithms for this specific application should be simple and not need high computational power in order to allow execution of the code on inspectors' laptop computers, and enable field deployment and quick execution in video analysis. Many current algorithms require significant computational power such as graphical-processing unit (GPU) optimized machines, which may not be accessible in field, or cloud computing, which may be limited due to information protection protocols.

In the scope of this research, a machine learning algorithm is being developed focusing on object detection, which is critical to addressing the above challenges. This algorithm is based on You-Only-Look-Once (YOLO) machine learning model and is referred as YOLO for Nuclear Safeguards or YOLO-SG. Section 2 provides a high-level background of deep machine learning in object detection. Section 3 discusses the details of YOLO-SG, and Section 4 presents the test results of YOLO-SG.

2. DEEP MACHINE LEARNING FOR OBJECT DETECTION

Object detection (does an image contain an object?) and localization (where in an image is an object?) have been widely studied in computer vision and machine learning. Initially, features were manually extracted from an image edge and contour detection. These features were often discovered using a convolution, which is a mathematical operator that is essentially a function that looks for a given feature in a region of the image. The convolution is examined in all areas of the image. These features are then provided as input to a machine learning algorithm. One type of machine learning algorithm used was a neural network. A neural network is inspired by neural processing in the brain where nodes representing neurons are arranged in layers. Typically, neural networks contain three layers: an input layer, a hidden layer and an output layer. This approach was successful but two major modifications in neural networks have led to significant improvements in their performance in object detection and localization. First, using multiple hidden layers has significantly improved the performance of neural networks and is referred to as deep learning or deep neural networks due to a "deeper" neural network. Second, convolutions were encoded in the neural network such that rather than hand crafting the convolutions, they were learned by the neural network. These types of neural networks are deep neural networks and are called convolutional neural networks (CNNs).

CNNs have achieved better than human performance and several benchmark data sets, such as COCO [2][3]. However, CNNs often have very high computational cost in training and execution. The You Only Look Once (YOLO) model [4] was designed to speed up the detection and localization of objects in images. Prior to YOLO, other methods [5] - [7] would create several (on the order of thousands) proposals boxes (suggestions that an object might be in that region). All the region proposals were then passed through a trained deep neural network to extract features and then a classifier was trained to determine if an object is in that region of the image. YOLO treats localization and classification simultaneously. It uses a regression model on the feature maps and directly obtains bounding box location, size and class scores.

YOLO provides fast execution times. However, as with most deep learning approaches, training times can be prohibitively large, require large amounts of training data on the order of 1000's to 10,000's of images, and setting the correct parameters for the algorithms (the number of layers, number of nodes per layer, type of nodes, etc.) is difficult. Transfer learning is an effective approach to train models using a smaller data set [8]. Transfer learning is based on the premise that the underlying features in images (edges, contours, shapes, etc.) can be shared amongst different tasks. Transfer learning uses a deep neural network model trained on a different task (e.g. detecting birds) with a large number of annotated samples. In deep neural networks, the last layer is for classifying

the objects. Every previous layer can be viewed as extracting features. Transfer learning exploits this fact by retraining the final layers using a new task (e.g. detecting safeguard relevant objects) that may have only a few annotated examples. Since the common lower level features can be shared, the number of annotated examples can be small. Thus, even with few images of the objects, a new model can be trained efficiently.

YOLO and transfer learning provide the foundation for addressing the limitations of working within the constraints of IAEA-inspected facilities. These are that nuclear facilities and objects of interest vary from one facility to another, there is limited available training data, and that an efficient algorithm should allow execution on an inspector's laptop for real-time processing.

3. YOLO FOR NUCLEAR SAFEGUARDS

Building on the foundation principles in Section 2, our algorithm uses a CNN to provide state-of-the-art image processing capability, YOLO (a type of an architecture of CNN) for fast execution of object identification, and transfer learning to deal with few training examples and to speed up training times, which the team calls YOLO-SG. More detail of YOLO-SG is described below.

3.1. YOLO Model

For the object detection task of the project, YOLOv3 model was chosen as the starting point due to its efficiency and high accuracy [9]. YOLOv3 has incorporated many state-of-the-art techniques, such as residual blocks [10], anchor box [7], fully convolutional network [11] and pyramid design [12], which have been proven to be effective for multi-scale detections. Currently, YOLOv3 is state-of-the-art in object detection and in the speed in which it can detect and localize the objects. The key point is its efficiency without lack of performance as shown in FIG. 1. While other methods do achieve slightly better performance (FPN FRCN), that slightly better gain in performance comes at an inflated cost in inference time. More details of YOLO can be found in the appendix.

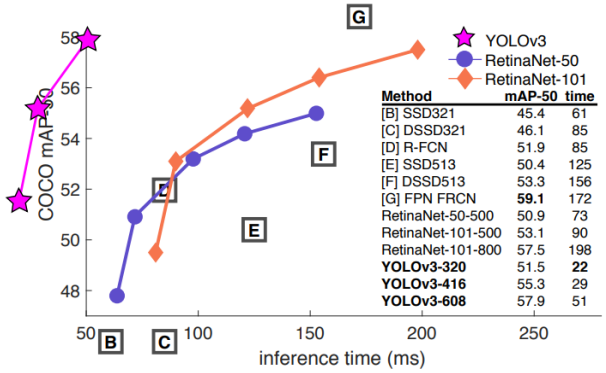


FIG. 1. Inference time vs performance. Lower time (to the left) is better and high performance is better. YOLO is by far the fastest and still achieves good results. [9]

3.2. Transfer Learning

As explained in Section 2, due to the data-driven nature of the deep learning method, the performance of a model can be improved significantly by simply providing more training data (given enough learning capacity of the model). However, annotating millions of images is an expensive task and is infeasible for safeguards which have additional constraints, such as sensitivity policies. Further, training a high-capacity model with only a few data points will lead to over fitting (i.e. the model memorizing the data set) and sub-optimal performance during testing. To address the issue, transfer learning is used [8]. In the computer vision community, there are several large open-source benchmark data sets containing thousands of annotated images of everyday objects such as dogs, cats, cars, houses, plates, etc. The team starts with the weights of YOLO that was trained on one of these benchmark data sets. These weights represent the initial settings of the algorithm. Given the large data set, the weights are tuned very well for detecting basic features of objects, e.g. edges and surfaces and basic shapes. The

weights related to advanced features, e.g. object classification, are updated in this development by training on the set of images containing annotated images of the safeguards relevant objects.

3.3. Use Cases

In conducting surveys with the IAEA safeguards inspectors, it was conveyed that some of the most time-consuming surveillance review processes that inspectors undertake as part of the in-field safeguards activities are related to the transfer of spent nuclear fuel from a wet storage facility (typically the cooling pond immediately adjacent to a nuclear reactor core) into dry storage or transportation casks consisting of drying the containers at or near wet storage facility, and then moving the containers to a remote storage or processing area. Challenges to this surveillance data review are a result of several factors including, but not limited to: 1) the safeguards-importance of the material being moved; 2) the busy nature of the scene, with people, cranes, containers, and fuel assemblies in motion; 3) the necessity to track objects through multiple camera viewpoints, which is difficult even when an object is not in motion; 4) the long duration of the transfer activities, with an individual cask transfer taking up to one to two weeks; and, 5) significant time pressure on inspectors to conclude surveillance review activities. The effort to detect and classify objects of safeguards relevance in video surveillance is intended to free up inspector time to focus on other activities and increase the likelihood of detecting anomalous events. In this project, two test facilities that simulate the above use cases are being used to produce data for algorithm evaluation.

3.3.1. *Waste Repackaging and Storage*

A scenario of waste repackaging and storage was identified at BNL and is being used to simulate the waste transfer process within a reactor hall. FIG. 2 shows the high-bay building where the radioactive waste materials are packaged. Two main objects are showed in the insertion of the figure. The pig container (a) is used to transfer radioactive waste from laboratories to the packaging facility. Another container, a 55-gallon drum, is used for waste packaging. When the waste is transferred and the 55-gal container is sealed, the container is stored in the storage area in the same building until it is shipped out. There are also other containers with different shapes in this operation area, which can be used to simulate the busy nature of nuclear facilities. Also captured in the photo is the crane to move these heavy objects. One NGSS camera and one commercial off-the-shelf (COTS) camera are used to cover the waste transfer area. Another COTS camera is used to monitor the waste storage area.

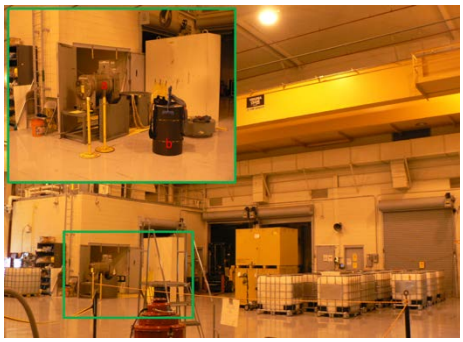


FIG. 2. Waste packaging facility. Shown in the insertion: (a) waste transfer container and (b) 55-gallon container for packaging

3.3.2. *Move Dry Storage or Transportation Casks*

To support more efficient processing of the safeguard surveillance data surrounding cask transfer, SNL has developed an experiment that mimics some aspects of the cask transfer activity. In the experiment, distinct white drums, shown in FIG. 3a, serve as proxies to nuclear spent fuel casks. These are moved into and out of a large, concrete floor vault (3.0 m (W) × 4.6 m (L) × 7.6 m (D)) using a combination of pallet jacks and facility and gantry cranes (FIG. 3b). Early experimental campaigns will feature a dry floor vault, which will give the team experience with the test equipment and test procedures, as shown in FIG. 3c. In later campaigns, the vault will be

filled with water to produce a more realistic environment of a spent fuel pond at a nuclear reactor. Video cameras, including NGSS cameras and a secondary system of commercial off-the-shelf (COTS) cameras, are deployed around the floor vault in a manner simulating the setup of surveillance cameras at a safeguarded operational nuclear power reactor. Specifically, the NGSS cameras are setup to capture this use case from multiple angles (perspectives) and to eliminate any areas (dead zones) where the container could not be seen by any of the cameras.

FIG. 3. (a) Proxy spent fuel cask, (b) overhead crane, and (c) dry simulated spent fuel pool.

4. METHOD

4.1. Dataset

The more annotated data that is available for training, the better the trained model will be. However, the IAEA identified data confidentiality and model training in externally-controlled environments as problematic, which drove the team's data collection methodology. To have high fidelity with the real-world constraints, the team used digital cameras to capture images and videos of objects of interests from different orientations and distance inside and outside of the testing facilities. This is analogous to situations where images of the objects of interest may be available, but not within the actual facility. Depending on the availability, photos with different background were also taken for some objects of interest. These images and videos were labelled and used as the training data set. In total, about 650 images have been taken and labelled in the current development.

Selected videos from the NGSS cameras and COTS cameras were labelled using VitBAT software [13], a software package that specifically supports annotating video and tracking items across image sequences. This data set is used as test data during algorithm evaluation.

4.2. Evaluation Metrics

The most common metric for measuring the performance of object detection models is mean average precision (mAP), which is widely used by benchmarks of object detection in the computer vision community. The average precision (AP) is measured per object class, which is the area under the precision-recall curve. Here precision and recall are defined as the followings.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Where TP, FP and FN are true positive, false positive and false negative value respectively. The data points are ranked by their confidence scores. It can be viewed as a summary of the precision-recall curve. The mean AP is the average over all object classes.

4.3. Results

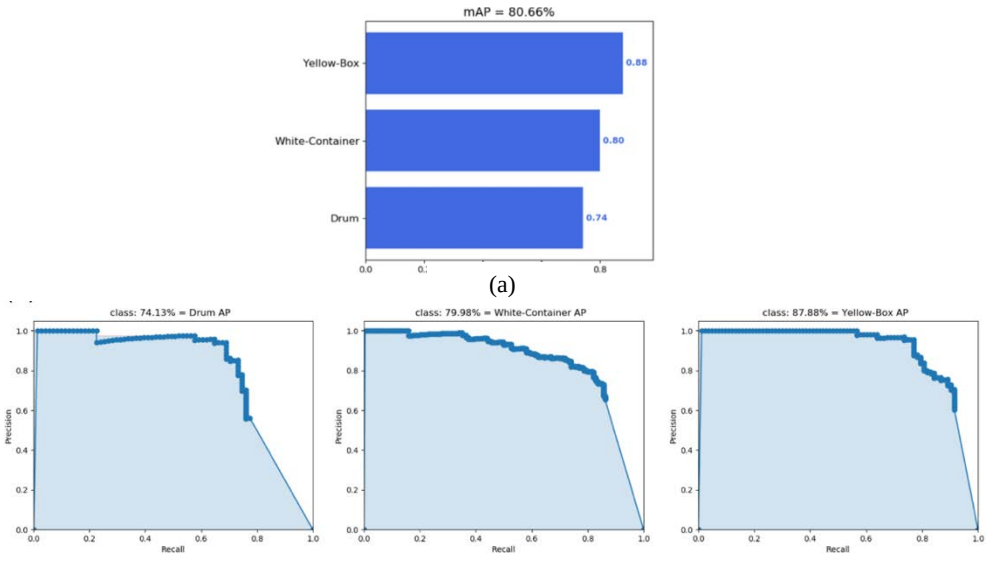
FIG. 4 shows the example images collected at the testing facilities and annotated by the YOLO-SG algorithm. FIG. 4(a) is a photo of waste repackaging facility and has three classes of objects detected, 55-gallon drum, white plastic containers and yellow box containers. FIG. 4(b) and (c) show images where objects are

identified at simulated spent fuel pool. As can be seen, YOLO-SG is able to detect and localize the objects of interest after being trained on fewer than 300 images. Bounding boxes around the objects of interest, such as those produced in our results, should significantly increase the efficiency of review for an inspector.

(a) (b) (c)

FIG. 4. Examples of images annotated by YOLO-SG. (a) Waste repackaging facility. (b-c) Simulated spent fuel pool.

FIG. 5(a) shows the mAP values of the evaluation results at waste repackaging facility. An averaged value of 0.80 is measured. FIG. 5(b) shows the detailed precision-recall curves for each object class. Further investigation shows that the precision-recall curve and the mAP value is closely related to labelling of data set. For data set in which objects' boundaries are labelled precisely, a higher mAP value can be achieved. Given this finding, the data set for this evaluation is being re-visited. A guidance of labelling images for this project will be generated based on re-evaluation results, which will facilitate the transfer of the algorithm from development phase to real deployment.



(a)

(b)

FIG. 5: Test results of YOLO-SG. (a) The Average Precision score per class with the mean-AP of 80.66%. (b) The precision-recall curves per class.

5. CONCLUSION

For the spent fuel transfer model, a proof-of-concept implementation of the YOLOv3 model was implemented using the full dataset for training and testing to evaluate basic operability of the model which the team refers to as YOLO-SG. Initial results are promising and work continues as of this writing. YOLO-SG addresses the problem of localizing objects of interest in images within the constraints of a lack of training data (due to confidentiality issues) and fast execution time. Successful demonstration of this algorithm will help bridge state-of-the-art computer vision techniques with safeguards video review and continue to make the review process more efficient and less error prone.

ACKNOWLEDGEMENTS

The authors thank Joseph Carbonaro and Robert McNair at BNL for encouraging discussions and their guidance in preparing test beds for data collection. The authors are grateful to Glen Todzia, Edward Richards and John Aloi Jr. at BNL for their efforts in arranging data collection. The authors also thank Mary Arnhart and Phillip Kay at SNL for supporting data collecting and annotation, and Greg Baum, Don Hanson and Wayne Garcia at SNL for facilitating in-situ experiments.

This material is based upon work supported by the U.S. Department of Energy's National Nuclear Security Administration (NNSA) Office of Nuclear Safeguards and Security (NA-241). The manuscript has been authored by Brookhaven National Laboratory managed by Brookhaven Science Associates, LLC under Contract No. DE-SC0012704 with the U.S. Department of Energy, and by Sandia National Laboratories which is managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

REFERENCES

[1] Safeguards Techniques and Equipment: 2011 Edition, International Nuclear Verification Series 1 (Rev. 2), IAEA, Vienna, 2011.

[2] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R.m Jays, J., Rerona, P., Romanan, D., Zitnick, C.L. Dollar, P., "Microsoft COCO: Common Objects in Context", European Conference on Computer Vision, (Proc. European, Zurich, Switzerland, 2014), Springer Nature, Switzerland AG., 2014, pp. 740–755.

[3] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A., "The Pascal Visual Object Classes (VOC) Challenge", Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, Jun. 2010.

[4] Krizhevsky, A., Sutskever, I., Hinton, G.E., "ImageNet classification with deep convolutional neural networks", Neural Information Processing Systems (NIPS) (Advances in Neural Information Processing Systems 25, Nevada, USA, 2012), Neural Information Processing Systems Foundation, Inc., La Jolla, USA, 2012.

[5] Girshick, R., Donahue, J., Darrell, T., Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation", IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Proc. Int. Conf., Columbus, Ohio, 2014), IEEE, Washington, DC, USA (2014).

[6] Girshick, R., "Fast R-CNN", IEEE Conference on Computer Vision (ICCV) (Proc. Int. Conf., Santiago, Chile, 2015), IEEE, Washington, DC, USA (2015).

[7] Ren, S., He, K., Girshick, R., Sun, J., "Faster R-CNN: Towards real-time object detection with region proposal networks", Neural Information Processing Systems (NIPS) (Advances in Neural Information Processing Systems 28, Montreal, Canada, 2015), Neural Information Processing Systems Foundation, Inc., La Jolla, USA, (2015).

[8] Yosinki, J., Cline, J., Bengio, Y., Lipson, H., "How transferable are features in deep neural networks?", Neural Information Processing Systems (NIPS) (Advances in Neural Information Processing Systems 27, Montreal, Canada, 2014), Neural Information Processing Systems Foundation, Inc., La Jolla, USA, (2014).

[9] Redmon, J., Farhadi, A., "YOLOv3: An Incremental Improvement", Technical Report, http://arxiv.org/abs/1804.02767, (2018).

[10] He, K., Zhang, X., Ren, S., Sun, J., "Deep Residual Learning for Image Recognition," ArXiv E-Prints, Dec. 2015.

[11] Shelhamer, E., Long, J., Darrell, T., "Fully Convolutional Networks for Semantic Segmentation", IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 640–651, Apr. 2017.

[12] Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., "Feature Pyramid Networks for Object Detection," IEEE Conference on Compuer Vision and Pattern Recognition (CVPR) (Proc., Honolulu, USA, 2017), IEEE, Washington DC, USA, 2017.

[13] ViTBAT software, https://vitbat.weebly.com/.

APPENDIX
YOLO CNN MODEL

The overall network architecture of YOLOv3 consists of three components: feature extractor network, feature pyramid network, and detection and classification heads, as seen in FIG. A and are described in the following. The entire YOLOv3 network is trained in an end-to-end fashion.

The *feature extractor network* is to transform the raw image input to a feature tensor, using a series of convolution and non-linear activation functions. The term "features" refers to the useful information extracted from the image that may be used to characterize objects. Before the deep learning revolution, the features are extracted by applying carefully hand-crafted operators or functions called kernels. With the reinvention of the neural networks, especially convolutional neural networks, the kernels became learnable parameters. The state-of-the-art feature extractors are the residual networks [10] and its variations, which help tackle the "vanishing gradient problem" when the neural network becomes very deep. The feature extractor network of YOLOv3 contains 23 residual blocks of various filter sizes, intermediated by five down-sampling convolution layers. This architecture is effective in extracting features as shown by its good performance in image classification tasks on ImageNet, and efficient comparing to much deeper residual networks (ResNet-101).

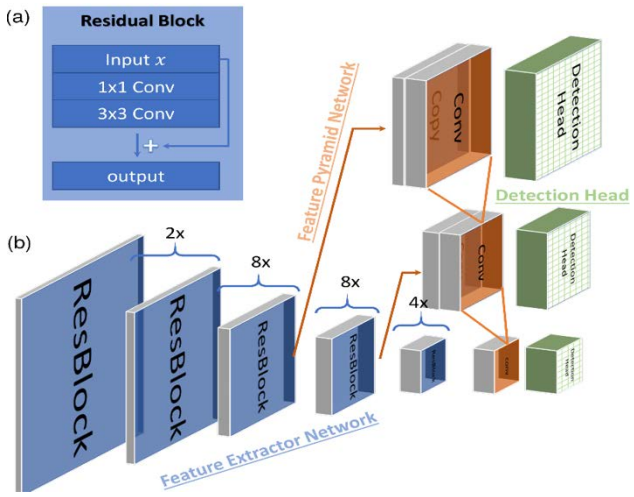


FIG A. A residual block consists of two convolution layers and a shortcut connection from the input to the output. (b) The YOLOv3 network consists of three components 1) the feature extractor network (blue), 2) the feature pyramid network (orange), and 3) the detection heads (green).

The *feature pyramid network* brings the feature tensor to different scales using convolution layers for down-sampling and plain up-sampling. Compared with the YOLO v2, the most prominent modification of YOLOv3 is its ability to detect objects at different scales. The architecture is very similar to the feature pyramid network [12]. After the last layer of the feature extractor network, the features are abstract - the highest filter sizes yet the smallest spatial dimensions. Hence, the output is good for classification tasks, but not for detection because subtle changes of locations and postures have been abstracted out, and small objects shrink to one single pixel. To tackle this problem, the feature pyramid network up-samples the most abstract features and concatenate with the features multiple layers earlier where the more details are preserved. As shown in FIG. A, YOLOv3 has three stages of the pyramid network for small, medium and large objects. The flexible design of the model allows adding more pyramid network stages for more various scales to fit specific applications.

The last component of the YOLOv3 model is the *detection head* responsible for making decisions based on the extracted features. Each detection head has three convolution layers connected to each stage of the pyramid network, as seen in FIG. A. Anchor boxes are pre-defined stereotypes of bounding boxes with different aspect ratios. For each "pixel", or coordinate in spatial dimension of the features, B different anchor boxes are bounded. One of the tasks is to regress the deviation (adjustment) of the box centres and sizes (4 scalars in total). The model also dedicates one scalar as the "objectiveness" score. In addition, C scalars represent classification classes. Therefore, in total, there are B×(4+1+C) output scalars per "pixel" in feature space. The spatial dimensions of the features depend on the design of the network and the stage of the pyramid networks. For example, if a YOLOv3 configuration decides to take the input of size 416-by-416, after going through the feature extractor, the spatial dimension is down-sampled 32-fold, resulting a 13-by-13-by-filter_size feature tensors. For each of the 13-by-13 "pixel", the detection head outputs a vector of size B×(4+1+C).