Proceedings - 5th IMA Conference on Mathematics in Defence
23 November 2017, The Royal Military Academy Sandhurst, Camberley, UK

Deep object classification in low resolution LWIR imagery via transfer learning

By Abbott R.[1], Del Rincon J.M.[1], Connor B.[1], and Robertson N.[2]

[1]ECIT, Queen's University Belfast, BT3 9DT
[2]Thales Optronics Glasgow, G51 4BZ
{*rabbott02, j.martinez-del-rincon, n.robertson*}*@qub.ac.uk*
{*Barry.Connor*}*@uk.thalesgroup.com*

Abstract

In this paper we present an object detection system based on YOLOv2 and transfer learning which is compared to Keras faster Regions with Convolutional Neural Networks (Keras faster R-CNN). We evaluate our systems using two infrared datasets: a small dataset of low resolution thermal images and a large dataset of high resolution thermal images both containing the object classes; people and land vehicles. Both detectors are trained on the large dataset of high resolution thermal images and tested using the low resolution images. Fine tuning on the small dataset is implemented to increase the accuracy of the detectors. This research will be of great interest to the defence community as they could save a lot of time and money collecting and annotating data.

1. Introduction

The security and defence industry is becoming increasingly reliant on intelligent signal processing methods to achieve 24 hour surveillance capabilities. The military are one of the largest applications for computer vision. This includes detection of enemy soldiers, vehicles and guiding missiles. Modern military concepts such as "battlefield awareness", use various sensors, including image sensors, to provide information about a combat scene which can be used to support strategic decisions. By developing computer vision with 24 hour surveillance we can reduce the burden on human operators thereby allowing them to make safer decisions.

The first obstacle in developing 24 hour surveillance is localising and classifying objects in images i.e. object detection. Infrared images can be used in these systems as they provide data in the day, night and in low visibility. Normally detection systems are are trained and evaluated in controlled conditions in high resolution images so real life object detection is difficult due to occlusion, clutter, low-resolution images and long-range targets. The most popular solution to the problem of object detection is using deep convolutional neural networks due to unprecedented performance. However, there is a lack of research that involves object detection in infrared images and under operational conditions which are required for 24 hour surveillance systems.

Twenty-four hour surveillance systems require accurate object detectors. Object detection systems require a large annotated training dataset to achieve high accuracies. This is because convolutional neural networks (CNNs) do not typically generalize well to data not present in the training set. We therefore need a large dataset in order to get as many different angles/views of objects in as many different situations as possible.

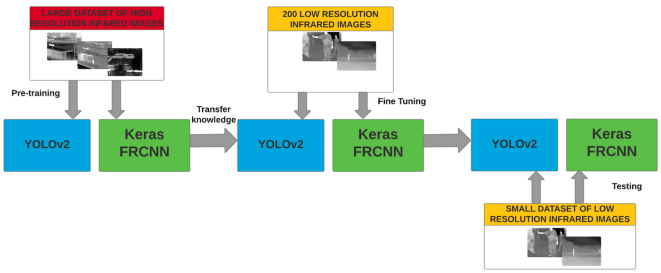

FIGURE 1. Network diagram of Transfer learning and fine tuning process. A large dataset of high resolution images are trained on Keras faster R-CNN and YOLOv2. All the weights produced are reused and a small dataset of low resolution images are trained on top in a process called fine tuning. The system is then ready to be tested on the low resolution images.

These datasets are extremely expensive to collect in both money and time, especially for those in defence who heavily rely on these systems. This makes it very difficult to collect enough realistic data to train a neural network from scratch to obtain accurate results. Transfer learning is a solution to this problem as it requires only a small dataset of images. Transfer learning involves training a CNN on a very large dataset, then extracting the features produced in the convolutional layers, and finally training a new model on these extracted features using a small dataset in a process called fine tuning. This technique shows impressive results on a number of vision tasks as shown in Donahue *et al.* and is adapted in this paper for a new domain (Automatic target recognition) and sensor (Long Wavelength InfraRed). The theory behind Transfer learning is that low level features which have been learned through training are the same for all objects. Therefore these features can be used to help classify new objects and existing objects in different resolution images. When applying transfer learning to different resolution images, the weights need to be adapted to compensate for the different size of images.

It is in this context that we propose the use of transfer learning to adapt two state-of-art object detectors (Keras faster R-CNN and YOLOv2) to low resolution images with the future aim of developing a 24 hour surveillance system. Keras faster R-CNN and YOLOv2 are robust to training on different sized images and so learning across different resolution images allows successful classification without further costly and time consuming work to change the weights.

We take as our starting point the weights from an object detector trained on a pre-existing large dataset of high resolution infrared images and re-train the model with a very small dataset of low resolution infrared images. A diagram depicting this set-up is shown in Figure 1. Our work would be of significant interest to the military as accurate and low cost object detection systems are crucial.

2. Related Work

In the past hand crafted models such as Bag of Words were very popular for visual recognition. However, deep learning has clearly surpassed these conventional methods. Deep learning is used in a diverse number of areas such as in driver-less cars, medicine (Greenspan *et al.*) and astrophysics (Baldi *et al.*). It can be thought of as a subfield of machine learning which is able to learn complex and hierarchical representations from raw data. In the modern era of deep learning, complex architectures have won many computer

(a) (b) (c) (d)

FIGURE 2. Images a) and b) were taken with the Foxhound camera (resolution of 320x240) and c) and d) were taken with Catherine MP (resolution 1280x1024)

recognition challenges and have surpassed previously unattainable performance with regards to localization, classification, detection, tracking, pose estimation, segmentation or image captioning with respect to other older methods. The success of deep learning has been made possible by the increase in computational resources such as open source libraries, GPUs and the increase of labeled data. In 2012 Krivhesky *et al.* created a large and deep convolutional neural network which won the ImageNet large-scale visual Recognition challenge. This CNN had a top five error rate of 15.4% which was significantly lower than previous years and 10.8% lower than the next best entry that year. In 2014 Simonyan *et al.* created a convolutional neural network with 19 layers after investigating the effect of depth on the accuracy in image recognition tasks. This model achieved 7.3% error rate and was influential as it highlighted that networks need to be deep in order for hierarchical representations of visual data to be successful.

Object detection is a more difficult problem than object classification or recognition and requires greater computational power. Regions with convolutional neural networks (R-CNN) by Girshick *et al.*(2013) was one of the first systems to combine object localisation and classification tasks. It uses selective search to bound an object in an image (region proposal) and then a CNN is used to classify. This detector works well but is slow because for every image there are multiple region proposals and each has to go through the CNN to be classified. Fast R-CNN Girshick *et al.*(2015) solves this problem by running the CNN computation once and then sharing the information across all the region proposals. Although this increased the speed, the region proposal method was still slow. Faster R-CNN was created to solve this issue. Region proposals depend on features which are calculated through the CNN. The insight with faster R-CNN was to use the CNN results for region proposals.

YOLO is a new object detector by Redmon *et al.* (2015). It combines object localization and classification into one neural network to produce bounding boxes and class probabilities for objects in images. The architecture works in real-time at 45 frames per second. FAST YOLO, which is a smaller version of the network can achieve 155 frames per second. YOLO makes more localisation errors compared to other state-of-the-art systems Redmon *et al.* (2015) but is less likely to predict a false positive in the background of the image. YOLO outperforms other detection methods such as deformable parts model (DPM) and R-CNN when generalizing from natural images to other domains.

An extension of YOLO, called YOLOv2, is proposed in Redmon *et al.* (2016) and provides both higher accuracy and faster performance. This is achieved by increasing the resolution, applying batch normalisation and multi-scale training, optimising the dimension clusters and using anchor boxes during the convolution. Interestingly, training the classifier on high resolution images increases the mAP by nearly 4%. YOLOv2 is trained to detect objects across a variety of input resolutions defined every 10 batches by choosing a different image resolution. The resulting network is then resized to that

dimension and training is continued. Therefore, YOLO is robust to running on images of different sizes.

There is a lot of work using deep neural networks (DNNs) for face recognition using thermal images for example, in Peng *et al.*. However, there is a lack of research using thermal images for object detection. Rodger *et al.* (2016) develops a CNN trained on short-mid range high resolution infrared images (taken on Thales Catherine MP thermal imager) containing the following objects classes; person, land-vehicle, helicopter, aeroplane, unmanned aerial vehicle and false alarm. This network was successful at classifying other short-mid range objects in unseen images but struggled to generalize to long range targets which indicated problems identifying targets in low resolution conditions. Their solution was to introduce a new long-range class and train the network again. They achieve accuracies of 95.7%. We aim to build upon this area of research, address the lack of work using infrared images in detection and expand it to low resolution imagery.

3. Experiments

3.1. *Datasets*

Two different datasets are used in this paper to validate our proposal:

(a) A large dataset of high resolution images taken on Thales thermal imager Catherine MP. This data consists of around 3200 images of people and land vehicles taken at long range from a look down perspective. Partial and total occlusions by trees and building are present in the data. The data also contains images of vehicles traveling along a road at long range from a face on perspective.

(b) A small dataset of low resolution images taken by the universally deployed Thales Foxhound thermal imager. These images show close and long range people in indoor and outdoor scenes as well as some close range images of cars and vans. There are approximately 700 images in total.

The Catherine MP dataset will be used to train the system while the Foxhound data will be used for fine tuning and to validate the performance.

3.2. *Method*

In this paper, we evaluate the accuracies of object detection using a small dataset of low resolution images and transfer learning. We train two detectors, Keras faster R-CNN and YOLOv2 on a large dataset of high resolution images and then use fine tuning to increase the accuracy for testing on the low resolution images. The weights from all the layers in the networks trained on the high resolution images are used as the objects in the low resolution images are the same. YOLOv2 starts training with a weights file pre-trained on the ImageNet dataset. Keras faster R-CNN starts training from scratch.

3.3. *Architecture of Keras FRCNN and YOLOv2*

YOLOv2 is a light-weight architecture as it uses a convolutional network for both localisation and classification. The classifier part of the network is called darknet-19 which consists of 19 convolutional layers and 5 max pooling layers. Darknet-19 is trained on the ImageNet for 160 epochs at resolution 224x224. Standard data augmentation tricks are used such as saturation/hue etc. The network is then fine-tuned for 10 epochs on resolution 448x448. When training for detection the network is modified by removing the last convolutional layer and replacing it with three 3x3 convolutional layers (with 1024 filters) and one 1x1 convolutional layer with the number of outputs for detection. Faster R-CNN, uses the VGG-16 model to classify the objects and 3 additional layers for

region proposals. YOLOv2 has the most efficient architecture as it frames detection as a regression problem and doesn't require such a complex structure that Faster R-CNN has. The two detection systems use different loss functions to optimize the weights produced. YOLOv2's cost function simultaneously solves for object localisation and classification tasks. The following cost function is used:

$$loss = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (x_i - \hat{x_i})^2 + (y_i - \hat{y_i})^2$$
$$+\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (\sqrt{w_i} - \sqrt{\hat{w_i}})^2 + (\sqrt{h_i} - \sqrt{\hat{h_i}})^2 + \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (C_i - \hat{C_i})^2$$
$$+\lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{noobj} (C_i - \hat{C_i})^2 + \sum_{i=0}^{S^2} 1_i^{obj} \sum_{c \in classes} (p_i(c) - \hat{p_i})^2$$

The sum squared error is used in this cost function as it is easy to optimize. The first two terms calculate the error of the bounding boxes. 1_{ij}^{obj} equals 1 if the jth bounding box predictor in cell i is responsible for that detection. The x and y are the coordinates of the bounding box, and w and h are the width and height respectively. In order to reflect that small deviations in large boxes matter less than small boxes, the square root of the width and heights are taken. The third and fourth terms penalize the difference in confidence of having an object in the grid, and the final term is penalizing the difference in class probability. The parameters λ_{coord} and λ_{noobj} are equal to 5 and 0.5 respectively. The value of λ_{coord} is to ensure a larger contribution of the bounding box and classification error to the overall cost function, and λ_{noobj} is set to penalize less for the confidence of identifying an object when there is not one present. The value 1_i^{obj} equals 1 if an object appears in cell. S^2 represents the number of cells that YOLOv2 splits the image into for the detection process.

In faster R-CNN a binary class label is assigned to each anchor describing if an object is present or not. A positive label is assigned to the anchor which has the biggest intersection over union (IoU) with the ground truth or an anchor which has a IoU of > 0.7. A negative label is given to an anchor if its IoU is less that 0.3 for all ground truth. The other anchors do not contribute. Using these rules, Faster RCNN minimizes the following loss function:

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, \hat{p_i}) + \lambda \frac{1}{N_{reg}} \sum_i \hat{p_i} L_{reg}(t_i, \hat{t_i}) \tag{3.4}$$

In this loss function, i represents the index of an anchor in a mini-batch and p_i is the predicted probability of anchor i being an object. The value $\hat{p_i}$ is the ground truth label and $\hat{t_i}$ represents the vector containing the four coordinates of the bounding boxes. L_{cls} is the log loss over the two classes; object verses not object. L_{reg} is the regression loss and $L_{reg} = R(t_i - \hat{t_i})$ where R is the robust loss function smooth $L1$:

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if} \quad |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{3.5}$$

4. Results

Both object detectors were initially trained on the Catherine MP dataset and tested on the Foxhound dataset. Then a small subset of around 200 of the Foxhound images were used for fine tuning. The results are discussed in the following subsections.

4.1. *YOLOv2 Results*

YOLOv2 which was pre-trained on ImageNet was trained for one day on Catherine MP data, going through 90000 iterations altogether. A weights file is saved every 10000 iterations. After testing the precision at each of the iterations, the highest precision produced on the Catherine MP test set was 92% and 52% was achieved on the Foxhound test set. This result of 52% is a good starting precision considering how different the datasets are. When analysing the errors and miss-detections, the following challenges were discovered: a) YOLOv2 does not correctly classify close up images of people. It confuses them with land vehicles and in some cases does not register there is an object present as seen in Figure 3. b) Objects not present in the training data such as a wheel barrow, hoover and a person on a chair all are mistaken for land vehicles, also shown in Figure 3.

To overcome these initial mis-classifications, fine tuning was implemented and YOLOv2 was trained with around 200 Foxhound images. The test set precision of Foxhound images increased to 84%. Figure 4 shows these results. As you can see from Figure 3 the results have greatly improved, however, the hoover is still being classified as a land vehicle, although this is understandable as it is a new class not present in the training data. The main improvement this fine tuning process made are that close range images of people are now being classified correctly and precision has significantly increased.

4.2. *Results from Keras faster RCNN*

After one week of training Keras faster R-CNN initially gives an mAP of 83% on the Catherine MP test set and 26% on the Foxhound test set. This initial low performance in both the high and low resolution images in comparison to YOLOv2 could be explained by the lack of pre-training in other images as it is trained from scratch in Catherine MP. After fine-tuning was implemented, the Foxhound test set mAP is raised to 87%. Figure 4 shows a chart of the results.

5. Discussion

Our results show that Keras faster R-CNN is the marginally better detector with a mean average precision of 87% compared to 84% with YOLOv2. While the accuracy of these detectors could be improved, these results are promising considering how small the datasets were in comparison to other research (which use tens of thousands of images for training), and the fact the datasets were taken in different locations and at different ranges. These results are a starting point to develop this research and make the detection system better. They also show that transfer learning followed by fine tuning produces good results from real world data in low resolution.

YOLOv2's strength is that it is significantly quicker to train. YOLOv2 took one day to train in comparison to Keras faster RCNN which took five days to train to a suitable test set accuracy. Therefore although Keras faster R-CNN is 3% more accurate, YOLOv2 is the better detector as it is much faster to train.

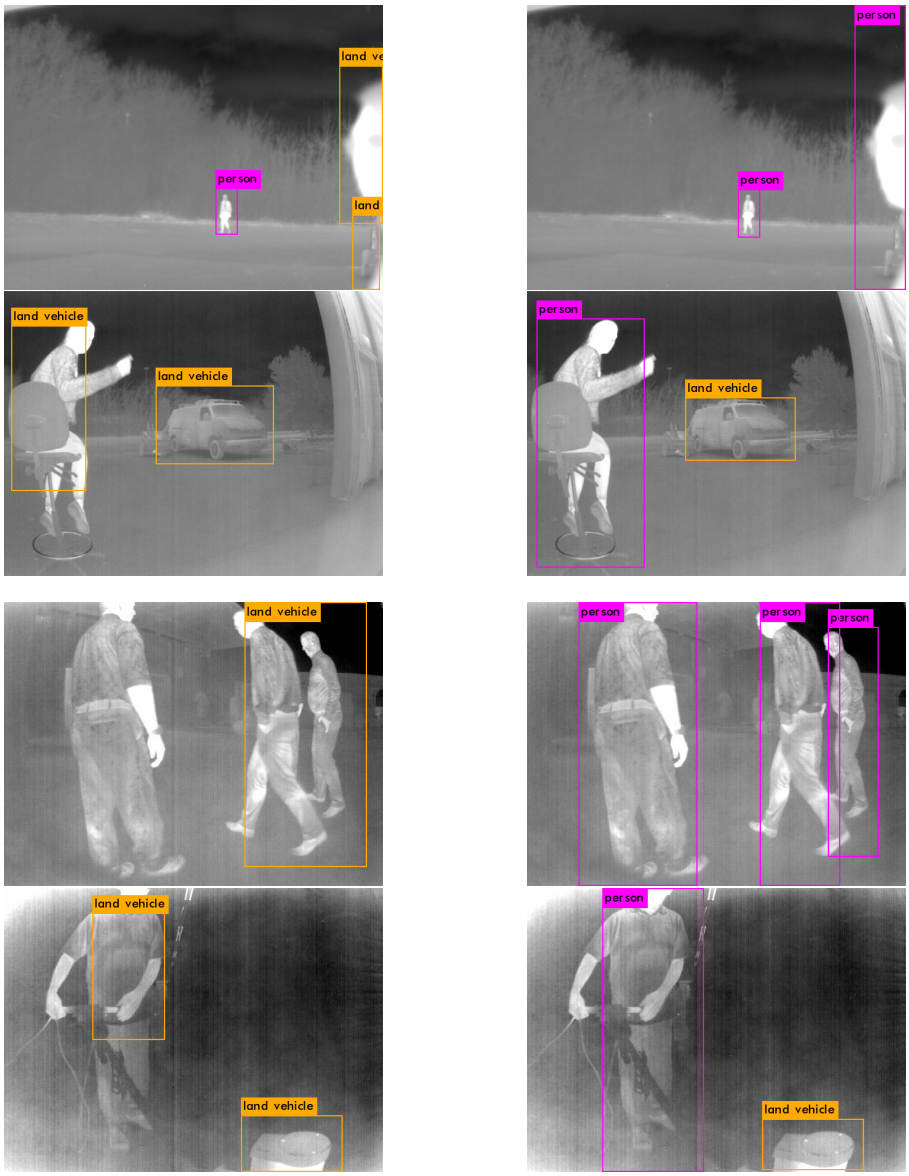

FIGURE 3. This Figure shows Foxhound images that have been tested on YOLOv2 trained on Catherine MP images before (left hand side) and after (right hand side) fine tuning (on a small set of Foxhound images). Before fine tuning most of the people and the hoover are all classified incorrectly as land vehicles. After fine tuning the person on a chair and all the people are correctly classified as people and not land vehicle. The hoover, however, is still being classified as a land-vehicle.

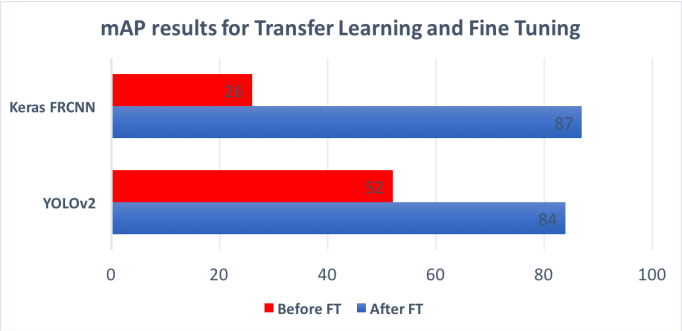

FIGURE 4. Mean average precision results for Keras faster R-CNN and YOLOv2 tested on Foxhound images after training on Catherine MP images.

6. Conclusion

The originality of our work lies in focusing on tasks with a small infrared dataset. We show one large dataset can be used for pre-training for other small datasets of different resolution producing high accuracies. This will save a lot of time and money collecting and annotating large datasets. This work is important as one of the reasons Computer Vision tasks cannot accurately classify objects is due to a lack of training data. This paper will be interesting for the defence community as it shows how transfer learning in combination with modern off-the-shelf object detectors based on Deep neural networks can be applied for surveillance and military applications in thermal imagery. Although the accuracy may not be as high as when training on a large dataset, this study is interesting because shows the trade off there is between speed (using small datasets) and accuracy (using large datasets).

6.1. *Future Work*

We propose to increase the accuracy of YOLOv2 by designing a novel cost function to reduce the number of false positives produced and to improve the transfer learning process.

7. Acknowledgements

We acknowledge the support of the Engineering and Physical Sciences Research Council (EP-SRC) Grant number EP/KO14277/1, and the University Research Collaboration (UDRC) in Signal Processing.

REFERENCES

Redmon, J. and Divvala, S. and Girshick, R. and Farhadi, A. 2015 You Only Look Once: Unified, Real-Time Object Detection. *ArXiv e-prints*.

Redmon, J. and Farhadi, A. 2016 YOLO9000: Better, Faster, Stronger *CoRR*.

Rodger, I. & Connor, B. & Robertson, N. 2016 Classifying objects in LWIR imagery via CNNs *Proc.SPIE* **9987**, 99870H-99870H-14.

Donahue, J. and Jia, Y. and Vinyals, O. and Hoffman, J. and Zhang, N. and Tzeng, E. and Darrell, T. 2013 DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition *ArXiv e-prints*.

Fergus, R. and Perona, P. and Zisserman, A. 2003 Object class recognition by unsupervised scale-invariant learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* **2**, 264–271.

Krizhevsky, A. and Sutskever, I. and Hinton, G. E. 2012 ImageNet Classification with Deep Convolutional Neural Networks *Advances in Neural Information Processing Systems 25* 1097–1105.

Simonyan, K. and Zisserman, A. 2014 Very Deep Convolutional Networks for Large-Scale Image Recognition *CoRR* **abs/1409.1556**

Girshick, R.B and Donahue, J. and Darrell, T. and Malik, J. 2013 Rich feature hierarchies for accurate object detection and semantic segmentation *CoRR* **abs/1409.1556**.

Girshick, R. 2014 Fast R-CNN *CoRR* **abs/1504.08083**.

Ren, S. and He, H. and Girshick, R. and Sun, J. 2015 Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks* 91-99.

Dejong, G. and Mooney, R. 1986 Explanation-Based Learning: An Alternative View *Machine Learning* **1** 145–176.

Mhlenbein, H. 2009 Computational Intelligence: The Legacy of Alan Turing and John von Neumann" *Computational Intelligence: Collaboration, Fusion and Emergence* 23–24.

Peng, C. and Wang, C. and Chen, T. and Lui, G. 2016 NIRFaceNet: A Convolutional Neural Network for Near-Infrared Face Identification *Information*.

Greenspan, H.; Summers, R.M.; Ginneken, B. van 2016 Guest Editorial: Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique *IEEE Transactions on Medical Imaging*.

Baldi, P. and Sadowski, P. and Whiteson, D. 2014 Searching for exotic particles in high-energy physics with deep learning *Nature Communications*.

Kong, T. and Yao, A. and Chen, Y. and Sun, F. 2016 HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection *CoRR*