Object Detection for Body Worn Video

Kaiyan Peng, Carson Hu, Jaydeep Singh, Mingxi Sun, Duo Zhang and Kan Yao
Mentor: Andrea Bertozzi, Supervisor: Bao Wang
Department of Applied Mathematics, University of California Los Angeles

Introduction

Deep learning object detection

We apply Faster-RCNN, a deep convolutional neural network (CNN) based object detection algorithm, to the field of body-worn video: we want to locate people in the image and distinguish them between police and non-police individuals. Along with creating a functional object detector, we want to improve the results of Faster-RCNN. The basic detection pipeline is shown in figure 1.

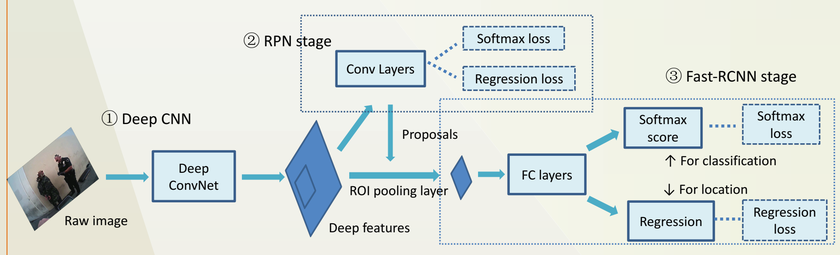


Fig1: We first feed in the image, which goes through the Deep Convolutional Network (ConvNet). The output features then go through the Region Proposal Network (RPN) stage, which will return proposed Regions of Interest (ROI). After combining the features and the region proposals, the Fast-RCNN stage will determine the label of different regions (police, others, background) and output a more precise bounding box location through regression.

Dataset

We sample the images from around 700 body worn videos offered by the Los Angeles Police Department at a rate of 1 frame per second. In total we sample around 1.0M images from the videos. We then

- manually identify frames with objects of either class present
- annotate those frames as is shown in figure 2.

Fig2: We annotate each person with a class label, and a bounding box circumscribing as much of the person as appears in frame

After a verification process, 30K frames comprise our training and testing set.

Experiment Setup and Primary Results

Experiment setup:

- An alternate optimization way to train Faster-RCNN network
  - step 1 : RPN stage, 80k iterations
  - step 2 : Fast-RCNN stage, 40k iterations
  - step 3 : RPN stage, 80k iterations
  - step 4 : Fast-RCNN stage, 40k iterations
- For the Deep ConvNet layers, we have tried ZF, VGG16 & ResNet-50 models.

Accuracy measure:

We use the notions of Average Precision (AP) and Mean Average Precision (mAP). If the overlap over union area of the predicted bounding box and the ground truth bounding box is greater than 0.5, and they share the same label, we count it as a true positive. In case of duplicate detection, only the predicted bounding box with the highest confidence will be calculated as a true positive. The mAP is the mean of the AP of these two classes.

Primary results:

- ZF:
  - Police: 0.891
  - Others: 0.873
  - mAP: 0.882
- VGG16:
  - Police: 0.901
  - Others: 0.895
  - mAP: 0.898
- ResNet-50:
  - Police: 0.9
  - Others: 0.895
  - mAP: 0.897

Fig3: Successful examples. The left is the detection results for others, and the right is the detection results for police.

Fig4: Problems. The left result mislabels a shadow as a person. The right misses many people.

Reference

[1] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. CoRR , abs/1506.01497, 2015.
[2] Zuoqiang Shi, Stanley Osher, and Wei Zhu. Low dimensional manifold model for image proccessing. Cam report 16-04, UCLA, 2016
[3] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. CoRR abs/1603.02754 (2016).

Methods and Results

Method 1: Image enhancement through LDMM

Suspecting that the tile effect of low-resolution images impedes detection, we adopt Low-dimensional Manifold Model (LDMM) for denoising. Based on the patch manifold model of images and the fact that the patch manifolds of many natural images have low dimensional structures, LDMM uses the dimension of the patch manifold as the regularization to solve the ill-posed problem of recovering the original image, which gives the optimization problem:

$$\min_{f\in\mathbb{R}^{m\times n},\mathcal{M}\subset\mathbb{R}^d} \dim(\mathcal{M}),\ \text{subject to } y=\Phi f+\varepsilon, \mathcal{P}(f)\subset\mathcal{M}$$

Where f: the original image we try to recover, of size $m\times n$
Φ: damage like missing pixels
ε: noise
y: the image we observe
$\mathcal{P}(f)$: the patch set of image f
$\mathcal{M}$: the low dimensional smooth patch manifold embedded in $\mathbb{R}^d$

Enhanced images' results are shown in figure 5 & table 1.

Fig5: LDMM results, left: the original image, right: the repaired image. The tile effect is greatly reduced

	ZF		VGG16		ResNet50	
	Original	Repaired	Original	Repaired	Original	Repaired
AP (police)	0.891	0.892 ↑	0.901	0.899 ↓	0.9	0.897 ↓
AP (others)	0.873	0.872 ↓	0.895	0.895 --	0.895	0.895 --
mAP	0.882	0.882 --	0.898	0.897 ↓	0.897	0.896 ↓

Table1: Detection results on the original images and enhanced images.

Discussion: While LDMM does a great job for denoising from humans' perspective, it somewhat decreases the detection accuracy. The noise actually increases the detector's robustness.

Method 2: Loss Function improvement

To decrease the intra-class variance and increase the inter-class distance of feature distribution, we introduce new penalties(*) with intuition from the K-means algorithm. Our new loss:

$$\mathcal{L}=\mathcal{L}_s+\frac{\lambda}{2}\mathcal{L}_{cen}+\frac{\mu}{2}\mathcal{L}_{con}$$

Where

$$\mathcal{L}_s=-\sum_{i=1}^{m} log\frac{e^{W_{y_i}^T x_i+b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i+b_j}} \quad \text{(Softmax Loss)}$$

$$\mathcal{L}_{cen}=\frac{1}{2}\sum_{i=1}^{m}\| x_i-c_{y_i}\|_2^2 \quad \text{(Center Loss*)}$$

$$\mathcal{L}_{con}=\frac{1}{2}[\mathrm{M}-\sum_{\substack{i\le m,j\le n\\ y_i\neq j}}\| x_i-c_j\|_2^2]_+ \quad \text{(Contrastive Loss*)}$$

Here, λ,μ: loss weight
m: total number of features
n: total number of classes
x_i: the i th feature
y_i: the class of feature x_i
c_j: the center of the class j features
M: a constant
W_j , b_j: the weight term and the bias term for j th class in softmax loss

The resulting feature distribution with Principal Component Analysis (PCA) applied for different loss weights is shown in figure 6.

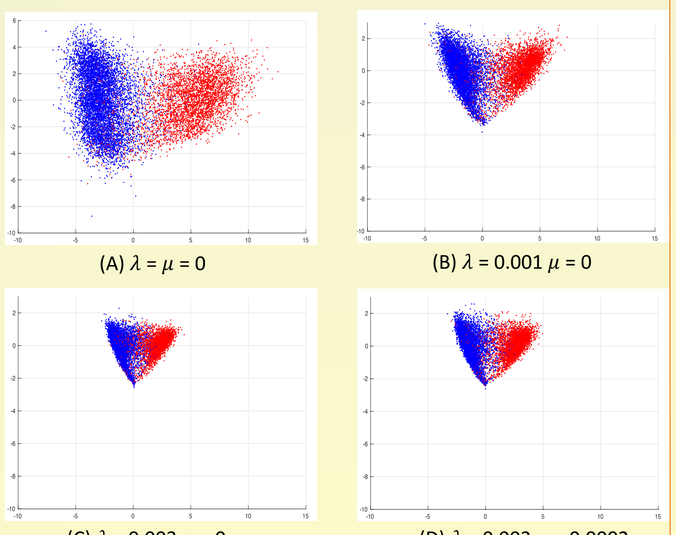
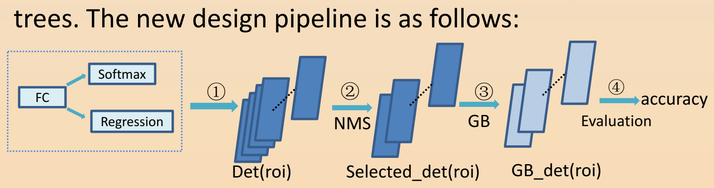


Fig6: Feature distribution with new loss functions. After extracting the 4096 dimension features right before the softmax layer, we apply PCA for visualization. The blue points denote others, while the red represent police.

Discussion: From A, B and C, it is clear that the center loss decreases intra-class variance as larger center-loss weight gives more compact features. One side effect of center loss is it also decreases inter-class distance, which makes a contrastive loss that can pull centers apart necessary. From C and D, contrastive loss contributes a little to separating two classes, but the appropriate parameters are still under exploration.

Method 3: Classifier reinforcement through Gradient Boosting (GB)

With the concern that the softmax classifier is not strong enough to distinguish between features of the two classes, we add in another classifier – Gradient Boosting (GB). GB is a machine learning technique for prediction in the form of an ensemble of classification and regression trees. The new design pipeline is as follows:



① Collect information for each region of interest, denoted as Det(roi), includes the features(a 4096-D vector), softmax scores, predicted bounding box coordinates, the ground truth label
② Decrease duplicate detection by Non-Maximum-Suppression(NMS) and get Selected_det(roi)
③ Feed the GB with selected information, substitute the softmax score with the GB score and get GB_det(roi)
④ Feed the evaluator with the GB_det(roi), get accuracy

	ZF		VGG16		ResNet50	
	Original	GB	Original	GB	Original	GB
AP (police)	0.891	0.895 ↑	0.901	0.902 ↑	0.9	0.904 ↑
AP (others)	0.873	0.886 ↑	0.895	0.899 ↑	0.895	0.899 ↑
mAP	0.882	0.891 ↑	0.898	0.901 ↑	0.897	0.902 ↑

Table 2: Accuracy using new pipeline with GB

Discussion: New pipeline gives positive results for every test.

Conclusion and Future Work

We have applied Faster-RCNN to the problem of police detection in body-worn video. We have tried three directions to improve detection accuracy: image enhancement, feature learning improvement and classifier reinforcement. We have demonstrated that noise in the images makes more robust detectors and Gradient Boosting surpasses the softmax classifier in distinguishing deep features. As for loss functions, more experiments are necessary in this promising aspect.

Looking into the future, there are still multiple fields under exploration:

- Further improve feature learning by designing more stable loss functions
- Apply object detection to other tasks like object tracking and people reidentification.

Acknowledgements

Special thanks to Bao Wang, Andrea Bertozzi and Jeffrey Brantingham for supervision and support.
Big thanks to the UCLA Computational Research Training Program in Computational and Applied Mathematics with funding from NIJ Grant Number: 2014-R2-CX-0101, to the Los Angeles Police Department for supplying the video data for the project, and to the CSST program for fund and support