Final Project Report

Juan Carlos Cruz - ira406

ME 6543 Machine Learning and Data Analytics

Abstract

This paper implements and evaluates a traditional computer vision approach to object detection using Histogram of Oriented Gradients (HOG) features with Support Vector Machine (SVM) classification on a modern industrial safety dataset. The SH17 dataset, containing 8,099 annotated images with 75,994 instances across 17 classes, is used to train and evaluate a person detector. A linear SVM with stochastic gradient descent optimization is implemented, processing 64,990 total features from 10,920 positive and 54,070 negative samples. The model's performance is evaluated using COCO metrics and compared against modern YOLO architectures trained on the same dataset. While the HOG-SVM approach achieved limited performance (mAP50: 1.1%, mAP50-95: 1.0%) compared to deep learning methods (best mAP50: 70.9%), analysis of the detection boxes reveals that predictions tend to fall within ground truth areas, suggesting potential for improvement through enhanced post-processing and parameter tuning.

Introduction

Object detection is a core task in computer vision research and development. Modern deep learning approaches like convolutional neural networks (CNN) have been the standard approach to object detection from RGB images since their capabilities for training with GPU hardware was demonstrated in 2012 (AlexNet). Since then architectures such as YOLO and ResNet have been shown to be highly performant towards this task.

In this report my goal is to take a look at previously common methods for approaching the object detection task and see how they perform compared to more modern approach. This paper implements and evaluates a HOG-SVM based person detector on a recent work-environment focused dataset, the Safe Human dataset consisting of 17 different objects (SH17) [1].

Literature Review

Object detection from images involves two tasks: object classification and object locating. Both of these tasks involve the extraction of meaningful features from an image. Various techniques exists in literature to determine feature sets from an image.

This paper uses the Histogram of Oriented Gradients (HOG) method which captures local gradient structures from images (1467360). Other approaches to extracting features include Scale-Invariant Feature Transform and Speeded Up Robust Features algorithms.

Once features have been extracted from an image various options exists for then using those features to make a prediction on new data. Support Vector Machines (SVM) have commonly been used for this tasks as their able to efficiently classify higher dimensional data through the use of space transformations.

For a similar task of face detection, AdaBoost proved to be an effective method.

Problem Description

The SH17 dataset contains 8,099 annotated images with 75,994 instances across 17 classes. A person detection model is then implemented using HOG features extraction method and a classified with a linear SVM model. The performance of the model is evaluated using COCO benchmarks and compared to benchmarks obtained in the the SH-17 paper which uses YOLO9 and YOLO10 architectures.

Method

For the implementation notebooks, the code in (plvs2023hogdetection) is used as a primary reference.

The dataset splits the images into training and validation data along with their respective annotations. The annotations are parsed so only the "person" annotation data is used. The images are loaded and converted to greyscale, though it's worth noting that (HOGpaper) found this conversion shows little performance improvements.

The preparation of training data requires both positive and negative samples. Positive samples are extracted from the annotated person bounding boxes and assigned a label value of 1, while negative samples are derived from image regions outside these boxes and assigned a label value of 0. A sliding window approach generates feature extraction windows, using a fixed window size of (64, 128) pixels, which has been shown effective for person detection in previous studies Once a sampling window is calculated we then use the skimage HOG implementation to generate the feature data for that window. The HOG is The process is repeated for all the person bounding boxes in the image and for non-person regions.

An issue encountered by doing this is that without any limits on sample size we get a very large amount of negative samples for the dataset, greatly outnumbering the positive samples. This could potentiatliayy lead to overfittingm, but more practically this often caused the training notebook to crash as the more than 100 batch files were generated of a few GBs of size. Initially a Linear Support Vector Classification (SVC) was implemented for this training, but when the data batches were passed to the model for fitting, it would raise a memory input error. At that point, I investigate alternative optimizer approaches and found thus a SGD optimizer was used instead – this is discussed further in the Training section. In addition to the change of optimizer, a ratio of 5 negative to every positive sample was added to the sample generation to limit the number of negative samples.

For training, an sklearn pipeline is created using stochastic gradient descent (SGD) learning method and fit with a linear SVM.

As described in the preprocessing section, the large amount of samples resulted in a exceeding the memory capabilities in the original implementaion. Given the large amount of sample data available, the SGD optimizer method is chosen as recommended in the Sklearn guide

The default parameters are used for training including a tolerance of 0.001 and a max iteration of 1000. During the model fitting step the max iterations were met, thus convergence was not achieved in our training. Since SGD is sensitive to the feature scaling, the model inputs are scaled with the the standard scalar API of sklearn.

In total, 6479 images were loaded as training data with the following output: Total features: 64990 Positive samples: 10920 Negative samples: 54070 Feature dimension: 3780

Implementation

For the evaluation pipeline, the sliding window approach is again used but incorporates a image scaling pyramid to handle size variations of the target objects similar to that used in (plvs2023hogdetection). In the sliding window the HOG feature extractor method is again used with the same parameters as in the training step.

The trained SVM model provides confidence scores for each window using decision$_{f}unction(), whichreturnsthe$

Without post-processing the above method generates many "positive" predictions as seen in Figure X. This is problematic as each prediction is considered to be person thus negatively impacting the COCO metric calculations with the false positives. To account for this, post processing methods are added to standardizes the detection aspect ratios, merge overlapping boxes, and the do a non-maximum suppression to keep only the highest scoring boxes. It's worth noting that since the model used a scaler in it's pipeline, features with scores above 0.0 are considered within the "person" classification, while those below 0.0 are negative.

After 2 post processing steps, the prediction boxes are consolidated and merged resulting in results such as Figure X.

Results

The evaluation system implements COCO metrics including precision and recall at IoU threshold 0.5, mean average precision (mAP) at IoU 0.5, and mAP across IoU thresholds 0.5-0.95.

The results of the developed model and the YOLO models trained in SH-17 paper are displayed in Table 1 which comes from the SH-17 repository ().

The top performing model is bolded.

As can be seen the model performed terribly on the COCO metrics. One of the key reasons for the terrible metrics seems to be the excess prediction boxes being generated by the prediction code, and while these boxes do fall within the ground truth area as seen in Figure X, each is counted as a "person" class detection and thus skews metric calculation.

As described in the previous section, steps were taken to account for this such as NMS, box merging, and feature score threshold. Further improvements of the boxes post processing and hyper parameter tuning for the SGD optimizer could be done, but are beyond the scope of this report.

Model	Params (M)	Images	Instances	P (%)	R (%)	mAP50 (%)	mAP50-95 (%)
HOG-SVM	-	1620	15358	3.3	12.2	1.1	1.0
Yolo-8-n	3.2	1620	15358	67.5	53.6	58.0	36.6
Yolo-8-s	11.2	1620	15358	81.5	55.7	63.7	41.7
Yolo-8-m	25.9	1620	15358	77.1	60.5	66.6	45.7
Yolo-8-l	43.7	1620	15358	76.7	62.9	68.0	47.0
Yolo-8-x	68.2	1620	15358	77.1	63.1	69.3	47.2
Yolo-9-t	2.0	1620	15358	75.0	52.6	58.5	37.5
Yolo-9-s	7.2	1620	15358	73.6	60.2	65.3	42.9
Yolo-9-m	20.1	1620	15358	77.4	62.0	68.6	46.5
Yolo-9-c	25.5	1620	15358	79.6	60.8	67.7	46.5
Yolo-9-e	58.1	1620	15358	**81.0**	**65.0**	**70.9**	**48.7**
Yolo-10-n	2.3	1620	15358	66.8	53.2	57.2	35.9
Yolo-10-s	7.2	1620	15358	75.8	57.0	62.7	40.9
Yolo-10-m	15.4	1620	15358	71.4	61.4	65.7	43.8
Yolo-10-b	19.1	1620	15358	77.7	59.1	65.8	45.1
Yolo-10-l	24.4	1620	15358	76.0	61.8	67.4	46.0
Yolo-10-x	29.5	1620	15358	76.8	62.8	67.8	46.7

Conclusion

In this report, the author wanted to compare traditional machine learning approaches to more modern deep learning methods. A linear SVM classifier with a SGD optimizer was chosen in this case, as linear SVM have previously been used in many successful applications of person object detection. Features from the images were extracted using the HOG approach.

The final model performed extremely poorly when evaluated on the COCO object detection metrics. From qualitative inspection of the resulting prediction boxes, the author notes that the predictions do tend to fall within the ground truth area but there still remain very high scoring outliers which indicate further model tuning, training and post processing steps for detection boxes are needed.

AI Use Disclaimer

Claude 3.5 Sonnet was used in code generation for this paper. Code was edited and verified by the author.

References

[1] Hafiz Mughees Ahmad and Afshin Rahimi. Sh17: A dataset for human safety and personal protective equipment detection in manufacturing industry, 2024.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[4] Sam Plvs. Object detection via hog-svm, 2023.

[5] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: Image processing in Python. *PeerJ*, 2:e453, 2014.