

CS229 Final Project Report

A Multi-Task Feature Learning Approach to Human Detection

Tiffany Low tlow@stanford.edu

Abstract

We focus on the task of human detection using unsupervised pre-trained neural networks. The goal is to use multi-task feature learning to pre-train the network to identify people given image data. Intuitively, by learning features to identify subparts of human figures, such as arms, legs or torsos, these features can then be used for the learning task of classifying people. We train smaller convolutional networks on a dataset comprising of annotated video data of people in a variety of environments and poses and on existing datasets of labeled body part data in still images. The shared features that are learnt as a result of the multi-task feature learning are then applied to learning humans in an object classification task.

Related Work/Motivation:

Human detection has strong commercial and defense applications for a variety of purposes, including surveillance, analytics and entertainment. The task of detecting humans with high precision and recall is still a challenge. Multi-task learning has been used to boost accuracy and learning rate in classification tasks that are related. Unlike single-task learning, the goal is to simultaneously learn across all tasks, implicitly depending on a shared feature representation. This has been successfully used in applications such as detecting generic 3D objects, human faces/expressions (Torralba, Murphy and Freeman 2007) and 2D symbols (Kremp, Geman and Amit 2002).

There exists a large body of work based on recursive models (Zhu, Chen, et al. 2010, Zhu, Lin, et al. 2008), and on spring-based models for object classification/detection (Ramanan 2011). These approaches rely on a priori knowledge of the human model when tackling the object classification task. In particular, unsupervised learning has been used with a combination of levels of recursion (Zhu, Chen, et al. 2010) to learn the hierarchical dictionaries for a class of 26 object models automatically, with a shared common dictionary of 5 features. This multi-task learning approach results in both part-sharing and appearing-sharing, enabling more efficient learning and inference.

Dataset

The datasets were compiled from five sources: LAMDA (Sapp, Jordan and Taskar n.d.), CVPR (Hofmann and Gavrilu 2009), ETHZ (Eichner and Ferrari n.d.), H3D (Bourdev and Malik n.d.) and VideoPose (Sapp, Weiss and Taskar n.d.) datasets. For each of these datasets, the relevant body parts were identified and an appropriate bounding box of the relevant scale drawn over the body parts. The generated patches were then evaluated to removed occluded data (particularly from the CVPR dataset).

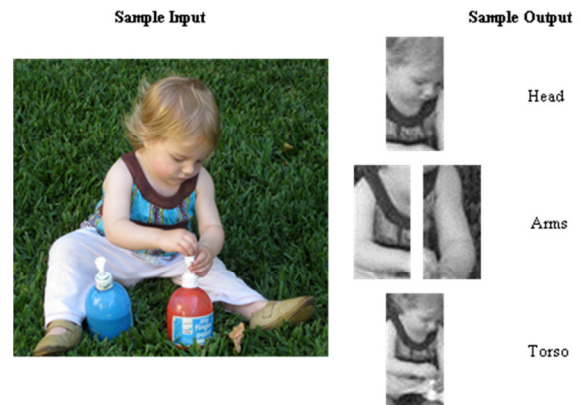


Figure 1. Sample image from H3D dataset with output training patches generated for head, arm and torso datasets.

The raw images for each of these body parts were on the order of 5000 training examples. There are 6 body part datasets: head, torso, arms and legs. The images are rescaled to patches of 64x32 pixels. This aspect ratio was chosen to better train patches for the classification task of detecting human figures, which have approximately a 2:1 ratio.

Convolutional Neural Networks:

Convolutional Sparse Networks as developed by (Kavukcuoglu, et al. n.d.), allow for faster training with similar performance as compared to traditional steepest descent sparse coding. The convolution operator allows for structures at a given orientation within the data to be modeled independent of their locations in the image.

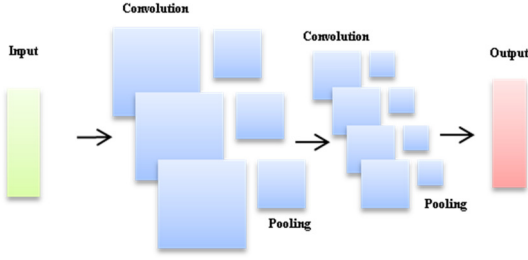


Figure 2. Sample CNN with two layers, one output node.

The work is on a multi-task training algorithm to produce a common set of features across the datasets for the purposes of a human detecting neural network. Motivated by work from (Zhu, Chen, et al. 2010), the network is trained using all the datasets simultaneously, to obtain common feature sets. This feature set is then evaluated as an input to the human detection network. The human detection system is a convolutional neural network (ConvNet), implemented by Zou, W. The networks trained use a single layer of hidden nodes, with features such as max-pooling and maximum suppression to boost classification accuracy.

Network Training:

All networks were 1-layer CNNs using 8 feature maps with dimensions 16x8, a pooling factor of 2 with response maps of size 8x4. For each output node, there are a total of 128 weights. An additional round of training was conducted using 16 feature maps for each network.

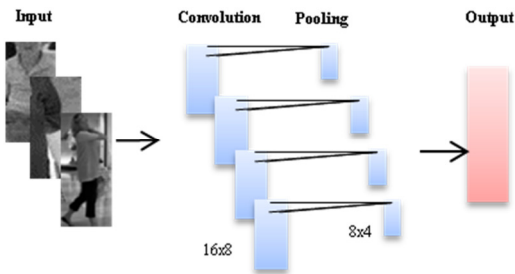


Figure 3. Network structure used for human classification using part data

The training sets consisted of positive training examples taken from the INRIA training set, negative examples from the INRIA training set, and training sets for the various parts (arms, legs, head and torso). Networks were trained for 300 iterations or until the learning rate was cut-off. In this case, the maximum iteration count was reached first for all networks trained. Refer to Figure 5. for explanations of the different networks trained. A total of 10 different types of configurations were considered.

Pipeline:

Given a sample input image, sliding windows of various sizes are run over the image. The input patch is then run through the detection network to observe if a person has been detected in that region. The activation maps over all the sliding window sizes are then agglomerated into bounding boxes greedily, starting with the box with the highest probability and then either merging it to nearby boxes or eliminating overlapping boxes otherwise. The output bounding boxes are then passed as input to the detection unit, which evaluates a hit if the bounding box approximates the annotation by more than 80%, or otherwise as a miss.

Experimental Results:

For the part-trained datasets, their performance was validated against a test set of patches from the part databases which were set aside from the training set (Figure 4.). The performance on the test and training sets demonstrate reasonable accuracy on part detection, given the limited feature size.

	8 Feature Maps	16 Feature Maps
PartsOnly	68.66	48.04
Head	60.0	60.49
Torso	60.53	40.15
Arms	91.82	93.02
Legs	62.17	25.91
PeopleParts	67.19	60.26
Head	60.27	40.33
Torso	55.73	58.22
Arms	88.77	85.47
Legs	62.95	49.30

Figure 4. Percentage accuracy on test set for part classification

The performance of the different networks was tested on a sample taken from the INRIA test set. These samples label people with bounding boxes of a similar aspect ratio (2:1), but only people who are upright or are clearly in the frame (cropped people parts are considered negative examples). First, sliding

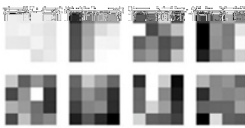
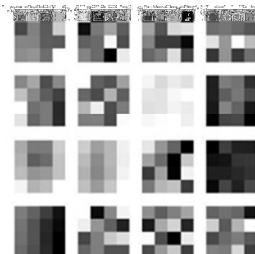

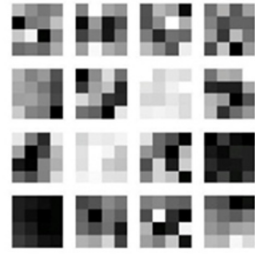

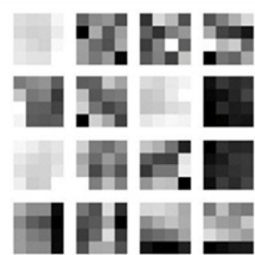

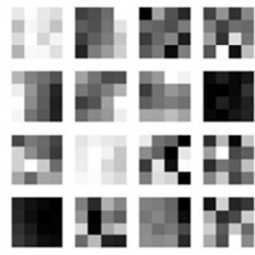

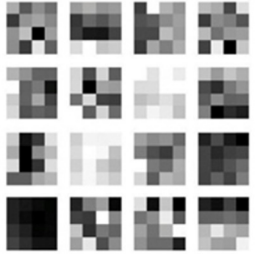
	8 Feature Maps	16 Feature Maps
Control Two output classes: People and Negatives. This was the baseline network trained using only INRIA data.		
PartsOnly Five output classes: Heads, Arms, Legs, Torsos and Negatives. Only the part dataset and the negative training dataset was used for training.		
PeopleNegative Two output classes: People and Negatives. The parts datasets were used to extend the negative training example sets.		
PeopleParts Six output classes: People, Heads, Arms, Legs, Torsos and Negatives. Only the part dataset and the negative training dataset was used for training.		
PartsInitial Two output classes: People and Negatives. The network is initialized using the trained network from PartsOnly. Training set used is the INRIA dataset alone.		

Figure 5. Network configurations used for experiments.

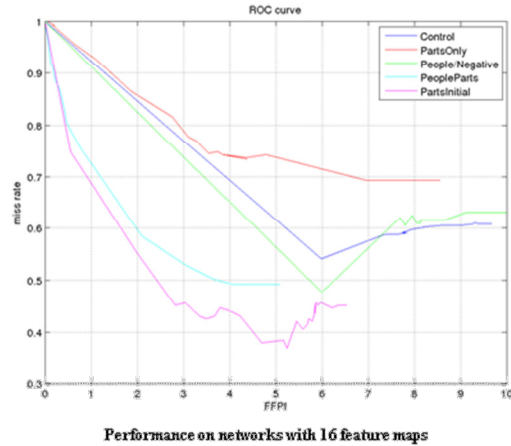
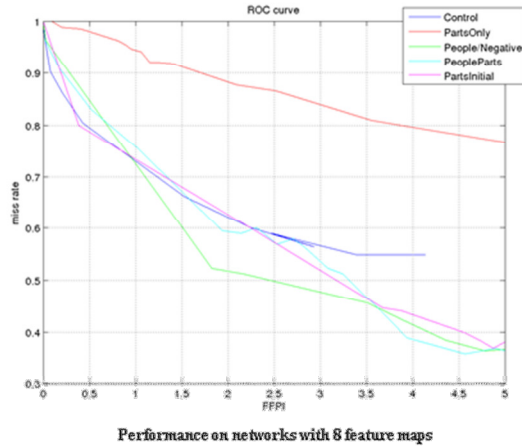


Figure 6. ROC Curves for 8-feature map networks and 16-feature map networks.

window detection was run at several image scales. Bounding boxes were only drawn for those parts that were above a variable threshold value. Based on the threshold value, we can obtain an ROC curve of the performance of the differently trained networks (Figure 6) in terms of their miss rate and false positive rate.

We observe that the addition of part data to the training set results in a significant performance gain for the trained networks. The parts-based networks achieve a lower miss rate to a minimum of 0.35, although at the cost of a higher FFPI. This suggests that these networks identify additional regions for people parts. As shown in Figure 7, the parts-based networks identify people who are only partially visible in the image (and thus not marked as positive in the INRIA test set). For more discussion on some sample false positives and bounding box limitations, refer to Figures 7 through 10, which give some bounding box annotations returned by the PartialInitial network.

It is unsurprising that the dataset trained only on part data and negative examples performs poorly on people detection. Here, the activation nodes for part labels are considered labels for person detection. However, we observe that using such a part-trained network as initialization weights for people detection results in significantly better performance on the 16 feature map networks.

Conclusion

The inclusion of part classification into the training of people detection networks is shown to improve performance of such detectors. A simple extension to the project would be to train part classifiers to reach

some degree of performance on a cross-validation set, and then applying this result to train a people detection network. Similarly, a network with a larger number of feature maps (or an additional convolutional layer) could be trained.

Acknowledgements

The project would not have been possible without the guidance and support of Will Zou and the collaboration with Rukmani Ravi (involved as part of research).

References

- Bourdev, Lubomir, and Jitendra Malik. H3D Dataset. n.d. <http://www.eecs.berkeley.edu/~lbourdev/h3d/> (accessed October 15, 2011).
- Eichner, M., and V Ferrari. ETHZ PASCAL Stickmen. n.d. http://www.vision.ee.ethz.ch/~calvin/ethz_pascal_stickmen/ (accessed October 15, 2011).
- Gourier, N, D Hall, and Crowley. "Estimating Face Orientation from Robust Detection of Salient Facial Features." ICPR International Workshop on Visual Observation of Deictic Gestures. 2004.
- Hofmann, M, and M Gavrilu. "Multi-view 3D Human Pose Estimation combining Single-frame Recovery, Temporal Integration and Model Adaptation." CVPR, 2009: 2214--2221.
- Kavukcuoglu, Koray, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michael Mathieu, and Yann LeCun. "Learning Convolutional Feature Hierarchies for Visual Recognition." n.d. Kremp, S., D. Geman, and Y. Amit. "Sequential Learning of Reusable Parts." 2002.
- Sapp, Ben, Chris Jordan, and Ben Taskar. LAMDa - Limbs Annotated from Movies Dataset. n.d. <http://vision.grasp.upenn.edu/video> (accessed October 10, 2011).
- Sapp, Benjamin, David Weiss, and Ben Taskar. Video Pose. n.d. <http://vision.grasp.upenn.edu/cgi-bin/index.php?n=VideoLearning.VideoPose> (accessed October 1, 2011).
- Tang, Lei. "Multi-Task Learning." n.d. www.public.asu.edu/~ltang9/presentation/multitask.pdf.
- Torralba, Antonio, Kevin Murphy, and William Freeman. "Sharing visual features for multiclass and multiview." IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 2007

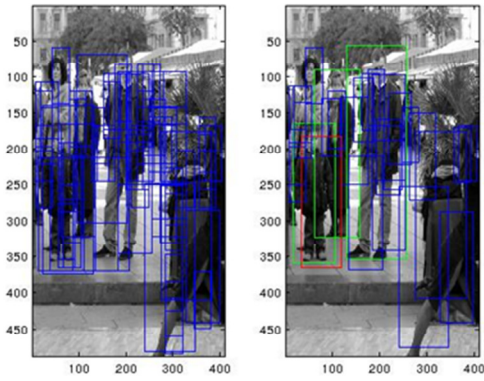


Figure 7. Detection of persons who are partially occluded (lady in the background, lady in the foreground).

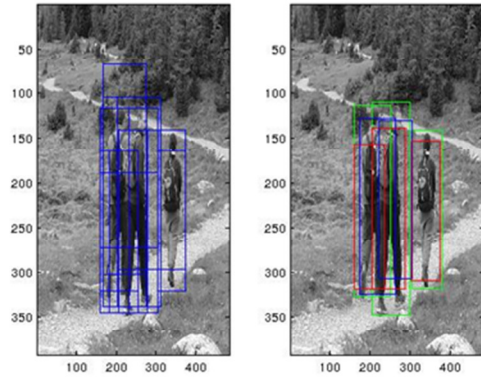


Figure 8. Successful example of person detection using part-trained network.

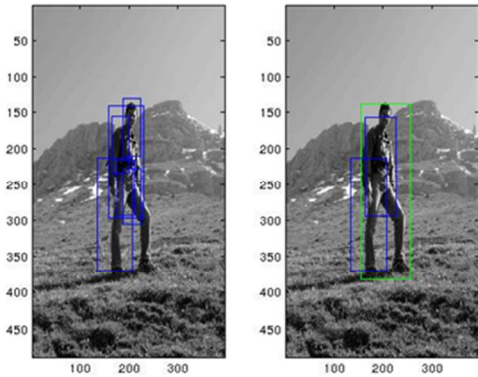


Figure 9. Miss due to incorrect bounding box resolution.

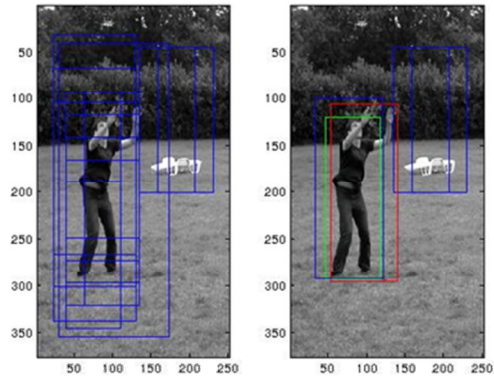


Figure 10. False positive of a park bench.