Master Thesis Specification
Multi-task Human Image Parsing

Student: Magnus thor Benediktsson, bened@kth.se
Supervisor: Hossein Azizpour
Examiner Danica Kragic

February 26, 2016

1 Background and Objective

Human detection is central to many computer vision tasks such as surveilance or assisted driving. Following the success of ConvNets on classification tasks recent work has focused on applying them to pose estimation[6][7] and semantic segmentation[4],[5] with state-of-the-art performance.

It is reasonable to expect semantic segmentation of humans and estimation of human joint locations to have similiar features. The main goal of this thesis will be to train a single multi-task fully convolutional network(FCN)[4][**?**] for human image parsing such as semantic segmentation and pose estimation. It is interesting to see whether the multi-tasking will benefit the individual tasks.

The sharing of features among various tasks may help regularize the network and dampen overfitting. This may enable us to learn larger networks. More tasks that could be incorporated include Human detection, Human Action Recognition etc. There is also an computational efficiency benefit from performing two tasks with a single network as many computations will be shared accross tasks.

Stat-of-the-art ConvNet pose estimators incorporate multiple network in connected in a cascade like manner to refine the joint locations[6][7]. We intend to construct a single network with a recursive structure to refine model predictions.

2 Research Question & Method

We want to investigate whether networks will benefit from multi-tasking. Can sharing weights between tasks reduce overfitting thus enable us to learrn larger networks. Can cues from one task improve the performance of another?

The first target will be two tasks, human semantic segmentation and human pose estimation. We will implement a single baseline architecture for both human segmentation and pose estimation seperately, assuming a single person in the image.

Next step is to train the same architecture for both tasks and compare to the individually trained networks.

The first challenge we face is how to train such a network without data that is both annotated for human segmentation and pose estimation. An initial approach would be to follow the example of [**?**] and alternate the training task between iterations. This would allow us to use separate training sets for each task.

Possible extensions to this include the use multiple scales, adding more tasks and including more than one person.

3 Evaluation & News Value

The dataset for Human Pose Estimation task will be the MPII[2] and for the Semantic Segmentation we will use Microsoft COCO[3]. We will compare the performance of the network to the baseline networks trained individually on their respective test sets.

The news value of this project would be the capacity of a network to multi-task on human related tasks and perhaps the performance improvement on individual tasks due to regularization from the sharing of weights.

4 Pilot Study

The literature study will focus on semantic segmentation and pose estimation using ConvNets as well as object detection and a general survey of the recent advances of neural networks. Included in the pilot study the student will participate in online tutorials on deep learning and the implementation of the baseline architectures mentioned above.

5 Conditions & Schedule

The experiments will be implemented in TensorFlow[1] amd run on a GPU provided by CVAP.

References

[1] Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Man, Rajat Monga, Sherry Moore, Derek Murray, Jon Shlens, Benoit Steiner, Ilya Sutskever, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Oriol Vinyals, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *None*, page 19, 2015.

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, Bernt Schiele, and Max Planck. 2D Human Pose Estimation : New Benchmark and State of the Art Analysis : Supplementary Material.

[3] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5):740–755, 2014.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2014.

[5] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning Deconvolution Network for Semantic Segmentation. *Arvix*, 1:1520–1528, 2015.

[6] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann Lecun, and Christoph Bregler. Efficient Object Localization Using Convolutional Networks. *Cvpr*, page 2014, 2015.

[7] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. 2016.