Master Thesis Specification
Multi-task Human Image Parsing Using Convolutional Networks

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

March 2, 2016

1 Background and Objective

Deep Convolutional Networks (ConvNet) have in recent years enjoyed a remarkable success at various vision related tasks such as image classification [4, 13, 6], object detection [10] and semantic segmentation [8, 9]. Conventionally object detection and localization are approached with Region based Convolutional Networks (R-CNN) [3] that use a separate separate algorithm such as Selective Search [16] to generate region proposal within an image upon which the detection ConvNet is applied.

More recently [10] introduced Region Proposal Network (RPN) that shares a Fully Convolutional Network (FCN) [8] with a a detection network that is able to simultaneously generate region proposals and extract features for classification of those regions. This method eliminates the need for a separate region proposal algorithm that had become a bottle neck of developing state-of-the-art object detectors [10]. Furthermore, the RPN efficiently models the two tasks of object proposal and detection with a single network.

Human detection is central to many computer vision tasks such as surveillance or assisted driving. It is reasonable to expect different human related vision tasks such as human semantic segmentation [8, 9] and joint locations [15, 17] to share many cues and features that can be efficiently combined in a single multi-task network solving more than one. Just as the object detection and region proposals from [10].

The main goal of this thesis is to explore the possibility of training a Deep Neural Network that efficiently solves multiple human related tasks. Training a single network for multiple tasks may affords the network to be trained on more data which may help regularize the network and it can be interesting to compare the performance of a multi-task network to a single-task one or if it can enable us to train larger models.

We will begin by focusing on the tasks of human semantic segmentation and pose estimation and then possibly extend towards human detection and human action recognition.

2 Research Question & Method

Section and its subsections will be the first priority of the thesis. Depending on the time available we will focus on the later questions which are ordered after priority.

2.1 Multi-Task Architecture

We will investigate architectures for a single baseline network that can be trained to solve two tasks separately. We will use the architectures of [8, 9] for segmentation and [15, 17] for pose estimation as starting points to design our architecture.

Next step is to train the architecture on both tasks simultaneously and compare its performance to the individually trained networks.

2.1.1 Optimization of the Architecture

Recent developments of novel architectures have shown to be beneficial to various aspects of ConvNets. Most notably [4, 12] have developed methods for optimizing extremely deep networks. Batch Normalization [5] has been successfully applied to ConvNets to alleviate the problem of covariate shift, it suggests that networks trained with Batch Normalization do not need to apply dropout [11] which has been an important regularization mechanism for DNN's in recent years. [13, 14] have introduced a variety of so called inception modules that break down 2D-convolutions into a set of 1D-convolutions in order to optimize model size.

2.1.2 Recursive Architecture

Stat-of-the-art ConvNet pose estimators incorporate multiple network in connected in a cascade like manner to refine the joint locations [15, 17]. We intend investigate the possibility of achieving similar behavior with a network with a recursive structure to refine model predictions. This can be looked first for the single-task trained network and later the multi-task network.

2.2 Regularization

2.2.1 Data Augmentation

Training sets are commonly enlarged with label-preserving transformations [6][this should perhaps point towards refs. within Krizhevsky] in order to reduce overfitting. Such transformations include adding random jitter to images, horizontal reflections or image cropping. Furthermore, we will consider more task specific data augmentation such as warping poses and segmentation masks.

2.2.2 Combined Training Sets

As a bi-product of multi-tasking training is that the model has access to more training data which may help battle overfitting. We want to know if performance of each individual task can be enhanced by training a multi-task network.

2.3 Feature Sharing

Features learned by by the network for one task may be latent in other tasks. Connecting diverged parts of the network at later stages may improve performance of individual tasks. One can reason semantic segmentation can help decrease false positives for landmark estimation by learning general body shapes through mining hard negatives. Or that human pose estimation can help to increase true positives of semantic segmentation by providing higher level of information about body parts.

3 Evaluation & News Value

The dataset for the human pose estimation task will be the MPII[2] and for the semantic segmentation we will use Microsoft COCO[7]. We will compare the performance of the network to the baseline networks trained individually on their respective test sets.

The capacity of a network to efficiently model multiple human related tasks is a new value onto itself as it could speedup systems that seek to solve both as demonstrated by [10]. Furthermore, if we can realize some of the speculated performance enhancements discussed above.

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. All cited papers in this will be included. Specifically the student will familiarize himself with various deep network architectures by reading up on the most recent developments to facilitate the design of the baseline architecture and further optimizations.

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] and 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] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based Convolutional Networks for Accurate Object Detection and Segmentation. *Pami*, 38(1):1–16, 2014.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *Arxiv.Org*, 7(3):171–180, 2015.

[5] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv*, 2015.

[6] Alex Krizhevsky, IIya Sulskever, and Geoffret E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information and Processing Systems (NIPS)*, pages 1–9, 2012.

[7] 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.

[8] 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.

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

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *ArXiv 2015*, pages 1–10, 2015.

[11] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15:1929–1958, 2014.

[12] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training Very Deep Networks. *arXiv*, pages 1–9, 2015.

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. *arXiv preprint arXiv:1409.4842*, pages 1–12, 2014.

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *arXiv preprint*, 2015.

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

[16] J. R R Uijlings, K. E A Van De Sande, T. Gevers, and A. W M Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

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