Chapter 6

Further Analysis

6.1 Future Works

For a stable real time object tracking with no unsteadiness of the marked object in the output a object detection rate (cat, no cat) of 95% is not enough. Therefore I suggest the following optimizations.

Data augmentation is something every sophisticated state of the art classifying network does before processing the dataset. It is a simple way to simulate effectively a 10 to 100 times bigger input dataset by performing some affine transformation on the input or especially helpful for training the deconvolutional part of the network by masking some areas of the images with gray patches so that the network can train to recognize partially not recognizable features and is able to focus on other parts of an object so that the marked output of the deconvolution captures a better silhouette of the object. Deep learning in general is definitely capable of continuing halve recognized features internally like Deep Dream proved.

The next step on the list is to try out different kind of optimizations I left out which are part of the original in 2015 published ResNet. The original ResNet also had a special architectural change for 50-layer, 101-layer, 152-layer networks I didn't try out [HZRS15, page 5, table 1.]. I tested over 100 different variations of the ResNet with and without checking other sources and came to the conclusion, that finding new optimizations is still possible but not worth the effort when there are known optimizations left to implement, that might on first glance be extensive to implement. But changing some parameters of the network is a good thing always to test, because small changes in the network parameters lead to big changes in the outcome, which is not the case for the randomly determined starting weight parameters of the network. Sometimes a change in the step size lead to a 10% better score and I didn't had any clue why. Training Neural Networks means to test a lot of scenarios without knowing always whats going on. This is also the reason why Neural Networks had a bad standing among scientists, because they tend to behave unempirical and are difficult to examine. But a deeper understanding of Neural Networks will help in the future to get some of the unknown nature of them under control.

Another improvement of the network could be dropout layers. They are not part of the original ResNet, but they have been proven to reduce overfitting effectively.

Many papers suggest to train the network with HSV color space images instead of RGB colored images. Solely because a human has RGB receptors on the retina doesn't mean that this is the optimal color space for object classification. Another change, which is implemented with a few changes, would be to add the 3 HSV color layers to the 3 RGB color layers so that the network receives 6 input layers. Some edges and patterns are

better visible in different color spaces, however I received a lower score on CIFAR-10 by using just the 3 HSV layers.
Speaking of CIFAR-10. Another dataset might be a good choice for testing the network. Because CIFAR-10 is in comparison with its 60000 images a very small dataset and a complex network could perform better on a bigger dataset like ImageNet with 14197122 images or on dataset with bigger sized input images like the SIFT10M Dataset.
Another way to create a bigger input dataset is to use depth information of the input image as an additional fourth input layer. Those depth information can be retrieved by a stereoscopically camera like the Kinect cam, which also uses a infrared laser projector combined with a monochrome CMOS Sensor to captures 3D video data even if the captured area has a bad illumination. If we want to use object tracking for robotic applications, where an autonomous robot has to evade moving objects by assuming the trajectory of those objects, we need that depth information. Most of todays applications are not capable of outmaneuver other moving objects nicely without standing still every time passes another trajectory, because for that they need an excellent object tracking ability in addition to the capability of real time outmaneuvering. This object tracking ability has to be able to subtract out the own movement on the input images. 360 degree cameras or parabolic mirror cameras would be needed in addition to that if we want to realize autonomous driving with this approach.
Differential Camera Tracking or infrared cameras could also be used for additional input information especially to create self generating ground truth dataset.
With all this optimizations to do a reliable object tracking should be makeable. The last future work I would consider for this approach is to build a simulation, which generates all the input data for a specific scenario, like car driving or crowd evasion. This would enable the AI agent to interact with the simulation, so that we can use reinforcement learning, because for reinforcement learning we always need something to interact with to form its policy graph.

6.2 Summary and Conclusion

To summarize this thesis I will recapitulate the main thoughts of every chapter and the main thoughts I had during the implementation of the approach. In the first chapter I gave an short overview about the motivation and challenges of the thesis. The motivation was very simple: Implementing an deconvolutional version of a ResNet without loosing the advantages of the ResNet. The challenges were versatile and complex. In chapter [2], I introduced other approaches from published related works, which solve some of this challenges. For implementing those solutions it is important to understand the technical and mathematical background of CNNs and Deconvolutional Neural Networks. But for solving the problem with those irreversible nonlinear properties of the ResNet those backgrounds weren't enough. Thinking out of the box was necessary and so I came up with[1] a morphologic network transformation, that linearized the ResNet and perseveres the pretrained training progress of the ResNet. After that I tested the deconvolution with mixed results. At first I saw nothing at all on the object localization output of the network, because of different inference patterns that interfere with the output. But after resolving this problem, I was a bit skeptical about the marked

[1] This idea was based on the thought of using a linear network like the VGG Net instead of the ResNet, after I tried different approaches for reverting the ResNet for a couple of weeks with no good results. But before I shut the ResNet down completely I gave it a last chance as a pretraining network. However I look forward to try out some other ways for a true inversion of the ResNet.

areas, because they were far away from capturing the silhouette of an object perfectly like trainable segmentation networks. But the deconvolutional part of the network is just an information extraction part and nevertheless useful for lesser accurate tracking methods, that just needs some coordinates that lie inside an object space. Those information are contained by the network in any case and should be used before going through the hassle of creating a ground truth dataset for every application necessary. After optimizing the output of the deconvolution I got results which were finally satisfiable for carrying out a test about object tracking. I didn't implement all the details for a real time object tracking, but with the data I achieved on single unrelated images I'm sure related images and more data would be beneficial in any way for the accuracy of the object tracking, that this theoretical approach with the optimizations I suggest in chapter [6.1] would succeed.

Bibliography

[BKC15] V. Badrinarayanan, A. Kendall, and R. Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation". In: *ArXiv e-prints* (Nov. 2015). arXiv: 1511.00561 [cs.CV].

[Ben15] Rodrigo Benenson. *What is the class of this image ?* 2015. URL: http://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html#43494641522d3130.

[chr17] (chrisbasoglu). *Object detection using Fast R-CNN.* 2017. URL: https://github.com/MicrosoftDocs/cognitive-toolkit-docs/blob/live/articles/Object-Detection-using-Fast-R-CNN.md.

[DV16] V. Dumoulin and F. Visin. "A guide to convolution arithmetic for deep learning". In: *ArXiv e-prints* (Mar. 2016). arXiv: 1603.07285 [stat.ML].

[Fer15] Prof. Dr. A. Ferrein. "Einführung in die Künstliche Intelligenz Lernen mit Beispielen — Neuronale Netze (unpublished script)". In: (2015).

[FLR+17] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. "DSSD : Deconvolutional Single Shot Detector". In: *ArXiv e-prints* (Jan. 2017). arXiv: 1701.06659 [cs.CV].

[GDDM13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *ArXiv e-prints* (Nov. 2013). arXiv: 1311.2524 [cs.CV].

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* http://www.deeplearningbook.org. MIT Press, 2016.

[Gra14] B. Graham. "Fractional Max-Pooling". In: *ArXiv e-prints* (Dec. 2014). arXiv: 1412.6071 [cs.CV].

[GMW15] M. D. Gregory, S. V. Martin, and D. H. Werner. "Improved Electromagnetics Optimization: The covariance matrix adaptation evolutionary strategy." In: *IEEE Antennas and Propagation Magazine* 57.3 (June 2015), pp. 48–59. ISSN: 1045-9243. DOI: 10.1109/MAP.2015.2437277.

[HZRS15] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *ArXiv e-prints* (Dec. 2015). arXiv: 1512.03385 [cs.CV].

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *twtygqyy/resnet-cifar10.* 2016. URL: https://github.com/twtygqyy/resnet-cifar10.

[IS15] S. Ioffe and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *ArXiv e-prints* (Feb. 2015). arXiv: 1502.03167 [cs.LG].

[JSD+14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. "Caffe: Convolutional Architecture for Fast Feature Embedding". In: *arXiv preprint arXiv:1408.5093* (2014).

[KF15] A. Karpathy and L. Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* June 2015, pp. 3128–3137. DOI: 10.1109/CVPR.2015.7298932.

[KB14] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *ArXiv e-prints* (Dec. 2014). arXiv: 1412.6980 [cs.LG].

[Kri12] Alex Krizhevsky. "Learning Multiple Layers of Features from Tiny Images". In: (May 2012).

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25.* Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[LMP01] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning.* ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. ISBN: 1-55860-778-1. URL: http://dl.acm.org/citation.cfm?id=645530.655813.

[LKF10] Y. LeCun, K. Kavukcuoglu, and C. Farabet. "Convolutional networks and applications in vision". In: *Proceedings of 2010 IEEE International Symposium on Circuits and Systems.* MISSING PUBLISHER, May 2010, pp. 253–256. DOI: 10.1109/ISCAS.2010.5537907.

[LBD+90] Yann Lecun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L.D. Jackel. "Handwritten digit recognition with a back-propagation network". In: *Advances in Neural Information Processing Systems (NIPS 1989), Denver, CO.* Ed. by David Touretzky. Vol. 2. Morgan Kaufmann, 1990.

[LFDA16] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. "End-to-End Training of Deep Visuomotor Policies". In: *Journal of Machine Learning Research* 17.39 (2016), pp. 1–40. URL: http://jmlr.org/papers/v17/15-522.html.

[LAE+15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. "SSD: Single Shot MultiBox Detector". In: *ArXiv e-prints* (Dec. 2015). arXiv: 1512.02325 [cs.CV].

[LSD14] J. Long, E. Shelhamer, and T. Darrell. "Fully Convolutional Networks for Semantic Segmentation". Nov. 2014.

[MP43] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (Dec. 1943), pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/BF02478259. URL: https://doi.org/10.1007/BF02478259.

[MHGK14] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. "Recurrent Models of Visual Attention". In: *ArXiv e-prints* (June 2014). arXiv: 1406.6247 [cs.LG].

[NHH15] H. Noh, S. Hong, and B. Han. "Learning Deconvolution Network for Semantic Segmentation". In: *ArXiv e-prints* (May 2015). arXiv: 1505.04366 [cs.CV].

[ODO16] Augustus Odena, Vincent Dumoulin, and Chris Olah. "Deconvolution and Checkerboard Artifacts". In: *Distill* (2016). DOI: 10.23915/distill.00003. URL: http://distill.pub/2016/deconv-checkerboard.

[Ola14a] Christopher Olah. "Conv Nets: A Modular Perspective". In: (July 2014). URL: http://colah.github.io/posts/2014-07-Conv-Nets-Modular/.

[Ola14b] Christopher Olah. "Neural Networks, Manifolds, and Topology". In: (Apr. 2014). URL: http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/.

[RHGS15] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *ArXiv e-prints* (June 2015). arXiv: 1506.01497 [cs.CV].

[Roj96] Raúl Rojas. *Neural Networks: A Systematic Introduction.* New York, NY, USA: Springer-Verlag New York, Inc., 1996. ISBN: 3-540-60505-3.

[Rud16] S. Ruder. "An overview of gradient descent optimization algorithms". In: *ArXiv e-prints* (Sept. 2016). arXiv: 1609.04747 [cs.LG].

[SMDH13] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. "On the Importance of Initialization and Momentum in Deep Learning". In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28.* ICML'13. Atlanta, GA, USA: JMLR.org, 2013, pp. III-1139–III-1147. URL: http://dl.acm.org/citation.cfm?id=3042817.3043064.

[SLJ+14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going Deeper with Convolutions". In: *ArXiv e-prints* (Sept. 2014). arXiv: 1409.4842 [cs.CV].

[Tal12] N.N. Taleb. *Antifragile: Things That Gain from Disorder.* Incerto. Random House Publishing Group, 2012. ISBN: 9780679645276. URL: https://books.google.de/books?id=5fqbz%5C_qGi0AC.

[Tur10] et al. Turian. *t-SNE visualizations of word embeddings.* http://metaoptimize.s3.amazonaws.com/cw-embeddings-ACL2010/embeddings-mostcommon.EMBEDDING_SIZE=50.png. [Online; accessed 11-September-2017]. 2010.

[TRB10] Joseph Turian, Lev Ratinov, and Yoshua Bengio. "Word Representations: A Simple and General Method for Semi-supervised Learning". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* ACL '10. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 384–394. URL: http://dl.acm.org/citation.cfm?id=1858681.1858721.

[Wik17] Wikipedia. *AI winter — Wikipedia, The Free Encyclopedia.* [Online; accessed 8-September-2017]. 2017. URL: https://en.wikipedia.org/w/index.php?title=AI_winter&oldid=796465213.

[YPW+07] H. Yang, M. Pollefeys, G. Welch, J. M. Frahm, and A. Ilie. “Differential Camera Tracking through Linearizing the Local Appearance Manifold”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition.* June 2007, pp. 1–8. DOI: 10.1109/CVPR.2007.382978.

[ZF13] M. D Zeiler and R. Fergus. “Visualizing and Understanding Convolutional Networks”. In: *ArXiv e-prints* (Nov. 2013). arXiv: 1311.2901 [cs.CV].

[ZKTF10] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. “Deconvolutional networks”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* June 2010, pp. 2528–2535. DOI: 10.1109/CVPR.2010.5539957.

[ZC12] Y. Zhang and T. Chen. “Efficient inference for fully-connected CRFs with stationarity”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition.* June 2012, pp. 582–589. DOI: 10.1109/CVPR.2012.6247724.

[Zha16] Zhang; Jianming; Lin; Zhe; Brandt; Jonathan; Shen; Xiaohui; Sclaroff; Sta; Zhang. “Top-down Neural Attention by Excitation Backprop”. In: *European Conference on Computer Vision(ECCV).* 2016.

FH AACHEN
UNIVERSITY OF APPLIED SCIENCES