Compression Methods for Monocular Depth Estimation

Group 3 | M.B. Büller | T. van Dijk | R.R. Feith | J.J. Hagenaars | D.B. de Jong | M.J. Mollema

TU Delft

Introduction

- **CNN + monocular camera** are used for depth estimation
- **State-of-the-art performance**, even compared to stereo imaging
- **Due to computational burden**: not applicable in small, on-board systems
- **Our goal**: reduce computational complexity while maintaining accuracy

"How are the **accuracy and size** of the MonoDepth network[1] affected by the various **model compression methods**?"

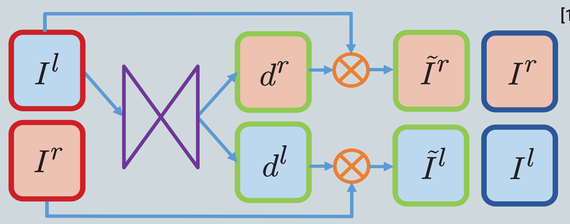
Methods

Removing layer(s)
- Reduce parameters by removing intermediate layer(s)
- Train from scratch on KITTI

Pruning
- Reduce memory footprint by setting small weights to 0 (increase sparsity)[3]
- Retrain from checkpoint on KITTI subset

SqueezeNet[2]
- Reduce 3x3 filter channels by preceding them with multiple 1x1 filters
- Train from scratch on KITTI

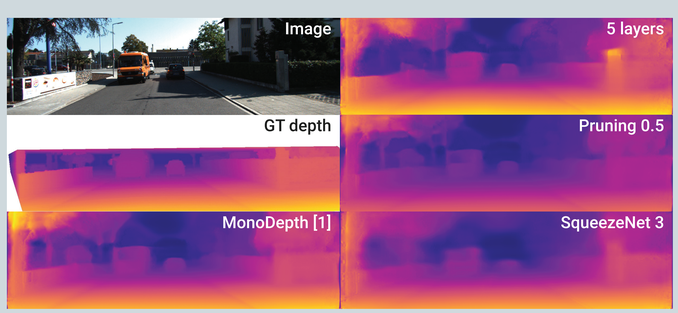
MonoDepth Network

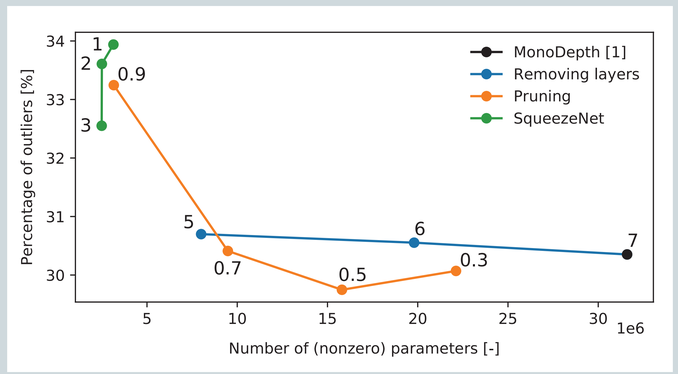


[1]

Left image used to predict disparities for both images, thus enforcing **mutual consistency**

Results





Discussion

Removing layers
- Reducing number of parameters up to 25% without significant loss in performance
- Disparity maps become less 'smooth' with decreasing number of layers

Pruning
- Similar and sometimes even better performance up to a sparsity of 0.7, which might indicate overfitting of original network
- However, clear visual reduction in detail and contrast from a sparsity of 0.5 onwards

SqueezeNet
- Removes a lot of parameters, up to 90% of original network
- Suffers from a somewhat larger reduction in performance
- Disparity map becomes more blurry, likely due to the change in decoder (the encoder of the original Squeezenet was good)

References

[1] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In CVPR, volume 2, page 7, 2017.

[2] F. N. Iandola et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. arXiv preprint arXiv:1602.07360, 2016.

[3] M. Zhu and S. Gupta. To prune or not to prune: exploring the efficacy of pruning for model compression. arXiv preprint arXiv:1710.01878, 2017.