Review on “Cascade Residual Learning: A Two-stage Convolutional Neural Network for Stereo Matching”

Authors: J. Pang, W. Sun, J. SJ. Ren, C. Xang, Q. Yan

# Short Summary

The authors of this paper propose a novel cascading CNN architecture containing two stages to address the challenge of generating high-quality disparities for ill-posed regions in a stereo matching task. This network may be referred to as Cascade Residual Learning (CRL).

The first stage is essentially an up-convolution module that produces fine-grained disparities. It builds upon a previous work DispNetC, a CNN which possesses an hour-glass structure with skip connections and a correlation layer at the end. The authors modify this network by including extra deconvolution layers to magnify disparity yielding estimates that are at the same size of the input image. The first stage is called “**DispFulNet**”. The input into this layer is a stereo (left and right) pair of images and outputs an initial disparity as well as a synthesized version of the left image with that disparity. The error between the original left image and the synthesized version are also fed into the second layer.

In the second stage, the disparity is corrected using residual signals at multiple scales. The output of this stage is a residual signal which is used the generate the new residual by taking the sum of the initial disparity and the residual. This allows the network to focus on learning the residual instead of trying to learn the disparity directly (experimentally shown to improve performance). This stage also has an hour-glass structure and produces residuals across multiple scales. The second stage is called “**DispResNet**”.

The method ultimately achieves state-of-the-art on the KITTI 2015 stereo dataset (as of August, 2017) and takes 0.47 seconds with a Nvidia GTX 1080 to obtain a disparity image.

# Main Contributions

* Proposed Cascade Residual Learning (two stage approach) for estimating disparity
  + Proposed first stage DispFulNet which improves upon DispNetC
  + Proposed second stage DispResNet to learn residuals and experimentally show that this boosts performance
* Achieved state-of-the-art on KITTI 2015 stereo dataset at time of publishing (August, 2017)

# High-Level Evaluation of Paper

First and foremost, this paper does not do a good job of explaining what disparity is (or why one would want to estimate it) which makes understanding its contents very difficult to follow for a reader that hasn’t studied computer vision before. While I may not be the intended audience for such a paper, the central topic of the paper, at the very least, should be accessible by any reader. Otherwise, although the paper is dense and requiring CV knowledge, it’s possible to follow along and understand what the authors are proposing. On this note, the paper places more emphasis on experimental design and discussion of results. Fortunately, these sections are explained well though the tables are a bit difficult to interpret at first glance. Additionally, it does not seem like the code for the proposed network is publicly available; if it is, the link is missing in the paper.

# Discussion on Evaluation Methodology

In terms of designing the experiment, the authors use a complex training schedule with different datasets and fine-tuning. While it’s theoretically possible to recreate their results, it wouldn’t be easy to do so. Their model is first compared to the individual building blocks that were used to construct the Cascade Residual Learning (CRL) architecture. Evaluation is performed on the FlyingThings3D, MiddleBury and KITTI 2015 datasets. While their proposed model performs the best, some of their claims are a bit exaggerated given that most of their improvements yield a very small improvement <1 error. For example, DispFulNet over DispNetC yields approximately 0.2 improvement in EPE though sometimes up to 2% in 3PE. It’s good that both metrics are provided as they allow for some interpretability. Over DispNetC, CRL obtains 0.52 improvement in EPE and 3.47 in 3PE. It’s not quite clear to me whether these are significant improvements over prior works as the improvements could be explained by the complex training methodology.

# Possible Directions for Future Work

As mentioned in the paper, a possible future direction would be to include new mechanisms in the network (such as a left-right consistency check) to further performance. However, these would need to be designed in a robust fashion to yield a tangible improvement as CNN performance is already high.