Critical Analysis Report in Single Image Super Resolution

Dong et al's SRCNN[1] marked the beginning of a new trend of using convolutional neural nets (CNNs) in image super resolution (SR). Following the introduction of the SRCNN, one trend was to build deeper networks. Kim et al[2] showed their very deep network (VDSR) of 20 weight layers resulted in significantly improved performance compared to 3 layers in SRCNN (+0.87 dB at x2 scale factor). The approach used context from larger regions of the image into account and their model was able to work at arbitrary (continuous) scale factors, while SRCNN required a single model per scale factor. Most importantly, they showed that the attempts by Dong et al. improve performance with deeper networks had failed due to learning rates that were too low, and that their deeper models had likely failed to converge. Their usage of adjustable gradient clipping additionally meant they could train their much deeper network in just 4 hours compared to SRCNN's several days.

Evidence that deeper networks could improve performance opened the door to further work that further deepened the networks and made architectural improvements changes to deep networks to improve the performance of these deep CNN models. Memnet[3] increased the number of layers to 80, at the time the deepest network in super resolution. and introduced a memory block. This was designed to address a problem seen in previous models where each layer is mainly influenced by the layer directly before it – resulting in mid or high frequency information being lost between layers.Their solution was to mimic long term memory effects by mining persistent memory through an adaptive learning process – specifically, by a gate unit that adaptively controlled how much previous states should be reserved and how much of a given state should be stored.

Another trend, inspired by advances in image classification tasks made by He et al with ResNet[4], J. Kim et al. proposed DRCN[5]. Due to the multiplicative nature of chained gradients, other deep models suffered from exploding or vanishing gradient problems, where gradients (respectively) grow or fall very quickly, and the model is not able to improve it's performance[6]. The DRCN used recursive layers and skip connections to alleviate these problems and most subsequent models used residual learning (see table 2 [7])

SRCNN relied on upsampling the low resolution (LR) image using bicubic interpolation and then reconstructing the detail in a high dimensional HR feature space. As the detail was reconstructed in a higher dimensional space, SRCNN requires more parameters than if it were operating in LR dimensional space - thus greatly increasing the computational complexity and resulting in sub real time performance. C. Dong et al[8] proposed their FSRCNN which could run in real time on a generic CPU. Their approach consisted of replacing the bicubic interpolation with a deconvolution layer and adding a shrinking layer at the beginning of their network and an expanding layer at the end. Unlike VDSR, their network could not be used at different scaling factors, but it has the advantage over SRCNN that only the deconvolution layer needed to be fine-tuned between different scaling factors and the remaining convolution layers could be frozen – meaning a single model could work at different scale factors without requiring an entirely new network.

This was expanded on by Lai et al's Laplacian Pyramid Super-Resolution Network (LapSRN)[9]. In this network each pyramid level takes a coarser resolution feature map as input, predicts the high frequency residuals and uses transposed convolutions to upsample

to a higher level. Most previous SR approaches construct HR images in a a single step, in contrast theirs progressively upsamples the LR image. This additionally meant that LapSRN was able to produce images at different scaling factors by using or bypassing residuals at different levels, as required. Furthermore, they also show that their model is faster at inference than most previous approaches and with a similar speed to FSRCNN.

One, as yet unmentioned, aspect of these models is the loss function used to guide model optimisation. Early models, including SRCNN used pixel loss (ie comparing pixel-wise differences between images) based on the L2 norm. Lai et al show that the L2 norm results in perceptually worse results, instead using the Charbonnier loss function[9]. This is because the L2 norm penalises larger errors at the expense of permitting smaller errors – resulting in smoothing artifacts. From this point on, many new proposed systems rely on the L1 norm as opposed to the dominant L2 norm beforehand (see table 2 [7]).

Although there were many improvements in these models, they all share a common feature in that they rely on supervised learning to predict HR images. Following Goodfellow et al.[10] Ledig et al presented work[11] as part of a new trend in using Generative Adversarial Networks (GANs) to synthesise HR images. They improve on other work[12][13] to develop a perceptual loss function that consists of the weighted sum of a content loss and an adversarial loss component. Thus, their network attempts to fool a discriminator network by favouring solutions that reside on the manifold of natural image. Their method is able to hallucinate fine structure at very high scale factors(8x) where the algorithm needs to generate plausible detail that couldn't be inferred from the LR image – this is explicitly supported by Lai et al's highlighting that previous SR algorithms, including their own, is unable to generate this detail. Further work with GANs could provide a route into producing systems that are able to generate detail far beyond what could be inferred from the LR image. However, as Ledig et al. acknowledge, their work may not be suitable for medical or surveillance applications. If this trend continues, more work on the ethics of the fine line between enhancing images and hallucinating convincing detail would be interesting and necessary.

This work was furthered by GAN networks that can learn from unsupervised, unpaired LR/HR images[14]. This represents a new direction to solve the problem of the difficulty of generating HR/LR datasets – traditionally solved by downsampling HR images – but with the disadvantage that models learn the inverse process of the predefined degredation. Other work[15] has shown that models trained on particular types of images do not generalise well, so work to allow a wider range of datasets could lead to performance improvements.

The use of GANs showed the deficiencies of existing image quality assessment techniques. PSNR is one of the most popular measurements[7] but Ledig at al showed that their SRGAN had a lower PSNR score than existing methods but had superior performance according to human perception with Mean Opinion Score (MOS) testing. Techniques to discriminate at a more perceptual level includes content loss [16][17] which, instead of looking at pixel-wise differences, considers the semantic differences between images using a pre-trained image classification network. Other approaches to assess detail include texture loss [18][19] which is more concerned with ensuring textures are accurately reproduced. Image quality assessment is an unsolved problem that requires further work. An ideal solution would take large and small scale regions into account and be able to achieve human level perceptual discrimination between different models to ensure it is applicable to a wide range of models.

References

1: C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deepconvolutional network for image super-resolution,” in ECCV,2014.

2: Kim, J., Kwon Lee, J. and Mu Lee, K., 2016. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1646-1654).

3: Tai, Y., Yang, J., Liu, X. and Xu, C., 2017. Memnet: A persistent memory network for image restoration. In Proceedings of the IEEE international conference on computer vision (pp. 4539-4547).

4: He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

5: Kim, J., Lee, J.K. and Lee, K.M., 2016. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1637-1645).

6: Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult." IEEE transactions on neural networks 5, no. 2 (1994): 157-166.

7: Wang, Z., Chen, J. and Hoi, S.C., 2020. Deep learning for image super-resolution: A survey. IEEE transactions on pattern analysis and machine intelligence.

8: Dong, C., Loy, C.C. and Tang, X., 2016, October. Accelerating the super-resolution convolutional neural network. In European conference on computer vision (pp. 391-407). Springer, Cham.

9: Lai, W.S., Huang, J.B., Ahuja, N. and Yang, M.H., 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 624-632).

10: Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial networks. arXiv preprint arXiv:1406.2661.

11: Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. and Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4681-4690).

12: Johnson, J., Alahi, A. and Fei-Fei, L., 2016, October. Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision (pp. 694-711). Springer, Cham.

13: Bruna, J., Sprechmann, P. and LeCun, Y., 2015. Super-resolution with deep convolutional sufficient statistics. arXiv preprint arXiv:1511.05666.

14: Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C. and Lin, L., 2018. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 701-710).

15: Takano, N. and Alaghband, G., 2019. Srgan: Training dataset matters. arXiv preprint arXiv:1903.09922.

16: Johnson, J., Alahi, A. and Fei-Fei, L., 2016, October. Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision (pp. 694-711). Springer, Cham.

17: Dosovitskiy, A. and Brox, T., 2016. Generating images with perceptual similarity metrics based on deep networks. arXiv preprint arXiv:1602.02644.

18: Gatys, L.A., Ecker, A.S. and Bethge, M., 2015. Texture synthesis using convolutional neural networks. arXiv preprint arXiv:1505.07376.

19: Gatys, L.A., Ecker, A.S. and Bethge, M., 2016. Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2414-2423).