Structure Analysis of Deep Image Prior

Zhuofan Xi

Abstract—While deep convolutional neural network's power of solving inverse image problem is mostly imputed to the ability to find patterns within large datasets, the Deep Image Prior paper [1] introduces a fresh new idea. They simply "train" a randomly initialized network on a corrupted image before overfitting and get satisfactory restoration results. In this project I analyze the impact of network structure on the output by modifying some components in the original network and finally build a new model with slightly better performance.

Index Terms—Computational Photography, U-net, ResNet, Denoising

1 Introduction

DENOISING is a classical but still highly active and useful topic in computational imaging. The goal is to recover a clean image from a noisy observation. One common assumption is that the noise is additive white Gaussian noise with potentially unknown standard deviation. Over the past decades models have been exploited for modeling image priors such as state-of-the-art BM3D method [2]. Recently CNN learning based method also show great performance in denoising problems. Though CNN models do not require explicit prior information, most CNN models require training on large datasets. As a result many believe that the power of CNNs stems from their ability to extract underlying patterns in datasets and may give unpredicted results if they have never seen similar samples before.

Deep Image Prior [1] challenges this notion by showing that fitting a randomly initialized network on the noisy image can produce surprisingly great results. However there are two issues to be resolved. One is that [1] says the structure of the network does have an impact on the result but they choose U-net like model without giving explanations. The other one is the problem of overfitting. Deep neural networks typically have at least millions of parameters and therefore are prone to overfitting if only trained on a single image. In this project I mainly deal with these issues and finally build a model that absorbs the valueable components of the original architecture and alleviates overfitting.

2 Related Work

Over the past decade numerous CNN architectures have been proposed. So far ResNet [3] designed by He et al. is one of the most performant network for image classification tasks. ResNet basically consists of sequential residual blocks that only differ from plain convolutional blocks (Conv, BatchNorm and activation layers) in that they add the input and output of block together. This identity mapping solves the degradation training problem in very deep neural networks. They also investigated many different types of shortcut connections in [4]. Then Zhang et al. adapt residual network to the denoising problem [5] by training network on a 400 images dataset to directly predict the Gaussian noise. Their model does not have downsampling and upsampling operations and the outputs of blocks have the same size.

Another popular network is U-net [6]. It is also the default choice of Deep Image Prior. This hour-glass shaped model contains encoding and decoding process and skip connections between them. Variants of U-net are reviewed in [7]. One inspiring example is fusing ResNet and U-net together to get ResUnet and it proves feasible in [8].

Two major techniques to help training are batchnorm [9] and dropout [10]. However combination of these two weapons fails to obtain extra reward in many architectures, and [11] suggests that dropout layer can be applied after BN.

3 Proposed Method

3.1 Evaluation Metrics

Although original paper demonstrates effectiveness of the network in many imaging problems, this project mainly focuses on denoising task. Therefore primary evaluation metrics is the Peak Signal to Noise Ratio (PSNR) of the network output with respect to the ground truth. Due to the randomness in initialization and training process, PSNR of each iteration output may fluctuate and the weighted averaged output

$$AVG_i = \alpha AVG_{i-1} + (1-\alpha) IMG_i$$

is a better criteria with less variance. In the experiments I set $\alpha = 0.99$. Note that the model will first learn the image then overfit the noise and PSNR will first increase then decrease. I use **peak** averaged PSNR to measure performance.

In addition to the peak PSNR, flatness around peak point, or in other words number of iterations where PSNR is close to peak, should also be a qualitative component in evaluation. Since we do not have ground truth in practice, we will not know when the model starts to overfit. So the more iteration PSNR stays around peak, the better result we will get.

• *Zhuofan Xi is a master student at Institute for Computational and Mathematical Engineering, Stanford University.*
E-mail: zfxi@stanford.edu

3.2 Remove hourglass structure

The original paper uses 2×2 upsampling and downsampling with convolution stride=2 to achieve hourglass shape. What if we remove the sampling operations and therefore keep more feature map information? After some experiments I find that the hourglass shape does not improve final PSNR ratio but greatly reduces training time. This is because the convolution on the original 512×512 image size is very time consuming.

3.3 ResNet-like structure

Another characteristics of U-net is the skip connection. Without skip connections the network becomes plain sequential convnet. So to analyze impact of skip connections I change the network to ResNet. The simplified 3-block structure is shown in Fig 1. Red square contains layers in a residual block. In numerical experiments I use 5 blocks.

3.4 Final Architecture

ResNet and U-net have their own advantages. Fusing them together may be a good idea and I implement this idea into my final model. Batchnorm layers are incorporated in each block: Conv→BN→Act→Conv→BN→Act. Despite reported failures of combining dropout and batchnorm together in many architectures, I still believe it is worth trying. Dropout layers are placed between building blocks. A simplified 3-layer residual U-net is shown in Fig 2. Black block represents a residual encoding block. Orange line represents the add operation within a residual block. Green block represents a residual decoding block. Red line represents skip connections from encoding block to decoding block. In numerical test I use 5 layers. To compare the performance of using residual blocks and dropout layers, I make them as selective options.

4 Experimental Results

One technical issue is that the model does not converge on some GPUs in `torch.cuda.FloatTensor` precision. Therefore I conduct experiments on NVIDIA Tesla V100 GPU with `torch.cuda.DoubleTensor`. As stated before, ResNet here has 5 residual blocks with 128 channels each; both original U-net used in Deep Image Prior paper and my residual U-net have 5 layers with 128 channels each. This setting is meant to keep numbers of parameters similar.

PSNR curve during training is shown in Fig 3. From the plot we can see that residual U-net makes steady progress towards peak performance and stays for more iterations before small decrease. U-net reaches overfitting more quickly and if we do not stop in the halfway, we may still get noisy image. Configurations and numbers are listed in table 1.

TABLE 1: Denoising Results (5000 iters)

Model	# Parameters	One-Pass PSNR	Averaged
original U-net	2,217,831	30.60	32.53
ResNet	1,665,411	27.31	29.90
res U-net (p=0)	1,706,331	30.72	32.68
res U-net (p=0.1)	1,706,331	31.91	32.97

5 Conclusion

ResNet training ($\approx 1.3s/iter$) is time-consuming compared to U-net and the newmodel ($\approx 0.28s/iter$). This is because with downsampling each block needs to convolve with 512×512 feature maps and convolution time is quadratic to downsampling rate. Also deep stack of convolutional layers erases high frequency component and outputs low PSNR blurry image. U-net outperforms ResNet largely due to the skip connections especially those in shallow layers. Skip connections preserve both edges and noise.

Residual U-net combines merits of the two models. One the one hand, denoising is close to identity mapping and therefore residual blocks are desired. They enhance convergence of the network. On the other hand, hourglass structure of U-net reduces training time per iteration by reducing the input size of middle layers.

Dropout layer benefits in two ways. First it significantly improves the one-pass PSNR ratio of the network. Evaluation mode of network can be viewed as an average of exponential number of trained networks. Second it slows down overfitting noise.

6 Future Work

One possible improvement is to apply the newly built network to other applications described in the original paper and make adjustments. Generative network is also a promising architecture class to be explored on this topic.

Acknowledgments

I would like to thank my project mentor Jeong Joon Park for providing useful feedback and references. Also thank EE367 teaching team for delivering such a wonderful course.

References

[1] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9446–9454.

[2] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] ——, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[5] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[7] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, 2021.

[8] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

Fig. 1: Three Blocks ResNet.

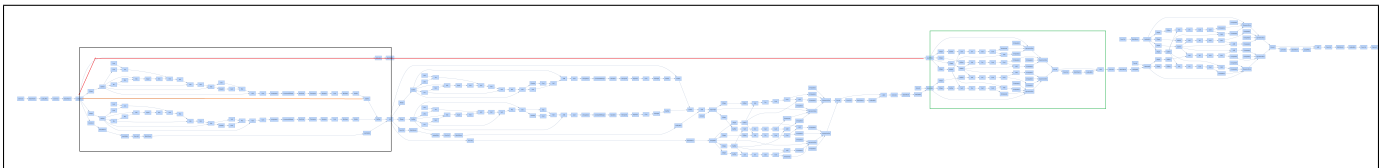
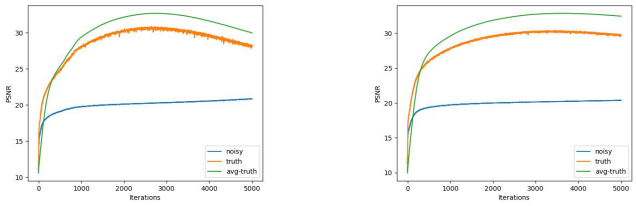


Fig. 2: Three Blocks ResNet.



(a) Original U-net (b) Residual U-net

Fig. 3: PSNR curve

[10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[11] X. Li, S. Chen, X. Hu, and J. Yang, "Understanding the disharmony between dropout and batch normalization by variance shift," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2682–2690.