Literature Survey: Deep Image Prior

Saravanabalagi Ramachandran
saravanabalagi@hotmail.com

Abstract

This report presents a short literature review of the paper Deep Image Prior, describing what ideas and methods were presented, what objectives were set and how these methods were used to achieve the objectives mentioned. It also discusses how results of this method compares to that of its existing counterparts. Further, advantages and downsides of the methodology and its usefulness and applicability in production environments are also discussed.

1. Introduction

Deep Image Prior introduces the idea of using randomly initialized untrained convolutional neural network that uses corrupted image as its training data and parameters of the network as prior to solve inverse problems like denoising, inpainting, super-resolution and flash no-flash reconstruction. Image statistics are captured by the way network is structured and not by learning from multiple images.

2. Method

2.1. Core Idea

It's not always the learning that makes the neural networks perform better than traditional methods. Structure of the network plays a more significant role in generalization than learning. Structure needs to be tweaked to *resonate* with the pattern found in data, so as to get the most out of it.

2.2. How it works

Randomly initialize a neural network and train it with just the corrupted image. As the network learns, it tries to produce the same image and we stop training so the image produced is sufficiently closer to the corrupted image but without noise (artifacts, grains, jagged edges, etc.,). When not stopped it would overfit eventually producing the exact corrupted image. With no learning required, by controlling the structure of the network like number of hidden layers, number of hidden units, and hyperparameters such as learning rate and stopping, this technique can be cleverly used to get rid of noise, artifacts, jagged edges, and even for inpainting smaller regions.

This can be closely compared to an autoencoder or the generator part of GANs, which share strikingly same specifics with the CNNs used in Deep Image Prior, as they also encode the image data like the other two, see Fig 1.

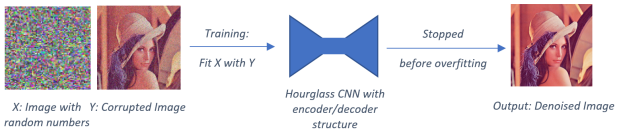

Figure 1. A very high level view of Deep Image Prior network

2.3. Why it works

Before capturing the minute details about the noise, the network is forced to change the whole bunch of random numbers into something that is as close to the corrupted image provided, which prevents it from learning the characteristics of noise at earlier stages. It is reluctant to pick up noise which contains high impedance in presence of more useful stats of the image that do not impede learning.

Learning improves and gets closer and closer to the corrupted image with every iteration until the network learns the image perfectly by overfitting it. Hence using the right learning rate and stopping at the right iteration should recover an image that is much closer to the corrupted image but is not close enough that it has all the noise. As there is no training from a large dataset with many images, therefore, the structure of the network plays a very significant role in what the network learns about the image before it gets to overfit.

2.4. Technical Explanation

Inverse problems like denoising, inpainting and super resolution can be represented as the energy minimization problem.

$$x^* = \min_x E(x; x_0) + R(x) \tag{1}$$

where x is the ground truth or original image and x_0 is the corrupted image, $E(x; x_0)$ is a task-dependent data term, and R(x) is the regularizer.

Deep neural networks learn a decoder function $x = f_\theta(z)$ which maps a random code vector z to an image x with its parameter θ. In Deep Image Prior, the regularizer term $R(x)$ is replaced with the implicit prior capture by the network. Since there is no pre-training, this prior is effectively handcrafted like in Total Variation (TV) norm. Hence, the data term minimizer and decoder function becomes

$$\theta^* = \arg\min_\theta E(f_\theta(z); x_0) \tag{2}$$

$$x^* = f_{\theta^*}(z) \tag{3}$$

where θ^* starts at random and is optimized using gradient descent in general. Keeping the iterations low is the key to not having θ capture noise due to overfitting.

3. Results and Comparison

3.1. Metrics

Peak Signal to Noise Ratio (PSNR) between the corrupted image and the ground truth was used as a standard metric to measure how good the image is. Higher the PSNR value, better is the image restoration technique used.

$$MSE = \frac{1}{m\,n\,c}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\sum_{k=0}^{c-1}[x(i,j) - x_0(i,j)]^2 \tag{4}$$

$$PSNR = 10 \cdot \log_{10}\frac{MAX_x^2}{MSE} \tag{5}$$

where both images x and x_0 have same resolution m x n with c channels and MAX_x refers to the maximum value a pixel can have in the image x.

3.2. Denoising

For denoising, data term corresponds to $E(x; x_0) = \|x - x_0\|^2$. For Blind Image denoising, where the pattern of noise is unknown, the results are compared on (Matteo Maggioni & Egiazarian, 2014) with 9 images. With Color Block Matching and Denoising (CBM3D) it catches up closely after averaging restored images from last iterations, and further averaging over two optimization runs. However, with Non-local means, it outmanoeuvres. The same method could also be modified to incorporate noise model information, if known.

3.3. Super Resolution

This inverse problem is solved by setting the data term in 1 to $E(x; x_0) = \|d(x) - x_0\|^2$, where $d(x)$ is the downsampling operation that scales x by factor t. With absolutely no training, it performs much better than bicubic upsampling, but just below learned SRResNet and LapSRNs models. Limited to using just the unscaled image, it could not, however, compete with GAN based trained SRGAN and EnhanceNet.

3.4. Inpainting

The data term for inpainting is given by $E(x; x_0) = \|(x - x_0) \odot m\|^2$, where $\odot$ is the Hadamard's product. It produced convincing results in text overlayed image, (diffused) missing pixels and (hole or) region inpainting, which outsmarted Convolutional Sparse Coding, Shepard Networks and performed on par with Globally and Locally Consistent Image Completion which is trained network. However, it cannot fill in larger holes or patches in faces as it would require the network to fill in data that can usually only be filled by training on similar images.

3.5. Natural Pre-Images

Deep Image Prior helps produce dramatically better inverted deep image representations than TV norm which produces noisy images, while still remaining unbiased towards a particular training set, unlike trained upconvolutional neural networks by (Dosovitskiy & Brox, 2015) which tend to bias towards learned data and regress towards the mean.

3.6. Flash No-Flash Reconstruction

This method can also be used to produce images with lighting similar to no-flash image but with reduced noise using details obtained from flash image. It is better at avoiding lighting pattern leaks than joint bilateral filtering technique.

4. Advantages and Downsides

4.1. Pre-Training

No training is required. Corrupted image is the only image the network would require to adjust its weights.

4.2. Variety of Applications

Deep Image Prior method has been tested to work well for Denoising images, Super Resolution, Inpainting, Natural Pre-Image, Flash No-Flash reconstruction and so on.

4.3. Processing Time

Time taken to process one image will be very high, as the network has to be run exclusively for reproducing each image, with both forwardprop and the expensive backprop, unlike supervised networks which will only do the forwardprop in production environments.

4.4. One ring to rule them all?

It was very well explained how the network should be structured, how depth, learning rate and stopping criteria impacts capturing of noise stats, but it begs the question of whether we can have *one* network for a variety of images to achieve results closer to what was shown by networks that were trained. Although it gives great results, handcrafting a network for each image type might not be feasible in production environments.

5. Conclusion

While we have multiple methods that work well when there are loads of training data, Deep Image Prior shows a way to build networks that can work with zero pre-training. Although it takes a few minutes even on GPUs to de-corrupt images, it outperforms existing methods that require no training, clearly emphasizing that the structure of the network captures image stats. It is not just a good trade-off when we got no dataset in hand to make a trained model but it also opens up the space for further research on unsupervised deep learning approaches.

References

Dosovitskiy, Alexey and Brox, Thomas. Inverting convolutional networks with convolutional networks. *CoRR*, abs/1506.02753, 2015. URL http://arxiv.org/abs/1506.02753.

Matteo Maggioni, Enrique SÃąnchez-Monge, Alessandro Foi Aram Danielyan Kostadin Dabov Vladimir Katkovnik and Egiazarian, Karen. CBM3D dataset with 9 images, 2014. URL http://www.cs.tut.fi/~foi/GCF-BM3D/index.html.