Adversarial Framework for General Image Inpainting

Wei Huang[1] and Hongliang Yu[1]

Tsinghua University, Beijing, China
huang-w15@mails.tsinghua.edu.cn, hlyu@mail.tsinghua.edu.cn

Abstract. We present a novel adversarial framework to solve the arbitrarily sized image random inpainting problem, where a pair of convolution generator and discriminator is trained jointly to fill the relatively large but random "holes". The generator is a symmetric encoder-decoder just like an hourglass but with added skip connections. The skip connections act like information shortcut to transfer some necessary details that discarded by the "bottleneck" layer. Our discriminator is trained to distinguish whether an image is natural or not and find out the hidden holes from a reconstructed image. A combination of a standard pixel-wise L2 loss and an adversarial loss is used to guided the generator to preserve the known part of the origin image and fills the missing part with plausible result. Our experiment is conducted on over 1.24M images with uniformly random 25% missing part. We found the generator is good at capturing structure context and performs well in arbitrary size images without complex texture.

Keywords: Inpainting · GAN · skip connections.

1 Introduction

Image inpainting, or "hole filling" problem, aims to reconstruct the missing or damaged part of an input image. Classical inpainting algorithms in computer vision mainly fall into three categories: information diffusion and examplar-based filling.

Partial differential equation (PDE) is the foundation of the diffusion algorithms [3] [2][14] [12]. They are also referred as variational methods. Through iterations, the information outside a hole is continuously propagated into the hole, while preserving the continuity of the isophote. They can tackle cracks and small holes well, but produce blur artifacts when faced with large and textured region.

The exemplar-based algorithm, on the other hand, is able to reconstruct large region and remove large unwanted objects by patch matching and filling. The straight forward exemplar approach applies a carefully designed prioritizing filling [5][4] or a coherence optimization strategy [15][1]. The limitations are also obvious. It has difficulty in handling curved structures. When proper similar patches do not exist, it will not produce reasonable result.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_50

Recently, convolution network and adversarial training are also introduced into inpainting problem. Compared with the classical algorithms above, the network based algorithms are born to understand high-level semantic context, which brings it capability to tackle harder problems, like semantics required prediction. The context-encoder[13], for example, consisting of an encoder and a decoder, can predict the large squared missing center of an image.

In this paper, we aim to solve a more challenging semantic inpainting problem, the arbitrarily sized image with random holes. We adopt the adversarial training to suppress the multi-modal problem and get sharper result. Instead of predicting a single likelihood to evaluate whether an image is fake like most of other GAN works do, the discriminator provides a pixel-wise evaluation by outputting a single channel image. If the generator does not work well, the discriminator is supposed to point out the origin missing region. The output is visually explainable, as one can clearly figure out the contour of the hole mask if the inpainted image is unsatisfactory and vice versa.

In our experiment, we found the generator is good at capturing structure context and performs well in arbitrary size images without complex texture. As for the failed cases, mainly the complex texture with tiny variations in the intensity, the generator will produce reasonable but blur result.

2 Method

2.1 Aderversarial framework

The adversarial framework in this paper is based on Deep Convolutional Generative Adversarial Networks (DCGAN). A generator G and a discriminator D are trained jointly to for two opposite goals. When the training is stopped, the G is supposed to reconstruct a damaged image in high quality.

Generator The generator G is an hourglass encoder-decoder consisting of basic convolution blocks (Conv/FullConv-BatchNorm-LeakyReLU/ ReLU), but with shortcut connections to propagate detail information in the encoder directly to the corresponding the symmetric layer of the decoder (Figure 1).

Endoder The encoder performs down sampling using $4*4$ convolution filters with stride of 2 and padding of 1. The encoder drops out what it considered useless for reconstruction and squeezes the image information liquid into a "concept". When it passes the bottleneck layer of the encoder, the feature map size is reduced to $1 * 1$. The number of the filters in this the bottle layer (m), decides the channel capacity ($m*1*1$) of the whole encoder-decoder pipeline. The activation function for the encoder is LeakyReLU with negative slope of 0.2.

Dedoder The structure of the decoder is completely symmetric toward the encoder except the output layer. There are 3 added shortcut connection directly

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_50

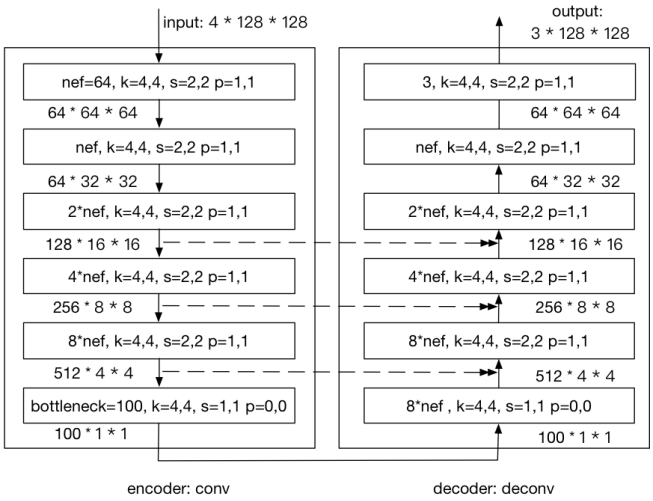

Fig. 1. Generator Architecture: an hourglass encoder-decoder with 3 added shortcut connections. The damaged image and the selected mask is passed through the encoder. The decoder then reconstructs the image without holes. k(kernel size), s(stride), p(padding) are parameters of spatial convolution / deconvolution layer.

join the 3 decoder layers closest to the bottleneck layer with their corresponding symmetric encoder layers. The final layer is supposed to output the desired reconstructed RGB image without holes. The activation function are Tanh for the final layer and ReLU for the rest layers.

Discriminator The discriminator D is a 5-layer stack of convolution blocks (Conv-BatchNorm-LeakyReLU). All the convolution layers have the same number of 3*3 filters with stride of 1 and padding of 1. The shape of the information fluid is the same as the shape of input throughout the network. The activation layers are Sigmoid for the final layer and LeakyReLU with negative slope of 0.2 for the rest layers.

As the final step of the inpainting is to fusion the output with the origin damaged image, we specify a high-level goal, to make the "hole" indistinguishable. Instead of predicting a single likelihood to evaluate whether an image is fake like most of other GAN works do, the discriminator here is trained to find the flaw of the output of G or the hidden holes.

The discriminator is supposed to output all ones when faced with natural images and output the given hole masks (ones/white for known regions and zeroes/black for the holes) otherwise. Compared to a single number, this single channel image output is visually explainable. And it can provide more targeted guidance for each input pixel and comprehensive judgment.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_50

2.2 Objective

The generator is trained to regress the ground truth content of the input image. It is well known that the L2 loss (Equation 1) - and L1, prefers a blurry solution to a clear texture[10]. It is an effective way to capture and rebuild the low frequency information.

$$L_2(G) = ||x - G(x')||_2^2 \tag{1}$$

The adversarial loss is introduced to get a sharper result. The objective of a GAN can be expressed as Equation 2. x' is the damaged input image and x is the corresponding ground truth. $\hat{M}$ is the hole mask. The holes are filled with zeroes, and the known regions are filled with ones. The all **1**s mask means no holes at all. The G is trained to minimize this objective, while the D is trained to maximize it.

$$L_{GAN}(G, D) = E_x||\hat{M} - D(G(x'))||_2^2 + E_{x'}||\mathbf{1} - D(x)||_2^2 \tag{2}$$

The total loss function is a weighted average of a reconstruction loss and an adversarial loss (Equation 3). We assign a quite large weight ($\lambda = 0.999$) upon the reconstruction loss.

$$L(G, D) = \lambda L_2(G) + (1 - \lambda)L_{GAN}(G, D) \tag{3}$$

2.3 Masks for training

As our framework is supposed to support damaged region with arbitrary shape or position, we need to train the networks with numerous random mask. For the efficiency, the inputs within a mini-batch share the same mask and all the masks are sampled from a global pattern pool.

The global is generated as follows: 1) create a fix-sized uniform random distribution (range from 0 to 1) matrix; 2) scale it to a given large size(10000 * 10000); 3) mark the region with value less than a threshold as "holes" (ones/white) and the rest as "known region" (zeroes/black). The *scaling ratio* and the *loss threshold* are two important hyper parameters for the global pattern pool. The scaling ratio controls the continuity of the holes. Larger scaling ratio generates scatter result.

3 Experiment

3.1 Training details

This work is implemented in Torch and trained using Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz with TITAN X (Pascal). We train the G and the D jointly for 500 epochs, using stochastic gradient solver, ADAM, for optimization. The learning rate is 0.0002.

The training dataset is 100 classes of ILSVRC2010 training dataset (over 1.24M natural images). The natural images are scaled down at the same proportion so that the maximum of the width and the height is no more than 350px.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_50

Then, 64 random crops of 128∗128 from different images consist of a mini-batch. These crops share the same mask. During the training process, we assume the hole area of mask should be in between 20% - 30%.

The G receive input with size of $128 * 128 * 4$, where the first three channels are the RGB data and the last channel is a binary mask. The ones in the binary mask indicate the holes while the zeros indicate the known region. The missing region of RGB data specified by the mask will be filled with a constant mean value (R:117, G:104, B:123). We also experimented the gray value (R:127, G:127, B:127) and found no significant difference between them in improving the performance of the generator. The G consists of 10 convolution layers and the bottleneck size is 4000.

3.2 Evaluation

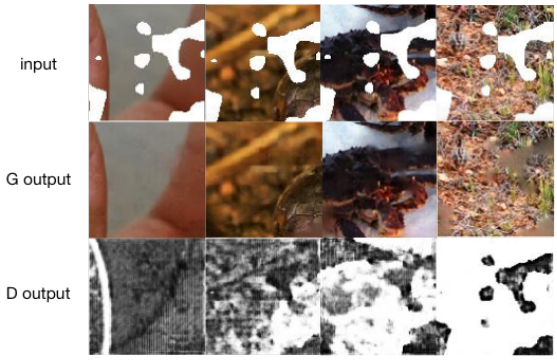

Fig. 2. The comparison of the G output and the D output. The output is visually explainable, because the blur or unnatural parts is darker than the normal parts.

In prior works in GAN, the D outputs a single probability indicating whether the input is a natural image. In this random inpainting problem, we find it requires elaborate design of weight assign among the hole regions and the known regions when updating the parameters of both D and G. What's more, the output of D may be not consistent with human intuition.

Our method, on the contrary, trains D to find the pixel-wise flaw of G output. It turns out that the output of D is visual explanatory and the optimization is easier. One can clearly figure out the contour of the hole mask if the inpainted image is unsatisfying and vice versa (Figure 2).

We evaluate our inpainting framework using images from the ILSVRC2010 validation dataset (the "barn spider" and black and "gold garden spider"). As the generator only receives 128*128 images, we split the input into a batch of 128*128 crops if the input image is too large. Afterwards, the result will be tiled to create the origin-sized image. We found the G generates plausible result when

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_50

faced with linear structure, curved lines and blur scenes. But it is difficult for G to handle complex texture with tiny variations in the intensity. The background regions in Figure 3 are inpainted so well that the G successfully fools the D, while the spider body regions full of low contrast details are handled poorly.

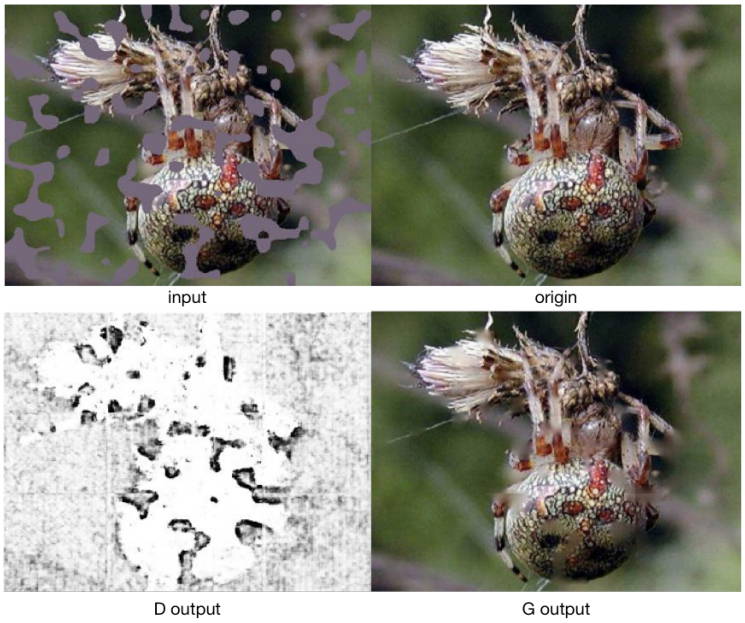

Fig. 3. Uniform random inpainting example from ILSVRC2010. The origin images are taken apart into $128 * 128$ crops and inpainted respectively.

4 Conclusions

In this paper, we aim to provide a unified solution for random image inpainting problems. Unlike the prior works, the output of D is visually explainable and the G is modified to adapt the general inpainting tasks. Trained in a completely unsupervised manner, without carefully designed strategy, the GAN networks learn basic common sense about natural images. The results suggest that our method is a promising approach for many inpainting tasks.

References

1. C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: a randomized correspondence algorithm for struc- tural image editing. *ACM Trans. Graphics, vol. 28, no. 3*, pages 24:1–24:11, 2009.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_50

2. M. Bertalmio, A.L. Bertozzi, and G. Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 1:I–355–I–362, 2001.
3. Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. *Proceedings of Conference on Computer Graphics and Interactive Techniques*, pages 417–424, 2000.
4. W. Cheng, C. Hsieh, and S. Lin. Robust algorithm for exemplar-based image inpainting. *IEEE Transactions on Image Processing, 13(9)*, pages 1200–1212, 2005.
5. A. Criminisi, P. Prez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing, 13(9)*, pages 1200–1212, 2004.
6. W. Dong, G. Shi, and X. Li. Nonlocal image restoration with bilateral variance estimation: A low-rank approach,. *IEEE Trans. Image Process., vol. 22, no. 2*, pages 700–711, 2013.
7. M. Elad, J. L. Starck, P. Querre, and D. L. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (mca). *Appl. Comput. Harmon. Anal., vol. 19, no. 3*, pages 340–358, 2005.
8. Q. Guo, S. Gao, X. Zhang, Y. Yin, and C. Zhang. Patch-based image inpainting via two-stage low rank approximation. *IEEE Transactions on Visualization and Computer Graphics, 2626(c)*, 2017.
9. L. He and Y. Wang. Iterative support detection-based split bregman method for wavelet frame-based image inpainting. *IEEE Trans. Image Process. vol. 23, no. 12*, pages 5470–5485, 2014.
10. A. B. L. Larsen, S. K. Snderby, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv1512.09300*, 2015.
11. W. Li, L. Zhao, Z. Lin, D. Xu, and D. Lu. Non-local image inpainting using low-rank matrix completion. *Computer Graphics Forum, vol. 34, no. 6*, pages 111–122, 2015.
12. Manuel M Oliveira, Brian Bowen, Richard McKenna, and Yu-Sung Chang. Fast Digital Image Inpainting. *International Conference on Visualization, Imaging and Image Processing*, pages 261–266, 2001.
13. D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.
14. Alexandru Telea. An Image Inpainting Technique Based on the Fast Marching Method. *Journal of Graphics Tools*, 9(1):23–34, 2004.
15. Y. Wexler, E. Shechtman, and M. Irani. Space-time completion of video,. *IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 3*, pages 463–476, 2007.
16. M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin. Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE Trans. Image Process., vol. 21, no. 1*, pages 5470–5485, 2012.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_50