Generative Image Inpainting with Contextual Attention

Yufeng Jiang

Abstract

Recent deep learning based approaches have shown promising results for the challenging task of inpainting large missing regions in an image. These methods can generate visually plausible image structures and textures, but often create distorted structures or blurry textures inconsistent with surrounding areas. This is mainly due to ineffectiveness of convolutional neural networks in explicitly borrowing or copying information from distant spatial locations. Based on this reasion, authors propose a new deep generative model-based approach which can not only synthesize novel image structures but also explicitly utilize surrounding image features as references during network training to make better predictions. The model is a fully convolutional neural network which can process images with multiple holes at arbitrary locations and with variable sizes during the test time.

1. Introduction

Filling missing pixels of an image, often referred as image inpainting or completion, is an important task in computer vision. It has many applications in photo editing, image-based rendering and computational photography [4]. The core challenge of image inpainting lies in synthesizing visually realistic and semantically plausible pixels for the missing regions that are coherent with existing ones.

Rapid progress in deep convolutional neural networks (CNN) and generative adversarial networks (GAN) [1] inspired recent works to formulate inpainting as a conditional image generation problem where high-level recognition and low-level pixel synthesis are formulated into a convolutional encoder-decoder network.

They present a unified feed-forward generative network with a novel contextual attention layer for image inpainting. Their proposed network consists of two stages. The first stage is a simple dilated convolutional network trained with reconstruction loss to rough out the missing contents. The contextual attention is integrated in the second stage.

2. Improved Generative Inpainting Network

Authors first construct their baseline generative image inpainting network by reproducing and making several improvements to the recent state-of-the-art inpainting model [3] which has shown promising visual results for inpainting images of faces, building facades and natural images.

Coarse-to-fine network architecture The network architecture of their imporved model is shown in Figure 1.They follow the same input and output configurations as in [3] for training and inference.

Global and local Wasserstein GANs Different from previous generative inpainting networks [3], they propose to use modigied version of WGAN-GP. They attach the WGAN-GP loss to both global and local outputs of the second-stage refinement network to enfore global and local consistency, inspired by [3].

Specifically, WGAN uses the $Earth - Mover$ distance distance $W(\mathbb{P}_r, \mathbb{P}_g)$ for comparing the generated and real data distributions. Its objective function is constructed by applying the $Kantorovich - Rubinstein$ duality:

$$\min_G \max_{D \in \mathcal{D}} E_{\mathbf{x} \sim \mathbb{P}_r}[D(\mathbf{x})] - E_{\tilde{\mathbf{x}} \sim \mathbb{P}_r}[D(\tilde{\mathbf{x}})] \tag{1}$$

where $\mathcal{D}$ is the set of 1-Lipschitz function and $\mathbb{P}_g$ is the model distribution implicitly defined by $\tilde{\mathbf{x}} = G(\mathbf{z})$. $\mathbf{z}$ is the input to the generator.

Gulraijani *et al.* [2] proposed an imporved version of WGAN with a gradient penalty term

$$\lambda E_{\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{\mathbf{x}}}}(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2 \tag{2}$$

where $\hat{\mathbf{x}}$ is sampled from the straight line between points sampled from distribution $\mathbb{P}_g$ and $\mathbb{P}_r$. The reason is that the gradient of D^* at all points $\hat{\mathbf{x}} = (1-t)\mathbf{x} + t\hat{\mathbf{x}}$ on the straight line should point directly towards current sample $\hat{\mathbf{x}}$, meaning $\nabla_{\hat{\mathbf{x}}} D^*(\hat{\mathbf{x}}) = \frac{\tilde{\mathbf{x}} - \hat{\mathbf{x}}}{\|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\|}$.

For image inpainting, they only try to predict hole regions, thus the gradient penalty should be applied only to pixels inside the holes. This can be implemented with multiplication of gradients and input mask $\mathbf{m}$ as follows:

$$\lambda E_{\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{\mathbf{x}}}}(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}}) \odot (\mathbf{1} - \mathbf{m})\|_2 - 1)^2, \tag{3}$$

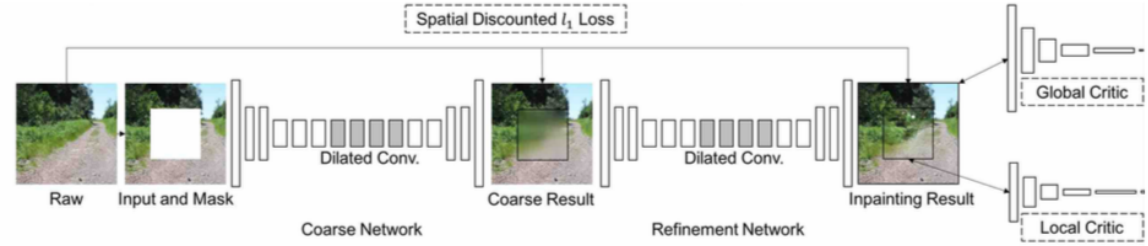

Figure 1. Overview of the improved generative inpainting framework. The coarse network is trained with reconstruction loss explicitly, while the refinement network is trained with reconstruction loss, global and local WGAN-GP adversarial loss.

where the mask value is 0 for missing pixels and 1 for elsewhere. λ is set to 10 in all experiments.

3. Spatially discounted reconstruction loss

Inpainting problems involve hallucination of pixels, so it could have many plausible solutions for any given context. In challenging cases, a plausible completed image can have patches or pixels that are very different from those in the original image. As they use the original image as the only ground truth to compute a reconstruction loss, strong enforcement of reconstruction loss in those pixels may mislead the training process of convolutional network.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1

[2] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017. 1

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[4] A. Levin, A. Zomet, S. Peleg, and Y. Weiss. Seamless image stitching in the gradient domain. In *ECCV*, 2004. 1