An Improved Image Inpainting Method Based on Feature Similarity Context Encoder

Zhiqiang Li¹, Zhouzhong Zhang¹ and Hongchen Guo² *

¹School of Computer Science, Beijing Institute of Technology, China
²Network Information Technology Centre, Beijing Institute of Technology, China.
Email: lizq@bit.edu.cn, 2220160673@bit.edu.cn, guohongchen@bit.edu.cn *

Abstract. With the development of deep learning techniques, image inpainting is a hot and popular field to use information around the residual images for repairing the missing area. Context encoder (CE) makes great progress in this field. However, it will cause the issue that the model tends to be overfitting when repairing the image. The existing method can’t address it very well. In this paper, we propose a feature similarity context encoder (FSCE), which integrates the feature extraction network with the encoder. FSCE uses feature loss to improve the performance of features and avoid overfitting problem. Various experiments between FSCE and CE are conducted on VOC 2007 and Paris Street View datasets. The results demonstrate that our FSCE model achieves superior performance in both the texture of output and dealing with the overfitting issue. FSCE model outperforms CE in the aspect of PSNR and SSIM.

1. Introduction

Image inpainting is a classic problem in the field of image processing and computer vision, which extracts information from residual images to repair missing parts. With the development of deep learning and the success of the convolution neural network in the image classification, the hot point of research has gradually changed from the traditional cell domain to the large area image filling. The task of image inpainting is still challenging due to the complexity of the image content in large area.

There are some kinds of traditional methods to repair the images. PDE models [1] use the edge information of the area such as the isphoto direction, and then propagates along the direction of the equal light line to the missing area. Texture synthesis techniques [2] use similar texture or texture patches around the missing area to restore images. However, the traditional methods have the problems that they have bad performance in the large missing area, and they are excessively dependent on parameter selection. So, the deep learning methods to inpaint the image are more popular recently. Convolution neural network (CNN) is used to restore images with dirt or rain [3]. Context encoder [4] makes great success in the image inpainting, which uses an encoder-decoder CNN to restructure the missing area, the model uses L2 loss and adversarial loss to improve the quality of the missing image regions. Although the context encoder has the desirable result, there remains a problem that the model is over fitting, the texture features of the input image is overlearned and make the result more complicated and less realistic.

In this paper, we propose a new context encoder with feature similarity comparison (FSCE). It is based on the context encoder and improves the feature extraction in the encoder. FSCE uses feature similarity comparison in the encoder, comparing between the features of input and the feature extracted from real image of missing region. The ability to transfer from the features of peripheral known image to the feature of the missing area is improved. By training two feature extracted
networks alternately, the limitation that the feature from the missing area should be extracted from the existed model can be avoided, and the ability of reconstruction can be improved.

To show the effectiveness of our method, the comparison experiment between context encoder and our method are carried out on the VOC2007 and Paris Street View respectively. The experimental results show that the improved network can capture the main features more accurately. Compared with the baseline, our model can prevent the overfitting issue effectively. The images generated by our method have more real subjective feelings. In addition, the experiments prove the effectiveness of the FSCE model over basic CE in different metrics.

2. Related Work

In recent years, many researches have contribution to the image inpainting. Autoencoder is applied to denoising the image and blind inpainting early due to its capability of image reconstruction. Denoising autoencoder[5] can repair the picture interfered by noise such as Gaussian noise or salt-and-pepper noise. Sparse denoising autoencoder can improve the performance of removing the text on images by extracting more accurate features [6]. While there is a limitation on using full connected layer to extract the feature in the image, CNN is combined with the image inpainting because CNN makes great progress in image feature extraction. Eigen et al.[7] Proposed a simple convolution neural network to learn how to map corrupted image patches to the clean patches, the model uses the L2 loss as the loss function. However, it has only two hidden layers that it can’t extract enough features, it can’t repair the large missing area in the image. Pathak et al.[4] presented an encoder-decoder network to generate the missing area, whole image can be input into the network and the missing area can be output, the similar network is also used to remove the wavy lines in the face verification[7], restoring the occluded face behind the occlusion[8] and so on.

Many researches have focus on the loss function of the CNN model to get more realistic prediction of the missing image. Huber loss is used to improve verification performance [9], it can be viewed as a variant of L2 loss. Adversarial loss is produced by a discriminative model [10]. By propagating the error of misjudgment from the discriminative model. The generative model can produce more realistic picture [11]. By combining L2 loss and adversarial loss [4] the prediction is more accurate and more semantic compared with only L2 loss. However, the Adversarial loss probably leads to overfitting of the model, there is a lot of redundant texture that reduces the authenticity of the content in the generative image, the lines in the image are complicated and twisted. These results in the decline of semantic information.

To get more realistic texture feature, Yang et al.[12] Proposed a texture network to produce a local texture constraint loss, which extracts some of the feature maps from the repaired images, find the most similar feature patches in the neighbour of missing area and use it to repair the texture in the missing area. Besides, perceptual loss is applied in the repaired images[13], which are computed by the feature maps extracted by CNN on output and real missing image. The perceptual loss can make the feature of missing area similar to the feature of real picture. These loss functions propagate the error back to the entire network so that output picture can be more realistic. However, the overfitting of the model still exists because the encoder part in the encoder-decoder network cannot extract better features, and the methods mentioned above cannot improve the encoder effectively.

In this paper, we propose the FSCE model that adds the feature loss in the encoder, reducing the overfitting caused by the adversarial loss and speeding up the generation of missing feature. The model can learn the more important semantic information rather than reduplicate texture.

3. Method

3.1. Framework Overview

To solve the problem that context encoder has model overfitting, we propose a new model called feature similarity context encoder (FSCE) for image inpainting. It is based on the context encoder and we add the similarity comparison in the encoder network. By comparing the feature similarity of the middle feature layer, the features extracted from the neighbourhood around the missing image are more accurately transferred to the missing image features, thus improving the accuracy of the features.
and obtaining a higher quality of final inpainting image. The FSCE model is illustrated in figure 1. It consists of a context encoder network and a feature extraction network.

![Figure 1. Architecture of the feature similarity context encoder.](image)

The context encoder contains two parts: encoder and decoder. The encoder network extracts the potential features by using the information from the residual images, in which the images with hole are input and the potential features of the existed images are output. The decoder network uses the extracted features to generate the missing image contents. The feature extracted from the encoder is compared with the feature from the real missing picture to avoid the over-fitting phenomenon. The feature extraction network can ignore the interference features. We describe the details of FSCE model in the following, including the context encoder network, the feature extraction network and training process of the FSCE model.

### 3.2. Context Encoder Network

The context encoder network in our model consists of convolution layers and deconvolution layers, the encoder uses five convolution layers to extract the feature from input image, the convolution kernel is size of 4*4 with stride size 2 and padding size 1 around the input feature maps, the number of convolution kernels on the next layer is two times that of the last layer and the first layer has 64 kernels, so that the weight and height of output feature map will be half of the original image and its number is twice as much as input.

The input image size limit is 128 * 128 in the network, input picture is incomplete picture with missing parts, which is located in centre of image input image and the size of missing area is 64 * 64, accounting for 1/4 size of input image. There are 512 output feature maps with size of 4*4, then they will be flattened to a 4000-dimensional vector by a full-connection layer, which is seen as the vector to generate the missing area. Batch normalization layer is added into each convolution layers and leaky ReLU is adopted as activation function. The decoder has a symmetric structure with the encoder, using deconvolution layers and ReLU instead. the output is the predicted missing region with size of 64*64. The encoder and decoder are seen as generative models, the discriminator is similar to the encoder in the convolution kernels, but it is used to distinguish whether the image is real or generated by model.
3.3. Loss Function

The loss of the whole network consists of three parts: feature similarity loss, L2 loss and adversarial loss. In the encoder network, we use a feature extraction network to get the feature vector of the real image in missing region. The encoder is denoted as E, let M be a binary mask for input image, the value of M in the missing region will be 1 and the other will be 0, so that the feature can be denoted as $E((1 - M) \odot x)$, the feature extraction network is denoted as $E_F$, the feature similarity loss $L_F(x)$ can be shown in the equation (1).

$$L_F(x) = ||E_F(x) - E((1 - M) \odot x)||_2^2 \quad (1)$$

In our model L2 loss is applied to the feature similarity comparison. It can be replaced by other methods of similarity comparison. The feature extraction network is similar to the discriminator, the difference between them is that the output of feature extraction network is the same dimension to the encoder so that we can get the error between them, helping the encoder to get more feature of the missing region image. It is also helpful to avoid the model overfitting because it comes from such an intuition that the feature extracted by the encoder tend to the real value and reduce the complicated texture generated in order to deceive the discriminator.

In the loss function of the encoder-decoder network, the L2 error between the true image of the missing part and the predicted image is adopted, which can roughly recover the outline of the missing picture. Given an image x, the generative network is denoted as F, so that the generated image can be denoted as $F((1 - M) \odot x)$, the L2 loss is shown in equation (2).

$$L_{rec}(x) = ||M \odot (x - F((1 - M) \odot x))||_2^2 \quad (2)$$

Where $\odot$ represents element-wise product operation. L2 loss can give objects a rough prediction, but it is difficult to capture for the details of high frequency images, so L2 loss usually gets obscure inpainting patterns and thus do not show accurate textures. In order to make up for the details of the repairing, the generative adversarial network is introduced. The encoder-decoder network is regarded as the generative model and the classification model that judge the true and false of between the real missing image and generative model’s predicted images is added as the discriminant network. The model is trained by adversarial loss. The discriminant network and the generative model improve each other so that the generative model can greatly improve the quality of the predicted picture details. Denoting the discriminator as D, the adversarial loss of the generative network is shown in equation (3).

$$L_{adv} = \max_D E_{x \in X} \left[ \log D(x) + \log \left(1 - D \left(F((1 - M) \odot x)\right)\right) \right] \quad (3)$$

In the formula, the context encoder network is required to be changed in the direction of maximizing the discriminator network error. By updating the generative model’s parameters, the discrimination network error rate is higher. The gradient descent is carried out to reduce the error of the generative network. Generated image cause misjudgement, which indirectly improve the authenticity of the generated image. The total loss consists of L2 loss and adversarial loss. Denoting the $\lambda_{rec}$ as the ratio of L2 loss to the total, the loss for context encoder network $L$ is shown in equation (4).

$$L = \lambda_{rec} L_{rec} + (1 - \lambda_{rec}) L_{adv} \quad (4)$$

3.4. Training Process

Because we use a feature extraction network in the encoder instead of existed texture extraction network, a training strategy is designed to combine feature extraction network with context encoder network. Mini-batch gradient descent is used to update the model parameters. In each batch of each iteration, firstly, the discriminator is updated by minimize the cross entropy loss. Secondly, compute the L2 loss and adversarial loss when the parameters of discriminator remain the same. Then combine the feature extraction network and decoder to be trained as context encoder network, while only updating the parameters in feature extraction network. After training the real feature will be obtained. The finally training step is updating the encoder with the feature similarity loss.
4. Experiments and Results

4.1. Datasets and Evaluation Metrics
The models are trained and evaluated on PASCAL VOC2007 and Paris Street View respectively. PASCAL VOC2007 is the standard dataset, which contains 9963 images and 20 categories. Paris Street View is a real dataset from google street view, which is a collection of 15000 pictures. It has more complicated street scenes, and the content in the image are more abundant than PASCAL VOC2007.

Three metrics are adopted in our comparison: Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM). When the value of MSE is smaller and PSNR is higher, it means the quality of restoration in pixel level is better. While the SSIM value tend to 1, it means the image is more similar to the real image in the aspects of brightness, contrast and structure of image.

4.2. Experiment Details and Baseline
The FSCE model is compared with two other image inpainting model as baseline, one is the context encoder trained with only L2 loss, the other is the context encoder with L2 loss and adversarial loss, which is the state-of-the-art in the image inpainting using deep learning framework. The same parameters are used in the model to show the effectiveness of feature similarity loss. The image is scaled to 128*128 to fit the input dimension. And the missing area is a 64*64 region in the centre of the input, the model is initialized by the pre-training model for ImageNet to accelerate convergence. ADAM is used to update the parameters, where the beta1 is set to 0.5, learning rate is set to 0.002 and others is default. In the training process, the batch size is 64, $\lambda_{rec}$ is set to 0.99 and model has 100 iterations. 1000 images not used in the training is used to evaluate the model.

4.3. Result Analysis
The result is shown in the figure 2. We can see our model avoid the overfitting well. The unnecessary texture in first and last row of picture is removed in our FSCE model. and in second row of the picture, the line of the building restored by FSCE model is more vertical and border of road is more explicit than the output by CE. It means that the feature similarity loss makes model extract more accurate feature.
Figure 2. Three method comparison, (a) input, (b) CE with only L2 loss, (c) CE with L2+adv loss and (d) FSCE respectively.

The quantitative comparison is shown in Table 1. We find that the FSCE model has the better performance in the different metrics in VOC dataset, which is 0.2dB more than context encoder with adversarial loss. It is normal that all the method drop out in three metrics when using real dataset, while the PSNR of our method still has a little higher than context encoder. In the SSIM index, there are not obvious difference between our FSCE and CE.

Table 1. Quantitative comparison in VOC2007 and Paris Street View datasets with MSE, PSNR and SSIM.

| Dataset                | Metric    | CE+L2 | CE+L2+adv | FSCE  |
|------------------------|-----------|-------|-----------|-------|
| PASCAL VOC2007         | MSE (%)   | 1.54  | 1.29      | 1.23  |
|                        | PSNR (dB) | 18.11 | 18.87     | 19.07 |
|                        | SSIM      | 0.760 | 0.797     | 0.799 |
| Paris Street View      | MSE (%)   | 1.57  | 1.32      | 1.28  |
|                        | PSNR (dB) | 18.02 | 18.79     | 18.89 |
|                        | SSIM      | 0.699 | 0.733     | 0.733 |

5. Conclusions
We propose a feature similarity context encoder to avoid the overfitting phenomenon in the image inpainting task. The feature loss can make encoder learn more accurate feature and improve the quantity of output. Alternating training strategy effectively updates the generating network and feature extraction network. The experiments between the real and standard datasets illustrate that our FSCE model are competitive with CE model and its capability of avoiding overfitting is better than CE model.
6. References

[1] Bertalmio M, Sapiro G, Caselles V and Ballester C 2000 Image inpainting In Proc. of the 27th annual Conf. on Computer graphics and interactive techniques pp 417 - 424.

[2] Criminisi A, Pérez P and Toyama K 2004 Region filling and object removal by exemplar-based image inpainting IEEE Transactions on image processing 13 (9) 1200 -1212.

[3] Eigen D, Krishnan D and Fergus R 2013 Restoring an image taken through a window covered with dirt or rain In Computer Vision (ICCV) 2013 IEEE International Conference on pp 633-40.

[4] Pathak D, Krahenbuhl P, Donahue J, Darrell T and Efros A A 2016 Context encoders: Feature learning by inpainting In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 2536-44.

[5] Vincent P, Larochelle H, Bengio Y and Manzagol P A 2008 Extracting and composing robust features with denoising autoencoders In Proceedings of the 25th international conference on Machine learning pp 1096-103.

[6] Xie J, Xu L and Chen E 2012 Image denoising and inpainting with deep neural networks In Advances in neural information processing systems pp 341 - 9.

[7] Zhang S, He R, Sun Z and Tan T 2016 Multi-task ConvNet for blind face inpainting with application to face verification In Biometrics (ICB) 2016 International Conference on pp 1-8.

[8] Zhang J, Kan M, Shan S and Chen X 2016 Occlusion-free face alignment: deep regression networks coupled with de-corrupt autoencoders In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 3428-37.

[9] Zhang S, He R, Sun Z and Tan T 2018 Demeshnet: Blind face inpainting for deep meshface verification IEEE Transactions on Information Forensics and Security 13 (3) 637- 47.

[10] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S and Bengio Y 2014 Generative adversarial nets In Advances in neural information processing systems pp 2672-80

[11] Yeh R A, Chen C, Lim T Y, Schwing A G, Hasegawa-Johnson M and Do M N 2017 Semantic image inpainting with deep generative models In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 5485-93.

[12] Yang C, Lu X, Lin Z, Shechtman E, Wang O and Li H 2017 High-resolution image inpainting using multi-scale neural patch synthesis In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[13] Zhao G, Liu J, Jiang J and Wang W 2017 A deep cascade of neural networks for image in painting deblurring and denoising Multimedia Tools and Applications 1- 16.