Schedae Informaticae Vol. 27 (2018): 81–91
doi: 10.4467/20838476SI.18.007.10412

Image Inpainting with Gradient Attention

Michał Sadowski[1], Aleksandra Grzegorczyk[2]
[1]Faculty of Mathematics and Computer Science
Jagiellonian University, Kraków, Poland
[2]Samsung R&D Institute Poland
e-mail: m.sadowski@student.uj.edu.pl, a.grzegorczy@samsung.com

Abstract. We present a novel modification of context encoder loss function, which results in more accurate and plausible inpainting. For this purpose, we introduce gradient attention loss component of loss function, to suppress the common problem of inconsistency in shapes and edges between the inpainted region and its context. To this end, the mean absolute error is computed not only for the input and output images, but also for their derivatives. Therefore, model concentrates on areas with larger gradient, which are crucial for accurate reconstruction. The positive effects on inpainting results are observed both for fully-connected and fully-convolutional models tested on MNIST and CelebA datasets.

1. Introduction

Image inpainting is the process of filling missing or corrupted regions in images based on surrounding image information so that the result looks visually plausible. It is widely used to rebuild damaged photographs, remove unwanted objects and texts, or replace them.

Recently, deep learning techniques have been applied successfully to the problem of inpainting by Pathak et al. [18]. They introduced context encoder (CE, a convolutional neural network trained to generate the contents of an arbitrary image region, conditioned on its surroundings), which is able to fill-in missing regions in natural images. Since its publication, various modifications of this method have been proposed [8, 12, 21, 22, 23].

The overall architecture of context encoder is a simple encoder-decoder pipeline. The encoder takes an input image with missing regions and produces a latent feature representation of that image. The decoder takes this feature representation and produces the missing image content. Context encoders is trained by regressing to the ground-truth content of the missing region. The reconstruction loss is responsible for capturing the overall structure of the missing region, while adversarial loss tries to make prediction look real.

In this paper, we propose a feasible modification of context encoder reconstruction loss function, which focuses model attention on objects' edges, what in consequence results in more plausible inpainting. We provide its theoretical background and experimentally prove its relevance in inpainting task.

The paper is structured as follows. Section 2. reviews related approaches. In Section 3. we introduce gradient attention component in reconstruction loss. Sections 4. presents the experimental setup and results. We conclude the paper in Section 5..

2. Related Work

In this section, we first present state-of-the-art for image inpainting, and then we analyze previous approaches to autoencoder loss modification.

Image inpainting. Existing methods for inpainting problem can be divided into several categories such as structural inpainting [2, 13], textures synthesis [1, 2], and example-based methods [3, 4]. Structural inpainting uses geometric approaches to fill-in the missing information in the region. Textures synthesis inpainting algorithms use similar textures approaches, under the constraint that image texture should be consistent. Example-based image inpainting attempts to infer the missing region through retrieving similar patches or through learning-based model. The classical inpainting method can produce plausible output, however, they cannot handle hole-filling task, since the missing region is too large for local non-semantic methods to work well.

Over the recent years, convolutional neural networks have significantly advanced the image classification performance [10]. Motivated by the generative power of deep neural network, Pathak et al. [18] used it as the backbone of their hole-filling approach. They introduced context encoder (CE), which is a type of conditional generative adversarial net, GAN [16]. The overall architecture is a simple encoder-decoder pipeline, which is trained based on reconstruction and adversarial loss (the latter obtained from discriminator). This approach inspired many other researchers, and in result various modifications have been proposed. Yang et al. [21] proposed modification for high-resolution image inpainting, which uses two loss functions, the holistic content loss (conditioned on the output of the pre-trained content network) and the local texture loss (derived by pre-trained texture network). They also propose to use cascade coarse-to-fine strategy. Li et al. [12] introduced additional local discriminator to distinguish the synthesize contents in the missing region (in contrast to global discriminator, which analyzes whole generated image). Moreover, they use the parsing network (pre-trained model which remains fixed) to ensure more photo-realistic images. Zhao et al. [23] proposed a cascade neural network, consisting two parts, where the result of inpainting GAN is further processed by deblurring-denoising network in order to remove the blur and noise. In one of the newest approaches, Iizuka et al. [8] proposed a fully-convolutional network with dilated convolutions, modified training procedure, as well as global and local discriminators. This architecture was then improved by Yu et al. [22] by introducing the first part of the network dedicated for coarse approximation, and contextual attention, which borrows texture from the background.

Modification of Loss Function. Reconstruction loss in case of autoencoders architectures is usually computed as mean square error (MSE) or mean absolute error (MAE). However, it can result in blurry output images, therefore additional components in loss function were proposed to preserve sharp edges and details in reconstruction.

One possibility is to use generative adversarial networks, GAN [5]. They were applied by Pathak et al. [19] to image inpainting task, where reconstruction loss was combined with adversarial loss. The similar combination was used in follow-up papers (already described above). Alternatively, loss function can be extended to include perceptual loss, applied by Johanson et al. [9] to style-transfer and super resolution tasks. It utilize high level features extracted from pretrained network, which should be similar for real and fake images. It was already applied to inpainting task by Liu at al. [14] with very good results for irregular holes. Recently, Guo et al. [7] introduced gradient sensitive loss for image super-resolution. They used image gradient magnitude to create mask, separating low- and high-frequency areas of the image. They proved that gradient can give additional information when combined with mean absolute error, reconstructing high frequency content. Alternative approach was presented by Nguyen et al. [17], who computed loss function based on mean absolute error of 2D Fourier transforms of images. It was justified by the specific type of analyzed images and resulted in improved recovery of high frequency information.

3. Image Inpainting with Gradient Attention

It is very common to use information about image gradients ∇I in various methods of processing image I. This concept derives from the fact that the human eye is much more sensitive to gradient than to overall intensity of image. For instance, Pérez et al. [20] introduced method of combining region from the source image in the target image with maximum preservation of the source's gradient. It leads to output image with a realistic appearance thanks to solving Poisson's equation with Dirichlet boundary conditions, which aims to find a new target image that can produce the gradient from the source image. Encouraged by those results, we propose gradient attention component to suppress the common problem of inconsistency in shapes and edges between the inpainted region and its context.

Let us assume that standard mean absolute error (MAE) loss function is defined as follows:

$$l_1(I,R) = \left\| I - R \right\|_1 ,$$

where I and R are the input and output of the network, respectively. Then, gradient attention component is described as:

$$grad\_l_1(I,R) = \left\| \nabla I - \nabla R \right\|_1 .$$

For functions with discrete, two-dimensional domain (such as the channels of RBG image), a gradient in (i,j) pixel can be calculated as:

$$\nabla I_{i,j} = (I_{i,j} - I_{i-1,j}, I_{i,j} - I_{i,j-1}),$$

where $I_{i,j}$ is value of image in pixel (i,j). Then, reconstruction loss function is defined as:

$$\mathcal{L}_{rec}(I,R,\alpha) = l_1(I,R) + \alpha \cdot grad\_l_1(I,R),$$

where α is the constant weight of gradient attention component.

In order to visualize difference between l_1 and $grad\_l_1$ components, let us consider Fig. 1, generated for training process of CelebA inpainting model. One can observe that $|\nabla I - \nabla R|$ image (the second row), which is summed-up to obtain $grad\_l_1$, concentrates on eyebrows and tip of the nose, which are crucial in face perception. On the other hand, $|I - R|$ image (the first row) is distracted by less important characteristics, like cheeks and hair.

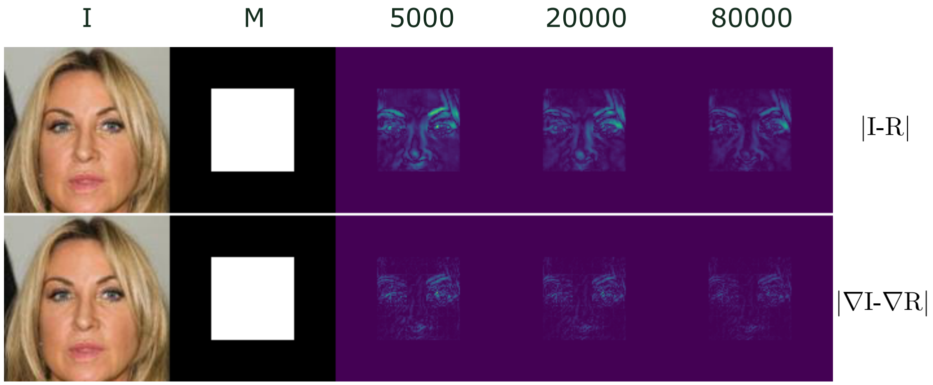

Figure 1. The impact of $grad\_l_1$ component and adversarial loss. The first and the second columns correspond to original image and mask, respectively. The remaining columns correspond to $|I - R|$ and $|\nabla I - \nabla R|$ (the first and the second rows, respectively) in successive iterations of training.

Let M be binary mask with 1s in missing region (with several pixels margin) and 0 in the remaining area. Then, reconstruction loss for image inpainting task is defined as:

$$\mathcal{L}_{inp}(I,R,M,\alpha) = M \odot \mathcal{L}_{rec}(I,R,\alpha) + (1-M) \odot l_1(I,R),$$

where $\odot$ is the element-wise product operation. Moreover, generative adversarial loss for image inpainting is defined as:

$$\mathcal{L}_{inp}^{adv} = \mathcal{L}_{inp} + \beta \mathcal{L}_{adv},$$

where $\mathcal{L}_{adv}$ is generator loss based on simultaneously trained discriminator, used in GAN training [5], and parameter β is its constant weight.

4. Experiments

In order to validate the effect of the proposed approach on inpainting results, we trained three models, without and with gradient attention component. Those three models are summarized in Table 1. In the first experiment, we evaluate the proposed loss function $\mathcal{L}_{inp}$ on simple fully-connected model, trained for inpainting task on MNIST dataset [11]. Images size is 28x28 and mask is 14x14 located in center of image. We set minibatch size as 128 and choose Adam optimizer with parameters learning rate 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We train model for 14000 iterations. In the second experiment, we apply loss function $\mathcal{L}_{inp}$ to more complicated fully-convolutional neural network, which was successfully used in recent state-of-the-art inpainting methods [8, 22], to inpainting task on CelebA dataset [15]. Instead of using only the standard convolutional layers, some of them are replaced with dilated convolution layers, which increase the size of the receptive fields. In this experiment, no generative adversarial loss was used. Finally, in the third experiment, we test $\mathcal{L}_{inp}^{adv}$ on the same fully convolutional network, however this time with adversarial loss WGAN-GP [6] which is broadly used in image inpainting models. The weight of the adversarial component was set to 0.001, which is common value in the case of inpainting task. We tested different non-zero values of α in order to fit the accurate proportions between l_1 and $grad\_l_1$ components. Model was trained for 120000 iterations. Both, in the second and the third experiments, we use 128×128 centrally cropped images. Mask is located in the center with size of 64×64. Batch size is set to 16 with Adam optimizer of learning rate 0.0001, $\beta_1 = 0.5$, and $\beta_2 = 0.9$.

The results of the first experiment trained on MNIST dataset are shown in Fig. 2. One can observe, that both, l_1 and $grad\_l_1$ losses decrease if the latter loss is used in training. This trend is confirmed by qualitative results, as shapes in the area of inpainting have sharper edges when gradient attention component is taken into consideration.

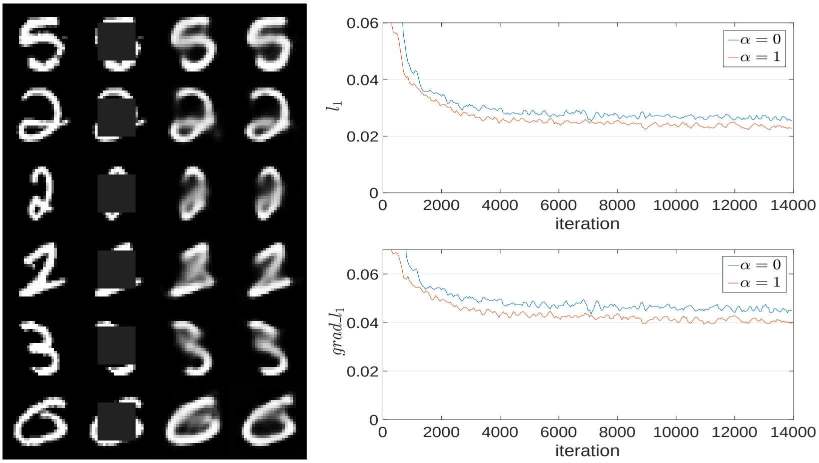

Figure 2. Results for MNIST inpainting model. On the left, successive columns correspond to: original image, mask, and output of model trained with $\alpha = 0$ and $\alpha = 1$. Plots on the right show l_1 and $grad\_l_1$ (upper and lower plots, respectively) across training iterations for $\alpha = 0$ (blue curve) and $\alpha = 1$ (red curve).

Experiment name	MNIST inpainting	CelebA inpainting	CelebA adversarial inpainting
input size	28x28	128x128	128x128
mask size	14x14	64x64	64x64
mask position	center	center	center
autoencoder layers	fc-256 fc-128 fc-64 fc-32 fc-64 fc-128 fc-256	conv3-32 (1,1) conv3-64 (2,1) conv3-64 (1,1) conv3-128 (2,1) conv3-128 (1,1) conv3-128 (1,1) conv3-128 (1,2) conv3-128 (1,4) conv3-128 s (1,8) conv3-128 (1,16) conv3-128 (1,1) conv3-128 (1,1) deconv3-64 (2,1) conv3-64 (1,1) deconv3-32 (2,1) conv3-32 (1,1)	conv3-32 (1,1) conv3-64 (2,1) conv3-64 (1,1) conv3-128 (2,1) conv3-128 (1,1) conv3-128 (1,1) conv3-128 (1,2) conv3-128 (1,4) conv3-128 s (1,8) conv3-128 (1,16) conv3-128 (1,1) conv3-128 (1,1) deconv3-64 (2,1) conv3-64 (1,1) deconv3-32 (2,1) conv3-32 (1,1)
discriminator layers	-	-	conv3-64 (2,1) conv3-128 (2,1) conv3-256 (2,1) conv3-256 (2,1) fc-1
database name	MNIST	CelebA	CelebA
train set size	60000	202099	202099
test set size	10000	500	500

Table 1. The summary of three models used in our experiments. “fc-N” corresponds to fully-connected layer with N neurons, while “convK-C(S,D)” refers to convolutional layer with kernel size K, C channels, stride S and dilation D.

We present the results for CelebA inpainting without and with adversarial loss in Fig. 3 and 4, respectively. Results of those experiments demonstrate that $grad\_l_1$ focuses model attention on objects' edges, resulting in more plausible inpainting, whether adversarial loss component is used or not. This trend is also confirmed by quantitative results, as $grad\_l_1$ loss in Fig. 3 and 4 significantly decreases when training with positive α. The α parameter itself is not crucial, however, one can observe slight differences in the areas with high gradient for models trained with $\alpha = 0.3$ and $\alpha = 1$.

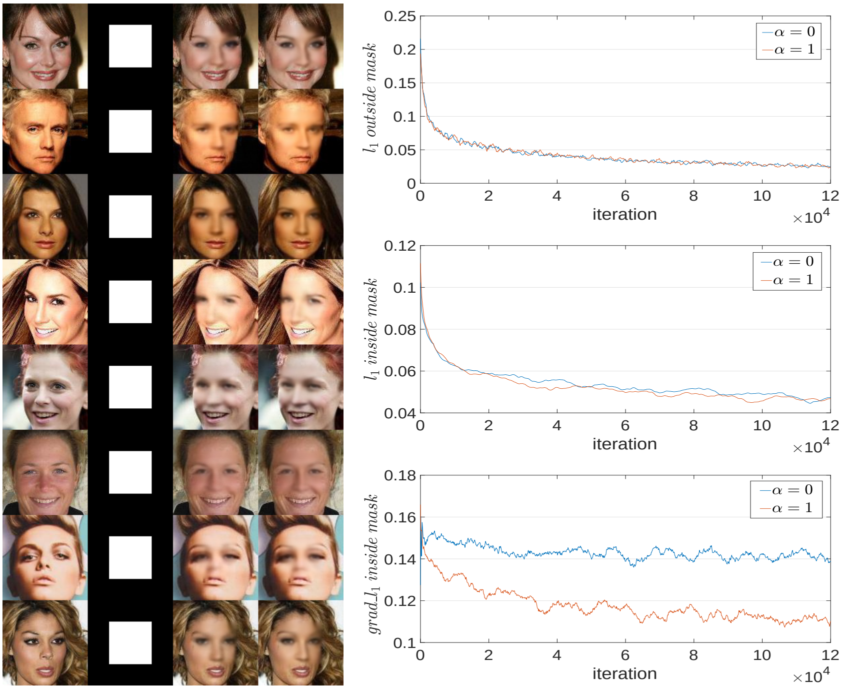

Figure 3. Results for fully-convolutional model trained on CelebA without adversarial component. In the picture on the left successive columns correspond to: original image, mask, output of model trained with $\alpha = 0$ and $\alpha = 1$. Plots on the right show l_1 *outside mask*, l_1 *inside mask* and $grad\_l_1$ *inside mask* (upper, middle and lower plots, respectively) across training iterations for $\alpha = 0$ (blue curve) and $\alpha = 1$ (red curve).

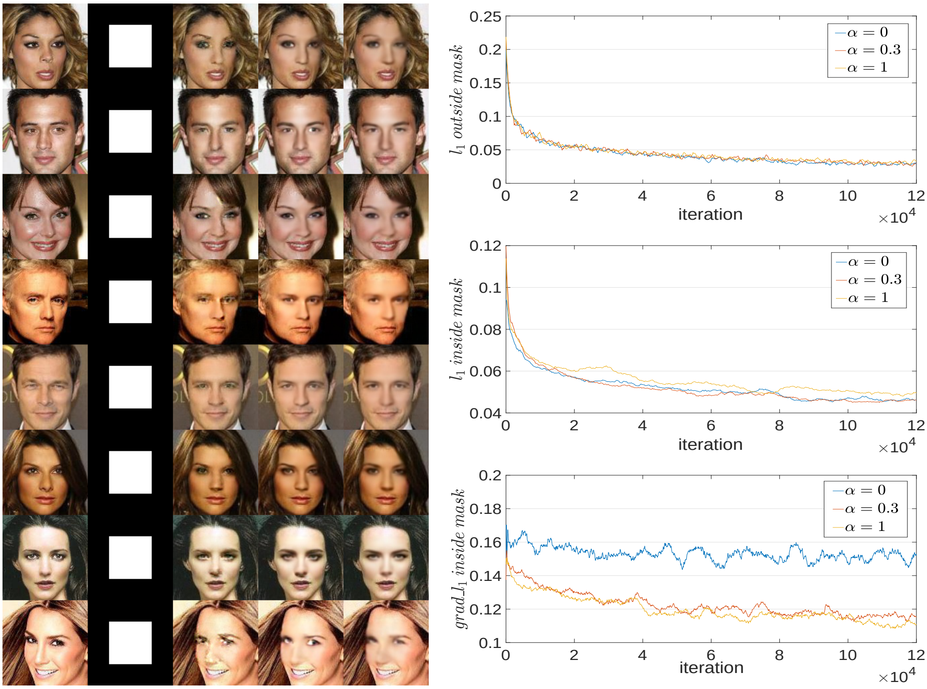

Figure 4. Results for fully-convolutional model trained on CelebA with adversarial component. In the picture on the left successive columns correspond to: original image, mask, output of model trained with $\alpha = 0$, $\alpha = 0.3$, and $\alpha = 1$. Plots on the right show l_1 *outside mask*, l_1 *inside mask* and *grad_*l_1 *inside mask* (upper, middle and lower plots, respectively) across training iterations for $\alpha = 0$ (blue curve), $\alpha = 0.3$ (red curve), and $\alpha = 1$ (orange curve).

5. Conclusion

We proposed a novel modification of context encoder loss function, which involves additional usage of gradient attention component besides mean absolute error in autoencoder loss function. The experiments, conducted on MNIST and CelebA datasets, demonstrate increased attention on areas with higher gradient, such as edges of the objects. They also confirm the positive effect, whether the adversarial loss component is used or not. In the future, we plan to apply this strategy to other tasks using autoencoder architecture. Moreover, we intend to use it other methods to indicate the crucial parts of the image.

6. References

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pages 417–424. ACM Press/Addison-Wesley Publishing Co., 2000.

[2] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher. Simultaneous structure and texture image inpainting. IEEE transactions on image processing, 12(8):882–889, 2003.

[3] A. Criminisi, P. Perez, and K. Toyama. Object removal by exemplar-based inpainting. In Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, volume 2, pages II–II. IEEE, 2003.

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

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

[6] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans, 2017.

[7] Y. Guo, Q. Chen, J. Chen, J. Huang, Y. Xu, J. Cao, P. Zhao, and M. Tan. Dual reconstruction nets for image super-resolution with gradient sensitive loss, 2018.

[8] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. ACM Transactions on Graphics (TOG), 36(4):107, 2017.

[9] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In ECCV, 2016.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[11] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010.

[12] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, volume 1, page 6, 2017.

[13] D. Liu, X. Sun, F. Wu, S. Li, and Y.-Q. Zhang. Image compression with edge-based inpainting. IEEE Transactions on Circuits and Systems for Video Technology, 17(10):1273–1287, 2007.

[14] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In ECCV, 2018.

[15] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), 2015.

[16] M. Mirza and S. Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[17] T. Nguyen, Y. Xue, Y. Li, L. Tian, and G. Nehmetallah. Deep learning approach to fourier ptychographic microscopy, 2018.

[18] 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.

[19] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2536–2544, 2016.

[20] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. ACM Trans. Graph., 22(3):313–318, July 2003.

[21] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 1, page 3, 2017.

[22] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. arXiv preprint, 2018.

[23] G. Zhao, J. Liu, J. Jiang, and W. Wang. A deep cascade of neural networks for image inpainting, deblurring and denoising. Multimedia Tools and Applications, pages 1–16, 2017.