Image Inpainting using GANs with Partial Convolutions

Nico Jersch, Justin Lundgren, Wilhelm Bolin, {njersch, julu1, wbolin}@Stanford.edu

Image Inpainting

- Image inpainting is the task of accurately filling in a removed part of an image with suitable imagery that blends in with the rest of the image
- Our model combines a **GAN with partial convolutions**, which is a convolution method that incrementally updates both the image and the mask
- This allows the network to learn the broader context of the image and accurately inpaint missing parts and detailed features.

Dataset

Our model is trained on a subsample of the dataset Places 365. It consists of a large set of images displaying cities, buildings, parks among other things. Of the over 8 million total images, we use a subset of 30,000 images for training and 1,000 for testing.

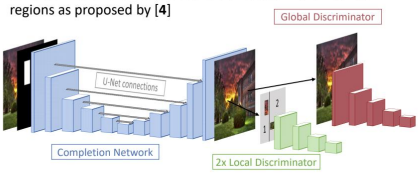
Pre-processing

- The images are scaled down to 128x128x3 (RGB)
- The masked regions are replaced by the mean value
- All pixel values are normalized by division of 255

Method & Model

Key features of our architecture:

- **Encoder-decoder structure:** The image is embedded into a lower dimensional representation (encoding) and then reconstructed into the high-dimensional output space (decoding) [**1**]
- **U-Net architecture** with skip connections from encoding to decoding layers so not all the information has to travel through the encoder-decoder bottleneck [**2**]
- **Local and global discriminators** to improve local consistency around the inpainted regions as suggested by Iizuka [**3**]
- **Partial convolutions** that take into account the mask to avoid conditioning on arbitrary placeholder values in the masked regions as proposed by [**4**]



Layer	K	S	F	Activation
PConv	5	2	64	ReLu
PConv	5	2	128	ReLu
PConv	3	2	256	ReLu
PConv	3	2	256	ReLu
PConv	3	2	256	ReLu
PConv	3	1	256	Leaky ReLu
PConv	3	1	256	Leaky ReLu
PConv	3	1	128	Leaky ReLu
PConv	3	1	64	Leaky ReLu
PConv	3	1	3	Leaky ReLu
PConv	1	1	3	Sigmoid

Table 1: Generator structure, $K :=$ kernel size, $S :=$ stride, $F :=$ amount of filters.

Layer	K	S	F	Activation
Conv	5	2	32	ReLu
Conv	5	2	64	ReLu
Conv	5	2	64	ReLu
Conv	5	2	128	ReLu
Conv*	5	2	128	ReLu
FC	-	-	-	ReLu

* 5th layer only exists for global.

Table 2: Global and local discriminator structure, K, S and F are defined as in table 1.

Loss Functions

- $\mathcal{L}^{+}_{MSE} = \frac{1}{n_I} \|(\mathbf{I}_{out} - \mathbf{I}_{in}) \odot (\mathbf{1} - \mathbf{M})\|^2$
- $\mathcal{L}^{-}_{MSE} = \frac{1}{n_I} \|(\mathbf{I}_{out} - \mathbf{I}_{in}) \odot \mathbf{M}\|^2$
- $\mathcal{L}_{MSE} = \mathcal{L}^{+}_{MSE} + \frac{1}{6}\mathcal{L}^{-}_{MSE}$
- $\mathcal{L}_D = -\log D(\mathbf{I}_{out}) - \log(1 - D(G(\mathbf{I}_{in})))$
- $\mathcal{L}_G = \mathcal{L}_{MSE} - \frac{1}{2000} \log D(G(\mathbf{I}_{in}))$

where $\mathbf{I}_{out}$ is the ground truth, **M** the mask and $\mathbf{I}_{in}$ the masked input image. Different losses are used in different phases of training. $\mathbf{n}_I$ is the number of pixels in the input image.

Partial convolutional layers comprise two steps:
(1) A masked and renormalized convolution step where the convolution p_c is

$$p_c = \begin{cases} \mathbf{W}^\top(\mathbf{X} \odot \mathbf{M})c + \mathbf{b} & \text{if } \mathbf{1}^\top\mathbf{M}\mathbf{1} > 0, \\ 0 & \text{else} \end{cases}$$

where **X** is the sliding window and **M** the corresponding mask from the previous layer. **c** is a rescaling factor and **W** and **b** the filter weights.
(2) A mask-update step where the mask value for the given convolution is expressed as:

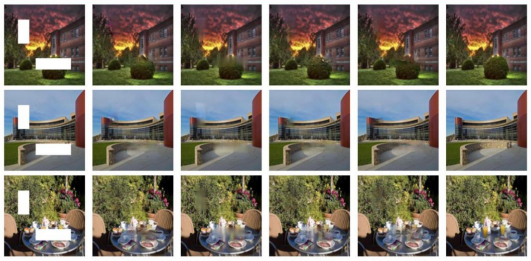
$$m = \begin{cases} 1 & \text{if } \mathbf{1}^\top\mathbf{M}\mathbf{1} > 0, \\ 0 & \text{else} \end{cases}$$

Training in three phases

Phase 1: use only weighted MSE loss (L_{MSE}) for 25,000 iterations
Phase 2: train only discriminators on L_D for 6,000 iterations
Phase 3: train both generator and discriminators for 94,000 iterations

200 2000 10000 50000

Results



Input (a) Full model (b) MSE only (c) No LD (d) No PC Ground truth

Significance of adversarial loss
Training only on mean squared error (MSE) produces blurry results

(a) Adversarial (b) MSE only GT

Significance of partial convolutions
Reduce color discrepancies and blur
Increases checkerboard artifacts

(a) With PC (b) Without PC

Significance of local discriminators
Only slight differences observable
Tend to produce finer details

(a) With PC (b) Without PC

Conclusion

Main results

- To the best of our knowledge, we are the first to train a GAN with partial convolutions
- Adversarial training is absolutely crucial to reproduce the finer details of an image
- Good results can be obtained with significantly smaller networks than state-of-the-art approaches
- Partial convolutions improve the results compared to typical convolutions

Future work

- With more computational resources, the model could have been trained to inpaint irregular shapes
- Since partial convolutions successively fill the holes from the edges inward, the size of the inpainted region is limited by the depth of the network
- Future work could train a **deeper network on irregular masks** to perform well on larger and randomly selected masks

References

[1] D. Pathak et al. "Context Encoders: Feature Learning by Inpainting". In: *Berkeley University* (2016). URL: https://arxiv.org/pdf/1604.07379.pdf.

[2] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Sixth International Conference on Learning Representations* (2015). URL: https://arxiv.org/abs/1505.04597.

[3] S. Iizuka, E. Simo-Serra, and H. Ishikawa. "Globally and Locally Consistent Image Completion". In: *ACM Trans. Graph. 36, 4, Article 107* (2017). URL: :%20http://dx.doi.org/10.1145/3072959.3073659.

[4] G. Liu et al. "Image Inpainting for Irregular Holes Using Partial Convolutions". In: *Nvidia* (Dec. 2018). URL: https://arxiv.org/abs/1804.07723v2.