The Analysis of Image Inpainting Based on Pix2Pix Model and Mix Loss

Xu Yan
School of Statistic and Data Science, Nankai University, Tianjin, China

Keywords: Image Translation, Pix2Pix, Loss Function.

Abstract: This paper investigates the application of image translation as an important use case for generative adversarial networks (GAN), which has received widespread attention from scholars in recent years. This study builds an Image-to-image translation (Pix2Pix) model with a U-net as the generator and PatchGAN as the discriminator and observes the performance by adjusting the loss function of the generator. First experiment is modifying the scale factors for GAN Loss and Mean Absolute Error loss (L1 Loss), and the second is exchanging L1 Loss with square loss function (L2 Loss). After comparing the image authenticity and detail processing of different results, it is noticed that an overall better translation is achieved when the scale factor is set to 1:100. If finer detail handling is required, lowering the scale factor to 1:10 can be beneficial. However, it's also found that including L2 Loss in the generator loss function do not yield favorable results. It provides guidance for future choices of hyperparameters for the pix2pix model and lays the foundation for further research into loss functions.

1 INTRODUCTION

Image processing is a technique that can repair damaged portions of target images, reconstructing them to generate high-quality, deeply semantically approximated original images. In recent years, with the advancement of computer computational capabilities and rapid development of machine learning, achievements in computer vision have greatly enhanced scientific technology and human quality of life. Deep learning-based image processing techniques play a crucial role in many practical applications, such as object removal in image editing, restoration of old photos, repair of occluded portions of specific objects, facial restoration, and more (Lecun 1998). Currently, it is one of the main focal points of research in the field of computer vision.

The inception of Convolutional Neural Networks (CNN) enabled to extracting image semantics and features, making it one of the earliest neural network models employed in image processing. Moreover, due to its capability to extract image features, CNNs can also be utilized in tasks like texture synthesis and image style transfer (Gatys et al 2015 & GatyS et al 2016). The introduction and widespread application of Generative Adversarial Networks (GAN) have further enhanced the visual outcomes of image inpainting (Goodfellow et al 2014). In the realm of image processing using GANs. Yu et al. introduced the concept of gated convolutions, which elevated the effectiveness of image restoration (Yu et al 2019). However, due to the larger model size and a high number of parameters, training costs are significantly increased. Also, because of the uncertainty in filling the missing regions of damaged images using regular GAN methods, it is challenging to determine the inpainting area, which can even lead to severe restoration errors. A two-stage visual consistency network was proposed, consisting of a mask prediction module and a robust restoration module (Wang et al 2020). This approach significantly improved the model's generalization ability. By incorporating image semantic understanding by introducing an attention mechanism, the precision of the restored images can be further enhanced (Yu et al 2018). While the former often leads to image discontinuity issues, Liu et al. introduced a coherence semantic attention mechanism, as mentioned in reference, which focuses on the interrelation of deep features in the area to be restored (Liu et al 2019). This effectively resolves color discontinuities and boundary distortions. Image-to-image translation with Conditional Adversarial Networks (Pix2Pix) is an

Yan, X.
The Analysis of Image Inpainting Based on Pix2Pix Model and Mix Loss.
DOI: 10.5220/0012804900003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 438-443
ISBN: 978-989-758-705-4

image translation technique based on GAN. It is used to transform images from one type to another type. The same model architecture exhibits varying gradients under different loss functions, thereby affecting training efficiency and outcomes. Indeed, by modifying the loss function expression, you can enhance the clarity and coherence of generated images.

The main objective of this study is to modify the loss function formula of the generator to observe the effects of different loss functions on image translation results. The pix2pix model employs Mix Loss, a combination of GAN loss and Mean Absolute Error loss (L1 loss), to define the generator's loss function. This article aims to explore the advantages and limitations of this defined approach. First, the differences are observed in training effects between adding L1 loss to the generator's loss function and not having L1 loss. Second, it conducting multiple experiments by adjusting the ratio of L1 loss to GAN loss in the generator of the pix2pix model. The prediction results are obtained for the same input data from different models trained for the same number of iterations. Third, the training results of the model at the same epoch under different coefficient ratios are observed and compared. In addition, this paper employs No-Reference Blind Video Quality Assessment (NR-BVQA) in combination with human subjective perception to assess the coherence and authenticity of generated images. The experimental results demonstrate that using only the GAN loss can lead to gradient explosion, and if the L1 loss ratio is too small, it can result in severe image distortion, while if the L1 loss ratio is too large, it can cause the image to become overly blurry. In particular, if L1 loss is replaced with square loss function (L2 loss), although it can accelerate convergence, the final results may become excessively smooth and lose image details.

2 METHODOLOGY

2.1 Dataset Description and Preprocessing

The dataset used in this study called “facade” is sourced from Kaggle's pix2pix dataset (Dataset). This dataset consists of paired images, with each pair containing a photograph of a building and its corresponding sketch at the same resolution. In this dataset, the network is trained to translate hand-drawn sketches of buildings. The entire dataset comprises 400 pairs of training data and 106 pairs of test data. Figure 1 shows an example of the training data.

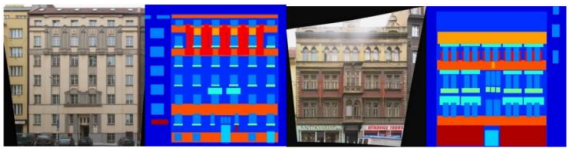

Figure 1: Images from the facade dataset (Original).

2.2 Proposed Approach

Under the same model, different choices of loss functions can impact the convergence of the loss function, the model's learning rate, and the accuracy of predictions, significantly influencing the model's performance. This paper places a strong emphasis on analyzing the performance of the pix2pix model under different generator loss functions. This analysis includes modifying the hyperparameters of the loss function and changing the composition structure of the generator loss. By observing the prediction results of the same sketch under different loss functions, it seeks to explore the strengths and weaknesses of the loss function choices as discussed in the original paper (Isola et al 2017). In the end, the paper aims to provide insights into selecting the most suitable loss function for different tasks. The experimental workflow is illustrated in Figure 2.

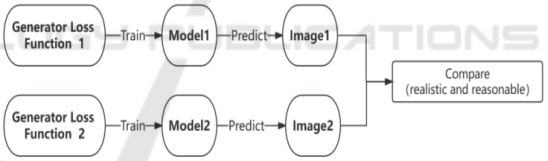

Figure 2: The pipeline of the study (Original).

2.2.1 L1 Loss

The L1 Loss is calculated by adding the absolute differences between each predicted value and its corresponding target value, and then taking the average. The mathematical expression for L1 Loss is:

$$L1\ loss = \frac{\Sigma|y_{pred} - y|}{n} \quad (1)$$

where the y_{pred} represent the prediction of y and y represent the true value of y. One characteristic of L1 Loss is that it's insensitive to outliers because it employs absolute differences. This property makes it perform well on datasets with a considerable amount of noise. L1 Loss can be used to measure the absolute difference between generated images and target images. This helps improve the stability of the

generator, making the generated images closer to the target images. During training, L1 Loss serves as an important feedback signal, assisting the generator in gradually producing more realistic images.

2.2.2 L2 Loss

L2 Loss, also known as Mean Squared Error, is indeed a loss function commonly used for regression problems. The mathematical expression for L2 Loss is as follows:

$$L2\ loss = \frac{\Sigma(y_{pred}-y)^2}{n} \quad (2)$$

The advantage of L2 loss is that it is a smooth, continuously differentiable function, which makes it easy to handle in optimization algorithms like gradient descent. Additionally, it is typically a convex function, implying it has a global minimum. However, L2 loss calculates errors using squared terms, which means it is more sensitive to large errors because squaring amplifies the impact of these errors.

2.2.3 GAN Loss

Adversarial Loss as one of the primary implementations of Generator Loss, is typically represented in the specific form of Binary Cross-Entropy Loss. It pushes the model's training by having the generator and discriminator engage in mutual competition. The generator aims to create more realistic data, while the discriminator strives to differentiate between real and generated data. This adversarial training approach leads to continuous improvement in the generator's ability to produce more authentic data. However, GAN loss training is often less stable compared to traditional supervised learning because the competition between the generator and discriminator can result in oscillations during the training process. Therefore, it's crucial to carefully select hyperparameters and employ various techniques to stabilize the training. Additionally, during the experiments, we also attempted to replace L1 Loss with L2 Loss and observed the training results.

2.2.4 Unet

The generator in the pix2pix model uses a U-net network architecture. It is a deep learning convolutional neural network architecture consisting of an encoder and a decoder. The first half is used for feature extraction, while the second half is used for upsampling. Its specific network architecture is shown in Figure 3. In this architecture, the down-sampling path consists of convolutional layers and pooling layers, which are used to reduce image resolution, decrease the spatial size of the image, and simultaneously extract image features. In contrast, the up-sampling path serves the opposite purpose and has a complementary architecture compared to the down-sampling path. U-Net also employs skip connections to connect feature maps of different depths between the encoder and decoder. This helps in transmitting both low-level and high-level features, addressing the common issue of information loss. The U-net network's five pooling layers enable it to achieve multi-scale feature recognition in images, making it highly effective for semantic image segmentation.

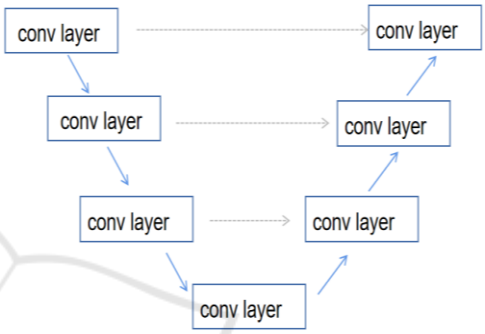

Figure 3: Unet architecture (Original).

2.3 Implementation Details

In this experiment, the choice of using the Adam optimizer is justified. Adam has relatively low demands on storage and computational resources, making it advantageous for deep neural networks dealing with large-scale data and parameters. Additionally, Adam's ability to adaptively adjust the learning rate can accelerate the model's training process. Simultaneously, we trained each model for a total of 20 epochs to ensure that we could observe variations in training results among different models and also assess differences in training efficiency. In the selection of hyperparameters, the experiments conduct four different ratios of GAN Loss to L1 Loss, namely 1:100, 1:200, 1:10, and 1:1, and observe the impact of GAN Loss and L1 Loss on the training results under these various ratio combinations.

3 RESULTS AND DISCUSSION

In the results section, the paper showcases and discussed the image translation outcomes under various loss functions. Keeping the training dataset consistent with the pix2pix dataset and maintaining all

training parameters except the loss functions unchanged, different models will attempt to translate the same test image at the same number of training iterations and then compare the translated images to evaluate the differences between different models. This chapter will consist of two parts: Scale Factor Selection and Choice of L1 and L2.

3.1 The Performance of Scale Factor Selection

As can be seen from Figure 4-7, under 20 epochs, each model has provided its respective training results. From the loss function curves, it is evident that when the L1 loss proportion is relatively higher, there is a noticeable downward trend in the loss function. This leads to good prediction performance even before complete convergence is achieved. Conversely, when the L1 loss proportion is too small, the loss function exhibits strong oscillations, and there is no sign of convergence even with increased training cycles. This behavior is attributed to the nature of the GAN loss itself, which struggles to converge without the presence of pre-training data. From the translation of the hand-draw images, it can be observed that the translation results are better when the scale factor is set at 1:10 and 1:100. In comparison, when the L1 loss proportion is too small, although the image resolution is higher, it leads to the generation of more inconsistent regions. On the other hand, in cases where the L1 loss proportion is higher, since the L1 loss measures the absolute difference between the original and predicted images, it results in both reduced image resolution and enhanced image smoothness. Therefore, adjusting the scale factor to balance resolution and smoothness is a key aspect of the experiment.

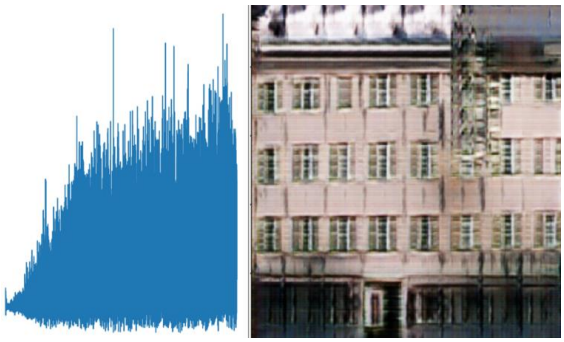

Figure 4: Loss Function Curves and Results When Generator Loss = 10GAN Loss + L1 Loss (Original).

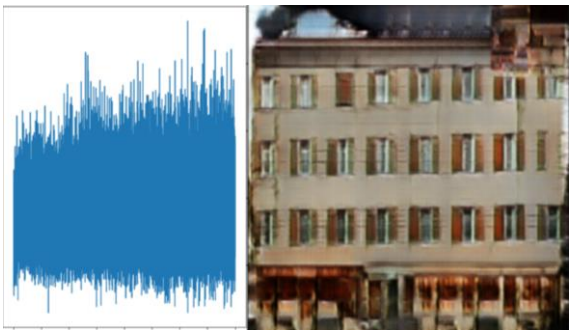

Figure 5: Loss Function Curves and Results When Generator Loss = GAN Loss + 10L1 Loss (Original).

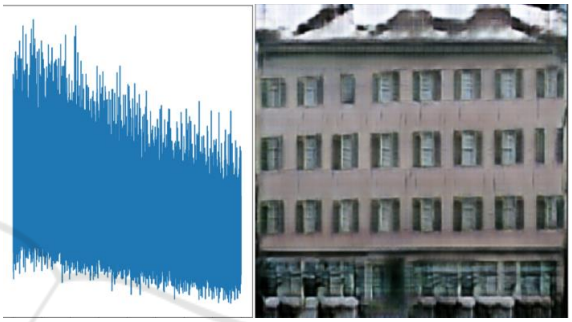

Figure 6: Loss Function Curves and Results When Generator Loss = GAN Loss + 100L1 Loss (Original).

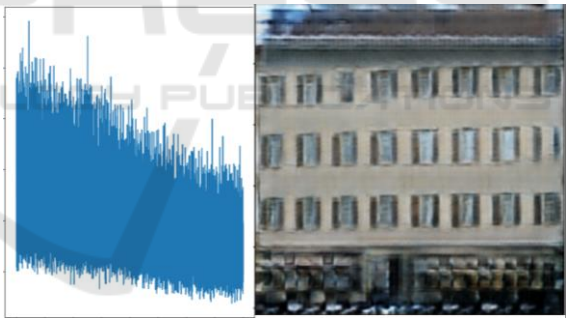

Figure 7: Loss Function Curves and Results When Generator Loss = GAN Loss + 200L1 Loss (Original).

3.2 The Performance of Choice of L1 and L2

As can be seen from Figure 8-11, L1 loss tends to generate relatively smooth results but may result in the loss of some details, whereas L2 loss tends to produce more precise translated images but may also make the generated images more susceptible to noise. As can be seen from figure 8-11, when the scale factor is set to 1:100, it can be observed that L1 loss converges significantly while L2 loss does not. This is because L2 loss is more likely to get stuck in local minima and may struggle to achieve a smaller loss, whereas L1

loss, being a convex function, does not face this issue. At the same time, it can also be observed from the results that when using L2 loss, the images are noticeably blurrier, and artifacts are introduced. This occurrence is because the square operation in L2 loss penalizes large errors but is relatively insensitive to small errors. As a result, it may not perform well in terms of fine detail.

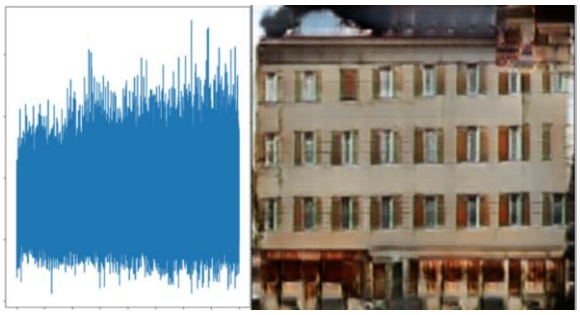

Figure 8: Loss Function Curves and Results When Generator Loss = GAN Loss + 10L1 Loss (Original).

Figure 9: Loss Function Curves and Results When Generator Loss = GAN Loss + 10L2 Loss (Original).

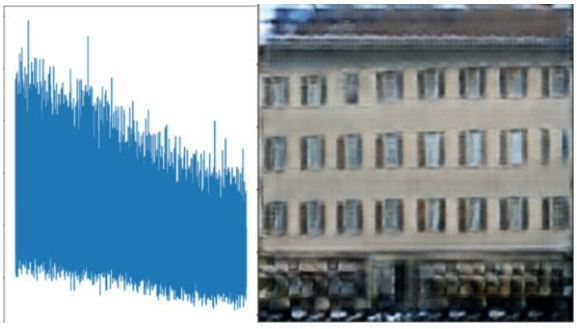

Figure 10: Loss Function Curves and Results When Generator Loss = GAN Loss + 100L1 Loss (Original).

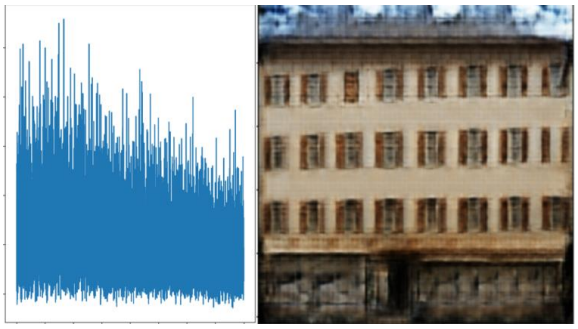

Figure 11: Loss Function Curves and Results When Generator Loss = GAN Loss + 100L2 Loss (Original).

In summary, properly setting the ratio between GAN Loss and L1 Loss can indeed contribute to the training and translation performance of an image translation network. Additionally, under specific requirements, adjusting the values of these scale factors can be used to control the model's translation effect effectively.

4 CONCLUSION

This study presents the generation results of the Pix2Pix model under different generator loss functions. First, it constructs a Pix2Pix network with a U-net as the generator and PatchGAN as the discriminator. Second, it trains the model for image translation ability using the Kaggle Pix2Pix dataset as a training set. Then, it adjusts the scale factors for GAN loss and L1 loss to observe overall better translation results. The findings indicate that a higher ratio of L1 loss results in lower resolution, while a higher ratio of GAN loss leads to inconsistent regions in the generated images. The study reveals that setting the scale factor to 1:100 results in images that combine realism and good resolution. On the other hand, a scale factor of 1:10 can be employed to improve resolution, particularly for finer details. Finally, the research identifies that L2 loss does not promote model convergence and is less suitable for handling details compared to L1 loss. Therefore, it is not recommended for use in this model. In the future, considering the incorporation of other loss functions, such as SSIM loss, into the generator loss function is a promising avenue for enhancing image translation results.

REFERENCES

Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, “Gradient-based Learning Applied to Document Recognition,” Proceedings of the IEEE, vol. 86, 1998, pp. 2278-2324.

L. Gatys, A. S. Ecker, M. Bethge, “Texture synthesis using convolutional neural networks, Advances in neural information processing systems, 2015, p. 28

L. A. GatyS, A. S. Ecker, M. Bethge, “Image style transfer using convolutional neural networks, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2414-2423.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. “Generative adversarial nets,” Advances in neural information processing systems, 2014, p. 27.

J. Yu, Z. Lin, J. Yang, et al. “Free-form image inpainting with gated convolution,” Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4471-4480.

Y. Wang, Y. C. Chen, X. Tao, et al. “Vcnet: A robust approach to blind image inpainting,” Computer Vision–ECCV 2020: 16th European Conference, Glasgow, 2020, pp. 752-768.

J. Yu, Z. Lin, J. Yang, et al. “Generative image inpainting with contextual attention,” Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5505-5514.

H. Liu, B. Jiang, Y. Xiao, et al. “Coherent semantic attention for image inpainting,” Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4170-4179.

Dataset https://www.kaggle.com/datasets/vikramtiwari/pix2pix-dataset

P. Isola, J. Y. Zhu, T. Zhou, et al. “Image-to-image translation with conditional adversarial networks,” Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125-1134.