RESOURCE EFFICIENT IMAGE INPAINTING TINY PAPER SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, many breakthroughs have been established in the field of image inpainting, thanks to the advent of Adversarial and Diffusion networks. However, these approaches have a crucial drawback, they utilize superfluously complex models or training algorithms, especially those based on Vision Transformers (ViT's) Dosovitskiy et al. (2020). Though these methods have proven to produce good results, the increase in resource consumption with respect to the increase in performance is not a scalable solution. This paper diverges from the above-mentioned norm by proposing a simple U-net architecture supplemented with Discreet-Wavelet-Transformation (DWT) based token mixing algorithm, Wavemix Jeevan et al. (2022), as an alternative. Finally, the paper illustrates that the proposed model outperforms the Large Mask inpainting model (LaMa) Suvorov et al. (2021) on various metrics, with less than half the parameter count and consuming considerably lower time during training and evaluation cycles.

1 INTRODUCTION

Image inpainting is the restoration process, where blemishes, holes and other defects of an image are filled in to match the available context, thereby replenishing the image's quality. A number of learning-based approaches have been proposed to tackle this problem which guarantee excellent performance, but they require high computing resources, especially for networks based on ViT Dosovitskiy et al. (2020). To solve this issue, a new class of networks called token mixers, or Metaformers Yu et al. (2021) has been proposed. Wavemix Jeevan et al. (2022) is one such Metaformer. These alternatives consume a fraction of the resources as ViTs but their performance is competitive with that of ViTs, in simple tasks such as classification and segmentation. That being said, there is relatively less research done on their performance on image generative/restorative tasks. This paper is an investigation and extension to the literature on the application of such alternatives, in specific, Wavemix to the task of image inpainting.

One key aspect of inpainting that differentiates it from problems like classification and segmentation is that it is an ill-posed unsupervised learning task. Unsupervised because the input and output can be generated by simply using a binary mask on any image, refuting the need for additional information in the form of labels. Ill-posed because any considerably large defect in any image could have multiple equally valid and contextually correct inpainted solutions. These aspects make the problem much harder to train on and to evaluate.

2 THE APPROACH

The model proposed in the paper is an approach that is inspired by wavemix Jeevan et al. (2022) as a base model. This choice is motivated by the unprecedented success of wavemix in approacing the SOTA for different datasets on the task of parameter efficient image classification [1]. A few key changes were made to the original architecture to accommodate the image generation task. The main change is the removal of the reduction convolution as the first layer within waveblock. Although this reduced the number of learnable parameters, it allowed for a complete image representation to pass through the network. Secondly, transpose convolutional layers were added to the architecture to

[1] https://github.com/brohrer/parameter$_e ff iciency_l eaderboard$

supplement the MLP layers. Lastly, the loss function was replaced with a weighted linear combination of Structural Symmetry Index Measure (SSIM) and L1 (3:1) to capture the model's coherence and regeneration capabilities.

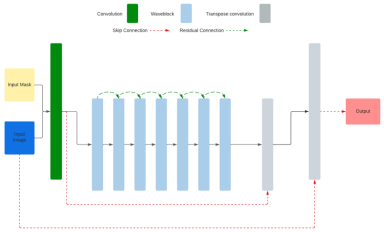


Figure 1: Auto-encoder architecture

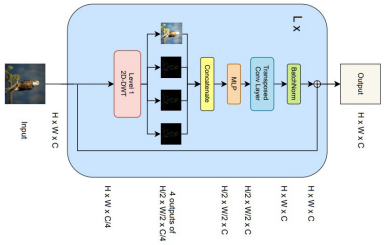


Figure 2: Waveblock architecture

Within the context of the paper, STL10 dataset (Coates et al., 2011), more specifically, the unlabeled images within the dataset were used for training, validation, and testing of various models. It comprises 100,000 images of 96px X 96px resolution with three channels (RGB). The choice of this dataset over the more conventional Imagenet is a result of constraints over the availability of time and resources. All the results, and conclusions have been established on the same dataset. Lastly, the masks for the validation and test set were pre-generated and saved to ensure uniformity while testing and evaluating different models, but the masks for the train set were generated on the fly. The masked image is fed as input to the model, and the output is evaluated against the unmasked image.

3 Result

The table bellow illustrates the relative performance of the proposed model, a traditional deep convolutional neural network and LaMa Suvorov et al. (2021). From the table it is clear that the wavemix model outperforms the baseline models (CNN & LaMa) on the metrics chosen for evaluation. The model has a faster training and inference time. Following this observation, one possible justification was conjured; ironically, it is possible that the sheer number of excess parameters in the baseline led to a case overfitting on the train set, which ultimately led to poor performance on test set. Although many augmentations and mask randomization operations were performed to overcome this situation, it is still a possibility that cannot be ruled out without further investigation.

Table 1: Standardized evaluation against baseline

Model	**# param (x10^6)**	**FID Score ↓**	**SSIM ↑**	**Train Time** [2]	**Inference Time** [3]
Deep CNN	12	86.373	0.850	11	0.52
LaMa	27	70.294	0.893	81	0.89
Wavemix (ours)	14	64.784	0.902	23	0.73

4 Conclusion

This paper sheds light upon resource efficiency for the less explored task of image inpainting. The performance of the proposed model is comparable to much larger models,on the given dataset, but the quality of inpainting drops significantly as the size of the mask increases. This indicates that the model is able to capture global context, but it is unable to produce semantically coherent output due to lack of generative ability, a possible solution for that being, a simple change of architecture to that of a GAN or a diffusion model. Thus, this paper establishes the potential of Metaformer models for the task of image inpainting.

[2] minutes per epoch
[3] seconds per 100 images

ACKNOWLEDGEMENTS

I wish to record a deep sense of gratitude to Prof. Amit Sethi, my supervisor for his valuable guidance and constant support towards my Master's study and related research. I also wish to extend my gratitude towards Pranav Jeevan P for his support throughout my research.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. *Journal of Machine Learning Research - Proceedings Track*, 15:215–223, 01 2011.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

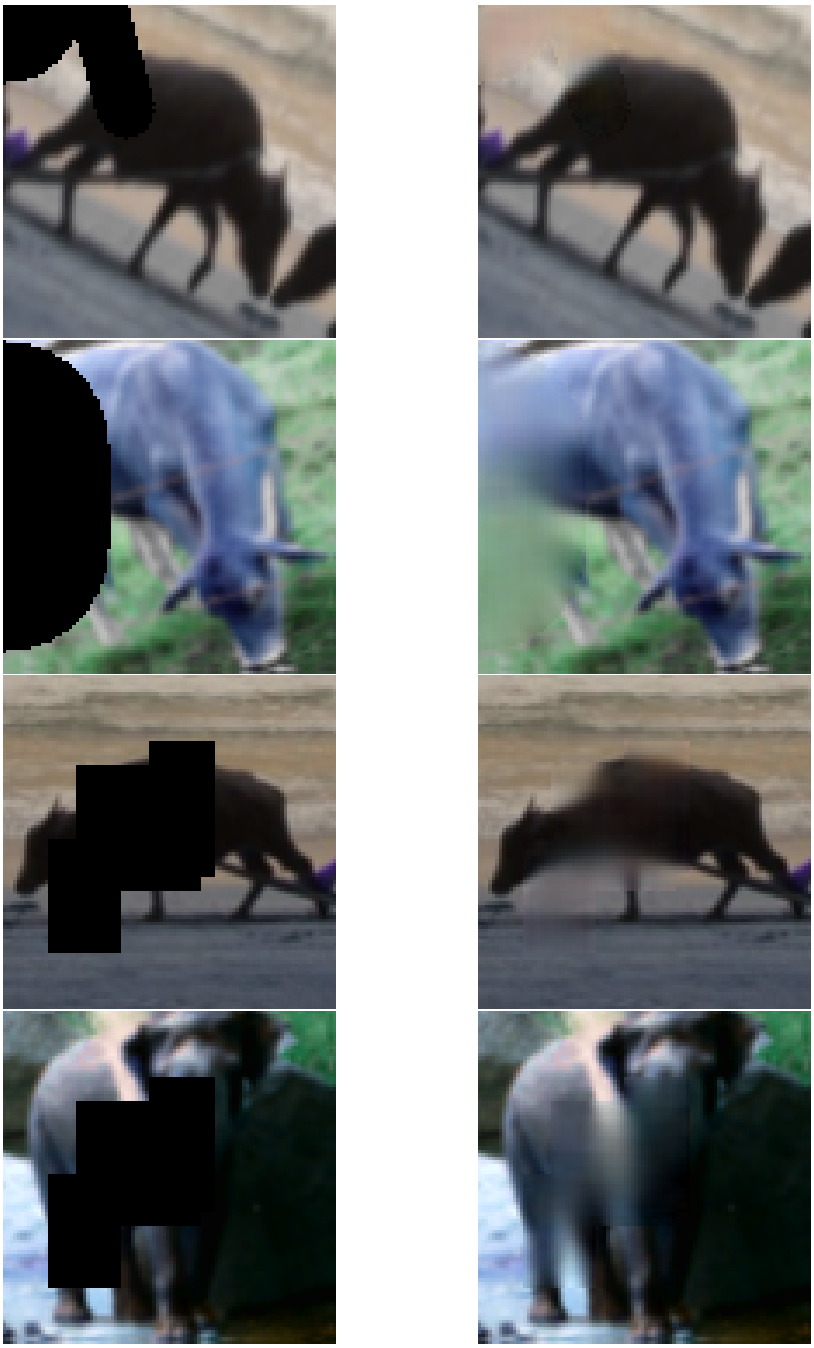
Pranav Jeevan, Kavitha Viswanathan, and Amit Sethi. Wavemix-lite: A resource-efficient neural network for image analysis, 2022. URL https://arxiv.org/abs/2205.14375.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions, 2021. URL https://arxiv.org/abs/2109.07161.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. *CoRR*, abs/2111.11418, 2021. URL https://arxiv.org/abs/2111.11418.

A APPENDIX: VISUAL EXAMPLES

The images on the following page are an illustration of the performance of the proposed model for visual evaluation (cherry-picked samples). One key observation from the samples is the ability of the model to understand context very well but its inability to produce sharp and high-quality infilled images,



(a) Input — Masked Image

(b) Output — Inpainted Image

Figure 3: Visual Performance of proposed model

B APPENDIX: MODEL SENSITIVITY ANALYSIS

This section is dedicated to the study of the effect of mask characteristics on the performance of the model, more specifically, the effect of mask size and mask randomization. As expected, the models performed worse on bigger masks, compared to smaller masks, as in the former, the model has little to no context from the rest of the image to properly fill in the mask. Therefore, the model

performance was severely affected on larger masks, but at the same time, training on too thin of a mask meant the model lost its generative abilities. This concept is illustrated in the image below.

(a) 8X8 mask (even contours the reflection)

(b) 12X12 mask (pixelated & blurry output)

Figure 4: Model sensitivity to mask size

Similarly, the effect of randomization followed the expectation, i.e., it helped the model learn better and generalize to various masks, during the inference time. The effect of randomizing the masks during training is subtle, but it is very much noticeable during inference, thanks to the generalizability that it imparts, due to this visual representation of this observation is not possible.