Non-Lossy Ground Truth Comparison via Convolutional Auto-Encoders for Rethinking Image Inpainting via a Mutual Encoder-Decoder with Feature Equalizations

Sander Gielisse — a.s.gielisse@student.tudelft.nl
Misha Katona — m.j.katona@student.tudelft.nl

April 2021

1 Introduction

In this short paper a new algorithm variant is proposed and evaluated for the 'Rethinking Image Inpainting via a Mutual Encoder-Decoder with Feature Equalizations' (MEDFE) paper by Hongyu Liu et al Liu et al. [2020]. All the code from the original paper is available through the author's GitHub repository, with data publicly available. Instead of doing a reproduction, we instead propose a variant model that is a possible improvement over the original MEDFE model. At the end of this paper we show that the proposed implementation performed possibly better on the CelebA dataset and therefore should be considered for further investigation.

2 Paper Recap

Image in-painting is used to reconstruct corrupted or lost sections of an image. These regions with no information are filled in using the surrounding 'uncompromised' image. This problem has conventionally been solved using encoder-decoder based CNN's where the texture and the structure are filled in step-by-step by two different networks. The paper proposes a novel solution where intermediate texture and structure representations are formed, which, after channel and spatial equalization, are fed to the decoder for more consistent results.

These two representations are extracted from the encoder part of the network, where the first 3 layers are used for the textures and the last 3 layers for the structure. This can be seen in Figure 1. Furthermore, in order for the network to learn both texture and structure representations, an intermediate ground truth is used to ensure that the two features are correctly learned.

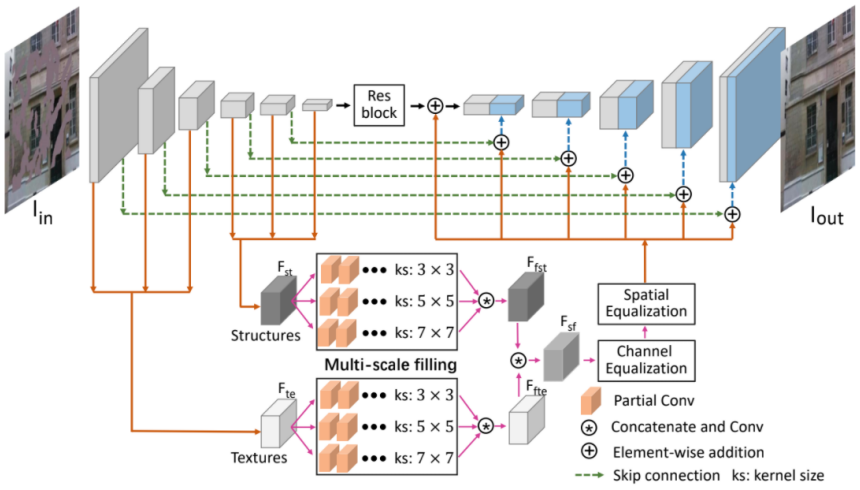

Figure 1: Image in-painting network overview

3 Purpose

The objective of this paper is to improve on the learning of the intermediate ground truths for the texture and structure feature representations. One aspect of the network that was not addressed is the information loss when the structure and texture ground truths are compared with the intermediate representations generated by the network.

The input and output images of the network are set to be 256 by 256 pixels over 3 channels (RGB), however the intermediate representation for both features are represented by 32 by 32 pixels, but with 256 channels. The approach taken by the paper is to use a 1x1 convolution to decrease the amount of channels to 3, which is a very lossy operation. Then, when the intermediate representations are evaluated with respect to the ground truth, another lossy compression is needed; that of the ground truth image. The authors of the paper used a simple bilinear sampling technique to shrink the images, which leads to a significant information loss. This is especially problematic for the textures ground truth as textures are usually high frequency fine detail, and therefore important information is assumed to be lost. Figure 2 gives a schematic overview of the approach of the original paper.

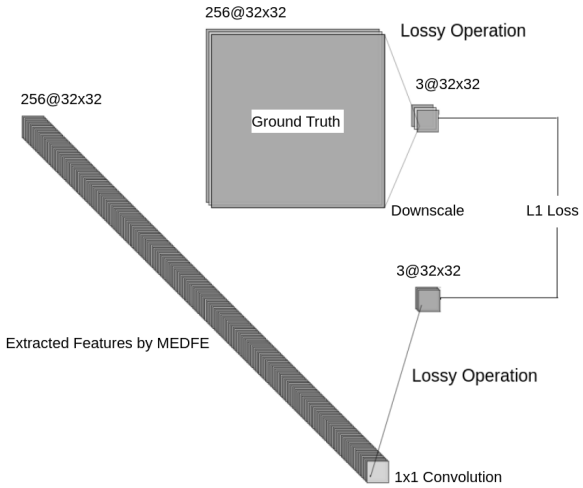

Figure 2: The L1 loss constraint in the original work.

We argue that this L1 loss is not a meaningful way of enforcing the entire found feature space to correlate with textures or structures. For example, the network could theoretically use just 3 out of the 256 channels to learn the desired texture or structure representation, where the 1x1 convolution would learn to literally copy those specific 3 channels. This would lead to a good L1 loss, but would allow the other 253 channels to represent any other value the network wants to. Because of this, we argue that even though the obtained [3, 32, 32] (channels, height, width) feature map correlates with the correct features, there is no way of forcing that the full feature map [256, 32, 32] also correlates with the correct features.

In this work we propose a solution to allow for the network to be trained with a near lossless representation of the ground truths without large alterations to the network and no change in intermediate resolution of two features.

4 Implementation

The goal of the new proposed implementation is in finding a way to compare the features extracted by the network which has a shape of [256, 32, 32], with the ground truth image which has a shape of [3, 256, 256], as can be seen in Figure 1. With the main emphasis being on conservation of information from the ground truth image when used for the training of the two intermediate features.

A convolutional auto-encoder is proposed which can learn a [256, 32, 32] representation from a [3, 256, 256] ground truth image. In practice we find that this convolutional auto-encoder that performs an almost lossless encoding and decoding, showing a near-zero L1 loss when comparing the input with the output image. This convolutional auto-encoder is trained completely separate from the network beforehand. The decoder is discarded and the encoder is loaded as part of the originally proposed network.

Two separate convolutional auto-encoders are trained; one for textures and one for structures, which can be seen in Figure 3.

We find that using a smaller intermediate bottleneck such as [256, 8, 8] forces additional information loss, which is mostly high-frequency information (textures). We find this an easier way of generating a representation of the structures for the image and thus replace the originally proposed relativistic total variation. One of the reasons why we do this is because the relativistic total variation was shown to be computationally inefficient; a simple GPU-optimized encoder network outperforms the relativistic total variation by a large margin in our experiments performance-wise, but further research is needed to confirm these findings. The figure below gives an overview of the proposed use of an encoder to do an almost lossless comparison of intermediate ground truth values.

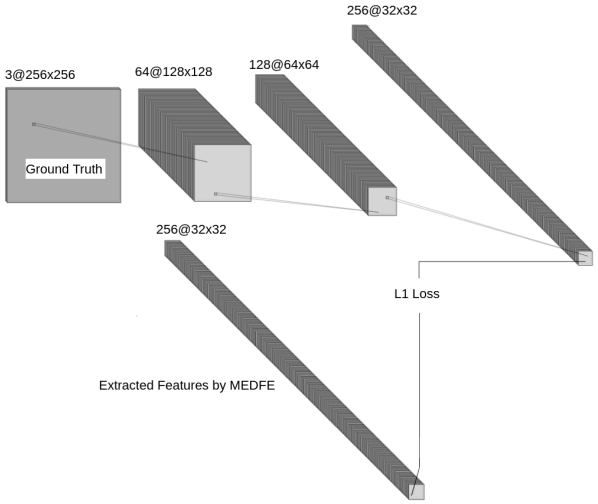

Figure 3: Proposed L1 loss where all operations are almost lossless.

No further hyper-parameter tuning was done for our model compared to the originally proposed approach MEDFE, except for multiplying the L1 loss with a constant value of 8; this way our L1 loss gets values that are approximately in the same range as their original L1 loss, suggesting they then thus play an equivalent role in the weight updates.

5 Results

Our model was trained for a total of 5 epochs on an NVIDIA RTX 3070 GPU. The total training time was around 3 days on this single GPU. Figure 4 on the page below shows some results from our proposed model on the test data. The images were 256 x 256 pixels, where the center 128 x 128 pixels were masked and so the network had to reconstruct.

5.1 Source Code

The implementation of our proposed model is a combination of the original MEDFE repository (`https://github.com/KumapowerLIU/Rethinking-Inpainting-MEDFE`) and the CycleGAN-and-pix2pix repository (`https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix`), of which we just only used the global set-up for the convolutional auto-encoder. Our combined repository with some modifications such as loading the encoder into MEDFE as well as changed ground truth calculations can be found at `https://github.com/SanderGielisse/MEDFE-CAE`.

Figure 4: Ground truth, MEDFE and our MEDFE + CAE results.

6 Evaluation

Next, we do an evaluation on the results produced using the proposed convolutional auto-encoder. On visual inspection, we are happy with the obtained results. For some images there are some minor visual mistakes, but overall we find it difficult to distinguish between the ground truth and inpainted image. As a way to do a more objective comparison, we calculate the FID score for the inpainted images with respect to the ground truth images. An observation here is that an FID score requires at least 2048 images. However, the original paper mentions the use of a random selection of 500 images and does not reveal any detail about the way their FID score is calculated. They could have calculated their FID score on fewer than 2048 samples, but FID scores Heusel et al. [2017] generated from fewer than 2048 samples are unlikely to correlate with the 'observed realism' it was designed for.

A alternative approach to comparing was to generate 2048 images ourselves using their pre-trained models which are available on GitHub. However, when simply cloning the repository, loading the model and doing predictions, there was clearly something wrong as the results looked awful. The reason for this might have been our possibly different approach on pre-processing the training data, or a mistake on their end. We did not further investigate this, so we were still unable to find a meaningful comparison for the FID score. Manual inspection of their generated faces would also have given a nice comparison, however unfortunately here we are again limited to the examples from their paper, which for CelebA are just two faces. Training their model till convergence was our only option, so we decided this would be our approach here. The training time for their model was again around 3 days.

We calculate our FID score on 2048 randomly selected samples from the test set and find our FID score to be 5.996. The FID score for their proposed original model was 6.279, so we consider this a nice improvement (lower is better). Visual inspection by us reveals situations in which are proposed variant outperforms the original one, but there are also cases where this is not the case. Unbiased selection of which image looks visually better by image experts would have been nice for for a significant subset of the test data, but this is outside of the scope of this project. In our biased opinion we often favor our results over the original model.

7 Discussion

It is quite impressive that the proposed implementation in this paper was able to outperform the original paper. However this was done on a dataset with only a very narrow type of data, that being celebrity faces with a missing central square. Further testing is required so that statistically significant improvement can be decisively concluded, both with increasing the number of runs, as well as the variety of data.

However even without statistical significance we hope that this paper can show insight that distance loss functions (such as L1) after lossy operations are not necessarily as one would expect. One should consider how information loss is handled with image based deep learning. With our implementation the intermediate ground truths are forced to learn the features that we want them to learn, in comparison to the original version in which there was a large degree of freedom for the network in those intermediate representations. The use of an auto-encoder is just an example on how information loss can be avoided when working with incompatible shapes, but many more options might exist. Since the new intermediate ground truth has the same shape as the network uses, this enables it to constrain the network in a possibly more meaningful way. This could mean that the network is forced to learn a better or more useful representation to be used for the two features, which layers later in the network could benefit from.

References

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. *arXiv preprint arXiv:2007.06929*, 2020.