Homework 2 report

Advanced Machine Learning course, taught by Prof. Fabio Galasso
A.Y. 2020/21

Riccardo Ceccaroni
1884368

Simone Ercolino
1587229

Romeo Lanzino
1753403

Dario Ruggeri
1741637

November 2020

Abstract

This work has as main purpose to improve the results of classification on small images, and as a corollary attempts to propose a novel feature loss for tacling the task of Super Resolution.
The main task tackled is the classification of the 1000 classes of ILSVRC2012 (Imagenet dataset) after a substantial downscaling, with a linear reduction of more than $3\times$ *(scale factor of 0.32) on the sides of already rescaled to* 256×256 *pixels images (note that the average size of images in Imagenet is* 482×418 *pixels).*
The main feature of the model is the use of a pre-trained Super Resolution network (ESRGAN [1]) in order to reconstruct the downscaled images before classification with the Resnet50 network pre-trained on Imagenet. In particular, the Super Resulotion network is fine-tuned using only a classification based loss, in an hybrid Teacher-Student framework with Resnet50 as teacher of the ESRGAN network.
This approach brought to promising results on both the classification on low resolution images and the super resolution. Results that leave room for further improvement, but are already worth of being presented.

Introduction

The problem of recovering information (especially object details) from low resolution images is common to possibly any Computer Vision task.
Our work goes in the direction of super-sampling a tiny image for the purpose of classification; for this task, we take as upper-bound baseline the cross-entropy obtained on predictions made by Resnet50 on the original 256×256 crops, that are used as ground truth for the training of our model.
The framework we developed may be easily applied to other scenarios, where we may have different classes to predict, by doing the fine-tuning of DarioNet using other pre-trained models, so with same feature loss, but different teacher network; or even by completely changing the task, for example by using a model for image segmentation or object detection (in this case the feature loss should also change, and possibly the results we have achieved with our experiment may not hold anymore).
In addition, the proposed framework gives a possibly new approach (different from the latest and widely used perceptual loss) for dealing with the task of super resolution. The proposed hybrid Teacher/Student network, named Darionet, gives good visual results on the super resolution of small images, with good definition of edges and small presence of artifacts (mostly caused by the very low resolution of the input). Results are compared with the pre-trained RRDB architecture Table 1, that gives a lower bound baseline for both classification and Super Resolution tasks.

Related work

Several researches have been carried on in the last years on the topic of Super Resolution.
The ESRGAN [1] network enhanced with the residual blocks RRDB is, as far as we currently know, the state of the art in the field; it has been trained on the Div2K dataset [3] a small dataset of two thousand large images, mostly artistic photos.
More researches are done toward video super resolution, using also transformers architectures. One of the most recurrent issues in the filed are regarding the losses. Indeed one of the most used loss for this application, the PSNR, is broadly considered unstable for training and seems to lead to blurred images. [1] The alternative to PSNR oriented methods seems to be the training using perceptual loss that is composed by the $L1$ loss between the reconstructed image and the original one and the a loss based on the activation maps based on a pretrained model that works on the images. [1]

Proposed method explained

As it's shown in the image 1 the model pipeline is a sequence of:

1. A Scaler, a model that simply performs a downsampling of the image, reducing the image by one third of its original dimensions;

2. Then the little image is fed to DarioNet; a

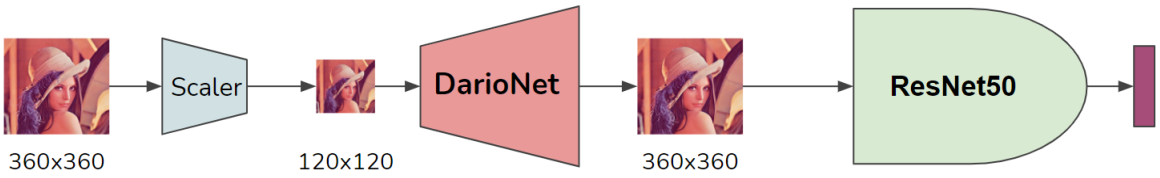


Figure 1: *Model structure*
First the downsampling, then the supersampling with DarioNet and finally the classification with ResNet50

model obtained by the finetuning of the ESRGAN network on the ImageNet2012 dataset;

3. Finally the supersampled image coming out from DarioNet is directed to ResNet50 pretrained in order to get the prediction for the reconstructed image.

The model is finetuned (only the DarioNet architecture) in a teacher student framework where the loss is the MSE between the scores (before the softmax) of ResNet50 applied on the original image and the scores of the same classifier applied on the reconstructed image.
Then the performances are evaluated on the test dataset to see how much our novel architecture lose in terms of classification accuracy.

Dataset and Benchmark

The dataset used for the fine-tuning is ImageNet2012 [2].

- Train: 15% of the ImageNet2012 train dataset, almost 21GB;
- Validate: 50% of the ImageNet2012 validate dataset, around 3.5GB;
- Test: the other 50% of the ImageNet2012 validate dataset, around 3.5GB (since the test dataset of ImageNet2012 doesen't have the labels).

Since the best classification performances we could achieve are the one of ResNet50 applied on the original images, so we traced the performances of our model in respect of that one.
On the other end, the classifier applied on the super-sampled images with either the simple bicubic upscaler or SRGAN pretrained were our lower benchmarks, so the best performances that we may achieve on small images without DarioNet. (Table 1)

Experimental results

The experiments with Darionet offered many possible metrics for evaluating its performance in both SR and classification tasks, with respect to classification results on the original images coming from Resnet50 and Super Resolution results coming out of the original ESRGAN. Since our main concern was to evaluate the possibility to make a successful (re-)classification after down-scaling, most of the metrics used are linked to Resnet50 predictions, which is indeed the true final stage of our end-to-end pipeline. So here the reported MSE values are those calculated with the two vectors of predictions coming out of Resnet when passing the original images and the reconstructed images. So in this framework, predictions on original images are to be considered the ground truth to be reconstructed from the input after applying the Darionet pipeline. This is also closely linked to the cross-entropy of each of the two predictions vector with the ground truth vector of labels, which is also reported, and in particular the difference between these two cross-entropies is used as a validation loss, so that when the minimum for this loss is reached on the validation set, Resnet50 correctly classifying (on average) the highest possible rate of images coming form Darionet that were also correctly classified from the original 256×256 pixels crops. The effects of training on these metrics, as well as on the average PSNR and Cross-entropy of the predictions on Darionet upscaled images with

Name of the model	Average Cross Entropy loss	Average PSNR (Db)	Accuracy	Total time (test) (minutes)
Model 1 (Original images)	0.9112	∞	0.7676	1.5
Model 2 (bilinear upscaling)	1.9864	**25.2466**	0.5446	1.5
Model 3 (ESRGAN super resolution)	1.6946	22.5088	0.6063	28
Model 4 (**DarioNet** super resolution)	**1.0931**	25.2211	**0.7275**	28

Table 1: *Comparison on the four models performances*

the vector of GT labels, are summarized in plots of their values over the 10 epochs of training in Figure 2. As plots show, the validation loss reaches a minimum on validation set at epoch 8, so the weights of the model were saved at that epoch.

By looking in details at the performances we may have some more information about the difference in performances between the two architectures.
In figure 3 we can see how the two architectures behaves also considering the hardness of the classification class.
We can see that on average we have a worsening of the performances with DarioNet (we expect that since it works on images that are $\frac{1}{3}$ of the original ones) but we see also that we actually have some classes on which the novel network performs better than the old one.
From the $3D$ plot we can go further and see that the worsening, or the improvement, on performances seems to not depend on the difficulty of the class, a good sign.
Next we can analyze the performances changes on the hardest classes of both the networks, figure 4. Can be seen that the difference between the difference between the two models oscillate, and tends to a worsening in performances but how we have seen also in the figure 3 it seems that there is not a strong correlation between hardness of the class and worsening in performances.
Finally we can make a brief comment on the plots shown in figure 5; as we where expecting the major worsening are greater that the bigger improvements. We should keep in mind that some of the classes in these plots may be there for a meter of chances; but others are there for specific reasons, for example we can observe that targets that shows geometric shapes seems to be well reconstructed and classified; on the other end objects classes that are recognized mostly by the details, for example dog breeds seems to be poorly reconstructed, mainly because we lose that information in the undersampling.
Below in the *Appendix* there are some images to show the supersampling performances of DarioNet on these extreme classes.

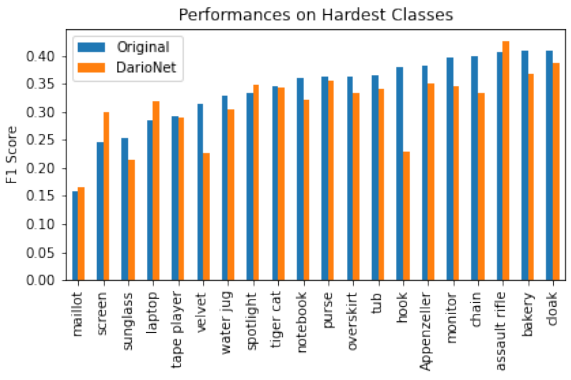


Figure 4: *The hardest classes to predict according to the $F1$ measure on the performances of ResNet50 on the original images*

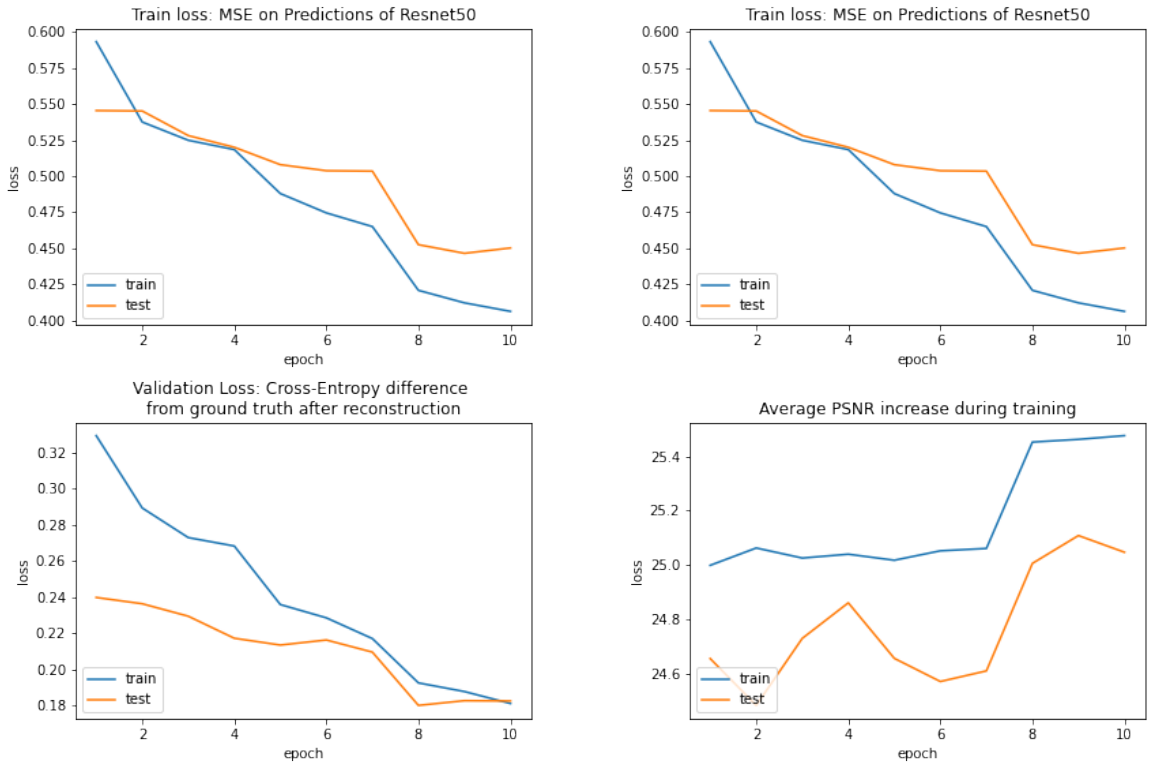


Figure 2: *Plots of the values of representative classification and Super Resolution metrics during training of 10 epochs*

Conclusions and Future work

Our starting idea was to evaluate the results of super-resolution on the results it gives on a different but related task. This comes from the simple idea that if the Super Resolution is able to reconstruct at least enough information of an image such that the same classifying network is able to predict (with a small loss) the same vector of probabilities starting from the original and reconstructed image, then its results should be, up to a certain degree, good enough also for a human observer.
But the framework we developed proved worthy of being explored also as a way to improve the results on the classification task starting from low resolution inputs, and this became also the primary interest of our work. This can be useful in the task of image compression (paired with an efficient compression algorithm, possibly another neural network), but also as is, as a general framework to improve classification on small images.
This could be done by testing Darionet against the tinyImagenet dataset classification, possibly with an even more advanced teacher network (Resnet50 proved to be already a good enough teacher to reach our goal, but a better accuracy of the classifying network improves also the learning capacity of Darionet), that should also be fine-tuned to recognize the 200 classes of the small dataset. Another consideration that we came up with is that possibly a different feature loss coming from a different task (more strictly related to the whole image structure), like semantic segmentation or detection, could offer a more more information with which Darionet could be trained, and this remains a very feasible task, having to make just small changes to the way in which the feature loss for training Darionet are extracted, and just having to change the pre-trained teacher network (if necessary).

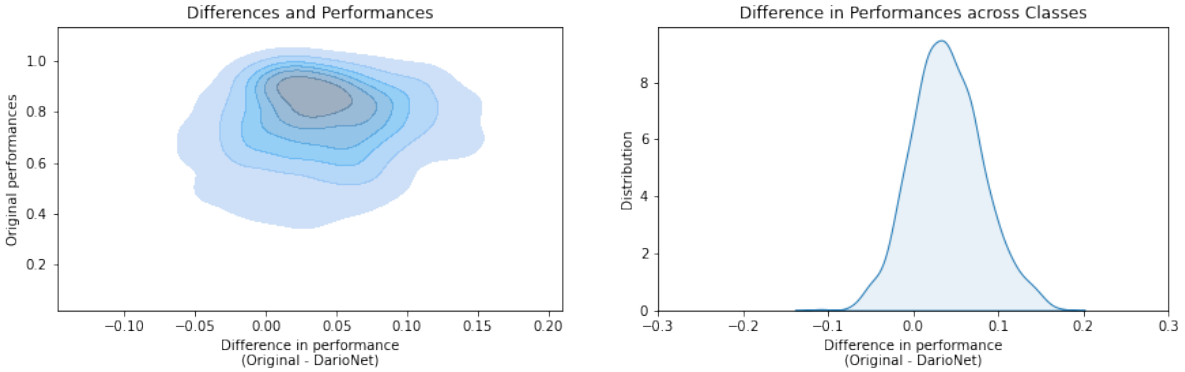


Figure 3: *Distribution of the difference in performances by classes between the ResNet50 applied on the original image and the same network applied on the reconstructed image supersampled with DarioNet*

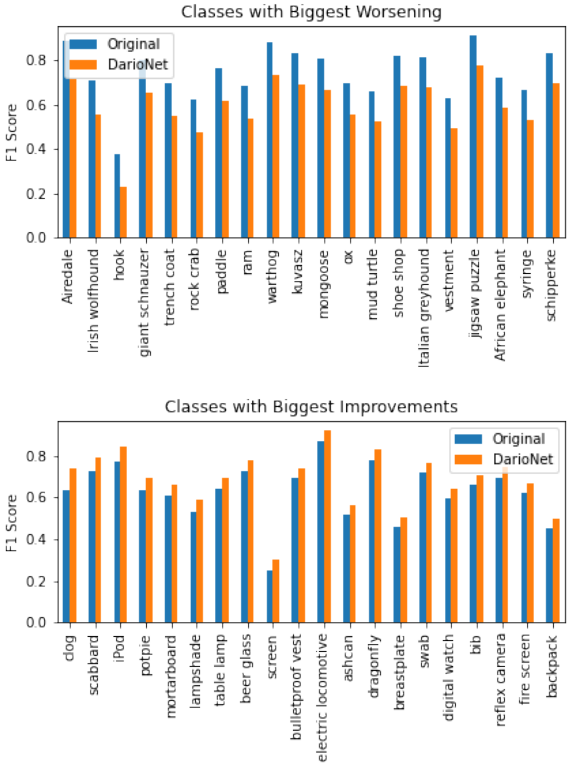


Figure 5: *At the left the classes with the biggest worsening in performances, at the right the one with the biggest improvements.*

References

[1] Wang et al., *ESRGAN: Enhanced super-resolution generative adversarial networks*, 2018.

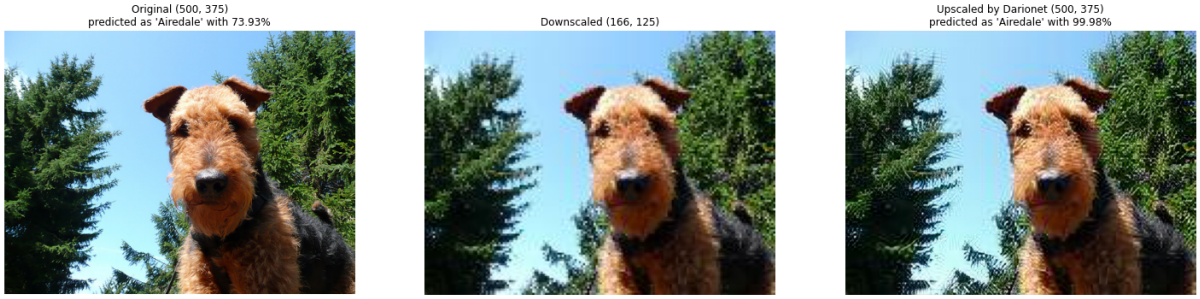
[2] Russakovsky et al., *ImageNet2012 dataset*, 2014.

[3] R. Timofte et al., *DIV2K dataset*, 2017.

Appendix

Here you can find some examples of reconstructions using Darionet.
At the left is shown the original image in the center the downsampled and at the right the reconstructions.
First are going to be shown images belonging to the classes where we had worsening in the reconstructed images (Figure 6), then are going to be presented images from classes where we had an improvement in the classification (Figure 7); according to the rankings shown in figure 5

Airedale, Airedale terrier



Irish wolfhound

hook, claw

Figure 6: classes for which we had the major worsening

clog, geta, patten, sabot

scabbard

iPod

Figure 7: classes for which we had the major improvements