02465 DEEP LEARNING, DTU COMPUTE, FALL 2020

IMAGE SUPER-RESOLUTION USING HIERARCHICAL VAES

Paolo Zifferero (s200149), Prasad Jagtap (s200109), Paul Connetable (s151404)

Source code: *github.com/prasad-jagtap/02456-deep-learning*

ABSTRACT

We propose a method for image super-resolution (SR), where we learn an end-to-end mapping between the low and high-resolution images. The mapping is carried out using CNN & VAEs separately. The framework represented as a deep convolutional neural network (CNN) takes the low-resolution image as the input and outputs the high-resolution one. But unlike traditional methods that handle each component separately, our method jointly optimizes all layers. Our CNN model demonstrates state-of-the-art restoration quality and achieves practical on-line usage. The behavior of optimization-based super-resolution methods is principally driven by the choice of the objective function. We explore different network structures and parameter settings to achieve trade-offs between performance and speed. We believe that variational auto-encoders (VAEs) are the powerful framework for unsupervised learning, however, work in the past has been restricted to the models with one or two layers of fully factorized latent variables, limiting the flexibility of the latent representation. We consider believing our deep residual network might be able to recover photo-realistic textures from the down-sampled images on public benchmarks. The project code is available at `https://github.com/prasad-jagtap/02456-deep-learning`

***Index Terms*—** Super-resolution, CNN, VAE

1. INTRODUCTION

Increasing an image resolution or reconstructing a high-resolution image from a coarse resolution image is a very active area of research. This is called *image super-resolution*, and can be used to improve the image quality of movies or pictures, as well as store them in coarser resolution resulting to accommodating lesser space. This challenge was carried out at the beginning by *bicubic interpolation* or *sparse-coding-based* methods. Quickly, Deep Learning models proved their efficiency at solving this ill-posed problem. At first, Convolutional Neural Networks were proposed to tackle this problem [1], by letting them learn the features of the image, and reconstruct them in higher resolution. Later, the use of VAEs was proposed to improve quality of the reconstructed images [2]. GANs have also been used to improve the perceptual quality of the images, to make them look more realistic [3]. The latest state of the art super-resolution neural network [4] is a deep hierarchical VAE. The more recent models like the first ones cited above have proven to be much more adequate to the task, and have also been able to generate new images, separate an image's content & style, and use them separately as in google's deep-dream.

The preference lies around leveraging the CIFAR-10 data-set for the project, which consists in a collection of 32 by 32 RGB images. The CIFAR-10 photo classification problem is a standard data-set used in deep learning, although, the data-set is effectively solved, it can be used as the basis for learning and practicing to develop, evaluate, and use convolutional deep learning neural networks for image classification from scratch. This includes developing a robust test harness for estimating the performance of the model, exploring improvements to the model, saving the model and later load it to make predictions on available data.

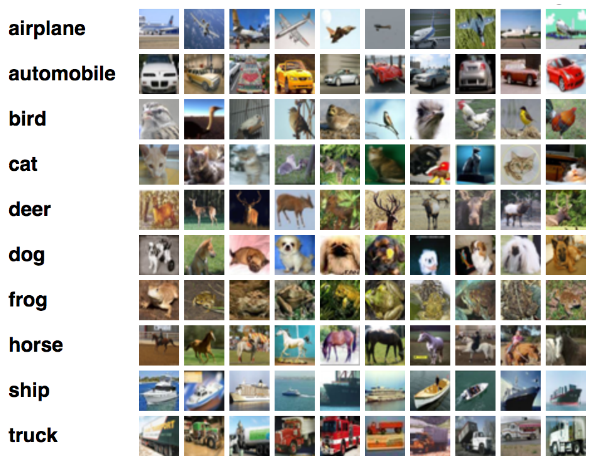


Fig. 1. Example of a CIFAR-10 dataset

In order to train the networks on Super Resolution, the original images has to be treated as high-resolution image targets. Later create the low-resolution images by applying a *Gaussian filter* and *sub-sampling* the high-resolution images, to obtain 8 by 8 RGB images. The *Gaussian filter* is deemed particularly important as it allows to preserve the information of a pixel by spreading it to the neighboring ones. Hence, while performing down-sampling, no information loss would be recorded.

2. BACKGROUND

2.1. Image Super-Resolution

Regularly, single image super-resolution algorithms can be categorized into four types – prediction models, edge based methods, image statistical methods and patch based (example-based) methods [1]. It is a hypothesis that patch based method achieve optimum performance.

The internal patch based method exploit the self similarity characteristics and generate flawless patches from the inserted images. The dictionaries are directly presented as low/high-resolution patch pairs, and the nearest neighbour of the input patch is found in the low-resolution space, with its corresponding high-resolution patch used for reconstruction. Other mapping functions such as simple function, kernel regression, random forest and anchored neighborhood regression are proposed to further improve the mapping accuracy and speed.

2.2. Deep Learning for Image Restoration

There have a few studies of using deep learning techniques for image restoration. Mostly, the convolutional neural network is applied for removing noisy patterns, while proposed to embed auto-encoder networks in their super-resolution pipeline under the notion internal example-based (patch-based) approach [5]. The corresponding deep model is however, not designed to be an end-to-end elucidation, each layer of the cascade requires individual optimization of the self-similarity search process and the auto-encoder. Whereas, on the other hand, the Super Resolution-CNN optimizes an end-to-end mapping, as well, the Super Resolution-CNN is faster when it comes to speed. It is not only a best practical method but quantitatively superior method [1].

2.3. Variational Auto-encoders (VAE)

The framework of variational auto-encoders (VAEs) provides a principled method for jointly learning latent-variable models and corresponding inference models. However, the main drawback of this approach is the blurriness of the generated images. Some studies link this effect to the objective function, the log-likelihood. Further studies have been produced to enhance VAEs by adding a random variable that is a down-scaled version of the original image and still use the log-likelihood function as the learning objective. Further, by providing the down-scaled image as an input to the decoder, it can be used in a manner similar to the super-resolution. The VAE with the bijective prior showcases an excellent performance on the natural image reconstruction task, which is contrary to the performance that is often provided in the literature [6].

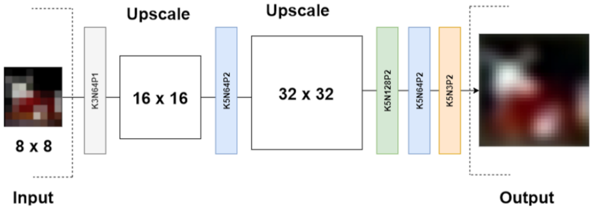
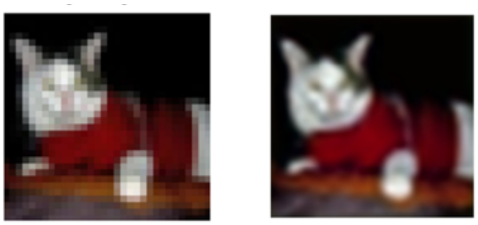


Fig. 2. Structure of the SR convolutional network



(a) LR input image (b) Reconstructed HR image

Fig. 3. Results obtained by feeding a bigger image to the SRCNN.

3. SR WITH A CNN

The first successfully implemented model able to perform SR has been a convolutional neural network. It took inspiration from the work of Dong *et al.* [1], being fully convolutional and having adapted learning rates and activation functions for each layer. The model proposed in this paper, however, performs up-scaling in between convolutions, since the operation allows gradient descent to go through. The input image undergoes a convolution before being up-scaled by a factor of 2, then the same process is repeated once again to bring the image dimensions to the target ones. Three more convolution layers end the model. The structure is shown in figure 2.

The network has been trained on the CIFAR-10 data-set, allowing fast training times. A downside of this data-set is that once its images are down-sampled by a factor of 4, going from 32 by 32 to 8 by 8, a significant amount of information and detail is lost. As a result, the reconstructed images will result often blurry and hardly recognisable. To verify that the network was able to perform image SR properly, once trained, it has been fed with 32 by 32 images that contained much more information to start with. The results, shown in figure 3, prove that the model is extremely valid and can reconstruct features that are hidden in the LR image, as it did with the cat's pupils.

4. SR WITH A VAE

As a first step, before trying more complex networks based on VAEs, a normal VAE was used for Super Resolution. The goal of a VAE is to shrink input data into a latent space representation, of a lower dimension, through an encoder. A decoder performs the wanted application based on the latent space representation. It could be a classification, or an input reconstruction for example. The different observations (i.e. the latent space, and the observation) are represented in a probabilistic framework. If we denote $\mathbf{x}$ the input data, $\mathbf{z}$ the latent space, the VAE learns through a reparameterization trick, the best approximate posterior model $q_\phi(\mathbf{z}|\mathbf{x})$ and reconstruction model $p_\theta(\mathbf{x}|\mathbf{z})$ models, based on a prior $p_\theta(\mathbf{z})$. In practice, it is only possible to compute an Evidence Lower Bound (ELBO) as a loss to the network. We used a modified ELBO loss in this work defined by;

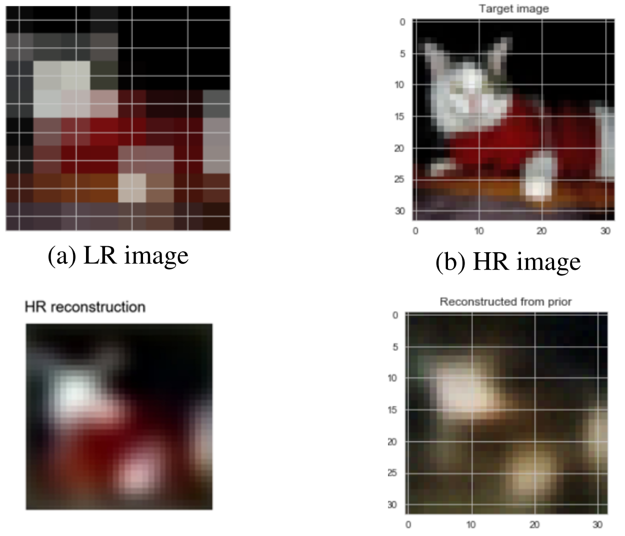
$$\mathcal{L}^{\beta}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[log p_\theta(\mathbf{x}|\mathbf{z})] \; - \; \beta \, \mathcal{D}_{KL}(q_\phi(\mathbf{z}|\mathbf{x})|p(\mathbf{z})) \tag{1}$$

In an SR problem, the goal is to increase the resolution of an image, and in this case, the encoder is used to lower the size representation of a high resolution image to the smaller latent space, and the decoder must try to reconstruct as best as possible the input image. While this is used to train the network's encoder and decoder, only the low resolution image is available in real case scenarios. The corresponding LR image is therefore used as a prior to the VAE's latent space. Our VAE's architecture is shown in figure 4. The encoder shrinks the HR 32 by 32 RBG image into our latent space, which is a 192 long vector. The prior used is the low resolution 8 by 8 RGB image, which is reduced to one dimension and transformed through one linear layer into the latent space. The selected length of the latent space corresponds to the number of information in the prior. A normal distribution with a fixed low standard deviation (at 0.01) has been chosen for the observation model. The posterior distribution (in the latent space) and the prior are also represented by two normal distributions. The decoder has a very close structure to the CNN shown above.

The VAE results shown below have been obtained with a batch size of 16, and $\beta = 0.5$. Similar results have been obtained with higher β values, but required an increased batch size for training. The optimizer used is an Adam optimizer with a learning rate of 10^{-3}, and the loss used is the opposite of the beta ELBO mean.

5. RESULTS

The VAE was significantly longer to train than the CNN, and needed several epochs to train properly on the data-set. Figure 5 shows the different metrics we used to monitor the network training, over 50 epochs. The training values are displayed in blue, and the validation values in orange. The trained network was then used with only the prior information to try to reconstruct HR images. The reconstructed cat with a Christmas sweater image by the VAE is presented in figure 6, along side the LR, HR and CNN reconstructed image. A series of other reconstructed images is shown in the appendix 8.



(a) LR image

(b) HR image

(c) CNN reconstruction

(d) VAE reconstruction

Fig. 6. Results obtained on the cat with a Christmas sweater.

The MSE loss on the whole test data for the CNN and the VAE based reconstructions is presented in table 1. It is important to note that neither this value nor the PSNR which is often used to quantify the quality of reconstructed images convey the whole information [3, 4]. It is often preferred to perform tests with volunteers to rate the perceived reconstruction quality [3, 4].

	CNN	VAE
MSE loss	0.0095	0.0121

Table 1. MSE between the HR images and the reconstructed images for the whole test data from the CIFAR-10 data-set.

6. DISCUSSION AND CONCLUSION

In this work, we implemented two different SR algorithms, based on a CNN and on a VAE respectively. We trained and tested them on the CIFAR-10 data-set, and were able to compare them. The CNN was much easier and faster to train. A single epoch was enough to get optimal results, and we only used a couple of epochs to train our best performing CNN. It is also easy to generalize its results to any image size, as

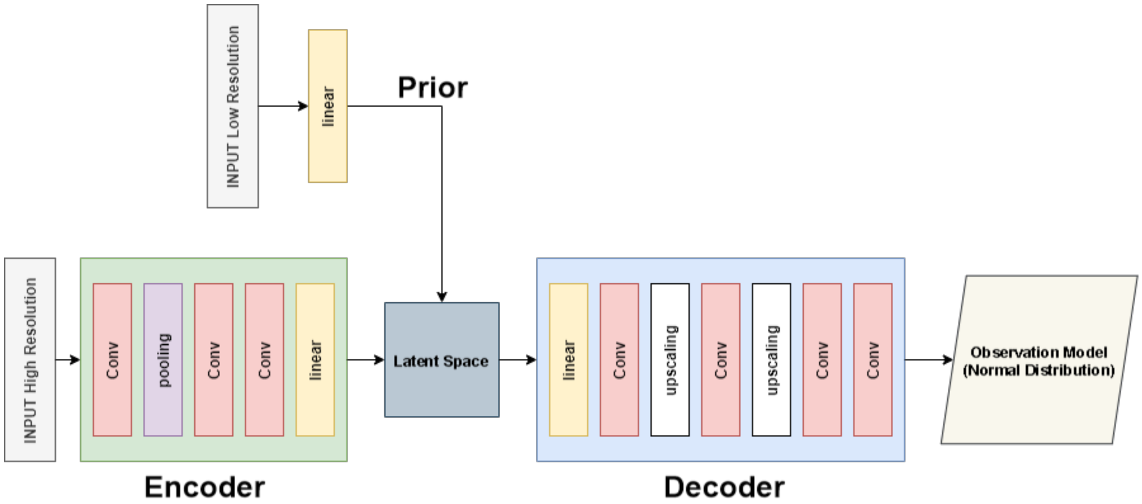


Fig. 4. Our VAE network structure

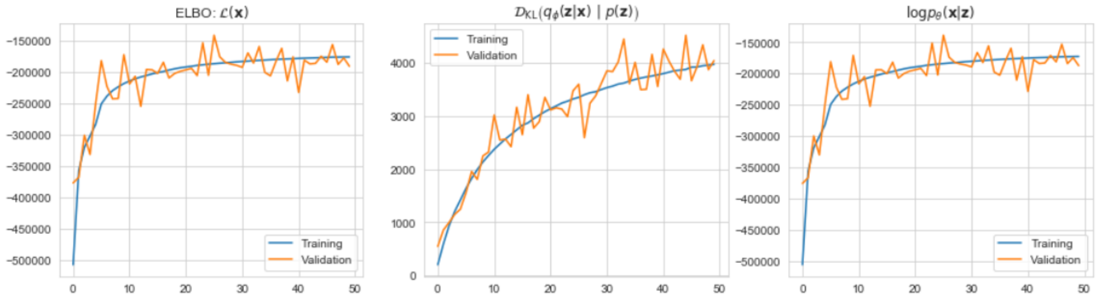


Fig. 5. Evolution of the ELBO, $\mathcal{D}_{KL}(q_\phi(\mathbf{z}|\mathbf{x})|p(\mathbf{z}))$, and $logp_\theta(\mathbf{x}|\mathbf{z})$ on 50 epochs. The blue lines represent the values obtained on the training data and the orange ones the values obtained on the validation data.

it is composed only of convolutionnal and upscaling layers, as shown in figure 3. The results we obtained by applying our CNN on the original HR images to upscale them further are very promising and show that the features learned are still relevant for higher quality images.

The VAE network was much more difficult to train, and needed several dozens of epochs to train. As it is also composed of both an encoder and a decoder, its structure is also larger than the CNN, making it slower to train on an epoch. We also tried a great number of different configurations with different batch sizes, β values, latent space sizes, observation distribution models, learning rates, and changing the loss function of the network by adding a MSE loss part. While this makes VAEs very powerful tools, they also make them more challenging and time-consuming to train. We managed to train a VAE with results close to the CNN when looking at the MSE, while not only using a standard modified ELBO loss.

7. REFERENCES

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, “Image super-resolution using deep convolutional networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[2] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther, “Ladder variational autoencoders,” in *Advances in neural information processing systems*, 2016, pp. 3738–3746.

[3] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

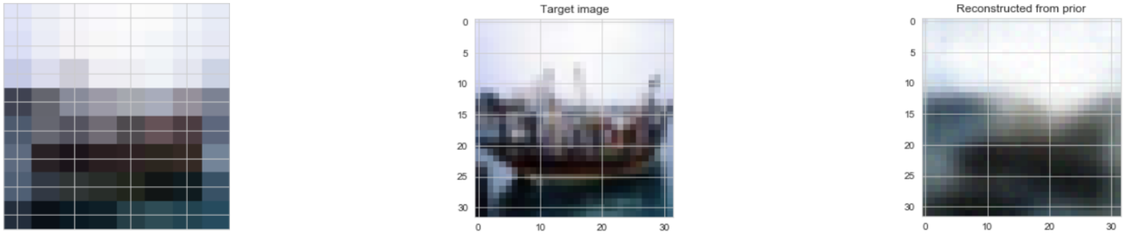
[4] Arash Vahdat and Jan Kautz, “Nvae: A deep hi-

erarchical variational autoencoder," *arXiv preprint arXiv:2007.03898*, 2020.

[5] S. Bagon and D. Glasner, "Super-resolution from a single image," *IEEE International Conference on Computer Vision*, pp. 349–356, 2009.

[6] Ioannis Gatopoulos, Maarten Stol, and Jakub M Tomczak, "Super-resolution variational auto-encoders," *arXiv preprint arXiv:2006.05218*, 2020.

8. APPENDIX

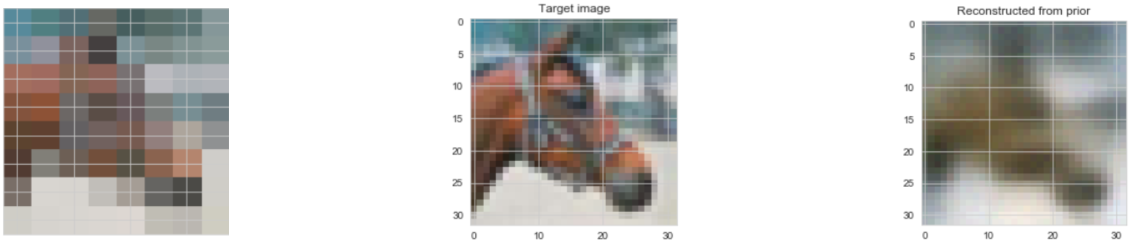


(a) LR image

(b) HR image

(c) VAE reconstruction

Fig. 7. Results obtained on a boat.



(a) LR image

(b) HR image

(c) VAE reconstruction

Fig. 8. Results obtained on a horse head.