
Super Resolution of Diffusion Tensor Images

Yamen Mubarka, Christopher Moffitt, Jayanth Kocherlakota, Yan-Ran Wang, PhD, Haruka Itakura, MD, PhD*
Department of Computer Science
Stanford University
ymubarka@stanford.edu, cmoffitt@stanford.edu, jkoch26@stanford.edu

Abstract

Diffusion Tensor Imaging (DTI) is a valuable diagnostic imaging technique often used to study the human brain. Young children are often unable to remain still for long periods of time as necessary for DTI imaging procedures. We propose a super resolution model capable of transforming a low-resolution (LR) DTI image to a 6-minute high-resolution (HR) image. We obtained DTI brain images of 9 patients from the Oncology and Brain Sciences labs of Stanford Professors Haruka Itakura and Tamar Green. We introduced gaussian noise and downsampled these DTI images to train our Convolutional Neural Network (CNN) model. We chose the ResNet architecture with a General Adversarial Network trained to differentiate between super-resolved images and the original HR images. This architecture demonstrated strong performance, measured by perceptual difference in three public benchmark datasets. We trained our model with slices from 7 patients and tested its performance, measured by peak-signal-to-noise and the Structural Similarity Index Measure, against the Bicubic standard in slices from 2 patients. The SR-GAN underperformed the Bicubic model on quantitative metrics, but recreated images perceptually similar to the original HR image. Further research evaluating alternative structures over larger datasets of DTI images should be performed to determine the optimal super resolution model.

1 Introduction

Diffusion Tensor Imaging is a magnetic resonance imaging technique that measures the diffusion of water in tissue to produce neural tract images. DTI makes it possible to estimate the location and orientation of white matter tracts and is commonly used to study the human brain. We researched Super Resolution (SR) models to improve the quality of 3-minute Low Resolution (LR) DTI images. A successful model would allow patients to be screened more quickly during imaging tests. The required long patient standstill time is an acute concern in pediatric diagnostics.

The input data for our model is a low resolution grayscale DTI image of a human brain. The goal of our super resolution model is to increase the quality and detail of the 3-minute LR images to match that of 6-minute high-resolution (HR) images. In the typical SR framework, the LR image Y is modeled as a downgrade function of X , the HR image and the blurry kernel k and noise term n .

$$y = (x \otimes k) \downarrow_s + n,$$

Neural networks have empirically been successful in high-dimensional image classification problems [1]. We explore various neural network architectures in order to transform the LR DTI image into a

*NOTE: Feel free to reach out to the authors at the email addresses listed above.

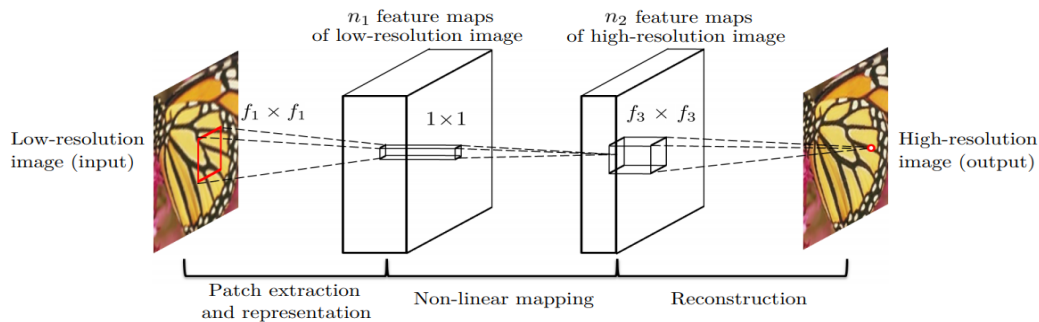
corresponding HR image. The model is trained on a set of (LR, HR) pairs and seeks to minimize the loss function (defined in section 4) between the HR image and the true picture.

2 Related Work

Researchers historically used computationally cheap techniques like regression and interpolation to map LR images to HR images [2]. Interpolation uses known data to estimate values at unknown pixels through techniques such as nearest neighbor or bilinear. Today, mainstream SR algorithms are of three categories: interpolation, reconstruction, and learning. Interpolation methods such as bicubic are computationally cheap but use only local information in the LR image to compute pixel values of the HR image, leading to large bias. Reconstruction methods do not scale well and are computationally expensive. Learning-based models rely on training data and range from Markov random fields to random forest and other methods.

Recent experiments by Dong et al and Wang et al have supported the use of convolutional neural networks (CNN) in single image SR [3, 4]. Dong et al trained the CNN on the same dataset as the traditional regression-based experiments (91 images consisting of 24,800 sub-images) [2] and demonstrated superior results, measured by MSE. The CNN first pre-processes LR images by using bicubic interpolation to increase the number of pixels in the LR image to match the desired size of the HR image. Then, the CNN splits the image into various ‘image patches’, or small overlapping subsets of the larger picture, and stores each patch in a high-dimensional vector. These vectors containing representations of overlapping pixels are the features of the CNN.

Super Resolution using CNN



Source: *Image Super-Resolution Using Deep Convolutional Networks* [3]

The first convolutional layer extracts a set of feature maps for each patch. Each additional hidden layer of the neural network maps a high-dimensional vector onto another high-dimensional vector through non-linear activation functions such as ReLu. The final layer of the network aggregates the patch vectors and reconstructs the high-resolution image.

CNNs share parameters and have few connections, reducing the parameters required to be trained; imaging data is usually high dimensional, so weight matrices in each network layer have a high number of parameters to train. Increasing the layers (depth) of a CNN improves its performance because early network layers detect edges while later layers detect entire objects. Experiments by Krizhevsky et al have demonstrated the efficacy of deep CNNs in image classification competitions, with regularization techniques such as dropout reducing overfitting and test error [5].

Though conventional techniques like batch-normalization can speed the training process, deep neural networks with many layers experience the ‘vanishing gradient’ problem, since the gradients of early network layers are the product of many partial derivatives of activation functions by the chain rule. He et al propose a Residual block architecture with skipped connections performing identity mappings, increasing the value of partial derivatives [6]. The architecture supports up to 152 network layers and achieves lower error than traditional CNNs on the ImageNet test set, used in image classification competitions. Given the strong performance of CNNs in image classification and the advantages of deeper networks, we chose to implement the ResNet architecture.

Most experiments use MSE, defined below, as the loss function and use stochastic gradient descent with standard backpropagation to train network weights [3].

$$L(\Theta) = \frac{1}{n} \sum_{i=1}^n \|F(\mathbf{Y}_i; \Theta) - \mathbf{X}_i\|^2,$$

However, MSE is based on pixel-wise image differences, and may not capture all perceptual differences. Many experiments struggled with high upscaling factors and had poor texture detail in reconstructed images [7, 8]. Therefore we used a generative adversarial network (GAN) and also examined the SSIM metric. The GAN is trained to differentiate between the super-resolved images and the original high-resolution images. The generator reproduces artificial LR images, and the network layers are applied to an upsampled version of the LR image.

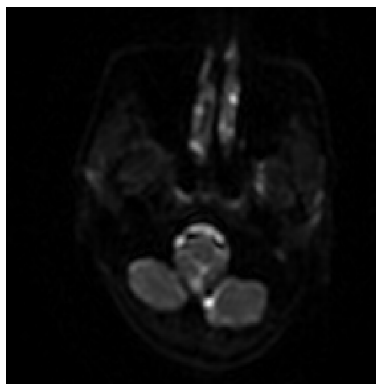
Since the discriminator must classify between real and fake images, the model can generate solutions more similar to real images. Ledig et al show a significant increase in perceptual quality using SRGAN with Mean-Opinion Score tests from three public benchmark datasets. We chose SRGAN because its test scores are closer to the HR images than those of any other method [9].

3 Dataset and Features

We obtained the dataset from the Oncology and Brain Sciences Lab at Stanford and worked with the Laboratory of Quantitative Imaging and Artificial Intelligence at Stanford University [10]. To access patient medical images, we completed CITI Training required by Stanford School of Medicine and received IRB Approval to conduct our research.

We received 6 minute HR images for 9 patients, with 1925-1960 image slices per patient (1 patient had 868 slices). The images are grayscale pictures of the human brain. Each image slice has dimensions of 256 x 256, is 130 KB (total dataset is 4.5 GB) and represents a different layer of the brain; the input image z axis represents the number of slices, rather than RGB values. We did not receive 3 minute DTI images representing the LR element of the aforementioned (LR, HR) pair so we ‘reverse-labeled’ our data to train our model.

DTI Slice: Single Data Sample



Source: Laboratory of Quantitative Imaging and Artificial Intelligence at Stanford University[10]

To create LR images paired with the original, we first convert the DICOM, or Digital Imaging and Communications in Medicine images to Numpy arrays. DICOM is the standard for storing and transmitting medical images of any kind. We represent HR images as numpy arrays with shape [256, 256, 3] by normalizing / rescaling and creating three channels from the grayscale image (for each individual slice). Then we introduce gaussian noise and downsample by a factor of 4 to create LR images. The numpy arrays for the newly created paired LR images and HR images are stored as portable network graphics (PNG) files on our local computers. After preprocessing the data, we loaded it onto AWS. Since images are de-identified, we use Amazon’s EB2 for storage.

We trained our final model with 7 patients, with 2 patients in the development / test set. We did not create a separate dev and test set; our goal was to maximize the size of the training set, while evaluating our model on multiple patients.

4 Methods

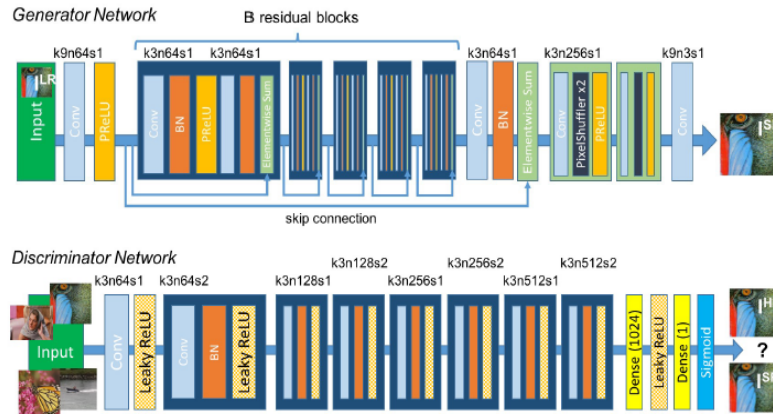
We use a 16 block ResNet architecture with a skip-connection and replace MSE-based loss with loss from pre-trained VGG network with features extracted. We define a discriminator network which is alternatively optimized along with the generator network. We train the generative model G to fool a discriminator model D trained to distinguish between super-resolved images and real images. Our goal is for the generator to create images that are perceptually indistinguishable from real images. Per Goodfellow et al we optimize D and G in an alternating manner [11].

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I^{HR} \sim p_{\text{train}}(I^{HR})} [\log D_{\theta_D}(I^{HR})] + \mathbb{E}_{I^{LR} \sim p_G(I^{LR})} [\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))]$$

We use the architecture proposed by Ledig et al [9], in which each block includes two convolutional layers with 3x3 kernels and 64 feature maps followed by batch-normalization and ParametricRelu as the activating function, as initially proposed by Gross and Wilber [12]. The discriminator trained to solve the above equation. It has 8 convolutional layers with many filters, similar to the VGG network. The feature maps are followed by a final sigmoid activation function to determine the probability that the given image is real.

The GAN architecture is implemented in Tensorflow and Tensorlayer. We found an online GitHub repository which implemented the SR-GAN model on natural images and used this as the baseline for our DTI SR model[13]. We trained our model on AWS, given the workload is compute-intensive.

GAN Architecture



Source: Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network [9]

Metrics

We will be looking at two metrics: 1. Peak Signal-to-Noise Ratio (PSNR) and 2. Structural Similarity index (SSIM). These metrics are widely used in image reconstruction techniques in the medical field. PSNR is the ratio between the maximum value of a signal and the MSE. It is used as a pixel-by-pixel comparison which works well but is very local and does not take global visual differences into account. SSIM looks at structure similarities between images. We use the below equation for PSNR:

$$\text{PSNR} = 20 * \log_{10} \frac{\text{MAX}}{\text{MSE}^{1/2}}$$

However, two distorted images with the same MSE may have very different levels of perceptual similarity to the original picture. SSIM is based on the assumption the human visual system is able to extract structural information. It compares local patterns of pixel intensities normalized for luminance and contrast, since the structures of objects should be independent of both factors. We use the same method as Wang et al and separate the similarity measurements into luminance contrast and structure to create the below structural similarity index [7].

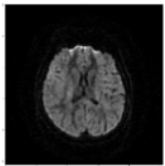
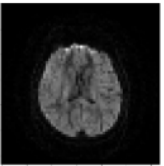
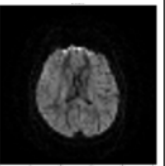
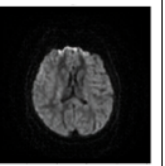
$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

5 Experiments/Results/Discussion

We trained our SR-GAN model with slices from 7 patients with the following hyperparameters. Training took about a day to reach 7 epochs. We slightly reduced the number of epochs because the training time was very high. After training, we then tested the model on the 2 patients in our test set, including 3,920 image slices. We compared our results against the well-known bicubic super-resolution model. The results are the following:

	PSNR	SSIM
SR-GAN Model	33.838600	0.783874
Bicubic Benchmark	34.811728	0.889286
Select SR-GAN Hyperparameters		
Batch Size		8
Learning Rate		1E-04
Number of Epochs		7

Below is a sample of our results looking at a single DTI slice.

Original HR slice (label)	Synthesized LR slice (input)	Bicubic benchmark	Generated Image from SR-GAN Model
			
		PSNR: 35.60127 SSIM: 0.901932	PSNR: 34.67988 SSIM: 0.82892

As evidenced by the sample images and metrics data, our SR-GAN model comes very close to approaching, but does not yet surpass the Bicubic model, based on the mean PSNR and SSIM. The mean is calculated across all slices from the 2 patients in the test set. Interestingly however, our SRGAN model surpasses the Bicubic model qualitatively in its apparent structure and definition.

We also trained our SR-GAN model using the DIV2K natural images and applied this model with different weights to our DTI images. This model achieved a mean PSNR of 33.785 and SSIM of 0.889, better than our SR-GAN model trained with DTI images. This suggests we need to increase the size of our training set or train for more epochs. We should also consider alternative metrics beyond PSNR and SSIM because they do not reflect the complexity perceptual difference.

6 Conclusion/Future Work

Further research evaluating alternative structures over larger datasets of DTI images should be performed to determine the optimal super resolution model. Also, data augmentation techniques such as random crops, mirroring, or rotation could be explored to increase the samples in the training set.

7 Contributions

Yamen Mubarka - Primary person responsible for training model, including tuning hyperparameters and deciding optimal performance metrics. Yamen extensively debugged code for multiple model candidates on both AWS and his local machine. Yamen introduced Git as a collaboration method to track group changes.

Christopher Moffit - Primary person responsible for pre-processing data, including transforming the DICOM data into Numpy Arrays, and increasing the number of channels for the SR-GAN Model. Chris modularized such steps in Jupyter notebooks, allowing code to be easily re-used. Chris also assisted with training the model on his local machine.

Jayanth Kocherlakota - Primary person responsible for literature review, paper write-up, video, and dividing team responsibilities. Jay considered alternative model architectures in prior research. Jay also assisted with setting up AWS infrastructure and helped Chris with pre-processing data.

References

- [1] W. Yang , X. Zhang , Y. Tian , W. Wang , J.-H. Xue , Q. Liao , Deep learning for single image super-resolution: a brief review, *IEEE Trans. Multimedia* (2019) (early access)
- [2] Timofte, R., De Smet, V., Van Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: *ICCV*, pp. 1920–1927 (2013)
- [3] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014*, pages 184–199. Springer, 2014
- [4] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang. Deep networks for image super-resolution with sparse prior. In *IEEE International Conference on Computer Vision (ICCV)*, pages 370–378, 2015.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.
- [6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3, 4
- [7] Wang, Z.; Bovik, A.C.; Rahim Sheikh, H.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* 2004, 13
- [8] P. Gupta, P. Srivastava, S. Bhardwaj, and V. Bhateja. A modified psnr metric based on hvs for quality assessment of color images. In *IEEE International Conference on Communication and Industrial Application (ICCIA)*, pages 1–4, 2011.
- [9] Ledig, Christian et al. “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network.” 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017): 105-114.
- [10] Stanford Oncology and Brain Sciences Lab
- [11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014. 3, 4, 6
- [12] S. Gross and M. Wilber. Training and investigating residual nets, online at <http://torch.ch/blog/2016/02/04/resnets.html>. 2016. 4
- [13] @articletensorlayer2017, author = Dong, Hao and Supratak, Akara and Mai, Luo and Liu, Fangde and Oehmichen, Axel and Yu, Simiao and Guo, Yike, journal = *ACM Multimedia*, title = *TensorLayer: A Versatile Library for Efficient Deep Learning Development*, url = <http://tensorlayer.org>, year = 2017
- [14] Z. Lu, Y. Chen. Single Image Super Resolution based on a Modified U-net with Mixed Gradient Loss. *DeepAi*.