DEEP RESIDUAL LEARNING FOR COMPRESSED SENSING MRI

Dongwook Lee, Jaejun Yoo and Jong Chul Ye

Bio Imaging and Signal Processing Lab., Dep. of Bio and Brain Engineering, KAIST

ABSTRACT

Compressed sensing (CS) enables significant reduction of MR acquisition time with performance guarantee. However, computational complexity of CS is usually expensive. To address this, here we propose a novel *deep residual learning* algorithm to reconstruct MR images from sparsely sampled k-space data. In particular, based on the observation that coherent aliasing artifacts from downsampled data has topologically simpler structure than the original image data, we formulate a CS problem as a residual regression problem and propose a deep convolutional neural network (CNN) to learn the aliasing artifacts. Experimental results using single channel and multi channel MR data demonstrate that the proposed deep residual learning outperforms the existing CS and parallel imaging algorithms. Moreover, the computational time is faster in several orders of magnitude.

Index Terms— Compressed sensing MRI, deep learning, residual learning, CNN

1. INTRODUCTION

In MR acquisition, efficient acceleration scheme is important. Nowadays, parallel MR imaging and compressed sensing MRI (CS-MRI) are the two most important tools in accelerating the speed of MR acquisition. Generalized Autocalibrating Partially Parallal Acquisitions (GRAPPA) is a representative parallel technique that interpolates the missing k-space data by exploiting the diversity of coil sensitivity maps. CS-MRI reconstructs high resolution image from randomly sampled k-space data by exploiting the sparsity of the data in transformed domain. CS algorithms are commonly formulated as a penalized inverse problems that minimize the tradeoff between data fidelity term in k-space and sparsity penalty in the image domain. In recent annihilating filter-based low-rank Hankel matrix approach (ALOHA) [1], CS-MRI and parallel MRI can be unified as k-space interpolation [1]. One of the limitations of these CS algorithms is, however, that the computational complexity is usually high.

Recently, deep learning using CNNs have achieved tremendous success in classification problems [2] as well as regression problems [3]. The exponential expressivity under a given network complexity (e.g. VC-dimension or Rademacher complexity [4]) has been attributed to its success [5, 6].

Wang et al [7] was the first to apply deep learning to CS-MRI. They trained the deep neural network from the downsampled reconstruction images to learn fully sampled reconstruction. Then, they used the deep learning output either as an initialization or a regularization term in classical CS approaches. Deep network architecture using unfolded iterative CS algorithm was also proposed [8]. Rather than using handcrafted regularizers, the authors in [8] tried to learn a set of optimal regularzers using a reaction diffusion model.

Unlike these existing deep learning researches for CS-MRI, this paper is mainly interested in deep *residual* learning [9]. As shown in Fig. 1, the main idea is to learn aliasing artifacts rather than the aliasing-free fully sampled reconstruction. Once the aliasing artfacts are estimated, aliasing-free image is then obtained by subtracting the estimated aliasing artifact. This network architecture is proposed based on our conjecture that aliasing artifacts from uniformly undersampled patterns may have simpler topological structure such that learning residual is easier than learning the original aliasing free images. Using the persistent homology analysis from computational topology [10], we show that this conjecture is true. Accordingly, we investigate several architectures of residual learning and show that deconvolution network with contracting path - which is often called U-net structure [11] for image segmentation - was the most effective in estimating the aliasing artifacts. Experimental results using single and multi channel CS-MRI show significant improvement in performance over the existing state-of-the-art MR reconstruction algorithms.

2. THEORY

2.1. Generalization bound for residual learning

As shown in Fig. 1, based on an input $X \in \mathcal{X}$ and a label $Y \in \mathcal{Y}$ generated by a distribution D, we are interested in estimating a regression function $f : X \to Y$ in a functional space $\mathcal{F}$ that minimizes the risk $L(f) = E_D\|Y - f(X)\|^2$. A major technical issue is, however, that the associated probability distribution D is unknown. Moreover, we only have access to a finite sequence of independent, identically distributed training data

$$S = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$$

Grant sponsor: Korea Science and Engineering Foundation, Grant number NRF-2016R1A2B3008104.

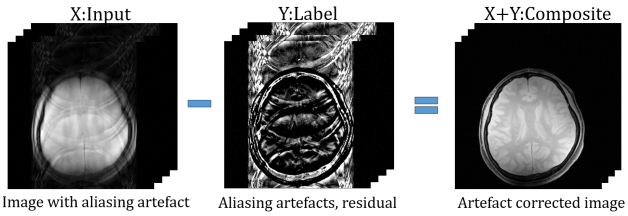

Fig. 1. Residual learning of aliasing artifacts.

such that only an empirical risk $\hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \|Y_i - f(X_i)\|^2$ is available. Direct minimization of empirical risk is, however, problematic due to the potential of overfitting.

To address this issues, statistical learning theory has been developed to bound the risk of a learning algorithm in terms of complexity measures (eg. VC dimension and shatter coefficient) and the empirical risk. Rademacher complexity [4] is one of the most modern notions of complexity that is distribution dependent and defined for any class real-valued functions. Specifically, with probability $\geq 1-\delta$, for every function $f \in \mathcal{F}$,

$$L(f) \leq \underbrace{\hat{L}_n(f)}_{\text{empirical risk}} + \underbrace{2\hat{R}_n(\mathcal{F})}_{\text{complexity penalty}} + 3\sqrt{\frac{\ln(2/\delta)}{n}} \quad (1)$$

where the empirical Rademacher complexity $\hat{R}_n(\mathcal{F})$ is defined to be

$$\hat{R}_n(\mathcal{F}) = E_\sigma \left[\sum_{f\in\mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right) \right],$$

where $\sigma_1, \cdots, \sigma_n$ are independent random variable uniformly chosen from $\{-1, 1\}$. Therefore, to reduce the risk, we need to minimize both the empirical risk (i.e. data fidelity) and the complexity terms in (1) simultaneously. In neural network, the value of empirical risk is determined by the representation power of network [5], whereas the complexity term is determined by the structure of a network [4].

2.2. Topological structure of residual manifold

There have been several studies explaining the benefit of depth in neural networks [5, 6]. In deep networks, the representation power grows exponentially with respect to the number of layers while it grows at most polynomially in shallow ones [5]. Therefore, with the same number of resources, theoretical results supports that deep architecture is prefered to shallow one, and it increases the performance of the network by reducing the empirical risk in (1). Thus, if the manifold of label Y is simple enough to meet the representation power of a given network architecture, then the empirical risk as well as the risk upper bound can be reduced.

The complexity of a manifold is a topological concept, so it should be analyzed using topological tools. Here, we employed the recent computational topology tool called *persistent homology* [10]. Specifically, as shown in Fig.2(a), we

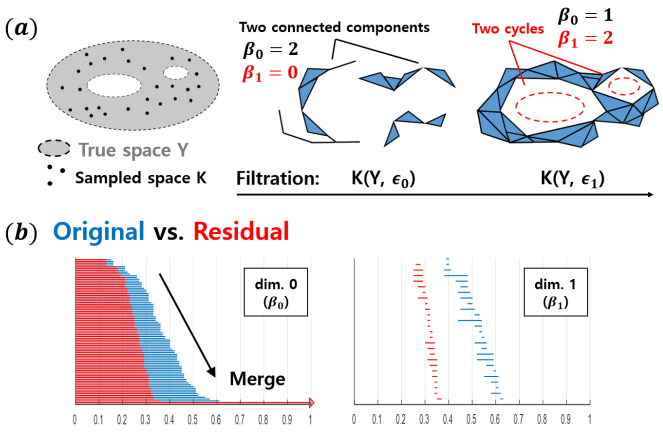

Fig. 2. (a) Point cloud data K of true space Y and its configuration over ϵ-distance filtration. (b) Zero and one dimensional barcodes of the original (blue) and residual (red) of one channel MR data. Similar results were shown in four channel MRI data.

can infer the topology of the data by varying the distance measure ϵ. As allowable distance ϵ increases, point clouds merge together and finally become a single cluster. During this filtration process [10], Betti numbers (β_m) are calculated which are the numbers of m-dimensional holes of a manifold. Specifically, β_0 and β_1 are the number of connected components and cycles, respectively. Therefore, the point clouds with high diversity will merge slowly, which can be reflected as a slow decrease in Betti numbers. This trend can be illustrated using so-called *barcode* of Betti numbers [10], and persistent homology investigates the topology of a manifold using barcodes.

To compare the topological structure of residual (aliasing artifact) and the original image space, we calculated Betti numbers using a toolbox called JAVAPLEX (http://appliedtopology.github.io/javaplex) to calculate Betti numbers. To generate a point cloud, each label image (either aliasing artifact image or aliasing-free MR image from fully sampled data) is regarded as a point in a high dimensional space, and then we calculated Euclidean distance between each point and normalized the value by the maximum distance. The topological complexity of the original and residual image spaces were compared by the change of Betti numbers in Fig. 2(b). Indeed, β_0 and β_1 of residual image manifold decreased faster to a single cluster, which informs that the residual image manifold has a simpler topology than the original one. In experimental results in Section 5, we confirmed that the prediction by the Betti number fully reflects reconstruction performance.

3. RESIDUAL LEARNING ARCHITECTURE

To construct a residual learning architecture, we utilize the convolution, batch normalization, rectified linear unit (ReLU), and contracting path connection with concatenation [9, 11]. Specifically, let x^l denote the l-th layer input and

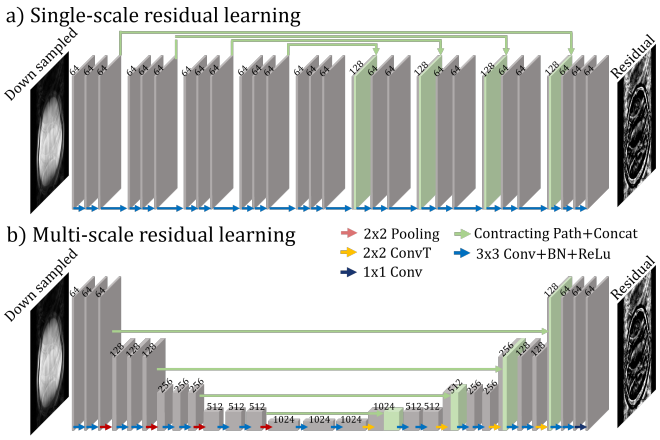

Fig. 3. (a) Single scale residual learning with a modified deconvolution network framework[12] with symmetric contracting path. (b) Multi scale residual learning with U-net architecture.

w^l, b^l represent weights and bias of l-th convolution layer, respectively. Then, the l-th layer of the network performs following operation, repeatively, $f^l(x^l) := \sigma(BN(w^l * x^l + b^l))$, where $\sigma(\cdot)$ is the ReLu function, and BN is a batch normalization. In contracting path, we concatenate x_l along channel dimension [11].

Fig. 3 illustrates the several network configuration we have investigated for residual learning. Fig. 3(a) is single scale residual learning with a modified deconvolution network framework[12]. Fig. 3(b) is a multi-scale residual learning with additional pooling and unpooling(conv transpose) layers on top of Fig. 3(a). To make the number of network features similar to Fig. 3(a), the number of channels are doubled after the pooling layer. This architecture is often called the U-net [11]. In the following, we investigate the performance of each network configuration for residual learning.

4. MATERIALS AND METHODS

4.1. MR dataset

We used brain MR image dataset, which consists of total 81 axial brain images from 9 subjects. The data were acquired in cartesian coordinate with a 3T MR scanner that has four Rx coils (Siemens, Verio). The following parameters were used for SE and GRE scans: TR 3000-4000ms, TE 4-20ms, slice thickness 5mm, 256× 256 acquisition matrix, 4 coils, FOV 240× 240, FA 90 degrees. For the dataset, we split the training and test data set by randomly choosing about 80% of total images for training and about 20% for testing. For data augmentation, we generated 32 times more training samples by rotating, shearing and flipping the images. For single channel experiments, we chose 1 channel data from the four channel data.

4.2. Network training

The original k-spaces were retrospectively downsampled by 4 times with 13 ACS (autocalibration signal, 5 percents of total PE) lines in the k-space center. As shown in Fig. 1, the residual is constructed as the difference between the reconstruction images from fully sampled data and the down-sampled data. During the training, residual images were used as labels (Y) whereas the aliased images from down-sample data were used as input (X). Since MR images are complex valued and standard CNNs are real-valued, we trained the two residual networks: one for the magnitude and the other for the phase of the images. Both networks have the same residual learning structure (however, due to the page limit, we only show the magnitude results).

The network was implemented using MatConvNet toolbox(ver.20, http://www.vlfeat.org/matconvnet/) in MATLAB 2015a environment (Mathworks, Natick). We used a GTX 1080 graphic processor and i7-4770 CPU (3.40GHz). The weights of convolutional layers were initialized by Gaussian random distribution with Xavier method to achieve proper scale. This helped to prevent the signal from exploding or vanishing in the early stage of learning. The stochastic gradient descent (SGD) method with momentum was used to train the weights of the network and minimized the loss function. It took about 9 hours for training the network.

To verify the performance of the network, for multi-channel dataset, we compared the reconstruction results with those of GRAPPA. We also compared the ALOHA [1] reconstruction as the state-of-the-art CS algorithm for both single and 4-channel reconstructions.

5. RESULTS

In single channel experiment with x4 acceleration (Fig. 4(b)), there exists significant amount of aliasing artifacts from the zero-filled reconstruction. Moreover, due to coherent aliasing artifacts from uniform downsampling, most of the existing CS algorithm failed and only ALOHA was somewhat successful with slightly remaining aliasing artifacts. However, the residual learning results clearly showed very accurate reconstruction visually and quantitatively by removing the coherent aliasing artifacts. In four channel parallel imaging experiments in Fig. 4(c), GRAPPA shows the strong aliasing artifacts due to the insufficient number of coils and ACS lines. ALOHA reconstruction could remove most of the aliasing artifacts, but the results were not perfect due to the coherent sampling. However, the proposed method provided near perfect reconstruction.

In Fig. 5(a)(b), the convergent plots for a test data set from single and four channel reconstruction are illustrated. Among the various residual learning architectures, the multi-scale residual learning (Fig. 3(b)) provided the best reconstruction results. Moreover, residual learning was significantly better than direct image learning with the same U-net architecture as shown in Fig. 5.

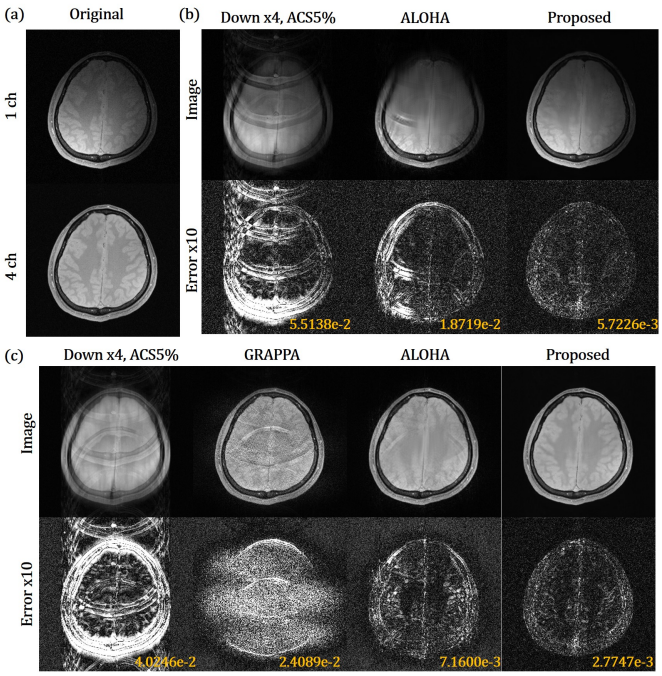

Fig. 4. (a) Original image: (top) single channel, (bottom) 4-channel. (b) Single channel reconstruction results at x4 acceleration. (c) 4 channel reconstruction at x4 acceleration. The resulting normalized mean square error(NMSE) is displayed at the bottom of each figure.

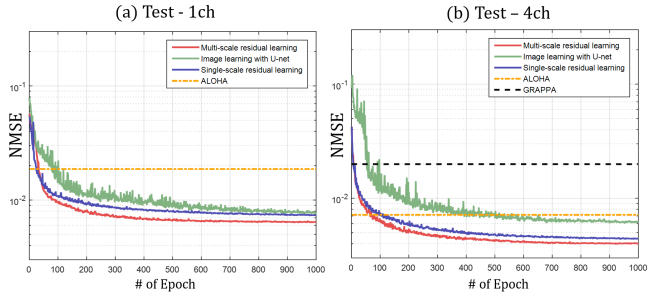

Fig. 5. NMSE convergence graph for test data. (a) single channel reconstruction, (b) four channel reconstruction.

The reconstruction time of GRAPPA was about 30 seconds for multi channel image under the aforementioned hardware setting. The reconstruction time for ALOHA was about 10 min for four channel and about 2 min for single channel data. The proposed network only took less than 41 ms for multi channel image and about 30 ms for single channel image.

6. DISCUSSIONS AND CONCLUSIONS

In this article, we have presented an accelerated MRI reconstruction method from uniform downsampled MR brain data using residual learning. By learning aliasing artifacts that have simpler topology, the resulting risk of the proposed residual learning can be reduced so that we can obtain more accurate MR images. The proposed method works on not only multi channel data but also single channel data. Even with severe coherent aliasing artifacts, the proposed residual learning successfully learns the aliasing artifacts, whereas the existing parallel and CS reconstruction fails. Moreover, compared to existing algorithms which need heavy computational cost, the proposed network produces the results instantly.

7. REFERENCES

[1] Kyong Hwan Jin, Dongwook Lee, and Jong Chul Ye, "A general framework for compressed sensing and parallel MRI using annihilating filter based low-rank hankel matrix," *IEEE Trans. on Computational Imaging*, (in press), 2016.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[3] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *arXiv preprint arXiv:1608.03981*, 2016.

[4] Peter L Bartlett and Shahar Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

[5] Matus Telgarsky, "Benefits of depth in neural networks," *arXiv preprint arXiv:1602.04485*, 2016.

[6] Monica Bianchini and Franco Scarselli, "On the complexity of neural network classifiers: A comparison between shallow and deep architectures," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1553–1565, 2014.

[7] Shanshan Wang, Zhenghang Su, Leslie Ying, Xi Peng, Shun Zhu, Feng Liang, Dagan Feng, and Dong Liang, "Accelerating magnetic resonance imaging via deep learning," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 514–517.

[8] K Hammernik, F Knoll, D Sodickson, and T Pock, "Learning a variational model for compressed sensing MRI reconstruction," in *Proceedings of the International Society of Magnetic Resonance in Medicine (ISMRM)*, 2016.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[10] Herbert Edelsbrunner and John Harer, "Persistent homology-a survey," *Contemporary Mathematics*, vol. 453, pp. 257–282, 2008.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[12] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.