SFT-KD-Recon: Learning a Student-friendly Teacher for Knowledge Distillation in Magnetic Resonance Image Reconstruction

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Deep cascaded architectures for magnetic resonance imaging (MRI) acceleration have shown remarkable success in providing high-quality reconstruction. However, as the number of cascades increases, the improvements in reconstruction tend to become marginal, indicating possible excess model capacity. Knowledge distillation (KD) is an emerging technique to compress these models, in which a trained deep 'teacher' network is used to distill knowledge to a smaller 'student' network, such that the student learns to mimic the behavior of the teacher. Most KD methods focus on effectively training the student with a pre-trained teacher that is unaware of the student model. We propose SFT-KD-Recon, a student-friendly teacher training approach along with the student as a prior step to KD to make the teacher aware of the student's structure and capacity and enable aligning the teacher's representations with the student. In SFT, the teacher is jointly trained with the unfolded branch configurations of the student blocks using three loss terms - teacher-reconstruction loss, student-reconstruction loss, and teacher-student imitation loss, followed by KD of the student. We perform extensive experiments for MRI acceleration in 4x and 5x under-sampling, on the brain and cardiac datasets on five KD methods using the proposed approach as a prior step. We consider the DC-CNN architecture and setup teacher as D5C5 (141765 parameters), and student as D3C5 (49285 parameters) denoting 2.87:1 compression. Results show that (i) our approach consistently improves the KD methods with improved reconstruction performance and image quality, and (ii) the student distilled using our approach is competitive with the teacher, with the performance gap reduced from 0.53 dB to 0.03 dB.

Keywords: Knowledge Distillation (KD), Student-Friendly Teacher KD for reconstruction (SFT-KD-Recon), MRI, deep cascaded convolutional neural networks (DC-CNN).

1. Introduction

Modern deep neural networks have achieved outstanding performance in various medical imaging tasks such as image reconstruction, super-resolution, and object detection. Specifically in Magnetic Resonance Imaging (MRI) acceleration, the top five performing methods of fastMRI were implemented by using cascades of deep learning (DL) models (Muckley et al., 2021), with greater depth and complexity yielding higher output fidelity. However, the gain in performance per added cascade is not linear (Ramanarayanan et al., 2020a) over discrete increments in model size but rather becomes marginal asymptotically, implying possible excess capacity and potential compressibility of cascade models like DC-CNN (Schlemper et al., 2017) and DC-UNet (Sun et al., 2019).

© 2023 Anonymous.

The field of model compression research addresses the problem of achieving the least network size with minimum reduction in performance. To obtain lightweight models, there exist recent techniques such as (1) model pruning [(Li et al., 2016), (He et al., 2017)], where the model weights are sparsified to minimize redundancies, (2) lightweight network design [(Howard et al., 2017), (Zhang et al., 2018a)], such as substituting convolutions with separable convolutions, and (3) knowledge distillation (KD) methods [(Hinton et al., 2015), (Romero et al., 2014), (Yim et al., 2017b), (Gao et al., 2018)], where a trained deep 'teacher' network is used to distill representations to a smaller 'student' network, and the student learns to mimic the behavior of the teacher network, retaining the structure (Fig. 1a).

KD is an interesting method, as besides distilling from a pre-trained teacher network (Hinton et al., 2015), the training framework can accommodate the training of the teacher model along with the student. Such approaches can render the teacher aware of the student's structure and capacity, and offer scope for aligning representations with the student's layer features. A real-world analogy is neuro-linguistic programming, wherein educators understand what motivates students and orient the teaching to suit them (Tosey and Mathison, 2010). For example, Student-Friendly Teacher (SFT) training (Park et al., 2021) is a recent method that introduces a prior step of jointly training the teacher along with the student (named student-aware training), followed by routine feature-based knowledge distillation. This approach uses modular block-structured architectures for the teacher and the student and augments the student branches to the teacher during student-aware teacher training. This has been shown to improve top-1 classification accuracy in standard image classification datasets using ResNet-50 as a teacher and ResNet-34 as a student with a compression factor of 32 % (Park et al., 2021).

For MRI reconstruction, we consider the deep cascaded convolutional neural network (DC-CNN) owing to its block-structured configuration and good-quality reconstruction. For KD, we consider the DC-CNN with deeper blocks as the teacher and shallower blocks as the student. We perform SFT training by enabling the teacher to interact with the student block-wise. This student-oriented approach enables the low-level features of the student to match with the corresponding features of the teacher, improving both the teacher and the student. Furthermore, we effectively reuse the mutual knowledge (Zhang et al., 2018b) learned by the student for the subsequent distillation process. Hence, we have chosen the SFT-KD framework to ensure consistency between the teacher and the student for optimal knowledge distillation in MRI reconstruction. Our contributions are as follows:

1. We propose SFT-KD-Recon, a Student-Friendly Teacher learning framework for enhancing knowledge distillation for MRI restoration tasks. The proposed framework enables knowledge distillation to be oriented to the student network.

2. The proposed framework improves the teacher and provides an initialization to the student layers for the subsequent stage of distillation. Our method is straightforward and can be easily plugged into the standard KD methods for high fidelity.

3. We demonstrate the effectiveness of SFT-KD-Recon on various KD methods for MRI reconstruction and super-resolution tasks. Experiments reveal (i) the superiority of the proposed student-friendly teacher over conventional teacher models, and (ii) the consistency of SFT-KD-Recon in improving the student accuracy across five KD methods for image reconstruction and super-resolution.

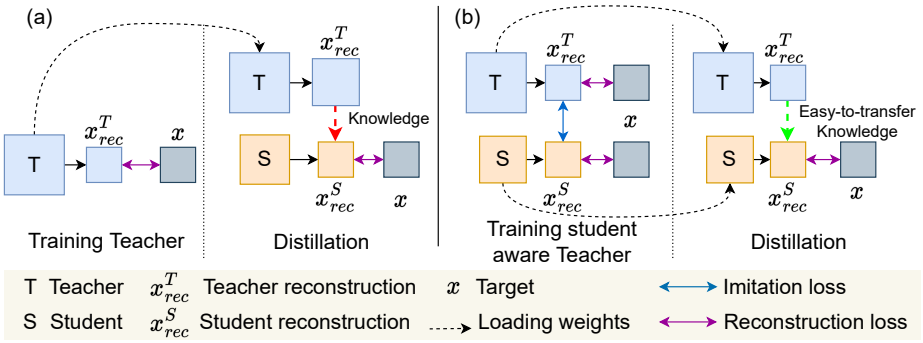

Figure 1: Comparison between the standard KD and SFT-KD-Recon. (a) The standard KD trains teacher alone and distills knowledge to student. (b) SFT-KD-Recon trains the teacher along with the student branches and then distills effective knowledge to student.

2. Related work

MRI Reconstruction: Several data-driven deep learning architectures have been proposed for the problem of MRI reconstruction (Wang et al., 2016; Lee et al., 2017; Schlemper et al., 2017; Sun et al., 2019; Huang et al., 2019). These methods are cascaded with interleaved CNN and data consistency blocks with residual connections. The DC-CNN (Schlemper et al., 2017) is the foremost deep cascaded architecture that learns the mapping between under-sampled and fully-sampled images. Subsequently, more complex architectures with dilated networks (Sun et al., 2018), attention mechanism (Huang et al., 2019), and dense connections (Wu et al., 2018) and adaptive reconstruction (Ramanarayanan et al., 2020b) have shown promising performance for various anatomies, contrasts and acceleration factors for reconstruction.

Knowledge Distillation: The concept of knowledge distillation was first introduced by (Hinton et al., 2015) for model compression in deep neural networks. Previous works have shown a range of KD methods to distill knowledge to the student networks. Attention transfer (AT) (Zagoruyko and Komodakis, 2016) distills the response patterns in the teacher to the student feature maps, and has shown to provide the best KD results for MRI reconstruction (Murugesan et al., 2020). FitNets (Romero et al., 2014) uses the pre-trained teacher's hint layer and student's guided layer and minimizes the loss between these layer representations. In the case of flow of solution procedure (FSP) (Yim et al., 2017a), the student preserves the pairwise similarities with the teacher in the representation space of the first and the penultimate layer. In Similarity-preserving (SP) KD (Tung and Mori, 2019), the student mimics the similarity map of the intermediate layers of the teacher. In Correlation Congruence (CC) (Peng et al., 2019), information at the input instance level and the correlation between instances are transferred to the student.

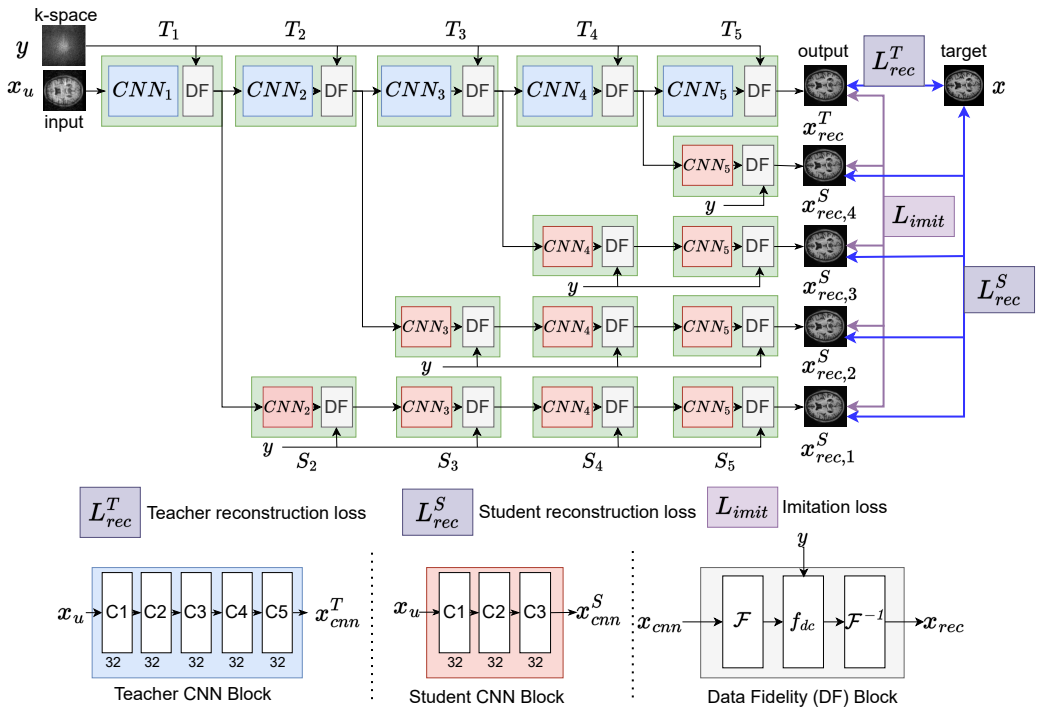

Figure 2: Student-Friendly training of the teacher. The teacher DC-CNN has five blocks, each having CNN with five convolution layers followed by DF layer, and the student DC-CNN has five blocks, each having three convolution layers and a DF layer. The teacher is trained with three loss terms - L^T_{rec}, L^S_{rec} (blue arrows), and L_{imit} (violet arrows). Note that all the blocks of the student learn initial weights except the first block during SFT training.

3. Methodology

3.1. Problem formulation for deep learning based MRI reconstruction

The deep-learning-based MRI reconstruction can be formulated as an optimization problem (Huang et al., 2019) and is given by:

$$\min_{x} \|x - f_{cnn}(x_u \mid \theta)\|_2^2 + \lambda \|\mathcal{A}x - y\|_2^2 \tag{1}$$

where $x \in \mathbb{C}^N$ denotes the desired image, $y \in \mathbb{C}^M$ is the under-sampled k-space measurements. Here $\mathcal{A} : \mathbb{C}^N \to \mathbb{C}^M$ represents the forward operator of the MRI acquisition and is given by $\mathcal{A} = \mathcal{M} \circ \mathcal{F}(x)$ where $\circ$ indicate Hadamard product. $\mathcal{M}$ is the under-sampling mask and $\mathcal{F}(.)$ is the 2D Fourier transform. Here, f_{cnn} is the CNN that learns the mapping between the zero filled (ZF) or under-sampled image x_u and the fully sampled image x and is parameterized by θ. The data fidelity (DF) operation in the k-space domain is performed after each CNN block to ensure consistency with the acquired k-space and is given by:

$$\hat{X}_{rec} = \begin{cases} \hat{X}_{cnn}(k) & k \notin \Omega \\ \frac{\hat{X}_{cnn}(k) + \lambda \hat{X}_u(k)}{1+\lambda} & k \in \Omega, \lambda \to \infty \end{cases} \tag{2}$$

where $x_{cnn} = f_{cnn}(x_u \mid \theta)$, $\hat{X}_{cnn} = \mathcal{F}(x_{cnn})$, $\hat{X}_u = \mathcal{F}(x_u)$, $x_{rec} = \mathcal{F}^{-1}(\hat{X}_{rec})$, Ω is the index set of the acquired k-space measurements and λ is the data fidelity weight.

3.2. Proposed Student-friendly Teacher-KD Framework

The SFT approach aims to train the teacher model collaboratively with the student, prior to distillation. During teacher training, the teacher learns along with the student branches to obtain representations tailored to the student. Figure 2 shows that SFT training consists of a teacher network along with multiple branches from the student network. Let $\{T_{\mathrm{n}}\}_{n=1}^{N}$ and $\{S_{\mathrm{n}}\}_{n=1}^{N}$ denote the blocks in the teacher and the student respectively. Here N denotes the number of blocks in the two networks. The figure shows a case with $N = 5$. Here, T_1 block is associated with $S_2 - S_3 - S_4 - S_5$ branch, T_2 is associated with $S_3 - S_4 - S_5$, and so on. By infusing student intermediate branches with the corresponding blocks of the teacher, each teacher block improves its knowledge based on the respective student branch and also the knowledge of the subsequent teacher blocks. For instance, T1 learns according to the $S_2 - S_3 - S_4 - S_5$ branch and T_2 to T_5 blocks. We consider three loss terms:

(1) ***Teacher-target reconstruction loss:*** The $L1$ loss between the teacher's prediction x_{rec}^{T}, and the target x, given by

$$L_{rec}^{T} = \left\| x - x_{rec}^{T} \right\|_1 \tag{3}$$

(2) ***Student-target reconstruction loss:*** The average $L1$ loss taken over each student branch output $x_{rec,i}^{S}$, $i = 1, 2, ..N-1$ and the target, given by

$$L_{rec}^{S} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\| x - x_{rec,i}^{S} \right\|_1 \tag{4}$$

(3) ***Student-teacher imitation loss:*** The $L1$ loss that minimizes the inconsistency between the predictions of the teacher and the student (branch-wise), given by

$$L_{imit} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\| x_{rec}^{T} - x_{rec,i}^{S} \right\|_1 \tag{5}$$

The overall training process is illustrated in Algorithm 1

Algorithm 1: SFT-KD-Recon procedure:

- Step 1: Train the teacher DC-CNN, f_{cnn}^{T} parameterized by θ^{T}, along with multiple student branches with loss $L_{SFTN} = L_{rec}^{T} + L_{rec}^{S} + L_{imit}$.
- Step 2: Load the student blocks weights θ^{S} obtained during SFT training in Step 1, and train the student network, f_{cnn}^{S} using reconstruction loss $L_{rec}^{S} = \left\| x - x_{rec}^{S} \right\|_1$ and distillation loss (depending upon the KD method) between teacher and student.

4. Experiments and Results

4.1. Dataset Description and Evaluation metrics

We have used two publicly available datasets to evaluate the proposed method for MRI reconstruction. (1) **The Cardiac MRI dataset**, released as part of the Automated Cardiac Diagnosis Challenge (ACDC) (Bernard et al., 2018), consists of 150 patient records (1841 slices) for training and 50 patient records (1076 slices) for validation. Each slice is cropped to its central 150×150 region. (2) **The Brain MRI dataset**, MRBrainS Dataset (Mendrik et al., 2015) consists of 7 volumes of T1 MRI split into 5 volumes (240 slices) for training and the remaining (96 slices) for validation. Each slice is of size 240×240. The under-sampled k-space and under-sampled images are retrospectively obtained using fixed Cartesian under-sampling masks for 4x and 5x acceleration factors (Ramanarayanan et al., 2020a).

We use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics as our evaluation metrics.

4.2. Implementation Details

For KD, we use DC-CNN configurations $D5C5$ as the teacher and $D3C5$ as the student with a compression factor of 65%. Here $D3C5$ means each cascade block has 3 convolution layers and there are 5 cascade blocks. Each layer in the two networks has 32 channels with ReLU activations and 3x3 filters. The cascaded network have alternating CNN blocks and DF units. We have chosen five KD methods to demonstrate the effectiveness of our SFT approach, namely AT (Attention Transfer) (Zagoruyko and Komodakis, 2016), Fitnets (Romero et al., 2014), Similarity Procedure (SP) (Tung and Mori, 2019), Flow of Solution Procedure (FSP) (Yim et al., 2017a), and CC (Correlation congruence) (Peng et al., 2019).

Models are implemented in PyTorch (v1.12) on a 24GB RTX 3090 GPU. For every step mentioned in the training algorithm, models are trained for 150 epochs using the Adam optimizer, with a learning rate of $1e^{-3}$. Our SFT framework is built upon KD-MRI [1] and the KD methods are taken from the RepDistiller repository [2].

4.3. Results and discussion

We compare our SFT-KD-Recon (Student trained using our procedure) with Std-KD (standard KD i.e. student trained using pre-trained teacher (Hinton et al., 2015)), student, and teacher in 4x and 5x setup on cardiac and brain datasets. Table 1 provides the quantitative analysis of the reconstruction performance of the proposed SFT approach applied to five different KD methods. From the table, our observations are as follows. (i) Student-friendly teacher performs consistently better than the conventional teacher with the highest improvement margin of 0.15 dB in PSNR and 0.002 in SSIM. (ii) SFT-KD is consistently better than Std-KD in the five KD methods. (iii) In terms of PSNR, the performance drop in the student relative to the teacher is 0.56 dB without KD. By using Std-KD, this gap is 0.53 dB. With the proposed SFT-KD, the gap is significantly minimized to 0.03 dB. We note that during SFT training, multiple configurations of the students ((i) S_5 alone, (ii) S_4 and

1. https://github.com/Bala93/KD-MRI
2. https://github.com/HobbitLong/RepDistiller

Table 1: Comparison of our framework with standard KD framework for MRI Reconstruction on MRBrainS and cardiac datasets. In all the KD methods, the student distilled from the SFT outperform the ones trained from the standard teacher.

Dataset	**Model**	**4x**		**5x**	
		Std-KD	**SFT-KD-Recon**	**Std-KD**	**SFT-KD-Recon**
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
MRBrainS	ZF	31.38/0.6651	-	29.93/0.6304	-
	Teacher	40.05/0.9785	**40.23/0.9799**	39.11/0.9715	**39.16/0.972**
	Student	39.49/0.9759	-	38.49/0.9663	-
	Fitnets	39.52/0.9762	**40.08/0.9790**	38.70/0.9681	**38.95/0.9699**
	AT	39.76/0.9769	**40.07/0.9789**	38.85/0.9691	**38.92/0.9699**
	CC	39.64/0.9762	**40.06/0.9789**	38.60/0.9675	**38.86/0.9691**
	FSP	39.45/0.9756	**40.01/0.9786**	38.32/0.9643	**38.82/0.9692**
	SP	39.53/0.9762	**39.93/0.9782**	38.46/0.9661	**38.81/0.9692**
Cardiac	ZF	24.27/0.6996	-	23.82/0.6742	-
	Teacher	32.15/0.9108	**32.26/0.9126**	31.25/0.8964	**31.33/0.8968**
	Student	31.68/0.9013	-	30.59 /0.8826	-
	AT	31.95/0.9060	**32.03/0.9070**	30.88/0.8879	**30.93/0.8884**
	Fitnets	31.90/0.9050	**31.95/0.9067**	30.59/0.8811	**30.86/0.8871**
	CC	31.83/0.9049	**31.96/0.9070**	30.73/0.8859	**30.94/0.8889**
	FSP	31.71/0.9024	**31.91/0.9060**	30.51/0.8817	**30.82/0.8862**
	SP	31.58/0.9004	**31.91/0.9060**	30.73/0.8860	**30.83/0.8865**

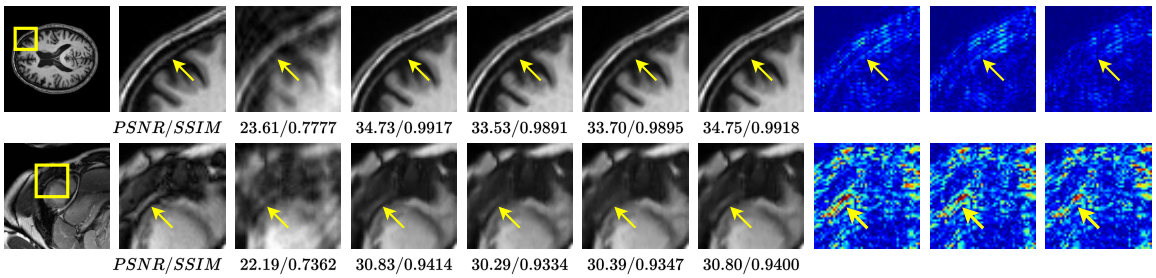

Figure 3: Visual results (from left to right): target, target inset, ZF, teacher, student, Std-KD, SFT-KD-Recon, student residue, Std-KD residue, SFT-KD-Recon residue with respect to the target, for the brain (top) and cardiac (bottom) with 4x acceleration. We note that in addition to lower reconstruction errors, the SFT-KD distilled student is able to retain finer structures better when compared to the student and Std-KD output.

S_5, (iii) S_3, S_4 and S_5, and so on) are collaboratively involved. The teacher and each student configuration are primarily directed by a supervised learning loss. Each configuration of student branch is a sub-network that steers towards a common objective. Each student sub-network starts with different initialization and provides predictions that serve as extra information for the teacher to optimize to a more robust set of features (Zhang et al., 2018b). This joint learning of the teacher makes it oriented toward student and transfers effective knowledge during the distillation stage as shown in our experiments. Interestingly, the distilled student performs as good as the original teacher in three KD methods, namely AT, Fitnets, and CC for 4x acceleration.

Among all the KD methods, the SFT-KD-Recon for AT shows the best performance improvements in all cases of anatomy and acceleration. Using AT, we further analyzed the performance of SFT-KD-Recon with and without initializing student weights learned during SFT training. Our PSNR/SSIM results for AT are as follows. Teacher: 40.05/0.9785, Student: 39.49/0.9759, KD: 39.76/0.9769, SFT-KD (random initializations): 40.01/0.9786, SFT-KD with initializations: 40.07/0.9789. This shows that the weights learned by student branches during SFT training are helpful to provide better initializations to the student.

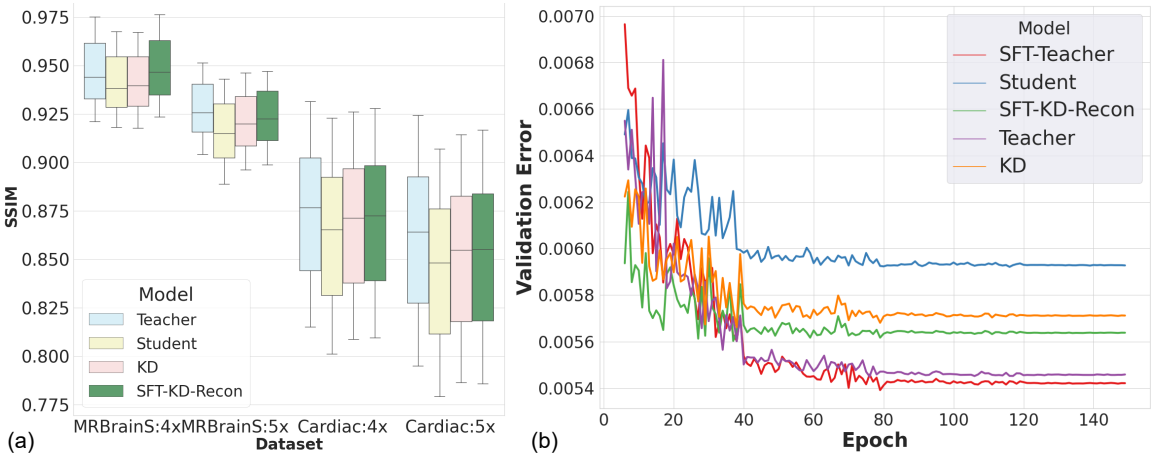

Figure 4: (a) SSIM Box plots of KD, SFT-KD-Recon with respect to teacher and student across the brain and cardiac datasets for 4x and 5x acceleration factors. (b) Reconstruction loss of teacher, student, SFT-Teacher, KD, SFT-KD-Recon on the validation set for the cardiac dataset, 4x acceleration factor.

Figure 3 shows the visual results comparing the target image, ZF, teacher, student, KD, and SFT-KD-Recon for the brain and cardiac respectively. Our observations are (i) SFT-Recon exhibits better structure recovery as compared to KD. The arrows pointing to the middle frontal gyrus region of the brain show that our method can recover details better than student and the Std-KD student. In the cardiac image, the cardiac mass pointed by the arrows for student and KD is more faded when compared to our approach. The box plots in Figure 4 show better results for our proposed approach (higher SSIM value (green) in figure 4(a) and lower reconstruction loss (green) in (b)) than student and KD.

5. Conclusion

In this work, we introduced SFT-KD-Recon, a student-friendly teacher training approach for improved knowledge distillation in the MRI restoration tasks. In the proposed approach, the teacher network is learned along with the student in a block-structured manner to improve the accuracy of the student during KD. Extensive experimentation for MRI reconstruction and super-resolution tasks on the brain and cardiac datasets using various KD methods show that the proposed SFT training can align the teacher to the student and can significantly improve the performance of knowledge distillation in MRI.

References

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018.

Qinquan Gao, Yan Zhao, Gen Li, and Tong Tong. Image super-resolution using knowledge distillation. In *Asian Conference on Computer Vision*, pages 527–541. Springer, 2018.

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1389–1397, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. Eprint arXiv:1503.02531, 2015.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. URL https://arxiv.org/abs/1704.04861.

Q. Huang, D. Yang, P. Wu, H. Qu, J. Yi, and D. Metaxas. Mri reconstruction via cascaded channel-wise attention network. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1622–1626, April 2019. doi: 10.1109/ISBI.2019.8759423.

Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

D. Lee, J. Yoo, and J. C. Ye. Deep residual learning for compressed sensing mri. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 15–18, April 2017. doi: 10.1109/ISBI.2017.7950457.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets, 2016. URL https://arxiv.org/abs/1608.08710.

Adriënne Mendrik, Koen Vincken, Hugo Kuijf, Marcel Breeuwer, Willem Bouvy, Jeroen de Bresser, Amir Alansary, Marleen de Bruijne, Aaron Carass, Ayman El-Baz, Amod Jog, Ranveer Katyal, Ali Khan, Fedde Lijn, Qaiser Mahmood, Ryan Mukherjee, Annegreet Opbroek, Sahil Paneri, Sérgio Pereira, and Max Viergever. Mrbrains challenge:

Online evaluation framework for brain image segmentation in 3t mri scans. *Computational Intelligence and Neuroscience*, 2015:1–16, 12 2015. doi: 10.1155/2015/813696.

Matthew J. Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, Simon Arberet, Dominik Nickel, Zaccharie Ramzi, Philippe Ciuciu, Jean-Luc Starck, Jonas Teuwen, Dimitrios Karkalousos, Chaoping Zhang, Anuroop Sriram, Zhengnan Huang, Nafissa Yakubova, Yvonne W. Lui, and Florian Knoll. Results of the 2020 fastmri challenge for machine learning mr image reconstruction. *IEEE Transactions on Medical Imaging*, 40(9):2306–2317, 2021. doi: 10.1109/TMI.2021.3075856.

Balamurali Murugesan, Sricharan Vijayarangan, Kaushik Sarveswaran, Keerthi Ram, and Mohanasankar Sivaprakasam. Kd-mri: A knowledge distillation framework for image reconstruction and image restoration in mri workflow. In *Medical Imaging with Deep Learning*, pages 515–526. PMLR, 2020.

Dae Young Park, Moon-Hyun Cha, Daesin Kim, Bohyung Han, et al. Learning student-friendly teacher networks for knowledge distillation. *Advances in Neural Information Processing Systems*, 34:13292–13303, 2021.

Nikolaos Passalis and Anastasios Tefas. Probabilistic knowledge transfer for deep representation learning. *CoRR, abs/1803.10837*, 1(2):5, 2018.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2019.

Sriprabha Ramanarayanan, Balamurali Murugesan, Keerthi Ram, and Mohanasankar Sivaprakasam. DC-WCNN: A Deep Cascade of Wavelet Based Convolutional Neural Networks for MR Image Reconstruction. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020a.

Sriprabha Ramanarayanan, Balamurali Murugesan, Keerthi Ram, and Mohanasankar Sivaprakasam. MAC-ReconNet: A Multiple Acquisition Context based Convolutional Neural Network for MR Image Reconstruction using Dynamic Weight Prediction. In *MIDL*, volume 121, pages 696–708, 06–08 Jul 2020b. URL http://proceedings.mlr.press/v121/ramanarayanan20a.html.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Jo Schlemper, Jose Caballero, Joseph V. Hajnal, Anthony Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for mr image reconstruction. In *Information Processing in Medical Imaging*, pages 647–658, 2017. ISBN 978-3-319-59050-9.

Liyan Sun, Zhiwen Fan, Yue Huang, Xinghao Ding, and John Paisley. Compressed sensing mri using a recursive dilated network. In *Proc. AAAI Conf. Artif. Intell.*, 2018.

Liyan Sun, Zhiwen Fan, Xinghao Ding, Yue Huang, and John Paisley. Joint cs-mri reconstruction and segmentation with a unified deep network. In *Information Processing in Medical Imaging*, pages 492–504, 2019. ISBN 978-3-030-20351-1.

Paul Tosey and Jane Mathison. Neuro-linguistic programming as an innovation in education and teaching. *Innovations in Education and Teaching International*, 47(3):317–326, 2010. doi: 10.1080/14703297.2010.498183. URL https://doi.org/10.1080/14703297.2010.498183.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

S. Wang, Z. Su, L. Ying, X. Peng, S. Zhu, F. Liang, D. Feng, and D. Liang. Accelerating magnetic resonance imaging via deep learning. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 514–517, April 2016. doi: 10.1109/ISBI.2016.7493320.

Huafeng Wu, Yawen Wu, Liyan Sun, Congbo Cai, Yue Huang, and Xinghao Ding. A deep ensemble network for compressed sensing mri. In *Proc. Neural Inf. Process.*, pages 162–171, 2018. ISBN 978-3-030-04167-0.

Junho Yim, Donggyu Joo, Ji-Hoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. pages 7130–7138, 07 2017a. doi: 10.1109/CVPR.2017.754.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7130–7138, 2017b. doi: 10.1109/CVPR.2017.754.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018a.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018b.

Appendix A. MRI Reconstruction

A.0.1. Tables and figures

Table 2: Ablative study comparing SFT and SFT-KD-Recon for AT method on MRBrainS dataset and 4x acceleration. Here our framework which trains with better initializations ensures better performance. In general, fine-tuning often outperforms training from scratch because the student during SFT training already has a generous amount of knowledge.

Evaluation Metric	Teacher	Student	KD	SFT	SFT-KD-Recon
PSNR	40.05 ± 2.023	39.49 ± 1.778	39.76 ± 1.899	40.01 ± 1.985	**40.07 ± 1.983**
SSIM	0.9785 ± 0.00655	0.9759 ± 0.005949	0.9769 ± 0.006172	0.9786 ± 0.006202	**0.9789 ± 0.0062**

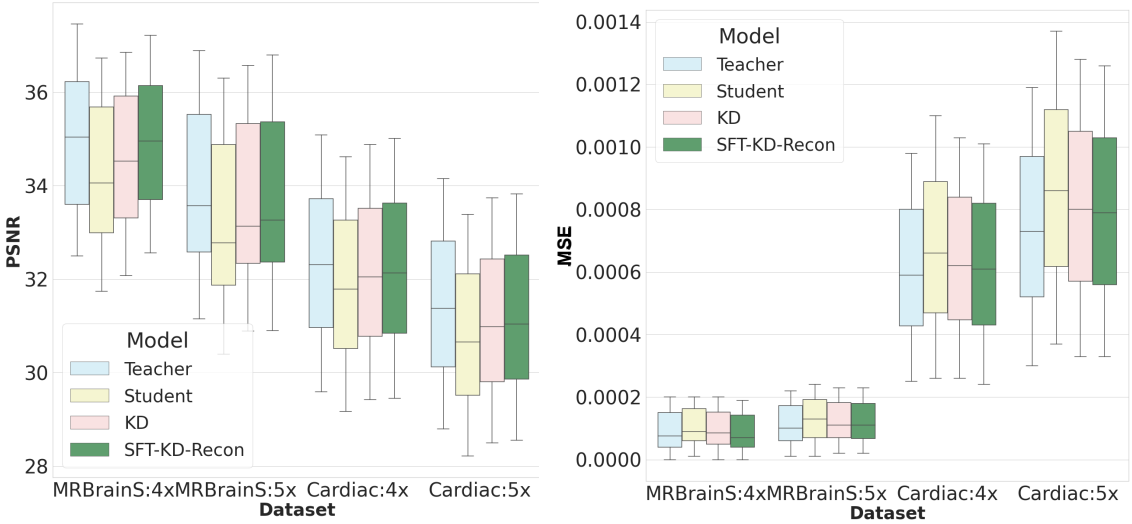

Figure 5: PSNR and MSE Box plots of KD, SFT-KD-Recon with respect to teacher and student across validation data of MRBrainS and cardiac datasets for 4x and 5x acceleration factors. From the plots, it is evident that our approach improves student performance in terms of PSNR also. Our approach (green) has more PSNR and less mean square error than student and KD

Appendix B. MRI Super Resolution:

B.1. MRI Super-Resolution architecture

MRI Super-Resolution involves reconstructing a high-resolution (HR) image from a low-resolution (LR) image. In general, interpolation fails to recover the loss of high-frequency details (fine edges). So, deep learning architectures like VDSR (Kim et al., 2016) were proposed to restore the LR image. VDSR architecture consists of n blocks of convolution and ReLU with a residual connection between the input and the output. LR image is interpolated to match the dimension of the HR image. We consider VDSR with 12 convolution layers with a residual connection between input and output for the teacher network and 4 convolution layers with a residual connection between input and output for the student

network. We evaluate the performance of the student network using our approach and standard approach by considering various baseline KD methods such as AT (Zagoruyko and Komodakis, 2016), Fitnets (Romero et al., 2014), CC (Peng et al., 2019), Neuron selectivity transfer (NST) (Huang and Wang, 2017), Factor transfer (FT) (Kim et al., 2018), Probabilistic knowledge transfer (PKT) (Passalis and Tefas, 2018).

B.2. Experiments and Results:

B.2.1. Dataset Description and Evaluation metrics

Dataset Description: **Calgary dataset[Complex-valued]:** The human brain dataset was obtained with 12-channel receiver coil and then combined to fetch a single-coil acquisition. The dataset contains 35 volumes of T1 each of size 256 x 256. We consider the center 110 slices from each volume, which provided 25 volumes (2750 slices) and 10 volumes (1100 slices) for training and validation, respectively.

We use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics as our evaluation metrics.

B.2.2. Experiments and results:

Table 3: Comparison of our framework with standard KD framework for MRI super-resolution on MRBrainS dataset for 4x acceleration.

RESOLUTION TASK					
Dataset	**Model**	**4x**			
		Std-KD		SFT-KD-Recon	
		PSNR	**SSIM**	PSNR	SSIM
Calgary	Teacher	30.40+/- 1.345	**0.8753 +/- 0.01291**	30.49 +/- 1.262	**0.8767 +/- 0.01268**
	Student	29.94 +/- 1.423	**0.8656 +/- 0.01412**	-	-
	AT	29.97 +/- 1.448	**0.8662 +/- 0.01404**	**30.02 +/- 1.392**	**0.8673 +/- 0.01364**
	Fitnets	29.95 +/- 1.38	**0.8658 +/- 0.01373**	**30.01 +/- 1.381**	**0.8672 +/- 0.01360**
	CC	29.98 +/- 1.423	**0.8730 +/- 0.01277**	**29.99 +/- 1.425**	**0.8663 +/- 0.01328**
	PKT	29.95 +/- 1.435	**0.8659 +/- 0.01381**	**30.01 +/- 1.421**	**0.8676 +/- 0.01349**
	factortransfer	29.99 +/- 1.427	**0.8665 +/- 0.01332**	**30.04 +/- 1.387**	**0.8677 +/- 0.01360**
	NST	29.98 +/- 1.364	**0.8662 +/- 0.01374**	**29.99 +/- 1.414**	**0.8663 +/- 0.01382**

Table 3 shows the quantitative analysis of the distilled student performance of the proposed SFT approach applied to various KD methods. From the table, we observe that SFT-KD performs consistently better than the routine KD approach and the student-friendly teacher performs better than the conventional teacher. In terms of PSNR, the performance drop in the student from the teacher is around 0.46 dB without KD. By using normal KD methods, this drop reduces to only around 0.41 dB. With the proposed SFT-KD, the gap is significantly minimized to 0.36 dB.