Active MR k-space Sampling with Reinforcement Learning

Luis Pineda[1], Sumana Basu[2], Adriana Romero[1], Roberto Calandra[1], and Michal Drozdzal[1]

[1] Facebook AI Research
[2] McGill University

Abstract. Deep learning approaches have recently shown great promise in accelerating magnetic resonance image (MRI) acquisition. The majority of existing work have focused on designing better reconstruction models given a pre-determined acquisition trajectory, ignoring the question of trajectory optimization. In this paper, we focus on learning acquisition trajectories given a fixed image reconstruction model. We formulate the problem as a sequential decision process and propose the use of reinforcement learning to solve it. Experiments on a large scale public MRI dataset of knees show that our proposed models significantly outperform the state-of-the-art in active MRI acquisition, over a large range of acceleration factors.

Keywords: Active MRI Acquisition · Reinforcement Learning

1 Introduction

Magnetic resonance imaging (MRI) is a powerful imaging technique used for medical diagnosis. Its advantages over other imaging modalities, such as computational tomography, are its superior image quality and zero radiation exposure. Unfortunately, MRI acquisition is slow (taking up to an hour) resulting in patient discomfort and in artifacts due to patient motion.

MRI scanners sequentially acquire k-space measurements, collecting raw data from which an image is reconstructed (e.g., through an inverse Fourier transform). A common way to accelerate MRI acquisition is to collect less measurements and reconstruct the image using a partially observed k-space. Since this results in image blur or aliasing artifacts, traditional techniques enhance the image reconstruction using regularized iterative optimization techniques such as compressed sensing [6]. More recently, the inception of large scale MRI reconstruction datasets, such as [22], have enabled the successful use of deep learning approaches to MRI reconstruction [19, 3, 12, 1, 26, 23, 7, 18]. However, these methods focus on designing models that improve image reconstruction quality for a *fixed* acceleration factor and set of measurements.

In this paper, we consider the problem of optimizing the sequence of k-space measurements (i.e., trajectories) to reduce the number of measurements

taken. Previous research on optimizing k-space measurement trajectories is extensive and include Compressed Sensing-based techniques [13, 11, 24, 2], SVD basis techniques [27, 9, 28], and region-of-interest techniques [20]. However, all these solutions work with *fixed trajectories* at inference time. Only recently, on-the-fly acquisition trajectory optimization methods for deep-learning-based MRI reconstruction have emerged in the literature [4, 25]. On the one hand, [4] showed that jointly optimizing acquisition trajectory and reconstruction model can lead to slight increase in image quality for a *fixed* acceleration factor with subject-specific acquisition trajectories. On the other hand, [25] introduced *active* MRI acquisition, where both acceleration factor and acquisition trajectory are subject-specific. Their proposed approach performs trajectory optimization over a *full range* of possible accelerations but makes a myopic approximation during training that ignores the sequentiality of k-space acquisition.

In contrast, we focus on Cartesian sampling trajectories and expand the formulation of [25]. More precisely, we specify the active MRI acquisition problem as a Partially Observable Markov Decision Process (POMDP) [14, 5], and propose the use of deep reinforcement learning [8] to solve it.[3] Our approach, by formulation, optimizes the reconstruction over the whole range of acceleration factors while considering the sequential nature of the acquisition process – future scans and reconstructions are used to determine the next measurement to take. We evaluate our approach on a large scale single-coil knee dataset [22][4] and show that it outperforms common acquisition heuristics as well as the myopic approach of [25]. Our contributions are: 1) formulating active MRI acquisition as a POMDP; 2) showing state-of-the-art results in active MRI acquisition on a large scale single-coil knee dataset; 3) performing an in-depth analysis of the learned trajectories. The code to reproduce our experiments is available at: https://github.com/facebookresearch/active-mri-acquisition.

2 Background

Partially Observable Markov Decision Processes (POMDP) A POMDP is a highly expressive model for formulating problems that involve sequential decisions [5, 10, 14]. A solution to a POMDP is a *policy* that maps history of observations (in our case, the k-space measurements) to actions (in our case, an index of k-space measurement to acquire). A POMDP is formally defined by: (1) a set of *states,* summarizing all the relevant information to determine what happens next; (2) a set of *observations,* containing indirect information about

[3] For a comprehensive overview of reinforcement learning applied to healthcare data and medical imaging please refer to [21].

[4] Data used in the preparation of this article were obtained from the NYU fastMRI Initiative database (fastmri.med.nyu.edu) [22]. As such, NYU fastMRI investigators provided data but did not participate in analysis or writing of this report. A listing of NYU fastMRI investigators, subject to updates, can be found at: fastmri.med.nyu.edu. The primary goal of fastMRI is to test whether machine learning can aid in the reconstruction of medical images.

the underlying states, which are themselves not directly observable; (3) a set of *actions,* that cause the transition between states; (4) a *transition function,* specifying a probability distribution for moving from one state to another, after taking an action; (5) an *emission function,* defining the probability of obtaining each possible observation after taking an action in some state; (6) a *reward function,* specifying the numeric feedback obtained when moving from one state to another; and (7) and a *discount factor,* defining the impact of immediate versus more distant rewards.

Double Deep Q-Networks (DDQN) (DDQN) [17] is a state-of-the-art deep reinforcement learning method for solving high-dimensional POMDPs. Since active MRI acquisition involves discrete actions (e.g., indexes of k-space measurements to acquire), we focus on the Double DQN (DDQN) [17] algorithm, given its training stability and recent success. In DDQN, policies are not explicitly modelled. Instead, a *value network* is used to predict the *value* of each possible action—defined as the expectation of the future cumulative reward. A policy is then recovered by greedily choosing actions with maximal estimated value. The value prediction problem is posed as a supervised regression problem, trained to minimize the temporal difference error [15] over data sampled from a *replay memory buffer*, which contains tuples of states, actions and rewards obtained by running the learned policy in an exploratory way. The target values used to update the value network are computed as bootstrapped estimates of the estimated value. Further details can be found in [17].

3 Learning Active MR k-space Sampling

Let $\mathbf{y} \in \mathbb{C}^{M \times N}$ be a complex matrix representing a fully sampled k-space, and $\mathbf{x}$ be a reconstruction of $\mathbf{y}$ obtained via Inverse Fourier Transform, $\mathcal{F}^{-1}$; i.e., $\mathbf{x} = \mathcal{F}^{-1}(\mathbf{y})$. We simulate partially observed k-space by masking $\mathbf{y}$ with a Cartesian binary mask $\mathbf{M}$, resulting in $\tilde{\mathbf{y}} = \mathbf{M} \odot \mathbf{y}$. We denote the corresponding *zero-filled* reconstruction as $\tilde{\mathbf{x}} = \mathcal{F}^{-1}(\tilde{\mathbf{y}})$. In this paper, we use a deep-learning-based reconstruction model $\mathbf{r}(\tilde{\mathbf{x}}; \phi) \rightarrow \hat{\mathbf{x}}$ that takes $\tilde{\mathbf{x}}$ as input, and outputs a de-aliased image reconstruction $\hat{\mathbf{x}}$. The reconstruction model is a convolutional neural network (CNN) parametrized by ϕ; in particular, we use the network proposed in [25]. Finally, we use subindices to denote time, e.g., $\hat{\mathbf{x}}_t = \mathbf{r}(\mathcal{F}^{-1}(\mathbf{M}_t \odot \mathbf{y}))$ represents the reconstruction obtained at time step t of the acquisition process.

3.1 Active MRI Acquisition as POMDP

In our active acquisition formulation, the goal is to learn a *policy* function $\pi(\hat{\mathbf{x}}_t, \mathbf{M}_t; \theta) \rightarrow a_t$ that, at time t, maps the tuple of de-aliased image reconstruction $\hat{\mathbf{x}}_t$, and mask representing observed frequencies $\mathbf{M}_t$, to the *action* a_t. In Cartesian active MRI acquisition, actions are represented by the unobserved columns of the fully sampled k-space, $\mathbf{y}$ (the ones for which the mask $\mathbf{M}_t$ has value of 0). Once an action is observed, both the mask and the de-aliased image reconstruction are updated to $\mathbf{M}_{t+1} = \mathbf{M}_t + \mathbf{M}^{a_t}$ and $\hat{\mathbf{x}}_{t+1} = \mathbf{r}(\mathcal{F}^{-1}(\mathbf{M}_{t+1} \odot \mathbf{y}))$,

where $\mathbf{M}^{a_t}$ is a binary matrix with all zeros except of the column indicated by the action a_t. The parameters θ of the policy are optimized to select a sequence of actions (k-space measurements) $[a_t, a_{t+1}, ..., a_T]$ that minimize the acquisition cost function (i.e., maximize the reward function) over all future reconstructions up to time step T. In this work, we consider acquisition costs of the form $f(\mathcal{C}(\hat{\mathbf{x}}_1, \mathbf{x}), \mathcal{C}(\hat{\mathbf{x}}_2, \mathbf{x}), ..., \mathcal{C}(\hat{\mathbf{x}}_T, \mathbf{x}))$, where $\mathcal{C}$ represents a pre-defined cost of interest (e.g., Mean Squared Error or negative SSIM), and f is a function that aggregates the costs observed throughout an acquisition trajectory (e.g., sum, area under the metric curve, final metric value at time T). Thus, the objective is to minimize the aggregated cost over the *whole range* of MRI acquisition speed-ups. Further, we assume that f can be expressed as a discounted sum of T rewards, one per time step. Under this assumption, we formally introduce the active acquisition problem as a POMDP defined by:

- **State set**: The set of all possible tuples $\mathbf{s}_t \triangleq \langle \mathbf{x}, \mathbf{M}_t \rangle$. Note that the ground truth image, $\mathbf{x}$, is hidden and the current mask, $\mathbf{M}_t$, is fully visible.
- **Observation set**: The set of all possible tuples $\mathbf{o}_t \triangleq \langle \hat{\mathbf{x}}, \mathbf{M}_t \rangle$.
- **Action set**: The set of all possible k-space column indices that can be acquired, i.e., $\mathcal{A} \triangleq \{1, 2, ..., W\}$, where W is the image width. Since sampling an already observed k-space column does not improve reconstruction, we specify that previously observed columns are invalid actions.
- **Transition function**: Given current mask $\mathbf{M}_t$ and a valid action $a_t \in \mathcal{A}$, the mask component of the state transitions deterministically to $\mathbf{M}_{t+1} = \mathbf{M}_t + \mathbf{M}^{a_t}$, and $\mathbf{x}$ remains unchanged. After T steps the system reaches the final time step.
- **Emission function**: In our formulation, we assume that the reconstruction model returns a deterministic reconstruction $\hat{\mathbf{x}}_t$ at each time step t; thus, the observation after taking action a_t at state $\mathbf{s}_t$ is defined as $\mathbf{o}_t \triangleq \hat{\mathbf{x}}_{t+1}$ (i.e., the reconstruction after adding the new observed column to the mask).
- **Reward function**: We define the reward as the decrease in reconstruction metric with respect to the previous reconstruction: $R(\mathbf{s}_t, a_t) = \mathcal{C}(\hat{\mathbf{x}}_{t+1}, \mathbf{x}) - \mathcal{C}(\hat{\mathbf{x}}_t, \mathbf{x})$. This assumes that $f(\mathcal{C}(\hat{\mathbf{x}}_1, \mathbf{x}), \mathcal{C}(\hat{\mathbf{x}}_2, \mathbf{x}), ..., \mathcal{C}(\hat{\mathbf{x}}_T, \mathbf{x})) \triangleq \mathcal{C}(\hat{\mathbf{x}}_T, \mathbf{x})$. We found this reward to be easier to optimize than rewards corresponding to more complex aggregations.
- **Discount factor**: We treat the discount, $\gamma \in [0, 1]$, as a hyperparameter.

Note that the above-mentioned POMDP is *episodic*, since the acquisition process has a finite number of steps T. At the beginning of each episode, the acquisition system is presented with an initial reconstruction, $\hat{\mathbf{x}}_0$, of an unobserved ground truth image $\mathbf{x}$, as well with an initial subsampling mask $\mathbf{M}_0$. The system then proceeds to iteratively suggest k-space columns to sample, and receives updated reconstructions from $\mathbf{r}$.

3.2 Solving the Active MRI Aquisition POMDP with DDQN

We start by defining a subject-specific DDQN value network. As mentioned in Section 2, POMDP policies are functions of observation histories. However, in active Cartesian MRI acquisition, the whole history of observations is captured

by the current observation $\mathbf{o}_t$. Thus, we use $\mathbf{o}_t$ as single input to our value network. We design the value network architecture following [25]'s evaluator network, which receives as input a reconstructed image, $\hat{\mathbf{x}}_t$, and a mask $\mathbf{M}_t$. To obtain the reconstructed image $\mathbf{x}_t$ at each time step, we use a pre-trained reconstruction network. Additionally, we also consider a dataset-specific DDQN variant, which only takes time step information as input (which is equivalent to considering the number of non-zero elements in the mask $\mathbf{M}_t$), and thus, uses the same acquisition trajectory for all subjects in the dataset.

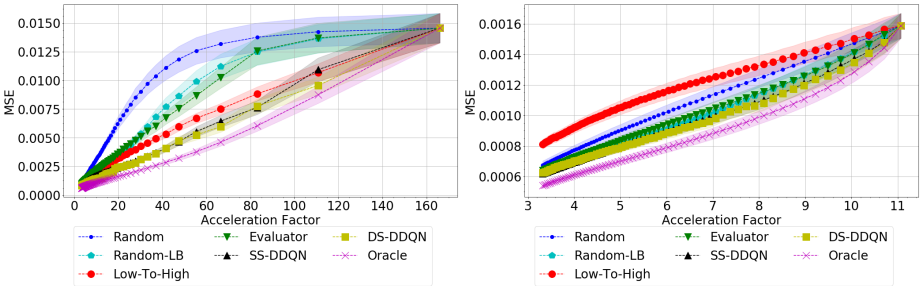
In both cases, we restrict the value network to select among valid actions by setting the value of all previously observed k-space columns to $-\infty$. Additionally, we use a modified ϵ-greedy policy [16] as exploration policy to fill the replay memory buffer. This policy chooses the best action with probability $1-\epsilon$, and chooses an action from the set of valid actions with probability ϵ.

4 Experimental Results

Datasets and baselines. To train and evaluate our models, we use the single-coil knee RAW acquisitions from the fastMRI dataset [22], composed of 536 public training set volumes and 97 public validation set volumes. We create a held-out test set by randomly splitting the public validation set into a new validation set with 48 volumes, and a test set with 49. This results in 19,878 2D images for training, 1785 images for validation, and 1851 for test. All data points are composed of a 640×368 complex valued k-space matrix with the 36 highest frequencies zero-padded. In all experiments, we use vertical Cartesian masks, such that one action represent acquisition of one column in the k-space matrix. Hence, the total number of possible actions for this dataset is 332.

We compare our approach to the following heuristics and baselines: (1) Random policy (RANDOM): Randomly select an unobserved k-space column; (2) Random with low frequency bias (RANDOM-LB): Randomly select an unobserved k-space column, favoring low frequency columns; (3) Low-to-high policy (LOWTOHIGH): Select an unobserved k-space column following a low to high frequency order; and (4) Evaluator policy (EVALUATOR): Select an unobserved k-space column following the observation scoring function introduced by [25]. To ensure fair comparisons among all methods, we fix same number of low-frequency observations (for details see Section 4). Moreover, for reference, we also include results for a one-step oracle policy that, having access to ground truth at test time, chooses the frequency that will reduce $\mathcal{C}$ the most (denoted as ORACLE).

Training details. In our experiments, we use the reconstruction architecture from [25] and train it using negative log-likelihood on the training set. Following [25], the reconstructor is trained on the whole range of acceleration factors, by randomly sampling masks with different patterns and acceleration factors each time an image is loaded. Similarly to [25], we also force a fixed number of low frequencies to be always observed, and train two versions of the reconstruction model: one that always observes 30 low frequencies—referred to as Scenario-30L—, corresponding to $\sim 3\times - 11\times$ accelerations, and another that



(a) Scenario-2L (b) Scenario-30L

Fig. 1: Reconstruction MSE vs. acceleration factor for different strategies. Both DDQN models outperform all baselines and heuristics for the vast majority of acceleration factors considered.

Metric	RANDOM	RANDOM-LB	LOWTOHIGH	EVALUATOR	SS-DDQN	DS-DDQN
MSE ($\times 10^{-3}$) ↓	8.90 (0.41)	8.24 (0.37)	9.64 (0.45)	8.33 (0.38)	8.00 (0.35)	**7.94 (0.35)**
NMSE ($\times 10^{-1}$) ↓	3.02 (0.16)	2.93 (0.16)	3.13 (0.17)	3.06 (0.17)	2.88 (0.17)	**2.87 (0.16)**
PSNR ($\times 10^{2}$) ↑	2.23 (1.28)	2.25 (1.75)	2.21 (1.23)	2.24 (1.33)	**2.27 (1.34)**	2.26 (1.35)
SSIM ↑	4.77 (0.06)	4.82 (0.07)	4.71 (0.06)	4.78 (0.07)	**4.86 (0.07)**	**4.86 (0.07)**

Table 1: Average test set AUC with 95% confidence intervals (one AUC value/image) for 6 different active acquisition policies (Scenario-30L).

always observes 2 low frequencies—referred to as Scenario-2L—, corresponding to extreme acceleration factors of up to 166× acceleration. In contrast to [25], we train the reconstructor and the policy networks in stages, as we noticed that it leads to better results. Note that none of the architectures used in the experiments uses complex convolutions. For both DDQN approaches, we experimented with the reward defined in Section 3.1 and four choices of $\mathcal{C}$: Mean Squared Error (MSE), Normalized MSE, negative Peak Signal to Noise Ratio (PSNR) and negative Structural Similarity (SSIM). For each metric, we trained a value network using a budget of $T = 100 - L$ actions, where $L = 2$ or $L = 30$ is the number of initial fixed low frequencies, a discount factor $\gamma = 0.5$, and a replay buffer of 20,000 tuples. Each training episode starts with a mask with a fixed number of low frequencies L. The DDQN models are trained for 5,000,000 state transition steps, roughly equivalent to 3.6 iterations over the full training set for $L = 30$, or 2.6 for $L = 2$. We periodically evaluate the current greedy policy on the validation set, and use the best validation result to select the final model for testing. To reduce computation, we used a random subset of 200 validation images when training DDQN.

Metric	RANDOM	RANDOM-LB	LOWTOHIGH	EVALUATOR	SS-DDQN	DS-DDQN
MSE ↓	1.97 (0.17)	1.73 (0.15)	1.39 (0.10)	1.70 (0.15)	1.31 (0.11)	**1.24 (0.10)**
NMSE ↓	20.6 (0.43)	18.7 (0.41)	17.5 (0.37)	17.7 (0.42)	15.5 (0.35)	**15.1 (0.34)**
PSNR ($\times 10^3$) ↑	3.63 (0.02)	3.73 (0.02)	3.76 (0.02)	3.79 (0.02)	3.89 (0.02)	**3.90 (0.02)**
SSIM ↑	73.2 (1.19)	75.3 (1.22)	73.3 (1.19)	76.0 (1.23)	**78.0 (1.25)**	**78.0 (1.25)**

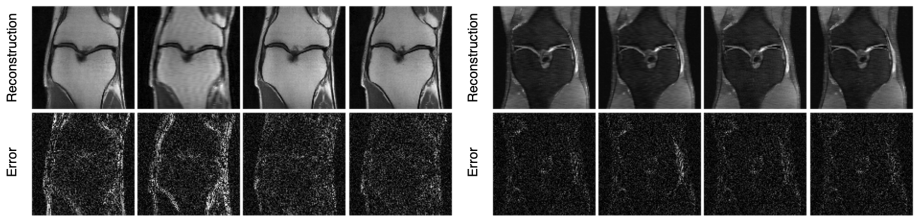
Table 2: Average test set AUC with 95% confidence intervals (one AUC value/image) for 6 different active acquisition policies (Scenario-2L).

4.1 Results

Fig. 1 (a-b) depicts the average test set MSE as a function of acceleration factor, with Scenario-2L on Fig. 1 (a) and Scenario-30L on Fig. 1 (b).[5] Results show that both DDQN models outperform all considered heuristics and baselines for the vast majority of acceleration factors. In Scenario-30L, the mean MSE obtained with our models is between 3-7% lower than the best baseline (EVALUATOR), for all accelerations between 4× and 10×. For the case of extreme accelerations (Scenario-2L), our best model (dataset-specific DDQN) outperforms the best heuristic (LOWTOHIGH) by at least 10% (and up to 35%) on all accelerations below 100×. Note that for this scenario and the MSE metric, the performance of RANDOM and EVALUATOR deteriorated significantly compared to Scenario-30L. To further facilitate comparison among all methods, we summarize the overall performance of an acquisition policy into a single number, by estimating, for each image, the area under curve (AUC), and averaging the resulting values over the test set. Results are summarized in Tab. 1 for Scenario-30L and Tab. 2 for Scenario-2L. In all cases, the DDQN policies outperform all other models. In Scenario-30L, the improvements of our models over the best baselines range from 0.55% to 2.9%, depending on the metric. In Scenario-2L, the improvements range from 2.68% to 11.6%. Further, paired t-tests (pairing over images) between the AUCs obtained with our models and those of the best baseline, for each metric, indicate highly significant differences, with p-values generally lower than 10^{-8}. Interestingly, we found that the data-specific DDQN slightly outperforms the subject-specific one for most metrics. While this seems to suggest that subject-specific trajectories are not necessary, we point out that the gap between the performance of ORACLE and the models considered indicates the opposite. In particular, policy visualizations for ORACLE (see Section 4.2) show wide subject-specific variety. Therefore, we hypothesize that the better performance of DS-DDQN is due to optimization and learning stability issues.

Qualitative results. Fig. 2 shows examples of reconstructions together with error maps for four different acquisition policies (RANDOM, LOWTOHIGH, EVALUATOR, subject-specific DDQN). We display the 10× and 8× acceleration for Scenario-2L and Scenario-30L, respectively. Looking at the error maps (bottom row), the differences in reconstruction quality between the subject-specific DDQN and the rest of the policies are substantial, and the reconstruction is visibly sharper and more detailed than the ones obtained with the baselines.

[5] Plots for NMSE, SSIM and PSNR are available in the supplementary material.



(a) Scenario-2L (b) Scenario-30L

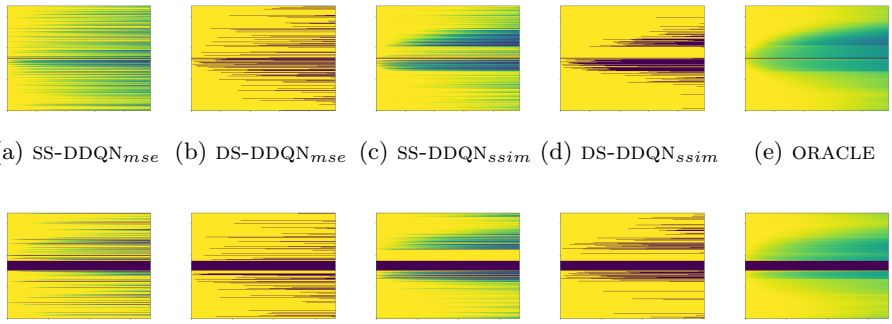
Fig. 2: Reconstructions and error maps under 4 acquisition policies (from left:RANDOM, LOWTOHIGH, EVALUATOR, SS-DDQN) at 10× acceleration for Scenario-2L and 8× for Scenario-30L. The images depict magnitude information. Note that the subfigures (a) and (b) depict different knee images. Additional images are shown in the supplementary material.

4.2 Policy Visualization

Fig. 3 illustrates DDQN policies on the test set for DDQN trained with MSE and negative SSIM acquisition costs. Each row in the heat maps represents a cumulative distribution function of the time (x-axis) at which the corresponding k-space frequency is chosen (y-axis); the darker color the higher the values. Note that low frequencies are closer to the center of the heat map, while high frequencies are closer to the edges. In the dataset-specific DDQN heat maps, each row instantly transitions from light to dark intensities at the time where the frequency is chosen by the policy (recall that this policy is a deterministic function of time). In the subject-specific DDQN, smoother transitions indicate that frequencies are not always chosen at the same time step. Furthermore, one can notice that some frequencies are more likely to be chosen earlier, indicated by a dark intensity appearing closer to the left side of the plot. Overall, for both models and costs, we observe a tendency to start the acquisition process by choosing low and middle frequencies, while incorporating high frequencies relatively early in the process. However, when comparing MSE-based to SSIM-based policies, we observe that the SSIM policy is more biased towards low frequencies and it seems to take advantage of k-space Hermitian symmetry – only few actions selected in the upper-center part of the SSIM heat maps.

5 Conclusion

In this paper, we formulated the active MRI acquisition problem as a Partially Observable Markov Decision Process and solved it using the Double Deep Q-Network algorithm. On a large scale single-coil knee dataset, we learned policies that outperform, in terms of four metrics (MSE, NMSE, SSIM and PSNR), simple acquisition heuristics and the scoring function introduced in [25]. We also



(a) SS-DDQN$_{mse}$ (b) DS-DDQN$_{mse}$ (c) SS-DDQN$_{ssim}$ (d) DS-DDQN$_{ssim}$ (e) ORACLE

(f) SS-DDQN$_{mse}$ (g) DS-DDQN$_{mse}$ (h) SS-DDQN$_{ssim}$ (i) DS-DDQN$_{ssim}$ (j) ORACLE

Fig. 3: Policy visualizations for DDQN models: Scenario-2L (a-e) and Scenario-30L (e-j). Subfigures (e) and (j) shows oracle policy obtained with SSIM criteria. See main text for details. See supplementary for additional results.

observed that the dataset-specific DDQN slightly outperforms the subject-specific DDQN, and that a gap still exists between our models and the best possible performance (illustrated by an oracle policy). This performance gap encourages further research to improve models and algorithms to address the active MRI acquisition problem. Finally, it is important to note that our experimental setup is simplified for the purpose of model exploration (e.g. we do not consider all the practical MRI phase-encoding sampling issues).

Acknowledgements. The authors would like to thank the fastMRI team at FAIR and at NYU for engaging discussions. We would like to express our gratitude to Amy Zhang and Joelle Pineau for helpful pointers and to Matthew Muckley for providing feedback on an early draft of this work.

References

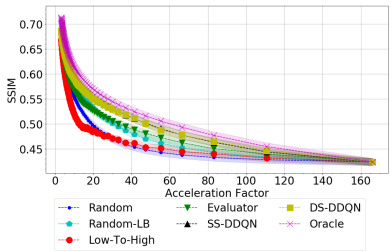
1. Feiyu Chen, V. Taviani, Itzik Malkiel, Joseph Cheng, Jonathan Tamir, Jamil Shaikh, Stephanie Chang, Christopher Hardy, John Pauly, and Shreyas Vasanawala. Variable-density single-shot fast spin-echo mri with deep learning reconstruction by using variational networks. *Radiology*, 289:180445, 07 2018.
2. Baran Gözcü, Rabeeh Karimi Mahabadi, Yen-Huan Li, Efe Ilıcak, Tolga Çukur, Jonathan Scarlett, and Volkan Cevher. Learning-based compressive mri. *IEEE transactions on medical imaging*, 37(6):1394–1406, 2018.
3. Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P. Recht, Daniel K. Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *arXiv preprint arXiv:1704.00447*, 2017.
4. Kyong Hwan Jin, Michael Unser, and Kwang Moo Yi. Self-supervised deep active accelerated mri. *arXiv preprint arXiv:1901.04547*, 2019.
5. Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

6. Michael Lustig, David Donoho, and John Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 58:1182–95, 12 2007.
7. Kai Lønning, Patrick Putzky, Jan-Jakob Sonke, Liesbeth Reneman, Matthan Caan, and Max Welling. Recurrent inference machines for reconstructing heterogeneous mri data. *Medical Image Analysis*, 53, 04 2019.
8. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
9. Lawrence P Panych, Claudia Oesterle, Gary P Zientara, and Jürgen Hennig. Implementation of a fast gradient-echo svd encoding technique for dynamic imaging. *Magnetic resonance in medicine*, 35(4):554–562, 1996.
10. Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.
11. Saiprasad Ravishankar and Yoram Bresler. Adaptive sampling design for compressed sensing mri. In *Engineering in Medicine and Biology Society (EMBC)*, pages 3751–3755, 2011.
12. J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503, 2018.
13. Matthias Seeger, Hannes Nickisch, Rolf Pohmann, and Bernhard Schölkopf. Optimization of k-space trajectories for compressed sensing by bayesian experimental design. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 63(1):116–126, 2010.
14. Edward Jay Sondik. The optimal control of partially observable markov processes. Technical report, Stanford Univ Calif Stanford Electronics Labs, 1971.
15. Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.
16. Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
17. Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*, 2016.
18. Shanshan Wang, Huitao Cheng, Leslie Ying, Taohui Xiao, Ziwen Ke, Hairong Zheng, and Dong Liang. Deepcomplexmri: Exploiting deep residual network for fast parallel mr imaging with complex convolution. *Magnetic Resonance Imaging*, 68:136 – 147, 2020.
19. 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.
20. Seung-Schik Yoo, Charles RG Guttmann, Lei Zhao, and Lawrence P Panych. Real-time adaptive functional mri. *Neuroimage*, 10(5):596–606, 1999.
21. Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: A survey. *ArXiv preprint arXiv:1908.08796*, 2019.
22. Jure Zbontar, Florian Knoll, Anuroop Sriram, Matthew J. Muckley, Mary Bruno, et al. fastMRI: An open dataset and benchmarks for accelerated MRI. In *ArXiv preprint arXiv:1811.08839*, 2018.
23. Pengyue Zhang, Fusheng Wang, Wei Xu, and Yu Li. Multi-channel generative adversarial network for parallel magnetic resonance image reconstruction in k-space. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer*

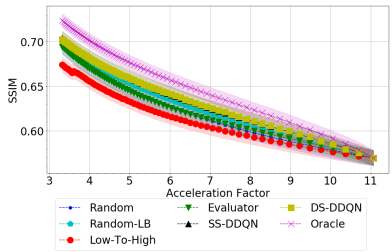
Assisted Intervention – MICCAI 2018, pages 180–188, Cham, 2018. Springer International Publishing.

24. Yudong Zhang, Bradley S Peterson, Genlin Ji, and Zhengchao Dong. Energy preserved sampling for compressed sensing mri. *Computational and mathematical methods in medicine*, 2014, 2014.
25. Zizhao Zhang, Adriana Romero, Matthew J. Muckley, Pascal Vincent, Lin Yang, and Michal Drozdzal. Reducing uncertainty in undersampled MRI reconstruction with active acquisition. In *The IEEE Conference on CVPR*, June 2019.
26. Bo Zhu, Jeremiah Zhe Liu, Bruce Rosen, and Matthew Rosen. Image reconstruction by domain transform manifold learning. *Nature*, 555, 03 2018.
27. Gary P Zientara, Lawrence P Panych, and Ferenc A Jolesz. Dynamically adaptive mri with encoding by singular value decomposition. *Magnetic Resonance in Medicine*, 32(2):268–274, 1994.
28. Gary P Zientara, Lawrence P Panych, and Ferenc A Jolesz. Applicability and efficiency of near-optimal spatial encoding for dynamically adaptive mri. *Magnetic resonance in medicine*, 39(2):204–213, 1998.

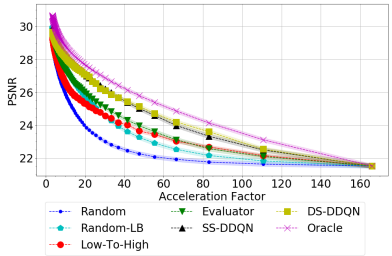
Supplementary



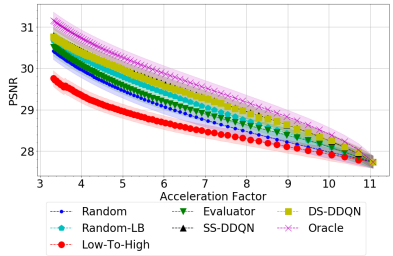
(a) SSIM: Scenario-2L



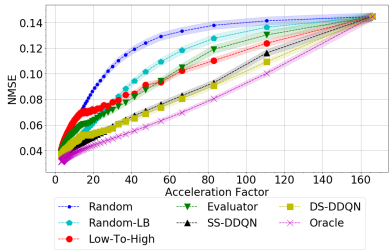
(b) SSIM: Scenario-30L



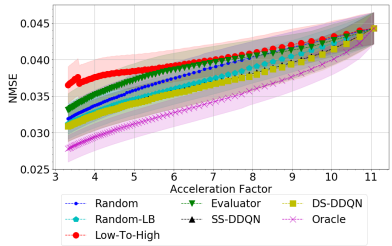
(c) PSNR: Scenario-2L



(d) PSNR: Scenario-30L

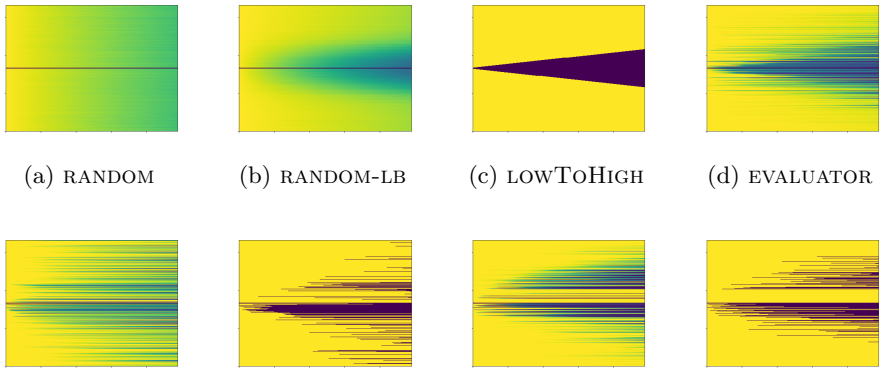


(e) NMSE: Scenario-2L



(f) NMSE: Scenario-30L

Fig. 4: Reconstruction quality vs. acceleration factor for different acquisition strategies. Our DDQN models outperform all baselines for most acceleration factors in all metrics. DDQN models are trained with rewards based on the corresponding metric. For MSE metric, see the main body of the paper.



(a) RANDOM (b) RANDOM-LB (c) LOWTOHIGH (d) EVALUATOR

(e) SS-DDQN$_{nmse}$ (f) DS-DDQN$_{nmse}$ (g) SS-DDQN$_{psnr}$ (h) DS-DDQN$_{psnr}$

Fig. 5: Policy visualizations for all heuristics and baselines for Scenario-2L (a-d) as well as DDQN policies trained with NMSE and PSNR. For DDQN policies trained with MSE and SSIM, see the main body of the paper.

Fig. 6: Example of image reconstructions and error maps under all different acquisition policies (Scenario-2L). Top to bottom: 4X, 16X, 64X acceleration. Left to right: RANDOM, RANDOM-LB, LOWTOHIGH, EVALUATOR, SS-DDQN, DS-DDQN, ORACLE.