MRI Image Reconstruction Using Vision Transformers

Prabhat Kumar Jharia
Signal Processing (Electrical Engineering)
Indian Institute of Science
Bengaluru, India
prabhatkj@iisc.ac.in

Muthuvel Arigovindan
Electrical Engineering
Indian Institute of Science
Bengaluru, India
mvel@iisc.ac.in

***Abstract*—The Vision Transformer (ViT) has shown promise in image classification tasks due to its self-attention mechanisms, which capture long-range dependencies. In this study, we explore its application in accelerated magnetic resonance imaging (MRI) reconstruction using the fastMRI dataset, which, despite its smaller scale, serves as a benchmark for such tasks. Our results demonstrate that a ViT-based model can achieve reconstruction accuracy comparable to the U-Net, while offering superior throughput and reduced memory consumption, making it more efficient for real-time applications.**

To address the challenge of limited MRI data, we propose a novel pre-training strategy using large natural image datasets like ImageNet. This approach enhances the ViT's performance on MRI tasks, significantly improving training efficiency and generalization across different anatomical variations. With as few as 100 MRI images, the pre-trained ViT outperforms pre-trained convolutional networks and other state-of-the-art methods in image quality.

Our research highlights the potential of Vision Transformers for accelerated MRI reconstruction, demonstrating their ability to deliver high-quality results with fewer training images and improved computational efficiency. The proposed pre-training strategy further strengthens the applicability of ViTs in clinical imaging, paving the way for more robust and efficient solutions in medical practice.

***Index Terms*—: Accelerated MRI, Transformer, pre-training, image reconstruction**

I. Introduction

Magnetic resonance imaging (MRI) is a widely used medical imaging technique known for its excellent soft-tissue contrast and safety, enabling reliable detection of various diseases such as tumors, hemorrhages, and infections. However, the inherently slow data acquisition process in MRI leads to prolonged examination times, which can be particularly challenging for patients who have difficulty remaining still, such as children. Even small movements during the scan can introduce image artifacts, further complicating the process.

To address this issue, MRI is often accelerated by collecting only a few undersampled measurements. Reconstruction algorithms must then leverage prior knowledge about MRI images to recover high-quality scans from the limited data. Traditionally, this prior knowledge has been incorporated through sparse representations of the images. More recently, deep learning-based methods have demonstrated superior reconstruction quality and speed compared to traditional approaches . These methods typically rely on convolutional neural networks (CNNs), which benefit from inductive biases that allow for impressive data efficiency.

However, the Vision Transformer (ViT), a convolution-free architecture with minimal inductive bias, has recently outperformed CNNs in image classification tasks when trained on large datasets. This suggests that the inductive biases in CNNs may limit performance in scenarios where abundant data is available. In such cases, the ViT can potentially learn more effective features directly from the training data. While the ViT and self-attention mechanisms have been widely explored for image classification and processing, research on their application to accelerated MRI is still limited.

In this paper, we investigate the potential of the ViT for accelerated MRI reconstruction. Our findings demonstrate that even when trained on the 35k-70k images of the fastMRI dataset, the ViT achieves reconstruction performance comparable to or better than the U-net, a widely used CNN-based baseline in state-of-the-art MRI reconstruction methods. Moreover, the ViT offers almost twice the throughput and reduced memory consumption compared to the U-net. We also show that pre-training the ViT on large natural image datasets like ImageNet significantly enhances its performance on MRI tasks, achieving competitive results with the state-of-the-art and surpassing pre-trained U-nets after fine-tuning on the fastMRI dataset. Particularly in low-data regimes, where only 100 MRI images are available for fine-tuning, the pre-trained ViT still delivers sharp and detailed reconstructions, outperforming the U-net. Additionally, we demonstrate that pre-trained ViTs exhibit greater robustness to anatomical variations compared to current CNN-based methods.

Our work highlights the promising potential of the Vision Transformer for accelerated MRI reconstruction, showing that it not only delivers high-quality results with fewer training images but also offers computational efficiency, making it a strong candidate for real-time clinical applications.

II. Dataset

In this study, we utilize the fastMRI dataset (Zbontar et al., 2019), which is currently the largest publicly available dataset for MRI reconstruction tasks. The fastMRI dataset consists of knee and brain MRI scans, providing a diverse set

of anatomical regions to evaluate the performance of reconstruction models. This dataset is widely used in the research community as a benchmark for evaluating MRI reconstruction algorithms, particularly in the context of accelerated MRI, where the goal is to reconstruct high-quality images from undersampled measurements.The dataset can be accessed at fastMRI dataset

The knee dataset includes a total of 35,000 MRI slices for training and 7,000 slices for validation. These slices were collected from knee scans of various patients, capturing a range of anatomical variations and pathologies. The knee dataset is useful for training models on images with relatively consistent anatomical features and can be challenging due to the relatively smaller size and higher variability within the slices compared to brain scans.

The brain dataset, on the other hand, is larger and more complex, containing 70,000 slices for training and 21,000 slices for validation. Brain MRIs encompass a broader range of anatomical structures and variations, including the cortex, subcortical regions, and deep structures like the brainstem. These variations in tissue types and structures present a more challenging task for reconstruction algorithms, requiring models to effectively capture fine details and complex anatomical relationships to produce high-quality reconstructions.

The fastMRI dataset provides a comprehensive and diverse set of training and validation images, allowing for the evaluation of reconstruction models across different anatomical regions and varying scan complexities. By using this dataset, we aim to assess the performance of the Vision Transformer (ViT) in comparison to traditional deep learning approaches, such as U-net, for accelerated MRI reconstruction.

III. Problem Formulation

During an accelerated MRI scan, electromagnetic waves are measured by several receiver coils. These measurements are commonly referred to as *k-space measurements*, which can be expressed as:

$$y_i = PFS_i x^* + z_i \quad \text{for} \quad i = 1, \ldots, C.$$

Where:

- $x^* \in \mathbb{C}^n$ is the unknown, vectorized image that we aim to reconstruct.
- $S_i \in \mathbb{C}^{n \times n}$ represents the sensitivity maps associated with the C receiver coils. This is typically realized as a diagonal matrix.
- $F \in \mathbb{C}^{n \times n}$ denotes the 2D discrete Fourier transform (DFT) matrix.
- $P \in \mathbb{R}^{m \times n}$ contains $m < n$ rows of an $n \times n$ identity matrix and describes the undersampling operation.
- $z_i \in \mathbb{C}^n$ represents the additive white Gaussian noise.

In the case of multi-coil MRI ($C > 1$), the setup involves multiple receiver coils, while for single-coil MRI ($C = 1$), there is only one receiver coil. The goal is to reconstruct the image x^* from the undersampled measurements y_i.

IV. Vision Transformer (ViT) Architecture for Image Reconstruction

The Vision Transformer (ViT), introduced by Dosovitskiy et al. (2020), adapts the original Transformer encoder (Vaswani et al., 2017) for image classification tasks. The core idea behind the ViT is to map an input image to features by transforming it into a sequence of image patches, which are then processed by a Transformer encoder. Below is a detailed explanation of the ViT architecture and its adaptation for image reconstruction.

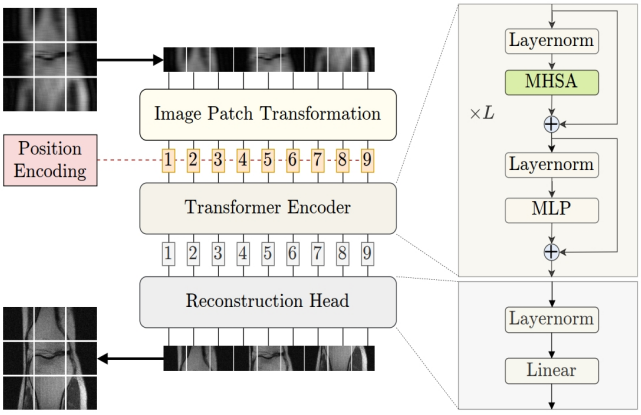

Fig. 1. Architecture of ViT.

The ViT operates as follows:

1) **Input Image Patchification:** The input image is spatially divided into N equally sized patches, each representing a small portion of the image.
2) **Patch Embedding:** Each image patch is linearly transformed into a d-dimensional feature vector, called a patch embedding, using a trainable linear transformation.
3) **Positional Embedding:** Since the Transformer encoder does not inherently capture spatial information, N learnable position embeddings are added to the patch embeddings. These d-dimensional vectors encode the absolute position of the patches within the image.
4) **Classification Token:** A learnable classification token is prepended to the sequence of patch embeddings. This token serves as a representative feature for the entire image and is used for classification purposes.
5) **Transformer Encoder:** The sequence of $N + 1$ feature vectors (patch embeddings plus classification token) is fed into the Transformer encoder. The encoder consists of L layers, each containing:
   - A **Multi-Head Self-Attention (MHSA)** block, which captures dependencies between patches.
   - A two-layer **MLP block** that processes each feature vector independently.
   - **Layer normalization** is applied before each block, and **residual connections** are used after each block to aid in training.
6) **Output:** At the output of the Transformer encoder, only the final representation of the classification token is

passed to a classification head, typically implemented as an MLP or a linear layer, to predict the class label of the image.

A. Key Differences from Other Vision Transformers (ViTs)

1) **GPSA Mechanism:**
   - Integrates Generalized Position-Specific Attention (GPSA) for incorporating positional locality.
   - Allows hybrid attention by blending positional and content-based information through learnable gating parameters.
2) **Support for Variable Patch and Input Sizes:**
   - Adapts to non-square images and varying patch sizes.
   - Enables better handling of irregular image dimensions during processing.
3) **Flexible Positional Embeddings:**
   - Uses bilinear interpolation for positional embeddings to support arbitrary image sizes.
4) **Reconstruction-Oriented Design:**
   - Designed for image reconstruction tasks, with a specialized sequence-to-image transformation.
   - Custom head for patch-level regression instead of classification.
5) **Locality Strength Parameter:**
   - Controls the intensity of local attention initialization in GPSA, promoting locality.
6) **Attention Map Insights:**
   - Provides methods to extract and analyze attention maps and their distance-based weights.
7) **Mixed Attention Block:**
   - Alternates between GPSA (for initial layers) and Multi-Head Self-Attention (MHSA) in deeper layers.

These modifications make the model versatile and particularly suitable for image reconstruction and tasks requiring localized attention.

B. Image Reconstruction Process

Once the Transformer encoder processes the sequence of image patches, the resulting output is passed through the reconstruction head, which generates reconstructed image patches. These patches are then combined to form the full-sized image, which serves as the final output of the model.

C. Model Architecture Summary

Unlike other approaches that integrate convolutions with specialized versions of Transformers, our model is convolution-free and relies solely on the standard Transformer encoder. This architecture is most similar to the Image Processing Transformer (Chen et al., 2021), which also uses a standard Transformer encoder but incorporates a convolutional image patch transformation and reconstruction head.

By using this architecture, we aim to leverage the power of Transformers to directly perform image reconstruction, eliminating the need for convolutional layers typically seen in traditional image reconstruction methods.

V. WORKING OF THE MODEL AND LOSSES

A. Mathematical Formulation of Losses

1) **SSIM Loss:** The Structural Similarity Index Measure (SSIM) is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (1)$$

where:
   - μ_x, μ_y: Means of x and y,
   - σ_x^2, σ_y^2: Variances of x and y,
   - σ_{xy}: Covariance of x and y,
   - C_1, C_2: Stabilizing constants to avoid division by zero.

   The loss form is:

$$\text{SSIMLoss}(x, y) = 1 - \text{SSIM}(x, y)$$

2) **Perceptual Loss:** Perceptual loss compares high-level feature representations of the input and target images using a pre-trained neural network (e.g., VGG). It is defined as:

$$\text{PerceptualLoss}(x, y) = \sum_{l \in \text{layers}} \frac{1}{N_l} \|\phi_l(x) - \phi_l(y)\|^2 \quad (2)$$

where:
   - $\phi_l(x)$: Feature map of layer l of the VGG network for input x,
   - N_l: Number of elements in the feature map of layer l,
   - $\|\cdot\|^2$: Mean squared error between feature maps.

3) **Pixel-wise Loss:** Pixel-wise loss compares the pixel intensity values of the input and target images directly. Using Mean Squared Error (MSE), it is defined as:

$$\text{PixelWiseLoss}(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2 \quad (3)$$

where:
   - x_i, y_i: Intensity values of the i-th pixel in x and y,
   - N: Total number of pixels in the image.

   Alternatively, Mean Absolute Error (MAE) is:

$$\text{MAE}(x, y) = \frac{1}{N} \sum_{i=1}^{N} |x_i - y_i| \quad (4)$$

4) **Combined Loss:** The combined loss is a weighted sum of SSIM Loss, Perceptual Loss, and Pixel-wise Loss:

$$\begin{aligned} \text{TotalLoss}(x, y) = \alpha \cdot \text{SSIMLoss}(x, y) \\ + \beta \cdot \text{PerceptualLoss}(x, y) \\ + \gamma \cdot \text{PixelWiseLoss}(x, y) \end{aligned} \quad (5)$$

TABLE I
EXPERIMENTAL COMBINATIONS FOR FINE-TUNING STRATEGIES

Exp ID	Layer Freezing	Learning Rate	Loss Function	Final SSIM, PSNR
E1	Patch Embedding + 25% Transformer Frozen	Variable	L2 Loss + SSIM Loss + Perceptual Loss	
E2	Patch Embedding + 50% Transformer Frozen	Variable	L2 Loss + SSIM Loss + Perceptual Loss	0.7441, 32.94
E3	Fine-Tune All Layers	Variable	L2 Loss + SSIM Loss + Perceptual Loss	0.7389, 32.89

where:

- $\alpha =$ ssim_weight: Weight for SSIM loss,
- $\beta =$ perceptual_weight: Weight for Perceptual loss,
- $\gamma =$ pixel_weight: Weight for Pixel-wise loss.

B. Metrics Used for Qualitative Analysis

1) **Peak Signal-to-Noise Ratio (PSNR):** The Peak Signal-to-Noise Ratio (PSNR) is defined as:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \tag{6}$$

where:

- MAX is the maximum possible pixel value of the image.
- MSE is the Mean Squared Error, calculated as:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2, \tag{7}$$

where x_i and $\hat{x}_i$ represent the original and reconstructed pixel values, respectively, and N is the total number of pixels.

2) **Normalized Mean Squared Error (NMSE):** The Normalized Mean Squared Error (NMSE) is given by:

$$\text{NMSE} = \frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{\sum_{i=1}^{N} x_i^2}, \tag{8}$$

where:

- x_i represents the original pixel values.
- $\hat{x}_i$ represents the reconstructed pixel values.
- N is the total number of pixels in the image.

VI. RESULTS

Reconstruction SSIM of different models in TABLE-2 when trained on the fastMRI dataset for 4-fold accelerated multi-coil (MC) and single-coil (SC) MRI, and their empirical computational costs during inference measured by throughput and maximal possible batch size.

VII. EXPERIMENTS

The experiment evaluates TABLE-1 Vision Transformers for MRI reconstruction by pretraining on ImageNet and fine-tuning on MRI data, analyzing different layer freezing strategies, learning rates, loss functions, and Final SSIM.

TABLE II

Metric	ViT-L	ViT-M	ViT-S
MC-Knee	0.908 ± 0.118	0.907 ± 0.118	0.903 ± 0.118
SC-Knee	0.744 ± 0.250	0.744 ± 0.249	0.740 ± 0.248
SC-Brain	0.828 ± 0.148	0.826 ± 0.148	0.823 ± 0.148
SC-Knee(our)	0	0	0.740 ± 0.248
Throughput	97.4 img/s	183.32 img/s	442.96, 152.88 img/s
Batch size	272	380	440, 24

VIII. FUTURE DEVELOPMENT

The future scope of this work lies in advancing Vision Transformers (ViT) for domain-specific tasks like MRI reconstruction. Improved transfer learning strategies, such as adaptive freezing based on layer importance, can enhance domain adaptation. Incorporating unsupervised pretraining (e.g., contrastive learning) on MRI-specific datasets may further improve fine-tuning performance. Hybrid architectures combining transformers with CNNs could exploit both local and global features, improving reconstruction quality. Integrating physics-based constraints into ViT for k-space processing in MRI can bridge the gap between model-based and learning-based approaches. Lastly, deploying lightweight ViT models using quantization and pruning enables real-time MRI reconstruction on resource-constrained devices.

REFERENCES

[1] Kang Lin, and Reinhard Heckel, "Vision Transformers Enable Fast and Robust Accelerated MRI," International Conference on Medical Imaging with Deep Learning, PMLR 172:774-795, 2022.

[2] 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. Magnetic Resonance in Medicine, 2018..

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations,2020.