BOSTON UNIVERSITY

BRIGHAM HEALTH BWH BRIGHAM AND WOMEN'S HOSPITAL

HARVARD MEDICAL SCHOOL

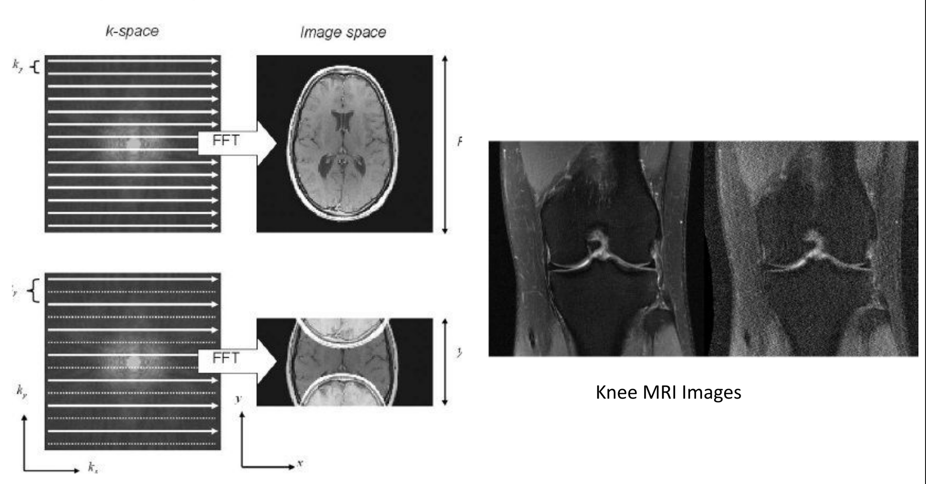
A Deep Learning Approach Using Transformers for MRI Reconstruction of Undersampled k-spaces

Kyler Larsen[1,2], Arghya Pal[2], Yogesh Rathi[2]

A&M Consolidated High School, College Station, TX, USA[1], Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA [2]

Introduction

- Modern Magnetic Resonance Imaging (MRI) scans are time consuming and precarious, since the patients must remain still in a confined space for extended periods of time.
- Experts have experimented with undersampled k-spaces, trying to use deep learning to predict the fully sampled result.



Knee MRI Images

Fully sampled vs Undersampled K-space

- Current MRI reconstruction primarily makes use of the deep learning architecture UNet (Ernst et al 2021). UNet is a model created in 2015; while the models are updated and can still be accurate, new architectures have become more advanced.
- None of these studies experiment with masked image modeling for prediction/reconstruction. This study makes use of Masked Image Modelling through a modified version of the Simple Masked Image Modeling (SimMIM) architecture.
- This study hypothesizes that due to its superior ability to extract features from patch-sized images, Masked Image Modeling will be able to accurately reconstruct MRI images from undersampled k-spaces simulated through masking.

Methods

- This study makes use of knee images from Facebook's fastmri dataset, split 80/20.

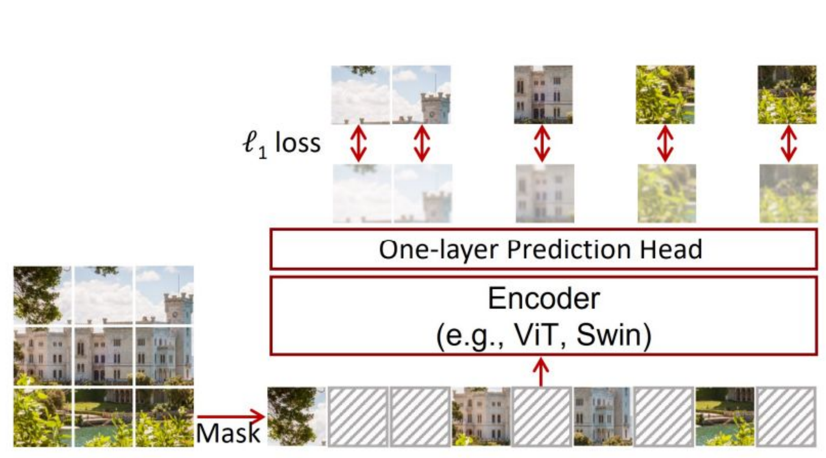
	Volumes		Slices	
	Multi-coil	Single-coil	Multi-coil	Single-coil
training	973	973	34,742	34,742
validation	199	199	7,135	7,135
test	118	108	4,092	3,903
challenge	104	92	3,810	3,305

fastmri dataset distribution paired with MR image

- The data was then augmented by random cropping/stretching to reduce overfitting probabilities.
- Since the baseline model was built to classify, it had to be re-engineered for prediction optimization.
- To experiment, several different encoders were tested and hyperparameters were changed to find the optimal values.
- The masking and patch functions were also modified to better preserve details by reducing parameters like size or masking ratio.
- The model was evaluated on metrics of L1-loss, gradient normalization, and structural similarity for both training and validation after each change.

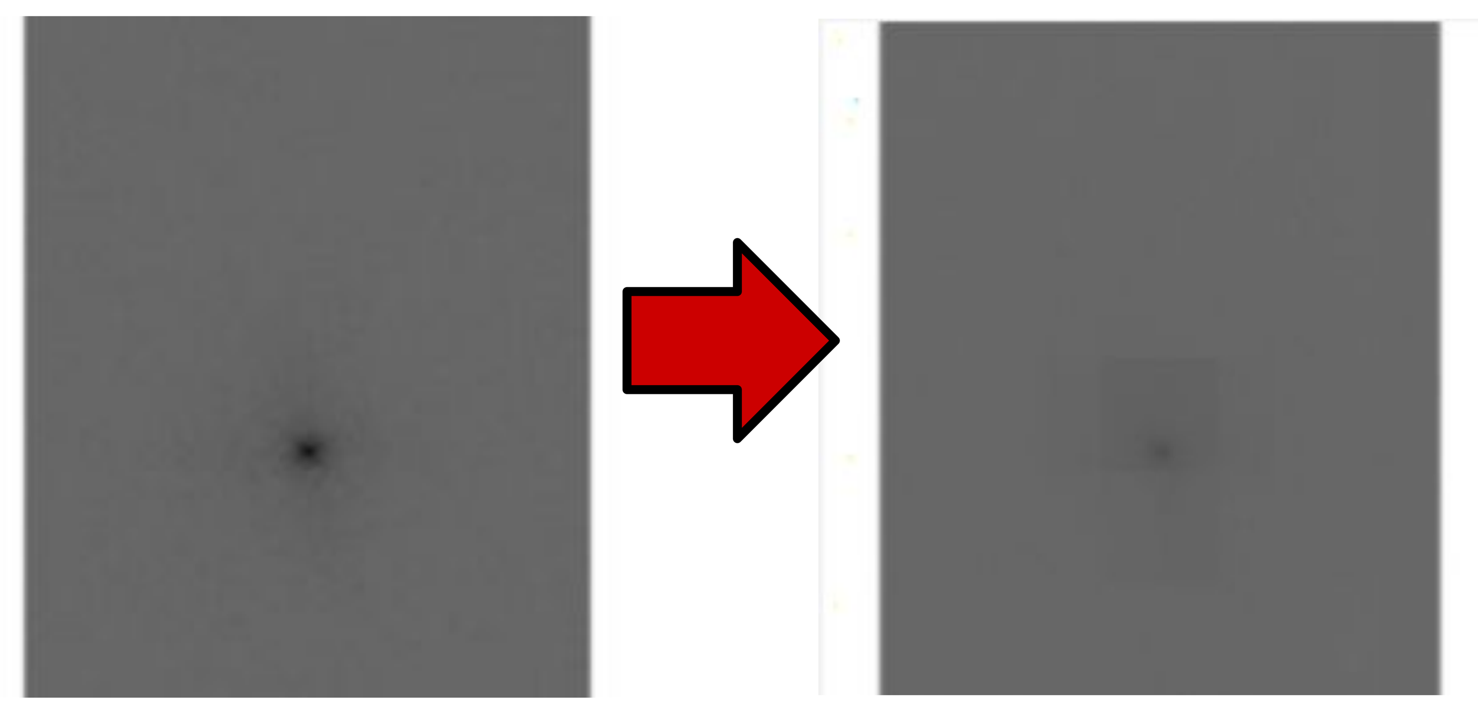
$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad ; \quad \sum_{i=1}^{n} |y_{true} - y_{predicted}|$$

- Parameters were then adjusted and reevaluated until optimal image and metric results were met.



SimMIM Architecture (Xie et al 2019)

Results



Fully Sampled K-space

Reconstructed K-space

The model was trained on approximately 5580 of the knee MRI images, with 1372 used for validation. The above images show a fully sampled k-space compared to the model's predicted k-space from the simulated undersampling. From the model, the structural similarity of the reconstructed k-spaces reached over 99.5%, such as the one shown above, with validation loss values of under 0.01.

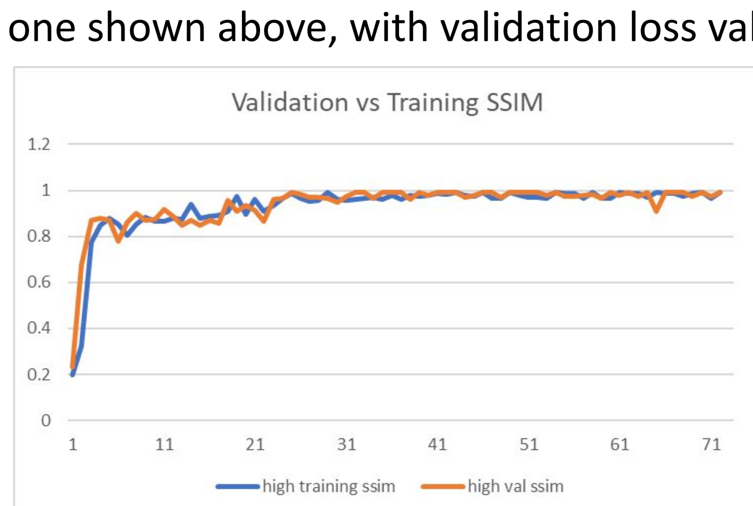


Fig 1.1 (Swin Encoder)

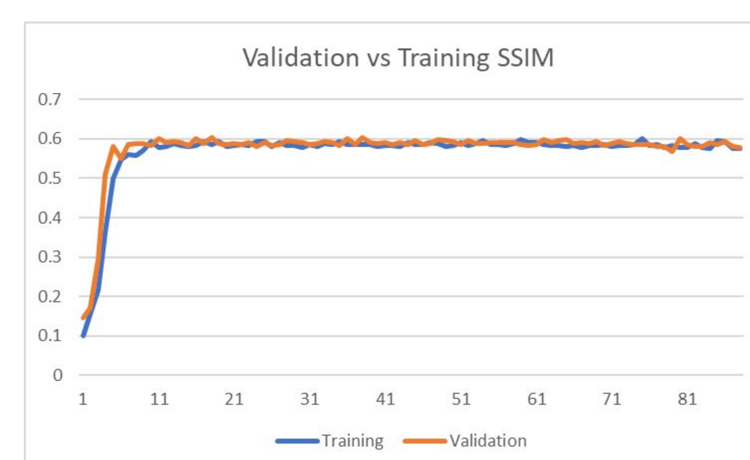


Fig 1.2 (ViT Encoder)

Figures 1.1 and 1.2 (above) illustrate the trend of structural similarity (SSIM)over time for two encoders used in the study. From the left graph, representing the Swin encoder, both the validation and training SSIM increased quickly and remained above 90% from after epoch 5, and continually increased by small amounts before reaching its highest values of >99.5%. The right figure illustrates the use of a Vision (ViT) encoder. The figure shows that although the model improved early, the SSIM flattened at under 60%, meaning the model was not performing well on the reconstruction.

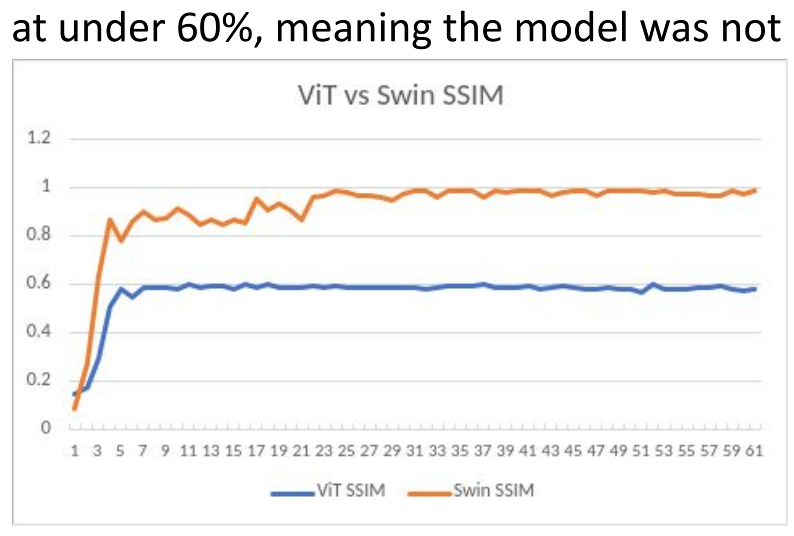


Figure 2.1: Vision vs Swin encoder validation structural similarity

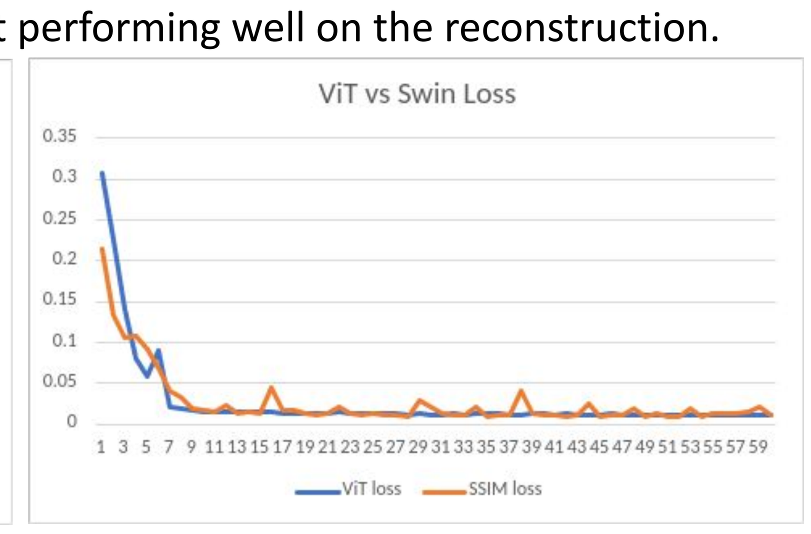


Figure 2.2: Vision vs Swin validation loss values

The above figures 2.1 and 2.2 illustrate a direct comparison between the two encoders. It is evident the Swin Encoder performs more optimally by the SSIM graph alone. However, this coincidentally finds that loss is not an absolute metric for MRI reconstruction since the Swin and ViT have similar loss trends and values even though they produce significantly different outputs. Figures 3.1 and 3.2 below also illustrate the prevalence of overfitting in the Vision Encoder, meaning it's performance was hampered by dataset memorization.

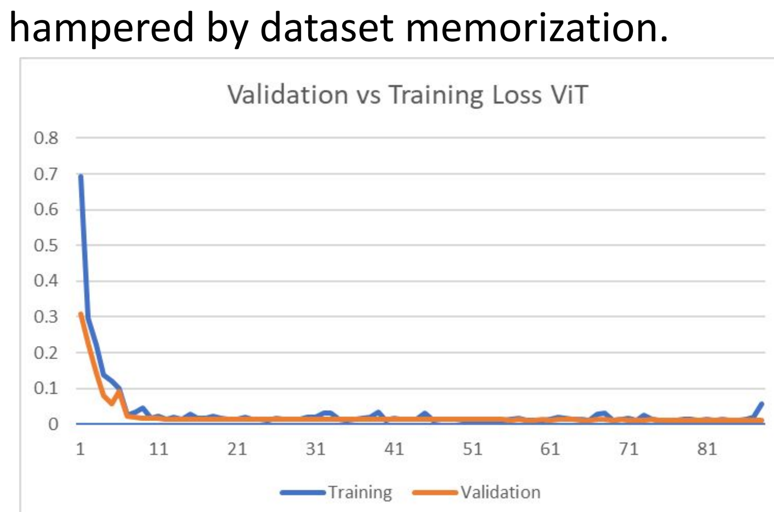


Figure 3.1

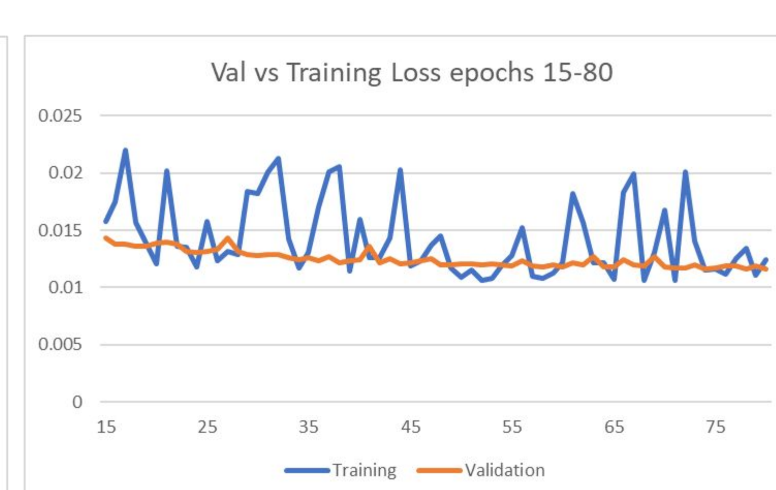
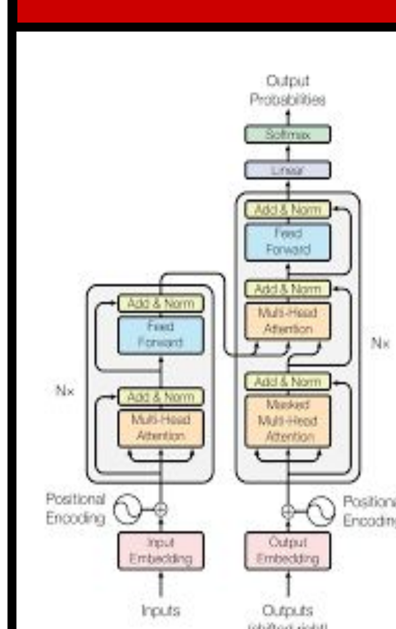


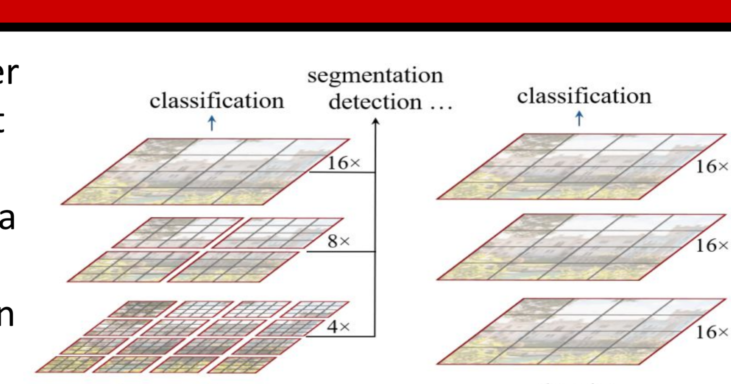
Figure 3.2

Encoder-Decoder Architecture



Basic Encoder Architecture (Xie et al 2019)

The function of an encoder is to extract the important features (attention) from an image and format it in a readable way for the decoder. The decoder then takes this input and outputs a prediction.



Swin vs Vision encoder architectures (Xie et al 2019)

The Swin Transformer works by continually splitting the image into patches, reconstructing the missing parts in the individual patches, and then remerging the patches and completing reconstruction. The Vision transformer works similarly, but does not prioritize reconstruction, rather attention and positional embedding (for classification).

Discussion

- The model performed the best at reconstructing the extremities of the k-space compared to the center
- While both models continually improved, the Swin encoder-based model far outperformed the vision encoder for reconstruction.
- The ViT encoder slightly overfit while the Swin encoder never did, as shown in Figure 4, supporting the idea that the Swin model works for MRI reconstruction

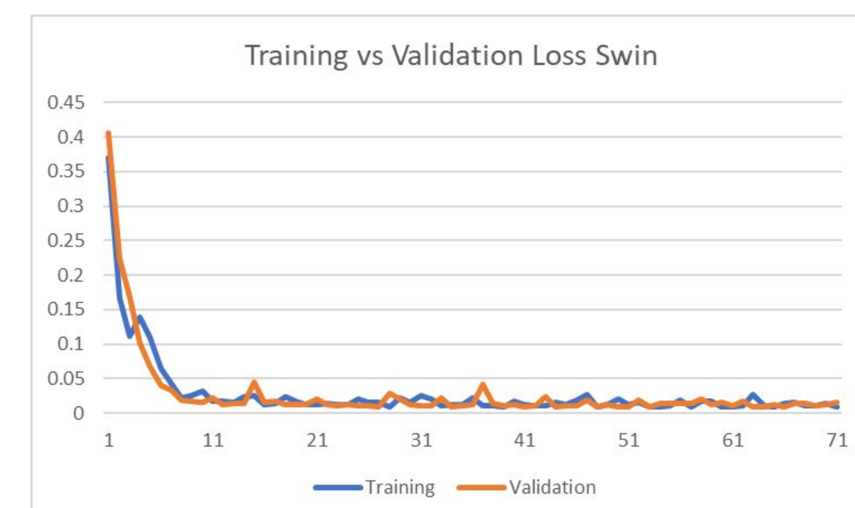


Figure 4

Conclusion

The hypothesis of this study was that the Masked Image Modeling architecture would perform well on k-space reconstruction due to its superior feature extraction. This means that the null hypothesis would be that the model performs subpar on MRI reconstruction. Overall, the model performed well on reconstructing the extremities which control the fine details. The Swin transformer performed significantly better than the Vision transformer as the primary encoder, producing SSIM values almost 40% greater. The production of reconstructed k-spaces more than 99.5% similar to the original, fully sampled k-space provides enough evidence to reject the null hypothesis, meaning this study concludes that the MIM model does work for basic k-space reconstruction.

References

Blaimer, M., Breuer, F., Mueller, M., Heidemann, R. M., Griswold, M. A., & Jakob, P. M. (2004). SMASH, SENSE, PILS, GRAPPA: how to choose the optimal method. *Topics in magnetic resonance imaging : TMRI*, *15*(4), 223–236. https://doi.org/10.1097/01.rmr.0000136558.09801.dd

Moratal, D., Vallés-Luch, A., Martí-Bonmatí, L., & Brummer, M. (2008). k-Space tutorial: an MRI educational tool for a better understanding of k-space. *Biomedical imaging and intervention journal*, *4*(1), e15. https://doi.org/10.2349/biij.4.1.e15

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Arxiv.* doi:10.48550/ARXIV.1505.04597

Ernst, P., Chatterjee, S., Rose, G., Speck, O., & Nürnberger, A. (2021). *Sinogram upsampling using Primal-Dual UNet for undersampled CT and radial MRI reconstruction*. doi:10.48550/ARXIV.2112.13443

Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M. J., … Lui, Y. W. (2018). *fastMRI: An Open Dataset and Benchmarks for Accelerated MRI*. doi:10.48550/ARXIV.1811.08839

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., … Hu, H. (2021). *SimMIM: A Simple Framework for Masked Image Modeling*. doi:10.48550/ARXIV.2111.09886

Acknowledgements

I would like to thank Dr. Arghya Pal and Dr. Kevin Cho for helping me through this project. I would also like to thank the Harvard Medical School and Boston University for allowing me this opportunity to conduct laboratory research. Lastly, I would like to thank my parents and family for continuing to nurture me and support my education.