Lost in Translation: The Critical Role of Attention in Adaptive Image Reconstruction

L.T.N. Wickremasinghe,[1] D.S.P.A. Niwarthana,[1] P.M.P.H. Somarathne,[1] A. Thieshanthan,[1] Brandon Weissbourd, [2] C.U.S. Edussooriya,[1,+] and Dushan N. Wadduwage [3,*]

[1] *Department of Electronic and Telecommunication Engineering, University of Moratuwa, Sri Lanka*
[2]*The Picower Institute for Learning and Memory, Massachusetts Institute of Technology, USA*
[3] *Centre for Advanced Imaging, Faculty of Arts and Sciences, Harvard University, Massachusetts, USA*
[+] *chamira@uom.lk*
[*] *wadduwage@fas.harvard.edu*

Abstract: We propose a novel attention-based deep learning framework for reconstructing images from compressively sensed measurements. Our method leverages knowledge of the sampling basis and structured sampling patterns, highlighting the importance of learning adaptive features from measurements.

1. Main Text

Compressive Sensing (CS) is transforming the way we acquire and process data by enabling high-quality reconstruction from a limited number of measurements by exploiting signal sparsity. The integration of deep learning models into compressive sensing has led to significant advances in recent years. By drawing inspiration from traditional convex optimization algorithms and incorporating ideas of projected gradient descent, these models are now capable of producing high-quality image reconstructions with impressive accuracy. These exciting developments represent a powerful fusion of two cutting-edge technologies providing a promising avenue for achieving high-quality image reconstructions in numerous fields including medical imaging.

Recent research has shown that traditional projected gradient algorithms such as iterative shrinkage-thresholding algorithm (ISTA) can be unrolled to learn flexible models for image reconstruction. [1,2]. However, we recognize that these approaches have led to the CS problem being framed as an image-to-image translation problem, which may not be optimal for all applications. The intuition follows from a single pixel camera setup, where a sampling matrix, with each row representing a particular measurement vector, is used to obtain a single pixel measurement through a linear mapping. Given an imaging specimen, the set of measurements available from sampling do not contribute equal information. This highlights the importance of developing algorithms/models that can adaptively pay more attention to certain measurements than others to solve the measurement-to-image problem intelligently.

The ideas for adaptively deciding the degree of relative importance of each measurement naturally fits into the work of transformers [3] and attention mechanisms. Particularly, the image-to-image tasks such as segmentation have been improved with Vision Transformers [4]. Extending this idea beyond the image-pixel domain, we develop a model that is adaptable to measurements, where features learned through attention can faithfully reconstruct images from whichever measurements are available. This approach ensures that the learning is not limited by a specific sampling scheme, and that the model can fill in "measurement-gaps" regardless of how samples are obtained. Treating each sampled measurement as a representative of the sampled image and the measurement vector from a particular sampling basis; we decouple the measurement value from this sensing vector, and thereby learn weights adaptively. By doing so, we improve the learning directly from the measurements, which would not be possible without the attention mechanism. This, as we validate through our experiments, plays a critical role in significantly improving the accuracy of the image reconstruction, enabling faithful reconstructions even around 16x compression.

We see that our architecture is able to contextually generate embeddings that are agnostic to the sampling algorithm and contains a sufficient information flow to generate a faithful reconstruction. The model is trained on the CIFAR-10 dataset [5], which contains 32×32 pixel images. During the training, images are sampled randomly by selecting rows of the sampling matrix. The measurement, along with the selected sensing vector set is then input into our model, which predicts the image in pixel space. The experiments that complement the model design confirms our hypothesis, that such an adaptive learning improves the reconstruction performance. Figure 1 depicts some cross platform use-cases of our architecture. In each case, the same model is tested without task-specific fine-tuning, showing greater generalizability.

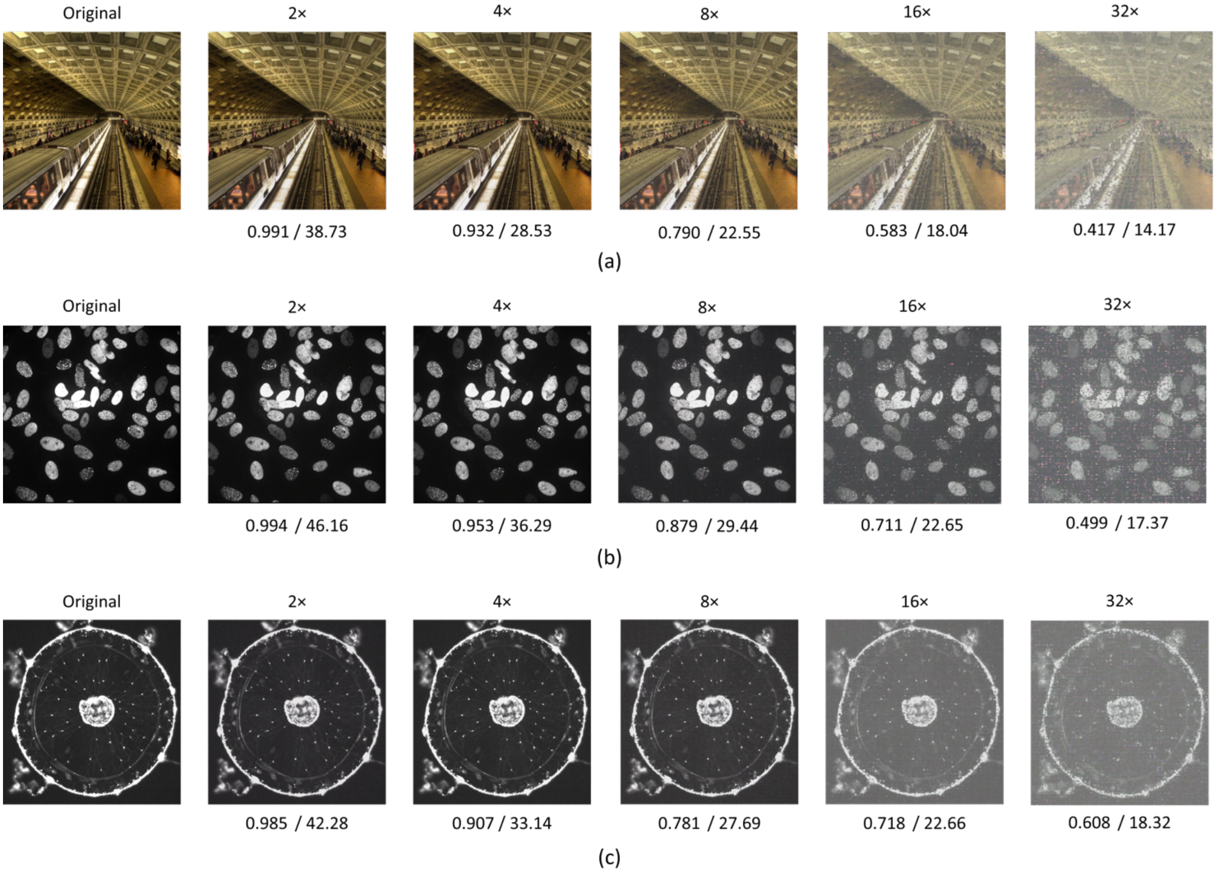

Fig. 1. Reconstruction of multi modal images, sampled at various compression factors. The Structured Similarity Index Measure (SSIM)/ Peak Signal to Noise Ratio (PSNR) is reported below each image. (a) Multichannel images from the Urban100 [6] image super-resolution dataset to validate our approach for macro-scale imaging systems. (b) Bone Osteosarcoma cell imaged using a spinning disk confocal microscope at 63× magnification using an objective with 1.4 numerical aperture. (c) Jellyfish images acquired using a wide-field fluorescence microscope.

References

1. J. Zhang and B. Ghanem, "ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* (IEEE Computer Society, Los Alamitos, CA, USA, 2018), pp. 1828–1837.
2. C. A. Metzler, A. Mousavi, and R. G. Baraniuk, "Learned D-AMP: Principled neural network based compressive image recovery." in *NIPS,* I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, eds. (2017), pp. 1772–1783.
3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems,* (Curran Associates Inc., Red Hook, NY, USA, 2017), NIPS'17, p. 6000–6010.
4. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations,* (2021).
5. A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. rep., University of Toronto, Department of Computer Science (2009).
6. J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* (2015), pp. 5197–5206.