Diagnostically Lossless Compression of Medical Images

Anonymous Authors[1]

Abstract

Medical images (e.g. X-rays) are often acquired at high resolutions with large dimensions in order to capture fine-grained details. In this work, we address the challenge of compressing medical images while preserving fine-grained features needed for diagnosis, a property known as *diagnostic losslessness*. To this end, we (1) use over one million medical images to train a domain-specific neural compressor and (2) develop a comprehensive evaluation suite for measuring compressed image quality. Extensive experiments demonstrate that large-scale, domain-specific training of neural compressors improves the diagnostic losslessness of compressed images when compared to prior approaches.

1. Introduction

Medical images are essential diagnostic tools in clinical practice. Since medical conditions are often characterized by the presence of small features (e.g. microcalcifications, fractures), images are acquired with high spatial resolution in order to capture the required level of detail (Huda & Abrahams, 2015). However, high-resolution medical images often have large dimension sizes, particularly when covering a large anatomical region; this can pose a problem for computer-aided diagnosis (CAD) by resulting in increased or even intractable computational complexity (Freire et al., 2022; Tan & Le, 2019). As a result, effective compression approaches are necessary for enabling computationally feasible analysis of medical images.

Previous strategies for lossy compression include (1) *storage methods*, which optimize for low bitrates (e.g. JPEG2000), and (2) *scaling methods*, which compress images into structured representations with reduced input dimensions (e.g. neural compressors). Representations generated by storage methods generally need to be decompressed before subsequent usage and do not improve computational efficiency of CAD; consequently, we focus solely on scaling methods in this work. In particular, recent work has demonstrated that large-scale neural compressors (e.g. autoencoders) trained on millions of natural images can effectively compress images into downsized latent representations while preserving key visual features, leading to improvements in downstream computational efficiency (Rombach et al., 2022).

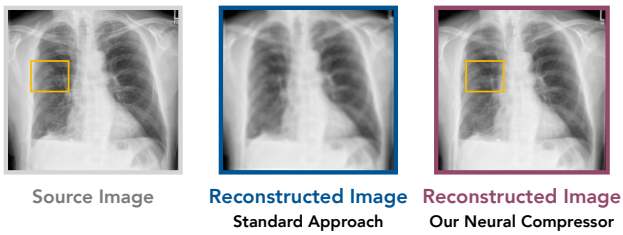

Figure 1. We train a neural compressor that better preserves diagnostic features in medical images (e.g. fracture shown in yellow).

However, medical image compression is complicated by the need for *diagnostic losslessness*, meaning that compressed images should preserve all features needed for diagnosis (European Society of Radiology, 2011). Whereas natural images generally only require the preservation of larger, global features for accurate image interpretation, medical images consist of small, fine-grained features that must be retained in order to enable effective clinical diagnoses. As a result, there are two key challenges associated with applying large-scale neural compressors to the medical image domain. First, existing neural compressors are trained on natural images, which represent a significant domain shift from medical images. In particular, the role of domain-specific training on performance of large-scale neural compressors remains unclear. Second, evaluating compression approaches with respect to diagnostic losslessness is challenging due to the lack of available evaluation suites. Prior works predominantly evaluate the quality of compressed images using perceptual metrics, which do not specifically measure the preservation of clinically-relevant features.

In this work, we address these challenges by introducing the first large-scale domain-specific variational autoencoder (VAE) designed for compression of high-resolution medical images. We use over one million medical images to

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

train several domain-specific VAEs at various levels of compression. Then, we introduce a suite of quantitative and qualitative metrics for evaluating our method with respect to diagnostic losslessness. While compression methods are typically evaluated from a rate-distortion perspective (Shannon, 1948), we instead propose a benchmark that includes 5 fine-grained classification tasks and an expert reader study in addition to standard perceptual quality assessments.

Our experiments demonstrate that domain-specific training of neural compressors improves the diagnostic losslessness of compressed images. When comparing compressed images generated by our domain-specific VAE to those from existing neural compressors, we obtain an average performance improvement of 5.7% across our fine-grained classification tasks. Our expert reader study confirms these findings qualitatively. Additionally, we show that commonly-used perceptual metrics are insufficient for measuring diagnostic losslessness, demonstrating the critical need for finer-grained evaluations.

Related Work: Prior works on neural compression have predominantly focused on the natural image setting (Rombach et al., 2022; Dubois et al., 2021). In the medical domain, some previous works have applied various neural compression techniques to X-rays, MRI scans, and pathology images (Dupont et al., 2022; Tellez et al., 2019; Sushmit et al., 2019; Tellez et al., 2020); however, no prior studies have explored large-scale training of neural compressors on medical images. Additionally, prior medical compression methods are generally evaluated with rate-distortion and perceptual quality metrics (Dupont et al., 2022; Sushmit et al., 2019). No systematic evaluation framework for diagnostic losslessness has been previously introduced for medical images.

2. Our Approach

In this section, we introduce our domain-specific VAE for medical image compression (Section 2.1). We also discuss our evaluation suite for benchmarking compression methods with respect to diagnostic losslessness (Section 2.2).

2.1. Neural Compression of Medical Images

We begin with a training dataset $\mathcal{D} = \{x_i\}_{i=1}^n$ consisting of n grayscale medical images $x_i \in \mathcal{X}$. Each image x_i has dimensions $H \times W$ with a single channel, expressed as $x_i \in \mathbb{R}^{H \times W \times 1}$. Our goal is to learn a stochastic mapping $g : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ represents a low-dimensional latent space. Specifically, for a compression factor f and image x_i, a latent sample $z_i \in \mathcal{Z}$ has dimension $(H/f) \times (W/f) \times C$, where C represents a pre-specified number of channels. In combination with g, we also learn a mapping $h : \mathcal{Z} \to \hat{\mathcal{X}}$, which reconstructs the image from the latent sample.

Motivated by prior work on large-scale neural compressors (Rombach et al., 2022), we learn functions g and h using a fully convolutional VAE with a combination of a perceptual loss, a patch-based adversarial objective, and a penalty based on the Kullback-Leibler (KL) divergence. We train six neural compressors with varying values of f (4, 8, and 16) and C (1, 4, and 16).

The training dataset $\mathcal{D}$ consists of X-ray data from two modalities: chest X-rays and full-field digital mammograms (FFDM). We select these modalities because (a) chest X-rays are well-studied with large amounts of publicly-available data and (b) FFDMs are a challenging class of images due to large dimensions and the presence of fine-grained features critical for diagnoses (e.g. microcalcifications). We use images from two chest X-ray datasets and six FFDM datasets: MIMIC-CXR (Johnson et al., 2019), CANDID-PTX (Feng et al., 2021), EMBED (Jeong et al., 2022), CSAW-CC (Sorkhei et al., 2021), RSNA Mammography (RSNA, 2023), VinDr-Mammo (Nguyen et al., 2022), INBreast (Moreira et al., 2012), and CMMD (Cai et al., 2023). The final dataset comprises 1,021,356 images.

2.2. Evaluation Suite

We use three evaluation tasks for quantitatively and qualitatively assessing compression methods with respect to diagnostic losslessness: (1) fine-grained classification, (2) expert reader study, and (3) perceptual quality metrics.

First, we evaluate the preservation of fine-grained features in compressed images z with five classification tasks: malignancy detection, calcification identification, and BI-RADS classification on FFDMs (Nguyen et al., 2022; Cai et al., 2023); bone age prediction on hand X-rays (Halabi et al., 2019); and fracture detection on pediatric wrist radiographs (Nagy et al., 2022). Classification accuracy is assessed using a high-resolution network (HRNet) (Wang et al., 2020) with supervised linear probing (Zhang et al., 2016).

Next, we qualitatively assess the information loss resulting from compression. We perform a study with two expert radiologists, where each reader is presented with 50 unique reconstructed chest X-rays ($\hat{x}$) paired with the ground-truth image (x). All X-rays include at least one fracture. Ratings are given on a 5-point Likert-scale for image fidelity, diagnostic losslessness, and the presence of artifacts.

Finally, we evaluate image fidelity by using standard perceptual quality metrics to compare reconstructed images $\hat{x}$ with the original x. We report Fréchet Inception Distance (FID), peak signal-to-noise ratio (PSNR), and multi-scale structural similarity index measure (MS-SSIM). Since FID is computed using an Inception V3 network (Szegedy et al., 2015) that is not adapted for medical images, we introduce two variants based on CLIP and BiomedCLIP (Radford et al., 2021; Zhang et al., 2023; Chambon et al., 2022).

Method	Compression				AUROC ↑				
	f	C	γ	FLOPS	Malignancy	Calcification	BI-RADS	Bone Age	Wrist Fracture
Full Size	1	1	1	607.33	65.5	63.3	63.4	80.4	73.7
Bicubic Interpolation	4	1	16	37.96	**65.9**	58.9	61.3	81.8	**71.0**
SD VAE	4	3	5.3	37.98	65.6	54.6	59.5	75.9	66.9
Ours	4	1	16	37.96	65.8	**62.2**	**63.2**	81.7	70
Ours	4	4	4	37.99	65.4	60.9	61.3	**82.5**	69.7
Bicubic Interpolation	8	1	64	9.49	61.5	56.9	61.2	73.0	**67.9**
SD VAE	8	4	16	9.50	59.6	57.3	57.3	67.7	61.9
Ours	8	1	64	9.49	**62.6**	**59.1**	**61.9**	**73.8**	65.4
Ours	8	4	16	9.50	61.5	57.5	59.4	67.2	62.1
Bicubic Interpolation	16	1	256	2.37	**59.2**	53.4	**59.7**	**63.9**	**61.2**
SD VAE	16	16	16	2.38	56.4	52.7	56.0	60.5	56.9
Ours	16	1	256	2.37	59.0	**55.1**	58.9	62.4	59.9
Ours	16	16	16	2.38	54.5	53.5	56.0	62.3	56.2

Table 1. Fine-grained classification results. We compare our trained neural compressor (Ours) to a conventional image downsizing approach (Bicubic Interpolation) as well as an existing large-scale neural compressor trained on natural images (SD VAE). Here, f represents the compression factor per dimension, C represents the number of latent channels, γ represents the compression ratio, and FLOPS represent GigaFLOPS using a high-resolution network (HRNet_w64) with a single output class.

3. Experiments

We use our evaluation suite to compare our domain-specific neural compressor to a conventional image downsizing approach as well as an existing large-scale neural compressor trained on eight million natural images (Rombach et al., 2022). Our experiments show that (1) domain-specific training yields compressed images that better capture fine-grained features, (2) expert radiologists qualitatively confirm these findings, and (3) commonly-used perceptual metrics do not effectively measure diagnostic losslessness.

3.1. Fine-Grained Classification

We evaluate the quality of compressed images across 5 classification tasks that measure the preservation of fine-grained features. In Table 1, we compare three methods: bicubic interpolation, a large-scale neural compressor trained on natural images (SD VAE), and our domain-specific neural compressor (Ours). We evaluate each approach with three different compression factors f (4, 8, and 16) and varying numbers of latent channels C. The original, full-size input images are 1024 pixels along the longest dimension, meaning that compressed images range in size from 256 pixels ($f = 4$) to 32 pixels ($f = 16$) along the longest dimension.

As shown in Table 1, our domain-specific VAE consistently outperforms the SD VAE on fine-grained classification tasks. On average across the 5 tasks, our domain-specific VAE demonstrates a 6.33% improvement over the SD VAE at a compression factor of 4, a 6.25% improvement at a compression factor of 8, and a 4.50% improvement at a compression factor of 16. Additionally, our domain-specific VAE outperforms bicubic interpolation across most tasks at $f = 4$ and $f = 8$; however, bicubic interpolation is consistently superior at $f = 16$. Our findings suggest that domain-specific training of neural compressors is critical for improving diagnostic losslessness of compressed medical images.

However, we note that none of the methods evaluated in this study exhibit perfect diagnostic losslessness. Across the majority of tasks, we observe significant drops in performance between classification models trained with full-size images and those trained with compressed images (e.g. up to 20 points on bone age classification). Additionally, we observe that increasing C decreases the compression ratio yet leads to performance degradations; this suggests that the classification models are unable to effectively reason over the extra information stored in the latent channels. Our findings demonstrate the need for more effective compression methods as well as downstream models that can effectively reason over multi-channel latent samples.

Method	Compression			Perceptual Metrics				
	f	C	γ	FID-Inc ↓	FID-CLIP ↓	FID-BiomedCLIP ↓	PSNR ↑	MS-SSIM ↑
Bicubic Interpolation	4	1	16	$28.03_{\pm 0.20}$	$7.52_{\pm 0.03}$	$227.20_{\pm 3.55}$	$33.48_{\pm 0.08}$	$0.973_{\pm 0.00}$
SD VAE	4	3	5.3	$3.23_{\pm 0.03}$	$4.36_{\pm 0.03}$	$\mathbf{3.66}_{\pm 0.09}$	$38.37_{\pm 0.08}$	$0.992_{\pm 0.00}$
Ours	4	1	16	$6.28_{\pm 0.02}$	$0.36_{\pm 0.02}$	$10.00_{\pm 0.17}$	$33.45_{\pm 0.10}$	$0.973_{\pm 0.00}$
Ours	4	4	4	$\mathbf{2.57}_{\pm 0.02}$	$\mathbf{0.10}_{\pm 0.01}$	$5.04_{\pm 0.04}$	$\mathbf{38.85}_{\pm 0.12}$	$\mathbf{0.996}_{\pm 0.00}$
Bicubic Interpolation	8	1	64	$79.29_{\pm 0.47}$	$13.13_{\pm 0.07}$	$840.56_{\pm 1.18}$	$29.81_{\pm 0.08}$	$0.913_{\pm 0.00}$
SD VAE	8	4	16	$7.37_{\pm 0.03}$	$4.30_{\pm 0.03}$	$\mathbf{12.10}_{\pm 0.35}$	$\mathbf{33.80}_{\pm 0.10}$	$0.971_{\pm 0.00}$
Ours	8	1	64	$19.09_{\pm 0.39}$	$0.99_{\pm 0.02}$	$45.64_{\pm 1.39}$	$29.71_{\pm 0.07}$	$0.926_{\pm 0.00}$
Ours	8	4	16	$\mathbf{6.57}_{\pm 0.07}$	$\mathbf{0.38}_{\pm 0.00}$	$13.51_{\pm 0.30}$	$33.35_{\pm 0.09}$	$\mathbf{0.973}_{\pm 0.00}$

Table 2. Perceptual quality assessments. We compare our trained neural compressor (Ours) to a conventional image downsizing approach (Bicubic Interpolation) as well as an existing large-scale neural compressor trained on natural images (SD VAE). We report five metrics that measure the perceptual quality of the reconstructed image.

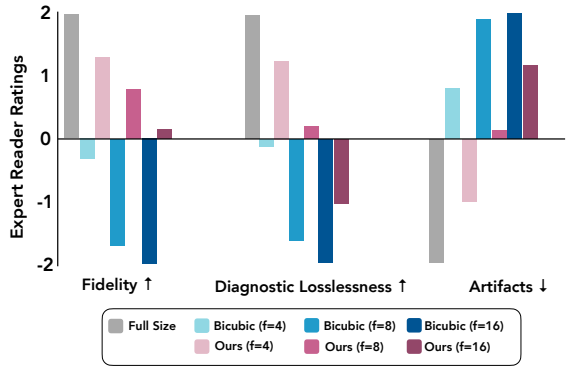

Figure 2. Results from expert reader study.

3.2. Expert Reader Study

We qualitatively assess image fidelity, diagnostic losslessness, and the presence of artifacts for bicubic interpolation ($f \in \{4, 8, 16\}$) and our domain-specific VAEs with equivalent compression factors (Figure 2). Image fidelity for our neural compressor was 2.1 points higher than bicubic interpolation ($p < 0.05$). Diagnostic losslessness scores also favored our neural compressor by 1.37 points ($p < 0.05$). Artifacts (e.g. blurring, hallucinations) were more frequent in interpolated images (+1.48 points; $p < 0.05$). Our results suggest that our neural compressor better preserves critical diagnostic features than bicubic interpolation, a conventionally-used downsizing approach.

3.3. Perceptual Quality Assessments

We evaluate image fidelity by comparing original and reconstructed images using standard perceptual quality metrics. We evaluate three compression methods (bicubic interpolation, SD VAE, and our domain-specific compressor) with two different compression factors f (4 and 8). As shown in Table 2, we observe that across most metrics, our domain-specific VAE outperforms both bicubic interpolation and SD VAE. We also note a general trend that increasing C improves image perceptual quality.

However, the perceptual quality metrics exhibit some inconsistencies. In particular, results in Table 2 suggest that SD VAEs offer better image fidelity than bicubic interpolation, yet our analysis in Section 3.1 demonstrates the opposite: bicubic interpolation better captures important diagnostic features. Similarly, our domain-specific VAE trained with $f = 4$ and $C = 1$ achieves similar PSNR and MS-SSIM scores to bicubic interpolation with $f = 4$; however, our results from Section 3.2 show that radiologists perceive our VAE to exhibit better fidelity. These findings suggest that perceptual quality metrics, which are conventionally used to evaluate compression approaches, are inadequate for capturing diagnostic losslessness and should be supplemented with finer-grained evaluations.

4. Discussion

In this work, we explore neural compression on medical images with the goal of achieving diagnostic losslessness. To this end, we introduce (1) a large-scale domain-specific VAE for compression of high-resolution medical images and (2) a suite of quantitative and qualitative metrics for evaluating compressed image quality with respect to diagnostic losslessness. Our results demonstrate that large-scale, domain-specific training of neural compressors improves the quality of compressed medical images. Future directions include expanding our evaluation suite to additional tasks and modalities and designing classification methods to better capture signal from multi-channel latents.

References

Cai, H., Wang, J., Dan, T., Li, J., Fan, Z., Yi, W., Cui, C., Jiang, X., and Li, L. An online mammography database with biopsy confirmed types. *Scientific Data*, 10(1):123, 2023.

Chambon, P., Bluethgen, C., Delbrouck, J.-B., Van der Sluijs, R., Połacin, M., Chaves, J. M. Z., Abraham, T. M., Purohit, S., Langlotz, C. P., and Chaudhari, A. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.

Dubois, Y., Bloem-Reddy, B., Ullrich, K., and Maddison, C. J. Lossy compression for lossless prediction. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14014–14028. Curran Associates, Inc., 2021.

Dupont, E., Loya, H., Alizadeh, M., Goliński, A., Teh, Y. W., and Doucet, A. Coin++: Data agnostic neural compression. *arXiv preprint arXiv:2201.12904*, 2022.

European Society of Radiology, . Usability of irreversible image compression in radiological imaging. A position paper by the European Society of Radiology (ESR). *Insights Imaging*, 2(2):103–115, Apr 2011.

Feng, S., Azzollini, D., Kim, J. S., Jin, C.-K., Gordon, S. P., Yeoh, J., Kim, E., Han, M., Lee, A., Patel, A., et al. Curation of the candid-ptx dataset with free-text reports. *Radiology: Artificial Intelligence*, 3(6):e210136, 2021.

Freire, P. J., Srivallapanondh, S., Napoli, A., Prilepsky, J. E., and Turitsyn, S. K. Computational complexity evaluation of neural network applications in signal processing. *arXiv preprint arXiv:2206.12191*, 2022.

Halabi, S. S., Prevedello, L. M., Kalpathy-Cramer, J., Mamonov, A. B., Bilbily, A., Cicero, M., Pan, I., Pereira, L. A., Sousa, R. T., Abdala, N., Kitamura, F. C., Thodberg, H. H., Chen, L., Shih, G., Andriole, K., Kohli, M. D., Erickson, B. J., and Flanders, A. E. The rsna pediatric bone age machine learning challenge. *Radiology*, 290(2):498–503, 2019. doi: 10.1148/radiol.2018180736. PMID: 30480490.

Huda, W. and Abrahams, R. B. X-ray-based medical imaging and resolution. *American Journal of Roentgenology*, 204(4):W393–W397, 2015.

Jeong, J. J., Vey, B. L., Reddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Ilies, G., Smith, G., et al. The emory breast imaging dataset (embed): A racially diverse, granular dataset of 3.5 m screening and diagnostic mammograms. *arXiv preprint arXiv:2202.04073*, 2022.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., and Cardoso, J. S. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

Nagy, E., Janisch, M., Hržić, F., Sorantin, E., and Tschauner, S. A pediatric wrist trauma x-ray dataset (grazpedwri-dx) for machine learning. *Scientific Data*, 9(1):222, May 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01328-z.

Nguyen, H. T., Nguyen, H. Q., Pham, H. H., Lam, K., Le, L. T., Dao, M., and Vu, V. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *MedRxiv*, pp. 2022–03, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

RSNA, . Rsna screening mammography breast cancer detection. *Kaggle.com*, 2023.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Sorkhei, M., Liu, Y., Azizpour, H., Azavedo, E., Dembrower, K., Ntoula, D., Zouzos, A., Strand, F., and Smith, K. Csaw-m: An ordinal classification dataset for benchmarking mammographic masking of cancer. *arXiv preprint arXiv:2112.01330*, 2021.

Sushmit, A. S., Zaman, S. U., Humayun, A. I., Hasan, T., and Bhuiyan, M. I. H. X-ray image compression using convolutional recurrent neural networks. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 1–4. IEEE, 2019.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Tellez, D., Litjens, G., van der Laak, J., and Ciompi, F. Neural image compression for gigapixel histopathology image analysis. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):567–578, 2019.

Tellez, D., Höppener, D., Verhoef, C., Grünhagen, D., Nierop, P., Drozdzal, M., Laak, J., and Ciompi, F. Extending unsupervised neural image compression with supervised multitask learning. In *Medical Imaging with Deep Learning*, pp. 770–783. PMLR, 2020.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 649–666. Springer, 2016.

Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

A. Implementation Details for Our Neural Compressor

A.1. Data Preprocessing

All datasets were collected in Digital Imaging and Communications in Medicine (DICOM) file format. First, we extract the raw pixel data from the DICOM files. We then compute a mask based on iterative binary thresholding and extract the region of interest from the original pixel array. In accordance with the metadata, we either apply the `RescaleIntercept` and `RescaleSlope` DICOM attributes or the Modality LUT. Finally, the images are windowed based on the `WindowCenter` and `WindowWidth` attributes, and inverted based on the `PhotometricInterpretation` if needed. All preprocessing was performed using the *pyvoxel* package for Python (`https://github.com/pyvoxel/pyvoxel`). The final dataset was divided into a training and validation set with non-overlapping patients.

A.2. Training Details

Our domain-specific VAEs were trained for 15000 steps on 16 NVIDIA A100 GPUs across two nodes with a batch size of 64. All training was performed using full precision. To preserve aspect ratio and fine-grained features, we use random (resized) crops of the high-resolution images as inputs.

B. Extended Details on Evaluation Suite

We use three tasks - fine-grained classification, an expert reader study, and perceptual quality metrics - to evaluate compression methods with respect to diagnostic losslessness. Below, we provide extended details on our evaluation suite.

B.1. Fine-Grained Classification

We use the following 5 fine-grained classification tasks to evaluate the preservation of diagnostic features in compressed images:

- *Malignancy Detection*: We classify images from the CMMD mammogram dataset (Cai et al., 2023) into two classes: presence of a malignancy and absence of a malignancy.
- *Calcification Identification*: We classify images from the CMMD mammogram dataset (Cai et al., 2023) into two classes: presence of calcification and absence of calcification.
- *BI-RADS Classification*: We classify images from the VinDR-Mammo mammogram dataset (Nguyen et al., 2022) into five classes corresponding to BI-RADS scores. BI-RADS scores are assigned by clinicians to score mammogram findings on a scale from 1 (no cancer detected) to 5 (>95% likelihood of cancer).
- *Bone Age Prediction*: We classify images from the RSNA Bone Age dataset (Halabi et al., 2019) into 20 classes corresponding to patient age in years (ranging from 0 to 19).
- *Fracture Detection*: We classify pediatric forearm X-rays from the GRAZPEDWRI-DX dataset (Nagy et al., 2022) into two classes: presence of a wrist fracture and absence of a wrist fracture.

We note that the bone age prediction and fracture detection tasks involve images with anatomical regions that do not appear in the training set for our neural compressor; as a result, these serve tasks as effective out-of-distribution evaluation settings.

For each classification task, we divide the dataset into a train and test set, ensuring that no samples in the test set are used for training the neural compressor. We then train a high-resolution convolutional neural network (Wang et al., 2020) with supervised linear probing (Zhang et al., 2016) to perform the classification task. We use an AdamW optimizer with a batch size of 128 and a learning rate of 1e-4, and we train with fp16 precision for 100 epochs across 8 NVIDIA A100 GPUs.