Imitating Radiological Scrolling: A Glocal-Lobal Attention Model for 3D Chest CT Volumes Multi-Label Anomaly Classification

Theo Di Piazza[1,2] THEO.DIPIAZZA@CREATIS.INSA-LYON.FR
Carole Lazarus[3]
Olivier Nempont[3]
Loic Boussel[1,2]
[1] *UCBL1, INSA Lyon, CNRS, INSERM, CREATIS UMR 5220, U1294, Villeurbanne, France*
[2] *Department of Radiology, Croix-Rousse Hospital, Hospices Civils de Lyon, Lyon, France*
[3] *Philips Clinical Informatics, Innovation Paris, France*

Editors: Under Review for MIDL 2025

Abstract

The rapid increase in the number of Computed Tomography (CT) scan examinations has created an urgent need for automated tools, such as organ segmentation, anomaly classification, and report generation, to assist radiologists with their growing workload. Multi-label classification of Three-Dimensional (3D) CT scans is a challenging task due to the volumetric nature of the data and the variety of anomalies to be detected. Existing deep learning classification methods, relying on standard Convolutional Neural Networks or Vision Transformers, do not explicitly model the radiologist's navigational behavior while scrolling through CT scan slices. In this study, we present CT-Scroll, a novel glocal-local attention model specifically designed to emulate the scrolling behavior of radiologists during the analysis of 3D CT scans. Our approach is evaluated on two public datasets, demonstrating its efficacy through comprehensive experiments and an ablation study that highlights the contribution of each model component.

Keywords: Multi-label classification, Computed-Tomography, Attention Mechanism.

1. Introduction

Computed Tomography (CT) provides detailed imaging of the human body, enabling radiologists to thoroughly examine various anatomical regions, identify abnormalities, and guide patient care from initial diagnosis to follow-up (Mazonakis and Damilakis, 2016). However, the growing number of CT scans (Broder and Warshauer, 2006) and the associated workload for radiologists have created a pressing need for automated methods to assist in analyzing these volumes (Chen et al., 2022). In medical imaging, and particularly in CT scans, substantial progress has been made in leveraging deep learning techniques to support radiologists in tasks such as segmentation (Gu et al., 2022), image restoration (Yuan et al., 2023), classification (Draelos et al., 2021), and more recently, report generation (Hamamci et al., 2024b). As illustrated in Figure 1, multi-label anomaly classification from Three-Dimensional (3D) CT volumes remains a challenging task due to the significant variability in the anomalies that need to be detected.

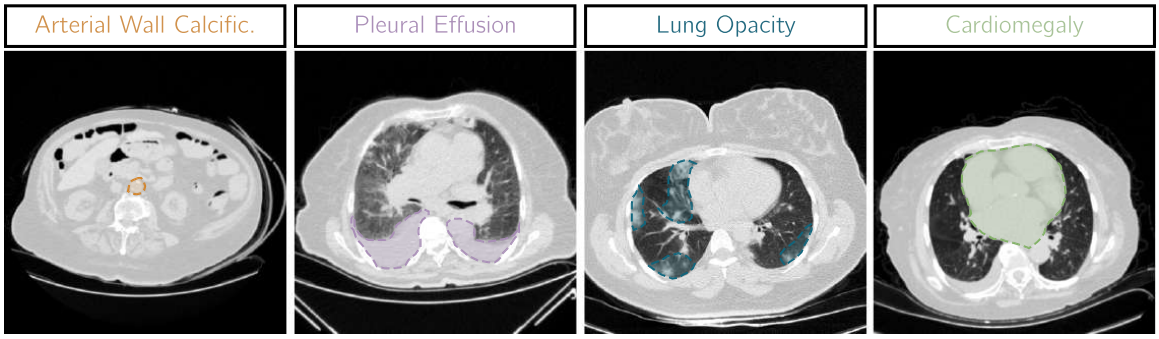

Figure 1: **Examples of 4 axial CT scan slices** with anomalies of varying sizes from the CT-RATE dataset.

To process volumetric data and extract meaningful visual features, early approaches predominantly relied on 3D convolutional neural networks (CNNs) to capture spatial dependencies within volumetric data effectively (Singh et al., 2020). Alternatively, some studies adopted conventional 2D architectures by treating a volume as a sequence of slices and subsequently fusing the extracted features (Draelos et al., 2021). CNNs excel at capturing local spatial features, and their hierarchical structure facilitates the progressive learning of features, from low-level patterns to high-level semantic representations. More recently, attention mechanisms (Vaswani et al., 2023), initially introduced in Natural Language Processing, have demonstrated exceptional performance across diverse text-related tasks (Touvron et al., 2023). This paradigm has been adapted to visual data, including 2D and 3D imaging, by representing images as sequences of 1D tokens derived from flattened 2D or 3D patches. In particular, Vision Transformers (ViTs) (Dosovitskiy et al., 2021) leverage attention mechanisms to model global context by enabling interactions across different regions of an image. This capability is particularly advantageous for applications requiring a comprehensive understanding of global contexts, making ViTs a promising alternative for complex medical imaging tasks. However, the local receptive fields of CNNs limit their ability to capture global contextual information across large 3D volumes, while ViTs can be computationally expensive when applied to high-dimensional volumetric data and often require large-scale pre-training on extensive datasets to achieve competitive performance (Hamamci et al., 2024c). When radiologists analyze a CT scan, they typically navigate through axial slices to have a global understanding of the volume before focusing on specific anatomical regions of interest (Goergen et al., 2013). If an area appears abnormal, the radiologist often revisits the same slices repeatedly, carefully examining the local context to confirm the diagnosis. Inspired by this diagnostic approach and leveraging the strengths of alternating attention (Warner et al., 2024), originally introduced in NLP, we present a novel alternating global-local attention module, termed the *Scrolling Block* (SB), illustrated in Figure 2. This module integrates both global and local information through a Sliding Window Attention (SWA) (Child et al., 2019; Beltagy et al., 2020) mechanism specifically designed for 3D CT volumes. Our contributions are summarized as follows:

- The introduction of a global-local attention model designed to imitate radiology navigation in 3D CT scans, enhancing multi-label anomaly classification while being

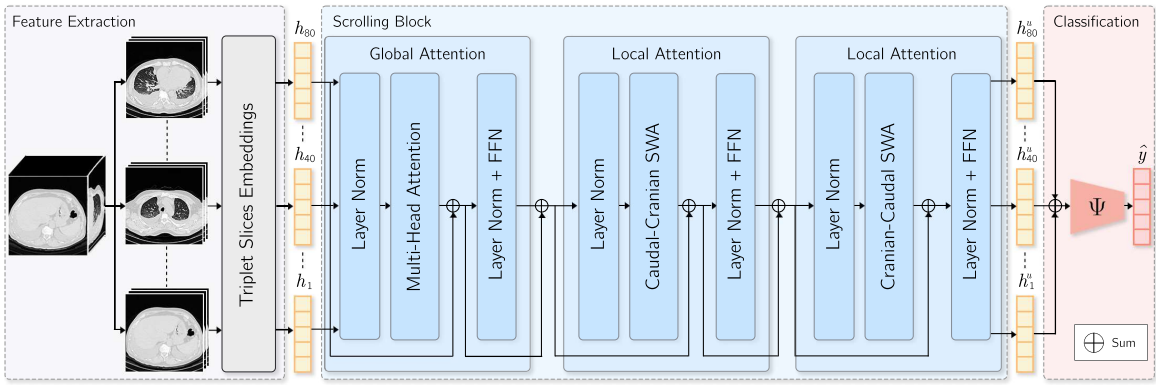

Figure 2: **The CT-Scroll architecture** consists of three main components. (1) Axial slices of the volume are grouped into triplets and processed by a ResNet followed by a GAP layer, producing a vector representation per triplet. (2) The Scrolling Block then refines these embedded visual tokens using both global and local attention mechanisms. (3) Finally, the aggregated representations are fed into a classification head to predict anomalies.

achievable with limited computational resources (single GPU, $<$ 24-hour training time).

- An extensive evaluation on two public datasets for a multi-label anomaly classification task using chest 3D CT scans.
- A comprehensive ablation study to assess the contribution of each component.

2. Related Work

2.1. 3D Visual Encoder for Medical Imaging

In the domain of 3D feature extraction, significant efforts have been made across various application areas such as remote sensing, robotic manipulation, and autonomous driving (Sarker et al., 2024). In medical imaging, particularly with 3D CT scans, conventional 3D convolutional neural networks have been widely employed for segmentation (Ilesanmi et al., 2024) and classification (Ho et al., 2021) tasks. Given the computational complexity of 3D convolutional operations, CT-Net (Draelos et al., 2021) proposes grouping consecutive slices of a CT volume into triplets, which are then passed through a ResNet (He et al., 2015) followed by a small 3D CNN to extract a compact vector representation, subsequently fed into a classification head. The adaptation of Vision Transformers (Chen et al., 2021) and Swin Transformers (Yang et al., 2023) to 3D volumes has enabled the interaction of visual tokens corresponding to different patches of the volume via self-attention mechanisms. Positional embeddings are used to preserve spatial information, facilitating better understanding of the volume structure. More recently, CT-ViT (Hamamci et al., 2024c) was introduced as a 3D-CT Vision Transformer used to generate 3D CT volumes from free-form medical text prompts. CT-ViT learns compact latent representations of 3D volumes by leveraging self-attention and causal attention mechanisms to address the challenges posed by CT scans with varying cranio-caudal coverage.

2.2. Global and Local Attention

Global Attention In both Natural Language Processing (NLP) and computer vision, Transformer-based models leverage global attention (Luong et al., 2015), where each embedded token interacts with all other tokens through the self-attention mechanism. This allows for comprehensive contextualization, capturing long-range dependencies and integrating global semantic information into the token representations (Devlin et al., 2019).
Local Attention Despite its effectiveness, global attention suffers from quadratic complexity with respect to sequence length, making it computationally expensive for long sequences in NLP (Beltagy et al., 2020). To address this, local attention mechanisms such as windowed attention were introduced, restricting each token's receptive field to a local neighborhood, thereby improving efficiency while preserving essential contextual information. In computer vision, local attention has been successfully adapted in models like Swin Transformer (Liu et al., 2021), where image patches interact within localized windows. This hierarchical approach enables efficient processing of high-resolution images and enhances the model's ability to handle objects with varying scales.
Alternating Attention Recent advancements in large language models (LLMs) (Touvron et al., 2023) have demonstrated the benefits of alternating global and local attention to improve efficiency and contextual modeling. For instance, ModernBERT (Warner et al., 2024) integrates architectural innovations inspired by recent LLMs (Team et al., 2024), alternating between global and local attention layers to balance long-range context aggregation with fine-grained local dependencies.

3. Dataset

CT-RATE dataset. We leverage the publicly available CT-RATE dataset (Hamamci et al., 2024a) to train and evaluate our proposed method. This dataset comprises 3D non-contrast chest CT scans, annotated with 18 anomalies extracted from radiology reports using a RadBERT classifier (Yan et al., 2022). The dataset is partitioned as follows: 17,799 unique patients corresponding to 34,781 CT volumes for the train set, 1,314 unique patients, corresponding to 3,075 CT volumes for the validation set and 1,314 unique patients, corresponding to 3,039 CT volumes for the test set.
Rad-ChestCT dataset. To extend our evaluation, we utilize the Rad-ChestCT dataset (Draelos et al., 2021), which consists of 1,344 3D non-contrast chest CT scans annotated with 83 anomalies extracted using a SARLE labeler from radiology reports. Among these anomalies, we evaluate our method on the 11 anomalies shared with the CT-RATE dataset.
Processing. For both datasets, all CT volumes are preprocessed to ensure uniformity and consistent input characteristics across datasets, enabling robust training and evaluation. Each volume is either center-cropped or padded to achieve a resolution of $240 \times 480 \times 480$ with an in-plane resolution of 0.75 mm and 1.5 mm in the z-axis. Hounsfield Unit (HU) values are clipped to the range $[-1000, +200]$, before normalization to $[-1, 1]$.

4. Method

When a radiologist navigates along the longitudinal axis of a CT volume (Patel and De Jesus, 2024), they scroll through axial slices to detect anomalies. Initially, they perform a

global assessment to develop a comprehensive understanding of the volume before revisiting specific slices that may contain abnormalities. Upon identifying a potential anomaly, radiologists frequently scroll back and forth across adjacent slices to incorporate local contextual information, refining their assessment by leveraging both global structure and local details. As illustrated by Figure 2, we propose a method that extracts vector representations from triplets of slices and models their interactions using global and local attention blocks. These attention mechanisms are designed to imitate the scrolling behavior of radiologists, capturing global and local contextual relationships across adjacent slices. The extracted features are then fused to predict the presence of anomalies effectively.

4.1. Triplet Slices Embedding

Similar to CT-Net (Draelos et al., 2021), the slices of the initial volume $x \in \mathbb{R}^{240\times480\times480}$ are grouped in triplets, where each triplet consists of three consecutive slices. This results in a 4D tensor with dimensions $(80 \times 3 \times 480 \times 480)$. For each triplet $x_i^t \in \mathbb{R}^{3\times480\times480}$ $(i \in \{1, \ldots, 80\})$, a feature map is extracted using a ResNet (He et al., 2015) pre-trained on ImageNet (Russakovsky et al., 2015), noted f_{ResNet}, and passed through a Global Average Pooling (GAP) layer f_{GAP} to obtain a vector representation for the triplet, noted $h_i \in \mathbb{R}^{512}$ $(i \in \{1, \ldots, 80\})$, such that:

$$h_i = (f_{\text{GAP}} \circ f_{\text{ResNet}})(x_i^t), \quad \forall\, i \in \{1, \ldots, 80\}\,. \tag{1}$$

We employ Global Average Pooling instead of a linear projection or a 3D reducing convolutional layer to significantly reduce the total number of trainable parameters while preserving the local information encoded in the feature maps (Li et al., 2023).

4.2. Scrolling Block

These vector representations $h = \{h_i\}_{i=1}^{80}$, considered as visual tokens associated with the triplet slices, are then fed into a *Scrolling Block* (SB), denoted as f_{SB}. A Scrolling Block consists of three Transformer encoders (Vaswani et al., 2023). The first encoder, denoted as f_{G}, employs global self-attention, enabling each token to aggregate information from the entire volume. The second and third encoders, denoted as $f_{\text{CAU}\to\text{CRA}}$ and $f_{\text{CRA}\to\text{CAU}}$, use Caudal-Cranial and Cranial-Caudal Sliding Window Attention, respectively, emulating the radiologist's scrolling behavior along the longitudinal axis to focus on local contextual information. For each triplet slice, the corresponding visual token can only interact with visual tokens associated with $q \in \mathbb{N}^+$ slices above it (for Caudal-Cranial modeling) or below it (for Cranial-Caudal modeling) along the longitudinal axis, as illustrated by Figure 3. Our method leverages both global attention, capturing long-range dependencies across slices, and local attention, refining contextual representations within localized regions. This design effectively models both short- and long-range interactions along the cranial-caudal axis, mirroring the way radiologists navigate through CT scans for clinical assessment. Each Transformer encoder is followed by a residual connection (He et al., 2015), a normalization layer (Ba et al., 2016), and a FeedForward Network leveraging GeGLU, a Gated Linear Unit (GLU)-based activation function that has demonstrated consistent empirical improvements over standard activation functions (Shazeer, 2020). This Scrolling Block module generates

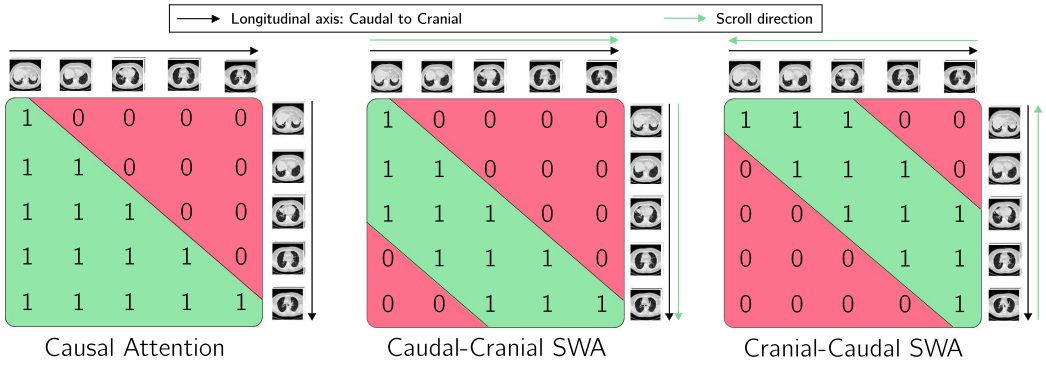

Figure 3: **Causal and Sliding Window Attention Masks.** A mask of shape (n, n) preventing attention to certain positions. 1 indicates that the corresponding position is allowed to attend, 0 otherwise. Example with $n = 5$ triplet slices and a sliding window of size $q = 3$.

updated visual tokens with a dimension of 512, denoted as $\{h_i^u\}_{i=1}^{80}$, such that:

$$h^u = \{h_1^u, \ldots, h_{80}^u\} = f_{\text{SB}}(h) = (f_{\text{CRA}\to\text{CAU}} \circ f_{\text{CAU}\to\text{CRA}} \circ f_{\text{G}})(h)\,. \tag{2}$$

Aggregation. The resulting vector representations are aggregated through summation and passed to a classification head, implemented as a lightweight Multilayer Perceptron, denoted as Ψ, which predicts a logit vector $\hat{y} \in \mathbb{R}^{18}$, such as:

$$\hat{y} = \Psi\left(\sum_{i=1}^{80} h_i^u\right)\,. \tag{3}$$

The model is trained for multi-label classification using a binary cross-entropy loss function (Goodfellow et al., 2016).

5. Implementation Details

The model was trained for 50,000 steps with a batch size of 4, using the AdamW optimizer and a cosine scheduler with a warm-up phase of 20,000 steps and a maximum learning rate of 10^{-4}. Training was conducted on a GPU with 48GB of memory.

6. Experimental Results

6.1. Quantitative results

We evaluate the model's performance using standard metrics: AUROC, F1-Score, precision, and accuracy. For classification, we determine the threshold that maximizes the F1-Score for each of the 18 labels on the validation set, as F1-Score is the harmonic mean of precision and recall (Rainio et al., 2024). On the test set, we compute the average of each metric across all labels, as well as the weighted average based on label frequencies in the test set. Reported mean and standard deviation metrics were computed over five independent runs with different random seeds to ensure robustness. As shown in Table 1, our method

Dataset	Method	AUROC	Accuracy	F1 Score	Weighted F1 Score	Precision
CT-RATE	Random Predictions	49.88 ± 0.62	49.89 ± 0.31	27.78 ± 0.51	33.13 ± 0.33	49.85 ± 1.12
	3D CNN	76.49 ± 0.28	73.22 ± 0.50	46.86 ± 0.31	51.70 ± 0.27	38.46 ± 0.54
	CT-ViT	73.92 ± 1.17	70.83 ± 0.17	45.01 ± 0.85	49.65 ± 0.88	35.59 ± 0.46
	Swin3D	<u>79.94</u> ± 0.15	75.95 ± 0.25	50.64 ± 0.25	54.68 ± 0.21	42.07 ± 0.56
	CT-Net	79.37 ± 0.27	<u>77.37</u> ± 0.40	<u>51.39</u> ± 0.50	<u>56.37</u> ± 0.32	<u>43.51</u> ± 0.68
	CT-Scroll (ours)	**81.80** ± 0.22	**79.49** ± 0.45	**53.97** ± 0.21	**58.08** ± 0.28	**48.34** ± 1.49
Rad-ChestCT	Random Predictions	50.49 ± 0.75	50.47 ± 0.37	32.71 ± 0.66	42.48 ± 0.79	27.95 ± 0.45
	3D CNN	66.53 ± 0.58	60.67 ± 1.01	44.37 ± 0.32	55.44 ± 0.26	39.77 ± 0.86
	CT-ViT	66.68 ± 0.86	63.72 ± 1.75	47.07 ± 0.74	<u>59.40</u> ± 0.34	38.94 ± 0.41
	Swin3D	70.71 ± 0.41	63.63 ± 1.01	48.29 ± 0.56	58.71 ± 0.53	40.92 ± 0.80
	CT-Net	<u>70.77</u> ± 0.38	<u>64.46</u> ± 2.01	<u>48.73</u> ± 0.83	59.14 ± 0.67	<u>41.75</u> ± 1.07
	CT-Scroll (ours)	**73.07** ± 0.54	**66.84** ± 0.52	**49.94** ± 0.49	**59.57** ± 0.53	**44.29** ± 0.58

Table 1: **Quantitative evaluation on the CT-RATE and Rad-CT-Chest test sets.** Reported mean and standard deviation metrics were computed over 5 independant runs. **Best** results are in bold, <u>second best</u> are underlined.

achieves an F1-Score of 53.97 (+Δ5.02% over CT-Net) and an AUROC of 81.80 (+Δ3.06% over CT-Net) on the CT-RATE test set. On the Rad-ChestCT test set, CT-Scroll achieves a Δ+9.58% improvement in AUROC over CT-ViT, a Δ+3.34% increase over Swin3D and a Δ+3.24% increase compared to CT-Net. A paired t-test between our method and CT-Net on all metrics yielded p-values below 0.01, demonstrating the statistical significance of these improvements.

6.2. Ablation study

Impact of the Scrolling Block module. To evaluate the effectiveness of the proposed Scrolling Block, we compare its performance against various traditional modules by replacing the Scrolling Block with these alternatives. Table 2 presents the performance of our models and the contribution of each architectural component. Replacing a small 3D convolutional layer (Draelos et al., 2021) with a Global Average Pooling layer (Li et al., 2023) for dimensionality reduction yields a +Δ2.39% improvement in AUROC while significantly reducing inference time. Introducing self-attention through Transformer Encoders (Vaswani et al., 2023) to allow interactions between visual tokens corresponding to triplet slices achieves an F1-score of 53.32, marking a Δ+1.25% increase, compared to not using self-attention. Integrating local attention via a standard Sliding Window Attention mechanism (Beltagy et al., 2020), after an initial global attention module, leads to a Δ+0.64% improvement in AUROC and a Δ+0.62% increase in F1-score compared to the global-attention-only configuration. Introducing local attention limits the interaction between CT scan slices within the same spatial neighborhood, which could enable the model to learn more fine-grained feature representations, ultimately enhancing anomaly classification performance. Finally, incorporating the Scrolling Block leads to an AUROC of 81.80 (+Δ0.68% increase over global-attention-only configuration) and an F1-score of 53.97 (+Δ1.22% increase over global-attention-only configuration), all while maintaining low computational complexity.

Method	Feat. Extractor	Reduction	Interactions	AUROC	F1 Score	Params (M)	FLOPs (T)	Infer. Time (ms)
3D CNN	-	-	-	76.49 ± 0.28	46.86 ± 0.31	0.3	0.388	1.58 ± 1.02
CT-ViT	-	-	-	73.92 ± 1.17	45.01 ± 0.85	37	0.500	13.66 ± 2.34
Swin3D	-	-	-	79.94 ± 0.15	50.64 ± 0.25	28	0.905	14.79 ± 1.91
CT-Net	ResNet-18	**3D Conv.**	None	79.37 ± 0.27	51.39 ± 0.50	15	1.344	16.54 ± 2.32
-	ResNet-18	**GAP**	None	81.21 ± 0.40	52.66 ± 0.41	12	1.335	3.65 ± 0.75
-	ResNet-18	**GAP**	**Tr. Enc. (causal attention)**	81.45 ± 0.21	52.98 ± 0.44	16	1.337	5.11 ± 0.92
-	ResNet-18	**GAP**	**Tr. Enc. (global attention)**	81.25 ± 0.07	53.32 ± 0.22	16	1.337	5.57 ± 0.98
-	ResNet-18	**GAP**	**Tr. Enc. (global + local)**	81.77 ± 0.06	53.65 ± 0.39	16	1.337	5.26 ± 1.15
CT-Scroll	ResNet-18	**GAP**	**Scrolling Block**	**81.80** ± 0.22	**53.97** ± 0.21	16	1.337	5.46 ± 1.55

Table 2: **Comparison of performance across different modules.** We use traditional Transformer Encoders with matching layer counts and computational costs to ensure fair comparisons across setups. The Params column corresponds to the number of trainable parameters (in millions, M). FLOPs (T) column refers to the number of floating-point operations (in tera, T). Infer. Time (ms) corresponds to the average inference time per sample (in milliseconds, ms), estimated with a NVIDIA RTX A6000 GPU.

Impact of the Sliding Window Size. To determine the optimal window size for contextual understanding, we systematically vary the window size q of the SWA and measure its impact on performance. Table 3 presents the model's performance across various window sizes. CT-Scroll with a window size of 16 yields a Δ+0.68% improvement both in AUROC and in Accuracy, and a Δ+1.22% enhancement in F1-Score compared to global attention.

Window size	AUROC	Accuracy	F1 Score	Precision	Infer. Time (ms)
4	81.54 ± 0.09	78.94 ± 0.37	53.47 ± 0.03	46.99 ± 0.48	5.42 ± 0.87
8	81.45 ± 0.25	79.02 ± 0.26	53.93 ± 0.30	47.46 ± 0.54	5.42 ± 1.01
16	**81.80** ± 0.22	**79.49** ± 0.85	**53.97** ± 0.21	**48.34** ± 1.49	5.46 ± 1.55
32	81.53 ± 0.14	79.42 ± 0.19	53.50 ± 0.17	47.59 ± 0.51	5.54 ± 0.92
64	81.46 ± 0.14	79.42 ± 0.72	53.37 ± 0.27	47.75 ± 0.84	5.54 ± 0.90
Global	81.25 ± 0.07	78.95 ± 0.66	53.32 ± 0.22	46.85 ± 0.48	5.57 ± 0.98

Table 3: **Impact of the sliding window size.** The sliding window size, denoted as q, corresponds to the number of triplet slices considered during the computation of the attention mechanism. *Global* indicates global attention, where each token attends to all others across the full set of 80 tokens.

7. Conclusion

In this work, we introduce CT-Scroll, a hybrid model for 3D CT Volumes that extracts triplet slices embeddings through a 2D convolutional network and facilitates interactions between these representations using both global and local attention mechanisms, imitating the radiologist's behavior. CT-Scroll is trained on a multi-label classification task and evaluated on two public datasets, with a focus on chest CT volumes. In addition to enhancing multi-label anomaly classification performance (+Δ19.91% increase over CT-ViT, +Δ6.58% increase over Swin3D and +Δ5.02% increase over CT-Net in F1-Score), CT-Scroll demonstrates low computational complexity. Future work could explore the integration of region-specific information to further enhance classification performance or the incorporation of a lightweight 3D CNN module to take full advantage of the third dimension.

Acknowledgments

We acknowledge (Hamamci et al., 2024a) for providing the CT-RATE dataset and (Draelos et al., 2021) for providing the Rad-ChestCT dataset.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization, July 2016. URL http://arxiv.org/abs/1607.06450. arXiv:1607.06450 [cs, stat].

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, December 2020. URL http://arxiv.org/abs/2004.05150. arXiv:2004.05150.

Joshua Broder and David M. Warshauer. Increasing utilization of computed tomography in the adult emergency department, 2000-2005. *Emergency Radiology*, 13(1):25–30, October 2006. ISSN 1070-3004. doi: 10.1007/s10140-006-0493-9.

Junyu Chen, Yufan He, Eric C. Frey, Ye Li, and Yong Du. ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration, April 2021. URL http://arxiv.org/abs/2104.06468. arXiv:2104.06468 [eess].

Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C. Thai, Kathleen Moore, Robert S. Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, 79: 102444, July 2022. ISSN 1361-8423. doi: 10.1016/j.media.2022.102444.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating Long Sequences with Sparse Transformers, April 2019. URL http://arxiv.org/abs/1904.10509. arXiv:1904.10509.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

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, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Rachel Lea Draelos, David Dov, Maciej A. Mazurowski, Joseph Y. Lo, Ricardo Henao, Geoffrey D. Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical Image Analysis*, 67:101857, January 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101857. URL https://www.sciencedirect.com/science/article/pii/S1361841520302218.

Stacy K. Goergen, Felicity J. Pool, Tari J. Turner, Jane E. Grimm, Mark N. Appleyard, Carmel Crock, Michael C. Fahey, Michael F. Fay, Nicholas J. Ferris, Susan M. Liew, Richard D. Perry, Ann Revell, Grant M. Russell, Shih-Chang S. C. Wang, and Christian

Wriedt. Evidence-based guideline for the written radiology report: methods, recommendations and implementation challenges. *Journal of Medical Imaging and Radiation Oncology*, 57(1):1–7, February 2013. ISSN 1754-9485. doi: 10.1111/1754-9485.12014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL https://www.deeplearningbook.org/.

Pengfei Gu, Yejia Zhang, Chaoli Wang, and Danny Z. Chen. ConvFormer: Combining CNN and Transformer for Medical Image Segmentation, November 2022. URL http://arxiv.org/abs/2211.08564. arXiv:2211.08564 [cs].

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Mehmet K. Ozdemir, and Bjoern Menze. A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities, March 2024a. URL http://arxiv.org/abs/2403.17834. arXiv:2403.17834 [cs].

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging, March 2024b. URL http://arxiv.org/abs/2403.06801. arXiv:2403.06801 [cs, eess].

Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Dogan, Muhammed Furkan Dasdelen, Chinmay Prabhakar, Hadrien Reynaud, Sarthak Pati, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. GenerateCT: Text-Conditional Generation of 3D Chest CT Volumes, March 2024c. URL http://arxiv.org/abs/2305.16037. arXiv:2305.16037 [cs].

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs].

Thao Thi Ho, Taewoo Kim, Woo Jin Kim, Chang Hyun Lee, Kum Ju Chae, So Hyeon Bak, Sung Ok Kwon, Gong Yong Jin, Eun-Kee Park, and Sanghun Choi. A 3D-CNN model with CT-based parametric response mapping for classifying COPD subjects. *Scientific Reports*, 11(1):34, January 2021. ISSN 2045-2322. doi: 10.1038/s41598-020-79336-5. URL https://www.nature.com/articles/s41598-020-79336-5. Publisher: Nature Publishing Group.

Ademola E. Ilesanmi, Taiwo O. Ilesanmi, and Babatunde O. Ajayi. Reviewing 3D convolutional neural network approaches for medical image segmentation. *Heliyon*, 10 (6):e27398, March 2024. ISSN 2405-8440. doi: 10.1016/j.heliyon.2024.e27398. URL https://www.sciencedirect.com/science/article/pii/S2405844024034297.

Zhenwei Li, Mengying Xu, Xiaoli Yang, Yanqi Han, and Jiawen Wang. A Multi-Label Detection Deep Learning Model with Attention-Guided Image Enhancement for Retinal Images. *Micromachines*, 14(3):705, March 2023. ISSN 2072-666X. doi: 10.3390/mi14030705. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10054796/.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, August 2021. URL http://arxiv.org/abs/2103.14030. arXiv:2103.14030 [cs].

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation, September 2015. URL http://arxiv.org/abs/1508.04025. arXiv:1508.04025 [cs].

Michalis Mazonakis and John Damilakis. Computed tomography: What and how does it measure? *European Journal of Radiology*, 85(8):1499–1504, August 2016. ISSN 1872-7727. doi: 10.1016/j.ejrad.2016.03.002.

Paula R. Patel and Orlando De Jesus. CT Scan. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024. URL http://www.ncbi.nlm.nih.gov/books/NBK567796/.

Oona Rainio, Jarmo Teuho, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086, March 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-56706-x. URL https://www.nature.com/articles/s41598-024-56706-x. Publisher: Nature Publishing Group.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, January 2015. URL http://arxiv.org/abs/1409.0575. arXiv:1409.0575 [cs].

Sushmita Sarker, Prithul Sarker, Gunner Stone, Ryan Gorman, Alireza Tavakkoli, George Bebis, and Javad Sattarvand. A comprehensive overview of deep learning techniques for 3D point cloud classification and semantic segmentation. *Machine Vision and Applications*, 35(4):67, July 2024. ISSN 0932-8092, 1432-1769. doi: 10.1007/s00138-024-01543-1. URL http://arxiv.org/abs/2405.11903. arXiv:2405.11903 [cs].

Noam Shazeer. GLU Variants Improve Transformer, February 2020. URL http://arxiv.org/abs/2002.05202. arXiv:2002.05202 [cs].

Satya P. Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balázs Gulyás. 3D Deep Learning on Medical Images: A Review, October 2020. URL http://arxiv.org/abs/2004.00218. arXiv:2004.00218 [q-bio].

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple

Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving Open Language Models at a Practical Size, October 2024. URL http://arxiv.org/abs/2408.00118. arXiv:2408.00118 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023. URL http://arxiv.org/abs/2302.13971. arXiv:2302.13971 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, December 2024. URL http://arxiv.org/abs/2412.13663. arXiv:2412.13663 [cs].

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting Transformer-based Language Models to Radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, July 2022. doi: 10.1148/ryai.210258. URL https://pubs.rsna.org/doi/full/10.1148/ryai.210258. Publisher: Radiological Society of North America.

Yu-Qi Yang, Yu-Xiao Guo, Jian-Yu Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3D: A Pretrained Transformer Backbone for 3D Indoor Scene Understanding, August 2023. URL http://arxiv.org/abs/2304.06906. arXiv:2304.06906 [cs].

Yijie Yuan, Matthew Tivnan, Grace J. Gang, and J. Webster Stayman. Deep Learning CT Image Restoration using System Blur Models. *Proceedings of SPIE–the International Society for Optical Engineering*, 12463:124634J, February 2023. ISSN 0277-786X. doi: 10.1117/12.2655806. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10760795/.