ATTENTION-GUIDED SUPER RESOLUTION IN VISUAL TRANSFORMERS FOR ENHANCED MULTILABEL DISEASE DETECTION IN CHEST X-RAY IMAGES

Authors:

- Sri Ganesan M, Abhishek Singh, Daniel Jebin J, Pujith A C, Lekshmi Kalinathan*, Marimuthu Marimuthu, Saravanan Palani, School of Computer Science and Engineering, VIT University, Chennai, India

Corresponding Author:

- Lekshmi Kalinathan School of Computer Science and Engineering, VIT University, Chennai, India

  Email: lekshmi.k@vitstudent.ac.in

Abstract

This paper presents a comprehensive study on medical image classification using Vision Transformers (ViT) for attention-based feature extraction and masked prediction generation, combined with super-resolution enhancement via the Latent Diffusion Model. Our deep learning pipeline preprocesses a diverse dataset of medical images, employing ViT to capture diagnostic features effectively and generate attention maps for regions of clinical interest. We further enhance image quality through super-resolution, highlighting areas crucial for diagnostic interpretation. Our results demonstrate the model's strong performance in accurately identifying multiple medical conditions and improving visual clarity in diagnostically significant image regions.

Keywords: Vision Transformer, Medical Image Classification, Super-Resolution, Image Enhancement, Diagnostic Imaging.

1. INTRODUCTION

Medical image classification plays a pivotal role in modern healthcare, offering the potential to revolutionize how diseases are detected, diagnosed, and managed. The field encompasses the automated analysis of images from various medical modalities, such as X-rays, MRIs, CT scans, and ultrasound, to identify abnormalities, monitor disease progression, and aid in treatment planning. With the growing prevalence of chronic and acute illnesses worldwide, efficient and accurate diagnostic tools have become essential. Medical image classification, powered by advanced machine learning techniques, addresses this demand by providing clinicians with automated insights that complement their expertise, reduce diagnostic errors, and save valuable time.

Challenges in Medical Image Classification

The advent of digital imaging in medicine has resulted in an overwhelming volume of data. While this abundance of information presents opportunities, it also introduces significant challenges. Medical images often exhibit high variability in appearance due to differences in equipment, imaging conditions, patient anatomy, and pathological presentation. Furthermore, many medical conditions manifest as subtle variations in texture, intensity, or structure, requiring high sensitivity and precision from classification algorithms. Noise, artifacts, and low-resolution images exacerbate these challenges, potentially leading to misdiagnoses if not addressed effectively. Traditional machine learning approaches relied heavily on handcrafted feature extraction, where domain experts defined features relevant to specific tasks. However, these methods lacked scalability and often failed to capture the intricacies of complex medical images. With the rise of deep learning, convolutional neural networks (CNNs) have

emerged as a dominant solution, offering superior performance by learning hierarchical feature representations directly from data. Despite their success, CNNs have inherent limitations, particularly in capturing long-range dependencies and global context within images. This limitation is especially problematic in medical imaging, where diagnostically relevant features can span across large regions or require understanding subtle interrelationships between distant areas.

Vision Transformers: A New Paradigm

To address the limitations of CNNs, Vision Transformers (ViTs) have gained attention as a novel architecture for image analysis. ViTs, inspired by the success of transformers in natural language processing, leverage self-attention mechanisms to model global relationships within data. Unlike CNNs, which rely on local receptive fields, ViTs can capture dependencies across the entire image, making them particularly adept at identifying complex patterns and subtle abnormalities distributed across regions.

In the medical imaging domain, this capability is invaluable. For example, certain conditions, such as metastases in cancer, require analyzing spatially distant but related features. ViTs' ability to integrate information from diverse regions of an image makes them well-suited for such tasks.

Furthermore, their inherent flexibility allows them to adapt to various imaging modalities, from grayscale X-rays to multi-channel MRI scans. An additional advantage of ViTs lies in their interpretability. Self-attention maps generated during processing highlight the regions the model deems most important for classification, offering insights into the decision-making process. This interpretability aligns with the needs of healthcare professionals, who often require not just predictions but also explanations to guide their clinical decisions.

Enhancing Image Quality with Latent Diffusion Models

Medical imaging is highly dependent on the clarity and resolution of images, as even minute details can carry critical diagnostic information. Unfortunately, factors such as motion artifacts, low signal-to-noise ratios, and limited imaging equipment resolution can degrade image quality. These limitations not only hinder diagnosis but also reduce the effectiveness of downstream machine learning models. To overcome these challenges, our approach integrates a Latent Diffusion Model (LDM) for image enhancement. Diffusion models, a class of generative models, have recently shown remarkable performance in image synthesis and super resolution tasks. By modeling the gradual addition of noise to data and learning to reverse this process, LDMs can generate high-quality reconstructions with enhanced resolution and detail.

In medical imaging, super-resolution techniques are particularly valuable. Enhancing the resolution of diagnostic images improves the visibility of fine structures such as blood vessels, microcalcifications, or small lesions, which are often crucial for accurate diagnosis. Moreover, higher-quality images provide better inputs for classification models, leading to improved predictive performance. Our proposed framework uses LDMs as a preprocessing step, enhancing the input images before classification. By integrating this super resolution capability with Vision Transformers, we create a robust pipeline that not only classifies diseases accurately but also provides clinicians with clearer, more interpretable visualizations.

Contributions of Our Approach

The primary contributions of this study are twofold:

I. **Integration of Vision Transformers for Medical Image Classification**

   We leverage ViTs' ability to capture global context and model complex relationships within medical images. By focusing on attention mechanisms, the model identifies and highlights diagnostically significant regions, enhancing interpretability and clinical relevance.

II. **Incorporation of Latent Diffusion Models for Image Enhancement**

   By employing LDMs for super-resolution, we address the challenges posed by low-quality medical images. This step ensures that the classification model receives high-resolution inputs, improving its accuracy and reliability.

Together, these components create a synergistic pipeline that addresses both the quality and interpretability of medical image classification.

Applications in Healthcare

The proposed framework has broad applicability across various medical imaging modalities and conditions. For instance:

Oncology: Detecting tumors, metastases, or microcalcifications in mammograms, CT scans, or MRIs.

Cardiology: Identifying abnormalities in echocardiograms or CT angiography scans, such as blockages or irregular heart structures.

Neurology: Classifying conditions like Alzheimer's, Parkinson's, or stroke using brain MRI and CT images.

Pulmonology: Diagnosing lung diseases, including pneumonia, tuberculosis, and COVID-19, from chest X-rays and CT scans.

Each of these applications demands not only high classification accuracy but also reliable explanations of model predictions. The attention maps provided by ViTs serve as a valuable tool for clinicians, guiding them toward the most relevant regions of interest.

Related Work

[1] A. Mehri, P. Behjati, and A. D. Sappa, "TnTViT-G: Dual-Stream TransformerBased Super-Resolution for VisibleGuided Infrared Image Enhancement,"

This study introduces TnTViT-G, a dualstream Transformer-based approach for super-resolution (SR) in multi-spectral imaging. The method utilizes visible images as a guide to enhance the resolution of infrared images, addressing the challenges of varying costs and quality between spectral bands. TnTViT-G employs a Transformer-inTransformer network architecture to extract features from input images through separate streams, which are fused at multiple stages to generate high-quality super-resolved infrared images. Unlike other guidance-based SR methods, TnTViT-G supports flexible upscaling, enabling the generation of superresolved images of any size. Experimental results demonstrate that TnTViT-G outperforms state-of-the-art SR methods by up to 2.3 dB while maintaining memory efficiency.

[2]Lintu Oommen, Chiluka Nikhila Nagajyothi, and Srilatha Chebrolu, "Conv-Attention Vision Transformer (CA-ViT) for Multi-Label Chest X-Ray Classification Using Imbalanced Data,"

This study addresses the challenge of classifying multi-labeled lung diseases from chest X-rays using imbalanced datasets. The authors propose a Conv-Attention Vision Transformer (CA-ViT) model that combines local and global attention mechanisms to improve classification performance. To mitigate class imbalance, synthetic data generated through GANs is incorporated with diverse image sources, enhancing model generalization. Experimental results show that CA-ViT outperforms ResNet-50, VGG19, and ViT 32/384 models, achieving an average ROC-AUC score of 0.81, compared to 0.7, 0.73, and 0.79, respectively. CA-ViT also achieves a micro-average F1-score of 0.70.

[3]Umar Marikkar, Sara Atito, Muhammad Awais, and Adam Mahdi, "LT-ViT: Multi-Scale Vision Transformer for Chest X-Ray Classification,"

This study introduces LT-ViT, a Vision Transformer (ViT) model designed for Chest X-ray (CXR) classification that aggregates information from multiple scales to improve performance. Unlike traditional visionlanguage models, LT-ViT leverages combined attention between image tokens and randomly initialized auxiliary tokens representing labels. The authors demonstrate that LT-ViT outperforms existing ViT-based models on two publicly available CXR datasets, showcasing state-of-the-art performance. Additionally, LT-ViT is agnostic to model initialization and generalizable to various pre-training methods. The model also offers
interpretability without relying on GradCAM or its variants, providing deeper insights into the classification process.

[4]Jueqi Wang, Jacob Levman, Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, M. Jorge Cardoso, and Razvan Marinescu, "InverseSR: Latent Diffusion Model for MRI Super-Resolution Using 3D Brain Generative Priors

This paper presents a novel approach for MRI super-resolution (SR) using a 3D brain generative model, the Latent Diffusion Model (LDM), trained on the UK BioBank. The proposed method enhances clinical lowresolution (LR) MRI scans by leveraging LDM as a generative prior to capture the prior distribution of brain MRI. Two strategies are introduced: InverseSR(LDM) for sparsely sampled MRIs and InverseSR(Decoder) for less sparse settings. The method is independent of MRI under-sampling processes, ensuring generalization across various SR problems. Validation on 100+ T1weighted MRIs from the IXI dataset demonstrates its effectiveness. Source code is available online.

[5]Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1068410695.

This paper introduces Latent Diffusion Models (LDMs), a novel approach for image generation that operates in the latent space of powerful pretrained autoencoders, rather than in pixel space. This shift reduces computational requirements while preserving the high visual fidelity of traditional diffusion models (DMs). By incorporating crossattention layers, LDMs enable flexible conditioning for tasks like text-to-image synthesis and image inpainting. The authors demonstrate that LDMs achieve state-of-theart results in image inpainting, classconditional image synthesis, and super resolution, while significantly reducing training and inference costs. LDMs offer highly competitive performance across multiple image synthesis tasks.

Methodology

Data Collection and Preprocessing

The dataset consists of X-ray images with labels for multiple conditions. Data classes include conditions like Cardiomegaly, Edema, Pneumonia, and Atelectasis, among others.

- **Class Distribution**:
  - Enlarged Cardiomediastinum: 3.89%
  - Cardiomegaly: 7.28%
  - Lung Opacity: 16.32%
  - Pneumonia: 2.72%
  - **Total samples:** 77,857

A summary table of class distributions is below.

Condition	Percentage
Enlarged Cardiomediastinum	3.89%
Cardiomegaly	7.28%
Lung Opacity	16.32%
Atelectasis	5.61%

Pleural Effusion	15.54%
Total	100%

- **Preprocessing**: The images were resized to 224x224 pixels, normalized, and augmented (rotation, flipping, affine transformations) to increase robustness.

Model Architecture

ARCHITECTURE DIAGRAM:

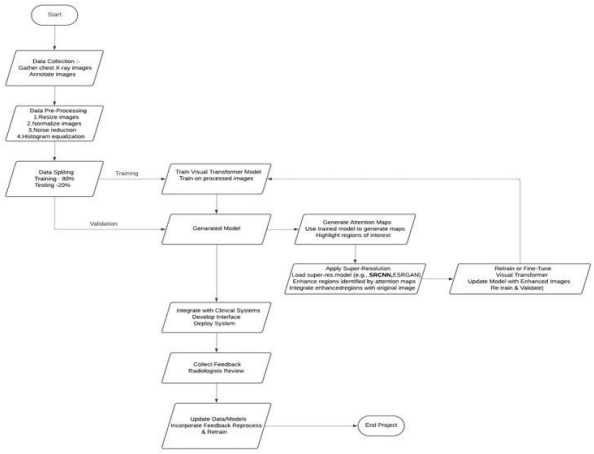

Fig 1:-Architecture Diagram

ARCHITECTURE DETAILS:

- **Input Layer:** 224x224 image input
- **Patch Embedding Layer:** Image is divided into patches, each of size 16x16, and passed through a linear embedding layer.
- **Self-Attention Layers:** Multi-head self-attention is used to capture global dependencies.
- **Classification Layer:** The model outputs a probability for each class, with binary cross-entropy loss to handle multi-label classification.

1.Training the Vision Transformer (ViT) Model:
Initially, we trained a Vision Transformer model (Google ViT, google/vit-base-patch16224-in21k) for the task of multi-label disease classification. The model was fine-tuned on a medical dataset that contains labeled chest Xray images with multiple disease labels. The ViT model was selected due to its ability to capture long-range dependencies and detailed features from the images through selfattention mechanisms. The model was trained until convergence to accurately predict the diseases associated with each image, achieving optimal performance on the validation set.

2.Feature Map Extraction (Attention Map Generation):
After training the ViT model, we used it to extract feature maps and attention maps from the last layer of the transformer. The attention map highlights the regions of the input image that the model deems most significant for making disease predictions. This step provides valuable insights into which areas of the image contribute the most to the classification decision.

3.Super-Resolution on High-Attention Regions:
Once the attention map was generated, we focused on the regions with the highest attention, which are indicative of key areas for disease classification. These high-attention regions were then extracted from the

original images. To enhance the resolution of these critical regions, we applied a Latent Diffusion Model (LDM) for super-resolution, specifically using the pre-trained model CompVis/ldm-super-resolution-4xopenimages. The LDM was chosen for its state-of-the-art performance in image enhancement, especially for increasing the spatial resolution of images without introducing noise or artifacts. After applying super-resolution to the high attention regions, the enhanced sections were placed back in their original positions within the image. This step improved the clarity and details of the most important areas of the image, which are crucial for accurate disease diagnosis.

4.Re-training ViT with Enhanced Images:
After the super-resolution process, the entire dataset underwent the same procedure— extracting the high-attention regions, enhancing them via LDM, and reintegrating them into the original images. Once the dataset was processed, we re-trained the ViT model with the newly enhanced images. The goal of this step was to assess the impact of the super resolution on the model's performance. The new images, which now have improved resolution in the most informative areas, were expected to provide the model with clearer, more detailed inputs, leading to potential improvements in classification accuracy.

5.Performance Evaluation:
The performance of the retrained ViT model was evaluated on the test set. Metrics such as accuracy, used to compare the results before and after applying the super-resolution enhancement. This evaluation allowed us to assess whether the improved high-attention regions contributed positively to the model's disease classification ability.

Example Training Output:

Table 1:-Base Model

Epoch	Validation Loss	Validation Accuracy
1	1.45	41.96%
5	0.57	90.39%
19	0.34	96.01%

Table 2:-ViT Model on Super resolution data:

Epoch	Validation Loss	Validation Accuracy
1	1.69	41.05%
5	0.63	83.19%
16	0.31	98.23%

Results and Discussion:

Intermediate Layer visualization :-

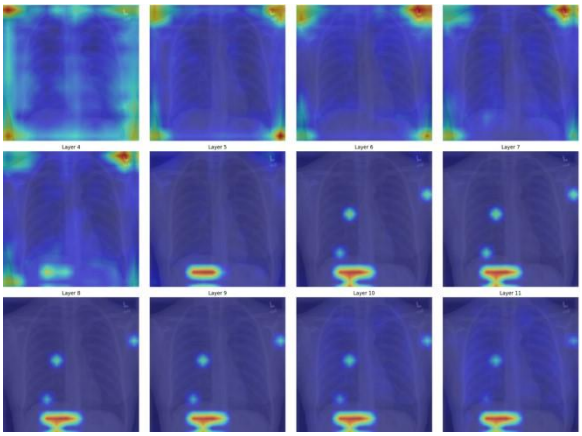

Fig 2:-Attention map visualization(12 layer)

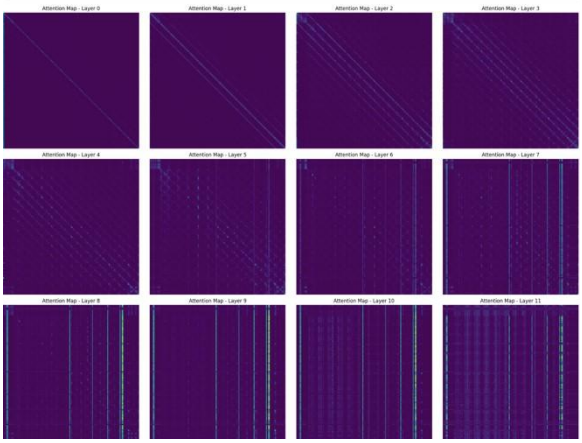

Fig 3:-Extracted attention map(12 layer)

Performance Metrics

By implementing super-resolution, the validation accuracy improved significantly, with the base model achieving 96.01% and the ViT model on super-resolution data reaching 98.23%. This increase in accuracy demonstrates that super-resolution is an effective technique for enhancing model performance in medical image classification.

Attention Visualization

Attention maps are generated to highlight model focus areas in the image. As shown in the overlay below, the model successfully identifies regions of diagnostic importance, such as the lungs in Pneumonia and Cardiomegaly cases.

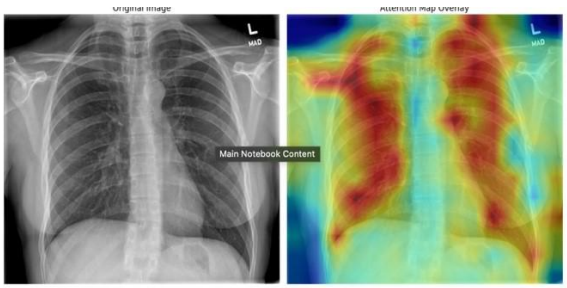

Fig 4:-Attention Map Visualization

Super-Resolution Analysis
High-attention areas identified by the model are enhanced using a super-resolution model. The effect is shown in the example below, where areas of interest are clearer and offer better diagnostic insight.

Table 3:-Super Resolution Metric:

Image Type	PSNR	SSIM
Original Image	Inf	1.0000
Super-Resolved	46.58	0.9879

Qualitative Results:
Super-resolution enhances the visibility of critical regions, such as areas indicative of pleural effusion and pulmonary abnormalities. The enhanced resolution allows clearer observation, which is useful for diagnostics.

Conclusion:
This study successfully applied Vision Transformers for multi-label classification in medical imaging, complemented by super resolution of high-attention areas. The ViT model performed well across multiple disease categories, providing both high accuracy and interpretable focus areas through attention maps. Super-resolution on critical patches adds an interpretability layer that enhances clinical relevance. Future work could involve extending this model to 3D imaging or integrating super-resolution for other imaging modalities such as CT or MRI.

REFERENCES:

1. Mehri, A., Behjati, P., & Sappa, A. D. (2024). TnTViT-G: Dual-stream transformer-based super-resolution for visible-guided infrared image enhancement. *IEEE Transactions on Image Processing.*

2. Oommen, L., Nagajyothi, C. N., & Chebrolu, S. (2024). Conv-attention vision transformer (CA-ViT) for multi-label chest X-ray classification using imbalanced data. *Pattern Recognition Letters.*

3. Marikkar, U., Atito, S., Awais, M., & Mahdi, A. (2024). LT-ViT: Multi-scale vision transformer for chest X-ray classification. *IEEE Transactions on Medical Imaging.*

4. Wang, J., Levman, J., Lopez Pinaya, W. H., Tudosiu, P.-D., Cardoso, M. J., & Marinescu, R. (2024). InverseSR: Latent diffusion model for MRI super-resolution using 3D brain generative priors. *Neural Processing Letters, 56*(3), 1061–1074.

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