”Transformer Decoder for Chest X-ray Image Captioning Using Deep Feature Extraction”

Abstract

Chest x-ray are widely used in hospitals to help doctors diagnose lung problems. Since the outbreack of covid19, especially during second wave and winter season, it has become even more important to quickly detect the disease. To help doctors and reduce their workload, we use Deep Learning to automatically analyze chest X-ray .This study, propose a method can look at a chest X-ray image and automatically generate a medical report. First, we use a model called Vision Transformer (ViT) to understand overall features of image. use another model called CheXNet, which is good at identifying chest-related diseases, to extract detailed medical features.These features are combined and sent to a Transformer decoder, which creates a meaningful text description of what is seen in the image. This helps doctors by giving them a quick, accurate summary of the patient’s condition, making the diagnosis process faster and more reliable. **keyword**:Chest X-ray,COVID-19,Medical Image,Vision Transformer (ViT),CheXNet,Transformer Decoder,Automated Diagnosis, Medical Image, Radiology

1 Introduction

Chest X-ray Imaging: Principles and Diagnostic Significance chest x ray imaging represents one of the most fundamental and frequently employed diagnostic modalities in contemporary clinical medicine. Due to its widespread availability, rapid image acquisition, low cost, and minimal radiation exposure, it continues to serve as a primary tool for the evaluation of thoracic diseases across diverse healthcare settings, from emergency departments to outpatient clinics. Chest radiographs provide a two-dimensional (2D) projection of complex three-dimensional (3D) thoracic anatomy, including the lungs, heart, major vessels, diaphragm, mediastinum, bony structures, and pleural spaces. The diagnostic utility of CXR imaging is grounded in the principle of differential radiographic density. This principle allows radiologists to discern normal anatomical variations and pathological alterations based on how various tissues attenuate X-ray

beams. Air-filled structures, such as healthy lung parenchyma, appear radiolucent or black, while denser tissues—such as bone, fluid, or consolidated lung tissue—manifest as progressively whiter opacities on the radiograph. This fundamental contrast enables the identification of key pathological hallmarks associated with pulmonary infections, neoplastic growths, pleural abnormalities, and vascular or cardiac conditions. Despite the inherent value of chest X-rays, interpretation remains an intricate task, challenged by the overlapping of anatomical structures, variable patient positioning, technical inconsistencies (e.g., exposure, projection angles), and subtle early-stage pathologies. Moreover, disease-specific radiographic manifestations can vary significantly in their appearance and severity, sometimes producing overlapping visual patterns that further complicate diagnosis. Consequently, accurate interpretation demands substantial expertise and experience, often necessitating the integration of clinical findings with imaging data. Even among seasoned radiologists, inter-observer variability poses a persistent limitation, prompting increased interest in automated diagnostic systems utilizing deep learning to enhance consistency and accuracy. Radiographic Variability Across Disease States The visual representation of thoracic disease on chest X-ray imaging is deeply influenced by the underlying pathological processes. Below, we examine the radiographic hallmarks and pathophysiological basis of four major pulmonary conditions—pneumonia, COVID-19, consolidation, and pleural effusion—which serve as core focus areas for automated image-to-text translation systems in medical AI. Pneumonia: Pathogenesis and Radiologic Characteristics Pneumonia is a common and potentially life-threatening respiratory infection characterized by inflammation of the alveoli—the small air responsible to oxygen-carbon dioxide exchange in the lungs. This inflammation leads to the filling of alveolar spaces with exudate, pus, and cellular debris, significantly impairing pulmonary function and oxygenation. The etiological spectrum of pneumonia is broad, encompassing bacterial agents (such as Streptococcus pneumoniae), viruses (e.g., influenza virus, respiratory syncytial virus), and opportunistic fungal pathogens (e.g., Pneumocystis jirovecii, particularly in immunocompromised individuals). Clinically, pneumonia manifests with symptoms such as fever, productive or non-productive cough, pleuritic chest pain, tachypnea, dyspnea, and general malaise. The condition disproportionately affects the extremes of age—the very young and the elderly—as well as individuals with underlying health conditions such as chronic obstructive pulmonary disease , diabetes, or immunodeficiency. Radiographically, pneumonia is characterized by areas of increased pulmonary opacity on chest X-rays, indicative of alveolar consolidation. These opacities may be confined to a single lobe (lobar pneumonia), distributed in multiple segments (segmental), or diffusely scattered (bronchopneumonia). The classic radiologic finding is the presence of homogeneous, well-defined opacities, often accompanied by air bronchograms—radiolucent tubular structures representing air-filled bronchi within fluid-filled alveoli. Pneumonia significant cause of global morbidity and mortality, and radiographic imaging plays a vital role in its diagnosis, monitoring, and therapeutic management. COVID-19: Radiological Insights into a Global Pandemic Coronavirus Disease 2019 COVID-19 is a highly contagious respira-

tory illness caused by the novel coronavirus SARS-CoV-2. First reported in Wuhan, China, in December 2019, COVID-19 rapidly escalated into a global pandemic, placing an unprecedented burden on healthcare systems worldwide. The disease exhibits a broad clinical spectrum, ranging from asymptomatic infection to severe pneumonia, acute respiratory distress syndrome (ARDS), multi-organ failure, and death. The primary mode of transmission is via respiratory droplets and aerosols, with the lungs being the principal target organ due to the expression of ACE2 receptors. Radiologic imaging, including both chest X-ray and computed tomography , played essential role in the triage, diagnosis, and longitudinal assessment of COVID-19 patients, especially in settings where polymerase chain reaction testing is delayed or unavailable. On chest X-rays, COVID-19-related pneumonia is often characterized by bilateral, peripheral ground-glass opacities , diffuse patchy infiltrates, and reticular or nodular patterns. In advanced cases, extensive consolidation may be observed, particularly in the lower lung zones. Unlike bacterial pneumonia, which often presents as localized consolidation, COVID-19-related pulmonary involvement tends to be more diffuse and asymmetric. The pulmonary damage seen in severe COVID-19 is frequently attributed to a dysregulated immune response, including the so-called “cytokine storm,” which leads to widespread alveolar damage, increased capillary permeability, and interstitial edema. Chest X-ray imaging thus serves not only as a diagnostic tool but also as a monitoring modality to track disease progression and evaluate response to therapeutic interventions, including antivirals, corticosteroids, and supportive oxygen therapy. Consolidation: A Radiologic Sign of Underlying Disease In radiological parlance, consolidation refers to the replacement of normally aerated alveolar spaces with pathologic substances such as pus (as in infection), blood (as in hemorrhage), fluid (as in edema), or neoplastic cells (as in malignancy). This phenomenon results in a loss of the normal air-tissue interface, producing regions of increased radiodensity on imaging studies. Consolidation is most commonly associated with infectious processes such as pneumonia, but it may also signify non-infectious pathologies including pulmonary infarction, neoplasms, and autoimmune conditions such as organizing pneumonia. On chest X-rays, consolidation appears as a well-demarcated or patchy homogeneous white opacity that often obscures the underlying pulmonary vascular markings. It may be associated with additional features such as air bronchograms, silhouette sign (loss of the normal border between heart and lung), and volume loss or expansion depending on the disease mechanism. Importantly, the morphology and distribution of consolidation can yield diagnostic clues. For instance, focal lobar consolidation suggests bacterial pneumonia, whereas bilateral diffuse consolidation might point toward viral infection or ARDS. Thus, identifying the pattern, density, and extent of consolidation is pivotal in the differential diagnosis of pulmonary conditions and in guiding appropriate clinical management. Pleural Effusion: Imaging Features and Clinical Relevance Pleural effusion is defined as the pathological accumulation of fluid within the pleural space—the narrow compartment between the parietal and visceral pleura that envelops the lungs. This condition may arise due to a wide array of systemic and local factors, including congestive heart

failure, pneumonia, malignancy, pulmonary embolism, and connective tissue diseases like lupus or rheumatoid arthritis. Clinically, pleural effusion can manifest as dyspnea, chest pain, and diminished breath sounds on auscultation. The severity of symptoms typically correlates with the volume and rapidity of fluid accumulation. On chest radiographs, pleural effusion is indicated by blunting of the costophrenic angles, a classic meniscus sign, and, in larger effusions, a homogeneous opacity obscuring the underlying lung parenchyma. In lateral decubitus positioning, the fluid may shift with gravity, further confirming its free-flowing nature. Effusions are broadly classified as transudative—resulting from systemic conditions such as heart failure or hypoalbuminemia—and exudative—caused by local inflammation, infection, or malignancy. Accurate differentiation between these types is crucial for clinical decision-making and is often guided by imaging findings combined with pleural fluid analysis via thoracentesis. Challenges in CXR Interpretation and the Role of AI Despite its invaluable diagnostic role, chest X-ray interpretation remains fraught with challenges. Variability in image acquisition (e.g., anterior-posterior vs. posterior-anterior views), patient factors (e.g., obesity, inability to inspire deeply), and subtle early-stage disease findings contribute to potential diagnostic uncertainty. Furthermore, overlapping features among different diseases—for instance, diffuse opacities in both COVID-19 and ARDS—make visual differentiation difficult even for experienced radiologists. In response to these limitations, artificial intelligence (AI)-driven solutions have emerged as powerful tools to augment diagnostic accuracy. Deep learning models, particularly convolutional neural networks (CNNs) and vision transformers (ViTs), have demonstrated considerable promise in analyzing CXR images. These models can automatically detect radiographic patterns associated with specific diseases, highlight regions of interest using attention maps, and even generate textual interpretations akin to radiology reports. By learning from large annotated datasets, such AI systems can extract hierarchical features that transcend simple pixel-level differences, capturing the complex visual patterns indicative of specific pathological states. In particular, transformer-based models have shown an ability to integrate image features with natural language generation, enabling the translation of visual inputs into coherent textual descriptions—an approach that underpins the goal of chest X-ray image-to-text transformation projects.

2 Related Work

- Pnemonia detection using deep learning

a deep learning model that achieves radiologist-level performance in detecting pneumonia from chest X-rays. The model is a 121-layer Dense Convolutional Network (DenseNet-121) trained on the ChestX-ray14 dataset, which contains over 100,000 frontal-view chest X-rays labeled with 14 different pathologies. To demonstrate how a deep learning system can match or even outperform expert radiologists. To achieve this, the authors trained CheXNet to predict all 14

pathologies and then fine-tuned it specifically for pneumonia detection. A key contribution is the comparison between CheXNet and four radiologists, where the model performed slightly better than the average expert in terms of F1 score. CheXNet uses transfer learning, where a model pretrained on ImageNet is adapted to chest X-ray images. The paper also employs class activation maps to visualize regions in the X-rays that are most relevant to the model's predictions, aiding interpretability. The results showed that CheXNet can serve as a reliable tool for screening and triage, especially in areas with limited access to radiologists. The study suggests that deep learning has the potential to support or partially automate medical image interpretation.

- "Automated Chest X-ray Radiology Report Generation"

a model for automated generation of radiology reports from chest x-ray, aiming to replicate human-level descriptive capability using deep learning. The system combines a Convolutional Neural Network (CNN) for image feature extraction and a Recurrent Neural Network, specifically Long Short-Term Memory, for report generation. The dataset used is IU X-Ray, which contains X-ray images paired with structured radiology reports. The authors propose a two-stage pipeline: (1) image encoder using CNN to obtain image embeddings, and (2) report decoder using LSTM to generate textual findings. Attention mechanisms are integrated to help the decoder focus on relevant image areas while generating each word. Performance is measured using BLEU, METEOR, and ROUGE scores. The model showed promising results in terms of linguistic fluency and medical accuracy, though it still struggles with rare findings and fine-grained nuances. The paper emphasizes the potential of AI in clinical documentation, reducing workload and improving consistency. Limitations include data scarcity and challenges in accurately modeling diverse medical terminology.

- "Transformer-Based Chest X-ray Report Generation"

This work explores the application of Transformer architectures for generating radiology reports from chest X-rays, aiming to improve upon traditional RNN-based methods. The authors propose a Vision Transformer (ViT) + Transformer decoder model that directly generates full reports. Using the MIMIC-CXR dataset, the system maps image patches to embeddings via a ViT encoder, then feeds these into a Transformer decoder to produce natural language reports. This setup allows for better handling of long-range dependencies in text and fine-grained image features. Results are evaluated using BLEU, ROUGE-L, and CIDEr. The proposed method outperforms RNN-based baselines and matches clinical accuracy in many cases. Visualizations of attention weights show that the model effectively links image regions to relevant report content. The study concludes that pure Transformer models, while computationally intensive, are superior in coherence, accuracy, and scalability for medical text generation. Challenges remain in aligning predictions with clinically correct language and incorporating domain-specific knowledge.

- "Clinically Accurate Chest X-ray Report Generation with Knowledge Graphs"

This paper enhances chest X-ray report generation by incorporating medical knowledge graphs into a Transformer-based pipeline. The model, called KERP (Knowledge Enhanced Report Parser), integrates domain-specific knowledge to improve accuracy and reduce factual errors. KERP uses a three-step process: (1) a graph encoder creates medical entity embeddings from the knowledge base, (2) a visual encoder extracts features from images, and (3) a Transformer decoder generates reports using a fusion of visual and graph-based knowledge. The dataset used is MIMIC-CXR. The system outperforms other methods in generating clinically accurate and coherent reports, especially in rare or subtle disease cases. Evaluation includes BLEU, ROUGE, and a newly proposed clinical accuracy score. This approach highlights the importance of domain-specific knowledge in medical AI systems. The fusion of structured medical knowledge with image features leads to reports that are more aligned with real clinical interpretations. Limitations include the static nature of the graph and incomplete knowledge coverage.

- ”AlignTransformer: Alignment-Aware Transformer for Chest X-ray Report Generation”

Align Transformer is a novel architecture designed to improve alignment between image features and textual descriptions in chest X-ray report generation. The key idea is to explicitly model the alignment between image regions and phrases in the report, which traditional models often ignore. The model uses a standard CNN (e.g., ResNet-101) to encode the image, followed by an alignment-aware Transformer decoder that emphasizes cross-modal relationships. It introduces an alignment loss function to guide the training process towards better correspondence between visual and textual elements. Using the IU X-Ray and MIMIC-CXR datasets, the model achieves higher BLEU, METEOR, and ROUGE scores compared to state-of-the-art baselines. Visualizations show that the model better grounds textual tokens in specific image regions, making the reports more interpretable. The paper concludes that alignment-aware modeling significantly enhances clinical relevance and interpretability of generated reports. Limitations include increased complexity and longer training times.

- ”Exploring the Limits of Chest X-ray Report Generation with GPT”

IN This investigates the capabilities of large language models (LLMs), especially GPT, in generating radiology reports from chest X-ray images. The authors integrate image features from CNNs or ViTs with GPT-style decoders to examine how well general-purpose LLMs perform in a medical setting. The architecture includes an image encoder followed by a frozen or fine-tuned GPT decoder trained to generate findings, impressions, and recommendations. The model is evaluated using standard metrics like BLEU, ROUGE, and clinical correctness by expert radiologists. While GPT performs well in terms of fluency and general structure, it often hallucinates medical facts not grounded in the image. Fine-tuning with medical data reduces these errors but doesn’t eliminate them. The paper also discusses prompt engineering and transfer learning as ways to

adapt general models to medical tasks. The study concludes that while GPT shows promise, domain adaptation and medical grounding are crucial for clinical safety. LLMs are not yet reliable as standalone diagnostic tools but could assist radiologists as writing aids.

- ”Uncertainty-Aware Chest X-ray Report Generation”

In This proposes an uncertainty-aware approach to chest X-ray report generation. The key idea is to quantify the confidence of the AI system when generating each sentence in the report, helping doctors identify which parts are more reliable. The system combines a CNN image encoder, a Transformer-based decoder, and an uncertainty estimation module. This module uses techniques like Monte Carlo Dropout to produce confidence intervals for the generated content. Trained on the dataset, the model shows comparable language quality to previous methods but provides extra information about prediction reliability. This is crucial in clinical settings where overconfidence in incorrect results can be harmful. By integrating uncertainty scores with generated text, the model enables clinicians to better interpret and validate AI outputs. Limitations include increased computational cost and difficulty in calibrating uncertainty measures.

3 Proposed Methodology

details of the proposed model for robust and effcient classifcation of Covid-19 disease from input chest x ray.

1. **Dataset description** A dataset based on chest X-rays is used in this study. To complete the classification task, 10,874 X-ray images in PNG (portable network graphics) format are used. The size of the input image is set to $224 \times 224 \times 3$.One dataset is created by combining the three different chest radiographs of lung diseases. All photos are from publicly available sources. Three categories are used to group all of the samples: training , testing , and validation. A strong and deep effcient model is developed.

2. **Data Preprocessing**

   1. The goal of image resizing and scaling is to uniformize input dimensions throughout the dataset. Method: Depending on the model (e.g., ResNet, DenseNet, ViT), all CXR images are downsized to a specific resolution, usually 224×224 or 512×512 pixels.
   Rescaling: Pixel values are frequently standardized using mean and standard deviation (e.g., ImageNet values) or normalized to a [0, 1] range.

   2. Contrast Enhancement Histogram Equalization: This technique disperses intensity values to improve contrast. Contrast Limited Adaptive Histogram Equalization, is a better local technique that is frequently applied in medical imaging to improve soft tissue contrast.

   3. Diminution of Noise Gaussian blurring, also known as median filtering,

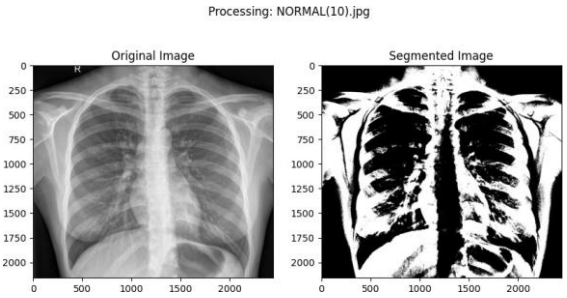

Figure 1: Preprocessing image

eliminates random noise without obscuring significant patterns.aids in removing artifacts from images, which is crucial when utilizing portable or low-quality X-ray equipment.

3. **Dataset description** This work uses a specially curated dataset of frontal chest X-ray (CXR) images along with matched textual radiology reports in the cross-modal translation from radiography to descriptive text. The data set is obtained from publicly accessible data sources like the ChestX-ray14, normal X-ray image, and COVID19, pneumonia, consolidation, pleural effusion datasets.

4 Ferature Extraction

- Feature extraction serves as an important aspect of bridging the visual and textual modalities in our chest X-ray image-to-report transformation pipeline. In this work, we utilize the merits of two of the most current deep convolutional and transformer-based architectures—Vision Transformer (ViT-B/16) and CheXNet (DenseNet-121)—to extract high-level, semantically dense feature representations from preprocessed chest radiographs.

- Vision Transformer is a transformer-based model that uses the self-attention operation for image patches as an alternative to convolutional neural networks. We specifically use the ViT-B/16 variant, which separates input images into non-overlapping 16×16 patches, embeds them into linear space, and processes them via stacked encoder transformer blocks. Input Image Size: $224 \times 224 \times 3$ Patch Size: 16×16 (196 patches in total) Output Dimension: 768 (for the [CLS] token and each patch) Feature Vector Utilized: Output embedding for the [CLS] token Output Shape: (1, 768) The ViT model is pre-trained on ImageNet-21k and fine-tuned on ImageNet-1k. We obtain the last [CLS] token embedding after passing the

image through all the encoder layers, which captures a global contextual representation of the image appropriate for downstream tasks like report generation.

- CheXNet (DenseNet-121) CheXNet is a DenseNet-121 model pretrained on the ChestX-ray14 dataset alone for classification of thoracic diseases. Its convolutional backbone can extract spatially dense clinical abnormality-relevant features.

  Input Image Size: $224 \times 224 \times 3$

  Final Convolution Output: (1024, 7, 7)

  Adaptive Average Pooling: Pooled output to (1024, 1, 1)

  Flattened Feature Vector: 1024-dimensional

  Output Shape: (1, 1024)

  In order to use CheXNet as a feature extractor, we remove the classification head and employ the penultimate feature map. The feature tensor is pooled and flattened to get a dense feature representation encapsulating the diagnostic content of the image.

5 Model Architecture

CheXNet Model Architecture

Input: $3 \times 224 \times 224$ (Chest X-ray image, RGB) 1. Initial Convolution and Pooling Layers

Layer Type	Output Shape	Kernel/Stride/Pad	Description
Conv2d	$64 \times 112 \times 112$	7×7 / 2 / 3	Initial convolution layer
BatchNorm2d	$64 \times 112 \times 112$	-	Batch normalization
ReLU	$64 \times 112 \times 112$	-	Activation
MaxPool2d	$64 \times 56 \times 56$	3×3 / 2 / 1	Downsampling

Table 1: Neural network layer specifications.

Dense Block 1 + Transition Layer 1

Component	**Output Shape**	**Description**
Dense Block 1	$256 \times 56 \times 56$	6 dense layers
Transition Layer 1	$128 \times 28 \times 28$	1×1 conv + avg pool

Table 2: Placeholder caption for component description and output shape

Dense Block 2+ Transition Layer 2

- Input: 128 × 28 × 28 (from Transition Layer 1)
- Number of Layers: 12 Dense Layers
- Growth Rate: 32 (Each layer adds 32 channels)
- Output Channels: 128 (input) + 12 × 32 (new channels) = 512 channels

Component	Output Shape	Description
Dense Block 2	512 × 28 × 28	12 dense layers
Transition Layer 2	256 × 14 × 14	1×1 conv + avg pool

Table 3: Description of Network Components

Vision Transformer (ViT-B/16)

The ViT-Base model meets performance on par by utilizing global attention mechanisms to represent long-distance relations across image areas. Its design eschews convolution operations altogether and instead relies on patch embeddings, self-attention, and deep Transformer encoders to obtain semantic representations of visual information. Its design is especially potent when pretrained on big datasets and fine-tuned for applications such as medical image interpretation, including diagnosis on chest X-rays.

Layer No.	Layer Type	Input Shape	Output Shape
1	Input Image	(3, 224, 224)	
2	Patch Split + Flatten	(3, 224, 224)	(196, 768)
3	Linear Projection	(196, 768)	(196, 768)
4	Class Token [CLS]	(196, 768)	(197, 768)
5	Position Embedding	(197, 768)	(197, 768)
6-29	Transformer Encoder ×12	(197, 768)	(197, 768)
	— LayerNorm	(197, 768)	(197, 768)
	— Multi-Head Attention	(197, 768)	(197, 768)
	— Skip Connection	(197, 768)	(197, 768)
	— LayerNorm	(197, 768)	(197, 768)
	— MLP (Linear → GELU → Linear)	(197, 768)	(197, 768)
	— Skip Connection	(197, 768)	(197, 768)
30	Final LayerNorm	(197, 768)	(197, 768)
31	CLS Token Extraction	(197, 768)	(768,)
32	Classification Head (optional)	(768,)	(num_classes,)

Table 4: Model Architecture Layer Details

Vision Transformer Algorithm

1. Patch Embedding

Convert an input image $I \in R^{3\times H\times W}$ into a sequence of flattened patches:

$$x_p = Flatten(Patch(I)) \in R^{P^2\cdot C}$$

Apply a trainable linear projection:

$$z_0 = x_p W_e + b$$

2. Self-Attention Mechanism

For each patch embedding x, compute query, key, and value vectors:

$$Q = xW_Q, \quad K = xW_K, \quad V = xW_V$$

Compute the attention weights and apply them to the values:

$$Attention(Q, K, V) = softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

3. Multi-Head Self-Attention (MHSA)

Split the input into h heads, perform attention in parallel, and concatenate the results:

$$MHSA(X) = Concat(head_1, \ldots, head_h)W_O$$

4. Feed-Forward Network (FFN)

A two-layer MLP with a GELU activation function:

$$FFN(x) = Linear_2(GELU(Linear_1(x)))$$

Algorithm: ViT-Base

Input: RGB Image $x \in R^{3\times 224\times 224}$
Output: Feature vector $f \in R^{768}$ or classification vector $y \in R^C$ function ViT-Base(x):
Divide x into 16×16 non-overlapping patches
Flatten patches and apply linear projection → z
Add class token [CLS] to z
Add positional encoding to z
for each of the 12 Transformer Encoder layers:
Apply LayerNorm → Multi-Head Attention → Residual
Apply LayerNorm → MLP → Residual
end for

Extract [CLS] token output as feature vector f
if classification task then
Apply classification head → y
return y
else
return f
end if
end function

Mathematical Description of the Grad-CAM

The Gradient-weighted Class Activation Mapping (Grad-CAM) visualizes the spatial importance of each region of an input image for a specific class prediction. It does this by computing the gradient of the output class score with respect to the feature maps of a convolutional layer.

Let:

- y^c be the class score (e.g., probability or logit) for class c.
- $A^k \in R^{H \times W}$ be the k-th feature map of a convolutional layer.
- α_k^c be the importance weight for feature map k with respect to class c.

Step 1: Compute Gradients

Compute the gradient of the class score with respect to the feature maps:

$$\frac{\partial y^c}{\partial A^k}$$

Step 2: Global Average Pooling Over Gradients

Compute the importance weights:

$$\alpha_k^c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y^c}{\partial A_{ij}^k}$$

Step 3: Compute Weighted Combination of Feature Maps

The class activation map is obtained as:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k \right)$$

Here, ReLU is applied to retain only positive influences that contribute positively to the class score.

6 Result and Discussion

Self-Attention in Vision Transformers (Viit)

- Vision Transformers (ViTs) use self-attention to weigh the importance of different image patches relative to one another.
- The image is divided into patches (e.g., 16×16), embedded, and fed into a Transformer encoder.
- At each layer, self-attention maps determine how each patch attends to every other patch.
- These maps can be aggregated (e.g., using attention rollout) to visualize overall focus.
- Attention Rollout Technique:
- This method propagates attention across layers to determine how the output class token depends on input patches. It provides a holistic view of spatial dependencies learned by the Transformer.
- Clinical Relevance of Attention Maps Attention heatmaps serve not only as interpretability tools but also aid in:
- Feature localization: Helps the model attend to pathologically relevant structures.
- Model trustworthiness: Provides clinicians visual evidence for AI-driven decisions.
- Training supervision: In weakly supervised learning, attention maps act as pseudo-labels.
- Dataset annotation: Radiologists can validate attention maps to refine annotations.

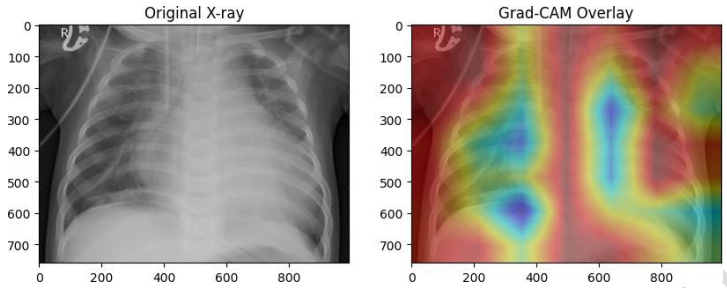

Figure 2: Grad-CAM

Prediction and Attention Visualization Result

Figure 3 illustrates an example output of the proposed image-to-text generation model for chest X-rays. On the left, the input is a frontal chest X-ray image of a normal subject, displaying clear lung fields without any radiographic signs of pathology. In the center, the colored grid represents an attention heatmap overlay derived from the Transformer decoder during the text generation process. The heatmap highlights the regions of the image that were most influential in the model's prediction, particularly focusing on the central thoracic zone corresponding to the lung fields and mediastinum.

On the right, the generated textual report reads:

"The lungs are clear. No pleural effusion, pneumothorax or focal air-space disease."

This output demonstrates the model's ability to not only identify normal anatomical structures but also to rule out critical pathologies such as pleural effusion, pneumothorax, or focal consolidation. The attention map further confirms that the model is attending to medically relevant areas of the chest X-ray during inference, thereby reinforcing the interpretability and clinical plausibility of the generated report.

This result exemplifies the effectiveness of the Vision Transformer (ViT) feature extractor combined with a Transformer decoder in producing coherent, medically accurate, and interpretable radiological summaries.

This generated report mirrors common language used by radiologists in normal chest X-ray assessments. The absence of findings such as pleural effusion, pneumothorax, and air-space disease (e.g., pneumonia or consolidation) indicates a normal study. The specificity and clarity of this output demonstrate the model's capacity for both diagnostic accuracy and clinically relevant language generation.

chest X-ray image-to-text transformation using a Vision Transformer and Transformer decoder. The left panel shows the input image; the middle panel is the attention heatmap generated during decoding; the right panel shows the automatically generated report. The attention mechanism effectively focuses on clinically relevant thoracic regions.

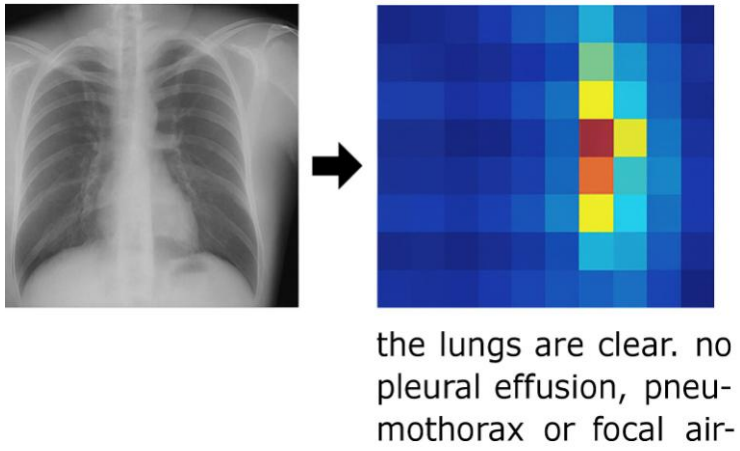

Figure 3: Text transformation

Gradient-weighted class Activation Mapping (Grad-CAM)

As shown in fig.4 Grad-CAM visualizes the spatial importance of each region of an input image for a specific class prediction. It does so by computing the gradient of the output class score with respect to the feature maps of a convolutional neural network (CNN), and generating a heatmap that localizes the most discriminative regions. In fig.4 Show the Grad-CAM image in different region show is in different color

Context in chest x-ray Image-to-Test Project

- The input image ***I*** is a preprocessed chest X-ray that has undergone segmentation and enhancement.
- A CNN backbone (e.g., ResNet or hybrid ViT with convolutional stem) is used to extract image features A^k.
- A Transformer decoder generates medical reports based on these features.
- Grad-CAM is applied to the CNN encoder to identify the spatial regions that most strongly influenced the encoded features for a given class (e.g., pneumonia, pleural effusion).

This visualization serves as an interpretability tool to validate that the model is attending to clinically relevant anatomical structures, thereby enhancing model transparency in a critical domain like radiology.

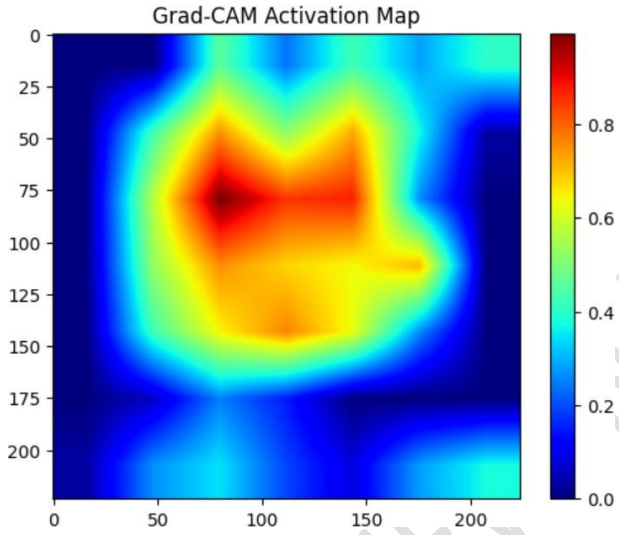

Figure 4: Grad-CAM

Mapping and Visualization

- The resulting heatmap $L^c_{Grad-CAM}$ is upsampled to the input resolution and overlaid on the original image.
- A jet colormap is used to visualize:
  - **Red/Yellow** regions $\Rightarrow L^c_{Grad-CAM} \approx 1$: High influence zones.
  - **Blue** regions $\Rightarrow L^c_{Grad-CAM} \approx 0$: Low influence zones.

This visualization serves as an interpretability tool to validate that the model is attending to clinically relevant anatomical structures, thereby enhancing model transparency in a critical domain like radiology.

7 References

1. Hage Chehade, Aya, et al. ”A systematic review: Classification of lung diseases from chest X-ray images using deep learning algorithms.” SN Computer Science 5.4 (2024): 405.

2. Al-qaness, Mohammed AA, et al. ”Chest x-ray images for lung disease detection using deep learning techniques: a comprehensive survey.” Archives of Computational Methods in Engineering 31.6 (2024): 3267-3301.

3. 2D-to-3D: A Review for Computational 3D Image Reconstruction from X-ray Images

4. Mann, Mukesh, et al. ”Utilization of deep convolutional neural networks for accurate chest X-ray diagnosis and disease detection.” Interdisciplinary Sciences: Computational Life Sciences 15.3 (2023): 374-392.

5. Hage Chehade, Aya, et al. ”A systematic review: Classification of lung diseases from chest X-ray images using deep learning algorithms.” SN Computer Science 5.4 (2024): 405.

6. Al-Hammuri, Khalid, et al. ”Vision transformer architecture and applications in digital health: a tutorial and survey.” Visual computing for industry, biomedicine, and art 6.1 (2023): 14.

7. Shin, Hoo-Chang, et al. ”Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

8. Wang, Xiaosong, et al. ”Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

9. Ahmed, Saad Bin, Roberto Solis-Oba, and Lucian Ilie. ”Explainable-ai in automated medical report generation using chest x-ray images.” Applied Sciences 12.22 (2022): 11750.

10. Rajasenbagam, T., S. Jeyanthi, and J. Arun Pandian. ”Detection of pneumonia infection in lungs from chest X-ray images using deep convolutional neural network and content-based image retrieval techniques.” Journal of Ambient Intelligence and Humanized Computing (2021): 1-8.

11. Rajpurkar, Pranav, et al. ”Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning.” arXiv preprint arXiv:1711.05225 (2017).

12. Zeyer, Albert, et al. ”A comparison of transformer and lstm encoder decoder models for asr.” 2019 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE, 2019.

13. Goyal, S., and R. Singh. ”Detection and classification of lung diseases for pneumonia and Covid-19 using machine and deep learning techniques. J Ambient Intell Human Comput 14, 3239–3259 (2023).”

14. Asnake, Nigus Wereta, Ayodeji Olalekan Salau, and Aleka Melese Ayalew. ”X-ray image-based pneumonia detection and classification using deep learning.” Multimedia Tools and Applications 83.21 (2024): 60789-60807.

15. Puttagunta, Muralikrishna, and S. Ravi. ”Medical image analysis based on deep learning approach.” Multimedia tools and applications 80.16 (2021): 24365-24398.

16. Rana, Meghavi, and Megha Bhushan. ”Machine learning and deep learning approach for medical image analysis: diagnosis to detection.” Multimedia Tools and Applications 82.17 (2023): 26731-26769.

17. Automated COVID-19 Detection from Chest X-Ray Images: A High-Resolution Network (HRNet) Approach

18. Guefrechi, Sarra, et al. ”Deep learning based detection of COVID-19 from chest X-ray images.” Multimedia tools and applications 80 (2021): 31803-31820.

19. Tekerek, Adem, and Ismael Abdullah Mohammed Al-Rawe. ”A novel approach for prediction of lung disease using chest x-ray images based on DenseNet and MobileNet.” Wireless Personal Communications (2023): 1-15.

20. Cho, Kyungjin, et al. ”CHESS: Chest X-Ray Pre-trained model via self-supervised contrastive learning.” Journal of Digital Imaging 36.3 (2023): 902-910.