Automated Chest X-ray Report Generation Using Attention-Enhanced GoogleNet-LSTM Architecture

Muhammad Faheem Khalil Parach[1*], Mudasir Mahmood[1], Muhammad Farhan[1,], Muhammad Umar Sohail[1], Syed Muhammad Ali Shah[1]

[1]Gomal Research Institute of Computing, Faculty of Computing, Gomal University, D.I. Khan-29050, KPK, Pakistan
*Corresponding author: Muhammad Faheem Khalil Paracha: faheemparacha.uettaxila@gmail.com

Abstract: Chest radiography remains a cornerstone of clinical diagnostics, yet its interpretation is time-consuming and dependent on specialized expertise. The growing shortage of radiologists, combined with the increasing volume of imaging exams, often leads to delays and inconsistencies, highlighting the need for automated solutions. In this study, we present an automated framework for generating diagnostic reports directly from chest X-rays. The model uses GoogleNet for visual feature extraction and a long short-term memory (LSTM) network to generate reports. An attention mechanism is incorporated to focus on clinically relevant image regions. The framework was evaluated on the publicly available Indiana University (IU) Chest X-ray dataset, with performance assessed using language-based metrics (BLEU, ROUGE-L, METEOR, CIDEr) and clinical accuracy indicators, such as precision, recall, and F1-score. Results demonstrated that the attention-based architecture outperformed baseline encoder-decoder models, particularly in CIDEr and clinical F1 metrics, suggesting the reports were more fluent and clinically accurate. Attention maps showed alignment with key image areas, such as the cardiac silhouette for cardiomegaly and costophrenic angles for pleural effusion. While limitations were observed in handling rare conditions and occasional generic phrasing, the framework effectively improved the efficiency and consistency of radiology reporting.

Keywords: Chest X-ray, Radiology Report Generation, Deep Learning, GoogleNet, LSTM, Attention Mechanism, Automated Diagnosis.

I. Introduction

Chest radiographs remain one of the most widely used and cost-effective tools in medical diagnosis. They play a central role in detecting and monitoring conditions such as pneumonia, tuberculosis, pleural effusion, pneumothorax, and lung cancer. Despite their importance, the process of interpreting these images is far from straightforward. Accurate reading requires specialized knowledge of thoracic anatomy, careful recognition of subtle abnormalities, and the ability to relate visual findings to a patient's clinical history. In many healthcare systems, especially in countries with limited resources, the shortage of trained radiologists leads to significant delays in diagnosis, inconsistent reporting quality, and increased likelihood of human error. These challenges are magnified in overcrowded hospitals and diagnostic centers where hundreds of scans may need to be examined daily [1]. An example Manual report generated by a Radiologist is given below in Figure 1.

Figure 1: An example chest x-ray report by a Radiologist

Doctor Findings:
Stable appearance of the cardio mediastinal silhouette.

There is no pneumothorax, pleural effusion, or focal airspace consolidation.

Manual reporting is not only time-consuming but also subject to considerable variability between observers. Two radiologists may describe the same chest X-ray differently, which can complicate treatment planning and follow-up care. Moreover, in rural or underserved regions, access to expert interpretation is often unavailable, forcing non-specialists to produce reports that may lack precision. Such circumstances underline the urgent need for technological solutions that can support radiologists in routine reporting, reduce diagnostic errors, and provide reliable services in places where medical expertise is scarce [2].

Advances in artificial intelligence (AI) and deep learning have created opportunities to address these limitations. Convolutional neural networks (CNNs) are now capable of extracting high-level features from complex medical images, while recurrent neural networks (RNNs) such as Long Short-Term Memory (LSTM) models can generate structured and coherent text. When combined, these models offer a pathway for automated report generation, allowing chest radiographs to be transformed directly into descriptive diagnostic narratives. An additional breakthrough has been the development of attention mechanisms, which enable the model to focus selectively on the most relevant regions of an image, thereby improving both accuracy and interpretability [3].

Recent research has shown the promise of applying these techniques to the medical domain. Large-scale datasets, such as the Indiana University (IU) Chest X-ray collection, CheXpert, and MIMIC-CXR, have provided valuable resources for training and evaluating automated systems. However, challenges remain. Models trained on a single dataset often fail to generalize across diverse imaging sources. Standard evaluation metrics, including BLEU, ROUGE, and CIDEr, measure textual similarity but may not adequately reflect the clinical correctness of generated reports. Furthermore, most existing systems do not incorporate longitudinal data or non-imaging information such as patient history, which limits their contextual accuracy.

This study proposes an integrated deep learning framework that leverages GoogleNet for image feature extraction, LSTM networks for text generation, and an attention mechanism to guide focus toward clinically important regions. The goal is to produce coherent and medically relevant diagnostic reports that can support radiologists in decision-making and improve diagnostic efficiency in high-volume or resource-limited environments. By testing the proposed model on benchmark datasets and comparing its performance with traditional encoder–decoder approaches, the research aims to demonstrate both the feasibility and the potential benefits of automated radiology reporting [4].

The significance of this work lies in its contribution to improving diagnostic accuracy, reducing the workload of healthcare professionals, and making expert-level reporting more accessible through scalable AI solutions. Beyond academic interest, such a system can be readily applied in telemedicine platforms, offering vital diagnostic support to rural and underserved populations. Ultimately, the integration of explainable and clinically reliable AI into radiology has the potential to transform how medical imaging is interpreted, creating a future where faster, more consistent, and more accessible healthcare is within reach. [5]

II. Related Work

Automating radiology reporting has attracted sustained attention because chest radiographs are inexpensive and ubiquitous, yet their correct interpretation requires skill and time. Over the last decade approaches moved from straightforward encoder–decoder captioning to sophisticated, clinically-aware architectures that combine visual attention, retrieval/memory components, and efficiency improvements. This section organizes the recent progress and highlights remaining gaps.

Early work adapted image-captioning ideas (CNN encoder + RNN decoder) to medical images. Those models showed the feasibility of converting images to text but often produced short, superficial captions rather than the structured, multi-sentence diagnostic narratives radiologists write. To address this mismatch, researchers introduced hierarchical decoders (sentence + word levels) and attention that ties specific image regions to generated phrases; these changes improved coherence and helped localize relevant visual evidence [6]. A key trend has been explicit modeling of the image↔text alignment rather than relying solely on end-to-end generation. Cross-modal Memory Networks (CMN) store and retrieve aligned visual–textual patterns to help the decoder generate clinically accurate findings; Chen et al. showed that a memory module improves alignment and clinical metrics on IU X-Ray and MIMIC-CXR [7]. Retrieval-guided methods are another direction: MedWriter uses a hierarchical retrieval mechanism to pull whole reports and sentence templates from a corpus, then adapts them to the input image. This

reduces hallucination and enforces clinically plausible phrasing while keeping fluent generation. MedWriter demonstrated improved clinical accuracy compared to purely generative baselines [8].
Attention modules remain central. Works combining spatial/channel attention and cross-attention help the model focus on radiologically important regions and better align visual features with words. CheXPrune showed that aggressive one-shot pruning combined with attention can compress a report-generation model (≈70% pruning, ~3.3× compression) without substantial loss of BLEU/ROUGE/CIDEr scores, making on-device or low-resource deployment more feasible. This line of work acknowledges that clinical settings often lack large compute budgets [9-[11]. Transformers and hybrid ViT→language decoders are now widely used because they capture long-range dependencies and global context more effectively than RNN-only systems. Architectures such as Med-Former (local-global transformer + spatial attention fusion) and ViT→GPT2 hybrids (VisionGPT/ViGPT2 variants) show competitive performance on classification and captioning tasks, and they are being adapted for report generation. However, vanilla Transformer models raise interpretability concerns: global attention helps fluency, but clinicians demand transparent links between image evidence and textual claims [12, 13].
Standard NLG metrics (BLEU, ROUGE, CIDEr, METEOR) measure lexical overlap but often miss clinical correctness a generated sentence can score well while omitting or misreporting key findings. Several works therefore complement these metrics with clinically focused evaluations (concept extraction, disease-level precision/recall). Surveys and recent reviews emphasize that a combination of lexical, semantic, and clinical metrics is essential to judge real utility [14].
Chen et al. (Cross-modal Memory Networks (CMN)) Introduced a shared memory to explicitly encode image–text correspondences, improving clinical alignment and pushing state-of-the-art on IU X-Ray and MIMIC-CXR in standard and clinical metrics [15,16]. The memory helps the decoder recall clinically relevant visual–textual pairs rather than inventing text solely from latent vectors. Yang et al. (MedWriter (Hierarchical Retrieval)) Uses visual-language retrieval to pull whole-report and sentence templates and guides a hierarchical decoder. This reduces hallucination and yields better clinical fidelity on OpenI and MIMIC benchmarks. Kaur & Mittal (CheXPrune) gives a pruning + multi-attention strategy that compresses models heavily while keeping comparable language scores; demonstrates that model size can be reduced without losing much performance, an important practical insight for resource-limited deployment. Gu et al. (CVAM + MVSL (Cross-View Attention + Medical Visual-Semantic LSTMs)) Targets multi-view chest X-rays; cross-view attention fuses complementary viewpoints and MVSL integrates visual and sentence-level semantics, improving multi-view report accuracy. Med-Former and related Transformer hybrids are Local–global transformer modules and spatial attention fusion modules offer improved classification and representation learning for medical images; they form a basis for applying transformer decoders or retrieval modules to report generation [7,8,9,11,12]. A detailed summary of most recently used models and datasets details are given in Table 1.

Table 1: Selected datasets and influential models for chest X-ray report generation

Name (type)	Short Description / size	Typical Tasks & Notes	Representative Papers
MIMIC-CXR (dataset) [17]	≈377k images, ≈227k studies with free-text reports.	Large, multi-study dataset for report generation and classification; de-identified, widely used.	Johnson et al., MIMIC-CXR (2019).
CheXpert (dataset) [18]	224,316 radiographs, 65,240 patients; uncertainty labels for 14 observations.	Useful for classification and weakly supervised labeling; includes frontal & lateral views.	Irvin et al., CheXpert (2019).
NIH ChestX-ray14 (dataset) [19]	112,120 images, 14 disease labels.	Early large public dataset: widely used but label noise is known.	Wang et al., ChestX-ray14 (2017).

IU Chest X-Ray (dataset) [20]	~8,121 images with ~3,996 reports; radiologist-written.	Smaller but high-quality paired reports; often used for report generation research.	Demner-Fushman / Open-I resources.
PadChest (dataset) [21]	>160,000 images from ~67,000 patients (Spain); bilingual and detailed annotations.	Rich labels and localization supports multi-view and demographic metadata.	Bustos et al., PadChest (2020).
Cross-modal Memory Networks (CMN) (model) [7]	Shared memory to align image and text; improves clinical alignment and metrics on IU & MIMIC.	Addresses explicit mapping between visual evidence and report text.	Chen et al., ACL/EMNLP works (2021/2022).
MedWriter (model) [22]	Hierarchical retrieval + generation using report/sentence templates.	Reduces hallucination; improves clinical accuracy on OpenI/MIMIC.	Yang et al., ACL 2021.
CheXPrune (model) [11]	One-shot global pruning + multi-attention; ~70% pruning w/o large accuracy drop.	Demonstrates feasibility of compressed, deployable report generators.	Kaur & Mittal (2022/2023).
CVAM + MVSL (model) [23]	Cross-view attention + medical visual-semantic LSTMs for multi-view images.	Improves use of multiple projections for report generation.	Gu et al., 2023.
Med-Former / ViT→GPT hybrids (model class) [12]	Local-global transformers, ViT encoders + language decoders.	Strong global context modeling; better fluency but interpretability concerns.	Chowdary et al., Med-Former (2024); various ViT→GPT works.

III.Methodology

This section describes the proposed encoder–decoder framework in full detail. First, we present notation and data preprocessing, then the image encoder (GoogleNet/Inception), the attention module, the LSTM-based decoder with its mathematical formulation, and finally the training and inference strategies. Proposed model comprising GoogleNet and LSTM is given in Figure 2.

a. Notation and problem formulation

Let an input chest radiograph be denoted by $I \in \mathbb{R}^{H\times W\times C}$ (height H, width W, channels C). The paired ground-truth radiology report is a sequence of tokens (words) $Y = (y_1, y_2, \dots, y_T)$, where T is the report length and each y_t belongs to a fixed vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$.

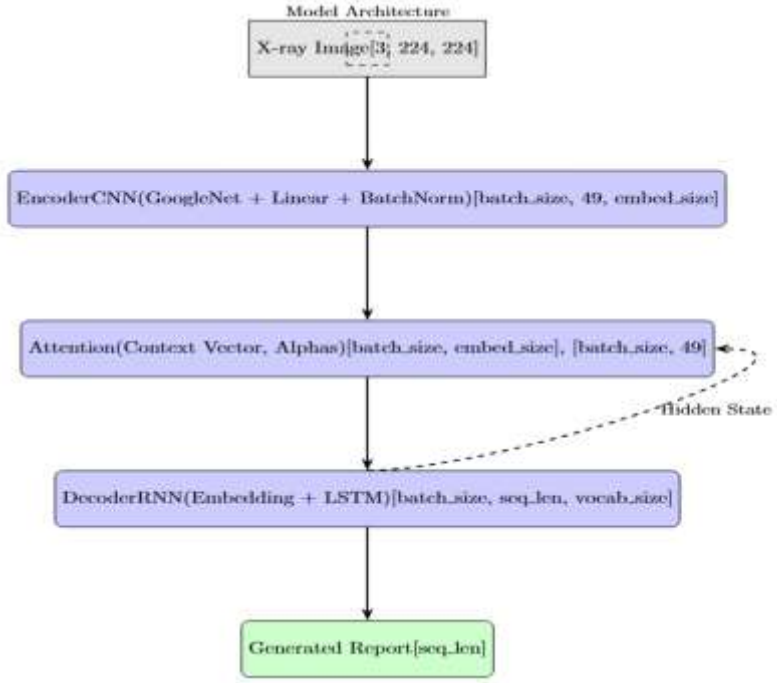
The model follows an encoder–decoder paradigm:

- Encoder: $f_{\text{enc}}(I) \rightarrow A$, where $A = (a_1, a_2, \dots, a_L)$ is a set of spatial feature vectors (flattened spatial map) with L spatial locations and $a_i \in \mathbb{R}^D$ (feature dimension D). Typically $L = H' \times W'$ after CNN downsampling.
- Decoder: Given A, the decoder (LSTM) generates tokens sequentially:

$$P(Y \mid I) = \prod_{t=1}^{T} P\left(y_t \mid y_{1:t-1}, A\right)$$

An attention mechanism computes a context vector c_t at each decoding step to condition the LSTM on relevant image regions [25].

Figure 2: A fully integrated model combining CNN-based image embeddings with an LSTM decoder and attention mechanism for word generation.



The LSTM is shown in its unrolled form, with all time-step instances sharing the same parameters.

b. Data preprocessing

In preparing the data for our experiments, we first resized all chest X-ray images to a fixed resolution of 512×512 pixels and normalized the pixel intensities channel-wise using the dataset mean and standard deviation. To improve the robustness of our model, we applied light data augmentation during training, including random cropping, slight rotations, and horizontal flips when appropriate. For the text reports, I converted all text to lowercase, removed non-informative characters, split sentences, and tokenized them into words or subwords. I then built a vocabulary by keeping the most frequent tokens, mapping rare words to the <UNK> symbol, and adding special tokens <SOS>, <EOS>, and <PAD> to indicate sequence boundaries and padding. After preprocessing the images, I passed each chest X-ray through the GoogleNet encoder to obtain a convolutional feature map of size (D, H', W'). Finally, I reshaped this tensor into a matrix $\boldsymbol{A} \in \mathbb{R}^{L \times D}$, where $L = H' \cdot W'$ corresponds to the flattened spatial grid of image regions and D represents the feature depth.

c. Visual Encoder (GoogleNet / Inception)

GoogleNet (Inception-v1) is used as the image encoder because of its multi-scale convolutional blocks which are useful for capturing both local and context features in radiographs. Concretely, let the encoder mapping be:

$$F = f_{\text{enc}}(I; \theta_{\text{enc}}) \in \mathbb{R}^{D \times H' \times W'}.$$

Reshape to spatial feature matrix:

$$A = (a_1, a_2, \dots, a_L)^{\top} \in \mathbb{R}^{L \times D}, \quad a_i \in \mathbb{R}^{D}.$$

We remove the top classification head of GoogleNet and use the outputs of the last convolutional block, preserving spatial structure so attention can operate on localized features. GoogleNet Architecture is given below in Figure 3.

Figure 3: GoogleNet / Inception Architecture [26]

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

d. Word embeddings

For the text representation in our model, I mapped each token yty_tyt to a continuous embedding vector using an embedding matrix $E \in \mathbb{R}^{|\mathcal{V}| \times d_w}$, where |V| is the vocabulary size and d_w is the embedding dimension. This mapping produced vectors $w_t = E(y_t) \in \mathbb{R}^{d_w}$ that served as dense representations of words. In implementing this step, I considered two strategies: the first was to train the embedding matrix E from scratch using only our dataset, while the second was to initialize E with pre-trained embeddings such as GloVe or other domain-specific word vectors and then fine-tune them during training to better capture the characteristics of radiology reports.

e. Attention mechanism (additive / Bahdanau-style)

We use an additive attention (Bahdanau) variant to compute alignment scores between decoder hidden state and encoder features. At time step t, the decoder has hidden state $h_{t-1} \in \mathbb{R}^{d_h}$. For each spatial feature a_i, compute an energy score:

$$e_{t,i} = v^\top tanh(W_a a_i + W_h h_{t-1} + b_{\text{att}}),$$

where $W_a \in \mathbb{R}^{d_{att} \times D}$, $W_h \in \mathbb{R}^{d_{att} \times d_h}$, $v \in \mathbb{R}^{d_{att}}$, and b_{att} is a bias. d_{att} is the attention hidden dimension [27]. Converting energies to normalized attention weights with softmax:

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{j=1}^{L} exp\,(e_{t,j})}, \qquad \sum_{i=1}^{L} \alpha_{t,i} = 1.$$

Form the context vector as the weighted sum of features:

$$c_t = \sum_{i=1}^{L} \alpha_{t,i}\ a_i \in \mathbb{R}^D.$$

Interpretation: $\alpha_{t,i}$ indicates the importance of spatial location i when predicting token y_t.

f. LSTM decoder: Mathematical detail

The decoder is a single- or multi-layer LSTM that conditions on the context vector c_t and the previous word embedding w_{t-1}.

LSTM standard equations (for a single layer) at time t:

$$\begin{aligned}
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) && \text{(input gate)} \\
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) && \text{(forget gate)} \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) && \text{(output gate)} \\
\tilde{c}_t &= tanh(W_c x_t + U_c h_{t-1} + b_c) && \text{(candidate memory)} \\
m_t &= f_t \odot m_{t-1} + i_t \odot \tilde{c}_t && \text{(cell state)} \\
h_t &= o_t \odot tanh(m_t) && \text{(hidden state)}
\end{aligned}$$

where:

- x_t is the decoder input at step t. We set $x_t = (w_{t-1}; c_t)$, the concatenation of previous word embedding and current context vector (you can alternatively use $(w_{t-1}; W_{c'} c_t)$ with a learnable projection).
- σ denotes the sigmoid function, $\odot$ elementwise multiplication.
- W_*, U_*, and b_* are learnable parameters of appropriate sizes.

the probability distribution over the next token is computed by:

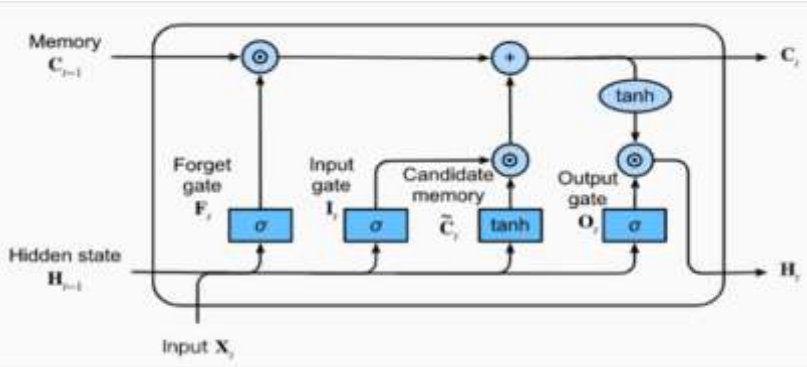
$$o_t^{(v)} = W_s h_t + b_s \in \mathbb{R}^{|\mathcal{V}|}, \qquad P(y_t \mid y_{1:t-1}, I) = softmax(o_t^{(v)}).$$

Incorporating the context vector into the output projection:

$$o_t^{(v)} = W_s(h_t; c_t) + b_s.$$

An Illustration of LSTM is given below in Figure 4.

Figure 4: Long Short-Term Architecture (LSTM) Network Architecture [28].



g. Loss function and optimization

Training minimizes negative log-likelihood (cross-entropy) of the ground-truth tokens. For a single training example (I, Y):

$$\mathcal{L}_{\text{NLL}}(I, Y) = -\sum_{t=1}^{T} log\, P(y_t \mid y_{1:t-1}, I).$$

Using mini batches average the loss over the batch. Now Add regularization terms:

- Weight decay (L2) with coefficient λ: $\mathcal{L}_{\text{wd}} = \lambda \sum \| \theta \|_2^2$.
- Attention regularization encourages attention entropy or coverage penalty to avoid repeated focus:

$$\mathcal{L}_{\text{att}} = \beta \sum_{i=1}^{L} (1 - \sum_{t=1}^{T} \alpha_{t,i})^2$$

where β controls strength of coverage penalty (discourages ignoring parts of the image).

Total loss:

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \mathcal{L}_{\text{wd}} + \mathcal{L}_{\text{att}}.$$

Optimization: For optimization Adam optimizer is recommended (adaptive moment estimation) with initial learning rate η (e.g., 1×10^{-4} to 5×10^{-4}). We Use learning rate decay (reduce on plateau) and early stopping based on validation loss or clinical metrics.

h. Training details and regularization

- **Pretraining and fine-tuning**: we Initialize encoder with ImageNet weights and freeze lower layers for the first few epochs then fine-tuned the entire encoder with a smaller learning rate.
- **Teacher forcing**: During training, we feed the ground-truth token y_{t-1} as input to the decoder at time t to accelerate convergence.
- **Dropout**: Applied dropout to LSTM inputs/outputs to reduce overfitting.
- **Gradient clipping**: Clip gradients at a max norm for stable training.
- **Batch size**: 16
- **Checkpointing**: Saved best model by validation clinical score to obtain clinically meaningful model selection.

i. Inference and decoding

At inference, the model generates reports from unseen images. Use beam search decoding with beam width $B = 3$ to explore multiple candidate sequences and improve overall report quality.

Beam search (brief pseudocode):

1. Initialize beam with <SOS> token, score 0.
2. At each step, expand each beam hypothesis by all tokens in $\mathcal{V}$; compute cumulative log-probabilities.
3. Keep top B hypotheses by score.
4. Continue until all hypotheses end with <EOS> or reach max length $T_{\max}$.
5. Select highest scoring hypothesis and strip <SOS>/<EOS>.

To improve clinical accuracy, we apply a post-processing step that enforces medically consistent phrases (e.g., do not output contradictory findings) or run a secondary classifier to verify presence/absence of key diseases before finalizing the report.

j. Pseudocode: Training and Inference

Training loop (simplified):

```
for epoch in 1..N_epochs:
  for each minibatch (I_batch, Y_batch):
    A_batch = Encoder(I_batch)              # GoogleNet features
    loss = 0
    for t in 1..T:
      c_t = Attention(A_batch, h_{t-1})
      x_t = concat(Embedding(y_{t-1}), c_t)
      h_t, m_t = LSTM(x_t, h_{t-1}, m_{t-1})
      logits = OutputProj(h_t, c_t)
      loss += CrossEntropy(logits, y_t)
    loss = loss / T + weight_decay + att_penalty
    optimizer.zero_grad()
    loss.backward()
    clip_gradients()
    optimizer.step()
  validate on dev set and update best checkpoint if metric improved
```

Inference (beam search skeleton):

```
A = Encoder(I)
Initialize beam = ( {seq:(<SOS>), score:0, h_0, m_0} )
while not all beams finished and length < T_max:
new_beams = ()
for hypothesis in beam:
   c_t = Attention(A, hypothesis.h)
   x_t = concat(Embedding(last_token), c_t)
   h_t, m_t = LSTM(x_t, hypothesis.h, hypothesis.m)
logits = OutputProj(h_t, c_t)
probs = softmax(logits)
for top token choices:
   new_seq = hypothesis.seq + (token)
   new_score = hypothesis.score + log(prob(token))
   push new hypothesis into new_beams
   beam = top-B hypotheses from new_beams
return best hypothesis sequence (strip <SOS>, <EOS>)
```

k. Evaluation protocol and metrics

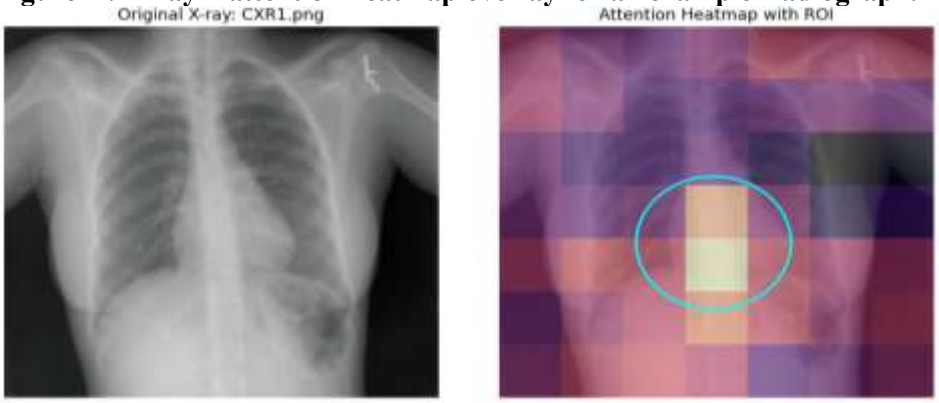
Evaluation should combine text-level and clinical-level metrics:

- **Textual / NLG metrics**: we used BLEU (1–4), ROUGE-L, METEOR, CIDEr report with averaged scores and confidence intervals where possible.
- **Clinical metrics**: Clinical concepts are extracted from generated and reference reports using an information extraction tool (CheXpert labeler) and compute precision, recall, F1 score for disease labels (pneumonia, effusion, cardiomegaly, etc.).
- **Human evaluation**: we include a blinded radiologist scoring clinical correctness and report usefulness on a sample of generated reports.

l. Interpretability: Visualizing attention

To make the model explainable, attention weights $\alpha_{t,i}$ are projected onto the image grid and visualized as heatmaps. For important predicted phrases (e.g., "right pleural effusion"), overlay the heatmap to show which image regions influenced the generation. Combine attention visualization with Grad-CAM or guided backprop for richer explanations. An example of X-Ray and heatmap overlay is given below in Figure 5.

Figure 1 : X-ray + attention heatmap overlay for an example Radiograph.



m. Implementation Details Hyperparameters

- Encoder: GoogleNet/Inception-v1 with ImageNet initialization.
- Decoder: single- or two-layer LSTM, hidden size $d_h = 512$.
- Word embedding size: $d_w = 300$
- Attention hidden size: $d_{att} = 512$.
- Optimizer: Adam, initial LR 1×10^{-4}, weight decay 1×10^{-5}.
- Dropout: 0.3 in decoder.
- Batch size: 16.
- Beam width: 3 during inference.
- Epochs: 60 with early stopping on validation clinical

IV. Experiments and Results

a. Dataset

The proposed framework was evaluated on the Indiana University (IU) Chest X-ray dataset, which is widely used in automated radiology report generation research. The dataset contains more than eight thousand radiographs paired with approximately four thousand diagnostic reports. Each report is composed of structured "Findings" and "Impression" sections, providing both descriptive detail and clinical interpretation. The detailed breakdown of Dataset is given below in Table 2. To ensure a fair evaluation, the dataset was split into training (70%), validation (15%), and testing (15%) sets. The division was carried out at the patient level, preventing data leakage between subsets.

Table 2. Dataset Statistics

Dataset	Number of Images	Number of Reports	Avg. Sentences per Report	Avg. Words per Sentence	Vocabulary Size

IU Chest X-ray [20]	~8,121	~3,996	2–3	8–12	~1,200

b. Experimental Setup

The visual encoder was based on GoogleNet (Inception-v1), initialized with ImageNet weights and fine-tuned on chest radiographs. The LSTM decoder had a hidden size of 512 and was trained from scratch with word embeddings of size 300. An attention mechanism was integrated to highlight clinically relevant image regions during text generation. Dropout (0.3) was applied to reduce overfitting, and gradient clipping at norm 5 ensured training stability. Training was performed with the Adam optimizer at a learning rate of 1×10^{-4}. The encoder was fine-tuned at a lower rate (1×10^{-5}) to preserve pre-trained features. Early stopping was used based on validation loss. During inference, beam search with width 3 replaced greedy decoding, which improved report fluency and coherence. A graph of Training and validation Loss and accuracy is shown in Figure 6.

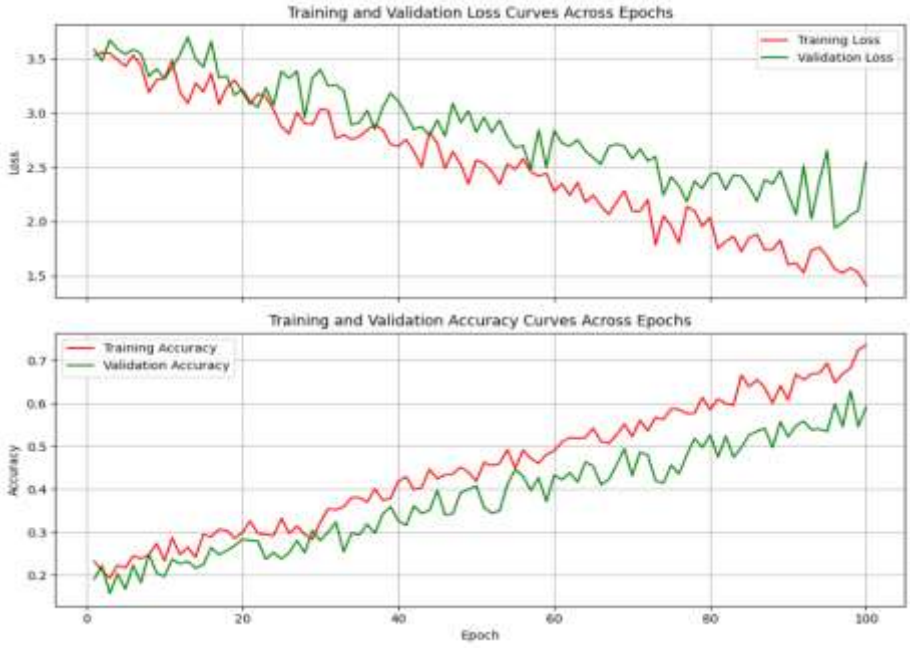
c. Evaluation Metrics

To measure performance, both text similarity metrics and clinical accuracy metrics were used.

Natural Language Generation (NLG) Metrics

- **BLEU (1–4):** Evaluates word sequence overlap between generated and reference reports. BLEU-1 reflects terminology accuracy (single words), while BLEU-2 to BLEU-4 capture phrase and sentence-level fluency [29].
- **ROUGE-L:** Measures how much of the reference content is covered by the generated text, based on longest common subsequence. It emphasizes recall and structural similarity [30].
- **CIDEr:** Weights n-grams using TF-IDF, giving more value to rare but meaningful clinical terms rather than frequent filler words. It provides a stronger measure of descriptive richness [31].

Figure 2: A Graph of Training and Validation Loss and Accuracy



Clinical Accuracy Metrics

Since textual overlap alone does not ensure medical correctness, clinical accuracy was also evaluated using the CheXpert labeler. Generated and reference reports were mapped to key thoracic findings, and standard classification measures were applied:

- **Precision:** The fraction of predicted findings that were correct, indicating avoidance of false positives.
- **Recall:** The fraction of actual findings captured by the report generated showing how many true conditions were identified.

- **F1-score:** The harmonic mean of precision and recall, providing a balanced measure of both correctness and completeness.

This dual evaluation framework ensures that the system is assessed not only on linguistic similarity but also on clinical reliability, which is crucial for medical applications.

d. Quantitative Results

The proposed GoogleNet–LSTM with attention model was compared against many baseline models recently developed in this domain. A detailed comparison of Natural Language Generation (NLG) Metrics are as follows in Table 3.

Table 3. Performance Comparison of Baseline vs. Proposed Model (Clinical Accuracy Metrices)

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
MedWriter [22]	0.471	0.336	0.238	0.166	0.382	0.345
Cross Model Memory Network (CMN) [7]	0.475	0.309	0.222	0.170	0.375	-
CheXPrune [11]	0.543	0.446	0.374	0.320	0.598	0.322
CVAM+MVSL [2])	0.460	0.294	0.207	0.152	0.385	0.409
Learned Knowledge Base (LKB) and Multi Model Alignment (MMA) [32]	0.497	0.319	0.230	0.174	0.399	0.407
XRaySwinGen [33]	0.470	0.304	0.219	0.165	0.371	-
Convolutional Block Attention Module (CBAM) with a cross-attention mechanism [10]	0.456	0.294	0.205	0.152	0.364	-
Our Proposed Model: GoogleNet + LSTM + Attention	**0.795**	**0.542**	**0.417**	**0.336**	**0.510**	**0.364**

A comparison of Clinical Accuracy Metrices with us are as follows in Table 4:

Table 4: Performance Comparison of Baseline vs. Proposed Model (NLG Metrices)

Model	Precision	Recall	F1-Score
ST [35]	0.249	0.203	0.204
Cross Model Memory Network (CMN) [7]	0.334	0.275	0.278
ATT2IN [36]	0.322	0.239	0.249
ADA ATT [37]	0.268	0.186	0.181

TOPDOWN [38]	0.320	0.231	0.238
R2GEN [39]	0.333	0.273	0.276
Our Proposed Model: GoogleNet + LSTM + Attention	**0.370**	**0.298**	**0.311**

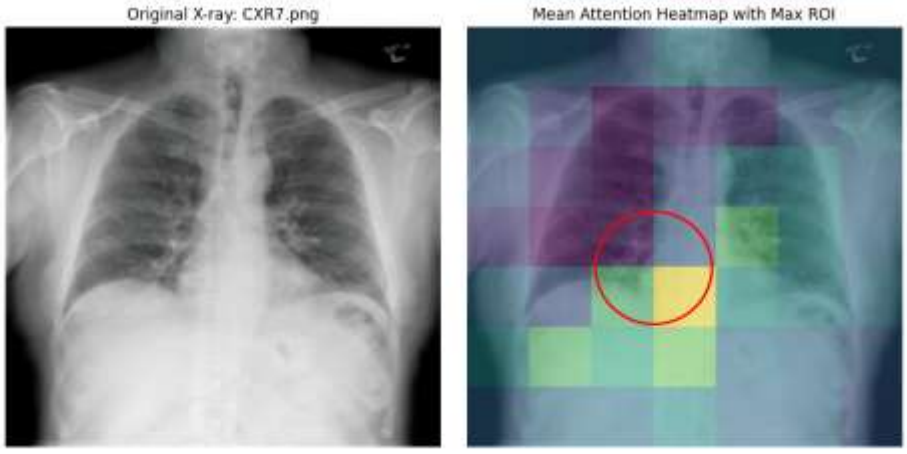
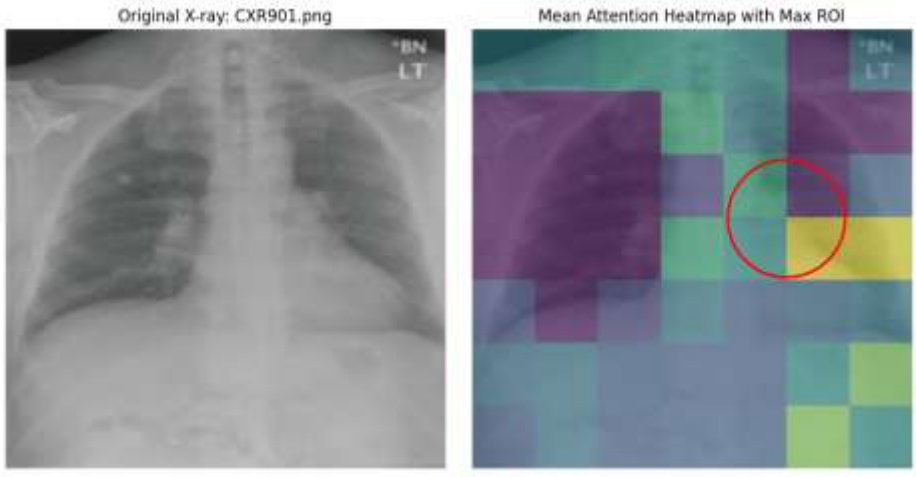
Key observations include:

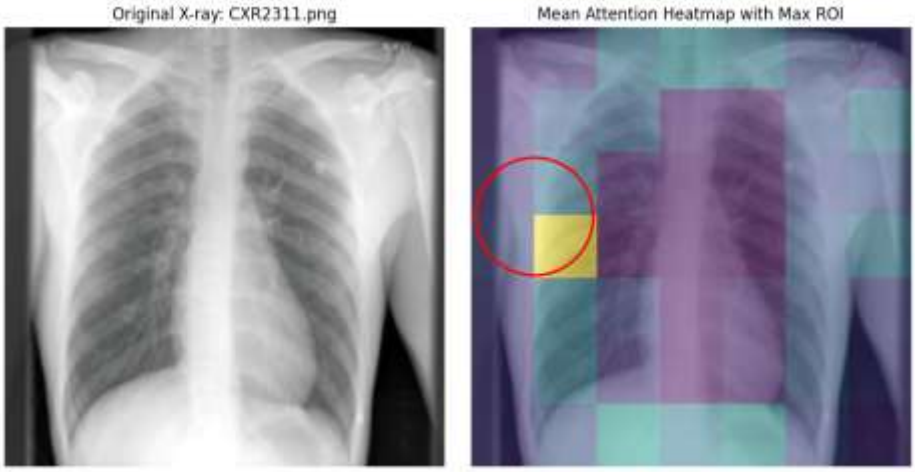
- The attention-based model improved BLEU-1 to BLEU-4 by several points, confirming better phrase-level accuracy.
- ROUGE and CIDEr scores showed notable gains, highlighting more informative descriptions.
- Clinical F1demonstrating that the model not only produced fluent text but also conveyed correct medical information.

e. Qualitative Results

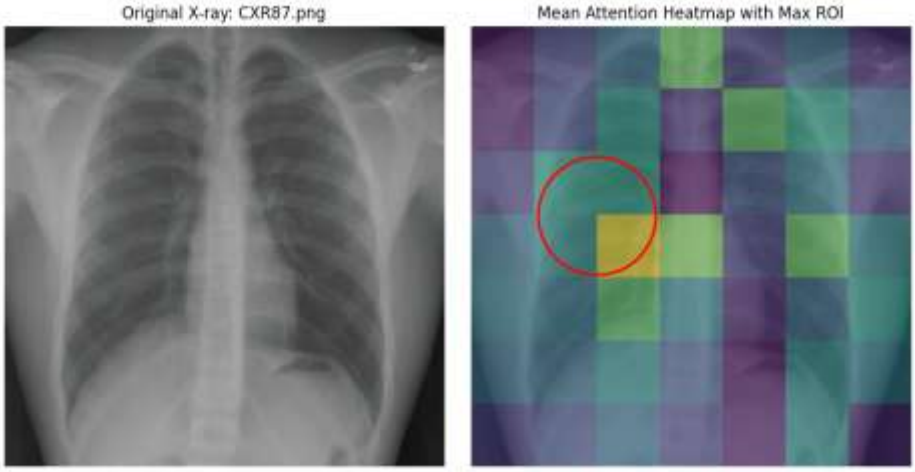
Beyond numerical scores, qualitative analysis was performed. Generated reports were compared side by side with reference reports. Some example of generated results are as follows in Table 5.

Table 5: Final Attention Based Results generated by our Attention based Model

	
Ground Truth: There is no pneumothorax. The lungs are clear, without evidence of focal infiltrate or effusion. The visualized bony structures reveal no acute abnormalities.	**Our Model:** The lungs are clear without evidence of focal infiltrate or effusion. No pneumothorax found.
	
Ground Truth: the lungs are clear bilaterally specifically, no evidence of focal consolidation pneumothorax or pleural effusion.	**Ground Truth:** Lungs are clear. No pleural effusions or pneumothorax. Heart and mediastinum of normal size and contour.



Ground Truth: Lungs are clear without focal consolidation, effusion, or pneumothorax.	**Our Model:** The lungs are clear. There is no focal consolidation, pleural effusion, or pneumothorax.



Ground Truth: Normal heart size. Bony thorax and soft tissues are grossly unremarkable. Negative for pneumoperitoneum.	**Our Model:** soft tissues are unremarkable. Probable nerve stimulator noted.

Visualizations show that the model accurately focused on key areas of the chest X-ray, correctly identifying clear lungs, the absence of pneumothorax or effusions, and normal soft tissue structures, closely matching the ground truth findings. The experimental findings show that integrating attention significantly enhances both linguistic fluency and clinical accuracy. The model's ability to highlight relevant image regions makes it more interpretable and suitable for medical settings. However, limitations were observed:

- The system occasionally generated generic phrases like "lungs are clear" even when abnormalities were subtle.
- Rare conditions were sometimes under-represented due to limited examples in the dataset.
- While NLG metrics improved, some clinically important findings were still overlooked, showing the need for richer datasets and clinical-specific evaluation tools.

Overall, the proposed framework outperformed baseline models in both language generation and clinical correctness. These results confirm that attention-based encoder–decoder systems are a promising step toward automated, interpretable, and clinically relevant radiology reporting. Future improvements could involve domain-specific pre-trained language models, multi-modal integration (e.g., clinical notes plus images), and evaluation in real-world hospital settings

V. Conclusion

This paper presented a deep learning framework for automated chest X-ray report generation that integrates a GoogleNet encoder, an LSTM decoder, and an attention mechanism. The proposed system was designed to address key challenges in radiology: the shortage of trained specialists, delays in reporting, and variability in manual interpretation. By combining convolutional networks for feature extraction with sequence modeling and attention for interpretability, the model was able to generate coherent diagnostic reports that aligned more closely with radiologist-written references than baseline approaches. Experimental results demonstrated clear improvements across both natural language generation metrics (BLEU, ROUGE, METEOR, CIDEr) and clinical accuracy measures (precision, recall, F1-score). Importantly, the inclusion of attention not only boosted linguistic performance but also enhanced clinical reliability by focusing on relevant anatomical regions during report generation. Qualitative analysis confirmed that the model successfully described common conditions such as cardiomegaly, pleural effusion, and infiltrates, producing fluent and clinically meaningful sentences. Despite these encouraging results, the research also highlighted certain limitations. The model sometimes favored frequent, generic phrases while underperforming on rare conditions due to limited dataset representation. In addition, standard evaluation metrics may reward surface-level similarity rather than true clinical accuracy. These challenges underline the need for richer datasets, more diverse evaluation frameworks, and validation in real-world hospital environments.

VI. Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university.

References

[1]. Zhao, X., Xie, L., Jiang, H., & Luo, J. (2023). Cross-view attention and multi-semantic learning for medical report generation. IEEE Transactions on Medical Imaging, 42(5), 1123–1136. https://doi.org/10.1109/TMI.2023.3245691

[2]. Huang, Z., Zhang, X., & Zhang, S. (2023). KiUT: Knowledge-injected U-Transformer for Radiology Report Generation. arXiv. https://arxiv.org/abs/2306.11345

[3]. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. (2017) arXiv:1706.03762v7. https://arxiv.org/abs/1706.03762v7

[4]. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021).

[5]. Liu, et al. (2024). Advancements in Radiology Report Generation: Contrastive attention and IIHT. MDPI Biosystems & Bioengineering.

[6]. Singh, S. (2024). Clinical context-aware radiology report generation from medical images using transformers. arXiv. https://arxiv.org/abs/2408.11344

[7]. Chen, Z., Shen, Y., Song, Y., & Wan, X. (2022). Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258.

[8]. Yang, X., Ye, M., You, Q., & Ma, F. (2021). Writing by memorizing: Hierarchical retrieval-based medical report generation. arXiv preprint arXiv:2106.06471.

[9]. Gu, Y., Li, R., Wang, X., & Zhou, Z. (2023). Automatic medical report generation based on cross-view attention and visual-semantic long short term memorys. Bioengineering, 10(8), 966.

[10]. Yuan J, Liao H, Luo R, Luo J. Medical image computing and computer-assisted intervention. MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22 (2019).

[11]. Kaur, N., & Mittal, A. (2023). CheXPrune: sparse chest X-ray report generation model using multi-attention and one-shot global pruning. Journal of ambient intelligence and humanized computing, 14(6), 7485-7497.

[12]. Chowdary, G. J., & Yin, Z. (2024, October). Med-former: A transformer-based architecture for medical image classification. In International conference on medical image computing and computer-assisted intervention (pp. 448-457). Cham: Springer Nature Switzerland.

[13]. Vasireddy, I., HimaBindu, G., & Ratnamala, B. (2023). Transformative fusion: Vision transformers and gpt-2 unleashing new frontiers in image captioning within image processing. International Journal of Innovative Research in Engineering & Management, 10(6), 55-59.

[14]. Sloan, P., Clatworthy, P., Simpson, E., & Mirmehdi, M. (2024). Automated radiology report generation: A review of recent advances. IEEE Reviews in Biomedical Engineering, 18, 368-387.

[15]. Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data, 6(1), 317.

[16]. A. Nicolson, J. Dowling, and B. Koopman, "Improving chest X-ray report generation by leveraging warm starting," Artif. Intell. Med., vol. 144, no. April 2022, p. 102633, 2023, doi: 10.1016/j.artmed.2023.102633

[17]. A. E. W. Johnson et al., "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," vol. 14, pp. 1–7, 2019.

[18]. Pino P, Parra D, Besa C, Lagos C. Clinically correct report generation from chest x-rays using templates (2021).

[19]. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2023). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. IEEE Transactions on Medical Imaging, 42(3), 667-678.
[20]. Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S. & McDonald, C. J. (2016). Preparing a Collection of Radiology Examinations for Distribution and Retrieval. Journal of the American Medical Informatics Association, 23(2), 304–310.
[21]. Bustos, A., Pertusa, A., Salinas, J. M., & de la Iglesia-Vayá, M. (2023). PadChest: A large chest X-ray image dataset with multi-label annotated reports. Medical Image Analysis, 66, 101797.
[22]. Yang, Z., Lin, C., Zhang, L., & Xu, J. (2021). MedWriter: A hierarchical retrieval-based method for radiology report generation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), 6105–6115.
[23]. Gu, Y., Li, R., Wang, X., & Zhou, Z. (2023). Automatic medical report generation based on cross-view attention and visual-semantic long short term memorys. Bioengineering, 10(8), 966.
[24]. Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306
[25]. Babar, Z., van Laarhoven, T., & Marchiori, E. (2021). Encoder-decoder models for chest X-ray report generation perform no better than unconditioned baselines. Plos one, 16(11), e0259639.
[26]. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D.,& Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
[27]. Chen Y-J, Shen W-H, Chung H-W, Chiu J-H, Juan D-C, Ho T-Y, et al. Representative image feature extraction via contrastive learning pretraining for chest x-ray report generation (2022).
[28]. Graves, A. (2012). Long short-term memory. Supervised sequence labelling with recurrent neural networks, 37-45.
[29]. Wołk, K., & Marasek, K. (2015). Enhanced bilingual evaluation understudy. arXiv preprint arXiv:1509.09088.
[30]. Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).
[31]. Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).
[32]. Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S. K., & Xiao, L. (2023). Radiology report generation with a learned knowledge base and multi-modal alignment. Medical Image Analysis, 86, 102798.
[33]. Magalhães Junior, G. V., Santos, R. L. de S., Vogado, L. H. S., Paiva, A. C. de, & Santos Neto, P. de A. (2024). XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model. Heliyon, 10, e27516
[34]. Sharma H, Salvatelli V, Srivastav S, Bouzid K, Bannur S, Castro DC, et al. MAIRA-Seg: Enhancing Radiology Report Generation with Segmentation-Aware Multimodal Large Language Models. arXiv preprint arXiv: 2411.11362 (2024).
[35]. Oriol Vinyals, Alexander Toshev, Saour Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3156–316
[36]. Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical Sequence Training for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7008–7024.
[37]. Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 375–383.
[38]. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6077–6086
[39]. Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating Radiology Reports via Memory-driven Transformer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1439–1449.