

Article

Chest X-Ray Medical Report Generation Using a CNN—Transformer Model with Maximum Attention

Mei-Hua Hsieh ¹, Shih-Po Lin ² and Chen-Chiung Hsieh ^{2,*} 

¹ Department of Product Design, School of Arts and Design, Sanming University, Sanming 365000, China; 20220204@fjssmu.edu.cn

² Department of Computer Science and Engineering, Tatung University, Taipei City 10452, Taiwan; stephenline0210@gmail.com

* Correspondence: cchsieh@gm.ttu.edu.tw; Tel.: +88-62-2182-2928 (ext. 6571)

Abstract

Medical imaging, particularly chest X-rays, plays a vital role in radiological diagnosis. However, interpreting these images and generating detailed diagnostic reports is a time-consuming task for clinicians. To address this challenge, this study proposes an automated image captioning framework for chest X-ray images, aiming to reduce clinical workload and enhance diagnostic efficiency. The proposed approach employs convolutional neural networks (CNNs) for visual feature extraction and a modified Transformer architecture—referred to as the Medical Transformer—for structured report generation. Three CNN models, namely InceptionV3, ResNet152V2, and Inception–ResNetV2, were evaluated as feature extractors. The attention mechanisms, Bahdanau, Luong, and scaled dot product, were activated by ReLU or Tanh functions to identify the optimal configuration, i.e., the maximum attention is used. Experiments were conducted using the Indiana University Chest X-ray dataset, which contains 7466 images paired with corresponding diagnostic reports. The proposed approach employs image augmentation to accommodate input variability, utilizes Inception–ResNetV2 for feature extraction, and integrates the Medical Transformer with maximum attention mechanisms to achieve optimal performance in medical report generation. Evaluation metrics include BLEU (BLEU-1 to BLEU-4 scores of 0.720, 0.669, 0.648, and 0.600, respectively), METEOR (0.741), and BERTScore ($F_{\text{BERT}} = 0.787$), demonstrating superior performance compared to baseline models and the state of the art. These results validate the effectiveness of the proposed Medical Transformer framework in generating accurate and clinically relevant medical image captions.

Keywords: deep learning; image captioning; diagnostic reports; attention mechanism; Transformer



Academic Editors: Chi-hung Chuang and Chih-Lung Lin

Received: 3 September 2025

Revised: 18 October 2025

Accepted: 19 October 2025

Published: 21 October 2025

Citation: Hsieh, M.-H.; Lin, S.-P.; Hsieh, C.-C. Chest X-Ray Medical Report Generation Using a CNN—Transformer Model with Maximum Attention. *Electronics* **2025**, *14*, 4123. <https://doi.org/10.3390/electronics14204123>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the context of an increasingly strained healthcare system—exacerbated by crises such as the COVID-19 pandemic—the need for efficient diagnostic support tools has become more urgent than ever. Healthcare professionals, particularly physicians and nurses, face elevated workloads and constant exposure to infection risks. The surge in patient numbers often results in understaffed medical facilities, placing additional pressure on clinicians who must perform a wide range of tasks, including patient diagnosis, ward rounds, image interpretation, report writing, and medication prescription.

Among these responsibilities, interpreting chest X-ray images and composing diagnostic reports are particularly time-consuming and cognitively demanding. Accurate analysis

requires meticulous attention to detail to avoid overlooking pathological signs, while report writing demands precision and adherence to standardized formats to ensure clarity and consistency across institutions. Given that patients may receive care from multiple providers, the quality and readability of diagnostic reports are critical for continuity of care.

Automating the generation of medical reports from X-ray images offers a promising solution to alleviate this burden. Unlike face-to-face consultations, medical imaging provides a structured foundation for automated interpretation. Chest X-rays are widely used to visualize internal anatomical structures, including the lungs, bronchi, heart, blood vessels, and thoracic cavity. Figure 1 illustrates a typical chest X-ray image used in clinical practice.

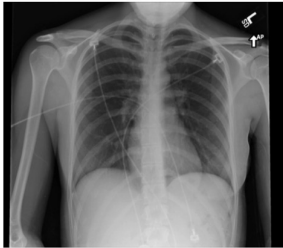
	Comparison: None Indication: Chest pain, feels out of it. Findings: The Cardio mediastinal silhouette and pulmonary vasculature are within normal limits in size. The lungs are clear of focal airspace disease, pneumothorax, or pleural effusion. There are no acute bony findings. Impression: No acute cardiopulmonary findings.
	Impression: COPD. No acute pulmonary disease. Findings: The lungs are clear. There is hyperinflation of the lungs. There is no pleural effusion or pneumothorax. The heart and mediastinum are normal. The skeletal structures are normal. Labels: Hyperinflation; Chronic obstructive; Copd; Pulmonary disease
	The Trachea is midline. The heart is large unchanged from prior exam. Slightly widened mediastinum secondary to cardiomegaly and a tortuous aorta is accentuated by ap portable. There are low lung volumes causing bibasilar atelectasis and bronchovascular crowding. The lungs do not demonstrate focal infiltrate or effusion. There is no pneumothorax. The visualized bony structures reveal no acute abnormalities.

Figure 1. Sample chest X-ray images from the open-access Indiana University Chest X-ray Collection. Source: <https://openi.nlm.nih.gov/> [1].

Recent advances in artificial intelligence, particularly in deep learning, have enabled significant progress in medical image analysis. Neural network architectures tailored to specific tasks—such as object classification and natural language processing—have been successfully integrated into image captioning frameworks. These developments pave the way for automated systems capable of generating clinically relevant and linguistically coherent diagnostic reports from radiological images.

While image captioning offers richer descriptive capabilities than traditional image classification, it often comes at the cost of reduced accuracy and increased computational time. This study focuses on the generation of diagnostic reports from chest X-ray images [2], tasks significantly more complex than captioning natural scene photographs [3]. Several challenges make this domain particularly demanding:

- **Grayscale Imaging:** Chest X-rays are monochromatic, limiting the visual cues available for machine interpretation.
- **Extended Text Output:** Diagnostic reports are typically longer than standard image captions, requiring a broader vocabulary and more sophisticated sentence construction.

- Structured Format Requirements: Medical reports must conform to strict formatting standards to ensure clarity and consistency across clinical settings.
- Limited and Paired Data: Compared to natural image datasets, chest X-ray datasets are smaller and must be paired with corresponding expert-written reports, making data collection and annotation labor-intensive.

To address these challenges, this research explores various neural network architectures for feature extraction and integrates multiple attention mechanisms in the Transformer. Through image augmentation, systematic parameter tuning, and comparative evaluation, the optimal model configuration is selected. Meanwhile, common reference-based metrics BLEU (Bilingual Evaluation Understudy) [4], METEOR [5], and BERTScore [6] are used for assessing N-gram overlap, word alignment, and semantic similarity, respectively. METEOR complements BLEU by assessing both the accuracy and fluency of the generated text, considering word order and semantic alignment. BERTScore is an embedding-based metric that uses contextual embedding from pre-trained models like BERT to measure semantic similarity between generated and reference texts. The proposed approach achieves a BLEU-1 score exceeding 0.6 and a METEOR score above 0.6, outperforming some state-of-the-art [7], which reports a BLEU-1 score of 0.498.

In summary, our main contributions are:

- A new architecture is proposed for medical report generation, based on a combination of Inception–ResNet152V2 for feature extraction and the Transformer accepting multimodal data.
- The attention part in Transformer is replaced with the proposed maximum attention mechanism. This modified Transformer is trained with multimodal data consisting of extracted image features and the corresponding medical text report.
- Image augmentation is employed to enhance data diversity and improve model generalization. Each augmented image is paired with its original medical report, helping to mitigate overfitting and achieve higher evaluation metrics in report generation.

This paper is organized into five sections. Section 1 introduces research motivation, objectives, and key challenges. Section 2 presents a literature review, highlighting the intersection of computer vision and natural language processing and discussing relevant neural network models. Section 3 details the research methodology, including dataset descriptions and the proposed model architecture. Section 4 reports experimental results and performance metrics. Finally, Section 5 concludes the study and outlines potential directions for future research.

2. Related Works

To address the concurrent processing of chest X-ray images and associated medical reports, this study employs convolutional neural network (CNN) architectures for visual feature extraction and natural language models for automated text processing. Representative models commonly used in image classification and text generation are reviewed in the following subsections. Most related state of the art for medical report generation are also surveyed.

2.1. Image Classification

Image classification is one of the most fundamental tasks in computer vision. It typically involves using Convolutional Neural Networks (CNNs) [8] to extract visual features from input images, which are then passed through Fully Connected (FC) layers for classification. Early approaches focused on binary classification—determining the presence or absence of a specific object. Over time, these methods evolved into multi-class classification frameworks capable of distinguishing among multiple object categories.

- **Inception Architectures:** The Inception module, introduced within GoogLeNet [9] by the Google team in 2014, marked a significant advancement in deep learning architecture. InceptionV1 [9] laid the foundation, followed by InceptionV2 and InceptionV3 [10] in 2015, each introducing architectural refinements like asymmetric convolution, auxiliary classifiers, and feature map reduction.
- **Residual Networks:** ResNetV1 [11] was introduced by He et al. in 2015 to address the vanishing gradient problem in deep networks. The key innovation was the use of residual connections, which allow gradients to flow more effectively through the network, bypassing certain layers. However, ResNetV1 encountered optimization difficulties in extremely deep architectures. ResNetV2 [12] was proposed in 2016 as an improved version, redesigning the residual block structure to achieve lower error rates while maintaining depth. Residual connections help preserve gradient flow during backpropagation, preventing the gradient from diminishing to near-zero values in deeper layers.
- **Inception–ResNet Architectures:** In 2016, Google introduced Inception-ResNetV2 [13], which combines the strengths of Inception modules with residual connections. This hybrid architecture integrates the modular design of Inception with the gradient-preserving capabilities of ResNet, addressing both gradient vanishing and moment instability. Inception–ResNetV2 incorporates several advanced techniques, including residual connections, asymmetric convolutions, parallel computation, and dimensionality control.

Despite these innovations, training instability can arise when the number of filters exceeds 1000. Even with reduced learning rates and additional Batch Normalization layers, the network may produce suboptimal results during early training stages. To address this, Szegedy et al. [13] introduced Activation Scaling before residual addition, stabilizing the network and improving convergence.

2.2. Natural Language Attention Mechanisms

Natural Language Processing (NLP) leverages machine learning techniques to enable computers to interpret and generate human language. Core applications include machine translation, image captioning, and speech recognition. For a model to effectively learn and process human language, it must first comprehend the semantic meaning of individual words and analyze their linguistic properties.

The Sequence-to-Sequence (Seq2Seq) model [14], first introduced by the Google team in 2014, was designed to map input sequences to output sequences and was initially applied to machine translation. The Seq2Seq architecture comprises two main components: an encoder and a decoder, both typically implemented using recurrent neural networks (RNNs), gated recurrent units (GRUs), or long short-term memory (LSTM) networks. The encoder processes the input sequence and encodes it into a fixed-length context vector, which encapsulates the semantic information of the entire sequence. This context vector is then passed to the decoder, which generates the output sequence one token at a time. While the encoder processes the entire input sequence before producing the context vector, the decoder generates each output token based on the previously generated token, enabling sequential prediction.

While the Seq2Seq architecture effectively addresses the limitations of traditional alignment-based machine translation, it still suffers from two notable drawbacks. First, when processing long input sequences, recurrent neural networks (even advanced variants such as GRUs and LSTMs) [15] tend to lose important contextual information over time. This degradation occurs because the fixed-length context vector cannot adequately capture all relevant details from lengthy inputs. Second, the standard Seq2Seq model treats all input

tokens with equal importance, lacking the ability to selectively focus on more informative or contextually relevant words.

To mitigate these issues, Bahdanau et al. [16] introduced the attention mechanism, which enhances the decoder's ability to dynamically focus on different parts of the input sequence during generation. Instead of relying solely on a single context vector, the attention mechanism computes a weighted sum of the encoder's hidden states at each decoding step. These weights are determined by the relevance of each input token to the current decoding state, allowing the model to generate a context vector that varies over time. This dynamic context vector is then used as input to the decoder, enabling more accurate and context-sensitive output generation.

Luong et al. [17] further advanced attention mechanisms by introducing two variants: the global attention model and the local attention model. The global attention mechanism closely resembles the approach proposed by Bahdanau et al., with key differences in implementation. Specifically, Luong's model employs matrix multiplication to compute alignment scores, whereas Bahdanau's method uses additive operations. Additionally, Bahdanau's attention derives context vectors based on the decoder's previous hidden state, while Luong's model utilizes the current hidden state during decoding.

The local attention mechanism, on the other hand, restricts focus to a subset of the encoder's hidden states at each decoding step. This design aims to reduce computational overhead by avoiding full-sequence attention. However, empirical results indicated that when the input sequence is relatively short, the computational savings are minimal. Consequently, the global attention model has become more widely adopted in practice due to its simplicity and effectiveness.

The integration of attention mechanisms into sequence-to-sequence models significantly advanced the capabilities of natural language processing (NLP). However, early implementations still relied on convolutional neural networks (CNNs) or recurrent neural networks (RNNs), which posed efficiency challenges. In particular, RNNs process input sequences sequentially, scanning from beginning to end, which limits parallelization and slows down computation.

To overcome these limitations, the Google team introduced the Transformer architecture in 2017 [18]. Like the Seq2Seq model, the Transformer consists of an encoder–decoder structure. However, it eliminates the need for RNNs, LSTMs, or GRUs by relying entirely on self-attention mechanisms within both the encoder and decoder. The core of this mechanism is the scaled dot-product attention, which computes attention scores based on the similarity between query (Q), key (K), and value (V) vectors—concepts originally formalized in Luong's attention model [16].

One limitation of single-head attention is its inability to capture diverse contextual relationships simultaneously. To address this, the Transformer introduces Multi-Head Attention, which applies multiple parallel attention layers to the input. Each head independently transforms the input matrices (Q, K, V) and computes attention scores, allowing the model to focus on different aspects of the input sequence concurrently. The outputs from all heads are then concatenated and linearly transformed to produce the final representation. Unlike RNN-based models, the Transformer processes entire sequences in parallel, eliminating the dependency on previous outputs during training. This parallelism not only accelerates computation but also enhances model performance by enabling richer contextual modeling through multi-head attention. As a result, the Transformer has become the dominant architecture in modern NLP applications.

2.3. Medical Image Captioning

Medical imaging is essential for clinical diagnosis, yet interpreting and summarizing scans remains time-consuming and a bottleneck in clinical workflows. Yin et al. [19] identified two key challenges in generating medical image reports: difficulty in detecting all abnormalities—especially rare ones—and the complexity of multi-paragraph medical narratives compared to natural image captions. To address this, they proposed a Hierarchical Recurrent Neural Network (HRNN) comprising a Sentence RNN and a Word RNN. The Sentence RNN generates semantic topic vectors, which guide the Word RNN in producing coherent sentences. While HRNN improved diagnostic accuracy, its sequential processing and layered structure limit computational efficiency and parallelization, making it less scalable than newer architectures.

Beddiar et al. [20] highlighted the complexity of title generation for medical images, emphasizing that it poses greater challenges than captioning general images. Unlike natural images, medical image interpretation requires domain-specific knowledge, including familiarity with specialized medical terminology. Moreover, the generated descriptions must closely align with a clinician's diagnostic perspective to be clinically useful. Conventional image captioning methods often rely on identifying objects adjacent to visual symbols to construct narratives. While effective for general imagery, this strategy tends to produce inaccurate and low-quality descriptions when applied to medical contexts. To overcome these limitations, Beddiar et al. [20] proposed a novel framework tailored for medical image captioning. Their approach begins by classifying semantic features using a Multi-Label Classifier (MLC) [21]. VGG-16 [22] is employed to extract visual features from the medical images. These semantic vectors, along with image features and preliminary captions, are concatenated and fed into an LSTM network [23] for training. This method is applicable across various anatomical regions, including the brain, chest, and limbs. However, the concatenation of multiple feature vectors results in high-dimensional input sequences. This can lead the LSTM to prioritize frequently occurring words while discarding less common but potentially critical terms, thereby affecting the accuracy and richness of the generated descriptions.

Chen et al. [7] observed that multilayer recurrent neural networks (RNNs) fail to leverage the structural consistency commonly found across medical image reports. To address this limitation, they introduced the Memory-driven Transformer architecture [7], designed to capture and utilize the repetitive patterns and semantic structures inherent in medical reporting more effectively. The proposed model comprises the Encoder retaining the standard Transformer configuration, while the Decoder is enhanced with Relational Memory and Memory-driven Conditional Layer Normalization (MDCLN). However, the inclusion of Relational Memory introduces additional parameter matrices, potentially increasing the model's complexity and computational overhead compared to standard Transformer models.

Wijerathna et al. [24] proposed a hybrid architecture for chest radiograph caption generation, integrating CheXNet [25] with the Memory-driven Transformer [7]. Their framework leverages CheXNet for high-quality feature extraction from chest X-ray images, while the Memory-driven Transformer is employed to generate semantically rich and clinically relevant textual reports. This combination yielded significantly higher evaluation scores compared to other contemporary models. However, a key limitation lies in CheXNet's domain specificity. Since it was trained exclusively on thoracic diseases, its generalizability to non-thoracic medical imaging tasks—such as brain or abdominal scans—may be constrained, potentially diminishing its effectiveness in broader diagnostic applications.

Selivanov et al. [26] proposed a model for automatic clinical image captioning that integrates radiological scan analysis with structured patient records. It leverages two language models—Show-Attend-Tell and GPT-3—to generate detailed radiology summaries, including pathology descriptions, locations, and corresponding 2D heatmaps. Evaluated on IU X Ray, MIMIC-CXR, and MS-COCO datasets using standard language metrics, the model demonstrates strong performance in chest X-ray captioning.

Liu et al. [27] proposed a label-correlated contrastive learning method to enhance report quality. First, a refined similarity matrix is constructed by comparing multi-label classifications across reports. Next, image features and decoder-generated semantic embeddings are projected into a shared hidden space. Contrastive learning is then applied using these representations and the similarity matrix, assigning greater weight to “hard” negatives—samples sharing more labels with the target. Finally, this contrastive framework is integrated with an attention mechanism to generate diagnostic reports.

Ramadan and Akay [28] introduced the use of Convolutional Vision Transformers (CvT) for medical image captioning (MIC), combining convolutional local feature extraction with transformer-based global context modeling. Aimed at enhancing report quality and clinical relevance, this hybrid approach is among the first to apply CvT to MIC tasks. Evaluated on IU X-Ray and MIMIC-CXR datasets using NLP and clinical efficacy (CE) metrics, the model achieves notable improvements on MIMIC-CXR, with F1-score and Precision increasing by 7.9% and 8.3%, respectively—indicating stronger clinical feature representation.

Recent research in medical image captioning highlights the effectiveness of combining convolutional neural networks (CNNs) for visual feature extraction with Transformer-based architectures for text generation. Early RNN-based models like HRNN improved diagnostic accuracy but suffered from inefficiency due to sequential processing. Subsequent approaches introduced multi-label classification, memory-driven Transformers, and contrastive learning to enhance semantic alignment and report quality. Hybrid frameworks integrating models such as CheXNet, GPT-3, and CvT demonstrated strong performance on chest X-ray datasets, though challenges remain in generalizability and computational complexity. These findings support the use of CNN-Transformer hybrids for generating clinically relevant radiology reports.

3. Proposed Methodology

This study aims to alleviate the workload of radiologists by employing an automated image captioning approach to generate diagnostic reports from chest radiographs, one of the most used modalities in medical imaging [2]. For less experienced clinicians, accurately interpreting radiological images can be challenging, while for seasoned professionals, composing detailed reports is often time-consuming and labor-intensive. The diagnostic reports generated in this research are structured into three key sections: Indication, Findings, and Impression.

- Indication refers to the clinical context or presenting symptoms of the patient. In the generated reports, this section is inferred from visual features in the X-ray image and may include demographic information and symptomatic descriptions, such as “45-year-old male with chest pain.”
- Findings detail the model’s observations of anatomical structures and potential abnormalities, including phrases like “Lungs clear,” “Pericardial fat,” or “Pulmonary edema.”
- Impression synthesizes the information presented in the Findings and offers a concise summary or clinical recommendation, reflecting the diagnostic conclusion.

Figure 2a illustrates an example of a generated medical report, demonstrating the complexity and length of the textual output. Similarly to other radiology image captioning

tasks (as shown in Figure 2b), medical report generation demands significantly more detailed and domain-specific language, making it a more challenging task.

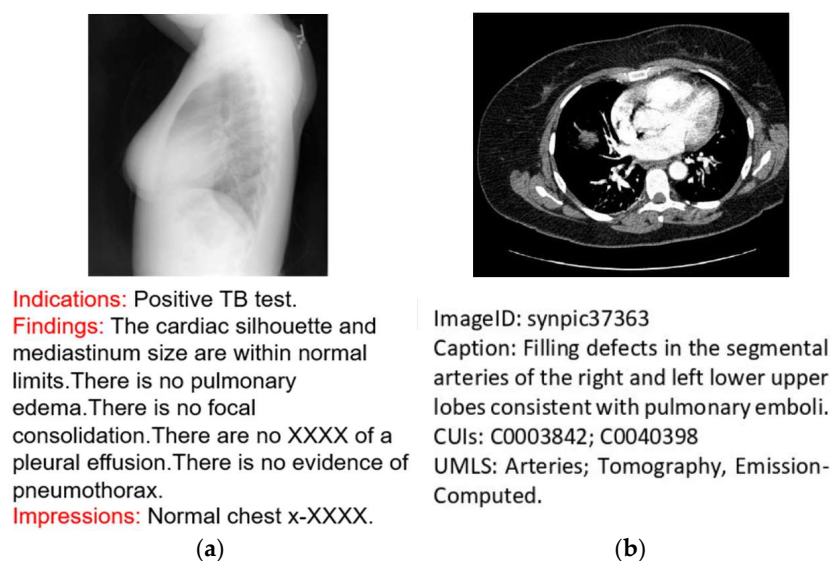


Figure 2. Medical reports are typically authored by radiologists with specialized expertise in interpreting diagnostic images. Shown here are example reports corresponding to (a) a chest X-ray image [1] and (b) a radiological image from the ImageCLEF dataset [29].

3.1. System Flowchart

This section describes the system's architecture, followed by the descriptions of the deployed neural network models. Finally, since the results of natural language processing cannot be evaluated using the accuracy rate as in object detection, the evaluation metrics of image caption generation will also be reviewed.

Figure 3 illustrates the system flowchart developed in this study. Upon execution, the program begins by loading the input medical X-ray images and simultaneously initiates the pre-processing stage. This pre-processing involves resizing the images and converting them into numerical feature vectors suitable for model input.

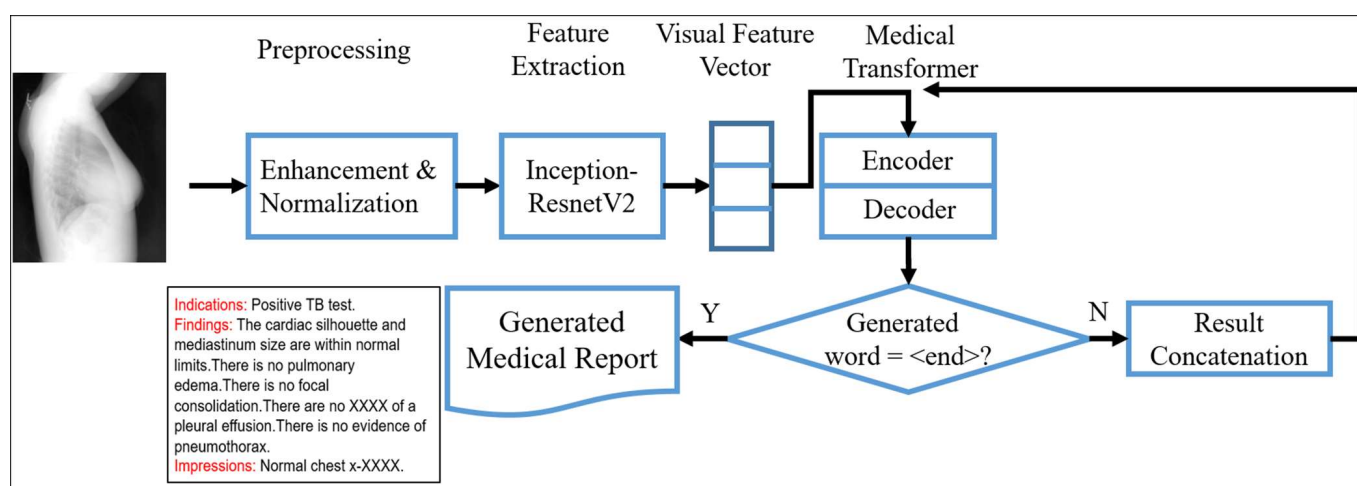


Figure 3. System flowchart.

Subsequently, the Inception–ResNetV2 architecture is employed to extract high-level image features. These features are then fed into a Transformer-based model to generate the diagnostic report. The generation process begins with a predefined token, ‘<start>’, which

prompts the Transformer to predict the subsequent word based on both the image features and the previously generated text.

The model continues to iteratively predict and append words to the output string until the '<end>' token is produced. Each newly predicted word is concatenated to the result string and reintroduced into the model alongside the image features to inform the next prediction. Once the '<end>' token is reached, the final generated report is compared against the ground truth using evaluation metrics, including BLEU-1 through BLEU-4 and METEOR. The program then concludes its execution.

3.2. Neural Network Models

This section outlines the two core neural network models trained in the context of medical image captioning: feature extraction and diagnostic report generation.

3.2.1. CNN Models for Feature Extraction

A wide range of convolutional neural network (CNN) architectures has been developed in recent years, demonstrating exceptional performance in tasks such as object detection and classification. In this study, three CNN models—InceptionV3, ResNet152V2, and Inception–ResNetV2—were selected for comparison based on their documented performance metrics, including accuracy, network depth, computational efficiency, memory consumption, and parameter count, as reported on the Keras official website [30].

These models were pre-trained on the ImageNet dataset, where the original top layer is a fully connected layer configured for classification across 1000 categories. However, the objective of this research is not image classification, but rather feature extraction for generating diagnostic reports using attention-based mechanisms. To adapt the models for this purpose, the top classification layer is removed, and the second-to-last layer is repurposed as the output layer. This modification enables the extraction of high-level image features, which serve as input to the subsequent caption generation framework.

3.2.2. Medical Report Generation-Proposed Medical Transformer

This study focuses on generating diagnostic reports by evaluating and adapting the widely used Transformer model, which incorporates distinct attention mechanisms, specifically Bahdanau (additive) attention [16] and Luong (multiplicative) attention [17]. To enhance the alignment between generated and reference diagnostic reports, a modified architecture—referred to as the Medical Transformer—is proposed.

The overall structure of the Medical Transformer is depicted in Figure 4. While the encoder component remains unchanged, modifications are introduced in the decoder. The attention mechanism employed in the Transformer is based on the Scaled Dot-Product Attention, defined in Equation (1). In this formulation, Q , K , and V denote the Query, Key, and Value matrices, respectively, which are learned during training. d_K is the dimension of the Key. The dot product of Q and K captures the relevance between input and output sequences, and the result is scaled by the square root of the Key dimension to prevent gradient vanishing. The scaled scores are then normalized using the Softmax function and multiplied by V , as shown in Equation (1), to produce the final attention output. The Medical Transformer extends the Scaled Dot-Product attention by integrating the additive attention mechanism introduced by Bahdanau et al., as formulated in Equations (2) and (3). Both equations utilize matrix addition between the Query Q and Key K representations to compute attention scores. The primary distinction between the two lies in the activation function applied during the computation.

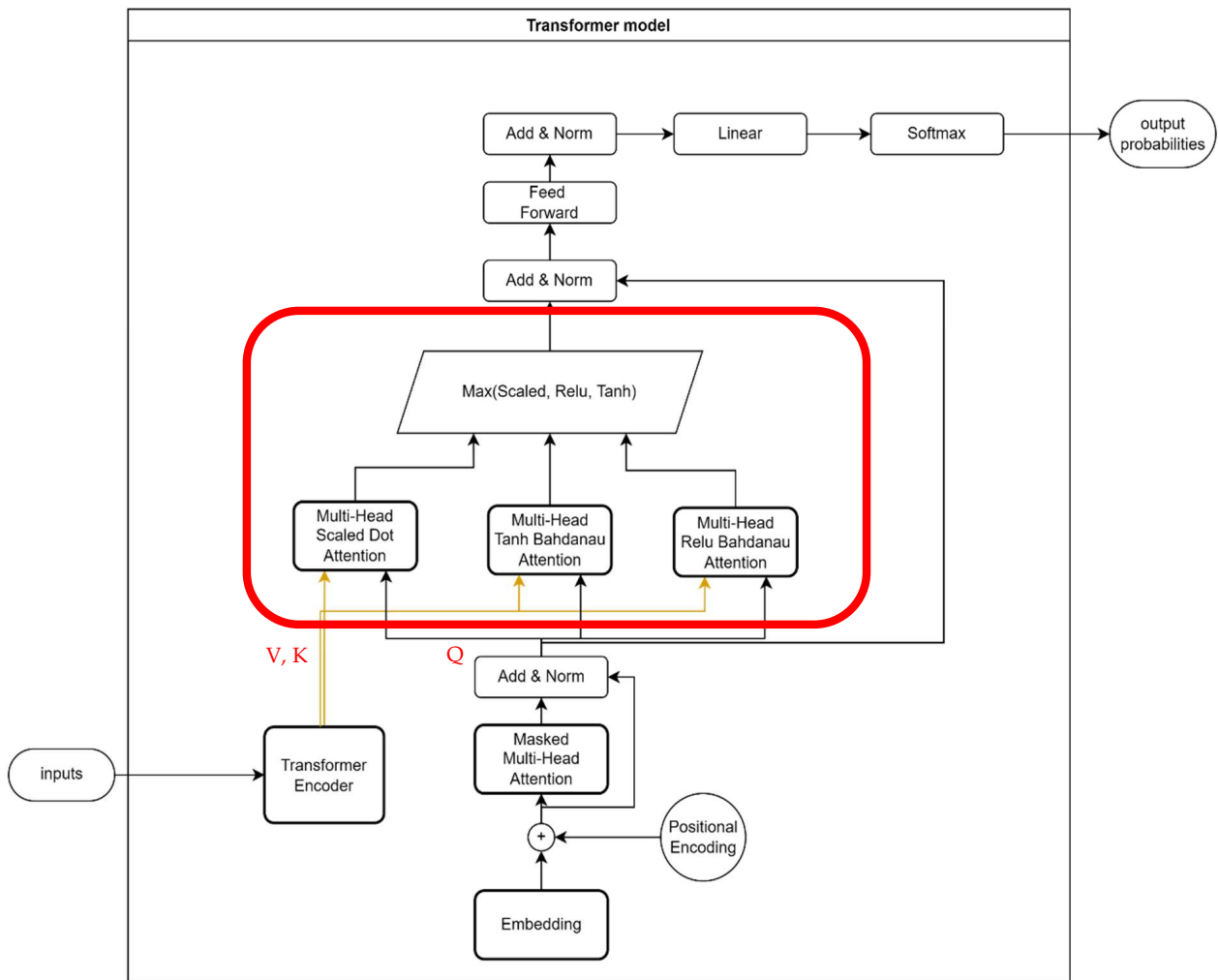


Figure 4. The architecture of the proposed Medical Transformer.

As indicated by the red box in Figure 4, Equation (4) presents the final attention mechanism adopted in this study, which finds the maximum Attention among Scaled Dot-Product attention, Tanh Bahdanau attention, and Relu Bahdanau attention. By evaluating the outputs of each mechanism and selecting the most effective result at each decoding step, the model dynamically enhances its alignment capabilities. This hybrid approach addresses the limitations of individual attention strategies and contributes to improved accuracy and robustness in diagnostic report generation. This is similar to the maximum pooling function in CNN, which could receive the most important attention of the candidate attentions. Other functions may not suggest better attention.

$$Scaled_Dot_Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

$$Tanh_Bahdanau_Attention(Q, K) = Tanh(Q + K) \quad (2)$$

$$Relu_Bahdanau_Attention(Q, K) = Relu(Q + K) \quad (3)$$

$$Maximum_Attention(Q, K, V) = Max((1), (2), (3)) \quad (4)$$

3.3. Evaluation Metrics

This section presents several widely adopted evaluation metrics for image captioning models. Given that the output of image captioning is natural language text, conventional

object detection metrics—such as accuracy, recall, precision, and F1-score—are insufficient for a comprehensive assessment. Instead, specialized linguistic and semantic metrics are required to evaluate the quality and relevance of generated captions.

BLEU (Bilingual Evaluation Understudy) [4] is a widely used metric for evaluating machine-generated text, particularly in machine translation and image captioning tasks. BLEU assesses the similarity between a candidate sentence produced by a model and one or more reference sentences (ground truth), with the final score ranging from 0 to 1. A higher score indicates greater overlap with the reference text. The metric is based on N-gram precision, where different levels of N-grams—such as Unigram (1-g), Bigram (2-g), Trigram (3-g), and 4-g—are used to evaluate both word-level accuracy and sentence fluency. BLEU scores are typically computed across multiple N-gram levels and may be averaged using weighted combinations to reflect overall performance, as shown in Equation (5).

$$BLEU = BP \times \exp\left(\frac{1}{n} \sum_{i=1}^N P_n\right) \quad (5)$$

$$BP = \begin{cases} 1, & \text{if MT output length} > \text{Ground truth output length} \\ \exp\left(1 - \frac{\text{MT output length}}{\text{Ground truth output length}}\right), & \text{otherwise} \end{cases} \quad (6)$$

$$P_n = \frac{\sum_{C \in \{\text{candidates}\}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{C' \in \{\text{candidates}\}} \sum_{n\text{-gram}' \in C'} \text{Count}(n\text{-gram}')} \quad (7)$$

$$\text{Count}_{\text{clip}} = \min(\text{Count}, \text{Max_Ref_Count}) \quad (8)$$

BP defined in Equation (6) is the brevity penalty applied when the candidate sentence generated by the model is shorter than the ground truth sentence, discouraging overly concise outputs. The modified precision P_n , shown in Equation (7), corresponds to the N-gram level n , where the *candidates* refer to the machine-generated translation. An N-gram is a contiguous sequence of n words, and the term *Count* represents the frequency of each N-gram in the candidate. $\text{Count}_{\text{clip}}$, defined in Equation (8), limits the count of each N-gram to its maximum occurrence in the ground truth, preventing overestimation due to repetition. While BLEU is computationally efficient and straightforward to implement, it primarily focuses on matching word frequencies and does not account for grammatical correctness, word order, or semantic equivalence (e.g., synonyms). Typically, BLEU employs up to 4-g precision, which limits its ability to evaluate longer and more complex sentences comprehensively. Nonetheless, it remains effective for assessing shorter, structurally simpler outputs.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) [5] was developed to address several limitations of the BLEU metric. While BLEU emphasizes precision and overlooks aspects such as word order, grammar, and synonymy, METEOR incorporates these linguistic features to provide a more comprehensive evaluation of machine-generated text. Unlike BLEU, METEOR calculates both precision and recall and combines them into a harmonic mean (F1-score), thereby balancing the trade-off between completeness and accuracy. Additionally, METEOR leverages WordNet to identify synonyms, allowing for more flexible matching between candidate and reference sentences.

The METEOR score is computed using Equation (9), with the F_{mean} defined in Equation (10). After aligning the candidate and reference sentences using unigram matches, precision (P) and recall (R) are calculated as shown in Equations (11) and (12), respectively. In these equations, m denotes the number of unigram matches, t is the total number of words in the candidate, and r is the total number of words in the reference. A penalty factor is then applied, as defined in Equation (13), to account for discrepancies in word order. The penalty is based on the number of contiguous matched segments $\#chunks$ and the

total number of unigram matched m . This adjustment penalizes fragmented alignments, encouraging coherent and well-ordered sentence generation.

$$METEOR = (1 - Penalty) \times F_{mean} \quad (9)$$

$$F_{mean} = \frac{10 \times P \times R}{R + 9 \times P} \quad (10)$$

$$P = \frac{m}{t} \quad (11)$$

$$R = \frac{m}{r} \quad (12)$$

$$Penalty = 0.5 \times \left(\frac{\#chunks}{\#unigram_matched} \right)^3 \quad (13)$$

The evaluation metrics BLEU and METEOR focus on word order and frequency, without fully capturing semantic accuracy. This limitation may result in misleading assessments, such as identifying a disease where none exists, or vice versa. We incorporate a more semantically aware evaluation method BERTScore [6], to provide a more comprehensive assessment of diagnostic quality. The complete score matches each token in x to a token in $\hat{x}$ to compute recall, and each token in $\hat{x}$ to a token in x to compute precision. We use greedy matching to maximize the matching similarity score, where each token is matched to the most similar token in the other sentence. We combine precision and recall to compute an F1 measure. For a reference x and candidate $\hat{x}$, the recall, precision, and F1 scores are:

$$R_{BERT} = \frac{1}{|x|} \times \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \quad (14)$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \times \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \hat{x}_j^T x_i \quad (15)$$

$$F_{BERT} = 2 \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \quad (16)$$

4. Experimental Results

The hardware training environment for this study consists of a PC with an Intel i7-11700 CPU (Intel, Santa Clara, CA, USA), an NVIDIA GeForce RTX 3070Ti Suprim 8G GPU (NVIDIA, Santa Clara, CA, USA), and 32GB of RAM. This section describes the datasets used for training and the test results after training in the experiments.

4.1. Dataset and Image Augmentation

This study utilized the Indiana University Chest X-ray Collection [27], consisting of 7466 chest X-ray images, each accompanied by its corresponding diagnostic report. The diagnostic reports include seven sections, but this research only utilized three for report generation: Indication, Findings, and Impression. Other sections are privacy-related and not used. When training with the IU Dataset, the distribution between the training and test sets was 80% for training and 20% for testing. As shown in Figure 5, data augmentations such as horizontal flipping, rotation, and resizing were applied to the images to prevent overfitting. Despite these augmentations, the diagnostic reports of the augmented images remained identical to those of the original images. In the end, the total dataset comprised 29,632 images.

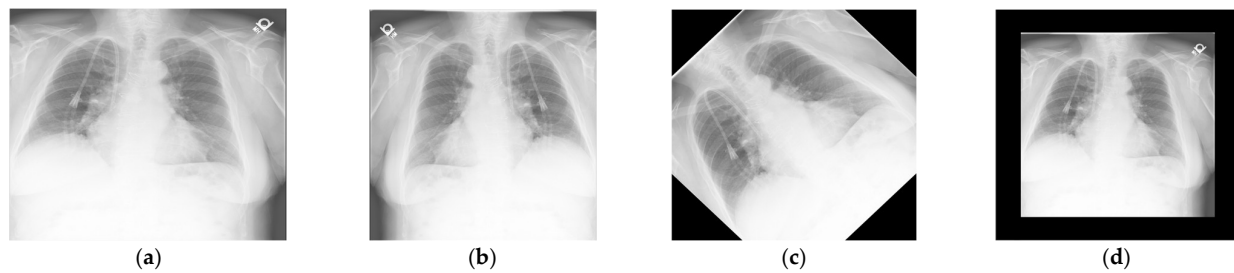


Figure 5. Images for data augmentation. (a) Original image. (b) Horizontal flipping. (c) Rotation. (d) Resizing.

4.2. Training Results

Three feature extraction models and four attention mechanisms are used in various combinations. The three feature extraction models are ResNet152V2, InceptionV3, and Inception–ResNetV2, all using “ImageNet weight” as the pre-trained weights. The four attention mechanisms are Bahdanua Attention [16], Luong Attention [17], Self-Attention [18], and the proposed Maximum Attention.

4.2.1. Feature Extraction Model Selection

The Transformer has emerged as a state-of-the-art architecture across a wide range of natural language processing tasks, and its effectiveness is reaffirmed in this study. Following extensive parameter tuning, the optimal configuration was identified for each feature extraction model. Each Transformer training session was conducted over 1000 epochs, with an average duration of approximately 10 to 12 h.

Leveraging parallel computation, the Transformer demonstrates a substantial advantage over recurrent neural network (RNN)-based models in terms of training efficiency. Figure 6a–c depict the training loss curves for the Transformer paired with ResNet152V2, InceptionV3, and Inception–ResNetV2, respectively. As shown at the last row in bold of Table 1, Inception–ResNetV2 consistently outperforms the other two feature extractors across all evaluation metrics.

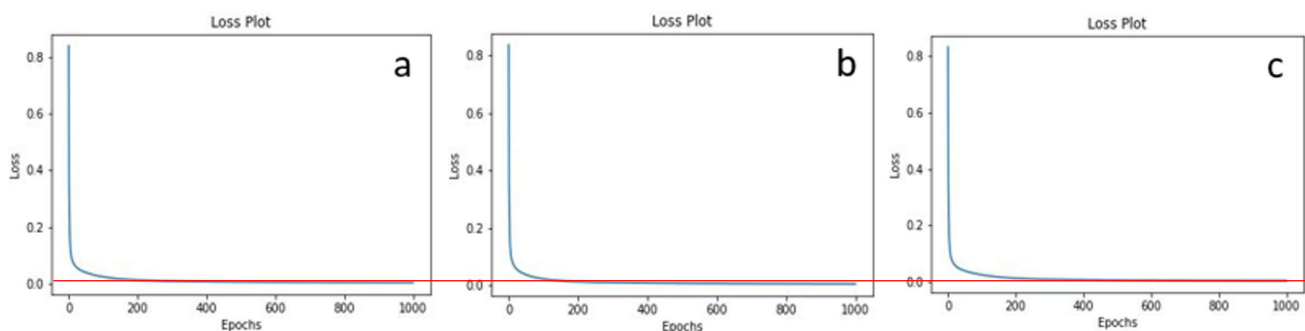


Figure 6. Training loss of Transformer with (a) ResNet152V2, (b) InceptionV3, (c) Inception–ResNetV2, as feature extraction. All the CNNs evaluated have the potential to serve as effective feature extractors. Based on the rate at which these models approach the red horizontal line, model (c) converges the fastest.

Table 1. Transformer performance across different feature extraction models.

Feature Extraction	Training (hr)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BERTScore		
							P _{BERT}	R _{BERT}	F _{BERT}
ResNet152V2	12	0.624	0.556	0.528	0.464	0.653	0.832	0.551	0.663
InceptionV3	10	0.655	0.592	0.566	0.507	0.679	0.864	0.597	0.706
Inception–ResNetV2	12	0.702	0.649	0.627	0.577	0.726	0.895	0.652	0.754

Notably, the BLEU-1 score achieved by the Transformer–Inception–ResNetV2 combination reaches 0.63, surpassing the performance of the Memory-driven Transformer. Based on these results, we conclude that pairing Inception–ResNetV2 with the Transformer architecture yields the most effective configuration for diagnostic report generation in this study and offers strong potential for similar medical imaging tasks.

4.2.2. Proposed Medical Transformer

The Medical Transformer was trained over 1000 epochs, with each training session requiring approximately 10 to 12 h. Figure 7a–c illustrate the training loss curves for the Medical Transformer when paired with ResNet152V2, InceptionV3, and Inception–ResNetV2, respectively. As shown in Table 2, Inception–ResNetV2 consistently delivers the highest performance among the evaluated feature extraction models. Notably, its integration with the Medical Transformer results in improved BLEU, METEOR, and BERTScore scores compared to the original Transformer architecture, further validating its effectiveness for diagnostic report generation.

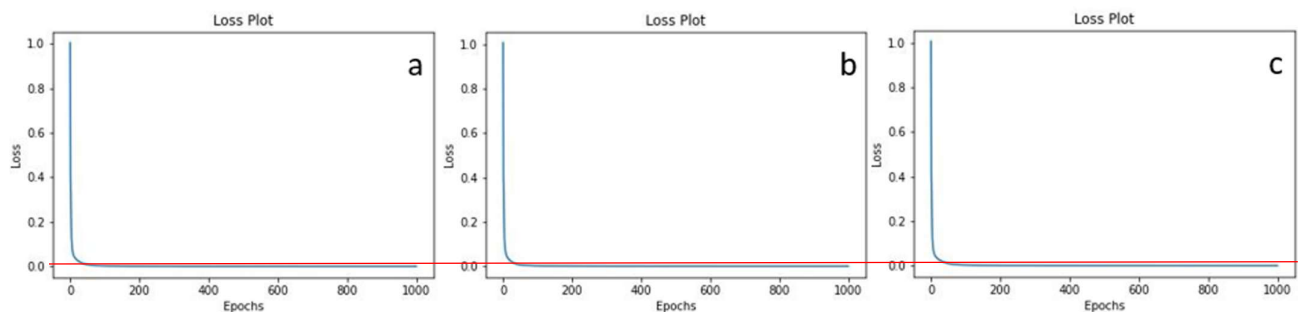


Figure 7. Training loss of the proposed Medical Transformer with (a) ResNet152V2, (b) InceptionV3, (c) Inception–ResNetV2, as feature extraction. All the tried CNNs are excellent feature extractors. As indicated by the horizontal red line, (c) exhibits the fastest convergence.

Table 2. Evaluation of Medical Transformer using different visual encoders.

Feature Extraction	Training (hr)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BERTScore		
							P _{BERT}	R _{BERT}	F _{BERT}
ResNet152V2	12	0.598	0.525	0.495	0.426	0.629	0.846	0.542	0.661
InceptionV3	12	0.624	0.556	0.527	0.463	0.651	0.875	0.579	0.697
Inception–ResNetV2	12	0.720	0.669	0.648	0.600	0.741	0.924	0.685	0.787

4.2.3. Comparisons of Different Attention Mechanisms

As summarized in Table 3, Bahdanau Attention and Luong Attention were trained over 100 epochs, each requiring approximately 7 h. Their highest BLEU-1 scores reached 0.313 and 0.293, respectively. In contrast, both the Transformer and the proposed Medical Transformer were trained for 1000 epochs, with a training duration of approximately 12 h. These models achieved significantly higher BLEU-1 scores of 0.702 and 0.720, respectively.

Table 3. Comparisons of the evaluation of various attention mechanisms.

Attention Types	(Hr., Epoch)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BERTScore		
							P _{BERT}	R _{BERT}	F _{BERT}
Bahdanau (Additive)	7, 100	0.313	0.217	0.179	0.089	0.354	0.336	0.347	0.341
Luong (Multiplicative)	7, 100	0.293	0.200	0.161	0.073	0.334	0.312	0.325	0.318
Transformer (Baseline)	12, 1000	0.702	0.649	0.627	0.577	0.726	0.895	0.652	0.754
Proposed (Maximum)	12, 1000	0.720	0.669	0.648	0.600	0.741	0.924	0.685	0.787

The results demonstrate that Transformer-based architectures outperform RNN-based models with attention mechanisms in both training efficiency and evaluation metrics. Furthermore, the proposed Medical Transformer surpasses the standard Transformer in BLEU-4, METEOR, and BERTScore performance, validating the effectiveness of its architectural enhancements. Based on these findings, the Medical Transformer is selected as the core neural network architecture for the diagnostic recognition system developed in this study.

4.3. Testing Result Analysis

This subsection presents the recognition results obtained using the Inception-ResNetV2 feature extractor in combination with the Medical Transformer. Figure 8a,b demonstrate cases where the predicted captions are identical to the reference captions, resulting in perfect evaluation scores—BLEU-1 through BLEU-4, METEOR, and F_{BERT} all reaching 1.0.

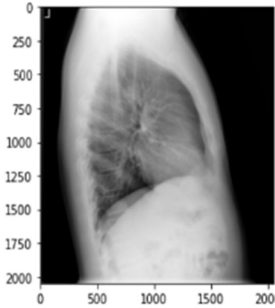
(a)		Real Caption: startseq indications indication history of tobacco use endseq startseq findings the lungs appear clear there are no suspicious appearing pulmonary nodules or masses there is no evidence of pneumonia the heart pulmonary xxxx appear normal pleural spaces are clear mediastinal contours are normal endseq startseq impressions no acute cardiopulmonary disease endseq. Predicted Caption: startseq indications indication history of tobacco use endseq findings the lungs appear clear there are no suspicious appearing pulmonary nodules or masses there is no evidence of pneumonia the heart pulmonary xxxx appear normal pleural spaces are clear mediastinal contours are normal endseq startseq impressions no acute cardiopulmonary disease endseq. BLEU-1 score: 1.0 BLEU-2 score: 1.0 BLEU-3 score: 1.0 BLEU-4 score: 1.0 METEOR score: 1.0 F_{BERT} score: 1.0
(b)		Real Caption: startseq indications shortness of breath endseq startseq findings stable cardiomegaly calcified tortuous aorta no focal air space disease stable chronic lung changes no large pleural effusion or pneumothorax diffuse degenerative changes in the thoracic spine endseq startseq impressions no acute cardiopulmonary abnormalities endseq. Predicted Caption: startseq indications shortness of breath endseq startseq findings stable cardiomegaly calcified tortuous aorta no focal air space disease stable chronic lung changes no large pleural effusion or pneumothorax diffuse degenerative changes in the thoracic spine endseq startseq impressions no acute cardiopulmonary abnormalities endseq. BLEU-1 score: 1.0 BLEU-2 score: 1.0 BLEU-3 score: 1.0 BLEU-4 score: 1.0 METEOR score: 1.0 F_{BERT} score: 1.0

Figure 8. (a) The prediction is identical to the actual caption. (b) The prediction is like the actual caption.

However, not all results exhibit such high accuracy. Based on an analysis of the lower-performing cases, three potential factors contributing to suboptimal outcomes have been identified:

i. Length Mismatch: Reference Longer Than Prediction

When the reference caption is substantially longer than the predicted output, the brevity penalty imposed by evaluation metrics such as BLEU, METEOR, and BERTScore increases, resulting in lower scores. For example, in Figure 9a, the Indication section of the predicted report deviates from the reference, and the Finding only describes normal anatomical features, omitting pathological details. Consequently, the Impression differs significantly from the ground truth. Similarly, in Figure 9b, the Finding fails to identify any abnormalities, resulting in a shorter and semantically incomplete report, which in turn leads to reduced evaluation scores.

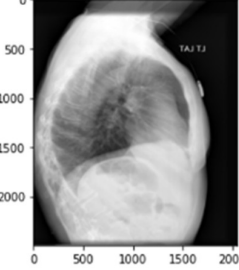
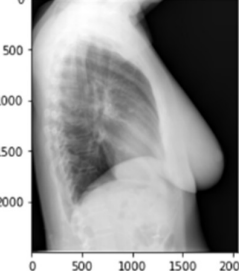
(a)		Real Caption: startseq indications chest pain after xxxx endseq startseq findings there is a cortical irregularity along the anterior margin of the sternum in addition there is a focal retrosternal hypodense convexity the cardiac silhouette is within normal limits the thoracic aorta is tortuous however the mediastinal contours are within normal limits there is no pneumothorax there is no large pleural effusion there is streaky xxxx kyphosis there is hyperinflation of the lungs endseq startseq impressions core irregularity along the anterior margin of the sternum may represent an age indeterminate nondisplaced fracture in addition focal lentiform hyperensity along the xxxx aspect of the sternum may represent callus formation left basilar atelectasis otherwise clear lungs endseq. Predicated Captions: startseq indications xxxx year old with syncope endseq startseq findings the heart and mediastinal contours are stable there is minimal patchy right lower lobe airspace disease identified no pleural effusion or pneumothorax endseq startseq impressions 1 patchy right lower lobe airspace disease may be due to atelectasis or infiltrate endseq. BLEU-1 score: 0.089 BLEU-2 score: 0.057 BLEU-3 score: 0.041 BLEU-4 score: 6.6×10^{-77} METEOR score: 0.162 F_{BERT} score: 0.184
(b)		Real Caption: startseq indications chest pain endseq startseq findings the cardiac silhouette upper mediastinum and pulmonary vasculature are within normal limits there is no acute pulmonary consolidation large effusion or pneumothorax there is minimal left basilar atelectasis there are small bilateral pulmonary nodules measure approximately 5 mm in size in the right midlung and left upper lung xxxx these are not well appreciated on the lateral projection endseq startseq impressions 1 no focal pulmonary consolidation or effusion minimal left basilar atelectasis 2 bilateral pulmonary nodules these were not present on the prior study and may represent sequela of infection but could represent neoplastic process correlation with history of primary malignancy is recommended further evaluation xxxx of the thorax could be performed if clinically indicated endseq. Predicted Caption: startseq indications xxxx endseq startseq findings the heart is normal in size the mediastinum is unremarkable the lungs are clear endseq startseq impressions no acute disease endseq. BLEU-1 score: 0.232 BLEU-2 score: 0.143 BLEU-3 score: 0.010 BLEU-4 score: 0.005 METEOR score: 0.154 F_{BERT} score: 0.284

Figure 9. (a) The prediction in the Finding part is incomplete, with too short a length, and the Impression part is inaccurate. (b) The prediction missed detecting the disease in the Finding part, leading to the error in the Impression part.

ii. Length Mismatch: Prediction Longer Than Reference

Conversely, when the reference caption is unusually short, the model may generate a longer output, again triggering a penalty. In Figure 10a, the predicted Indication incorrectly mentions lower back pain, and the Finding includes several conditions that are not present in the reference. This over-generation results in a longer caption and a lower assessment score. In Figure 10b, the predicted Finding describes multiple anatomical regions, whereas the reference states “nan,” leading to a stark length discrepancy and a poor evaluation outcome. However, F_{BERT} would suggest a higher similarity from the token viewpoint for these two cases.

iii. Rare Vocabulary and Data Sparsity

Some recognition failures stem from the infrequent occurrence of specific terms in the training dataset. For instance, in Figure 11a, words such as “bleed,” “uncalcified,” and “adjacent” appear in the reference but are rarely seen in the training corpus, making them unlikely to be generated. In Figure 11b, the phrase “curvature of the thoracolumbar spine” refers to a rare condition that the model fails to detect. Moreover, while the reference describes several normal anatomical features, the prediction omits them entirely, substituting “nan” instead. These omissions contribute to a substantial length and semantic gap, resulting in a BLEU-1 score as low as 0.0189.

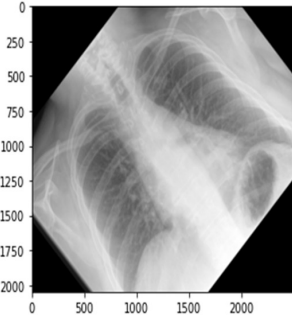
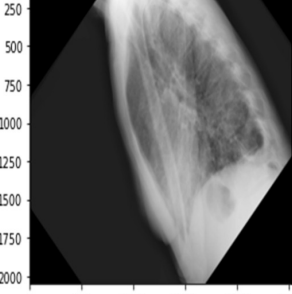
(a)		Real Caption: startseq indications endseq startseq findings the lungs are clear the heart and pulmonary xxxx are normal the pleural spaces are clear mediastinal contours are normal endseq startseq impressions no acute cardiopulmonary disease endseq. Predicated Caption: startseq indications low back pain endseq startseq findings the heart size and pulmonary vascularity appear within normal limits the lungs are free of focal airspace disease no pleural effusion or pneumothorax is noted degenerative changes are noted in the spine the descending thoracic aorta is mildly tortuous the mediastinum appears somewhat prominent endseq startseq impressions 1 prominent mediastinum may be due to mediastinal fat comparison films if available would be useful to determine if this is a chronic appearance 2 clear lungs endseq. BLEU-1 score: 0.289 BLEU-2 score: 0.188 BLEU-3 score: 0.149 BLEU-4 score: 0.068 METEOR score: 0.476 FBERT score: 0.523
(b)		Real Caption: startseq indications nan endseq startseq findings nan endseq startseq impressions nan endseq. Predicated Caption: startseq indications the patient is a xxxx year old woman preop lung transplant endseq startseq findings left sided medication injection xxxx has its tip projecting at the cavoatrial junction the trachea is midline extensive bilateral bronchiectasis cystic changes and scarring represents sequela from the patient's cystic fibrosis no evidence of focal pulmonary infiltrate or pleural effusion no large pneumothorax has developed in the interim the overlying bony structures reveal no acute abnormalities the heart size is normal endseq startseq impressions 1 extensive pulmonary bronchiectasis and scarring from cystic fibrosis not significantly xxxx from prior 2 left sided medication injection xxxx. BLEU-1 score: 0.08 BLEU-2 score: 0.064 BLEU-3 score: 0.06 BLEU-4 score: 0.0 METEOR score: 0.338 FBERT score: 0.57

Figure 10. (a) Both the Indication and the Finding of the recognition result diagnose the condition that the patient does not have. (b) The Finding of the identification result describes all the problematic areas. In contrast, the actual result is only represented by nan, which is a big difference in the length of the two descriptions, resulting in a lower evaluation score.

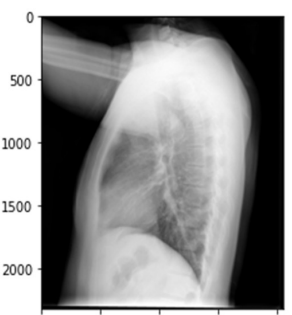
(a)		Real Caption: startseq indications xxxx bleed endseq startseq findings in the interval a 3 cm uncalcified mass has developed in the posterior segment of the right upper lobe in addition on the pa view an 8 mm opacity is adjacent to the left xxxx of the heart this opacity cannot be well identified on the lateral view it may be artifactual but another mass on the left cannot be excluded mediastinum is normal with no evidence for adenopathy heart size normal noted xxxx of an unchanged hiatal hernia endseq startseq impressions xxxx right upper lobe mass suspicious for neoplasm ct of chest abdomen and head would be helpful for further evaluation endseq. Predicated Caption: startseq indications xxxx endseq startseq findings there is a stable closure device projected over the heart and mediastinum are otherwise normal there is stable xxxx scarring of left mid lung ther lungs are otherwise clear there is no infiltrate effusion mass or pneumothorax endseq startseq impressions stable appearance of the chest endseq. BLEU-1 score: 0.190 BLEU-2 score: 0.101 BLEU-3 score: 0.071 BLEU-4 score: 1.13×10^{-78} METEOR score: 0.214 FBERT score: 0.376
(b)		Real Caption: startseq indications xxxx year old male who complains of xxxx with inhalation and left sided chest pain left mid quadrant pain concern for stones endseq startseq findings chest normal heart size mediastinal silhouette is unremarkable no focal infiltrates or masses no pneumothorax or visible pleural fluid no free intraperitoneal air in the diaphragm osseous structures unremarkable abdomen there are no dilated loops of bowel to suggest obstruction no air fluid levels or free intraperitoneal air no suspicious calcifications there is xxxx curvature of the thoracolumbar spine otherwise the osseous structures are grossly unremarkable endseq startseq impressions 1 chest no radiographic evidence of acute cardiopulmonary abnormality 2 abdomen nonobstructive bowel xxxx pattern endseq. Predicetd Caption: startseq indications tb endseq startseq findings nanendseq startseq impressions heart size normal lungs clear no evidence of tuberculosis no change prior exam endseq. BLEU-1 score: 0.0189 BLEU-2 score: 0.013 BLEU-3 score: 0.009 BLEU-4 score: 1.29×10^{-79} METEOR score: 0.14 FBERT score: 0.352

Figure 11. (a) The recognition result does not show the rarer words. (b) The recognition result is omitted, and the description is too concise.

Poor recognition performance is primarily driven by limitations in the training data. When the reference captions differ significantly in length or contain rare terminology not well represented in the dataset, the model struggles to generate accurate and comprehensive outputs. Enhancing dataset diversity and balancing caption lengths are essential for improving recognition fidelity.

However, FBERT demonstrates superior semantic similarity in cases where the predicted and reference texts differ in length. From a pragmatic perspective, BERTScore captures the intended meaning beyond literal word matching by accounting for the speaker's intent and contextual nuances. This enables the model to recognize implied or “invisible” meanings that are not explicitly stated, offering a more human-aligned evaluation of semantic equivalence.

4.4. Comparisons with SOTA

This subsection presents the comparisons of the surveyed references in Section 2 with respect to the deployed approach, the used dataset, and the resultant performance. Since ImageCLEF [29] is for multiple scenarios or domains, it is not used in this study for evaluation. Table 4 shows that our approach achieved the highest BLEU and METEOR scores. This should be attributed to several reasons. The first is image augmentation, widely used in image classification. Here, Inception–ResNet152V2 extracts the important visual features after training with all variant images of the same syndromes. Along with the proposed maximum attention mechanism, which again relates the visual features to the corresponding medical report keywords. Together, the last row in bold demonstrated our superior performance.

Table 4. Comparisons of the proposed approach with some SOTA approaches.

Approach	Methods	Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Yin et al. [19]	HRNN/LSTM	IU X-Ray	0.445	0.292	0.201	0.154	0.175
Beddiar et al. [20]	VGG16 + LSTM	ImageCLEF/multiple domains	0.423	-	-	-	-
			(BLEU)				
Chen et al. [7]	ResNet + Memory-driven Transformer	IU X-Ray	0.470	0.304	0.219	0.165	0.187
Wijerathna et al. [24]	CheXNet + Memory-driven Transformer	IU X-Ray	0.498	0.32	0.229	0.169	0.205
Selivanov et al. [26]	SAT + GPT3	IU X-Ray	0.520	0.390	0.296	0.235	-
Liu et al. [27]	Label correlated contrastive learning	IU X-Ray	0.508	0.321	0.229	0.172	-
Ramadan & Akay [28]	CvT + GPT2	IU X-Ray	0.414	0.265	0.190	0.144	-
Ours	Augmentation + CNN + Medical Transformer	IU X-Ray	0.720	0.669	0.648	0.600	0.741

5. Conclusions and Future Works

This study systematically evaluated three feature extraction models and four attention mechanisms for the task of diagnostic report generation from chest X-ray images. Among the evaluated configurations, the highest performance was achieved by combining image augmentation—used to accommodate chest X-ray image variability—with Inception–ResNetV2 for feature extraction and the proposed Medical Transformer for report generation. This setup yielded BLEU-1 to BLEU-4 scores of 0.720, 0.669, 0.648, and 0.600, respectively, along with a METEOR score of 0.741 and a BERTScore F_{BERT} of 0.787. These results exceeded the target threshold of 0.5 established for this study and

outperformed the standard Transformer architecture, underscoring the effectiveness of the Medical Transformer's computational design. Given its superior performance, the Medical Transformer shows promise for broader applications beyond chest radiography.

Despite significant advancements in medical technology, the recent global pandemic has underscored the persistent shortage of healthcare personnel. The models and methods proposed in this research aim to alleviate the burden on clinicians by automating aspects of diagnostic reporting and supporting clinical decision-making. As even experienced practitioners are susceptible to oversight, especially under time constraints, automated systems can serve as valuable references to reduce the likelihood of diagnostic errors.

While the Inception-ResNetV2 and Medical Transformer pairing demonstrated strong results, the experimental findings also revealed instances of suboptimal report generation. Future work should explore the integration of larger and more diverse datasets, as well as the adoption of novel feature extraction architectures, to further enhance model robustness and generalization.

Currently, the proposed system is limited to chest X-ray interpretation. To extend its utility across radiological domains, future research should aim to generalize the model to other anatomical regions, such as the brain, limbs, and abdomen. Expanding the scope of diagnostic coverage would enable radiologists to allocate more time to complex cases and reduce the risk of missed diagnoses, thereby improving overall clinical efficiency and patient outcomes.

In future work, we aim to address the challenge of patient-specific anatomical variability and anomaly identification by incorporating structured clinical metadata—such as age, sex, height, and medical history—into the model's input pipeline. This multimodal integration would enable more personalized and context-aware diagnostic reporting. From a commercial perspective, we recognize the importance of ethical compliance, legal accountability, and clinical safety. Any future deployment of this system will require rigorous validation, transparent decision-making, and adherence to regulatory standards to ensure responsible AI usage in healthcare environments.

Author Contributions: Conceptualization, C.-C.H.; Data curation, M.-H.H.; Formal analysis, C.-C.H.; Funding acquisition, M.-H.H.; Methodology, S.-P.L. and C.-C.H.; Project administration, M.-H.H.; Software, S.-P.L.; Validation, M.-H.H. and C.-C.H.; Writing—original draft, S.-P.L.; Writing—review & editing, C.-C.H. All authors have read and agreed to the published version of the manuscript.

Funding: Sanming University. Research Foundation for Advanced Talents, grant number: 22YG11S.

Data Availability Statement: The used dataset was reported in [28], including links to publicly archived datasets at <https://pubmed.ncbi.nlm.nih.gov/26133894/> (accessed on 23 August 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Demner-Fushman, D.; Kohli, M.D.; Rosenman, M.B.; Shooshan, S.E.; Rodriguez, L.; Antani, S.; Thoma, G.R.; McDonald, C.J. Preparing a Collection of Radiology Examinations for Distribution and Retrieval. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 304–310. [CrossRef] [PubMed]
2. Jing, B.; Xie, P.; Xing, E. On the Automatic Generation of Medical Imaging Reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 2577–2586.
3. Hodosh, M.; Young, P.; Hockenmaier, J. Framing Image Description as a Ranking Task: Data, Models, and Evaluation Metrics. *J. Artif. Int. Res.* **2013**, *47*, 853–899. [CrossRef]
4. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6 July 2002; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318.

5. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 23 June 2007; Association for Computational Linguistics: Stroudsburg, PA, USA, 2007; pp. 65–72.
6. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020. [\[CrossRef\]](#)
7. Chen, Z.; Song, Y.; Chang, T.H.; Wan, X. Generating Radiology Reports via Memory-Driven Transformer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 16–20 November 2020; pp. 1439–1449.
8. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [\[CrossRef\]](#)
9. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.; Liu, W.; et al. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [\[CrossRef\]](#)
10. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2 December 2015; pp. 2818–2826.
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [\[CrossRef\]](#)
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *Computer Vision—ECCV 2016*; ECCV 2016. Lecture Notes in Computer Science, 9908; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016. [\[CrossRef\]](#)
13. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI’17), San Francisco, CA, USA, 4–9 February 2017; AAAI Press: Washington, DC, USA, 2017; pp. 4278–4284.
14. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems—(NIPS’14), Volume 2, Cambridge, MA, USA, 8 December 2014; MIT Press: Cambridge, MA, USA, 2014; pp. 3104–3112.
15. Chung, J.; Gulcehre, C.; Cho, K.H.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555. [\[CrossRef\]](#)
16. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR. *arXiv* **2015**, arXiv:1409.0473.
17. Luong, M.T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 1412–1421.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008.
19. Yin, C.; Qian, B.; Wei, J.; Li, X.; Zhang, X.; Li, Y.; Zheng, Q. Automatic Generation of Medical Imaging Diagnostic Report with Hierarchical Recurrent Neural Network. In Proceedings of the IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019; pp. 728–737.
20. Beddiar, D.R.; Oussalah, M.; Seppänen, T.; Jennane, R. ACapMed: Automatic Captioning for Medical Imaging. Online. *Appl. Sci.* **2022**, *12*, 11092. [\[CrossRef\]](#)
21. Jesse, R.; Pérez-Cruz, F. Deep Learning for Multi-label Classification. *arXiv* **2014**, arXiv:1502.05988. [\[CrossRef\]](#)
22. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
23. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Wijerathna, V.; Raveen, H.; Abeygunawardhana, S.; Ambegoda, T.D. Chest X-ray Caption Generation with CheXNet. In Proceedings of the Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, 27–29 July 2022; pp. 1–6.
25. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225.
26. Selivanov, A.; Rogov, O.Y.; Chesakov, D.; Shelmanov, A.; Fedulova, I.; Dylov, D.V. Medical Image Captioning via Generative Pretrained Transformers. *Sci. Rep.* **2023**, *13*, 4171. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Liu, X.; Xin, J.; Dai, B.; Shen, Q.; Huang, Z.; Wang, Z. Label Correlated Contrastive Learning for Medical Report Generation. *Comput. Methods Programs Biomed.* **2025**, *258*, 108482. [\[CrossRef\]](#) [\[PubMed\]](#)

28. Ramadan, W.; Akay, B. A Hybrid Deep Learning Approach for Improving Medical Image Captioning using Convolutional Vision Transformer. In Proceedings of the 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), Antalya, Turkiye, 7–9 August 2025; pp. 1–7. [\[CrossRef\]](#)
29. Ionescu, B.; Müller, H.; Drăgulescu, A.-M.; Rückert, J.; Ben Abacha, A.; de Herrera, A.G.S.; Bloch, L.; Brüngel, R.; Idrissi-Yaghir, A.; Schäfer, H.; et al. Overview of the ImageCLEF 2024: Multimedia Retrieval in Medical Applications. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*; CLEF 2024. Lecture Notes in Computer Science; Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Di Nunzio, G.M., Soulier, L., Galuščáková, P., Seco de Herrera, A.G., Faggioli, G., Ferro, N., Eds.; Springer: Cham, Switzerland, 2024; Volume 14959. [\[CrossRef\]](#)
30. Keras. Keras Application. Available online: <https://keras.io/api/applications/> (accessed on 10 September 2025).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.