Biomedical Signal Processing and Control 111 (2026) 108340

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Radiology report generation from a singular perspective using transformers with Knowledge Distillation

Asad Mansoor Khan [a,*], Mashood Mohammad Mohsan [a], Muhammad Usman Akram [a], Taimur Hassan [b], Sajid Gul Khawaja [a], Adil Qayyum [c]

[a] National University of Sciences and Technology, Islamabad, 44000, Pakistan
[b] Department of Electrical, Computer, and Biomedical Engineering, Abu Dhabi University, United Arab Emirates
[c] Resident Radiologist, Combined Military Hospital, Rawalpindi, Pakistan

ARTICLE INFO

Keywords:
Medical imaging
Report generation
Knowledge distillation
Pre-training

ABSTRACT

Nearly two billion chest X-rays (CXRs) are performed annually, making them the most used imaging technique in radiology for the diagnosis of pulmonary disorders. The accompanying report with the findings from a chest X-ray forms a crucial part of the examination. By providing an accurate report, healthcare professionals can be enabled to make better decisions about the care being provided. To this end, we propose an end-to-end radiology report generation framework built on transformers trained on text reports in conjunction with visual characteristics of the chest X-ray to generate a reliable report that astutely describes the findings from a single CXR taken either from the Anterior-Posterior or Posterior-Anterior position. A foundation model is utilised to perform Knowledge Distillation (KD) in conjunction with the Encoder which is fine-tuned during the training phase. In addition, using a large corpus of radiology reports to pre-train the foundation model in an unsupervised manner is shown to improve the performance on smaller datasets. This training methodology results in comparable performance to architectures that employ a lot more parameters. The proposed framework is evaluated on multiple datasets including the Indiana University dataset, MIMIC dataset, MIMIC-PRO dataset, and BRAX dataset. The incorporation of KD results in an increase of BLEU-1 score for Indiana dataset by 4% and BERTScore by 7.5%. Similarly, pre-training on larger datasets in combination with KD, further increases BLEU-1 score for Indiana dataset by 7.2% and BERTScore by 3%. For MIMIC dataset, comparable performance is achieved for the Findings and the Impression sections of the report while the proposed framework outperforms other techniques when both of these sections are combined. For MIMIC-PRO dataset, an s_{emb} score of 0.4069 while a RadGraph F1 score of 0.1165 is achieved outperforming other techniques in the literature. Finally, the proposed framework is also evaluated on locally gathered dataset and BRAX subset without any re-training or fine-tuning resulting in BLEU-1 score of 0.3827 and a BERTScore of 0.4392 for the former and BLEU-1 score 0.1671 of and a BERTScore of 0.2186 for latter showing generalisation ability.

1. Introduction

For the diagnosis of pulmonary disease, radiologists rely on chest X-rays (CXR) since they are a low-cost, readily available method with the added advantage that they can even be brought to the patient if the circumstances necessitate [1,2]. Despite the greater sensitivity and diagnostic efficacy of Computed Tomography (CT) scans compared to CXR [3], the widespread availability and practicality of the CXR significantly enhance its utility as a diagnostic tool. Additionally, the reduced radiation exposure further bolsters its appeal. Owing to these reasons, the demand for CXR evaluation and interpretation surpasses the capacity of radiologists to address it effectively [4,5]. Adding to the utility of CXR, are the associated medical reports that provide supplementary insights that may not be apparent from the imaging itself thus making such a report quite useful. Automated report generation from a single or a pair of CXRs can allow for rapid evaluation of a patient's condition as not only it can be generated quickly but can be made less error-prone as compared to a non-experienced radiologist [6]. The use of diverse deep learning-based techniques has become more prevalent

* Corresponding author.
E-mail addresses: asad.mansoor@ceme.nust.edu.pk (A.M. Khan), mmohsan.cse20ceme@student.nust.edu.pk (M.M. Mohsan), usman.akram@ceme.nust.edu.pk (M.U. Akram), taimur.hassan@ku.ac.ae (T. Hassan), sajid.gul@ceme.nust.edu.pk (S.G. Khawaja), qayyum.adil@gmail.com (A. Qayyum).

https://doi.org/10.1016/j.bspc.2025.108340
Received 6 February 2024; Received in revised form 2 November 2024; Accepted 27 June 2025
Available online 16 July 2025

where the problem can either be modelled as a retrieval problem because the radiology report follows a template [7] or a generation problem where a free-text report is generated from scratch.

Li et al. [8] employed the retrieval-based approach and proposed a Knowledge-Driven Encode, Retrieve, Paraphrase (KERP) framework. This framework adopted a sequential process involving graph transformers at each architectural phase using either the latent space output or the graph output from the preceding module. By the use of a Graph Transformer (GTR), the features in the latent space are used for generating an abnormality graph. The abnormality graph is again transformed using GTR, resulting in a sequence of templates that are enriched with case-specific information along with improving the text itself to make it more dynamic using the *Paraphrase* module. The researchers validated their research on two datasets: the Indiana University Chest X-ray dataset [9] and a private dataset. Li et al. [10] also proposed a Hybrid retrieval-generation reinforced (HRGR) agent - a similar approach to [8] of encoding visual features extracted through a CNN [11] as a context vector which was used to generate topic states by the use of stacked RNN layers with attention. A retrieval policy module was then used to determine whether a template should be retrieved or a new sentence generated depending on the probability. The researchers trained this module using hierarchical reinforcement learning.

The conventional sequence-to-sequence architecture was expanded by Chen et al. [12] by the introduction of two additional modules. In the proposed framework, latent space features extracted from an image are regarded as input sequences. A relational memory module was added to store the pattern information in order to take advantage of similarities between the images. ResNet101 [13] was used as the feature extractor and the feature length was capped at 2048. Using a Multilayer Perceptron to predict the $\Delta\gamma$ and $\Delta\beta$ from the matrix from the memory module, the second module called Memory-driven Conditional Layer Normalisation (MCLN) is responsible for improving the generalisation capabilities of the decoding process. [14] utilised Long Short-Term Memory (LSTM) [15] in conjunction with CNN and attention module to generate X-ray reports. They modified the VGG19 architecture by integrating additional fully connected layers at the end, yielding a feature vector of length 256. Employing an embedding layer, the latent space vector is converted to embeddings where one dimension represents the maximum size of the findings in [9] dataset. An attention mechanism generates a context vector prior to the embeddings being used by the LSTM for the generation of the output words, which is then combined with the hidden state to forecast the following word. The inclusion of a context vector allows the framework to selectively focus on the information that is most relevant to the region of interest. Similar to [12], two datasets [9,16] were used for gauging the performance of the proposed methodology.

Liu et al. [17] used the Knowledge Graph Auto-Encoder (KGAE), an unsupervised method that relies on a pre-built knowledge graph, to reduce the reliance on datasets that contain image-report pairs. The knowledge graph, integral to both the encoder and decoder, is constructed using [16] reports instead of the images, with the abnormalities acting as nodes and the normalised co-occurrence of these abnormalities in the reports serving as edge weights. ResNet50 [13] and Transformer [18] are used to extract the embeddings which are then used as queries to the pre-built knowledge graph. Similarly, a knowledge-driven decoder is used to generate textual reports from the graph representations of the image. The researchers also modified their proposed approach to make it semi-supervised and supervised by incorporating some of the image-report pairs in the generation of the knowledge graph. Their findings showed that the supervised approach outperformed both the semi-supervised and the unsupervised approach. In order to mimic the process of report writing by the radiologists, Liu et al. [19] proposed a Posterior and Prior Knowledge Exploring and Distilling (PPKED) approach. Three primary components formed this approach: Multi-domain Knowledge Distiller (MKD), Prior Knowledge Explorer (PrKE), and Posterior Knowledge Explorer (PoKE). PoKE was used to extract the abnormality information from the image embeddings and the embeddings from a bag of Tags T containing the most common abnormality tags. Using multi-head attention and Feed Forward Network (FFN), the output from the image and tag embeddings was normalised and added to generate the final output of PoKE. The next component, PrKE, uses the results of PoKE to generate two more outputs: Prior Working Experience created by generating embeddings from the text reports of the 100 closest images using cosine similarity from the training corpus, and Prior Medical Knowledge created by combining the results of PoKE with embeddings from a knowledge graph created using the same abnormality tags. MKD serves as the decoder, utilising the outputs from PoKE and PrKE.

Keeping in the same vain as incorporating previous knowledge, [20] proposed including two different forms of knowledge — general domain knowledge and image-specific knowledge — in the report generation task, the former of which is independent of the input and the latter depends on the input. The manually constructed RadGraph [21] was utilised for generic domain knowledge and as a relation extractor. Latent-space features were used to retrieve similar reports from a report pool to obtain image-specific information. The novel knowledge-enhanced attention module aggregates the embeddings generated from the generic domain knowledge and those extracted from the visual feature extractor. These embeddings are amalgamated into triplets derived from similar reports, further processed through Clinical BERT [22], and aggregated alongside the visual features. This concatenation facilitates the creation of image-specific domain knowledge. For report generation, the decoder from [18] is used with the generic knowledge, specific knowledge, and the visual features concatenated as input.

Srinivasan et al. [23] suggested a report-generation methodology that utilised Image Level Chest Features (ICLF) extracted using a CNN and Tag Level Chest Features (TCLF) extracted using Multi-Head Attention (MHA) by using only the lung area clipped using the Single Shot Detector (SSD) [24]. The CNN used for ICLF was modified such that it could classify the image as normal or abnormal based on overlapping patches to retrieve the tags for only diseased images. Triplet loss was used for training this sub-module. Using a concatenated version of ICLF passed through MHA, the top 16 tags were selected. The framework employed two [18]-esque encoders: one processing TCLF for embeddings and the other handling ICLF. Similarly, two decoders are used as well; one for generating *Findings* output based on both embeddings while the other generates *Impressions* based on output features from the first decoder. In addition to the aforementioned approaches, Large Language Models (LLMs) have also been utilised for the generation of radiology reports. Zhong et al. [25] utilised Llama2 for report generation that was trained on 0.3 million images from six different institutes. In addition to providing the report for CXRs, ChatRadio-Valuer was also adapted for the diagnosis of disease for musculoskeletal, head and neck regions based on specific prompts. The fine-tuned Llama2 model outperformed several other LLMs including GPT3.5 and GPT4. To gauge the performance of state-of-the-art LLMs for radiology report interpretation, Liu et al. [26] used 32 LLMs to interpret the radiology reports in [16] and OpenI dataset. ROUGE-L, ROUGE 1 and ROUGE-2 [27] were employed as the performance metrics with Zero-shot, One-shot and Five-shot evaluation methods. Claude2 was able to outperform all other models in a Zero-shot setting.

The majority of report generation frameworks for chest X-rays often rely on multiple views of the chest cavity, with only a scant few relying on a singular view, whether it Posterior-Anterior (PA) or Anterior-Posterior (AP). Additionally, many of these approaches tend to employ either large-scale models or a hybrid approach combining report retrieval and report generation. To address this gap, our main research contributions in this article are as follows:

1. We propose a report generation framework that is capable of generating a CXR report based on a singular view such as AP or PA.

2. The proposed framework employs foundation model fine-tuning on CXR reports for use as a Teacher model to train a smaller Student model. We also employ Knowledge Distillation (KD) so that a smaller Student model can learn better CXR representation.
3. We show that pre-training on larger CXR datasets with reports improves performance when used in conjunction with Knowledge Distillation providing reasonable performance at a relatively small size and simple architecture.
4. As there is no standard test set available for BRAX, therefore, we have provided manually annotated reports by radiologists for 100 images sourced from this dataset which can also be used for validating future frameworks and experimentation.
5. Locally collected dataset with radiology reports (findings) is used to validate the proposed framework.

The remaining paper is organised as follows: related work is described in Section 2. Section 3 Materials and Methods describes the data set and the proposed architecture. The results are given in Section 5. Section 6 provides the discussion of the results and finally Section 7 concludes the paper. Moreover, the source code of the framework proposed in this paper is available at: https://github.com/asadkhan1221/radiology_report_generation/tree/main

2. Related work

Our proposed framework generates a CXR report that provides the findings from a single CXR captured either from the AP or PA viewing position. Central to this framework are three components: an encoder–decoder architecture employed in the report-generation module, fine-tuning of a foundation model, and Knowledge Distillation (KD) performed using the encoder and the foundation model trained in the previous step. Keeping this in view, the current section provides an overview of the relevant literature.

Leveraging the transformer architecture put forth by [18] for primarily sequential data, Vision Transformer (ViT) [28] modified it to work with images by creating embeddings of size d of the flattened image patches. To these embeddings, positional encoding embeddings are added to keep a record of the position of each patch, and a learnable embedding representing the class of the image is prepended. Within the encoder block in the transformer architecture, two distinct sub-layers, namely the Multi-head Self Attention and the Feed Forward Network are present. For Self Attention (SA), instead of using the input embeddings directly, *Key (K), Query (Q)* and *Value (V)* are generated by projecting the input with different weight matrices. Eqs. (1) and (2) detail the operations that are performed on the *Key, Query* and *Value*. Skip connections are also incorporated to maintain consistent latent vector size followed by layer normalisation. ViT was shown to provide better performance than convolutional neural networks for image classification tasks [28]. ViT's versatility has found applications in various domains. [29] paired a Convolutional Neural Network with ViT where the former extracted feature maps and visual tokens from those feature maps and the latter modelled relationships between the visual tokens. This fusion approach was employed for both classification and semantic segmentation and was shown to improve the results compared to the convolution-only counterpart significantly while also using fewer parameters. In another study, [30] implemented a novel training strategy aimed at reducing both training time and data requirements for ViT models. Their proposed approach also incorporated knowledge distillation. Moreover, the study highlighted the advantageous integration of established optimisation and regularisation techniques typically employed in CNNs, demonstrating their utility in enhancing the training efficacy of ViTs.

$$\begin{aligned}
&\text{Similarity Score} = E = \frac{Q \cdot K^T}{\sqrt{D_q}},\\
&\text{Attention Weights} = A = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{D_q}}\right),\\
&\text{Attention(Q, K, V)} = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{D_q}}\right) \cdot V,
\end{aligned} \tag{1}$$

$$\begin{aligned}
&\text{Multi-Head Attention}(Q, K, V) = \text{Concat}(Head_1, Head_2, \ldots, Head_N) \cdot W^O\\
&\text{where Head}_i = Softmax\left(\frac{Q_i \cdot K_i^T}{\sqrt{D_q}}\right) \cdot V_i
\end{aligned} \tag{2}$$

The decoder architecture mirrors the encoder in that Masked Multi-Head Self Attention (MMHA), another sub-module, is included alongside Multi-Head Self Attention and Multilayer Perceptrons. In the Masked MHA layer, some of the values in the product of the Attention Weights and the Value V are masked limiting prediction dependency to preceding outputs, and offsets embeddings by one position. The MHA sub-module of the decoder computes the Self Attention utilising Keys K and Values V from encoder block's hidden state and Queries Q generated from the decoder input. Skip connections and normalisation layers, akin to the encoder, are applied by adding input to sub-layer output and normalising. Positional encoding embeddings are used similarly.

In computer vision, a model can learn broad representations of various objects by being pre-trained on a big labelled dataset with a large number of samples for a large number of classes [31]. The same model can then be fine-tuned for a comparable task using these weights. In the same vein, it is possible to train large language representation models for upstream tasks using unlabelled data and then fine-tune them for a particular downstream task — an approach that has shown performance improvements [32–35]. During training, certain words or phrases are masked in the input data, requiring the model to predict and fill in these blanks. BioBERT has applied the same approach to BERT [35] by pre-training the aforementioned model on medical corpora from sources such as PubMed and PMC articles totalling 18 billion words [36]. It was then further fine-tuned for three downstream tasks namely biomedical named entity recognition, biomedical relation extraction, and biomedical question answering where it achieved better performance compared to BERT.

Researchers have applied various types of Knowledge Distillation including response-based [37–40], feature-based [13,41–43] and relation-based [44–47] KD to take advantage of knowledge learnt by a larger, more complex model owing to its size and the number of parameters [48]. The smaller model (student) learns not only from the labelled data but also from the larger model (teacher). By distilling the knowledge encapsulated in the Teacher's predictions, representations, and learned relationships, the student model gains a more comprehensive understanding of the data distribution and underlying patterns. Sun et al. [49] proposed a novel KD method where the student is able to learn either from a series of specific k layers output of the teacher model or learns from every k^{th} layer in the model. This methodology was demonstrated to be effective at providing comparable performance to the large model at a fraction of the cost both in terms of training efficiency and parameter size. Jiao et al. [50] also employed BERT [35] as a teacher model to train a student model termed TinyBERT at only 13% the size of the teacher model while retaining close to 97% performance. They were able to achieve this performance by a two-step knowledge distillation process from [35]; initially, a general Teacher model is used to train a general Student model. Subsequently, the process is repeated utilising a task-specific Teacher and a task-specific Student.

3. Materials and method

This section describes the data sets, pre-processing for the reports, proposed framework, and the training methodology.

3.1. Data sets

Every chest X-ray examination is accompanied by a report that summarises the key findings of the CXR providing a thorough analysis of the chest cavity's condition from which a label for the CXR can also be obtained [51,52]. Recently, datasets with a focus on report generation have also become available and they serve as an important training resource for transformer-based and other architectures for the generation of reports. On three different data sets, the proposed methodology has been validated.

Gathered from different hospitals affiliated with the Indiana University School of Medicine, this dataset is a collection of 7784 frontal and lateral chest radiographs [9] available in the DICOM format. Accompanying these CXRs are 3927 unique reports that contain the key findings of the scan under different sections. The overall number of words is 1,22,096 and there are 3257 unique terms in the lexicon with over 75% of the paragraphs being unique.

MIMIC [16] dataset includes 377,110 images acquired from 227,835 radiography studies of 65,379 patients making it the largest collection of chest X-rays that is publicly accessible. The CXRs in this dataset were predominantly obtained in frontal (PA and AP) and lateral projections utilising a variety of different equipment. The screening may include one or more images with one or more projections for a single patient. Reports with the label *No finding* have the maximum number of around a 140,000 followed by*support devices*. *Pleural effusion* and *lung opacity* have the second and third most samples respectively when *support devices* are ignored. There are 324,641 words total in the dataset, with 145 words and 642 characters on average in each report. Each provided report has three main sections: findings, impressions, and free-form text without any heading. The findings provide an elaborate description of the key findings in the CXR while the impressions are far more concise. A report may contain both the findings, impressions, and free-form text but one or more may also be absent from a report.

The references to earlier CXR studies that may or may not be contained in the dataset are one issue that such a huge dataset poses. These prior references may limit the deep learning models' capacity to learn the representation, leading to reports that make references to erroneous data. Previous data might be ignored because the major goal of a radiological report is to highlight the findings in the most recent scan. This issue has been resolved by [53] by automatically rewriting the reports and eliminating past references. This dataset which can be considered a modified version of [16] contains 316,437 reports for an equivalent number of images in the train portion and 2188 reports in the test set containing both the frontal and lateral views.

The original split for MIMIC [16] and MIMIC-PRO [53] have been used and the split used for Indiana University [9] is taken from [54]. For all the datasets, only the AP and PA view positions are kept while all other view positions are discarded reducing the number of samples that are used for training. The textual data must be cleaned as part of the standardisation process by getting rid of special characters, unused spaces, and redacted patient data. The removed personal information of the patients is represented by $xxxx$ in [9] and ___ in [16]. Furthermore, after the removal of all unnecessary characters, the remaining characters are converted to lowercase. The standardisation process from [20] was adapted for the [16,53] datasets while for the Indiana dataset, [54] was used along with the training, validation, and test split. In addition, the vocabulary for each of the datasets — generated from the entire text corpus including findings, impressions, and free text — was determined individually, and for [16] the words with an occurrence of ≤ 3 were dropped from the reports. This resulted in 9396 unique words for [16] and 5615 for [53]. The findings and impressions sections are combined for [16], and if either is missing, the other is used. The free text section is utilised if either or both are missing. A local dataset has also been acquired from Rawalpindi, Pakistan from January 2020 to November 2020 and has been used for testing. This dataset contains 1054 frontal chest X-rays which is the same number of patients. The resolution of the scans is 3072 × 3072 pixels and they have been captured using 'FXRD-1717NB' machine. In addition to the local dataset, a test set of 100 images from BRAX [52] dataset has been manually annotated for report generation by local radiologists as well.

3.2. Proposed framework

Our proposed framework generates a CXR report that provides the findings from a single CXR taken either from the Anterior-Posterior or Posterior-Anterior viewing position. An encoder and a decoder are employed in the report generation module. The training process also makes use of fine-tuning of a foundation model that is then used to perform Knowledge Distillation using the encoder. The sections below describe the different parts of the framework in detail. An overview of the entire framework is provided in Fig. 1.

3.2.1. Foundation model and pre-training

Making use of the same approach that [36] used to train Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT), we pre-train BioBERT on the CXR reports using the Masked Language Modelling approach which was inspired by [55]. Although BioBERT is already pre-trained on a medical corpus that is suited for the medical domain, this training corpus lacks CXR reports. Therefore, pre-training it on the CXR report corpus should improve the performance for the downstream task of knowledge distillation.

We make use of pre-trained $\text{BioBERT}_{\text{BASE}}$ having 137 million parameters where approximately 108 million parameters are trainable. The model contains 12 layers with the size of the latent vector from the last hidden layer being 768 and it also incorporates a token classification layer. This model is trained with the maximum number of 512 tokens and 35% of the tokens are masked during this process. The learning rate is kept as 5×10^{-5} with a weight decay of 0.01. As the result of applying a softmax function to the full vocabulary, the model outputs the probability of the masked token. Fig. 2 shows a part of a report is masked for training the foundation model.

3.2.2. Knowledge distillation

In order to further improve the performance of the proposed framework, features-based knowledge distillation was adopted. Using a similar approach to [56], the visual encoder in the report generation module acts as the student model while BioBERT which has been pre-trained on a particular dataset reports acts as the teacher model.

In contrast to the masked language model training of BioBERT [36], the complete associated text is tokenised and no token is masked. Concurrently, the corresponding image is passed through the ViT [28] to obtain the image embeddings. The last hidden state from BioBERT and the image embeddings are then used to calculate the Kullback–Leibler [57] divergence loss given by Eq. (3).

$$L_{KL}(y_{\text{pred}}, y_{\text{true}}) = y_{\text{true}} \cdot \log\left(\frac{y_{\text{pred}}}{y_{\text{true}}}\right) = y_{\text{true}} \cdot \left(\log y_{\text{true}} - \log y_{\text{pred}}\right) \quad (3)$$

$$L_{KL}(y_{pred}, y_{true}) = Softmax(y_{true}) \cdot (log(Softmax(y_{true})) - log(Softmax(y_{pred}))) \quad (4)$$

To ensure that the KL divergence is mathematically correct, a log operation is applied to the embeddings from each model preceded by a Softmax operation as shown in Eq. (4). The KL divergence loss makes sure that the hidden states, which are regarded as features, allow the reduction of the difference between the teacher's and the student's model activation, where the embeddings from the former are regarded

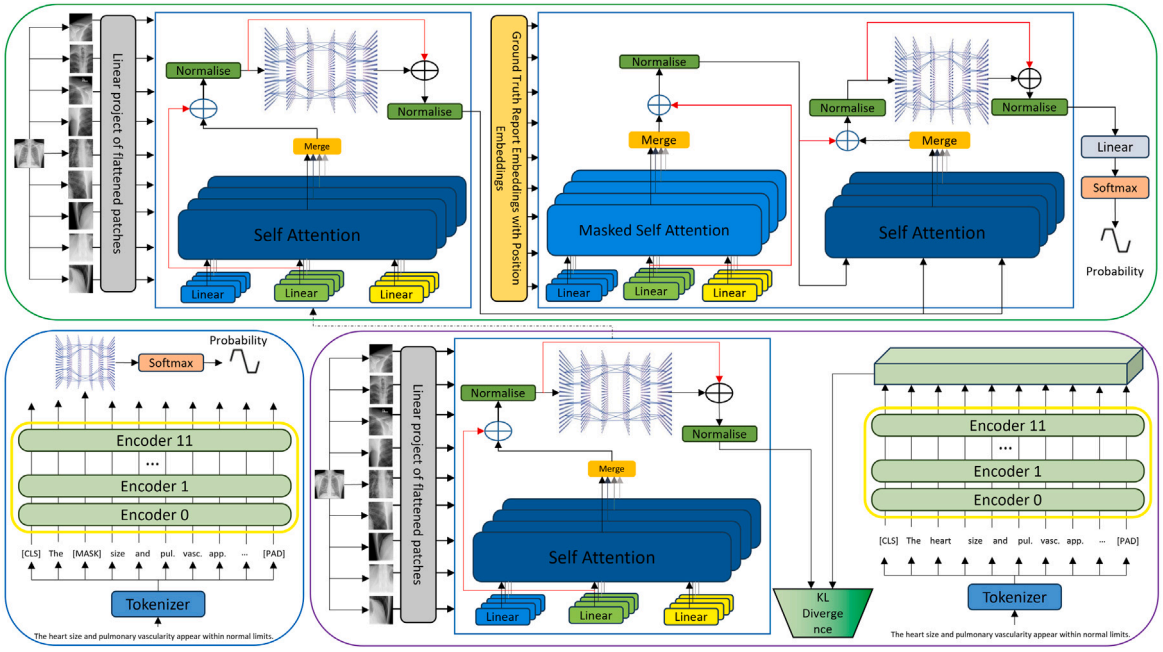

Fig. 1. Proposed report generation architecture. In the first step, the teacher BioBERT model (Bottom Left) is fine-tuned on the target dataset. Knowledge distillation (Bottom Right) utilises the trained teacher model to reduce the loss in the ViT student model. Finally, this trained ViT is used as the Encoder in the report generation module (Top Centre) in combination with the decoder. The initial normalisation layer in the ViT architecture is omitted for brevity.

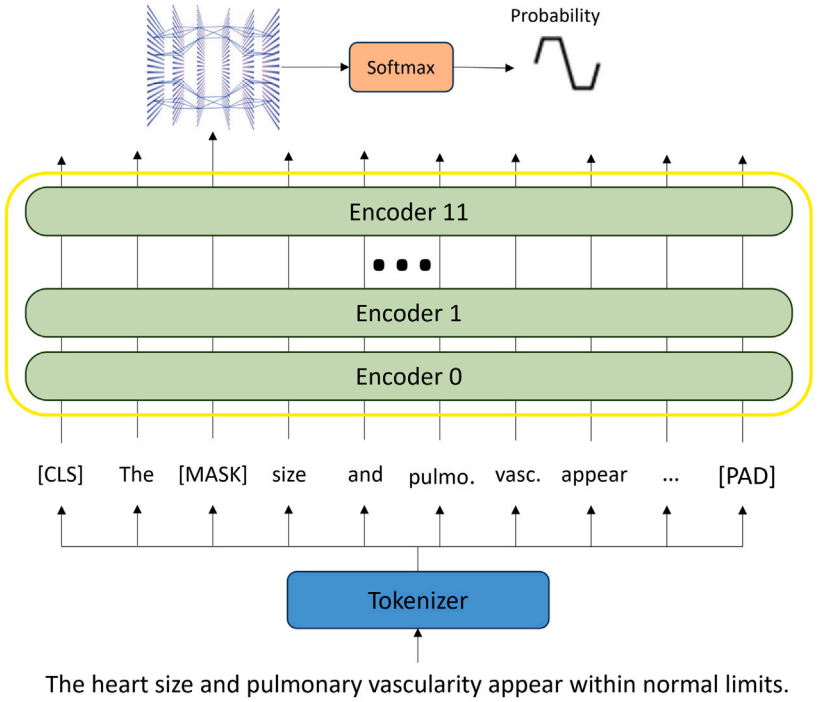

Fig. 2. Part of the CXR report is masked to train the foundation model in an unsupervised manner. The trained foundation model is then used as the teacher in the Knowledge Distillation step.

as the target while those from the latter are regarded as input. Fig. 3 shows the training time configuration of the Teacher-Student model. The teacher model is removed during inference and the trained student model is used as the encoder in training the encoder–decoder module in the report generation framework.

The encoder model ViT [28], which behaves as the student, is trained using the foundation model serving as the teacher, once the foundation model has been trained using the processed dataset. The KL divergence [57] uses the embeddings generated by the final hidden layer of each of these models, which have the same dimensions. The encoder for training the report-generation sub-module is the student model from the training with the minimum loss.

For the student model, We make use of ViT-Base [28] architecture consisting of 12 layers and 12 heads and containing approximately 86

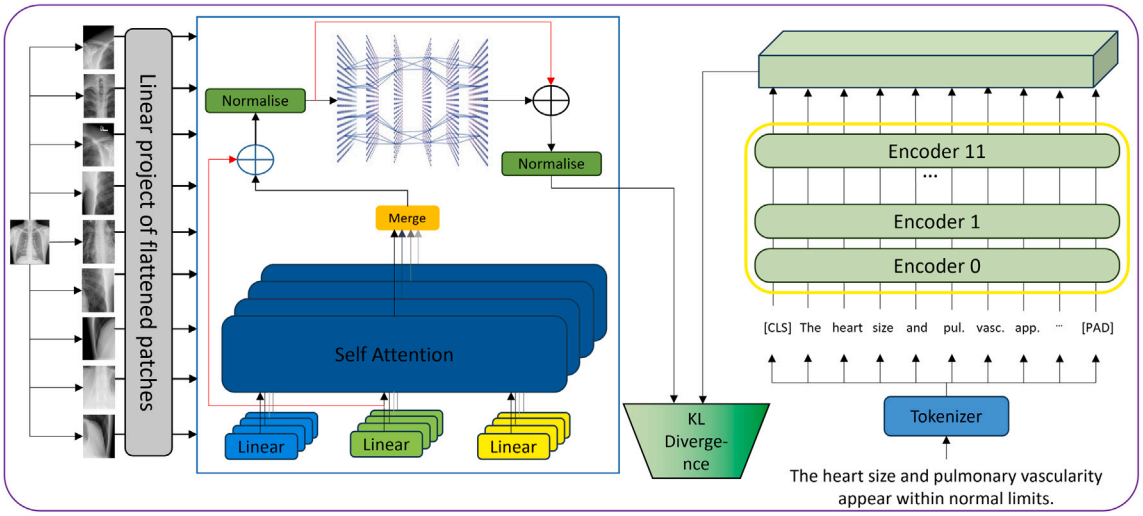

Fig. 3. For a CXR-report pair, the latent space features are generated by both the ViT and the BioBERT which are used to as features to increase the similarity between the two by reducing the loss using KL Divergence.

million parameters. The input image size for the ViT is set to 224×224 and 16 patches per image are used. The latent vector's dimensionality from ViT's final hidden layer mirrors that of $\text{BioBERT}_{\text{BASE}}$ at 768, eliminating the need for dimensional adjustments. The collective parameter count for the combined ViT and $\text{BioBERT}_{\text{BASE}}$ models reaches approximately 194 million, indicating a substantial capacity for learning and adaptation. 5×10^{-4} is the learning rate that is used for training with no weight decay and a batch size of 128.

3.2.3. Training of the downstream task: Report generation

Using the distilled student ViT-Base from the previous training step, the report generation sub-module is trained. ViT acts as the encoder that extracts the visual features in conjunction with MiniLM [58] which acts as the decoder that takes the latent space vectors from the encoder and uses it as the hidden state. The transformer model in [58] has itself been trained using distillation by using a larger teacher model based on BERT [35] achieved by using Kullback–Leibler Divergence [57] between the scaled dot products of Query-Key and Value-Value from the last self-attention module of both the teacher and the student model. The MiniLM that is employed in this configuration has 12 layers, a hidden state dimension of 384, and only 41 million parameters, bringing the total number of parameters for the sub-module close to 127 million. The tokenizer from MiniLM is the one used for tokenization with a varying maximum length for different datasets. AdamW [59] optimiser is used with a learning rate 5×10^{-5} and the training batch size is varied between 2 to 8. Fig. 4 shows the configuration of the report generation module.

4. Experimental setup

The proposed framework was trained and evaluated on three different datasets: Indiana University Chest X-ray [9], MIMIC [16] and MIMIC-PRO [53]. To gauge the performance of the proposed framework, every model in the framework was trained at least three times, and the best-performing model was retained. The details of the hyperparameters have been added in Table 4. The models were trained using PyTorch in Python on two systems; one with 64 GB RAM and two Nvidia RTX 2070 GPUs and the other with 128 GB RAM and a single Nvidia RTX 3090 Ti GPU. The performance metrics that have been calculated for this framework include both semantic similarity and lexical similarity. The BLEU score [60] and ROUGE-L [27] are used for lexical similarity while BERTScore [61] is used for semantic similarity. RadCliQ [62] provides a combination of both lexical and semantic similarity. The computation of the aforementioned and other metrics has been done using [62]. For BLEU, ROUGE-L, RadGraph F1 and s_{emb}, a higher score (↑) is better while for RadCliQ a lower score (↓) is better.

In order to verify the utility of Knowledge Distillation and pre-training using a large dataset, two ablation studies were performed. This section lists the findings of those studies.

4.1. Ablation study for knowledge distillation

The strength of the proposed framework stems from Knowledge Distillation where what the teacher model has learnt can be used to improve the performance of a smaller student model despite their architectural differences and distinct purposes. In order to evaluate the effect of Knowledge Distillation, using the IU dataset [9], the proposed framework was trained with and without utilising this training step. Table 1 shows the effects of using this strategy. The BERTScore and BLEU-1 scores both increased with the use of knowledge distillation, going from 0.4725 to 0.5482 and 0.4449 to 0.4848, respectively, as seen in Table 1.

4.2. Ablation study for pre-training

In order to utilise a large amount of data and account for the lower number of samples used in the Indiana University dataset, two experiments were conducted. In the first experiment (*Exp 1*), the foundation model was trained on the larger MIMIC dataset while the knowledge distillation and the report generation sub-module were trained using the IU dataset. In the other experiment (*Exp 2*), only the report generation sub-module was trained on the IU dataset, and the training of the foundation model and knowledge distillation were both performed using the MIMIC dataset. As before, the split given in [54] has been used.

While the difference between the number of samples in MIMIC and MIMIC-PRO is only around 16%, the difference in unique words in the text corpus between the two datasets is nearly 40%. Therefore, keeping this in view, in a similar manner to the pre-training performed on the Indiana University dataset, the same two experiments are conducted as before. Table 2 shows the results of this approach.

From Table 2, it can be seen that both types of pre-training result in improvement of scores across all evaluation metrics. Just using the foundation model trained on MIMIC dataset and re-training the

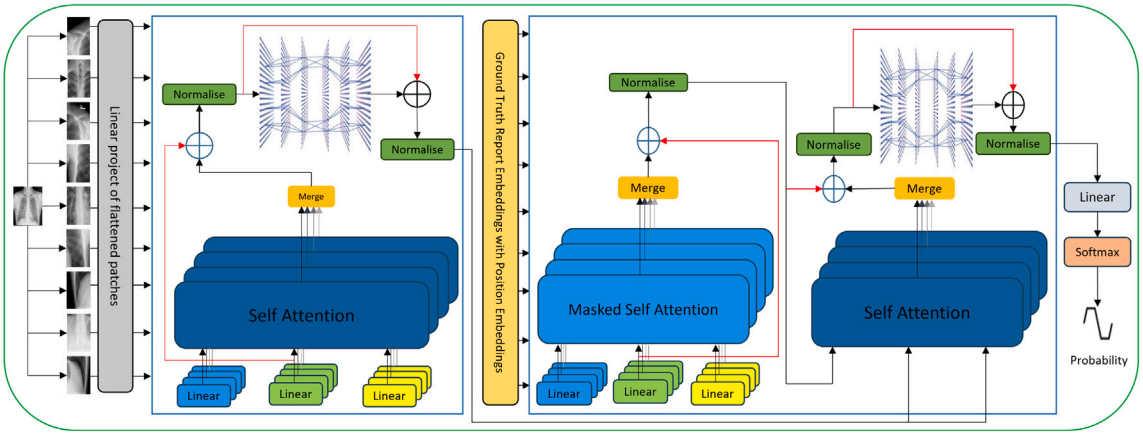

Fig. 4. Multiple layers are used in both the encoder and the decoder with each layer containing multiple attention heads. Each successive layer in the encoder receives the input from the previous layer while the last layer passes the output to each decoder layer. The figure illustrates the last encoder and decoder layer.

Table 1
Effects of using knowledge distillation in training by using the foundation model as teacher and ViT as the student model on the Indiana University [9] data set.

Technique	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	RadCliQ ↓	BERTScore ↑
Without knowledge distillation	0.4449	0.2921	0.2070	0.1518	0.3562	2.767	0.4725
With knowledge distillation	0.4848	0.3183	0.2280	0.1670	0.3573	2.3954	0.5482

Table 2
Performance of proposed framework on Indiana University [9] and MIMIC-PRO [53] data sets with pre-training on MIMIC [16] data set for different sub-modules.

Technique	Dataset	Data partition	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	RadCliQ ↓	BERT-score ↑
Exp 1	[9]	–	0.5572	0.4032	0.3028	0.2172	0.4211	2.1943	0.5800
	[53]	I (AP & PA)	0.1181	0.0602	0.0329	0.0192	0.118	3.9468	0.2243
		I	0.1157	0.0581	0.0314	0.0180	0.1139	3.9590	0.2245
Exp 2	[9]	–	0.5476	0.3397	0.3450	0.2218	0.4291	2.1600	0.5730
	[53]	I (AP & PA)	0.1098	0.0555	0.0303	0.0172	0.1130	3.9393	0.2235
		I	0.1081	0.0542	0.0294	0.0165	0.1099	3.9480	0.2245

Knowledge Distillation module on IU data set results in higher BLEU-1, BLEU-2, and BERTScore as compared to the other methodology. Conversely, BLEU-3, BLEU-4, ROUGE, and RadCliQ show improvement when the ViT distilled using MIMIC foundation model is used. Improvement in these metrics can be observed for the MIMIC-PRO data set where the scores improve as a result of pre-training which can be attributed to the fact that while the difference between the number of samples in MIMIC and MIMIC-PRO is only around 16%, the difference in unique words in the text corpus between the two datasets is nearly 40%.

5. Results

The following section details the results for different datasets that were used for evaluation.

5.1. MIMIC dataset evaluation

For the MIMIC dataset, the Findings and the Impressions sections were combined to form the reports for the chest X-rays. The use of this combination provides a two-fold advantage: it not only increases the length of the individual report but also increases the number of samples that would have been otherwise discarded due to the missing Findings or Impressions sections. Furthermore, only PA and AP views were kept resulting in 237,887 training samples and 1958 validation samples. The foundation model and the knowledge distillation sub-module were trained using the training and evaluated using the test samples. Table 3 shows the results of the proposed framework on only Findings, only Impressions, and Findings combined with Impressions.

Along with the quantitative results shown in Table 3, the qualitative results from randomly selected samples from the MIMIC dataset are shown in Table 5corresponding to results on Findings, Impressions, and combined Findings and Impressions sections. The differences are shown in red while the other colours are used for similarities. The order of the sentences generated varies from that of the ground truth. Furthermore, different verbiage is employed to express the same concept.

A comparison with state-of-the-art techniques is undertaken to assess the efficacy of the proposed report generation methodology. Table 7 shows the performance of the proposed methodology against different techniques for the under consideration evaluation metrics for just the Findings section of the reports.

Similar to Tables 7, 8 shows the performance of the proposed methodology against different techniques the Impressions section which is a brief summary of the findings section. Table 9 shows the

Table 3
Performance of proposed on MIMIC [16] dataset. The framework is trained on a combination of Findings and Impressions sections while the results are computed on different sections of the report. F represents *Findings*, F+I represents *Findings & Impressions* while I represents *Impressions*. The absence of (AP & PA) denotes that all samples were used for testing.

Data partition	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	RadCliQ ↓	BERTScore ↑
F (AP & PA)	**0.3104**	**0.1801**	**0.1167**	**0.0818**	**0.2213**	**3.6145**	0.3484
F	0.3048	0.1762	0.1145	0.0803	0.2197	3.6360	**0.3507**
F + I (AP & PA)	0.3007	0.1720	0.1099	0.0763	0.2142	3.7210	0.3192
F + I	0.2959	0.1689	0.1084	0.0725	0.2134	3.745	0.3207
I (AP & PA)	0.1625	0.0839	0.0488	0.0312	0.1404	4.1372	0.2405
I	0.1512	0.0767	0.044	0.0276	0.1344	4.2053	0.2344

Table 4
Hyperparameters of different stages of the report generation framework.

Stage	Models	Layers	Parameters (Millions)	Latent dimensions	Tokens	Learning rate	Learning scheduler	Loss function	Activation	Image size	Input dimensions	Optimiser	Batch size
Foundation Model	BioBERT	12	137, 108	768	512	5×10^{-5}	Constant	Cross-entropy	Softmax	–	–	AdamW	4
Knowledge Distillation	BioBERT ViT Base	12,12	194 (108, 86)	768	512	5×10^{-4}		KL Divergence	–	224	3		128
Report Generation	ViT Base MiniLM	1212	127 (86, 41)	768, 384	512, 50/80	5×10^{-5}		Cross-entropy	GELU				2 to 8

Table 5
The text in red represents portions of the text that are absent in the generated report for the MIMIC [16] data set. Conversely, the other coloured sentences demonstrate the similarities between the ground truth and the generated report. The model is unable to correctly identify the distance of different tubes in the image. The combination of Findings and Impression sections not only improves the number of unique words in vocabulary but also improves the quality of the generated reports as well. Best viewed in colour.

	Findings	Impressions	Findings + Impressions
Actual	pa and lateral views of the chest provided . lung volumes are somewhat low. allowing for this there is no focal consolidation effusion or pneumothorax. the cardiomediastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.	as compared to the previous radiograph the patient has been intubated. the tip of the endotracheal tube projects approximately 48 cm above the carinal. the course of the nasogastric tube is unremarkable the tip projects over the middle parts of the stomach. the right internal jugular venous introduction sheet is unchanged . increasing atelectasis at the left lung bases. unchanged appearance of the right lung.	no focal consolidation is seen there is no pleural effusion or pneumothorax. the cardiac and mediastinal silhouettes are unremarkable. no acute cardiopulmonary process.
Predicted	pa and lateral views of the chest provided. there is no focal consolidation effusion or pneumothorax. the cardiomediastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen. no acute intrathoracic process.	as compared to the previous radiograph the patient has been intubated. the tip of the endotracheal tube projects approximately 4 cm above the carina. the patient has also received a nasogastric tube. the course of the tube is unremarkable the tip of the tube is not included on the image. the right internal jugular vein catheter is	the lungs are clear without focal consolidation. no pleural effusion or pneumothorax is seen. the cardiac and mediastinal silhouettes are unremarkable. no acute cardiopulmonary process.
Actual	frontal and lateral views of the chest demonstrate low lung volumes which accentuate bronchovascular markings. there is no focal consolidation pleural effusion or pneumothorax. the hilar and mediastinal silhouettes are unremarkable. heart size is top normal. partially imaged upper abdomen demonstrates prominent air-filled large bowel loops.	in comparison with the study of there again are bilateral pleural effusions with compressive atelectasis much more prominent on the left. continued enlargement of the cardiac silhouette with mild elevation of pulmonary venous pressure . pacer leads are unchanged.	cardiac mediastinal and hilar contours are normal. pulmonary vasculature is normal. no focal consolidation pleural effusion or pneumothorax is present. there are no acute osseous abnormalities. no acute cardiopulmonary process.
Predicted	frontal and lateral views of the chest demonstrate low lung volumes which accentuate bronchovascular markings. there is no pleural effusion focal consolidation or pneumothorax. hilar and mediastinal silhouettes are unremarkable. heart size is normal. there is no pulmonary edema. partially imaged upper abdomen is unremarkable. a	in comparison with the study of there is little overall change. again there is substantial enlargement of the cardiac silhouette with pulmonary edema and bilateral pleural effusions with compressive basilar atelectasis more prominent on the left. pacer device remains in place.	heart size is normal. the mediastinal and hilar contours are normal. the pulmonary vasculature is normal. lungs are clear. no pleural effusion or pneumothorax is seen. there are no acute osseous abnormalities. no acute cardiopulmonary abnormality.

comparison of the combined Findings and Impression sections with the literature. Only a few studies have chosen to use this combined strategy.

5.2. Indiana university dataset evaluation

Keeping the same approach as that used for MIMIC, the proposed framework was also trained on the Indiana University chest X-rays and reports. However, instead of combining the Findings and the Impressions, only Findings were used in training. In addition, only PA and AP views of the CXRs are kept while others are discarded. Images that do not contain the findings section are also discarded. 3200 training samples, 500 validation samples, and 300 test samples were produced as a result. Table 10 shows the results of the proposed framework with the inclusion of Knowledge Distillation.

Table 6
For the Indiana dataset [9], the text in red shows that even when different verbiage is used in the generated report, the same meaning is conveyed. For MIMIC-PRO [53] and BRAX [52], the text highlighted in red represents portions of the text that are absent in the generated report. Conversely, the other highlighted sentences demonstrate the similarities between the ground truth and the generated report. The reports for BARX [52] subset have been generated without any retraining or fine-tuning. For all datasets, it can be observed that different verbiage has been used for the same statement in the actual and the predicted reports.

	Indiana [9]	MIMIC-PRO [53]	BRAX [52]
Actual	the cardiomediastinal silhouette is within normal limits for size and contour the lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax no acute bone abnormality	There are small bilateral pleural effusions, which combine with signs of mild pulm- onary edema. Findings suggest small bilateral pleural effusions, pulmonary vascular conges- tion and bibasilar atelectasis.	lungs are clear heart size appear normal normal study .
Predicted	the cardiomediastinal silhouette is within normal limits for size and contour the lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax no acute bony abnormality	mild pulmonary edema with small bilateral pleural effusions and bibasilar atelectasis.	no acute cardiopulmonary pathology.
Actual	the cardiomediastinal silhouette is within normal limits for size and contour the lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax no acute osseus abnormality	Moderate congestive heart failure with mo- derate pulmonary edema, small bilateral pl- eural effusions, and bibasilar airspace opac- ities likely reflective of atelectasis. Findings compatible with acute pulmonary edema with possible layering pleural effusions.	heart size is enlarged with well defined margins may suggest moderate pericardial effusion however need correlation with echocardiogram lungs are clear cp angles are sharp .
Predicted	the cardiomediastinal silhouette is within normal limits for size and contour the lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax osseous structures are intact	bilateral lower lobe opacities and enlargement of the cardiac silhouette consistent with mod- erate pulmonary edema. probable small bilateral pleural effusions and bibasilar atelectasis.	massive cardiomegaly with history of pericardial effusion. echocardiogram is recommended for further evaluation.

Table 7
Performance of the proposed framework on the findings section against different techniques in literature for MIMIC [16] data set. The absence of (AP & PA) denotes that all samples were used for testing.

Technique	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	RadCliQ ↓	BERTScore ↑
Vinyals et al. [63]	0.256	0.157	0.102	0.070	0.249	–	–
Xu et al. [64]	0.304	0.177	0.112	0.077	0.249	–	–
Anderson et al. [65]	0.280	0.169	0.108	0.074	0.250	–	–
Liu et al. [66]	0.334	0.217	**0.140**	**0.097**	**0.281**	–	–
Lu et al. [67]	0.302	0.189	0.122	0.082	0.259	–	–
Rennie et al. [68]	0.314	0.199	0.126	0.087	0.265	–	–
Krause et al. [69]	0.321	0.203	0.128	0.089	0.266	–	–
Harzig et al. [70]	0.328	0.204	0.127	0.090	0.267	–	–
Jing et al. [6]	0.329	0.206	0.133	0.095	0.273	–	–
Huang et al. [71]	**0.337**	0.211	0.136	0.095	0.274	–	–
Jeong et al. [72]	–	**0.220**	–	–	–	**3.585**	**0.353**
Nicolson et al [73]	–	0.196	–	–	–	3.617	0.347
Proposed (PA & AP)	0.310	0.180	0.116	0.081	0.221	3.614	0.348
Proposed	0.304	0.176	0.114	0.080	0.219	3.636	0.350

Table 8
Performance of the proposed framework on the impressions section against different techniques in literature for MIMIC [16] data set. The absence of (AP & PA) denotes that all samples were used for testing.

Technique	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	RadCliQ ↓	BERTScore ↑
Ramesh et al. [53]	–	–	–	–	–	–	0.229
Jeong et al. [72]	–	0.084	–	–	–	**3.781**	**0.287**
Endo et al. [74]	–	0.055	–	–	–	4.121	0.193
Li et al. [75]	–	0.030	–	–	–	4.313	0.190
Miura et al. [76]	–	**0.087**	–	–	–	–	0.227
Chen et al. [12]	–	0.059	–	–	–	–	0.186
Yan et al. [77]	–	0.064	–	–	–	–	0.188
Nicolson et al. [73]	–	0.066	–	–	–	–	0.192
Proposed (PA & AP)	0.162	0.083	0.048	0.031	0.140	4.137	0.240
Proposed	0.151	0.076	0.044	0.027	0.134	4.205	0.234

Table 9
Performance of the proposed framework on the combined Findings and Impressions section against different techniques in literature for MIMIC [16] data set. The absence of (AP & PA) denotes that all samples were used for testing.

Technique	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	RadCliQ ↓	BERTScore ↑
Jeong et a. [72]	–	0.161	–	–	–	3.835	0.287
Yan et al. [77]	–	0.144	–	–	–	3.986	0.275
Chen et al. [41]	–	0.137	–	–	–	4.051	0.271
Proposed (PA & AP)	0.300	**0.172**	0.109	0.076	0.214	**3.721**	0.319
Proposed	0.151	0.076	0.044	0.027	0.134	3.745	**0.320**

Table 10
Performance of the proposed framework on the findings for Indiana University [9] data set.

BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	RadCliQ ↓	BERTScore ↑
0.4848	0.3183	0.2280	0.1670	0.3562	2.7670	0.4725

Table 11
Performance of the proposed framework on the findings section against different techniques in literature for Indiana University [9] data set.

Technique	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	RadCliQ ↓	BERTScore ↑
Liu et al. [19]	0.483	0.315	0.224	0.168	0.376	–	–
Liu et al. [78]	0.492	0.314	0.222	0.169	0.381	–	–
Liu et al. [66]	0.473	0.305	0.217	0.162	0.378	–	–
You et al. [79]	0.484	0.313	0.255	0.173	0.379	–	–
Nooralahzadeh et al. [80]	0.486	0.317	0.232	0.173	0.390	–	–
Yang et al. [20]	0.496	0.317	0.232	0.173	0.390	–	–
Kale et al. [81]	0.423	0.256	0.194	0.165	**0.444**	–	–
Qin et al. [82]	0.492	0.321	0.235	0.181	0.384	–	–
Wang et al. [83]	0.497	0.357	0.279	**0.225**	0.414	–	–
Proposed	**0.557**	**0.403**	**0.302**	0.217	0.421	2.194	0.58

Table 12
Performance of proposed on MIMIC [16] dataset. The framework is trained on a combination of Findings and Impressions sections while the results are computed on different sections of the report. I represents *Impressions* and the absence of (AP & PA) denotes that all samples were used for testing.

Data partition	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	RadCliQ ↓	BERTScore ↑
I (AP & PA)	0.0988	0.0496	0.0274	0.0162	0.1064	4.0684	0.2074
I	0.0972	0.0481	0.0262	0.0154	0.1034	4.0766	0.2085

A comparison with state-of-the-art techniques is undertaken to assess the efficacy of the proposed report generation methodology. Table 11 shows the performance of the proposed methodology against different techniques for the under-consideration evaluation metrics.

Along with the quantitative results and comparison with different techniques in literature, the qualitative results from randomly selected samples from the IU dataset are shown in Table 6 where the differences are shown in red.

5.3. MIMIC-PRO dataset evaluation

For MIMIC-PRO, only the Impression section has been rewritten to remove references to previous reports which results in an overall reduced length of the report [53]. In a similar fashion to the other data sets, only the PA and AP views were used in this dataset for training. There are 2065 samples in the test dataset which are either PA or AP and 2188 samples if all views are considered, whereas 199,396 samples make up the training dataset. Table 12 shows the results of the proposed framework.

Along with the quantitative results and comparison with different techniques in literature, the qualitative results from randomly selected samples from the MIMIC-PRO dataset are shown in Table 6.

A comparison with state-of-the-art techniques is undertaken to assess the efficacy of the proposed report generation methodology. Table 13 shows the performance of the proposed methodology against different techniques for the under-consideration evaluation metrics.

5.4. BRAX subset evaluation

Just like with the local dataset, the trained models were evaluated on the BRAX subset of 100 images, for which reports were provided by a radiologist. The models were used without any retraining or fine-tuning on the target dataset to assess its performance. Table 14 shows the metrics under consideration for three models that were trained on [9,16], and [53] respectively.

From Table 14, it can be seen that the model trained on the Indiana dataset outperforms the models trained on the other two datasets in all metrics except RadCliQ. This difference in performance can be explained by the overwhelming ratio of normal samples in the Indiana dataset. Furthermore, the model trained on MIMIC outperforms both models for RadCliQ by a significant margin. For BERTScore, the difference between models trained on Indiana and MIMIC dataset is just 1.16% which is significantly less than the difference observed in BLEU-1, BLEU-2, and BLEU-3 scores. Similarly, for the BLEU-4 score, the difference between the model trained on the Indiana dataset and the model trained on MIMIC dataset is far less.

Another thing to note is that the variation in the generated reports is least from the model that has been trained on the Indiana dataset which is due to less variability of the Indiana corpus itself. The reports that have been generated by the model trained on MIMIC have a tendency to have hallucinatory references to prior examinations as such reports are quite common in the dataset corpus. This problem is largely solved by the use of model trained on MIMIC-PRO which does not contain such references. However, as the MIMIC-PRO text corpus focuses on just the *Impressions* portion of the report, the generated reports tend to be relatively brief and follow the structure of the *Impressions* section instead of the *Findings* section which more closely resembles the structure of BRAX reports. Table 6 shows the radiologist report and the reports generated by MIMIC-PRO [53] model.

5.5. Local dataset evaluation

Without any fine-tuning or retraining, the trained models were used on the locally gathered dataset to assess the generalisability of the proposed framework. The models trained on both the MIMIC [16] and Indiana [9] dataset were used for the local dataset where the model trained on IU dataset performed better than the one trained on MIMIC. Using this trained model, a BLEU-1 score of 0.3827 and a BERTScore of 0.4392 was achieved on the local dataset as shown in Table 15.

6. Discussion

One of the limitations of the framework presented here lies in the hardware and the time requirements for training the complete framework as it contains several sub-modules. However, at the time of inference, the architecture is simplified which reduces the number of

Table 13
Performance of the proposed framework on the against different techniques in literature for MIMIC-PRO [53] data set. For the different techniques used in the proposed methodology, the absence of (AP & PA) denotes that all samples were used for testing.

Technique	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	s_{emb} ↑	RadGraph F1 ↑	BERTScore ↑
Ramesh et al. (Report Retrieval) [53]	–	–	–	–	0.3601	0.0925	0.2160
Ramesh et al. (Sentence Retrieval, k = 1) [53]	–	–	–	–	0.3967	0.0864	0.2159
Ramesh et al. (Sentence Retrieval, k = 2) [53]	–	–	–	–	0.3859	0.1056	**0.2351**
Ramesh et al. (Sentence Retrieval, =3) [53]	–	–	–	–	0.3779	0.1118	0.2254
Proposed (AP & PA)	0.0988	0.0496	0.0274	0.0162	0.3739	0.0983	0.2074
Proposed	0.0972	0.0481	0.0262	0.0154	0.3735	0.0904	0.2085
Proposed (AP & PA)[Exp 1]	0.1181	0.0602	0.0329	0.0192	0.4000	**0.1165**	0.2243
Proposed [Exp 1]	0.1157	0.0581	0.0314	0.0180	0.3988	0.1083	0.2245
Proposed (AP & PA)[Exp 2]	0.1098	0.0555	0.0303	0.0172	**0.4069**	0.1125	0.2235
Proposed [Exp 2]	0.1081	0.0542	0.0294	0.0165	0.4059	0.1048	0.2245

Table 14
Performance of proposed framework on BRAX [52] subset. Models trained on three different datasets (Indiana University [9], MIMIC [16] and MIMIC-PRO [53]) were used.

Model trained on	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	RadCliQ ↓	BERTScore ↑
Indiana [9]	**0.1671**	**0.0816**	**0.0422**	**0.0169**	**0.1085**	4.3325	**0.2186**
MIMIC [16]	0.1295	0.0570	0.0261	0.0139	0.1057	**4.0717**	0.2070
MIMIC-PRO [53]	0.1013	0.0391	0.0129	0.0009	0.0716	4.2896	0.1875

Table 15
Performance of proposed framework on Local data set. The framework trained on the Indiana University [9] data set was used.

BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	RadCliQ ↓	BERTScore ↑
0.3827	0.2367	0.1624	0.1153	0.2745	2.5256	0.4392

parameters required. This, in addition to the proposed framework not relying on report retrieval, makes the framework a relatively simple approach at the time of inference.

Due to most of the lungs having a normal presentation in the MIMIC [16], BRAX [52], MIMIC-PRO [53] and the Indiana [9] datasets and the findings section following a systematic structure, most of the associated reports are repetitive and similar thus lacking information regarding different abnormalities. While this allows the machine learning approaches to learn better representations of normal images and is in line with the natural distribution of the samples in most of the CXR datasets [9,16,52,53], critical findings can be missed in the report due to the lack of abnormal samples. This issue also plagues the proposed framework here as evidenced by the sample generated reports present in Table 5. One attempt to address this is the use of only abnormal radiology reports in the MIMIC dataset to train a model as proposed by [84]. To improve the model's ability to identify abnormalities in CXRs, it may be helpful to fine-tune a model that was trained on the complete MIMIC corpus using this aberrant-only dataset.

Knowledge Distillation is shown to improve the performance as evident by the improvement in performance on the Indiana dataset in Table 1. Furthermore, it has been demonstrated that pre-training for report generation enhances performance on smaller datasets relative to the bigger pre-training dataset, in addition to Knowledge Distillation. The performance for both [9,53] datasets increases — as evident by Table 2 — as a result of pre-training on [16]. The larger corpus of the MIMIC dataset, both in terms of samples and vocabulary allows the framework to learn better representations.

The use of Knowledge Distillation allows for a model with a relatively small size to perform reasonably well. In contrast to our proposed strategy, the architecture proposed in [74] makes use of a report retrieval strategy which necessitates that the embeddings from a large CXR reports corpus are extracted and then compared with the embeddings extracted from the test image. This results in large memory and time requirements which can be reduced by a compression technique [74]. Even though the proposed framework takes less computational resources [85] than the retrieval approach [74], it is still able

to outperform the in s_{emb} and RadGraph F1 score. Even in BERTScore, the difference is a just little over 1% as can be seen in Table 13. Another thing to consider is the report template discrepancy that can be found in the gathered local dataset and the international datasets that were used to train the framework. The difference in the style of how the findings are provided in a radiology report can impact the performance of automated frameworks. This difference can been in Table 14 for the BRAX reports created by local experts. The length of the reports in the different datasets can have greater impact on metrics like BLEU scores while metrics like BERTScore are affected less by it.

One more thing to consider is that the BLEU-1 score captures semantic similarity. For reports that are closer in content, the BLEU-1 score is also high. However, this is affected by the ratio of the normal reports in the testing data and the structure of the findings section that is followed for a dataset. During experimentation, some models, while underperforming on the BLEU-1 metric (approximate BLEU-1 score of 0.41 on Indiana [9] dataset) had a greater number of unique generated reports with more variability in the verbiage of the reports. This variability resulted in a lower BLEU-1 score, despite having more unique words and less repetition.

While pre-training has been performed for the Foundation model to accommodate for the lack of CXR reports in the original training corpus, pre-trained weights have been used in case of BioBERT and ViT while MiniLM has been trained from scratch. Training these larger models from scratch on the CXR corpus might improve the overall performance of the proposed framework. However, such a change to the training strategy will increase the training time significantly.

7. Conclusion

We present a radiology report generation framework from a single perspective using a transformer encoder–decoder with features-based knowledge distillation and demonstrate reasonably good performance on all the datasets. We also demonstrate that using pre-training and knowledge distillation improves the report generation results which is evident from an increase in metrics such as the BLEU score. Our method also shows that it is possible to achieve variation in the generated reports while still doing well on the aforementioned measures. Masked language pre-training of the teacher language model with a particular dataset allows for better domain and dataset representation. The parameters of the student models are then updated based on loss calculated using the latent space features of the student and the teacher model. Generating an automated report that represents an accurate picture of the condition of the chest cavity is an important step in providing better care to patients.

CRediT authorship contribution statement

Asad Mansoor Khan: Writing – original draft, Validation, Software, Methodology, Data curation. **Mashood Mohammad Mohsan:** Writing – review & editing, Software, Methodology, Formal analysis, Data curation. **Muhammad Usman Akram:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Taimur Hassan:** Writing – review & editing, Validation, Project administration. **Sajid Gul Khawaja:** Writing – review & editing, Validation, Supervision. **Adil Qayyum:** Writing – review & editing, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] A.M. Tahir, M.E. Chowdhury, A. Khandakar, T. Rahman, Y. Qiblawey, U. Khurshid, S. Kiranyaz, N. Ibtehaz, M.S. Rahman, S. Al-Maadeed, et al., COVID-19 infection localization and severity grading from chest X-ray images, Comput. Biol. Med. 139 (2021) 105002.

[2] S. Manna, J. Wruble, S.Z. Maron, D. Toussie, N. Voutsinas, M. Finkelstein, M.A. Cedillo, J. Diamond, C. Eber, A. Jacobi, et al., COVID-19: a multimodality review of radiologic techniques, clinical utility, and imaging features, Radiol.: Cardiothorac. Imaging 2 (3) (2020).

[3] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z.A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li, et al., Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection, Radiology (2020).

[4] T. Dall, R. Reynolds, K. Jones, R. Chakrabarti, W. Iacobucci, The complexities of physician supply and demand: projections from 2017 to 2032, Assoc. Am. Med. Coll. (2019) 86.

[5] S. Raoof, D. Feigin, A. Sung, S. Raoof, L. Irugulpati, E.C. Rosenow III, Interpretation of plain chest roentgenogram, Chest 141 (2) (2012) 545–558.

[6] B. Jing, P. Xie, E. Xing, 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), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2577–2586, http://dx.doi.org/10.18653/v1/P18-1240, URL: https://aclanthology.org/P18-1240.

[7] Y. Hong, C.E. Kahn, Content analysis of reporting templates and free-text radiology reports, J. Digit. Imaging 26 (2013) 843–849.

[8] C.Y. Li, X. Liang, Z. Hu, E.P. Xing, Knowledge-driven encode, retrieve, paraphrase for medical image report generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6666–6673.

[9] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, C.J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, J. Am. Med. Inform. Assoc. 23 (2) (2016) 304–310.

[10] Y. Li, X. Liang, Z. Hu, E.P. Xing, Hybrid retrieval-generation reinforced agent for medical image report generation, Adv. Neural Inf. Process. Syst. 31 (2018).

[11] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[12] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating radiology reports via memory-driven transformer, 2020, arXiv preprint arXiv:2010.16056.

[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[14] M. Sirshar, M.F.K. Paracha, M.U. Akram, N.S. Alghamdi, S.Z.Y. Zaidi, T. Fatima, Attention based automated radiology report generation using CNN and LSTM, PLoS One 17 (1) (2022) e0262209.

[15] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[16] A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz, S. Horng, MIMIC-CXR-JPG-chest radiographs with structured labels (version 2.0. 0), PhysioNet (2019).

[17] F. Liu, C. You, X. Wu, S. Ge, X. Sun, et al., Auto-encoding knowledge graph for unsupervised medical report generation, Adv. Neural Inf. Process. Syst. 34 (2021) 16266–16279.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[19] F. Liu, X. Wu, S. Ge, W. Fan, Y. Zou, Exploring and distilling posterior and prior knowledge for radiology report generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13753–13762.

[20] S. Yang, X. Wu, S. Ge, S.K. Zhou, L. Xiao, Knowledge matters: Chest radiology report generation with general and specific knowledge, Med. Image Anal. 80 (2022) 102510.

[21] S. Jain, A. Agrawal, A. Saporta, S.Q. Truong, D.N. Duong, T. Bui, P. Chambon, Y. Zhang, M.P. Lungren, A.Y. Ng, et al., Radgraph: Extracting clinical entities and relations from radiology reports, 2021, arXiv preprint arXiv:2106.14463.

[22] E. Alsentzer, J.R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, 2019, arXiv preprint arXiv:1904.03323.

[23] P. Srinivasan, D. Thapar, A. Bhavsar, A. Nigam, Hierarchical x-ray report generation via pathology tags and multi head attention, in: Proceedings of the Asian Conference on Computer Vision, 2020.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 21–37.

[25] T. Zhong, W. Zhao, Y. Zhang, Y. Pan, P. Dong, Z. Jiang, X. Kui, Y. Shang, L. Yang, Y. Wei, et al., Chatradio-valuer: a chat large language model for generalizable radiology report generation based on multi-institution and multi-system data, 2023, arXiv preprint arXiv:2310.05242.

[26] Z. Liu, T. Zhong, Y. Li, Y. Zhang, Y. Pan, Z. Zhao, P. Dong, C. Cao, Y. Liu, P. Shu, et al., Evaluating large language models for radiology natural language processing, 2023, arXiv preprint arXiv:2307.13693.
[27] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.
[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[29] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, P. Vajda, Visual transformers: Token-based image representation and processing for computer vision, 2020, arXiv preprint arXiv:2006.03677.
[30] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.
[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[32] A.M. Dai, Q.V. Le, Semi-supervised sequence learning, Adv. Neural Inf. Process. Syst. 28 (2015).
[33] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving Language Understanding with Unsupervised Learning, Technical Report, OpenAI, 2018.
[34] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, 2018, arXiv preprint arXiv:1801.06146.
[35] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[36] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.
[37] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015, arXiv preprint arXiv:1503.02531.
[38] J. Ba, R. Caruana, Do deep nets really need to be deep? Adv. Neural Inf. Process. Syst. 27 (2014).
[39] S.W. Kim, H.-E. Kim, Transferring knowledge to smaller network with class-distance loss, 2017.
[40] Q. Ding, S. Wu, H. Sun, J. Guo, S.-T. Xia, Adaptive regularization of labels, 2019, arXiv preprint arXiv:1908.05474.
[41] D. Chen, J.-P. Mei, Y. Zhang, C. Wang, Z. Wang, Y. Feng, C. Chen, Cross-layer distillation with semantic calibration, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 7028–7036.
[42] G. Zhou, Y. Fan, R. Cui, W. Bian, X. Zhu, K. Gai, Rocket launching: A universal and efficient framework for training well-performing light net, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.
[43] X. Jin, B. Peng, Y. Wu, Y. Liu, J. Liu, D. Liang, J. Yan, X. Hu, Knowledge distillation via route constrained optimization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1345–1354.
[44] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, Z. Zhang, Correlation congruence for knowledge distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5007–5016.
[45] F. Tung, G. Mori, Similarity-preserving knowledge distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1365–1374.
[46] N. Passalis, M. Tzelepi, A. Tefas, Probabilistic knowledge transfer for lightweight deep representation learning, IEEE Trans. Neural Netw. Learn. Syst. 32 (5) (2020) 2030–2039.
[47] W. Park, D. Kim, Y. Lu, M. Cho, Relational knowledge distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3967–3976.
[48] J. Gou, B. Yu, S.J. Maybank, D. Tao, Knowledge distillation: A survey, Int. J. Comput. Vis. 129 (2021) 1789–1819.
[49] S. Sun, Y. Cheng, Z. Gan, J. Liu, Patient knowledge distillation for bert model compression, 2019, arXiv preprint arXiv:1908.09355.
[50] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, Q. Liu, Tinybert: Distilling bert for natural language understanding, 2019, arXiv preprint arXiv:1909.10351.
[51] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 590–597.
[52] E.P. Reis, J.P. de Paiva, M.C. da Silva, G.A. Ribeiro, V.F. Paiva, L. Bulgarelli, H.M. Lee, P.V. Santos, V.M. Brito, L.T. Amaral, et al., BRAX, Brazilian labeled chest x-ray dataset, Sci. Data 9 (1) (2022) 1–8.
[53] V. Ramesh, N.A. Chi, P. Rajpurkar, Improving radiology report generation systems by removing hallucinated references to non-existent priors, in: Machine Learning for Health, PMLR, 2022, pp. 456–473.
[54] M.M. Mohsan, M.U. Akram, G. Rasool, N.S. Alghamdi, M.A.A. Baqai, M. Abbas, Vision transformer and language model based radiology report generation, IEEE Access 11 (2022) 1814–1824.
[55] W.L. Taylor, "Cloze procedure": A new tool for measuring readability, Journal. Q. 30 (4) (1953) 415–433.
[56] I. Najdenkoska, X. Zhen, M. Worring, L. Shao, Uncertainty-aware report generation for chest X-rays by variational topic inference, Med. Image Anal. 82 (2022) 102603.
[57] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.
[58] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, Adv. Neural Inf. Process. Syst. 33 (2020) 5776–5788.
[59] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2017, arXiv preprint arXiv:1711.05101.
[60] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[61] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2019, arXiv preprint arXiv:1904.09675.
[62] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E.P. Reis, E.K.U.N. Fonseca, H.M.H. Lee, Z.S.H. Abad, A.Y. Ng, et al., Evaluating progress in automatic chest x-ray radiology report generation, MedRxiv (2022) 2022-2008.
[63] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.
[64] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, PMLR, 2015, pp. 2048–2057.
[65] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, 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, 2018, pp. 6077–6086.
[66] F. Liu, S. Ge, X. Wu, Competence-based multimodal curriculum learning for medical report generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3001–3012, http://dx.doi.org/10.18653/v1/2021.acl-long.234, URL: https://aclanthology.org/2021.acl-long.234.
[67] J. Lu, C. Xiong, D. Parikh, R. Socher, 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, 2017, pp. 375–383.
[68] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7008–7024.
[69] J. Krause, J. Johnson, R. Krishna, L. Fei-Fei, A hierarchical approach for generating descriptive image paragraphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 317–325.
[70] P. Harzig, Y.-Y. Chen, F. Chen, R. Lienhart, Addressing data bias problems for chest x-ray image report generation, 2019, arXiv preprint arXiv:1908.02123.
[71] X. Huang, F. Yan, W. Xu, M. Li, Multi-attention and incorporating background information model for chest x-ray image report generation, IEEE Access 7 (2019) 154808–154817.
[72] J. Jeong, K. Tian, A. Li, S. Hartung, F. Behzadi, J. Calle, D. Osayande, M. Pohlen, S. Adithan, P. Rajpurkar, Multimodal image-text matching improves retrieval-based chest X-ray report generation, 2023, arXiv preprint arXiv:2303.17579.
[73] A. Nicolson, J. Dowling, B. Koopman, Improving chest X-ray report generation by leveraging warm starting, Artif. Intell. Med. (2023) 102633.
[74] M. Endo, R. Krishnan, V. Krishna, A.Y. Ng, P. Rajpurkar, Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model, in: Machine Learning for Health, PMLR, 2021, pp. 209–219.
[75] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, PMLR, 2022, pp. 12888–12900.
[76] Y. Miura, Y. Zhang, E.B. Tsai, C.P. Langlotz, D. Jurafsky, Improving factual completeness and consistency of image-to-text radiology report generation, 2020, arXiv preprint arXiv:2010.10042.
[77] A. Yan, Z. He, X. Lu, J. Du, E. Chang, A. Gentili, J. McAuley, C.-N. Hsu, Weakly supervised contrastive learning for chest x-ray report generation, 2021, arXiv preprint arXiv:2109.12242.
[78] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, X. Sun, Contrastive attention for automatic chest X-ray report generation, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 269–280, http://dx.doi.org/10.18653/v1/2021.findings-acl.23, URL: https://aclanthology.org/2021.findings-acl.23.
[79] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, X. Wu, Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, Springer, 2021, pp. 72–82.
[80] F. Nooralahzadeh, N.P. Gonzalez, T. Frauenfelder, K. Fujimoto, M. Krauthammer, Progressive transformer-based generation of radiology reports, 2021, arXiv preprint arXiv:2102.09777.

[81] K. Kale, P. Bhattacharyya, M. Gune, A. Shetty, R. Lawyer, KGVL-BART: Knowledge graph augmented visual language BART for radiology report generation, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 3401–3411, http://dx.doi.org/10.18653/v1/2023.eacl-main.246, URL: https://aclanthology.org/2023.eacl-main.246.

[82] H. Qin, Y. Song, Reinforced cross-modal alignment for radiology report generation, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 448–458, http://dx.doi.org/10.18653/v1/2022.findings-acl.38, URL: https://aclanthology.org/2022.findings-acl.38.

[83] Y. Wang, Z. Lin, Z. Xu, H. Dong, J. Tian, J. Luo, Z. Shi, Y. Zhang, J. Fan, Z. He, Trust it or not: Confidence-guided automatic radiology report generation, 2021, arXiv preprint arXiv:2106.10887.

[84] J. Ni, C.-N. Hsu, A. Gentili, J. McAuley, Learning visual-semantic embeddings for reporting abnormal findings on chest X-rays, 2020, arXiv preprint arXiv:2010.02467.

[85] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S.C.H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, Adv. Neural Inf. Process. Syst. 34 (2021) 9694–9705.