
Citation

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37376413

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility

A DISSERTATION PRESENTED
BY
Katherine Tian
TO
THE DEPARTMENTS OF COMPUTER SCIENCE AND STATISTICS
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS WITH HONORS
IN THE JOINT CONCENTRATION OF
COMPUTER SCIENCE AND STATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2023

Abstract

Automatic radiology report generation has the potential to improve patient care and reduce diagnosis delays. Deep learning approaches have shown promising progress but are still not accurate enough for clinical deployment. In this thesis, we investigate and develop two approaches for report generation, one retrieval-based and one generation-based, both of which leverage deep vision-language pre-training.

Our retrieval-based method uses a multimodal encoder and contrastive loss to learn pre-trained radiology image and text representations, followed by a learned image-text matching similarity metric for retrieval. This method achieves state-of-the-art results on clinical accuracy and natural language metrics including CheXpert vector disease profile similarity and BLEU2 score. We also conduct an expert evaluation study on a subset of samples, where we collect radiologists’ error annotations on our generated reports, a baseline method’s generated reports, and human-written reports. The study confirms that our method improves significantly upon the baseline, and we will release the dataset of error annotations to aid future research into types of generated report errors and alignment of evaluation metrics with human radiologists’ assessment.
For our generation-based method, we use a querying transformer module for modality alignment between an image encoder and a text decoder. We also investigate a novel prompting method to generate both impression and findings report sections with the same model to increase efficiency. The model is trained on a mixed report section dataset and can be prompted to generate both report sections with similar performance to separate single-section models. Finally, we study the impact of different pre-training methods for the querying transformer and find that unlocking the image encoder during pre-training helps with domain adaptation and clinical accuracy but not natural language metrics.
Contents

0 Introduction

1 Deep Learning Background
  1.1 Natural Language Processing ...................................... 7
    1.1.1 Transformer Architecture .................................. 8
    1.1.2 Unsupervised pre-training .................................. 9
  1.2 Computer Vision ................................................. 10
    1.2.1 Vision Architectures ....................................... 10
    1.2.2 Image Contrastive Learning ................................ 11
  1.3 Multimodal Learning ............................................ 12
    1.3.1 Image Captioning ........................................... 13
    1.3.2 Image-Text Contrastive Learning .......................... 14
    1.3.3 Contrastive-Tuning (Locking) .............................. 16

2 Radiology Report Generation Background .......................... 17
  2.1 Task Description ............................................... 18
  2.2 Related Work .................................................. 18
    2.2.1 Retrieval-Based Report Generation ........................ 19
    2.2.2 Generation-Based Report Generation ........................ 20
    2.2.3 Reward-Based Report Generation ............................ 21

3 X-REM: A Retrieval-Based Approach ............................... 23
  3.1 Motivation and Overview ...................................... 24
  3.2 Methods ....................................................... 25
    3.2.1 Data and Preprocessing .................................... 25
    3.2.2 Representation Learning ................................... 26
    3.2.3 Learned Similarity Score ................................ 27
    3.2.4 Retrieval Pipeline ........................................ 28
  3.3 Experiments and Results ...................................... 28
    3.3.1 Implementation Details .................................... 28
    3.3.2 Evaluation Metrics ........................................ 29
    3.3.3 Comparison to Baseline Methods ........................... 30
  3.4 Expert Evaluation Study ....................................... 32
    3.4.1 Study Method .............................................. 32
    3.4.2 Study Results ............................................. 33
List of Contributions for the Work in Chapter 3

The work in Chapter 3 comes from a team project. Katherine Tian contributed the code for data pre-processing and adaptation of the pre-trained model for radiology. Jaehwan Jeong contributed the implementation of the learned similarity score, retrieval pipeline, and model tuning. Sina Hartung and Andrew Li designed the expert evaluation study and recruited radiologists for the study. Sina Hartung, Andrew Li, and Katherine Tian collected expert annotations. Katherine Tian performed the data analysis and prepared the ReFiSco dataset for release.
Acknowledgments

I would first like to express my deepest gratitude to my thesis advisor Professor Pranav Rajpurkar for your guidance and feedback throughout two research projects. Your tireless energy and commitment to research are incredible, and your expertise in radiology and deep learning has been invaluable to my academic growth. Over time, you have connected me with other researchers and medical experts whom I’ve learned from as well. Thank you for being an amazing mentor and a role model. I also want to thank my thesis co-advisor Professor Finale Doshi-Velez. You have been an essential part of my journey with machine learning during all four years of college. Your thoughtfulness in both research and teaching is inspiring and helped shape the way I think about new problems. I’m thankful for all of your support. Both of you have helped me deepen my love of research.

I would like to thank Dr. Liyue Shirley Shen (and soon-to-be professor!) for mentoring me on this project as well and for providing feedback and encouragement on my thesis. I enjoyed our discussions and getting to the bottom of new research decisions with you. I want to thank my team members Jaehwan Jeong, Sina Hartung, and Andrew Li. You are incredible, and I miss working with you all.

Outside of this project, I would like to thank my CS academic advisors Professor
David Parkes, Professor Madhu Sudan, and Professor James Mickens for your advice and encouragement throughout college. Thank you to past research mentors Dr. Raaz Dwivedi and Professor Susan Murphy for your mentorship and for approaching research with so much joy.

To my friends, I’m so grateful to have your constant faith in me and to be able to count on you. I will cherish our many memories together. Finally, I want to extend a huge thank you to my parents and little sister for your unending love and support. I would not be where I am today without you.
Introduction

Chest radiography is the most common form of medical imaging in the world and can provide important diagnostic information on life-threatening cardiopulmonary conditions in a non-invasive manner [1, 2]. However, interpreting a chest X-ray accurately requires trained radiologists’ expert medical understanding, which may not be highly available [3]. Providing accurate, automated radiology reports using machine learning
will significantly reduce costs, ease burdens on radiologist workflows, and minimize
diagnosis delays, especially in low-resource settings [4, 2]. One potential workflow is
having a model generate a report first and then having an experienced radiologist
review and edit the report as necessary. The computer-aided approach to diagnosis
can improve the efficiency and accuracy of radiology diagnoses [5, 6].

Over recent decades, the digitization of image and text data has increased rapidly.
This development, along with the improvement of compute resources, allows for building
increasingly complex and high-performing computational models, such as deep
learning algorithms [7]. There have been several key advancements in deep learning,
such as transformer models [8] and unsupervised pre-training methods [9], which
have pushed forward our ability to interpret data and automate tasks. Thus, the
goal of this work is to work with advanced vision-language deep learning methods,
transferring them to the medical domain in order to automatically generate free text
radiology reports from patients’ radiology images.

Existing modern deep learning methods for generating radiology report generation
include two families of methods: retrieval-based and generation-based [2]. For a given
chest X-ray, retrieval-based approaches often design a similarity metric used to select
the most relevant existing text as the report. Modern retrieval-based frameworks
often use similarity based on neural representations of radiology image and text data
[10]. On the other hand, generative-based approaches often use an image encoder and
text decoder architecture to generate the report token by token with the decoder [2].
Many current works seek to use advanced architectures for the encoder and decoder or target modality alignment between the image embedding and text embedding to improve performance [11, 12, 13]. There are pros and cons of both methods. While retrieval-based methods have guaranteed clinical coherency since they select from human-written texts, they may be limited in variety. Generation-based approaches are more flexible and may be able to describe previously unseen or rare conditions [2]. However, generating domain-specific phrases can be challenging.

There are several challenges facing current methods for radiology report generation. First, although existing methods have recently improved coherency and natural language metrics, there is still much room for improvement with clinical accuracy before these models can be deployed in practice [10]. Models for generating reports sometimes hallucinate information and lack the deep medical understanding needed to self-evaluate the correctness of generated reports [14]. Secondly, medical domain data is very different from general domain data. For example, reports have specialized medical jargon and phrasing, and general object detectors do not generalize well to chest X-rays [15]. Moreover, there is class imbalance among the conditions—“no finding” appears much more often than any other symptom [15]. Lastly, one helpful step toward improving models’ clinical accuracy is developing suitable evaluation metrics to measure clinical efficacy. Current automated metrics are a good proxy for report quality but might not always align with radiologists’ assessment of report accuracy and quality [16].
In this thesis, we present a thorough investigation of both categories of report generation—retrieval and generation. First, we identify two areas of improvement for retrieval methods: (1) improving the quality of image and text embeddings used for calculating similarity and (2) increasing the efficacy of the similarity metric for retrieval. We introduce a novel retrieval-based approach called Contrastive X-Ray REport Match (X-REM), which addresses these weaknesses to achieve state-of-the-art evaluation results. In X-REM, we use a new multimodal encoder to learn pre-trained image and text embeddings, develop a learned similarity metric based on a learned image-text matching score, and build a final selection and filtering pipeline to generate a radiology report. Next, we design and conduct an expert evaluation study to collect radiologists’ annotations of errors in generated reports from X-REM, reports generated from a previous retrieval baseline method, and human-written reports. The sources of the reports were randomized and hidden from the radiologists. The study confirmed the improvement of our method against the previous baseline, yet also revealed the gap between generated reports and human-written reports. We release the dataset of annotations, called Report Fix and Score Dataset (ReFiSco-v0), to aid in future research.

Finally, we present a generative-based approach called Bootstrapping Language-Image Pre-training for Chest X-rays (CXR-BLIP), where we pre-train and fine-tune a modality-alignment query transformer module in between the image encoder and text decoder for report generation. We develop a new method to prompt the same
model to generate multiple report sections to save storage without losing performance compared to single-section models. We also study the effects of different pre-training variants and discover that unlocking the visual encoder during pre-training increases clinical accuracy but not natural language performance.

To provide a roadmap for the rest of this thesis, we first provide background on relevant modern deep learning advances for processing image and text data and automating vision-language tasks in Chapter 1. Next, in Chapter 2, we describe relevant works for the radiology report generation task. In Chapter 3, we present our new retrieval-based method and associated human expert evaluation study. We then describe our novel generative-based approach and discuss the results in Chapter 4. Finally in Chapter 5, we provide a summary of our contributions and directions for future work.
Deep learning refers to the subset of machine learning which uses highly parameterized models based on neural networks. A deep learning method consists of a model architecture, i.e., the layout of neurons in a neural network, as well as a training procedure and loss function for updating the model weights [17]. Deep learning models can automatically extract useful feature representations from the input data for
downstream tasks, such as classification or image captioning [18]. Most deep learning algorithms can be categorized as supervised, unsupervised, or reinforcement learning. Supervised algorithms use feature data to predict associated labels, while unsupervised algorithms learn patterns or structures from unlabeled data only, and reinforcement learning (RL) aims to learn decision-making based on maximizing a reward [17, 19]. Using large-scale datasets, deep learning methods have demonstrated impressive performance on a wide range of benchmark tasks relating to understanding text and image data [20, 21]. Recently, there have been several key advancements in deep learning, which make it feasible to automate complex tasks, such as high-stakes ones in medicine. In this section, we will present important deep learning architectures and training procedures for text data, image data, and multimodal data, which are relevant for radiology report generation.

1.1 Natural Language Processing

Natural language processing (NLP) is the computational analysis of natural language data such as text or speech, with applications in tasks including machine translation or question answering. A common way to teach machine learning models an understanding of text is by mapping chunks of text, such as sentences, words, or sub-words, to numerical vector representations that we call embeddings, which can be used for further computational analysis [22]. Two developments contributed to the strength
of modern-day text representations: the transformer architecture and unsupervised
pre-training methods.

1.1.1 Transformer Architecture

Because text data is sequential in nature, e.g. a sequence of words or tokens, previous
common deep learning architectures for text included Recurrent Neural Networks
(RNNs) [23] and Long Short-Term Memory models (LSTMs) [24], which sequentially
feed in tokens. These models attempt to balance remembering information from the
most recent token and previous tokens but struggle with vanishing gradient issues
for long-range dependencies in text. In 2017, the transformer model was published
by Google Research, which is based on layers of attention mechanisms [8]. Attention
mechanisms use query, key, and value parameter matrices (denoted $Q, K, V$) per to-
ken to model the interactions between each pair of tokens in the input sequence via
matrix multiplication, as given by

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V,$$

where $d$ is a scaling factor. The strength of transformers is that they are able to more
flexibly model long-term word dependencies, which leads to a better understanding of
language. Also, because the architecture is not sequential like an RNN, parallelized
computations in the transformer decreased runtime, allowing for larger models to be
used. Today, transformers are widely regarded as the default state-of-the-art deep learning architecture for handling text.

1.1.2 Unsupervised pre-training

Unsupervised learning algorithms aim to learn structure or representations based on input data without labels. In 2018, the BERT (Bidirectional Encoder Representations from Transformers) model achieved a breakthrough in NLP task performance by using a new unsupervised pre-training method on top of transformers [25]. The method first collects a large corpus of unlabeled text, including data from Wikipedia. Then, BERT introduces a clever yet simple unsupervised training procedure called masked language modeling (MLM). The MLM pre-training task is to randomly mask a percentage of the words and have the model predict the missing word based on the remaining words – this task proved to be very helpful in training high-quality contextual text embeddings. Finally, the BERT pre-trained model can be fine-tuned with labeled data for specific language tasks, such as question answering, to achieve high performance.

Since then, MLM and variations have been used in more advanced unsupervised pre-training methods for text and are still dominant today [26]. A key benefit of this training paradigm is that it is effective even without requiring data labels, which are costly to obtain and limited. For example, supervised question-answering or translation datasets require manual annotation of each question’s answer or each sentence’s
translation and often contain on the order of 100K samples [27]. Instead, pre-training algorithms can scale up and learn from a much larger training set of generic text of 10 million or more samples to broadly improve language understanding in a task-agnostic way. The model is more generalizable and can be fine-tuned further on different task-specific labeled datasets to achieve state-of-the-art results on numerous NLP tasks. The scale-up of this training procedure also comes with using larger architectures, in what we call large language models (LLMs).

1.2 Computer Vision

Similarly, in computer vision, researchers have designed deep learning architectures and training methods suitable for understanding images. Modern computer vision algorithms often process images as arrays of pixels (width \times height \times number of color channels) and seek to implicitly extract higher-level image features, such as edges or shapes, to use for tasks such as image classification, object detection, and image segmentation.

1.2.1 Vision Architectures

One dominant vision architecture is the Convolutional Neural Network (CNN). Each convolutional layer in the CNN scans across the image to extract certain higher-order local features [28]. The performance of computer vision models has been benchmarked
on the ImageNet object classification challenge, a dataset of 14 million labeled images [21]. In 2012, the AlexNet model, one of the earlier deep learning successes, with 5 convolutional layers achieved top-5 accuracy of 85%, a stark improvement from the previous approach [29]. In 2015, the ResNet (residual network) model, which uses 50 layers and skip connections, achieved 96.5% accuracy on the test set and is still used as a solid baseline vision model today [30]. A recent alternative architecture that is gaining traction is the Vision Transformer (ViT), which treats image patches or pixels as sequential data for processing. Recent research often explores using ViTs at scale and with self-supervision [31].

1.2.2 Image Contrastive Learning

Just as in NLP, unsupervised pre-training methods, which are also called self-supervised learning methods, have also been developed for images. In particular, a type of unsupervised pre-training called contrastive learning has achieved impressive results. At a high level, contrastive learning seeks to guide similar pairs of samples close to each other in the embedding space.

One image contrastive learning framework is SimCLR (A Simple Framework for Contrastive Learning of Visual Representations), which achieves 85% ImageNet accuracy after fine-tuning on only 1% of the labels [32]. First, for each image in the dataset, SimCLR applies data augmentations (random cropping, random color distortions, and random Gaussian blur), which change the image slightly but not its label.
Then, during unsupervised pre-training, contrastive loss maximizes similarity between augmentations of the same image while minimizing similarity between augmentations of different images. The loss function for a positive pair $i$ and $j$ is below

$$l_{i,j} = -\frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1, k\neq i}^{2N} \exp(\text{sim}(z_i, z_k)/\tau)},$$

where $z$ is an encoding of the data, $\tau$ is an adjustable temperature parameter, and $\text{sim}(\cdot)$ is the cosine similarity function. These loss terms are summed across all positive pairs in a batch. Finally, this method was fine-tuned on a small proportion of ImageNet labels to achieve results comparable to supervised methods. This method successfully develops powerful image representations.

### 1.3 Multimodal Learning

Multimodal learning is machine learning algorithms that attempt to leverage multiple types of data together, such as vision and language. Being able to analyze vision and language data together opens up the door to automating more complex tasks and improving contextual understanding. For example, multimodal algorithms may ground language understanding to images, i.e. the word “cat” refers to the image of a cat. Multimodal tasks include image-text retrieval [33], image captioning [34], and visual question answering [35]. Additionally, training data from across modalities can also improve performance on single-modality tasks.
1.3.1 Image Captioning

A fundamental approach to image-to-text generation is to use an image encoder and text decoder deep learning architecture, such as in the “Show and Tell” paper [36]. An example image captioning approach uses an image encoder, such as a CNN, which maps the input image \( X \) to an image representation \( x \). Then, the image representation is fed as input to a text decoder, such as an LSTM, to output the first generated caption token \( \hat{y}_1 \). For each subsequent step, to generate the each token \( \hat{y}_k \), we feed in information from \( x \) and \( y_1, \ldots, y_{k-1} \) to the decoder again. We repeat until the decoder outputs a stop token. The parameters of the encoder and decoder can be jointly trained to optimize the conditional log-likelihood with the following loss:

\[
L_{\text{captioning}} = -\log P(Y|x) = -\sum_{k=1}^{K} \log P(y_k|y_1, \ldots, y_{k-1}, x),
\]

where \( Y = (y_1, \ldots, y_K) \) is a target sequence.

Recent state-of-the-art generative approaches often use CNN or ViT-based image encoders and transformer-based text decoders. Moreover, this framework can be improved with unsupervised pre-training and methods for aligning modalities, which we discuss in later sections.
1.3.2 Image-Text Contrastive Learning

One prominent multimodal paradigm for deep learning is image-text contrastive learning. A key framework was published in ConVIRT (Contrastive VIsual Representation Learning from Text) for medical tasks [9]. Here, the authors leverage image-text pairs that appear in data naturally, such as medical images $X_i$ and their associated reports $Y_i$. ConVIRT first applies image augmentations, say $f$, and text augmentations $g$, which were limited to avoid distorting clinical accuracy. Then, it uses contrastive loss to maximize the similarity between augmentations of an image $f(X_i)$ and augmentations of the image’s associated text $g(Y_i)$, while minimizing the error similarity of unpaired image and text data points. After pre-training on paired image and text data, ConVIRT demonstrated better medical classification performance than strong supervised baselines even when fine-tuned on small amounts of labeled data.

OpenAI’s CLIP (Contrastive Language–Image Pre-training) model extended ConVIRT’s image-text contrastive learning framework from medical datasets to natural datasets [37]. Non-domain-specific image-text pairs naturally appear abundantly as image and text captions publicly available on the internet, without the need for manual annotation. First, the CLIP paper collected a large dataset of 400 million image-text pairs publically available on the internet for training. For comparison, the 2015 ResNet was trained on 1.2 million ImageNet samples, and one of ConVIRT’s medical datasets contains 300K images. Without using any ImageNet labels, CLIP
is able to match the fully supervised ResNet-50 ImageNet classification accuracy. In addition to performance gains, one important benefit of CLIP is that it embeds image and text data into the same representation space.

This ability to transfer to downstream tasks without training on labels is called “zero-shot.” Previously, as in the BERT paradigm, pre-trained embeddings still needed to be fine-tuned on smaller subsets of data for a specific task. Now, in contrastive learning-based approaches, these embeddings can potentially do well on several tasks without fine-tuning on downstream datasets—an incredible feat.

An example of how to use a pre-trained CLIP for zero-shot image classification is as follows. Given an image $X$ and all possible classes as strings $y_1, \ldots, y_k$, use the pre-trained model to predict the similarity of each pair $(X, y_i)$. The label with the highest score is the one chosen as a classification, requiring no further training. This is an early example of prompt engineering.

Another work called ALBEF (ALign representations BEfore Fusing) extends these ideas in a vision-language pre-trained model for high performance on downstream multimodal tasks [26]. The authors propose several architecture design choices to obtain these results, including (1) a multimodal encoder with cross-modal attention to fuse image and text features together for the joint representation, and (2) momentum distillation, i.e., keeping a momentum version of the model that takes the moving average of its parameters and using it to generate additional pseudo-targets. Momentum distillation can be used to improve learning from noisy data, such as internet
data in this case. In terms of training, ALBEF uses a loss function that adds MLM and image-text contrastive losses. With additional task-specific fine-tuning, ALBEF achieves state-of-the-art results on several multimodal tasks, including visual question answering (VQA), visual entailment (VE), and visual grounding (VG).

1.3.3 Contrastive-Tuning (Locking)

A growing paradigm in multimodal learning is contrastive-tuning [38], which tunes models with locked unimodal components on contrastive losses rather than end-to-end multimodal contrastive training from scratch. By leveraging existing unimodal models, this approach can drastically improve compute time and data efficiency, as well as generalization to out-of-distribution, for multimodal learning. The paper found that on natural images, training with locked image and unlocked language components performed well. Some other examples include FROMAGe (Grounding Language Models to Images for Multimodal Generation) [39], which takes a frozen pre-trained LLM and fine-tunes the input and output layers for visual grounding in multimodal reasoning tasks, and BLIP-2 (Bootstrapping Language Image Pre-training 2) [40], which introduces a cross-modal transformer module in between a visual encoder and LLM and trains the model in two pre-training stages where either the visual or language component is frozen.
In this chapter, we provide relevant background and related works for research in automatic radiology report generation. First, we describe details of the report generation data and task setup. Then, we dive into related research for radiology report
2.1 Task Description

Typically, the target radiology report for the report generation task contains multiple fixed sections, which include

- an INDICATION, which is a description of the context for why the X-ray was requested,
- a FINDINGS section, which contains descriptions of the radiologist’s observations of the image, and
- an IMPRESSION section, which is a summary of the relevant findings for potential diagnosis or next steps.

Each report is typically associated with one or more chest X-ray images taken from different views, such as frontal or lateral. Radiology report-generation methods typically aim to generate the findings section, impression section, or both given the study image(s). Studies might also consider conditioning on the extra information from the indication section to guide the report generation. Commonly used radiology image and report datasets include MIMIC-CXR ($N = 227,827$ reports) and IU X-Ray ($N = 7,470$ reports) [3, 41].

2.2 Related Work

There is a growing body of literature on the automatic interpretation of radiology data, including deep learning methods for classifying conditions and generating re-
ports. Some benefits of generating free-text reports instead of outputting a disease classification are that reports have the potential to convey the severity of the diagnosis, an explanation of the symptom including the location of the affliction, and other flexible information outside a fixed number of given disease categories.

For automatic report generation, we discuss three types of methods for report generation: retrieval, generation, and reward optimization. Retrieval-based methods frame report generation as an image-text retrieval task: design a mechanism to retrieve the most relevant text from a training corpus as the report for a given radiology image. The generation-based approach, similar to medical image captioning, works as follows: given a medical image, produce its report sequentially conditional on previous tokens and the image. Finally, another branch of work explores designing helpful rewards for high-quality reports and using reinforcement learning to optimize deep learning end-to-end architectures to maximize the rewards. These reward-based approaches can be used independently or to further improve another end-to-end method.

2.2.1 Retrieval-Based Report Generation

Modern retrieval-based systems use neural image and text representations at the report level or sentence level to help identify relevant texts. For a new input image, the system encodes its visual representation, compares it with the language embedding for every text in the corpus, and then retrieves the most similar text. The goal is to design a retrieval system to consistently select clinically accurate texts for new chest
X-rays as a suitable report. One such method is CXR-RePaiR (Contrastive X-ray Report Pair Retrieval), which fine-tunes CLIP image and text representations for the radiology domain using MIMIC-CXR image-impression pairs [10]. Then, the method uses cosine similarity between the image and text representations to retrieve texts, achieving high clinical accuracy for report impressions. Since the model’s output is taken from existing written snippets, the method helps ensure the generated text is coherent. Moreover, by retrieving from an existing corpus, the method takes advantage of the limited possibility space of radiology report findings.

2.2.2 Generation-Based Report Generation

Generation-based methods for radiology report generation can conceptualize the task as a medical image captioning task. They often take inspiration from the state-of-the-art image-to-text generation approaches used for natural image captioning task benchmarks. For example, a prior method for radiology report generation (R2Gen) trains a ResNet CNN image encoder and novel enhanced transformer-based decoder for report generation. The novel decoder is a Memory-Driven Transformer, which uses memory-driven conditional layer normalization to incorporate a relational memory that can enhance the transformer’s ability to learn patterns and generate text [11].

One recent direction of work among generative methods is to use architectures that facilitate modality alignment between the image encoder and text decoder to improve performance. For example, the Meshed-Memory Transformer (M2Trans)
used for report generation in the IFCC approach uses a memory-augmented attention
mechanism that extends keys and values to process image feature information along
with the text [42, 12]. Another specialized architecture is the Cross-Modal Memory
Network (CMN) [13].

In terms of training innovations, WCL [43] uses a weakly supervised contrastive
loss with “hard negatives” to achieve good results, and CvT2DistilCPT2 [44] finds
that warm-starting encoder-decoder training from domain-specific pre-trained check-
points helps improve performance.

2.2.3 Reward-Based Report Generation

These report-generation architectures can also be trained end-to-end using modern re-
inforcement learning techniques with customizable non-differentiable reward functions.
Within reinforcement learning, the policy-gradient method REINFORCE gives us the
expected gradient of a non-differentiable reward function \( r(y) \) as

\[
\nabla_\theta = -E_{y \sim p_\theta} [r(y) \nabla_\theta \log p_\theta(y)],
\]

which is estimated in practice using a single Monte Carlo sample \( y \) as

\[
\nabla_\theta = -r(y) \nabla_\theta \log p_\theta(y).
\]
We can take the difference between the reward and a reference baseline reward without changing the gradient [45]. For image captioning and report generation, we use an RL algorithm called self-critical sequence training (SCST), which is a variant of the REINFORCE algorithm that uses a model’s output to normalize rewards instead of estimating a baseline [46].

As an example in report generation, a previous method uses the M2Trans captioning architecture and designs novel rewards to encourage factual completeness and consistency of generated reports, based on entity match and entailment. Then, the authors optimize the architecture end-to-end on these rewards along with BERTScore using SCST [12]. The authors compute the loss for a reward $r$ as

$$\nabla_{\theta} L(\theta) = -\nabla_{\theta} \log P(y_s|\mathbf{x})((r(y_s) - r(\hat{y}_g)),$$

where $\mathbf{x}$ is a chest radiology image, $y_s$ is a sampled report and $\hat{y}_g$ is a generated report. Another method uses SCST to facilitate modality alignment over a CMN architecture [47].
In this chapter, we present a novel retrieval-based method for report generation, called Contrastive X-ray REport Match (X-REM), which was co-developed with Jaehwan Jeong. We discover that using a multimodal encoder to learn medical pre-trained representations and a learned image-text matching similarity metric significantly improves retrieval-based accuracy. We also conduct an expert evaluation study.
and release a new dataset of report error scores and fixes (ReFiSco-v0) to better understand the alignment between automatic evaluation metrics and radiologist assessment of reports.

3.1 Motivation and Overview

A previous work studied the theoretical upper bound on retrieval-based performance via the oracle report which maximizes each evaluation metric [16]. The authors found that the upper bound is high, but prior retrieval-based approaches are still far from maximizing performance. Thus, the aim of X-REM is to narrow this performance gap while maintaining the benefits of retrieval-based approaches.

Existing retrieval-based report-generation methods often leverage neural image and text representations. Given an image embedding, the method selects the text with the most similar language embedding to retrieve as the report. Modern representation learning literature provides promising results on image-text retrieval for natural images. However, there is room for improvement, especially when adapting similar methods to radiology. In our work X-REM, we use a deep pre-trained vision-language model, which uses cross-modal attention and masked language modeling (MLM), image-text contrastive (ITC), and image-text matching (ITM) training losses. We fine-tune the model on radiology data to generate high-quality image representations of chest X-rays and text representations of their associated radiology reports. More-
over, we develop a novel learned neural score as a similarity metric for the retrieval method. Finally, we present performance metrics and a human expert evaluation study and demonstrate X-REM’s performance improvement over strong baselines and state-of-the-art models.

3.2 Methods

3.2.1 Data and Preprocessing

For training and testing our model, we use MIMIC-CXR, one of the largest available radiology datasets [3]. The dataset contains 227,835 imaging studies for 65,379 patients, with a total of 377,110 images. In each study, the patient may have one or more X-rays, possibly from different views, such as frontal and lateral. The images are originally collected in a Digital Imaging and Communications in Medicine (DICOM) format and have been compressed to JPG files for ease of analysis [48]. Unless the report is missing, each imaging study is associated with a free-text radiologist-written report describing the imaging study.

The train and test splits are created such that the same patient does not have studies in both the train and test sets. For developing our model, we select one image from each study to pair with the report. If available, we select a frontal view X-ray image. The training set, which we used for pre-training and as our retrieval corpus, contains 185,538 studies with impressions, 123,839 studies with findings, and 123,814
Figure 3.1: An overview of the training step of the X-REM language-image model. Pre-training radiology image and text representations is discussed in Section 3.2.2, and fine-tuning for ITM similarity scoring is discussed in Section 3.2.3. Matching Loss* in the pre-training step is different from the fine-tuning objective in terms of the negative samples used.

studies with both report sections.

3.2.2 Representation Learning

As the first step of our retrieval method, we train high-quality, aligned radiology image and text embeddings using the ALBEF pre-training model from Salesforce, which was introduced in Section 1.3.2. An overview of this training step is shown on the left diagram in Figure 3.1. The model contains a unimodal image encoder, a unimodal text encoder, and a multimodal encoder with cross-modal attention, which fuses together the unimodal image and text representations. The model is trained on the sum of ITC, MLM, and ITM losses. The ITC loss is applied to the image and text unimodal embeddings to first align them, and the MLM and ITM loss are applied to the multimodal encoder to achieve better multimodal understanding. For more detail on the multimodal losses, the multimodal MLM loss predicts the masked words using both the image and text information. The ITM loss uses a fully connected layer and softmax to predict whether an image-text pair is associated (i.e., matched) or not and
takes the prediction cross-entropy loss. We select ALBEF as a suitable encoder for our retrieval method because of its features for enhanced multimodal understanding and its resulting high performance when adapted for relevant downstream tasks such as image-text retrieval, VQA, and visual entailment.

To adapt this pre-trained model to the radiology domain, we train the model using these losses on radiology image-report pairs from the MIMIC-CXR dataset, starting from the ALBEF checkpoint pre-trained on natural images. We use the original ALBEF architecture, a BERT$\text{base}$ text encoder and ViT-B/16 visual encoder [25, 31].

3.2.3 Learned Similarity Score

In this section, we introduce the similarity function we design for our retrieval report generation method. Image-text retrieval relies on a similarity function $f(x, y)$, where $x$ is an image representation and $y$ is a text representation. We define a novel learned image-text matching score, which we use as the similarity function $f$. Our idea is to use an image-text matching model $M$ to learn the matching score. We train the model $M$ to classify whether a radiology image $x$ and report $y$ is associated with the same study or not, outputting logits $(p_+, p_-)$. Then, we take the logit of the positive class as the image-text matching score. In our case, we use a multi-layer perceptron binary classification head on top of our ALBEF representations as $M$. We directly fine-tune our pre-trained encoder along with the classification head on the ITM task as our similarity scorer. This training step is depicted on the right side of Figure 3.1.
3.2.4 Retrieval Pipeline

Our retrieval system has a final report selection and filtering step controlled by hyperparameters $i, j, \text{ and } k$ (default $i = 50, j = 5,$ and $k = 2$), as shown in Figure 3.2. Since computing our neural ITM score for each item in the full corpus would be computationally expensive and impractical, we first use embedding cosine similarity to pre-select the top $i$ candidate reports. Then, we compute ITM score for only these $i$ reports and then select the top $j$ reports. Finally, we concatenate at most $k$ of the retrieved texts together after applying a natural language inference (NLI) filter to exclude contradictory or redundant texts.

3.3 Experiments and Results

3.3.1 Implementation Details

We conduct pre-training of the ALBEF representations on MIMIC-CXR data for 60 epochs with a starting learning rate of 1e-5. Then, we conduct ITM fine-tuning for 8 epochs with a starting learning rate of 2e-5. We used input image size 256x256
for pre-training and 384x384 for fine-tuning. Using 4 NVIDIA RTX 8000 GPUs, the pre-training stage took 2 days, and the fine-tuning stage took 4 days.

3.3.2 Evaluation Metrics

Let us describe the evaluation metrics we used to evaluate our model. Common quantitative metrics for assessing generated radiology reports include both natural language (NLG) metrics and clinical accuracy (CA) metrics. NLG metrics measure how well the words of our generated report match the words in the ground truth report. However, NLG metrics might not capture important clinical semantics. For example, “no evidence of pneumonia” and “evidence of pneumonia” may score similarly on NLG metrics but would lead to very different diagnoses in practice. Thus, we utilize other works that design automated clinical accuracy computation. Our goal is to deliver human-readable and accurate free-text reports, so we consider both NLG and CA metrics for evaluation.

For NLG metrics, we report BLEU-2 and BERTScore [49]. BLEU-2 counts the number of unigrams and bigrams, with overlap allowed, between the generated and ground truth reports. BERTScore measures cosine similarity between the token embeddings from the generated report and the ground truth report. For CA metrics, we report CheXbert vector similarity [50] and RadGraph F1 [51]. CheXbert vector similarity first extracts a 14-dimensional vector that indicates the presence of 13 common symptoms or a no-finding observation for each report, then computes the similarity
between the vectors for the generated and ground truth reports. RadGraph F1 extracts a knowledge graph of entities and relations from each report, then computes an F1 score to measure the similarity between the graphs of the generated and ground truth reports. Lastly, we use RadCliQ score, which is a combination of these 4 metrics (BLEU2, BERTScore, CheXbert, and Radgraph F1) that is intended to measure report error, so lower is better. We keep in mind that these automated metrics may not perfectly quantify the quality of generated reports according to radiologists. A previous study examined which metrics are more aligned with radiologists' evaluations and found that RadCliQ score was most aligned with radiologists’ evaluations [16].

### 3.3.3 Comparison to Baseline Methods

We compare X-REM to a selection of strong baseline and state-of-the-art report generation deep learning models on these quantitative metrics on the MIMIC-CXR test set. For direct comparisons, we re-train our model on MIMIC-CXR to generate impression sections, findings sections, or findings and impression sections concatenated to match the output of the other state-of-the-art methods.

On impressions, we compare X-REM to a previous SOTA retrieval method CXR-RePaIR, which fine-tuned CLIP embeddings on MIMIC-CXR data for impressions. Secondly, we also fine-tune the BLIP model on impressions to serve as a generative baseline of comparison against X-REM. BLIP is an image captioning model pre-trained on ITC and ITM loss, like X-REM, but is a generative-based method rather
Table 3.1: Comparison of X-REM to previous report generation models. Models trained on data $I$, $F$, $I + F$ generate the impressions, findings, and both the impression and the findings sections of MIMIC-CXR, respectively ($M^2$ has been additionally trained on CheXpert). Results with * are taken from [16], which evaluated the models on the identically-preprocessed MIMIC-CXR test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>RadClinQ ↓</th>
<th>RadGraph $F_1$ ↑</th>
<th>CheXbert ↑</th>
<th>BERTScore ↑</th>
<th>BLEU2 ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-REM</td>
<td>I</td>
<td>3.781</td>
<td>0.133</td>
<td>0.384</td>
<td>0.287</td>
<td>0.084</td>
</tr>
<tr>
<td>CXR-RePaiR*</td>
<td>I</td>
<td>4.121</td>
<td>0.090</td>
<td>0.379</td>
<td>0.193</td>
<td>0.055</td>
</tr>
<tr>
<td>BLIP</td>
<td>I</td>
<td>4.313</td>
<td>0.046</td>
<td>0.309</td>
<td>0.190</td>
<td>0.030</td>
</tr>
<tr>
<td>X-REM</td>
<td>I + F</td>
<td>3.835</td>
<td>0.172</td>
<td>0.351</td>
<td>0.287</td>
<td>0.161</td>
</tr>
<tr>
<td>WCL*</td>
<td>I + F</td>
<td>3.986</td>
<td>0.143</td>
<td>0.299</td>
<td>0.275</td>
<td>0.144</td>
</tr>
<tr>
<td>R2Gen*</td>
<td>I + F</td>
<td>4.051</td>
<td>0.134</td>
<td>0.286</td>
<td>0.271</td>
<td>0.137</td>
</tr>
<tr>
<td>$M^2$ Trans†*</td>
<td>F</td>
<td>3.277</td>
<td>0.244</td>
<td>0.452</td>
<td>0.386</td>
<td>0.220</td>
</tr>
<tr>
<td>X-REM</td>
<td>F</td>
<td>3.585</td>
<td>0.181</td>
<td>0.381</td>
<td>0.353</td>
<td>0.186</td>
</tr>
<tr>
<td>CvT2DistilGPT2*</td>
<td>F</td>
<td>3.617</td>
<td>0.183</td>
<td>0.375</td>
<td>0.347</td>
<td>0.196</td>
</tr>
</tbody>
</table>

than retrieval-based [52]. For findings and impression, we compare X-REM to WCL and R2Gen, two previous state-of-the-art generative methods. On just findings, we compare X-REM to SOTA method M2Trans and strong baseline CvT2DistilGPT2.

The results are shown in Table 3.1. X-REM makes significant improvement upon the previous retrieval-based method CXR-RePaiR, which we attribute to our novel learned similarity metric and improved modality alignment in the pretrained embeddings. Moreover, X-REM outperforms the other generative approaches except for M2Trans, demonstrating the viability of learned retrieval methods with modern pre-training methods. However, we also caveat that M2Trans has been trained on an external dataset, CheXpert, in addition to MIMIC-CXR.
3.4 Expert Evaluation Study

To supplement standard automated quantitative evaluation, we conduct a human expert evaluation study to further analyze the quality of X-REM with clinical context. This study has been approved by an institutional review board. In this study, we compare X-REM with an ML baseline and a human benchmark.

3.4.1 Study Method

We conduct the human evaluation on a selected subset of 60 studies from MIMIC-CXR. For each study, we compile three reports: one generated from our method (X-REM), one generated from a machine learning baseline (CXR-RePaiR), and one taken from a human benchmark (MIMIC-CXR). Both X-REM and CXR-RePaiR were retrained on the same MIMIC-CXR training set without the 60 samples.

We recruit four radiologists to provide evaluations on the reports. To each expert, we present one image and one report for each of the 60 studies. Each report is randomly and independently chosen from one of the three sources. The radiologist is blinded to the source. We ask each radiologist to assess the error severity of their assigned reports, so each study receives four expert annotations. The study set-up is summarized in Figure 3.3.

Next, we describe how the error of each report annotation is assessed. Each report is broken down into lines, and the radiologist is asked to score the error of each line.
Figure 3.3: Human study randomization process. For each study, we choose a report from the three sources and presented the chest X-ray and radiology report to human radiologists for evaluation.

of the report. The radiologists select from five possible error categories (No error, Not actionable, Actionable nonurgent error, Urgent error, or Emergent error), which we map to error scores 0, 1, 2, 3, 4, respectively, in increasing order of severity. We aggregate severity per line to a single severity score per report using the following two metrics: (1) maximum error severity (MES) across all lines in the report, which captures the worst error of any line in the report, and (2) average error severity (AES) per line in the original report. AES is the sum of error severity across lines standardized by the number of lines, which avoids punishing longer reports.

3.4.2 Study Results

Our analysis finds that our method X-REM increases the number of zero-error reports by 70% and provides statistically strong improvements over the previous ML baseline.
Table 3.2: Human Evaluation Study Results. The table shows the empirical cumulative distribution for each report source on MES and AES scoring. For example, 87% of reports generated by X-REM received max error severity of 3 or less.

We present the empirical distribution of MES and AES across annotations for each source in Table 3.2. According to the radiologists in our study, 17% of reports generated by X-REM have zero mistakes, which surpasses 10% from the baseline but lags behind the human benchmark’s 35%.

**Paired Comparison:** In order to directly assess the improvement of CXR-RePaiR over the baseline, we analyze the 40 studies that have at least one annotation from both CXR-RePaiR and the baseline. We report how often X-REM improves upon the baseline and the magnitude of error severity decrease.

We find that in 62.5% of studies, X-REM has the same or lower MES compared to the baseline. On average across shared cases, our method reduces MES from 2.29 to 2.11 compared to the baseline. We also conduct a paired t-test on the mean MES and observe a p-value of 0.043 (t-stat 1.77). On the AES metric, 70% of shared reports have an equal or lower AES compared to the baseline. On average, our method reduces AES from 2.48 to 1.76 compared to the baseline. A paired t-test on the mean AES yields a p-value of 0.0062 (t-stat 2.62).
CXR-BLIP: A Generative-Based Approach

In this chapter, we present an independently-developed generative-based approach called CXR-BLIP for radiology report generation. We use a querying transformer (Q-Former) module as a cross-modal transformation between an image encoder and
text decoder and investigate a novel prompting method to generate both impression
and findings sections with the same model. We find that the prompting model was
able to maintain performance and hold inferential power for generating both sections
simultaneously.

4.1 Overview

We aim to address some limitations of the retrieval-based method in the previous
section. First, retrieval methods are inflexible and rely on a limited set of existing
sentences, resulting in a performance ceiling. Generative models do not have this
restriction and may be able to learn more intricate relationships between visual and
textual features. Secondly, X-REM needs to be re-trained for generating findings or
impression sections, which requires extra compute and storage. In this chapter, we
develop a generative model that can flexibly generate both findings and impressions
when prompted.

In our method, shown in Figure 4.1, we use a state-of-the-art ViT encoder and
LLM decoder. To facilitate modality alignment, we train a Q-Former image and text
transformer module in between the encoder and decoder with a pre-training and fine-
tuning stage. Finally, we present a new prompting method to train one single model
to generate either findings or impression sections when prompted.
4.2 Methods

4.2.1 Data Preprocessing

The original MIMIC-CXR images are 2544x3056 pixels. We resize the images to 224x224 for pre-training and 364x364 for fine-tuning, maintaining the aspect ratio of the image and adding padding on the sides. Then, we construct a dataset of image-impression pairs out of all the reports that had an available impressions section and a similar image-findings dataset. The impression dataset has 316,441 train samples, 2571 validation samples, and 3704 test samples, and the findings dataset has 260,159 train samples, 2026 validation samples, and 3518 test samples. We train on either the impression dataset, findings dataset, or both datasets mixed together.
4.2.2 Model Architecture

In this section, we describe the end-to-end architecture we use for report generation. We took the BLIP2 default architecture, which uses the ViT-G/14 from EVA-CLIP for the vision encoder [53]. It has 40 layers, a hidden dimension of 1408, and 16 attention heads, for a total of 1.011 billion parameters, and it takes in an input image size of 224x224. The LLM decoder is Meta AI’s 2.7B OPT Transformer [54], which has 32 layers and 32 attention heads. It is a pre-trained decoder-only causal language model. The Q-Former module uses 12 self-attention layers, GELU activations, and dropout with \( p = 0.1 \). For the learnable queries, we use 32 queries each with an embedding dimension of 768. The size of the learnable queries (32x768) is much smaller than the large pretrained VE output feature space of VIT-L/14 (257x1408) in order to create an information bottleneck.

4.2.3 Querying Transformer and Contrastive Tuning

For this method, we leverage the contrastive-tuning vision-language task paradigm used for image captioning from BLIP2 as a starting point [40]. The method allows us to adapt frozen off-the-shelf large pre-trained visual encoders (VEs) and large language models (LLMs) for multimodal tasks like vision-language reasoning. We adapt this methodology to the radiology domain and the report generation task.

In order to successfully mix and match frozen VEs and LLMs, BLIP2 provides
cross-modal alignment using a Querying Transformer (Q-Former) module in between the VE and LLM, as shown in Figure 4.2. The Q-Former is an image and text transformer where the image and text submodules share the same self-attention layer weights in order to facilitate modality alignment. Moreover, the Q-Former is paired with learnable query embeddings, which are input to the image submodule. The queries interact with the frozen VE via additional cross-attention layers and interact with the paired text via the self-attention layers. Thus, the queries serve to create an information bottleneck and extract the most relevant image features to pass to the LLM decoder.

We conduct one pre-training stage on the Q-Former using ITC, ITM, and generation losses. However, in our case, we use an unlocked VE for domain adaptation. Then, we unlock the VE for fine-tuning on the report generation task with a generation loss. Originally, the BLIP2 Q-Former is trained for the image captioning task in two locked pre-training stages and one unlocked fine-tuning stage. However, since the
second pre-training stage uses the same generation loss as fine-tuning on the image-captioning task, we wish to skip the second pre-training step and directly unlock the VE to allow for visual domain adaptation. Our training stages are described in the subsequent sections.

4.2.4 Pre-training Stage: Representation Learning

In this step, we connect a pre-trained unimodal VE to the Q-Former and train the Q-Former to extract image feature representations from the VE that are most relevant to the text. This stage uses three pre-training loss functions: image-text contrastive (ITC) loss, image-grounded text generation (ITG) loss, and image-text matching (ITM) loss.

- The ITC loss is computed on output query representations from the Q-Former image transformer and paired text embeddings from the Q-Former text transformer.

- ITG uses a causal mask to train the Q-Former text transformer to generate the texts with a captioning loss. This way, it guides the learnable queries to capture visual information relevant to generating the paired text.

- Using the output query embedding as input to a linear ITM classifier, the prediction logit is used as a matching score for the ITM loss function. ITM loss aims to help with image-text alignment.

We construct a mixed dataset of separate findings and impressions sections for pre-training. Specifically, we create a set of all the image-findings pairs and a set of all the image-impression pairs. Moreover, we also tag each findings and impression
text with a special “<findings>” or “<impression>” token at the start of the text so that the model can distinguish between the two report sections. Since the impression section is typically a summary of the most important findings, the two sections contain very related information, and training on data from both sections can provide helpful data augmentation on top of training on just one section. Also, there are many reports that may have one of the findings or impression section missing, which we previously had to filter out during concatenation but we can now leverage during training, increasing our number of available training data points from 221,531 to 576,660, which should improve performance.

4.2.5 Fine-tuning for Report Generation

For this training step, we now connect the Q-Former to an unlocked VE encoder and frozen LLM decoder. We project the output queries to the text embedding dimension before feeding it into the LLM. Here, the output query vectors serve as visual prompts to condition the LLM on information from the given image. This step uses a text-generation loss between the target report and generated report.

Many previous report-generation methods that aim to generate both findings and impression sections do so by concatenating the two sections, such as [55]. However, these methods then cannot separate the two sections for evaluation or other purposes. Instead, we propose a new method to prompt a single LLM to produce impression or findings sections, thereby saving storage and time. We accomplish this by training
on a mixed dataset of both impressions and findings with separate prompts. Since the LLM is frozen, we do not train the prompt embeddings. Instead, to emphasize the distinction between sections to the model, we chose relatively long prompts for fine-tuning for impressions and findings:

- “The following is a summary of the most important observations of the chest x-ray image. <impression> ”
- “The following is a description of an expert radiologist’s observations of the chest x-ray image. <findings> ”

During training, we concatenate each report section’s corresponding prompt to the front of the text to prompt the decoder for the rest of the generation, and we apply the text generation loss only to newly generated tokens, excluding the prompt. We try fine-tuning on impressions directly, findings directly, and the mixed dataset of both sections with identifying prompts.

4.3 EXPERIMENTS AND RESULTS

4.3.1 IMPLEMENTATION DETAILS

We conduct each training stage until the validation loss stopped decreasing, which took 6 epochs for pre-training and 10-14 epochs for fine-tuning depending on the dataset. Using 4 rtx8000 GPUs with 48G VRAM, training takes around 4 hours per epoch. We used a max token length of 150, beam search size of 3 for pre-training and 5 for fine-tuning, and starting learning rate of 5e-6.
Table 4.1: Evaluation of the CXR-BLIP method on impression (rows 1-2) and findings (rows 3-4). All rows were trained with a locked pre-training stage and an unlocked fine-tuning stage. Rows 2 and 4 were trained on a mixed dataset and evaluated by prompting for impressions or findings respectively.

<table>
<thead>
<tr>
<th>Train</th>
<th>Eval</th>
<th>RadCliQ ↓</th>
<th>RadGraph F1 ↑</th>
<th>CheXbert ↑</th>
<th>BERTScore ↑</th>
<th>BLEU2 ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>I</td>
<td>4.560</td>
<td>0.053</td>
<td>0.352</td>
<td>0.076</td>
<td>0.025</td>
</tr>
<tr>
<td>Mixed</td>
<td>I</td>
<td>4.493</td>
<td>0.041</td>
<td>0.349</td>
<td>0.106</td>
<td>0.021</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>4.590</td>
<td>0.081</td>
<td>0.222</td>
<td>0.143</td>
<td>0.068</td>
</tr>
<tr>
<td>Mixed</td>
<td>F</td>
<td>4.563</td>
<td>0.085</td>
<td>0.211</td>
<td>0.158</td>
<td>0.073</td>
</tr>
</tbody>
</table>

4.3.2 Mixed-Dataset Model Prompting Results

In this section, we compare the results of prompting a single mixed model to results from separate standard models, shown in Table 4.1. All rows of this table use a locked-VE pre-training stage and an unlocked-VE fine-tuning stage. Recall that RadGraph F1 and CheXbert are clinical accuracy metrics, while BERTScore and BLEU2 are natural language metrics, and the direction of the arrow, up or down, indicates whether higher or lower is better. Rows 1 and 3 use a standard data pipeline: in row 1, the model is fine-tuned and evaluated on impressions, and in row 3, the model is fine-tuned and evaluated on findings. Rows 2 and 4 show results based on the model which was fine-tuned on the mixed-section dataset. Reports are generated for rows 2 and 4 by prompting for impressions or findings respectively.

We observe that despite not training solely on the evaluation task, the mixed-dataset model has comparable performance to the single-dataset models. The mixed model improves RadCliQ and BERTScore for impressions, and RadCliQ and RadGraph F1 for findings, while degrading only very slightly across other metrics. This
suggests that our model is capable of holding useful inferential information for generating both impressions and findings. Also, the impression and findings-generation tasks are related, so training on both sets of data together could provide helpful augmentation. We believe that improving such multi-task models is promising for future work.

### 4.3.3 Training Stages Ablation

In this section, we also conduct an ablation test on the impact of different training stages on performance, reporting the results in Table 4.2. For this set of ablations, we consider only the impression section. In row 1, we skip any medical pre-training and directly fine-tune the model on MIMIC-CXR data starting from the general domain pre-trained checkpoint. In row 2, we pre-train the Q-Former on MIMIC-CXR while keeping the VE locked, before fine-tuning on radiology data as well. In row 3, we unlock both the Q-Former and VE during pre-training on MIMIC-CXR, then fine-tune on MIMIC-CXR.

We notice that adding a pre-training stage with a locked VE degrades performance

<table>
<thead>
<tr>
<th>Pt</th>
<th>Ft</th>
<th>RadCliQ ↓</th>
<th>RadGraph F$_1$ ↑</th>
<th>CheXbert ↑</th>
<th>BERTScore ↑</th>
<th>BLEU2 ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>U</td>
<td>4.219</td>
<td>0.044</td>
<td>0.351</td>
<td>0.197</td>
<td>0.033</td>
</tr>
<tr>
<td>L</td>
<td>U</td>
<td>4.560</td>
<td>0.060</td>
<td>0.351</td>
<td>0.073</td>
<td>0.025</td>
</tr>
<tr>
<td>U</td>
<td>U</td>
<td>4.324</td>
<td>0.054</td>
<td>0.376</td>
<td>0.140</td>
<td>0.030</td>
</tr>
</tbody>
</table>

**Table 4.2**: Ablation on training stages for impressions. In columns Pt (pre-training) and Ft (fine-tuning), L and U denote whether the VE was locked or unlocked during that column’s training stage, and no letter means the training stage was not conducted on MIMIC-CXR.
from only fine-tuning. This makes sense because the image representations from out-of-domain VE may not be helpful enough for guiding the Q-Former. For the sake of ablation, we also tried a version of the model with a locked-VE pre-training 2 stage between pre-training and fine-tuning as in the original BLIP-2 procedure, which resulted in incoherent reports, so we exclude it from the table. This confirms that pre-training the Q-Former, especially on text generation loss, with a locked out-of-domain VE is unhelpful.

Unlocking the VE during the pre-training stage does improve performance on most metrics compared to locked pre-training, confirming that training to adapt the VE in addition to the Q-Former for the medical domain is helpful. Compared to fine-tuning only, the unlocked pre-training stage showed mixed results, improving our clinical accuracy metrics, Radgraph F1 and CheXbert, and not our natural language metrics, BERTScore and BLEU2. This result demonstrates that additional pre-training with an unlocked VE potentially allows for better representations that can increase clinical accuracy.

Our impression models surpass the baseline model CXR-RePair-Select on CheXbert score (3.343) and demonstrate the potential of our model. Moreover, when comparing to our baseline models in Table 3.1, we note that we trained only on the train set, while many of the other models were trained on the extra samples from the validation set in addition to the train set, giving them a performance advantage. However, this model does not yet achieve state-of-the-art results overall. For further improvement,
we wish to swap out the current bootstrapped encoder and decoder with a suitable medical image encoder and text decoder, which will provide closer adaptation to radiology. We plan to do this in the near future.
In this thesis, we did a deep dive into automated radiology report generation, exploring how well modern machine learning methods transfer to the medical domain. We propose two novel methods using deep learning and pre-training for radiology report generation.
5.1 Summary of Contributions

In our retrieval-based method (Chapter 3), to generate high-quality automated reports, we pre-train a multimodal vision-language model on radiology image and text data to learn effective radiology embedding representations, leverage a learned similarity metric for retrieval based on a neural image-text matching score, and create a retrieval framework that pre-selects using cosine similarity and post-processes using an NLI filter. Moreover, to facilitate better alignment of generated reports to radiologists’ opinions, we conduct an expert evaluation study on a subsample of radiology imaging studies. We collect radiologists’ annotations of errors on reports generated by our method, generated by a baseline generation method, and taken from a human benchmark. We compile the annotations of report errors and error scores in the dataset Report Fix and Score (ReFiSco-v0) to further research toward studying types of errors and alignment of evaluation metrics with human assessment. Based on automatic natural language and clinical accuracy evaluation metrics as well as error scoring from the human expert evaluation study, our retrieval-based approach achieved state-of-the-art results compared to other existing models.

In our generative-based method (Chapter 4), we explore the possibility of connecting a unimodal image encoder and unimodal text decoder with a modality-alignment transformer module called the Q-Former for radiology report generation. We adapt the Q-Former and visual encoder to the radiology domain by pre-training on a dataset
of mixed report sections and fine-tuning the end-to-end encoder-decoder pipeline for report generation. We design and implement a prompting method for flexibly generating both impression and findings sections from one model instead of training one for each section and find that the LLM decoder is capable of being prompted for both reports. Lastly, we observe that unlocking the visual encoder during pre-training improves clinical accuracy but not natural language metrics.

5.2 Future Work

In Chapter 3, we present a novel retrieval-based report generation that achieves state-of-the-art results on standard quantitative evaluation metrics in the field. However, there still exists a notable gap between reports written by human radiologists and reports generated by ML models including X-REM. The purpose of our human evaluation was to measure the current size of the gap, and we plan to release our ReFiSc dataset to aid future research toward improving alignment with radiologists’ assessment. Secondly, we only developed our model on the MIMIC-CXR train and val splits and tested it on the MIMIC-CXR test split, but in the future, we would like to evaluate X-REM’s generalization capability to other datasets, such as IU X-Rays.

In Chapter 4, we introduce a novel pre-trained generative image-captioning framework for radiology report generation. We see replacing the VE and LLM in our framework with frozen medical pre-trained models as an important next step to capture
further unrealized performance gains. Although we currently unlock the VE and train the Q-Former to help extract relevant medical visual features, starting with a medical VE that has already been pre-trained on radiology data will likely improve these image features further since salient objects in natural images look very different from important aspects of chest X-rays. For text decoding, we use a flexible sub-word tokenizer that is currently able to generate many medical terms and phrases, but we could try a pre-trained medical LLM for additional benefits. Specifically, we may want to replace the VE and LLM components with the BioViL image encoder and CXR-BERT decoder from Microsoft Health [15] or the radiology-finetuned CLIP encoder and decoder from CXR-RePaiR [10]. Also, we would like to try unlocking and training the prompt embeddings in the LLM for more fine-tuned mixed models, an idea inspired by [39].
References


[6] Yongsik Sim, Myung Jin Chung, Elmar Kotter, Sehyo Yune, Myeongchan Kim, Synho Do, Kyunghwa Han, Hanmyoung Kim, Seungwook Yang, Dong-Jae Lee,


[40] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.


Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019.


[53] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale, 2022.
