TU WIEN Informatics

Multimodal Modeling of Chest X-rays

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Daniel Blasko
Matrikelnummer 12202215

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Inf. Dr.rer.nat. Thomas Lukasiewicz
Mitwirkung: Maxime Kayser, BSc

Wien, 27. November 2024

Daniel Blasko	Thomas Lukasiewicz

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Multimodal Modelling of Chest X-Rays

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Daniel Blasko
Registration Number 12202215

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Inf. Dr.rer.nat. Thomas Lukasiewicz
Assistance: Maxime Kayser, BSc

Vienna, November 27, 2024

Daniel Blasko	Thomas Lukasiewicz

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Daniel Blasko

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 27. November 2024

Daniel Blasko

Acknowledgements

The work I present in this thesis would not have been possible without the support and guidance of many people to whom I am genuinely thankful.

First, I would like to sincerely thank Maxime Kayser, who actively supervised my work, for his continued support, availability and invaluable advice throughout this project. Your dedication to meeting with me regularly and whenever it was needed, and your patience in discussing ideas and helping me navigate challenges, made all the difference. You inspired me to go the extra mile and to always aim higher, and your guidance has truly taught me a lot. I also want to thank Professor Thomas Lukasiewicz for his supervision, which helped shape this thesis into a body of work that I am proud to present. I greatly appreciate the oversight and feedback you gave me.

Additionally, I would like to thank my family and friends for their encouragement along this journey. To my parents, a particular thank you for the constant support that you gave me in every possible way, it meant so much to me. Things would not have been possible without it, and your belief in me pushed me to persevere even when challenges arose. As to my friends, thank you for always being there to listen, to encourage me, and to remind me to find balance whenever I needed it.

This thesis is not just a reflection of my own efforts, but of the collective support and kindness of the people who were by my side. I am deeply grateful to all of you.

Kurzfassung

Röntgenaufnahmen des Brustkorbs sind ein zentrales Instrument in der medizinischen Diagnostik, doch ihre Auswertung stellt erhebliche Herausforderungen dar, die den Einsatz zuverlässiger computerunterstützter Verfahren notwendig machen. Diese Studie erforscht die Anwendung von Deep Learning zur Verbesserung der Interpretierbarkeit und Genauigkeit der Diagnosen von Brust-Röntgenbildern durch die Erzeugung von Erklärungen in natürlicher Sprache (NLEs). Bestehende einstufige neuronale Netzwerke erweisen sich zwar als effektiv, jedoch mangelt es ihnen oft an Erklärbarkeit, was ihre Akzeptanz in der Klinik beeinträchtigt. Zur Lösung dieses Problems stellen wir ein innovatives „Explain-then-Predict"-Modell vor, das die BLIP-2-Architektur mit einer Q-former-Komponente kombiniert, um NLEs während des diagnostischen Vorgangs zu erstellen und zu evaluieren. Im Unterschied zu bisherigen Methoden, die nachträglich Erklärungen liefern, ohne diagnostische Ergebnisse zu beeinflussen, nutzt unser Modell die NLEs, um seine Vorhersagen zu untermauern und zu rechtfertigen, wodurch die Erklärungen mit dem klinischen Denken harmonisiert und das Vertrauen in die automatisierte Diagnostik gestärkt wird.

Unser primäres Forschungsziel ist die Evaluierung, inwiefern ein Modell das Bild-Text-Kontrastlernen einsetzen kann, um treue NLEs zu generieren, die unmittelbar die Klassifikationsgenauigkeit verbessern. Wir entwickeln einen multimodalen Ansatz, der für jedes diagnostische Etikett NLEs erzeugt, die anschließend vom Q-former hinsichtlich ihrer Relevanz und Genauigkeit im Vergleich zum zugehörigen Röntgenbild bewertet werden. Dieses Modell wird end-to-end auf dem MIMIC-NLE-Datensatz trainiert und verwendet ein innovatives Trainingsregime, das die Erstellung von Erklärungen sowie deren Bewertungsgenauigkeit verbessert.

Empirische Ergebnisse belegen, dass unser Ansatz die Leistung der besten aktuellen Methoden zur Brust-Röntgenklassifikation erreicht und gleichzeitig Erklärungen bietet, die intrinsisch mit den diagnostischen Ergebnissen verknüpft sind. Dies fördert nicht nur ein tieferes Verständnis der Entscheidungsfindung des Modells, sondern steigert auch den praktischen Nutzen des Modells in realen klinischen Einsatzgebieten. Die Beiträge dieser Arbeit weisen auf eine vielversprechende Richtung für zukünftige Forschungen in der medizinischen Bildgebung hin, mit einem Schwerpunkt auf der Integration von aussagekräftigen Modellen, die sowohl die Interpretierbarkeit als auch die Genauigkeit diagnostischer KI-Systeme verbessern.

Abstract

Chest X-rays are a foundational tool for medical diagnostics, and yet interpreting them takes radiologists' time and is subject to challenges, prompting the development of reliable computer-assisted methods. This thesis investigates how interpretability of deep-learning-based chest X-ray diagnostics can be improved without compromising accuracy. It does so through the generation and utilization of natural language explanations (NLEs). Existing single-stage neural networks are clinically effective but often lack explainability, which limits their clinical adoption. To address this, we propose a novel "explain-then-predict" approach that leverages the BLIP-2 architecture and its Q-former component to generate NLEs and evaluate their relevance during the diagnostic process. Unlike previous methods that generate post-hoc explanations that do not affect the diagnostic outcomes, our solution incorporates the generated NLEs to guide its predictions, aligning explanations with clinical reasoning and enhancing explanation faithfulness by design.

Our work evaluates the extent to which an NLE-generating model can leverage image-text contrastive learning to measure how relevant a generated NLE is to an X-ray. We introduce a multimodal framework that generates NLEs for each diagnostic label, which are then assessed for relevance against the corresponding X-ray image by the Q-former. This model is further trained end-to-end to refine both the generation of explanations and the diagnosis accuracy.

Empirical results show that our approach matches state-of-the-art chest X-ray classification performance, while also providing explanations that are intrinsically tied to the diagnostic output. This allows to get an understanding of the model's outputs while enhancing its utility in clinical settings. The contributions of this work suggest a promising direction for future research in computer-assisted medical imaging analysis, focusing on the integration of explanatory models that can enhance both the interpretability and accuracy of such deep-learning-based diagnostic systems.

Contents

CHAPTER 1

Introduction

Chest X-rays are the most frequently prescribed type of chest radiologic diagnostic tool due to their low cost and invasivity, but their interpretation remains challenging even for experienced radiologists [1]. Computer vision-based X-ray diagnosis tools can help reduce misdiagnoses [2] and ease radiologists' workload by automating some parts of their workflow. The most common approach for this are single-stage neural networks that directly map an input X-ray image to diagnosis labels [3].

1.1 Motivation And Problem Statement

Single-stage classifiers, despite their success in medical image analysis, suffer from a lack of explainability for their predictions, which limits their adoption in real clinical settings. Explainability is crucial in such environments, where trustworthiness, transparency, and an understanding of model decisions are essential for safe and effective usage [4]. In the absence of explanations for their predictions, models are difficult to trust, particularly when the consequences of their errors can be life-threatening.

The opacity of such models also does not allow to detect and mitigate their potential biases that can for example be related gender, ethnicity, or socio-economic background and can have serious consequences [5]. For instance, if a model is biased towards certain demographic characteristics, it could result in misdiagnosis or inappropriate treatments for underrepresented groups, leading to significant ethical and medical challenges.

Different explainability techniques have been designed to probe black-box models. A popular approach for image classifiers, which are typically used for X-ray diagnosis, is to compute saliency maps that allow users to visualize the degree to which every pixel of an image contributed to the classification prediction [6]. While these methods illustrate which parts of the input most influence the output, a limitation to their usefulness is that they highlight *where* there model focused without justifying *why* those particular

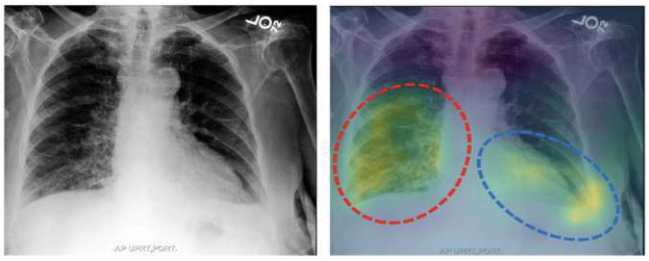

Figure 1.1: Example of how attention maps can be visualized for chest X-ray processing models in [7]. The attention focuses on an area of pulmonary edema (encircled in red, left) and on small pleural effusions (encircled in blue, right).

areas were relevant. This limitation makes it challenging to assess whether the model's focus is medically valid or if it is relying on spurious correlations.

An example of such spurious correlations arises in [8], where a skin lesion classifier was found to associate the presence of rulers in images with malignancy, leading to incorrect conclusions. Similarly, [9] demonstrated that the use of gentian violet surgical skin markers in dermoscopic images caused a drastic reduction in a deep learning model's diagnostic performance, dropping its specificity from 84.1% to 45.8%. This emphasizes the dangers of using non-explainable models in clinical practice: a model can be basing its decisions on irrelevant artifacts, potentially being more harmful than beneficial, and this can go unnoticed without model explainability techniques.

Additionally, most current explainability methods for image classifiers, like saliency maps, are generated *a posteriori*, after the model made its prediction, independently from the training process. This lack of integration means that these explanations do not influence how the model is trained or how it forms its decision boundaries. In practice, this implies that models can still make predictions based on clinically irrelevant features and are not constrained to focus on clinically relevant areas during training.

In the context of chest X-ray diagnosis, the problem is compounded by the complexity of the images and the wide range of pathologies they can reveal. Chest X-rays are one of the most common diagnostic tools in clinical practice, used for detecting thoracic abnormalities and conditions such as pneumonia, pulmonary edema or pleural effusion for example. Making correct diagnoses requires detailed examination of subtle features that can significantly vary across medical conditions. That is why it is crucial that deep learning models not only achieve high accuracy but also provide explanations that align with clinical reasoning. For instance, to diagnose pneumonia, the model should focus on regions of the lung that show opacities or consolidation rather than irrelevant artifacts.

The motivation for this work stems from these limitations in available explainability techniques and from the urgent need for more trustworthy AI systems in healthcare. An effective explainable model should offer faithful, clinically meaningful explanations that constrain the model's learning process to focus on relevant features, thereby ensuring both interpretability and clinical validity.

1.2 Aim Of The Work

In the work described in this document, the overall aim is to generate chest X-ray diagnoses with explanations that do not suffer from the limitations described for saliency maps. The approach we take to this is inspired by [10], which started research in the direction of neural networks that generate explanations for their decisions in natural language. This was motivated by the observation that humans tend to learn from explaining concepts themselves, hinting that natural language explanation (NLE) generation could be beneficial to the overall task learning process [10].

This paradigm would be very desirable for chest X-ray (CXR) diagnoses, as radiologists are used to debating such explanations in radiology reports [11]. NLEs could enable radiologists to challenge a model's outputs more naturally, and to get a more intuitive understanding of what a model's prediction was based on. Moreover, unlike saliency maps, text-based explanations can clarify the *reasoning* behind a decision, not just attribute importance to input features. The task of generating NLEs for chest X-ray diagnoses has been introduced in [12] based on similar motivations. The authors publish the MIMIC-NLE dataset, composed of 38,003 image-NLE pairs, along with baseline NLE generation models.

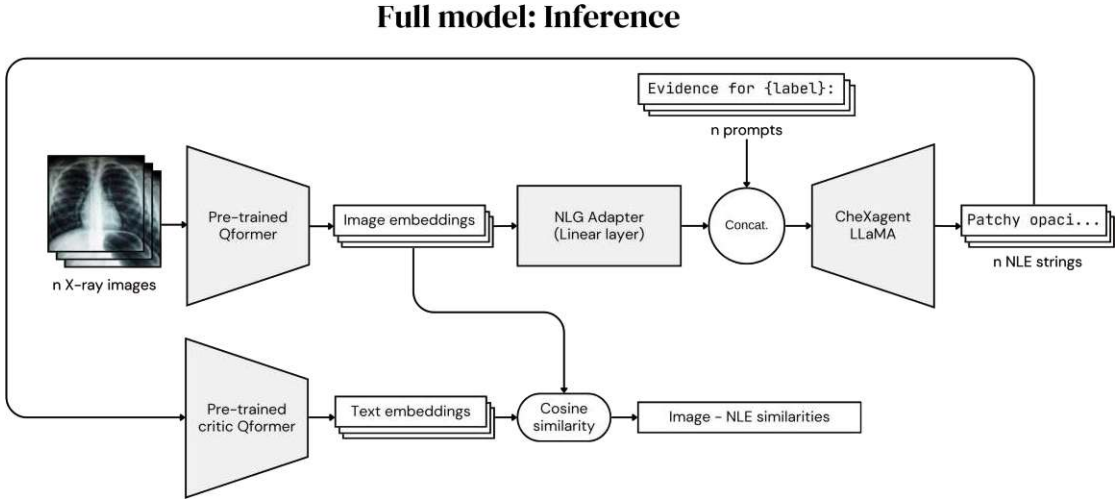

Figure 1.2: At inference, the model is prompted for an NLE for each pathology for a given chest X-ray image to autoregressively decode candidate NLEs. Those are fed into the critic model (Q-former) to compute their similarity to the X-ray image: the diagnosis is made based on the similarity to the image (faithful NLEs have higher similarities).

We are both the first to apply modern, large-scale vision-language models (VLMs) to this problem and to consider an "explain-then-predict" framework that encourages faithfulness of the NLEs. Figure 1.2 depicts the architecture of our model. We use modules from BLIP-2 [13], such as the Q-Former, and use those to generate natural language explanations for labels the model is prompted with.

To diagnose a given chest X-ray image, our model is prompted to generate a natural language explanation for each possible diagnosis label. All candidate explanations that have been generated are then fed back into the critic Q-former block of the model, which has been pre-trained to maximize the similarity of image and text embeddings of true image-NLE pairs and to minimize the similarity of false pairs. Thus, that critic outputs low similarities for NLEs that are irrelevant or unfaithful to the image and high similarities for explanations that accurately reflect the X-ray. This way, the final classification output is based on the explanation and its relevance to the input, and not vice versa. The explanations act as an information bottleneck in the prediction process, making the explanations for positive labels faithful by design, as the output is determined based on them and how they relate to the medical image.

By carrying out this work, the goal is to answer the following research question: *To what extent can image-text contrastive models improve classification performance and faithfulness in a self-rationalising approach to chest X-ray analysis?* To do so, we look into the three following research sub-questions:

- *To what degree can self-rationalisation match state-of-the-art chest X-ray classifiers?*
- *How effectively can a contrastive vision-language model capture NLE relevance?*
- *How much can a contrastive critic enhance NLE generation quality and faithfulness?*

1.3 Contributions

In the effort to answer these research questions, we make the following contributions:

- We detect that unfaithful NLEs, even those that have been generated without taking the chest X-ray into account at all, are convincing and lead to acceptable natural language generation metrics. We therefore highlight the need for a critic that can capture if an NLE is relevant to a chest X-ray to avoid generating unfaithful NLEs that go undetected.
- We demonstrate that multimodal models measuring image-text similarity between NLEs and CXR images are capable of separating true and false chest X-ray/NLE pairs when trained on an image-text contrastive objective for X-ray/NLE pairs, both for ground-truth and model-generated NLEs. We find that the Q-former component proposed in the BLIP-2 architecture can be used as an NLE critic given that it was pre-trained with an image-text-contrastive loss. We also validate the Q-former's critic capabilities and its benefit over some alternative vision-language models as it achieves better retrieval and critic performance than a CLIP [14] model trained on the same data and off-the-shelf medical CLIP models.
- Based on those findings, we propose a new "explain-then-predict" approach to CXR classification that makes predictions by generating candidate explanations in natural language for all diagnoses, and measuring their relevance to the chest X-ray to decide on the outcome. Our approach is based on the BLIP-2 architecture [13], as we find that its Q-former component's image-text representation learning capabilities lead to a better explanation critic than CLIP to judge if an explanation is true for a given X-ray. That Q-former can also be leveraged as part of the NLE generation pipeline that is based on the commonly used BLIP-2 vision language model architecture. We are the first to leverage this larger framework specifically for the task of NLE generation for chest X-rays, as well as to employ its Q-former component for a self-rationalized approach to classification. We show that this architecture is beneficial as it leads to better NLE generation performance than alternative, more straightforward image captioning approaches, and that it, to our knowledge, currently achieves the best NLE-generation performance on the MIMIC-NLE dataset. We also verify that training it is beneficial in comparison to using off-the-shelf CXR-specialized vision-language-models like CheXagent [15].
- Based on this, we leverage the critic capabilities of the Q-former and train the pipeline described in the previous point end-to-end. We introduce an explicit information bottleneck by generating explanations for each potential diagnosis, and making the classification decision based on the Q-former output similarity between the generated explanation for a given disease and the chest X-ray. By doing so, we match state-of-the-art chest X-ray classification performance [1] on our dataset while generating faithful NLEs by design.

[1] We consider the best performing non-ensemble approach on the CheXpert [16]

- We also show that representation learning performed for the NLE generation and chest X-ray classification tasks is beneficial to each other. A vision encoder pre-trained for NLE generation leads to better chest X-ray classification in a simple setting, and more importantly, in our full pipeline, we find that further propagating the critic (classification) loss to the natural language generation part of the network end-to-end is beneficial. By conditioning NLE generation on the critic's loss, we further improve classification performance of the full pipeline and outperform the state-of-the-art chest X-ray classifier.

1.4 Overview of the methodological approach

To design, train and validate the full model we introduce, we focus on the MIMIC-NLE dataset [12]. Built upon the MIMIC-CXR dataset [18], it contains 38,003 image-NLE pairs explaining the presence of thoracic pathologies and findings representing 14 different labels. We use the official split of the dataset where the three splits have the same class distribution, and contain 37,016, 273 and 714 images for the train, validation and test splits respectively.

The inference flow of the full model is illustrated in Figure 1.2. The model components are inspired from the BLIP-2 [13] architecture, but leveraged differently: a vision transformer, a Q-former (which is a BERT [19] encoder), a projection layer and a LLaMA-2 [20] language model. As the CheXagent vision-language model [15] is based on the same architecture, and pre-trained on a large corpus of chest X-ray related datasets and tasks, we initialize our vision encoder and language model with the CheXagent weights to take advantage of their pre-training. This does not lead to data leakage as they use the same official split of MIMIC-CXR and MIMIC-NLE in their dataset as we do [15]. The components of the full pipeline are trained on the MIMIC-NLE dataset in the following four stages.

1. **Stage 1 - Q-former, aligning representations of image and text**: In our first stage of training, we focus on training the Q-former component of the network. The goal of this step is to train the Q-former to align the representations of corresponding images (X-rays) and texts (explanations), to maximize the similarity of corresponding image-text embedding pairs while minimizing the similarity of false pairs. As illustrated in Figure 4.2, we do this by training the Q-former on a combination of the three same objectives as for BLIP-2 [13]: image-text contrastive, and image-text matching and image-grounded generation losses. We train on all image - NLE pairs contained in the MIMIC-NLE train split while preprending `"Evidence for {LABEL}"`, with the label of the given NLE, to the explanations.

multi-label chest X-ray classification benchmark (`https://paperswithcode.com/sota/multi-label-classification-on-chexpert`): DeepAUC [17]. We re-train the model, with parameter tuning, on our dataset to ensure fair comparability.

2. **Stage 2 - NLE generation**: The next goal is to learn to generate natural language explanations (text sequences) given an input X-ray image and prompt of the form "`Evidence for {LABEL}`". Just as in the second stage of BLIP-2 training [13] and as illustrated in Figure 3.2, we leverage the Q-former pre-trained in the previous stage to embed the image, add a projection layer ("adapter") to re-project the latent representation before concatenating it to the embedded prompt and feeding the entire sequence into the pre-trained autoregressive decoder language model. The pipeline is trained using a cross-entropy loss (a language generation task) while unfreezing the Q-former and adapter parameters, on all NLEs contained in the train split of the MIMIC-NLE dataset.

3. **Stage 3 - end-to-end training of the critic**: In the last two stages, we assemble all components into a pipeline that solves our main task, classifying chest X-rays while generating natural language explanations for positive labels. While Figure 1.2 describes the full pipeline that first generates explanations for every candidate label and then classifies based on the similarity of the explanation to the X-ray, this training stage focuses on further training the critic (Q-former) as shown in Figure 5.1. We consider all chest X-rays in the training dataset, and for every image, generate candidate NLEs for all labels, make the classification prediction based on them, and propagate the binary cross-entropy loss from the multi label classification task back to the critic Q-former that is kept unfrozen in the pipeline.

4. **Stage 4 - end-to-end training of NLE generation**: The final stage of training further conditions NLE generation on the critic (classification) signal. This is achieved by considering the same model pipeline and data as in Stage 3, but this time freezing the critic Q-former while training the parameters of the adapter block of the NLE generation part of the network on a combination of the binary cross-entropy loss (multi-label classification) and the cross entropy loss (language generation) as illustrated in Figure 5.2. As the sampling operation involved in decoding NLEs to feed into the critic would not be differentiable in the pipeline, we employ the Gumbel-Softmax trick to approximate the latent representation of the decoded sequences, and project them through a trained linear layer before feeding them into the Q-former. That linear layer aligns these representations with what would be the representation of the decoded sequence after tokenization with the Q-former's tokenizer.

The models obtained from each stage are evaluated on the MIMIC-NLE test split against different baselines and state of the art models:

1. **Stage 1**: To verify if the Q-formers image-text contrastive capabilities can be leveraged to detect false X-ray/NLE pairs, we measure the area under the curve (AUC) score for all possible pairs in our evaluation dataset, and we verify the separability with a Mann-Whitney U-test as well. We do the same thing for model-generated NLEs as well. In addition, we validate the Q-formers ranking (and thus

image-text representation) capabilities by measuring recall at 1, 5 and 10. To confirm that the Q-former is beneficial in addition to a simpler CLIP model trained with an image-text contrastive loss only, we compare these evaluation metrics with such a model trained on the same data.

2. **Stage 2**: We validate the NLE generation capabilities of our model by generating explanations for every positive label of all X-rays contained in the test split of the dataset, and measuring BLEU1, BLEU4, ROUGE1 and ROUGEL scores considering the ground-truth explanations. These scores are compared to those of baselines of different levels of complexity that we train: a language model pre-trained on medical data (GPT-2 or LLaMA-2) where images are embedded with a chest X-ray specific ResNet50 and passed through a projection layer before being prepended to the LM input, zero-shot prompting of CheXagent [15] or training a medical language model on the textual prompts only without inputting any image signal, to verify how such an unfaithful model would compare and behave.

3. **Stage 3 and Stage 4**: The evaluation of the final stage tests the full pipeline, and thus **evaluates how perform for the main task we want to solve**: chest X-ray classification. We generate NLEs for every label on each chest X-ray of the testing dataset. Based on the relevance of the NLEs to the X-ray images, we make final classification predictions, and measure per-class and mean AUC scores considering the ground truth labels. We measure those AUC scores for our different baselines and alternative models: simple classifiers (different vision encoders with a classification head), zero-shot prompting of CheXagent [15] and DeepAUC [17], the currently state-of-the-art non-ensemble chest X-ray classifier.

For further details on the methodology, setup and results for each of our training stages, please refer to Chapter 3, Chapter 4 and Chapter 5.

1.5 Structure Of The Thesis

To communicate our work and results, this document is organized as follows:

The reader is first familiarized with the main concepts involved in this work, how they have been covered in the literature, as well as with published work that is related to our project in **Chapter 2**.

Then, we describe how we approach the task of generating natural language explanations for positive labels on chest X-rays in **Chapter 3**, where we also identify how unfaithful explanations can still lead to good natural language generation evaluation metrics.

Based on this observation, we explore how similarity between image (X-ray) and text (explanation) representations can capture faithfulness of explanations to chest X-rays to build an explanation critic in **Chapter 4**.

In **Chapter 5**, we document how we integrate the NLE generator and the critic in a single pipeline to perform X-ray classification in an "explain-then-predict" approach. We also describe how training it end-to-end leads to the explanation generation model benefitting from the critic signal.

Finally, we summarize our key findings, highlight the limitations of our results and the future work they could benefit from in **Chapter 6**.

CHAPTER 2

Related Work And Key Concepts

The multimodal model introduced in this work involves techniques from the fields of explainable AI and vision, text, and multimodal representation learning. To describe the literature in these fields and introduce important concepts, Section 2.1 will summarize the history and current state of vision and language representation learning, including joint image-text modeling and their application in chest X-ray analysis. Section 2.2 will introduce explainable AI, summarizing post-hoc and self-explaining approaches to AI model interpretability. It will also cover natural language explanations: their generation, usage and evaluation. Finally, it will discuss the application of explainable AI to chest X-ray processing, highlighting unresolved aspects that inspired this thesis.

2.1 Vision and language representation learning

First, we discuss vision and language representation learning, which forms the foundation for many modern deep-learning applications by enabling machines to understand and integrate visual and textual information. These representations serve as the basis for tasks such as image captioning, visual question answering or multimodal retrieval, many of which are relevant to the topics discussed in this thesis. The aim is not only to provide a comprehensive overview of the literature but also to critically analyze the strengths and limitations of different methods. By understanding the historical context and current trends, we identify the gaps and opportunities that this thesis seeks to address. This foundational knowledge sets the stage for the subsequent sections, where these principles are applied to the specific challenges of vision and language representation learning in the medical domain, particularly in the analysis of chest X-rays.

2.1.1 In the natural domain

While deep learning is currently applied to a multitude of modalities, focus will be placed on learning representations of the two types involved in the work described in

this thesis: images (*in our case X-rays*) and text (*patient context, diagnosis explanations, X-ray reports...*). First, the literature about learning visual and textual representations independently will be described before elaborating on how they are unified in multimodal models that handle both text and images as part of their inputs and outputs.

2.1.1.1 Vision models

Learning good visual representations is essential to solve different downstream tasks on images, such as segmentation or object detection. While before 2012, signal processing techniques and filters were used to extract properties from images to create their representations, the AlexNet published in that year [21] jump-started the dominant use of deep learning for visual representation learning by significantly outperforming shallow approaches on the 2012 ImageNet challenge.

Visual feature extractors: Inspired by the performance achieved by the AlexNet, many convolutional neural network (CNN) architectures were created to learn ever better visual representations, such as the GoogLeNet [22] or the VGG-Net [23]. A major breakthrough was achieved in 2015 with the ResNet [24] that introduced the use of skip connections in CNNs. These connections sum the input signal of a layer to its output to create the final output. Among other benefits, this allows later layers to simply output 0 to forward the input when the network is too deep for a particular task, reducing the need to tailor the model size too precisely. This family of models is still widely used in contemporary approaches as a convolution-based visual feature extraction backbone.

An alternative architecture to CNNs for deep-learning-based visual representation learning was introduced in 2021: the Vision Transformer (ViT) [25]. While at the time of publication of [25], they were found to scale better than CNNs, some argue that hybrid models such as Swin-T [26] or MLP-Mixer [27] work better at scale, or that CNNs can be just as powerful as ViTs [28], while a different direction of research still believes that scaling pure vision transformers is more promising [29].

(Pre)training visual feature extractors: The authors of [30] empirically found scaling laws for computer vision that show that model, data and compute scaling provides consistent performance gains in diverse downstream tasks. As more compute became more accessible with time, supervised training on labelled data became a limitation as obtaining such data is expensive. This is why weakly- and self-supervised training has become widely used to pretrain large vision backbones.

For example, SimCLR [31] introduced a self-supervised training method where two views of the same unlabeled image are created (*through transformation*), and the representations of both views are constrained to be similar in the representation space (*contrastive approach*). Non-contrastive approaches have also been adopted to eliminate the need for negative samples, such as Bootstrap Your Own Latent [32], where two neural networks,

an online and a target network, are used to learn effective visual representations by predicting the latent representation of one augmented view from another. An alternative approach to self-supervised training of e.g. vision transformers has been inspired by masked language modeling: the masked autoencoder (MAE) [33]. Patches of the image are hidden, and based on the learned representation of the available patches, the network learns to predict the hidden ones through a decoder.

Using the pretrained backbones to solve computer vision tasks: Most pure computervision tasks solved by deep neural networks are one of the following.

- *Classification:* given a set of classes, an input image has to be classified into one or multiple of those classes.
- *Object detection:* given a set of classes, instances of those classes have to be detected in an input image, and each instance should be localized by a bounding box. This task is typically solved in the literature with models such as Faster R-CNN [34], Mask R-CNN [35] (CNN-based) or DETR [36] (ViT-based).
- *Segmentation:* given a set of classes, determine the class of each pixel of the image (thus creating masks of the different classes, optionally of their instances, in the image). This task is typically solved in the literature with models such as the U-Net [37], Mask R-CNN [35] (CNN-based), DETIC [38] or Segment Anything Model (SAM) [39] (ViT-based).

The different architectures used to solve these tasks generally have the same structure. A vision backbone, which is a network that learns visual representations (*as described in the previous subsection*) and a task specific module that takes the visual representation output by the backbone, and uses it as an input to the task specific network. In most settings, transfer learning is performed by taking a pretrained backbone that has already learned to extract high-level features, and the complete network (*with some layers potentially frozen*) is further trained end-to-end. The simplest example of this is in image classification, where the image representation is fed into a simple fully connected network that outputs class logits.

2.1.1.2 Language models

Similarly to the visual modality, strong language representations are the foundation to good performance on downstream natural language processing (NLP) tasks. While the transformer architecture introduced in 2017 revolutionized language modeling [40], different approaches can be used to learn text representations as well.

Text representation learning: Initially, shallow models relied on text feature extraction methods like bag of words or TF-IDF. But word embedding approaches like Word2Vec [41] and GloVe [42] democratized the learning of dense vector representations for words based on their co-occurrence patterns in text, allowing deep networks to encode semantic similarity between words. Similarly, methods like Skip-Thought [43] or InferSent [44] learned such representations of full sentences or documents through a process of predicting neighboring sentences instead of words.

However, the publication of the Transformer architecture in 2017 [40] started a wave of new, foundational models in textual representation learning. The "Bidirectional Encoder Representations from Transformers" (BERT) [19] leverages the transformer encoder block to generate context-sensitive embeddings where each token attends to the full sequence at each layer. This led to very powerful text representations, that power downstream tasks like retrieval or classification for example. This approach inspired other transformer encoder-only models such as T5 [45] or RoBERTa [46] which are still very commonly used text encoders.

Language modeling: Language modeling is the task of predicting the upcoming, most probable word in a sequence of text - most often iteratively to generate the full sequence. It is the problem behind many generative AI applications in the textual modality, and solutions have made great progress with the use of transformers in this field too.

Earlier approaches to language modeling relied on n-gram statistics to predict the most probable word, but these techniques failed to capture long-range dependencies in the sequences and the semantic relations between the words. This is why, just as for representation learning where transformers allowed to capture those relations better as we mentioned for BERT earlier, transformer-based language models were democratized after 2017. These language models are most often either decoder-only or encoder-decoder transformer architectures, and follow scaling laws that support how scaling compute, data and model size has a large potential, as shown for example by the Chinchilla scaling laws [47]. Thus, such models have been scaled to very large amounts of parameters, and have been pretrained on Internet-scale text corpora in a self-supervised manner similarly to vision pretraining, for example through masked language modeling and next token prediction [20]. Those models are then further instruction-tuned to follow user-instructions and to be aligned to human preferences, and optionally fine-tuned to domain specific data.

While transformer-based language models currently make state-of-the-art progress on an approximatively monthly basis, some large language model families that achieve excellent performance on diverse text-based tasks through prompting and few-shot learning only include ChatGPT (GPT-4) [48], Mistral [49], LLaMa (currently LLaMa 3) [20] or Gemini [50].

2.1.1.3 Vision-language models: fusing vision and language representations

Some tasks however require inputs of both text and image modalities. This is why some families of models aim to learn representations of image and text in a common representation space, and to combine the representations of image and text for downstream tasks where the input and output are of both modalities.

Image-text joint representation learning: When learning aligned text and image representations, typically for vision-language pretraining, the objective is that representations of images and texts referring the same thing end up close to each other in the latent space, while unrelated embeddings should be as distant as possible. This idea is at the core of contrastive learning, which was democratized when its effectiveness at scale was proven at the publication of "Contrastive Language-Image Pre-training" (CLIP) [14]. It introduces a learning objective and sampling strategy to pre-train an image and a text encoder jointly.

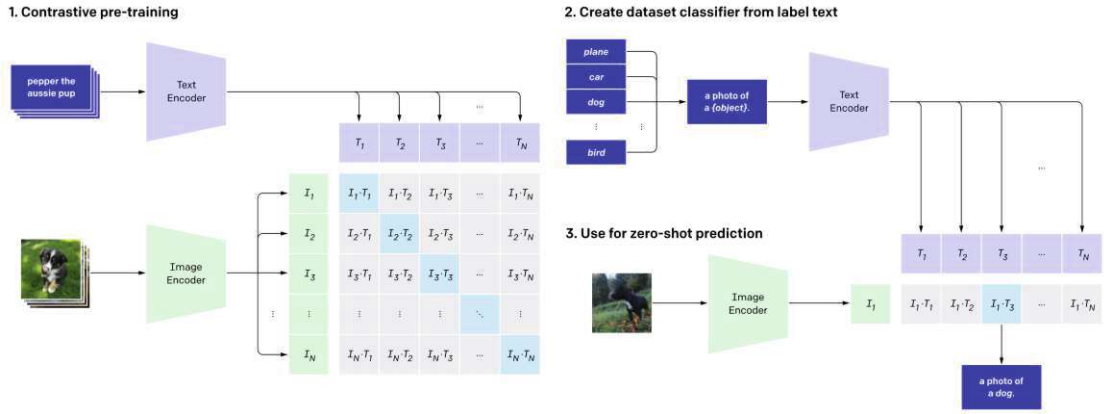

Figure 2.1: Overview of the CLIP pretraining strategy. Source: `https://openai.com/research/clip`.

As illustrated in Figure 2.1, for each batch, n image/text pairs are sampled from the pre-training dataset. All images and texts are processed through their corresponding encoders. Then, the dot product between all image and text embeddings is computed, measuring their similarities. The learning objective aims to maximize the similarity of the true image/text pairs, while minimizing the similarity to the rest of the batch. This pushes embeddings of corresponding image and text to be similar. The authors of CLIP observed that this training strategy scales to billions of image-text pairs very efficiently, enables competitive zero-shot transfer to image classification on unseen classes, and leads to vision backbones that compete with state-of-the-art label-supervised ones [14]. Alternatively, a very competitive joint representation is obtained through a similar approach based on a pairwise sigmoid loss, eliminating the need for the overall pairwise similarities of the batch for normalization, in SigLIP [51]. The removal of that constraint

unlocks larger batch sizes on equal compute while also training better with smaller batches [51].

More data efficient approaches to such image-text pretraining have also been proposed. [52] introduce more data-efficient contrastive language-image pre-training through joint learning with self-supervision. For example, [52] introduces "Multi-View Supervision" where it uses text and image augmentation to take advantage of positive pairs in addition to other self-supervised objectives.

Finally, [53] proves that a competitive joint representation can be learnt through captioning as a pretraining task rather than contrastive learning, and that contrastive learning is not the only way to learn aligned image-text representations. Instead of training a vision and text encoder jointly with a contrastive loss, they train the vision transformer encoder with a transformer decoder tasked to generate captions for the image, using a cross-entropy loss. During pretraining, the model switches between two caption generation modes: the captions are either generated autoregressively with a causal self-attention mask, or through parallel decoding, where the mask forces the decoder to predict the full caption based on the image exclusively. This pretraining strategy leads to vision backbones that match or outperform equivalent CLIPs in few-shot and supervised classification, and outperform them in captioning or visual question answering tasks [53].

Model architectures for multimodal tasks: Common multimodal tasks include multimodal classification (*classifying based on image and text inputs*), visual question answering (VQA) (*answering natural language questions based on an input image*), multimodal chatting (*chatbots that support image inputs too*) or image captioning (*generating natural language descriptions of input images*).

One approach to solving these tasks is to train multimodal models from scratch. The architectures used for these tasks are similar, with the exception of classification tasks. The image captioning task for example has long been solved through architectures trained specifically for this multimodal task of outputting text based on an input image. Popular research for image captioning [54], [55], [56], [57] relied on feeding the output of a CNN encoder into a recurrent neural network (RNN) decoder to generate the captions sequentially. More modern approaches use a transformer image encoder and text decoder in a similar setup [53]. Alternatives to this encoder-decoder approach have also been proposed, such as for example the fusion-encoder [58] or the dual encoder [14].

But pretraining from scratch on image-text is hard to scale due to the limited availability of such data. For this reason, many methods take pretrained vision and language models, optionally freeze some of their parts, and further train them together by combining them in different ways in order to take advantage of the capacities learned on pure image and text pretraining respectively [59] [60]. Different approaches are used to combine the representations from both modalities. The overall idea is to have modality-specific encoders, and feed their output into a fusion module. This module combines these latent

representations and aligns them to the output modality. A common approach for the fusion module is to use a weighted sum of the latent representation's features with learnt weights, while an alternative is to simply concatenate the two feature vectors and pass them through a fully connected network, sometimes including attention mechanisms, for reprojection. The output vector of this fusion module is then fed into the task specific model used for classification or language generation for example. However, fusion can happen at different stages.

Some specific approaches to combining pretrained modules to create vision-language models that proved excellent scaling and performance include Flamingo [61], LLaVA [62] and BLIP [63].

- Flamingo adds cross-attention layers in the pretrained LLM to inject the visual features. These layers are trained on a large image and text dataset with a language modeling loss [61].

- Alternatively, LLaVA combines a pre-trained LLM and vision encoder with a vision-language connector network [62]. This is a lightweight fully-connected network that learns to project the visual representations into the LLMs input embedding space. The model, including this network, go through supervised training on image-text pairs to learn this projection through VQA and image captioning tasks.

- "Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation" (BLIP) [63] is an approach to leveraging pretrained vision and language models to train a vision-language model that can be prompted for different vision-language-tasks that was introduced in 2022. It introduced a pretraining framework that combines three kinds of losses at pretraining. Image-text contrastive (ITC) involving the image and text unimodal encoders to encourage similar latent representations of related images and captions, image-text matching (ITM) involving the image-grounded text encoder to learn fine-grained alignment between vision and language, and a captioning loss [63].

- BLIP-2 [13] was proposed the following year, taking a different approach. It bootstraps pre-trained vision encoders and language models by aligning the visual representations to the language model's representation via a Querying Transformer ("Q-former"). This module is a lightweight transformer architecture (originally a pretrained BERT model [13]), that learns textual queries which attend to the image latent in order to generate an image representation that is relevant to the language model. This allows a very parameter-efficient training of the pipeline in comparison to the size of the involved models, leading to beating state-of-the-art performance with significantly less trainable parameters [13]. The model is trained in two stages. The first stage bootstraps vision-language (VL) representation learning by freezing the vision encoder and training the Q-former with the three combined losses described for BLIP, while the second stage bootstraps vision-to-language

generation from the frozen language-model by training the full pipeline end-to-end with a text generation objective.

Other large scale vision-language models (VLMs) that also bootstrap such pretrained modules and that have been scaled enough on curated data to develop a perception and understanding of text and images include Qwen-VL [64], PaLI-X [65] or CogVLM [66]. Evaluating such multimodal is not fully unified yet, though benchmarks like SEED-Bench [67] and MME [68] are being adopted. These are datasets of manually generated instruction-answer samples to limit data leakage to foundation model pretraining.

2.1.2 In the medical domain

Medical diagnosis is the process of attributing a pathology to a patient's symptoms. To do this, medical doctors take the patient's context into account (*medical history, lifestyle, symptoms...*) and enrich it by requesting different sorts of medical imaging (*MRI, fMRI, CT, X-ray...*). Since this thesis works with chest X-ray (CXR) images, this Section will mainly focus on X-ray imaging data.

X-rays are numerically available as images, while patient context can be represented as (*optionally structured*) text. Therefore, models that leverage patient context and/or X-rays need to learn and use representations of the two modalities discussed in the previous Section 2.1.1 for the downstream tasks on (chest) X-rays that are introduced in the following subsections.

2.1.2.1 Chest X-ray classification for automated diagnosis

Due to the cost-effectiveness and low invasiveness (low doses of radiation in comparison to CT), X-rays are very commonly prescribed to diagnose chest pathologies. However, accurate reading of this imagery requires years of training and medical experience. This is why research has been pursuing generating diagnoses of chest X-rays automatically through different deep-learning-based approaches.

The most common approach to automated CXR classification in the literature is to use an image encoder (e.g. a CNN) and feed its output into a classification module (most often a fully connected network). This was introduced in [69] for chest X-rays, where such an architecture was leveraged to classify view orientations of CXRs. [70] then used the same approach to classify pulmonary tuberculosis. From there on, publications like [71] or [72] benchmarked different CNN architectures for this task, and introduced different architectural tweaks and data augmentations to improve performance. Alternatively to CNNs, [73] introduced attention-driven spatial transformer networks to classify anomalies in CXRs. LT-ViT [74] also obtained competitive performance on multiple CXR classification datasets by using a vision-transformer as encoder [74]. While some took advantage

of ensembling ([75], [76]) to further improve performance metrics, the overall approach to solving the task remained the same by adding classifiers on top of vision backbones.

The CheXpert [16] benchmark can be considered to compare the performance of the most recent approaches to multi-label classification on chest X-rays in the literature. On the 14 abnormality labels present on the dataset, the ten best performing publications obtain an average area under the curve (AUC) score of 0.933 to 0.928 [77], making them all very similarly effective in practice. While seven of those 10 publications use ensembling techniques to optimise these performance scores [78] [79], DeepAUC-v1 [17] and two versions of Conditional-Training-LSR [78] obtain a 0.93 AUC score with a single model that achieves the current state-of-the-art performance in multi-label chest X-ray classification. [78] obtain this through a CNN that leverages dependencies among abnormality labels and label smoothing as uncertain labels are frequent in the benchmark, while the approach of [17] is based on optimizing a CNN for AUC maximization through a margin-based min-max surrogate loss function. The work of this thesis includes a chest X-ray diagnosis (as classification) task as we generate natural language generations for each label, and a critic selects the relevant explanations, implicitly selecting positive labels. Thus, we will compare our classification performance to these aforementioned state-of-the-art approaches to CXR classification.

Alternatively to pure image classification, some research efforts approached X-ray diagnosis by including localization information in the diagnoses. [80, 81, 82, 83] did this as an object detection task, identifying bounding boxes delimiting areas corresponding to a given class (e.g. abnormality label in the case of CXRs). Strongly supervised approaches used algorithms like YOLO variants or Faster R-CNN to localize abnormalities in chest X-rays [84]. More recent works like AnaXnet trained two blocks independently, a Faster R-CNN that localizes anatomical regions, and a graph convolutional network that classifies the presence of abnormalities in each bounding box (multiclass classification) [85]. Due to limited availability of bounding box annotations in X-rays, works like [82] [81] [80] used weakly- or semi-supervised learning to solve this task. Similarly, [86, 87, 88] tried solving this as a segmentation task, predicting the membership of each pixel to abnormality classes. For example, [87] used a DenseNet architecture to segment cardiomegaly in chest X-rays while [86] used a U-Net structure for this segmentation task.

However, these different approaches to chest X-ray diagnosis present some flaws that limit their adoption in medical settings. The deep-learning models lack explainability for their predictions, leading to confirmation bias in doctors relying on them and difficulty to understand their mistakes [5]. Moreover, as they are not constrained on the information from the chest X-rays they use to make their predictions, they often use irrelevant areas of chest X-rays (such as artifacts specific to some imaging hardware used in parts of the dataset) that a professional would not have used to make their diagnosis [12]. Attempts to solve these limitations will be presented in Section 2.2.3.

2.1.2.2 Medical vision-language foundation models and chest X-ray report generation

Biomedical language models such as PubMedBERT [89], [90] or Med-PaLM [91, 92] have proven that large language models trained or fine-tuned on medical text corpora led to very competitive performance on medical question answering problems such as the US Medical Licensing Examination where Med-PaLM 2 obtained state of the art performance [92].

As most medical tasks involve images and text (*in many cases, a medical professional includes medical imaging in their diagnosis process*), similar medical foundation models trained image and text data could be leveraged for many downstream tasks.

Medical vision-language foundation models: One approach to building such foundation models is to train as general and versatile medical models as possible. The most notable in this category are Med-Flamingo [93], LLaVA-Med [94] and Med-PaLM M [95]. These are generalist vision-language models that have been trained on different types of imaging and corresponding medical text including dermatology, pathology or radiology. This allows for broad downstream use with little to no further training on various medical tasks. Med-Flamingo [93] is an OpenFlamingo model [96] trained on medical textbook images and text for few-shot visual question answering (VQA), while LLaVA-Med [94] is an adaptation of the previously described LLaVA [62] on multimodal data from PubMed. Further instruction tuning allow the use of this foundation model as a medical generalist chatbot [94]. Med-PaLM M [95] is a version of PaLM-E [97] fine-tuned on a large biomedical dataset, leading to a foundation model supporting different types of medical images, text and genomics. Similarly, XrayGPT trains a combination of the vision encoder from MedCLIP and a pre-trained Vicuna LM on medical data.

Image-text contrastive foundation models have also been trained on medical or even X-ray-specific data. For example, MedCLIP [98] is such a model that was trained on unpaired medical image and text data. For chest X-rays specifically, GloRIA [99] learns to align embeddings of words of reports and regions of images, while BioVIL [100] combined a similar local alignment strategy with a classical contrastive global alignment between the full chest X-rays and reports. The currently best performing CLIP-based model for chest X-rays is CXR-CLIP [101], which obtains this performance through diverse data-efficient vision-language pretraining techniques, a broader dataset including an image-label dataset where labels are augmented to natural language prompts, and the use of a combination of image, text, and image-text contrastive losses in each batch.

Some foundation models are more specialized for better performance on a specific family of tasks, while still being more general than task-specific models. For example, some foundation models focus on different CXR tasks through a single model. ELIXR [102] is such a model that aligns a CXR-specialized image encoder to a medical LLM (PaLM 2 [92]) for tasks like medical VQA.

MAIRA-1 [103] is an alternative state-of-the-art radiology-specific multimodal foundation model that takes a similar approach. It aligns a CXR-specific image encoder with a Vicuna-7b LLM previously fine-tuned on medical text, leading to state-of-the-art CXR report generation performance metrics [103].

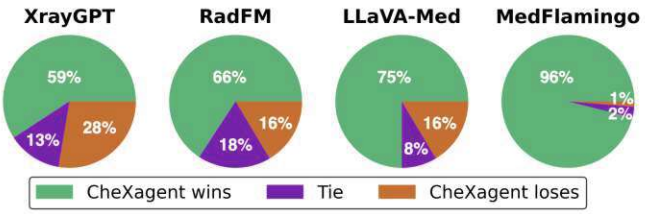

Figure 2.2: GPT-4 evaluation of reports generated by CheXagent and competing medical foundation models on the MIMIC-CXR dataset. Source: [15, Figure 3].

Finally, a recently published and competitively performing CXR foundation model is CheXagent [15]. This is obtained by first pre-training a clinical LLM on CXR reports, a vision encoder on CXR images, and then training a bridger network between those two components. This bridger follows the BLIP-2 approach described in Section 2.1.1.3. A Q-former bridging the two networks is trained in two stages, with the difference that the first stage includes only image captioning and image-text contrastive losses, while BLIP-2 also uses an image-text matching loss. The model then performs competitively on a large array of CXR-related tasks: view classification, binary disease classification, single disease identification, multi diseases identification, visual question answering, image-text reasoning and report generation [15]. Figure 2.2 illustrates how this CXR-specialized model most often beats other multimodal medical foundation models mentioned in this section.

Chest X-ray report generation: Physicians communicate their findings and observations from medical imaging through medical reports. The production of such a report, on average, takes five to ten minutes [104]. To reduce this time, one of the main tasks pursued through deep learning on chest X-rays is report generation. The goal is to assist radiologists in making faster and accurate diagnoses by generating CXR reports in natural language based on an input CXR image. A successful solution generates clinically accurate reports that correctly describe the patient's condition and symptoms, with fluent, realistic and human-readable language [105].

Historically, this task was most often solved with report-generation-specific models, similar to those used in image captioning. Early approaches like [106, 107] relied on CNNs to extract features from CXR images that were then fed into recurrent neural networks (RNNs) to generate the natural language reports. Attempts at improving this overall architecture have been made based on different ideas. [108] and [109] classify the top-k most probable diseases in the X-ray in a first stage, and then feed that information into the decoder for more accurate reports. Works like [110] and [111] try to alleviate

the fact that RNNs are not best at generating long sentences and paragraphs [112] by replacing the RNN with a Transformer-based decoder, while [113] does so through a hierarchical RNN architecture. To integrate prior medical knowledge in report generation models, [114, 115, 116] built knowledge graphs to guide report generation with that knowledge.

A different direction of improvement on these report-generation-specific architectures is through the addition of a visual bottleneck, as recently done in [117, 118]. RGRG, introduced in [117], builds upon the typical encoder-decoder architecture by addiction region-of-interest (ROI) detection and selection stages. The CXR image is input into a Faster R-CNN object detector that was trained to output bounding boxes and embeddings for 29 anatomical regions in chest X-rays. Those regions are then fed into a region selection module (a binary classifier) as well as an abnormality classifier at training to include the abnormality information in the learnt embeddings. Only the four regions with maximal logits output by the region selection classifier are fed into the transformer-decoder that generates sentences for each ROI independently, forming the CXR report altogether. This approach allows the model to focus on relevant and important areas of the X-ray only, which led to state-of-the-art clinical efficacy and natural language generation metrics at the time of publication. [118] built upon this idea leveraging a similar anatomical region detecting Faster R-CNN, but using a finding classifier head instead of a region selection one. Thus, each bounding box is assigned an anatomical region label as well as diagnosis labels. To generate the report, these regions of interest and their labels are encoded as triples in natural language (*for example "Opacity LOCATED_AT Spine"*), and the top-k region embeddings are concatenated to their corresponding embedded triples to form a single sequence fed to the transformer-decoder for report generation. The main difference to RGRG [117] is that the decoder takes all selected regions of interest as an input for full report generation instead of generating independent sequences for each region, allowing it to reason over the different findings. This led to a slight improvement in clinical efficacy and natural language generation metrics over RGRG [118].

But increasingly often, medical multimodal foundation models such as those described at the beginning of this Subsection are used for this medical image-to-text task. Designed for natural language generation based on image-text input, these models generalize well and turn out to often outperform the task-specific architectures described in the previous paragraph. For example, the previously described MAIRA-1 and LLaVA-Med foundation models [103], trained for multiple CXR-related task, outperform the best performing CXR-specific architectures such as [117, 118] both in clinical efficacy and natural language generation [103, 94].

Evaluating CXR report generation: To compare these different approaches to chest X-ray report generation, two families of metrics are used: **clinical efficacy** (CE) metrics that measure how good the model is at finding the medical symptoms and pathologies in the X-ray, and **natural language generation** (NLG) metrics that measure how fluent,

realistic and close to human writing the generated text is.

- **Clinical efficacy metrics**: These metrics are those that would be used for the underlying medical task without language generation. In the case of report generation, the reports communicate the findings of diagnoses and symptoms, which is a (multi-label) classification task. Hence, the clinical efficacy metrics used are classification metrics like weighted AUC score, F1 score, or accuracy.
- **Natural language generation metrics**: The metrics used to measure how fluent and realistic the generated text is are those usually used to evaluate language models and image captioners. They include the following metrics.
  - *BLEU*: Based on the ratio of common 1-to-n-grams between the ground-truth and generated text. For example, BLEU-3 takes anything between unigrams and trigrams into account. The reported metric is usually the average BLEU score across all generated texts.
  - *ROUGE-L*: Based on the non-consecutive word longest common subsequence (LCS) between the ground-truth and generated text. The longer the common subsequence, the higher is the ROUGE-L score. Similarly to a F-score, ROUGE-L is computed using the precision and recall based on the LCS.
  - *CIDEr*: Based on the overlap in words between the ground-truth and generated text. Computes the TF-IDF weight vector of unigrams to 5-grams of the two sequences, and then the cosine similarity between the four vector representations of both texts. The CIDEr score is the mean of those four similarities.
  - *METOR*: Measures the similarity between the generated and ground-truth text considering exact word matches, synonym matches, stemming matches, and paraphrase matches. Additionally, it incorporates precision, recall, and alignment-based scoring, penalizing differences in word order to provide a more nuanced evaluation than for example BLEU.

2.2 Explainable AI

Literature from the domain of psychology has highlighted how humans rely on explanations for learning by building inferences to enrich their prior knowledge [119]. Meanwhile, [4] highlights how explainability is essential for intelligent systems to be trusted, especially in medical settings. Explainability is also needed to detect biases like racism or sexism that AI systems can develop [120]. While this shows how generating explanations for decisions made by neural networks is important, as deep learning models grow larger, it gets increasingly difficult to explain how they used their input to make their prediction. We therefore need to make sure that these systems are interpretable to attain their widespread use and public trust. In this section, we will describe how the explainability of modern neural networks is approached in the literature. We will begin by describing

the two categories of approaches to explainable AI (XAI), then discuss natural language explanations (NLEs), to finally explore how these XAI concepts are used in deep learning applied to chest X-rays.

2.2.1 Different approaches to explainable AI

Making neural networks explainable is approached in two fundamental ways in the literature. *Post-hoc explanation techniques* are the most common. They are independent of the model training, and aim at explaining predictions of a trained target model by probing it in different ways. The other approach to explainability are *self-explaining* models. These are models designed to generate explanations for their predictions at inference, and that were trained with that explainability objective as well.

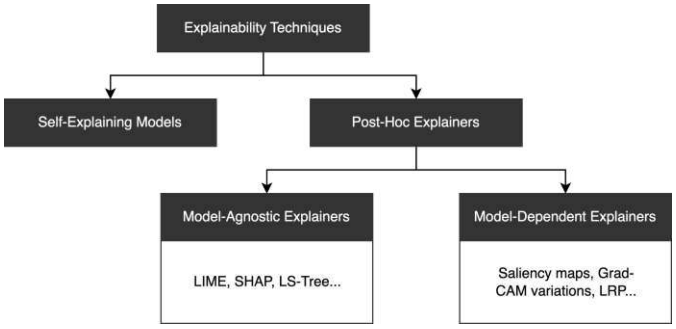

Figure 2.3: Categorization of the explainability methods mentioned in this section.

2.2.1.1 Post-hoc explainers

Post-hoc explanation generation is a family of techniques that are used to explain the predictions of already trained and frozen models. They are completely independent from training, and thus, do not constrain the model to take these predictions into account in its training objective. An example of such methods is LIME [121], where an independent, explainable model, such as linear regression, is trained on a neighborhood of the model prediction it explains. In this section, we will mostly focus on post-hoc explainers for visual and multimodal models, as this thesis focuses on visual input (a chest X-ray image).

These post-hoc explainers can further be divided into two subcategories, as illustrated in Figure 2.3. Model-agnostic explainers solely rely on calling the model on various inputs, while model-dependent explainers are based on having access to the model's architecture and trained weights. Common examples of the first category are LIME [121], KernalSHAP [122] or the more recent LS-Tree [123]. These have the advantage of being applicable on a broader amount of models as they are independent from their architecture. Notable explainers from the second category are LRP [124], Grad-CAM

[125], Deep- and MaxSHAP [126] as well as saliency maps [127]. The latter are obtained by computing the gradient of the class of an input sample and deriving the saliency maps from a first-order Taylor expansion of the image [127]. While they are not applicable as often as model-agnostic explainers, they have the potential to generate more accurate explanations: explainers that only probe the models through different inputs can infer correlations between the inputs and outputs that do not have to represent how the model works inside, while these model-dependent explainers often base their explanations on the model's inner weights and activations.

These techniques base their explanations on features of the model input. Most often, they quantify the importance of each feature for a prediction on a given input, or select a subset of the features that played a significant role in put prediction. In the case of images, these explanations are usually at the pixel level, while for natural language, they are usually at the granularity of tokens. An explanation based on importance weights for a natural language sequence would be a vector of importance scores of the length of the sequence, or a heatmap for an image. A subset-based explanation would be the important tokens of the sequence in the case of a natural language sentence.

2.2.1.2 Self-explaining models

Self-explaining models take a different approach to explainability in comparison to post-hoc explanation methods by designing the models to also output explanations for their predictions. These are models that have an explanation-generating module as part of their architecture. Typically, they are composed of a predictor module (that generates the task specific output) and an explanation generation module (that also takes the predictor module output as an input) [128]. Due to this design, the main difference with post-hoc explanation approaches is that generating explanations is part of the training objective of the model. However, there are also exceptions in such models that are solely supervised on the task-specific prediction and do not include the explanation generation in the training objective, as for example [129], or counterfactual explanations that are post-hoc explanations describing the minimal amount of modifications to be made to the input to change the model's prediction [130].

We group the predictor module of self-explaining models in two categories: "predict-then-explain" or "explain-then-predict". We will refer to models that generate a prediction and feed the prediction, optionally along other signals like the input or outputs of inner layers, as "predict-then-explain" (PTE). In contrast, "explain-then-predict" (ETP) self-explaining models are models that first generate explanations for different possible outputs, and then take those potential explanations into account to make their final prediction.

Predict-then-explain: Most rationalized models are PTE approaches to self-explaining AI. In these models, such as [131, 132, 133, 134, 135], explanation generation is an extension to the existing task architecture. A task-module predicts the task-specific output, and subsequently, an explanation generation module generates the explanation

in the form of a subset of the input or natural language for example [136]. The potential flaw of these approaches is that the explanation is still generated after the choice of label, possibly leading to unfaithful explanations that justify wrong outputs in a persuasive way. In clinical settings, this would have a high potential of inducing confirmation bias.

Explain-then-predict: In ETP models, the explainer module is trained to output an explanation based on a premise (input) and hypothesis (one of the potential task-specific outputs). The task-specific output, such as a label, is then predicted by a module that never gets the full input, and makes it prediction only based on the explainer's output (the candidate explanation). This creates an information bottleneck for the task-specific prediction and forces the explainer to generate explanations that contain relevant and faithful information. While in some works, such a setup led to a decrease in task-specific performance metrics in comparison to PTE [137, 131], the generated explanations were more faithful and useful.

This configuration also has the theoretical advantage of allowing the model to potentially "reason" over explanations before outputting its prediction. This is similar to what was recently achieved with **Chain-of-Thought (CoT) prompting** [138] in LLMs, where allowing the language model to output its reasoning before generating the final answer led to better behavior in some downstream tasks [138]. Similarly, a vision-language ETP model, ReVisE [139], improves its explanations for VQA iteratively through such an "explain-then-predict" approach that reminds of CoT prompting: a BLIP-2-like architecture generates grounded answers based on an image and natural language question as well as a rationale (explanation) for that answer. Iteratively, a new answer is generated taking the output rational into account as an input, until convergence where the answer does not change anymore. One could also argue that some of the report-generation models cited in Section 2.1.2.2 also are ETP approaches. As described in Section 2.1.2.2, RGRG [117] and the architecture introduced in [118] first find abnormal areas in the chest X-ray and generate reports based on this subset of the input only. The selected abnormal areas can act as visual grounding, a form of explanation, and the task-specific output is generated based on that explanation only. While these explain-then-predict systems are rarer than predict-then-explain ones, they will be discussed in further detail and with more examples in Section 2.2.2.

2.2.2 Natural Language Explanations

A particular type of explanations generated by self-explaining models are **natural language explanations (NLEs)**. These are sentences in natural language that provide arguments supporting a prediction the way a human would. For example, a NLE for a classifier labeling an image as "animal" in a binary setting could be "Four legged silhouette with brown furs. Humans do not have fur and stand on two legs". Such explanations are very valuable as they could describe the model's reasoning in a human-intelligible way, even to people that are not familiar with AI systems.

2.2.2.1 Generating and leveraging natural language explanations

Generating NLEs: [10] and [140] spurred research in the direction of neural nets generating explanations for their decisions in natural language, inspired by the fact that humans learn from explanations and examples, but also from creating explanations themselves. This research direction was further supported by the creation of datasets containing natural language explanations. The ACT-X and VQA-X datasets [141] contained natural language and visual explanations for VQA and visual activity recognition, and were published with the PJ-X model that generates a prediction with feature-based and natural language explanations. The BDD-X dataset [142] contains natural language explanations for choices made by a self-driving car and was used to train a car-controlling model explaining its choices in natural language. In most cases, the architecture is task-specific, with the addition of an explanation-generation module which is a language model like GPT-2 [143] which takes the model input, along with the task-specific output in some cases, as an input for NLE generation.

But NLEs are not only used in pure natural language processing as in [137], or in visual question answering. Intelligible explanations for visual reasoning are particularly important in real-world settings [144], as [145] showed for assistive technologies or [146] for interactive learning for example. [147, 148, 149] built models generating NLEs in computer vision settings, while [12, 150, 151, 142] did so in multimodal ones. For example, the e-UG model [136], based on the works of [152], uses the UNITER vision-language transformer-based encoder as a task-specific-module that predicts a label. The intermediate representations based on which the label was generated are fed to the explanation-generation module which is GPT-2 [143] language model.

Additional benefits of integrating NLE generation: Including NLE generation in an architecture can be beneficial in more ways than just by improving transparency. For example, training with NLEs can also enhance model training. Adding natural language explanations of reasoning to a model's training data is a way of integrating further external knowledge into the model, as it also learns the content of these explanations. This was for example leveraged in [153] where NLEs generated by GPT-3 [154] were used to improve the reasoning capabilities of smaller language models. Additionally, [154] introduced the ability of LLMs to perform in-context learning (ICL), which is making

accurate predictions on data that was not seen in training by being prompted with a few examples. It turns out that performing in-context learning with NLEs provided in the examples (reffered-to as "X-ICL" [155]) leads to better performance on tasks requiring complex reasoning [155, 156, 154].

2.2.2.2 Evaluating natural language explanations

Evaluation methods and metrics for extractive explanations of machine learning models [157, 158, 159, 160] are not suitable for natural language explanations which are generated rather than extracted from the input. The simplest and most common approach to evaluating NLEs is to measure the similarity between the generated and ground-truth explanation with metrics like those described in Section 2.1.2.2's paragraph on evaluating report generation, for example BLEU [161]. Some publications like [142] also perform human evaluation, but this approach is challenging and expensive to scale. However, these NLE-generation evaluation metrics measure the persuasiveness of the explanations rather than their faithfulness. Faithfulness is defined as "the accuracy with which the explanation describes the decision-making process of the target model. Faithfulness of an explanation should not be confused with the property of an explanation to provide ground-truth argumentation for solving the task at hand, which is independent of a model decision-making process" in [162]. [131] introduces two conditions that are required for a NLE to be faithful: feature importance agreement and robustness equivalence. To measure them, the model input is altered and the change in model output is measured to observe the degree of label-NLE association [149]. The underlying idea is that an explanation can be faithful only if it is closely tied to the predicted label, and they should therefore be similarly affected by noise or by the removal of important explanation features for robustness equivalence and feature importance agreement respectively.

Domain-specific NLE faithfulness metrics also have been introduced as fields like medicine rely on more precise and factual explanations. In the chest X-ray analysis field for example, the "CLinical EVidence" (CLEV) score measures the clinical accuracy of these explanations using the CheXbert labeler to extract the evidence labels mentioned in the NLE and ground-truth explanation, and computing the accuracy of their detection as faithful NLEs should be based on the same evidence and findings in chest X-ray diagnosis [12].

To make NLE-generation evaluation more unified and comparable, benchmarks are being developed and adopted, such as e-SNLI [137] in the natural language field and e-ViL [136] for vision-language tasks.

2.2.3 Explainability in AI for chest X-rays

Model explainability is especially crucial in medical settings because mistakes are directly affect human lives, and when they happen, they need to be detected and understood immediately to correct them and the underlying model. Moreover, medical professionals need to understand the model's reasoning to challenge it based on their own knowledge

to limit confirmation bias. But how are these explainable AI techniques applied to deep-learning-based X-ray analysis, and are they sufficient for those constraints? In this section, we will begin by describing how post-hoc explainers and self-explaining models are used in radiology AI to then outline the limitations of these current applications that drive and inspire the work described in this thesis.

2.2.3.1 How post-hoc explainers are applied to chest X-rays

Post-hoc explanation techniques are the most common approach to explainable AI for chest X-ray analysis [163]. In this image processing task, the intuitive way to add explainability is to employ tools such as saliency maps to visually highlight the areas that contributed the most to a prediction [164]. Similar heatmaps are generated for automated CXR diagnosis using GradCAM variants in [165, 166, 167, 168]. This post-hoc explanation techniques generates the heatmaps through a weighted combination of forwarding activation maps fed through a ReLU activation function [169]. Different post-hoc visual explainers were compared in studies such as [170] where GradCAM and LIME are tested on a chest X-ray classification model, or [171] where SHAP and GradCAM++ are employed and compared to radiologist annotations.

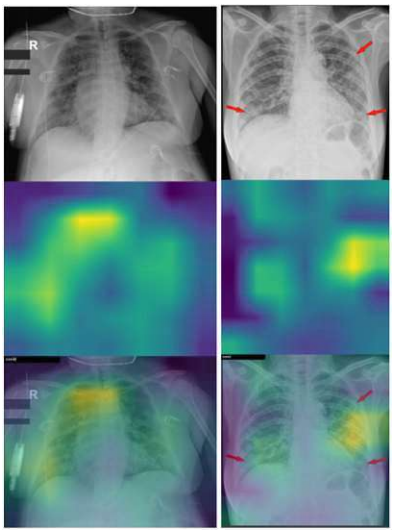

Figure 2.4: Examples of GradCAM activation maps on chest X-rays. Source: [166, Figure 9].

Some more recent approaches, such as the LT-ViT [74], use transformers as vision encoders which enables them to visualize the importance of image-regions without GradCAM and related variants by directly visualizing the smoothened attention maps of specific tokens (the *[CLS]* token for overall contribution or the *[LBL]* token for class-specific information). Even though this post-hoc explainability is directly built-into the model in comparison to the other post-hoc explainers that were mentioned, it still is performed

a posteriori and does not constrain the model to learn from relevant areas. Therefore, self-explaining approaches are still required to address these limitations.

2.2.3.2 Self-explaining models for chest X-rays

Though less common, self-explaining models are applied to chest X-rays to add explainability to automated CXR processing as well. Early works in this direction learnt to extract explicit rules for X-ray classification, making the model decisions inherently explainable. For example, [172] performs classical image feature extraction through programmatic filters and then extracts "IF-THEN" rules based on these features for explainable rule-based classification. Similarly, [169] learns a decision-tree to classify chest X-rays.

But more modern deep-learning-based self-explaining models for chest X-rays also exist. As mentioned in Section 2.2.1.2, some two-stage approaches to CXR report generation like [118, 117] perform visual grounding of the clinical observations they generate by design as they first detect abnormal or anatomical areas, and only feed those to the report generation module. Since the generated content was made based on that area only, these regions of interest can be used as visual explanations generated before the task-specific prediction (report generation) was made, making this an explain-then-predict approach. An alternative visually-self-explaining model that generates heatmaps like GradCAM for example is the CXRNet [173]. It introduces an encoder-decoder-encoder architecture for CXR classification, where the first encoder extracts features from the X-ray, the decoder reconstructs the input X-ray with an importance heatmap overlaid, and the last encoder extracts feature from this generated image to include that information in the final classification output. This pipeline is trained with label-only supervision without requiring heatmap annotations, thus making this a self-explaining explain-then-predict alternative to post-hoc visual explainability techniques like GradCAM.

Generating explanations in natural language (NLEs) for chest X-rays has not been explored as much yet. This task differs from report generation as generating NLEs isolates the reasoning capabilities of the model to specifically explain the presence of individual labels. Very recently, [174] has explored this through a hybrid approach between post-hoc explaining and self-explaining models for localized natural language generation. It introduces a model that learns to align both chest X-rays with report-level embeddings and chest X-ray patches with sentence-level embeddings through combined global and localized contrastive losses to then explain predictions by computing similarity scores between areas of the X-rays and different descriptions. This leads to localized explanation generation that is not purely post-hoc as it is integrated to the loss at training. However, the task of generating natural language explanations for predictions on chest X-rays has been explicitly introduced and formalized in [12]. The publication releases the MIMIC-NLE dataset, containing more than 38,000 NLEs for thoracic pathologies extracted from the largest chest X-ray report dataset, MIMIC-CXR [18]. It also describes and benchmarks baseline models for this new task that were explicitly trained to generate

natural language explanations for chest X-ray diagnoses (two CXR-captioning-based approaches adapted from TieNet [107] and RATCHET [175] and a model combining GPT-2 and a DenseNet-121 called "DPT"). As a part of this benchmark, it presents a new metric to evaluate the quality of the generated NLEs in addition to the report generation described in Section 2.1.2.2: the CLEV (CLinical EVidence) score. The CheXbert labeler [176] is used to extract the evidence labels mentioned in the generated and ground-truth NLEs, and computes the accuracy over all generated NLEs as clinically accurate NLEs should include the same findings as the ground-truth ones [12].

2.2.3.3 Limitations of current applications of explainable AI to chest X-rays and how they inspired the work done in this thesis

As discussed in this section, attempts at explaining automated chest X-ray analysis systems are most commonly made through post-hoc explaining techniques that do not constrain the model to learn from relevant features and only generate those explanations a posteriori. Even the self-explaining models that have been mentioned always apply a predict-then-explain approach. While [118, 117] are an exception to a certain degree, they do not generate NLEs that explain the reasoning but only visual areas that were taken into account, and they do not express how and why those areas were used. The NLE generation baselines introduced in [12] also all have a predict-then-explain approach where the explanation generator is conditioned on the output of the task-specific predictor. To our knowledge, explain-then-predict approaches to NLE generation for chest X-ray are yet to be introduced. Moreover, [177] argues that the current predict-then-explain paradigm in vision-language settings leads to completely independent language and vision-language models. The disconnection between explanation generation and task-specific answering prevents the final prediction from leveraging the reasoning and information contained in the explanation.

This is what drives our approach to generate natural language explanations for CXR diagnoses in an explain-then-predict paradigm. By first generating an explanation for each diagnosis (label) and then selecting the ones that are actually true and relevant for the input X-ray, the critic (classifier) leverages the content of the explanations when making diagnoses. This in turn constrains the model to learn to generate clinically faithful explanations that are truly relevant to the image in order to be selected by the critic, and places explanation generation at a central role in the learning of the model which is different to current approaches to explainable chest X-ray diagnosis. These explainable diagnoses that were made using on the reasoning contained in the explanation then have the potential to be challenged by medical professionals to avoid confirmation bias when using such a system as assistance in a clinical setting.

CHAPTER 3

Generating Natural Language Explanations for Chest X-ray diagnoses

To solve our overall goal of predicting CXR diagnoses with and based on natural language explanations, the first step is to be able to generate accurate NLEs for positive pathologies on chest X-rays. This is why, in this first chapter, we describe our effort in building a vision-language model that takes a chest X-ray image and a textual prompt specifying a label (pathology) and generates an accurate explanation for why that label is positive for the medical image in natural language.

3.1 Problem setting

Natural language explanations for chest X-ray diagnoses offer radiologist-friendly explanations that reflect how these professionals articulate their findings [12]. They can help identify potential biases or errors in the model's reasoning, leading to safer and more reliable diagnoses in a real world setting. This is why, in our effort of building a system that outputs faithful natural language explanations for each output diagnosis, we begin by solving the task of generating explanations for labels that we know are positive for a given chest X-ray.

Solving this task can be formalized as training a multimodal machine learning model f that processes both visual and textual inputs to generate a natural language sequence. The inputs and output of the model are defined as follows:

- Let $\mathbf{I}$ represent an input (chest X-ray) image, encoded as a tensor suitable for processing by the model.

- Let **s** represent a textual prompt in the form of a string, specifically structured as "Evidence for {LABEL}", where {LABEL} is dynamically replaced with the relevant label.

The output of the model is a natural language sequence **y**, which is a decoded string representing the natural language explanation for the given label and X-ray. The interaction between the inputs and the output can be described by the function f:

$$\mathbf{y} = f(\mathbf{I}, \mathbf{s}; \boldsymbol{\theta}) \tag{3.1}$$

where $\boldsymbol{\theta}$ represents the parameters of the model, which may include weights of neural networks, settings for the encoding and decoding mechanisms, and other relevant model-specific parameters.

The authors introducing the task of generating NLEs for CXR diagnoses propose baselines for this exact task [12]. Their proposed approaches revolve around in a first step classifying the positive labels, and then generating NLEs for those labels only. Since our dataset also only contains NLEs for positive labels, and because we want to generate NLEs justifying the presence of a pathology to see if they are actually faithful to the image to make our classification decision in our overall approach, we learn to generate NLEs for positive labels only in this stage. As the problem has so far been tackled with simpler models, we also explore if a larger and more sophisticated architecture can more accurately model these NLEs.

3.2 Method

To solve this problem, we base our work on the BLIP-2 vision-language architecture for our model f due to its promise of enabling more nuanced interactions between the text and image modalities while bootstrapping pre-trained base vision and language models [13].

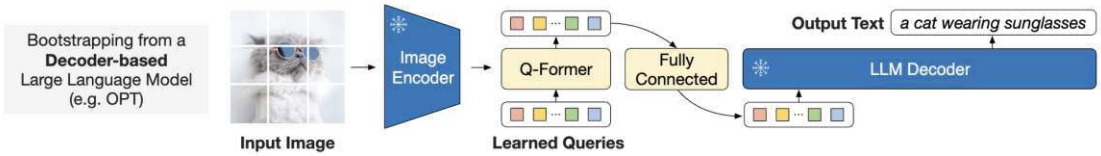

Figure 3.1: Illustrating how BLIP-2 bootstraps a trained image encoder and decoder-based LLM. Image from [13].

As illustrated in Figure 3.1, this architecture improves the integration of image and text information through a Query Transformer ("Q-former"). This component is essentially a BERT model, which introduces a fixed number of learnt query tokens that are not derived from actual text inputs but are instead designed to capture various aspects of the image. These query tokens interact with the image features through a cross-attention mechanism

within the Transformer structure to extract and represent the most salient features for the task at hand. The resulting image representation is fed through a projection layer (referred to as adapter) before being concatenated to the tokenized input sequence (prompt) of the frozen LLM to align it with the model's input space.

BLIP-2 is trained in two stages. The first one pre-trains the Q-former component based on a combination of three losses (image-text matching, image-text contrastive and captioning) to align image and text representations. This will be further covered in Chapter 4, and we here focus on the second stage shown in Figure 3.1 which focuses on vision-language modeling.

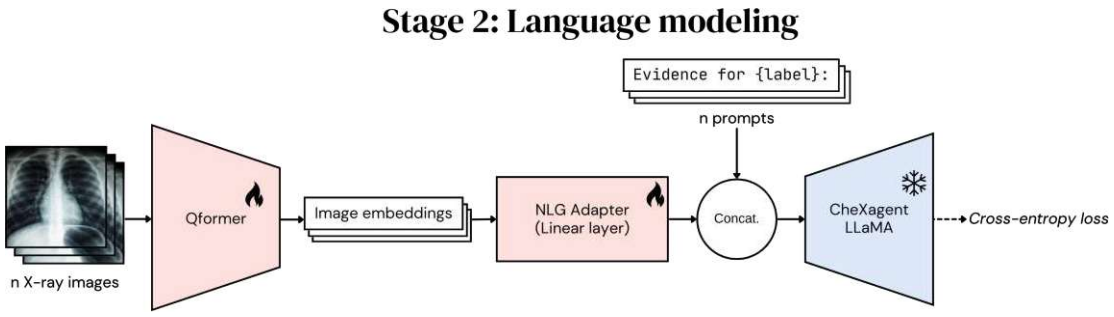

Figure 3.2: Language modeling stage (stage 2) of the training of our model. The Q-former and language-model-input adapter are trained at this stage, using a cross-entropy loss on the token logits (language modeling training).

Figure 3.2 represents the model we train to learn to generate an NLE for a given chest X-ray and label. We use the same setup as the BLIP-2 model: a vision encoder, a Q-former, an adapter linear layer and a large language model. We use EVA-CLIP-g [178] for the vision encoder, BERT [19] for the Q-former and LLaMA-2 [20] as a language model architecture. Based on the observation that CheXagent [15], a chest X-ray-specialized vision-language foundation model, is also based on the same BLIP-2 architecture and pre-trained on a large set of chest X-ray data [1], we choose to initialize the weights of our vision encoder and language model based on CheXagent's. This allows us to bootstrap their CXR-specialized vision encoder and clinical language model as we work on the same domain. The Q-former and adapter components however are initialized randomly, as we train them to specifically model the task of NLE generation. For this second stage training focused on language modeling, we initialize the Q-former with the weights from our first stage of training focused on the Q-former itself, which is discussed in Chapter 4.

[1]While MIMIC-CXR and MIMIC-NLE are included, they use the same official split as we do [15], avoiding data leakage.

To train this model to generate natural language explanations, we use the training split of the MIMIC-NLE dataset that is further described in Section 3.3. As the dataset only contains NLEs for positive labels, the training samples consist of every image-NLE pair of the train split, associated with their respective labels. To train the model in Figure 3.2, we feed an image and prompt as an input, and train for autoregressive language modeling with a cross-entropy loss just as the language modeling stage of BLIP-2 [13]. We thus learn to auto-regressively decode NLE sequences from the prompt and image-information-carrying tokens. The prompt to the model is of the form `Evidence for {LABEL}` where the specific label is inserted. This form of prompting allows to generate explanations for different labels, but also allows the addition of more information in the future, such as patient context for example. Just as BLIP-2, we only train the Q-former and adapter parameters at this stage, bootstrapping the already chest-X-ray-specialized vision encoder and language model.

3.3 Experimental setup

3.3.1 Dataset and processing

As for all of our experiments, we use the MIMIC-NLE dataset [12]. It extends the MIMIC-CXR dataset [18] by extracting natural language explanations for the positive diagnoses of a subset of its chest X-ray images. The NLEs in the dataset focus on explaining diagnoses that are positive or uncertain, as negative findings generally don't require case-specific explanations. Those explanations were automatically extracted from the radiology reports provided in MIMIC-CXR using a BERT-based labeler, a set of clinical explanation keywords, and an empirically and clinically validated set of extraction rules. The creators begin by extracting the Findings and Impression sections from the radiology reports, which contained the descriptive portions of the reports. They then filter the extracted sequences to remove noise and information that is not visible in the images, like patient history. Afterwards, the CheXbert [179] labeler is used to identify the 14 chest X-ray labels of the dataset that can be mentioned in every extracted sentence. To determine which labels in an NLE are being explained and which are the evidence, the authors designed an evidence graph that formalizes label combinations with high-confidence relationships. The graph shows which labels can act as evidence for other labels, for example, Consolidation is considered evidence for Pneumonia. Based on this graph, mutually exclusive rules are defined to formalize valid NLEs based on the label and the present of explanation keywords: for example, the combination of Consolidation and Pneumonia is considered a valid NLE even without an explanation keyword because the evidence relationship is generally clear [12].

This process yields a dataset of 38,003 image-NLE pairs or 44,935 image-diagnosis-NLE triplets as some NLEs can explain multiple diagnoses. They are split into 37,016 training, 273 development and 714 testing image-NLE pairs. The label distribution is the same across splits, and the label counts and proportions are for example described for the test split of the dataset in Figure 3.3.

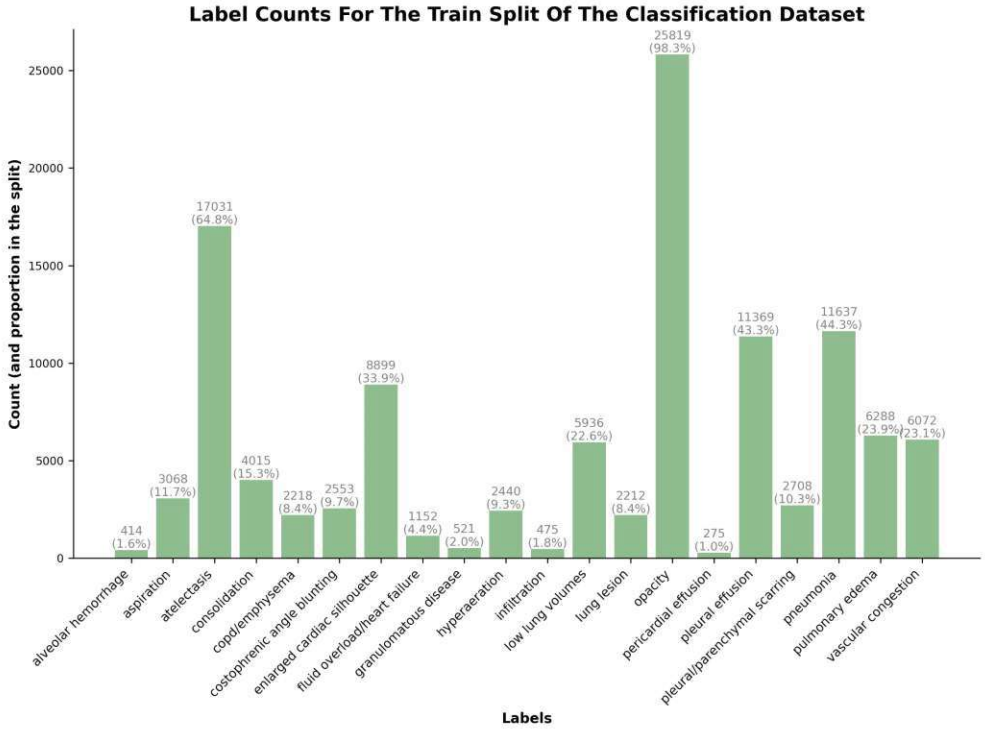

Figure 3.3: Class counts and proportions for the train split of the MIMIC-NLE dataset.

We only keep half of the labels, specifically those with the highest proportions of positive samples also associated with NLEs: aspiration, pneumonia, pulmonary edema, vascular congestion, atelectasis, pleural effusion, and COPD/emphysema. This selection primarily impacts evidence classes[2], which naturally generate fewer NLEs. We consider this advantageous, as focusing on NLE generation for evidence classes would be less beneficial. Our decision to filter labels is driven by our aim to showcase the potential of our approach when there is sufficient data for all selected labels, rather than striving to build a model that comprehensively covers all pathologies. In our NLE-generation training and evaluation process, we consider image-NLE pairs of the dataset as samples. Additionally, we also consider the label associated to each pair to craft the prompt for which we learn to generate that NLE for the given X-ray. We process the images in 224x224 resolution, in grayscale for the DenseNet vision encoder and in RGB for other models. During the training process, we apply data augmentations to the chest X-rays such as randomly flipping the images horizontally, rotating them by up to 10 degrees and applying random affine transformations with no rotation but with translations and scaling, converting them to tensor format. At both train- and test-time, we normalize the pixel values based on the training data's mean and standard deviation values.

[2]Types of findings or abnormalities that are directly observable in the radiographic image and that suggest or indicate the presence of a pathological condition.

3.3.2 Evaluation metrics

To measure and compare natural language generation performance, we focus on some typical natural language generation metrics, BLEU-1, BLEU-4, ROUGE-1 and ROUGE-L scores, which are also measured in [12]. The BLEU-n scores quantify the correspondence of n-grams between the generated and ground-truth NLEs, while ROUGE-1 measures the overlap in unigrams and ROUGE-L is based on the longest common subsequence to take coherence and order of textual elements into account.

3.3.3 Baselines

To verify if the added complexity of the BLIP-2 architecture is beneficial for chest X-ray NLE generation, we compare our results to multiple baselines.

1. We also implement and train a simpler baseline approach to vision-language modeling illustrated in Figure 3.4 by connecting an image encoder to a pre-trained language model only through a lightweight adapter linear layer. We train two models of this form.

   - The first one combines the chest X-ray pre-trained DenseNet-121 vision encoder provided by TorchXRayVision [180] with the BioGPT language model, a GPT-2-based model pre-trained on large-scale biomedical literature [90] through a simple projection layer. For this model, all LM, vision encoder and adapter parameters are trainable.
   - The second baseline uses the same components but replaces the BioGPT language model with the CheXagent LLaMA-2-based language model used in our main model. In this case, only the vision encoder and adapter parameters are trainable, keeping the pre-trained large language model frozen.

2. We also introduce a "blind" baseline, which consists of only the BioGPT language model where all parameters are trainable. This model does not take any visual input, thus generating an NLE solely based on the "`Evidence for {LABEL}`" prompt, thus generating unfaithful NLEs by design as it ignores the chest X-ray itself. This baseline serves as a sanity check verifying how much an unfaithful model that does not ground its explanation in the medical image can still generate persuasive NLEs that could lead to good natural language generation (NLG) metrics.

3. To determine how well a chest X-ray specialized foundation vision-language model performs off the shelf on this NLE generation task, we also consider a baseline where we prompt CheXagent for explanations for positive labels. Note that the MIMIC-NLE dataset was included in its pre-training corpus, so this is not purely "zero-shot" as it has already encountered natural language explanations. We do not further train this baseline.

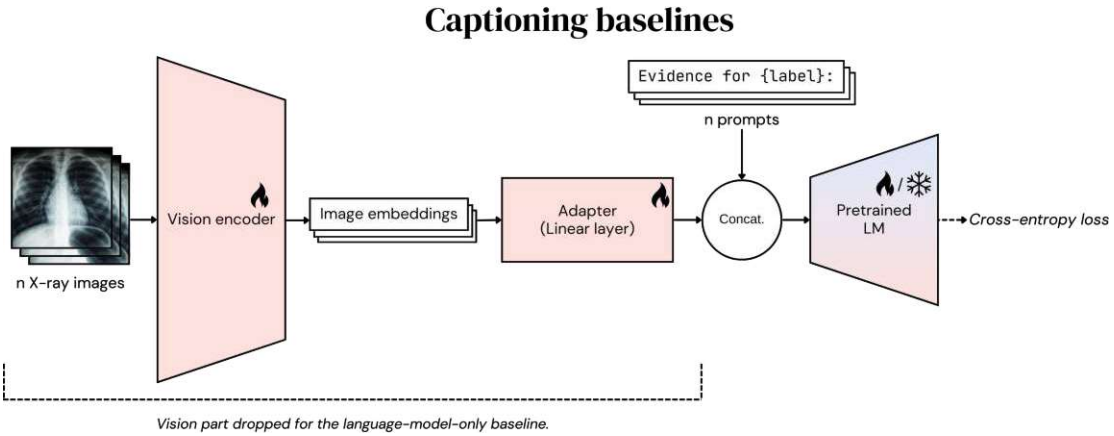

Figure 3.4: Structure of our captioning baselines for NLE generation. A vision encoder's output is fed into a projection layer to create image tokens that are concatenated to the tokenized prompt to create the LLM input. Vision encoder, adapter, and for some cases language model parameters are trained using a cross-entropy loss.

4. Finally, we also consider the BLEU and ROUGE scores reported in the publication that introduced the NLE generation task for chest X-rays and the MIMIC-NLE dataset [12]. We do not re-run their models, and only report their values as a check that we do not score significantly below the baselines that they introduce, and to verify that our added complexity is beneficial to a certain degree.

3.3.4 Training setup

We implement our model based on the LAVIS library's [181] official implementation of the BLIP-2 model. The vision encoder and language model weights are loaded in 16-bit floating point precision and kept frozen. The Q-former and adapter weights are in 32-bit floating point precision. We train for 100 epochs, while implementing early stopping based on the NLG metrics on the dev split of MIMIC-NLE, leading to the best model being stopped at epoch 58. We obtain our best results by freezing the Q-former weights for the first two epochs, first only training the adapter that has been randomly initialized, and keeping the Q-former weights training for the rest of the training. We use the AdamW optimizer with a weight decay factor of 0.01 and separate learning rates of $2e^{-6}$ for the Q-former and $5e^{-5}$ for the adapter. These hyperparameters, as well as the benefit of using distinct learning rates, have been determined through tuning on the dev dataset. The training is run on a single Nvidia H100 GPU with 80GB of VRAM using a batch size of 32.

As for the baselines, we use 32-bit floating point precision for all trainable parts of the networks. For the BioGPT + DenseNet baseline, we use a batch size of 128, a weight decay factor of 0.02 and a cosine-decay-scheduled learning rate ranging from $1e^{-5}$ to $1e^{-6}$, and the best model is early stopped at epoch 9. The BioGPT-only baseline uses

the same configuration but the learning rate ends at $5e^{-6}$ instead and early stops at epoch 18. Finally, the CheXagent LM + DenseNet baseline is trained with a batch size of 64, a weight decay factor of 0.02 and a cosine-decay-scheduled learning rate ranging from $1e^{-4}$ to $9e^{-6}$. The best model is trained for 84 epochs.

3.4 Results and discussion

The training dynamics of our model are stable. As shown in Figure 3.5, the cross-entropy loss does go down without large spikes while the validation BLEU and ROUGE scores grow before plateauing, and even slightly starting decreasing, hinting at overfitting. This is not a problem as we use our early-stopped checkpoint at step 400 of the 58th epoch which led to the best NLG metrics on the validation split of MIMIC-NLE. The baselines we have trained displayed very similar training dynamics. During validation, we also track the image-text contrastive capabilities of the Q-former for which it was pretrained in the previous stage (as discussed in Chapter 4). As expected given the fact that this stage is trained only with a captioning objective, image-text contrastive capabilities decrease as the validation AUC goes down slightly the longer we train (*12% after 100 epochs*). This not a problem for the BLIP-2 architecture as the purpose of the Q-former is to align the image and text representations before the captioning training. However, in our case where we want to reuse the image-text contrastive capabilities of the Q-former in the final system, this observation motivates our decision to use two different sets of Q-former weights in our final system for the critic and natural language generation Q-formers.

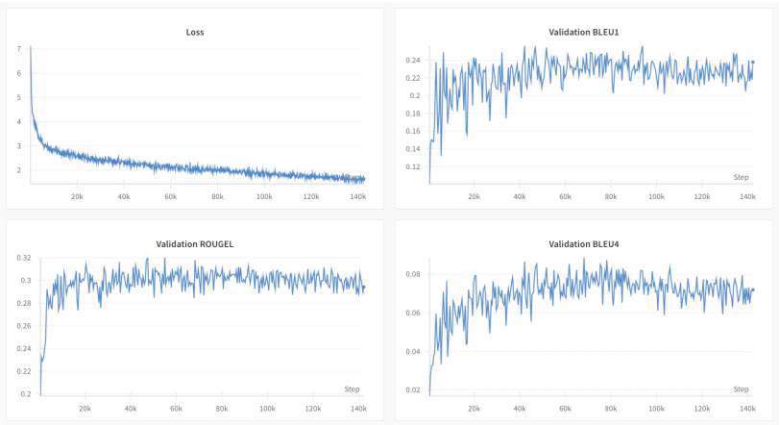

Figure 3.5: Training loss and validation BLEU-1, BLEU-4 and ROUGE-L scores over the 100 epochs of training.

The results of the evaluation on the test-split of the MIMIC-NLE dataset of our trained model, as well as of our different baselines introduced in 3.3.3, are summarized in Table 3.1.

Model	BLEU-1 ↑	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-L ↑
BioGPT LM Only	0.1047	0.0225	0.1791	0.1588
BioGPT LM + vision proj.	0.1941	0.044	0.2376	0.2059
CheXagent LM + vision proj.	**0.2391**	**0.0819**	0.3088	0.2791
CheXagent zero-shot	0.1426	0.0329	0.2493	0.2376
RATCHET, best in [12]	0.225	0.047	-	0.222
Ours	0.2312	0.0764	**0.3323**	**0.3034**

Table 3.1: NLG metrics of our model and baselines for the task of NLE generation for positive labels on the official test split of the MIMIC-NLE dataset. The best value for each metric is in bold.

A first observation that can be made is that we validate that we at least match or beat the metrics reported for the best baseline (RACHET) in [12]. While this is not a direct comparison, this was a first sanity check to ensure that we do meaningful work on this task, and that the complexity we add in our approaches is beneficial. Furthermore, all of our approaches (our simpler baselines and our proposed model) that take the X-ray image as an input exhibit significantly better NLG scores than the zero-shot prompting of the CheXagent foundation model, which has scores that are close to the blind baseline with textual prompts only. This is observed even though CheXagent is a strong chest X-ray generalist model, that also includes MIMIC-NLE in its pretraining dataset, meaning that it encountered NLEs at training. This shows that it still seems beneficial to train custom models for this natural language explanation generation task for chest X-ray daignoses, as ev en our much smaller and simpler baselines report significantly higher NLG scores.

Overall, we consider our approach as the best performing out of the evaluated models, closely followed by the baseline combining a DenseNet with the CheXagent language model. While that baselines has slightly better BLEU scores, our approach is the most balanced and consistent across metrics: it has significantly higher ROUGE scores while BLEU-1 and BLEU-4 are very close. The performance of our model also confirms that the added complexity of the Q-former seems justified for NLE generation. It is also important to note that the NLG metrics we report here a proxy for NLE quality rather than an exact measure, further highlighting the need to look at the whole picture rather than difference on a single metric. NLEs are a specific type of text that is usually only one sentence, and where some specific medical keywords are more important than others. BLEU and ROUGE do not measure whether the generated text provides medically accurate and relevant information, only whether it resembles the reference texts in structure and vocabulary. Those metrics focus on exact matches of words or phrases and do not account for the meaning or the semantic correctness of the content. This is particularly critical in medical NLEs, where different wordings can convey the same essential information but might not be recognized by these metrics.

Finally, an important result of these experiments is that "blindly" (without taking the X-ray image into account) generating NLEs, by only generating an explanation for the given label with a language model, leads to roughly half as good NLG metrics. However, the generated explanations are still convincing and do not lead to completely bad metrics. This highlights the danger of a model generating NLEs that sound plausible but that are not faithful to the medical image, which would be a particularly critical problem in a clinical setting that needs to be detected. This motivates our work on designing a "critic" part of our system that is able to differenciate between NLEs that are faithful to the X-ray and explanations that are not, which we explore in the following chapter.

CHAPTER 4

Building a critic: Can image-text similarity capture faithfulness of natural language explanations?

While we have shown that we can generate natural language explanations for chest X-ray diagnoses, our work has also highlighted the risk of decoding persuasive unfaithful NLEs that do not base their content on the X-ray's content. This is what drives the need to build a critic model that is able to detect unfaithful or irrelevant NLEs, explanations that are not true in regard to the medical image.

Our goal here is to determine if a critic can capture NLE faithfulness, making it able to separate true image-NLE pairs from false ones. This capability is key in our overall "explain-then-predict" approach to chest X-ray diagnosis, as this idea of making a label prediction based on the relevance of the generated explanation to the image hinges on a critic that is able to reject explanations for labels that are not present in the image, because the explanations for those labels will inherently be unfaithful. In this chapter, we investigate how well such a critic can be achieved through image-text similarity, using a vision-language model that learns joint image-text representations like CLIP or the Q-former component of our BLIP-2-based network.

4.1 Problem setting

The task we are attempting to solve is to determine if an NLE is faithful and relevant to the visual content of a chest X-ray (if it corresponds to the image). This problem of differentiating between relevant or faithful NLEs from unfaithful ones can be formalized as a multimodal binary classification task, where a model f processes image and text

inputs to output a classification label determining if the NLE is true for the given image. The inputs and output of the model are as follows:

- Let **I** represent an input (chest X-ray) image, encoded as a tensor suitable for processing by the model.
- Let **n** represent an NLE preprended with a textual prompt specifying the label it explains in the form of a string, specifically structured as "Evidence for {LABEL}: {EXPLANATION}".

The output of the model is a probability $y \in [0; 1]$ that the NLE n is true for the chest X-ray image I. The interaction between the inputs and the output can be described by the function f:

$$\mathbf{y} = f(\mathbf{I}, \mathbf{n}; \boldsymbol{\theta}) \tag{4.1}$$

where $\boldsymbol{\theta}$ represents the parameters of the model, which may include weights of neural networks, settings for the encoding and decoding mechanisms, and other relevant model-specific parameters.

4.2 Method

While this problem could be approached many different ways and with different types of models, we base our method on the idea of reusing a component of our already complex natural language generation pipeline. Since we include a model that was trained including an image-text contrastive loss, the Q-former, in our captioning architecture, we explore how the image-text similarity capabilities of that component capture faithfulness of NLEs to an image. The overall approach is to train the Q-former on our image-NLE pairs, as it would be done for a BLIP-2 model in any case, and to then evaluate if the representations learnt by that model can capture image faithfulness, for example with true NLEs being more similar to their corresponding images than random false NLEs reliably.

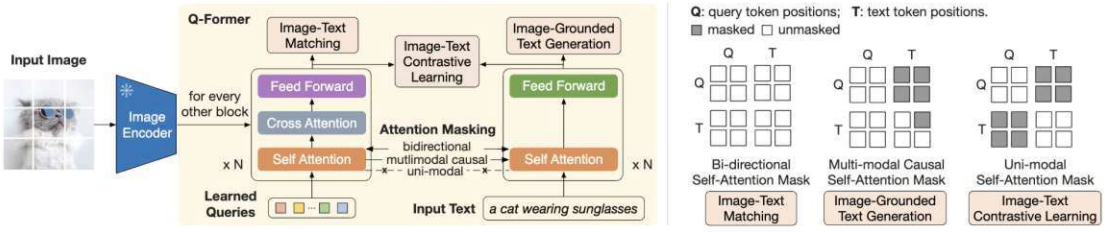

Figure 4.1: Q-former pre-training for BLIP-2, from [13].

The Q-former component combines our vision encoder features with a BERT transformer-based text encoder to optimize cross-modal understanding through cross-modal attention layers where learnt query tokens can attend information from both text and image embeddings. This setup allows the model to focus on relevant parts of an image given a

text query (and vice versa), enhancing the model's understanding of how text and image content correspond to each other. As illustrated in Figure 4.1 and Figure 4.2, we use the same combined learning objective to pre-train our randomly initialized Q-former as BLIP-2 [13] while training all parameters of the Q-former and keeping the vision encoder frozen. This combined loss is a sum of an image-text contrastive, and image-text matching and a text-generation loss. The text generation loss is used to enhance the linguistic understanding of the model. It trains the model to predict words that have been intentionally masked in the text, based on the context provided by both the remaining words and the corresponding image, to ensure that the model can integrate both textual and visual information to improve language generation. The image-text matching loss trains an added lightweight head to output if the input pair of image and text correspond to each other. Finally, the image-text contrastive loss aims to align the embeddings of the text and the images in a shared multimodal space. This helps the model learn a robust representation where corresponding images and texts are close to each other, while non-corresponding pairs are farther apart, through a contrastive learning framework. This is essential for aligned image and text representations that are then used in the language generation pipeline, but we also leverage this property of the learnt representation to repurpose the Q-former as a critic.

The combined learning objective $\mathcal{L}_{\text{total}}$ is specified as follows:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{ITC}} + \lambda_2 \mathcal{L}_{\text{ITM}} + \lambda_3 \mathcal{L}_{\text{IGT}}, \tag{4.2}$$

where λ_1, λ_2, and λ_3 are hyperparameters that control the relative contributions of each loss component.

$$\mathcal{L}_{\text{ITC}} = -\frac{1}{N} \sum_{i=1}^{N} \left[\log \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau)} + \log \frac{\exp(\text{sim}(\mathbf{t}_i, \mathbf{v}_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{t}_i, \mathbf{v}_j)/\tau)} \right], \tag{4.3}$$

where $\mathbf{v}_i$ and $\mathbf{t}_i$ represent the image and text embeddings for the i-th sample, respectively, $\text{sim}(\cdot, \cdot)$ is a similarity function (e.g., cosine similarity), τ is a temperature parameter that controls the sharpness of the distribution, and N is the batch size.

$$\mathcal{L}_{\text{ITM}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log p(\text{match}|\mathbf{v}_i, \mathbf{t}_i) + (1 - y_i) \log(1 - p(\text{match}|\mathbf{v}_i, \mathbf{t}_i)) \right], \tag{4.4}$$

where $y_i \in \{0, 1\}$ is the ground truth label indicating whether the image and text pair match, and $p(\text{match}|\mathbf{v}_i, \mathbf{t}_i)$ is the predicted probability that the image $\mathbf{v}_i$ matches the text $\mathbf{t}_i$.

$$\mathcal{L}_{\text{IGT}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \log p(\mathbf{w}_{i,t}|\mathbf{v}_i, \mathbf{w}_{i,<t}), \tag{4.5}$$

where $\mathbf{w}_{i,t}$ represents the t-th word in the target sequence for the i-th sample, $\mathbf{w}_{i,<t}$ represents all words before the t-th word, and T is the length of the target sequence.

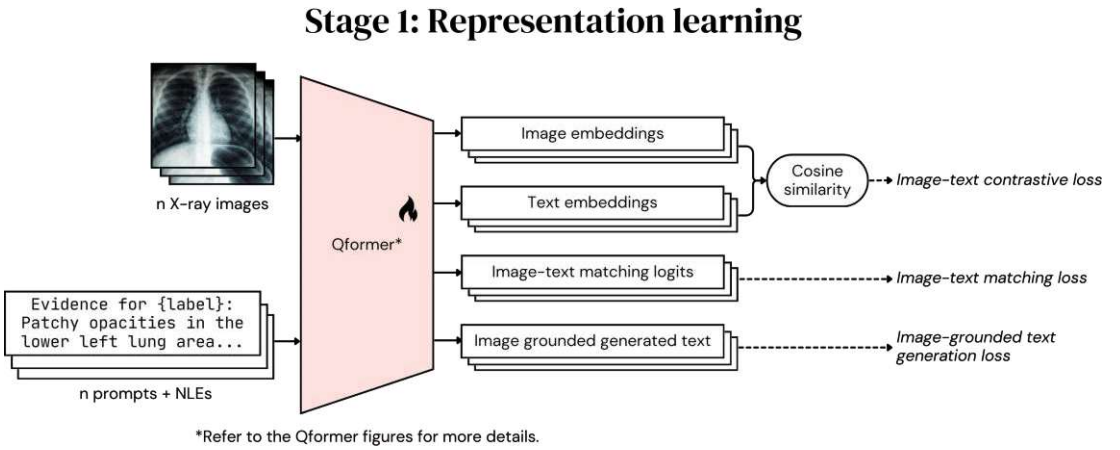

Figure 4.2: Our Q-former training process: both an image and an NLE with its corresponding prompt are encoded with the Q-former. A combination image-text contrastive, image-text matching and text-generation losses is computed and backpropagated.

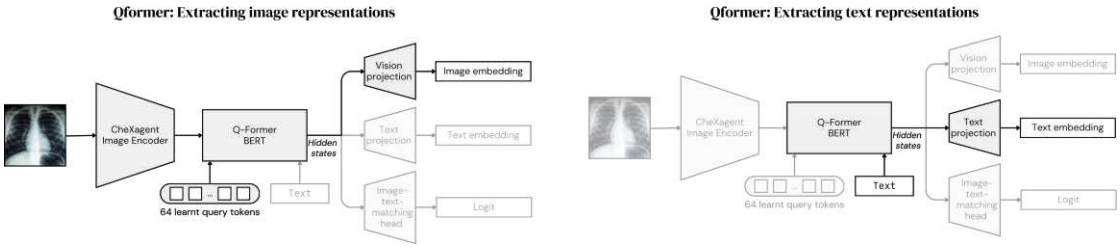

Figure 4.3: Illustrations of how our Q-former is used to embed images and text.

The idea is that, as we constrain our image (X-rays) and text (NLEs) representations to be similar to eachother when they correspond, and dissimilar when they do not, we could leverage the similarity between an image and a generated candidate NLE to determine if the generated explanation is relevant and using the combined BLIP-2 loss on our entire train split of MIMIC-NLE, considering all pairs of X-rays and their corresponding NLEs as training samples. We use the resulting model weights as Q-former base weights for the NLE-generation pipeline described in Chapter 3, but also investigate if the image-text similarity properties of the trained Q-former are able to discriminate between true and false (unfaithful) image-NLE pairs. To do so, we measure the similarity between all, true and false, image-NLE pairs. We do this for both the ground-truth NLEs of that dataset, and for NLEs generated by our model and the blind baseline that have been trained and described in Chapter 3. This allows us to verify if faithfulness is also captured for model-generated NLEs (as the blind model generates unfaithful NLEs by design) also are detected by this model, and not just false pairs from the dataset itself. We then verify, through multiple metrics, if true and false pairs can be separated and detected accurately based on those similarities.

If the Q-former captures NLE relevance and faithfulness effectively, the true and false pairs from the ground-truth dataset and generated by our proposed model should be separable based on similarities, while the NLEs generated by the blind model should not be as even the "true pairs" (NLEs generated for the corresponding X-ray) should be unfaithful as the NLE was not generated based on the image's content. In that case, false pairs and NLEs generated for their image should not have significantly different similarities as both are not faithful to the medical images.

4.3 Experimental setup

4.3.1 Dataset and processing

The training is performed on the test split of the MIMIC-NLE dataset, while hyperparameters are determined from tuning on the validation split and all models are evaluated on the test split of the dataset described in Section 3.3.1. We apply the same preprocessing steps and data augmentations for these experiments. We consider all image-NLE pairs of the dataset, where each NLE is prepended with its a prompt containing its corresponding label of the form `Evidence for {LABEL}: {NLE}`. This is to ensure that the x-ray-NLE similarities learnt take the label information into account.

4.3.2 Evaluation metrics

To evaluate the retrieval performance of the representations learnt by the Q-former we train, we measure Recall at k (Recall@k) for $k \in 1, 5, 10$.

More importantly, we quantify to what degree true and false image-NLE pairs are separable through the image-text similarity by measuring the AUC score for the binary classification true-or-false pair classification task. In addition, we verify the statistical significance of the separability of both classes with a Mann-Whitney U-test. We also include the mean similarity of positive and negative pairs as an additional information about the similarity-value differences between those two groups.

4.3.3 Baselines

To confirm that the combination of three losses to train our Q-former is beneficial, we also train a baseline purely using a CLIP loss. As illustrated in Figure 4.4, in this setup, we use the Q-former to encoder the text while the image is only embedded with the vision encoder. That latent representation is directly used in the CLIP loss, without attending to the Q-formers query tokens. The representations are learnt solely using the CLIP loss that constraints related image- and text-embeddings to have a high cosine similarity while unrelated ones are forced to be as distant as possible [14].

When evaluating retrieval performance, we additionally include the pre-trained PubMed CLIP model [182] as an off-the-shelf, zero-shot baseline [1]. This CLIP-based model has

[1]The authors have made three variations of the model available, using ResNet-50, ResNet-50×4 and

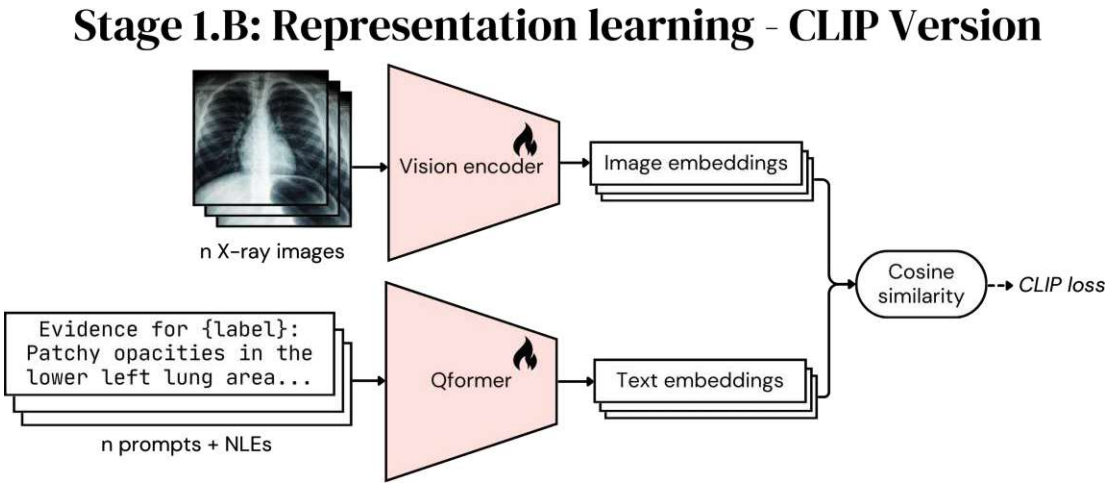

Figure 4.4: Q-former-alternative baseline trained on a single CLIP loss: the image is embedded with the vision encoder only, and does not attend to the Q-former query tokens.

been trained on the Radiology Objects in COntext dataset that provides PubMed-mined multimodal medical data from different types of medical imaging for various physiological regions [183].

4.3.4 Training setup

The implementation of the Q-former is based on the LAVIS library [181]. The vision encoder weights are loaded from the CheXagent checkpoint [15] in 16-bit floating point precision and frozen at training, while all other Q-former parameters are in 32-bit floating point precision and randomly initialized. We train 64 learnable query-tokens for the Q-former. Training is performed for 35 epochs, but the best performing model is early stopped at the 26th epoch. We achieved our best results by only training the vision and text representation projection layers and the query tokens for the first six epochs, and then unfreezing all parameters of the Q-former. We train on 4 Nvidia Titan RTX GPUs, providing 24GB of VRAM each, using a data-distributed-parallel strategy. This allows us to obtain a total batch size of 128. We train using the AdamW optimizer with a weight decay factor of 0.01 and a learning rate of $5e^{-7}$ for the projection layers and $2e^{-6}$ for the BERT part of the Q-former. We also clip the gradients to a maximum norm of 1.0.

As for the CLIP-loss-only baseline, we train the model on a single Nvidia H100 GPU with 80GB of VRAM using a batch size of 256. The training configuration is the same, except that we use a learning rate of $5e^{-4}$ for the image and text projection layers.

ViT32 as vision backbones. We used the ViT32-based model, available at `https://huggingface.co/flaviagiammarino/pubmed-clip-vit-base-patch32`.

4.4 Results and discussion

Training the Q-former was very sensitive to setup variations and hyperparameter changes. In many scenarios, loss was very spiky, decreased slowly, we observed fast overfitting and image-text matching did not improve. As shown in Figure 4.5, we have achieved more stable training where all three losses converge. The key aspects to succesful Q-former training were most importantly a large batch size, gradient clipping and progressively unfreezing layers as we train. Keeping the BERT part of the Q-former frozen for the first 5 epochs and only training the query tokens and vision and text projection layers, and then unfreezing the full model (as visible in the loss-drop at step 8200 on Figure 4.5) significantly improved learning dynamics. The image-text contrastive loss is more stable than image-text matching, which did not converge at all at smaller batch sizes. Placing a focus on it by scaling it or down-weighting other components of the loss did not help. We also also attempted oversampling NLEs from minority classes at training which did not improve dynamics. Large batch sizes led it to start to converge, and increases in batch size always improved stability in image-text matching loss. If the trend we saw by increasing our batch size continues, it could probably learn with even more stable dynamics with more hardware, and the fact that BLIP-2 training was done using way larger batches on multiple GPUs [13] supports this supposition. But the representation learning constrained by this image-text matching loss still is beneficial the experiments we ran excluding this loss led to both worse image-text contrastive performance and worse downstream NLE generation performance.

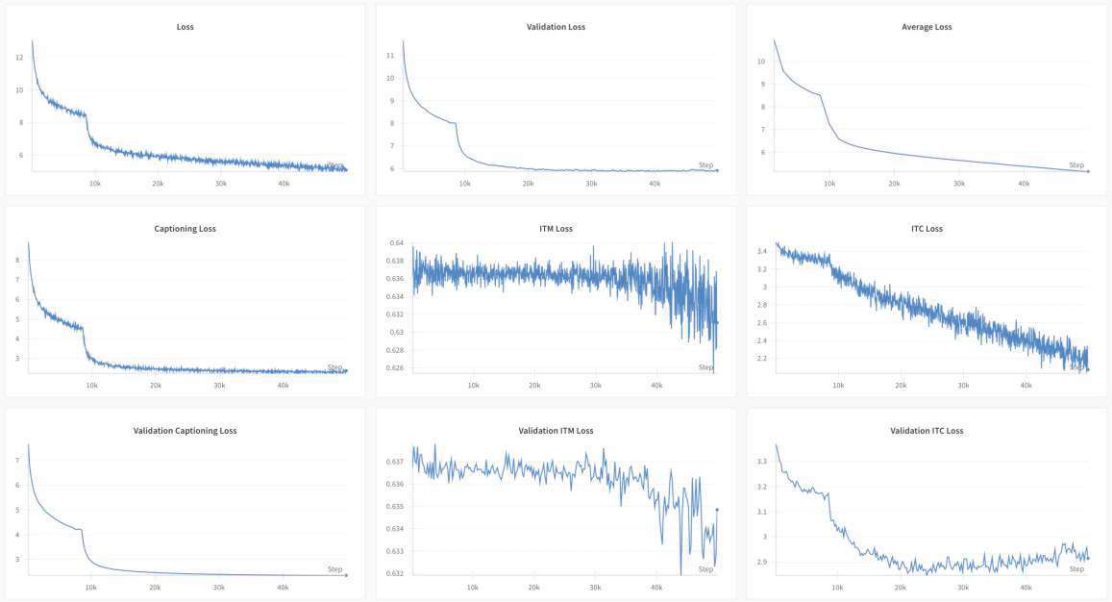

Figure 4.5: Evolution of the combined and individual losses during Q-former training.

The recall scores obtained by our model, reported in Table 4.1, beat both our CLIP-loss-only trained baseline and the off-the-shelf PubMed CLIP model.

Model	R@1 i2t ↑	R@1 t2i ↑	R@5 i2t ↑	R@5 t2i ↑	R@10 i2t ↑	R@10 t2i ↑
PubMed CLIP (0-shot)	0.0013	0.0013	0.0064	0.0064	0.014	0.0216
Ours (CLIP loss)	0.0102	0.0152	0.033	0.047	0.0788	0.0801
Ours (BLIP-2 loss)	**0.0178**	**0.0191**	**0.0801**	**0.0648**	**0.1182**	**0.1067**

Table 4.1: Retrieval performance (recall at k, referred to as R@k) of our Q-former pretrained on a pure CLIP loss or a BLIP-2 combined loss, and of a pre-trained PubMed CLIP model on all chest X-ray/NLE pairs of of the MIMIC-NLE test split. The best value for each metric is in bold.

Model	NLE source	Mean pos. NLE sim.	Mean neg. NLE sim.	AUC ↑	Separability (Mann-Whitney)
Ours, CLIP loss	Ground truth	0.3084	0.2487	0.7314	✓
Ours, BLIP-2 loss		0.3474	0.2301	0.792	✓
PubMed CLIP		0.2535	0.2535	0.5	×
Ours, CLIP loss	BioGPT LM only	0.2531	0.2369	0.5676	✓
Ours, BLIP-2 loss		0.2199	0.2019	0.5499	✓
Ours, CLIP loss	Our NLE generator	0.3154	0.249	0.7701	✓
Ours, BLIP-2 loss		0.3677	0.2432	**0.8452**	✓

Table 4.2: Separation capabilities of true and false image-NLE pairs, for both ground-truth and generated NLEs. For the different models, mean similarity for negative and positive pairs and AUC is reported, as well as the result of a statistical Mann-Whitney U-test for separability. The best AUC is in bold.

This confirms that the combined loss adding the image-text matching and language modeling losses is beneficial to representation learning for retrieval, and thus, for critic performance as it is based on the retrieval capabilities. Moreover, as even the baseline we train significantly beats the zero-shot use of PubMed CLIP, we confirm that training our own models for this specific task is beneficial.

The results of our investigation of the critic's capabilities to separate true and false image-NLE pairs based on image-text similarity are summarized in Table 4.2. As the zero-shot PubMed CLIP baseline did not have good retrieval results, it is not able to capture NLE relevance at all here: there is no difference in mean similarity between positive and negative image-NLE pairs, we get an AUC score of 0.5 for the binary classification task based on the similarity and the Mann-Whitney U-test fails. However, for both ground-truth NLEs and for the NLEs generated by our NLE-generation pipeline, both our BLIP-2 loss trained and CLIP loss trained models learn features such that the true and false pairs are separable based on image-text similarity (confirmed by both AUC and Mann-Whitney U-tests). There is a clear distinction in similarity, as illustrated by the mean similarities, between both kinds of pairs. The best AUC score (and difference in mean similarities) is achieved by the Q-former trained with the combined BLIP-2 loss

in both settings, confirming that this combined loss is beneficial to learn stronger image and text representations that capture NLE relevance and faithfulness better.

Additionally, we observe that there is only a small difference in mean similarity for positive and negative pairs using NLEs generated the blind model (BioGPT language model only taking the textual prompt as an input), and AUC scores are also very close to 0.5. This model generates unfaithful NLEs even for "true" pairs (generating an NLE for a label that is true for the given chest X-ray) by design as it never gets the image signal to generate the explanation. The image-text-similarity-based critic models seem to capture this as well, as the unfaithful generated explanations do not have very different similarities from randomly shuffled false NLEs, while our visually-grounded NLE generators create faithful NLEs for true pairs that can be separated from the unfaithful NLEs generated for labels that are negative for the X-rays.

Our Q-former is thus able to capture the faithfulness of an explanation to a medical image based on image-text representation similarities, and can therefore be leveraged as an explanation relevance critic in our self-rationalized approach to chest X-ray diagnosis.

CHAPTER 5

Leveraging the critic: An end-to-end trained "explain-then-predict" self-rationalized approach to chest X-ray diagnosis

We validated that our model can generate NLEs for positive labels of chest X-rays (Chapter 3). Additionally, we demonstrated that the Q-former component can serve as a critic that measures if an NLE is faithful and relevant to the visual content of a chest X-ray in Chapter 4. Assuming that these components perform as expected, we construct a self-rationalized "explain-then-predict" approach to CXR diagnosis, where candidate NLEs are generated for each possible label, and the diagnosis decision is made based on the faithfulness of the explanation to the radiological image. This chapter describes how we implement and validate this method.

5.1 Problem setting

In our full pipeline, the main task we solve is multi-label chest X-ray classification, while additionally generating natural language explanations for positive diagnoses. The goal is to build a model that accepts a chest X-ray image as an input, and that outputs class probabilities for the seven classes of the dataset, along with natural language explanations for the classes having probabilities exceeding a certain threshold. We highlight that classification is the main task being solved and evaluated, and alternative models are therefore not required to generate NLEs for positive diagnoses.

5.2 Method

We aim to assess whether learning to generate natural language explanations enhances chest X-ray classification performance. In order to achieve this, we integrate our NLE generation and critic components into a self-rationalized pipeline. We adopt an "explain-then-predict" approach, as illustrated in Figure 1.2, which outlines the inference flow. To generate predictions for a chest X-ray, we decode an NLE for every possible label, and then measure the image-NLE similarities with the critic Q-former. We base label predictions on similarities that reflect the relevance of generated explanations to the image: we first explain every pathology, and predict based on how relevant and truthful to the image the explanations are.

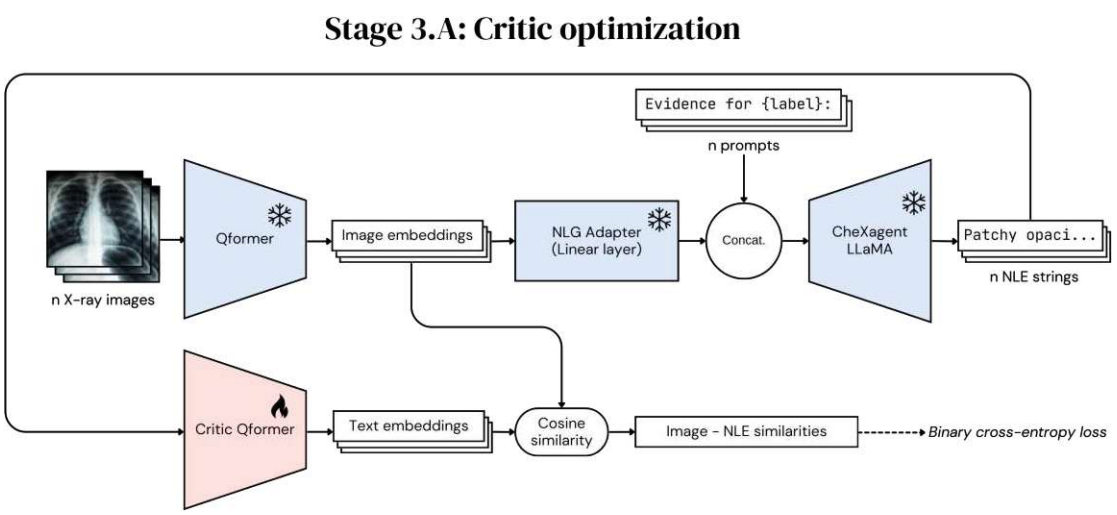

Figure 5.1: First stage of end-to-end training: the critic Q-former is optimized based on its discrimination performance of generated NLEs.

Since this pipeline assembles previously trained components, it can function without additional training. We however decide to further train the end-to-end model. As mentioned in Section 3.4, we use two separate sets of weight for the Q-former used for natural language generation and for the one used as a critic, as further optimizing the Q-former NLG degraded critic capabilities. In memory constrained environments, a single Q-former can be used with a tradeoff to be made by freezing the Q-former at a given point during nle-generation training to preserve it. We first further train the critic Q-former for the classification task within the end-to-end framework. As pictured in Figure 5.1, we do this by, for every image of each batch, decoding NLEs for all 7 labels of the dataset, and embedding the image and candidate NLEs with the critic Q-former. Similarities between the image and NLEs are measured and used as output logits for the labels corresponding to the NLEs. We compute a binary cross-entropy loss for those predictions, and propagate it back through the network to the critic Q-former that is the only part of the pipeline with trainable parameters at this stage.

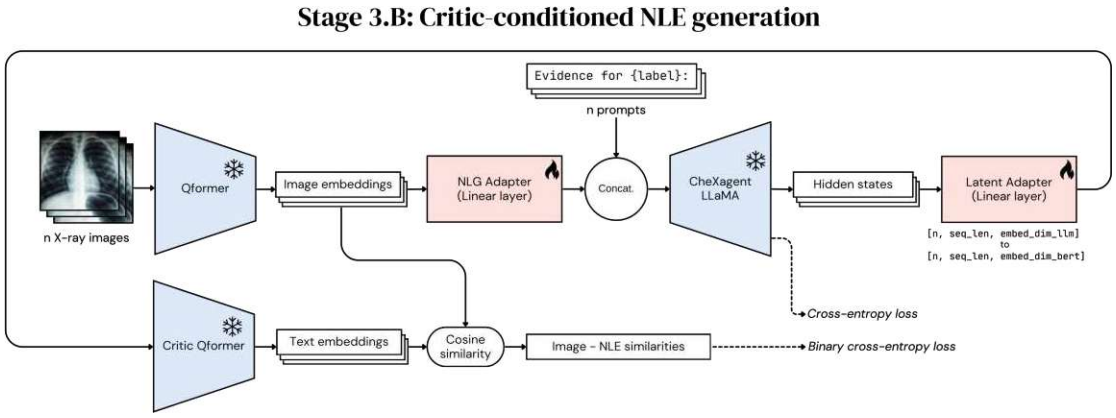

Figure 5.2: Second (optional) stage of end-to-end training: the natural language generation pipeline is conditioned on the critic's loss using the Gumbel-Softmax trick to allow backpropagation into the NLG components.

We then additionally find that we can also leverage the critic's gradients to improve the NLE generation part of the pipeline itself in this end-to-end setting in an additional training stage following the critic optimization. This is achieved by using the same setup, but this time integrating the natural language generation cross-entropy loss in addition to the classification binary cross-entropy loss and only training the NLG adapter parameters (the image-embedding projection layer before the language model input), as summarized in Figure 5.2. However, this configuration implies backpropagating the gradients from the critic back to early components of the NLG part of the network that are placed before the language model, while sampling tokens is not differentiable. To circumvent this, we leverage the Gumbel-Softmax trick [184]. Instead of decoding a sequence by sampling tokens from the language model output probability distribution, we obtain a continuous and differentiable approximation of the language model's output by perturbing the logits with Gumbel noise and applying a softmax function, which allows us to backpropagate gradients through the LLM's outputs and to train the entire pipeline end-to-end without breaking the flow of gradient information. Specifically, we define the categorical distribution with class probabilities π_i for $i = 1, \ldots, k$, where k is the number of possible output classes (tokens). To draw samples that are differentiable, we add Gumbel noise g_i to the logits z_i of the language model and apply the softmax function to approximate the sampling. This can be described as follows:

$$y_i = \frac{\exp\left((z_i + g_i)/\tau\right)}{\sum_{j=1}^{k} \exp\left((z_j + g_j)/\tau\right)}, \tag{5.1}$$

where g_i are i.i.d. samples from the Gumbel distribution, i.e.,

$$g_i = -\log\left(-\log\left(U_i\right)\right), \quad U_i \sim \text{Uniform}(0, 1). \tag{5.2}$$

The temperature parameter $\tau > 0$ controls the smoothness of the approximation. As $\tau \to 0$, the Gumbel-Softmax distribution approaches a one-hot categorical sample,

whereas for higher values of τ, the distribution becomes smoother, allowing for more effective gradient-based optimization during training. The approximation of the model output is fed through a linear layer (referred to as "LM adapter") to re-align it with the critic Q-former's input space, as the Q-former is trained taking BERT-tokenized sequences as an input.

This method has the benefit of generating explanations that are faithful to the model by design as the prediction was made based on the explanations themselves and how they relate to the image. Moreover, using the Q-former we trained as part of the BLIP-2-based NLE-generation pipeline as a critic, leveraging the strong features we learn in the first stage of training, also has the advantage of not requiring the use of further computing resources on training an additional critic model from scratch, and instead getting more value of that stage of training that is required before learning to generate NLEs.

5.3 Experimental setup

5.3.1 Dataset and processing

We perform all experiments on the official splits on the MIMIC-NLE dataset. For this stage of training, we obtain the best results by dropping the data augmentations we used in the previous steps. We only normalize the pixel values based on the train split's value distribution. At training and inference, each chest X-ray image of the dataset is considered as a sample. For critic tuning, the image is provided with the values of the 7 class labels. As for the NLG training based on the critic, we also include the NLEs for positive labels when they are available to be able to compute the language modeling cross-entropy loss.

5.3.2 Evaluation metrics

To measure multi-label chest X-ray classification performance, we rely on the area under the curve (AUC) score. We focus on this score as it quantifies the model's ability to distinguish between classes across different classification thresholds, whereas scores like F1, accuracy or recall depend on chosing an optimal classification threshold which is not the focus here. Moreover, works in the literature most often report AUC scores for chest X-ray classification. Using the same metric enables fairer comparison. We report AUC for each individual label as well as the average score across classes for each model.

5.3.3 Baselines

To evaluate the classification performance of our method, we include multiple CXR classification baselines. We train three simple classification baselines that follow the pattern pictured in Figure 5.3: the X-ray is embedded with a vision encoder, and the embedding is fed into a classification head that is either a projection layer or a multi-layer perceptron. All parameters are trainable and a binary cross-entropy loss is employed for training. We train three such baselines: one uses CheXagent's pre-trained vision encoder,

another uses the chest-X-ray-pretrained DenseNet-121 provided by TorchXRayVision [180], and the third one uses the same DenseNet architecture but initializes the weights with those from the captioning baseline based on the DenseNet and BioGPT we have trained in Chapter 3. The last baseline allows us to explore if the features learnt when training for NLE generation are beneficial to classification performance.

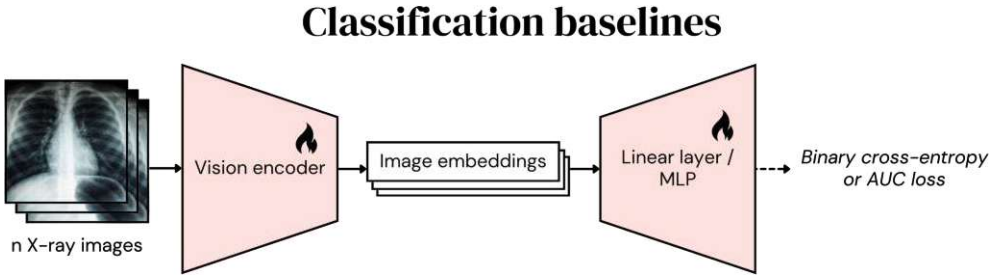

Figure 5.3: Overall architecture of our classification baselines: image features are extracted and fed into a simple classification head while training all model parameters.

We also include zero-shot prompting of CheXagent as a baseline. The foundation model is prompted for the presence of each disease, and the class probability is derived from the logits of "yes" or "no" equivalent tokens when answering the question. For positive labels, we also prompt the model to generate natural language explanations.

Finally, we re-train the currently state-of-the-art non-ensemble chest X-ray classification model on our dataset to fairly verify how our method compares to the best performing classification-specialized models. As mentioned in Chapter 1, we choose the DeepAUC [17] method that tops the CheXpert multi-label chest X-ray classification benchmark when ensemble-methods are not taken into account [77]. This approach further trains a pre-trained DenseNet-201 with a classification head by directly optimizing the AUC score instead of using a cross-entropy loss which may not align with maximizing AUC, especially in medical imaging where class imbalances are common [17].

5.3.4 Training setup

For the first stage of end-to-end training where only the critic Q-former is tuned, we initialize the network with the weights from NLE-generation training described in Chapter 3 and the critic Q-former with the weights obtained by pre-training the Q-former as described in Chapter 4. All weights are loaded in 32-bit floating point precision except for the vision encoder that is in 16-bit floating point precision. Training is performed with a batch size of 128 images on 8 Nvidia A100 GPUs equipped with 80GB of VRAM each, leading to a total batch size of 1024. The AdamW optimizer is employed using a weight decay factor of 0.01 and a cosine-scheduled learning rate ranging between $5e^{-6}$ and $5e^{-7}$, and gradients are clipped to a norm of 0.75. Early stopping occurs during the 18th epoch of training.

As for the additional stage of end-to-end training where the NLG part of the network learns from the critic's signal, we initialize the network from the resulting checkpoint of the first stage of end-to-end training. We train on the same hardware, but using a batch size of 8, leading to a total batch size of 64. We still clip the gradients to a norm of 0.75, but use a learning rate of $3e^{-5}$ for the LM adapter that reprojects the Gumbel-Softmax approximation and $1e^{-5}$ for the adapter that is part of the NLG pipeline. Unfreezing the NLG Q-former at this stage has not been beneficial in our experiments. The AdamW weight decay factor is set to 0.05.

Both DenseNet-based baselines are trained with a batch size of 128 using a learning rate of $1e^{-4}$. The version using NLE-generation-pretrained weights is trained for 4 epochs, while the other model is early stopped after 7 epochs. The CheXagent-vision-encoder-based baseline uses a batch size of 32 for 3 epochs with a learning rate of $1e^{-5}$.

Finally, the DeepAUC baseline is trained for 20 epochs using a batch size of 32, with the AUC optimizer's learning rate set to 0.1, an epoch decay of 0.002, a decay of 1.2, a margin of 1 and a weight decay of $1e^{-5}$. These parameters have been based on the descriptions provided in the DeepAUC publication [17] and further parameter tuning on the MIMIC-NLE dev set to provide a fair comparison to our method.

5.4 Results and discussion

The training process for the critic Q-former converges smoothly, with a stable increase in per-class AUC scores. The additional stage of end-to-end training behaved with less stability in early experiments due to the fact that for many positive labels, NLEs are not available in the MIMIC-NLE dataset. At smaller batch sizes, this led to often having batches with very little NLEs or no NLEs at all. Similarly, some batches happened to not have any positive labels. Larger batches were key in this setting, and this required optimization and some engineering tricks to achieve as this stage of end-to-end training is particularly memory-heavy. During the first stage of end-to-end training, as the NLG components are all frozen, we were able to decode NLEs for all candidate labels of all images of the dataset once and cache them for all subsequent epochs, allowing large batch sizes, which is not possible in this second stage that trains NLE generation components. With a batch size of 8×8, the training stabilized: losses converged more smoothly, validation AUC scores grew and over 30 epochs of training, only a single batch was encountered where no NLE was available, and there was no batch without positive labels.

The classification performance on the test split of MIMIC-NLE, measured by per-class and mean AUC, is reported in Table 5.1 and Table 5.2. A first observation is that learning to generate NLEs for classification labels seems to be beneficial to the task of chest X-ray classification. The DenseNet-based baseline where the DenseNet checkpoint from our captioning baseling was used to initialize the vision encoder weights outperforms the same DenseNet baseline where the same vision encoder was initialized with the chest X-ray specialized weights used to initialize the captioning baseline as well. As expected, the state-of-the-art DeepAUC model beats the performance of all of our baselines.

Class	DeepAUC	CheXagent	Ours (critic tuned)	Ours (NLG tuned)	CheXagent vis. encoder	DenseNet vis. encoder	DenseNet vis. encoder (from captioning weights)
Pulmonary Edema	0.79	0.79	**0.82**	**0.82**	0.66	0.77	0.78
Vascular congestion	0.71	0.68	0.72	**0.73**	0.51	0.67	0.68
Pneumonia	0.66	0.62	0.68	**0.69**	0.59	0.61	0.60
Aspiration	0.62	0.62	**0.63**	0.62	0.55	0.44	0.46
Atelectasis	**0.70**	0.64	0.68	0.69	0.68	0.65	0.68
Pleural effusion	**0.85**	0.83	0.81	0.82	0.81	0.80	0.81
copd/emphysema	**0.90**	0.71	0.89	0.88	0.77	0.74	0.84

Table 5.1: Per-class chest X-ray classification performance for each label. AUC scores for our baselines (vision encoder + projection layer or MLP, zero-shot CheXagent prompting), the state-of-the art CXR classifier (DeepAUC) and our end-to-end trained model (both with only critic end-to-end training and additional natural language generation training conditioned on the critic loss). The AUC value for each class (row) is in bold.

Metric	DeepAUC	CheXagent	Ours (critic tuned)	Ours (NLG tuned)	CheXagent vis. encoder	DenseNet vis. encoder	DenseNet vis. encoder (from captioning weights)
Mean AUC (all classes)	0.747	0.699	0.747	**0.750**	0.653	0.669	0.693

Table 5.2: Overall classification performance comparison with mean AUC scores.

Model	BLEU-1 ↑	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-L ↑
Ours, critic tuned	**0.2184**	**0.0584**	**0.2853**	0.2585
Ours, NLG tuned	0.2015	0.0548	**0.2853**	**0.2611**

Table 5.3: NLG metrics after each stage of end-to-end training of the pipeline. NLEs are only evaluated for positive samples. The best value for each metric is in bold.

Our explain-then-predict pipeline, after further training the critic, matches the overall performance of that state-of-the-art model, confirming that **our self-rationalized approach is able to solve the classification task effectively while additionally generating faithful explanations in natural language by design**.

We also show that by further training the natural language generation part of the pipeline based on the critic's signal, we align the generated NLEs with the critic's diagnoses better which leads to beating the best X-ray classification model, DeepAUC, achieving the best classification performance across all classes, with a slightly better mean AUC score of 0.75 on the MIMIC-NLE test split. Conditioning natural language explanation on the critic's signal is beneficial to the overall classification task. With this second stage of end-to-end training, we show that a self-rationalized chest X-ray classifier can beat state of the art CXR classification models while generating faithful NLEs. These NLEs can be considered faithful as the model makes its prediction based on them and their relevance to the input image directly, but also because in Chapter 4, we have shown that unfaithful NLEs (generated by the blind baseline) are not sufficient to perform the classification task. Our end-to-end system achieving the best classification performance therefore further supports the fact that that the generated NLEs are faithful.

However, the improvement of the classification pipeline by further adapting natural language explanation generation based on the critic's loss does not explicitly reflect in better NLG scores for the generated NLEs. As shown in Table 5.3, the BLEU and ROUGE scores remain very similar, with slightly lower BLEU scores and higher ROUGE scores. As we mentioned earlier, these metrics do not perfectly measure NLE quality in a clinical setting and are more of a proxy of good NLE generation, so this might not fully reflect how this stage influences NLE quality itself. However, given how the first stage of end-to-end training where only the critic is optimized is already sufficient to match the state-of-art classifier performance, and how this last stage is by far the longest most compute-intensive training stage of our method, this last step of further training the NLG pipeline on critic signal can be considered optional in practical applications. Based on the system's priority, a tradeoff can be made: it might be worth doing this part of training if classification performance is key and resources are available, but in a resource-constrained setting, the model achieves state-of-the-art performance with only the first step of end-to-end training.

CHAPTER 6

Conclusion And Future Work

In this thesis, we explored a novel "explain-then-predict" approach to chest X-ray (CXR) diagnosis that integrates natural language explanation (NLE) generation into the diagnosis process. Our method addresses the limitations of traditional single-stage neural networks and post hoc interpretability techniques, which often lack the transparency and explainability essential for medical applications. We aimed to construct a self-rationalizing system that not only classifies CXRs but also provides faithful and medically relevant explanations for its decisions.

6.1 Key results

Through a series of experiments described in the thesis, we achieved the following key results:

1. **NLE Generation and Faithfulness**: We demonstrated the feasibility of generating natural language explanations for CXR diagnoses, using a vision-language architecture based on BLIP-2. Our model outperformed simpler baseline models, such as those using dense vision encoders and medical language models, as well as published results on the MIMIC-NLE dataset, on BLEU and ROUGE metrics. However, we also noted the limitations of these metrics in capturing the clinical relevance of explanations. Notably, we observed that simple language models trained without considering the image input produced unfaithful but persuasive NLEs. This finding underscores the importance of being able to capture NLE faithfulness.

2. **Image-Text Similarity as a Critic**: In response to the challenge of unfaithful explanations, we introduced an NLE critic module that judges the relevance of an explanation to a chest X-ray by leveraging the Q-former component of the BLIP-2

architecture as it performed as a better critic than alternative vision-language models. We trained the Q-former using a combination of image-text contrastive, image-text matching, and captioning losses to align the representations of images and texts. Our experiments showed that the critic effectively captured the relevance of NLEs to CXRs, separating faithful explanations from unfaithful ones accurately. Our method showed that false NLEs, whether generated by unfaithful models or randomly shuffled pairs, had significantly lower image-text similarity than true NLEs.

3. **Self-Rationalized "Explain-Then-Predict" Approach to classification**: By combining NLE generation with a critic that measures the faithfulness of explanations, we built a self-rationalized diagnostic model that first generates explanations and then makes classification predictions based on the relevance of those explanations to the input image. We further trained the entire system end-to-end in two stages. First, we trained the critic on the classification task using the generated NLEs, and we then conditioned the NLE generation on the critic's feedback. These steps improved classification performance, achieving a mean AUC score of 0.75 and outperforming the state-of-the-art DeepAUC [17] classifier, while also generating faithful NLEs.

4. **Multimodal Learning Benefits**: We also observed that training the vision-language model in an end-to-end manner, where the critic's feedback was propagated to the NLE generation module, led to improvements in both classification performance and explanation quality. Additionally, initializing the vision encoder of a classification model with weights resulting from learning NLE generation also led to better performance than weights that result from classification pretraining. The integration of multimodal learning showed that NLE generation and chest X-ray classification tasks benefit from each other, as improvements in one domain contributed positively to the other.

5. **Evaluation on the MIMIC-NLE Dataset**: Throughout our experiments, we used the MIMIC-NLE dataset, which contains 38,003 image-NLE pairs. Our model consistently outperformed simpler models and other baselines in terms of natural language generation metrics, such as BLEU and ROUGE, as well as classification performance, achieving competitive results with the current best-performing CXR classifiers. Notably, our method also improved the interpretability of the model's predictions, providing a robust, explainable solution for real-world clinical settings.

In summary, we have developed a novel multimodal architecture that not only improves chest X-ray classification accuracy but also generates natural language explanations that are both faithful to the medical image and relevant to the clinical decision-making process. Our "explain-then-predict" approach provides a step toward more explainable and usable machine learning models in healthcare, particularly in the context of medical imaging.

6.2 Limitations and future works

While our approach yields promising results, several areas offer potential for further improvement. Firstly, our model was trained and evaluated on a subset of pathologies where a sufficient amount of NLEs was available in our dataset, limiting its generalizability to less represented conditions. This choice demonstrated the technique's capabilities when data is available. Future work could explore handling rarer pathologies using few-shot learning, data augmentation, or synthetic data generation. Of course, further work on creating larger NLE datasets for chest X-rays would be ideal.

Secondly, the improvement in AUC scores in comparison to the state-of-the-art baseline that we achieve are small. These incremental improvements however seem to be typical on benchmarks like CheXagent, where the differences in the leaderboard are in a similar magnitude [77]. More importantly, our goal was not to surpass the state-of-the-art classification model but to ensure our self-rationalized approach achieved comparable performance without sacrificing accuracy for explainability. In that sense, our goal is achieved and this small improvement in AUC is not a problem and does not change our conclusions. However, a useful follow-up work would also be to test our method on other datasets to further explore how the method generalizes on data with different pathologies and different biases.

While we integrate information from two different modalities in our model, another opportunity for future work would be to explore how additional information could be included in the model and how it could improve performance. For example, patient context could be added as unstructured input in the prompt or through a structured format, or longitudinal data could be supported by the model as it is often available in radiological settings.

Additionally, it would be beneficial to validate the model's performance in real-world clinical settings in another study. While the MIMIC-NLE dataset offers a solid foundation for testing, real-world data from clinical radiology departments could provide further insights into the model's effectiveness and reliability in diverse scenarios. Collaboration with radiologists and clinicians in a user study to validate the generated explanations would help ensure that the model's outputs are both clinically relevant and useful for medical professionals in practice.

Lastly, our approach remains computationally expensive, especially the last end-to-end training stage involving the Gumbel-Softmax trick to propagate the critic loss to NLE-generation components. The full inference pipeline is large (about 8.5b parameters) and by the nature of the explain-then-predict approach, to predict labels for a single image, as many sequences have to be decoded as there are labels. While the decoding cost is offset by the very short sequence length, this remains a costly architecture for a classification task. Training the entire solution has a non-negligible environmental impact.

We estimate that training the entire pipeline produces 88 kg of CO_2 equivalents [1]. However, 82% of those emissions are generated by the last stage of end-to-end training that we have proposed as optional as explained in Section 5.4. Therefore, we highlight the importance of weighing the benefit of this step when using this method, as well as the opportunity for further research in making this step more efficient, potentially by further exploring using PPO for this step as we mentioned attempting in Section 5.4. While we believe the computational and environmental cost is justified in clinical settings where faithful explanations significantly enhance classification labels' utility, there is a lot of potential for further work that makes this setup more efficient. For example, scaling down the architecture and its components and how much that impacts performance could be explored, as well as distilling our model's knowledge into a smaller one. Generating more synthetic NLEs for training could also potentially help offset a loss in performance when scaling down the model size. As this thesis focused on showing the capabilities of the approach, we believe there are a lot of opportunities to make it even more efficient while preserving its advantages.

[1] This is based on the formula provided by [185], using an estimation of CO_2 per kWh in Austria provided by [186] as the hardware we used is hosted in that country. This calculation estimates the cost of training all components of the pipeline once, and does not take iteration and other experiments into account.

Overview of Generative AI Tools Used

In this document, text-generation models and chatbots accessed through ChatGPT [187] or Claude [188] have only been used as an aid for specific sentence improvement and refinement, to iterate on formulations or to brainstorm key ideas, or to refine the latex presentations of equations that formalize some concepts in this work, and not to generate blocks of multiple sequences or full paragraphs.

In general, to address specific sentences or chunks that might have needed improvement, the following prompt template has been used:

```
I will give you parts of my master thesis manuscript.
Please carefully read through them and list any mistakes, poor
wording, or issues in writing that should be addressed.
List every modification you would make as a bullet point.
Content:  {CONTENT}
```

The output was a bullet list of potential mistakes and problems that were then manually integrated and taken into account instead of letting the model rewrite the sequences.

Additionally, ChatGPT and DeepL [189] have been used to assist in the translation and cleaning up of the German version of the Abstract, by using DeepL to generate a rough translation, cleaning and improving it manually, and prompting ChatGPT with that German version and the English abstract to fix mistakes, poor formulations and any discrepancies in content.

List of Figures

List of Tables

Bibliography

[1] Cindy S. Lee, Paul G. Nagy, Sallie J. Weaver, and David E. Newman-Toker. "Cognitive and System Factors Contributing to Diagnostic Errors in Radiology". In: *American Journal of Roentgenology* 201.3 (Sept. 2013), pp. 611–617. ISSN: 0361-803X. DOI: 10.2214/AJR.12.10375. (Visited on 10/14/2024).

[2] Jarrel C. Y. Seah et al. "Effect of a Comprehensive Deep-Learning Model on the Accuracy of Chest x-Ray Interpretation by Radiologists: A Retrospective, Multireader Multicase Study". In: *The Lancet Digital Health* 3.8 (Aug. 2021), e496–e506. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(21)00106-0. (Visited on 11/25/2024).

[3] Dulani Meedeniya, Hashara Kumarasinghe, Shammi Kolonne, Chamodi Fernando, Isabel De la Torre Díez, and Gonçalo Marques. "Chest X-ray Analysis Empowered with Deep Learning: A Systematic Review". In: *Applied Soft Computing* 126 (Sept. 2022), p. 109319. ISSN: 1568-4946. DOI: 10.1016/j.asoc.2022.109319. (Visited on 11/25/2024).

[4] Maria Frasca, Davide La Torre, Gabriella Pravettoni, and Ilaria Cutica. "Explainable and Interpretable Artificial Intelligence in Medicine: A Systematic Bibliometric Review". In: *Discover Artificial Intelligence* 4.1 (Feb. 2024), p. 15. ISSN: 2731-0809. DOI: 10.1007/s44163-024-00114-7. (Visited on 11/26/2024).

[5] Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. "Machine Learning in Medicine: Addressing Ethical Challenges". In: *PLoS Medicine* 15.11 (Nov. 2018), e1002689. ISSN: 1549-1277. DOI: 10.1371/journal.pmed.1002689. (Visited on 11/26/2024).

[6] Alessandro Wollek, Robert Graf, Saša Čečatka, Nicola Fink, Theresa Willem, Bastian O. Sabel, and Tobias Lasser. "Attention-Based Saliency Maps Improve Interpretability of Pneumothorax Classification". In: *Radiology: Artificial Intelligence* 5.2 (Mar. 2023), e220187. ISSN: 2638-6100. DOI: 10.1148/ryai.220187. (Visited on 11/27/2024).

[7] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. "Multi-Modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training". In: *IEEE Journal of Biomedical and Health Informatics* 26.12 (Dec. 2022), pp. 6070–6080. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2022.3207502. (Visited on 11/27/2024).

[8] Akhila Narla, Brett Kuprel, Kavita Sarin, Roberto Novoa, and Justin Ko. "Automated Classification of Skin Lesions: From Pixels to Practice". In: *Journal of Investigative Dermatology* 138.10 (Oct. 2018), pp. 2108–2110. ISSN: 0022-202X. DOI: 10.1016/j.jid.2018.06.175. (Visited on 11/26/2024).

[9] Julia K. Winkler et al. "Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition". In: *JAMA Dermatology* 155.10 (Oct. 2019), pp. 1135–1141. ISSN: 2168-6068. DOI: 10.1001/jamadermatol.2019.1735. (Visited on 11/26/2024).

[10] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. *Program Induction by Rationale Generation : Learning to Solve and Explain Algebraic Word Problems*. May 2017. (Visited on 04/28/2024).

[11] William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P. Bradley, and Lyle J. Palmer. *Producing Radiologist-Quality Reports for Interpretable Artificial Intelligence*. June 2018. DOI: 10.48550/arXiv.1806.00340. arXiv: 1806.00340. (Visited on 10/14/2024).

[12] Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartlomiej Papiez, and Thomas Lukasiewicz. "Explaining Chest X-Ray Pathologies in Natural Language". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li. Vol. 13435. Cham: Springer Nature Switzerland, 2022, pp. 701–713. ISBN: 978-3-031-16442-2 978-3-031-16443-9. DOI: 10.1007/978-3-031-16443-9_67. (Visited on 11/27/2024).

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models". In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, July 2023, pp. 19730–19742. (Visited on 11/27/2024).

[14] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, July 2021, pp. 8748–8763. (Visited on 11/27/2024).

[15] Zhihong Chen et al. *CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation*. Jan. 2024. arXiv: 2401.12208 [cs]. (Visited on 04/22/2024).

[16] Jeremy Irvin et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 590–597. ISSN: 2374-3468. DOI: 10.1609/aaai.v33i01.3301590. (Visited on 11/27/2024).

[17] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. "Large-Scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 3020–3029.

ISBN: 978-1-66542-812-5. DOI: 10.1109/ICCV48922.2021.00303. (Visited on 11/27/2024).

[18] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. "MIMIC-CXR, a de-Identified Publicly Available Database of Chest Radiographs with Free-Text Reports". In: *Scientific Data* 6.1 (Dec. 2019), p. 317. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0322-0. (Visited on 05/14/2024).

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. (Visited on 11/27/2024).

[20] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. Feb. 2023. DOI: 10.48550/arXiv.2302.13971. arXiv: 2302.13971 [cs]. (Visited on 05/01/2024).

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. (Visited on 05/01/2024).

[22] Christian Szegedy et al. "Going Deeper with Convolutions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, June 2015, pp. 1–9. ISBN: 978-1-4673-6964-0. DOI: 10.1109/CVPR.2015.7298594. (Visited on 11/27/2024).

[23] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *3rd International Conference on Learning Representations (ICLR 2015)* (2015). (Visited on 11/27/2024).

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 770–778. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.90. (Visited on 11/27/2024).

[25] Alexey Dosovitskiy et al. *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. June 2021. arXiv: 2010.11929 [cs]. (Visited on 05/01/2024).

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002. ISBN: 978-1-66542-812-5. DOI: 10.1109/ICCV48922.2021.00986. (Visited on 11/27/2024).

[27] Ilya O Tolstikhin et al. “MLP-Mixer: An All-MLP Architecture for Vision”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 24261–24272. (Visited on 11/27/2024).

[28] José Maurício, Inês Domingues, and Jorge Bernardino. “Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review”. In: *Applied Sciences* 13.9 (Jan. 2023), p. 5521. ISSN: 2076-3417. DOI: 10.3390/app13095521. (Visited on 05/01/2024).

[29] Mostafa Dehghani et al. “Scaling Vision Transformers to 22 Billion Parameters”. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, July 2023, pp. 7480–7512. (Visited on 11/27/2024).

[30] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. “Big Transfer (BiT): General Visual Representation Learning”. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Vol. 12350. Cham: Springer International Publishing, 2020, pp. 491–507. ISBN: 978-3-030-58557-0 978-3-030-58558-7. DOI: 10.1007/978-3-030-58558-7_29. (Visited on 11/27/2024).

[31] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. “A Simple Framework for Contrastive Learning of Visual Representations”. In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 1597–1607. (Visited on 11/27/2024).

[32] Jean-Bastien Grill et al. “Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 21271–21284. (Visited on 11/27/2024).

[33] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. “Masked Autoencoders Are Scalable Vision Learners”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 15979–15988. ISBN: 978-1-66546-946-3. DOI: 10.1109/CVPR52688.2022.01553. (Visited on 11/27/2024).

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc., 2015. (Visited on 11/27/2024).

[35] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. “Mask R-CNN”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322. (Visited on 11/27/2024).

[36] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. “End-to-End Object Detection with Transformers”. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Vol. 12346. Cham: Springer International Publishing, 2020, pp. 213–229. ISBN: 978-3-030-58451-1 978-3-030-58452-8. DOI: 10.1007/978-3-030-58452-8_13. (Visited on 11/27/2024).

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28.

[38] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. "Detecting Twenty-Thousand Classes Using Image-Level Supervision". In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner. Vol. 13669. Cham: Springer Nature Switzerland, 2022, pp. 350–368. ISBN: 978-3-031-20076-2 978-3-031-20077-9. DOI: 10.1007/978-3-031-20077-9_21. (Visited on 11/27/2024).

[39] Alexander Kirillov et al. "Segment Anything". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 3992–4003. ISBN: 9798350307184. DOI: 10.1109/ICCV51070.2023.00371. (Visited on 11/27/2024).

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. (Visited on 11/27/2024).

[41] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*. Sept. 2013. DOI: 10.48550/arXiv.1301.3781. arXiv: 1301.3781 [cs]. (Visited on 05/01/2024).

[42] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. (Visited on 05/01/2024).

[43] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. "Skip-Thought Vectors". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15. Cambridge, MA, USA: MIT Press, Dec. 2015, pp. 3294–3302. (Visited on 11/27/2024).

[44] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 670–680. DOI: 10.18653/v1/D17-1070. (Visited on 11/27/2024).

[45] Colin Raffel et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *J. Mach. Learn. Res.* 21.1 (Jan. 2020), 140:5485–140:5551. ISSN: 1532-4435.

[46] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. July 2019. DOI: 10.48550/arXiv.1907.11692. arXiv: 1907.11692 [cs]. (Visited on 05/01/2024).

[47] Jordan Hoffmann et al. “Training Compute-Optimal Large Language Models”. In: ().

[48] OpenAI et al. *GPT-4 Technical Report*. Mar. 2024. DOI: 10.48550/arXiv.2303.08774. arXiv: 2303.08774 [cs]. (Visited on 05/01/2024).

[49] Albert Q. Jiang et al. *Mistral 7B*. Oct. 2023. DOI: 10.48550/arXiv.2310.06825. arXiv: 2310.06825 [cs]. (Visited on 05/01/2024).

[50] Gemini Team et al. *Gemini: A Family of Highly Capable Multimodal Models*. Apr. 2024. DOI: 10.48550/arXiv.2312.11805. arXiv: 2312.11805 [cs]. (Visited on 05/01/2024).

[51] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. “Sigmoid Loss for Language Image Pre-Training”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 11941–11952. ISBN: 9798350307184. DOI: 10.1109/ICCV51070.2023.01100. (Visited on 11/27/2024).

[52] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. “Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm”. In: (2022).

[53] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. “Image Captioners Are Scalable Vision Learners Too”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., May 2024, pp. 46830–46855. (Visited on 11/27/2024).

[54] L. Srinivasan and Dinesh Sreekanthan. “Image Captioning-A Deep Learning Approach”. In: 2018. (Visited on 05/02/2024).

[55] Hyeryun Park, Kyungmo Kim, Jooyoung Yoon, Seongkeun Park, and Jinwook Choi. “Feature Difference Makes Sense: A Medical Image Captioning Model Exploiting Feature Difference and Tag Information”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Ed. by Shruti Rijhwani, Jiangming Liu, Yizhong Wang, and Rotem Dror. Online: Association for Computational Linguistics, July 2020, pp. 95–102. DOI: 10.18653/v1/2020.acl-srw.14. (Visited on 05/02/2024).

[56] Dexin Zhao, Zhi Chang, and Shutao Guo. “A Multimodal Fusion Approach for Image Captioning”. In: *Neurocomputing* 329 (Feb. 2019), pp. 476–485. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2018.11.004. (Visited on 05/02/2024).

[57] David Lyndon, Ashnil Kumar, and Jinman Kim. "Neural Captioning for the ImageCLEF 2017 Medical Image Challenges". In: ().

[58] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. "Align before Fuse: Vision and Language Representation Learning with Momentum Distillation". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 9694–9705. (Visited on 05/02/2024).

[59] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. "UNITER: UNiversal Image-TExt Representation Learning". In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Cham: Springer International Publishing, 2020, pp. 104–120. ISBN: 978-3-030-58577-8. DOI: 10.1007/978-3-030-58577-8_7.

[60] Xiujun Li et al. "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks". In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Cham: Springer International Publishing, 2020, pp. 121–137. ISBN: 978-3-030-58577-8. DOI: 10.1007/978-3-030-58577-8_8.

[61] Jean-Baptiste Alayrac et al. "Flamingo: A Visual Language Model for Few-Shot Learning". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., Apr. 2024, pp. 23716–23736. ISBN: 978-1-71387-108-8. (Visited on 11/27/2024).

[62] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual Instruction Tuning". In: *Advances in Neural Information Processing Systems* 36 (Dec. 2023), pp. 34892–34916. (Visited on 11/27/2024).

[63] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation". In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR, June 2022, pp. 12888–12900. (Visited on 11/27/2024).

[64] Jinze Bai et al. *Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond*. Oct. 2023. DOI: 10.48550/arXiv.2308.12966. arXiv: 2308.12966 [cs]. (Visited on 05/02/2024).

[65] Xi Chen et al. "On Scaling Up a Multilingual Vision and Language Model". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2024, pp. 14432–14444. ISBN: 9798350353006. DOI: 10.1109/CVPR52733.2024.01368. (Visited on 11/27/2024).

[66] Weihan Wang et al. *CogVLM: Visual Expert for Pretrained Language Models*. Feb. 2024. DOI: 10.48550/arXiv.2311.03079. arXiv: 2311.03079 [cs]. (Visited on 05/02/2024).

[67] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. "SEED-Bench: Benchmarking Multimodal Large Language Models". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2024, pp. 13299–13308. ISBN: 9798350353006. DOI: 10.1109/CVPR52733.2024.01263. (Visited on 11/27/2024).

[68] Chaoyou Fu et al. *MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models*. Mar. 2024. DOI: 10.48550/arXiv.2306.13394. arXiv: 2306.13394 [cs]. (Visited on 05/02/2024).

[69] Alvin Rajkomar, Sneha Lingam, Andrew G. Taylor, Michael Blum, and John Mongan. "High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks". In: *Journal of Digital Imaging* 30.1 (Feb. 2017), pp. 95–101. ISSN: 1618-727X. DOI: 10.1007/s10278-016-9914-9. (Visited on 05/02/2024).

[70] Paras Lakhani and Baskaran Sundaram. "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks". In: *Radiology* (Apr. 2017). DOI: 10.1148/radiol.2017162326. (Visited on 05/02/2024).

[71] Yu-Xing Tang et al. "Automated Abnormality Classification of Chest Radiographs Using Deep Convolutional Neural Networks". In: *npj Digital Medicine* 3.1 (May 2020), pp. 1–8. ISSN: 2398-6352. DOI: 10.1038/s41746-020-0273-z. (Visited on 05/02/2024).

[72] Anant Bhatt, Amit Ganatra, and Ketan Kotecha. "COVID-19 Pulmonary Consolidations Detection in Chest X-ray Using Progressive Resizing and Transfer Learning Techniques". In: *Heliyon* 7.6 (June 2021), e07211. ISSN: 24058440. DOI: 10.1016/j.heliyon.2021.e07211. (Visited on 05/02/2024).

[73] Joana Rocha, Sofia Cardoso Pereira, João Pedrosa, Aurélio Campilho, and Ana Maria Mendonça. "STERN: Attention-driven Spatial Transformer Network for Abnormality Detection in Chest X-ray Images". In: *Artificial Intelligence in Medicine* 147 (Jan. 2024), p. 102737. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2023.102737. (Visited on 04/22/2024).

[74] Umar Marikkar, Sara Atito, Muhammad Awais, and Adam Mahdi. "LT-ViT: A Vision Transformer for Multi-Label Chest X-ray Classification". In: *2023 IEEE International Conference on Image Processing (ICIP)*. Oct. 2023, pp. 2565–2569. DOI: 10.1109/ICIP49359.2023.10222175. arXiv: 2311.07263 [cs]. (Visited on 04/22/2024).

[75] Oishy Saha, Jarin Tasnim, Md. Tanvir Raihan, Tanvir Mahmud, Istak Ahmmed, and Shaikh Anowarul Fattah. "A Multi-Model Based Ensembling Approach to Detect COVID-19 from Chest X-Ray Images". In: *2020 IEEE REGION 10 CONFERENCE (TENCON)*. Nov. 2020, pp. 591–595. DOI: 10.1109/TENCON50793.2020.9293802. (Visited on 05/02/2024).

[76] Sagar Deep Deb, Rajib Kumar Jha, Kamlesh Jha, and Prem S Tripathi. "A Multi Model Ensemble Based Deep Convolution Neural Network Structure for Detection of COVID19". In: *Biomedical Signal Processing and Control* 71 (Jan. 2022), p. 103126. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2021.103126. (Visited on 05/02/2024).

[77] *Papers with Code - CheXpert Benchmark (Multi-Label Classification)*. (Visited on 05/04/2024).

[78] Hieu H. Pham, Tung T. Le, Dat Q. Tran, Dat T. Ngo, and Ha Q. Nguyen. "Interpreting Chest X-rays via CNNs That Exploit Hierarchical Disease Dependencies and Uncertainty Labels". In: *Neurocomputing* 437 (May 2021), pp. 186–194. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.03.127. (Visited on 11/27/2024).

[79] Chak Fong Chong, Xu Yang, Tenglong Wang, Wei Ke, and Yapeng Wang. "Category-Wise Fine-Tuning for Image Multi-label Classification with Partial Labels". In: *Neural Information Processing*. Ed. by Biao Luo, Long Cheng, Zheng-Guang Wu, Hongyi Li, and Chaojie Li. Singapore: Springer Nature, 2024, pp. 332–345. ISBN: 978-981-9981-45-8. DOI: 10.1007/978-981-99-8145-8_26.

[80] Haoqin Ji et al. "Point Beyond Class: A Benchmark for Weakly Semi-supervised Abnormality Localization in Chest X-Rays". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li. Cham: Springer Nature Switzerland, 2022, pp. 249–260. ISBN: 978-3-031-16437-8. DOI: 10.1007/978-3-031-16437-8_24.

[81] Yan Han, Chongyan Chen, Ahmed Tewfik, Benjamin Glicksberg, Ying Ding, Yifan Peng, and Zhangyang Wang. "Knowledge-Augmented Contrastive Learning for Abnormality Classification and Localization in Chest X-Rays With Radiomics Using a Feedback Loop". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 2465–2474. (Visited on 05/03/2024).

[82] Xi Ouyang, Srikrishna Karanam, Ziyan Wu, Terrence Chen, Jiayu Huo, Xiang Sean Zhou, Qian Wang, and Jie-Zhi Cheng. "Learning Hierarchical Attention for Weakly-Supervised Chest X-Ray Abnormality Localization and Diagnosis". In: *IEEE Transactions on Medical Imaging* 40.10 (Oct. 2021), pp. 2698–2710. ISSN: 1558-254X. DOI: 10.1109/TMI.2020.3042773. (Visited on 05/03/2024).

[83] Yongwon Cho, Young-Gon Kim, Sang Min Lee, Joon Beom Seo, and Namkug Kim. "Reproducibility of Abnormality Detection on Chest Radiographs Using Convolutional Neural Network in Paired Radiographs Obtained within a Short-Term Interval". In: *Scientific Reports* 10.1 (Oct. 2020), p. 17417. ISSN: 2045-2322. DOI: 10.1038/s41598-020-74626-4. (Visited on 05/03/2024).

[84] Yongwon Cho, Young-Gon Kim, Sang Min Lee, Joon Beom Seo, and Namkug Kim. "Reproducibility of Abnormality Detection on Chest Radiographs Using Convolutional Neural Network in Paired Radiographs Obtained within a Short-Term Interval". In: *Scientific Reports* 10.1 (Oct. 2020), p. 17417. ISSN: 2045-2322. DOI: 10.1038/s41598-020-74626-4. (Visited on 05/03/2024).

[85] Nkechinyere N. Agu, Joy T. Wu, Hanqing Chao, Ismini Lourentzou, Arjun Sharma, Mehdi Moradi, Pingkun Yan, and James Hendler. "AnaXNet: Anatomy Aware Multi-label Finding Classification in Chest X-Ray". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert. Cham: Springer International Publishing, 2021, pp. 804–813. ISBN: 978-3-030-87240-3. DOI: 10.1007/978-3-030-87240-3_77.

[86] Minki Kim and Byoung-Dai Lee. "Automatic Lung Segmentation on Chest X-rays Using Self-Attention Deep Neural Network". In: *Sensors* 21.2 (Jan. 2021), p. 369. ISSN: 1424-8220. DOI: 10.3390/s21020369. (Visited on 05/03/2024).

[87] Qiwen Que et al. "CardioXNet: Automated Detection for Cardiomegaly Based on Deep Learning". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. July 2018, pp. 612–615. DOI: 10.1109/EMBC.2018.8512374. (Visited on 05/03/2024).

[88] Mohammad Eslami, Solale Tabarestani, Shadi Albarqouni, Ehsan Adeli, Nassir Navab, and Malek Adjouadi. "Image-to-Images Translation for Multi-Task Organ Segmentation and Bone Suppression in Chest X-Ray Radiography". In: *IEEE Transactions on Medical Imaging* 39.7 (July 2020), pp. 2553–2565. ISSN: 1558-254X. DOI: 10.1109/TMI.2020.2974159. (Visited on 05/03/2024).

[89] Yu Gu et al. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *ACM Transactions on Computing for Healthcare* 3.1 (Oct. 2021), 2:1–2:23. DOI: 10.1145/3458754. (Visited on 05/05/2024).

[90] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. "BioGPT: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining". In: *Briefings in Bioinformatics* 23.6 (Nov. 2022), bbac409. ISSN: 1477-4054. DOI: 10.1093/bib/bbac409. (Visited on 05/05/2024).

[91] Karan Singhal et al. "Large Language Models Encode Clinical Knowledge". In: *Nature* 620.7972 (Aug. 2023), pp. 172–180. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06291-2. (Visited on 11/27/2024).

[92] Karan Singhal et al. *Towards Expert-Level Medical Question Answering with Large Language Models*. May 2023. DOI: 10.48550/arXiv.2305.09617. arXiv: 2305.09617 [cs]. (Visited on 05/05/2024).

[93] Michael Moor et al. "Med-Flamingo: A Multimodal Medical Few-shot Learner". In: *Proceedings of the 3rd Machine Learning for Health Symposium*. PMLR, Dec. 2023, pp. 353–367. (Visited on 05/06/2024).

[94] Chunyuan Li et al. “LLaVA-med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., May 2024, pp. 28541–28564. (Visited on 11/27/2024).

[95] Tao Tu et al. *Towards Generalist Biomedical AI*. July 2023. arXiv: 2307.14334 [cs]. (Visited on 05/06/2024).

[96] Anas Awadalla et al. *OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models*. Aug. 2023. DOI: 10.48550/arXiv.2308.01390. arXiv: 2308.01390 [cs]. (Visited on 05/06/2024).

[97] Danny Driess et al. “PaLM-E: An Embodied Multimodal Language Model”. In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. ICML’23. Honolulu, Hawaii, USA: JMLR.org, July 2023, pp. 8469–8488. (Visited on 11/27/2024).

[98] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. “MedCLIP: Contrastive Learning from Unpaired Medical Images and Text”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3876–3887. DOI: 10.18653/v1/2022.emnlp-main.256. (Visited on 11/27/2024).

[99] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. “GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3942–3951. (Visited on 05/05/2024).

[100] Benedikt Boecking et al. “Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing”. In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner. Cham: Springer Nature Switzerland, 2022, pp. 1–21. ISBN: 978-3-031-20059-5. DOI: 10.1007/978-3-031-20059-5_1.

[101] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun Kyoung Hong, Woonhyunk Baek, and Byungseok Roh. “CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training”. In: vol. 14221. 2023, pp. 101–111. DOI: 10.1007/978-3-031-43895-0_10. arXiv: 2310.13292 [cs]. (Visited on 04/22/2024).

[102] Shawn Xu et al. *ELIXR: Towards a General Purpose X-ray Artificial Intelligence System through Alignment of Large Language Models and Radiology Vision Encoders*. Sept. 2023. DOI: 10.48550/arXiv.2308.01317. arXiv: 2308.01317 [cs, eess]. (Visited on 05/06/2024).

[103] Shaury Srivastav et al. “MAIRA at RRG24: A Specialised Large Multimodal Model for Radiology Report Generation”. In: *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*. Ed. by Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 597–602. DOI: 10.18653/v1/2024.bionlp-1.50. (Visited on 11/27/2024).

[104] Baoyu Jing, Pengtao Xie, and Eric 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)*. 2018, pp. 2577–2586. DOI: 10.18653/v1/P18-1240. arXiv: 1711.08195 [cs]. (Visited on 05/06/2024).

[105] Hoang Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. “Automated Generation of Accurate & Fluent Medical X-ray Reports”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3552–3569. DOI: 10.18653/v1/2021.emnlp-main.288. (Visited on 04/22/2024).

[106] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru 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. (Visited on 05/07/2024).

[107] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong lu, and Ronald Summers. “TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays”. In: *IEEE CVPR 2018*. Jan. 2018. DOI: 10.1109/CVPR.2018.00943.

[108] Preethi Srinivasan, Daksh Thapar, Arnav Bhavsar, and Aditya Nigam. “Hierarchical X-Ray Report Generation via Pathology Tags and Multi Head Attention”. In: *Proceedings of the Asian Conference on Computer Vision*. 2020. (Visited on 05/07/2024).

[109] Baoyu Jing, Pengtao Xie, and Eric 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)*. 2018, pp. 2577–2586. DOI: 10.18653/v1/P18-1240. arXiv: 1711.08195 [cs]. (Visited on 05/07/2024).

[110] Preethi Srinivasan, Daksh Thapar, Arnav Bhavsar, and Aditya Nigam. “Hierarchical X-Ray Report Generation via Pathology Tags and Multi Head Attention”. In: *Proceedings of the Asian Conference on Computer Vision*. 2020. (Visited on 05/07/2024).

[111] Justin Lovelace and Bobak Mortazavi. "Learning to Generate Clinically Coherent Chest X-Ray Reports". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 1235–1243. DOI: 10.18653/v1/2020.findings-emnlp.110. (Visited on 05/07/2024).

[112] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li 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. (Visited on 05/07/2024).

[113] Baoyu Jing, Pengtao Xie, and Eric 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)*. 2018, pp. 2577–2586. DOI: 10.18653/v1/P18-1240. arXiv: 1711.08195 [cs]. (Visited on 05/07/2024).

[114] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. "Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 6666–6673. ISSN: 2374-3468. DOI: 10.1609/aaai.v33i01.33016666. (Visited on 05/07/2024).

[115] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. "When Radiology Report Generation Meets Knowledge Graph". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.07 (Apr. 2020), pp. 12910–12917. ISSN: 2374-3468. DOI: 10.1609/aaai.v34i07.6989. (Visited on 05/07/2024).

[116] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian 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. (Visited on 05/07/2024).

[117] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. "Interactive and Explainable Region-guided Radiology Report Generation". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 7433–7442. DOI: 10.1109/CVPR52729.2023.00718. arXiv: 2304.08295 [cs]. (Visited on 04/22/2024).

[118] Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeffrey Dalton, and Alison Q. O'Neil. "Finding-Aware Anatomical Tokens for Chest X-Ray Automated Reporting". In: *Machine Learning in Medical Imaging: 14th International Workshop, MLMI 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, Oct. 2023, pp. 413–423. ISBN: 978-3-031-45672-5. DOI: 10.1007/978-3-031-45673-2_41. (Visited on 11/27/2024).

[119] Tania Lombrozo. "The Structure and Function of Explanations". In: *Trends in Cognitive Sciences* 10.10 (Oct. 2006), pp. 464–470. ISSN: 1364-6613, 1879-307X. DOI: 10.1016/j.tics.2006.08.004. (Visited on 05/07/2024).

[120] Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, Jan. 2018, pp. 77–91. (Visited on 05/07/2024).

[121] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. (Visited on 04/28/2024).

[122] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. (Visited on 05/08/2024).

[123] Jianbo Chen and Michael Jordan. “LS-Tree: Model Interpretation When the Data Are Linguistic”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04 (Apr. 2020), pp. 3454–3461. ISSN: 2374-3468. DOI: 10.1609/aaai.v34i04.5749. (Visited on 05/08/2024).

[124] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation”. In: *PLOS ONE* 10.7 (July 2015), e0130140. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0130140. (Visited on 05/08/2024).

[125] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74. (Visited on 11/27/2024).

[126] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. (Visited on 05/08/2024).

[127] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: (Dec. 2013). (Visited on 11/27/2024).

[128] Tao Lei, Regina Barzilay, and Tommi Jaakkola. “Rationalizing Neural Predictions”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by Jian Su, Kevin Duh, and Xavier Carreras. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 107–117. DOI: 10.18653/v1/D16-1011. (Visited on 11/27/2024).

[129] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. “INVASE: Instance-wise Variable Selection Using Neural Networks”. In: *International Conference on Learning Representations*. Sept. 2018. (Visited on 04/28/2024).

[130] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR". In: *Harvard Journal of Law & Technology (Harvard JOLT)* 31 (2017/2018), p. 841.

[131] Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. "Measuring Association Between Labels and Free-Text Rationales". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10266–10284. DOI: 10.18653/v1/2021.emnlp-main.804. (Visited on 11/27/2024).

[132] Nuno M. Guerreiro and André F. T. Martins. "SPECTRA: Sparse Structured Text Rationalization". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6534–6550. DOI: 10.18653/v1/2021.emnlp-main.525. (Visited on 11/27/2024).

[133] Lei Sha, Oana-Maria Camburu, and Thomas Lukasiewicz. "Learning from the Best: Rationalizing Predictions by Adversarial Information Calibration". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.15 (May 2021), pp. 13771–13779. ISSN: 2374-3468. DOI: 10.1609/aaai.v35i15.17623. (Visited on 05/09/2024).

[134] Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. "Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4094–4103. DOI: 10.18653/v1/D19-1420. (Visited on 11/27/2024).

[135] Jasmijn Bastings, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2963–2977. DOI: 10.18653/v1/P19-1284. (Visited on 11/27/2024).

[136] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. "E-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 1224–1234. ISBN: 978-1-66542-812-5. DOI: 10.1109/ICCV48922.2021.00128. (Visited on 11/27/2024).

[137] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. "E-SNLI: Natural Language Inference with Natural Language Explanations". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. (Visited on 05/09/2024).

[138] Jason Wei et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., Apr. 2024, pp. 24824–24837. ISBN: 978-1-71387-108-8. (Visited on 11/27/2024).

[139] Jiaxin Ge, Sanjay Subramanian, Trevor Darrell, and Boyi Li. "From Wrong To Right: A Recursive Approach Towards Vision-Language Explanation". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1173–1185. DOI: 10.18653/v1/2023.emnlp-main.75. (Visited on 11/27/2024).

[140] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8779–8788. (Visited on 04/28/2024).

[141] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8779–8788. (Visited on 05/09/2024).

[142] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. "Textual Explanations for Self-Driving Vehicles". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 563–578. (Visited on 05/09/2024).

[143] Alec Radford, Jeff Wu, R. Child, D. Luan, Dario Amodei, and I. Sutskever. "Language Models Are Unsupervised Multitask Learners". In: 2019. (Visited on 05/09/2024).

[144] Peter Anderson et al. "Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3674–3683. (Visited on 05/09/2024).

[145] Pierre Dognin et al. "Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge". In: *Journal of Artificial Intelligence Research* 73 (Jan. 2022), pp. 437–459. ISSN: 1076-9757. DOI: 10.1613/jair.1.13113. (Visited on 05/09/2024).

[146] Ishan Misra, Ross Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, and Laurens van der Maaten. “Learning by Asking Questions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 11–20. (Visited on 05/09/2024).

[147] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. “Grounding Visual Explanations”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 264–279. (Visited on 05/09/2024).

[148] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. “From Recognition to Cognition: Visual Commonsense Reasoning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 6720–6731. (Visited on 05/09/2024).

[149] Bodhisattwa Prasad Majumder, Oana Camburu, Thomas Lukasiewicz, and Julian Mcauley. “Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations”. In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR, June 2022, pp. 14786–14801. (Visited on 11/27/2024).

[150] Björn Plüster, Jakob Ambsdorf, Lukas Braach, Jae Hee Lee, and Stefan Wermter. *Harnessing the Power of Multi-Task Pretraining for Ground-Truth Level Natural Language Explanations*. Mar. 2023. DOI: 10.48550/arXiv.2212.04231. arXiv: 2212.04231 [cs]. (Visited on 05/09/2024).

[151] Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. “NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 8322–8332. (Visited on 05/09/2024).

[152] Ana Marasović, Chandra Bhagavatula, Jae sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. “Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 2810–2829. DOI: 10.18653/v1/2020.findings-emnlp.253. (Visited on 11/27/2024).

[153] Shiyang Li et al. *Explanations from Large Language Models Make Small Reasoners Better*. Oct. 2022. DOI: 10.48550/arXiv.2210.06726. arXiv: 2210.06726 [cs]. (Visited on 05/09/2024).

[154] Tom Brown et al. “Language Models Are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. (Visited on 05/09/2024).

[155] Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. “Using Natural Language Explanations to Improve Robustness of In-context Learning”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku,

Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 13477–13499. DOI: 10.18653/v1/2024.acl-long.728. (Visited on 11/27/2024).

[156] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. "Self-Consistency Improves Chain of Thought Reasoning in Language Models". In: *The Eleventh International Conference on Learning Representations*. Sept. 2022. (Visited on 11/27/2024).

[157] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. (Visited on 05/09/2024).

[158] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. Mar. 2017. DOI: 10.48550/arXiv.1702.08608. arXiv: 1702.08608 [cs, stat]. (Visited on 05/09/2024).

[159] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. "A Benchmark for Interpretability Methods in Deep Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. (Visited on 05/09/2024).

[160] Alon Jacovi and Yoav Goldberg. "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. (Visited on 11/27/2024).

[161] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. "Explain Yourself! Leveraging Language Models for Commonsense Reasoning". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4932–4942. DOI: 10.18653/v1/P19-1487. (Visited on 11/27/2024).

[162] Oana-Maria Camburu. *Explaining Deep Neural Networks*. Oct. 2021. arXiv: 2010.01496 [cs]. (Visited on 04/28/2024).

[163] Bas H. M. van der Velden, Hugo J. Kuijf, Kenneth G. A. Gilhuijs, and Max A. Viergever. "Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis". In: *Medical Image Analysis* 79 (July 2022), p. 102470. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102470. (Visited on 11/26/2024).

[164] Bas H. M. van der Velden, Hugo J. Kuijf, Kenneth G. A. Gilhuijs, and Max A. Viergever. "Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis". In: *Medical Image Analysis* 79 (July 2022), p. 102470. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102470. (Visited on 05/14/2024).

[165] Tanvir Mahmud, Md Awsafur Rahman, and Shaikh Anowarul Fattah. "CovXNet: A Multi-Dilation Convolutional Neural Network for Automatic COVID-19 and Other Pneumonia Detection from Chest X-ray Images with Transferable Multi-Receptive Feature Optimization". In: *Computers in Biology and Medicine* 122 (July 2020), p. 103869. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2020.103869. (Visited on 05/14/2024).

[166] Luca Brunese, Francesco Mercaldo, Alfonso Reginelli, and Antonella Santone. "Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays". In: *Computer Methods and Programs in Biomedicine* 196 (Nov. 2020), p. 105608. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2020.105608. (Visited on 05/14/2024).

[167] Jared A. Dunnmon, Darvin Yi, Curtis P. Langlotz, Christopher Ré, Daniel L. Rubin, and Matthew P. Lungren. "Assessment of Convolutional Neural Networks for Automated Classification of Chest Radiographs". In: *Radiology* 290.2 (Feb. 2019), pp. 537–544. ISSN: 0033-8419. DOI: 10.1148/radiol.2018181422. (Visited on 05/14/2024).

[168] Pranav Rajpurkar et al. "Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists". In: *PLOS Medicine* 15.11 (Nov. 2018), e1002686. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002686. (Visited on 05/14/2024).

[169] Shiva prasad Koyyada and Thipendra P. Singh. "An Explainable Artificial Intelligence Model for Identifying Local Indicators and Detecting Lung Disease from Chest X-ray Images". In: *Healthcare Analytics* 4 (Dec. 2023), p. 100206. ISSN: 2772-4425. DOI: 10.1016/j.health.2023.100206. (Visited on 05/13/2024).

[170] Sivaramakrishnan Rajaraman, Sema Candemir, George Thoma, and Sameer Antani. "Visualizing and Explaining Deep Learning Predictions for Pneumonia Detection in Pediatric Chest Radiographs". In: *Medical Imaging 2019: Computer-Aided Diagnosis*. Vol. 10950. SPIE, Mar. 2019, pp. 200–211. DOI: 10.1117/12.2512752. (Visited on 05/14/2024).

[171] Lin Zou et al. "Ensemble Image Explainable AI (XAI) Algorithm for Severe Community-Acquired Pneumonia and COVID-19 Respiratory Infections". In: *IEEE Transactions on Artificial Intelligence* 4.2 (Apr. 2023), pp. 242–254. ISSN: 2691-4581. DOI: 10.1109/TAI.2022.3153754. (Visited on 05/14/2024).

[172] Ivanoe De Falco, Giuseppe De Pietro, and Giovanna Sannino. "Classification of Covid-19 Chest X-ray Images by Means of an Interpretable Evolutionary Rule-Based Approach". In: *Neural Computing and Applications* 35.22 (Aug. 2023),

pp. 16061–16071. ISSN: 1433-3058. DOI: 10.1007/s00521-021-06806-w. (Visited on 04/22/2024).

[173] Xin Zhang et al. "CXR-Net: An Encoder-Decoder-Encoder Multitask Deep Neural Network for Explainable and Accurate Diagnosis of COVID-19 Pneumonia with Chest X-ray Images". In: *IEEE Journal of Biomedical and Health Informatics* 27.2 (Feb. 2023), pp. 980–991. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2022.3220813. arXiv: 2110.10813 [cs, eess]. (Visited on 04/22/2024).

[174] Hao Yang et al. "Multimodal Self-Supervised Learning for Lesion Localization". In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. May 2024, pp. 1–5. DOI: 10.1109/ISBI56570.2024.10635268. (Visited on 11/27/2024).

[175] Benjamin Hou, Georgios Kaissis, Ronald M. Summers, and Bernhard Kainz. "RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert. Cham: Springer International Publishing, 2021, pp. 293–303. ISBN: 978-3-030-87234-2. DOI: 10.1007/978-3-030-87234-2_28.

[176] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. "Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 1500–1519. DOI: 10.18653/v1/2020.emnlp-main.117. (Visited on 11/27/2024).

[177] Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. "NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 8312–8322. ISBN: 978-1-66546-946-3. DOI: 10.1109/CVPR52688.2022.00814. (Visited on 04/22/2024).

[178] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. *EVA-CLIP: Improved Training Techniques for CLIP at Scale*. Mar. 2023. DOI: 10.48550/arXiv.2303.15389. arXiv: 2303.15389. (Visited on 11/09/2024).

[179] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. *CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT*. Oct. 2020. DOI: 10.48550/arXiv.2004.09167. arXiv: 2004.09167 [cs]. (Visited on 05/14/2024).

[180] Joseph Paul Cohen et al. "TorchXRayVision: A Library of Chest X-ray Datasets and Models". In: *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*. PMLR, Dec. 2022, pp. 231–249. (Visited on 11/27/2024).

[181] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. "LAVIS: A One-stop Library for Language-Vision Intelligence". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Ed. by Danushka Bollegala, Ruihong Huang, and Alan Ritter. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 31–41. DOI: 10.18653/v1/2023.acl-demo.3. (Visited on 11/27/2024).

[182] Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. "PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?" In: *Findings of the Association for Computational Linguistics: EACL 2023*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1181–1193. DOI: 10.18653/v1/2023.findings-eacl.88. (Visited on 09/29/2024).

[183] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. "Radiology Objects in COntext (ROCO): A Multimodal Image Dataset". In: *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Ed. by Danail Stoyanov et al. Vol. 11043. Cham: Springer International Publishing, 2018, pp. 180–189. ISBN: 978-3-030-01363-9 978-3-030-01364-6. DOI: 10.1007/978-3-030-01364-6_20. (Visited on 11/12/2024).

[184] Eric Jang, Shixiang Gu, and Ben Poole. "Categorical Reparameterization with Gumbel-Softmax". In: *International Conference on Learning Representations*. Feb. 2017. (Visited on 11/27/2024).

[185] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. *Quantifying the Carbon Emissions of Machine Learning*. Nov. 2019. DOI: 10.48550/arXiv.1910.09700. arXiv: 1910.09700. (Visited on 11/15/2024).

[186] *CO2 Emissions per kWh in Austria - Nowtricity*. (Visited on 11/15/2024).

[187] *ChatGPT*. (Visited on 11/27/2024).

[188] *Claude*. (Visited on 11/27/2024).

[189] *DeepL Traduction – DeepL Translate : le meilleur traducteur au monde*. (Visited on 11/27/2024).