DEGREE PROJECT IN TECHNOLOGY,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN, JULY 18, 2022

The Ability of Visual and Language Explainable Models to Resemble Domain Expertise

Using the Local Surrogate Explainability Technique

Petrus Oskarsson

KTH ROYAL INSTITUTE OF TECHNOLOGY
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Authors

Petrus Oskarsson | petruso@kth.se
Information and Communication Technology
KTH Royal Institute of Technology

Place for Project

Stockholm, Sweden
Amazon Web Services

Examiner

Henrik Boström | bostromh@kth.se
Stokholm, Sweden
KTH Royal Institute of Technology

Supervisor

Amir H. Payberah |payberah@kth.se

Stokholm, Sweden

KTH Royal Institute of Technology

Abstract

Recent advances in vision and language models have taken inspiration from the language transformer network BERT, with promising results on visual and language tasks. In parallel, studies show that learning from the joint vision and language embeddings is effective in learning clinical tasks, especially learning from radio-graph and radiology reports. However, there is a constant need for model transparency in the healthcare field, and state-of-the-art vision and language models struggle to explain made predictions. One prominent technique to explain predictions of deep learning models is using a local surrogate model, which separates the machine learning explanations from the machine learning model. In addition, the inclusion of domain expertise has been shown to be an essential success factor for machine learning models to make an entrance into the medical field. This thesis work explores the feasibility of resembling domain expertise when using the local surrogate explainability technique in combination with an underlying vision and language model to generate multi-modal visual and language explanations. A case study has been carried out to explain vision and language models trained to predict thoracic findings from radio-graphs and radiology reports. More specifically, we trained an UNITER and a VisualBERT network on the machine learning task and then trained explainable models to generate explanations for model predictions. Next, we collected explanations from domain experts and finally compared those with explanations from the explainable model. The results show low similarity compared to domain expertise. Nevertheless, the results also suggest that the particular case study task of explaining thoracic findings is challenging as annotations from domain experts indicate that there is ambiguity on what is the ground truth in terms of explanations. Furthermore, despite the low similarity scores, the explainable models seem to some extent have captured signals in explaining predictions, and generated explanations can serve as helpful feedback for data scientists and machine learning engineers in the field.

Keywords

explainable machine learning, multi-modality, vision and language models, healthcare

Sammanfattning

De senaste framstegen inom syn- och språkmodeller har hämtat inspiration från språktransformatornätverket BERT och utvisar lovande resultat på visuella och språkliga uppgifter. Parallellt visar studier att lärande från den gemensamma inbäddningen av syn och språk är effektivt för att lära sig kliniska uppgifter, särskilt lärande från röntgenbilder och röntgenrapporter. Det finns dock ett ständigt behov av modelltransparens inom vård och medicin, och *state-of-the-art* syn- och språkmodeller har svårt att förklara sina prediktioner. En framträdande teknik för att förklara prediktioner inom djupinlärning är att använda en lokal surrogatmodell, som särskiljer maskininlärningsförklaringar från maskininlärningsmodellen. Dessutom har inkluderingen av domänexpertis visat sig vara en viktig framgångsfaktor för maskininlärningsmodeller inom medicinska fältet. Detta examensarbete undersöker möjligheten att efterlikna domänexpertis vid användet av den lokala surrogattekniken i kombination med en underliggande syn- och språkmodeller för att generera multimodala syn- och språkförklaringar. En fallstudie har genomförts för att förklara syn- och språkmodeller som tränats för att prediktera thoraxfynd från röntgenbilder och röntgenrapporter. Mer specifikt tränade vi ett UNITER- och ett VisualBERT-nätverk på maskininlärnings-uppgiften och tränade sedan förklarande modeller för att generera förklaringar till modell-prediktioner. Därefter samlade vi in motsvarande förklaringar från domänexperter och jämförde dem med förklaringar från förklaringsmodellen. Resultaten visar låg likhet jämfört med domänexpertis. Däremot tyder resultaten också på att för den specifika fallstudie-uppgiften av att förklara thorax-fynd är utmanande eftersom annoteringar från domänexperter indikerar på tvetydighet gällande vad som är *sanna* förklaringar. Vidare, även om resultaten visar låga likhetsvärden, så verkar förklaringarnsmodellerna ha upptagit en viss signal till att förklara prediktioner, och genererade förklaringar kan fungera som användbar feedback för data scientists och maskininlärningsingenjörer i fältet.

Nyckelord

förklarande maskininlärning, multimodalitet, syn- och språkmodeller, sjukvård

Acknowledgements

From KTH Royal Institute of Technology, I would like to thank my supervisor Dr. Amir Hossein Payberah, for consistent guidance throughout the work and for introducing me to the exciting world of multi-modal explainable predictions. In addition, I want to thank my examiner Dr. Henrik Boström for lifting my work with valuable feedback during the process. Finally, I would like to thank Dr. Karl Eriklson, who made his Ph.D. studies in multi-modal explainable predictions, for sharing his insights in the research field.

From Amazon Web Services (AWS), I would like to thank my manager Vinita Chay for giving me the opportunity to do my research internship in her organization and for helping me relate the research to the industry. I also want to thank my mentor Dr. Manusnan Harinsuit for giving me invaluable coaching throughout the work. Moreover, I want to thank Anna Zapaishchykova for investing her time in going deep with me into the topic of visual and language explainable predictions. Similarly, I want to thank Ujjwal Ratan for investing his time into my work and sharing his senior guidance. Lastly, I want to thank Gustavo Beck for endless discussions and for sharing his strong machine learning expertise.

Finally, I want to send special gratitude to Erica Fred (Karolinska Hospital), Madeleine Broberg (Karolinska Institute), and Amilia Werner (University of Linköping) for their invaluable work in identifying explainable features from the radio-graphs and radiology reports. I want to thank them for sharing their time and knowledge because this thesis work would not have been possible without them.

Acronyms

ML	Machine Learning
AWS	Amazon Web Services
GDPR	General Data Protection Regulations
MIMIC-CXR	Medical Information Mart for Intensive Care - Chest X-ray
AUC	Area Under the Curve
LIME	Local Interpretable Model-Agnositc Explanations
SHAP	SHapley Additive exPlanations
NLP	Natural Language Processing
R-CNN	Region-based Convolutional Neural Network
BERT	Bidirectional Encoder Representations from Transformers

Contents

Chapter 1

Introduction

This chapter introduces the thesis study by describing the context, motivation, and positioning of the work. In addition, the chapter gives an overview of the methodology, describes the delimitations, and outlines the content of the remaining chapters.

1.1 Background

Multi-modal Machine Learning (ML) refers to training on several input sources that have different representations and contain complementary information in relation to the ML task [43]. Moreover, studies show that ML models trained on multi-modal inputs outperform models that learn from one modality alone [43] [42] [17] [39]. For instance, AWS has performed a study on multi-modal ML models trained on health data [42] [41]. The study compares the performance of multi-modal versus unimodal models to predict the survival outcome of patients diagnosed with non-small cell lung cancer. The results shows that the multi-modal healthcare model outperforms the models trained on singular data modalities alone.

Applying multi-modal ML in the healthcare field is an active area of research and looks to have a promising impact on patient care [38]. In practice, medical practitioners consider multiple sources of information for patient diagnosis [20] and, intuitively, healthcare ML models should also benefit from learning from multiple modalities. One up-and-coming area of multi-modal ML in the medical field is to learn from the joint input space of vision and language. In fact, several studies in the healthcare domain have confirmed the potential of learning from the joint vision and language embedding for clinical tasks, and especially the combination of radio-graphs and radiology reports [7] [29] [30].

Besides that, multi-modal and language models have made some recent advancements in the ML research field. Traditionally, multi-modal vision and language architectures have learned the image representation by using convolutional neural networks and have learned the text embedding via recurrent neural networks [53]. After that, the architectures have performed multi-modal fusion to achieve a joint representation [53]. However, lately, multi-modal and language models have taken inspiration from the Bidirectional Encoder Representations from Transformers (BERT) architecture [27]. BERT is a transformer language model and adopts a pre-training transfer learning approach [12]. Researchers have been extending this architecture by adopting the pre-train transfer learning approach to a multi-modal setting [9] [52] [31] [26]. Recent vision and language models learn a joint image-text embedding by first performing pre-training on multiple large-scale vision and language data sets and then fine-tuning this on downstream tasks [9] [52] [31] [26]. Some of the most prominent architectures with this approach are *UNITER [9], LXMERT [52], VisualBERT [26]* that have shown state-of-the-art performance on several vision and language tasks [9] [52] [31] [26].

1.2 Problem

Even if recent vision and language models have a promising impact in the medical field, just like many other state-of-the-art deep learning models, they cannot manifest how and why a model has made a certain decision. As a result, such models appear as *black boxes* [2]. Due to the lack of transparency, such models commonly struggle to enter the medical field as the ability to reason and explain is critical to earning clinicians' trust [10] [37] [5]. In addition, compliance and regulations make the transparency of ML models utterly important in the field [15] [37]. For instance, the General Data Protection Regulations (GDPR) regulations state the right to receive *"meaningful information about the logic involved"* for automated decision-making [15].

However, even unimodal models suffer from a lack of transparency [51]. Furthermore, for a multi-modal setting, the complexity of understanding what information and factors a model has based a decision on increases as the diversity of the input data increases [23].

Explainable ML is an active area of research and refers to answering *what a model has learned* as well as *how and why a model made a prediction* [35]. The purpose is to increase model transparency and detect eventual biases that a model has learned [35]. One prominent technique to overcome the problem of deep learning models manifesting as black boxes is to separate the ML model and ML explanations [16]. *Local surrogate explainable models* generate explanations to model predictions by training an inherently interpretable model on the underlying model's outputs. The explanations are specific around an instance, i.e., a local prediction, and the technique allows ML practitioners to separate the ML explanations from the model architecture [16].

In addition, as the field of explainable ML has evolved, there is limited work on how to evaluate ML explanations [35] [6]. Yet, Doshi-Velex and Kim suggest that for applications that requires extensive human domain expertise, the evaluation of ML explanations should also involve domain expertise [13]. In addition, the inclusion of domain expertise has shown to be an important success factor for ML systems to earn trust in the medical field [40] [8].

While there is limited work on multi-modal visual and language explanations, previous work has been carried out on comparing explainability across different types of modality fusion for vision and language learning [3]. However, to the best of our knowledge, there is no previous work done on evaluating visual and language explanations in relation to domain expertise.

1.3 Purpose

Recent advances in vision and language learning look to have a promising impact in the medical field [7] [29] [30]. However, in an industry with a constant need for transparency, such models commonly struggle to explain predictions [10] [37] [5] [15] [37]. Nevertheless, one prominent technique to overcome the problem of black-box deep learning models is to use the local surrogate explainability technique to explain a made prediction [16]. In addition, the inclusion of domain expertise is both suggested to be used to evaluate ML explanations on advanced clinical tasks [35] [6], and have shown to be a critical success factor for ML system to enter into the medical field [40] [8]. Propelled by these factors, the research question addressed in this thesis work is:

Considering the combination of an underlying ML model and an explainability technique, can the local surrogate explainability technique be used to resemble domain expertise for explaining multi-modal vision and language predictions?

1.4 Goal

The goal of this work is to create an understanding of the feasibility of using the local surrogate explainability technique to generate multi-modal explanations in terms of both image and text explanations that resemble domain expertise. In doing so, the work aims to evaluate an ML system as a whole, meaning the combination of an underlying vision and language model and the local surrogate explainability technique and its ability to resemble domain expertise. To achieve this goal, the following steps are carried out:

1. Train vision and language models on the task of predicting thoracic findings, namely a VisualBERT and an UNITER model [26] [9].
2. Build explainable models using the local surrogate technique to explain predictions made by the vision and language models. More specifically, build two types of explainable models: one that is trained by perturbing both modalities simultaneously and one that combines the output of two models that have been trained separately on the vision contra the language modalities.
3. Collect annotations of the text and image modalities from domain experts that represent their explanations for identifying thoracic findings.
4. Compare and evaluate explanations from the explainable models with that of domain experts.

1.4.1 Benefits, Ethics, and Sustainability

ML practitioners may advantageously use the multi-modal vision-and-language explainable model described in our work in the medical field to explain predictions of vision and language models. Such explanations provide insights and serve as an interactive feedback loop for learning by highlighting decisive parts of the radiology and image modality. In addition, explanations can be used to provide explanations of made predictions to medical practitioners, ultimately helping to increase the transparency of vision and language models. Generally speaking, as ML models become more transparent, the likelihood of detecting eventual biases increases [35]. For example, the model could have possibly learned ethical and gender biases that would have served unfairly in patient care. Moreover, the reports and X-ray images used in the studies are anonymized [11]. Nevertheless, such patient data is potential sensitive information and should be treated accordingly. Therefore the study will not try to uncover or discuss any unethical or sensitive information or parts of the patient data. To be added is that the data set used in the study is publicly available [11]. More details of the data set is given in Section 3.2.2.

1.5 Methodology

For a detailed explanation of the choice of research methodology as well as the application of the research methodology, please advise Chapter 3. However, this is a qualitative and abductive study with an applied research methodology [18]. Moreover, the research strategy adopted is a case study approach[18] where we explain predictions made by vision and language models trained on the clinical task of predicting thoracic findings. First, we specifically train vision and language models on the multi-label classification task of predicting thoracic findings from a multi-modal input consisting of radio-graphs and radiology reports. After that, we build explainable models to generate a multi-modal explanation of predictions made by these models. Lastly, we compare these explanations with explanations collected from domain experts.

1.6 Delimitations

The thesis work does not aim to train state-of-the-art performing models on the issued ML task, but instead, the focus is on explainable models. If the underlying models happen to generalize poorly and have learned in-significant patterns, explanations should indicate such behavior. The explainable models used in this study will highlight what seem to be decisive parts of the text and image modality. However, if this does not resemble domain expertise, this could indicate that the model, for instance, contains biases, which can still be an insightful result.

Moreover, the size of the issued data set serves as a limitation in our study. Suppose that the models could train on a larger corpus of radiology reports and X-ray images. In that case, they could likely better learn the ML problem and likely learn patterns that better resemble domain expertise. However, given the project scope, we could only use the OpenI data set[11] which is limited in size. Please advise Section 3.2.2 for more details on the OpenI data set.

In addition, the study have trained two types of vision and language models, namely VisualBERT[26] and UNITER[9], and will perform experiments with explainable models on these two models only. However, we also considered training an LXMERT model [52], which has a dual-stream architecture, as opposed to VisualBERT and UNITER, which both have a single-stream architecture. It would have been interesting to compare the explanations of the LXMERT model to that of the other two, but we struggled to download the pre-trained weights of the LXMERT model, and given the project timeline, we did not include this model in the study.

Furthermore, for the choice of explainability techniques to explain predictions made by the vision and language models, we also considered using SHAP [32]. However, we ended up building on the Local Interpretable Model-Agnositc Explanations (LIME) technique to generate our multi-modal vision and language explanations. We could not include the SHAP technique in our experiments, given the project scope. Still, the reason for prioritizing to build on the LIME technique is that arguably the explanations generated by this technique are intuitive for an audience. The simple method of perturbing the input data and measuring its impact on the model outputs is often described as one of the major advantages of the LIME technique [35]. Please advise Section 2.2.3 for a further description of the LIME explainability technique.

Lastly, important to note is that in this work we do not distinguish between interpretable and explainable ML. While some previous work do distinguish between these two terms [49], there is arguably no consensus on the distinction between explainable and interpretable ML [34]. For simplicity, we chose to make no distinction between the two terms and use them interchangeably. Important to note, however, is that this works aims to explain prediction made by seemingly black-box models rather than trying to completely abandon the underlying model and directly train an interpretable ML model instead.

1.7 Outline

Chapter 2 gives an extended background by diving deep into related work and topics in vision and language learning and explainable ML.

Chapter 3 describes the choice of research method and the application of the research method with an in-depth description of the carried out experiments.

Chapter 4 gives a detailed description of the visual and language models.

Chapter 5 showcases and discusses the results.

Chapter 6 concludes the work and opens up for future work.

Chapter 2

Extended Background

This chapter describes the previous work that this thesis builds on, and provides the reader with the necessary background knowledge for the remaining parts of the report. Section 2.1 provides background on vision and language learning, while Section 2.2 helps the reader to understand the necessary components of explainable ML.

2.1 Vision and Language Learning

2.1.1 Introduction to Vision and Language Models

With inspiration from the masked language modeling and next sentence prediction that occur in the BERT transformer network [12], vision and language transformer networks such as VisualBERT, UNITER, and LXMERT have been developed [31] [9] [52]. These models all take on the masked language modeling approach to learn a joint image-text embedding and perform large-scale pre-training on vision and language data sets. The pre-trained weights can then fine-tune on specific vision and language tasks and such vision and language models have domonstrated state-of-the-art performance [31] [9] [52]. The transformer module of these vision and language models inputs pre-processed representations of the text and image modality that involves processing the text through a language encoder and extracting the visual feature representation with a visual encoder network [9] [26].

VisualBERT Network

VisualBERT is a pre-trained model for joint vision and language representations [26]. The model's architecture is an extension to the prominent Natural Language Processing (NLP) model BERT [12]. Just like BERT, the model consists of a stack of attention layers. However, VisualBERT also integrates object detection models such as Fast Region-based Convolutional Neural Network (R-CNN) in addition to the transformer layers [26]. The object detection model extracts visual features, and together with the text input, they serve as input to the VisualBERT model and are jointly processed throughout the transformers. As in the BERT network, the model uses self-attention mechanisms to align visual features with elements of the text [26]. The architecture allows a rich interaction between words and visual features such that the model can capture the intrinsic association between text and images. All in all, the model learns to capture a joint image and text representation tasks [26].

Moreover, the training approach that the VisualBERT model adopts is twofold: 1) First, parts of the text input are masked and the model learns to predict these by considering the remaining text and visual inputs. 2) Second, the model learns to match text and image inputs [26]. By applying this training procedure, the model learns transferable text and visual representations, which beneficially may be used for vision and language downstream tasks [26].

UNITER Network

The UNITER model is another large-scale pre-trained model for joint vision and language embedding. It also adopts a transformer architecture and makes use of the self-attention mechanism. With inspiration from BERT [12], the UNITER model is pre-trained on four tasks: Masked Language Modelling, Masked Region Modelling, Image-Text Matching, and Word Region Alignment [9]. With a visual embedding in the form of visual feature vectors and boxes and language encoding as text tokens, the transformer architecture learns a joint image and text embedding. Moreover, the UNITER model applies a masked language modeling conditioned on the entire image and text input instead of randomly masking joint image and text pieces. Furthermore, the architecture uses Optimal Transport for the Word Region Alignment pre-training task. The Optimal Transport mechanism optimizes distribution matching by minimizing the cost of transporting the image regions to word regions (and vice versa) which aims for more fine-grained image-word alignments. The Optimal Transport is one of the key enablers for this network to produce a fine-grained joint embedding of the modalities [9].

LXMERT Network

Both VisualBERT and UNITER apply a single-stream architecture, where a single transformer is used to learn the joint image and text embeddings [26] [9]. However, LXMERT (Learning Cross-Modality Encoder Representations from Transformers) has a dual-stream architecture, where the architecture has two transformer streams that learn to encode each of the modalities separately and then use a cross-attention layer to achieve a joint the embedding [52]. Moreover, as for VisualBERT and UNITER, LXMERT also performs large-scale pre-training on data sets of image-and-text with inspiration from BERT, and the model may advantageously be used for several transfer-learning vision and language tasks.

2.1.2 Related Work on Vision and Language Learning

Li et al. have performed a study on using X-ray images and radiology reports to train vision and language models on the task of predicting thoracic findings in their work *”A comparison of pre-trained vision-and-language models for multi-modal representation learning across medical images and reports* [27]. The study uses four pre-trained vision and language models: LXMERT, VisualBERT, UNITER, and PixelBERT, and trains on a downstream task of classifying thoracic findings. Their models train and evaluate using the Medical Information Mart for Intensive Care - Chest X-ray (MIMIC-CXR) data set [22] and test generalization capabilities on the OpenI data set [11]. The authors show that the use of pre-trained vision and language models increases the performance on the particular ML task. In addition, the study confirms that models that learn from the joint embedding of X-ray images and text from radiology reports outperform models that only train on one of the modalities. Further, the result demonstrates that the VisualBERT model has the best generalization capability with the highest Area Under the Curve (AUC) score across 11 out of 13 thoracic findings. After that, UNITER and LXMERT perform second and third best in terms of average AUC across the classes.

2.2 Explainable Machine Learning

2.2.1 Introduction to Explainable Machine Learning

ML models, especially deep learning networks, often cannot manifest how and why a model made a prediction and therefore commonly appear as black-boxes [2]. Explainable ML helps to make models more transparent by answering *what a model has learned* as well as *how and why a model has made a prediction* [35]. Explainability techniques may be segmented across three major splits: *local or global, model-specific or model-agnostic, and intrinsic or post-hoc* [35]. Local explainability techniques explain individual predictions, while global techniques explain entire model behaviors. Moreover, model-specific techniques refer to when the explanations derived are specific to the model's architecture, such as the weights of a regression model. Model-agnostic methods, on the other hand, work across models independently of the model-architecture [35]. Moreover, an intrinsic technique refers to when the model itself is explainable, and as a consequence, the technique is also model-specific. Finally, a post-hoc method is an external method that is most often used after training to generate explanations [35]. Post-hoc methods are often also model-agnostic since the method is applied after training and are therefore often independent of the model in use.

Furthermore, explanations for unimodal vision or languge models can be either based on visualizations showing important parts of the image input or language-based displaying what pieces of the text are more decisive for the prediction [19]. Multimodal vision and language models need explanations from both modalities, namely images and text explanations [23].

2.2.2 Evaluation of Explainable Machine Learning

Driven by the lack of unity on how to evaluate explainable ML, Doshi-Velex and Kim [13] suggest three approaches on how to evaluate explainable ML: application-grounded, human-grounded, and functionally-grounded. Below follows a short description of each approach [13]:

1. *Application grounded evaluation* refers to evaluating by conducting experiments involving human domain expertise within a real applications. The idea is to evaluate the quality of explanations in relation to its context. The authors gives

the example that assume there is a special application in mind such as working with doctors to diagnose patients for a certain disease - the best way to evaluate explanations is to evaluate it with respect to the task: doctors performing diagnosis [13]. The application-grounded evaluation approach requires domain experts within the application task. Such evaluations can be costly and difficult to set up. In addition, when evaluating ML explanations against human expertise, one should also consider how well human produced explanations assist other humans in performing on the same task [13].

2. *Human grounded evaluation* is an approach that involves simpler human-subject experiments that still maintain the essence of the application task. The idea is to simplify the application task in such a way that it still captures the quality of an explanation. For instance, a human subject might be presented with pairs of explanations and asked to chose the best quality as opposed to explaining from scratch without any guidance [13]. Since this approach does not require highly trained domain experts, it is generally simpler and cheaper than application-grounded evaluation.

3. *Functionally grounded evaluation* refers to an approach where no human expertise is involved. Instead, a proxy is used to evaluate the quality of explanations. The main challenge of this approach is how to define a proxy that well assesses the explanation quality in relation to the application task [13]. However, once a good proxy is defined, it can be advantageously used to optimise an explainable model.

2.2.3 Surrogate Explainable Models

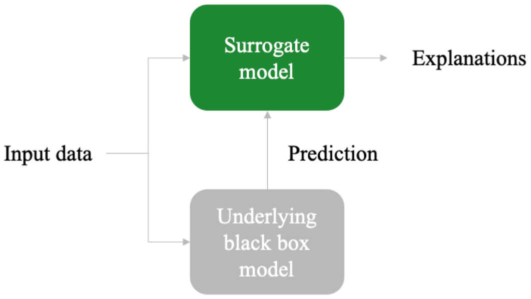


Figure 2.2.1: Illustrating the surrogate explainable technique.

The technique of using an explainable surrogate model refers to replacing a black-box model with a simpler interpretative model such as linear regression or decision tree models [35]. The surrogate model's training can be thought of as learning to predict the ouptuts of the original model. The goal is to mimic the behavior of the black-box model and make use of the explainable architecture to generate explanations.

Moreover, global surrogate models refer to explaining all model predictions, i.e., generating explanations for general model behavior. Local surrogate models, on the other hand, refer to explaining individual-made prediction by the black-box model [35]. The authors of the LIME explainability technique, suggest that it may be too complex to approximate global behaviour of an advanced underlying ML model, but for a local neighborhood, it is more reasonable to expect an interpretable surrogate model to capture local fidelity, or mimic the behaviour of the underlying model, for a single made prediction [47].

LIME

Local Interpretable Model-agnostic Explanations (LIME) is an explainability technique that provides explanations by learning an interpretable model locally around a single prediction [47]. The goal is to train an explainable model over an interpretable representation, typically a coalition vector, that mimics the behavior of the underlying *black-box* model. A coalition vector is a binary vector with 1 representing that a feature is present and 0 representing that a feature is absent. The technique is model-agnostic, so it can be applied to understand predictions of any black-box model. LIME explanations are generated under the following optimization problem [35]:

$$g(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{2.1}$$

Given Equation 2.1, an explanation model, $g(x)$, is produced to the input data x while minimizing the loss L and keeping the regularization term $\Omega(g)$ low. The loss L measures the closeness of the explanation to the original prediction, and $\Omega(g)$ represents model complexity. G is a set of interpretable models such as decision trees, linear models, or rule lists. For instance, for a decision tree choice of explainable models, the $\Omega(g)$ might be represented by the depth of the tree, while for a linear regression model, it could be the $L1$ or $L2$ penalty on the weights. π_x measures proximity by defining the size of the neighborhood around the instance x.

Furthermore, the steps to train an explainable model with LIME can briefly be summarized as:

1. Define an interpretable representation of features presence or absence, typically a binary vector with the size of the number of features.
2. Perform perturbations to generate samples around the instance x by randomly inactivating features from the interpretable representation of x.
3. Compute the distance of the each perturbation to x using the proximity function π_x for each perturbation where each sample is weighted with a proximity function measuring the closeness to the original instance x.
4. Compute the model's output for x and all perturbations.
5. Train a simple interpretable model under the constraint given in Equation 2.1.
6. Use the interpretable architecture of the simple model, $g(x)$, to generate explanations. For instance, explanations can be adhered from the coefficients of a linear model.

SHAP

SHapley Additive exPlanations (SHAP) is yet another commonly used local surrogate explainability technique. SHAP computes the impact of each feature for a particular prediction similar to LIME [32]. More specifically, the method adopts a game theoretic approach which computes shapley values from coalitional game theory [35]. This method aims to fairly assess how to distribute the credits for a prediction. In the context of game theory, the feature values represent players in a coalition. Further, SHAP generate explanations with an additive feature attribution method [35]:

$$g(x) = \phi_0 + \sum_{j=1}^{M} \phi_j x_j \tag{2.2}$$

In Equation 2.2, g represents the explainable model with a coalition vector, x. ϕ_j represents the feature attribution for a feature j as the feature's shapley value. In addition, the authors of SHAP give theoretical justifications that the class of additive feature importance measures has a single unique solution [32]. Furthermore, SHAP proposes different ways to approximate the Shapley values. One of them is KernalSHAP that is shown to be a special case of LIME for a certain choice of parameters [35].

2.2.4 Related Work on Local Surrogate Explainable Models

Alvi has carried out related work in the space of explaining multi-modal vision and language models in the work *Explainable Multimodal Fusion* [3] that serves as an important source of inspiration for this study. Alvi first compares single versus dual stream vision and language models on the visual entailment task. A visual entailment task consists of image-sentence pairs, and the task is to predict whether an image semantically entails the text in a sentence [55]. In addition, the study explores which architecture of single versus dual stream performs better in terms of explainability. For this, the author uses the local surrogate explainability technique to generate multi-modal vision and language explanations for the visual entailment task.

More specifically, the author makes use of LIME to train the local surrogate explainable models. Next, perform text and image perturbations and then train an inherently interpretative model to generate explanations. Further, the author makes the image perturbations by using the LIME feature of extracting superpixels from an image, referring to a group of pixels in the image used to represent different features in the image [3]. The author inactivates features by transforming superpixels into black pixels.

Furthermore, new visual features are extracted, and together with the language embedding, these are fed to the vision and language model for prediction. By doing so many times, a local surrogate model train on the data points to eventually generate explanations [3]. It is important to note that Alvi highlights a significant limitation to her work: the computationally cost of re-extracting visual features for every new image perturbation. The computational burden of this action limited the number of explanation experiments and the number of perturbations for each experiment she could carry out [3].

Chapter 3

Methods

This chapter describes the choice of research method and application of the research method. Section 3.1 elaborates on method choices, while Section 3.2 gives details on the implementation workflow with descriptions of the data set, training of vision and language models, training of the explainable models, as well as the evaluation of generated explanations.

3.1 Choice of Research Method

In order to decide on the most suitable research methods for our study, we advised Anne Håkansson's work *Portal of Research Methods and Methodologies for Research Projects and Degree Projects* which provides a portal of research methods and methodologies aimed at helping students in their degree project [18].

Håkansson states that the first overarching research methodology choice is to decide between a quantitative or qualitative method. Quantitative studies are numerical and aim to explore and test a hypothesis by measuring and quantifying variables. On the other hand, qualitative studies are non-numerical and aim to understand contexts, viewpoints, and behaviors to form a hypothesis, theories, or develop inventions and artifacts. The data sets for qualitative studies are generally smaller than for quantitative ones [18]. In our study, we want to form an understanding of the feasibility of creating image and text explanations that resembles domain expertise by using the local surrogate explainable technique in combination with an underlying vision and language model. Inherently, the intuitiveness of explanations are rather subjective and difficult to formalize [56]. Since we want to build an understanding and reach a tentative hypothesis on the feasibility of creating vision and language explanations with the help of local surrogate models, we concluded that a qualitative research methodology is the most suitable.

Moreover, this is an *applied research* study, a method that refers to trying to solve specific practical problems. We considered other research methods such as experimental research, non-experimental research, descriptive research, analytical research, and fundamental research [18]. However, we concluded that the applied research method is the most suitable as we want to address the specific problem of the lack of transparency of promising vision and language models in the medical field.

Håkansson also describes that authors must choose between an inductive, deductive, and abductive research approach. An inductive approach seeks to, from observations, formulate a general conclusion that is likely but not necessarily certain. In contrast, a deductive approach makes certain conclusions by verifying or falsifying a hypothesis. The deductive approach is quantitative by nature and requires a large enough data set to make a conclusion. An abductive approach uses both the inductive and deductive approaches. From an incomplete data set and set of observations, the abductive research approach seeks to find the hypothesis that most likely serves as an explanation for the set. We choose to adopt the *abductive* research approach as we, from the limited case study observations, form the most likely conclusion to our issued research question.

Next, with the motivation to address the lack of transparency of promising vision and language models, yet the constant need for explainability in the medical field, we choose to adopt a *case study strategy*. A case study refers to an empirical study on a practical phenomenon [18]. Our case study explains predictions made by vision and language model trained on the task to predict thoracic findings from radio-graphs and radiology reports. By evaluating the generated explanations, we will then formulate an answer to our research question.

Moreover, since we adopt a case study strategy, a suitable data collection method is a *case study* methodology [18]. We have made an explicit study on a thoracic findings data set, namely the OpenI data set [11]. Please advise Section 3.2.2 from an in-depth description of the data set used in the study. From this data set, we have trained the vision and language models and used the input data to generate explanations. However, in addition to applying a case study data collection methodology, we also use *questionnaires* [18] to collect annotations of explanations by domain experts. For this, we design a labeling environment that Section 3.2.5 describes.

3.2 Application of Research Method

3.2.1 Workflow Overview

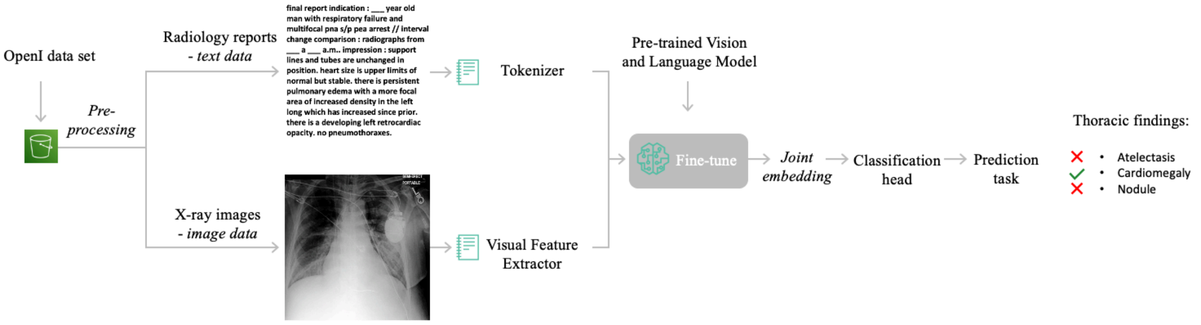
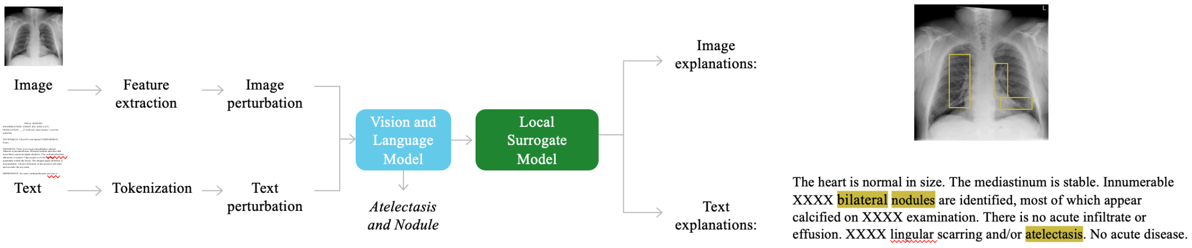

OpenI data set
Pre-processing
Radiology reports - text data
final report indication : ___ year old man with respiratory failure and multifocal pna s/p pea arrest // interval change comparison : radiographs from ___ a ___ a.m.. impression : support lines and tubes are unchanged in position. heart size is upper limits of normal but stable. there is persistent pulmonary edema with a more focal area of increased density in the left long which has increased since prior. there is a developing left retrocardiac opacity. no pneumothoraxes.
X-ray images - image data
PORTABLE
Tokenizer
Visual Feature Extractor
Pre-trained Vision and Language Model
Fine-tune
Joint embedding
Classification head
Prediction task
Thoracic findings:
✗ Atelectasis
✓ Cardiomegaly
✗ Nodule

Figure 3.2.1: Schematic figure illustrating the training of the vision and language models.

Figure 3.2.1 illustrates the workflow of training the vision and language models. First of all, we have carried out pre-processing steps on the OpenI data set. The major steps are tokenization of the text modality and visual feature extraction of the images. Please advise Section 3.2.2 for more details on the data set and pre-processing steps. The tokens and visual features together make up the input space for the vision and language models used in the study. The vision and language models are fine-tuned on the downstream task of classifying three thoracic findings: Atelectasis, Cardiomegaly, and Nodule. The models learn to project the text and visual features to a latent space with the same dimensions [9] [26]. Finally, a classification head is added to the joint image and text embedding to perform the multi-class classification task. The classification head generates a probability matrix across the classes of thoracic findings, and with that, we generate the predictions. Please advise Section 3.2.3 for further details on model training and evaluation. We downloaded the pre-trained weights for the two architectures used, namely UNITER and VisualBERT. Both of these models have in previous studies demonstrated state-of-the-art performance on vision and language tasks [9] [26]. The transformer module of both models inputs text tokens and visual features in terms of box positions and associated visual embedding vectors.

Therefore, these models were considered suitable for our explainability experiments that specifically aim to perturb those features. Please advise 3.2.4 for more details on the perturbations.



The heart is normal in size. The mediastinum is stable. Innumerable XXXX bilateral nodules are identified, most of which appear calcified on XXXX examination. There is no acute infiltrate or effusion. XXXX lingular scarring and/or atelectasis. No acute disease.

Figure 3.2.2: Schematic figure illustrating the training of a local surrogate explainable model.

Once the training was completed, we built explainable models to explain the predictions by the vision and language models. Figure 3.2.2 illustrates the workflow of training an explainable model. In essence, visual features and tokens are perturbed, and proximity to the original vector is computed for each perturbation. After that, new model outputs are generated by feeding the underlying vision and language model with the perturbed input vectors. Next, we compute the prediction loss between the new model outputs in relation to the original model output. The original model output refers to the output from feeding the model with the original input vector. Finally, we fit a linear regression model and use the interpretable architecture to generate explanations by using the weights of the linear surrogate model.

The reason for perturbing specifically pre-processed tokens and visual features as oppose to raw data is twofold: First, it explores an alternative pathway targeting the limitation in Alvi's work of having to re-extract visual features for every new image perturbation [3]. Second, it could serve as an interactive feedback loop by perturbing the direct inputs to the transformer modules of the VisualBERT and UNITER model and generating explanations considering specifically those features. Highlighting of important features in explanations could serve as helpful feedback for data scientist and ML engineers by giving guidance for model training. For instance, explanations could indicate that the model puts attention on irrelevant features which could give guidance on the pre-processing steps of generating tokens and visual features from the radiology reports and X-ray images.

Moreover, our implementation is inspired by how LIME generates explanations [47], but we have made extensions to a multi-modal case as well as perturbing visual features as opposed to perturbing pixel regions of an image.

3.2.2 Data Set and Pre-Processing

We used the publicly available OpenI data set for model training and evaluation. The OpenI data set, created by the Indiana University, contains 3851 matching data points with radiology reports and chest X-ray images, with labels of 14 thoracic findings from unique patients. Radiologists have manually performed the labelling of the thoracic findings [11]. We also considered using the MIMIC-CXR data set, which also contains radiology reports and chest X-ray images as well as similar labels [22]. In fact, after exploring the MIMIC-CXR data set and performing our pre-processing steps, the data set had 150 000 data instances, so significantly larger than the OpenI data set. Yet, the labels MIMIC-CXR data set are generated automatically from images and reports using ChexPert [21] and NegBio [44]. In contrast, the labels from the OpenI data set are produced by professional annotators and can therefore be viewed as more reliable. In addition, we had the visual features of the X-ray images in the OpenI data set readily available for training [27], while for the MIMIC-CXR we would have had to implement a visual feature extractor object network to extract the features ourselves. Extracting the visual features of the MIMIC-CXR data set turned out to involve several technical challenges. Given the project timeline and the more reliable labels, we chose to build our study using only the OpenI data set.

Both of the transformer modules of UNITER and VisualBERT inputs text tokens and visual features [26] [9]. We pre-process the text from the radiology reports through a BERT encoder network to produce text tokens [12]. The visual features were extracted using the *bottom up top down attention* network Detectron2 [54]. The resulting visual features contains visual boxes paired with visual embedding vectors. In total, each image is represented with <36 x 4> box position vectors and <36 x 2048> visual feature embedding vectors.

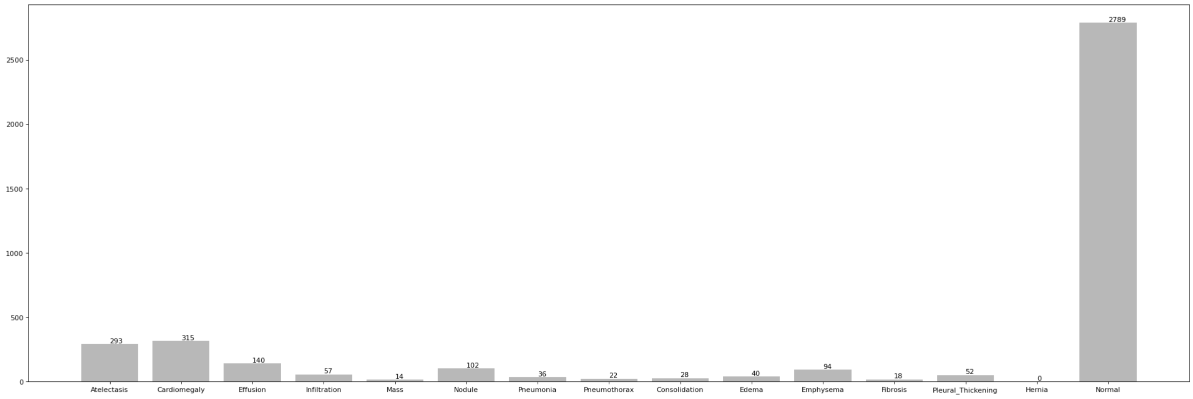


Figure 3.2.3: Displays data distribution across the classes of thoracic findings before down-sampling.

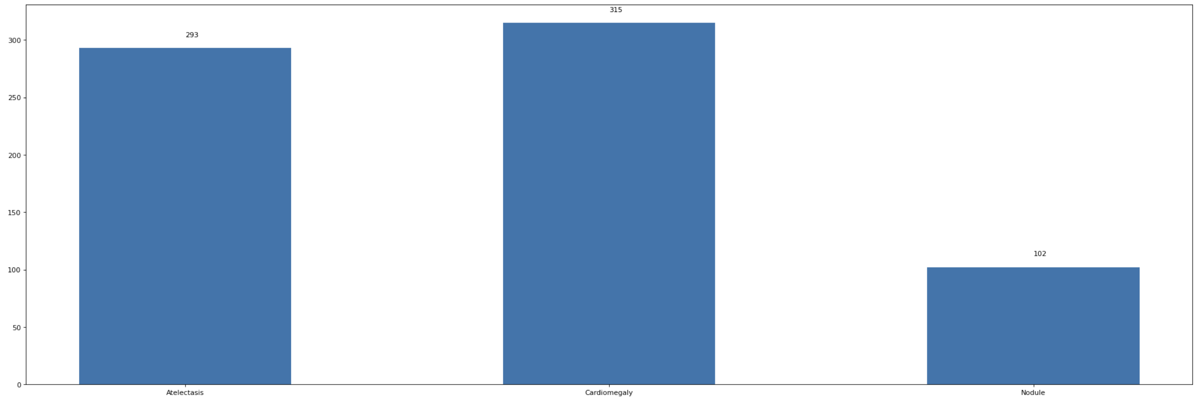


Figure 3.2.4: Displays data distribution across classes of thoracic finding post down-sampling.

Moreover, given the relatively unbalanced class distribution of the OpenI data set, as can be seen in Figure 3.2.3, the following three labels were selected out of the 14 thoracic findings: Atelectasis, Cardiomegaly, and Nodule. These labels were among the most dominant ones in the data set and helped to balance the class distribution and simplified model training as can be seen in Figure 3.2.4. The down-sampling, i.e., transforming the data set down to three classes, reduced the data set to a total of 786 data points (where 139 were negative, i.e., none of the three findings). Further, the data set was divided into a train set of 594 data points and a test set with 192 instances. The split into train and test was done to preserve the same class distribution to the largest extent possible.

3.2.3 Train and Evaluate Vision and Language Models

For training, we load pre-trained weights of the VisualBERT and UNITER networks which both have been trained mainly on the COCO data set [28] as well as the VQA

2.0 [1] and Visual Genome [24] data sets. For both models, we use the visual question answering architecture [26][9] and set the number of questions equal to the number of classes, which enables us to treat it as a multi-label classification task. This approach has been adopted in previous work of predicting thoracic findings from the OpenI data set [27]. Moreover, we fine-tune with seven epochs on the VisualBERT model and use eight epochs for the UNITER model chosen to minimize the training loss. As a result of the fine-tuning, the models learn a joint image-and-text representation. Moreover, we use a classification head with a linear network to generate a probability matrix across the thoracic findings. For both models, we use a weight decay of 5e-4 and a learning rate of 1e-5. In addition, both of the transformers optimize with the Adam optimizer, and we use cross entropy as the loss function. Finally, we report a confusion matrix and AUC score on the test data set. We train using PyTorch inside the Amazon SageMaker ML platform with a GPU accelerated *ml.g4dn.xlarge* instance powered by one NVIDIA T4 GPU.

3.2.4 Build Explainable Models

Explainable model:	Approach:	Description:
Separate Perturbations	Perturb only Text Tokens while keeping Visual Features Fixed + Perturb only Visual Features while keeping Text Tokens Fixed	Finds text and image explanation by combining the outcome of one explainable model that perturbs only the visual embedding and one that perturbs only the tokens.
Simultaneous perturbations	Perturb both Modalities Simultaneously	Finds text and image explanations by training an explainable model that perturbs both modalities simultaneously.

Table 3.2.1: Describes the two explainable models of simultaneous and separate perturbations.

As aforementioned, we fit a simple model to explain individual predictions made by the vision and language models by perturbing the input data. Our goal is to generate multimodal explanations, i.e., text and image explanations. To do so, we need to perturb both modalities to fit a simple model that can generate explanations for each modality. Intuitively, there are two ways of achieving this goal, either we perturb the vision and language modality separately and combine the generated explanations afterward, or we perturb both modalities simultaneously and build an explainable model for that. With this motivation, we carried out three types of perturbations:

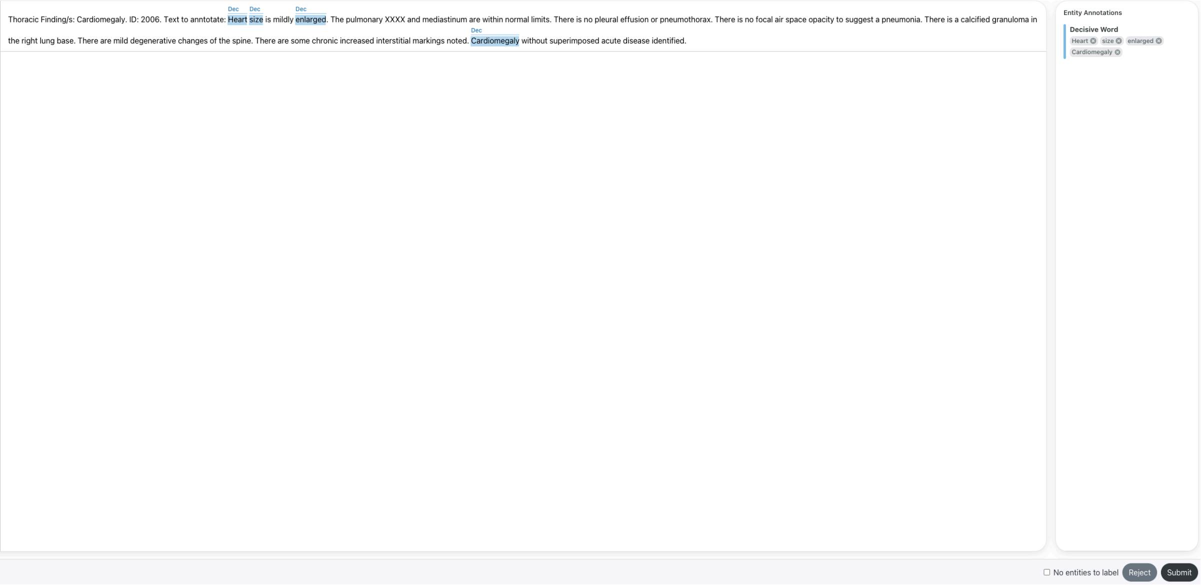
- First, we perturb only the tokens while keeping the visual features fixed.
- Next, we perturb the visual features while keeping the tokens fixed.
- Lastly, we perturb both modalities simultaneously.

The outcomes of the first two together serve as a multi-modal explanation, while the latter one does so on its own. For more details on the visual and language explainable models, please visit Chapter 4.

3.2.5 Evaluate Explainable Models

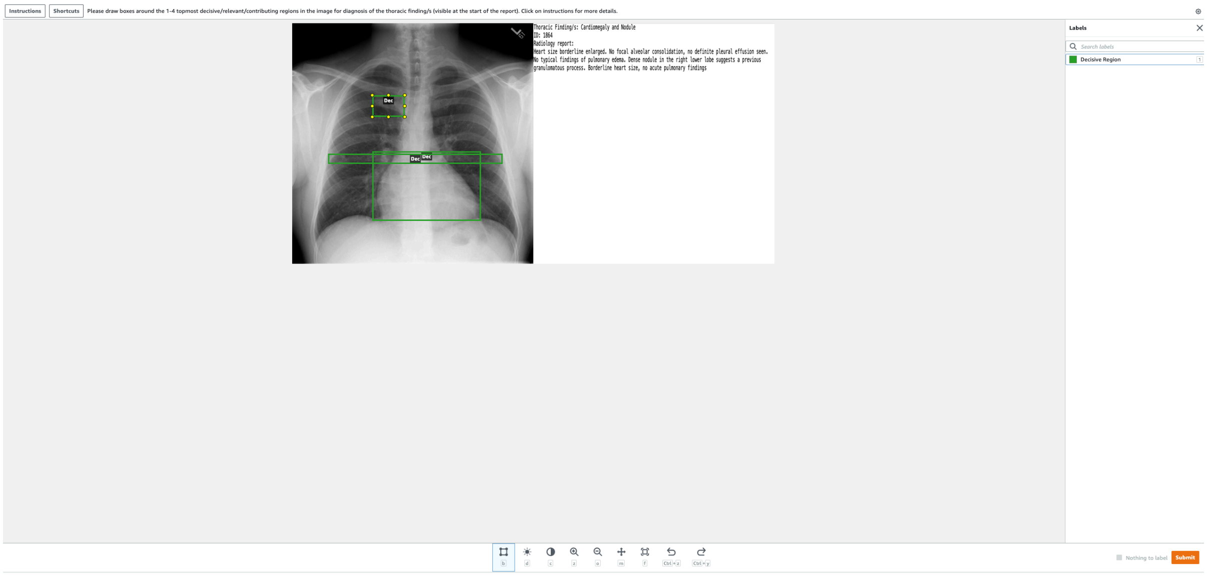
To evaluate generated explanations from the explainable models, we let radiology domain experts highlight important words and important parts of the images [14] [4] [33]. Referring to the three evaluation approaches of explainable ML by Doshi-Velex and Kim[13] presented in Section 2.2.2, we make use of the application-grounded evaluation approach. In our case study, we have a special application in mind, namely to explain model predictions of thoracic findings. Moreover, the identification of explainable features for this ML task requires domain expertise. Accordingly, to evaluate the explanations from the explainable model, it should be logical to evaluate against human-subject domain experts performing on the same task.

For this, we collected annotations from radiology domain experts on 46 sample data points. The sample data points were extracted from the test set to mimic the class distribution of the test data set. We created an annotation environment for domain experts to highlight the X-ray images and radiology reports. In total, we managed to collect explanations from three domain experts [14] [4] [33].



Thoracic Finding/s: Cardiomegaly. ID: 2006. Text to anntotate: Heart size is mildly enlarged. The pulmonary XXXX and mediastinum are within normal limits. There is no pleural effusion or pneumothorax. There is no focal air space opacity to suggest a pneumonia. There is a calcified granuloma in the right lung base. There are mild degenerative changes of the spine. There are some chronic increased interstitial markings noted. Cardiomegaly without superimposed acute disease identified.

Figure 3.2.5: Displays an example of a text annotation job from the annotation tool.



Please draw boxes around the 1-4 topmost decisive/relevant/contributing regions in the image for diagnosis of the thoracic finding/s (visible at the start of the report). Click on instructions for more details.

Thoracic Finding/s: Cardiomegaly and Nodule
ID: 1864
Radiology report:
Heart size borderline enlarged. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. Dense nodule in the right lower lobe suggests a previous granulomatous process. Borderline heart size, no acute pulmonary findings

Labels

🟩 Decisive Region 1

□ Nothing to label

Figure 3.2.6: Displays an example of an image annotation job from the annotation tool.

Figure 3.2.5 and 3.2.5 shows an example of both a text and image labelling job from the created annotation environment. The labeling jobs were created using the serverless cloud tool Amazon SageMaker Ground Truth [50]. We launched one labeling job for the image modality and one for the text modality and sent a link to the domain experts leading to a web application to perform the annotations. Before the domain experts made their annotations, we organized a meeting to demonstrate with examples how to annotate both the X-ray images and radiology reports. Specifically, we asked the domain experts to, given the presence of thoracic findings in a sample, draw bounding boxes around the most decisive or pathological regions in the image, as well as highlight the set of most important words for identifying the specified thoracic findings.

To evaluate the success of our explanations generated by the explainable models, we compared them with the annotations made by the domain experts. We measure the degree of overlap for the text modality by computing the Jaccard similarity [25] between identified decisive words of the explainable models compared to words highlighted by the domain experts as in Equation 3.1. For the image modality, we compute the intersection of the union [46] of identified regions of the explainable model and regions annotated by the domain experts as described in Equation 3.2.

$$Sim_{Text}(D_w, E_w) = \frac{D_w \cap E_w}{D_w \cup E_w} \tag{3.1}$$

In Equation 3.1, D_w represents the identified set of explainable words for the domain expert, while E_w refers to the set of important words highlighted by the explainable models.

$$Sim_{Image}(D_b, E_b) = \frac{D_b \cap E_b}{D_b \cup E_b} \tag{3.2}$$

In Equation 3.2, D_b refers to drawn boxes by the domain expert, while E_b represents boxes identified by the explainable model.

Except for comparing the similarity of the explanations from the explainable model and the domain experts, we made two additional comparisons to serve as benchmarks. The first one was a comparison between the domain experts. To better understand the degree of similarity of the explainable model results to annotations of domain experts, it should be helpful to measure the similarity between domain experts [13]. Therefore, we computed the text and image similarity between the domain experts as an additional benchmark. The other additional comparison was to produce a random baseline by randomly picking a language and image explanations. We randomly choose decisive words and visual boxes to represent random explanations. Next, we compared the similarity of those random explanations to that of domain experts.

Chapter 4

Visual and Language Explainable Models

This part gives a more detailed description on the visual and language explainable models. The Section 4.1 is more elaborate, while the Sections 4.2 and 4.3 extends the first description.

4.1 Perturb only Text Tokens while keeping Visual Features Fixed

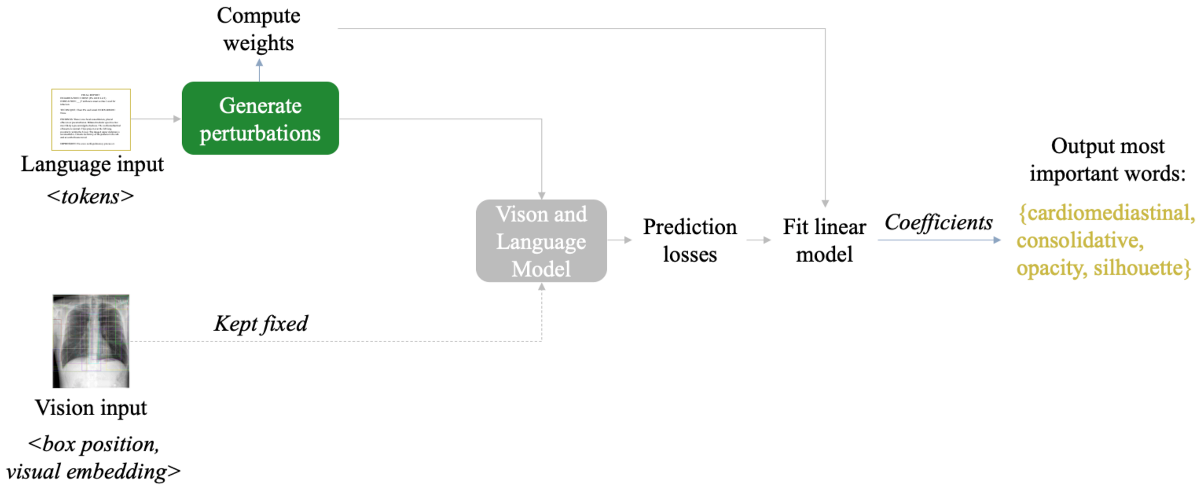


Figure 4.1.1: Schematic figure illustrating the training of an explainable model perturbing only the text modality while keeping the image modality fixed.

Figure 4.1.1 illustrates the implementation workflow of training an explainable model from only perturbing the text modality. Our implementation is inspired by the python LIME implementation [48]. The explainable model takes as input:

- *Original prediction* - The probability matrix generated from the original inputs.
- *Model* - The model used to generate the original prediction.
- *Text* - The radiology report.
- *Text tokens* - The original pre-processed tokens.
- *Visual feature boxes* - The positions of visual feature boxes in the input image.
- *Visual feature embedding vectors* - The original feature vectors associated with each box.

In addition, the model has a number of hyperparameters:

- *Number of samples* - The number of perturbations for a single prediction to be explained.
- *Number of important words* - The number of important words to highlight in the explanation output.
- *Number of features* - The number of top-ranked perturbation vectors to extract important words from.
- *Distance metric* - The distance metric used to compute distance of a perturbed vector to the original vector.
- *Kernel width* - A float value between 0 and 1 that impacts the weights assigned to a perturbation vector where a smaller number will give more weight to perturbed vectors that are closer to the original input vector.

Moreover, the model generates text explanations by performing the following training steps:

- First, the model generates perturbations of the original text token vector by randomly turning on and off tokens and stores the associated words that have been inactivated. We represent each perturbation as a binary vector where each vector position represent whether a token is active or not. Further, we compute the pairwise distance between perturbation vectors and the original vector. Note that the original input vector is represented by an array where all positions are set to one as this refers to all features being active.

- Moreover, we assign a weight to each perturbation sample by feeding the computed distances into a kernel function as in the LIME implementation [47]. The kernel functions assigns more weight on samples closer to the original vector. The logic behind giving more weights to closer vectors is that if such a vector ends up having a large impact on the model output, then those few tokens that have been inactivated in the sample are likely to be decisive ones [47].

- Next, we feed the perturbed vectors to the model. For this, we use a *wrapper function* that inputs the text tokens while constantly keeping the visual features fixed, and use these inputs to make model inference. Finally, we compute the cross entropy loss between the outputted probability matrix to that of the original prediction.

- Finally, we fit a linear model on the perturbed feature vectors, using the computed weights, and the prediction losses. Then we take out the vectors associated with the top-ranked coefficients. From these perturbed vectors, we collect the inactivated words, and extracts the most frequent ones to serve as the text explanation.

4.2 Perturb only Visual Features while keeping Text Tokens Fixed

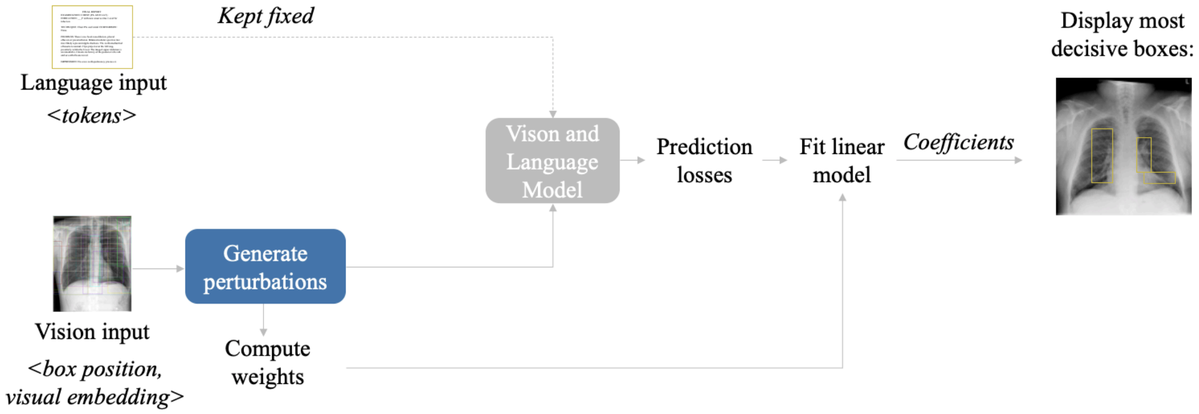


Figure 4.2.1: Schematic figure illustrating the training of an explainable model perturbing only the image modality while keeping the text modality fixed.

Figure 4.2.1 illustrates the training of the explainable model that only perturbs the image modality. The implementation is similar as described in Section *Perturb only Text Tokens while keeping Visual Features Fixed*. However, this time we keep the text tokens fixed and only perturb the visual features. The model takes the same inputs as previously described except that the hyperparameters differ slightly: *Number of important words* is removed, but another parameter is added, *Probability p*, which is the probability parameter used in a binomial function to generate perturbed visual features.

The visual features consist of box positions and visual embedding vectors mapped together. Important to note is that there is a fixed size of 36 visual boxes and visual embedding vectors for each image. More specifically, the choice of possible image explanations is limited to the set of 36 boxes to choose from.

Furthermore, we use a binomial distribution function to generate perturbations of the original visual features where all ones represent the original input vector while the perturbations also contain zeros representing inactivated visual features. The *Probability p* hyperparameter determines the probability of inactivating a feature. The inactivation of a feature means that the associated visual feature vector's elements are all set to zeros.

To the best of our knowledge there is no previous work that uses the approach of perturbing the visual features as a mean to generate visual explanations using a local surrogate explainable model. The logic is similar to how LIME generates image perturbations, but LIME perturbs pixel regions instead of visual boxes [47].

Furthermore, as described in Section *Perturb only Text Tokens while keeping Visual Features Fixed*, we assign a weighted score to each perturbation. The *wrapper function* keeps the text tokens fixed this time, while varying the visual features, and performs model inferences accordingly. Finally, we fit a linear model and, this time, use the coefficients to find the top most decisive perturbations of visual features and collect the inactivated boxes for those perturbations. As an output, the model draws the most decisive boxes onto the input image, which serves as the image explanation.

4.3 Perturb both Modalities Simultaneously

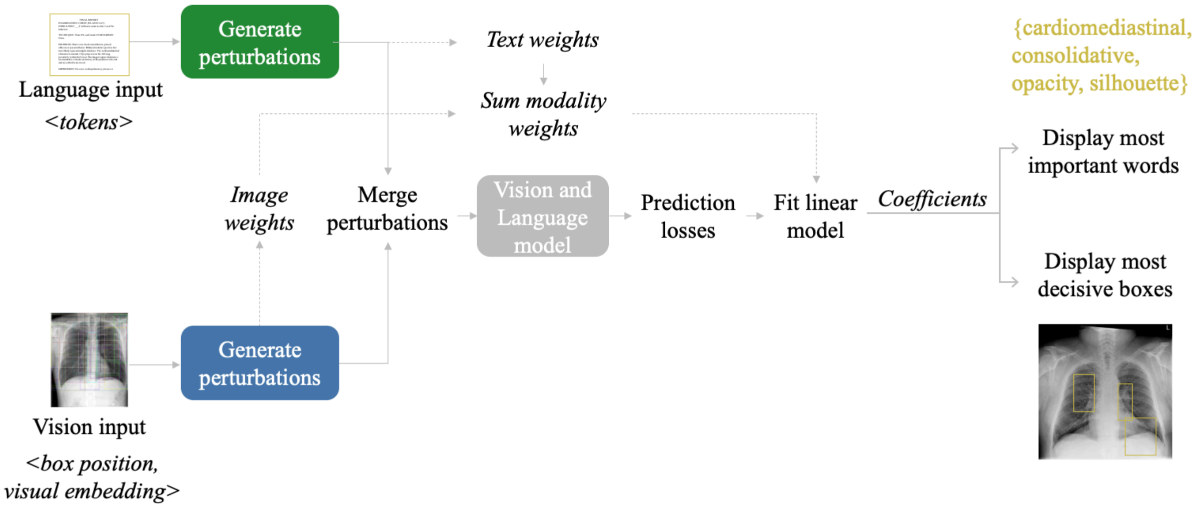


Figure 4.3.1: Schematic figure illustrating the training of an explainable model perturbing both modalities simultaneously.

Figure 4.3.1 illustrates our implementation of an explainable model that perturbs both the vision and language modality simultaneously. The implementation is a combination of the ones described in the Sections *Perturb only Text Tokens while keeping Visual Features Fixed* and *Perturb only Visual Features while keeping Text Tokens Fixed*. The inputs of this model are the union of the two input sets of the two previously described models. In addition, the multi-modal explainable model has hyperparameters specifying the number of important perturbed feature vectors to consider for the text explanation, as well as for the image explanation. Additionally, the hyperparameters of the multi-modal explainable model allow to specify what kernel width and what distance metric to use with dedicated parameters targeting the text and image modality separately.

First, the same number of perturbations are generated for the tokens and the visual features, and weights are computed for each sample of each modality. Next, the weights from each modality are first normalized and then summed up, and then the resulting weight vector after the summation is also normalized. Next, the token- and visual perturbations are concatenated, each sample is fed for model inference, and prediction loss is computed. Finally, we fit a linear model and extract the top ranked vectors with regards to associated coefficients. Similar to the two previously described models, we then output the most decisive words and visual boxes, and, together, they serve as a multi-modal explanation.

4.4 Training Details

We train the explainable models using PyTorch inside the Amazon SageMaker with a GPU accelerated *ml.g4dn.xlarge* instance as model inference of the vision and language models requires CUDA.

Chapter 5

Results and Discussion

This chapter presents and discusses the results. The first part reports the underlying model performance on the test data set, while the second part presents and discusses the evaluation of the explainable models.

5.1 Underlying Model Performance

Even if the focus of our work is on explainable models, a necessary pre-requisite for the case study was to train vision and language models on the clinical task of predicting thoracic findings. As aforementioned, two vision and language models were trained on this task, one with the VisualBERT architecture and one with the UNITER architecture. Figure 5.1.1 displays the ROC curves for the models across the findings. Both of these figures report performance on the test set. To be noted is that this is a multi-label classification task, so a particular instance may contain one or more thoracic findings. The negative examples, i.e., the samples with none of the three findings, were handled as negative classes in the evaluation.

As can bee seen in Figure 5.1.1, the VisualBERT and UNITER report AUC scores above 0.97 for all three thoracic findings. Yet, as described in Section 3.2.2, the data set is relatively small, so there is a risk of learnt biases which is further discussed in Section 5.2.

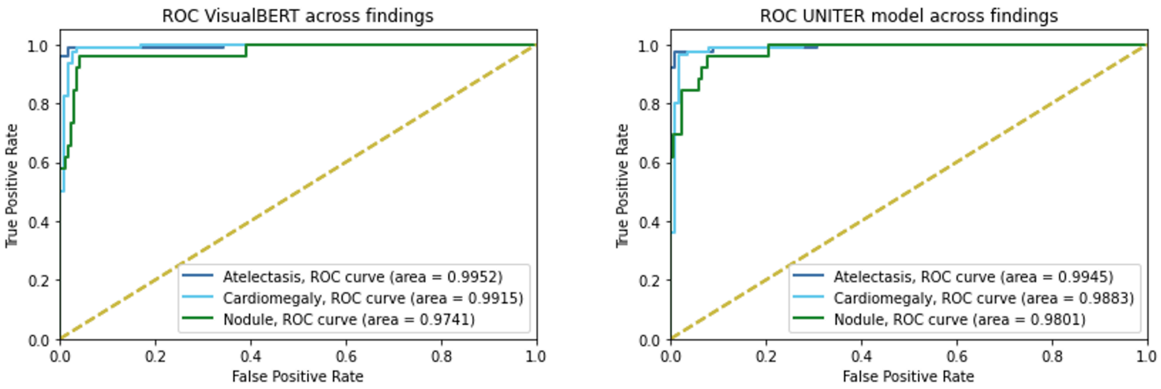


Figure 5.1.1: ROC curve for VisualBERT and UNITER models across findings.

5.2 Evaluation of Explainable Models

To evaluate the results of the explainable models, explanations from domain experts were collected and used as a benchmark. Table 5.2.1 presents the similarity of the text respectively the image explanations for each underlying model and for each type of explainable model. In addition, Table 5.2.4 shows average similarity scores across the variety of underlying models, perturbation types, and domain experts.

Text explanations refer to identified important words given a prediction of a thoracic finding, and the similarity is computed as in Equation 3.1, i.e., the intersection over the union of identified words from the explainable model and domain experts. Image explanations refer to identified important regions of the X-ray images and the similarity is the intersection of union between the explainable model explanations and that of domain experts as described in Equation 3.2. However, please advise Section 3.2.4 for a more detailed description on these metrics and the evaluation approach of explanations.

Table 5.2.1: Displays similarity of text explanations and image explanations between explainable model and domain experts across experiments and type of underlying models.

		Domain expert 1		**Domain Expert 2**		**Domain Expert 3**	
Underlying Model:	**Experiment:**	**Text similarity:**	**Image similarity:**	**Text similarity:**	**Image similarity:**	**Text similarity**	**Image similarity**
UNITER	Simultaneous Perturbations	0.083	0.119	0.085	0.156	0.096	0.238
UNITER	Separate Perturbations	0.103	0.102	0.122	0.172	0.138	0.261
VisualBERT	Simultaneous Perturbations	0.073	0.091	0.079	0.016	0.100	0.261
VisualBERT	Separate Perturbations	0.128	0.102	0.171	0.172	0.117	0.302
	Average:	0.097	0.104	0.114	0.165	0.113	0.270

Average text similarity across domain experts:	0.108
Average image similarity across domain experts:	0.178

As can be seen in Table 5.2.1, the similarity scores indicate that the explanations from the explainable model and domain experts are far from identical. The text similarity is rather low, close to 11%, while the image similarity on average is higher, yet only around 18%. However, in discussion with radiology domain experts [14] [33] [4], the experts independently mentioned that reasoning and explanations vary from one domain expert to another for this particular task. To measure the level of variation between annotations gathered from the experts, text and image similarity were computed between the domain experts. The results can be seen in Table 5.2.2. While these comparisons received higher similarity scores compared to those in 5.2.1, the scores are still far from an identical score, i.e., a similarity score of 100%. Rather, the similarity scores support notable variations between domain experts' explanations.

Table 5.2.2: Displays the similarity of text explanations and image explanations between domain experts.

	Text Similarity:	**Image Similarity:**
Domain Expert 1 and 2	0.62	0.38
Domain Expert 1 and 3	0.36	0.32
Domain Expert 2 and 3	0.40	0.34

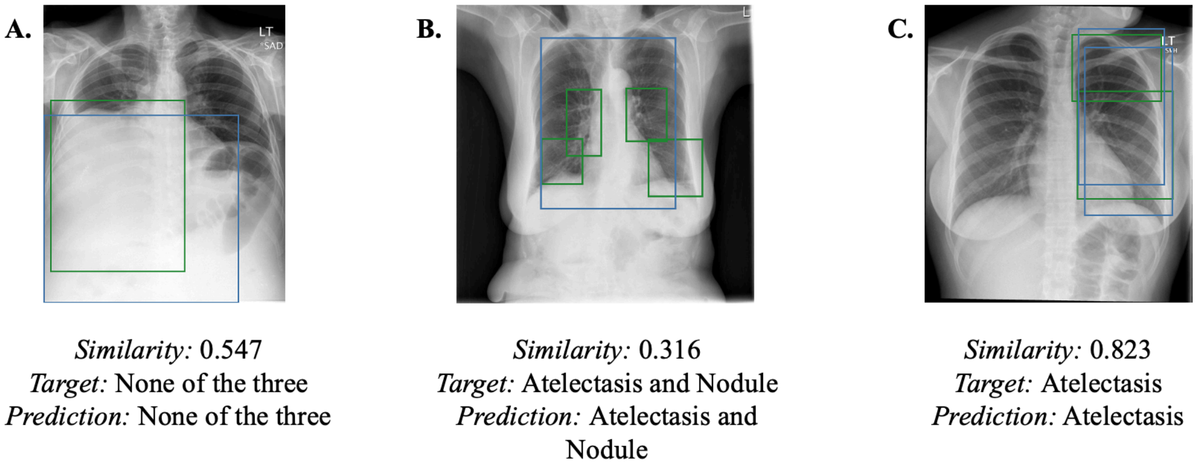
Moreover, suppose the outcomes of the explainable models turned out to be very close to explanations from domain experts. In that case, we might conclude that the explanations well resemble that of domain expertise. Yet, given the notable differences between domain experts, as can be seen in Table 5.2.2, such a result seems difficult to reach. Rather, these results indicate that this is a difficult task to explain and that the ground truth is ambiguous in terms of true explanations. More specifically, the results suggest that there is ambiguity among domain experts on what are contributing features from radio-graphs and radiology reports to explain thoracic findings. With this motivation, the similarity scores between the domain experts could represent an upper bound for the explainable model rather than a perfect match of 100% similarity. So, if the model would reach the same level of similarity as between the domain expert, we might conclude that explanations resemble domain expertise. However, to make further conclusions on our explanations, it could be beneficial to compare with a lower bound as well. Instead of generating explanations via our explainable model, we randomly picked words from the radiology reports and visual boxes extracted from the X-ray images to represent random explanations. Then, computing similarity scores between random explanations and explanations from domain experts serve as a baseline or a lower bound. Table 5.2.3 showcases the similarity scores for the resulting random baseline.

Table 5.2.3: Displays the random baseline showing text and image similarity domain experts for randomly choosing words and visual boxes.

Domain Expert:	Text Similarity:	Image Similarity:
Domain Expert 1	0.038	0.059
Domain Expert 2	0.051	0.107
Domain Expert 3	0.041	0.130
Average:	0.043	0.099

As can be seen in Table 5.2.3 when randomly choosing explanations of words and visual boxes, the similarity scores are lower than those of the explainable model. Even though the difference between the similarity scores of the random baseline and the explainable model is not massive, one can distinguish a general trend. The results indicate that although there is a difference between the similarity scores of the explainable model and the similarity scores between domain experts, it is still better than the random baseline. Hence, although the generated explanations had relatively low similarity scores to domain experts, the explainable model seems to some extent have captured signal in explaining the predictions.

Even though the similarity scores measure the similarity to domain experts, to learn more about the explanations of the explainable models, it should be helpful to examine some examples of image and text explanations. Figure 5.2.1 and 5.2.2 showcases three examples each of image explanations and Figure 5.2.3 and Figure 5.2.4 display three example each of text explanations.

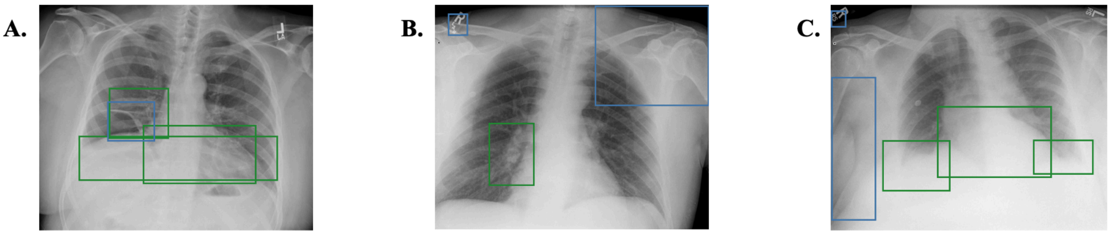


Similarity: 0.547
Target: None of the three
Prediction: None of the three

Similarity: 0.316
Target: Atelectasis and Nodule
Prediction: Atelectasis and Nodule

Similarity: 0.823
Target: Atelectasis
Prediction: Atelectasis

Figure 5.2.1: Displays three examples of explanations from explainable models (blue boxes) and explanations from domain expert (green boxes).



Similarity: 0.140
Target: Atelectasis and Cardiomegaly
Prediction: Atelectasis

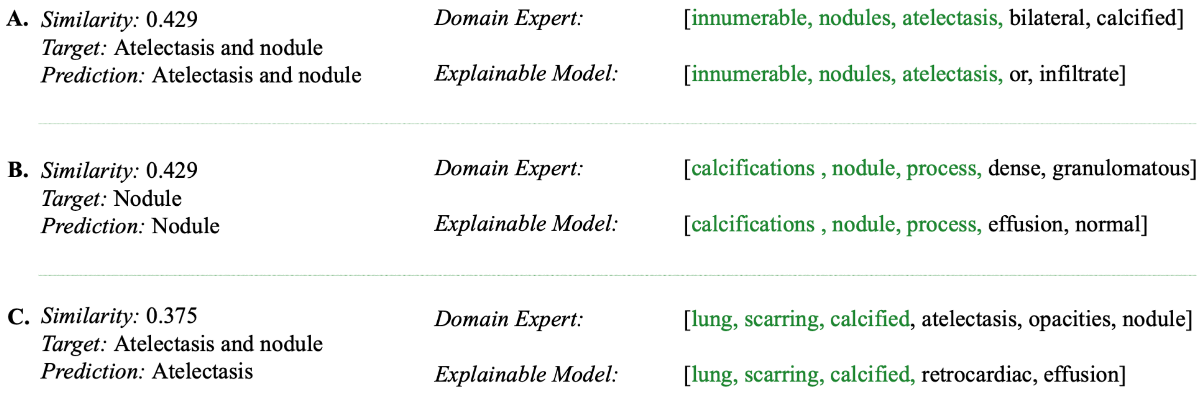
Similarity: 0
Target: Atelectasis
Prediction: Atelectasis

Similarity: 0
Target: Atelectasis and Cardiomegaly
Prediction: Atelectasis and Cardiomegaly

Figure 5.2.2: Displays three examples of explanations from explainable models (blue boxes) and explanations from domain expert (green boxes).

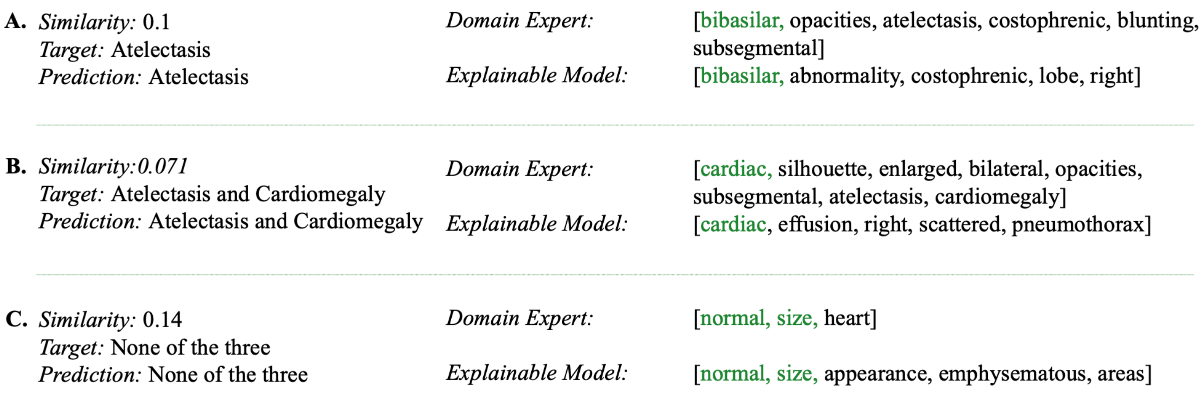
When studying Figure 5.2.1, it seems that even though the model somewhat matches domain expertise, the boxes look too large and rigid compared to those of the domain experts. There is a limited set of visual boxes that the explainable model could choose from in our experiments. The fixed amount of visual boxes likely limits how well the generated explanations could resemble that of domain expertise. Moreover, example *A* in Figure 5.2.2 is a particularly interesting one to study. For this example, the underlying UNITER model predicted only atelectasis, but in reality, the ground truth says both atelactasis and cardiomegaly. The domain experts were given the ground truth and drew their boxes accordingly, represented by the green boxes in the image. As can be seen, the domain expert highlighted the heart region, the horizontal space between the ribs, as well as a region at the mid-left where the latter contains the atelectasis region [33]. Now, looking at the blue box drawn by the explainable model, this one seems to more or less also capture the atelectasis region. However, the explainable model did not highlight the heart region, which typically is associated with cardiomegaly [14]. In this example, it seems that the explanations from the explainable model compared to domain expertise could explain why the model only predicted atelectasis but missed to include cardiomegaly.

Moreover, when studying example *B* and *C* in Figure 5.2.2, one can see that the model seems to focus on the outer parts of the image as these were highlighted as the most contributing region to the made prediction. Also, these examples highlight letters and numbers in the corner of the images. Such explanations could suggest that the model has captured some signal in these letters and numbers, which generally should be irrelevant for the ML task. Potentially, the model has learned bias in this case. Our data set is relatively small, and we cannot disregard learned biases. One natural way to overcome this is to train the models on a more extensive data set, and it should be a natural next step for future work. Nevertheless, another hypothesis to deal with what looks to be learned biases in detecting the letters and numbers could be to add pre-processing steps to the X-ray images. From looking at the explanations provided by the domain experts, we noticed that they consistently are centered around the thorax region. Hence, the outer part of the images seems to be less relevant for predicting the thoracic findings. We hypothesize that by helping the underlying model as well as the explainable model to focus on the thoracic regions of the images, the explainable models could better resemble explanations of explainable models. Unfortunately, we could not explore such guidance in this work. However, future work could explore eventual improvements by, for instance, blurring outer regions of the X-ray images or applying other denoising pre-processing techniques. Such pre-processing techniques would typically mean that new visual features need to be extracted from the images, and the underlying models and explainable models should retrain accordingly.



A. *Similarity:* 0.429 *Target:* Atelectasis and nodule *Prediction:* Atelectasis and nodule	*Domain Expert:* *Explainable Model:*	[innumerable, nodules, atelectasis, bilateral, calcified] [innumerable, nodules, atelectasis, or, infiltrate]
B. *Similarity:* 0.429 *Target:* Nodule *Prediction:* Nodule	*Domain Expert:* *Explainable Model:*	[calcifications , nodule, process, dense, granulomatous] [calcifications , nodule, process, effusion, normal]
C. *Similarity:* 0.375 *Target:* Atelectasis and nodule *Prediction:* Atelectasis	*Domain Expert:* *Explainable Model:*	[lung, scarring, calcified, atelectasis, opacities, nodule] [lung, scarring, calcified, retrocardiac, effusion]

Figure 5.2.3: Displays three examples of words from domain experts and from the explainable models.



A. *Similarity:* 0.1 *Target:* Atelectasis *Prediction:* Atelectasis	*Domain Expert:* *Explainable Model:*	[bibasilar, opacities, atelectasis, costophrenic, blunting, subsegmental] [bibasilar, abnormality, costophrenic, lobe, right]
B. *Similarity:0.071* *Target:* Atelectasis and Cardiomegaly *Prediction:* Atelectasis and Cardiomegaly	*Domain Expert:* *Explainable Model:*	[cardiac, silhouette, enlarged, bilateral, opacities, subsegmental, atelectasis, cardiomegaly] [cardiac, effusion, right, scattered, pneumothorax]
C. *Similarity:* 0.14 *Target:* None of the three *Prediction:* None of the three	*Domain Expert:* *Explainable Model:*	[normal, size, heart] [normal, size, appearance, emphysematous, areas]

Figure 5.2.4: Displays three examples of words from domain experts and from the explainable models.

Furthermore, Figure 5.2.3 and 5.2.4 display three examples of text explanation each. For examples *A* and *B* in Figure 5.2.3 the explainable model manages to capture a majority of the words that the domain expert highlighted, and also the underlying model predicted the correct findings. However, the model predicted atelectasis for example *C* but failed to predict the thoracic finding nodule. Interestingly enough, the explainable words from the explainable model missed to include *nodule* that the domain expert had highlighted in the radiology report.

While studying the examples in Figure 5.2.4, it is notable that the length of the domain experts' explanation words varies. However, the size of the explainable model remains the same. One major limitation of the text explanations is that the number of words to output is a pre-defined hyperparameter of the explainable model. However, when collecting the language explanations from domain experts, they were allowed to vary the number of words that represent an explanation. In practice, this means that the explainable model always generates a fixed number of words, in this case, five words, while the domain experts' explanations may vary in size from one to eight words. Likely, the fixed size of outputted words harms the text similarity score. Therefore, we considered using the coefficients associated with each of the important vectors to aggregate the importance score of each word and rank the outputted words. Following, one can use this ranking to match the same number of words as the domain experts before computing the similarity. We hypothesize that this would improve our scores but choose not to follow this path. The reason for not doing so is that the information on the size of domain experts' important words would not be available in practice and would serve unfairly in the evaluations. Nevertheless, this is a limitation to our work, and future work is encouraged to study how to fairly vary the size of outputted words of the explainable models. Possibly, introducing a threshold on what words to output could help, but future work needs to explore how to determine such a threshold. In addition, it is worth noting that even the LIME python implementation takes the size of the number of explainable words to output as a pre-defined input parameter [48].

Table 5.2.4: Displays average text and image similarity for both underlying vision and language model, as well as for each type of perturbation. The averages derives from the similarity scores presented in 5.2.1.

	Average Text Similarity	Average Image Similarity
UNITER	0,101	0,173
VisualBERT	0,109	0,173
Simultaneous Perturbations	0,088	0,122
Separate Perturbations	0,104	0,192

Moreover, one interesting take from the experiments presented in Table 5.2.1 is the comparison between separate and simultaneous perturbations for training the explainable models. One of the main questions that arose when we designed the explainable model was how to generate the multi-modal explanations. Either one could train separate simple models with their perturbations and then combine the output explanations of the two, or one could train a joint explainable model that trains on perturbations from both modalities. Taking a closer look at the similarity scores in Table 5.2.4, one can distinguish that there seems to be a trend that the explainable models with separate perturbations generally receive higher similarity scores than simultaneous perturbation models. For the models with separate perturbations, each surrogate model for the image and text modality could be tuned independently. In contrast, there is an interaction between the hyperparameters from each modality for the explainable model with simultaneous perturbations. Please advise Section 3.2.4 for more details on hyperparameters. We hypothesize that the interaction of hyperparameters of perturbing both modalities simultaneously makes the fine-tuning more complex and challenging to optimize. In addition, it can also be that one of the modalities contains more signal than the other making the dynamics between the simultaneous perturbation more complex, while the annotation from the domain experts does not capture such relationships. The domain experts annotated the modalities separately, which might also favor the separate perturbation technique. For future work, it would be interesting to investigate the strength of the signal from each modality for explaining a prediction.

In contrast, when studying the similarity scores in Table 5.2.1 between the UNITER and the VisualBERT model, we could not distinguish any general trend. Instead, the similarity scores from explaining predictions across the two underlying vision and language models seem indifferent. Hence, the results suggest that there is no difference between the success of explanations between using an UNITER or a VisualBERT underlying vision and language model for this particular case study.

Similarly, another interesting discussion is on ways to perturb features in the input data. Local surrogate explainability techniques like LIME and SHAP build one central assumption: it is possible to turn on and off features for model predictions [47] [32]. While this is an effective way to evaluate feature contributions, it could be worth considering whether such inactivation of features is logical for the ML task. While turning on and off the absence of words and measuring their impact is a common application, there is limited work on how to inactivate the visual boxes. Our approach is to inactivate a visual box by replacing all the elements in the associated visual embedding vector with zeros. As for now, it is difficult to distinguish how this choice impacts the explainable models. Nevertheless, it is worth questioning whether the vision and language models are designed to input a visual embedding vector consisting of all zeros. If we were working with pixels, replacing values with zeros would mean that we would get black-colored regions, but for visual embedding, the representation is less intuitive. One other approach could be to compute the mean and standard deviations of values in a visual embedding vector and then distort the elements a multiple of standard deviations away from the mean. Another approach could be to randomize the elements of a visual embedding vector. However, future work is encouraged to explore different ways to represent the inactivation of features for explaining vision and language predictions.

Furthermore, a limitation to be recognized is the visual feature extraction network used in the study. The visual features consist, for every image, of 36 bounding boxes with an associated visual embedding vector for each box. These visual features were extracted using the bottom up and top down attention network Detectron2 [54]. The feature extraction network is pre-trained on the visual genome data set, a general non-medical specific data set [54]. Since the extracted bounding boxes directly impact the choice of possible explanations to choose from, it should be fair to assume that this impacts the success of the explainable model. Namely, the explainable model perturbs the extracted boxes, and from that, it finds what bonding box seems to best explain a prediction. Therefore, we hypothesize that if a medical pre-trained, or even better, a chest X-ray pre-trained network would have been used for the visual box extraction, the explanations would better resemble domain expertise. Such a network could extract boxes that better represent relevant image features. For instance, if such a network could detect an area with high opacity, that could be a reasonable explanation for the thoracic finding of atelectasis [14].

However, to the best of our knowledge, such a network does not yet seem to have been developed. Nevertheless, for instance, there is one chest X-ray pre-trained network, ChexNet [45]. The problem is that the visual features produced by this network do not have meaningfully associated locations in the radio-graph image [45]. The network does extract visual embeddings that could be used for training and inference for vision and language models [36]. However, our explainable model requires that the visual features have a location attached to it. Otherwise, if a visual feature turned out to be especially important for a prediction, the model could not meaningfully show the explanation in the image. We hypothesize that the absence of a medical pre-train network with logical positions associated is due to a lack of data in the field [36]. However, given a large enough corpus of annotated objects in chest X-ray images with positions, such a visual extraction network could be trained and further be used to improve our explainable model.

In general, it is interesting to note that even if the explainable model does not successfully match domain expertise, it can still be helpful for data scientists and ML engineers to improve their models. For instance, if annotations from domain experts suggest that the signal is centered around the chest area for predicting these findings, while the explainable model suggests that the model finds the signal in the corners of the image, this can be helpful feedback. If this is the case, possibly guiding the model towards decisive regions could help to improve model generalization. Overall, comparing the explainable model outcome with domain expertise helps to get insights into whether what the model seems to think are contributing features, do match with what domain expertise view as contributing features. For instance, such insights could suggest to denoise training and inference data for better generalization capabilities or give guidance on hyperparameter-tuning.

Chapter 6

Concluding Remarks

This thesis work has explored the feasibility of resembling domain expertise when using the local surrogate explainability technique in combination with an underlying vision and language model to generate multi-modal vision and language explanations. A case study was carried out to explain predictions made by vision and language models trained to predict thoracic findings from radio-graphs and radiology reports. The results indicate that the particular case study task of explaining thoracic findings is challenging as annotations from domain experts suggest that there is ambiguity on what is the ground truth in terms of explanations. Nevertheless, the results indicate that the explainable model has to some extent, captured signals in explaining the predictions. The resulting similarity scores were relatively far from the similarity levels between domain experts, yet above a random baseline representing the lower bound. In addition, the explainable model looks to capture some useful feedback for model improvement. For instance, explanations could suggest pre-processing data and retraining to better guide the model toward thoracic regions. In addition, miss-matches of explanations from the explainable model and domain expertise could potentially serve as an explanation for false negatives. Additionally, the results suggests that the experiments with separate perturbations technique outperform that of simultaneous perturbations in terms of similarity to domain experts. More importantly, the thesis work has identified opportunities for ways to improve the multi-modal vision and language explainable model presented below:

- Perform a similar study, but on another clinical task where there is more unity on the ground truth in terms of explanations. Our results suggest that there is ambiguity on what the ground truth is in terms of explanations of contributing features for predicting thoracic findings. As a result, this seems to be a difficult problem for an explainable model to learn. Future work is encouraged to study other clinical ML tasks or in other fields that have a more precise definition of the most explainable features.

- Replace the feature extraction network with a medical pre-trained network. However, there are also two other opportunities for future work before that. Firstly, collect a large enough medical data set with annotated objects. Then, together with domain expertise, set up a data collection process to perform annotations of medical objects from radio-graph images. Secondly, use such a data set to train a feature extraction network with meaningful image locations associated with each embedding. For instance, one could train a *bottom-up attention top-down* network like Detectron2 [54] to produce visual boxes with visual embeddings and then use it to train a vision and language model. Given the use of such a network to extract visual features of the radio-graphs, we hypothesize that it can help to generate more relevant explanations for the ML task.

- Explore ways to fairly vary the number of important words generated by the explainable model. We hypothesize that if the number of outputted words could vary in size to better match the number of words from domain experts, that would improve the text similarity scores. However, our study left out this as we could not conclude a fair way of achieving this. Therefore, we suggest future work exploring ways to fairly vary the output size of important words.

- Investigate the strength of different modalities when explaining predictions. Our results indicate that separate perturbations outperform simultaneous with regard to resembling domain expertise. However, we conclude that the hyperparameter tuning of the simultaneous perturbations was more challenging to optimize and that there can be a different amount of signal between the two modalities. In future work, it would be highly interesting to explore ways to capture the strength of each modality when creating vision and language explanations from each modality.

- Finally, study ways to inactivate features and the impact on explanations. Prominent techniques like LIME and SHAP builds on one central premise: it is possible to inactivate features for making predictions [47] [32]. However, especially for the visual boxes, it is non-intuitive how to logically inactive the visual embedding. Our work did so by replacing the elements in the visual embedding with zeros. However, future work is encouraged to investigate other ways to inactivate features and measure their impact on resulting explanations.

Bibliography

[1] Abacha, Asma Ben, Hasan, Sadid A, Datla, Vivek V, Liu, Joey, Demner-Fushman, Dina, and Müller, Henning. “VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019.” In: *CLEF (Working Notes)* 2 (2019).

[2] Adadi, Amina and Berrada, Mohammed. “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”. In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[3] Alvi, Jaweriah. “Explainable Multimodal Fusion”. In: *DiVA* 46 (2021).

[4] Amilia Werner. *Radiology domain expert interview*. 2022-05-16.

[5] Asan, Onur, Bayrak, Alparslan Emrah, Choudhury, Avishek, et al. “Artificial intelligence and human trust in healthcare: focus on clinicians”. In: *Journal of medical Internet research* 22.6 (2020), e15154. DOI: 10.2196/15154.

[6] Bibal, Adrien and Frénay, Benoıt. “Interpretability of machine learning models and representations: an introduction.” In: *ESANN.* (2016).

[7] Biswal, Siddharth, Zhuang, Peiye, Pyrros, Ayis, Siddiqui, Nasir, Koyejo, Sanmi, and Sun, Jimeng. “EMIXER: End-to-end Multimodal X-ray Generation via Self-supervision”. In: *arXiv preprint arXiv:2007.05597* (2020). DOI: 10.48550/arXiv.2007.05597.

[8] Bussone, Adrian, Stumpf, Simone, and O’Sullivan, Dympna. “The role of explanations on trust and reliance in clinical decision support systems”. In: *2015 international conference on healthcare informatics*. IEEE. 2015, pp. 160–169. DOI: 10.1109/ICHI.2015.26.

[9] Chen, Yen-Chun, Li, Linjie, Yu, Licheng, Kholy, Ahmed El, Ahmed, Faisal, Gan, Zhe, Cheng, Yu, and Liu, Jingjing. *UNITER: UNiversal Image-TExt Representation Learning*. 2019. DOI: `10.48550/arXiv.1909.11740`. arXiv: `1909.11740 [cs.CV]`.

[10] Cutillo, Christine M, Sharma, Karlie R, Foschini, Luca, Kundu, Shinjini, Mackintosh, Maxine, and Mandl, Kenneth D. “Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–5.

[11] Demner-Fushman, Dina, Kohli, Marc D, Rosenman, Marc B, Shooshan, Sonya E, Rodriguez, Laritza, Antani, Sameer, Thoma, George R, and McDonald, Clement J. “Preparing a collection of radiology examinations for distribution and retrieval”. In: *Journal of the American Medical Informatics Association* 23.2 (2016), pp. 304–310.

[12] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. In: *Proceedings of the 2019 Conference of the North* (2019). DOI: `10.18653/v1/n19-1423`. URL: `http://dx.doi.org/10.18653/v1/N19-1423`.

[13] Doshi-Velez, Finale and Kim, Been. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. DOI: `10.48550/ARXIV.1702.08608`. URL: `https://arxiv.org/abs/1702.08608`.

[14] Erica Fred. *Radiology domain expert interview*. 2022-05-03, 2022-05-06.

[15] GDPR, EU. *Art. 15 GDPR Right of access by the data subject*. 2018. URL: `https://gdpr.eu/article-15-right-of-access/` (visited on 10/23/2021).

[16] Guidotti, Riccardo, Monreale, Anna, Ruggieri, Salvatore, Turini, Franco, Giannotti, Fosca, and Pedreschi, Dino. “A survey of methods for explaining black box models”. In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42. DOI: `10.1145/3236009`.

[17] Guillaumin, Matthieu, Verbeek, Jakob, and Schmid, Cordelia. “Multimodal semi-supervised learning for image classification”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, pp. 902–909. DOI: `10.1109/CVPR.2010.5540120`.

[18] Håkansson, Anne. “Portal of Research Methods and Methodologies for Research Projects and Degree Projects”. In: *Proceedings of the International Conference on Frontiers in Education : Computer Science and Computer Engineering FECS'13*. QC 20131210. CSREA Press U.S.A, 2013, pp. 67–73. ISBN: 1-60132-243-7. URL: `http://www.world-academy-of-science.org/worldcomp13/ws`.

[19] Hind, Michael, Wei, Dennis, Campbell, Murray, Codella, Noel CF, Dhurandhar, Amit, Mojsilović, Aleksandra, Natesan Ramamurthy, Karthikeyan, and Varshney, Kush R. “TED: Teaching AI to explain its decisions”. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 123–129. DOI: `10.1145/3306618.3314273`.

[20] Huang, Shih-Cheng, Pareek, Anuj, Seyyedi, Saeed, Banerjee, Imon, and Lungren, Matthew P. “Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines”. In: *npj Digital Medicine* 3.1 (Oct. 2020), p. 136. ISSN: 2398-6352. DOI: `10.1038/s41746-020-00341-z`. URL: `https://doi.org/10.1038/s41746-020-00341-z`.

[21] Irvin, Jeremy, Rajpurkar, Pranav, Ko, Michael, Yu, Yifan, Ciurea-Ilcus, Silviana, Chute, Chris, Marklund, Henrik, Haghgoo, Behzad, Ball, Robyn, Shpanskaya, Katie, et al. “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597. DOI: `10.1609/aaai.v33i01.3301590`.

[22] Johnson, Alistair, Pollard, Tom, Mark, Roger, Berkowitz, Seth, and Horng, Steven. “Mimic-cxr database”. In: *PhysioNet10* 13026 (2019), C2JT1Q.

[23] Joshi, Gargi, Walambe, Rahee, and Kotecha, Ketan. “A Review on Explainability in Multimodal Deep Neural Nets”. In: *IEEE Access* 9 (2021), pp. 59800–59821. DOI: `10.1109/ACCESS.2021.3070212`.

[24] Krishna, Ranjay, Zhu, Yuke, Groth, Oliver, Johnson, Justin, Hata, Kenji, Kravitz, Joshua, Chen, Stephanie, Kalantidis, Yannis, Li, Li-Jia, Shamma, David A, et al. “Visual genome: Connecting language and vision using crowdsourced dense image annotations”. In: *International journal of computer vision* 123.1 (2017), pp. 32–73.

[25] Leskovec, Jure, Rajaraman, Anand, and Ullman, Jeffrey David. *Mining of massive data sets*. Cambridge university press, 2020.

[26] Li, Liunian Harold, Yatskar, Mark, Yin, Da, Hsieh, Cho-Jui, and Chang, Kai-Wei. *VisualBERT: A Simple and Performant Baseline for Vision and Language*. 2019. DOI: 10.48550/arXiv.1908.03557. arXiv: 1908.03557 [cs.CV].

[27] Li, Yikuan, Wang, Hanyin, and Luo, Yuan. "A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 1999–2004. DOI: 10.48550/arXiv.2009.01523.

[28] Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C Lawrence. "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[29] Litjens, Geert, Kooi, Thijs, Bejnordi, Babak Ehteshami, Setio, Arnaud Arindra Adiyoso, Ciompi, Francesco, Ghafoorian, Mohsen, van der Laak, Jeroen A.W.M., van Ginneken, Bram, and Sánchez, Clara I. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2017.07.005. URL: https://www.sciencedirect.com/science/article/pii/S1361841517301135.

[30] Liu, Guanxiong, Hsu, Tzu-Ming Harry, McDermott, Matthew, Boag, Willie, Weng, Wei-Hung, Szolovits, Peter, and Ghassemi, Marzyeh. "Clinically Accurate Chest X-Ray Report Generation". In: *Proceedings of the 4th Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens. Vol. 106. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 249–269. URL: https://proceedings.mlr.press/v106/liu19a.html.

[31] Lu, Jiasen, Batra, Dhruv, Parikh, Devi, and Lee, Stefan. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. 2019. DOI: 10.48550/arXiv.1908.02265. arXiv: 1908.02265 [cs.CV].

[32] Lundberg, Scott and Lee, Su-In. *A Unified Approach to Interpreting Model Predictions*. 2017. DOI: 10.48550/arXiv.1705.07874. arXiv: 1705.07874 [cs.AI].

[33] Madeleine Broberg. *Radiology domain expert interview*. 2022-05-08.

[34] Marcinkevičs, Ričards and Vogt, Julia E. *Interpretability and Explainability: A Machine Learning Zoo Mini-tour*. 2020. DOI: 10.48550/ARXIV.2012.01805. URL: https://arxiv.org/abs/2012.01805.

[35] Molnar, Christoph. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2019, pp. 17–42.

[36] Monajatipoor, Masoud, Rouhsedaghat, Mozhdeh, Li, Liunian Harold, Chien, Aichi, Kuo, C-C Jay, Scalzo, Fabien, and Chang, Kai-Wei. "BERTHop: An Effective Vision-and-Language Model for Chest X-ray Disease Diagnosis". In: *arXiv preprint arXiv:2108.04938* (2021). DOI: 10.48550/arXiv.2108.04938.

[37] Mourby, Miranda, Ó Cathaoir, Katharina, and Collin, Catherine Bjerre. "Transparency of machine-learning in healthcare: The GDPR European health law". In: *Computer Law Security Review* 43 (2021), p. 105611. ISSN: 0267-3649. DOI: https://doi.org/10.1016/j.clsr.2021.105611. URL: https://www.sciencedirect.com/science/article/pii/S0267364921000844.

[38] Muhammad, Ghulam, Alshehri, Fatima, Karray, Fakhri, Saddik, Abdulmotaleb El, Alsulaiman, Mansour, and Falk, Tiago H. "A comprehensive survey on multimodal medical signals fusion for smart healthcare systems". In: *Information Fusion* 76 (2021), pp. 355–375. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2021.06.007. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001330.

[39] Ngiam, Jiquan, Khosla, Aditya, Kim, Mingyu, Nam, Juhan, Lee, Honglak, and Ng, Andrew Y. "Multimodal Deep Learning". In: *ICML*. 2011, pp. 689–696. URL: https://icml.cc/2011/papers/399_icmlpaper.pdf.

[40] Nourani, Mahsan, King, Joanie, and Ragan, Eric. "The role of domain expertise in user trust and the impact of first impressions with intelligent systems". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 8. 2020, pp. 112–121.

[41] Olivia Choudhury, Andy Schuetz and Hsieh, Michael. "Building Scalable Machine Learning Pipelines for Multimodal Health Data on AWS". In: *AWS for industries* (2021), p. 1.

[42] Olivia Choudhury, Andy Schuetz and Hsieh, Michael. “Training Machine Learning Models on Multimodal Health Data with Amazon SageMaker”. In: *AWS for industries* (2021), p. 1.

[43] Parcalabescu, Letitia, Trost, Nils, and Frank, Anette. “What is Multimodality?” In: (2021). arXiv: `2103.06304 [cs.AI]`.

[44] Peng, Yifan, Wang, Xiaosong, Lu, Le, Bagheri, Mohammadhadi, Summers, Ronald, and Lu, Zhiyong. “NegBio: a high-performance tool for negation and uncertainty detection in radiology reports”. In: *AMIA Summits on Translational Science Proceedings* 2018 (2018), p. 188.

[45] Rajpurkar, Pranav, Irvin, Jeremy, Zhu, Kaylie, Yang, Brandon, Mehta, Hershel, Duan, Tony, Ding, Daisy, Bagul, Aarti, Langlotz, Curtis, Shpanskaya, Katie, et al. “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning”. In: *arXiv preprint arXiv:1711.05225* (2017). DOI: `10.48550/arXiv.1711.05225`.

[46] Rezatofighi, Hamid, Tsoi, Nathan, Gwak, JunYoung, Sadeghian, Amir, Reid, Ian, and Savarese, Silvio. “Generalized intersection over union: A metric and a loss for bounding box regression”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 658–666.

[47] Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. “”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, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144.

[48] Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. “”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, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144.

[49] Rudin, Cynthia. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[50] Services, Amazon Web. *Amazon SageMaker Ground Truth*. 2020. URL: `https://aws.amazon.com/sagemaker/data-labeling/` (visited on 05/05/2022).

[51] SHARMA, YUKTI, VERMA, ABHINAV, RAO, KRISSTINA, and ELURI, VIVEK. “Reasonable Explainability ‘for Regulating AI in Health”. In: (2020).

[52] Tan, Hao and Bansal, Mohit. “LXMERT: Learning Cross-Modality Encoder Representations from Transformers”. 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)* (2019). DOI: `10.18653/v1/d19-1514`. URL: `http://dx.doi.org/10.18653/v1/D19-1514`.

[53] Wang, Xiaosong, Peng, Yifan, Lu, Le, Lu, Zhiyong, and Summers, Ronald M. “TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2018). DOI: `10.1109/cvpr.2018.00943`. URL: `http://dx.doi.org/10.1109/CVPR.2018.00943`.

[54] Wu, Yuxin, Kirillov, Alexander, Massa, Francisco, Lo, Wan-Yen, and Girshick, Ross. *Detectron2*. `https://github.com/facebookresearch/detectron2`. 2019.

[55] Xie, Ning, Lai, Farley, Doran, Derek, and Kadav, Asim. *Visual Entailment: A Novel Task for Fine-Grained Image Understanding*. 2019. DOI: `10.48550/arXiv.1901.06706`. arXiv: `1901.06706 [cs.CV]`.

[56] Zhou, Jianlong, Gandomi, Amir H, Chen, Fang, and Holzinger, Andreas. “Evaluating the quality of machine learning explanations: A survey on methods and metrics”. In: *Electronics* 10.5 (2021), p. 593. DOI: `10.3390/electronics10050593`.