

Class Based Strategies for Understanding Neural Networks

by

Devinder Kumar

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2020

© Devinder Kumar 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Mehran Ebrahimi
Associate Professor,
Faculty of Science, Ontario Tech University

Supervisor(s): Alexander Wong
Associate Professor,
Systems Design Engineering, University of Waterloo

Graham W. Taylor
Associate Professor,
Systems Design Engineering, University of Waterloo

Internal Member: John Zelek
Professor,
Systems Design Engineering, University of Waterloo

Internal Member: Katharine Scott
Assistant Professor,
Systems Design Engineering, University of Waterloo

Internal-External Member: Mark Crowley
Assistant Professor,
Electrical & Computer Engineering, University of Waterloo

This thesis consists of material all of which I authored or co-authored. Please see the statement of contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Content from 6 papers are used in this thesis. I was the co-author with major contributions on designing the interpretability methods:

A. Ziletti, D. Kumar, M. Scheffler, L. M. Ghiringhelli, “Insightful classification of crystal structures using deep learning”, Nature Communications, Vol. 9(1), p.2775, 2018

This paper is incorporated in Chapter 3 of this thesis.

D. Kumar, V. Menkovski, G. W. Taylor, A. Wong, “Understanding anatomy classification through attentive response maps”, IEEE 15th International Symposium on Biomedical Imaging (ISBI), 2018

This paper is incorporated in Chapter 3 of this thesis.

D. Kumar, V. Sankar, D. A. Clausi, G. W. Taylor, A. Wong, “SISC: End-to-end Interpretable Discovery Radiomics-Driven Lung Cancer Prediction via Stacked Interpretable Sequencing Cells”, IEEE Access Journal, 2019

This paper is incorporated in Chapter 4 of this thesis.

D. Kumar, A. Wong, G. W. Taylor, “Explaining the unexplained: A class-enhanced attentive response (clear) approach to understanding deep neural networks”, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W), 2017

This paper is incorporated in Chapter 5 of this thesis.

D. Kumar, G. W. Taylor, A. Wong, “Discovery Radiomics With CLEAR-DR: Interpretable Computer Aided Diagnosis of Diabetic Retinopathy, IEEE Access Journal, 2019

This paper is incorporated in Chapter 5 of this thesis.

Best Paper Award at Transparent and Interpretable Workshop at 30th Neural Information Processing (NIPS), 2017

D. Kumar, I. Ben Daya, K. Vats, J. Feng, G. W. Taylor, A. Wong, “Beyond Explainability: Leveraging Interpretability for Improved Adversarial Learning”, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W), 2019

This paper is incorporated in Chapter 6 of this thesis.

Abstract

One of the main challenges for broad adoption of deep learning based models such as Convolutional Neural Networks (CNN), is the lack of understanding of their decisions. In many applications, a simpler, less capable model that can be easily understood is favorable to a black-box model that has superior performance. Hence, it is paramount to have a mechanism for deep learning models such as deep neural networks to explain their decisions.

To resolve this explainability issue, in this thesis the main goal is to explore and develop new class-enhanced support strategies for visualizing and understanding the decision-making process of deep neural networks. In particular, we take a three level approach to provide a holistic framework for explaining deep neural networks predictions.

In the first stage (Chapter 3), we first try to answer the question: based on what information neural networks make their decision and how it relates to a human expert's domain knowledge? To this end, we propose to introduce attentive response maps. The attentive response maps are able to show: 1) The locations in the input image that are contributing to decision making and 2) the level of dominance of such locations. Through various experiments we elaborate how through attention response maps, we are able to visualize the decision making process of deep neural networks and show where the neural networks were able to or failed to use landmark features similar to a human expert's domain knowledge.

In second stage (Chapter 4), we propose a novel end-to-end design architecture for obtaining end-to-end explanations through attentive response maps. Towards the end of this stage, we explore some of the shortcomings of the attentive response maps in failing to explain some of the complex scenarios.

In the last stage, (Chapter 5), we try to overcome the shortcomings of the binary attention maps introduced in the first stage. Towards this goal, a **CL**ass-**E**nhanced **A**ttentive **R**esponse (CLEAR) approach was introduced to visualize and understand the decisions made by deep neural networks (DNNs) given a specific input based on spatial support. CLEAR facilitates the visualization of attentive regions and levels of interest of DNNs during the decision-making process. It also enables the visualization of the most dominant classes associated with these attentive regions of interest. As such, CLEAR can mitigate some of the shortcomings of attention response maps-based methods associated with decision ambiguity, and allows for better insights into the decision-making process of DNNs.

In the last Chapter of this thesis (Chapter 6), we draw conclusions about the introduced class based explanation strategies and discuss some interesting future directions, including

a formulation for class based global explanation that can be used for discovering and explaining the concepts identified by trained deep neural networks using human attribute priors.

Acknowledgements

First, I would like to earnestly thank my two co-supervisors, professors, Alexander Wong and Graham W. Taylor for their immense support without which this would not have been possible. Thank you for your guidance and for teaching me things related to academics and beyond. I will always be grateful for your support, and the help that allowed me to overcome various challenges that I faced during my PhD.

I would like to thank Prof. Mehran Ebrahimi from Ontario Tech University for accepting to be my external examiner and for his time to review this thesis. I would also like to sincerely thank my Ph.D. committee members Prof. Mark Crowley, from Electrical and Computer Engineering, Prof. John Zelek, and Prof. Katharine Scott from Systems Design Engineering department for their time and commitment.

I am grateful to all the members of VIP lab at UWaterloo for supporting me and for the fruitful and sometimes over the top discussions in the lab. To my friends, Nikhil, Mrigank & Manvit, thanks for helping me think through a lot of things in my personal life and for your continuous support.

At the end, I would like to thank my family. My parents for supporting me through every decision of mine and letting me do whatever I wanted to do without any restrictions. Immense credit goes to both of them: my mother Savita Yadav, who is the strongest person I know and has always acted as a moral guiding light throughout my life. My late father Narender Kumar Yadav, who unfortunately passed away during my PhD, from whom I get the uncanny and humorous side of mine. My brother, Ravi for guiding me through difficult decisions and stepping in when things got out of hand and lastly to my new born niece Kavya, who came into our family after a sad period and quickly became the beacon of new hope and made our family full of smiles & laughter once again.

Dedication

This is dedicated to my parents:
Mom & Dad (will always miss you!),

Table of Contents

List of Tables	xiii
List of Figures	xv
1 Introduction	1
1.1 Problem Definition and Challenges	3
1.2 Objectives	4
1.3 Contributions	5
1.4 Thesis Structure	6
2 Background & Related Work	8
2.1 Explainable AI	9
2.1.1 Modes Of Explanation	10
2.1.2 Types Of Explanation	10
2.2 Convolutional Neural Networks	11
2.2.1 Convolutional Layer	11
2.2.2 Activation Function	12
2.2.3 Pooling Layer	13
2.2.4 Batch-Normalization Layer	15
2.2.5 Dropout	15
2.2.6 Fully Connected (FC) Layer	15

2.2.7	CNN Architecture	16
2.3	Types Of Explanation Methods	17
2.4	Local Explanations via Instance-Based Methods	17
2.4.1	Visualization via Propagation Based Methods	18
2.4.2	Text Based Local Explanations	22
2.5	Global Explanations Based Methods	22
2.6	Summary	24
3	Domain Knowledge Based Architecture Design Using Attentive Response Maps	25
3.0.1	Article Details	26
3.0.2	Context	26
3.0.3	Contribution	27
3.1	Introduction	28
3.2	Methodology - Attention Maps Formation	28
3.2.1	Formulation	29
3.3	Experiments- Crystal Structures Classification	30
3.3.1	Motivation	30
3.3.2	Dataset Explanation	33
3.3.3	Experiment Design	33
3.3.4	Results	34
3.4	Discussions	44
3.5	Experiments - Human Anatomy Classification	45
3.5.1	Dataset Explanation	49
3.5.2	Experiment Design & Results	49
3.6	Discussions	51
3.7	Summary	54

4	End To End Interpretable Architecture Design	57
4.0.1	Article Details	58
4.0.2	Context	58
4.0.3	Contribution	58
4.1	Introduction	59
4.1.1	Motivation: End-to-End Deep Radiomics Sequencers	59
4.2	Methodology	60
4.2.1	Dataset Explanation- LIDC	60
4.2.2	Interpretable Sequencing Cells	62
4.2.3	Interpretability Through Critical Response Maps	64
4.3	Experiments And Results	67
4.3.1	Experimental Setup	68
4.3.2	Cancer Prediction Performance	69
4.3.3	Discussion	71
4.4	Summary	73
5	Class-Enhanced Attentive Response (CLEAR)	76
5.0.1	Details Of Articles	77
5.0.2	Context	77
5.0.3	Contributions	78
5.0.4	Recent Developments	78
5.1	Introduction	79
5.1.1	Methodology Overview - Class Enhanced Attentive Response (CLEAR)	80
5.1.2	Formulation	81
5.2	Experiments- Generic Image datasets	84
5.2.1	Dataset Explanations - MNIST & SVHN	84
5.2.2	Experiments Design	85
5.2.3	Qualitative Results	85

5.2.4	SVHN	87
5.2.5	Quantitative Results	87
5.2.6	Discussions	91
5.3	Experiments - Diabetic Retinopathy	91
5.3.1	Diabetic Retinopathy: Motivation	91
5.3.2	Diabetic Retinopathy Dataset	94
5.3.3	Experimental Setup: Training A Discovery Radiomics Sequencer	95
5.3.4	Qualitative Experiments: CLEAR for Interpretable CAD	96
5.3.5	Discussions	96
5.4	Summary	98
6	Conclusions & Future Work	100
6.1	Thesis Contribution Highlights	101
6.1.1	Limitations	102
6.2	Future Work	103
6.2.1	Human Attributes Prior Based Concept Explanations	103
6.2.2	Multi-modal Data Explanations	107
6.2.3	Beyond Explainability	107
	References	110
	APPENDICES	125
A	Interpretabel Crystal Structure Classification	126
A.1	How To Represent A Material	126
A.2	Crystal Formation Formulation	128

List of Tables

3.1	Architecture of the convolutional neural network used in this work.	34
3.2	Accuracy in identifying the correct (most similar) crystal class in the presence of defects.	39
3.3	Results: Accuracy in percent for three different networks trained on the ImageClef 2009 annotation task	49
4.1	Number of samples for corresponding malignancy scores, for three different datasets. The datasets were created based on how the radiologist malignancy rating 3 was treated i.e., Ignored (I), treated as benign (B), or treated as malignant (M).	68
4.3	Performance comparison between tested cancer prediction methods for the ignored (I) dataset. Best results are highlighted in bold	70
4.2	Dataset distribution for the three different dataset configuration obtained from the LIDC-IDRI dataset before and after data augmentation (as described in Section 4.3.1).	70
4.4	Performance comparison between tested cancer prediction methods for the benign (B) dataset. Best results are highlighted in bold	71
4.5	Performance comparison between tested cancer prediction methods for the malignant (M) dataset. Best results are highlighted in bold	71
5.1	Architecture of our CNN used for MNIST Classification.	85
5.2	Architecture of our CNN used for SVHN Classification.	86

5.3 Evaluation to re-validate the effectiveness and contribution of identified strong features on accuracy.	87
---	----

List of Figures

1.1	Examples of two different scenarios where model explainability is crucial from both spatial and temporal perspective.	2
1.2	Examples of handwritten digits from MNIST are shown to prove shortcomings of binary heatmaps.	4
2.1	Example of max-pooling using a 2x2 stride.	13
2.2	Example of Unpooling using a 2x2 stride.	14
2.3	Example of CNN architecture design.	16
2.4	Different methods of back propagating through ReLU non-linearity along with the formal formulation for propagating an output back through a ReLU unit in layer l . Here, f is the feature map in the forward pass and R is the reconstructed feature map in the backward pass.	19
2.5	Different methods for creating visualization maps [40]. The figure shows the visualization maps and original images. All images were classified correctly.	21
2.6	Class activation map example.	22
3.1	The model workflow of automatic crystal-structure classification.	32
3.2	Illustration of two dimensional diffraction fingerprint formation.	35
3.3	Schematic representation of the convolutional neural network (ConvNet) used for crystals classification	37
3.4	Neural network predictions on structural transitions.	41
3.5	Visualizing the convolutional neural network (ConvNet) attentive response maps for crystal classification.	43

3.6	overview of the proposed visualization framework for understanding and visualizing human anatomy prediction.	46
3.7	Architecture of three different CNNs with different capacities that were used in this human anatomy classification study.	47
3.8	Correspondence between anatomical descriptions found in the literature that are used by human experts.	48
3.9	Attentive response maps overlaid on the original images from the last conv layer of the deeper network with no data augmentation for foot and hand class.	50
3.10	Focus area of the top 5 attentive response maps from the top 5 most activated units from last conv layer of the shallow network.	50
3.11	Illustration of the effectiveness of the attentive response maps in highlighting key medically relevant landmarks	52
3.12	Ablation study to test the effectiveness of attentive response maps	53
3.13	Individual attentive response maps for top 9 activated units from the last conv layer of the deeper network with augmentation for the hand class example	55
4.1	Overview of the proposed deep stacked interpretable sequencing cell (SISC) architecture used as the radiomic sequencer within a discovery radiomics framework	61
4.2	Overview of the end-to-end interpretable discovery radiomics-driven framework for lung cancer prediction.	63
4.3	Example critical response maps for malignant cases.	65
4.4	Example critical response maps for benign cases.	66
4.5	Receiver operating curve (ROC) for the ignored (I) dataset for 10 different cross validation runs.	72
4.6	Receiver operating curve (ROC) for the benign (B) dataset for 10 different cross validation runs.	72
4.7	Receiver operating curve (ROC) for the malignant (M) dataset for 10 different cross validation runs.	73
4.8	Comparing the best ROC curves for the three different datasets: ignored (I), benign (B) and malignant (M).	74

4.9	Accuracy, sensitivity, and specificity for the “ignored” (I) dataset for three different training sample sizes.	75
5.1	Examples of handwritten digits from MNIST dataset to show the failure cases for binary heatmaps	80
5.2	The procedure for generating CLass-Enhanced Attentive Response (CLEAR) maps.	82
5.3	CLEAR maps example images from the MNIST dataset.	88
5.4	CLEAR maps example images from the SVHN dataset.	89
5.5	Examples of handwritten digits from MNIST are shown, along with: 1) the decision made by the CNN, 2) heatmaps used in existing visualization methods, 3) the proposed CLass-Enhanced Attentive Response (CLEAR) maps	90
5.6	Three different scenarios for grading diabetic retinopathy: 1) without CAD, 2) CAD system without interpretability, 3) interpretable CAD via CLEAR.	92
5.7	Architecture of the convolutional radiomic sequencer used in the deep radiomic sequencer discovery process.	95
5.8	Abnormalities present for various cases of diabetic retinopathy	96
5.9	CLEAR maps examples for correctly (a) and mis-classified (b) examples for all diabetic retinopathy grades.	97
6.1	Examples of human defined attributes	104
6.2	Examples of class specific global explanation process	105
6.3	The illustration shows an example how explainable methods can be used to first obtain attention maps which in turn can be used with generative methods to obtain annotations or at-least as initial seed for segmentation algorithm.	108

List of Acronyms

ACE	Automated Concept Explanation
CAM	Class Activation Map
ARM	Attention Response Map
CAD	Computer Aided Diagnostics
CLEAR	Class Enhanced Attentive Response
CLEAR-DR	Class Enhanced Attentive Response for Diabetic Retinopathy
CNN	Convolutional Neural Network
CT	Computed Tomography
DNN	Deep Neural Network
FHI	Fritz Haber Institute
LIDC	Lung Image Database Consortium
NOMAD	Novel Materials Discovery
ROC	Receiver Operating Characteristic
SISC	Stacked Interpretable Sequencing Cells
SVHN	Street View House Numbers
TCAV	Testing with Concept Activation Vectors
XAI	Explainable AI

Chapter 1

Introduction

[That] transparency is “absolutely critical” for applications in science, but it is also important for many commercial applications. For example, in many countries, banks that deny a loan have a legal obligation to say why - something a deep-learning algorithm might not be able to do.

- Prof. Zoubin Ghahramani, University of Cambridge [\[83\]](#)

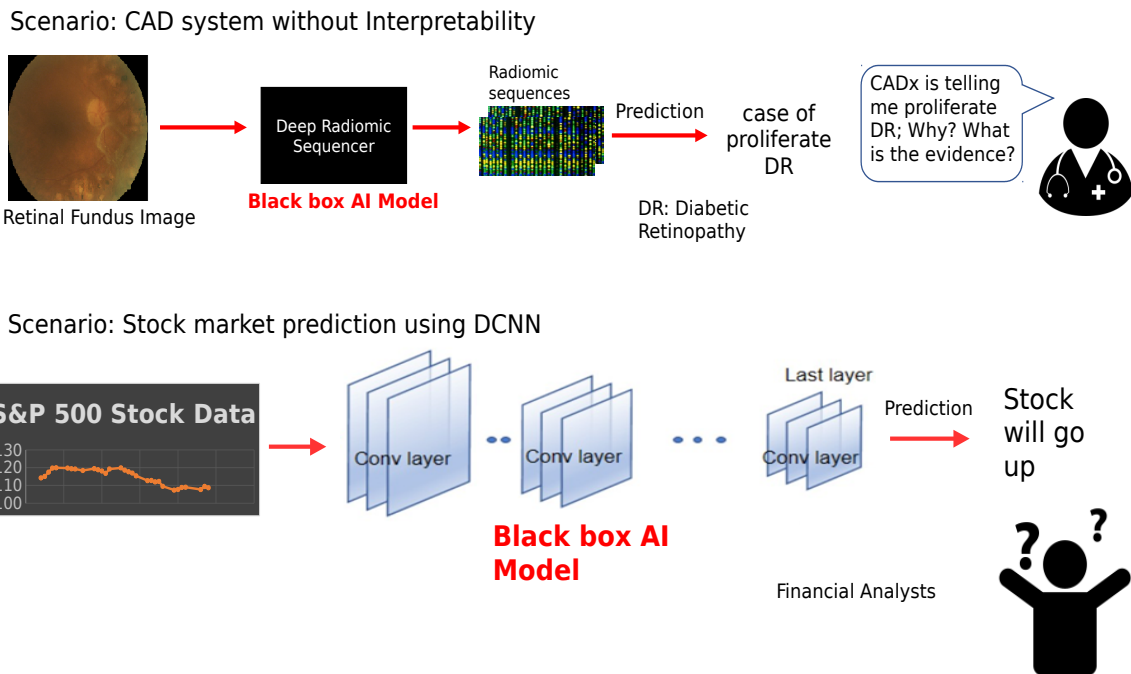


Figure 1.1: Examples of two different scenarios where model explainability is crucial from both spatial and temporal perspective. In both scenarios it is evident that using a black box model for prediction is neither sufficient nor helpful for the end user. It is thus highly necessary to use explainable model in such scenarios, where decision making is critically important.

In recent years, we have seen tremendous success in the field of artificial intelligence (AI). In particular, many of the recent advances have been related to one particular area of machine learning: deep neural networks (DNNs). DNNs have been shown to outperform previous machine learning techniques for a variety of tasks, such as fine-grained classification [14,132], self-driving cars [10], captioning and answering questions about images [2,77], and even defeating human champions at Go [107]. Although DNNs have demonstrated tremendous effectiveness at a wide range of tasks, when they fail, they often fail spectacularly, producing unexplainable and incoherent results that can leave one to wonder what caused the DNN to make such decisions. This lack of transparency and interpretability of DNNs during the decision-making process is largely due to their complex nature, where individual neural responses, unlike other interpretable decision-making processes such as decision trees, provide very little insight as to what is actually going on.

1.1 Problem Definition and Challenges

The lack of transparency in the decision-making process of DNNs is a significant bottleneck in their widespread adoption in industry, such as healthcare, defense, cyber-security, etc. as shown in Fig. 1.1, where the error tolerance is very low and the ability to interpret, understand, and trust decisions is critical. As such, a way to peer inside a DNN and see why it made a decision the way it did can have tremendous potential for pushing towards explainable AI, where a human expert gains the ability to understand, interpret and verify the decisions made.

Recently, a number of researchers have been exploring the understanding and interpretation of decisions made by DNNs, in particular by Convolutional Neural Networks (CNNs), through spatial information support, by asking the following question: *based on what information in the image is the CNN making a decision?* To tackle this question, much recent work has focused on understanding the decision-making process of networks in the spatial context by leveraging heatmaps that provide information about which areas of the image is used by the CNNs to make a particular decision. These approaches have produced some promising results in revealing what is important to a decision made by a CNN. More details regarding the relevant works are provided in Chapter 2. However, there are certain shortcomings relating to such approaches. Some of which are mentioned below:

- A common limitation with such heatmap-based visualization approaches to understanding the decision-making process of CNNs is that of **decision ambiguity**, where one can gain insight into **which** regions of interest are important for making decisions, but gives no insight as to **why** such regions of interest are important. An example of this is shown in Fig. 1.2. As a result, these methods leave the “thought process” of the CNN largely ambiguous.
- Another limitation of these methods is that they don’t always provide a co-relation or justification from the human point of view. That is to say that no information is provided as to how the identified region relates to human domain knowledge.
- Most of the methods can only be applied to the convolution layers i.e., there is a lack of end-to-end explainable deep networks. For reliable explanations, the explainability framework needs to be end-to-end from output space to input space.
- Most of the current methods also offer local level explanations i.e., they only provide information for a given instance and nothing is conveyed as to how it relates to global information in terms of what the deep network has learned or how it is making decisions at the dataset level.










Input		Output	Heatmap	Interpretation
	→  →	3 ✓		Focuses on right areas: Looks correct!
	→  →	2 ✗		Focuses on wrong part, curve might be two; but why not 3 or 5 or 6?
	→  →	3 ✗		Probably focuses on correct part, but why 3?

Figure 1.2: Examples of handwritten digits from MNIST are shown, along with: 1) the decision made by the CNN, 2) heatmaps used in existing visualization methods, and 3) what can be interpreted based on the heatmaps. While the heatmaps used in existing approaches show which information in the image works for (positive focus: hot regions) or against (negative focus: green), it is evident that the heatmaps are insufficient to fully interpret and explain the decision made by the CNN.

1.2 Objectives

In an attempt to mitigate the problem of decision ambiguity and other challenges mentioned above, we want to take a step towards “explaining the unexplained” with regards to the decision-making process of CNNs. Thus, the main objective of this thesis is to explore and develop new class-enhanced support strategies for visualizing and understanding the decision-making process of deep neural networks. Using the class-enhanced explainability framework, we want to address the challenges laid out in the previous section as follows:

- The first objective of the thesis to introduce a framework to address the question: “What kind of features are used to make predictions by neural networks and how do

they relates to human domain knowledge?”

- While addressing the above question, the thesis also aims to provide an explainability framework that offers end-to-end explanations from the output predictions to input space which is lacking in some of the current methods.
- Another major objective is to address the “unexplained” scenarios as shown in Fig 1.2.
- Lastly, as most of the trusted explainability methods are for local explanations whereas ideally for building trustworthy, reliable & satisfactory explanations, we need methods that can provide both global and local explanation.

1.3 Contributions

The described objectives in this research lead to the following contributions:

- The thesis proposes a framework based on attention response maps, that is able to first show a response for any given convolution layer in the input space and show how the identified information relates to human expert domain information. This thesis also provides experiments to show as to how through the use of attention response maps we can identify when in the training of deep neural networks do they start using human-expert identifiable landmarks.
- To turn the above process into an end-to-end procedure, we introduce a novel end-to-end interpretable prediction pipeline. The presented deep architecture comprised of stacked interpretable sequencing cells (SISC). The proposed SISC architecture is shown to outperform previous approaches while providing more insight into its decision making process. The SISC based architecture achieves state-of-the-art results and also offers prediction interpretability in the form of attention response maps in an end-to-end manner for binary class predictions.
- As mentioned in the above Section, there is a need to create a spatial context visualization method that goes beyond the current binary heatmap based approaches as they provide the attentive regions of interest only. Therefore in this thesis, we introduce the **CL**ass-**E**nhanced **A**ttentive **R**esponse (CLEAR) approach that goes beyond what existing heatmap-based approaches [5, 81, 135] can provide. The CLEAR approach allows for the visualization of not only the attentive regions of interest and

corresponding attentive levels of CNNs during the decision-making process, but also the corresponding dominant classes associated with these attentive regions of interest. As such, compared to heatmaps, CLEAR visualization is much more effective at conveying where and why certain regions of interest influence the decision-making process. We further demonstrate the effectiveness of the proposed CLEAR approach, both quantitatively and qualitatively, by conducting a number of experiments using different publicly available datasets across domains.

- As noted above, for a complete and holistic explainable framework we need both local and global explanations. Hence, in the last Chapter, the thesis also presents a formulation for class based global explanation that can be used for discovering and explaining the concepts identified by trained deep neural networks using human attribute priors.

1.4 Thesis Structure

The thesis is organized in six Chapters and is structured as follows:

- Chapter 2 introduces the necessary definitions and concepts related to explainable AI and DNNs. The Chapter also includes related work including various current approaches for visualization and understanding the decision making process of DNNs along with their mathematical definitions.
- Chapter 3 presents our proposed framework of attention maps to judge how DNNs incorporate domain knowledge during training and while making decisions.
- As most of the above methods, including attention maps are not end-to-end, the Chapter 4 forwards an end-to-end interpretable DNNs architecture framework for learning and explaining binary classification.
- Chapter 5 first presents the shortcomings of the binary attention maps as presented in the previous Chapter in providing holistic explaining. Motivated by this, we propose a new framework for understanding their decision making process to address the unexplainability problem from a spatial support level through a multi class enhanced attention maps approach.
- In the last Chapter 6, we highlight the key contributions and some of the limitations of the thesis research. The Chapter also presents some interesting future directions

along which the thesis research can be extended. As the previous Chapters are based on local explainability methods, in Chapter 6, we also provide a formulation for class based global explanations that can help create a holistic framework for thorough explanations.

Chapter 2

Background & Related Work

“If you had a very small neural network, you might be able to understand it. But once it becomes very large, and it has thousands of units per layer and maybe hundreds of layers, then it becomes quite un-understandable.”

- Prof. Tommi Jaakkola, MIT [\[91\]](#)

As noted in Chapter 1, contemporary deep AI systems offer a lot of promise and are quite effective in solving complex problems. However, their effectiveness is limited by the inability of these systems to explain their decisions and predictions in an understandable manner to their users (humans). Thus, to solve this problem for explainability in deep AI systems, in this Chapter we aim to first define what explainable AI is. We then cover the basic fundamental background related to a particular type of deep neural network architecture known as Convolutional Neural Networks (CNN). We define and cover CNNs, as this is the architecture type that has been used throughout the thesis to define and present our explainability formulations. After this, we elaborate on what are the different approaches that can be used to solve the inherent un-explainability problem in the modern deep AI systems, in particular for CNNs and the recent work done by AI researchers in this regard.

2.1 Explainable AI

Explainable AI (X-AI) refers to the systems, programs, algorithms, techniques or methods in the applications of artificial intelligence that are inherently able to provide explanations that can be understood by human domain experts. Here, understanding from the human domain experts perspective refers to the situation where the human expert is able to understand, appropriately trust, and effectively manage the evolving AI solution. Thus, for a AI system to be explainable, it needs to have the ability to explain its rationale, explain how and why it fails and how it will behave in uncertain future scenarios¹, all of this while maintaining a high level of performance. Hence, if an AI system contains all the above mentioned attributes and is acceptable to human domain experts (also referred to as end-users or users), we can refer to it as explainable.

Usually for humans to trust a system, it is imperative to show that the system uses the same domain knowledge or *features* as humans to arrive at a particular decision. To achieve this goal, the XAI can use different approaches or modes through which XAI is able to explain its rationale. A detailed classification of the modes of explanation is presented in the section below.

¹<https://www.darpa.mil/program/explainable-artificial-intelligence>

2.1.1 Modes Of Explanation

DARPA, the creator of the XAI program ², has classified the different modes of explanation in the following four different categories. These categories are:

1. **Analytic (didactic) statements:** In this mode, the explainable AI system uses natural language to describe the various elements and context that support a given prediction. For example, for identifying a zebra, the XAI can produce a statement as: There is an animal with stripes.
2. **Visualization:** The XAI systems in this mode directly highlight the portions of the raw data (e.g. group of pixels in an input image) to provide rationale for a given prediction. This mode usually allows the human expert to form their own subjective opinion for perceptual understanding. For example, for a given zebra image, it can highlight the stripe features on the back of a zebra.
3. **Cases:** In this mode, the system provides specific examples that are similar to the given input to support the prediction being made. For example, the person in a given image is smiling because it looks like these other images of people who are also smiling.
4. **Rejection of Alternative Choices:** In this case, the XAI systems produces argument against less preferred answers based on analytics, cases and data points. These methods are sometimes also referred to as counterfactual methods.

2.1.2 Types Of Explanation

Using the different modes, explanations can be provided at two different levels. These two levels of explanations are presented in detail below. A detailed discussion regarding the relevant recent work pertaining to their levels is discussed in Section 2.3.

- **Local Explanations:** In this type of explanation method, explanations and justifications are produced by the XAI system for a given instance i.e., for a particular input. Hence, these type of methods are sometimes also referred to as instance-based methods. For example, for a given input image of a cat that is correctly identified, rationale is produced as such to show what is the evidence in the given input based on which the prediction is being made.

²The author and Dr. Graham W. Taylor are part of this program and the research was funded partly with the associated funds.

- **Global Explanations:** In these type of XAI methods, explanations are produced on a class or dataset level. These methods usually aim to provide a rationale at the *global* level as to what the XAI system has learned about a class or dataset. For a given class level, these methods aim to answer, what is the interpretation of the learned XAI system with regards to that particular class. For example, for the class cat, what are the attributes a cat should have to be identified as a cat by a given model. On the other extreme, at the dataset or model level, these methods aim to explain what each of the learned components such as weights, other parameters, and structures entail to with respect to the given dataset.

2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are currently one of the most popular deep learning architectures, especially for image based visual recognition tasks. More formally, a CNN is a particular type of deep learning based feed forward neural network, which can take an input and assign prediction for a particular task for objects present in the input. The idea of architectural design for CNNs is loosely inspired by the information process that happens in the visual cortex of human brain. In the visual cortex, each individual neuron responds to stimuli only for a certain region of the visual field known as a receptive field. Hence, CNNs use the idea of a receptive field to identify and optimize its various parameters. While CNN architectures have many optimizable parameters that make them unique from each other, CNNs are typically composed of few key constituent layers and operations. The main constituents of the typical CNN architecture design are: convolutional layer, pooling layer, activation function, batch-normalization, dropout and fully connected layer. All of these individual components are explained below in detail with the full architecture design explained at the end of the Section.

2.2.1 Convolutional Layer

In discrete mathematics, the convolution of a 1D signal f with another signal g is defined as follows:

$$o[n] = f[n] * g[n] = \sum_{u=-\infty}^{\infty} f[u]g[n - u]. \quad (2.1)$$

Here, n & u are discrete variables. However, this definition can be extended for 2D

convolution, as required for pixel based convolutions:

$$o[m, n] = f[m, n] * g[m, n] = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} f[u, v]g[m - u, n - v]. \quad (2.2)$$

where m and n index the two dimensions of the original signal, and u, v index the second signal. However, when considering convolutions applied on images, the input is no longer $(-\infty, \infty)$. The input is bound to finite numbers, i.e., size of the input images and associated filters.

2.2.2 Activation Function

Activation functions are responsible for inducing non-linear properties in neural networks. These functions are responsible in deciding whether a neuron should be fired or not i.e. to see if the information from the previous neuron is relevant or not. Usually an activation function is applied to the linear output of every convolution to prevent the network from collapsing to a single layer. As shown below, we multiply the input I with the weight W of the neuron, add the bias b , and then apply a particular activation function σ (non-linear function). The transformed output O is then sent to the next layer.

$$O = \sigma(W * I + b). \quad (2.3)$$

Some of the popular activation functions are: sigmoid, tanh and ReLU (rectified linear unit). While traditionally in past, neural networks have used tanh, and sigmoid as activation functions, many of the state-of-the art networks are now using Rectified Linear Units (ReLU).

ReLU consists of a maximum operator, that takes an input x and if $x > 0$, lets the information/signal pass through it as seen in the equation below.

$$R(x) = \max(x, 0). \quad (2.4)$$

As the ReLU activation function, doesn't constrain the input into a set fixed upper bound unlike sigmoid and tanh, it helps in avoiding and rectifying the vanishing gradient problem.

2.2.3 Pooling Layer

Pooling layer is generally applied after convolution and non-linear activation layers or functions. Pooling layer in a CNN architecture is responsible for reducing the spatial size of the feature maps at a given layer. In the spatial reduction process, local regions (window size) of the previous layers are replaced with statistics that summarize the neighbouring outputs, meaning the size of every input region and eventually the input feature map is spatially reduced. The spatial reduction is usually done using one of the two operations: max or average. In the max-pooling operation, pooling is done over a window as shown in Fig 2.1, and the maximum activation value of the a window size is selected.

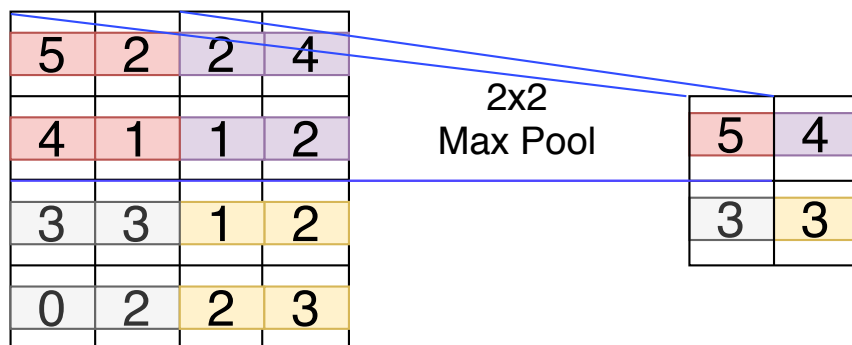


Figure 2.1: Example of max-pooling using a 2x2 stride.

Mathematically, the same can be expressed by considering a pooling layer with square input matrix of size M_{in} , which outputs a square matrix of size M_{out} . Assuming choosing a stride that leads to no overlapping of filters, the input is divided into pooling regions $p_{i,j}$ of a stride of size $k \times l$.

Max pooling then becomes a max operation applied element wise to given region, and tiled across the feature map with the given stride.

$$M_{out} = \max_{(k,l) \in p_{i,j}} M_{in}. \quad (2.5)$$

Here, k, l is the stride, where usually k has the same value as l .

In average pooling, the average of the activation values in the window is used. The pooling operation is also done to decrease the computational power required to process the data through dimensionality reduction and also to propagate only most dominant and relevant information further down the network.

Most commonly, max pooling is used over average pooling in various modern CNN architectures.

Unpooling

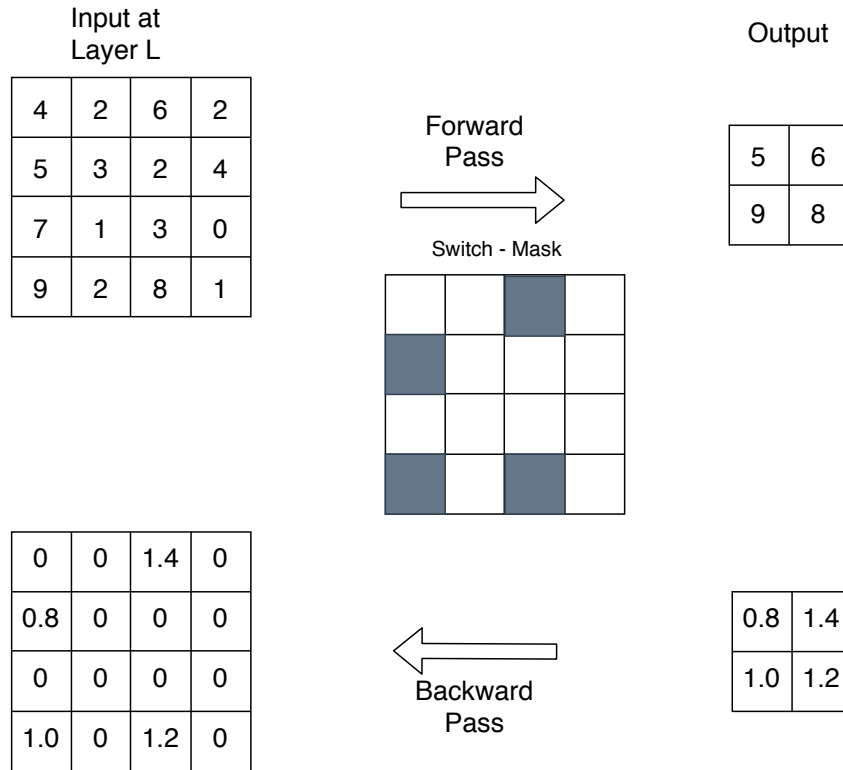


Figure 2.2: Example of Unpooling using a 2x2 stride.

In the above part we described the pooling operation. In this thesis, presented methods also use an operation known as “unpooling”. Hence, we describe it in detail here.

As explained above, in the forward pass, for the max-pooling operation we choose a window stride. Then, we choose the max value in that window to create a max-pooled output. In the unpooling operation, two additional operations are done. First, in the forward pass, we save the location of the max-values as a binary “switch” matrix. In the backward pass, we just the replace the open “switch” values at layer l from the values of layer $l + 1$. An example of this is shown in Fig. 2.2

2.2.4 Batch-Normalization Layer

Batch normalization (BN) layer [51] normalizes the output of the previous layer (x) by subtracting the mean (μ) from the output and dividing it with the standard deviation of the output matrix (σ^2) for a given batch (b). Similar to the seminal paper on BN [51], mathematically for a batch (b) with k examples, it can be expressed as:

$$\mu_b \leftarrow \frac{1}{k} \sum_{i=1}^k x_i \quad ; \quad \sigma_b^2 \leftarrow \frac{1}{k} \sum_{i=1}^k (x_i - \mu_b)^2 \quad (2.6)$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_b}{\sqrt{\sigma_b^2 + \epsilon}} \quad (2.7)$$

For better optimization, a shift (α) and scale factor (β) is applied (the two trainable parameters) at the end after the normalization operation.

$$o_i \leftarrow \alpha \hat{x}_i + \beta \quad (2.8)$$

2.2.5 Dropout

Dropout's idea is based on probability. The method temporarily “drops out” certain neurons during training of the neural network. Mathematically, we use a probability p , to select whether to drop a neuron temporarily. Usually the value range for p varies from 0.25 – 0.5. During inference, dropout is disabled.

2.2.6 Fully Connected (FC) Layer

Fully Connected (FC) layers are generally used closer to the output layer to capture global context and model high-level concepts. In a fully connected layer, as the name suggest, each neuron in the previous layer is connected to the every neuron in the next layer. This layer can be realised as a matrix multiplication and adding of a bias term. Consider a neural network with L hidden layers expressed in matrix form:

$$O_l = \sigma_l(W_l I_{l-1} + b_l). \quad (2.9)$$

Here, l is a particular layer in the given network, O_l is the output of the layer l , σ is the activation function, W_l is the weight matrix of the layer, I_{l-1} is the output of the previous layer and b_l is the bias.

There is another constituent layer or operation known as batch normalization in CNN. As most of the methods described in the thesis work in the inference mode, where the batch normalization operation parameters are fixed, hence it is not covered in detail here.

2.2.7 CNN Architecture

Combining the different components mentioned in the Sections above, we can obtain the architecture as described in Fig. 2.3. In the given example, the architecture of the CNN consists of three convolutional layers, two max pooling layers and two fully connected layers in the end. Activation functions are assumed to be implicitly applied after each convolutional operation, therefore not shown explicitly in the figure. For information processing, first the input image is fed into the first convolutional layer. Where the convolutional filters are applied to the different channels in the image, and summed up individually to form feature maps equal to the number of convolutional filters in the given layer. Each map is then passed first through activation function and then along the max-pooling layers. This process is repeated two more times. After the last convolutional, the feature maps outputs are flatten and then passed along to the fully connected layers. At the end, we pass the obtained activations through classifier (such as softmax activation, for multi class classification) to produce a prediction (such as a label here).

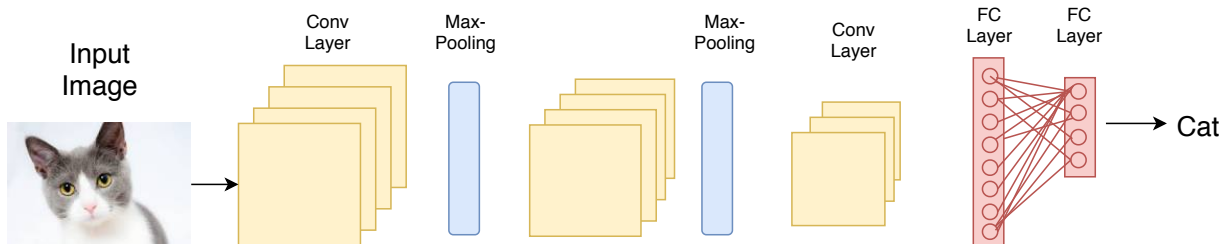


Figure 2.3: Example of CNN architecture design. Architecture of the deep convolutional network here consists of 3 convolutional layers and 2 max pooling layers, and two fully connected layers. The network takes an image as input (cat) and outputs a class label.

2.3 Types Of Explanation Methods

The proposed thesis research aims to investigate both local and global approaches to get a more comprehensive, well-rounded perspective of how deep neural networks make decisions. There has been a significant body of work in recent years in the domain of local explanations via visualizing and understanding CNNs from spatial-level support only. Also, as the thesis aims at the shortcoming with the heatmap based spatial-level support approaches, hence the literature pertaining to spatial support is described more in detail as compared to other methods in this Chapter.

This literature can be broadly divided into two groups:

- Instance Based Methods: approaches that mainly focus on understanding the decision-making process of trained networks for a specific instance [5,81,108,112,129,132,135].
- Global Understanding Based Methods: approaches that aim to understand the global structure of a trained network and its internal working scenarios [6,28,38,127].

Our presented work in Chapters 3, 4 and 5 can be considered as belonging to the first category, while the human attribute based explainability framework in Chapter 6 relates to the second category. Hence, we focus on approaches belonging to both categories and discuss relevant work pertaining to both groups. To follow the chronological order of the types of methods, we first present works related to local explainability and then briefly mention the few recent approaches belonging to second category as well.

2.4 Local Explanations via Instance-Based Methods

These methods are based on interpreting individual decisions made by a trained CNN for a particular image instance. This approach can also be known as post-hoc interpretability. Even though these approaches don't exactly elude to how a model works, they nonetheless provide useful information to the end user of the model. One of the major advantages of these approaches for interpretability is that we can interpret opaque models after-the-fact without sacrificing predictive performance [72].

Instance-based methods can be further divided into two categories, visualization via propagation methods and text explanation methods. These methods are explained below. It is important here to point out that all of the proposed approaches in this thesis belongs to visualization via propagation methods, hence they are explained in more detail.

2.4.1 Visualization via Propagation Based Methods

Approaches in this category use variants of either back-propagating or forward propagating information to create visualization maps. Such approaches use trained network structure itself for their visualizations. The common factor between different methods in this approach is that each method finds the contribution of each pixel in the input image by starting at an activation of interest in a particular layer of a network and then iteratively computing the contributions of each unit in the lower layer to that activation. This way, by moving backwards through the network from a particular unit to the input image, the contribution level of each pixel in the input image can be obtained leading to the creation of a visualization of the most attentive features that are most relevant to the activation of interest.

One of the first methods in this category was proposed by Simonyan et al. in [108]. The authors used back-propagated partial derivatives of the class score with respect to pixel values while masking out the negative entries in the feature maps to create class saliency maps. The authors reason that the derivative of the class score in the input space defines the relative importance of the input pixels for the resulting class score. From a different angle, this process is equivalent to finding pixels that if changed even slightly can produce a large effect on the final class score values. The underlying assumption is that the pixels constituting the objects are far more important than the pixels from unrelated objects or parts of an image.

Zeiler & Fergus [129] proposed a deconvolution-based method to project the activations from feature space back to the input space (pixels) recursively. Deconvolution networks, initially proposed for unsupervised training of CNNs were shown by Zeiler & Fergus that they can also be used to reverse the external stimuli to show which pattern and pixels in the input image are responsible for observed activations or output. Given a high level feature map in a layer, the deconvnet approach inverts the data flow of the given CNN for a particular unit in the layer. This is done by making all the units except the one we are interested in zeros. It results in a visualization map showing which regions in the image are mostly activating that particular unit. Also, in order to deal with max-pooling layers, that are not invertible, the authors first do a forward pass to compute *switches* which are binary maps with positions of maxima within each pooling region. These switches are then used in the backward pass to obtain discriminative reconstruction. This process is usually known as unpooling as explained in Section 2.2.3.

Springenberg et al. [112] provided another gradient-based visualization method, which builds upon the work of [108] and [129]. The method presented by Springenberg et. al. is mostly similar to [108] and [129] except how data goes through the non-linear ReLU

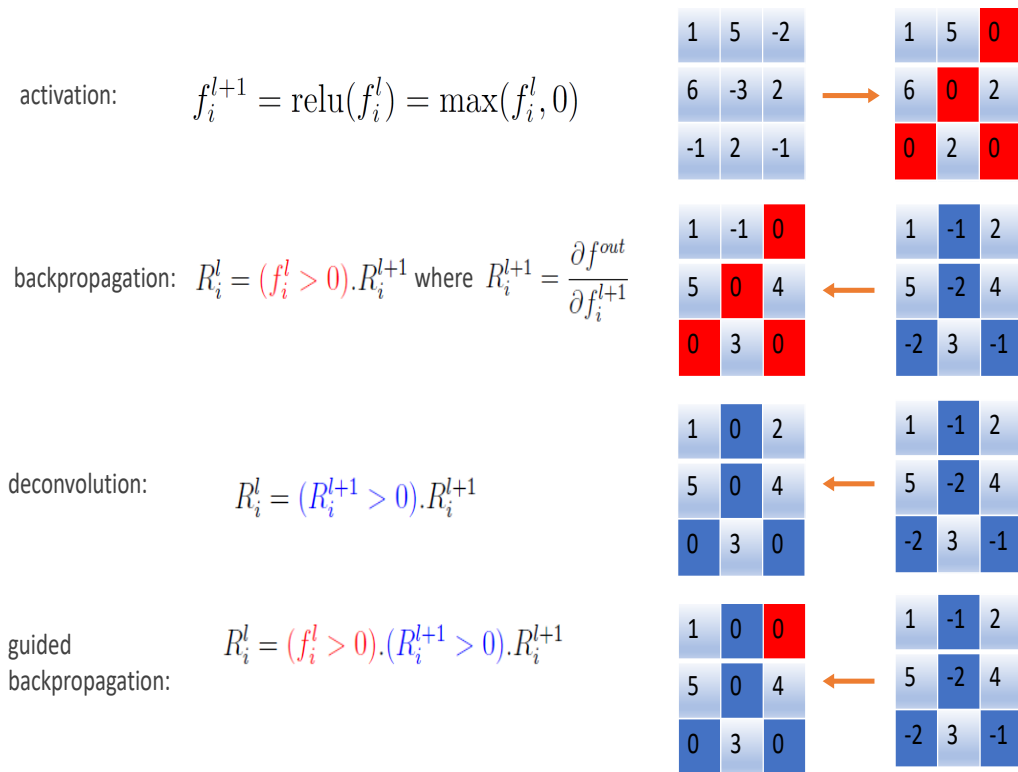


Figure 2.4: Different methods of back propagating through ReLU non-linearity along with the formal formulation for propagating an output back through a ReLU unit in layer l . Here, f is the feature map in the forward pass and R is the reconstructed feature map in the backward pass.

function in the backward pass and the fact that the authors used a network without max-pooling. The authors called this approach guided backpropagation. The deconvnet approach is equivalent to a backward pass through the network, except that when propagating through a non-linearity, its gradient is solely computed based on the top gradient signal, ignoring the bottom-up input. In the case of the ReLU non-linearity this amounts to setting to zero certain entries based on the top gradient. Rather than masking out values corresponding to negative entries of the top gradient (deconvnet) or bottom data

(back-propagation), guided back-propagation masks out the values for which at least one of these values is negative. This process is explained in equations 2.1-2.4 and Fig 2.5. This leads to sharper visualizations. The guided propagation approach also strongly shows the efficacy of networks with global average pooling for image classification and visualization for interpretability.

$$\text{Activation : } f_i^{l+1} = \text{ReLU}(f_i^l) = \max(f_i^l, 0). \quad (2.10)$$

$$\text{Backpropagation : } R_i^l = (f_i^l > 0) \cdot R_i^{l+1}; R_i^{l+1} = \frac{\partial f^{out}}{\partial f_i^{l+1}}. \quad (2.11)$$

$$\text{Deconvolution : } R_i^l = (R_i^{l+1} > 0) \cdot R_i^{l+1}. \quad (2.12)$$

$$\text{Guided Backpropagation : } R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1}. \quad (2.13)$$

In the above equations, R_i^l is the response at the l_{th} of given network, and R_i^{l+1} is the response from the layer ahead of l_{th} layer. f^l is the feature map out for filter at l_{th} layer.

However, the above mentioned methods did not provide any meaning to the assignments other than that they should form a coherent set of interpretable pixels that are responsible for certain outputs.

To visually discern unique features for a particular category of image, Zhou et. al. [132] created a *Class Activation Map* (CAM) using CNNs with a global average pooling layer. This *class activation map* was also used for localizing objects within the image as shown in Fig 2.6. The CAM indicates the discriminative regions in the input image used by a CNN to classify the image with respect to a particular category. Zhou et. al. added one more additional layer between GAP layer and softmax to learn weights associated with a particular class in the all convolutional layers. Using CAMs as a prior, the authors were also able to improve localization accuracy.

Recently Selvaraju et.al [99] proposed a method known as Grad-CAM in which the authors combined the guided-backpropagation [112] and Class Activation Map (CAM) methods [132] to weight the visualization maps provided by the CAM method with the gradients provided by guided backpropagation.

Also, some other related works in this area include Bach et al. [5] and Montavon et al. [81] which aimed at finding a general approach to visualize non-linear classifiers, leading to interesting heatmap generation. Recently, similar to the occlusion-based methodology

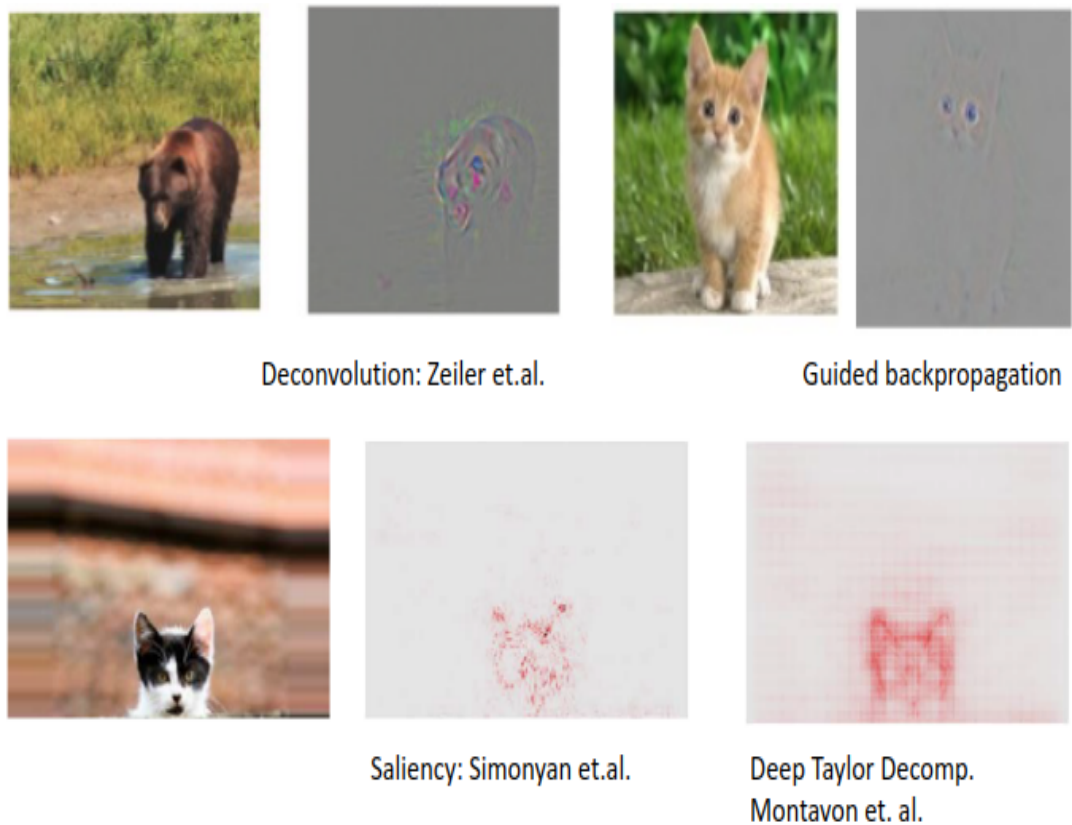


Figure 2.5: Different methods for creating visualization maps [40]. The figure shows the visualization maps and original images. All images were classified correctly.

for creating heatmaps in [129], Zintgraf et al. [135] proposed a method based on multivariate conditional sampling over image patches to visualize and interpret individual decisions of CNNs as binary saliency maps to represent information that contributes for or against the decision. This technique extends upon the work of Robnik-Sikonja & Kononenko [92] converting it from a univariate approach to a multivariate one.

In our work (Chapter 5), instead of only obtaining feature maps, we attribute meaning to each pixel in the back-projected response in the input space using a class-based approach. Also, unlike [5, 81] or [135], that provide heatmaps or binary heatmaps for correctly classified samples, we create CLEAR maps that are more interpretable (Fig. 5.5) for both correctly or misclassified cases. Finally, compared to the per-class maps created in [132], CLEAR maps show multiple class-specific contributions at once.

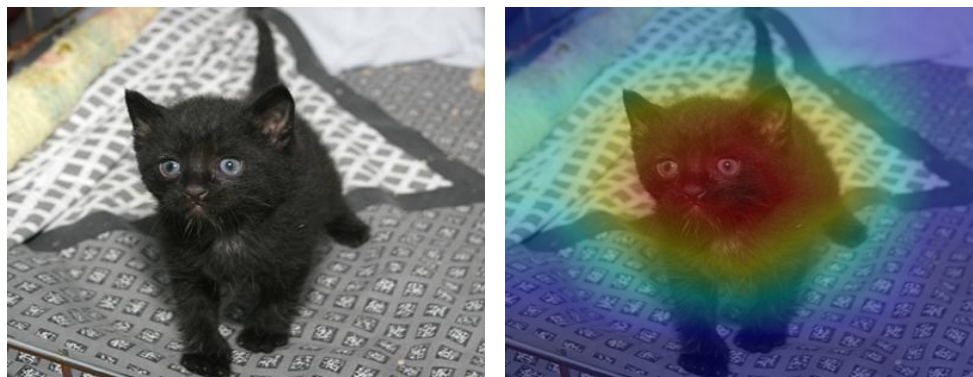


Figure 2.6: Class activation map example.

2.4.2 Text Based Local Explanations

Humans generally provide explanations using words. Using this motivation, Krening et.al. [59] used a two network approach where a CNN was used for prediction and a second, recurrent neural network was used to generate text based explanations. These explanations are trained to maximize the likelihood of previously observed ground truth explanations from humans.

In recommender systems, McAuley & Leskovec [78] used text to explain the decisions of the latent factor. The authors used simultaneous training of a latent factor model for ranking and a topic model over product reviews. Park et. al. [87] used an attention based VQA model to provide explanation for human activities. Recently, Selvaraju et.al. [99] also provided a VQA model with their Grad-CAM method to provide text based explanations for instances.

2.5 Global Explanations Based Methods

Many methods in this domain try to understand the decision-making process of deep networks by measuring their operating characteristics; for example, finding an input that maximizes the response of a particular neuron [28], measuring the network’s invariance to certain kinds of data augmentation [38], or determining global decision structure [6]. Other methods seek to find image instances from a database that maximally activate particular neurons or the posterior class probability of a given network [127].

Recently, there has been an emerging line of research in the sub-field of global explanations methods based on human concepts. In this line of research, the aim is to provide explanations in terms of human understandable high level “concepts” instead of providing justifications for a given instance via feature based visualization methods (explained in instance based methods above). As this is an emerging sub-field, only a handful of research studies have been published so far. In Zhou et. al. [8], authors created a database of 1200 odd pre-defined concepts for generic natural images. Using the database, they evaluated specific nodes in deep neural networks to probe which concept do each of the specific node respond to. This is an important starting point, but this requires a large fixed database of concepts, which is difficult to adopt for various other specialized domains such as medical imaging. Kim et. al. [57] introduced another method known as TCAV, where they use human provided concepts for each instance and test the sensitivity of prediction with respect to the given concept. For example, how sensitive a prediction of “zebra” is to the presence of the concept of stripes. One of the major shortcomings of this method is that the human user needs to provide concepts for each given instance (image in this case). The same has been pointed out in their later work in [34]. In [34], the authors proposed to use an automated method to discover different levels of concepts for a given neural network. The proposed Automated Concept Explanation (ACE) method uses three different levels of super-pixel segmentation to get “important” parts of a given image and then use the trained neural network for discovering concepts. This is an interesting approach in automating concept discovery. However, in this approach there is still a considerable amount of subjectivity involved with regards to the discovered concepts as the human expert still needs to go through the discovered concept and try to understand what exactly has been discovered. This defeats the purpose of automated concept discovery as the human expert is the end-user and she/he still needs to approve the discovered concept. Also, though the ACE approach works well, it cannot be used explicitly to compare different trained networks as the concepts are discovered uniquely to a given network.

To overcome some of above mentioned challenges, in Chapter 6 we provide a framework with human attribute prior based concept explanation method. In our proposed method, the human expert uses few exemplar images for a given class to define “concepts” specific for their domain. Using the defined concepts, the proposed method automatically evaluates if the predictions being made by given neural network for a specific class is based on human expert defined concepts or not. In the Chapter, we also present arguments on how the proposed approach can be used to provide a quantitative metric to explanations and in theory can be used to evaluate and compare different trained neural networks and various local instance based visualization methods.

2.6 Summary

In the present Chapter, we covered some of the recent and relevant work in the field of explainability for both local and global explanations. In Chapter 3, we will go into detail regarding some of the shortcomings of current heatmap based (visualization) local explainability methods. This includes the scenarios where the current heatmap-based approaches fail and how through an introduction of layer wise response method, we can solve some of the shortcomings. Experiments using three different datasets are provided in Chapter 3 to introduce the shortcomings of the current methods and present solutions for some of them.

Chapter 3

Domain Knowledge Based Architecture Design Using Attentive Response Maps

At the same time, Deep Patient is a bit puzzling. It appears to anticipate the onset of psychiatric disorders like schizophrenia surprisingly well. But since schizophrenia is notoriously difficult for physicians to predict, Dudley wondered how this was possible. He still doesn't know. The new tool offers no clue as to how it does this. If something like Deep Patient is actually going to help doctors, it will ideally give them the rationale for its prediction, to reassure them that it is accurate and to justify, say, a change in the drugs someone is being prescribed. "We can build these models," Dudley says ruefully, "but we don't know how they work."

- "The Dark Secret at the Heart of AI", Wired Magazine, 2017

Prologue to Articles

The Chapter borrows contribution from two different papers, details of which are mentioned below.

3.0.1 Article Details

Insightful classification of crystal structures using deep learning, A. Ziletti, D. Kumar, M. Scheffler, L. M. Ghiringhelli, *Nature Communications*, Vol. 9(1), p.2775, 2018

Understanding anatomy classification through attentive response maps, D. Kumar, V. Menkovski, G. W. Taylor, A. Wong, *IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, 2018

Personal Contribution For the second paper, I devised the study on the x-ray human anatomy medical image dataset, and conducted various experiments to explore the significance of attention maps beyond individual node outputs and wrote the manuscript. After our first work, I was invited to work with physicists at FHI, Berlin. For this work, I wrote the code for the visualization and the initial code for training the CNN model used in this study. Dr. Angelo Ziletti was the lead on acquiring the data, study setup and eventually making the graphs and visualization from my code more presentable. He also wrote most of the manuscript, except the parts related to explainability.

3.0.2 Context

The idea of working on explainable AI (XAI) came up during my internship at Philips Research at their HQ in Eindhoven, Netherlands. At that time, there was very limited work being done in the XAI domain. Realizing its potential impact and with motivation from a mentor at Philips research, I decided to work in this field. As the attentive visualization methods were the first forayed direction in this area, I decided to pursue the same. First, under the guidance of my mentor at Philips research, Dr. Vlado Menkovskiv (now a Prof. at TU Eindhoven), and later with my PhD supervisors, we decided to first explore how attentive methods could be leveraged to understand the learning process of deep neural networks. Later, I collaborated with physicists on a similar study for crystal structure classification. I was invited by physicists at FHI, Max Plack, Berlin, who were impressed by our first study and wanted to explore it for their new material discovery project known as NOMAD.

3.0.3 Contribution

In our human anatomy study, we explore and show case how attention maps can be formed beyond individual neurons to understand the neural network's decision process. We present experimentation to show where in the training process does the hierarchy of features emerge and at which stage do neural networks start to use landmarks that correlate with human domain knowledge.

Most of the current methods in the crystal structure identification study fail on slight noise/imperfect data and require a tolerance threshold to produce coherent results. Our procedure does not require any tolerance threshold, and it is very robust to defects (even at defect concentrations as high as 40%). First, we introduce a way to represent crystal structures (by means of images, i.e. two-dimensional maps of the three-dimensional crystal structures, see below), then we present a classification model based on convolutional neural networks, and finally we unfold the internal behavior of the classification model through visualization. An interactive online tutorial for reproducing the main results of this work is also provided to further the development of advanced methods in this field.

The articles used in this Chapter have gained a total of 67 citations to date (based on Google Scholar).

3.1 Introduction

The previous Chapter outlined various attention based visualizations for explaining local predictions for deep neural networks. However there are various shortcomings associated with such methods, some of which are listed below:

- These methods only convey the attentive region for individual nodes in a particular layer. A holistic layer wise representation is absent.
- No additional information is provided for the attentive regions such as how the attentive regions relate to human domain expertise.
- No information is shown regarding how the learning process of neural networks changes with or during training and how it relates to human domain expertise.
- There is a lack of overall end-to-end explanations that are provided with these methods.

To overcome some of the challenges and shortcomings mentioned above, in this Chapter we introduce the concept of attention response maps. First, we show how attention maps can be used to form an interpretability response as attentive regions for any convolutional layer in a given CNN. We then relate these attentive regions used by deep neural networks for making decisions to the human domain expertise through various carefully designed experiments and ablation studies. Along with this, we also conduct experiments to show where in the training process of such neural network’s learning process does the hierarchy of features emerge and at which stage of training, do neural networks start to use landmarks that correlate with human domain knowledge.

In the following sections, we first lay out the basic framework and formulation for forming the attentive response maps for any given convolutional layer for a trained network. We then show the relevance of attentive regions through two different experimental design studies to show the efficacy of the proposed framework.

3.2 Methodology - Attention Maps Formation

This section presents the generic framework and formulation of the attentive response maps for any convolution layer in a CNN. The below mentioned formulation is then applied to two different studies that deal with classification experiments as explained in detail by their respective experiments section later in the Chapter.

3.2.1 Formulation

To explain the formulation of the attentive response maps, first consider a single layer of a CNN. Let \hat{h}_l be the deconvolved output response of the single layer l with n unit weights w . The deconvolution output response at layer l then can be then obtained by convolving each of the feature maps z_l with unit weights w_l and summing them as:

$\hat{h}_l = \sum_{k=1}^n z_{k,l} * w'_{k,l}$. Here $*$ represents the convolution operation. For notational brevity, we can combine the convolution and summation operation for layer l into a single convolution matrix G_l . Hence the above equation can be denoted as: $\hat{h}_l = G_l z_l$.

For multi-layered CNNs, we can extend the above formulation by adding an additional un-pooling operation U as described in Section 2.2.3 and [129]. Thus, we can calculate the deconvolved output response from feature space to input space for any layer l in a multi-layer network as:

$$R_l = G_1 U_1 G_2 U_2 \cdots G_{l-1} U_{l-1} G_l z_l. \quad (3.1)$$

For attentive response maps, we specifically calculate the output responses from individual units of the last conv. layer of a network. Hence, given a network with last layer L containing n top activated units, we can calculate the attentive response map; $R(\underline{x}|f)$ (where \underline{x} denotes the response back-projected to the input layer, and thus an array the same size as the input) for any unit f ($1 \leq f \leq n$) in the last conv layer as:

$$R(\underline{x}|f) = G_1 U_1 G_2 U_2 \cdots G_{L-1} U_{L-1} G_L^f z_L. \quad (3.2)$$

Here G_L^f represents the convolution matrix operation in which the unit weights w_L are all zero except that at the f^{th} location.

Given the set of individual attentive response maps, we then compute the dominant attentive response map, $\hat{D}(\underline{x})$, by finding the value at each pixel that maximizes the attentive response level, $R(\underline{x}|f)$, across all top n units:

$$\hat{D}(\underline{x}) = \underset{f}{\operatorname{argmax}} R(\underline{x}|f). \quad (3.3)$$

The above formulation is used to form attentive response maps for two different studies namely crystal structure classification and human anatomy classification. We first present the whole experimental design, formulation and results pertaining to the crystal structure

classification study, make some observations and then proceed to the human anatomy classification study to show how the learning process of the network changes with the change in architecture. We then show how deep neural networks with sufficient depths use the same landmarks as human experts. We further solidify our observations in both classification scenarios by using carefully designed ablation experiments.

3.3 Experiments- Crystal Structures Classification

3.3.1 Motivation

Crystals play a crucial role in materials science. In particular, knowing chemical composition and crystal structure - the way atoms are arranged in space - is an essential ingredient for predicting properties of a material [15, 30, 86]. Indeed, it is well-known that the crystal structure has a direct impact on materials properties [84]. Just to give a concrete example: in iron, carbon solubility (important for steel formation) increases nearly forty times going from body-centered-cubic (bcc) α -Fe (ferrite) to face-centered-cubic (fcc) γ -Fe (austenite) [110]. From the computational point of view, identification of crystal symmetries allows, for example, to construct appropriate k -point grids for sampling, generate paths between high-symmetry points in band structure calculations, or identify distortions for finite-displacement calculations.

Given the importance of atomic arrangement in both theoretical and experimental materials science, an effective way of classifying crystals is to find the group of all transformations under which the system is invariant; in three-dimensions, these are described by the concept of space groups [43]. Currently, to determine the space group of a given structure, one first determines the allowed symmetry operations, and then compares them with all possible space groups to obtain the correct label; this is implemented in existing symmetry packages such as FINDSYM [115], Platon [111], Spglib [27, 39], and most recently the self-consistent, threshold-adaptive AFLOW-SYM [47]. For idealized crystal structures, this procedure is exact. But in most practical applications atoms are displaced from their ideal symmetry positions due to (unavoidable) intrinsic defects or impurities or experimental noise. To address this, thresholds need to be set in order to define how loose one wants to be in classifying (namely, how much deviation from the ideal structures is acceptable); different thresholds may lead to different classifications (see for instance Table 3.2). So far, this was not a big problem because individual researchers were manually finding appropriate tolerance parameters for their specific dataset.

However, our goal here is to introduce an automatic procedure to classify crystal structures starting from a set of atomic coordinates and lattice vectors; this is motivated by the advent of high-throughput materials science computations, thanks to which millions of calculated data are now available to the scientific community (see the Novel Materials Discovery (NOMAD) Laboratory [23] and references therein). Clearly, there is no universal threshold that performs optimally (or even sub-optimally) for such a large number of calculations, nor a clear procedure to check if the chosen threshold is sound. Moreover, the aforementioned symmetry-based approach fails - regardless of the tolerance thresholds - in the presence of defects such as, for example, vacancies, interstitials, antisites, or dislocations. In fact, even removing a single atom from a structure causes the system to lose most of its symmetries, and thus one typically obtains the (low symmetry, e.g. $P1$) space group compatible with the few symmetry operations preserved in the defective structure. This label - although being technically correct - is practically always different from the label that one would consider appropriate (i.e. the most similar space group, in this case the one of the pristine structure). Robustness to defects, however, is paramount in local and global crystal structure recognition. Grain boundaries, dislocations, local inclusions, and in general all crystallographic defects can have a large impact on macroscopic materials properties (e.g. corrosion resistance [24, 97]). Furthermore, atom probe tomography - arguably the most important source of local structural information for bulk systems - provides three-dimensional atomic positions with an efficiency up to 80% [32] and near-atomic resolution; which, on the other hand, means that at least 20% of atoms escaped detection, and the uncertainty on their positions is considerable.

Here, we propose a procedure to efficiently represent and classify potentially noisy and incomplete three-dimensional materials science structural data according to their crystal symmetry (and not to classify x-ray diffraction images, or powder x-ray diffraction data [88]). These three-dimensional structural data could be for example atomic structures from computational materials science databases, or elemental mappings from atom-probe tomography experiments. Our procedure does not require any tolerance threshold, and it is very robust to defects (even at defect concentrations as high as 40%). First, we introduce a way to represent crystal structures (by means of images, i.e. two-dimensional maps of the three-dimensional crystal structures, see below), then we present a classification model based on convolutional neural networks, and finally we unfold the internal behavior of the classification model through visualization. An interactive online tutorial for reproducing the main results of this work is also provided [134].

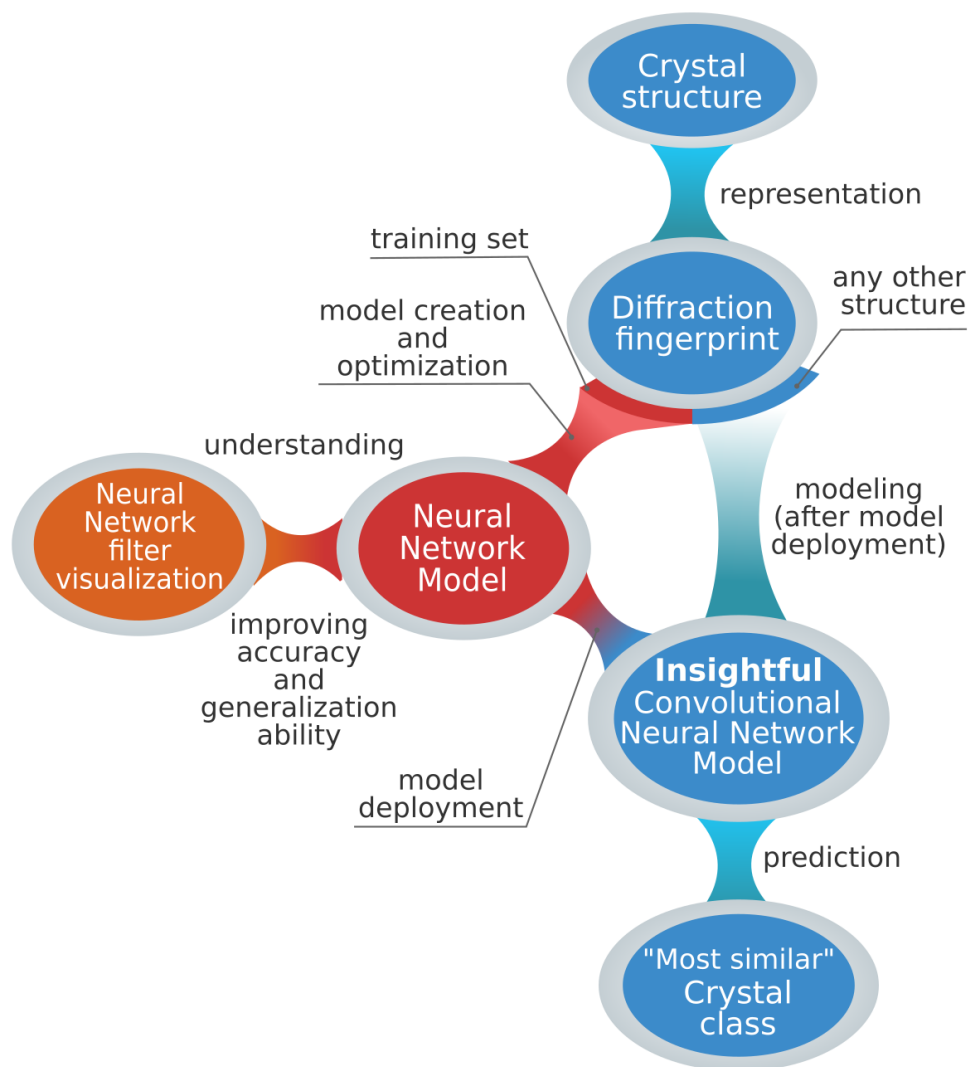


Figure 3.1: The model workflow of automatic crystal-structure classification. First, every crystal structure is represented by its two-dimensional diffraction fingerprint. Then, a small subset of these structures is used as training set to generate a classification model. In particular, a convolutional neural network is used, and optimized by minimizing the training set classification error. However, this is in general not enough to have a sound and generalizable model. Thus, we unfold the neural network internal operations by visualization, and ensure that the model arrives at its classification decision on physically motivated grounds. Finally, a classification model is deployed, and crystal structures can be directly and efficiently classified without any additional model optimization.

3.3.2 Dataset Explanation

Our pristine dataset consists of materials from the AFLOWLIB elemental solid database [16] belonging to centrosymmetric space groups which are represented with more than 50 configurations in the database. Specifically, we extract structures that have a consistent space group classification for different symmetry tolerances. This gives us crystal structures belonging to the following space groups: 139 (bct), 141 (bcc), 166 (rh), 194 (hex), 221 (sc), 225 (fcc), 227 (diam), and 229 (fct). From this, we apply the defective transformations (random displacements, vacancies, and chemical substitutions) to the pristine structures; the resulting dataset is used as a test set. For this defective dataset we use labels from the pristine structures because the materials' class will unlikely be changed by the transformations above. To quantify this, let us consider the transformation of bcc into sc crystals for the case of random vacancies as an illustrative example. As stated, sc structure can be obtained removing all atoms laying at the center of the bcc unit cell (see Fig.3.2b). Therefore, for a structure comprising N atoms, one needs to remove exactly the $N/2$ atoms which are at the center of the cubic unit cell (note that each corner atom is shared equally between eight adjacent cubes and therefore counts as one atom). For $N/2$ randomly generated vacancies, the probability of removing all and only these central atoms is $P_N = 2 \left[\binom{N}{N/2} \right]^{-1}$ which - for the structure sizes considered in this work - leads to negligible probabilities ($P_{64} \approx 10^{-18}$, $P_{128} \approx 10^{-38}$). The same holds for chemical substitutions: even if in principle they could change the space group (e.g. diamond to zincblende structure), the probability of this happening is comparable with the example above, and therefore negligible. Finally, in the case of displacements, atoms are randomly moved about their original positions, and due to this randomness it is not possible to obtain any long-range re-organization of the crystal, necessary to change the materials' class; moreover, for large displacements, the system becomes amorphous (without long-range order).

3.3.3 Experiment Design

Neural network architecture and training procedure

The architecture of the convolutional neural network used in this work is detailed in Table 3.1. Training was performed using Adam optimization with batches of 32 images for 5 epochs with a learning rate of 10^{-3} , and cross-entropy as cost function.

Layer type	Specifications
Convolutional Layer	(Kernel: 7x7; 32 filters)
Convolutional Layer	(Kernel: 7x7; 32 filters)
Max Pooling Layer	(Pool size: 2x2, stride: 2x2)
Convolutional Layer	(Kernel: 7x7; 16 filters)
Convolutional Layer	(Kernel: 7x7; 16 filters)
Max Pooling Layer	(Pool size: 2x2, stride: 2x2)
Convolutional Layer	(Kernel: 7x7; 8 filters)
Convolutional Layer	(Kernel: 7x7; 8 filters)
Fully connected Layer + Dropout	(Size: 128; dropout: 25%)
Batch Normalization	(Size: 128)
Softmax	(Size: 7)

Table 3.1: Architecture of the convolutional neural network used in this work.

3.3.4 Results

For the sake of completeness, we first present the results associated with how to represent a material and obtain the various crystal structure figures used in the study succinctly.

How to represent a material

The first necessary step to perform any machine learning and/or automatized analysis on materials science data (see Fig. 3.1) is to represent the material under consideration in a way that is understandable for a computer. This representation termed as “descriptor” [33] should contain all the relevant information about the system needed for the desired learning task.

In the case of crystal-structure recognition, it is essential that the descriptor captures the system’s symmetries in a compact way, while being size-invariant in order to reflect the infinite nature of crystals. Periodicity and prevailing symmetries are evident and more compact in reciprocal space, and therefore we introduce an approach based on this space. Details of how the representation is formed are explained in Appendix A.1

However, a disadvantage of the two-dimensional diffraction fingerprint (as shown in Fig 3.2) is that it is not unique across space groups. This is well-known in crystallography: the diffraction pattern does not always determine unambiguously the space group of a

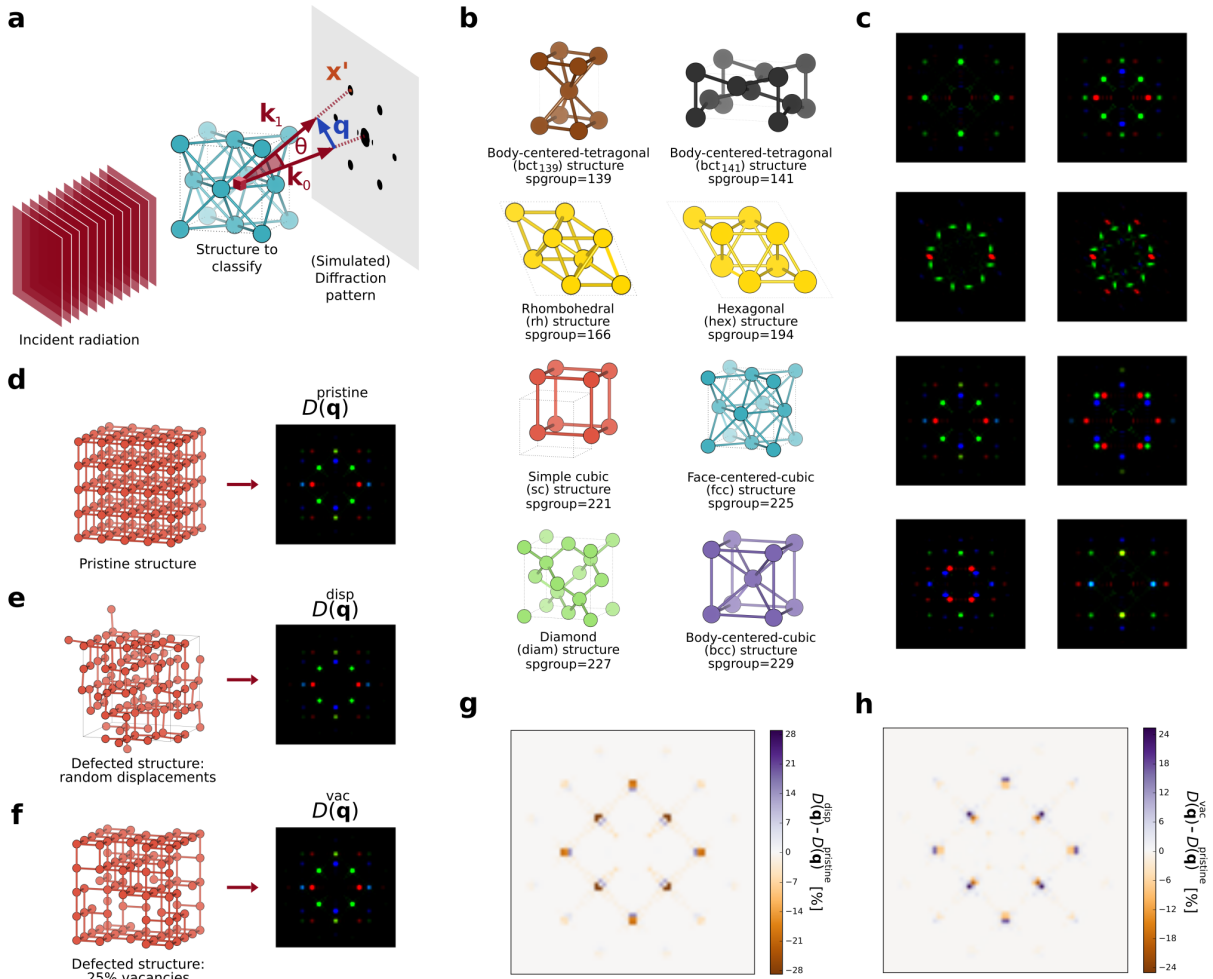


Figure 3.2: The two-dimensional diffraction fingerprint. (a) Schematic representation of the two-dimensional diffraction fingerprint calculation. An incident plane wave is scattered by the material, and the diffraction pattern on a plane perpendicular to the incident radiation is computed. (b) Prototypes of the crystal classes considered in this work. (c) Examples of two-dimensional diffraction patterns for materials belonging to each of the eight classes. The ordering is the same as b. Rhombohedral and hexagonal structures have the same two-dimensional diffraction fingerprint. (d)-(e)-(f) A pristine simple-cubic structure (d), the same structure with 25% of vacancies (e), and with atoms displaced randomly according to a Gaussian distribution with standard deviation of 0.08 Å (f), together with their diffraction fingerprints. (g) (h) Difference between the diffraction fingerprints of the defective e-f and the pristine structure d.

crystal. This is primarily because the symmetry of the diffraction pattern is not necessarily the same as the corresponding real-space crystal structure; for example, Friedel’s law states that if anomalous dispersion is neglected a diffraction pattern is centrosymmetric, irrespective of whether or not the crystal itself has a centre of symmetry. Thus, the diffraction fingerprint D_F cannot represent non-centrosymmetric structures by construction. The non-uniqueness of the diffraction pattern $I(\mathbf{q})$ across space groups also implies that crystal structures belonging to different space groups can have the same diffraction fingerprints. Nevertheless, from Fig. 3.2c we notice that out of the eight crystal structure prototypes considered (covering the large majority of the most thermodynamically stable structures formed in nature by elemental solids, only the rhombohedral and hexagonal structures whose real-space crystal structures are quite similar have the same two-dimensional diffraction fingerprint.

The classification model

Having introduced a way to represent periodic systems using scattering theory, we tackle the problem of their classification in crystal classes based on symmetries. A first (and naive) approach to classify crystals - now represented by the diffraction descriptor D_F - would be to write specific programs that detect diffraction peaks in the images, and classify accordingly. Despite appearing simple at first glance, this requires numerous assumptions and heuristic criteria; one would need to define what is an actual diffraction peak and what is just noise, when two contiguous peaks are considered as one, how to quantify relative peak positions, to name but a few. In order to find such criteria and determine the associated parameters, one in principle needs to inspect all (thousands or even millions of) pictures that are being classified. These rules would presumably be different across classes, require a separate and non trivial classification paradigm for each class, and consequently lead to a quagmire of ad-hoc parameters and task-specific software. In addition, the presence of defects leads to new peaks or alters the existing ones (see Fig. 3.2g and 3.2h), complicating matters even further. Thus, this approach is certainly not easy to generalize to other crystal classes, and lacks a procedure to systematically improve its prediction capabilities.

However, it has been shown that all these challenges can be solved by deep-learning architectures [68, 98]. These are computational non-linear models sequentially composed to generate representations of data with increasing level of abstraction. Hence, instead of writing a program by hand for each specific task, we collect a large amount of examples that specify the correct output (crystal class) for a given input (descriptor image D_F), and then minimize an objective function which quantifies the difference between the predicted and the correct classification labels. Through this minimization, the weights (i.e. parameters)

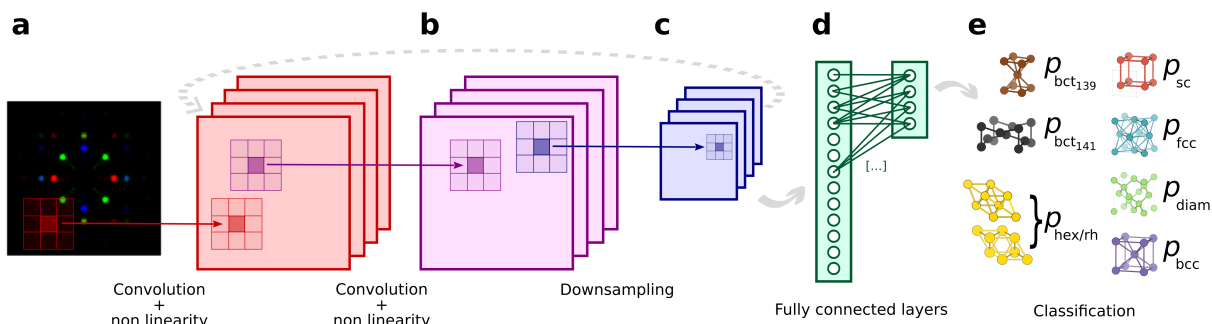


Figure 3.3: Schematic representation of the convolutional neural network (ConvNet) used for crystals classification. (a) A learnable filter (also called kernel) is convolved across the image, and the scalar product between the filter and the input at every position is computed. This results in a two-dimensional activation map (in red) of that filter at each spatial position, which is then passed through a rectified linear unit (ReLU). (b) The same procedure as point a is applied to this activation map (instead of the original image), producing another activation map (in purple). (c) A down-sampling operation (in blue) is performed to coarse-grain the representation. Six convolutional and two down-sampling (max-pooling) layers are stacked sequentially (see Methods for additional details). (d) The output of the convolutional/down-sampling layers sequence is passed to fully-connected layers (regularized using dropout) to complete the classification procedure. (e) The ConvNet outputs the probabilities that the input image, and therefore the corresponding material, belongs to a given class. Minimizing the classification error, the above-mentioned filters are learned - through back-propagation - and they will activate when a similar feature (e.g. edges or curves for initial layers, and more complex motifs for deeper layers) appears in the input.

of the neural network are optimized to reduce such classification error [48, 49]. In doing so, the network automatically learns representations (also called features) which capture discriminative elements, while discarding details not important for classification. This task known as feature extraction usually requires a considerable amount of heuristics and domain knowledge, but in deep learning architectures is performed with a fully automated and general-purpose procedure [68]. In particular, since our goal is to classify images, we use a specific type of deep learning network which has shown superior performance in image recognition: the convolutional neural network (ConvNet) [61, 69]. A schematic representation of the ConvNet used in this work is shown in Fig. 3.3.

As detailed in Chapter 2, CNNs are inspired by the multi-layered organization of the visual cortex: filters are learned in a hierarchical fashion, composing low-level features (e.g. points, edges or curves) to generate more complex motifs. In our case, such motifs encode the relative position of the peaks in the diffraction fingerprint for the crystal classes considered, as we will show below.

The model performance

For every calculation in the AFLOWLIB elemental solid database [16], we determine its space group using a symmetry-based approach [27, 39] as implemented by the Spglib code. We then extract all systems belonging to centrosymmetric space groups which are represented with more than 50 configurations. This gives us systems with the following space group numbers: 139, 141, 166, 194, 221, 225, 227, and 229. For the case of elemental solids presented here, these space groups correspond to body-centered-tetragonal (bct, 139 and 141), rhombohedral (rh, 166), hexagonal (hex, 194), simple cubic (sc, 221), face-centered-cubic (fcc, 225), diamond (diam, 227), and body-centered-cubic (bcc, 229) structures. This represents a rather complete dataset since it includes the crystal structures adopted by more than 80% of elemental solids under standard conditions. It is also a challenging dataset because it contains 10,517 crystal structures comprising 83 different chemical species, cells of various size, and structures that are not necessarily in the most stable atomic arrangement for a given composition, or even at a local energy minimum. This last point in particular could potentially be a problem for the symmetry-based approach: when crystals are not in a perfect arrangement, it can fail in returning the correct labels. In fact, if atoms are slightly displaced from their expected symmetry positions, the classification could return a different space group because symmetries might be broken by this numerical noise. To avoid this, we include in the pristine dataset only systems which are successfully recognized by the symmetry-based approach to belong to one of the eight classes above, thus ensuring that the labels are correct. We refer to the above as the pristine dataset; the dataset labels are the aforementioned space groups, except for rh and hex structures, which we merge into one class (hex/rh) since they have the same diffraction fingerprint (see Fig. 3.2c).

We apply the workflow introduced here (and schematically shown in Fig. 3.1) to this dataset. For each structure, we first compute the two-dimensional diffraction fingerprint D_F ; then, we train the ConvNet on (a random) 90% of the dataset, and use the remaining 10% as test set. We obtain an accuracy of 100% on both training and test set, showing that the model is able to perfectly learn the samples and at the same time is capable of correctly classifying systems which were never encountered before. The ConvNet model optimization

	Random Displacements (σ)						Vacancies (η)				
	0.001Å	0.002Å	0.005Å	0.01Å	0.02Å	0.06Å	1 %	2 %	15 %	25 %	
Spplib (tight)	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	
Spplib (medium)	73.70	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	
Spplib (loose)	99.99	99.99	99.99	75.22	0.00	0.00	0.01	0.00	0.00	0.00	
This work	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	

Table 3.2: Accuracy in identifying the correct (most similar) crystal class in the presence of defects created from test set. The defective structures are calculated randomly displacing atoms according to Gaussian distribution with standard deviation σ (left), or removing $\eta\%$ of the atoms (right). The accuracy values shown in the table are in percentages.

(i.e. training) takes 80 minutes on a quad-core Intel(R) Core(TM) i7-3540M CPU, while one class label is predicted - for a given D_F - in approximately 70 ms on the same machine (including reading time). The power of machine learning models lies in their ability to produce accurate results for samples that were not included at training. In particular, the more dissimilar test samples are from the training samples, the more stringent is the assessment of the model generalization performance. To evaluate this, starting from the pristine dataset, we generate heavily defective structures introducing random displacements (sampled from Gaussian distributions with standard deviation σ), randomly substituting atomic species (thus forming binaries and ternaries alloys), and creating vacancies. This results in a dataset of defective systems, for some of which even the trained eyes of a materials scientist might have trouble identifying the underlying crystal symmetries from their structures in real space (compare for example, the crystal structures in Fig.3.2d with 3.2e and 3.2f).

As mentioned in the Introduction and explicitly shown below, symmetry-based approaches for space group determination fail in giving the correct (most similar) crystal class in the presence of defects. Thus, strictly speaking, we do not have a true label with which to compare. However, since in this particular case the defective dataset is generated starting from the pristine, we do know the original crystal class for each sample. Hence, to estimate the model generalization capability, we label the defective structures with the class label of the corresponding pristine (parental) system. This is a sensible strategy given that displacing, substituting or removing atoms at random will unlikely change the materials' crystal class. Using the ConvNet trained on the pristine dataset (and labels from the pristine structures), we then predict the labels for structures belonging to the defective dataset. A summary of our findings is presented in Table 3.2, which comprises results for $10,517 \times (6 + 4) = 105,170$ defective systems.

When random displacements are introduced, Spglib accuracy varies considerably according to the threshold used; moreover, at $\sigma \geq 0.02 \text{ \AA}$ Spglib is never able to identify the most similar crystal class, regardless of threshold used. Conversely, the method proposed in this work always identifies the correct class up to σ as high as 0.06 \AA . Similar are the results for vacancies: Spglib accuracy is $\sim 0\%$ already at vacancies concentrations of 1%, while our procedure attains an accuracy of 100% up to 40% vacancies, and greater than 97% for vacancy concentrations as high as 60% (Table 3.2). Since no defective structure was included at training, this represents a compelling evidence of both the model robustness to defects and its generalization ability.

If random changes will unlikely modify a crystal class, it is however possible to apply targeted transformations in order to change a given crystal from one class to another. In particular, starting from a bcc one can obtain a sc crystal removing all atoms at the

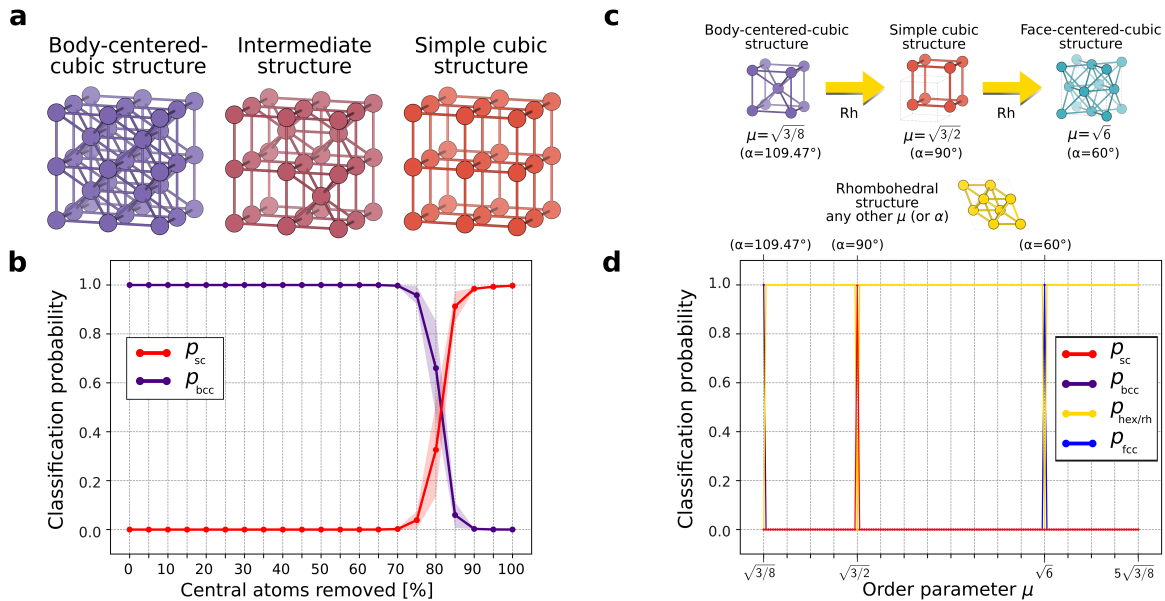


Figure 3.4: Neural network predictions on structural transitions. (a)(b) Body-centered-cubic (bcc) to simple cubic (sc) structural transition. (a) Examples of a bcc, an intermediate bcc/sc, and a sc structure. (b) Distributions of classification probability for the bcc (purple) and sc (red) classes as a function of the percentage of central atoms being removed (see text for more details). The shaded area corresponds to a range of one standard deviation above and below these distributions. (c)(d) Structural transition: transition path including rhombohedral, body-centered-cubic, simple-cubic and face-centered-cubic structures.

center of the bcc unit cell (Fig.3.2b, and 3.4a). We remove different percentages of central atoms (from 0% to 100%, at 10% steps) from a subset of bcc structures in the pristine dataset; this gives us a collection of structures which are intermediate between bcc and sc by construction (see Fig. 3.4a center for a concrete example).

Let us now recall that the output of our approach is not only the crystal class, but also the probability that a system belongs to a given class; this quantifies how certain the neural network is regarding its classification. The probability of the aforementioned structures being bcc (purple) or sc (red) according to our model are plotted in Fig.3.4b as function of the percentage of central atoms removed (the shaded area indicates the standard deviation of such distributions). This percentage can be seen as a order parameter of the bcc-to-sc structural phase transition. If no atoms are removed, the structures are pure bcc, and the model indeed classifies them as bcc with probability 1, and zero standard deviation. At first, removing (central) atoms does not modify this behavior: the structures are seen by the model as defective bcc structures. However, at 75% of central atoms removed, the neural network judges that such structures are not defective bcc anymore, but are actually intermediate between bcc and sc. This is reflected in an increase of the classification probability of sc, a corresponding decrease in bcc probability, and a large increment in the standard deviation of these two distributions. When all central atoms are removed, we are left with pure sc structures, and the model classifies again with probability 1, and vanishing standard deviation: the neural network is confident that these structures belong to the sc class.

We conclude our model exploration applying the classification procedure to a structural transition path encompassing rhombohedral, body-centered-cubic, simple-cubic and face-centered-cubic structures. From the AFLOW Library of Crystallographic Prototypes, we generate rhombohedral structures belonging to space group 166. To test our model on this structural-transition path, we generate crystal structures, and use the neural network trained above to classify these structures. The results are shown in Fig. 3.4d. Our approach is able to identify when the prototype reduces to the high-symmetry structures mentioned above (at μ_{bcc} , μ_{sc} , and μ_{fcc}), and also correctly classify the structure as being rhombohedral for all other values of μ . This is indeed the correct behavior: outside the high symmetry bcc/sc/fcc the structure goes back to hex/rh precisely because that is the lower symmetry family (μ not equal to μ_{bcc} , μ_{sc} , or μ_{fcc}).

Opening the black-box using attentive response maps

Our procedure based on diffraction fingerprints and ConvNet correctly classifies both pristine and defective examples but are we obtaining the right result for the right reason? And

how does the ConvNet arrive at its final classification decision?

To answer these questions, we need to unravel the neural network’s internal operations; a challenging problem which has recently attracted considerable attention in the deep learning community [129]. The difficulty of this task lies in both the tendency of deep learning models to represent the information in a highly distributed manner, and the presence of non-linearities in the network’s layers. This in turn leads to a lack of interpretability which hindered the widespread use of neural networks in natural sciences: linear algorithms are often preferred over more sophisticated (but less interpretable) models with superior performance.

To shed light on the ConvNet classification process, we resort to visualization: using the fractionally strided convolutional technique as formulated in Section 3.2, we back-project attentive response maps (i.e. filters) in image space. Such attentive response maps - shown in Fig. 3.5 - identify the parts of the image which are the most important in the classification decision.

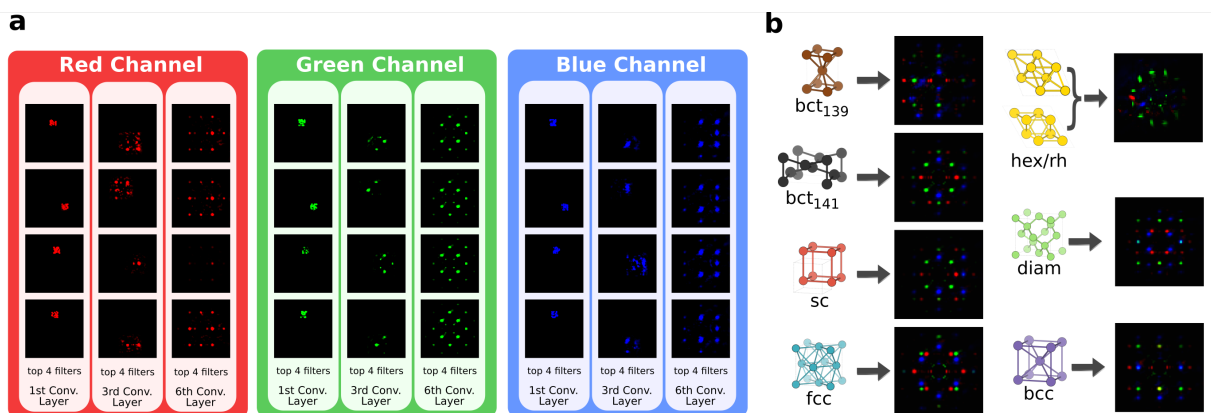


Figure 3.5: Visualizing the convolutional neural network (ConvNet) attentive response maps. (a) Attentive response maps from the top four most activated filters of the first, third and last convolutional layers for the simple-cubic class. The brighter the pixel, the most important is that location for classification. Comparing across layers, we notice that the ConvNet filters are composed in a hierarchical fashion, increasing their complexity from one layer to another. At the third convolutional layer, the ConvNet discovers that the diffraction peaks, and their relative arrangement, are the most effective way to predict crystal classes. (b) Sum of the last convolutional layer filters for all seven crystal classes: the ConvNet learned crystal templates automatically from the data.

The top four most activated (i.e. most important) filters from the first, third and last convolutional layers for each of the three color channels are shown in Fig. 3.5a for the sc class. The complexity of the learned filters grows layer by layer, as demonstrated by the increasing number of diffraction peaks spanned by each motif. The sum of the last convolutional layer filters for each class is shown in Fig. 3.5b; they are class templates automatically learned from the data by the ConvNet. From the figure, there are interesting observations that can be made:

- Along the lines of hypothesis and what has been observed in the past, the network builds the unique features through hierarchy of feature across multiple layers for identifying crystal structure.
- Certain channels are more responsive to the stimuli than others.

Comparing Fig.3.2c and 3.5b, we see that our deep learning model is able to autonomously learn, and subsequently use, the same features that a domain expert would use. This not only confirms the soundness of the classification procedure, but also explains its robustness in terms of generalization.

3.4 Discussions

We have introduced a way of representing crystal structures by means of (easily interpretable) images. Being based on reciprocal space, this descriptor termed two-dimensional diffraction fingerprint compactly encodes crystal symmetries, and possesses numerous attractive properties for crystal classification. In addition, it is complementary with existing real-space based representations [7], making possible to envision a combined use of these two descriptors. Starting from these diffraction fingerprints, we use a CNN to predict crystal classes. As a result, we obtain an automatic procedure for crystal classification which does not require any user-specified threshold, and achieves perfect classification even in the presence of highly defective structures. In this regard, we argue that since materials science data are generated in a relatively controlled environment defective datasets represent probably the most suitable test to probe the generalization ability of any data-analytics model. Given the solid physical grounds of the diffraction fingerprint representation, our deep learning model is modest in size, which translates in short training and prediction times. Finally, using recently developed visualization techniques, we uncover the learning process of the neural network. Thanks to its multi-layered architecture, we demonstrate that the network is able to learn, and then use in its classification decision the same landmarks

a human expert would use. Further work is needed to make the approach proposed here unique across space groups and to widen its domain of applicability to non-centrosymmetric crystals, which can exhibit technologically relevant ferroelectric, piezoelectric or nonlinear optical effects. In accordance with the principle of reproducible research, we also provide an online tutorial [134] where users can interactively reproduce the main results of this work (but also produce their own) within the framework of the NOMAD Analytics-Toolkit. As an outlook, our method could also be applied to the problem of local micro-structure determination in atomic probe tomography experiments, with the ultimate goal of discovering structural-property relationships in real materials.

3.5 Experiments - Human Anatomy Classification

Automated classification of human anatomy is an important pre-requisite for many computer aided diagnosis systems. The changing complexity and variability of anatomy throughout the human body makes it an arduous task to classify various anatomical parts. Recently, deep learning methods, in particular CNNs have shown to outperform other methods in such tasks. Hence, in this study, we use CNN based architecture for automated human anatomy classification.

In order to understand the decision making process of deep CNNs in the medical domain and to construct an informed approach to designing models for medical diagnosis, we first build three different deep CNN models with different architectures and hyper-parameters: a shallow CNN, a deeper CNN without data augmentation and a deeper CNN with data augmentation inspired by the work of Razavian et al. [89]. The network architecture for each model is depicted in Fig. 3.7. After successfully training the above mentioned networks, we examine which part of a particular input image from an anatomy class, particularly the spatially distributed information, is used in the decision process of the CNN. It is done by visualizing attentive response maps from the top n most *activated* units of the last convolutional layer in the above described models, similarly to Bau et. al. [8] and as formulated in the Formulation section of the Chapter. The top n units are used to visualize the parts of the input image that the network considers *important*. The formation of attentive response maps are done by projecting the top unit activations back to image space. The back projection to input space is achieved by using the fractionally strided convolution, also known as the transposed convolution, and sometimes incorrectly termed the deconvolution technique [129] as shown in Fig. 3.6. To explain the formulation for the formation of attentive response maps, let us consider a multi-layered neural network with n layers.

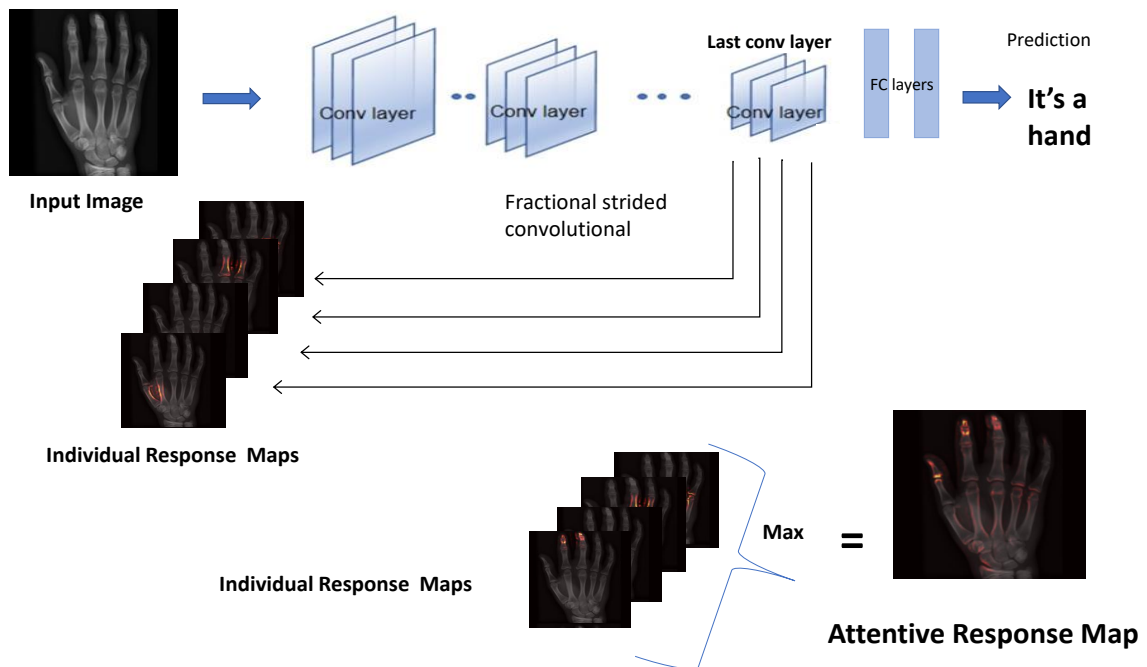


Figure 3.6: overview of the proposed visualization framework for understanding and visualizing human anatomy prediction. As an illustrative case in this study, a new x-ray anatomy image is fed into the convolutional neural network to obtain the prediction (in this case, body-part classification) and through the fractionally strided convolution, individual attentive response maps are computed through top- n units from the last convolutional layer in the network. By computing the max operation per pixel across all the individual maps, we obtain the attentive response map. The attentive response map shows: 1) The locations in the input image that are contributing to decision making and 2) the level of dominance of such locations.

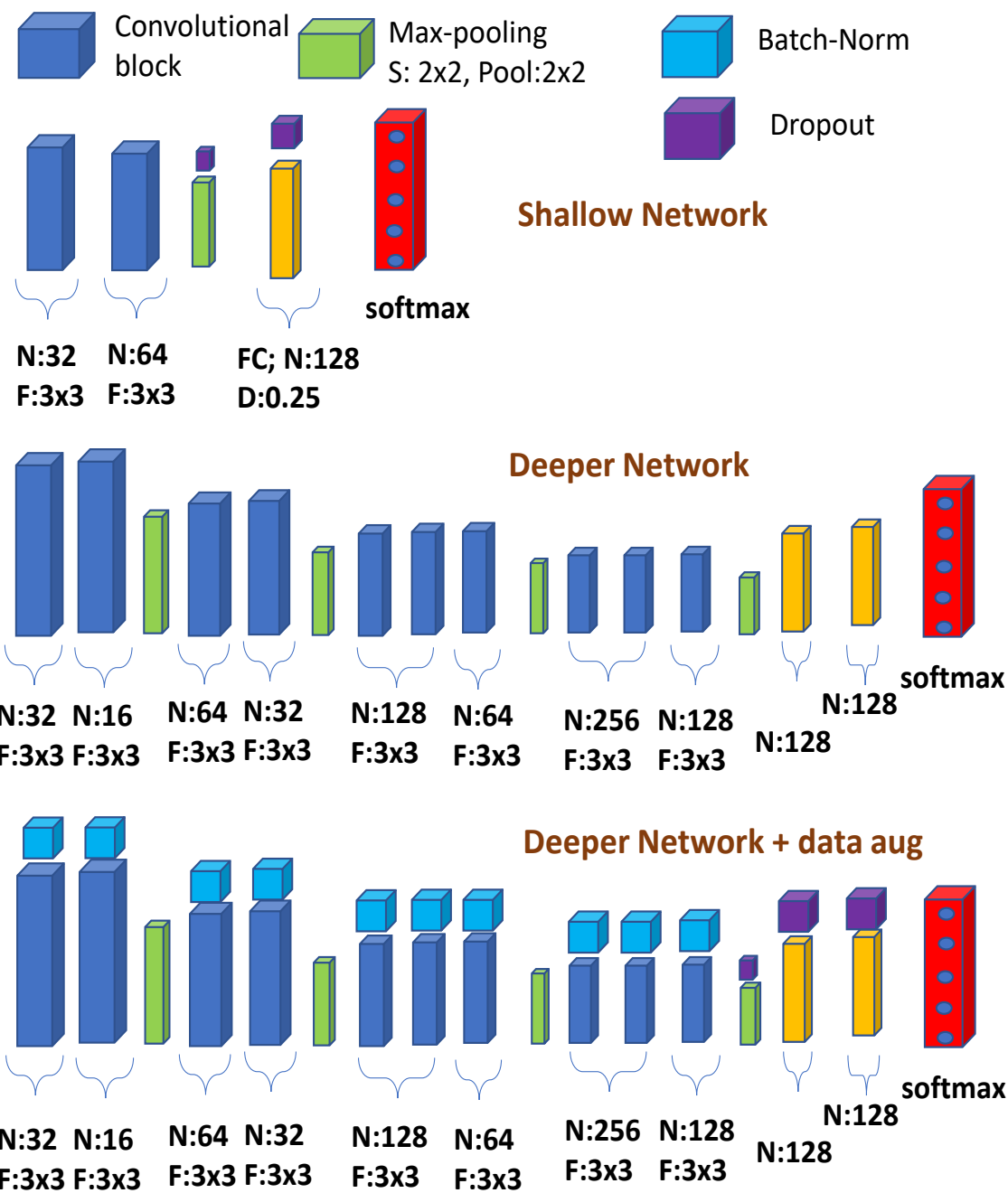


Figure 3.7: Architecture of three different CNNs with different capacities that were used in this human anatomy classification study. The experiment was done to show what different kind of features are used for making predictions by different network architectures.



(a) Anatomy description of foot found in literature, unique bone structures pertaining to class are indicated.



(b) Foot X-ray image from ImageCLEF dataset.



(c) Attentive response map from top 25 filters from last conv layer of network overlaid on original image



(d) Anatomy description of hand found in literature, unique bone structures pertaining to class are indicated.



(e) Hand X-ray image from ImageCLEF dataset.



(f) Attentive response map from top 25 filters from last conv layer of network overlaid on original image

Figure 3.8: Correspondence between anatomical descriptions found in the literature that are used by human experts ((a) & (d)) and the attentive response maps overlaid on the original images ((b) & (e)) from the last conv layer of the deeper network with data augmentation ((c) & (f)) for the foot and hand class. It can be observed that the deeper neural network uses the same landmarks as a human expert for anatomy classification. Best viewed in color.

3.5.1 Dataset Explanation

To visualize and understand the decision making of a deep neural network, we used anatomy classification from X-ray images as an example use-case. To train our three different convolutional neural networks, radiographs from the ImageClef 2009 Medical Image Annotation task ¹ were used. This data set consists of a wide range of X-ray images for clinical routine, along with a detailed anatomical classification. For uniform training without any bias, we removed the hierarchical class representation and removed the classes consisting of less than 50 examples. Using this, we ended up with 24 unique classes e.g. foot, knee, hand, cranium, thoracic spine etc., from the full body anatomy.

3.5.2 Experiment Design & Results

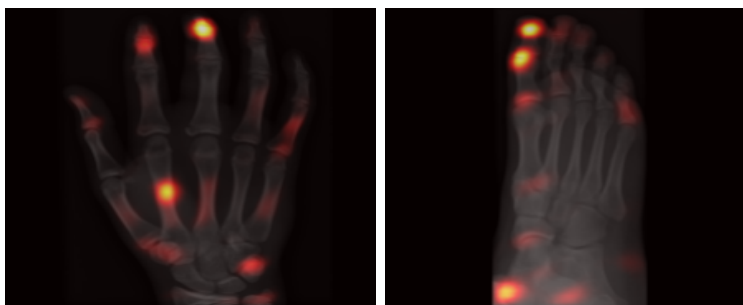
For training the three networks described in Fig. 3.7, we resized the images to 224×224 . For evaluation, we divided the ImageClef dataset (14,676) images into randomly selected training and test sets with 90 % and 10 % of the data respectively. For the third (deeper) network specifically, we used various data augmentation techniques ranging from cropping, rotation, translation, shearing, stretching and flipping. We trained the three networks for all the 24 classes simultaneously. The results obtained by training the three models are shown in Table 3.3.

Table 3.3: Results: Accuracy in percent for three different networks trained on the ImageClef 2009 annotation task

Shallow Net	Deeper Net	Deeper Net+data aug
71.1	90.36	95.62

We visualized the internal activations of the models on test data through attentive response maps. More specifically, we combined the attentive response maps of the top $n = 25$ units from the last convolutional layer and overlaid them on the original image. In this way we constructed the focused attentive response maps that can be easily examined by a human expert. The $n = 25$ was chosen empirically as it produced attentive response maps closer to the anatomical landmarks with least number of units. The results are shown in 3.8, 3.9, 3.10 and 3.11 for foot, hand and knee classes from ImageClef dataset. In the end, we also conducted an ablation study by only keeping the unique bone structures identified through attentive response maps for various human anatomy classes and passing

¹<http://www.imageclef.org/2009/medanno>



(a)

Figure 3.9: Attentive response maps overlaid on the original images from the last conv layer of the deeper network with no data augmentation for foot and hand class. It can be observed that this network fails to use the same landmarks as a human expert for anatomy classification, as shown in Fig. 3.8: (a) & (c). Best viewed in color.



Figure 3.10: Focus area of the top 5 attentive response maps from top 5 most activated units from last conv layer of the shallow network. For clarity, instead of top 25 only top 5 units are shown separately. It is evident that the network doesn't learn any medically relevant landmarks. Best viewed in color.

the images again for classification through CNN. An example of this is shown in Fig. 3.12. In Fig. 3.12, even after removing rest of structures and only keeping part of cuneiforms and metatarsals bones, the trained CNN is able to still classify the human anatomy class correctly.

3.6 Discussions

Through the attentive maps, we first examine the correlation of those regions obtained through visualizing the dominant attentive response maps with identified regions and shapes of image landmarks that are mentioned in the medical radiology literature. With the qualitative assessment, we can establish that the same landmarks that are described in the medical image literature are also used by the CNN. For example, in Fig. 3.8, 3.9, 3.10 and 3.11, we observe that the particular outlines of bones are used to detect the human anatomy part in the image rather than some background information. In Fig. 3.11, we can observe the obtained attentive response map (part (b)), and the medically relevant landmarks from medical literature outlining the unique bone structures for a knee (part (c) & (d)). Observing these images, we can conclude that the CNN is identifying the unique bone structures for knee, namely the neck of tibia and fibula bones and the joint. These are the main identifying bone structures for a knee. Hence, it can be established that CNN is identifying a knee through medically relevant landmarks.

We use this to guide the decisions for the model architecture and learning algorithm. We can furthermore use this method to detect biases or limitations of the models. In certain examples of mis-classification (Fig. 3.13), we can observe that the information used for making decisions is part of an artifact rather than the object in the image. From Fig. 3.8, we know that the trained CNN model identifies hand by the unique bones structures of Phalanges (specifically Distal Phalanx and Interphalanges joints). As these landmarks are missing in Fig. 3.13, the CNN model mis-classifies this particular examples. Hence, through attentive response maps we can identify the key landmarks for certain classes and its limitations in the absence of such landmarks. This understanding can inform us about the possible adjustments to the pre-processing of data augmentation procedures needed to remove the bias from the model or adjust for the prediction being made.

In Fig. 3.8, we show a correspondence between the obtained attentive response maps and the anatomical landmarks from the medical literature ² for the anatomy of hand and

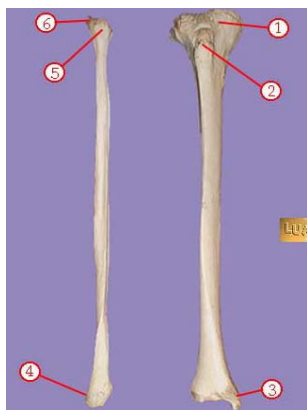
²http://www.meddean.luc.edu/lumen/meded/radio/curriculum/bones/Structure_Bone_teach_f.htm



(a) Original Knee X-ray Image



(b) Attentive Response Maps from top 4 nodes



(c) Knee unique landmarks



(d) Knee landmarks on actual x-ray

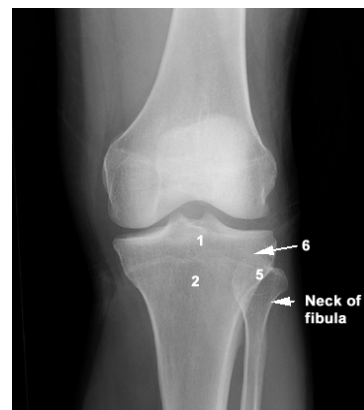


Figure 3.11: Illustration of the effectiveness of the attentive response maps in highlighting key medically relevant landmarks for knee class. (a) original image, (b) shows attentive response map for correctly identified original image, and (c) represents the key bone landmarks points in knee from medical literature. (d) presents the key landmarks on an actual knee x-ray. Observing (b) and (d), we can see how attentive response maps highlighting the key landmarks (points 1, 2, 5 & 6) related to the neck of tibia and fibula bones in a knee. Thus, proving that the CNN model is relying on relevant medical landmarks for making predictions.

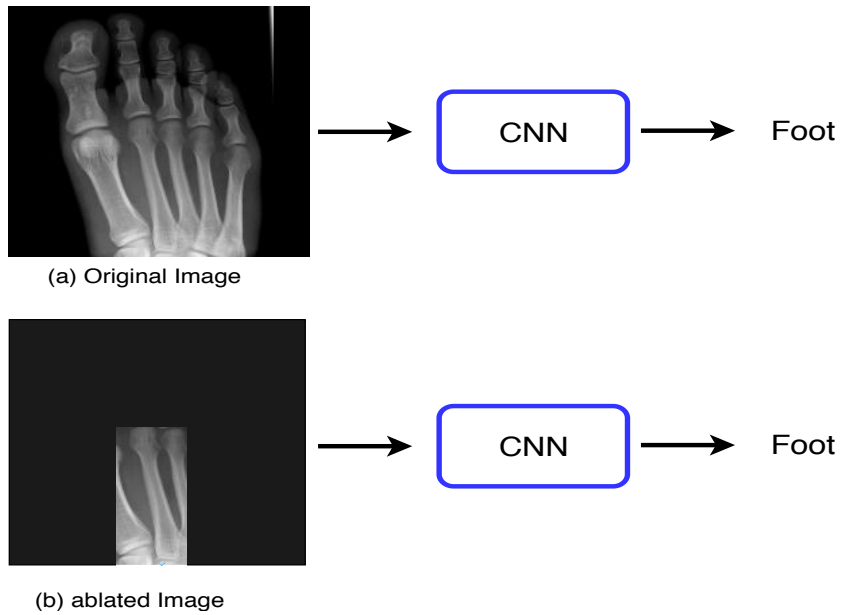


Figure 3.12: Illustration of ablation study conducted to test the effectiveness and efficacy of the key landmarks identified by attentive response maps. In (b) we can observe that even after removing rest of the bone structures other than some part of the cuneiforms and metatarsals bones, the network is still able to identify the human anatomy correctly. Cuneiforms and metatarsals were identified as key landmarks by attentive response maps earlier and are indeed medically unique bones in foot. Hence, this proves the network is relying on the key bone structure landmarks for making predictions.

foot. Particularly for the foot image, we can observe that the edges of the metatarsals' shaft has been used together with the distal phalanges, navicular, cuboid, tibia, and fibula. Similarly for the hand, three of the distal phalanges, many of the heads of joints, metacarpals' shafts as well certain carpals. In contrast to this, in Fig. 3.9 and Fig. 3.10 we can observe that the shallow and deep network trained without specific data augmentation fails to learn such specific landmarks. These models use broader ranges that are clearly not as specific as the information used in the first model. From the above visual results as well as the performance of the final model we come to the conclusion that sufficiently deep neural network models can be successfully trained to use the same medical landmarks as a human expert while attaining superior performance.

To summarize, the following observations can be made:

- Attentive response maps are useful for highlighting if the trained model is able to use medically relevant landmarks or not for its predictions. This can eventually guide network architecture design and training procedure.
- Attentive response maps can help in identifying key features associated with a class.
- Attentive response maps can also potentially help in detecting bias and limitations of trained models. This can assist the end user in understanding when and why there can be a failure case.

3.7 Summary

In the domain knowledge experiments specially via the human anatomy classification, we show that the design of the model architectures for deep CNN and the training procedure does not necessarily need to be a trial-and-error process, solely focused on optimizing the test set accuracy. Through attentive response map visualization, we managed to incorporate domain knowledge and overall managed to achieve a much more informed decision process, which finally resulted in a model with superior performance. This approach is applicable to many different image analysis applications of deep learning that are unable to easily leverage the potentially large amount of available domain knowledge. Furthermore, visually understanding the information involved in the model decision allows for more confidence in its performance on unseen data.

In the aforementioned formulation and experimental studies in the present Chapter, some of the shortcoming of the current visualization based methods were solved through

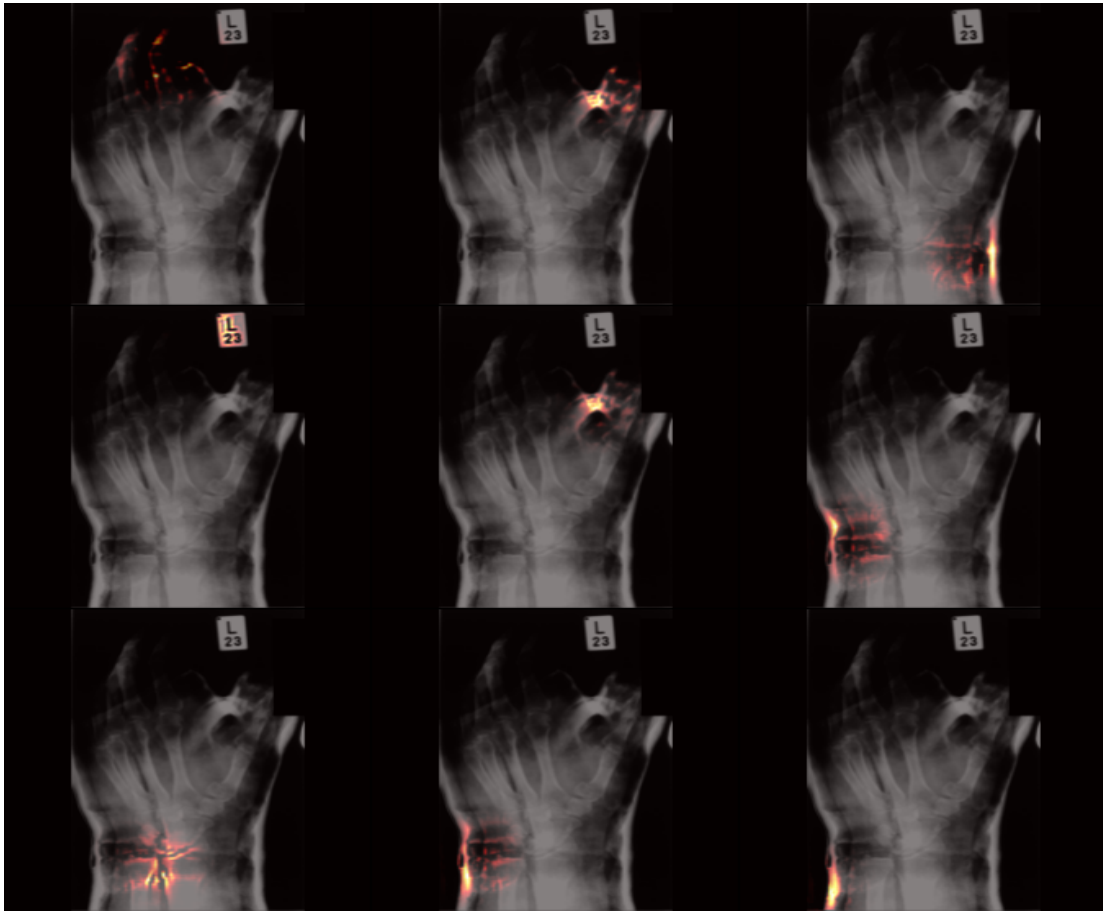


Figure 3.13: Individual attentive response maps for top 9 activated units from the last conv layer of the deeper network with augmentation for the hand class example, mis-classified as cranium. From the figure, it is evident that the top 9 most activated units are focusing on the wrong information present in the signal. Best viewed in color.

attentive response maps. However, similar to the previous methods, the above described attentive response maps methods wasn't end-to-end. Here, end-to-end refers to the scenarios where explanations are formed from the last layer of neural networks, just before the prediction layer. To build trust in human experts using such systems and for providing effective explainability, especially for critical sectors such as health-care, these methods have to be end-to-end. The next Chapter of the thesis aims to solve precisely this shortcoming of the current methods including attention based methods. In Chapter 4, we present an end-to-end approach and architectural design for forming the attention response maps. The following study is presented through forming and explaining the decision made by deep radiomics sequencers, that act as computer aided diagnostics (CAD) system for identifying malignancy of lung nodules.

Chapter 4

End To End Interpretable Architecture Design

“Interpretability is a domain-specific notion, so there cannot be an all-purpose definition. Usually, however, an interpretable machine learning model is constrained in model form so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity, or physical constraints that come from domain knowledge.”

- Adrian Colyer, Venture Partner, Accel.

Prologue to Article

4.0.1 Article Details

SISC: End-to-end Interpretable Discovery Radiomics-Driven Lung Cancer Prediction via Stacked Interpretable Sequencing Cells, D. Kumar*, V. Sankar*, D. A. Clausi, G. W. Taylor, A. Wong, *IEEE Access Journal*, 2019

Personal Contribution In this particular study, we created a process for obtaining attention maps end-to-end. This work was a joint work with Vignesh Sankar. I designed the study, and network architecture and wrote the interpretability code. I also conducted the majority of the experiments and obtained the results. Vignesh processed the dataset and conducted some additional experiments and co-wrote the manuscript with me.

4.0.2 Context

The SqueezeNet architecture introduced the concept of the “fire module”. Inspired by the architecture of the fire module, we wanted to introduce a similar modular based architecture that can produce state-of-the-art results while remaining end-to-end interpretable. Hence, similar to the fire module, in this work we introduce the concept of a sequencing cells-based (SISC) architecture. We choose to use lung nodule malignancy as the case study for the SISC architecture, as to build upon my previous work in this domain.

4.0.3 Contribution

In this part of the thesis, we present an end-to-end based architectural design for forming attention response maps. We introduce a SISC architecture comprising of an interpretable sequencing cells module, for building models with state-of-the-art performance while incorporating interpretability within each module. We also form critical response maps generated through a stack of interpretable sequencing cells which highlight the key critical regions leveraged in the prediction process.

4.1 Introduction

Chapter 3 proposed how attentive response maps can be used to form an interpretability response as attentive regions for any convolutional layers in a given CNN. The Chapter then extended the proposed framework by relating the obtained attentive response maps to human domain knowledge. However, similar to other current approaches, the proposed framework also lacks end-to-end explanations. Therefore, there is still a need to have an end-to-end framework.

To address this issue, in this Chapter of the thesis, we propose a unique *sequencing cell* based architectural design to make the whole process end-to-end. Overall, this Chapter forwards a framework that can enable human domain experts to observe end-to-end explanations on a layer wise (including the last layer) basis instead of just on a per neuron level and also enables them to see how the given evidence for making a prediction relates to particular domain knowledge. The end-to-end architecture’s efficacy in producing state-of-the-art interpretable results is proved by modeling the end-to-end architecture in the form of radiomics sequencer for detailed experimentation on a medical domain study of lung nodule malignancy.

4.1.1 Motivation: End-to-End Deep Radiomics Sequencers

In recent times, radiomic sequencers with deep convolutional architectures (discovery radiomics) that are discovered in an end-to-end manner were shown to consistently achieve state-of-the art performance in medical imaging analysis. The use of such radiomic sequencers within the discovery radiomics framework is particularly effective for lung cancer prediction using CT imaging. This is due to the availability of very large annotated CT scan data sets such as LIDC-IDRI [4], which enables highly discriminative radiomic features to be discovered directly from this wealth of data.

Despite the effectiveness of discovery radiomics-based approaches from a diagnostic performance perspective, a key challenge that still remains is the difficulty in interpreting the rationale behind their predictions. As such, one can view such radiomics-based approaches as ‘black box’, and the lack of transparency in their decision-making processes makes it difficult for radiologists to verify, validate, and ultimately trust the predictions being made. To enable the widespread adoption of discovery radiomics within CAD systems, one needs to improve radiologists’ trust by providing interpretable reasoning behind the predictions made by radiomic sequencers.

Also, it is important to note here that the previous interpretable techniques as shown in previous Chapter and in the past, were not end-to-end. These method only worked up to the final convolutional layers. As such, we need to have end-to-end visualization of the attentive response maps for accurate diagnosis and interpretability and eventually to build higher confidence and trust.

Motivated by this, we propose an end-to-end interpretable discovery radiomics-driven lung cancer prediction pipeline for the binary prediction case. Specifically, the main contributions of our approach are:

- the introduction of the SISC architecture (Fig. 4.1), comprising interpretable sequencing cells, for building radiomic sequencers with state-of-the-art performance for lung cancer prediction, and
- interpretable lung cancer predictions in the form of critical response maps generated through a stack of interpretable sequencing cells which highlight the key critical regions leveraged in the prediction process (Fig. 4.2).

4.2 Methodology

This section describes the design and implementation of a radiomics sequencer. The modular design of the SISC architecture (see Section 4.2.2) enables a significant reduction in the design search space while improving classification performance. Furthermore, the introduction of interpretable sequencing cells allow us to achieve end-to-end interpretability through the generation of critical region maps to aid the clinician in the decision-making process (see Section 4.2.3). The LIDC-IDRI dataset (see Section 4.2.1) is used to validate the proposed radiomic sequencer architecture.

4.2.1 Dataset Explanation- LIDC

The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) [4] published a structured and categorized repository of computed tomography (CT) scans to assist the development of CAD methods for automated lung cancer diagnosis. The dataset consists of 1018 thoracic CT scans, where each scan is processed by four radiologists at both blinded and un-blinded stages. In the blinded stage, each radiologist reviews the CT scans without inputs from other radiologists. In the second, un-blinded stage, each

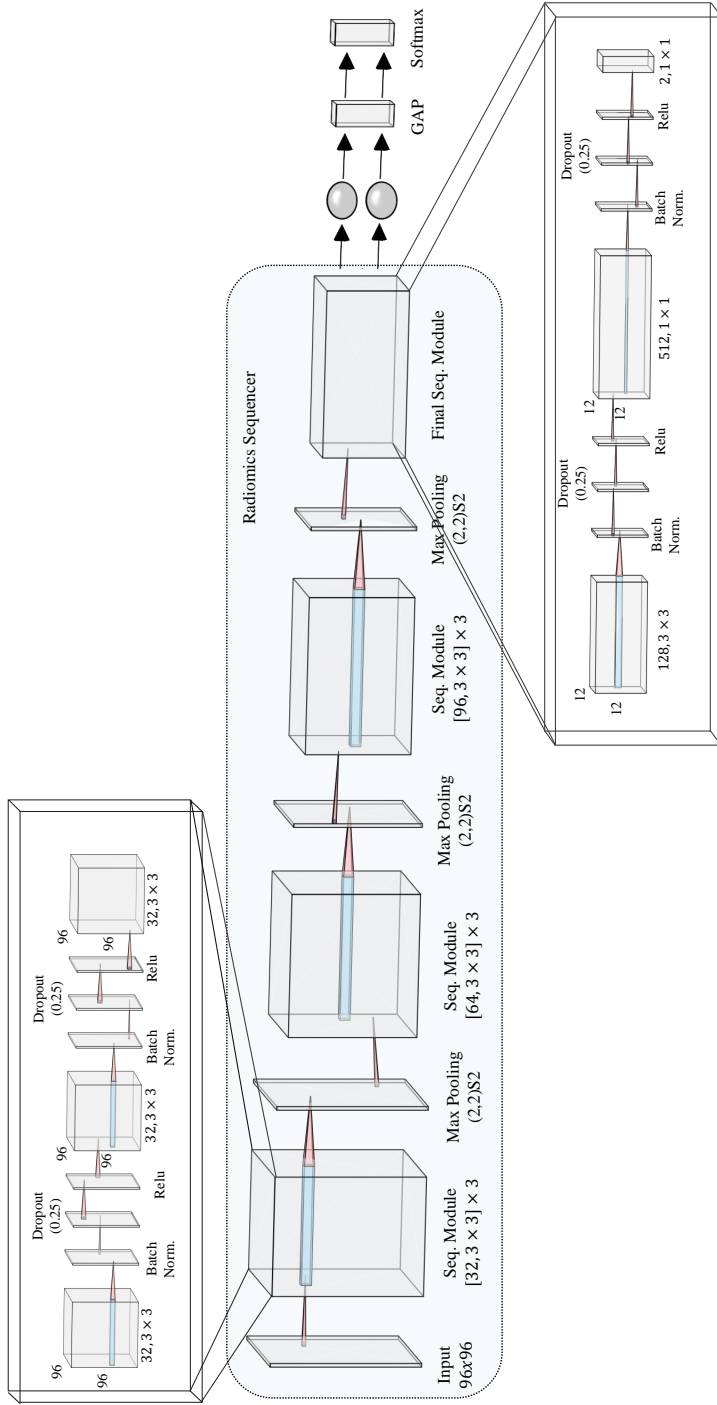


Figure 4.1: Overview of the proposed deep stacked interpretable sequencing cell (SISC) architecture used as the radiomic sequencer within a discovery radiomics framework. The SISC radiomic sequencer is formed by stacking interpretable sequencing cells together, each comprised of different specialized convolutional layers along with max-pooling and dropout operations. A typical interpretable sequencing cell consists of a block of convolutional layers followed by batchnorm, dropout, and ReLU layers, repeated three times. The final interpretable sequencing cell is flexible and can be changed based on the input and task. Here, the final layer in the last cell consists of two 1×1 convolutions as this study is focused on binary lung cancer classification.

radiologist is shown the results of other radiologists from the blinded stage and is given a chance to change their initial evaluations. The two stage process was designed to provide the best estimate of the nodule characteristics.

The suspected lung lesions in the LIDC dataset are divided into three categories: i) Non-nodule ≥ 3 mm, ii) Nodule ≥ 3 mm, and iii) Nodule < 3 mm, where the diameter is measured as the length of the lesion’s longest axis. For each category, different nodule characteristics are included in the dataset. Similar to previous methods, we decided to only use Nodule ≥ 3 mm. For the Nodule ≥ 3 mm category, the required malignancy score along with the nodule location and contour information by at least one radiologist are included. Therefore, the Nodule ≥ 3 mm category is used for our experiments. A total of 2,669 nodules are reported in the dataset under the Nodule ≥ 3 mm category.

For each nodule, its characteristics are provided by at most four radiologists. The final malignancy score for each nodule is obtained by combining the scores from all of the radiologists. As suggested in [45], the average score rounded to the nearest integer was taken as the final malignancy score. In each slice of the given nodule, a 96×96 pixel window was cropped at the nodule center. The size was determined to accommodate for all the variability in the nodule contour and also to include sufficient background information. The same malignancy score was assigned for all the slices in the given nodule. A total of 14,433 nodule images along with their corresponding malignancy score were extracted from the dataset.

4.2.2 Interpretable Sequencing Cells

A modular design strategy was leveraged to construct the proposed SISC radiomics sequencer, where the underlying architecture is comprised of a deep stack of interpretable sequencing cells with similar micro-architectures. More specifically, an interpretable sequencing cell as introduced in this study comprises a block of convolutional layers along with max-pooling and dropout operations, all optimized using the available data. The proposed SISC radiomic sequencer is then constructed by stacking the interpretable sequencing cells together in a depth-wise manner. The aim is to reduce the design search space while improving classification accuracy, thus enabling optimized design of sequencer architectures in a more predictable manner. The micro-architecture of an interpretable sequencing cell is defined by three convolutional layers separated by batch normalization [51] and drop-out [113]. Furthermore, the ReLU [36] activation is used after each convolutional layer. The interpretable sequencing cell is optimized by sharing the same architectural values. For example, all the dropout layers in a given interpretable sequencing cell have the same dropout rate.

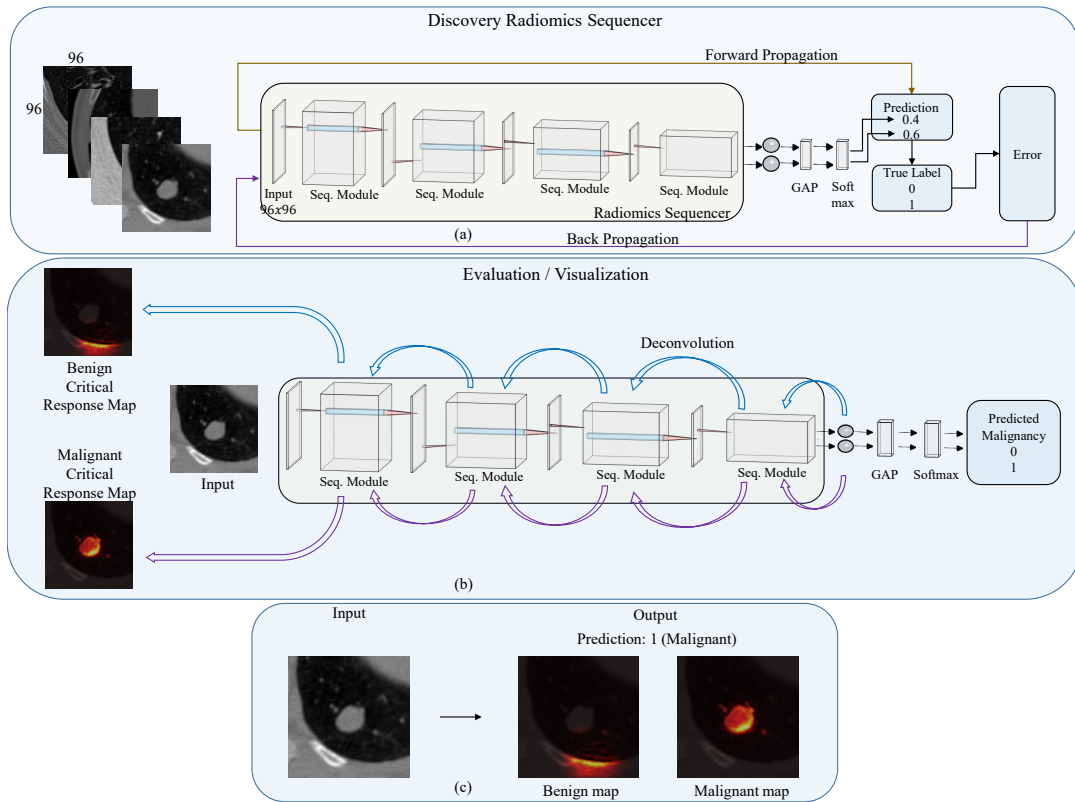


Figure 4.2: Overview of the end-to-end interpretable discovery radiomics-driven framework for lung cancer prediction. Part (a) shows the sequencer discovery process, where a specialized radiomic sequencer, comprised of a deep stack of interpretable sequencer cells, is discovered for the given set of CT lung nodule data. Part (b) presents the cancer prediction process, where the discovered radiomic sequencer is used to make a prediction based on CT data and how interpretable critical response maps are generated through the stack of sequencer cells. In part (b), the input CT data of a new patient is fed into the radiomic sequencer to generate a radiomic sequence and perform prediction on whether it is a benign and malignant case. To generate the critical response map, the output of the last layer in the sequencing cell of the radiomic sequencer is back-propagated through each sequencing cell using the method described in Section 4.2.3 for each of the possible prediction states (benign and malignant). As such, we obtained two critical response maps, each highlighting the critical regions used by the sequencer for making predictions regarding whether the given input nodule is benign or malignant. The last part (c) is the interface seen by the end user, which shows the given input, the prediction and evidence used to obtain the particular prediction through critical response map.

In this study, the proposed SISC radiomic sequencer is comprised of four interpretable sequencing cells stacked together in a depth-wise manner as shown in Fig 4.1. The number of channels is increased as we go deeper into the SISC architecture whereas the size of the kernel is fixed for the first three cells. Each cell is followed by a max pooling layer. The dropout, batch normalization parameters and number of cells are optimized with the LIDC-IDRC dataset. The final cell of the SISC radiomic sequencer is defined as shown in Fig. 4.1. The number of kernels in the final convolutional layer of the final cell is equivalent to the total number of classes to enable end-to-end interpretability via critical response maps, which will be further discussed in the following section. The final convolutional layer is followed by a global average pooling layer, which is then followed by a softmax output layer.

The proposed SISC radiomic sequencer formed by the stacking of interpretable sequencing cells with learned parameters is shown in Fig 4.1. We can observe that the repeated modular approach allows us to compactly define the sequencer with a minimum number of configurable architectural parameters. The modular approach also leads to state-of-the-art results as described in Section 4.3.

4.2.3 Interpretability Through Critical Response Maps

To enable interpretability and explainability in the decision-making process of the proposed SISC radiomic sequencer, we take inspiration from [67] and [66] and introduce an approach where critical response maps are generated through the entire stack of interpretable sequencing cells. An critical response map provides spatial insights on critical regions in the CT scan and their level of contribution to a particular prediction made. Here, an individual critical response map is generated for each possible prediction (benign and malignant).

Using these critical response maps, the clinician can not only validate the the evidence behind the predictions made using the proposed SISC radiomic sequencer, but the maps also help in locating relevant regions in the CT scan responsible for either a malignant or benign nodule prediction. An example pair of critical response maps can be seen in Fig. 3.6.

For example, in the case of a malignant nodule, a successful and reasonable prediction should lead to the malignant critical map highlighting the nodule regions. As such, critical response maps may potentially help radiologists to have greater confidence in the CAD system and in aiding them with their clinical diagnosis decisions.

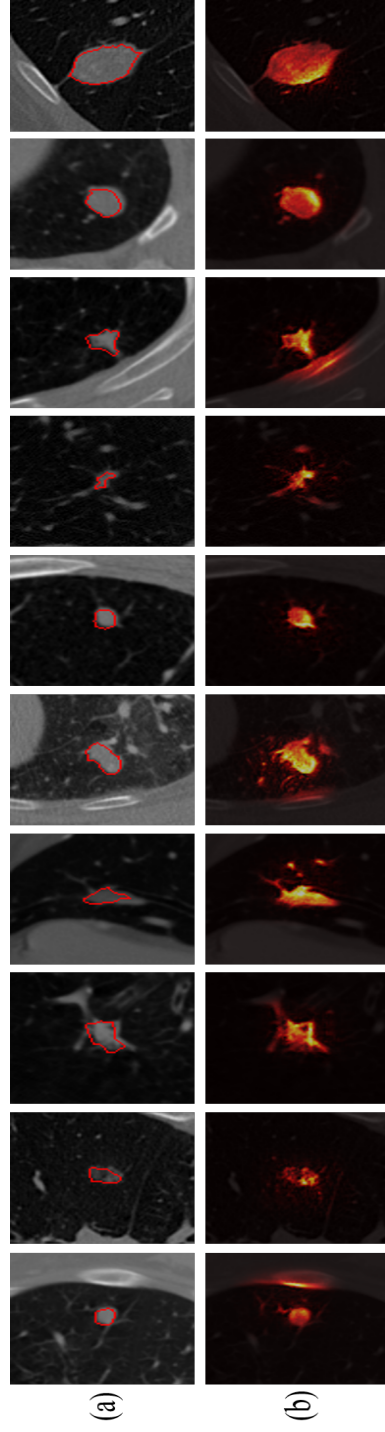


Figure 4.3: Example critical response maps for malignant cases. (a) original 96×96 pixels malignant nodule sub-image taken from lung CT slices with the radiologist-provided best contours for a given patient CT slice image, and (b) corresponding critical response maps showing the malignant critical regions been used for correctly predicting malignant nodules. It can be seen that for almost all the example cases, the proposed SISC radiomic sequencer uses clinically relevant markers when achieving correct predictions. Therefore, the use of critical response maps can potentially improve the overall confidence of the clinician on the discovered SISC radiomic sequencer.

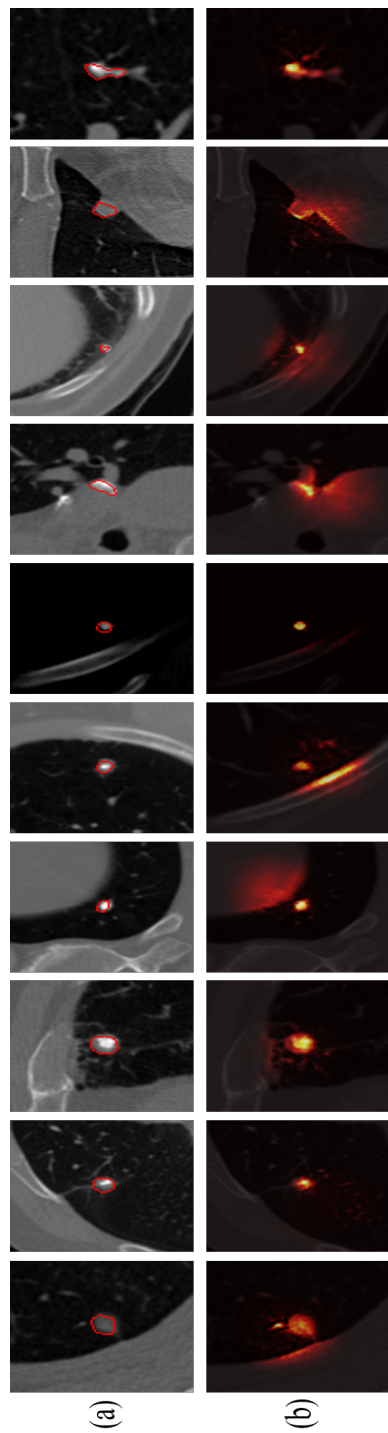


Figure 4.4: Example critical response maps for benign cases. (a) original 96×96 pixels malignant nodule sub-image taken from lung CT slices with the radiologist-provided best contours for a given patient CT slice image, and (b) corresponding critical response maps showing the benign critical regions used for correctly predicting benign nodules. It can be seen that for almost all the example cases, the proposed SISC radiomic sequencer uses clinically relevant markers when achieving correct predictions. Therefore, as with the previous figure, the use of critical response maps can potentially improve the overall confidence of the clinician on the discovered SISC radiomic sequencer.

The critical response map generation process can be described as follows. Let the critical response maps $A(x|c)$ for a given CT scan slice image x for each prediction c be computed via back-propagation from the last layer of the last interpretable sequencing cell in the proposed SISC radiomic sequencer. The notation used in this study are based on Chapter 3 for consistency. As shown in Fig 4.1, the last layer in the interpretable sequencing cell at the end of the proposed SISC radiomic sequencer contains $N = 2$ nodes, equal to the number of possible predictions (i.e., benign and malignant). The output activations of this layer are followed by global average pooling and then a softmax output layer. So, to create the critical maps for each possible prediction, the back-propagation starts with the individual prediction nodes in the last layer to the input space. For a single layer l , The deconvolved output response \hat{h}_l is given by,

$$\hat{h}_l = \sum_{k=1}^K f_{k,l} * p_{k,l}. \quad (4.1)$$

where f_k is the feature map from the k_{th} convolutional filter. K is the total number of filters in the layer l and p_l is the kernel of layer l . the symbol $*$ represents the convolution operator. For simplicity, The convolution and summation can be combined as $\hat{h}_l = D_l f_l$. Therefore, the critical response map $C(x|c)$, for a given prediction c is defined as,

$$C(x|c) = D_1 U'_1 D_2 U'_2 \dots D_{L-1} U'_{L-1} D_L^c F_L. \quad (4.2)$$

Where U' is the un-pooling operation as described in [130] and D_L^c is the convolution operation at the last layer with kernel p_L replaced by zero except at the c^{th} location corresponding to the prediction c .

4.3 Experiments And Results

In this section, we will evaluate and discuss the efficacy of the proposed SICS radiomic sequencer for the purpose of lung cancer prediction on two main fronts: i) cancer prediction performance of the proposed sequencer compared to state-of-the-art, and ii) interpretability of the cancer predictions made by the proposed sequencer through the generated critical response maps.

Table 4.1: Number of samples for corresponding malignancy scores, for three different datasets. The datasets were created based on how the radiologist malignancy rating 3 was treated i.e., Ignored (I), treated as benign (B), or treated as malignant (M).

Malignancy score	Size	I	B	M
1	1376	3981	9638	3981
2	2605			
3	5657	Ignored		
4	3192	4795	4795	10452
5	1603			

4.3.1 Experimental Setup

The setup of the experiments in this study can be described as follows. The cropped lung nodule images with their corresponding malignancy scores are obtained from the LIDC-IDRI dataset as explained in section 4.2.1. The distribution of the malignancy scores in the dataset is described in Table 4.1. Malignancy scores 1 and 2 are considered as benign, whereas scores 4 and 5 are considered as Malignant. The malignancy score 3 can be either considered as benign, malignant or ignored depending on how it was done by previous studies in this field. In this study, we have created three different datasets where the score 3 is considered malignant (dataset: ‘M’), benign (dataset: ‘B’), and ignored (Dataset: ‘I’). The final dataset distribution is shown in Table 4.1. Each dataset was further divided into 80% training data, 10% for validation and 10% testing data. The pre-processing and dataset distribution is similar to Xie et. al. [126].

Table 4.2 shows the distribution of training data for the three datasets. We can observe that dataset ‘B’ and ‘M’ are not evenly distributed. To mitigate this imbalance, data augmentation was performed on each of the datasets to balance the number of examples associated with each class. Furthermore, the data augmentation performed also acts to enhance the variability and generalizability of the radiomic sequencer. In particular, ran-

dom horizontal shifts, vertical shifts, and rotations were applied along with vertical and horizontal flips to construct the augmented training dataset. Since the size of the training data has a huge impact on the performance of the proposed radiomic sequencer, three different augmented datasets with varying size were created for each of the ‘M’, ‘B’ and ‘I’ datasets, as shown in Table 4.2. The performance of dataset ‘I’ under different levels of data augmentation is shown in Fig 4.9. We can observe that the $\approx 15k$ size yielded the best performance. Going forward, we have finalized the dataset size to be $\approx 15k$ for further hyper-parameter tuning and validations for the ‘M’, ‘B’ and ‘I’ datasets.

Different data normalization techniques such as standard deviation, Min-Max, and ZCA whitening [60] were also applied to the lung nodule images. It was found that Min-Max normalization yielded the best results. After optimizing the hyper-parameter using the validation set, batch normalization was implemented with momentum= 0.99 and a dropout layer was used with rate= 0.25. The Adam optimizer was used with learning rate = $1e^{-5}$ and batch size as 128. The proposed radiomic sequencer was learnt for approximately 200 epochs and evaluated against the test dataset. The final results were reported by averaging over 10-fold cross validation for all three datasets.

4.3.2 Cancer Prediction Performance

To evaluate the cancer prediction performance of the proposed SISC radiomic sequencer, we computed the sensitivity, specificity, accuracy, and AUC-ROC of the proposed sequencer and compared it with five other state-of-the-art approaches. The average lung cancer prediction performance of the proposed radiomic sequencer for dataset ‘M’, ‘B’ and ‘I’ are shown in Table. 4.5 & 4.4 & 4.3 respectively. The AUC curves of the 10 different cross validation runs for each dataset are shown in Figs. 4.5-4.7. The best performing AUC curve from each dataset is shown in Fig. 4.8. From the results, we can observe that, for the dataset ‘I’, by leveraging the proposed SISC radiomic sequencer, we were able to achieve comparable performance with the current state-of-the-art method proposed by Xie et al. [126]. For datasets ‘B’ and ‘M’, the proposed sequencer is able to outperform the accuracy results from Xie et al. [126]. The comparison of the existing and current state-of-the-art methods with the proposed SISC radiomic sequencer is given in Tables 4.3- 4.5. The comparison methods are based on the previous studies in this field, employing the same data pre-processing steps for fair comparison amongst the methods. Based on these experimental results, it can be observed that the proposed SISC radiomic sequencer can provide strong cancer prediction performance that exceeds state-of-the-art in all but one case, where in that case the performance is comparable to state-of-the-art.

Table 4.3: Performance comparison between tested cancer prediction methods for the ignored (I) dataset. Best results are highlighted in **bold**.

Method	Accuracy	Sensitivity	Specificity	AUC
Han et al [45]	85.59	70.62	93.02	89.25
Dhara et al [22]	88.38	84.58	90.03	95.76
Shen et al [105]	87.14	77.00	93.00	93.00
Sun et al [117]	-	-	-	88.23±1.70
Xie et al [126]	89.53±0.09	84.19±0.09	92.02±0.01	96.65±0.01
Ours (SISC)	89.36±1.20	90.28±2.00	88.25±2.00	96.01±0.70

Table 4.2: Dataset distribution for the three different dataset configuration obtained from the LIDC-IDRI dataset before and after data augmentation (as described in Section 4.3.1).

Dataset	Grade	Before Data Aug	15k	30k	60k
I	Malignant	3836	9836	15836	30836
	Benign	3184	9184	15184	30184
	Total	7020	19020	31020	61020
B	Malignant	3836	7672	19180	38360
	Benign	7712	7712	21712	35712
	Total	11548	15384	40892	74072
M	Malignant	8361	8361	16361	33444
	Benign	3187	6374	15935	28683
	Total	11548	14735	32296	62127

Table 4.4: Performance comparison between tested cancer prediction methods for the benign (B) dataset. Best results are highlighted in **bold**.

Method	Accuracy	Sensitivity	Specificity	AUC
Han et al [45]	87.36	73.75	93.37	93.79
Dhara et al [22]	87.69	80.00	89.30	94.44
Xie et al [126]	87.74±0.03	81.11±0.85	89.67±0.09	94.45±0.01
Ours (SISC)	88.57±1.70	78.32±8.12	93.66±2.06	94.34±0.08

Table 4.5: Performance comparison between tested cancer prediction methods for the malignant (M) dataset. Best results are highlighted in **bold**.

Method	Accuracy	Sensitivity	Specificity	AUC
Kumar et al [63]	75.01	83.35	-	-
Han et al [45]	70.97	53.61	89.41	76.26
Dhara et al [22]	71.17	53.47	89.74	79.74
Sharma et al [102]	84.13	91.69	73.16	-
Xie et al [126]	71.93±0.04	59.22±0.04	84.85±0.10	81.24±0.01
Ours (SISC)	84.17±1.50	90.71±4.01	67.00±8.64	89.06±1.20

4.3.3 Discussion

Here, we will investigate the efficacy of the proposed SISC radiomic sequencer in terms of interpretability of the lung cancer predictions made. Fig. 4.3 shows example critical response maps generated in an end-to-end manner for several example malignant nodule

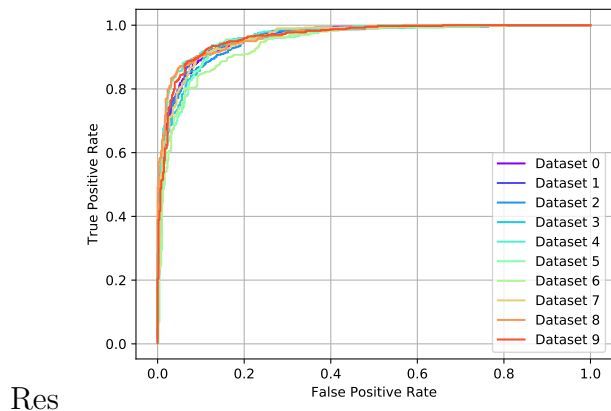


Figure 4.5: Receiver operating curve (ROC) for the ignored (I) dataset for 10 different cross validation runs.

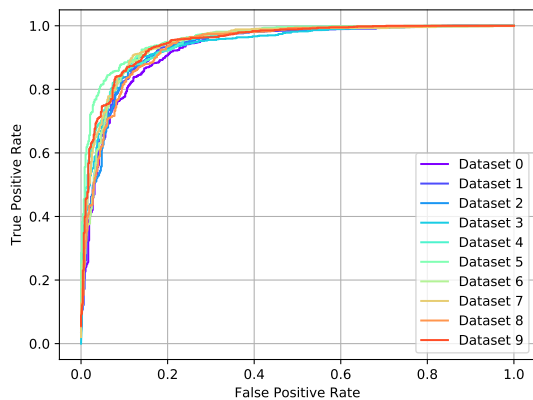


Figure 4.6: Receiver operating curve (ROC) for the benign (B) dataset for 10 different cross validation runs.

images that were correctly predicted to be malignant. From the figure, it can be observed that the proposed SISC radiomic sequencer is able to successfully identify the nodule regions in the given CT slices without being explicitly directed to do so. As shown in Fig. 4.3, the highlighted region’s contour closely matches the contours given by the radiologists, and in some cases provide improved contour localization than that provided by the radiologists. The proposed SISC radiomic sequencer is able to successfully highlight a

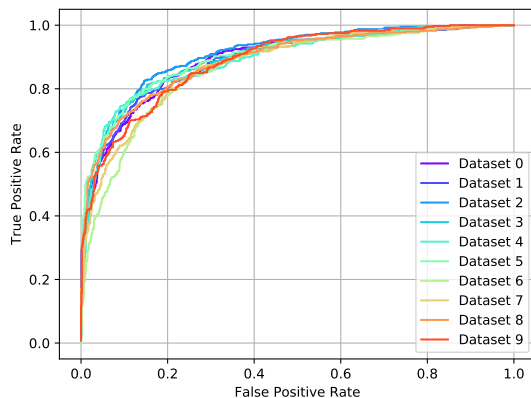


Figure 4.7: Receiver operating curve (ROC) for the malignant (M) dataset for 10 different cross validation runs.

wide range of nodules with different shapes and sizes. We can also infer the discriminative nature of the proposed sequencer by observing that the highlighted regions in the critical response maps that contribute highly to a malignancy prediction. This helps to gain better insight in the rationale behind the malignancy prediction. Similar observations can also be observed for benign nodules as shown in Fig. 4.4.

Due to the end-to-end interpretable nature of the proposed SISC radiomic sequencer, the critical response maps produced through the entire stack of interpretable sequencing cells can potentially help improve the confidence of radiologists working with a CAD system. Furthermore, the critical response maps can also assist radiologists to more consistently and rapidly spot abnormal nodules within the large volume of a CT scan, as well as understand the nature and characteristics most linked to malignancy.

4.4 Summary

In this Chapter, we introduce a novel end-to-end interpretable discovery radiomics-driven lung cancer prediction framework. This framework is enabled by the proposed radiomic sequencer: a deep stacked interpretable sequencing cell (SISC) architecture comprised of interpretable sequencing cells. Experimental results show that the proposed SISC radiomic sequencer is able to not only achieve state-of-the-art results in lung cancer prediction, but also offers prediction interpretability. Interpretability is offered in the form of critical

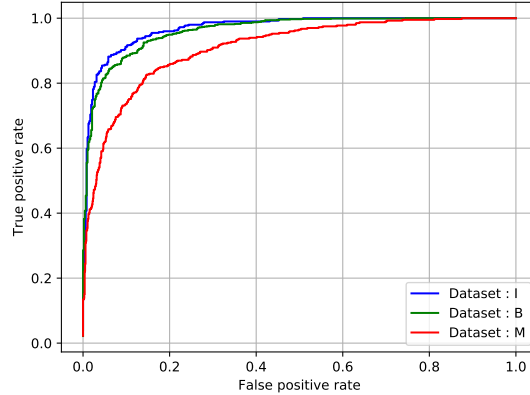


Figure 4.8: Comparing the best ROC curves for the three different datasets: ignored (I), benign (B) and malignant (M).

response maps generated through the stack of interpretable sequencing cells. The sequencing cells highlights the critical regions used by the sequencer for making predictions. The critical response maps are useful for not only validating the predictions of the proposed SISC radiomic sequencer, but also provide improved radiologist-machine collaboration for improved diagnosis.

However, there are certain scenarios which still remain unexplained. In the next Chapter, we will explore some of the examples where the current binary attention maps fail to produce explanations and why there is a need to move from binary class approach to a multi class approach. To this effect, we introduce and explain a new class-enhanced attentive response (CLEAR) maps and show their efficacy in explaining scenarios that cannot be explained from simple binary attention maps.

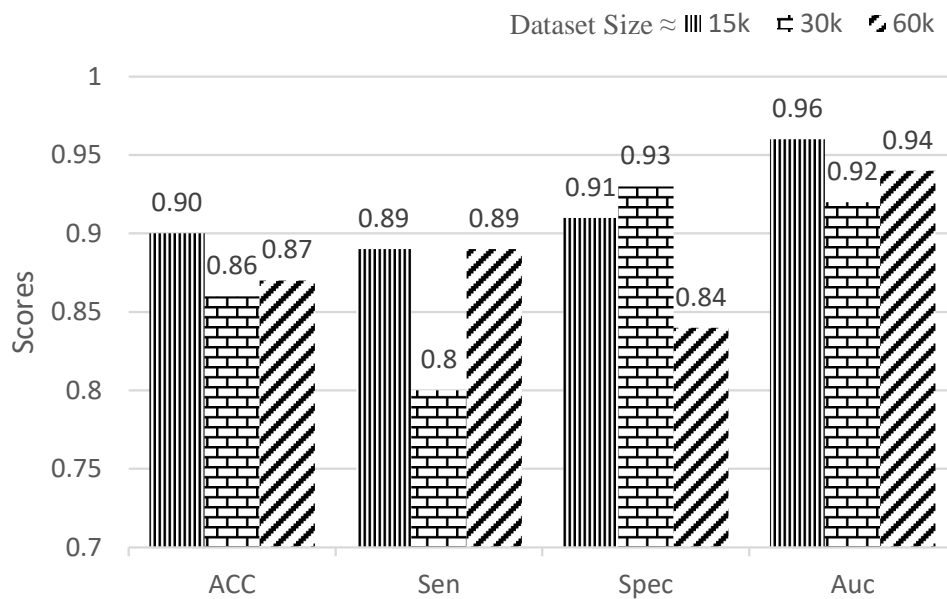


Figure 4.9: Accuracy, sensitivity, and specificity for the “ignored” (I) dataset for three different training sample sizes. It can be observed that the $\approx 15k$ size performed the best for 3 out of 4 metrics, including AUC; hence it was chosen as the default training size for the three different dataset categories: ignored (I), benign (B), and malignant (M).

Chapter 5

CLass-Enhanced Attentive Response (CLEAR)

“Second, much like banks are required by law to “know their customer,” engineers that build systems need to know their algorithms. For example, Eric Haller, head of Datalabs at Experian told us that unlike decades ago, when the models they used were fairly simple, in the AI era, his data scientists need to be much more careful. “In the past, we just needed to keep accurate records so that, if a mistake was made, we could go back, find the problem and fix it,” he told us. “Now, when so many of our models are powered by artificial intelligence, it’s not so easy. We can’t just download open-source code and run it. We need to understand, on a very deep level, every line of code that goes into our algorithms and be able to explain it to external stakeholders.””

- Greg Satell, Harvard Business Review

Prologue to Articles

This Chapter derives contributions from the below mentioned two papers.

5.0.1 Details Of Articles

Explaining the unexplained: A class-enhanced attentive response (CLEAR) approach to understanding deep neural networks, D. Kumar, A. Wong, G. W. Taylor, *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, 2017

Discovery Radiomics With CLEAR-DR: Interpretable Computer Aided Diagnosis of Diabetic Retinopathy D. Kumar, G. W. Taylor, A. Wong, *IEEE Access Journal*, Vol. 7, 2019

Best Paper Award at Transparent and Interpretable Workshop at 30th Neural Information Processing (NIPS), 2017

Personal Contribution The idea of using multi-class enhanced maps instead of single node or binary heatmaps came up during the discussions with Alex regarding extending our attentive response maps work. I first proposed multi-class maps and Alex suggested to use HSV space for representation. The idea was further refined through extensive discussions with Graham and Alex. I implemented the CLEAR formulation and trained all the convolutional neural network models for three different datasets namely MNIST, SVHN & Stanford dog dataset. Through discussions with Graham & Alex, I devised the controlled experiments and obtained quantitative and qualitative results for evaluating the efficacy of CLEAR maps in providing explanations for scenarios that were previously unexplained via binary heatmaps.

5.0.2 Context

The goal of CLEAR was to overcome the shortcomings introduced by the various heatmap visualization methods and the binary heatmap approach introduced by Zintgraf et. al (ICLR 2017). Our CLEAR approach builds on the previous work introduced by Zeiler et. al. (ICCV 2014) and our attentive response map method (explained in Chapter 3). In Zeiler et. al. the authors use their method to devise responses for individual nodes only for a single class as

attention in input space. CLEAR however, does layer-wise multi class end-to-end attentive response in the input space.

5.0.3 Contributions

All of the heatmap visualization methods (such as CAM, Grad-CAM etc.) only highlight the regions of attention and provide no meaning to the assignments other than that they should form a coherent set of interpretable pixels. In **CL**ass-**E**nhanced **A**ttentive **R**esponse (CLEAR), we provide an approach that not only provides the relevant attentive regions but also provides meaning to the constituent pixels by assigning class-based information for each individual pixel. This allows for explaining really complex scenarios that cannot be explained via node-based visualization.

Also, Gondal et.al. [37] leveraged CAM maps to highlight the lesion areas for diabetic retinopathy; however, this approach provides no interpretation of grading information and thus is limited in providing clinical insight on grading decisions. Motivated by the need for clinical interpretability, we proposed **CLEAR-DR**, a novel interpretable CAD system based on the notion of **CL**ass-**E**nhanced **A**ttentive **R**esponse **D**iscovery **R**adiomics for the purpose of clinical decision support for diabetic retinopathy. CLEAR-DR not only generates discriminative radiomic sequences for making grading decisions for diabetic retinopathy as a use case, but also visually interprets and understands these decisions via information back-propagation. The back-propagation is done through the discovered radiomic sequencer by embedding the CLEAR approach. This process is designed to enable grade-level interpretability and can also help in reducing inter-observer and intra-observer variability while speeding up the overall diagnostic process.

5.0.4 Recent Developments

During the development of CLEAR, there was a lack of feature attribution approaches that were end-to-end methods. Therefore, we had to primarily rely on the fully convolutional neural network architectures. However, recently a few approaches (SHAP, Deep Taylor Decomposition etc.) have been proposed that can provide feature attributions in an end-to-end manner for networks with fully connected layers as well. This eliminates the need to rely on fully convolutional network for providing end-to-end explanations and paving the way for the use of CLEAR to be used with any type of network architecture via recent back-propagation based feature attribution approaches.

The CLEAR article has gained a total of 29 citations to date (based on Google Scholar).

5.1 Introduction

As noted in Chapters 2 and 3, much of the recent works have focused on understanding the decision-making process of networks by leveraging heatmaps that provide information about which areas of the image are used by the deep CNN (DCNN) to make a particular decision. These approaches have produced some promising results in revealing what is important to a decision made by a DCNN. More details regarding the relevant works are provided in Chapter 2. However, a common limitation with such heatmap-based approaches to understanding the decision-making process of DCNNs is that of **decision ambiguity**, where one can gain insight into **which** regions of interest are important for making decisions, but gives no insight as to **why** such regions of interest are important. As a result, these methods leave the “thought process” of the DCNN largely ambiguous. An example of this is shown in Fig. 5.1.

In this Chapter, we attempt to mitigate the problem of this particular decision ambiguity, and take a step towards “explaining the unexplained”, with regards to the decision-making process of DCNNs, through the introduction of **CLass-Enhanced Attentive Response** (CLEAR) maps that go beyond what existing heatmap-based approaches [5, 81, 135] can provide. The proposed CLEAR maps allow for the visualization of not only the attentive regions of interest and corresponding attentive levels of DCNNs during the decision-making process, but also the corresponding dominant classes associated with these attentive regions of interest. As such, compared to heatmaps, CLEAR maps are much more effective at conveying where and why certain regions of interest influence the decision-making process. An example of this is shown in Fig. 5.5. We further demonstrate the effectiveness of the proposed CLEAR maps, both quantitatively and qualitatively, by conducting a number of experiments using three different publicly available datasets.

This section explains the procedure for generating the proposed **CLass-Enhanced Attentive Response** (CLEAR) maps. The main goal of CLEAR maps is to convey the following information:

- the attentive regions of interest in the image responsible for the decision made by the CNN;
- the attention levels at these regions of interest so that we understand their level of influence over the decision made by the CNN;
- the dominant class associated with these attentive regions of interest so that we can better understand **why** a decision was made.










Input		Output	Heatmap	Interpretation
	→  →	3 ✓		Focuses on right areas: Looks correct!
	→  →	2 ✗		Focuses on wrong part, curve might be two; but why not 3 or 5 or 6?
	→  →	3 ✗		Probably focuses on correct part, but why 3?

Figure 5.1: Examples of handwritten digits from MNIST are shown, along with: 1) the decision made by the CNN, 2) heatmaps used in existing visualization methods, and 3) what can be interpreted based on the heatmaps. While the heatmaps used in existing approaches show which information in the image works for (positive focus: hot regions) or against (negative focus: green), it is evident that the heatmaps are insufficient to fully interpret and explain the decision made by the CNN.

5.1.1 Methodology Overview - Class Enhanced Attentive Response (CLEAR)

The procedure for generating CLEAR maps can be summarized as follows (see Fig. 5.2). First, individual attentive response maps are computed for each kernel associated with a class by back-projecting activations from the output layer of the DCNN. Based on this set of attentive response maps, two different types of maps are computed: 1) a dominant attentive response map, which shows the dominant attentive level for each location in the image; and 2) a dominant class attentive map, which shows the dominant class involved in the decision-making process at each location. Finally, the dominant attentive response map and the dominant attentive class map are combined visually by using color and intensity

to produce the final CLEAR map for a given image.

Inspired by the effectiveness of the previously introduced architectural design in Chapter 3 and ALL-CNN [112] on different datasets, we leveraged a similar network architecture for building the DCNN used for classification in this part of the thesis. While, for clarity, we describe the procedure for computing individual attentive response maps based on the previously defined architecture, the procedure will generalize to other DCNNs provided class-specific responses can be computed in input (pixel) space. The network is composed primarily of convolutional, ReLU, and max-pooling layers. Towards the output of the DCNN, the last convolutional layer contains a set of kernels equal to the number of classes, and then global averaging is performed before passing these energy values to the softmax output layer which represents categories. As such, each kernel can be thought of as being associated with a particular class.

5.1.2 Formulation

The first step of CLEAR is to compute a set of individual attentive response maps, one for each of the classes learned by the DCNN, which we will denote as $\{R(\underline{x}|c)|1 \leq c \leq N\}$, where N is the number of classes. This is achieved in the current realization of CLEAR by back-propagating the responses of each kernel in the last convolutional layer from feature space to the input space to form each attentive response map, thus extending upon the idea presented in Chapter 3. To explain the formulation for the formation of CLEAR maps, first consider a single layer of a DCNN. Let \hat{h}_l be the deconvolved output response of the single layer l with K kernel weights w . The deconvolution output response at layer l then can be then obtained by convolving each of the feature maps z_l with kernels w_l and summing them as:

$$\hat{h}_l = \sum_{k=1}^K z_{k,l} * w_{k,l}. \tag{5.1}$$

Here $*$ represents the convolution operation. For notational brevity, we can combine the convolution and summation operation for layer l into a single convolution matrix G_l . Hence the above equation can be denoted as: $\hat{h}_l = G_l z_l$.

For multi-layered DCNNs, we can extend the above formulation by adding an additional un-pooling operation U as described in [130]. Thus, we can calculate the deconvolved output response from feature space to input space for any layer l in a multi-layer network as:

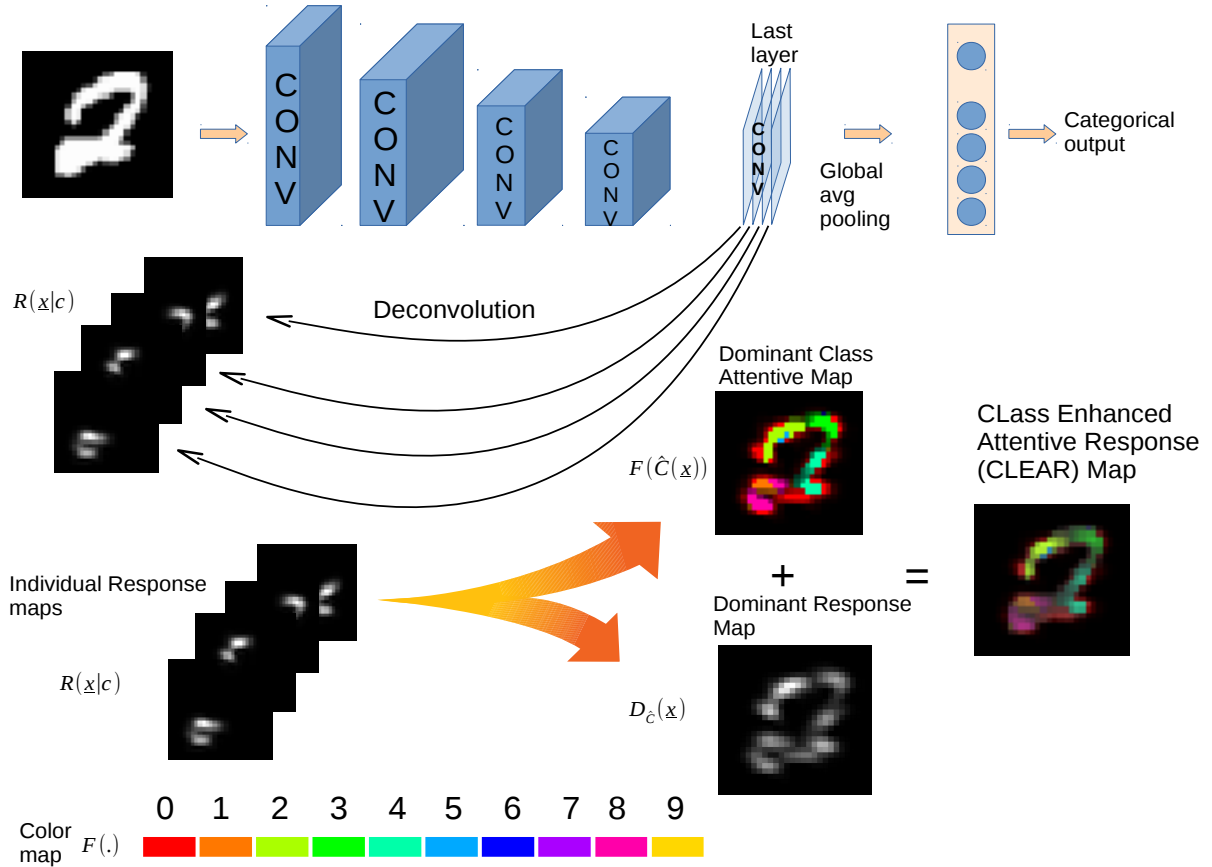


Figure 5.2: The procedure for generating **C**lass-**E**nhanced **A**ttentive **R**esponse (CLEAR) maps. First, individual attentive response maps are computed for each class based on the last layer of the DCNN. Based on this set of attentive response maps, two different types of maps are computed: 1) a dominant attentive response map, which shows the dominant attentive level for each location in the image, and 2) a dominant class attentive map, which shows the dominant class involved in the decision-making process at each location. Finally, the dominant attentive response map and the dominant attentive class map are combined to produce the final CLEAR map for a given image.

$$R_l = G_1 U_1 G_2 U_2 \dots G_{l-1} U_{l-1} G_l z_l. \quad (5.2)$$

For CLEAR maps, we specifically calculate the output responses from individual kernels of the last layer of a network. Hence, given a network with last layer L containing $K = N$ kernels, we can calculate the attentive response map; $R(\underline{x}|c)$ (where \underline{x} denotes the response back-projected to the input layer, and thus an array the same size as the input) for any class-specific kernel c ($1 \leq c \leq N$) in the last layer as:

$$R(\underline{x}|c) = G_1 U_1 G_2 U_2 \dots G_{L-1} U_{L-1} G_L^c z_L. \quad (5.3)$$

Here G_L^c represents the convolution matrix operation in which the kernel weights w_L are all zero except that at the c^{th} location.

Given the set of individual attentive response maps, we then compute the dominant attentive class map, $\hat{C}(\underline{x})$, by finding the class at each pixel that maximizes the attentive response level, $R(\underline{x}|c)$, across all classes:

$$\hat{C}(\underline{x}) = \underset{c}{\operatorname{argmax}} R(\underline{x}|c). \quad (5.4)$$

Given the dominant attentive class map, $\hat{C}(\underline{x})$, we can now compute the dominant attentive response map, $D_{\hat{C}}(\underline{x})$, by selecting the attentive response level at each pixel based on the identified dominant class, which can be expressed as follows:

$$D_{\hat{C}}(\underline{x}) = R(\underline{x}|\hat{C}). \quad (5.5)$$

To form the final CLEAR map, we map the dominant class attentive map and the dominant attentive response map in the HSV color space as follows, then transform back into the RGB color space:

$$\begin{aligned} H &= F(\hat{C}(\underline{x})), \\ S &= 1, \\ V &= D_{\hat{C}}(\underline{x}). \end{aligned} \quad (5.6)$$

Here $F(\cdot)$ is the color map dictionary that assigns an individual color to each dominant attentive class, c . Fig. 5.2 shows an example of the CLEAR map overlaid on the image.

In this Chapter we present results that illustrate the efficacy of CLEAR maps for understanding and interpreting the decision-making of CNNs. For this, we used image datasets from two different domains:

- Generic Image Domain

- Medical Image Domain

For the generic image domain experiments, we conducted qualitative and quantitative experiments on three different datasets: the commonly used benchmarks MNIST and Street View House Numbers (SVHN) dataset. Qualitative experiments on medical image domain were conducted using the retinal fundus images from diabetic retinopathy dataset [53]. In the following section, we explain individually the experimental setup and the results obtained for each domain.

5.2 Experiments- Generic Image datasets

5.2.1 Dataset Explanations - MNIST & SVHN

This section explains the two generic image datasets used in this part of the study.

MNIST

The Modified National Institute of Standards and Technology (MNIST) dataset is a collection of handwritten digits that is commonly used for the training and testing of new machine learning algorithms. The training and testing set for MNIST were formed by taking a larger NIST dataset and mixing its training & testing set in half and half proportion. In total, MNIST dataset contains 60,000 images in the training set and 10,000 images in testing set. In total, there are 10 classes, one for each digit 0 to 9. All images are grey level of size 28×28 .

SVHN

The Street View House Numbers (SVHN) dataset is a real world dataset of digits, similar to MNIST. The dataset was collected from the house numbers via Google Street View images. The images are cropped to a size of 32×32 , slightly larger than MNIST and is significantly harder as it the images come from real world problem i.e., recognizing digits and numbers in natural scene images especially as it has cases where there are multiple digits in the same image as shown in the first column of Fig. 5.4. In total, there are 10 classes of images for numbers 0 to 9 i.e., one for each digit. Digit '1' has label 1, '9' has label 9 and '0' has label 10. In total there are 73257 digits images for training, 26032 digits for testing, and 531131 additional, somewhat less difficult samples, to use as extra training data if required.

Table 5.1: Architecture of our CNN used for MNIST Classification.

Conv Layer	(3x3, 32x)
Conv Layer	(3x3, 32x)
Conv Layer	(3x3, 32x)
Conv Layer	(3x3, 32x)
MaxPool Layer	(2x2, 2x2 stride)
Conv Layer	(3x3, 64x)
Conv Layer	(1x1, 10x)
Global average pooling	(10)
Softmax	(10)

5.2.2 Experiments Design

Setup

To conduct experiments on three different datasets, we trained three different CNN architectures with all convolutional layers. For training on MNIST and SVHN, we set our network architecture similar to [112], as it has shown to perform very effectively for a variety of datasets. To train these networks, we used the default train and test split. We achieved an accuracy of 99.26% and 92.6% for the MNIST and SVHN datasets, respectively.

In both networks, as the last layer (convolutional layer) was linearly connected to the softmax activation function, each kernel can be considered to represent one separate class. It is important to note that the aim was to understand and interpret the decision of a trained network; hence we did not strive to achieve the best architecture and state-of-the-art results for each dataset. Using the previously mentioned setup, we conducted the following experiments.

5.2.3 Qualitative Results

In this set of experiments, we first create binary heatmaps and the proposed CLEAR maps for individual images in the three different datasets. The binary heatmaps represent which information in the image was used for or against the true class versus other image classes during classification. The binary heatmaps were formed by overlaying the output response from the kernel representing the true class as “hot” regions and response of the rest of

Table 5.2: Architecture of our CNN used for SVHN Classification.

Conv Layer	(3x3, 32x)
Conv Layer	(3x3, 32x)
Conv Layer	(3x3, 32x)
MaxPool Layer	(2x2, 2x2 stride)
Conv Layer	(3x3, 64x)
Conv Layer	(3x3, 64x)
Conv Layer	(3x3, 64x)
MaxPool Layer	(2x2, 2x2 stride)
Conv Layer	(3x3, 128x)
Conv Layer	(1x1, 128x)
Conv Layer	(1x1, 10x)
Global average pooling	(10)
Softmax	(10)

the kernels in the last layer, represented by green regions. The response for the rest of the kernels is formed by performing max operation across the individual output responses. Thus, in the binary heatmaps, the hot regions and green regions represent the information for and against the actual class respectively, that was used for decision-making by the network. The binary heatmaps are constructed similarly to [135] and [81]. The CLEAR map formation is explained in Section 5.1.2 and Fig. 5.2.

For the SVHN, we also create an additional binary map. This map replaces the varying values in the binary heatmaps with a constant value. In the binary map, red and blue regions represent the information used for and against the class, respectively. We create these maps for visual clarity, as sometimes it is harder to visualize the green regions in the binary heatmaps.

Some of the randomly chosen results for the MNIST dataset are shown in Fig. 5.3. This figure shows examples of correctly classified and misclassified examples by the network. From these results, observations that can be made are: 1) Looking at the example sets for digit 0, although positive support is contributed by the same bottom curved features in both examples, only in one case is the image correctly identified as zero. Looking at the CLEAR maps, we can see the dominant activations for the correctly classified example corresponds to class 0, whereas for the misclassified case they correspond to class 5. 2) Similarly, for digit 7 and 8 it is difficult to interpret the decision output of the CNN, but

looking at the CLEAR maps make them more interpretable.

5.2.4 SVHN

Presented similarly to the MNIST dataset, results obtained for the SVHN dataset are shown in Fig. 5.4. Some interesting observations are as follows: 1) For the misclassified 0 digit, the heatmap overwhelmingly focuses on the correct curves; but the network still misclassifies it. This is counter-intuitive to human interpretation. But when observing the CLEAR maps, we see that almost all the strong activations are for classes other than 0. 2) For the digit 9, it is difficult to interpret the binary heatmaps, as the positive kernel focuses on the digit 1, but it still correctly classifies the digit as 9 with high confidence. Observing the CLEAR maps, we see that most of the dominant activation in the focus areas belong to digit 9, including the ones for digit 1.

5.2.5 Quantitative Results

To re-validate our observations for the MNIST and SVHN datasets, we conducted two different quantitative experiments. In the first experiment, we removed all parts of the image, except for regions responsible for the activations of the kernel associated with the class of the image (*positive kernel*). We call these regions *strong features* associated with the class. For the MNIST dataset, we replace the digit with the background and for the SVHN dataset, we replace the region with a gray patch.

Table 5.3: Evaluation to re-validate the effectiveness and contribution of identified strong features on accuracy.

Accuracy(%)	MNIST	SVHN
Full image	99.26	92.60
with only strong features	79.89	69.12
without strong features	43.45	54.46

In the second experiment, we do the opposite: we remove the regions responsible for the kernel associated with the true class of the input image and keep the rest of the image. Results are shown in Table 5.3, and demonstrate that the identified strong features are vital for correctly classifying a particular class. For the case where the network is still able

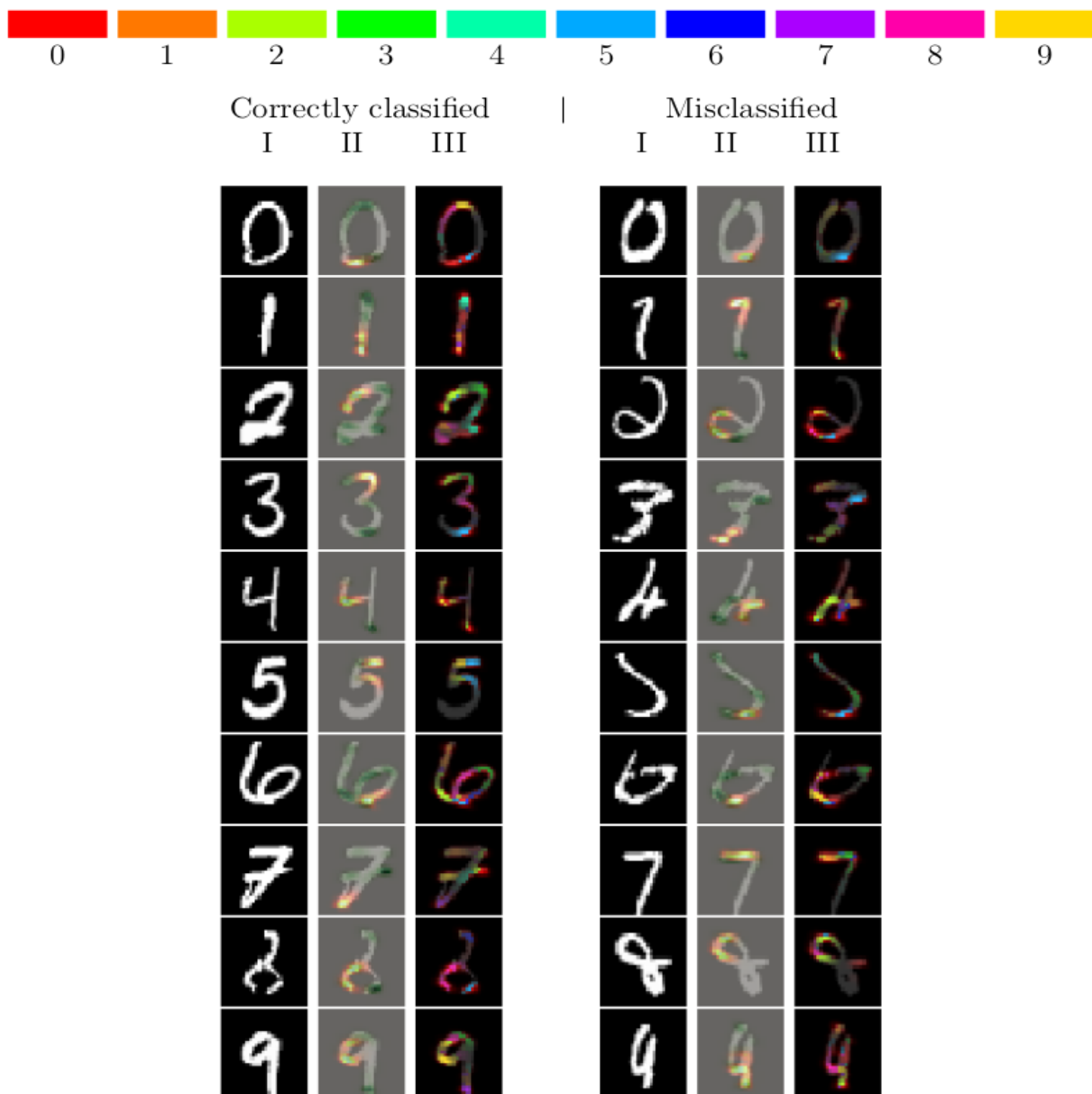


Figure 5.3: Example images from the MNIST dataset. Each row represents two sets of examples for digit 0-9: correctly classified example (left) and misclassified example (right). Each example set consists of the (I) original image, (II) heatmap results (where hot regions are focus of positive kernel and green represents dominant pixel results for the rest of kernels in the last layer) and (III) CLEAR maps. The color map on top shows the associations of different colors with their respective classes in the CLEAR map.

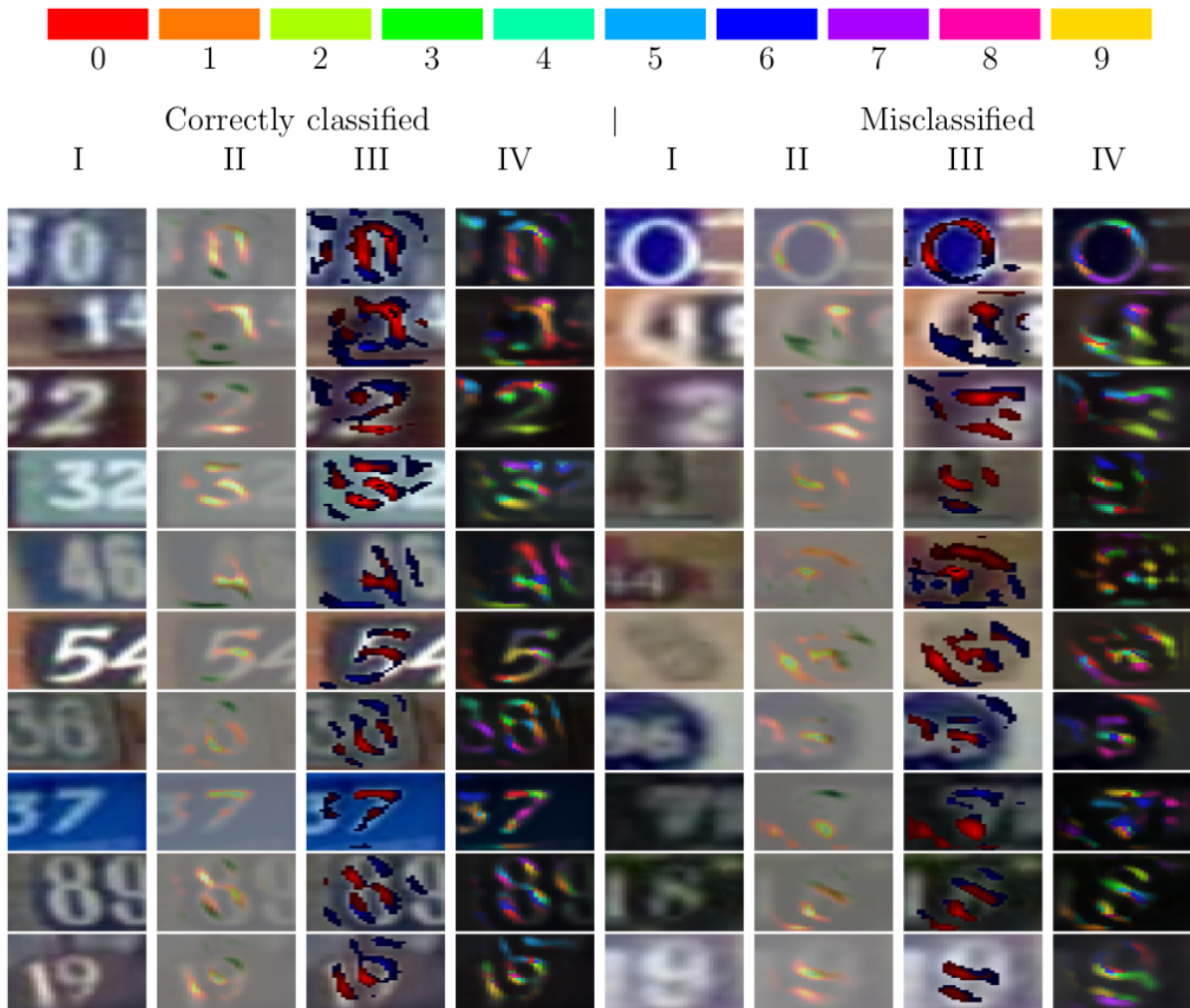


Figure 5.4: Correctly classified (left) and misclassified (right) images from the SVHN dataset. Each row represents two sets of examples for digit 0-9. Each example set consists of the (I) original image, (II) heatmap results (where hot regions are focus of positive kernels, and green regions for the rest of kernels), (III) binary map (red represents information for and blue represents information against the given image class) and (IV) CLEAR map respectively. The color map at the top shows the associations of different colors with their respective classes in the CLEAR map.

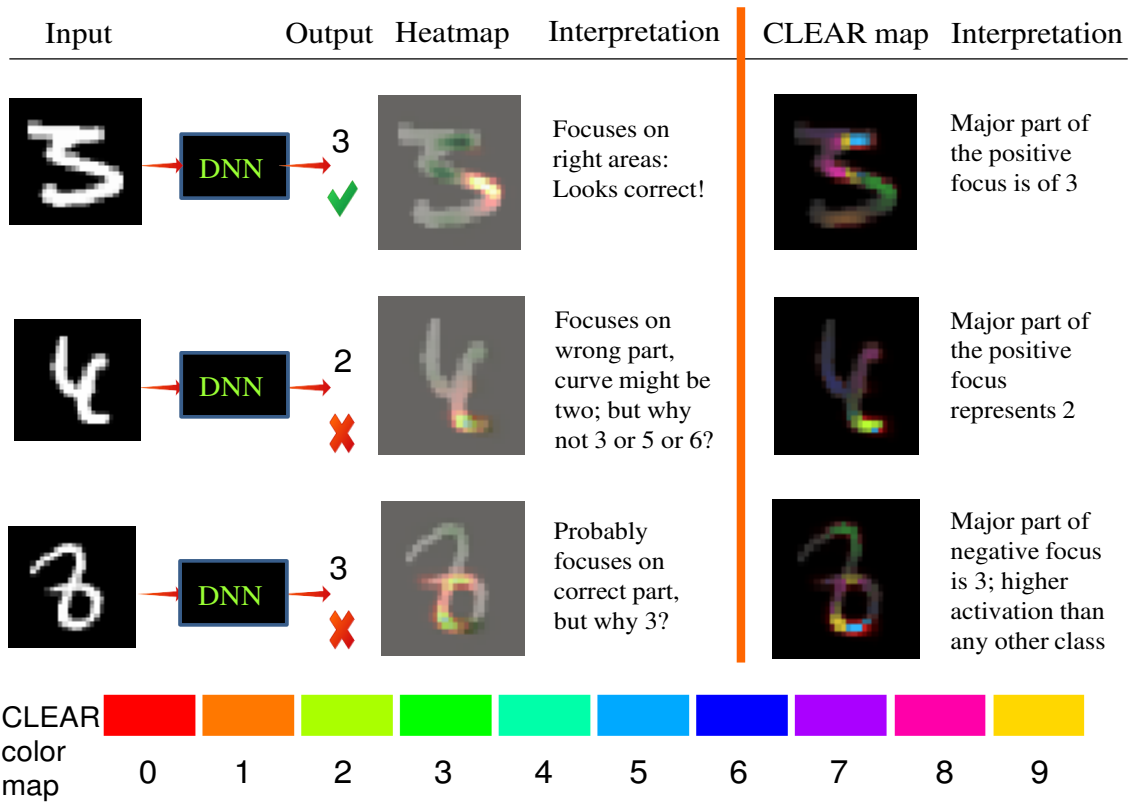


Figure 5.5: Examples of handwritten digits from MNIST are shown, along with: 1) the decision made by the CNN, 2) heatmaps used in existing visualization methods, 3) the proposed **C**lass-**E**nhanced **A**ttentive **R**esponse (CLEAR) maps, and 4) what can be interpreted based on the heatmaps and the proposed CLEAR maps. While the heatmaps used in existing approaches show which information in the image works for (positive focus: hot regions) or against (negative focus: green) a particular decision made by the CNN, the proposed CLEAR map allows for the visualization of the attentive regions of interest, the corresponding attentive levels, as well as the dominant class for each attentive region of interest that the CNN uses during the decision-making process. Each individual color in the CLEAR map represents the corresponding dominant attentive class at that location. Correspondence between colors and the dominant classes can be derived by the color map given at the bottom. In these examples, it is evident that the heatmaps are insufficient to fully interpret and explain the decision made by the CNN, whereas the proposed CLEAR maps can explain the decision-making process more effectively through a multi-factor visualization approach.

to classify without the strong features, albeit with half of the accuracy in comparison to the above case, an argument can be made that for these cases, the network focuses again on similar or redundant features. An example is digit 3, where there are redundant strong curve features.

5.2.6 Discussions

This section discusses some general points associated with the CLEAR maps approach: 1) It is interesting to note that in Fig. 5.2, there is sparsity in the individual response maps from the last layer kernels. We observed the same pattern for all datasets considered. Evidence for classes tends to come from very specific localized regions. 2) In the current realization of our approach, we use deconvolution responses with only fully convolutional networks. We would like to point out that even though end-to-end learning in this case is only possible with Fully Convolutional Nets (FCN), our approach can be extended to be used with different network architectures with the use different response methods, such as Layer-wise Relevance Propagation (LRP) [5], Deep Taylor decomposition [81], or prediction differential analysis [135].

The above experiments proved the efficacy of the proposed CLEAR method in explaining cases which remain unexplained through other visualization based methods. In the following sections, we present another study that uses CLEAR in the medical imaging domain specifically for the diabetic retinopathy, the leading cause of blindness in the world.

5.3 Experiments - Diabetic Retinopathy

To prove the efficacy of CLEAR in the medical imaging domain, we use CLEAR to create a Computer Aided Diagnostic (CAD) system that helps clinicians detect and diagnose diseases faster and more accurately. In particular, we apply the CLEAR approach for diabetic retinopathy for which the motivation and detailed setup along with some results are explained in the below subsections.

5.3.1 Diabetic Retinopathy: Motivation

Diabetic retinopathy is a medical condition that causes damage to the retina due to diabetes. It is the leading cause of blindness in the world. Traditionally, clinical diagno-

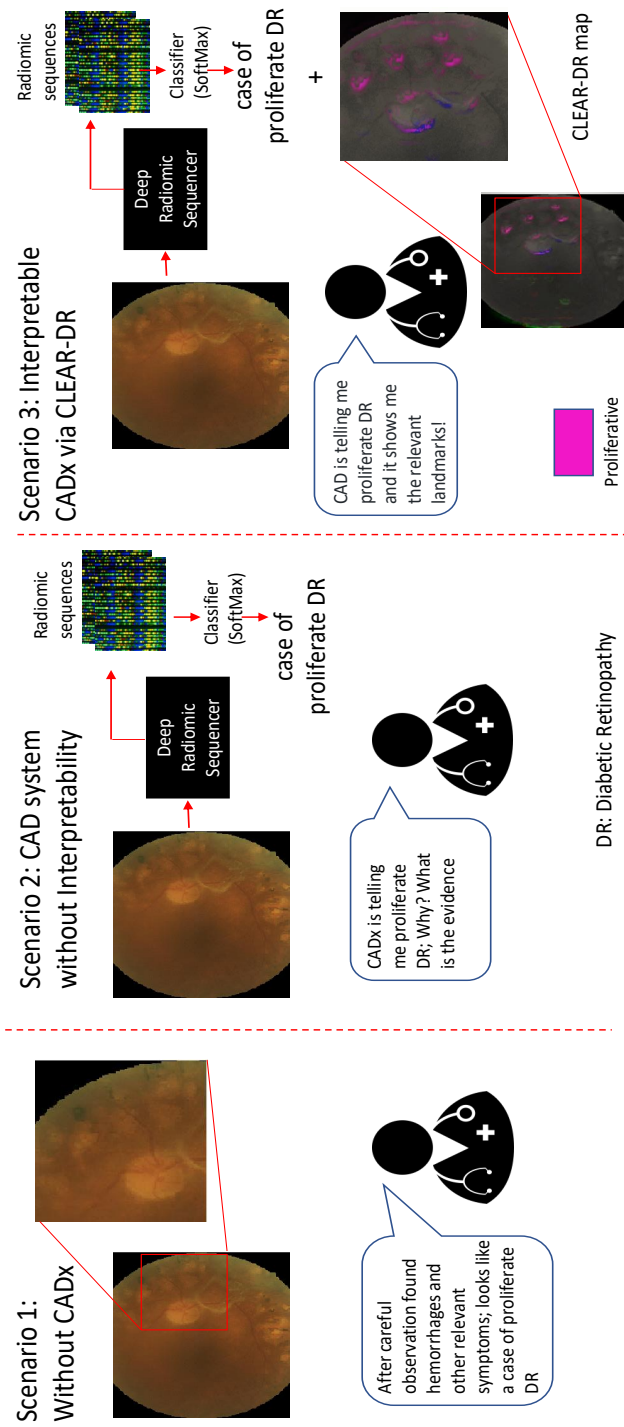


Figure 5.6: Three different scenarios for grading diabetic retinopathy: 1) without CAD, 2) CAD system without interpretability, 3) interpretable CAD via CLEAR. The proposed CLEAR-DR system improves clinical interpretability by providing effective visual interpretations of the decision-making process. The CLEAR-DR maps allows for the visualization of i) the attentive regions of interest responsible for grading decisions made by the deep radiomic sequencer; ii) their level of contribution to the grading decision; and iii) the dominant grade associated with each attentive region of interest. This visualization enables clinicians to better understand the rationale behind the grading decision made by the deep radiomic sequencer.

sis for diseases such as diabetic retinopathy is highly subjective based purely on a clinician’s experience. As such, these diagnoses have high inter- and intra-observer variability. The prevalence of computer-aided diagnosis (CAD) systems to support clinicians in their decision-making process has risen, enabling faster and more accurate diagnostic decisions with lower variability. In particular, radiomics-driven CAD has become an increasingly more prevalent area of research focus, where radiomic sequences consisting of a large number of image-based features are extracted and used to help clinicians make more informed decisions, and provide a virtual second opinion [1]. However, traditional radiomic sequences comprise largely of generic, hand-crafted features, which may be limiting in characterizing unique disease traits.

More recently, the concept of *Discovery Radiomics* has been shown to be particularly promising for oncology decision support by directly discovering radiomic sequencers based on medical imaging data [63], resulting in radiomic sequences that are tailored to characterizing unique disease traits. A particularly powerful use of Discovery Radiomics is for the discovery of deep radiomic sequencers, which leverage deep neural network (DNN) architectures to learn and extract subtle, latent features associated with key disease characteristics. Although these CAD systems are largely uninterpretable, such DNN-based approaches have shown considerable promise in detecting diabetic retinopathy [41, 122]. Though these CAD systems are largely uninterpretable as there is no mechanism in their method to explain the predictions.

Motivated by the need for clinical interpretability, we implemented CLEAR for the purpose of clinical decision support for diabetic retinopathy. CLEAR not only generates discriminative radiomic sequences for making grading decisions for diabetic retinopathy as a use case, but also visually interpret and understand these decisions via information back-propagation. The back-propagation is done through the discovered radiomic sequencer by embedding the CLEAR approach proposed by Kumar et. al. [67]. This process is designed to enable grade-level interpretability. As shown in Fig. 5.6, CLEAR can also help in reducing inter-observer variability and intra-observer variability while speeding up the overall diagnostic process. The main contribution of the proposed CLEAR CAD system is as follows:

- To the best of the authors’ knowledge, this is the first interpretable deep radiomic sequencer-driven CAD system proposed that enables the visualization of multi-class medical diagnosis grading processes.
- The study shows a direct qualitative correlation between the medically relevant landmarks that human experts use for diabetic retinopathy grading and the landmarks

used by CLEAR for classifying different grades of diabetics through retinopathy images.

This section presents the experimental setup and the qualitative experiments performed to show the efficacy of the CLEAR maps via the discovery radiomics framework. We conducted experiments on the Kaggle diabetic retinopathy dataset [53] using a CNN-based deep radiomic sequencer. Details about the dataset and training are explained below.

5.3.2 Diabetic Retinopathy Dataset

The Kaggle diabetic retinopathy dataset [53] consists of high-resolution retinal fundus images with varying degrees of illumination conditions captured using different types of cameras. The retinal fundus images in the dataset were clinically annotated with five different grades related to the presence of diabetic retinopathy. The five grades of diabetic retinopathy are as follows: 0: Negative, 1: Mild, 2: Moderate, 3: Severe, and 4: Proliferative. The dataset consists of images from both right and left eyes. Mild noise is present in both the images and ground truth labels.

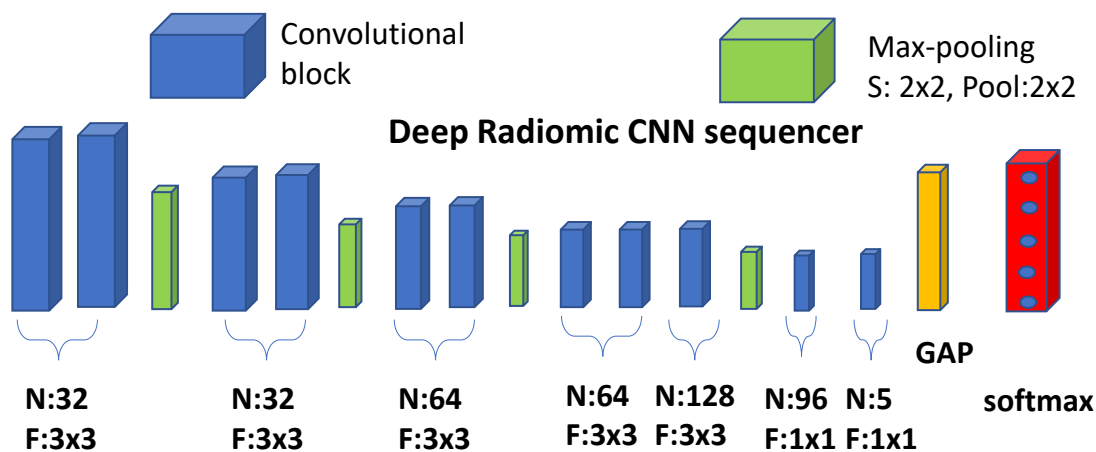


Figure 5.7: Architecture of the convolutional radiomic sequencer used in the deep radiomic sequencer discovery process. The radiomic sequencer is embedded in the sequencer discovery process, which augments a set of fully convolutional layers, a rectified linear unit layer, max-pooling, global average pooling (GAP) and a loss layer at the end of the sequencer for the learning process.

5.3.3 Experimental Setup: Training A Discovery Radiomics Sequencer

To create and train a deep radiomic sequencer for diabetic retinopathy, we use a CNN as shown in Fig. 5.7. To train this deep radiomic sequencer, we selected retinal fundus images for one eye (right) only and performed an automatic selective cropping to remove the background information. The use of a single eye led to 53,354 images in total. For evaluation purposes, we divided the dataset into 90% and 10% of the dataset for training and testing respectively. We augment the dataset by performing horizontal and vertical flipping along with channel-wise normalization. Using the above setup, we trained the deep radiomic sequencer and achieved an accuracy of 73.2% overall. It is important to note here that the goal of this study is to create an interpretable system for diabetic retinopathy, and thus the focus is on interpretability of grading decisions made using the deep radiomic sequencer. As such, the accuracy of the proposed CAD system can be improved further by leveraging alternative DNN architectures and other optimization approaches.

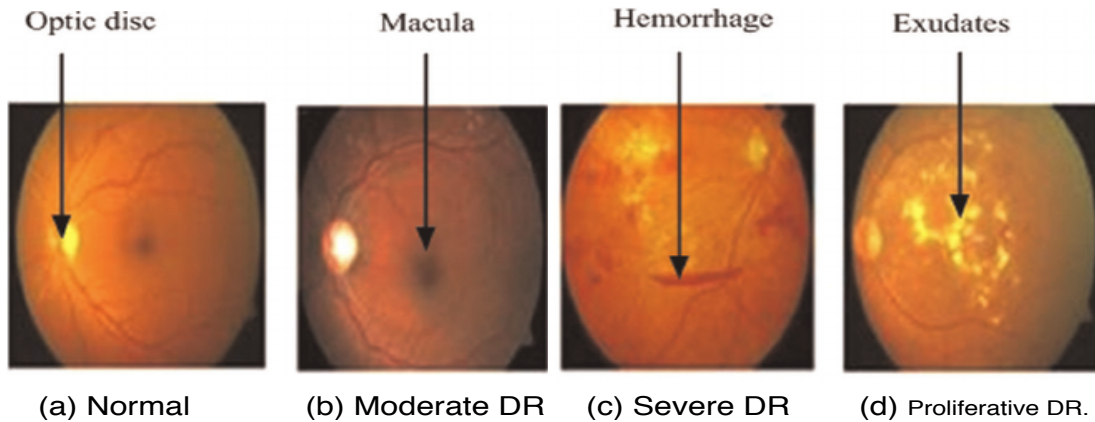


Figure 5.8: Figure showing a normal fundus image (a) and different grades of diabetic retinopathy ((b),(c) & (d)) identified by the presence of various types of abnormalities in the retinal fundus image.

5.3.4 Qualitative Experiments: CLEAR for Interpretable CAD

To demonstrate the efficacy of interpretability with the CLEAR system for diabetic retinopathy, we took the above discovered deep radiomic sequencer and created CLEAR maps using the same procedure shown in Fig. 5.2 for all diabetic retinopathy grades for both scenarios i.e., cases where the CLEAR CAD system correctly predicts a diabetic’s grade (Fig. 5.9(a)) and cases where it failed to identify the correct grade (Fig. 5.9(b)). Observing the individual cases in Fig. 5.9, it is evident that CLEAR maps are able to explain both scenarios i.e., where the grade was either correctly or mis-classified. For both scenarios, it provides the attention areas that highlight the associated abnormalities as shown in Fig. 5.8. Thus, giving a rationale in both cases for particular decisions that the clinician can reason with.

5.3.5 Discussions

Specific observations can be made from Fig. 5.9. For example, in the correctly classified diabetic retinopathy case in Fig. 5.9(a), it can be observed that the deep radiomic sequencer mainly focuses on the veins near the eye-balls in the retinal fundus image as there is an absence of abnormalities. Hence, it only focuses on normal nerve near optic disc. A similar observation about the abnormality and the extent of it can be made in other correctly identified cases. Specifically, for proliferative case in Fig. 5.9(a), CLEAR-DR

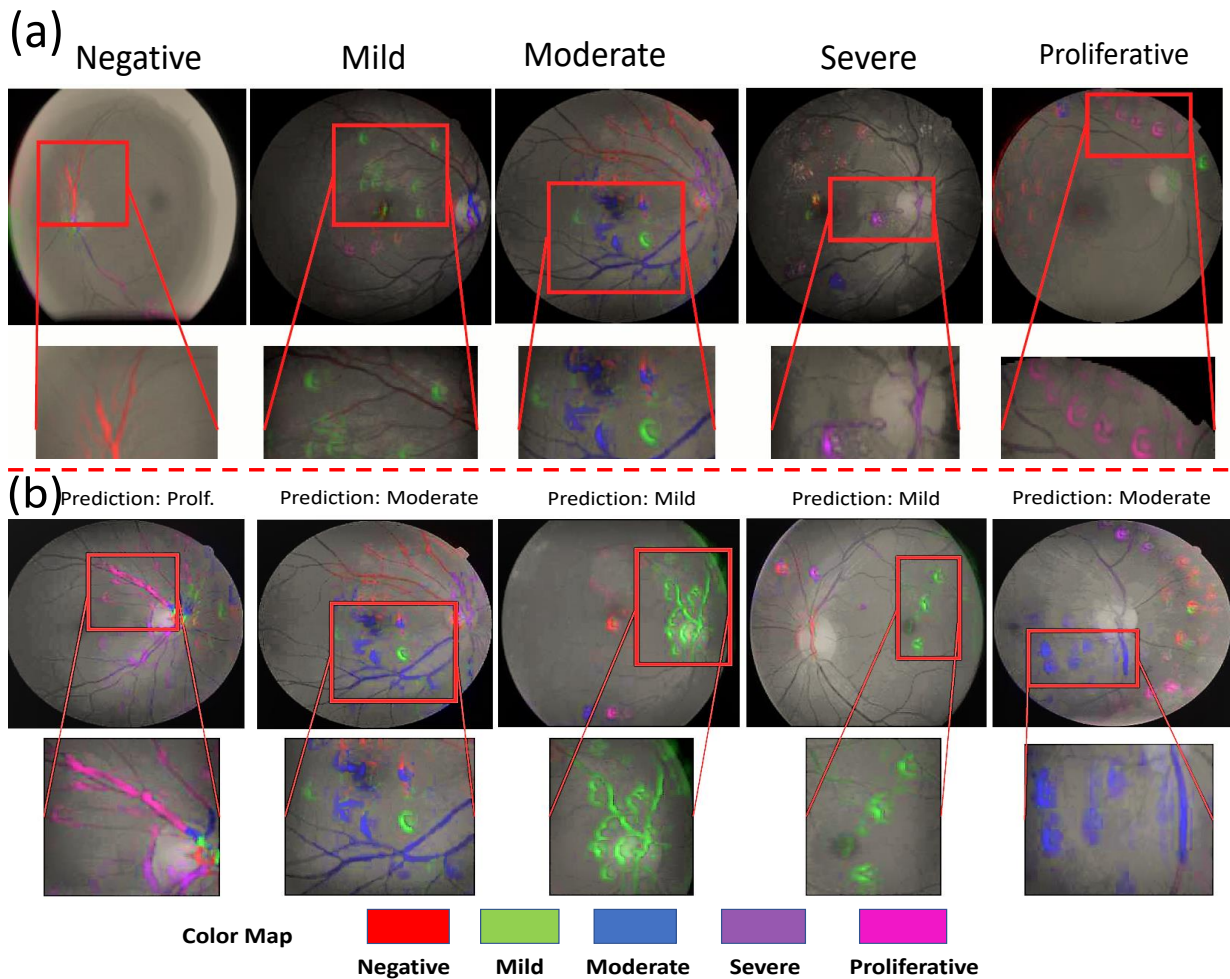


Figure 5.9: Correctly (a) and Mis-classified (b) examples for all diabetic retinopathy grades. Each color represents a single grade, as identified by the color map at the bottom of the figure. As well, the red box indicates the most attentive region used for grade prediction. It can be observed that the attentive regions used by the deep radiomic sequencer for making correct decisions corresponds to medically relevant landmarks, thus providing additional evidence for the proposed prediction. Best viewed in color and zoomed in.

system correctly identifies the exudates in the fundus image, which is usually associated with proliferative diabetic retinopathy. Similarly, in the same image, macula is detected for the moderate case. In the mis-classified case of mild diabetic retinopathy in Fig. 5.9 (b), it can be seen the deep radiomic sequencer fails to focus on the correct abnormalities. For example, in the normal case, it identifies the nerves around the disc as abnormalities, and considers it as proliferative case. If such an image is shown to clinician, they can see the evidence and reject it right away.

Based on these results and observations, it is evident that CLEAR maps show a direct correlation between the medically relevant landmarks for identifying the condition of diabetic retinopathy and the attentive areas used by the CLEAR CAD system for grading diabetic retinopathy. Thus, we argue that the CLEAR maps are effective for understanding and interpreting the classification decisions made by a CAD system and also for providing a reason for their decision making process in clinical settings.

5.4 Summary

In this Chapter, a novel approach to better understanding and visualizing the decision-making process of DNNs was introduced in the form of CLASS-Enhanced Attentive Response (CLEAR) maps. CLEAR maps are designed to enable the visualization of not only the areas of interest that predominantly influence the decision-making process, but also the degree of influence as well as the dominant class of influence in these areas. This multi-faceted look at the decision-making process allows for a better understanding of not only where but why certain decisions are made by DCNNs compared to existing heatmap-based approaches.

Experiments using three different publicly available datasets (two generic image dataset and one medical) were performed and show the efficacy of CLEAR maps both quantitatively and qualitatively. For the generic image datasets, we demonstrated that strong areas of interest identified with CLEAR maps play a pivotal role in the correct classification of the class.

For the diabetic retinopathy sequencer via CLEAR-DR, we show a direct correlation between the medically relevant landmarks used by human experts for grading and the visual features identified and used by the CLEAR-DR system for diabetic retinopathy grading. Thus in the medical domain, the proposed approach has great potential to reduce inter- and intra-observer variability and to accelerate the overall screening and diagnosis process while improving consistency and accuracy in clinical settings.

In the last Chapter of this thesis (Chapter 6), we draw conclusions about the introduced class based explanation strategies and discuss some interesting future directions, including a formulation for class-based global explanation that can be used for discovering and explaining the concepts identified by trained deep neural networks using human attribute priors.

Chapter 6

Conclusions & Future Work

“Call it the Hamlet strategy: lending a deep neural network the power of internal monologue, so that it can narrate what’s going on inside. But do the concepts that a network has taught itself align with the reality that humans are describing, when, for example, narrating a baseball highlight? Is the network recognizing the Boston Red Sox by their logo or by some other obscure signal, like “median facial-hair distribution,” that just happens to correlate with the Red Sox? Does it actually have the concept of “Boston Red Sox” or just some other strange thing that only the computer understands? It’s an ontological question: Is the deep neural network really seeing a world that corresponds to our own?”

- Trevor Darrell, NY Times article

The goal of this research was to study the effectiveness of the current interpretability approaches for explaining the decision making process of deep neural networks and address some of their shortcomings. In this regard, in Chapter 3, we address some of their shortcomings pertaining to absence of correlation between the obtained binary heatmaps and human domain knowledge. We also propose to make the heatmap formation process end-to-end in Chapter 4. In Chapter 5, we showed that using a multi-class enhanced approach for forming attentive regions can produce more comprehensive explanations for scenarios that remained explained using the previous heatmaps methods including attention response maps as introduced in Chapter 3. As the previous chapters contributed towards local explanation methods, in this final chapter, we also introduce a framework for producing class level global explanations as potential interesting direction to explore. Therefore, this Chapter first summarizes the main contributions presented in this thesis and offers some promising directions for future research including a human attributes prior-based global explanation framework.

6.1 Thesis Contribution Highlights

The main contribution of this thesis can be summarized as follows:

- **Domain knowledge discovery via attention response maps (Chapter 3):** We proposed the concept of attention response maps. First, we showed how attention maps can be used to relate the attentive region used by a deep neural network for making decisions to human domain expertise. Along with this, we also conducted experiments and presented specific evidence to show where in the training process of such neural network’s learning process does the hierarchy of features emerge and at which stage of training, deep neural networks start to use landmarks that correlate with human domain knowledge.
- **End-to-end architecture design for attention response maps (Chapter 4):** We introduced a novel end-to-end interpretable architectural design framework for attentive response maps. This framework is enabled by the proposed radiomic sequencer: a deep stacked interpretable sequencing cell (SISC) architecture. Experimental results show that the proposed SISC radiomic sequencer is able to not only achieve state-of-the-art results, but also offers prediction interpretability in the form of critical response maps generated through the stack of interpretable sequencing cells. The critical response maps highlights the critical regions used by the sequencer for making predictions. The critical response maps are useful for not only validating

the predictions of the proposed SISC radiomic sequencer, but also provide improved human expert-machine collaboration for improved decision support.

- **Class-enhanced attentive response maps (CLEAR) (Chapter 5):** We proposed the CClass-Enhanced Attentive Response (CLEAR) maps for better understanding and visualizing the decision-making process of DNNs. CLEAR maps are designed to enable the visualization of not only the areas of interest that predominantly influence the decision-making process, but also the degree of influence as well as the dominant class of influence in these areas. This multi-faceted look at the decision-making process allows for a better understanding of not only where but why certain decisions are made by CNNs compared to existing heatmap-based approaches. Thus, CLEAR is able to explain scenarios that remain unexplained by binary attention maps.

6.1.1 Limitations

The qualitative and quantitative experimental results demonstrated that the proposed methods produce quite interpretable results. However, there are some limitations accompanied with these methods which should be considered:

1. **Attention Response Maps:** One of the major limitations of the attentive response maps is their inability to automatically compare to a human expert’s domain knowledge. It still requires a human in the loop to compare the output prediction of these maps and relate it to domain knowledge. Eliminating the human in the loop can result in removing some of the subjectivity still involved in the analysis. We attempt to resolve this situation to an extent in this chapter but we believe it can be explored further.
2. **CLEAR:** One of the limitations of our proposed CLEAR framework is visualization in scenarios with large number of classes (> 20). We did not strive to show the CLEAR maps for all scenarios with that number of classes, as doing so would make it extremely difficult to interpret the decision outputs. For such cases, perhaps showing the top 10 most activated class or several different maps with N classes would be a better approach. In the current realization of our approach, we use deconvolution responses with only fully convolutional networks. We would like to point out that even though end-to-end learning in this case is only possible with Fully Convolutional

Nets (FCN), our approach can be extended to be used with different network architectures with the use of different response methods, that allows such representation to be made for all kind of network architectures.

6.2 Future Work

The proposed methods in this thesis open several new directions for future work. Here we describe the main topics.

6.2.1 Human Attributes Prior Based Concept Explanations

Most of the recent works (as discussed in Chapters 2-5) in the domain of explainability of neural networks pertain to local explanations via feature importance methods. In particular, research in this sub-domain of XAI tends to depend on answering a particular question: *what evidence is present in a given instance or sample that justifies the prediction?*. This is an important question to answer to arrive at an explanation for a given input. However, there are certain limitations attached with such an approach.

One of the prominent limitations of local feature importance based methods is the lack of objectivity in their explanations. The end user (human domain expert) still needs to look at the evidence presented in the form of group of pixels and make his or her own subjective explanation. Another limitation of these method is the absence of human labelled *attributes*. Providing such explicit attributes as concepts can in-part remove the problem associated with the lack of objectivity and can also provide capabilities to *measure* explanations in terms of human expert defined important and discriminate concepts. Hence, there is a need to have methods that can provide global explanation in terms of human defined concepts.

In this section of the thesis, we aim to propose a solution for the above mentioned limitations of the local feature or attention based methods by providing a framework that can generate class specific “global explanations” using human provided attributes as priors. We believe it is an important and interesting research direction to pursue for obtaining a holistic explanation framework. In particular, we propose to use human provided attributes as exemplar priors for concepts instead of automatically discovering concepts as in Ghorbani et. al. [34]. As discussed earlier, we do so because the end-user (human expert) is the final authority in accepting and rejecting the provided explanations. Hence, it is advisable to do this premortem than postmortem. Therefore, starting with human defined attributes

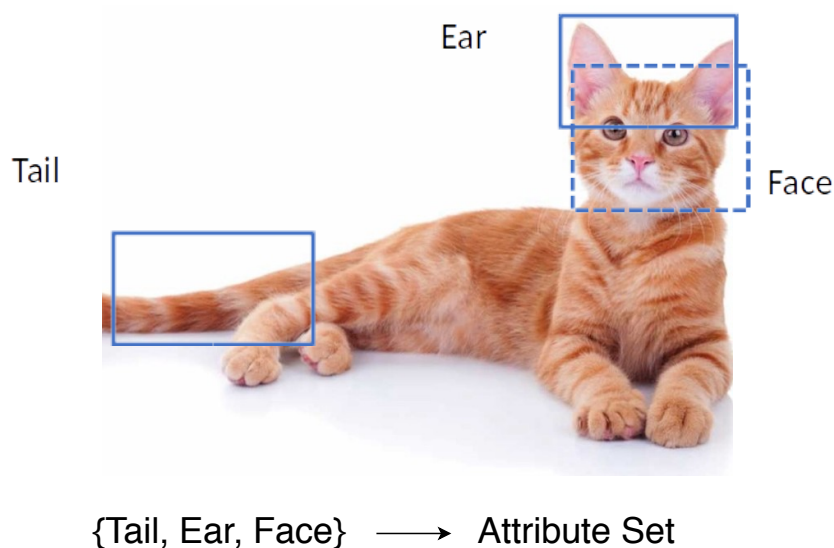


Figure 6.1: Illustration of human defined attributes. In the given image, a human user labels the critical attributes associated with identifying a cat. These attributes can be collected to form a attribute set that can be later used as priors for the formation of class specific global explanations.

(as shown in Fig. 6.1) allows us to remove the subjective analysis of discovered concepts later on by the human expert and can also help in additional analysis such as providing a quantifiable metric for comparing different attention based methods and different neural network architectures (discussed later).

The overview and detailed formulation for our proposed human attributes prior based concept explanation framework is explained below.

Methodology

An overview of the proposed framework is shown in Fig. 6.2. As shown in the figure, we first need human experts to provide attributes as priors. The attributes are meant to be defined as the most discriminative features or attributes that define a given class of objects based on the domain knowledge of the human expert. Hence, to select these attributes the human expert is shown few exemplar cases to select and define/label these attributes. Once the human domain expert defines all the required n attributes, we move to the next steps in the process.

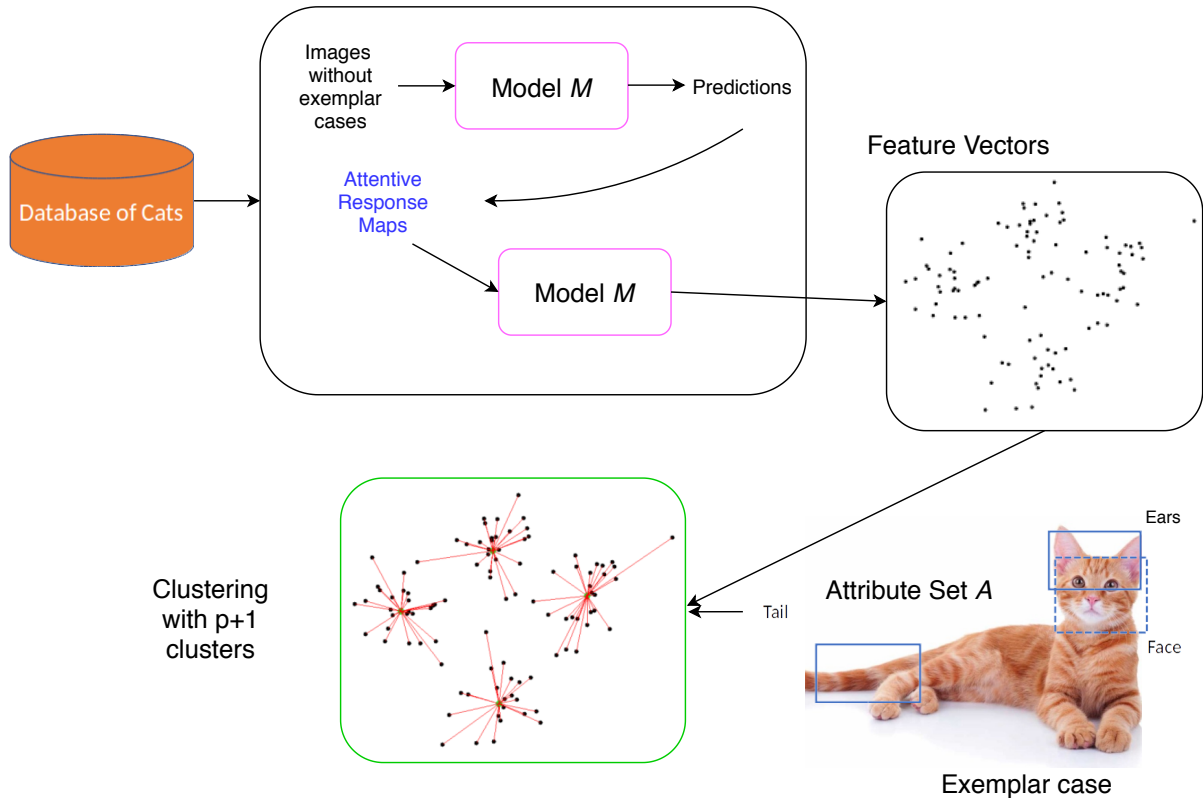


Figure 6.2: Illustration of how class specific global explanations can be formed using human defined attributes. The formed clusters can be used to identify which attentive response maps align with which human expert defined attributes.

For the given class of objects, we select all the examples of that particular class and pass them through a trained neural network which is trained to identify/classify objects including the given class. We then form attention maps for all the correctly identified images for given classes in the dataset excluding the exemplar cases. In the next steps, we take the exemplar attributes and labelled regions and cluster the obtained attentive regions for the rest of the samples, with the attributes acting as the initial centroids. We start with $p + 1$ clusters for all the p attributes along with a cluster for other background features. At the end, we apply a distance metric to quantify how a given trained network makes decisions and whether decisions are based on the human expert defined attributes or something else.

Formulation

Consider a set of images I for a particular class c in a given database. Hence, the set of all the images of class c will be:

$$I_c = \{i_1^c, i_2^c, i_3^c, \dots, i_n^c\}. \quad (6.1)$$

Using some random number of images from I_c we form an exemplar set, I_e where $I_e \subset I_c$. The images from set I_e are shown to the human expert one by one to extract the domain specific attributes a_i where a_i is i th attribute. The images are shown until the human expert is satisfied that all the discriminative features associated with a given class c are identified. Hence, we then collect all attributes, to form the attribute set A_h , where it is defined as:

$$A_h = \{a_{h_1}, a_{h_2}, a_{h_3}, \dots, a_{h_p}\}. \quad (6.2)$$

The length of set A_h of the above defined set is p . We then take the set of images for the given class from the database other than the images contained in the exemplar set I_e , i.e., images in $I_c \setminus I_e$. These images are then sent to a trained neural network M , to extract the spatial attentive regions using a instance based local explanation method. Let us call the collected set of spatial attentive regions for N images of class c as S_c :

$$S_c = \{s_{i_1}, s_{i_2}, s_{i_3} \dots, s_{i_N}\}; i \subset I_c \setminus I_e. \quad (6.3)$$

The attentive regions set S_c is then passed through M to obtain attentive regions feature vectors, thus obtaining the vectorized set \mathbf{S}_c . In the next step, we attempt to form $p + 1$ clusters to classify the spatial attentive region identified in S_c . We use the p plus one more random defined attribute as the initial cluster cluster for our spectral clustering process to

cluster \mathbf{S}_c . At the end, we calculate the Euclidean distance of each image in a given cluster p , with the initial cluster centroid as obtained from A_h . For the $p + 1$ th cluster, we don't calculate the score. Thus, we can obtain a quantitative metric how all the defined local explanations for images in a given class c pertains to the human expert defined attributes.

The above defined quantitative metric then can be used to compare different neural networks architectures as well as different visualization methods. Thus, the proposed framework can act as a comparative tool for XAI methods, which is sorely lacking currently.

6.2.2 Multi-modal Data Explanations

Humans do most of their learning and recognition tasks through multi dimensional data and multi-modal data. For example, to understand a cat, we look at how it moves, what sounds it makes, what it looks like in three dimensions etc. Therefore, building explainable neural network models with just 2D images and trying to relate it to human learning is quite restrictive. Hence, in future it would be interesting to look at the XAI methods through a multi-modal perspective.

6.2.3 Beyond Explainability

One of the major questions that still needs to be explored is in the direction of where else these interpretability methods can be used? Is there a possibility of using these methods for tasks other than the interpretability such as improving the learning process of deep networks or creating adversarial examples. We propose to explore in the following directions as future work in this domain:

Improved Adversarial Learning

There has been some attention in exploring the use of gradient-based interpretability for purposes other than purely explainability. For example, Zhou et. al. [132] proposed the use of gradient-based interpretability to improve the localization performance of deep neural networks. A number of research studies [42, 70] have leveraged sensitivity maps produced via gradient-based interpretability as initialization for the task of segmentation. However, leveraging gradient-based interpretability for tasks beyond explainability is still not well explored outside of these few examples, making further investigations into alternative directions for leveraging insights gained through interpretability ripe for exploration.

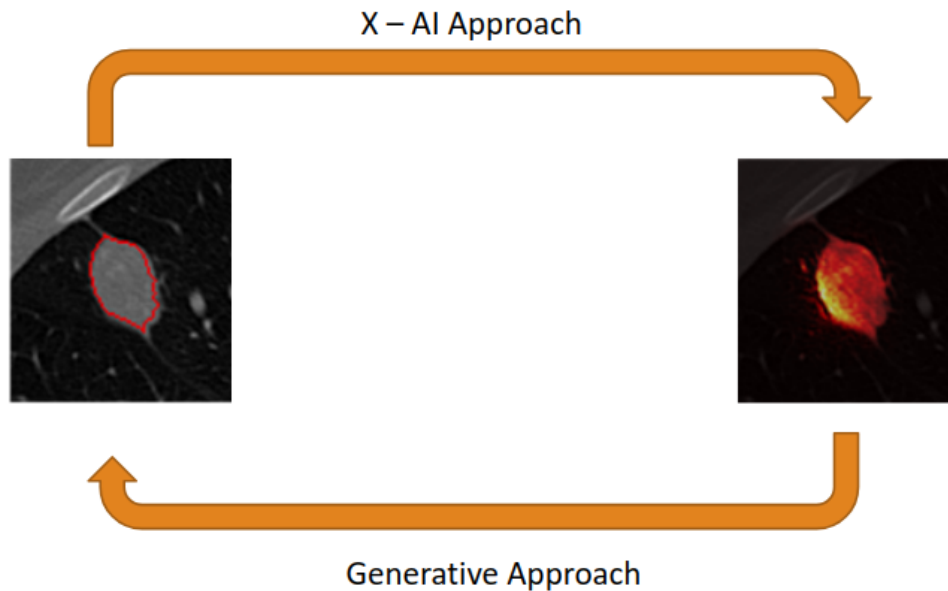


Figure 6.3: The illustration shows an example how explainable methods can be used to first obtain attention maps which in turn can be used with generative methods to obtain annotations or at-least as initial seed for segmentation algorithm.

It would be an interesting avenue to leverage gradient-based interpretability in the realm of adversarial examples, where the goal is to produce delicately perturbed inputs designed to mislead machine learning models towards incorrect predictions. More specifically, it would be important to look at the concept of spatially constrained one-pixel adversarial perturbations, guided by gradient-based interpretability such that insights gained via interpretability is used to aid adversarial attacks. One-pixel adversarial perturbations [116] is an extreme case of adversarial examples where only one pixel is modified to fool a model into providing the wrong prediction. This pixel is found through Differential Evolution [19], where a population of candidate pixels is randomly modified to create children that compete with its parents for fitness in the next iteration; this fitness criterion being the probabilistic predicted label. The optimal pixels for one-pixel adversarial perturbations usually lie in positions of interest. This observation motivates us to leverage gradient-based interpretability to constrain the differential evolution initialization; we posit that, by ensuring that the initial population of pixels lie in positions of interest as given by generated sensitivity maps, the optimization algorithm for generating one-pixel adversarial perturbations can converge faster with fewer iterations. Furthermore, by guiding it towards areas of interest, the produced attacks may also be more visually difficult to perceive. It

would be an interesting approach to test out our initial hypothesis further with elaborate experimentation.

Generating Annotations

Another potential use of the explainability methods can be in automatically generating annotations for segmentation, especially in the cases where getting annotations can be difficult both in terms of time and cost (An illustration of this is shown in Fig. 6.3). For example, getting an experienced radiologist to provide large number of annotations of cancerous regions that can be then used to train deep neural networks involves a large amount of cost and time. Also, there are certain times where intra or even inter observer variability is introduced in such a process. Using an approach, where you train a model for a classification task, and then use an interpretability method to identify or label annotations (Fig. 6.3) can prove to remove such observer related variability while saving time and reducing cost.

References

- [1] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Cavalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5, 2014.
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.
- [3] Samuel G Armato, Karen Drukker, Feng Li, Lubomir Hadjiiski, Georgia D Tourassi, Justin S Kirby, Laurence P Clarke, Roger M Engelmann, Maryellen L Giger, George Redmond, et al. Lungx challenge for computerized lung nodule classification. In *Journal of Medical Imaging*, volume 3, page 044506. International Society for Optics and Photonics, 2016.
- [4] Samuel G Armato, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. In *Medical physics*, volume 38, pages 915–931. Wiley Online Library, 2011.
- [5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015.

- [6] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÅžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.
- [7] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.
- [8] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.
- [9] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.
- [10] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Praseon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jikai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.
- [11] Mikio L Braun, Joachim M Buhmann, and Klaus-Robert MÅžller. On relevant dimensions in kernel feature spaces. *Journal of Machine Learning Research*, 9(Aug):1875–1908, 2008.
- [12] Mario Buty, Ziyue Xu, Mingchen Gao, Ulas Bagci, Aaron Wu, and Daniel J Mollura. Characterization of lung nodule malignancy using hybrid shape and appearance features. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 662–670. Springer, 2016.
- [13] Audrey G Chung, Mohammad Javad Shafiee, Devinder Kumar, Farzad Khalvati, Masoom A Haider, and Alexander Wong. Discovery radiomics for multi-parametric mri prostate cancer detection. In *arXiv preprint arXiv:1509.00111*, 2015.
- [14] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1153–1162, 2016.
- [15] Stefano Curtarolo, Dane Morgan, Kristin Persson, John Rodgers, and Gerbrand Ceder. Predicting crystal structures with data mining of quantum calculations. *Physical review letters*, 91(13):135503, 2003.

- [16] Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H Taylor, Lance J Nelson, Gus LW Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, et al. Aflowlib. org: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science*, 58:227–235, 2012.
- [17] Yaojun Dai, Shiju Yan, Bin Zheng, and Chengli Song. Incorporating automatically learned pulmonary nodule attributes into a convolutional neural network to improve accuracy of benign-malignant nodule classification. In *Physics in Medicine and Biology*. IOP Publishing, 2018.
- [18] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.
- [19] S. Das and P. N. Suganthan. Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation*, 15(1):4–31, Feb 2011.
- [20] Antonio Oseas de Carvalho Filho, Aristofanes Corrêa Silva, Anselmo Cardoso de Paiva, Rodolfo Acatauassú Nunes, and Marcelo Gattass. Classification of patterns of benignity and malignancy based on ct using topology-based phylogenetic diversity index and convolutional neural network. In *Pattern Recognition*, volume 81, pages 200–212. Elsevier, 2018.
- [21] Raunak Dey, Zhongjie Lu, and Yi Hong. Diagnostic classification of lung nodules using 3d neural networks. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 774–778. IEEE, 2018.
- [22] Ashis Kumar Dhara, Sudipta Mukhopadhyay, Anirvan Dutta, Mandeep Garg, and Niranjan Khandelwal. A combination of shape and texture features for classification of pulmonary nodules in lung ct images. In *Journal of digital imaging*, volume 29, pages 466–475. Springer, 2016.
- [23] Claudia Draxl and Matthias Scheffler. Nomad: The fair concept for big data-driven materials science. *MRS Bulletin*, 43(9):676–682, 2018.
- [24] Maria Jazmin Duarte, Julia Klemm, Sebastian Oliver Klemm, Karl Johann Jakob Mayrhofer, Martin Stratmann, Sergiy Borodin, Aldo H Romero, Milad Madinehei, Daniel Crespo, Jorge Serrano, et al. Element-resolved corrosion analysis of stainless-type glass-forming steels. *Science*, 341(6144):372–376, 2013.

- [25] Nóirín Duggan, Egil Bae, Shiwen Shen, William Hsu, Alex Bui, Edward Jones, Martin Glavin, and Luminita Vese. A technique for lung nodule candidate detection in ct using global minimization methods. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 478–491. Springer, 2015.
- [26] Ayman El-Baz, Matthew Nitzken, Fahmi Khalifa, Ahmed Elnakib, Georgy Gimel'farb, Robert Falk, and Mohammed Abo El-Ghar. 3d shape analysis for early diagnosis of malignant lung nodules. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 772–783. Springer, 2011.
- [27] Ulli Englert. Symmetry relationships between crystal structures. applications of crystallographic group theory in crystal chemistry. by ulrich müller. *Angewandte Chemie International Edition*, 52(46):11973–11973, 2013.
- [28] Dumitru Erhan, Aaron Courville, and Yoshua Bengio. Understanding representations learned in deep architectures. *Department d'Informatique et Recherche Operationnelle, University of Montreal, QC, Canada, Tech. Rep*, 1355, 2010.
- [29] Macedo Firmino, Giovanni Angelo, Higor Morais, Marcel R Dantas, and Ricardo Valentim. Computer-aided detection (cade) and diagnosis (cadx) system for lung cancer with likelihood of malignancy. In *Biomedical engineering online*, volume 15, page 2. BioMed Central, 2016.
- [30] Christopher C Fischer, Kevin J Tibbetts, Dane Morgan, and Gerbrand Ceder. Predicting crystal structure by merging data mining with quantum mechanics. *Nature materials*, 5(8):641, 2006.
- [31] Center for Disease Control and “Lung cancer statistics” Prevention. <https://www.cdc.gov/cancer/lung/statistics/>. 2016.
- [32] Baptiste Gault, Michael P Moody, Julie M Cairney, and Simon P Ringer. Atom probe crystallography. *Materials Today*, 15(9):378–386, 2012.
- [33] Luca M Ghiringhelli, Jan Vybiral, Sergey V Levchenko, Claudia Draxl, and Matthias Scheffler. Big data of materials science: critical role of the descriptor. *Physical review letters*, 114(10):105503, 2015.
- [34] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pages 9273–9282, 2019.

- [35] Github. Keras visualization. <https://github.com/raghakot/keras-vis>, 2017. [Online; accessed 19-Oct-2017].
- [36] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.
- [37] Waleed M Gondal, Jan M Köhler, René Grzeszick, Gernot A Fink, and Michael Hirsch. Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. *arXiv preprint arXiv:1706.09634*, 2017.
- [38] Ian Goodfellow, Honglak Lee, Quoc V Le, Andrew Saxe, and Andrew Y Ng. Measuring invariances in deep networks. In *Advances in neural information processing systems*, pages 646–654, 2009.
- [39] RW Grosse-Kunstleve. Algorithms for deriving crystallographic space-group information. *Acta Crystallographica Section A: Foundations of Crystallography*, 55(2):383–395, 1999.
- [40] Felix Grün, Christian Rupprecht, Nassir Navab, and Federico Tombari. A taxonomy and library for visualizing learned features in convolutional neural networks. *arXiv preprint arXiv:1606.07757*, 2016.
- [41] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.
- [42] M. L. Ha, G. Franchi, M. Moller, A. Kolb, and V. Blanz. Segmentation and shape extraction from convolutional neural networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1509–1518, March 2018.
- [43] Th Hahn. International tables for crystallography, volume a of international tables for crystallography. *International Union of Crystallography, Chester, England*, 2006.
- [44] Fangfang Han, Huafeng Wang, Guopeng Zhang, Hao Han, Bowen Song, Lihong Li, William Moore, Hongbing Lu, Hong Zhao, and Zhengrong Liang. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. In *Journal of digital imaging*, volume 28, pages 99–115. Springer, 2015.

- [45] Fangfang Han, Guopeng Zhang, Huafeng Wang, Bowen Song, Hongbing Lu, Dazhe Zhao, Hong Zhao, and Zhengrong Liang. A texture feature analysis for diagnosis of pulmonary nodules using lide-idri database. In *Medical Imaging Physics and Engineering (ICMIPE), 2013 IEEE International Conference on*, pages 14–18. IEEE, 2013.
- [46] JB Heaton, NG Polson, and JH Witte. Deep learning in finance. *arXiv preprint arXiv:1602.06561*, 2016.
- [47] David Hicks, Corey Oses, Eric Gossett, Geena Gomez, Richard H Taylor, Cormac Toher, Michael J Mehl, Ohad Levy, and Stefano Curtarolo. Aflow-sym: platform for the complete, automatic and self-consistent symmetry analysis of crystals. *Acta Crystallographica Section A: Foundations and Advances*, 74(3):184–203, 2018.
- [48] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.
- [49] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
- [50] Peng Huang, Seyoun Park, Rongkai Yan, Junghoon Lee, Linda C Chu, Cheng T Lin, Amira Hussien, Joshua Rathmell, Brett Thomas, Chen Chen, et al. Added value of computer-aided ct image features for early lung cancer diagnosis with small pulmonary nodules: A matched case-control study. In *Radiology*, volume 286, pages 286–295. Radiological Society of North America, 2017.
- [51] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [52] Akihiro Ishibazawa, Taiji Nagaoka, Atsushi Takahashi, Tsuneaki Omae, Tomofumi Tani, Kenji Sogawa, Harumasa Yokota, and Akitoshi Yoshida. Optical coherence tomography angiography in diabetic retinopathy: a prospective pilot study. *American journal of ophthalmology*, 160(1):35–44, 2015.
- [53] Kaggle. diabetic retinopathy challenge. <https://www.kaggle.com/c/diabetic-retinopathy-detection/>, 2015. [Online; accessed 19-July-2017].
- [54] Amir-Hossein Karimi, Audrey G Chung, Mohammad Javad Shafiee, Farzad Khalvati, Masoom A Haider, Ali Ghodsi, and Alexander Wong. Discovery radiomics via

a mixture of deep convnet sequencers for multi-parametric mri prostate cancer classification. In *International Conference Image Analysis and Recognition*, pages 45–53. Springer, 2017.

- [55] Aydın Kaya and Ahmet Burak Can. A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics. In *Journal of biomedical informatics*, volume 56, pages 69–79. Elsevier, 2015.
- [56] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.
- [57] Been Kim, Justin Gilmer, Martin Wattenberg, and Fernanda Viégas. Tcav: Relative concept importance testing with linear concept activation vectors. 2018.
- [58] Bojan Kolosnjaji, Apostolis Zarras, George Webster, and Claudia Eckert. Deep learning for classification of malware system call sequences. In *Australasian Joint Conference on Artificial Intelligence*, pages 137–149. Springer, 2016.
- [59] Samantha Krening, Brent Harrison, Karen M Feigh, Charles Lee Isbell, Mark Riedl, and Andrea Thomaz. Learning from explanations using sentiment and advice in rl. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1):44–55, 2017.
- [60] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. In *Citeseer*, 2009.
- [61] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [63] D Kumar, AG Chung, MJ Shafiee, F Khalvati, MA Haider, and A Wong. Discovery radiomics for pathologically-proven computed tomography lung cancer prediction. In *14th Int. Conf. Image Analysis and Recognition*. Springer, 2017.
- [64] Devinder Kumar, Audrey G Chung, Mohammad J Shaifee, Farzad Khalvati, Ma-soom A Haider, and Alexander Wong. Discovery radiomics for pathologically-proven

- computed tomography lung cancer prediction. In *International Conference Image Analysis and Recognition*, pages 54–62. Springer, 2017.
- [65] Devinder Kumar, Mohammad Javad Shafiee, Audrey G Chung, Farzad Khalvati, Masoom Haider, and Alexander Wong. Discovery radiomics for pathologically-proven computed tomography lung cancer prediction. In *Image Analysis and Recognition: 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5–7, 2017, Proceedings*, volume 10317, page 54. Springer, 2017.
- [66] Devinder Kumar, Graham W Taylor, and Alexander Wong. Opening the black box of financial ai with clear-trade: A class-enhanced attentive response approach for explaining and visualizing deep learning-driven stock market prediction. In *arXiv preprint arXiv:1709.01574*, 2017.
- [67] Devinder Kumar, Alexander Wong, and Graham W Taylor. Explaining the unexplained: A class-enhanced attentive response approach to understanding deep neural networks. *Computer Vision & Pattern Recognition Workshop*, 2017.
- [68] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [69] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.
- [70] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. *arXiv preprint arXiv:1902.10421*, 2019.
- [71] Wei Li, Peng Cao, Dazhe Zhao, and Junbo Wang. Pulmonary nodule classification with deep convolutional neural networks on computed tomography images. In *Computational and mathematical methods in medicine*, volume 2016. Hindawi, 2016.
- [72] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.
- [73] Shuang Liu, Yiting Xie, Artit Jirapatnakul, and Anthony P Reeves. Pulmonary nodule classification in lung cancer screening with three-dimensional convolutional neural networks. In *Journal of Medical Imaging*, volume 4, page 041308. International Society for Optics and Photonics, 2017.

- [74] Xinglong Liu, Fei Hou, Hong Qin, and Aimin Hao. Multi-view multi-scale cnns for lung nodule type classification from ct images. In *Pattern Recognition*. Elsevier, 2018.
- [75] Xiaolei Ma, Haiyang Yu, Yunpeng Wang, and Yinhai Wang. Large-scale transportation network congestion evolution prediction using deep learning theory. *PLoS one*, 10(3):e0119044, 2015.
- [76] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *IEEE CVPR*, pages 5188–5196, 2015.
- [77] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9, 2015.
- [78] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.
- [79] V. Menkovski, Z. Aleksovski, A. Saalbach, and H. Nickisch. Can pretrained neural networks detect anatomy? *arXiv preprint arXiv:1512.05986*, 2015.
- [80] R Meyes, M Lu, C Waubert de Puiseau, and T Meisen. Ablation studies to uncover structure of learned representations in artificial neural networks. pages 185–191, 2019.
- [81] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.
- [82] Hiroki Nagai and Young Hak Kim. Cancer prevention from the perspective of global cancer burden patterns. In *Journal of thoracic disease*, volume 9, page 448. AME Publications, 2017.
- [83] Nature news. Can we open the black box of AI? <https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731/>, 2016. [Online; accessed 03-09-2017].
- [84] John Frederick Nye et al. *Physical properties of crystals: their representation by tensors and matrices*. Oxford university press, 1985.

- [85] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. In *IEEE Transactions on pattern analysis and machine intelligence*, volume 24, pages 971–987. IEEE, 2002.
- [86] Gregory B Olson. Designing a new material world. *Science*, 288(5468):993–998, 2000.
- [87] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1612.04757*, 2016.
- [88] Woon Bae Park, Jiyong Chung, Jaeyoung Jung, Keemin Sohn, Satendra Pal Singh, Myoungcho Pyo, Namsoo Shin, and K-S Sohn. Classification of crystal structure using a convolutional neural network. *IUCrJ*, 4(4):486–494, 2017.
- [89] A. S Razavian, J. Sullivan, S. Carlsson, and A. Maki. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016.
- [90] Anthony P Reeves, Yiting Xie, and Artit Jirapatnakul. Automated pulmonary nodule ct image characterization in lung cancer screening. In *International journal of computer assisted radiology and surgery*, volume 11, pages 73–88. Springer, 2016.
- [91] MIT Technology Review. The Dark Secret at the Heart of AI. <https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/>, 2017. [Online; accessed 19-July-2017].
- [92] Marko Robnik-Šikonja and Igor Kononenko. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600, 2008.
- [93] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [94] H. R Roth, C. T Lee, H. C. Shin, A. Seff, L. Kim, J. Yao, L. Lu, and R. M Summers. Anatomy-specific classification of medical images using deep convolutional nets. In *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 101–104. IEEE, 2015.

- [95] Geoffrey D Rubin, John K Lyo, David S Paik, Anthony J Sherbondy, Lawrence C Chow, Ann N Leung, Robert Mindelzun, Pamela K Schraedley-Desmond, Steven E Zinck, David P Naidich, et al. Pulmonary nodules on multi-detector row ct scans: performance comparison of radiologists and computer-aided detection. *Radiology*, 234(1):274–283, 2005.
- [96] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.
- [97] Mary P Ryan, David E Williams, Richard J Chater, Bernie M Hutton, and David S McPhail. Why stainless steel corrodes. *Nature*, 415(6873):770, 2002.
- [98] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.
- [99] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *IEEE International Conference on Computer Vision (ICCV)*, 2016.
- [100] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram van Ginneken. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. In *IEEE transactions on medical imaging*, volume 35, pages 1160–1169. IEEE, 2016.
- [101] Mohammad Javad Shafiee, Audrey G Chung, Devinder Kumar, Farzad Khalvati, Masoom Haider, and Alexander Wong. Discovery radiomics via stochasticnet sequencers for cancer detection. *arXiv preprint arXiv:1511.03361*, 2015.
- [102] Manu Sharma, Jignesh S Bhatt, and Manjunath V Joshi. Early detection of lung cancer from ct images: nodule segmentation and classification using deep learning. In *Tenth International Conference on Machine Vision (ICMV 2017)*, volume 10696, page 106960W. International Society for Optics and Photonics, 2018.
- [103] Shiwen Shen, Alex AT Bui, Jason Cong, and William Hsu. An automated lung segmentation approach using bidirectional chain codes to improve nodule detection accuracy. In *Computers in biology and medicine*, volume 57, pages 139–149. Elsevier, 2015.

- [104] Shiwen Shen, Simon X Han, Denise R Aberle, Alex AT Bui, and William Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. In *arXiv preprint arXiv:1806.00712*, 2018.
- [105] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging*, pages 588–599. Springer, 2015.
- [106] Wei Shen, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. In *Pattern Recognition*, volume 61, pages 663–673. Elsevier, 2017.
- [107] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- [108] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- [109] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [110] William F. Smith, Javad Hashemi, and Francisco Presuel-Moreno. *Foundations of materials science and engineering*. McGraw-Hill Publishing, 2006.
- [111] Anthony L Spek. Structure validation in chemical crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 65(2):148–155, 2009.
- [112] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *ICLR*, 2015.
- [113] N. Srivastava, G. E Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [114] Stanford. CS231:Convolutional Neural Network Class. <https://cs231n.github.io/understanding-cnn/>, 2017. [Online; accessed 19-Oct-2017].

- [115] Harold T Stokes and Dorian M Hatch. Findsym: program for identifying the space-group symmetry of a crystal. *Journal of Applied Crystallography*, 38(1):237–238, 2005.
- [116] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019.
- [117] Wenqing Sun, Bin Zheng, Xia Huang, and Wei Qian. Balance the nodule shape and surroundings: a new multichannel image based convolutional neural network scheme on lung nodule diagnosis. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 101343L. International Society for Optics and Photonics, 2017.
- [118] Kenji Suzuki. A supervised ‘lesion-enhancement’ filter by use of a massive-training artificial neural network (mtann) in computer-aided diagnosis (cad). In *Physics in Medicine & Biology*, volume 54, page S31. IOP Publishing, 2009.
- [119] Kenji Suzuki, Samuel G Armato, Feng Li, Shusuke Sone, and Kunio Doi. Effect of a small number of training cases on the performance of massive training artificial neural network (mtann) for reduction of false positives in computerized detection of lung nodules in low-dose ct. In *Medical Imaging 2003: Image Processing*, volume 5032, pages 1355–1367. International Society for Optics and Photonics, 2003.
- [120] Kenji Suzuki, Samuel G Armato, Feng Li, Shusuke Sone, et al. Massive training artificial neural network (mtann) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. In *Medical physics*, volume 30, pages 1602–1617. Wiley Online Library, 2003.
- [121] Kenji Suzuki, Junji Shiraishi, Hiroyuki Abe, Heber MacMahon, and Kunio Doi. False-positive reduction in computer-aided diagnostic scheme for detecting nodules in chest radiographs by means of massive training artificial neural network1. In *Academic Radiology*, volume 12, pages 191–201. Elsevier, 2005.
- [122] Hidenori Takahashi, Hironobu Tambo, Yusuke Arai, Yuji Inoue, and Hidetoshi Kawashima. Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. *PloS one*, 12(6):e0179790, 2017.
- [123] Zhiqiong Wang, Junchang Xin, Peishun Sun, Zhixiang Lin, Yudong Yao, and Xiaosong Gao. Improved lung nodule diagnosis accuracy using lung ct images with uncertain class. In *Computer methods and programs in biomedicine*, volume 162, pages 197–209. Elsevier, 2018.

- [124] Ted W Way, Lubomir M Hadjiiski, Berkman Sahiner, Heang-Ping Chan, Philip N Cascade, Ella A Kazerooni, Naama Bogot, and Chuan Zhou. Computer-aided diagnosis of pulmonary nodules on ct scans: Segmentation and classification using 3d active contours. In *Medical physics*, volume 33, pages 2323–2337. Wiley Online Library, 2006.
- [125] Alexander Wong, Audrey G Chung, Devinder Kumar, Mohammad Javad Shafiee, Farzad Khalvati, and Masoom Haider. Discovery radiomics for imaging-driven quantitative personalized cancer decision support. In *Journal of Computational Vision and Imaging Systems*, volume 1, 2015.
- [126] Yutong Xie, Jianpeng Zhang, Yong Xia, Michael Fulham, and Yanning Zhang. Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest ct. In *Information Fusion*, volume 42, pages 102–110. Elsevier, 2018.
- [127] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.
- [128] Matheus Zanon, Gabriel Sartori Pacini, Vinicius Valério Silveiro de Souza, Edson Marchiori, Gustavo Souza Portes Meirelles, Gilberto Szarf, Felipe Soares Torres, and Bruno Hochhegger. Early detection of lung cancer using ultra-low-dose computed tomography in coronary ct angiography scans among patients with suspected coronary heart disease. In *Lung Cancer*, volume 114, pages 1–5. Elsevier, 2017.
- [129] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.
- [130] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2528–2535. IEEE, 2010.
- [131] Tingting Zhao, Huafeng Wang, Lihong Li, Yifang Qi, Haoqi Gao, FangFang Han, Zhengrong Liang, Yanmin Qi, and Yuan Cao. A hybrid cnn feature model for pulmonary nodule differentiation task. In *Imaging for Patient-Customized Simulations and Systems for Point-of-Care Ultrasound*, pages 19–26. Springer, 2017.
- [132] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

- [133] Wentao Zhu, Chaochun Liu, Wei Fan, and Xiaohui Xie. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In *arXiv preprint arXiv:1801.09555*, 2018.
- [134] Kumar D. Scheffler M. Ziletti, A. and L. M. Ghiringhelli. Tutorial for insightful classification of crystal structures using deep learning. 2018.
- [135] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *ICLR*, 2017.

APPENDICES

Appendix A

Interpretabel Crystal Structure Classification

A.1 How To Represent A Material

In the case of crystal-structure recognition, it is essential that the descriptor captures system's symmetries in a compact way, while being size-invariant in order to reflect the infinite nature of crystals. Periodicity and prevailing symmetries are evident - and more compact - in reciprocal space, and therefore we introduce an approach based on this space. For every system, we first simulate the scattering of an incident plane wave through the crystal, and then we compute the diffraction pattern in the detector plane orthogonal to that incident wave. This is schematically depicted in Fig. 3.2a. For completeness, the formulation behind this procedure is explained in appendix A.2.

For each structure we first construct the standard conventional cell. Then, we rotate the structure 45° clockwise and counterclockwise about a given crystal axis (e.g. x), calculate the diffraction pattern for each rotation, and superimpose the two patterns. Any other choice of rotation angle is in principle valid, provided that the diffraction patterns corresponding to different crystal classes do not accidentally become degenerate. This procedure is then repeated for all three crystal axes. The final result is represented as one RGB image for crystal structure, where each color channel shows the diffraction patterns obtained by rotating about a given axis (i.e. red (R) for x -axis, green (G) for y -axis, and blue (B) for z -axis). Each system is thus described as an image, and we term this descriptor two-dimensional diffraction fingerprint (D_F). We point out that this procedure does not require to already know the crystal symmetry, and x , y , and z are arbitrary, e.g.

determined ordering the lattice vectors by length(or whatever the chosen criterion). For additional computational details on the descriptor D_F , please refer to the section Methods.

Moreover, its dimension is independent of the number of atoms and the number of chemical species in the system being represented. This is an important property because machine learning models trained using this descriptor generalize to systems of different size by construction. This is not valid for most descriptors: for example, the Coulomb matrix dimension scales as the square of atoms in the largest molecule considered [96], while in symmetry functions-based approaches [9] the required number of functions (and thus model complexity) increases rapidly with the number of chemical species and system size. Being based on the process of diffraction, the diffraction fingerprint mainly focuses on atomic positions and crystal symmetries; the information on the atomic species - encoded in the form factor f_a^λ in Eq. A.1 - plays a less prominent role in the descriptor. As a result, materials with different atomic composition but similar crystal structure have similar representations. This is the ideal scenario for crystals classification: a descriptor which is similar for materials within the same class, and very different for materials belonging to different classes. Finally, the diffraction fingerprint is straightforward to compute, easily interpretable by a human (it is an image, see Fig. 3.2c), has a clear physical meaning (Eqs. A.1 and A.2), and is very robust to defects. This last fact can be traced back to a well-known property of the Fourier transform: the field at one point in reciprocal space (the image space in our case) depends on all points in real space. In particular, from Eq. A.1 we notice that the field Ψ at point \mathbf{q} is given by the sum of the scattering contributions from all the atoms in the system. If for example, some atoms are removed, this change will be smoothen out by the sum over all atoms and spread over - in principle - all points in reciprocal space. Practically, with increasing disorder new low-intensity peaks will gradually appear in the diffraction fingerprint due to the now imperfect destructive interference between the atoms in the crystal. Examples of highly defected structures and their corresponding diffraction fingerprint are shown in Fig. 3.2e-3.2f. It is evident that the diffraction fingerprint is indeed robust to defects. This property is crucial in enabling the classification model to obtain a perfect classification even in the presence of highly defective structures (see below).

A.2 Crystal Formation Formulation

The amplitude Ψ , which originates from the scattering of a plane wave with wave-vector \mathbf{k}_0 by N_a atoms of species a at positions $\{\mathbf{x}_j^{(a)}\}$ in the material can be written as:

$$\Psi(\mathbf{q}) = r^{-1} \sum_a f_a^\lambda(\theta) \left[\sum_{j=1}^{N_a} r_0 \exp(-i\mathbf{q} \cdot \mathbf{x}_j^{(a)}) \right]. \quad (\text{A.1})$$

where r_0 is the Thomson scattering length, $\mathbf{q} = \mathbf{k}_1 - \mathbf{k}_0$ is the scattering wave-vector, \mathbf{x}' the corresponding position in the detector plane, and $r = |\mathbf{x}'|$ (see Fig.3.2a). Assuming elastic scattering, we have that $|\mathbf{k}_0| = |\mathbf{k}_1| = 2\pi/\lambda$, where λ is the wavelength of the incident radiation. The quantity $f_a^\lambda(\theta)$ is the so-called x-ray form factor; it describes how an isolated atom of species a scatters incident radiation with wavelength λ and scattering angle θ . Since x-rays are scattered by the electronic cloud of an atom, its amplitude increases with the atomic number Z of the element [?]. Following the successful application of scattering concepts in determining atomic structures, we propose the diffraction pattern intensity as the central quantity to describe crystal structures:

$$I(\mathbf{q}) = A \cdot \Omega(\theta) |\Psi(\mathbf{q})|^2. \quad (\text{A.2})$$

where $\Omega(\theta)$ is the solid angle covered by our (theoretical) detector, and A is a (inessential) constant determined by normalization with respect to the brightest peak (see section Methods).

Despite its rather complicated functional form (see Eqs. A.1 and A.2), the descriptor D_F is one image for each system being represented (data point); the eight crystal classes considered in this work (see below) and examples of their calculated two-dimensional diffraction fingerprints are shown in Fig. 3.2b and Fig. 3.2c, respectively. This descriptor compactly encodes detailed structural information (through Eq. A.1) and - in accordance with scattering theory - has several desirable properties for crystal-structure classification, as we outline below.

It is invariant with respect to system size: changing the number of periodic replicas of the system will leave the diffraction peak locations unaffected. This allows to treat extended and finite systems on equal footing, making our procedure able to recognize global and local order, respectively. We exploit this property, and instead of using periodically repeated crystals, we calculate D_F using clusters of approximately 250 atoms. These clusters are constructed replicating the crystal unit cell (see Methods). By using finite samples, we explicitly demonstrate the local structure recognition ability of our procedure. The diffraction fingerprint is also invariant under atomic permutations: re-ordering the list of atoms in the system leads to the same D_F due to the sum over all atoms in Eq. A.1.