Master's Thesis

"On the Explainability of Neural Network Models to Classify Rare Genetic Syndromes from Frontal Facial Images"

Aswinkumar Vijayananth

Submitted to Hochschule Bonn-Rhein-Sieg,
Department of Computer Science
in partial fullfilment of the requirements for the degree
of Master of Science in Autonomous Systems

Supervised by

Prof. Dr Paul G. Plöger
Prof. Dr Ralf Thiele
Prof. Dr. med. Dipl. Phys. Peter Krawitz

November 2022

I, the undersigned below, declare that this work has not previously been submitted to this or any other university and that it is, unless otherwise stated, entirely my own work.

Date

Aswinkumar Vijayananth

Abstract

Your abstract

Acknowledgements

Thanks to ....

Contents

List of Figures

List of Tables

1

Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

1.1 Motivation

1.1.1 ...

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

1.1.2 ...

1.2 Challenges and Difficulties

1.2.1 ...

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor.

Nulla in ipsum. Praesent eros nulla, congue vitae, euismod ut, commodo a, wisi. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Aenean nonummy magna non leo. Sed felis erat, ullamcorper in, dictum non, ultricies ut, lectus. Proin vel arcu a odio lobortis euismod. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Proin ut est.

Aliquam odio. Pellentesque massa turpis, cursus eu, euismod nec, tempor congue, nulla. Duis viverra gravida mauris. Cras tincidunt. Curabitur eros ligula, varius ut, pulvinar in, cursus faucibus, augue.

1.2.2 ...

1.2.3 ...

1.3 Problem Statement

1.3.1 ...

Etiam pede massa, dapibus vitae, rhoncus in, placerat posuere, odio. Vestibulum luctus commodo lacus. Morbi lacus dui, tempor sed, euismod eget, condimentum at, tortor. Phasellus aliquet odio ac lacus tempor faucibus. Praesent sed sem. Praesent iaculis. Cras rhoncus tellus sed justo ullamcorper sagittis. Donec quis orci. Sed ut tortor quis tellus euismod tincidunt. Suspendisse congue nisl eu elit. Aliquam tortor diam, tempus id, tristique eget, sodales vel, nulla. Praesent tellus mi, condimentum sed, viverra at, consectetuer quis, lectus. In auctor vehicula orci. Sed pede sapien, euismod in, suscipit in, pharetra placerat, metus. Vivamus commodo dui non odio. Donec et felis.

Etiam suscipit aliquam arcu. Aliquam sit amet est ac purus bibendum congue. Sed in eros. Morbi non orci. Pellentesque mattis lacinia elit. Fusce molestie velit in ligula. Nullam et orci vitae nibh vulputate auctor. Aliquam eget purus. Nulla auctor wisi sed ipsum. Morbi porttitor tellus ac enim. Fusce ornare. Proin ipsum enim, tincidunt in, ornare venenatis, molestie a, augue. Donec vel pede in lacus sagittis porta. Sed hendrerit ipsum quis nisl. Suspendisse quis massa ac nibh pretium cursus. Sed sodales. Nam eu neque quis pede dignissim ornare. Maecenas eu purus ac urna tincidunt congue.

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.

Maecenas non massa. Vestibulum pharetra nulla at lorem. Duis quis quam id lacus dapibus interdum. Nulla lorem. Donec ut ante quis dolor bibendum condimentum. Etiam egestas tortor vitae lacus. Praesent cursus. Mauris bibendum pede at elit. Morbi et felis a lectus interdum facilisis. Sed suscipit gravida turpis. Nulla at lectus. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Praesent nonummy luctus nibh. Proin turpis nunc, congue eu, egestas ut, fringilla at, tellus. In hac habitasse platea dictumst.

Vivamus eu tellus sed tellus consequat suscipit. Nam orci orci, malesuada id, gravida nec, ultricies vitae, erat. Donec risus turpis, luctus sit amet, interdum quis, porta sed, ipsum. Suspendisse condimentum, tortor at egestas posuere, neque metus tempor orci, et tincidunt urna nunc a purus. Sed facilisis blandit tellus. Nunc risus sem, suscipit nec, eleifend quis, cursus quis, libero. Curabitur et dolor. Sed vitae sem.

Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Maecenas ante. Duis ullamcorper enim. Donec tristique enim eu leo. Nullam molestie elit eu dolor. Nullam bibendum, turpis vitae tristique gravida, quam sapien tempor lectus, quis pretium tellus purus ac quam. Nulla facilisi.

Duis aliquet dui in est. Donec eget est. Nunc lectus odio, varius at, fermentum in, accumsan non, enim. Aliquam erat volutpat. Proin sit amet nulla ut eros consectetuer cursus. Phasellus dapibus aliquam justo. Nunc laoreet. Donec consequat placerat magna. Duis pretium tincidunt justo. Sed sollicitudin vestibulum quam. Nam quis ligula. Vivamus at metus. Etiam imperdiet imperdiet pede. Aenean turpis. Fusce augue velit, scelerisque sollicitudin, dictum vitae, tempor et, pede. Donec wisi sapien, feugiat in, fermentum ut, sollicitudin adipiscing, metus.

Donec vel nibh ut felis consectetuer laoreet. Donec pede. Sed id quam id wisi laoreet suscipit. Nulla lectus dolor, aliquam ac, fringilla eget, mollis ut, orci. In pellentesque justo in ligula. Maecenas turpis. Donec eleifend leo at felis tincidunt consequat. Aenean turpis metus, malesuada sed, condimentum sit amet, auctor a, wisi. Pellentesque sapien elit, bibendum ac, posuere et, congue eu, felis. Vestibulum mattis libero quis metus scelerisque ultrices. Sed purus.

Donec molestie, magna ut luctus ultrices, tellus arcu nonummy velit, sit amet pulvinar elit justo et mauris. In pede. Maecenas euismod elit eu erat. Aliquam augue wisi, facilisis congue, suscipit in, adipiscing et, ante. In justo. Cras lobortis neque ac ipsum. Nunc fermentum massa at ante. Donec orci tortor, egestas sit amet, ultrices eget, venenatis eget, mi. Maecenas vehicula leo semper est. Mauris vel metus. Aliquam erat volutpat. In rhoncus sapien ac tellus. Pellentesque ligula.

Cras dapibus, augue quis scelerisque ultricies, felis dolor placerat sem, id porta velit odio eu elit. Aenean interdum nibh sed wisi. Praesent sollicitudin vulputate dui. Praesent iaculis viverra augue. Quisque in libero. Aenean gravida lorem vitae sem ullamcorper cursus. Nunc adipiscing rutrum ante. Nunc ipsum massa, faucibus sit amet, viverra vel, elementum semper, orci. Cras eros sem, vulputate et, tincidunt id, ultrices eget, magna. Nulla varius ornare odio. Donec accumsan mauris sit amet augue. Sed ligula lacus, laoreet non, aliquam sit amet, iaculis tempor, lorem. Suspendisse eros. Nam porta, leo sed congue tempor, felis est ultrices eros, id mattis velit felis non metus. Curabitur vitae elit non mauris varius pretium. Aenean lacus sem, tincidunt ut, consequat quis, porta vitae, turpis. Nullam laoreet fermentum urna. Proin iaculis lectus.

Sed mattis, erat sit amet gravida malesuada, elit augue egestas diam, tempus scelerisque nunc nisl vitae libero. Sed consequat feugiat massa. Nunc porta, eros in eleifend varius, erat leo rutrum dui, non convallis lectus orci ut nibh. Sed lorem massa, nonummy quis, egestas id, condimentum at, nisl. Maecenas at nibh. Aliquam et augue at nunc pellentesque ullamcorper. Duis nisl nibh, laoreet suscipit, convallis ut, rutrum id, enim. Phasellus odio. Nulla nulla elit, molestie non, scelerisque at, vestibulum eu, nulla. Ut odio nisl, facilisis id, mollis et, scelerisque nec, enim. Aenean sem leo, pellentesque sit amet, scelerisque sit amet, vehicula pellentesque, sapien.

1.3.2 ...

1.3.3 ...

2

Background

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

2.1 Rare Genetic Syndromes

2.2 Phenotypes

2.3 Human Phenotype Ontology

3

State of the Art

3.1 Explainability Vs Interpretability Methods

Use as many sections as you need in your related work to group content into logical groups

3.2 Taxonomy of Explainability Methods

3.3 Explainability Methods for Neural Networks

3.3.1 Types

3.3.2 Choice of Methods for Further Evaluation

Occlusion Sensitivity Maps

Occlusion sensitivity Mapping is a model-agnostic perturbation based method, which generates explanations by manipulating parts of the input image. The approach is computationally expensive, O(#simultaneous occlusions * #features * #ablations_per_eval * 1/#strides), and is included in this work to verify if Gestalt Matcher model is focusing on key facial features, or simply using the surrounding context to produce predictions. This is achieved by systematically occluding different portions of the input image with a black square or rectangular mask, and computing the difference in outputs (logit scores of the target class). In this work, we use a black square mask of dimensions 10x10. Important portions of the input when occluded, result in relatively larger logit score differences, than the trivial ones. The differences are plotted on the image, yielding the occlusion sensitivity maps.

Deconvolution

Zeiler and Fergus proposed the “Deconvolution” approach to visualize and provide insights into the functions learned by intermediate layers of a CNN. It is one of the earliest attribution techniques, which produces visualizations based on computing gradient of loss function with respect to a given input. The work acts a baseline till date for development and evaluation of new pixel attribution techniques.
The method uses a deconvolution counterpart for every building block of a CNN, to obtain reverse mapping

from features to input pixels. The idea of deconvolution was first introduced by Zeiler et al. [1], as a way to perform unsupervised learning. In order to obtain attribution maps using the Deconvolution approach, the first step is to attach each block of convnet with its deconvolution counterpart as shown in the figure 1. Every activation except the ones belonging to the class of interest is set to zero. The activation value is then backpropagated through the deconvolution blocks such as unpooling, rectification and transposed convolution, all the way to the input layer. Deconvolution blocks act according to a pre-defined set of rules. The transposed convolution block performs the inverse of convolution operation by using transposed versions of the same filters. This is equivalent to flipping a given filter both in vertical and horizontal directions. In order to backtrack activations through max-pooling layer (.i.e. using the unpooling layer), indices corresponding to maximum activations in every layer, are first stored during the forward pass and later retrieved during the back propagation phase. However, the use of indices or switches from the forward pass, constrains the visualization on the input image [2].
Authors test their method on an AlexNet [] trained on the ImageNet [3], Caltech-101 [] and Caltech-256 [] and PASCAL2012 [] datasets. As a first step, they visualize the top 9 feature maps of the each of the first five layers, to show the proportional increase of complexity in features with respect to their receptive fields. The visualizations are obtained by backtracking the strongest activation of a feature map for most of the data samples, all the way until a given input, using the deconvolution rules.The paper also discusses about the proportionality between the time taken for a given layer to learn features and its corresponding depth. Further, it shows that features learned by top layers are more invariant to transformations like translations, rotations and scale changes.
The work evaluates itself by qualitatively comparing its resulting attribution maps with occlusion sensitivity maps. Occlusion sensitivity maps ([**?**]erturbation method) are obtained by systematically occluding portion of an image and analyzing the given classifier's output, to determine the most discriminative regions as shown in the first image of Figure 3.1.

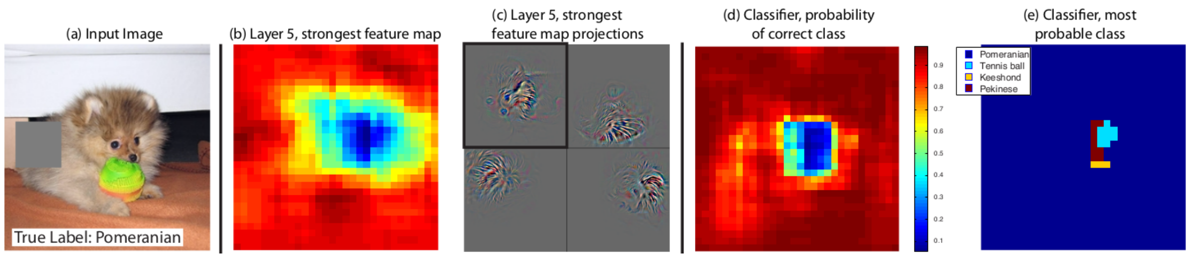

Figure 3.1: An illustration of occlusion sensitivity mapping

Saliency Maps

Simonyan et al. propose two visualization techniques with intents to generate an image which maximizes the class score, and to compute a class-specific saliency map for a given input. The first technique numerically optimizes the input image while the other computes the spatial support of a given class in an input. This work is one of the earliest to leverage saliency maps for the task of weakly-supervised object segmentation. Authors demonstrate the proposed techniques by applying to a deep convnet trained on the ILSVRC-2013 dataset. [46]

Class Model Visualization

The intention of this technique is to numerically generate an image which is representative of the target category with respect to the convolutional net's class scoring model. This is achieved by finding a L2-regularised image such that the logit S_c of a given class c is maximized:

where λ refers to the regularization parameter and I is a local optimum, which can be found with help of back propagation. The optimization process uses the mean image of the data set as the initial value. The work also mentions about the prominence of visualizations produced by using logit scores over soft-max/unnormalized class scores.

Image-Specific Class Saliency

The objective of this technique is to rank pixels of an input image, based on their impact on class scores (S_c). Authors provide a couple of interpretations for the class score values/ logits with respect to which saliency maps are created:

1. A linear approximation of the function learned by neural network in a local neighbourhood of the input image.

2. Higher the saliency associated with a pixel, lesser it needs to be altered in order to increase it respective class's score.

The derivative of class score with respect to input image is found using back propagation as described by the equation below:

In order to obtain the saliency map for a multi-channel image, the maximum magnitude of gradient for a given position across channels is used. Class saliency maps thus produced are used as initial points to compute object segmentation mask using the GraphCut algorithm. (**??**raphcut). Foreground and background portions are considered as Gaussian Mixture Models and the former is estimated from pixels with saliency value higher than the 95% quantile of the image's saliency distribution.On the other hand, the latter is estimated from pixels with saliency smaller than 30% quantile.
The work evaluates its outputs on test split of the data set for the localization task in the ILSVRC-2013 [] challenge, where it achieves 46.4% top-5 error in spite of its simplicity. In hindsight, apart from the strategy used to reverse the ReLU layer this approach is equivalent to Deconvnet.

Guided Backpropagation

Springenberg et al. proposed a new variant of the deconvolution approach in their work, as a means to analyze their "All Convolutional Net" architecture, which replaces max-pooling layers by convolutional layers with increased stride. The first objective of this work was to empirically prove the equivalence (in terms of predictive performance) between a max-pooling layer and a convolutional layer with an increased stride. This was achieved by evaluating a custom cnn model with max-pooling layers against its convolutional counterpart on three datasets: CIFAR-10, CIFAR-100 and ILSVRC 2012. In all cases, performances of the all convolutional models were on par with their max pooling counterparts. The second objective was to determine the quality of representations learned by the intermediate layers of the all convolutional neural network models. In order to achieve this, authors proposed a visualization approach, which can be considered as a slight modification of the Deconvnet [] technique.

Back propagation through ReLU

One of the most significant difference between Saliency Maps [], Deconvnet [] and Guided backprop [] approaches is the strategy used by these methods to backpropagate gradients through the ReLU layer.

- Saliency maps approach backpropagates gradients of positions with respective to non-negative activations.
- On the other hand, the deconvnet approach allows only positive gradients to flow in reverse direction.
- The guided-backprop approach combines the above mentioned methods and masks out values for which at least one of activation or gradient values is negative. This is performed with an intention to avoid the reverse of negative gradients of neurons which reduce the activation of the target neuronal unit.

Figure b illustrates differences between back propagation strategies with help of an example feature map.

The term “guided backpropagation” comes from the use of the additional navigation signal, to selectively back propagate only the positive gradients of the positively activated neuronal units. Though guided backprop was devised to show the learning capability of the all convolutional network architectures, authors show the effectiveness of the technique on the ones with max-pooling units. Guided backprop produced significantly more accurate representation, especially for higher layers, when compared to Deconvnet and Saliency maps.

Deep LIFT

Shrikumar et al. propose an attribution technique, which assigns scores to input pixels based on their contribution to change in activation of each neuronal unit with respect to a reference value, which is chosen based on the problem in hand. Authors claim the technique to be computationally inexpensive, and yield meaningful representations in comparison with other methods. Besides, the technique is devised in a way that it is suitable for neural net variants apart from CNN like Recurrent Neural Networks (RNN). Intuitively, Deep LIFT seeks to explain the deviation in output from some reference output in terms of deviation in input from its respective reference. The motivation to use a reference value, arises from the need to handle the “saturation problem”.

Saturation problem

The saturation problem can be explained intuitively with an example. The figure illustrates a simple neural network whose outputs saturates, when the sum of its inputs (i_1 and i_2) exceeds 1. Application of any perturbation or gradient based attribution method, to this scenario would lead to creation of undesirable artifacts. For the methods of earlier type, perturbing inputs will not cause any changes in the output. On the other hand, gradient based methods will suffer from lack of gradients in the region where $i_1 + i_2 > 1$.

Working

Deep LIFT handles this problem by using a reference value, enabling it to backpropagate the importance signal even in zero and discontinuous gradient situations.
The contributions computed by DeepLIFT is analogous to the idea of partial derivatives, except for the fact that change in input is computed with respect to a finite value (activation difference) in the earlier, in contrast to an infinitesimal value in the latter. The concept of "multipliers" is used to achieve the same. x is the input neuron with a difference from reference Δx, and t is the target neuron with a difference from reference Δt. C is then the contribution of Δx to Δt. Multipliers obey chain rule Along with the custom chain rule, the paper also defines a set of rules for neurons of every kind of neural network layer, to assign contribution scores their inputs:

- Linear rule for linear layers such as the fully-connected and convolutions
- Rescale rule for non-linear transformations such as ReLU, tanh or sigmoid
- Reveal-cancel rule which enables the measurement of marginal effect of having an input over all possible orderings, similar to Shapely values [].

The technique also takes in to account that a zero contribution could be due to cancellation of positive and negative contributions from a pair of entities. The application of DeepLIFT also depends on the choice of target neuron's layer, logit or softmax layer in a classifier network, for example. Authors demonstrate the approach by applying it to a CNN trained on MNIST dataset [] and an RNN trained on simulated genomic data.
The benchmarking on MNIST classifier for various methods was performed on basis of the change in log-odds score, when a selected pixels of a given image belonging to class c0 were erased to convert it to an image of some target class ct. The upper section of figure illustrates the result of masking pixels ordered as the most significant for converting to target classes 3 and 6. As graphically represented in bottom part of figure, DeepLIFT outperformed other backpropagation based methods.

GradCAM

Selvaraju et al. propose GradCAM, a pixel-attribution technique leveraging the gradients of a given target class with respect to a convolution layer. This approach can be considered as a generalization of CAM for

CNNs with fully connected layers. Besides, the technique is applicable to neural network architectures for tasks such as image captioning and visual-question answering. The work also comes up with a means to convert class-agnostic fine grained visualization approaches like Guided-Backprop and Deconvolution to be class-discriminative in nature.
Grad-CAM is one of the most widely used approaches in obtaining attribution map based explanations, for neural network models used with medical data (provide references). Concisely put, Grad-CAM creates a visualization of which parts of an input image a convolutional layer "looks" for a certain class prediction. The working of Grad-CAM can be described in the following steps:

1. Perform a forward pass of the input image through CNN to obtain logit scores for all classes.

2. Except for the logit activation of the class of interest, set other activations to zero.

3. Backpropagate the gradient of the class of interest all the way to the chosen convolutional layer containing k feature maps Ak. Del yc /delAk

4. **Global pooling:** For every feature map, weigh their pixels based on the gradient value, and obtain their weighted mean value scaled by constant Z where y_c refers to logit score for class c, A^k refers to the k^{th} activation map of dimensions ixj and α refers to the computed weightage.

5. Obtain the α weighted average of all pooled feature maps and apply a ReLu to filter out negative values, if any present. where Lc refers to the localization map produced by the Grad-CAM for the class of interest c and Ak refers to the kth feature map.

6. An element-wise multiplication operation is applied on the scaled version (input image dimensions) of Grad-CAM's maps, and outputs of fine-grained visualization approaches like Guided-Backpropagation and Deconvolution, to obtain fine-grained yet class discriminative visualizations. These combinations are termed as Guided-GradCAM and Guided-Deconvolution approaches respectively.

7. Guided Grad-CAM visualizations that are both high-resolution and class-discriminative"

8. Typically, last convolutional layers of a neural network model are chosen for Grad-CAM as they contain high-level semantics and detailed spatial information (**??**ame paper). This is because the attribution maps produced by the methods becomes progressively qualitatively worse as we use earlier convolutional layers which have relatively lesser receptive fields.

9. In a classification network, logit scores of the target class are used for gradient computations. However, any differentiable activation value can be backpropagated. The embedding based Grad-CAM discussed in the section, draws motivation from this fact.

10. In the original implementation of the attribution method, a Rectified Linear Unit (ReLU) function is applied on heat maps, to obtain only regions that positively affect the given prediction. However, to get better insights into prediction decisions, this work uses unfiltered heat maps, which consists of negatively correlated regions as well.

11. The work is evaluated from the perspective of three different tasks: weakly-supervised localization, weakly-supervised segmentation and Pointing Game experiment ([?]eep features for discriminative localization). The approach is also evaluated and bench marked on its ability to generate discriminative, trust-worthy, faithful and interpretable attribution maps. A couple of neural network models with different architectures (AlexNet and VGG-16) are used for evaluation, to determine the method's performance consistency across architectures.

12. Finally, the work also demonstrates the association of a given concept with a neuron, similar to the ones presented by (Visualizing and understanding convolutional networks, Object detectors emerge in deep scene CNNs.)

13. Authors also demonstrate it use in analyzing failure modes in neural network models and identifying bias in dataset.

14. The approach is considered to be computationally in-expensive when compared to perturbation based methods such as LIME or SHAP, yet producing interpretable and discriminative attribution maps.

4

Methodology

How you are planning to test/compare/evaluate your research. Criteria used.

4.1 Datasets

4.2 Choice of Syndromes for Evaluation

4.3 Neural Network Models

4.4 Design of Experiments

4.4.1 Overview

4.4.2 Objectives

5

Solution

5.1 Implementation Details

5.1.1 Patient-wise Attribution Maps

5.1.2 Attribution Maps for Clinical Evaluation

5.1.3 Similarity Maps

5.1.4 Composite Faces

5.1.5 Generic Attribution Maps

5.2 Implementation details

6

Evaluation and Results

6.1 Metrics

Describe the experiments/evaluation you are performing to analyse your method.

6.2 Qualitative Analysis

6.2.1 Patient-wise Maps

6.2.2 Syndrome-wise Maps

6.2.3 Composite Faces

6.3 Effect of Class Imbalance on the Explainability of Models

6.4 Quantitative Analysis

6.4.1 Use of Occlusion Sensitivity Maps

6.4.2 Eye-tracking based Evaluation

6.4.3 Similarity Mapping

6.5 Evaluation Summary

Describe the results of your experiments in detail.

7

Conclusions

7.1 Contributions

7.2 Lessons learned

7.3 Future work

A

Design Details

Your first appendix

B

Parameters

Your second chapter appendix

References

[1] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *2011 international conference on computer vision*. IEEE, 2011, pp. 2018–2025.

[2] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, p. e0130140, Jul. 2015.

[5] A. Binder, M. Bockmayr, M. Hägele, S. Wienert, D. Heim, K. Hellweg, M. Ishii, A. Stenzinger, A. Hocke, C. Denkert *et al.*, "Morphological and molecular breast cancer profiling through explainable machine learning," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 355–366, 2021.

[6] D. Castelvecchi, "Can we open the black box of ai?" *Nature News*, vol. 538, no. 7623, p. 20, 2016.

[7] S. A. Hicks, J. L. Isaksen, V. Thambawita, J. Ghouse, G. Ahlberg, A. Linneberg, N. Grarup, I. Strümke, C. Ellervik, M. S. Olesen *et al.*, "Explaining deep neural networks for knowledge discovery in electrocardiogram analysis," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.

[8] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn to pay attention," in *International Conference on Learning Representations*, 2018.

[9] H. Lee, S. Yune, M. Mansouri, M. Kim, S. H. Tajmir, C. E. Guerrier, S. A. Ebert, S. R. Pomerantz, J. M. Romero, S. Kamalian *et al.*, "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets," *Nature biomedical engineering*, vol. 3, no. 3, pp. 173–182, 2019.

[10] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2021.

[11] C. Molnar, *Interpretable Machine Learning*, 2019.

[12] F. Nunnari, M. A. Kadir, and D. Sonntag, "On the overlap between grad-cam saliency maps and explainable visual features in skin cancer images," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2021, pp. 241–253.

[13] X. Qian, J. Pei, H. Zheng, X. Xie, L. Yan, H. Zhang, C. Han, X. Gao, H. Zhang, W. Zheng *et al.*, "Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning," *Nature biomedical engineering*, vol. 5, no. 6, pp. 522–532, 2021.

[14] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. W. Battaglia, and T. P. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4967–4976.

[15] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine learning for healthcare conference*. PMLR, 2019, pp. 359–380.

[16] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[17] Z. Xue, S. Antani, L. R. Long, D. Demner-Fushman, and G. R. Thoma, "Window classification of brain ct images in biomedical articles," in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 1023.

[18] C. Li, D. Konomis, G. Neubig, P. Xie, C. Cheng, and E. Xing, "Convolutional neural networks for medical diagnosis from admission notes," *arXiv preprint arXiv:1712.02768*, 2017.

[19] F. Schwendicke, T. Golla, M. Dreher, and J. Krois, "Convolutional neural networks for dental image diagnostics: A scoping review," *Journal of Dentistry*, vol. 91, p. 103226, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0300571219302283

[20] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, no. 11, pp. 1–13, 2018.

[21] B. Ait Skourt, A. El Hassani, and A. Majda, "Lung ct image segmentation using deep neural networks," *Procedia Computer Science*, vol. 127, pp. 109–113, 2018, pROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING IN DATA SCIENCES, ICDS2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918301157

[22] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of covid-19 cases using deep neural networks with x-ray images," *Computers in biology and medicine*, vol. 121, p. 103792, 2020.

[23] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[24] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi *et al.*, "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[25] "Explainable AI," Mar. 2021, [Online; accessed 4. May 2022]. [Online]. Available: https://www.ibm.com/watson/explainable-ai

[26] T.-C. Hsieh, A. Bar-Haim, S. Moosa, N. Ehmke, K. W. Gripp, J. T. Pantel, M. Danyel, M. A. Mensah, D. Horn, S. Rosnev *et al.*, "Gestaltmatcher facilitates rare disease matching using facial phenotype descriptors," Nature Publishing Group, Tech. Rep., 2022.

[27] "GestaltMatcher Database," May 2022, [Online; accessed 5. May 2022]. [Online]. Available: https://db.gestaltmatcher.org/publications

[28] "Home - Face2Gene," Mar. 2022, [Online; accessed 5. May 2022]. [Online]. Available: https://www.face2gene.com

[29] Y. Gurovich, Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Basel-Salmon, P. M. Krawitz, S. B. Kamphausen, M. Zenker *et al.*, "Identifying facial phenotypes of genetic disorders using deep learning," *Nature medicine*, vol. 25, no. 1, pp. 60–64, 2019.

[30] "Human Phenotype Ontology," Apr. 2022, [Online; accessed 6. May 2022]. [Online]. Available: https://hpo.jax.org/app

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition.* Ieee, 2009, pp. 248–255.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[33] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection," *arXiv preprint arXiv:2003.10769*, 2020.

[34] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning.* PMLR, 2018, pp. 2668–2677.

[35] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[36] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[37] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017, https://distill.pub/2017/feature-visualization.

[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[39] Q. Ferry, J. Steinberg, C. Webber, D. R. FitzPatrick, C. P. Ponting, A. Zisserman, and C. Nellåker, “Diagnostically relevant facial gestalt information from ordinary photos,” *elife*, vol. 3, p. e02020, 2014.

[40] I. E. Nielsen, D. Dera, G. Rasool, N. Bouaynaya, and R. P. Ramachandran, “Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks,” *arXiv preprint arXiv:2107.11400*, 2021.

[41] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, “Layer-wise relevance propagation: an overview,” *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[42] L. Chen, J. Chen, H. Hajimirsadeghi, and G. Mori, “Adapting grad-cam for embedding networks,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2794–2803.

[43] D. Matthew Zeiler and F. Rob, “Visualizing and understanding convolutional neural networks.” ECCV, 2014.

[44] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan *et al.*, “A unified and generic model interpretability library for pytorch, 2020,” 2009.

[45] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.

[46] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[47] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *International conference on machine learning.* PMLR, 2017, pp. 3145–3153.

[48] Z. Zhang, Y. Song, and H. Qi, “Age progression/regression by conditional adversarial autoencoder,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5810–5818.