THESIS TITLE

by

Elizabeth Dethy

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2018

Author .....
Department of Electrical Engineering and Computer Science
September X, 2018

Certified by .....
Daniel J. Weitzner
Principal Research Scientist
Thesis Supervisor

Accepted by .....
Someone
Chair, Master of Engineering Thesis Committee

THESIS TITLE

by

Elizabeth Dethy

Submitted to the Department of Electrical Engineering and Computer Science
on September X, 2018, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Engineering

Abstract

In this thesis, I created and utilized a dataset to determine the interpretability of open source facial recognition algorithms as judged by the NetworkDissection technique. Determining the interpretability of these networks required constructing a dataset with labelled examples of disentangled facial features such as nose, eye, mouth, etc. Additionally, I created a dataset of skin tones to test for skin tones as a concept in the networks.

In addition to the technical work of this thesis, I researched the policy implications of network interpretability techniques and whether or not policy should require that networks provide meaningful insight into their behavior.

Thesis Supervisor: Daniel J. Weitzner
Title: Principal Research Scientist

Acknowledgments

This is the acknowledgements section. You should replace this with your own acknowledgements.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1

1.1.1

1.1.2

<++>

Chapter 2

Motivation

The goal of this thesis is to begin to bridge the gap between the policy community's work on machine learning accountability and the technical research into explainability. There are gaps in both the policy considerations of the explainability of machine learning as well as the technical research in the area

At a high level, in this section these are what I want the take-aways to be

1. There is a disconnect in the assumptions and direction in the technology explainability research

2. There is a profound lack of policy direction in requiring explanations for the use of these models (the focus has not been on explainability but rather accountability in general with different groups skewing to more or less regulation)

3. Technologists claim humans may not trust the technology without understanding how it produces it outputs, but the prevalence of such models proves otherwise

4. Technologists need to push the boundaries and limits of their explainability techniques, ideally to policy relevant areas and policy people need to realize that explainability should be written off as technically infeasible

5. Research in this area is limited by the "sample of one" problem and the lack of consistency around definitions and how to quantify explainability

Setting aside, for a moment, the European Union's General Data Protection Regulation that seems to enshrine a "right to an explanation", no laws exist to specifically regulate the use of machine learning in sensitive contexts. Areas of policy where laws may require the explainability of automated decision are only those for areas where explanations are already required. The vagueness of the GDPR, and the anticipation that the courts will be left to decide what ultimately counts as a valid explanation for an automated decision, has left researchers to create their own definitions of explainability and the reasons why explainability is necessary for machine learning algorithms.

The extensive work undertaken by researchers to increase human trust in these algorithms has proceeded alongside calls for increased fairness and transparency in cases where opaque algorithms are used. In the past two years, numerous examples of bias in widely used algorithms has gained widespread attention. The first was Julia Angwin's ProPublica piece analyzing bias in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a criminal assessment tool. The second was work by Joy Bowlami at MIT who revealed bias in the gender identifier algorithms provided by four major technology companies.

The public response, primarily outcry over the discriminatory nature of supposed objective systems, has led to new working papers on algorithmic accountability for think tanks, researchers, and non profits. An interdisciplinary effort at Harvard between the Law School, the School of Engineering and Applied Sciences, the Department of Psychology and Center for Brain Science, and the Ceneter for Internet and Society, produced a paper explaining what they believed to be the role of the explaination in AI Accountability.

There is an assumption in the technical literature and new reports about explainabil-

ity in machine learning that a human cannot trust the output of a model if they do not understand how it reaches its decision. While this may be true, opaque algorithms are used by human decision makers in extremely sensitive contexts today. Cities across the country have used criminal risk assessments in sentencing and black-box models to predict if a child is at-risk of mistreatment. In each of the aforementioned uses, a human is ultimately responsible for using the output of the algorithm in making a decision about an individual's life. In the former case, how long the defendent should spend in jail, and in the latter case, whether the child's family needs to be visited and monitored by a social worker.

Clearly, there is a gap if technology researchers claim a human cannot effectively use the output of an algorithm without understanding how it made its decision, yet the models are used in highly sensitive areas where individuals use the models even without understanding how the output was produced.

Chapter 3

Importance of Explanations in Real-World Usage

Main take aways for this chapter or flow for this chapter

1. In areas where there is little to no oversight, and humans are using the output of black box algorithms in sensitive areas, explanations should be required for all model outputs

2. Specifically, child services or some other extremely sensitive area

3. There is precedence for establishing humans must be able to effectively use the technology tools at their disposal (esp human factors in flight deck design - though I see that this could be too off topic and not the most relevant example)

4. There is also precendece for government regulation driving the technology forward rather than hamstringing it. See fuel efficiency standards that resulted in manufacturers producing cars that produced less smog. Regulation does not always hamper an industry or adoption of technology, even if some say it will

There are established precedents for creating regulations and guidelines to ensure humans can safely and effectively utilize the technology at their disposal. This includes both training measures for the individuals who will be using the technology as well as requirements of the technology itself. A critical example of this is aircraft

flight deck equipment design. The Federal Aviation Association publishes explicit guidelines for creating compliant flight deck system controls.

A circular published by the FAA in 2011 aimed to provide "guidance for the installation and airworthiness approval of flight deck system control devices, from a primarily human factors perspective." Federal regulations mentioned in published in the appendix of the report mandate things like, "Each control must operate with the ease, smoothness, and positiveness appropriate to its function" and "Controls must be located, arranged, and identified to prevent the possibility of confusion and subsequent inadvertent operation." Aircraft systems, even as automated as they are today, still rely on a pilot's understanding to safely and effectively work.

While the guidance on producing explanations for machine learning networks is minimal, there has been increased focus in the literature on two concepts. Firstly, the GDPR and the "right to an explanation", and secondly, the best way to realize the benefits of machine learning in many problem areas while providing for accountability and fairness in their use. The discussion of fairness and accountability was sparked by Julia Angwin's ProPublica piece that purported there was systemic bias in the COMPAS system against black individuals. More recently, Joy Bowlami at MIT demonstrated bias against black females in the facial recognition algorithms provided by four major technology companies. Additionally, the role machine learning plays in everything from health care decisions to determining if a child is in an at-risk home, has increased the prevalence of these opaque algorithms in our everyday lives.

Chapter 4

Related Work

The past few years have seen an explosian of papers from the technical community related to providing an explanation for a neural network's decision. These methods range from manipulating features of the input to see how they related to the output, to attempting to gain a deep understanding of the internals of the network. Work in this area includes research into developing interpretable facial recognition models.

This thesis focuses on applying a singular method of understanding the internals of a neural network to facial characterization networks. Thus, I will briefly survey three techniques focused on developing a comprehension for the internals of a deep neural network and research related to interpretable facial recognition in particular. The three techniques I will discuss are explanation generation, saliency mapping, and representation analysis.

4.0.1 Explanation Generation

Explanation generation is a technique used to provide human-readable explanations for the decisions of a neural network. As the network is trained on the desired task, a second network is trained using textual explanations to generate an explanation. Models trained in this way perform well on the assigned task and providing a human understandable explanation. Despite this, adding a second neural network to generate the explanation increases the complexity of the overall model and it is not known if

the generated explanations are correct or merely producing explanations that sound good to a human.

4.0.2 Salience Mapping

Saliency mapping is a technique that can be applied to deep neural networks for computer vision problems to identify features in an image relevant to classifying it. The output is a heat map of input sensitivity.

For a given image, portions of the image are occluded to determine which regions affect the network's output. Salience mapping provides useful information to understand why the network made a particular classification. However, saliency mapping does not provide insight into the higher level decision making of the network such as why a network classified a cat image as a cat and a bird image as a bird.

4.0.3 Representation Analysis

Representation analysis pertains to determining learned concepts in the hidden units of the network. The goal is to, for a given network, determine the concepts individual layers in the network learn. The technique to do this is called NetworkDissection and it is discussed in greater detail in **Section X**. I used the network dissection method to provide the interpretability of facial characterization networks.

4.0.4 Facial Recognition Interpretability

Work is being done in this space as well.

Chapter 5

Network Dissection

Network Dissection is a technique to quantitatively identify human-interpretable concepts learned by a neural network created at MIT by Bolei Zhou and David Bhao. Network Dissection uses a set of semantic concepts to align those concepts with the representations learned by individual convolutional filters of a deep neural network. It aligns the semantic concepts of a provided dataset against the activated sections of images obtained by a forward pass through the network.

Network Dissection proceeds in six steps, including preprocessing to construct the dataset. The preprocessing step takes a set of datasets with labels and generates a **Broad and Densely Lablled Databset (Broden)**. The second through fifth steps obtain the activation maps for each image in the dataset passed through the network and identify the label that best matches the activated parts of the images. The final steps rank the interpretability of each unit using Intersection Over Union and creating a visualization of the interpretable units.

5.0.1 Generating Densely Labelled Dataset with Segmentations

5.0.2 Forward Pass to get Activation Layer

Each image in the generated dataset is run through a forward pass to the desired convolutional layer of the network being interrogated. If multiple convolutional layers are being examined, a forward pass is completed to each of the desired layers. The output of the forward pass is an activation map that is saved in a mmap file to be used in the next step.

5.0.3 Calculation of Distribution of Individual Unit Activations

For each unit, the quantiles of the activation maps are sorted and saved. The top 0.5% activation maps for each unit are used in further evaluation to idenfity and label any disentangled concepts for the unit.

Each of the activation maps in the 0.005 quantile are scaled up using bilinear interpolation to the size of the original image. The interpolants are anchored at the center of the unit's receptive field.

Chapter 6

Assessing the Usefulness of NetworkDissection on Small Neural Networks with Small Feature Sets

6.1 Overview

The goal of this thesis was to ascertain the usefulness of the NetworkDissection method on a smaller network with a smaller feature set. To do so, I used two facial characterization networks as the target for Network Dissection.

NetworkDissection probes for disentangled features in a deep neural n

6.2 Quantifying Facial Characteristic Interpretability Using Network Dissection

6.3 Experimental Setup

To quantify the interpretability of facial recognition network I probed three facial characterization networks with two datasets. The first goal was to determine if disentangled facial features showed up in pre-trained models. The second goal was to

determine differences in results between different architectures, training task, and network output.

6.3.1 Datasets

Two datasets were used to determine the interpretability of the networks being probed. The first was the Broden dataset constructed for the NetworkDissection experiments described in the paper. The second was a dataset specifically created representing facial features and skin tones. The Broden dataset was described in the previous chapter on Network Dissection.

Broden Dataset

Move to Network Dissection Chapter The Broden dataset was compiled using the ADE, PASCAL, PASCAL Parts, PASCAL Context, OpenSurfaces, and DTD datasets. Broden generated segmentations of each image in the dataset that labelled each pixel with one or more label for the category the pixel represents: color, part, object, or texture.

Facial Feature Dataset Generation

Two disparate datasets were used to generate the **Note: I need a name for this dataset** BrodenFace dataset. I used the Labelled Faces in the Wild dataset with dlib's 68 coordinate facial landmark detector to label disengtangled units of a face. Specifically, I labelled the eyes, eyebrows, nose, mouth, and jaw features. I used dlib's facial bounding box feature to label a face attribute.

The labels for the lfw dataset were stored in two generated files. The first, denoted with the suffix appended to the original filename '_face'contained indicated the pixels in the image corresponding to the individual's face. The second, denoted similarly with the suffix '_facepart'contained the pixels labelled for the distinct facial features.

The process to generate the labels proceeded as follows. For each image in the lfw dataset, use the dlib library's bounding box feature to identify the part of the image

Facial Region	BGR Value
mouth	(0,0,1)
eyebrow	(0,0,2)
eye	(0,0,3)
nose	(0,0,4)
jaw	(0,0,5)

Table 6.1: BGR Values Corresponding to Facial Features

with the face. An image with BGR values (0,0,0) is created with the same size as the face image. The pixels in the new image that correspond to the pixes in the bounding box on the original image are given an BGR value of (0,0,1). The new image is then saved in BGR format.

In the second step of the process we provided labels for the facial features, I used the dlib facial landmark detector to identify the areas of the face corresponding to the eyes, eyebrows, nose, mouth, and jaw in the image. I used the opencv module to to set BGR values on the enclosed pixels in the respective regions. The BGR values used to label the facial regions are indicated in **Table x**.

Skin Tone Dataset Generation

The dataset of skin tone images was compiled from the UCI skin tone segmentation dataset. The dataset contained the 245057 samples of BGR values corresponding to both skin tone and not skin tones from face images. For the purposes of skin segmentation, I used the 50859 samples of skin tone BGR values.

Using eighteen representative RGB values for skin tone, I grouped the skin tone samples by Euclidean distance. For each of the values in the dataset, its distance to each representative color group was calculated. The sample was placed into the group it was closest to. For each grouped sample, I used OpenCV to create a 227x227 image with each pixel's RGB value that of the dataset. The designated grouping of the sample was denoted in the filname for the generated image. The nomenclature used for the skin tone groupings where 'color0' to 'color19'. Each sample was named according to its grouping and sample number (i.e. 'color0_0.jpg').

Table X shows the number of samples that were assigned to each grouping.

Figure X shows samples from each grouping of skin tones.

6.3.2 Generating BrodenFace Dataset

The BrodenFace dataset used to probe the facial characterization networks contained labelled images from the facial feature dataset, skin tone dataset, and the dtd dataset. The facial feature and skin tone datasets have been described in the previous two sections. The DTD dataset is a collection of textures in the wild.

I enhanced the joinseg utility in NetworkDissection to create segementations for the new facial feature, skin tone, and dtd datasets.
The final dataset contained labelled examples of the following categories: face_part, skintone, texture, and color.

add description of what it does here

6.3.3 Models Interrogated

Two open-source pre-trained facial characterization networks were probed. Both were created and trained by Gil Levi, et. al, and described in the paper, 'Age and gender classification using convolutional neural networks'.

Both networks were convolutional neural networks with the same architecture. **describe the network architecture**

6.4 Results

Chapter 7

Policy Implications of Interpretability

7.1 Policy Implications of Interpretability

7.1.1 uh huh ok cool

7.1.2 so there's something

Appendix A

Tables

Table A.1: Armadillos

Armadillos	are
our	friends

Appendix B

Figures

Figure B-1: Armadillo slaying lawyer.

Figure B-2: Armadillo eradicating national debt.