NEURO140
Final Report

Applying perceptual modulation for qualitative interpretability of deep learning models in sex identification on retinal fundus images

Kamil Kon

May 7th, 2024

Abstract

In this project, we explore the prospects of using targeted adversarial attacks as a method of explainability for CNNs. A 19,000 images large dataset consisting of retinal fundus images has been trained on a ResNet50 architecture for a binomial sex classification. Despite the poor performance of the model on the small dataset, application of pixel perturbations aligns with the results of other saliency maps used in earlier research and shows promise for future applications.

1 Opening remarks

As I hand in this final project, the results are not finished to the extent I'd expect when first undertaking it. I approached NEURO140 with almost no experience of coding in Python (not to mention machine learning) and from the beginning facing it as mostly an exploration of the field. I leave this project as a computational neuroscience concentrator, but the departure from the course is not a departure from this work and I certainly hope that we will be back in touch when a potential project can become a research paper living up to the potential of the topic.

2 Hypothesis

Since its discovery by Poplin et al. in 2018, the ability of CNN's to determine sex from retinal fundus images with extremely high accuracy has been a recurring topic for research, especially considering the inability of professional opthalmologists to complete the same task. The nature of black box deep learning models, however, deem the workings through which a CNN makes its determination not interpretable. Throughout the years, various explainability methods have been applied in research to explore which parts of the retina contain the anatomical differences related to sex. This project will contribute to this academic debate by using adverserial modulation as a method of explainability.

The project will explore the hypothesis that images of the retinal fundus modified by adverserial attacks introducing specific and targeted pixel perturbations can be used for qualitative analysis and enhance human understanding of the significant areas similarily to other explainability models such as GradCAM. Furthermore, the promising research on use of pixel perturbations to simulate human vision gives hope that this explainability model can outperform others in accuracy and clarity of answer.

3 Literature review

In an attempt to create a non-intrusive and cheap diagnostic method for cardiovascular risk factors, Poplin et al. (2018) discovered that machine learning models working with datasets of retinal fundus photographs succeed in detecting various risk factors, including gender. Significantly for the field of machine learning, deep learning models have been found to perform much better than professional ophthalmologists, with certain research studies achieving to create models with an accuracy of AUC 0.99 (Taha et al. 2022). The weak human performance in this case is explained by the lack of obvious or salient features in the retinal fundus images which can be used for sex determination, creating the need for an explanation of the workings of black-box machine learning models doing the classification.

In the currently existing research exploring explainability of such models, a plethora of articles point out various areas of the retina as responsible for sex

determination, as briefly outlined in the table below:

Study	**Identified anatomic feature**
Poplin et al. 2018	optic disc, retinal vessels, macula (also some signal distributed throughout the retina)
Korot et al. 2021	fovea, optic nerve, vascular arcades
Ilanchezian et al. 2021	macula (female features), optic disc (male features)
Dieck et al. 2020	optic disc (80 percent of the cases) and macula (50 percent)
Rim et al. 2020	optic disc, retinal vessels
Betzler et al. 2021	optic disc

In their experimentation with explainability methods, both Poplin and Rim et al. (2020) used soft attention for creation of saliency maps, resulting in the identification of the optic disc and retinal vessels as the most significant elements. Dieck et al. (2020) screened for novel anatomic features by subsequently occluding parts of the test image to detect significance through analyzing which occlusions affect the performance of the model the most, leading to the conclusion that the optic disc is the most significant with some significance of the macula (although in contrast to the optic disc, no clear structural diferrences within the macula could be identified). Korot et al. (2021) was the first one to employ a region-based saliency map (XRAI) that suggested the significance of the optic nerve and vascular arcades. Furthermore, the inclusion of samples with foveal pathology in the dataset and comparision in performance between a healthy and disease-inflicted dataset led to the additional conclusion that the fovea is a salient region for sex classification. Ilanchezian et al. (2021) similarly used an explanation technique which considers local evidence without considering global relationships called BagNet, which concluded that the macula and the optic disc contain sex-indicative characteristics for females and males respectively. Betzler et al. (2021) used the fairly common gradient-based saliency mapping method known as GradCAM, with which it concluded that the optic disc is of interest for gender prediction.

Gaziv et al. (2023) contributes to the field of CNN explainability by introducing the concept of "robustified ANNs" which, by being trained on datasets featuring samples subject to adverserial attacks, "have internal neural representations more closely aligned with the ones of primate ventral stream" and can be hence used to understand the vulnerabilities of natural vision. Gaziv's experiments with Targeted Modulation (attempting to make models and humans misclassify an image as some other, specific category) showed the capacity of robustified models to also "shift human percepts toward prescribed target categories from arbitrary start images" in a pixel budget-efficient manner. The results support the researcher's conclusion of there existing "wormholes" in image spaces, meaning specific areas targeted for perturbations that can change

the perceived category, and the notion that robustified ANNs are strong enough to locate such wormholes. In the field of explainability, named wormholes serve as markers for features determinant for classifications and hence, in qualitative analysis, features that "make something what it is."

4 Methodology

The retinal fundus images used in the research project came from the public ODIR dataset [original size 4784, actual size after cleaning up images with retinal diseses 3099] and the semi-public BREST dataset [size 16266, no cleaning up commited]. Although the original outline of the project seeked to create a composite model through an ensembling method, where the weights of the composite model would be a mean sum of the best performing individual models for the ODIR and the BREST dataset respectively. Although such an approach has shown to create better generalizing models (of significance to the ultimate goal of this study) and better-performing models on a small amount of data, no success was achieved in creating a well-performing model on the ODIR dataset in the first place (with the peak accuracy 95% for training and 60% for validation) and hence a different approach was taken in the form of creating a composite dataset of 19,364 images to train a singular model on. The loaded data was pre-processed using tools from the cv2 library to crop the images around the retina by finding retinal contours and then scaled to 300x300 resolution. Within the PyTorch framework, the data was then augmented using a random horizontal flip with p=0.5 as well as normalization based on the mean and standard deviation pixel values found for the ODIR dataset:

$$\mu_{(r,g,b)} = [0.3004, 0.1870, 0.1007] \tag{1}$$

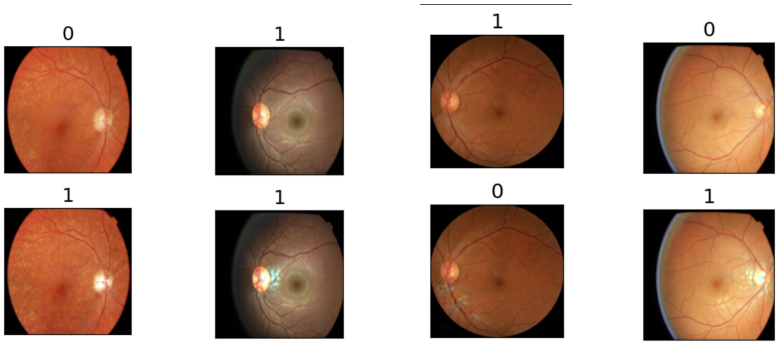
$$\sigma_{(r,g,b)} = [0.2777, 0.1806, 0.1034] \tag{2}$$

The data, split into a training and validation set (80/20) was trained on a ResNet-50 architectured, preloaded on ImageNet weights and modified last layer to feature a linear layer going from 256 neurons to a single output neuron.
Upon completion of training of the model, the weights were exported and used with another ResNet50 model in a separate Jupyter Notebook where the robustness library was used. The adverserial model training was conducted with all the values set at the default values suggested in the documentation, for the training to then be conducted for 5 epochs.

The code can be accessed via the following GitHub repository:
https://github.com/kamiljkon/Neuro140FP

5 Results and discussion

As expected, the trained ResNet50 did not achieve great results due to a plethora of factors, the size of the dataset being the primary one, achieving peak 62.3%



(a) M to F (b) F to F (c) F to M (d) M to F

accuracy in validation. Similarily, during the adverserial model training, peak adverserial accuracy was only 56.8%.

In the figure above are selected examples of targeted modulation attacks. Although the poor performance of the original model and the quality of the data do not allow any large scale qualitative analysis and using this as a sole explainability method, the figures above show a lot of promise as they introduce targeted attacks covering areas that previous research has shown to be significant for sex classification.
In figure (a), clear pixelization can be seen in the optic disc, implying that the optic disc is a significant area for female sex classification, as implied by almost all of the studies outlined in the literature review. In figure (b), a similar modification can be observed although horizontally shifted, introducing an interesting prospect (that once again, due to the quality of the model can not be taken too seriously) that the location of the optic disc might be of significance for sex classification. Figures (c) and (d) shows pixel perturbations along the major lower blood vessels, confirming the observations of Poplin, Korot and Rim.
In summary, the results confirm the hypothesis that targeted pixel perturbations can be applied as an explainability method, as proven by the correlation in terms of anatomical features pointed out with other explainability methods used in research. Due to the model's low performance, however, the model observed in example images (see the github repo for more) might not outperform other sorts of saliency maps.

6 Improvements

Being conscious of the current under-performance of the models and self-awareness of the project's shortcomings, there are improvements to the applied approach that could be implemented upon further work on the project:

- **Dataset**: as other research on datasets of similar size show only somewhat better performance of the model, one of the key improvements for future development of the project would be definitely obtaining access to a

larger amount of datasets of retinal fundus images. Furthermore, a proper and meticulous pre-processing of the data removing possible datapoints that would decrease performance of the image, as has been done in the case of the ODIR dataset, would be another natural improvement. Even within the framework of the small dataset used here, a more detailed pre-processing in form of cleaning up of the BREST images and balancing of classes could improve the model performance.

- **Training time**: due to hardware limitations (inability to run either of the models on more than one worker within the dataloader), neither the original training of the ResNet50 model, nor the adverserial training could be performed to the full potential of the dataset due to too-short, and not experimented though, training times.
- **Comparision with other explainability methods**: a natural next step for this project would be conducting a qualitative analysis of targeted modulation as an explainability methods through comparision to other explainability methods, such as GradCAM or standard occlusion, working on the same set of sample images.