Improving Model Opinion Interpretability

Similarity measurement of input to average training data

(SMI2ATD)

G.H. de Carvalho

s3450295

April 2021

Bachelor Project Proposal

Artificial Intelligence

University of Groningen,
The Netherlands

Supervisors:

Prof. Dr. L.R.B. Schomaker
Artificial Intelligence,
University of Groningen

Ming Wai Yeung
PhD Student, University
Medical Center Groningen

Jan Walter Benjamins
PhD Student, University
Medical Center Groningen

1 Introduction

Explainable AI has been receiving increasing attention in the last 10 years. This branch of AI tries to tackle problems that exist in the interface between users and AI systems. The aim revolves around describing model behaviour or decision making, understanding when a system is mostly right or very wrong in its output, how to correct the error and how trustworthy a system is [19]. This is especially relevant in the field of Medical Sciences, where the margin of error can be the threshold between life and death, a successful treatment or a failed one, and a correct diagnosis and an incorrect one. This makes it an extremely sensitive area to deploying AI systems in practice or amidst existing workflows. But also, one that would prospectively benefit many times from doing so due to data dependence and abundance, technological involvement and many opportunities to automate and improve existing medical techniques.

Explainability used to mean a trade-off between prediction accuracy and model interpretability, but not anymore, with many technologies being developed and having been developed in the last decade that auxiliate model performance and deployment. For example, in deep learning a technique called *deep explanation* has been developed that maps an NLP model (RNN) capable of generating written explanations of models like CNNs and their decisions (T. Darrell, P. Abeel (UCB)) thus teaching models how to learn semantic associations in data features (H. Sawhney (SRI Sarnoff)). The two models output a classification together with the associated image descriptions of discriminative features that led to the output label [15]. In model interpretation, stochastic And-Or-Graphs (AOG) has been developed by Song-Chun Zhu, from UCLA (*University of California, Los Angeles*) with the same aim as the previous research. This technique creates a five-layer AOG that maps semantic meanings between learned patterns [16]. Another technique called Probabilistic Program Induction attempts to describe parameters involved in character generation (e.g. in written symbols) by using a generative model that recognizes these symbols by developing an explanation of how a new character may be created through a sequence of probable strokes that is deduced from the training cases [17]. Finally, there are model induction techniques. These refer to techniques that attempt to infer explainable models from other more complex models that are viewed as black-boxes[14] (uninterpretable models). One such technique is LIME (Local Interpretable Model-agnostic Explanations)[18]. This algorithm explains classifier predictions using locally approximate interpretable models that are then selected by a method that chooses a set of locally faithful representative instances of explanations of the whole model.

To illustrate an example of an interpretability problem tackled by this thesis, take a 250 by 250 image. Initially, it will be represented by 62500 features in a CNN. These get reduced through convolutions and max-pooling, for example, into a smaller feature representation. In the case of W. Bai, et al. 's [4] model used in this thesis, 16 convolutional layers downsample the feature map every two or three convolutions by a factor of 2. This means that the model will have reduced the image to approximately 245 features at the bottleneck layer. At this point, the model reached the minimal feature representation of the data that it can use to reliably classify the data. These are 245 reasons why the image is classified with a particular label as the output of the model, all possibly contributing differently to the classification. If we take a visualization technique to try and make sense of this feature space that means 245 dimensions would be seen - an explanation space that is made too complex to interpret by the 242 extra dimensions added on

Statistics famously reduces facts about complex data composed by inordinate amounts of singular data points to readily understood numbers (e.g. μ as a distribution's mean, the T value of a T-test or the P-value). There is then a need for such summarizing values or concepts that work as sense-making mechanisms that allow for peering into the machine's mind, figuratively speaking. Especially in a field where most explanatory techniques involve model training and deployment which can easily lead to a recursion problem (which model explains the model that is supposed to explain the production model…?).

This is one of the central problems regarding ML models interpretability - they are regarded as black-box models [14]. In such models, independently of knowledge concerning the mathematical principles involved in the system or model, there is no explicit knowledge representation that is readily understood by humans, nor is there knowledge regarding underlying reasoning, the causal chains that lead to the output or their explanations that do not involve model making [19]. There is a need to work towards making AI systems transparent, explainable and error sensitive (GDPRand ISO/IEC 27001, 2018). Moreover, in the case of Medicine as a domain of application, this is especially relevant, where data complexity and heterogeneity is rampant. Medicine can then be seen as one of the best thresholds for AI explainability because professionals will not accept anything below a gold standard. This implies that models have been extensively tested, error margins must be negligible and it has to be possible for professionals involved to understand the reasoning behind the model's opinion.

In this paper, a training-independent technique will be deployed to further substantiate the quality of model output, its generalizability, and interpretability. Using an MRI pipeline consisting of CNNs [4] trained on the Uk Biobank dataset [6]. The measurement will elucidate how familiar the model is with a new image by measuring the difference between it and the training dataset used (the feature knowledge of the model) so as to enable some interpretation of the quality of the models' output as well as measuring its generalizability before trying a large batch of new images from a novel dataset. This is supposed to deploy simple mathematics, no model production nor training. If the measurement - the distance between images - is too large, then the model probably doesn't generalize to the novel dataset. Then, a good dimensional space is required to enable feature extraction. PCA is the method of choice for many when it comes to dimensionality reduction of data (including image data). This is the benchmark used in this thesis. To show that using the bottleneck layers of a trained model is a better representation than the previous method when it comes to dimensionality reduction steps. These two methods will be deployed to obtain a feature vector of an input cardiac magnetic resonance image. Then, the similarity measurements will be used on both to compare it to a prototype vector that represents the data, this can be obtained by simply averaging the image vectors used in the training dataset element-wise. This is the same as calculating the centroid of the data. The vector that generates the results that are closer to the ground truth is the best according to this quantitative measure. Ground-truth will be the input image chosen. Images not included in the UKBB [6] dataset should have a large distance from the prototype and vice-versa: images from the UKBB dataset should show a short distance from the prototype. Of course, if it is the case that both vectors yield results contrary to the ground-truth then the experiment is taken to have failed (e.g. the distance between an image used in training the model is used but the resulting distance measurement is very large). After training the model on UMCG data, the measurement is deployed again and it is expected to have shortened.

2 Research Questions

The aforementioned ML pipeline [4] will be deployed for processing and analysing cardiovascular magnetic resonance images using CNNs from the UK Biobank [6], working towards showing transfer learning and model deployment. This pipeline is referred to as the '*ukbb_cardiac*' which can be found in the following Github repository. The models used and developed in the pipeline have gold-standard classification accuracy in the UK Biobank dataset. To check to what extent the model is reusable UMCG data will be fed to it.

Research Question 1: Does the *ukbb_cardiac* toolbox generalise (achieve gold-standard 84.3% classification accuracy) to a UMCG dataset from the UK Biobank?

Hypothesis: The *ukbb_cardiac* toolbox will not generalise to UMCG data.

To test the SMI2ATD as a preventive check of model generalizability a second research question is asked. For this hypothesis, known methods in computing distances of similarity between two images will be calculated (e.g. Hamming distance, Mahalanobis distance, and Euclidean distance) to compare performance between using PCA-generated vector images and a feature vector directly read from the trained model's bottleneck layer.

Research Question 2: A training independent technique that produces a similarity measurement of novel input to average training data point has better accuracy when deploying a feature vector read from the bottleneck of the trained model than when using PCA for image feature reduction as measured by benchmark distances.

Hypothesis: Deploying a feature vector read from the bottleneck of the trained model is better than using PCA for image feature reduction as measured by benchmark distances. Using the bottleneck generated vector as input results in a data-independent technique that produces a similarity measurement of novel input to the average training data point better than using a PCA reduction. This is the case since the former is capable of non-linear dimensional reduction in the case of large feature spaces, while the latter is not.

3 Methods

i) Replicating an automated cardiovascular MRI analysis pipeline with fully convolutional neural networks and testing the results of transfer-learning using a new dataset

The first objective is to test W. Bai, et al. 's[4] pipeline with UMCG (Universitair Medisch Centrum Groningen) data and verify if it generalizes well from the UK Biobank dataset to UMCG data.

A. Implement W. Bai, et al. 's [4] pipeline (particularly the short-axis model).
B. Test the performance of the model on a new dataset generated from a different site with a different machine protocol (UMCG). This should be done using the following quantitative assessments also deployed in W. Bai, et al.[4]:
    a. Distance measures;

C. Perform transfer learning using UMCG data, more specifically the short-axis model that classifies left atrium (LA) and right atrium (RA) on CMR images. Compare performance with that obtained in B.

(ii) Visualization wrapper to monitor the training process

Apply real-time learning curve plotting using tensorboard *(ii)* for a more explainable training procedure. The goal is to create a live graph of the two learning curves (training and validation learning curve), thus assisting in model interpretability as well as increasing research interaction with the learning process.

A. Incorporate a tool to monitor training processes using tensorboard. This enables tracking of cost metrics (loss) as well as visualization of these two learning curves (training and validation learning curve).

(iii) Similarity measurement of input to average training data (SMI2ATD)

Given a new image, the program should check how similar the image is to the data that it was trained on. This will serve as a similarity measurement of novel input data to the training data so that the model can make an informed judgement on the quality or certainty of the new input data's similarity to its prior knowledge, and thus the subsequent classification.

A. Calculate average vector from training data sample (PCA-prototype-vector or PCAPV) by using PCA to reduce each image and average the resulting vectors element-wise.
B. Repeat 1. Using model bottleneck layer instead of PCA for dimensionality reduction (model-knowledge-prototype-vector or MKPV).
C. Use PCAPV and MKPV to reduce a UMCG image.
D. Calculate similarity measure using traditional metrics (e.g. Euclidean distance) between UMCG reduced vectors and…
    a. PCAPV;
    b. MKPV.
E. Check differences between the results from *D-a)* and *D-b)*.
F. Train the model on UMCG data and repeat A to E. Training strategy follows from the transfer-learning conducted in W. Bai, et al.[4]: randomly splitting data into 80% subjects for fine-tuning and 20% for evaluation (This will be implemented in part *i) C)*.
G. Compare the final results between vectors generated before training on UMCG data and after training.

4 Relevance

Deep learning (DL) is the most widely used Machine Learning (ML) approach, seeing plenty of applications across virtually every field, most prominently in visual object recognition and natural language processing (NLP). Models learn useful representations and features automatically, directly from data, bypassing an otherwise manual and conceptually difficult step of feature design and extraction. These automatic techniques have advantages particularly for high-dimensional data such as statistical images used in medical research. MRI is one of the fastest-growing research fields where ML meets a better-than-human performance in a clinical application, more specifically in the analysis of the output of

a complex imaging technique [13]. CNNs, or convolutional neural networks, are the preferred algorithm in MRI analysis[8]. In a nutshell, automated MRI analysis using a fully convolutional network (FCN) architecture predicts a pixel-wise image segmentation (semantic segmentation) by applying a number of convolutional filters onto an input image. The convolutional layer maps local connectivity allowing the network to learn the spatial local correlation of the input. Where for each pixel there is a higher correlation to its neighbouring pixels than to distant pixels. The model then learns image features from fine to coarse levels using these convolutions and combines multiscale features for predicting the label class at each pixel.

One problem that assails the multidisciplinary field of ML, in general, is the problem of model interpretability. Prospective applications of AI systems and efforts by the ML research community that could improve and automate diagnosis, for example, which can save lives, see these efforts cut short (rightly so) by Medical governing bodies and representatives on the basis of model interpretability and lack of generalizability (e.g. claiming that such a system is a ’black-box’, where it is ultimately impossible, not to know, but to completely understand causal links and processes that generate the output).

SVMs (Support Vector Machines) and other automatic techniques often serve preliminary data analysis before feeding data to the main classifier [4]. By comparing new input to the training set used to fit the model in question, these methods usually give a dichotomous answer (yes or no) to the question "is this data similar enough to the training set?". This is so that researchers know when further learning is required. Two problems are relevant in the context of this thesis: these methods need to be retrained on new data every time they are used - just like the classification model they serve. This is a problem because data is scarce, expensive and time-consuming to handle, as is the training involved in such preliminary models. The second problem arises from the fact that this binary check, like the served classifier, does not give reasons for its result, thus making it what some in the field refer to as a 'black-box'. This is analogous to saying that the process or computation that led to the output after feeding the model with some input is hidden, unexplained or too complex to make sense out of.

SMI2ATD tries to address the first problem by answering the same question without needing training time - it's training independent. And it tackles the second problem through its simplicity - no longer does a model need to be explained or understood since the model is removed and in its place simpler well known mathematical computations are used instead (e.g. Hamming distance, Mahalanobis-Abstand distance, and Euclidean distance).

The idea here is to work towards AI explainability and reinforce the use of a technique that is more general than using a ML model in checking the quality of the output of another model. This increases generalizability since the method is training-independent, thus reducing the resources that would have been necessary for creating a preliminary model that is not reusable (e.g. time, data collection, money and computational resources).

Furthermore, in the recurrent case of model-expert disagreement, where the model classifies a certain input as x but a human expert disagrees, taking the input to be a y instead, this technique helps elucidate

how certain the model is of its conclusion so as to solve the disagreement between the human and the machine.

5 Resources and Support

The *ukbb_cardiac* pipeline consists of a data preparation module for training, one for validation and another for classification of MRI images from the UK Biobank dataset[6]. After testing it using a different validation dataset provided by the UMCG, a different dataset will be used for retraining the model. To do this, the UMCG will also provide a machine learning-ready computer. Live plotting will be deployed using *tensorboard* - a python library for visualizing training built on top of *TensorFlow*. More data will probably be necessary so as to have a trifecta approach to data partitioning, enabling at least two holdout sets (one for training and validating from the UK BB, another from the UMCG for validation and the third one for further validation).

In regards to the distance measurements, these are useful in ML for comparing two objects of the same domain in terms of their relative difference. The library *SciPy* has readily available methods for calculating these. They are at the core of some very useful algorithms in ML, like KNN (K Nearest Neighbors), K-means and SVMs. Euclidean distance computes the distance of two real-valued vectors. It is the square root of the sum of squared differences across the two vectors element-wise. The Manhattan distance calculates the shortest distance between two coordinates - it can be useful, for instance, in the place of a Euclidean measurement if the vectors exist in an integer feature space instead of floating-point feature space. It is the sum of absolute differences across vectors. For the effects of this thesis, Euclidean distance will be sufficient since segmentation patterns should be similar enough across subjects and the calculation involves the reduced representations of the images, not an unprocessed image. The underlying assumption being that a small difference in the vector space equals a small difference in the actual image.

Finally, for evaluating the accuracy of the retrained model using UMCG data, the same quantitative measurements deployed by W. Bai, et al.[4] will be used, namely, the Dice metric outputs a value between 0 and 1 and is defined as:

$$\frac{2|A \cap B|}{|A|+|B|},$$

used to calculate the overlap between an automated segmentation A and manual segmentation B[4], the higher the outputted value the closer the segmentations, thus the better the result.

The two other metrics are the contour distance and Hausdorff distance. The former evaluates the mean distance and the latter calculates the maximum distance between segmentation contours. Here, the lower the distance metric, the better the agreement between segmentations for both metrics.

6 Planning

Dates	**Tasks**
1 Feb. - 1 Mar.	Literature review, research thesis topic, check available datasets.
1 Mar. - 1 Apr.	Discuss findings from review with UMCG and RUG supervisors, agree on a thesis and write the proposal.
1 Apr. - 15 Apr.	Clone Git repository and fix the library inconsistencies and incompatibilities (dependency issues). Visualize and understand the data from both sets (UK BB and UMCG).
15 Apr. - 1 May	Test pipeline with UMCG data and check poor generalization (I). Deploy similarity measure (II) using PCA and the model bottleneck layers to check the correlation between the resulting classification of UMCG data and the measurement results.
1 May. - 15 May	Retrain the model using UMCG data while deploying real-time plotting to improve explainability and observe the learning curves (II).
15 May - 15 June	Deploy similarity measure (II) using PCA and the model bottleneck layers again and show a correlation between the measurement and the classification results.
15 June - 1 July	Finalize experiments and code. Write the thesis.

7 References

[1] Michel Jose Anzanello and Flavio Sanson Fogliatto.“Learning curve models and applications: Literature review and research directions”. English. In: International Journal of Industrial Ergonomics41.5 (2011), pp. 573–583.DOI:10.1016/j.ergon.2011.05.001.

[2] W. Bai et al. “Recurrent Neural Networks for Aortic Image Sequence Segmentation with SparseAnnotations”. In: Medical Image Computing and Computer Assisted Intervention? MICCAI 2018.Vol. 11073. Lecture Notes in Computer Science. The final authenticated version is available online at https://doi.org/10.1007/978-3-030-00937-367. Cham: Springer, Sept. 2018, pp. 586–594.DOI:10.1007/978-3-030-00937-3_67.URL:https://openaccess.city.ac.uk/id/eprint/23149/.

[3] Wenjia Bai et al. “A population-based phenome-wide association study of cardiac and aortic structure and function”. In: Nature Medicine 26.10 (Oct. 2020),pp. 1654–1662.ISSN: 1546-170X.DOI:10.1038/s41591-020-1009-y.URL:https://doi.org/10.1038/s41591-020-1009-y.

[4] Wenjia Bai et al. “Automated cardiovascular magnetic resonance image analysis with fully convolutional networks”. In: Journal of Cardiovascular Magnetic Resonance20.1 (Sept. 2018), p. 65.ISSN: 1532-429X.DOI:10.1186/s12968-018-0471-x.URL:https://doi.org/10.1186/s12968-018-0471-x.

[5] Alejandro Barredo Arrieta et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”.In: Information Fusion58 (2020), pp. 82–115.ISSN:1566-2535. DOI: https://doi.org/10.1016/j . inffus . 2019 . 12 . 012.URL: https :/ / www . sciencedirect. com / science /article/pii/S1566253519308103.

[6] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In:Na-ture562.7726 (Oct. 2018), pp. 203–209.ISSN: 476-4687.DOI:10.1038/s41586-018-0579-z.URL:https://doi.org/10.1038/s41586-018-0579-z.

[7] Hao Li et al.Pruning Filters for Efficient ConvNets.2017. arXiv:1608.08710 [cs.CV].

[8] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: Medical Image Analy-sis42 (Dec. 2017), pp. 60–88.ISSN: 1361-8415. DOI:10 . 1016 / j . media . 2017 . 07 . 005. URL:http://dx.doi.org/10.1016/j.media.2017.07.005.

[9] Adnan Mustafa. “A Complete Probabilistic Model for the Quick Detection of Dissimilar Binary Images by Random Intensity Mapping”. In: WSEAS Transactions on Signal Processing13 (Oct. 2017), pp. 208–214.

[10] Adnan Mustafa. “A Modified Hamming DistanceMeasure for Quick Detection of Dissimilar BinaryImages”. In: Jan. 2015.DOI:10.1109/ICCVIA.2015.7351897.

[11] Steffen E. Petersen et al. “Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UKBiobank population cohort”. In: Journal of Cardio-vascular Magnetic Resonance19.1 (Feb. 2017), p. 18.ISSN: 1532-429X.DOI:10.1186/s12968-017-0327-9.URL:https://doi.org/10.1186/s12968-017-0327-9.

[12] H. Shin et al. “Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning”. In: IEEE Transactions on Medical Imaging35.5 (2016), pp. 1285–1298. DOI:10 . 1109 / TMI . 2016 .2528162.

[13] K. R. Siegersma et al. “Artificial intelligence in cardiovascular imaging: state of the art and implications for the imaging cardiologist”. In: Netherlands Heart Journal 27.9 (Sept. 2019), pp. 403–413.ISSN: 1876-6250.DOI:10.1007/s12471-019-01311-1.URL:https://doi.org/10.1007/s12471-019-01311-1.

[14] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, & Douglas B. Kell. (2017). What do we need to build explainable AI systems for the medical domain?

[15] Hendricks, L.A, Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating Visual Explanations, arXiv:1603.08507v1 [cs.CV] 28 Mar 2016.

[16] Si, Z. and Zhu, S. (2013). Learning AND-OR Templates for Object Recognition and Detection. IEEE Transactions On Pattern Analysis and Machine Intelligence. Vol. 35 No. 9, 2189-2205.

[17] Lake, B.H., Salakhutdinov, R., & Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. Science. VOL 350, 1332-1338.

[18] Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. CHI 2016 Workshop on Human Centered Machine Learning. (arXiv:1602.04938v1 [cs.LG] 16 Feb 2016).

[19] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, & Francisco Herrera. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.