Visualisation Research Center (VISUS) & Institute of Signal Processing and System Theory (ISS)

Masterarbeit

Deep Visualization for MR based Biological Age Estimation

Shashank Salian

Course of Study:	Computer Science
Examiner:	Prof. Dr.-Ing. Bin Yang, Prof. Dr. Daniel Weiskopf
Supervisor:	Karim Armanious, M.Sc., Dipl.-Inf. Tanja Munz, M.Sc., Sherif Abdullatif, M.Sc.
Commenced:	May 18, 2020
Completed:	November 18, 2020

Abstract

Deep Learning (DL) architectures proposed over recent years have been successfully applied to human data (such as medical data or facial data) to estimate age with reasonably good accuracy. However, the reasoning underlying the models' age predictions are still obscure. To address this challenge, Explainable AI (XAI) methods such as Deep Visualization (DV) have been applied to the deep learning models. Existing literature in the DV domain mainly considers image classification and object detection tasks involving two-dimensional input images. In this work, we show how DV can be applied to models handling high-dimensional medical data, and thereby identify a medical condition in a subject. The objective of this thesis is to perform a comparative study of the state-of-the-art DV techniques and adapt them for age estimation models. We use an iterative training framework to estimate the Biological Age (BA) from human brain MRI trained over a sizeable dataset across various age groups and Clinical Dementia Ratings (CDRs). We identify salient regions that affect aging by visualizing the gray matter volumes of the subjects. Further, we combine fine-grained and coarse-grained visualization techniques namely, Saliency Map Order Equivalent (SMOE) Scale saliency maps and GradCAM++ respectively to generate improved visualization maps. The differences in salient regions are observed across subjects with varying CDRs. The regions in the brain significantly affected by Alzheimer's Disease (AD) are identified based on the relative intensities in the visualization maps between healthy subjects and potential AD patients. In this work, we analyze how the application of Deep Visualization techniques bolster the theory of accelerated aging of the brain in the subjects having ailments, thereby assisting in timely diagnosis.

Keywords and phrases: Biological Age (BA), Class Activation Map (CAM), Clinical Dementia Rating (CDR), Deep Visualization (DV), Explainable Artificial Intelligence (XAI), Magnetic Resonance Imaging (MRI).

List of Figures

2.1 ConvLSTM architecture 17
2.2 BA estimation from retinal imaging 18
2.3 Overview of CAM, GradCAM, GradCAM++ 22
2.4 Noisy input 23
2.5 SMOE Scale saliency maps 24
2.6 SMOE Scale overview 25

3.1 Demographic distribution 27
3.2 CDR and Gender distribution 28
3.3 Preprocessing steps 30
3.4 AC-PC reorientation 31
3.5 Segmentation 32
3.6 Rigidly aligned segments 32
3.7 Initial DARTEL template 33
3.8 Final DARTEL template 34
3.9 Final preprocessed gray matter 34

4.1 Age-Net network architecture 38
4.2 Iterative data cleaning procedure 38
4.3 High-level view of the visualization procedure 40
4.4 Detailed visualization stages 42
4.5 Input chunk generation 43

5.1 Anatomy of the Limbic system 46
5.2 Healthy activated regions 47
5.3 Visualization maps comparison 48
5.4 Test-set visualization maps 49

A.1 Comparison of GradCAM++ with SmoothGradCAM++ 60

List of Tables

2.1 brain-EAD for GM. 19
2.2 brain-EAD for WM. 19

3.1 CDR counts. 29
3.2 Average age of subjects with dementia. 29

Acronyms

AC	Anterior Commissure
AD	Alzheimer's Disease
ADRC	Alzheimer Disease Research Center
AI	Artificial Intelligence
ANS	Autonomic Nervous System
BA	Biological Age
CA	Chronological Age
CAM	Class Activation Map
CDR	Clinical Dementia Rating
CNN	Convolutional Neural Networks
CSF	Cerebrospinal Fluid
COWC	Cars Overhead with Context
ConvLSTM	Convolutional Long Short Term Memory
DARTEL	Diffeomorphic Anatomical Registration through Exponential Lie Algebra
DL	Deep Learning
EAD	Effective Age Difference
FWHM	Full Width at Half Maximum
GAP	Global Average Pooling
GM	Gray Matter
KAR	Keep And Retrain
LRP	Layerwise Relevance Propagation

LSTM Long Short Term Memory

MAE Mean Absolute Error
ML Machine Learning
MNI Montreal Neurological Institute
MR/MRI Magnetic Resonance / Magnetic Resonance Imaging

NIFTI Neuroimaging Informatics Technology Initiative

OASIS Open Access Series of Imaging Studies

PC Posterior Commissure
PD Parkinson's Disease

ReLU Rectified Linear Unit
RF Radio Frequency
ROAR Remove And Retrain

SMOE Saliency Map Order Equivalent
SPM Statistical Parameter Mapping

TE Echo Time
TR Repetition Time

VQA Visual Question Answering

WM White Matter

XAI Explainable Artificial Intelligence

Contents

1 Introduction **13**
- 1.1 Motivation . . . 13
- 1.2 Contribution . . . 15

2 Related Work **17**
- 2.1 Recent attempts in Biological Age Estimation . . . 17
- 2.2 GradCAM . . . 20
- 2.3 GradCAM++ . . . 20
- 2.4 SmoothGradCAM++ . . . 23
- 2.5 Efficient Saliency Maps for Explainable AI . . . 24

3 Datasets and Data Preprocessing **27**
- 3.1 Datasets . . . 27
- 3.2 Preprocessing . . . 30

4 Methods & Experiments **37**
- 4.1 Biological Age Estimation Framework . . . 37
- 4.2 Visualization Experiments on BA Estimation Model . . . 39

5 Visualization Results and Significant regions correlated with Dementia **45**

6 Conclusion and Outlook **51**

Bibliography **53**

A Appendix **59**
- A.1 OASIS dataset metadata . . . 59
- A.2 Visual comparison: SMOE with GradCAM++ vs. SMOE with Smooth-GradCAM++ . . . 59
- A.3 Iterative framework configuration . . . 60

1 Introduction

1.1 Motivation

Age has been one of the key parameters while assessing the health of a person. The proof of age is also one of the essential information required in various legal, criminal, and administrative tasks globally. However, there are instances when the age of a person is unknown, for instance, displaced migrants, orphaned children, or loss of documents due to natural calamity or man-made disasters. In such cases, medical data are often the last resort, these are popularly referred to as *bio-markers*. Medical data such as skeletal imaging (dental. hand and clavicle bone) [ŠKPU17], retinal imaging [LWL+19], brain imaging [AAB+20], and bio-markers such as the calcium content in the arteries [MNB+09], or DNA methylation measure [MGLV20] have been used to estimate age.

The chronological age (CA) of a person is the time elapsed since their birth. Whereas, biological age (BA) is a complex concept to define and is not as obvious as computing the CA. There are multiple factors such as existing diseases, lifestyle, genetics, which affect how our organs age. Organ-based BA would also enable us to identify ailments related to a specific organ which leads to abnormal aging of that organ. The biological age of an organ may diverge from the chronological age of a subject due to the unhealthy condition of that organ. To solve this problem, medical imaging data is an ideal candidate as it enables us to focus on specific regions in the body. This is important for determining the BA of an organ, unlike data which generalizes the aging of the body as a whole.

The most commonly used techniques to obtain medical imaging data are using X-rays, CT scans, MRI scans. X-rays are commonly used for detecting bone fractures, dental regions, and infections such as pneumonia. They produce 2D images. CT scans also use X-rays, however, it can generate 3D images and can also be used to detect tumors or damage in soft tissues. MRI scan is generally used for scanning the softer tissues such as tendons, ligaments, spine, and internal organs such as the brain. MR scanners produce better quality images than CT scans or X-rays. Moreover, MRI scans are safer than X-rays and CT scans as MR scanners use strong magnets and radio-frequency pulses, unlike X-rays. X-rays are harmful to the softer tissues due to their ionizing nature and its prolonged exposure could be cancerous. Therefore MR scanners are preferred even though it takes longer time to scan compared to X-rays.

The brain is the most vital organ in our body as it controls all the voluntary and involuntary processes within our body. Therefore, medical imaging of the brain such as MRI has been used to identify various physical and psychological ailments in the body. These ailments could lead to accelerated aging of the brain due to alteration or even degeneration of brain tissues. Medical condition such as dementia leads to impairment of memory, psychological & physical balance, and cognition. People can suffer from one or more neurological conditions which lead to dementia such as Alzheimer's Disease, Lewy Body Dementia, Vascular Disease, Frontotemporal Dementia, Huntington's Disease, or mixed dementia. Alzheimer's Disease (AD) is one of the major causes of dementia. The severity of dementia is determined by the Clinical Dementia Rating (CDR) which is a 5 point rating namely: 0, 0.5, 1, 2, 3 where the value 0 is for cognitively normal subjects and 3.0 for severe Alzheimer's Disease. Such medical conditions lead to divergence in the chronological age from the biological age owing to the premature alterations in the brain anatomy due to the disease. Hence, the atypical aging characteristics identified when estimating BA from brain MRI can be a good indicator of a life-threatening disease such as dementia and therefore AD.

With significant progress in deep learning over the recent years, DL architectures are capable of extracting complex non-linear features within the medical data and thus solving problems which are tedious and error-prone using conventional methods. Estimating age from medical reports is one such problem that can be solved using DL. Unlike, machine learning (ML) approaches, DL does not require hand-crafted features and DL techniques such as CNNs can extract hierarchical information from the data. Recent work by the ISS institute proposed a novel DL architecture [AAB+20], [AAS+20] to estimate chronological age. It is further used in an iterative training based BA estimation framework to estimate the biological age from the brain MRI. However, the reasoning involved while making predictions using such sophisticated deep learning models should be justifiable. Gaining trust is very important for AI applications especially in domains such as medical, defense, finance, autonomous driving, aerospace, or any field which directly involves human safety or economic stability. The AI system should be responsible for making fair and transparent decisions. In simple words [BDD+20] aptly defines Explainable AI as follows: "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make it's functioning clear or easy to understand."

[BDD+20] further highlighted the goals of Explainable AI which are as follows:

- Trustworthiness
- Causality
- Transferability
- Informativeness
- Confidence
- Fairness

- Accessibility
- Interactivity
- Privacy Awareness

With the relatively modern AI techniques such as Deep Learning, it has become increasingly difficult to interpret the models' actions, and are often regarded as 'black-box models'. This is primarily due to an inherent non-linearity of these models. Hence, Explainable AI (XAI) for DL is becoming obligatory. DL models involving Convolutional Neural Networks(CNNs) have tried to mimic human visual cognition. There exist various types of XAI techniques [BDD+20] such as *model simplification*, *feature relevance estimators*, *text explanations*, *local explanation*, *model visualization* that are applied to deep learning models. *Model Visualization* techniques are commonly applied for the CNN based models to provide visual explanations. Over the years various *model visualization* techniques have been proposed for visual tasks such as image classification, object detection, image captioning, VQA (Visual Question Answering), image segmentation to name a few. Some of the widely used techniques (generally an overlap between *feature relevance estimators* and *model visualization*) among these include Guided Backpropagation [SDBR15], Deconvolution [ZF13], Integrated Gradients [STY17], Layerwise Relevance Propagation (LRP) [BBM+15], Class Activation Maps (CAM) [ZKL+16], GradCAM [SCD+19], GradCAM++ [CSHB18]. They are also sometimes collectively referred to as *Deep Visualization techniques*. Deep Visualization (DV) techniques have evolved significantly over recent years and are still evolving to provide more accurate visual explanations while reducing the computational and memory requirements.

DV can be applied to the age estimation models to identify significant regions in the input which influenced the model's decision while predicting age. The corresponding alterations in the internal anatomy of the brain can be visualized using deep visualization methods. However, existing deep visualization methods have been commonly demonstrated on classification tasks or object detection tasks with 2-Dimensional image data and seldom with higher dimensional medical imaging data. We need fine-grained visualizations that impart greater confidence when applied to images with intricate regions such as the human brain. To the best of our knowledge, these visualization methods have not been applied to organ-based age estimation models. In this thesis, a comparative study of the current state-of-the-art deep visualization techniques have been performed and have been applied to the BA estimation framework.

1.2 Contribution

In this thesis we perform a detailed study of the state-of-the-art DV techniques namely: GradCAM [SCD+19], GradCAM++ [CSHB18], Saliency Map Order Equivalent (SMOE) Scale saliency maps [MCF20] and apply these techniques to the medical imaging data used by the BA estimation framework, i.e. brain MRI in our case. We further combine two

visualization techniques to generate better visualization maps by leveraging the strengths of multiple techniques. We combined SMOE Scale saliency maps with GradCAM++ maps to visualize the activated regions in the gray matter of the subject as inferred by the model. This thesis needed to adapt these visualization techniques and re-implement to apply it to 3D medical imaging data: Open Access Series of Imaging Studies (OASIS) [MWP+07], [MFC+10], [LBM+19] data-set for the BA estimation model.

Before the implementation and application of the above visualization procedure, we perform a multi-step preprocessing procedure (refer to Section 3.2). The preprocessing procedure is applied to a cohort dataset with more than 2000 MR scans from more than 1200 subjects across ages and different stages of dementia which included subjects from OASIS-1 & OASIS-3 datasets [MWP+07], [LBM+19].

Moreover, we extend the above visualization techniques to produce visualizations for all three planes of the brain namely: axial, coronal, sagittal, and not just the axial plane. This enables us to visualize highly activated regions across all three planes for healthy subjects and subjects with dementia. Additionally, this visualization procedure was applied for multiple voxel chunks across each of these planes. This is done since different chunks aided in identifying different regions in the gray matter that may otherwise have been occluded or absent in other chunks.

We apply the above visualization techniques to both non-outliers and outliers scans as identified by the iterative training procedure using the BA estimation framework. We then analyze the regions in the gray matter along with its network activation intensities in the visualization maps. These regions were analyzed for atrophy (indicated by the decreased intensity in the map) and its corresponding biological regions were identified using a brain atlas [MPV09]. Subsequently, we discuss how the identified regions correspond to the areas affected by dementia, specifically Alzheimer's Disease.

2 Related Work

In this section, we discuss some of the recent attempts in biological age estimation and their strengths and shortcomings. We further discuss in detail the DV literature for some of the state-of-the-art deep visualization approaches.

2.1 Recent attempts in Biological Age Estimation

Over recent years, there have been significant attempts for estimating BA from non-imaging as well as imaging data. A recent approach to estimate BA using lifestyle factors such as physical activity is proposed by [RA19]. The authors use physical activity data mea-

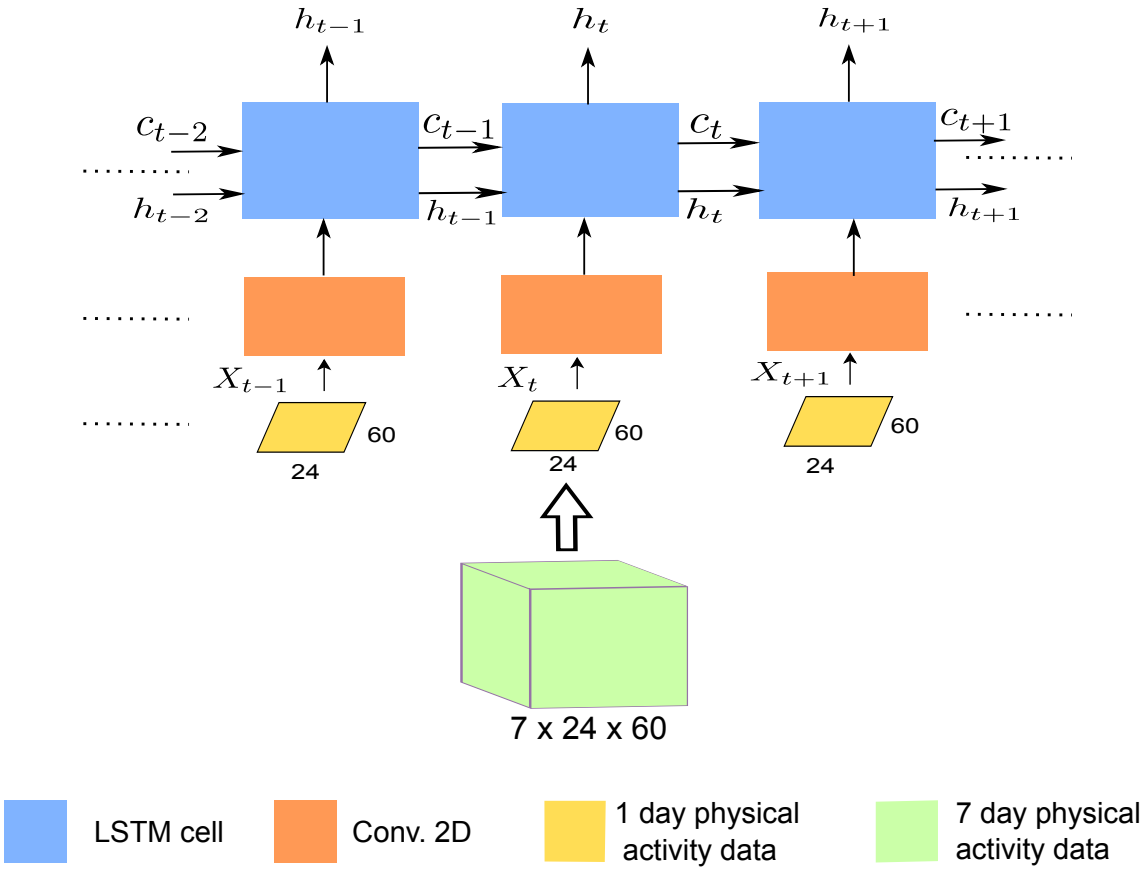


Figure 2.1: Structure of ConvLSTM proposed by [RA19].

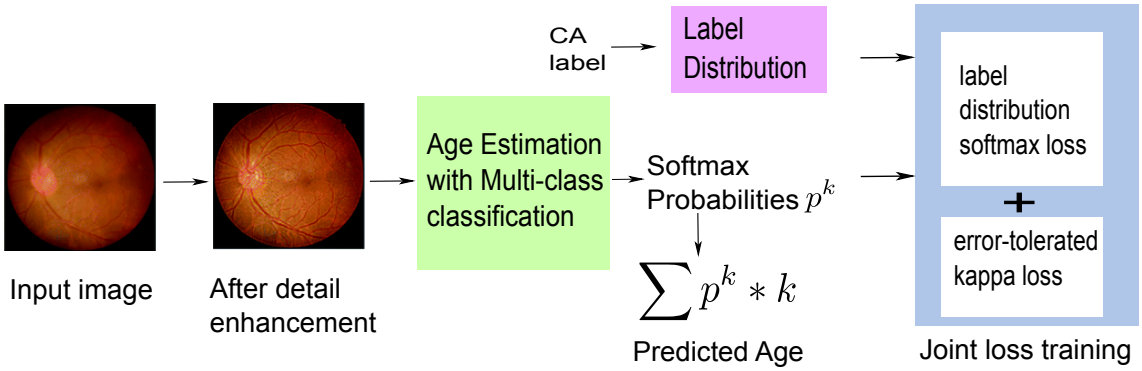


Figure 2.2: High-level overview of the BA estimation from retinal imaging. [LWL+19]. In this figure, k indicates k^{th} label type as it is modeled as a multi-class classification problem.

sured by accelerometers and use a model based on convolutional long short term memory (ConvLSTM) proposed by [SCW+15]. The structure of ConvLSTM used in this work is depicted in Figure 2.1. [RA19] adapted this architecture for solving a regression task to estimate BA instead of predicting the next activity sequence. The rationale underlying the use of LSTM architecture is due to the time-series nature of the data of size 7 x 24 x 60 (days x hours x minutes). The use of 2D convolutions is to identify local patterns from the activity data. The time duration of measurement of data was relatively short (i.e., 1 week) which is inadequate to measure long-term effects of lifestyle on aging. As far as we know, the sensor used had limited water resistance, therefore physical activities such as swimming or taking a shower may not have been recorded. This method of BA estimation may not give us the true biological age as it uses chronological age as labels to train the BA estimation model. Moreover, this approach would give us a BA estimate for the whole body and it is not organ-specific.

One of the recent organ-based approaches is using retinal imaging as depicted in Figure 2.2. [LWL+19] proposed BA estimation using retinal imaging on the premise the authors' state as follows: "The retina is considered to share anatomical and physiological similarities with the brain, and rich information related to aging can be visualized non-invasively from retinal imaging." Interestingly, [LWL+19] propose age estimation as a multi-class classification problem based on the VGG-19 network [SZ15]. Important thing to note is that the label, i.e. CA is not one-hot encoded. Instead, the label is represented as a gaussian distribution centered at the CA, and the loss computed as a result is referred to as label distribution softmax loss L_s. An additional loss function is computed to account for error tolerance of the model, this is referred to as error-tolerated weighted kappa loss L_κ which is explained in detail in [LWL+19]. The authors combined these loss functions to propose a joint loss function as shown in Equation (2.1) [LWL+19] and Figure 2.2.

$$L_{joint-loss}(X) = \lambda L_s(X) + (1 - \lambda)L_\kappa(X) \tag{2.1}$$

Disease	brain-EAD
AD	+9.29 ± 6.43
PD	+1.50 ± 6.03

Table 2.1: brain-EAD for GM.

Disease	brain-EAD
AD	+8.85 ± 6.62
PD	+2.47 ± 5.85

Table 2.2: brain-EAD for WM.

The authors used a weighting factor of $\lambda = 0.9$. The argument as described in [LWL+19], underlying the use of this joint loss function is to improve model robustness and account for the time-continuous nature of human aging. The proposed network was trained only on the healthy subjects' data to avoid learning atypical aging effects while estimating BA. The model was evaluated on both healthy and unhealthy subjects (for instance, diabetic patients). The model achieved an MAE of 3.39 years for healthy subjects and 4.21 years for unhealthy subjects. The limitation of the proposed technique is that the authors consider clinical health as the basis for using CA as labels for the BA estimation.

[BMS+20] proposed brain age estimation for patients suffering from AD and Parkinson's Disease (PD) using T1-weighted MRI scans. Two independent support vector regression (SVR) models with linear kernels are trained for gray matter (GM) and white matter (WM) respectively. The models are trained on preprocessed GM and WM voxel intensities to estimate the brain age. The preprocessing procedure is similar to the procedure described in Section 3.2. The model was validated by computing the MAE, Root Mean Squared Error (RMSE), and the correlation coefficient between the estimated brain age and chronological age. Additionally, the authors measured the brain effective age difference (brain-EAD) which is computed below:

$$EAD = Estimated\ Brain\ Age - Chronological\ Age \tag{2.2}$$

The authors observed the mean brain-EAD for AD subjects to be significantly higher than subjects with PD for both GM and WM models as listed in Tables 2.1 and 2.2. This showed that AD caused accelerated brain aging compared to PD. Therefore, organ-based age estimation of the brain can be a good indicator for detecting atypical brain aging in AD. It is important to note here, that chronological age was used as age labels, hence this method may not give a true estimate of the brain age.

[ŠPU19] proposed to estimate biological age using hand bone MRI. The authors trained two different models, namely a random forest regressor with hand-crafted features and a deep CNN. However, this work relied on the physical maturity of the bones and is aimed towards adolescents and young adults. Therefore, it cannot be applied to a wider demographic. Other approaches such as [ŠKPU17] relied on medical experts, (e.g. radiologists, dentists) to assign biological age labels, which may be inaccurate and subjective.

As far as we know there is no reliable way to obtain biological age labels. Therefore, biological age estimation remains an open problem.

2.2 GradCAM

Gradient Weighted Class Activated Map (GradCAM) is a state-of-the-art Class Activation Map-based deep visualization technique proposed by Selvaraju et al. [SCD+19]. This method can be used for any CNN-based model architectures, unlike CAM (Class Activation Map) [ZKL+16] which requires fully-convoluted architectures with a global average pooling (GAP) as the penultimate layer in the network. Therefore, to apply CAM we need to alter the network architecture and this hence involves retraining. CAM can be considered as a special case of GradCAM for fully convoluted networks. GradCAM has had applications in image classification, tasks, image captioning, and VQA tasks.

To compute a GradCAM localization map for a particular class, first, the gradient of the class score y^c (i.e. before applying softmax activation) is obtained with respect to the activation feature map A^k of a convolution layer (generally the last convolution layer). As we will discuss in Section 4.1 that we are using a regression network, hence, we do not have any softmax layer to be considered. y^c is not restricted to be only class scores, it can be any differentiable activation. The gradients are summed and averaged across all dimensions i.e. global average pooled to obtain w_k^c where Z is the number of pixels in the feature map. This can be mathematically expressed as follows [SCD+19]:

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2.3}$$

w_k^c act as weight coefficients and it signifies the feature importance of k^{th} activation map for class output c. Subsequently, a weighted summation is performed using these weights and activation feature map A^k. A ReLU operation is performed on this weighted sum to obtain the resulting GradCAM map as follows [SCD+19]:

$$L_{GradCAM}^c = ReLU\left(\sum_k w_k^c A^k\right) \tag{2.4}$$

The ReLU operation ensures that non-class regions are suppressed from the visualization map. The map needs to be up-sampled to match in size with the original input image. Though GradCAM is very popular, however, the visualizations are coarse and as of now, it is a fairly old technique. It has been evolved over the years to propose better visualization techniques such as GradCAM++ and SmoothGradCAM++ which will be discussed in the following subsections.

2.3 GradCAM++

GradCAM++ proposed by Chattopadhyay et al. [CSHB18] is an improvement over the GradCAM technique. It generates better-localized maps and produces improved visualization for multiple occurrences of an object in the same image. These multiple occurrences

have different spatial footprints owing to their differences in position, orientations, or views. Hence, different spatial footprints would activate different feature maps and it is important to utilize the majority of them in the final map. Therefore, the authors proposed pixel-wise weighted gradients to compute the weights w_k^c as seen in the following equations [CSHB18]:

$$y^c = \sum_k w_k^c \sum_i \sum_j A_{ij}^k \tag{2.5}$$

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot ReLU\left(\frac{\partial y^c}{\partial A_{ij}^k}\right) \tag{2.6}$$

Here w_k^c are the weighting coefficients indicating the feature importance of feature map k for individual class c, whereas α_{ij}^{kc} are the pixel-wise weighting coefficients for the gradients. The authors proposed using only positive gradients as indicated by the ReLU operation in the above equation. According to the authors, this ensures a higher preference for visual features which increases the output neuron's activation. The weighting coefficients α_{ij}^{kc} are derived by computing higher order partial derivatives derivatives of y^c w.r.t. A_{ij}^k using Equations (2.5) and (2.6) to get the following expression in Equation (2.7). This is followed by the GradCAM++ localization map as expressed in Equation (2.8) [CSHB18]:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 y^c}{(\partial A_{ij}^k)^2}}{2\left(\frac{\partial^2 y^c}{(\partial A_{ij}^k)^2}\right) + \sum_a \sum_b A_{ab}^k \left(\frac{\partial^3 y^c}{(\partial A_{ij}^k)^3}\right)} \tag{2.7}$$

$$L_{ij\,GradCAM++}^c = ReLU\left(\sum_k w_k^c A_{ij}^k\right) \tag{2.8}$$

Interesting observation to note is that the GradCAM map can be deduced from the GradCAM++ map by replacing α_{ij}^{kc} in Equation (2.6) with $\frac{1}{Z}$ from Equation (2.3). In GradCAM++, the computation of the higher-order derivatives of the gradients is further simplified. An exponential function is applied to y^c, making it infinitely differentiable and it requires the corresponding convolution layer to have a linear or ReLU activation. Applying the exponential function converts the higher-order partial derivatives to multiplicative powers as shown below [CSHB18]:

$$S^c = exp(y^c) \tag{2.9}$$

$$\frac{\partial S^c}{\partial A_{ij}^k} = exp(y^c)\frac{\partial y^c}{\partial A_{ij}^k} \tag{2.10}$$

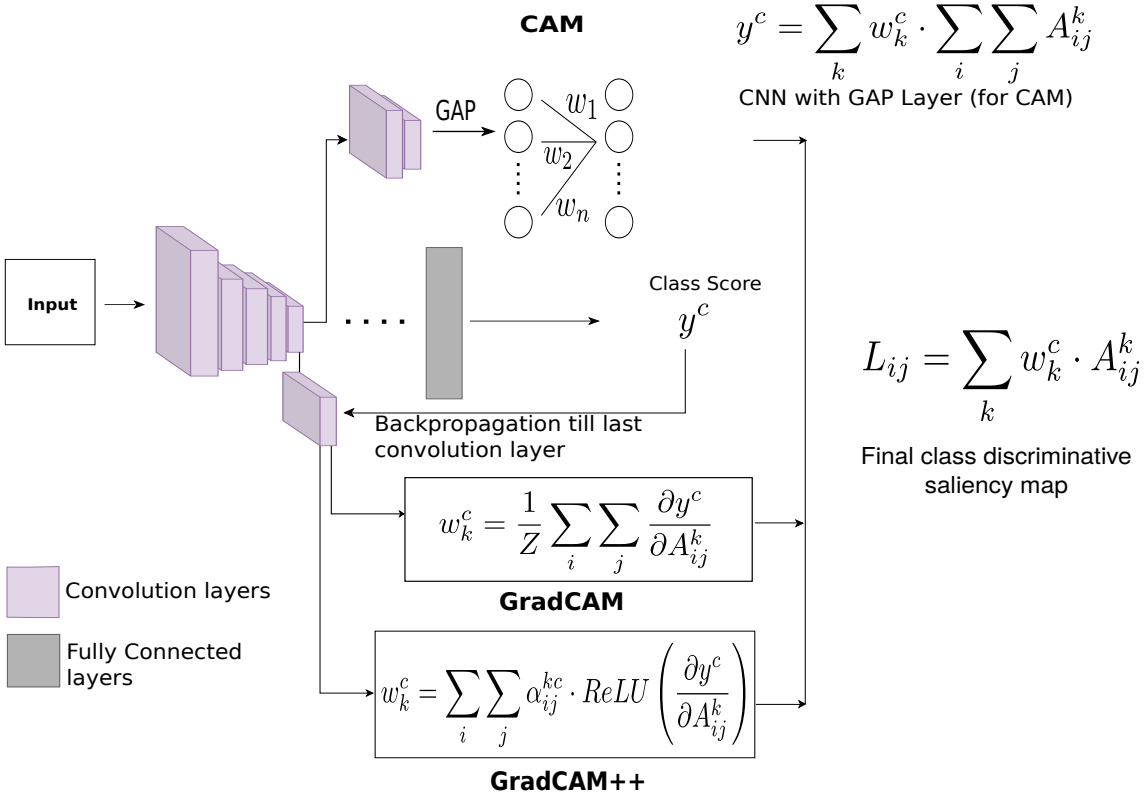


Figure 2.3: Overview of CAM [ZKL+16], GradCAM [SCD+19], GradCAM++ [CSHB18].

$$\frac{\partial^2 S^c}{(\partial A_{ij}^k)^2} = exp(y^c)\left(\frac{\partial y^c}{\partial A_{ij}^k}\right)^2 \tag{2.11}$$

The reasoning underlying the use of ReLU or any linear activation is to eliminate second-order or higher-order derivatives of y^c [CSHB18]. As a result, in Equation (2.11) we have omitted the $\frac{\partial^2 y^c}{(\partial A_{ij}^k)^2}$ term. Similarly, we compute the third derivative and apply all three derivatives in Equation (2.7) to get:

$$\alpha_{ij}^{kc} = \frac{\left(\frac{\partial S^c}{\partial A_{ij}^k}\right)^2}{2\left(\frac{\partial S^c}{\partial A_{ij}^k}\right)^2 + \sum_a \sum_b A_{ab}^k \left(\frac{\partial S^c}{\partial A_{ij}^k}\right)^3} \tag{2.12}$$

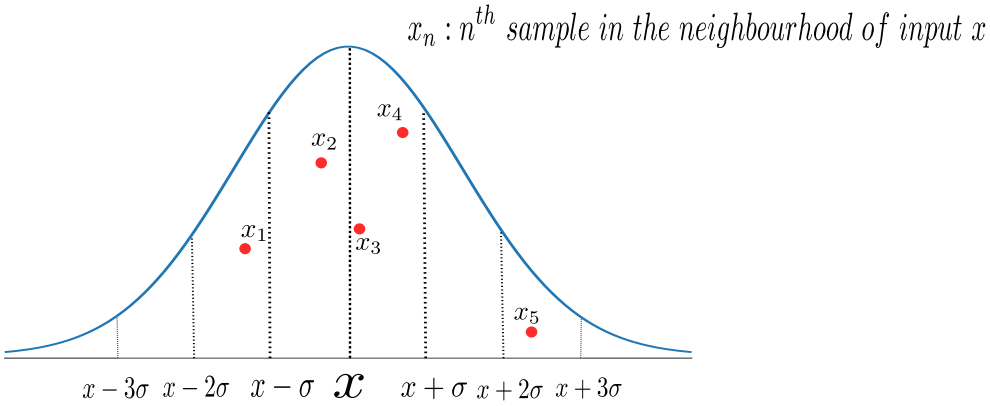


Figure 2.4: Normal distributed noised samples from the neighbourhood of input x for $N = 5$.

Figure 2.3 depicts a structural overview of CAM, GradCAM, GradCAM++. The authors [CSHB18] extended their visualization to apply to spatio-temporal tasks such as to videos and 3D action recognition. The maps obtained by GradCAM, GradCAM++ can be multiplied element-wise with visualizations obtained from Guided Backpropagation [SDBR15] or Deconvolution [ZF13] to generate fine-grained visualizations.

2.4 SmoothGradCAM++

SmoothGradCAM++ proposed by Omeiza et al. [OSCW19] is an inference based visualization technique obtained by combining SmoothGrad [STK+17] and GradCAM++ [CSHB18]. It aims to generate sharper and better-localized visualizations. The only difference to GradCAM++ is that here for every input, a fixed number of samples are generated from a normal distribution. These samples lie in the neighborhood of the input with some added noise (standard deviation). This is in effect taking N random samples from a normal distribution centered at input x with some standard deviation σ^2 as noise, $x + \mathcal{N}(0, \sigma^2)$.

$$\frac{\partial y^c}{\partial A_{ij}^k} = \frac{1}{N}\sum_{n=1}^{N} D_n^k \tag{2.13}$$

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} ReLU\left(\frac{1}{N}\sum_{n=1}^{N} D_n^k\right) \tag{2.14}$$

In Equation (2.13) D_n^k is the partial derivative of y_n^c w.r.t A_{ij}^k. y_n^c is obtained from the forward pass by using the n^{th} sample in the neighbourhood of input x as depicted in Figure 2.4. The subsequent steps are the same as when computing the GradCAM++ maps.

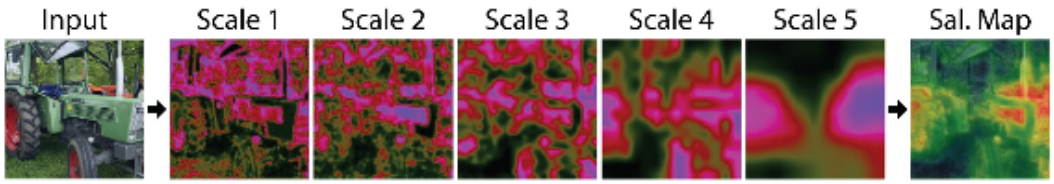


Figure 2.5: Example of Saliency maps visualization at various scales and the combined Saliency Map.[MCF20].

SmoothGradCAM++ involves greater time overhead compared to GradCAM++ because for every input x we compute the gradients for N samples. Therefore, as the number of samples increases, the computation time increases. Additionally, the choice of the number of samples N, and the standard deviation σ (noise) varies across its applications.

2.5 Efficient Saliency Maps for Explainable AI

CAM based methods such as CAM, GradCAM, GradCAM++ generally use Guided Backpropagation to add fine-grained details to the saliency maps. However, using Guided Backpropagation requires computing a full backward pass through the entire network which adds significant computational and memory overheads. While the CAM based methods require backward pass over just a few layers at the end of the network. Mundhenk et. al [MCF20] proposed an efficient method to compute the saliency maps which can also be used with CAM based methods to generate fine-grained and well-localized saliency maps.

This efficient saliency map method is aimed to be relatively efficient, provide reasonable information localization, and generate relatively sparse patterns with high peaks. The saliency maps are computed at the end of each scale group in the network, i.e. computing a saliency map for each scale output. E.g., if the input size is 224 x 224 then various scale groups (after convolution followed by downsampling, through the network) would be of sizes 112 x 112, 56 x 56, 28 x 28, 14 x 14, 7 x 7 respectively as depicted in Figures 2.5 and 2.6. These intermediate maps can be computed using the forward pass at inference. Hence, it is more efficient. The individual maps are computed for each scale output $T \in \mathbb{R}^{+,h\,x\,w\,x\,r}$ obtained after batch normalization and activation such as ReLU. To compute the information localization, a function $\psi(.)$ is applied on the output activation T. The ReLU activation ensures that there is no negative input for the log computation in Equation (2.15). The result is a 2D saliency map $S \in \mathbb{R}^{+,h\,x\,w}$ with dimensions h x w. This is mathematically formulated as follows [MCF20]:

$$\psi(x) = \frac{1}{r}\sum_{k=1}^{r} x_k \cdot \left[\log_2\left(\frac{1}{r}\sum_{k=1}^{r} x_k\right) - \frac{1}{r}\sum_{k=1}^{r}\log_2\left(x_k\right)\right] \tag{2.15}$$

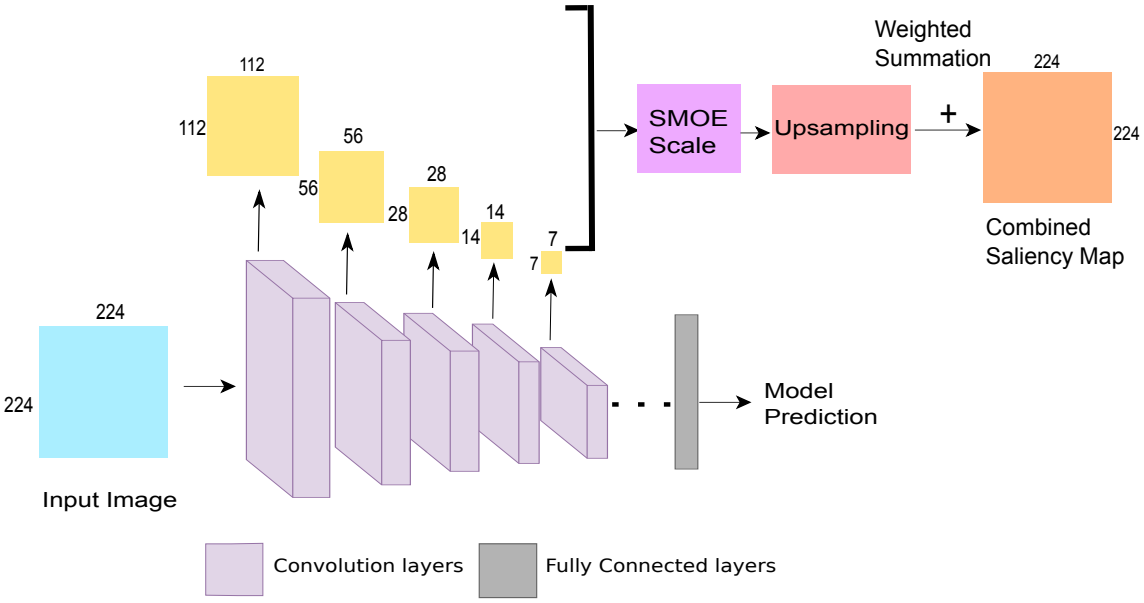


Figure 2.6: Application of SMOE saliency maps in a CNN model [MCF20]. Here we depict a traditional CNN based network to show various convolution blocks at which the saliency maps are computed and are then upsampled to the original input image size. Then these maps are weight-summed to generate a combined saliency map.

(2.16) $S = \psi(T)$

In Equation (2.15) k denotes the k^{th} channel, whereas r denotes the total number of channels. Further, we substitute $\frac{1}{r}\sum_{k=1}^{r} x_k$ as the mean μ and simplify it as follows [MCF20]:

(2.17) $$\psi(x) = \frac{1}{r}\sum_{k=1}^{r} \mu \log_2\left(\frac{\mu}{x_k}\right)$$

The maps obtained by this simplification are called Saliency Map Order Equivalent Scale maps. Equation (2.17) looks similar to entropy, however, it is not the same as there are no probabilities involved. It is analogous to information gain in the sense that it is the mean activation times the information we would gain from the mean activation given the individual activations [MCF20]. SMOE maps preserve equivalence for two maps when e.g. an independent monotonic function is applied to the saliency maps.

Once the maps are computed for each scale, a normal cumulative distribution function $\gamma(s;\mu,\sigma)$ is used to normalize the maps from 0 to 1. Here s is the individual saliency map and μ & σ are the mean and standard deviation respectively for these individual saliency maps. As these maps are at different scales they have different dimensions, hence, they are up-sampled to the size of the input image. Subsequently, these maps are weighted and

combined to compute a single combined saliency map as depicted in Figures 2.5 and 2.6 and as seen in Equation (2.18) below [MCF20]:

$$L_{i,jSMOE-Scale} = \frac{\sum_{k=1}^{r} \gamma(s_{i,j,k}; \mu_k, \sigma_k) \cdot w_k}{\sum_{k=1}^{r} w_k} \tag{2.18}$$

The weighting coefficients w_k can be chosen as progressively increasing (e.g., [1, 2, 3, 4, 5]) as the saliency maps computed later in the network has less irrelevant information. The authors of [MCF20] suggested using equal weights (i.e., [1, 1, 1, 1, 1]) when combining the saliency maps with other visualization techniques such as GradCAM, GradCAM++.

To evaluate the quality of the saliency maps, it is important that the maps have higher activations for more significant regions and vice-versa. Therefore, the authors used Keep and Retrain (KAR) and Remove and Retrain (ROAR) [HEKK19] accuracy scores. A good visualization technique will have higher KAR accuracy and lower ROAR accuracy. For the KAR accuracy score, the most significant 10%, 30%, 50%, 70%, 90% of the input pixels were kept and the remaining pixels were masked. For the ROAR the most significant 10%, 30%, 50%, 70%, 90% of the input pixels were removed. To avoid artifacts caused by masking, the authors retrained a ResNet-50 [HZRS15] model from scratch using the masked images.

[MCF20] evaluated the SMOE Scale statistics (Equations (2.15) and (2.17)) against other statistical measures to compute the maps such as standard deviation, truncated normal entropy, Shannon entropy, normal mean to name a few. All the statistical measures were evaluated on multiple data-sets namely: Image Net [DDS+09], CSAIL Places [ZLX+14], COWC (Cars Overhead with Context) [MKSB16]. The proposed SMOE scale statistic performed better overall across the datasets compared to the other statistical methods mentioned above.

The authors also evaluated and compared with other DV methods such as Gradient Heatmaps, Guided Backpropagation (GB), and Integrated Gradients. A ResNet-50 [HZRS15] model was trained on Image Net [DDS+09] and the KAR and ROAR accuracy scores were computed. SMOE Scale achieved higher KAR scores than other techniques. Whereas, Guided Backpropagation and SMOE Scale had similar ROAR scores, nonetheless, SMOE Scale is computationally much faster than GB due to the reasons mentioned earlier in this section. Further, [MCF20] combined SMOE Scale with CAM based method such as GradCAM++. This hybrid visualization approach produced the best results in comparison to all the former methods and with reasonable speed and modest resource needs.

3 Datasets and Data Preprocessing

Here we introduce the datasets and perform brief analysis on it. Subsequently, we discuss the various data preprocessing steps in detail.

3.1 Datasets

We intended to use a publicly available Neuroimaging dataset for our task. Hence, we used the Open Access Series of Imaging Studies (OASIS) dataset [MWP+07], [MFC+10], [LBM+19]. OASIS has three kinds of Neuroimaging datasets primarily for Normal aging and Alzheimer's Disease:

OASIS-1: Age 18-96 containing all right-handed subjects, OASIS-2: Age 60-96, OASIS-3: Age 42-97 containing right-handed, left-handed, and ambidextrous subjects.

The dataset uses the Neuroimaging Informatics Technology Initiative (NIFTI) file format. The NIFTI file format uses either a single or dual file storage format. The single format stores the data in (.nii) format containing the header and images. Whereas, the dual format stores data in two separate files as header (.hdr) and image (.img) respectively. The images from

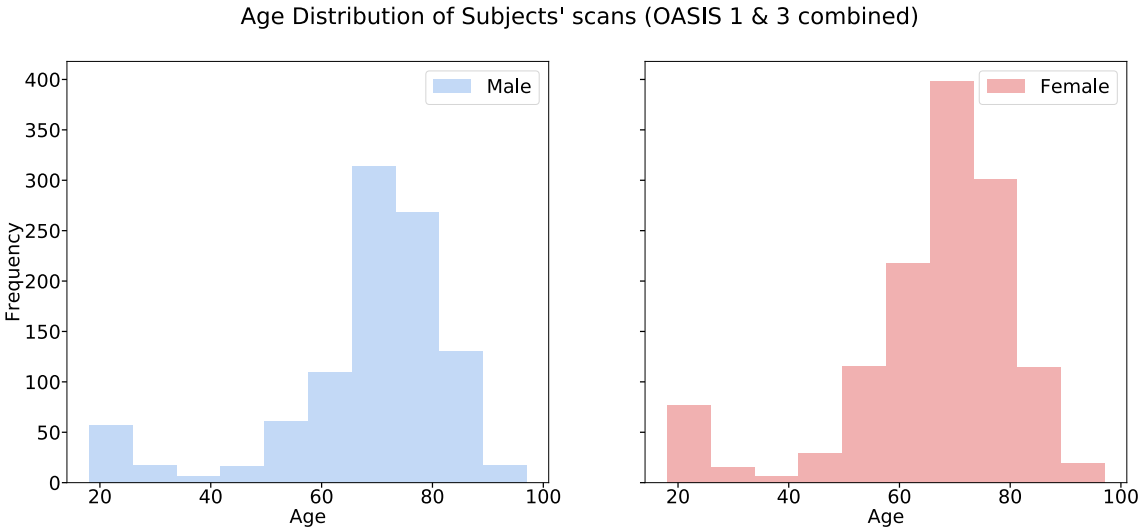


Figure 3.1: Subject demographics of the dataset. Each histogram represents the age distribution of the scans for (a) Male subjects and (b) Female subjects respectively.

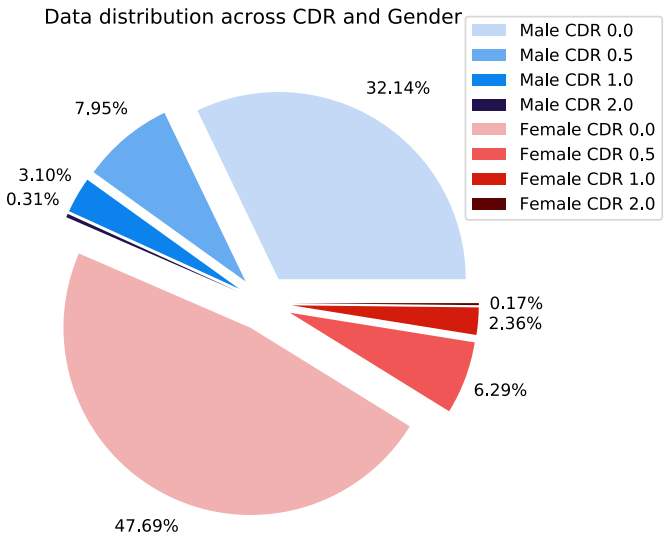


Figure 3.2: Distribution of subjects according to CDR and for each gender. The dataset as is did not contain any subjects with severe dementia, i.e. CDR 3 and had a very small proportion of subjects with moderate dementia, i.e. CDR 2.

OASIS 1 were converted from (.img) to (.nii) format by loading and saving them in (.nii) using neuroimaging libraries like nibabel [BMH+20]. Therefore we had all images in (.nii) format before advancing into the data preprocessing steps. To perform a detailed analysis of the dataset additional metadata such as Age, Handedness, CDR (Clinical Dementia Rating), and Gender were procured. This metadata for each subject was obtained in .csv format from the OASIS data browser. We have dropped the records which had missing age information.

After observing the subject demographics across all three datasets, we decided to merge OASIS-1 and OASIS-3 as it will cover a broad age group, i.e. from 18 to 97 years. However, we excluded OASIS-1 subjects > 42 years, to avoid any possible subject overlaps with subjects from OASIS-3 as the subjects were anonymized. The age distribution of the subjects categorized by gender is shown in Figure 3.1. In all we have 2290 scans, we only used T1 weighted scans. We will discuss T1 weighted scans later in this section. An important point to note is that a subject can have multiple scans at different ages, e.g. at 65 years, at 69 years, and at 72 years wherein the subjects' CDR may improve or in most cases deteriorate with age. The CDR values range as follows [0, 0.5, 1.0, 2.0, 3.0], here CDR = 0 implies healthy subjects whereas higher values indicate dementia with increasing severity. The pie chart in Figure 3.2 represents the CDR distribution among male and female subjects in the dataset. However, OASIS-3 MR Sessions data did not have associated CDR values and these had to be obtained from the OASIS-3 ADRC Clinical Data (collected by Alzheimer Disease Research Center Washington University St. Louis). Moreover, this data had a varied number of scans when compared with MR Sessions per subject. Hence,

CDR	Count
0.0	1828
0.5	326
1.0	125
2.0	11
3.0	0

Table 3.1: CDR counts.

Sub-Category	Average Age
Male CDR 0.5	76.602
Male CDR 1.0	76.59
Male CDR 2.0	76.544
Female CDR 0.5	74.942
Female CDR 1.0	75.472
Female CDR 2.0	69.5

Table 3.2: Average age of subjects with dementia.

from ADRC Clinical Data, the CDR values which correspond to the age closest to the age in MR Sessions data for every subject were taken. The total size of these 2290 scans was around 50 GB (uncompressed) wherein OASIS-1 accounted for 179 scans and OASIS-3 accounted for 2111 scans. In total there were 996 males and 1294 female subjects.

The distribution of subjects based on CDR values after preprocessing is listed in Table 3.1. The values listed in Table 3.2 indicate that the severity of dementia does not necessarily increase with age. Instead, this just indicates that a certain age-range is maybe more probable to suffering from dementia.

What are T1 weighted scans?
During an MRI Session, when a subject enters the scanner the protons inside the body align with the axis of the strong magnetic field produced inside the scanner. Then an external Radio Frequency (RF) energy signal is applied to the designated region. This excites the protons and is aligned at an angle to the magnetic field. Additionally, it causes the protons to spin in phase thereby causing resonance. After a brief amount of time, the external radio frequency signal is removed. The protons realign to the magnetic field of the MR scanner and lose their spin resonance. As a result, the emitted signal is measured by the scanner. The two types of emitted signals due to the realignment and dephasing are namely: T1 signal and T2 signal [Llo17]. T1 signal strength is directly proportional to the speed of the realignment of the protons. Whereas, T2 signal strength is inversely proportional to how quickly the proton spins go out of phase. The T1 and T2 signal strengths are exploited accordingly to produce T1 weighted or T2 weighted scans. The variations in T1 and T2 signal strengths are affected by two important properties i.e. Repetition Time (TR) and Time to Echo (TE). Here, TR is the time interval between two successive RF pulses applied in a sequence and TE is the time interval between RF pulse applied and receipt of its echo. T1 weighted scans have shorter TR and TE duration. The values of these parameters used for obtaining the scans can be referred to in Appendix A.1.

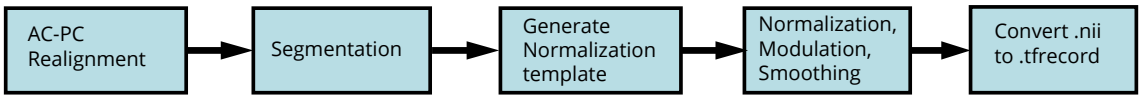


Figure 3.3: OASIS dataset [MWP+07], [MFC+10], [LBM+19] Preprocessing Steps.

3.2 Preprocessing

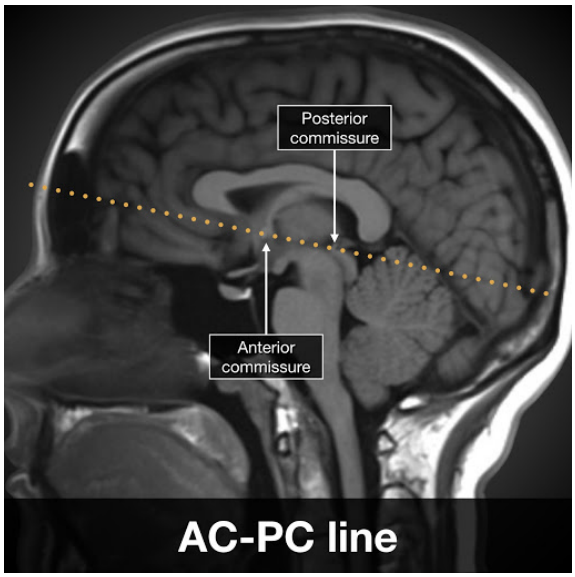
The preprocessing involved multiple stages as depicted in Figure 3.3 namely:

- Decompressing and Converting all files to .nii formats. (Python)
- AC-PC realignment using the statistical parameter mapping 12 [ABC+14], [ABC+20] (SPM 12) package. (R/MATLAB).
- Segmentation of samples into Gray matter, White matter, CSF (MATLAB). We have only used Gray Matter segments in this thesis.
- Diffeomorphic Anatomical Registration through Exponential Lie Algebra (DARTEL) [Ash07] template generation along with flow field generation. (MATLAB)
- Spatial normalization, modulation, and smoothing: We used the template generated in the previous step along with the flow field generated for each gray matter segment to normalize them. (MATLAB)
- Conversion of the preprocessed images to a serialized format here we used TFRecord format. (Python)

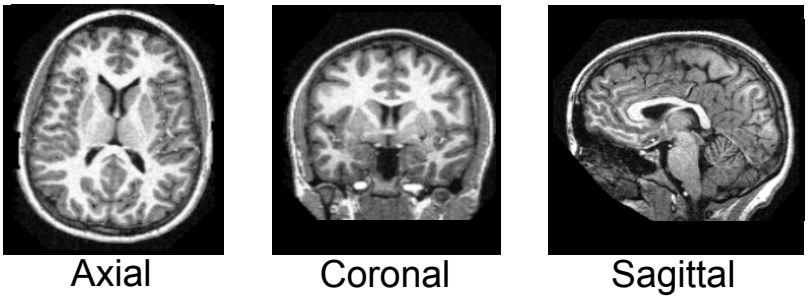
Software environment used: Python 3.6.x, R 3.5.3, MATLAB R2020a with SPM 12 package [ABC+20]. Using R for AC-PC realignment is optional, it was used because there is a simple to use R language wrapper for acpc_reorient based on MATLAB SPM 12, using the SPM12r package [Mus19].

AC-PC realignment
Anterior Commissure - Posterior Commissure (AC, PC) are two reference points inside the brain. The AC lies in between the line joining the left and right temporal lobe, whereas the PC lies in between the third and fourth ventricles in the posterior part of the brain just above the cerebral aqueduct. The AC-PC line acts as a reference plane to which all images are aligned so that the data is standardized as depicted in Figure 3.4. AC-PC realignment could be performed using SPM 12 package in MATLAB [ABC+20] or using SPM 12 package in R [Mus19].

Segmentation into gray matter, white matter and cerebrospinal fluid (CSF) segments
After reorienting the images, image segmentation (using SPM 12 segmentation module [ABC+14]) was performed on the realigned .nii images. As a result, the gray matter (c1*.nii), white matter (c2*.nii), and CSF (c3*.nii) segments were generated along with their respective rigidly aligned segments namely rc1*.nii, rc2*.nii, rc3*.nii files respectively.



(a) AC-PC reference line [Gal19].



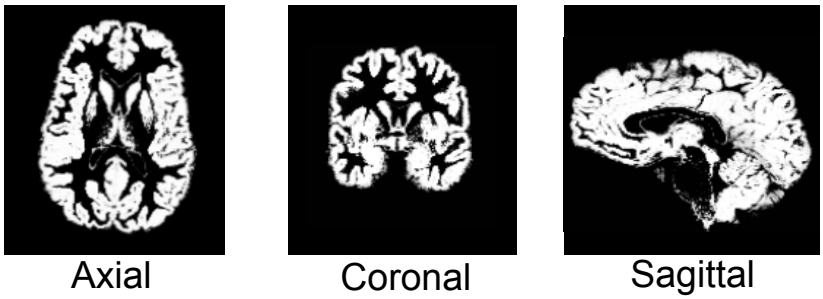
(b) AC-PC Realignment on a sample image.

Figure 3.4: (a) AC-PC reference line and (b) Results after AC-PC Realignment on an OASIS dataset image.

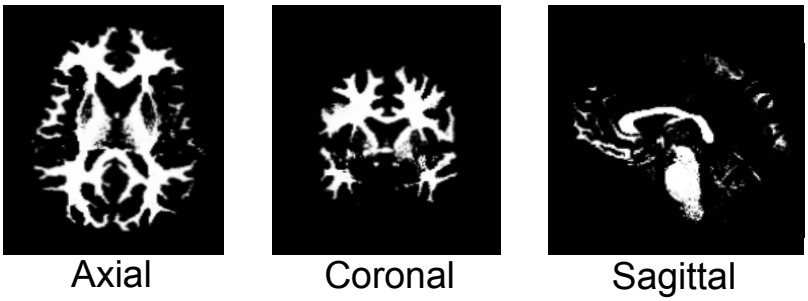
The gray and white matter segments and their rigidly aligned counterparts are shown in Figure 3.5 and Figure 3.6. The CSF segment is not shown here as we did not use it for further preprocessing. The imported rigidly aligned segments are used for the template generation step. Segmentation and DARTEL template generation steps are resource-intensive. It roughly processes 10-20 images per hour on a VM on a server cluster. Hence, the segmentation was divided across 4 VMs to enable faster processing. However, the DARTEL template generation cannot be divided in such a manner as it generates a single template for the entire dataset. If we parallelized this then there may be multiple sets of templates generated.

Generate DARTEL Normalization Template

In this step, we used DARTEL to compute a deformation field (also referred to as flow fields) for each subject. These deformation fields are used to perform spatial normalization.

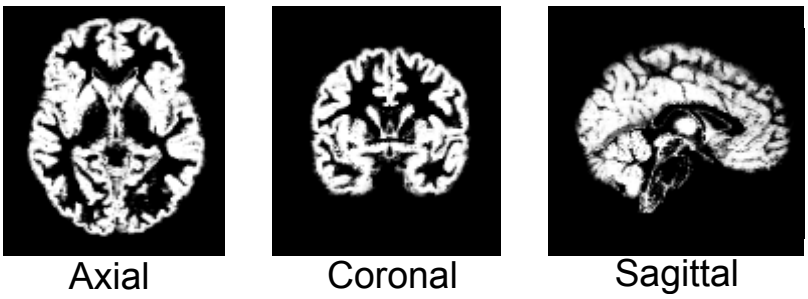


(a) Gray matter segment from OASIS-1 subject 0014 after segmentation.

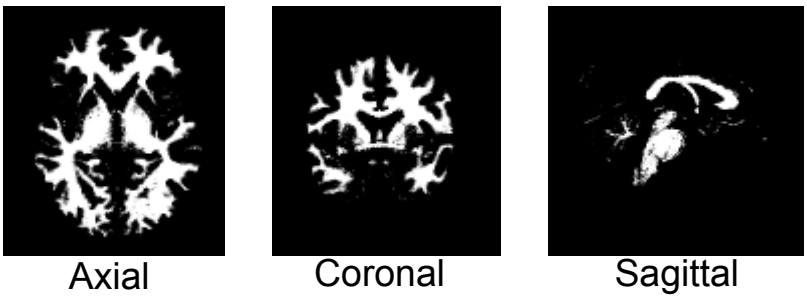


(b) White matter segment from OASIS-1 subject 0014 after segmentation.

Figure 3.5: Gray and White matter segments obtained from segmentation.

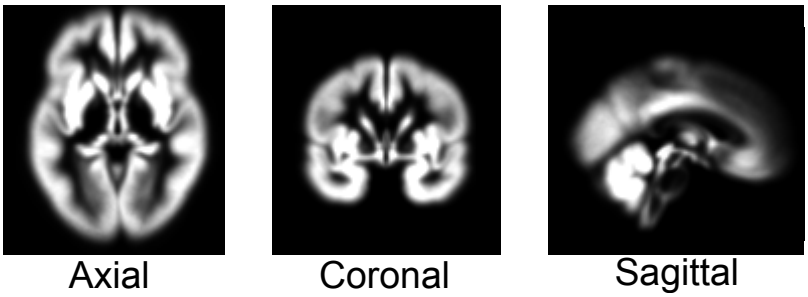


(a) Rigidly aligned gray matter segment from OASIS-1 subject 0014.

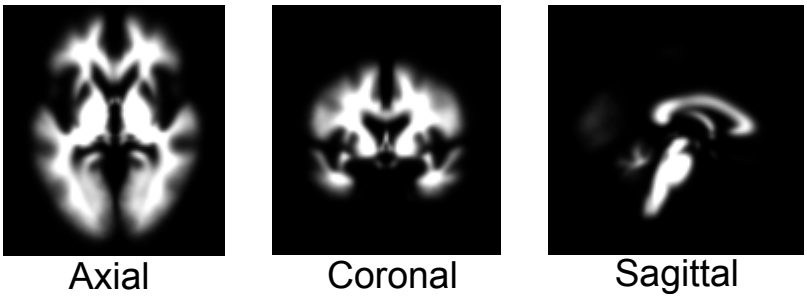


(b) Rigidly aligned white matter segment for OASIS-1 subject 0014.

Figure 3.6: Rigidly aligned Gray and White Matter segments obtained from segmentation.



(a) Initial template for gray matter segment.



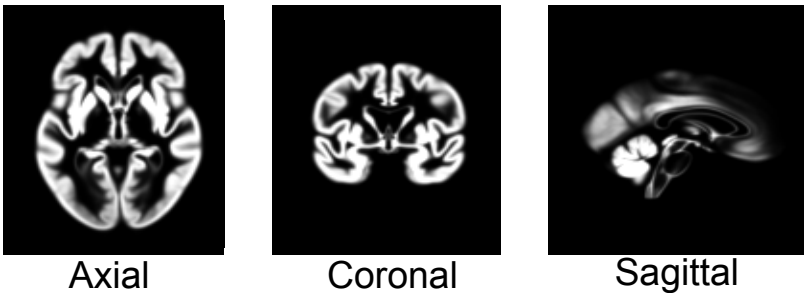
(b) Initial template for white matter segment.

Figure 3.7: Initial DARTEL template.

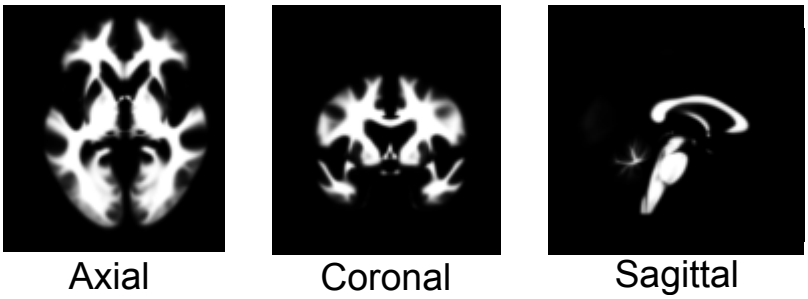
To estimate the deformations an initial template file Template_*0*.nii was generated for the entire dataset using the rigidly aligned segments of all the subjects (rc1*.nii, rc2*.nii files). Subsequently, each subjects' segments were registered with the template. After this entire iteration, a new Template was generated and was used as an initial template for the next iteration. This process of building and registering templates was repeated over 6 iterations. The subsequent iterations were slower as they required more time-steps to generate the deformations. The result of this procedure was a set of templates, the initial and final templates generated can be seen in Figure 3.7 and Figure 3.8 respectively. The final template appears visually sharper than the initial template. These template files are generally multi-volume files i.e. in any neuroimaging file viewer we have to specify the volume number i.e. 1 (corresponding to gray matter), 2 (corresponding to white matter). Besides the templates, a flow-field file namely uc1*.nii (for the gray matter) was generated for every subject. These files were used to parameterize the deformations for the gray matter segments (c1*.nii files) in the spatial normalization step. We did not use the flow-fields for white matter (uc2*.nii files).

Spatial Normalization with Modulation and Smoothing

Human brains are of different shapes and sizes. Hence, the MRI data needs to be normalized so that a particular region in the brain of a particular subject is in the same position for another subject. Hence, the templates were registered on the segments and the deformation fields (also referred to as flow fields) were generated using DARTEL as described in



(a) Final template for gray matter segment.



(b) Final template for white matter segment.

Figure 3.8: Final DARTEL template.

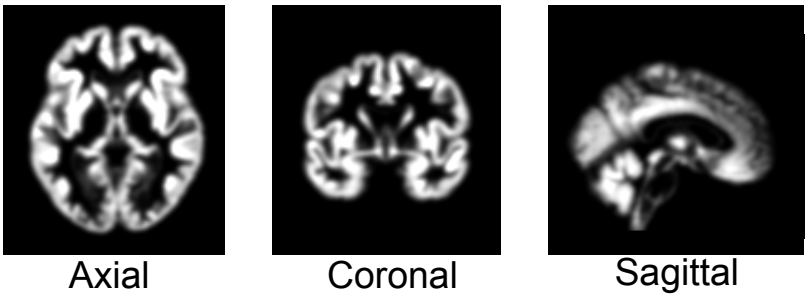


Figure 3.9: Final preprocessed gray matter for OASIS-1 subject 0014.

the previous step. We then performed spatial normalization on the gray matter segment (c1*.nii files) images from the DARTEL template space to Montreal Neurological Institute (MNI) template space which is an affine transformation. While normalizing we performed modulation i.e., "preserve amount", this step preserves the total amount of signal from each region of the scan even if the images are expanded or shrunk during the transformations [ABC+14]. Further, we performed smoothing using a Gaussian kernel with a Full Width at Half Maximum (FWHM) magnitude of 4mm along x, y, z axes. The repository by [Soc17] provided a good starting point to implement the normalization procedure from DARTEL templates. The resulting image (smwc1*.nii files) from our preprocessing procedure is shown for a sample subject in Figure 3.9. The preprocessed images have dimensions 121 x 145 x 121.

Conversion to TFRecord format

Finally, the preprocessed images were converted into TFRecord format [MAP+15]. It enables the storage of the data as protocol buffers for efficient serialization and storage of structured data. After the serialization process, these files were divided into training and testing folders. However, this split was not trivial, because a subject could have multiple scans. Hence, we did not split the data based on scans but based on subjects. This means multiple scans of a subject should not scatter across training and testing sets, this ensures the model does not gain an unfair advantage by training on subjects' scans whose other scans may belong to the test set. The overall size of the preprocessed dataset in the .tfrecord format was around 20 GB wherein each .tfrecord file was around 8-9 MB in size.

4 Methods & Experiments

In this section, we introduce and describe Age-Net architecture. Next, we discuss the iterative training framework to obtain cleaned data free from outliers. Finally, we explain the visualization procedure in detail and apply it to the BA estimation model to obtain visualizations on healthy and outlier subjects.

4.1 Biological Age Estimation Framework

The Age-Net architecture proposed by [AAB+20], [AAS+20] is a regression network that performed better than the existing state-of-the-art networks to predict the chronological age (CA) from MRI data. The Age-Net architecture is depicted in Figure 4.1. The MRI input of size 121 x 145 x 121 is fed into the model as 20 non-overlapping 3D chunks each of size 121 x 145 x 6. Feeding the input as smaller chunks instead of an entire 3D volume needs comparatively less memory while training. Moreover, it does not require on the fly data augmentation. The architecture starts with a stem network with 3 x 3 x 3 convolutions and max-pooling layers. This is followed by four consecutive connected modules. Here each module consists of two inflated versions of Inception v1 [SLJ+14] blocks followed by a fire module proposed in SqueezeNet [IHM+16]. The standard inception modules were adapted to handle the 3D inputs by inflating all convolution filters and pooling kernels. The fire module uses squeeze-and-expand layers which reduces the total number of training parameters. The squeeze-and-expand layers consist of 1 x 1 x 1 and 3 x 3 x 3 convolutions followed by a concatenation layer. In each of the blocks, every convolution layer uses L2 regularization and is followed by a batch-normalization layer and a ReLU activation.

After the 4 consecutive connected modules discussed above, a global average pooling (GAP) is applied to reduce the high-dimension tensor to a vector with 512 features. Next, we concatenated the gender of the subject with the result from the GAP layer. This is followed by three dense layers of size 512, 256, and 128 features respectively. Finally, a regression layer is used to generate the age estimate. The architecture for the same is depicted in Figure 4.1. The architecture is explained in greater detail in [AAS+20].

The authors of [AAB+20], [AAS+20] have further used the Age-Net model in a proposed iterative training framework to obtain cleaned data to train the BA estimator. The data feeding strategy is the same as it was in the CA estimation discussed before. As depicted in Figure 4.2, the training dataset is further divided into training and validation subsets. In every iteration, the model is trained on the training subset and is predicted on the

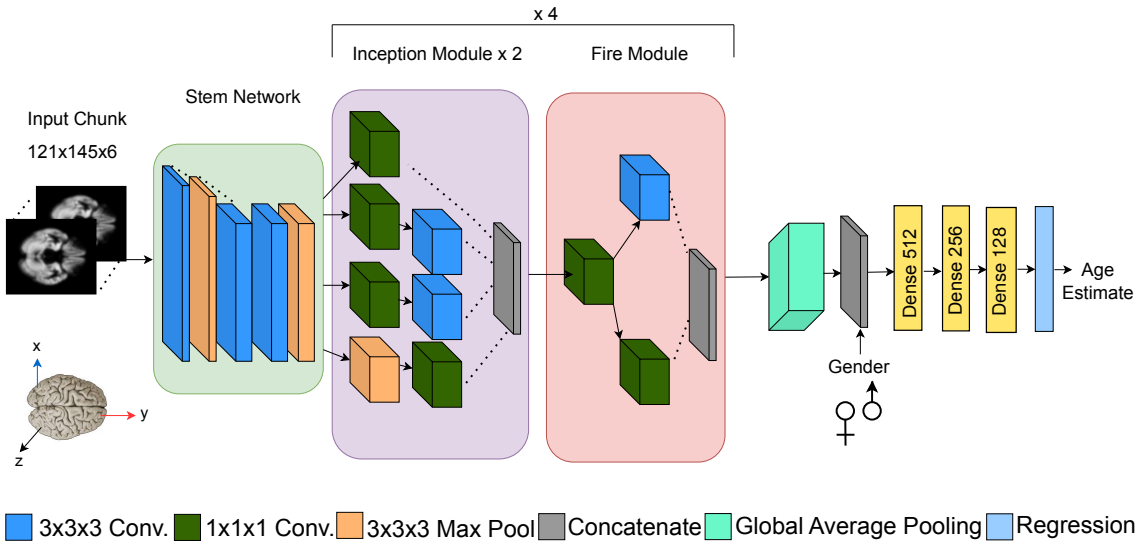


Figure 4.1: Age-Net Network Architecture [AAB+20], [AAS+20].

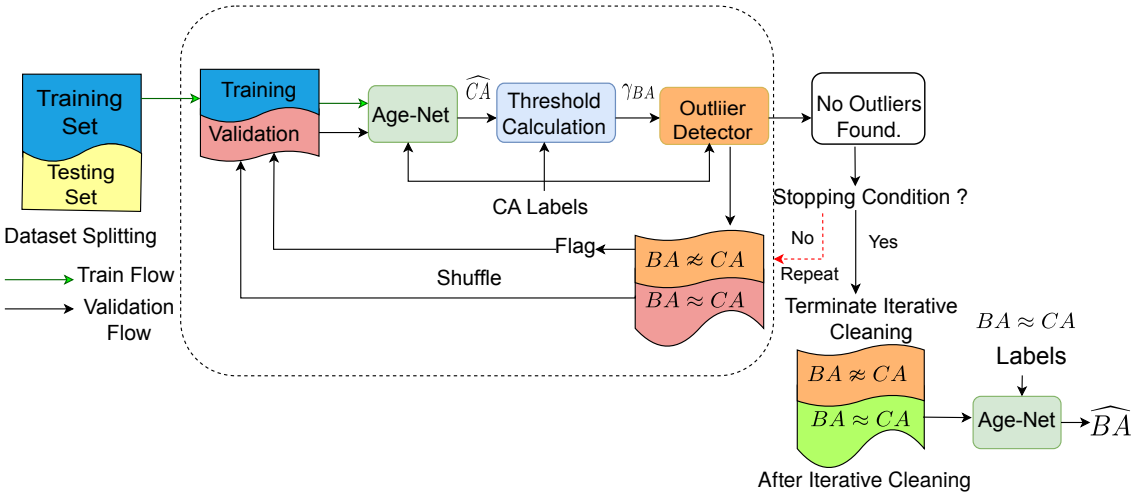


Figure 4.2: Detailed overview of the iterative data cleaning procedure. [AAB+20], [AAS+20]

validation subset to segregate subjects with a significant predicted age deviation $\left|\widehat{CA} - CA\right|$ are identified as outlier subjects. As we feed the input as K chunks, we obtained the predicted CA for every n^{th} subject by computing a mean on the predicted age for individual chunks which is shown as follows [AAS+20]:

$$\widehat{CA}_n = \frac{1}{K}\sum_{i=1}^{K} \widehat{CA}_{n,i} \tag{4.1}$$

The deviation and the threshold can be defined as follows [AAS+20]:

$$\underline{D} = \left|\underline{\widehat{CA}} - \underline{CA}\right| \quad , \quad \underline{\gamma_{BA}} = R \cdot \underline{\sigma} \tag{4.2}$$

Here $\underline{\widehat{CA}}$ and $\underline{\sigma}$ are the predicted CA and corresponding standard deviation vector for all the subjects. R is a constant value set to 1.96 which represents 95% of the confidence interval of the mean predictions of each subject, assuming the chunk predictions follow a normal distribution. Every n^{th} subject is flagged as an outlier if the deviation for this subject is greater than the corresponding threshold $D_n > \gamma_{BA,n}$ [AAS+20]. After every iteration, the training and the validation subsets are shuffled as depicted in Figure 4.2. This iterative training procedure is continued until an empirical stopping condition is satisfied. This stopping condition states that the iterative procedure can be terminated if no new outliers have been flagged in three consecutive data-cleaning iterations trained with different initializations. Finally, the subjects who have been flagged as outliers more than once are removed from the training set.

The rationale behind the iterative training is that the cleaned training dataset obtained after this procedure would have subjects with $BA \approx CA$. Therefore we can use the CA as biological age labels to train the BA estimation model on this cleaned training data. The outliers obtained from the iterative framework are the subjects which according to Age-Net exhibit atypical aging characteristics. These atypical aging characteristics could be owed to abnormalities in the brain gray matter which may indicate ailments, which we will discuss in detail in the later section. This training process being an iterative procedure, it takes longer to train compared to the CA estimation training. The training configuration for the iterative procedure [AAS+20] is given in Appendix A.3.

It is interesting to note that few subjects were healthy CDR 0 according to the dataset but considered as outliers by the iterative cleaning process. These cases have also been visualized to identify the regions which are likely suffering from atrophy and are further discussed in Chapter 5.

4.2 Visualization Experiments on BA Estimation Model

We visualized the gray matter chunks from the outlier set (CDR 0, 0.5, 1, 2) and the healthy (CDR 0) subjects from the non-outlier set as depicted in Figure 4.3. Efficient Saliency Maps for Explainable AI [MCF20] proposed by Mundhenk et. al, is a gradient-free technique and is hence computationally more efficient in comparison to gradient-based techniques such as GradCAM [ZKL+16], GradCAM++ [CSHB18]. A combination of more than one technique can significantly improve the visualizations. Therefore we combined the strengths of two techniques, for our dataset we implemented and applied a combination of SMOE Scale saliency maps [MCF20] and GradCAM++ [CSHB18] as the SMOE Scale saliency maps are comparatively fine-grained whereas maps generated by gradient-based class activation maps (such as GradCAM, GradCAM++) are highly localized.

The Saliency Map Order Equivalent (SMOE) Scale approach proposed in [MCF20] was used to generate the saliency maps as described in Section 2.5. The saliency maps were generated at different spatial scales (i.e. for convolution output activation tensors with

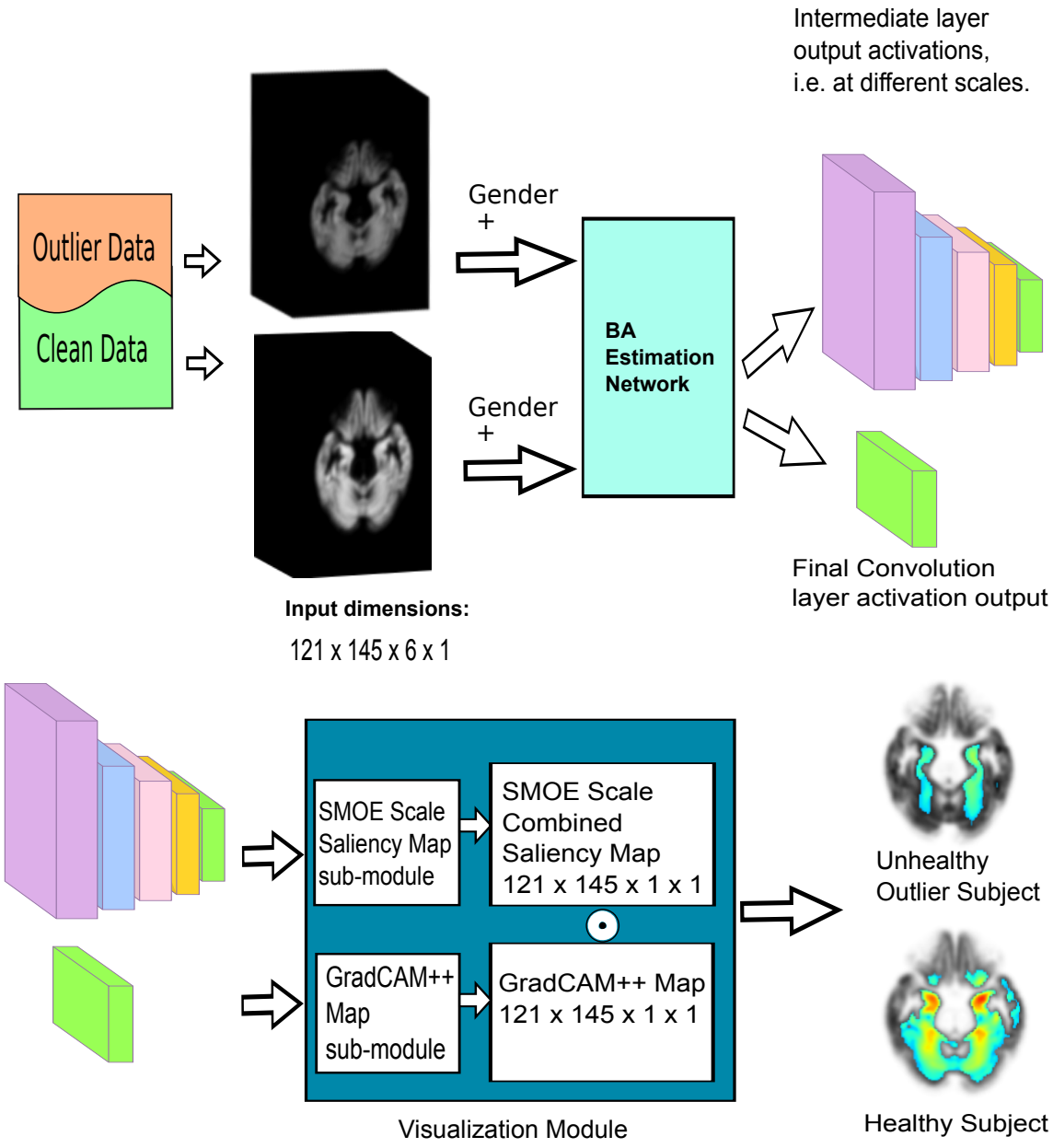


Figure 4.3: A high-level view of the visualization procedure.

different dimensions) as depicted in Figure 4.3. A total of 5 intermediate maps were generated for our network. In our case, the input chunk being 3 dimensional, the mean was computed along the depth axis of the spatial outputs before computing the above maps. After computing the maps, these individual maps were normalized using a normal cumulative distribution function as implemented in [MCF20]. The resulting maps were then interpolated to the same $height$ x $width$ as that of the input image chunk, i.e. (121, 145) in our case. Then these maps were weight-summed to generate a combined saliency map, we used equal weights while performing the weighted summation.

Subsequently, for GradCAM++ [CSHB18] the back-propagated gradients from the model output till the last convolution layer were computed. Then a weighted summation of gradients is performed to obtain w_k^c. The weighting coefficients α^{kc} for the same were computed using the higher-order derivatives of the gradients as proposed in [CSHB18] and as explained in Section 2.3. The GradCAM++ map is obtained by multiplying output activation maps from the last convolution layer with the computed w_k^c. The resultant map was passed through a ReLU to include only the positive values and then resized to the size of the input image chunk. Subsequently, the maps were scaled to 0 to 1, as implemented in [CSHB18]. It is important to note that here we are computing GradCAM++ on 3D inputs, hence the mathematical formulation from Section 2.3 was adapted as follows:

$$w_k^c = \sum_i \sum_j \sum_l \alpha_{ijl}^{kc} ReLU\left(\frac{\partial y^c}{\partial A_{ijl}^k}\right) \tag{4.3}$$

$$\alpha_{ijl}^{kc} = \frac{\frac{\partial^2 y^c}{(\partial A_{ijl}^k)^2}}{2\left(\frac{\partial^2 y^c}{(\partial A_{ijl}^k)^2}\right) + \sum_a \sum_b \sum_d A_{abd}^k \left(\frac{\partial^3 y^c}{(\partial A_{ijl}^k)^3}\right)} \tag{4.4}$$

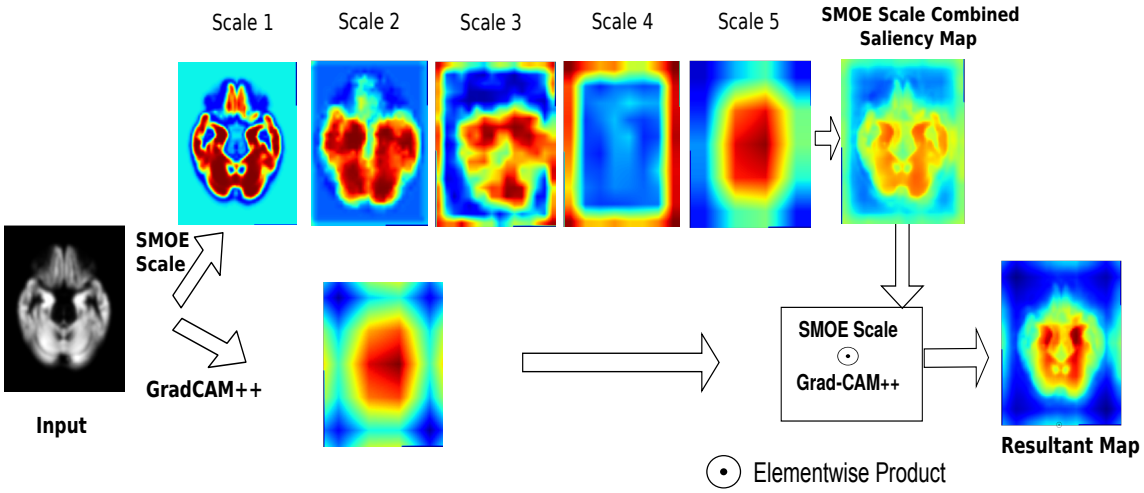
$$L_{ijlGradCAM++}^c = ReLU\left(\sum_k w_k^c A_{ijl}^k\right) \tag{4.5}$$

To visualize the map in 2D, the mean was computed along the depth axis of the 3D GradCAM++ map, therefore, transforming the map with dimensions (121, 145, 6) to (121, 145).

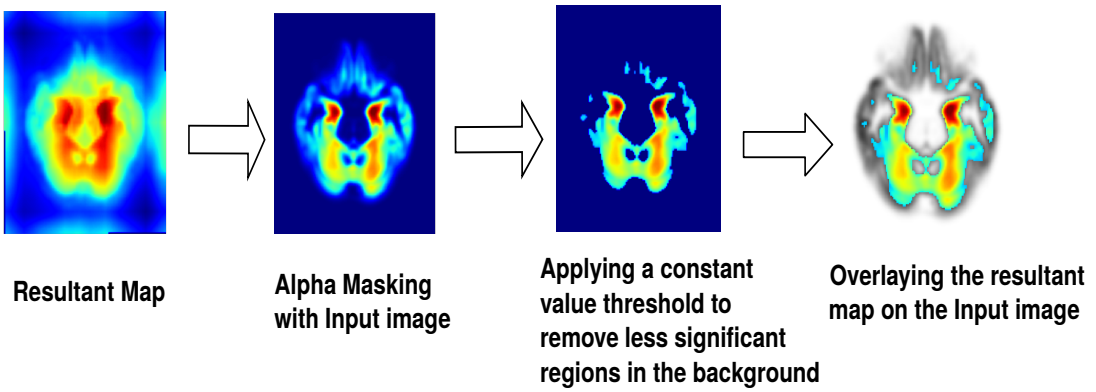
Once both maps were obtained, a pixel-wise product operation between the maps was performed to generate the final localization map L_{final_map} which is our hybrid visualization map. Using the Equations (2.8) and (2.18) we can formulate the final hybrid saliency map as follows:

$$L_{final_map} = L_{SMOE-Scale} \odot L_{GradCAM++} \tag{4.6}$$

Figure 4.4a illustrates the SMOE Scale Saliency maps at different scales and the combined map. Additionally, it shows the GradCAM++ map and its product with the SMOE saliency map to generate the final saliency map. We know that GradCAM++ generates accurate



(a) Generation of SMOE Scale [MCF20] and GradCAM++ [CSHB18] Maps and the resulting hybrid visualization map. Here the SMOE scale maps at different scales have been resized to *height* x *width* of the input image.



(b) Cleaning the visualization map and overlaying on the gray matter input.

Figure 4.4: Detailed visualization stages (a) Generation of individual SMOE scale saliency maps and the combined map. This is then multiplied element-wise with GradCAM++ map to generate the visualization map. (b) Here further post-processing is applied to generate a cleaner map without any background and then overlaid on the gray matter slice.

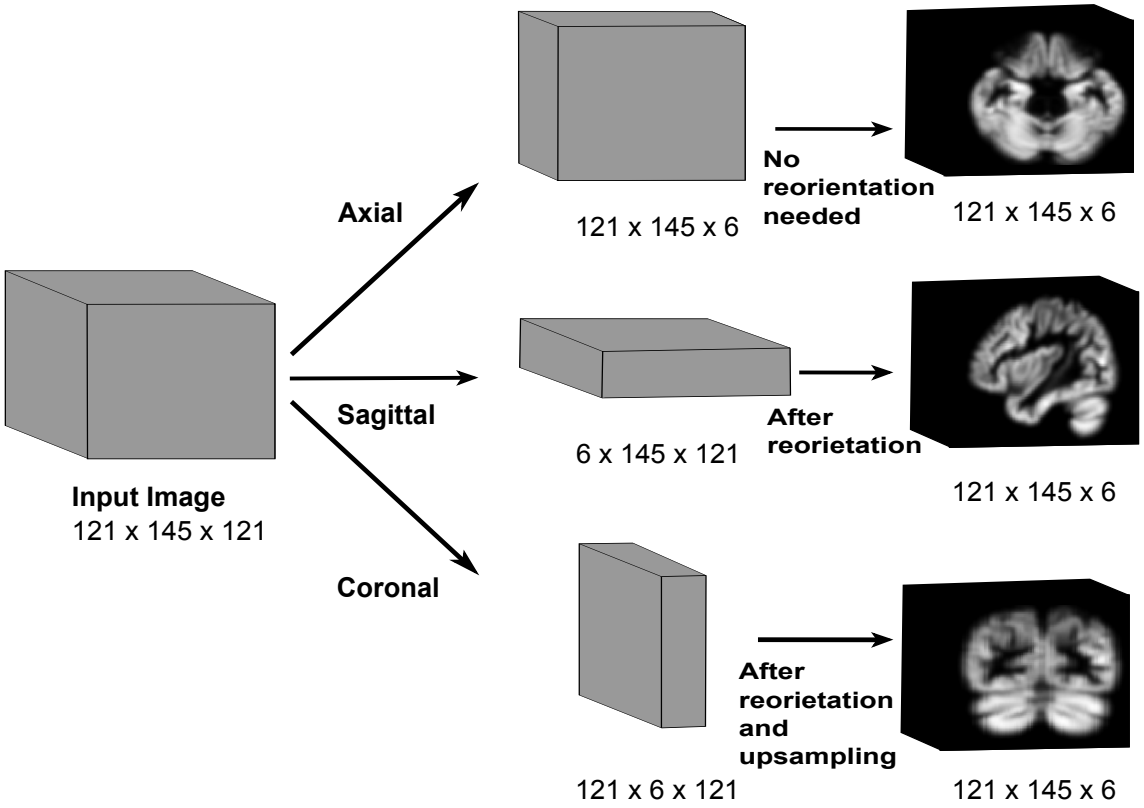


Figure 4.5: Generating an input chunk from the preprocessed input across axial, coronal, and sagittal planes.

visualization maps for locating objects in object classification or object detection type of tasks [CSHB18]. However, Figure 4.4a shows that the GradCAM++ map is very coarse and is not suitable for identifying intricate regions within a single object, hence using this map alone is not suitable for our task. Whereas, the SMOE Scale map highlights internal regions of an object much better. Therefore combining these two maps (as shown in Equation (4.6)) generated a map that preserved the fine-grained nature of the SMOE Scale map and at the same time preserved the localized higher intensity regions from the GradCAM++ map.

Further, an alpha masking with the input gray matter was performed with the resultant map as performed in [MCF20]. Finally, an appropriate constant threshold was applied to the map (to include more significant regions) and overlaid on the input gray matter slice as depicted in Figure 4.4b. For a preprocessed input image of a subject, the above visualization procedure was applied to multiple gray matter chunks across all three planes:

- Axial
- Coronal
- Sagittal

We will further illustrate how did we generate the coronal and sagittal input chunks to produce visualizations for the same. Given our input size 121 x 145 x 121, we feed the trained Age-Net model with non-overlapping 3D chunks each of size 121 x 145 x 6, (e.g. 6^{th} axial chunk would be expressed as [121, 145, 30:36]). The generation of the input chunks across the three planes is depicted in Figure 4.5. By comparing with regions on the brain atlas [MPV09], the significant input chunks to be visualized were identified. Multiple chunks were needed to be considered because different chunks could contain different regions that may have been affected by dementia and ultimately Alzheimer's Disease. The regions were visualized for chunks 6 to 12. These results will be discussed in detail in Chapter 5.

Other methods such as GradCAM and SmoothGradCAM++ were also applied with the SMOE Scale saliency maps approach. However, in comparison to GradCAM++, GradCAM is a relatively older technique with results equal or inferior to GradCAM++ [CSHB18]. On the other hand SmoothGradCAM++ is an extension of GradCAM++ as explained in the Related Work Section 2.4. In SmoothGradCAM++, for a given input a fixed number of random samples are generated from a normal distribution centered on this input with a constant standard deviation as noise. Once the gradients have been computed for these samples, a gradient averaging is performed as proposed in [OSCW19] to generate the final map for an input. This process is hence computationally expensive and when we compare the maps with the ones obtained using GradCAM++, there are no visible improvements as shown in Appendix A.2. Moreover, the parameters, i.e. number of samples and standard deviation need to be selected beforehand.

The model training was performed on the ISS server cluster and University Hospital Tübingen server cluster. All the visualizations were computed on the Google Colaboratory environment. This being a visualization project, using a notebook based environment was helpful for quick prototyping and experimentation. The notebooks can be found in the GitHub [1] repository linked below.

[1] https://github.com/shashank3110/Master_Thesis_BA_DeepVis

5 Visualization Results and Significant regions correlated with Dementia

The billions of cells in the human body are much younger than our actual age (CA) because most types of cells have a shorter lifespan and are periodically replenished multiple times across our lifetime. The nerve cells in the human body unlike the majority of cells with shorter lifespan are evolved to have a much longer lifespan and are considered to last a lifetime for humans. Hence, with normal aging, though the brain may shrink, there is no significant loss in the number of neurons inside the brain and the rate of brain atrophy is subtle. However, in the case of AD, there is accelerated atrophy causing significant loss of neurons in the brain and as these neurons get damaged they lose connections with other neurons and these cells eventually die. The severity increases as this process spread throughout the gray matter in the brain.

Subjects diagnosed with Alzheimer's Disease have shown accelerated memory loss and aberrant emotional balance. The limbic system is a collective region in the brain which is primarily responsible for memory and emotions. It consists of several regions mainly the hippocampus, entorhinal cortex, amygdala, parahippocampal gyrus, cingulate gyrus, corpus callosum, and the regions in and around the thalamus & the hypothalamus. The anatomy of the Limbic System is depicted in Figure 5.1. From the onset of Alzheimer's Disease, the regions in the medial temporal lobe especially the hippocampus and the entorhinal cortex show significant atrophy since the early stages. The early stages of the disease are commonly associated with neuron degeneration in these two regions [SHR17]. However, studies such as [PDM+12] suggest that the amount of atrophy witnessed in the amygdala is comparable to that in the hippocampus for very mild dementia (CDR 0.5) to mild dementia (CDR 1.0) subjects.

The functions of the regions discussed above give us a better understanding of why these regions in particular are responsible for AD. The hippocampus is responsible for forming new memory and converting short-term memory to long-term memory. The entorhinal cortex and the hippocampus are also responsible for spatial memory, hence helping in navigation [SRN+11]. Whereas the amygdala is responsible for emotions such as anger, fear, anxiety, and aggression to name a few. Degeneration in the amygdala could lead to impulsive behavior, hyper-orality, hence causing an emotional imbalance. The cingulate gyrus belonging to the cingulate cortex is mainly responsible for processing emotions and regulating endocrine functions in response to the emotions [JD20]. Additionally, the cingulate gyrus is also responsible for manifesting emotions to actions & controlling imagination and orientation [JD20]. The thalamus is responsible for relaying our senses

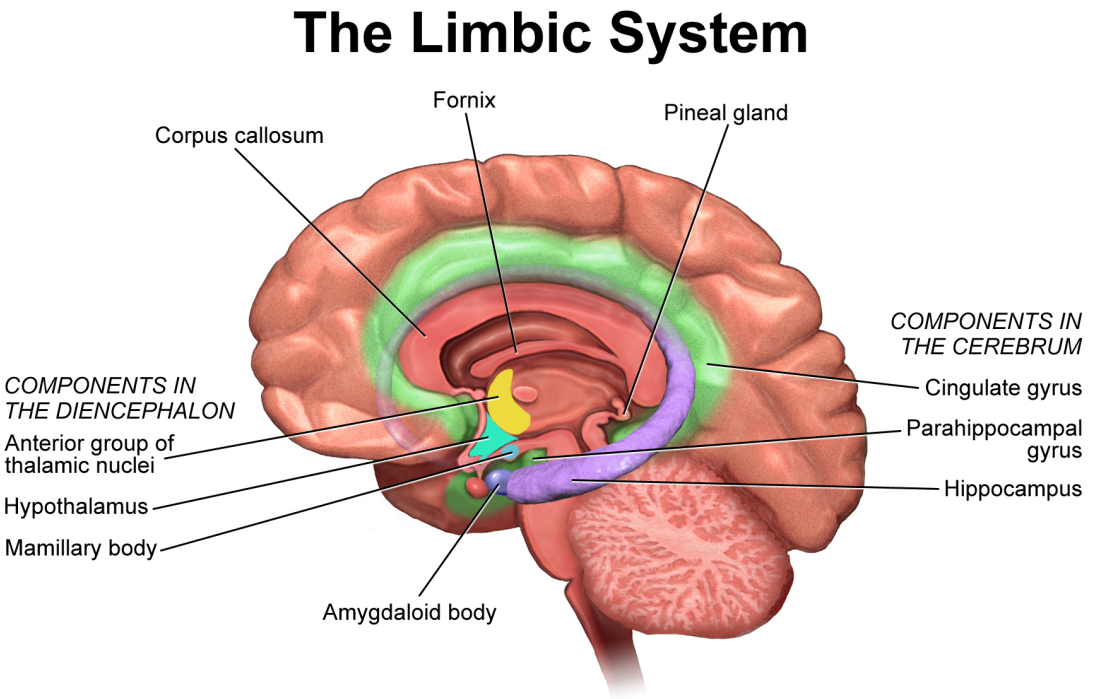


Figure 5.1: Anatomy of the Limbic system [sta14].

(except smell) to the appropriate regions in the cerebral cortex which are responsible for processing these sensory signals [MS92]. Whereas the hypothalamus regulates the autonomic nervous system (ANS), the ANS is responsible for involuntary actions such as respiratory control, digestion, and reflexes such as coughing and sneezing [MS92]. It also triggers the endocrine system which in turn regulates the hormones' secretion and is hence responsible for processing emotions into actions [MS92]. As the severity of AD increases, the extent of neuron degeneration can spread to other regions in the cerebral cortex across the temporal and frontal lobes. The visualization maps obtained by combining the SMOE Scale saliency map [MCF20] with the GradCAM++ map [CSHB18] were computed (as explained in Section 4.2) on the healthy subjects from the non-outliers' set and on the subjects from the outliers' set. These were obtained from iterative data cleaning by the BA estimation framework. The most significant regions observed in general for the healthy non-outlier subjects have been highlighted in Figure 5.2. This figure contains visualization maps of healthy CDR 0 subjects.

A significant number of highly activated regions in the visualizations of healthy subjects correlate to the regions in the limbic system, hence, indicating a correlation with Alzheimer's Disease. These regions along with some regions in the cerebral cortex and the cerebellum showed reduced activation in visualization maps for the outlier subjects as depicted in Figure 5.3. This firmly indicates tissue atrophy in these regions.

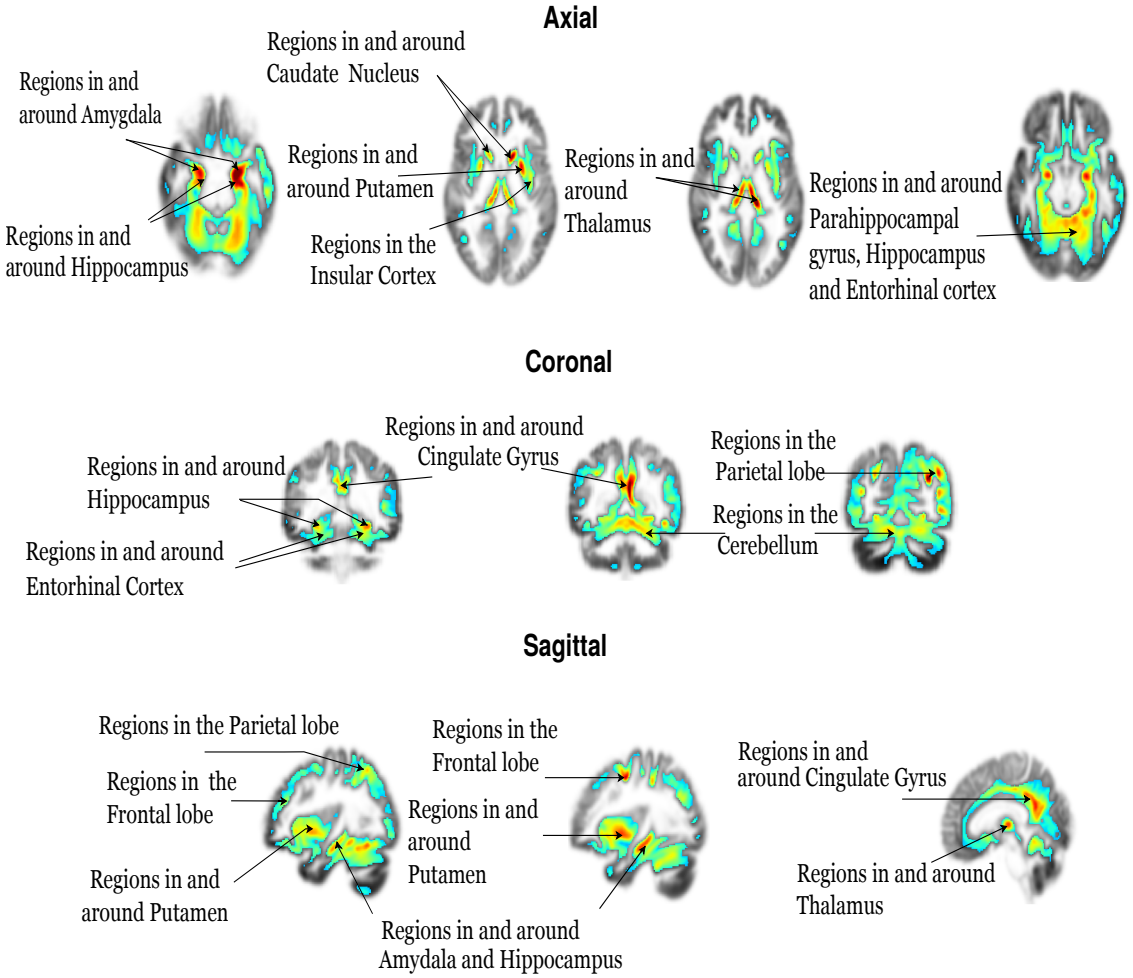


Figure 5.2: Identifying activated significant regions in the visualization maps for healthy CDR 0 subjects across different chunks and planes. Regions in red and yellow indicate regions of higher activation compared to the regions in green and blue.

The healthy CDR 0 subjects from the non-outlier set showed higher activation in the specific regions (regions highlighted in Figure 5.2) of the gray matter compared to the subjects in the outlier set as shown in Figure 5.3. The outlier set contained subjects with atypical aging having CDR 0.5, 1, 2, and even subjects with CDR 0. As expected as the severity of Alzheimer's Disease progressed, the intensity of the maps in these regions reduced substantially.

A notable observation is that there was a substantial difference in the visualization maps between CDR 0 non-outliers and CDR 0 outliers, this indicates that these outlier subjects though have CDR 0 but are likely to be diagnosed with Alzheimer's Disease in near future or are suffering from mild-Alzheimer's Disease. Such observations would assist in diagnosing the disease far earlier and suggest relevant precautions and treatments and prevent it from accelerating.

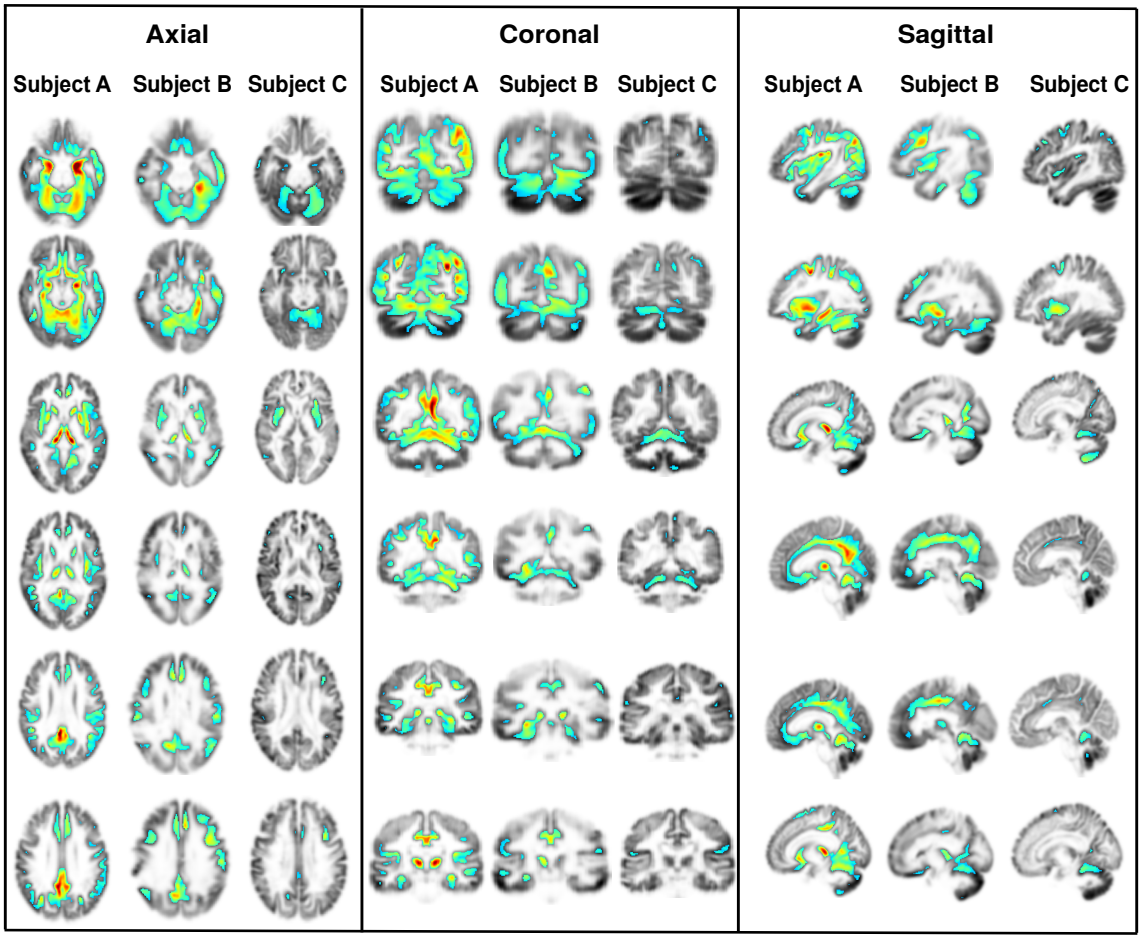


Figure 5.3: Visualization maps comparison for 3 different subjects across various chunks and planes from the outlier and clean subsets obtained from the iterative framework. Subject A:[CDR 0, CA 58.71, Healthy]; Subject B:[CDR 0, CA 72.13, Outlier]; Subject C: [CDR 2, CA 77.71, Outlier]. The regions in this figure can be compared with the activated regions in Figure 5.2. We can see that subject A being a healthy non-outlier has far more highly activated regions for a given chunk and in a given plane when compared to the outlier subjects B & C.

We further visualized the subjects from the test-set as shown in Figure 5.4. Here we have shown subjects both with low and very high deviations of the BA estimation model's predicted age from the CA of the subjects. An important thing to note is that the low or high deviation in predicted ages for the subjects in the test-set does not imply that they are healthy or outliers. As discussed in Section 4.1 the iterative cleaning was performed on the data from the training set. In the above figure, we can see that the CDR 0 subjects with high deviation in predicted ages have visualizations with weak activations compared to CDR 0 subjects with low deviation in predicted age.

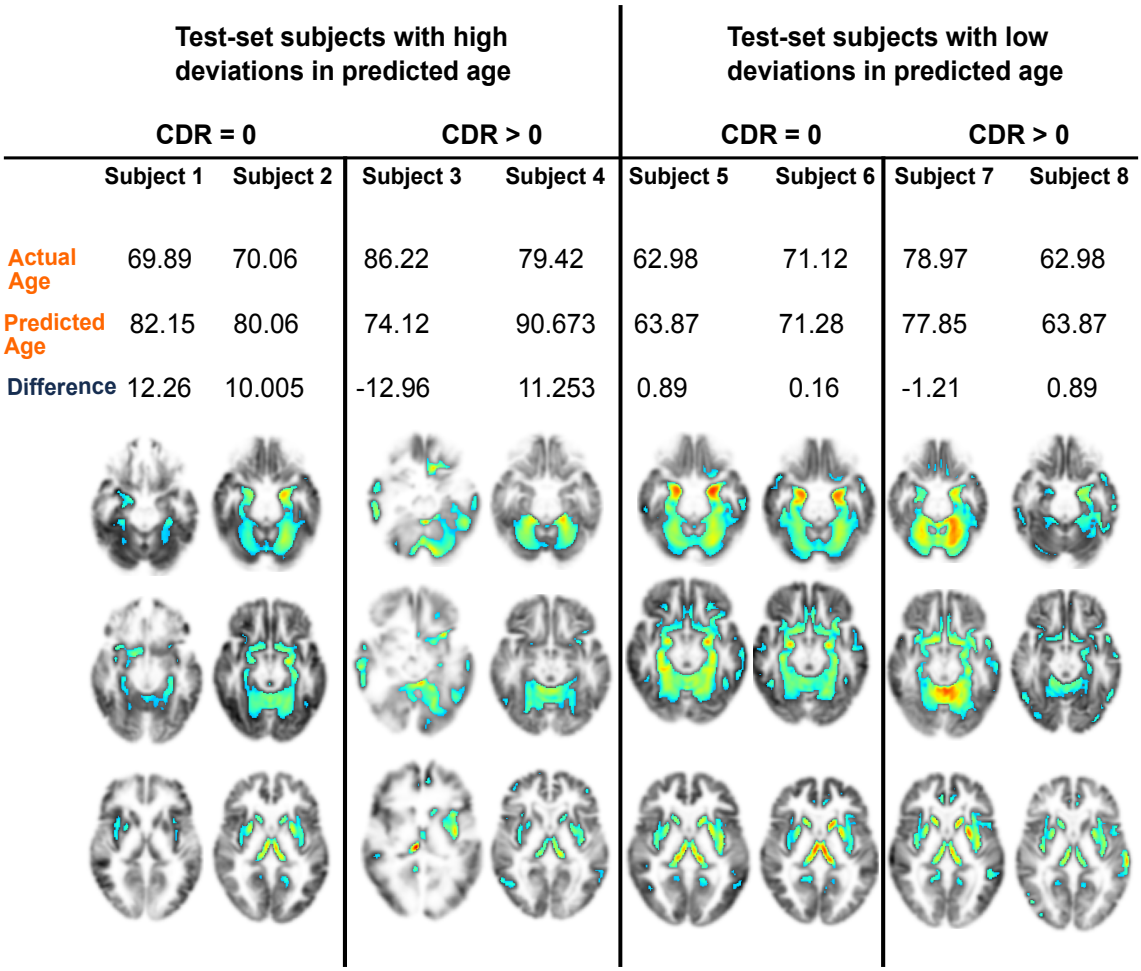


Figure 5.4: Visualization Maps Comparison across axial chunks for subjects from test-set.

6 Conclusion and Outlook

Deep Visualization techniques have been improving over the years to provide fine-grained, well-localized visualizations that are faster and computationally efficient to compute. These techniques are not restricted to traditional vision tasks like object classification and have been evolved to apply to tasks like VQA, image captioning, image segmentation, attention explanations to name a few.

In this thesis, we adapted some of the deep visualization techniques and applied them to the Biological Age Estimation framework. Through this work, we were able to identify the regions in the input data i.e. in the brain gray matter whose atrophy lead to accelerated or abnormal aging of the brain. This is indicated by the biological age being significantly greater than the actual age (chronological age). In addition to this, we were able to analyze how the reduced intensities in these regions as shown by our generated visualization maps significantly correlated with the regions affected by Alzheimer's Disease. The thesis also bolsters the efficacy of the iterative outlier detection in the BA estimation framework as some subjects though clinically healthy (i.e. CDR 0), their visualizations showed some characteristics of AD and are hence labeled as outliers. This would assist healthcare professionals to identify the onset of dementia specifically Alzheimer's Disease before it turns very severe.

This thesis provided a comprehensive application of the visualization techniques by considering the 3D aspect of the data and extending it to the visualization of all three planes: axial, coronal, sagittal. Moreover, visualization at every chunk level helped us identify and cover most of the intricate regions in the brain which otherwise may have been occluded in some chunks.

However, there are probable shortcomings and opportunities to improve further. Firstly, the dataset could have been more balanced with more subjects with moderate and severe dementia (CDR 2 & 3). Moreover, additional data about the lifestyle of the subjects (i.e. dietary and recreational habits) could be supplemental features for a thorough analysis. The existing BA estimation model has only been trained with axial chunks. We can further extend it to train on coronal and sagittal chunks for a better understanding of the input data resulting in better outlier detection. Our work can be further extended to identify significant region activations at the level of individual feature maps instead of only at layers. This would help us identify more intricate sub-regions individually.

Overall this thesis is a starting step in the direction of building reliable AI systems for assisting radiologists and other specialists in efficiently analyzing medical data. These systems would also aid them in identifying critical medical conditions in their subjects

and advise timely remedies. The AI-based systems are not here to replace the human specialists (i.e. doctors, radiologists) but to assist them in doing their work more efficiently and productively in this stressful and safety-critical profession. Explainable AI techniques help further gain the trust of these specialists and the general public on such AI-based healthcare systems.

Bibliography

[AAB+20] K. Armanious, S. Abdulatif, A. R. Bhaktharaguttu, T. Küstner, T. Hepp, S. Gatidis, B. Yang. "Organ-based Chronological Age Estimation based on 3D MRI Scans". In: (2020). arXiv: 1910.06271 [eess.IV] (cit. on pp. 13, 14, 37, 38).

[AAS+20] K. Armanious, S. Abdulatif, W. Shi, S. Salian, T. Küstner, D. Weiskopf, T. Hepp, S. Gatidis, B. Yang. "Age-Net: An MRI-Based Iterative Framework for Biological Age Estimation". In: (2020). arXiv: 2009.10765 [eess.IV] (cit. on pp. 14, 37–39, 60).

[ABC+14] J. Ashburner, G. Barnes, C.-c. Chen, J. Daunizeau, G. Flandin, K. Friston, S. Kiebel, J. Kilner, V. Litvak, R. Moran, et al. *SPM12 Manual The FIL Methods Group (and honorary members)*. 2014. URL: https://www.fil.ion.ucl.ac.uk/spm/doc/manual.pdf (cit. on pp. 30, 34).

[ABC+20] J. Ashburner, G. Barnes, C.-c. Chen, J. Daunizeau, G. Flandin, K. Friston, S. Kiebel, J. Kilner, V. Litvak, R. Moran, et al. 2020. URL: https://www.fil.ion.ucl.ac.uk/spm/doc/ (cit. on p. 30).

[Ash07] J. Ashburner. "A fast diffeomorphic image registration algorithm". In: *NeuroImage* 38.1 (Oct. 2007), pp. 95–113. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2007.07.007. URL: http://dx.doi.org/10.1016/j.neuroimage.2007.07.007 (cit. on p. 30).

[BBM+15] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PLOS ONE* 10.7 (July 2015), pp. 1–46. DOI: 10.1371/journal.pone.0130140. URL: https://doi.org/10.1371/journal.pone.0130140 (cit. on p. 15).

[BDD+20] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (June 2020), pp. 82–115. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012. URL: http://dx.doi.org/10.1016/j.inffus.2019.12.012 (cit. on pp. 14, 15).

[BMH+20] M. Brett, C. J. Markiewicz, M. Hanke, M.-A. Côté, B. Cipollini, P. McCarthy, C. P. Cheng, Y. O. Halchenko, M. Cottaar, S. Ghosh, et al. *nipy/nibabel: 3.0.2*. Version 3.0.2. Mar. 2020. DOI: 10.5281/zenodo.3701467. URL: https://doi.org/10.5281/zenodo.3701467 (cit. on p. 28).

[BMS+20] I. Beheshti, S. Mishra, D. Sone, P. Khanna, H. Matsuda. "T1-weighted MRI-driven Brain Age Estimation in Alzheimer's Disease and Parkinson's Disease". In: *Aging and disease* 11.3 (2020), p. 618. ISSN: 2152-5250. DOI: 10.14336/ad.2019.0617. URL: http://dx.doi.org/10.14336/AD.2019.0617 (cit. on p. 19).

[CSHB18] A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Mar. 2018). DOI: 10.1109/wacv.2018.00097. URL: http://dx.doi.org/10.1109/WACV.2018.00097 (cit. on pp. 15, 20–23, 39, 41–44, 46).

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*. 2009 (cit. on p. 26).

[Gal19] F. Gallard. *AC-PC line*. 2019. URL: https://radiopaedia.org/play/15508/entry/253440/case/50730/studies/56187 (cit. on p. 31).

[HEKK19] S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim. "A Benchmark for Interpretability Methods in Deep Neural Networks". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett. Curran Associates, Inc., 2019, pp. 9737–9748. URL: http://papers.nips.cc/paper/9167-a-benchmark-for-interpretability-methods-in-deep-neural-networks.pdf (cit. on p. 26).

[HZRS15] K. He, X. Zhang, S. Ren, J. Sun. "Deep Residual Learning for Image Recognition". In: (2015). arXiv: 1512.03385 [cs.CV] (cit. on p. 26).

[IHM+16] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, K. Keutzer. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size". In: (2016). arXiv: 1602.07360 [cs.CV] (cit. on p. 37).

[JD20] F. R. Jumah1, R. H. Dossani. "Neuroanatomy, Cingulate Cortex". In: (2020). URL: https://www.ncbi.nlm.nih.gov/books/NBK537077/ (cit. on p. 45).

[LBM+19] P. J. LaMontagne, T. L. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. Vlassenko, et al. "OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease". In: (Dec. 2019). DOI: 10.1101/2019.12.13.19014902. URL: http://dx.doi.org/10.1101/2019.12.13.19014902 (cit. on pp. 16, 27, 30, 59).

[Llo17] G. Lloyd-Jones. In: (2017). URL: https://www.radiologymasterclass.co.uk/tutorials/mri/mri_signal (cit. on p. 29).

[LWL+19] C. Liu, W. Wang, Z. Li, Y. Jiang, X. Han, J. Ha, W. Meng, M. He. "Biological Age Estimated from Retinal Imaging: A Novel Biomarker of Aging". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (2019), pp. 138–146. ISSN: 1611-3349. DOI: 10.1007/978-3-030-32239-7_16. URL: http://dx.doi.org/10.1007/978-3-030-32239-7_16 (cit. on pp. 13, 18, 19).

[MAP+15] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/ (cit. on p. 35).

[MCF20] T. N. Mundhenk, B. Y. Chen, G. Friedland. "Efficient Saliency Maps for Explainable AI". In: (2020). arXiv: 1911.11293 [cs.CV] (cit. on pp. 15, 24–26, 39, 41–43, 46).

[MFC+10] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, R. L. Buckner. "Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults". In: *Journal of Cognitive Neuroscience* 22.12 (Dec. 2010), pp. 2677–2684. ISSN: 1530-8898. DOI: 10.1162/jocn.2009.21407. URL: http://dx.doi.org/10.1162/jocn.2009.21407 (cit. on pp. 16, 27, 30).

[MGLV20] A. B. Márquez-Ruiz, L. González-Herrera, J. d. D. Luna, A. Valenzuela. "DNA methylation levels and telomere length in human teeth: usefulness for age estimation". In: *International Journal of Legal Medicine* 134.2 (Jan. 2020), pp. 451–459. ISSN: 1437-1596. DOI: 10.1007/s00414-019-02242-7. URL: http://dx.doi.org/10.1007/s00414-019-02242-7 (cit. on p. 13).

[MKSB16] T. N. Mundhenk, G. Konjevod, W. A. Sakla, K. Boakye. "A Large Contextual Dataset for Classification, Detection and Counting of Cars with Deep Learning". In: (2016). arXiv: 1609.04453 [cs.CV] (cit. on p. 26).

[MNB+09] R. L. McClelland, K. Nasir, M. Budoff, R. S. Blumenthal, R. A. Kronmal. "Arterial Age as a Function of Coronary Artery Calcium (from the Multi-Ethnic Study of Atherosclerosis [MESA])". In: *The American Journal of Cardiology* 103.1 (Jan. 2009), pp. 59–63. ISSN: 0002-9149. DOI: 10.1016/j.amjcard.2008.08.031. URL: http://dx.doi.org/10.1016/j.amjcard.2008.08.031 (cit. on p. 13).

[MPV09] J. K. Mai, G. Paxinos, T. Voss. *Atlas of the Human Brain*. 2009. ISBN: 9780123736031. URL: http://www.thehumanbrain.info/ (cit. on pp. 16, 44).

[MS92] I. of Medicine, N. A. of Sciences. *Discovering the Brain*. Ed. by S. Ackerman. Washington, DC: The National Academies Press, 1992. ISBN: 978-0-309-46799-5. DOI: 10.17226/1785. URL: https://www.nap.edu/catalog/1785/discovering-the-brain (cit. on p. 46).

[Mus19] J. Muschelli. *spm12r: Wrapper Functions for "SPM" (Statistical Parametric Mapping) Version 12 from the "Wellcome" Trust Centre for "Neuroimaging"*. 2019. URL: https://github.com/cran/spm12r (cit. on p. 30).

[MWP+07] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, R. L. Buckner. "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults". In: *Journal of Cognitive Neuroscience* 19.9 (Sept. 2007), pp. 1498–1507. ISSN: 1530-8898. DOI: 10.1162/jocn.2007.19.9.1498. URL: http://dx.doi.org/10.1162/jocn.2007.19.9.1498 (cit. on pp. 16, 27, 30, 59).

[OSCW19] D. Omeiza, S. Speakman, C. Cintas, K. Weldermariam. "Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models". In: (2019). arXiv: 1908.01224 [cs.CV] (cit. on pp. 23, 44).

[PDM+12] S. P. Poulin, R. Dautoff, J. C. Morris, L. F. Barrett, B. C. D. on behalf of the Alzheimer's Disease Neuroimaging Initiative. "Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity". In: (2012). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3185127/ (cit. on p. 45).

[RA19] S. A. Rahman, D. A. Adjeroh. "Deep Learning using Convolutional LSTM estimates Biological Age from Physical Activity". In: *Scientific Reports* 9.1 (Aug. 2019). ISSN: 2045-2322. DOI: 10.1038/s41598-019-46850-0. URL: http://dx.doi.org/10.1038/s41598-019-46850-0 (cit. on pp. 17, 18).

[SCD+19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: http://dx.doi.org/10.1007/s11263-019-01228-7 (cit. on pp. 15, 20, 22).

[SCW+15] X. SHI, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, W.-c. WOO. "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett. Curran Associates, Inc., 2015, pp. 802–810. URL: http://papers.nips.cc/paper/5955-convolutional-lstm-network-a-machine-learning-approach-for-precipitation-nowcasting.pdf (cit. on p. 18).

[SDBR15] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller. "Striving for Simplicity: The All Convolutional Net". In: (2015). arXiv: 1412.6806 [cs.LG] (cit. on pp. 15, 23).

[SHR17] S. E. Setti, H. C. Hunsberger, M. N. Reed. "Alterations in Hippocampal Activity and Alzheimer's Disease". In: (2017). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5972837/ (cit. on p. 45).

[ŠKPU17] D. Štern, P. Kainz, C. Payer, M. Urschler. "Multi-factorial Age Estimation from Skeletal and Dental MRI Volumes". In: *Machine Learning in Medical Imaging*. Ed. by Q. Wang, Y. Shi, H.-I. Suk, K. Suzuki. Cham: Springer International Publishing, 2017, pp. 61–69. ISBN: 978-3-319-67389-9 (cit. on pp. 13, 19).

[SLJ+14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. "Going Deeper with Convolutions". In: (2014). arXiv: 1409.4842 [cs.CV] (cit. on p. 37).

[Soc17] U. Social Cognitive Affective Neuroscience. *spm12-dartel*. 2017. URL: https://github.com/scanUCLA/spm12-dartel (cit. on p. 34).

[ŠPU19] D. Štern, C. Payer, M. Urschler. "Automated age estimation from MRI volumes of the hand". In: *Medical Image Analysis* 58 (Dec. 2019), p. 101538. ISSN: 1361-8415. DOI: 10.1016/j.media.2019.101538. URL: http://dx.doi.org/10.1016/j.media.2019.101538 (cit. on p. 19).

[SRN+11] J. Suh, A. J. Rivest, T. Nakashiba, T. Tominaga, S. Tonegawa. "Entorhinal Cortex Layer III Input to the Hippocampus Is Crucial for Temporal Association Memory". In: (2011). ISSN: 1095-9203. URL: https://doi.org/10.1126/science.1210125 (cit. on p. 45).

[sta14] B. staff. *Medical gallery of Blausen Medical 2014*. 2014. URL: 10.15347/WJM/2014.010 (cit. on p. 46).

[STK+17] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg. "SmoothGrad: removing noise by adding noise". In: (2017). arXiv: 1706.03825 [cs.LG] (cit. on p. 23).

[STY17] M. Sundararajan, A. Taly, Q. Yan. "Axiomatic Attribution for Deep Networks". In: (2017). arXiv: 1703.01365 [cs.LG] (cit. on p. 15).

[SZ15] K. Simonyan, A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: (2015). arXiv: 1409.1556 [cs.CV] (cit. on p. 18).

[ZF13] M. D. Zeiler, R. Fergus. "Visualizing and Understanding Convolutional Networks". In: (2013). arXiv: 1311.2901 [cs.CV] (cit. on pp. 15, 23).

[ZKL+16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba. "Learning Deep Features for Discriminative Localization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (cit. on pp. 15, 20, 22, 39).

[ZLX+14] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva. "Learning Deep Features for Scene Recognition using Places Database". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 487–495. URL: http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database.pdf (cit. on p. 26).

A Appendix

A.1 OASIS dataset metadata [MWP+07], [LBM+19]

Scanner:
Magnetic strength = 1.5 Tesla (1.5T Scanner)
Magnetic strength = 3 Tesla (3T Scanner)

Echo Time (TE):
OASIS 1: 4 ms
OASIS 3: 4 ms (1.5T scanner) / 3.16 ms (3T scanner)

Repetition Time (TR):
OASIS 1: 9.7 ms
OASIS 3: 9.7 ms (1.5T scanner) / 2.4 ms (3T scanner)

Flip Angle:
OASIS 1: 10
OASIS 3: 10 (1.5T scanner) / 8 (3T scanner)

Age Range (in years):
Youngest subject: 18
Oldest subject : 97.11

A.2 Visual comparison: SMOE with GradCAM++ vs. SMOE with SmoothGradCAM++

Figure A.1 shows that the visualization maps obtained by combining SMOE with SmoothGradCAM++ and GradCAM++ appear visually similar. However, SmoothGradCAM++ is much slower as explained in section Section 2.4 because we compute the gradients for n samples in the neighborhood of each input.

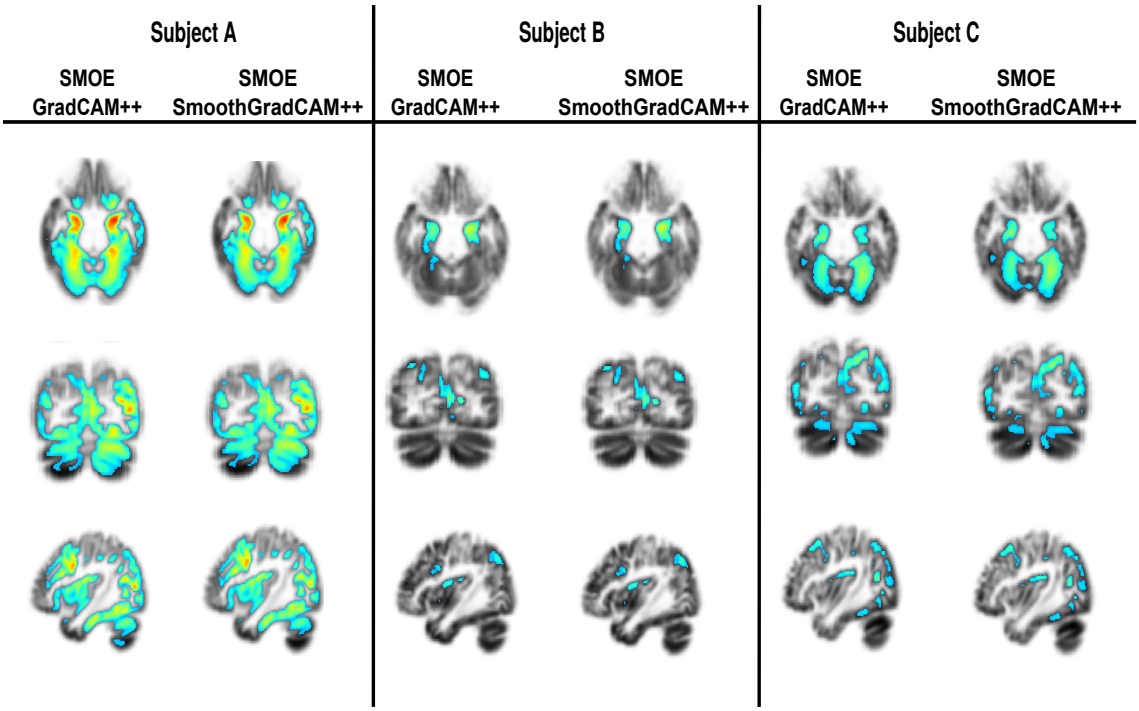


Figure A.1: Comparison of visualizations on combining SMOE with GradCAM++ and SmoothGradCAM++ (smoothing parameters: N = 5 samples, $\sigma = 0.15$) for three diffferent subjects visualized across all three planes for chunk 6.

A.3 Iterative framework configuration [AAS+20]

Learning rate: 0.0000062

Optimizer: Adam

Number of epochs: 60

Early stopping: If MAE does not drop by at least 0.005 over 10 epochs.

Number of iterations: 17

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature