Pázmány Péter Katolikus Egyetem
Információs Technológiai és Bionikai Kar

DIPLOMAMUNKA

Brain age prediction based on convolutional neural networks trained on T1 weighted MRI volumes

Kemenczky Péter
Computer Science Engineering MSc

2019

Témavezetők: Vakli Pál, PhD
Horváth András, PhD

Nyilatkozat

Alulírott Kemenczky Péter, a Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Karának hallgatója kijelentem, hogy ezt a szakdolgozatot meg nem engedett segítség nélkül, saját magam készítettem, és a szakdolgozatban csak a megadott forrásokat használtam fel. Minden olyan részt, melyet szó szerint, vagy azonos értelemben, de átfogalmazva más forrásból átvettem, egyértelműen a forrás megadásával megjelöltem. Ezt a Szakdolgozatot más szakon még nem nyújtottam be.

Kemenczky Péter

Abstract

The human brain changes across the whole lifespan, with considerable individual differences in the ageing processes and the associated risk for neurodegenerative diseases. With deep learning methods, especially with deep convolutional networks we can estimate the chronological age of healthy people with high precision (i.e. from 3 to 4 years average absolute difference) from their anatomical brain magnetic resonance imaging (MRI) volumes. According to recent findings, the brain-PAD score (the deviation of the brain-predicted age from chronological age) is a suitable measure of the aforementioned age-related alterations and disease risk and with this method the age-related structural effects of diseases could be investigated in the human brain. Up to now, migraine is not examined with this method however its impact in brain morphology is proven therefore the experiment is justified.

In this thesis I describe a convolutional neural network that is trained to predict the brain age based on T1 weighted MRI volumes. I then use this network to test the hypothesis that migraine is associated with accelerated or decelerated structural deterioration of brain structure as reflected in increased brain age. The neural network was trained on public datasets and it was retrained and fine-tuned for scanner and imaging sequence specific predictions with a dedicated dataset. Besides the MRI volumes of healthy subjects, the test data contains MRI records of migraine sufferers to test the hypothesis whether the migraine disorder can cause accelerated or decelerated brain ageing. I tested if the average brain-PAD score of the migraine test group is significantly higher than that of the healthy group. Whether it is higher for migraineurs that means their structural brain ageing has accelerated compared to healthy people. In the thesis I present a regression activation visualization technique for convolutional neural networks with which the regions of the input could be highlighted that play significant role in brain age prediction.

The deep convolutional neural network introduced in the thesis predicts brain age with an absolute error of 3.627 years on the validation set after retraining, while on the healthy test set the mean absolute error was 3.87 years. On the migraine test group the mean

absolute error was 4.06 years. T-test did not show significant difference between the means of the brain-PAD distributions of the two groups, therefore the null hypothesis had to be rejected, that is, the structural brain ageing of migraineurs and healthy people does not differ significantly. Contrary to the aforementioned results the activation visualization technique found differences between the two groups located in the temporal region on the left side of the brain, however the intensity difference is negligible to the activation intensities.

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani Vakli Pálnak, akinek szakmai támogatása óriási segítséget jelentett a dolgozat elkészítésében és a Természettudományi Kutatóközpontban (TTK) végzett munkához.
Köszönöm Vidnyánszky Zoltánnak, a TTK Agyi Képalkotó Központ vezetőjének, hogy biztosította az infrastruktúrát és lehetőséget adott, hogy az Agyi Szerkezet és Dinamika Kutatócsoport (ASzDK) tagjaként végezhettem a szakmai munkámat.
Köszönöm Horváth Andrásnak, hogy a teljes képzés alatt segített projektjeim elkészítésében, és köszönök minden szakmai segítséget az ASzDK korábbi és jelenleg is ott dolgozó tagjainak.

Contents

Chapter 1

Introduction

1.1. General Introduction

Machine learning (ML) and artificial intelligence (AI) are the parts of our everyday life and shaping our future. A vast part of technological and scientific advancements of the last years can be related to different machine learning methods. It can help to improve the user experience of average people on computer or on mobile phones (as Facebook's face recognition) or helps in shopping, however it is used in professional areas, e.g. in astronomy [1] AI can seek for specified astronomical events, in agriculture machine learning can be used for monitoring flowers [2], and also it can be used in economy, law or medical researches.

In clinical researches machine learning (especially deep learning) is used for image segmentation [3], tumor detection [4], drug research [5] and for many different purposes [6]. One of them is brain age prediction from brain MRI images. Cole et al. [7] use deep convolutional neural networks for predicting brain age from T1-weighted MRI data. Based on their article 'brain-predicted age' can be used as a biomarker of individual differences in the ageing process of the brain and its deviation from healthy data is associated with cognitive impairment and disease [8].

1.2. Desciption of the task

The human brain changes from birth until death. As all the tissues of the human body, the brain is also effected by ageing. Many cognitive functions, e.g. multitasking with deadlines [9], decision making or problem solving [10] are deteriorating with aging, however normal ageing (free from dementia) do not affect vocabulary and language knowledge [11]. The mentioned cognitive decline is caused by structural and chemical changes. The

structural changes are heterogeneous over the brain, however the average volume reduction in most brain areas is between 0.5% and 1% per year [12]. Beyond neural cell death, the shrinkage of neurons and reduction of synaptic spines are responsible for brain volume reduction.

In [12] authors summarize the results of cross sectional studies about the age-related morphometrical changes in brain. Based on the article, among the others, the amygdala, thalamus and diencephalic structures shrink linearly over the life, the cerebellum also shrinks linearly however the reduction is not significant, the striatum non-linearly related to age and for example the brain stem is not affected by ageing.

Structural changes can be noticed on T1 weighted MRI volumes and can be analysed by deep convolutional neural networks. In the recent years Cole et al. [7] [13] proved that these machine learning methods are good for predicting brain age with high precision. They used a number of T1 weighted brain MRI volumes of healthy individuals labelled with the participants' chronological age to train a convolutional neural network for age regression. To validate the model, they used tenfold cross-validation. Based on a sufficiently accurate model, the network was trained on the entire training data and the efficiency of the model was measured on a test set collected from new healthy participants'. The mean absolute error on the test data was 4.93 years [13].

As mentioned, the deviation of the brain-predicted age from the chronological age is associated with cognitive impairment and disease [13]. There are several diseases that cause age-related changes in the brain, such as Alzheimer's or Parkinson's disease [13], and there are many diseases that cause structural changes in brain however their relation to the ageing process is not clarified, yet. One of them is migraine that causes headaches regularly. The pain starts abruptly and affects half-side of the brain. It comes with light and sound sensitivity, and nausea. Sometimes so-called aura precedes the migraine pain, that is a subjective feeling accompanied by hallucinations and optical illusions [14]. It is proven that the effects of the migraine can affect many brain areas, e.g. the thalamus [15] which is shrinking volumetrically.

The goal of the project is to test the hypothesis that the migraine disorder can cause accelerated or decelerated structural deterioration in brain structure as reflected in increased brain age. The reference or control point is the healthy brain structure for a given age. To test this, a deep convolutional neural network was trained for healthy brain age prediction from T1 weighted MRI volumes collected from publicly available datasets. For more accurate brain age prediction the network was fine-tuned for scanner and sequence

specific regression and then it was tested on a dedicated test set with the brain scans of healthy individuals and migraineurs. Significant difference between the mean value of the brain-PAD scores of the two groups means that the speed of the structural brain ageing differs between migraineurs and healthy people.

1.2.1 The justification of the project

The explosive growth of the neural network methods in the last decade helped largely the brain research. Many functional, structural and clinical discovery was made however many questions about the brain functions, the functional relations, the effects of diseases are unanswered, yet. Machine learning and in this case brain age prediction helps to understand the effects of several diseases. In [13] Cole et al. compares the brain-PAD scores of diseases such as Alzheimer's disease, HIV or Schizophrenia however migraine is not examined with this method, yet. Many publications deal with the structural and functional effects of migraine however none of them deals with its possible acceleration effect on brain-ageing. Because migraine is one of the most prevalent illnesses in the world and more than one tenth of the population is affected by the disease it is justified to examine migraine from this side.

1.3. Details of the project

Based on the proposal I had to review the literature of deep convolutional networks and connected to it I had to examine the possible training methods including transfer learning. The project required to review the basic theory of magnetic resonance imaging. For training the convolutional network I had to collect publicly available T1 weighted MRI dataset from healthy people. The project required to study publications about brain age prediction.

Based on the related literature I had to implement a deep convolutional network to predict the age of the subjects from their T1 weighted brain MRI volumes and connected to migraine research I had to compare the brain-predicted age between migraine patients and healthy controls. Last but not least, I had to visualize the regions of the brain that the network used for brain age prediction.

Detailed task analysis Like all machine leaning methods brain age prediction also needs a vast amount of training data. The necessary condition of good training data is the good health of the subject and at least as good data quality as the test set. To

find publicly available T1 weighted MRI datasets with good enough quality that consist of records from healthy subjects is a difficult task. Magnetic resonance imaging is time consuming and expensive, therefore many datasets consist only of a few records. The larger published datasets were recorded with several MRI scanners and different imaging sequences. Larger data was mainly set out with the aim of researching a given disease, therefore a huge part of the sets is unusable because they are coming from patients with different diseases.

Considering the variance caused by the different scanners and imaging sequences, the optimization performance can be improved by applying transfer learning. With this method the neural network is trained with publicly available data that is different in quality and then fine-tuned by using data with similar characteristics (same MR scanner and recording sequence) as the test data.

The differences between the brain-predicted age of healthy and migraine patients can be compared with statistical methods. To test whether the brain of migraine patients is older than that of healthy people we can use a t-test to compare the brain-predicted age distribution between the two groups.

The activation visualization of the convolutional neural network could highlight the brain regions that are affected by ageing and the brain areas appearing on the activation map could be compared to the mostly age-affected brain areas examined in publications. Besides that, the heatmaps determined for migraine and healthy brain images could help to differentiate the brain areas that are affected by migraine and related to age.

1.4. Structure of the thesis

Brief summary of the structure of the thesis:

Chapter 2: Theoretical background In this chapter I review the theory and literature related to neural networks, magnetic resonance imaging, brain age prediction and migraine.

Chapter 3: Methods In chapter "Methods" I summarize and introduce the used data, the deep convolutional network used for prediction, and I provide further details regarding the training strategy and statistical methods.

Chapter 4: Results In chapter 4 I specify the results of the method and examine the output of activation visualization.

Chapter 5: Discussion In Discussion I summarize the results achieved and interpret them. I highlight the limitations of the method and I propose further research methods.

Chapter 2

Theoretical background

In this chapter I introduce the theoretical background necessary to understand brain age prediction based on neural networks, the theory of neural networks and magnetic resonance imaging.

Machine Learning (ML) includes specific algorithms and statistical models built for performing a specific problem without being explicitly programmed. We can distinguish supervised (task driven method with known input-output pairs), unsupervised (data driven method with known input and unknown output) and reinforcement learning (reward-based training).

As its name implies, a supervised machine learning model needs to be trained with a set of input-output data (*training set*) that represent the possible input-output data distribution well. Then the model is tested on a dedicated *test set* sampled from the given data space to measure accuracy the extent to which the model is able to generalize to novel samples that it has not seen during the training process. The efficiency of the model can be measured during training with *validation set* sampled from the training set. Hyperparameter tuning with the training set would lead to overfitting therefore validation data that is not used for optimization provides more better hyperparameter tuning.

Deep neural networks can be used for supervised learning and for many classification and regression problems deep neural networks are the best solutions.

2.1. Neural Networks

The idea behind artificial neural networks (ANN or NN) is based on the structure of the human nervous system and multipolar neurons. According to the basic principle,

the ANN is made up from small computational units, that connect to each other for solving complex problems without being programmed with task-specific rules. Properly constructed ANNs can learn a mapping between non-stochastic inputs and their outputs, thus the neural networks are universal approximators [16].

The small units that build up the neural networks are called artificial neurons (AN). The artificial neurons have one or more binary inputs and their output depends on the inputs, the learned weights and the activation function of the cell.

2.1.1 Artificial neuron

The artificial neuron gets N number of input values $(x_1, ..., x_N)$ from the input source or from the previous artificial neuron. An AN cell summarizes its weighted inputs and a bias value and applies an activation function $(f(.))$ on the calculated sum. If the $\mathbf{w}$ and $\mathbf{x}$ vectors are column vectors and contain the w_k and x_k values, respectively, with the bias value as w_0 and the corresponding $x_0 = 1$ then the formula (2.1) shows the mathematical representations of the artificial neuron.

$$y = f\left(\sum_{i=1}^{N} w_i x_i + b\right) = f\left(\sum_{i=0}^{N} w_i x_i\right) = f\left(\mathbf{w}^T \mathbf{x}\right) \tag{2.1}$$

The perceptron is one of the simplest ANN architectures. It is based on the Linear Threshold Unit (LTU), which is a modified artificial neuron that uses real numbers as inputs [17]. Perceptron is composed of a single LTU layer and each neuron connected to all the inputs. The perceptrons are often referred as the building units of ANNs for simplicity.

Activation function

The necessary characteristic of the activation function is to be continuously differentiable. The optimization algorithm uses gradient information to find local minima, therefore without this characteristic the optimization algorithms are not able to compute the gradient accurately. The ideal activation functions are nonlinear.
Initially, the idea behind the activation function was to compute whether an artificial neuron fires or not for a given input x, as in the case of biological neurons. The sigmoid or logistic function returns a value y between zero and one [18] [19] [20] and can be calculated by the eq.:(2.2). This function works similarly to the biological neuron; when the input is high enough, the output is one, the AN fires and for lower inputs the function

generates a smaller number. The upper left plot of figure 2.1 shows the output values generated with given input values.

The tanh is similar to the sigmoid function, however, it maps to the $[-1, 1]$ interval [21]. Its input-output relation can be seen on the lower left part of fig.2.1.

$$f(x) = \frac{1}{1 + \exp -x} = \frac{\exp x}{\exp x + 1} \tag{2.2}$$

Nowadays the most widely applied activation functions for neural networks are Rectified Linear Unit (ReLU) functions. For input x lower than 0 the function returns with 0 and for positive numbers $f(x) = x$ (eq.:(2.3)). The parametric version of ReLU is the same for positive inputs, while for negative input the response is the weighted value of the input. The weighting constant is called α and for $\alpha \leq 1$ eq.:(2.4) stands. In case of $\alpha = 0.01$ the function is called Leaky ReLU which is the most famous parametric ReLU function used in applications. The right side of figure 2.1 shows the ReLU and the parametric ReLU activation functions.

$$f(x) = max(0, x) \tag{2.3}$$

$$f(x) = max(x, \alpha x) \tag{2.4}$$

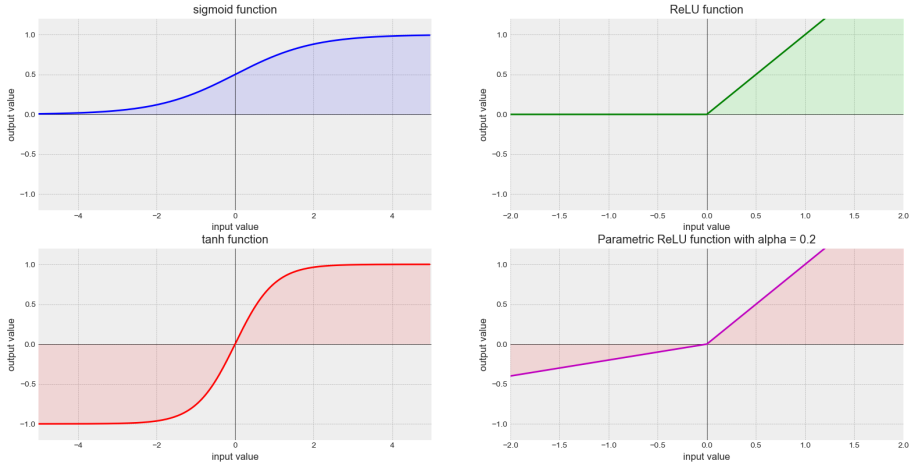


Figure 2.1: Four of the most common activation functions. The sigmoid (upper left), the tanh function (bottom left), the ReLU (upper right) and the Parametric ReLU (bottom right).

2.1.2 The layers of Neural Networks

The artificial neurons of an ANN are forming layers and getting their inputs from the previous layer or from the input data. The output of the ANs of a given layer is the input of the next layer or it is the response of the network, the output. Nowadays there are many different layer types like dense (or fully-connected), convolutional or long short-term memory (LSTM). In image processing and classification, convolutional neural networks (CNN or DCN from deep convolutional network) are the most famous and most useful network types. The CNNs contain convolutional and dense neuron layers [22] [23] and usually contain so-called pooling layers.

Convolutional layers

The analytical formula of the continuous convolution of f and g functions can be seen in eq.:(2.5) [24] [25].

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) \cdot g(t - \tau)\, d\tau \tag{2.5}$$

In case of discrete functions the eq.:(2.5) changes a little. The integration is a summation in the discrete space, therefore the formula is shown on (2.6).

$$(f * g)(n) = \sum_{m=-\infty}^{\infty} f(m) \cdot g(n - m) \tag{2.6}$$

Consider an image I as the f function. The pixels of I are arranged in a two dimensional grid. Assume, that we would like to convolve the image I with an other image, called "kernel" K, we have to change the formula of the discrete convolution to be true for two dimensional functions. If image I has $[m * n]$ pixels and kernel K has $[(2*p+1)*(2*r+1)]$ values, the convolution formula that describes the Y output image is shown on eq.(2.7) [22] and the figure 2.2 represents the operation.

$$I * K = \sum_{k=-p}^{p} \sum_{l=-r}^{r} K(k, l) \cdot I(x - k, y - l) = Y(x, y) \tag{2.7}$$

If the size parameters of the kernel p and r are greater than zero, the center point of the kernel $K(p + 1, r + 1)$ can run on pixels (x, y) of image $I(x, y)$ where $x \in [p, n - p]$ and $y \in [r, m - r]$, therefore we lose information about the edges of I. To avoid the information loss it is necessary to use border padding.

Padding is a function $pad()$ that is applied on I to expand its size. The size of $pad(I)(x, y)$ will be $[n + 2p, m + 2r]$ and it contains the original image I between the

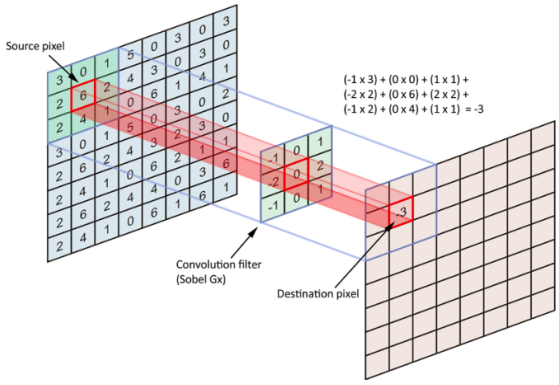


Figure 2.2: The 2D discrete convolution. Source: [26]

pixels $x \in [p, n+p]$ and $y \in [r, m+r]$. The edges of the $pad(I)$ image is depend on the applied rule that is usually zero padding in case of convolutional networks, i.e., zero values are used one the edges. (In Tensorflow machine learning library it is called same padding.) If we are not concerned about the loss of information we can use valid padding for which the $I(x, y)$ values are convolved where $x \in [p, n-p]$ and $y \in [r, m-r]$.

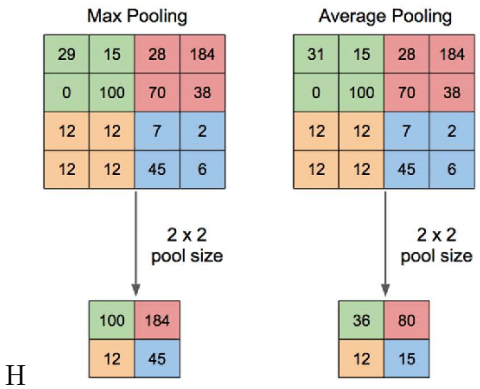
In neural networks the convolution kernels are artificial neuron groups that scan through the image to calculate a so-called feature map, that are the inputs of the next neuron layer. If a convolutional layer has k number of kernels that scans the input then the layer generates k different 2D feature maps. The k number gives the depth or the number of filters of the layer.

With step size adjustment the shape of the output could be manipulated. This function of the convolution layers are called stride size. Assume, that the layer input is a $[H_{in} * W_{in} * depth]$ size feature map and the strides are $[H_{stride}, W_{stride}]$ on the two dimensions. The kernel will step H_{stride} pixels on the first dimension and W_{stride} pixels on the second dimension in each iteration, therefore the size of the output can be calculated with eq.(2.8), where the image is expanded with $H_{padding}$ and $W_{padding}$ pixels on the edges and the convolution kernel is a $[H_{kernel} * W_{kernel}]$ size array.

The convolution formula can be easily modified to be apply to 3D tensors, like magnetic resonance volumes.

Pooling layers

The pooling layers are usually arranged after the convolution layers. This operation is usually used to reduce the spatial dimensions of the feature maps except their depth [27]. With convolutional layers, the dimension reduction is possible however the parameters of the layer have to be optimized, therefore they are expensive computationally. The



H

Figure 2.3: The Max pooling and the average pooling operation. source: [28]

dimension reduction of the pooling means fewer parameters that results in a gain in computational performance or reduces the resource requirements.

The pooling layer uses a pooling size that defines a window similar to the convolution kernel. This window scans the input of the layer and determines a metric for each location. There are different pooling operations like average-pooling when the layer computes the average of the values in the running window and sends it to the next layer or the max-pooling when the max elements in the window are forwarded. As in case of the convolutional layer, it is possible to adjust the stride to modify the step size of the moving window.

$$\begin{pmatrix} H_{out} \\ W_{out} \end{pmatrix} = \begin{pmatrix} floor\left(\frac{H_{in} - H_{kernel} + 2H_{padding}}{H_{stride}} + 1\right) \\ floor\left(\frac{W_{in} - W_{kernel} + 2W_{padding}}{W_{stride}} + 1\right) \end{pmatrix} \tag{2.8}$$

Dense layers

The dense or fully-connected layers are the simplest ANN layers. This layer type consist of ANs that connect to each unit of the previous layer. Basically, most of neural networks contain at least one hidden dense layer and an output dense layer at the end of the network. The dense layers can handle 1D vector inputs, therefore if the last layers of the convolution network are fully-connected layers it is necessary to flatten the last feature map.

The figure 2.4 shows a simple neural network consists of an input layer, three hidden dense layer and an output dense layer.

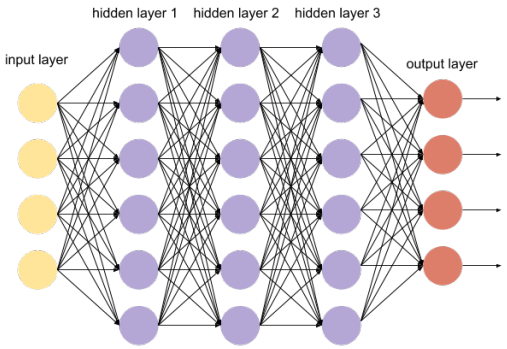


Figure 2.4: Neural network with three hidden dense layer and an output layer. Source: [29]

Convolutional Neural Network

The convolutional neural networks (CNNs) are invented in the 1999, however in 1982, their predecessor, the Neocognitron of Kunihiko Fukushima [30] have already used similar solutions. Y. LeCun and Y. Bengio published the LeNet-5 architecture in 1999 [31] that contains the same convolutional layers as DCNs nowadays. These structures are based on the visual cortex of mammals.

The main goal with these networks was to process images however they are used for several different machine learning tasks, as time series prediction [32] or regression from 3D images [7].

CNNs consist of convolutional layers, subsampling layers and fully connected layers. The convolutional layers can capture spatial dependencies. The subsampling layers are usually max-pooling layers, however several networks use convolutional layers to reduce the spatial dimension with stride adjustment. With consecutive convolutional and subsampling layers the spatial dimensions of the feature maps will be a fraction of the shape of the input image. With the dimension reduction, computational power can be saved therefore the depth of the convolutional layers can be expanded to recognise more features.

The fully connected layers after the last feature map do the regression and the classification using the recognised features.

The figure 2.5 shows a simple convolutional neural network with convolutional and dense layers.

We can handle the neural networks as black boxes, because it is difficult or impossible to interpret the performed transformations. The parameters of the model can be observed during and after the training however the number of parameters (usually in the order of

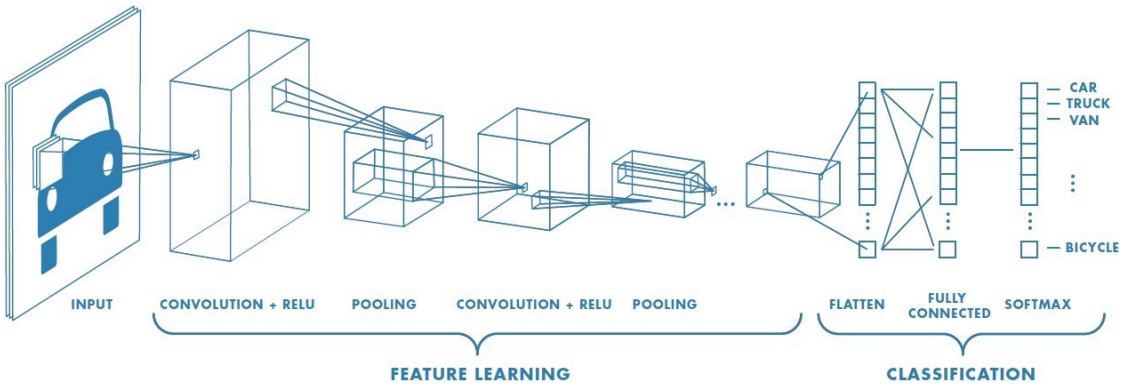


Figure 2.5: Neural network with several convolutional layers, one hidden and an output dense layer. Source: [33]

millions) makes it impossible to draw a conclusion about the precise function of the inner parts of the model. The eq.:(2.9) shows the simple connection between the input x and prediction $\hat{y}$ if the neural network is handled as a function and marked with $f_w(.)$.

$$\hat{y} = f_w(x) \tag{2.9}$$

2.1.3 Optimization

Let us consider a simple supervised learning problem with an unknown data distribution $P((x_{real}, y_{real}))$. Assume that data pairs $(x_{sampled}, y_{sampled})$ are sampled from the given data distribution. $x_i \in x_{sampled}$ values are the input values of the learning system (in our case a neural network) and we would like to predict $y_i \in y_{sampled}$ target values. Let us consider a cost function $\mathcal{L}(y, \hat{y})$ which measures the distance between y_i values and $\hat{y}_i = f_w(x_i)$ predictions. If the $P((x_{sampled}, y_{sampled}))$ represent the unknown $P((x_{real}, y_{real}))$ distribution well, the $\mathcal{L}(y_{sampled}, \hat{y}_{sampled})$ and $\mathcal{L}(y_{real}, \hat{y}_{real})$ will be close to each other [34].

If $\mathcal{F}$ is a set that contains the possible states of the neural network, we are seeking for the $f_w(.) \in \mathcal{F}$ parametrized by w weights where the $\mathcal{L}(y, f_w(x))$ is minimal [35]. In practice, we can only minimize the average loss function on the training dataset, thus we have to assume that the training data represents the the unknown data distribution $P((x_{real}, y_{real}))$ well.

The Gradient Descent algorithm If we have (x_i, y_i) training samples where $i \in [1, n]$, with the so-called gradient descent algorithm we can compute the weights in $t+1^{th}$ time step with the eq.:(2.10), where γ parameter is called learning rate. The method

changes the weights w_i slightly in the opposite direction as the loss gradient vector points.

$$w_{t+1} = w_t - \gamma \frac{1}{n} \sum_{i=1}^{n} \nabla_w \mathcal{L}(y_i, f_{w_t}(x_i)) \tag{2.10}$$

The Stochastic Gradient Descent Optimizer The stochastic gradient descent (SGD) algorithm is similar to the gradient descent algorithm but instead of iterating over the sampled data and computing the true gradient, the algorithm picks an example from the data randomly in each iteration and computes the stochastic gradient (eq.:(2.11)). The theory of the SGD assumes that after enough iteration the stochastic gradient and the true gradient will be close to each other [34] [35] [36], furthermore the computational cost is cheaper and simple.

$$w_{t+1} = w_t - \gamma \nabla_w \mathcal{L}(y_t, f_{w_t}(x_t)) \tag{2.11}$$

There exist several different improved versions of SGD algorithm. In [37] the authors introduce different algorithms which were published before 2017. The improved versions are usually based on adaptive learning rates. The optimizers find more optimal gradient directions, find minima faster or radically decrease the learning rate when the state is close to a local optima.

The RMSProp Optimizer Let us mark the computed gradient for each parameter w_i in time step t with $g_{t,i}$ (eq.:(2.12)).

$$g_{t,i} = \nabla_{w_i} \mathcal{L}(y_t, f_{w_{t,i}}(x_t)) \tag{2.12}$$

To adapt the learning rate we compute the squared running average of g and mark it with $E[g^2]_t$ (eq.:(2.13)). With using the running average the parameter update is made by the eq.:(2.14).

$$E[g^2]_{t+1} = 0.9E[g^2]_t + 0.1g_{t+1}^2 \tag{2.13}$$

$$w_{t+1} = w_t - \frac{\gamma}{\sqrt{E[g^2]_t + \epsilon}} g_t \tag{2.14}$$

The Adam Optimizer The ADAptive Moment optimizer (ADAM) stores the decaying average of previous gradients m_t (eq.:(2.15)), similar to the Momentum optimizer [37] and the decaying average of the squared previous gradients v_t (eq.:(2.16)), like RMSProp. To

compute the parameter update with eq.:(2.17) we use the bias-corrected moments $\hat{m}_t$ and $\hat{v}_t$. In [38] the default parameter values are $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$.

$$\begin{aligned} m_{t+1} &= \beta_1 m_t + (1 - \beta_1) g_{t+1} \\ \hat{m}_t &= \frac{m_t}{1 - \beta_1^t} \end{aligned} \tag{2.15}$$

$$\begin{aligned} v_{t+1} &= \beta_2 v_t + (1 - \beta_2) g_{t+1}^2 \\ \hat{v}_t &= \frac{v_t}{1 - \beta_2^t} \end{aligned} \tag{2.16}$$

$$w_{t+1} = w_t - \frac{\gamma}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \tag{2.17}$$

Exponential decay With learning rate decay the optimizer can dynamically set the value of the learning rate. More sophisticated optimizers can adapt the learning rate however manually adjusted learning rate can help the convergence. The exponential decay is decreasing the learning rate during training based on eq.:(2.18) where $\gamma(t)$ is the learning rate in the t^{th} time point, λ is the decay rate and τ is the value of decay steps.

$$\gamma(t) = \gamma(0) * \lambda^{t/\tau} \tag{2.18}$$

2.1.4 Loss function

In the subsection 2.1.3 the loss function $\mathcal{L}(.)$ had an important role to optimize a learning system, however, no concrete loss function was mentioned. There exist classification and regression supervised problems with corresponding classification and regression losses. The main point of the thesis—to predict the brain age of patients based on MR volumes—is a regression problem, therefore I deal with only regression loss functions. An important characteristic of the loss functions is its differentiability. Every gradient descent optimizer calculates the gradient of the loss function and thus requires a non-differentiable loss function.

Mean Squared Error

The mean squared error (MSE) or L2 loss is a statistical method that measures the average of the squared distances between two sample sets. In machine learning the mean squared error is an often-used regression error that computes the distance between the predicted data $\hat{y}$ and the original target data y (eq.:(2.19)) [39] [40].

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_i (y_i - \hat{y}_i)^2 \tag{2.19}$$

Mean Absolute Error

The mean absolute error (MAE) or L1 loss computes the absolute distance between $\hat{y}$ and y (eq.:(2.20)) [40].

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_i |y_i - \hat{y}_i| \tag{2.20}$$

Huber Loss

The Huber loss is a smoothed version of the MAE with adjustable steepness (eq.:(2.21)).

$$HUBER(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{if } |y - \hat{y}| \leq \delta \\ \delta|y - \hat{y}| - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases} \tag{2.21}$$

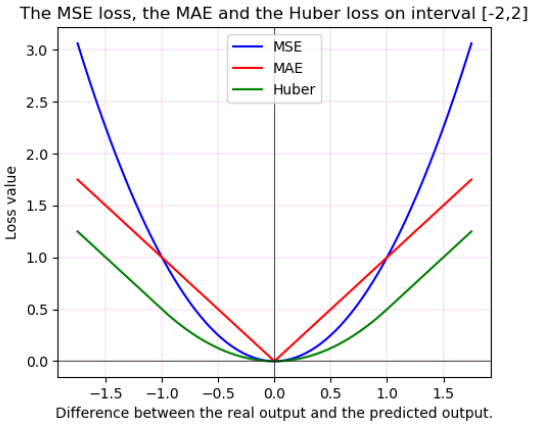


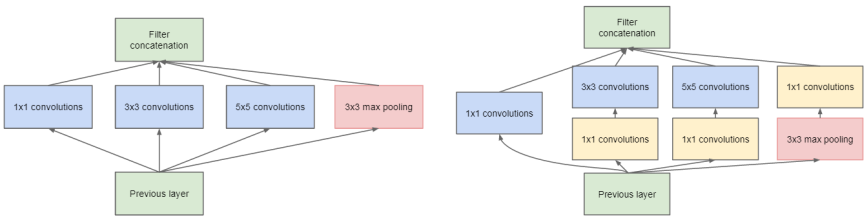
Figure 2.6: The mean squared error, the mean absolute error and the huber function

2.1.5 Inception modules

In 2014 a group of researchers introduced a new convolutional neural network architecture called GoogLeNet [41]. This structure contains so-called inception modules which are a group of convolutional layers with different characteristics. The modules are functioning as bottleneck layers, that is, they reduce the dimensionality by returning fewer feature maps than the input has. The structure help the layer to learn spatial characteristics easier, therefore the module can improve the efficiency of the network.

The layers of the module are aligned on the same level, get their inputs from the same source and their outputs are concatenated and delivered to the next layer or module. The idea behind the inception module is that we could not know which type of convolution layer set works better on a problem, thus we should let the network decide which layers are useful. With this state-of-art idea the network parameters increase drastically however the efficiency of the system will be better.

The naive version (left side of figure 2.7) of the inception modules of GoogLeNet are containing convolutional layers with $[1 \times 1]$, $[3 \times 3]$ and $[5 \times 5]$ kernel sizes. The fourth layer is a max pooling layer which has a $[3 \times 3]$ running window however the stride size is $[1 \times 1]$, therefore there are no dimension reduction.
The module working with dimension reduction (right side of figure 2.7) is uses $[1 \times 1]$ convolutional kernels before the $[3 \times 3]$ and $[5 \times 5]$ layers and after the max pooling layer to reduce the number of feature maps.



(a) Inception module, naïve version (b) Inception module with dimension reductions

Figure 2.7: The inception modules of GoogLeNet [41]. On the left the naive version of the module can be seen while on right the module with dimension reduction.

2.1.6 Transfer learning

To solve complex problems with artificial NN requires complex networks. The Vapnik–Chervonenkis theory [34] [36] defines a relation between the complexity of the neural network and the number of the required data. The Vapnik–Chervonenkis dimension (d_{VC}) is a measure of the capacity of the function space $\mathcal{F}$ that contains those $f_w(.)$ functions that can be learned by the network, therefore it is an indicator of the network complexity. According to the theory, the required number of data N to train a network with d_{VC} complexity is calculated by $N \approx 10000 d_{VC}$, however, in practice $N \approx 10 d_{VC}$ [34].

To compute d_{VC} for deep convolutional networks is a difficult task but we can see without calculation that the required data size is huge for a network that is complex enough to recognize and distinguish high resolution images.

To collect the required data for exact tasks like brain age prediction from T_1 weighted MR volumes is difficult or infeasible because the experiment is expensive and time-consuming. Transfer learning can be a solution for this problem.

Transfer learning is the method when a trained neural network is fine-tuned with a few samples for a target problem [42] [43]. In the first phase the neural network is trained on a dataset called source domain, therefore it learns the basic structural elements of the inputs. Then, in the second phase the model or a part of the model is retrained with the target dataset (target domain). The phrase "transfer" refers to the network that transfers the knowledge about the problem in its weights learned in the first phase to the target domain. The source and the target domain have to have similar characteristics for efficient transfer learning.

It is an important question that which layers could be used in the retrained network and which layers have to be retrained. In [44] the authors examine how the efficiency develops if they retrain a given number of layers from a pretrained network and retrain the other layers. They found that *fragile co-adapted features* appear after a few layers and if the first k layers are frozen and the remaining layers are retrained then the performance drops on the target domain. Important conclusion that transfer and fine-tuning on the whole network can improve the generalization.

The method could be effective when the target domain contains small number of samples or the available training run-time is short.

2.2. Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a non-invasive biological imaging technique. It generates three dimensional anatomical records about the target organ [45] [46] [47]. It is a popular method used for clinical diagnostics and research. During the measurement there are no ionizing radiation as in the case of Computer Tomography (CT), therefore it is safer however it takes more time to record volumes. The MR scanner is capable of taking structural and functional (fMRI) images. The image 2.8 shows a 3 Tesla MR scanner, the Siemens MAGNETOM Spectra.

During MR imaging the subject is placed into a strong homogeneous magnetic field B_0 that forces the protons of the target organ to align parallel with B_0. At equilibrium, the parallel net magnetization vector is called M_0 and it is equal to the longitudinal magnetization (M_Z) while the X and Y components of M are zero. With enough energy injected, if the frequency is equal to the energy difference between the spin states, it is

possible to saturate the spin system and make $M_Z = 0$. Later, the system will realign to the equilibrium state after the stimulation energy is turned off.

$$M_Z(t) = M_Z(0)e^{-t/T_1} + M_0(1 - e^{-t/T_1}) \tag{2.22}$$

With T_1 constant that is called spin lattice relaxation time the M_Z value can be computed in continuous time with the formula (2.22). Figure 2.9 shows the ratio of the M_Z component with respect to time. The T_1 times of different tissues are different.

The T_2 constant can show how the magnetic component perpendicular to B_0 disappears after a short time from eq.(2.23).

$$M_\perp(t) = M_\perp(0)e^{-t/T_2} \tag{2.23}$$

Figure 2.8: Siemens MAGNETOM Spectra. Source: [48]

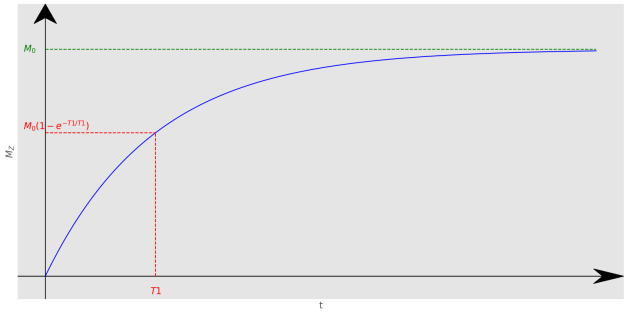


Figure 2.9: The behavior of M_Z magnetization vector of spins after electromagnetic stimulation perpendicular to M_Z with respect to time t.

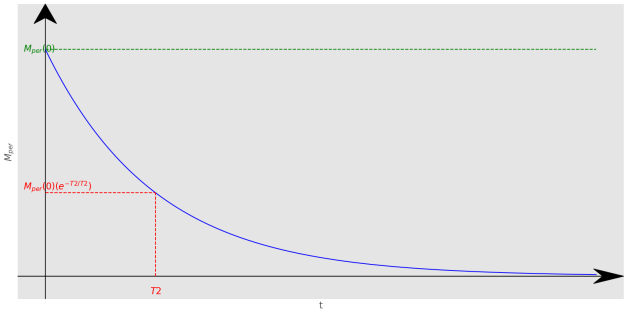
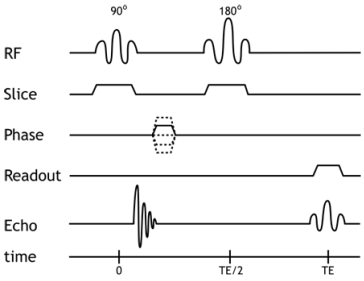


Figure 2.10: The magnetic components perpendicular to B_0 will disappear and the relaxation time is connected to T_2 relaxation constant.



The Free Induction Decay (FID) experiment can introduce the T_2^* relaxation time. In the FID experiment the spins are rotated from the z direction with 90° to the xy plane. T_2^* shows the decay after the radiofrequency (RF) pulse [47].

When applying a sequence of RF pulses can cause a phenomenon called spin echo, marked with T_E. After applying a 90° pulse and later a 180° pulse, an echo signal appears in T_E time after the first impulse and $T_E = 2 * \tau$ if τ is the time between the two pulses [47].

Applying several spin echo measurements with equal T_E times if the time between the echo signals is T_R (Repetition Time), the measured MR signal is given by the equation (2.24). The M_0 is the net magnetization proportional to the proton density in the different tissues.

$$S_{MR} = M_0 \left(1 - e^{-\frac{T_R}{T_1}} \right) e^{-\frac{T_E}{T_2}} \tag{2.24}$$

Choosing the T_E and T_R values differently the contrast of the recorded MR image will be different [47].

- If $T_R \gg T_1$ and $T_E \ll T_2$ then the relaxation time is negligible therefore S_{MR} will show the proton density (**PD measurement**).
- If $T_R \gg T_1$ and $T_E \approx T_2$ then the exponential that depends on the T_1 constant is approximately zero therefore it only depends on the T_2 time (**T2 weighted imaging**)
- If $T_R \approx T_1$ and $T_E \ll T_2$ then the exponential that depends on T_2 is approximately one therefore it only depends on T_1 time (**T1 weighted imaging**)

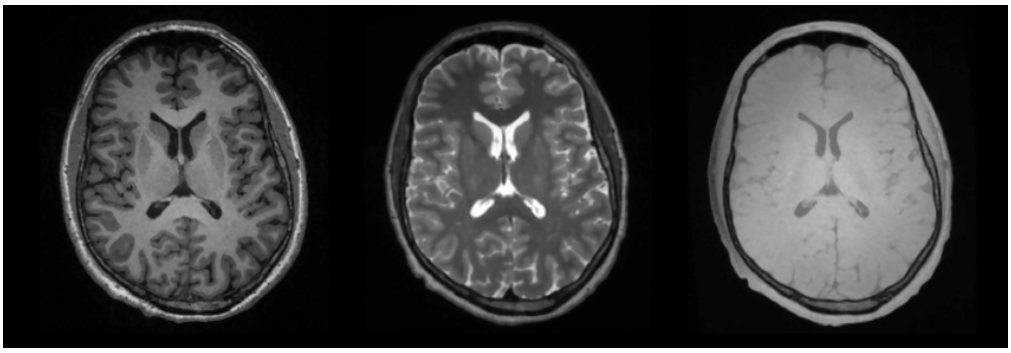


Figure 2.11: From left to right T_1, T_2 and PD weighted MR images. Source: ResearchGate: Axial slice acquired with T1, T2 and PD weighted MR imaging

The MNI transformation is image transformation that rotates the MRI volume into the MNI space. The MNI space defines the boundaries around the brain, expressed in millimeters, from a set origin [49]. The transformation is template dependent therefore the conversion could result different structures. The [50] guarantees a detailed review about the MNI templates. The figure 2.12 shows a raw, defaced MRI volume on the left and the MNI transformed version of the same volume on the right.

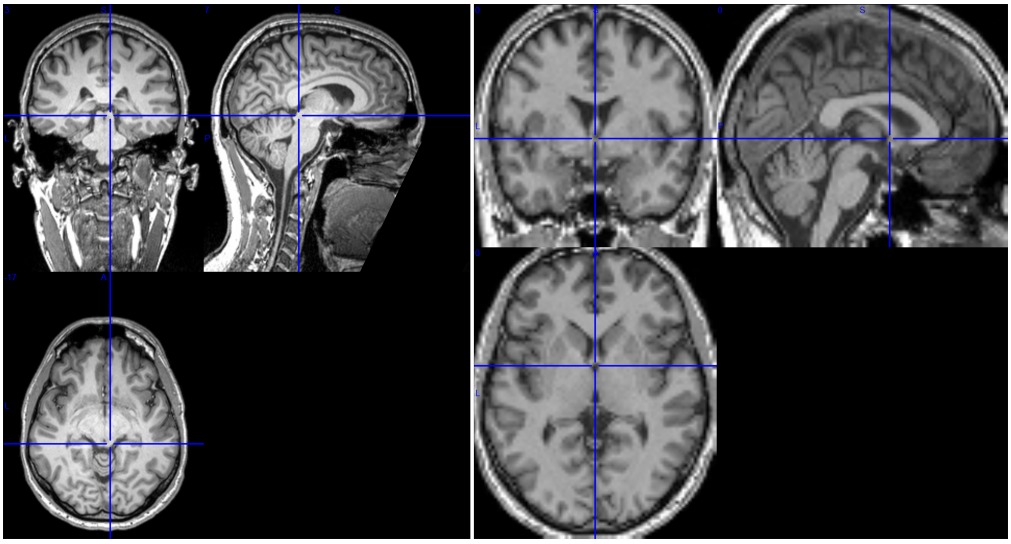


Figure 2.12: Defaced T1 weighted MRI volume on the left and MNI transformed image on right.

2.3. Deep networks for MR Imaging

The MR volumes are 3D tensor arrays that contain the measured S_{MR} values. With the generalization of the convolutional formula (subsection 2.1.2) these tensors can be used as inputs of a 3D convolutional neural network. The application fields of deep learning

methods in MR and fMRI researches is exponentially growing in every year. In [51] and [52] authors use 2D deep convolutional networks for brain slice segmentation while in [53] a 3D model is used for segmenting the whole brain. In [54] and in [55] 3D convolutional networks were used for predicting Alzheimer's disease, while in [7] the author, Cole et al. uses a 3D convolutional network to predict brain age from structural images.

2.4. Brain age prediction

The brain changes across the whole lifespan. These changes are related to cognitive performance. Older people are more prone to have neurodegenerative diseases, Alzheimer disease or dementia [56]. For seniors it is often harder to find the right words in a conversation, they would have problems with paying attention for more things at the same time [57]. Cognitive aging could also come with positive effects, for example learning new things could be easier [58].

With aging, several cognitive functions decline significantly. Executive cognitive functions, like multitasking with deadlines [9], decision making or problem solving [10] and working memory [59] are deteriorating with aging. However, vocabulary, language knowledge and verbal performance are not influenced significantly by healthy aging [11].

These changes do not happen at the same time for everybody. The speed of decline for a specific function is different and the degree of deterioration also varies [60]. The brain age prediction is based on the structural changes that cause the functional decline.

These differences can be discovered visually on MRI or CT (Computer Tomography) between a young brain and an old brain (fig. 2.13). During the development the sequence of progressive (e.g., cell growth, myelination) and regressive (e.g., synaptic pruning) processes cause the structural changes in the brain. Later, during aging the structural changes are caused by cell death and atrophy [13]. The volume of the grey matter (mainly the cortical areas) decreases steadily while the volume of the subcortical areas (white matter) develop until the midlife then they start to regress [61]. Different cellular and molecular alterations (e.g. neurite outgrowth or calcium signalling changes) contribute to the decline of cognitive functions and to the structural changes.

For a doctor it is easy to find the difference between the brain of the 90 years old patient and the brain of a 20 years old patient however to find the younger one from a 22 years old and a 27 years old brain is a hard task even for a neurobiologist. In [7] Cole uses convolutional neural network to predict the age of patients from their brain MRI volumes with average difference 4.16 years.

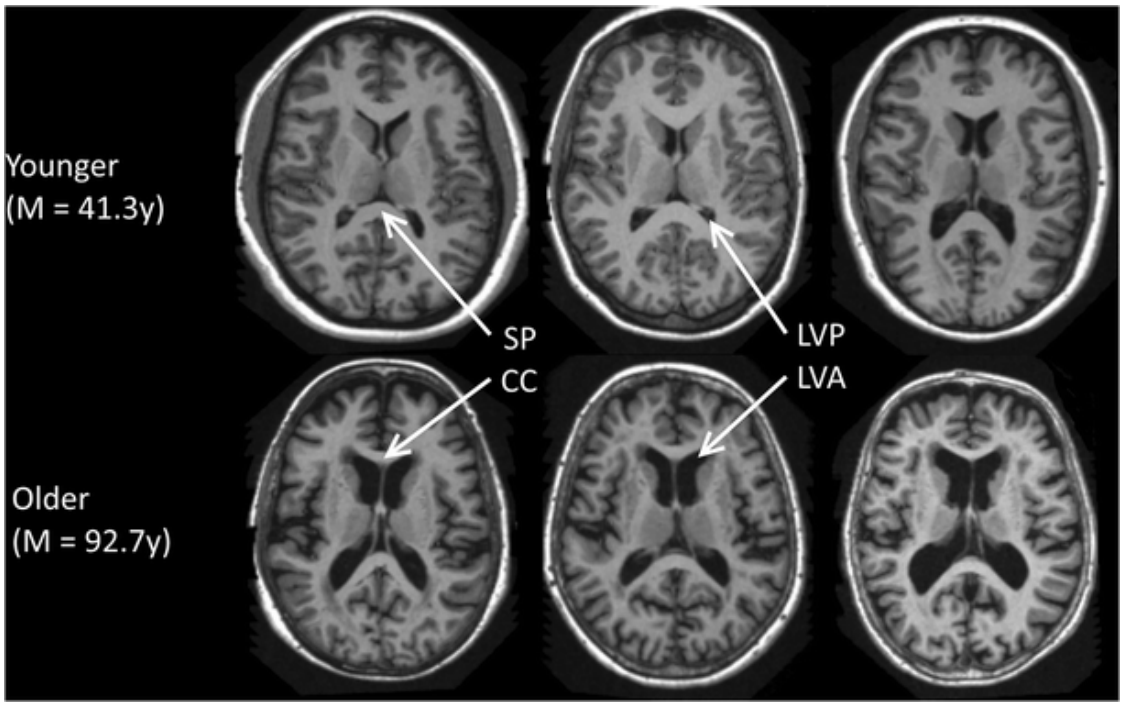


Figure 2.13: The T1 weighted MRI records about the brain of three young patients and three old patients. The images show larger lateral ventricles (LVP and LVA) for the older subjects. The tissue loss can be easily predictable. source: OASIS dataset, [62]

In [13] the author shows that an 'older'-appearing brain relates to advanced physiological and cognitive ageing and the risk of mortality. The brain age prediction can help to understand the age-associated structural changes in the brain and examine the effects of different diseases. Furthermore, it can help to identify the risk of cognitive diseases in clinical practice.

As mentioned before, brain age prediction is performed by convolutional neural network. With a number of training data that are T1 weighted brain MRI volumes the network is trained for predicting the chronological age of healthy individuals. A part of the training data is used as validation set (in the articles author uses tenfold cross-validation) that provides an unbiased evaluation for training therefore it helps to fine-tune the hyperparameters of the network for more accurate optimization and monitor the performance during training. The fitted model can be used for estimating the brain-predicted age on test samples. Brain-PAD metric is calculated by subtracting the chronological age from brain-predicted age (eq. (2.25)) and it could be used as a biomarker of an individual's brain health. Positive brain-PAD score implies older brain structure than the chronological age would justify and negative brain-PAD means younger brain. In [13] the author compares the average brain-PAD scores of different diseases, for example ma-

jor depression or Alzheimer's disease patients have highly positive brain-PAD score, and obesity also cause accelerated brain ageing while bipolar disorder patients have negative brain-PAD in average. Based on their review and the mentioned scores, brain-PAD correlates with several brain diseases and poor physical health. The method showed that having an older brain relates to advanced physiological and cognitive ageing and the risk of mortality.

$$Brain\_PAD = PredictedBrainAge - ChronologicalAge \tag{2.25}$$

2.5. Migraine and structural changes in the brain

Migraine is one of the most prevalent illnesses in the world [63]. More than 12% of the population suffers from it. Migraine can cause suddenly appearing pain, usually on one side of the head. It can be accompanied with light and sound sensitivity, nausea and vomiting. The pain can last for hours or for days [64]. For many sufferers (25% of patients) the pain could be preceded by aura that means hallucinations, optical illusions or, among the others, vision loss. Migraine usually occurs once or twice a month however many sufferers live with chronic daily migraine (more than 15 migraine days in a month).

Several diseases can cause structural changes in brain, e.g. Alzheimer's disease or Parkinson's disease however many publications are also dealing with the morphological changes in the brain caused by migraine [65] [66]. In [67] the authors summarized the results in migraine research until 2013. They used the database of PubMed to find articles where migraine is associated with white matter abnormalities (WMAs), infarct-like lesions (ILLs) or volumetric changes in gray (GM) and white matter (WM). They used 13 clinic-based and 6 population-based studies and in the most publications structural changes was found however they did not find any volumetric structural difference that correlates with age. In [15] Magon et al. examined the morphological abnormalities of thalamic subnuclei in migraine on MRI records. Based on the article the volumetric reduction in thalamus is significant. In [68] authors collected publications that examine the volumetric changes of the whole gray matter and they summarized the results. Among other brain areas they have consistently found volumetric changes in right claustrum, left cingulated gyrus or amygdala.

In the mentioned publications the authors all found structural changes in brain however they did not test or they did not find any age-related alteration. In 2015 Maleki et al. published their experiment where they have found age-related structural changes in

the brain. Their results indicated that in contrast to healthy subjects migraine patients show lack of thinning in the insula by age [69]. (The publication is based only on female patients.)

There are many pieces of evidence that the pain and functional effects of migraine cause lifelong impacts on sufferers. Brain age prediction could be an appropriate method to prove that the mentioned impacts of migraine accelerate (or decelerate) the ageing of the sufferers' brain.

2.6. Grad-CAM

Gradient-weighted Class Activation Mapping or Grad-CAM is a class-discriminative visualization method to highlight the regions of the target image that are significantly important in a classification task. Grad-CAM uses the class specific information in the last convolutional layer to produce a saliency map about the input images [70] [71]. To determine the saliency map, we have to compute the global average pooling from the gradient between the one-hot encoded output of the image and the last convolutional layer. With this step we get n values marked with w_k^c where $k \in [1, n]$ and n is the depth of the last convolutional layer (eq.:(2.26)) and the feature maps of the layer are marked by A^k [71].

$$w_k^c = \frac{1}{i * j} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{i,j}^k} \tag{2.26}$$

Multiplying these w_k^c values with the last feature maps generated by the network for the target input, adding them up and using the positive part of the sum gives back the saliency map $L_{i,j}^c$ (figure 2.14). $L_{i,j}^c$ is calculated with eq.(2.27) for a particular class c. $L_{i,j}^c$ values in a location (i, j) are directly related to a given class. The size of the map is dependent on the size of the feature map output of the final convolutional layer, therefore to fit it to the input image it has to be resized.

$$L_{i,j}^c = ReLU\left(\sum_k w_k^c A_{i,j}^k\right) \tag{2.27}$$

The Guided Backpropagation (on figure 2.14 the upper flow) is a pixel-space gradient visualization method. It can be generated by computing the gradient between the loss and the input. The product of the two maps (Guided Backporpagation and Grad-CAM) is the Guided Grad-CAM that shows not only the image parts that are important in decision but highlights the particular variance on the image.

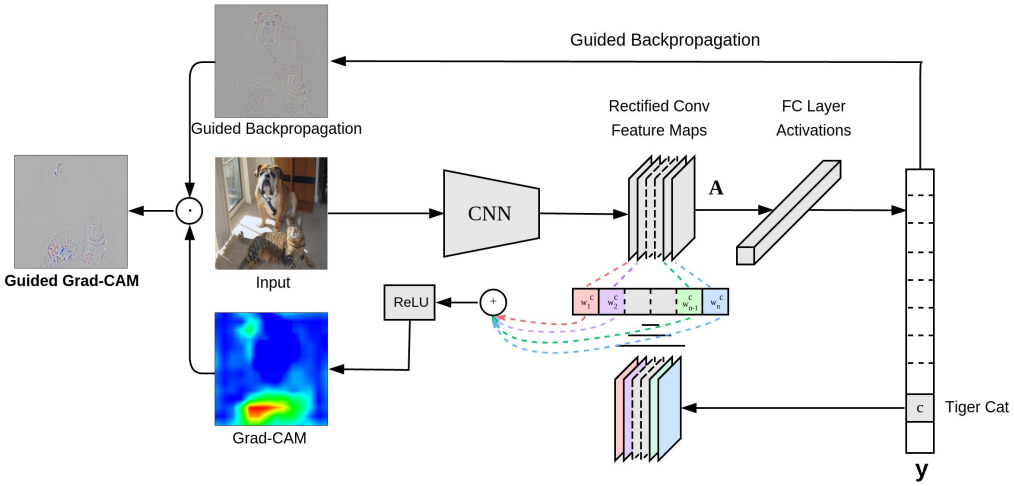


Figure 2.14: Gradient-weighted Class Activation Mapping (Grad-CAM) workflow based on [70].

Regression Activation Mapping (RAM) is a method based on CAM [72], which localizes the disciminative regions toward the prediction outcomes. In the article [72] the authors use a convolutional neural network without fully connected layer before the output layer because it makes it difficult to identify the important features. The last convolutional layer is followed by a GAP layer and it is connected to the output layer, therefore the number of the trainable parameters of the output layer is equal to the number of the feature maps of the last convolutional layer.

From the feature maps of the last convolutional layer $A^k_{i,j}$ the predicted output $\hat{y}$ is computed with the formula (2.28).

$$\hat{y} = \sum_k w_k \sum_i \sum_j A^k_{i,j} \tag{2.28}$$

The RAM maps can be computed with the formula (2.29). As in case of GradCAM the maps usually have smaller size than the size of the input therefore to use them as heatmaps they have to be upsampled.

$$L^{\hat{y}}_{i,j} = \sum_k w_k A^k(i,j) \tag{2.29}$$

Chapter 3

Methods

In this chapter I will introduce the dataset which was used to train the neural network, the implemented network itself, the training and evaluation strategy, and the statistical methods used for proving and testing the hypotheses.

3.1. Datasets

The dataset used for training and evaluation contains 1620 T1 weighted magnetic resonance brain images from different subjects labeled with gender and age information. The gender is coded in binary format, the 0 means female patient and the 1 means male patient. The age is a floating point number for each subject. Because the parts of the dataset come from different sources, the precision of the age is different. When the date of the assessment and the date of birth was available (in-house datasets), the age was determined and coded in double precision format. In case of open-source datasets I could only use the published ages coded in integers. The MR volumes are stored in 3D arrays and the size of the records is $79 * 95 * 79$.

3.1.1 Migraine dataset

According to our hypothesis the age of the brain of a migraine person is higher than that of a healthy person. To test this I had to train a neural network on T1 weighted MR volumes to predict the brain age of healthy people and compare brain-predicted age between a test group of migraine patients and healthy controls.

The migraine dataset (referred as migraine test data and healthy test data) was collected by the SE-NAP 2 Genetical Brain Imaging Migraine Research Group (SE-NAP 2 Genetikai Agyi Képalkotó Migrén Kutatócsoport) at the Research Centre for Natural

Sciences which contains 84 MR volumes from migraine patients and 69 volumes from control subjects (figure 3.1).

The MRI volumes were acquired using Siemens Magnetom Prisma scanner [73] 3T MRI scanner and 12-channel headcoils. The high-resolution T1-weighted anatomical images were acquired using a 3D magnetizationprepared rapid gradient echo (MPRAGE) sequence and 2-fold GRAPPA acceleration with a Partial Fourier factor of 7/8 ($1mm$ isotropic voxels; slice thickness/slice gap = $1/0.5mm$; TR = $2400ms$; TE = $3.06ms$; FA = $9o$; FOV = $256ms$).

70 patients of the migraine group are female and the other 14 subjects are male. In the control group the number of females is 42, while the number of males is 27.

The age distribution of the groups can be seen in figure 3.2. As the figure shows the number of the older control subjects is low, while the distribution of the age of the migraine group is more uniform. The youngest subject of the migraine group was 19 years old when the volume was recorded and the oldest migraine patient was 49. In the control group the youngest subject was 20 years old and the oldest was 48 years old.

The dataset was transformed into the MNI space that strictly defines the boundaries of the brain on the image [49] rotates the raw image to the appropriate degree and removes the image parts outside of the defined boundaries.

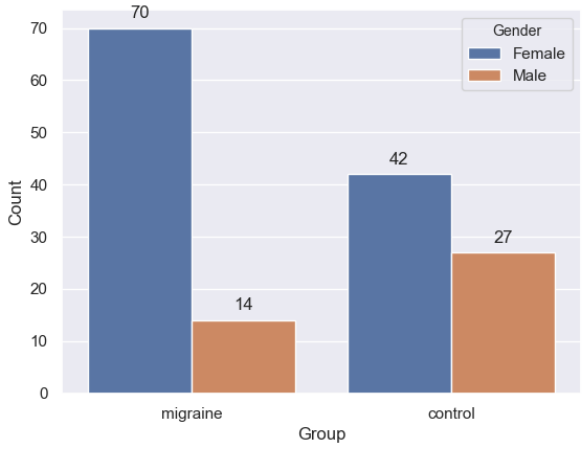


Figure 3.1: The figure shows the number of the migraine and control subjects grouped by gender.

3.1.2 Public dataset

The INDI (International Neuroimaging Data-sharing Initiative) [74] guarantees the possibility to access many published MR and fMRI datasets. The '1000 Functional Con-

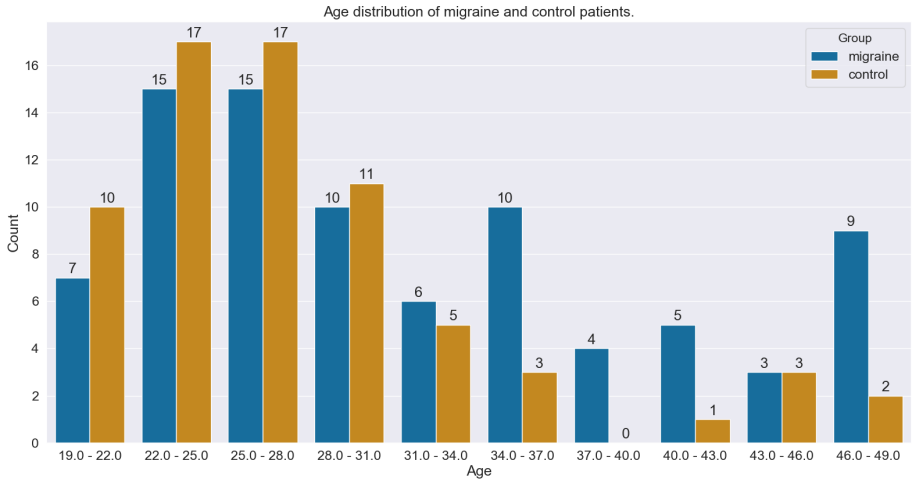


Figure 3.2: The figure shows the age (years) distribution of the migraine and control subjects.

nectomes' Project (FCP) [75] contains more than 1200 records from 33 different sources. The volumes of this project were part of the part of the public dataset. I could use 905 volumes from the project because many of the data were recorded with old MRI scanners and record was distorted (the thickness on a dimensions was the multiple of others) or their resolution was too small (above 2.5 mm slice thickness) to get relevant information thus I excluded these records from the dataset. (I transformed the volumes into MNI space. When the script was able to handle the resolution I leaved the data in the set, when it was not able to do it I excluded the volume.) The biggest part of this dataset is collected from young subjects, more precisely 60% of the data belongs to subjects between 16 and 23 years.

Another dataset that can be accessed from the INDI site is the SALD (Southwest University Adult Lifespan Dataset) [76]. This set contains 494 good resolution images (1 mm slice thickness) from which I could use 493. For brain age examination this set is one of the fundamental public datasets because it contains volumes from subjects from all age groups between 19 and 80 years.

The third part of the public set is a small (222 images), randomly selected part of the control (healthy) group of the ADNI (Alzheimer's Disease Neuroimaging Initiative) dataset [77]. With this data I wanted to mix some older brain examples into the public set because the age distribution of the two previously mentioned datasets show a salient peak around 20 years.

The final public dataset (referred later as public train data) contains 1620 T1 weighted

MR volumes with mean 39.17 years and 20.92 years standard deviation. 931 patients are females and 689 are males. Figure 3.3 shows volume count as a function of gender and data source.

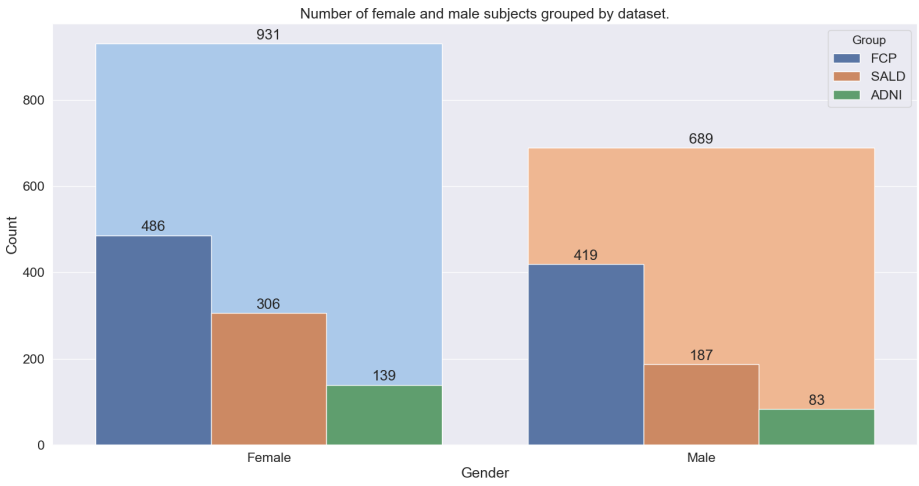


Figure 3.3: The distribution of the gender in the public dataset.

The age distribution of the dataset can be seen in figure 3.4. The plot shows that the dataset contains a high number of youngsters while above 30 years the distribution is more uniform.

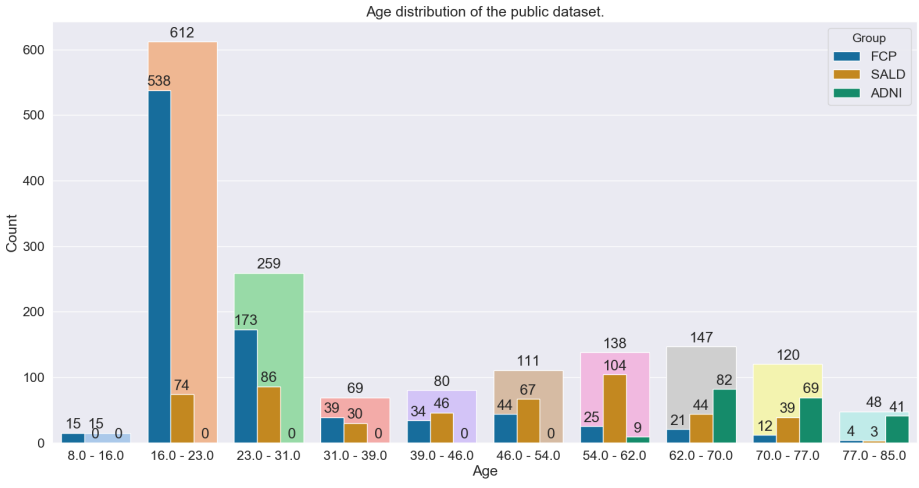


Figure 3.4: The figure shows the age distribution of the public data.

Last, but not least, figure 3.5 shows the age distribution of the public dataset grouped by gender. The mean and the standard deviation of the subject age is close to each other in both groups. The mean of the age of women is 38.82 years, for men this value is 39.63

years, the standard deviation is 20.79 years and 21.09 years for the female and male group.

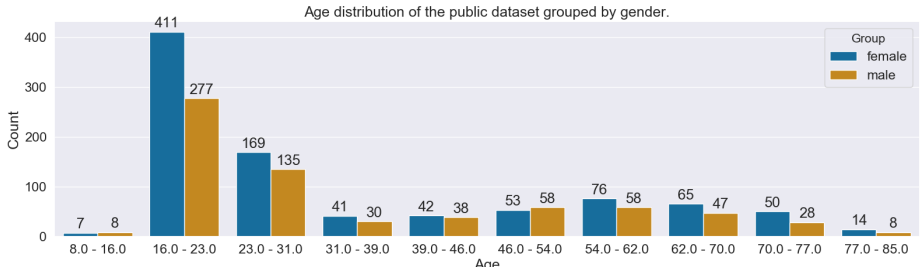


Figure 3.5: The figure shows the age distribution of the public data grouped by gender.

Because the public dataset is coming from several sources in which the brain volumes were recorded with different MR scanners and different recording sequences I separated the training phase into a pretrain and a retrain phase. In the pretrain phase I used public datasets to let the network know the general structure of the brain and in the retrain phase I used a different data collected with the same scanner and sequence (in-house data) as the migraine and control dataset to fine-tune the network for the detailed characteristics. In the section 3.3 I specify the details of the usage of the datasets during training and evaluation.

3.1.3 The in-house dataset

The in-house dataset of the Research Centre for Natural Sciences contains 157 T1 weighted MR volumes of healthy subjects, from which the number of females is 93 and the number of males is 64. The research center usually examines age-related changes in brain structure and function between the old and the young brain thus the in-house dataset contains the brain volumes of seniors and youngsters. Figure 3.6 shows the age distribution of the dataset.

In-house data were acquired with the same scanner and sequence settings as the migraine dataset. The specifications are detailed in the migraine dataset section (3.1.1). The data was also preprocessed with the same method as migraine dataset. It is referred as in-house data later.)

3.2. The network

For the implementation I used Python 3.6 programming language and the Spyder [78] development environment. Python uses scientific libraries for mathematical computa-

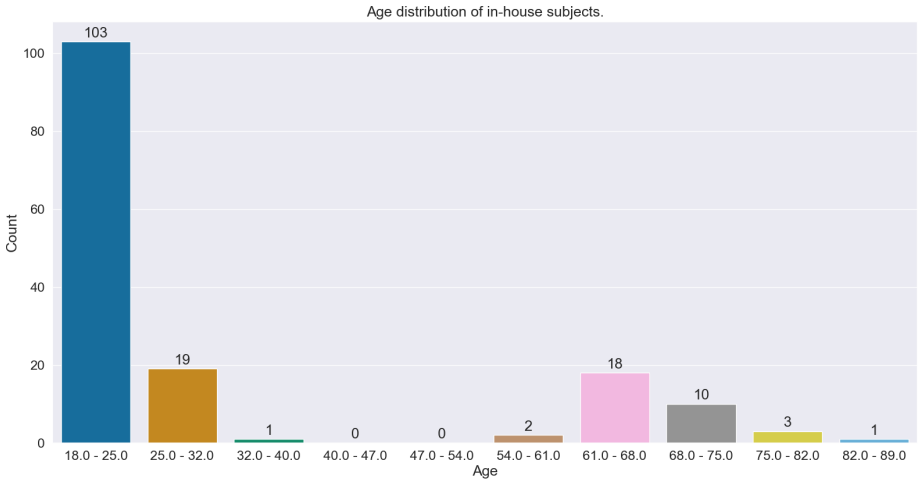


Figure 3.6: The figure shows the age distribution of the in-house data.

tion (numpy, scipy) and there are several machine learning library (TensorFlow, Keras, PyTorch, Theano or scikit-learn) that can guarantee the possibility to implement complex deep neural networks with GPU support. I used Tensorflow 1.13 [79] which is an open-source end-to-end ML library.

Tensorflow gives the possibility to implement network structures from the basic functions and classes of the library or handle the system with simplified structures as estimators [79] [80] (figure 3.7).

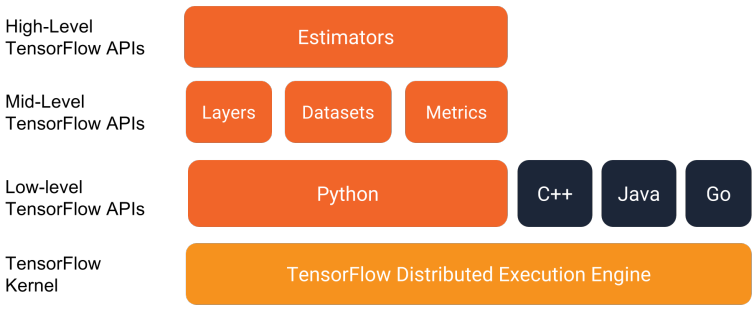


Figure 3.7: The API levels of the Tensorflow 1.13 with low level tensor definitions and high level estimators.

I implemented and tested many network structures from simple convolutional networks with a few layers to more complex ones. Each network used the 3D MRI volume and the gender code (0 or 1) as input and their response was the age as a single output value.

3.2.1 net1

The first implemented network was a simple CNN with four 3D convolutional layers and two fully connected (dense) layers. The convolution layers are followed by a max pooling layer with stride size $[2 \times 2 \times 2]$ and a dropout layer. The dropout layers are the input of the next convolutional layers. The value that defines the gender was concatenated to the flattened output of the last convolutional layer. The output of the last layer is one value, the predicted brain age of the input volume. net1 contains 13.879.681 trainable parameters. Table 3.8 shows the explicit characteristics of the layers that contain trainable parameters.

Name		Type	Shape	Number of parameters	Input layer
conv1	kernel	conv3D	[7x7x7x1x8]	2 744	Input image
	bias		[8]	8	
pool1		maxpool3D	[2x2x2]	0	conv1
do1		dropOut		0	pool1
conv2	kernel	conv3D	[5x5x5x8x64]	64 000	do1
	bias		[64]	64	
pool2		maxpool3D	[2x2x2]	0	conv2
do2		dropOut		0	pool2
conv3	kernel	conv3D	[3x3x3x64x256]	442 368	do2
	bias		[256]	256	
pool3		maxpool3D	[2x2x2]	0	conv3
do3		dropOut		0	pool3
conv4	kernel	conv3D	[3x3x3x256x512]	3 538 944	do3
	bias		[512]	512	
pool4		maxpool3D	[2x2x2]	0	conv4
do4		dropOut		0	pool4
flatten		Flatten		0	do4
concat1		concatenation		0	flatten, gender
dense1	kernel	Dense	[76801x128]	9 830 528	concat1
	bias		[128]	128	
dod1		dropOut		0	dense1
denseo	kernel	Dense	[128x1]	128	dod1
	bias		[1]	1	

Figure 3.8: The layer structure of net1. The gender value is concatenated to the flattened vector after the last convolutional layer. The total number of parameters is 13.879.681.

3.2.2 net2

The second network was based on the inception module theory detailed in the subsection 2.1.5. net2 is similar to net1 however the second convolution-pooling-dropout group was exchanged to an inception module inspired by the article [41] of Google. The inception module contains three convolutional layers with $[1 \times 1 \times 1]$, $[3 \times 3 \times 3]$ and $[5 \times 5 \times 5]$ kernels. Each of them has 32 filters. One dropout layer follows the convolutions. The fourth node is a deconvolutional layer with $[5 \times 5 \times 5]$ kernel, $[2 \times 2 \times 2]$ stride size and 16 filters. The max pooling layer restores the size of the feature map with $[2 \times 2 \times 2]$ pooling size (figure A.1). The last step is the concatenation along the axis of feature map depth.

net2 contains 83.015.473 trainable parameters, table 3.9 shows the details.

Name		Type	Shape	Number of parameters	Input layer
conv1 - pool1 - do1		conv3D + maxpool3D + dropOut	conv3D: [7x7x7x1x8]	2 752	Input image
inc-conv2_1	kernel	conv3D	[1x1x1x8x32]	256	do1
	bias		[32]	32	
do2_1		dropOut		0	inc-conv2_1
inc-conv2_2	kernel	conv3D	[3x3x3x8x32]	6 912	do1
	bias		[32]	32	
do2_2		dropOut		0	inc-conv2_2
inc-conv2_3	kernel	conv3D	[5x5x5x8x32]	32 000	do1
	bias		[32]	32	
do2_3		dropOut		0	inc-conv2_3
pool2_4		maxpool3D	[2x2x2]	0	do1
inc-deconv2_4	kernel	deconv3D	[5x5x5x16x8]	16 000	pool2_4
	bias		[16]	16	
concat2		concatenation		0	do2_1, do2_2, do2_3, inc-deconv2_4
conv3 - pool3 - do3		conv3D + maxpool3D + dropOut	conv3D: [3x3x3x112x256]	774 400	concat2
conv4 - pool4 - do4		conv3D + maxpool3D + dropOut	conv3D: [3x3x3x112x256]	3 539 456	do3
flatten		Flatten		0	do4
concat		concatenation		0	flatten, gender
denseo		Flatten + Dense + dropOut + Dense	Dense: [76801x128] Dense: [128x1]	9 830 785	concat

Figure 3.9: The layer structure of net2. The inception module is framed by dash-dotted lines. The layers used in net1 were grouped and marked in the table. The orange groups are the convolution, pooling and dropout groups and the dark green is the classification group, the two dense layers and one dropout layer. The total number of parameters is 83.015.473.

3.2.3 net3

Since the training and testing loss of the second network decreased compared to the first network I wanted to test a bigger structure, which is a larger version of net2 with more parameters that can utilize the resources of the drivers. The A.4 figure shows the graph structure of the final DCN, net3, that was trained for the task. The exact structure of the inception nodes can be seen in the Appendix (figures A.1, A.2 and A.3).

The number of the trainable parameters in the network is 95.342.105. Table 3.10 contains the detailed parameter description of the network for each layer.

net3 contains the first convolutional and inception module of net2, while the rest of the network is different. The inception2 is a custom module shown in figure A.2 that uses convolutions for shape reduction. The last convolutional part of the network contains two nodes. The first node uses max pooling layers for spatial dimension reduction while the other node uses convolutional layers with strides (figure A.3). The outputs of the nodes are concatenated, flattened and merged with the gender value. The last two dense layers are the same as the dense layers of net2.

Name		Type	Shape	Number of parameters	Input layer
conv1 - pool1 - do1		conv3D + maxpool3D + dropOut	conv3D: [7x7x7x1x8]	2 752	Input image
concat2		inc-conv2_X module with 7 layers		55 280	do1
conv_7_1	kernel	conv3D	[1x1x1x112x128]	14 336	concat2
	bias		[128]	128	
dropout_7_1		dropOut		0	conv_7_1
pool_8_1		maxpool3D	[2x2x2]	0	dropout_7_1
dropout_8_1		dropOut		0	pool_8_1
conv_7_2	kernel	conv3D	[1x1x1x112x128]	14 336	concat2
	bias		[128]	128	
dropout_7_2		dropOut		0	conv_7_2
conv_8_2	kernel	conv3D + dimension reduction [2x2x2]	[3x3x3x128x128]	442 368	dropout_7_2
	bias		[128]	128	
dropout_8_2		dropOut		0	conv_8_2
conv_7_3	kernel	conv3D	[1x1x1x112x128]	14 336	concat2
	bias		[128]	128	
dropout_7_3		dropOut		0	conv_7_3
conv_8_3	kernel	conv3D + dimension reduction [2x2x2]	[5x5x5x128x128]	2 048 000	dropout_7_3
	bias		[128]	128	
dropout_8_3		dropOut		0	conv_8_3
conv_78_4	kernel	conv3D + dimension reduction [2x2x2]	[3x3x3x112x64]	193 536	concat2
	bias		[64]	64	
concat3		concatenation		0	dropout_8_1, dropout_8_2, dropout_8_3, conv_78_4
batch_normalization		batchNormalization		20	concat3
conv_11	kernel	conv3D	[3x3x3x448x512]	6 193 152	batch_normalization
	bias		[512]	512	
dropout_12		dropOut		0	conv_11
pool_13		maxpool3D	[2x2x2]	0	dropout_12
conv_14	kernel	conv3D	[3x3x3x512x1024]	14 155 776	pool_13
	bias		[1024]	1 024	
dropout_15		dropOut		0	conv_14
pool_16		maxpool3D	[2x2x2]	0	dropout_15
conv_17_2	kernel	conv3D + dimension reduction [2x2x2]	[1x1x1x448x256]	114 688	batch_normalization
	bias		[256]	256	
dropout_18_2		dropOut		0	conv_17_2
conv_19_2	kernel	conv3D + dimension reduction [2x2x2]	[5x5x5x256x1024]	32 768 000	dropout_18_2
	bias		[1024]	1 024	
dropout_20_2		dropOut		0	conv_19_2
concat_21		concatenation		0	pool_16, dropout_20_2
flatten		Flatten		0	concat_21
concat		concatenation		0	flatten, gender
denseo		Flatten + Dense + dropOut + Dense	Dense: [76801x128] Dense: [128x1]	9 830 785	concat

Figure 3.10: The layer structure of net3. The inception modules are framed by dash-dotted line. The layers used in net2 was grouped and marked in the table. The orange groups are the convolution, pooling and dropout groups, the turquoise is the inception module of net2 and the dark green is the classification group, the two dense layers and one dropout layer. The total number of parameters is 95.342.105.

3.2.4 Activation functions

The last artificial neuron of the networks that returns with the predicted brain age uses linear activation that means the output is the same as the sum of the weighted input. The rest of the neurons uses Parametric ReLU activation. The alpha parameter of the function was set to 0.2.

3.3. Training and evaluation strategy

Only from the voxel intensities of an MR volume a trained neural network can conclude the age of the subject. In the recorded MR volumes the whole head can be seen above the neck, but I wanted to rule out the possibility that the network estimates the age from the face, from the bone density or from any tissue outside the brain. The research team of the Brain Imaging Center provided the opportunity to use their script which is able to transform the MR volumes into the MNI space. On the transformed volumes the cranial cavity can be seen, the script removes the non-brain parts from the image and rotates it into the appropriate angle. For training and testing I used the transformed volumes to exclude the mentioned problem.

3.3.1 Usage of datasets

In the section 3.1 I introduced three sets of data, the public dataset, the in-house dataset and the migraine dataset. I define three phases: pretrain, retrain and test. In the pretrain phase I used the public dataset to train the neural network. As I mentioned in section 3.1 the number of in-house dataset volumes is not enough to train a good brain age predictor, therefore I used this phase to map the structure of the brain and store it in the convolutional network. I chose net3 network to fine-tune based on the pretrain performances of the three networks. Later, in the retrain phase, I applied transfer learning to fine tune the network for the in-house dataset. Since the test dataset was collected with the same MR scanner and with similar recording settings as the retrain dataset, this phase can improve the efficiency of the network in predicting brain age on the test set.

In both train phases I randomly separated the datasets into training and validation sets. The training set contained the 85% of the whole set and the validation set was the remaining 15% of the data. In the testing phase the brain age of the patients of the migraine and control group was predicted by the trained CNN. The predictions of the two groups was compared to each other with different statistical methods detailed in the

subsection 3.3.3.

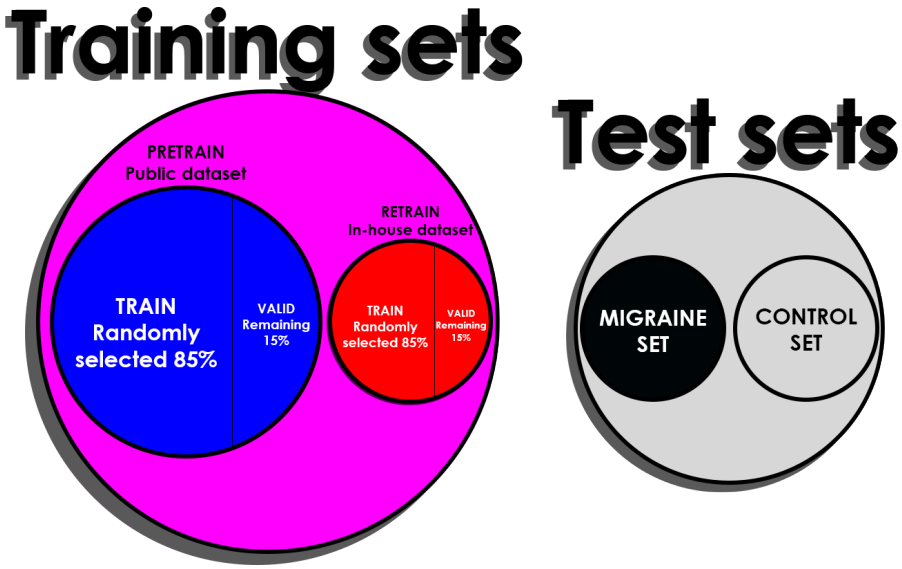


Figure 3.11: The sets of the training and testing phase.

3.3.2 Training parameters

The loss function that the networks had to minimize was the Mean Squared Error (subsection 2.1.4). In the pretrain phase the built-in Adam Optimizer [38] and Momentum Optimizer [81] of Tensorflow was used for optimization. The convergence time and the results of the Momentum Optimizer was better, thus I used the Momentum instead of Adam. The momentum parameter was set to 0.8 while the initial learning rate of the network was set to 1e−6 for each networks. For bigger initial learning rates the optimization was failed and for smaller learning rates the convergence time was high.
I used exponential decay for adjusting the learning rate detailed in subsection 2.1.4. From formula (2.18), the λ value was set to 0.98 and the τ value was set to 80000.
The number of training epochs in the pretrain phase was 50 in case of net1. Because the number of trainable parameters of net2 is significantly larger I set the epoch size of net2 to 60 and for the same reason the training epochs of net3 was 70. In the retrain phase the training epochs was 20.

The final network used in retrain phase was chosen based on the testing results after the pretrain phase. The test metric was the MSE loss and it was measured on the validation set of the public dataset.

3.3.3 Testing and statistical analysis

In the testing phase the migraine and the control test data was analysed by the trained neural network and for each MRI volume-gender input pairs the network has assigned an age prediction. Between the real age and the predicted age the mean absolute difference was calculated by the formula (2.20) where $\mathbf{y}$ contains the real age values and the $\hat{\mathbf{y}}$ contains the predicted ones. The mean absolute error of the test dataset is marked by $d_{a;M}$ for migraine set and $d_{a;C}$ for control patients. $d_{r;M}$ and $d_{r;C}$ are calculated by the formula (3.1) and it shows the bias in the neural network predictions, if d_r value is significantly differ from zero.

$$d_r = \frac{1}{n}\sum_i (\hat{y}_i - y_i) \tag{3.1}$$

The hypotheses assumes difference between the brain age of the migraine and the control group therefore the $\mathbf{d_{r;M}}$ and $\mathbf{d_{r;C}}$ (eq.(3.2)) were compared to each other with t-test. The null hypothesis of the two-sample t-test assumes that the mean of the two population is equal [82]. The necessary conditions of the t-test are the normally distributed sample sets and equal variances. Welch's t-test is an alternative version of t-test, which does not assume equal population variances [83] therefore can be used if the variances are different.

$$\mathbf{d_r} = \hat{\mathbf{y}} - \mathbf{y} \tag{3.2}$$

Normally distributed populations. A normal (or Gaussian or Gauss or Laplace–Gauss) distribution can be characterised by its mean value μ and standard deviation σ (or variance σ^2) and it is noted by $\mathcal{N}(\mu, \sigma^2)$ [84]. The probability density function of normal distribution can be formulized by (3.3).

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3.3}$$

In statistics the distribution of the data is unknown but in many cases the normality characteristic can be assumed. To test the assumption the Shapiro–Wilk normality test can be used. The null hypothesis of the test is that the $i \in [1, ..., n] : x_i$ samples come from a normally distributed population. The test statistic W is computed with the formula (3.4) [85] where $x_{(i)}$ is the i^{th} value of the ordered $\mathbf{x}$ vector and $\overline{x}$ is the mean of samples x_i. The $\mathbf{a}$ can be computed with eq.(3.5) if $\mathbf{V}$ is the covariance matrix of the order statistics of a sample of n standard normal random variables with expectation

vector **m**. If the p-value calculated from W is less than the chosen significance limit the null hypothesis have to be rejected, i.e. the samples are not from normal distribution.

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.4}$$

$$\mathbf{a} = (a_1, ..., a_n)^T = (m^T V^{-1} V^{-1} m)^{-1/2} m^T V^{-1} \tag{3.5}$$

The distributions also can be characterized by their skewness and kurtosis [86]. The skewness measures the asymmetry of the distribution. If it is 0 the distribution is symmetric and if it is much higher of lower than 0 it is asymmetric. The skewness of the random variable X can be computed with the formula (3.6). The kurtosis can be computed with equation (3.7) and as the skewness the kurtosis measures the shape of the distribution. It shows whether the distribution is heavy-tailed or light-tailed compared to a normal distribution. The higher the kurtosis value is, the heavier the tail.

$$Skewness[X] = E\left[\frac{(X-\mu)^3}{\sigma^3}\right] \tag{3.6}$$

$$Kurtosis[X] = E\left[\frac{(X-\mu)^4}{\sigma^4}\right] \tag{3.7}$$

Correlation analysis. The Pearson Correlation Coefficient (PCC) shows the strength of the linear relationship between the two random variables y and $\hat{y}$ [87]. The coefficient is defined by the equation (3.8), where $E(y, \hat{y})$ is the cross-correlation between the two variables, and σ_y and $\sigma_{\hat{y}}$ are the variances. For $P(y, \hat{y})$ is true that the value is $P(y, \hat{y}) \in [-1, 1]$ and if $P^2(y, \hat{y})$ is close to zero the linear correlation is weak and if it is close to one than the correlation is strong.

$$P(y, \hat{y}) = \frac{E(y, \hat{y})}{\sigma_y \sigma_{\hat{y}}} \tag{3.8}$$

Equal variances. Levene's test is a suitable solution for comparing the variances of two sample sets. Its null hypothesis is that the population variances are equal for the groups. Based upon the formula (3.9) if the variances are equal, Levene's test statistic is close to zero. In the eq.(3.9) N is the number of all samples from all groups, N_i is the sample number of group i, c is the number of groups, $Y_{i,j}$ is the j^{th} sample of i^{th} group and $\overline{Y_{i\cdot}}$ is the median of samples in group i. The null hypothesis, i.e. the variances are equal, have to be rejected if the p-value calculated from L is smaller than the chosen significance level.

$$L = \frac{N-c}{c-1}\left(\frac{\sum_{i=1}^{c} N_i \left(\frac{1}{N_i}\sum_{j=1}^{N_i}(|Y_{ij}-\overline{Y_{i\cdot}}|) - \frac{1}{N}\sum_{j=1}^{c}\sum_{k=1}^{N_j}(|Y_{jk}-\overline{Y_{j\cdot}}|)\right)^2}{\sum_{i=1}^{c}\sum_{j=1}^{N_i}\left(|Y_{ij}-\overline{Y_{i\cdot}}| - \frac{1}{N_i}\sum_{k=1}^{N_i}(|Y_{ik}-\overline{Y_{i\cdot}})|\right)^2}\right) \tag{3.9}$$

T-test. Whether it is assumed that two normally distributed population has the same mean value if the variances are equal it is possible to use t-test to check the hypotheses [82]. The formula to compute the t value of t-test can be seen on (3.10) if s_p^2 is computed by eq.(3.11) [88].

$$t = \frac{\mu_1 - \mu_2}{s_p^2\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}} \tag{3.10}$$

$$s_p^2 = \frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2} \tag{3.11}$$

The Welch-test is an alternative version of t-test for populations coming from distributions with unequal variances [83]. The test value t_W can be computed with eq.(3.12).

$$t_W = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}} \tag{3.12}$$

3.3.4 Activation Visualization

Regression Activation Mapping (RAM) detailed in chapter 2 was used as activation visualization technique. RAM uses the feature map of the last convolutional layer of the network and uses the weights of the only dense layer, the output layer. Unlike RAM, net3 uses two dense layers after the convolutions, therefore the method had to be changed.

In the new method I used the trained net3 and its weights to copy them into a new network that has the same convolutional structure as net3 however it has only one dense layer at the end. More precisely, the new network comprised all layers of net3 up to *concat_21* (3.10), a global average pooling layer and a new output dense layer. All weights inherited from net3 were frozen and only the output dense layer was optimized to predict brain-age.

As the new network was trained, the last feature maps and the weights of the dense layer were extracted for all test MRI volumes. The RAM feature maps were computed with the formula (2.29) and were averaged over the test set (separated for migraine and healthy groups). Based on the equation (2.29) RAM maps have the same size as the

last feature maps, that is, in case of net3, $[5 * 6 * 5]$. As table 3.10 shows the number of feature maps used for computing RAM is 2048. To visualize the significant parts of the input image for predicting brain-age, RAM maps have to be resized. In python the zoom function of 'scipy.ndimage' was used [89].

During training the convolutional networks learn what type of features should be extracted with respect to the inputs to make the most accurate predictions. The deeper the convolutional layer, the more complex the extracted features. Based on the previous statement if a deeper convolutional layer is used for computing RAM then it can visualize more complex features.

Whether I neglect the information loss and drop the last module of net3 (in figure A.3) the resolution of the computed heatmap could be much bigger. Because net3 lose two convolutional layers on both nodes it loses higher level feature knowledge. As in the case of the previous method I removed the rear part of net3, that is, the layers that was copied from net3 are the layers up to *batch_normalization* (3.10) and changed it to average pooling and output dense layer. Only the new output dense layer was optimized in this step. The size of the last feature map and heatmap is $[20 * 24 * 20]$ while the number of usable feature maps decreased to 448.

Feature scaling method was used for heatmap normalization, that is, the heatmap intensities were scaled to $[0, 1]$ range (eq.(3.13)).

$$\hat{y} = \frac{y - min(y)}{max(y) - min(y)} \tag{3.13}$$

3.3.5 Hardware and Software

The computer used for training comprised an Intel i7-6700 CPU, 32 GB DDR4 RAM and an NVIDIA Quadro M4000 GPU. With this devices the run-time of the training lasted for a few days. For net1 the run-time was about 8 hours, for net2 it was around a day and for net3 it was 2 days.

The scripts are available on github.
https://github.com/KemenczkyP/Age-estimation-from-MRI-anatomy

Chapter 4

Results

In this chapter I detail the results of the training and the statistical tests. The neural network used for fine-tuning was chosen based on the pretrain validation results. The mean squared error and the mean absolute error was monitored on the validation set during the whole pretrain process and after the pretrain phase. In the next section I describe the results of pretrain.

4.1. Pretrain

4.1.1 net1

The training procedure of net1 was detailed in subsection 3.3.2. In the pretrain phase the initial learning rate of the Momentum optimizer was set to 1e−6 with exponential decay. The training epochs was 50 on the public set. The momentum parameter of the optimizer was set to 0.8. The figure 4.1 shows the epochs-loss graph of net1 during the first training phase (pretrain). The mean absolute difference between the real age and the predicted brain age on the public validation set was 3.782 years. In table 4.1 the corresponding column shows the pretrain errors (MAE and MSE) of net1.

4.1.2 net2

The pretrain hyperparameters of net2 were the same as in the case of net1. The initial learning rate of the Momentum optimizer was set to 1e−6 with exponential decay. Because the number of trainable parameters of net2 is much higher than before (figure 3.9), the number of training epochs was increased by 10, to 60 epochs. The momentum parameter of the optimizer was set to 0.8.

The MSE on the public validation set was decreased slightly compared to net1, from

25.9 to 25.1. The MAE on the validation set however increased by 0.055 years which is a negligible difference. Table 4.1 shows the MAE and MSE pretrain errors of net2.

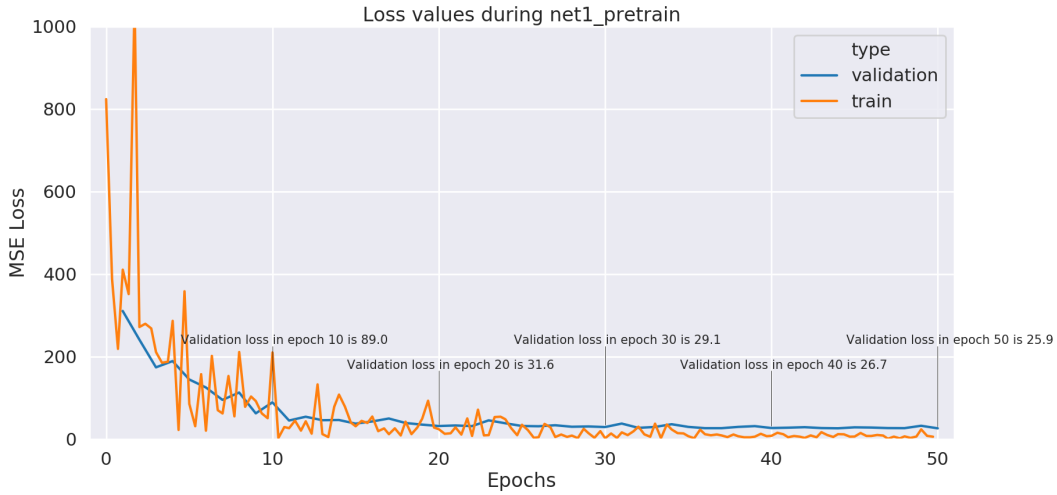


Figure 4.1: The training and the validation loss can be seen on the figure. The validation loss at the end of the training was 25.9 and the MAE, i.e. the mean absolute difference between the real age and the predicted brain age was 3.782 years.

4.1.3 net3

The number of training epochs was adjusted to 70 because of the increasing network parameters (figure 3.10). The pretrain procedure and the training hyperparameters remained unchanged. The figure 4.2 shows the training and validation loss during the training period.

The public validation loss decreased a lot compared to net2 pretraining. The mean squared error on the public validation set was 23.3 and the mean absolute erros was 3.627 years.

Table 4.1 shows the validation loss values and the MAE values on the public validation set and figure 4.3 shows the scatter plots of the results. The three plots show that the predicted brain age is on average greater than the real age for younger subjects and it is on average smaller for older patients. Both mean squared error and mean absolute error decreased compared to net1 and net2 therefore net3 was chosen for retrain and test the distributions of brain-PAD scores.

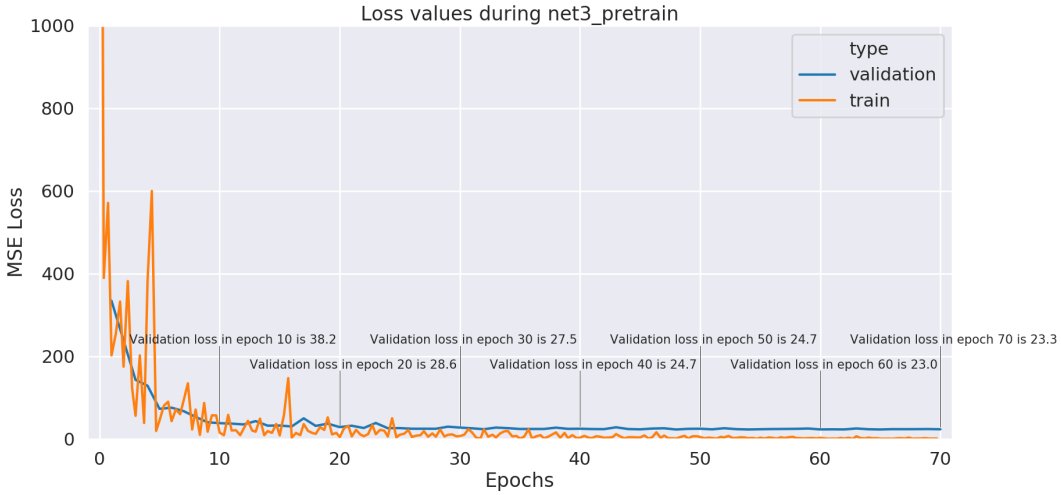


Figure 4.2: The training and the validation loss can be seen on the figure in case of net3. The validation loss at the end of the training was 23.3 and the MAE, i.e. the mean absolute difference between the real age and the predicted brain age was 3.627 years.

			network		
Phase	Metric	Dataset	net1	net2	net3
Pretrain	MSE	public VALID	25.9	25.1	23.3
	MAE	public VALID	3.782	3.837	3.627

Table 4.1: Measured MSE and MAE metrics during pretrain.

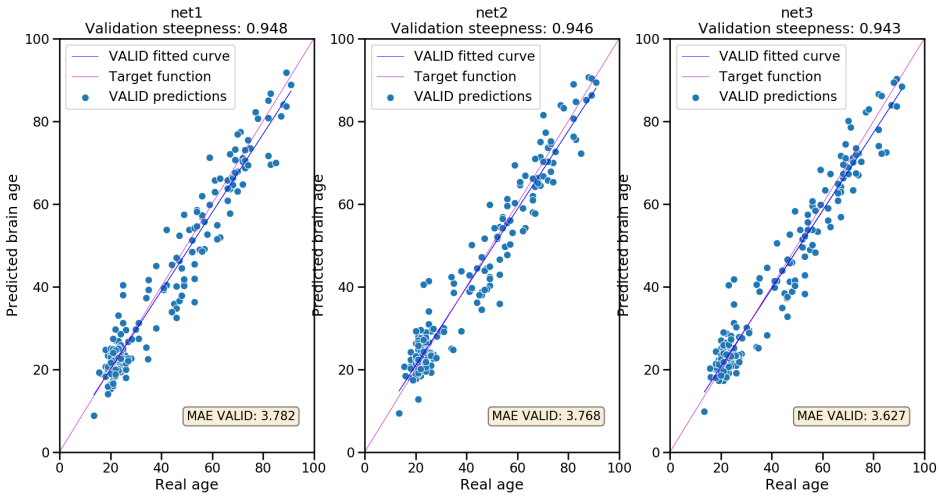


Figure 4.3: The figures show the predicted brain age compared to the real age at the end of the pretrain phase for the 3 networks. The blue marks are the public validation data and the blue line is the fitted curve on data points. The magenta lines show the optimal ($predictedage = chronologicalage$) values.

4.2. Retrain

As mentioned in subsection 3.3.1, the in-house dataset and the test dataset was collected with the same MR scanner and similar setting, therefore this step could significantly improve the efficiency. In section 3.1.3 the details of the in-house dataset were demonstrated.

The retrain phase of net3 lasted for 20 epochs on the in-house training set. The scatter plot of the predicted values can be seen on figure 4.4. The steepness of the fitted line on predictions is close to one, that means after retrain the network is not estimating greater values for younger brains and smaller age for older subjects. The fitted line computed with linear regression is under the target curve therefore the mean difference is under zero.

The migraine and the healthy control data were used after retrain as test dataset.

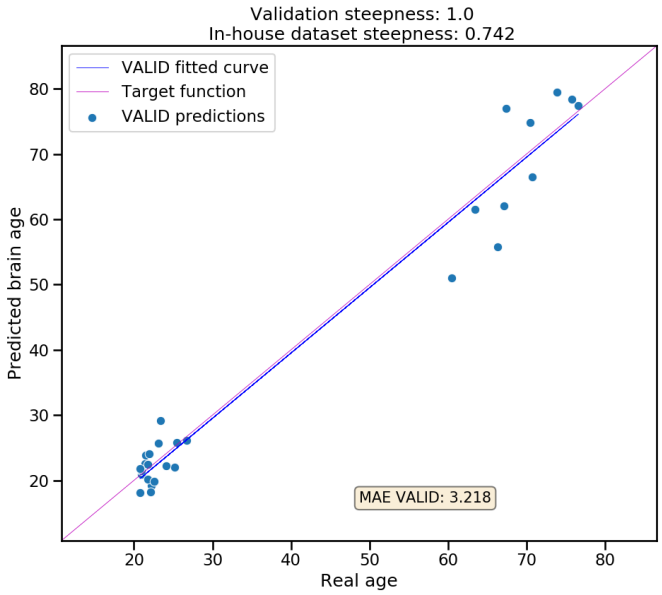


Figure 4.4: Chronological age - Predicted brain age plot on the in-house validation set after retraining net3. As the plot shows the prediction is biased since the line fitted on predictions is under the target line on the whole range. The steepness of the fitted line is close to 1 therefore the model can follow the ageing of the brain.

4.3. Testing

For testing I used the anatomical brain images of healthy and migraine subjects, detailed in subsection 3.1.1. The left plot of figure 4.6 shows the results of the healthy test dataset. The phenomenon that the younger brains are estimated older and the olders are estimated younger appeared for this dataset however the mean absolute error is 3.87 years, which is a good result compared to other solutions in literature [7] [13]. The mean difference between the chronological age and brain-predicted age is higher than zero (1.9 years), however it is caused because there are more younger subjects and their brain are predicted older. For examining the correlation between the chronological age and the predicted brain age Pearson's test was used. The p-value is close to zero (7.11e−12) therefore the correlation is significant with Pearson coefficient 0.712. The steepness of the fitted line computed with linear regression is 0.607.

For the migraine group the test results are shown on the middle plot of figure 4.6. The mean difference between the predicted brain age and real age is 1.514 years, that is smaller than for the healthy group and the mean absolute error is 4.059 years. Pearson's correlation coefficient computed between the chronological age and the predicted brain age for migraine group was resulted in 0.825 with p-value 0. The steepness of the fitted line is 0.809.

The third, right plot shows the predictions for both groups, the differences are clearly visible.

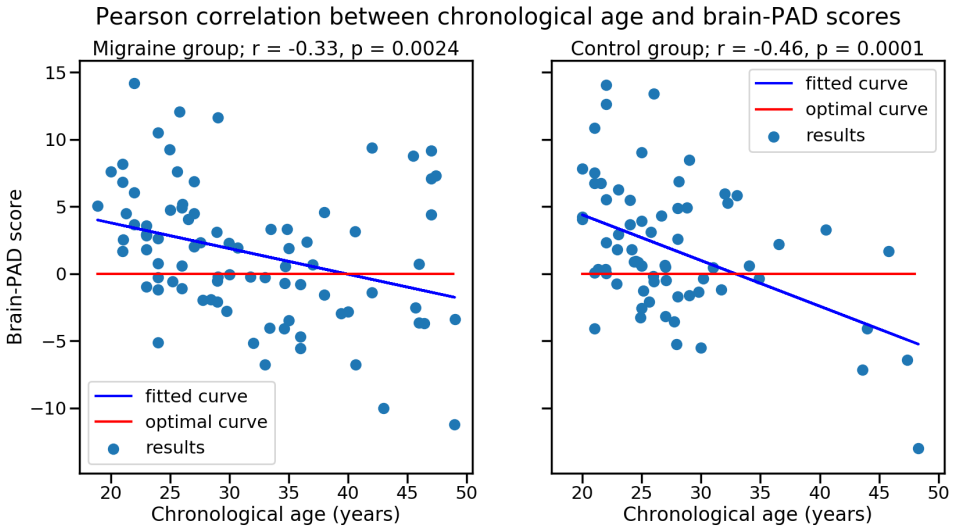


Figure 4.5: The brain-PAD scores for the two groups with respect to the chronological age.

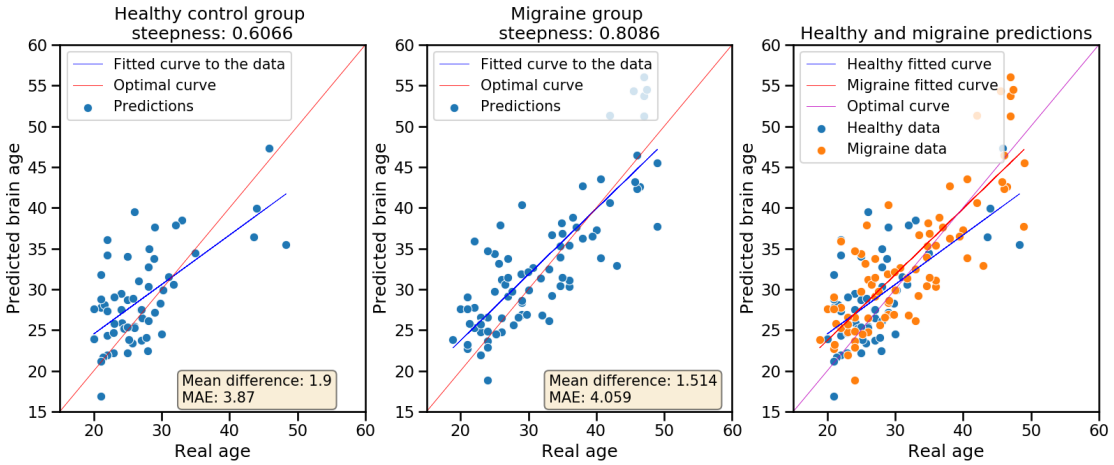


Figure 4.6: The plots show the test results of net3 on chronological age - brain-predicted age plots. The left figure shows the predicted ages for the healthy testing group, while the middle one shows the predictions for the migraine group. The right plot shows the difference between the two groups, where the blue data belongs to the healthy group and the red belongs to the migraine group.

4.4. Group analysis

According to the hypothesis, the average brain-predicted age of the migraine and healthy groups differ. To test the hypothesis the difference of the predicted brain age and the real age has to be examined, that is, the distributions of the brain-PAD score (eq.(2.25)) have to be compared.

The figure 4.5 shows the brain-PAD scores for the mentioned groups with respect to the chronological age. Predicted brain age is lower than the chronological age for older subjects and it is higher for younger subjects for both groups.

I used two-sample t-test to compare the brain-PAD scores of the groups. The necessary condition of the t-test is the normally distributed dataset, therefore I used Shapiro-Wilk test for examining the brain-PAD distributions. The results of the test can be seen in table 4.2. Because both p-values are larger than the 0.05 significance threshold we accept the null hypothesis that the data is normally distributed.

		group	
Test		migraine	healthy
Shapiro-Wilk	statistic	0.994	0.979
	p-value	0.976	0.342

Table 4.2: The results of Shapiro-Wilk test measured on the migraine and healthy brain-PAD data.

Figure 4.7 shows the histograms of the migraine and the control brain-PAD scores with fitted kde plots in the upper row and the corresponding Q-Q plots in the bottom row. The values that characterize the distribution (mean, variance, skewness and kurtosis) are shown on the histogram plots. The mean values of the two distributions are close to each other, the difference is 0.251 years. The variancevalues are also similar, the distance between them is 0.618. For the migraine group, both the kurtosis and the skewness are close to zero, that means the distribution is light-tailed and is symmetric. For the control group the kurtosis is relatively high, therefore the distribution is high-tailed.

I tested the equality of brain-PAD score variances between the two groups equality with Levene's test. The test statistic is 0.162 and the p-value is 0.688 (table: 4.3). Since the test statistic is close to zero and the p-value is larger than the limit value 0.05 I accepted that the two populations has the same variance and the standard t-test can be used.

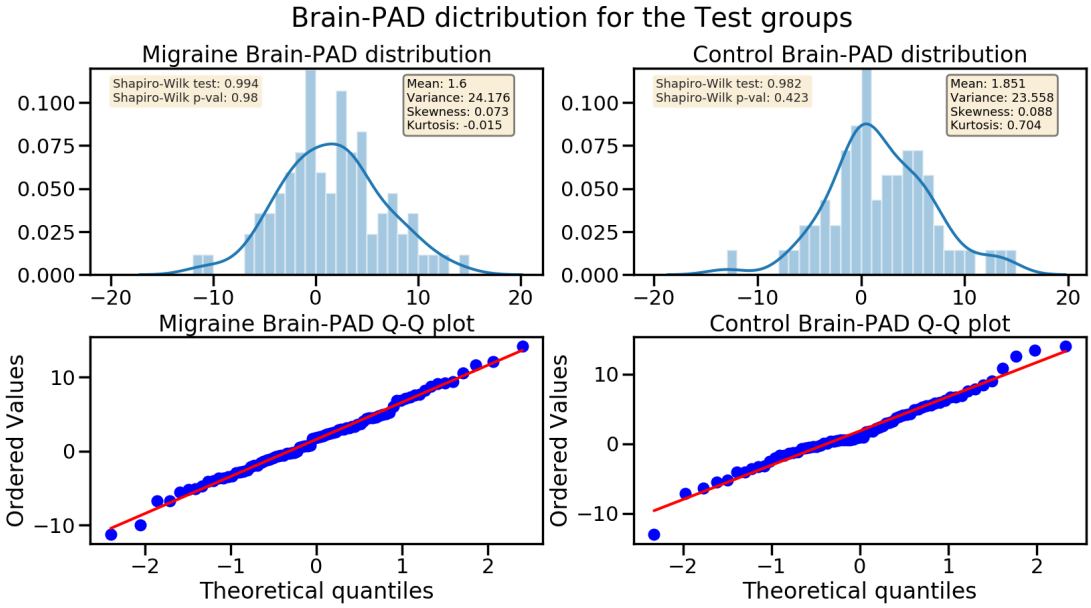


Figure 4.7: The left plots show the histogram of the brain-PAD scores of the migraine group with the corresponding Q-Q plot. The right part shows the histogram and Q-Q plot that belongs to the healthy group.

The t-test statistic was 0.315 and the corresponding p-value 0.753 (table: 4.3). Thus, brain-PAD scores did not differ significantly between migraine subjects and healthy controls.

Test		Test statistics
Levene's test	statistic	0.162
	p-value	0.688
t-test	statistic	0.315
	p-value	0.754

Table 4.3: The results of Levene's test and t-test for comparing brain-PAD distributions between migraine subjects and healthy controls.

4.4.1 Activation Visualization

Based on the method detailed in section 3.3.4 the RAM generates [5∗6∗5] heatmaps whose resolution is too small to highlight small circumscribed regions of the input volume. In figure 4.8 the activations and the corresponding brain regions can be seen for the migraine group and the healthy controls. Although the small brain areas cannot be seen with the resized heatmap the results certify that the network's predictions are largely based on

the overall structure of the brain. The figures in Appendix (A.5, A.6 and A.7) show that the eyes and the mouth cavity have no significant role in regression.

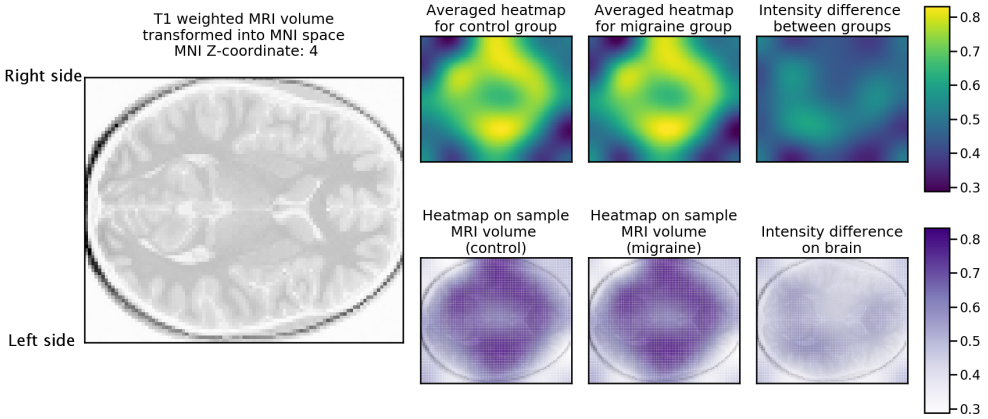


Figure 4.8: The results of RAM visualization (slice: MNI Z-coordinate 4). On the upper row the normalized heatmaps are shown for the two groups and the difference between the two groups. The green color means the larger activation and the blue is the smaller activation. On the bottom row the heatmaps are projected onto the MRI volume (Subject: 22-year-old woman). The more purple color means the greater activation.

An interesting result of RAM method is that despite the t-test showed that brain-predicted age does not differ between the two groups, a small intensity difference can be seen between the migraine and the healthy heatmaps (right side of figure 4.8). This difference is mainly located to the temporal lobe in left hemisphere.

The second method detailed in section 3.3.4 generates $[20 * 24 * 20]$ size heatmaps. The complexity of the network is reduced, however, the highlighted brain regions in figure 4.9 are consistent with the brain regions from section 2.5. The difference between migraine and healthy people decreased however all on the heatmap slices the skull bone has remarkable role in prediction (figures A.8, A.9).

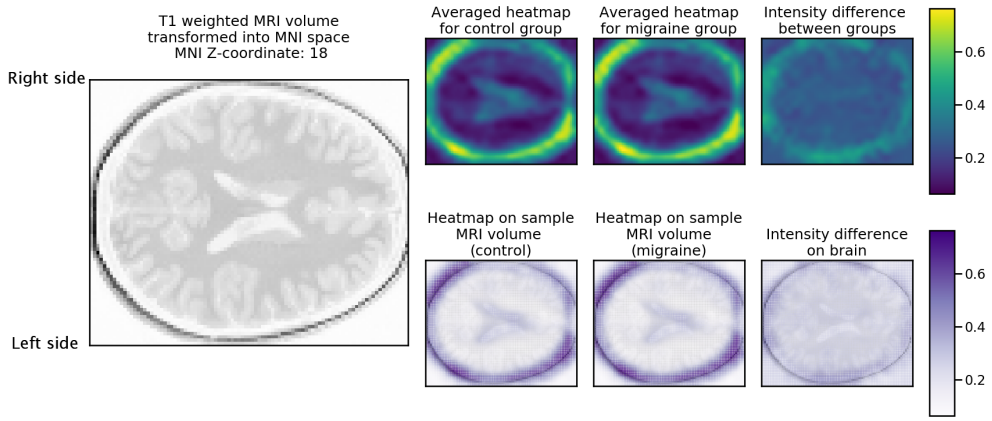


Figure 4.9: The results of RAM visualization (slice: MNI Z-coordinate 18) with the larger heatmap ($[20 * 24 * 20]$). Subject: 22-year-old woman

Chapter 5

Discussion

In medical research and applications machine learning plays an important role from year to year. Convolutional neural networks are used for medical imaging research like MR volume segmentation [3] or brain age prediction [13]. Machine learning can help medical technology develop in the next years.

Brain-age prediction [7] uses a number of training data that are T1 weighted brain MRI volumes to train a convolutional neural network for predicting the chronological age of healthy individuals. Validation error is monitored with a dedicated dataset (validation set) similar to the training set or tenfold cross-validation can be used for more accurate optimization and monitoring the performance during training. The fitted model can be used for estimating the chronological age of individuals by their brain-predicted age on unseen brain MRI samples (test set). Brain-PAD metric calculated by subtracting the chronological age from brain-predicted age (eq. (2.25)) could be used as a biomarker of an individual's brain health. Positive brain-PAD score implies older brain structure than the chronological age would justify and negative brain-PAD means younger brain. Several diseases are examined with brain-predicted age and brain-PAD score distribution (Alzheimer's disease, Parkinson's disease, HIV [13]) however migraine that also influences brain morphometry was not tested before.

Brain-age prediction is feasible based on convolutional neural networks. The methods in literature are accurate enough to use the predictions as ageing biomarkers which are helpful to measure the effects of cognitive diseases and to determine the risk of age-related deterioration and death. This biomarker could be helpful in both clinical and research applications. In clinical environment the method could be used to warn about disease risk or abnormal cognitive changes while in research more and more diseases could be examined in the future.

Article [69] emphasized that the thinning of the insula is slower for female migraineurs than for healthy women. This article implied that locally in this area the ageing of the brain of migraine sufferers decelerated compared to healthy individuals.

Based on the results described in this thesis, statistical difference was not found between the brain-PAD score distributions of migraine patients and healthy individuals. The neural activations of the convolutional network was examined and RAM activation visualization showed a small difference related to regression between the groups in the temporal lobe in the left hemisphere. (The images on the plot are flipped vertically therefore on the pictures the right sides of the temporal lobe are highlighted.) In [90] the author found strong correlation between the temporal lobe and the migraine disease however their observations highlight the right and right middle temporal lobe. The heatmaps generated from the last convolutional layer of net3 (*concat_21* in figure 3.10) showed that the network's predictions are largely based on the overall structure of the brain however the heatmap generated from the layer *batch_normalization* (3.10) with larger feature map size indicated that the skull bone also plays an important role in brain age regression.

Contrary to [69], where the author found evidence of decelerated brain aging in case of migraine sufferers, that is, the thinning of the insula of migraineur women is slower than for healthy individuals, the method described in this thesis did not find any probative evidence that the migraine can influence the ageing of the brain. Based on activation visualization a small difference can be seen on the left brain side of the temporal lobe therefore the problem requires further investigations.

5.1. Limitations of the method

All machine learning tasks need a vast amount of data that describe and represent the real distribution of the data space. The described public datasets in section 3.1 are collected from several sources and their quality were very diverse. Ideally, the training dataset would comprise uniformly distributed data by age and sex for the problem of brain age prediction. In the case of in-house dataset the age distribution is bimodal that means there are a few number of data belonging to middle aged people and few data for senior individuals.

For testing the null hypothesis accurately, that the migraine illness influences the aging of the brain, the control and migraine groups should contain the T1 weighted brain MRI volumes of individuals with similar age and the same gender characteristics. It

implies that the two groups should contain the same number of subjects with similar age and sex distributions for accurate hypothesis testing.

In articles [7] [13] the author used segmented brain MR anatomy to compute brain-PAD scores and examined the results separately for grey matter and white matter besides the whole brain. Since for net3 the RAM heatmap highlighted the skull bone for the larger feature map, training the network only on segmented or masked anatomical brain volumes where skull is removed could improve the performance of regression and whether the speed of the ageing of migraineurs' brain is accelerated or decelerated the new method could highlight the difference.

5.2. Conclusion

In many publications the authors prove that migraine can cause astrophy. These articles are measuring the structural differences between migraine patients and healthy individuals however the accelerated structural brain ageing is not examined. Based on the detailed method and the results the hypothesis that the migraine and the healthy brain aging differently is not supported. Heatmaps generated by regression activation mapping show a small difference in the structural basis of brain age prediction between the brain of healthy individuals and migraine patients. The temporal lobe in the left hemisphere appears to play a more important role in brain age prediction for migraine patients than for healthy individuals. Since we did not find statistical difference between the two groups however the activation visualization show a small difference between migraine and healthy aging the problem requires further investigations.

5.3. Summary

In the thesis I described the useful definitions related to neural network training and the important mathematical and technical details of convolutional neural networks. I reviewed task specific literature connected to CNNs as transfer learning, inception modules and class specific and regression activation visualization. In chapter 2 I described the magnetic resonance imaging with the corresponding definitions and mathematical formulas. I expounded the past researches about brain age prediction and presented some important details about migraine.

I sought for publicly available T1 weighted MRI datasets for training and collected data from three different sources.

In chapter methods I detailed three convolutional neural networks that were trained on publicly available T1 weighted brain MRI data for predicting brain age and one of the networks was fine-tuned for scanner and imaging sequence specific prediction. On the test set the measured efficiency of the trained network (net3) is comparable to the results in literature.

The healthy test set was also used as the control data to the brain MRI records of migraine sufferers. With statistical methods I examined the brain-PAD distribution similarity between the two groups and based on the methods I rejected the null hypothesis, i.e. the migraine disorder does not cause age-related structural changes in the brain.

Lastly, I visualized the network activations and used them as heatmaps on brain volumes. With two different heatmap resolutions I visualized the network activations with respect to the input volume.

Based on the method detailed in this thesis more neurocognitive diseases could be examined and their brain age related effects could be revealed.

Bibliography

[1] N. M. BALL and R. J. BRUNNER, "Data mining and machine learning in astronomy", *International Journal of Modern Physics D*, vol. 19, no. 07, pp. 1049–1106, 2010. DOI: 10.1142/S0218271810017160. eprint: https://doi.org/10.1142/S0218271810017160. [Online]. Available: https://doi.org/10.1142/S0218271810017160.

[2] (2019). Machine learning in agriculture: How ai helps solve the industry's most pressing challenges, [Online]. Available: https://objectcomputing.com/industries/agriculture/machine-learning-in-agriculture (visited on 11/21/2019).

[3] P. McClure, N. Rho, J. A. Lee, J. R. Kaczmarzyk, C. Y. Zheng, S. S. Ghosh, D. M. Nielson, A. G. Thomas, P. Bandettini, and F. Pereira, "Knowing what you know in brain segmentation using bayesian deep neural networks", *Frontiers in Neuroinformatics*, vol. 13, p. 67, 2019, ISSN: 1662-5196. DOI: 10.3389/fninf.2019.00067. [Online]. Available: https://www.frontiersin.org/article/10.3389/fninf.2019.00067.

[4] F. Ã–zyurt, E. Sert, E. Avci, and E. Dogantekin, "Brain tumor detection based on convolutional neural network with neutrosophic expert maximum fuzzy sure entropy", *Measurement*, vol. 147, p. 106 830, 2019, ISSN: 0263-2241. DOI: https://doi.org/10.1016/j.measurement.2019.07.058.

[5] L. Scotti, H. Ishiki, F. Mendonça Junior, M. Silva, and M. Scotti, "Artificial neural network methods applied to drug discovery for neglected diseases", *Combinatorial chemistry high throughput screening*, vol. 18, Aug. 2015. DOI: 10.2174/1386207318666150803141219.

[6] K. Shailaja, B. Seetharamulu, and M. Jabbar, "Machine learning in healthcare: A review", Mar. 2018, pp. 910–914. DOI: 10.1109/ICECA.2018.8474918.

[7] J. Cole, R. Poudel, D. Tsagkrasoulis, M. Caan, C. Steves, T. Spector, and G. Montana, "Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker", *NeuroImage*, vol. 163, Dec. 2016. DOI: 10.1016/j.neuroimage.2017.07.059.

[8] J. H. Cole, S. J. Ritchie, M. E. Bastin, M. C. ValdÃ©s HernÃ¡ndez, S. MuÃ±oz Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, N. R. Wray, P. Redmond, R. E. Marioni, J. M. Starr, S. R. Cox, J. M. Wardlaw, D. J. Sharp, and I. J. Deary, "Brain age predicts mortality", *Molecular Psychiatry*, vol. 23, no. 5, pp. 1385–1392, 2018, ISSN: 1476-5578. DOI: 10.1038/mp.2017.62. [Online]. Available: https://doi.org/10.1038/mp.2017.62.

[9] I. Todorov, F. Del Missier, and T. MÃntylÃ, "Age-related differences in multiple task monitoring", *PloS one*, vol. 9, no. 9, e107619–e107619, Sep. 2014, PONE-D-13-53177[PII], ISSN: 1932-6203. DOI: 10.1371/journal.pone.0107619. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25215609.

[10] D. L. Murman, "The impact of age on cognition", *Seminars in hearing*, vol. 36, no. 3, pp. 111–121, Aug. 2015, 00674[PII], ISSN: 0734-0451. DOI: 10.1055/s-0035-1555115. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27516712.

[11] F. Craik and T. Salthouse, *The Handbook of Aging and Cognition: Third Edition*. Taylor & Francis, 2011, ISBN: 9781136872143. [Online]. Available: https://books.google.hu/books?id=YeJ4AgAAQBAJ.

[12] A. Fjell and K. Walhovd, "Structural brain changes in aging: Courses, causes and cognitive consequences", *Reviews in the neurosciences*, vol. 21, pp. 187–221, Jan. 2010. DOI: 10.1515/REVNEURO.2010.21.3.187.

[13] J. H. Cole and K. Franke, "Predicting age using neuroimaging: Innovative brain ageing biomarkers", *Trends in Neurosciences*, vol. 40, no. 12, pp. 681–690, 2017, ISSN: 0166-2236. DOI: https://doi.org/10.1016/j.tins.2017.10.001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016622361730187X.

[14] J. Huang, T. G. Cooper, B. Satana, D. I. Kaufman, and Y. Cao, "Visual distortion provoked by a stimulus in migraine associated with hyperneuronal activity", *Headache: The Journal of Head and Face Pain*, vol. 43, no. 6, pp. 664–671, 2003. DOI: 10.1046/j.1526-4610.2003.03110.x. eprint: https://

headachejournal.onlinelibrary.wiley.com/doi/pdf/10.1046/j.1526-4610.2003.03110.x. [Online]. Available: https://headachejournal.onlinelibrary.wiley.com/doi/abs/10.1046/j.1526-4610.2003.03110.x.

[15] S. Magon, A. May, A. Stankewitz, P. J. Goadsby, A. R. Tso, M. Ashina, F. M. Amin, C. L. Seifert, M. M. Chakravarty, J. Müller, and T. Sprenger, "Morphological abnormalities of thalamic subnuclei in migraine: A multicenter mri study at 3 tesla", *Journal of Neuroscience*, vol. 35, no. 40, pp. 13 800–13 806, 2015, ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.2154-15.2015. eprint: https://www.jneurosci.org/content/35/40/13800.full.pdf. [Online]. Available: https://www.jneurosci.org/content/35/40/13800.

[16] E. Lotfi and A. A. Rezaee, "A competitive functional link artificial neural network as a universal approximator", *Soft Computing*, vol. 22, no. 14, pp. 4613–4625, Jul. 2018, ISSN: 1433-7479. DOI: 10.1007/s00500-017-2644-1. [Online]. Available: https://doi.org/10.1007/s00500-017-2644-1.

[17] A. Gron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st. O'Reilly Media, Inc., 2017, ISBN: 1491962291, 9781491962299.

[18] R. HECHT-NIELSEN, "Iii.3 - theory of the backpropagation neural network**based on âœnonindentâ by robert hecht-nielsen, which appeared in proceedings of the international joint conference on neural networks 1, 593â"611, june 1989. â© 1989 ieee.", in *Neural Networks for Perception*, H. Wechsler, Ed., Academic Press, 1992, pp. 65–93, ISBN: 978-0-12-741252-8. DOI: https://doi.org/10.1016/B978-0-12-741252-8.50010-8. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780127412528500108.

[19] (2019). 7 types of neural network activation functions: How to choose?, [Online]. Available: https://missinglink.ai/guides/neural-network-concepts/7-types-neural-network-activation-functions-right/ (visited on 10/22/2019).

[20] (2019). Activation functions in neural networks, [Online]. Available: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6 (visited on 10/22/2019).

[21] (). A comprehensive guide to convolutional neural networks — the eli5 way, [Online]. Available: http://mathworld.wolfram.com/HyperbolicTangent.html (visited on 10/20/2019).

[22] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network", in *2017 International Conference on Engineering and Technology (ICET)*, Aug. 2017, pp. 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186.

[23] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology", *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018, ISSN: 1869-4101. DOI: 10.1007/s13244-018-0639-9. [Online]. Available: https://doi.org/10.1007/s13244-018-0639-9.

[24] (). Convolution, [Online]. Available: http://mathworld.wolfram.com/Convolution.html (visited on 10/20/2019).

[25] (). Convolution: An exploration of a familiar operator's deeper roots, [Online]. Available: https://towardsdatascience.com/convolution-a-journey-through-a-familiar-operators-deeper-roots-2e3311f23379 (visited on 10/20/2019).

[26] (2018). An intuitive guide to convolutional neural networks, [Online]. Available: https://www.freecodecamp.org/news/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050/ (visited on 10/20/2019).

[27] A. Giusti, D. C. Cireşan, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Fast image scanning with deep max-pooling convolutional neural networks", in *2013 IEEE International Conference on Image Processing*, Sep. 2013, pp. 4034–4038. DOI: 10.1109/ICIP.2013.6738831.

[28] M. Yani, S. Irawan, and M. S.T., "Application of transfer learning using convolutional neural network method for early detection of terry's nail", *Journal of Physics: Conference Series*, vol. 1201, p. 012 052, May 2019. DOI: 10.1088/1742-6596/1201/1/012052.

[29] (). What do you mean by dense layer and drop out layer in keras neural network?, [Online]. Available: https://www.i2tutorials.com/deep-learning-interview-questions-and-answers/what-do-you-mean-by-dense-layer-and-drop-out-layer-in-keras-neural-network/ (visited on 10/20/2019).

[30] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position", *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, Apr. 1980, ISSN: 1432-0770. DOI: 10.1007/BF00344251. [Online]. Available: https://doi.org/10.1007/BF00344251.

[31] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning", in *Shape, Contour and Grouping in Computer Vision*, London, UK, UK: Springer-Verlag, 1999, pp. 319–, ISBN: 3-540-66722-9. [Online]. Available: http://dl.acm.org/citation.cfm?id=646469.691875.

[32] S. Yi, J. Ju, M.-K. Yoon, and J. Choi, *Grouped convolutional neural networks for multivariate time series*, 2017. arXiv: 1703.09938 [cs.LG].

[33] (2018). A comprehensive guide to convolutional neural networks, [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 (visited on 10/20/2019).

[34] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012, ISBN: 1600490069, 9781600490064.

[35] L. Bottou, "Large-scale machine learning with stochastic gradient descent", in *Proceedings of COMPSTAT'2010*, Y. Lechevallier and G. Saporta, Eds., Heidelberg: Physica-Verlag HD, 2010, pp. 177–186, ISBN: 978-3-7908-2604-3.

[36] Y. S. Abu-Mostafa, "The vapnik-chervonenkis dimension: Information versus complexity in learning", *Neural Computation*, vol. 1, no. 3, pp. 312–317, 1989. DOI: 10.1162/neco.1989.1.3.312. eprint: https://doi.org/10.1162/neco.1989.1.3.312. [Online]. Available: https://doi.org/10.1162/neco.1989.1.3.312.

[37] S. Ruder, *An overview of gradient descent optimization algorithms*. 2016. [Online]. Available: http://arxiv.org/abs/1609.04747.

[38] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980.

[39] D. M. Allen, "Mean square error of prediction as a criterion for selecting variables", *Technometrics*, vol. 13, no. 3, pp. 469–475, 1971. DOI: 10.1080/00401706.1971.10488811. eprint: https://amstat.tandfonline.com/doi/pdf/10.1080/00401706.1971.10488811. [Online]. Available: https://amstat.tandfonline.com/doi/abs/10.1080/00401706.1971.10488811.

[40] A. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology", *ArXiv*, vol. abs/1809.03006, 2018.

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", in *Computer Vision and Pattern Recognition (CVPR)*, 2015. [Online]. Available: http://arxiv.org/abs/1409.4842.

[42] S. J. Pan and Q. Yang, "A survey on transfer learning", *IEEE Trans. on Knowl. and Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, ISSN: 1041-4347. DOI: 10.1109/TKDE.2009.191. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2009.191.

[43] P. Vakli, R. J. Deák-Meszlényi, P. Hermann, and Z. Vidnyánszky, "Transfer learning improves resting-state functional connectivity pattern analysis using convolutional neural networks", *GigaScience*, vol. 7, no. 12, Nov. 2018, giy130, ISSN: 2047-217X. DOI: 10.1093/gigascience/giy130. eprint: http://oup.prod.sis.lan/gigascience/article-pdf/7/12/giy130/27030389/giy130.pdf. [Online]. Available: https://doi.org/10.1093/gigascience/giy130.

[44] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?", in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2014, pp. 3320–3328. [Online]. Available: http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf.

[45] S. W. Atlas, *Magnetic Resonance Imaging of the Brain and Spine*. LWW, 2016, ISBN: 1469873206, 9781469873206.

[46] J. P. Hornak, *The Basics of MRI*. Interactive Learning Software, Henietta, NY, 1996. [Online]. Available: https://www.cis.rit.edu/htbooks/mri/.

[47] (). Mr/ct/röntgen/pet/spect képalkotás, [Online]. Available: http://vision.itk.ppke.hu/?q=node/923 (visited on 10/22/2019).

[48] (2019). Magnetom spectra: Its the key to 3t., [Online]. Available: https://www.siemens-healthineers.com/magnetic-resonance-imaging/3t-mri-scanner/magnetom-spectra (visited on 10/20/2019).

[49] (). About the mni space(s), [Online]. Available: https://www.lead-dbs.org/about-the-mni-spaces/ (visited on 10/20/2019).

[50] M. Brett, I. Johnsrude, and A. Owen, "Opinionthe problem of functional localization in the human brain", *Nature reviews. Neuroscience*, vol. 3, pp. 243–9, Apr. 2002. DOI: 10.1038/nrn756.

[51] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images", *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1240–1251, May 2016. DOI: 10.1109/TMI.2016.2538465.

[52] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using u-net based fully convolutional networks", in *Medical Image Understanding and Analysis*, M. Valdés Hernández and V. González-Castro, Eds., Cham: Springer International Publishing, 2017, pp. 506–517, ISBN: 978-3-319-60964-5.

[53] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller, "Deep mri brain extraction: A 3d convolutional neural network for skull stripping", *NeuroImage*, vol. 129, Jan. 2016. DOI: 10.1016/j.neuroimage.2016.01.024.

[54] A. Payan and G. Montana, *Predicting alzheimer's disease: A neuroimaging study with 3d convolutional neural networks*, 2015. arXiv: 1502.02506 [cs.CV].

[55] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, "Residual and plain convolutional neural networks for 3d brain mri classification", in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, Apr. 2017, pp. 835–838. DOI: 10.1109/ISBI.2017.7950647.

[56] I. J. Deary, J. Corley, A. J. Gow, S. E. Harris, L. M. Houlihan, R. E. Marioni, L. Penke, S. B. Rafnsson, and J. M. Starr, "Age-associated cognitive decline", *British Medical Bulletin*, vol. 92, no. 1, pp. 135–152, Sep. 2009, ISSN: 0007-1420. DOI: 10.1093/bmb/ldp033. eprint: http://oup.prod.sis.lan/bmb/article-pdf/92/1/135/951616/ldp033.pdf. [Online]. Available: https://doi.org/10.1093/bmb/ldp033.

[57] C. McAlister and M. Schmitter-Edgecombe, "Naturalistic assessment of executive function and everyday multitasking in healthy older adults", *Aging, Neuropsychology, and Cognition*, vol. 20, no. 6, pp. 735–756, 2013, PMID: 23557096. DOI: 10.1080/13825585.2013.781990. eprint: https://doi.org/10.1080/

13825585.2013.781990. [Online]. Available: https://doi.org/10.1080/13825585.2013.781990.

[58] A. Williamson, "Â˜you're never too old to learn!â^{TM}: Thirdâage perspectives on lifelong learning", *International Journal of Lifelong Education*, vol. 16, no. 3, pp. 173–184, 1997. DOI: 10.1080/0260137970160302. eprint: https://doi.org/10.1080/0260137970160302. [Online]. Available: https://doi.org/10.1080/0260137970160302.

[59] F. McNab, P. Zeidman, R. B. Rutledge, P. Smittenaar, H. R. Brown, R. A. Adams, and R. J. Dolan, "Age-related changes in working memory and the ability to ignore distraction", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 20, pp. 6515–6518, May 2015, 1504162112[PII], ISSN: 1091-6490. DOI: 10.1073/pnas.1504162112. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25941369.

[60] "Does ageing inevitably lead to declines in cognitive performance? a longitudinal study of elite academics", *Personality and Individual Differences*, vol. 23, no. 1, pp. 67–78, 1997, ISSN: 0191-8869. DOI: https://doi.org/10.1016/S0191-8869(97)00022-6. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0191886997000226.

[61] Z. Geng, H. Liu, L. Wang, Q. Zhu, Z. Song, R. Chang, and H. Lv, "A voxel-based morphometric study of age- and sex-related changes in white matter volume in the normal aging brain", *Neuropsychiatric Disease and Treatment*, vol. 12, p. 453, Feb. 2016. DOI: 10.2147/NDT.S90674.

[62] N. J. Davis, "Brain stimulation for cognitive enhancement in the older person: State of the art and future directions", *Journal of Cognitive Enhancement*, vol. 1, no. 3, pp. 337–344, Sep. 2017, ISSN: 2509-3304. DOI: 10.1007/s41465-017-0036-1. [Online]. Available: https://doi.org/10.1007/s41465-017-0036-1.

[63] [Online]. Available: https://migraineresearchfoundation.org/.

[64] (2019). Mayo clinic: Migraine, [Online]. Available: https://www.mayoclinic.org/diseases-conditions/migraine-headache/symptoms-causes/syc-20360201 (visited on 11/21/2019).

[65] M. A. Rocca, A. Ceccarelli, A. Falini, B. Colombo, P. Tortorella, L. Bernasconi, G. Comi, G. Scotti, and M. Filippi, "Brain gray matter changes in migraine patients with t2-visible lesions", *Stroke*, vol. 37, no. 7, pp. 1765–1770, 2006. DOI: 10.1161/

01.STR.0000226589.00599.4d. eprint: https://www.ahajournals.org/doi/pdf/10.1161/01.STR.0000226589.00599.4d. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/01.STR.0000226589.00599.4d.

[66] N. Pavese, R. Canapicchi, A. Nuti, F. Bibbiani, C. Lucetti, P. Collavoli, and U. Bonuccelli, "White matter mri hyperintensities in a hundred and twenty-nine consecutive migraine patients", *Cephalalgia*, vol. 14, no. 5, pp. 342–345, 1994. DOI: 10.1046/j.1468-2982.1994.1405342.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1468-2982.1994.1405342.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1468-2982.1994.1405342.x.

[67] A. Bashir, R. B. Lipton, S. Ashina, and M. Ashina, "Migraine and structural changes in the brain", *Neurology*, vol. 81, no. 14, pp. 1260–1268, 2013, ISSN: 0028-3878. DOI: 10.1212/WNL.0b013e3182a6cb32. eprint: https://n.neurology.org/content/81/14/1260.full.pdf. [Online]. Available: https://n.neurology.org/content/81/14/1260.

[68] Z. Jia and S. Yu, "Grey matter alterations in migraine: A systematic review and meta-analysis", *NeuroImage. Clinical*, vol. 14, pp. 130–140, Jan. 2017, ISSN: 2213-1582. DOI: 10.1016/j.nicl.2017.01.019.

[69] N. Maleki, G. Barmettler, E. A. Moulton, S. Scrivani, R. Veggeberg, E. L. H. Spierings, R. Burstein, L. Becerra, and D. Borsook, "Female migraineurs show lack of insular thinning with age", *Pain*, vol. 156, no. 7, pp. 1232–1239, Jul. 2015, ISSN: 1872-6623. DOI: 10.1097/j.pain.0000000000000159.

[70] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization", in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[71] A. Chattopadhay, A. Sarkar, P. Howlader, and 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, pp. 839–847. DOI: 10.1109/WACV.2018.00097.

[72] Z. Wang and J. Yang, "Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation", *ArXiv*, vol. abs/1703.10757, 2017.

[73] (2019). Magnetom prisma: The 3t powerpack for exploration., [Online]. Available: https://www.siemens-healthineers.com/magnetic-resonance-imaging/3t-mri-scanner/magnetom-prisma (visited on 10/20/2019).

[74] (2019). Nitrc, [Online]. Available: https://www.nitrc.org/include/about_us.php (visited on 10/20/2019).

[75] (2019). Fcp classic data sharing samples, [Online]. Available: http://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html (visited on 10/20/2019).

[76] (2019). Southwest university adult lifespan dataset (sald), [Online]. Available: http://fcon_1000.projects.nitrc.org/indi/retro/sald.html (visited on 10/20/2019).

[77] (2019). Alzheimer's disease neuroimaging initiative, [Online]. Available: http://adni.loni.usc.edu/ (visited on 10/20/2019).

[78] (2019). Spyder, [Online]. Available: https://www.spyder-ide.org/ (visited on 10/22/2019).

[79] (2019). Tensorflow, [Online]. Available: https://www.tensorflow.org/ (visited on 10/22/2019).

[80] T. Verhulsdonck. (2018). An advanced example of the tensorflow estimator class, [Online]. Available: https://towardsdatascience.com/an-advanced-example-of-tensorflow-estimators-part-1-3-c9ffba3bff03 (visited on 10/22/2019).

[81] J. Duda, "SGD momentum optimizer with step estimation by online parabola model", *arXiv e-prints*, arXiv:1907.07063, arXiv:1907.07063, Jul. 2019. arXiv: 1907.07063 [cs.LG].

[82] W. Haynes, "Student's t-test", in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds. New York, NY: Springer New York, 2013, pp. 2023–2025, ISBN: 978-1-4419-9863-7. DOI: 10.1007/978-1-4419-9863-7_1184. [Online]. Available: https://doi.org/10.1007/978-1-4419-9863-7_1184.

[83] Z. Lu and K.-H. Yuan, "Welch's t test", in. Jan. 2010, pp. 1620–1623. DOI: 10.13140/RG.2.1.3057.9607.

[84] (). Normal distribution, [Online]. Available: http://mathworld.wolfram.com/NormalDistribution.html (visited on 11/08/2019).

[85] P. Royston, "Approximating the shapiro-wilk w-test for non-normality", *Statistics and Computing*, vol. 2, no. 3, pp. 117–119, Sep. 1992, ISSN: 1573-1375. DOI: 10.1007/BF01891203. [Online]. Available: https://doi.org/10.1007/BF01891203.

[86] K. V. MARDIA, "Measures of multivariate skewness and kurtosis with applications", *Biometrika*, vol. 57, no. 3, pp. 519–530, Dec. 1970, ISSN: 0006-3444. DOI: 10.1093/biomet/57.3.519. eprint: http://oup.prod.sis.lan/biomet/article-pdf/57/3/519/702615/57-3-519.pdf. [Online]. Available: https://doi.org/10.1093/biomet/57.3.519.

[87] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient", in *Noise Reduction in Speech Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–4, ISBN: 978-3-642-00296-0. DOI: 10.1007/978-3-642-00296-0_5. [Online]. Available: https://doi.org/10.1007/978-3-642-00296-0_5.

[88] G. D. Ruxton, "The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test", *Behavioral Ecology*, vol. 17, no. 4, pp. 688–690, May 2006, ISSN: 1045-2249. DOI: 10.1093/beheco/ark016. eprint: http://oup.prod.sis.lan/beheco/article-pdf/17/4/688/17275561/ark016.pdf. [Online]. Available: https://doi.org/10.1093/beheco/ark016.

[89] (2019). Scipy, [Online]. Available: https://scipy.org/ (visited on 10/20/2019).

[90] T. J. Schwedt, V. Berisha, and C. D. Chong, "Temporal lobe cortical thickness correlations differentiate the migraine brain from the healthy brain", *PLOS ONE*, vol. 10, no. 2, pp. 1–12, Feb. 2015. DOI: 10.1371/journal.pone.0116687. [Online]. Available: https://doi.org/10.1371/journal.pone.0116687.

[91] (2019). Tensorboard, [Online]. Available: https://www.tensorflow.org/tensorboard (visited on 11/21/2019).

Appendix A

Appendix

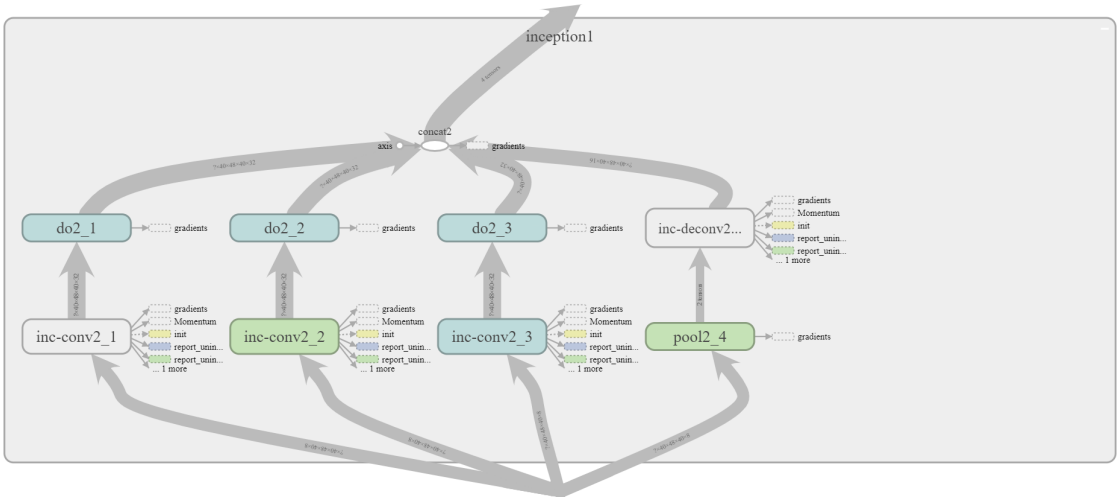


Figure A.1: The inception module of net2 and the first inception module of net3 contains three convolutional layers with $1 * 1 * 1$, $3 * 3 * 3$ and $5 * 5 * 5$ kernels. Each of them has 32 filters. One dropout layer follows the convolutions. The fourth node of the inception module is a deconvolutional layer with $5 * 5 * 5$ kernel, $2 * 2 * 2$ stride size and 16 filters. The max pooling layer restores the size of the feature map with $2 * 2 * 2$ pooling size. The last step is the concatenation along the axis of feature map depth. (generated by Tensorboard [91])

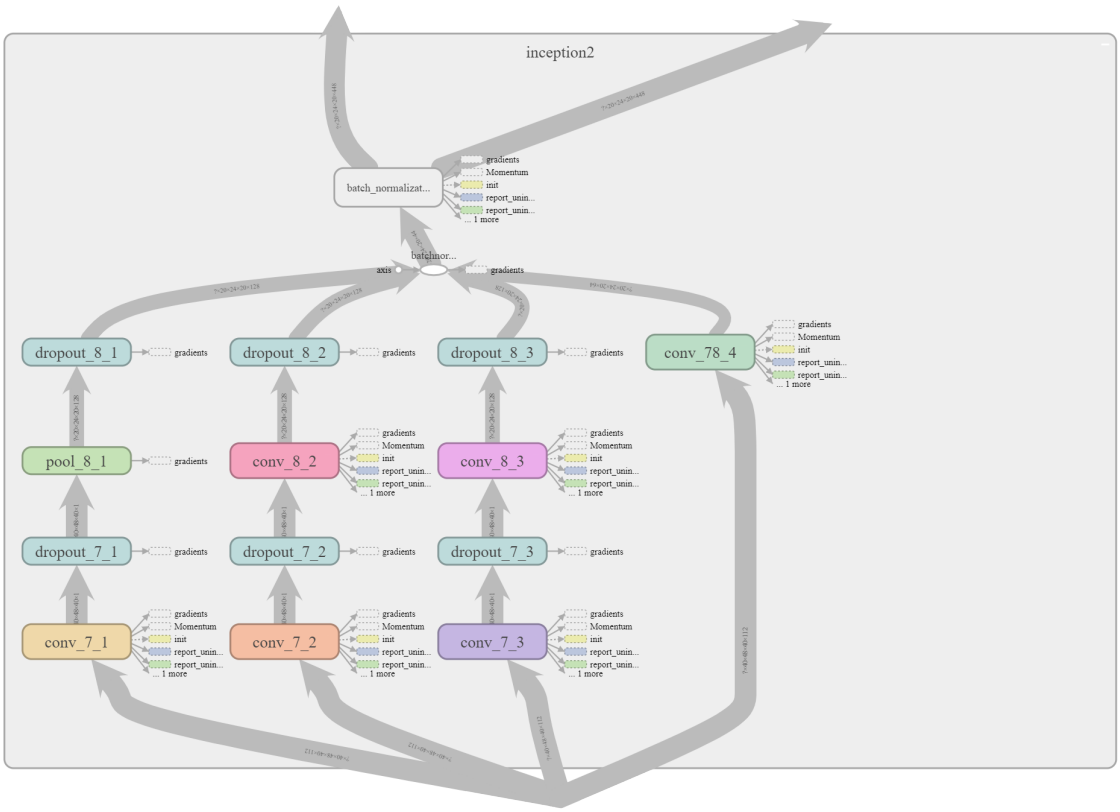


Figure A.2: The second inception module of net3 also contains four nodes (generated by Tensorboard [91]). The first convolution layers of the nodes (conv_7_1, conv_7_2 and conv_7_3) have $1 * 1 * 1$ kernel size and 128 filters. On the first node the convolutional layer is followed by a dropout layer, a max pooling layer with pool size 2 and one more dropout layer. The second and third node also contains dropout layer, however the conv_8_2 and conv_8_3 have $3 * 3 * 3$ and $5 * 5 * 5$ kernels and $2 * 2 * 2$ strides. Their depth is 128. The last node with the conv_78_4 layer has $3 * 3 * 3$ convolutional kernel, $2 * 2 * 2$ stride size and 64 filters. The last step is the concatenation along the axis of feature map depth.

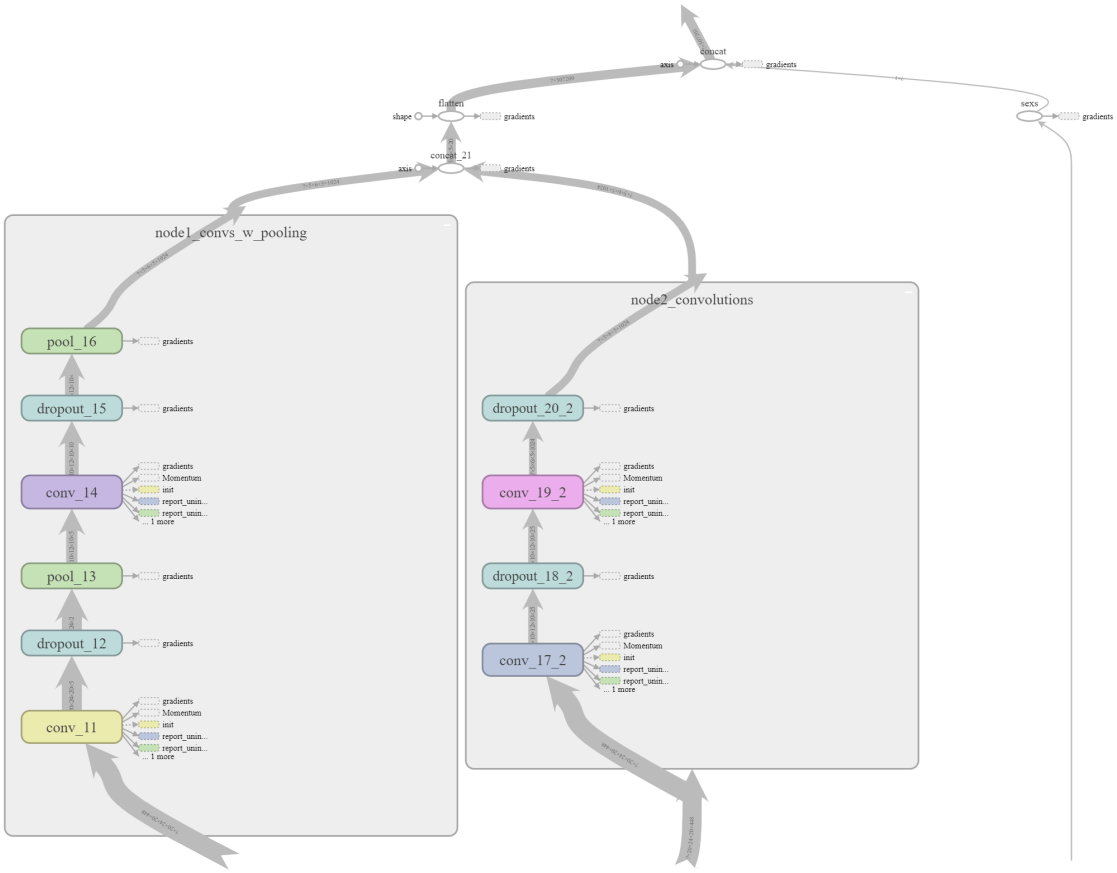


Figure A.3: The last module has two nodes (generated by Tensorboard [91]). The first node (node1_convs_w_pooling) has two convolution layers followed by dropout and max pooling layers. Both convolution layers has $3*3*3$ kernel, and 512 and 1024 filters, consequently. The max pooling layers are working with $2*2*2$ pooling size. The second note contains two convolutional layers. The first has $1*1*1$ kernel and the second has $5*5*5$ with 256 and 1024 filters. The shape reduction is realized by the convolutional layers with $2*2*2$ stride size. After concatenation and a flatten layer the gender bit is merged to the vector.

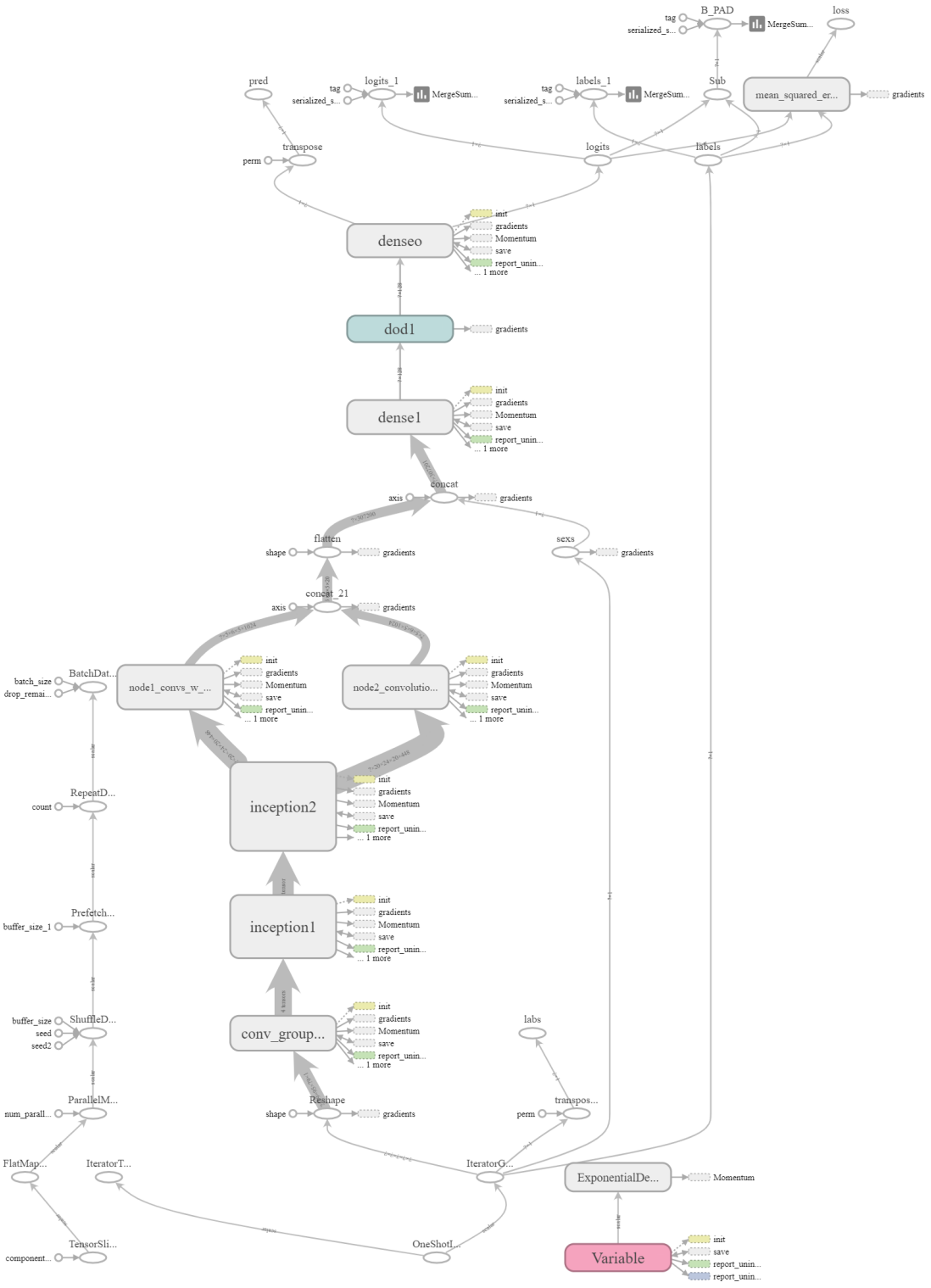


Figure A.4: The structure of the third network with inception modules generated by Tensorboard [91]

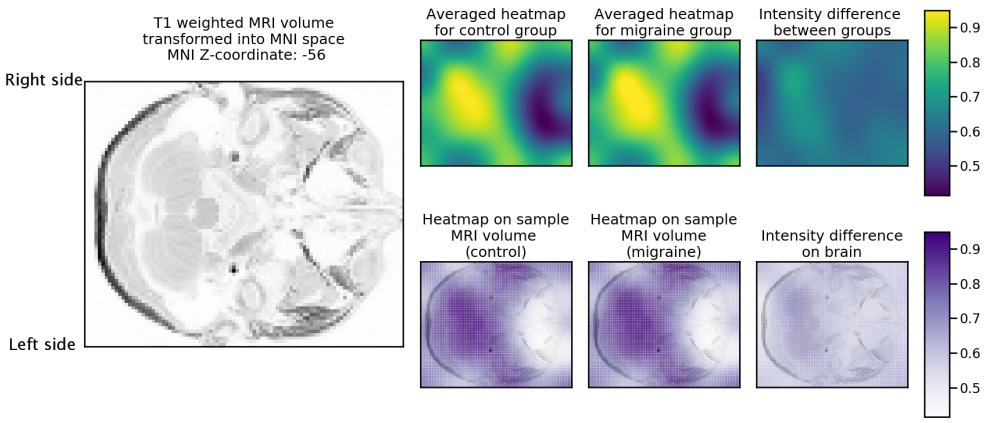


Figure A.5: The results of RAM visualization (slice: MNI Z-coordinate -56) with heatmap size $[5 * 6 * 5]$. On the upper row the normalized heatmaps are shown for the two groups and the difference between the the two groups. On the bottom row the heatmaps are projected onto the MRI volume (Subject: 22-year-old woman).

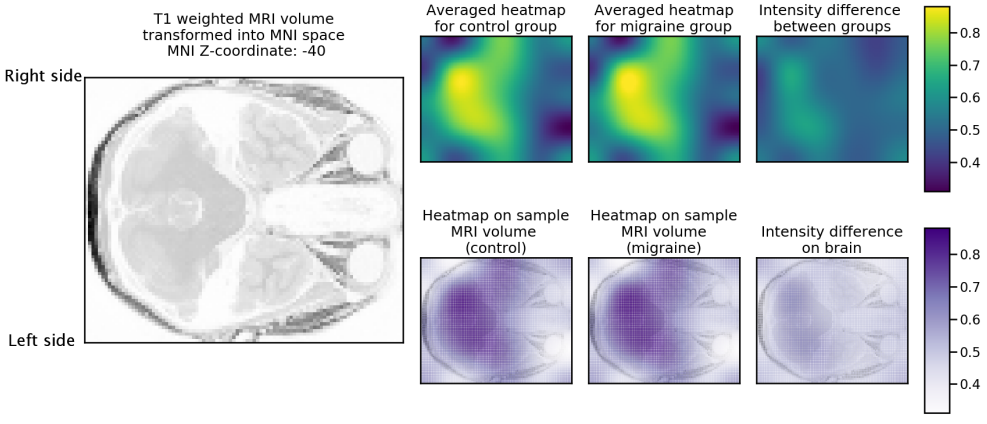


Figure A.6: The results of RAM visualization (slice: MNI Z-coordinate -40) with heatmap size $[5 * 6 * 5]$. Subject: 22-year-old woman

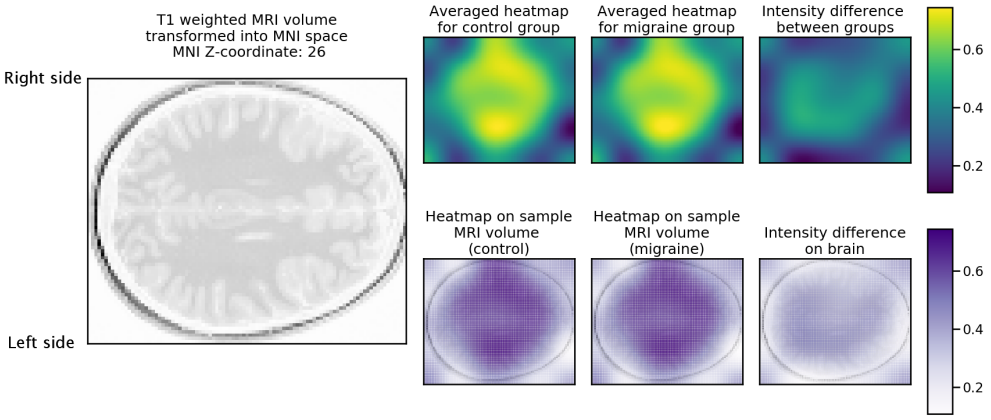


Figure A.7: The results of RAM visualization (slice: MNI Z-coordinate 26) with heatmap size $[5 * 6 * 5]$. Subject: 22-year-old woman

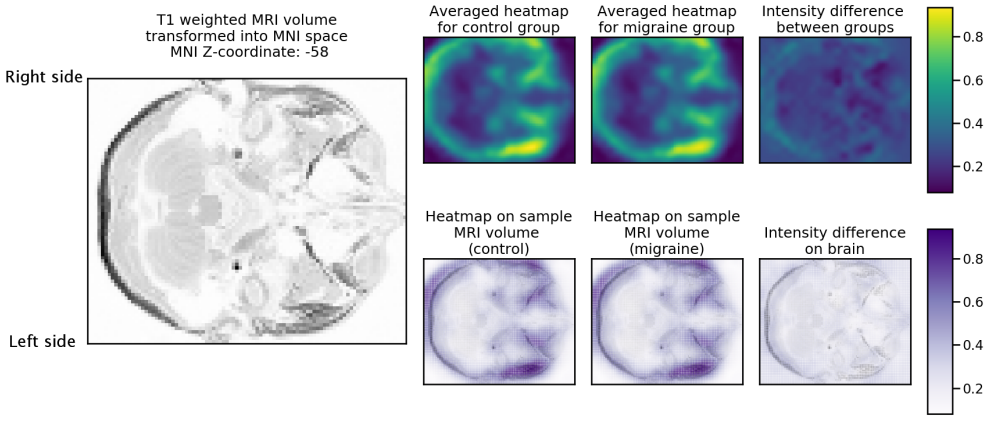


Figure A.8: The results of RAM visualization (slice: MNI Z-coordinate -58) with heatmap size $[20 * 24 * 20]$. Subject: 22-year-old woman

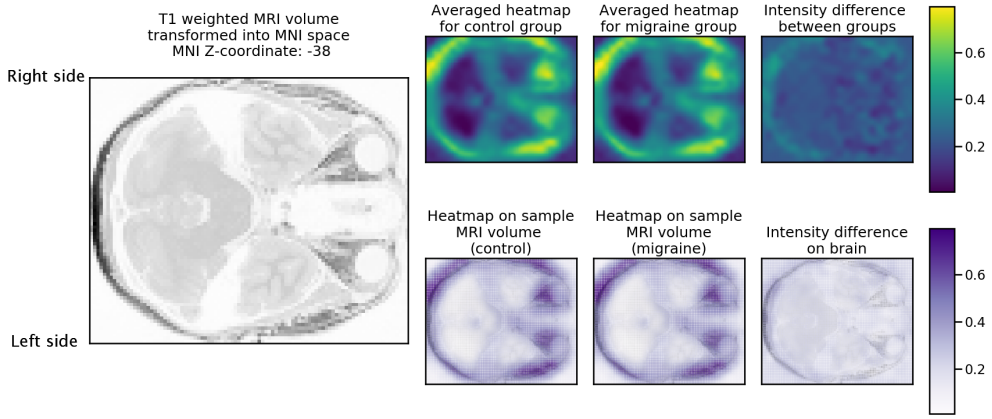


Figure A.9: The results of RAM visualization (slice: MNI Z-coordinate -38) with heatmap size $[20 * 24 * 20]$. Subject: 22-year-old woman