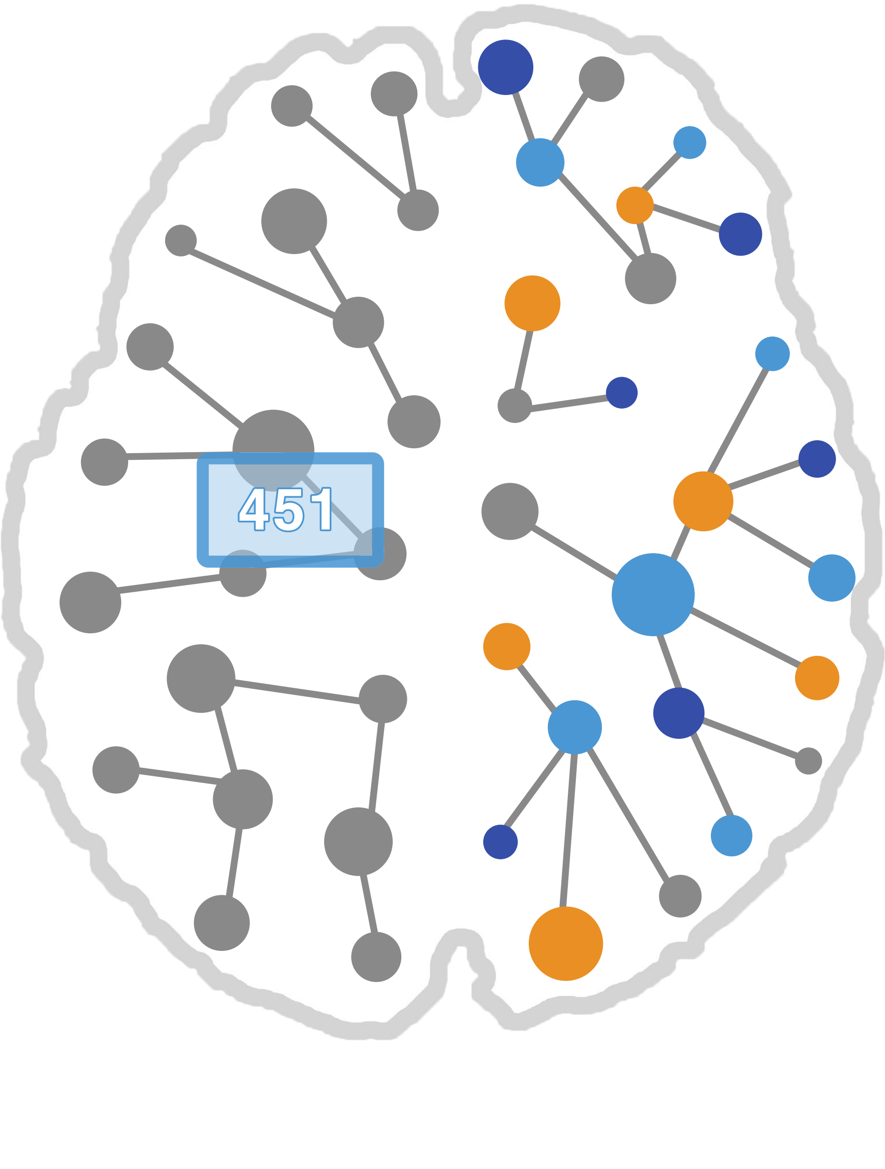
When Weak Becomes Strong

Robust Quantification of White Matter Hyperintensities on Brain MRIs



O. Werner

When Weak Becomes Strong

Robust Quantification of White Matter Hyperintensities on Brain MRIs

by

O. Werner

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday, October 30th, 2020 at 9:00 AM.

Student number:	4167880	
Thesis committee:	Prof. W. J. Niessen,	TU Delft, supervisor
	Prof. M. de Bruijne,	Erasmus MC, supervisor
	Dr. F. Dubost,	Erasmus MC, daily supervisor
	Dr. D. M. J. Tax,	TU Delft
	Dr. F. M. Vos,	TU Delft
	Dr. M. Staring,	TU Delft

Abstract

In clinical practice, as a first approximation, the severity of an abnormality on an image is often determined by measuring its volume. Researchers often first segment this abnormality with a neural network trained by voxel-wise labels and thereafter extract the volume. Instead of this indirect two steps approach, we propose to train neural networks directly using the volumes as image-level label and predict the volume directly. Using image-level labels to train automatic abnormality prediction could decrease the labeling burden for clinical experts, which is both expensive and time consuming. In this report, a neural network that consisted of a segmentation part and an appended regression part was compared with the indirect segmentation approach. It was investigated if networks trained with image-level labels have the same performance of image-level prediction as networks trained with voxel-wise labels. The neural networks were trained on a large local dataset to quantify white matter hyperintensity (WMH) burden from brain MRI, and their performance was evaluated on a held-out test set. Furthermore, generalization properties were compared by applying the trained networks on four independent public datasets. The networks trained with image-level labels achieved volume quantification that was slightly better than their counterpart on the held-out test set. The attention maps of these networks showed that the networks were able to focus on the surroundings of the WMH, and hence learned meaningful image features. Nevertheless, the attention maps were not suitable to achieve a compatible segmentation. In terms of generalization towards external datasets, the advantage of weak labels for volume quantification did not hold as there was no significant difference between the performance of the label types. The results suggest that neural networks optimized with image-level labels were able to directly predict WMH volume as well as neural networks trained with voxel-wise labels. Subsequently, we also studied networks that were optimized on both image-level and voxel-wise labels. Those networks reached a lower performance, which suggested that the tasks and their image features learned were not similar enough.

Contents

List of Figures

List of Tables

1 Introduction

It is not the strongest of the species that survives (...). It is the one that is the most adaptable to change.

Charles Darwin

Recent deep learning research showed that neural networks are capable of learning strong image features and have near-human performance on prediction or segmentation of objects on image scans. Most deep learning methods need a lot of training data to converge. This is especially challenging in medical image analysis as annotations have to be done by clinical experts. Especially, strong annotations (segmentations or voxel-wise annotations) are generally very time consuming and expensive to obtain. Therefore, studies have looked into the feasibility of training a neural network with *weak* labels (Cheplygina et al., 2019; de Bruijne, 2016). These labels are often cheaper and less time-consuming to annotate, can be produced in larger quantities, and are often either already available, i.e. Electronic Medical Records (EMR) (Chaganti et al., 2019) or could be annotated by non-experts (McKenna et al., 2012).

The term weak labels has many definitions. Zhou provided one definition with three different kinds of weak supervision: incomplete, inexact and inaccurate supervision.
For the first label type, *incomplete supervision*, which is also often referred to as *semi-supervised*, the dataset consists of both labeled and unlabeled data.
Second, *inexact supervision*, includes labels that do not fully cover or explain the image. For example, an image depicting a street with a car, a bus, and predestrians only has the label *car*. Sun et al. showed that the absence of parts of the label did not imply that the label was not suitable for training. The network can learn that next to the provided label, other labels should not be excluded and are still relevant. Similar to this inexact supervision, Tan et al. defined weak labels as labels where only a subset of the relevant labels are annotated and hence some labels are missing.
Third, inaccurate supervision consists of labels that are challenging to annotate and lead to relatively inaccurate approximations of the ground-truth. This is the case for certain applications e.g. brain lesions, for which there is no feasible way to obtain the ground truth and annotating is difficult for clinical experts leading to a relatively low interrater and intrarater agreement (Zou et al., 2004). Another example is the use of crowd-based annotations, in

which a group of non-experts rates or annotates a dataset. Their annotations are merged into one by averaging or majority voting (Maier-Hein et al., 2014).
Labels can also be considered weak when the target labels contain more information than the training labels. The term weak does not relate to the quality of the label itself as the previous definitions, but to the relation between the train label with the target label. An example is for instance predicting segmentations when trained with image-level labels (Jia et al., 2017).

Furthermore, if only a part of the image is annotated with voxel-wise annotations, then this can also be considered a weak label (Koch et al., 2017). This can be categorized as incomplete supervision, i.e. a part of the voxels have labels, and another part is unlabeled. Scribbles (Lin et al., 2016) and bounding boxes (Dai et al., 2015) are two kinds of weak labels that include classification and localization information of the voxels they pass or surround. They are easy and fast to create for a rater, while providing much information. These labels are a combination of inexact and inaccurate supervision, as they do not fully explain the target and only provide an approximation of its localization. In addition, weak labels can consist of one or multiple words to describe an image. In computer vision, these are often the class labels of the images (Noh et al., 2015), while in medical image analysis they are often clinical reports extracted from EMR (Shin et al., 2016). These labels are also called semantic labels. Semantic labels are a special case of inexact supervision. It is hard to describe an object of an image perfectly with just words, hence a picture is worth a thousand words.
Subsequently, labels created by other pretrained models are also considered weak labels (Wang et al., 2019b).
Lastly, weak or strong labels can be defined as image-level labels and voxel-level labels, respectively. In this definition weak labels are often used for regression and classification purposes, while strong labels are used for segmentation. The image-level weak label can be regarded as inexact supervision because it is coarse, or as incomplete supervision as it only covers a group of voxels. Several examples of this type of weak labels are volume (Kervadec et al., 2019), number of occurrences (Dubost et al., 2017b) or severity score (Arvaniti and Claassen, 2018). These kind of weak labels were investigated within this study. For clarification, within this report, the image-level labels were considered *weak* labels, i.e. volume or number of occurrences. The voxel-wise annotations were considered to be *strong* labels.

In clinical practise, the volume of an abnormality is often used to determine a scale of severity. This quantitative number can be readily used as a weak label for an automated approach. When using an automated approach, there are two ways to predict the volume: the direct and the indirect approach. A regression network that is trained with image-level labels consisting of the total volume predicts the volume directly, hence the direct approach. With the indirect approach, a segmentation network is used that first predicts a segmentation from which the volume is extracted. The direct approach could have better performance, as it tries to directly optimize on the volume. It can find image features or biomarkers that help predict the volume that do not necessarily delineate the regions of the abnormalities. The indirect approach, on the other hand, has voxel-wise labels to learn from. These labels contain much more information about the abnormality than the image-level labels.

Aside from the decrease in annotation burden, weak labels could also be beneficial for changes in protocols of image acquisition. The image-level labels are more robust to changes, as the abnormality does not change due to a change in protocol. The voxel-wise annotation, however, might be skewed, translated or undergo any other form of rigid and non-rigid transformation.

In this thesis we address the hypothesis that direct optimization on image-level labels could have similar performance as an image-level label prediction via voxel-wise labels. The generalization properties of both methods were evaluated with the use of public datasets other than the training data. Furthermore, it is investigated whether a network could benefit when optimizing on different label types simultaneously to predict image-level labels. The main medical application studied was the prediction of the White Matter Hyperintensity (WMH) burden, i.e. the WMHs volume, WMH severity score, or the number of WMH lesions. Other medical applications investigated were the MS lesion volume for generalization properties, and the cardiac right ventricle to evaluate the performance on heterogeneous tissue. As a controlled environment, an artificial dataset was used as well.

1.1. Contributions

Our contributions were fourfold. First we showed that the WMH volume prediction by networks optimized with image-level labels had a slight but significant increase in performance compared to the networks optimized with voxel-wise labels on the test set of the training data. Second, we showed that these image-level label networks learned other, but meaningful image features than the voxel-wise label networks. Subsequently, these image features were not confounded with the most obvious WMH confounders, i.e. intracranial volume, ventricle volume and white matter volume. Third, we evaluated the generalization properties of image-level label networks and voxel-wise label networks and found no significant difference. Lastly, we showed that networks trained with both image-level and voxel-wise labels simultaneously had a worse performance. This could indicate that the tasks, i.e. volume prediction and segmentation, did not share the same optimal solution.

1.2. Outline

First, related research is discussed in the section below. Thereafter, the clinical importance of the quantitative image biomarker WMHs and MS lesions are discussed. In chapter 4, the proposed method is described. The data used for training and testing is addressed in chapter 5. After, the implementation and evaluation of the experiments is discussed in chapter 6. The experiments and their results are shown in chapter 7. Subsequently, the results are discussed in chapter 8.

2
Related research

Many Medical Image Analysis papers made use of weak labels for training neural networks in the recent years. The tasks for these networks have a great variety. They range from classification, detection, regression to segmentation. Similar to our method for WMH severity prediction, Bortsova et al.; Hussain et al.; Jia et al.; Kervadec et al.; Luo et al. used volumes as weak labels to optimize their neural networks. Kervadec et al. trained a regression network to predict volumes of the left ventricle on cardiac scans, and use the network to predict volume labels on unlabeled images. An additional segmentation network used these volume labels as constraint to learn in a semi-supervised setting. Jia et al. used coarse volume labels. Their labels existed of area sizes relative to the image size. They used these labels, together with binary classification labels to supervise the network after each convolution layer in a multi-supervised setting. Their network was trained to learn both classification and segmentation of histopatology scans. As the areas were coarse defined, they claimed that the creation was not a big burden on the clinical experts, while the performance of the network with this additional supervision increased significantly. Hussain et al. used a network to first find the boundary slices of the kidney and hence the region of interest (ROI). The centre slice of the ROI was thereafter used as input for a second regression network that predicted the volume. Their method is different from our proposed method as they did not use the complete 3D scan to predict the volume and used an additional CNN for extracting the ROI. Moreover, the paper lacked a fair comparison to a fully supervised network.
Related to the WMH quantification with weak labels, Xie and Tao proposed a machine learning pipeline that learned with bounding boxes produced my non-experts. Based on the histograms of the image patches with lesions, they thresholded the WMH regions. Their methods produced relatively high Dice scores with an average of 0.71. Schlegl et al. formulated the difficulty of using image-level labels in medical imaging analysis. They claimed that networks optimized with volume based image-level labels learn features related to the targets in the image, but have a difficulty in medical fields as the target structures are small relative to the image sizes.

Our method was copied from the proposed method in Dubost et al.. Within this paper, however, the network was optimized with the number of occurrences of perivascular spaces instead of volume. As there were no voxel-wise annotations, they could also not do a comparison with a fully-supervised model. The produced attention maps showed reasonable localization.

Our experiments on simultaneous optimizing on different label types are a form of combined supervision, also known as hybrid supervision or mixed label supervision. In this kind of supervision, networks use multiple kinds of labels, e.g. both strong (i.e. voxel-wise) and weak labels. Often, the number of weak labels is several times larger than the number of strong labels. Arvaniti and Claassen proposed a method which optimized on few voxel-wise annotated images and many severity score rated images. Their method used both types for supervision or just the weak label supervision depending on the availability of the labels. The method created a confidence score based on the probability of the weak label supervised prediction. This was done to moderate the impact of the unsure prediction on the backpropagation. Subsequently, there are networks that use two kinds of weak labels. By adding a second weak label for supervision, the proposed methods added additional information for the network to learn. This often required some prior knowledge of the application. The effect was described as positive and significant in the methods with mixed labels. Jia et al. used multiple supervisions, i.e. the network was supervised after each convolution, with both classification and area constraint losses. This is also called deep supervision. They showed that the area constraints, binned in 5 relative sizes of 20, 40, 60, 80 and 100% of the image, had a significant improvement on the classification performance. As the areas were coarsely defined, they claimed that obtaining this label was not a big burden for the clinical experts. While they showed a significant segmentation and classification improvement over the baseline that used just one weak label, their method was still underperforming compared to the fully supervised method. The method of Hwang and Kim had both classification and localization tasks. For the latter, voxel-wise annotations were used to create a coarse loss for the distance between annotated and predicted pathology. While coarse annotations might create the same level of performance of the loss function, they did not examine if their method worked with coarse annotations only as well. In their method, the task branches shared most of the network layers. They claimed that early during training, the backpropagation of the localization loss had a negative influence, and that tweaking the weight of the classification and localization losses during training was important. Yoo et al. used an additional weak label, i.e. gender, to help the network predict the age from face images. They saw that if the network learned to predict both gender and age, the performance for the age prediction increased. Wang et al., using a similar method, mentioned that using the additional weak label helped with confounders. Chaganti et al. extracted semantic image information from Eletronic Medical Reports (EMR) to produce an additional feature vector as input, next to the orbital CT scans. The network's performance on predicting the optic nerve volumes increased significantly with this additional information and outperformed a multi-atlas method.

Most weak label based papers compared their methods to other methods that also optimized on weak labels. The papers that made a comparison with fully supervised methods for regression, classification and detection tasks, showed an increase or similar performance in favor of the fully supervised methods. In case of segmentation, the weak label networks underperformed significantly. Subsequently, we didn't find any papers that looked into the comparison of the generalization properties of the image-level label prediction between the label types. In our understanding, this is the first study to do so.

3

Clinical background

The proposed method was applied to the medical application of quantifying the imaging biomarker White Matter Hyperintensities. As multiple sclerosis lesions look very similar to WMHs on brain MRIs, datasets containing multiple sclerosis were used as external datasets for evaluation. The sections below describe the clinical importance and properties of both WMHs and MS lesions.

3.1. White Matter Hyperintensities

White Matter Hyperintensities, also known as leuko-araiosis (white, decrease in density) or white matter disease, was first described by Hachinski et al., and have a presumed vascular origin (Wardlaw et al., 2015). They appear hyperintense, i.e. increased brightness, on T2-weighted scans and hypodense, i.e. decreased brightness, on CT scans. As the cerebrospinal fluid (CSF) also appears white on T2-weighted scans, Fluid Attenuated Inversion Recovery (FLAIR) scans are preferred. A distinction is made between the occurrence in periventricular white matter (PVWM) and deep white matter (DWM). The severity score proposed by Fazekas et al., is based on these localization and the severity of the WMHs. Other studies, like the Rotterdam Scan Study, quantified the WMH by total volume per image scan (Ikram et al., 2017).
While this abnormality occurs more often in the aging population, there is still a high variety in severity (Wardlaw et al., 2015). Vascular risk factors, such as diabetes, smoking, and hypertension increase the prevalence of WMHs. WMHs have been a promising biomarker for all kinds of cerebral diseases (Chutinet and Rost, 2014). Au et al. (2006); de Groot et al. (2015) showed a strong association between WMHs and cardiovascular mortality. Debette and Markus compiled research based on population studies to show the relation between WMHs and the risk of stroke, cognitive and emotional dysfunction, dementia, and death. Wardlaw et al. claimed, based on cross-sectional and longitudinal studies, that WMHs help predict an increased risk of stroke, depression, death, and impaired movement. While there is also a strong correlation with vascular dementia and WMHs, the impact of WMHs on neurodegenerative diseases remains unclear (Debette and Markus, 2010). In the case of deep white matter hyperintensities (DWMHs), Khalaf et al.; Taylor et al. researched the correlation between WMHs and Late Life Depression (LLD). They found that the depression was not only more severe but also less sensitive to treatment.

3.2. Multiple sclerosis

Multiple sclerosis (MS) is the inflammatory demyelination of axons, and occurs mostly in young people (Rush et al., 2015). MS is an autoimmune disease. The symptoms depend on the localization and, in case of white matter MS, range from numbness, fatigue, loss of coordination, to visual impediments. Wardlaw et al. proposed to add the terminology of presumed vascular origin to the WMHs and therefore WMHs do not include multiple sclerosis. The WMHs and MS look, however, very similar on the MRI scans and are hard to distinguish. Note that the average age of patients diagnosed with either WMHs or multiple sclerosis differ considerably, as multiple sclerosis occurs mostly in younger patients whereas WMHs occur mostly in the older population.

4

Method

We propose a method based on the GP-Unet architecture (Dubost et al., 2017a) to train neural networks with the weak (i.e. image-level) labels. This network architecture enabled the neural networks that were trained with either weak or strong (i.e. voxel-wise) labels to be very similar. This had two advantages. First, a fair comparison of the performance could be made between the two network types. Second, it enabled us to propose a method in which this network optimizes on both label types simultaneously. The GP-Unet made use of a global pooling layer after a common U-net structure. Different kinds of global pooling methods are discussed in the next section. The GP-Unet is discussed more in detail in section 4.2. Different from the original paper, we proposed an adaptation of the regression part of the GP-Unet. This is explained in section 4.3.

4.1. Global Pooling methods

There are multiple global pooling methods used in Computer Vision and Medical Image Analysis (Bortsova et al., 2018; Chaganti et al., 2019; Yao et al., 2018; Zhou et al., 2016). Hwang and Kim showed better results for the global average pooling when compared to global max pooling. Kolesnikov and Lampert stated that max pooling underestimates and average pooling overestimates the foreground pixels. Zhou et al. showed that the global max pooling mostly focused on the discriminate part of the target object, while the global average pooling helped the network learn the extend of the target object. Their results found similar classification performances for both global pooling methods, but an improvement of the localization of the global average pooling. Overall there was an agreement that the global average pooling was a relatively good performing method to pool values to a lower dimension. Another reason for global average pooling came from the application towards volume prediction. The global average pooling layer sums all voxel values of the input feature map, and divides it by the total number of voxels. In other words, if the pooling layer is applied to a binary segmentation, an average voxel intensity of the segmentation is computed. As there is a linear relation between the average voxel intensity of the segmentation and the target volume, it made sense to use a global average pooling layer for volume prediction.

4.2. GP-Unet

The network used for this research was inspired by the GP-Unet paper from Dubost et al.. The GP-Unet has a U-net shape with additional global pooling and fully connected layers. The network architecture of the GP-Unet is shown in Figure 4.1. The U-net shape is an architecture that has been proven to work well for medical segmentation tasks (Ronneberger et al., 2015). It consists of a downsampling or encoding part where the input is downscaled to provide a larger field of view, and create feature maps that encode the meaningful features in a lower resolution. An appended upsampling or decoder part scales the feature maps to the dimension of the input image. Due to skip connections between layers of the encoder and decoder parts, the downsampled feature maps can be upscaled without a loss of resolution. This enables a U-net to segment and predict smaller structures. Dubost et al. appended a global pooling and fully connected layer to the output of the U-net to create a regression network with prior hidden segmentation layers. Their implementation was weak label supervised and used to predict the number of occurrences of perivascular spaces in brain MRIs. They showed that their method had a higher sensitivity in detection than saliency or intensity-based methods, or similar architectures without the upsampling part. In addition, the global pooling method is also used frequently in computer vision (Kolesnikov and Lampert, 2016; Kwak et al., 2017; Zhou et al., 2016).

The output of a U-net shape network is a segmentation. This segmentation often consists of near binary values. Hence, when a global average pooling layer is applied to the segmentation, it computes the sum of the voxels of the segmentation, divided by the sum of the total voxels, i.e. the average voxel intensity. If this value is properly scaled with the total number of voxels, one acquires the volume of the segmentation. This inspired us to use the GP-Unet method to directly quantify the volume of the target instead of the number of occurrences.

4.3. Order of convolution and pooling

The regression part of the GP-Unet consists of a convolotional layer and a global pooling layer. The ordering of these two layers can be interchanged. This resulted in two different methods to regress the U-net shape part of the network. This new architecture is shown in Figure 4.2. The first method involves projecting each feature map to a single scalar using global pooling. A fully connected layer makes from this vector a single output value. This was the main method used and was proposed by others (Dubost et al., 2019; Zhou et al., 2016). In the second method, feature maps are first convolved into one feature map. This combined feature map is thereafter merged by a global pooling layer into a single output value. This method could be more intuitive for volume quantification, as it first creates a segmentation or attention map from which it extracts the volume. Note that since only the ordering changed, the total number of parameters is still the same. For easy distinguishing, this method is hereafter called Global Pooling Aggregate (GPA).

4.4. Generating segmentation from attention maps

The image-level label optimized networks predict a scalar value. To be able to use this network for segmentation, an attention map was computed. The attention map was generated by multiplying the weights learned in the last fully connected layer with the feature maps,

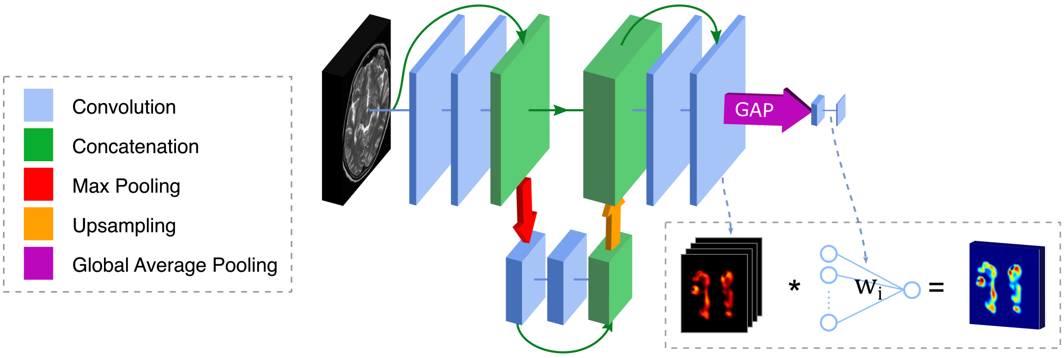


Figure 4.1: The proposed network architecture. It is a shallow U-net with concatenations and an appended regression part, which consists of a global average pooling and a fully connected layer. The weights learned in the fully connected layer are used compute a linear combination of the feature maps before the global pooling. This weighted multiplication creates the attention map of the network.

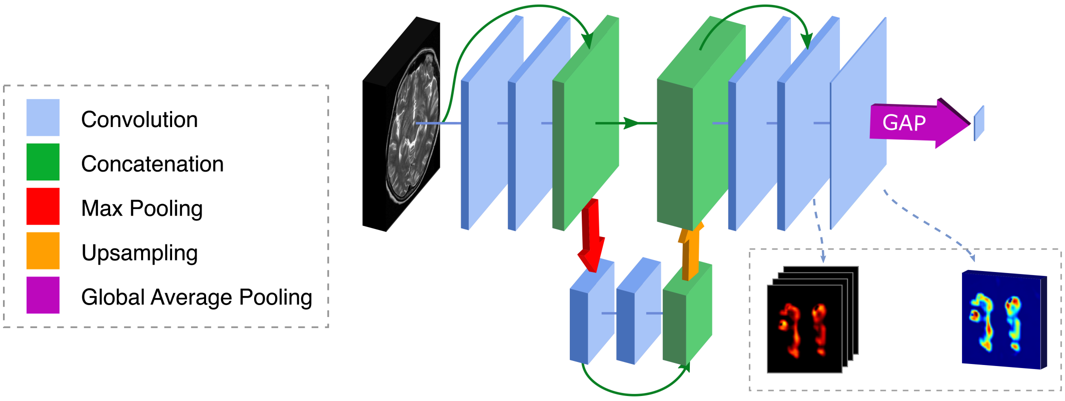


Figure 4.2: The proposed network architecture with a different ordering of the regression part (GPA). The regression part consists of a fully connected convolution to one feature map, and thereafter a global average pooling layer to a single output value. The single attention map before global pooling can be seen as the attention map of the network.

as shown in the dashed rectangular section of Figure 4.1. The resulting attention map is non-binary and hence has to be thresholded to create a segmentation. This threshold was chosen such that the volume of the segmentation corresponds to the predicted label.

5 Data

The characteristics of the datasets used for training and testing are given below. A simple, self-created dataset was created to test the proposed method in a controlled environment. Subsequently, WMH and MS lesions datasets have been used to evaluate the performance on a medical application and test the generalization properties. For easy comparison, an overview of the WMH and MS datasets is shown in Table 5.1. Lastly, a cardiac MRI dataset was used as a more heterogeneous dataset. This dataset helped to evaluate the proposed method on a harder medical application with heterogeneous tissue targets.

5.1. Artificial dataset

The performance of the proposed network can be sensitive to the kind of application, i.e. the medical dataset used for training. For a fundamental evaluation, simple artificial datasets were created. These 3D datasets consisted of foreground spheres of varying size and location on an empty background. The first dataset consisted only of one sphere per image. Due to varying size and location of the spheres, the spheres might or might not fit completely within the image borders. The second dataset consisted of two random spheres. These spheres had a different intensity, either one or two, and their intensities were summed on the points they intersect. Example images of these datasets are shown in Figure 5.1. For each dataset, 100 images with the size of $128 \times 128 \times 32$ voxels were generated. For both datasets, all voxels with the intensity of 1 were considered the targets. The datasets will be called Art1 and Art2 respectively in the rest of the report.

5.2. Rotterdam Scan Study

The Rotterdam Scan Study (RSS), acquired from Ikram et al., is a large population study from the Erasmus MC in Rotterdam, the Netherlands. For this study, a cohort of brain MRIs was used that consisted of 4336 FLAIR scans that have been acquired from the same number of patients using the same scanner and protocol (De Leeuw et al., 2001). The images were acquired from a 1.5 T GE scanner with a reconstructed voxel resolution of $0.49 \times 0.49 \times 2.5\ \text{mm}^3$. The images were pre-processed with skull extraction and bias field correction (Smith, 2002). The annotations have been made with an automatic thresholding algorithm proposed by De Boer et al. (2009). The annotations were thereafter corrected by multiple clinicians. Several example images with the corresponding annotations are shown in Figure 5.2.

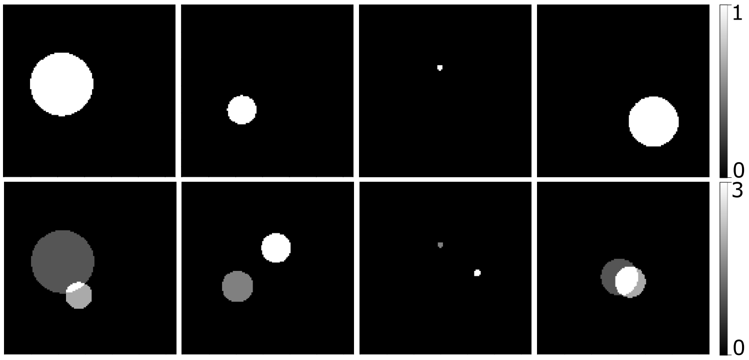


Figure 5.1: Example slices of the artificial datasets. The images were sliced in the middle of the third dimension. The upper row shows 4 examples of the first artificial dataset with just one sphere per image. The second row shows 4 examples of the second artificial dataset with two spheres per image. Note that they have a different intensity range.

5.3. Public datasets

To test generalization properties of the proposed networks, multiple public datasets available through segmentation challenges, were used. These challenges were centered around segmentation and/or classification tasks of annotated WMHs or MS lesions. Although the MS lesions are a different pathology than the WMHs, both look very similar on the MRI scans. Therefore, the public datasets with MS lesions were included as datasets to test generalization. The image modality was FLAIR for all datasets. Some datasets included different modalities as well, but these were not used in our experiments. Some challenges had a part of their data withhold. We used only the part released to the public. The sizes of these datasets were significantly smaller than that of the RSS, with dataset sizes of 15, 21 or 60 images. Due to small sizes, these datasets were only used for testing and not for training.

5.3.1. WMH Segmentation challenge

The WMH Segmentation challenge (WMHSeg) dataset, obtained from Kuijf et al., consisted of 60 FLAIR images and was published in 2017. The images were gathered from three sites with 20 images each. The patients originated from UMC Utrecht, NUHS Singapore and VU Amsterdam. Each site used a different scanner, a 3 T Philips Achieva, a 3 T Siemens TrioTim, and a 3 T GE Signa HDxt respectively. The voxel sizes ranged from 0.94 - 1.00 × 0.94 - 0.100 × 1.20 - 3.00 mm^3. The protocol settings of the repetition time, echo time and inversion time were not equal. The images were bias-corrected during pre-processing. The manual annotations were created by a clinical expert and peer-reviewed by a second expert. Example slices with these annotations are shown in Figure 5.3.

5.3.2. Longitudinal Multiple Sclerosis Lesion Segmentation challenge

The Longitudinal Multiple Sclerosis Lesion Segmentation challenge (LongMSLes) dataset, obtained from Carass et al., consisted of 21 images taken from 5 subjects. It was published in 2015. Four subjects had four scans taken over time, while the fifth had five. The scans were acquired via a 3 T Philips scanner with a voxel size of 0.82 × 0.82 × 2.2 mm^3. In pre-processing, the images were inhomogeneity-corrected and skull stripped. The annota-

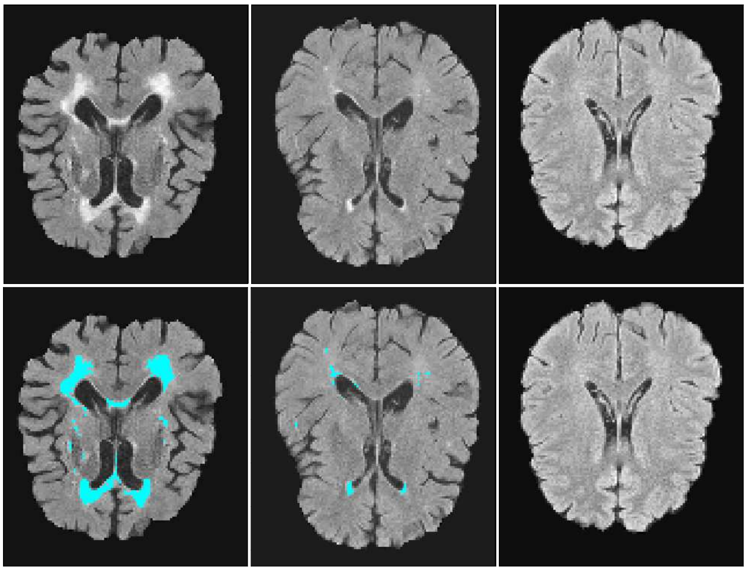


Figure 5.2: Example image slices of the RSS dataset. The images were sliced in the middle of the longitudinal axis. The upper row shows the image slices and the lower row the images with the corresponding WMH annotations as overlap in cyan. From left to right, the brain scans have a high severity of WMHs, a low severity of WMHs, and no WMHs.

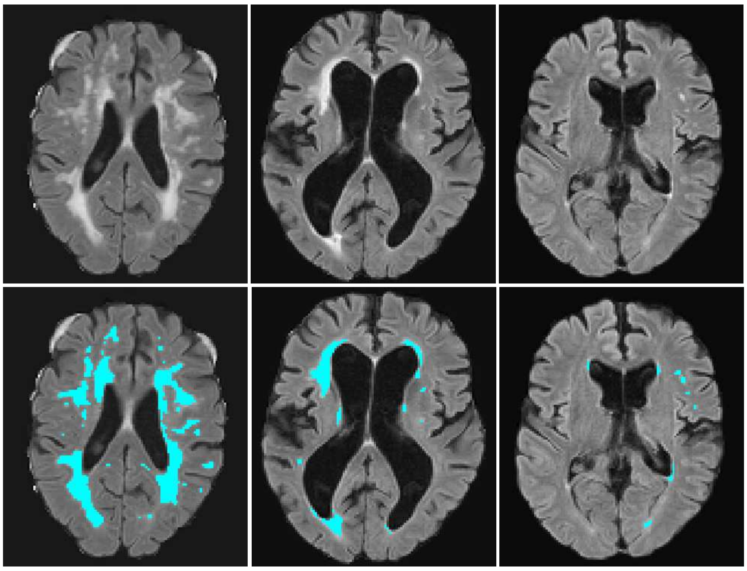


Figure 5.3: Example image slices of the WMHSeg dataset. The slices are in axial view and were sliced in the middle of the longitudinal axis of the original 3D scan. The upper row shows the image slices and the lower row the images with the corresponding WMH annotations as overlap in cyan. From left to right, the brain scans have a high severity of WMHs, a medium severity of WMHs, and a low severity of WMHs. Note that the example image on the left still contains some remnants of the skull. The skull stripping was not completely successful for every image.

tions were provided by merging the ratings of two clinical experts. Examples are shown in Figure 5.4.

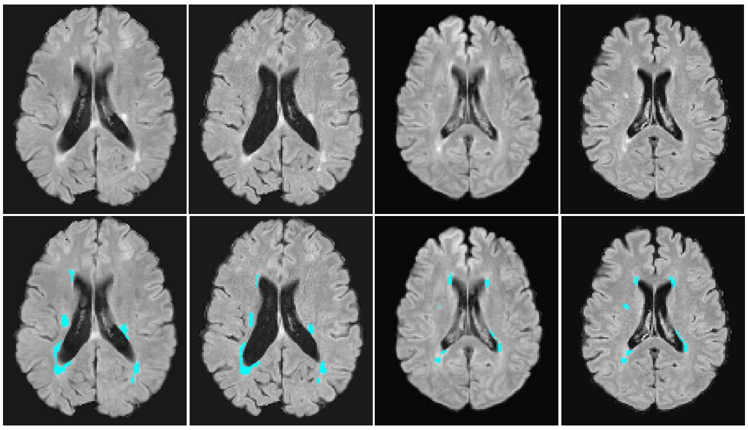


Figure 5.4: Example image slices of the LongMSLes dataset. The slices are in axial view and were sliced in the middle of the longitudinal axis of the original 3D scan. The upper row shows the image slices and the lower row the images with the corresponding multiple sclerosis annotations as overlap in cyan. The image scans of the first two columns and the last two columns are from the same subject. The first two show a subject with a high severity of MS lesions, and the last two show a subject with a low severity of MS lesions.

5.3.3. MS Segmentation challenge

The MS Segmentation challenge (MSSeg) dataset was acquired from Commowick et al., consisted of 15 scans, and was released in 2016. The scans were gathered from three different French hospitals and were acquired with three different scanners. The scanners were made by a 3 T Siemens Verio, a 1.5 T Siemens Aera, and a 3 T Philips Ingenia. The voxel spacing was 0.5 - 1.03 × 0.5 - 1.03 × 0.7 - 1.1 mm^3. Images were skull stripped and bias-corrected during pre-processing. The manual annotations were created by a consensus of seven clinical experts.

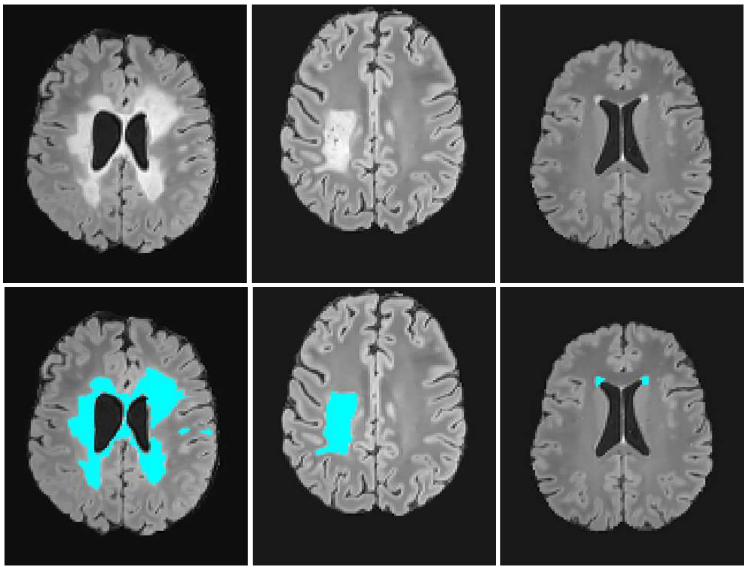


Figure 5.5: Example image slices of the MSSeg dataset. The slices are in axial view and were sliced in the middle of the longitudinal axis of the original 3D scan. The upper row shows the image slices and the lower row the images with the corresponding multiple sclerosis annotations as overlap in cyan. From left to right, the brain scans have a high severity of MS lesions, a medium severity of MS lesions, and a low severity of MS lesions.

5.3.4. MS Lesion Segmentation challenge

Styner et al. created the MS Lesion segmentation challenge (MSLes) dataset and released it in 2008. The 15 scans were acquired from two different hospital with a 3 T Siemens Allegra and a 3 T Siemens scanner. The voxel spacing was 0.5 × 0.5 × 0.5 mm^3. No preprocessing was done by the dataset creators. Therefore, BET (Smith, 2002) was used for skull stripping and bias-field correction. The annotations were made by either one or two clinical experts. Samples of the dataset with the annotations are shown in Figure 5.6.

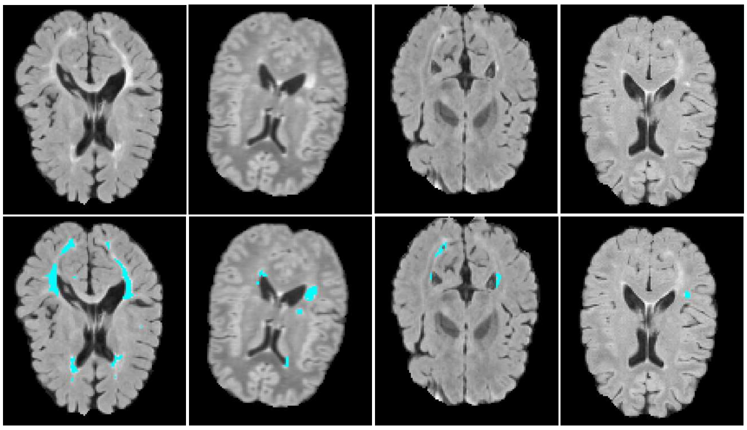


Figure 5.6: Example image slices of the MSL es dataset. The slices are in axial view and were sliced in the middle of the longitudinal axis of the original 3D scan. The upper row shows the image slices and the lower row the images with the corresponding multiple sclerosis annotations as overlap in cyan. From left to right, the brain scans have a high severity of MS lesions, a medium severity of MS lesions, and a low severity of MS lesions. The image scans of the last two columns have been made of the same subject.

Table 5.1: Overview of the properties of the WMH and MS datasets. (RSS = Rotterdam Scan Study, WMHSeg = WMH Segmentation challenge, LongMSLes = Longitudinal Multiple Sclerosis Lesion Segmentation challenge, MSSeg = MS Segmentation challenge, MSLes = MS Lesion Segmentation challenge.)

Dataset	# Images	Voxel spacing [mm^3]	Scanner	Sites
RSS	4336	0.49 × 0.49 × 2.5	1.5 T GE	1
WMHSeg	60	0.94 - 1.00 × 0.94 - 0.100 × 1.20 - 3.00	3 T Philips, 3 T Siemens, 3 T GE	3
LongMSLes	21	0.82 × 0.82 × 2.2	3 T Philips	1
MSSeg	15	0.5 - 1.03 × 0.5 - 1.03 × 0.7 - 1.1	3 T Philips, 3 T Siemens, 1.5 T Siemens	3
MSLes	15	0.5 × 0.5 × 0.5	2 types of 3 T Siemens	2

5.4. ACDC

The proposed method could be sensitive to different kinds of datasets. Therefore, instead of the homogeneous targets of the WMH and MS datasets, a heterogeneous cardiac MRI was used to evaluate the prediction of the proposed method on harder targets. The Automated Cardiac Diagnosis Challenge (ACDC) (Bernard et al., 2018) dataset consists of 100 MR images from the heart. All images are from separate subjects from Dijon, France. The MR images were annotated by one clinical expert, which provided the voxel-wise labels of the right ventricle, the left ventricle, and the myocardium. The data was created using two different Siemens scanners, one of 1.5 T and one of 3.0 T. The voxel spacing ranges between 1.37 mm and 1.68 mm. Per subject, multiple scans were made over a small time frame to cover (partially) the cardiac cycle. The first and last images of the cycle were extracted to create a dataset of 200 images. Example images are shown in Figure 5.7. All images from the ACDC dataset were resampled towards the median size of 144 × 144 × 4 voxels. The annotated images were stripped into binary label images of the left ventricular cavity, the right ventricular cavity, and the myocardium to create different target labels.

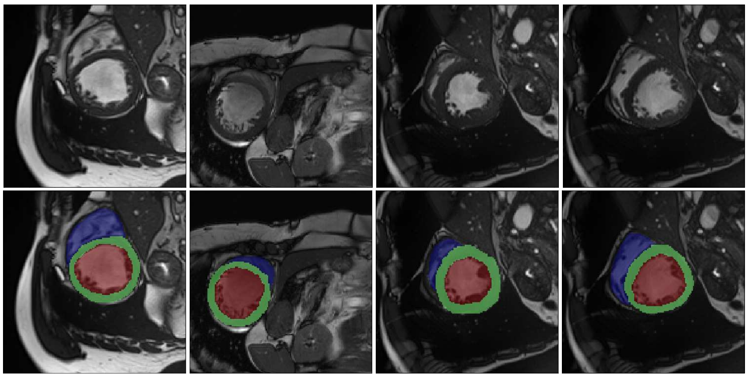


Figure 5.7: Example slices of the ACDC dataset. The images were sliced in the middle of the longitudinal axis. The upper row shows the image slices and the lower row the images with the corresponding annotations as overlap. The right ventricular cavity is shown in blue, the myocardium in green and the left ventricular cavity in red. Note that the third and fourth columns were images of the same subject but on different time points. Although the images were extracted from the same subject within a small time frame, the images showed a substantial difference in volume sizes and localization of the targets.

6
Implementation

When a measure becomes a target, it ceases to be a good measure.

Charles Goodhart

6.1. Preprocessing

The data was preprocessed to ensure less differences between images within and between datasets. As the RSS was the biggest dataset, its voxel spacing of 1.42 x 1.42 x 2.327 mm^3 was used as the default for the other WMH and MS datasets. The resampling was done with trilinear interpolation for the images and the nearest neighbor for the label images. After resampling, each image was cropped to only enclose the region of interest, consisting of the upper region of the brain above the eye border. The region was coarsely determined visually and the same regions were extracted from each image of a dataset. This resulted in image sizes of 112 x 128 x 32 voxels for the sagittal, frontal and longitudinal axis respectively. Percentile normalization was used to correct for intensity differences, i.e. the values were linearly scaled between 0 and 1 based on a low and a high threshold value of the image. To correct for outlier voxels, the 1% and 99% percentile values were used as threshold values. After normalization, the variance of the annotated voxel intensities varied with 0.057 with a mean of 1.04 over all five WMH and MS datasets.
The images of the ACDC dataset were normalized with percentile normalization (1%-99%) as well. There were no further preprocessing steps for the ACDC data.

6.2. Data augmentation

A data generator was used to create more training images using rigid data augmentation techniques. The methods used were translation (with a maximum shift of 20% of the length for each dimension), rotation ($\pm 54^\circ$ for each axis) and random mirroring for each dimension. For consistency, the data generator used the same random seed for all experiments to select and alter images. All train images were picked in a random order each epoch. The datasets were split in train, validation and test sets consisting of 60%, 20% and 20% of the images, respectively. The train data was used for training, the validation data was used to evaluate and keep the best network weights during training, and the test data was used for computing the performance after training. All splitting was done patient wise in case patients had multiple image scans. This prevented image scans of the same patient to be in both train and test sets and hence counteracted bias.

6.3. Network

The network was inspired by the GP-Unet paper from Dubost et al. The GP-Unet has a U-net shape with an additional regression part consisting of a global pooling layer and a convolution layer as detailed in section 4.2. The network's architecture is shown in Figure 4.1. Each convolution layer had a batch-normalization layer (based on the type of experiment, see also subsection 6.6.3) and a Rectified Linear Unit (ReLU) activation layer appended. The shallow default depth of two (i.e. just one max-pooling layer) was chosen as it enabled fast training with (almost) no difference in performance.

6.4. Hyperparameters

The input was padded with zero values so the output size of each convolution layer was the same as the input size. The number of feature maps after the first convolution was 32. The number of feature maps increased due to concatenation and max-pooling to a maximum of 128. Before the regression part, the number of feature maps was 32. The downscaling and upscaling was a factor of 2 in each dimension. The size of the kernels was 3 for each dimension. As optimizer, Adadelta with Keras' default parameters was used (Zeiler, 2012). Adadelta was designed to be less sensitive to the initial learning rate. Each batch contained 4 (augmented) images. The randomization of the data generator (see section 6.2) was fixed with a random seed, meaning that the experiments had the same order and the same augmentations of the images.

6.5. Loss function

For the regression networks, a mean-squared error loss was used defined as

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(a_i - p_i)^2, \tag{6.1}$$

in which N is the number of images, a_i the label, e.g. WMH volume, for the ith image, and p_i the predicted value for the ith image.

For the segmentation networks, a Dice score loss was used, which measures the overlap between the predicted segmentation and the annotations. The Dice score loss is defined as

$$\text{Dice score loss} = 1 - 2\frac{|A \cap P|}{|A| + |P|}, \tag{6.2}$$

in which A is the annotated segmentation and P the predicted segmentation. This holds however only for binary segmentation. For non-binary predictions, the Dice score loss can be rewritten in terms of voxels, as

$$\text{Dice score loss} = 1 - \frac{2\sum_{i=1}^{N} a_i p_i}{\sum_{i=1}^{N} a_i^2 + \sum_{i=1}^{N} p_i^2}, \tag{6.3}$$

in which i is the ith voxel of the image and N the total number of voxels in the image. This Dice score loss was used for evaluation of the segmentation networks, i.e. networks optimized with voxel-wise labels.

6.6. Evaluation

The performance of the networks was measured on test data. This data was not used for either training or validation (during training). To further assess generalization properties, other external datasets were used. The metrics used will be discussed in subsection 6.6.1 below. The metrics have been bootstrapped with 1000 runs to find the 5% and 95% confidentiality intervals. Bootstrap hypothesis testing was used to determine whether the difference in metric scores were significant. With bootstrap hypothesis testing, $1 \cdot 10^6$ runs were used. To measure the consistency, most experiments have been conducted twice or thrice with the same parameters but different random initialization of the network weights. For the experiments on WMH datasets, an additional control test was used to adjust for known confounders, discussed in subsection 6.6.2 below.

6.6.1. Metrics

Several metrics have been used to quantify to performance of the trained networks. To evaluate the prediction of WMH volume, the intraclass correlation (ICC) score (Koo and Li, 2016), the Pearson coefficient, the mean squared error (MSE), and the average volume difference (AVD) have been used. The Pearson coefficient measures the correlation (i.e. linearity) between the predicted and annotated volume. One disadvantage of the Pearson coefficient is that it doesn't measure scaling. Therefore the ICC score, in this case ICC (2,1) (Koo and Li, 2016), was also used. The AVD is defined as

$$\text{AVD} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|a_i - p_i\|}{a_i} \cdot 100\%, \tag{6.4}$$

in which N is the number of images, a_i the label, e.g. WMH volume, for the ith image, and p_i the predicted value for the ith image. It measures the relative difference between predicted and annotated volume while being independent of localization of the segmentation.

The segmentation networks were evaluated by using the Dice score. Note that the Pearson coefficient, ICC score and Dice score range from 0 to 1, with 1 the best possible score, while the MSE and AVD have a range of $[0, +\infty]$, with 0 the best possible score.

6.6.2. Confounders

There can be multiple confounding variables when trying to predict WMHs. The most obvious confounders are intracranial volume (ICV), ventricle volume (VV), white matter volume (WMV), and grey matter volume (GMV). Instead of learning to quantify the WMHs, the network could learn to predict one or multiple of these confounders and still end up with a reasonable prediction. To evaluate whether they influenced the predictions of the regression networks, linear regression with ordinary least squares was used to adjust for them. Their relationship was computed with the coefficient of determination and its significance was assessed by the p-value. The volumes of the confounders were extracted from the brain scans with the FreeSurfer registration tool (Desikan et al., 2006).

6.6.3. Influence of different setups

Performance of deeper and larger networks. A main concern of the proposed architecture was the shallowness. Many state-of-the-art network architectures have more depth, and thus more parameters and an increased receptive field (Kuijf et al., 2019). While the proposed network performs similarly to the highest-ranking methods in the WMH Segmentation challenge (Kuijf et al., 2019), experiments were conducted with increasing depth. We increased the depth of the GP-Unet in Figure 4.1 by an additional group of layers, similar to the lower U-net part, that was appended to the network just before the upsampling part. The architecture gained an additional max-pooling layer, a block of convolutions and concatenation, and an upsampling layer. Note that there is also an additional skip connection for the upsampling layer.

Influence of batch-normalization on training. Batch-normalization (Ioffe and Szegedy, 2015; Santurkar et al., 2018) is a wide-used and successful method for segmentation tasks, which normalizes the activation layer of the previous hidden layer. It could help with the hyperparameter search, makes the neural network more robust to changes in the hyperparameters, i.e. a larger range of hyperparameter options are suitable, and help deeper networks converge. Each convolution layer in the architecture proposed in section 4.2 had an optional batch-normalization layer appended. The influence of this layer was assessed for both voxel-wise label optimized and image-level label optimzed networks.

7

Experiments & Results

In this section, the experiments and their results are discussed. For easy reference, the terminology of weak labels was used for image-level labels, e.g. WMH volume, and strong labels for voxel-wise labels, i.e. segmentation. The experiments are discussed as follows. First, the performance on the artificial datasets is compared between different kinds of setups of both weak and strong label networks in section 7.1. The main focus of this research, however, was the quantification of WMH volume. The volume prediction of WMH volume by strong and weak label networks is discussed in section 7.2. Networks trained with both label types instead of one are discussed in section 7.3. Instead of WMH volume, other kinds of weak labels for WMH quantification are assessed in section 7.4. The generalization properties towards external datasets are shown in section 7.5. In section 7.6, it is shown that the weak labels networks were not influenced by obvious confounders. Section 7.7 details whether the weak label networks can be used for WMH segmentation. In section 7.8, the relationship between the performance of weak label networks with or without batch-normalization and the size of the dataset is assessed. Subsequently, harder targets of the ACDC dataset were used to evaluate the performance of the proposed method on heterogeneous tissue targets in section 7.9. Lastly, a more detailed evaluation of the relations between the metrics is discussed in section 7.10.

7.1. Artificial data

The artificial datasets provide simple targets in a controlled environment. They can help finding the properties and limitations of training with weak (image-level) labels. The first dataset (see section 5.1), Art1, was used for the following setups: weak, weak with additional depth (depth of 3, as the depth of 2 is the default), weak with batch-normalization, and strong with batch-normalization. The objective was to predict the total number of voxels of the sphere in the image. The second dataset, Art2, was used to compare the following setups: weak, strong, and strong with batch-normalization. For Art 2, the target volume consisted of all voxels with the intensity of 1. These voxels belonged to one of the two spheres in the image.

The performance on the test sets is shown in Table 7.1. Predicted segmentations (strong label networks) or thresholded attention maps (for weak label networks) on the test set are shown in Figure 7.1 and Figure 7.2 for Art1 and Art2, respectively.

For predicting the total volume, both weak and strong label networks performed similarly

Table 7.1: Performance of different experiment setups of networks trained on the artificial datasets. The best score per metric is highlighted in bold.

Experiment Setup	Pearson	AVD	ICC	MSE	Dice
Art1 - weak	**1.00**	17	**1.00**	459755	0.83
Art1 - weak with add. depth	**1.00**	15	**1.00**	352748	0.80
Art1 - weak with batch-normalization	0.99	106	**1.00**	4039416	0.59
Art1 - strong with batch-normalization	**1.00**	**0**	**1.00**	**0**	**1.00**
Art2 - weak	**1.00**	5	**1.00**	20254	0.91
Art2 - strong	0.68	142	0.81	173148601	0.77
Art2 - strong with batch-normalization	**1.00**	**0**	**1.00**	**0**	**1.00**

and achieved high metric scores. The weak label networks were significantly worse for segmentation. The result of the network with additional depth layers was a decrease in Dice score, while there was a increase in performance for the AVD. The difference was, however, not significant. The weak label setup with batch-normalization decreased the performance for the AVD and Dice scores, which implied that batch-normalization was not beneficial for weak label training.

A performance difference was expected between both datasets, as the Art2 dataset was slightly more complicated. The segmentation performance (Dice score) of the weak label network was, however, better when trained on Art2 than on Art1.

The worse performance on Art1 was mainly due to images with small volume labels. The images with low Dice scores for the Art1 experiments contained small spheres, and hence had a small volume. These volumes were smaller than 2200 voxels, while the average of the sphere volumes was around 19,000 voxels. The model had difficulty predicting these small volumes, and predicted 0 instead. The test set of 20 images included 3 images with small spheres that were wrongly predicted as empty. Although this is a relatively large part of the test set, this didn't reflect in the Pearson or ICC scores (both 1.00).

Art2 also contained images with small target volumes. The number of images with small volume labels and the size of the labels was comparable between both datasets. The weak label network trained on Art2 did a better job to predict the target volumes. While it also had the worst performance on images with small volume labels, the model never predicted 0 volume and therefore achieved significantly higher Dice scores. The strong label networks with batch-normalization achieved the highest and perfect performance.

Lastly, the strong label network without batch-normalization had trouble distinguishing the different spheres based on intensity. Instead of predicting the correct volume of the target sphere, it predicted the volume of all spheres combined. This can be seen as a local minimum from which it didn't deviate during training. From these experiments, it seemed that batch-normalization had a positive influence on the performance of the strong label networks, but a decreasing performance on the weak label networks.

7.2. WMH volume quantification: direct estimation versus segmentation

Within this section, the weak and strong label types have been compared for WMH volume quantification. The strong networks were optimized on the segmentation. During testing,

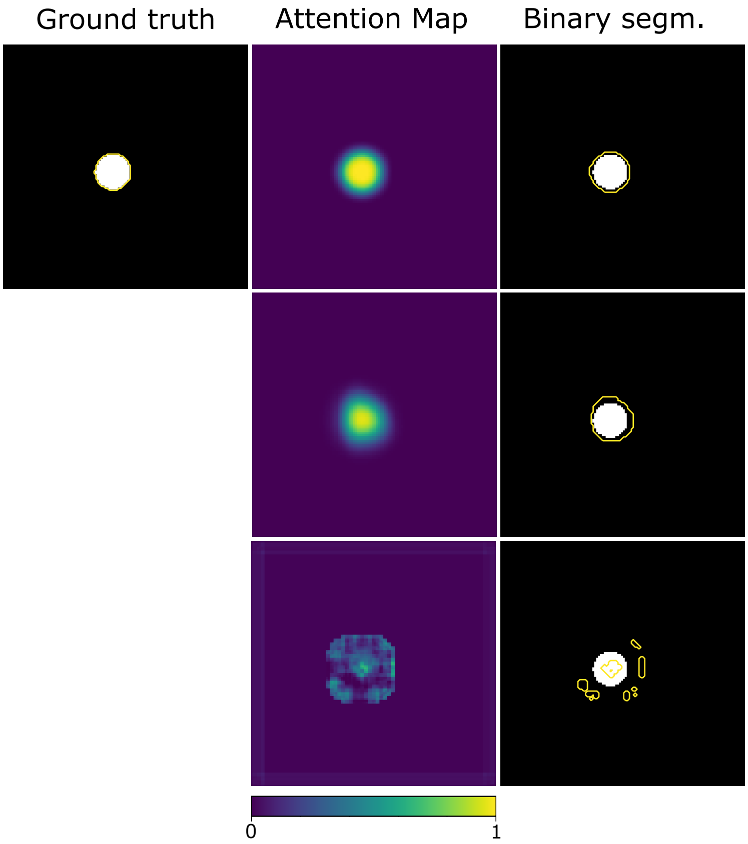


Figure 7.1: Attention maps created by weak label experiments on the Artificial 1 dataset. The ground truth is shown in the top-left cell. The second column shows the attention maps and the third column the binary segmentation created by thresholding the attention map. The rows depict, from top to bottom, the weak label experiment, the weak label experiment with additional depth, the weak label experiment with batch-normalization.

the volume was extracted from these segmentations. The weak networks, however, were directly optimized on the volume. While their labels contained less information, this direct optimization could outperform the indirect one. It is less prone to biases within the voxel-wise annotations and can learn other meaningful features that are not bound to the objects themselves to quantify the volume. The networks had been trained on the RSS dataset and thereafter tested on a separate test set. These results are shown in Table 7.2. The ICC score was slightly better for the weak label networks and the AVD was better for the strong label networks. The AVD was lower for the weak label networks as they overestimated images with a low target volume, resulting in relatively high AVD scores. In general, the weak label networks performed better on high target volumes, and the strong label networks performed better on low target volumes. The MSE scores were better for the weak label networks. This was expected as the weak label networks were directly optimized on this metric (see also section 6.5). The predictions of the weak and strong networks are shown in Figure 7.3. The decrease in volume prediction for the strong networks seemed to correspond with a slight underestimation. This is especially visible for the images with high severity of WMHs.

To validate the significance of the ICC score, bootstrap hypothesis testing (BHT) was used. As each experiment had three repetitions, BHT was used to evaluate the significance

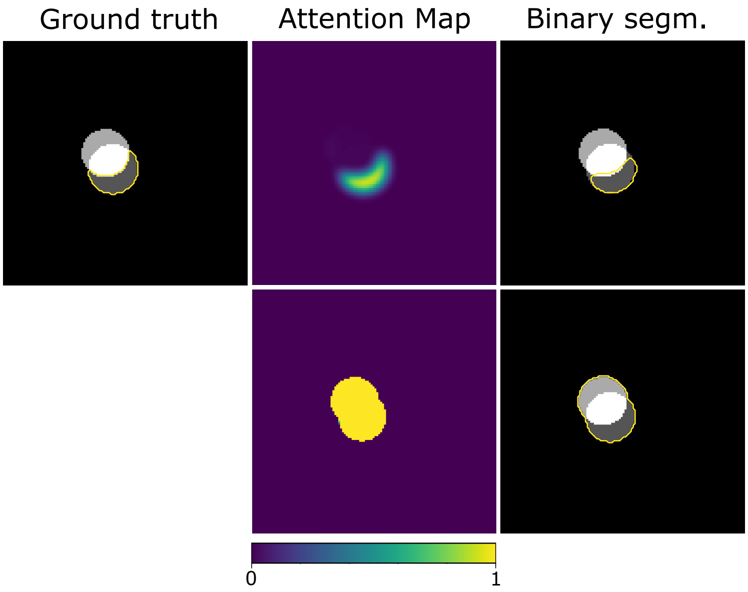


Figure 7.2: Attention maps created by experiments on the Artificial 2 dataset. The ground truth is shown in the top-left cell. The second column shows the attention maps and the third column the binary segmentation. The top row shows the weak label experiment, and the bottom row the strong label experiment without batch-normalization.

Table 7.2: Performance of the weak and strong networks on the RSS test set. The best score per metric is highlighted in bold. (rep. = repetition)

	ICC	Pearson	AVD	MSE
Weak rep. 1	**0.989**	**0.990**	36.1	**89984**
Weak rep. 2	**0.989**	**0.990**	26.5	90277
Weak rep. 3	0.987	0.988	25.5	112962
Strong rep. 1	0.977	0.988	18.5	175658
Strong rep. 2	0.974	0.987	**16.7**	200326
Strong rep. 3	0.975	0.987	17.2	194333

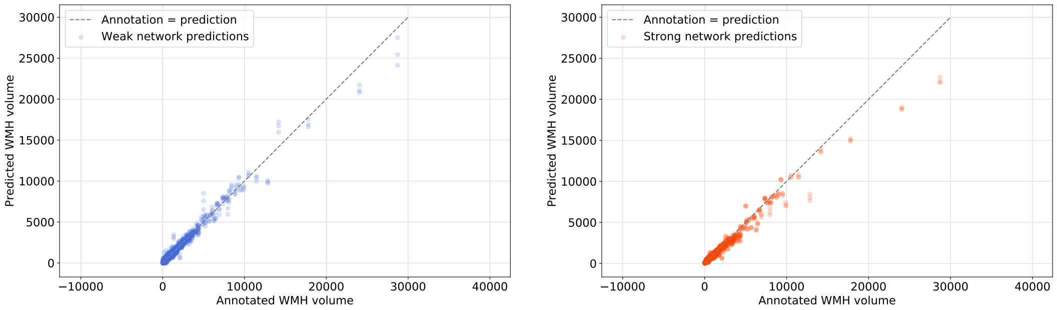
of the difference between each combination. The results are shown in Table 7.3. The p-value showed that the difference between the label types experiments was significant for all but one, i.e. considering p-values below the threshold of 0.05 as significant.

Table 7.3: The p-values of the Bootstrap hypothesis testing of the ICC score performance on the RSS test set of the three weak label networks vs the three strong label networks. A p-value higher than 0.05 means that there is no statistically significant difference. (rep. = repetition)

	Strong rep. 1	Strong rep. 2	Strong rep. 3
Weak rep. 1	0.014	0.011	0.012
Weak rep. 2	0.034	0.021	0.025
Weak rep. 3	0.053	0.031	0.039

7.2.1. The influence of different experiment setups

In the previous section, the networks had a depth of 2 and did not use batch-normalization. In this section, setups were evaluated with additional depth, the addition of batch-normalization, and the Global Pooling Aggregate (GPA) method (as described in section 4.3).



(a) The predictions of the weak networks.

(b) The predictions of the strong networks.

Figure 7.3: Predictions of the weak and strong networks versus the annotated WMHs of the RSS test set. Predictions of three repetitions are plotted per graph.

The ICC scores of the different setups are shown in Figure 7.4. All networks that used batch-normalization layers had a relatively high decrease in performance. The weak label networks, that were directly optimized on the WMH volume, had a slightly higher ICC score compared to the strong label networks. Subsequently, the ordering of the regression layers had no significant impact on the performance. Moreover, the deeper networks improved the performance in case of the strong networks. For the weak label networks, there was no higher performance, although this would also be hard to achieve as the ICC scores were near 1.

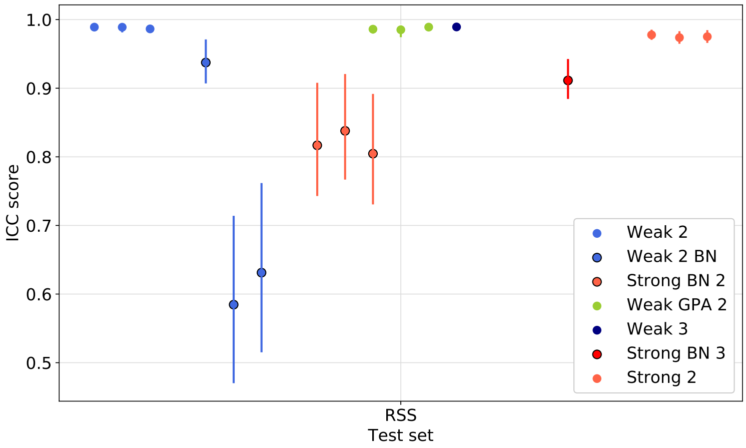


Figure 7.4: Performance of different experiment setups on the RSS test set. The weak label networks are shown in blue, the weak label network with different ordering of regression layers in green, and the strong label networks in red. If the network used batch-normalization, its points have a black edge. The number in the legend depicts the depth of the network. The bootstrapped upper and lower 95% bounds are shown as a line. Each setup had three repetitions, with the exception of the deeper networks.

7.3. Combined supervision

A benefit of the proposed network architecture is the possibility of combining the segmentation and regression part. This raised the question whether the network can learn from both parts simultaneously. Direct optimization on the regression part could have other benefits than the optimization on the voxel-wise labels of the segmentation part. The tasks are highly correlated, but could also regulate each other. On the other hand, the volume extracted from the strong labels is not added information to the network. It could extract this information from the voxel-wise labels during training itself. Within this section, experiments were conducted to verify whether training with both labels at the same time could increase the performance of the networks. For training, two different scenarios were evaluated: the simultaneous approach in which the network trained with both label types as input, and a sequential approach in which the network was pre-trained with one label type and fine-tuned with the other.

7.3.1. Simultaneous approach

The loss function of the network was altered to a combined loss function of the mean-squared error and the Dice score loss (see section 6.5). As these loss functions have different ranges, a scaling parameter α was used to bound the MSE in the range between 0 and 1. This combined loss for optimizing on weak and strong labels was defined as

$$\text{Weak-Strong Combined loss} = \alpha \frac{1}{N} \sum_{i=1}^{N} (a_i - p_i)^2 - \frac{1}{N} \sum_{i=1}^{N} \frac{2 \sum_{j=1}^{M} a_j p_j}{\sum_{j=1}^{M} a_j^2 + \sum_{j=1}^{M} p_j^2}, \qquad (7.1)$$

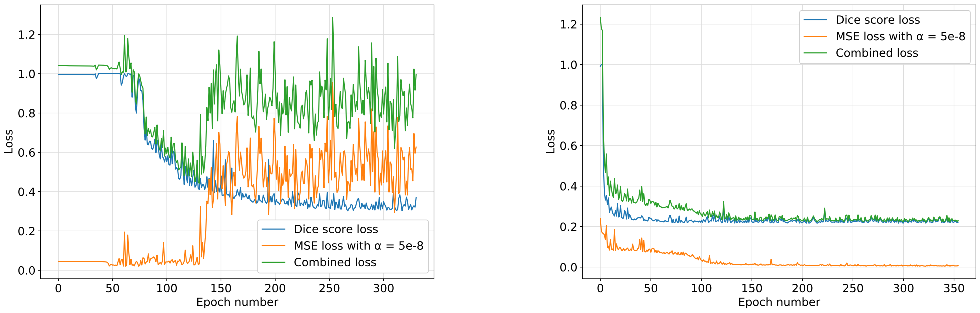
in which N is the number of images, i the ith image, M is the number of voxels, j the jth voxel of the image, and a and p are the annotated or predicted label, respectively.

Alpha was chosen by looking at the maximum MSE values from earlier experiments. Three different values have been used, $1 \cdot 10^{-7}$, $5 \cdot 10^{-8}$, and $1 \cdot 10^{-8}$. For these values, the MSE had a respectively high, medium, or very small influence on the weak-strong combined loss. The results of experiments with different dataset sizes are shown in Table 7.4. The training curves of the experiments with an α of $5 \cdot 10^{-8}$ are shown in Figure 7.5a (small dataset) and Figure 7.5b (big dataset).

For the smaller dataset size, the performance of the networks increased with a lower influence of the MSE. The best volume prediction, measured by ICC, Pearson and MSE, was achieved for network optimized with the big dataset of weak labels. In terms of segmentation, the combined supervision with $\alpha = 1 \cdot 10^{-8}$, i.e. a very small influence of the weak labels, had the highest Dice. The AVD was also the highest for this experiment. These results were not significantly better than the network trained with only strong labels. Figure 7.5a shows a negative influence of the MSE on the weak-strong combined loss while the Dice loss still decreases. This was not expected as the tasks should be more similar and should benefit of the additional regularization. Figure 7.5b doesn't show this sudden increase in MSE. However, there was still a plateau between the epochs of 30 and 70, where the Dice score loss was already optimized and the MSE stayed in a local optimum. While the network learned to achieve a better performance of MSE after this plateau, the Dice score loss stayed constant. Overall, there was no added benefit of using the combined supervision experiments in comparison to the strong label networks.

Table 7.4: Performance of networks trained with both label types simultaneously. The small datasets were trained with 78 images and the big dataset with 2602 images. The α is the scaling parameter for the weak-strong combined loss. The best score per metric is highlighted in bold.

Experiment Setup	ICC	Pearson	AVD	MSE	Dice
Small dataset, $\alpha = \infty$ (weak label)	0.212	0.682	89.6	3755408	0.008
Small dataset, $\alpha = 1 \cdot 10^{-7}$	0.299	0.559	27.8	3508723	0.585
Small dataset, $\alpha = 5 \cdot 10^{-8}$	0.382	0.623	28.3	3110495	0.638
Small dataset, $\alpha = 1 \cdot 10^{-8}$	0.727	0.886	22.5	1543791	0.731
Small dataset, $\alpha = 0$ (strong label)	0.756	0.911	20.9	1399379	0.735
Big dataset, $\alpha = \infty$ (weak label)	**0.989**	**0.990**	36.1	**89984**	0.139
Big dataset, $\alpha = 1 \cdot 10^{-7}$	0.956	0.985	17.0	310453	0.778
Big dataset, $\alpha = 5 \cdot 10^{-8}$	0.969	0.988	17.3	229977	0.780
Big dataset, $\alpha = 1 \cdot 10^{-8}$	0.958	0.985	**16.2**	297480	**0.785**
Big dataset, $\alpha = 0$ (strong label)	0.974	0.987	16.7	200326	0.784



(a) Network trained with a small dataset size.

(b) Network trained with a big dataset size.

Figure 7.5: Evaluation of the training and validation loss of networks trained with both label types simultaneous. The scaling parameter α was set to $5 \cdot 10^{-8}$.

7.3.2. Sequential approach

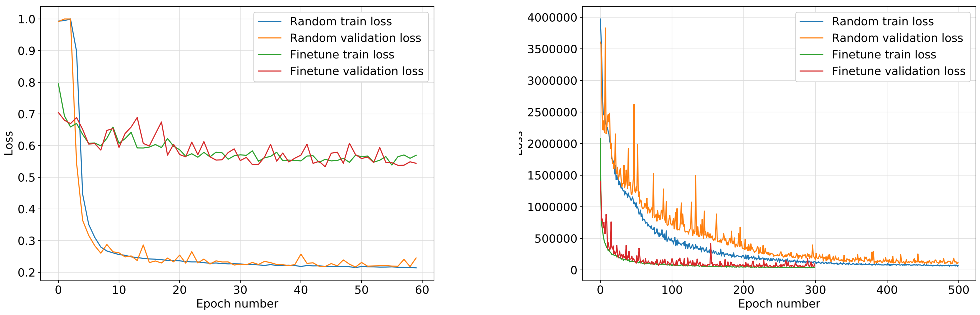
Instead of optimizing the networks with both labels during training, in the sequential approach the networks were pre-trained with one label type, e.g. weak labels, and thereafter fine-tuned with the other, e.g. strong labels. Pre-trained networks could be beneficial as they are strong initializations for further training. Their weights already extract image features related to the sequential task, which could help convergence.

Two experiments with the sequential approach were conducted. In the first experiment, the network was pre-trained with weak labels and fine-tuned with strong labels. In the second experiment, the network was pre-trained with strong labels and fine-tuned with weak labels. The results of both experiments are shown in Table 7.5. The first experiment with weak label pre-training and strong label fine-tuning had a decrease of performance for all metrics compared to the randomly initialized networks trained with strong labels. The training curves are shown in Figure 7.6a. While the pre-trained network had a head start with good initialization, it failed to converge to a better optimum and stagnated. The experiment with strong label pre-training and weak label fine-tuning achieved better volume quantification (i.e. ICC, Pearson, and MSE) and better segmentation performance (i.e. Dice) compared to the network without pre-training. The training curves, shown in Figure 7.6b, indicated a faster convergence than networks with random initialization and towards a better optimum.

This indicated that good initialization helped the network with both convergence speed and performance. While the strong labels were beneficial for pre-training, it seemed that the weak label pre-trained networks learned image features that were not beneficial for segmentation.

Table 7.5: Performance results of the sequential fine-tune experiments on the test set. As a comparison, the scores of a weak label network and a strong label network without pre-training were added. The best score per metric is highlighted in bold. (Pt. = Pretraining, ft. = fine-tuned.)

Experiment Setup	ICC	Pearson	AVD	MSE	Dice
Pt. with weak labels, ft. with strong labels	0.606	0.772	31.1	2142347	0.467
Pt. with strong labels, ft. with weak labels	**0.991**	**0.992**	21.6	**75100**	0.434
Strong label experiment without pt.	0.974	0.987	**16.7**	200326	**0.784**
Weak label experiment without pt.	0.989	0.990	36.1	89984	0.139



(a) Pre-trained with weak labels and fine-tuned with strong labels.

(b) Pre-trained with strong labels and fine-tuned with weak labels.

Figure 7.6: Train and validation curves of the sequential fine-tuning experiments. The blue and orange lines show the train and validation curves from the networks with random initialized weights, and the green and red lines the train and validation curves of the networks pre-trained with the other label type. On the left, the networks were trained with strong labels and the Dice loss, and on the right, the networks were trained with weak labels and the MSE loss.

7.4. Other weak labels for WMH volume quantification

Most weak label experiments were about WMH volume quantification and hence trained on the volume labels. However, there are multiple kinds of image-level labels which can be available for training. As detailed in section 4.2 and Figure 4.1, our proposed method is especially suitable for volume prediction. It would therefore be interesting to see whether the weak label network could learn to predict the WMH burden while trained with other kinds of weak labels. In the following sections, the number of WMH lesions and a self-created severity score were used as weak labels to train a weak label network. In subsection 7.4.2, the network was optimized with multiple weak labels to see whether the network could benefit from both kinds simultaneously. The Dice score evaluation was omitted in this section.

7.4.1. Number of WMH lesions as weak label

In this case, the number of connected components, i.e. the number of lesions or WMH regions, were extracted and used as weak training labels. The extracted number of components were not linearly correlated with the annotated volume, as shown in Figure 7.8. Therefore, this image-level label could be harder to optimize on for the network.

The trained network with the number of WMH lesions was able to reproduce the number of WMH regions on the test set with an ICC score of 0.894. There were no outliers, but it had a wider spread compared to the WMH volume quantification experiments. Examples of the network's attention maps are shown in Figure 7.7. The network found some correct localization of the WMH regions, but also many false positives (an example in the upper row) and false negatives (an example in the bottom row).

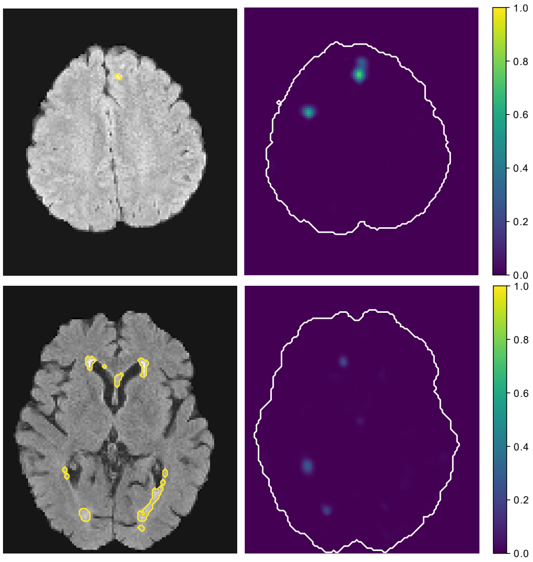


Figure 7.7: Attention maps produced by networks trained with the number of WMH regions. The left column shows the annotated WMHs, the right column the corresponding attention maps.

7.4.2. Combined weak labels

While both image-level labels were extracted from the same manual annotations, they had a relatively low Pearson correlation of 0.43. The relationship between both label types is shown in Figure 7.8. As the labels depict different kinds of information about the WMH burden, it raised the question if the network's performance could improve when optimized on both labels.

We created an architecture in which the U-net-like network had two instead of one regression branches, i.e. one for each weak label. Hence, both regression outputs shared the input layers and split after the global average pooling into two separate fully connected layers. This could force the network to learn image features that were feasible for both weak labels. Both labels were optimized with the MSE, but as the volumes were larger than the numbers of components, a scaling parameter α was introduced to make the sizes of the MSE more equal. This resulted in the loss function

$$\text{Weak-Weak Combined loss} = \frac{1}{N}\sum_{i=1}^{N}(a_{i,vol} - p_{i,vol})^2 + \alpha\frac{1}{N}\sum_{i=1}^{N}(a_{i,num} - p_{i,num})^2, \quad (7.2)$$

where vol denotes the volume labels and num denotes the number of components labels.

The results of the experiments, tested on the RSS test set, are listed in Table 7.6. The performance of the WMH volume prediction of the networks optimized with two image-level labels was slightly worse compared to the networks trained with only the WMH volumes as labels. The performance of the ICC Score, the Pearson coefficient and the MSE were lower, while the performance of the AVD was slightly higher. Even for smaller datasets, where the additional information from the second image-level label could help against the information scarcity, the performance did not improve.

Table 7.6: Performance of the networks on the RSS test set optimized with two different image-level labels: WMH volume and WMH lesions (number of connected components). The big dataset consisted of 2602 train images and the small dataset consisted of 260 images. The α is a scaling parameter to equalize the MSE of both labels. For comparison, the networks trained with only the WMH volume as a weak label are added. The best score per metric is highlighted in bold.

	ICC	Pearson	AVD	MSE
Big dataset, $\alpha = 8 \cdot 10^3$	0.981	0.986	38.7	148328
Big dataset, $\alpha = 4 \cdot 10^4$	0.984	0.985	**29.2**	135571
Small dataset, $\alpha = 4 \cdot 10^4$	0.843	0.881	94.4	1090038
Big dataset, volume label only	**0.989**	**0.990**	36.1	**89984**
Small dataset, volume label only	0.905	0.916	54.0	720211

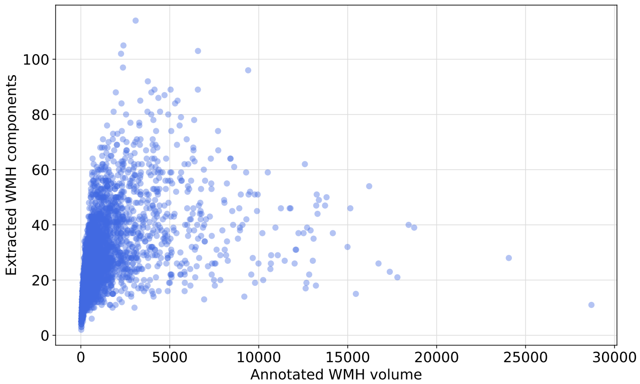


Figure 7.8: Relationship between the WMH volume and the number of components. The Pearson coefficient was 0.43.

7.4.3. WMH Severity Score as a weak label

In clinical practice, a severity score is often used to help with diagnosis. An example, in the case of WMH, is the Fazekas scale (Fazekas et al., 1987) The Fazekas scale is a visual scoring system describing the localization and the severity, i.e. the size and number, of WMHs in the brain scan. The results from section 7.2 showed that direct optimization on

the task improves performance. This raises the question whether it is possible to train a network directly on a severity score. Unfortunately, Fazekas scores were not available in our dataset and hence we created labels that resemble Fazekas scores. These labels were computed from the WMH annotated scans. A linear scale was used to categorize each scan to a number from 0-4 based on the total volume of the WMHs. This could be interpreted as a downsampled version of the weak labels used earlier. As there are only 5 values the label can take, there is a huge loss of information in comparison to the total volume. These arbitrary labels were created for the RSS and used to optimize a weak label network. As a linear scale was used, images with a large WMH volume were omitted from training and testing. This resulted in a training set of 1997 images, i.e $\sim 77\%$ of the total training set.
On the test set, the trained network achieved an ICC score of 0.92 between the predicted and created severity labels. As the model predicted continuous labels between 0 and 4, the predicted values could be scaled back towards a volume. The ICC score between the annotated WMH volume and the scaled severity score prediction was 0.92, the Pearson coefficient was 0.94 and the AVD was 43%. Although the performance was lower than the WMH prediction, this still showed the possibility of training good performing networks with just a severity score.

7.5. Generalization properties of different label types

The generalization was measured by evaluating the performance of models trained on the large RSS dataset on public datasets. Various setups of weak label trained networks were compared with the various setups of strong label trained networks. For each variant, three networks trainied with different random weight initialization were applied to this unseen data.

The networks trained on the RSS were applied on four public datasets to predict the WMH volume. Several different experiment setups were compared to test the influence of label type, batch-normalization, network depth, and in case of the weak labels, the ordering of the regression layers (GPA, see section 4.3). These were the same networks used for the volume quantification on the RSS test set in section 7.2.
The performance in terms of ICC scores between the annotated and predicted volumes for the different label types, as well as the influence of batch-normalization, is shown in Figure 7.9. The weak networks without batch-normalization seemed to perform better on the MSSEG and the MSLes datasets. For the LongMSLes dataset, the strong network without batch-normalization had a slight edge. On the WMHSeg the difference was relatively small. Bootstrap hypothesis testing showed no significant difference between the performance of the weak and strong label networks. Overall, the network with batch-normalization had worse generalization performance than those without.

The comparison between the weak label networks is shown in Figure 7.10. The networks with batch-normalization had a significantly worse performance. The other networks had similar performance and were not significantly different.

The comparison between the strong label networks is shown in Figure 7.11. These results show that also the strong networks perform better without batch-normalization. Subsequently, an additional depth layer was not beneficial. These deeper networks performed either similarly or worse.

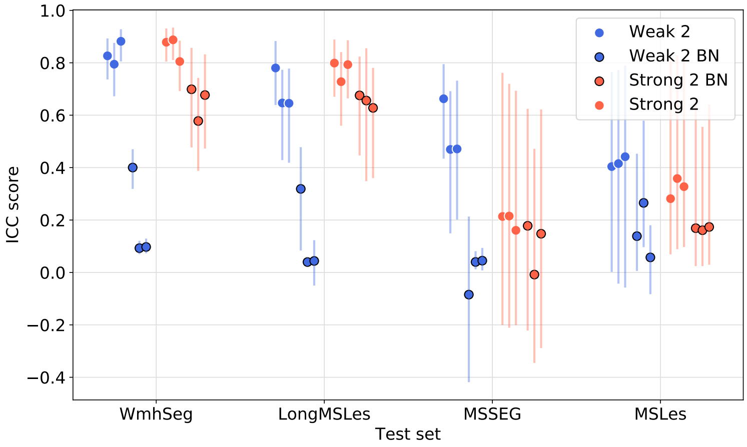


Figure 7.9: Generalization performance of volume quantification for networks with different label types and either batch-normalization or not. The depth of the networks was the default value of two. The bootstrapped upper and lower 95% bounds are shown as a line. Each setup had three repetitions, i.e. trained with three different random initializations of the network weights.

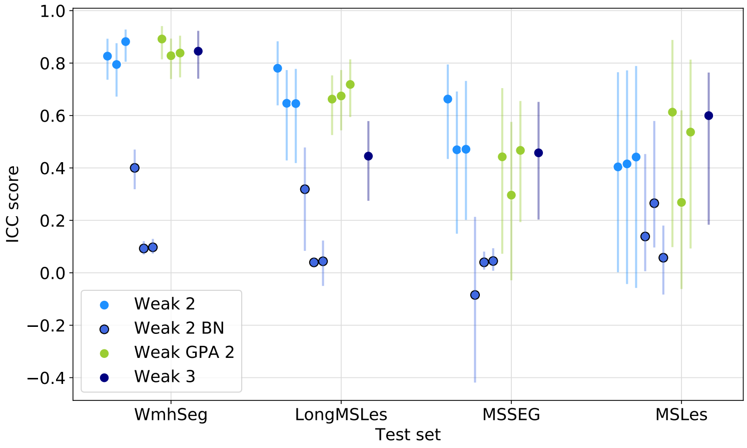


Figure 7.10: Generalization performance of volume quantification for weak label networks. The weak label networks are shown in blue and the weak label networks with a different regression layer (GPA) ordering are shown in green. If the network used batch-normalization, its points have a black edge. The number in the legend depicts the depth of the network. The bootstrapped upper and lower 95% bounds are shown as a line. Each setup had three repetitions, with the exception of the deeper networks.

7.6. Influence of confounders

The weak label trained networks could be prone to several confounding variables. Instead of focusing on the WMH regions, it could learn to focus on these confounders instead. For WMH, the most obvious confounders exist of white matter volume (WMV), ventricle volume (VV) and intracranial volume (ICV).

After adjusting for these confounders, the test (proposed in subsection 6.6.2) showed still a significant correlation between the predictions and the annotated ground truth with a Pearson coefficient of 0.99, a coefficient of determination of 0.98 and a p-value smaller than $1 \cdot 10^{-4}$. The confounders showed a lower correlation with the annotations. The Pearson coefficients were 0.47 for the WMV, 0.41 for the VV, and 0.47 for the ICV. Therefore, the

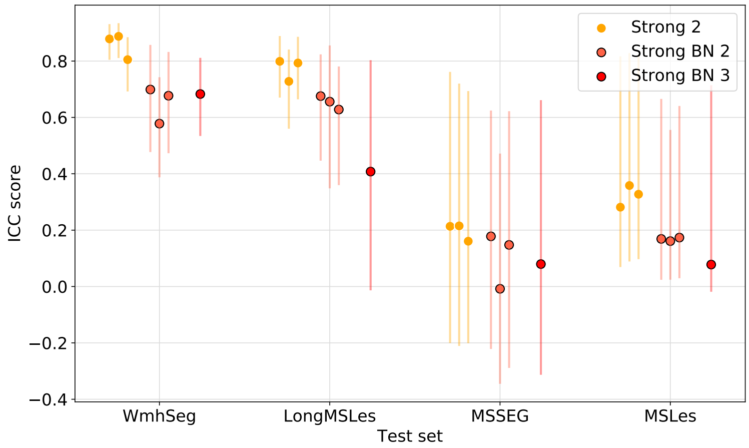


Figure 7.11: Generalization performance of volume quantification for strong label networks. The more shallow strong label networks are shown in orange and deeper strong models in red. If the network used batch-normalization, its points have a black edge. The number in the legend depicts the depth of the network. The bootstrapped upper and lower 95% bounds are shown as a line. Each setup had three repetitions, with the exception of the deeper network.

weak label network learned other meaningful image features and did not focus significantly on confounding variables.

7.7. Weak labels for WMH segmentation

Although weakly generated segmentations were not the focus of the experiments, the attention maps created could be thresholded to a binary segmentation. These can be used to compute a Dice score and evaluate whether image-level labels can be used for voxel-wise label prediction. The weak label networks trained on the RSS were used to evaluate the segmentation produced by weak label supervision. Figure 7.12 shows segmentations produced by both strong and weak label networks. The segmentations created by thresholding the attention maps of the weak label networks without batch-normalization had relatively low Dice scores with a maximum of 0.613 and an average of 0.25 ± 0.13. A weak label network with batch-normalization achieved a maximum Dice score of 0.730 and an average of 0.36 ± 0.13. It seemed that batch-normalization had a worse effect on the volume quantification (as discussed in section 7.2), but improved the segmentation.

The attention maps in Figure 7.12 show that the weak label network focused on the WMH regions, but did not attempt to delineate them. The attention maps had often a small shift and seemed to focus on regions around the WMHs themselves. This could also indicate that the weak label networks learned other meaningful image features to predict the WMH volume. However, the weak label networks could not be used to segment WMHs. High Dice scores were not possible without adding additional constraints or more information per label.

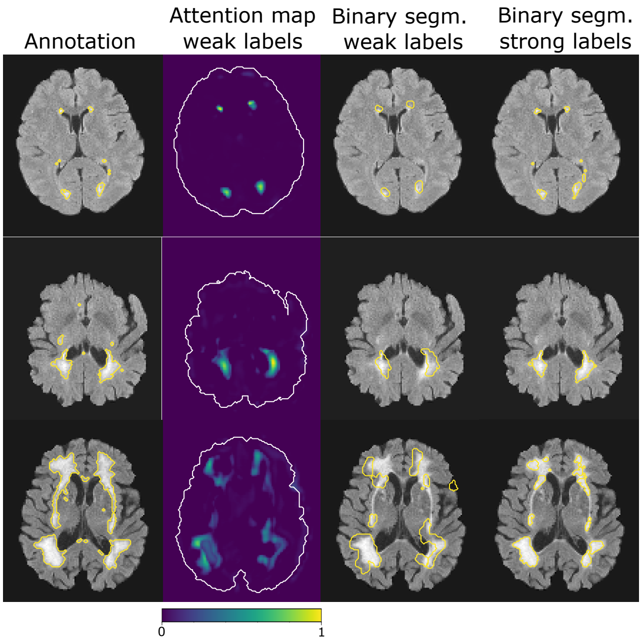


Figure 7.12: **Examples of segmentations on the RSS test set**. The rows show different subjects. The columns from left to right show an axial slice of the MRI scan overlaid with the annotated WMH contours, the attention map of a network trained with weak labels (contour of the brain shown), the corresponding predicted segmentations overlaid on the MRI scan and the last column shows the predicted segmentations by a network trained with strong labels overlaid on the MRI scan. The outputs of the networks trained with strong labels were almost binary due to the sigmoid activation, and hence were almost identical to the binary segmentations shown.

7.7.1. Different voxel-wise loss function for fair segmentation evaluation

The segmentation of both weak and strong label networks were evaluated with the Dice score, while the strong label networks were optimized with the Dice score loss as well. To make a more fair comparison, the strong label networks were optimized with another voxel-wise loss function, i.e. binary cross-entropy loss. The performance on the test set was similar as the networks optimized with the Dice loss. The differences between the networks were not substantial in terms of ICC score, Pearson, AVD or Dice score.

7.8. Batch-normalization dependency of dataset sizes

The influence of batch-normalization was examined more in-depth in this section. On the RSS, the weak label networks seemed to perform worse with the additional batch-normalization layers, whereas earlier experiments conducted on smaller datasets (not part of this report) indicated that batch-normalization was beneficial. Therefore, the weak label network with a depth of 2 was used as a baseline and its performance with and without batch-normalization was evaluated as a function of the dataset size. The results are shown in Figure 7.13. Each network was repeated once with a different random initialization for consistency.

For the larger dataset sizes, the results showed indeed a better volume quantification for the experiments without batch-normalization. However, as the dataset sizes grew smaller, the networks with batch-normalization had a better performance with a turning point at 120 images.

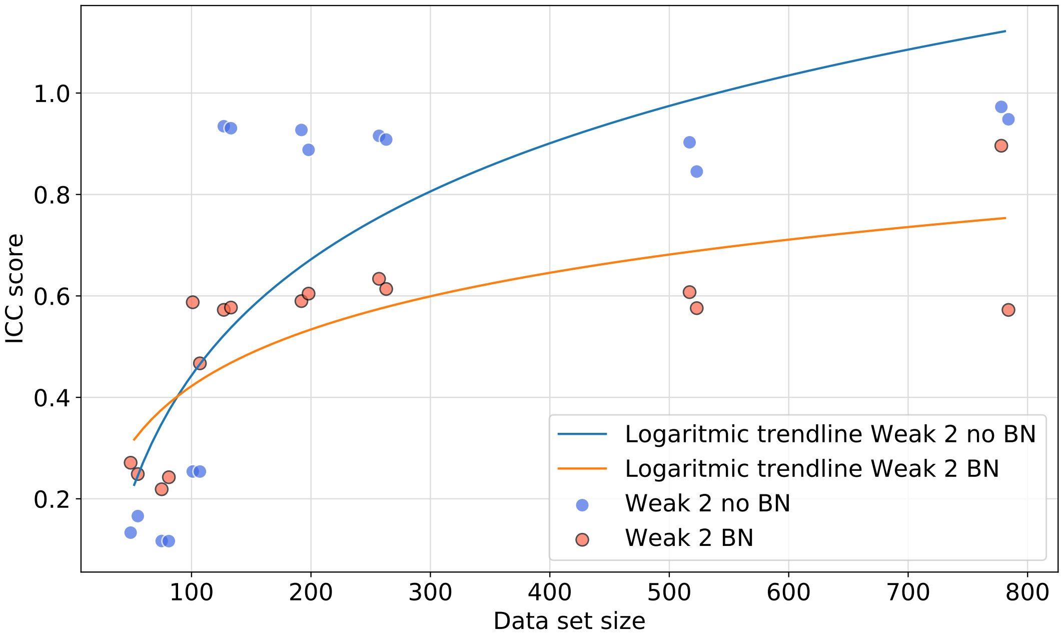


Figure 7.13: Performance of weak label networks with and without batch-normalization as a function of the dataset size. The networks without batch-normalization are shown in blue, the one with are shown in orange. Each experiment was conducted twice. For visualization purposes, the two points are slightly separated on the horizontal axis. They have however the same amount of training images. For clarification, the *Weak 2* denotes that it was a weak label network with a depth of 2. (BN = batch-normalization.)

7.9. Heterogeneous tissue targets (ACDC) for weak label training

All previous experiments tried to predict WMH or MS lesion volumes. These structures are relatively easy to distinguish from surrounding tissue; the WMHs or MS lesions are hyper-intense and homogeneous. Therefore, the ACDC set was used to judge if the proposed method also worked on other clinical applications with different tasks. The challenge of the ACDC set is the heterogenous tissue of the target regions. On the other hand, each image has just one target, whereas the RSS dataset contains many targets.

The volume of the left ventricular cavity (LVC) was used as the target. The annotated labels were used to compute the volumes of the left ventricular cavity. These image-level volume labels were used for training weak label networks. The network setups were varied with label types, network depth, the use of batch-normalization, and, in case of the weak label networks, the ordering of the regression layers (GPA). Results from the aforementioned experiments are shown in Table 7.7. The weak label networks had a poor performance predicting the LVC volume. The GPA method had an worse influence on the metrics. This is in contrast with earlier findings of the WMH volume prediction with weak GPA networks. On the other hand, the additional depth layer helped the weak label networks substantially. Batch-normalization had not a big impact on the end results.

The strong label networks had a relatively high performance on predicting the LVC volume. Deeper strong label networks had the best performance. One interesting aspect was the influence of batch-normalization for these strong label networks. The shallower networks had a decrease in performance due to batch-normalization. For the deepest network, however, the network did not converge without the batch-normalization layers. This might be due to the size of the network, i.e. the number of parameters to learn.

These results indicated that the weak label networks were more sensitive to the application than the strong label networks. For this hard case, the weak label networks could not find the correct target volumes in the heterogeneous data. Although the strong label networks had also more difficulties with predicting the LVC than the WMHs in previous experiments, they had relatively high performance. The network depth had a bigger positive impact on the performance on this dataset than the WMH datasets. An explanation could be that due to the complexity of the target and surrounding tissue, the network needed to be larger, i.e. more layers, to extract the right image features.

The attention maps (in case of weak label networks) or the segmentations (in case of the strong label networks) are shown in Figure 7.14. The attention maps of the weak label networks indicated why the performance was poor. The weak label network without batch-normalization focused on the wrong tissue regions, in this case the left ventricle and fat/skin tissue on the left part of the image. The weak label network with batch-normalization produced an attention map that didn't show a clear focus. The segmentations of the strong label network with a depth of 3 included some false positive regions, but found the contours of the target. The usage of batch-normalization seemed not a big factor for the localization performance of the shallower networks. For the depth of 4, the network without batch-normalization didn't find the target, which is clearly visible by the segmentation that focused on the dark surrounding tissue. The segmentation of the network with batch-normalization showed less false positives and improved the contour localization compared to the shallower networks.

Table 7.7: Performance results of different setups on the ACDC dataset with the LVC as target. The best performance per metric is shown in bold. (BN = batch-normalization, GPA = Global Pooling Aggregate, *Weak labels only)

Label type	Depth	BN	GPA*	ICC	Pearson	AVD	MSE	Dice
Weak	2	No	No	0.112	0.232	61.1	3681653	0.032
Weak	2	Yes	No	0.109	0.421	57.0	3803541	0.078
Weak	2	No	Yes	0.034	0.099	61.7	3603139	0.000
Weak	2	Yes	Yes	0.000	0.160	95.6	15329028	0.002
Weak	3	No	No	0.366	0.446	59.2	3029672	0.059
Weak	3	Yes	No	0.323	0.632	50.4	2684648	0.149
Weak	3	No	Yes	0.005	0.050	60.3	3812212	0.002
Weak	3	Yes	Yes	0.001	0.311	95.7	15347302	0.000
Strong	2	No	-	0.537	0.609	41.8	2720944	0.570
Strong	2	Yes	-	0.457	0.658	43.8	3610647	0.621
Strong	3	No	-	0.750	0.907	**26.5**	1494108	0.793
Strong	3	Yes	-	0.741	0.901	28.9	1532117	0.806
Strong	4	No	-	-0.004	-0.145	1230.7	835722538	0.104
Strong	4	Yes	-	**0.751**	**0.940**	30.5	**1492674**	**0.861**

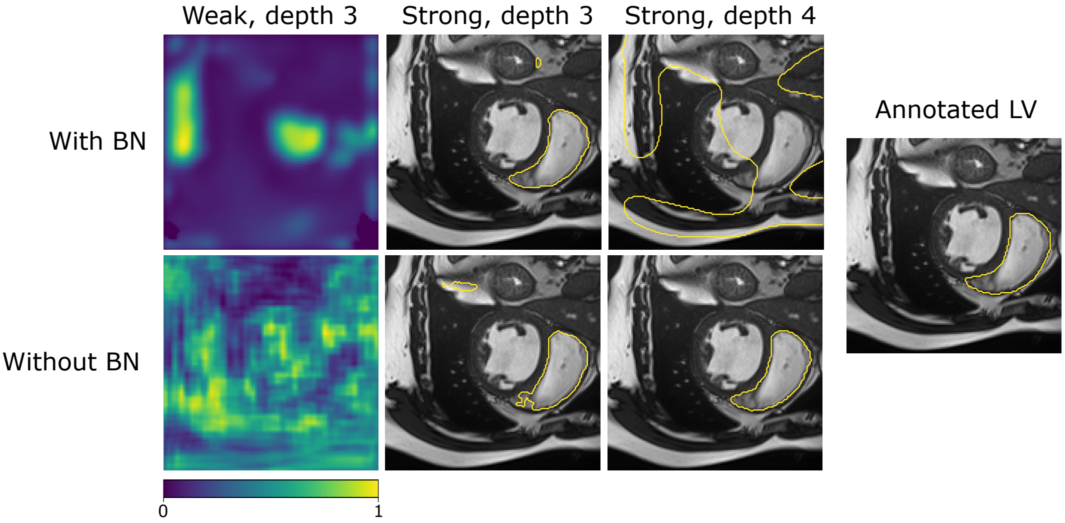


Figure 7.14: Attention maps and segmentations of networks trained on the ACDC dataset. The annotated LV is shown on the right. For the weak label networks, the attention maps are shown. For the strong label networks, the segmentations are shown.

7.10. Correlation of metrics

In subsection 6.6.1, several metrics for evaluation were discussed. As these metrics all assess the performance of the same task, i.e. volume quantification, they should be correlated over all experiments. A scatterplot matrix of the different metrics was made using the scores of the experiments conducted for this thesis to visualize the relationships between the metrics. This scatterplot matrix is shown in Figure 7.15. The metric scores of 375 experiments were used for the graph, of which 254 were weak label experiments and 121 were strong label experiments.

To determine the correlation between the metrics, the coefficient of determination was computed for linear correlations. To evaluate exponential correlations, first the logarithmic values of the metrics were computed before computing the coefficient of determination.

The scatter plots illustrated that not all metrics have clear correlations. Hence, a high score with one metric did not implicate a high score with another, and vise versa. The most clear relationship was the linear relationship of the ICC score and the Pearson coefficient, which was expected as they both measure a linear relationship between two variables. The coefficient of determination between ICC and Pearson was 0.81 and Spearman's rank coefficient was 0.95. Subsequently, the ICC score and the Dice of the strong label experiments also had a linear and exponential relation, with the coefficient of determinations of 0.60 and 0.62 respectively. The relation between the ICC score and the Dice of the weak label experiments, however, had no correlation with a coefficient of determination of 0.14 and a Spearman's rank coefficient of 0.38. The coefficient of determination for other combinations of metrics was below 0.5.

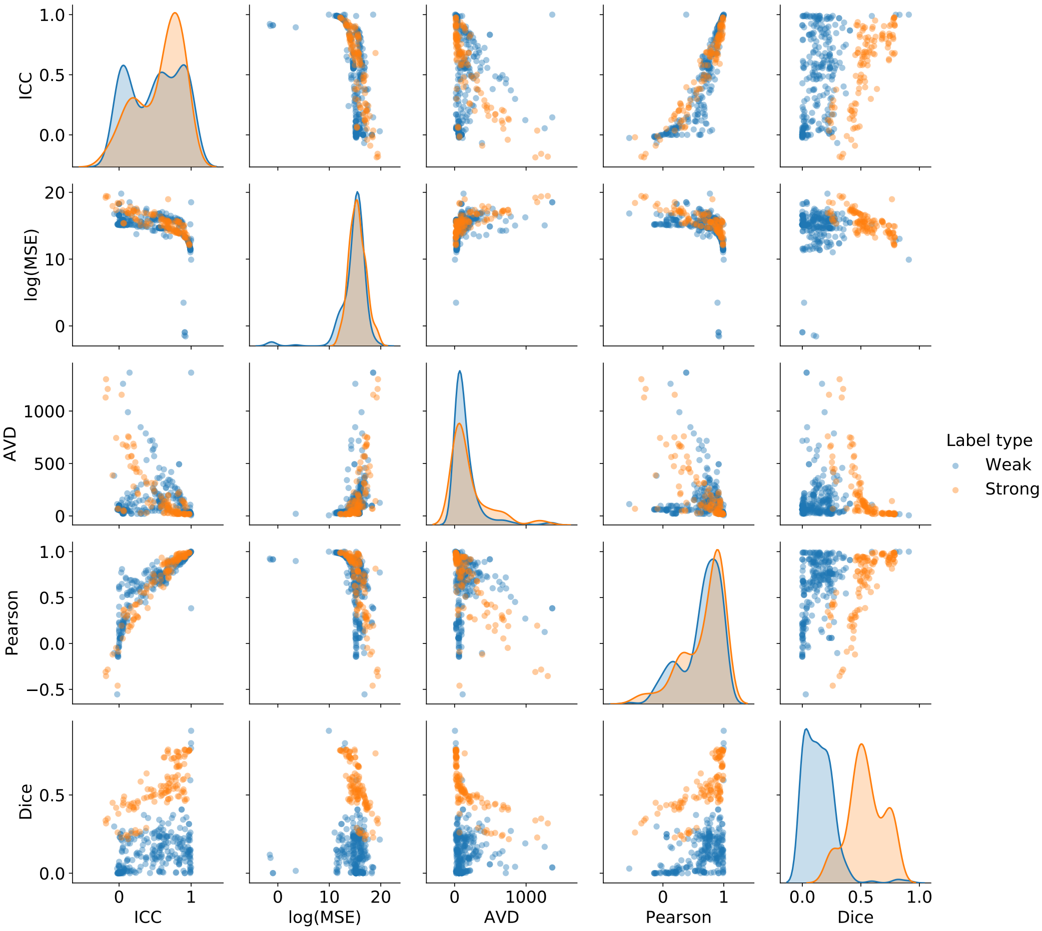


Figure 7.15: Scatter plot matrix of metric scores of different experiments. The metric scores of the weak label networks are shown in blue, and the metric scores of the strong label networks are shown in orange. Note that the MSE is shown with a log scale to show more diversity.

8

Discussion

The performance on the test dataset of the RSS showed that weak (image-level) label networks achieved a slight, but significant increase in ICC score performance for the quantification of the WMH volume compared to the strong (voxel-wise) label networks. The slight increase was probably due to its direct optimization. However, the AVD score was slightly higher (i.e. worse) for the weak label networks as they made more prediction errors for images with low WMH severity. The predictions of the weak label networks did not have a bias to overestimating the WMH volume. This finding is opposite to Kolesnikov and Lampert (2016), who suggested that global average pooling overestimated the targets size.

The generalization experiments to other datasets showed that the performance of network optimized with either label type was similar, with no significant differences. The sizes of the public datasets were probably too small to see a statistically significant difference. These results showed that weak labels, i.e. image-level labels, could be used for quantifying an image biomarker with the same performance as fully supervised networks. Creating fully annotated image scans by clinical experts could therefore be superfluous, and hence both time and money can be saved. Our results are in contrast with previous literature. Jia et al. (2017) and Schlegl et al. (2015) did not achieve similar performance as the fully supervised method with weak label optimization. The literature uses different datasets (e.g. in size and information per label) and methods (e.g. network architecture). For a fair comparison, our proposed method should be applied to their datasets in future research.

The generalization performance towards external datasets of both weak and strong networks was the highest on the WMHSeg dataset. This could be due to the anatomical differences between the WMH and MS as the networks were trained on a WMH annotated dataset. A clear correlation between the scanner types or voxel spacing and performance was not visible, as all external datasets used different settings and scanners than the RSS. The performance and consistency on the MS datasets varied per dataset. The generalization towards LongMSLes was relatively good, whereas MSSeg and MSLes had overall poor performance.

The weak label network learned meaningful image features. The generated attention maps showed the network focusing on regions surrounding WMHs. While these attention maps had poor overlap scores with the annotations, we showed that the network learned to identify the target region while using only the volume as an image-level label. This was in line with the conclusion of Zhou et al. (2016). Adjusting the weak label networks for intuitive confounders did not change the significance of the correlation between the predicted and annotation WMHs. This showed that the weak label network was not overfitting on the confounders, and extracted other meaningful image features or biomarkers to predict its score.

The performance of weak label networks was significantly worse than that of strong labels when trying to predict the volume of the left ventricle of the ACDC dataset. One of the reasons could be the dataset size. ACDC had 120 train images, in comparison to the 2602 train images of the RSS. As weak labels consist of less information per label, we need to compensate with quantity, i.e. more labels, to gain sufficient information for good performance. Subsequently, the heterogeneous tissues of the ACDC dataset made it hard for the network to know which region contributes to the image-level label. All kinds of combinations could be feasible. Therefore, we hypothesize that weak label networks are more sensitive to the application than strong label networks. The complexity and the relative target size (Schlegl et al., 2015) could influence the convergence of the weak label networks. We propose that it should be investigated whether and how this sensitivity decreases with data set size.

Both types of combined supervision on both the weak WMH volume label and the strong voxel-wise label showed decreasing performance. In the case of the simultaneous optimization, the difference in performance between the strong label network and both label networks was most visible for the smaller dataset. This difference reduced when the contribution of the weak label loss, tuned by the scaling parameter alpha, was reduced. The results suggested that both label types did not share the same image features for their task. Intuitively, these tasks are very similar. The attention maps of the weak label network, however, showed different imaging features extracted than the segmentation networks.

One possible reason might be that our weak labels, e.g. WMH volume, were extracted from the voxel-wise labels and hence did not contribute to the total information. Other literature on combined supervision extracted the additional labels from other sources (Chaganti et al., 2019; Hwang and Kim, 2016; Jia et al., 2017; Schlegl et al., 2015; Yoo et al., 2018). In contrast to our results, the literature reported an increase in performance when optimizing on multiple label types. Kervadec et al. (2019) extracted the weak labels from the voxel-wise label as well. They did, however, train a network with these labels to create regularization constraints for a second network.

In the case of the sequential optimization, the pre-training with weak labels decreased the performance of the network when fine-tuned with strong labels. Randomly initialized segmentation networks outperformed the weak label pre-training. While the initial segmentation scores of the pre-trained network were better than the random initialized network, we hypothesize that the weak labels pushed the pre-trained network towards a local minimum. The network was not able to escape from this minimum during fine-tuning.

In contrast, the strong label pre-trained model could be fine-tuned with weak labels to achieve better volume predictions. The direct optimization contributed to slightly better ICC scores after fine-tuning. The downside, however, was a larger decrease in Dice score, and thus localization. This could also indicate that the weak and strong label optimizations find different image features useful for prediction. The experiments did not show any advantage of using both label kinds for optimization. There are, however, different methods proposed by literature which were not applied with our network architecture and application. One example is the alternated training with both label types (Chen et al., 2019). The authors showed that optimizing alternatively on one label or the other converged better to a general solution that the simultaneous approach.

Using batch-normalization was not always beneficial. For the application of WMHs, the batch-normalization had a worsened effect on both weak and strong label networks' performance for a large dataset size. For smaller dataset sizes of the RSS, adding batch-normalization layers helped with the volume quantification.
In the case of the ACDC dataset, the batch-normalization layers resulted in worse volume quantification but better segmentation. For deeper networks, batch-normalization helped with convergence. Moreover, for the ACDC dataset, the best overall performance for both volume quantification and segmentation of the right ventricle was achieved by the deeper networks with batch-normalization. Overall, the influence of batch-normalization seemed sensitive to the task, dataset, and network depth. We propose therefore that for each new experiment setup to run experiments with and without batch-normalization and be careful to extrapolate performance trends towards other applications.

The correlation of metrics, as discussed in subsection 6.6.1, was worse than anticipated. A single metric could give a distorted view of the performance model. We suggest that its preferable to evaluate multiple metrics to show performance, especially for the prediction of image-level labels.

Subsequently, we showed that the neural networks did not only learn WMH quantification with its volume as a weak label, but with the number of WMH lesions or the severity score as weak labels as well. The attention maps produced by the number of lesions as the weak label showed that the network could learn localization. It mainly showed that the network could learn patterns without being restricted to one certain kind of weak label. The severity score, on the other hand, could be used to predict the WMH volume with relative high performance. It showed that the network was able to predict the volume with labels that consisted of less information. As severity scores are used in clinical practice, e.g. Fazekas scores, this could be a good label alternative to manual WMH annotations for training models for volume quantification.

Overall, the results showed the potential of quantifying image biomarkers with just image-level labels to train neural networks. This could open a new direction of dataset labeling in which image-level labels suffice. The task of segmentation with image-level label optimized networks seemed not feasible with current methods. Furthermore, our method for combined supervision had a negative influence on the WMH volume prediction. However, we do not omit the possibility that another method of combined supervision could be beneficial and urge further research in this direction. We recommend alternative optimization Chen et al. (2019) and extracting labels from different sources of information (e.g. Chaganti et al. (2019)) as most promising directions. Subsequently, we showed that our method was sensitive to the medical application. To make the methods feasible for clinical practice, different methods should be researched that are more robust to the application. Another important factor for the clinical practise, which was not evaluated in this report, is the reproducibility of the predictions. Next to accurancy, reproducibility is very important for making clinical applications. Research should be conducted to compare the reproducibility of the volume predictions between the image-level label and the voxel-wise label optimized networks.

8.1. Conclusion

The results showed that voxel-wise annotations could be superfluous for quantifying imaging biomarkers, i.e. WMH volume, and that image-level labels suffice. The creation of image-level labels instead of voxel-wise annotation could decrease the labeling burden of clinical experts and hence reduce costs. For the quantification of WMH volume, the best results were obtained using the WMH volume itself as a weak label, while we showed that other image-level labels like a WMH severity score could suffice as well. Subsequently, our implementation of combined supervision decreased the performance of the WMH quantification. The proposed method worked well for the application of WMH, but underperformed when it tried to predict the volume of heterogeneous tissue targets.

9 Acknowledgements

I had an amazing time working on this research and would like to thank everyone who made it possible.

I would like to thank my supervisors Florian and Marleen for making this graduation project a reality. Their support was well balanced between helping, critical evaluation, and letting me work independently.

Florian helped me a lot during my project. His initial ideas intrigued me to continue on his preliminary work. His positive attitude towards results was a huge motivation in continuing certain research directions while the benefit or impact was still not clear. Furthermore, I liked our discussions and he taught me a lot about doing research and formulating new hypotheses and claims. It was great working together on our MICCAI submission and the journal paper. I wish him the best in his new position at Stanford. Merci beaucoup!

I want to thank Marleen for helping me find a suitable project within her group. Furthermore, her knowledge and critical attitude were a great contribution to this thesis. I will not pursue an academic path, but there is no doubt about whom I would have liked to be my promoter in case of a PhD.

Kim was also a huge help during my time at BIGR. She was sort of my unofficial supervisor, as she was very interested in my work and a huge motivator. I really enjoyed our, sometimes lengthy, discussions and bantering about all kinds of small things related to deep learning.

The MBM group, i.e. Antonio, Deep, Gerda, Gijs, Robin, Shuai for all help and recommendations they provided during my stay. I enjoyed their presentations and literature discussions which helped broaden my knowledge passed my niche research direction.

Although not part of the MBM group, Martijn was very helpful regarding discussions about statistics and literature, and small things in general.

I'm thankful for all the help and organization Wiro provided. I'm looking forward to future collaborations with you at my new position.

I want to thank the members of BIGR, PhDs, and students, for all the fun lunch and coffee breaks, where we often couldn't refrain from continuing talking about research-related topics.

I would want to thank Frans for helping me find and contact Marleen to land this graduation project. His counseling before my graduation project helped me re-orientate and pursue my preferred direction of AI in healthcare.

Lastly, I want to like to thank my family and friends who supported me during my thesis and their interest in my research project.

Bibliography

Arvaniti, E. and Claassen, M. Coupling weak and strong supervision for classification of prostate cancer histopathology images. *arXiv preprint arXiv:1811.07013*, 2018.

Au, R., Massaro, J. M., Wolf, P. A., Young, M. E., Beiser, A., Seshadri, S., D'Agostino, R. B., and DeCarli, C. Association of white matter hyperintensity volume with decreased cognitive functioning: the framingham heart study. *Archives of neurology*, 63(2):246–250, 2006.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G., et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

Bortsova, G., Dubost, F., Ørting, S., Katramados, I., Hogeweg, L., Thomsen, L., Wille, M., and de Bruijne, M. Deep learning from label proportions for emphysema quantification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11071 LNCS:768–776, 2018. ISSN 16113349. doi: 10.1007/978-3-030-00934-2_85.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C. H., et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.

Chaganti, S., Bermudez, C., Mawn, L. A., Lasko, T., and Landman, B. A. Contextual deep regression network for volume estimation in orbital ct. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 104–111. Springer, 2019.

Chen, S., Bortsova, G., Juárez, A. G.-U., van Tulder, G., and de Bruijne, M. Multi-task attention-based semi-supervised learning for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–465. Springer, 2019.

Cheplygina, V., de Bruijne, M., and Pluim, J. P. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, may 2019. ISSN 13618423. doi: 10.1016/j.media.2019.03.009.

Chutinet, A. and Rost, N. S. White matter disease as a biomarker for long-term cerebrovascular disease and dementia. *Current treatment options in cardiovascular medicine*, 16 (3):292, 2014.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S. C., Girard, P., Ameli, R., Ferré, J.-C., et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific reports*, 8(1): 1–17, 2018.

Dai, J., He, K., and Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015.

De Boer, R., Vrooman, H. A., Van Der Lijn, F., Vernooij, M. W., Ikram, M. A., Van Der Lugt, A., Breteler, M. M., and Niessen, W. J. White matter lesion extension to automatic brain tissue segmentation on mri. *Neuroimage*, 45(4):1151–1161, 2009.

de Bruijne, M. Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis*, 33:94–97, 2016. ISSN 13618423. doi: 10.1016/j.media.2016.06.032.

de Groot, M., Ikram, M. A., Akoudad, S., Krestin, G. P., Hofman, A., van der Lugt, A., Niessen, W. J., and Vernooij, M. W. Tract-specific white matter degeneration in aging: the rotterdam study. *Alzheimer's & Dementia*, 11(3):321–330, 2015.

De Leeuw, F., de Groot, J. C., Achten, E., Oudkerk, M., Ramos, L., Heijboer, R., Hofman, A., Jolles, J., Van Gijn, J., and Breteler, M. Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. the rotterdam scan study. *Journal of Neurology, Neurosurgery & Psychiatry*, 70(1):9–14, 2001.

Debette, S. and Markus, H. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *Bmj*, 341:c3666, 2010.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

Dubost, F., Bortsova, G., Adams, H., Ikram, A., Niessen, W. J., Vernooij, M., and De Bruijne, M. GP-Unet: Lesion detection from weak labels with a 3D regression network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10435 LNCS:214–221, 2017a. ISSN 16113349.

Dubost, F., Bortsova, G., Adams, H., Ikram, A., Niessen, W. J., Vernooij, M., and De Bruijne, M. Gp-unet: Lesion detection from weak labels with a 3d regression network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 214–221. Springer, 2017b.

Dubost, F., Adams, H., Yilmaz, P., Bortsova, G., van Tulder, G., Ikram, M. A., Niessen, W., Vernooij, M., and de Bruijne, M. Weakly Supervised Object Detection with 2D and 3D Regression Neural Networks. pages 1–14, 2019.

Fazekas, F., Chawluk, J. B., Alavi, A., Hurtig, H. I., and Zimmerman, R. A. Mr signal abnormalities at 1.5 t in alzheimer's dementia and normal aging. *American journal of roentgenology*, 149(2):351–356, 1987.

Hachinski, V., Potter, P., and Merskey, H. Leuko-araiosis: an ancient term for a new problem. *Canadian Journal of Neurological Sciences*, 13(S4):533–534, 1986.

Hussain, M. A., Amir-Khalili, A., Hamarneh, G., and Abugharbieh, R. Segmentation-free kidney localization and volume estimation using aggregated orthogonal decision cnns. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 612–620. Springer, 2017.

Hwang, S. and Kim, H.-E. Self-transfer learning for fully weakly supervised object localization. *arXiv preprint arXiv:1602.01625*, 2016.

Ikram, M. A., Brusselle, G. G., Murad, S. D., van Duijn, C. M., Franco, O. H., Goedegebure, A., Klaver, C. C., Nijsten, T. E., Peeters, R. P., Stricker, B. H., et al. The rotterdam study: 2018 update on objectives, design and main results. *European journal of epidemiology*, 32(9):807–850, 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Jia, Z., Huang, X., Eric, I., Chang, C., and Xu, Y. Constrained deep weak supervision for histopathology image segmentation. *IEEE transactions on medical imaging*, 36(11): 2376–2388, 2017.

Kervadec, H., Dolz, J., Granger, É., and Ayed, I. B. Curriculum semi-supervised segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 568–576. Springer, 2019.

Khalaf, A., Edelman, K., Tudorascu, D., Andreescu, C., Reynolds, C. F., and Aizenstein, H. White matter hyperintensity accumulation during treatment of late-life depression. *Neuropsychopharmacology*, 40(13):3027–3035, 2015.

Koch, L. M., Rajchl, M., Bai, W., Baumgartner, C. F., Tong, T., Passerat-Palmbach, J., Aljabar, P., and Rueckert, D. Multi-atlas segmentation using partially annotated data: methods and annotation strategies. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1683–1696, 2017.

Kolesnikov, A. and Lampert, C. H. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016.

Koo, T. K. and Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

Kuijf, H. J., Biesbroek, J. M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M. J., Casamitjana, A., et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019.

Kwak, S., Hong, S., Han, B., et al. Weakly supervised semantic segmentation using superpixel pooling network. In *AAAI*, volume 1, page 2, 2017.

Lin, D., Dai, J., Jia, J., He, K., and Sun, J. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016.

Luo, G., Dong, S., Wang, K., Zuo, W., Cao, S., and Zhang, H. Multi-views fusion cnn for left ventricular volumes estimation on cardiac mr images. *IEEE Transactions on Biomedical Engineering*, 65(9):1924–1934, 2017.

Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H. G., Eisenmann, M., and Speidel, S. Can masses of non-experts train highly accurate image classifiers? In *International conference on medical image computing and computer-assisted intervention*, pages 438–445. Springer, 2014.

McKenna, M. T., Wang, S., Nguyen, T. B., Burns, J. E., Petrick, N., and Summers, R. M. Strategies for improved interpretation of computer-aided detections for ct colonography utilizing distributed human intelligence. *Medical image analysis*, 16(6):1280–1292, 2012.

Noh, H., Hong, S., and Han, B. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Rush, C. A., MacLean, H. J., and Freedman, M. S. Aggressive multiple sclerosis: proposed definition and treatment algorithm. *Nature Reviews Neurology*, 11(7):379, 2015.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, pages 2483–2493, 2018.

Schlegl, T., Waldstein, S. M., Vogl, W.-D., Schmidt-Erfurth, U., and Langs, G. Predicting semantic descriptions from medical images with convolutional neural networks. In *International Conference on Information Processing in Medical Imaging*, pages 437–448. Springer, 2015.

Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., and Summers, R. M. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506, 2016.

Smith, S. M. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., and Warfield, S. 3d segmentation in the clinic: A grand challenge ii: Ms lesion segmentation. *Midas Journal*, 2008:1–6, 2008.

Sun, Y.-Y., Zhang, Y., and Zhou, Z.-H. Multi-label learning with weak label. In *Twenty-fourth AAAI conference on artificial intelligence*, 2010.

Tan, Q., Yu, G., Domeniconi, C., Wang, J., and Zhang, Z. Incomplete multi-view weak-label learning. In *IJCAI*, pages 2703–2709, 2018.

Taylor, W. D., Steffens, D. C., MacFall, J. R., McQuoid, D. R., Payne, M. E., Provenzale, J. M., and Krishnan, K. R. R. White matter hyperintensity progression and late-life depression outcomes. *Archives of general psychiatry*, 60(11):1090–1096, 2003.

Wang, J., Knol, M. J., Tiulpin, A., Dubost, F., de Bruijne, M., Vernooij, M. W., Adams, H. H., Ikram, M. A., Niessen, W. J., and Roshchupkin, G. V. Gray matter age prediction as a biomarker for risk of dementia. *Proceedings of the National Academy of Sciences*, 116 (42):21213–21218, 2019a.

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Amin, S., and Liu, H. A clinical text classification paradigm using weak supervision and deep representation. *BMC medical informatics and decision making*, 19(1):1, 2019b.

Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R. I., T O'Brien, J., Barkhof, F., Benavente, O. R., et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12(8):822–838, 2013.

Wardlaw, J. M., Valdés Hernández, M. C., and Muñoz-Maniega, S. What are white matter hyperintensities made of? relevance to vascular cognitive impairment. *Journal of the American Heart Association*, 4(6):e001140, 2015.

Xie, Y. and Tao, X. White matter lesion segmentation using machine learning and weakly labeled mr images. In *Medical Imaging 2011: Image Processing*, volume 7962, page 79622G. International Society for Optics and Photonics, 2011.

Yao, L., Prosky, J., Poblenz, E., Covington, B., and Lyman, K. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*, 2018.

Yoo, B., Kwak, Y., Kim, Y., Choi, C., and Kim, J. Deep facial age estimation using conditional multitask learning with weak label expansion. *IEEE Signal Processing Letters*, 25(6):808–812, 2018.

Zeiler, M. D. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Zhou, Z.-H. A brief introduction to weakly supervised learning. *National Science Review*, 5 (1):44–53, 2018.

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells III, W. M., Jolesz, F. A., and Kikinis, R. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic radiology*, 11(2):178–189, 2004.