Fusing Unsupervised and Supervised Deep Learning for White Matter Lesion Segmentation

Christoph Baur[1] C.BAUR@TUM.DE
Benedikt Wiestler[3]
Shadi Albarqouni[1]
Nassir Navab[1,2]
[1] *Computer Aided Medical Procedures (CAMP), TU Munich, Germany*
[2] *Whiting School of Engineering, Johns Hopkins University, Baltimore, United States*
[3] *Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, TU Munich, Germany*

Abstract

Unsupervised Deep Learning for Medical Image Analysis is increasingly gaining attention, since it relieves from the need for annotating training data. Recently, deep generative models and representation learning have lead to new, exciting ways for unsupervised detection and delineation of biomarkers in medical images, such as lesions in brain MR. Yet, Supervised Deep Learning methods usually still perform better in these tasks, due to an optimization for explicit objectives. We aim to combine the advantages of both worlds into a novel framework for learning from both labeled & unlabeled data, and validate our method on the challenging task of White Matter lesion segmentation in brain MR images. The proposed framework relies on modeling normality with deep representation learning for Unsupervised Anomaly Detection, which in turn provides optimization targets for training a supervised segmentation model from unlabeled data. In our experiments we successfully use the method in a Semi-supervised setting for tackling domain shift, a well known problem in MR image analysis, showing dramatically improved generalization. Additionally, our experiments reveal that in a completely Unsupervised setting, the proposed pipeline even outperforms the Deep Learning driven anomaly detection that provides the optimization targets.

Keywords: Deep Learning, Anomaly Detection, Unsupervised, Semi-Supervised, Supervised, White Matter Lesion Segmentation, Multiple Sclerosis

1. Introduction

Deep Learning for medical image analysis is still impeded by a general lack of labeled training data. Especially for medical image segmentation, the creation of pixel-level annotations is a very tedious, time-consuming and costly task, which often has to be carried out by domain experts. Although it has been shown that in some cases supervised models can be trained from very small training datasets (Ronneberger et al., 2015), usually large amounts of labeled training data are required to achieve compelling model performance. This is also the case for automatic segmentation of white matter lesions (WML) in brain MR images. WML, a result of demyelination of cells in the white matter of the brain, are important biomarkers for underlying degenerative neurological diseases such as Multiple Sclerosis and can vary greatly in size, shape and location (Carass et al., 2017). Supervised Deep Learning based WML segmentation methods (Brosch et al., 2016; Valverde et al., 2017; Roy et al., 2018) do not only have to cope with this wide variety of lesion appearances, but

additionally are confronted with the problem of domain shift: In contrast to CT data, intensities in MR images do not have a clear physical interpretation, and generally there is a discrepancy between data distributions of images produced with different MR scanners. It is this discrepancy which makes segmentation methods in MR data hardly generalize to new devices, and labeled training data from different scanners to deal with this issue might not be readily available.

To generally overcome these burdens, the community has made numerous efforts towards Unsupervised and Semi-Supervised Deep Learning, i.e. learning without any labeled data and learning from both labeled and unlabeled data, respectively. A promising approach towards this direction is pseudo-labelling (Lee, 2013), where supervised models are fine-tuned from labeled data together with unlabeled samples, for which labels have been predicted with the same model. Another approach, however limited to patch-based classification, are so-called Ladder networks (Rasmus et al., 2015). More recent works leveraged adversarial networks for Domain Adaptation by either explicitly enforcing domain invariant feature representations (Kamnitsas et al., 2017) or encouraging the model to also produce realistic segmentation masks on unlabeled samples (Dong et al., 2018). (Ganaye et al., 2018) employ semantic constraints to improve robustness of a brain structure segmentation model and (Jiang et al., 2018) use tumor-aware MR image synthesis from CT images to train a model for tumor segmentation from both labeled and synthetic data. At the example of MS lesion segmentation, (Baur et al., 2017) proposed a Semi-Supervised Deep Learning framework for Domain Adaptation of fully convolutional segmentation networks by encouraging domain invariant feature representations on randomly sampled embeddings.

A recent trend to overcome the burden of pixel-level annotations is to leverage deep generative models and deep representation learning for the task of Unsupervised Anomaly Detection (UAD) in medical images. Under the assumption that "healthy" data is readily available at hospitals, these approaches model the distribution of healthy anatomy and try to detect anomalies as outliers from the modeled distribution directly in image space. In early work (Schlegl et al., 2017), GANs were proposed to detect anomalies in small retinal OCT patches. For head CT, Sato et al. (2018) showed promising initial results using 3D Auto-Encoders and Pawlowski et al. (2018) studied the effects of averaging multiple Monte-Carlo dropout reconstructions in Bayesian Auto-Encoders for anomaly detection. For UAD in brain MR images, Chen and Konukoglu (2018) showed promising results for detecting large lesions with Constrained Adversarial AEs, and (Baur et al., 2018) found that spatial Auto-Encoders enable UAD at high resolution, ultimately allowing such models to also detect small MS lesions.

We propose a novel framework for WML segmentation that can benefit from both labeled and unlabeled data. Therefore, we combine i) a spatial Auto-Encoder, which performs UAD in brain MR images, and ii) a supervised segmentation network. We show that, in addition to labeled data, the anomaly detections obtained from the Auto-Encoder on unlabeled data can be leveraged for Unsupervised Domain Adaptation. As a proof-of-concept, we also show that the segmentation network can be trained from UAD results alone, which performs considerably better than the actual UAD approach, leaving us with a novel approach for Unsupervised Deep Learning as well.

2. Methodology

2.1. Overall Concept

Our framework consists of an Auto-Encoder (AE), which is used for UAD, and a UNet-like model for supervised image segmentation (Figure 1). In a first step, the AE is optimized for compressing

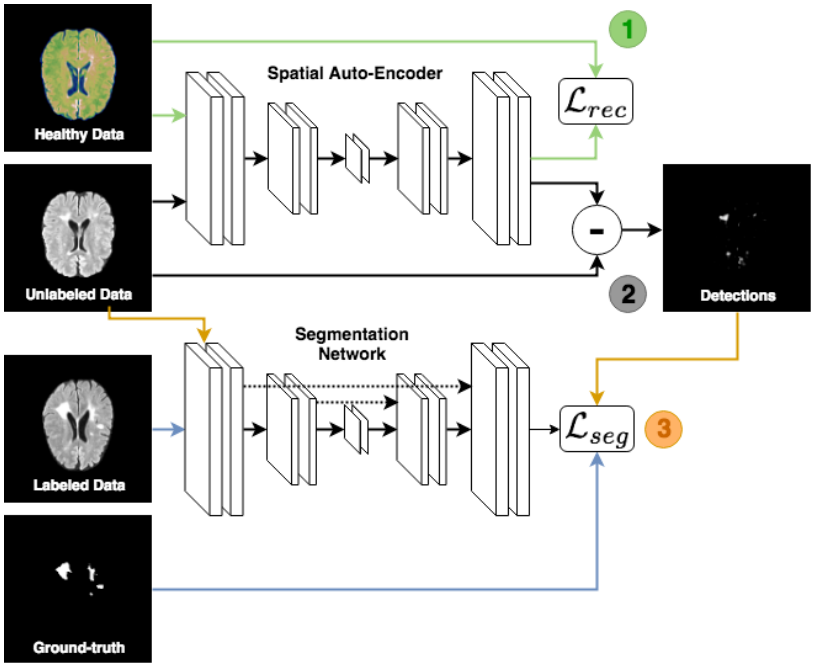

Figure 1: The proposed framework at a glance. Step 1: Training of a spatial AE on healthy data; Step 2: Inference on unlabeled data to obtain delineations; Step 3: Training of a supervised model from both labeled data with ground-truth and unlabeled data with UAD delineations.

and reconstructing images of healthy anatomy. Afterwards (step 2), it is used to detect and delineate anomalies in previously unseen, unlabeled data. In step 3, the UNet is trained in a supervised manner for pixel-wise WML, by jointly using labeled training data $\mathscr{X}_L$ with ground-truth $\mathscr{Y}_L$ as well as the unlabeled training data $\mathscr{X}_U$, for which the anomaly detection provides an "artificial ground-truth" $\mathscr{S}$.

2.2. Capturing normality for anomaly detection

Similar to (Baur et al., 2018), we train a 2D spatial Auto-Encoder to capture the notion of anatomical brain normality (see Figure 2 for a depiction of the network architecture). Given a set of healthy training data $\mathscr{X}_H$, we therefore optimized an AE for the following reconstruction objective:

$$\mathscr{L}_{rec}(\mathbf{x},\hat{\mathbf{x}}) = \ell_1(\mathbf{x},\hat{\mathbf{x}}) + \ell_2(\mathbf{x},\hat{\mathbf{x}}) + \lambda_{gdl}\,\mathrm{gdl}(\mathbf{x},\hat{\mathbf{x}}) \tag{1}$$

The terms ℓ_1 and ℓ_2 constitute the pixel-wise Manhattan and Euclidean distances between input image $\mathbf{x} \in \mathscr{X}_H$ and reconstruction $\hat{\mathbf{x}}$. In contrast to (Baur et al., 2018), which used an adversarial network to promote the reconstruction of realistic images, we used

$$\mathrm{gdl}(\mathbf{x},\hat{\mathbf{x}}) = \sum_{i,j} ||x_{i,j} - x_{i-1,j}| - |\hat{x}_{i,j} - \hat{x}_{i-1,j}|| + ||x_{i,j-1} - x_{i,j}| - |\hat{x}_{i,j-1} - \hat{x}_{i,j}|| \tag{2}$$

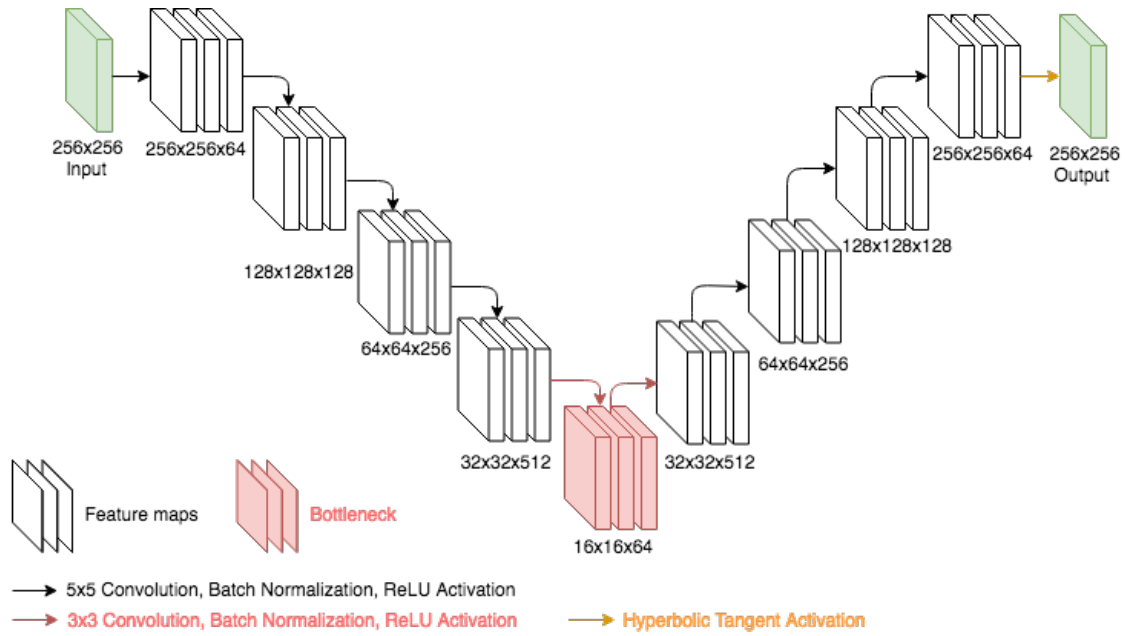

Figure 2: The architecture of the 2D Auto-Encoder used in our experiments.

the so-called gradient difference loss (Mathieu et al., 2015), weighted by λ_{gdl}. All these losses combined encourage the model not only to reconstruct coherent, but also very crisp images.

For detecting anomalies in a query sample $\mathbf{x}^* \notin \mathscr{X}_H$ not being part of the Auto-Encoders training set, $\mathbf{x}^*$ is propagated through the model and a pixel-wise residual between $\mathbf{x}^*$ and its reconstruction $\hat{\mathbf{x}}^*$ is computed:

$$\mathbf{r} = m(\max(\mathbf{x}^* - \hat{\mathbf{x}}^*, 0))$$

where $m(\cdot,\cdot)$ is a non-linear $5 \times 5 \times 5$ median filter for emphasizing connected anomalous structures and simultaneously removing unwanted, small residual pixels which might be high in intensity and lead to increased False Positive (FP) responses. Importantly, many of such potential FP residuals are already avoided by optimizing for the gdl-term, but the filtering is necessary. Further, we set any negative residuals to zero using ReLu to avoid detections of anomalies which do not resemble white matter lesions, as such lesions are usually hyper-intense in the FLAIR images we use. The resulting residuals are further binarized into images $\mathbf{s}$ via thresholding, i.e.

$$\mathbf{s} = \mathbf{r} \geq t$$

and collected in a set of anomaly labels $\mathscr{S}$. How we choose the threshold t is explained in the experiments section.

3. Experiments and Results

3.1. Dataset

For our experiments, we make use of two different datasets. We utilize the labeled data provided in the publicly available MICCAI 2008 MS lesion segmentation challenge dataset. The data ac-

quired at University of North Carolina ($\mathscr{D}_{UNC}$) and the Children's Hospital Boston ($\mathscr{D}_{CHB}$) comprise FLAIR, T1 and T2-weighted images from 10 subjects per site.

Further, we use a non-public dataset, generously provided by our clinical partners at Klinikum Rechts der Isar, consisting of FLAIR and T1-weighted MR acquisitions of 68 healthy subjects ($\mathscr{D}_{healthy}$) as well as 49 subjects which were diagnosed with MS ($\mathscr{D}_{MS}$). For the latter, expert delineations of MS lesions were provided. All images have been acquired with a Philips Achieva 3T scanner.

Preprocessing Prior to any Deep Learning, all acquisitions have been projected to the SRI24 ATLAS space (Rohlfing et al., 2009), denoised using CurvatureFlow, Skull-Stripped using ROBEX (Iglesias et al., 2011) and normalized into the range of $[0;1]$. While we train our models only from FLAIR images, we utilize the T1-modalities for co-registration and skull-stripping. Table 1 provides details about our training, validation and testing splits on the respective datasets.

Table 1: Training, Validation & Testing subjects of our datasets as well as additional subjects which are considered unlabeled in our experiments

Dataset	**Train** $\mathscr{X}_L$	**Val** $\mathscr{X}_{VAL}$	**Test** $\mathscr{X}_{TEST}$	**Additional** $\mathscr{X}_U$
$\mathscr{D}_{healthy}$	68	-	-	-
$\mathscr{D}_{MS}$	15	5	10	19
$\mathscr{D}_{CHB}$	6	2	2	-
$\mathscr{D}_{UNC}$	6	2	2	-

We utilize all FLAIR images of the healthy subjects in $\mathscr{D}_{healthy}$ for training the AE. Further, we randomly split D_{CHB} and D_{UNC} each into training, validation and testing subjects. Similarly, out of $\mathscr{D}_{MS}$ we utilize 15 randomly chosen subjects and their ground-truth segmentations for training a supervised segmentation model, 5 for validation and 10 subjects to test the models performance. We consider the remaining 19 subjects as unlabeled and utilize our AE to obtain an artificial "ground-truth" $\mathscr{S}$. The 5 labeled validation subjects with MS lesions are also used to choose an operating point for the UAD.

3.2. Auto-Encoder

A spatial AE, refered to as *UAD*, has been trained for 150 epochs from entire axial MR slices (256×256px) $\in \mathscr{D}_{healthy}$ with a learning-rate of 0.01. For the first 30 epochs, we set $\lambda_{gdl} = 0.0$ to allow the model to converge to coherent reconstructions, and then set it to 100.0 to make the model focus more on reconstructing fine details. Afterwards, the MS lesion validation set $\mathscr{X}_{VAL} \in \mathscr{D}_{MS}$ is processed by the AE to determine an Operating Point (OP) $t = 0.0187$ which maximizes the DICE-Score on $\mathscr{X}_{VAL}$. In succession, all non-empty slices (determined via skull-stripping) of the additional 19 "unlabeled" subjects have been processed with the UAD model to detect anomalies, i.e. to generate our artificial ground-truth $\mathscr{S}$ (see Figure 3 for an anomaly detection example from this set).

Similarly, another model $UAD_{no-\mathrm{gdl}}$ has been trained with fixed $\lambda_{gdl} = 0.0$ throughout the entire 150 training epochs to study the impact of the gradient-difference-loss component. The anomaly detection performance of the two models on the testing subjects $\mathscr{X}_{TEST} \in \mathscr{D}_{MS}$ is reported in Table 2.

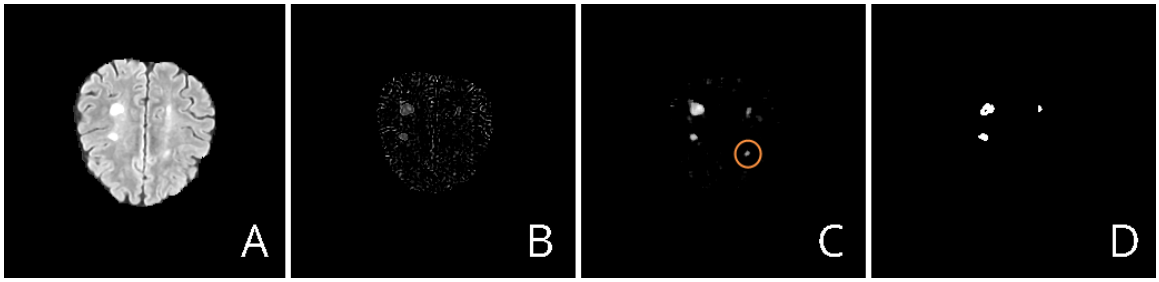

Figure 3: Anomaly detections provided by the AE. A: Input slice; B: Unprocessed Residuals; C: Postprocessed Residuals (with a FP encircled in orange); D: Ground-truth segmentation(s)

Notably, the UAD model performs better than its ablated counterpart $\text{UAD}_{no-\text{gdl}}$ in all measures, which shows that the training with the gradient-difference-loss is indeed beneficial for anomaly detection.

3.3. Unsupervised and Semi-Supervised Deep Learning

To investigate the general suitability of our framework for Unsupervised and Semi-Supervised DL, we first conducted a set of experiments on data from a single domain. Therefore, we trained multiple UNet segmentation models using the $\mathscr{D}_{MS}$ dataset, for which a larger number of subjects was available. The models comprise i) a supervised model $\mathbf{A}_{\mathscr{Y}_L}$, trained only on labeled data $(\mathscr{X}_L, \mathscr{Y}_L) \in \mathscr{D}_{MS}$, ii) a supervised model $\mathbf{A}_{\mathscr{Y}_L+\mathscr{Y}_U}$, trained with $(\mathscr{X}_L, \mathscr{Y}_L)$ as well as the additional data $(\mathscr{X}_U, \mathscr{Y}_U) \in \mathscr{D}_{MS}$ with its real ground-truth $\mathscr{Y}_U$, and iii) an unsupervised model $\mathbf{A}_{\mathscr{S}}$, trained only from additional "unlabeled" data $\mathscr{X}_U$ and artificial ground-truth $\mathscr{S}$. Further, we trained a semi-supervised model $\mathbf{A}_{\mathscr{Y}_L+\mathscr{S}}$ with $(\mathscr{X}_L, \mathscr{Y}_L)$ and $(\mathscr{X}_U, \mathscr{S})$. All models have been trained for 50 epochs from 128×128px sized patches extracted around MS lesions. The OP is again determined on the validation data. Performances of the models on the testing set are reported in Table 2.

Table 2: Unsupervised, Semi-Supervised and Supervised Deep Learning experiments. Note: DICE is the overall Dice-Score, DICE ($\mu \pm \sigma$) is the statistics over the Dice-Scores obtained per subject and AUPRC is the Area under the Precision-Recall-Curve

Model	**DICE**	**DICE** ($\mu \pm \sigma$)	**AUPRC**	**Training Subjects**
UAD	0.6343	0.6156 ± 0.0972	0.6157	all in $\mathscr{D}_{healthy}$
$\text{UAD}_{no-\text{gdl}}$	0.6101	0.5831 ± 0.0989	0.5989	all in $\mathscr{D}_{healthy}$
$\mathbf{A}_{\mathscr{Y}_L}$	0.7259	0.7026 ± 0.0635	0.7537	15 $(\mathscr{X}_L, \mathscr{Y}_L)$
$\mathbf{A}_{\mathscr{S}}$	0.6792	0.6643 ± 0.0775	0.6964	19 $(\mathscr{X}_U, \mathscr{S})$
$\mathbf{A}_{\mathscr{Y}_L+\mathscr{S}}$	0.7057	0.6815 ± 0.0743	0.7254	15 $(\mathscr{X}_L, \mathscr{Y}_L)$ + 19 $(\mathscr{X}_U, \mathscr{S})$
$\mathbf{A}_{\mathscr{Y}_L+\mathscr{Y}_U}$	**0.7338**	$\mathbf{0.7148 \pm 0.0591}$	**0.7642**	15 $(\mathscr{X}_L, \mathscr{Y}_L)$ + 19 $(\mathscr{X}_U, \mathscr{Y}_U)$

3.4. Semi-Supervised Domain Adaptation

Next, we investigated the suitability of our approach for the task of Domain Adaptation by comparing a semi-supervised model against supervised baselines. Therefore, we trained a supervised model $\mathbf{B}_{\mathscr{Y}_L}$ using the training set $\mathscr{X}_L \in \mathscr{D}_{CHB}$. We further trained a semi-supervised model $\mathbf{B}_{\mathscr{Y}_L+\mathscr{S}}$ using $\mathscr{X}_L$ as well as the unlabeled data $\mathscr{X}_U \in \mathscr{D}_{MS}$ with artificial labels $\mathscr{S}$, and a supervised upper bound model $\mathbf{B}_{\mathscr{Y}_L+\mathscr{Y}_U}$ using the same data $\mathscr{X}_L$ and $\mathscr{X}_U$, however optimizing for the real ground-truth $\mathscr{Y}_U$ of $\mathbf{X}_U$. The very same experiments have been performed with $\mathscr{D}_{UNC}$ as well. Again, all models have been trained for 50 epochs on 128×128px sized patches cropped around MS lesions. The respective performances on the testing sets of both domains are reported in Table 3 and Table 4.

Table 3: Domain Adaptation experiments for $\mathscr{D}_{CHB} \rightarrow \mathscr{D}_{MS}$

	MSSEG-CHB			**MS**		
Model	**DICE**	**DICE** ($\mu \pm \sigma$)	**AUPRC**	**DICE**	**DICE** ($\mu \pm \sigma$)	**AUPRC**
$\mathbf{B}_{\mathscr{Y}_L}$	0.4473	0.4472 ± 0.0003	0.3649	0.3975	0.3752 ± 0.0769	0.3185
$\mathbf{B}_{\mathscr{Y}_L+\mathscr{S}}$	0.5756	0.5423 ± 0.0580	0.5843	0.6751	0.6547 ± 0.0802	0.6927
$\mathbf{B}_{\mathscr{Y}_L+\mathscr{Y}_U}$	0.5590	0.5278 ± 0.0580	0.54081	0.7203	0.6935 ± 0.0646	0.7597

Table 4: Domain Adaptation experiments for $\mathscr{D}_{UNC} \rightarrow \mathscr{D}_{MS}$

	MSSEG-UNC			**MS**		
Model	**DICE**	**DICE** ($\mu \pm \sigma$)	**AUPRC**	**DICE**	**DICE** ($\mu \pm \sigma$)	**AUPRC**
$\mathbf{B}_{\mathscr{Y}_L}$	0.3924	0.3903 ± 0.0047	0.3170	0.3622	0.3428 ± 0.0698	0.3059
$\mathbf{B}_{\mathscr{Y}_L+\mathscr{S}}$	0.5634	0.5314 ± 0.0622	0.5649	0.6746	0.6611 ± 0.0780	0.6905
$\mathbf{B}_{\mathscr{Y}_L+\mathscr{Y}_U}$	0.5877	0.5628 ± 0.0467	0.5804	0.7195	0.6945 ± 0.0708	0.7433

3.5. Discussion

As Table 2 shows, training a supervised model ($\mathbf{A}_{\mathscr{S}}$) only from artificial ground-truth performs considerably better than the UAD which actually produces this artificial ground-truth $\mathscr{S}$. We believe this occurs due to the fact that the supervised model is trained for an explicit objective, whereas the UAD approach has no knowledge about the task at hand. In comparison to the supervised model $\mathbf{A}_{\mathscr{Y}_L}$ and the semi-supervised model $\mathbf{A}_{\mathscr{Y}_L+\mathscr{S}}$, we notice that $\mathbf{A}_{\mathscr{S}}$ is slightly inferior, which might be due to FPs in segmentations (see Figure 3 C) provided by the UAD for training the segmentation network. The FP in $\mathscr{S}$ are possibly learned and again reflected by the segmentation model. This effect might be overcome or at least weakened when using the continuous UAD output rather than binarizations. Interestingly, when using both $(\mathscr{X}_L, \mathscr{Y}_L)$ and $(\mathscr{X}_U, \mathscr{S})$ to train the semi-supervised model $\mathbf{A}_{\mathscr{Y}_L+\mathscr{S}}$, the resulting network is also inferior to the supervised model $\mathbf{A}_{\mathscr{Y}_L}$, although additional data has been provided. Again, we amount this to FP in $\mathscr{S}$. In fact, training data from 15 subjects seems to provide enough information for obtaining a model which performs already well, such that there might hardly be additional information in UAD delineations for the model to exploit. Simultaneously, the imperfections in $\mathscr{S}$ might be learned, though, and potentially confuse the model.

For the task of Domain Adaptation, however, our framework shows great potential. A model $\mathbf{B}_{\mathscr{D}_L}$ originally trained from 6 labeled subjects coming from dataset $\mathscr{D}_{CHB}$ generalizes poorly to testing data from domain $\mathscr{D}_{MS}$, but when leveraging the artificial ground-truth $\mathscr{S}$ provided by the UAD and also training the segmentation network from the originally unlabeled data, we witness great improvements on both domains. On the source domain $\mathscr{D}_{CHB}$, we even outperform the upper bound model $\mathbf{B}_{\mathscr{D}_L+\mathscr{D}_U}$ which has been trained from labeled data of both domains. Similarly, although not outperforming the upper bound model, a positive trend is also noticed in the experiments involving the $\mathscr{D}_{UNC}$ dataset, which provides clear evidence that UAD delineations can be very beneficial for Domain Adaptation.

4. Conclusion

We presented a novel framework which combines unsupervised deep representation learning and supervised deep learning into a pipeline which can be used for both Semi-supervised and completely Unsupervised Deep Learning. We believe that this approach can be useful beyond the presented use case of WML segmentation in brain MR, as long as the unsupervised anomaly detection provides labels at reasonable quality. In future work, we would like make use of the continuous UAD output rather than the binarized detections to make sure lower confidence anomalies have proportionally less impact on the segmentation performance of the supervised model. It might also be beneficial to follow the idea in (Dong et al., 2018) and employ a discriminator on the supervised segmentation network to regularize the model by encouraging good, realistic segmentations.

Acknowledgments

We thank our clinical partners from Klinikum Rechts der Isar for generously providing us with their dataset.

References

Christoph Baur, Shadi Albarqouni, and Nassir Navab. Semi-supervised deep learning for fully convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 311–319. Springer, 2017.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. *arXiv preprint arXiv:1804.04488*, 2018.

Tom Brosch, Lisa YW Tang, Youngjin Yoo, David KB Li, Anthony Traboulsee, and Roger Tam. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.

Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.

Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018.

Nanqing Dong, Michael Kampffmeyer, Xiaodan Liang, Zeya Wang, Wei Dai, and Eric Xing. Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 544–552. Springer, 2018.

Pierre-Antoine Ganaye, Michaël Sdika, and Hugues Benoit-Cattin. Semi-supervised learning for segmentation under semantic constraint. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 595–602. Springer, 2018.

J E Iglesias, Cheng-Yi Liu, P M Thompson, and Zhuowen Tu. Robust Brain Extraction Across Datasets and Comparison With Publicly Available Methods. *IEEE Transactions on Medical Imaging*, 30(9):1617–1634, 2011.

Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Gig S Mageras, Joseph O Deasy, and Harini Veeraraghavan. Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 777–785. Springer, 2018.

Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging*, pages 597–609. Springer, 2017.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013.

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

Nick Pawlowski, Matthew CH Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan Stevenson, Aneesh Khetani, Tom Newman, et al. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders. In *International Conference on Medical Imaging with Deep Learnin (MIDL 2018)*, 2018.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554, 2015.

Torsten Rohlfing, Natalie M Zahr, Edith V Sullivan, and Adolf Pfefferbaum. The SRI24 multichannel atlas of normal adult human brain structure. *Human Brain Mapping*, 31(5):798–819, December 2009.

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

Snehashis Roy, John A Butman, Daniel S Reich, Peter A Calabresi, and Dzung L Pham. Multiple sclerosis lesion segmentation from brain mri via fully convolutional neural networks. *arXiv preprint arXiv:1803.09172*, 2018.

Daisuke Sato, Shouhei Hanaoka, Yukihiro Nomura, Tomomi Takenaga, Soichiro Miki, Takeharu Yoshikawa, Naoto Hayashi, and Osamu Abe. A primitive study on unsupervised anomaly detection with an autoencoder in emergency head ct volumes. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105751P. International Society for Optics and Photonics, 2018.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.

Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluis Ramio-Torrenta, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.