MEDDOCS Open Access Publisher

ISSN: 2637-885X

Journal of Radiology and Medical Imaging

Open Access | Research Article

Multi-Modal Brain and Ventricle Segmentation Using Weakly Supervised Transfer Learning

Jorge Barrios Ginart[1]*; Benjamin P Ziemer[1]*; Tomi Nano[1]; Kerem Can Turgutlu[2]; Abdalla Ibrahim[3]; Yannet Interian[2]; Anish Dalal[2]; Robert Sandor[2]; Julie Leseur[4]; Martin Vallières[5]; Taman Upadhaya[1]; Steve Braunstein[1]; Gilmer Valdes[1]; Michael McDermott[6]; Javier Villanueva-Meyer[7]; Olivier Morin[1]

[1]Radiation Oncology, University of California San Francisco, San Francisco CA, USA.

[2]University of San Francisco, Data Science, San Francisco, CA, USA.

[3]The D-Lab, Department of Precision Medicine, GROW- School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, The Netherlands.

[4]Radiation Oncology, Centre Eugène Marquis, Rennes, France, USA.

[5]Department of Computer Science, Université de Sherbrooke, Sherbrooke, Quebec, Canada.

[6]Division of Neuroscience, Herbert Wertheim College of Medicine, Florida International University, Miami FL, USA.

[7]Radiology and Biomedical Imaging, University of California San Francisco, San Francisco CA, USA.

Corresponding Author(s): Olivier Morin

Associate Professor, University of California San Francisco, 505 Parnassus Avenue, L08/L75, 0226, San Francisco, CA 94143.

Tel: 415-353-9302, Fax: 415-353-8697;

Email: Olivier.Morin@ucsf.edu

*Both authors contributed equally to this work.

Received: Aug 16, 2021

Accepted: Sep 13, 2021

Published Online: Sep 16, 2021

Journal: Journal of Radiology and Medical Imaging

Publisher: MedDocs Publishers LLC

Online edition: http://meddocsonline.org/

Keywords: Weakly supervised learning; Transfer learning; Semantic segmentation; Convolutional neural networks; Brain ventricles.

Abbreviations: CNN: Convolutional Neural Network; TL: Transfer Learning; noTL: No Transfer Learning (training de novo, from scratch); DS: Dice Score; PVD: Percent Volume Difference.

Abstract

Purpose: To quantify performance of brain and ventricle segmentation using weakly supervised Transfer Learning (TL) and cross-modality CT-to-MR deep learning models trained with coarse-grained versus fine-grained data to enable accurate segmentation using small datasets.

Methods: An IRB approved, retrospective study using MR and CT images was performed. Three datasets consisting of roughly 2500 total images with coarse or fine annotations (labels) were used for training, validation and testing of Convolution Neural Networks (CNNs) models. The best CNN architecture was used to investigate TL performance on segmentation, influence of training set size for TL accuracy and effectiveness of cross-modality TL. Dice Score (DS) and Percent Volume Difference (PVD) were used to quantify segmentation accuracy. Two-sided Wilcoxon signed rank with $p<0.05$ indicated statistical significance.

Results: Deeper, wider models outperformed other architectures for segmentation tasks. Ventricle segmentation models trained with fine-grained data improved DS from 0.75 to 0.82 and PVD from 21.5% to 8.02% over coarse-grained models, both statistically significant. DS and PVD improved when using TL over noTL (0.86 vs. 0.83, $p<0.01$ and 6.87% vs. 8.02%, $p=0.80$, respectively). Both MR-to-MR and cross-modality MR-to-CT TL models trained with as few as 20 images showed similar results to models trained with 100 images and vastly outperformed small training size *de novo* models. Additionally, the cross-modality TL showed statistically significant improved results over noTL models and slightly lower DS and PVD than within-modality models.

Cite this article: Ginart J B, Ziemer B P, Nano T, Turgutlu K C, Ibrahim A, et al. Multi-Modal Brain and Ventricle Segmentation Using Weakly Supervised Transfer Learning. J Radiol Med Imaging. 2021: 4(1); 1052.

Conclusion: Brain and ventricle segmentation using deep and wide CNN networks outperformed shallower CNN models. Within-modality and cross-modality TL models achieved similar or superior performance compared to noTL models and TL models showed these results when trained with as little at 20% of the data of noTL models.

Introduction

Ventricular volume is routinely assessed in various acute and chronic neurological disorders [1-5]. As manual segmentation of ventricular image studies is time consuming and unrealistic in clinical practice, assessment of ventricular volume is routinely performed by qualitative inspection or by use of measures such as Evans' index, frontal horn index, or fronto-occipital horn ratio [6-9]. These methods are prone to inter- and intra-observer variability because they rely on subjective measurements and are of limited clinical utility where small changes in ventricular size carry the potential to impact patient management. To improve clinical confidence, an accurate, automated ventricle segmentation algorithm, and subsequent volume measurement, is necessary. Conventional techniques of auto-segmentation using atlas-based methods are problematic due to the need of multiple specialized software capabilities (rigid and/or deformable registration, atlas banking) [10]. Deep Learning (DL) could provide an accurate, quantitative description of ventricular volume changes with a single algorithm.

The application of DL for auto-segmentation generally requires vast amounts of richly-labelled data for superior results [11,12], but has shown to be effective with small to modest data sets [13,14]. However, medical imaging data can be difficult to access, is often poorly labeled (i.e. segmented) and creating fine-labeled data expends significant resources. This deficiency can be overcome with weakly supervised machine Transfer Learning (TL). Weakly supervised learning is effective in three situations: incomplete, inexact and inaccurate supervision [15-17]. In most medical data sets, all three scenarios occur simultaneously. TL can be described as using a model trained for a specific task with one dataset to guide the learning of another task with a different dataset. Of note, the pretrained dataset is much larger than the subsequent set and the two tasks do not need to be identical [18]; TL has improved natural language processing when using pretrained models [19]. The application of TL requires further study for the task of auto-segmentation on multi-modality medical images.

This study had three objectives

1. Compare previously published DL architectures and optimization strategies.
2. Quantify the benefit of TL for the segmentation of brain and ventricles on magnetic resonance (MR) imaging.
3. Quantify the effectiveness of cross-modality TL for the segmentation of brain and ventricles (MR to computed tomography, CT).

The main hypothesis was that DL models with deeper, wider (more parameters and outputs) architectures and transfer learning with a larger pool of fine-labelled training data will yield superior performance for brain and ventricles segmentation tasks.

Materials and methods

A two-step process was used for the segmentation task. First, specific DL strategies were developed to segment the brain from the skull and scalp. Second, with the brain segmented, independent strategies were implemented for ventricle segmentation. Figure 1 illustrates the proposed brain and ventricle auto-segmentation design, as well as the evaluation of TL (MR-to-MR and MR-to-CT). Datasets, segmentation workflow, model architectures and training are subsequently described.

Data

With institutional review board approval, head MR and CT images from patients diagnosed with a variety of pathologies (brain cancer, traumatic injury, vascular lesions) were used. All images were cropped or padded to 128 slices of 256x256 voxels and interpolated to 3x1x1 mm. Images were normalized by subtracting the mean, dividing by the standard deviation and scaling values to 0 to 1.

Three distinct datasets were used: MR_{coarse}, MR_{fine} and CT_{fine}. The subscripts of "coarse" and "fine" reflect either coarse or fine annotated labeling of the images. The MR_{coarse} dataset included 2143 training and 15 validation image studies. The MR_{fine} dataset consisted of 112 training and 15 validation image studies. The testing set for MR_{coarse} and MR_{fine} consisted of 35 images studies and further divided into $Test_{simple}$ and $Test_{complex}$ comprised of 15 simple and 20 complex cases. $Test_{simple}$ had no pathology within the ventricles while $Test_{complex}$ contained cases with lesions, hemorrhage and/or surgical cavities altering normal ventricular morphology. CT_{fine} included 107 training, 26 validation and 27 testing images. Similarly, the CT test set was split into 9 simple and 18 complex cases. Training, validation and testing images were manually annotated by expert radiologists with fine-grained labels to serve as ground truth.

Segmentation workflow

The proposed weakly supervised TL workflow was applied to both brain and ventricle segmentation. A total of 110 MR_{fine} images were used to create a segmentation atlas using the commercial software MIM (MIM Software Inc. v6.9). These atlas-based brain and ventricle segmentation models were applied to MR_{coarse} to generate a coarse dataset – i.e. coarsely segmented brain or ventricle images. This data was used to train coarse models and served as the starting point for TL models. The noTL models were built by training an algorithm *de novo* using the MR_{fine} or CT_{fine} dataset. The TL model was created by further training the coarse model with MR_{fine} and CT_{fine} datasets. Within modality TL refers to using the pretrained MR_{coarse} model and refining it with the MR_{fine} dataset and cross-modality refers to using the pretrained MR_{coarse} model refined with the CT_{fine} dataset.

Model architectures

Several deep learning Convolutional Neural Network (CNN) architectures were tested in this work. Further details of all 11 model variants can be found in the supplemental materials.

A 3D variant of the UNet architecture [17,20] was implemented. This architecture used *convolution → activation → normalization → dropout* blocks and varying activation functions (*ReLu*, *PReLu*), normalization layers (*BatchNorm*, *InstanceNorm*, *GroupNorm*) and *dropout* rates. Different normalizations were used to overcome *BatchNorm* problems with small batch sizes due to GPU memory constraints [21]. Trilinear interpolation was used to reduce the number of parameters and decrease model inference time.

MeshNet convolutional models used dilated convolutions to increase receptive field while maintaining number of model parameters low allowing for faster training and inference [22]. This implementation used PReLu activation layer and GroupNorm normalization.

Residual 3D UNet is similar to 3D UNet, but adds skipped connections, identity mappings and pre-activation residual layers to enable deeper networks to improve performance [23,24]. This architecture used *convolution → activation function → normalization → dropout* as the main block, pre-activation residual blocks for skipped connections, a PReLu function and InstanceNorm.

Commonly used loss functions of binary cross entropy and Dice loss were implemented. Binary cross entropy required more iterations before convergence possibly due to an imbalance between foreground and background pixels in the volumetric data.

Model training and evaluation

One cycle policy, mixed precision and distributed training was performed on 8 Nvidia 2080-Ti GPUs and used the fast.ai (https://www.fast.ai) PyTorch distribution [25]. Additionally, gradual unfreezing, discriminative fine-tuning, custom splits and Adam optimization were used to achieve good performance and well-behaved loss convergence [21,24]. Early stopping and learning rate reduction were used to mitigate overfitting. Architectures were not modified when fine-tuning. Evaluation was assessed using Sorensen-Dice Score (DS) and Percent Volume Difference (PVD) from ground-truth. Two-sided Wilcoxon signed-rank test was used for hypothesis testing and $p<0.05$ was used as the significance threshold. The code and dataset can be requested at: http://www.medomics.ai/applications/ventricle-segmentation.

TL training set dependencies

The best performing model over all tasks was investigated as a function of training set size. Models without Transfer Learning (noTL) and TL models were trained with 10, 20, 40, 60, 80 and 100 randomly selected images and repeated 10 times for both MR and CT datasets. Validation Dice scores were used for comparisons to test the robustness of the model created when varying training size, as opposed to model performance on test cases.

Results

Only the best performing models - MeshNet, wide 3D UNet and wide 3D Residual UNet (baseline 9, baseline 6 and baseline 11, respectively) - are shown in the main manuscript. The details of all 11 model variants are shown in Table S1.

Brain segmentation

Dice Scores (DS) and Percent Volume Differences (PVD) for the brain segmentation task are shown in Table 1 and illustrates that most comparisons between the three architectures and models showed consistently decent results on brain segmentation with DS and PVD of roughly 0.97 and 2%, respectively. The wide 3D UNet and wide 3D Residual UNet models performed better on average in both DS and PVD than MeshNet models. However, all models had DS > 0.93 and PVD < 5%.

Ventricle segmentation

Overall, models trained from fine-grained data showed better performance than models trained on coarse data. Figure 2 shows results of model performance with respect to DS and PVD for the $Test_{simple}$ and $Test_{complex}$ cases; the corresponding table of the same results is shown in Table 2. The DS wide 3D UNet and wide 3D Residual UNet trained on fine MR data performed significantly better than the MeshNet fine MR model (wide 3D UNet fine vs. MeshNet fine, $p = 0.03$; wide 3D Resnet fine vs. MeshNet fine, $p < 0.01$) and all of the coarse MR trained models. In addition, there were large improvements in the PVD for 3D Residual UNet coarse model (21.5%) to fine model (8.02%) with similar improvements 3D UNet. For the 3D UNet and 3D Residual UNet models, the complex test cases showed slightly lower DS compared to the simple cases (3D UNet 0.79 vs. 0.86; 3D Residual UNet 0.75 vs. 0.82, for complex and simple cases, respectively), but large PVD (>25%) when compared with the simple cases. A single slice of the three MR segmentation models for the best and worst in the $Test_{simple}$ and $Test_{complex}$ cases is shown in Figure 3.

The analogous results for the fine trained CT model are shown in Figure 4 and Table 3 and the significance tests for all architectures, fine and coarse models are presented in supplemental Table S3. As expected, models trained with fine CT data outperformed the models trained on a large set of coarsely labeled MR data. For 3D Unet and 3D Resnet, DS and PVD evaluated on the $Test_{simple}$ cases were fair (around 0.75 and 18%, respectively). The results on the $Test_{complex}$ cases for both coarse and fine model were poor with DS around 0.6 and PVD of 45%. However, the overall trend of better performance using baseline 9 and 11 models was seen in the cross-modality CT data, seen below.

Correspondingly, a single slice of the three CT segmentation models for the best and worst in the $Test_{simple}$ and $Test_{complex}$ sets is shown in Figure 5.

Within- and cross-modality TL results

The results for varying the training dataset size for noTL and TL for MR and CT are shown in Figure 6. The upper panels of Figure 6 show the results using the validation DS and the lower panels show the $Test_{simple}$ case DS.

The DS results for the TL MR models were 0.86 ± 0.01 (mean $\pm$ SD) with only a slight improvement as the training sample increased, as seen in the top-left panel of Figure 6. The noTL model was 0.49 ± 0.12 when using 20 samples and improved to 0.83 ± 0.02 when using 100 samples. All MR noTL and TL DS for the same sample sizes showed statistically significance differences. The same trend was present in the CT noTL and TL comparisons. The DS for the TL CT models were 0.81 ± 0.01 (mean $\pm$ SD) with, again, only a slight improvement as the training sample increased, as seen in the top-right panel of Figure 6. Using 20 samples from the training set, the CT noTL DS was 0.44 ± 0.01 and when using 100 samples the noTL DS was 0.76 ± 0.02. Likewise, all of the CT DS results were statistically different.

Transfer learning DS showed statistical significance for $Test_{simple}$ cases trained using the full MR dataset (noTL: 0.83, TL: 0.86, $p < 0.01$). The PVD were not significantly different between noTL and TL models (8.02% and 6.87%, respectively) using the full dataset. However, the TL model trained with 20 random samples yielded a PVD of 7.71% while the noTL model trained with 20 random samples showed a PVD of 140.4% ($p < 0.01$). Similarly, both DS (0.76 vs. 0.81, $p < 0.01$) and PVD (19.15% vs. 12.79%, $p=0.04$) were found to have significant difference between full dataset trained CT noTL and TL models, respectively. Using only 20 samples in the training, the noTL and TL showed results of 0.44 vs. 0.79 and 206.5% vs. 16.15% (both statistically signifi-

cant) for DS and PVD, respectively.

The results for the $Test_{complex}$ cases, in general, were inferior than the $Test_{simple}$ cases for both within- and cross-modality models. The DS were 0.75, 0.79 and 0.78 for the full samples noTL MR, full sample TL MR and 20 sample TL MR models, respectively; the PVD values were 28.65%, 25.90% and 27.42%. For reference, the DS and PVD were 0.62 and 45.31% for the 20 sample noTL MR model. For the corresponding CT models, the DS / PVD were 0.70 / 28.99%, 0.71 / 28.03%, and 0.67 / 34.80% for the full samples noTL, full sample TL and 20 sample TL models, respectively. And, the DS and PVD were 0.50 and 105.4% for the 20 sample noTL CT model.

In general, the proposed TL workflow showed improvement in both within-modality and cross-modality models. Additionally, the TL models trained on only 20 samples performed better or equivalent to noTL models trained with a fact or five more data.

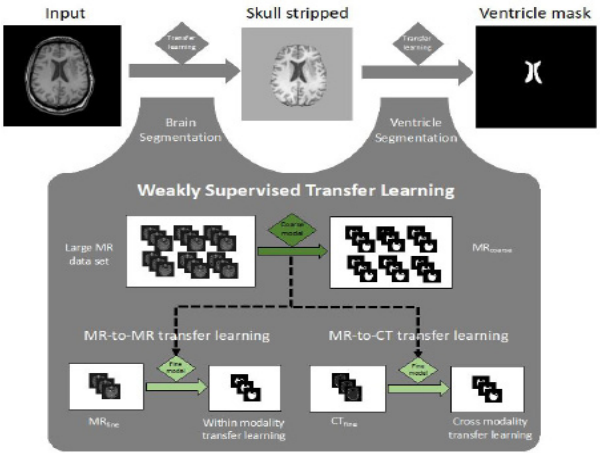

Figure 1: An illustration of the weakly supervised transfer learning workflow for brain and ventricle segmentation. Both the brain and ventricle tasks follow a similar workflow, but utilize different input image sets – the "input" and "skull stripped," respectively. Datasets are shown with rectangles and algorithmic processes are represented at arrows and diamonds.

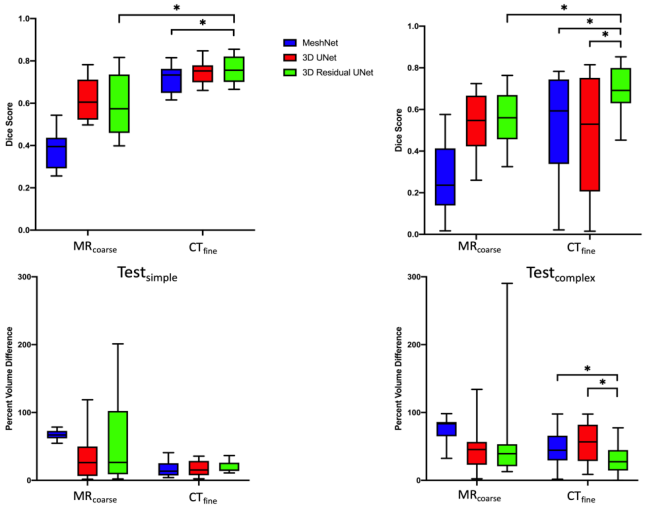

Figure 2: Results for MeshNet, 3D Unet and 3D ResNet noTL MR ventricle segmentation models. The left column shows results for easy test cases and the right shown hard test cases; the top row shows the dice scores and the bottom row shows percent volume differences (inter-quartile ranges are shown). The tabular form along with all statistical significance tests is shown in Table 2 and supplemental Table S5.

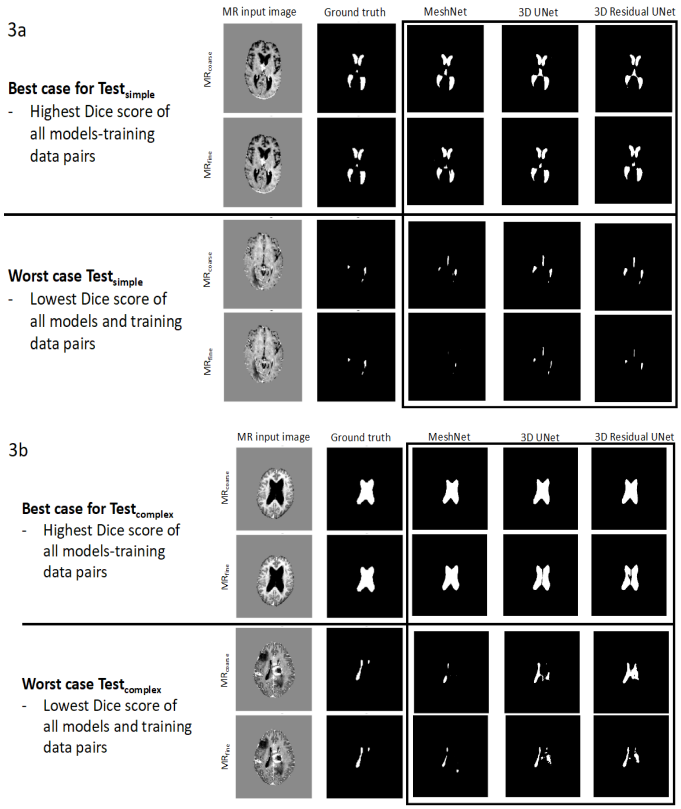

Figure 3: (a): Single slice ventricle segmentation noTL MR predictions using MeshNet, 3DUNet, and 3D Residual UNet model architectures shown for the image set with the highest and lowest dice score from $Test_{simple}$. The top rows show the best and worse dice scores for the MR_{coarse} dataset and the bottom rows show the MR_{fine}. The results of the segmentation for the full image set can be accessed in the supplemental material. **(3b)**: The corresponding figure showing results for the $Test_{complex}$ case.

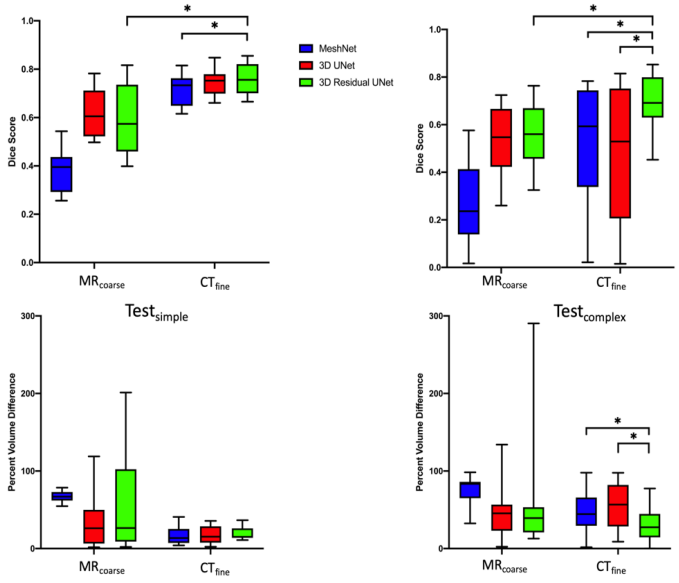

Figure 4: Results for MeshNet, 3D Unet and 3D ResNet noTL CT ventricle segmentation models. The left column shows results for easy test cases and the right shown hard test cases; the top row shows the dice scores and the bottom row shows percent volume differences (inter-quartile ranges are shown). The tabular form along with all statistical significance tests is shown in supplemental Table S4 and Table S5.

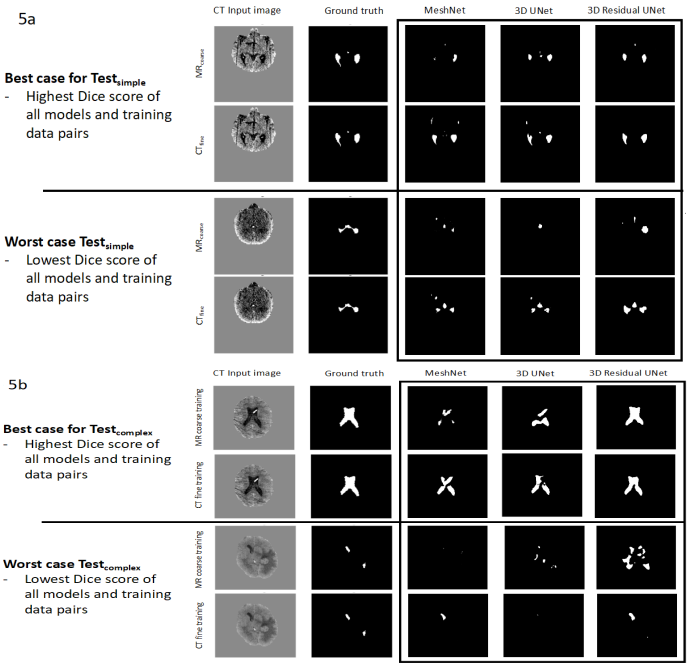

Figure 5: (a): Single slice ventricle segmentation noTL CT predictions using MeshNet, 3DUNet, and 3D Residual UNet model architectures shown for the image set with the highest and lowest dice score from $Test_{simple}$. The top rows show the best and worse dice scores for the MR_{coarse} dataset and the bottom rows show the CT_{fine}. The results of the segmentation for the full image set can be accessed in the supplemental material. **(b)**: The corresponding figure showing results for the $Test_{complex}$ case.

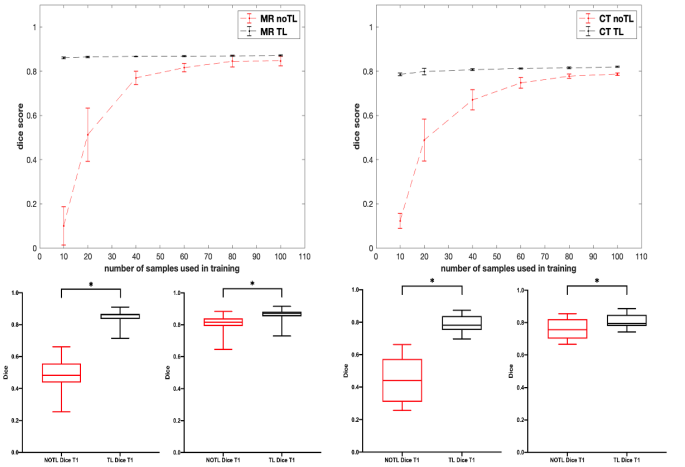

Figure 6: (a): The results of noTL compared to TL for both MR and CT models are shown here; the noTL and TL results are shown in red and black, respectively in this whole figure. The top row illustrates the dice score of the validation set as a function of the number of samples used in the training for the MR dataset (left) and CT dataset (right). The points show the average dice score for the correspond dataset and the error bar represents two times the standard deviation. The bottom row shows the box-and-whisker dice scores evaluated on the $Test_{simple}$ cases: MR and CT showed on the left and right, correspondingly. In each panel, the left shows results for training with 20 samples and the right shows the entire dataset. The differences were statistically significant for both sets of comparisons.

Brain segmentation results:

Table 1: Brain segmentation results for 3 different model architectures using coarse MR data, fine MR data and fine CT data. Each model was trained without transfer learning (*de novo*) Shown are the Dice score and percent volume difference (average +/- standard deviations).

Architecture	Metric	MR_{coarse}	MR_{fine}	CT_{fine}
3D UNet	Dice	0.96 +/- 0.01	0.96 +/- 0.01	0.98 +/- 0.01
	% volume difference	1.16 +/- 0.86	2.24 +/- 1.46	1.60 +/- 0.86
MeshNet	Dice	0.96 +/- 0.01	0.94 +/- 0.01	0.93 +/- 0.03
	% volume difference	1.51 +/- 1.39	3.99 +/- 2.82	4.94 +/- 2.77
3D Residual UNet	Dice	0.97 +/- 0.01	0.96 +/- 0.01	0.97 +/- 0.01
	% volume difference	1.22 +/- 0.87	3.22 +/- 2.23	3.06 +/- 1.74

Brain segmentation results:

Table 2: Corresponding tabular form of Figure 2 ventricle segmentation results. Results of the three best performing variants of 3D UNet, MeshNet and 3D Residual UNet Dice scores and percent volume difference evaluated for the simple and complex cases on the MR_{coarse} and MR_{fine} models. The mean +/- the standard deviation is shown.

		MR_{coarse}		MR_{fine}	
		$Test_{easy}$	$Test_{difficult}$	$Test_{easy}$	$Test_{difficult}$
3D UNet (baseline 9)	Dice	0.79 +/- 0.08	0.73 +/- 0.19	0.86 +/- 0.06	0.79 +/- 0.13
	% volume difference	26.74 +/- 29.00	24.06 +/- 10.96	7.96 +/- 8.61	21.97 +/- 11.79
MeshNet (baseline 6)	Dice	0.78 +/- 0.07	0.67 +/- 0.20	0.76 +/- 0.07	0.65 +/- 0.16
	% volume difference	18.87 +/- 16.27	37.22 +/- 16.22	19.40 +/- 10.31	43.58 +/- 13.87
3D Residual Unet (baseline 11)	Dice	0.80 +/- 0.08	0.73 +/- 0.11	0.82 +/- 0.07	0.75 +/- 0.12
	% volume difference	21.52 +/- 26.84	32.38 +/- 20.27	8.02 +/- 9.29	28.65 +/- 8.74

Table 3: Corresponding tabular form of Figure 4 ventricle segmentation results. Results of the three best performing variants of 3D UNet, MeshNet and 3D Residual UNet Dice scores and percent volume difference evaluated for the simple and complex cases on the MR_{coarse} and CT_{fine} models. The mean +/- the standard deviation is shown.

		MR_{coarse}		CT_{fine}	
		$Test_{easy}$	$Test_{difficult}$	$Test_{easy}$	$Test_{difficult}$
3D UNet (baseline 9)	Dice	0.62 +/- 0.10	0.52 +/- 0.15	0.75 +/- 0.06	0.48 +/- 0.30
	% volume difference	35.18 +/- 36.69	46.23 +/- 29.41	17.57 +/- 11.71	55.02 +/- 30.31
MeshNet (baseline 6)	Dice	0.38 +/- 0.09	0.28 +/- 0.18	0.72 +/- 0.07	0.54 +/- 0.23
	% volume difference	67.10 +/- 7.38	75.33 +/- 16.83	17.30 +/- 12.08	46.62 +/- 27.70
3D Residual Unet (baseline 11)	Dice	0.60 +/- 0.15	0.56 +/- 0.13	0.76 +/- 0.06	0.70 +/- 0.11
	% volume difference	55.36 +/- 68.20	60.64 +/- 74.81	19.15 +/- 8.50	30.00 +/- 21.05

Discussion

Overall, this analysis found deeper, wider CNNs had improved performance over other architectures for brain and ventricle segmentation tasks. Our findings are in agreement with other reports which utilized deeper and wider networks for various medical applications, such as segmentation in head and neck cancer cases [11,12,13,14,26,27]. Deeper, wider brain models showed minor, non-statistically significant improvement due to the simplicity of the task. For ventricle segmentation, deep, wide architectures DS and PVD showed significant improvement emphasizing that applying incorrect architectures can greatly affect segmentation task performance. This work presents a superior alternative to current qualitative and semi-quantitative methods for ventricular segmentation by introducing an autonomous, quantitative method that achieves significant improvement compared to current semantic ventricular segmentation solutions.

Deep, wide models trained with one hundred fine-grained images outperformed models that were trained with an order of magnitude more coarse-grained images. Specifically, the PVD results were far superior in fine models and also slightly better than other reported results [28,29]. $Test_{complex}$ DS for fine models were similar to coarse models. However, fine model PVD was roughly a factor of three better than corresponding coarse models. $Test_{complex}$ results could be further improved with a model trained on a finely labeled dataset of ventricles containing lesions, surgical cavities or other pathologies.

The TL workflow was equivalent or superior to the noTL approach for investigated tasks. The TL approach was equivalent in brain segmentation and was superior for the ventricle segmentation task. Within-modality TL ventricle segmentation DS were statistically superior than *de novo* models for $Test_{simple}$ and better, although non-statistically significant, PVD. Results were roughly equivalent on complex cases. Cross-modality TL also showed improvement compared to noTL approach and could be further improved with a pathology-specific fine-grained training dataset. The use of both TL methodologies could significantly aid in the creation of models to facilitate better care to patients in the neurologic community.

Weakly supervised TL was superior to *de novo* training with respect to the training data set size. The TL experiments showed significant DS and PVD improvements for all dataset sizes tested. An acceptable model can be created with a low volume of finely annotated data when combined with a larger set of coarse-labeled data. This is especially useful for small institutions where the creation of a finely labeled data set can expend substantial clinical resources and/or a large volume of patient data is unavailable for model creation. It has been reported in other studies that models trained with one institution's data underperforms when tested at a separate institution [29]. A coarse segmentation model created with data from multiple institutions could be created first and then a clinic-specific, fine-grained TL model then produced for certain, specific segmentation tasks thereby surmounting this previously reported issue.

The results of this study highlight the feasibility of a technique that could have wide-ranging impact for the neuroimaging community. In addition to being important in decisions of ventricular drainage management, accurate quantification of changes in ventricular volume is crucial in the assessment of neurodegenerative and psychiatric disease, as well as white matter disorders and in specific radiation treatments [1-5,31-34]. Future work would include further optimization of models to improve performance. Image selection could be optimized for the coarse-grained model [35] and combined with fine-grained datasets in the presence of significant ventricle distortion and/or catheters present to yield superior segmentation models. In addition, these improvements would subsequently lead to a superior weakly supervised TL model. Further refinement of architectures parameters and inclusion of combined loss functions could also improve resulting within- and cross-modality TL models.

Limitations of this study include the relatively small test sizes and lack of complete optimization of all of the experiments performed. The smallest number of test cases for the simple studies were roughly 10% of training size. Regarding the optimization, additional permutations of architecture output channels, normalization, activation, etc. could be investigated to completely optimize the algorithm for a given task. However, these concerns are mitigated by the fact that a main purpose of the study was the relative, inter-comparison of several variations of the multiple architectures investigated in this study. Nevertheless, the overall results of the best model showed very good performance with regards to both DS (~0.86) and PVD (~7%) compared to ground truth, and were comparable or superior to other reports [22,23].

Conclusion

Weakly supervised TL for segmentation tasks utilizing deep, wide networks trained with a small volume of fine-grained data demonstrates superior performance compared to shallow networks trained with larger volumes of coarse-grained data without the application of TL. These improvements were judged with two clinically relevant metrics (DS and PVD) and were seen in both within-modality and cross-modality TL models. Including fine-grained image data with both abnormal ventricle shapes and the presence of ventriculoperitoneal shunts in further training sets will improve models leading to a quantitative tool to aid clinical decisions. The benefits of TL models combined with relatively small amounts of fine-detailed datasets needs to be further explored for multiple segmentation tasks.

Key points

- Deep and wide residual CNNs perform better than other baseline architectures for the clinical tasks investigated in this study.
- TL methods when combined with weak supervision reduces the need of generating large, fine labelled data sets that can cost significant clinical resources
- In particular tasks, cross-modality TL, e.g. TL from MR to CT, and within-modality TL, e.g. from MR to MR shows superior performance than noTL models trained with moderate-sized datasets.

Summary statement

Weakly supervised transfer learning for deep learning tasks are equivalent or superior to atlas-based approaches, and transfer learning models can overcome the need for large, fine-detailed datasets in medical imaging domains where data collection can expend significant clinical resources or be completely unavailable.

References

1. Hughes J D, Puffer R, Rabinstein A A: Risk factors for hydrocephalus requiring external ventricular drainage in patients with intraventricular hemorrhage. J Neurosurg. 2015; 123: 1439-1446.
2. Anderson RC, Grant JJ, de la Paz R, Frucht S, Goodman RR: Volumetric measurements in the detection of reduced ventricular volume in patients with normal-pressure hydrocephalus whose clinical condition improved after ventriculoperitoneal shunt placement. J Neurosurg. 2002; 97:73-79.
3. Pinggera D, Kerschbaumer J, Petr O, Ortler M, Thomé C, Freyschlag C F: The volume of the third ventricle as a prognostic marker for shunt dependency after aneurysmal subarachnoid hemorrhage. World Neurosurg. 2017; 108: 107-111.
4. Apostolova LG, Beyer M, Green AE, Hwang KS, Morra JH, Chou Y, et al: Hippocampal, caudate, and ventricular changes in Parkinson's disease with and without dementia. Movement Disorders. 2010; 25:687-695.
5. Ferrarini L, Palm WM, Olofsen H, van Buchem MA, Reiber JHC, Admiraal-Behloul F: Shape differences of the brain ventricles in Alzheimer's disease. NeuroImage. 2006; 32: 1060-1069.
6. Toma AK, Holl E, Kitchen ND, Watkins LD. Evans' index revisited: The need for an alternative in normal pressure hydrocephalus. Neurosurgery. 2011; 68: 939-944
7. Ragan DK, Cerqua J, Nash T, McKinstry RC, Shimony JS, et al: The accuracy of linear indices of ventricular volume in pediatric hydrocephalus. Pediatr Neurosurg. 2015; 15: 547-551.
8. Barr AN, Heinze WJ, Dobben GD, Valvassori GE, Sugar O: Bicaudate index in computerized tomography of Huntington disease and cerebral atrophy. Neurology. 1978; 28: 1196-1200.
9. Jamous M, Sood S, Kumar R, Ham S: Frontal and occipital horn width ratio for the evaluation of small and asymmetrical ventricles. Pediatr Neurosurg. 2003; 39:17-21.
10. Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy, Med Phys. 2014; 41: 050902.
11. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, et al. Deep learning and alternative learning strategies for retrospective real-world clinical npj Digital Medicine. 2019: 2.
12. Wang F, Casalino LP, Khullar D. Deep learning in medicine—promise, progress, and challenges. JAMA Intern. Med. 2019; 179; 293-294.
13. Chan JW, Kearney V, Haaf S, Wu S, Bogdanov M, et al. A Convolutional Neural Network Algorithm for Automatic Segmentation of Head and Neck Organs-at-Risk Using Deep Lifelong Learning. Med Phys. 2019; 46: 2204-2213.
14. Men K, Chen X, Zhang Y, Dai J, Yi J, et al. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning CT images. Front Oncol. 2017; 7: 315.
15. Zhou Z. A brief introduction to weakly supervised learning, Natl Sci Rev. 2018; 5:44-53.
16. Mahajan D, Girshick R, Ramanathan V, He K, Paluri M, et al. Exploring the Limits of Weakly Supervised Pretraining arXiv:1805.00932. 2018.
17. Çiçek O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation arXiv: 1606.06650.
18. Tan C, Sun F, Kong T, Zhang W, Yang C, et al. A Survey on Deep Transfer Learning arXiv: 1808.01974.
19. Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification arXiv: 1801.06146.
20. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge arXiv: 1802.10508.
21. Wu Y, He K. Group Normalization arXiv: 1803.08494.
22. Fedorov A, Johnson J, Damaraju E, Ozerin A, Calhoun V, Plis S, End-to-end learning of brain tissue segmentation from imperfect labeling arXiv: 1612.00940.
23. He K, Zhang X, Ren S, Sun J. Identity Mappings in Deep Residual Networks arXiv: 1603.05027.
24. He T, Zhang Z, Zhang H, Zhang Z, Xie J, et al. Bag of Tricks for Image Classification with Convolutional Neural Networks arXiv: 1812.01187.
25. Smith LN. A disciplined approach to neural network hyper-parameters arXiv: 1803.09820.
26. Shin HC, Roth HR, Gao M, Lu L, Xu Z, et al. Mollura D, Summers RM Deep convolutional neural networks for Computer-Aided Detection: CNN architectures, dataset characteristics and transfer learning IEEE Trans. Med. Imaging. 2016; 35:1285-1298.
27. Kearney V, Ziemer BP, Perry A, Wang T, Chan J, et al. Attention Aware Discrimination for MRI to CT Translation Using Cycle-Consistent Generative Adversarial Networks (accepted by Radiology AI 9/2019).

28. Huff TJ, Ludwig PE, Salazar D, Cramer JA. Fully automated intracranial ventricle segmentation on CT with 2D regional convolution neural network to estimate ventricular volume, Int J Comput Assist Radiol Surg. 2019.

29. Liu J, Huang S, Ihar V, Ambrosius W, Lee LC, et al. Automatic model-guided segmentation of the human brain ventricular system from CT images. Acad Radiol. 2010; 17: 718-726.

30. AlBadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: impact of cross-institutional training and testing, Med Phys. 2018; 45:1150-1158.

31. Virhammar J, Laurell K, Cesarini KG, Larsson EM. Increase in callosal angle and decrease in ventricular volume after shunt surgery in patients with idiopathic normal pressure hydrocephalus, J Neurosurg. 2018; 130: 130-135.

32. Novak G, Fox N, Clegg S, Nielsen C, Einstein S, et al. Changes in Brain Volume with Bapineuzumab in Mild to Moderate Alzheimer's Disease J Alzheimers Dis. 2016; 49: 1123-1134.

33. Guo JY, Huhtaniska S, Miettunen J, Jääskeläinen E, Kiviniemi V, et al. Longitudinal regional brain volume loss in schizophrenia: Relationship to antipsychotic medication and change in social function, Schizophrenia Research. 2015; 168: 297-304.

34. Ghione E, Bergsland N, Dwyer MG, Hagemeier J, Jakimovski D, et al. Aging and Brain Atrophy in Multiple Sclerosis. J Neuroimaging. 2019; 29: 527-535.

35. Aljabar P, Heckermann RA, Hammers A, Hajnal JV, Rueckert D. Multi-altas based segmentation of brain images: Atlas selection and it effect on accuracy, Neuroimage. 2009; 46: 726-738.