Transfer Learning to Improve Pediatric Spinal Cord Gray Matter Segmentation

Ashwin Kumar

Department of Computer Science, Vanderbilt University

1. INTRODUCTION

Spinal cord anatomy consists of many different anatomical structures,[1,2] nuclei,[1,2] laminae,[2,3] and ascending and descending tracts[2,4] though the structures can be broadly differentiated between gray matter (GM) and white matter (WM).[2,5] Specifically, GM is involved in such vital functions as motor movement, memory, and emotions.[6] Therefore, spinal cord GM segmentation can provide us with valuable spinal cord spatial information, allow us to quantify GM atrophy due to neuropathologies and injury,[7] facilitate image analysis,[8] and can be used to extract multi-parametric MRI metrics.[7]

The spinal cord however is an inherently small structure in shape and cross-sectional size, which makes it difficult to both conduct MRI and post-acquisition segmentation.[5] Segmenting GM and WM is an important step for accurate tissue analysis in magnetic resonance imaging.[9] Specifically, prior to deep learning, such techniques as iterative non-local statistical fusion[10] and groupwise multi-atlas segmentation[11] did not achieve near-human performance and maintained Dice performance less than 0.80. Through the advent of deep learning, several automatic deep learning based segmentation algorithms have been developed[9,12–16] to automatically segment GM in the spinal cord with performances superior to non-deep learning approaches.

The Spinal Cord Gray Matter (SCGM) Challenge[9] featured six independently developed methods, with and without deep learning, that had good overall results to detect the GM butterfly but the methods had variable performance among the tested segmentation metrics. Following the challenge, Perone et al. developed a spinal cord GM segmentation technique by taking advantage of deep learning[13] and was publicly implemented part of *sct_deepseg_gm* in the Spinal Cord Toolbox (SCT).[17] Their proposed architecture was based on Atrous Spatial Pyramid Pooling (ASPP)[18] and uses dilated convolutions.[19] Specifically, their architecture works with 2D slice-wise axial images and consists of (i) two layers of 3 x 3 convolutions, (ii) two layers of dilated convolution of r=2, (iii) six parallel branches with two layers each of a 1 x 1 convolution and four different dilated convolutions, and (iv) a block of 2 layers of 1 x 1 convolutions with dense output. There were no residual connections and each layer was followed by batch normalization[20] and dropout.[21] They further implemented a surrogate loss function called Dice Loss to handle the unbalanced nature of GM segmentation. Their model was trained on 80 healthy adult subjects, which was split between training (40) and test (40) sets. Their proposed method performed superior to the methods submitted to the SCGM challenge on several such metrics as dice similarity coefficient (DSC), mean surface distance (MSD), Hausdorff surface distance (HSD), skeletonized Hausdorff distance (SHD), etc.

Though these algorithms have performed well on the adult spinal cord datasets, spinal cord segmentation algorithms trained for the adult cord have had sub-optimal translation to the pediatric population even for cross-sectional area segmentation.[22] Segmenting gray matter in the pediatric population is important especially to detect such gray matter diseases as acute disseminated encephalomyelitis,[23] multiple sclerosis, and acute transverse myelitis.[24] Therefore, the goal of this project was to develop a more accurate segmentation model that can facilitate pediatric gray matter segmentation. More clearly,

the *sct_deepseg_gm*[13] model was trained on a clinical pediatric spinal cord dataset by initializing the model architecture with the SCGM challenge pretrained weights, which was then trained or "transfer learned" on the pediatric dataset. The project was able to show that transfer learning on the pediatric dataset and data augmentation improves pediatric gray matter segmentation performance. The code for this project is publicly available at: https://github.com/ashwinkumargb/pediatric_gmseg.

2. METHODS

2.1 Dataset and Pre-Processing

High resolution multi-echo fast field echo (mFFE) scans were retrospectively acquired from pediatric spinal cord scans from 68 patients at the Monroe Carrel Jr. Children's Hospital. All studies were performed under local institutional review board approval (AAA_171784). Imaging was conducted using a 3T whole-body MR imaging scanner (Philips Healthcare, Best, Netherlands). mFFE volumes consisted of 25-70 mm variable lengths and were imaged in the cervical and thoracic levels. The volumes were then resampled to have axial slice dimensions of 256 x 256, first-pass segmented using *sct_deepseg_gm*[13] model, and then manually corrected by AK.

2.2 Dataset Split

The dataset was then balanced to handle skew from different cord regions. The obtained data had an increased subject count in the cervical and upper-thoracic cord region but not much representation in the lower-thoracic cord region. Therefore, relatively equal vertebral representation from the cervical, upper-thoracic, and lower-thoracic cord was maintained by pseudorandomly, manually assigning subject scans to training, validation, and testing datasets. After skew correction, the initially 68 subjects were divided into a roughly 60/20/20 split resulting in training (41), validation (13), and testing (14) datasets.

2.3 Data Augmentation

Data augmentation was conducted to increase the size of training and validation datasets 16-fold. The augmenter pipeline was designed to apply the following augmenters in 50% of all cases: crop and pad -10% and 10% of the height/width; affine transformation consisting of scaling on the x- and y-axis by 80-110% of their size, translating x- and y-axis by -10 to +10% of their size, rotating image by -25 to +25 degrees; Gaussian blur with sigma between 0 and 1.5; local means blur with kernel sizes between 2 and 4; sharpening and increasing lighting of images; adding Gaussian noise by scaling from 0 to mean intensity of dataset images; and improving and worsening the contrast. The dataset images were of type `float32` and the dataset ground truth segmentation were of type `uint8` for efficient memory utilization.

2.4 Model Description and Training

To reiterate, the deep dilated CNN model[13] architecture was used for training, and the pretrained weights were loaded based on the model trained on the SCGM challenge dataset. The model was then modified to ensure compatibility with Tensorflow (v1.8.0), CUDA (v11.1), and Keras (v2.20). The following models and data combinations were trained: (1) pretrained model without data augmentation, (2) model (no pretraining) without data augmentation, (3) pretrained model with data augmentation, and (4) model (no pretraining) with data augmentation. The models were trained on the Vanderbilt ACCRE high-performance cluster to take advantage of GPU computing for efficiency. The DSC surrogate loss function was also implemented to evaluate model performance during training and validation. Specifically, early stopping was implemented based the validation dice loss with a no improvement minimum change marker of 0.001 after 5 epochs.

2.5 Performance Metrics

To evaluate the performance of the models, performance metrics were chosen based on the categorical representation of overlap, distance, and statistics.[9] The overlap segmentation metrics included the Dice similarity coefficient (DSC), Jaccard index (JI), and conformity coefficient (CC). The distance based segmentation metric include mean surface distance (MSD) and Hausdorff surface distance (HSD). The statistical based metrics include true positive rate (TPR), true negative rate (TNR), and positive predictive value (PPV). All metrics were evaluated on a held-out test set (60/20/20 split). DSC was decided as the most important metric because it remains insensitive to imbalancing on GM segmentation targets and is employed by many medical imaging works.[25,26]

3. RESULTS

The final results from the baseline model and four model and data combination are described in section 2.4 and were evaluated on the performance metrics described in section 2.5. These results are visualized in Figure 1.

4. DISCUSSION AND CONCLUSION

4.1 Evaluation of Results

As mentioned in section 2.5, the overlap segmentation metrics include DSC, JI, and CC with an emphasis placed on DSC. The DSC median scores between baseline, pretrained, and the augmented pretrained models were relatively similar with the no pretrained model and augmented and not pretrained model performing worse comparatively (Fig. 1). Though the mean DSC between baseline, pretrained, and augmented pretrained models differed significantly with augmented pretrained model having a better mean DSC comparatively. Through visual analysis, it is clear that though the median DSC may not have much improvement between the baseline and augmented pretrained model, the mean DSC improves, suggesting accuracy improvements, and the IQR reduces, suggesting precision improvements. Further, the non-pretrained and pretrained models had greater precision compared to the baseline model, which maintained worse mean DSC values. The Jaccard Index showed similar trends to the DSC performance metric. The CC metric though showed differences especially between baseline and the models trained on the pediatric dataset regardless of pretraining. Although median values between the baseline and pretrained values were relatively similar, the baseline mean was negative. This means that on average in the baseline model that there was a lower ratio between mis-segmented and correctly segmented images. As observed in the DSC and JI metrics, the precision and mean of the baseline model was lower compared to all models trained on the pediatric datasets.

The distance segmentation metrics include MSD and HSD. The MSD median scores between baseline, pretrained, and the augmented pretrained models were relatively similar, but the mean MSD scores significantly improved in the pretrained models especially in the augmented pretrained model (Fig. 1). Further, the precision improved in the pretrained and augmented models compared to the baseline model. This suggests that the there is a reduced mean distance between mask countours in the augmented pretrained model. The HSD values further followed a similar trend to MSD though the non-pretrained model performed worse compared to all models. This shows that the longest euclidean distance between mask counters was most precise and accurate in the augmented pretrained model.

The statistical based metrics include TPR, TNR, and PPV (Fig. 1). The TPR shows that all four models trained on the pediatric dataset had a better mean and median compared to the baseline model with the augmented pretrained model having the highest median and mean values. This means that

the pediatric trained models reduce under-segmentation on the pediatric dataset. The TNR values were relatively similar between all models with a range between medians and means of around 0.05. Consequently, it appears that all models have a well segmented background. The PPV results suggest that the baseline model tends to be less prone to over-segmentation compared to the models trained on the pediatric dataset.

Based on the above results (Fig. 1), there exists evidence that the augmented pretrained model performs better compared to baseline and the other pediatric dataset trained models. Further, it is clear that pretraining improves model performance and data augmentation can help though the combination of data augmentation and pretraining remains most efficacious. There also does not need to be much pediatric data to obtain good performance as noted by the well performing non-augmented models, which provides credence to the architecture.[13] Perhaps the reason the baseline model had high median values was due to a subset of pediatric regions showing similarities to the adult spinal cord, but when pediatric specific differences emerged, the performance varied greatly. Most notably, the augmented pretrained model improves both accuracy and precision and had the best DSC performance compared to the other models.

4.2 Future Direction

Since the augmented pretrained model works better compared to baseline, it could be beneficial to disseminate the model to the larger imaging community, so that other researchers working with pediatric datasets can have more accurate, automatic segmentation. It would further be important to validate the results through more robust statistical analysis and determine whether the observed effects are significant.

4.3 Limitations

The ground truth segmentation for the training, validation, and testing targets was conducted by AK, so not having multiple raters to evaluate ground truth may result in the model having more biases. Further, the model data is designed for healthy pediatric patients and it would be beneficial to add pediatric patients with neuropathologies to enable the model to learn about segmenting GM pathologies.

REFERENCES

[1] Cramer, G. D. and Darby, S. A., "Clinical anatomy of the spine, spinal cord, and ans," (2017).

[2] Byrne, J. and Dafny, N., "Neuroscience online: An electronic textbook for the neurosciences," *Department of Neurobiology and Anatomy, The University of Texas Medical School at Houston* (1997).

[3] Chung, K., Carlton, S., Westlund, K., and Briner, R., "Immunohistochemical localization of seven different peptides in the human spinal cord," *Journal of Comparative Neurology* **280**(1), 158–170 (1989).

[4] Kuypers, H. G., "The descending pathways to the spinal cord, their anatomy and function," *Progress in brain research* **11**, 178–202 (1964).

[5] Stroman, P. W., Wheeler-Kingshott, C., Bacon, M., Schwab, J., Bosma, R., Brooks, J., Cadotte, D., Carlstedt, T., Ciccarelli, O., Cohen-Adad, J., et al., "The current state-of-the-art of spinal cord imaging: methods," *Neuroimage* **84**, 1070–1081 (2014).

[6] Mercadante, A. A. and Tadi, P., "Neuroanatomy, gray matter," (2020).

[7] De Leener, B., Taso, M., Cohen-Adad, J., and Callot, V., "Segmentation of the human spinal cord," *Magnetic Resonance Materials in Physics, Biology and Medicine* **29**(2), 125–153 (2016).

[8] Patil, D. D. and Deore, S. G., "Medical image segmentation: a review," *International Journal of Computer Science and Mobile Computing* **2**(1), 22–27 (2013).

[9] Prados, F., Ashburner, J., Blaiotta, C., Brosch, T., Carballido-Gamio, J., Cardoso, M. J., Conrad, B. N., Datta, E., Dávid, G., De Leener, B., et al., "Spinal cord grey matter segmentation challenge," *Neuroimage* **152**, 312–329 (2017).

[10] Asman, A. J., Smith, S. A., Reich, D. S., and Landman, B. A., "Robust gm/wm segmentation of the spinal cord with iterative non-local statistical fusion," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 759–767, Springer (2013).

[11] Asman, A. J., Bryan, F. W., Smith, S. A., Reich, D. S., and Landman, B. A., "Groupwise multi-atlas segmentation of the spinal cord's internal structure," *Medical image analysis* **18**(3), 460–471 (2014).

[12] Alsenan, A., Youssef, B. B., and Alhichri, H., "A deep learning model based on mobilenetv3 and unet for spinal cord gray matter segmentation," in [*2021 44th International Conference on Telecommunications and Signal Processing (TSP)*], 244–248, IEEE (2021).

[13] Perone, C. S., Calabrese, E., and Cohen-Adad, J., "Spinal cord gray matter segmentation using deep dilated convolutions," *Scientific reports* **8**(1), 1–13 (2018).

[14] Porisky, A., Brosch, T., Ljungberg, E., Tang, L. Y., Yoo, Y., Leener, B. D., Traboulsee, A., Cohen-Adad, J., and Tam, R., "Grey matter segmentation in spinal cord mris via 3d convolutional encoder networks with shortcut connections," in [*Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*], 330–337, Springer (2017).

[15] Horváth, A., Tsagkas, C., Andermatt, S., Pezold, S., Parmar, K., and Cattin, P., "Spinal cord gray matter-white matter segmentation on magnetic resonance amira images with md-gru," in [*International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*], 3–14, Springer (2018).

[16] Paugam, F., Lefeuvre, J., Perone, C. S., Gros, C., Reich, D. S., Sati, P., and Cohen-Adad, J., "Open-source pipeline for multi-class segmentation of the spinal cord with deep learning," *Magnetic resonance imaging* **64**, 21–27 (2019).

[17] De Leener, B., Lévy, S., Dupont, S. M., Fonov, V. S., Stikov, N., Collins, D. L., Callot, V., and Cohen-Adad, J., "Sct: Spinal cord toolbox, an open-source software for processing spinal cord mri data," *Neuroimage* **145**, 24–43 (2017).

[18] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H., "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587* (2017).

[19] Yu, F. and Koltun, V., "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122* (2015).

[20] Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in [*International conference on machine learning*], 448–456, PMLR (2015).

[21] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research* **15**(1), 1929–1958 (2014).

[22] Kumar, A., Vandekar, S., Schilling, K., Bhatia, A., Landman, B. A., and Smith, S., "Mapping pediatric spinal cord development with age," in [*Medical Imaging 2022: Image Processing*], **12032**, 286–292, SPIE (2022).

[23] Baum, P. A., Barkovich, A. J., Koch, T., and Berg, B., "Deep gray matter involvement in children with acute disseminated encephalomyelitis.," *American journal of neuroradiology* **15**(7), 1275–1283 (1994).

[24] Lu, V. M. and Niazi, T. N., "Pediatric spinal cord diseases," *Pediatrics In Review* **42**(9), 486–499 (2021).

[25] Milletari, F., Navab, N., and Ahmadi, S.-A., "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in [*2016 fourth international conference on 3D vision (3DV)*], 565–571, IEEE (2016).

[26] Drozdzal, M., Chartrand, G., Vorontsov, E., Shakeri, M., Di Jorio, L., Tang, A., Romero, A., Bengio, Y., Pal, C., and Kadoury, S., "Learning normalized inputs for iterative estimation in medical image segmentation," *Medical image analysis* **44**, 1–13 (2018).

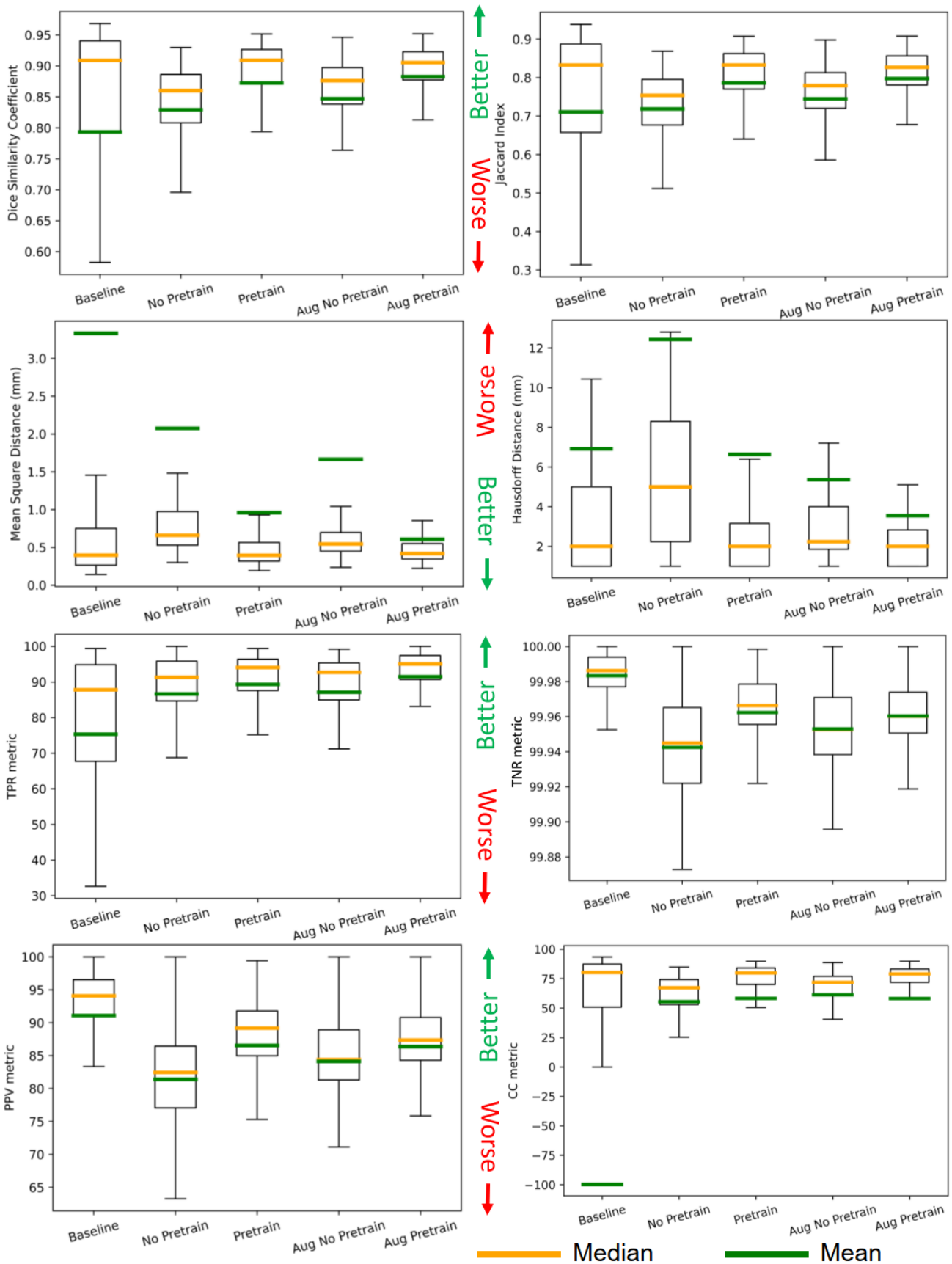

Figure 1. Test set evaluation on the pediatric dataset using the baseline and models listed in section 2.4. For fair comparison, the metrics are similar to the ones used in[10,13] and the results from the overlapping, distance, and statistical based performance metrics are visualized here. It is important to note that MSD and HSD are both in millimeters and that lower values mean better results.