Label Aware Denoising Pretraining

D. Ninalga[1]

[1] University of Toronto, Toronto, Canada

Abstract— **Most large-scale pre-trained image models are not designed with segmentation or medical imaging in mind. Hence, practitioners often use specialized augmentation techniques such as CarveMix and denoising pretraining objectives to initialize and train their models. However, these methodologies may misappropriate model resources for learning task-irrelevant information as they do not incorporate label information. We propose Label Aware Denoising Pretraining (LADP), a deep learning model pretraining technique for hypoxic-ischemic encephalopathy lesion segmentation, which causes severe motor and cognitive disability and high mortality in neonates. LADP uses the region-of-interest extraction method from CarveMix to impart increasing levels of noise to regions surrounding lesion contours. In this way, models efficiently learn better representations for a few key areas most relevant to the downstream task.**

Keywords— **Segmentation, Hypoxic-ischemic Encephalopathy, Denoising Pretraining**

I. Introduction

Hypoxic-ischemic encephalopathy (HIE) is the fifth leading cause of worldwide deaths of children according to the World Health Organization [1]. HIE results in a brain injury in neonates and occurs in approximately 1 - 5 per 1000 births [2, 3]. Neonates diagnoses with mild HIE are four times more likely to develop either cerebral palsy, epilepsy, mental retardation, or die before the age of six [4]. In severe cases, 93% of neonates report multiple organ failures [5], and about 22% die due to bleeding complications [6]. The diagnosis and prognosis of HIE is a multi-factorial process that most often involves neuroimaging [7]. In particular, neuroimaging with magnetic resonance imaging (MRI) has the most prognostic importance as it allows for accurate detection of lesions related to HIE [8]. MRIs are used in studies to predict long-term outcomes, identify common patterns, and inform treatment decisions in practice [9]. Hence, strong prediction and segmentation tools for detecting hypoxic-ischemic lesions may help further the understanding of the associated neurological factors, assist in prognosis, and ultimately help guide patient care.

As surveyed by [10], many automated segmentation techniques developed for isolating brain lesions in MRI using deep learning have been proposed over the last few years. Indeed, many of these proposed techniques often incorporate two components: pre-training and transfer-learning [11]. These methods can significantly improve sample efficiency, benefiting the training of deep learning models that typically require thousands of samples for good performance. *Self-pretraining* [12] is a recently proposed transfer-learning technique where models are pre-trained directly on the downstream task data. Models pre-trained exclusively for ImageNet classification [13], the most popular transfer-learning technique, often suffer from a degradation in performance segmentation tasks [14]. Given the low incidence rate of the disease, the size of datasets and the statistical power of MRI studies of HIE is limited. Hence, we use self-pretraining to initialize our models.

Recently, [15] demonstrated that self-pretraining using the denoising pretraining objective (without ImageNet pretraining) can outperform their ImageNet pre-trained counterparts in several image segmentation tasks. The denoising pretraining objective, which has its roots in denoising auto-encoders [16], typically trains deep learning segmentation models to predict the uncorrupted version of a noisy image. Inspired by [15], we will explore the denoising pretraining procedures in the HIE segmentation setting. In particular, we consider the recently proposed *Decoder Denoising Pretraining* (DDP) [17] a state-of-the-art denoising framework for improving segmentation performance.

Our goal in is paper is to demonstrate that incorporating label information during a denoising (self-)pretraining can further enrich learned representations with task-relevant information and improve results. In broad terms, we use the region-of-interest extraction method from the popular augmentation technique CarveMix [18]. In this way, models can focus resources to learn stronger representations only in a few key areas that are the most relevant to the downstream task.

II. Methods

A. Dataset

To evaluate our method, we use the dataset provided for the 1st Boston Neonatal Brain Injury Dataset for Hypoxic Ischemic Encephalopathy (BONBID-HIE) Lesion Segmentation Challenge [19]. The dataset provides 133 expertly annotated annotations of brain lesions in MRI scans of neonates born between 2001 and 2018. The scans are provided in 3D apparent diffusion coefficient (ADC) maps in addition to a newly developed Z_{ADC}, which normalizes the voxels ADC maps relative to the maps of healthy neonates. In this paper, we will use the publicly available training split (85 volumes) converted into 2D images to train and evaluate our models.

B. Pre-processing

Given a skull-stripped ADC map $x_{ss} \in \mathbb{R}^{h\times w}$ and a Z_{ADC} map $x_z \in \mathbb{R}^{h\times w}$ we perform Z-score normalization [20] to normalise the pixel values where the background values are assigned a constant value of -6. The resulting two normalised maps z_{ss} and z_{ss} are concatenated together to create the initial input image $x \in \mathbb{R}^{2\times h\times w}$. Finally, each image is upscaled to 256×256 pixels before being fed to the model.

C. Methodology

Our goal is for the model to develop stronger representations of regions of interest during the pretraining phase. Following the recent derivations for diffusion using non-isotropic Gaussian noise [21], our framework modifies DDP by noising the vector inputs x using

$$x' = \sqrt{\gamma}x + \sqrt{1-\gamma}\sqrt{\mathbf{I}(\sigma_a, \sigma_b|c)}\varepsilon \tag{1}$$

where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{I}(\sigma_a, \sigma_b|c)$ is a diagonal matrix whose diagonal elements are the standard deviations of the independent noise applied to each pixel. The diagonal elements of $\mathbf{I}(\sigma_a, \sigma_b|c)$ are σ_a for pixels within a region of interest c (e.g. labeld areas) and σ_b otherwise. Our image noising scheme equates to adding noise to pixels sampled from $\mathcal{N}(0, \sigma_a)$ for the regions of interest and $\mathcal{N}(0, \sigma_b)$ in the remaining regions.

Consistent with [18], we take the ROI c to be the annotated areas with brain lesions including areas adjacent to lesion contours. Following CarveMix, given an annotation $y \in \mathbb{R}^{w\times h}$ we define an indicator map for the region of interest for an image as

$$\mathscr{C}(y) = \begin{cases} 1 & D(\cdot|y) \leq \lambda \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $D(\cdot|y)$ is the distance between a pixel and the contour of the annotated area. Notably, $D(\cdot|y)$ is negative for annotated pixels within a lesion. CarveMix stochastically samples λ from

$$\lambda \sim \frac{1}{2}U\Big(-\frac{1}{2}\Big|D(\cdot|y)_{min}\Big|, 0\Big) + \frac{1}{2}U\Big(0, \Big|D(\cdot|y)_{min}\Big|\Big) \tag{3}$$

where $D(\cdot|y)_{min} = \min_x D(x|y)$ is an indication of lesion size for a given annotation y.

In practice, we first obtain a random matrix $\mathbf{W}^y_{ij} \in \mathbb{R}^{2\times h\times w}$ before merging it with the image. Where we have

$$\mathbf{W}^y_{kij} \sim \begin{cases} \mathcal{N}(0, \sigma_a^2) & \text{if } \mathscr{C}(y)_{ij} = 1 \\ \mathcal{N}(0, \sigma_b^2) & \text{otherwise.} \end{cases} \tag{4}$$

Using this construction, we can add in the noise pixel-wise so that given an image $\mathbf{x} \in \mathbb{R}^{2\times h\times w}$ we can apply the noising transform

$$\mathbf{x}' = \sqrt{\gamma}\mathbf{x} + \sqrt{1-\gamma}\mathbf{W}^y \tag{5}$$

Consistent with DDP, the model is trained to predict the noise $\mathbf{W}^y$ using the L2 loss. Altogether, given a U-net [22] segmentation model with encoder f_θ and decoder g_ϕ, we self-pretrain the parameters using the loss

$$\mathbb{E}_x\mathbb{E}_{\mathbf{W}^y}\Big\| g_\phi(f_\theta(\sqrt{\gamma}\mathbf{x} + \sqrt{\mathbf{1-\gamma}}\mathbf{W^y})) - \mathbf{W}^y \Big\|_2^2. \tag{6}$$

Similar to DDP we set $\gamma = 0.95$ and based on experiments we fix $\sigma_a = 1.5$ and $\sigma_b = 0.8$.

D. Training & Testing

Our algorithm uses TernausNet [23] a UNet classifier that is self-pretrained using the denoising procedure outlined in Section C.Subsequently, the model is fine-tuned to predict segmentation labels using the following weighted segmentation loss:

$$L_{\text{ft}}(p, y) = L_{\text{BCE}}(p, y) + L_{\text{Dice}}(p, y) + 3L_{\text{Focal}}(p, y) \tag{7}$$

where $L_{\text{BCE}}, L_{\text{Dice}}$, and L_{Focal} are the binary cross-entropy, dice, and Focal loss [24] respectively. Each model is trained until saturation on a validation set, where we only use the dice metric to measure performance.

During the self-pretraining and fine-tuning phases, we apply a simple augmentation where images are randomly flipped horizontally and vertically. Other augmentation strategies typically reduce performance. For both phases, the model is optimized using the Adam with a learning rate of 0.0001 for 30 epochs with a batch size of 16.

To create a prediction on a test image we use a test-time augmentation strategy. First, we aggregate the model outputs created by first applying the various flip transforms seen during training. We then average the outputs to create the final segmentation. The submitted algorithm is a voxel-wise voting ensemble [25] of eight identical models, each self-pretrained using LADP.

E. Evaluation metrics

We evaluate our method using the DICE overlap [26] evaluated with lesions occupying $< 1\%$, $1\% \sim 5\%$, and $> 5\%$ of brain volume. Additionally, we will measure surface distance metrics to evaluate the similarity between the surface contours of the predicted and ground truth segmentations. Namely, the Mean Average Surface Distance (MASD) and the Normalized Surface Dice (NSD) [27]. Similar to the BONBID-HIE challenge's ranking policy, we rank models based on the model's ranking on each metric.

III. Results

A. Cross-Validation

Table 1 outlines the results of our 5-fold cross-validation experiments. Indeed, LADP has superior performance on metrics measured across all validation samples. However, our method is marginally worse than the best methods when evaluated only on brain volumes with lesions occupying $< 1\%$, $1\% \sim 5\%$, or $> 5\%$ of brain volume. Similar to DDP[17], freezing the encoder parameters (θ in Eq. 6) during the pretraining phase provides the best results.

B. Ablation

We conducted an ablation study to assess the individual impact of each of our design choices, where we averaged the results using two experiments using 90% of the public data for training and the remainder for testing. Using TernausNet [23] as the baseline, the results in Table 2 show that our flip augmentation strategy during training yields the most substantial relative improvement. In contrast, LADP confers a slight advantage (+0.6 dice points) when tracing performance across all validation samples, however, the improvement is twice as large (1.4 dice points) for brain volumes with lesions occupying $< 1\%$ of total volume. This suggests that the denoising objective is more effective for learning the segmentation of smaller structures.

C. Overall-Performance

Based on mean rank, our method placed 2nd on the hidden test set for the BONBID-HIE lesion segmentation challenge. Similar to our cross-validation results, our method yields the best dice results when measured across all test samples but produces slightly worse but competitive results when evaluated on volumes with lesioned areas occupying only a small percentage.

IV. CONCLUSION

We have proposed LADP, a denoising pretraining framework for HIE lesion segmentation with the goal of training models with a strong representation in a few key areas relevant to the downstream task. At its core, LADP uses the segmentation labels and the region-of-interest extraction method from CarveMix in a denoising self-pretraining framework. In our experiments, LADA has superior results relative to state-of-the-art techniques when averaging across all test samples. Additionally, we suggested that the improvement is largely explained by the relatively improved performance on volumes with lesioned areas occupying less than one percent. We validated our methods using the 1st Boston Neonatal Brain Injury Dataset for Hypoxic Ischemic Encephalopathy Lesion Segmentation Challenge in which our final algorithim placed 2nd overall.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. S. A. Zanelli, "Hypoxic-ischemic encephalopathy," https://emedicine.medscape.com/article/973501-overview#a2, 2018, accessed: 2023-12-04.
2. E. M. Graham, K. A. Ruis, A. L. Hartman, F. J. Northington, and H. E. Fox, "A systematic review of the role of intrapartum hypoxia-ischemia in the causation of neonatal encephalopathy." *American journal of obstetrics and gynecology*, vol. 199 6, pp. 587–95, 2008.
3. J. J. Kurinczuk, M. White-Koning, and N. Badawi, "Epidemiology of neonatal encephalopathy and hypoxic-ischaemic encephalopathy." *Early human development*, vol. 86 6, pp. 329–38, 2010, pMID: 20554402.
4. A. E. Törn, S. Hesselman, K. Johansen, J. Ågren, A.-K. Wikström, and M. Jonsson, "Outcomes in children after mild neonatal hypoxic ischaemic encephalopathy: A population-based cohort study," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 130, pp. 1602 – 1609, 2023, pMCID: 37199188.
5. B. M. Michniewicz, S. R. A. Saad, L. M. Karbowski, J. Gadzinowski, M. Szymankiewicz, and D. Szpecht, "Organ complications of infants with hypoxic ischemic encephalopathy before therapeutic hypothermia." *Therapeutic hypothermia and temperature management*, 2020, pMCID: 33155883.
6. M. A. Pakvasa, A. M. Winkler, S. E. G. Hamrick, C. D. Josephson, and R. M. Patel, "Observational study of haemostatic dysfunction and bleeding in neonates with hypoxic–ischaemic encephalopathy," *BMJ Open*, vol. 7, 2017.
7. C. E. J. Parmentier, L. S. de Vries, and F. Groenendaal, "Magnetic resonance imaging in (near-)term infants with hypoxic-ischemic encephalopathy," *Diagnostics*, 2022.
8. A. Cabaj, M. Bekiesińska-Figatowska, and J. Madzik, "Mri patterns of hypoxic-ischemic brain injury in preterm and full term infants – classical and less common mr findings," *Polish Journal of Radiology*, vol. 77, pp. 71 – 76, 2012, pMCID: PMC3447438.
9. J. L. Wisnowski, P. Wintermark, S. L. Bonifacio, C. D. Smyser, A. J. Barkovich, A. D. Edwards, L. S. de Vries, T. E. Inder, and V. Chau, "Neuroimaging in the term newborn with neonatal encephalopathy."

	Dice(↑)			Overall			
Methodology	$< 1\%$	$1\% - 5\%$	$> 5\%$	MASD(↓)	NSD(↑)	Dice(↑)	Avg. Rank
CarveMix[28]	**51.92**	69.51	65.72	2.8364	76.60	59.19	3.17
DAE [16]	51.24	70.50	**65.92**	2.7483	75.46	59.08	3.17
DDP[17]	48.11	**71.13**	64.80	3.0238	75.51	57.25	4.83
DDP[17]	50.65	70.72	65.44	2.7617	76.72	58.71	3.33
LADP (ours)	48.17	69.83	65.88	2.9015	76.14	57.23	4
LADP (ours)	51.64	70.71	65.46	**2.4886**	**78.10**	**59.34**	**2**

Table 1: **Comparison with State-of-the-art**: The best-performing method is **bolded** and the second best-performing is underlined. Methods where only the decoder is pretraind are highlighted in grey.

Method	$< 1\%$	Δ	Overall	Δ
Baseline	42.5	-	49.5	-
+Augment	45.8	3.3	51.2	1.7
+Test-time Aug	46.4	0.6	51.7	0.5
+LADP	47.8	1.4	52.3	0.6

Table 2: **Ablation Results**: Dice overlap and relative improvement (Δ) after adding each design choice. Results are given for brain volumes with lesions only occupying $< 1\%$ of brain volume in addition to the dice measured across all validation samples.

Seminars in fetal & neonatal medicine, p. 101304, 2021, pMCID: PMC9135955.

10. E. S. Biratu, F. Schwenker, Y. Megersa, and T. G. Debelee, "A survey of brain tumor segmentation and classification algorithms," *Journal of Imaging*, vol. 7, 2021, pMCID: PMC8465364.
11. D. Karimi, S. Warfield, and A. Gholipour, "Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations," *Artificial intelligence in medicine*, vol. 116, p. 102078, 2021, pMCID: PMC8164174.
12. K. Krishna, S. Garg, J. P. Bigham, and Z. C. Lipton, "Downstream datasets make surprisingly good pretraining corpora," *ArXiv*, vol. abs/2209.14389, 2022.
13. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
14. P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. S. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang, "Wilds: A benchmark of in-the-wild distribution shifts," 2021.
15. A. El-Nouby, G. Izacard, H. Touvron, I. Laptev, H. Jegou, and E. Grave, "Are large-scale datasets necessary for self-supervised pre-training?" 2021.
16. P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *International Conference on Machine Learning*, 2008.
17. E. A. Brempong, S. Kornblith, T. Chen, N. Parmar, M. Minderer, and M. Norouzi, "Denoising pretraining for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4175–4186.
18. X. Zhang, C. Liu, N. Ou, X. Zeng, Z. Zhuo, Y. Duan, X. Xiong, Y. Yu, Z. Liu, Y. Liu, and C. Ye, "Carvemix: A simple data augmentation method for brain lesion segmentation," *NeuroImage*, vol. 271, p. 120041, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811923001878
19. R. Bao, Y. Song, S. V. Bates, R. J. Weiss, A. N. Foster, C. J. Cobos, S. Sotardi, Y. Zhang, R. L. Gollub, P. E. Grant, and Y. Ou, "Boston neonatal brain injury dataset for hypoxic ischemic encephalopathy (bonbid-hie): Part i. mri and manual lesion annotation," *bioRxiv*, 2023. [Online]. Available: https://www.biorxiv.org/content/early/2023/07/03/2023.06.30.546841
20. B. M. Ellingson, T. M. Zaw, T. F. Cloughesy, K. M. Naeini, S. Lalezari, S. Mong, A. Lai, P. L. Nghiemphu, and W. B. Pope, "Comparison between intensity normalization techniques for dynamic susceptibility contrast (dsc)-mri estimates of cerebral blood volume (cbv) in human gliomas," *Journal of Magnetic Resonance Imaging*, vol. 35, 2012.
21. V. Voleti, C. Pal, and A. Oberman, "Score-based denoising diffusion with non-isotropic gaussian noise models," 2022.
22. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.
23. V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," 2018.
24. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
25. L. Lam and C. Y. Suen, "Application of majority voting to pattern recognition: an analysis of its behavior and performance," *IEEE Trans. Syst. Man Cybern. Part A*, vol. 27, pp. 553–568, 1997.
26. C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248.
27. S. Nikolov, S. Blackwell, A. E. Zverovitch, R. Mendes, M. Livne, J. D. Fauw, Y. Patel, C. Meyer, H. Askham, B. Romera-Paredes, C. J. Kelly, A. Karthikesalingam, C. Chu, D. Carnell, C. Boon, D. D'Souza, S. A. Moinuddin, B. Garie, Y. McQuinlan, S. Ireland, K. Hampton, K. Fuller, H. Montgomery, G. E. Rees, M. Suleyman, T. Back, C. O. Hughes, J. R. Ledsam, and O. Ronneberger, "Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study," *Journal of Medical Internet Research*, vol. 23, 2020, pMCID: PMC8314151.
28. X. Zhang, C. Liu, N. J. Ou, X. Zeng, X. Xiong, Y. Yu, Z. Liu, and C. Ye, "Carvemix: A simple data augmentation method for brain lesion segmentation," *NeuroImage*, vol. 271, 2021, 36933626. [Online]. Available: 10.1016/j.neuroimage.2023.120041