Attention Shift: Interpretability Study of Texture-based Data Augmentation in Training U-Net Models for Brain Image Segmentation

Suhang You
ARTORG, Department of Medicine
University of Bern
suhang.you@artorg.unibe.ch

Mauricio Reyes
ARTORG, Department of Medicine
University of Bern
mauricio.reyes@med.unibe.ch

Abstract

Texture smoothing has recently become a promising data augmentation method to enhance the performance of deep learning segmentation methods in medical image analysis. However, a deeper understanding of this phenomenon has not been investigated. In this study, we investigated this phenomenon using a controlled experimental setting, using datasets from the Human Connectome Project, in order to mitigate the inhomogeneity of data confounders to the network, and investigate possible explanations as to why model performance changes when applying different levels of total variation smoothing during data augmentation. Through experiments we confirm previous findings regarding the benefits of smoothing during data augmentation, but further report that the regime of improvement is limited and it changes in relation to the selected imaging protocol. We also found that smoothing during data augmentation produces a spatial attention shift also associated with different performance levels of the trained segmentation model.

1 Introduction

Data augmentation has been introduced in medical image analysis to improve the performance of many applications such as brain image segmentation. From simple strategies, such as random rotation, cropping [9, 7], and contrast modification [1, 13] to generative models for pseudo-data generation [8, 10, 2], these studies have reported improved segmentation results. Meanwhile, complex strategies to optimize data augmentation have also been proposed in recent years. As discussed in [4], Convolutional Neural Networks (CNN) are biased towards texture, which can be used to improve network performance via targeted data selection and augmentation. In recent works, the authors of [18] and [14] reported on the benefit of using a smoothing-based data augmentation approach in biomedical image segmentation tasks. These works argue that high-frequency features are not important cues compared to semantic boundaries, and trained models are biased towards high-frequency features instead of semantic boundaries. Through smoothing operations, the bias is reduced and the segmentation performance can be improved. The work of [18] reported improved segmentation results on biomedical images through superpixel-based smoothing as data augmentation and the authors in [14] reported improved segmentation results using a Total Variation smoothing-based data augmentation for white matter segmentation and lesion segmentation tasks. Despite the benefit of using a smoothing-based data augmentation, further understanding as to how data augmentation affects a segmentation network, such as the U-Net [11] has not been investigated in more detail. The work of [14] did not explore different regimes of smoothing to report whether the benefits of smoothing disappear or could even worsen model performance, and the work of [18] randomly samples smoothing levels based on prior knowledge enforcing appropriate smoothed images generated. In order to investigate the effect of smoothing-based data augmentation on model

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

performance, we designed an experimental setup under controlled conditions constructing a large dataset of 21'000 synthetically generated brain MRI datasets, stemming from 500 real brain images from the Human Connectome project, on 42 different simulated imaging protocols. Through this controlled experimental setup we aimed at mitigating the effect of potential confounder effects such as the uncontrolled heterogeneity of vendor protocols, present on publicly available multi-center datasets, and other effects such as different anatomical information on imaging protocols, etc. Next to analyzing the regime of improvement attained by smoothing-based data augmentation, we analyzed how spatial attention of these segmentation models changed, using saliency maps [15] to investigate the relation between pixel attention and segmentation performance.

In summary, our contribution are: (1) We analyzed different regimes of smoothing-based data augmentation for segmentation models using a dedicated and controlled imaging dataset setup. (2) We investigated the spatial attention of segmentation models trained under different smoothing-based data augmentation levels.

2 Materials and methods

Datasets construction Our datasets are constructed utilizing BrainWeb: Simulated Brain Database [3] and datasets from the Human Connectome Project (HCP) datasets [17]. Datasets generated from BrainWeb provides vendor variability in a controlled manner where the same brain in different repetition time (TR) and echo time (TE) values can be manually simulated for a given scanning sequence. In our study, we used spin-echo sequence for T1 with 42 different combinations of TR and TE values characterizing commonly used ranges (TR $\in [300, 800]ms$ and TE $\in [10, 40]ms$), simulating 42 different vendor protocols for the same human brain (i.e. pseudo vendors). HCP is a worldwide initiative providing MRI imaging datasets characterizing brain anatomy. We randomly selected 500 brains and used accompanying transformation files to normalize each brain to MNI space. This step was intended to normalize background-to-foreground information which can confound trained models and allows the following registration. In order to simulate that each of the 500 cases underwent each of 42 pseudo vendors, we non-rigidly registered [6] every pseudo-vendor to every case, leading to a total of 21'000 simulated brain images. For model training, we adopted a 4:1 split for training, validation and test sets,resulting in 320, 80, 100 brains, accordingly. We constructed each of the three sets by randomly and homogeneously selecting brains from the 42 different pseudo-vendors, and 5 slices per brain.

Total variation smoothing as data augmentation Total Variation (TV) smoothing is based on a denoising method from [12]. Given an image $f \in \mathbb{R}^n$, the smoothed image $u \in \Omega \subset \mathbb{R}^n$ is found when minimizing

$$\underset{u \in BV(\Omega)}{\arg\min} \|u\|_{TV(\Omega)} + \alpha \int_{\Omega} (f(x) - u(x))^2 dx, \tag{1}$$

where $BV(\Omega)$ is the bounded variations of domain Ω and $\|\cdot\|_{TV(\Omega)}$ denotes the TV norm of u. TV smoothing parameter $\alpha \in \mathbb{R}^+$ is a weight parameter. In our experiments, we used split-Bregman based implementation [5] in which the smaller α is the stronger smoothing applied to the image. We selected multiple values in order to enforce proper- and over-smoothing cases.

Saliency maps for segmentation For a binary classification CNN $F : \mathbb{R}^n \to [0, 1]$, the saliency map for a label of input image $x \in \mathbb{R}^n$ in the Integrated Gradient (IG) method [16] is defined as

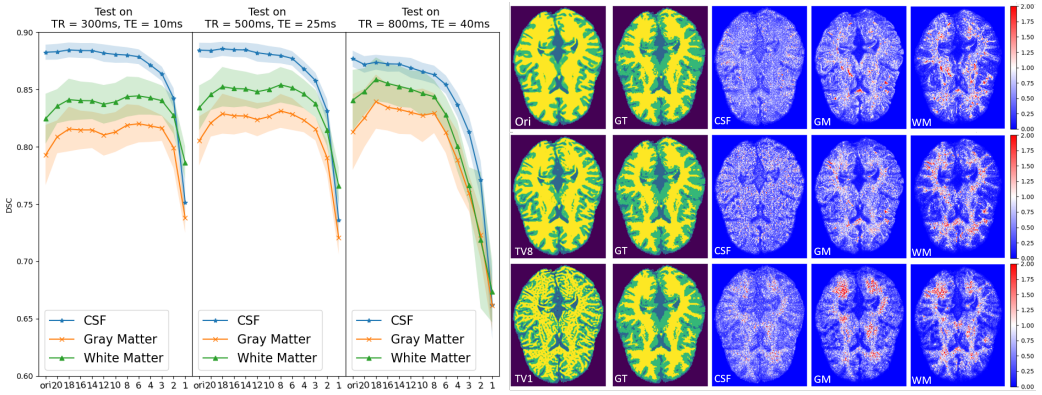
$$IG_i(x) ::= (x_i - x_i{}') \times \int_{\beta=0}^{1} \frac{\partial F(x' - \beta(x - x'))}{\partial x_i} d\beta \tag{2}$$

where x' is the baseline image and $IG_i(x)$ is the integrated gradients in dimension i. Based on this we constructed saliency maps for segmentation results by calculating a patch-wise IG instead of against a binary class $[0, 1]$ (In our task, we segmented three classes: white matter, gray matter and CSF). To achieve that, we slid a window in the soft-max output at the size of receptive field to one pixel in the bottleneck of the U-Net with a stride size of 8 (three poolings in U-Net). For one input slice, we accumulated 18×18 saliency maps against these patches to calculate our saliency maps.

Experiment settings We used the U-Net architecture to segment cerebral fluid (CSF), gray matter (GM) and white matter (WM). Models were trained with 250 epochs and the epoch with the lowest

validation loss was selected for testing. To reduce stochasticity during training, all training used the same initialization seed. Training and testing was repeated and averaged 20 times with different random sample selections. We used dice similarity coefficient (DSC) to assess segmentation performance at different levels of smoothing. We used GeForce GTX 1080Ti GPU during experiments.

3 Results and conclusion



(a) (b)

Figure 1: 1a) Test DSC vs. α. x-axis in each plot from left to right the applied α decrease. Smaller α with stronger smoothing applied. Test on Vendors of: left, TR = 300ms, TE= 10ms; middle, TR = 500ms, TE = 25ms; right, TR = 800ms, TE = 40ms. 1b) One slice example from the vendor: TR = 500ms, TE = 25ms. Columns from left to right: segmentation, ground truth, saliency maps of CSF, GM and WM. Rows from top to bottom: results from original image, smoothed images with α = 8 and 1. Higher values of the saliency maps correspond to higher pixel attribution.

Figure 1a shows the mean and standard deviation of DSCs in three tested vendors. For all test vendors, the segmentation performance for CSF generally degraded when applying TV smoothing, with stronger smoothing further degrading is. For GM and WM, smoothing benefits the segmentation in a given range ($\alpha \in$ [8,20]) then it deteriorates as the smoothing increases. We noticed that the effect of TV smoothing to different vendors varies, which suggests that in practice a more limited range of smoothing could be applied to multi-vendor test sets. Across all tested pseudo-vendors we noticed similar findings but with specific regimes of benefit, suggesting that in practice one might need to verify an optimal level of smoothing depending on vendor protocol.

To elucidate the spatial attention of trained models at different levels of smoothing, we calculated and compared saliency maps for each pseudo-vendor. One example slice is shown in Figure 1b. At α =8, GM and WM reach the best mean DSC. Conversely, at α =1, all three class segments (GM,WM,CSF) are worse than the original DSC. We observed a tendency of segmentation results from over-segmentation of the WM, for models trained with original images (i.e. no smoothing), to under-segmentation of WM for models trained with α =1. In terms of spatial attention, we observed an overall attention shift, where WM pixels are more attended as smoothing increases, while the attention to the other two tissues decreases. This phenomenon explains the increase DSC change for WM and GM where they shift from over-segmentation to under-segmentation, and from under-segmentation to over-segmentation, respectively.

Our work extends previous works as it shows that smoothing improves the performance of segmentation in a limited range and the selection of smoothing parameter is important. As smoothing increases, we show that U-Net models shift their attention to white matter, which is associated with a change from over- to under-segmentation of white matter. Interestingly, these findings suggest that U-Net models employ multi-class information to segment each class (i.e. multi-class attention), which in exacerbated sub-optimal scenarios, such as an over-smoothing, leads to the U-Net model to an over-attention. These observations relate to findings in shortcut learning [4] in classification models where sub-optimal data setups (e.g. data bias) can lead models to shift their attention to exploit spurious correlations in the data. As follow-up work, we intend to characterize these factors quantitatively in order to provide deeper understanding these first observations.

Broader impact

Beyond performance-related objectives, we believe it is important to elucidate how and why current data augmentation strategies affect the training and performance of segmentation models. We hope that our work can promote new evidence to help us better understand how deep learning network react to changes to the data, and hopefully stimulate the medical imaging deep learning community towards evidence-based studies improving the interpretability of deep learning models.

Acknowledgments and Disclosure of Funding

The project was supported by the Swiss Personalized Health Network (SPHN) initiative.

References

[1] Monika Agarwal and Rashima Mahajan. Medical images contrast enhancement using quad weighted histogram equalization with adaptive gama correction and homomorphic filtering. *Procedia computer science*, 115:509–517, 2017.

[2] Varghese Alex, Mohammed Safwan KP, Sai Saketh Chennamsetty, and Ganapathy Krishnamurthi. Generative adversarial networks for brain lesion detection. In *Medical Imaging 2017: Image Processing*, volume 10133, page 101330G. International Society for Optics and Photonics, 2017.

[3] Chris A. Cocosco, Vasken Kollokian, Remi K.-S. Kwan, G. Bruce Pike, and Alan C. Evans. Brainweb: Online interface to a 3d mri simulated brain database. *NeuroImage*, 5:425, 1997.

[4] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2019.

[5] Pascal Getreuer. Rudin-Osher-Fatemi Total Variation Denoising using Split Bregman. *Image Processing On Line*, 2:74–95, 2012. `https://doi.org/10.5201/ipol.2012.g-tvd`.

[6] Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. Fsl. *NeuroImage*, 62(2):782–790, 2012. 20 YEARS OF fMRI.

[7] Yan Liu, Strahinja Stojadinovic, Brian Hrycushko, Zabi Wardak, Steven Lau, Weiguo Lu, Yulong Yan, Steve B Jiang, Xin Zhen, Robert Timmerman, et al. A deep convolutional neural network-based automatic delineation strategy for multiple brain metastases stereotactic radiosurgery. *PloS one*, 12(10):e0185844, 2017.

[8] Tony CW Mok and Albert CS Chung. Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks. In *International MICCAI Brainlesion Workshop*, pages 70–80. Springer, 2018.

[9] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A. Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE Transactions on Medical Imaging*, 35(5):1240–1251, 2016.

[10] Mina Rezaei, Konstantin Harmuth, Willi Gierke, Thomas Kellermeier, Martin Fischer, Haojin Yang, and Christoph Meinel. A conditional adversarial network for semantic segmentation of brain tumor. In *International MICCAI Brainlesion Workshop*, pages 241–252. Springer, 2017.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[12] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[13] Mouna Sahnoun, Fathi Kallel, Mariem Dammak, Chokri Mhiri, Kheireddine Ben Mahfoudh, and Ahmed Ben Hamida. A comparative study of mri contrast enhancement techniques based on traditional gamma correction and adaptive gamma correction: Case of multiple sclerosis pathology. In *2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–7. IEEE, 2018.

[14] Rasha Sheikh and Thomas Schultz. Feature preserving smoothing provides simple and effective data augmentation for medical image segmentation. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 116–126, Cham, 2020. Springer International Publishing.

[15] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[16] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[17] D.C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T.E.J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S.W. Curtiss, S. Della Penna, D. Feinberg, M.F. Glasser, N. Harel, A.C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S.E. Petersen, F. Prior, B.L. Schlaggar, S.M. Smith, A.Z. Snyder, J. Xu, and E. Yacoub. The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4):2222–2231, 2012. Connectivity.

[18] Yizhe Zhang, Lin Yang, Hao Zheng, Peixian Liang, Colleen Mangold, Raquel G. Loreto, David P. Hughes, and Danny Z. Chen. Spda: Superpixel-based data augmentation for biomedical image segmentation. In M. Jorge Cardoso, Aasa Feragen, Ben Glocker, Ender Konukoglu, Ipek Oguz, Gozde Unal, and Tom Vercauteren, editors, *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 572–587. PMLR, 08–10 Jul 2019.