SMOCAM: SMOOTH CONDITIONAL ATTENTION MASK FOR 3D-REGRESSION MODELS

Salamata Konate[†,‡], Léo Lebrat[⋆,‡], Rodrigo Santa Cruz[⋆,‡], Pierrick Bourgeat[⋆], Vincent Doré[⋆,], Jurgen Fripp[⋆], Andrew Bradley[‡], Clinton Fookes[‡], and Olivier Salvado[†,‡]*

[⋆] CSIRO Health and Biosecurity, The Australian eHealth Research Centre, [†] CSIRO Data61, [‡]Image and Video Laboratory QUT, [*] Department of Nuclear Medicine and Centre for PET, Australia.

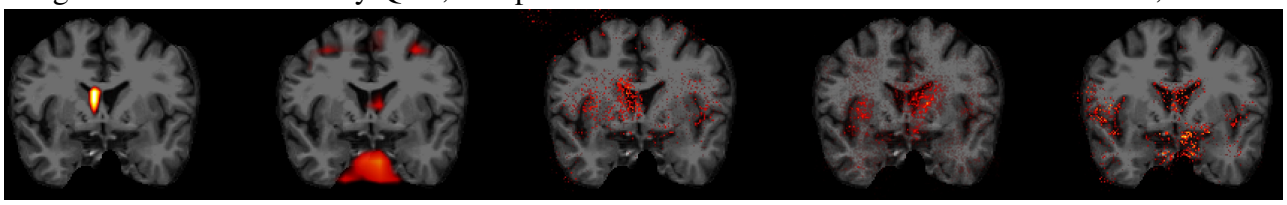

Fig. 1: Saliency maps for the volume prediction of the right lateral ventricle. Saliency methods from left to right: SMOCAM (ours), Grad-CAM, DeconvNet, gradient-based saliency, and Guided backprop.

ABSTRACT

Despite the pervasive growth of deep neural networks in medical image analysis, methods to monitor and assess network outputs, such as segmentation or regression, remain limited. In this paper, we introduce SMOCAM (**SMO**oth **C**onditional **A**ttention **M**ask), an optimization method that reveals the specific regions of the input image taken into account by the prediction of a trained neural network. We developed SMOCAM explicitly to perform saliency analysis for complex regression tasks in 3D medical imagery. Our formulation optimises an **3D**-attention mask at a given layer of a convolutional neural network (CNN). Unlike previous attempts, our method is relatively fast (40s per output) and is suitable for large data such as 3D MRI. We applied SMOCAM on a CNN that predicts Brain morphometry from 3D MRI which was trained using more than 5000 3D brain MRIs. We show that SMOCAM highlights neural network's limitations when cases are underrepresented and in cases with large volume asymmetry.

***Index Terms*—** Deep Learning, Explainability of regression models, Saliency maps.

1. INTRODUCTION

Neural networks used for machine learning have shown very high accuracy but low explainability, which have hindered their translation to medical applications. Drawing from recent progress in computer vision, one can use spatial sensitivity analysis to produce saliency maps, showing the locations in the image that "explain" network outputs. For instance, gradient-based saliency methods like Simonyan et al. [1], DeconvNet [2], Grad-CAM [3], and Guided-backprop [4] use gradient information and re-normalization to compute saliency maps highlighting the image regions that influence the prediction the most. While these methods are lightweight, they tend to produce scattered saliency maps that are difficult to interpret as shown in Figure 1.

Recently, optimization-based saliency methods have been proposed [5–7]. They optimize the localization and shape of a mask to obtain a more informative explanation map. For example, in the context of medical image analysis, the work of Fong and Vedaldi [5] has been successfully applied to the automatic diagnosis of Alzheimer's disease (AD) [8]. However, that method remains limited to tissue probability maps and simple binary classification problems like AD patients vs. Healthy Control.

In this paper, we propose a novel approach to perform saliency analysis when a CNN has been trained to regress morphometric measurements from non-registered 3D MRI, such as cortical thickness, gray matter volume, or cortical curvature. A recent optimization method by Fong et al. [6] yields a computationally intensive solution that is challenged by large 3D volumetric data. Rather, we improved on the work of Taha et al. [7]: we optimize an attention mask within the neural network, more specifically, on a feature map produced by a convolutional layer. The resulting conditional attention mask is then processed back to the original image resolution for saliency analysis. We name our approach SMOCAM (**SMO**oth **C**onditional **A**ttention **M**ask).

In the following, we describe the CNN model used for brain morphometry regression from 3D brain MRI. We then present our proposed SMOCAM method, highlighting some specific necessary adjustments for obtaining a fast and reliable optimization. Finally, we use SMOCAM's saliency masks to analyze the predicted volume of the ventricle, putamen, and hippocampi.

Thanks to Maxwell plus `http://maxwellplus.com/`.

2. BRAIN MORPHOMETRY USING CNN

Morphometric measurements of the brain's anatomical structures are important non-invasive biomarkers associated with neurodegenerative disorders [9, 10]. Traditionally, image processing pipelines such as the popular FreeSufer [11] can take several hours per MRI to produce cortical surfaces that are then processed to predict clinically meaningful metrics such as grey matter volume, cortical thickness, or brain surface curvature. Recently, processing time has been reduced to less than one hour by novel CNN based segmentation model [12] or devising novel formulation for cortical surface estimation [13]. Another promising approach is to directly estimate those metrics from the 3D MRI without a need for segmentation or cortical surface estimation. We are interested in those later methods and use in this paper the model proposed by Rebsamen et al. [14]. One limitation of regressing imaging biomarkers directly from a scan is the lack of explainability. It is unclear when a network produces an estimate of what imaging information has been used. For example, common sense dictates that the left hippocampus area should be used to estimate the left hippocampus volume, but we show here that it is not always the case as demonstrated by SMOCAM.

2.1. CNN for brain morphometry estimation

The regression model that we have used in this paper is based on the VGG framework [14]. It consists of three 3D convolutional layers with ReLU activation function and max-pooling followed by 3 fully connected layers. The neural network is then trained to minimize the mean squared error on the training set using batches of 6 MRI-images and the ADAM optimizer. The final model is selected for the best intra-class correlation coefficient on the validation set using early-stopping. This model produces 165 morphometric measurements from T1 weighted 3D-MRI. We refer the reader to the original publication [14] for more details.

2.2. Data

The data comprises T1w MR scans from two public datasets: the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [15] and the Australian Imaging, Biomarkers and Lifestyle (AIBL) study [16]. The training set comprises $5,632$ MR-images of dimension $164\times172\times202$. The ground truth bio-markers are estimated using FreeSurfer: 29 subcortical structure volumes and 68 mean thickness and mean curvatures. All are using the parcellations obtained from Freesurfer V6.0 cross-sectional pipeline (more details on the dataset can be found in [17]).

3. SMOCAM

Given a 3D input image I and a pretrained deep neural network regression model denoted as DNN : $I \rightarrow \mathcal{R}$, SMOCAM computes an attention mask $\mathcal{M}$ for a given layer, that is then interpolated to obtain a saliency map at the same dimensions of the input image such that it assigns higher values to voxels that are more relevant to the predictions. The choice of the layer depends on the structure of interest. Optimizing a mask for the last layer would encompass a large part of the whole brain and would not be specific to any brain structure in particular. We thus compute the attention mask at a feature map with dimensions $L_d \times H_d \times W_d \times D_d$ and then upsample it to the input image dimensions to perform saliency analysis. As described in Figure 2, the feature map F_d at a layer d is element-wise multiplied $\odot$ with a mask of same dimensions $\mathcal{M}_d$. The masked feature map $F_d \odot \mathcal{M}_d$ is passed through the remaining layers of the network to obtain the modified predictions $\text{DNN}_{\mathcal{M}_d}$. In the case of the brain morphometry regression model described in Section 2.1, we compute a saliency mask for each output measurement independently.

Therefore, given a fixed L^2 budget $\|\mathcal{M}_d\|_2 = 1$, our goal is to find a saliency mask $\mathcal{M}_d$ such that the prediction after masking the feature map $\text{DNN}_{\mathcal{M}_d}$ is as close as possible to its original value DNN. Formally,

$$\underset{\substack{\mathcal{M}_d \\ \text{s.t. } \|\mathcal{M}_d\|_2=1}}{\arg\min} \quad \mathbf{d}(\text{DNN}_{\mathcal{M}_d}, \text{DNN}), \tag{1}$$

where $\mathbf{d}$ is a normalized distance function. We solve this problem using its Tikhonov form by adding a regularization term $\mathcal{R}_\text{s}$ that promotes spatial smoothness of the mask $\mathcal{M}_d$,

$$\underset{\mathcal{M}_d}{\arg\min}\, \mathbf{d}(\text{DNN}_{\mathcal{M}_d}, \text{DNN}) + \lambda \mathcal{R}_{\|\bullet\|_2}(\mathcal{M}_d) + \gamma \mathcal{R}_\text{s}(\mathcal{M}_d), \tag{2}$$

where,

$$\mathbf{d}(a,b) = \sigma^{-1}\|a-b\|_1, \quad \mathcal{R}_{\|\bullet\|_2}(\mathcal{M}_d) = |1 - \|\mathcal{M}_d\|_2|,$$
$$\mathcal{R}_\text{s}(\mathcal{M}_d) = \frac{1}{3}(\|\nabla_x \mathcal{M}_d\|_1 + \|\nabla_y \mathcal{M}_d\|_1 + \|\nabla_z \mathcal{M}_d\|_1),$$

with σ the standard deviation of the output measurement. The resulting optimization problem is non-convex and has two hyper-parameters γ and λ that balances smoothness with the L^2-norm. We now present a few optimization tricks that help convergence towards satisfactory minima:

Step-size selection: We optimize the cost function of Equation (2), here denoted by $f(\mathcal{M}_d)$, using a gradient method. We found the selection of the gradient step-size τ to be important: too large we noticed oscillation, while too small we observed stagnation of the solution. To avoid arbitrary heuristics in selecting the step-size or a decay rate, we used an adaptive step-size with Armijo condition [18]: $f(\mathcal{M}_d) - f(\mathcal{M}_d - \tau\nabla\mathcal{M}_d) > 0$. Usually, this test is performed within a backtracking line search that can be time-consuming. To avoid this extra computation, we test the condition once; if it is not met, we decrease the value of τ with the update $\tau = 0.99\tau$.

Stopping criterion: Similarly to [7], our stopping criterion is the stagnation of cost function: $|F(\mathcal{M}_d^{i-50}) - F(\mathcal{M}_d^i)| \leq 10^{-5}$

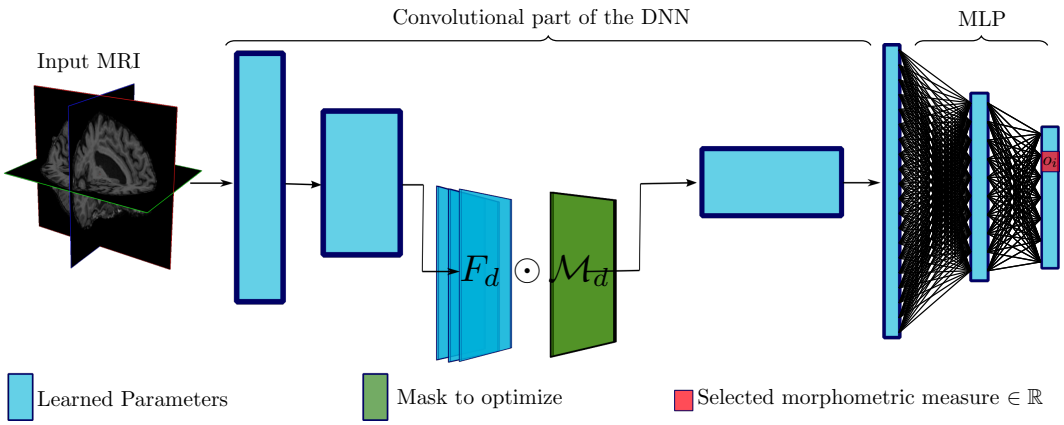

Fig. 2: Given a feature map F_d at a depth d of dimension $L_d \times H_d \times W_d \times D_d$, we piece-wise multiply each layer $(L_i)_{i=1..n}$ with the mask $\mathcal{M}_d$ of dimension $1 \times H_d \times W_d \times D_d$. The resulting feature map is then plugged back to the next convolutional layer. We also select a particular output o_i (red square) which, in our context is a single morphometric measurement.

Mask size[1]	55^3	27^3	14^3	6^3
Time for 1k iterations	59 s	56.6 s	42 s	38.2 s

Table 1: Run-time analysis for different mask size.

where $\mathcal{M}_d^i$ is the optimized saliency mask at the i iteration. We also fix the maximal number of iterations to $2,000$.

Initialization: Random initializing the saliency mask leads to different solutions since our problem is not convex and has a myriad of local minima. For this reason, we set the initial mask to be uniform $\mathcal{M}_d^0 = 1.1\frac{\mathbb{1}}{\|\mathbb{1}\|_2}$.

Obtaining the final saliency mask: The saliency obtained after optimization is usually not sparse with many very small coefficients ($\approx 10^{-4}$). One approach is to promote sparsity with a penalty term of the type $\|\mathcal{M}_d\|_1$, but this method involves a third hyper-parameter to select. Instead, we opt for hard thresholding the values below the 95th percentile.

Implementation details: SMOCAM is implemented using Pytorch and the `autograd` package. The runtime of such method depends on the location of the required mask and on the number of iteration required by the gradient optimization to converge. In our experiment, the mask's size is $14\times14\times17$ and it takes on average 950 iterations to achieve convergence resulting in 40 seconds on average per scan. In Table 1, we provide a run-time analysis for different layers, see [14] for the neural network architecture details.

4. ANALYZING BRAIN MORPHOMETRIC MEASUREMENTS WITH SMOCAM

In this section, we present a quantitative analysis of morphometric measurements using SMOCAM, the brain morphometry regression model presented in Section 2.1, and 200 MRI scans randomly chosen from the test split of the dataset described in Section 2.2. Using this dataset and model, we run SMOCAM with a regularization coefficient of $\lambda = 1$, mask smoothness coefficient of $\gamma = 30$ (λ and γ are problems dependent and determined using a grid-search), and learning-rate of $\tau = 10^{-2}$ to obtain saliency masks for the prediction of a chosen morphometric measurement. Then, we intersect these generated masks with the ground-truth brain parcellations to obtain the model's attention at a given brain region.

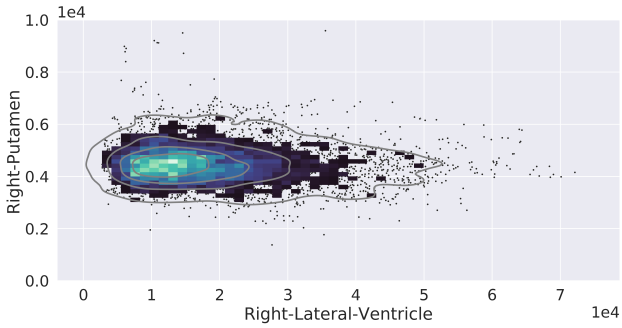

Fig. 3: Bivariate plot between the right lateral ventricle (x axis) and the right putamen (y axis) volumes within the training dataset. The dispersion of the distribution of these measurements in terms of coefficient of variation is 55% for the right lateral ventricle and 15% for the right putamen.

More specifically, we first focus on the prediction of the volumes for the right lateral ventricle and right putamen. The evaluated brain morphometry regression model presents an excellent performance ($ICC = 0.95$) for the former and a lower performance ($ICC < 0.65$) for the latter. We hypothesized that the performance of the network is correlated with the variability of the measurement in the training data, which

[1]To shorten the notation we used the dimensions of the sagittal plane, the mask size is larger in the coronal plane.

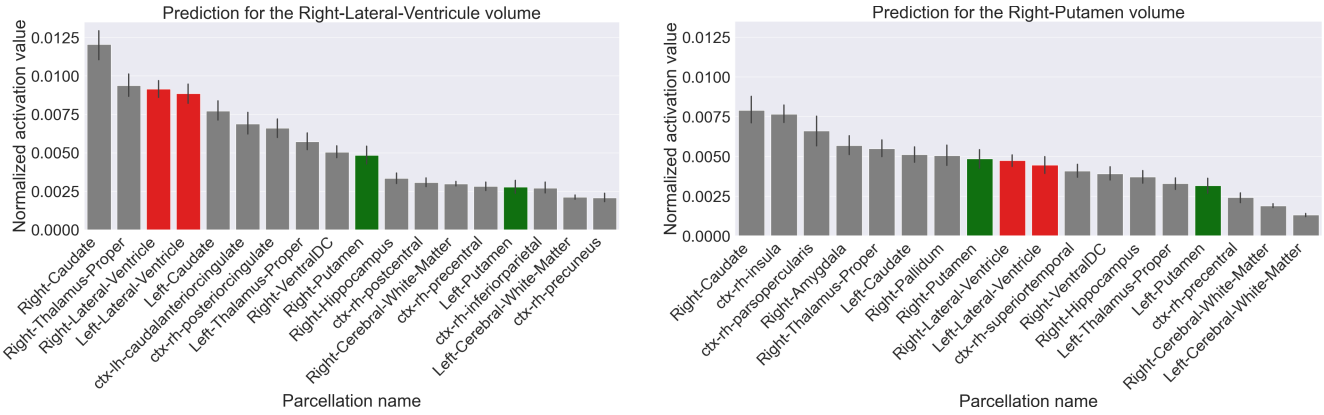

Fig. 4: Average attention per anatomical structure (Red for the ventricles, green for the putamina and grey for the other structures) normalized by their area of the right lateral ventricle (left) and the right putamen (right).

is lower for the putamen than for the ventricle as shown in Figure 3. Indeed, Figure 4 shows the average attention by anatomical structure normalized by their area. SMOCAM's saliency masks focus on the brain regions related to the predicted measurements and their neighborhood. The attention is also correlated to the variability of the region of interest, the more its value fluctuates within the training dataset, the more the region contains information which is relevant for the neural network to build its prediction. The prediction of the ventricle is therefore more accurate than the right putamen.

We also noticed that for predicting the volume of the right lateral ventricle, the attention is equivalently spread across the right and the left structures, with no significant difference between the two sides (p-value = 0.3). This does not prevent the network from producing accurate predictions for each side since both sides of the ventricle volume are highly correlated (Pearson correlation coefficient of 0.914 and ICC of 0.893). However, it is clearly a flawed estimate of individual volumes and should not be used as such for further analysis to investigate disease progression or etiology.

Finally, we describe a limitation uncovered by SMOCAM for the volume prediction of the Hippocampi. Figure 5 shows that the attention for the prediction of both hippocampi in this subject was localized on the left hemisphere. Indeed, the neural network estimated a hemispherical asymmetry of 0.10cm^3 whereas the ground truth variation was 10 times bigger. We hypothesized that only one hippocampus was used for both left and right predictions because in average both sides are well correlated and that large asymmetries are underrepresented in our training dataset (see Figure 6).

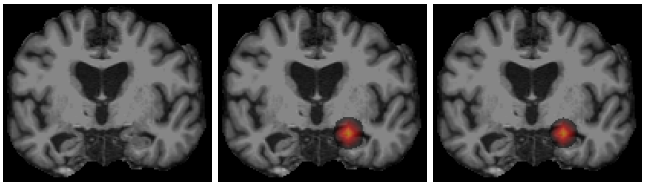

Fig. 5: Patient with large assymetry of the ventricule 1.12cm^3 (left), attention mask for the right hippocampus volume (middle), attention mask for left hippocampus volume (right).

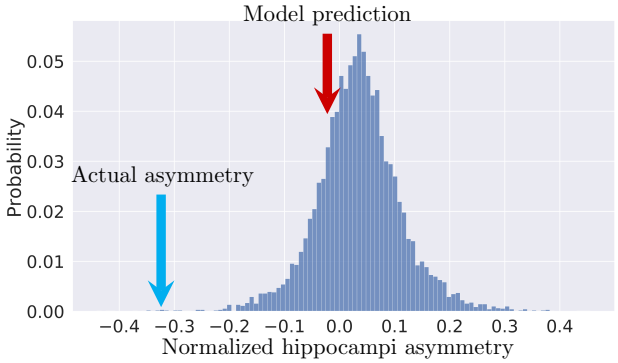

Fig. 6: Normalized hippocampi asymmetries (right hemisphere minus left hemisphere volume), position of the individual within the distribution of the training dataset (blue) and prediction made by the deep-learning model (red).

5. CONCLUSION

This manuscript introduces SMOoth Conditional Attention Mask (SMOCAM) a saliency method for neural network based regression models from 3D imagery. We showed that SMOCAM revealed serious limitations and issues when estimating bio-markers from 3D T1w MRI of the brain, and should therefore be considered before concluding about the accuracy of CNN based regression methods.

Compliance with Ethical Standards

This research was approved by CSIRO ethics 2020_068_LR.

References

[1] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[2] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[4] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *arXiv preprint arXiv:1704.02685*, 2017.

[5] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3429–3437.

[6] R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2950–2958.

[7] A. Taha, X. Yang, A. Shrivastava, and L. Davis, "A generic visualization approach for convolutional neural networks," *arXiv preprint arXiv:2007.09748*, 2020.

[8] E. Thibeau-Sutre, O. Colliot, D. Dormont, and N. Burgos, "Visualization approach to assess the robustness of neural networks for medical image classification," in *Medical Imaging 2020: Image Processing*, vol. 11313. International Society for Optics and Photonics, 2020, p. 113131J.

[9] C. Pettigrew, A. Soldan, Y. Zhu, M.-C. Wang, A. Moghekar, T. Brown, M. Miller, M. Albert, B. R. Team *et al.*, "Cortical thickness in relation to clinical symptom onset in preclinical ad," *NeuroImage: Clinical*, vol. 12, pp. 116–122, 2016.

[10] P. Nopoulos, V. A. Magnotta, A. Mikos, H. Paulson, N. C. Andreasen, and J. S. Paulsen, "Morphology of the cerebral cortex in preclinical huntington's disease," *American Journal of Psychiatry*, vol. 164, no. 9, pp. 1428–1434, 2007.

[11] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[12] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter, "Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline," *NeuroImage*, p. 117012, 2020.

[13] R. Santa Cruz, L. Lebrat, P. Bourgeat, C. Fookes, J. Fripp, and O. Salvado, "Deepcsr: A 3d deep learning approach for cortical surface reconstruction," *arXiv e-prints*, p. arXiv 2010.11423, 2020.

[14] M. Rebsamen, Y. Suter, R. Wiest, M. Reyes, and C. Rummel, "Brain morphometry estimation: From hours to seconds using deep learning," *Frontiers in neurology*, vol. 11, p. 244, 2020.

[15] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[16] C. C. Rowe, K. A. Ellis, M. Rimajova, P. Bourgeat, K. E. Pike, G. Jones, J. Fripp, H. Tochon-Danguy, L. Morandeau, G. O'Keefe *et al.*, "Amyloid imaging results from the australian imaging, biomarkers and lifestyle (aibl) study of aging," *Neurobiology of aging*, vol. 31, no. 8, pp. 1275–1283, 2010.

[17] R. S. Cruz, L. Lebrat, P. Bourgeat, V. Doré, J. Dowling, J. Fripp, C. Fookes, and O. Salvado, "Going deeper with brain morphometry using neural networks," *arXiv preprint arXiv:2009.03303*, 2020.

[18] L. Armijo, "Minimization of functions having lipschitz continuous first partial derivatives," *Pacific Journal of mathematics*, vol. 16, no. 1, pp. 1–3, 1966.