The effect of skull-stripping on transfer learning for 3D MRI models: ADNI data

Polina Druzhinina[1] POLINA.DRUZHININA@SKOLTECH.RU
Ekaterina Kondrateva[1] EKATERINA.KODRATEVA@SKOLTECH.RU
Maxim Sharaev[1] M.SHARAEV@SKOLTECH.RU
[1] *Skoltech, Skolkovo Institute of Science and technology (Moscow, Russia)*

Abstract

In recent years, with the improvement of data collection and preprocessing, as well as the development of deep learning algorithms, there have been more opportunities for applying artificial intelligence to different areas, including neuroimaging. Various model learning pipelines are emerging to study the degree of cognitive impairment in diseases such as Alzheimer's disease (AD). In this study, we explore knowledge transfer for the stability of the 3D computer vision models (CNN) for the classification of AD on ADNI data. To assess the model performance, and the quality of learned patterns and examine the ways of models overfitting we utilize conventional 3DCNN interpretation methods and swap tests. We imply that skull-stripping and knowledge transfer strategies can significantly impact the robustness and reproducibility of learned patterns, and suggest to apply swap tests to ensure the model stability.

Keywords: 3D CNN, ADNI, skull-stripping, GradCAM, Guided Backpropagation

1. Introduction

Three-dimensional structural magnetic resonance imaging (MRI) processing studies are particularly beneficial as they aid in the early detection and treatment of many diseases. MR imaging is a huge asset for for the early diagnosis of disorders such as Alzheimer's disease (AD). This is a progressive neurodegenerative disorder associated with brain damage, leading to progressive dementia. Alzheimer classification on ADNI data becomes a benchmark claffigication task on brain MRI data with state-of-the-art ML approaches. Yet, the prediction of AD itself is considered clinically meaningless, whereas the Mild Cognitive Impairment (MCI) classification, its subtypes, and biomarkers exploration is a topic of great interest. The investigation of MCI subtasks remains problematic and challenging: the MCI classification are far less accurate than AD, and can be echansed by model transfer the benchmark models.

There are numerous stuies devoted to MCI recognition, based on transfer learning (TL) from AD; numerous approaches with different data preprocessing. and TL strategies are compared in terms of classification accuracy. Yet, the stability and robustness of the learned patterns for transfer were never questioned.

In the current study, we explore the effect of data preprocessing (skull-stripping) and transfer learning strategy on the AD classification results. And examine the models on their stability with swap test and 3DCNN interpretation methods.

2. Materials and Methods.

Data description. The publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) database [1] was used in this study. We choose namely structural T1 MRI images which was spatially normalized, masked, and N3 corrected. For this study we use only baseline scans (the first session) of all dataset releases (ADNI 1,2,Go,3) for 2 diagnostic groups: Controls Normal (CN) - healthy participants, and patients with Alzheimer's disease (AD). The dataset consists 303 subjects including 163,140 scans of AD and CN correspondingly.
Experimental design. There were two types of preprocessing. In the first case, we used the bias field correction (N4ITK), linear (affine) registration in the MNI space (SyN algorithm), and intensity scaling (MinMax). To optimize the calculations, the background was removed, and the size of the final image was [169 × 208 × 179]. The entire preprocessing process was performed using the Clinica [2] library. The second pipeline repeated the steps of the previous one but adds the skull-stripping (SS) step computed with FSL BET.
Model. The training model for the ADNI dataset at the 3D-subject level was 3D CNN, consisting of 5 convolutional blocks and 3 fully connected layers. For the experiments with transfer learning the data were previously trained on an autoencoder (AE), as in SOTA from Clinica. The CNN loss function was cross-entropy, with softmax log normalization and negative log-likelihood loss, while for an AE was used the mean-squared error loss. Optimizer is Adam with hyperparameters betas = (0.9, 0.999), epsilon = 1e-4. The maximum number of epochs is 50, we also follow 20 epochs of patience as the criterion of early stopping. 5-fold cross-validation was performed with the Clinica library.
Model interpretation and swap tests. To study the impact of transfer learning and SS on training models and their performance, we use the basic classical methods of interpretation, namely GradCam and GB.

In order not to be misleading by the visual appeal of heat-maps, we use swap tests for additional verification: the data randomization and model randomization tests (Adebayo et al., 2018). Thus, such sanity check is able to reflect the dependence of the interpretation method on the model or on the training sample. If the approach is independent, it becomes clear that the obtained significance maps cannot be trusted. We assume that the criterion for passing the test is how the significance maps of the interpretation change depending on the randomization of the label or the corruption of the model weights. If the significance map does not differ from the one generated on true labels, or with real weights parameters, then this model stability and robustness is questioned.

3. Results and Discussion

At first we explored, how skull stripping affect classification quality. We show, that this step influences performance and helps model convergence (ROC AUC scores 84.0 and 88.45 for test and validation compared 73.0 and 75.6 of the experiment without skull stripping). But at the same time, it responsible for image corruption and loose of information, which could be critical to prediction of AD pathology. We experiment with different transfer learning strategies (Table 1) and try to find the balance between better performance with

[1] http://adni.loni.usc.edu/about/
[2] https://github.com/aramis-lab/clinica/

Table 1: The average ROC AUC over 5 folds on test/train/val for AD/ CN task on ADNI data. The comparison of experiments with TL (without AE/ with AE) and skull stripping (wo SS/ with SS). Experiments with passed randomization (swap) tests are marked in green

	wo AE	with AE w/o SS	with AE with SS
. 3D CNN w/o SS	73.0/ 90.0/ 75.6	77.0/ 98.9/ 86.8	78.5/ 99.2/ 84.8
3D CNN with SS	84.0/ 100.0/ 88.5	84.5/ 100.0/ 88.3	77.0/ 89.3/ 80.7

fast convergence and level of redusing informative features and regions on the border of the sculp. For that, we took two canonical methods of CNN interpretation, and performed swap tests on the best performing models to verify the stability of learning patterns.

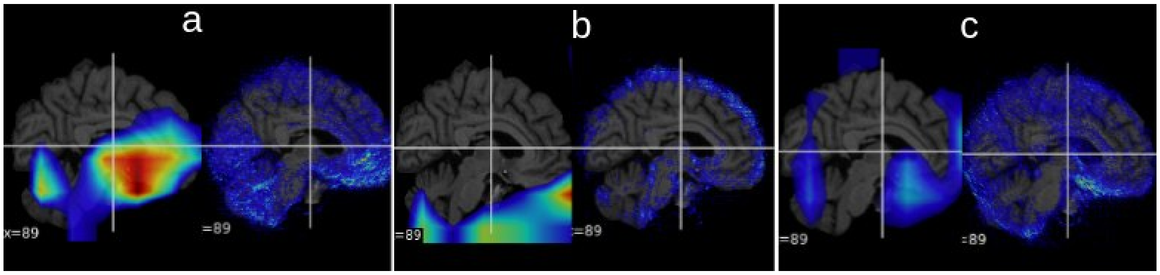

Figure 1: GradCAM and GB attention map on model with TL on AE w/o SS and 3D CNN with ss (a): the original model attention map with real AD disorder biomarkers; (b): model randomization test - GradCam pays attention on non-brain regions, GB tends to highlight edges (c): label randomization test - GradCAM map is corrupted.

We observe, that although all experiments with 3DCNN with skull stripping (Table 1 -low raw) passed randomization tests, the most stable stable interpretation during randomization tests was exhibited by model with AE w/o skull-stripping and 3DCNN with SS. This strategy identified more AD-specific biomarkers than a model trained on SS data without transfer learning, and has a fairly high quality compared to others (84.5 ROC AUC on test). We conclude that classical methods of interpretation are both unstable for randomisation tests, and GB is tough to interpret in clinical perspective. Yet these classical methods could be used to ensure stability of learned patters while transfer.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.