Interpreting Convolutional Networks for MRI-based Diagnosis of Alzheimer’s Disease

Johannes Rieke  Fabian Eitel  Martin Weygandt  John-Dylan Haynes  Kerstin Ritter
Charité – Universitätsmedizin Berlin, Bernstein Center for Computational Neuroscience Berlin, Berlin Center for Advanced Neuroimaging, Department of Neurology, Excellence Cluster NeuroCure
johannes.rieke@gmail.com

Abstract
Interpreting convolutional neural networks (CNNs) is challenging, but critical in areas such as medical decision making. In this study, we train a 3D CNN to detect Alzheimer’s disease based on structural MRI scans of the brain. Then, we apply several methods that explain the network’s classification decisions by highlighting relevant areas in the input image. We analyze the methods’ differences and find that our network indeed focuses on brain regions associated with Alzheimer’s disease.

1 Introduction
Alzheimer’s disease (AD) is the most common underlying disease for dementia in the elderly and leads to severe tissue loss (i.e. atrophy) in the brain, starting in the hippocampus. However, recognizing AD on MRI scans is a challenging task, especially during early stages. In this context, machine learning models provide great potential to capture even slight tissue alterations. State-of-the-art models for image classification are convolutional neural networks (CNNs), which have recently been applied to medical imaging data for various use cases [1], including AD detection. Although CNNs deliver good classification results, they are difficult to interpret. In medical decision making, however, it is critical to explain the behavior of a machine learning model and let medical experts verify the diagnosis. A number of interpretation methods have been suggested that highlight regions in an input image with strong influence on the classification decision [2–5]. In this work, we compare some of these interpretation methods on a 3D CNN, which was trained to classify structural MRI scans of the brain into AD patients and normal elderly controls (NCs).

2 Related Work
A number of machine learning models have been applied to automated Alzheimer detection [6]. [7–9] are similar to our approach (i.e. 3D CNN, full-brain structural MRI, AD vs. NC). We found two studies that apply interpretation methods to AD classification: [9] employ the occlusion method and show focus on hippocampus and ventricles. [5] use a segmentation-based occlusion (similar to our brain area occlusion), but reach inconclusive results. To the best of our knowledge, this is the first study that comprehensively compares different interpretation methods on CNNs for AD detection.

3 Methods
Our dataset consists of 969 structural MRI scans (475 AD, 494 NC) of 344 subjects (193 AD, 151 NC) from the Alzheimer’s Disease Neuroimaging Initiative (ADNI, http://adni.loni.usc.edu/). All MRI scans have been registered to a 1 mm isotropic ICBM template, resulting in a volume size of 193 x 229 x 193. For the test set, we randomly selected 30 patients with AD and 30 patients with NC.

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.
Figure 1: Relevance heatmaps for all interpretation methods, averaged over AD samples in the test set. Red indicates relevance, i.e. a red area was important for the network’s classification decision. Numbers indicate slice positions (out of 229 coronal slices).

Our model consists of four convolutional layers (with filter size 3 x 3 x 3 and 8/16/32/64 feature maps) and two fully-connected layers (128/64 neurons). We apply batch normalization and pooling after each convolution and dropout of 0.8 before the first fully-connected layer. The network has two output neurons with softmax activation. We train with cross-entropy loss and the Adam optimizer (learning rate 0.0001, batch size 5). Before feeding the brain scans to the network, we remove the skull and normalize each voxel to have mean 0 and standard deviation 1 across the training set.

We apply four different interpretation methods to our network: Sensitivity Analysis via Backpropagation [2], Guided Backpropagation [3] (for both methods, we take the absolute value of the gradient), Occlusion [4] (we use a patch of size 40 x 40 x 40 with value 0 and calculate the difference between unoccluded and occluded probability), and Brain Area Occlusion (a modification of occlusion, where we occlude an entire brain area based on the AAL atlas, inspired by [5]). All of these interpretation methods produce a relevance heatmap for each input image, which highlights image regions that are relevant for the classification decision. PyTorch implementations of all interpretation methods will be made available at [http://github.com/jrieke/cnn-interpretability].

4 Results

Our network achieves a classification accuracy of 81 % and a ROC AUC of 83 % on the test set. This is comparable to recent studies for similar models [7–9]. Please note that we did not perform hyperparameter optimization, because our focus was on the interpretation methods.

In Fig. 1, we show the average heatmaps over all AD samples in the test set. The averages over NC samples look similar, so we omit them here. Based on these heatmaps, we list in Table 1 the four most relevant brain areas for each interpretation method, calculated as the sum of the relevance in each brain area (according to the AAL atlas).

We can see that the main focus of the network is on the temporal lobe, especially its medial part. This brain area, containing the hippocampus and other structures associated with memory, has been empirically linked to AD [10]. The hippocampus itself is usually one of the earliest areas affected...
Table 1: Most relevant brain areas per interpretation method, averaged over AD samples in the test set. Values in brackets give fraction of complete relevance across the brain.

<table>
<thead>
<tr>
<th>Sensitivity Analysis (Backpropagation)</th>
<th>Guided Backpropagation</th>
<th>Occlusion</th>
<th>Brain Area Occlusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>TemporalMid (6.1 %)</td>
<td>TemporalMid (7.0 %)</td>
<td>TemporalMid (12.1 %)</td>
<td>TemporalMid (29.7 %)</td>
</tr>
<tr>
<td>TemporalInf (5.9 %)</td>
<td>TemporalInf (5.7 %)</td>
<td>TemporalInf (9.2 %)</td>
<td>TemporalInf (14.8 %)</td>
</tr>
<tr>
<td>Fusiform (4.6 %)</td>
<td>FrontalMid (4.2 %)</td>
<td>Fusiform (6.2 %)</td>
<td>TemporalSup (4.4 %)</td>
</tr>
<tr>
<td>CerebelumCrus1 (3.8 %)</td>
<td>Fusiform (3.9 %)</td>
<td>ParaHippocampal (5.4 %)</td>
<td>Hippocampus (4.1 %)</td>
</tr>
</tbody>
</table>

by AD [11]. In our experiments, we observe some relevance on the hippocampus, but usually the whole area around it is crucial for the network’s decision. This may be explained by the fact that our samples contain only advanced forms of the disease. In addition to temporal regions, we observe some relevance attributed to other areas across the brain (especially in the gradient-based interpretation methods). We find that the distribution of relevance varies between patients: Some brains have strong relevance in the temporal lobe, while in others, the cortex plays a crucial role.

Although all interpretation methods focus on similar brain areas, we can spot some differences: Occlusion and Brain Area Occlusion are more focused on specific regions, while relevance in the gradient-based methods seems more distributed. Obviously, the occlusion-based approaches cannot deal with large areas of distributed relevance (e.g. in the cortex), because these areas will never be covered up completely by the occlusion patch. Therefore, we recommend to apply gradient-based instead of occlusion-based interpretation methods for use cases where the relevance is presumably distributed across the input image. Moreover, we find that Brain Area Occlusion is admittedly a very natural approach for our context, but it suffers from the fact that only one brain region is covered up at a time. In our case, this leads to very high relevance for the temporal lobe, but hardly any relevance for other brain structures.

5 Conclusion

By using several interpretation methods, we show that our CNN indeed focuses on brain regions associated with AD, in particular the medial temporal lobe. Further research might identify potential subgroups of patients and address the sensitivity of methods for preclinical cases (i.e. mild or subjective cognitive impairment).

References