Studying the Interpretability of Brain MR Image Deep Learning Modelling

Data Science Institute
COLUMBIA UNIVERSITY

Justin Kennedy[1], Siyuan Shi[2], Jiook Cha[3]
[1]Data Science, [2]Data Science, [3]Pyschiatry, Columbia University

Data Science Capstone Project
with Professor Jiook Cha

Overview

In this project, we looked to identify parts of the brain that are relevant towards prediction of specific levels of cognitive impairment, from cognitively normal to diagnosed alzheimer's disease patients. To do this, we applied several machine learning algorithms in aggregation on patient brain MR image sequences to generate region-specific heatmaps of each patient sample. In building off the resulting interpretations, this project has implications on forecasting an individual's neurocognitive development, cognition, and behaviors.

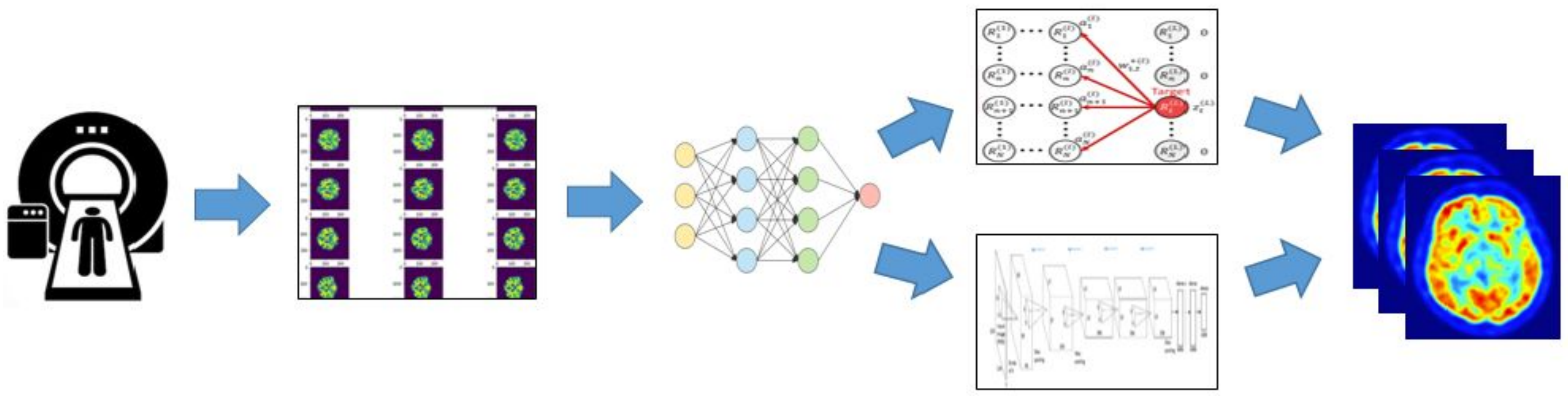
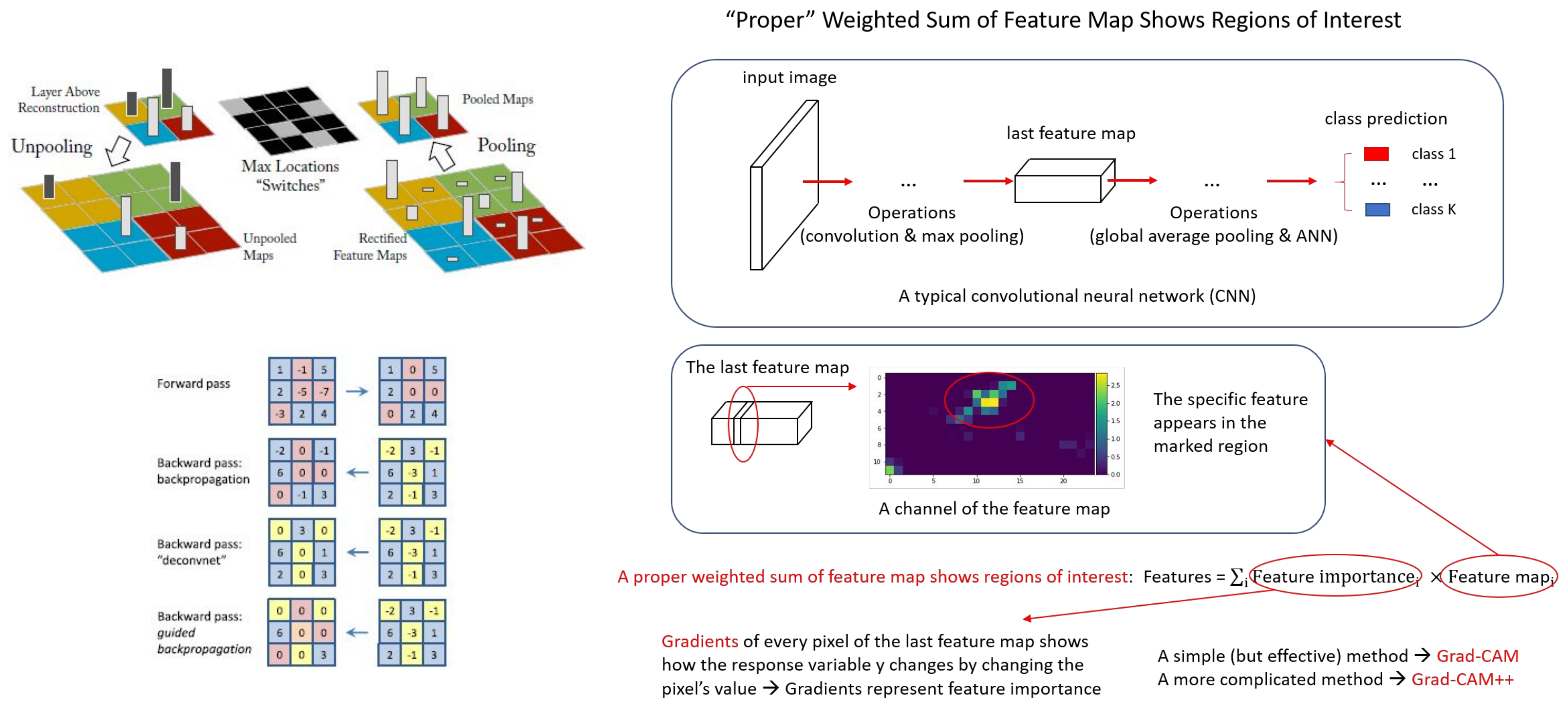


Figure 1. A patient's 3D MRI sequence is passed into a 3D CNN model that predicts the patient's cognitive group. Multiple MLI algorithms are applied to the CNN to produce a 'summary score' 3D pixel-wise contribution heatmap.

Method

Layer-wise relevance propagation (LRP) works by creating a heatmap of per-pixel contributions to an input image's output label by computing relevances of neurons in each layer in a backward pass. In Guided Backpropagation (GBP), the gradient is backpropagated and all negative gradients are set to 0 to identify the neurons that contribute positively to the predicted group. Grad-CAM(++) works by constructing a weighted sum of feature maps using gradients in the network.


Layer Above Reconstruction
Unpooling
Max Locations "Switches"
Pooled Maps
Pooling
Unpooled Maps
Rectified Feature Maps
Forward pass
Backward pass: backpropagation
Backward pass: "deconvnet"
Backward pass: *guided backpropagation*
"Proper" Weighted Sum of Feature Map Shows Regions of Interest
input image
last feature map
class prediction
class 1
class K
Operations (convolution & max pooling)
Operations (global average pooling & ANN)
A typical convolutional neural network (CNN)
The last feature map
The specific feature appears in the marked region
A channel of the feature map

A proper weighted sum of feature map shows regions of interest: Features = $\sum_i$ Feature importance$_i$ × Feature map$_i$

Gradients of every pixel of the last feature map shows how the response variable y changes by changing the pixel's value → Gradients represent feature importance

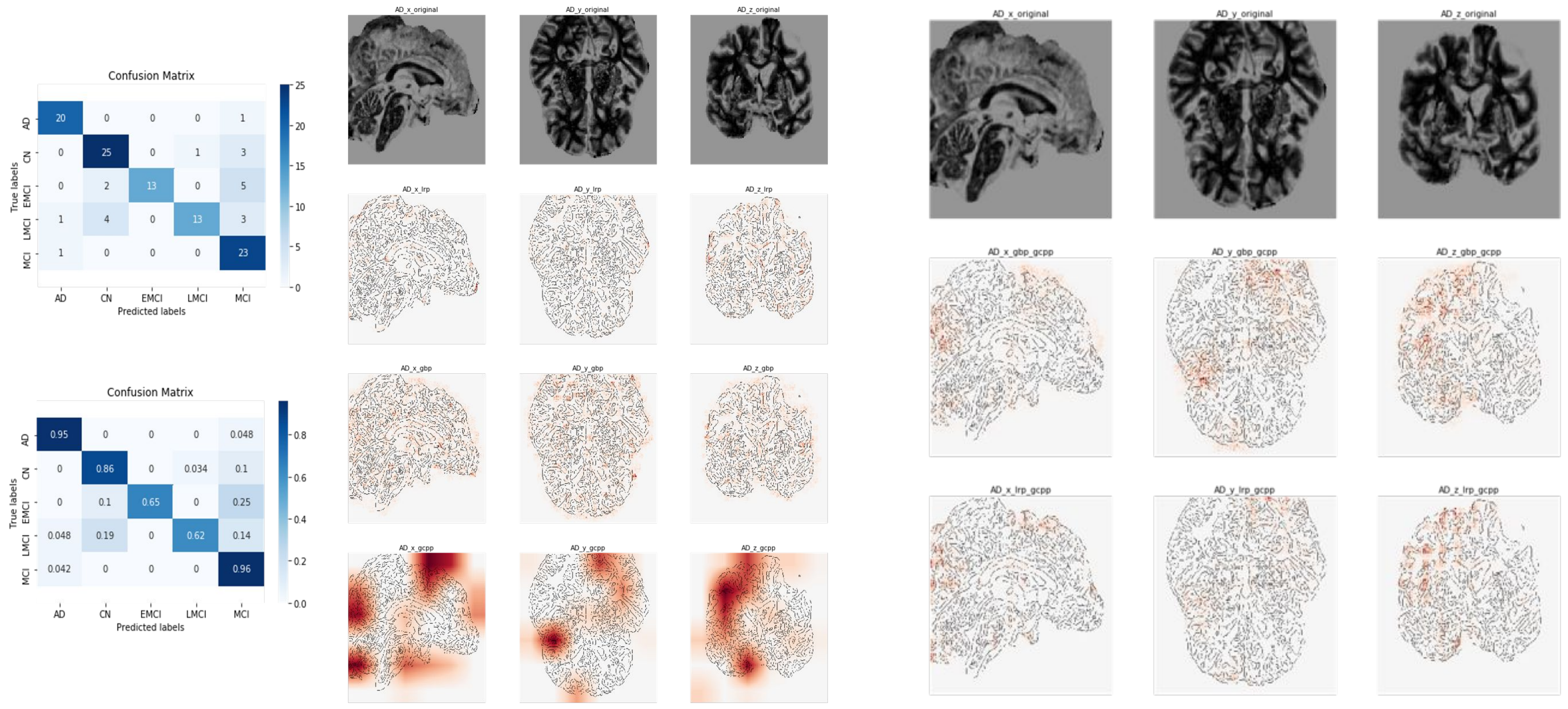
A simple (but effective) method → Grad-CAM
A more complicated method → Grad-CAM++

Figure 2. Core ideas of MLI algorithms. Left: guided backpropagation. Right: Grad-CAM (Grad-CAM++).

Results

A subsample of 133 instances was chosen to test the MLI algorithms on. The model's predictions on the sample are summarized as a confusion matrix in Figure 3.

Here we show the resulting MLI interpretations of one of the instances. The LRP and GBP algorithms give per-pixel details; the colored pixels are important for making the prediction. The Grad-CAM (Grad-CAM++) result focuses on the region-level details; the colored regions contain features that the model thinks belong to the predicted class.



Confusion Matrix

True labels \ Predicted labels	AD	CN	EMCI	LMCI	MCI
AD	20	0	0	0	1
CN	0	25	0	1	3
EMCI	0	2	13	0	5
LMCI	1	4	0	13	3
MCI	1	0	0	0	23

Confusion Matrix

True labels \ Predicted labels	AD	CN	EMCI	LMCI	MCI
AD	0.95	0	0	0	0.048
CN	0	0.86	0	0.034	0.1
EMCI	0	0.1	0.65	0	0.25
LMCI	0.048	0.19	0	0.62	0.14
MCI	0.042	0	0	0	0.96

Figure 3. Left: confusion matrix of the subsample. Middle: input image, LRP, GBP, and Grad-CAM++ results. Right: Layer-wise relevance Grad-CAM++ & Guided Grad-CAM++.

Conclusion

Though difficult to assess the accuracy of the resulting MLI heatmaps, especially given their dependence on the accuracy of the model they're based on, we can observe a consistency between separate MLI methods in certain brain regions being commonly specified as predictive of cognitive group. As the model's accuracy improves, these results could help infer area-specific characteristics related to cognitive conditions.

Acknowledgments

We would like to thank Seungwook Han and Professor Jiook Cha for providing helpful feedback over the course of this project.

References

Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," PLoS ONE 10(7): e0130140, 2015.

"Grad-CAM: Visual Explanations from Deep ...." 21 Mar. 2017, https://arxiv.org/abs/1610.02391.

"Grad-CAM++: Generalized Gradient-Based ...." https://ieeexplore.ieee.org/document/8354201.