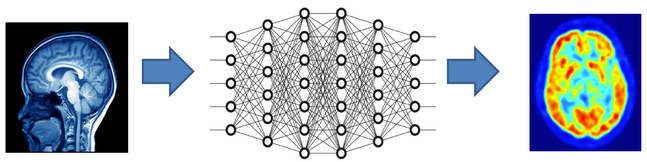
Interpretable Convolutional Neural Networks for Alzheimer's Detection

Michael Wang
mzwang14@Stanford.edu

https://youtu.be/xsXKoIQVO0w

Abstract/Introduction

- Alzheimer's Disease is the 6th leading cause of death
- Magnetic Resonance Imaging can be used for detection
- Difficult to detect with purely behavioral metrics – easily confused with normal aging
- Implemented 2D and 3D CNN's along with Grad-CAM algorithm for visualization
- 76% Accuracy, .78 F1 score



Dataset

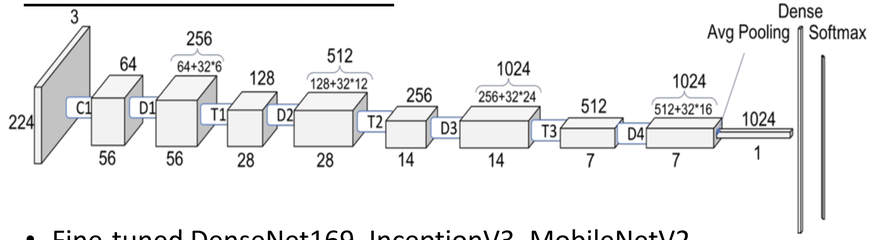
- 489 Demented, 609 normal cognition, 3114 NIFTI scans
- Psychiatric evaluations in CSV formats, labelled by matching each scan with closest psychiatric evaluation
- Normalized with N4 Bias Correction, often used to correct non uniformity in MRI images
- Used whole scan for 3D, middle horizontal slice for 2D

	max CDR					
min CDR	0	0.5	1	2	3	Grand Total
0	609*	192	39	12	2	854
0.5		66	61	45	5	177
>1			31	31	5	67
Grand Total	609	258	131	88	12	1098

Unchanged CDR = 0 represents cognitively healthy population

Model/Training

DenseNet architecture



- Fine-tuned DenseNet169, InceptionV3, MobileNetV2 performed the best(replaced Dense layers)
- 80-10-10 Train-Dev-Test split
- For each model, tuned dropout rate, augmentation, learning rate
- **Loss**: Binary Cross Entropy **Optimizer:** Adam **Batch size:** 23

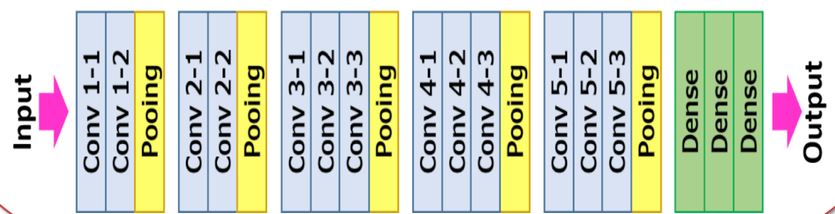
Grad-CAM

- Similar to regular Class Activation Maps, but uses activation map of final convolutional layer to determine important pixels
- Performed before flattening layer, so no need to retrain network

$$S^c = \frac{1}{K}\sum_i\sum_j\sum_k \frac{\partial y^c}{\partial A^j_{ik}} A^j_{ik}$$

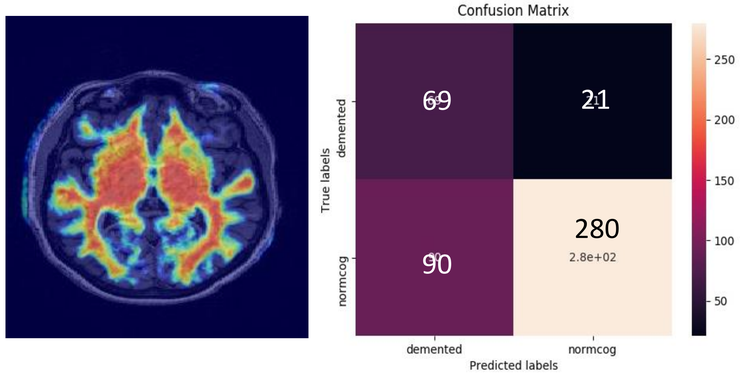
3D Models

- Implemented ground up VoxCNN, based on 3d VGG
- Major memory issues. Batch size of 4, which did not ensure samples of each class being in each batch



Results

	DenseNet 169	Inception V3	MobileNet V2
Best Dropout Rate	.5	.4	.6
Best Learning Rate	.001	.001	.0001
Data Augmentation	Yes	No	Yes
Test Accuracy Avg(weighted)	.74	.76	.75
Test Precision Avg(weighted)	.83	.83	.83
Test Recall Avg(weighted)	.74	.76	.75
Test F1 Avg(weighted)	.77	.78	.77
Test AUC	.81	.8	.76



3D model performed no better than random – likely due to memory constraints

Conclusion

- Fine-tuned several existing image recognition models, best results being from Inception V3 at .78 F1 score
- 3D networks unsuccessful, too much time needed to train and tune hyperparameters
- Grad-CAM used to visualize output.

Next Steps

- Increased memory resources could lead to better performance with 3D models
- Use Grad-CAM output to aid diagnosis
- Perform intermediary steps rather than a purely end to end approach