Distributed Weight Consolidation: A Brain Segmentation Case Study

Patrick McClure[1], Charles Y. Zheng[1], Jakub R. Kaczmarzyk[2], John A. Lee[1], Satrajit S. Ghosh[2], Dylan Nielson[1], Peter Bandettini[3], Francisco Pereira[1]

1) Machine Learning Team, National Institute of Mental Health; 2) McGovern Institute for Brain Research, Massachusetts Institute of Technology; 3) Section on Functional Imaging Methods, National Institute of Mental Health

Summary

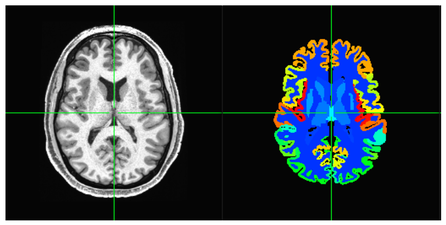
Problem: Training a network on distributed data from multiple sites that cannot share their data

Potential Solution: Use continual learning to sequentially update a network one site at a time using the previous network as a prior

Drawback: Each site will have to wait for all previous sites before training a network on their data

Our Proposal: Use continual learning at each site and then combine the resulting networks into a new prior for further training

Data and Architecture



Input

FreeSurfer Target

0 “background”
1 “white matter”
⋮
10 “hippocampus”
11 “amygdala”
⋮
48 “ctx-transversetemporal”
49 “ctx-insul”

Labels [1]

Data

HCP [2]: ~900 training and ~100 test MRI volumes

NKI [3]: ~1000 training and ~100 test MRI volumes

Buckner [3]: ~200 training and ~20 test MRI volumes

WU120 [4]: ~100 training and ~10 test MRI volumes

ABIDE [5]: ~200 test MRI volumes

Architecture

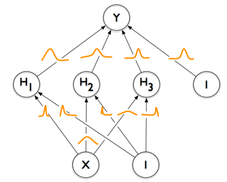
MeshNet [1,5]: An 8-layer 3d dilated convolutional neural network

Methods

Variational Bayesian Neural Networks [6,7,8,9]

$$p(W|D_1) = \frac{p(D_1|W)p(W)}{p(D_1)}$$

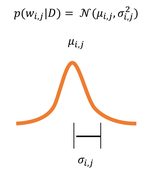
$$\underset{V}{\operatorname{argmin}} - \int q_V(W)\ \log p(D_1|W)\, dW + KL[q_V(W)||p(W)]$$



Variational Continual Learning (VCL) [10]

$$p(W|D_{1:T}) = \frac{p(D_T|W)p(W|D_{1:T-1})}{p(D_T)}$$

$$\underset{V}{\operatorname{argmin}} - \int q_V(W)\ \log p(D_T|W)\, dW + KL[q_V(W)||p(W|D_{1:T-1})]$$



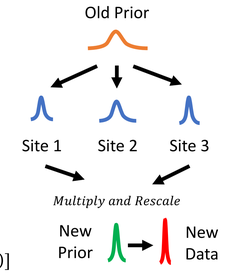
Distributed Weight Consolidation (DWC)

$$D_T = \{D_T^1, \ldots, D_T^S\}$$

$$p(W|D_{1:T-1}, D_T^s) = \frac{p(D_T^s|W)p(W|D_{1:T-1})}{p(D_T^s)}$$

$$p(W|D_{1:T}) = \frac{1}{p(W|D_{1:T-1})^{S-1}} \prod_{s=1}^{S} p(W|D_{1:T-1}, D_T^s)$$

$$\underset{V}{\operatorname{argmin}} - \int q_V(W)\ \log p(D_{T+1}|W)\, dW + KL[q_V(W)||p(W|D_{1:T})]$$



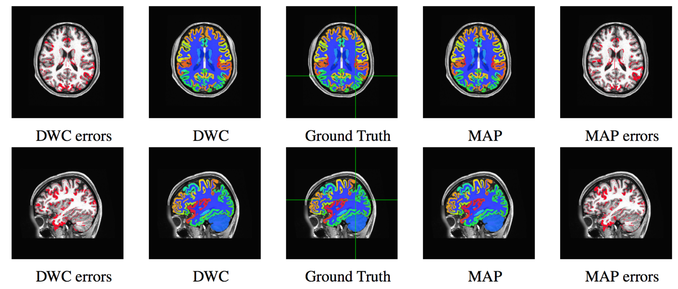
Numerical Results

Average Dice Score across Classes

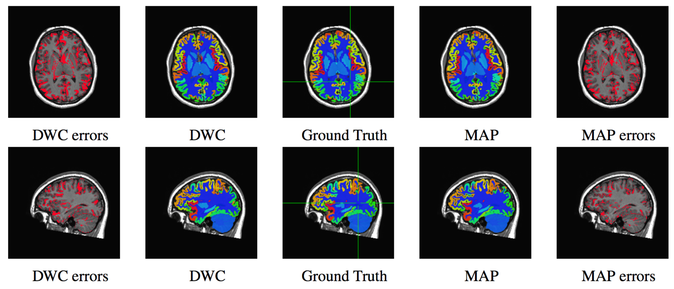
$Network$	H	N	B	W	Avg.	A
H_{MAP}	82.25	65.88	67.94	70.88	72.92	55.25
N_{MAP}	71.2	72.19	70.73	73.06	71.66	66.67
B_{MAP}	65.69	50.17	82.02	68.87	59.25	50.23
W_{MAP}	70.18	66.27	72.2	76.38	59.25	62.83
$H \rightarrow N$	75.40	73.24	71.77	73.17	74.03	64.62
$H \rightarrow B$	73.85	56.79	79.49	68.53	65.78	49.27
$H \rightarrow W$	77.07	67.63	76.15	77.26	67.63	62.31
$H \rightarrow N \rightarrow B \rightarrow W$	77.42	71.46	79.70	79.82	74.86	63.3
$H \rightarrow W \rightarrow B \rightarrow N$	78.04	78.15	75.79	79.50	77.99	70.79
$H \rightarrow N + B + W$ (DWC w/o FT)	78.28	73.52	78.02	77.37	75.95	65.56
$Ensemble$	79.13	72.32	80.02	78.84	75.94	66.27
$H \rightarrow N + B + W \rightarrow H$ (DWC)	80.34	73.64	77.46	78.10	76.82	66.21
$HNBW_{MAP}$	81.38	77.99	80.64	79.54	79.62	70.76

Visual Results

HCP



NKI



Conclusions

There are many problems for which accumulating data into one accessible dataset for training can be difficult or impossible, such as for clinical data. It may, however, be feasible to share models derived from such data.

DWC:

- Allows for using Bayesian continual learning non-sequentially
- Performs as well or better than or better than the MAP ensemble

References

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com