Distributed Weight Consolidation: A Brain Segmentation Case Study

Patrick McClure[1], Charles Y. Zheng[1], Jakub R. Kaczmarzyk[2], John A. Lee[3], Satrajit S. Ghosh[2], Dylan Nielson[3], Peter Bandettini[4], Francisco Pereira[1]

1) Machine Learning Team, National Institute of Mental Health; 2) McGovern Institute for Brain Research, Massachusetts Institute of Technology; 3) Data Science and Sharing Team, National Institute of Mental Health; 4) 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

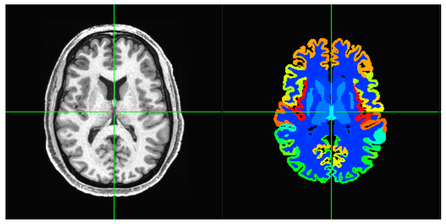
Potential Solution: Use continual learning to sequentially update a network one site at a time using the previous network as a prior

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

Example Brain Segmentation Training Pair



Data

Datasets: 4 datasets used for training and test (HCP, NKI, Buckner, and WU120) and 1 held out dataset only for testing (ABIDE)

Dataset	Training Examples	Testing Examples
HCP[3]	~900	~100
NKI[4]	~1000	~100
Buckner[5]	~200	~20
WU120[6]	~100	~10
ABIDE[7]	0	~2,000

Architecture

MeshNet[2]: An 8-layer 3d dilated convolutional neural network

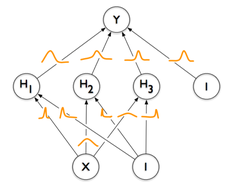
Layer	Filter	Padding	Dilation	Non-linearity
1	$96x3^3$	1	1	ReLU
2	$96x3^3$	1	1	ReLU
3	$96x3^3$	1	1	ReLU
4	$96x3^3$	2	2	ReLU
5	$96x3^3$	4	4	ReLU
6	$96x3^3$	8	8	ReLU
7	$96x3^3$	1	1	ReLU
8	$50x1^3$	1	1	Softmax

Methods

Variational Bayesian Neural Networks[8,9]

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

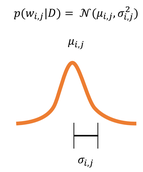
$$\underset{V}{\operatorname{argmin}} - \int q_V(W)\ \log p(D_1|W)\, dW + KL[q_V(W)||p(W)]$$



Variational Continual Learning (VCL)[10]

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

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



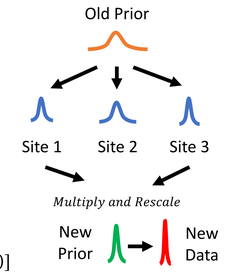
Distributed Weight Consolidation (DWC)

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

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

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

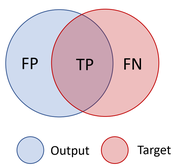
$$\underset{V}{\operatorname{argmin}} - \int q_V(W)\ \log p(D_{T+1}|W)\, dW + KL[q_V(W)||p(W|D_{1:T})]$$



Numerical Results

$$Dice = \frac{2TP}{2TP + FN + FP}$$

(Reported Dice scores are averaged across the 50 output classes)



Average test Dice scores for MAP networks trained on each site

Network	HCP (H)	NKI (N)	Buckner (B)	WU120 (W)	HNBW Avg.	ABIDE
H_{MAP}	82.25	65.88	67.94	70.88	72.92	55.25
N_{MAP}	71.20	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	72.20	76.38	68.76	62.83

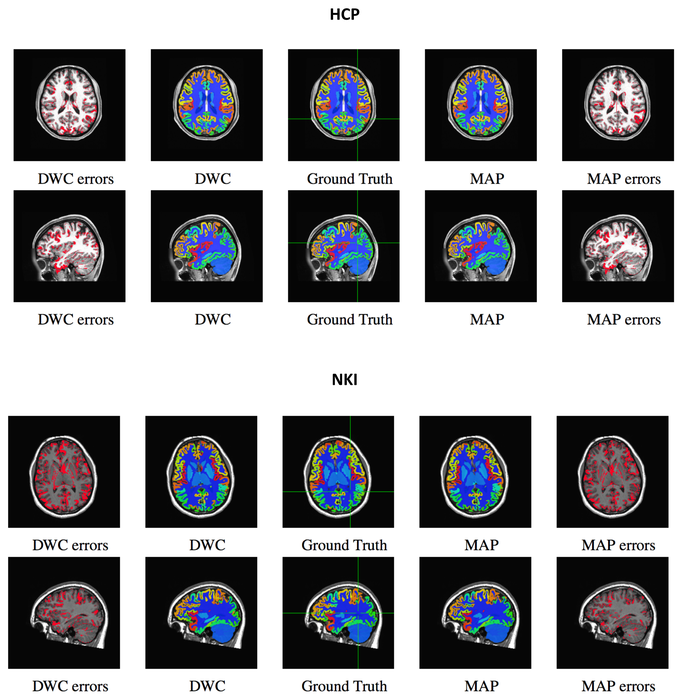
Average test Dice scores for networks trained on all sites

Network	HCP (H)	NKI (N)	Buckner (B)	WU120 (W)	HNBW Avg.	ABIDE
Ensemble	79.13	72.32	80.02	78.84	75.94	66.27
H->N+B+W->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

*$HNBW_{MAP}$ is a MAP network trained on the union of the H, N, B, and W training datasets

Visual Results

Example segmentations and errors for the DWC and $HNBW_{MAP}$ networks



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. We propose DWC, a method for combining these shared models into a single network within the Bayesian continual learning framework.

DWC:

- Allows for using Bayesian continual learning non-sequentially
- Performs as well as or better than the MAP ensemble
- Scales better with number of sites than the MAP ensemble
- Site-specialized networks train faster than MAP networks

References

[1] Fischl. Freesurfer. Neuroimage, 62(2), 2012.
[2] Fedorov et al. Almost instant brain atlas segmentation for large-scale studies. arXiv preprint arXiv:1711.00457, 2017.
[3] Van Essen et al..The Wu-Minn human connectome project: an overview. NeuroImage, 80:62–79, 2013.
[4] Nooner et al. The NKI-Rockland sample: a modelfor accelerating the pace of discovery science in psychiatry. Frontiers in Neuroscience, 6:152, 2012.
[5] Biswal et al. Toward discovery science of human brain function. Proceedings of the National Academy of Sciences, 107(10):4734–4739, 2010.
[6] Power et al. Temporal interpolation alters motion in fMRI scans: Magnitudes and consequences for artifact detection. PloS one, 12(9):e0182939, 2017.
[7] Martino et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry, 19(6):659, 2014.
[8] Graves. Practical variational inference for neural networks. In Advances in Neural Information Processing Systems, pages 2348–2356, 2011.
[9] Blundell et al. Weight uncertainty in neural network. In International Conference on Machine Learning, pages 1613–1622, 2015.
[10] Nguyen et al. Variational Continual Learning. In International Conference on Learning Representations, 2018.