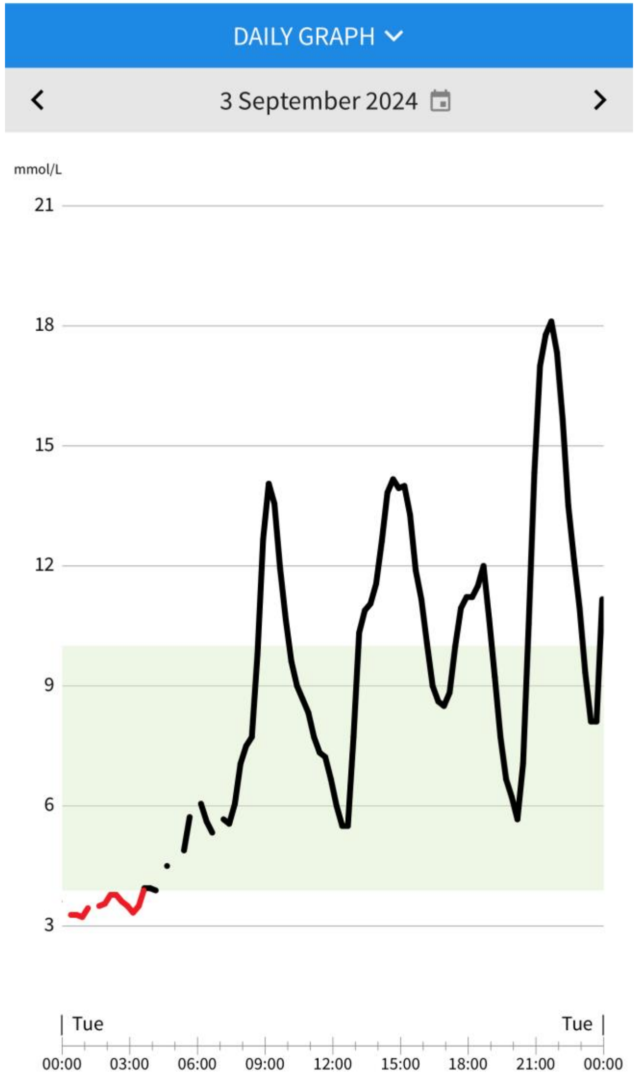
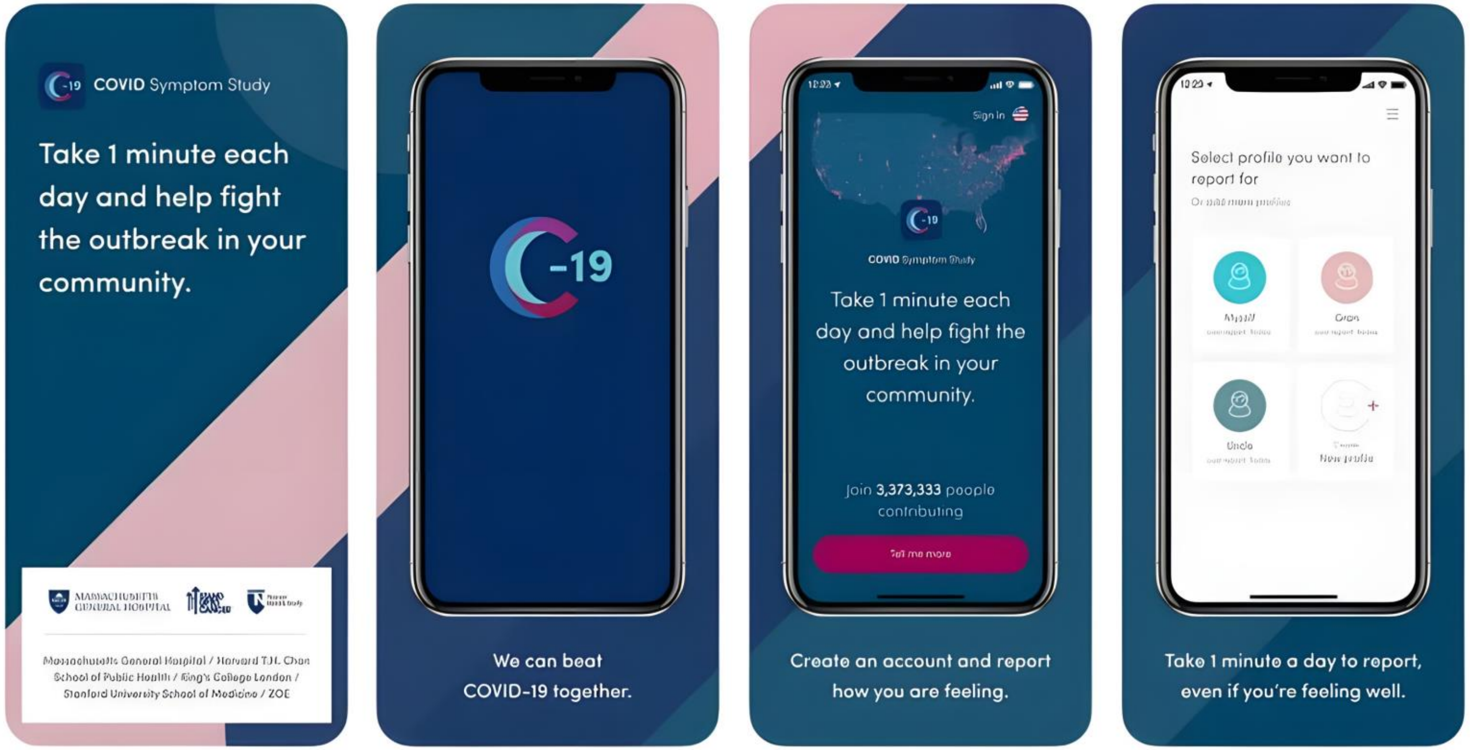
Towards Scalable, Flexible, and Interpretable Self-Supervised Learning for Multiview Biomedical Data

James Chapman


DAILY GRAPH
3 September 2024
mmol/L
21
18
15
12
9
6
3
Tue
Tue
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
00:00


COVID Symptom Study
Take 1 minute each day and help fight the outbreak in your community.
Massachusetts General Hospital / Harvard T.H. Chan School of Public Health / King's College London / Stanford University School of Medicine / ZOE
-19
We can beat COVID-19 together.
Sign in
COVID Symptom Study
Take 1 minute each day and help fight the outbreak in your community.
Join 3,373,333 people contributing
Create an account and report how you are feeling.
Select profile you want to report for
Uncle
Take 1 minute a day to report, even if you're feeling well.

Context

- Started working on *Deep Learning for Brain Behaviour Associations*

- Discovered that neuroimaging data is *really* high-dimensional and hard to work with

- And sample sizes are still small… but growing

- … and most of the relevant code was in MATLAB

Research Objectives

•Develop regularized Canonical Correlation Analysis (CCA) methods for improved interpretability

•Create efficient algorithms for large-scale datasets

•Extend CCA to deep learning and self-supervised settings

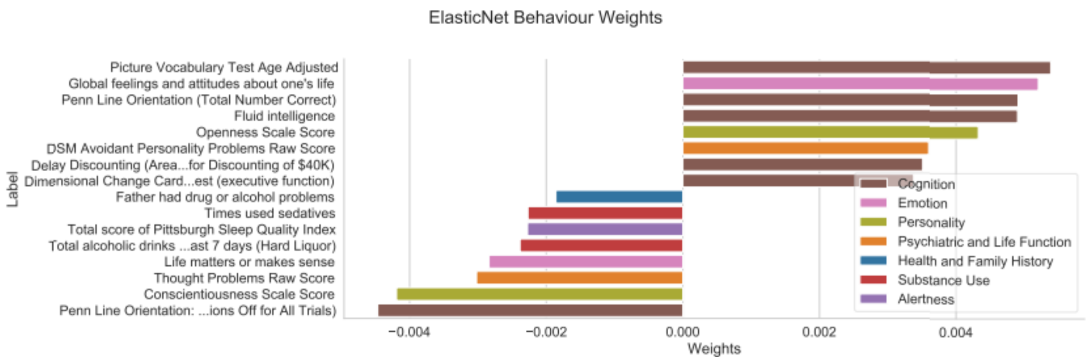
•Provide open-source implementations for the research community

FRALS Framework

- Addresses a known weakness of existing sparse CCA (Partial Least Squares) methods

- Incorporates structured priors into CCA models through regularization (e.g., elastic net)

- Enhances interpretability of CCA models



$$\underset{u^{(i)}}{\operatorname{argmin}} \left\{ \|X^{(i)}u^{(i)} - t\|_2^2 + \lambda_i P_i(u^{(i)}) \right\}$$

Weights & Loadings

- Categorized methods into explicit and implicit latent variable models
- Demonstrated the theoretical and practical robustness of loadings
- Produced meaningful multiview simulated data with known ground truth at the scale of neuroimaging studies

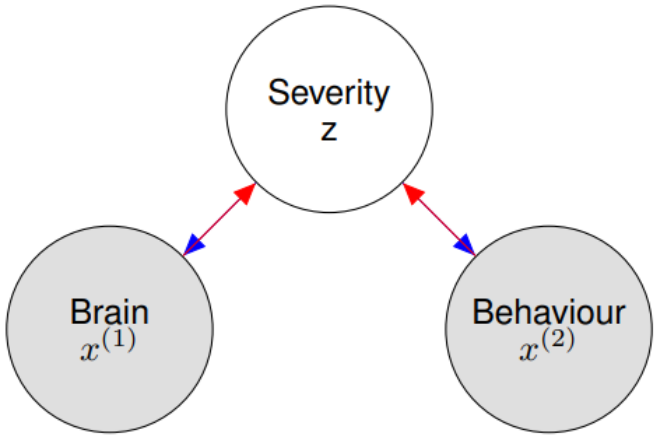


Figure IV.1: ***Forward and Backward Multiview Models:*** The generative/forward and discriminative/backward approaches in CCA.

GEP-EY Algorithm

- Efficient solution for generalized eigenvalue problems

- Enables application of CCA and PLS to large-scale datasets (e.g., UK Biobank)

- Outperforms existing methods in terms of scalability and convergence

Figure V.1: Comparison of the time taken to solve CCA using `eigh` and our CCA-EY method.

Deep Learning Extensions

- DCCA-EY: Novel formulation of Deep CCA for stochastic settings

- SSL-EY: New self-supervised learning method competitive with state-of-the-art with less tuning

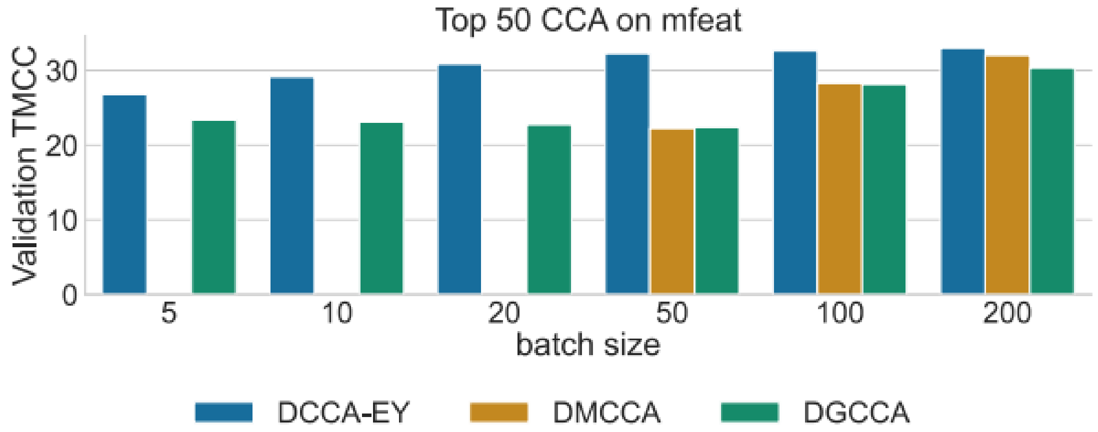
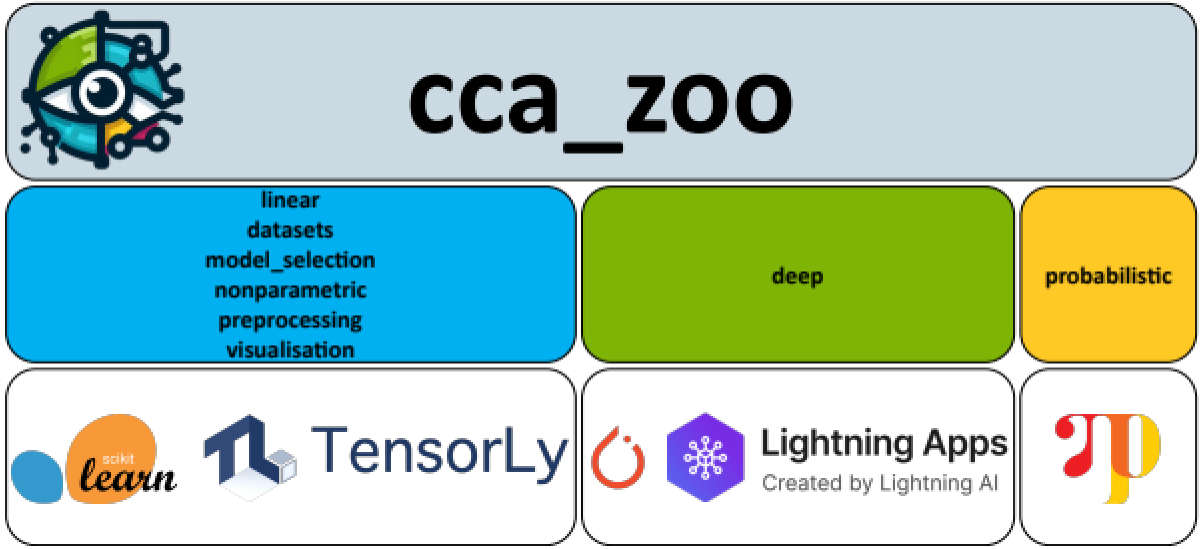


Figure VI.8: Deep Multi-view CCA on mfeat: Comparison across various mini-batch sizes using the Validation TMCC metric.

CCA-Zoo Software

- Comprehensive Python library for multiview learning methods
- Implements various CCA, PLS, and related techniques
- Facilitates broader adoption and innovation in the research community



Impact

- Can run (Multiview) CCA on UK Biobank on your laptop
- Can run Deep (Multiview) CCA on a single GPU
- 190 stars on GitHub from people who have built on my small contributions