DOI: 10.3217/978-3-99161-050-2-215

Population Transformer: Learning Population-level Representations of Neural Activity

G. Chau[1*♦], C. Wang[2♦], S. Talukder[1], V. Subramaniam[2], S. Soedarmadji[1], Y. Yue[1], B. Katz[2], A. Barbu[2]

[1] *California Institute of Technology;* [2] *MIT CSAIL, CBMM*

♦Equal contribution, *Corresponding email: gchau@caltech.edu

Introduction: We present a self-supervised framework that learns population-level codes for arbitrary ensembles of neural recordings such as intracranial electroencephalography (iEEG) at scale. We address key challenges in scaling models with neural time-series data, namely, sparse and variable electrode distribution across subjects and datasets. The Population Transformer (PopT) [1] stacks on top of pretrained representations of temporal activity (e.g. BrainBERT [2]) and enhances downstream decoding by learned aggregation of data channels. The pretrained PopT lowers the amount of data required for downstream decoding experiments, while increasing accuracy, even on held-out subjects and tasks. Beyond decoding, we visualize the learned attention weights to show how they can be used to extract neuroscience insights from large amounts of data. Full paper, code, and weights available at [1].

Material, Methods and Results: **Data**. iEEG data was collected from 10 subjects (total 1,688 electrodes, with a mean of 167 electrodes per subject) who watched 26 movies (19 for pretraining, 7 for downstream decoding) [3]. The downstream decoding tasks were auditory-linguistic features extracted from the movie audio and transcripts. **Model**. PopT consists of a transformer encoder stack, where the input tokens are a [CLS] token and temporal embeddings (E) from an ensemble of channels at time t. The 3D coordinates of each channel are added to each temporal embedding. **Pretraining**. Our self-supervised learning task has two discriminative components: (1) ensemble-wise – the model determines if activities from two channel ensembles occurred consecutively, and (2) channel-wise – the model identifies outlier channels that have been swapped with a different timepoint's activity. **Decoding**. We fine-tune and decode with the intermediate [CLS] embedding attached to a new single linear layer for the specific decoding task. **Results**. We find that fine-tuning and decoding with a pretrained PopT outperforms all baselines of aggregating individual channel embeddings across channel ensemble sizes (Fig 1a). Inspecting the attention weights on our model fine-tuned on decoding periods of Speech vs Non-speech reveals highly responsive language areas such as Wernicke's area (Fig 1b). We further show that pretraining allows one to reach similar performance levels as other models with far fewer samples and compute steps (Fig 1cd).

Conclusion: Our work reveals how beneficial self-supervised pretraining on large datasets can be for downstream decoding, neuroscience discovery, while being sample and compute efficient.

Acknowledgments and Disclosures: We would like to thank our funding sources: Caltech Chen Institute, Caltech Carver Mead New Adventures Fund, and MIT Center for Brains, Minds, and Machines.

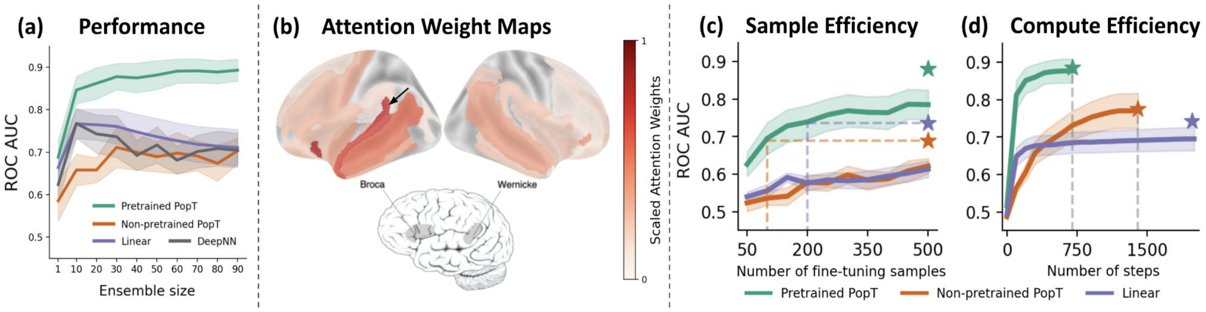

Figure 1: Results on the Speech vs Non-speech decoding task. (a) Downstream decoding performance vs ensemble size, (b) attention weight maps from the [CLS] token after fine-tuning, (c,d) sample & compute efficiency of fine-tuning using pretrained vs non-pretrained models.

References:

[1] Chau G, Wang C, Talukder S, Subramaniam V, Soedarmadji S, Yue Y, Katz B, Barbu A. Population Transformer: Learning Population-level Representations of Neural Activity. arXiv preprint arXiv:2406.03044. 2024 Jun 5.

[2] Wang C, Subramaniam V, Yaari AU, Kreiman G, Katz B, Cases I, Barbu A. BrainBERT: Self-supervised representation learning for intracranial recordings. arXiv preprint arXiv:2302.14367. 2023 Feb 28.

[3] Wang C, Yaari AU, Singh AK, Subramaniam V, Rosenfarb D, DeWitt J, Misra P, Madsen JR, Stone S, Kreiman G, Katz B. Brain Treebank: Large-scale intracranial recordings from naturalistic language stimuli. arXiv preprint arXiv:2411.08343. 2024 Nov 13.

Published by Verlag der Technischen Universität Graz