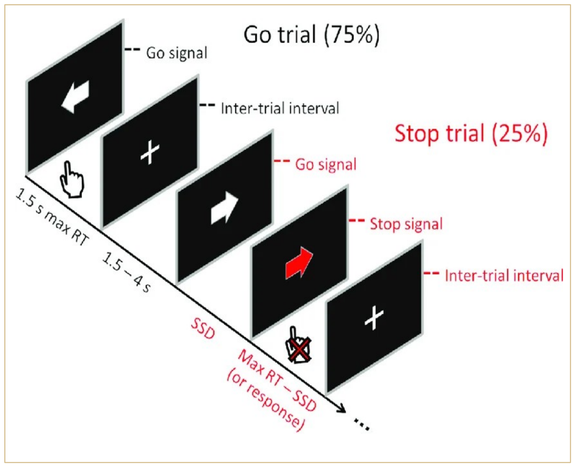
Decoding Inhibitory Control

Deep learning on brain data from Stop-Signal Task (SST)

Author: Maanas Karwa
Mentor: Prof. Yu-Chin Chiu

OVERVIEW

When you "cancel" an action right after initiating it, that's called inhibitory control. We used deep learning on brain signal data from a well-known experiment to identify the response structure of inhibitory control. We achieved high accuracy and a structure matching the empirical response, and also identified left vs right-hand responses along the way.

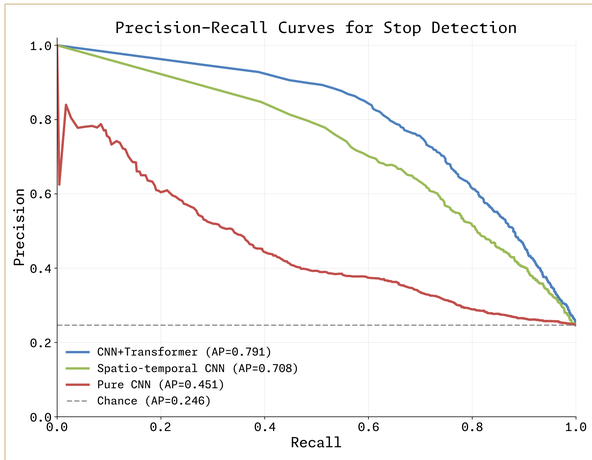


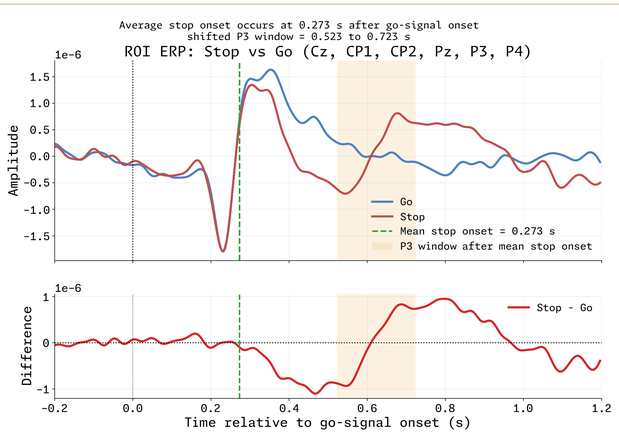
Experiment

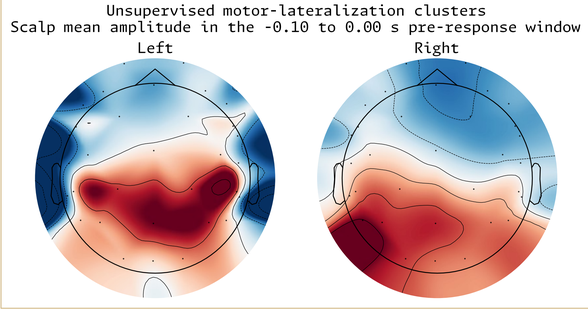
- Subjects **shown left/right arrow**, respond by pressing corresponding key.
- In 25% of trials, show stop signal right after left/right arrow. Subject expected to not press any key; **"stop" trial**.
- **Difficulty of trial** controlled by time between initial signal and stop signal; adjusted using **staircased delay (SSD)**.
- Record subjects' scalp EEG data throughout experiment. EEG has good temporal, low spatial resolution.
- **Expected response: frontocentral P3** around 300 ms after stop signal.
- Plausible explanations for the underlying brain activity: **interactive race model**, proactive control, pause-then-cancel.
- **Why does this experiment matter?** Provides insight into inhibitory control behavior, which is linked to behaviors like ADHD, addiction, compulsive behavior, etc.

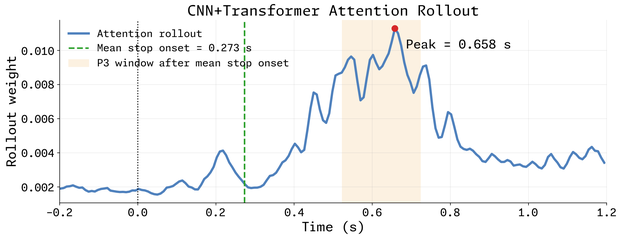
Decoding Methods

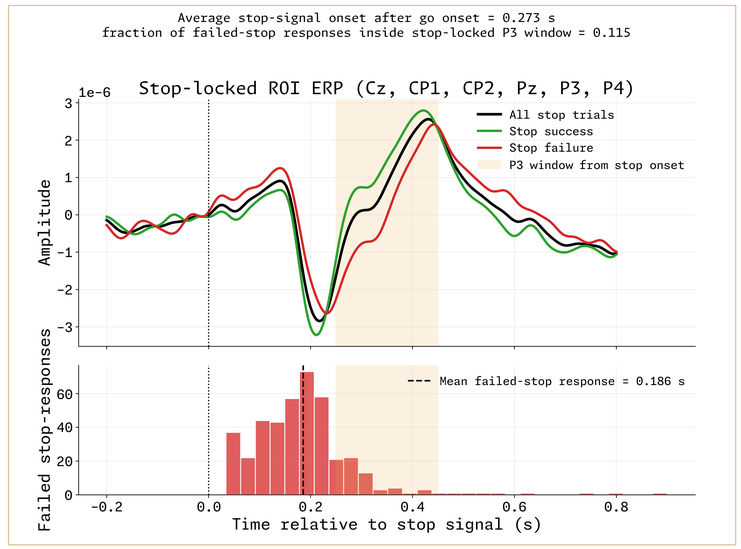
- Trained **supervised deep neural networks**, treating entire spatio-temporal signal for trial as input. Align trials by go signal onset.
- Three models: baseline CNN, spatio-temporal CNN, **CNN+transformer**.
- Evaluation scheme: **LOSO (leave one subject out);** train on all subjects' trials except one subject's, evaluate on that subject.
- CNN-transformer model yielded highest performance: **81% accuracy**, 0.90 ROC-AUC, 0.79 PR-AUC.
- **Interpretability** on CNN+transformer model: data occlusion, attention analysis, relevance propagation.
- **Masked different time windows** to see where the most signal was. Masking **600-800 ms** window causes 20% accuracy drop. Attention heatmaps revealed similar insights on importance of this window.
- Trials are aligned by go signal onset & not stop signal onset; 600 ms after trial onset is ~300 ms after stop signal onset, matching empirical P3 response structure.
- Also ran unsupervised clustering to classify left vs right-arrow go-trials.











Observations

- **Stop-success** and **stop-failure** trials **almost identical** in P3 window. Similar response observed regardless of whether subject actually stopped.
- Ran EEG channel-level ablation, found that most important channels for decoding aren't Cz or FCz (frontocentral channels), instead are more parietally located channels (P7, P8, TP10).

Conclusion & Future Work

- Improved results on decoding SST responses using **deep neural networks** and a **data-driven** approach.
- High accuracy, can be scaled up easily, and requires **no manual intervention** (such as running ICA and finding components by hand).
- Can **decode new subjects** without much calibration, since evaluation scheme (LOSO) explored zero-shot classification on new subjects.
- Future work: **decode other tasks with high accuracy**, including when **neural response is not empirically known.**

References

- Verbruggen et al. (2019) https://elifesciences.org/articles/46323
- Senkowski et al. (2023) https://pmc.ncbi.nlm.nih.gov/articles/PMC11166755
- Logan & Cowan (1984) https://psycnet.apa.org/record/1984-27832-001
- Elchlepp et al. (2016) https://pubmed.ncbi.nlm.nih.gov/26859519
- Diesburg & Wessel (2021) https://pubmed.ncbi.nlm.nih.gov/34293402
- Ceceli et al. (2021) https://pubmed.ncbi.nlm.nih.gov/36396402
- ENIGMA (2025) https://neurips.cc/virtual/2025/loc/san-diego/132694