10733 - Representation and Generation in Neuroscience and AI

Spring 2025, Syllabus

November 14, 2024

Instructors Leila Wehbe, lwehbe@cmu.edu, and Aran Nayebi, anayebi@cs.cmu.edu. Office hours are by appointment or right after class.

Day, time, location: Mondays and Wednesdays at 3:30pm-4:50pm, BH A36.

This PhD-level course explores the intersection of Neuroscience and AI (NeuroAI). The aim of this course is to provide students with the foundational concepts and the methodological tools to perform research in NeuroAI. This course is aimed to help students with a machine learning background or a neuroscience background to form a solid basis in NeuroAI research. The focus will be on acquiring the skills to both build models of brain activity that can perform intelligent tasks/behaviors, and to understand the implications of these models for the brain.

Specifically, this course evaluates the use of new AI methods as models for the brain, and the utility of inspiration from the brain for building better AI. The course has two main components: **(a)** a critical look at the current practice of computational neuroscience through lectures, readings and discussions and **(b)** an active, practical component (though homework and projects) that allows students to form their opinions based not only on readings and ideas imposed upon them, but also through their own experience and understanding. Specifically, the class has a project component, which is expected to be substantial enough that the work could lead to a publication at the end of the semester if successful.

The focus of the class will be on representations in the brain and in AI models, spanning multiple species, specifically in the areas of sensory systems, motor, language, and higher-cognitive areas. Students will learn the most recent approaches for using AI models as models of the brain, along with the arguments for and against these types of approaches. The course will also touch upon generation in both AI models and in the brain, and discuss possibilities of integrating them.

Key Topics: NeuroAI, encoding models, fMRI, MEG, invasive recordings, language, vision, NLP, computer vision, embodied AI.

Learning Objectives: The aim of this class is to enable students to:

- know the state of the art in modeling brain responses using AI and key findings,
- identify key fallacies in modeling and ways to avoid them,
- form computational cognitive neuroscience research questions and a plan to address them,
- have practical experience in modeling brain responses.
- have the foundational skills for research in NeuroAI.

Prerequisites Intermediate statistics and/or machine learning, familiarity with deep learning, linear algebra, programming experience with Python, and vector manipulation as in PyTorch, NumPy, and/or TensorFlow. Basic knowledge in neuroscience is helpful, but not required. For students lacking a background in neuroscience, introductory readings will be provided at the start of the semester. Official course prerequisites include any of the following: 10301, 10315, 10601, 10701, 10715, 10707, 10417, 10617, 10414, 10714, or 11785. If you believe your background from other coursework or experience is equivalent, please contact the instructor. This course also offers an opportunity for students from diverse academic backgrounds to receive mentorship in computational neuroscience. If you have any doubts about your readiness for this class, just ask us!

Credits and assessment structure The course is a 12 unit class. The graded components are as follows: (these percentages might change a bit before the start of the Spring semester)

- 15% Reading. Each week, a paper will be assigned, and the class will begin with a short quiz on that paper (the 10 best scores will be counted towards the grade).
- 15% Participation. You will be graded on your participation in class discussions and in-class debates.
- 20% Homework.
- 50% Class Project. The project is expected to target a new scientific problem or approach an existing problem in a new way. Success in evaluating the project hypothesis is not necessary for a good project. The final deliverable will be a NeurIPS-style paper and a class presentation. There will be multiple milestones for the project, which we will announce in class.

Schedule/topics

- Week 1: Intro to Neuroscience - Big Questions - Earlier hand-tuned approaches - Marr's levels
- Week 2: Methods - Metrics - encoding models - decoding - statistical tests
- Week 3: Vision: Task-optimized models of the visual system (feedforward CNNs/Transformers)
- Week 4: Sensory: Extensions to other modalities (audition, olfaction) - recurrence self-supervised learning (SSL)
- Week 5: Motor: Task-optimized models of the motor system
- Week 6: Foundation Models - Decoding - BCI - end-to-end models
- Week 7: Language: fast overview of existing work - interpretable feature spaces
- Week 8: Language: Using deep language models
- Week 9: Language: Interpreting encoding models
- Week 10: Higher Cognition: memory - reinforcement learning (RL) / events
- Week 11: Higher Cognition: world models
- Week 12: Epistemology - what is this for?
- Week 13: Extra topics - What's next? - Agent-based NeuroAI
- Week 14: Project presentations

Some of the papers we will study

- Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A. and Just, M.A., 2008. Predicting human brain activity associated with the meanings of nouns. science, 320(5880), pp.1191-1195.
- Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B. and Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Current biology, 21(19), pp.1641-1646.
- Wang, J., Miller, K., Marblestone, A., 2020, Where Neuroscience meets AI (And What's in Store for the Future), Neurips Tutorial 2020.
- Naselaris, T., Kay, K.N., Nishimoto, S. and Gallant, J.L., 2011. Encoding and decoding in fMRI. Neuroimage, 56(2), pp.400-410.
- Kriegeskorte, N., Mur, M. and Bandettini, P.A., 2008. Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience, p.4.
- Kornblith, S., Norouzi, M., Lee, H. and Hinton, G., 2019, May. Similarity of neural network representations revisited. In International conference on machine learning (pp. 3519-3529). PMLR.

- Wehbe, Leila, Ashish Vaswani, Kevin Knight, and Tom Mitchell. "Aligning context-based statistical models of language with brain activity during reading." In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 233-243. 2014.
- Yamins, D.L. and DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience, 19(3), pp.356-365.
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F. and Schmidt, K., 2018. Brain-score: Which artificial neural network for object recognition is most brain-like?. BioRxiv, p.407007.
- Maheswaranathan, N.*, McIntosh, L.T.*, Tanaka, H.*, Grant, S.*, Kastner, D.B., Melander, J.B., Nayebi, A., Brezovec, L., Wang, J., Ganguli, S., and Baccus, S.A. 2023. Interpreting the retinal neural code for natural scenes: from computations to neurons. *Neuron*, 111: 2742-2755.
- Nayebi, A., Sagastuy-Brena, J., Bear, D.M., Kar, K., Kubilius, J., Ganguli, S., Sussillo, D., DiCarlo, J.J., and Yamins, D.L.K. 2022. Recurrent connections in the primate ventral visual stream mediate a tradeoff between task performance and network size during core object recognition. *Neural Computation*, 34: 1652-1675.
- Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M.C., DiCarlo, J.J., and Yamins, D.L.K. 2021. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 118.
- Nayebi, A.*, Kong, N.C.L.*, Zhuang, C., Gardner, J.L., Norcia, A.M., and Yamins, D.L.K. 2023. Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *PLOS Computational Biology*, 19: 1-36.
- Nayebi, A., Attinger, A., Campbell, M.G., Hardcastle, K., Low, I.I.C., Mallory, C.S., Mel, G.C., Sorscher, B., Williams, A.H., Ganguli, S., Giocomo, L.M., and Yamins, D.L.K. 2021. Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 34.
- Nayebi, A., Rajalingham, R., Jazayeri, M., and Yang, G.R. 2023. Neural foundations of mental simulation: future prediction of latent representations on dynamic scenes. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 70548–70561.
- Huth, A.G., De Heer, W.A., Griffiths, T.L., Theunissen, F.E. and Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), pp.453-458.
- Guest, O. and Martin, A.E., 2023. On logical inference over brains, behaviour, and artificial neural networks. Computational Brain & Behavior, pp.1-15.
- Schrimpf, M., Blank, I.A., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J.B. and Fedorenko, E., 2021. The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences, 118(45), p.e2105646118.
- St-Yves, G., Allen, E.J., Wu, Y., Kay, K. and Naselaris, T., 2023. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. Nature communications, 14(1), p.3329.
- Toneva, M., Mitchell, T.M. and Wehbe, L., 2022. Combining computational controls with natural text reveals aspects of meaning composition. Nature computational science, 2(11), pp.745-757.
- Reddy, A.J. and Wehbe, L., 2020. Syntactic representations in the human brain: beyond effort-based metrics. BioRxiv, pp.2020-06.
- Makin, J.G., Moses, D.A. and Chang, E.F., 2020. Machine translation of cortical activity to text with an encoder–decoder framework. Nature neuroscience, 23(4), pp.575-582.
- Willett, F.R., Kunz, E.M., Fan, C., Avansino, D.T., Wilson, G.H., Choi, E.Y., Kamdar, F., Glasser, M.F., Hochberg, L.R., Druckmann, S. and Shenoy, K.V., 2023. A high-performance speech neuroprosthesis. Nature, pp.1-6.
- Cross, L., Cockburn, J., Yue, Y. and O'Doherty, J.P., 2021. Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. Neuron, 109(4), pp.724-738.

- Francl, A. and McDermott, J.H., 2022. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. Nature human behaviour, 6(1), pp.111-133.
- Tang, J., LeBel, A., Jain, S. and Huth, A.G., 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. Nature Neuroscience, pp.1-9.
- Caucheteux, C., Gramfort, A. and King, J.R., 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. Nature human behaviour, 7(3), pp.430-441.

Course policies

Attendance This course is a discussion-based course and therefore attendance is essential for benefiting from it and contributing to it. Attendance will not be directly graded but participation will. We understand that some events such as conferences will lead to absence. Please communicate your absence in advance to the course staff. There will be some allowance in the grade for a few absences.

Collaboration Discussion of class material is heavily encouraged. Collaboration in homework assignments is allowed as long as it's properly reported. Project collaboration is expected (projects are done in groups), and is allowed across groups as well.

Academic Integrity We have a zero tolerance policy for violation of class policies. If you are in any doubt in regards to the policy, please clarify with the course staff before proceeding.

- Any deviation from the rules will be dealt with according to the severity of the case. For example: plagiarising content or taking help from someone without acknowledging their contribution on the paper.
- In line with university policy, all instances of cheating/plagiarism will be reported to your academic advisor and the dean of student affairs. See the university policy on academic integrity.

Late Assignments The maximum earnable points for each assignment will drop by 20% per late day.

Additional information

Accommodations for Students with Disabilities

If you have a disability and are registered with the Office of Disability Resources, I encourage you to use their online system to notify me of your accommodations and discuss your needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

Statement of Support for Students' Health & Well-being

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

If you or someone you know is feeling suicidal or in danger of self-harm, call someone immediately, day or night (CaPS: 412-268-2922, Resolve Crisis Network: 888-796-8226). If the situation is life threatening, call the police (On-campus CMU Police: 412-268-2323, Off-campus Police: 911).