Brain2Model Learning: Training sensory and decision models with human neural activity as a teacher

Tomas Gallo Aquino
Columbia University
New York, NY 10027
tg2863@columbia.edu

Victoria Liu
Columbia University
New York, NY 10027
ql2491@columbia.edu

Habiba Azab
Baylor College of Medicine
Houston, TX 77030
azab@bcm.edu

Raissa Mathura
Baylor College of Medicine
Houston, TX 77030
Raissa.Mathura@bcm.edu

Andrew J. Watrous
Baylor College of Medicine
Houston, TX 77030
andrew.watrous@bcm.edu

Eleonora Bartoli
Baylor College of Medicine
Houston, TX 77030
bartoli@bcm.edu

Benjamin Y. Hayden
Baylor College of Medicine
Houston, TX 77030
benjamin.Hayden@bcm.edu

Paul Sajda
Columbia University
New York, NY 10027
ps629@columbia.edu

Sameer A. Sheth*
Baylor College of Medicine
Houston, TX 77030
sameer.Sheth@bcm.edu

Nuttida Rungratsameetaweemana*
Columbia University
New York, NY 10027
nr2869@columbia.edu

Abstract

Cognitive neuroscience shows that the human brain creates low-dimensional, abstract representations for efficient sensorimotor coding. Importantly, the brain can learn these representations with significantly fewer data points and less computational power than artificial models require. We introduce Brain2Model Learning (B2M), a framework where neural activity from human sensory and decision-making guides the training of artificial neural networks, via contrastive learning or latent regression. We provide a proof-of-concept for B2M in memory-based decision-making with a recurrent neural network and scene reconstruction for autonomous driving with a variational autoencoder. Our results show that student networks benefiting from brain-derived guidance can either converge faster, achieve higher predictive accuracy, or both, compared to networks trained in isolation. This indicates that the brain's representations can be useful for artificial learners, facilitating efficient learning of sensorimotor representations, which would be costly or slow through purely artificial training.

1 Introduction

Recent work in cognitive neuroscience reveals that neural populations encode low-dimensional, abstract representations that enable decision-making and sensorimotor behavior [1, 2, 3, 4, 5]. A key objective in the field is to create models that predict neural activity while uncovering the computational principles behind these representations. Advances have been achieved in vision and language, where convolutional and transformer-based models elucidate neural responses in corresponding brain areas

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: BrainBodyFM.

[6, 7, 8, 9, 10]. These findings indicate that artificial neural networks and biological systems may adopt similar representational strategies for perception and decision-making.

Despite advances in large-scale artificial models in vision [11, 12, 13, 14] and decision-making [15, 16, 17], training neural networks to achieve generalizable abstractions typically demands extensive supervision or large foundation models, which are slow and costly to obtain. In contrast, the brain learns complex representations from sparse and ambiguous data, efficiently guiding flexible decision-making and enabling rapid adaptation in novel settings [18, 19]. However, relatively little work has focused on leveraging brain-derived representations to shape the internal learning dynamics and representations of artificial models.

2 Related Work

Leveraging representations from large pre-trained models for efficient training of new task-specific models has been extensively explored in transfer learning literature, achieving success in both supervised [20, 21, 22, 23, 24] and reinforcement learning [25, 26, 27, 28]. Different transfer approaches have been proposed based on the type of knowledge being transferred and the differences between the data available to the source and target models [29, 23, 24, 30, 31], facilitating knowledge transfer across complex decision domains.

Past work has explored using sensory-based brain information to train convolutional neural networks (CNNs) [32, 33, 34, 35]. For instance, in McClure and Kriegeskorte (2016, the authors defined representational neural dissimilarity matrices (DSMs) leveraging data in an MNIST/CIFAR100 visual recognition task and utilized them to transfer information from a teacher model to a student CNN. Despite this relevant work in the field, some outstanding gaps remain regarding the flexibility and applicability of these methods. Utilizing DSMs requires *a priori* defining of discrete trial classes. Additionally, leveraging higher-order cognitive functions as transfer signals, such as memory or decision-making, beyond CNN architectures, remains an underexplored research direction.

3 Methods

To address these gaps, we propose Brain2Model Learning (B2M), a framework to guide model training by encouraging models to achieve similar representations to the human brain. B2M leverages human brain data collected while participants performed similar tasks to artificial models. We guide model training by augmenting standard learning objectives with a loss which incentivizes similarity between model representations and brain representations of reduced dimensionality. Concretely, we add a brain transfer function, which acts as a brain-derived regularization objective:

$$\mathcal{L}_{total} = (1-\alpha)\mathcal{L}_{task} + \alpha\mathcal{L}_{transfer} \tag{1}$$

where $\mathcal{L}_{task}$ is the standard loss function for the artificial learner in a task, $\mathcal{L}_{transfer}$ is the loss function comparing brain and artificial representational similarity, and α is the transfer weight hyperparameter.

We propose two alignment strategies for obtaining $\mathcal{L}_{transfer}$, whose applicability depends on whether artificial models are trained exactly on the same examples as seen by humans, or approximations of these examples: brain contrastive learning, and brain latent transfer (see Appendix, Extended Methods). In short, brain contrastive learning can be applied when there is a direct match between trials experienced by the model and the brain. It relies on a contrastive objective between brain and model representation pairs. Conversely, brain latent transfer can be applied to similar but not necessarily matching trial structures between brains and models, and relies on minimizing a minimum squared error distance between model and brain latent representations.

4 B2M improves RNN learning in memory-based decision making task

4.1 Memory task and brain data

Given novel contexts or a novel instruction set, humans are able to quickly adapt and correctly follow new goals [36, 37, 38, 39]. To understand the role of human brain activity in rapid goal switching,

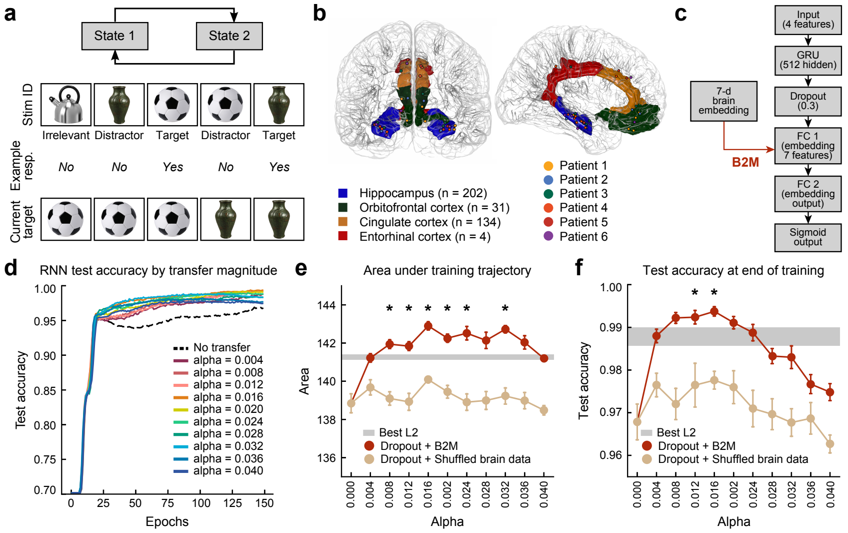

Figure 1: B2M for memory RNN. (a) Task design. The task alternated between two states, indicating which of the stimuli was the current target. In each episode, two stimuli were designated as potential targets, and one stimulus was always irrelevant. Patients were prompted to respond whether the current image was the target. The task alternated to the other state once a target was correctly identified. (b) Single neuron recording across epilepsy patients. (c) RNN for memory-based decision making. The first fully connected layer was jointly embedded with human brains via B2M. (d) Mean RNN test accuracy over epochs for different values of B2M strength α, including no B2M (dashed black curve). (e) Area under training trajectory for different values of B2M strength α. (f) Test accuracy at the end of training for different values of B2M strength α, for true brain embeddings (red) and shuffled brain (beige). Error bars represent standard error of the mean. All shaded areas indicate mean ± s.e.m. for the best run with L2 regularization and $\alpha = 0$. Stars indicate values of $\alpha = 0$ for which standard B2M outperforms both L2 regularization and B2M with shuffled brain data (Mann-Whitney U-Test, one sided, $p < 0.05$).

we utilized data collected from a memory-based goal switching task (Fig. 1a) in which intracranially implanted epilepsy patients (N=17 sessions in 6 patients) had to memorize the current target stimulus and watch a sequence of visual stimuli. If the current target was presented, the correct response would be to press an accept button; otherwise, the correct response would be to press a reject button. As soon as the participant correctly accepted the current target, the target stimulus would shift to another stimulus in the episode, and the current target would alternate between these two until the end of the episode. Episodes had a variable number of steps (stimulus presentations) and 3 possible stimuli: 2 target candidates and one distractor stimulus. In total, 400 steps were given to each participant.

We obtained a total of 371 neurons (31 in orbitofrontal cortex, 202 in hippocampus, 4 in entorhinal cortex, and 134 in cingulate cortex, Fig. 1b), which were co-registered to a shared 7-dimensional latent space via the CEBRA framework [40]. This dimensionality was chosen since it was the minimum available number of neurons in any given session. The neural representation consisted of one embedding per time point per trial for each session. Additional details on spike data preprocessing are provided in the Appendix.

4.2 Model architecture and training

To model an artificial agent performing this task, we implemented a single-layer gated recurrent unit (GRU) network designed for sequence modeling with alignment to brain-derived embeddings (Fig. 1c). Input size was 4, to provide a one-hot coded version of the 3 possible stimuli, plus one element for when no stimulus was on the screen (i.e., pre-stimulus window). The architecture consisted of a GRU layer (hidden size=512, num layers=1), a dropout layer (rate 0.3), transfer temperature = 0.1, a linear embedding projection layer (hidden dimensions: 7), an output layer mapping the embedding to a single continuous output, followed by a scaled sigmoid activation mapping outputs to the range [-1.5, 1.5]. This structure allowed the model to jointly produce outputs **while internally**

encoding representations that could align with neural embeddings, via direct brain contrastive learning. We presented trial sequences to the network in a one-to-one mapping with episodes seen by human participants, by one-hot coding the presented stimulus and showing the context of each episode to the model as a one-hot coded version of the current target stimulus at the beginning of each episode. We then tested the network in 1000 simulated trial sequences previously unseen by human participants. Each configuration (i.e., each value of $\alpha = [0, 0.004, ..., 0.040]$) was trained with 10 random initialization seeds. Models were optimized with Adam (learning rate = 10^{-4}) for 150 epochs with a batch size of 1. Accuracy was measured by comparing predictions to binary targets.

4.3 Results

We measured network performance by accuracy in left-out trials, determined by whether the agent correctly accepted or rejected stimuli in each step (Fig. 1d). To summarize training efficiency, we measured test accuracy after 150 epochs and the area under the test accuracy trajectories across epochs. We performed two additional control analyses. First, we performed B2M with brain data shuffled along the trial axis, to test whether gains derived from the correct structure of brain activity for each memory trial. Second, to compare with a standard regularization objective, we fit the model without B2M but with L2 regularization, for a grid search of 10 values of λ: $[10^{-7}, ..., 10^{-3}]$. We find a range (lowest: 0.008, highest: 0.032) of α values for which the area under B2M training trajectories outperform both shuffled controls and the top-performing L2, $\lambda = 2.1 \cdot 10^{-6}$ (Fig. 1e, Mann-Whitney U-Test, one-sided, $p < 0.05$). The displayed standard error was computed across 10 initialization seeds. Additionally, for $\alpha = [0.012, 0.016]$, B2M test accuracy at the end of training outperformed both top-performing L2 and shuffled controls (Fig. 1f).

5 B2M for naturalistic scene reconstruction in VAE for driving task

5.1 Virtual reality driving task and EEG recordings

We next turn to a naturalistic visual reconstruction task using non-invasive electroencephalogram (EEG) data. We explored B2M in a distinct model architecture, variational autoencoders (VAEs), and applied latent transfer to accommodate the lack of one-to-one trial correspondence between human and artificial data. We tested B2M on a dataset of human participants (N=11 sessions in 9 participants, with written informed consent, IRB approved) performing a vehicle driving task in a virtual reality (VR) environment, while they had 64-channel EEG activity recorded (Fig. 2a). This task has been previously shown to elicit activity in dorsolateral prefrontal cortex and anterior cingulate that are predictive of driving behavior [41].

One possible task to model in this setting is the sensorimotor processing that occurs with the goal of safely driving through a city. As such, we adapt a previously established [42] VAE trained with the specific purpose of creating a visual embedding of urban scenes, to be later passed onto a reinforcement learning agent (Fig. 2b). The VAE has already been described elsewhere [42] (see Appendix for implementation and extended methods).

For B2M, we used temporally aligned brain signals (EEG) and downsampled VR video frames (8Hz), reshaped to 160×80 pixels, totaling 133120 frames across all participants. **Since the VAE train set (simulated environment) and the human train set in VR are not exactly matching, we employed Brain Latent Transfer for B2M**. To present video/EEG data to the model, we treated each temporally aligned video/EEG example pair as a unique input, batched them (batch size: 256), and presented one batch at a time, in tandem with each training image batch.

Network hyperparameters were kept the same as in the original implementation [42], except for embedding dimensionality, which we changed to 64, in alignment with EEG inputs. Training was repeated across 10 random initializations for each $\alpha = [0, 0.02, ..., 0.2]$ value, yielding a total of 110 full training runs. Each model was trained for 100 epochs, using the Adam optimizer (learning rate = 10^{-4}). After each epoch, models were evaluated on a held-out validation set of visual scenes.

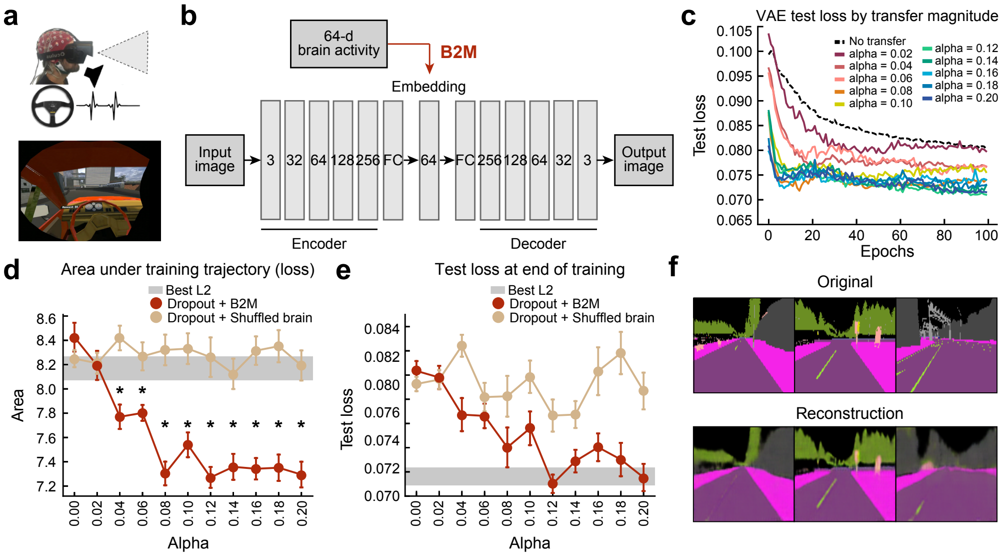

Figure 2: B2M for driving scene reconstruction with VAE. (a) Top: Participants steered a vehicle in VR with simultaneous EEG recordings; Bottom: VR driving scene example. (b) The VAE for scene reconstruction consisted of an encoder-embedding-decoder architecture. The embedding layer was jointly embedded with EEG data from human participants driving in VR. (c) Mean test loss for different values of B2M strength α, including no B2M (dashed black curve). (d) Area under training trajectory for different values of B2M α. (e) Test accuracy at the end of training for different values of B2M α, for true brain embeddings (red) and shuffled brain (beige). Error bars represent standard error of the mean. All shaded areas indicate mean ± s.e.m. for the best run with L2 regularization and $\alpha = 0$. Stars indicate values of $\alpha = 0$ for which standard B2M outperforms both L2 regularization and B2M with shuffled brain data (Mann-Whitney U-Test, one sided, $p < 0.05$). (f) Examples of original and reconstructed scenes, produced with B2M ($\alpha = 0.1$).

5.2 Results

In a small percentage of runs, we observed diverging test loss ($loss > 1$) that did not recover over the course of training (4 out of 110 runs, 3.6%). We excluded these runs from the subsequent analysis and visualization.

We measured network performance by reconstruction loss in left-out naturalistic driving scenes, determined by mean square error between target image and source image (Fig. 2c). To summarize training efficiency, we measured test accuracy after 150 epochs and the area under the test accuracy trajectories across epochs. As before, controlled for shuffled brain data and L2 regularization (λ: $[10^{-5}, ..., 10^{-2}]$). We found a range of $\alpha = [0.04, ..., 0.20]$ for which B2M outperforms both L2 regularization and shuffled brain in area under the loss trajectory (Mann-Whitney U-Test, one sided, $p < 0.05$, Fig. 2d). However, the top-performing L2 regularization ($\lambda = 10^{-5}$) matched B2M performance at the end of training (Fig. 2e). This indicates that while it is possible for standard regularization to achieve comparable accuracy, B2M converges more efficiently. The displayed standard error was computed across 10 initialization seeds. For visualization purposes, we include examples of original and reconstructed driving scenes with brain transfer, for $\alpha = 0.1$ (Fig. 2e).

6 Discussion

In this work, we demonstrate, as a proof of principle, that low-dimensional brain representations can be leveraged to guide neural network training in complex cognitive tasks involving sensory processing, memory, and flexible decision-making. This approach raises several promising avenues for future exploration. First, future work may systematically investigate the scalability of brain-guided training: what is the sample efficiency of brain priors, and how does performance scale with the dimensionality of neural embeddings?

Second, an open question remains regarding the transferability and utility of different brain recording modalities. Given their varying signal-to-noise ratios and spatiotemporal resolution, it is critical to benchmark the relative effectiveness of invasive (e.g., depth recordings) versus non-invasive (e.g., fMRI, EEG) data in shaping model representations across cognitive domains.

Third, this work requires a deeper investigation into the geometry of the brain-induced low-dimensional spaces, to determine exactly which representations are imparted to models. By characterizing how artificial models restructure their internal representations in response to brain-derived constraints, we can gain insight into both the alignment between biological and artificial computation and the types of inductive biases these embeddings confer.

We anticipate B2M can scale by leveraging increasingly more abundant neural datasets and emerging neural foundation models that distill common representational principles across individuals and tasks. In this way, B2M could become an alignment step that brings brain-derived structure into new models without requiring new recordings each time.

More broadly, this work could open a path toward more data-efficient training, where functionally aligned neural recordings guide the acquisition of generalizable representations in artificial agents. By continuing to bridge neuroscience and machine learning through shared principles of representation and computation, we may build novel strategies for systems that learn uniquely human skills.

References

[1] David Sussillo, Mark M Churchland, Matthew T Kaufman, and Krishna V Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, 18(7):1025–1033, 2015.

[2] Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, 32:148–155, 2015.

[3] Juan A Gallego, Matthew G Perich, Lee E Miller, and Sara A Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017.

[4] Silvia Bernardi, Marcus K Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C Daniel Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967, 2020.

[5] W Jeffrey Johnston, Justin M Fine, Seng Bum Michael Yoo, R Becket Ebitz, and Benjamin Y Hayden. Semi-orthogonal subspaces for value mediate a binding and generalization trade-off. *Nature Neuroscience*, 27(11):2218–2230, 2024.

[6] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[7] Jerry Tang, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in neural information processing systems*, 36:29654–29666, 2023.

[8] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

[9] Eshed Margalit, Hyodong Lee, Dawn Finzi, James J DiCarlo, Kalanit Grill-Spector, and Daniel LK Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451, 2024.

[10] Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, 87:101244, 2024.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[16] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[17] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[18] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[19] Brenden M Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*, 2019.

[20] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI proceedings-international joint conference on artificial intelligence*, volume 22, page 1541. Citeseer, 2011.

[21] Lixin Duan, Dong Xu, and Ivor Tsang. Learning with augmented features for heterogeneous domain adaptation. *arXiv preprint arXiv:1206.4660*, 2012.

[22] Joey Zhou, Sinno Pan, Ivor Tsang, and Yan Yan. Hybrid heterogeneous transfer learning through deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[23] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3:1–40, 2016.

[24] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 270–279. Springer, 2018.

[25] Matthew E Taylor, Peter Stone, et al. Transfer learning via inter-task mappings for temporal difference learning. *Journal of Machine Learning Research*, 8(9), 2007.

[26] Beomjoon Kim, Amir-massoud Farahmand, Joelle Pineau, and Doina Precup. Learning from limited demonstrations. *Advances in Neural Information Processing Systems*, 26, 2013.

[27] Wojciech M Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1331–1340. PMLR, 2019.

[28] Zhuangdi Zhu, Kaixiang Lin, Anil K Jain, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13344–13362, 2023.

[29] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[30] Paul Sajda, Sameer Saproo, Victor Shih, Sonakshi Bose Roy, and David Jangraw. Systems and methods for deep reinforcement learning using a brain-artificial intelligence interface, September 12 2023. US Patent 11,755,108.

[31] Ilia Sucholutsky and Tom Griffiths. Alignment with human representations supports robust few-shot learning. *Advances in Neural Information Processing Systems*, 36:73464–73479, 2023.

[32] Patrick McClure and Nikolaus Kriegeskorte. Representational distance learning for deep neural networks. *Frontiers in Computational Neuroscience*, 10:131, 2016.

[33] Ruth C Fong, Walter J Scheirer, and David D Cox. Using human brain activity to guide machine learning. *Scientific Reports*, 8(1):5397, 2018.

[34] Satoshi Nishida, Yusuke Nakano, Antoine Blanc, Naoya Maeda, Masataka Kado, and Shinji Nishimoto. Brain-mediated transfer learning of convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5281–5288, 2020.

[35] Zhe Li, Wieland Brendel, Edgar Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian Sinz, Zachary Pitkow, and Andreas Tolias. Learning from brains how to regularize machines. *Advances in neural information processing systems*, 32, 2019.

[36] Elliot H Smith, Guillermo Horga, Mark J Yates, Charles B Mikell, Garrett P Banks, Yagna J Pathak, Catherine A Schevon, Guy M McKhann, Benjamin Y Hayden, Matthew M Botvinick, et al. Widespread temporal coding of cognitive control in the human prefrontal cortex. *Nature Neuroscience*, 22(11):1883–1891, 2019.

[37] Margaret M Henderson, John T Serences, and Nuttida Rungratsameetaweemana. Dynamic categorization rules alter representations in human visual cortex. *Nature Communications*, 16:3459, 2025.

[38] Kalman A Katlowitz, Shraddha Shah, Melissa C Franch, Joshua Adkinson, James L Belanger, Raissa K Mathura, Domokos Meszéna, Elizabeth A Mickiewicz, Matthew McGinley, William Muñoz, et al. Learning and language in the unconscious human hippocampus. *bioRxiv*, pages 2025–04, 2025.

[39] Tomas G Aquino, Jeffrey Cockburn, Adam N Mamelak, Ueli Rutishauser, and John P O'Doherty. Neurons in human pre-supplementary motor area encode key computations for value-based choice. *Nature Human Behaviour*, 7(6):970–985, 2023.

[40] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, pages 1–9, 2023.

[41] Sharath Koorathota, Jia Li Ma, Josef Faller, Linbi Hong, Pawan Lapborisuth, and Paul Sajda. Pupil-linked arousal correlates with neural activity prior to sensorimotor decisions. *Journal of Neural Engineering*, 20(6):066031, 2023.

[42] Raza Asad Idrees. Implementing a deep reinforcement learning model for autonomous driving. Bachelor's thesis, Budapest University of Technology and Economics, 2022.

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[44] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[45] R Quian Quiroga, Zoltan Nadasdy, and Yoram Ben-Shaul. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation*, 16(8):1661–1687, 2004.

[46] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

Technical Appendix

Extended Methods

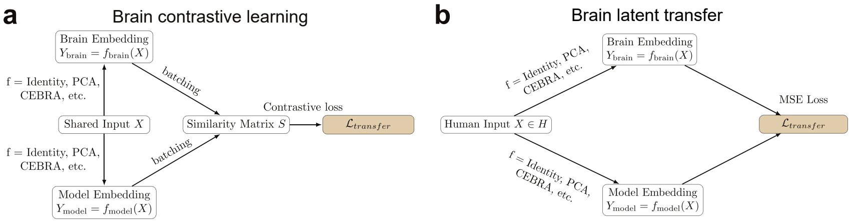
a
Brain contrastive learning
Brain Embedding
$Y_{\text{brain}} = f_{\text{brain}}(X)$
f = Identity, PCA, CEBRA, etc.
batching
Shared Input X
Similarity Matrix S
Contrastive loss
$\mathcal{L}_{transfer}$
f = Identity, PCA, CEBRA, etc.
batching
Model Embedding
$Y_{\text{model}} = f_{\text{model}}(X)$
b
Brain latent transfer
f = Identity, PCA, CEBRA, etc.
Brain Embedding
$Y_{\text{brain}} = f_{\text{brain}}(X)$
MSE Loss
Human Input $X \in H$
$\mathcal{L}_{transfer}$
f = Identity, PCA, CEBRA, etc.
Model Embedding
$Y_{\text{model}} = f_{\text{model}}(X)$

Figure 3: Obtaining $\mathcal{L}_{transfer}$ in B2M. (a) Brain contrastive learning. A shared input is transformed into both brain embeddings and artificial model embeddings of equal dimensionality. From input batches containing several examples, we compute a similarity matrix between brain and model embeddings of different inputs, which is utilized to obtain a transfer loss value. We designate neural-model embedding pairs computed from the same input example as positive pairs, while pairs computed from separate examples are designated negative pairs. (b) Brain latent transfer learning. An input X from the human train set $\mathcal{H}$ is used to produce brain embeddings and artificial embeddings, produced from a model trained with an artificial train set $\mathcal{A}$. These embeddings are subsequently compared with a mean square error distance to obtain a transfer loss.

Brain Contrastive Loss

One goal of B2M is to maximize the mutual information between neural representations and their counterparts in artificial models. This aims to assist artificial learning by encouraging models to find advantageous sensorimotor representational subspaces that might have been acquired by the human brain over the course of task learning and, ultimately, evolution. We propose to achieve this by adapting contrastive learning via an InfoNCE loss framework, such as the one adopted in SimCLR, [43, 44], which aims to maximize mutual information between similar data points (Fig. 3a).

Given a sensory or contextual input $X \in \mathcal{I}$, with $X \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_n}$, where $\mathcal{I}$ is a train set presented to both humans and artificial models, we propose aligning brain and artificial representations of X by maximizing the mutual information between embeddings $Y_{brain} \in \mathbb{R}^E$ and $Y_{model} \in \mathbb{R}^E$, where E is the embedding dimensionality. These embeddings are produced by non-linear transfer functions such that $Y_{brain} = f_{brain}(X)$ and $Y_{model} = f_{model}(X)$, indicating compressed neural and artificial representations of sensorimotor inputs, respectively. Any applicable dimensionality reduction method can be utilized as f, or even the identity function, as long as the dimensionality between brain and model embeddings is matched.

For this, we define a batch of b examples $X \in \mathcal{I}$, $[X_1, X_2, ..., X_b]$, and their respective neural and artificial embeddings $B_{neural} = [Y_{(neural,1)}, Y_{(neural,2)}, ..., Y_{(neural,b)}]$ and $B_{model} = [Y_{(model,1)}, Y_{(model,2)}, ..., Y_{(model,b)}]$. From these, given a temperature hyperparameter τ, we obtain a similarity matrix $\mathcal{S} = \frac{1}{\tau}(B_{neural} \times B_{model}^T)$, which represents the similarity between neural and artificial example pairs, including both pairs obtained from the same example X_i, but also pairs obtained from different examples. Then, we define positive contrastive pairs $Y_{(neural,i)}$ and $Y_{(model,i)}$, and negative contrastive pairs $Y_{(neural,i)}$ and $Y_{(model,j)}$ for all $i \neq j$. With these, we define, for a given anchor example i, with $\mathcal{L}_{transfer} = \sum_i \mathcal{L}_{transfer,i}$:

$$\mathcal{L}_{transfer,i} = -\log \frac{exp(S_{i,i})}{exp(S_{i,i}) + \sum_{i \neq j} exp(S_{i,j})} \tag{2}$$

Brain Latent Transfer Loss

In decision-making tasks, future states are often dependent on decisions made in past episodes, as well as the outcomes that occurred in them. For this reason, it is potentially challenging to assemble a train set of episodes and environment states for an artificial agent that will exactly match the train sets experienced by human participants in a task, given that the agent is free to act differently from their human counterparts during training. To circumvent this, instead of exact episode matching, we propose matching brain embeddings obtained during exposure to an input X to the artificial embeddings obtained during exposure to the same input X, as long as X is an approximation of the examples contained in the artificial train set (Fig. 3b).

Concretely, given a sensory or contextual input $X \in \mathcal{H}$, with $X \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_n}$, where $\mathcal{H}$ is the train set presented to humans, we assume brain activity produces an embedding Y_{brain} via a non-linear transfer function f_{brain}, such that $Y_{brain} = f_{brain}(X)$, with $Y_{brain} \in \mathbb{R}^E$, in which E is the embedding dimension. Additionally, we assume the artificial model learns from examples $X' \in \mathcal{A}$, $X' \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_n}$, where $\mathcal{A}$ is the train set presented to the artificial model, related but not necessarily equal to $\mathcal{H}$. Throughout its learning process, the artificial model finds a non-linear transfer function f_{model} which produces its own embedding $Y'_{model} = f_{model}(X')$, with $Y'_{model} \in \mathbb{R}^E$.

Then, we propose to achieve brain-to-model transfer by performing a latent transfer between brain and model embeddings. Concretely, for an input $X \in \mathcal{H}$ previously presented to humans, we obtain its embedding produced by the artificial model $Y_{model} = f_{model}(X)$ and minimize its mean square error distance to its known corresponding brain embedding $Y_{brain} = f_{brain}(X)$:

$$\mathcal{L}_{transfer} = \frac{1}{E} \sum_{i=1}^{E} (Y_{model,i} - Y_{brain,i})^2 \tag{3}$$

Memory Task Extended Methods

Invasive neural data was acquired with IRB approval and informed consent, utilizing the Blackrock system. Standard spike sorting was performed with WaveClus [45]. We applied the following pre-processing steps to temporally align neural data and RNN activity: (1) we aligned all spikes to stimulus presentation and created two time periods: pre-stimulus (-1s to 0, resolution: 10ms) and post-stimulus (0s to reaction time). (2) We normalized post-stimulus spikes by reaction time to always fit into a 100-element rate vector, and concatenated pre-stimulus and post-stimulus normalized spikes into a spiking rate vector. (3) We filtered spike rates by convolving the rate vector of each step with a causal exponential filter (kernel: 20 zeros followed by $e^{-0.5x}$, $x = [0, 0.5, 1, ..., 9.5, 10]$).

To learn shared representations across multiple behavioral sessions, we used the Contrastive Embedding by Relative Arrangement (CEBRA) framework [40] (Apache License). CEBRA is a self-supervised learning algorithm that maps high-dimensional neural activity to a low-dimensional space by leveraging temporal structure and optional contextual supervision. Below, we describe the full procedure used to generate multi-session embeddings from neural recordings across 17 behavioral sessions.

We aggregated neural activity from all neurons in the 17 distinct recording sessions. Each session contained neural population firing rates stored as 3D arrays with shape $(N_{steps}, N_{times}, N_{neurons})$, where N_{steps} represents the number of unique stimuli presented to a participant, considering each stimulus constitutes a step. For each session, data were reshaped into 2D arrays of shape $(N_{steps} \times N_{times}, N_{neurons})$ and paired with time labels repeated across trials. These reshaped arrays represent temporally ordered sequences of neural activity and serve as the input to CEBRA.

We instantiated a CEBRA model configured for multi-session contrastive learning. The following hyperparameters were used: model architecture: offset10-model, batch size: 512, learning rate: $3 \cdot 10^{-4}$, temperature mode: auto with a minimum temperature of 0.1, embedding dimensionality: 7, maximum training iterations: 15,000, distance metric: cosine similarity, supervision: conditional sampling based on relative time (i.e., time-delta), utilizing time within step as a supervising feature.

Driving Task Extended Methods

Healthy adults (N=11 sessions in 9 participants, with written informed consent, IRB approved) completed a boundary-avoidance driving task inside a virtual-city environment rendered through an HTC Vive Pro Eye headset. Seated at a Logitech G920 steering wheel with accelerator and brake pedals, participants piloted a virtual car while continuous “fog” opacity dynamically modulated visual uncertainty on a trial-by-trial staircase. Crashes with road boundaries incurred point penalties to encourage timely, accurate steering. Scalp EEG was recorded throughout with a 64-channel BioSemi ActiveTwo system (Ag/AgCl active electrodes, international 10–20 system, $impedances < 50k\Omega$) at 2048Hz; a lossless screen-capture of the VR scene was recorded via the Unity engine; steering-wheel position, pedal inputs, and headset-embedded eye-tracking data were time-synchronized with the EEG stream for later source and connectivity analyses.

Additionally, the VAE model utilized in for visual scene reconstruction in this task consists of an encoder-embedding-decoder architecture (Fig. 2b), that reconstructs input target urban scenes into matching outputs. These scenes were previously obtained in the CARLA driving simulator environment [46] and made publicly available [42], with 12000 preset training images and 2000 test images, which are examples of driving scenes. We changed the embedding dimensionality from the original VAE to 64 dimensions, to directly match the 64 channels recorded in EEG sessions. The encoder and decoder each contain 5 convolutional/leaky ReLU layers and one fully connected layer (encoder dimensions: $[3, 32, 64, 128, 256]$, decoder dimensions: $[256, 128, 64, 32, 3]$, fully connected dimensions: 1024.). We also adapted the original model to include dropouts ($p = 0.1$) in each convolutional and fully connected layer. The last convolutional encoder layer also contains a batch normalization step. Training visual scenes to be reconstructed were presented in batches (batch size: 32) of 160×80 pixel images.

Limitations

The observed improvements in learning performance could still be partially attributable to a regularization effect introduced by structured noise in the neural data, rather than reflecting brain information transfer alone. Future work should explore more refined controlled ablation studies (e.g., using permuted or synthetic neural data with matched noise statistics to disentangle representational transfer from implicit regularization effects).

In a small subset of training runs, Brain Latent Transfer produced instability, occasionally resulting in catastrophic model divergence. This could potentially arise when the neural training data and model training data differ substantially in input distribution, potentially exposing the model to conflicting or out-of-distribution (OOD) latent signals. Addressing this challenge may require more robust alignment techniques, such as OOD detection or reweighting schemes to reconcile mismatched training domains. Additionally, systematic benchmarking across architectures and input domains will be essential for assessing B2M's scalability.

Furthermore, the benefits of B2M were tested on RNNs and VAEs. It remains unclear whether these findings extend to other architectures, such as reinforcement learning agents, or task modalities beyond vision and memory-based decision making. Further work must be done to establish extended generalization. Systematic benchmarking across architectures and input domains will be essential for assessing B2M's scalability.

Finally, while B2M improves testing performance, it is not yet determined what portions of information are being transferred and whether the brain-derived embeddings encode high-level abstractions, low-level features, or task-specific biases. Developing tools to interpret and visualize the aligned latent spaces will be useful for understanding the semantic content of the transfer signal.

Ethical Considerations

Despite its promise, this line of work introduces important ethical considerations. Critically, the use of human brain data for training artificial models raises issues of privacy, consent, and data stewardship. All neural data used in this study were anonymized and collected under IRB approval with informed consent, and care must be taken that these standards are upheld even if the economic viability of B2M in large-scale projects is demonstrated.

Additionally, techniques that align AI systems with neural representations could, in theory, be misused in contexts such as surveillance or cognitive-behavioral manipulation. Although we do not release pretrained models, any future release will include usage terms to prevent misuse in the context of human participants protection. We encourage the community to proactively discuss ethical governance frameworks and emphasize full transparency, human participants protection, consent, and user autonomy in downstream applications. Additionally, care must be taken that human participants are compensated fairly for the data they provide for building better models, which could ultimately provide significant economic potential at scale.

Computer Resources

All experiments were performed on a Lambda Labs server, with the following characteristics: 192 CPUs, AMD Ryzen Threadripper PRO 7995WX 96-Cores, 5390MHz (maximum), graphics card: NVIDIA Corporation AD102GL [RTX 6000 Ada Generation], 512GB RAM, 18 TB HD storage.

On this machine, the 220 memory task runs (RNN: B2M and noise) took 14.16h in total, whereas the 220 driving task runs (VAE: B2M and noise) took 113.59h in total.

Human Participants

For the memory-based decision-making task, we collected intracranial data from 6 human epilepsy patients, undergoing seizure monitoring prior to surgery. Patients read the following text on the screen, substituting target-stim1 and target-stim2 for each episode with the actual target stimuli for that episode:

A scientist is studying different patterns. Your job is to help him. Objects will appear one after another, and he wants you to take a picture when you spot a particular pattern in their sequence. (For example, a Vase followed by a Flower.) We'll always tell you what pattern to look for. Every few objects, there will be a new pattern to watch for. When you spot the pattern, press your photograph button to take a picture. Otherwise, press the appropriate button to skip that object. We'll tell you which button is which on each trial. The objects can appear in any order, including several of the same type in a row. A new object will appear every two seconds or so. Press 'LEFT' to take a picture, and press 'RIGHT' to skip this object. Wait for the first [target-stim1], then take a picture of it. Then wait for the first [target-stim2], and take a picture of it. Alternate taking pictures of one [target-stim1] and one [target-stim2].

For the VR driving task, we collected EEG data from 9 healthy human participants. They had prior driving experience, normal or corrected-to-normal vision, and reported they were not prone to motion sickness. Participants were verbally instructed to drive a car along a road in VR for a fixed amount of time, avoiding collisions. They were told that collisions would deduct an amount of money from their total reward bonus, which was displayed on the car's dashboard. They could accelerate, brake, and steer with realistic input controls. Participants were compensated at a rate of 20 USD/h for 3 hours. EEG contact positioning was the same for all participants, as displayed in Fig. 4

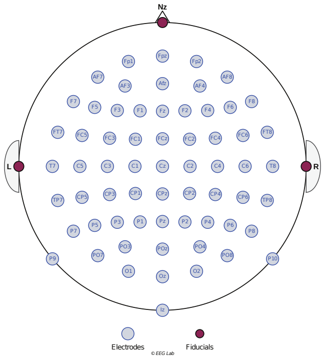

Figure 4: EEG standard contact map. All participants in the VR driving task underwent EEG recording, with electrodes positioned along the same standard contact grid. Nz indicates the direction of the front of the head.