ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Predictive Coding Dynamics Enhance Model-Brain Similarity

Manshan Guo[1,2], Michael Samjatin[1], Bhavin Choksi[1],
Sari Saba-Sadiya[1], Radoslaw M. Cichy[2], and Gemma Roig[1]

1- Goethe Universität, Frankfurt 2- Freie Universität, Berlin

Abstract. Predictive coding–a popular theory in neuroscience–has garnered significant attention in the machine learning community aiming to incorporate brain-inspired components in neural networks. While various proposals have demonstrated the ability of predictive dynamics to render robustness and entail human-like perception of illusions, it remains unclear if they improve the alignment between brain and artificial representations. Here, we systematically investigate the conditions under which brain-inspired modifications in predictive processing improve alignment between model and neural representations in the brain. Our results reveal that the feedback component significantly increases similarity between model representations and those found in higher-level visual brain areas, especially when processing complex visual scenes.

1 Introduction and Related Works

Predictive coding theory in neuroscience posits that the brain continuously predicts and refines an internal model of the world, allowing it to anticipate sensory inputs based on prior experiences and current context[1]. This popular hierarchical process integrates top-down and bottom-up information and has inspired machine learning researchers looking to develop visual models with brain-like performance and characteristics [2, 3, 4]. The incorporation of predictive coding was demonstrated to improve the performance of neural networks, even promoting human-like behavior [5].

These findings, in the context of using AI models to explain brain data, raise an obvious question: *Does the incorporation of these bio-inspired dynamics improve the alignment between the model and brain representations?*. Prior studies have investigated this question[6, 7]. Using representational similarity analysis, (i.e., RSA), [7] demonstrated that predictive models trained in an unsupervised manner explained brain visual representations better than their supervised counterparts. [6] showed that unsupervised predictive coding networks [4] with higher RSA scores are correlated with better performance in next frame prediction and object matching. While these works showed promising trends, they test the alignment only at the level of representational geometry, use smaller brain datasets, and use network architectures where tweaking the impact of feedback and feedforward components is relatively difficult.

In this work, we address these issues by quantifying the brain-similarity of a novel and flexible predictive coding network with a large-scale fMRI dataset. Specifically, using the model made available by the *predify* python package [2]–a

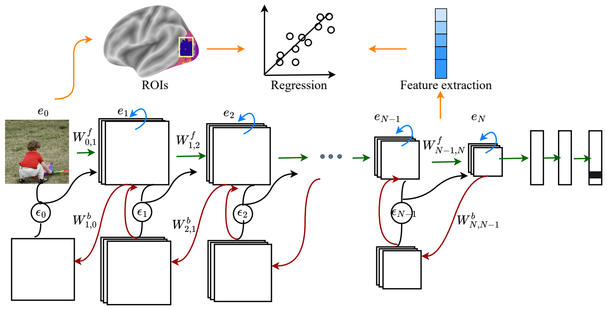

Fig. 1: We add predictive coding dynamics to VGG16 and extract activations from five layers. For each timestep, we extract layer activations and train ridge regressions to predict fMRI data from different ROIs

VGG16 performing predictive coding—we modeled representations in different brain regions of interest (ROIs) obtained from the Natural Scenes Dataset (NSD) [8]. We report results demonstrating improved ability to predict fMRI data over recurrent (time-)steps compared to a feedforward VGG16[1].

2 Methods

Predictive VGG16 (PVGG16) Predify integrates predictive coding dynamics into models by augmenting a pretrained classification model (here VGG16), with feedback connections trained in an unsupervised fashion. As illustrated in Figure 1, these feedback connections divide the model into a sequence of consecutive predictive coding (PCoder) modules (here 5 PCoders in PVGG16). The output of each PCoder, $e_N(t)$, over timesteps, incorporates input from feed forward, backward, and recurrent connections using the following equation:

$$e_N(t+1) = \overset{\text{feedforward}}{\beta_N[W^f_{N-1,N}e_{N-1}(t+1)]} + \overset{\text{memory}}{(1-\beta_N-\lambda_N)e_N(t)} - \underset{\text{error}}{\alpha_N\nabla\epsilon_{N-1}(t)} + \underset{\text{feedback predictions}}{\lambda_N[W^b_{N+1,N}e_{N+1}(t)]}$$

where β_N, λ_N $(0 \le \beta_N + \lambda_N \le 1)$, α_N are layer-specific coefficients controlling the weight of the feedforward, feedback and error-correction signals respectively. ϵ_{N-1} is the Mean Square Error between $e_{N-1}(t)$ and $W^b_{N,N-1}e_N(t)$ (the top-down prediction) at time step t and is used to train the feedback connections in-line with the predictive coding principle. We refer the readers to the original paper for additional details.

Natural Scenes Dataset The Natural Scenes Dataset (NSD)[8] contains high quality fMRI responses to natural scenes from the Common Object in Context

[1] All code to reproduce the results available at : `https://github.com/cvai-roig-lab/Predictive-Coding-Dynamics-Enhance-Model-Brain-Similarity`

(COCO) database [9]. Out of eight, we used a subset of five subjects (subjects 1,2,4,5,7) with high signal-to-noise ratio (SNR) [8]. The average number of trials number were 9629.

Feature Extraction and Brain Alignments We use the 'Predify'[2] and 'Net2Brain' toolboxes [10] to extract features from the five PCoders in PVGG16 and their corresponding baseline layers in VGG16. To estimate the alignment between model activations and fMRI responses, we used a linear regressor regularized with a ridge penalty as implemented in 'Net2Brain', which is a standard approach for voxel-wise encoding models [10]. The ridge regressor was trained on the training split (80 % of the total data) using nested cross-validation.

To align model activations, we divide with different brain ROIs into two broad groups – (i) *early visual cortex* (ECV) including the early- and mid-level visual cortex (*V1*, *V2*, *V3*, *hV4*), and (ii) *the high level visual cortex* (HCV), consisting of body-selective (*EBA*, *FBA-2*), face-selective (*OFA*, *FFA-1*, *FFA-2*, *mTL-faces*, *aTL-faces*), place-selective (*OPA*, *PPA*, *RSC*), and word-selective (*OWFA*, *VWFA-1*, *VWFA-2*, *mfs-words*) regions.

The trained ridge regressors are evaluated on held-out data (test set). The fMRI encoding accuracy is calculated by measuring the similarity between predicted and actual signals with Pearson correlation. The final alignment scores are the average across subjects and ROIs. To visualize directly the significance of predictive coding to augmenting brain similarity, the final alignment scores (mean Pearson coefficients) of PVGG16 are normalized by their respective feedforward counterparts. Normalized accuracy above 1 indicates that predictive coding promotes brain-model similarity.

Image Complexity Metrics We investigate the impact of predictive coding on complex visual scene processing. Following previous research, we used entropy to measure the complexity of test images [11]. We separately evaluated brain alignment for the images with the top and bottom 10% complexity values.

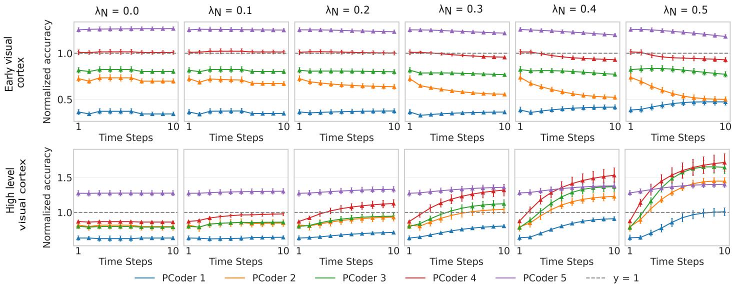

Fig. 2: Mean normalized encoding accuracy in early visual, and high level visual, cortex by PVGG16 ($\beta_N = 0.3$, $\lambda_N \in [0.1, 0.2, 0.3, 0.4, 0.5]$). The error bars represent the standard error and ▲ denotes that normalized encoding accuracy is significantly different(p $\leq$ 0.05) from the baseline.

3 Experiments and Results

Predictive coding selectively improves alignment in higher visual regions: We first measured the brain alignment by using the default parameters provided in predify ($\beta_N = 0.3, \lambda_N = 0.3, \alpha_N = 0.1$; Figure2 column 4). We observed that, in the early visual cortex, on average, predictive dynamics did not alter, or marginally decreased, the brain-model alignment across all the pcoders. In contrast, predictive dynamics consistently improved the alignment in the high level visual cortex, indicating that the integration of feedback information selectively helps the alignment for higher brain regions. Interestingly, for both early visual and high level visual cortex, PCoder5 consistently ranked above its feedforward counterpart. The slower changes in this PCoder can be attributed to the lack of any feedback (resulting in a strong memory component) to this layer. But its relatively higher alignment with the brain data is intriguing; previous works have reported that very late layers in networks show a decreased alignment with the brain data, possibly due to a preferential shift towards categorical representations necessary for the final discriminative task[12]. Our results indicate that the generative predictive coding iterations might alleviate this discrepancy, further aligning the models in processing hierarchy with the brain data.

Stronger feedback improves high-level visual cortex alignmnent: We hypothesized that the observed alignment can be further improved by increasing the feedback in the networks. To test this, we systematically increased the λ_N from 0.3 to 0.5 by keeping other coefficients fixed and measured the brain-model alignment. We observed that stronger feedback significantly and consistently improved the alignment in the high level visual cortex over timesteps. Indeed, as explained earlier, PCoder5 showed little change in its alignment (though surprisingly in the upward direction). The same tests in the early visual cortex showed an opposite trend, where higher feedback decreased the normalized accuracy values over timesteps.

Complex Scene Perception requires Predictive Coding Dynamics: Feedback plays a crucial role in visual scene processing, particularly in integrating contextual information and resolving ambiguities in complex visual stimuli. Prior research demonstrate that more complex scenes elicit increased feedback activity in the brain, highlighting the importance of top-down modulation during object detection in natural scenes [13]. This feedback supports the processing of intricate scene components by enhancing representational integration in higher visual areas. Building on these insights, we investigated whether feedback connections between PCoders could enable PVGG16 to exhibit brain-like processing of complex scenes. We selected models trained with $\lambda_N = 0.5$ which yielded highest similarity scores for late PCoders in Figure 2, and re-analyzed their predictive accuracy on test images of high and low complexity, respectively.

Predictive coding dynamics consistently improved alignment with brain representations in the high level, but not the early visual cortex when processing

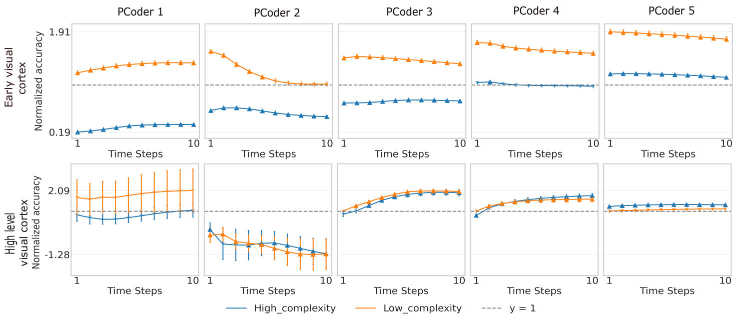

Fig. 3: Normalized mean encoding accuracy of the early visual and high level visual cortex for PVGG16 ($\beta_N = 0.3$, $\lambda_N = 0.5$), on subsets of images with low and high complexity values, estimated using entropy. The error bars represent the standard error and ▲ denotes that normalized encoding accuracy is significantly different(p $\leq$ 0.05) from the baseline.

high complexity scenes. Additionally, in the ECV, predictive coding improved brain-similarity for low complexity images, while having no (or even negative) influence on brain similarity for complex images (figure 3). This indicates that for complex images feedback dynamics transform model representations to be more akin to those employed in late-stage visual processing. These results support the notion that feedback mechanisms are essential for neural networks to correlate components within complex scenes and extract meaningful semantic information, paralleling the dynamics observed in biological vision systems.

4 Discussion and Conclusion

The results of our experiments reveal that predictive coding dynamics enhance alignment with representations in late, rather than early, brain regions. Specifically, our findings suggest that feedback signals–designed to simulate top-down cognitive processes–are the primary drivers of this improved alignment. This indicates that the incorporation of feedback information transforms the model representations to become similar to those found in higher brain regions responsible for complex scene processing. In contrast, early visual areas show weaker alignment with predictive coding models. This may indicate that the feedback information, as modeled in the current form, is less crucial when modeling the initial stages of visual processing.

In conclusion, our findings show that predictive coding dynamics enhance alignment with cognitive representations with strong integration of top-down signals. However, the effects of predictive coding may be counterproductive when trying to simulate representations that are mostly the results of feedforward processing. To fully leverage the potential of predictive coding for achieving human-like visual processing, it is essential to carefully consider the specific cognitive processes leveraged by humans to complete this task.

Finally, the current methods remain limited by the fact that the fMRI brain data used here have low temporal resolution, incompletely constraining the temporally active models. Future work using brain data with high temporal resolution, such as acquired using EEG/MEG or ECoG are a promising direction.

Acknowledgements

This project was funded by the German Research Foundation (DFG) - DFG Research Unit FOR 5368 (GR) awarded to GR, Deutsche Forschungsgemeinschaft (DFG; CI241/1-1, CI241/3-1, and CI241/7-1) awarded to RMC, and a European Research Council (ERC) starting grant (ERC-2018-STG 803370) awarded to RMC. We are grateful for access to the computing facilities of the Center for Scientific Computing at Goethe University and Freie universität Berlin. M. Guo is supported by a PhD stipend from the China Scholarship Council (CSC).

References

[1] Rajesh Rao and Dana Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2, 1999.

[2] Bhavin Choksi, Milad Mozafari, Callum Biggs O'May, Benjamin Ador, Andrea Alamia, and Rufin VanRullen. Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. In *Neural Information Processing Systems*, 2021.

[3] Haiguang Wen, Kuan Han, Junxing Shi, Yizhen Zhang, Eugenio Culurciello, and Zhongming Liu. Deep predictive coding network for object recognition. In *International conference on machine learning*, pages 5266–5275. PMLR, 2018.

[4] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[5] Zhaoyang Pang, Callum Biggs OMay, Bhavin Choksi, and Rufin VanRullen. Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *Neural Networks*, 144:164–175, 2021.

[6] Nathaniel Blanchard, Jeffery Kinnison, Brandon RichardWebster, Pouya Bashivan, and Walter Scheirer. A neurobiological evaluation metric for neural network model search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[7] Marcio Fonseca. Unsupervised predictive coding models may explain visual brain representation. *arXiv preprint arXiv:1907.00441*, 2019.

[8] Emily Allen, Ghislain St-Yves, Yihan Wu, Jesse Breedlove, Jacob Prince, Logan Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 2022.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference of Computer Vision*, 2014.

[10] Domenic Bersch, Martina Vilas, Sari Sadiya, Timothy Schauml, Kshitij Dwivedi, Radoslaw Cichy, and Gemma Roig. Net2brain: A toolbox to compare artificial vision models with human brain responses. *Frontiers in Neuroinformatics*, 2025.

[11] Jaume Rigau, Miquel Feixas, and Mateu Sbert. An Information-Theoretic Framework for Image Complexity. In Laszlo Neumann, Mateu Sbert, Bruce Gooch, and Werner Purgathofer, editors, *Computational Aesthetics in Graphics, Visualization and Imaging*. The Eurographics Association, 2005.

[12] Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human it well, after training and fitting. *BioRxiv*, pages 2020–05, 2020.

[13] Iris IA Groen, Sara Jahfari, Noor Seijdel, Sennay Ghebreab, Victor AF Lamme, and H Steven Scholte. Scene complexity modulates degree of feedback activity during object detection in natural scenes. *PLoS computational biology*, 14(12):e1006690, 2018.