Learning to predict visual brain activity by predicting future sensory states

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep predictive coding networks are neuroscience-inspired unsupervised learning models that learn to predict future sensory states. We build upon the *PredNet* implementation by Lotter, Kreiman, and Cox (2016) to investigate if predictive coding representations are useful to predict brain activity in the visual cortex. We use representational similarity analysis (RSA) to compare PredNet representations to functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) data from the Algonauts Project (Cichy et al., 2019). In contrast to previous findings in the literature (Khaligh-Razavi & Kriegeskorte, 2014), we report empirical data suggesting that unsupervised models trained to predict frames of videos without further fine-tuning may outperform supervised image classification baselines in terms of correlation to spatial (fMRI) and temporal (MEG) data.

1 Introduction

Currently, convolutional neural networks trained on image recognition tasks are the best performing models to account for brain activity during visual object recognition (Schrimpf et al., 2018). The performance of such supervised models on recent benchmarks led to the idea that supervised learning may be a requirement to explain visual cortex activity, especially in higher cortical areas (Khaligh-Razavi & Kriegeskorte, 2014). In this work, we report experimental results suggesting that predictive coding models trained on unlabeled video data may outperform supervised baselines, yielding internal representations with higher correlation to representation dissimilarity matrices (RDM) (Kriegeskorte, Mur, & Bandettini, 2008) obtained from human fMRI and MEG data.

This paper summarizes our two main contributions. First, we find that a small predictive coding model trained on 4 hours of unlabeled video outperforms supervised baselines (He, Zhang, Ren, & Sun, 2016; Krizhevsky, Sutskever, & Hinton, 2012) in terms of correlation to RDMs across two fMRI and MEG datasets (Cichy et al., 2019). Moreover, we show that as train the model on additional videos from the Moments in Time dataset (Monfort et al., 2018), the model internal representations become more similar to brain activity. Thus, our results suggest a promising avenue to build better models of human visual processing by scaling the training to bigger models and larger amounts of unlabeled video data.

Our work share similar methods with the *human-model similarity (HMS)* proposed by Blanchard, Kinnison, RichardWebster, Bashivan, and Scheirer (2019) as they use the same PredNet implementation and representational similarity analysis to measure the correlation between predictive coding representations and human fMRI. However, their work addresses the use of RDMs to inform hyperparameter search and model generalization. We focus on the comparison of the unsupervised predictive coding model versus supervised baselines and the effect of scaling unsupervised training on the correlation scores.

Implementation details and code can be provided on request.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

2 Methods

Predictive coding networks We build on the *PredNet* implementation by Lotter et al. (2016), which was shown to perform well on unsupervised learning tasks using video data. Inspired by the predictive coding theory (Friston & Kiebel, 2009), their model relies on the idea that to predict the next video frame, a model needs to capture latent structure that explains the image sequences. The PredNet architecture consists of recurrent convolutional layers (Xingjian et al., 2015) that propagate bottom-up prediction errors which are used by the upper-level layers to generate new predictions. For implementation details, please refer to the PredNet architecture description by Lotter et al. (2016).

Unsupervised training We evaluate predictive coding models trained on different quantities of unlabeled videos. The main idea is that the more data we use to train the model, the more "common sense" it should get about how events unfold in the world and, as a consequence, it should be better at disentangling latent explanatory factors. Using as starting point a PredNet pre-trained on the KITTI dataset (Geiger, Lenz, Stiller, & Urtasun, 2013), we further train the model with unlabeled videos from the Moments in Time dataset (Monfort et al., 2018), a large-scale activity recognition dataset. Additionally, we report correlation scores for a PredNet with random weights and a version trained from scratch using just videos from the Moments in Time dataset.

The predictive coding model is trained in an unsupervised way to predict the next frames using a top-down generative model. The errors between predictions and the actual frames are propagated bottom-up to update the prior for new predictions. In terms of architecture, we follow the same hyperparameter settings used in the original PredNet implementation proposed by Lotter et al. (2016), with four modules (PredNet-4) consisting of 3×3 convolutional layers with 3, 48, 96, and 192 filters and input frames with dimensions 128×160. We also train a larger 5-layer model (PredNet-5) with 3, 48, 96, 192, and 192 filters and input frames with a higher resolution of 256×256 pixels. The videos are subsampled at ten frames per second, and the network input is a sequence of ten frames for which the model generates ten frame predictions.

Brain data Feature extraction and evaluation are performed on two datasets consisting of 92 silhouette object images Cichy, Pantazis, and Oliva (2014) and 118 object images on natural background (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016) as well brain data recorded using fMRI (EVC and IT) and MEG (early and late in time) in response to visual stimuli. These datasets are curated to include representative visual categories such as animate/inanimate, human/non-human, and face/non-face, which are categorical divisions observed in high-level cortical areas in primates (Kriegeskorte, Mur, Ruff, et al., 2008).

Feature extraction Each layer l and timestep t of the PredNet model has representation units R_l^t, which are extracted as features to be compared to brain data. No further preprocessing or dimensionality reduction is performed. Since the Algonauts datasets consists of images and not videos, we repeat each image ten times to make the input compatible with the PredNet architecture.

Extracted features are then transformed to representational dissimilarity matrices (RDM) as described by Kriegeskorte, Mur, and Bandettini (2008). RDM serves as a common space to compare representations of different models and capture the dissimilarity (1 minus Pearson correlation) of internal representations generated for each pair of images in the dataset. To create RDMs from features, we use the Python implementation from the development kit provided by Cichy et al. (2019).

Evaluation The resulting RDMs for all model variants are compared to the RDMs from different brain regions, namely fMRI data from the early visual cortex (EVC) and the inferior temporal cortex (IT), and also MEG data for early and late stages of visual processing. The similarity of RDMs is computed in terms of Spearman's correlation, as defined by Kriegeskorte, Mur, and Bandettini (2008) and implemented in the Algonauts Python development kit.

3 Results and Discussion

Predictive coding yields representations similar to brain data The predictive coding model used in our experiments has many significant architectural differences comparing to most image recognition feedforward models. First, it is designed to model temporal sensory data, such as

Table 1: Correlation scores for model layers with best average correlation. Evaluation is performed on the **92-images dataset** in terms of models RDM noise normalized squared correlation (Spearman) to fMRI and MEG RDMs. All PredNet features are from recurrent timestep 10, and all results are statistically significant ($p < 0.05$). The noise ceiling corresponds to a 100% score.

		Noise Normalized R^2 (%)			
		fMRI		MEG	
Model Name	Pre-training	EVC	IT	Early	Late
AlexNet (conv2)	ImageNet	16.79	7.11	4.88	15.09
AlexNet (conv4)	ImageNet	13.98	10.66	2.72	16.68
ResNet-50 (block2)	ImageNet	10.80	1.28	12.46	8.33
ResNet-50 (block4)	ImageNet	12.45	10.02	5.29	12.60
PredNet-4 (layer 1)	random weights	5.25	1.09	2.33	3.81
PredNet-4 (layer 4)	KITTI (1h)	39.90	13.13	29.52	20.16
PredNet-4 (layer 4)	KITTI + Moments (4h)	47.03	**15.27**	32.22	16.78
PredNet-5 (layer 3)	Moments (6h)	**51.48**	14.52	**40.35**	**22.89**

Table 2: Correlation scores for model layers with best average correlation. Evaluation is performed on the **118-images dataset** in terms of models RDM noise normalized squared correlation (Spearman) to fMRI and MEG RDMs. All PredNet features are from recurrent timestep 10, and all results are statistically significant ($p < 0.05$). The noise ceiling corresponds to a 100% score.

		Noise Normalized R^2 (%)			
		fMRI		MEG	
Model Name	Pre-training	EVC	IT	Early	Late
AlexNet (conv2)	ImageNet	11.20	1.91	20.18	8.45
ResNet-50 (block2)	ImageNet	9.93	4.38	8.64	2.53
PredNet-4 (layer 1)	random weights	8.60	**9.77**	2.42	3.00
PredNet-4 (layer 3)	KITTI (1h)	17.14	6.93	20.15	7.74
PredNet-4 (layer 4)	KITTI + Moments (4h)	22.93	4.94	44.48	**18.24**
PredNet-5 (layer 3)	Moments (6h)	**24.73**	5.20	**44.68**	15.20

sequences of video frames or audio spectrograms. Second, the model does not rely on labeled data and learn just by minimizing the error between predicted and actual sensory data. Most importantly, it is trained on a dataset of videos portraying activities such as *cooking*, *walking*, and *dancing*, which differ significantly from the carefully curated 92 and 118-images datasets used to measure brain activity. Despite those differences, the model trained on just 1 hour of videos from the KITTI dataset and 6.9 million parameters outperforms an AlexNet model (≈ 60 million parameters, pre-trained on ImageNet) in terms of correlation to for both fMRI and MEG data (92-images dataset, Table 1) and fMRI data (118-images dataset, Table 2).

Scaling unsupervised learning improves correlation with brain data The noise normalized correlation scores (Tables 1 and 2) shows that learning more about how events unfold over time improves correlation scores. Correlation to brain representation continues to improve as we train the PredNet-4 and PredNet-5 with up to 6 hours of videos from the Moments in Time dataset, except for IT scores, which do not exhibit a clear trend. Since PredNet-5 input has higher dimensionality (256×256) and more layers, more experiments would be needed to separate the impact of the amount of training data and model size on the correlation scores.

4 Final remarks and future directions

In this paper, we report empirical results suggesting that unsupervised predictive coding models may be a better starting point than supervised image recognition models to explain visual brain data. While our results are still far from the noise ceilings, there is still much room for improvement by

scaling the model architecture and training data (AlexNet has about nine times more parameters than the PredNet-4 model).

Despite not being able to fully capture the semantic information found in the inferior temporal (IT) area of the cortex such as the clustering of animate/inanimate and face/face-non visual stimuli, we stress that we do not perform any fine-tuning using brain data or recombination of features from different layers as done in previous research (Khaligh-Razavi & Kriegeskorte, 2014). To address this limitation, we aim to investigate methods to use RDMs as a training signal and jointly optimize prediction loss and correlation to the brain representations. This approach would be an extension of the *Representation Distance Learning* (RDL) method proposed by McClure and Kriegeskorte (2016), adapted to the predictive coding architecture.

References

Blanchard, N., Kinnison, J., RichardWebster, B., Bashivan, P., & Scheirer, W. J. (2019). A neurobiological evaluation metric for neural network model search. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5404–5413).

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, *6*, 27755.

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature neuroscience*, *17*(3), 455.

Cichy, R. M., Roig, G., Andonian, A., Dwivedi, K., Lahner, B., Lascelles, A., ... Oliva, A. (2019). The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. *arXiv preprint arXiv:1905.05675*.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, *32*(11), 1231–1237.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. Retrieved from `https://www.frontiersin.org/article/10.3389/neuro.06.004.2008` doi: 10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*.

McClure, P., & Kriegeskorte, N. (2016). Representational distance learning for deep neural networks. *Frontiers in computational neuroscience*, *10*, 131.

Monfort, M., Zhou, B., Bargal, S. A., Andonian, A., Yan, T., Ramakrishnan, K., ... others (2018). Moments in time dataset: one million videos for event understanding. *arXiv preprint arXiv:1801.03150*.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... others (2018). Brain-score: which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* (pp. 802–810).