A Deep Generative Model Imitating Predictive Coding in Human Brain

Abstract

In recent years, the development of deep generative models for prediction has been attracting attention. In our study, we focus on predictive coding, a concept from the neuroscience literature that hypothesizes the brain is constantly making predictions of sensory input, and attempt to develop a human brain-like prediction model. PredNet is a deep learning model based on the concept of predictive coding, however, it can predict the next image with the same prediction interval, but not with any intervals. On the other hand, Temporal Differential Variational Auto-Encoder (TD-VAE) can predict the next images with any prediction intervals, although it is not a model reflecting human brain function. In this work, we develop a new human brain-like prediction model by unifying PredNet and TD-VAE, combining both predictive coding and flexible interval prediction abilities in one single model. Through experiments on the KITTI Vision Benchmark, we confirmed that our proposed model can predict the next images correctly with flexible prediction intervals. We also investigated the correlation between the feature values of representation layers in the model architecture and human brain activity data evoked under natural video stimulation.

1 Introduction

Recently, machine learning and deep learning technologies have been widely used in the development of fundamental technologies for decoding human brain and brain-machine interface (BMI). Predictive coding is a theoretical hypothesis that the brain continually generates predictions of sensory input and compares to its actual input. PredNet [8] is a neural network model based on the concept of predictive coding. It learns to predict the next future image frames in a video sequence by mimicking the process hypothesized to happen in the cerebral cortex. The model consists of hierarchical neural networks with dynamic processes of both bottom-up and top-down. Although PredNet predicts the next scenes sequentially, it is not able to generate predictions with flexible prediction intervals. On the other hand, Temporal Differential Variational Auto-Encoder (TD-VAE) [3] is a deep generative model that can predict the next scene with flexible prediction intervals by introducing a state variable that represents beliefs in its model. In this research, we combine the strengths of PredNet, a model that imitates the prediction mechanism that happens in the cerebral cortex, and TD-VAE, a model that can generate predictions with flexible prediction intervals. We built a model that mimics the prediction function at the cerebral cortex and also enables it to make predictions with flexible prediction intervals. We evaluate the effectiveness of our proposed model by conducting two experiments: one is to confirm whether our proposed model can predict the next scenes with flexible intervals with the KITTI Vision Benchmark dataset [2], and the other is to investigate whether there appears predictive coding function in our model by observing the correlation between the actual brain activity data, e.g., fMRI data, with the values of the hidden states of our model.

Preprint. Under review at the 2nd Shared Visual Representations in Human and Machine Intelligence (SVRHM) Workshop at NeurIPS 2020.

2 PredNet vs. TD-VAE

The overview of PredNet is illustrated in Figure 1. PredNet consists of a hierarchical representation of a series of repeating stacked modules. Each module has four internal components: an input convolutional layer, a recurrent convolutional representation layer, a convolutional prediction layer, and an error representation. In each module, the representation layer represents a state for prediction and the input layer deals with the input information. The prediction layer generates an internal prediction state and the error layer takes the difference between the prediction and the input and outputs an error representation. In PredNet, the top-down and bottom-up process functions to generate predictions – predictions generated at the upper layers of the model are passed to the lower layers through the representation modules, on the contrary, the errors detected in the lower layers are passed to the upper layers. This mechanism corresponds to the behaviors of the generalized equation of state, therefore, it is possible to make predictions.

Temporal Differential Variational Auto-Encoder (TD-VAE) [3] is a deep generative model that incorporates belief state into Partially Observable Markov Decision Process (POMDP) [5]. In TD-VAE, Long-Short Term Memory (LSTM) [4], which deals with the information sequentially provided, helps to achieve prediction tasks. TD-VAE also observes input data and predicts the behavior of the system about the forecast target by using state variables called 'belief' in the background. The whole flow of TD-VAE observes input information and propagates them to the belief layer that the behavior of TD-VAE's system is incorporated, and does an inference of prediction at an arbitrary time via prediction module. The inference information is generated through the decoder module as predictions at an arbitrary time. The outline of TD-VAE is illustrated in Figure 2.

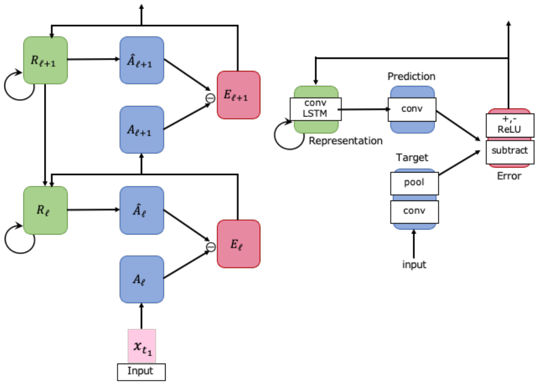


Figure 1: PredNet (left:hierachical representation of a series of repeating stacked modules; right: content of a module)

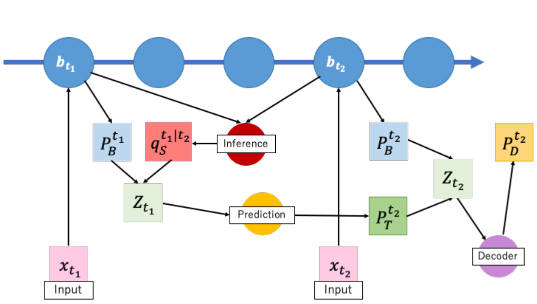
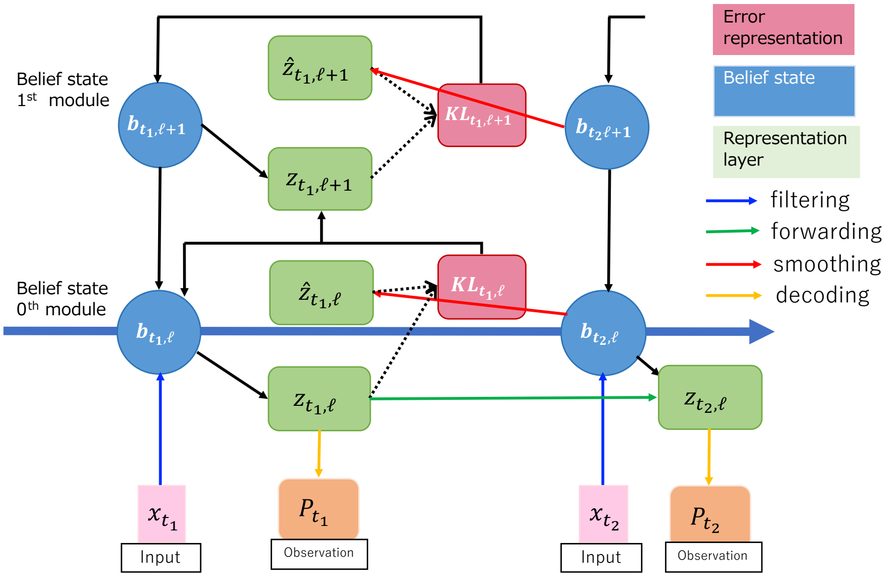


Figure 2: TD-VAE

3 Predictive coding model with flexible prediction interval ability

In this research, we constructed a new deep learning model for the next image prediction by integrating the function of TD-VAE with that of PredNet. The overview of our proposed model is illustrated in Figure 3. Let us assume the current time is t_1 and the future time we would like to predict is t_2. The input information to our model is sequential data and is directly input to the layer called belief state. The belief state which corresponds to a layer defined as the behavior of the model can be objectively observed, and also can be regarded as it stores all the information observed from the past to the present. First, $z_{t_1,l}$, the l-th layer's a latent variable at time t_1, is generated by $b_{t_1,l}$, the belief state of the same layer at time t_1. Here, $z_{t_1,\ell}$ indicates the behaviors of an observation target at time t_1. $z_{t_2,\ell}$ is a latent variable of the same layer at time t_2. $z_{t_2,\ell}$ is predicted by the transition probability of $p(z_{t_2,\ell}|z_{t_1,\ell})$. Because this model learns a series of states in advance, it can use a smoothing distribution $p(\hat{z}_{t_1,\ell}|b_{t_2,l})$ to estimate $\hat{z}_{t_1,\ell}$. This enables the model possible to predict the next states with various prediction intervals, unlike PredNet. Because the inferred $z_{t_1,\ell}$ and the inferred $\hat{z}_{t_1,\ell}$ should be the same at time t_1, As well as the l-th layer, the same procedure happens on the $l+1$-th layer, and then the error signal between $z_{t_1,\ell+1}$ and $\hat{z}_{t_1,\ell+1}$ is propagated to the upper-layer. In this way, the learning of the model is performed by reducing the error signals iteratively propagated to the upper layers. Then the latent state z_{t_2} at time t_2 is predicted from the state $z_{t_1,\ell}$ at time t_1, through the prediction layer. The output image is generated from the state z_{t_2} through the decoder layer. This mechanism to propagate error from the bottom-up layer and propagate the inferred value to the belief state from the upper layer imitates the bottom-up and top-down processing performed in the human cortex. This model predicts the latent variable to generate an observation object. This means that the belief state learns how the latent variable generates observation objects through learning sequential input information at the training phase and then becomes able to predict latent variables to generate observation objects at the inference phase.



$$L = KL\left(z_{t_1,\ell}, \hat{z}_{t_1,\ell}\right) + KL\left(z_{t_1,\ell+1}, \hat{z}_{t_1,\ell+1}\right) + \log P(z_{t_2,\ell}|b_{t_2,\ell})$$
$$-\log P\left(z_{t_2,\ell}|z_{t_1,\ell}\right) - \log P(P_{t_2}|z_{t_2,\ell})$$

Figure 3: Our model combining both predictive coding and flexible interval prediction abilities

4 Experiment

We conducted two experiments to compare the predicted images between PredNet and our model, and also to confirm whether there is a correlation between the brain activity data observed while human subjects were watching a movie and the internal state represented at the latent layers of the prediction models, i.e., PredNet and our proposed model. Table 1 shows the experimental settings for training PredNet and our proposed model.

Table 1: Architectural configurations of PredNet and our model.

model name	PredNet	Our model
# Layers	4	4
Size of convolutional filter	3 × 3 (for all convolutions)	3 × 3 (for all convolutions)
# Channels	From lower module, 3, 48, 96, 192	From lower module, 3, 48, 96, 192
Optimization algorithm	Adam [6] (α=0.001,β_1=0.9,β_2=0.999)	Adam [6] (α=0.0005,β_1=0.9,β_2=0.9995)
learning rate decay	After the midpoint α=0.0001	After the midpoint α=0.0001

4.1 Experiment 1: confirmation of the prediction ability

We verified the prediction ability between PredNet and our model especially in terms of flexibility for prediction intervals. In the experiment, we used the dataset of the KITTI Vision Benchmark Suite [2] for training and testing the models. This dataset is the world's largest automotive benchmark dataset and was developed by Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago. It contains videos taken from a car, which are prepared as 10 frames per second, i.e., one image every 0.1 seconds, for each of the five driving situation categories. The information about 15 items such as object types, the orientation of the object seen from a camera. etc. is annotated to each image. In the implementation of PredNet, we used PyTorch framework [10] and the hyperparameters adopted by the prior study [8].

The prediction intervals were set as 0.1 seconds. Figure 4 shows the results of PredNet and our model comparing with the original movie.

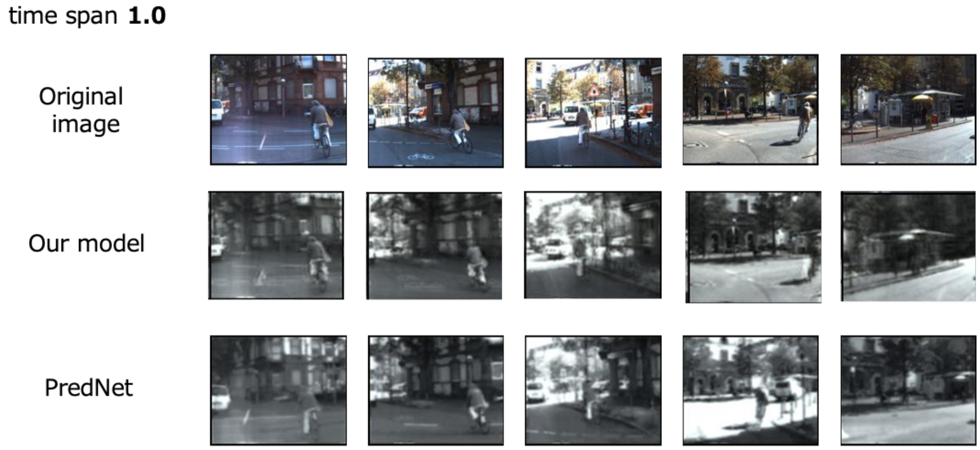


Figure 4: An example of prediction result between PredNet and our model comparing with the scenes of the original movie

The experimental results show that the accuracy of the prediction by PredNet decreases as the prediction interval increases. On the other hand, we could confirm that our model could predict almost the same scene as the one of the original movie.

4.2 Experiment 2: Correlation between model representations and brain activities

We trained PredNet and our model on the same natural movies used in [9]. As preprocessing for the movies of this dataset to be used in the models, we extracted a series of still images from a video at 10fps and downsampled the size of the images to 160×120 pixels. To confirm the correlation between the internal states of the representation layers of both models and human brain activities, we obtained the correlation coefficient between the brain activity data and each internal state of the representation layer, R_0,R_1,R_2, and R_3 in Figure 5 by estimating the value of the internal state of each layer by means of ridge regression – due to resource limitations, we did not estimate the value of layer R_1, since the internal state has very high dimensions (230,400). As the brain activity data we used in the experiments, we employed the image stimuli observed by fMRI from the subjects who were watching a movie. The data has 65,665 dimensions corresponding to the cerebral cortex part among all 96×96×72 observed voxels by fMRI. We constructed a regression model with the data of 4,497 representation-brain activity pairs by means of ridge regression, and evaluated the model on 300 pairs. We evaluated the obtained model in terms of mean square error. An overview of the process of regressing the value of the internal state of each layer from brain activity data for PredNet and our model is illustrated in Figure 5 and Figure 6, respectively.

Table 2 shows the correlation coefficients between the actual values of the inner representations acquired during a prediction task and the corresponding values of the representation layers of R_0, R_2, and R_3, estimated from brain activities using ridge regression. As α, a weight for regularization term in ridge regression, we tried values in the range of $\{0.50,\ 1.0 \times 10^3,\ 2.5 \times 10^4\}$.

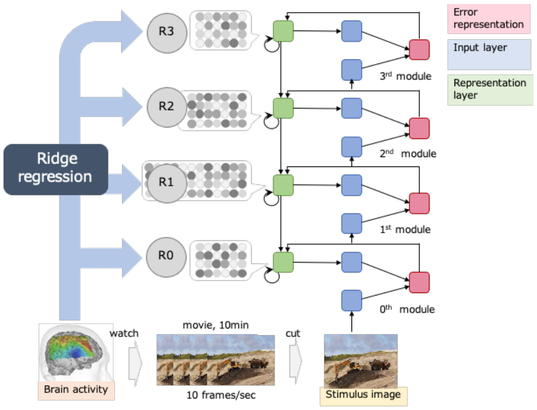


Figure 5: Correlation (PredNet)

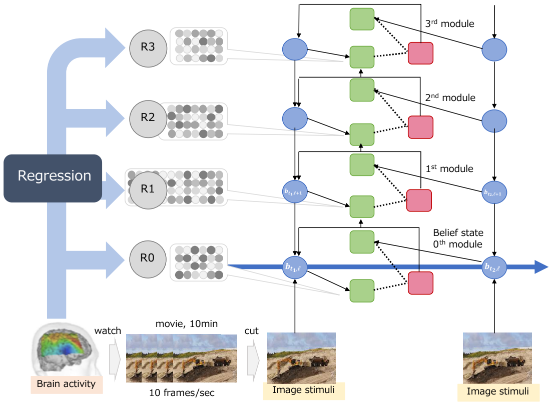


Figure 6: Correlation (our model)

Table 2: Correlation between the feature values and its estimated values from brain activity data in each model.

	PredNet			Our Model		
α	0.5	1K	25K	0.5	1K	25K
R0	0.2623	0.2971	0.3207	0.2635	0.2983	0.3291
R2	0.0925	0.1459	0.1955	0.0003	0.0012	0.0016
R3	0.0254	0.1217	0.1871	0.0004	0.0009	0.0012

For PredNet and our proposed model, the correlation coefficient between the feature value of the lowest layer, R_0, estimated from brain activity data, and the actual value of R_0 is approximately 0.32, when α is 2.5×10^4. From this result, we could say that there is a somewhat significant correlation to those in the field of neuroscience. Furthermore, comparing PredNet and our model, as for the correlation coefficient of R_0, we confirmed that our model is slightly higher than PredNet, while there is no correlation for the other layers, i.e., R_2 and R_3. This is because our proposed model infers with only the lowest layer, i.e., R_0 in the same way as TD-VAE. Thus, we could not confirm the correlation between the feature values of those layers.

5 Conclusion

In this study, we have proposed a new model for predictive coding in the brain by integrating the architecture of PredNet with that of TD-VAE so that the model can predict with flexible prediction intervals. Our model is modified so that it can predict the behaviors of the latent variable which generates prediction objects, unlike PredNet can predict observation objects, i.e., sensory input. Through the experiments, we have confirmed that our model outperformed PredNet in the task of predicting images of a movie. Moreover, we investigated whether there is a correlation between the feature values of the representation layers in both PredNet and our model and the brain activity data from the fMRI while a subject was watching a movie. Through this experiment, we have confirmed that there is a somewhat significant correlation between human brain activity data and the feature values of the lowest layer, R_0, in both PredNet and our proposed model. As future work, we will further improve the proposed model and quantitatively evaluate the prediction accuracy.

References

[1] Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005.

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[3] Karol Gregor and Frederic Besse. Temporal difference variational auto-encoder. *CoRR*, abs/1806.03107, 2018.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[5] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains, 1998.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[8] William Lotter, Gabriel Kreiman, and David D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *CoRR*, abs/1605.08104, 2016.

[9] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.

[10] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[11] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79, 1999.

[12] Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 802–810. Curran Associates, Inc., 2015.