Towards Bridging Human and Artificial Cognition: Hybrid Variational Predictive Coding of the Physical World, the Body and the Brain

André Ofner
Research Focus Cognitive Sciences
University of Potsdam
Potsdam, Germany
ofner@uni-potsdam.de

Sebastian Stober
Artificial Intelligence Lab
Otto von Guericke University
Magdeburg, Germany
stober@ovgu.de

Abstract

Predictive coding and its generalization to active inference offer a unified theory of brain function. The underlying predictive processing paradigm has gained significant attention within artificial intelligence research for its representation learning and predictive capacity. Here, we suggest that it is possible to integrate human and artificial generative models with an artificial neural network that predicts sensations simultaneously with their representation in the brain. Guided by the principles of active inference, we propose a recurrent hierarchical predictive coding model that jointly predicts stimuli, electroencephalogram and physiological signals under variational inference. We suggest that in a shared environment, the artificial inference process can learn to predict and integrate the human generative model. We evaluate the model on a publicly available dataset of subjects watching one-minute long video excerpts and show that the model can be trained to predict physical properties such as the amount, distance and motion of human subjects in future frames of the videos. Our results hint at the possibility of bi-directional active inference across human and machine.

1 Introduction

Predictive processing has been used to explain a large variety of phenomena in human cognition within neuroscience and psychology. Prominently, the notion of predictive coding refers to the idea that perception involves hierarchically organized generative models that aim to predict incoming sensations by expectation error propagation [1]. The more general framework of active inference suggests that perception and action exist in a closed loop, maintaining an agent's internal (probabilistic) generative model of the physical world [2]. These ideas have found traction in machine learning (ML) and a variety of artificial predictive coding and active inference models exist [3, 4]. It has been suggested to evaluate and enhance machine learning models by comparing internal activation with the human brain and ML is used to classify, predict and learn shared embeddings of stimuli and brain activation [5, 6]. Here, we propose to integrate human and artificial cognition directly, with the intention to learn a joint predictive model of the world that augments human representations. Following ideas from active inference, we suggest a multi-modal generative model that learns to predict future states using predictive coding and variational inference. Deep convolutional neural networks are used to parameterize a low-dimensional latent space for multiple time steps. Their latent representations are modulated by previous time-steps with predictive coding.

Three core assumptions underlie the model: 1) The human brain performs hierarchical predictive processing. 2) Information about expectations, intended actions, their outcome and (mis-)match with

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

the incoming sensations is observable in neuroimaging data. 3) These observations can be integrated into an artificial generative model and enhance its predictive capacity.

2 Hybrid variational predictive coding

Following ideas from predictive coding and active inference, we suggest a multi-modal sequential generative model that learns to predict future states using predictive coding and variational inference. Stimuli and EEG signal are processed independently by generating two views from a shared latent embedding z using variational inference: $p(stimulus, eeg, z) = p(z)p(stimulus|z)p(eeg|z)$.

For the sake of simplicity, we add any additional physiological signal as input to the EEG encoder. This structurally follows the multi-view variational autoencoder (MVAE) described by Deep Variational Canonical Correlation Analysis (VCCA), which has been demonstrated to effectively learn shared embeddings of multiple modalities [7]. Following the VCCA principle, the priors $p(z)$, $p(stimulus \mid eeg)$, and $p(eeg \mid z)$ are set to be Gaussian. The projections $E[z \mid stimulus]$ and $E[z \mid eeg]$ of the maximum likelihood solution exist within a shared space that maximizes their correlations. We use deep convolutional neural networks (CNNs) to parameterize the means of $p_\Theta(eeg \mid z)$ and $p_\Theta(stimulus \mid z)$ and the approximate posteriors $q_\phi(z \mid eeg, stimulus)$.

Training with this shared embedding can be done in analogy to variational autoencoders with variational inference by sampling from $q_\phi(z \mid eeg)$. We optimize the lower bound of the log likelihood $L(eeg, stimulus; \theta, \phi)$ with stochastic backpropagation by optimizing the sum of reconstruction losses and the Kullback-Leibler (KL) divergence between the learned $q_\phi(z \mid eeg, stimulus)$ and $p(z)$ using the reparameterization trick [8].

In order to extend the MVAE to process a total of n consecutive time-steps, we iteratively feed inputs into the encoders and compute a total reconstruction loss. For each time-step, an arbitrary selection of encoders can be active. Decoding from the latent space however is always executed for all modalities.

The inputs of the first step are directly used to compute the latent embedding. For time-steps 2 to n, a hierarchy of predictive coding layers process the latent embeddings of previous time-steps and predict the current embedding. This module extends the hierarchical convolutional predictive coding network introduced by Lotter et al. (PredNet) to multimodal processing and variational inference [4].

Like in the original PredNet, each layer l of the predictive coding module features recurrent convolutional network units R^l that are used to compute predictions $\hat{A}^l$ for each layer. These predictions are compared with a target for the corresponding layer A^l. For the lowest layer, the targets are approximate posteriors $q_\phi(z \mid eeg, stimulus)$. For higher layers, the targets are the error E^l between A^l and $\hat{A}^l$. The recurrent representation units R^l receive information about the error E^l of their layer as well as top-down feedback from the representation units in the next higher level of the network R^{l+1}. The error units and the layer-wise predictions are computed with CNNs and the recurrent representations are convolutional LSTMs. We iteratively feed the latent embeddings of time-steps 1 to 3 as inputs and use the resulting predictions $\hat{A}^l$ of the lowest predictive coding layer for variational inference for time-steps 2 - 4. As a result, the latent embeddings for time-steps 2 - 4 are replaced with their predicted counterparts. For accurate predictions, the model must encode the inputs into representations that minimizes the surprise for the next steps and is informative for the predictive coding of upcoming states. We suggest that this forces the network to learn temporal representations of the human physiology, brain and its interaction with the environment that are congruent with the model's own perception.

We refer to this approach of learning a shared generative model that aims at integrating the model's own predictions and the human generative model by means of predictive processing with hybrid predictive coding (HPC). We suggest that predictive coding of (neuro-)physiological signal resembles interoceptive predictive coding, i.e. inference on internal states of the body, which seems to play a crucial role for human cognitive capacity [9].

3 Physical presence and motion prediction with hybrid representations

We used the publicly available DEAP dataset to evaluate the model for its ability to predict future physical states [10]. For this, EEG signal recorded of 22 subjects while watching 40 one-minute long

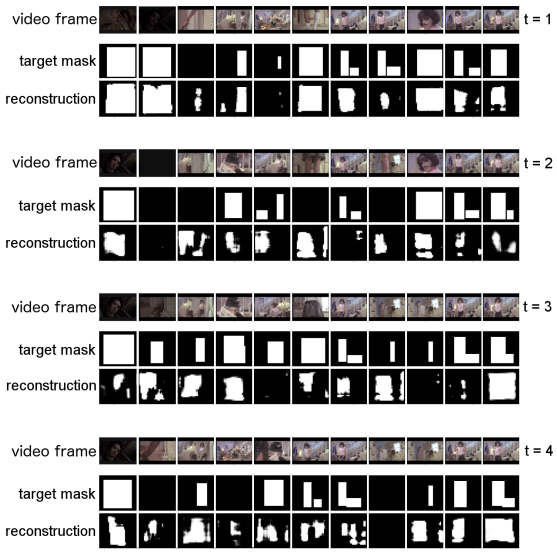

Figure 1: Examples for HPC predicted segmentation masks indicating the position of human subjects within a presented video excerpt. Each of the 11 presented independent examples corresponds to 4 sequential inference steps (from top to bottom). Masks for step 1 are reconstructed with target masks available at the encoders, while step 2-4 are predicted using only EEG and physiological signal.

excerpts of music videos (pop songs) as well as the presented visual stimuli was used as input to the model. EEG was recorded with a sampling rate of 512 Hz in 32 channels. Physiological signals were recorded in 8 channels and contained electrooculography (EOG) and electromyography (EMG) signal during stimulus presentation. The corresponding electrodes were mounted around the eyes, mouth and the shoulder blades. More detailed information about the recording procedure can be found in the DEAP publication. For each subject, the EEG and physiological signal was split into segments of 1 sec duration and the first frame of each second of video was extracted. To evaluate the model performance for predictive object recognition and tracking, we used a pre-trained version of the VGG network to replace each video frame with a segmentation mask framing human subjects if present [11]. This reduces the complexity of visual input, however the recorded brain signal still refers to the complex stimuli. The data for each subject was split by video identity and divided into subsets for training, validation and testing. The test dataset contained only previously unseen stimuli.

We iteratively fed 4 consecutive seconds of EEG and physiological data to the HPC encoders. The preprocessed visual stimulus was only presented for the first step, i.e. steps 2-4 used only EEG and physiological inputs. The total loss was computed as the sum of the individual MVAE reconstruction losses, the KL divergence and the summed reconstruction loss for all latent embeddings processed by the predictive coding module for each step. Reconstruction losses were computed as the mean squared error (MSE) between predictions and targets. We trained for 2000 epochs using the ADAM optimizer and evaluated performance for single subjects and across subjects by visually inspecting the prediction quality.

The network tended to predict the existence of human subjects more frequently than annotated using the VGG network. Interestingly, many of these predictions were wrongly annotated by the VGG network (mostly due to bad lighting) but still correctly interpolated by the HPC network. In longer scenes without any visible human subjects, the HPC network tended to predict many false positives with large fluctuation between frames. If one or multiple humans were visible, the HPC predictions tended to be more sparse compared to the VGG. Judging from visual inspection, the HPC network seemed to improve the quality of its predictions within the 4 time-steps and often chose to not rely on visually guided interpolation. Examples for reconstructions within a single subject are shown in Figure 1). As there is no way for the model to infer whether a subject will move or appear/disappear into the frame, these results indicate that the network learns to replace visual predictions with information from the brain and body.

In this experiment, the information used for the predictions might stem from various sources. For example, information about the initial distance and size of an object could be inferred either from the given video frame or from its representation in the brain. For future frames however, no visual input is provided. This means, that any change in amount, distance or motion in the environment has to be inferred from the physiological representation directly.

4 Conclusion

We proposed a hybrid variational recurrent predictive coding model that learns a shared generative model that integrates artificial and human predictive processes. For this, HPC performs variational inference on a joint latent representation of physical environment, human physiology and brain signal. We demonstrated that the model can be used to predict the content of future frames of videos with respect to existence, number and motion of human subjects. Future work will try to close the inference loop with bi-directional processing, e.g. by allowing humans to have access to visualizations of the learned hybrid generative model while the model adapts in real-time, possibly with reinforcement learning.

5 Acknowledgments

This research is funded by the Federal Ministry of Education and Research of Germany (BMBF).

References

[1] Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1521):1211–1221, 2009.

[2] Rick A Adams, Stewart Shipp, and Karl J Friston. Predictions not commands: active inference in the motor system. *Brain Structure and Function*, 218(3):611–643, 2013.

[3] Kai Ueltzhöffer. Deep active inference. *arXiv preprint arXiv:1709.02341*, 2017.

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

[6] Changde Du, Changying Du, and Huiguang He. Sharing deep generative representation for perceived image reconstruction from human brain activity. *arXiv preprint arXiv:1704.07575*, 2017.

[7] Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

[8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[9] Anil K Seth, Keisuke Suzuki, and Hugo D Critchley. An interoceptive predictive coding model of conscious presence. *Frontiers in psychology*, 2:395, 2012.

[10] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.