Hybrid Variational Predictive Coding as a Bridge between Human and Artificial Cognition

André Ofner[1] and Sebastian Stober[2]

[1] University of Potsdam, Potsdam, Germany
[2] 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 in 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 a predictive coding network that processes sensations simultaneously with the signature of predictive coding found in human neuroimaging data. We propose a recurrent hierarchical predictive coding model that predicts low-dimensional representations of stimuli, electroencephalogram and physiological signals with variational inference. We suggest that in a shared environment, such hybrid predictive coding networks learn to incorporate the human predictive model in order to reduce prediction error. We evaluate the model on a publicly available EEG dataset of subjects watching one-minute long video excerpts. Our initial results indicate that the model can be trained to predict visual properties such as the amount, distance and motion of human subjects in videos.

Introduction

Predictive processing has been used to explain a large variety of phenomena in human cognition within neuroscience and psychology. The notion of predictive coding refers to the idea that perception involves hierarchical predictive models with expectation error propagation (Friston and Kiebel, 2009). The more general framework of active inference suggests that perception and action exist in a closed loop, maintaining an agent's generative model of the world (Adams et al., 2013). These ideas have recently found traction in machine learning (ML). ML has been used to classify, predict and learn shared embeddings of stimuli and brain activation (Du et al., 2017). We propose to interface human and artificial inference on the basis of predictive coding as a shared principle. A predictive coding based artificial neural network processes human neurophysiological data simultaneously with visual stimuli that are perceived by both human and machine. The processed neurophysiological signals reflect predictive coding based inference in the human brain. This means that the artificial model fuses predictions about changes in the shared environment and the corresponding physiological response from human inference. We suggest that this allows the network to subsume its own and human predictions in a joint generative model, in a process referred to as hybrid predictive coding (HPC). Here we focus on augmenting artificial predictive coding using electroencephalography (EEG). The suggested generative model learns to predict compressed representations of multi-modal sensory states in the future by means of prediction error minimization. Deep convolutional neural networks and variational inference are used to parameterize the low-dimensional latent space at each time step.

Hybrid variational predictive coding

We introduce an architecture that processes stimuli, EEG and physiological signals by generating multiple views from a shared latent embedding z: $p(stimulus, eeg, z) = p(z)p(stimulus|z)p(eeg|z)$. EEG and physiological signal are treated as a single view, denoted with $p(eeg)$. The distributions $p(z)$, $p(stimulus \mid eeg)$, and $p(eeg \mid z)$ are set to be Gaussian. The expectations $E[z \mid stimulus]$ 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 using variational inference can be done 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 $p(z)$. In order 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 (Lotter et al., 2016).

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

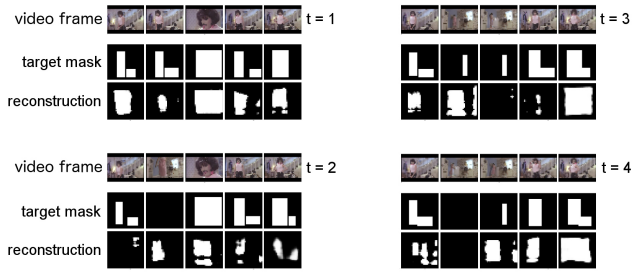

Figure 1: Predicted segmentation masks indicating the position of human subjects within presented videos. Each of the 5 presented examples corresponds to 4 sequential inference steps. Masks for step 1 are reconstructed with target masks available. Step 2-4 are predicted using only EEG and physiological signal.

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 $n-1$ as inputs and use the resulting predictions $\hat{A}^l$ of the lowest predictive coding layer for variational inference for time-steps $2-n$. Their embeddings are replaced with the predicted counterparts. This way, the model encodes the inputs into representations that minimize the surprise for future steps. We suggest that this forces the network to learn temporal representations of the human physiology and brain signals 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 as hybrid predictive coding. 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 (Seth et al., 2012).

Predicting with hybrid representations

We used the publicly available DEAP dataset to evaluate the model for the ability to predict future states (Koelstra et al., 2012). 32 channel EEG was recorded of 22 subjects while watching 40 one-minute long excerpts of music videos as well as the presented visual stimuli were provided as inputs to the model. Electrooculography (EOG) and electromyography (EMG) signals were recorded. The electrodes were mounted around the eyes, mouth and the shoulder blades. EEG and physiological signals were split into segments of 1 sec duration. The first frame of each second of video was extracted. We used a pre-trained image segmentation network to replace each video frame with a segmentation mask marking human subjects if present. This reduces the complexity of visual input, but the EEG signal still refers to the complex stimuli. The data for each subject was split by video identity. The test set 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 loss was computed as the sum of the reconstruction losses and the KL divergence for each step.

The network tended to predict the existence of human subjects more frequently than annotated using the segmentation network. Interestingly, many of these predictions were wrongly annotated by the segmentation network but still correctly interpolated by the HPC network. In longer scenes without 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 in comparison. Upon visual inspection, HPC 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 guide visual predictions with information from the brain and body. Information about the initial distance and size of an object could be inferred from the given video frame or from its physiological representation. For future frames however, no visual input is provided. Change in amount, distance or motion in the environment has to be inferred from the physiological representation.

Conclusion

We proposed a hybrid variational predictive coding architecture that interfaces artificial and human predictive coding. HPC performs predictive coding based inference about a shared visual environment, human physiology and brain signal. The same stimuli are perceived by human and machine, allowing the model's predictions to be modulated by human inference. Our initial results using an EEG dataset suggest that such a model can be used to predict aspects of the visual content of future frames of videos, such as the movement of human subjects.

Acknowledgments

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

References

Adams, R. A., Shipp, S., and Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Structure and Function*, 218(3):611–643.

Du, C., Du, C., and He, H. (2017). Sharing deep generative representation for perceived image reconstruction from human brain activity. *arXiv preprint arXiv:1704.07575*.

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

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2012). Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31.

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

Seth, A. K., Suzuki, K., and Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in psychology*, 2:395.