Deep neural networks can predict human behavior in arcade games

Holger Mohr (holger.mohr@tu-dresden.de)
Department of Psychology, Technische Universität Dresden
Zellescher Weg 17, 01069 Dresden, Germany

Radoslaw M. Cichy (rmcichy@zedat.fu-berlin.de)
Department of Education and Psychology, Freie Universität Berlin,
Habelschwerdter Allee 45, 14195 Berlin, Germany

Hannes Ruge (hannes.ruge@tu-dresden.de)
Department of Psychology, Technische Universität Dresden
Zellescher Weg 17, 01069 Dresden, Germany

Abstract:

In a standard experimental paradigm typically used in cognitive neuroscience, time is discretized into distinct events and the presented stimulus material is sampled from a low number of categories. While this approach allows conducting highly controlled experiments, its ecological validity is limited, as in real life the human brain has to operate on a continuous time scale and also has to process complex stimuli that typically cannot be assigned to a low-dimensional stimulus space. The encoding model approach has been introduced to address these issues by using high-dimensional sets of stimulus features as a means to analyze neuroimaging data from complex and time-continuous tasks. Recently, activations from deep neural networks (DNNs) were proposed to serve as features in the encoding model approach. However, it has been argued that such DNN-based features might be uninformative for human neuroimaging data, as the behavior of a trained DNN does not necessarily have to resemble human behavior on a given task. Here, we present preliminary evidence (N = 1) that DNN activations from the top network layer can predict human behavior with high fidelity in three different Atari 2600 arcade games based on a linear model. These findings clear the way for extending this type of analysis to neuroimaging data, testing whether DNN activations extracted from hidden layers explain variance in the fMRI signal of task-related brain regions.

Keywords: encoding models; deep neural networks; arcade games; Atari; neuroimaging; FMRI

Introduction

To isolate the neural correlates of specific cognitive processes, the factorial experimental design has been established as the workhorse in cognitive neuroimaging (Friston et al., 1995). The factorial design approach assumes that the presented stimulus material is carefully selected such that stimulus dimensions of no interest are balanced across pre-defined experimental conditions, and the stimuli are typically presented in a sequence of discrete temporal events (trials), with balanced ordering across experimental conditions. While this highly controlled experimental approach has provided many interesting insights into the neurofunctional architecture of the human brain, we argue that the ecological validity of this approach is limited, as in real life the human brain has to operate on a continuous time scale and process stimuli sampled from dynamic environments without an underlying low-dimensional categorial structure.

To model the neural activation underlying human behavior in complex, time-continuous tasks, we have to depart from factorial designs and look for alternative analysis approaches. One such approach, the encoding model approach, has been initially employed to characterize neural activation patterns underlying visual processing of complex, naturalistic stimuli (Kay et al., 2008). In the Kay et al. study, stimuli were not sampled from a low number of different categories, but instead each stimulus was represented by a high-dimensional feature vector, and these features were used as predictors in a linear model. Using the encoding model approach, Kay et al. showed that it is possible to predict the presented stimuli with high accuracy based on fMRI activation patterns in early visual cortex. Subsequently, it was shown that the encoding model approach also works in the time-continuous domain by decoding short video clips from activation patterns in early visual cortex (Nishimoto et al., 2011). The encoding model approach has also been used in tasks not requiring visual processing, for example to characterize the semantic representations of words by predicting neural activations associated with story listening (Huth et al., 2016).

In the aforementioned studies, the features used for prediction in the encoding models were carefully

chosen for the given task, for example Gabor wavelets for visual processing or a basic dictionary for semantic encoding. Interestingly, it was recently shown that manually designed features can be replaced by activations from deep neural networks (DNNs) both in the visual and auditory domain (Güclü and van Gerven, 2015; Cichy et al., 2016; Kell et al., 2018). The DNN-informed version of the encoding model approach assumes that a DNN architecture exists that can be trained to perform the given experimental task at least at human level performance. The stimulus material presented to the subjects is also processed through the trained DNN, and the resulting activations are used as features in the encoding model in order to predict fMRI time series or activation patterns (van Gerven, 2017). Thus, the DNN-informed encoding model approach can be seen as a specific realization of the general idea of employing DNNs for computational modeling in cognitive neuroimaging, as recently proposed by several authors (Naselaris et al., 2018; Kriegeskorte and Douglas, 2018; Cichy and Kaiser, 2019).

Over the last couple of years, research on deep learning has been extended from visual and auditory processing to more complex task requiring dynamically generated motor responses, for example in arcade games or spatial navigation tasks (Mnih et al., 2015; Jaderberg et al., 2018; Banino et al., 2018). While it is straightforward to implement the DNN-informed encoding model approach for such more complex tasks from a technical perspective, interesting novel questions arise in this context, as there are no guarantees that task-performing DNNs match with humans at the behavioral level, i.e. motor responses generated by the DNNs might substantially deviate from human-generated responses. As a consequence, features extracted from hidden layers might be uninformative for neuroimaging data.

Here we test to which extent activations extracted from the top layer of DNNs trained to perform different arcade games can be used to predict human behavior on the respective games. As the employed DNNs generated Q-values as outcomes at the top layer, we equivalently test whether the response distributions generated by the DNNs can be used to predict human responses. We present preliminary evidence that the DNN-informed encoding model approach can be used to predict human motor responses with high accuracy in each of the three tested arcade games.

Methods

Sample We present data from one subject (author H.M.).

Tasks Three Atari 2600 games were chosen to cover different types of arcade games, see Figure 1. Before starting to play, the subject read the manuals of the three games. After approximately one hour of training in each of the games, videos were recorded while the subject performed 5 blocks of 5 min length of each game. The games were performed on a desktop computer using a standard keyboard. Games were presented via the Atari Learning Environment (ALE, Bellemare et al., 2013) using a modified Python script originally written by Ben Goodrich (ALE Python interface). Left, right, bottom, up responses were given via the respective arrow keys (right hand) and the fire response via the 'z' button (left hand). The game screen was presented at a resolution of 1280 x 840 pixels (width x height), and game states were updated at a frequency of 60 hz.

Figure 1: The three Atari 2600 arcade games used as experimental tasks. In Breakout, the player controls a paddle at the bottom of the screen, steering a ball towards the wall at the top of the screen in order to remove bricks from the wall. In Enduro, the player steers a car along a race track. In Space Invaders, the player controls a spacecraft and fights against yellow aliens.

Deep neural networks The architecture of the employed DNNs was the same as described in the original Atari DQN paper (Mnih et al., 2015), with 3 convolutional steps and 2 fully connected processing steps from layer 1 to layer 6. Trained versions of the Enduro and Space Invaders networks were downloaded from the GitHub repository of Parisotto et al., 2016, while the Breakout DNN was trained on 40 million frames on a Nvidia Quadro P2000 graphics card using the same training procedure (RMSprop) and parameters as in Mnih et al., 2015. After training via reinforcement learning on the respective tasks, the DNNs' weights were fixed, i.e. no steps were taken to adapt the DNNs' behavior to human behavior during training.

Data processing Before being submitted to the DNNs, the recorded videos were downsampled to 15 hz by taking the maximum over every 3rd and 4th

frame, as in Mnih et al., 2015. Moreover, the screen resolution was downsampled to 84 x 84 pixels and RGB colors were converted to grayscale as in Mnih et al., 2015. One gaming block of 5 min corresponded to 4,500 processing steps for the network. The networks' top layer activation values (i.e. Q-values) were transformed into response probabilities via the softmax function using a temperature of $\tau = 1$:

$$p_i = \frac{\exp\left(\frac{q_i}{\tau}\right)}{\sum_k \exp\left(\frac{q_k}{\tau}\right)}$$

For Breakout, the 4 response options were no action, fire, left, right. For Enduro, the 9 response options were no action, fire, right, left, down, down-right, down-left, right-fire, left-fire. For Space Invaders, the 6 response options were no action, fire, right, left, right-fire, left-fire. The probabilities of the responses were upsampled to 18,000 frames per block by repeating each probability vector four times, as the human responses were recorded at 60 hz. For Breakout, human responses were binary coded for no action, fire, right and left. For Enduro, human responses were binary coded for no action, fire (also including right-fire and left-fire), right (also including right-fire, down-right), left (also including left-fire, down-left), and down (also including down-right, down-left). For Space Invaders, human responses were binary coded for no action, fire (also including right-fire and left-fire), right (also including right-fire), and left (also including left-fire).

As the presented pilot study was conducted to evaluate the DNN-informed encoding model approach from a behavioral perspective for future fMRI studies, both the human responses and the response probabilities generated by the DNNs were convolved with a Gaussian kernel of FWHM = 5.3 s, corresponding to the FWHM of the standard hemodynamic response function (HRF) in the SPM software package. The convolution step was implemented to remove information in the high frequency domain, as this information can also be expected to be absent in the BOLD signal in future fMRI studies. The lengths of the time series were reduced to 17,100 frames per block by the convolution step.

Encoding model and cross-validation The encoding model was a standard linear regression model with a convolved binary human response vector as outcome variable and convolved DNN-generated response probability variables as predictors. The model was fitted via ordinary least squares. A separate model was fitted for each human response variable. This simple modeling approach was chosen to show that a linear model can be sufficient for prediction of human behavior, as in future fMRI studies the encoding model will also be linear. The models were fitted on four blocks and based on this fit a prediction for the human responses for the left-out fifth block was computed, see Figure 2. This procedure was repeated such that each block was left out once (5-fold cross-validation). To quantify the predictive accuracy of the model, the Pearson correlation between the predicted and actual time series was computed on the left-out blocks. As an alternative measure, the mean squared error (MSE) of the predicted time series was computed. These values were averaged across left-out blocks and response variables to provide a summary measure of predictive accuracy for each game.

Permutation testing The whole cross-validation procedure described above was repeated N = 100,000 times with randomized human responses to generate a null distribution for the correlation coefficient and the MSE for each game. Human responses were randomly shuffled in each block and convolved with the Gaussian kernel. The resulting randomized time series were rescaled to have the same minima and maxima as the original time series (note that otherwise the randomized time series would have considerably less variance than the original time series due to the convolution step, which would not be an appropriate null model for the MSE). The same 5-fold cross-validation scheme as in the original analysis was implemented to obtain Pearson's correlation values and MSE values sampled under the null hypothesis.

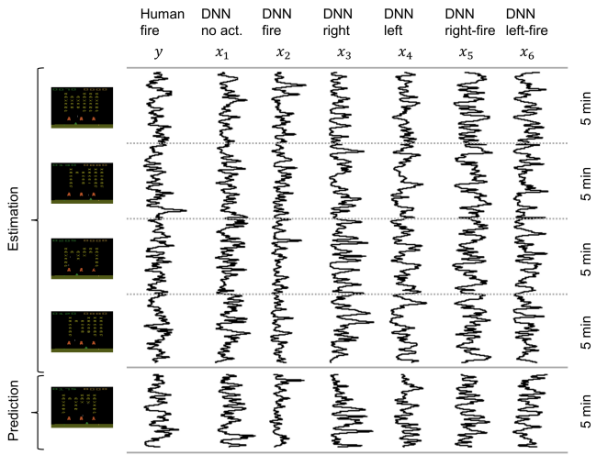

Figure 2: Illustration of the encoding model and 5-fold cross-validation scheme. A linear regression model was fitted to a human response variable (fire response in Space Invaders in this example) on four blocks using the response probabilities from a DNN trained on this game. Based on this model, a prediction for the human response variable of the left-out 5th block was computed and compared with the actual time series.

Game	Pearson's r	p(r)	MSE	p(MSE)
Breakout	0.46	< 0.00001	0.0012	< 0.00001
Enduro	0.47	< 0.00001	0.0139	< 0.00001
Space Invaders	0.23	< 0.00001	0.0107	0.00006

Table 1: Results

Results

The results are depicted in Table 1. Predictions of human motor responses were significantly above chance for all three games both in terms of Pearson correlation and MSE.

Discussion

The presented results provide preliminary evidence that human behavior in arcade games can be predicted with high accuracy using a DNN-informed encoding model approach. While the presented behavioral data were recorded at a high sampling rate, the predictions were based on time series convolved with a Gaussian kernel corresponding in width to the canonical HRF in SPM, which means that human behavior can be predicted at a temporal resolution covered by the BOLD response. Thus, this finding clears the way for an extrapolation of this approach to neuroimaging data. As the encoding model consists of a simple linear regression model, an extension to hidden layers encompassing more features also seems to be feasible but would require regularization of the encoding model, as implemented before in other studies (Kay et al., 2008; Schoenmakers et al., 2013; Güclü and van Gerven, 2015; Huth et al., 2016).

Acknowledgments

This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), grant SFB 940, project Z2.

References

Banino, A., Barry, C., Uria, B., et al. (2018). Vector-based navigation using grid-like representations in artificial agents. Nature *557*, 429–433.

Bellemare, M.G., Naddaf, Y., Veness, J., and Bowling, M. (2013). The Arcade Learning Environment: An Evaluation Platform for General Agents. J. Artif. Intell. Res.

Cichy, R.M., and Kaiser, D. (2019). Deep Neural Networks as Scientific Models. Trends Cogn. Sci. *23*, 305–317.

Cichy, R.M., Khosla, A., et al. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci. Rep. *6*.

Friston, K.J., Holmes, A.P., et al. (1995). Statistical parametric maps in functional imaging: A general linear approach. Hum. Brain Mapp. *2*, 189–210.

van Gerven, M.A.J. (2017). A primer on encoding models in sensory neuroscience. J. Math. Psychol. *76*, 172–183.

Güclü, U., and van Gerven, M.A.J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. J. Neurosci. *35*, 10005–10014.

Huth, A.G., de Heer, W.A., et al. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature *532*, 453–458.

Jaderberg, M., Czarnecki, W.M., et al. (2018). Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. ArXiv.Org *abs/1807.01281*.

Kay, K.N., Naselaris, T., et al. (2008). Identifying natural images from human brain activity. Nature *452*, 352–355.

Kell, A.J.E., Yamins, D.L.K., et al. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron *98*, 630-644.e16.

Kriegeskorte, N., and Douglas, P.K. (2018). Cognitive computational neuroscience. Nat. Neurosci. *21*, 1148–1160.

Mnih, V., Kavukcuoglu, K., et al. (2015). Human-level control through deep reinforcement learning. Nature *518*, 529–533.

Naselaris, T., Bassett, D.S., et al. (2018). Cognitive Computational Neuroscience: A New Conference for an Emerging Discipline. Trends Cogn. Sci. *22*, 365–367.

Nishimoto, S., Vu, A.T., et al. (2011). Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. Curr. Biol. *21*, 1641–1646.

Parisotto, E., Ba, L.J., and Salakhutdinov, R. (2016). Actor-Mimic: Deep Multitask and Transfer Reinforcement Learning. Arxiv.Org *abs/1511.06342*.

Schoenmakers, S., Barth, M., Heskes, T., and van Gerven, M. (2013). Linear reconstruction of perceived images from human brain activity. NeuroImage *83*, 951–961.