Brain-Mediated Transfer Learning of Convolutional Neural Networks

Satoshi Nishida,[1] **Yusuke Nakano,**[1] **Antoine Blanc,**[1] **Naoya Maeda,**[2]
Masataka Kado,[2] **Shinji Nishimoto**[1]

[1]Center for Information and Neural Networks (CiNet),
National Institute of Information and Communications Technology (NICT) and Osaka University, Osaka, Japan
[2]NTT DATA Corporation, Tokyo, Japan
{s-nishida, ynakano, blanc, nishimoto}@nict.go.jp, {naoya.maeda, masataka.kado}@nttdata.com

Abstract

The human brain can effectively learn a new task from a small number of samples, which indicates that the brain can transfer its prior knowledge to solve tasks in different domains. This function is analogous to transfer learning (TL) in the field of machine learning. TL uses a well-trained feature space in a specific task domain to improve performance in new tasks with insufficient training data. TL with rich feature representations, such as features of convolutional neural networks (CNNs), shows high generalization ability across different task domains. However, such TL is still insufficient in making machine learning attain generalization ability comparable to that of the human brain. To examine if the internal representation of the brain could be used to achieve more efficient TL, we introduce a method for TL mediated by human brains. Our method transforms feature representations of audiovisual inputs in CNNs into those in activation patterns of individual brains via their association learned ahead using measured brain responses. Then, to estimate labels reflecting human cognition and behavior induced by the audiovisual inputs, the transformed representations are used for TL. We demonstrate that our brain-mediated TL (BTL) shows higher performance in the label estimation than the standard TL. In addition, we illustrate that the estimations mediated by different brains vary from brain to brain, and the variability reflects the individual variability in perception. Thus, our BTL provides a framework to improve the generalization ability of machine-learning feature representations and enable machine learning to estimate human-like cognition and behavior, including individual variability.

Introduction

Transfer learning (TL) is an effective framework that improves the generalization ability of machine learning in pattern recognition (Pan and Yang 2009). In TL, feature representations trained with large-scale, high-quality data in a specific task are transferred to other tasks in different domains with insufficient training data. Recently, TL with convolutional neural networks (CNNs) has received increasing attention owing to its high generalization ability across various task domains (Tan et al. 2018). Despite the

fact that the training of CNNs requires massive amounts of labeled datasets, CNNs acquire generic feature representations available for various types of pattern recognition through TL (Cimpoi et al. 2016; Donahue et al. 2014; Girshick et al. 2014; Oquab et al. 2014; Razavian et al. 2014; Toshev and Szegedy 2014; Xiao et al. 2014). For example, Donahue et al. (2014) used a CNN feature space, pre-learned with an object recognition dataset, for tasks different from the original challenge, such as scene classification. Using higher-layer features of the pre-trained CNN, they trained classifiers for those new tasks and achieved state-of-the-art performance. In this way, TL improves the generalization ability of machine learning more effectively when using more generic feature representations. However, the generalization ability of machine learning is still much lower than that of the human brain.

The human brain has a sophisticated ability to generalize the knowledge acquired through limited experiences to cognition and behavior under novel situations. This indicates that the human brain can effectively use its own internal representations of cognitive information across different task domains. Furthermore, such representations can be used as a medium of brain decoding that infers various types of cognition from brain activity, such as semantic cognition (Mitchell et al. 2008; Nishida and Nishimoto 2018), affections (Kim et al. 2015; Peelen, Atkinson, and Vuilleumier 2010), economic decision (Hampton and O'doherty 2007; Knutson et al. 2007), human mass behavior (Dmochowski et al. 2014; Falk, Berkman, and Lieberman 2012). Therefore, machine-learning feature representations may become more generic by incorporating the brain representations into them, leading to further improvements in the generalization ability of machine learning through TL.

Recently, several studies have proposed techniques to guide pattern recognition with CNNs by combining CNN features with human brain activity (Fong, Scheirer, and Cox 2018; Spampinato et al. 2017). For example, Fong et al. (2018) trained a classifier of images using higher-layer features of a CNN while guiding the training by voxel responses to the images measured using functional magnetic resonance imaging (fMRI). To make the decision surface of the classifier more consistent with brain representations, the train-

ing was weighted by voxel responses. They demonstrated higher performance in image recognition for the classifier with voxel response weighting than for the classifier without it. These studies suggest that CNN feature representations can improve by combining them with human brain activity. However, the recognition task they targeted was image recognition, in which CNNs had already shown splendid performance by themselves. In addition, when their techniques are applied to other types of recognition tasks, they require brain data dedicated to each of the tasks. Therefore, whenever their techniques are applied to a novel task, new measurements of brain data specified to that task are required. These limitations prevent their techniques from improving the generalization ability of machine learning.

To address these issues, we propose, in this paper, brain-mediated TL (BTL). In the BTL, the voxel response data measured from the human brain in an fMRI experiment are combined with CNN features to be used for various types of recognition tasks. First, CNN features for audiovisual inputs are transformed into features of the individual brains using a pre-learned linear mapping between them. Then, the transformed features are used to train learners that estimate arbitrary pairs of audiovisual inputs and corresponding labels. We apply this BTL to the estimation of labels to which human subjective cognition is strongly related. Our results demonstrate that the BTL shows higher estimation performance than regular TL and the brain-by-brain variation of estimation by the BTL reflects individual variability in subjective cognition.

Model

Procedure

Our modeling procedure for BTL is divided into the four steps (Figure 1): (1) the feature extraction from the inputs of movies and/or sounds via CNNs, (2) the prediction of voxel responses to the inputs using models that linearly transform the extracted features to the voxel responses (CNN-to-voxel [cnn2vox] models), (3) the modification of the predicted responses using models that predict voxel response from the history of preceding voxel responses (voxel-to-voxel [vox2vox] models; Figure 1A), and (4) the estimation of cognitive labels from the predicted responses using models that have linear association between them (voxel-to-label [vox2lab] models; Figure 1B).

The cnn2vox and vox2vox models are pretrained using small datasets of movie-evoked voxel responses collected from individual brains with fMRI. Once the training is done, these models predict voxel responses to arbitrary audiovisual inputs by transforming the CNN features to the response of individual brains with no additional brain measurement. Supposing that voxel responses convey brain feature representations of the inputs, this process corresponds to the transformation from CNN features to brain feature representations. Then, the vox2lab model is trained using paired datasets of predicted voxel response and cognitive labels linked with audiovisual inputs. Since this training does not require datasets used in fMRI experiments, our method allows acquiring the association between arbitrary pairs of audiovisual inputs and cognitive labels through the feature transformation.

CNN Feature Extraction

In this study, VGG-16 (Simonyan and Zisserman 2014) and SoundNet (Aytar, Vondrick, and Torralba 2016), which are pretrained and released on the web, are used to extract visual and acoustic features, respectively, from movies. To extract visual features from movies via VGG-16, which was originally applied to static images with a fixed size of 224×224 pixels, the movies are decomposed into frames and resized to the same size. Then, unit activations of intermediate layers when inputting the movie frames are calculated and pooled for each second. Finally, the maximum activation value of each unit for each second is used as the visual features of the movies. This study uses eight layers of pool1–5 and fc6–8 and obtains the visual features for each layer. To extract acoustic features from the same movies, the sound waves of the inputs are resampled with the fixed frequency of 44100 Hz and decomposed into each second. Then, unit activations of intermediate layers in SoundNet when inputting the sound waves are calculated as the acoustic features of the movies. This study uses seven layers of conv1–5 and fc 6–7 and obtains the acoustic features for each layer. Finally, these processes produce eight series of visual features and seven series of acoustic features from time series of movies.

Cnn2vox Model

The construction of the cnn2vox, vox2vox, and vox2lab models is based on the voxelwise modeling technique (Naselaris et al. 2011). Using a time series of features and voxel responses, the cnn2vox model acquires the linear mapping from a CNN feature space to a response space of each voxel through statistical learning. The learning objective is to estimate weights of N voxels, denoted by $\mathbf{W}_{\mathrm{cv}} = \{\mathbf{w}_{\mathrm{cv}(1)}, \cdots, \mathbf{w}_{\mathrm{cv}(N)}\}$, of the linear model: $\mathbf{R} = f(\mathbf{X})\mathbf{W}_{\mathrm{cv}} + \epsilon$, where $\mathbf{R} = \{\mathbf{r}_1, \cdots, \mathbf{r}_N\}$ is a series of responses in each of N voxels, $\mathbf{X}$ is a series of audiovisual inputs, $f(\mathbf{X})$ is its feature representation with the dimensionality of D, and ϵ is isotropic Gaussian noise. A set of linear temporal filters is used to capture the hemodynamic delay in the response (Nishimoto et al. 2011). The matrix of $f(\mathbf{X})$ is constructed by concatenating four sets of D-dimensional feature vectors with temporal shifts of 3, 4, 5, and 6 s. This means that voxel response at a time point t, denoted by $\mathbf{R}_{(t)}$ $(t = 1, \cdots, T)$, is modeled by a weighted linear combination of the preceded series of features:

$$\mathbf{R}_{(t)} = \sum_{k=3,4,5,6} f(\mathbf{X}_{(t-k)})\mathbf{W}_{\mathrm{cv},k} + \epsilon,$$

where $\mathbf{W}_{\mathrm{cv},k}$ denotes the weights corresponding to the delay k. The weight estimation is performed using L2-regularized linear least-squares regression. The optimal regularization parameter for each model is determined by 10-fold cross validation of training data and shared across all voxels. In this study, $f(\mathbf{X})$ represents a series of unit activations induced by movies for each layer of VGG-16 or Sound-Net. Since the very large number of units in lower layers of
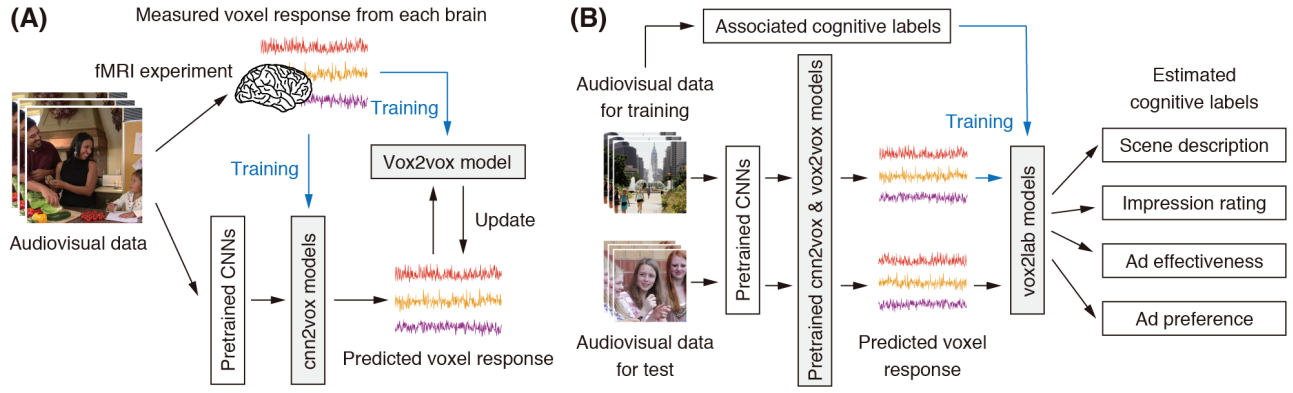
Figure 1: A schematic of the proposed brain-mediated transfer learning (BTL) method. (A) Voxel-response prediction from audiovisual inputs. (B) Label estimation using predicted response.

the CNNs took too much computational cost for the regression process, the dimensionality of unit-activation features for each layer is reduced in advance by principal component analysis (PCA) on training datasets. This study reduces the dimensionality, D, to 1000. Finally, eight models for eight VGG-16 layers (pool1–5 and fc6–8) and seven models for seven SoundNet layers (conv1–7) are constructed for each brain.

Each of the estimated linear models predicts voxel responses to new movie inputs. Then, the predicted voxel responses from individual models are integrated using linearly weighted average for each voxel. The weight for a given voxel is calculated based on the prediction accuracy (Pearson's correlation coefficient between measured and predicted voxel responses) for that voxel calculated during the cross-validation in model training. In particular, the weight for the i-th model, denoted by w_i, is determined by $w_i = a_i / \sum_j^{15} a_j$, where a_i is the prediction accuracy of the model. This integration process produces a single series of predicted voxel responses to the new movie inputs.

Vox2vox Model

The vox2vox model predicts response in one voxel at a given time point from responses in a group of voxels at the preceding time points. Hence, this model captures endogenous properties of voxel responses, such as intrinsic connectivity between different brain regions (Fox and Raichle 2007), whereas the cnn2vox model captures exogenous properties of voxel responses, such as stimulus selectivity. Although the vox2vox model is not indispensable for BTL, modifying response prediction by the vox2vox model improves estimation performance in some cases.

In the vox2vox model, a response in each of N voxels at a time point t, denoted by $\mathbf{R}_{(t)}$, is modeled by a weighted linear combination of responses in the set of M voxels preceded by 1, 2, and 3 s:

$$\mathbf{R}_{(t)} = \sum_{k=1,2,3} \mathbf{R}'_{(t-k)} \mathbf{W}_{\mathrm{vv},k} + \epsilon.$$

The M voxels are selected on the basis of the cnn2vox model prediction accuracy on model training data. In this

study, the top 2000 voxels with the highest prediction accuracy after the weighted average of all the models are used as the M voxels. The regression procedure is the same as the one used in the cnn2vox model. Response predictions from the cnn2vox and vox2vox models are combined by weighted sum. The weight is determined by the relative accuracy of each prediction for each voxel.

Vox2lab Model

The vox2lab model estimates cognitive labels associated with audiovisual inputs from predicted voxel responses. In this model, a series of z-scored cognitive labels at a time point t, denoted by $\mathbf{L}_{(t)}$ $(t = 1, \cdots, T')$, is regressed by a series of predicted responses to the inputs in the set of N voxels with the hemodynamic delay, k, of 3, 4, and 5s (Nishida and Nishimoto 2018):

$$\mathbf{L}_{(t)} = \sum_{k=3,4,5} \hat{\mathbf{R}}_{(t+k)} \mathbf{W}_{\mathrm{vl},k} + \epsilon.$$

The regression procedure is the same as the one used in the other models, except that the regularization parameters are determined separately for each dimension of label vectors. This model learns the association between predicted voxel responses (but not measured voxel responses) and cognitive labels. This allows us to associate arbitrary pairs of audiovisual data and corresponding labels, even when they are not used for training either the cnn2vox or vox2vox models.

However, the regression occasionally requires too much computational cost when the paired data have a large number of samples. In such cases, the number of voxels used for the modeling is reduced in advance by PCA on the training datasets. Then, the top PCs are used for the training of the three models. In this study, the top 300 PCs are used as voxels to predict (i.e., $N = 300$) for only one estimation task (see below). In addition, the top 10 PCs with the highest prediction accuracy are used for the regressors of the vox2vox model (i.e., $M = 10$). Even in this case, the other parameters and the regression procedure remain the same.

Data

Movie

Two sets of movies were provided by NTT DATA Corp. (Tokyo, Japan). One includes 368 Japanese ad movies broadcasted on the web between 2015 and 2018 (web ad movies). The movies were divided into 7200 s and 1200 s to collect voxel responses for the training (training dataset) and test (test dataset), respectively, of all the three models. The other includes 2452 Japanese ad movies broadcasted on TV between 2011 and 2017 (TV ad movies). Out of them, only 420 movies were used to collect voxel responses for the training of the cnn2vox and vox2vox models. The remaining 2032 movies were not used for fMRI experiments but for the training of the vox2lab model and the test of its estimation performance. The movies are all unique, include a wide variety of product categories (see supplemental material[1]), and have the same resolution (1280×720 pixels) and frame rate (30 Hz). The length of the movies is typically either 15 or 30 s. They are also accompanied by PCM sounds with the sampling rate of 44100 Hz and are normalized so that they have the same RMS level.

fMRI Data

fMRI responses to the movies were collected from Japanese participants using a 3T MRI scanner. Forty and twenty-eight participants were assigned to fMRI experiments with the web ad movies and those with the TV ad movies, respectively. The experimental protocol is approved by the ethics and safety committees of NICT. For the modeling in each participant, the fMRI data were preprocessed and all voxels within the whole cortex were extracted. For more details, see supplemental material.

Cognitive Labels

The four types of cognitive labels associated with the movie datasets are used for testing the performance of BTL and comparing it with the performance of other methods: (1) scene descriptions, (2) impression ratings, (3) ad effectiveness indices, and (4) ad preference votes. All the label sets reflect rich perceptual information and/or complex behavioral outcomes related to human subjective cognition. The label sets (1)–(3) are linked to the web ad movies, whereas the label set (4) is linked to the TV ad movies. The label sets (1), (3), and (4) were provided from NTT DATA Corp., whereas the label set (2) was collected on psychological experiments conducted in our lab. In the following, the details of each cognitive label set are described (for more details, see supplemental material).

Scene description data were collected from human annotators as manual descriptions given for every 1-s scene of the web ad movies. They were instructed to describe each scene using more than 50 Japanese characters. The descriptions contain a variety of expressions reflecting not only their objective perceptions but also their subjective perceptions (e.g., impression, feeling, association with ideas). To evaluate the scene descriptions quantitatively, the descriptions

were transformed into vectors of word2vec (Mikolov et al. 2013). Individual words in each description were projected into the pretrained word2vec vector space. Then, the word vectors obtained from all descriptions within each scene were averaged. This procedure yielded one 100-dimensional vector for each 1-s scene.

Impression rating data were collected from manual ratings on 30 different impression items conducted by human annotators. The ratings were given for every 2-s scene of the web ad movies. While the annotators sequentially watched 2-s separate clips of the movies, they evaluated each item on a scale of 0 to 4.The mean impression ratings in every 2-s scene were obtained by averaging multiple ratings and then oversampled to obtain time series of rating labels in every 1-s scene.

Ad effectiveness index data for the web ad movies were obtained as two types of mass behavior indices collected on the web. One index is click rate, that is, the fraction of viewers who clicked the frame of a movie and jumped to a linked web page. The other index is view completion rate, that is, the fraction of viewers who continued to watch an ad movie until the end without choosing a skip option. Although a single value of each index was assigned to each ad, time series of indices in every 1-s scene were obtained by filling all scenes in an ad with an identical index value assigned to the ad.

Ad preference vote data for the TV ad movies were collected for commercial investigation using questionnaires to large-scale testers. Each tester was asked to freely recall a small number of her/his favorite TV ads from among the ads recently broadcasted. The total number of recalls of an ad was regarded as its preference value. Although one preference value was assigned to each ad, time series of the value in every 1-s scene were obtained by filling all scenes in an ad with an identical value assigned to the ad. Since the preference data are distributed in a similar form of a gamma distribution, the logarithm of the data was taken. Only for this dataset, due to the large number of samples for training the vox2label model in the dataset (2032 movies, 45375 s), the voxel dimensionality is reduced to 300 in advance by PCA for computational efficiency of the modeling (see also BTL).

Reference Methods

To compare the performance of label estimation with BTL, two other methods are implemented.

Regular TL. TL estimates cognitive labels without the transformation from CNN features to voxel responses. Although the CNN feature extraction using VGG-16 and SoundNet is the same as the one used in BTL, CNN features are directly used to regress cognitive labels. In particular, a label at a time point t, denoted by $\mathbf{L}_{(t)}$, is modeled by a weighted linear combination of CNN features of the input at the same time point. For the neutral comparison with BTL, the reduction of CNN feature dimension to 1000 and the regression is performed in the same manner as used in BTL.

Table 1: Accuracy of voxel-response prediction

| Cnn2vox | | | Prediction accuracy |
VGG-16	SoundNet	Vox2vox	(mean $\pm$ SEM)
✓			0.096 ± 0.0039
✓		✓	0.098 ± 0.0040
	✓		0.054 ± 0.0018
	✓	✓	0.055 ± 0.0018
✓	✓		0.106 ± 0.0039
✓	✓	✓	0.108 ± 0.0039

In contrast to this model, the BTL models contain linear temporal filters that capture the hemodynamic response delay. This may enable only BTL to estimate labels from a series of features at multiple time points. If so, the comparison between BTL and TL becomes unfair. To avoid this, another TL model is also intorduced so that a label at a given time point is estimated by a combination of CNN features at -1, 0, and 1 s from that time point: $\mathbf{L}_{(t)} = \sum_{k=-1,0,1} f(\mathbf{X}_{(t-k)})\mathbf{W}_{\mathrm{TL},k} + \epsilon$.

These TL models are constructed using each of the 15 layers in VGG-16 and SoundNet. The label estimation from each layer model for new movie inputs is integrated using linearly weighted average for each dimension of label vectors. The averaging weight of i-th model, denoted by w_i, is determined according to label estimation accuracy of the model during the cross-validation in the model training, denoted by a_i. In particular, the averaging weight is represented as $w_i = a_i / \sum_j^{15} a_j$. This integration process produces a single series of label estimation to the new movie inputs.

Brain Decoding (BD).   BD is an effective method that estimates human perception and cognition induced by complex audiovisual inputs (Nishida and Nishimoto 2018). In the BD of this study, cognitive labels are directly estimated by analyzing measured voxel responses to movies. The model form is the same as in the vox2lab model of BTL except that the measured voxel responses are used instead of the predicted responses. The regression procedure is the same as the one used in BTL.

Results

Performance in Voxel-response Prediction

To confirm that our models appropriately extract audiovisual features of individual brains, we first examined the performance of cnn2vox and vox2vox models in terms of the prediction of voxel responses to web ad movies (Table 1). The prediction accuracy was evaluated by Pearson's correlation between the predicted and measured voxel responses in the test fMRI dataset, which have not been used for model training. Each accuracy was averaged over the whole cortex and across all participants.

The accuracy is higher when the predictions from both VGG-16 and SoundNet cnn2vox models were integrated than when only either of them was used. This is due to the fact that the VGG-16 and SoundNet features distinctively model audiovisual responses of voxels in different cortical regions. Indeed, the VGG-16 and SoundNet models accurately predict voxel responses in visual and auditory areas, respectively (see supplemental material). This is consistent with the previous findings that CNNs trained by datasets of specific modality predict voxel responses in brain regions involved in the processing of that modality (Güçlü and van Gerven 2017; 2015; Kell et al. 2018). Thus, our cnn2vox models well capture different modality of feature representations in the cortex. In addition, combining the prediction from the vox2vox models with the prediction from the cnn2vox models slightly improves the accuracy (Table 1; t-test, $p < 10^{-8}$), indicating that endogenous properties of voxel response represented by the vox2vox models augment the model ability to describe the brain feature representation by adding them to exogenous properties of voxel response.

Label Estimation Tasks

Task1: Estimation of Semantic Perception.   The first task is the estimation of semantic perception induced by each movie scene using the scene description data. All possible combinations of the cnn2vox and vox2vox models were calculated for BT, whereas for TL, all possible combinations of VGG-16 and SoundNet were calculated. Furthermore, for a fair comparison with BTL, the TL with the estimation from multiple time points was also computed (see Reference Methods). The performance of BTL and BD was evaluated using the mean estimation derived from the estimation of each participant model. Out of all the methods, the performance of BTL with the VGG-16 cnn2vox model is highest (Table 2). The significance of performance differences between methods was tested with a bootstrapping test (see supplemental material). The best performance in BTL is significantly higher than the best performance in TL and the performance in BD ($p < 0.0001$).

A previous study on brain decoding demonstrated that decoded contents varying brain to brain correlate with individual variability in semantic perception (Nishida and Nishimoto 2018). To test whether brain-by-brain variability in contents estimated by BTL has a similar correlation, we evaluated the correlation between individual variability in estimated contents and that in scene descriptions. In this test, BTL with VGG-16 cnn2vox and vox2vox models was used. The individual variability was evaluated by mean pairwise Pearson's correlation distance of estimation- or description-derived word2vec vectors between all possible pairs of brains or annotators. We used a 2-s window slid in 1-s steps to calculate the pairwise distance as is the case with the previous study (Nishida and Nishimoto 2018). Then, Pearson's and Spearman's correlation coefficients were calculated between the pairwise distances of estimated contents and those of scene descriptions. As a result, there are significant correlations not only for BD, consistent with the previous study (Nishida and Nishimoto 2018), but also for BTL (Figure 2; t-test, $p < 0.0001$). In addition, the correlation coefficients for

	CNN			Estimation performance				
	VGG-16	SoundNet	Vox2vox	Task1	Task2	Task3-1	Task3-2	Task4
BTL	✓			**0.546**	0.498	0.505	0.440	0.354
	✓		✓	**0.546**	0.500	0.502	0.444	0.360
		✓		0.146	0.234	0.094	0.223	0.244
		✓	✓	0.146	0.235	0.093	0.227	0.252
	✓	✓		0.536	0.516	0.446	0.376	0.381
	✓	✓	✓	0.536	**0.517**	0.447	0.375	**0.387**
			Time point					
TL	✓		Single	0.504	0.450	0.478	0.217	0.320
	✓		Multiple	0.537	0.482	**0.510**	0.288	0.334
		✓	Single	0.127	0.208	0.067	0.134	0.176
		✓	Multiple	0.143	0.232	0.065	0.166	0.213
	✓	✓	Single	0.504	0.471	0.475	0.245	0.336
	✓	✓	Multiple	0.535	0.501	0.502	0.303	0.365
BD				0.396	0.349	0.287	**0.468**	-

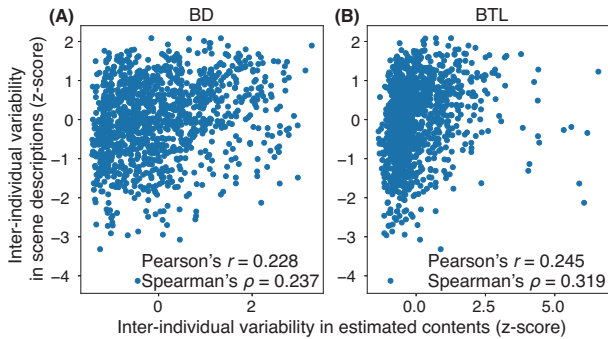Table 2: Estimation performance in all the tasks

Figure 2: Correlation of individual variability in semantic perception. Each dot depicts the variability in one scene.

BTL (Figure 2B) are rather higher than those for BD (Figure 2A). These results suggest that the models constructed from individual brains through BTL capture the individual variability in semantic perception.

Task2: Estimation of Impression Ratings. The second task is the estimation of impression perception induced by each scene of the web ad movies using the impression rating data. The estimation performance is mean values averaged over the 30 impression items. Out of all the methods, the performance of the BTL with the VGG-16 and SoundNet cnn2vox models and the vox2vox model is highest (Table 2). The best performance in BTL is significantly higher than the best performance in TL and the performance in BD (bootstrapping test, $p < 0.0001$). In this case, in contrast to the estimation of semantic perception, the SoundNet cnn2vox model improves the performance. This indicates that this task requires auditory information and suggests that BTL can extract auditory features from inputs and effectively use them for label estimation.

Task3: Estimation of Ad Effectiveness. The third task is the estimation of mass behavior that reflects the effectiveness of each clip of the web ad movies. The effectiveness was evaluated by two behavioral indices collected on the web: click rate (task3-1) and view completion rate (task3-2). For the estimation of click rate, of all the methods, the TL with the VGG-16 layers and the estimation from multiple time points shows the highest performance (Table 2, task3-1), although its performance is not significantly higher than the best-performing BTL (bootstrapping test, $p = 0.246$). Accordingly, in this case, BTL does not outperform TL. However, this is not surprising because the result that the estimation performance of BD is notably low in this task indicates that brain representations are not efficient for the estimation of click rate. Since BTL guides the estimation through brain representations, the inefficiency of brain representations negatively influences the estimation performance. Indeed, assuming that the vox2vox model makes feature representations more brain-like, the observation that the presence of the vox2vox model reduces the estimation performance (bootstrapping test, $p < 0.05$) may be consistent with this notion.

In contrast to the estimation of click rate, among all the methods, the highest performance in the estimation of view completion is shown by BD (Table 2, task3-2), although its performance is not significantly higher than the best-performing BTL (bootstrapping test, $p = 0.215$). Following this, BTL shows the second-highest performance despite the fact that TL shows much lower performance. This again indicates that BTL effectively uses brain representations to guide the estimation

Task4: Estimation of Ad Preference. The fourth task is the estimation of mass viewers' preference to each clip of the TV ad movies using the ad preference data. In this task, the vox2lab model of BTL was trained using movie datasets
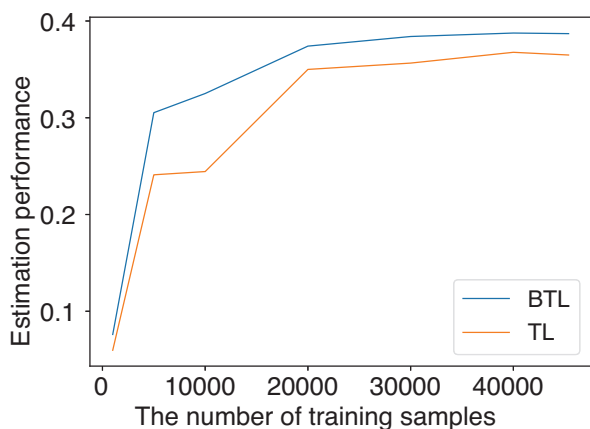
Figure 3: The effect of training sample size on the performance in ad preference estimation. The transition of estimation performance against the training sample size is shown separately for each of BTL (blue) and TL (orange).

completely separated from movie datasets for the training of the cnn2vox and vox2vox models. Hence, the BTL performance in this task reflects how the transformation between CNN and brain features by BTL is effectively generalized to novel datasets. The estimation performance was evaluated using 10-fold cross-validation.

Of all the methods, BTL with the VGG-16 and SoundNet cnn2vox models and the vox2vox model shows the highest performance (Table 2). The best performance in BTL is significantly higher than the best performance in TL (bootstrapping test, $p < 0.0001$). In this case, the presence of the vox2vox model improves the estimation performance.

In this task, the training of the vox2lab model in BTL and the TL model was conducted on data much larger than the ones used in the other tasks. Hence, to examine the impact of the size of the training data on label estimation, we evaluate the performance of the label estimation by changing the sample size of training data (Figure 3). The estimation performance of both the models gradually improves with the increase of the training sample size. Although the performance of each method seems to reach a plateau when all the samples were used, BTL still outperforms TL. This indicates that the superiority of BTL is not due to the difference between BTL and TL in terms of the sample size required for model training.

Conclusion

We proposed BTL for guiding the estimation of cognitive labels from complex audiovisual inputs through the transformation of CNN features to brain representations. To evaluate the performance of BTL and compare it with the performance of the other methods, we conducted four different types of recognition tasks. Our results in these recognition tasks demonstrated that BTL outperforms regular TL especially in the tasks in which brain representations are effective for the estimation. In addition, our results also suggest that the BTL estimation captures individual variability in percep-

tion. Therefore, BTL with CNNs potentially provides a powerful method to improve the generalization ability of CNNs in the estimation of human-like cognition and behavior, including individual differences.

The primary aim of this study was not to develop a machine-learning method that has state-of-the-art performance in pattern recognition. We rather aimed to examine the possibility that effective use of brain representations compensates for the shortcomings of conventional machine-learning methods. In particular, the present study focused on generalizability, which is one of the brain's prominent abilities that is better than machine learning, and we demonstrated that the addition of brain information improves the generalizability of machine learning. For this reason, we used standard CNNs (VGG-16 and SoundNet) and applied a simple TL algorithm with or without brain information. Nevertheless, it is also of interest to investigate whether BTL with more sophisticated CNNs and TL algorithms further improves the generalizability of machine learning.

The modeling in BTL is based on building voxelwise encoding models (Naselaris et al. 2011). In this line of research, there are many ongoing attempts to enhance modeling performance in terms of brain-response prediction (Çelik et al. 2019; Nunez-Elizalde, Huth, and Gallant 2019; Wen and Li 2016). Such enhancement potentially enables to utilize brain feature representations more efficiently by BTL, resulting in better performance in the estimation of human-like cognition and behavior, including their individual differences. In addition, the voxelwise modeling basically allows the prediction of brain response from any sort of feature representations. This means that BTL can also be applied to other machine-learning methods different from CNNs. Therefore, BTL has the potential to improve the generalization ability of overall machine-learning methods. We believe that BTL makes a key contribution to further developments in the field of machine learning.

References

Aytar, Y.; Vondrick, C.; and Torralba, A. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 892–900.

Çelik, E.; Dar, S. U. H.; Yılmaz, Ö.; Keleş, Ü.; and Çukur, T. 2019. Spatially informed voxelwise modeling for naturalistic fMRI experiments. *NeuroImage* 186:741–757.

Cimpoi, M.; Maji, S.; Kokkinos, I.; and Vedaldi, A. 2016. Deep filter banks for texture recognition, description, and segmentation. *International Journal of Computer Vision* 118(1):65–94.

Dmochowski, J. P.; Bezdek, M. A.; Abelson, B. P.; Johnson, J. S.; Schumacher, E. H.; and Parra, L. C. 2014. Audience preferences are predicted by temporal reliability of neural processing. *Nature Communications* 5:1–9.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, 647–655.

Falk, E. B.; Berkman, E. T.; and Lieberman, M. D. 2012. From neural responses to population behavior: neural focus group predicts population-level media effects. *Psychological Science* 23(5):439–445.

Fong, R. C.; Scheirer, W. J.; and Cox, D. D. 2018. Using human brain activity to guide machine learning. *Scientific Reports* 8(1):5397.

Fox, M. D., and Raichle, M. E. 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews. Neuroscience* 8(9):700–711.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 580–587.

Güçlü, U., and van Gerven, M. A. J. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience* 35(27):10005–10014.

Güçlü, U., and van Gerven, M. A. 2017. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage* 145:329–336.

Hampton, A. N., and O'doherty, J. P. 2007. Decoding the neural substrates of reward-related decision making with functional MRI. *Proceedings of the National Academy of Sciences of the United States of America* 104(4):1377–1382.

Kell, A. J.; Yamins, D. L.; Shook, E. N.; Norman-Haignere, S. V.; and McDermott, J. H. 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* 98(3):630–644.e16.

Kim, J.; Schultz, J.; Rohe, T.; Wallraven, C.; Lee, S.-W.; and Bulthoff, H. H. 2015. Abstract representations of associated emotions in the human brain. *Journal of Neuroscience* 35(14):5655–5663.

Knutson, B.; Rick, S.; Wimmer, G. E.; Prelec, D.; and Loewenstein, G. 2007. Neural predictors of purchases. *Neuron* 53(1):147–156.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their com-

positionality. *Advances in Neural Information Processing Systems* 26:3111–3119.

Mitchell, T. M.; Shinkareva, S. V.; Carlson, A.; Chang, K.-m.; Malave, V. L.; Mason, R. A.; and Just, M. A. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320(5880):1191–1195.

Naselaris, T.; Kay, K. N.; Nishimoto, S.; and Gallant, J. L. 2011. Encoding and decoding in fMRI. *NeuroImage* 56(2):400–410.

Nishida, S., and Nishimoto, S. 2018. Decoding naturalistic experiences from human brain activity via distributed representations of words. *NeuroImage* 180(A):232–242.

Nishimoto, S.; Vu, A. T.; Naselaris, T.; Benjamini, Y.; Yu, B.; and Gallant, J. L. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology* 21(19):1641–1646.

Nunez-Elizalde, A. O.; Huth, A. G.; and Gallant, J. L. 2019. Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage* 197:482–492.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 1717–1724.

Pan, S. J., and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1345–1359.

Peelen, M. V.; Atkinson, A. P.; and Vuilleumier, P. 2010. Supramodal representations of perceived emotions in the human brain. *Journal of Neuroscience* 30(30):10127–10134.

Razavian, A. S.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. CNN features off-the-shelf: An astounding baseline for recognition. In *CVPR*, 806–813.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv* 1409.1556.

Spampinato, C.; Palazzo, S.; Kavasidis, I.; Giordano, D.; Souly, N.; and Shah, M. 2017. Deep learning human mind for automated visual classification. In *CVPR*, 4503–4511.

Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; and Liu, C. 2018. A survey on deep transfer learning. In *ICANN*, 270–279.

Toshev, A., and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 1653–1660.

Wen, Z., and Li, Y. 2016. A spatial-constrained multi-target regression model for human brain activity prediction. *Applied Informatics* 3(1):10.

Xiao, J.; Ehinger, K. A.; Hays, J.; Torralba, A.; and Oliva, A. 2014. SUN database: Exploring a large collection of scene categories. *International Journal of Computer Vision* 119(1):3–22.