2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Decoding Brain Activity Features to Recognize Distorted Objects

Yuchou Chang and Mert Saritac, *Member, IEEE*

Abstract— **Brain decoding is able to make human interact with an external machine or robot for assisting patient's rehabilitation. Brain generic object recognition ability can be decoded through multiple neuroimaging modalities like functional magnetic resonance imaging (fMRI). On the other hand, external machine may wrongly recognize objects due to distorted noisy or blurring images caused by many factors, and therefore deteriorate performance of brain-machine interaction. In order to create better machine, generalization capability of human brain is transferred to classifier for enhancing classification accuracy of distorted images. Since homology existing between human and machine vision has been demonstrated, through decoding neural activity features of fMRI signals into feature units of convolutional neural network layers, an enhanced object recognition method is proposed to integrate brain activity into classifier for increasing classification accuracy. Experimental results show that the proposed method is able to enhance generalization capability of distorted object recognition.**

I. Introduction

Brain decoding is able to make human interact with a machine for assisting a patient's rehabilitation [1] through monitoring brain activities with brain-machine interface (BMI) [2]. Through machine learning analysis, brain activities can be decoded using neuroimaging modalities such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) to interpret mental states when people see, imagine, and dream. In human-machine interaction, joint human-machine cognitive system not only builds better performing machines [3], but also accurately identifies synchronous motion between human and machine for enabling more adept human-robot collaboration [4]. Recent study demonstrates that a homology exists between human and machine vision through hierarchies of individual areas from lower to higher visual cortex and the convolutional neural network (CNN) layers [5]. CNN has been used to study the human visual system from focusing on interpretable responses of single neurons to population-level descriptions of how visual information is represented and transformed for performing visual tasks [6].

The relationship between brain representations and hierarchical structures of layers of CNN makes it possible to not only predict the brain states in awake and anesthetized non-human primate resting-state functional magnetic resonance imaging (rsfMRI) data [7], but also transfer human knowledge to create better machines [8]. On the other hand, human vision system has strong generalization capability across a wide variety of input changes such as different illuminations, noise, and blur [9], in compared to the weak generalization ability of deep neural network. The difference of visual perception capabilities between human and CNN makes it difficult to build joint perception and joint actions for the future human-machine teaming [10]. For this reason, how to transfer generalization capability from human to machine is critical to design and create better machines.

Horikawa et al. have found that a brain decoding model trained on a limited set of object categories generalizes to decode arbitrary object categories [5]. Beside perception, imagined object categorization is also achieved by imagining about object images using the commonality of feature-level representations between perception and imagery. Therefore, arbitrary object categories imagined by human subjects can also be predicted from fMRI signals in the human visual cortical activities [5], even if a human subject only imagines the object and does not percept the object images. Based on human fMRI signals, a brain decoder may provide the possibility to make intelligent machine accurately recognize object in natural images with different image distortions in unstructured environment. For example, outdoor factory inspection robots in the haze environment have degraded performance due to blurred or distorted haze images in robotic vision system, leading to unsatisfactory results [11]. A brain decoding model may assist to enhance generic object recognition accuracy for this type of outdoor industry systems with imagined haze-free images.

In this paper, to enhance object recognition accuracy in distorted images, we investigate transferring the generalization capability of human visual system to a vision-based machine system with joint perception. Through a regression from brain signals to feature representations of distorted images in the pre-trained CNN model, a brain decoding model is built for improving classification accuracy of image distortion in machine vision. The remaining of the paper is structured as follows. Section II presents the related work about fMRI-based brain decoding and object recognition. The materials and methods are given in Section III. The experimental results and conclusion are presented in Sections IV and V.

II. Related Work

A. Brain Decoding with CNN Modeling

CNN has been used as a model of human visual system to gain insight and understanding about biological vision [6]. To decode brain activities, a decoder is trained using natural images, which are viewed or imagined by human subjects. Brain fMRI signals on visual cortex are acquired and regressed to CNN feature space, and then features are correlated to object categories [5]. The trained decoder has generalization capability without needing training data for zero-shot learning. Besides static natural images, brain is also able to represent

Yuchou Chang is with the Department of Computer and Information Science at University of Massachusetts Dartmouth, North Dartmouth, MA 02047 USA (phone: 508-999-8475; e-mail: ychang1@umassd.edu).

Mert Sarıtaç is with the Department of Computer and Information Science at University of Massachusetts Dartmouth, North Dartmouth, MA 02047 USA (e-mail: msaritac@umassd.edu).

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

dynamic natural vision in visual cortical areas which are correlated to different layers of CNN [12]. Wen et al use AlexNet [13], a classical CNN with 8 layers stacked architecture, to extract hierarchical visual features from the video clips stimuli, and fMRI signals are directly decoded to represent object categories in semantic space. Furthermore, machine learning algorithms have improved the performance of decoding brain activities and outperformed the traditional decoding methods [15].

B. *Improving Machine Performance with Assistance of Brain Decoding*

Due to strong generalization ability of human brain, brain imaging has been used for improving machine performance through decoding brain activities. For example, mind is read and decoded to transfer human visual capabilities to enhance computer vision tasks such as automated visual classification and allow machine to utilize human brain-based features [8, 16]. In [17], a brain imaging classification via EEG signals is proposed by integrating both implicit and explicit learning modalities. Accuracy is improved using multimodality information from both brain signals and image content. Under the CNN-based brain decoder network, a new concept of "brain-media" is proposed and both the brain domain and visual domain are exploited to increase the adversarial learning accuracy [14]. However, all those methods use EEG as brain imaging modality which has low spatial-resolution in compared to other imaging modalities like fMRI, and therefore it is difficult to build a correspondence between visual cortical areas and hierarchical representations of CNN. We will use fMRI to improve machine vision classification accuracy of distorted natural images with brain generalization capability support.

III. Materials and Methods

To transfer generalization capability of human brain to external machine, we used existing fMRI dataset which is acquired when human subjects see natural objects (e.g. car and coffee cup) in MR scanner. Then, as shown in Fig. 1, fMRI activity is used to train a decoder using training set of distorted images with blur and noise. The external machine (a humanoid robot in Fig. 1) with human brain activity decoding support is expected to have better ability of object recognition on distorted objects.

A. *Materials*

The fMRI scan is an expensive process in compared to other brain imaging modalities. For this reason, it's impossible to acquire millions of functional MR images, unlike cheap cost of natural image acquisition. We use BOLD5000 dataset [17] to build brain decoder. The BOLD5000 is a large-scale, slow event-related fMRI dataset collected on 4 subjects, who observed over 5000 images during 15 scanning sessions. Those images are extracted from Scene images [18], COCO dataset [19], and ImageNet dataset [20]. Total 10 regions of interests (ROIs) are acquired in BOLD5000 dataset and used to extract fMRI activities, which are parahippocampal place area (PPA), retrosplenial cortex (RSC), occipital place area (OPA), lateral occipital complex (LOC), early visual (EV) on left and right hemispheres, respectively. Those 5 types of ROIs are closely related to human visual system. In the dataset, BOLD signal was extracted from each voxel of 10 ROIs. We directly use the pre-processed BOLD signals extracted from those 10 ROIs for decoding brain activity. They can be downloaded from the OpenNeuro [30].

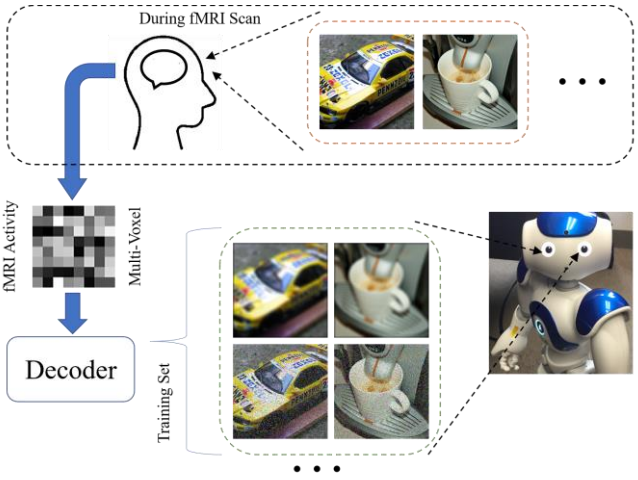

Figure 1. Transferring generalization capability of human brain to external machine via a decoder training by mapping fMRI activity data to distorted images features generated by pre-trained CNN.

In the BOLD5000 dataset, a slow event-related fMRI design enables a variety of visual feature, categories, and semantics encoded in neural representations on three image datasets, since hemodynamic response time restricts the fMRI's temporal resolution. Blood-oxygen-level-dependent (BOLD) signals are acquired by using a T2*-weighted gradient recalled echo-planar imaging (EPI) pulse sequence, with in-plane resolution = 2 × 2 mm; 106 × 106 matrix size, 2 mm slice thickness, field of view (FOV) = 212 mm, repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, and flip angle = 79 degrees [17]. To enhance image diversity, for those three image datasets, there are 1000 scene images, 2000 images from COCO dataset, and 1916 images from ImageNet dataset. The fMRIPrep [27-29] was used to preprocess the data with the ROI masks.

B. *Methods*

We manually selected 6 categories of objects from ImageNet dataset used in fMRI scans, since the set of images used in BOLD5000 is a small subset of all images in ImageNet. We do not use scene image and COCO images, because scene images have more semantic meanings rather than specific objects and COCO dataset has rich context information. The 6 categories of objects are "dog", "bird", "goat", "insect", "fish", and "monkey". Those 6 groups contain large categories. For example, the "dog" big category contains some sub-categories of dogs such as "Mexican hairless" and "Chihuahua" in the original ImageNet dataset. There are two reasons to build big category: (1) each sub-category has a small number of images used for fMRI scans and they are difficult for classification due to small sample size; (2) grouping sub-categories into a big category can be used for differentiating generalization ability human brain and CNN.

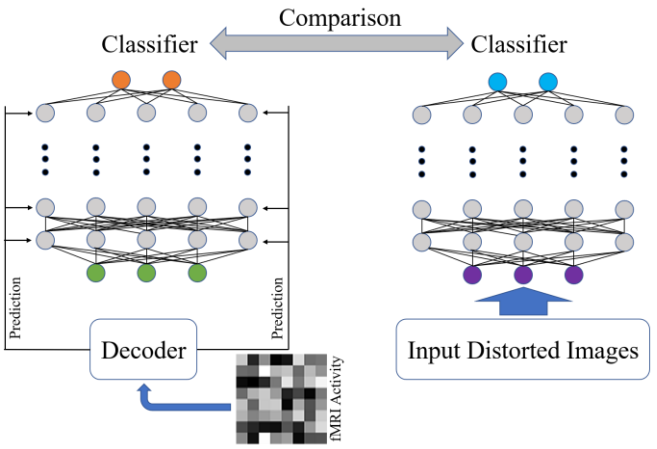

Figure 2. Classification comparison between brain decoded features and directly computed CNN features from distorted images. The fMRI activity using multi-voxel information on visual cortex ROIs is regressed through the decoder to CNN features of multiple layers.

To obtain distorted object images, both noise and blur are added into natural images used in BOLD5000 dataset. Noise and blur are added in images to evaluate the influence of distortions in object recognition performance for both human brain and CNN. Impulse noise [26] often deteriorates image quality due to defects of hardware or camera sensors, so impulse noise is added to images to simulate this type of defects in outdoor environment. The noise density is set as 0.6. In addition, a convolution filter with window sizes 30 is used to for blurring original images. The pre-trained AlexNet model is used to extract features of original images and distorted images from 5 convolution layers, which was trained with images in ImageNet. Similar to brain decoder in the reference [5], BOLD signals acquired from each of 10 ROIs are used to train a linear regression model as:

$$y = X\beta + \varepsilon, \qquad (1)$$

where y is predicted feature vectors of individual feature layers of CNN as shown in the left sub-figure of Fig. 2, X represents scalar values of fMRI signal magnitudes of ROI voxels, β denotes weights of voxels, and ε is bias. Similar to [5], we also select 1000 feature units to reduce computational costs, since 5 layers of AlexNet contain a large number of features. For comparison, distorted images are fed into the pre-trained network and obtain the same 1000 feature units as shown in the right sub-figure of Fig. 2. Without loss of generality, 1000 feature unites are randomly selected on 5 convolutional layers. Support vector machine (SVM) is used to classify both types of features.

IV. Results

Generalization capability is transferred from human brain to the classifier SVM through decoding brain activity into feature vectors on convolutional layers of CNN. Distorted image features decoded from brain and extracted from the pre-trained neural network model are compared for evaluating two types of features' performance. The experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board.

A. Pairwise Classification for Feature Evaluation

Pairwise classification is used to improve feature selection performance by classifying a pair of feature vectors [27]. In this work, pairwise classification on 6 categories of objects creates n(n-1)/2 = 15 SVM classifiers (where n is 6 here) for evaluating feature performance. We use MATLAB for programming pairwise classification and dimension reduction of principal component analysis (PCA). Cross validation (CV) is used for each SVM classifier, since CV is able to provide more information about classifier performance through resampling procedure. The 15% instances are used as test data for calculating cross-validated classification errors. The average of 15 classification accuracy is calculated as

$$Accuracy = 1 - \frac{\sum_{i=1}^{N} Classification\_Error_i}{N}, \qquad (2)$$

where $Classification\_Error_i$ is each classifier error and N is 15 for 6 categories. It is seen in the TABLE I that the brain decoded features-based classification accuracy outperforms CNN features on both undistorted and distorted images. Both noisy and blurring images degrade the classification accuracy and they have similar performance.

TABLE I. Average of Pairwise Classification Accuracy

	6 Categories of Objects in Undistorted / Distorted Images			
	Brain Decoded	***Undistorted***	***Noise***	***Blur***
Avera.	100%	83.59%	81.77%	82%

B. Dimension Reduction for Identifying Discriminant Ability

To identify discriminant ability between CNN features and brain decoded features from distorted images, PCA is used to reduce dimensions of both types of feature vectors from 1000 dimensions to 2 dimensions. As shown in Fig. 3, two principal components from brain decoded features show better discriminant ability than CNN features extracted from a convolution layer, since 6 categories are easily and clearly separated on the right figure. The stronger discriminant ability provided by brain decoded features explains that pairwise classification accuracy better than CNN features of distorted images.

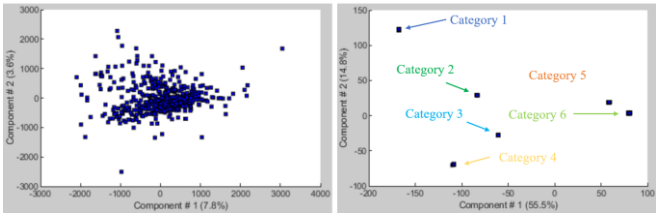

Figure 3. PCA-based dimension reduction on both CNN features (left figure) and brain decoded features (right figure) for observing their discriminant abilities. Brain decoded features show the stronger discriminant capability in compared to CNN features on the left figure, since 6 categories are easily and clearly separated.

C. Discussion from Transfer Learning Perspective

Human brain activity space and image feature space spanned by convolutional neural network have different data distributions, so that joint distribution is needed for transferring learning ability from source domain to target domain [28]. Since the homology between human and machine vision has been demonstrated [5], neural data on

brain visual cortical areas and feature units on layers of CNN may have similar distributions with linear regression on the training data for supervised classification, although both brain and CNN have different mechanism. Generalization capability using unsupervised classification on out-of-distribution (OOD) data may have worse performance than supervised mode [29], since CNN's ability to replicate human visual perception relies category labels for training. To overcome this limit, subspace-based transfer learning techniques [30] may be helpful to merge brain space and CNN feature space into a common subspace for transferring generalization capability. The selection of feature units on convolutional layers may also influence classification accuracy, because discriminative patches are helpful to enhance performance [26].

V. Conclusion

In conclusion, a distorted object recognition method using human brain feature decoding is proposed. Through transferring the generalization ability of human brain activity to CNN convolution layers, discriminant ability of brain decoded features enhances the classification accuracy of distorted images. To obtain the transferred generalization capability, a linear regression-based decoder is created for mapping fMRI signals on ROIs to feature units of convolution layers of CNN. Experimental results show that the proposed method improves the classification performance of different object images distorted by noise and blur. In the future work, brain-machine interaction will be studied using human brain decoded features .

References

[1] J. L. Contreras-Vidal, M. Bortole, F. Zhu, K. Nathan, A. Venkatakrishnan, G. E. Francisco, R. Soto, and J. L. Pons, "Neural decoding of robot-assisted gait during rehabilitation after stroke," *Am J Phys Med Rehabil*, vol. 97, no. 8, pp. 541-550, 2018.

[2] M. A. Lebedev and M. A. L. Nicolelis, "Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation," *Physiol Rev*, vol. 97, pp. 767-837, 2017.

[3] D. D. Woods, "Cognitive technologies: the design of joint human-machine cognitive systems," *AI Magazine*, vol. 6, no. 4, pp. 86, 1985.

[4] T. Iqbal, M. J. Gonzales, and L. D. Riek, "Joint action perception to enable fluent human-robot teamwork," *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2015.

[5] T. Horikawa and Y. Kamitani, "Generic decoding of seen and imagined objects using hierarchical visual features," *Nature Communications*, vol. 8, Article number: 15037, 2017.

[6] G. W. Lindsay, "Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future," *J Cogn Neurosci*, vol. 6, pp. 1-15, 2020.

[7] A. Grigis, J. Tasserie, V. Frouin, B. Jarraya, and L. Uhrig, "Predicting cortical signatures of consciousness using dynamic functional connectivity graph-convolutional neural networks," *bioRxiv - Neuroscience*, 2020.

[8] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, J. Schmidt, and M. Shah, "Decoding brain representations by multimodal learning of neural activity and visual features," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Early Access, 2020.

[9] R. Geirhos, C. R. Medina Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7549-7561, 2018.

[10] J. Laird, C. Ranganath, and D. Samuel Gershman "Future directions in human machine teaming workshop," *Department of Defense, Office of Prepublication and Security Review*, 2020.

[11] J. Li, L. Zhuo, H. Zhang, G. Li, and N. Xiong, "Effective data-driven technology for efficient vision-based outdoor industrial systems," *IEEE Trans. Industrial Informatics*, vol. 16, no. 7, pp. 4344-4354, 2020.

[12] H. Wen, J. Shi, Y. Zhang, K. H. Lu, J. Cao, and Z. Liu, "Neural encoding and decoding with deep learning for dynamic natural vision," *Cerebral Cortex*, vol. 28, no. 12, pp. 4136-4160, 2018.

[13] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, 2012.

[14] J. Jiang, A. Fares, and S. Zhong, "A brain-media deep framework towards seeing imaginations inside brains," *IEEE Trans. Multimedia*, Early Access, 2020.

[15] J. I. Glaser, A.S. Benjamin, R. H. Chowdhury, M. G. Perich, L. E. Miller, and K. P. Kording, "Machine learning for neural decoding," *eNeuro.*, vol. 7, no. 4, ENEURO.0506-19.2020, 2020.

[16] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] N. Chang, J. A. Pyles, A. Marcus, A. Gupta, M. J. Tarr, and E. M. Aminoff, "BOLD5000, a public fMRI dataset while viewing 5000 visual images," *Scientific Data*, vol. 6, Article number: 49, 2019.

[18] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: large-scale scene recognition from abbey to zoo," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485-3492, 2010.

[19] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *European Conference on Computer Vision*, pp. 740-755, 2014.

[20] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[21] T. Shongwey, A. J. Han Vinck, and H. C. Ferreira, "On impulse noise and its models," *18th IEEE International Symposium on Power Line Communications and Its Applications*, 2014.

[22] S. Li and S. Oh, "Improving feature selection performance using pairwise pre-evaluation," *BMC Bioinformatics*, v.17, article number: 312, 2016.

[23] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," *IEEE International Conference on Computer Vision*, 2013.

[24] T. Golan, P. C. Raju, and N. Kriegeskorte, "Controversial stimuli: Pitting neural networks against each other as models of human cognition," *Proceedings of the National Academy of Sciences of the United States of America*, vol.117, no.47, pp.29330-29337, 2020.

[25] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge and Data Engineering*, vol.22, no.10, pp.1345-1359, 2010.

[26] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[27] O. Esteban, D. Birman, M. Schaer, O. Koyejo, R.A. Poldrack, and K.J. Gorgolewski, "MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites," PloS ONE, vol.12, e0184661, 2017.

[28] O. Esteban, C.J. Markiewicz, R.W. Blair, C.A. Moodie, A. Ilkay Isik, A. Erramuzpe, et al., "fMRIPrep: a robust preprocessing pipeline for functional MRI," *Nature Methods*, vol.16, pp.111-116, 2019.

[29] O. Esteban, R. Ciric, K. Finc, R.W. Blair, C.J. Markiewicz, C.A. Moodie, et al., "Analysis of task-based functional MRI data preprocessed with fMRIPrep," *Nature Protocols*, vol.15, pp.2186-2202, 2020.

[30] OpenNeuro: A free and open platform for sharing MRI, MEG, EEG, iEEG, ECoG, ASL, and PET data. URL: https://openneuro.org/.