Repeated Exemplar Leakage in EEG Category Decoding

Jack A. Kilgallen (jkilgallen@protonmail.com)
Hamilton Institute, Maynooth University
Maynooth, Co. Kildare, Ireland

Barak A. Pearlmutter (barak@pearlmutter.net)
Department of Computer Science & Hamilton Institute, Maynooth University
Maynooth, Co. Kildare, Ireland

Jeffrey Mark Siskind (qobi@purdue.edu)
Elmore Family School of Electrical and Computer Engineering, Purdue University
West Lafayette, Indiana 47907-2035, United States of America

Abstract

Within neuroimaging research, it is a common practice to perform multiple trials using a single stimulus when working with noisy modalities such as electroencephalography (EEG). For many types of analyses, this practice is unproblematic. However, when attempting to decode object category information from EEG signals *(category decoding)*, we show that this practice can lead to a form of leakage that can inflate a model's performance when stimuli are shared across the training and test sets. We demonstrate this phenomenon by training several existing EEG decoding models on a dataset of EEG recordings from human subjects where multiple trials were recorded for each object within a category. We also develop a statistical framework to quantify the extent of this leakage. Our results reveal that per 1% increase above chance in the category decoding accuracy of a model trained on a dataset with repeated stimuli, the model's true generalization accuracy only increases by approximately 0.66%. This raises concerns about the validity of several EEG category decoding studies, and may have implications for brain computer interface (BCI) applications being developed on the basis of these studies.

Keywords: EEG; decoding; machine learning; leakage

Introduction

In neuroimaging studies it is a common practice to present a stimulus multiple times to a subject in order to reduce noise in the recorded signals. This practice is particularly common in electroencephalography (EEG) studies, where the signals are often noisy, and the signal-to-noise ratio can be improved by averaging over multiple trials such as in event-related potential (ERP) studies (Davis, 1939). However, when the analysis being performed is identifying the category of object observed by the subject *(category decoding)*, recording multiple trials of a single object from a category *(exemplar)* can lead to a form of leakage when exemplars are shared across the training and test sets. While the ongoing explosion of studies which apply machine learning techniques to neuroimaging data has yielded many promising results, there is a lack of awareness of this issue within the literature. In this study we demonstrate both the existence of this phenomenon and develop a statistical framework to quantify the extent of this leakage. We apply our framework to several existing EEG category decoding models within the literature which have been trained on a dataset which features repeated exemplars.

Materials and Methods

The Stanford University Dataset

The Stanford University Dataset (Kaneshiro et al., 2015) is a dataset of EEG recordings taken from 10 subjects while they viewed 72 images evenly distributed across 6 categories: Human Body (HB), Human Face (HF), Animal Body (AB), Animal Face (AF), Fruit/Vegetable (FV) and Inanimate Object (IO). To reduce the impact that noise would have on their analysis 72 trials were recorded per exemplar per subject and exemplars were presented in random order. This gives a total of 5,184 trials per participant. The data was recorded using a 128 channel EEG system with a sampling rate of 1 kHz. The EEG signals were then preprocessed using a high-pass fourth-order Butterworth filter to attenuate frequencies below 1 Hz, and a low-pass Chebyshev Type I filter to attenuate frequencies above 25 Hz. Ocular artifacts were removed using the Bell and Sejnowski (1995) Infomax independent component analysis algorithm, and finally the data was subsampled to 62.5 Hz to reduce the computational cost of the analysis. Coinciding with the publication of their paper the authors also made the preprocessed data available online.

Literature Review

To establish the extent to which the repeated exemplar leakage is present within the published literature, a reverse citation search was performed on the Stanford University Dataset. The search returned 19 articles[1] which made use of the Stanford University Dataset. These articles were then reviewed

[1] Ahmadieh et al. (2023); Bagchi and Bathula (2021, 2022); Bobe et al. (2018); Deng et al. (2023); Fares et al. (2020); Jiao et al. (2019); Kalafatovich and Lee (2021); Kalafatovich et al. (2020, 2023); Kaneshiro et al. (2015); Karimi-Rouzbahani et al. (2021); Karimi-Rouzbahani and Woolgar (2022); Kong et al. (2020); Luo et al. (2023); McCartney et al. (2022, 2019); Yavandhasani and Ghaderi (2022); Zheng et al. (2020)

to determine if the dataset was used to train a category decoding model, and if so whether their evaluation methodology was likely to be affected by the leakage. This revealed that out of 19 studies including the original which made use of the dataset, 13 were likely affected by the leakage.[2]

EEG Category Decoding Models

To capture the true effect of the leakage on published results we selected six of the EEG category decoding models found in our literature review for use in our experiment. The models selected were: Linear Discriminant Analysis (LDA) (Kaneshiro et al., 2015), Wide Convolutional Neural Network (WCNN) (Bagchi & Bathula, 2021), Attention-Driven Convolutional Neural Network (ADCNN) (Kalafatovich et al., 2020), EEG Convolutional Transformer (EEG-CT) (Bagchi & Bathula, 2022), Two-stream Convolutional Neural Network (TSCNN) (Kalafatovich et al., 2023), Reusable LSTM Network (RLN) (Deng et al., 2023). It should be stated that inclusion of a model in the analysis does not imply it is believed to be more likely affected by the leakage, merely that the model descriptions or code provided by the authors were sufficiently detailed for use in our analysis.

Evaluation Methodologies

To capture the difference in category decoding performance due to the leakage caused by repeated exemplars we used two different methodologies to evaluate each model's accuracy. One in which trials relating to each exemplar appear with equal frequency in the training and test sets *(the overlapping methodology)*, and another in which models are trained on trials relating to 11 exemplars per category, and then tested on the remaining exemplars *(the disjoint methodology)*. This allows us to generate two sets of accuracy results for our models, the overlapping accuracy which is inflated by leakage, and the disjoint accuracy which is not. 12-fold cross validation was used in both methodologies so that under the disjoint methodology each exemplar was used in a test set exactly once.

Statistical Framework

To capture the difference in performance, the accuracy results were aggregated at the subject level for each methodology to allow for a direct comparison in their results. The disjoint methodology accuracy was then predicted using a linear mixed model (LMM) with overlapping methodology as a fixed effect and model architecture and subject as random effects according to

$$\begin{gathered} Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_i + v_j + \varepsilon_{ij} \\ u_i \sim \mathcal{N}(0, \sigma_u^2) \quad v_j \sim \mathcal{N}(0, \sigma_v^2) \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{gathered} \tag{1}$$

In this model, Y_{ij} represents the percentage accuracy above chance without the leakage for each subject and model architecture combination. X_{ij}, the fixed effect, is the percentage accuracy above chance when the leakage is present. The random effects u_i and v_j capture the variability across subjects and model architectures, respectively. The error term ε_{ij} accounts for the residual variability. The primary term of interest is β_1 which explains the expected increase in disjoint accuracy given a 1% increase in overlapping accuracy.

Results

Table 1 gives a summary of our fitted model. The value fitted for the β_1 parameter indicated that per 1% increase above chance accuracy in a model's reported accuracy the true generalization accuracy only increases by 0.6614%. This indicates that there is a significant and systematic difference in accuracies due to the leakage introduced by sharing exemplars across the training and test set. Given that the highest reported accuracy is approximately 54.28% (Kalafatovich et al., 2023) this means that the accuracy of some models may have been inflated by approximately 12.73%.

Additionally, Fig. 1 breaks down the category decoding accuracy by stimulus category and reveals a substantial difference in performance for each classifier under the two methodologies on the individual categories. In particular, it appears the accuracy of these models is largely driven by the performance on the Human Face category. This raises the question of how feasible it is for a classification algorithm to learn the representation of a category as contrived as Inanimate Objects, given only trials relating to 11 exemplars as input. Moreover, it can be seen in the figure that there is a significantly higher standard deviation in the accuracy of the models under the disjoint methodology. This suggests that the accuracy of each of the models is highly dependent on the stimulus presented. This raises further concerns about the generalizability of the models to new stimuli, and the feasibility of applying such models to BCI applications.

Acknowledgments

This publication emanated from research supported in part by Science Foundation Ireland (SFI) research grants 16/RI/3399, 18/CRT/6049, and 20/FFP-P/8853, and the United States National Science Foundation Grant Number 1734938-IIS. We thank Hari M. Bharadwaj for suggesting that we look into the issue discussed in this paper as a potential confound.

Table 1: LMM Results Summary

Effect	Estimate	p-value
Fixed Effects		
Disjoint Accuracy	0.6614	2.306×10^{-14}
Random Effects		
Variance: Subject	1.4680	
Variance: Model	1.4187	

[2] Ahmadieh et al. (2023); Bagchi and Bathula (2021, 2022); Bobe et al. (2018); Deng et al. (2023); Fares et al. (2020); Jiao et al. (2019); Kalafatovich and Lee (2021); Kalafatovich et al. (2020, 2023); Luo et al. (2023); Yavandhasani and Ghaderi (2022); Zheng et al. (2020)

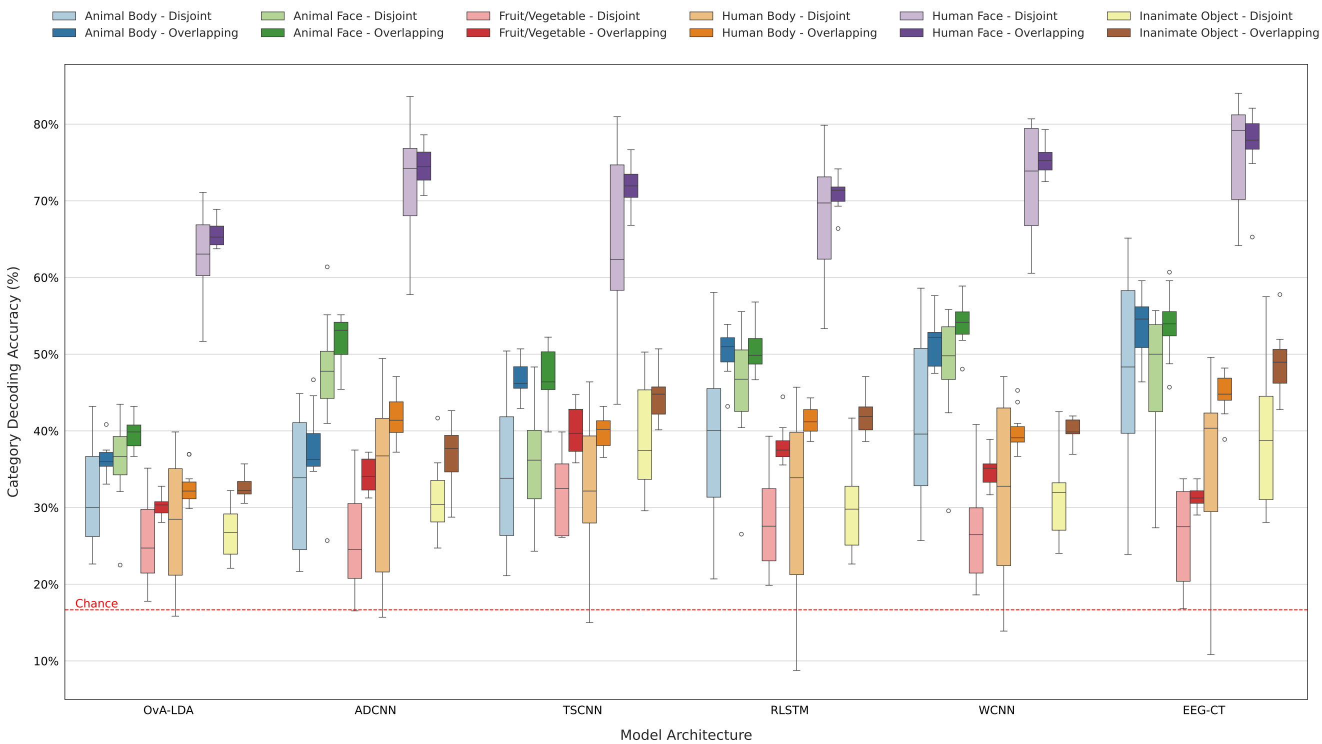

Figure 1: Comparison of accuracy under disjoint *vs* overlapping exemplar methodologies for each model and category

References

Ahmadieh, H., Gassemi, F., & Moradi, M. H. (2023, October). A hybrid deep learning framework for automated visual image classification using EEG signals. *Neural Computing and Applications*, *35*(28), 20989–21005. doi: 10.1007/s00521-023-08870-w

Bagchi, S., & Bathula, D. R. (2021, August). Adequately wide 1D CNN facilitates improved EEG based visual object recognition. In *2021 29th European Signal Processing Conference (EUSIPCO)* (pp. 1276–1280). Dublin, Ireland: IEEE. doi: 10.23919/EUSIPCO54536.2021.9615945

Bagchi, S., & Bathula, D. R. (2022, September). EEG-ConvTransformer for single-trial EEG-based visual stimulus classification. *Pattern Recognition*, *129*, 108757. doi: 10.1016/j.patcog.2022.108757

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–59.

Bobe, A. S., Alekseev, A. S., Komarova, M. V., & Fastovets, D. (2018, November). Single-trial ERP feature extraction and classification for visual object recognition task. In *2018 Engineering and Telecommunication (EnT-MIPT)* (pp. 188–192). Moscow, Russia: IEEE. doi: 10.1109/EnT-MIPT.2018.00049

Davis, P. A. (1939). Effects of acoustic stimuli on the waking human brain. *Journal of neurophysiology*, *2*(6), 494–499.

Deng, Y., Ding, S., Li, W., Lai, Q., & Cao, L. (2023, April). EEG-based visual stimuli classification via reusable LSTM. *Biomedical Signal Processing and Control*, *82*, 104588. doi: 10.1016/j.bspc.2023.104588

Fares, A., Zhong, S.-h., & Jiang, J. (2020, October). Brain-media: A dual conditioned and lateralization supported GAN (DCLS-GAN) towards visualization of image-evoked brain activities. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 1764–1772). Seattle WA USA: ACM. doi: 10.1145/3394171.3413858

Jiao, Z., You, H., Yang, F., Li, X., Zhang, H., & Shen, D. (2019, August). Decoding EEG by visual-guided deep neural networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 1387–1393). Macao, China. doi: 10.24963/ijcai.2019/192

Kalafatovich, J., & Lee, M. (2021, February). Subject-independent object classification based on convolutional neural network from EEG signals. In *2021 9th International Winter Conference on Brain-Computer Interface (BCI)* (pp. 1–4). Gangwon, Korea (South): IEEE. doi: 10.1109/BCI51272.2021.9385333

Kalafatovich, J., Lee, M., & Lee, S.-W. (2020, October). Decoding visual recognition of objects from EEG signals based on attention-driven convolutional neural network. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2985–2990). Toronto, ON, Canada. doi: 10.1109/SMC42975.2020.9283434

Kalafatovich, J., Lee, M., & Lee, S.-W. (2023). Learning spatiotemporal graph representations for visual perception using EEG signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 97–108. doi: 10.1109/TNSRE.2022.3217344

Kaneshiro, B., Perreau Guimaraes, M., Kim, H.-S., Norcia, A. M., & Suppes, P. (2015, August). A representational similarity analysis of the dynamics of object processing using single-trial EEG classification. *PLOS ONE*, *10*(8), e0135697. doi: 10.1371/journal.pone.0135697

Karimi-Rouzbahani, H., Shahmohammadi, M., Vahab, E., Setayeshi, S., & Carlson, T. (2021, August). Temporal variabilities provide additional category-related information in object category decoding: A systematic comparison of informative EEG features. *Neural Computation*, 1–46. doi: 10.1162/neco_a_01436

Karimi-Rouzbahani, H., & Woolgar, A. (2022, March). When the whole is less than the sum of its parts: Maximum object category information and behavioral prediction in multiscale activation patterns. *Frontiers in Neuroscience*, *16*, 825746. doi: 10.3389/fnins.2022.825746

Kong, N. C. L., Kaneshiro, B., Yamins, D. L. K., & Norcia, A. M. (2020, July). Time-resolved correspondences between deep neural network layers and EEG measurements in object processing. *Vision Research*, *172*, 27–45. doi: 10.1016/j.visres.2020.04.005

Luo, J., Cui, W., Xu, S., Wang, L., Li, X., Liao, X., & Li, Y. (2023). A dual-branch spatio-temporal-spectral transformer feature fusion network for EEG-based visual recognition. *IEEE Transactions on Industrial Informatics*, 1–11. doi: 10.1109/TII.2023.3280560

McCartney, B., Devereux, B., & Martinez-del Rincon, J. (2022, June). A zero-shot deep metric learning approach to Brain-Computer Interfaces for image retrieval. *Knowledge-Based Systems*, *246*, 108556. doi: 10.1016/j.knosys.2022.108556

McCartney, B., Martinez-del Rincon, J., Devereux, B., & Murphy, B. (2019, March). Towards a real-world brain-computer interface for image retrieval. *bioRxiv preprint*. Retrieved 2023-09-04, from https://www.biorxiv.org/content/10.1101/576983v1 doi: 10.1101/576983

Yavandhasani, M., & Ghaderi, F. (2022, July). Visual object recognition from single-trial EEG signals using machine learning wrapper techniques. *IEEE Transactions on Biomedical Engineering*, *69*(7), 2176–2183. doi: 10.1109/TBME.2021.3138157

Zheng, X., Cao, Z., & Bai, Q. (2020). An evoked potential-guided deep learning brain representation for visual classification. In *International conference on neural information processing (iconip)* (Vol. 1333, pp. 54–61). Cham: Springer International Publishing. doi: 10.1007/978-3-030-63823-8_7