Comments

Still an Ineffective Method With Supertrials/ERPs—Comments on "Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features"

Hari M Bharadwaj, Ronnie B. Wilbur, and Jeffrey Mark Siskind, *Senior Member, IEEE*

***Abstract*—A recent paper claims that a newly proposed method classifies EEG data recorded from subjects viewing ImageNet stimuli better than two prior methods. However, the analysis used to support that claim is based on confounded data. We repeat the analysis on a large new dataset that is free from that confound. Training and testing on aggregated supertrials derived by summing trials demonstrates that the two prior methods achieve statistically significant above-chance accuracy while the newly proposed method does not.**

***Index Terms*—EEG, human vision, neuroimaging, neuroscience, brain-computer interface, object classification.**

I. Introduction

A recent paper [15] presents a novel neural-network architecture, EEGChannelNet, for determining object class from EEG signals recorded from human subjects observing ImageNet [5] images as stimuli. *Inter alia*, it claims that EEGChannelNet can decode object class from EEG signals better than prior work, in particular two prior classifiers: EEGNet [9], and SyncNet [11].

> *Finally, we compare classification performance achieved by our EEG encoder and other state-of-the-art methods, namely [3][1] [21], [22], using high-frequency gamma band data, i.e., 55-95 Hz. EEG classification accuracy on the test split is given in Tab. 2 and shows that our approach reaches an average classification accuracy of 48.1%, outperforming previous methods, such as EEGNet, which, only achieves a maximum accuracy of 31.9%.*
>
> [15, Section 7.3 last paragraph, footnote, citations, and table in the original]

In their Table 2, they claim that EEGNet obtains 31.9% accuracy and SyncNet obtains 31.7%. These accuracy numbers were obtained on the dataset reported in Spampinato et al. [17], containing 50 ImageNet stimuli for each of 40 classes. Li et al. [10] already demonstrated that this dataset is confounded due to its nonrandomized collection method. Ahmed et al. [2, inline unnumbered tables 9, 10, 15, and 16] demonstrates that on a new dataset of the same size with randomized trials, EEGChannelNet and SyncNet yield chance accuracy, while EEGNet yields chance accuracy on some subjects and accuracy marginally above chance on some other subjects. Ahmed et al. [1, Table 2] already demonstrated that on a new dataset with randomized trials that is 20× larger, 1000 ImageNet stimuli for essentially the same 40 classes, EEGChannelNet and SyncNet yield chance accuracy, while EEGNet yields 7.0% accuracy, a value that is statistically significant above chance.

All of the above trained and tested classifiers on individual trials: a recording from a single stimulus presentation. Prior work has investigated training and testing EEG classifiers on supertrials or ERPs that are obtained by averaging across multiple trials independently per channel and per time point [4], [7], [8], [18]. This can lead to improved classification accuracy by increasing the signal-to-noise ratio. Here, we investigate how this can impact the claims from Palazzo et al. [15] regarding EEGChannelNet, EEGNet, and SyncNet.

II. Method

We preprocess the dataset from Ahmed et al. [1] exactly as described in Ahmed et al. [1].[1] This includes z-scoring the data on a per-channel and per-run basis. After breaking the data up into individual trials and randomly shuffling as described in Ahmed et al. [1], we partition the entire dataset into disjoint covering sets of N trials, on a per-class basis. Each partition is averaged per channel and per time point to yield a supertrial. The dataset is then split into five equal-sized disjoint sets of supertrials that cover the dataset. We then replicate the study from Ahmed et al. [1, Table 2] with supertrials instead of trials, with five-fold leave-one-portion-out cross validation. Critically, with this construction, the supertrials in the training and test sets are constructed from disjoint sets of trials. We repeat this same method on all six subjects of the image rapid event data from Li et al. [10] and replicate the study of Ahmed et al. [2, inline unnumbered table 9] with supertrials instead of trials, with five-fold leave-one-portion-out cross validation.[2]

Manuscript received 6 August 2021; revised 22 April 2023; accepted 27 June 2023. Date of publication 4 July 2023; date of current version 3 October 2023. This work was supported in part by US National Science Foundation under Grant 1734938-IIS. Recommended for acceptance by K. M. Lee (EIC). *(Corresponding author: Jeffrey Mark Siskind.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Purdue University IRB under Protocol No. 1208012608.

Hari M Bharadwaj is with the Department of Communication Science and Disorders, University of Pittsburgh, Pittsburgh, PA 15260 USA (e-mail: hari.bharadwaj@pitt.edu).

Ronnie B. Wilbur is with the Department of Speech, Language, and Hearing Sciences and the Department of Linguistics, Purdue University, West Lafayette, IN 47907 USA (e-mail: wilbur@purdue.edu).

Jeffrey Mark Siskind is with the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: qobi@purdue.edu).

Digital Object Identifier 10.1109/TPAMI.2023.3292062

[1]All code and data needed to replicate all results in this paper are available at https://ieee-dataport.org/open-access/dataset-perils-and-pitfalls-block-design-eeg-classification-experiments and https://ieee-dataport.org/open-access/dataset-object-classification-randomized-eeg-trials.

[2]To mimic Spampinato et al. [17], the studies of Ahmed et al. [1], [2] formed ten portions, with one portion per fold arbitrarily labeled as a validation set and one portion per fold arbitrarily labeled as a test set. But these were treated identically and averaged as we did not perform hyperparameter search. This has minor consequences discussed below.

TABLE I
REPLICATION OF THE ANALYSIS FROM AHMED ET AL. [1, TABLE 2] (LEFT) AND AHMED ET AL. [2, INLINE UNNUMBERED TABLE 9] (RIGHT) FOR VARIOUS SIZES N OF SUPERTRIALS

N	LSTM	k-NN	SVM	MLP	1D CNN	EEGNet	SyncNet	EEGChannelNet
1	2.2%	2.1%	5.0%*	2.5%	5.1%*	7.0%*	2.5%	2.5%
2	2.4%	2.5%	5.5%*	2.5%	6.9%*	9.6%*	2.4%	2.5%
4	2.0%	2.2%	6.3%*	2.4%	9.4%*	12.8%*	3.5%*	2.5%
5	2.3%	2.4%	5.9%*	2.7%	8.8%*	12.9%*	3.8%*	2.5%
8	2.8%	2.5%	3.5%*	2.8%	3.8%*	15.6%*	3.8%*	2.5%
10	2.6%	2.6%	2.9%	2.5%	3.5%*	17.0%*	4.2%*	2.6%
20	2.3%	2.1%	3.0%	2.8%	2.4%	17.6%*	3.7%*	2.7%
25	2.4%	3.4%	2.9%	2.6%	2.6%	16.6%*	4.0%*	2.4%
40	2.1%	2.8%	2.3%	2.3%	2.9%	14.6%*	3.1%	2.5%
50	2.8%	1.6%	2.1%	2.3%	2.4%	13.9%*	3.0%	2.3%
100	2.0%	1.0%	1.8%	3.0%	3.0%	6.5%*	1.5%	3.8%

N	subject	LSTM	k-NN	SVM	MLP	1D CNN	EEGNet	SyncNet	EEGChannelNet
1	1	1.2%	1.4%	2.7%	1.8%	2.4%	1.0%	1.9%	2.5%
	2	1.5%	1.8%	3.5%*	1.0%	3.1%	3.3%	3.4%	2.5%
	3	1.2%	1.6%	2.1%	1.1%	3.2%	4.4%*	3.5%*	2.5%
	4	2.0%	1.2%	3.1%	0.9%	2.4%	2.7%	2.3%	2.4%
	5	1.3%	1.8%	2.1%	1.9%	2.0%	1.8%	2.5%	2.5%
	6	1.1%	1.2%	3.3%	1.2%	2.7%	4.3%*	2.7%	2.1%
2	1	1.8%	1.8%	3.8%	1.8%	0.9%	1.1%	2.5%	2.1%
	2	1.1%	1.6%	4.3%*	0.9%	2.1%	4.9%*	2.3%	2.6%
	3	1.5%	2.6%	2.1%	1.4%	1.7%	4.7%*	2.0%	2.0%
	4	2.1%	2.4%	2.8%	1.6%	1.9%	2.8%	2.5%	2.2%
	5	1.1%	0.8%	4.5%*	1.5%	1.5%	1.6%	2.1%	2.1%
	6	1.4%	1.5%	4.4%*	1.4%	1.6%	4.3%*	3.1%	2.1%
5	1	1.8%	1.2%	7.0%*	0.8%	0.8%	0.8%	2.0%	1.8%
	2	2.3%	1.2%	6.5%*	0.5%	1.5%	4.3%	2.8%	1.8%
	3	1.3%	1.0%	6.8%*	1.3%	1.5%	2.8%	1.0%	1.0%
	4	1.5%	0.2%	7.8%*	0.3%	1.5%	3.8%	1.0%	3.0%
	5	0.8%	1.0%	6.8%*	0.8%	1.3%	0.8%	0.3%	1.8%
	6	0.3%	1.2%	7.3%*	1.5%	1.0%	2.5%	1.3%	2.0%
10	1	0.5%	1.5%	2.5%	0.0%	0.5%	0.5%	1.5%	1.0%
	2	0.0%	1.0%	4.5%	0.0%	0.0%	1.5%	1.0%	2.5%
	3	0.0%	3.0%	4.0%	0.0%	1.0%	2.0%	1.5%	2.0%
	4	1.0%	0.5%	3.5%	0.0%	1.0%	3.0%	1.0%	1.5%
	5	0.5%	1.5%	2.5%	0.0%	0.5%	1.0%	0.0%	2.0%
	6	1.0%	1.0%	4.0%	0.0%	0.0%	2.0%	1.0%	2.0%

The first row ($N = 1$) is from the original papers without supertrials. Starred values indicate statistical significance above chance ($p < 0.005$) by a binomial cmf. Note that when N gets larger, the number of test samples gets smaller, increasing quantization noise in the accuracy estimates, thus requiring higher accuracy to achieve significance. The results for EEGNet and SyncNet use our own previously reported [1] implementations in Pytorch translated from the original in Keras and Tensorflow. The results for EEGChannelNet used the code from https://github.com/perceivelab/eeg visual classification, modified slightly to support 96 channels instead of 128 and a duration of 512 samples instead of 440.

III. RESULTS

We evaluate this method for supertrials of various sizes (Table I). Supertrial sizes N were chosen to divide 1000 for the dataset from Ahmed et al. [1] and 50 for the dataset from Li et al. [10] so as to partition the trials per class.[3] Note that there is a tradeoff. Larger N can increase the S/N ratio further but yields smaller training and test sets. This leads to a behavior where there is an approximate unimodal concave region around a local maximum. The sweet spot can vary by classifier and size of dataset. For the dataset from Ahmed et al. [1], it appears to be around $N = 4$ for SVM and 1D CNN, $N = 20$ for EEGNet, and $N = 10$ for SyncNet, while for the dataset from Li et al. [10], it appears to be around $N = 5$. Note that the larger size of the dataset from Ahmed et al. [1] can afford a larger N. Also note that supertrials can allow SVM, 1D CNN, EEGNet, and SyncNet to achieve statistically significant above-chance classification accuracy, but appear unable to help LSTM, k-NN, MLP, and EEGChannelNet. This largely concurs with Ahmed et al. [1, Tables 2 and 3a, Fig. 3].

IV. DISCUSSION

Ahmed et al. [1, Fig. 3b, Table 3b] previously observed diminishing returns with larger training sets (larger than about 60% of their dataset of 40,000 samples). It might be the case that with supertrials, one could achieve even higher classification accuracy if one had even more training data. In fact, the observed sweet spots might occur simply because of the tradeoff, i.e., positions above which there are too few training samples despite the fact that the S/N ratio gets better. It might be that if one had more data, an even higher value of N would be better.

[3]Note that we perform an unweighted average of accuracy over all validation and test sets over folds. For the dataset from Ahmed et al. [1], for $N = 8$ and $N = 40$, and for the dataset from Li et al. [10], for $N = 2$ and $N = 10$, this means that the number of supertrials is not divisible by 10 and cannot be partitioned equally into validation and test sets. So for these analyses, the number of supertrials in each of the training, validation, and test sets varies across fold and the number of supertrials in the validation and test sets may be unequal within fold. The unweighted average might introduce a small bias. For all other analyses, the number of supertrials is divisible by 10, so exactly 80% are taken as the training set and exactly 10% are taken for each of the validation and test sets in each fold, so there is no bias. Nonetheless, for all analyses, in each fold, each trial occurs in exactly one supertrial and each supertrial occurs in exactly one of the training, validation, or test sets.

Here, we sample trials to form supertrials without replacement, i.e., each trial is in exactly one supertrial. But one could sample with replacement, allowing trials to be in more than one supertrial in combination with different sibling trials. This would allow forming a larger set of supertrials from the same set of trials. Greene and Hansen [7] use this approach. However, to maintain independence between training and test data, it is critical to ensure that the respective supertrials are formed from disjoint sets of trials; it is not clear if this was done in [7].

Here, we form supertrials by aggregating trials from a single subject. One could form supertrials by aggregating trials from multiple subjects.

Here, we form supertrials for both the training and test sets by aggregating the same number N of trials. One could aggregate different numbers of trials to form supertrials in the training and test sets, potentially using supertrials only for training but still performing single-trial classification for test.

Here, we aggregate supertrials by unweighted average in the time domain. One could average in the frequency domain, potentially considering only certain bands (e.g., induced responses), weighting some samples or bands more than others, or more generally averaging some nonlinear transform, learned or hard-coded, of single trials. Even more generally, one could employ classifiers with a multi-trial input feature space. Learned classifiers employing this likely would need to tie parameters across different trials in a supertrial, both to achieve position invariance of trials within a supertrial and to avoid over-fitting.

Differing sweet spots between different classifiers might be due to different classifiers having differing sensitivity to S/N ratio, differing sensitivities to training-set size, or some combination of the two. One could design analyses to tease these apart by varying one without the other.

We leave exploration of these variants to future work.

It is unclear why some methods, like SVM, the 1D CNN, EEGNet, and SyncNet, are able to achieve above-chance classification accuracy, and other methods, like EEGChannelNet, are not. Generally, there is little understanding in the community of why deep-learning methods work when they do and why they don't when they don't. That said, we offer some thoughts on the deficiencies of EEGChannelNet. EEGChannelNet employs 1D spatial convolution. This appears ill-motivated. Generally, one applies temporal/spatial convolution to model temporal/spatial invariance. However, there is little evidence that brain processing is spatially invariant. Further, the brain is 3D and the EEG

cap is 2D, but EEGChannelNet employs 1D spatial convolution. This seems odd. It might be reasonable to employ 2D analysis in the space of the EEG electrode array, or to attempt to reconstruct and analyze the underlying 3D signal from the 2D signal, but EEGChannelNet does neither. Finally, Palazzo et al. [15] do not specify the mapping from electrode topography to channel index. Thus we are unable to ascertain whether adjacent channel indices are spatially adjacent. For larger convolution kernels, they are almost certainly not.

While a block design, with adjacent trials coming from the same stimulus class, might be appropriate for some kinds of analyses [3], [6], [12], [13], [16], as they can increase signal-to-noise ratio, it is not appropriate for classification, either when temporally close trials are used in both the training and test sets or when the same stimulus-class order is used in both the training and test sets [10].

At this point, we know of no nonconfounded dataset and no training regimen on which EEGChannelNet achieves above-chance performance either with or without supertrials. We do know of several datasets on which several other classifiers in a variety of training regimens, both with and without supertrials, do achieve above-chance performance [1], [2], [10]. It is always possible that results may change with new methods and new datasets yet to be tried or yet to be collected.

V. Conclusion

Palazzo et al. [14] claim that the data collected in Li et al. [10] lacks class information due to lack of subject attentiveness during long sessions, and that classification failure is based on this. The data-collection method in Ahmed et al. [1] was similar to that in Li et al. [10], the sole differences being that (a) each stimulus was presented for 2 s with 1 s of blanking instead of 0.5 s with no blanking, (b) there were 400 stimuli per run for a total run length of 20:20, including 10 s of blanking at the start and end of each run instead of 2000 stimuli per run for a total run length of 23:20, including 10 s of blanking after every 50 trials, and (c) each session contained approximately 10 runs instead of 4–8 runs. In both cases, sessions lasted three to six hours, including capping, uncapping, and breaks between runs upon subject request. Table I demonstrates that the data of Ahmed et al. [1] and Li et al. [10] do contain class information; it is just that some classifiers successfully extract it and some do not. Thus our results here refute their claim.

Table I further demonstrates that:

- With and without supertrials, EEGChannelNet yields chance accuracy on a nonconfounded dataset 20× larger than that of [15].
- For some amounts of supertrial aggregation, EEGNet and SyncNet yield above chance accuracy.

This refutes the claim in [15] that EEGChannelNet outperforms EEGNet and SyncNet. Moreover, to the best of our knowledge, the classification accuracy of 17.5% obtained by EEGNet with $N = 20$ is the highest reported for a 40-class EEG classification task on ImageNet stimuli. Finally, this demonstrates that the datasets of Ahmed et al. [1] and Li et al. [10] do contain class information in the EEG signal; EEGNet, to some extent, and SyncNet, to a lessor extent, can extract that class information. EEGChannelNet cannot.

References

[1] H. Ahmed, R. B. Wilbur, H. M. Bharadwaj, and J. M. Siskind, "Object classification from randomized EEG trials," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3845–3854.

[2] H. Ahmed, R. B. Wilbur, H. M. Bharadwaj, and J. M. Siskind, "Confounds in the data—Comments on 'Decoding brain representations by multimodal learning of neural activity and visual features'," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9217–9220, Dec. 2022, doi: 10.1109/TPAMI.2021.3121268.

[3] P. A. Bandettini and R. W. Cox, "Event-related fMRI contrast when using constant interstimulus interval: Theory and experiment," *Magn. Reson. Med.: An Official J. Int. Soc. Magn. Reson. Med.*, vol. 43, no. 4, pp. 540–548, 2000.

[4] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence," *Sci. Rep.*, vol. 6, no. 1, pp. 1–13, 2016.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[6] L. E. Ethridge, S. Brahmbhatt, Y. Gao, J. E. Mcdowell, and B. A. Clementz, "Consider the context: Blocked versus interleaved presentation of antisaccade trials," *Psychophysiology*, vol. 46, no. 5, pp. 1100–1107, 2009.

[7] M. R. Greene and B. C. Hansen, "Disentangling the independent contributions of visual and conceptual features to the spatiotemporal dynamics of scene categorization," *J. Neurosci.*, vol. 40, no. 27, pp. 5283–5299, 2020.

[8] L. Isik, E. M. Meyers, J. Z. Leibo, and T. Poggio, "The dynamics of invariant object recognition in the human visual system," *J. Neuriophysiol.*, vol. 111, no. 1, pp. 91–102, 2014.

[9] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, 2018, Art. no. 056013.

[10] R. Li et al., "The perils and pitfalls of block design for EEG classification experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 316–333, Jan. 2021, doi: 10.1109/TPAMI.2020.2973153.

[11] Y. Li et al., "Targeting EEG/LFP synchrony with neural nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4620–4630.

[12] S. J. Luck, *An Introduction to the Event-Related Potential Technique*. Cambridge, MA, USA: MIT Press, 2014.

[13] F. M. Miezin, L. Maccotta, J. Ollinger, S. Petersen, and R. Buckner, "Characterizing the hemodynamic response: Effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing," *Neuroimage*, vol. 11, no. 6, pp. 735–759, 2000.

[14] S. Palazzo, C. Spampinato, J. Schmidt, I. Kavasidis, D. Giordano, and M. Shah, "Correct block-design experiments mitigate temporal correlation bias in EEG classification," 2020, *arXiv: 2012.03849*.

[15] 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 Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3833–3849, Nov. 2021, doi: 10.1109/TPAMI.2020.2995909.

[16] N. A. Roque, T. J. Wright, and W. R. Boot, "Do different attention capture paradigms measure different types of capture?," *Attention, Percep., Psychophys.*, vol. 78, no. 7, pp. 2014–2030, 2016.

[17] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6809–6817.

[18] X. Zheng, Z. Cao, and Q. Bai, "An evoked potential-guided deep learning brain representation for visual classification," in *Proc. Int. Conf. Neural Inf. Process.*, 2020, pp. 54–61.