Confounds in the data—Comments on “Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features”

Hamad Ahmed, R. B. Wilbur, Hari M Bharadwaj, and Jeffrey Mark Siskind, Senior Member, IEEE

Abstract—Neuroimaging experiments in general, and EEG experiments in particular, must take care to avoid confounds. A recent TPAMI paper uses data that suffers from a serious previously reported confound. We demonstrate that their new model and analysis methods do not remedy this confound, and therefore that their claims of high accuracy and neuroscience relevance are invalid.

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

1 INTRODUCTION

A recent paper [8] presents a novel neural-network architecture, EEGChannelNet, for determining object class from EEG signals recorded from human subjects observing ImageNet [1] images as stimuli. Inter alia, it claims:

1. EEGChannelNet can decode object class from EEG signals better than prior work.
2. A training regimen that jointly fine tunes an image classifier while training EEGChannelNet, using a triplet loss that associates both positive and negative image samples with EEG samples, leads to an improved EEG classifier.

Here, we present novel evidence to refute these claims. We note that prior work [6] has already demonstrated other problems, namely:

a. The data used in [8] (from Spampinato et al. [9]) suffers from a confound (training and test samples coming from the same block with stimuli from a single class) and thus exhibits abnormally high classification accuracy with many different classifiers. When analyzed across subjects to eliminate this confound, accuracy degrades to chance.

b. New data collected with a block design also exhibits abnormally high classification accuracy with all of the same classifiers. Accuracy degrades to chance when this new data is bandpass filtered. Likewise, accuracy degrades to chance with new data collected to eliminate the confound: randomized trials and trials where the training and test data have different class presentation order.

Li et al. [6] also noted the well-documented slow spectral change in EEG. No amount of filtering can remove the confound.

Here, we document problems with the classifiers and training regimen:

1. Their new classifier EEGChannelNet exhibits the same flawed characteristics as the LSTM used in Spampinato et al. [9], addressed in [6]. This refutes claim 1.
2. Two additional classifiers evaluated by Palazzo et al. [8], EEGNet [5] and SyncNet [7], also exhibit the flawed characteristics.
3. The joint training regimen exhibits the same flawed characteristics. This refutes claim 2.

All remaining claims [8] are contingent on the confounded data, which results in refutation of the entire paper.

2 METHOD

We attempted to follow the experimental method in [8] and [6] as closely as possible. The appendix in the supplementary material, available online, presents the details. In all cases, we report the average of accuracy on the validation and test sets after the full training regimen.

3 RESULTS

We report below the new results from EEGNet, SyncNet, and EEGChannelNet (abbreviated below as EECGN) along with the results from Li et al. [6].

We first replicate the experiment of Spampinato et al. [9] on the block-design data collected by them with their original splits where the test sets come from the same blocks as the training sets.

The numbers differ somewhat from [9] and [8] as we use a different code base. Nonetheless, the numbers are qualitatively similar in that all classifiers exhibit high EEG classification accuracy. We next replicate the experiment of [9] on the block-design data collected by them with different splits in a leave-one-subject-out cross-validation paradigm. This allows the test sets to come from different blocks than the training sets.

Note that accuracy drops to chance for all classifiers. The remaining tables report analyses done with our own collected data [6]. First, we replicate the experiment of [9] on data collected with a block design on six new subjects.

All authors are with Purdue University, West Lafayette, IN 47907 USA. H. Ahmed and J. M. Siskind are with the Elmore Family School of Electrical and Computer Engineering. R. B. Wilbur is with the Department of Speech, Language, and Hearing Sciences and the Department of Linguistics. H. M. Bharadwaj is with the Weldon School of Biomedical Engineering and the Department of Speech, Language, and Hearing Sciences.

Manuscript received

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3121268, IEEE Transactions on Pattern Analysis and Machine Intelligence

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
EEGNet, and SyncNet on that experiment. The above demonstrates that all classifiers can obtain high classification accuracy on data collected with a block design. We collected two runs of block data from subjects 2–5 and three runs of block data from subject 6. Next, we report the data from the second

and third

block runs. These concur with the third table above. As discussed in Li et al. [6], the analyses in [9] erroneously omitted the bandpass filtering described in that paper. We next repeat the analyses in the above three tables with bandpass filtering added, respectively.

Accuracy drops to chance for all classifiers. We next report analyses performed on data collected with randomized trials both with and without bandpass filtering. In other words, all stimuli in the first block are labeled with class 1, even though they reflect different object classes, all stimuli in the second block are labeled with class 2, even though they reflect different object classes, and so forth. Note that classification accuracy is high for all classifiers, without bandpass filtering, suggesting that they are classifying a spurious correlation between the EEG signal and the block, not the stimulus category. This can be unduly high even with bandpass filtering, as is often the case. The remaining tables report cross-block classification. For subjects 2–6, the first and second block runs presented the stimuli in the same order. For subject 6, the third block run presented the stimuli in a different order. First, we report the average results of training on the first block run and testing on the second, and vice versa, both with and without bandpass filtering.
bandpass filtering. These report analyses between different runs with the same stimulus presentation order. Note that classification accuracy with all classifiers is significantly lower than within-block analyses, but can be above chance. Finally, we report the corresponding results for the first and third block runs,

<table>
<thead>
<tr>
<th>Table</th>
<th>subject</th>
<th>LSTM</th>
<th>l-NN</th>
<th>SVM</th>
<th>MLP</th>
<th>ID CNN</th>
<th>EEGNet</th>
<th>SyncNet</th>
<th>EEGCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>6</td>
<td>2.3%</td>
<td>2.6%</td>
<td>2.4%</td>
<td>2.3%</td>
<td>2.2%</td>
<td>2.4%</td>
<td>2.6%</td>
<td>2.8%</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>2.5%</td>
<td>2.4%</td>
<td>2.4%</td>
<td>2.3%</td>
<td>2.2%</td>
<td>2.0%</td>
<td>2.2%</td>
<td>2.6%</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>2.3%</td>
<td>2.3%</td>
<td>2.2%</td>
<td>2.2%</td>
<td>2.1%</td>
<td>2.2%</td>
<td>2.1%</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

and for the second and third block runs.

<table>
<thead>
<tr>
<th>Table</th>
<th>subject</th>
<th>LSTM</th>
<th>l-NN</th>
<th>SVM</th>
<th>MLP</th>
<th>ID CNN</th>
<th>EEGNet</th>
<th>SyncNet</th>
<th>EEGCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>6</td>
<td>2.3%</td>
<td>2.6%</td>
<td>2.4%</td>
<td>2.3%</td>
<td>2.2%</td>
<td>2.4%</td>
<td>2.6%</td>
<td>2.8%</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>2.5%</td>
<td>2.4%</td>
<td>2.4%</td>
<td>2.3%</td>
<td>2.2%</td>
<td>2.0%</td>
<td>2.2%</td>
<td>2.6%</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>2.3%</td>
<td>2.3%</td>
<td>2.2%</td>
<td>2.2%</td>
<td>2.1%</td>
<td>2.2%</td>
<td>2.1%</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

These report analyses between different runs with different stimulus presentation order. Note that classification accuracy with all classifiers is at chance. These results demonstrate that there is a confound not only between training and test samples collected in close temporal proximity within the same block, but also a second confound between samples collected in different runs but with the same temporal offset from the beginning of the run. Collectively these results demonstrate that EEGNet, SyncNet, and EEGChannelNet exhibit exactly the same flawed pattern of behavior as the LSTM model from Spampinato et al. [9]. To summarize, the only experiment designs among those considered above that do not suffer from one or both confounds are the ones with randomized trials (the ninth and tenth tables) and cross-block with different stimulus presentation order (the fifteenth through eighteenth tables). EEGChannelNet accuracy is at chance on these. Since all of the analyses in [8] use the same flawed data as in [9], everything that follows from those analyses is suspect.

Palazzo et al. [8] compare EEGChannelNet with EEGNet [5] and SyncNet [7] and claim improved accuracy. The tables above demonstrate that any relative performance difference is artificial as EEGNet and SyncNet exhibit the same characteristics as EEGChannelNet on faulty data. We make no claim about EEGNet or SyncNet themselves or the experiments reported in Lawhern et al. [5] and Li et al. [7]. Our concerns arise solely for the use of EEGNet or SyncNet as described in [8] for analyzing the flawed data from [9]. It is interesting to note that the tenth table above indicates that EEGNet, along with the SVM and 1D CNN, achieve accuracy slightly above chance on randomized trials.

For joint training, the resulting image classifier always performs above chance, usually highly above chance, but the resulting EEG classifier exhibits the same broad characteristics as all other classifiers, namely high classification accuracy on designs that exhibit a confound (all tables above except the ninth, tenth, and fifteenth through the eighteenth) and chance on designs that do not (the ninth, tenth, and fifteenth through eighteenth tables).

4 CONCLUSION

We demonstrate here that the claims 1 and 2 in Palazzo et al. [8] cannot be maintained because they rely on the flawed dataset from Spampinato et al. [9]. Further, the classification experiments therein fail when repeated on properly collected data without this confound (the ninth, tenth, and fifteenth through eighteenth tables).

ACKNOWLEDGMENTS

This work was supported, in part, by the US National Science Foundation under Grants 1522954-IIS and 1734938-IIS, by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00341, and by Siemens Corporation, Corporate Technology. Any opinions, findings, views, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, official policies, or endorsements, either expressed or implied, of the sponsors. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

REFERENCES
