Determining the Optimal Model, Reaction Time Delay, and Preprocessing Technique for the Classification of Visual Stimuli from Mouse Visual Cortex Activity

Harjaisal Brar

Abstract

Improving classification accuracies of visual stimuli from neural data is important in developing future models, including those for transfer learning, which will allow further research in the areas of visual process and vision loss. In this paper, several models and solvers/kernels, time delays, and preprocessing techniques are tested. This study finds several parameters that can be used to maximize prediction accuracy, ultimately producing an average accuracy of 81% when evaluated on the test set.

Introduction

The application of machine learning to the analysis of neural data, particularly those from the visual cortex, has been growing in usage due to its ability to process large amounts of data to find classes that are, in most cases, too abstract to find manually. Researchers have applied such processes to predict the firing rates of neurons from a picture of the stimuli to predicting the class of an object in a mouse's visual field from neural data (Kindel et al., 2019; Iqbal et al., 2019).

Fine tuning is often used for computer vision tasks, such as the classification of medical imagery for diagnostic purposes. It has been found that using a previously trained model and applying transfer learning, rather than retraining from scratch for a relatively similar task, allows faster training and higher prediction accuracies (Zhou et al., 2017). However, it is essential that the data it is being applied to is related to that on which it was originally trained in order for the model to perform properly (Day & Khoshgoftaar, 2017).

In the vast majority of studies on using classifiers to decode neural data, a deep neural network, such as GoogleNet, is fine tuned on the stimuli and the neural data (Day & Khoshgoftaar, 2017; Zhou et al., 2017). Although this method is extremely effective, little research has gone into determining the optimal model type and preprocessing techniques in order to maximize prediction accuracies. This is essential for the development of a custom neural network in the future which would specifically be designed to classify neural data -- not image data -- allowing it to need less neural data for fine tuning and improving classification accuracy, since the data will be homogenous with the original training set. Such a model could be fine tuned for each use case, depending on what is necessary to fit the specifications.

The goal of the present study is to determine the best model and solver/kernel, time delay, and preprocessing technique to maximize the classification accuracy of visual stimuli from primary visual cortex data. Determining methods to increase the prediction accuracy of visual stimuli would improve the ability to classify these stimuli from neural data, enabling the application of such techniques to potentially decode dreams of animals and find solutions to blindness. These specific findings create a foundation for the development of a specific model

for transfer learning, which would enable future users to fine tune such a model for an exact use case involving decoding neural data, even with limited training data.

Methods

All data was obtained from the Allen Brain Map from the Allen Institute for Brain Science (n.d.). The data was recorded by researchers using Neuropixel probes, which use CMOS manufacturing to record brain activity over 960 recording sites per probe while only being 70 microns wide, making it possible to insert multiple probes into the brain. The data covered 58 experiments, each with up to 6 Neuropixel probes recording data.

Fifteen experiments were selected for this research by filtering mice which had a wt/wt genotype and were tested in the Brain Observatory 1.1 survey to ensure the results would be consistent and repeatable. Additionally, the Neuropixel probe from which data was analyzed was probeC, which was inserted into the primary visual cortex of the mice. This was because data from this probe was present in each sample and provided an overall assessment of visual cortex activity. Data was only used from stimulus blocks 2, 5, and 7 because these were where the drifting gratings were presented to the mice. isi_violations_maximum and amplitude_cutoff_maximum were set to infinity while presence_ratio_minimum was set to negative infinity to disable the default thresholds as recommended by the documentation. Spike times were downloaded from each session from the dataset (Allen SDK, n.d.). These dataframes contained the spike time, unit id, and the time since stimulus onset. Each unit id was put into a set and sorted in an increasing numerical order. Then, each spike time was added into the corresponding unit id's column in an array. Next, the start time and end time of the window in which spikes were counted was calculated by taking the stimulus start and end times and applying a delay term, which shifts the window to account for the reaction time of the mouse. The number of stimuli in this window was counted per unit and was added to a matrix. The orientation of the grating was also stored in a vector in degrees. There were 8 possible orientations, ranging from 0 to 315 degrees in 45 degree increments. This was repeated for each stimulus presentation in the session to form a class vector, containing each grating orientation, and a spike matrix, containing the number of spikes in each stimulus presentation window per stimulus presentation. Lastly, this is all repeated per session for each of the sessions, ultimately forming 15 class vectors and spike matrices.

The size of the delay was varied in order to determine if adding a delay to the time to account for reaction time was advantageous. Since mouse reaction time has been found to generally range between 20 ms and 40 ms for visual stimuli, these were chosen as the 2 delay values to compare to the delay of 0 ms (Jain et al., 2015).

Z-scoring and normalization are common preprocessing techniques used in machine learning in fields from Economics to Environmental Science to Medicine to improve model accuracies (Barboza et al., 2017; Cho & Lee, 2020; Ranjbarzadeh et al., 2021). By calculating the Z-score across each unit, it is possible to interpret the data by considering its distance from the mean, as opposed to its normal value. This could prove useful for analyzing spike counts,

since it will mean that overspiking units would not be considered as just large, singular points. Instead, the model would interpret the differences in the value from its mean, allowing it to see when changes happen rather than just its value at a point. By normalizing the data across each unit, all units would be considered more evenly, as units with high spike counts will now have the same range of values as units with low spike counts -- from 0 to 1. The data was Z-scored by using the method from the scipy.stats module across axis 0. This was then used as training data as opposed to the normal spike counts used in the control. A similar process was used to normalize the data by using the normalization method from sklearn.preprocessing.

Three general model types were assessed in this study-- a support vector machine (SVM), a Logistic Regression model, and a Multilayer Perceptron (MLP) neural network -- to test accuracies over a variety of options. Each model was imported from the sklearn library (scikit-learn, n.d.). First, the class vectors and spike matrices were imported and split using the train_test_split method from sklearn. Next, each model type was initialized with each solver or kernel present. The SVM included the linear, polynomial (poly), radial basis function (rbf), and sigmoid kernels. For the Logistic Regression model, the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (lbfgs), linear (liblinear), Newton conjugate gradient (newton-cg), Newton-Cholesky (newton-cholesky), stochastic average gradient (sag), and SAGA (saga) solvers were tested. Lastly, the Neural Networks were tested with the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (lbfgs), stochastic gradient descent (sgb), and Adam (adam) solvers. The models were trained on 90% of the data (randomly selected) to predict the orientation from the distribution and number of spikes during that presentation. Then, they were used to predict the orientation of the presentations kept separately as test data from the spike vectors. A percent accuracy was then calculated and stored for later analysis. This was repeated for each solver or kernel to allow the comparison of results to see which model and solver/kernel combinations had the highest accuracies.

In order to ensure consistent results and to increase the sample size, the data was reshuffled 10 times per model and solver/kernel, and accuracies were then averaged. This approach ensured that each model and kernel/solver combination's accuracy percentages were consistent and repeatable. Additionally, it decreased the standard error of the means, thereby improving the strength of the conclusions (Vaux et al., 2012). If this step were not to be performed, it would be more likely that the training/testing split could be skewed, favoring or opposing the conclusions. This reshuffling was achieved by looping over each of the model training and testing cycles 10 times and averaging all the resulting percentages to produce 1 value for each model and solver/kernel combination. This averaging was only performed after ensuring that no outliers existed between any of the 10 values obtained from the loops.

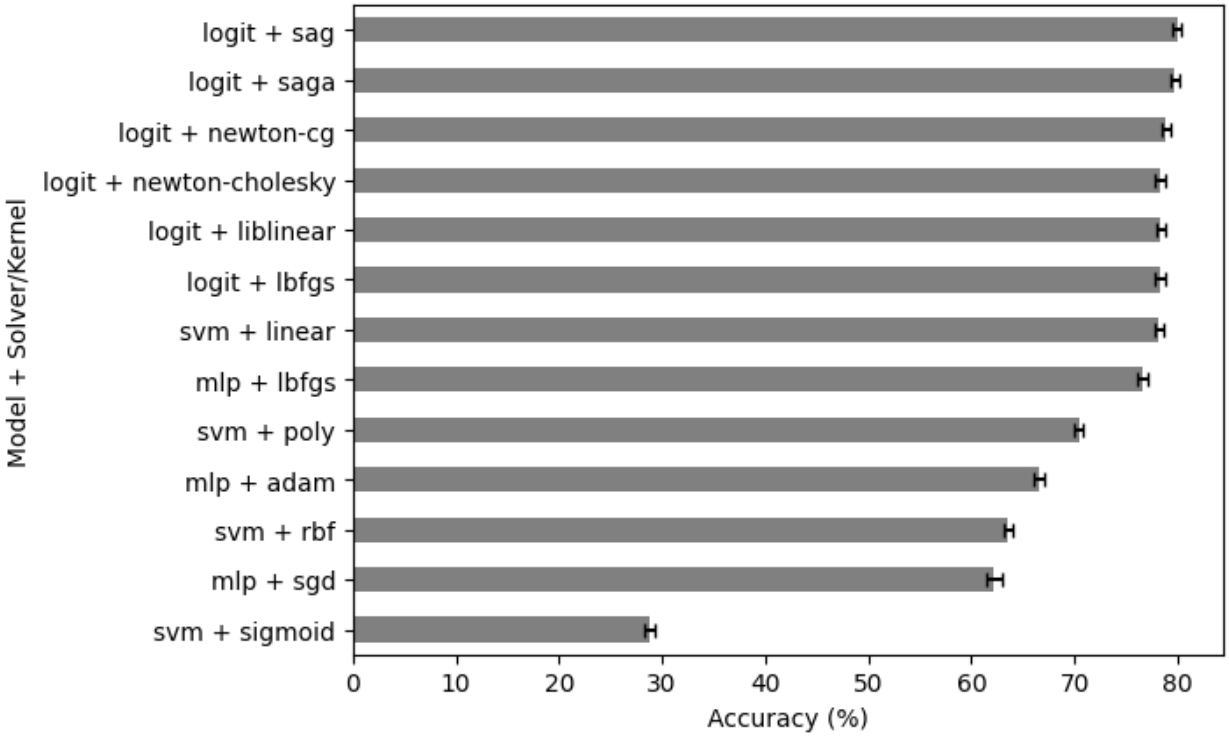
One of the samples proved to be outliers in this study, likely due to errors during data collection or processing in the lab. In order to objectively assess and remove this outlier, the Z-score of each percent result was calculated. Any results that were greater than 3 or less than -3 were removed (Andrade, 2021). Next, in order to determine the statistical significance of each result, a one-tailed t-test was performed with a p-value of 0.05 to determine if using the models

produced a significant increase in accuracy. The expected percent accuracy would be 12.5% since there are 8 classes, and the chance of randomly selecting the correct one is 1 in 8. To test the preprocessing and Z-scoring, one-tailed t-tests would once again be performed. However, this time, the expected value would be the average accuracy from the control group. Statistical tests were used on every set of results created in this study, including the accuracies for each combination of model and solver/kernel, and the effect of Z-scoring and normalization to ensure the data was consistent.

Results

Figure 1: Model and Solver/Kernel Results

The Accuracies of Various Model and Solver/Kernel Combinations with a 0 ms Delay and No Preprocessing Applied

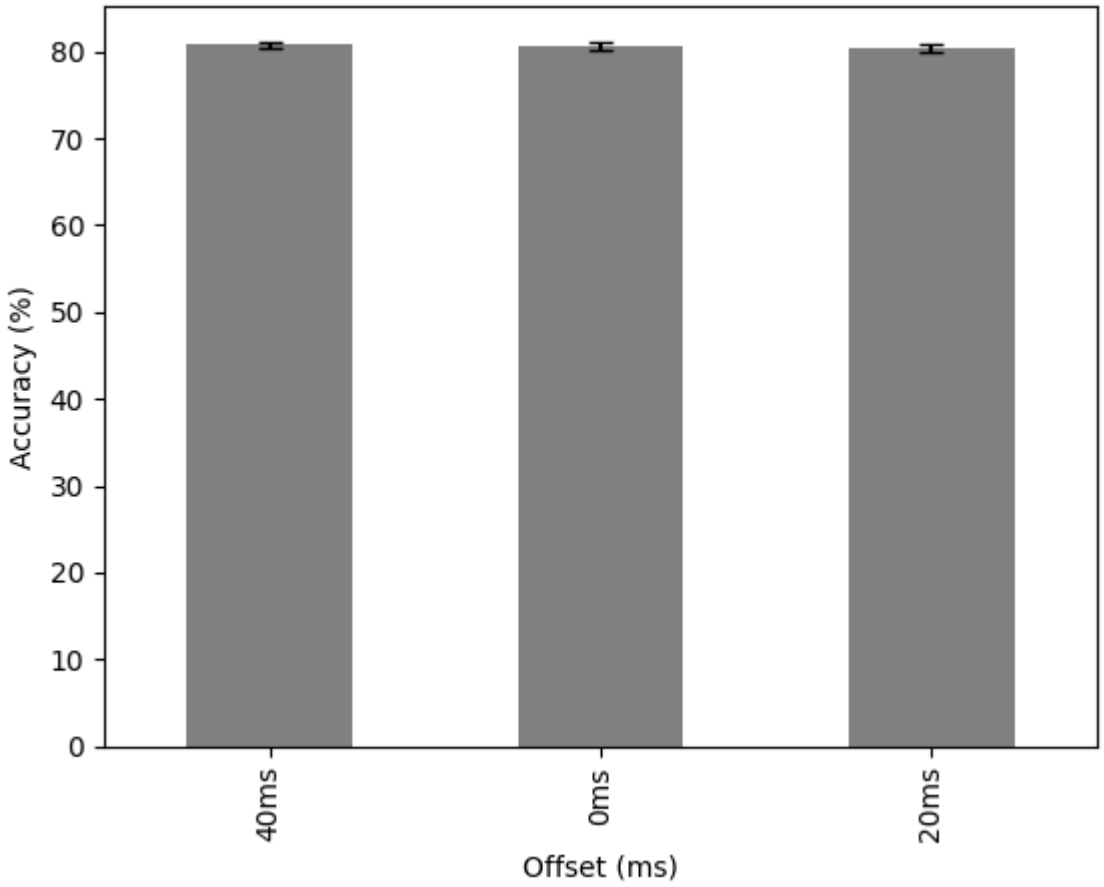


Note: Data was obtained by averaging the values for each sample over each loop, then averaging the values of the average sample accuracy. The error bars indicate the standard error of the mean.

Thirteen different model and solver or kernel combinations were tested over 10 loops for each of the 15 samples and the results were plotted on a bar graph (Figure 1). This showed the Logistic Regression Model with the stochastic average gradient (SAG) solver produced the most accurate predictions with an accuracy percentage of about 80%. However, similar results were also obtained with all other Logistic Regression solvers tested, as well as the linear SVM model.

Figure 2: Reaction Time Delay Results

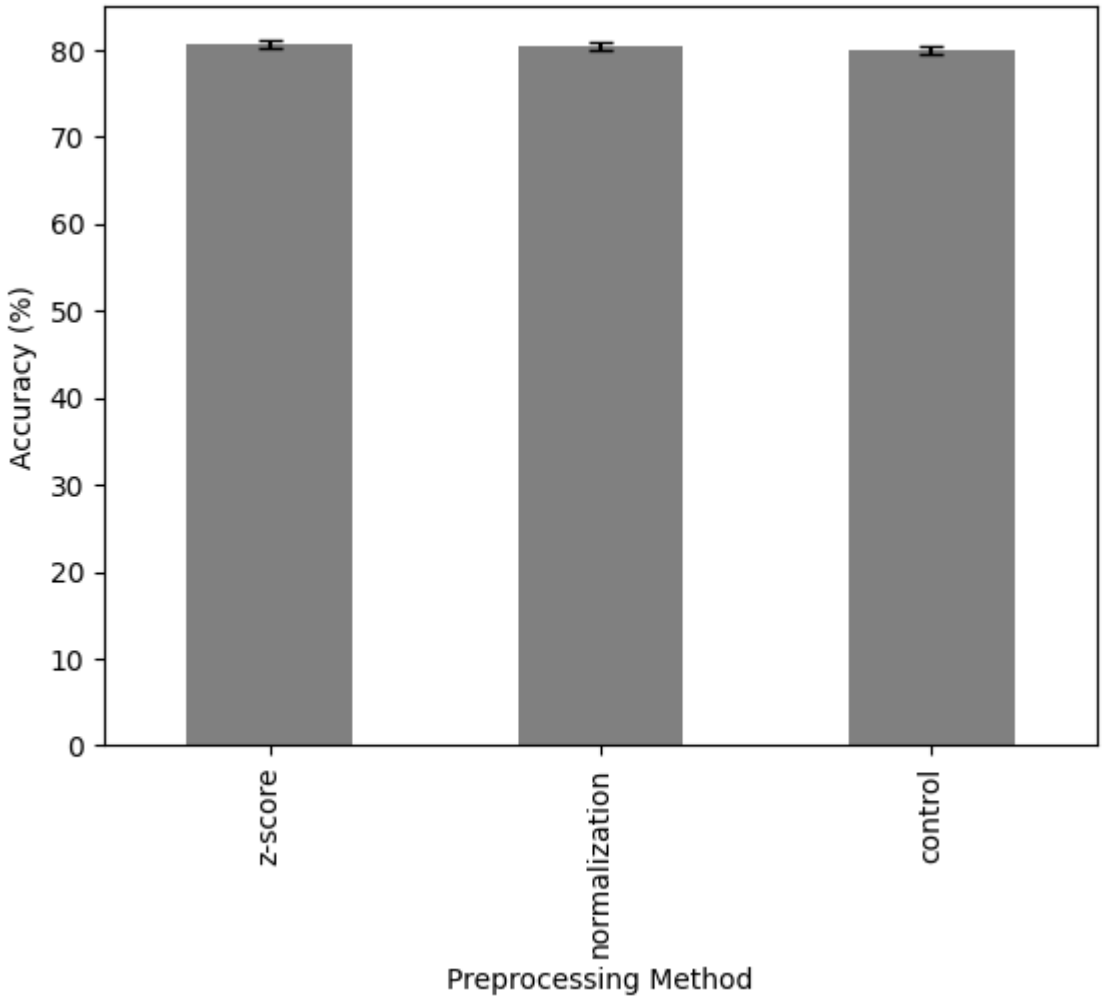
The Accuracies of the Logistic Regression Model with the SAG solver and the Input Data Z-Scored with Varying Reaction Time Delays



Note: Data was obtained by averaging the values for each sample over each loop, then averaging the values of the average sample accuracy. The error bars indicate the standard error of the mean.

Three different delay times were tested to see if accounting for the reaction time of the mouse would improve accuracy. These were tested on 15 samples over 10 loops. Of the three time delays that were tested, the 40 ms delay proved to classify the stimuli with the highest accuracy with an accuracy of about 81%. However, there were only small differences between the prediction accuracies for any of the delays.

Figure 3: Preprocessing Technique Results
The Accuracies of the Logistic Regression Model with the SAG solver and a 0 ms Time Delay with Varying Preprocessing Techniques



Note: Data was obtained by averaging the values for each sample over each loop, then averaging the values of the average sample accuracy. The error bars indicate the standard error of the mean.

Different preprocessing techniques were tested to see if they would improve model prediction accuracies as compared to the control, which had no preprocessing applied, namely normalization and Z-scoring. Over 15 samples and 10 loops, the Z-scoring proved to yield the highest prediction accuracies at around 81%. However, this was by a small amount.

Initially, when the values from each loop for individual samples were combined, outliers were tested for by calculating the Z-score of each element and determining if any elements had Z-scores over 3 or less than -3, as discussed in the methods. None of the Z-scores were out of this range, meaning no outliers were present, so all of the data was used in the next component.

Once the values from each loop from the individual samples were averaged, the data was once again tested for outliers using the same Z-score test. This time, sample 4 proved to be out of the range of the Z-scores for every combination of model and solver/kernel, time delay, and preprocessing applied, so it was an outlier. For this reason, this data was removed and was not considered for the rest of this study.

Now, the statistical tests were performed to test for the significance of the results. First, the significance of the models was tested using a one-tailed t-test with an expected value of 12.5%, since there are 8 classes present. The models produced a significant increase in accuracy (M = 70.8%, SD = 14.1) than would probabilistically be expected by random predictions, $t(12) = 14.86$, $p < .00001 < .05$. One-tailed t-tests were also applied to determine if the results were significantly improved by using the preprocessing techniques (M = 80.5%, SD = 0.109) or a time delay (M = 80.6%, SD = 0.278) as compared to each of their controls. The tests showed that there was a significant increase in accuracy by applying preprocessing techniques, $t(2) = 7.4615$, $p < .04241 < .05$, but there was not from using a time delay, $t(12) = -0.2121$, $p > .5 > .05$.

Discussion

The results showed that the Logistic Regression model with the SAG solver produced the highest accuracy predictions. In general, the linear models appeared to triumph over the non-linear models in this study, likely indicating that spike counts are linearly separable (Gherardi, 2021). The Logistic Regression models likely did better than the SVM linear model due to the presence of limited training data. If more data was present, it is possible that the results may be different, since there would be more time for the SVM and MLP models to properly converge. However, it can still be concluded that linear models, particularly those that use Logistic Regression, are optimal for classification of visual stimuli from spike counts when limited training data is present and all models were significantly more accurate than the expected value that would be obtained if predictions were randomly selected.

The 40 ms time delay produced the highest classification accuracy. However, this only occurred by a small amount. When tested for statistical significance, it was shown that there was no significant increase in prediction accuracy by implementing either time delay. This could occur for a few different reasons. First, the reaction time of each mouse to fast moving patterns, such as the drifting grating used in this study, could be lower than that for normal visual stimuli. Second, it is possible that, by adjusting the window so significantly, key spikes by neurons responsible for important factors such as brightness or movement, were skipped and were not considered as part of the data, making it significantly harder for the models to distinguish between the background spiking and the visual stimuli (Shmiel et al., 2005). Lastly, this could also have occurred due to individual variances between mouse reaction time due to numerous factors, such as age, experience with other visual stimuli, time of day, among others (Talboom et al., 2021). Ultimately, no conclusions can be drawn to indicate that a time delay improved prediction accuracy.

It was found that Z-scoring the input data before inputting it into the model for classification increased the prediction accuracies by a significant amount. The same was shown when the data was normalized, as well. Since these neurons were transformed across the rows, meaning they were essentially transformed over time, by applying these preprocessing techniques, the importance of each neuron was equated, so the model would treat each value equally instead of preferring a neuron which spikes more than the other. Furthermore, by using Z-scoring, the difference from its mean value was considered, meaning the value became positive when it spiked more than average and negative when it spiked less (Tanaka et al., 2022). This likely made it easier for the model to determine when a neuron is overspiking or underspiking, which can be used as a good benchmark to predict which class is being displayed.

Throughout this study, only one outlier was found, which was sample 4, which had very low prediction accuracy percentages. When a Z-score was calculated for all models, solvers/kernels, time delays, and preprocessing relative to the other samples, it was found to be less than -3 for every one, making it an outlier (Andrade, 2021). With further investigation, it was found that all values in the local field potential (LFP) data were constant over the duration of this study. This is likely indicating a data recording or processing error by the researchers who created the dataset. For this reason, this sample was dropped before the samples were averaged and final conclusions were drawn.

Some limitations of this analysis include the small amount of training data available and the lack of availability of labeled neural data from mice exposed to a broad range of visual stimuli. In this dataset, only about 660 orientations and spike vectors were available, since this was the maximum number of times these stimuli were shown to the mice. With more data, it is significantly more likely that the models will have improved classification accuracies, because they likely will have all converged and not overfit (Lopez et al., 2022). Another limitation was the lack of labeled neural data from a broad range of stimuli. This research was constrained to classifying the orientation of a moving grating. However, with a large amount of data encompassing many different forms of visual stimuli, the model would learn how to better process visual neural data. This would mean the model would have higher prediction accuracies, but could likely also be applied in a transfer learning setting to predict other visual stimuli with significantly less training data.

In the future, I would like to create a large, generalized neural network which can be fine tuned for a variety of classification tasks from visual neural data. This research could be applied in many different areas, from decoding the dreams of animals to finding new solutions to the loss of vision. Additionally, I would like to do similar analyses on thalamus, hippocampus, and midbrain activity to see if more information can be extracted, improving classification accuracies. Lastly, I would like to test to see if performing frequency analyses on the input data, such as calculating the Power Spectral Density to determine if they will function as preprocessing methods to improve classification accuracy (Boashash, 2016).

Conclusions

Through this project, Linear Models, particularly Logistic Regression with a Linear Solver, were discovered to be optimal for the classification of visual stimuli. Additionally, Z-scoring input data before training and predictions improves classification accuracy. Lastly, implementing a time delay to account for the animal's reaction time does not significantly impact the model's ability to accurately predict stimuli.

These findings could be applied to creating future models, specifically for the decoding of visual cortex activity, that could be used for transfer learning. These networks would be able to be fine tuned, making them perfect for a specific task and only requiring small amounts of data to train it for the specific role. Ultimately, these models would allow further advancement in our ability to understand and decipher the brain, shedding insights on visual processing and solutions to problems such as vision loss.

Bibliography

1. Allen SDK. (n.d.). Allen SDK Dev Documentation. Retrieved April 19, 2023, from https://allensdk.readthedocs.io/en/latest/
2. Andrade, C. (2021). Z Scores, Standard Scores, and Composite Test Scores Explained. *Indian Journal of Psychological Medicine*, *6*, 555–557. https://doi.org/10.1177/02537176211046525
3. Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006
4. Boashash, B. (2016). *Time-Frequency Signal Analysis and Processing* (2nd ed., pp. 521–573). Academic Press.
5. Cho, J. H., & Lee, H. (2020). Optimization of Machine Learning in Various Situations Using ICT-Based TVOC Sensors. *Micromachines*.
6. Day, O., & Khoshgoftaar, T. (2017). A survey on heterogeneous transfer learning. *Journal of Big Data*.
7. Gherardi, M. (2021). Solvable Model for the Linear Separability of Structured Data. *Entropy*, *3*, 305. https://doi.org/10.3390/e23030305
8. Iqbal, A., Dong, P., Kim, C. M., & Jang, H. (2019). Decoding Neural Responses in Mouse Visual Cortex through a Deep Neural Network. *ArXiv*. https://doi.org/10.48550/ARXIV.1911.05479
9. Jain, A., Bansal, R., Kumar, A., & Singh, K. (2015). A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. *International Journal of Applied and Basic Medical Research*, *2*, 124. https://doi.org/10.4103/2229-516x.157168
10. Kindel, W. F., Christensen, E. D., & Zylberberg, J. (2019). Using deep learning to probe the neural code for images in primary visual cortex. *Journal of Vision*, *4*, 29. https://doi.org/10.1167/19.4.29

11. López, O. A. M., López, A. M., & Crossa, J. (2022). *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (pp. 109–139). Springer Nature.
12. Ranjbarzadeh, R., Jafarzadeh Ghoushchi, S., Bendechache, M., Amirabadi, A., Ab Rahman, M. N., Baseri Saadi, S., Aghamohammadi, A., & Kooshki Forooshani, M. (2021). Lung Infection Segmentation for COVID-19 Pneumonia Based on a Cascade Convolutional Network from CT Images. *BioMed Research International*, 1–16. https://doi.org/10.1155/2021/5544742
13. *scikit-learn*. (n.d.). Scikit-Learn 0.16.1 Documentation. Retrieved April 19, 2023, from https://scikit-learn.org/stable/index.html
14. Shmiel, T., Drori, R., Shmiel, O., Ben-Shaul, Y., Nadasdy, Z., Shemesh, M., Teicher, M., & Abeles, M. (2005). Neurons of the cerebral cortex exhibit precise interspike timing in correspondence to behavior. *Proceedings of the National Academy of Sciences*, *51*, 18655–18657. https://doi.org/10.1073/pnas.0509346102
15. Talboom, J. S., De Both, M. D., Naymik, M. A., Schmidt, A. M., Lewis, C. R., Jepsen, W. M., Håberg, A. K., Rundek, T., Levin, B. E., Hoscheidt, S., Bolla, Y., Brinton, R. D., Schork, N. J., Hay, M., Barnes, C. A., Glisky, E., Ryan, L., & Huentelman, M. J. (2021). Two separate, large cohorts reveal potential modifiers of age-associated variation in visual reaction time performance. *Npj Aging*, *1*. https://doi.org/10.1038/s41514-021-00067-6
16. Tanaka, T., Nambu, I., Maruyama, Y., & Wada, Y. (2022). Sliding-Window Normalization to Improve the Performance of Machine-Learning Models for Real-Time Motion Prediction Using Electromyography. *Sensors*, *13*, 5005. https://doi.org/10.3390/s22135005
17. Vaux, D. L., Fidler, F., & Cumming, G. (2012). Replicates and repeats—what is the difference and is it significant? *EMBO Reports*, *4*, 291–296. https://doi.org/10.1038/embor.2012.36
18. *Visual Coding - Neuropixels*. (n.d.). Brain Map. Retrieved April 19, 2023, from https://portal.brain-map.org/explore/circuits/visual-coding-neuropixels
19. Zhou, Z., Shin, J., Zhang, L., Gurudu, S., Gotway, M., & Liang, J. (2017). Fine-tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit.*