Auditory Encoding Model with Convolutional Neural Network

Jonathan Gonzalez Martinez[1], Quynh-Anh Nguyen[1,2]

[1] Department of Mathematical Sciences, University of Indianapolis; [2] Center for Data Science, University of Indianapolis

Introduction

The auditory system enables us to understand the surrounding information, and human experiences auditory perception every day in the form of speech and music. The sound waves entering our ears are composed into frequency components, and the signals are sent to different brain areas. In the primary auditory cortex (A1), the inputs are integrated with information from other sensory systems. In fact, A1 plays an important role in information processing and the formation of auditory objects.

The encoding problem in A1 studies how A1 neurons response to external stimuli. Many studies of the auditory cortex use spectrotemporal receptive field (STRF) to investigate the relationship between the stimulus and the neural responses over time. One disadvantage of this approach is that it assumes linearity of auditory responses while different experimental and computational studies have indicated that neuronal response in the auditory system is not merely linear (Atencio et al., 2009; Sahani and Linden, 2002).

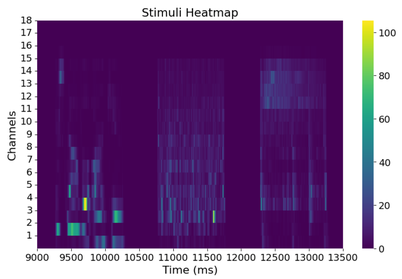
In our study, we deploy a **convolutional neural network framework** to investigate auditory cortical responses. Even though neural network is newer compared to many traditional models (that tend to involve systems of partial or ordinary differential equations), it has gained tractions recently due to its speed, availability of open datasets, and efficiency of training algorithms (Lindsay, 2021; Pennington and David, 2023). Also, we plan to study the structure of the convolution neural network. Investigating the network architecture and finding some common features will enable us to propose a potential mechanism of information processing in the auditory cortex.

Data

We obtained recording data from Zenodo (Pennington and David, 2023). The recording was performed in A1 and a secondary auditory field (PEG) on 5 ferrets. We only focus on the A1 recording for this study. The A1 recording has 22 sites with 849 units.

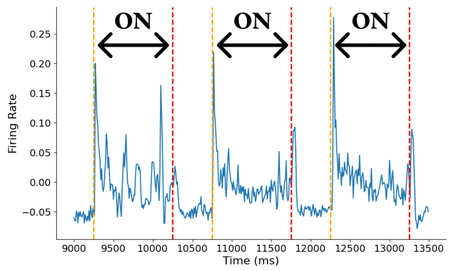
The stimuli is a collection of 595 natural sound samples. Each sample is 1-second long, followed by a 0.5-second of inter stimuli interval (ISI). Approximately 15% of the sounds were ferret vocalizations and environmental noises while the remaining 85% of the sounds were human speech, music, and environmental noises.

Here shows the heatmap of three samples from 900 ms to 1350 ms. The stimulus waveforms were transformed to log-spaced spectrograms with 18 channels ranging logarithmically from 200 Hz to 20,000 Hz. To be consistent with the obtained data from Zenodo (Pennington and David, 2023), we used the first 27 seconds as validation data.



The obtained data has firing rates for 849 A1 units. For the current model, we focus on the population firing rate, i.e. the mean firing rate across 849 units. We then normalize the population firing rate using its mean (0.08) and standard deviation (0.29).

The normalized population firing rate shown for 900 ms to 1350 ms. The yellow dashed lines indicate the onset of a sound. The red dashed lines indicate the end of a sound. It can be observed that the firing rate increases abruptly as soon as the sound is on. It then quickly diminishes even if the sound is still on.



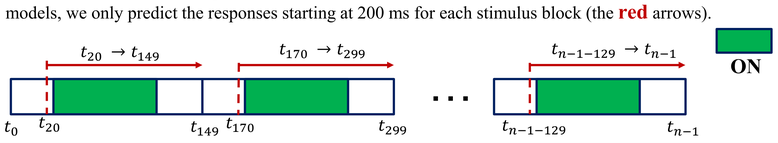
Acknowledgement: This study was supported by the Faculty Scholarship Grant, University of Indianapolis (Grant Number: 995049)

Convolutional Neural Network

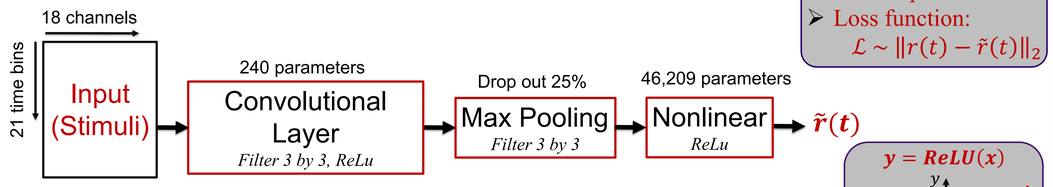
To study how A1 neurons response to external stimuli, we build convolutional neural network models that use stimuli as inputs to estimate the neural responses:

- The input is composed of the stimuli at that timepoint and the stimuli of the previous 20 timepoints (previous 200 ms, 1 time bin = 10 ms), resulting in a 21 (time dimension) by 18 (frequency dimension) matrix.
- The output is the normalized population firing rate $\tilde{\boldsymbol{r}}(\boldsymbol{t})$.
- The models are developed in the TensorFlow library of Python.

The Zenodo data was averaged across trials and aligned with respect to each stimulus. However, the actual order of each stimulus was randomized during the experiment (Pennington and David, 2023). Thus, we do not have a sequential time series from one stimulus block (1.5 sec) to the next. In our models, we only predict the responses starting at 200 ms for each stimulus block (the **red** arrows).

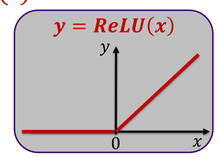


ONE-CONVOLUTIONAL LAYER MODEL (15 epochs)

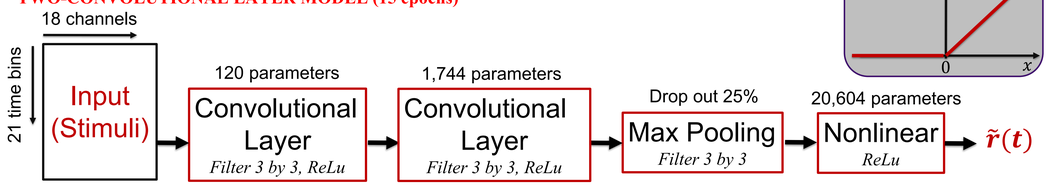


- Adam optimizer
- Loss function: $\mathcal{L} \sim \|r(t) - \tilde{r}(t)\|_2$

$y = ReLU(x)$

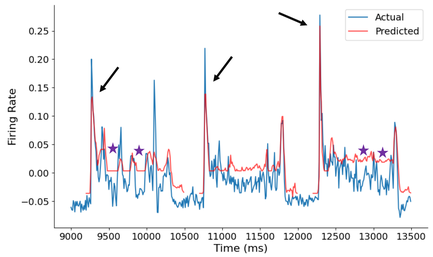


TWO-CONVOLUTIONAL LAYER MODEL (15 epochs)

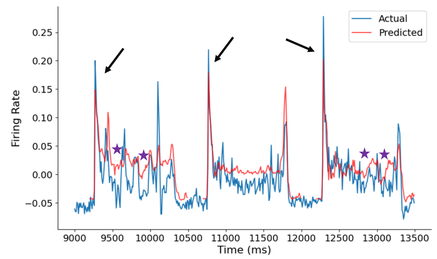


Results

ONE-CONVOLUTIONAL LAYER MODEL



TWO-CONVOLUTIONAL LAYER MODEL

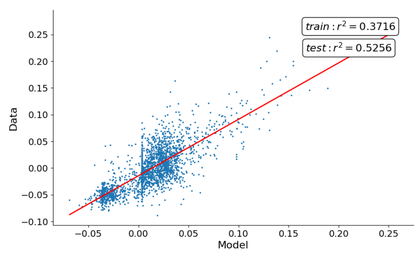


The outputs of both models (one-convolutional layer and two-convolutional layer) are shown above. The time plot of the two-convolutional layer model shows a slightly better fit than that of the one-convolutional layer model.

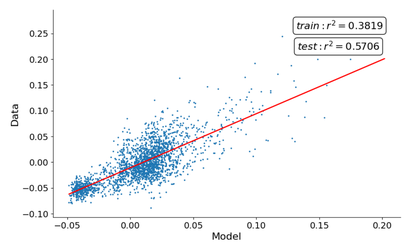
- In most cases, both models are able to capture the up and down trends of the normalized population firing rate. For example, both models predict the abrupt increase of the response at the stimulus onset (the **black** arrows).
- The two-convolutional layer model tends to predict the down trend more accurately than the one-convolutional layer model (the **purple** stars).
- While both models can capture the trend, the magnitude of the prediction tends to be less extreme than the magnitude of the actual response (in both up and down directions).
- The model predictions of the responses to some sound samples seem to be better than the predicted values for other sound samples. The response to the third sound sample gets a better prediction from both models as compared to the first two.

Note that there is no predicted responses at the first 200 ms of each sample block because our models do not apply for this period.

ONE-CONVOLUTIONAL LAYER MODEL



TWO-CONVOLUTIONAL LAYER MODEL

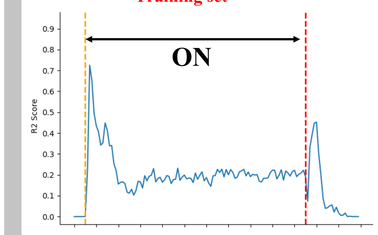


The outputs of both models (one-convolutional layer and two-convolutional layer) are shown above. The scatter plot between the predicted response and the actual response of the two-convolutional layer model shows a slightly better fit than that of the one-convolutional layer model. The R^2 values for both models are listed in the figures. The red lines $y = x$ show how close the fitted values are to the actual values.

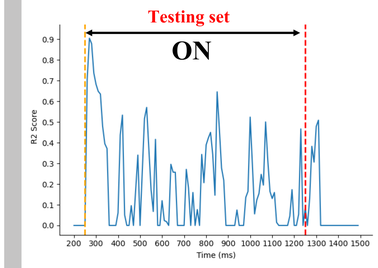
- In most cases, the predicted values follow the same increasing and decreasing trends as the empirical values because the scatter plots tend to hoover around the red line $y = x$.
- Moreover, the R^2 values for the testing data are always bigger than the R^2 values of the training data. We think that it is due to the fact that the testing data is much simpler and smaller than the training data. In fact, there are 577 sound samples in the training set while there are only 18 samples in the testing set. We use these as our training and testing sets to be consistent with the data we obtained from Zenodo (Pennington and David, 2023).

Results (Cont'd)

Training set



Testing set



For each stimulus block, we compare the predicted responses and the actual responses at each time point. We then compute the R^2 values across times. Here are the R^2 plots for both training and testing sets for the two-convolutional layer model.

- The R^2 value is the highest immediately after the onset of a sound sample. The model is able to capture the abrupt rise in the firing rate due to the start of a stimulus.
- The R^2 value is lower after that. In the training set, the R^2 value quickly decreases to about 0.2 within 150-200 ms after the stimulus onset. In the testing set, the R^2 value fluctuates between 0.6 and some really small values.
- The R^2 value improves quickly again immediately after the stimuli is off at 1250 ms. While the R^2 value here is smaller than the R^2 value at the onset of the stimulus, it is on the average higher then the R^2 value between 400 ms and 1250 ms.

In general, the model is able to predict the sudden changes in the firing rates due to the abrupt change in environment. However, while the stimulus is on, the model does not perform as well.

The result for one-convolutional layer model follows similar trend.

Conclusion and future works

In general, the convolutional neural networks perform better than the general linear models. Here the two-convolutional layer model is slightly better than the one-convolution layer model.

- In both models, the predicted responses tend to follow the same trend of the actual responses. However, the models cannot capture some of the extreme high and low values completely.
- Both models tend to fit the upshoot in the firing rate immediately after the onset of each sound sample really well compared to the rest of time course. This may be due to the fact that initially the neurons get excited mostly due to the onset of the sounds, but as time goes on there are other neuronal processes such as habituation and adaptation. Thus, our encoding modes (taking solely stimuli as inputs) do not capture the latter and more complicated neuronal processes.

Moving forward, we plan to improve our encoding models by:

- looking at different data sets with other types of stimuli such as pure tones, dynamics ripples, and natural sounds
- analyzing the fit across sound samples to investigate the relationship between response patterns and stimuli, i.e. representational similarity analysis.
- studying the patterns and structure of the parameter matrices.
- investigating the statistics of the stimuli to determine how these features can be used to inform the structures of the models.

References

Atencio, C.A., Sharpee, T.O., & Schreiner, C.E. (2009). Hierarchical computation in the canonical auditory cortical circuit, *Proceedings of the National Academy of Sciences, 106*(51), 21894-21899.

Lindsay G.W. (2021). Convolution neural networks as a model of the visual system: past, present, and future. *Journal of Cognitive Neuroscience, 33*(10), 2017-2031.

Pennington, J.R. & David, S.V. (2023). A convolution neural network provides a generalizable model of natural sound coding by neural populations in auditory cortex. *PLoS Computational Biology, 19*(5), e1011110.

Pennington, J.R. & David, S.V. (2023). Auditory cortex single unit population activity during natural sound presentation – dataset (1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7996574

Sahani, M. & Linden, J.F. (2002). How linear are auditory cortical responses? *Proceedings of the 15th International Conference on Neural Information Processing Systems*, 125-132.