Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Predicting Horizontal Direction of Eye Movement using Electroencephalography Data from EEGEyeNet's Visual Symbol Search Dataset

Semester Thesis

Srijan Saxena

saxenas@student.ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:
Ard Kastrati, Martyna Beata Płomecka, Anh Duong Vo
Prof. Dr. Roger Wattenhofer

January 31, 2023

Abstract

This research aims to perform the Left-Right Direction Prediction L/R task on the Visual Symbol Search data paradigm, as an extension of the work by [1]. Previously, the LR task was trained on the anti-saccade paradigm in the EEGEyeNet dataset, which was a structured experiment to record the direction of the gaze of participants after explicit cues. The VSS paradigm is a cognitive task with no explicit instructions for the participant to move their gaze in a certain direction. Thus, I explore whether the previously employed data preparation method and CNN model is robust enough to perform the L/R task on a dataset such as the VSS paradigm, while identifying and implementing necessary modifications. Ultimately, I was able to achieve an average accuracy of 0.9968, highlighting the robustness of the existing model as well as the efficacy of the proposed data preparation technique.

Contents

Abstract **i**

1 Introduction **1**

2 EEGEyeNet **2**

3 Data Processing **4**

3.1 Data Exploration 4

3.2 Data Preparation 6

4 Model **10**

5 Results **11**

5.1 Initial Results on Minimally Preprocessed Subset 11

5.1.1 Model Confidence 12

5.1.2 Data from Electrodes 125 and 128 13

5.1.3 Distribution of Incorrect Predictions by Subject 14

5.1.4 Errors and Eye Tracking Data 15

5.1.5 Key Insights 16

5.2 Results on the Minimally Preprocessed Dataset 16

5.2.1 Distribution of Incorrect Predictions by Saccade Amplitude 17

5.3 Results on the Maximally Preprocessed Dataset 17

6 Discussion **21**

7 Conclusion **22**

Bibliography **23**

Chapter 1

Introduction

Eye Tracking is a growing field of study due to its immense potential in areas such as UI/UX[2], behavioural research, assistance technology[3] etc. An active area of research in this field is the use of electroencephalographic (EEG) data to complement existing eye tracking solutions or to replace them with a low-cost alternative [4]. Any EEG based solution aiming to complement or replace existing Eye Tracking solutions needs to be very robust. In 2021, [1] released a powerful dataset of high quality EEG recordings over 3 different experiment paradigms called EEGEyeNet. Furthermore, they developed their own benchmark models for gaze estimation tasks from EEG signals. One of these gaze estimation tasks, the left-right direction task, focused on predicting which horizontal direction a participant's gaze went. In [1], they developed this direction estimation benchmark model by training it on the anti-saccade experiment paradigm, a structured experiment where a participant's gaze is controlled using cues. In this thesis, I wish to build upon their previous work by exploring the robustness of their model and data preparation techniques. I plan to try and apply their benchmark models to a dataset obtained from the Visual Symbol Search (VSS) experiment paradigm, where the participants were focused on solving cognitive and physical tasks with no prompts or cues to control their gaze. My contribution in this thesis is not to just explore the robustness of the existing benchmark models, but also to fully explore the data, build appropriate data preparation methods and optimize the accuracy for this task.

Chapter 2
EEGEyeNet

The EEGEyeNet database was constructed by recording hours of EEG using a 500Hz 128-channel EEG Geodesic Hydrocel system, from 356 healthy adults for over 3 different experiment paradigms [1]. Here I will introduce two of those experiment paradigms:

- Pro- and Antisaccade: Participants start by focusing on the central fixation square at centre of their screens. Every 1-3.5 seconds, a cue will appear for exactly 1 second on either the left or right of the central fixation square. In pro-saccade trials, participants are asked to immediately fixate on the cue, while during anti-saccade trials, they are asked to perform a saccade in the opposite direction. Once the cue disappears, the participants return to fixate on the central fixation square. As can be seen, this is a very rigid experiment that seeks to prompt and control the participant's gaze as much as possible. The participant is also likely only thinking about their gaze throughout the experiment.
- Visual Symbol Search: Participants are given 15 rows of symbol search questions at a time, wherein each row consists of 2 target symbols on the left and 5 search symbols on the right, and the participants press the "YES" button if either of the target symbols is present in the set of search symbols, otherwise they press "NO". VSS is a digital version of a clinical method to measure processing speed. Unlike Anti-saccade, the experiment doesn't restrict the participant as they are free to look wherever they want whenever they want. The participant is also likely not thinking about their gaze at all, instead focusing on solving the symbol search problem.

EEGEyeNet provides two different types of preprocessing on the raw EEG, using the openly available toolbox from [5]:

- Minimal preprocessing: Bad electrodes are identified and interpolated and the data is filtered with a 40 Hz high-pass filter and 0.5 Hz low-pass filter.

- Maximal preprocessing: Tries to remove contamination from a larger number of external artifacts such as the heart, muscles, eyes etc. using Independent component analysis along with a pre-trained classifier that "estimates the probability of a component reflecting artifactual activity"[1].

Minimally prepossessed data consistently performed better with benchmark models for gaze estimation tasks since it still included some ocular artifacts.

Chapter 3

Data Processing

3.1 Data Exploration

For the left-right direction task, I decided to divide the dataset into individual sequences, where each sequence is the EEG and Eye Tracker's recording during the occurrence of a consecutive fixation, saccade and fixation. The saccade is the important event for this task since that is when the eye actually moves. The preceding and proceeding fixations were included to account for any neurological processing that takes place before and after the saccade actually occurs.

My first aim was to analyze how the data is distributed and if the VSS dataset is even appropriate for the left-right direction task. The first distribution I analysed was the saccade angle, to ascertain whether the movement was primarily horizontal in nature. 3.1 shows the distribution of saccade angles plotted on an angular histogram. As can be observed, a majority of saccades are in the horizontal direction, having saccade angles close to either O degrees or 180 degrees. Furthermore, we can also observe that the left and right saccades are almost equally distributed, addressing initial concerns about a class imbalance.

Next, I tried to analyse the distribution of saccade amplitudes over the sequences. As can be observed in 3.2, the saccades in the Antisaccade Paradigm can be classified into 2 types based on their amplitudes: unprompted and uncontrolled microsaccades and the experiment-induced saccades. The left-right direction task on the anti-saccade paradigm was trained only on the latter experiment-induced saccades. Since the distance of the cues from the central fixation square was fixed, the saccade amplitudes had little variance and could also be considered (approximately) fixed.

However, since the saccades in the VSS paradigm were uncontrolled and unprompted, the distribution of saccade amplitudes could have a high variance. Thus, I wanted to explore how these amplitudes were currently distributed and if that variance might lead to lower model performance. 3.3 shows the distribution of saccade amplitudes (left) and the distribution of the horizontal component of saccade amplitudes (right) plotted as overlapping histograms for the left (blue)

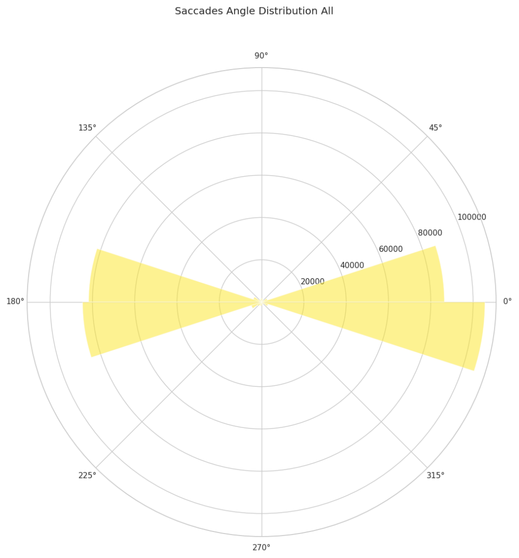
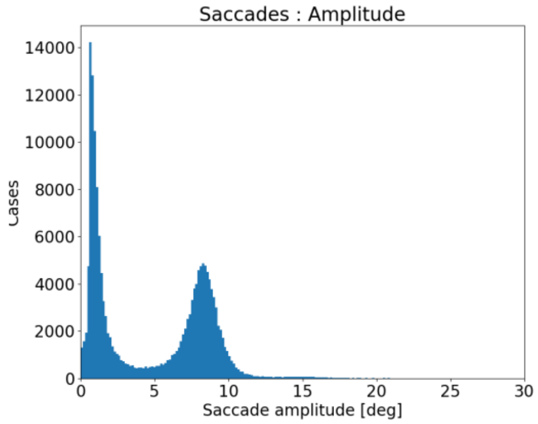


Figure 3.1: Saccade Angle Distribution

and right (orange) class. As expected, given the distribution of saccade angles shown earlier, the distributions of the saccade amplitudes and the horizontal component of the saccade amplitudes are very similar. We can also observe that the distribution of amplitudes is similar for both classes, with most saccades have amplitudes below 200 units. For saccades with amplitudes beyond 200 units, the distribution of amplitudes for both classes start to differ, with the right class having a noticeably higher frequency of amplitudes between 200-400 units, and the left class having a mildly higher frequency of amplitudes between 400-650 units.

There were some outlier sequences which had a calculated amplitude as high as 60,000, which is physically impossible. Thus, all sequences with amplitudes greater than 800 units were marked as outliers and removed.

Finally, I tried to analyse the distribution of the duration of the fixations and saccades that constitute each sequence. It is important to understand this distribution so that we can determine how to prepare the data and ascertain an appropriate fixed input sequence length for the model. The duration distributions for the sequence and its components were represented using violin plots, with separate plots for each class. 3.4 shows the initial plots, where the presence of extraordinary outliers can be observed. Ideally, singular saccades lasting for 30,000 time points (60 secs) shouldn't even be possible and it is possible that they arose as a result of equipment failure. Therefore, I removed all sequences where the total duration of the sequence was greater than 1200. 3.5 shows the distribution of this updated dataset, with 3.1 showing the exact percentile values.



(c) Distribution of the saccade amplitude

Figure 3.2: Saccade Amplitude Distribution (Antisaccade Paradigm) [1]

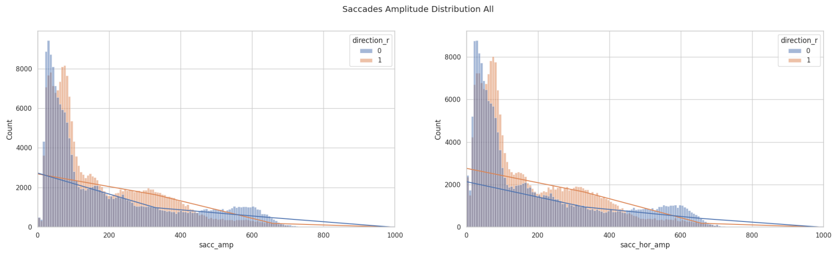


Figure 3.3: Saccade Amplitude Distribution (VSS Paradigm)

We can observe that the duration distributions for sequences belonging to both classes is very similar. This further reduces the probability of our model developing a class bias.

3.2 Data Preparation

In order to prepare the data for the model, I decided to divide the data into fixed-length samples constructed using the sequences defined in the previous section (occurrences of a consecutive fixation, saccade and fixation). Thus, each sample included a 600ms recording from each of the 128 EEG channels as well as the reference electrode, giving each sample a shape of 129 x 300.

The samples used for the left-right direction task on the anti-saccade paradigm

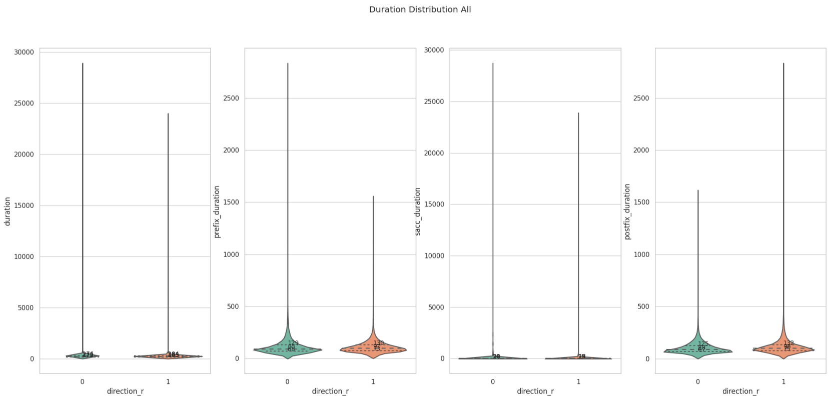


Figure 3.4: Saccade Duration Distribution for all sequences

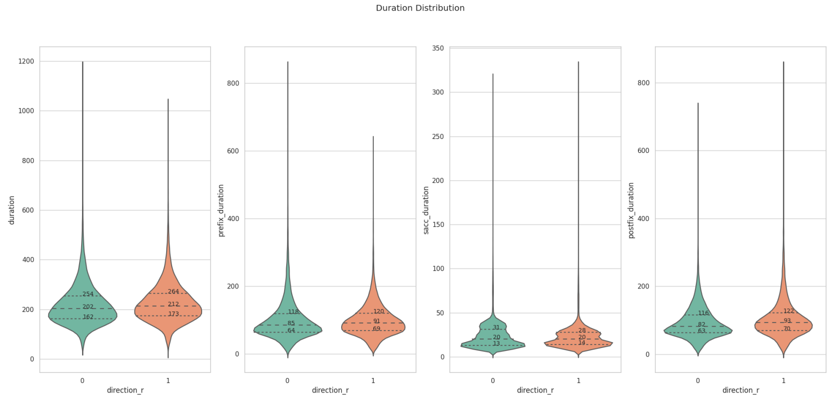


Figure 3.5: Saccade Duration Distribution without outliers

by [1] had a fixed sequence length of 500. This was appropriate for the anti-saccade paradigm since the gaze of the participant was controlled using a cue that lasted for exactly 1 second (500 data points). Thus, the 500 data points recorded during the cue's existence completely captured a sequence of a consecutive fixation, saccade, and fixation.

However, in the VSS dataset, all selected saccades are uncontrolled and unprompted and thus, as shown in 3.5, the length of the sequence varies greatly. This means that either a sequence will be too large for the chosen fixed sequence length (thus, requiring trimming) or be too small for the chosen fixed sequence length (thus, requiring padding). The aim was to choose a sequence length that is large enough to completely encapsulate most sequences, while not being so large that most of the sample is padding. Given that the 75th percentile for

Distribution	**25th**	**50th**	**75th**
Sequence Duration (L)	162	202	254
Sequence Duration (R)	173	212	264
Pre-fixation Duration (L)	64	85	118
Pre-fixation Duration (R)	69	91	120
Saccade Duration (L)	13	20	31
Saccade Duration (L)	14	20	28
Post-fixation Duration (L)	63	82	116
Post-fixation Duration (R)	70	93	122

Table 3.1: Distributions of Durations

the duration of a sequence is 254 for the left class and 264 for the right class (3.1), 300 was chosen as an appropriate fixed sequence length that could cover most sequences. The fixed sequence was prepared in such a way that the saccade started exactly at the 140th position. With a median saccade duration of 20, the saccade would be placed approximately at the center of the fixed distribution. Thus, only the preceding and proceeding fixations would be the subject of trimming and padding. Finally, the Y label was set to 1 if the direction recorded by the Eye Tracker for that duration was Right, otherwise the label was set to 0 if it was Left.

I created three different datasets using this preparation method:

1. Minimally Preprocessed Subset: Dataset prepared from a subset of the complete Minimally Preprocessed Dataset, consisting of 87 total participants with 1 experiment trial each. Relatively smaller subset allowed me to iterate and evaluate the model relatively quickly.

2. Minimally Preprocessed Dataset: Dataset prepared from the complete Minimally Preprocessed Dataset, consisting of 222 participants over 2 days with 2 trials per day.

3. Maximally Preprocessed Dataset: Dataset prepared from the complete Maximally Preprocessed Dataset, consisting of 222 participants over 1 day with 2 trials per day.

Since minimally preprocessed datasets had consistently outperformed maximally preprocessed datasets in [1], I chose to initially focus my experiments and evaluations on a subset of the minimally preprocessed dataset.

All outlier sequences identified during Data Exploration were removed.

Dataset	**Total**	**Train**	**Validation**	**Test**
Minimally Preprocessed Subset	40,092	32,268	3,912	3,912
Minimally Preprocessed Dataset	372,302	298,536	36,883	36,883
Maximally Preprocessed Dataset	189,363	151,895	18,734	18,734

Table 3.2: Datasets and their size

Chapter 4
Model

I used the CNN benchmark model designed by [1] as it is, without making any changes. Therefore, the architecture of the model was "a standard one-dimensional convolutional neural network with 12 layers and additive residual connections around blocks of three layers. Each layer consists of (1D-)convolution, batch normalization, ReLU activation and max pooling. In the convolutions we use 16 filters of size size 64, and for the pooling operation a kernel of size 2 and stride 1. Each residual connection performs a convolution followed by batch normalization."

The learning rate and the hyperparameters of the model were optimized using the validation dataset. I used Binary Cross Entropy Loss to train the model and used the Adam[6] Optimizer along with early stopping on the validation sets.

Chapter 5

Results

5.1 Initial Results on Minimally Preprocessed Subset

Here I shall present a summary on the results and evaluation of the model after multiple iterations on a fixed seed. Initially, the model returned an accuracy score of 0.9301 with the following metrics:

- False Positive Rate: 0.0512
- False Negative Rate: 0.9135
- Precision on Right: 0.9527
- Precision of Left: 0.9066

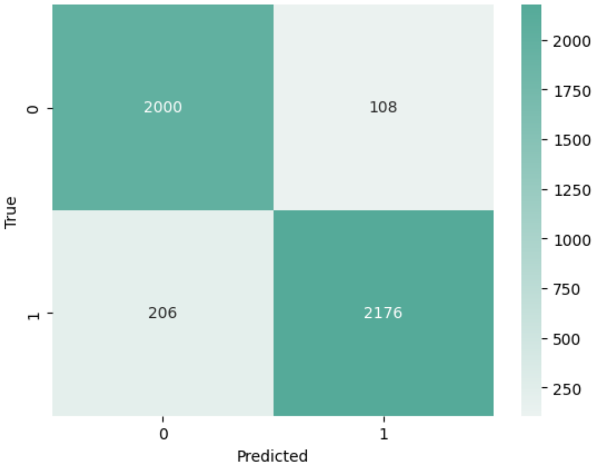


Figure 5.1: Confusion Matrix of the Initial Result on the Minimally Preprocessed Subset

These metrics can also inferred from the Confusion Matrix in 5.1. As we had hypothesized during Data Exploration, prediction on both classes performed almost equally well with no obvious signs of a class bias.

5.1.1 Model Confidence

Next, I evaluated how confident the model was in its correct and incorrect predictions, in an effort to explore areas for improvement. First, I plotted the probability values output by the model for all the correct predictions it made (5.2). For most of the predictions, the model was very confident as its probabilities were close to either 0 or 1. However, there was a conspicuous cluster of samples where the model was not confident at all, with samples from both classes returning probabilities close to 0.5. I also plotted the probability values output by the model for all the incorrect predictions it made (5.3). It can be observed that the model was confidently incorrect just as often as it in-confidently incorrect.

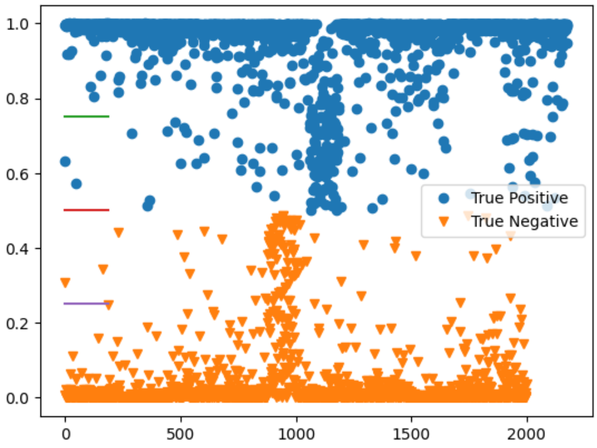


Figure 5.2: P(Y==1) for all correct predictions

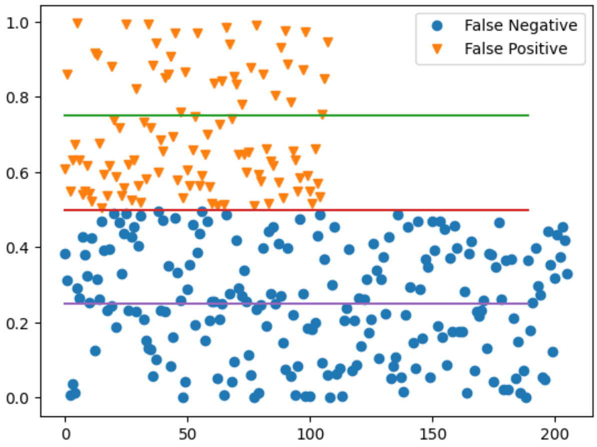


Figure 5.3: P(Y==1) for all incorrect predictions

5.1.2 Data from Electrodes 125 and 128

I decided to further analyze the relationship between True Positive, True Negative, False Positive and False Negative predictions by plotting their corresponding inputs. [7] showed that most of the important information for eye movement is highly concentrated in the frontal electrodes, with electrodes 125 and 128 being the most important. 5.4 shows the location of electrodes 125 and 128 near the eyes as well as the importance ranking of electrodes for a minimally preprocessed dataset.

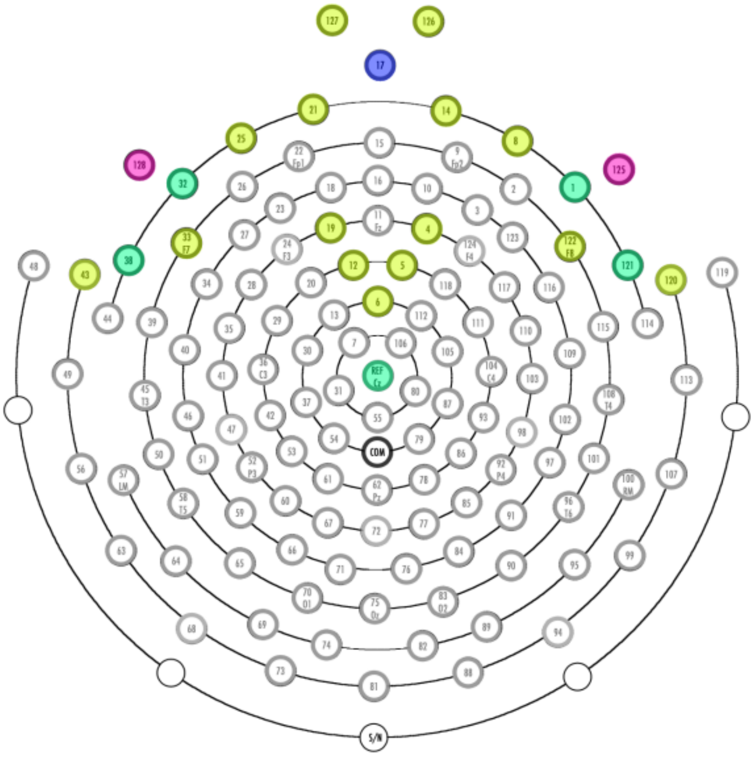


Figure 5.4: Electrode Clustering Visualization. This figure shows the electrode placement, with the most important electrodes for eye movement coloured pink, blue, teal and yellow in decreasing order of importance.[7]

Therefore, instead of plotting the input from all channels, I only plotted the inputs from electrodes 125 (5.5) and 128 (5.6). The first key observation is that the True Positive and True Negative samples have a clear visual pattern. For example, when the saccade starts (at time point 140), the plot of Electrode 125 for True Negatives drops significantly. The second observation is that there is a need for normalization, since even though a pattern exists, it does occur at the same scale with each sample. For example, the blue sample in the False Positive section of 5.5 is misclassified as Positive even though it exhibits the same pattern as the True Negative samples, just on a larger scale.

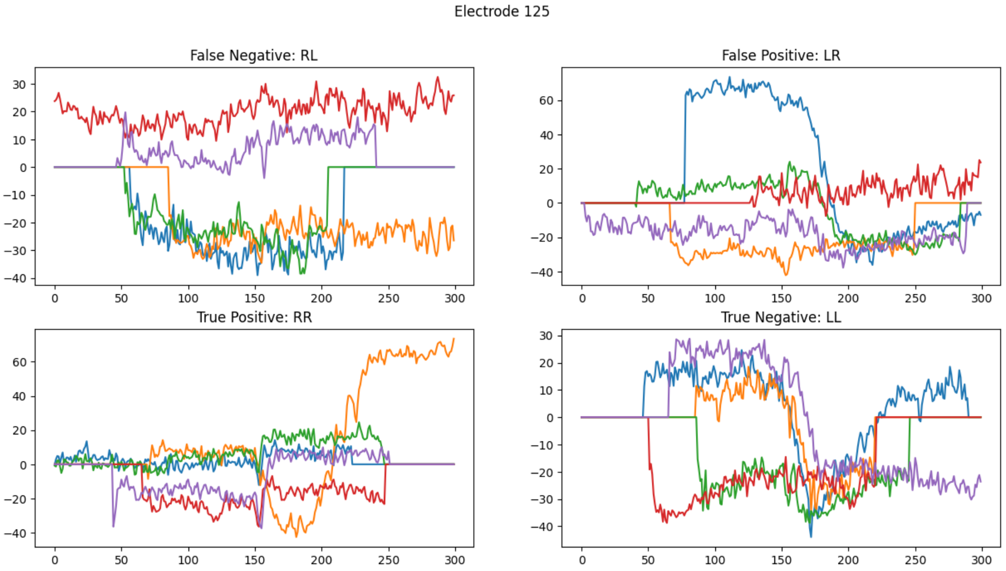


Figure 5.5: Plots of Electrode 125 in samples from the test dataset

5.1.3 Distribution of Incorrect Predictions by Subject

I also attempted to explore how correct and incorrect predictions were distributed amongst different subjects. Surprisingly, there were a few subjects that contributed towards a significant majority of incorrect predictions. For example, in one particular run, subject 'AR0' (labelled as 52 here) accounted for almost all of the prediction errors (5.7), despite having just as many total samples in the test dataset as other subjects (5.8).

Furthermore, inspecting the recordings from electrode 125 and 128 from these subjects revealed many exceptionally noisy samples (5.9. in fact, samples from these subjects were responsible for the conspicuous cluster of inconfident correct predictions seen in 5.2, and removing them completely tended to make the model much more confident in its correct predictions (5.10). Through multiple iterations, several other such subjects were found that were creating such out-sized errors. After analysing the preprocessing logs of the data from these subjects, it could be ascertained that Electrodes 125 and 128 for all of these subjects were detected as being of bad quality and thus, interpolated. It seems that the interpolation wasn't effective at modelling the true recordings from these electrodes, leading to such noisy samples and outsized errors.

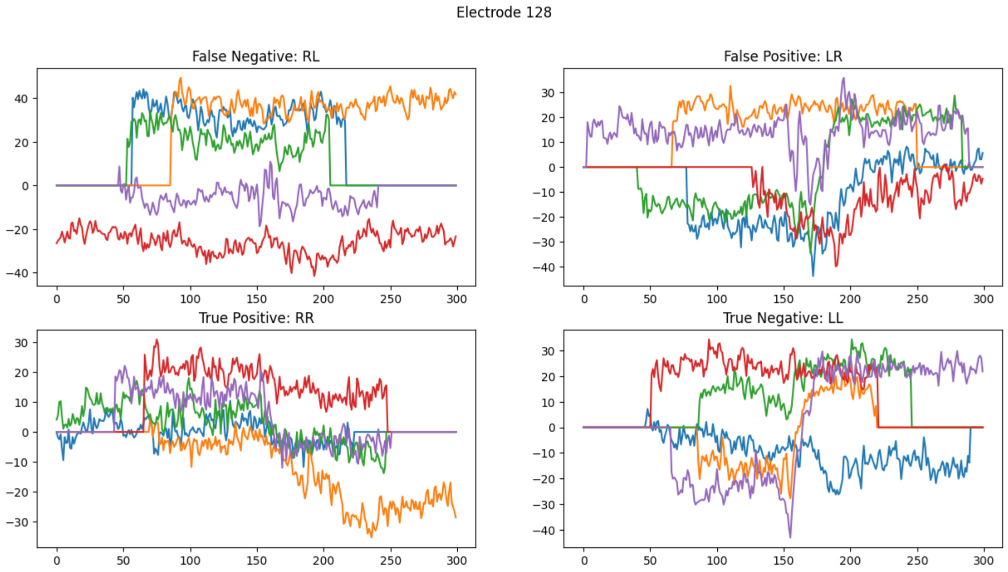


Figure 5.6: Plots of Electrode 128 in samples from the test dataset

5.1.4 Errors and Eye Tracking Data

Finally, I plotted the horizontal eye tracking data for False Positive and False Negative samples, to ascertain whether there were particular patterns of eye movement that the model was consistently failing to understand. The plots, shown in 5.11, go up if the saccade is in the "right" direction and go down if the saccade is in the "left" direction. The plots uncover two major sources of error. First is that the Eye Tracker sometimes fails and immediately goes to 0. This rapid change in value gets miscategorized as a left moving saccade within the VSS dataset, even though physically no such saccade occurs. The second source of error is that some singular saccades in the VSS dataset are multi-

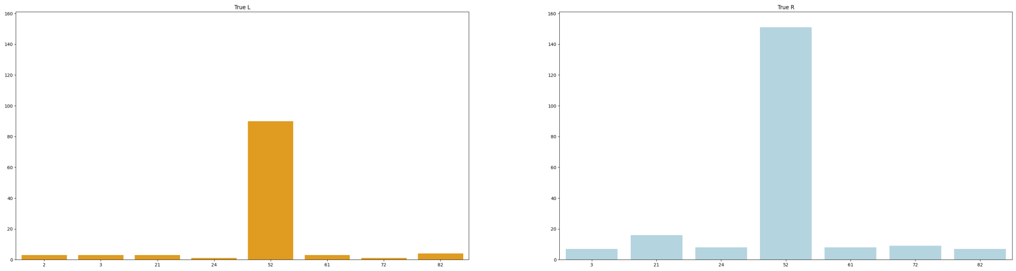


Figure 5.7: Number of prediction errors from each subject

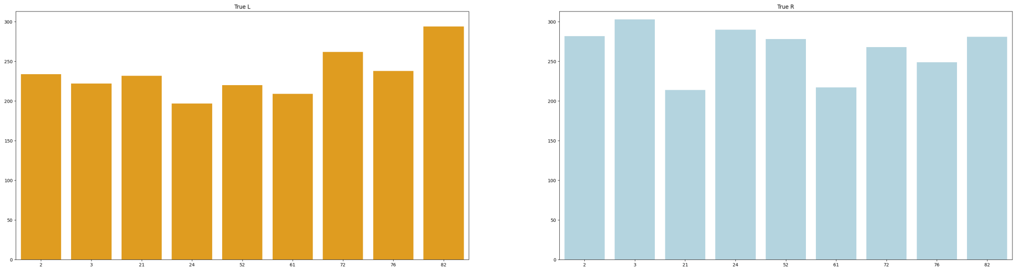


Figure 5.8: Number of samples in the test dataset from each subject

directional. Thus, defining the direction of a saccade based on just its start and end coordinates does not capture the actual physical movements taking place.

5.1.5 Key Insights

After training and evaluating the model on the minimally preprocessed subset several times, the following results emerge:

- Padding with zeroes provides better results than mirror padding or no padding.
- The padded data needs to be normalized using the Max Absolute Normalizer for best results.
- All subjects with interpolated electrodes 125 and 128 need to be removed from the training and test dataset.
- All multi-directional saccades and saccades where the x-coordinate recorded by the the Eye Tracker equals 0, need to be removed from the dataset.

5.2 Results on the Minimally Preprocessed Dataset

After implementing the insights derived from multiple iterations on the minimally preprocessed subset, I was able to achieve an average accuracy of 0.9968 on the complete Minimally Preprocessed Dataset.

- False Positive Rate: 0.0020
- False Negative Rate: 0.0040
- Precision on Right: 0.9983
- Precision of Left: 0.9950

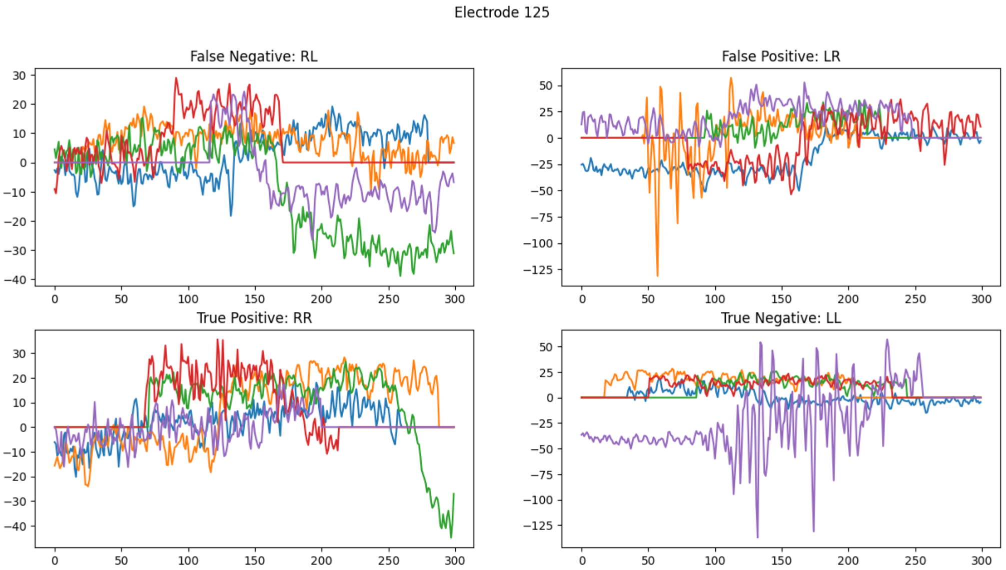


Figure 5.9: Plots of Electrode 128 in samples from subject 'AR0'

The average accuracy was even higher than the accuracy attained by any of the benchmark models in [1].

5.2.1 Distribution of Incorrect Predictions by Saccade Amplitude

As explained previously, I wanted to understand whether a dataset consisting of saccades with varying amplitudes would effect the model's ability to predict the direction of eye movement. 5.13 shows that the distribution of incorrect answers is very close to the original distribution of saccade amplitudes shown in 3.3 and thus, given the high score, it seems like the variance of the amplitude did not have a major effect on the model's ability to make predictions.

5.3 Results on the Maximally Preprocessed Dataset

Using the same model and data preparation technique, I was able to achieve an average accuracy of 0.9159 on the complete Maximally Preprocessed Dataset.

- False Positive Rate: 0.0918
- False Negative Rate: 0.0770
- Precision on Right: 0.9157

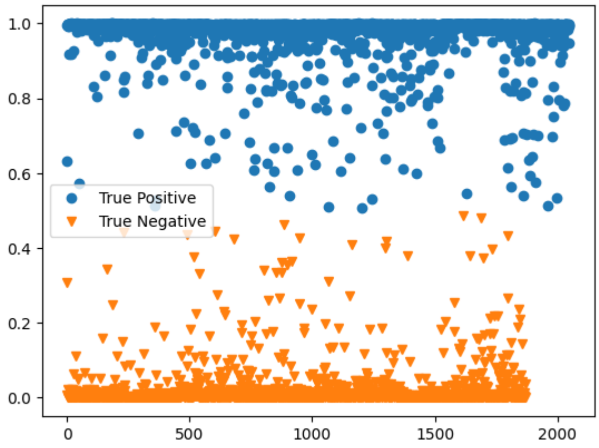


Figure 5.10: P(Y==1) for all correct predictions after removing samples from ’AR0’

- Precision of Left: 0.9161

Consistent with results from [1], the maximally preprocessed dataset performed worse than the minimally preprocessed dataset. Furthermore, most of the model’s correct prediction also had low confidence as seen in 5.14.

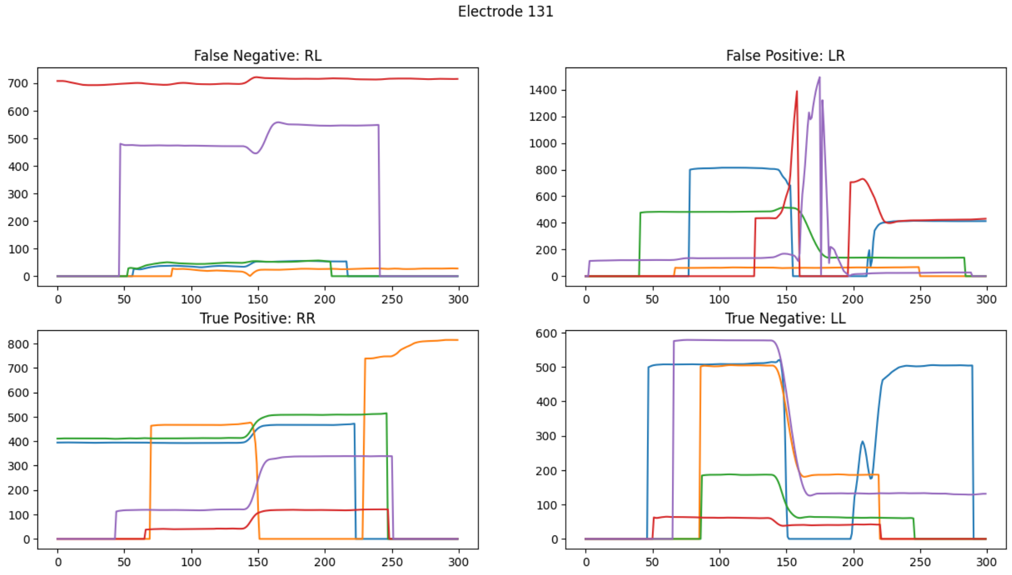


Figure 5.11: Horizontal Eye Tracker

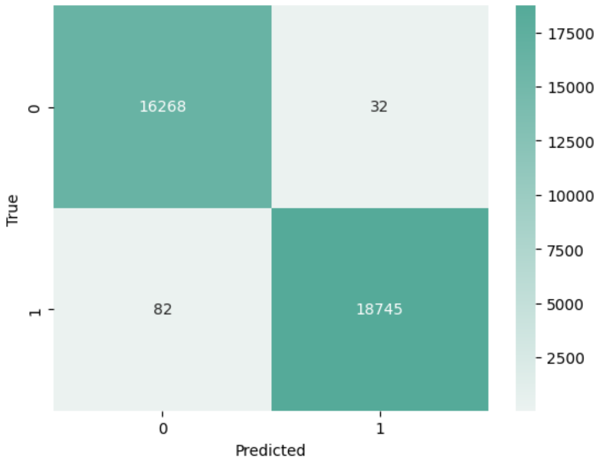


Figure 5.12: Confusion Matrix of the Result on the Minimally Preprocessed Dataset

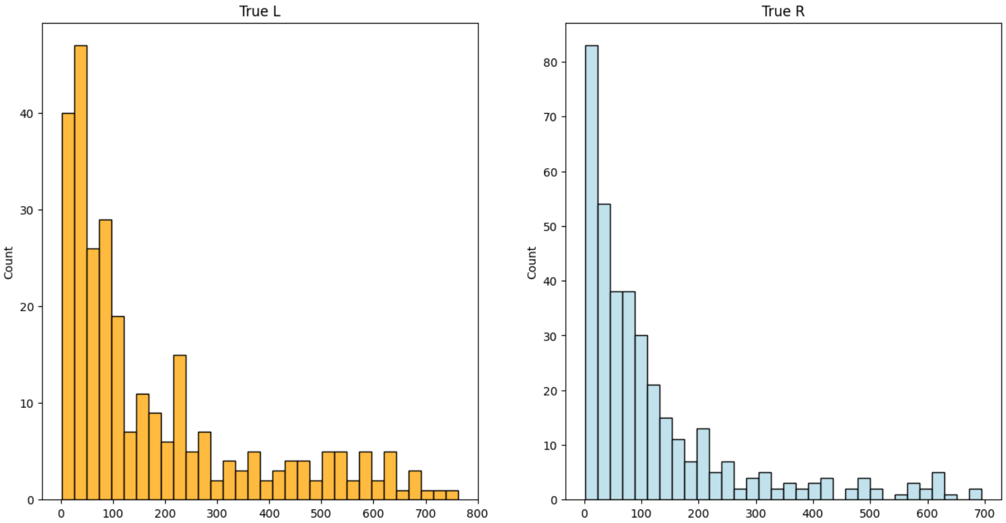


Figure 5.13: Distribution of Incorrect Predictions by Saccade Amplitude

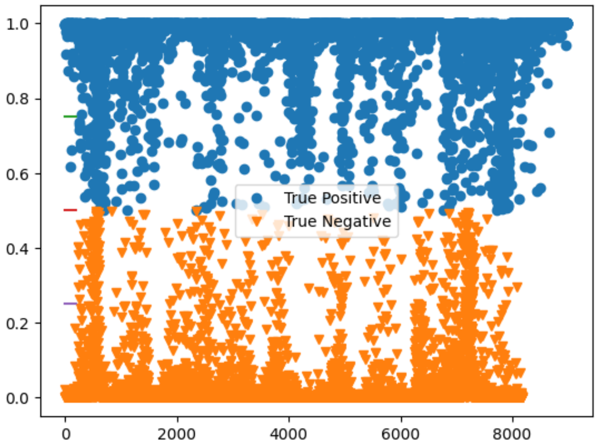


Figure 5.14: P(Y==1) for all correct predictions

CHAPTER 6

Discussion

We were able to successfully demonstrate that the CNN benchmark model described in [1] is robust enough that, with analysis based appropriate changes to data preparation, it can act as an excellent classifier in "realistic" datasets like the VSS, where the participant is focused on cognitive tasks and their gaze is unprompted and uncontrolled. A limitation of this thesis is that it was only able to establish this claim for the CNN benchmark model. Given more time, this research could be generalized to test the performance of modified and unmodified versions of other benchmark models on datasets such as the VSS as well. Even though the left-right direction task is the easiest of the gaze estimation tasks, these results pave the way for future work in implementing such models to estimate saccade angle, saccade amplitude, and absolute position on unstructured "realistic" datasets.

Chapter 7

Conclusion

For EEG based models to complement and/or replace expensive Eye Trackers, they need to be robust enough to perform well even on unstructured and noisy datasets such as the VSS, not just highly controlled laboratory experiments such as the Pro/Anti-Saccade Paradigm. My aim with this thesis was to not only test the robustness of one of the benchmark models but also to explore what changes in methodologies towards data preparation might be necessary to make the newer unstructured datasets compatible with the benchmark models. This thesis shows that an investigative and iterative approach to data exploration and model evaluation is necessary to transform data in such a way that not only errors are minimized but the model predictions are also confident.

Bibliography

[1] A. Kastrati, M. B. Płomecka, D. Pascual, L. Wolf, V. Gillioz, R. Wattenhofer, and N. Langer, "Eegeyenet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction." arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2111.05100

[2] J. R. Bergstrom and A. J. Schall, *Eye tracking in User Experience Design*. Elsevier, 2014.

[3] R. Rivu, Y. Abdrabou, K. Pfeuffer, A. Esteves, S. Meitner, and F. Alt, "Stare: Gaze-assisted face-to-face communication in augmented reality," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '20 Adjunct. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3379157.3388930

[4] M. Plöchl, J. Ossandón, and P. König, "Combining eeg and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data," *Frontiers in Human Neuroscience*, vol. 6, 2012. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2012.00278

[5] A. Pedroni, A. Bahreini, and N. Langer, "Automagic: Standardized preprocessing of big eeg data," *NeuroImage*, vol. 200, pp. 460–473, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811919305439

[6] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[7] A. Kastrati, M. B. Plomecka, J. Küchler, N. Langer, and R. Wattenhofer, "Electrode clustering and bandpass analysis of eeg data for gaze estimation," in *NeuRIPS 2022 Workshop on Gaze Meets ML*. s.l.: OpenReview, 2022-12, Conference Paper, gaze Meets ML Workshop 2022; Conference Location: New Orleans, LA, USA; Conference Date: December 3, 2022; Conference lecture on December 3, 2022.