SPIKE SORTING USING DEEP LEARNING

PhD thesis book

János Rokai, MD

János Szentágothai Doctoral School of Neurosciences Semmelweis University

Supervisor: Gergely Márton, Ph.D

Official reviewers: Péter Barthó, Ph.D

Csaba Dávid, Ph.D

Head of the Complex Examination Committee:

Alán Alpár, MD, D.Sc

Members of the Complex Examination Committee:

László Acsády, D.Sc

Zoltán Somogyvári, Ph.D

Budapest

2023

1. Introduction

Extracellular recordings in the central nervous system (CNS) provide information on neural activity patterns that can be valuable both for researchers in the field of neuroscience and for developers in the brain–computer interface industry. In order to analyze these neural patterns, the sources of single neuronal activities (single-units, spikes) need to be identified and clustered (spike sorting). To increase the precision of spike sorting and the number of recorded extracellular activity (spikes) from neurons, high-density neural microelectrode arrays (MEAs) are used, which are implanted into the CNS. The number of recording sites on the MEAs is growing rapidly by which the recorded data is also growing, making an automated, robust, input-source agnostic spike sorter an increasingly valuable asset.

Neural activities are usually recorded with sampling rates between 20–30 kHz. In order to remove local field potential, low- and high-frequency noises to more reliably identify single-cell activities, a bandpass frequency filter is applied to the recordings between 0,3–3 kHz (or 0,5–5 kHz).

Manual curation is no longer a viable option for interpreting raw data due to the increased number of channels, by which the time of the manual curation increases as well. Subjective bias is also present in manual curation based on the experience of the curator.

To automate the detection and clustering of the spikes, spike sorting algorithms are developed to speed up the processing of high-channel-number recordings. Conventional spike sorting algorithms comprise three main processes: spike detection, feature extraction, and clustering of the features. Spike detection is usually filtering the wideband data from local field potential signals, low and high-frequency noises. This filtering is often performed in the frequency domain and typically falls between 300 and 3000 Hz. Once the data has been filtered, action potentials can be detected using various methods, like threshold-based detection, energy operators or wavelet decomposition. After successful data filtering and action potential detection, the spikes need to be realigned for further analysis. One of the most popular dimensionality reduction methods is principal component analysis (PCA). The third step the clustering plays a crucial role in decoding extracellular action potentials. For clustering algorithm, a plethora of different methods were proposed, like Bayesian models, K-means clustering, Agglomerative clustering, Superparamagnetic clustering, ISO-SPLIT and others.

Modern state-of-the-art algorithms are performing offline spike sorting on a high-performance PC. However, a plethora of potential applications would benefit from on-site spike sorting. In order to evaluate the raw data on-site, allowing for building closed-loop systems, several hardware implementations were suggested to create an online embedded spike sorting system. The on-chip spike sorting solutions however sacrifice precision for speed, while they are also limited in the number of channels, they can efficiently process data from.

2. Objectives

- One of the aims was to develop a semi-supervised spike sorting system that leverages the promising technologies of unsupervised learning and loose supervision in the realm of deep learning. The system will incorporate detection, feature extraction, and sorting components, allowing for a flexible and adaptable approach to accommodate varying experimental conditions and datasets.
- Another focus of this thesis is to address the inflexibility of current deep learning solutions in spike sorting, which often rely on fixed numbers of clusters and waveform identifications. The goal is to develop a system that eliminates these constraints while harnessing the power of deep learning techniques. The system will adapt to changes in parameters across different experiments, leading to improved spike sorting performance and adaptability.
- Investigate the possibilities of deploying spike sorting systems in embedded environments, considering the limitations of relatively large computational power in current state-of-the-art solutions. The objective is to develop a spike sorting system that achieves performance comparable to state-of-the-art methods while being efficiently deployable in on-site and resource-constrained scenarios.

3. Methods

3.1. Datasets

The Fiath dataset consists of nine recordings obtained from a high-density silicon MEA with 128 channels targeting different neocortical areas. Spike labels were established using the KiloSort algorithm and manually refined. Additional custom scripts were applied to remove low-amplitude waveforms.

Paired datasets, such as the Kampff dataset, Boyden dataset or the Yger dataset, provide ground-truth references for extracellular recordings, but their use is limited to evaluating spike detection systems due to the availability of ground-truth labels for only one neuron.

The Hybrid Janelia dataset combines real-world recordings with synthetic data to create a comprehensive dataset for evaluating spike sorting solutions in conditions closely mimicking real-world extracellular recordings, offering recordings with different channel counts.

3.2. Metrics

To evaluate the performance of spike sorting algorithms, various evaluation metrics have been established. In this work, we discuss several commonly used evaluation metrics, including Recall, Precision, F1 Score (Micro, Macro, Weighted), and Accuracy. These metrics assess the algorithm's ability to correctly identify true positives, minimize false positives, and provide a balance between precision and recall. However, conventional evaluation metrics may not capture every aspect of spike sorting.

To address these limitations, we introduce custom evaluation methods that offer more comprehensive insights into the efficiency and efficacy of spike sorting algorithms. The xSpeed metric evaluates the relative performance of sorting speed compared to the actual recording duration, providing a quantitative indicator of algorithm efficiency. The Mean Embedding Similarity (MES) matrix and Distance Between Clusters (DBS) matrix assess cluster separability in the latent space generated by the model's encoder. A Combined matrix integrates both matrices, offering a holistic view of cluster separability based on spatial positioning and feature characteristics. Moreover, we introduce the Template Embedding Similarity (TES) matrix, which investigates the distances between embeddings of cluster-wise averaged waveforms.

3.3. Semi-supervised architecture

The basis of the semi-supervised model (ELVISort) is a β-Variational autoencoder, which is customized to fit the target task. The input of ELVISort is a 2D matrix of electrophysiological signals, where rows correspond to channels and columns correspond to sampling points in time. A subsidiary goal was to train the network to effectively reconstruct the different input patterns from their compressed representations, which are coded by the different states of the latent space of the autoencoder. A proper representation offers the possibility of distinguishing spikes originating from different sources. To achieve this, multiple branches are used while training the autoencoder to ensure the emergence of a well-balanced latent space which is useful for classification and sorting as well.

In spike analysis, time-domain feature extraction is as important as the inspection of space-domain-specific inter-channel relations. To exploit this concept, the main elements of ELVISort are long short-term memory (LSTM), bidirectional LSTM (Bi-LSTM) and 2D convolutional layers.

The encoder consists of two different branches: the LSTM-based branch processes data in the time domain, having a 2-dimensional matrix as input while the 2D CNN branch extracts spatiotemporal features from a 3-dimensional input. For the convolution branch 4 building blocks from GoogLeNet were included beside dropout and convolutional layers. The outputs of the LSTM and CNN branches are concatenated and combined non-linearly using fully connected layers. The last encoder layer outputs the mean and variance of the latent inference.

In the reconstruction branch, only LSTM elements were used. A custom layer was implemented to handle the inference of the latent variables based on their mean and variance approximated by the encoder. The latent space was constricted to a size of 32 in order to improve clustering. To further compress information, a hierarchical latent layout was used, moreover fully connected layers were applied to the latent variables to further decrease the size of the latent space.

3.4. Training of edgeTPU-compatible model

The training of edgeTPU-compatible model was conducted in two phases. First an unsupervised model was trained to extract relevant information from 1D waveforms. Subsequently, in the second phase, a supervised model was trained to detect spikes and predict the feature vector previously learned. For the unsupervised part, the nearest-neighbor contrastive learning (NNCLR) was chosen.

During training, pairs of inputs of the same cluster are given to the model and fed through the same encoder, which produces a feature vector for each input. These feature vectors are then processed by a projection head, and the NNCLR loss is calculated based on the output of this projection head.

The base model for NNCLR was constructed using Residual blocks, 1D convolution layers, Dense layers, and Batch normalization layers. Together, these layers work to transform the input sample into a compact, low-dimensional feature vector that represents the underlying patterns in the data. The model depth is quite shallow, to enhance stability and avoid overfitting. The input pairs for the self-supervised model were formed by 1-dimension waveform instances and the mean waveform of the given cluster.

For the supervised model, the single-shot detector was adopted, utilizing MobileNetV2 due to its lightweight architecture, edgeTPU support, and suitability for systems with limited computational resources. To enhance the model's performance, customizations were made to MobileNetV2, doubling the output dimensions while preserving the depth of the original model.

For the supervised model, inputs in the form of snippets were constructed, encompassing all channels from a specific recording and maintaining a fixed timespan. Within these snippets, the single-shot detector was tasked with identifying different spikes and predicting the feature-vector of each spike based on the previously learned feature-space from the NNCLR unsupervised method.

The model was deployed within an embedded environment, with the chosen EdgeTPU chip as the basis. This chip is built around a specialized Tensor Processing Unit (TPU) designed for deep-learning tasks and demonstrates impressive efficiency, consuming only 1 Watt for 2 Tera Operations Per Second. To achieve such efficiency, the chip only supports quantized models. Consequently, the model underwent quantized-aware training to minimize performance degradation during the quantization process.

To assess the model's speed, evaluations were conducted on two different TPU devices: the Coral Development Board Mini and the Coral USB Accelerator. Throughout the evaluation process, measurements were taken for both the inference speed of the model on the TPU chip and the additional time needed by the non-max suppression postprocessing step.

4. Results

4.1. Results of the semi-supervised solution

The performance of ELVISort was evaluated on various datasets, including Kampff, Fiath, and Hybrid Janelia. The algorithm's effectiveness was compared to state-of-the-art spike detectors and sorters, and it consistently performed well, outperforming some popular algorithms.

The model trained and tested on the Kampff dataset managed to produce excellent results, where it achieved an impressive F1 score of 0.964 and an accuracy of 95%.

The Fiath dataset was used to test ELVISort's detection and classification performance. The proposed method achieved an average F1 score of 85.55% for the validation set and 82.42% for the test set.

ELVISort's combined model was also tested on the Hybrid Janelia dataset, demonstrating consistent performance with and without the non-spike cluster. The algorithm's performance per cluster was evaluated, particularly regarding true positives (TP) versus falsely matched snippets (false positives + false negatives, FP+FN).

4.2. Results of the edgeTPU compatible deep learning solution

Using the NNCLR method, a highly separable latent space is obtained. To demonstrate the effectiveness of our approach in creating a general embedding space, and the overlapping clusters can be indeed resolved by using channel information, we generated similarity matrices to analyze the distinguishability of various clusters. In order to further examine the separability of the clusters, we also included channel-distance information between the clusters in our analysis. This was necessary because the hybrid recordings we used to train our model contained similar waveforms that were used to generate different clusters on different channels. The combination of both types of information resulted in a highly separable matrix, demonstrating the ability of our model to create a general embedding space that is able to effectively separate different waveforms. Feature prediction and the detection of the individual spikes were assessed separately as well. To assess the performance of the detection of our model, we used paired recordings. This allowed us to compare the results to those of other existing solutions. We demonstrate that our model performs very well in terms of spike detection and is able to generalize to new recordings with different electrode parameters and waveform types. In fact, the results show that our model performs better and more consistently than current state-of-the-art methods, even though it is specifically designed for use on embedded systems.

A separate assessment was made for the two hybrid datasets, where detection, sorting and the combination of the two was considered. The detection performance has a quite large gap between the two recordings: one of the probable explanations for this is that for the HS_64_12 recording cluster with the smallest SNR has an SNR value of 4.38, while for the HS_32_32 the minimum SNR is 0.34. The sorting of the found spikes show a more robust performance: while the Isosplit5 algorithm provides a faster sorting, the agglomerative clustering has a better performance on the generated feature space, however being the slower one.

We tested the inference speed of our model on 128-channel samples in three different scenarios: a completely PC-based setup, where high performance CPU and GPU is available; a hybrid setup where high performance CPU is coupled with a TPU-based USB Accelerator, and a development-board-based setup, where a lower performance CPU is coupled with a TPU. The first setup was obviously the fastest, while the last one was the slowest one.

5. Conclusions

- The application of semi-supervised deep learning methods to spike sorting has yielded promising results. The proposed deep learning model, ELVISort, leverages the β-VAE architecture to efficiently detect and sort spikes. ELVISort successfully reduces the input data size to less than 0.5% of its original dimensions, thereby achieving notable gains in memory and time efficiency during the clustering process. The model's performance was rigorously assessed using publicly available datasets, demonstrating commendable F1 scores on both the Hybrid Janelia and Kampff datasets. These findings underscore the potential of ELVISort as a valuable tool in the development of memory and time-efficient brain-computer interfaces in the future.
- The presented edgeTPU-compatible system is designed to be scalable in two key ways. First, it can be trained on multiple datasets simultaneously, allowing the system to be input-source agnostic, meaning it can be trained on data from different sources without requiring any prior knowledge of the recording conditions or electrode geometry. Second, the system can also be scaled in terms of architecture by using state-of-the-art methods to improve both the self-supervised and supervised components of the model.
- In the realm of deep learning solutions for edge devices, our model for edgeTPUs distinguishes itself as the pioneering deep learning-based solution capable of accommodating recordings with a high number of channels, while being deployable on embedded

systems, particularly on TPUs and at the same time, being able to exhibit a performance similar to existing state-of-art methods on unseen recordings.

6. Bibliography of the candidate`s publications

6.1. Papers closely related to the PhD dissertation

Rokai J, Ulbert I, Márton G. Edge computing on TPU for Brain Implant Signal Analysis. Neural Networks. 2023;162:212–24. doi:10.1016/j.neunet.2023.02.036

Bod RB, Rokai J, Meszéna D, Fiáth R, Ulbert I, Márton G. From end to end: Gaining, sorting, and employing high-density neural single unit recordings. Frontiers in Neuroinformatics. 2022;16. doi:10.3389/fninf.2022.851024

Rokai J, Rácz M, Fiáth R, Ulbert I, Márton G. ELVISort: Encoding latent variables for instant sorting, an artificial intelligence-based end-to-end solution. SSRN Electronic Journal. 2020; doi:10.2139/ssrn.3699796

Rácz M, Liber C, Németh E, Fiáth R, Rokai J, Harmati I, et al. Spike detection and sorting with Deep Learning. Journal of Neural Engineering. 2020;17(1):016038. doi:10.1088/1741-2552/ab4896

6.2. Other publications

Rokai J, Ulbert I, Márton G. Improving ECoG recordings with missing electrodes for classification through deep learning. FENS Regional Meeting, Algarve, Portugal, 2023

Rokai J, Ulbert I, Márton G. Deep learning-based spike sorting on edge devices. FENS Forum, Paris, 2022

Rokai J, Rácz M, Fiáth R, Ulbert I, Márton G. ELVISort: Encoding latent variables for instant sorting. IBRO Workshop, Budapest, 2022

Rokai J, Fiáth R, Ulbert I, Márton G - Two phase spike detection using deep learning. IBRO Workshop, Szeged, 2020

Matusz VI, Rácz M, Rokai J, Jorgosz NE, Molnár T, Maraki D, Ulbert I, Márton G. Head-mounted, wireless eyetracker for real-time gaze prediction utilizing machine-learning. 2020 IEEE International Conference on Human-Machine Systems (ICHMS). 2020

Rácz M, Liber C, Németh E, Fiáth R, Rokai J, Harmati I, Ulbert I, Márton G - Spike detection and sorting with Deep Learning. IBRO Workshop, Szeged, 2020