Coordinate-VAE: Unsupervised clustering and de-noising of peripheral nervous system data

Abstract

The peripheral nervous system represents the input/output system for the brain. Cuff electrodes implanted on the peripheral nervous system allow observation and control over this system, however, the data produced by these electrodes have a low signal-to-noise ratio and a complex signal content. In this paper, we consider the analysis of neural data recorded from the vagus nerve in animal models, and develop an unsupervised learner based on convolutional neural networks that is able to simultaneously de-noise and cluster regions of the data by signal content.

1 Introduction

Recent advances have made chronic observation [1] of, and limited control [2] over the peripheral nervous system possible. In order to characterise the dynamics of the signals passing to and from the brain, we wish to categorise the patterns of activity we observe within the peripheral nervous system.

However, consistent detection of single neuron activity remains a challenge with current cuff electrode technology suitable for *in vivo* neural data acquisition. The relative position of an extracellular recording electrode and neuronal axons close enough to sit above the noise floor, affects the polarity of presented signal components [3]. Their summation at the electrode occludes the presence of individual action potentials during periods of neuronal activity. Instead, local field potentials, the combination of many selective neuronal responses arriving concurrently at the electrode are observed. These population level responses are potentially informationally richer [4, 5], but preclude the use of conventional spike-sorting [6] methodologies on such data.

Instead, we develop here a method based on convolutional neural networks that simultaneously de-noises the data and categorises the observed signals. We show that we are able to train this model on approximately one hour of data taken from a single subject approximately twelve hours post surgical implantation. Moreover, we show that it is applicable without further training to data taken from a second subject, sixty days post surgical implantation.

2 Methods

Neural data are collected from two nerve cuffs implanted on the vagus nerve, each recording local field potential data at a sampling rate of 30000Hz using a chronically implanted ITX PNS (peripheral nervous system) implant (BIOS, Cambridge, UK). We begin by standardising the mean and standard deviation of the data coming from each cuff electrode, and applying a fifth-order Butterworth bandpass (50-1000Hz) filter. Given an arbitrary starting time, we divide the data into small time windows of equal size w with a constant step size s between windows. If $s < w$, the windows will overlap.

For each window, the goal is to reduce the data observed in that window to a one-hot latent vector. We train a variational auto-encoder [7] that encodes the information contained within a signal window into a sparse latent space vector using a Gumbel-Softmax distribution [8]. Models were trained in Tensorflow (v 1.12.2) on a single Nvidia Tesla K80. Hyperparameter tuning was carried out over one thousand evaluations for each model using a Tree-structured Parzen Estimator as implemented in the `hyperopt` package [9].

For the primary data set, the data were divided at random into training/validation/test sets comprising 70%, 20% and 10% of the data respectively. Models were trained over 2000 epochs, with training ending early if no improvement in the loss metric was seen for more than 200 epochs.

2.1 Model architecture

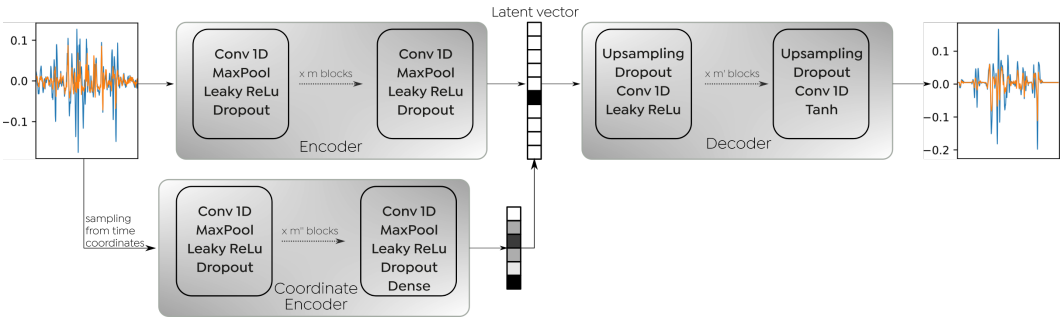


Figure 1: The architecture of the Coordinate-VAE. The input signal is encoded *via* a series of convolutional/pooling/leaky-relu/dropout blocks to a categorical latent vector representing the core process observed in the input signal. To allow the decoder to account for phase shifts, time warping, *et cetara*, a set of time coordinates for which the signal is closest to zero are sampled from each channel of the input signal. These pass through a 'coordinate encoder', before being concatenated with the latent vector. The decoder then upsamples with convolution to reconstruct the original signal.

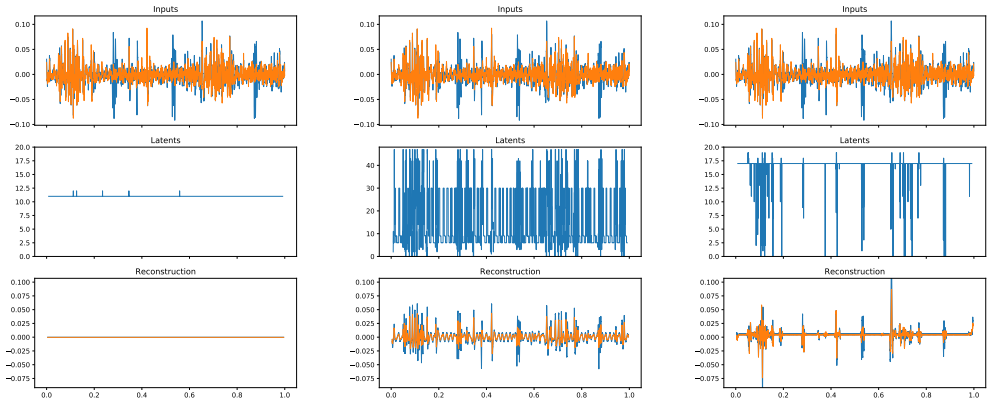
The basic model architecture is shown in Figure 1. We begin with a standard auto-encoder architecture with a Gumbel-Softmax [8] activation on the latent space. Encoding to the latent space is done through a series of convolutional blocks, where parameters for each block in the encoder are kept constant except for the number of filters in each convolutional layer, which doubles with each block. Pooling takes place in each block where this would not reduce the dimension of the data to less than the size of the convolutional layer. Decoding similarly follows a standard upsampling/convolutional scheme, with a hyperbolic tangent activation following the final convolutional layer. The temperature of the Gumbel layer is slowly annealed throughout training such that the temperature at epoch E is $2e^{-0.0003E}$. During inference, the temperature is fixed at 0.1. We define the loss as a weighted sum of the mean squared error on the reconstruction and the negative of the Kullback–Leibler divergence.

With a small (≈ 20 dimensional) one-hot latent space, a standard auto-encoder is unable to reconstruct any signal (Fig 2(a)). Given a sufficiently large (≈ 50) latent space, there is sufficient information to reconstruct the signal, but at the cost of retaining much of the noise, signal artefacts, and increasing the complexity of the latent space (Fig 2(b)). We solve this by allowing the leakage of some information directly from the original signal to the decoder, bypassing the latent space. For each channel c, we find the set of n time-coordinates at which the observed signal is closest to zero. To prevent memorisation of the signal based on these coordinates, we randomly sample a subset n' of these coordinates. Since these data can be ordered over time, we then apply a 1-d convolutional network to this input in a similar fashion to the encoder. Following these convolutional layers, we have a encoding of the signal as defined by the sampling from the coordinates. This 'coordinate encoding' is concatenated to the upsampled layers in each step of the decoder. This allows a small (20-dimensional) one-hot latent space to identify the signal present in the data while removing the noise (Fig 2(c)).

3 Results

Figure 2 demonstrates the ability of a Leaky-VAE model to effectively de-noise peripheral neural data and cluster the observed signals within a relatively simple latent space. Furthermore, we can apply the model trained on data from a single subject to other subjects.

Figure 3 shows the latent space and reconstructed signal from vagus nerve recordings from a second subject taken sixty days post surgical implantation. Despite the increased noise levels in this data set, the trained model is able to de-noise the signal and characterise the signals within the data. Data from this subject were human-labelled as containing regions of neural activity corresponding to respiration modulation. There is a clear correlation between respiration modulation events and the amplitude of



(a) No coordinate encoding, small latent space.

(b) No coordinate encoding, large latent space.

(c) Coordinate encoding, small latent space

Figure 2: Latent space and signal reconstructions inferred from a second of input data from the test set. The input data from the first (blue) and second (orange) cuffs is reduced to a single value in the latent space which evolves over time, and is used to reconstruct the original signal. Without the coordinate encoding, no reconstruction is possible using a small latent space (a). With a large latent space (b), reconstruction is possible but at the cost of a complex latent space and the reconstruction of noise in addition to signal. With a coordinate encoder (c) the latent space is relatively simple and the reconstruction is effectively de-noised.

the reconstructed signal, suggesting that the latent space is able to capture meaningful physiological signals from neural data.

4 Discussion

The recent developments of chronic neural interfacing implant systems that are able to record neural signals over period of months or years will create large sets of primarily unlabelled data, with numerous signals occurring over a range of time-scales. These data are currently largely uncharacterisable with standard methods (e.g. spike-sorting). Previous work in this field has relied on mixing categorical and real-valued latent vectors. Westhuizen *et al* [10] used an adversarial auto-encoder to project neural data to labels, incorporating an approximately one-hot encoding in the latent space from which labels may be derived, but also includes an approximately Gaussian vector to allow reconstruction. Since both vectors are trained simultaneously, the Gaussian component of the latent space may contain the labelling information for one or more true classes. InfoGAN [11],

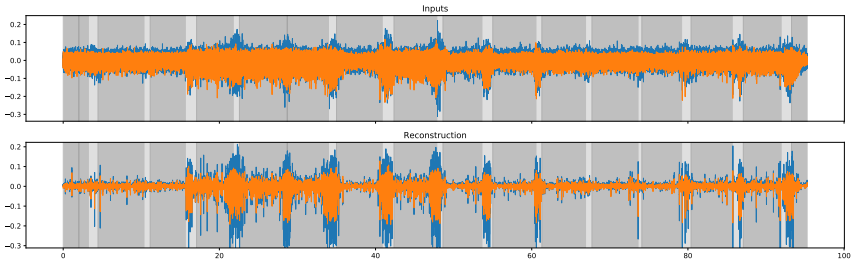


Figure 3: Input data and reconstructed signal over ≈ 100 seconds from a second subject. The trained model Data were human-labelled as containing respiration modulation neural activity (light-grey) or not (dark-grey). The reconstructed signal clearly identifies the respiration modulation events.

a GAN implementation in which the discriminator is required to identify components of the latent space, is similarly capable of one-hot latent representation of the data, but without constraints on the information carried within the one-hot encoding.

The Coordinate-VAE approach, in restricting the information available to the encoder creating the non-categorical portion of the latent space, allows unsupervised characterisation of the signals in time-series data, while simultaneously de-noising the signal. Models are transferable between individuals, suggesting that we may gain the ability to pre-train large models for the reduction to latent space representations. We might then rapidly train a final classifier or agent for monitoring or control of individual patients, as in Pandarianth *et al* [12] in which an auto-encoder is used as a dimension reduction technique on collections of neural spiking data acquired from macaque motor and pre-motor cortices, following which a GLM is used to map the complex latent space to spiking activity.

References

[1] Christopher Russell, Aidan D. Roche, and Samit Chakrabarty. Peripheral nerve bionic interface: a review of electrodes. *International Journal of Intelligent Robotics and Applications*, 3(1):11–18, mar 2019.

[2] Matthew Schiefer, Daniel Tan, Steven M Sidek, and Dustin J Tyler. Sensory feedback by peripheral nerve stimulation improves task performance in individuals with upper limb loss using a myoelectric prosthesis. *Journal of neural engineering*, 13(1):016001, feb 2016.

[3] György Buzsáki, Costas A. Anastassiou, and Christof Koch. The origin of extracellular fields and currents-EEG, ECoG, LFP and spikes, jun 2012.

[4] Guillaume Lajoie, Jean-Philippe Thivierge, and Eric Shea-Brown. Structured chaos shapes spike-response noise entropy in balanced neural networks. *Frontiers in Computational Neuroscience*, 8, oct 2014.

[5] Stefano Panzeri, Jakob H Macke, Joachim Gross, and Christoph Kayser. Neural population coding: combining insights from microscopic and mass signals. *Trends in cognitive sciences*, 19(3):162–72, mar 2015.

[6] Hernan Gonzalo Rey, Carlos Pedreira, and Rodrigo Quian Quiroga. Past, present and future of spike sorting techniques, oct 2015.

[7] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. 2013.

[8] Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. 2016.

[9] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. Hyperopt: a Python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1):014008, 2015.

[10] Jos van der Westhuizen, T E J Edwards, Raphael Schmetterling, Robert Tinn, Oliver E Armitage, Joan Lasenby, and Emil Hewage. Using adversarial autoencoders to infer actions from the peripheral nervous system, 2017.

[11] Lujie Chen, Artur Dubrawski, Donghan Wang, Madalina Fiterau, Mathieu Guillame-Bert, Eliezer Bose, Ata M. Kaynar, David J. Wallace, Jane Guttendorf, Gilles Clermont, Michael R. Pinsky, and Marilyn Hravnak. Using Supervised Machine Learning to Classify Real Alerts and Artifact in Online Multi-signal Vital Sign Monitoring Data. *Critical Care Medicine*, 44(7):e456, 2016.

[12] Chethan Pandarinath, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, Jonathan C. Kao, Eric M. Trautmann, Matthew T. Kaufman, Stephen I. Ryu, Leigh R. Hochberg, Jaimie M. Henderson, Krishna V. Shenoy, L. F. Abbott, and David Sussillo. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15(10):805–815, 2018.