CS230

Detecting Epileptic Seizures in Electroencephalogram Data

Nick Hershey
Department of Statistics
Stanford University
nhershey@stanford.edu

Abstract

Seizures are aberrations in the brain's normal electrical activity that produce disruptive physical symptoms. To diagnose and characterize seizures, clinicians measure the brain's electrical activity using electroencephalography (EEG), a method that involves placing electrodes on a patient's scalp to measure voltage differences across the brain. Due to the variety of seizure types and different presentations between people, no general model has been made to identify seizure from non-seizure EEG data. The goal of this paper is to build such a general recognition model. To do so, we used a data set of 99,721 EEG files recorded at Stanford Hospital over the last number of years, significantly more data than any other EEG classification paper has used. We examined the performance of logistic regression, dense neural networks, convolutional neural networks, and recurrent neural networks and found that the a 5-layer convolutional neural network followed by two dense layers performed best. We achieved 95% accuracy, significantly higher than any published work on a general cross-patient model.

1 Introduction

Affecting fifty million people worldwide, epilepsy is a chronic disorder of the central nervous system characterized by recurrent seizures Megiddo et al. (2016). During a seizure, aberrations in the brain's electrical activity produce physical symptoms ranging from convulsion to loss of memory to unconsciousness. An electroencephalogram (EEG) is a record produced by electrophysiological monitoring of the electrical activity of the brain. Electrodes are placed on the scalp and measure voltage fluctuations between the nodes as the net effect of millions of neurons in the brain. EEGs are used for diagnosis of a number of neurological disorders, including epilepsy, sleep disorders, comas and more.

Our goal task is to train a single neural network to classify an epoch of EEG data from any patient as being seizure or non-seizure. Previous attempts at EEG classification have been limited to

1. Building separate models for each patient.
2. Using highly cleaned benchmark data sets of only certain seizures and non-seizure activities.

We hope to overcome these shortcomings by using vastly more data than previous attempts. For example, Shoeb (2009) used 23 patients and 844 EEG files, whereas we have access to 12,385 patients and 136,363 EEG files.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

Computational detection of seizures from EEG has been attempted a number of times over the last 15 years. However, successes have been limited to being patient-specific, as in Shoeb (2009) or being on a very small and well-cleaned data set, as in Thodoroff et al. (2016); Fürbass et al. (2015); Hosseini et al. (2016). The main challenge is that **seizures present dramatically different EEG data for the various types of seizures, for different patients, and even for a given patient**. This is discussed in nearly every paper on the topic:

- "This is a challenging problem because seizure manifestations on EEG are extremely variable both inter- and intra-patient...Traditional methods typically do not perform very well on new patients, due to their low capacity to generalize well across different seizure patterns." Thodoroff et al. (2016)
- "Main challenge is finding reresentations that are invariant to inter- and intra-subject differences, as well as to inherent noise associated with EEG data collection." Bashivan et al. (2015)
- "The first challenge is due to the fact that EEG is not a stationary signal. Therefore, normal and seizure patterns may vary across different patients." Hosseini et al. (2016)

We are uniquely positioned to overcome this challenge because we have access to a data set that is **significantly larger** than anything used in other research. See Dataset discussion below for more details. Below is an in-depth review of previous model performance.

To begin, Tzallas et al. (2012) provides a good literature overview of the different computational techniques used over the last 15 years to identify seizures in EEG. These range from hand-engineered heuristics to various learning algorithms.

The most noteworthy achievement before the application of neural networks was the development of Filter Bank Common Spatial Pattern (FBCSP) Ang et al. (2008). These perform autonomous selection of key temporal-spatial discriminative EEG characteristics by passing EEG through frequency filters, a feature selection algorithm, and a classification algorithm.

In his PhD thesis, Shoeb (2009) had the first breakthrough for machine learning to detect seizures by building a patient-specific classifier. His model was trained on 2 or more seizures and tested on 844 hours of continuous scalp EEG from 23 pediatric epilepsy patients and detected 96% of 163 test seizures with a median 2 false detections per 24 hour period. Furthermore, the data set developed in this work was open sourced as the Children's Hospital of Boston-Massachusetts Institute of Technology dataset (CHB-MIT) and became used as a standard training data set. The CHB-MIT dataset contains 23 patients and 24 recordings (one patient had two reads) for a total 969 Hours of scalp EEG recordings with 173 seizures. The data set includes male and female patients aged 10-22 with a diversity of seizures (clonic, atonic, tonic).

Thodoroff et al. (2016) used this data set but with more complex Fourier video representations as described by Bashivan et al. (2015) and with a more complex convolutional-recurrent network. In detecting whether a 30 second segment of signal contains a seizure or not, their model outperformed Shoeb with 95-100% accuracy on patient-specific data. When testing on a cross-patient basis, the model achieved 85% accuracy.

Fürbass et al. (2015) used a slightly larger data set with 205 patients from 3 different epilepsy centers but only achieved an average sensitivity of 67% on the CHB-MIT.

Finally, Hosseini et al. (2016) achieved success but on a small clinical dataset of two epileptic patients (with 60 interictal and 60 preictal segments) that was developed by the University of Pennsylvania and the Mayo Clinic. Using a dimension reduction and a stacked autoencoder, they achieved accuracy 94%.

EEG data is used for a variety of clinical purposes, so researchers actively develop computational methods to identify a various activity. As such, representational and architectural advances for EEG data have often come from non-seizure related tasks. For example, when developing a model for a cognitive load task, Bashivan et al. (2015) proposes a video-like representation of EEGs whereby the electrodes are projected onto a 2-D plane and a Fourier transform of intensity in three hand-engineered buckets creates three channels to generate an EEG video that can then use any neural network from video analysis. Similarly, when developing a model for identifying movement-related information, Schirrmeister et al. (2017) developed a 1-dimensional convolutional neural network that inspired the convolutional neural network used in this paper.

3 Dataset

Put briefly, our data set is **huge and heterogeneous**. The data set consists of 136,363 electroencephalograms: 99,721 from adults measured at the Stanford Hospital and 36,642 from children measured at the Lucile Packard Children's Hospital. The data is remarkably heterogeneous:

- Type: intracranial and scalp
- Reason: routine, long-term EEG, ambulatory study, diagnosing epilepsy, diagnosing some other disorder, etc.
- Length: 5 seconds to 24 hours
- Frequency: 200 Hz to 500 Hz
- Age: 0 to 100+
- Channels: 3 to 142

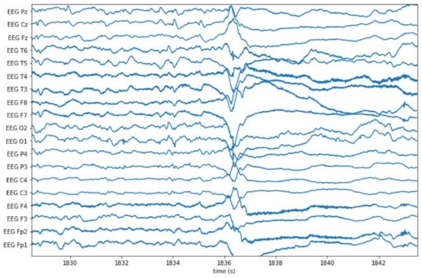

Figure 1: An example EEG in which a seizure begins around t=1836 seconds

Each electroencephalogram is stored in a hierarchical data format (HDF) containing anonymized data about the patient, metadata about the EEG read, the raw signals of the EEG read (a matrix with shape number of channels by length of EEG), and accompanying annotations with timestamps for the EEG. Because the data is not explicitly labeled as seizure and non-seizure, we use the accompanying annotations as a proxy for seizure labels. Unfortunately, the annotations rarely list the seizure end, so we only know the start of seizures.

It is worth re-emphasizing that unlike other data sets used in previous research, this data set contains an **immense variety of seizure and non-seizure data**. It contains every type of major seizure occurring in many different ways, and it contains non-seizure activity not included in any of the highly cleaned benchmark data sets such as eating, tooth-brushing, and exercising. This wide variety of activity improves the clinical usability of our final model because our results generalize quite well to all gathered EEG data.

4 Pre-Processing

Given the heterogeneity of the data, we filtered our data to be self-consistent by restricting our scope:

- Only including scalp EEG (i.e., excluding intracranial reads).
- Taking reads only from the 25 nodes in the International 10-20 System.
- Limiting length to minimum time of seizure: 10 seconds.
- Only including files with a sampling rate of 200 Hz.
- Standardizing each waveform to have mean 0 and standard deviation 1.

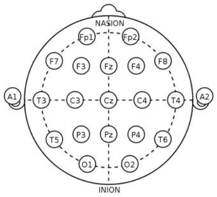

Figure 2: The nodes used in the International 10-20 System

With these filters in place, we defined our seizure and non-seizure data:

- **Seizures**: Every time a file contains a seizure annotation, slice the following 10 seconds. There are 11,641 unique files with 25,850 labeled seizures fitting the above filtering.

- **Non-seizures**: Randomly sample a 10-second slice from all files without a seizure annotation to get 25,850 non-seizures occurrences. This ensures that we train with a training balance of an equal number of seizure and non-seizure data points.

This ultimately gave us 51,700 matrices shaped 25 channels by 2000 = 200Hz x 10 seconds with a binary label, 1 for seizure and 0 for non-seizure. We split this data into 90% training, 5% development, and 5% test sets.

5 Proposed Architectures

We tested four different model architectures:

- **Baseline**: Flatten the matrix into a 50,000 dimensional vector and run logistic regression via a one-layer neural network with no hidden layers.
- **Dense Network**: Flatten the matrix into a 50,000 dimensional vector and run it through two hidden layers with size 1000 and 30 each with ReLU activation functions before outputting a single neuron with a sigmoid activation function.
- **Convolutional Network**: Run the 2000x25-dimensional matrix through five 1-dimensional convolutional layers each with 10x1 filters, 3x1 max-pooling, and exponential linear units. Flatten the remaining matrix and run it through one layer with batch norm, dropout, and ReLU before a final layer with a sigmoid activation function. This model is based off the work done by Schirrmeister et al. (2017). In each layer with input size (N, C_{in}, L_{in}) and output size (N, C_{out}, L_{out}), we apply:

  $$out(N_i, C_{out_j}) = bias(C_{out_j}) + \sum_{k=0}^{C_{in}-1} weight(C_{out_j}, k) \star input(N_i, k)$$

  where $\star$ is the cross-correlation operator, N is batch size, C is channel number, and L is sequence length.
- **Recurrent Network**: Process each of the 2000 inputs into two bi-directional LSTM units followed by mean pooling, a dense layer with batch norm and ReLU, and a final layer with a sigmoid activation. This is based on many-to-one recurrent networks on temporal data. At each time step t, each of the two hidden units calculates:

  $$\begin{gathered} i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\ f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\ g_t = tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\ o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\ c_t = f_t c_{t-1} + i_t g_t \\ h_t = o_t tanh(c_t) \end{gathered}$$

  where h_t, c_t, x_t, are the hidden state, cell, and previous layer's hidden state at time t. Also, i_t, f_t, g_t, o_t are the input, forget, cell, and out gates, respectively, at time t.

6 Results

As expected, the worst performing architecture was the logistic regression model. With just 50,000 parameters, it couldn't fit much beyond guessing.

The densely connected feed-forward neural network did significantly better than expected, achieving a test set accuracy and F1 score of 90.8% and 0.897, respectively. It's interesting that it could do so well without being designed to take advantage of the spatial or temporal features of an EEG. The architecture's flexibility likely was its key to success as it had over 1 billion parameters to tune.

The convolutional neural network was the top performing architecture. It achieved the highest test accuracy and F1 score at 95.2% and 0.943, respectively. This is as expected because this network

Table 1: The training and test accuracy and F1 score for the four model architectures described above. These results were found after iterating on exact model architectures and hyperparameters used. The metrics shown are at apparent convergence after 50 epochs of training.

Model	Data	Accuracy	F1 Score
Logistic	Train	56.5%	0.535
Logistic	Test	54.6%	0.533
Dense	Train	90.8%	0.921
Dense	Test	90.2%	.897
Convolutional	Train	96.5%	0.947
Convolutional	Test	95.2%	0.943
Recurrent	Train	60.1%	0.621
Recurrent	Test	58.3%	0.592

architecture was explicitly designed to take advantage of the temporal structure of an EEG with the added flexibility of a final densely-connected network.

The bi-directional LSTM network performed okay with an accuracy around 60%. Its lackluster performance is likely due to the known challenge of training an LSTM with 2000 time steps, larger LSTMs usually work with. We expect this network could perform significantly better if we used convolutional filters or down-sampling to reduce the number of input time steps before feeding it into an LSTM.

7 Discussion and Next Steps

Our top-performing architectures were the densely-connected and the convolutional networks. Both achieved accuracies and F1 scores over 90%, well-beyond the results of published work for a general EEG classifier used on any patient to classify any of the types of seizures. This is especially remarkable given the heterogeneity of the data, which contains ample non-seizure activity, many types of seizures, and many patients.

That said, we plan to pursue each of the following three steps and expect them improve our modeling capabilities:

1. **Improve Labels**: The current seizure annotations are not trustworthy and, more importantly, often missing. We can apply labeling functions to generate sound training data via weak supervision as described in Ratner et al. (2017).
2. **Improve Embeddings**: We currently use the raw waveform. However, Bashivan et al. (2015) proposes representing an EEG as a video by (1) projecting the nodes into two-dimensions, (2) taking the Fourier intensity of three discretized buckets over a t-second interval, (3) combining these three channels to create an image, and (4) combining images temporally to make a video. This captures temporal and spatial data in EEGs and will allow us to use standard video neural net architectures.
3. **Improve Network**: Our proposed models have been relatively shallow compared to what Dai et al. (2017) suggest is necessary to capture all information from raw waveforms.

Given its success on such a general data set, this work is already clinically useful as a tool to label EEGs. That said, there are two additional tasks we hope to tackle using deep learning on this data:

1. **Segmentation**: We would love to build a model that could segment an entire file into seizure and non-seizure activity. Unfortunately, the annotations in our EEG files usually only mark seizure start, so we plan to tackle this either using weak supervision as described above to generate labels for the training set or by running our current network and using time-dependencies to create continuity.
2. **Prediction**: An ultimate ambitious goal of the project is to forecast a seizure *t*-seconds before it occurs. Although we already have the infrastructure to test this, we do not expect great results as even the best clinicians cannot forecast seizures from EEG data with any improvement over guessing.

8 Contributions

I was the only student working on the project. I wrote the code that pre-processed the data and implemented and iterated on the deep learning architectures presented in the paper. All of the code used to build this model can be found at https://github.com/nhershey/cs230eeg. I also wrote the poster and final report.

That said, this project could not have been happened or been successful without the help of some collaborators and advisers:

- **Dr. Chris Lee-Messer, MD, PhD**: Dr. Lee-Messer is a pediatric neurologist at the Lucile-Packard Children's Hospital and a Clinical Assistant Professor of Neurology & Neurological Sciences at Stanford University. He provided access to the dataset used in this project and served as the expert on all things related to electroencephalograms, epilepsy, and seizures. He was instrumental in working with this large data set and determining what type of featurization made sense.
- **Jared Dunnmon, PhD**: Jared Dunnmon is a postdoctoral research fellow in the InfoLab at Stanford University focusing on weakly supervised machine learning. He helped me get set up on Stanford computing resources and provided advice on designing deep learning networks.
- **Paroma Varma, PhD**: Paroma is a third year Ph.D. student advised by Prof. Chris Ré and affiliated with the DAWN and Infolab groups focused on making machine learning easily usable for domain experts who do not have access to the massive datasets required for training complex models. She acted as a research mentor to me on this project.
- **Professor Christopher Ré, PhD**: Christopher (Chris) Ré is an associate professor in the Department of Computer Science at Stanford University in the InfoLab who is affiliated with the Statistical Machine Learning Group, Pervasive Parallelism Lab, and Stanford AI Lab. His work's goal is to enable users and developers to build applications that more deeply understand and exploit data. His contributions span database theory, database systems, and machine learning. He served as an adviser to this project.

References

Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. (2008). Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 2390–2397. IEEE.

Bashivan, P., Rish, I., Yeasin, M., and Codella, N. (2015). Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*.

Dai, W., Dai, C., Qu, S., Li, J., and Das, S. (2017). Very deep convolutional neural networks for raw waveforms. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 421–425. IEEE.

Fürbass, F., Ossenblok, P., Hartmann, M., Perko, H., Skupch, A., Lindinger, G., Elezi, L., Pataraia, E., Colon, A., Baumgartner, C., et al. (2015). Prospective multi-center study of an automatic online seizure detection system for epilepsy monitoring units. *Clinical Neurophysiology*, 126(6):1124–1131.

Hosseini, M.-P., Soltanian-Zadeh, H., Elisevich, K., and Pompili, D. (2016). Cloud-based deep learning of big eeg data for epileptic seizure prediction. In *Signal and Information Processing (GlobalSIP), 2016 IEEE Global Conference on*, pages 1151–1155. IEEE.

Megiddo, I., Colson, A., Chisholm, D., Dua, T., Nandi, A., and Laxminarayan, R. (2016). Health and economic benefits of public financing of epilepsy treatment in india: An agent-based simulation model. *Epilepsia*, 57(3):464–474.

Ratner, A. J., Bach, S. H., Ehrenberg, H. R., and Ré, C. (2017). Snorkel: Fast training set generation for information extraction. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1683–1686. ACM.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., and Ball, T. (2017). Deep learning with convolutional neural networks for brain mapping and decoding of movement-related information from the human eeg. *arXiv preprint arXiv:1703.05051*.

Shoeb, A. H. (2009). *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology.

Thodoroff, P., Pineau, J., and Lim, A. (2016). Learning robust features using deep learning for automatic seizure detection. In *Machine Learning for Healthcare Conference*, pages 178–190.

Tzallas, A. T., Tsipouras, M. G., Tsalikakis, D. G., Karvounis, E. C., Astrakas, L., Konitsiotis, S., and Tzaphlidou, M. (2012). Automated epileptic seizure detection methods: a review study. In *Epilepsy-histological, electroencephalographic and psychological aspects*. InTech.