Interpretability of Deep Learning Models for Classification of Epilepsy in EEG Recordings

Daniel Tang
Department of Computer Science
Stanford University
Palo Alto, CA 94305
danieldt@stanford.edu

Abstract

The application of deep-learning tools in clinical neuroscience has the potential to vastly increase the effectiveness and efficacy of current diagnostics. In order for these tools to be effective, they must be transparent, such that clinical experts can interpret the model's decisions based on key features identified by the model. Recent work has shown state-of-the-art performance in multi-class EEG detection of epilepsy, but relatively little work in the field has been done to analyze the interpretability of such models. Here, I implement a modified CNN that is able to achieve similar performance to state-of-the-art methods and use DeepLift, and GradCAM to allow users to interpret the decisions made by the model.

1 Problem Statement

With the increasing use of deep learning in basic research and clinical neuroscience, there is a growing need for interpretability in the application of these tools. [1] Given the inherent complexity and incomplete understanding of the nervous system, it is paramount that applied deep learning models in the neuroscience space be explainable, such that scientists and clinicians can understand model outcomes in the context of known biological mechanisms. Providing this interpretability will promote adoption of such tools in the clinic, identify potential mechanisms for further research, and allow for accountability under unexpected model results.

Here, we explore model interpretability in deep learning models of epilepsy detection. Epilepsy is generally monitored for in clinical contexts using EEG-recording, where EEG detection of epilepsy is used to inform further treatment, pre-surgical evaluation, and temporary lifestyle adjustments. Prior research has examined the use of a variety of pretrained and fine-tuned CNNs to classify EEG images according to the specific class of seizure (simple partial, complex partial, focal non-specific, generalized non-specific, absence, tonic, tonic-clonic, and non-seizure), with top performing models achieving 80-90% accuracy. [2] Given the visual nature of this approach and the relatively high performance of such models with respect to conventional feature and cluster based approaches, I chose to move forward with CNNs as the architecture of choice for this task. Recently, Raghu et al demonstrated state-of-the-art classification accuracy of 82 % on a 9-class classification task using a fine-trained GoogleNet model and 88.3% using feature extracted classification from the fine-tuned Inceptionv3 model. [2] The goal of this project will be to improve the interpretability of multi-class EEG classifiers by examining salient features extracted from DeepLift, and GradCAM. [3]

2 Dataset

The Temple University Hospital Epilepsy dataset consists of 32-channel EEG recordings labeled with time-stamps for seizure events and non-seizure artifacts. For each class of seizure, there are between

Seizure type	Description	Number of patients	Seizure events	Duration (h)
SP[a]	Partial seizures during consciousness; Type specified by clinical signs only	18	63	1.2
CP[a]	Partial seizures during unconsciousness; Type specified by clinical signs only	31	210	8.2
FN[b]	Focal seizures which cannot be specified by its type	63	542	22.56
GN[b]	Generalized seizures which cannot be further classified into one of the groups below	75	231	12.5
AB[a]	Absence discharges observed on EEG; the patient loses consciousness for a few seconds (Petit Mal)	21	63	1.3
TN[a]	Stiffening of the body during a seizure (EEG effects disappears)	19	29	1.3
TC[a]	At first stiffening and then jerking of the body (Grand Mal)	15	25	0.8
NS[a]	EEG without any type of seizure events	110	540[c]	35.6

Figure 1: Temple EEG Seizure Corpus: The corpus contains 200 GB of EEG recording and metadata files. There are 9 EEG categories within the data with each category containing over 0.8 hours of data each.

15-100 different patients with a total recording duration of 0.8-23 hrs of recording for each seizure class (Figure 1). The data corpus is organized by training vs test data, the type of electrode reference scheme used, patient ids, and recording session id.

3 Data Preprocessing and Loading

The data were stored as .edf files containing the raw EEG recording. Physician labels for events during the the EEG recording was stored as text files. Once imported, the raw EEG data were passed through a 0.1-44 Hz bandpass filter to reduce noise and then binned into 10s windows. A spectrogram was then produced for each EEG channel in each 10s window by calculating the short time fourier transform, using a Kaiser window of length 63 and shape parameter of 1, a 75 % window overlap, and an FFT length of 256. The spectrogram from each channel was then plotted into the same image, creating a merged spectrogram of all the channels for every 10s window. The labels were originally stored as date-time arrays indicating the start and stop time of episodes (baseline, seizure, noise) in seconds. These labels were converted to one-hot vectors for each 10s window. Each vector had a length of 10, representing non-seizure and 9 categorized seizure classes.

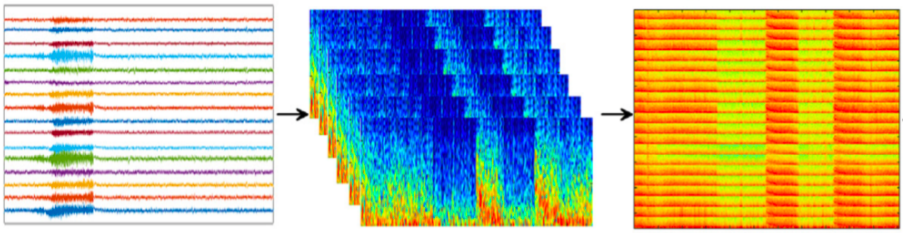


Figure 2: Data Processing Pipeline: (Left) Raw EEG data were pre-processed via a 0.1-44 hz bandpass filter prior to being binned into 10s windows. (Middle) Spectrograms for each channel were created using the STFT. (Right) The spectrograms for each channel were plotted together to form a merged spectrogram, which were the inputs into the CNN models.

Due to storage limitations, I was only able to download an older version of the corpus with less files. Upon processing the data, there was a lack of general seizure, absz, tnsz, and cnsz examples (n = 0) in the dataset such that I removed these classes from the model all together. I also identified severely imbalanced data, with a significantly larger proportion of baseline data compared to data from seizures (>95% of data were baseline). In order to account for the unbalanced data, I randomly under-sampled the baseline examples so that the number of baseline cases were within an order of magnitude to that of the other seizure types and used a weighted cost function. Once the data were balanced, the remaining classes each had between 100-2000 examples for training and testing.

4 Model

Raghu et al demonstrated state-of-the-art performance (82% accuracy) on a multi-class EEG classification using transfer learning alone (GoogleNet) and transfer-learning with feature extraction (InceptionV3). Here I will focus on the GoogleNet framework due to concern over the feature-selection scheme in the proposed InceptionV3 model.

I tested two different model architectures. The first architecture consisting of a modified GoogleNet network with a different output layer size. For the second architecture I changed the output of GoogleNet to a two-layer neural network with the first layer containing 512 neurons and the second layer containing the same number of neurons as output classes. The final output of both architectures was passed through a softmax function.

Training was performed using cross-entropy loss, Adam's optimizer, a learning rate of 1e-4, and over 25 epochs. The data were split into 90% train, 10% test and each class was weighted by the inverse relative class count.

5 Interpretability

]DeepLift, and GradCAM frameworks were implemented using the Captum package for pytorch and used to visualize feature importance within the input spectrograms leading to prediction.

6 Analysis and Evaluation Metrics

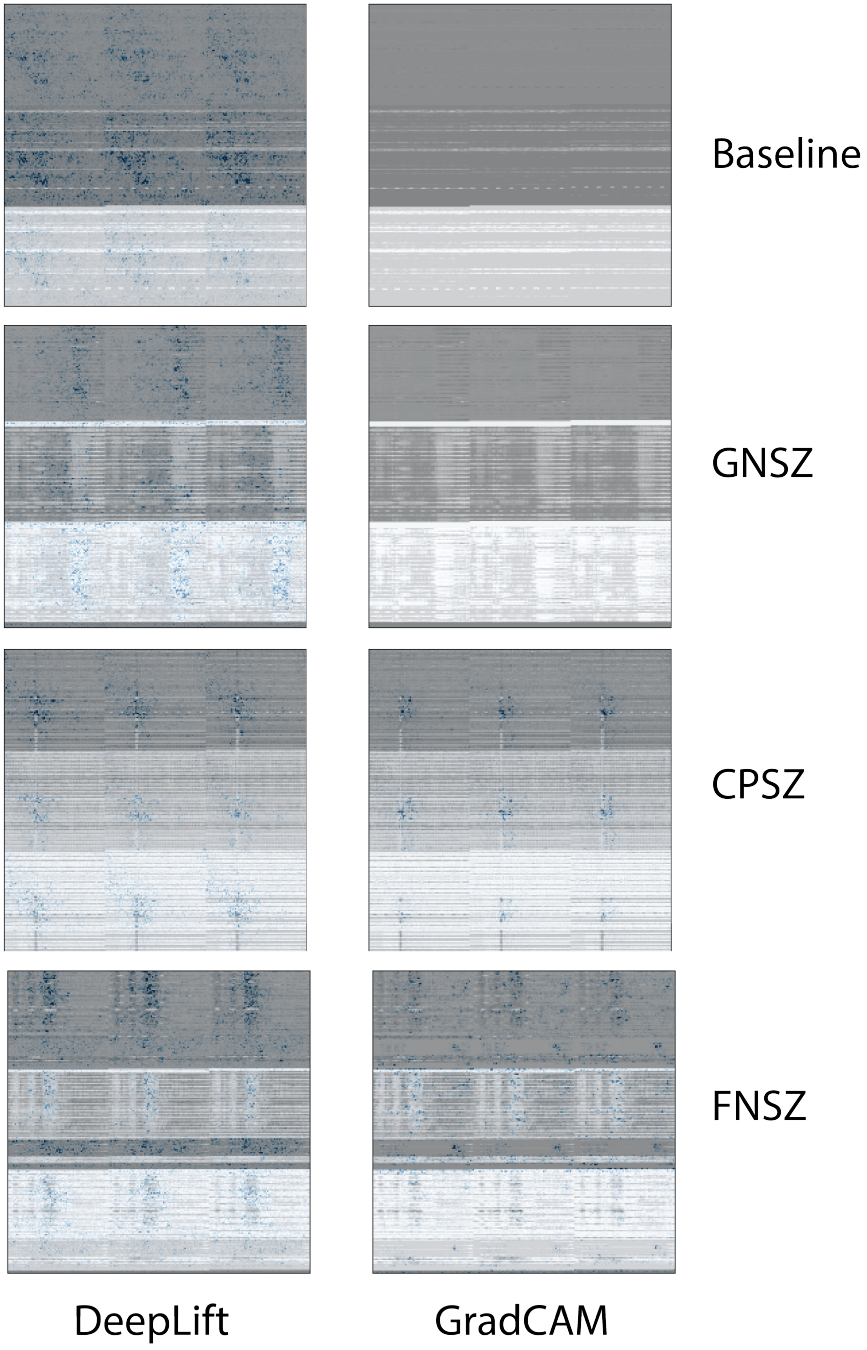
Model performance will be evaluated using accuracy and compared to both the Raghu et al model and a random baseline. Interpretability methods will produce salience maps of input spectrogram images to identify features of EEG that are predictive of epilepsy. The identified salient features will be compared against typical clinical guidelines for epilepsy detection outlined in the Temple Epilepsy Annotation Guidelines. [4] Analysis of the saliency maps will be conducted to examine differences in predictive features across different classes of epilepsy. Similarity of identified features will also be analyzed across DeepLift, and GradCAM by comparing the amount and degree of overlap between identified inputs.

Ideally, evaluation of the interpretability of the model would be quantified by a neurologist or EEG technician. We would provide the user with an explanation from the model and have the user assign

a seizure class. We would then compare the users' prediction accuracy as a measure of quality or degree of explainable information provided be the interpretable model.

7 Preliminary Results

Using the first architecture described by Raghu et al, I was able to achieve 76% accuracy. Using the 2-layer output architecture, performance was increased to 77%. When organizing accuracy by class, I observed a baseline accuracy of 69 %, fnsz accuracy of 85 %, gnsz accuracy of 56 %, spsz accuracy of 68 %, cpsz accuracy of 79%, and atsz accuracy of 72%.



Between DeepLift and GradCAM, DeepLift appeared to work more consistently in identifying salient features in the spectrogram. The features identified by DeepLift and GradCAM appear to be consistent for the CPSZ and FNSZ seizure examples. Furthermore, for each seizure type, DeepLift and GradCAM are able to extract differing important features, indicating that the model is in fact learning functional differences in the EEG spectrograms.

8 Remaining/Future Works

There's still a significant amount of analyses that could be done to continue this project. For one, incorporating the latest version of the Temple EEG corpus and identifying additional data sources would help improve model accuracy, particularly for under-represented classes in the dataset. I could also augment the data by training/testing the model on shorter spectrogram windows. More rigorous analysis of the performance for each class could also be performed, identifying failure mechanisms for poorer-performing classes.

One could also explain feature saliency or interpretability using LIME or counterfactual examples. In order to test how relevant these spectrogram surveys are, one can imagine either working with clinical experts with apriori knowledge of epilepsy to evaluate feature relevance or testing with a group of non-experts in a human-driven classification task, where participants are asked to match the saliency map to the correct class.

Interpretability methods will help to improve the adoptability of deep-learning tools in the clinic. However, there will need to be sufficient effort spent towards training and gaining feedback from clinical experts as these tools are being developed and validated for use in hospitals.

9 References

[1] Fellous J, Sapiro G, Rossi A, Mayberg H, Ferrante M, "Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation", Front. Neuroscience, 13 Dec 2019

[2] Raghu S, Sriraam N, Yemel Y, Rao S, Kubben P, "EEG based multi-class seizure type classification using convolutional neural network and transfer learning", Neural Networks, Vol 124, April 2020

[3] Ribeiro M, Singh S, Guestrin C, "Why Should I Trust You? Explaining the Predictions of Any Classifier", arXiv ,9 Aug 2016

[4] Ochal D, Rahman S, Ferrell S, Elseify T, Obeid I, Picone J, "The Temple University Hospital EEG Corpus: Annotation Guidelines ", Temple University Hospital ,15 April 2020