Improving understanding of EEG measurements using transparent machine learning models

Roadknight, C., Zong, G., Rattadilok, P.

University of Nottingham
UK | CHINA | MALAYSIA

University of Nottingham Ningbo China, 199 Taikang East Road, Ningbo, 315100, Zhejiang, China.

First published 2019

University of Nottingham
UK | CHINA | MALAYSIA

Improving understanding of EEG measurements using transparent machine learning models

Chris Roadknight, Guanyu Zong and Prapa Rattadilok

University of Nottingham Ningbo Campus
[chris.roadknight; zy19656; prapa.rattadilok]@nottingham.edu.cn

Abstract. Physiological datasets such as Electroencephalography (EEG) data offer an insight into some of the less well understood aspects of human physiology. This paper investigates simple methods to develop models of high level behavior from low level electrode readings. These methods include using neuron activity based pruning and large time slices of the data. Both approaches lead to solutions whose performance and transparency are superior to existing methods

Keywords: Deep Learning, Physiological data, CAPing.

1 Introduction

1.1 Physiological data

With the rise in popularity of smart devices such as smart watches, fitness trackers and EEG monitoring devices, high quality, high frequency physiological data is easier than ever to collect. With the promise of improving health, fitness and/or sleep this trend shows no sign of abating. They capture large amounts of high frequency data that largely goes unused. With the rise of availability of all forms of personal healthcare related data comes the requirement to do something with it. The desired output includes an understanding of the current and future state of the individual recording the data. This physiological data can come in the form of static values such as weight, height, gender and data that has more short term time dependence such as temperature, sweat levels and respiratory rate. There are also a range of measurements that vary rapidly in short time frames, these include electrical activity in the brain or heart.

The EEG signal is a voltage that can be measured on the surface of the head, this signal is related to coordinated neural activity. This is particularly powerful when groups of neurons fire at the same time. Neural activity varies depending on the patient's mental state and the EEG signal can detect such variation with a degree of noise.

1.2 Machine Learning

Machine learning takes many forms and means different things to different people. For the purposes of this research we will deem it any process where we use computational algorithms to learn relationships between causal data and resulting effects. By modelling these cause-effect relationships we can make predictions about labels associated

with a range of causal data. This is also deemed ‘supervised learning’ as we supervise the training of a model until we can no longer improve its accuracy. This trained model is then evaluated on data that was not used to train it, giving us an idea on how well the model generalized. Many machine learning techniques exist, this research uses some well-known methods such as Gradient Boosting and Deep Learning.

1.3 Deep Learning

Deep learning is an extension of classic single hidden layer neural networks. One form of deep learning is based on a multi-layer feedforward artificial neural network that uses backpropagation for stochastic gradient descent training. The network can contain a large number of hidden layers consisting of neurons. Adaptive learning rates, rate annealing, momentum training and regularization enable high prediction accuracy in many complex prediction scenarios. This kind of feed-forward ANN model is the most common type of deep neural network and will be used in this research

Within a practical problem-solving context using conventional machine learning techniques, researchers have discovered that although the existing models could perform well on synthetic or well-structured datasets, when working with raw natural data, like processing an image pixel-wise in a pattern recognition task, models often dont reach their optimality unless an extreme amount of effort is spent to convert the raw data into a suitable form of internal representation for the network to process [14]. Besides this, the challenges brought by natural datasets also include the curse of dimensionality, examples of physiological or medical datasets which contain large amount of attributes [15]. When the number of attributes of a dataset rises, the dimensions of data space also rise above conventional intuition and have unfamiliar properties. For networks which are built by human researchers this could be an extra barrier for understanding and analyzing the essence behind data [9]

The deep learning methodology aims to solve these two problems by both providing a simpler way to convert high-dimensional, raw data into feature vectors or other internal representations that has lower data dimensions, and, under some conditions, boosting the computational power of the model itself by allowing more computing units in the process.

Deep learning models can be difficult to interpret, due to their non-linearity. It is important to make these models as simple as possible while still retaining performance so as to be more human readable, this is sometimes called Mimic Learning [5]. The basic premise is to first ensure you have the simplest version of your model and then convert the underpinning complex mathematical processes into more understandable forms [7,8] The approach here is based on a method used for single hidden layer neural networks [9] but modifications have been made to enact the process on a deep neural network

2 Experimental Design

2.1 Dataset

EEG data has been widely used in machine learning-based classification problems. This research uses a publically available dataset [4]. In this data 10 college students were asked to wear a wireless single channel 'MindSet' EEG device [10] that measured activity over the frontal lobe. They were then asked to watch ten 2-minute long videos that ranged in complexity and then decide if they were confusing or not. This dataset contains 11 EEG based metrics:

a. The raw EEG measurement itself (Raw)
b. 8 frequency based transformation (Delta, Theta, Alpha 1, Alpha 2, Beta 1, Beta 2, Gamma 1, Gamma 2)
c. 2 proprietary functions (Attention, Mediation)

It also contains a label that is a subjectively assigned decision as to whether the video was confusing or not. This was largely ignored in this research. The data is collected for 2 minutes but only the middle minute is deemed usable. The data is binned at one sample every 0.5 seconds Even though the EEG is measured at a 512 Hz. (Table 1)

Table 1. Features generated from EEG data.

Features	Description	Sampling rate	Statistic
Attention	Proprietary measure of mental focus	1 Hz	Mean
Meditation	Proprietary measure of calmness	1 Hz	Mean
Raw	Raw EEG signal	512 Hz	Mean
Delta	1-3 Hz of power spectrum	8 Hz	Mean
Theta	4-7 Hz of power spectrum	8 Hz	Mean
Alpha1	Lower 8-11 Hz of power spectrum	8 Hz	Mean
Alpha 2	Higher 8-11 Hz of power spectrum	8 Hz	Mean
Beta1	Lower 12-29 Hz of power spectrum	8 Hz	Mean
Beta 2	Higher 12-29 Hz of power spectrum	8 Hz	Mean
Gamma1	Lower 30-100 Hz of power spectrum	8 Hz	Mean
Gamma2	Higher 30-100 Hz of power spectrum	8 Hz	Mean

There are 8 attributes based on different power spectrums, 2 proprietary measurements and a 2 Hz mean of the 512 Hz raw sample. Preprocessing this kind of data is always a key step in gaining an understanding of cause-effect relationships. This is usually carried out using statistical or time series approaches to smooth out unwarranted variation and noise. This step is hugely important as models need to be developed for underlying processes and not the overlaying noise.

In addition to the transformations of the raw EEG data detailed in the previous paragraph, we log transformed some of the attributes based on their distribution.

2.2 Machine Learning Approaches

The current 'best' performance of this dataset is yielded using 2 different types of Long Short Term Memory (LSTM) Recurrent Neural Networks [2, 16]. These approaches build on the back propagation algorithm that passes through layers, each layer summating to a transfer function. During the training process, error sent backward through the network can be amplified, which may lead to instability, oscillating weights, or vanishing gradients. Exploding gradients can be mitigated via truncation with correct transfer functions. Vanishing gradients are addressed using the Long Short-Term Memory RNN (LSTM) approach, introducing memory units to RNNs. The memory units help stop the error signal vanishing so that it is large enough to be back.

We compare the current leading methods discussed in the previous paragraph with 2 more sets of approaches: Flat Time Segmentation and Batch Processing. The 'Flat Time Segment' approach takes raw 0.5 second time slices of data and uses them as time independent training data. This method has no knowledge of whether the 0.5 seconds of data is the first, second or last time segment. This removes possibly useful temporal information but makes the resulting models somewhat simpler to understand.

The Batch processed approach takes a larger time segment and performs statistical analysis on the data producing static values such as mean, median and standard deviation of the metric over a 1 minute time period. This is a simpler way of retaining memory than the LSTM approach

The machine learning algorithms are all publically available approaches, largely used 'out-of-the-box' with little or no tuning. This means subsequent researchers should easily be able to reproduce our results. We use the H2o [6] and R [17] platforms. In H20 we use the Random Forest, Gradient Boosting, Naïve Bayes and Deep Learning algorithms and within R we use the e1071[18] package for SVM and the CARET [19] package for Classification Trees

2.3 Correlated Activity Pruning

There has been some success in optimizing deep and shallow neural networks using a pruning approach, whereby less useful links and nodes are removed. The most popular is the pruning method of Han et al. [11] which works by first training a network, setting all the weights to zero based on a fixed threshold, are then fine-tuning the remaining connections. An alternative method reduces values of trained weights by applying vector quantization techniques [12]. A distillation approach can be used to train a separate smaller network that imitates the original one [13]. Correlated activities can be used to condense networks [8, 9,10]. A 'brain damage' approach can be used that makes use of second-order information on the gradient of the cost function to remove connections [14]. Simpler approaches include using limited numerical precision [15] (eg. single bit per weight [16]) and lossy hash functions to force weight sharing [17].

The correlated activity approach has been applied to deep networks in a whole node merging manner [20, 21] but this approach fails to simplify for correlations in activity profiles for single connections between layers. For this work we applied a piecewise approach to the process. After the initial training and testing of the deep learning network, connections between all the nodes in layer n-1 to layer n were examined and a

Pearson correlation coefficient taken, the connections with the highest correlation coefficient were merged, the network was retrained and as long as the resulting accuracy was not significantly affected this was repeated (Figure 1)

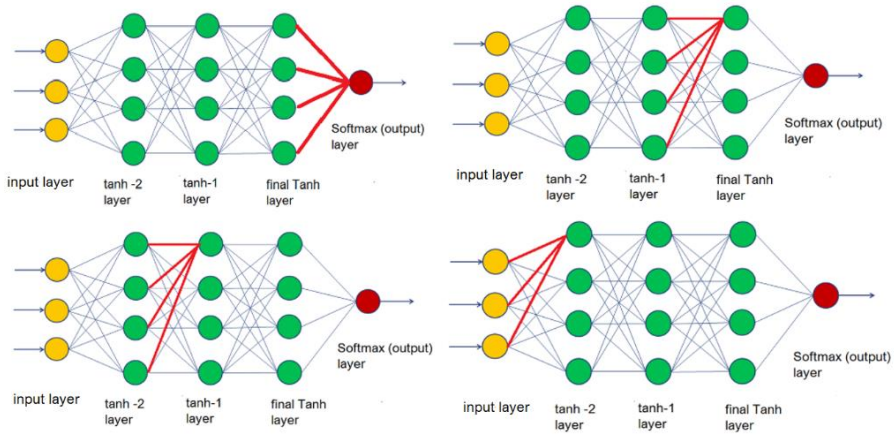

Fig. 1. Schematic of Correlated Activity Pruning process

3 Results

The results presented here will show how well different groups of methods compare to the existing 'leading' approaches. Accuracy is a suitable measure here as there is a balanced target ratio. Qualitatively we have assessed how difficult the machine learning approach is to apply and subsequently understand, this makes up our 'Complexity' metric. This is used for information only and it allows the reader an insight into our opinion on the difficulty (application and interpretation) each method brings. Figure 2 and Table 2 shows the performance of different methods. It can be seen that the three groups of methods discussed previously group together into three clear groups. The flat time segment approaches has a range of lower complexity methods but accuracy is inferior to the published 'best'. The memory enabled LSTM have relatively high complexity and better accuracy. The batch processed approach has the highest range of accuracy but also 5 methods that clearly outperform the LSTM approaches.

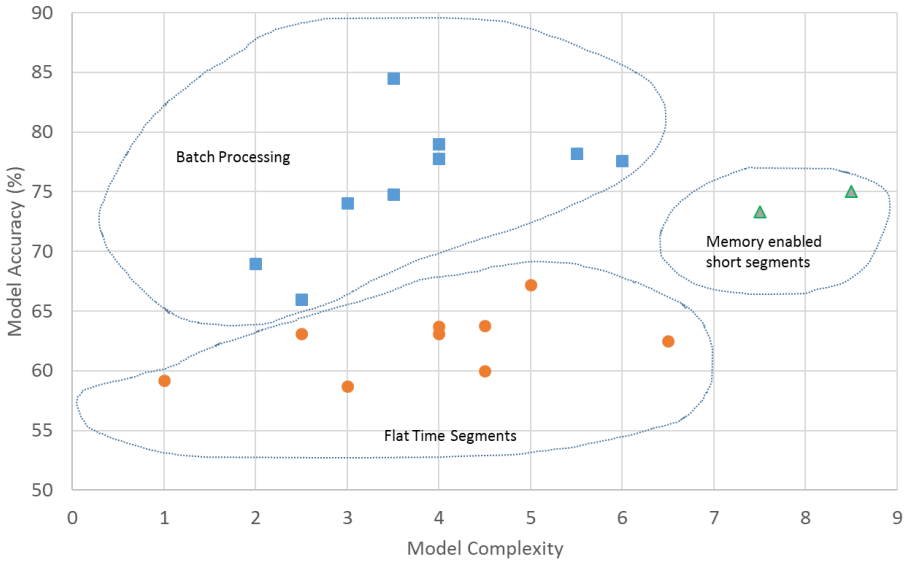

Fig. 2. Performance of the 3 different groups of approaches.

Table 2. Qualitative and quantitative performance of models

Model	Accuracy (%)	Complexity (model)	Complexity (process)	Complexity (Combined)
LSTM approaches				
CF-Bi-LSTM (Ni et al 2017)	73.3	7	8	7.5
Bi-LSTM (Wang et al 2018)	75	8	9	8.5
Flat Time Segmented				
SVM (LibSVM)	67.2	6	4	5
Gradient Boosting(H2o)	63.7	5	3	4
MLP (neuralnet	63.1	3	5	4
Naïve bayes)H2o)	58.7	5	1	3
SVM (e1071)	60	5	4	4.5
Random Forest(H2o)	63.8	4	5	4.5
Classification Tree (CARET)	63.1	2	3	2.5
Logistic Regression (R)	59.2	1	1	1
Deep Learning (H2o)	62.5	6	7	6.5
Batch Processed				
Gradient Boosting(H2o)	84.5	4	3	3.5
Naïve bayes (H2o)	66	4	1	2.5
Random Forest(H2o)	79	3	5	4
Deep Learning (H2o)	77.6	5	7	6
Classification Tree (CARET)	76	2	3	2.5
Gradient Boosting - Only Power Spectrum data	74.03	3	3	3
Naïve bayes (H2o)- Only Power Spectrum data	69	3	1	2
Random Forest(H2o) - Only Power Spectrum data	74.8	2	5	3.5
Deep Learning (H2o) - Only Power Spectrum data	78.2	4	7	5.5
CAPed Deep Learning (H2o) - Only Power Spectrum data	75	2	6	4
Classification Tree (CARET) - Only Power Spectrum data	73	1.5	2.5	2

Different deep learning methods exhibit a range of accuracies (Figure 3). We manually prune the training attributes so that 'No AMR' means no Attention, Mediation or Raw values were used. 'No Pre' means that the preassigned confusion value was not used (unused in any of the previous models). It can be seen that deep learning performance can vary between 56 to 81% depending on the attributes used, with our CAPed approach giving a reasonable 75% accuracy with a minimal architecture.

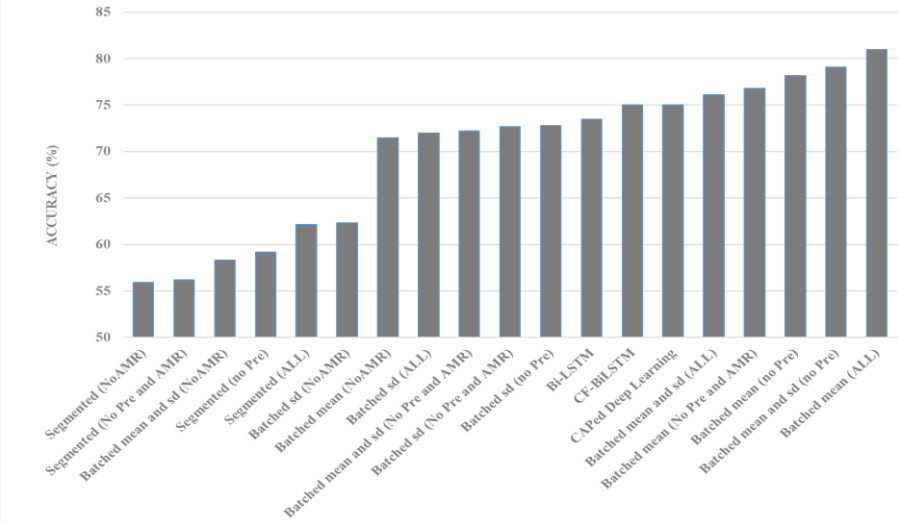

Fig. 3. Accuracy of a range of Deep Learning approaches

4 Conclusions

Modelling physiological data is a growing and important area of data modelling. Within this area, brain behavior and function is still poorly understood. This research shows that an understanding of what is happening when we observe different media of different complexity is more accurately model-able than previously thought. It appears that using publically available modelling tool we can achieve in excess of 80% accuracy in predicting if someone is 'confused' or not based on power spectrum values from a single EEG reading. Using a complicated memory based approach to weight optimization appears to be unnecessary if the correct window size for data aggregation statistical transformations are chosen. Further to this, by taking a small reduction in accuracy we can produce a minimal solution using pruning and merging of links with correlated activity.

This is the first publication to outline a correlated activity pruning approach based on link merging rather than whole node merging. This allows for more fine-tuned pruning of a deep learnt network. Secondly, this paper demonstrates that while the LSTM approach offers a performance improvement over crude time-slice segmentation, correctly dimensioned time slicing can offer both a simpler and more accurate solution for modelling the temporal EEG data used in this paper.

References

1. https://www.kaggle.com/wanghaohan/confused-eeg
2. Yin, Z., Zhao, M., Wang, Y., Yang, J., & Zhang, J. (2017). Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. Computer methods and programs in biomedicine, 140, 93-110.
3. Wang, H., Li, Y., Hu, X., Yang, Y., Meng, Z., & Chang, K. M. (2013, June). Using EEG to Improve Massive Open Online Courses Feedback Interaction. In AIED Workshops.
4. Ni, Z., Yuksel, A. C., Ni, X., Mandel, M. I., & Xie, L. (2017, August). Confused or not confused?: Disentangling brain activity from eeg data using bidirectional lSTM recurrent neural networks. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 241-246). ACM.
5. Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep?. In Advances in neural information processing systems (pp. 2654-2662).
6. H2o.ai company. H2o home - open source leader in AI and ML. https://www.h2o.ai/. Online, accessed 27-April-2019.
7. Roadknight, C., Palmer-Brown, D., and Al-Dabass, D. Simulation of correlation activity pruning methods to enhance transparency of ANNs. Int. J. Simulation 4 (1997), 68-74.
8. Roadknight, C. M., Palmer-Brown, D., and Mills, G. E. Correlated activity pruning (CAPing). In International Conference on Computational Intelligence (1997), Springer, pp. 591-592.
9. Roadknight, C. M., Balls, G. R., Mills, G. E., & Palmer-Brown, D. (1997). Modeling complex environmental data. IEEE Transactions on Neural Networks, 8(4), 852-862.
10. NeuroSky. NeuroSky's eSense meters and detection of mental state. 2009.
11. S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in Advances in Neural Information Processing Systems, 2015, pp. 1135–1143.
12. Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," arXiv preprint arXiv:1412.6115, 2014.
13. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
14. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.
15. Minli, Z., and Shanshan, Q. Research on the application of articial neural networks in tender offer for construction projects. Physics Procedia 24 (2012), 1781-1788.
16. Wang, H., Wu, Z., & Xing, E. P. (2018). Removing Confounding Factors Associated Weights in Deep Neural Networks Improves the Prediction Accuracy for Healthcare Applications. arXiv preprint arXiv:1803.07276.
17. Team, R. C. (2013). R: A language and environment for statistical computing.
18. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., & Leisch, M. F. (2009). Package 'e1071'. R Software package, avaliable at http://cran. rproject. org/web/packages/e1071/index. html.
19. Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of statistical software, 28(5), 1-26.
20. Babaeizadeh, Mohammad, Paris Smaragdis, and Roy H. Campbell. "A simple yet effective method to prune dense layers of neural networks." (2016).
21. Babaeizadeh, Mohammad, Paris Smaragdis, and Roy H. Campbell. "Noiseout: A simple way to prune neural networks." arXiv preprint arXiv:1611.06211 (2016).