Effectiveness of a Convolutional Neural Network in Sleep Arousal Classification Using Multiple Physiological Signals

Yinghua Shen
Analytics Engineering, Etsy, Inc., Brooklyn, NY, USA

Abstract

The 2018 PhysioNet Challenge utilizes 13 physiological signals collected during polysomnographic sleep studies to classify explicitly-defined arousal regions. The goal is to assign a probability of arousal at each sample for each test subject. Automatic detection of non-apnea arousals may help us better understand various causes of sleep disturbance and advance sleep arousal analysis.

Neural networks possess powerful feature-learning abilities to gain insights from complex datasets. Our approach is based on a deep convolutional neural network (CNN), which we trained with normalization, pooling, activation and dropout techniques in Python using Keras on top of Tensorflow. The CNN was trained on 737 patients' sleep data and validated on 185 patients' sleep data. Our model obtained AUROC performance score of 0.514293+/-0.054509 and AUPRC performance score of 0.501947+/-0.063199. In this paper, we discuss the strengths and limitations of our CNN in sleep arousal classification using a variety of physiological signals. We also present some possible directions for future work.

1. Introduction

Sleep is considered crucial for preserving daytime cognitive function and physiological well-being. Sleep insufficiency may be detrimental effects on our overall health and safety, and result in economic burden at both the individual and societal levels [1]. In fact, the Centers for Disease Control and Prevention (CDC) in the United States has declared insufficient sleep a "public health problem." According to a recent CDC study, more than a third of American adults are lacking enough sleep on a regular basis [2]. Furthermore, sleep disorders are commonly associated with other major medical problems such as chronic pain, mental illness and cardiovascular disease [1].

Diagnoses for sleep disorders are traditionally performed in sleep laboratories where various physiological signals of the sleeping subject are carefully reviewed by sleep experts. Apnea is one of the more well-studied sleep disorders, but it is not the only cause of sleep disturbance. Therefore, the 2018 PhysioNet/ Computing in Cardiology Challenge (henceforth referred to as "Challenge") seeks to detect non-apnea arousals during sleep using a variety of physiological signals collected during polysomnographic sleep studies [3]. A limited number of approaches for detecting sleep arousal using mainly electroencephalography (EEG), electrooculography (EOG), electromyography (EMG) and electrocardiology (EKG) already exists, such as applying wavelet analysis [4], time-frequency analysis and the support vector machine (SVM) classifier [5, 6] for the automatic detection of arousals during sleep.

In this study, we tackle the problem of classifying non-apnea arousals as proposed by the Challenge by using well-known deep learning techniques that have been successfully applied to other classification problems [7]. These techniques have proven to be effective in feature extraction, which is at the core of a useful classification algorithm. The proposed paper describes the dataset used in section 2. The data pre-processing, model architecture and training are discussed in sections 3. Finally, the results and discussions are explained in sections 4 and 5.

2. Materials

The training dataset for the Challenge consisted of 994 subjects, as described in [3]. 13 physiological signals were recorded, including EEG, EOG, EMG, EKG, oxygen saturation (SaO2), breathing effort and airflow. Except for SaO2, all signals were sampled to 200 Hz and were measured in microvolts. SaO2 was resampled to 200 Hz, and is measured as a percentage. In the training dataset, target arousal regions have been annotated with 1, non-target arousal regions with 0, and regions that will not be scored with -1. Figure 1 shows the percent of target arousal regions in a single patient. On average, only 4.88 percent of samples in all training subjects are target arousal regions, and the average length of target arousal regions is 6429 samples long. The distribution of target arousal regions in most training subjects is reasonable for the training of classifiers. The test dataset consisted of 989 subjects.

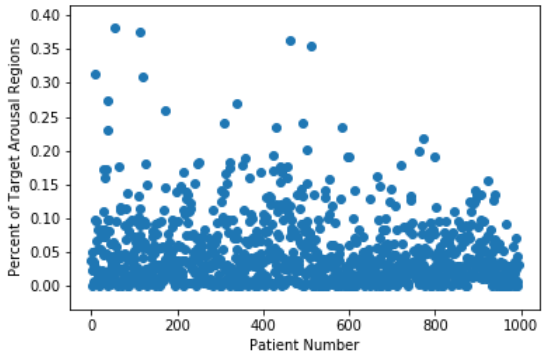

Figure 1. Percent of target arousal regions in all patients.

3. Methods

In this section, we describe the data pre-processing and deep learning approach used in the Challenge.

3.1. Data Pre-processing

We applied all 13 signals in order to account for all potential causes and attain accurate detection. We used zero-mean and unit-variance to scale each signal. We also truncated the given signals to a fixed window size (5,100,000 samples) for the CNN model.

The architecture of a CNN depends on the complexity of the problem, the amount of training data available and the resources needed to train a model. Given the large window size and the finite training data we have, we decided to output a probability of target arousal region for every 500 samples (henceforth *interval* refers to 500 consecutive samples). Therefore, we re-organized training data so that a probability is assigned to every 500 samples. To compute this probability, we discarded samples whose annotations are -1 and took the average of samples whose annotations are 0 or 1.

3.2. Problem Formulation

The target arousal region classification task is a sequence-to-sequence task that takes as input 13 physiological signals $X = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & x_{2,3} & \dots & x_{2,n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{13,1} & x_{13,2} & x_{13,3} & \dots & x_{13,n} \end{bmatrix}$ and outputs a scalar p between 0 and 1 such that p indicates the probability that interval of 500 samples contains target arousal region.

For a single patient in the training dataset, we optimize the loss function below, which was defined as the cross-entropy with L2 regularization:

$$\begin{aligned} L(\omega) &= \frac{1}{n}\sum_{i=1}^{13}\sum_{j=1}^{n} H(x_{ij}, \hat{x_{ij}}) + L2 \\ &= \frac{1}{n}\sum_{i=1}^{13}\sum_{j=1}^{n} x_{ij}\log\hat{x_{ij}} + (1-x_{ij})\log(1-\hat{x_{ij}}) \\ &\quad + \lambda||\omega||_2^2 \end{aligned}$$

where x_{ij} is the true possibility of the ij^{th} window, $\hat{x_{ij}}$ is the estimated probability of the ij^{th} window, ω is the scalar probability to be assigned, and λ is the weight decay parameter.

3.3. Model Architecture

A CNN is a supervised classification model in which low-level input is transformed through a network of filters and pooling layers. The feature produced by the model reflects properties of the data and the associated labels. Therefore, the predictive power of the model increases as more data is observed [8]. CNNs have achieved great success in areas such as computer vision and signal processing. We attempted to build a model that takes advantage of the strengths of CNN for the Challenge task.

The architecture used was chosen on the basis of other published models, as described in He 2015 [7]. The CNN takes as input 13 pre-processed physiological signals, and outputs a sequence of probabilities. The network contains 33 layers of convolutions and a fully connected layer and a softmax layer at the end. The high-level architecture of the network is shown in Figure 2.

The input is first sent to three 1D convolutional layers. Between each convolutional layer, we added max pooling layer that extracts the maximum value of the filters and provides the most informative feature. This helps to avoid redundency and reduce computational cost. The processed input is then sent to 16 residual blocks with two convolutional layers per block. All the convolutional layers have 64 filters, and each filter has a length of 16. The last layers are a fully-connected layer followed by softmax activation procedure, which produces a probability for each interval.

As our model has more parameters than a simple model, it is more prone to over-fit. Overfitting was reduced by using a number of regularization techniques, including batch normalization, dropout and early stopping.

3.4. Training

We split the training dataset provided by the Challenge into training and validation set. We used 737 subjects sleep data (80% of the data) for training and 185 subjects' sleep data for validation.

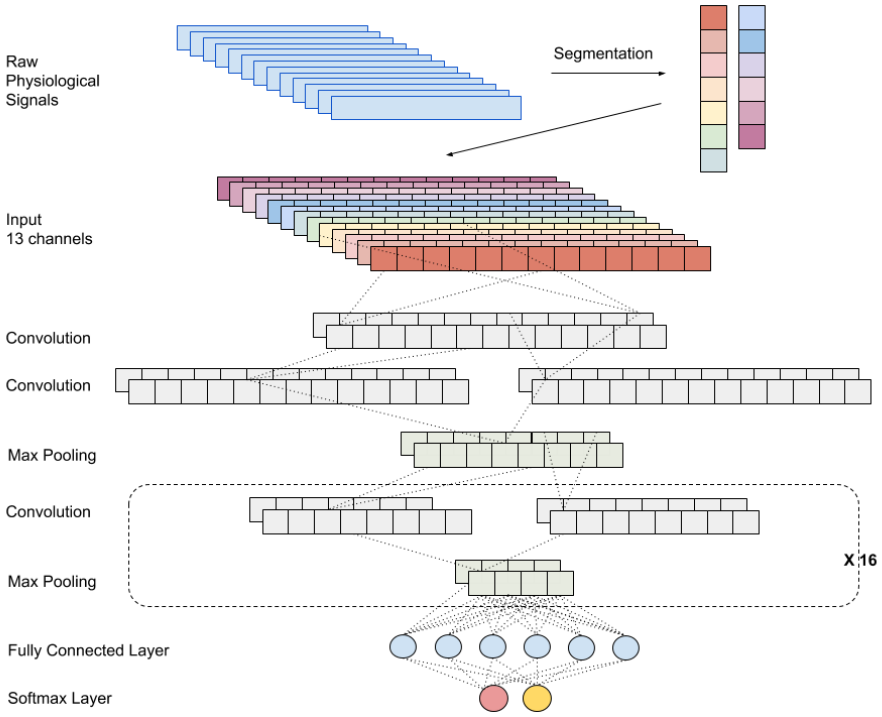

Figure 2. Put the figure legend here, clearly describing the figure.

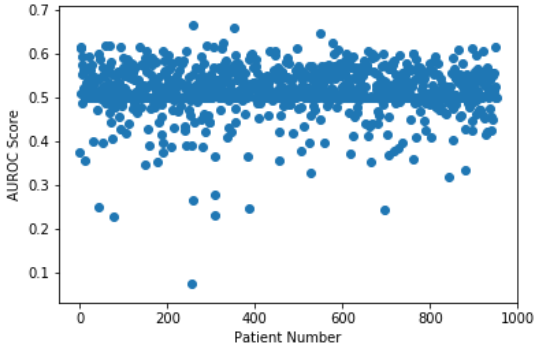

Figure 3. AUROC Score Distribution.

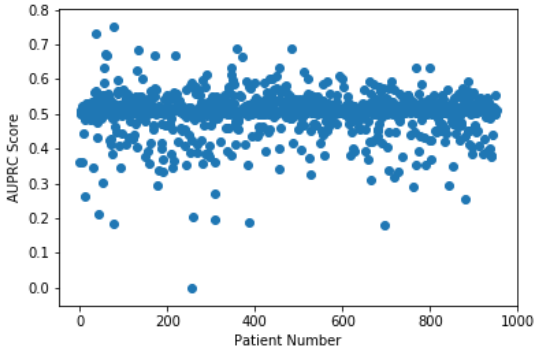

Figure 4. AUPRC Score Distribution.

During training, we used the Adam optimizer [9] with the default parameters and the mean squared error loss optimization [10]. We trained our model in Python using Keras on top of Tensorflow. A self-designed generator function is used to process multiple training subjects in parallel.

4. Results

We trained our networks from scratch and saved the best model as evaluated on the validation set. The results of cross validation on the training data provided by the Challenge obtained a testing accuracy of 0.7014 and validation accuracy of 0.7482 in Keras.

The final results are evaluated based on two evaluation metrics - AUROC (Area Under the curve of the Receiver Operating Characteristic) and AUPRC (Area Under the Precision Recall Curve). In other words, AUROC is the area under the curve where x is the false positive rate and y is the true positive rate; and AUPRC is the area under the curve where x is the recall and y is the precision.

Our best model obtained AUROC performance score of 0.514293+/-0.054509 and AUPRC performance score of 0.501947+/-0.063199. Performance distributions for individual training data subjects are shown in Figure 3 and 4.

5. Discussions

Our proposed model utilizes a 33-layer CNN for the target arousal classification task. Results obtained with the model show that it is weak in assigning a probability of target arousal to a given sample because both the AUROC and AUPRC scores were slightly above 0.5 at best.

A major reason for this poor performance is in data-preprocessing and CNN model design. We believe that the results could have improved if we put more considerations into transforming and scaling the different raw physiological data, as well as experimenting with more CNN models.

Another factor that we think contributed to the weak performance is highly imbalanced dataset. To deal with this, theoretically, we could add surrogate data that contains more target arousal regions. However, it is not very practical given the importance of continuity in analyzing individuals' sleep data [11].

Finally, we would like to discuss briefly whether or not deep learning might be suitable to process datasets such as the one provided by the Challenge, especially considering growing ethical concerns [12]. There has already been evidence that algorithms introduced in non-medical fields make problematic decisions that reflect biases inherent in the data used to train them. Such ethical issues regarding machine learning algorithms have already caused the financial sector to be exceedingly cautious about adopting AI technologies [13]. It is possible that we could build algorithms to compensate for known biases and properly deploy machine learning algorithms such that their full potential could be utilized. We think such considerations should be taken by every designer who create machine learning system for clinical use.

Acknowledgements

The author would like to thank Dr. Tom Pollard, Dr. Mohammad Ghassemi and Dr. Alistair Johnson for providing valuable insights and helpful feedback.

References

[1] Skaer TL, Sclar DA. Economic implications of sleep disorders. Pharmacoeconomics. 2010;28(11):1015-23.

[2] Prevalence of Healthy Sleep Duration among Adults–United States, 2014. Liu Y, Wheaton AG, Chapman DP, Cunningham TJ, Lu H, Croft JB.

[3] Mohammad M Ghassemi, Benjamin E Moody, Li-wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G. Mark, M Brandon Westover, Gari D Clifford, You Snooze, You Win: the PhysioNet/Computing in Cardiology Challenge 2018, Computing in Cardiology Volume 45. Maastricht, Netherlands, 2018. pp 1-4.

[4] De Carli F, Nobili L, Gelcich P, Ferrillo F. A method for the automatic detection of arousals during sleep. Sleep. 1999 Aug 1;22(5):561-72.

[5] Cho S, Lee J, Park H, Lee K. Detection of arousals in patients with respiratory sleep disorders using a single channel EEG. Conf Proc IEEE Eng Med Biol Soc. 2005;3:2733-5.

[6] Wallant DC, Muto V, Gaggioni G, Jaspar M, Chellappa SL, Meyer C, Vandewalle G, Maquet P, Phillips C. Automatic artifacts and arousals detection in whole-night sleep EEG recordings. J Neurosci Methods. 2016 Jan 30;258:124-33.

[7] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv151203385 2015;7(3):171180.

[8] LeCun, Y. et al. Deep learning. Nature 521, 436444 (2015).

[9] Kingma, Diederik P. and Ba, Jimmy. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG], December 2014.

[10] Linear Regression - Introduction to Optimization. Coursera, Michigan State University, www.coursera.org/lecture/intro-to-deep-learning/linear-regression-9lpTn.

[11] Muzet A, Werner S, Fuchs G, Roth T, Saoud JB, Viola AU, Schaffhauser JY, Luthringer R. Assessing sleep architecture and continuity measures through the analysis of heart rate and wrist movement recordings in healthy subjects: comparison with results based on polysomnography. Sleep Med. 2016 May;21:47-56.

[12] Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care Addressing Ethical Challenges. The New England journal of medicine. 2018;378(11):981-983. doi:10.1056/NEJMp1714229.

[13] AI for Finance Adoption Affected by Legal, Ethical Issues. SearchERP, searcherp.techtarget.com/feature/Legal-ethical-issues-could-slow-adoption-of-AI-for-finance.

Address for correspondence:

Yinghua Shen
117 Adams St., Brooklyn, NY 11201
yshen@etsy.com