Automated Detection of Sleep Arousals From Polysomnography Data Using a Dense Convolutional Neural Network

Matthew Howe-Patterson, Bahareh Pourbabaee, Frederic Benard

OMsignal Inc., Montreal, Canada

Abstract

In this work, a dense recurrent convolutional neural network (DRCNN) was constructed to detect sleep arousals using available Polysomnography (PSG) measurement channels provided in the 2018 Physionet challenge database. The electrocardiogram signal which is not necessary for sleep scoring is excluded from our analysis. Our model structure is composed of multiple dense convolutional units (DCU) followed by a bidirectional long-short term memory (LSTM) layer followed by a softmax output layer. The sleep events including sleep stages, arousal regions and multiple types of apnea-hypopnea/normal are manually annotated in 2018 Physionet challenge database which enable us to train our proposed network using a multi-task learning mechanism. Three binary cross-entropy loss functions corresponding to sleep/wake, arousal presence/absence and apnea/hypopnea presence/absence detections are summed up to generate our overall network loss function that is optimized using the Adam method.

The available dataset includes 994 subjects with manually annotated PSG data that were partitioned into separate training (794), validation (100) and consistent testing (100) records for each fold of cross-validation and an ensemble model. Our model performance was evaluated using two metrics: the area under the precision-recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC). To measure our model generalization, 4-fold cross-validation was also performed. For training, full night recording data was applied to our model. Finally, the average AUPRC and AUROC values associated with the arousal detection task were 0.505 and 0.922, respectively on our testing dataset. An ensemble of four models trained on different data folds improved the AUPRC and AUROC to 0.543 and 0.931, respectively. Our proposed algorithm achieves 0.573 AUPRC and 0.933 AUROC on an unseen dataset.

1. Introduction

Arousal is an abrupt change in the pattern of brain wave activity leading to a shift from deep sleep, which is commonly known as rapid eye movement (REM) sleep, to light sleep (NREM), or from sleep to wakefulness. It can become an issue if it happens constantly during sleep. According to the American Academy of Sleep Medicine (AASM) guidelines, arousal is an abrupt shift within Electroencephalogram (EEG) signal frequency bands including alpha, theta and greater than 16 Hz which lasts at least 3 seconds and is preceded with at least 10 seconds of stable condition. During REM stage, the arousal may also appear with an increase in chin Electromyogram (EMG) signal [1, 2].

The arousal has various causes such as spontaneous arousal, respiratory effort related arousal (RERA), bruxisms, hypo-ventilation, hypopneas, apneas (central, obstructive and mixed), vocalizations, snores, periodic leg movement, Cheyne-Stokes breathing or partial airway obstruction. Normally, RERA is the most common type of non-apnea and non-hypopnea arousal. Polysomnography (PSG) is an essential and standard method to investigate the sleep quality, and to detect any respiratory or non-respiratory related sleep disorders through measuring multiple physiological signals when the subject is asleep [3].

Recently, convolutional neural networks (CNN) have gained a lot of interest in physiological signal processing due to their strong capabilities in learning complex features by being directly applied on raw data without extracting any hand-crafted features [4, 5]. In this work, a dense recurrent convolutional neural network is proposed to detect arousal regions using PSG data provided in the 2018 Physionet challenge. Our network is a modified DenseNet that is proposed in [6] and is composed of multiple dense convolutional units (DCU), where each is a sequence of convolutional layers that are all connected to provide maximum information flow. It ends with a bidirectional long-short term memory layer (LSTM) with a residual skip connection and extra convolutions to convert the LSTM hidden states from forward and backward passes to the output shape. To compute the probability of arousal at each sam-

ple during training process as well as computing losses, a remapping mechanism is also proposed to simplify the network decision making process. Moreover, other task labels such as apnea-hypopnea/normal and sleep/wake are used as auxiliary tasks in a multi-task learning framework to share representations between related tasks and to improve our model generalization on our desired task which is the arousal detection.

2. Materials and Pre-Processing

The dataset includes PSG data from 1,985 subjects which were monitored at the MGH sleep laboratory for the diagnosis of sleep disorders. The data were partitioned into balanced training (n = 994), and test sets (n = 989), where the training data are provided publicly to design a model to detect target arousal regions. It includes multiple physiological signals that are all sampled at 200 Hz and are manually scored by certified sleep technicians at MGH sleep laboratory according to the AASM guidelines. More details regarding the dataset and available annotations for different sleep analysis purposes are provided in [7].

In this work, the PSG measurements (12 channels), excluding electrocardiogram signal (ECG), are used to design an arousal detector model. First, an anti-aliasing finite impulse response (FIR) filter is applied to all channels. Then, the channels are down-sampled to 50 Hz and are normalized by removing the mean and the root-mean-square (RMS) of signals in a moving 18-minute window, using fast Fourier transform (FFT) convolution. According to the AASM guidelines, the baseline breathing is established in 2 minutes. Normalizing over 18-minute interval ensures 90% overlap between the two ends of the baseline window. Our proposed normalization process is not applied to the oxygen saturation (SaO2) measurement that is only scaled to be limited in $(-0.5, 0.5)$ to avoid saturating the neural network with large values.

3. Arousal Detector Model

In this section, the DRCNN structure that is proposed to detect arousal regions is explained. Then, the multi-task learning framework is described in which all available annotations associated with the sleep/wake, arousal and apnea-hypopnea/normal events are employed to improve our network generalization.

3.1. DRCNN Network Structure

In this work, our proposed DRCNN is trained and evaluated using data down-sampled to 50 Hz to decrease computational effort and to fit a full night recording into memory to be applied to the network. The network is composed of multiple blocks, DCU1, DCU2 and LSTM which
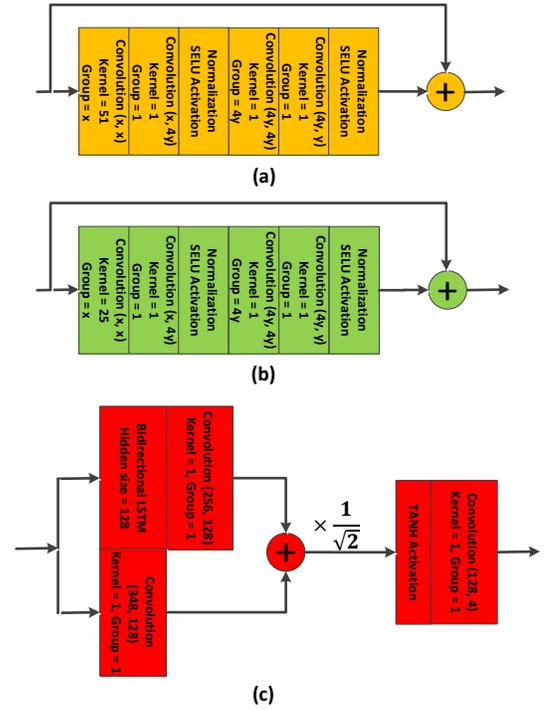

Figure 1. (a) DCU1, with no channel-wise normalization, (b) DCU2, with channel-wise normalization and (c) LSTM block, where x and y are the dimensions of input and output channels of DCUs.

are displayed in Figure 1. First, there are three DCU1s, each followed by a max-pooling layer to down-sample input signals to one entity per second. This is followed by eleven DCU2s. The DCU1s and DCU2s have similar structure comprising two sequences of two depthwise separable convolutional layers followed by the scaled exponential linear unit (SELU) activation functions.

In DCU2, weight normalization, channel-wise normalization and stochastic batch normalization [8] with a channel specific affine transform are also applied on convolutional layer outputs before using SELU activation function. Channel-wise normalization involves subtracting the mean and dividing by the standard deviation across the channel dimension independently for each time step. To extend the DCU2 receptive field, dilated convolutions are also deployed, where the dilation rates are first increased exponentially with the depth of the network along the first six DCU2s, and then are exponentially decreased along the remaining ones [9]. However, in DCU1, neither a channel-wise normalization nor a dilation factor is applied. Stochastic batch normalization is used in both DCU1 and DCU2.

Following the DCUs, a bidirectional long short-term memory (LSTM) layer with a residual skip connection

(linear 1×1 convolution) is also applied across the input channel temporal dimension. Finally, two more convolutional layers with 1×1 mapping are used to convert the LSTM hidden states from forward and backward passes to the output shape. The hyperbolic tangent (tanh) is also applied before the last convolutional layer. The overall structure of our proposed DRCNN is displayed in Figure 2.

3.2. Learning Mechanism

In this work, a multi-task learning mechanism is used to improve the generalization of our proposed arousal detector model and to learn more complex features through using other correlated tasks such as apnea-hypopnea/normal and sleep/wake. The ground truth corresponding to each task is a three-condition vector that is defined as follows:
• Arousal presence/absence detection task: (target arousal = 1, non-target arousal (apnea/hypopnea or wake) = -1, and normal = 0),
• Apnea-hypopnea/normal detection task: (obstructive apnea/hypopnea = 1, central/mixed apnea = -1, and normal = 0),
• Sleep/wake detection task: (sleep stages (REM, NREM1, NREM2, NREM3) = 1, wake = 0, and undefined stage = -1)

Considering the above three possible conditions associated with every task, 27 combinations can be defined. To investigate the distribution of the data associated with all combinations, the histogram of the labelled data is obtained. As it is displayed in Figure 3, only 13 combinations out of 27 were non-empty. To simplify the structure of the network output layer that computes joint probabilities, the non-empty bins are remapped to 4 bins that are displayed in green color in Figure 3. All the red bins corresponding to the beginning of the record before annotating the first sleep epoch (undefined sleep stage) are remapped to bin 0. The data associated with bin 0 are still processed by our model during training, however it does not contribute to the loss gradient.

It is by definition impossible to get a sleep disorder while the subject is awake (condition in bin 1 and bin 7). This happens because according to the AASM guidelines, the sleep stages are annotated in 30-second epochs. Therefore, it is necessary to update sleep/wake detection task labels upon reaching such a state. For this purpose, bin 1 and bin 7 are respectively remapped to bin 2 and bin 8. Similarly, bin 5 is remapped to bin 4 because when the arousal label is -1 and no apnea or hypopnea is present, the subject must be awake.

The last convolutional layer of our proposed DRCNN has four output channels that are soft-maxed to compute joint probabilities corresponding to bins 4, 8, 14 and 23. Then, the predicted arousal, apnea-hypopnea/normal and sleep/wake marginal probabilities are computed as:

Table 1. Cross-validation results, where each model is evaluated on its own validation dataset

Performance Metrics	Model 1	Model 2	Model 3	Model 4	Average
Arousal AUROC	0.922	0.922	0.913	0.921	0.919
Arousal AUPRC	0.557	0.505	0.524	0.529	0.528
Apnea-Hypopnea/Normal AUROC	0.956	0.958	0.960	0.972	0.961
Apnea-Hypopnea/Normal AUPRC	0.734	0.760	0.764	0.785	0.760
Sleep/Wake AUROC	0.959	0.958	0.961	0.937	0.953
Sleep/Wake AUPRC	0.826	0.834	0.853	0.767	0.820

P(arousal) = P(bin 23), P(non-arousal) = P(bin 4) + P(bin 8) + P(bin 14), P(apnea/hypopnea) = P(bin 8), P(no apnea and hypopnea) = P(bin 4) + P(bin 14) + P(bin 23), P(wake) = P(bin 4), and P(sleep) = P(bin 8) + P(bin 14) + P(bin 23).

To train our DRCNN, the apnea-hypopnea/normal and sleep/wake are used as auxiliary detection tasks. The total cross-entropy loss is computed as the weighted average of loss values corresponding to the desired and auxiliary tasks, where the arousal loss (desired task) weight is set to 2. The network weight parameters are optimized by using the Adam method which outperforms other optimization techniques in this work. In every epoch, one full-night recording is randomly selected and processed through the network. Then, to evaluate the performance of the network, the AUPRC and AUROC are obtained for validation data and the model is checkpointed if there is any improvement with any of the above scores. The full training process is repeated four times across different folds of training and validation data and finally the predictions of our four models are averaged to obtain ensemble model predictions.

4. Results and Discussion

The proposed DRCNN is applied to 12 PSG channels, excluding ECG signal. The network hyper-parameters and learning procedure are explained in Section 3. The PSG channels are first pre-processed as described in Section 2. To train our network, the available annotated data are divided into four folds, where each includes 794 training, 100 validation and 100 consistent testing records. Using a multi-task learning process, the AUPRC and AUROC are obtained for sleep/wake, arousal and apnea-hypopnea/normal detection tasks. Table 1 displays the performance metrics measured for each fold of cross-validation as well as the average performance on validation records across the 4 folds.

Using four trained models on different data folds, their corresponding predictions are averaged to form an ensemble model prediction. The ensemble model strategy improves the performance compared to the single model strategy. Table 2 displays single and ensemble model performance evaluation results on the consistent test set. It must be noted that the performance results are obtained for the up-sampled data to the original 200 Hz. Finally, applying the above ensemble model framework on an un-
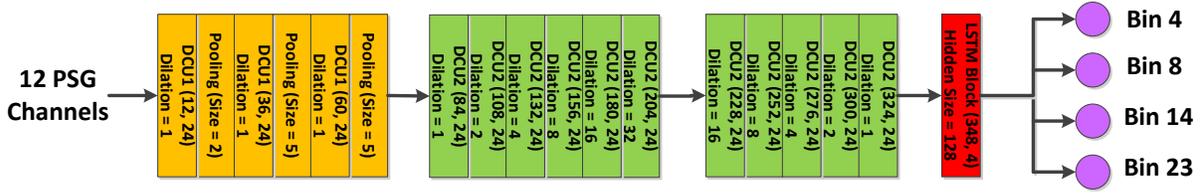
Figure 2. Proposed DRCNN architecture, including DCU1, DCU2 and LSTM block, where input and output channel dimensions are given in parentheses.

Figure 3. Bins remapping mechanism to simplify multi-task learning process

Table 2. Performance on testing records using single and ensemble model strategies

Performance Metrics	Model 1	Model 2	Model 3	Model 4	Ensemble
Arousal AUROC	0.921	0.923	0.923	0.922	0.931
Arousal AUPRC	0.492	0.497	0.519	0.511	0.543
Apnea-Hypopnea/Normal AUROC	0.951	0.955	0.954	0.965	0.965
Apnea-Hypopnea/Normal AUPRC	0.721	0.745	0.761	0.781	0.783
Sleep/Wake AUROC	0.958	0.957	0.958	0.944	0.960
Sleep/Wake AUPRC	0.831	0.822	0.822	0.771	0.832

seen dataset of 2018 challenge database results in 0.573 and 0.933 of AUPRC and AUROC, respectively.

5. Conclusion

In this paper, a modified version of the dense convolutional neural network comprising multiple convolutional and LSTM blocks is proposed to detect sleep arousal regions using 12 PSG channels that are provided in the 2018 Physionet challenge database. To improve our network generalization and to use information from other correlated tasks such as apnea-hypopnea/normal and sleep/wake, a multi-task learning procedure using hard parameter shar-

ing framework is also exploited in this work. Four DCRNN models are trained and evaluated on different subsets of training and validation data. Finally, an ensemble model is obtained through computing the average prediction of the above four models. The results confirm the superiority of our ensemble model against single model approach. On challenge unseen records, the ensemble model achieved 0.573 of AUPRC and 0.933 of AUROC.

References

[1] Halász P, Terzano M, Parrino L, Bódizs R. The nature of arousal in sleep. Journal of sleep research 2004;13(1):1–23.

[2] Berry RB, Albertario CL, Harding S, Lioyd RM, Plante DT, Quan SF, Troester MM, Vaughn BV. The AASM manual for the scoring of sleep and associated events. Rules Terminology and Technical Specifications Version 25 American Academy of Sleep Medicine ;2018.

[3] Chesson Jr, Andrew L, Ferber RA, Fry JM, Grigg-Damberger M, Hartse KM, Hurwitz TD, Johnson S, Kader GA, Littner M, Rosen G, et al. The indications for polysomnography and related procedures. Sleep 1997; 20(6):423–487.

[4] Pourbabaee B, Javan Roshtkhari M, Khorasani K. Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients. IEEE Transactions on Systems Man and Cybernetics Systems 2017; (99):1–10.

[5] Pourbabaee B, Howe-Patterson M, Reiher E, Benard F. Deep convolutional neural network for ECG-based human identification. CMBES Proceedings 2018;41.

[6] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition 2017;1(2):4700–4708.

[7] Ghassemi M, Moody BE, Lehman LH, Song C, Li Q, Sun H, Mark RG, Westover MB, Clifford GD. You snooze, you win: The physionet/computing in cardiology challenge. Computing in Cardiology Maastricht Netherlands 2018;45:1–4.

[8] Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using real NVP. arXiv preprint arXiv160508803 ;2016.

[9] Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv180301271 ;2018.