Automated Recognition of Sleep Arousal using Multimodal and Personalized Deep Ensembles of Neural Networks

Andrea Patane[1], Shadi Ghiasi[2], Enzo Pasquale Scilingo[2] and Marta Kwiatkowska[1]

[1]University of Oxford, Department of Computer Science; [2]Bioengineering and Robotics Research Center "E. Piaggio" & Dept. of Information Engineering, University of Pisa, Pisa, Italy

Abstract

***Background and Aim**: Monitoring physiological signals during sleep can have substantial impact on detecting temporary intrusion of wakefulness, referred to as sleep arousals, in order to improve the quality of sleep. To overcome the problems associated with the cubersome visual inspection of these events by sleep experts, automated sleep arousal recognition algorithms have been proposed.*

***Method**: As part of the Physionet/Computing in Cardiology Challenge 2018, this study proposes a deep ensemble neural network architecture for automatic arousal recognition from multi-modal sensor signals. Separate branches of the neural network extract features from electro-encephalography, electrooculography, electromyogram, breathing patterns and oxygen saturation level; and a final fully-connected neural network combines features computed from the signal sources to estimate the probability of arousal in each region of interest. We investigate the use of shared-parameter Siamese architectures for effective feature calibration. Namely, at each forward and backward pass through the network we concatenate to the input a user-specific template signal that is processed by an identical copy of the network.*

***Result**: The proposed architecture obtains an AUPR score of* 0.40 *on the hidden test set of the official phase of Physionet/CinC Challenge 2018. A score of* 0.45 *is obtained by means of 10-fold cross-validation on the training set provided.*

1. Introduction

Sleep affects health, mood and wellness, and studying it is important both from the theoretical and clinical point of view. The quality of sleep in patients with sleep disorders is degraded by frequent occurrences of sleep arousals, that is, temporary interruptions of wakefulness into sleep or spontaneous increase of the vigilance level [1]. Arousal stimulus may be associated with the pathophysiology of several sleep disorders (e.g. Apnea, snoring, periodic leg movement, and rapid increase of electromyogram), or may be unrelated to the pathological factors (i.e. spontaneous arousal). Polysomnography (PSG) is widely used in sleep laboratories to assess the structure and physiological changes of sleep, which makes it possible for sleep experts to monitor electroencephalogram (EEG), electromyogram (EMG), electrooculogram (EOG), electrocardiogram (ECG), breathing patterns, and other signals associated with chest, body and leg movements [2]. Generally, the scoring of arousal is done manually by sleep experts by inspecting several epochs of PSG recordings. This is, of course, a time-consuming and cumbersome task for medical technologists. Further, the outcome of the sleep scoring is crucially affected by the knowledge and experience of the human performing the scoring. As such, development of an automated arousal detection system from PSG, in the form of an efficient, fast and reliable algorithm, may provide a powerful aid to clinical practitioners.

In this study, we aim to use the current gold-standard diagnostic methods from manually annotated sleep arousals in PSG recordings to develop an automated sleep arousal detection system from a large amount of data provided by the Physionet/CinC Challenge 2018 [3]. We design a deep neural network (NN) architecture for multi-modal sleep arousal detection from EEG, EOG, EMG, Airflow and SaO2 signals. Namely, after pre-processing and data augmentation routines are applied, an ensemble of Convolutional Neural Networks (CNNs) automatically extracts relevant features separately from each input sensor channel. The feature vectors are then concatenated together and a sequence of fully-connected layers is used to estimate sleep arousal. Importantly, the architecture relies on the concept of shared-parameter Siamese networks to perform automatic feature calibration on-the-fly. Results on the challenge test set provided an Area Under the Precison-Recall curve (AUPR) score of 0.40 for the model discussed in this paper.

Related Works. Several automatic and semi-automatic detection algorithms have been proposed using different kinds of PSG signals as the input [4–8].

An automated method to discriminate arousal segments was presented by De Carli *et al.* [4]; the method relied on a combined EEG and EMG analysis. Again, EEG and EMG, pulse and SaO2 signals were used in a framework

based on data mining, which employed a meta-rule extraction to obtain arousal episodes [6]. Olsen *et al.* [5] proposed an autonomic arousal detection method based on feature learning using heart rate variability (HRV) analysis tools. The ECG signal was also used for obstructive sleep Apnea screening employing K-nearest neighbourhood and artificial neural networks as supervised classification algorithms in [7]. In another study [8], a diagnostic sleep Apnea system based on linear discriminant analysis used a combination of features based on heart rate variability analysis and SaO2.

2. Materials and Methods

In this section we briefly review the characteristics of the dataset provided by the challenge organisers, as well as the methods employed in our challenge submissions.

2.1. Dataset

The dataset provided for the Physionet/CinC challenge 2018 is split into a training and a test set. The training set is composed of 994 PSG recordings (including 6 EEG channels, EOG, 3 EMG channels, respiratory signal, SaO2 and ECG), while 989 recordings comprise the test set. The data was gathered from 1985 subjects who underwent an overnight recording sessions in the Massachusetts General Hospital (MGH).

2.2. Arousal annotations

Arousal annotations were provided only for all the samples included in the training set. According to the American Sleep Disorders Association (ASDA), alternations in EEG and EMG activities are the most significant indicators for arousal detection [9]. Furthermore, the American Academy of Sleep Medicine (AASM) defines the electroencephalographic arousal as an abrupt shift in electroencephalogram frequency, including alpha, theta, and/or frequencies greater than 16 Hz, lasting at least 3 seconds and with at least 10 seconds of previous stable sleep [2]. Another marker of arousal is related to episodes of arterial oxygen desaturation during room air breathing [8]. We plot in Figure 1 an example of 30 seconds of PSG recording with sleep arousal annotation provided in the training set.

2.3. Network Architecture

In this section we describe the NN model employed for the challenge, as well as signal pre-processing and input preparation.

Pre-processing. We investigate the use of a pre-processing step only for EEG channels, while the other signals are fed into the NN directly in the form provided in the dataset.

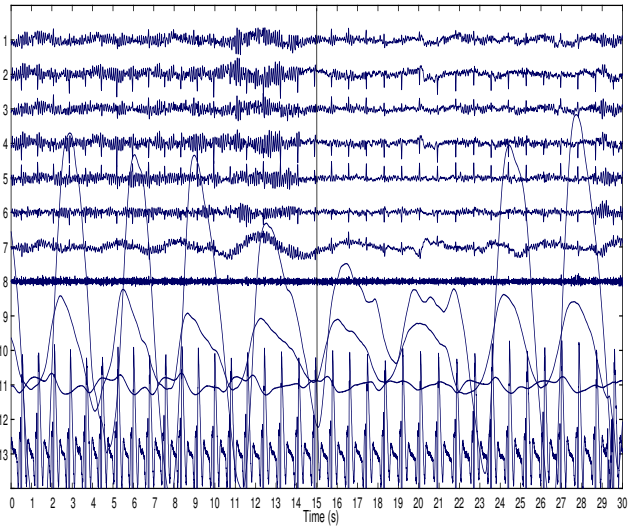

Figure 1. Example of a Polysomnography (PSG) recording during 30 seconds of sleep. The first 6 channels indicate EEG recordings, [F3-M2,F4-M1,C3-M2,C4-M1,O1-M2,O2-M1], and the 7th channel is electrooculography (EOG) signal named E1-M2 according to 10-20 standard system of EEG Placement. The rest represent electromyography (EMG), abdomen respiration, chest respiration, airflow, oxygen saturation (SaO2) and the electrocardiography (ECG), respectively. The black line at 15 second represents the onset of arousal according to the annotation provided by the PhysioNet/CinC Challenge 2018.

Regarding EEG pre-processing, first we employ a 6th order band pass filter in the $[0.5 - 45]$ Hz frequency range. Afterwards, we apply an automatic algorithm to remove candidate movement artefacts. Briefly, we compute amplitude distributions for non-overlapping 8 second time-windows of signal. We then recognise movement artifacts as those epochs above the 95th percentile of the amplitude distribution, and accordingly we discard them [10].

Windowing. We segment each signal into 30 second time windows with 50% overlap, further sub-sampling all the signals to 50 Hz and standardising them to zero mean and unit standard deviation. At each prediction step the NN processes a full time window at once giving a unique arousal score for the whole window. The results for adjacent overlapping slices of windows are then averaged together.

Data Augmentation. We heavily rely upon data augmentation both at training and testing time. We do this by randomly cropping each time window to a fixed size of 1400 consecutive time samples (i.e. 28 seconds at 50 Hz). This has the effect of reducing the dimensionality of the NN input space (hence reducing the number of weights that need to be learnt), as well as increasing the effective size of the training set available. At training time data

	L. 1	L. 2	L. 3	L. 4	L. 5	L. 6
Conv. Filters	8	8	8	16	32	32
Conv. Kernel	16	16	16	32	32	64
Max-pooling	✗	✓	✗	✓	✗	✓

Table 1. CNN for feature embedding of EEG, EOG, EMG and airflow signals. 33% dropout is used between layers. A CNN is trained for each of the input signals, and the outputs are concatenated together.

augmentation is done on-the-fly, making sure that every batch of data has as many aroused samples as non-aroused ones. This has the effect of re-balancing the dataset and forces the NN to give the same a-priori importance to both classes. Otherwise, since the non-arousal class is greatly over-represented in the dataset, the NN would give it preference, thus negatively affecting the AUPR score.

At test time, 10 rounds of data augmentation are applied to each time window and the final prediction is taken as the mean value of those, in an effort to average out the stochasticity that arises from signal cropping.

Architecture. The overall architecture is composed of an ensemble of NNs, where each NN is separately responsible for embedding each specific channel into a lower-dimensional vector space, i.e. the feature space, then merged together and fed through a Siamese architecture [11]. For most of the channels (that is, EEG, EOG, EMG and airflow channels) the embedding is obtained by processing the inputs with a CNN model, whose architecture is described in Table 1. This is a 6 layers one-dimensional CNN architecture, where we use Parametric ReLU nodes as activation functions [12], with three fully-connected layers stacked on top (of 256, 128 and 128 units each). On the other hand, the SaO2 signal is processed by means of only four fully-connected layers (of 512, 256, 64 and 64 units each). The rationale for this is that the latter did not benefit from feature embedding, as it seems mostly to be a baseline type of signal. Feature vectors obtained for each input signal are then concatenated together into a single overall feature vector. We build a shared-parameter Siamese NN on top of the latter, in order to achieve effective user-specific feature calibration [13]. This is done by making exact copies of the NNs described above, and applying them to specific templates extracted for each user. We empirically find good performance by selecting two templates for each user (see Section 3), and randomly looking at input samples close in time to that currently under analysis.

Finally, three fully-connected layers (of 1024, 512 and 256 units each) merge together the calibrated features in a non-linear fashion and provide the final prediction on the arousal level through a final soft-max activation function.

Implementation and Training. We train the overall NN architecture end-to-end relying on the Adam optimiser [14]. Training is performed for a maximum of 50 epochs, and we use early-stopping if the AUPR (which is the ranking score used in the challenge) on the validation set does not improve for 10 consecutive epochs. We finally use the model that obtained the best AUPR score on the validation set, among those explored by the optimisation algorithm throughout the learning process.

For the final entry we use 85% of the dataset for training and the rest for validation, while for cross-validation results we use 80% for training, 10% for validation and 10% for testing. We implement the model in Keras [15] using Tensorflow backend [16][1]. Training is done on an NVIDIA Tesla K80 GPU, with training time of about 22 hours.

3. Experimental Results

Table 2 lists a comparison of cross-validation results obtained on the training set provided by the challenge organisers. Namely, we compare single-modal classification results with multi-modal ones, and analyse how the AUPR is affected by the number of Siamese copies of the network, that is, 0-Siamese (i.e. standard non-Siamese architecture), 1-Siamese (i.e. standard Siamese architecture where two copies of the same network exist) and 2-Siamese (i.e. a variation on the standard Siamese network, in which we consider 3 copies of the same network) networks. Because of the challenge computation time constraints, we have limited our analysis to 5 channels only, namely: (i) C3-M2 (central) EEG; (ii) E1-M2 EOG; (iii) Abdominal EMG; (iv) Airflow; and (v) Oxygen Saturation level (SaO2). In fact, the multi-modal model is trained only on these 5 channels. We leave for future work an investigation of how to include the remaining channels into the proposed neural network architecture.

We observe how increasing the number of Siamese copies of the network generally increases the AUPR as well. In fact, the increase is higher when comparing 0-Siamese to 1-Siamese than when comparing 1-Siamese to 2-Siamese. Though the trend possibly persists when increasing the number of Siamese copies to a number $n > 2$, such an analysis was not considered for our challenge submission. Notice how different sensors have different first-order contributions to the final AUPR score. In fact, the 2-Siamese network built using only Abdominal EMG signal obtains already $\approx 84\%$ of the AUPR obtained by the multi-modal model; with Airflow being the second most important channel and EEG the least relevant. These scores, however, are relative to the particular architecture employed and do not necessarily generalise to other architectures.

The final result that the 2-Siamese multi-modal network obtains in the challenge test set is an AUPR of 0.40.

[1] An implementation can be found at `https://github.com/andreapatane/SiameseNet-PhNet2018Challenge`.

	C3-M2 EEG	E1-M2 EOG	ABD EMG	AIR Flow	SaO2	Multi modal
0	0.09	0.16	0.29	0.25	0.15	0.34
1	0.13	0.20	0.35	0.27	0.20	0.40
2	0.16	0.22	0.38	0.27	0.23	0.45

Table 2. Average AUPR score for cross-validation results on single-modal and multi-modal arousal classification. Different rows correspond to the number of Siamese copies of the network.

4. Discussion and Conclusions

In this work we have presented a neural network architecture for multi-modal sleep arousal detection. This was built by training an ensemble of CNNs for feature space embedding, and relying on a shared-parameter Siamese architecture to effectively enable feature-level calibration. While working directly on raw data for the other sensor channels, for EEG processing we crucially relied on frequency-based pre-processing of the signal, which allowed us to take advantage of the relationship that exists between the frequency shift of EEG and sleep arousal.

By means of cross-validation, we have empirically shown the advantages of the Siamese architecture compared to the standard one in the problem addressed here, and evaluated first-order effects of how single sensors contribute to the final model. The presented model obtains a final AUPR score of 0.40 in the hidden test set of the Physionet/CinC Challenge 2018.

Future work will investigate the inclusion of the remaining sensor channels into the network architecture, as well as the development of CNNs specifically tailored for each different sensor signal. Finally, empirical results suggest that further improvements can potentially be obtained by generalising the standard Siamese architecture and exploring different strategies for template generation, which we intend to study in future.

5. Acknowledgements

This project was partially funded by the EU's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 722022.

Address for correspondence:

Andrea Patane
andrea.patane@cs.ox.ac.uk
Department of Computer Science
University of Oxford.

References

[1] Bonnet M, Carley D, Carskadon M, Easton P, Guilleminault C, Harper R, Hayes B, Hirshkowitz M, Ktonas P, Keenan S, et al. ASDA Report. EEG arousals: scoring rules and examples. Sleep 1992;15(2):173–184.

[2] Berry RB, Brooks R, Gamaldo CE, Harding SM, Marcus C, Vaughn B, et al. The AASM manual for the scoring of sleep and associated events. Rules Terminology and Technical Specifications Darien Illinois American Academy of Sleep Medicine 2012;.

[3] Ghassemi MM, Moody BE, Lehman LwH, Song C, Li Q, Sun H, Mark RG, Westover BM, Clifford GD. You snooze, you win: the physionet/computing in cardiology challenge 2018. In Computing in Cardiology, volume 45. 2018; 1–4.

[4] De Carli F, Nobili L, Gelcich P, Ferrillo F. A method for the automatic detection of arousals during sleep. Sleep 1999; 22(5):561–572.

[5] Olsen M, Schneider LD, Cheung J, Peppard PE, Jennum PJ, Mignot E, Sorensen HBD. Automatic, electrocardiographic-based detection of autonomic arousals and their association with cortical arousals, leg movements, and respiratory events in sleep. Sleep 2018;41(3):zsy006.

[6] Shmiel O, Shmiel T, Dagan Y, Teicher M. Data mining techniques for detection of sleep arousals. Journal of neuroscience methods 2009;179(2):331–337.

[7] Mendez MO, Bianchi AM, Matteucci M, Cerutti S, Penzel T. Sleep apnea screening by autoregressive models from a single ECG lead. IEEE transactions on biomedical engineering 2009;56(12):2838–2850.

[8] Ravelo-García AG, Kraemer JF, Navarro-Mesa JL, Hernández-Pérez E, Navarro-Esteva J, Juliá-Serdá G, Penzel T, Wessel N. Oxygen saturation and RR intervals feature selection for sleep apnea detection. Entropy 2015; 17(5):2932–2957.

[9] EEG A. Scoring rules and examples: a preliminary report from the sleep disorders atlas task force of american sleep disorders association. Sleep 1992;15:173–184.

[10] Valenza G, Greco A, Bianchi M, Nardelli M, Rossi S, Scilingo EP. EEG oscillations during caress-like affective haptic elicitation. Psychophysiology 2018;e13199.

[11] Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R. Signature verification using a "siamese" time delay neural network. In Advances in neural information processing systems. 1994; 737–744.

[12] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision. 2015; 1026–1034.

[13] Patane A, Kwiatkowska M. Calibrating the classifier: siamese neural network architecture for end-to-end arousal recognition from ECG 2018;.

[14] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv14126980 2014;.

[15] Chollet F, et al. Keras. `https://github.com/keras-team/keras`, 2015.

[16] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`.