Automatic Sleep Arousal Identification From Physiological Waveforms Using Deep Learning

Daniel Miller, Andrew Ward, Nicholas Bambos

Stanford University, Stanford, CA, USA

Abstract

The 2018 PhysioNet Computing in Cardiology Challenge focused on diagnosing sleep disorders, motivated by enabling treatment to alleviate the associated mental and physical health consequences. The dataset consists of 1,985 patients monitored at an MGH sleep laboratory where vital signs were recorded, and arousal regions were annotated by experts. This work presents a deep-learning method to identify sleep arousals. In traditional machine learning, feature extraction is one of the most time-intensive considerations, requiring a great deal of domain expertise and experimentation. In contrast, deep learning techniques automatically learn variable interactions between pairs or groups of signals, and any relevant temporal dependencies. This allows such algorithms to automatically extract sleep patterns from rich physiological time series. The model presented here integrates ideas from several successful deep learning models to construct a multi-channel time-series convolutional-deconvolutional neural network. This network was trained using cross-entropy loss, and evaluated on a 20% held-out validation set. Hyper-parameters were selected on the AUPRC metric, and training utilized early stopping to prevent overfitting. The resultant model achieved an AUPRC of 0.369 and an AUROC of 0.855 on the final competition test set.

1. Introduction

Improving sleep quality is an important concern due to the significant detrimental impacts of poor sleep on health quality. Despite recent discoveries in the mechanisms governing human sleep patterns, many things about the reasons we sleep are still not understood [1, 2]. Despite these gaps, significant associations have been well-established between poor sleep quality and a wide range of negative outcomes [3,4]. The 2018 PhysioNet Computing in Cardiology (CinC) challenge seeks to develop automated methods for quantifying sleep states, motivated to discover the sources of non-apnea sleep arousals.

Machine learning (ML) provides an extremely flexible toolset for solving problems involving large-scale data. In many cases, ML algorithms enable decision support tools which assist front-line healthcare providers to make efficient clinical decisions from significantly more information than would otherwise be possible [5, 6].

Selecting a specific ML algorithm is initially governed by the size and dimensions of the available data, the expected patterns in the data, and the available computational resources. While significant advantages may be conferred by domain expertise and extensive feature engineering, as is done in traditional statistical learning models, these advantages may be rendered unnecessary by sufficiently powerful models in conjunction with large datasets. Deep learning models relax model constraints to increase flexibility and power, and ameliorate the consequent overfitting issues by reducing learning rates and training the models on many samples. Convolutional neural networks (CNN's) are an extremely efficient and parallelizable set of models for structuring temporally or spatially correlated data [7–10]. Although developed for computer vision and image processing, CNN's have demonstrated applications in many forms of time series data, including multi-channel physiological waveforms [11].

2. Methods

2.1. Overview

13 channels of annotated sleep waveforms were used as inputs to the model, including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiology (ECG), and oxygen saturation (SaO2), as shown in Figure 1. A Convolutional-Deconvolutional Neural Network was provided these waveforms and trained to predict the probability of arousal at each time step. The network was inspired by densely connected CNN and semantic segmentation networks which generate attention maps over an image [7–10]. This architecture allows for a flexible model with applications in many variable-length time series tasks, and captures interactions between temporally-correlated physiological signals in its convolutional filter weights. The final model

ISSN: 2325-887X DOI: 10.22489/CinC.2018.242

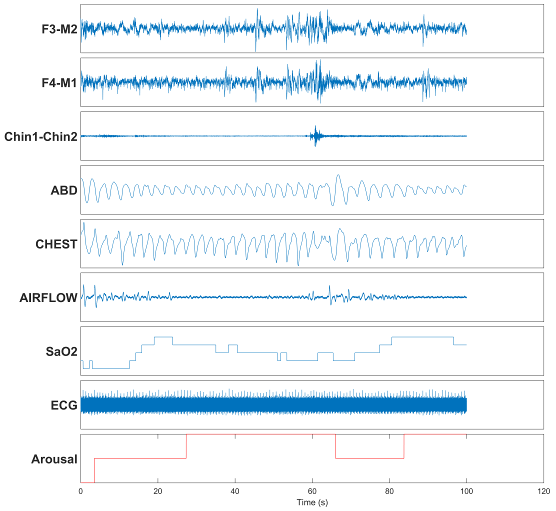

Figure 1. Time series waveforms given as inputs to the competition (blue), and labels determining whether the patient was in a sleep arousal phase (red).

was implemented in PyTorch and trained on an NVIDIA Tesla M60 GPU to 23 epochs in roughly 13 hours.

2.2. Model Architecture

The convolutional-deconvolutional network architecture used to identify arousal regions is shown in Figure 2. The final model comprised 8 convolutional layers, 8 deconvolutional layers, and a single fully connected layer. A softmax was applied to the final layer to produce probability estimates across the binary classes. Multi-channel 1-dimensional filter kernels were convolved along the time dimension. Each convolutional layer group contained—in order—a sequence of a 1-dimensional convolution layer, dropout regularization, batch normalization, and a ReLU nonlinearity. The final model included 1x2 maxpool layers following the ReLU's of the second and third groups. The model's Deconvolutional layers were comprised of of a 1-dimensional deconvolution, or transpose convolution, followed by a batch normalization layer, and a final ReLU.

The model also utilized "skip" cross-connections: the input of each deconvolutional layer included both the output of the previous deconvolutional layer and the output of the corresponding convolutional layer. The two inputs were concatenated in the channel dimension. These skip connections gave the later deconvolutional layers direct access to higher-level network features, preventing the network from needing to implicitly pass all low-level information through the entire network. Finally, the fully connected layer was added to allow the network to compare the output activations from the penultimate layer with each of the other outputs temporally in order to make a decision about each output element based on the entirety of a patient's information.

2.3. Training Phase

To train the model, the 994 patients in the labeled training dataset were randomly split into an 80% training split of 795 patients, and a 20% validation split of 199 patients. The amount of waveform data for each patient varied widely, ranging from approximately 3 million waveform values (~ 4 hours) to over 7 million values (~ 10 hours). The data for all patients' waveforms was -1-padded to a length of $7,159,808 = 437 \cdot 2^{14}$ which was both longer than the maximum sample length for any single patient in both the train and test set, and also a convenient multiple of a power of 2, facilitating hyper-parameter exploration, particularly in the number of layer groups, and both maxpool and convolutional strides.

The model was trained on an inverse class-frequency-weighted binary cross-entropy loss function over the output probabilities. The label frequency across the training set was 61.7% non-arousal (0), 4.46% arousal (1), and 33.84% not-scored (-1). Therefore, the loss function enforced a much higher penalty on misclassifying arousals than mis-classifying non-arousals. The model weights were updated using an Adam optimizer. All values labelled as -1, either from the test labelling or from the max-length padding, were masked from the loss function, and therefore did not contribute to the optimization gradients.

2.4. Hyperparameters

Table 1. Hyperparameter settings used for our final model. These values were tuned using the AUPRC of the 20% held-out validation data.

Hyperparameter	Value
learning rate	1×10^{-3}
dropout probability	0.05
conv kernel sizes	[63, 7, 7, 7, 7, 3, 3, 3]
conv output sizes	[16, 32, 32, 32, 32, 64, 64, 64]
conv strides	[32, 2, 2, 2, 2, 1, 1, 1]
deconv kernel sizes	[3, 3, 3, 7, 7, 7, 7, 63]
deconv output sizes	[64, 64, 64, 32, 32, 16, 16, 8]
deconv strides	[1, 1, 1, 2, 2, 4, 8, 16]

The training phase examined a wide range of hyperparameters, including the number of layers, convolutional filter kernel and stride sizes, optimizer learning rate, dropout probability, and combinations of maxpool layers and skip connections. The final network parameters are given in Table 2.4. The hyperparameters were chosen based off of the

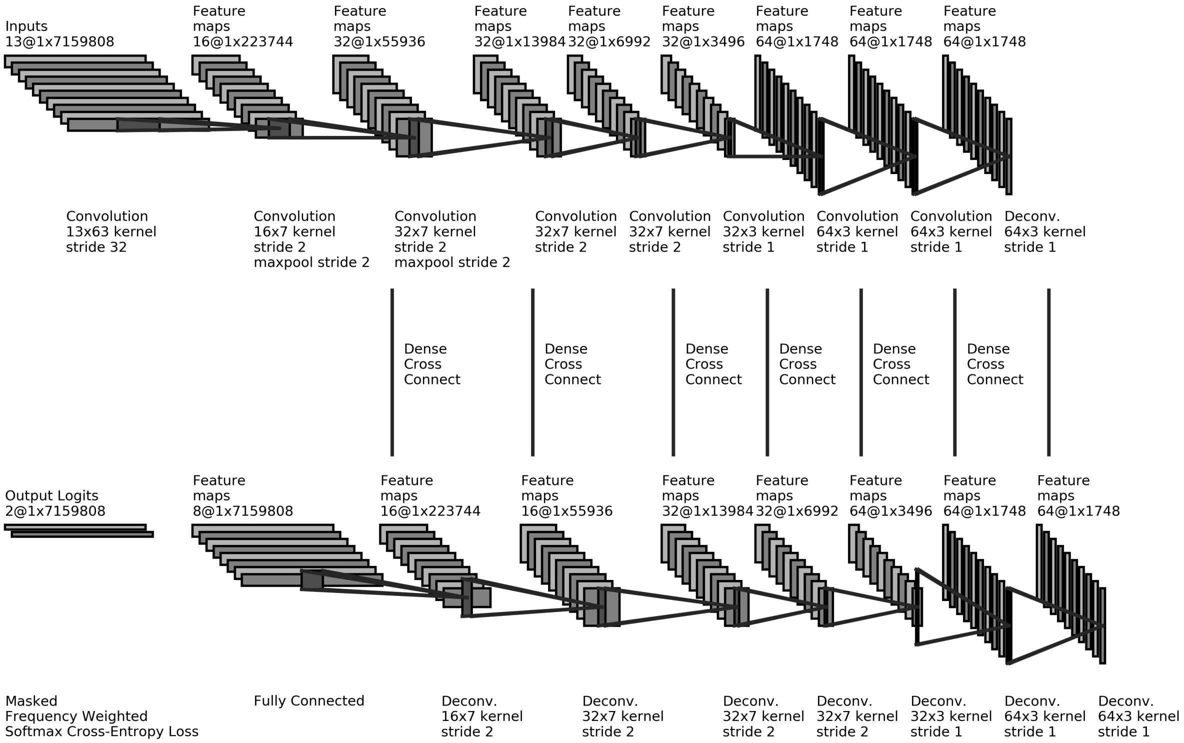

Figure 2. Convolutional-deconvolutional arousal identification network with cross-connections.

maximum validation AUPRC over the entire training run. Because each input to the model was an entire patient's data (13 waveforms, padded to a length of 7,159,808), the hardware resources were restricted to implementing stochastic gradient descent with batch sizes of 1. However, this was not the only restriction due to the choice of input. The sheer data size limited the number of the activations which could be stored in memory, and, since the activations decreased in size with every subsequent layer, this problem chiefly limited hyperparameter selection for the first convolutional (and last deconvolutional) layer. The first layer's large filter size of 63 and stride of 32 alleviated the top-level memory constraints.

3. Results

The final AUROC, AUPRC, and loss metrics are shown in Table 3. The training loss was notably lower than, and the AUROC/AUPRC metrics notably higher than their respective validation values, indicating an overfit to the training data, despite the dropout regularization in the convolutional filters. Further, the validation split metrics slightly outperformed the test set; indicating a either an overfitting of the validation set due to repeated hyper-parameter selection, or an artifact of the smaller validation split.

Table 2. Area under the receiver operating characteristic curve, area under the precision-recall curve, and the frequency-weighted cross-entropy loss for each split.

	AUROC	AUPRC	Loss
Train	0.914	0.496	0.051
Validation	0.856	0.406	0.078
Test	0.855	**0.369**	-

Figure 3 shows three metrics evaluated on both the training and validation set as the model trained. The training phase employed early stopping to prevent overfitting; the final model used the model weights from epoch 18, as the model achieved the highest AUPRC after this epoch. However, the training curves show that the training/validation loss diverge quickly, and the training/validation AUPRC

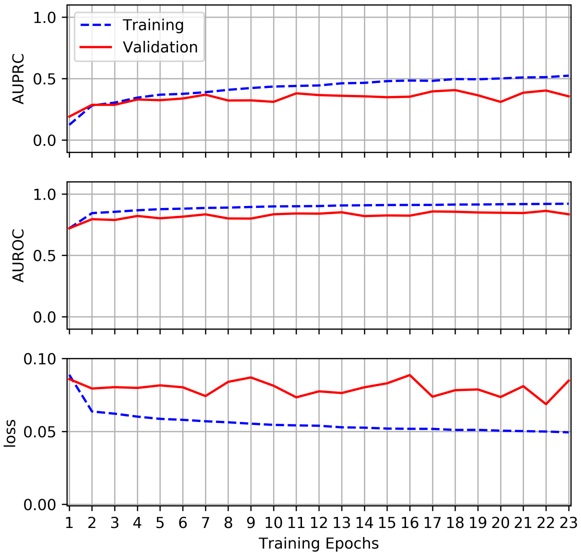

Figure 3. Training curves for the final model, showing the AUPRC (top), AUROC (middle), and loss (bottom) for the training (blue) and validation (red) splits. The training employed early stopping to prevent overfitting; the final model implemented the network weights after epoch 18.

diverges around epoch 7, indicating that the model began overfitting far before epoch 18. Although higher dropout rates decreased the model's performance, future work could explore other methods of model regularization, such as adding an L_2 weight penalty to the loss function.

4. Conclusions and Future Work

The convolutional-deconvolutional model demonstrated significant precision and recall on a complex medical diagnostics task with minimal reliance on domain-specific knowledge. As this model was trained on the raw data with no pre-processing, it may be applied to similar classification problems with minimal modification.

Acknowledgements

Daniel Miller is supported by Stanford University via the Stanford Graduate Fellowship. Andrew Ward is supported by the Department of Defense via the National Defense Science & Engineering Graduate Fellowship.

References

[1] Hardin PE, Hall JC, Rosbash M. Feedback of the drosophila period gene product on circadian cycling of its messenger rna levels. Nature 1990;343(6258):536.

[2] Nicholson C. Importance of sleep: Six reasons not to scrimp on sleep, 2006. Harvard Health Online Journal. Available from http://www.health.harvard.edu/press_releases/importance_of_sleep_and_health.

[3] Lee M, Choh A, Demerath E, Knutson K, Duren D, Sherwood R, Sun S, Chumlea WC, Towne B, Siervogel R, et al. Sleep disturbance in relation to health-related quality of life in adults: the fels longitudinal study. JNHA The Journal of Nutrition Health and Aging 2009;13(6):576–583.

[4] Nutt D, Wilson S, Paterson L. Sleep disorders as core symptoms of depression. Dialogues in Clinical Neuroscience 2008;10(3):329.

[5] Miller D, Scheinker D, Bambos N. A practical approach to machine learning for clinical decision support. In Cappanera P, Li J, Matta A, Sahin E, Vandaele NJ, Visintin F (eds.), Health Care Systems Engineering. Cham: Springer International Publishing. ISBN 978-3-319-66146-9, 2017; 111–120.

[6] Miller D. Automated identification of drug-drug interactions in pediatric congestive heart failure patients. CoRR 2017;abs/1702.04615. URL http://arxiv.org/abs/1702.04615.

[7] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015; 3431–3440.

[8] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In CVPR, volume 1. 2017; 3.

[9] Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE, 2017; 1175–1183.

[10] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention. Springer, 2015; 234–241.

[11] Miller D, Ward A, Scheinker D, Shin A, Bambos N. Physiological waveform imputation of missing data using convolutional autoencoders. In International Conference on E-health Networking, Application & Services (HealthCom). IEEE, 2018; To appear.

Address for correspondence:

Daniel Miller
350 Serra Mall Stanford CA 94305
danielrm@stanford.edu