EasyChair Preprint

№ 6910

Self-Attention Long-Term Dependency Modelling in Electroencephalography Sleep Stage Prediction

Georg Brandmayr, Manfred Hartmann, Franz Fürbass and Georg Dorffner

October 20, 2021

Self-Attention Long-Term Dependency Modelling in Electroencephalography Sleep Stage Prediction

Georg Brandmayr[✉1,2][0000−0001−6771−9843], Manfred Hartmann[2][0000−0002−1446−7600], Franz Fürbass[2][0000−0002−6744−2802], and Georg Dorffner[1][0000−0002−3181−2576]

[1] Section for Artificial Intelligence, Medical University of Vienna, Vienna, Austria
[2] Center for Health & Bioresources, AIT Austrian Institute of Technology GmbH, Vienna, Austria
georg.brandmayr@ait.ac.at

Abstract. Complex sleep stage transition rules pose a challenge for the learning of inter-epoch context with Deep Neural Networks (DNNs) in ElectroEncephaloGraphy (EEG) based sleep scoring. While DNNs were able to overcome the limits of expert systems, the dominant bidirectional Long Short-Term Memory (LSTM) still has some limitations of Recurrent Neural Networks. We propose a sleep Self-Attention Model (SAM) that replaces LSTMs for inter-epoch context modelling in a sleep scoring DNN. With the ability to access distant EEG as easily as adjacent EEG, we aim to improve long-term dependency learning for critical sleep stages such as Rapid Eye Movement (REM). Restricting attention to a local scope reduces computational complexity to a linear one with respect to recording duration. We evaluate SAM on two public sleep EEG datasets: MASS-SS3 and SEDF-78 and compare it to literature and an LSTM baseline model via a paired t-test. On MASS-SS3 SAM achieves $\kappa = 0.80$, which is equivalent to the best reported result, with no significant difference to baseline. On SEDF-78 SAM achieves $\kappa = 0.78$, surpassing previous best results, statistically significant, with +4% F1-score improvement in REM. Strikingly, SAM achieves these results with a model size that is at least 50 times smaller than the baseline.

Keywords: attention · sleep scoring · inter-epoch context.

1 Introduction

The visual scoring of sleep stages based on polysomnography (PSG) is essential for the diagnosis of many sleep disorders, but the instrumentation burden and time effort limit it's application. While expert systems could solve sleep scoring of PSG already in 2005 [1] with human level agreement, limitations were in high development effort and low flexibility for reuse. End-to-end learning via Convolutional Neural Networks (CNNs) enabled sleep scoring based on electroencephalography (EEG) only (the brain signal subset of PSG) and promises to reduce both burden

and time effort [22]. Figure 1 shows an example of nightly EEG with a hypnogram according to the rules of the American Association of Sleep Medicine (AASM). It depicts a time series of sleep stages, i.e., brain-states, based on 30 s segments termed *epochs* (not to be confused with training epochs). Soon after the first CNN solutions it was realized that sleep scoring is not only a pattern recognition problem, but also a sequence transduction problem. Complex sleep stage transition rules [4] pose a challenge for the learning of sleep EEG inter-epoch context with DNNs. The inset in Fig. 1 shows an example of a prolonged Rapid Eye Movement (REM) stage. Sequence transduction tasks are typically found in Natural Language Processing (NLP) and can be solved with encoder-decoder architectures based on Recurrent Neural Networks (RNNs). However, it was not until the solution of the vanishing gradient problem with LSTM, that many sequence tasks such as translation were improved dramatically [9]. Automatic sleep scoring was no exception, and today the modeling of inter-epoch context is dominated by bidirectional LSTMs [12, 20, 21]. However, their sequential nature makes them harder to train than feed-forward networks and the fixed size state may still limit representation of distant sequence elements. NLP research demonstrated that direct access to encoded sequence elements via the attention mechanism improves performance [3]. The Transformer model drops RNNs completely and uses a pure feed-forward, self-attention based, encoder-decoder mechanism for sequence transduction [23].

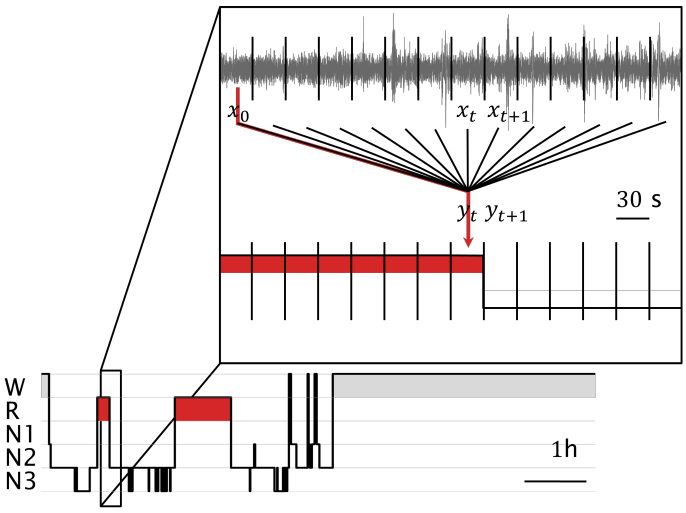

Fig. 1. Hypnogram sleep scoring from EEG with sleep stages W awake, R rapid eye movement (REM) sleep (red), N1 transitional sleep, N2 normal sleep and N3 deep sleep. The inset shows an inter-epoch context of 14 epochs (30 s long) for prolonged R scoring.

2 Motivation

In this paper we seek to replace LSTMs with a self-attention-based sequence encoder for epoch features in a sleep-scoring DNN, demonstrating an application to time series classification. We argue that directly accessing epoch features reduces the long-term dependency challenge, since no encoding into states is necessary. This could aid especially stage R (REM sleep), which can depend on distant epochs in the past and future. Since attention relates every output element with every input element, the computational complexity is quadratic in sequence length. This becomes an issue when scoring an entire night of EEG, opposed to NLP with short sequences [23]. As solution we propose to restrict the attention context L to a fixed size, moving scope. The inset in Fig. 1 shows an example of a restricted context of $L = 14$ epochs, where the prediction of sleep stage y_l depends not only on $\mathbf{X}_l$ but also on $\mathbf{X}_0$. With this approach, the computational complexity is reduced to linear, however at the price of a limited attention scope. Due to the biological limitation of sleep cycle length we conjecture that a full night scope is not required. Thus, it remains a tradeoff to choose a scope size L. We hypothesize, that the direct access to distant epochs adds more representative capacity, than what may be lost by limited scope.

3 Method

Competitive sleep scoring models typically use, with only few exceptions [18], a sequence to sequence approach [19–21] based on LSTMs. The proposed model SAM also uses this approach, but replaces bidirectional LSTMs with a sequence encoder stack, based on the encoder part of the Transformer. Figure 2 left shows the main building blocks. The model estimates from a length L sequence of EEG epochs $\boldsymbol{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(L)}]$ a sequence of sleep stages $y \in V^L$ with the set of sleep stages V. While some other work uses transformed versions of the EEG [6, 19], here we process raw EEG. The input $\boldsymbol{X} \in \mathbb{R}^{f_\mathrm{s} \times T \times C}$ is a high-resolution signal with sampling frequency f_s, C channels and T seconds duration. Overall, the output probability sequence $\mathbf{P} \in \mathbb{R}^{L \times K}$ with $K = |V|$ is estimated from the input $\boldsymbol{X}$ via the non-linear, parameterized function g:

$$\mathbf{P} = g(\boldsymbol{X}; \theta) = f_\mathrm{s}([f_\mathrm{e}(\mathbf{X}_0; \theta_\mathrm{e}), \ldots, f_\mathrm{e}(\mathbf{X}_L; \theta_\mathrm{e})]\,; \theta_\mathrm{s}) \tag{1}$$

with parameters $\theta = (\theta_\mathrm{e}, \theta_\mathrm{s})$ for embedder f_e and sequence encoder f_s.

3.1 Embedder

The epoch embedder operates on each sequence element individually, and thus treats the sequence like a batch, indicated by the dashed line in Fig. 2. It is a CNN designed to reduce a raw EEG epoch $\mathbf{X}_l$ to a vector $\mathbf{e}_l \in \mathbb{R}^{N_\mathrm{F}}$ of high level, representative features. Every CNN layer has the same kernel size and is followed by batch normalization [10] and ReLU activation, with padding and stride set for identical in- and output resolution. The actual reduction of resolution gets

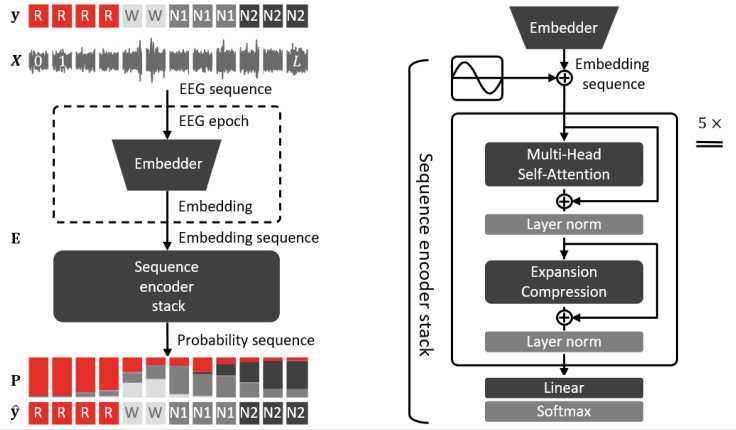

Fig. 2. Left: Model overview. The model predicts from the length-L EEG sequence $\boldsymbol{X}$ the sleep probability sequence $\mathbf{P}$ (W white, R red, N1–3 grayscale). **Right: Encoder.** Five layers of MHSA and feed-forward net model relational representations.

done via identical max pooling layers with kernel size two. We found identical resolution reduction performing better than aggressive resolution reduction [18]. Final max pooling concludes the embedder and reduces the finally N_F feature maps to dimension $1 \times N_\mathrm{F}$. Skip connections [8] after every second max pooling layer facilitate deep training and diversify the receptive field [14].

3.2 Sequence encoder stack

The embeddings, stacked to the sequence $\mathbf{E} \in \mathbb{R}^{L \times N_\mathrm{F}}$, are the input for inter-epoch context modelling in the sequence encoder stack. It uses, unlike most state-of-the-art solutions based on RNNs, only self-attention to model the inter-epoch dependencies in the encoded sequence $\mathbf{Z} \in \mathbb{R}^{L \times N_\mathrm{F}}$.

Attention. It solves the problem of access to past information without regard of the distance and avoids the bottleneck of squeezing the past into a single state vector. It computes relational representations of sequences via a form of content-addressed retrieval. Attention is a function $f(\mathbf{q}_l, \mathbf{k}, \mathbf{v})$ that maps a query vector $\mathbf{q}_l$ via key-value pairs to a retrieved context vector $\mathbf{c}_l$. With a query of dimension Q, $L \times Q$ key matrix $\mathbf{K}$ and $L \times V$ value matrix $\mathbf{V}$ the l-th context $\mathbf{c}_l$ of dimension V is:

$$\mathbf{c}_l = \boldsymbol{\alpha}_l \mathbf{V} \tag{2}$$

with $\boldsymbol{\alpha}_l = \mathrm{softmax}(f_\mathrm{score}(\mathbf{q}_l, \mathbf{K})$ and the scoring function f_score. Among different attention functions [16], we consider dot-product attention $f_\mathrm{score}(\mathbf{q}_l, \mathbf{K}) = \mathbf{q}_l^\mathsf{T} \mathbf{K}^\mathsf{T}$ for its favorable grouping to matrix operations. With input scaling and application

to a sequence of L queries at once:

$$f_{\mathrm{SDPA}}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathsf{T}}}{\sqrt{d_k}}\right) \tag{3}$$

scaled dot-product attention forms the basic attention operation. Attention is, due to its definition (3), agnostic to element permutation. To supply element ordering information a solution is to add positional embeddings to the input E. We use fixed sinusoids with position dependent frequency $u_{lk} = \sin l$ according to [23].

Multi-headed self-attention. The relational representation of different positions of a single sequence is obtained via $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{E}$ and called self-attention or intra-attention. It originates in NLP and has been successfully applied in tasks such as reading comprehension [5]. Multi-headed self-attention splits the attention operation along the model dimension into N_{H} attention heads applied in parallel to linear projections of the input $\mathbf{E}$:

$$f_{\mathrm{MHSA}}(\mathbf{E};\mathbf{W}) = [\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_{N_{\mathrm{H}}}]\,\mathbf{W}_{\mathrm{O}} \tag{4}$$

with head context $\mathbf{C}_i = f_{\mathrm{SDPA}}(\mathbf{E}\mathbf{W}_{\mathbf{Q}i}, \mathbf{E}\mathbf{W}_{\mathbf{K}i}, \mathbf{E}\mathbf{W}_{\mathbf{V}i})$, projection weights $\mathbf{W}_{\mathbf{Q}i}$, $\mathbf{W}_{\mathbf{K}i} \in \mathbb{R}^{M\times Q}$,$\mathbf{W}_{\mathbf{V}i} \in \mathbb{R}^{M\times V}$ and $\mathbf{W}_{\mathrm{O}} \in \mathbb{R}^{N_{\mathrm{H}}V\times M}$ reshaping inputs from and the output to the same model dimension M.

Architecture. According Fig. 2 right the sequence encoder stack is comprised of N_{s}-layers (i.e., parameters are duplicated) of f_{MHSA} followed by a 2-layer fully connected feed-forward net. This net $f_{\mathrm{EC}}(\mathbf{z}_l) = \max(0, \mathbf{z}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_1$ performs expansion and compression for each encoded sequence element. After the last encoder layer class probabilities are obtained with the same linear-softmax projection for each sequence element. Residual connections [8] facilitate model convergence during parameter optimization and layer normalization [2] avoids overfitting.

4 Experiments

We conduct experiments with the public Sleep-EDF Database Expanded sleep cassette study (SEDF-78) from Physionet [7] and the Montreal Archive of Sleep Studies subset 3 (MASS-SS3). In the following, we compare SAM with existing approaches on the single EEG sleep scoring task.

4.1 Datasets

Table 1 provides an overview of both datasets. Both cover healthy subjects and span a total age range from 20 to 101 years. While MASS-SS3 [17] is scored with the actual 5 class AASM standard, SEDF-78 [11] is scored with the older Rechtschaffen and Kales (R&K) standard. R&K has 6 sleep stages, but it is possible to merge the 2 stages S3 and S4 for close resemblance of N3. Thus, we evaluated both datasets with $V = \{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{R}\}$.

Table 1. Datasets.

Dataset	SEDF-78	MASS-SS3
n	78	62
F:M	41:37	34:28
mean age (range)	59.0 yrs (25-101)	42.5 yrs (20-69)
sleep disorders	none	AHI < 10
scoring standard	R&K	AASM
epoch duration	30 s	30 s
records	153	62
derivation	Fpz-Cz	EOG-F4
sampling rate	100 Hz	256 Hz

4.2 Preparation

MASS signals were resampled from 256 Hz to the model sampling rate $f_\mathrm{S} = 100$ Hz with a polyphase filter with up conversion 25 and down conversion 64 was used. The $C = 1$ input channels were bipolar derivations, with derivation Fpz-Cz for SEDF-78 and EOG-F4 for MASS-SS3. Both datasets were scored with epoch duration T= 30 s. To remove drifts and low frequency artifacts the data were filtered with a forward-backward, i.e., zero phase, Butterworth high pass filter of 5-th order with 0.1 Hz cutoff. According to this specification the l-th input epoch is $\mathbf{X}_l \in \mathbb{R}^{3000\times 1}$.

4.3 Model setup

According to the experimental EEG size of 3000 samples we chose 9 max pooling layers followed by final max pooling in the embedder. The resulting 20 CNN layers had kernel size 5. This contrasts other work, such as the successful DeepSleepNet with kernels up to size 400 [21]. After the first convolution feature maps increased super-linear from 8 to $N_\mathrm{F} = 64$. A complete specification is provided in the appendix Table 3. The sequence encoder layer was specified with model dimension $M = 64$, $N_\mathrm{H} = 4$ attention heads and feed forward expansion 200. We stacked $N_\mathrm{s} = 5$ identical layers (with different parameters) and chose an attention scope $L = 30$. This resulted in a total number of 360 k parameters. For SEDF-78 we reimplemented the DeepSleepNet [21] LSTM model as a baseline. Notably, this model has 22 M parameters.

4.4 Training

While some work applies separated and subsequent training of embedder and sequence encoder [21], we train SAM jointly, via cross-entropy loss:

$$\mathcal{L}\left(\theta; \boldsymbol{X}, \mathbf{y}\right) = -\frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K}\delta_{y_l k}\log g\left(\boldsymbol{X};\theta\right)_{lk} \tag{5}$$

with the label y_l one hot coded and the model 3. Unlike many other solutions we use uniform class weights. We optimize (3) with AdamW [15] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, a batch size of 32 and weight decay of 10^{-4}. The learning rate was set to 10^{-3} and followed a fixed schedule with 4 epochs ramp up, 4 epochs ramp down and a total duration of 20 epochs. Since no validation set was used, we always used the final model for testing. During training we use 10% dropout probability after MHSA and expansion-compression. The model is implemented in PyTorch with a proprietary data loader. Training and testing were performed on a 64 GB workstation with an NVIDIA 3090 GPU with 24 GB RAM. The complete CV protocol on SEDF-78 required 1.7 h for SAM, while the reimplemented DeepSleepNet required 12.2 h.

4.5 Protocol

Both datasets were evaluated with k-fold cross-validation (CV) with randomly split *subjects*. In accordance with literature we chose $k = 10$ for SEDF-78 and $k = 31$ for MASS-SS3. In all training and test runs, a single subject was treated atomic, i.e., a single subject's data, on record or epoch level, was never split over test and training. This avoided over-fitting due to correlated test- and training data. We randomly chose 4 folds (2 per dataset) to find optimal hyperparameters from a set of combinations.

We report Cohen's κ, specific F1-scores and the macro F1-score (MF1) from pooled subjects. Thus, every result is from a single confusion matrix. In comparison with our baseline LSTM, we can compare results at the subject level. We use a paired-sample t-test to show statistical significance at the $\alpha = 0.05$ level. Note that we do not compare to work that does not treat subjects atomic, uses a different k in CV, does not report comparable metrics or uses different electrode derivations.

Table 2. Results for the proposed SAM and the LSTM baseline DeepSleepNet compared to literature. Our work is indicated by *. The best results are boldfaced.

		Overall scores		Sleep stage F1-scores				
Dataset	Model	κ	MF1	W	N1	N2	N3	R
MASS-SS3	SAM* (0.36 M)	**0.80**	**82**%	87%	56%	**91**%	85%	88%
EOGL-F4	DeepSleepNet*	**0.80**	**82**%	**88**%	58%	**91**%	84%	88%
	DeepSleepNet [21] (22 M)	**0.80**	**82**%	87%	**60**%	90%	82%	**89**%
	IIT [20]	0.79	81%	85%	54%	**91**%	**87**%	85%
SEDF-78	SAM*	**0.78**	**79**%	**93**%	49%	**86**%	**82**%	**84**%
Fpz-Cz	DeepSleepNet*	0.76	77%	92%	48%	84%	80%	79%
	CNN-LSTM [12]	0.77	-	-	-	-	-	-
	U-Time [18]	0.75	76%	92%	**51**%	83%	75%	80%

5 Results and discussion

Table 2 shows overall agreement scores and sleep stage specific agreement for all datasets. On MASS-SS3 SAM achieves $\kappa = 0.80$ and MF1 = 82%, which is on par with the best reported result. While our model is less accurate in N1 it is more accurate in the clinically important N3 stage. The reimplemented LSTM baseline DeepSleep-Net achieves comparable results to the published version. On SEDF-78 our results surpass the DeepSleepNet and the literature with $\kappa = 0.78$ and MF1 = 79%. While all sleep stages except N1 improve between 1% and 2%, the largest improvement occurs in R with 4%.

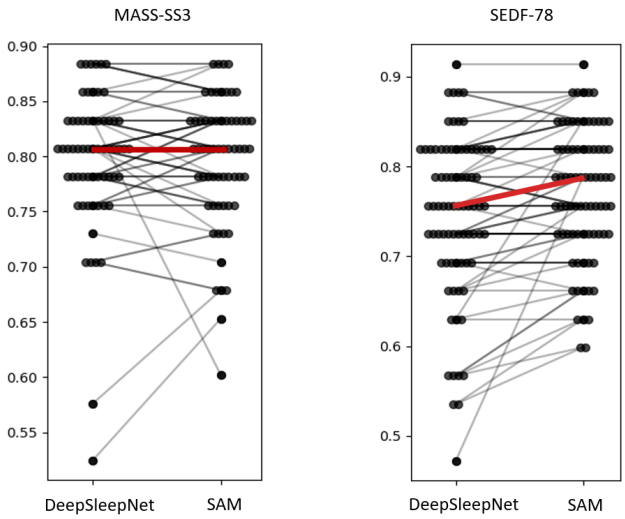

Fig. 3. Paired samples plot of binned subject-wise Cohen's κ for the proposed model SAM and the baseline LSTM DeepSleepNet for datasets MASS-SS3 and SEDF-78, with binned average in red.

Figure 3 shows paired samples plots of binned subject level κ. For MASS-SS3 there is no difference on average (horizontal red line), analogous to the pooled result (cf. Table 2). The paired-sample t-test confirms this expectation statistically with $p = 0.7 > \alpha$. For SEDF-78 most subjects show an increase by at least one bin. Also, on average, there is an increase of 1 bin, which has a size of 0.025. The t-test confirms a statistically significant difference with $p = 4 \times 10^{-5} < \alpha$.

Our reimplemented DeepSleepNet* achieves higher agreement on SEDF-78 (MF1 = 0.77) than U-Time, although their DeepSleepNet reimplementation (MF1 = 0.73) was inferior to U-Time. We carefully reimplemented the original layer details (e.g., CNN padding) and adapted learning rates for the single, joint training session. In accordance with comparable published and reimplemented

results on MASS-SS3 we assume that our higher baseline result on SEDF-78 is representative. Results show twice as large database-differences for DeepSleepNet ($\Delta\kappa = 0.04$) than for SAM ($\Delta\kappa = 0.02$), caused by two different reasons. First, we consider MASS-SS3 a simpler problem than SEDF-78 (EEG only), since the former provides the network with EEG and EOG—the most important information human experts use to score sleep—reflected by higher scores on MASS-SS3. Second, DeepSleepNet results are particularly high on MASS-SS3 compared to SEDF (both to SEDF-78 here and SEDF-20 in [21]). According the authors the DeepSleepNet architecture was optimized only on a MASS-SS3 subset, which may cause an architectural bias towards MASS-SS3.

While these are promising results, we acknowledge that they should be solidified with more subjects from an independent dataset. A question out of this work's scope are clinical benefits of the proposed method. Since raw agreement has no direct clinical relevance, improvements must be interpreted cautiously. However, since stage R is clinically important (e.g., for time-to-REM) our improvements could be relevant. Considered that SAM is more than 50 times smaller than DeepSleepNet it's parity on MASS-SS3 is remarkable. On the harder EEG only SEDF-78 experiment SAM could achieve a considerable REM improvement (+4% F1-R), albeit the small size. The results support the introductory hypothesis that REM accuracy benefits most from attention. We attribute this to the direct (i.e., not state encoded) access to distant embeddings and the distant-indifferent (i.e., constant) maximum path length of attention. Although we conceived SAM for sleep scoring, other tasks such as abnormality detection or movement intention detection as well may fit as well.

6 Conclusion

This contribution introduced SAM, a simple, and lightweight EEG local attention model. We showed that SAM with 360 k parameters is on par with the 22 M parameter state-of-the-art on the MASS-SS3 PSG sleep scoring task. On the harder SEDF-78 EEG sleep scoring task SAM achieves the new state-of- the-art performance and proves long-term dependency modelling benefits with a considerable improvement in the practically important REM sleep stage. On top of its effectiveness SAM is also efficient on EEG of arbitrary length since computational complexity scales linear with EEG length. SAM may be an important step towards reduced-instrumentation, but PSG-quality sleep scoring and we look forward to investigations in clinical use.

Appendix

Table 3 shows the embedder layer specification.

Table 3. Embedder layer specification based on CNN blocks (FCi) and residual blocks (FRi). Conv BN layers comprise CNN, batch norm and ReLU activation and are specified by kernel size, feature maps C_o, stride and padding.

Layer				Conv./Pooling			
ID	Group	Type	Out dim	Size	C_O	Stride	Padd.
1	FC1	Input	3000 x 1				
2		Conv BN	3000 x 8	5	8	1	2
3	FR2	Conv BN	3000 x 18	5	18	1	2
4		Max Pool	1500 x 18	2		2	0
5		Conv BN	1500 x 18	5	18	1	2
6		Conv BN	1500 x 18	5	18	1	2
7	FC3	Conv BN	1500 x 21	5	21	1	2
8		Max Pool	750 x 21	2		2	0
9	FR4	Conv BN	750 x 25	5	25	1	2
10		Max Pool	375 x 25	2		2	0
11		Conv BN	375 x 25	5	25	1	2
12		Conv BN	375 x 25	5	25	1	2
13	FC5	Conv BN	375 x 29	5	29	1	2
14		Max Pool	187 x 29	2		2	0
15	FR6	Conv BN	187 x 34	5	34	1	2
16		Max Pool	93 x 34	2		2	0
17		Conv BN	93 x 34	5	34	1	2
18		Conv BN	93 x 34	5	34	1	2
19	FC7	Conv BN	93 x 40	5	40	1	2
20		Max Pool	46 x 40	2		2	0
21	FR8	Conv BN	46 x 47	5	47	1	2
22		Max Pool	23 x 47	2		2	0
23		Conv BN	23 x 47	5	47	1	2
24		Conv BN	23 x 47	5	47	1	2
25	FC9	Conv BN	23 x 54	5	54	1	2
26		Max Pool	11 x 54	2		2	0
27	FR10	Conv BN	11 x 64	5	64	1	2
28		Max Pool	5 x 64	2		2	0
29		Conv BN	5 x 64	5	64	1	2
30		Conv BN	5 x 64	5	64	1	2
31	F11	Max Pool	1 x 64	5		5	0
32		Flatten	64				
33	C	Linear	5				

Acknowledgments

Asan Agibetov helped with LaTeX help and Kluge Tilmann provided computing infrastructure. This work was supported by the Austrian Research Promotion Agency (FFG) grant number 867615.

References

1. Anderer, P., Gruber, G., Parapatics, S., Woertz, M., Miazhynskaia, T., Klösch, G., Saletu, B., Zeitlhofer, J., Barbanoj, M.J., Danker-Hopfe, H., et al.: An e-health solution for automatic sleep classification according to rechtschaffen and kales: Validation study of the somnolyzer 24 x 7 utilizing the siesta database. Neuropsychobiology **51**(3), 115–133 (2005). https://doi.org/10.1159/000085205
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (Jul 2016)
3. Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings p. 1–15 (2015)
4. Berry, R.B., Brooks, R., Gamaldo, C.E., Harding, S.M., Marcus, C., Vaughn, B.V.: The aasm manual for the scoring of sleep and associated events. Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine **176**, 2012 (2012)
5. Cheng, J., Dong, L., Lapata, M.: Long short-term memory-networks for machine reading. EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings p. 551–561 (2016). https://doi.org/10.18653/v1/d16-1053
6. Dong, H., Supratak, A., Pan, W., Wu, C., Matthews, P.M., Guo, Y.: Mixed neural network approach for temporal sleep stage classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering **26**(2), 324–333 (2018)
7. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet. Circulation **101**(23) (Jun 2000)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2016-Decem**, 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (Nov 1997)
10. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. PMLR (2015)
11. Kemp, B., Zwinderman, A.H., Tuk, B., Kamphuisen, H.A., Oberyé, J.J.: Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the eeg. IEEE Transactions on Biomedical Engineering **47**(9), 1185–1194 (2000). https://doi.org/10.1109/10.867928
12. Korkalainen, H., Aakko, J., Nikkonen, S., Kainulainen, S., Leino, A., Duce, B., Afara, I.O., Myllymaa, S., Töyräs, J., Leppänen, T.: Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. IEEE journal of biomedical and health informatics **24**(7), 2073–2081 (2019)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)

14. Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T.: On the compactness, efficiency, and representation of 3d convolutional networks: Brain parcellation as a pretext task. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 10265 LNCS, p. 348–360 (2017)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
16. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing p. 1412–1421 (2015)
17. O'Reilly, C., Gosselin, N., Carrier, J., Nielsen, T.: Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research. Journal of Sleep Research **23**(6), 628–635 (2014)
18. Perslev, M., Jensen, M.H., Darkner, S., Jennum, P.J., Igel, C.: U-time: a fully convolutional network for time series segmentation applied to sleep staging. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. pp. 4415–4426 (2019)
19. Phan, H., Andreotti, F., Cooray, N., Chen, O.Y., De Vos, M.: Seqsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Transactions on Neural Systems and Rehabilitation Engineering **27**(3), 400–410 (2019)
20. Seo, H., Back, S., Lee, S., Park, D., Kim, T., Lee, K.: Intra- and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg. Biomedical Signal Processing and Control **61**, 102037 (2020)
21. Supratak, A., Dong, H., Wu, C., Guo, Y.: Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. IEEE Transactions on Neural Systems and Rehabilitation Engineering **25**(11), 1998–2008 (2017)
22. Tsinalis, O., Matthews, P.M., Guo, Y., Zafeiriou, S.: Automatic sleep stage scoring with single-channel eeg using convolutional neural networks. arXiv preprint arXiv:1610.01683 (Oct 2016), `http://arxiv.org/abs/1610.01683`
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)