Teacher Assistant-Based Knowledge Distillation Extracting Multi-level Features on Single Channel Sleep EEG

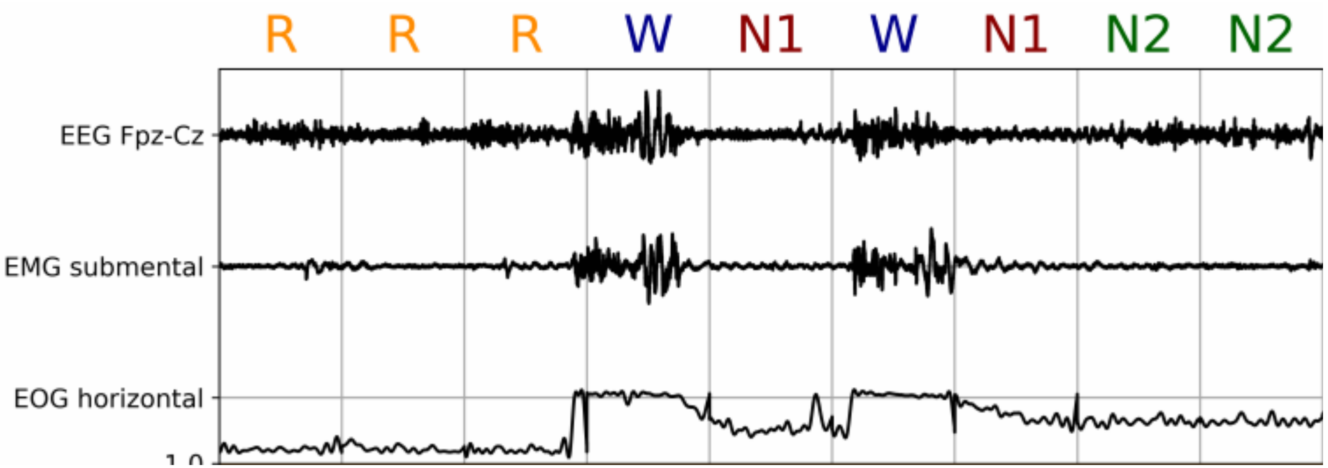
Heng Liang*, Yucheng Liu*, Haichao Wang†, Ziyu Jia*

*Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China

†Tsinghua-Berkeley Shenzhen Institute, Shenzhen, China

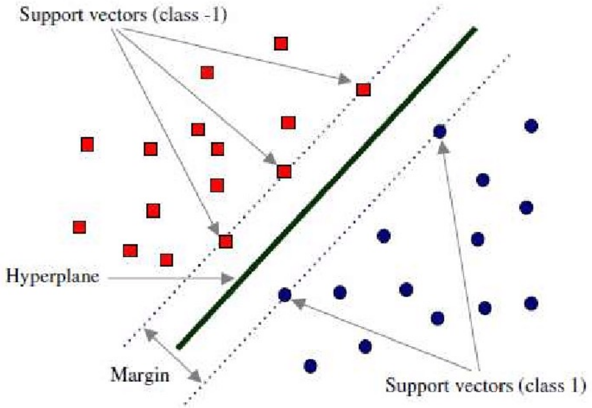
Sleep stage classification

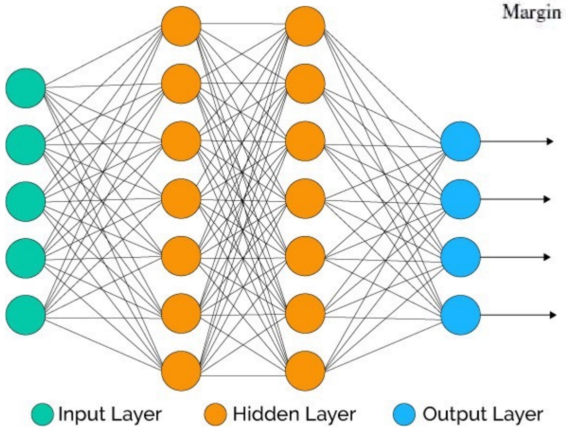
- **Background:**
  - The American Academy of Sleep Medicine classifies sleep into five main stages: W, N1, N2, N3, and REM.
  - The patient's 8-hour sleep data is processed and analyzed every 30s which is usually the length of a epoch to give the classification judgment results by the physician.
- **Importance:**
  - help doctors correctly diagnose narcolepsy, snoring, Alzheimer's, diabetes, depression, and other diseases.



Introduction

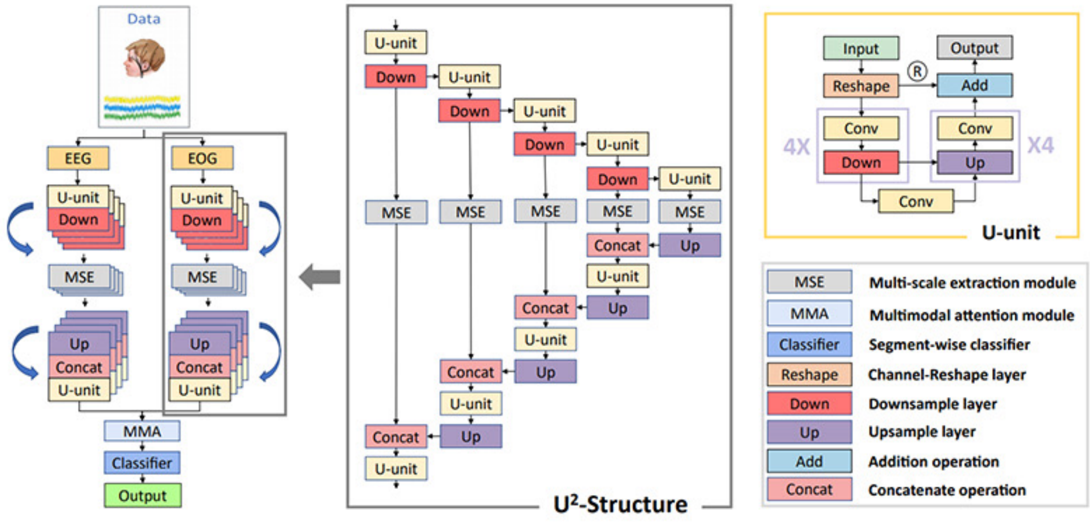
- **Manual extraction:**
  - Manual extraction of data features and analysis by medical experts.
  - High labor cost, Time-consuming and subjective results.
- **Machine learning:**
  - Fourier Transform, SVM, Riemannian geometry.
  - Require prior knowledge, Normal accuracy.
- **Deep learning:**
  - CNN，RNN，Transformer
  - No pre-processed data required, High accuracy.

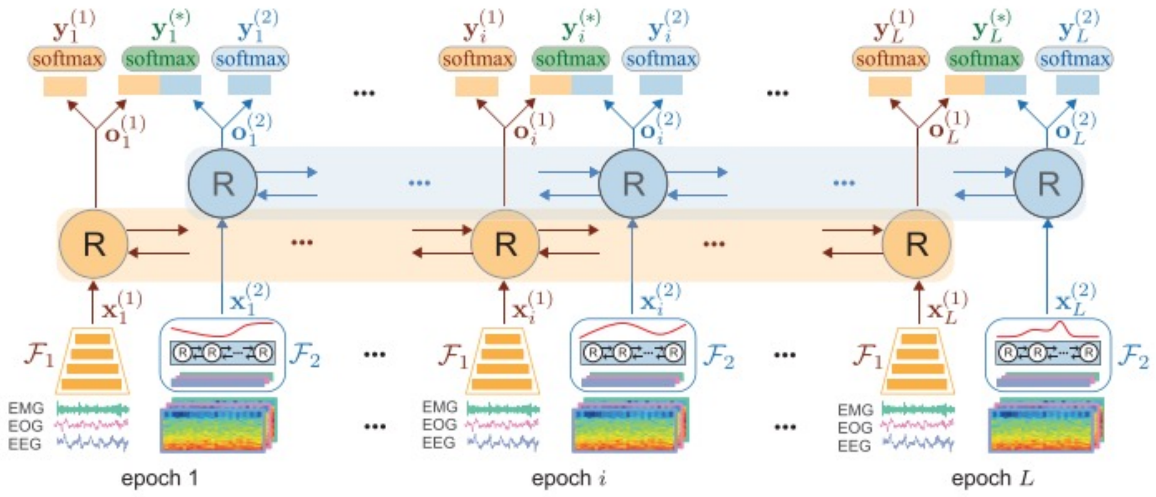




Related Work

- **Two typical deep learning architectures that are widely used：**
  - CNN-based: SalientSleepNet[1], MMCNN[2]
  - Hybrid architecture of CNN and RNN: DeepSleepNet[3], XsleepNet[4]





Related Work

Difficulty in applying existing deep learning models on wearable devices:

- Platform performance degradation and poor user experience.
- High computational cost.
- Large number of parameters, long training time.

Method	Parameters
SalintSleepNet[1]	$0.9 * 10^6$
DeepSleepNet[3]	$2.1 * 10^7$
XsleepNet[4]	$5.6 * 10^6$
TinySleepNet[7]	$1.3 * 10^6$
SleepEEGNet[8]	$2.6 * 10^6$

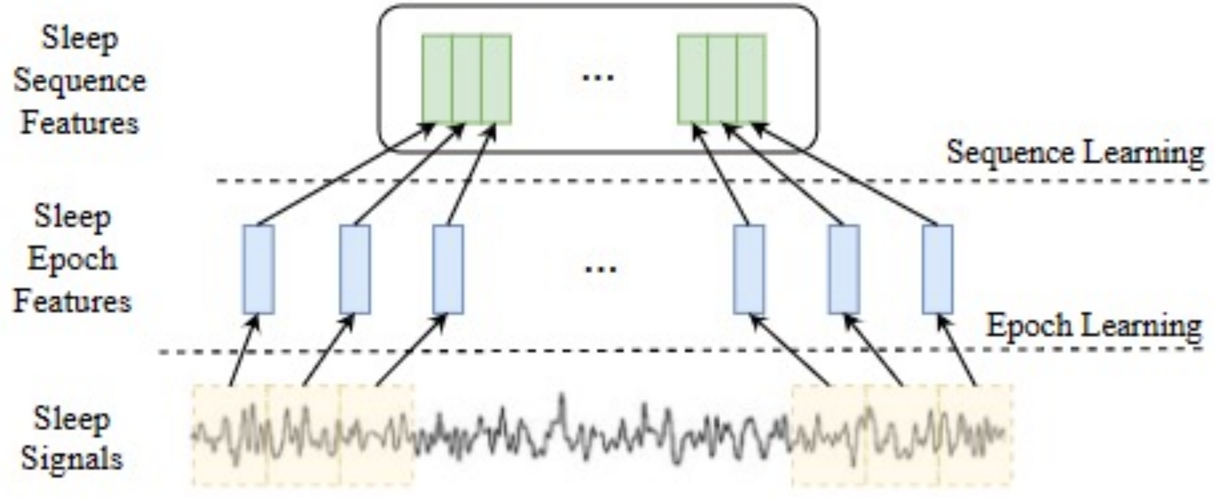
Model lightweight based on knowledge distillation

Motivation 1

There are two kinds of important features in the sleep signals:

- **Epoch-level** features: local characteristics of a single sleep epoch. For example, the N2 stage includes mainly sleep spindles and K complexes.
- **Sequence-level** features: transition rules between multiple sleep epochs. For instance, the N1 stage often serves as a transition stage between the W stage and other stages.

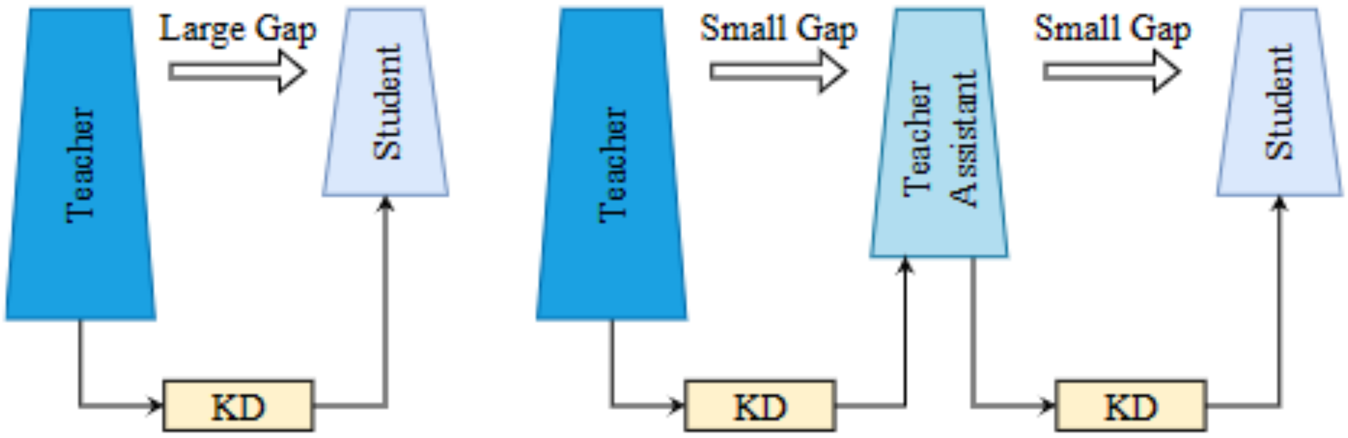
How can better transfer these two types of knowledge?



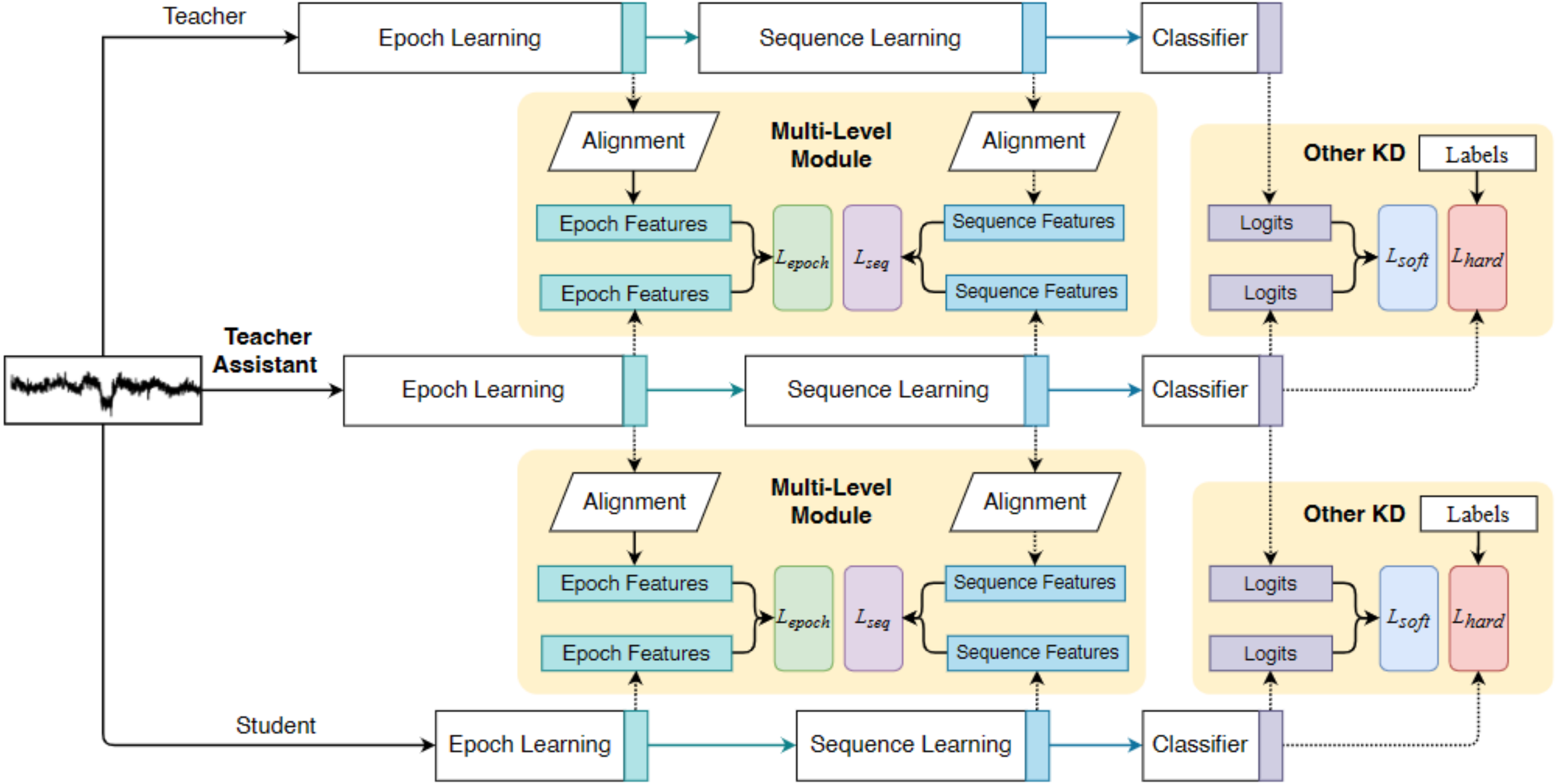
Motivation 2

In most cases, the teacher network is deep while the student network is shallow, which leads to excessive gap and the knowledge may be transferred inefficiently.

How to bridge the gap between the teacher and student model?



Methods：SleepKD



- ## 1. Multi-Level Module

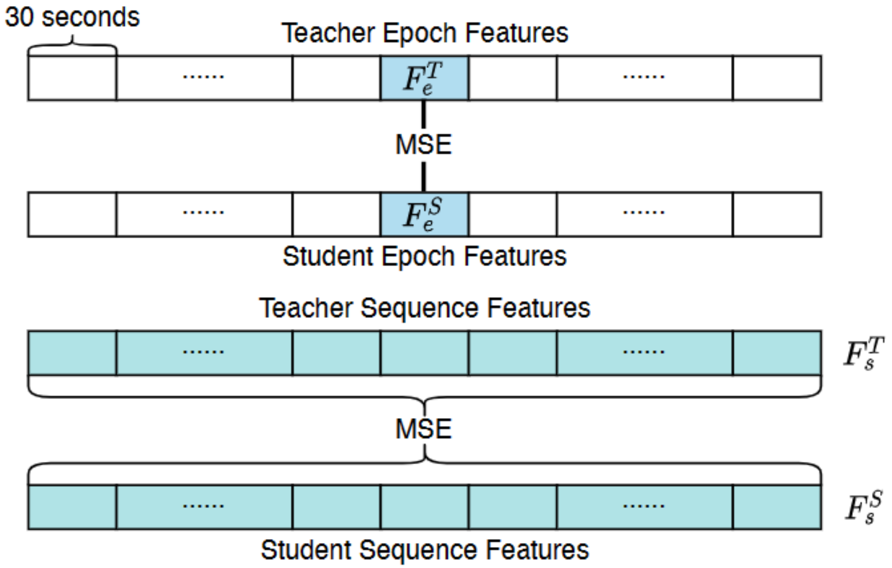
Mainly capture these two types of features:

- Epoch-level features:

$$\mathcal{L}_{epoch} = \mathcal{L}_{MSE}\left(\Phi(\boldsymbol{F}_e^T), \boldsymbol{F}_e^S\right)$$

- Sequence-level features:

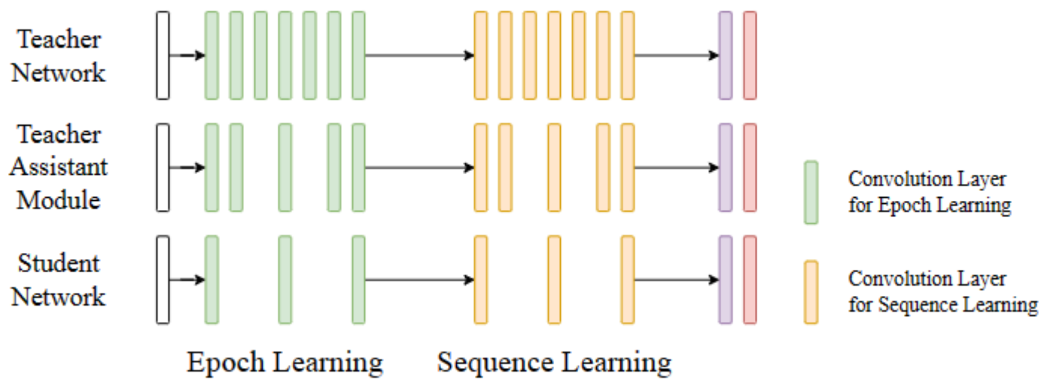
$$\mathcal{L}_{seq} = \mathcal{L}_{MSE}\left(\Phi(\boldsymbol{F}_s^T), \boldsymbol{F}_s^S\right)$$



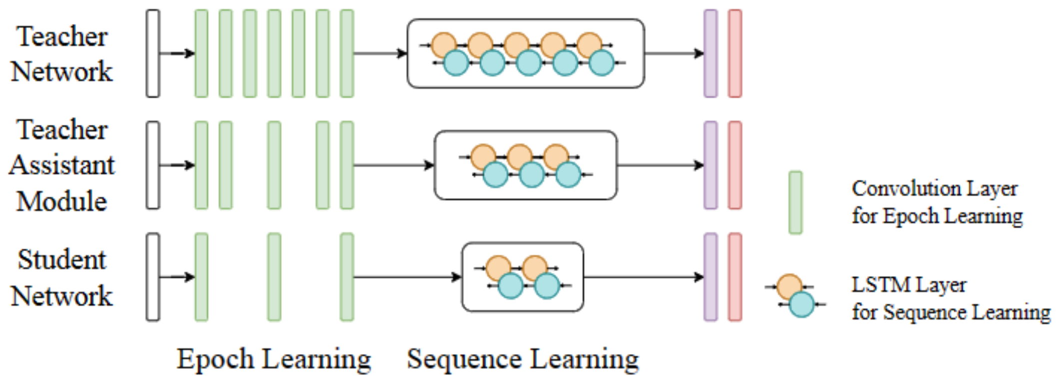
Methods

• 2. Teacher Assistant Module (TA Module)

- Traditional distillation: knowledge transfer is hindered when teacher and student network are too much different.
- TA Module: bridges the gap between teacher and student models and improves knowledge transfer



CNN-based architecture



Hybrid architecture of CNN and RNN

- **3. Other Knowledge Distillation Module**

  - **Soft label**: the probability distribution for each stage from the teacher model output.

$$\mathcal{L}_{soft} = D_{KL}\left(\boldsymbol{p}^{T} \| \boldsymbol{p}^{S}\right)$$

  - **Hard label**: One-hot encoded true labels in the original dataset.

$$\mathcal{L}_{hard} = \mathcal{L}_{CE}\left(\boldsymbol{y}, \boldsymbol{p}^{S}\right)$$

- **4. Module Integration**

$$\mathcal{L}_{Total} = \alpha\mathcal{L}_{epoch} + \beta\mathcal{L}_{seq} + \gamma\mathcal{L}_{soft} + \delta\mathcal{L}_{hard}$$

Experiments

- **SleepEDF[9]：**
  - The dataset contains the PSG data samples from 20 subjects (10 for males and 10 for females) in 2 days.
  - These recordings were manually classified into eight classes by sleep experts according to the R&K standard.
  - We merge the N3 and N4 stage into a single N3 stage according to the AASM manual.

- **ISRUC-III[10]：**
  - The dataset contains the PSG data samples from 10 subjects (1 for males and 9 for females) for a whole night in 8 hours.
  - The annotations of this dataset are scored by two professional experts.

Experiments

Baseline Methods：

- **Hinton-KD[11]**:Propose a simple way to improve the performance by distilling the knowledge of the complex model into a compact model with the output of the former.
- **Fitnets[12]:**Extend the idea of the traditional knowledge distillation by using both the output of the teacher network and the intermediate representation as a hint to the student.
- **NST[13]:**Implement a knowledge transfer loss function by minimizing the Maximum Mean Discrepancy between the feature map of the sophisticated model and the slimming model.
- **TAKD[14]:**Introduce a multi-step knowledge distillation by using teacher assistant (TA) whose size is between the teacher and student model.
- **DGKD[15]:**Devise the densely-guided knowledge method using multiple teacher assistant to fill the large gap between teacher and student model gradually.
- **DKD[16]:**Reformulate the classical KD method with non-target class knowledge distillation (NCKD) and target class knowledge distillation (TCKD).

Experiments

The comparison of the knowledge distillation baselines：

SalientSleepNet

Method	ISRUC-III		Sleep-EDF	
	Acc	F1-Score	Acc	F1-Score
KD	74.65	73.74	83.62	78.93
Fitnets	75.00	73.33	85.33	80.21
NST	75.68	75.46	83.67	77.85
TAKD	77.27	76.19	85.57	80.74
DGKD	76.70	73.68	85.19	78.86
DKD	76.70	73.73	84.64	78.96
SleepKD	**79.66**	**78.57**	**87.05**	**81.40**

DeepSleepNet

Method	ISRUC-III		Sleep-EDF	
	Acc	F1-Score	Acc	F1-Score
KD	80.22	74.54	81.28	64.41
Fitnets	81.11	75.05	80.59	65.83
NST	81.59	76.48	84.71	68.53
TAKD	81.59	76.46	83.97	67.87
DGKD	81.36	75.75	84.47	68.46
DKD	79.88	75.37	83.88	67.78
SleepKD	**83.29**	**77.29**	**85.66**	**69.46**

Experiments

Compression performance of SleepKD-based student model:

- The inference speed is effectively improved while maintaining performance.
- The memory usage and the number of parameters are significantly reduced.

SalientSleepNet

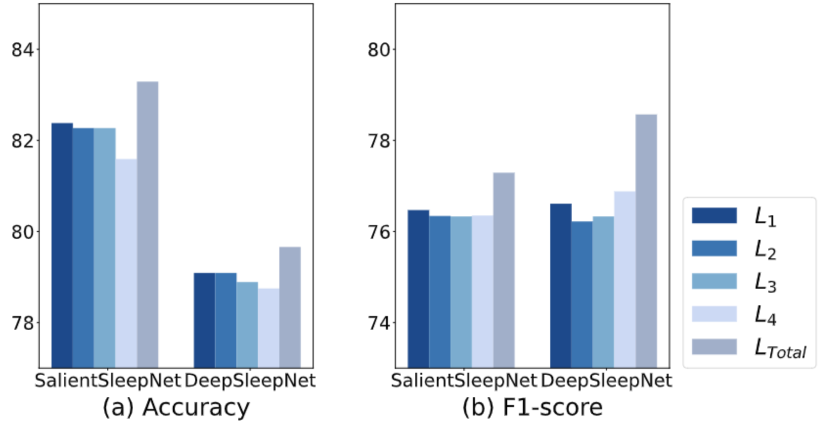
Metric	Teacher	Student
Accuracy	80.34%	79.66%
Memory	632.88MB	160.24MB
Parameters	474,662	120,181
Compression Ratio	74.68%	
Acceleration	6.85x	

DeepSleepNet

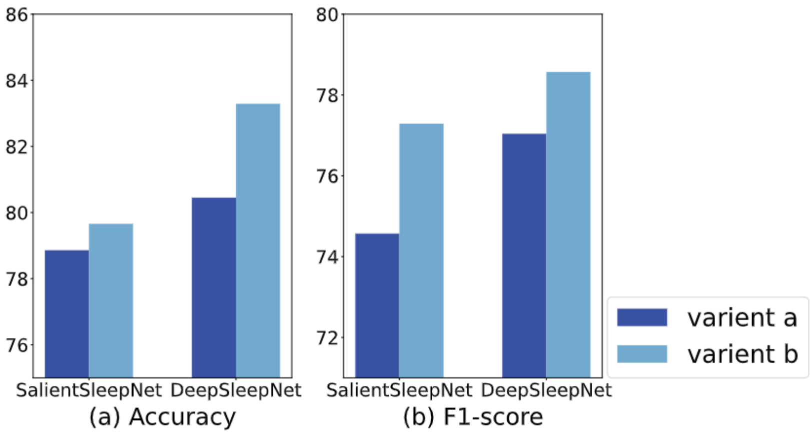
Metric	Teacher	Student
Accuracy	83.97%	83.29%
Memory	21.46MB	6.04MB
Parameters	5,502,474	1,552,906
Compression Ratio	71.78%	
Acceleration	5.59x	

Ablation Experiments

- **Ablation settings of each Loss term:**
  - $\mathcal{L}_1 = \mathcal{L}_{Total} - \mathcal{L}_{seq}$
  - $\mathcal{L}_2 = \mathcal{L}_{Total} - \mathcal{L}_{epoch}$
  - $\mathcal{L}_3 = \mathcal{L}_{Total} - \mathcal{L}_{soft}$
  - $\mathcal{L}_4 = \mathcal{L}_{Total} - \mathcal{L}_{hard}$



- **Ablation settings of the TA module:**
  - *varient a)*: Multi-Level Module
  - *varient b)*: Multi-Level Module + TA Module



Conclusion

Main contributions:

- We employ knowledge distillation on the multi-level sleep stage classification model for the first time and design the Multi-Level Module. It better transfers the features of single sleep stages and transition rules between multiple sleep stages.

- We design corresponding TA modules for different architectures. This can bridge the excessive gap between teacher and student network and the experiments show that SleepKD achieves excellent results on two popular architectures.

- SleepKD achieves state-of-the-art distillation performance compared to other distillation methods. In addition, we apply it to the transformer network and obtain state-of-the-art results.

Reference

中国科学院自动化研究所
INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES
TBSI 清华-伯克利深圳学院 Tsinghua-Berkeley Shenzhen Institute

[1] Jia Z, Lin Y, Wang J, et al. Salientsleepnet: Multimodal salient wave detection network for sleep staging[J]. arXiv preprint arXiv:2105.13864, 2021.

[2] Chambon S, Galtier M N, Arnal P J, et al. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2018, 26(4): 758-769.

[3] Supratak A, Dong H, Wu C, et al. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2017, 25(11): 1998-2008.

[4] Phan H, Chén O Y, Tran M C, et al. XSleepNet: Multi-view sequential model for automatic sleep staging[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5903-5915.

[5] Jia Z, Lin Y, Wang J, et al. GraphSleepNet: Adaptive Spatial-Temporal Graph Convolutional Networks for Sleep Stage Classification[C]//IJCAI. 2020: 1324-1330.

[6] Phan H, Mikkelsen K, Chén O Y, et al. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification[J]. IEEE Transactions on Biomedical Engineering, 2022, 69(8): 2456-2467.

[7] Supratak A, Guo Y. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG[C]//2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2020: 641-644.

[8] Mousavi S, Afghah F, Acharya U R. SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach[J]. PloS one, 2019, 14(5): e0216456.

[9] Imtiaz S A, Rodriguez-Villegas E. An open-source toolbox for standardized use of PhysioNet Sleep EDF Expanded Database[C]//2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2015: 6014-6017.

[10] Khalighi S, Sousa T, Santos J M, et al. ISRUC-Sleep: A comprehensive public dataset for sleep researchers[J]. Computer methods and programs in biomedicine, 2016, 124: 180-192.

[11] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[12] Romero A, Ballas N, Kahou S E, et al. Fitnets: Hints for thin deep nets[J]. arXiv preprint arXiv:1412.6550, 2014.

[13] Huang Z, Wang N. Like what you like: Knowledge distill via neuron selectivity transfer[J]. arXiv preprint arXiv:1707.01219, 2017.

[14] Mirzadeh S I, Farajtabar M, Li A, et al. Improved knowledge distillation via teacher assistant[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(04): 5191-5198.

[15] Son W, Na J, Choi J, et al. Densely guided knowledge distillation using multiple teacher assistants[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 9395-9404.

[16] Zhao B, Cui Q, Song R, et al. Decoupled knowledge distillation[C]//Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2022: 11953-11962.

Thanks!