IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Analysis of Transformer's Attention Behavior in Sleep Stage Classification and Limiting It to Improve Performance

Dongyoung Kim[1], Young-Woong Ko[1], Dong-Kyu Kim[2], and Jeong-Gun Lee[1], (Member, IEEE)

[1]Department of Computer Engineering, Hallym University, Chuncheon, South Korea. (email: kimdongyoung0218@hallym.ac.kr; jeonggun.lee@hallym.ac.kr)

[2]Department of Otorhinolaryngology-Head and Neck Surgery, Chuncheon Sacred Heart Hospital and Hallym University College of Medicine, Chuncheon, South Korea. (email: doctordk@naver.com)

Corresponding author: Jeong-Gun Lee (e-mail: jeonggun.lee@hallym.ac.kr) and Dong-Kyu Kim (e-mail: doctordk@naver.com).

"The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002)."

ABSTRACT The transformer architecture has been focused on many tasks like natural language processes, vision tasks and etc. The most important and general requirement of using the transformer-based architecture is that the model must be trained on a large-scale dataset before it can be fine-tuned for a specific task like classification, object detection and etc. However, in this paper, we find that the transformer architecture has better generalization capability to capture the features from data samples for sleep stage classification than CNN-based architectures, despite using a small-scale dataset without pretraining on large-scale dataset. This outcome contradicts the widely-held belief that a transformer architecture is more effective when trained on large datasets. In this paper, we investigate the attention behavior of a transformer model and demonstrate how global and local attentions influence an attention map in a transformer architecture. Finally, through experiments, we show that restricting global attention using *'**M**asked **M**ulti-**H**ead **S**elf-**A**ttention* (**M-MHSA**)' can lead to improved model generalization in sleep stage classification compared with the previous methodologies and original transformer-based architecture on three different datasets.

INDEX TERMS Sleep Stage Classification, Deep Learning, Transformer

I. INTRODUCTION

RECENTLY deep learning approaches have been focused on many tasks such as vision tasks, medical fields, natural language processing (NLP) and etc. In particular, convolutional neural networks (CNNs) have been proven highly effective for many computer vision tasks such as classification, object detection and segmentation tasks. Contrary to CNNs, recurrent neural networks (RNNs) have shown meaningful performance on NLP tasks. More recently, a transformer architecture has gained popularity in various tasks, particularly in NLP tasks. Transformer-based architectures, such as the transformer [1] and BERT [2] models, have been used successfully in these tasks due to their ability to capture both of local and global features through the use of multi-head self-attention (MHSA) which enables them to understand the relationships among the words in a sentence, unlike conventional CNNs and RNNs. Similarly, in computer vision tasks, the transformer-based architectures such as the Vision Transformer (ViT) [3] have been used for image classification tasks by using the attention mechanism to capture image features among the different patches of an image.

Many studies in sleep stage classification have used a CNN-based model architecture to predict sleep stages from an epoch[1] of polysomnography (PSG) using mainly signal channels such as Electroencephalogram (EEG), Electrooculography (EOG) and Electromyography (EMG) [4], [5]. Although some recent studies have used transformer-based architectures, they have received less attention than conventional CNN-based architectures [6], [7].

[1] A 30-second interval of PSG data is referred to as an '*epoch*', and sleep stage classifications are made for each epoch. It should not be confused with a '*train epoch*', which refers to the process of using all the data to train a model.

In general, transformer-based architectures require a large-scale dataset for pre-training before fine-tuning specific downstream tasks to achieve meaningful performance compared to conventional CNN-based architectures [8]. Therefore, we expect that transformer-based architectures will achieve lower performance compared to conventional CNN-based architectures when using small-scale training PSG records due to a lack of inductive bias [9]. However, the transformer-based architectures present higher performance compared to the CNN-based architectures even with small-scale training PSG records in sleep stage classification. This result contradicts the commonly accepted knowledge that transformer-based architectures require a large-scale training dataset for pre-training to perform better on specific tasks [8].

In this paper, first, we analyze the reason why transformer-based architecture can obtain the generalization capability compared to CNN-based architecture. Second, to have a deeper understanding of the attention behavior of a transformer model, we develop a novel approach to restrict the global attention in MHSA using a masking strategy applied to the attention map. Through comprehensive experimentation, we figure out that this restriction strategy can effectively enhance the model's generalization ability in sleep stage classification tasks, leading to performance improvements.

Our contributions can be summarized as follows:

1. In contrast to the well-accepted training guideline for a transformer architecture, we show that the transformer model performs better in sleep stage classification compared to conventional CNN-based architectures, even when using a small dataset without pre-training with an extra large-scale dataset.
2. We investigate the reason for the robustness of a transformer-based architecture and we evaluate the impact of local and global information by limiting the locality in the MHSA mechanism using an attention mask.
3. To enhance the generalization ability, we propose a novel approach to focus on the local information using **M**asked **M**ulti-**H**ead **S**elf-**A**ttention (**M-MHSA**). When applying the M-MHSA in transformer-based architecture, the model can improve performance.

II. RELATED WORKS

A. SLEEP STAGE CLASSIFICATION

In sleep stage classification, sleep experts determine the class of a sleep stage for a given epoch according to the Rechtschaffen and Kales (R&K) [10] or American Academy of Sleep Medicine (AASM) [11] criteria. In AASM criterion, the sleep stage was classified into 5 classes such as Wake, Non-REM1 (N1), Non-REM2 (N2), Non-REM3 (N3), and Rapid Eye Movement (REM). The sleep stage classification uses PSG data that consists of various biosensors information including EEG, EOG, EMG and etc. According to the AASM sleep standard, PSG records are usually segmented into 30 seconds to determine the specific sleep stages by sleep experts. To classify the sleep stage into 5 classes, sleep experts mainly utilize EEG, EOG and EMG channels.

B. CNN-BASED MODEL ARCHITECTURE

In numerous studies [12]–[15], researchers frequently employ CNN-based models to extract time-invariant features from an epoch data. In general, a CNN-based architecture is adopted well in capturing the 'temporal' information from the time-varying signal data. In DeepSleepNet [12], the authors utilize two convolutional layers of different sizes to extract distinct features. Typically, the larger CNN layer captures 'frequency' information from the signal (i.e., frequency components), while the smaller CNN layer focuses on capturing 'temporal' information (i.e., when certain of EEG patterns appear such as k-complex) [16]. Similarly, SleepEEGNet [13] utilizes the same CNN architecture as DeepSleepNet. However, they adopt an encoder-decoder architecture instead of employing an RNN. ResNet+LSTM [14] utilizes a ResNet architecture to extract features from the epochs of EEG signals. AttnSleep [15] employs attention mechanisms, including adaptive feature recalibration and causal convolutional multi-head attention, to extract meaningful features from epoch samples.

Typically, in order to achieve the generalized prediction performance with a small-scale training set, most of the approaches utilize CNN-based architectures, which have an inductive bias, instead of using transformer-based models [9].

C. TRANSFORMER-BASED MODEL ARCHITECTURE

The transformer is a widely-used deep learning model architecture designed specifically for processing sequential data such as NLP and time series analysis. It utilizes "self-attention" mechanisms, allowing the model to learn relationships between elements in the sequence without relying on recurrence or convolutions [1]. In recent years, in addition to sequential data processing, the transformer architecture has seen success in a variety of vision tasks.

It is worth noting that in general, transformer architectures are known to require a larger amount of data samples than CNN models in order to achieve comparable performance [1]–[3], [17]. This is because transformer-based models lack the inductive bias which is presented in the CNNs but the inductive bias can limit their ability to generalize to new data with fewer training samples [3], [8].

Sleep stage classification tasks, like vision tasks, also make use of attention mechanisms in transformer-based architectures. To extract inner and inter features from raw signals, two transformer-based encoder blocks are employed [6]. SleepTransformer [7] utilizes a transformer-based architecture to extract features using spectrogram images which are obtained from epoch samples by applying Fourier transform, and further it employs an entropy-based method to quantify uncertainty.

TABLE 1. Class distributions of SHHS public PSG dataset.

Dataset	# of PSG	# of epochs (samples)				
		Wake	N1	N2	N3	REM
Train	50	15221 (29.93%)	1929 (3.79%)	21569 (42.42%)	5409 (10.63%)	6713 (13.20%)
	100	28858 (28.49%)	3862 (3.81%)	42475 (41.93%)	12065 (11.91%)	14024 (13.84%)
	150	43664 (28.75%)	5669 (3.73%)	61714 (40.64%)	20029 (13.19%)	20762 (13.67%)
	200	58453 (29.05%)	7442 (3.69%)	82113 (40.81%)	25909 (12.87%)	27281 (13.55%)
	250	74942 (29.38%)	9367 (3.71%)	102566 (40.70%)	31652 (12.56%)	34327 (13.62%)
	300	90251 (29.77%)	11426 (3.76%)	122331 (40.35%)	38224 (12.6%)	40923 (13.49%)
	350	104151 (29.48%)	13103 (3.70%)	142490 (40.33%)	45352 (12.83%)	48177 (13.63%)
	400	118275 (29.31%)	14965 (3.70%)	165131 (40.93%)	50500 (12.51%)	54554 (13.52%)
Validation	832	244386 (29.0%)	30021 (3.6%)	342148 (40.6%)	109206 (12.9%)	117773 (13.9%)
Test	832	243759 (29.1%)	32696 (3.9%)	341805 (40.7%)	102379 (12.2%)	118150 (14.1%)

TABLE 2. Class distributions of Institution-A PSG dataset.

Dataset	# of PSG	# of epochs (samples)				
		Wake	N1	N2	N3	REM
Train	50	10930 (30.2%)	7652 (21.2%)	12118 (33.4%)	1511 (4.2%)	3965 (11.0%)
	100	21652 (30.0%)	15019 (20.7%)	24908 (34.3%)	3172 (4.4%)	7741 (10.7%)
	150	31531 (29.0%)	22187 (20.4%)	37937 (34.9%)	4828 (4.5%)	12108 (11.2%)
	200	43187 (29.8%)	29854 (20.6%)	50248 (34.7%)	5781 (4.0%)	15713 (10.9%)
	250	53094 (29.5%)	37517 (20.9%)	62665 (34.8%)	7209 (4.0%)	19505 (10.8%)
	300	63573 (29.3%)	44659 (20.6%)	76071 (35.1%)	8929 (4.1%)	23665 (10.9%)
	350	72610 (28.9%)	51085 (20.3%)	90216 (35.8%)	10465 (4.2%)	27210 (10.8%)
	400	86707 (30.0%)	58189 (20.1%)	101771 (35.2%)	11742 (4.1%)	30520 (10.6%)
Validation	91	19191 (28.6%)	14507 (21.6%)	23020 (34.2%)	2690 (4.0%)	7830 (11.6%)
Test	91	19160 (30.0%)	12779 (20.0%)	22542 (35.3%)	2624 (4.1%)	6720 (10.6%)

III. SLEEP STAGE CLASSIFICATION: DATASETS AND MODEL

A. DATASETS

We use three datasets for our analysis: the SHHS [18], [19] public dataset, Institution-A dataset and Institution-B dataset. Institution-A and Institution-B datasets are collected from different hospitals using a Noxturnal software system. Therefore, the sleep experts of each institution are different. The sampling rates of the signals in the SHHS, Institution-A and Institution-B are 125 Hz, 200 Hz and 200 Hz, respectively.

Table 1, 2 and 3 present the distribution of sleep stages on the datasets used in this experimental study. The number of PSG records in SHHS, Institution-A and Institution-B dataset are 5,550, 582 and 2,266, respectively. The Institution-A dataset includes various patient records with obstructive sleep apnea (OSA), encompassing normal, mild, moderate, and severe cases.

TABLE 3. Class distributions of Institution-B PSG dataset.

Dataset	# of PSG	# of epochs (samples)				
		Wake	N1	N2	N3	REM
Train	50	8452 (23.9%)	4756 (13.4%)	11913 (33.6%)	5098 (14.4%)	5207 (14.7%)
	400	64946 (23.0%)	37079 (13.1%)	92168 (32.7%)	48117 (17.1%)	39741 (14.1%)
Validation	340	57112 (23.7%)	32406 (13.5%)	77360 (32.1%)	39197 (16.3%)	34688 (14.4%)
Test	340	53837 (22.3%)	30935 (12.8%)	79638 (33.1%)	40347 (16.8%)	36143 (15.0%)

The ratios of men and women in the Institution-A dataset are 78.01% (454 records) and 21.99% (128 records), respectively. In addition, the mean and standard deviation of age in Institution-A dataset are 48.74 and 15.47, respectively. The statistical characteristic (mean and standard deviation) of the AHI-index in the Institution-A dataset is 47.07 ± 30.09. The individual statistical characteristic of the AHI-indices for normal, mild, moderate and severe cases are 1.71 ± 1.03 (21 records), 9.58 ± 3.10 (66 records), 22.65 ± 4.46 (113 records) and 63.26 ± 24.02 (382 records), respectively.

The demographic information for the "Institution-B" dataset is not available due to the strict data protection rules of the Institution-B. The dataset is divided into training, validation, and test sets with a 70:15:15 ratio on SHHS, Institution-A and Institution-B datasets. Therefore, the numbers of PSG records for a test set in SHHS, Institution-A and Institution-B are 832, 91 and 340, respectively. In addition, to verify the generalization ability, we utilize various numbers of PSG records for training the model (50, 100, ..., 400).

The distribution of sleep stages for the patients with severe OSA differs from that of the patients with normal OSA, as seen in the SHHS. In particular, the institution-A dataset consists of a relatively large distribution of N1. On the contrary, the distribution of N3 is smaller than SHHS dataset. We mainly use the "C3-A2" and "C3-M2" single EEG channels to classify sleep stages for SHHS and Institution datasets, respectively. Finally, two preprocessing approaches are applied to PSG record data. First, the band-pass filter is used to extract signal information within 0.5-35 Hz range, following the AASM criterion. Second, the z-score normalization is used to rescale the values to improve robustness and convergence speed.

B. MODEL ARCHITECTURE

To investigate the effect of deep learning models on the accuracy and the sensitivity with regard to the size of training dataset, we employ various backbone model architectures: DeepSleepNet, ResNet-based model and a transformer model. Through the experiments with those backbone models, we attempt to search the robust architecture when training the model with a limited number of samples.

Figures 1 and 2 present the detailed structures of the mod-

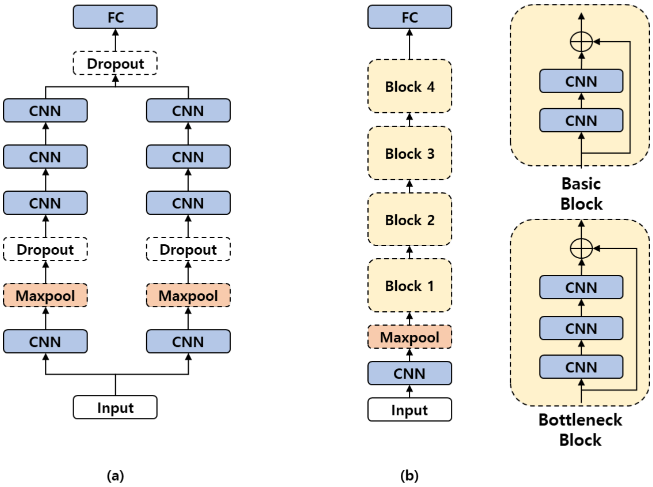

FIGURE 1. CNN-based model architectures: (a) a DeepSleepNet architecture and (b) a ResNet-based architecture.

els employed for sleep stage classification. In the CNN-based model architecture, we use DeepSleepNet architecture shown in Fig. 1(a) and ResNet-based CNN architecture shown in Fig. 1(b).

For the transformer-based model architecture shown in Fig. 2, we only utilize encoder blocks for extracting features from an epoch such as ViT [3]. Before inputting the signal into the encoder block in the transformer architecture, the signal is divided into sub-patches and then passed through an embedding layer. For dividing patches from an epoch, we use sliding window algorithm [20]. We use window size ($W = 200(sampling\ rate) \times 4(sec)$) and stride ($S = 200(sampling\ rate) \times 2(sec)$) to divide the patches from an epoch. According to the sliding window method, each patch has **4-second information** with 2-second overlapping between adjacent patches [21].

Therefore, the number of patches (P) can be calculated as follows:

$$P = \frac{200(sampling\ rate) \times 30(sec) - W}{S} + 1 \quad (1)$$

From Eq. 1, we know that **14 patches** (i.e., $(200 \times 30 - 800)/400 + 1$) are generated from an epoch sample. Then, the generated 14 patches pass through an embedding layer and corresponding embedding latent vectors are used as an input for a transformer model. For the embedding, we aim to incorporate the inductive biases formed by a CNN into the transformer model through the use of a CNN-based embedding layer instead of a linear projection such as the architecture used in [22].

As shown in Fig. 2, the CNN-based embedding layer block consists of 5 small-sized stacked convolution layers, each with a Rectified Linear Unit(ReLU) and batch normalization. In the transformer block, we employ three encoder blocks that consist of an MHSA and an FFN. In this work, 8 heads are used in the MHSA to extract various attention scores using queries, keys, and values.

C. MODEL TRAINING

In this paper, we not only focus on identifying a suitable model for extracting important features from an epoch sample, but also aim to achieve generalized performance with limited amounts of data. The limited number of data samples is a common problem in medical applications. As a result, it's crucial to build a model with high accuracy and strong generalization capabilities using only a small number of samples.

To assess the generalization abilities of different model architectures, we train the three different model architectures while keeping the hyperparameters constant. The Adam optimizer is used for training with a learning rate of 0.0001, and a batch size of 512. No data augmentation techniques are applied during training as it was observed that the model performance decreases when techniques such as flipping, jittering, scaling, and shifting are used. A cosine schedule is employed to regulate the learning rate over the training epochs.

D. NEW MASK DESIGN FOR CONTROLLING ATTENTION RANGE

The equations for an original MHSA can be expressed as follows [1]:

$$MHSA = concatenate(H_1, H_2, ..., H_N)W^O \quad (2)$$

$$H_i = SelfAttention(Q_i, K_i, V_i) \quad (3)$$

$$SelfAttention(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \quad (4)$$

In the equations, Q_i, K_i and V_i are the query, key and value embedding vectors, respectively, to utilize self-attention (H_i) for the i-th head in an MHSA and those embedding vectors are derived from input patches. d_k is the dimension of the keys. W^O is a weight matrix for a fully connected layer in MHSA. The original MHSA inside encoder layers does not use a mask filter to restrict the use of specific patches. Therefore, the original MHSA can utilize all the patch information without any constraint on the local and global relations between the patches.

In this paper, we observe that a transformer-based architecture attains generalized performance even though using the small-scale dataset for sleep stage classification. The observation goes against the commonly held notion that a transformer requires large amounts of data for achieving good accuracy due to its lack of inductive bias. To investigate the reason behind a transformer-based architecture outperforms a CNN-based architecture in our sleep stage classification, we propose 'Masked MHSA' (M-MHSA) which can control the range of the global information considered within the MHSA.

The masked attention mechanism used previously in transformer-based language models is typically implemented by adding a mask to the dot-product attention calculation.

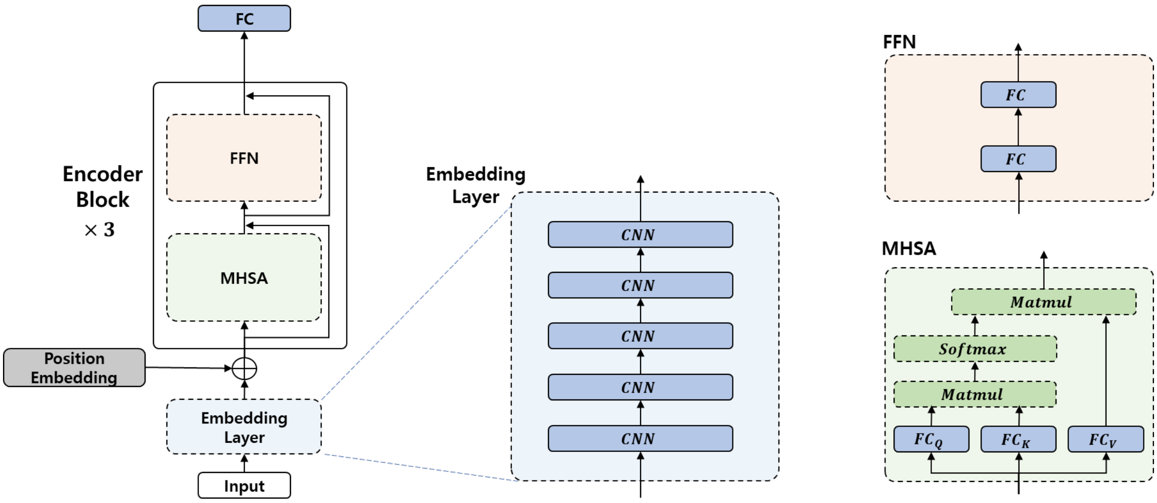

FIGURE 2. Overall of the transformer-based architecture. The transformer-based model consists of 3 modules embedding layer, MHSA and FFN.

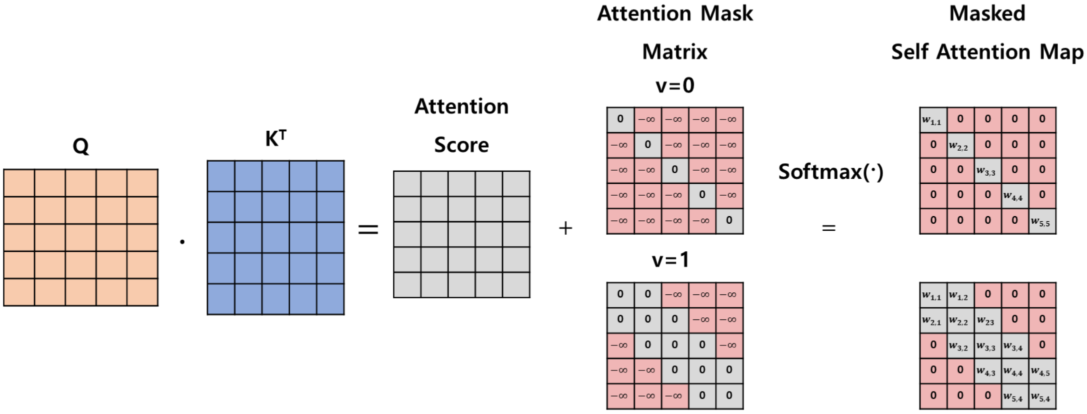

FIGURE 3. Adding an attention mask $\mathbf{M}^v$ to constraint the range of attention between patches in the masked self attention map.

The mask is applied to the attention scores, which are used to weight the representation of different elements in the sequence. The mask is usually implemented as a matrix with large negative numbers (-∞) in the positions where attention should be prohibited, and zeros in the positions where attention is allowed. The mask is added to the attention scores before they are passed through a softmax function to produce the attention weights. In conventional masked attention in decoder layers, the upper diagonal of the mask is set to large negative values to remove the attention scores for masked elements. This makes the attention weights for these elements close to zero after the softmax function, effectively removing their impact on the final attention output.

In this work, unlike the previous mask configuration, in order to constrain and control the range of attentions over near or far patches, a newly devised masked attention approach is proposed, where zeros are placed in the main diagonal and additional neighbor diagonals. The rest of the non-diagonal part is filled with large negative values. The resulting mask has a “band (or banded) matrix" shape, which is a well-known structure in matrix theory [23] but with a different value arrangement.

A mask, $\mathbf{M}^v$, for attention can be defined as an $n \times n$ matrix with *view width* v, where v is a positive integer, then the i-th row and j-th column element, $\mathbf{M}^v_{i,j}$ $(1 \leq i, j \leq n)$, in the matrix satisfy the following conditions.

$$\mathbf{M}^v = \begin{cases} M^v_{i,j} \;=\; -\infty, & if\ |i-j| > v \\ M^v_{i,j} \;=\; 0, & else \end{cases} \tag{5}$$

In other words, the elements of the matrix outside of the main diagonal and the diagonals v steps away from it are set to zero. The main diagonal is represented by the index difference, $|i-j| = 0$.

With the definition of the proposed mask, the masked self-attention ($MaskedSA$) can be calculated with Eq. 6 as follows:

$$MaskedSA(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}} + \mathbf{M}^v)V_i \tag{6}$$

$\mathbf{M}^v$ is an attention mask used to limit the scope of the attention map, as demonstrated in Fig. 3. When v is equal to 0, the transformer-based model can only utilize self-patch information. On the other hand, if v is greater than 0, the model can access much information from more distant patches ($2v$+1 patches), including the self-patch.

A small value of v results in the early transformer block only being able to consider information from nearby patches, as opposed to using a larger v of a mask. If v is not 0, however, the deeper transformer block can capture much more global information than the early transformer block, as the receptive field is larger in later layers compared to the early layers in CNNs. The v can be considered like as the size of a CNN filter in some sense. In our MHSA model, the proposed attention mask, $\mathbf{M}^v$, is added to the calculated original attention map, as shown in Eq. 6.

Our hypothesis is that the enhanced performance of the transformer-based architecture is contributed from its capability of effectively regulating the extraction of both local and global features based on input features from an epoch data sample, leading to better generalization compared to a conventional CNN-based architecture. Therefore, when applying strict local regulation in MHSA, the performance of the model will decline, while the performance of the model will increase when applying less regulation in the attention mask.

E. METHODOLOGY TO MEASURE DISTANCE BETWEEN DIFFERENT PATCHES (MEAN DISTANCE)

According to Fig. 4, a transformer-based architecture is shown to have more generalized model performance compared to conventional CNN-based architectures in sleep stage classification tasks without any pre-training step on a large-scale dataset. To prove our hypothesis that a transformer-based architecture can utilize global features well in addition to local features, we use a metric of "**mean attention distance**" to investigate how far a patch attends to other patches in an MHSA for the given input data, on average.

The mean attention distance is evaluated as the weighted average of the euclidean distances between the query patch and the other patches in the sequence, where the weights are given by the attention scores. The intuition behind this calculation is that the mean attention distance captures "how far away (globally related) the important patches are from the query patch, taking into account their relative importance as determined by the attention mechanism".

The mean attention distance (MD_i) for the i-th head of an MHSA can be calculated with Eq. 10 as follows:

$$AS(Q_i, K_i) = Softmax(\frac{Q_i K_i^T}{\sqrt{d_k}}) \tag{7}$$

$$Dist(Q_i, K_i) \in \mathbf{R}^{S\times S} \tag{8}$$

$$Dist(Q_i, K_i)_{j,z} = |j - z|_1 \tag{9}$$

$$MD_i = \frac{\sum_j^S \sum_z^S AS(Q_i, K_i)_{j,z} \cdot Dist(Q_i, K_i)_{j,z}}{S} \tag{10}$$

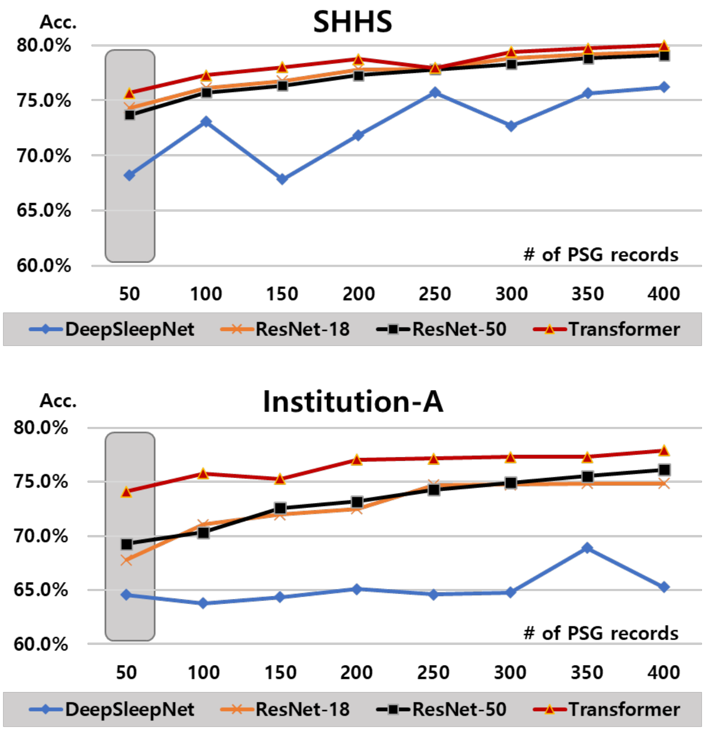

FIGURE 4. Performance of various model architectures in sleep stage classification into 5 classes (Wake, N1, N2, N3 and REM) when using a C3-A2 and C3-M2 signal channel for SHHS and Institution-A dataset, respectively.

$AS(Q_i, K_i)$ and $Dist(Q_i, K_i)$ are the matrices for the attention score of the i-th head and euclidean distance, respectively. The elements at the j-th row and z-th column of the matrices, $AS(Q_i, K_i)$ and $Dist(Q_i, K_i)$, are denoted by $AS(Q_i, K_i)_{j,z}$ and $Dist(Q_i, K_i)_{j,z}$, respectively.

IV. RESULTS

A. BASIC OBSERVATIONS

When only relying on epoch data, the choice of a model architecture used to extract the features can result in significant performance differences, as shown in Fig. 4. In this paper, we attempt to figure out what is the best backbone architecture for extracting features from an epoch signal in sleep stage classification, especially when the amount of data is limited. Our results show that the transformer-based architecture has more generalization capability than the conventional CNN-based architecture even when only a small-scale dataset is utilized. This result contradicts common expectations regarding transformers. Thus, we aim to investigate the reason for this phenomenon through further experiments.

B. GENERALIZATION CAPABILITY OF A TRANSFORMER-BASED MODEL

To analyze and assess the generalization capability of a transformer-based model while comparing with other deep neural network architectures, we train multiple models using a limited amount of patient PSG data. The models are evaluated in this experiment include DeepSleepNet, a ResNet-

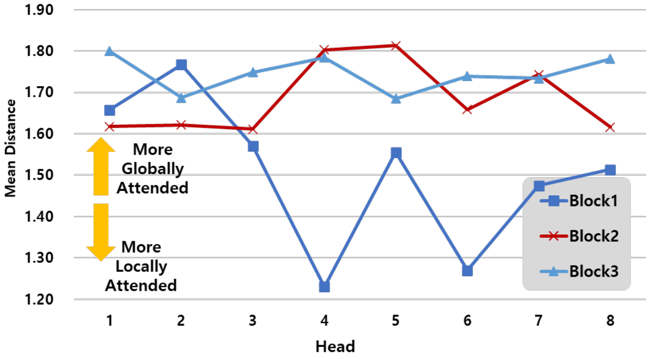

FIGURE 5. Mean attention distance of the eight heads in MHSAs at different encoder block layers. The result is obtained from the use of 400 PSG records.

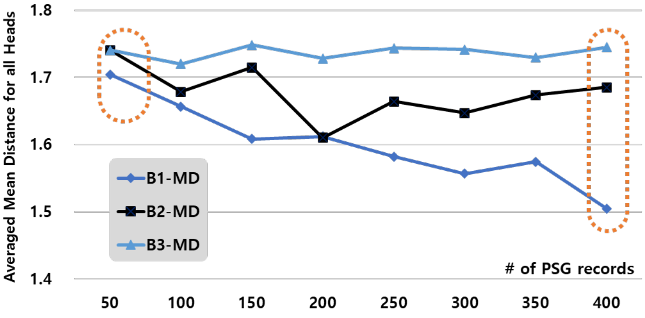

FIGURE 6. The average of mean attention distances in the heads of an MHSA for varying amounts of dataset in training.

based model and a transformer-based model. Typically, researchers extract meaningful latent features from an epoch sample using 1D-CNNs, which are well-suited for capturing time-invariant information.

Figure 4 shows the performance of the models using different amounts of training data. Interestingly, the transformer-based architecture has better generalized performance compared to CNN-based models, even with a small number of data samples and without pre-training processing about a large-scale dataset.

From the results of Fig. 4, we hypothesize that the generalized performance of the transformer-based architecture is due to its ability to effectively manage the extraction of well-balanced local and global features according to input features from an epoch data sample, resulting in improved generalization capability compared to a conventional CNN-based architecture. To validate the hypothesis, in the next subsection, we conduct multiple experiments to investigate the attention behavior of the MHSA in a transformer architecture in depth.

C. MEAN ATTENTION DISTANCE TO CONFIRM GLOBALITY IN MHSA

In this section, we analyze the MHSA to verify the globality based on mean distance. According to Eq. 10 in Sec. III-E, when the mean attention distance is low, a self-attention head focuses on local information more than global information. On the other hand, when the mean attention distance is high, the head focuses on global information more than local information. As you can see in Fig. 5, in the encoder block (**Block1**) at a shallow layer, there are cases where certain heads (i.e., Head 4 and Head 6) in the MHSA utilize local information more than global information, while in the other cases, different heads utilize global information more than local information.

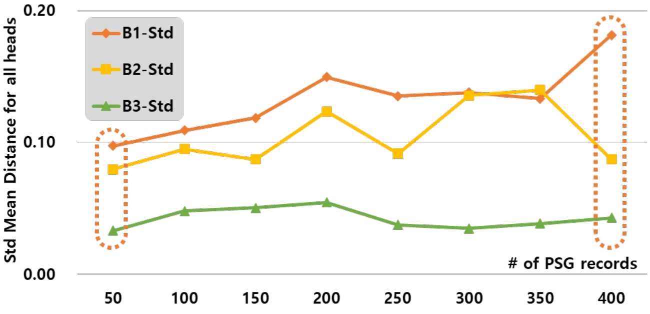

FIGURE 7. The standard deviation of mean attention distances in the heads of an MHSA for varying amounts of dataset in training.

This result implies that the MHSA (Block1) at the shallow encoder layer considers both local and global information. On the other hand, when the encoder layer becomes deeply located (i.e., Block2 and Block3), it can be seen that the mean attention distances of all heads become high, which means that all the heads are focusing on utilizing global features rather than local features. In other words, both local and global information are utilized in the MHSA (Block1) of the shallow layers while global information is more strongly utilized in the MHSA (Block2 and Block3) of the deeper layers. Compared to CNN architectures, the MHSAs of a transformer architecture can exploit the more globally correlated information between the patches of epoch data. From Fig. 5, we guess that the more globally attended information obtained through the transformer makes it possible to achieve higher prediction accuracy even with a small size dataset in our sleep stage classification.

Figures 6 and 7 show how the “averaged mean attention distance” and the “standard deviation of the mean attention distances” vary according to the size of training dataset. The averaged mean attention distance is the average value of all mean attention distance (MD_i) over all eight heads at a certain block layer. For an example, in Fig. 6, “B1-MD” is the average value of all the mean attention distance for all the heads in the Block1. On the other hand, in Fig. 7, “B2-Std” is a standard deviation of mean attention distances of all the heads in the MHSA of the Block2.

In general classification tasks, the layer that has the greatest influence is usually the layer closest to the output. So, the attention behavior in the MHSA of the output layer was investigated, and it was observed that the MHSA of the last encoder layer utilized global information primarily regardless of the amount of training data. As presented in Fig. 5, the mean attention distances of all the heads in Block3 (the block closest to output) are evaluated as high and it means that *the MHSA in the block near an output layer has an attention pattern of more globally correlated among patches*. Accordingly, the average mean attention distance of the Block3 (B3-MD) is always evaluated as high between 1.7 and 1.8 regardless of the amount of training data in Fig. 6.

Unlike B3-MD, B1-MD (the averaged mean attention distance of the Block1 which is the block closest to input) is reduced as the size of a dataset increases as shown in Fig. 6. The result implies that the MHSA near an input side has the *attention behavior of decreasing global attention and increasing local attention as the number of train data sample increases*. It is note worthy that *the globally activated attentions are observed in the MHSA of the encoder block near an input side when the size of data sample is small*.

In the case of the CNN, unlike the attention behavior of a transformer architecture, the convolution layer near an input size has a small and limited receptive field, which means that it uses only local information regardless of a dataset size. We believe that this is a key difference between our transformer architecture and a CNN architecture and the performance difference between the two architectures is derived from the difference particularly in the case of employing small-sized dataset.

From the experimental result of the sleep stage classification showing that a transformer architecture shows better performance than a CNN architecture when a small size dataset is available, we guess that *the global information works as a much critical feature for classifying a sleep stage when the dataset size is small* and such a property of the sleep stage classification is well exploited in the MHSA of a transformer architecture which has a capability of exploiting much global information extracted from the epoch sample than a CNN architecture.

D. BALACING BETWEEN GLOBAL AND LOCAL FEATURES USING MASKED MHSA BY CONTROLLING THE RANGE OF ATTENTION IN AN MHSA

It is shown that a transformer-based model can achieve more robust performance even when using a small-scale dataset, compared to other CNN-based models in our application domain. In the previous section, with the mean attention distance derived from the three different MHSAs at the different layers of a transformer model, we analyzed the attention behavior of the results and we infer the possible interpretation for the results.

To analyze the more detailed working mechanism of an MHSA in a transformer-based model, we investigate the patterns of patch interactions using the attention score matrix. After then, we use M-MHSA presented in Section III-D to constrain the degree of global attention in the M-MHSA by changing the value of view width, v, of the M-MHSA.

Figure 8 shows the reconstructed attention map utilizing the attention mask as given in Eq. 6. Adding an attention mask with a *larger* v (e.g., the attention map given at the right bottom with $v = 13$ in Fig. 8) to the original attention score results in the masked attention map that utilizes *more global* correlations between patches. On the other hand, adding a mask with a *smaller* v (e.g., the attention map given at the left top with $v = 0$ in Fig. 8) to the original attention score results in the masked attention map that focuses on *more local* patch interaction. In consequence, by changing the value of v, we can control the range of attention so that we can limit the globality of the attention in an M-MHSA. Then, the final masked self attention map can be adjusted to incorporate "the desired locality or globality level of the attentions between patches" into the M-MHSA as presented in detail in Fig. 3.

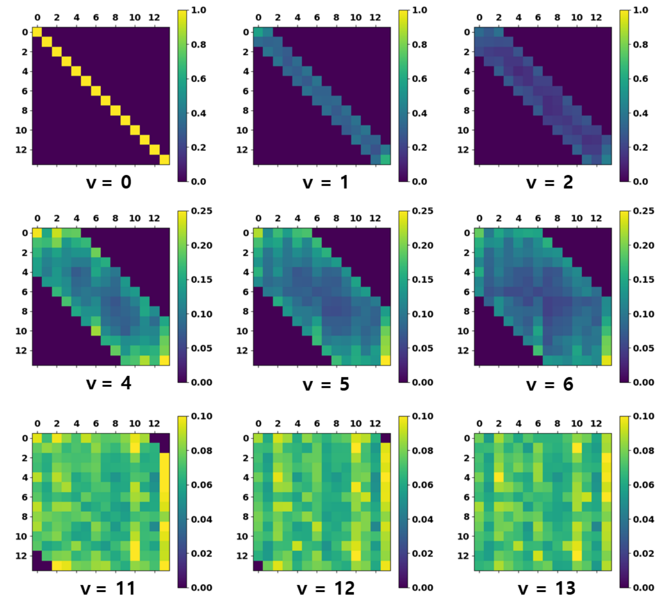

FIGURE 8. The observed patterns of attentions in an MHSA of the Block3 according to the various v of an attention mask. 50 PSG records are used for the training.

We investigate the impact of the attention range that can be controlled by the attention mask, $\mathbf{M}^v$, on the performance of a transformer architecture by conducting several experiments while varying the value of v in the attention mask of the M-MHSA. The experiments aim to figure out the performance impact of constraining the range of attentions of an M-MHSA in our sleep stage classification.

From Fig. 8, we observe that a transformer model tries to obtain global information as much as possible under a given constraint. When the attention range is limited by v, the attentions between the most distant patches are generally higher than the attention between the locally correlated patches (near diagonal in the figure). We think that we can find the best value of v of an M-MHSA in order to achieve an optimal performance.

In order to evaluate the importance of the attention range for the performance of a transformer architecture, we utilize an attention mask to control the attention range. Table 4 shows the accuracy performances of a sleep stage classification with a transformer architecture while using various v (from 0 to 13) of attention masks and various numbers of PSG records (from 50 to 400). Note that we use only single "C3-M2 channel" data from the PSG records.

When a view width, v, is set to zero to totally eliminate the attentions between two different patches in an epoch sample and to consider the case of utilizing only the inner-patch information (the attention map is presented with $v = 0$ in Fig. 8), as presented in Table 4, the performance of the transformer-based model is lower than that of the transformer-based model with other larger v trivially since no information between patches is utilized.

On the other hand, when v becomes larger than 0, the performance increases and the increase stops at a certain v (this is an *optimal* v *in terms of performance*) between 0 and 13. It is noteworthy that *when* v *is more than* ***2****, the model can achieve the performance similar to that obtained without an attention mask* (i.e., $v = 13$ and this is the case of fully utilizing all the attention information between patches so all possible local and global information are used).

Even though v is less than 13, limiting the attention range, the M-MHSA at deeper encoder layers can obtain more global attention information compared to the M-MHSA at the earliest layer near the input side. This is similar to the larger receptive field at deeper layers of a CNN-based architecture. It is also interesting to see that the model achieves higher performance with smaller v than that of using larger v for attention masks. We guess this happens because the restriction on the attention range of an M-MHSA *prevents overfitting* caused by too much global information and it also *injects a local inductive bias* into the encoder blocks of a transformer model [24].

From the results presented in Fig. 5, 6 and 7, the transformer-based architecture has attention to various ranges of attention information inside the early encoder block (i.e., Block1) while it has attention to global information inside the deep encoder block (i.e., Block3) when the model utilizes a sufficient amount of the dataset (i.e., 400 PSG records).

A moderately M-MHSA is shown to make a transformer model well-suited for the inherent range of inter-patch correlation in the input data samples for a specific application, i.e., sleep stage classification. Such a moderately constraint M-MHSA can make earlier layers to focus on local inter-patch interactions while the global attention between long distant patches is exploited well at deeper layers with more emphasis. Finally, the model can achieve higher performance than that without M-MHSA by controlling v properly in the M-MHSA.

Since the multi-channel sleep stage classification using multiple bio-signal channel data is also an important problem in addition to single channel sleep stage classification, we conduct another experiment with the multi-channel input signals which are available from PSG records. Table 5 shows the performances for the cases of using various v (from 0 to 13) of attention masks while increasing channels. Note that "**+ Channel-Name**" in Table 5 means that the channel is added additionally. So, the results in the second column, "C3-M2" implies the case of using the single C3-M2 channel. On the other hand, the third column, "+C4-M1", include the results of using two channels, C3-M2 and C4-M1. With continuous additions, the results in the last column, "+Chin" implies the case of using all the channels, C3-M2, C4-M1, E1-M2, E2-M1 and Chin.

TABLE 4. Model performance when using various v (0 to 13) for the attention masks in M-MHSA and different numbers (50 to 400) of PSG records including C3-M2 channel data. Red and blue colors mean first highest performance and second highest performance compared to different v with the same number of training PSG records, respectively.

v	# of PSG records							
	50	100	150	200	250	300	350	400
0	72.95%	73.23%	73.00%	74.89%	75.23%	74.77%	75.52%	75.36%
1	73.14%	74.38%	74.80%	76.57%	76.88%	76.29%	77.20%	77.32%
2	74.00%	75.55%	76.08%	76.68%	77.23%	77.18%	77.82%	77.90%
3	74.40%	75.75%	76.09%	**77.32%**	77.45%	77.56%	77.06%	77.78%
4	74.37%	**75.99%**	76.29%	76.63%	77.23%	77.21%	77.54%	77.91%
5	**74.62%**	75.83%	**76.33%**	77.23%	77.38%	77.32%	77.74%	77.97%
6	74.44%	75.81%	76.29%	77.23%	**77.52%**	77.34%	77.48%	77.91%
7	74.57%	75.74%	75.40%	77.07%	77.51%	77.60%	**77.92%**	**78.05%**
8	74.45%	75.91%	75.75%	76.61%	77.09%	77.47%	77.53%	77.99%
9	74.34%	75.69%	75.48%	76.44%	77.01%	**77.71%**	77.35%	77.82%
10	74.39%	75.90%	75.76%	77.03%	77.41%	77.70%	77.29%	77.73%
11	74.23%	75.73%	75.77%	77.01%	77.16%	77.43%	77.49%	77.87%
12	74.17%	75.73%	75.77%	77.09%	77.22%	77.40%	77.40%	77.86%
13	74.11%	75.78%	75.28%	77.04%	77.17%	77.29%	77.29%	77.92%

TABLE 5. Model performance when using various v for the attention masks in M-MHSA with multi-channel signal data. For the experiment, 50 PSG records are used for training. Red and blue colors mean first highest performance and second highest performance compared to different v with the same channels, respectively.

v	Channel				
	C3-M2	+ C4-M1	+ E1-M2	+ E2-M1	+ Chin
0	72.95%	72.88%	72.03%	73.33%	74.19%
1	73.14%	73.50%	73.18%	73.64%	74.24%
2	74.00%	73.89%	74.53%	74.62%	74.71%
3	74.40%	74.70%	74.48%	74.81%	75.13%
4	74.37%	**74.80%**	**74.95%**	75.15%	74.67%
5	**74.62%**	74.52%	74.44%	**75.41%**	74.72%
6	74.44%	74.75%	74.42%	75.00%	**75.30%**
7	74.57%	74.20%	74.72%	75.03%	74.73%
8	74.45%	74.31%	74.29%	74.92%	74.77%
9	74.34%	74.27%	74.43%	74.83%	75.18%
10	74.39%	74.35%	74.32%	74.66%	74.78%
11	74.23%	74.12%	74.66%	75.10%	74.64%
12	74.17%	74.28%	74.79%	74.82%	74.92%
13	74.11%	74.26%	74.84%	74.48%	74.95%

The experimental results in Table 5 is obtained from a model trained with 50 PSG records. The result of Table 5 is similar to Table 4 from the perspective that there exists an optimal value of v between 0 and 13. Lastly, we find out that the capability of utilizing proper range of attention in an M-MHSA of a transformer architecture helps to improve its performance and the use of an appropriate v of masked attention in the M-MHSA can enhance the generalizability of the model in sleep stage classification tasks.

TABLE 6. Model performance when using various model architectures when using 50 and 400 PSG records. The C3-A2 and C3-M2 channels are used for training SHHS and Institution-A/Institution-B, respectively. Red and blue colors mean first highest performance and second highest performance compared to different model architectures, respectively.

# of PSG records	Dataset (Used Channel)	Model Architecture	Performance						
			F1-score					W-Acc.	Acc.
			Wake	N1	N2	N3	REM		
50	SHHS (C3-A2)	DeepSleepNet	0.813	0.002	0.718	0.672	0.531	57.385	68.214(−7.510)
		ResNet-18	0.852	0.018	0.772	0.669	0.576	57.620	74.320(−1.404)
		ResNet-50	0.851	0.076	0.766	0.674	0.547	57.603	73.703(−2.021)
		Transformer	**0.866**	0.070	0.774	**0.703**	**0.636**	61.829	75.611(−0.113)
		Ours (M-MHSA, $v = 2$)	**0.866**	**0.162**	**0.783**	0.700	0.625	**62.143**	**75.724(−0.000)**
	Institution-A (C3-M2)	DeepSleepNet	0.753	0.449	0.750	0.578	0.498	61.196	64.547(−10.073)
		ResNet-18	0.789	0.468	0.771	0.689	0.463	63.319	67.768(−6.852)
		ResNet-50	0.801	0.354	0.779	0.624	0.549	61.544	69.277(−5.343)
		Transformer	0.829	0.572	0.799	0.707	**0.644**	71.115	74.109(−0.511)
		Ours (M-MHSA, $v = 5$)	**0.833**	**0.573**	**0.803**	**0.747**	0.628	**71.457**	**74.620(−0.000)**
	Institution-B (C3-M2)	DeepSleepNet	0.764	0.202	0.649	0.742	0.559	62.434	63.283(−8.517)
		ResNet-18	0.777	0.384	0.663	0.741	0.650	65.923	66.775(−5.025)
		ResNet-50	0.773	0.283	0.705	0.756	0.621	62.667	67.790(−4.010)
		Transformer	0.795	0.409	0.734	0.767	0.692	67.817	71.178(−0.622)
		Ours (M-MHSA, $v = 6$)	**0.799**	**0.444**	**0.739**	**0.776**	**0.695**	**68.968**	**71.800(−0.000)**
400	SHHS (C3-A2)	DeepSleepNet	**0.891**	0.180	0.787	0.686	0.683	64.534	76.202(−3.979)
		ResNet-18	0.886	0.185	0.810	**0.736**	0.721	66.514	79.353(−0.828)
		ResNet-50	0.881	0.138	0.810	0.724	0.714	65.108	79.074(−1.107)
		Transformer	**0.891**	0.189	0.813	0.731	0.747	67.268	79.982(−0.199)
		Ours (M-MHSA $v = 3$)	**0.891**	**0.202**	**0.815**	0.735	**0.750**	**67.498**	**80.181(−0.000)**
	Institution-A (C3-M2)	DeepSleepNet	0.835	0.506	0.617	0.701	0.614	66.818	65.258(−12.712)
		ResNet-18	0.834	0.595	0.817	0.662	0.593	67.583	74.845(−3.125)
		ResNet-50	0.842	0.585	0.821	0.726	0.646	71.080	76.125(−1.845)
		Transformer	0.855	0.610	**0.833**	**0.751**	0.693	74.546	77.916(−0.054)
		Ours (M-MHSA, $v = 6$)	**0.859**	**0.619**	0.829	0.745	**0.700**	**76.018**	**77.970(−0.000)**
	Institution-B (C3-M2)	DeepSleepNet	0.812	0.499	0.704	0.779	0.734	72.366	72.196(−3.444)
		ResNet-18	0.817	**0.515**	0.748	0.786	0.747	72.206	73.722(−1.918)
		ResNet-50	0.813	0.450	0.758	0.795	0.717	69.308	73.631(−2.009)
		Transformer	**0.833**	0.509	0.763	0.798	0.768	**73.270**	75.599(−0.041)
		Ours (M-MHSA, $v = 11$)	0.832	0.508	**0.765**	**0.799**	**0.770**	73.182	**75.640(−0.000)**

E. VALIDATION IN OTHER DATASETS

In addition to two datasets (SHHS and Institution-A), we use another dataset (Institution-B) that is collected from a different medical institution in order to show the *experimental reliability and consistency* on the performance superiority of a transformer architecture particularly for a sleep stage classification task. To evaluate the performance on the Institution-B dataset, the models are trained on the Institution-B training dataset using 50 and 400 PSG records and then evaluated on the test dataset from the Institution-B dataset.

Table 6 shows the performance obtained through evaluating the various model architectures with SHHS and our two datasets (Institution-A and Institution-B). The transformer-based architecture achieves higher performance compared to traditional CNN-based architectures when using both of 50 and 400 PSG records to train the model.

Specifically, even when using a small number of PSG records (50 PSG records) to train the models, the transformer-based architecture with our M-MHSA attains **1.404%**, **5.343%** and **4.010%** higher accuracy than different CNN-based architecture on SHHS, Institution-A and institution-B datasets, respectively. When we use a large number of PSG records to train the model, transformer-based architecture also achieves 0.828%, 1.845% and 1.918% higher accuracy on SHHS, Institution-A and Institution-B datasets, respectively. The proposed method achieves better weighted accuracy excluding Institution-B dataset with 400 PSG records compared to the original transformer method which does not include M-MHSA. Transformer-based architecture already obtains better overall performance compared to conventional CNN-based architecture. Specifically, one of the main goals is to improve the generalization ability utilizing a M-MHSA to reflect globality restriction in MHSA. According to Table 6, the proposed method outperforms different model architectures including original transformer-based architecture even with small-scale training PSG records in which case the probability of falling into overfitting is high.

V. CONCLUSION

In this paper, we analyzed three representative deep learning model architectures for sleep stage classification tasks and we aimed to determine a suitable backbone model architecture for extracting meaningful features from a PSG dataset, particularly under the limited data availability. Traditionally, the most prevalent backbone architectures utilized in sleep stage classification have been CNN-based architectures. Recently, transformer-based architectures have been introduced by some researchers. However, they have not received as much attention as CNN-based architectures. In our experiments, we observed that a transformer-based architecture outperforms conventional CNN-based architectures, even when using a small amount of data. This result is intriguing, as typically transformer architectures require a large dataset for pre-training before fine-tuning in a specific task. However, in our experiments on the sleep stage classification task, the transformer-based architecture achieved the highest performance without any pre-training step on the well-known public dataset (SHHS) and two our own datasets (Institution-A and Institution-B datasets).

To understand the reason for these uncommon results obtained from the performance evaluations, we conducted additional experiments and analyzed the detailed behavior of the transformer architecture. For the analysis, we particularly investigated patterns of attentions in the MHSA of the transformer architecture and we found that the MHSA mechanism in the transformer allows the model to extract more useful latent features from an epoch sample data depending on the number of train samples, i.e., the size of a dataset. In addition, by introducing an attention mask utilizing the shape of a band matrix for controlling the range of attention between patches, we can improve the classification performance further by assigning an optimal value to a view width, v, of an attention mask.

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[4] Md Mosheyur Rahman, Mohammed Imamul Hassan Bhuiyan, and Ahnaf Rashik Hassan. Sleep stage classification using single-channel eog. Computers in biology and medicine, 102:211–220, 2018.

[5] Fernando Andreotti, Huy Phan, Navin Cooray, Christine Lo, Michele TM Hu, and Maarten De Vos. Multichannel sleep stage classification and transfer learning using convolutional neural networks. In 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pages 171–174. IEEE, 2018.

[6] Tianqi Zhu, Wei Luo, and Feng Yu. Convolution-and attention-based neural network for automated sleep stage classification. International Journal of Environmental Research and Public Health, 17(11):4152, 2020.

[7] Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. IEEE Transactions on Biomedical Engineering, 69(8):2456–2467, 2022.

[8] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International conference on machine learning, pages 10347–10357. PMLR, 2021.

[9] Zhiying Lu, Hongtao Xie, Chuanbin Liu, and Yongdong Zhang. Bridging the gap between vision transformers and convolutional neural networks on small datasets. Advances in Neural Information Processing Systems, 35:14663–14677, 2022.

[10] Allan Rechtschaffen. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. United States Government Printing Office, pages 1–55, 1968.

[11] Conrad Iber. The aasm manual for the scoring of sleep and associated events: rules, terminology, and technical specification. 1st ed. Westchester, IL: American Academy of Sleep Medicine, 2007.

[12] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. 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.

[13] Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya. Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. PloS one, 14(5):e0216456, 2019.

[14] Yudong Sun, Bei Wang, Jing Jin, and Xingyu Wang. Deep convolutional network method for automatic sleep stage classification based on neurophysiological signals. In 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pages 1–5. IEEE, 2018.

[15] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29:809–818, 2021.

[16] Mike X Cohen. Analyzing neural time series data: theory and practice. MIT press, 2014.

[17] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[18] Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. Sleep, 20(12):1077–1085, 1997.

[19] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. Journal of the American Medical Informatics Association, 25(10):1351–1358, 2018.

[20] Vladimir Braverman. Sliding Window Algorithms, pages 2006–2011. Springer New York, New York, NY, 2016.

[21] Dongyoung Kim, Jeonggun Lee, Yunhee Woo, Jaemin Jeong, Chulho Kim, and Dong-Kyu Kim. Deep learning application to clinical decision support system in sleep stage classification. Journal of Personalized Medicine, 12(2):136, 2022.

[22] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. Advances in Neural Information Processing Systems, 34:30392–30400, 2021.

[23] Gilbert Strang. Fast transforms: Banded matrices with banded inverses. Proceedings of the National Academy of Sciences, 107(28):12413–12416, 2010.

[24] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv: 1806.01261, 2018.

DONGYOUNG KIM received his B.E. and M.E. degrees of computer engineering from Hallym University in 2019 and 2021, respectively. Currently, he is a Ph.D student in the Department of Computer Engineering at Hallym University. His research interests focus on deep learning for medical applications and deep learning based domain adaption for multi-institutional dataset.

YOUNG-WOONG KO received the B.S. and M.S. degrees in computer engineering and the Ph.D. degree in computer science from Korea University, Seoul, South Korea, in 1997, 1999, and 2003, respectively. From 2009 to 2010, he was a Visiting Scholar at the University of Pennsylvania, Philadelphia, PA, USA, and Visiting Scholar at the Harvard Medical School, Boston, MA, USA, from 2016 to 2017. He is a Professor with Hallym University, South Korea. Since 2003, he has been a Professor with the Department of Computer Engineering, Hallym University. He was the Dean of the Software College from 2018 to 2022. His lectures cover operating systems, computer architectures, and embedded system. His research interests include operating systems, embedded system design, high-performance parallel, and distributed computing.

DONG-KYU KIM is an associate professor in the Department of Otorhinolaryngology-Head and Neck Surgery, Hallym University College of Medicine, Chuncheon, Republic of Korea. He completed his residency and clinical fellowship at the Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Hospital. His major is sleep medicine, allergy, and rhinology, which is a medicine dealing with the nose and upper respiratory tract. As a researcher, his practice focuses on immunology and big-data analysis. To date, he published several papers in high-impact factor articles, such as the Journal of American Academy of Allergy and Clinical Immunology, Journal of American Academy of Allergy and Clinical Immunology: In Practice, and Thorax. He is also interested in the field of artificial intelligence and is a Vice Director of the Center for Artificial Intelligence at Hallym University Chuncheon Sacred Hospital. He also has served as a peer reviewer and a member of the editorial board in several medical journals.

JEONG-GUN LEE (Member, IEEE) received his B.S. degree in computer engineering from Hallym University in 1996, and M.S. and Ph.D. degree from Gwangju Institute of Science and Technology (GIST), Korea, in 1998 and 2005. Currently, he is a Full Professor in the Department of Computer Engineering at Hallym University. Prior to joining the faculty of Hallym University in 2008, he was a visiting postdoctoral researcher of the Computer Lab. at the University of Cambridge and a research professor at GIST. In 2014, he was a visiting scholar of the Computer Lab. at the University of Cambridge. His research interests focus on deep learning for medical applications, FPGA based deep learning accelerator designs, energy efficient heterogeneous computing, and GPU based parallel computing.