Supplement to "MrsFormer: Transformer with Multiresolution-head Attention"

A Additional details on the experiments

A.1 UEA time series classification

Datasets and metrics The benchmark (Bagnall et al., 2018) consists of 30 datasets. Following (Wu et al., 2022), we choose 10 datasets, which vary in input sequence lengths, the number of classes, and dimensionality, to evaluate our models on temporal sequences.

Models and baselines We adapt code from (Wu et al., 2022; Zerveas et al., 2021) for our experiments. Following the same setting from these papers, we set the number of heads and layers to 8 and 2, respectively. For the MrsFormers, we use the same set of scales at each layer, which is given by $\boldsymbol{s} = [1, 1, 2, 2, 4, 4, 8, 8]$. For MRA-2 and MRA-2-s models (Zeng et al., 2022), each head is approximated by blocks of scales $[1, 32]$ as suggested in their paper. The percentage of blocks with scale 1 in these MRA-2 models is set to 25% of the full attention matrix. Other hyperparameters have the same values as in (Wu et al., 2022) (for the PEMS-SF, SelfRegulationSCP2, and UWaveGestureLibrary tasks) and (Zerveas et al., 2021) (for other tasks). Hyperparameters for these tasks are presented in Table 6.

A.2 Long range arena benchmark

Datasets and metrics We adopt the tasks: Listops (Nangia & Bowman, 2018), byte-level IMDb reviews text classification (Maas et al., 2011), and byte-level document retrieval (Radev et al., 2013) in the LRA benchmark for our experiments. They consist of long sequences of length $2K$, $4K$, and $4K$, respectively. The evaluation protocol and metric are the same as in (Tay et al., 2021b).

Models and baselines We follow the same settings and adapt code for LRA task from (Zeng et al., 2022), which uses transformer with 2 heads and 2 layers. We choose the same set of scales $\boldsymbol{s} = [1, 2]$ for all the layers in MsFormer. Hperparameters for these tasks are presented in Table 7.

A.3 Image Classification on Imagenet

Dataset and metric: We perform classification task on ILSVRC-2012 ImageNet dataset to validate the performance of our model on large dataset. This dataset has 1000 classes and about 1.28 million images.

Models and baselines In this section, we apply the MrsFormer to the Deit model (Touvron et al., 2020) with 4 heads. Since Deit uses special class token $[CLS]$ for the classification, we do not downsample this token along with other tokens in the sequence. For our MrsFormers, we use the set of scales $\boldsymbol{s} = [1, 2, 2, 4]$ at each layer. We also study the MRA-2-s attention on this task. As reported in (Zeng et al., 2022), the MRA-2-s is a better model than the MRA-2 on the ImageNet image classification task since its sparse attention structure is more effective for modeling images.

B Proof of Theorem 1

Recall from Eqn. (14) that

$$\mathbf{H} \approx \mathbf{H}^{s,s'} = \uparrow_{s,1} ((\downarrow_{s,1}\downarrow_{s',2} \mathbf{A})(\downarrow_{s',1} \mathbf{V})).$$

Let $\mathbf{T}_s$ be the down-sampling operator (matrix multiplication) on the first dimension of a matrix corresponding to the scale s. $\mathbf{T}_s$ is the Kronecker product (or outer product) between an identity matrix $\mathbf{I}$ and the row vector $\frac{1}{s_i}\overrightarrow{\mathbf{1}}$ of size $1 \times s$, i.e. $\mathbf{T}_s = \mathbf{I} \otimes \frac{1}{s}\overrightarrow{\mathbf{1}}$. Under this notation, the up-sampling operator is the transpose of $\mathbf{T}_s$. In addition, the down-sampling operator on the second dimension of a matrix is also $\mathbf{T}_s^T$ but with the right multiplication instead. Then, we can rewrite the approximation $\mathbf{H}^{s,s'}$ as follows:

$$\mathbf{H}^{s,s'} = \mathbf{T}_s^T((\mathbf{T}_s\mathbf{A}\mathbf{T}_{s'}^T)(\mathbf{T}_{s'}\mathbf{V})) = (\mathbf{T}_s^T\mathbf{T}_s\mathbf{A}\mathbf{T}_{s'}^T\mathbf{T}_{s'})\mathbf{V}.$$

From the above equation, we have

$$\mathbf{H} - \mathbf{H}^{s,s'} = \left(\mathbf{A} - (\mathbf{T}_s^T\mathbf{T}_s\mathbf{A}\mathbf{T}_{s'}^T\mathbf{T}_{s'})\right)\mathbf{V}.$$

Dataset	dim. model	dim. feedforward	learning rate	batchsize
SelfRegulationSCP2	512	2048	0.001	16
PEMS-SF	512	2048	0.001	16
UWaveGestureLibrary	512	2048	0.001	16
EthanolConcentration	64	256	0.001	128
Handwriting	128	256	0.001	128
Heartbeat	64	256	0.001	128
JapaneseVowels size	128	256	0.001	128
SelfRegulationSCP1 size	128	256	0.001	128
SpokenArabicDigits size	64	256	0.001	128
FaceDetection size	128	256	0.001	128

Table 6: Hyperparameter configuration for UEA time series classification task.

Dataset	embedding dim	hidden dim	head dim	learning rate
listops	64	128	32	0.0001
retrieval	64	128	32	0.0001
text	64	128	32	0.0001

Table 7: Hyperparameter configuration for LRA task.

From the inequality with the Frobenius norm, we have

$$\|\mathbf{H} - \mathbf{H}^{s,s'}\|_F \leq \|\mathbf{A} - \mathbf{T}_s^T \mathbf{T}_s \mathbf{A} \mathbf{T}_{s'}^T \mathbf{T}_{s'}\|_F \|\mathbf{V}\|_2.$$

Therefore, it suffices to approximate the upper bound $\|\mathbf{A} - \mathbf{T}_s^T \mathbf{T}_s \mathbf{A} \mathbf{T}_{s'}^T \mathbf{T}_{s'}\|_F$. Let $\mathbf{A}^{s,s'} = \mathbf{T}_s^T \mathbf{T}_s \mathbf{A} \mathbf{T}_{s'}^T \mathbf{T}_{s'}$ and obviously $\mathbf{A}^{s,s'}$ contains blocks matrices of the same values. We can rewrite $\mathbf{A}$ and $\mathbf{A}^{s,s'}$ as block matrices of size $s \times s'$: $\mathbf{A} = [\mathbf{A}_{m,n}]_{m,n}$ and $\mathbf{A}^{s,s'} = [\mathbf{A}^{s,s'}_{m,n}]_{m,n}$ where $m = 0, 1, ..., \text{qlen}/s$, and $n = 0, 1, ..., \text{klen}/s'$. Note that all elements of $\mathbf{A}^{s,s'}_{m,n}$ have an identical value to the average of all elements of the sub-matrix $\mathbf{A}_{m,n}$.
Now we can decompose the above quantity into a sum of Frobenius norms:

$$\|\mathbf{A} - \mathbf{T}_s^T \mathbf{T}_s \mathbf{A} \mathbf{T}_{s'}^T \mathbf{T}_{s'}\|_F^2 = \sum_{m,n} \|\mathbf{A}_{m,n} - \mathbf{A}^{s,s'}_{m,n}\|_F^2.$$

Recall that from the hypothesis, we have

$$|\mathbf{A}_{i,j} - \mathbf{A}_{i\pm1,j}| \leq \delta, \ |\mathbf{A}_{i,j} - \mathbf{A}_{i,j\pm1}| \leq \delta. \tag{18}$$

Then, by applying Popoviciu's inequality, we have

$$\text{Var}\,[X] \leq \frac{(M-m)^2}{4},$$

where $m = \inf X$ and $M = \sup X$. Since matrix is finite, the infimum and the maximum become the maximum and minimum respectively. By Assumption 18, we can approximate the upper bound of $M - m$ as follows:

$$(M-m)^2 \leq (s + s' - 2)^2 \delta^2.$$

Integrate the sum, we find that

$$\|\mathbf{A} - \mathbf{A}^{s,s'}\|_F^2 \leq \frac{\text{qlen}}{s} \frac{\text{klen}}{s'} (s + s' - 2)^2 \frac{\delta^2}{4}.$$

When we plug in $\text{klen} = \text{qlen} = N$, we obtain a simpler version:

$$\|\mathbf{A} - \mathbf{A}^{s,s'}\|_F \leq \frac{s + s' - 2}{\sqrt{ss'}} \frac{N\delta}{2}.$$

As a consequence, we obtain the conclusion of the theorem.

C Additional Experiments

C.1 Combining MrsHA with other efficient attentions

In this section, we combine the proposed MrsHA architecture with other efficient attention mechanisms to demonstrate MrsHA can be combined with other efficient transformer to reduce memory and computation requirements. We run our experiments on 5 efficient transformer including

Table 8: Accuracy (%) of the models that combined MrsHa with other efficient transformers versus the accuracy of the original efficient transformers on the UEA Time Series Classification task. The combined models are indicated by the prefix "Mrs", results are averaged over 5 seeds (In this experiment, we use the set of scales $\boldsymbol{s} = [1, 1, 2, 2, 4, 4, 8, 8]$).

DATASET / MODEL	MRSLINFORMER (LINFORMER)	MRSLINEAR (LINEAR)	MRSFMM (FMM)	MRSPERFORMER (PERFORMER)	MRSLUNA (LUNA)
ETHANOLCONCENTRATION	32.70 (32.95)	35.49 (34.35)	34.47 (34.22)	33.21 (33.59)	33.71 (33.59)
FACEDETECTION	68.83 (68.53)	68.91 (68.46)	69.53 (68.97)	68.90 (68.96)	68.64 (68.92)
HANDWRITING	32.55 (32.47)	32.98 (33.29)	33.02 (31.57)	30.51 (30.47)	32.94 (32.32)
HEARTBEAT	75.12 (75.12)	75.45 (76.75)	76.42 (75.77)	75.61 (75.93)	75.61 (75.77)
JAPANESEVOWELS	98.56 (98.65)	99.46 (99.28)	99.64 (99.64)	99.01 (99.19)	99.46 (99.46)
PEMS-SF	87.67 (86.51)	83.43 (79.96)	86.9 (82.47)	84.59 (84.59)	81.31 (81.12)
SELFREGULATIONSCP1	92.61 (91.47)	91.13 (91.24)	93.06 (92.26)	91.13 (91.01)	91.24 (90.78)
SELFREGULATIONSCP2	55.19 (57.41)	54.26 (53.33)	54.82 (54.44)	54.44 (55.19)	55.74 (55.37)
SPOKENARABICDIGITS	98.91 (98.88)	98.76 (98.86)	99.48 (99.38)	99.02 (98.84)	99.03 (99.06)
UWAVEGESTURELIBRARY	86.25 (85.62)	82.19 (80.63)	86.46 (85.73)	85.10 (85.00)	86.25 (87.08)
AVERAGE ACCURACY	**72.84** (72.76)	**72.21** (71.61)	**73.38** (72.45)	72.15 (**72.28**)	**72.39** (72.35)

Table 9: Accuracy (%) of models that combined MrsHa with other efficient transformers versus accuracy of the original efficient transformers (in the parentheses) in LRA task. The combined models are indicated by the prefix "Mrs", results are averaged over 5 seeds (In this experiment, we use the set of scales $\boldsymbol{s} = [1, 2]$).

DATASET / MODEL	MRSLINFORMER (LINFORMER)	MRSLINEAR (LINEAR)	MRSFMM (FMM)	MRSPERFORMER (PERFORMER)	MRSLUNA (LUNA)
LISTOPS	36.93 (36.59)	36.97 (36.90)	37.77 (30.67)	37.12 (36.41)	37.03 (37.02)
RETRIEVAL	78.38 (78.17)	81.36 (81.13)	81.65 (80.91)	78.93 (78.67)	74.54 (69.55)
TEXT	57.39 (56.50)	66.57 (65.69)	68.39 (68.57)	65.20 (65.17)	64.51 (66.13)
AVERAGE ACCURACY	**57.57** (57.09)	**61.63** (61.24)	**62.60** (60.05)	**60.42** (60.08)	**58.69** (57.57)

Table 10: The resutls of the comparison between MrsFT-Transformer and FT-Transformer. The ↑ symbol denotes that the metric being reported is accuracy (the higher the better), the ↓ symbol denotes that the metric being reported is root mean square error (the lower the better).

DATASET / MODEL	FT-TRANSFORMER	MRSFT-TRANSFORMER
CALIFORNIA HOUSING ↓	**0.4671**	0.468
ADULT INCOME ↑	85.76	**85.87**
HELENA ↑	37.99	**38.23**
JANNIS ↑	72.46	**72.54**
HIGGS ↑	**72.50**	72.44
ALOI ↑	95.48	**95.52**
EPSILON ↑	**89.65**	89.58
YEAR ↓	8.905	**8.904**
COVERTYPE ↑	96.67	**96.84**
YAHOO ↓	**0.7567**	0.7586
MICROSOFT ↓	0.7474	**0.7468**

Linformer (Wang et al., 2020a), Linear transformer (Katharopoulos et al., 2020), FMM transformer (Nguyen et al., 2021), Performer (Choromanski et al., 2021) and Luna transformer (Ma et al., 2021). All experiments settings in this section follows directly from subsections 3.1 and 3.2 unless stated otherwise.

C.1.1 UEA TIME SERIES CLASSIFICATION

Results in Table 8 presents the accuracy of the combined and original models on the UEA Time Series Classification task. All the efficient transformers in this experiment either maintain comparable performance or experience a boost in average accuracy when combined with MrsHA.

C.1.2 LONG RANGE ARENA

In Listops experiments, we increase the number of training step from 5000 to 15000 to ensure convergence for all models. Table 9 further consolidates the advantage of the proposed MrsHA architecture. In fact, all the combined models obtain better average accuracy than the original models in the LRA task.

C.2 TABULAR DATA

We include a diverse set of 11 tabular dataset for our benchmarking: California Housing (Kelley Pace & Barry, 1997), Adult (Kohavi, 1996), Helena (Guyon et al., 2019), Jannis (Guyon et al., 2019), Higgs (Baldi et al., 2014), ALOI (Geusebroek et al., 2005), Epsilon (EP, simulated physics experiments), Year (Bertin-Mahieux et al., 2011), Covertype (Blackard & Dean, 1999), Yahoo (Chapelle & Chang, 2011), Microsoft (Qin & Liu, 2013). We follow all the train settings and use the default set of hyperparameters used in paper (Gorishniy et al., 2021) for all models. For simplicity, we omit the ensemble step from paper (Gorishniy et al., 2021). We report average accuracy over 5 random seed for both FT-Transformer (Gorishniy et al., 2021) and the combined model of MrsHA and FT-Transformer, which we denote MrsFT-Transformer.

Table 10 evidently shows that our combined model obtained better results in 7 over 11 tasks, while other tasks maintain comparable performance. This result consolidates the benefit of combining MrsHA with other transformer models in a diverse set of tasks.

C.3 EFFICIENCY WHEN COMBINING MRSHA WITH OTHER EFFICIENT TRANSFORMER

For illustration, we present FLOP and memory reduction ratios of train and test phases of our MrsFMM transformer comparing to the original FMM transformer for LRA retrieval task in Figure 5. Our model saves up to 35% of the original FLOP and has lower memory footprint, less than 65% and 85% of the original model for training and testing phases, respectively.

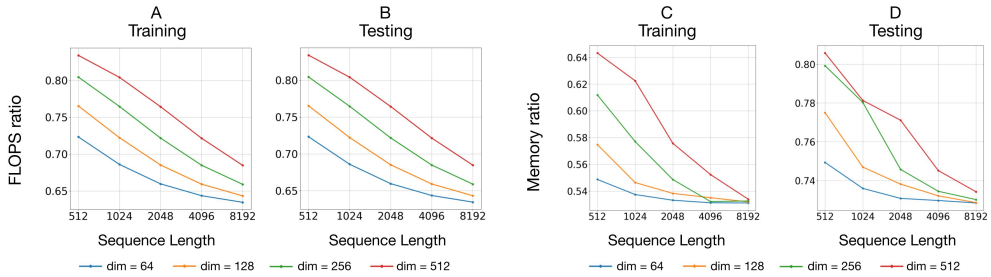

Figure 5: Training-inference FLOP ratios (A-B) and memory ratios (C-D) between the MrsFMM transformer and FMM transformer across different model dimensions and sequence lengths on the LRA retrieval task ($\boldsymbol{s} = [1, 2]$).

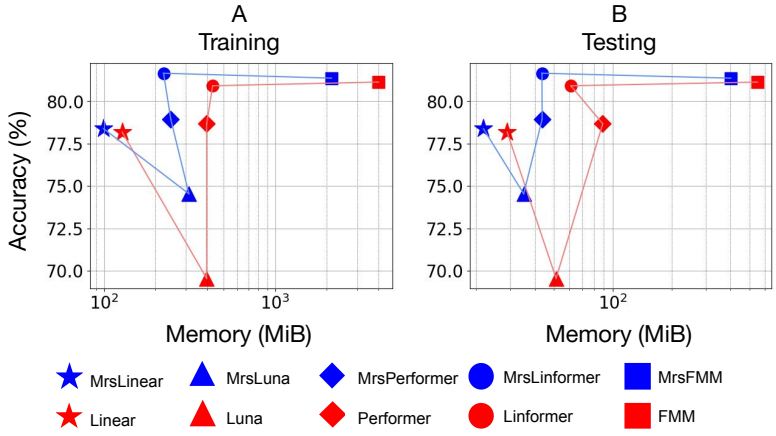

Figure 6: Scatter-plots for the relations between the memory usage and accuracy of the MrsHA-based efficient transformers vs. the baseline efficient transformers ($\boldsymbol{s} = [1, 2]$) trained for the LRA retrieval task.

C.4 SCATTER-PLOTS FOR THE RELATIONS BETWEEN THE MEMORY USAGE AND ACCURACY OF THE MRSHA-BASED EFFICIENT TRANSFORMERS VS. THE BASELINE EFFICIENT TRANSFORMERS

We have included the scatter-plots for the relations between the memory usage and accuracy of the MrsLuna, MrsLinformer, MrsPerformer, MrsLinear, and MrsFMM vs. the Luna, Linformer,

Performer, Linear, and FMM baselines trained for the LRA retrieval task in Figure 6. We observe that in both train and test cases, the scatter-plots of our MrsHA-based models are above and on the left of the scatter-plots of the baselines, suggesting that our MrsHA-based models are more memory efficient while achieving comparable or better accuracies than the baseline models.