MRSFORMER: TRANSFORMER WITH MULTIRESOLUTION-HEAD ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose the Transformer with Multiresolution-head Attention (MrsFormer), a class of efficient transformers inspired by the multiresolution approximation (MRA) for approximating a signal f using wavelet bases. MRA decomposes a signal into components that lie on orthogonal subspaces at different scales. Similarly, MrsFormer decomposes the attention heads in the multi-head attention into fine-scale and coarse-scale heads, modeling the attention patterns between tokens and between groups of tokens. Computing the attention heads in MrsFormer requires significantly less computation and memory footprint compared to the standard softmax transformer with multi-head attention. We analyze and validate the advantage of MrsFormer over the standard transformers on a wide range of applications including image and time series classification.

1 INTRODUCTION

The transformer architectures (Vaswani et al., 2017) is popularly used in natural language processing (Devlin et al., 2018; Al-Rfou et al., 2019; Dai et al., 2019; Child et al., 2019; Raffel et al., 2020; Baevski & Auli, 2019; Brown et al., 2020; Dehghani et al., 2018), computer vision (Dosovitskiy et al., 2021; Liu et al., 2021; Touvron et al., 2020; Ramesh et al., 2021; Radford et al., 2021; Arnab et al., 2021; Liu et al., 2022; Zhao et al., 2021; Guo et al., 2021), speech processing (Gulati et al., 2020; Dong et al., 2018; Zhang et al., 2020; Wang et al., 2020b), and other relevant applications (Rives et al., 2021; Jumper et al., 2021; Chen et al., 2021; Zhang et al., 2019; Wang & Sun, 2022). Transformers achieve state-of-the-art performance in many of these practical tasks, and the results get better with larger model size and increasingly long sequences. For example, the text generating model in (Liu et al., 2018a) processes input sequences of up to 11,000 tokens of text. Applications involving other data modalities, such as music (Huang et al., 2018) and images (Parmar et al., 2018), can require even longer sequences. Lying at the heart of transformers is the self-attention mechanism, an inductive bias that connects each token in the input through a relevance weighted basis of every other tokens to capture the contextual representation of the input sequence (Cho et al., 2014; Parikh et al., 2016; Lin et al., 2017; Bahdanau et al., 2014; Vaswani et al., 2017; Kim et al., 2017). The capability of self-attention to attain diverse syntactic and semantic representations from long input sequences accounts for the success of transformers in practice (Tenney et al., 2019; Vig & Belinkov, 2019; Clark et al., 2019; Voita et al., 2019a; Hewitt & Liang, 2019).

1.1 SELF-ATTENTION

The self-attention mechanism learns long-range dependencies via parallel processing of the input sequence. For a given input sequence $\mathbf{X} := [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N\times D_x}$ of N feature vectors, the self-attention transforms $\mathbf{X}$ into the output sequence $\mathbf{H} := [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_N]^\top \in \mathbb{R}^{N\times D_v}$ as follows

$$\mathbf{H} = \mathrm{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D}}\Big)\mathbf{V} := \mathbf{A}\mathbf{V}, \tag{1}$$

where $\boldsymbol{Q} := [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_N]^\top, \mathbf{K} := [\boldsymbol{k}_1, \cdots, \boldsymbol{k}_N]^\top$, and $\mathbf{V} := [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N]^\top$ are the projections of the input sequence $\mathbf{X}$ into three different subspaces spaned by $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{D\times D_x}$, and $\mathbf{W}_V \in \mathbb{R}^{D_v\times D_x}$, i.e. $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q^\top, \mathbf{K} = \mathbf{X}\mathbf{W}_K^\top, \mathbf{V} = \mathbf{X}\mathbf{W}_V^\top$. Here, in the context of transformers, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are named the query, key, and value matrices, respectively. The softmax function is applied to row-wise. The matrix $\mathbf{A} = \mathrm{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D}}\Big) \in \mathbb{R}^{N\times N}$ is the attention matrix, whose component a_{ij} for i, $j = 1, \cdots, N$ are the attention scores. The structure of the attention matrix $\mathbf{A}$ after training from data determines the ability of the self-attention to capture contextual representation for each token. Eqn. (1) is also called the scaled dot-product or softmax attention. In our paper, we call a transformer that uses this attention the softmax transformer.

1.2 Multi-head Attention (MHA)

In MHA, multiple heads are concatenated to compute the final output. Let H be the number of heads and $\mathbf{W}_O^{\text{multi}} = \left[\mathbf{W}_O^{(1)}, \dots, \mathbf{W}_O^{(H)}\right] \in \mathbb{R}^{D_v \times HD_v}$ be the projection matrix for the output where $\mathbf{W}_O^{(1)}, \dots, \mathbf{W}_O^{(H)} \in \mathbb{R}^{D_v \times D_v}$. The multi-head attention is defined as

$$\begin{aligned}\text{MultiHead}(\{\mathbf{H}\}_{h=1}^{H}) &= \text{Concat}(\mathbf{H}^{(1)}, \dots, \mathbf{H}^{(H)})\mathbf{W}_O^{\text{multi}\top} \\ &= \sum_{h=1}^{H} \mathbf{H}^{(h)}\mathbf{W}_O^{h\top} = \sum_{h=1}^{H} \mathbf{A}^{(h)}\mathbf{V}^{(h)}\mathbf{W}_O^{(h)\top}. \end{aligned} \tag{2}$$

The MHA extends the single-head attention and enables transformers to capture more diverse attention patterns. However, it has been shown that attention heads in MHA are redundant and tend to learn similar attention patterns, thus limiting the representation capacity of the model. Furthermore, additional heads increase the computational and memory costs, which becomes a bottleneck in scaling up transformers for very long sequences in large-scale practical tasks.

1.3 Contribution

Levaraging the idea of the multiresolution approximation (MRA) (Mallat, 1999; 1989; Crowley, 1981), we propose a class of efficient and flexible transformers, namely the Transformer with Multiresolution-head Attention (MrsFormer). At the core of MrsFormer is to use the novel Multiresolution-head Attention (MrsHA) that computes the approximation of the output $\mathbf{H}^h$, $h = 1, \dots, H$, of attention heads in MHA at different scales for saving computation and reducing the memory cost of the model. Our contribution is three-fold:

1. We derive the approximation of an attention head at different scales via two steps: i) Directly approximating the output sequence $\mathbf{H}$, and ii) approximating the value matrix $\mathbf{V}$, i.e. the dictionary that contains bases of $\mathbf{H}$.
2. We develop MrsHA, a novel MHA whose attention heads approximate the output sequences $\mathbf{H}^h$, $h = 1, \dots, H$, at different scales. We then propose MrsFormer, a new class of transformers that use MrsHA in their attention layers.
3. We empirically verify that the MrsFormer helps reduce the head redundancy and achieves better efficiency than the baseline softmax transformer while attaining comparable accuracy to the baseline.

Organization: We structure this paper as follows: In Section 2, we derive the approximation for the output sequence $\mathbf{H}^h$, $h = 1, \dots, H$, at different scales and propose the MrsHA and MrsFormer. In Section 3 and 4, we empirically validate and analyze the advantages of the MrsFormer over the baseline softmax transformer. We discuss related work in Section 5. The paper ends up with concluding remarks. More experimental details are provided in the Appendix.

2 Transformer with Multiresolution-head Attention

2.1 Background: Wavelet Transform and Multiresolution Approximations

The wavelet transform uses time-frequency atoms with different time supports to analyze the structure of a signals. In particular, it decomposes signals over dilated and translated copies of a fixed function φ. A dictionary of time-frequency atoms is obtained by scaling φ by s and translating it by t:

$$\mathcal{B} = \left\{\varphi_t^s = \frac{1}{\sqrt{s}}\varphi\left(\frac{x-t}{s}\right)\right\}_{t\in\mathbb{R}, s\in\mathbb{R}+}. \tag{3}$$

Here, s controls the dilation, i.e., the scale, and t controls the location, e.g., the time. Using this dictionary of time-frequency atoms, a signal $f \in \mathbf{L}^2(\mathbb{R})$ can be expanded in the following form:

$$f = \int_0^{+\infty}\int_{-\infty}^{+\infty} \alpha_t^s \varphi_t^s(x)\, dt ds. \tag{4}$$

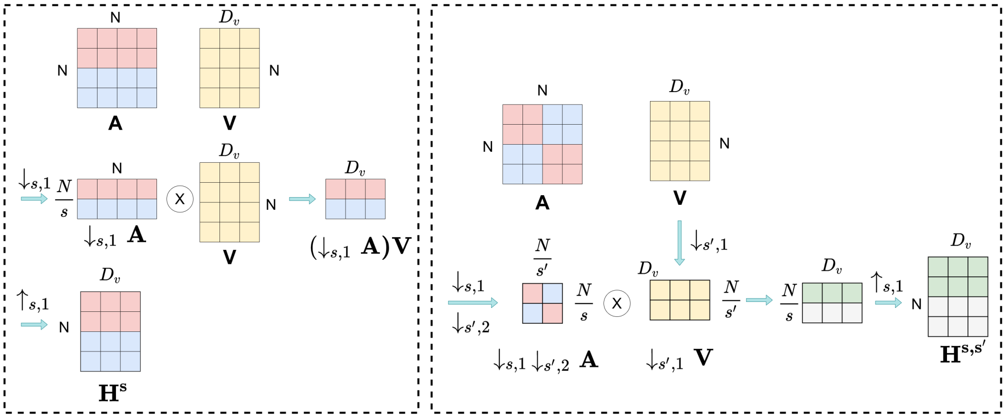

Figure 1: Illustration of Eqn. 11 (Left) and Eqn. 14 (Right).

The wavelet transform then maps the signal f to the coefficient α_t^s as follows

$$\alpha_t^s = \langle f, \varphi_t^s \rangle = \int_{-\infty}^{+\infty} f(x)(\varphi^*)_t^s \, dx, \tag{5}$$

where φ^* is the complex conjugate of φ. The coefficient α_t^s captures the measurement of the signal f at scale s and location t (Mallat, 1999).

2.2 Transformer with Multiresolution-head Attention

2.2.1 First Level Approximation: Approximating the Output Sequence H at Different Scales

Let $\mathcal{B}^s = \{\boldsymbol{\varphi}_t^s \in \mathbb{R}^N\}$ be a set of orthogonal expansion functions for possible translations at scale s where $s = 1, 2, 4, \ldots, N$. For simplicity, we assume that the sequence length $N = 2^k$. The expansion functions $\boldsymbol{\varphi}_t^s$ are chosen to be the boxcar functions as follows

$$\boldsymbol{\varphi}_t^s[i] = \begin{cases} 1 & \text{if } st - s < i \le st \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

for $s \in \{1, 2, 4, \ldots, N\}$ and $t \in \{1, \ldots, N/s\}$. At each scale s, we approximate the columns $\mathbf{H}[:, d]$, $d = 1, \ldots, D_v$, of the output sequence $\mathbf{H}$ as follows

$$\mathbf{H}[:, d] \approx \mathbf{H}^s[:, d] = \sum_{\boldsymbol{\varphi}_t^s \in \mathcal{B}^s} \alpha_{td}^s \boldsymbol{\varphi}_t^s, \tag{7}$$

where the coefficient α_{td}^s is computed as follows

$$\alpha_{td}^s = \frac{1}{s} \langle \boldsymbol{\varphi}_t^s, \mathbf{H}[:, d] \rangle. \tag{8}$$

Plug Eqn. (1) and Eqn. (8) into Eqn. (7), we obtain

$$\begin{aligned} \mathbf{H}[:, d] \approx \mathbf{H}^s[:, d] = \sum_{\boldsymbol{\varphi}_t^s \in \mathcal{B}^s} \frac{1}{s} \langle \boldsymbol{\varphi}_t^s, \mathbf{H}[:, d] \rangle \boldsymbol{\varphi}_t^s &= \sum_{t=1}^{N/s} \left(\frac{1}{s} \sum_{i=st-s+1}^{st} \mathbf{H}[i, d]) \right) \boldsymbol{\varphi}_t^s \\ &= \sum_{t=1}^{N/s} \left(\left(\frac{1}{s} \sum_{i=st-s+1}^{st} \mathbf{A}[i, :] \right) \mathbf{V}[:, d] \right) \boldsymbol{\varphi}_t^s \qquad (9) \\ &= \uparrow_{s,1} ((\downarrow_{s,1} \mathbf{A}) \mathbf{V}[:, d]). \qquad (10) \end{aligned}$$

Here, we employ the notations for downsampling and upsampling from signal processing. In particular, $\downarrow_{s,\ell}$ denotes the average pooling by the factor s along the ℓ^{th} dimension, and $\uparrow_{s,\ell}$ denotes the nearest-neighbor interpolation by the factor s along the ℓ^{th} dimension. Applying Eqn. (10) for $d = 1, \ldots, D_v$, we achieve the approximation of $\mathbf{H}$ at scale s as follows:

$$\mathbf{H} \approx \mathbf{H}^s = \uparrow_{s,1} ((\downarrow_{s,1} \mathbf{A}) \mathbf{V}). \tag{11}$$

An illustration of Eqn. 11 is given in Figure. 1 (Left).

Remark 1 (Approximating the columns of $\mathbf{H}$ independently) *As pointed out in (Nguyen et al., 2022), the features $\mathbf{H}[:, d]$ in the ouput sequence $\mathbf{H}$, as well as the features $\mathbf{V}[:, d]$ in the value matrix $\mathbf{V}$, $d = 1, \dots, D_v$, in the softmax attention are independent due to the use of the unnormalized Gaussian kernels with the isotropic covariance. This finding justifies our approach of approximating the columns of $\mathbf{H}$ independently.*

Remark 2 (Group-to-token attention) *The downsampling $\downarrow_{s,1} \mathbf{A}$ of the matrix $\mathbf{A}$ in Eqn. (11) computes the attentions between groups of tokens and individual tokens in the sequence.*

2.2.2 Second Level Approximation: Approximating the Head Bases $\mathbf{V}$ at Different Scales

In Eqn. (11) that approximates the output sequence $\mathbf{H}$ at scale s by $\mathbf{H}^s$, we can further approximate the bases $\mathbf{V}$, i.e., the value matrix, by its approximation at scale s'. Following the derivation in Section 2.2.1 above, we can derive the approximation $\mathbf{V}^{s'}[:, d]$ for the d^{th} columns of $\mathbf{V}$ as follows

$$\mathbf{V}[:, d] \approx \mathbf{V}^{s'}[:, d] = \sum_{t'=1}^{N/s'} \left(\frac{1}{s'} \sum_{j=s't'-s'+1}^{s't'} \mathbf{V}[j, d] \right) \boldsymbol{\varphi}_{t'}^{s'}. \tag{12}$$

Plugging Eqn. (12) into Eqn. (9), we obtain the second level approximation of the head output $\mathbf{H}$:

$$\begin{aligned}
\mathbf{H}[:, d] &\approx \mathbf{H}^{s,s'}[:, d] \\
&= \sum_{t=1}^{N/s} \left(\left(\frac{1}{s} \sum_{i=st-s+1}^{st} \mathbf{A}[i, :] \right) \sum_{t'=1}^{N/s'} \left(\frac{1}{s'} \sum_{j=s't'-s'+1}^{s't'} \mathbf{V}[j, d] \right) \boldsymbol{\varphi}_{t'}^{s'} \right) \boldsymbol{\varphi}_t^s \\
&= \sum_{t=1}^{N/s} \left(\sum_{t'=1}^{N/s'} \left(\frac{1}{s's} \sum_{i=st-s+1}^{st} \mathbf{A}[i, :] \boldsymbol{\varphi}_{t'}^{s'} \right) \left(\sum_{j=s't'-s'+1}^{s't'} \mathbf{V}[j, d] \right) \right) \boldsymbol{\varphi}_t^s \\
&= \sum_{t=1}^{N/s} \left(\sum_{t'=1}^{N/s'} \left(\frac{1}{s's} \sum_{i=st-s+1}^{st} \sum_{j=s't'-s'+1}^{s't'} \mathbf{A}[i, j] \right) \left(\sum_{j=s't'-s'+1}^{s't'} \mathbf{V}[j, d] \right) \right) \boldsymbol{\varphi}_t^s \\
&= \uparrow_{s,1} \left((\downarrow_{s,1} \downarrow_{s',2} \mathbf{A})(\downarrow_{s',1} \mathbf{V}[:, d]) \right).
\end{aligned} \tag{13}$$

Same as above, by applying Eqn. (13) for $d = 1, \dots, D_v$, we achieve the full approximation of $\mathbf{H}$ at scale s of $\mathbf{H}$ and scale s' of $\mathbf{V}$ as follows:

$$\mathbf{H} \approx \mathbf{H}^{s,s'} := \uparrow_{s,1} \left((\downarrow_{s,1} \downarrow_{s',2} \mathbf{A})(\downarrow_{s',1} \mathbf{V}) \right). \tag{14}$$

An illustration of Eqn. 14 is given in Figure. 1 (Right). Given the approximation $\mathbf{H}^{s,s'}$ of the attention matrix $\mathbf{H}$, we have the following upper bound on the approximation error.

Theorem 1 *Assume that $\delta > 0$ is chosen such that the attention matrix $\mathbf{A}$ satisfies the following inequalities $|\mathbf{A}_{i,j} - \mathbf{A}_{i\pm1,j}| \leq \delta$, $|\mathbf{A}_{i,j} - \mathbf{A}_{i,j\pm1}| \leq \delta$ for all $1 \leq i, j \leq N$. Then, we obtain that*

$$\|\mathbf{H} - \mathbf{H}^{s,s'}\|_F \leq \frac{(s + s' - 2)N\delta}{\sqrt{ss'}} \|\mathbf{V}\|_2,$$

where $\|.\|_F$ denotes the Frobenius norm and $\|.\|_2$ denotes the spectral norm of a matrix.

Proof of Theorem 1 is in Appendix B. The result of Theorem 1 shows that the approximation matrix $\mathbf{H}^{s,s'}$ approximates $\mathbf{H}$ exactly when $s = s' = 1$, which is true. In the coarsest scale when $s = s' = N$, the upper bound achieves the maximum value $(N - 1)\delta\|\mathbf{V}\|_2$.

Remark 3 (Group-to-group attention) *The downsampling $\downarrow_{s,1}\downarrow_{s',2} \mathbf{A}$ of the matrix $\mathbf{A}$ in Eqn. (14) computes the attentions between groups of tokens and groups of tokens in the sequence.*

2.2.3 Efficient Downsampling of the Attention Matrix $\mathbf{A}$

As shown in Eqn. (1), $\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{D}}\right)$. Since the softmax function needs access to the full matrix $\frac{\mathbf{QK}^\top}{\sqrt{D}}$, downsampling $\mathbf{A}$ via average pooling still requires to compute the full product $\frac{\mathbf{QK}^\top}{\sqrt{D}}$

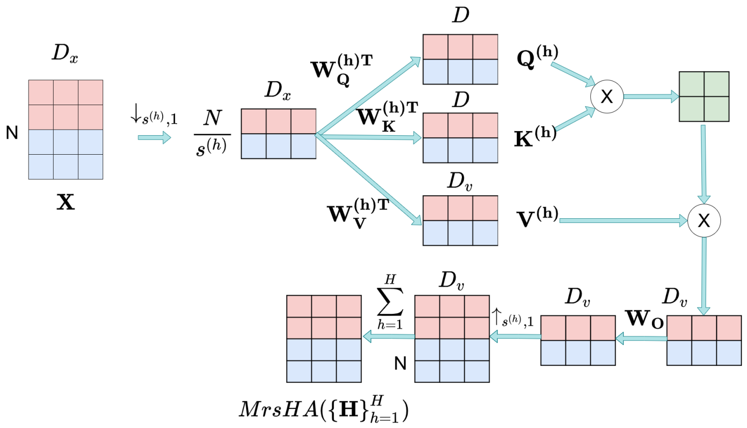

Figure 2: Illustration of Eqn. 17.

first. In order to avoid this redundant computation, we propose to compute the lower bound of this average pooling (due to the convexity of the exponential in the softmax function). In particular, we approximate the downsampling of $\mathbf{A}$ as follows:

$$\downarrow_{s,1}\downarrow_{s',2} \mathbf{A} \approx \text{softmax}\Big(\frac{\downarrow_{s,1}\downarrow_{s',2} (\mathbf{Q}\mathbf{K}^\top)}{\sqrt{D}}\Big) = \text{softmax}\Big(\frac{(\downarrow_{s,1} \mathbf{Q})(\downarrow_{s',1} \mathbf{K})^\top}{\sqrt{D}}\Big). \tag{15}$$

2.2.4 Transformer with Multiresolution-head Attention: Each Head Approximates the Attention at a Different Scale

In this section, we formally define our Multiresolution-head Attention (MrsHA) and Transformer with Multiresolution-head Attention (MrsFormer). MrsHA combines Eqn. (14) and (15) to implement the approximation of the output sequences $\mathbf{H}^{(h)}$, $h = 1, \dots, H$, at different scales s and s'.

Definition 1 (Multiresolution-head Attention) *Let H be the number of heads and $\mathbf{W}_O^{multi} = \Big[\mathbf{W}_O^{(1)}, \dots, \mathbf{W}_O^{(H)}\Big] \in \mathbb{R}^{D_v \times HD_v}$ be the projection matrix for the head outputs where $\mathbf{W}_O^{(1)}, \dots, \mathbf{W}_O^{(H)} \in \mathbb{R}^{D_v \times D_v}$. Given a set of scales $\{s^{(h)}, s'^{(h)}\}_{h=1}^H$ for the output $\mathbf{H}^{(h)}$ and the value matrix $\mathbf{V}^{(h)}$, $h = 1, \dots, H$, at each head, the MrsHA is an efficient attention mechanism that computes the approximation of $\mathbf{H}^{(h)}$ at scale $s^{(h)}$ using an approximation of $\mathbf{V}^{(h)}$ at scale $s'^{(h)}$ by the following attention formula:*

$$MrsHA(\{\mathbf{H}\}_{h=1}^H) = \sum_{h=1}^{H} \uparrow_{s^{(h)},1} \left(\text{softmax}\Big(\frac{(\downarrow_{s^{(h)},1} \mathbf{Q})(\downarrow_{s'^{(h)},1} \mathbf{K})^\top}{\sqrt{D}}\Big)(\downarrow_{s'^{(h)},1} \mathbf{V}^{(h)})\right) \mathbf{W}_O^{(h)\top}. \tag{16}$$

The MrsFormer is the class of transformers that use the MrsHA in their attention layers.

Remark 4 (Downsampling Q, K, and V) *Downsampling $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ can be efficiently implemented by downsampling the input sequence $\mathbf{X}$ before projecting it into the query matrix $\mathbf{Q}$, the key matrix $\mathbf{K}$, and the value matrix $\mathbf{V}$ via the linear transformations $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$, respectively. Eqn. (16) of the MrsHA then becomes*

$$\begin{aligned} &MrsHA(\{\mathbf{H}\}_{h=1}^H) \\ &= \sum_{h=1}^{H} \uparrow_{s^{(h)},1} \left(\text{softmax}\Big(\frac{(\downarrow_{s^{(h)},1} \mathbf{X}\mathbf{W}_Q^{(h)\top})(\downarrow_{s'^{(h)},1} \mathbf{X}\mathbf{W}_K^{(h)\top})^\top}{\sqrt{D}}\Big)(\downarrow_{s'^{(h)},1} \mathbf{X}\mathbf{W}_V^{(h)\top})\right) \mathbf{W}_O^{(h)\top}. \end{aligned} \tag{17}$$

An illustration of Eqn. 17 is given in Figure. 2.

Remark 5 (Choosing $s^{(h)}$ and $s'^{(h)}$) *$s^{(h)}$ and $s'^{(h)}$ are hyperparameters that can be tuned for each head. In our experiments, we use $s^{(h)} = s'^{(h)} = 2^{k^{(h)}}$, where $k^{(h)}$ is an integer.*

Remark 6 (Choosing the expansion functions φ_t^s and 1-D convolution) *In order to derive the MrsHA in Eqn. (16), we have chosen the expansion functions φ_t^s to be the boxcar functions. Other expansion functions, such as the wavelet bases or the triangular functions, can be used to derive different forms of the MrsHA. In a general case, the average pooling and the nearest-neighbor interpolation in Eqn. (16) and (17) can be replaced by the 1-D convolution operators with φ_t^s as the corresponding filters.*

3 Experimental Results

In this section, we empirically justify the advantages of our propsed MrsFormer model. We compare the performance of the MrsFormer with the baseline softmax transformer, the MRA-2 (Zeng et al., 2022), and the MRA-2-s (which is the sparse version of the MRA-2) on various benchmarks. Unlike our method, the MRA-2 and MRA-2-s perform multiresolution analysis for each head by approximating the attention matrix by blocks of different scales, while the MrsHA in our MrsFormer computes the approximation of each head $\mathbf{H}^h$ at a specific scale. The benchmarks studied in our experiments include 10 tasks from the UEA time series classification dataset (Bagnall et al., 2018), 3 tasks from Long Range Arena (Tay et al., 2021b) (LRA) benchmark, and ImageNet image classification task (Russakovsky et al., 2015). In addition, we also study the performance of the MrsHA when being combined with other attention mechanism such as the linear attention (Katharopoulos et al., 2020), the MRA-2 attention, and the MRA-2-s attention (Zeng et al., 2022). We aim to show that: (i) the MrsFormer can achieve better or comparable accuracy over the baseline softmax, MRA-2, and MRA-2-s transformers; (ii) the MrsFormer saves significant amount of FLOPs and memory compared to the baseline softmax transformer, and this advantage grows with the sequence length; (iii) the MrsHA can be combined with other attentions to achieve similar or better performance with better efficiency; and (iv) the MrsFormer reduces redundancy between heads comparing to the softmax baseline.

In our experiment, we keep the hyperparameters the same for all models for fair comparisons. All of our results are averaged over 5 runs with different seeds.

3.1 UEA Time Series Classification

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).

Results. We summarize the results in Table 1. The MrsFormer achieves bettter test accuracy than the baseline softmax transformer for 5 out of 10 tasks while being much more efficient. Among these tasks, the MrsFormer outperforms the baseline by at least 1% accuracy. For the remaining tasks, besides Handwriting, our model maintains an accuracy gap less than 0.8% compared to the baseline. Our model gets the best accuracy for 4 out of the 10 tasks. In addition, it achieves second best accuracy for 4 out of the remaining tasks. The MrsFormer achieves the average accuracy across all tasks. Note that among 8 heads at each layer, our model computes 6 of them with the size of only $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{16}$, $\frac{1}{16}$, $\frac{1}{64}$ and $\frac{1}{64}$ of the size of the corresponding heads in the baseline softmax transformer. Thus, the MrsFormer has a significant smaller FLOPS and memory usage compared to the baseline.

3.2 Long Range Arena

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.

Results. Table 2 summarizes our results. Although being an approximation of the softmax attention, it is evidently from Table 2 that MrsFormer can consistently achieve better than or comparable accuracy as the baseline softmax attention on the LRA tasks. The MRA-2 and MRA-2-s models (Zeng et al., 2022) are also included for comparison. Our MrsFormer's performance is comparable with these MRA baselines. Overall, the MrsFormer yields the best average accuracy across the LRA tasks.

Table 1: Accuracy (%) of the MrsFormer vs. the baseline softmax transformer on the UEA Time Series Classification task averaged over 5 seeds. The best model for each task is highlighted in bold, while the second best one is underlined. We also include the reported results for the softmax transformer from (Wu et al., 2022) and (Zerveas et al., 2021) (in parentheses). The MrsFormer attains the best average accuracy across all tasks while being much more efficient than the baseline softmax transformers.

Dataset / Model	Baseline softmax	MrsFormer	MRA-2	MRA-2-s
EthanolConcentration	32.08 (33.70)	**35.87**	34.35	34.48
FaceDetection	**68.70** (68.10)	68.23	68.28	68.24
Handwriting	**32.08** (30.50)	30.24	29.49	29.68
Heartbeat	75.77 (77.60)	**78.86**	77.24	78.05
JapaneseVowels	**99.46** (99.40)	99.10	99.01	99.01
PEMS-SF	82.66 (82.10)	84.2	**86.13**	82.85
SelfRegulationSCP1	91.46 (**92.50**)	91.81	91.70	92.04
SelfRegulationSCP2	54.72 (53.90)	**56.85**	55.56	56.29
SpokenArabicDigits	**99.33** (99.30)	98.73	98.60	98.62
UWaveGestureLibrary	84.45 (85.60)	**86.67**	**86.67**	86.56
Average Accuracy	72.07(72.27)	**73.06**	72.70	72.58

Table 2: Accuracy (%) of the MrsFormer vs. the baseline softmax transformer averaged over 5 seeds. The best model for each task is highlighted in bold, while the second best one is underlined. The MrsFormer attains the best average accuracy across all tasks while being much more efficient than the baseline softmax transformers.

Dataset / Model	Baseline softmax	MrsFormer	MRA-2	MRA-2-s
Listops	36.84 (37.10)	**37.52**	37.10 (37.2)	37.05 (37.4)
Retrieval	79.52 (79.6)	**80.22**	78.88 (79.6)	79.76 (80.3)
Text	64.93 (65.2)	65.05	**65.09** (65.4)	64.43 (64.3)
Average Accuracy	60.43 (60.63)	**60.93**	60.36 (60.73)	60.41 (60.67)

Table 3: Accuracy (%) of the MrsFormer DeiT vs. the baseline softmax DeiT and the MRA-2-s DeiT on the ImageNet image classification task. The MrsFormer DeiT outperforms the MRA-2-s DeiT and yields comparable accuracy to the softmax DeiT.

Model name	Top-1 accuracy	Top-5 accuracy
Softmax DeiT	72.178	91.126
MRA-2-s DeiT	70.784	90.154
MrsFormer DeiT	71.342	90.566

3.3 ImageNet

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.

Results: We present our results in Table 3. The MrsFormer DeiT's top-1 accuracy is about 0.5% higher than MRA-2-s DeiT and is the closest model to the performance of the softmax DeiT baseline. The performance gap of less than 1% of MrsFormer DeiT is very promising for applying the MrsFormer-based model in large scale tasks to reduce the computational and memory cost while maintaining comparable performance with the baseline transformer.

4 Empirical Analysis

In this section, we use the models trained on the LRA retrieval task for our analysis.

4.1 Efficiency Analysis

We study the efficiency of MrsFormer over the baseline softmax transformer. Figure 3 demonstrates the reduction ratio of train and test flops of the MrsFormer over the softmax transformer. Although in this experiment, we only approximate one head with scale $s = 2$ and preserve the other head the

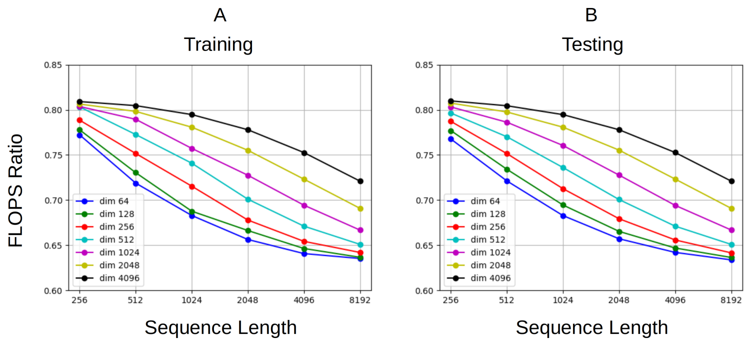

Figure 3: Training (A) and inference (B) FLOP ratios between the MrsFormer and the baseline softmax transformer across different model dimensions D (dim) and sequence lengths N on the LRA retrieval task. The MrsFormer requires fewer FLOPs compared to the baseline, and this advantage grows with the sequence length for very long sequences. Also, this efficient advantage of the MrsFormer holds for large-scale models with the large model dimension D.

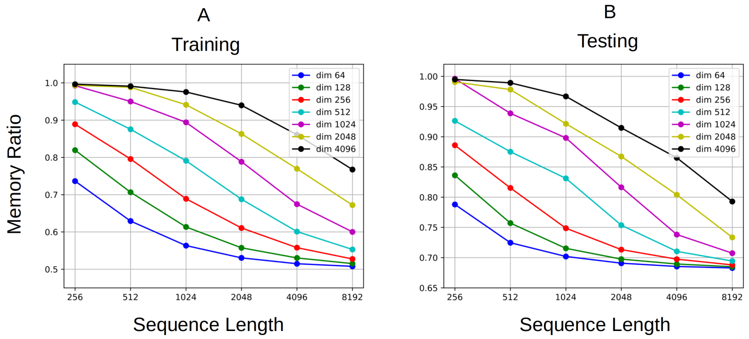

Figure 4: Training (A) and inference (B) memory ratios between the MrsFormer and the baseline softmax transformer across different model dimensions D (dim) and sequence lengths N on the LRA retrieval task. The MrsFormer requires fewer FLOPs compared to the baseline, and this advantage grows with the sequence length for very long sequences. Also, this efficient advantage of the MrsFormer holds for large-scale models with the large model dimension D.

Table 4: Layer-average mean and standard deviation of $\mathcal{L}_2$ distances between heads of the MrsFormer vs. the softmax transformer trained on the retrieval task. The MrsFormer obtains greater $\mathcal{L}_2$ distances between heads compared to the baseline, indicating that the MrsFormer captures more diverse attention patterns.

MetricModel	*Baseline Softmax*	MrsFormer
Mean	2.01	**2.68**
Std	0.39	**0.54**

same as in the baseline, the FLOP saving ratio over softmax attention still ranges from 18% up to more than 36% and grows with sequence length in both the training and testing phases. Figure 4 presents the memory saving ratio of the MrsFormer over the softmax transformer. This figure shows a similar trend of more memory saving when the sequence length increases. Our model achieves up to 49% and 31% decrease in memory usage in the training and testing phases, respectively. This indicates that our model scales well with long sequences and takes significantly less resource than the baseline softmax attention in both training and testing.

4.2 MrsFormer helps reduce head redundancy

To show that the MrsFormer captures more diverse attention patterns, we compare the average $\mathcal{L}_2$ distances between the heads of our trained MrsFormer model (on the retrieval task) and the softmax baseline. Table 4 reports the layer-average mean and standard deviation of distances between heads. Since the MrsFormer attains higher $\mathcal{L}_2$ distances, it reduces the risk of learning redundant heads compared to the softmax baseline.

4.3 Beyond the Softmax Attention: Combining MrsHA with other Attentions

The MrsHa is complementary to many other types of attentions. Therefore, a natural question arises is whether we can combine the MrsHa with other attentions besides the softmax attention? To answer this question, we combine the MrsHA with the MRA attention (Zeng et al., 2022) and the

Table 5: Accuracy (%) of the models that combined MrsHa with the MRA and linear attentions vs. the original MRA and linear transformers on the LRA retrieval task. The results are averaged over 5 seeds (In this experiment, we use the set of scales $\boldsymbol{s} = [1, 2]$).

MODEL NAME	COMBINED MODEL ACCURACY	ORIGINAL MODEL ACCURACY
MRA-2	**79.24**	78.88
MRA-2-S	**80.05**	79.76
LINEAR	**81.36**	81.13

linear attention (Katharopoulos et al., 2020) and train these combined models for the LRA retrieval tasks (Tay et al., 2021a) as in Section 3.2. The results are presented in Table 5. It is interesting to see from Table 5 that all combined models gain an improvment in test accuracy over the original models despite being an approximation. This observation suggests that the MrsHa can be applied to other attention mechanisms besides softmax to reduce computation and memory while maintaining the accuracy of the original models.

5 Related Work

Efficient Transformers To reduce the quadratic computational cost and memory usage of transformers, many efficient transformer models have been developed (Roy et al., 2021). Sparse transformers are a line of works in this branch, which explore and design the sparsity structure of attention matrix, resulting in more efficient models (Parmar et al., 2018; Liu et al., 2018b; Qiu et al., 2019; Child et al., 2019; Beltagy et al., 2020). Another class of efficient transformers is patterns integration, combining different attention patterns to cover a diverse and wide range of dependencies (Child et al., 2019; Ho et al., 2019). These patterns can be set as pre-specified or learnable during training, along with model parameters (Kitaev et al., 2020; Roy et al., 2021; Tay et al., 2020). In another attempt, multiple tokens can be accessed simultaneously with a side memory module, saving the cost of computing and memory storage(Lee et al., 2019; Sukhbaatar et al., 2019; Asai & Choi, 2020; Beltagy et al., 2020). In a different approach, observing that the attention matrices are low-rank, kernelization and low-rank approximation methods have been proposed to replace the softmax attention with more efficient attentions (Tsai et al., 2019; Wang et al., 2020a; Katharopoulos et al., 2020; Choromanski et al., 2021; Shen et al., 2021; Nguyen et al., 2021; Peng et al., 2021). From a signal processing perspective, wavelet-based and multiscale methods has been used lately to learn a multiresolution approximation of self-attention (Zeng et al., 2022; Fan et al., 2021; Tao et al., 2020; Li et al., 2022), which flexibly discover the coarse and fine attention patterns. Our approach decomposes the attention heads into coarse- and fine-scale heads, diversely modeling the dependencies between tokens and between group of tokens to reduce the computational and memory costs of the model in both training and testing.

Redundancy in Transformers Pre-trained transformers contain many redundant neurons and heads which can be pruned away for downstream tasks (Dalvi et al., 2020; Michel et al., 2019; Durrani et al., 2020). Studying the contextualized embeddings in these pre-trained networks shows the anisotropicity of the learned representation from these models under this redundancy (Mu & Viswanath, 2018; Ethayarajh, 2019). Multiple approaches have been proposed to reduce this redundancy and improve the efficiency of transformers, such as the knowledge distillation and sparse approximation Sanh et al. (2019); Sun et al. (2019); Voita et al. (2019b); Sajjad et al. (2020). Our MrsHA and MrsFormer represent the attention heads at different scales and are complementary to these methods.

6 Concluding Remarks

In this paper, we propose the MrsFormer, a class of efficient transformers that calculates the approximation of the attention heads at different scales using the Multiresolution-head Attention (MrsHA). The MrsFormer achieves better computational and memory cost than the corresponding softmax transformers baseline. Furthermore, the MrsFormer helps reduce the redundancy between attention heads and can be easily combined with other attention mechanisms. In the MrsFormer, we use the boxcar function to form a set of orthogonal expansion functions. It is natural to further develop the MrsFormer using other basis functions including the popular wavelets. Furthermore, in our derivation of the MrsHA and MrsFormer in Section 2.2, we employ the observation from (Nguyen et al., 2022) that the features $\mathbf{H}[:, d]$ in the output sequence $\mathbf{H}$ are independent. We leave the extenson of the MrsHA and MrsFormer to capture dependent output features as future work.

REFERENCES

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3159–3166, 2019.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6816–6826, 2021. doi: 10.1109/ICCV48922.2021.00676.

Akari Asai and Eunsol Choi. Challenges in information seeking qa: Unanswerable questions and paragraph retrieval. *arXiv preprint arXiv:2010.11915*, 2020.

Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByxZX20qFQ.

Anthony J. Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn J. Keogh. The UEA multivariate time series classification archive, 2018. *CoRR*, abs/1811.00075, 2018. URL http://arxiv.org/abs/1811.00075.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://www.aclweb.org/anthology/W19-4828.

James L Crowley. A representation for visual information. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA ROBOTICS INST, 1981.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. Analyzing redundancy in pretrained transformer models. *arXiv preprint arXiv:2004.04010*, 2020.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

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.

Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884–5888. IEEE, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. Analyzing individual neurons in pre-trained language models. *arXiv preprint arXiv:2010.02695*, 2020.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6824–6835, 2021.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://www.aclweb.org/anthology/D19-1275.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations*, 2018.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165. PMLR, 2020.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured attention networks. *arXiv preprint arXiv:1702.00887*, 2017.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pp. 3744–3753. PMLR, 2019.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4804–4814, 2022.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017. URL http://arxiv.org/abs/1703.03130.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=Hyg0vbWC-.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-1015.

Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/2c601ad9d2ff9bc8b282670cdd54f69f-Paper.pdf.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

Nikita Nangia and Samuel Bowman. ListOps: A diagnostic dataset for latent tree learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 92–99, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-4013. URL https://www.aclweb.org/anthology/N18-4013.

Tan Nguyen, Minh Pham, Tam Nguyen, Khai Nguyen, Stanley J Osher, and Nhat Ho. FourierFormer: Transformer meets generalized Fourier integral theorem. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tan M. Nguyen, Vai Suliafu, Stanley J. Osher, Long Chen, and Bao Wang. Fmmformer: Efficient and flexible transformer via decomposed near-field and far-field attention. *arXiv preprint arXiv:2108.02347*, 2021.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1244. URL https://www.aclweb.org/anthology/D16-1244.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4055–4064. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/parmar18a.html.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QtTKTdVrFBB.

Jiezhong Qiu, Hao Ma, Omer Levy, Scott Wen-tau Yih, Sinong Wang, and Jie Tang. Blockwise self-attention for long document understanding. *arXiv preprint arXiv:1911.02972*, 2019.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944, 2013.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021. doi: 10.1162/tacl_a_00353. URL https://www.aclweb.org/anthology/2021.tacl-1.4.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. Poor man's bert: Smaller and faster transformer models. *arXiv e-prints*, pp. arXiv–2004, 2020.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3531–3539, 2021.

Sainbayar Sukhbaatar, Edouard Grave, Guillaume Lample, Herve Jegou, and Armand Joulin. Augmenting self-attention with persistent memory. *arXiv preprint arXiv:1907.01470*, 2019.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*, 2019.

Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. Sparse Sinkhorn attention. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9438–9447. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/tay20a.html.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=qVyeW-grC2k.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=qVyeW-grC2k.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://www.aclweb.org/anthology/P19-1452.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. *arXiv preprint arXiv:1908.11775*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 63–76, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4808. URL https://www.aclweb.org/anthology/W19-4808.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://www.aclweb.org/anthology/P19-1580.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019b.

Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020a.

Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, et al. Transformer-based acoustic modeling for hybrid speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6874–6878. IEEE, 2020b.

Zifeng Wang and Jimeng Sun. TransTab: Learning Transferable Tabular Transformers Across Tables. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022.

Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing transformers with conservation flows. In *International Conference on Machine Learning*, 2022.

Zhanpeng Zeng, Sourav Pal, Jeffery Kline, Glenn M Fung, and Vikas Singh. Multi resolution analysis (mra) for approximate self-attention. In *International Conference on Machine Learning*, pp. 25955–25972. PMLR, 2022.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery; Data Mining*, KDD '21, pp. 2114–2124, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467401. URL https://doi.org/10.1145/3447548.3467401.

Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7829–7833. IEEE, 2020.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2019.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16259–16268, 2021.

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).

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.

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

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

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 $\mathbf{A}^{s,s'} = [\mathbf{A}_{m,n}^{s,s'}]_{m,n}$ where $m = 0, 1, ..., \text{qlen}/s$, and $n = 0, 1, ..., \text{klen}/s'$. Note that all elements of $\mathbf{A}_{m,n}^{s,s'}$ 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:

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

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.