A APPENDIX

A.1

Let $\mathcal{S}_N$ represent a sequence of N tokens, denoted as $w_{i}{}_{i=1}^{N}$, where w_i corresponds to the ith element in the sequence. The word embeddings for $\mathcal{S}_N$ are represented as $\mathbb{E}_N$, given by $\mathbf{x}_{i}{}_{i=1}^{N}$, where each $\mathbf{x}_i$ is a d-dimensional word embedding vector for token w_i, devoid of location information. Self-attention initially incorporates location information into the word embeddings, transforming them into queries, keys, and value representations.

$$\begin{aligned}\mathbf{Q}_m &= f_q(\mathbf{x}_m, m)\\ \mathbf{K}_n &= f_k(\mathbf{x}_n, n)\\ \mathbf{V}_n &= f_v(\mathbf{x}_n, n)\end{aligned} \tag{15}$$

where $\mathbf{Q}_m$, $\mathbf{K}_n$, and $\mathbf{V}_n$ are combined at positions m and n using functions f_q, f_k, and f_v, respectively. Typically, queries and keys are employed to calculate attentional weights, which are then used as coefficients for computing the weighted sum of VALUE representations to generate the output.

$$\begin{aligned}a_{m,n} &= \frac{\exp(\frac{\mathbf{Q}_m^\top \mathbf{K}_n}{\sqrt{d}})}{\sum_{n=1}^{N} \exp \frac{\mathbf{Q}_m^\top \mathbf{K}_n}{\sqrt{d}}}\\ \mathbf{o}_m &= \sum_{n=1}^{N} a_{m,n}\mathbf{V}_n\end{aligned} \tag{16}$$

Current position encoding methods for transformers primarily revolve around selecting the suitable function to formulate the equation 15. Let $\mathcal{P}$ represent the position coding operator introduced into the equation 15.

$$f_{t;t\in\{Q,K,V\}}(\mathbf{x}_i, i) := \mathbf{W}_{t;t\in\{Q,K,V\}}(\mathcal{P}(\mathbf{x}_i)) \tag{17}$$

Now bringing the HeterPos method into Equation 17, we can get

$$\begin{aligned}f_q(\mathbf{x}_m, m) &= \mathbf{W}_Q[\mathbf{x}^m, \sin(c\mathbf{x}^m e^{-m\ln(10000/d)}), \cos(c\mathbf{x}^m e^{-m\ln(10000/d)})]\mathbf{w}_{\text{PE}}^{(m)}\\ f_k(\mathbf{x}_n, n) &= \mathbf{W}_n[\mathbf{x}^n, \sin(c\mathbf{x}^n e^{-m\ln(10000/d)}), \cos(c\mathbf{x}^n e^{-n\ln(10000/d)})]\mathbf{w}_{\text{PE}}^{(n)}\\ f_v(\mathbf{x}_n, n) &= \mathbf{W}_V[\mathbf{x}^n, \sin(c\mathbf{x}^n e^{-n\ln(10000/d)}), \cos(c\mathbf{x}^n e^{-n\ln(10000/d)})]\mathbf{w}_{\text{PE}}^{(n)}\end{aligned} \tag{18}$$

We decompose $\mathbf{Q}m^\top\mathbf{K}n$ in Equation 18 using the analysis of Dai et al. (2019). To simplify the analysis, we do not consider the weight matrix and have:

$$\begin{aligned}&[\mathbf{x}^m, \sin(c\mathbf{x}^m e^{-m\ln(10000/d)}), \cos(c\mathbf{x}^m e^{-m\ln(10000/d)})]^\top \cdot\\ &\quad [\mathbf{x}^n, \sin(c\mathbf{x}^n e^{-n\ln(10000/d)}), \cos(c\mathbf{x}^n e^{-n\ln(10000/d)})]\end{aligned} \tag{19}$$

In this way, we can obtain Equation 7.

A.2 PROOF OF THEOREM 1

To better illustrate the derivation of our theorem, we first give the idea of the proof of Lemma 1. We give the following assumption:

- The activation function is L-Lipschitz,i.e., for any $x_1, x_2 \in \mathbb{R}^k, L_\sigma\|\sigma(x_1) - \sigma(x_2)\| \leq \|x_1 - x_2\|$.
- For any $x \in \mathbb{R}^d$ and $\mathbf{W} \in \mathbb{R}^{n\times d}$, we have $\|\mathbf{W}x\| \leq B_w\|x\|$
- softmax is continuously differentiable and its Jacobian satisfies for vectors $\theta_1, \theta_2 \in \mathbb{R}^p$, $\|\text{softmax}(\theta_1) - \text{softmax}(\theta_2)\| \leq 2\|\theta_1 - \theta_2\|_\infty$

Since the core part of the Transformer is the Attention mechanism, for this reason we need to first give an upper bound on the covering number of the Attention head.

Lemma 2 (Edelman et al. (2022)) $\forall\alpha \in [0,1]$, *the coverage number of the Attention head* $\mathcal{F}_{tf-head}$ *satisfies*

$$\begin{aligned} &\log\mathcal{N}_\infty(\mathcal{F}_{tf-head};\epsilon;\{(X^{(i)},z^{(i)})\}_{i=1}^m) \\ &\leq \inf{}_{\alpha\in[0,1]}[\log\mathcal{N}_\infty(\mathcal{F}_{QK};\frac{\alpha\epsilon}{2L_\sigma B_V B_X};\{(x_t^{(i)},z^{(i)})\}_{i\in[m],t\in[T]}) \\ &+\log\mathcal{N}_\infty(\mathcal{F}_V;\frac{(1-\alpha)\epsilon}{L_\sigma};\{x_t^{(i)}\}_{i\in[m],t\in[T]};\|\cdot\|_2)] \end{aligned} \tag{20}$$

where ϵ is any real number greater than 0, z is the additonal context of x, and B_V and B_X are upper bounds on the weights $\mathbf{W}_V$ and $\mathbf{X}$, respectively.

Then, in order to further derive an optimization upper bound on the above covering number upper bound, we make use of the conclusions on covering number upper bounds for the class of linear functions given by Edelman et al. (2022).

Lemma 3 *Let* $\mathcal{W} : \{\mathbf{W}\in\mathbb{R}^{d_1\times d_2} : \|\mathbf{W}^\top\|_{2,1}\leq B_W\}$*, for the function class* $\mathcal{F} : \{x\mapsto\mathbf{W}x : \mathbf{W}\in\mathcal{W}\}$*. For* $\forall\epsilon>0$ *and* $x^{(1)},...,x^{(N)}\in\mathbb{R}^{d_1}$ *satisfying* $\forall i\in[N], \|x^{(i)}\|\leq B_X$*,*

$$\log\mathcal{N}_\infty(\mathcal{F};\epsilon;x^{(1)},...,x^{(N)};\|\cdot\|_2)\lesssim\frac{(B_X B_W)^2}{\epsilon^2}\log(d_1 N) \tag{21}$$

For ease of arithmetic, we consider $\mathbf{W}_K\mathbf{W}_Q$ used in the computation of self-attention scores as a matrix with an upper bound of B_KQ, and for this purpose, using the Lemma 2 we can easily obtain upper bounds on the number of coverings for the class of linear transformation functions of the linear transformations $W_{KQ}X$ and W_VX.

$$\begin{aligned} \log N_\infty(\mathcal{F}_{QK};\epsilon_{QK};\{(x_t^{(i)},z^{(i)})\}_{i\in[m],t\in[T]}) &\lesssim\frac{(B_{QK}^{2,1}B_X)^2\log(\text{dmT})}{\epsilon_{QK}^2} \\ \log\mathcal{N}_\infty(\mathcal{F}_V;\epsilon_V;\{(x_t^{(i)},z^{(i)})\}_{i\in[m],t\in[T]}) &\lesssim\frac{(B_V^{2,1}B_X)^2\log(\text{dmT})}{\epsilon_V^2} \end{aligned} \tag{22}$$

Then we get by our assumptions that when we want to choose ϵ_{QK} and ϵ_V such that the sum of the above two terms is minimized, subject to

$$2L_\sigma B_V B_X\epsilon_V\leq\epsilon \tag{23}$$

the solution (without position encoding) to this optimization leads to an optimal bound of:

$$\log\mathcal{N}_\infty(\mathcal{F}_{tf};\epsilon;\mathbf{X}^{(1)},..,\mathbf{X}^{(M)})\lesssim(L_\sigma B_X)^2\cdot\frac{((B_V^{2,1})^{\frac{2}{3}}+(B_{QK}^{2,1}B_V B_X)^{\frac{2}{3}})^3}{\epsilon^2}\cdot\log(dmT) \tag{24}$$

With Lemma 1 proof ideas laid out, we now proceed to prove the relevant properties of the model after we introduce the positional encoding we devised. According to Equation 21, we have

$$\|\mathbf{Z}\| = \|[[\mathbf{X}_0, sin(\mathbf{X}_0), cos(\mathbf{X}_0), \mathbf{X}_1, ...]\mathbf{W}\| \tag{25}$$

From $\|\mathbf{X}\|\leq B_X$, we can easily obtain $\|\mathbf{Z}\|\leq B_X+d$ based on the triangular inequality. Therefore, according to Lemma 1, we can get the upper bound on the number of model coverings after introducing the positional encoding after HeterPos

$$\begin{aligned} \log\mathcal{N}\infty(\mathcal{F}(\mathbf{X};\epsilon;\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(n)},\|\cdot\|))\lesssim&\frac{((B^{2,1}V)^{2/3}+(2BQ,K^{2,1}B_V(B_X+d))^{2/3})^3}{\epsilon^2}\cdot \\ &(L\sigma(B_X+d))^2\log(nd) \end{aligned} \tag{26}$$

In order to obtain a further derivation of the relationship between the covering number and the Rademacher Complex, we give the following theorem.

Lemma 4 *Consider a real-valued function class $\mathcal{F}$ such that $|f| \leq A$ for all $f \in \mathcal{F}$. Then,*

$$\hat{\mathcal{R}} \leq c \cdot \inf_{\delta \geq 0} (\delta + \int_{\delta}^{A} \sqrt{\frac{\log \mathcal{N}_{\infty}(\mathcal{F}; \epsilon; \mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)})}{n}} dx) \tag{27}$$

We may assume that $\log \mathcal{N}_{\infty}(\mathcal{F}; \epsilon; \mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}) \leq \frac{\mathcal{N}_{\mathcal{F}}}{\epsilon^2}$. Based on Lemma 2, we have...

$$\begin{aligned}
\hat{\mathcal{R}}(\mathcal{F}; \mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}) &\leq c \cdot \inf_{\delta \geq 0} (\delta + \int_{\delta}^{A} \sqrt{\frac{\log \mathcal{F}; \epsilon; \mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)})}{n}}) dx \\
&\leq c \cdot \inf_{\delta \geq 0} (\delta + \int_{\delta}^{A} \sqrt{\frac{\mathcal{N}_{\mathcal{F}}}{\epsilon^2 n}} d\epsilon) \\
&= c \cdot \inf_{\delta \geq 0} (\delta + \sqrt{\frac{\mathcal{N}_{\mathcal{F}}}{n}} \int_{\delta}^{A} \int_{\delta}^{A} \frac{1}{\epsilon} d\epsilon) \\
&= c \cdot \inf_{\delta \geq 0} (\delta + \sqrt{\frac{\mathcal{N}_{\mathcal{F}}}{n}} \log(\frac{A}{\delta})) \\
&= c \cdot \sqrt{\frac{\mathcal{N}_{\mathcal{F}}}{n}} (1 + \log(A + \sqrt{\frac{n}{\mathcal{N}_{\mathcal{F}}}}))
\end{aligned} \tag{28}$$

The $\mathcal{N}_{\mathcal{F}}$ is $((B_V^{2,1})^{2/3} + (2B_{\mathrm{PE}} B_{Q,K}^{2,1} B_V (B_X + d))^{2/3})^3 \cdot (L_{\sigma} B_{\mathrm{PE}} (B_X + d))^2 \log(nd)$, $|f| \leq A$ for all $f \in \mathcal{F}$.

B APPENDIX

B.1 DETAILED EXPERIMENTAL SETUP

We employed feature vectors, class labels, and 10 random splits (48%/32%/20% for training/validation/testing) from the work of Yan et al. (2022) for all baseline models. And we conducted experiments over 2000 epochs and implemented early stopping if the validation loss decreased consistently for 200 consecutive epochs.

The parameters used for MPformer experiments on each dataset are shown in Table 3

Table 3: Parameters used in each data set.

Hyperparameter	Cora	Citeseer	Cornell	Texas	Wisconsin	Actor	Chameleon	Squirrel
Layer	1	1	1	1	1	1	1	1
Heads	2	2	2	2	2	2	2	2
Hidden dim	64	256	16	16	16	128	16	16
Epoch	2000	2000	2000	2000	2000	2000	2000	2000
Learning rate	0.001	0.001	0.01	0.01	0.01	0.0001	0.01	0.01
Weight decay	5e-4	1e-4	5e-4	5e-4	5e-4	5e-5	5e-4	5e-4
Transformer dropout	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Feature dropout	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

The environment in which we run experiments is:

- CPU information:24 vCPU AMD EPYC 7642 48-Core Processor
- GPU information:RTX 3090(24GB)