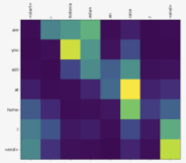
Primphormer: Efficient Graph Transformers with Primal Representations

Mingzhen He · Ruikai Yang · Hanling Tian · Youmei Qiu · Xiaolin Huang*

SHANGHAI JIAO TONG UNIVERSITY

Attention mechanism

$$\begin{cases} K(x_i, x_j) = \sigma(\langle q(x_i), k(x_j)\rangle) \\ o(x_i) = \sum_{j=1}^{N} v(x_j) K(x_i, x_j) \end{cases}$$



- **Asymmetry**

  Where does asymmetry come from?

  $$\begin{matrix} q(x_i) \ k(x_j) & & q(x_j) \ k(x_i) \\ \langle W_q x_i, W_k x_j \rangle & \neq & \langle W_q x_j, W_k x_i \rangle \end{matrix}$$

- **Computational overhead**

  Large-scale problem (huge N) $\longrightarrow$ $\mathcal{O}(N^2)$ Sometime infeasible

Primphormer

$$\min_{\substack{W_e, W_r \\ e, r}} \quad J = \frac{1}{2}\sum_i e_i^T \Lambda e_i + \frac{1}{2}\sum_j r_j^T \Lambda r_j - \mathrm{Tr}(W_e^T W_r),$$

$$\text{s.t.} \quad e_i = f_X W_e \phi_q(x_i), \quad r_j = f_X W_r \phi_k(x_j).$$

Key idea: Using primal-dual relationship to remodel the attention mechanism on graphs

$$\text{Primal} \quad \begin{cases} e(x) = f_X W_e \phi_q(x) \\ r(x) = f_X W_e \phi_k(x). \end{cases}$$

$$\text{Dual} \quad \begin{cases} e(x) = \sum_{j=1}^{N} F_X h_{r_j} \phi_k(x_j)^T \phi_q(x) := \sum_{j=1}^{N} \tilde{h}_{r_j} K(x, x_j), \\ r(x) = \sum_{i=1}^{N} F_X h_{e_i} \phi_q(x_i)^T \phi_k(x) := \sum_{i=1}^{N} \tilde{h}_{e_i} K(x_i, x). \end{cases}$$

The standard attention formulation

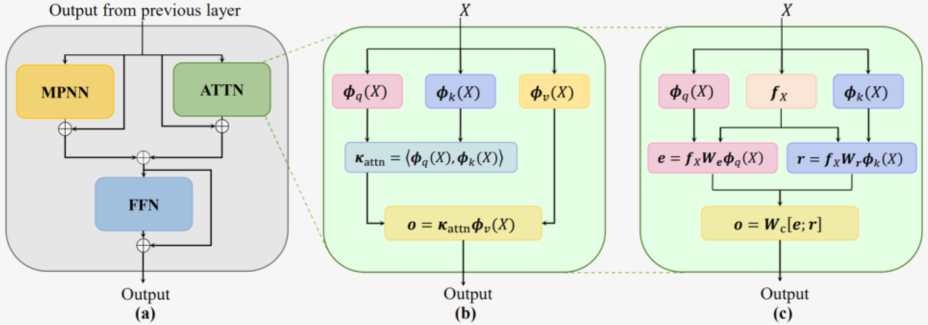


Figure 1 Illustrations of the architectures in one layer. **(a)** The GPS architecture. **(b)** The standard self-attention architecture. The attention score κ_{attn} involves pair-wise computations. **(c)** Primphormer eliminates the need for pair-wise computations by introducing the primal representation, resulting in a new computationally efficient GT.

Expressivity and UAT

Theorem 3.3. *For any continuous function $f : [0,1]^{d\times N} \to \mathbb{R}^{d\times N}$ and for each $\epsilon > 0$, there exists a Primphormer $\mathcal{T}_{\mathrm{PE}}$ with the positional encoding $\boldsymbol{E}$ such that*

$$\sup_{\boldsymbol{X}\in\mathcal{X}^N} \|f(\boldsymbol{X}) - \mathcal{T}_{\mathrm{PE}}(\boldsymbol{X})\|_\infty < \epsilon. \tag{3.2}$$

Theorem 3.4. *Let $G = (V, E, \ell)$ be a labeled graph with N nodes, and node feature matrix $\boldsymbol{X}^{(0)} := \boldsymbol{H} \in \mathbb{R}^{d\times N}$ consistent with the label ℓ. Then, for all iterations $t \geq 0$, there exists a parameterization of Primphormer such that*

$$C_t^1(v) = C_t^1(w) \iff \boldsymbol{X}^{(t)}(v) = \boldsymbol{X}^{(t)}(w), \tag{3.3}$$

for all nodes $v, w \in V$, where $C_t^1 : V \to \mathbb{N}$ is the coloring function of the 1-WL test at t-th iteration.

Corollary 3.5. *Primphormer is as powerful as Transformer in terms of distinguishing non-isomorphic graphs.*

Table 3 Comparison of attentions in GPS. Best results are colored in first, second, third. OOM means out of memory.

MODEL GPS	CIFAR10 ACCURACY↑	MALNET-TINY ACCURACY↑	PASCALVOC-SP F1↑	PEPTIDES-FUNC AP↑	OGBN-PRODUCTS ACCURACY↑
MPNN-ONLY	69.95 ± 0.499	92.23 ± 0.650	0.3016 ± 0.0031	0.6159 ± 0.0048	74.25 ± 0.214s
+TRANSFORMER	72.31 ± 0.344	93.50 ± 0.410	0.3748 ± 0.0109	0.6535 ± 0.0041	OOM
+BIGBIRD	70.48 ± 0.106	92.34 ± 0.340	0.2762 ± 0.0069	0.5854 ± 0.0079	73.82 ± 0.412
+PERFORMER	70.67 ± 0.338	92.64 ± 0.780	0.3724 ± 0.0131	0.6475 ± 0.0056	74.30 ± 0.211
+PRIM-ATTEN	71.57 ± 0.256	92.97 ± 0.228	0.3173 ± 0.0055	0.6447 ± 0.0046	74.47 ± 0.134
+EXPHORMER	**74.69 ± 0.125**	**94.02 ± 0.209**	**0.3975 ± 0.0037**	**0.6527 ± 0.0043**	**74.67 ± 0.179**
+PRIMPHORMER	**74.13 ± 0.241**	**93.62 ± 0.242**	**0.4602 ± 0.0077**	**0.6612 ± 0.0065**	**74.89 ± 0.281**

Table 4 Efficiency comparisons on running time and peak memory consumption.

MODEL	TIME (S/EPOCH)					PEAK MEMORY USAGE (GB)				
GPS	CIFAR.	MALNET.	PASCAL.	FUNC.	PROD.	CIFAR.	MALNET.	PASCAL.	FUNC.	PROD.
MPNN-ONLY	20.3	24.5	15.7	4.8	21.1	2.31	1.92	4.18	2.45	11.97
+TRANSFORMER	28.0	232.4	35.6	12.8	-	3.81	35.32	7.82	8.46	OOM
+BIGBIRD	55.2	325.6	52.3	51.9	93.9	2.81	2.71	4.99	4.99	17.29
+PERFORMER	50.8	73.5	49.7	21.7	22.7	10.5	11.59	6.14	7.71	16.14
+PRIM-ATTEN	32.1	62.5	25.7	7.9	22.6	2.74	2.58	4.74	3.38	13.63
+EXPHORMER	44.5	62.1	35.2	7.6	25.4	5.54	10.38	7.35	4.81	31.09
+PRIMPHORMER	32.6	61.9	25.3	7.7	22.1	2.74	2.86	4.72	3.41	13.35

Table 5 Results on the BREC benchmark. Basic, Regular, Extension, and CFI are subsets of the BREC benchmark. Experiments are averaged over 5 runs.

MODEL	PE	BAS.↑	REG.↑	EXT.↑	CFI↑	ALL↑
GRAPHORMER		16	12	41	10	79
APE-GT		50.6	31.3	62.4	1	145.3
3-WL		**60**	**50**	**100**	**60**	**270**
TRANSFORMER	LAP	47.2	39	65.2	**3**	154.4
PRIM-ATTEN		12.8	19	13.6	0.6	46
PRIMPHORMER		**51.6**	**42**	**72.4**	3	**169**
TRANSFORMER	SPE	59.8	49.4	98.6	5.2	213
PRIM-ATTEN		46.4	49	73	3	171.4
PRIMPHORMER		**60**	**50**	**100**	**9.4**	**219.4**

Takes away

- **Primal representation can be an efficient module**
- **Primphormer preserves good theoretical property**