Primal-Attention: Self-attention through Asymmetric Kernel SVD in Primal Representation

Yingyi Chen*, Qinghua Tao*, Francesco Tonin, Johan A.K. Suykens

ESAT-STADIUS, KU Leuven, Belgium   *Equal contribution

European Research Council
Established by the European Commission

KU LEUVEN

NEURAL INFORMATION PROCESSING SYSTEMS

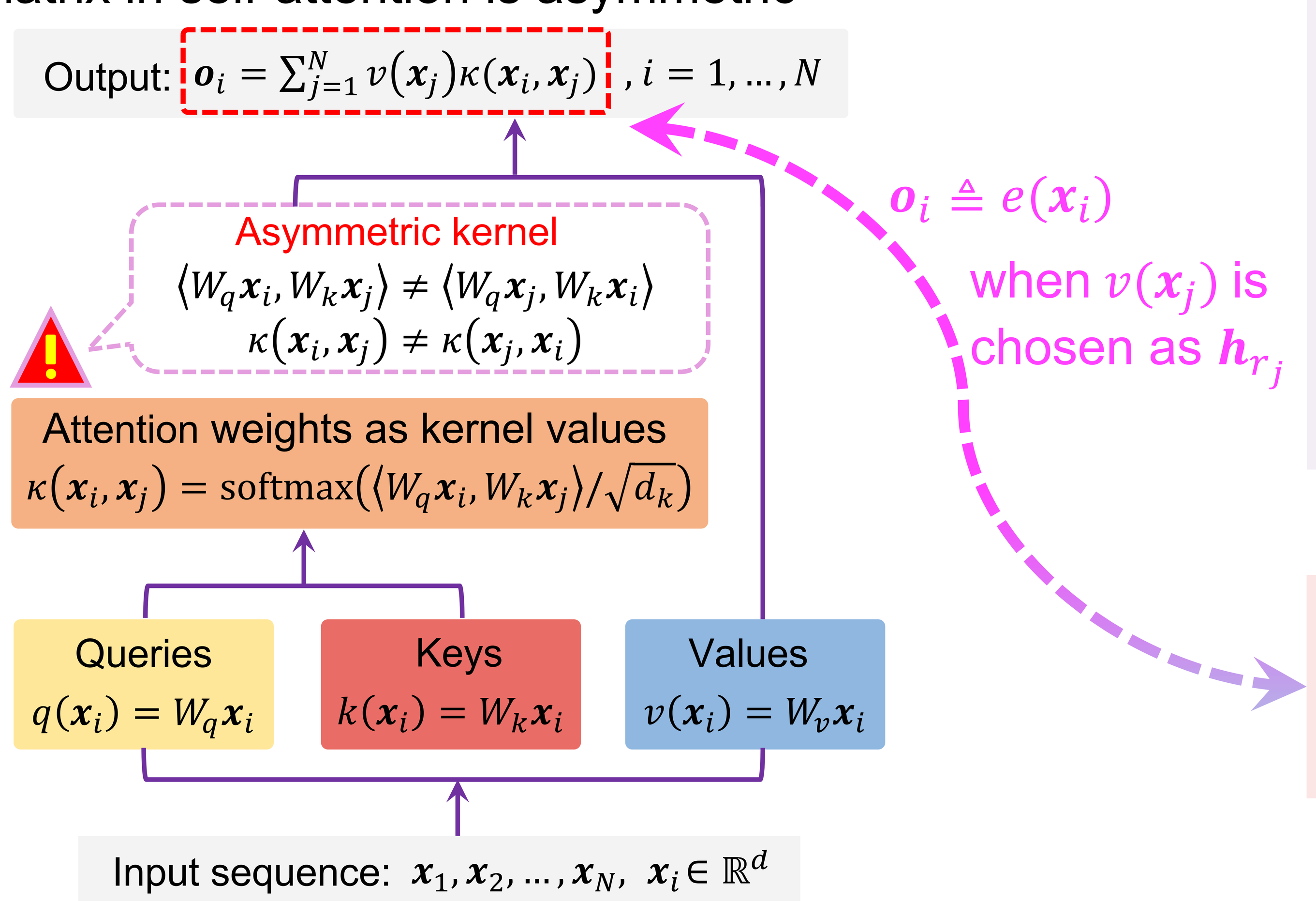
Summary

Represent and optimize self-attention through *asymmetric Kernel Singular Value Decomposition* (KSVD):

- A primal-dual representation of self-attention in Transformers is formulated;
- A new attention mechanism named Primal-Attention based on primal representation of KSVD is proposed;
- A KSVD optimization designed for Primal-Attention is implemented.

Canonical self-attention is with Asymmetric Kernel

- Attention matrix in self-attention is asymmetric



- For *asymmetric kernel*, the kernel trick from *Reproducing Kernel Banach Spaces* (RKBS) with $\kappa(\cdot,\cdot): \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ can be defined by the inner product of two feature maps from $\mathcal{B}_\mathcal{X}, \mathcal{B}_\mathcal{Z}$:

$$\kappa(\boldsymbol{x}, \boldsymbol{z}) = \langle \phi_x(\boldsymbol{x}), \phi_z(\boldsymbol{z}) \rangle, \ \forall \boldsymbol{x} \in \mathcal{X}, \phi_x \in \mathcal{B}_\mathcal{X}, \boldsymbol{z} \in \mathcal{Z}, \phi_z \in \mathcal{B}_\mathcal{Z}.$$

SVD and Shifted Eigenvalue Problem

- SVD factorizes a given matrix $A \in \mathbb{R}^{N \times M}$ by two sets of orthonormal eigenbases:

$$A = U \Sigma V^\top, \ \Sigma = \mathrm{diag}\{\sigma_1, \dots, \sigma_s\}, \ \underbrace{U \in \mathbb{R}^{N \times s}}_{\text{left singular vectors}}, \ \underbrace{V \in \mathbb{R}^{M \times s}}_{\text{right singular vectors}}$$

- Decomposition theorem (Lanczos, 1958): Any non-zero matrix $A \in \mathbb{R}^{N \times M}$ can be written as $A = \tilde{U} \tilde{\Sigma} \tilde{V}^\top$, where $\tilde{U} \in \mathbb{R}^{N \times s}$, $\tilde{V} \in \mathbb{R}^{M \times s}$, $\tilde{\Sigma} \in \mathbb{R}^{s \times s}$ are defined by the *shifted eigenvalue problem*:

$$A \tilde{V} = \tilde{U} \tilde{\Sigma}, \quad A^\top \tilde{U} = \tilde{V} \tilde{\Sigma},$$

$\tilde{U}^\top \tilde{U} = I_s$, $\tilde{V}^\top \tilde{V} = I_s$, $\tilde{\Sigma}$ is diagonal with positive numbers.

Primal-dual Representation of Self-attention with KSVD

- *Primal problem* with KSVD for self-attention: we extend SVD under *Least Squares Support Vector Machines* (Suykens et al., 2002) framework (Suykens, 2016) to a nonlinear version

$$\max_{W_e, W_r, \boldsymbol{e}_i, \boldsymbol{r}_j} J = \frac{1}{2} \sum_{i=1}^N \boldsymbol{e}_i^\top \Lambda \boldsymbol{e}_i + \frac{1}{2} \sum_{j=1}^N \boldsymbol{r}_j^\top \Lambda \boldsymbol{r}_j - \mathrm{Tr}(W_e^\top W_r)$$

$$\text{s.t. } \boldsymbol{e}_i = (f(X)^\top W_e)^\top \phi_q(\boldsymbol{x}_i), \ i = 1, \dots, N,$$
$$\boldsymbol{r}_j = (f(X)^\top W_r)^\top \phi_k(\boldsymbol{x}_j), \ j = 1, \dots, N,$$

Data-dependent projection weights	Feature maps related to queries and keys	Asymmetric attention kernel / Regu. coeff.	Projection scores w.r.t. queries, keys
$f(X)^\top W_e =: W_{e\vert X} \in \mathbb{R}^{p \times s}$	$\phi_q(\cdot): \mathbb{R}^d \to \mathbb{R}^p$	$K := [\langle \phi_q(\boldsymbol{x}_i), \phi_k(\boldsymbol{x}_j) \rangle]$	$\boldsymbol{e}_i := W_{e\vert X}^\top \phi_q(\boldsymbol{x}_i)$
$f(X)^\top W_r =: W_{r\vert X} \in \mathbb{R}^{p \times s}$	$\phi_k(\cdot): \mathbb{R}^d \to \mathbb{R}^p$	$\Lambda \in \mathbb{R}^{s \times s} \succ 0$ diagonal	$\boldsymbol{r}_j := W_{r\vert X}^\top \phi_k(\boldsymbol{x}_j)$

- *Dual problem* with KSVD for self-attention: with Lagrangian duality and KKT conditions, the dual problem of above leads to the shifted eigenvalue problem

$$K H_r = H_e \Sigma, \quad K^\top H_e = H_r \Sigma,$$

$H_e = [\boldsymbol{h}_{e_1}, \dots, \boldsymbol{h}_{e_N}]^\top \in \mathbb{R}^{N \times s}$, $H_r = [\boldsymbol{h}_{r_1}, \dots, \boldsymbol{h}_{r_N}]^\top \in \mathbb{R}^{N \times s}$ are dual variables serving as *left* and *right singular vectors*.

- *Primal-dual representation* of KSVD in self-attention:

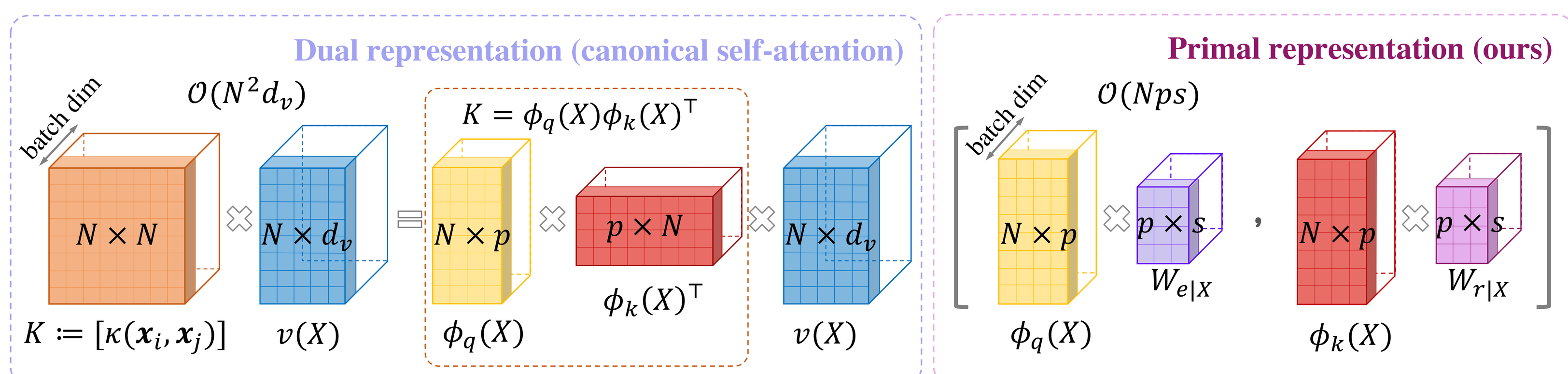
$$\text{Primal: } \begin{cases} e(\boldsymbol{x}) = W_{e|X}^\top \phi_q(\boldsymbol{x}) \\ r(\boldsymbol{x}) = W_{r|X}^\top \phi_k(\boldsymbol{x}) \end{cases}, \quad \text{Dual: } \begin{cases} \boxed{e(\boldsymbol{x}) = \sum_{j=1}^N \boldsymbol{h}_{r_j} \kappa(\boldsymbol{x}, \boldsymbol{x}_j)} \\ r(\boldsymbol{x}) = \sum_{i=1}^N \boldsymbol{h}_{e_i} \kappa(\boldsymbol{x}_i, \boldsymbol{x}) \end{cases}$$

Primal-Attention

- *Modelling:* leveraging primal representation of KSVD with ϕ_q, ϕ_k

$$\boldsymbol{o}_i := [\boldsymbol{e}_i; \boldsymbol{r}_i] = [W_{e|X}^\top \phi_q(\boldsymbol{x}_i); W_{r|X}^\top \phi_k(\boldsymbol{x}_i)],$$

experimentally, $\phi_q(\boldsymbol{x}) := q(\boldsymbol{x}) / \|q(\boldsymbol{x})\|_2, \phi_k(\boldsymbol{x}) := k(\boldsymbol{x}) / \|k(\boldsymbol{x})\|_2$, reducing time complexity from $\mathcal{O}(N^2 d_v)$ to $\mathcal{O}(Nps)$.



- *Optimization:* stationary solutions of KSVD for each head can be obtained by a zero-value of the primal objective $J = 0$

$$J(W_e, W_r, \Lambda) = \frac{1}{2} \sum_{i=1}^N \left\| (W_{e|X} \Lambda^{1/2})^\top \phi_q(\boldsymbol{x}_i) \right\|_2^2 + \frac{1}{2} \sum_{j=1}^N \left\| (W_{r|X} \Lambda^{1/2})^\top \phi_k(\boldsymbol{x}_j) \right\|_2^2 - \mathrm{Tr}(W_e^\top W_r).$$

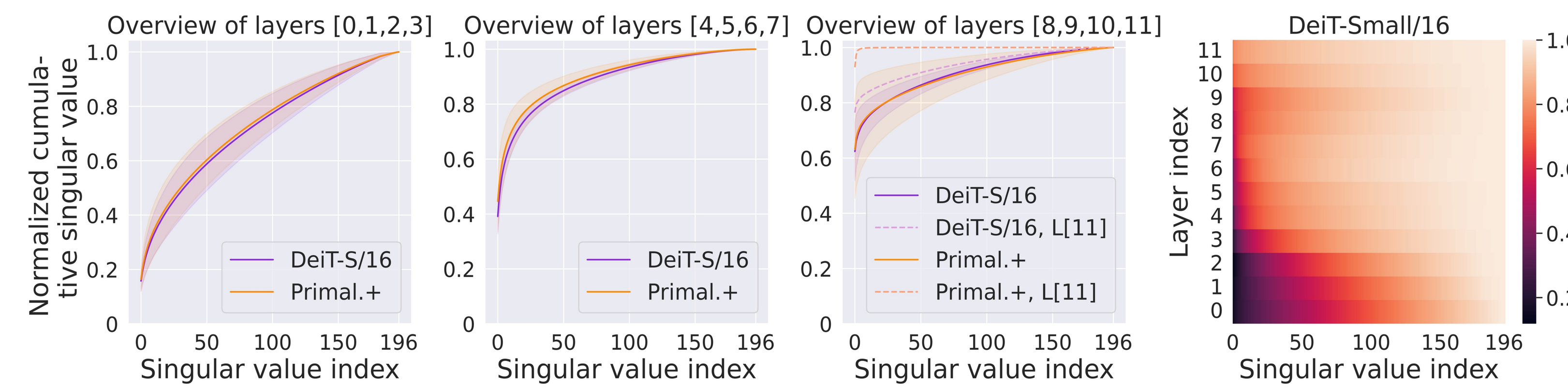
The *objective* of Primal-Attention is

$$\min \ L_{\mathrm{CE}} + \eta \sum_l J_l^2,$$

Σ_l denotes additive objectives J_l of all Primal-attention blocks, where J_l is implemented as the mean of all heads.

Numerical Experiments

- *Enhanced low-rank property*: spectrum analysis of the self-attention matrix on ImageNet-1K



- *Enhanced accuracy & efficiency*: Long-Range Arena Benchmark

Dataset (seq. length)	Trans-former	Re-former	Per-former	Lin-former	Nyström-former	Long-former	YOSO-E	**Primal.**	**Primal. +Trans.**
ListOps (2K)	37.1	19.1	18.8	37.3	37.2	37.2	37.3	37.3	37.3
Text (4K)	65.0	64.9	63.8	55.9	65.5	64.6	64.7	61.2	65.4
Retrieval (4K)	79.4	78.6	78.6	79.4	79.6	81.0	81.2	77.8	81.0
Image (1K)	38.2	43.3	37.1	37.8	41.6	39.1	39.8	43.0	43.9
Pathfinder (1K)	74.2	69.4	69.9	67.6	70.9	73.0	72.9	68.3	74.3
Avg. Acc. (%)	58.8	55.1	53.6	55.6	59.0	59.0	59.2	57.5	**60.4**

Model	Time (s/1K-steps)					Memory (GB)				
	ListOps	Text	Retrieval	Image	Pathfinder	ListOps	Text	Retrieval	Image	Pathfinder
Transformer	194.5 (1×)	694.8 (1×)	1333.7 (1×)	334.5 (1×)	405.5 (1×)	5.50 (1×)	21.24 (1×)	18.72 (1×)	5.88 (1×)	5.88 (1×)
Nyströmformer	68.4 (2.8×)	120.9 (5.7×)	235.5 (5.7×)	179.5 (1.9×)	221.2 (1.8×)	0.89 (6.2×)	1.69 (12.6×)	3.29 (5.7×)	1.93 (3.0×)	1.93 (3.0×)
Linformer	63.4 (3.1×)	116.5 (6.0×)	226.2 (5.9×)	158.5 (2.1×)	204.0 (2.0×)	1.73 (3.2×)	3.45 (6.2×)	6.33 (3.0×)	3.45 (1.7×)	3.45 (1.7×)
Performer	83.8 (2.3×)	157.5 (4.4×)	320.6 (4.2×)	211.4 (1.6×)	278.1 (1.5×)	1.67 (3.3×)	3.34 (6.4×)	6.28 (3.0×)	3.34 (1.8×)	3.34 (1.8×)
Reformer	87.0 (2.2×)	168.5 (4.1×)	339.9 (3.9×)	223.7 (1.5×)	286.7 (1.4×)	1.64 (3.3×)	3.29 (6.5×)	6.09 (3.1×)	3.29 (1.8×)	3.29 (1.8×)
Primal.+Trans.	113.4 (1.7×)	367.6 (1.9×)	546.5 (2.4×)	212.1 (1.6×)	263.2 (1.5×)	5.24 (1.1×)	20.7 (1.0×)	18.59 (1.0×)	5.35 (1.1×)	5.35 (1.1×)
Primal.	**56.5 (3.4×)**	**93.6 (7.4×)**	**185.3 (7.2×)**	**142.9 (2.3×)**	**180.0 (2.3×)**	**0.69 (7.9×)**	**1.37 (15.5×)**	**2.99 (6.3×)**	**1.39 (4.2×)**	**1.52 (3.9×)**

- Other benchmarks including *UEA time series classification*, *D4RL reinforcement learning*, *ImageNet-100*, *ImageNet-1K*, *WikiText-103* and more *ablation studies* can be found in the paper.

Paper:

arXiv:2305.19798

Code:

github.com/yingyichen-cyy/PrimalAttention

References:

Lanczos. "Linear systems in self-adjoint form." *The American Mathematical Monthly*, 1958.

Suykens et al. *Least Squares Support Vector Machines*. World scientific, 2002.

Suykens. "SVD revisited: A new variational principle, compatible feature maps and nonlinear extensions." *Applied and Computational Harmonic Analysis*, 2016.