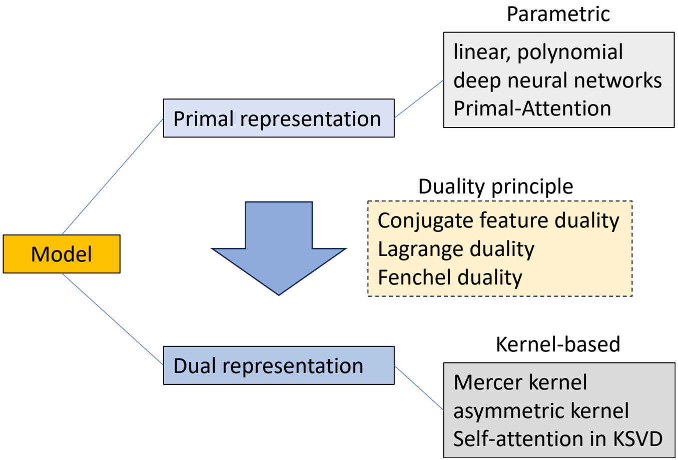
Asymmetric Kernels Meet Transformers: A Primal-Dual Approach to Self-Attention through Kernel Singular Value Decomposition

Francesco Tonin

Department of Electrical Engineering ESAT-STADIUS
KU Leuven. Kasteelpark Arenberg 10 B-3001 Leuven, Belgium
francesco.tonin@esat.kuleuven.be

EPFL LIONS, Lausanne
November 15, 2023

Core idea

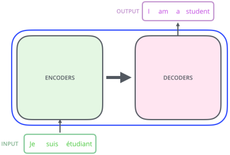


Core framework behind a series of three papers

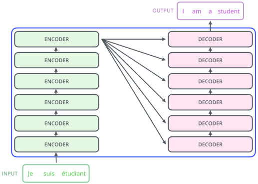
- Faster and robust/sparse Kernel PCA [Ton+23]
- Nonlinear SVD through asymmetric kernels [Tao+23]
- New representation of self-attention in Transformer [Che+23]

High-level look on Transformer: introduction

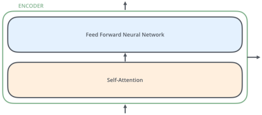
Transformer as autoencoder



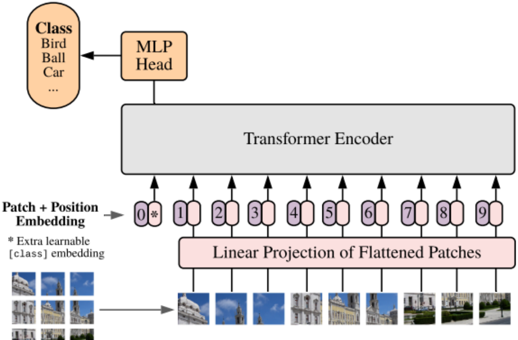
Multi-layer encoder and decoder



Transformer Encoder block



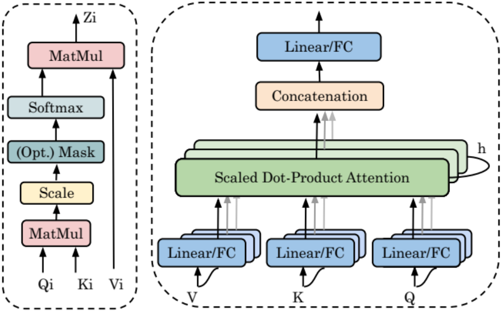
Vision Transformer (ViT)



High-level look on Transformer: self-attention

Queries, keys, values as linear projections of the input sequence:

$$q(x_i) = W_q x_i$$
$$k(x_i) = W_k x_i$$
$$v(x_i) = W_v x_i$$



(a) Scaled Dot-product Attention

(b) Multi-Head Self-Attention

(source from [Chi+23])

Self-Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Multi-Head Self-Attention (MHSA): concatenation of h parallel self-attention mechanisms

Challenges in Transformers: very large models

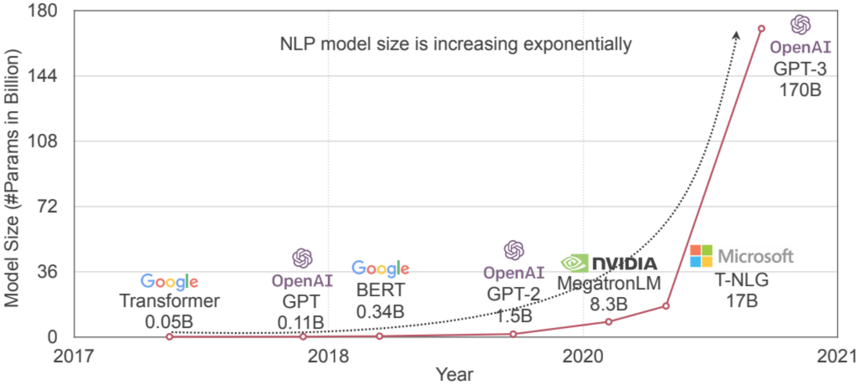
Computing self-attention is expensive for larger models

Mechanism	Computation	Memory
Self-Attention	$\mathcal{O}(N^2 d)$	$\mathcal{O}(N^2 + Nd)$
MHSA	$\mathcal{O}(N^2 d + Nd^2)$	$\mathcal{O}(N^2 h + Nd)$

with sequence length N, hidden size d, and h heads

Challenges in Transformers: very large models

Transformers are big!



(source from Song Han)

Challenges in Kernel PCA

Given datapoints $(x_i)_{i=1}^N$, feature map ϕ mapping into feature space $\mathcal{H}$ associated to kernel k, and number of components s

Kernel PCA

Find orthonormal directions $(w_j)_{j=1}^s \in \mathcal{H}^s$ that give the best rank s approximation of the empirical covariance in feature space

Challenges

- Speed: solved by truncated SVD of the kernel matrix $G = [k(x_i, x_j)]_{i,j=1}^N \rightarrow$ not scalable
- Robustness: KPCA only maximizes variance, how can we robustify solutions ?

Challenges in Kernel PCA

Given datapoints $(x_i)_{i=1}^N$, feature map ϕ mapping into feature space $\mathcal{H}$ associated to kernel k, and number of components s

Kernel PCA

Find orthonormal directions $(w_j)_{j=1}^s \in \mathcal{H}^s$ that give the best rank s approximation of the empirical covariance in feature space

Challenges

- Speed: solved by truncated SVD of the kernel matrix $G = [k(x_i, x_j)]_{i,j=1}^N \to$ not scalable
- Robustness: KPCA only maximizes variance, how can we robustify solutions ?

How can Kernel PCA help address efficiency and modelling problems in Transformers?

Challenges in Kernel PCA

Given datapoints $(x_i)_{i=1}^N$, feature map ϕ mapping into feature space $\mathcal{H}$ associated to kernel k, and number of components s

Kernel PCA

Find orthonormal directions $(w_j)_{j=1}^s \in \mathcal{H}^s$ that give the best rank s approximation of the empirical covariance in feature space

Challenges

- Speed: solved by truncated SVD of the kernel matrix $G = [k(x_i, x_j)]_{i,j=1}^N \rightarrow$ not scalable
- Robustness: KPCA only maximizes variance, how can we robustify solutions ?

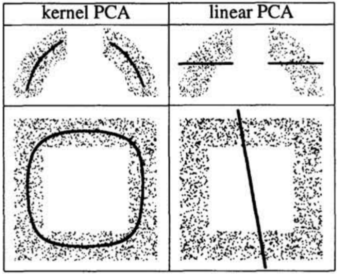
How can Kernel PCA help address efficiency and modelling problems in Transformers?

Through **asymmetric kernels**...

Kernel PCA

Corresponding paper:
Tonin, F., Lambert, A., Patrinos, P., & Suykens, J. (2023). Extending Kernel PCA through Dualization: Sparsity, Robustness and Fast Algorithms. *ICML 2023*.

Kernel Principal Component Analysis (KPCA)



(reproduced from [Mik+99])

Nonlinear extension of PCA by:

- Mapping input space to a high dimensional feature space $\mathcal{H}$
- Linear PCA is performed in the feature space induced by ϕ
- Applying the kernel trick
- Usual way to solve KPCA [Sch+98]: top s eigenvectors of kernel matrix $G \in \mathbb{R}^{N \times N}$ $\Rightarrow$ slow with larger N

LS-SVM approach to kernel PCA

LS-SVM formulation of KPCA with Lagrangian duality [Suy+02]

- Primal problem:

$$\boxed{\text{P}} \quad \min_{w,e} \frac{1}{2}\|w\|^2 - \frac{1}{2\lambda}\sum_{i=1}^{N} e_i^2 \quad \text{s.t.} \quad e_i = w^\top \phi(x_i)$$

- Lagrangian

$$\mathcal{L}(w, e; h) = \frac{1}{2\lambda}\sum_{i=1}^{N} e_i^2 - \frac{1}{2} w^\top w - \sum_{i=1}^{N} h_i \left(e_i - w^T (\phi(x_i))\right)$$

- Elimination of w, e in the optimality conditions gives

$$\boxed{\text{D}} \quad Gh = \lambda h,$$

with kernel trick $G_{ij} = \phi(x_i)^T \phi(x_j) = k(x_i, x_j)$

KPCA as difference of convex functions

Alternative formulation: variance maximization under orthonormality constraints

$$\text{KPCA problem:} \qquad \sup_{W \in \mathcal{S}_{\mathcal{H}}^{s}} \frac{1}{2} \|\Gamma W\|_{\mathrm{F}}^{2}$$

💡 **Key idea**: Rewrite KPCA as difference of convex functions

Proposition: Dual of difference of convex functions

Let $\mathcal{S}_{\mathcal{H}}^{s}$ be the Stiefel manifold of orthonormal s-frames in $\mathcal{H}$, and operator $\Gamma \colon \mathcal{H}^s \to \mathbb{R}^{N \times s}$, $\Gamma W = [\langle \phi(x_i), w_j \rangle]_{i,j=1}^{N,s}$. The problem

$$\inf_{W \in \mathcal{H}^s} \; g(W) - f(\Gamma W)$$

admits the dual formulation

$$\inf_{H \in \mathbb{R}^{N \times s}} \; f^\star(H) - g^\star(\Gamma^\sharp H)$$

and strong duality holds.

Solving the dual of KPCA

For the KPCA problem: $f = \frac{1}{2}\|\cdot\|_{\mathrm{F}}^2$ and $g = \iota_{\mathbb{S}^s_{\mathcal{H}}}$

Dual problem to KPCA

$$\inf_{H \in \mathbb{R}^{N \times s}} \frac{1}{2} \operatorname{Tr}(H^\top H) - \underbrace{\operatorname{Tr} \sqrt{H^\top G H}}_{:= \pi(H)}$$

Computing $\nabla\pi$ is possible:

$$\nabla\pi(H) = GHU^\top \operatorname{diag}\left(\frac{1}{\sqrt{\lambda(H^\top G H)}}\right) U$$

where U comes from the SVD of $H^\top G H$. Complexity:

- Computation of $H^\top G H$ in $\mathcal{O}(sN^2)$,
- SVD of $H^\top G H$ in $\mathcal{O}(s^3)$.

Consequences:

- SVD of $H^\top G H$ is cheap
- Can be solved with gradient-based algorithms

Experiments: faster KPCA

We solve our dual problem with L-BFGS and compare training time with full SVD, Lanczos method, and Randomized SVD (RSVD).

KPCA Training Time for multiple KPCA problems with fixed $\delta = 10^{-2}$ accuracy. Speedup factor w.r.t. RSVD.

Task	N	Time (s)				Speedup
		SVD	Lanczos	RSVD	Ours	Factor
Synth 1	7000	96.73	0.85	1.97	**0.53**	3.72
Protein	14895	868.64	3.46	6.70	**1.07**	6.25
RCV1	20242	-	6.04	12.50	**2.12**	5.90
CIFAR-10	60000	-	48.10	123.89	**13.51**	9.17

Beyond variance maximization

Typical loss function: $\frac{1}{2}\left\|\cdot\right\|_{\mathrm{F}}^2 \;\rightarrow\;$ ⚠ Sensible to outliers

Modified loss: $L = \frac{1}{2}\left\|\cdot\right\|^2 \,\square\, \Psi$
New KPCA objective:

$$\sup_{W \in \mathcal{S}_{\mathcal{H}}^{s}} L(\Gamma W)$$

With dual problem

$$\inf_{H \in \mathbb{R}^{N \times s}} \frac{1}{2}\,\mathrm{Tr}(H^\top H) + \Psi^\star(H) - \pi(H)$$

Ψ enforces desired properties of the solution, e.g., robustness or sparsity

Solving the dual problem

Use the DC algorithm [Tao+97], with current iterate $H^{(t)}$

- $Y = \nabla\pi(H^{(t)})$
- $H^{(t+1)} = \mathrm{prox}_{\Psi^\star}(Y)$

Enforce robustness by **extended Huber loss**

$$H_\kappa^p := \frac{1}{2}\left\|\cdot\right\|^2 \square\, \kappa \left\|\cdot\right\|_p$$

Fenchel conjugate of the p-norm is the indicator of q-ball, thus

$$\Psi := \kappa \left\|\cdot\right\|_p, \qquad \Psi^\star = \iota_{\mathcal{B}_\kappa^q}, \qquad \mathrm{prox}_{\Psi^\star}(Y) = \mathrm{Proj}_{\mathcal{B}_\kappa^q}(Y).$$

Effect: the coefficients H are forced to pertain to a certain ball (robustness)

Kernel SVD

Corresponding paper:
Tao, Q.*, **Tonin, F.***, Chen, Y., Patrinos, P., & Suykens, J. (2023). Nonlinear SVD with Asymmetric Kernels: feature learning and asymmetric Nyström method. *arXiv:2306.07040*.

SVD vs. KPCA

- Singular Value Decomposition of $A \in \mathbb{R}^{N \times M}$
  - $A = U \Sigma V^\top$
  - Two sets of orthonormal eigenbases U, V
- KPCA of data matrix A
  - Samples are the rows of A: $\{x_i \in \mathbb{R}^M\}_{i=1}^N$
  - Eigendecomposition of kernel matrix $G_{ij} = k(x_i, x_j)$, with symmetric kernel k
  - One set of eigenbases
- Research Question: *How to extend SVD to a nonlinear form through asymmetric kernels?*

Problem Formulation

Given a data matrix $A \in \mathbb{R}^{N \times M}$, it can be seen as an array w.r.t. either rows or columns:

- $\mathcal{X} = \{A[i,:] \triangleq x_i\}_{i=1}^{N}$
- $\mathcal{Z} = \{A[:,j] \triangleq z_j\}_{j=1}^{M}$

SVD gives two sets of embeddings for both $\mathcal{X}$ and $\mathcal{Z}$

KPCA provides only one set of features to rows $\mathcal{X}$

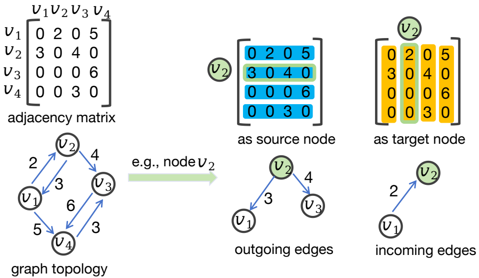


Figure: Example of asymmetric similarity in directed graphs.

Kernel SVD: variational principle

Instead of working with only one feature map of x_i as in KPCA, we apply two maps $\phi\colon \mathbb{R}^M \to \mathbb{R}^p, \psi\colon \mathbb{R}^N \to \mathbb{R}^p$ to both x_i and z_j:

$$x_i \in \mathbb{R}^M \mapsto \phi(x_i) \in \mathbb{R}^p, \quad z_j \in \mathbb{R}^N \mapsto \psi(z_j) \in \mathbb{R}^p.$$

KSVD Primal problem [Suy16]:

$$\boxed{\text{P}} \quad \begin{aligned} \max_{W_e, W_r, e_i, r_j} \quad & J = \frac{1}{2}\sum_{i=1}^{N} e_i^\top \Lambda e_i + \frac{1}{2}\sum_{j=1}^{M} r_j^\top \Lambda r_j - \operatorname{Tr}\left(W_e^\top W_r\right) \\ \text{s.t.} \quad & e_i = W_e^\top \phi(x_i),\ i = 1, \ldots, N, \\ & r_j = W_r^\top \psi(z_j),\ j = 1, \ldots, M \end{aligned}$$

Kernel SVD: asymmetric kernels

- Lagrangian

$$\mathcal{L}(W_e, W_r, e_i, r_j, h_{e_i}, h_{r_j}) = J - \sum_{i=1}^{N} h_{e_i}^\top \left(e_i - W_e^\top \phi(x_i) \right) - \sum_{j=1}^{M} h_{r_j}^\top \left(r_j - W_r^\top \psi(z_j) \right)$$

- Writing the conditions for optimality and eliminating W_e, W_r, e_i, r_j gives the shifted eigenvalue problem

$$\boxed{\text{D}} \quad \begin{array}{l} \left[\varphi(x_i)^T \psi(z_j) \right] H_r = H_e \tilde{\Lambda} \\ \left[\psi(z_j)^T \varphi(x_i) \right] H_e = H_r \tilde{\Lambda} \end{array}$$

Asymmetric kernel

The asymmetric kernel $\kappa : \mathbb{R}^M \times \mathbb{R}^N \to \mathbb{R}$ is defined by the inner product of two feature mappings:

$$\kappa(x, z) = \langle \phi(x), \psi(z) \rangle, \quad \forall x \in \mathbb{R}^M, z \in \mathbb{R}^N,$$

where the output spaces of ϕ, ψ are compatible in dimensionality.

KSVD: solution

Asymmetric $N \times M$ kernel matrix $G_{ij} = \phi(x_i)^\top \psi(z_j) = \kappa(x_i, z_j)$:

$$\boxed{\mathrm{D}} \quad \begin{aligned} G\, H_r &= H_e \tilde{\Lambda} \\ G^\top H_e &= H_r \tilde{\Lambda} \end{aligned}$$

Lanczos' decomposition theorem

Any non-zero rank-r matrix A can be written as $A = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$, with matrices $\tilde{U}, \tilde{\Sigma}, \tilde{V}$ defined by the shifted eigenvalue problem:

$$\begin{aligned} A\tilde{V} &= \tilde{U}\tilde{\Sigma}, \\ A^\top \tilde{U} &= \tilde{V}\tilde{\Sigma}, \end{aligned}$$

where $\tilde{U} \in \mathbb{R}^{N\times r}$ and $\tilde{V} \in \mathbb{R}^{M\times r}$ satisfy $\tilde{U}^\top \tilde{U} = I_r$ and $\tilde{V}^\top \tilde{V} = I_r$, and $\tilde{\Sigma} \in \mathbb{R}^{r\times r}$ is a positive definite diagonal matrix.

- Consequence: the KSVD solution is obtained by the SVD on the asymmetric kernel matrix G

KSVD: model representation

Model-based approach with two representations

$\mathcal{M}$ → **P**

$$e(x) = W_e^\top \phi(x)$$
$$r(z) = W_r^\top \psi(z)$$

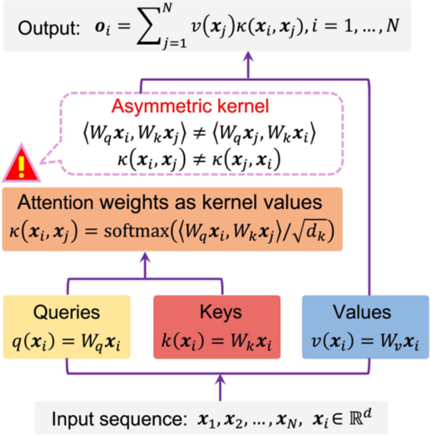
$\mathcal{M}$ → **D**

$$e(x) = \sum_{j=1}^{M} h_{r_j} \kappa(x, z_j)$$
$$r(z) = \sum_{i=1}^{N} h_{e_i} \kappa(x_i, z).$$

Primal-Attention

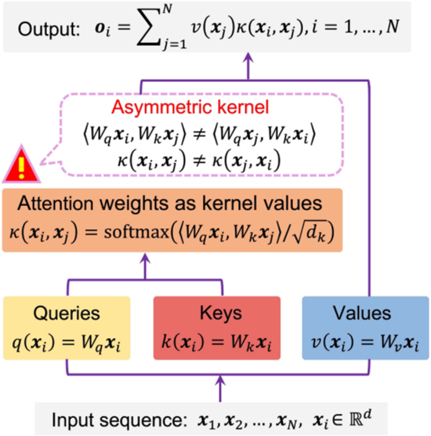
Corresponding paper:
Chen, Y.*, Tao, Q.*, **Tonin, F.**, & Suykens, J. (2023).
Primal-Attention: Self-attention through Asymmetric Kernel SVD in Primal Representation. *NeurIPS 2023*.

Self-attention is asymmetric



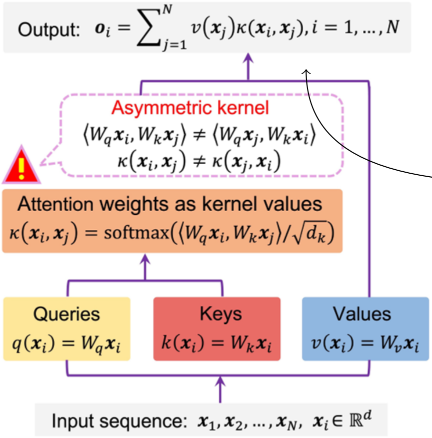
- Attention matrix can be seen as kernel matrix
- Previous works consider symmetric kernels [Tsa+19; Ngu+23]
- However, attention is asymmetric ⚠

Self-attention with asymmetric kernel



- We define two feature maps ϕ_q, ϕ_k related to queries and keys
- The asymmetric kernel for self-attention is $\kappa(x_i, x_j) = \langle \phi_q(x_i), \phi_k(x_j) \rangle$

Connecting self-attention and dual KSVD



Primal-dual representations of KSVD in self-attention:

P
$$\begin{cases} e(x) = W_e^\top \phi_q(x) \\ r(x) = W_r^\top \phi_k(x) \end{cases}$$

D
$$\begin{cases} e(x) = \sum_{j=1}^{N} h_{r_j} \kappa(x, x_j) \\ r(x) = \sum_{i=1}^{N} h_{e_i} \kappa(x_i, x) \end{cases}$$

- The values play the role of the right singular vectors of the attention matrix $v(x_j) =: h_{r_j}$
- Canonical self-attention only outputs e

Primal-Attention

Primal-dual representation of KSVD in self-attention:

$$\boxed{\text{P}} \quad \begin{cases} e(x) = W_e^\top \phi_q(x) \\ r(x) = W_r^\top \phi_k(x) \end{cases}, \qquad \boxed{\text{D}} \quad \begin{cases} e(x) = \sum_{j=1}^{N} h_{r_j} \kappa(x, x_j) \\ r(x) = \sum_{i=1}^{N} h_{e_i} \kappa(x_i, x). \end{cases}$$

Primal-Attention: leveraging primal representation with ϕ_q, ϕ_k:

$$o_i := [e_i; r_i] = \left[W_e^\top \phi_q(x_i); W_r^\top \phi_k(x_i) \right]$$

Primal-Attention

Primal-dual representation of KSVD in self-attention:

$$\boxed{\text{P}} \quad \begin{cases} e(x) = W_e^\top \phi_q(x) \\ r(x) = W_r^\top \phi_k(x) \end{cases}, \qquad \boxed{\text{D}} \quad \begin{cases} e(x) = \sum_{j=1}^{N} h_{r_j} \kappa(x, x_j) \\ r(x) = \sum_{i=1}^{N} h_{e_i} \kappa(x_i, x). \end{cases}$$

Primal-Attention: leveraging primal representation with ϕ_q, ϕ_k:

$$o_i := [e_i; r_i] = \left[W_e^\top \phi_q(x_i); W_r^\top \phi_k(x_i) \right]$$

In experiments we use cosine similarity kernel

$$\phi_q(x) := q(x)/\|q(x)\|_2 \quad \phi_k(x) := k(x)/\|k(x)\|_2$$

Primal-Attention

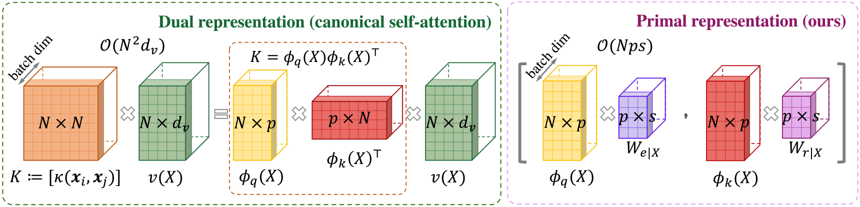
Primal-dual representation of KSVD in self-attention:

$$\boxed{\text{P}} \quad \begin{cases} e(x) = W_e^\top \phi_q(x) \\ r(x) = W_r^\top \phi_k(x) \end{cases}, \qquad \boxed{\text{D}} \quad \begin{cases} e(x) = \sum_{j=1}^{N} h_{r_j} \kappa(x, x_j) \\ r(x) = \sum_{i=1}^{N} h_{e_i} \kappa(x_i, x). \end{cases}$$

Primal-Attention: leveraging primal representation with ϕ_q, ϕ_k:

$$o_i := [e_i; r_i] = \left[W_e^\top \phi_q(x_i); W_r^\top \phi_k(x_i) \right]$$

➔ Result: time complexity reduced from $\mathcal{O}(N^2 d_v)$ to $\mathcal{O}(Nps)$



Primal-Attention objective

The Primal-Attention objective combines the task-oriented loss L and the KSVD primal objective J_l

$$J_{\text{PrimalAtt}} = L + \eta \sum\nolimits_l J_l^2,$$

where the second term adds objectives of all Primal-Attention blocks and J_l is implemented as mean over all heads

$$\begin{aligned} J_l(W_e, W_r, \Lambda) &= \frac{1}{2}\sum_{i=1}^{N} e_i^\top \Lambda e_i + \frac{1}{2}\sum_{j=1}^{N} r_j^\top \Lambda r_j - \operatorname{Tr}\left(W_e^\top W_r\right) \\ &= \frac{1}{2}\sum_{i=1}^{N} \|(W_e \Lambda^{\frac{1}{2}})^\top \phi_q(x_i)\|_2^2 + \frac{1}{2}\sum_{j=1}^{N} \|(W_r \Lambda^{\frac{1}{2}})^\top \phi_k(x_j)\|_2^2 - \operatorname{Tr}\left(W_e^\top W_r\right). \end{aligned}$$

Motivated by

Lemma (A zero-value objective with stationary solutions)

The solutions to the KSVD shifted eigenvalue problem in the dual representation lead to the zero-value primal objective J_l.

Experiments: higher reward in D4RL

D4RL benchmark: offline RL performance for continuous robot control tasks
Three different environments: HalfCheetah, Hopper and Walker, under three policies: Medium-Expert, Medium and Medium-Replay

Dataset	Environment	DT	Linear.	Re.	Per.	Cos.	Flow.	**Ours**
Medium -Expert	HalfCheetah	83.8±3.3	78.2±3.2	81.5±1.6	85.1±2.1	85.5±2.9	90.8±0.4	77.8±22.1
	Hopper	104.0±2.5	107.2±0.9	104.2±9.8	93.5±13.9	98.1±7.4	109.9±1.0	111.5±0.2
	Walker	107.7±0.6	67.2±27.3	71.4±1.8	72.6±2.4	100.5±14.5	108.0±0.4	108.9±0.1
Medium	HalfCheetah	42.4±0.1	42.3±0.2	42.2±0.1	42.1±0.2	42.1±0.3	42.2±0.2	43.0±0.0
	Hopper	64.2±1.1	58.7±0.4	59.9±0.7	59.7±7.5	59.8±3.8	66.9±2.5	74.5±0.6
	Walker	70.6±3.2	57.9±10.6	65.8±4.9	63.3±10.7	71.4±1.2	71.7±2.5	77.9±7.8
Medium -Replay	HalfCheetah	34.6±0.6	32.1±1.5	33.6±0.7	31.7±0.9	32.8±3.6	34.7±1.5	38.9±0.4
	Hopper	79.7±7.4	74.3±7.0	66.1±2.6	64.6±24.2	59.3±16.5	75.5±14.5	88.5±12.5
	Walker	62.9±5.0	62.1±7.4	50.1±3.5	61.3±6.7	60.5±9.9	62.0±3.1	76.8±10.3
Average Reward		72.2±**2.6**	64.4±6.5	63.9±2.9	63.8±7.6	67.8±7.6	73.5±2.9	**77.5**±6.0

Experiments: large-scale language modelling

Language modelling on WikiText-103. **157M** parameters

6 layers, 512 attention channels, 2048 FC channels, 267744 dictionary size $\rightarrow 6(4 \cdot 512^2 + 2 \cdot 512 \cdot 2048) + 512 \cdot 267744$

Models grow large quickly...

Model	Perplexity	Time (s/1K-steps)	Memory (GB)
Transformer	33.0	3108.4	9.0
Flowformer	**30.8**	3998.4	10.5
Primal+Trans.	31.0	**3104.0**	**8.9**

Conclusion

Take home message

Primal-dual model representations are powerful

- Faster KPCA algorithm and convolution with p-norms induces robustness
- Primal-dual representation of self-attention through KSVD avoids computing attention matrix
- Primal-Attention: higher accuracy & efficiency

Future perspectives

- Robust KSVD through dualization of difference of convex functions

Future perspectives

- Robust KSVD through dualization of difference of convex functions $\rightarrow$ robust self-attention ?

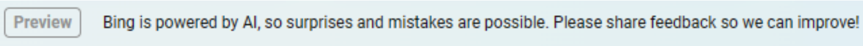
Future perspectives

- Robust KSVD through dualization of difference of convex functions $\rightarrow$ robust self-attention ?
- Uncertainty estimation in Transformers

Future perspectives

- Robust KSVD through dualization of difference of convex functions $\rightarrow$ robust self-attention ?
- Uncertainty estimation in Transformers

Would you trust a system that says it's unreliable?



Future perspectives

- Robust KSVD through dualization of difference of convex functions $\rightarrow$ robust self-attention ?
- Uncertainty estimation in Transformers
- Compressing LLMs for faster inference/adaptation through low-rank properties

Thanks for your attention!

- Support from ERC AdG E-DUALITY, KU Leuven, FWO projects, iBOF, Leuven.AI
- Thanks to Johan Suykens, Yingyi Chen, Alex Lambert, Qinghua Tao for their materials/slides, and to my advisors and collaborators at ESAT-STADIUS: Johan Suykens, Panos Patrinos; Sonny Achten, Arun Pandey, and others

References

References I

[Che+23] Yingyi Chen et al. “Primal-Attention: Self-attention through Asymmetric Kernel SVD in Primal Representation”. In: *Advances in Neural Information Processing Systems*. 2023 (cit. on p. 2).

[Chi+23] Krishna Teja Chitty-Venkata et al. “A Survey of Techniques for Optimizing Transformer Inference”. In: *Journal of Systems Architecture* (2023) (cit. on p. 4).

[Dos+21] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations*. 2021 (cit. on p. 3).

[Mik+99] Sebastian Mika et al. "Kernel PCA and De-Noising in Feature Spaces". In: *Advances in Neural Information Processing Systems*. MIT Press, 1999, pp. 536–542 (cit. on p. 11).

[Ngu+23] Tan Minh Nguyen et al. "A Primal-Dual Framework for Transformers and Neural Networks". In: *The Eleventh International Conference on Learning Representations*. 2023 (cit. on p. 26).

[Sch+98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. "Nonlinear Component Analysis as a Kernel Eigenvalue Problem". In: *Neural Computation* 10.5 (1998), pp. 1299–1319 (cit. on p. 11).

[Suy+02] Johan Suykens et al. *Least Squares Support Vector Machines*. World Scientific, Nov. 2002 (cit. on p. 12).

References III

[Suy16] Johan AK Suykens. "SVD Revisited: A New Variational Principle, Compatible Feature Maps and Nonlinear Extensions". In: *Applied and Computational Harmonic Analysis* 40.3 (2016), pp. 600–609 (cit. on p. 21).

[Tao+23] Qinghua Tao et al. *Nonlinear SVD with Asymmetric Kernels: feature learning and asymmetric Nyström method*. 2023. arXiv: `2306.07040 [cs.LG]` (cit. on p. 2).

[Tao+97] Pham Dinh Tao and Le Thi Hoai An. "Convex Analysis Approach to DC Programming: Theory, Algorithms and Applications". In: *Acta mathematica vietnamica* 22.1 (1997), pp. 289–355 (cit. on p. 17).

References IV

[Ton+23] Francesco Tonin et al. "Extending Kernel PCA through Dualization: Sparsity, Robustness and Fast Algorithms". In: *International Conference on Machine Learning*. PMLR. 2023 (cit. on p. 2).

[Tsa+19] Yao-Hung Hubert Tsai et al. "Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel". In: *EMNLP-IJCNLP*. Ed. by Kentaro Inui et al. ACL, Nov. 2019 (cit. on p. 26).