Self-Attention through Kernel-Eigen Pair Sparse Variational Gaussian Processes

Yingyi Chen*,[1] Qinghua Tao*,[1] Francesco Tonin[2] Johan A.K. Suykens[1]

*Equal contribution [1]ESAT-STADIUS, KU Leuven, Belgium [2]LIONS, EPFL, Switzerland (most work done at ESAT-STADIUS, KU Leuven)

KU LEUVEN

European Research Council
Established by the European Commission

Summary

Building uncertainty-aware self-attention in Transformers with efficiency:

- Large capacities of Transformers can lead to overconfident predictions, risking of safety-critical issues;
- Bayesian inference, a good uncertainty quantification tool, alleviates overconfidence by providing predictions with confidence scores;
- We propose a new Bayesian self-attention based on *Sparse Variational Gaussian Processes* (SVGP);
- The time-complexity of our Bayesian self-attention is further reduced to $\mathcal{O}(s), s < N$ with *Kernel Singular Value Decomposition* (KSVD).

Background I: SVGP

Gaussian Process (GP) represents real-valued function $f(\cdot): \mathcal{X} \rightarrow \mathbb{R}$ with Gaussian distributions based on $\kappa(\cdot,\cdot): \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, **positive-definite kernel**:

Prior: $f(\cdot) \sim \mathcal{GP}(0, \kappa(\cdot,\cdot)) \Rightarrow \boldsymbol{f} \sim \mathcal{N}(\boldsymbol{0}, K_{XX}),\ K_{XX} := [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)] \in \mathbb{R}^{N\times N}$

Posterior: $\boldsymbol{f}^* | X^*, X, \boldsymbol{y} \sim \mathcal{N}(K_{X^*X}(K_{XX} + \sigma^2 I_N)^{-1}\boldsymbol{y},$
$K_{X^*X^*} - K_{X^*X}(K_{XX} + \sigma^2 I_N)^{-1} K_{XX^*})$

- Time complexity of computing posterior is $\mathcal{O}(N^3)$.

Sparse Variational Gaussian Process (SVGP) variationally approximates GP posterior with s inducing variables $\{Z_1, \dots, Z_s\} \in \mathcal{X}$, $\boldsymbol{u}[i] = f(Z_i)$:

Prior: $\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{u} \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K_{XX} & K_{XZ} \\ K_{ZX} & K_{ZZ} \end{bmatrix}\right)$

Posterior: $\mathrm{q}(\boldsymbol{f}) = \mathcal{N}(K_{XZ}K_{ZZ}^{-1}\boldsymbol{m}_{\boldsymbol{u}}, K_{XX} - K_{XZ}K_{ZZ}^{-1}(K_{ZZ} - S_{\boldsymbol{uu}})K_{ZZ}^{-1}K_{ZX})$

- Posterior is based on $\mathrm{q}(\boldsymbol{f}) = \int \mathrm{p}(\boldsymbol{f}|\boldsymbol{u})\mathrm{q}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}$ with variational distribution

$$\mathrm{q}(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{m}_{\boldsymbol{u}}, S_{\boldsymbol{uu}}),\ \boldsymbol{m}_{\boldsymbol{u}} \in \mathbb{R}^s,\ S_{\boldsymbol{uu}} \in \mathbb{R}^{s\times s}.$$

- Evidence lower-bound: $\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{\mathrm{q}(\boldsymbol{f})}[\log \mathrm{p}(\boldsymbol{y}|\boldsymbol{f})] - \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \,||\, \mathrm{p}(\boldsymbol{u}))$
- Time complexity of computing posterior is $\mathcal{O}(s^3)$, $s < N$.

SVGP with Kernel-Eigen Features reduces time complexity by choosing inducing variables as the eigenvectors of K_{XX}, i.e., $\boldsymbol{u}[i] := \boldsymbol{v}_i$:

Prior: $\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{u} \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K_{XX} & H\Lambda \\ \Lambda H^\top & \Lambda \end{bmatrix}\right)$

Posterior: $\mathrm{q}(\boldsymbol{f}) = \mathcal{N}((H\Lambda)\Lambda^{-1}\boldsymbol{m}_{\boldsymbol{u}}, K_{XX} - (H\Lambda)\Lambda^{-1}(\Lambda - S_{\boldsymbol{uu}})\Lambda^{-1}(\Lambda H^\top))$

- $H := [\boldsymbol{v}_1, \dots, \boldsymbol{v}_s] \in \mathbb{R}^{N\times s}$ contains the eigenvectors to the top-s nonzero eigenvalues of K_{XX}, i.e., $\Lambda = \mathrm{diag}\{\lambda_1, \dots, \lambda_s\}$.
- Time complexity of computing posterior is $\mathcal{O}(s)$, $s < N$.

Background II: KSVD

Self-Attention corresponds to Asymmetric Kernel: let $\{x_i \in \mathbb{R}^d\}_{i=1}^N$ be the inputs, then the queries, keys and values are

$$q(\boldsymbol{x}_i) = W_q \boldsymbol{x}_i,\quad k(\boldsymbol{x}_i) = W_k \boldsymbol{x}_i,\quad v(\boldsymbol{x}_i) = W_v \boldsymbol{x}_i.$$

The canonical self-attention is with attention weights:

$$\kappa_{\mathrm{att}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathrm{softmax}\left(\langle W_q \boldsymbol{x}_i, W_k \boldsymbol{x}_j \rangle / \sqrt{d_k}\right),\quad i, j = 1, \dots, N,$$

where $\kappa_{\mathrm{att}}(\cdot,\cdot): \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ serves as kernel function. Notice that in general,

$$\langle W_q \boldsymbol{x}_i, W_k \boldsymbol{x}_j \rangle \neq \langle W_q \boldsymbol{x}_j, W_k \boldsymbol{x}_i \rangle \Rightarrow \kappa_{\mathrm{att}}(\boldsymbol{x}_i, \boldsymbol{x}_j) \neq \kappa_{\mathrm{att}}(\boldsymbol{x}_j, \boldsymbol{x}_i),$$

κ_{att} is asymmetric kernel function[1]. Output is $o(\boldsymbol{x}) = \sum_{j=1}^N v(\boldsymbol{x}_j)\kappa_{\mathrm{att}}(\boldsymbol{x}, \boldsymbol{x}_j)$.

Kernel-Eigen Pair Sparse Variational Process

Pair of Adjoint Eigenfunctions for Self-Attention: the self-attention corresponds to a shifted eigenvalue problem[1,2] w.r.t. attention matrix

$$K_{\mathrm{att}} H_r = H_e \Lambda \qquad K_{\mathrm{att}}^\top H_e = H_r \Lambda$$

Shifted eigenvalue problem w.r.t asymmetric kernel K_{att}.

Equiv.

$$(K_{\mathrm{att}} K_{\mathrm{att}}^\top) H_e = H_e \Lambda^2 \qquad (K_{\mathrm{att}}^\top K_{\mathrm{att}}) H_r = H_r \Lambda^2$$

Eigendecompositions w.r.t. symmetric kernels $K_{\mathrm{att}} K_{\mathrm{att}}^\top$, $K_{\mathrm{att}}^\top K_{\mathrm{att}}$.

Asymmetric…no SVGP

Symmetric…two SVGPs

Two SVGPs with adjoint kernel-eigen features:

SVGP w.r.t. right singular vectors

Prior: $\begin{pmatrix} \boldsymbol{f}^e \\ \boldsymbol{u}^e \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K_{\mathrm{att}} K_{\mathrm{att}}^\top & H_e \Lambda^2 \\ \Lambda^2 H_e^\top & \Lambda^2 \end{bmatrix}\right)$

Posterior: $\mathrm{q}(\boldsymbol{f}^e) \approx \mathcal{N}(\underbrace{e(X)\Lambda^{-1}\boldsymbol{m}_{\boldsymbol{u}}}_{\boldsymbol{\mu}^e}, \underbrace{e(X)\Lambda^{-2} S_{\boldsymbol{uu}} e(X)^\top}_{\Sigma^e := L^e (L^e)^\top})$

SVGP w.r.t. left singular vectors

Prior: $\begin{pmatrix} \boldsymbol{f}^r \\ \boldsymbol{u}^r \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K_{\mathrm{att}}^\top K_{\mathrm{att}} & H_r \Lambda^2 \\ \Lambda^2 H_r^\top & \Lambda^2 \end{bmatrix}\right)$

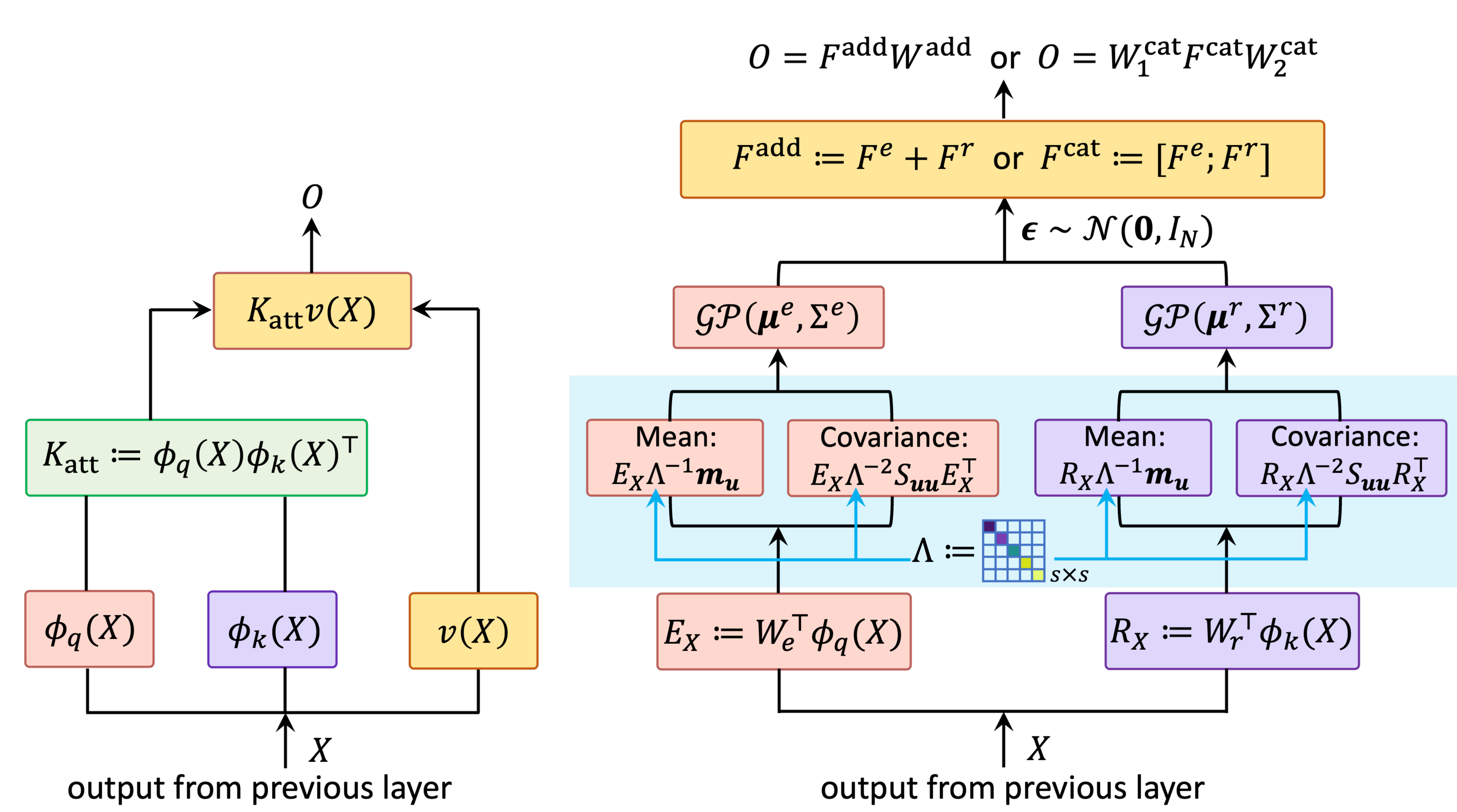
Posterior: $\mathrm{q}(\boldsymbol{f}^r) \approx \mathcal{N}(\underbrace{r(X)\Lambda^{-1}\boldsymbol{m}_{\boldsymbol{u}}}_{\boldsymbol{\mu}^r}, \underbrace{r(X)\Lambda^{-2} S_{\boldsymbol{uu}} r(X)^\top}_{\Sigma^r := L^r (L^r)^\top})$

Outputs of the two SVGPs are obtained by the Monte-Carlo sampling:

$$F^e = \boldsymbol{\mu}^e + L^e \boldsymbol{\epsilon},\quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I_N);\qquad F^r = \boldsymbol{\mu}^r + L^r \boldsymbol{\epsilon},\quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I_N).$$

Merge two SVGP outputs either by addition or concatenation schemes:

Addition: $F^{\mathrm{add}} := F^e + F^r \in \mathbb{R}^N$; Concatenation: $F^{\mathrm{cat}} := [F^e; F^r] \in \mathbb{R}^{2N}$.



(a) Canonical Transformer (b) KEP-SVGP

Self-Attention with KSVD: let $\kappa_{\mathrm{att}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi_q(\boldsymbol{x}_i), \phi_k(\boldsymbol{x}_j) \rangle$, then the primal-dual representations of self-attention with KSVD gives[1]

Equiv. under reg. $\mathcal{L}_{\mathrm{KSVD}}$

$$\text{Primal: } \begin{cases} e(\boldsymbol{x}) = W_e^\top \phi_q(\boldsymbol{x}) \\ r(\boldsymbol{x}) = W_r^\top \phi_k(\boldsymbol{x}) \end{cases} \quad \text{Dual: } \begin{cases} e(\boldsymbol{x}) = \sum_{j=1}^N \boldsymbol{h}_{r_j} \kappa_{\mathrm{att}}(\boldsymbol{x}, \boldsymbol{x}_j) \\ r(\boldsymbol{x}) = \sum_{i=1}^N \boldsymbol{h}_{e_i} \kappa_{\mathrm{att}}(\boldsymbol{x}_i, \boldsymbol{x}) \end{cases}$$

where $H_e := [\boldsymbol{h}_{e_1}, \dots, \boldsymbol{h}_{e_N}]^\top$, $H_r := [\boldsymbol{h}_{r_1}, \dots, \boldsymbol{h}_{r_N}]^\top \in \mathbb{R}^{N\times s}$ are column-wisely the left and right singular vectors of the attention matrix K_{att}.

Experiments

- Good, robust and efficient performances on *in-distribution*, *distribution-shift* and *out-of-distribution* benchmarks.
- Methods in comparison:
  - *i)* uncertainty estimation baselines implemented into transformers;
  - *ii)* deep kernel learning implemented into transformers;
  - *iii)* Bayesian transformers.
- Rationales behind KEP-SVGP's good performance in
  - *i)* Distribution-shift robustness: KSVD filters out noisy patterns;
  - *ii)* OOD detection: KSVD differentiates different eigen spaces.

In-distribution data:

Dataset	Method	ACC (↑)	AURC (↓)	AUROC (↑)	FPR95 (↓)
CIFAR-10 [Krizhevsky et al., 2009]	MSP [Hendrycks & Gimpel, 2017]	83.50 ± 0.43	42.60 ± 1.84	86.15 ± 0.35	66.51 ± 2.19
	Temp. Scaling [Guo et al., 2017]	83.50 ± 0.43	40.47 ± 1.63	86.55 ± 0.36	65.10 ± 2.23
	MC Dropout [Gal&Ghahramani, 2016]	83.69 ± 0.51	41.36 ± 1.45	86.18 ± 0.28	66.49 ± 1.96
	KFLLLA [Kristiadi et al., 2020]	83.54 ± 0.45	40.12 ± 1.65	86.70 ± 0.50	$\mathbf{63.13} \pm 1.75$
	SV-DKL [Wilson et al., 2016]	83.82 ± 0.58	39.78 ± 1.91	86.57 ± 0.38	65.02 ± 1.33
	KEP-SVGP (ours)	$\mathbf{84.70} \pm 0.61$	$\mathbf{35.15} \pm 2.65$	$\mathbf{87.20} \pm 0.65$	64.93 ± 1.41
IMDB [Maas et al., 2011]	MSP [Hendrycks & Gimpel, 2017]	88.17 ± 0.52	35.27 ± 3.04	82.29 ± 0.87	71.41 ± 1.57
	Temp. Scaling [Guo et al., 2017]	88.17 ± 0.52	35.27 ± 3.04	82.29 ± 0.87	71.08 ± 1.55
	MC Dropout [Gal&Ghahramani, 2016]	88.34 ± 0.65	34.62 ± 3.17	82.24 ± 0.83	71.65 ± 2.03
	KFLLLA [Kristiadi et al., 2020]	88.17 ± 0.52	35.20 ± 3.01	82.31 ± 0.86	71.07 ± 1.51
	SV-DKL [Wilson et al., 2016]	88.86 ± 1.04	59.84 ± 18.90	73.20 ± 5.56	69.91 ± 3.68
	SGPA [Chen&Li, 2023]	88.36 ± 0.75	33.14 ± 3.46	82.78 ± 0.44	70.85 ± 2.46
	KEP-SVGP (ours)	$\mathbf{89.01} \pm 0.14$	$\mathbf{30.69} \pm 0.69$	$\mathbf{83.22} \pm 0.31$	$\mathbf{68.15} \pm 0.95$

Distribution-shift data:

Out-of-distribution detection with AUROC (↑):

ID	CIFAR-10			CIFAR-100		
OOD	SVHN	CIFAR-100	LSUN	SVHN	CIFAR-10	LSUN
MSP	**86.56**	81.50	87.48	75.83	67.14	74.97
MC Dropout	**86.56**	81.67	88.19	76.62	**67.54**	74.94
KFLLLA	75.95	75.67	80.00	72.81	65.37	71.25
SV-DKL	75.48	76.81	82.02	74.35	65.72	72.03
KEP-SVGP (ours)	84.75	**82.32**	**91.50**	**79.98**	67.51	**78.22**

Paper:

Code:

References:

[1] Chen et al. "Primal-Attention: self-attention through asymmetric kernel SVD in primal representation." NeurIPS, 2023.

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

Correspondence to: Yingyi Chen (yingyi.chen@esat.kuleuven.be)