Flow-based Spectral Kernel Learning for Nonstationary Attention

香 港 大 學
THE UNIVERSITY OF HONG KONG

Zi Yang[1,*], Ying Li[2,*], Alejandro Lancho[3], Darío C. Larese[3], Pablo M. Olmos[3,†], Michael Minyi Zhang[1,†]

[1]The University of Hong Kong [2]National University of Singapore [3]Universidad Carlos III de Madrid

[*]Equal contribution [†]Corresponding authors

NeurIPS 2026

uc3m | Universidad Carlos III de Madrid

Motivation: Learn the Attention Kernel

- **Flexible similarities.** Fixed or stationary kernels restrict input-dependent token interactions.
- **Scalable attention.** Dense pairwise attention costs quadratically in sequence length L.
- **Our approach.** Learn a nonstationary spectral density and factorize attention with random features.

Kernel expressiveness can simplify attention.

1. Nonstationary Spectral Features

The generalized spectral representation uses a **bivariate** density $p_\theta(\omega_1, \omega_2)$. It models similarities beyond functions of $x - x'$.

Theorem 1. Unbiased signed features

Let k_θ be a real-valued kernel admitting the paper's spectral representation with a normalized spectral probability density. For even M, draw $M/2$ i.i.d. frequency pairs and define

$$\phi_\theta(x) = \frac{1}{\sqrt{2M}} \begin{bmatrix} \cos(\Omega_1 x) + \cos(\Omega_2 x) \\ \sin(\Omega_1 x) + \sin(\Omega_2 x) \end{bmatrix} \in \mathbb{R}^M.$$

Then the feature inner product is unbiased:

$$\mathbb{E}_\Omega\left[\phi_\theta(x)^\top \phi_\theta(x')\right] = k_\theta(x, x').$$

2. Flow-based Spectral Density Learning

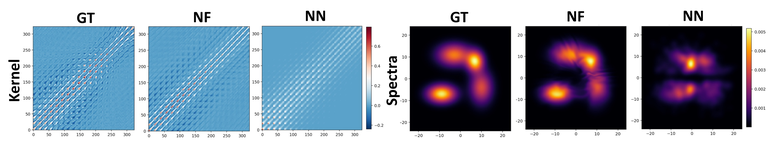
RealNVP flows map a Gaussian base to paired frequencies and provide a tractable spectral density:

$$z \sim \mathcal{N}(0, I), \quad (\omega_1, \omega_2) = F_\theta(z),$$
$$p_\theta(\omega_1, \omega_2) = p_0(z) \left|\det \frac{\partial F_\theta(z)}{\partial z}\right|^{-1}.$$

A bivariate Gaussian-mixture reference (NG-SM) regularizes the spectrum while the task loss learns useful token similarities:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{spec}} \mathrm{KL}(p_\theta \| p_{\text{NGSM}}).$$

Learned kernels and spectral densities



GT: ground truth; NF: normalizing flow; NN: implicit neural model. NF recovers more of the target spectral structure and kernel pattern.

Highlight. Learn the kernel through its spectrum; retain explicit features and density regularization.

3. Absorbing Query–Key Projections

Standard attention adapts similarity through W_Q, W_K. An expressive function family can instead approximate the projected similarity **directly on the input space**.

Theorem 2. Q/K projection absorption

Let $\mathcal{X}$ be compact and $\mathcal{K}$ uniformly dense in $C(\mathcal{X} \times \mathcal{X})$. Let $k \in C(\mathcal{U} \times \mathcal{V})$, with $\mathcal{U}, \mathcal{V}$ compact and $\mathcal{X}W_Q \subset \mathcal{U}$, $\mathcal{X}W_K \subset \mathcal{V}$. Define

$$f(x, x') = k(xW_Q, x'W_K).$$

For every $\epsilon > 0$, there is $\tilde{k}_\epsilon \in \mathcal{K}$ such that

$$\sup_{x, x' \in \mathcal{X}} \left| k(xW_Q, x'W_K) - \tilde{k}_\epsilon(x, x') \right| \le \epsilon.$$

Proof idea. The projected similarity is continuous on the compact input domain; uniform density yields the approximation.

Scope. The theorem assumes uniform density; it does not establish this property for a finite FSKA implementation.

Highlight. Move similarity adaptation into the learned kernel and remove explicit learned Q/K.

4. Fixed V and Diverse Attention Heads

Q/K controls similarity; V defines the aggregation space. Using raw X as values gives all heads the same representation space. Fixed orthogonal projections retain complementary value subspaces.

For head h, learn its kernel and keep $W_V^{(h)}$ fixed:

$$\alpha_{ij}^{(h)} = \frac{\widetilde{k}_{\theta_h}(x_i, x_j)}{\sum_{j'} \widetilde{k}_{\theta_h}(x_i, x_{j'})}, \qquad y_i^{(h)} = \sum_j \alpha_{ij}^{(h)} x_j W_V^{(h)},$$
$$(W_V^{(h)})^\top W_V^{(h')} = 0 \ (h \neq h'), \qquad y_i = \mathrm{Concat}_{h=1}^{H} y_i^{(h)}.$$

Design principle. Head-specific kernels learn *which tokens to aggregate*; fixed V preserves *where their information is represented.*

5. Positive Features, Linear Aggregation

Apply the positive feature map $\psi_h = \mathrm{ELU}(\phi_{\theta_h}) + 1$ to avoid signed cancellation. Let $v_j^{(h)} = x_j W_V^{(h)}$ (a row vector):

$$S_h = \sum_j \psi_h(x_j) v_j^{(h)}, \qquad z_h = \sum_j \psi_h(x_j),$$
$$\widetilde{k}_{\theta_h}(x, x') = \psi_h(x)^\top \psi_h(x'), \qquad y_i^{(h)} = \frac{\psi_h(x_i)^\top S_h}{\psi_h(x_i)^\top z_h}.$$

Linear cost: $O(LMd_v)$ aggregation per head. Theorem 1 applies to signed features *before* ELU+1, not to the transformed affinity or normalized attention ratio.

6. Ablation Study

Architectural choices on MNIST

Model	Q/K	Learn V	Flex. kernel	PRF	Accuracy
Transformer	Yes	Yes	No	–	0.8556
Transformer w/o QKV	No	No	No	–	0.5267
IKAN	Yes	Yes	Yes	No	0.8656
FSKA	No	No	Yes	Yes	**0.8722**
+ Direct RFF	No	No	Yes	No	0.3956
+ Learnable V	No	Yes	Yes	Yes	0.8725

- FSKA retains accuracy with simpler projections.
- Learning V adds little here (0.8722 to 0.8725).
- Direct signed RFF degrades performance.

7. Sentence Classification

Macro-F1 (%): mean ± SD over three runs.

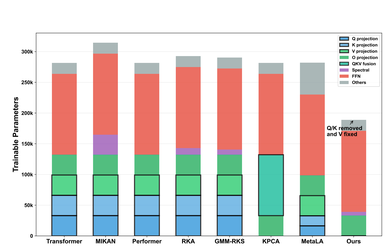
Model	DBPedia	AGNews	TREC	RT-Pol.	SST-2	SST-5
Transformer	98.13 ± 0.15	90.43 ± 0.07	86.14 ± 0.93	70.52 ± 0.45	76.07 ± 1.10	35.11 ± 1.46
MIKAN	97.74 ± 0.01	91.15 ± 0.24	85.47 ± 1.07	71.30 ± 0.92	76.04 ± 0.86	32.84 ± 0.39
Performer	98.12 ± 0.09	90.23 ± 0.31	85.70 ± 0.57	70.88 ± 0.62	77.24 ± 0.41	35.92 ± 0.96
RKA	98.26 ± 0.01	90.80 ± 0.28	85.56 ± 0.75	71.15 ± 0.72	77.08 ± 1.46	36.10 ± 0.49
GMM-RKS	96.76 ± 0.21	90.58 ± 0.41	85.36 ± 0.91	71.05 ± 1.02	76.81 ± 0.89	36.42 ± 0.84
KPCA	97.59 ± 0.14	90.20 ± 0.49	85.90 ± 0.96	72.16 ± 1.58	77.62 ± 1.13	36.97 ± 1.36
MetaLA	97.24 ± 0.15	91.14 ± 0.36	85.07 ± 1.02	70.28 ± 1.71	77.03 ± 0.69	35.27 ± 1.41
FSKA	98.33 ± 0.13	91.20 ± 0.37	86.21 ± 1.02	72.08 ± 1.44	78.74 ± 1.15	38.16 ± 1.40

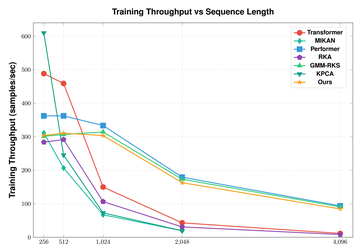
Highest mean macro-F1 on five of six datasets. FSKA improves SST-2 from 76.07 to 78.74 and SST-5 from 35.11 to 38.16 over Transformer.

8. Parameter and Runtime Efficiency

Trainable parameters

Training throughput





Left: original parameter breakdown. Removing learned Q/K and fixing V reduces projection parameters.
Right: higher throughput is faster. Synthetic LRA-style benchmark, lengths 256–4096; successful measurements only.

Adaptive similarities with simpler projections.

Conclusion. FSKA learns a regularized nonstationary kernel, simplifies attention projections, and retains feature-based linear aggregation.

Correspondence: pamartin@ing.uc3m.es mzhang18@hku.hk

Paper: openreview.net/forum?id=aP71hsUN95

Paper
OpenReview →