Fast Neural Kernel Embeddings for General Activations

Insu Han

Yale University

Joint work with Amir Zandieh, Jaehoon Lee,

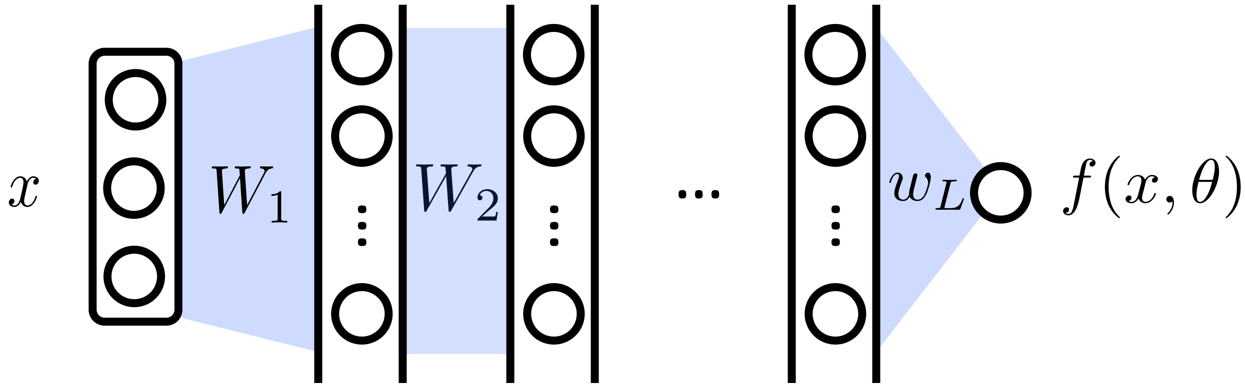
Roman Novak, Lechao Xiao, Amin Karbasi

Neural Kernels

- Fully-connected neural network:

$$f(x,\theta) = h_L^\top w_L, \quad h_\ell = \frac{1}{\sqrt{m}}\sigma(h_{\ell-1}^\top W_\ell), \quad h_0 = x$$

  - $\theta = (W_1, \dots, w_L)$: trainable parameters, m : network width



Neural Kernels

- Fully-connected neural network:

$$f(x,\theta) = h_L^\top w_L, \quad h_\ell = \frac{1}{\sqrt{m}} \sigma(h_{\ell-1}^\top W_\ell), \quad h_0 = x$$

  - $\theta = (W_1, \dots, w_L)$: trainable parameters, m : network width

- **Neural Tangent Kernel (NTK)**:

$$\mathop{\mathbb{E}}_{\theta \sim \mathcal{N}(0,I)} \left\langle \frac{\partial f(x,\theta)}{\partial \theta}, \frac{\partial f(y,\theta)}{\partial \theta} \right\rangle = \mathrm{NTK}_\sigma(x,y)$$

Neural Kernels

- Fully-connected neural network:

$$f(x,\theta) = h_L^\top w_L, \quad h_\ell = \frac{1}{\sqrt{m}}\sigma(h_{\ell-1}^\top W_\ell), \quad h_0 = x$$

  - $\theta = (W_1, \dots, w_L)$: trainable parameters, m : network width

- **Neural Tangent Kernel (NTK)** under infinite-width limit:

$$\lim_{m\to\infty} \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(0,I)} \left\langle \frac{\partial f(x,\theta)}{\partial\theta}, \frac{\partial f(y,\theta)}{\partial\theta} \right\rangle = \mathrm{NTK}_\sigma(x,y)$$

Neural Kernels

- Fully-connected neural network:

$$f(x,\theta) = h_L^\top w_L, \quad h_\ell = \frac{1}{\sqrt{m}}\sigma(h_{\ell-1}^\top W_\ell), \quad h_0 = x$$

  - $\theta = (W_1, \ldots, w_L)$: trainable parameters, m : network width

- **Neural Tangent Kernel (NTK)** under infinite-width limit:

$$\lim_{m\to\infty} \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(0,I)} \left\langle \frac{\partial f(x,\theta)}{\partial\theta}, \frac{\partial f(y,\theta)}{\partial\theta} \right\rangle = \mathrm{NTK}_\sigma(x,y)$$

- **Neural Network Gaussian Process (NNGP) Kernel**:

$$\lim_{m\to\infty} \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(0,I)} \langle f(x,\theta), f(y,\theta)\rangle = \mathrm{NNGP}_\sigma(x,y)$$

Neural Kernels

- Fully-connected neural network:

$$f(x,\theta) = h_L^\top w_L, \quad h_\ell = \frac{1}{\sqrt{m}}\sigma(h_{\ell-1}^\top W_\ell), \quad h_0 = x$$

  - $\theta = (W_1, \ldots, w_L)$: trainable parameters, m : network width

- **Neural Tangent Kernel (NTK)** under infinite-width limit:

$$\lim_{m\to\infty} \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(0,I)} \left\langle \frac{\partial f(x,\theta)}{\partial\theta}, \frac{\partial f(y,\theta)}{\partial\theta} \right\rangle = \mathrm{NTK}_\sigma(x,y)$$

- **Neural Network Gaussian Process (NNGP) Kernel**:

$$\lim_{m\to\infty} \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(0,I)} \langle f(x,\theta), f(y,\theta)\rangle = \mathrm{NNGP}_\sigma(x,y)$$

- Neural Kernels = {NTK, NNGP}

Neural Kernels

- Most infinitely wide neural kernels are based on the ReLU activation

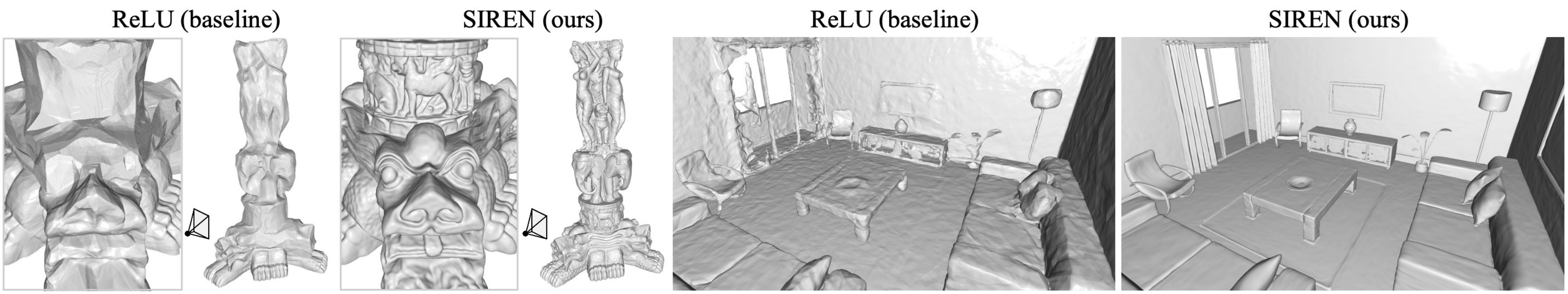
$$f(x, \theta) = h_L^\top w_L, \quad h_\ell = \frac{1}{\sqrt{m}} \sigma(h_{\ell-1}^\top W_\ell), \quad h_0 = x$$

Neural Kernels

- Most infinitely wide neural kernels are based on the ReLU activation

$$f(x,\theta) = h_L^\top w_L, \quad h_\ell = \frac{1}{\sqrt{m}}\sigma(h_{\ell-1}^\top W_\ell), \quad h_0 = x$$

- However, periodic activations are good for representing complex natural signals such as images, wavefields, video, sound [SMB+20]

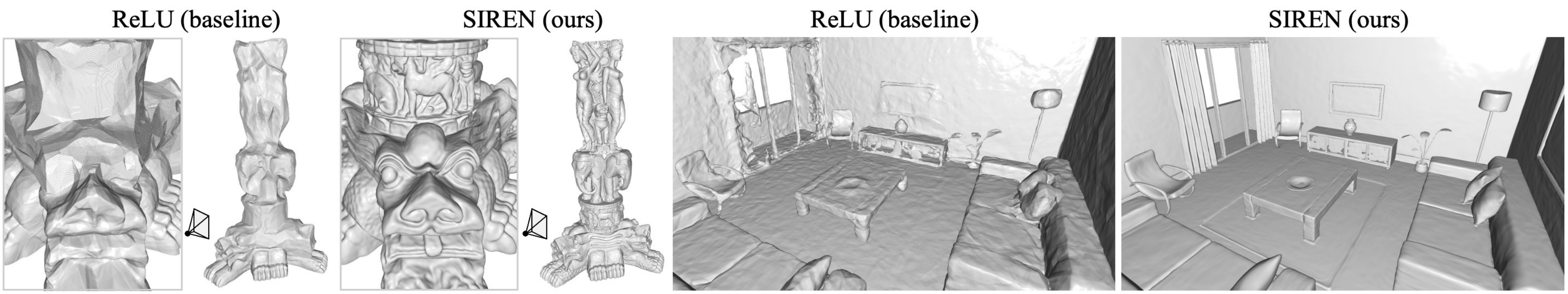


Neural Kernels

- Most infinitely wide neural kernels are based on the ReLU activation

$$f(x,\theta) = h_L^\top w_L, \quad h_\ell = \frac{1}{\sqrt{m}}\sigma(h_{\ell-1}^\top W_\ell), \quad h_0 = x$$

- However, periodic activations are good for representing complex natural signals such as images, wavefields, video, sound [SMB+20]



- **Questions:** can we compute neural kernels for general activations?

Contributions

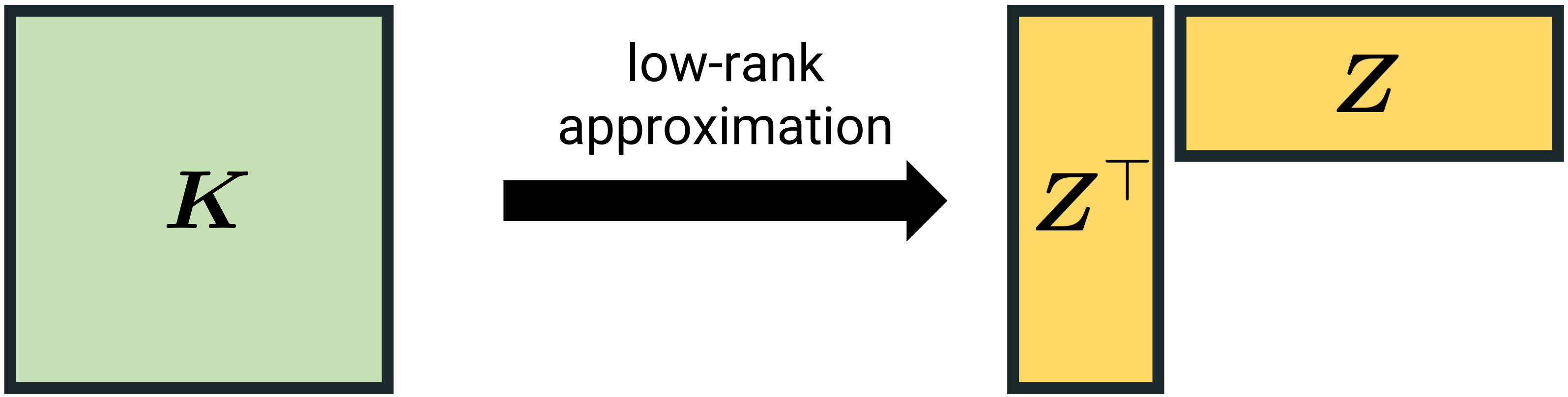
1. **Explicit expression of neural kernels for general activations**
   – We show the NTK expression without knowing the activation

Activation	$\sigma(t)$	Reference for the NNGP	Reference for the NTK
Rectified monomials	$t^q \cdot \mathbb{1}_{\{t \geq 0\}}$	[44]	[44]
Error function	$\operatorname{erf}(t)$	[43]	[5]
ABReLU (Leaky ReLU)	$-A\min(t,0) + B\max(t,0)$	[50, 51, 42]	[50, 51, 42]
Exponential	$\exp(At)$	[52, 46]	[52, 46]
Hermite polynomials	$h_q(t)$	[46]	This work
Sinusoidal	$\sin(At+B)$	[45, 47, 53]	This work
Gaussian	$\exp\left(-At^2\right)$	[43]	This work
GeLU	$\frac{t}{2}\left(1+\operatorname{erf}\left(\frac{t}{\sqrt{2}}\right)\right)$	[48]	This work
ELU	$\operatorname{step}(t)t + \operatorname{step}(-t)\left(e^t - 1\right)$	[48]	This work
Normalized Gaussian	Unknown	[54]	This work
RBF	$\sqrt{2}\sin(\sqrt{2A}t + \frac{\pi}{4})$	[45]	This work
Gabor	$\exp(-t^2)\sin(t)$	This work	This work
Monomial	t^q	This work	This work
Polynomial	$\sum_{j=0}^{q} a_j t^j$	This work	This work

Contributions

1. **Explicit expression of neural kernels for general activations**
   - We show the NTK expression without knowing the activation

2. **Fast neural kernel approximations by sketching algorithm**
   - Our algorithm runs in linear in the number of inputs/dimension
   - In practice, it runs up to **×106 faster** than the exact computation
   - For homogeneous activations, subspace embedding is guaranteed



Building Block for Neural Kernels

- **Dual kernel.** For every $x, y \in \mathbb{R}^d$ and a smooth $\sigma : \mathbb{R} \to \mathbb{R}$

$$K_\sigma(x, y) = \mathop{\mathbb{E}}_{w \sim \mathcal{N}(0,I)} [\sigma(\langle w, x \rangle)\sigma(\langle w, y \rangle)]$$

  – Neural kernels ⇒ composition of dual kernel with σ, σ'

Building Block for Neural Kernels

- **Dual kernel.** For every $x, y \in \mathbb{R}^d$ and a smooth $\sigma : \mathbb{R} \to \mathbb{R}$

$$K_\sigma(x, y) = \mathop{\mathbb{E}}_{w \sim \mathcal{N}(0,I)} \left[\sigma(\langle w, x\rangle)\sigma(\langle w, y\rangle)\right]$$

  - Neural kernels ⇒ composition of dual kernel with σ, σ'

- We derive a closed-form expression of K_σ when σ is a polynomial
  - Taylor series of σ can provide a power series of K_σ
  - Error bound of the truncated series is analyzed

Building Block for Neural Kernels

- **Dual kernel.** For every $x, y \in \mathbb{R}^d$ and a smooth $\sigma : \mathbb{R} \to \mathbb{R}$

$$K_\sigma(x, y) = \mathop{\mathbb{E}}_{w \sim \mathcal{N}(0, I)} \left[\sigma(\langle w, x \rangle) \sigma(\langle w, y \rangle) \right]$$

  - Neural kernels ⇒ composition of dual kernel with σ, σ'

- We derive a closed-form expression of K_σ when σ is a polynomial
  - Taylor series of σ can provide a power series of K_σ
  - Error bound of the truncated series is analyzed

- Computing the kernel matrix takes a quadratic time in the number of inputs
  - Huge memory space/infeasible computation time

Approximating Neural Kernels

- **Goal.** A fast and efficient kernel approximation

$$\{\mathrm{NTK}, \mathrm{NNGP}\}_\sigma(x, y) \approx \langle \phi(x), \phi(y) \rangle$$

Approximating Neural Kernels

- **Goal.** A fast and efficient kernel approximation

$$\{\mathrm{NTK}, \mathrm{NNGP}\}_\sigma(x,y) \approx \langle \phi(x), \phi(y) \rangle$$

- For homogeneous σ, i.e., $\sigma(at) = |a|\,\sigma(t)$ for all $a, t \in \mathbb{R}$
neural kernels = normalized dot-product kernels, e.g.,

$$\mathrm{NTK}_\sigma(x,y) = \|x\|\,\|y\|\,\kappa\left(\frac{\langle x, y\rangle}{\|x\|\,\|y\|}\right)$$

$\kappa : \mathbb{R} \to \mathbb{R}$ is an analytic function

- Feature map ϕ can be approximated by combining Taylor series on κ with randomized sketching algorithm

Conclusion

- We develop how to compute infinitely wide neural kernels
- We propose how to approximate these kernels using sketching

Activation	$\sigma(t)$	Reference for the NNGP	Reference for the NTK
Rectified monomials	$t^q \cdot \mathbb{1}_{\{t \geq 0\}}$	[44]	[44]
Error function	$\mathrm{erf}(t)$	[43]	[5]
ABReLU (Leaky ReLU)	$-A\min(t,0) + B\max(t,0)$	[50, 51, 42]	[50, 51, 42]
Exponential	$\exp(At)$	[52, 46]	[52, 46]
Hermite polynomials	$h_q(t)$	[46]	This work
Sinusoidal	$\sin(At+B)$	[45, 47, 53]	This work
Gaussian	$\exp\left(-At^2\right)$	[43]	This work
GeLU	$\frac{t}{2}\left(1+\mathrm{erf}\left(\frac{t}{\sqrt{2}}\right)\right)$	[48]	This work
ELU	$\mathrm{step}(t)t + \mathrm{step}(-t)\left(e^t - 1\right)$	[48]	This work
Normalized Gaussian	Unknown	[54]	This work
RBF	$\sqrt{2}\sin(\sqrt{2A}t + \frac{\pi}{4})$	[45]	This work
Gabor	$\exp(-t^2)\sin(t)$	This work	This work
Monomial	t^q	This work	This work
Polynomial	$\sum_{j=0}^{q} a_j t^j$	This work	This work