Sparse Kolmogorov-Arnold Networks

Preprint · August 2024

DOI: 10.13140/RG.2.2.10535.97448

1 author:

Gary Nan Tie
Mu Risk LLC

Sparse Kolmogorov-Arnold Networks

Gary Nan Tie, August 25, 2024

Abstract
Kolmogorov-Arnold Networks (KAN) are concatenated layers of univariate function matrices. We introduce two structured DCD matrix forms to sparsify layers, and learn the univariate functions (e.g. B-splines or wavelets).

Sparse Kolmogorov-Arnold Networks

KAN architecture

The shape of a Kolmogorov-Arnold Network (KAN) is represented by an integer array:

$[n_0, n_1, \ldots, n_L]$

where n_i is the number of nodes in the i^{th} layer of the computational graph. We denote the i^{th} neuron in the l^{th} layer by (l, i), and the activation value of (l, i) neuron by $x_{l,i}$. Between layer l and layer $l+1$, there are $n_l n_{l+1}$ activation functions. The activation function that connects (l, i) and $(l+1, j)$ is denoted by $\varphi_{l,j,i}$,

$l = 0, \ldots, L-1, \quad i = 1, \ldots, n_l, \quad j = 1, \ldots, n_{l+1}$

The pre-activation of $\varphi_{l,j,i}$ is $x_{l,i}$ and the post-activation of $\varphi_{l,j,i}$ is $\tilde{x}_{l,j,i} \triangleq \varphi_{l,j,i}(x_{l,i})$. The activation value of the $(l+1, j)$ neuron is the sum of all incoming post-activations:

$$x_{l+1,j} \triangleq \sum_{i=1}^{n_l} \tilde{x}_{l,j,i} = \sum_{i=1}^{n_l} \varphi_{l,j,i}(x_{l,i}), \quad j = 1, \ldots, n_{l+1}$$

In matrix form, this reads

$$x_{l+1} = \Phi_l \, x_l \quad \text{where} \quad \Phi_l = [\varphi_{l,j,i}]$$

is the matrix of univariate functions corresponding to the l^{th} KAN layer. A general KAN network is a composition of L layers: given an input vector $x \in \mathbb{R}^{n_0}$, the output of KAN is

$$\mathrm{KAN}(x) = (\Phi_{L-1} \circ \Phi_{L-2} \circ \ldots \circ \Phi_1 \circ \Phi_0)\, x$$

Sparse structured Φ_l matrices

Let filter $h = (h_1, \ldots, h_n)$ and circulant matrix

$$C(h) \triangleq \begin{bmatrix} h_1 & \cdot & \cdot & h_{n-1} & h_n \\ h_n & h_1 & & & h_{n-1} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot \\ h_2 & \cdot & \cdot & h_n & h_1 \end{bmatrix}$$

Let gate $g = (g_1, \ldots, g_n)$ and diagonal matrix

$$D(g) \triangleq \begin{bmatrix} g_1 & & & 0 \\ & g_2 & & \\ 0 & & \ddots & \\ & & & g_n \end{bmatrix}$$

Two observations:

(1) If matrix A has rank r, then A can be written as a sum of r rank-1 matrices. Now any rank-1 matrix is a DCD, so any matrix A of rank r is the sum of r DCD matrices.

(2) Observe that any Toeplitz matrix, often arising in signal processing, is the sum of a pair of DCDs, because any Toeplitz matrix can be decomposed into the sum of a circulant matrix C and a skew circulant matrix S; now note both C and S are DCD matrices. Moreover, any matrix A can be written as a product of Toeplitz matrices. Hence any matrix A can be written as a concatenation of (DCD+DCD) operators.

For these reasons, we are interested in DCD operators to sparsify Φ_l. We now propose two sparse structured matrix forms for the univariate function matrices Φ_l when $n_i = n \ \forall i$.

Unstructured Φ_l has n^2 functions to learn.

In the following, the filters h, r, s and gates f, g, p, q are B-splines or wavelets to be learned.

(1) TriAd $\Phi_L \triangleq \sum_{i=1}^{N} D(f^i)\, C(h^i)\, D(g^i)$

has $N 3n$ functions to learn

(2) Centipede $\Phi_L \triangleq T_N \ldots T_2 T_1$

where $T_i = [D(p^i)\, C(r^i)\, D(q^i) + D(f^i)\, C(s^i)\, D(g^i)]$

has $N 6n$ functions to learn.

So if hyperparameter N is chosen so that $N \ll n$ then $N 6n < n^2$ and Φ_L is relatively sparse.

$\square$

References

KAN:Kolmogorov–Arnold Networks
Liu, Z. et al, arXiv:2404.19756v4 [cs.LG] 16 Jun 2024

TriAd Hierarchy: Actually, signal processing is all you need
Nan Tie, G., ResearchGate Preprint · Aug 2023
DOI: 10.13140/RG.2.2.30634.59844/1