UMBC

A Walsh Hadamard Derived Linear Vector Symbolic Architecture

Mohammad Mahmudul Alam, Alexander Oberle, Edward Raff,
Stella Biderman, Tim Oates, James Holt

LPS | LABORATORY FOR PHYSICAL SCIENCES

Booz | Allen | Hamilton

Objectives

- Many VSAs have shown issues in numerical stability, computational complexity, or suboptimal performance when applied in neural network
- Introduced a novel VSA derived from Walsh Hadamard transformation, named Hadamard-derived Linear Binding (HLB)
- HLB supports commutative and associative binding in linear time and provides a numerically stable exact inverse while unbinding
- Defined a bimodal distribution which is proposed as initialization condition that avoids numerical instability
- Moreover, mathematically derived a correction term that improves the response of HLB during unbinding

Binding Definition

The Binding function is defined by replacing the Fourier transform in circular convolution with the Hadamard transform

$$H_1 = \begin{bmatrix} 1 \end{bmatrix} \qquad H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad \cdots \qquad H_{2^n} = \begin{bmatrix} H_{2^{n-1}} & H_{2^{n-1}} \\ H_{2^{n-1}} & -H_{2^{n-1}} \end{bmatrix}$$

$$\mathcal{B}(x, y) = \frac{1}{d} \cdot H(Hx \odot Hy)$$

A parameter ρ is defined that denotes the number of vector pairs bundled in a composite representation χ

$$\underset{\rho=1}{\chi} = \mathcal{B}(x_1, y_1) \qquad \underset{\rho=2}{\chi} = \mathcal{B}(x_1, y_1) + \mathcal{B}(x_2, y_2) \qquad \cdots \qquad \chi_\rho = \sum_{i=1}^{\rho} \mathcal{B}(x_i, y_i)$$

Unbinding

The unbinding is defined via an inverse function following the theorem. This will give a symbolic form of unbinding step that retrieves the original component x being searched for, as well as a necessary noise component η° which must exist whenever $\rho > 1$ items are bound together

Proof of Theorem 8.2.1. We start from the identity function $Hx \cdot Hx^{\dagger} = \mathbb{1}$ and thus $Hx^{\dagger} = \frac{\mathbb{1}}{Hx}$. Now using Equation 8.2 we get,

$$\mathcal{B}^{*}(\mathcal{B}(x_1, y_1) + \cdots + \mathcal{B}(x_\rho, y_\rho), y_i^{\dagger}) = \frac{1}{d} \cdot H((Hx_1 \odot Hy_1 + \cdots + Hx_\rho \odot Hy_\rho) \odot \frac{1}{Hy_i})$$

$$= \frac{1}{d} \cdot H(Hx_i + \frac{1}{Hy_i} \odot \sum_{\substack{j=1 \\ j \neq i}}^{\rho} (Hx_j \odot Hy_j)) = x_i + \frac{1}{d} \cdot H(\frac{1}{Hy_i} \odot \sum_{\substack{j=1 \\ j \neq i}}^{\rho} (Hx_j \odot Hy_j)) \quad \text{Lemma 3.1}$$

$$= \begin{cases} x_i & \text{if} \quad \rho = 1 \\ x_i + \eta_i^{\circ} & \text{else } \rho > 1 \end{cases}$$

Projection

- To reduce the noise component, Holographic Reduced Representations (HRR) utilizes a projection step that normalizes inputs in Fourier domain
- While such normalization is not helpful in Hadamard domain, we just apply the Hadamard transformation to the inputs as a projection step

Definition 8.2.2 (Projection). The projection function of x is defined by $\pi(x) = \frac{1}{d} \cdot Hx$

If we apply the Definition 8.2.2 to the inputs in Theorem 8.2.1 then we get

$$\begin{aligned}\mathcal{B}^*(\mathcal{B}(\pi(x_1), \pi(y_1)) + \cdots + \mathcal{B}(\pi(x_\rho), \pi(y_\rho)), \pi(y_i)^\dagger) &= \mathcal{B}^*(\frac{1}{d} \cdot H(x_1 \odot y_1 + \cdots x_\rho \odot y_\rho), \frac{1}{y_i}) \\ &= \frac{1}{d} \cdot H(\frac{1}{y_i} \odot (x_1 \odot y_1 + \cdots x_\rho \odot y_\rho))\end{aligned}$$

Simplification

- By applying the projection step, we get a different noise component η^{π}
- More interestingly, the retrieved output term does not contain any Hadamard matrix
- Therefore, the initial binding definition is redefined as $\mathcal{B}'(x, y) = x \odot y$ and unbinding as $\mathcal{B}^{*\prime}(x, y) = x \odot \frac{1}{y}$

$$H(\frac{1}{d} \cdot H(\frac{1}{y_i} \odot (x_1 \odot y_1 + \cdots x_\rho \odot y_\rho))) = \frac{1}{y_i} \odot (x_1 \odot y_1 + \cdots + x_\rho \odot y_\rho)$$

$$= \begin{cases} x_i & \text{if} \quad \rho = 1 \\ x_i + \sum_{j=1,\ j \neq i}^{\rho} \frac{x_j y_j}{y_i} & \text{else } \rho > 1 \end{cases} = \begin{cases} x_i & \text{if} \quad \rho = 1 \\ x_i + \eta_i^{\pi} & \text{else } \rho > 1 \end{cases}$$

Initialization Condition

- For the binding and unbinding operations to work, vectors need to have an expected value of zero
- Since during unbinding, values close to zero would destabilize the noise component and create numerical instability
- Thus, a Mixture of Normal Distribution (MiND) is defined with an expected value of zero
- But the expected absolute value is greater than zero where U is the Uniform distribution

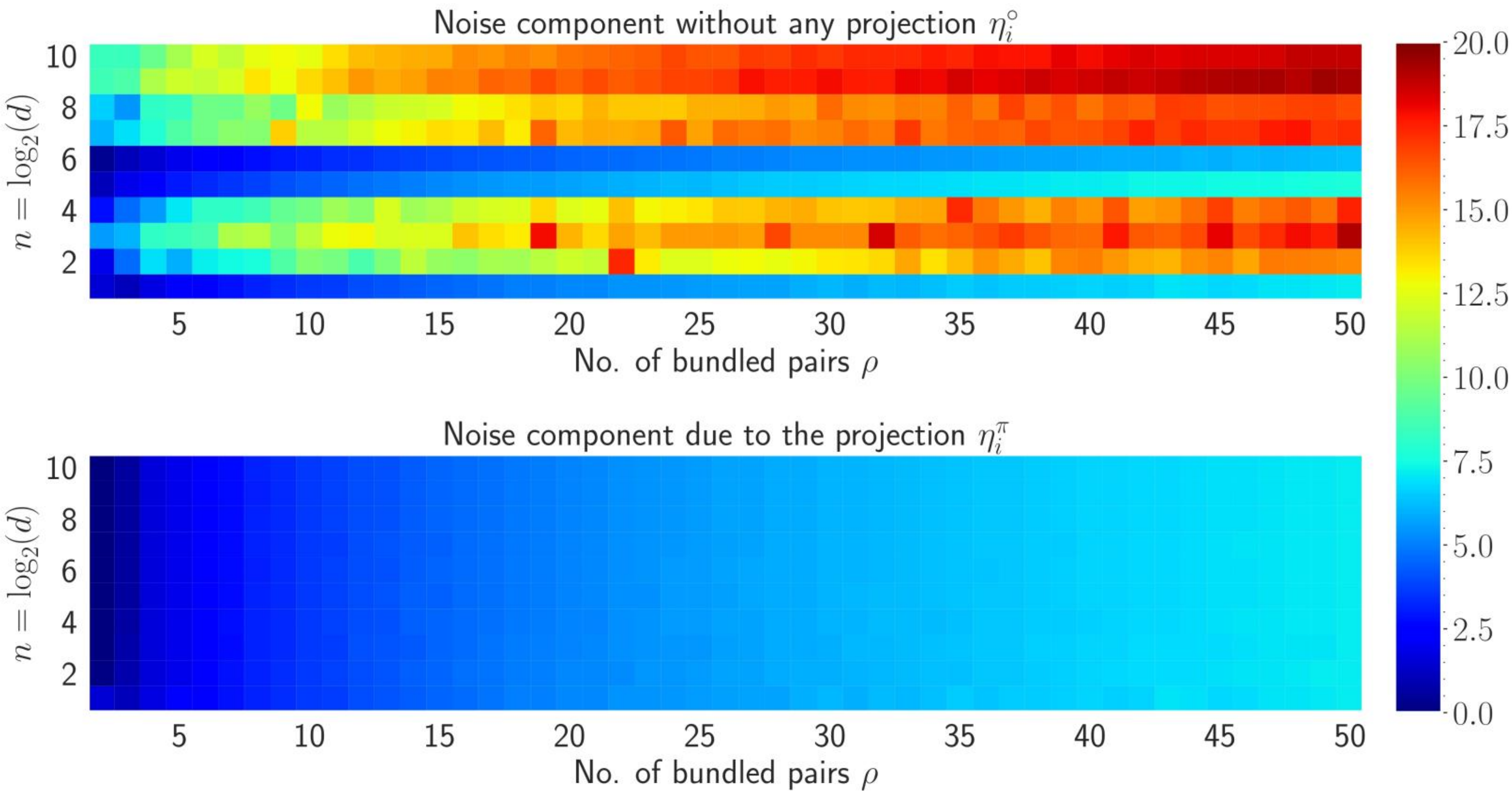
$$\Omega(\mu, 1/d) = \begin{cases} \mathcal{N}(-\mu, 1/d) & \text{if} \quad \mathcal{U}(0,1) > 0.5 \\ \mathcal{N}(\ \mu, 1/d) & \text{else}\ \mathcal{U}(0,1) \leqslant 0.5 \end{cases}$$

Properties

$$E[x] = 0, E[|x|] = \mu, and\ E[||x||_2] = \sqrt{\mu^2 d}$$

Noise Reduction

In expectation, it is proved that $\eta^{\pi} < \eta^{\circ}$ which is also verified by the experimental results.



Similarity Augmentation

Theorem 8.2.3 ($\phi - \rho$ Relationship). *Given $x_i, y_i \sim \Omega(\mu, 1/d)\ \forall\ i \in \mathbb{N} : 1 \leqslant i \leqslant \rho$, the cosine similarity ϕ between the original x_i and retrieved vector $\hat{x}_i$ is approximately equal to the inverse square root of the number of vector pairs in a composite representation ρ given by* $\phi \approx \frac{1}{\sqrt{\rho}}$

$$\phi = \frac{\sum^{d} x_i \cdot \hat{x}_i}{\|x_i\|_2 \cdot \|\hat{x}_i\|_2} = \frac{\sum^{d} x_i \cdot \left(x_i + \sum_{j=1,\ j\neq i}^{\rho} \frac{x_j y_j}{y_i} \right)}{\|x_i\|_2 \cdot \|x_i + \sum_{j=1,\ j\neq i}^{\rho} \frac{x_j y_j}{y_i}\|_2} = \frac{\sum^{d} x_i \cdot x_i + \sum^{d} x_i \cdot \left(\sum_{j=1,\ j\neq i}^{\rho} \frac{x_j y_j}{y_i} \right)}{\|x_i\|_2 \cdot \|x_i + \sum_{j=1,\ j\neq i}^{\rho} \frac{x_j y_j}{y_i}\|_2} \tag{7}$$

Employing Properties 3.1 we can derive that $\|x_i\|_2 = \sqrt{\sum x_i \cdot x_i} = \sqrt{\mu^2 d}$ and $\|\frac{x_j y_j}{x_i}\| = \sqrt{\mu^2 d}$. Thus, the square of the $\|x_i + \sum_{j=1,\ j\neq i}^{\rho} \frac{x_j y_j}{y_i}\|_2$ can be expressed as

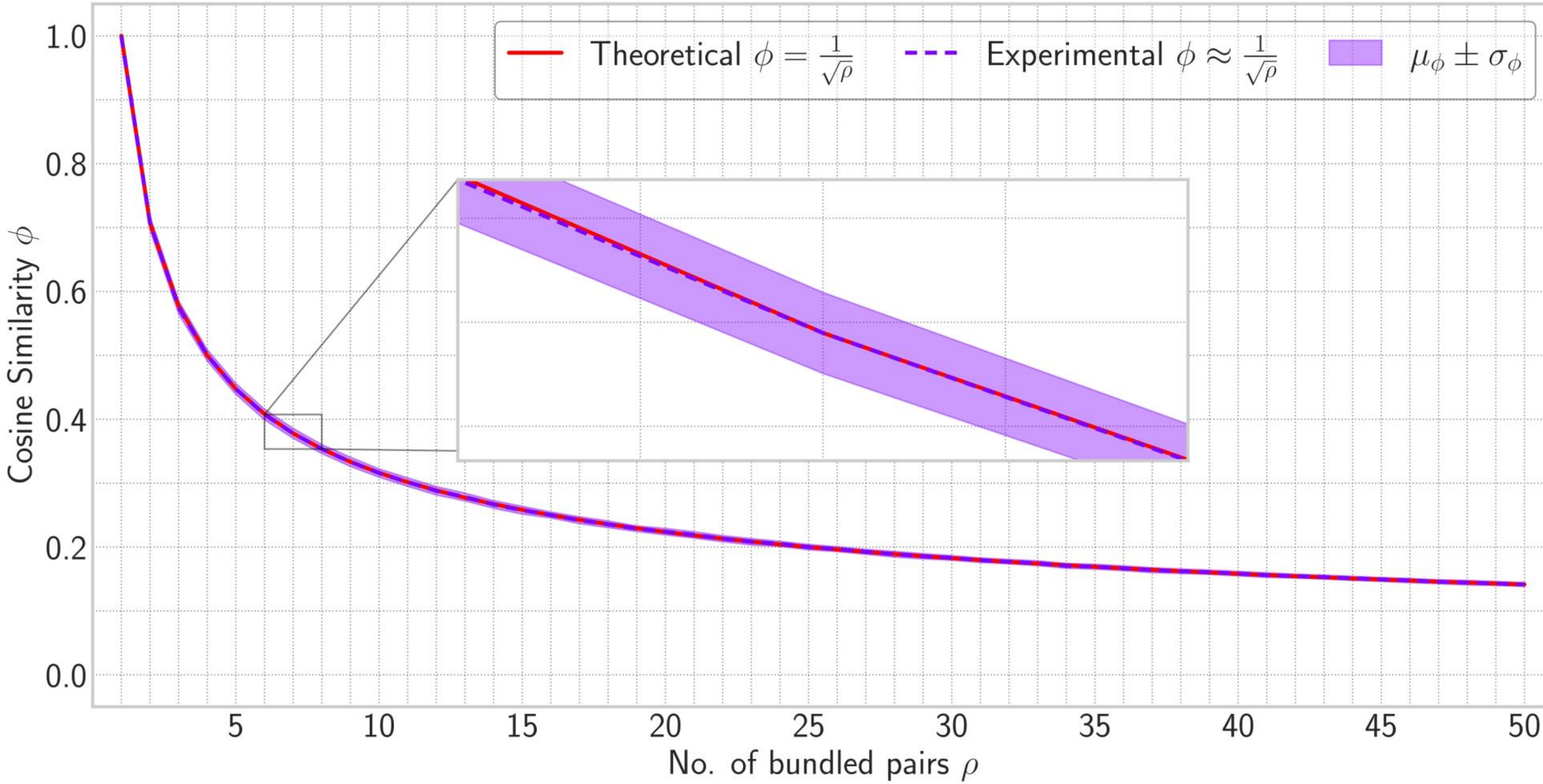
$$= \|x_i\|_2^2 + \sum_{j=1,\ j\neq i}^{\rho} \left\| \frac{x_j y_j}{y_i} \right\|_2^2 + 2 \cdot \underbrace{\sum^{d} x_i \left(\sum_{j=1,\ j\neq i}^{\rho} \frac{x_j y_j}{y_i} \right)}_{\alpha} + \underbrace{\sum^{d} \sum_{\substack{j=1 \\ j\neq i}}^{\rho-1} \sum_{\substack{l=1 \\ l\neq j}}^{\rho-1} \frac{x_j y_j}{y_i} \cdot \frac{x_l y_l}{y_i}}_{\beta} \tag{8}$$

$$= \mu^2 d + (\rho - 1) \cdot \mu^2 d + 2\alpha + 2\beta \quad = \rho \cdot \mu^2 d + 2\alpha + 2\beta$$

Therefore, using Equation 7 and Equation 8 we can write that

$$\mathrm{E}[\phi] = \frac{\mu^2 d + \alpha}{\sqrt{\mu^2 d} \cdot \sqrt{\rho \cdot \mu^2 d + 2\alpha + 2\beta}} \approx {}^{1} \frac{\mu^2 d}{\sqrt{\mu^2 d} \cdot \sqrt{\rho \cdot \mu^2 d}} = \frac{\mu^2 d}{\sqrt{\rho} \cdot \mu^2 d} = \frac{1}{\sqrt{\rho}} \qquad \square$$

Similarity Augmentation Contd.

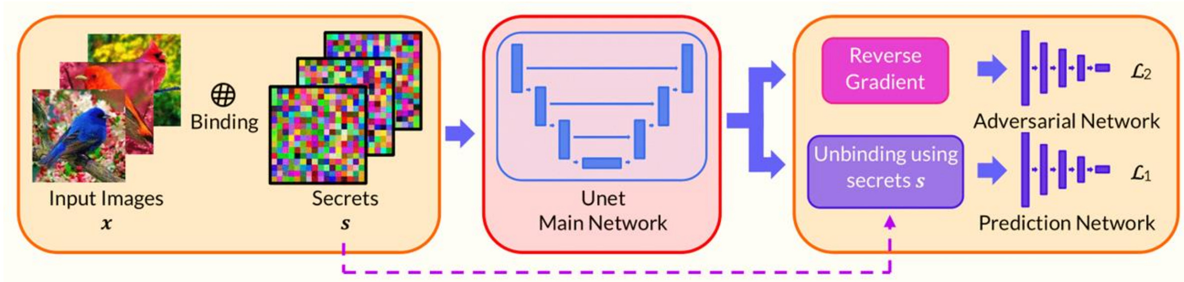


HLB in Deep Learning

- Experiments are performed in two deep learning applications using HLB
- In the first application a method called Connectionist Symbolic Pseudo Secrets (CSPS) that leverages binding and unbinding to obfuscate the nature of input and outputs of network using HRR
- The same experiments are performed via the properties of HLB
- Similarly, in the second application, different VSA's are used and compared to perform Extreme Multi Label (XML) classification

METHOD	BIND $\mathcal{B}(x, y)$	UNBIND $\mathcal{B}^*(x, y)$	INIT x
HRR	$\mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{x}) \odot \mathcal{F}(\boldsymbol{y}))$	$\mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{x}) \oplus \mathcal{F}(\boldsymbol{y}))$	$x_i \sim \mathcal{N}(0, 1/d)$
VTB	$V_y x$	$V_y^\top x$	$\tilde{x}_i \sim \mathcal{N}(0, 1) \rightarrow x = \tilde{\boldsymbol{x}}/\|\tilde{\boldsymbol{x}}\|_2$
MAP-C	$x \odot y$	$x \odot y$	$x_i \sim \mathcal{U}(-1, 1)$
MAP-B	$x \odot y$	$x \odot y$	$x_i \sim \{-1, 1\}$
HLB	$x \odot y$	$x \oplus y$	$x_u \sim \{\mathcal{N}(-\mu, 1/d), \mathcal{N}(\mu, 1/d)\}$

Experimental Results: CSPS



Mohammad Mahmudul Alam, et. al. "Deploying Convolutional Networks on Untrusted Platforms Using 2D Holographic Reduced Representations", ICML 2022, Baltimore, MD, USA

Connectionist Symbolic Pseudo Secrets (CSPS)
Accuracy

Dataset	Dims/ Labels	CSPS + HRR		CSPS + VTB		CSPS + MAP-C		CSPS + MAP-B		CSPS + HLB	
		Top@1	Top@5	Top@1	Top@5	Top@1	Top@5	Top@1	Top@5	Top@1	Top@5
MNIST	$28^2/10$	98.51	–	98.44	–	98.46	–	98.40	–	**98.73**	–
SVHN	$32^2/10$	88.44	–	19.59	–	79.95	–	92.43	–	**94.53**	–
CR10	$32^2/10$	78.21	–	74.22	–	76.69	–	82.83	–	**83.81**	–
CR100	$32^2/100$	48.84	75.82	35.87	61.79	56.77	81.52	57.76	84.63	**58.82**	**87.50**
MIN	$84^2/100$	40.99	66.99	45.81	73.52	52.22	78.63	57.91	82.81	**59.48**	**83.35**
GM		67.14	71.26	47.24	67.40	70.89	80.06	75.90	83.72	**77.17**	**85.40**

Experimental Results: XML

Extreme Multi-label Classification

- Each class is represented with a unique vector, network learns the composite representation
- Each class label is queried, and cosine similarity is used to identify which labels are present
- Evaluated in terms of normalized discounted cumulative gain (nDCG) and propensity-scored (PS) based normalized discounted cumulative gain (PSnDCG)

Dataset	Bibtex		Delicious		Mediamill		EURLex-4K	
Metrics	nDCG	PSnDCG	nDCG	PSnDCG	nDCG	PSnDCG	nDCG	PSnDCG
HRR	60.296	45.572	66.454	30.016	83.885	63.684	77.225	30.684
VTB	57.693	45.219	63.325	31.449	87.232	66.948	76.964	31.180
MAP-C	59.280	46.092	65.376	31.943	87.255	66.886	72.439	26.752
MAP-B	59.412	46.340	65.431	32.122	86.886	66.562	71.128	26.340
HLB	**61.741**	**48.639**	**67.821**	**32.797**	**88.064**	**67.525**	**77.868**	**31.526**

Dataset	EURLex-4.3K		Wiki10-31K		Amazon-13K		Delicious-200K	
Metrics	nDCG	PSnDCG	nDCG	PSnDCG	nDCG	PSnDCG	nDCG	PSnDCG
HRR	84.497	38.545	81.068	9.185	93.258	49.642	44.933	6.839
VTB	84.663	38.540	78.025	9.645	92.373	49.463	44.092	6.664
MAP-C	85.472	39.233	80.203	10.027	92.013	48.686	45.373	6.862
MAP-B	85.023	38.820	80.238	10.035	92.307	48.812	45.459	6.870
HLB	**88.204**	**43.622**	**83.589**	**11.869**	**93.672**	**50.270**	**46.331**	**6.952**

Concluding Remarks

- Proposed a novel VSA called HLB that offers significant benefits for both classical VSA tasks and differentiable systems.
- Additionally, proposed a new initialization condition called Mixture of Normal Distribution (MiND)
- Mathematically showed the cosine similarity ϕ is approximately equal to the inverse square root of the number of bundled vector pairs ρ
- HLB outperformed other VSAs in both CSPS and XML tasks across all datasets, highlighting its extensive potential in Neuro-symbolic AI

Thank you for your patience

Any Questions?

NEURAL INFORMATION PROCESSING SYSTEMS

2024