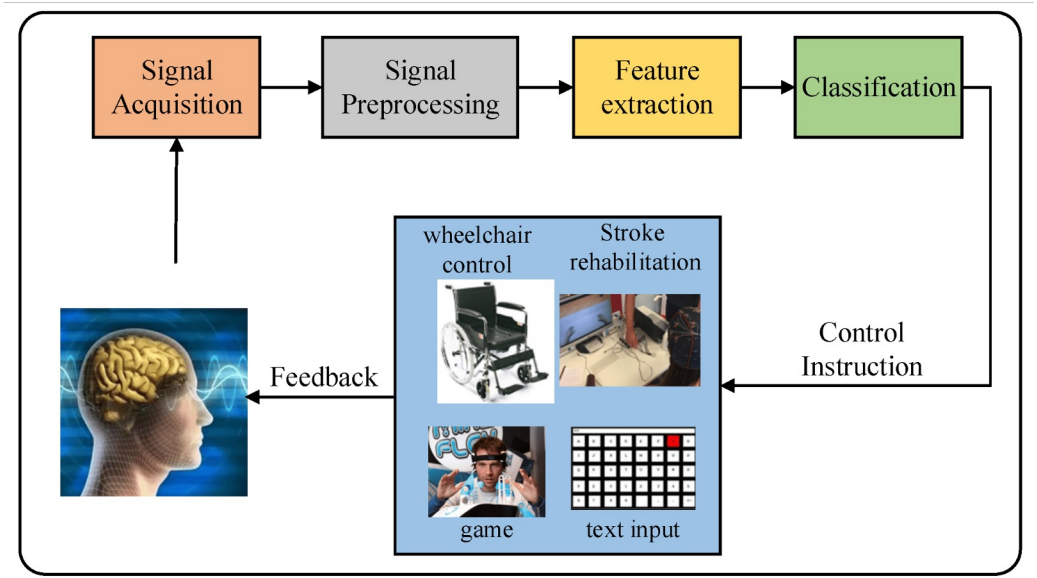
Scheduled Knowledge Acquisition on Lightweight Vector Symbolic Architectures for Brain-Computer Interfaces

Yejia Liu, Shijin Duan, Xiaolin Xu, Shaolei Ren.

TinyML 2024

Presenter: **Yejia Liu**

Challenges in brain-computer interfaces (BCIs) applications



- Complex brain signals
- Low latency requirement in real-time EEG-based BCIs

* Image credit to “Chang Z, Zhang C, Li C. Motor Imagery EEG Classification Based on Transfer Learning and Multi-Scale Convolution Network. *Micromachines*. 2022; 13(6):927. https://doi.org/10.3390/mi13060927.”

Motivation

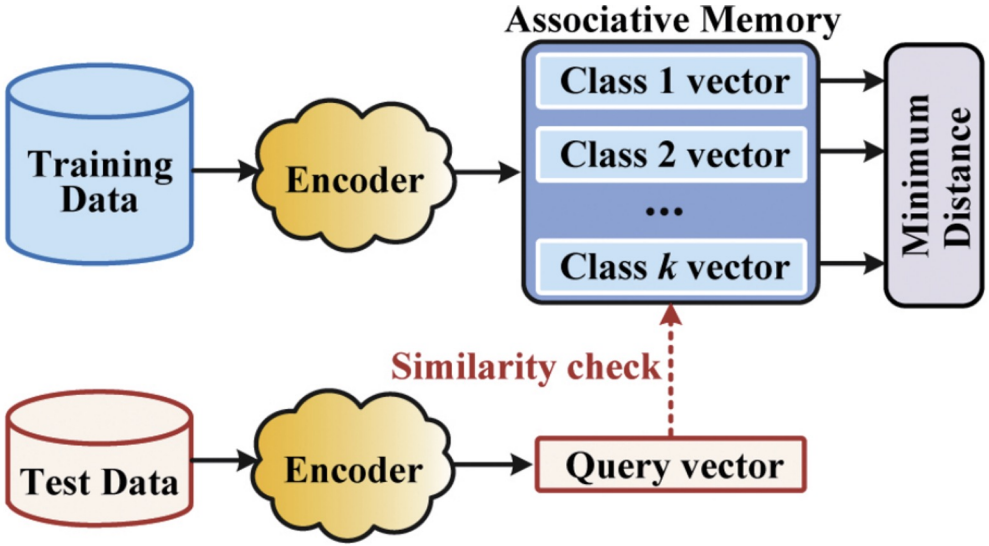
- Neural networks achieve high accuracy but are expensive in computational cost and have high latency, unaffordable for some implantable BCI devices with stringent power constraints
- Classic feature engineering methods (e.g. SVM) exhibits unsatisfactory accuracy
- **We need a method that has low latency but also offers satisfactory accuracy**

Background: Vector symbolic architecture (VSA)

- Neural networks are not able to decompose joint representations to obtain distinct objects
- Symbolic AI suffers from exhaustive rule searches
- VSA serves as a common language between neural networks and symbolic AI

Background: Hyper-dimensional computing (HDC/VSA)

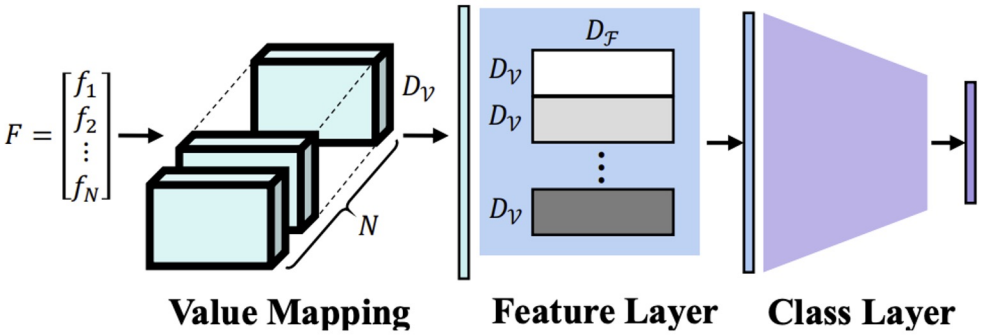
- (Bipolar) High-dimensional vectors
  - From a thousand to tens of thousands dimensionality
- Hardware-efficient operations
  - Element-wise additions and dot products



* Image credit to "Nunes, Igor et al. "GraphHD: Efficient graph classification using hyperdimensional computing." *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (2022): 1485-1490."

A possible feasible architecture for real-time lightweight BCIs: Low-dimensional computing (LDC) classifier

- Problems in HDC
  - Large model size
  - Low accuracy
- Low-dimensional classifier (LDC)
  - Systematic training procedure for higher accuracy
  - Low dimensional encodings for lightweight model



* Image credit to “Duan, Shijin & Xu, Xiaolin & Ren, Shaolei. (2022). A Brain-Inspired Low-Dimensional Computing Classifier for Inference on Tiny Devices.”

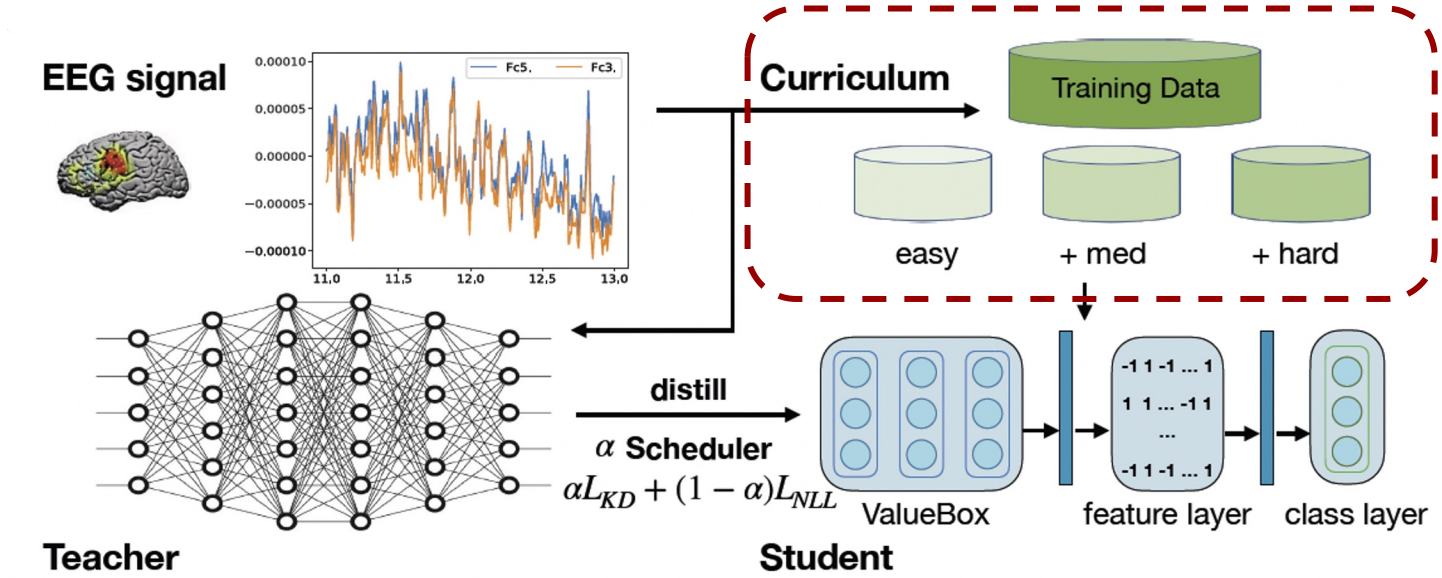
Applying LDC to BCIs

- Problem - low accuracy compared to DNNs
  - Accuracy of LDC classifiers still lags behind the neural networks (e.g. 3-layer MLP w/ hidden_size = 50)
- A solution - Knowledge distillation

$$\mathcal{L} := \alpha \mathcal{L}_{KD}(\mathbf{z}^s, \mathbf{z}^t) + (1 - \alpha)\mathcal{L}_{NLL}(\mathbf{z}^s, y),$$

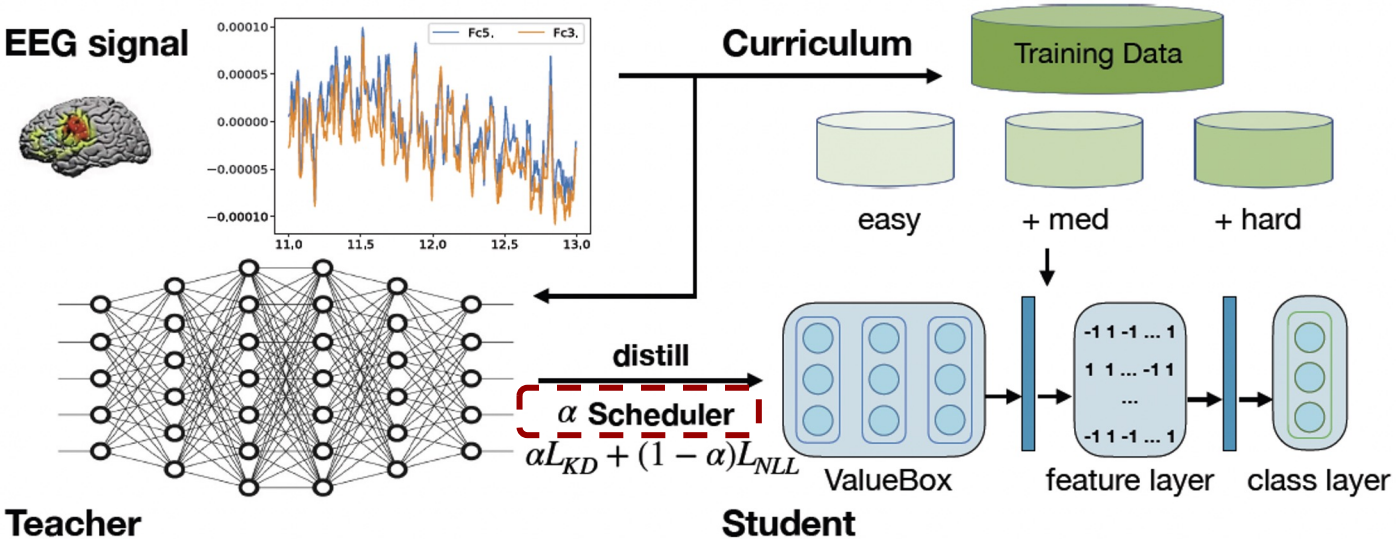
  - However, large capacity gap between “big” teachers and “small” students can result in ineffective knowledge
    - Intricate patterns and fine-grained data details captured by teachers hard to be comprehended by much smaller student architecture
    - Constant distillation level: Lack of adaption throughout the distillation process

ScheduledKD-LDC: Scheduling knowledge distillation on LDC classifiers



- Data level - Curriculum data ordering
  - Allow the student to build data representations step-by-step, from easy to hard

ScheduledKD-LDC: Scheduling knowledge distillation on LDC classifiers



- Learning procedure - α scheduler to manage distillation level
  - Begin with higher α for amplifying teacher's influence
  - Gradually decrease to foster student's independence

ScheduledKD-LDC: Our algorithm

Algorithm 1 Scheduled Knowledge Distillation

Input: Training data $\{x_i, y_i\}_{i=1}^{I}$; Total training epoch H; Pretrained teacher model f^t with θ_t; Student model f^s with randomly initialized θ_s; Initial balancing weight α; Difficulty ranking function t; Decay step k; Decay rate γ; Change point P; Order $o \in$ {"curriculum", "random", "anti-curriculum"}.

Output: Trained student model f^s.

Rank data: $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_I) \leftarrow \text{sort}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_I, s, o)$

while $h < H$ **do**

  if $h \geq P$ and $h \,\%\, k = 0$ **then**

    Exponentially decrease α: $\alpha \leftarrow \alpha \times \gamma^{\lceil \frac{h}{r} \rceil}$

  end if

  Update θ_s based on the curriculum: $\theta_s^h \leftarrow$ train-one-epoch$(\theta_s^{h-1}, \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_I\})$

end while

Evaluation Setup

- Evaluation metrics
  - Inference computation efficiency
    - Floating-point multiply-accumulate operations (FPMACs)
    - Binary multiply-accumulate operations (BMACs)
    - Model size
  - Accuracy
- Datasets
  - Motor Imagery - Classifying five movements from EEG signals
  - X11 and S4b - Classifying hand movements from EEG signals
  - ERN - A P300 speller task

Key results

- ScheduledKD-LDC has achieved a good balance between accuracy and efficiency

Dataset	Method	Accuracy (%)	BMACs ($\times 10^6$)	FPMACs ($\times 10^6$)	Model Size (KB)
Motor Imagery	EEGNet [34]	$85.33_{\pm 1.25}$	0	4.81	72.22
	DeepConvNet [34]	$92.83_{\pm 1.21}$	0	25.33	613.82
	Binarized-DeepConvNet	$57.68_{\pm 2.21}$	24.56	0.76	47.15
	SVM (HALO) [29]	$58.42_{\pm 0.45}$	0	1.02×10^{-3}	4.60
	MLP	$86.12_{\pm 1.44}$	0	32.90	125.88
	LeHDC	$76.02_{\pm 0.44}$	4.06	0	527.81
	LDC	$77.18_{\pm 0.89}$			
	KD-LDC	$77.89_{\pm 0.73}$			
	CTKD [35] w/ LDC	$78.85_{\pm 0.62}$	**0.13**	**0**	**16.89**
	ScheduledKD-LDC	**80.17**$_{\pm 0.83}$			
X11 and S4b	EEGNet [34]	$80.04_{\pm 1.09}$	0	5.98	105.38
	Binarized-EEGNet	$56.14_{\pm 1.83}$	5.76	0.21	12.77
	DeepConvNet	$83.71_{\pm 2.07}$	0	28.99	892.01
	SVM (HALO) [29]	$53.55_{\pm 0.43}$	0	1.50×10^{-3}	6.01
	LeHDC	$68.64_{\pm 0.22}$	5.93	0	754.68
	LDC	$69.16_{\pm 1.04}$			
	KD-LDC	$69.68_{\pm 0.83}$			
	CTKD [35] w/ LDC	$70.22_{\pm 0.64}$	**0.19**	**0**	**24.15**
	ScheduledKD-LDC	**71.83**$_{\pm 0.77}$			
ERN	EEGNet [34]	$82.84_{\pm 1.04}$	0	4.77	69.78
	Binarized-EEGNet	$59.63_{\pm 1.74}$	4.59	0.18	5.31
	DeepConvNet	$86.67_{\pm 1.34}$	0	27.68	632.56
	SVM (HALO) [29]	$55.80_{\pm 0.33}$	0	1.12×10^{-3}	4.38
	LeHDC	$72.63_{\pm 0.45}$	4.59	0	597.81
	LDC	$73.34_{\pm 0.87}$			
	KD-LDC	$73.86_{\pm 0.73}$			
	CTKD [35] w/ LDC	$74.42_{\pm 0.60}$	**0.15**	**0**	**19.13**
	ScheduledKD-LDC	**75.57**$_{\pm 0.62}$			

Key results

- ScheduledKD-LDC has achieved a good balance between accuracy and efficiency

Dataset	Method	Accuracy (%)	BMACs ($\times 10^6$)	FPMACs ($\times 10^6$)	Model Size (KB)
Motor Imagery	EEGNet [34]	85.33±1.25	0	4.81	72.22
	DeepConvNet [34]	92.83±1.21	0	25.33	613.82
	Binarized-DeepConvNet	57.68±2.21	24.56	0.76	47.15
	SVM (HALO) [29]	58.42±0.45	0	1.02×10^{-3}	4.60
	MLP	86.12±1.44	0	32.90	125.88
	LeHDC	76.02±0.44	4.06	0	527.81
	LDC	77.18±0.89			
	KD-LDC	77.89±0.73			
	CTKD [35] w/ LDC	78.85±0.62	**0.13**	**0**	**16.89**
	ScheduledKD-LDC	**80.17**±0.83			
X11 and S4b	EEGNet [34]	80.04±1.09	0	5.98	105.38
	Binarized-EEGNet	56.14±1.83	5.76	0.21	12.77
	DeepConvNet	83.71±2.07	0	28.99	892.01
	SVM (HALO) [29]	53.55±0.43	0	1.50×10^{-3}	6.01
	LeHDC	68.64±0.22	5.93	0	754.68
	LDC	69.16±1.04			
	KD-LDC	69.68±0.83			
	CTKD [35] w/ LDC	70.22±0.64	**0.19**	**0**	**24.15**
	ScheduledKD-LDC	**71.83**±0.77			
ERN	EEGNet [34]	82.84±1.04	0	4.77	69.78
	Binarized-EEGNet	59.63±1.74	4.59	0.18	5.31
	DeepConvNet	86.67±1.34	0	27.68	632.56
	SVM (HALO) [29]	55.80±0.33	0	1.12×10^{-3}	4.38
	LeHDC	72.63±0.45	4.59	0	597.81
	LDC	73.34±0.87			
	KD-LDC	73.86±0.73			
	CTKD [35] w/ LDC	74.42±0.60	**0.15**	**0**	**19.13**
	ScheduledKD-LDC	**75.57**±0.62			

Ablation studies: Analysis of the α scheduler

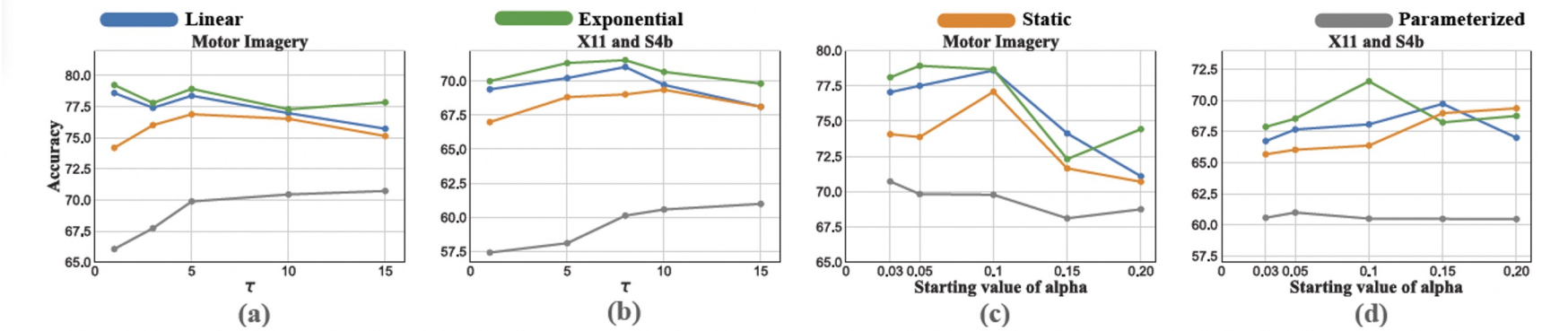


Figure 3: Comparison of different α setups, *without* employing data curriculum in the KD setting on the Motor Imagery, X11 and S4b datasets. (a), (b): With different temperatures τ. (c), (d): With different starting values of α.

- Using an α scheduler is more effective than a static α
- Exponential α scheduler performs slightly better than the linear one
- Parameterized α has the lowest accuracy

Ablation studies: Efficacy of curriculum data order

	Static α	Linear α	**Expo α**	param α
Curri KD-LDC	**78.04**$\pm$0.60	**79.57**$\pm$0.86	**80.17**$\pm$0.83	**72.83**$\pm$0.56
KD-LDC	77.89$\pm$0.73	78.59$\pm$0.64	78.92$\pm$0.57	70.72$\pm$0.47
Anti-curri KD-LDC	73.93$\pm$0.81	73.84$\pm$0.74	74.65$\pm$0.70	67.72$\pm$0.83
Curri LDC		**77.20**$\pm$0.57		
LDC		77.18$\pm$0.89		
Anti-curri LDC		73.43$\pm$1.20		

- Curriculum helps in the knowledge distillation setting
- Anti-curriculum (from hard-to-easy) adversely affects accuracy
- Curriculum does not help if not under the knowledge distillation setting

Ablation studies: Efficacy of curriculum data order

Table 3: Data ordering analysis on the Motor Imagery dataset. (a) Accuracy comparison when data ordered by loss of teacher model vs. student model; (b) Loss-based ranking intersection between teacher and student model.

	Order by Teacher Loss	**Order by Student Loss**
ScheduledKD LDC	74.14±0.66	**80.17±0.83**
Curri LDC	70.67±0.58	**77.20±0.57**
Curri LDC w/ KD	74.34±0.61	**78.04±0.60**

(a)

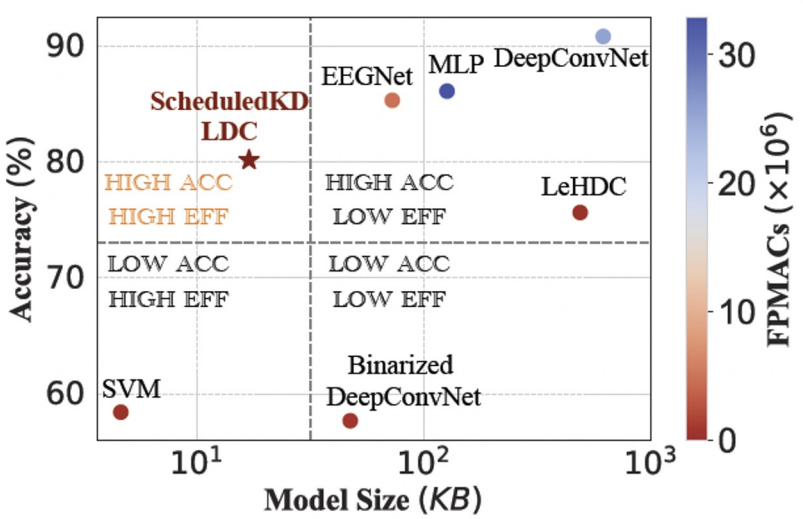
	Overlapped Rank (%)
Hardest 30%	30.06
Hardest 50%	50.25
Hardest 70%	70.72

(b)

- Using teacher model's loss to order data results in worse accuracy than using the student model's loss
- Student and teacher model have different perception on the hardness of data points

Summary & Takeaways

- ScheduledKD-LDC strikes a good balance between inference accuracy and efficiency for BCI applications
- α scheduler manages distillation level to allow the students to comprehend knowledge from the teacher
- The curriculum data order helps small student models to build their own data representation gradually




THANK
YOU

Copyright Notice

www.tinyml.org