Disentangling Multi-view Representations via Curriculum Learning with Learnable Prior

Kai Guo[1], Jiedong Wang[1], Xi Peng[1,2], Peng Hu[1], Hao Wang[1*]

[1]Sichuan University, [2]National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation

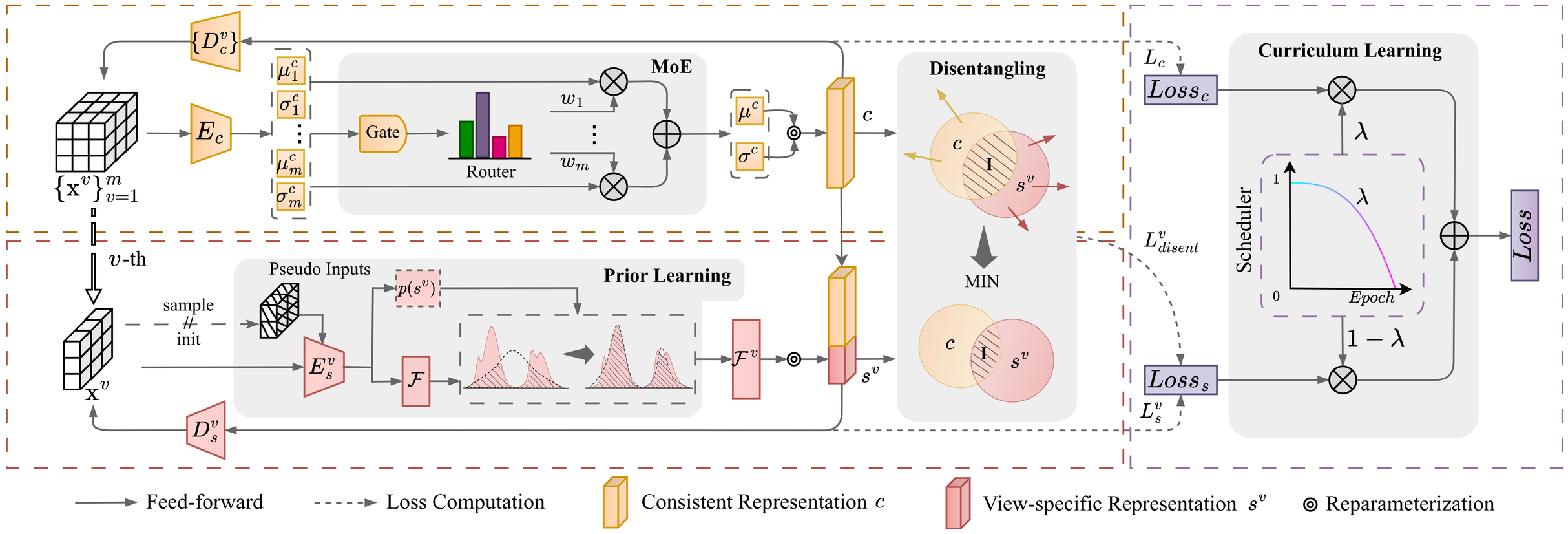


Figure 1: An overview of the proposed CL2P, which mainly consists of 1) Curriculum learning, 2) Mixture-of-Experts (MoE), 3) Prior learning, and 4) Disentangling module.

Problem Statement

(**Multi-view Representation Learning**) Given a set of multi-view data $\mathcal{D} = \{\mathbf{x}_i|\mathbf{x}_i^1, ..., \mathbf{x}_i^m\}_{i=1}^n$ with n samples and m views, the dataset is used to train a model. The trained model is then used to derive high-quality view-consistent representations and view-specific representations for downstream tasks.

Challenges/Issues: 1. No work on the "simple-first, complex-later" property of deep neural networks, 2. Limitations due to prior assumptions, 3. Fusion disentanglement between consistency and specificity.

Our Solution

Property. *Deep neural networks tend to prioritize memorization of simple instances first and then gradually memorize hard instances [Arpit et al., 2017].*

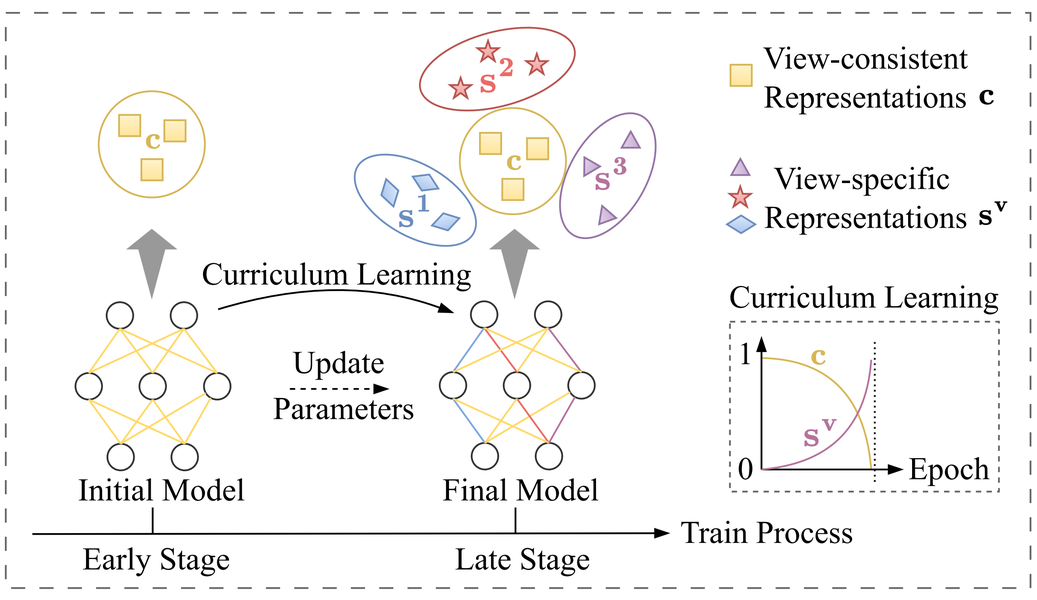


Figure 2: Curriculum learning over view-consistency and view-specificity (first c then s^v).

Proposed Model. Figure 1 illustrates the architecture of our CL2P, which consists of four components:

- **Curriculum learning**, adjusts the whole model to first learn the simple view-consistent representations c and then hard view-specific representations $\{s^v\}_{v=1}^m$ progressively.
- **Mixture-of-Experts (MoE)**, integrates all views into a consistent representation c.
- **Prior learning**, together with view-specific representation learning, drives optimal priors for different views.
- **Disentangling module**, reduces the redundancy between consistent representation c and specific representation $\{s^v\}_{v=1}^m$.

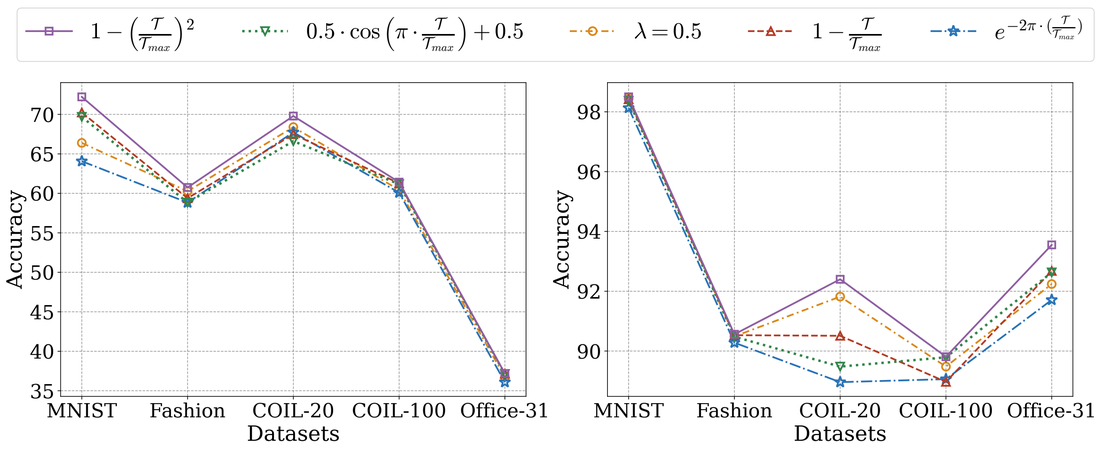
Experiments

Main Results. We evaluate CL2P against ten baselines for clustering and classification tasks on five datasets.

Table 1: Clustering and classification results (%) on five real-world datasets.

Method	Edge-MNIST		Edge-Fashion		Multi-COIL-20		Multi-COIL-100		Multi-Office-31	
	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC
Beta-VAE	53.92	94.00	55.21	80.84	77.80	88.74	80.62	79.27	28.12	40.00
Joint-VAE	14.58	95.30	24.10	79.82	64.46	53.68	70.71	36.20	27.24	25.67
MFLVC	29.00	55.30	27.03	41.95	75.07	55.84	81.97	29.84	34.79	32.02
GCFAgg	24.93	75.86	34.83	78.85	78.77	61.69	82.16	48.37	45.55	54.38
SCM	29.58	89.39	32.16	81.56	68.08	78.79	81.72	63.54	21.01	21.89
CSOT	32.35	54.83	37.80	58.86	63.15	35.06	75.17	33.16	15.48	25.10
MVAE	45.15	97.76	55.25	88.93	77.32	93.37	84.74	92.23	46.24	88.41
DVIB	20.00	76.19	23.51	71.68	67.43	71.86	73.13	72.69	22.10	37.03
Multi-VAE	61.71	95.76	40.69	84.69	79.91	88.31	70.01	74.93	33.30	61.46
MRDD	64.18	98.53	60.51	88.81	79.42	94.38	84.64	91.11	40.22	82.46
CL2P-C	58.91	**98.67**	62.45	89.35	**83.24**	**96.36**	**85.71**	**93.12**	37.31	73.99
CL2P-CS	**67.50**	98.50	**65.53**	**90.56**	80.39	92.40	82.74	89.81	**50.57**	**93.42**

Scheduler Ablation. We investigate multiple strategies to derive the adaptive trade-off parameter λ.

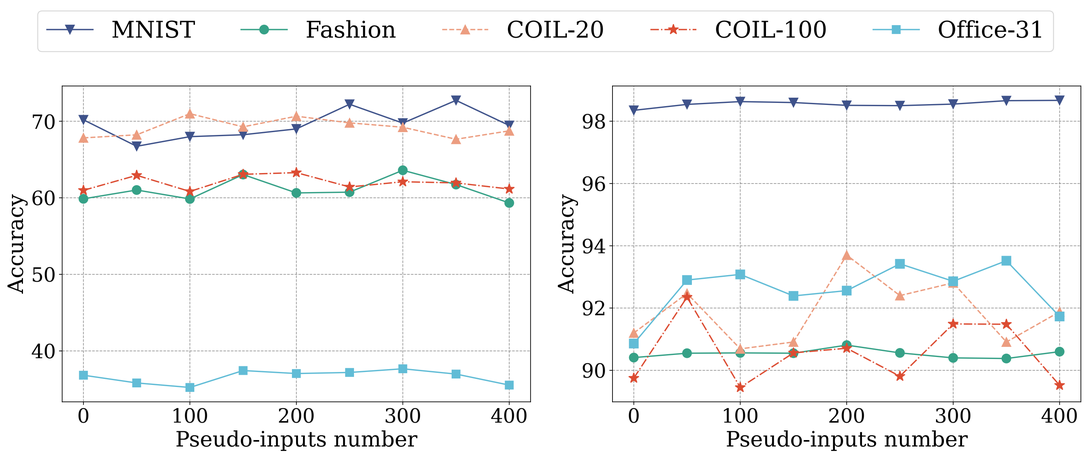


Clustering accuracy (%) of different schedulers on five datasets.

Classification accuracy (%) of different schedulers on five datasets.

Figure 3: Study of different schedulers on curriculum learning.

Parameter Study. We analyze the effect of the parameter K.



Clustering accuracy (%) of different K values on five datasets.

Classification accuracy (%) of different K values on five datasets.

Figure 4: Different metrics against the number K of pseudo-inputs.