第十九届中国图象图形学学会青年科学家会议

Title : Variational Data-Free Knowledge Distillation for Continual Learning

TPAMI 2023

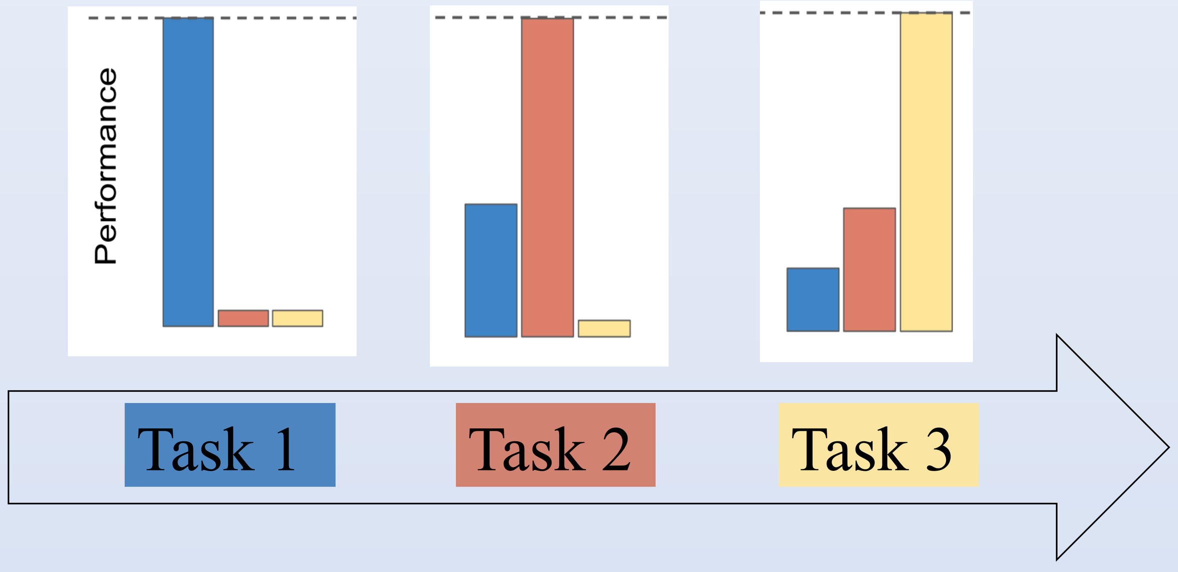
Authors: Xiaorong Li, Shipeng Wang, Jian Sun, Zongben Xu

Background

- Deep neural networks have achieved promising performance on a single task in a wide range of fields. However, they may not work well in an open environment where tasks are encountered continuously.
- Continual learning aims to preserve the performance of the neural network on previous tasks (i.e., stability) when learning new knowledge on a new task (i.e., plasticity).

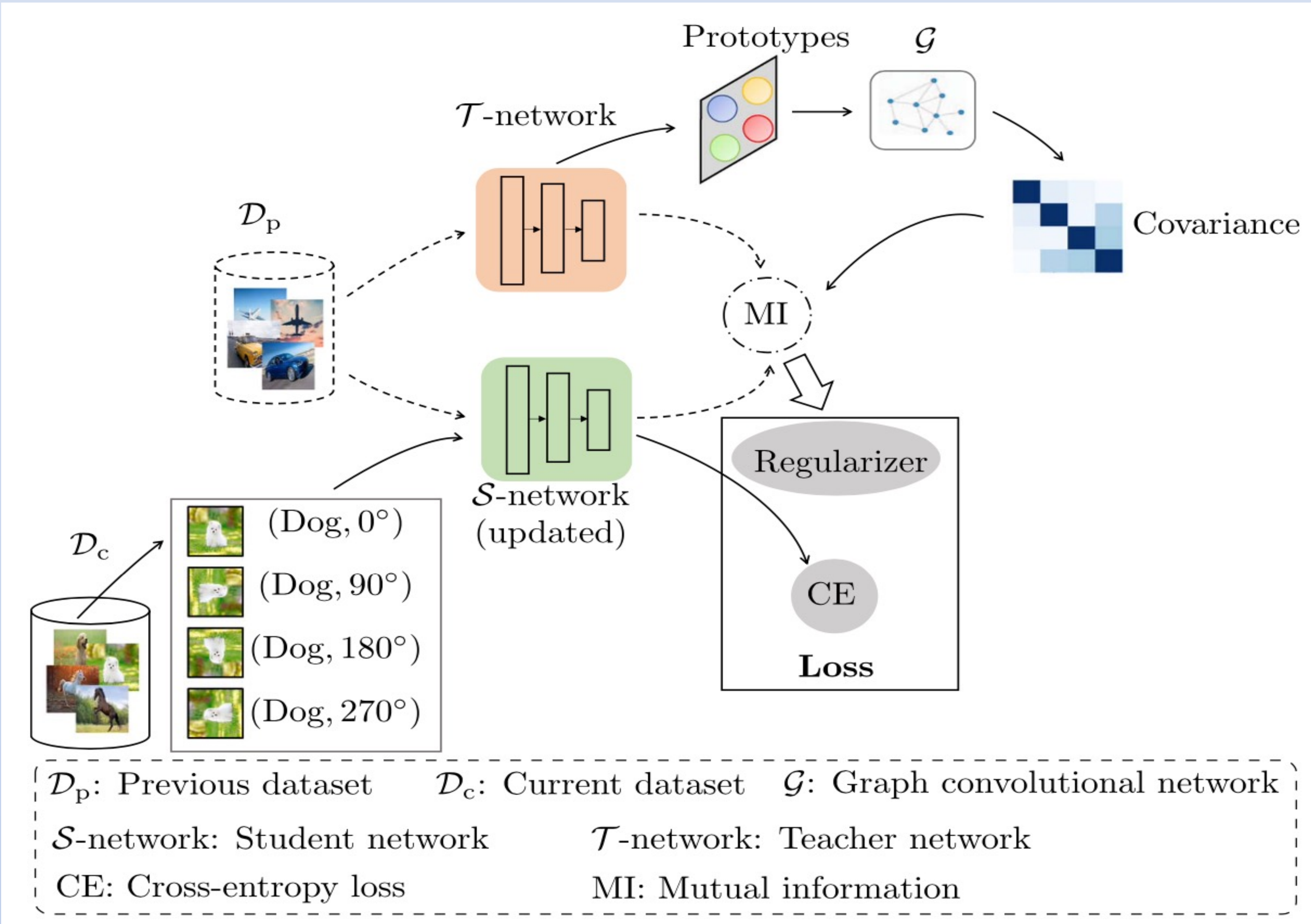
Challenge: Catastrophic Forgetting

The performance on previous tasks becomes worse when training networks on sequential tasks in continual learning



Method: VDFD

➢ Overview



➢ Notation

Input random variable x_i, outputs of student / teacher networks $s_i = f(x_i, w), t_i = f(x_i, w_i^*)$, mutual information between t_i and s_i, $I(t_i; s_i)$

➢ Mitigate Catastrophic Forgetting by VDFD

Maximize the variational lower bound of the mutual information

$$I(t_i; s_i) = H(t_i) - H(t_i|s_i)$$
$$\geq H(t_i) + \mathbb{E}_{t_i,s_i}[log\ q(t_i|s_i)]$$

Assuming that the variational distribution $q(t_i|s_i)$ is Gaussian distribution $\mathcal{N}(s_i, \Sigma_i)$

$$logq(t_i|s_i) = -\frac{1}{2}[(t_i - s_i)^\top \Sigma_i^{-1}(t_i - s_i) + \log|\Sigma_i|] + constant$$

Tackling the inaccessibility of data of previous tasks by first-order Taylor expansion

$$s_i \approx t_i + G_i^\top (w - w_i^*)$$

Variational data-free distillation loss

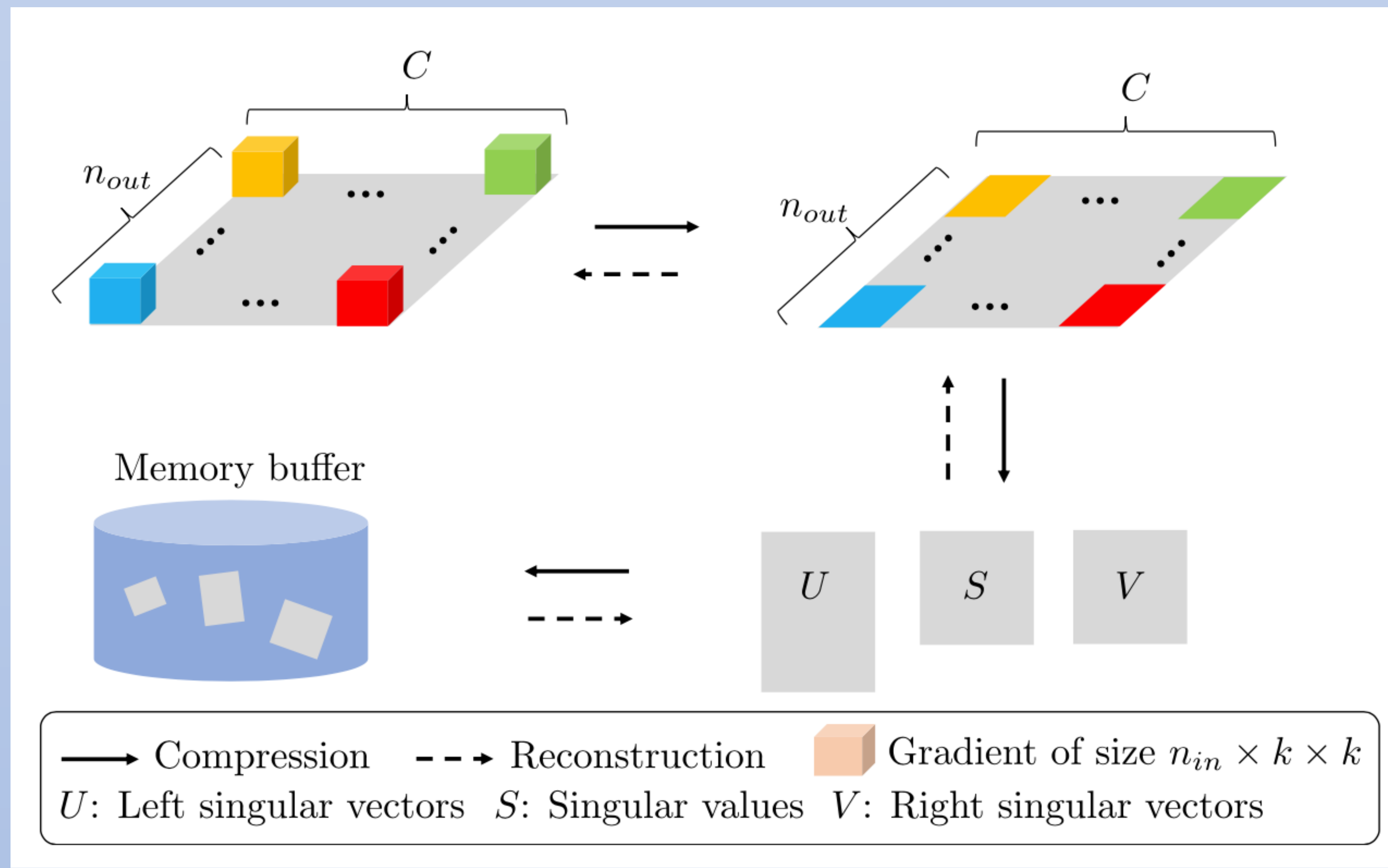
$$\mathcal{L}_{dis}(w; \mathcal{D}_i) \triangleq \mathbb{E}_{t_i}\left[\log|\Sigma_i| + (w - w_i^*)^\top G_i \Sigma_i^{-1} G_i^\top (w - w_i^*)\right]$$

➢ Modeling Covariance by GCN

$$\Sigma_i^{-1} = P^M P^{M\top} + \epsilon I$$

where P^M is the matrix containing all latent vectors of output nodes of GCN parameterized with θ

➢ Compressing the Gradients for memory efficiency



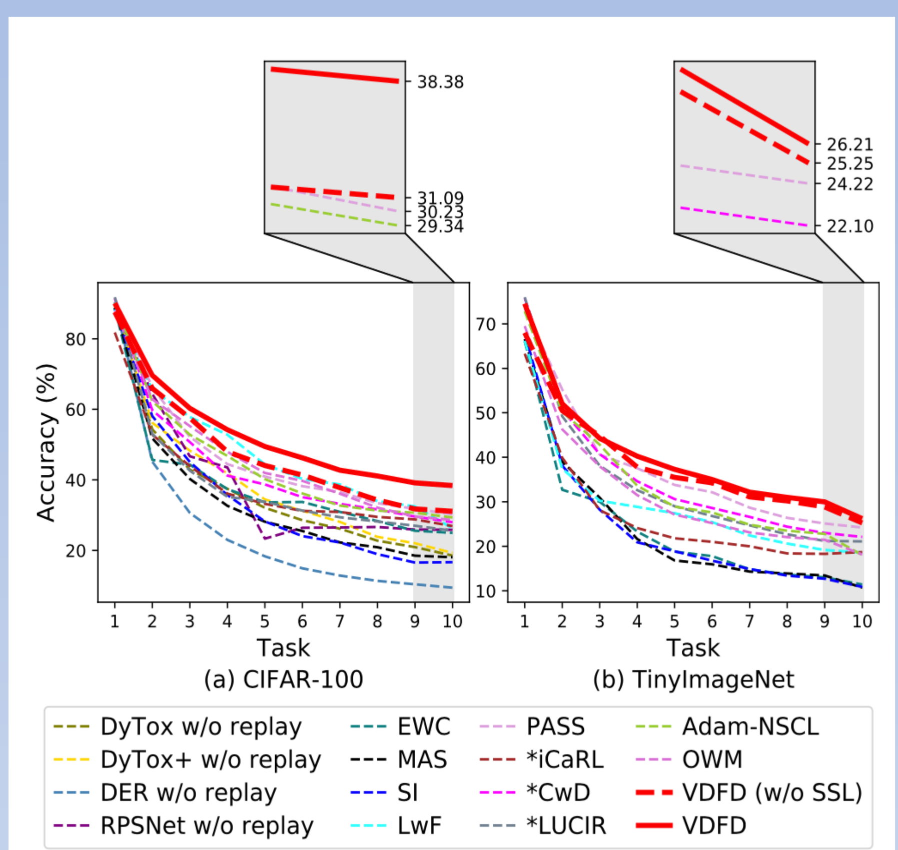
➢ Integrated Objective

$$\min_{w,\theta} \mathcal{L}_{CE}(w; \mathcal{D}_t) + \lambda \sum_{i=1}^{t-1} \mathcal{L}_{dis}(w, \theta; \mathcal{D}_i)$$

Experiments

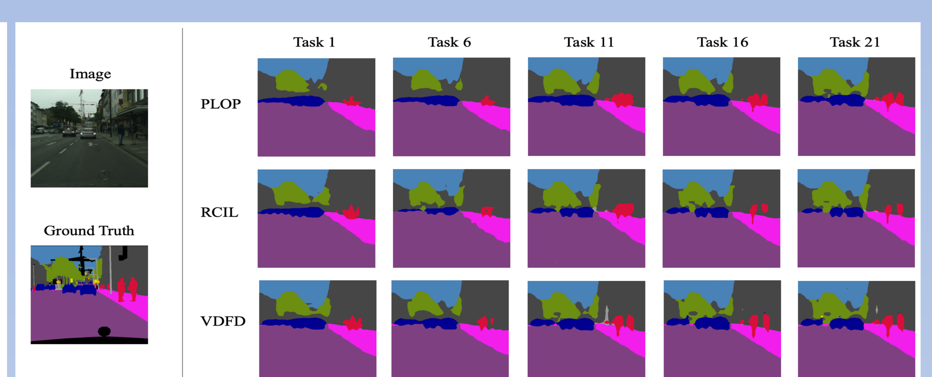
➢ Continual Learning for Image Classification

Method	10-split CIFAR-100		20-split CIFAR-100	
	ACC (%)	BWT (%)	ACC (%)	BWT (%)
InstAParam [45]	47.84	-11.92	51.04	-4.92
Packnet [37]	77.18	**-0.00**	67.50	**-0.00**
BLIP [29]	61.09	<u>-0.70</u>	68.17	-4.21
*GEM [28]	49.48	2.77	68.89	-1.20
*A-GEM [43]	52.73	-1.30	68.92	<u>-0.69</u>
*MEGA [44]	54.17	-2.19	64.98	-5.13
*CTN [27]	53.97	-7.27	67.56	-5.59
EWC [6]	71.28	-2.97	70.90	-3.03
MAS [19]	66.71	-4.61	63.63	-6.36
SI [33]	60.57	-5.17	59.76	-8.65
IMM [7]	62.67	-9.32	57.94	-9.41
MUC-MAS [41]	63.73	-3.88	67.22	-5.72
RankInc [30]	-	-	68.46	**-0.00**
AdNS [32]	77.21	-2.32	77.33	-3.25
LwF [20]	70.70	-6.27	74.38	-9.11
*iCaRL [21]	76.43	-4.87	75.75	-6.08
*GD [24]	66.90	-21.34	78.16	-14.39
PASS [9]	71.23	-5.21	74.43	-4.03
OWM [18]	68.89	-1.88	68.47	-3.37
Adam-NSCL [15]	73.77	-1.60	75.95	-3.66
VDFD (w/o SSL)	<u>79.23</u>	-2.93	<u>80.97</u>	-4.79
VDFD	**83.30**	-1.27	**85.84**	-1.53



➢ Continual Learning for Semantic Segmentation

Method	11-5 (3 tasks)	11-1 (11 tasks)	1-1 (21 tasks)
Fine-tuning	61.55	60.41	41.71
LwF [20]	61.74	60.44	42.93
iCaRL [21]	61.96	60.77	42.51
ILT [46]	61.79	60.45	42.92
MiB [51]	61.72	60.49	42.94
†PLOP [48]	63.51	62.05	45.24
†RCIL [49]	<u>64.30</u>	<u>63.00</u>	<u>48.90</u>
VDFD	**64.77**	**63.53**	**49.34**



Paper

Code

lixiaorong@stu.xjtu.edu.cn
wangshipeng8128@stu.xjtu.edu.cn
jiansun@xjtu.edu.cn
zbxu@xjtu.edu.cn