Class-Incremental Learning for Action Recognition in Videos

Jaeyoo Park **Minsoo Kang** **Bohyung Han**

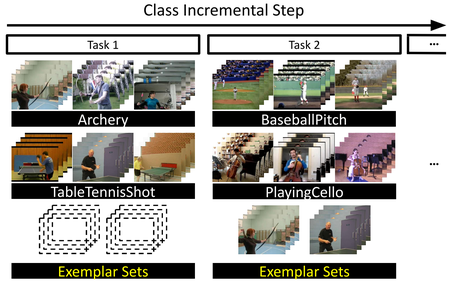
Computer Vision Lab, Seoul National University

2021 ICCV OCTOBER 11-17 VIRTUAL

Introduction

Motivation

We tackle catastrophic forgetting problem in the context of class-incremental learning for video recognition, which has not been explored actively despite the popularity of continual learning.



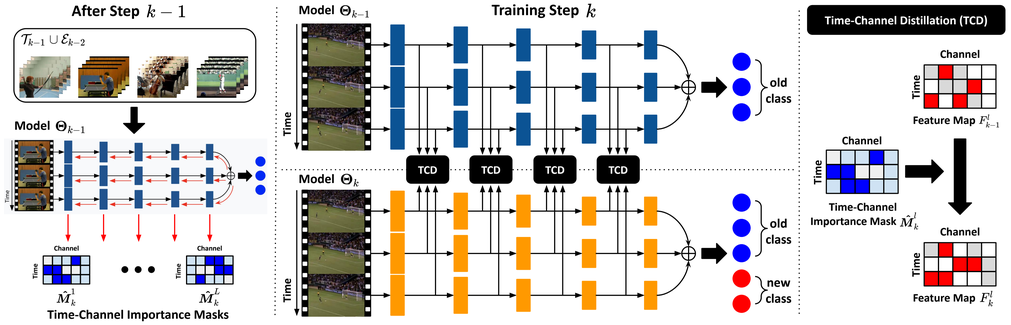
We propose a novel continual learning framework, called **Time-Channel Distillation (TCD).** Our main claim is that time-wise representation should be distilled with different weights depending on their relevance and uniqueness to target classes and maintained for better utilization in the future stages.

Contribution

- We introduce an efficient **class-incremental learning** technique for **action recognition in videos** by adopting a simple **frame-based feature representation** method to store exemplars for the tasks learned in the past.
- Our algorithm estimates **time-channel importances** and **distills knowledge with the importance weight** while **encouraging the diversity of the features in each frame** for regularization and enhance the performance of our target model.
- The proposed approach presents **remarkable accuracy gains on the multiple standard action recognition benchmarks with brand-new splits** compared to the existing methods designed in the image domain.

Our Approach

Overall Framework



Notation

- $\{\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_k, \cdots\}$: Sequence of tasks
- $\mathcal{T}_k$: A set of videos whose labels belong to the predefined classes in $\mathcal{C}_k$, where $(\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_{k-1}) \cap \mathcal{C}_k = \emptyset$
- $\mathcal{E}_k \subset (\mathcal{T}_1 \cup \cdots \cup \mathcal{T}_k)$: A small exemplar set for step k
- $\mathcal{T}'_k = \mathcal{T}_k \cup \mathcal{E}_{k-1}$: Training set for step k
- $\boldsymbol{F}^l_k$: Feature map for l^{th} layer at step k

Orthogonality between Frames

- Suppress correlation between the representations of individual frames to alleviate feature drift issue

$$\mathcal{L}^k_{\text{ortho}} = \sum_{l=1}^{L}\sum_{c=1}^{C_l} \|\boldsymbol{I}_T - \boldsymbol{F'}^l_{k,:,c}(\boldsymbol{F'}^l_{k,:,c})^\top\|^2_F$$

Time-Channel Distillation

- After step $k-1$, compute **Time-Channel importance mask**

$$M^l_{k,t,c} = \mathbb{E}_{(x,y)\sim\mathcal{T}_{1:k-1}} \|\nabla_{\boldsymbol{F}^l_{k-1,t,c}} \mathcal{L}^{k-1}_{\text{cls}}(x,y)\|^2_F$$

- Approximate $\mathcal{T}_{1:k-1}$ using $\mathcal{E}_{k-2}$ and $\mathcal{T}_{k-1}$
- Normalize the mask

$$\hat{M}^l_{k,t,c} = \frac{\tilde{M}^l_{k,t,c}}{\frac{1}{TC_l}\sum_{t=1}^{T}\sum_{c=1}^{C_l}\tilde{M}^l_{k,t,c}}$$

- Distillation objective at step k

$$\mathcal{L}^k_{\text{dist}} = \sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{c=1}^{C_l} \hat{M}^l_{k,t,c}\|\boldsymbol{F}^l_{k,t,c} - \boldsymbol{F}^l_{k-1,t,c}\|^2_F$$

Training Objective

$$\mathcal{L}^k_{\text{final}} = \mathcal{L}^k_{\text{cls}} + \alpha\mathcal{L}^k_{\text{dist}} + \beta\mathcal{L}^k_{\text{ortho}}$$

Evaluation Protocol

Benchmark

	UCF 101	HMDB 51	Sth-Sth V2
Total Classes	101	51	174
Initial Step Classes	51	26	84
Added Class(es) per Step	+2, +5, +10	+1, +5	+5, +10

Metric

- Average Incremental Accuracy
- ***CNN*** : Standard Classification
- ***NME*** : Nearest Mean Exemplars

Experimental Results

Results on UCF 101 & HMDB 51

	UCF101						HMDB51			
Num. of classes	10 × 5 stages		5 × 10 stages		2 × 25 stages		5 × 5 stages		1 × 25 stages	
Classifier	CNN	NME	CNN	NME	CNN	NME	CNN	NME	CNN	NME
Fine-tuning	24.97	—	13.45	—	5.78	—	16.82	—	4.83	—
LwFMC	42.14	—	25.59	—	11.68	—	26.82	—	16.49	—
LwM	43.39	—	26.07	—	12.08	—	26.97	—	16.50	—
iCaRL	—	65.34	—	64.51	—	58.73	—	40.09	—	33.77
UCIR	74.31	74.09	70.42	70.50	63.22	64.00	44.90	46.53	37.04	37.15
PODNet	73.26	74.37	71.58	73.75	70.28	71.87	44.32	48.78	38.76	46.62
TCD (Ours)	**74.89**	**77.16**	**73.43**	**75.35**	**72.19**	**74.01**	**45.34**	**50.36**	**40.07**	**46.66**
Oracle (Upper Bound)	84.15	83.37	83.96	83.20	83.82	83.16	55.03	55.98	54.89	55.32

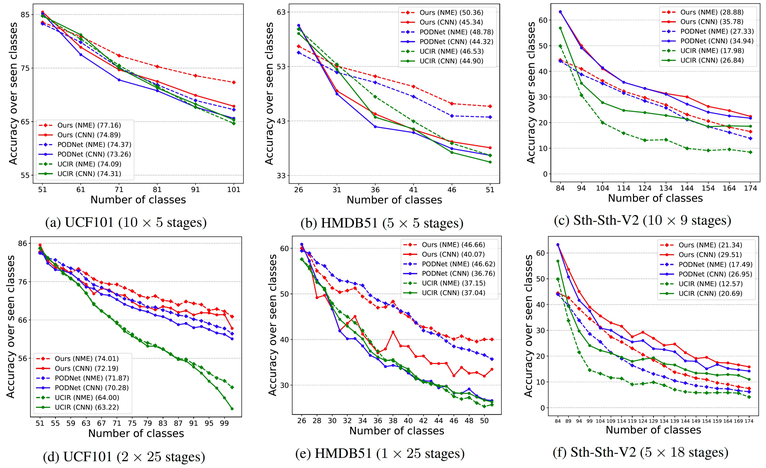
Results on Sth-Sth V2

Num. of classes	10 × 9 stages		5 × 18 stages	
Classifier	CNN	NME	CNN	NME
UCIR	26.84	17.98	20.69	12.57
PODNet	34.94	27.33	26.95	17.49
TCD (Ours)	**35.78**	**28.88**	**29.60**	**21.63**

Ablation Study (UCF 101 with 10 steps)

Objective function	CNN	NME
$\mathcal{L}^k_{\text{cls}} + \mathcal{L}'^k_{\text{dist}}$	71.21	73.24
$\mathcal{L}^k_{\text{cls}} + \mathcal{L}'^k_{\text{dist}} + \mathcal{L}^k_{\text{ortho}}$	72.31	74.42
$\mathcal{L}^k_{\text{cls}} + \mathcal{L}^k_{\text{dist}}$	72.61	74.81
$\mathcal{L}^k_{\text{cls}} + \mathcal{L}^k_{\text{dist}} + \mathcal{L}^k_{\text{ortho}}$ (Ours)	**73.43**	**75.35**

Average Accuracy over Seen Classes at each Incremental Step



(a) UCF101 (10×5 stages) (b) HMDB51 (5×5 stages) (c) Sth-Sth-V2 (10×9 stages)

(d) UCF101 (2×25 stages) (e) HMDB51 (1×25 stages) (f) Sth-Sth-V2 (5×18 stages)