Soft Computing Fusion with Applications

www.scfa.reapress.com

Soft. Comput. Fusion. Appl. Vol. 2, No. 2 (2025) 120–126.

Paper Type: Original Article

Video Class-Incremental Learning for Action Recognition

E. Mohanapriya[1,*], T.T. Mirnalinee[1]

[1] Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Tamil Nadu, India; mohanapriya2320034@ssn.edu.in; MirnalineeTT@ssn.edu.in.

Citation:

Received: 15 October 2024 *Revised: 10 March 2025* *Accepted: 07 May 2025*	Mohanapriya, E., & Mirnalinee, T. T. (2025). Video class-incremental learning for action recognition. *Soft computing fusion with applications, 2*(2), 120-126.

Abstract

In the domain of video-based action recognition, overcoming catastrophic forgetting while continuously learning new classes remains a major challenge. We propose a Video Class-Incremental Learning (VCIL) framework that addresses this issue by employing a teacher-student knowledge distillation strategy. Our approach leverages both response-based distillation, which aligns the student model's predictions with the teacher's softened outputs, and feature-based distillation, which ensures the student retains internal feature representations learned by the teacher. With the UCF101 action recognition dataset and a 3D ResNet backbone, our approach extracts spatiotemporal features to recognize actions in multiple incremental steps. Our model is tested with various settings (10×5, 5×10, 2×25) and has high accuracy for retaining knowledge of past classes and learning new courses. The results show that our approach is efficient in preventing forgetting and maintaining high performance on new tasks.

Keywords: Incremental learning, Action recognition, Knowledge distillation, Deep learning.

1 | Introduction

In a variety of applications, including computer vision, speech recognition, natural language processing, reinforcement learning, and more, deep learning networks have shown exceptional performance. However, considering real-world applications, the traditional training methods need to be trained on both new and old data from scratch because they are not designed to work with new streaming data in connection with the domain. For example, a face recognition system ought to have the capability to incorporate new users without requiring explicit training on previously learned faces. A system that retains knowledge while picking up new tasks, mimicking the operation of the human brain. Accordingly, new knowledge can be continuously acquired by incremental learning while existing knowledge is preserved.

Incremental learning is essential in the action recognition sector because it can adapt to real-world, dynamic situations, including smart cities, industrial settings, autonomous vehicles, healthcare settings, and more.

Corresponding Author: mohanapriya2320034@ssn.edu.in

https://doi.org/10.22105/scfa.v2i2.63

These advances in incremental learning can be made possible by deep learning models, which have some limitations but have made great progress in recognition and classification. A real-world classification system should learn new classes progressively instead of beginning from scratch each time new training data becomes available. However, this brings up a significant issue: Catastrophic forgetting. A deep learning model often overwrites the knowledge it has learned from earlier tasks when it is exposed to new data, which causes it to regress on earlier tasks.

However, there are disadvantages to training a system from scratch for new classes, such as the high cost of training and the requirement to store previous class data for future training. Therefore, as computer vision approaches artificial intelligence, adaptable methods are required to manage the complex and ever-changing environment. Knowledge distillation appears to be one of the more appealing approaches to tackle this problem. An approach for generalizing knowledge within a neural network to train another neural network is called distillation of knowledge, and it is independent of architecture. Training a small model, the student, on a transfer set with soft targets supplied by the complex model, the teacher, is the principle behind knowledge distillation. Training a model may lessen the forgetting to some extent.

2 | Related Works

Incremental learning

Incremental learning is related to continual learning, lifelong learning, transfer learning, multitask learning, and never-ending learning. These algorithms' main goal is not to forget the previously learned task and learn increasingly over time. Incremental learning is, therefore, an important area in the study of machine learning, as there is an increasing number of demands for systems that constantly update the information acquired, yet preserve the prior knowledge. This demand has sparked research on various learning paradigms, such as task-incremental learning, class-incremental learning, and domain-incremental learning [1].

Class-incremental learning

In the domain of image processing, class incremental learning has been researched, and numerous methods have been proposed. Researchers have proposed several strategies to mitigate catastrophic forgetting, which can be broadly categorized into the following techniques: 1) parameter regularization, 2) knowledge distillation, 3) exemplar-based methods, and 4) synthetic data generation. In parameter regularization methods, the task knowledge is preserved by penalizing updates to parameters deemed important for earlier tasks. Focus on preserving the knowledge of previous tasks by penalizing significant updates to parameters deemed important for earlier tasks. Elastic Weight Consolidation (EWC) [1], [2] exploits the Fisher information matrix to calculate the importance of the parameters, and significant changes are prevented through regularization. Synaptic Intelligence (SI) [3] calculates a measure of the contribution of the parameters to performance on each task as training progresses, allowing for selective updates of less critical weights. Learning without Forgetting (LwF) [4] penalizes deviations in earlier tasks' predictions when training on new ones. Knowledge distillation is the most broadly adopted technique in class-incremental learning frameworks. The "teacher-student" model architecture is employed. Here, the teacher is learned from earlier tasks, and such distilled knowledge from the teacher is used as guidance for the student, which is learned explicitly for newer tasks without forgetting previously gained knowledge. Logit-based distillation [5] allows aligning the output probability of the teacher and the student models to retain knowledge. Feature-level distillation [6] ensures that the student learns intermediate feature representations like the teacher. Exemplar-based methods concentrate on a small memory buffer that maintains exemplars from previous tasks, and these exemplars are replayed while learning new classes. The Incremental Classifier and Representation Learning (iCaRL) [7] method addresses the forgetting challenge by employing a nearest-mean-of-exemplars classification strategy that dynamically selects exemplars from incoming data, allowing the model to maintain performance across observed classes while learning new information. Herding strategies are based on the selection of exemplars based on their proximity to the class mean in feature space; hence, it is representative. Another alternative for

storing real exemplars is synthetic data generation by exploiting generative models such as GANs [6] or VAEs [8] for simulating data distributions from prior tasks.

Image-class incremental learning

This paper [9] presents a domain-incremental learning that could minimize catastrophic forgetting through the utilization of Random Vector Functional Link (RVFL) networks. RVFL is a lightweight neural network with fewer parameters but more computationally intense, hence its success in incremental learning settings. An approach to the core of Gram Matrix accumulation could be a way of retaining the previous knowledge learned without requiring access to past data. Important information regarding the previous domains is thus being encoded by the Gram Matrix, which enables the model to learn new domains while conserving performance on earlier ones. This paper compares the proposed method's performance with the joint training approach.

Another work [10] combines Bi-directional Guided Modulation (BGM) and meta-learning principles. The model is trained first on a large corpus of base classes to provide a robust starting point for recognition while getting the model ready for incremental sessions. To have the network be both plastic and stable, BGM presents a modulation network that computes gating masks in order to modulate activations throughout the classification network. These masks are guided through attention maps drawn from the weights of the network, to maintain significant parameters for previously acquired tasks when accommodating new classes. A meta-learning framework based on Model-Agnostic Meta-Learning (MAML) is employed for simulating incremental learning during training. The MAML is through bi-level optimization. The meta-learning framework has produced promising results on benchmark datasets such as CIFAR100, Mini-ImageNet, and CUB200. However, it can be prone to overfitting on particular task orders and depends on a known number of classes and examples.

Video-class incremental learning

The paper [11] has merged a variety of approaches: Knowledge distillation, exemplar management, and attention mechanisms to retain all knowledge learned before acquiring new tasks. Knowledge distillation techniques align feature representations of the current model with those of the previous model, weighted by importance masks that weigh critical temporal and channel dimensions in feature maps. Furthermore, exemplar selection strategies, such as the herding approach, are utilized to maintain representative samples in ways to approximate class distributions without excessive memory usage. In an effort to further enhance feature distinctiveness and counter representation drift, orthogonality regularization is applied to ensure that the temporal features are independent. TSM-based models have been used to process spatiotemporal data.

Class-incremental learning on video-based action recognition by distillation of various knowledge [12] uses two approaches to maintain network information in combination, which are network sharing and network knowledge distillation. The architecture used is a teacher-student framework. The teacher model is first trained on an initial set of classes, and then a student model is trained incrementally. Knowledge distillation is applied at the classification, spatial, and temporal levels to transfer knowledge from the teacher to the student. The loss function combines classification and distillation losses, ensuring that previously learned classes are not forgotten. This approach allows efficient learning of new classes without retraining from scratch. Three-stream architecture with pretrained AlexNet for spatial processing and a separate motion-CNN (TV-L1) for optical flow (Temporal processing), combined with an LSTM for temporal modeling. The proposed paper [13] focuses on a prompt-based incremental learning technique adapted for image-based tasks by using CLIP, a pretrained vision-language model. The methodology is in two stages. Task-specific prompts are optimized independently for each task to integrate new knowledge and retain prior information while keeping the pre-trained model frozen during inference. Task-agnostic prompt matching is used to dynamically select relevant prompts for the test example, enabling accurate classification based on visual-text feature similarity. Task-specific prompts consist of multi-granular prompts (Spatial, temporal, and comprehensive) stored in prompt pools. These are selected through feature matching using K-Means centroids and combined via operations like addition for task-specific integration. Task-agnostic prompts, designed to model temporal

relationships across video frames, are selected using a frame-level attention mechanism to ensure contextual consistency across frames. These prompts are then prepended to the video embeddings, allowing the pre-trained model to understand temporal variations better.

An important addition to the Class-Incremental Learning (CIL) methods is Class-Incremental Masking (CIM), introduced [14]. The proposed paper method focuses on adaptive exemplar compression, where only non-discriminative pixels in the image are downsampled while preserving discriminative features, allowing for more exemplars to be stored within the same memory budget. CIM utilizes Class Activation Maps (CAM) to locate discriminative pixels and employs bilevel optimization to tune masks through incremental steps while maintaining the model's ability to retain existing class information while acquiring new classes—comprehensive experiments on benchmarks such as ImageNet-100 and Food-101.

3 | Proposed Methodology

In this work, we introduce a class-incremental learning paradigm for action recognition from videos on the UCF101 dataset. The goal is to learn a deep neural network that incrementally learns new action classes while maintaining knowledge of previously acquired ones. Our method employs a 3D ResNet-18 model, specially crafted to extract the spatiotemporal features of video data. Incremental learning is organized in staged configurations (e.g., 10×5, 5×10) to enable smooth adaptation with the retention of past knowledge.

In order to counteract catastrophic forgetting, we incorporate knowledge distillation to make sure that the model keeps necessary information from earlier training phases. The process involves five major stages: data preprocessing, initialization of the base model, incremental learning, hybrid loss function, and training and evaluation.

Data preprocessing

The 101 action classes of the UCF101 dataset are dealt with incrementally. In every step, a subset of the classes is added with a balance across all the classes. A video is broken down into fixed frames, resized to 224 times 224 pixels, and normalized to $[-1, 1]$ to maintain consistency. Data augmentation is employed for model generalization enhancement, comprising temporal augmentation (Random frame selection with frame repetition for shorter videos) and spatial augmentation (Random cropping, flipping, brightness change, and contrast). The dataset is then divided into training, validation, and testing subsets for every incremental step.

Base model initialization

The base model is a 3D ResNet-18 architecture, optimized for action recognition. It employs 3D convolutional layers to capture both spatial and temporal information from videos. These layers decompose convolutions using Conv2Plus1D, which sequentially applies spatial and temporal convolutions for enhanced feature extraction. The network integrates residual connections to improve gradient flow and stability during training. Layer normalization and ReLU activation functions are used to strengthen feature learning. A final dense layer maps the extracted features to action classes. The model is initially trained on a subset of classes from the UCF101 dataset, establishing a baseline before incremental learning begins. Implementation is carried out using TensorFlow/Keras, with custom layers designed for specific operations.

Fine-tuning with knowledge distillation

To enable incremental learning, the model is expanded by adding new output neurons for newly incorporated classes. Previous layers are frozen to preserve previous knowledge and update only the new, added layers. To further address catastrophic forgetting, a knowledge distillation scheme is used. A previously trained model (Teacher network) passes its knowledge to an expanded model (Student network) using response-based and feature-based distillation.

I. Response-based distillation: Response-based distillation ensures that the student model aligns its predictions with the teacher model. The loss function is defined as follows:

$$L_{\text{response-distil}} = -\sum_{i=1}^{N} P_{\text{teacher}}(y_i) \log\ (P_{\text{student}}(y_i)), \quad (1)$$

where pteacher(y_i) represents the probability distribution predicted by the teacher model for class y_i, and pstudent(yi) is the corresponding prediction from the student model.

II. Feature-based distillation: Feature-based distillation ensures that the student model retains internal feature representations learned by the teacher model. The feature-based distillation is achieved by minimizing the difference between their intermediate feature maps. The feature-based distillation loss is defined as:

$$L_{\text{feature-distill}} = \frac{1}{N}\sum_{i=1}^{N} \|F_{\text{teacher}}(X_i) - f_{\text{student}}(x_i)\|^2, \quad (2)$$

where $f_{\text{teacher}}(x_i)$ and $f_{\text{student}}(x_i)$ represent the extracted feature embeddings for input x_i, and $\|.\|^2$ denotes the squared Euclidean norm.

4 | Implementation

Dataset

The UCF101 dataset, a widely used benchmark for action recognition, consists of 101 action classes across 13,320 videos. To facilitate incremental learning, the dataset was partitioned into staged configurations, where a subset of classes was introduced incrementally, such as 10 × 5 or 5 × 10. Each video was resized to 224 × 224 pixels and normalized within the range of [−1, 1]. The dataset was divided into training, validation, and testing subsets. To enhance generalization, data augmentation techniques, including random cropping, flipping, brightness adjustment, and temporal frame sampling, were applied during training.

Model architecture

The suggested system, developed with TensorFlow, is based on a 3D ResNet model with (2+1)D convolutions to effectively learn spatiotemporal features for action recognition with minimized computational complexity. This architecture separates 3D kernels into 2D spatial and 1D temporal convolutions to promote better temporal dynamics. Residual connections enable effective gradient flow and training of deep networks. The model was first trained on 10 classes for 100 epochs with Adam optimizer (Learning rate 0.0001, batch size 2) with categorical cross-entropy loss.

To simulate real-world scenarios, extra classes were added gradually in groups of 10. A knowledge distillation framework was applied to mitigate catastrophic forgetting, where the previous model (Teacher) provides softened probability distributions to guide the current model (Student). Two distillation techniques were used: response-based, which aligns output logits, and feature-based, which aligns intermediate features using L2 loss.

Evaluation protocol

We adopt the Top-1 accuracy on the validation set after each incremental update as our primary evaluation metric. To assess forgetting, we record the model's performance on previously learned classes before and after the addition of new classes. The forgetting rate is calculated as the difference in accuracy on old classes before and after each incremental step.

We evaluate the performance of both response-based and feature-based distillation techniques under varying incremental configurations. The effectiveness of each method is assessed in terms of overall accuracy and forgetting rate across configurations.

5 | Results

Experiments were conducted on the UCF101 dataset using three incremental configurations: 10×5, 5×10, and 2×25,

where 50 classes are introduced in groups of 5, 10, and 25, respectively. Our focus was on preserving knowledge across stages using response-based and feature-based distillation.

The results demonstrate that while response-based distillation alone is insufficient to maintain performance across all stages—especially when class diversity increases—feature-based distillation improves retention but still suffers from moderate forgetting. Comparative results are summarized below:

Table 1. Performance of distillation strategies (Accuracy/forgetting rate).

Method	10×5	5×10	2×25
Response distillation	69.34 / 11.00	70.08 / 40.26	74.30 / 26.04
Feature distillation	72.00 / 22.00	74.00 / 26.34	79.00 / 21.34

These results highlight that both methods have their strengths, with feature-based distillation generally outperforming response-based distillation in terms of knowledge retention. However, each method's performance varies with the incremental configuration, underscoring the importance of choosing an appropriate distillation strategy based on deployment scenarios.

6 | Conclusion

We present a class-incremental learning framework for video-based action recognition that effectively addresses catastrophic forgetting through the integration of response-based and feature-based knowledge distillation techniques. Our approach ensures that the student model retains critical information from previous stages by aligning both its predictions and internal feature representations with those of the teacher model. The proposed method demonstrates significant improvements in knowledge retention, with strong performance on previously learned classes while adapting to new ones. Evaluation results across different incremental configurations show the effectiveness of our framework in maintaining high recognition accuracy, showcasing its potential to overcome the challenges of incremental learning in dynamic environments.

Funding

No funding was received for this research.

Conflicts of Interest

The authors confirm there are no competing interests.

References

[1] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., … ., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

[2] Feng, T., Wang, M., & Yuan, H. (2022). Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR) (pp. 9427–9436)*. IEEE. https://doi.org/10.1109/CVPR52688.2022.00921

[3] Zenke, F., Poole, B., & Ganguli, S. (2017). Continual learning through synaptic intelligence. *Proceedings of the 34th international conference on machine learning* (pp. 3987–3995). PMLR. https://proceedings.mlr.press/v70/zenke17a.html

[4] Li, Z., & Hoiem, D. (2018). Learning without Forgetting. *IEEE transactions on pattern analysis and machine intelligence*, *40*(12), 2935–2947. https://doi.org/10.1109/TPAMI.2017.2773081

[5] Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the knowledge in a neural network*. https://doi.org/10.48550/arXiv.1503.02531

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … ., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the acm*, *63*(11), 139–144. https://doi.org/10.1145/3422622

[7] Rebuffi, S. A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017). iCaRL: Incremental classifier and representation learning. *Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2001–2010)*. IEEE. https://doi.org/10.1109/CVPR.2017.587

[8] Kingma, D. P. (2017). *Variational inference & deep learning.* [Thesis]. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://pure.uva.nl/ws/files/17891313/Thesis.pdf

[9] Liu, C., Wang, Y., Li, D., & Wang, X. (2024). Domain-incremental learning without forgetting based on random vector functional link networks. *Pattern recognition*, *151*, 110430. https://doi.org/10.1016/j.patcog.2024.110430

[10] Chi, Z., Gu, L., Liu, H., Wang, Y., Yu, Y., & Tang, J. (2022). Metafscil: A meta-learning approach for few-shot class incremental learning. *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 14166–14175). IEEE. https://doi.org/10.1109/CVPR52688.2022.01377

[11] Park, J., Kang, M., & Han, B. (2021). Class-incremental learning for action recognition in videos. *2021 IEEE/CVF international conference on computer vision (ICCV)* (pp. 13698–13707). IEEE. https://doi.org/10.1109/ICCV48922.2021.01344

[12] Maraghi, V. O., & Faez, K. (2022). Class-incremental learning on video-based action recognition by distillation of various knowledge. *Computational intelligence and neuroscience*, *2022*(1), 4879942. https://doi.org/10.1155/2022/4879942

[13] Pei, Y., Qing, Z., Zhang, S., Wang, X., Zhang, Y., Zhao, D., & Qian, X. (2023). Space-time prompting for video class-incremental learning. *2023 IEEE/CVF international conference on computer vision (ICCV)* (pp. 11932–11942).IEEE. https://doi.org/10.1109/ICCV51070.2023.01096

[14] Luo, Z., Liu, Y., Schiele, B., & Sun, Q. (2023). Class-incremental exemplar compression for class-incremental learning. *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 11371–11380). IEEE. https://doi.org/10.1109/CVPR52729.2023.01094