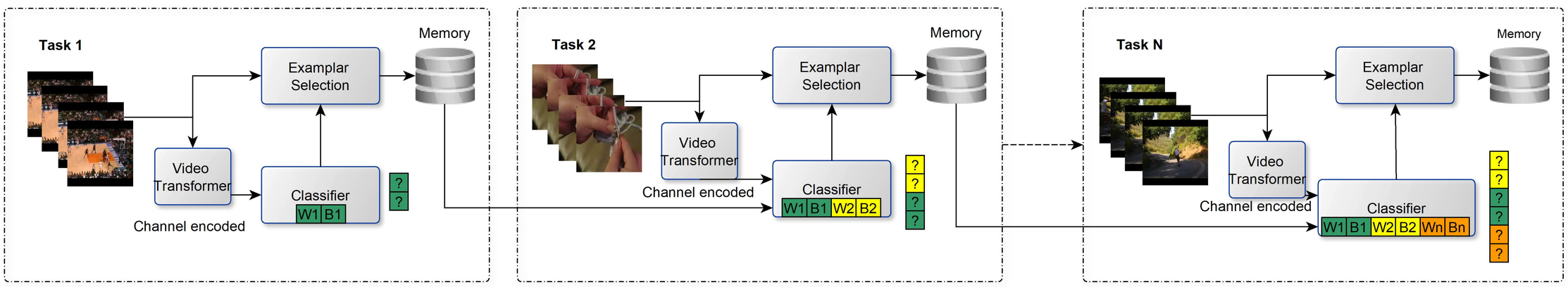
TEMPORAL TRANSFORMER ENCODER FOR VIDEO CLASS INCREMENTAL LEARNING

Nattapong Kurpukdee and Adrian G. Bors,
Department of Computer Science, University of York, York YO10 5DD, UK
Email : nattapong.kurpukdee@york.ac.uk, adrian.bors@york.ac.uk

ICIP abu dhabi 2024
IEEE International Conference on Image Processing
27 - 30 October 2024 | Abu Dhabi, UAE



A memory-based video transformer for Class Incremental Learning (CIL). The video CIL model learns a set of video classification tasks and memorizes the features of each class in a memory buffer, which is added when learning a new classes.

Abstract

- ✓ Current video classification approaches suffer from catastrophic forgetting when retrained on new databases.
- ✓ Continual learning aims to enable a classification system with learning from a succession of tasks without forgetting.
- ✓ We propose to use a characteristic spatiotemporal feature from videos extracted by a transformer-based model for video continual learning.
- ✓ To prevent catastrophic forgetting problems, we gradually built and trained a new classifier model with video data from new tasks combined with the memory data.
- ✓ Our proposed model is evaluated on standard action recognition datasets including UCF101 and HMDB51, which are split into sets of classes, to be learnt sequentially.

Methodology

Architecture

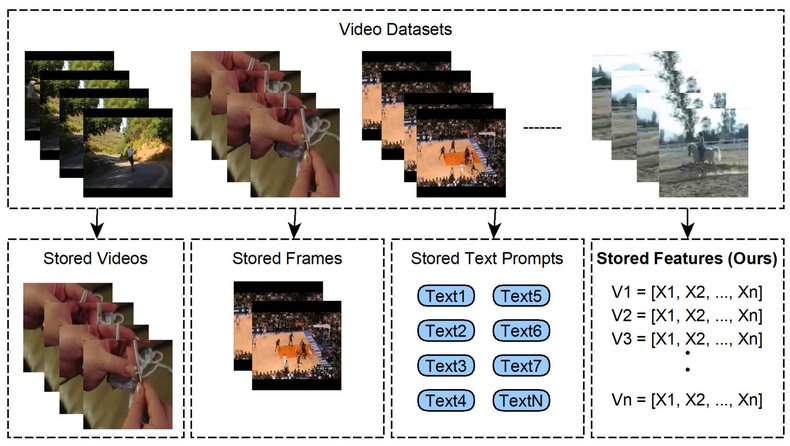
The overview of our proposed pipeline is illustrated in the figure above. A video transformer is considered for extracting spatiotemporal video features while a series of modules are built upon the transformer.

1. Initially, the model is composed of a video transformer model and a classification module, which is considered a Multi-Layer Perceptron (MLP) network.
2. The features extracted by the transformer from the data corresponding to a certain class are used for training a classifier.
3. After completing the training of the classifier, the features extracted by the transformer are stored in a memory buffer.
4. When new video samples become available through a continual incoming stream of video data, the classifier model is trained again on both the features extracted from the videos from the new task videos as well as those from the existing buffers.
5. New buffers with transformer-based features corresponding to each new task are added continuously to the memory buffer during the CIL.

Memory buffer management

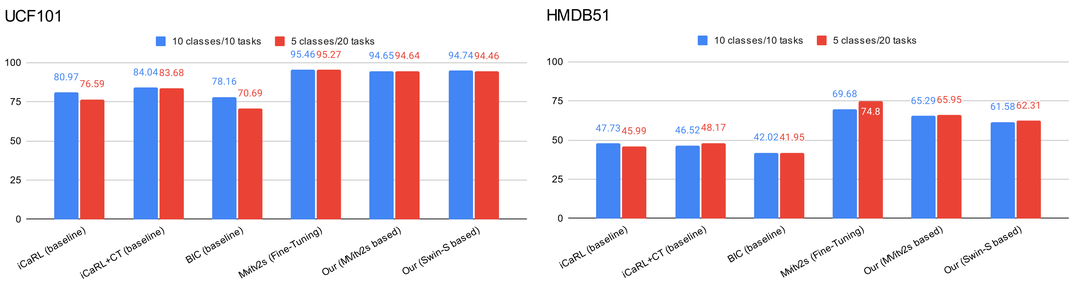
A certain video will be kept in the memory buffer. In our approach, we tackle the memory problem by proposing as following:

- ✓ We store features instead of the entire video.
- ✓ We reducing the amount of stored data by roughly 99% when compared with storing the entire video.



Results

Average Accuracy For Action recognition performance on UCF101 and HMDB51 with no initial learned classes.



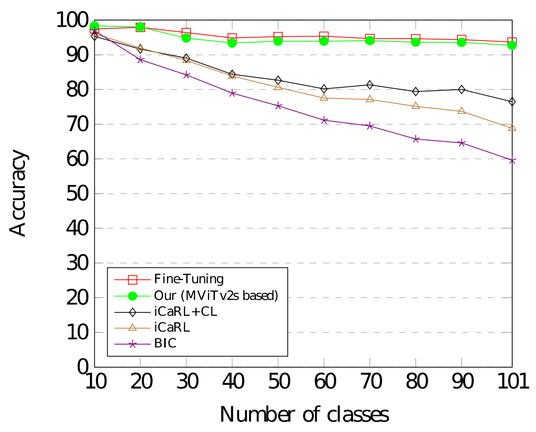
On UCF101 :

- ✓ achieves up to 10.70% at learning 10 classes per task.
- ✓ achieves up to 10.96% at learning 5 classes per task.

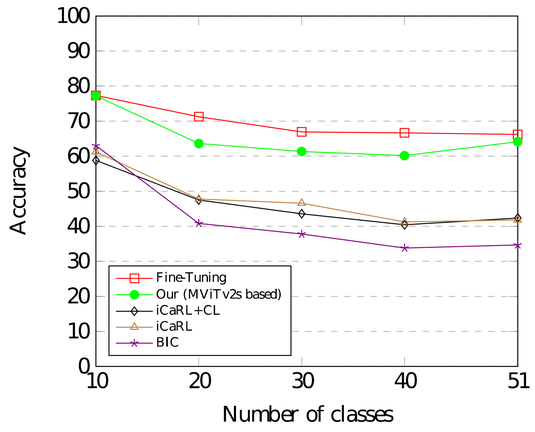
On HMDB51 :

- ✓ achieves up to 17.89% at learning 10 classes per task.
- ✓ achieves up to 20.39% at learning 5 classes per task.

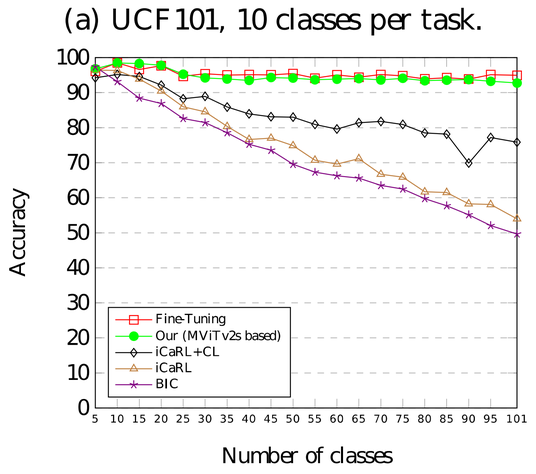
Results on UCF101 and HMDB51 when considering 10 (top) and 5 (bottom) classes per task.



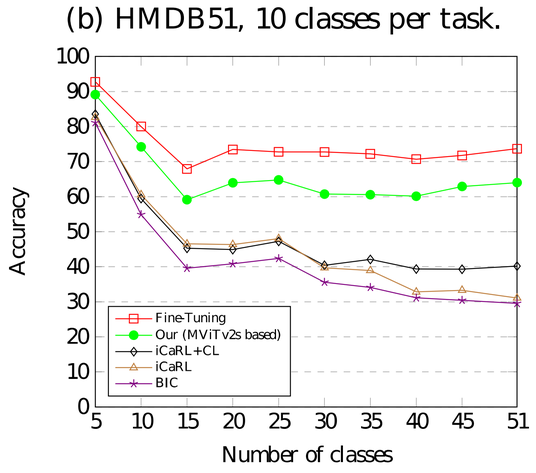
(a) UCF101, 10 classes per task.



(b) HMDB51, 10 classes per task.

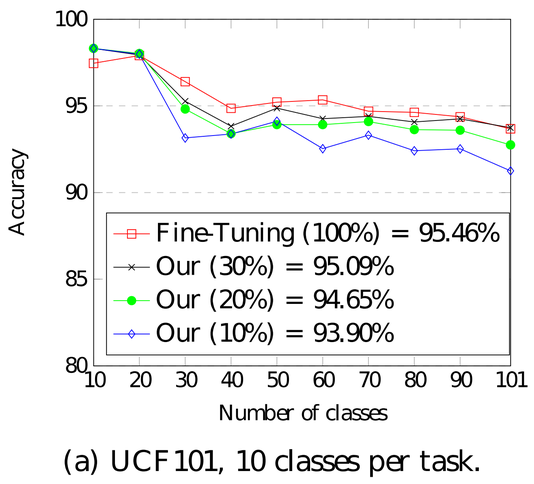


(c) UCF101, 5 classes per task.

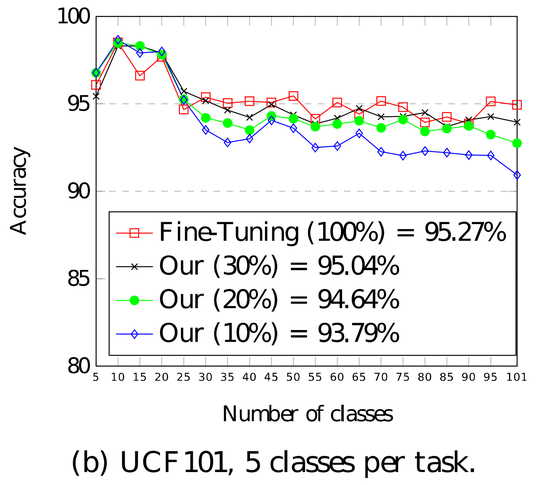


(d) HMDB51, 5 classes per task.

Results on UCF101 when varying the number of video features stored for the continual learning.



(a) UCF101, 10 classes per task.



(b) UCF101, 5 classes per task.

Conclusions

- ✓ We proposed to use the temporal transformer features with the memory-based architecture for video class incremental learning.
- ✓ We replace storing original videos with storing extracted features, thus significantly reducing the required memory.
- ✓ We replay extracted features, thus efficiently training time during video CIL.
- ✓ Our proposed method outperforms other video CIL models in both performance and efficiency, while not requiring an initial set of classes to initiate its training.

Acknowledgements

The authors would like to acknowledge to the Royal Thai Government Scholarship, the Office of Educational Affairs The Royal Thai Embassy for their funding support.

IEEE Signal Processing Society