Efficient Convolutional Network for Online Video Understanding

Dev Barbhaya[1], Ankit Yadav[2], Lochan Gupta[3], Utkarsh Agrawal[4], and Vinamra Shrivastava[4]

[1]Department of Biological Sciences and BioEngineering [2]Department of Mechanical Engineering [3]Department of Civil Engineering [4]Department of Electrical Engineering

1 Objective

The state of the art in video understanding suffers from two problems: (1) Important relationships within actions that spans over several seconds are missed as the major part of reasoning is performed locally in the video. (2) The processing of the whole video is not efficient and hampers fast video retrieval or online classification of long-term activities.

In this paper, network architectural that takes long term content into account and enables fast per-video processing is introduced. Rather than merging post-hoc fusion, this architecture aims at merging long term content. Sampling strategy, which exploits the neighboring frames which are largely redundant. The approach achieves effective performance while being 10x to 80x faster than state-of-the-art methods.

Firstly, a good initial classification of an action can already be obtained from just a single frame. We process only a single frame of a temporal neighborhood efficiently with a 2D convolutional architecture in order to capture appearance features of such frame. Secondly, to capture the contextual relationships between distant frames, we feed the feature representations of distant frames into a 3D network that learns the temporal context between these frames and so can improve significantly over the belief obtained from a single frame especially for complex long-term activities.

2 Stages

2.1 Video classification with deep learning

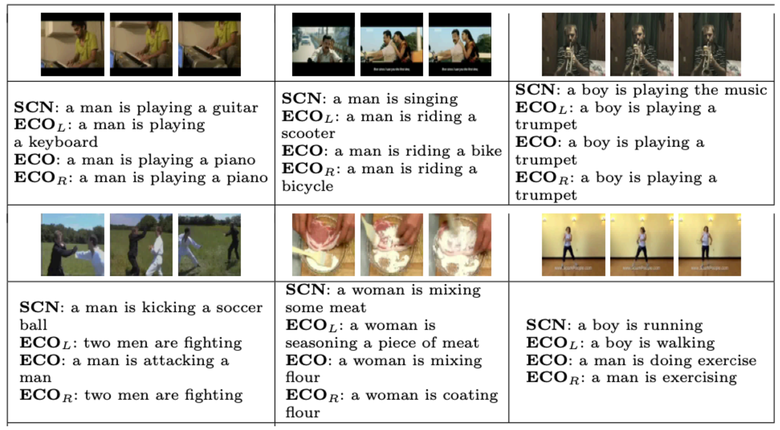
the use of a Resnet architecture with 3D convolutions was studied and it showed the improvements over their earlier c3d architecture. Extra feature/score aggregation reduces the speed of video processing and disables the method to work in a real-time setting.

2.2 Long-term representation learning

Expanding the temporal length of the input has two major drawbacks. (1) It is computationally expensive, and (2) still fails to cover the visual information of the entire video, especially for longer videos. Each video is split into N subsections of equal size. From each subsection a single frames is randomly sampled. The samples are processed by a regular 2D convolutional network to yield a representation for each sampled frame. These representations are stacked and fed into a 3D convolutional network, which classifies the action.

2.3 Video Captioning

A CNN pretrained on image classification or action recognition to generate features. It utilized a frame-based feature aggregation or a sliding window over the whole video to generate video-level features. The features are then passed to a recurrent neural network to generate the video captions via a learned language model.

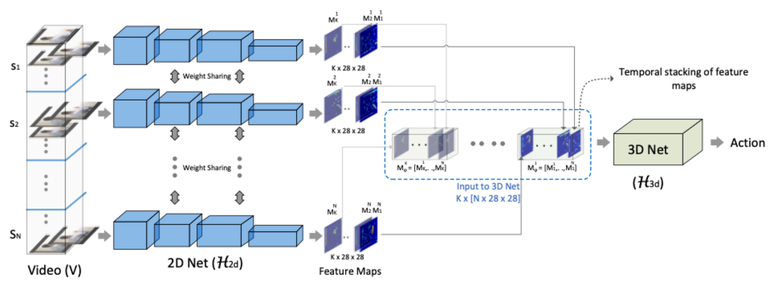


SCN: a man is playing a guitar **ECO**$_L$: a man is playing a keyboard **ECO**: a man is playing a piano **ECO**$_R$: a man is playing a piano	**SCN**: a man is singing **ECO**$_L$: a man is riding a scooter **ECO**: a man is riding a bike **ECO**$_R$: a man is riding a bicycle	**SCN**: a boy is playing the music **ECO**$_L$: a boy is playing a trumpet **ECO**: a boy is playing a trumpet **ECO**$_R$: a boy is playing a trumpet
SCN: a man is kicking a soccer ball **ECO**$_L$: two men are fighting **ECO**: a man is attacking a man **ECO**$_R$: two men are fighting	**SCN**: a woman is mixing some meat **ECO**$_L$: a woman is seasoning a piece of meat **ECO**: a woman is mixing flour **ECO**$_R$: a woman is coating flour	**SCN**: a boy is running **ECO**$_L$: a boy is walking **ECO**: a man is doing exercise **ECO**$_R$: a man is exercising

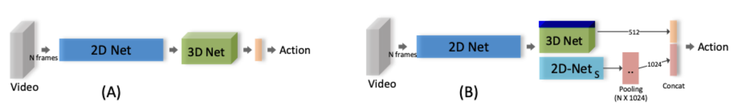
3 Long-term Spatio-temporal Architecture

3.1 ECO Lite and ECO Full

ECO Lite tends to waste capacity. Therefore, an extension of the architecture by using a 2D network in parallel. For the simple actions, this 2D network architecture can simplify processing and ensure that the static image features receive the necessary importance, whereas the 3D network architecture takes care of the more complex actions that depend on the relationship between frames.



Architecture overview of ECO Lite.

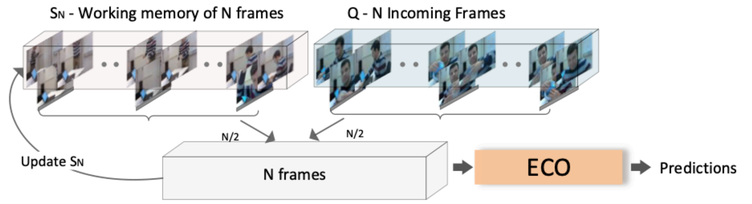


(A) ECO Lite architecture as shown in more detail in (B) Full ECO architecture with a parallel 2D and 3D stream.

3.2 Network details

2D-Net	3D-Net	2D-Net$_S$
For the 2D network (H_{2D}) that analyzes the single frames, we use the first part of the BN-Inception architecture (until inception-3c layer) .	For the 3D network H_{3D} we adopt several layers of 3D-Resnet18, which is an efficient architecture used in many video classification works .	For this network, we use the BN-Inception architecture from inception-4a layer until last pooling layer .
The output of H_{2D} for each single frame consist of 96 feature maps with size of 28 × 28.	The out- put of H_{3D} is a one-hot vector for the different class labels.	The last pooling layer will produce 1024 dimensional feature vector for each frame.

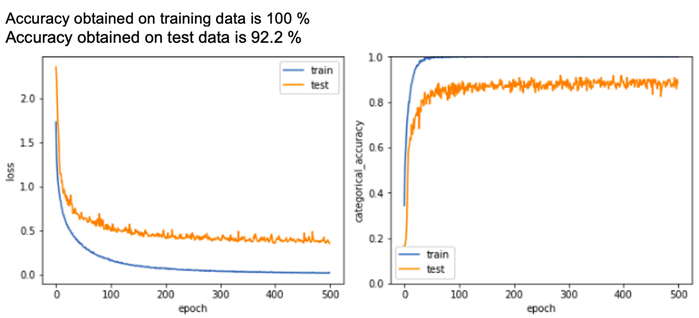
3.3 Training details



3.4 Test time inference

This network produces action labels for the whole video directly without any additional aggregation. We sample N frames from the video, apply only center cropping then feed them directly to the network, which provides the prediction for the whole video with a single pass.

4 Results



https://github.com/soulsucker4600/BCS_*ECO* – *Teams*1–