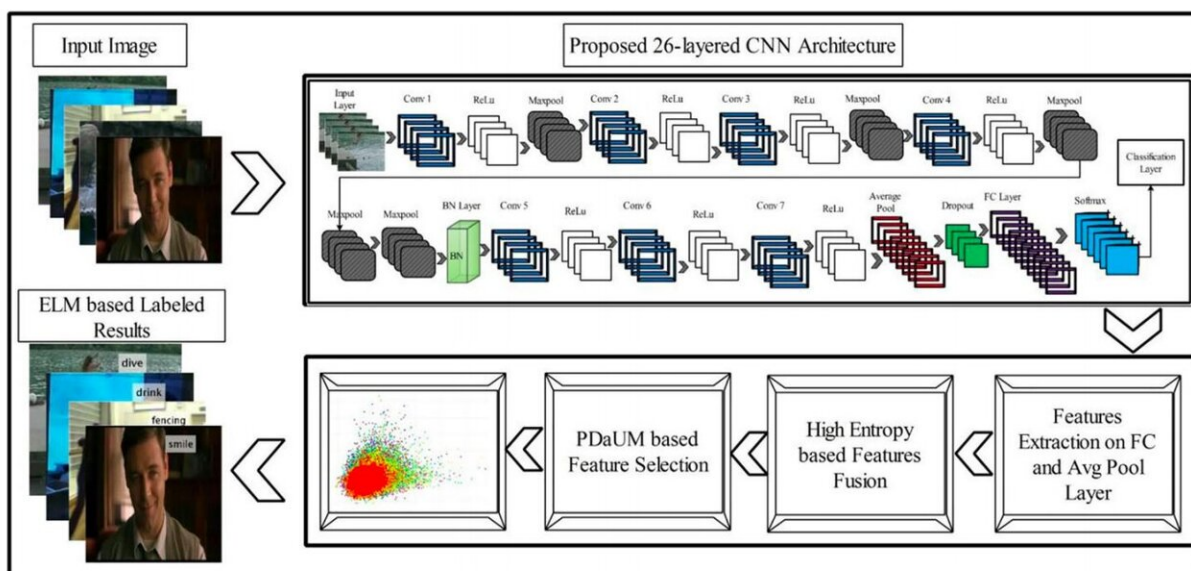# A 26-layer convolutional neural network for human action recognition

August 25 2020, by Ingrid Fadelli

The architecture for human action recognition based on a 26-layer CNN and PDaUM approach proposed by the researchers. Credit: Khan et al.

Deep learning algorithms, such as convolutional neural networks (CNNs), have achieved remarkable results on a variety of tasks, including those that involve recognizing specific people or objects in images. A task that computer scientists have often tried to tackle using deep learning is vision-based human action recognition (HAR), which specifically entails recognizing the actions of humans who have been

captured in images or videos.

Researchers at HITEC University and Foundation University Islamabad in Pakistan, Sejong University and Chung-Ang University in South Korea, University of Leicester in the UK, and Prince Sultan University in Saudi Arabia have recently developed a new CNN for recognizing human actions in videos. This CNN, presented in a paper published in Springer Link's Multimedia Tools and Applications journal, was trained to differentiate between several different human actions, including boxing, clapping, waving, jogging, running and walking.

"We designed a new 26-layered convolutional neural network (CNN) architecture for accurate complex action recognition," the researchers wrote in their paper. "The features are extracted from the global average pooling layer and fully connected (FC) layer and fused by a proposed high entropy-based approach."

When trying to recognize the actions of humans in images or videos, CNNs typically focus their analyses on a number of potentially relevant features. Some human actions, however, such as jogging and walking, can be very similar, which makes it harder for these algorithms to tell them apart, especially if they are focusing on features that are at the core of their similarities. To overcome this challenge, the researchers used an approach that merges a feature selection method called Poisson distribution with univariate measures (PDaUM).

The researchers observed that the features on which CNNs base their analyses can sometimes be irrelevant or redundant, which causes them to make incorrect predictions. To reduce the risk of this happening, their PDaUM approach only selects the strongest features for recognizing a particular human action and ensures that the CNN makes its final prediction based on these features.

The researchers trained and evaluated two distinct CNNs, an extreme learning machine (ELM) and a Softmax classifier, on four datasets, namely the HMDB51, UFC Sports, KH, and Weizmann datasets. These datasets contain several videos of humans performing different types of actions.

The researchers then compared the performance of their two CNNs , both of which were enhanced using their feature selection method. In their evaluations, the ELM classifier performed significantly better than the Softmax algorithm, recognizing human actions from videos with an accuracy of 81.4% on the HMDB51 dataset, 99.2% on the UCF Sports dataset, 98.3% on the KTH dataset and 98.7% on the Weizmann dataset.

Remarkably, the PDaUM-enhanced ELM classifier also outperformed all the existing deep learning techniques for HAR that the researchers compared it with, both in terms of accuracy and prediction time. These results thus highlight the potential of the feature selection method introduced by the researchers for improving the performance of CNNs on HAR tasks.

In the future, the ELM classifier and PDaUM approach introduced in this paper could enable the development of more effective tools for automatically distinguishing what humans are doing in both recorded and live video footage. These tools could prove valuable in several different ways, for instance, helping law enforcement agents to monitor the behavior of suspects on CCTV videos or allowing researchers to rapidly analyze large quantities of videos.