C3D Critique

James Peralta
Department of Computer Science
University of Calgary
Calgary, Canada
james.peralta@ucalgary.ca

I. Overview

The amount of videos on the internet has been growing rapidly due to platforms such as YouTube where users are posting thousands of videos every minute. This is a continuing trend so it is essential to understand and analyze these videos for purposes like search and recommendations. The computer vision community has been analyzing videos for a long time but most video descriptors are very specific to a task such as recognizing actions [2], abnormal event detection [3], and activity recognition [4]. The problem is that the types of videos that are posted by users can be very different between each other, so it is essential to create a generic video descriptor. A generic video descriptor is an algorithm that describes the features of a video and is able to do this for many different tasks. Tran et al [1] proposes to use a 3-dimensional convolutional neural network (3D CNN) as a general video descriptor because CNNs are currently the state of the art for image and video recognition tasks. There are three main contributions from this paper. They show that 3D CNNs are better for video analysis than 2D CNNs, 3x3x3 convolutional kernels produce the highest accuracy, and their 3D CNN is able to generalize to multiple tasks and outperform the current state of the art on 4 different tasks.

To prove that 3D CNNs perform better than 2D CNNs they performed a theoretical review and experimental analysis of the architecture. First they argued how 2D CNNs naturally lose their temporal information because they take as input an image and output an image. In contrast, 3D CNNs take as input a video and output a video so it is more equipped for handling temporal data. To prove this, they compared the performance of both architectures against each other on the UCF-101 dataset and saw that their 3D CNNs outperformed state of the art 2D CNNs. Now that they were convinced that 3D CNNs were the way to go, they wanted to figure out the optimal size of kernels to use. The kernels of a CNN are important because they are the filters that look for certain patterns in the input. To empirically find the best kernel size, they tested several sizes including 3x3x3, 3x3x5, 3x3x7, etc. They tested each on the UCF-101 dataset again and found the 3x3x3 was the optimal size.

Using the results of these two experiments they designed a generic video descriptor called “C3D” which is a 3D CNN which uses 3x3x3 kernels. To test how well this generic video descriptor worked, they introduced 6 different datasets. These datasets contained 4 different video analysis tasks; action recognition, action similarity labelling, scene recognition, and object recognition. For each task they created a classification model using the C3D as a feature extractor and compared its performance against the state of the art methods. They went on to show that their C3D was able to generalize to each task and outperform these state of the art methods in all 4 benchmarks. In conclusion, they were able to systematically create a state of the art general video descriptor using a novel Neural Network architecture.

II. Strengths

The research is solving a real problem that will become increasingly important. We no longer just have to index and search through just text. Videos are becoming a major source of information and will continue to be in the future. As this supply of videos continues to increase, it will be important to create algorithms that can shift through all of this data and create meaningful representations. Another strength was that most of their design decisions leveraged state of the art research done in the past. They were aware of previous literature where 2D CNNs have blown out all classical image recognition methods. On top of this, many people were achieving state of the art results in video analysis tasks by converting 2D image classifiers into 3D video classifiers. This was the motivation for expanding a 2D CNN into 3D and they also mathematically argued how 3D CNNs were able to retain temporal information better than 2D CNNs.

Furthermore instead of just arguing that their method worked the best, they carefully crafted experiments to test it against the many alternatives. For example, they tested different 3D CNNs against a 2D CNN in the task of Action Recognition on the UCF 101 dataset and saw that every variation actually performed better than the 2D CNN. For each dataset, they mentioned the size, amount of classes, and the current state of the art method. The results were displayed clearly using figures with precise titles and headings. All of their conclusions were backed up with experimental evidence. This paper has the chance to affect how we approach video recognition. Not only was their architecture novel and explained well, it was rigorously tested against the top methods in different tasks.

III. Weaknesses

The goal of this paper was to create a general video descriptor but they didn’t compare their method against any other general video descriptors. They only compared their method to the other methods in each specific task. They also mentioned that previous studies found that 3x3 kernels on the spatial dimensions were optimal [5] and only decided to experiment with the temporal dimensions in their research. This reduced the architecture search space and saved some time but

some experiments with differing heights and widths would help solidify their design choice. It is not enough to just assume that 3x3 kernels will also work on videos. The extra dimension in a 3D CNN may cause the classifier to behave differently so it is important to test other possible sizes such as 5x5x5. These larger kernels might be more efficient at handling videos because it requires fewer convolutional operations on the input.

Furthermore they didn't discuss any of the weaknesses of their approach. With any classifier, it is important to understand which examples the algorithm misclassified. This can help search for further improvements or help other researchers understand the constraints of the approach. They also didn't elaborate on their evaluation methods so it was not clear how they came up with their accuracy scores. This inhibits other researchers from directly comparing their algorithms with this one. A confusion matrix would visualize their results well because it not only illustrates the accuracy, it also quantifies the amount of false positives. The paper was fairly short so there was room to include all of these details, especially when introducing a new technology this is very important.

REFERENCES

[1] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," 12 2014.

[2] I. Laptev and T. Lindeberg. Space-time interest points. InICCV,2003. 1, 2

[3] O. Boiman and M. Irani. Detecting irregularities in images and invideo.IJCV, 2007. 1, 2

[4] D. B. Kris M. Kitani, Brian D. Ziebart and M. Hebert. Activity forecasting. InECCV, 2012. 1

[5] K. Simonyan and A. Zisserman. Very deep convolutional networksfor large-scale image recognition. InICLR, 2015. 3, 4