Beyond Short Snippets: Deep Networks for Video Classification

Joe Yue-Hei Ng[1], Matthew Hausknecht[2], Sudheendra Vijayanarasimhan[3], Oriol Vinyals[3], Rajat Monga[3], George Toderici[3],
[1]University of Maryland, College Park. [2]University of Texas at Austin. [3]Google, Inc.

Convolutional Neural Networks have proven highly successful at static image recognition problems such as the MNIST, CIFAR, and ImageNet Large-Scale Visual Recognition Challenge [5, 7, 8]. By using a hierarchy of trainable filters and feature pooling operations, CNNs are capable of automatically learning complex features required for visual object recognition tasks achieving superior performance to hand-crafted features. Encouraged by these positive results several approaches have been proposed recently to apply CNNs to video and action classification tasks [1, 3, 4, 6].

Video analysis provides more information to the recognition task by adding a temporal component through which motion and other information can be additionally used. At the same time, the task is much more computationally demanding even for processing short video clips since each video might contain hundreds to thousands of frames, not all of which are useful. A naïve approach would be to treat video frames as still images and apply CNNs to recognize each frame and average the predictions at the video level. However, since each individual video frame forms only a small part of the video's story, such an approach would be using incomplete information and could therefore easily confuse classes especially if there are fine-grained distinctions or portions of the video irrelevant to the action of interest.

Therefore, we hypothesize that learning a global description of the video's temporal evolution is important for accurate video classification. This is challenging from a modeling perspective as we have to model variable length videos with a fixed number of parameters. We evaluate two approaches capable of meeting this requirement: feature-pooling and recurrent neural networks. The feature pooling networks independently process each frame using a CNN and then combine frame-level information using various pooling layers. The recurrent neural network architecture we employ is derived from Long Short Term Memory (LSTM) [2] units, and uses memory cells to store, modify, and access internal state, allowing it to discover long-range temporal relationships. Like feature-pooling, LSTM networks operate on frame-level CNN activations, and can learn how to integrate information over time. By sharing parameters through time, both architectures are able to maintain a constant number of parameters while capturing a global description of the video's temporal evolution.

Since we are addressing the problem of video classification, it is natural to attempt to take advantage of motion information in order to have a better performing network. Previous work [4] has attempted to address this issue by using frame stacks as input. However, this type of approach is computationally intensive since it involves thousands of 3D convolutional filters applied over the input volumes. The performance grained by applying such a method is below 2% on the Sports-1M benchmarks [4]. As a result, in this work, we avoid implicit motion feature computation.

In order to learn a global description of the video while maintaining a low computational footprint, we propose processing only one frame per second. At this frame rate, implicit motion information is lost. To compensate, following [6] we incorporate explicit motion information in the form of optical flow images computed over adjacent frames. Thus optical flow allows us to retain the benefits of motion information (typically achieved through high-fps sampling) while still capturing global video information. Our contributions can be summarized as follows:

1. We propose CNN architectures for obtaining global video-level descriptors and demonstrate that using increasing numbers of frames significantly improves classification performance.
2. By sharing parameters through time, the number of parameters remains constant as a function of video length in both the feature pooling and LSTM architectures.
3. We confirm that optical flow images can greatly benefit video classification and present results showing that even if the optical flow images themselves are very noisy (as is the case with the Sports-1M dataset), they can still provide a benefit when coupled with LSTMs.

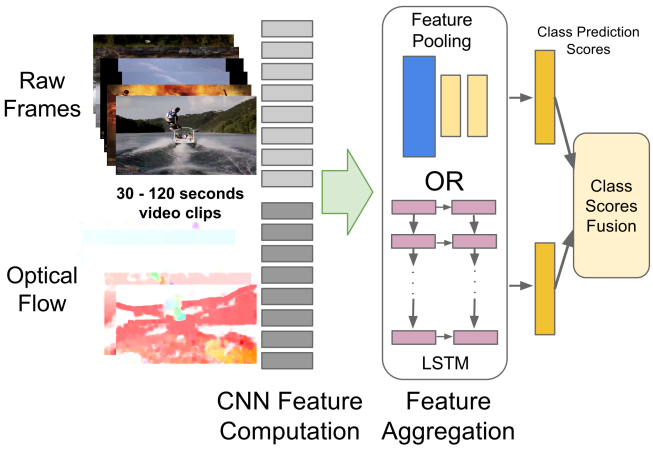

Figure 1: Overview of our approach.

Leveraging these three principles, our best networks exhibit significant performance improvements over previously published results on the Sports 1 million dataset (73.1% vs. 60.9%) and the UCF-101 datasets with (88.6% vs. 88.0%) and without additional optical flow information (82.6% vs. 73.0%).

[1] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential Deep Learning for Human Action Recognition. In *2nd International Workshop on Human Behavior Understanding (HBU)*, pages 29–39, November 2011.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computing*, 9(8):1735–1780, November 1997. ISSN 0899-7667.

[3] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *IEEE Trans. PAMI*, 35(1):221–231, January 2013. ISSN 0162-8828.

[4] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. CVPR*, pages 1725–1732, Columbus, Ohio, USA, 2014.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, pages 1097–1105, Lake Tahoe, Nevada, USA, 2012.

[6] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proc. NIPS*, pages 568–576, Montreal, Canada, 2014.

[7] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[8] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proc. ECCV*, pages 818–833, Zurich, Switzerland, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.