Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

Cheng Guan

August 24, 2018

Abstract

The paucity of videos in current action classification datasets (UCF-101 and HMDB-51) has made it difficult to identify good video architectures, as most methods obtain similar performance on existing small-scale benchmarks. This paper re-evaluates state-of-the-art architectures in light of the new Kinetics Human Action Video dataset. Kinetics has two orders of magnitude more data, with 400 human action classes and over 400 clips per class, and is collected from realistic, challenging YouTube videos.

The authors also introduce a new Two-Stream Inflated 3D ConvNet (I3D) that is based on 2D ConvNet inflation: filters and pooling kernels of very deep image classification ConvNets are expanded into 3D, making it possible to learn seamless spatio-temporal feature extractors from video while leveraging successful ImageNet architecture designs and even their parameters.

1. Introduction

One of the unexpected benefits of the ImageNet challenge has been the discovery that deep architectures trained on the 1000 images of 1000 categories, can be used for other tasks and in other domains. One of the early examples of this was using the fc7 features from a network trained on ImageNet for the PASCAL VOC classification and detection challenge [2, 5]. Furthermore, improvements in the deep architecture, changing from AlexNet to VGG-16, immediately fed through to commensurate improvements in the PASCAL VOC performance [6]. Since then, there have been numerous examples of ImageNet trained architectures warm starting or sufficing entirely for other tasks, e.g. segmentation, depth prediction, pose estimation, action classification.

In this paper The authors aim to provide an answer to this question using the new Kinetics Human Action Video Dataset [4], which is two orders of magnitude larger than previous datasets, HMDB-51 and UCF-101 . Kinetics has 400 human action classes with more than 400 examples for each class, each from a unique YouTube video.

2. Action Classification Architectures

A graphical overview of the five types of architectures they evaluate is shown in Fig.2

2.1. The Old I: ConvNet+LSTM

In theory, a more satisfying approach is to add a recurrent layer to the model [1], such as an LSTM, which can encode state, and capture temporal ordering and long range dependencies. The authors position an LSTM layer with batch normalization after the last average pooling layer of Inception-V1, with 512 hidden units. **A** fully connected layer is added on top for the classifier.

2.2. The Old II: 3D ConvNets

D ConvNets seem like a natural approach to video modeling, and are just like standard convolutional networks, but with spatio-temporal filters. They have been explored several times, previously [3, 8]. They have a very im- portant characteristic: they directly create hierarchical representations of spatio-temporal data. One issue with these models is that they have many more parameters than 2D ConvNets because of the additional kernel dimension, and this makes them harder to train.

2.3. The Old III: Two-Stream Networks

A different, very practical approach, introduced by Simonyan and Zisserman [7], models short temporal snapshots of videos by averaging the predictions from a single RGB frame and a stack of 10 externally computed optical flow frames, after passing them through two replicas of an ImageNet pre-trained ConvNet. The flow stream has an adapted input convolutional layer with twice as many input channels as flow frames (because flow has two channels, horizontal and vertical), and at test time multiple snapshots are sampled from the video and the action prediction is averaged.

2.4. The New: Two-Stream Inflated 3D ConvNets

With this architecture, The authors show how 3D ConvNets can benefit from ImageNet 2D ConvNet designs and,

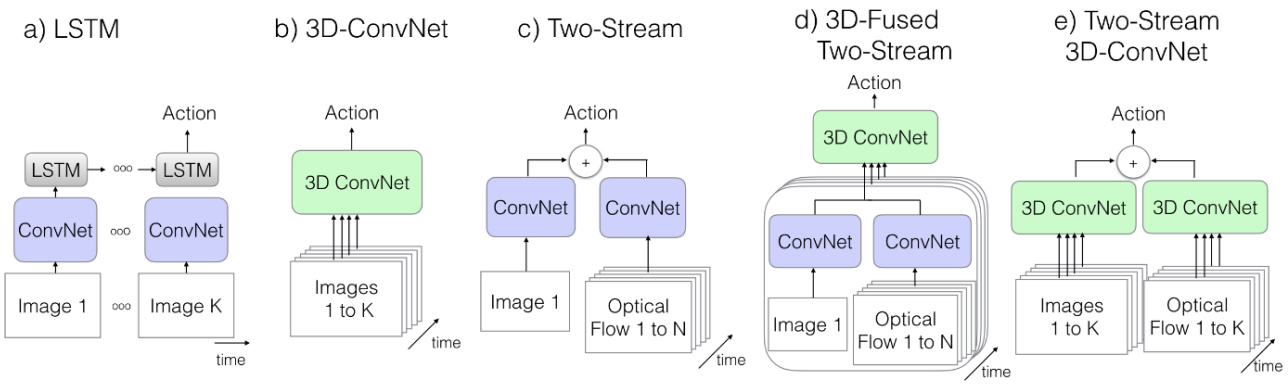

Figure 1. Video architectures considered in this paper. **K** stands for the total number of frames in a video, whereas **N** stands for a subset of neighboring frames of the video.

optionally, from their learned parameters. They also adopt a two-stream configuration here. While 3D ConvNets can directly learn about temporal patterns from an RGB stream, their performance can still be greatly improved by including an optical-flow stream.

Inflating 2D ConvNets into 3D. A number of very successful image classification architectures have been developed over the years, in part through painstaking trial and error. Instead of repeating the process for spatio-temporal models the authors propose to simply convert successful image (2D) classification models into 3D ConvNets. This can be done by starting with a 2D architecture, and inflating all the filters and pooling kernels endowing them with an additional temporal dimension.

Bootstrapping 3D filters from 2D Filters. Besides the architecture, one may also want to bootstrap parameters from the pre-trained ImageNet models. To do this, the authors observe that an image can be converted into a (boring) video by copying it repeatedly into a video sequence. The 3D models can then be implicitly pre-trained on ImageNet, by satisfying what they call the boring-video fixed point: the pooled activations on a boring video should be the same as on the original single-image input. This can be achieved, thanks to linearity, by repeating the weights of the 2D filters N times along the time dimension, and rescaling them by dividing by N.

Pacing receptive field growth in space, time and network depth. The boring video fixed-point leaves ample freedom on how to inflate pooling operators along the time dimension and on how to set convolutional/pooling temporal stride these are the primary factors that shape the size of feature receptive fields. Virtually all image models treat the two spatial dimensions (horizontal and vertical) equally pooling kernels and strides are the same.

Two 3D Streams. While a 3D ConvNet should be able to learn motion features from RGB inputs directly, it still performs pure feedforward computation, whereas optical flow algorithms are in some sense recurrent (e.g. they perform iterative optimization for the flow fields).

References

[1] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *ICCV*, 2015. 1

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[3] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *IEEE TPAMI*, 2013. 1

[4] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[5] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 1

[6] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[7] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1

[8] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *ECCV*, 2010. 1