3D Convolutional Neural Networks for Human Action Recognition

Liangjie Cao

Aug. 26, 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 bench- marks. This paper reevaluates 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. We provide an analysis on how current architectures fare on the task of action classification on this dataset and how much performance improves on the smaller benchmark datasets after pre-training on Kinetics. 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. We show that, after pre-training on Kinetics, I3D models considerably improve upon the state-of-the-art in action classification, reaching 80.9% on HMDB-51 and 98% on UCF-101.

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 [1]. Furthermore, improvements in the deep architecture, changing from AlexNet to VGG-16, immediately fed through to commensurate improvements in the PASCAL VOC performance [2]. 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 the video domain, it is an open question whether training an action classification network on a sufficiently large dataset, will give a similar boost in performance when applied to a different temporal task or dataset. The challenges of building video datasets has meant that most popular benchmarks for action recognition are small, having on the order of 10k videos.

2. Action Classification Architectures

While the development of image representation architectures has matured quickly in recent years, there is still no clear front running architecture for video. Some of the major differences in current video architectures are whether the convolutional and layers operators use 2D (image-based) or 3D (video-based) kernels; whether the input to the network is just an RGB video or it also includes pre-computed optical flow; and, in the case of 2D ConvNets, how information is propagated across frames, which can be done either using temporally-recurrent layers such as LSTMs, or feature aggregation over time.

2.1. The Old I: ConvNet+LSTM

The model is trained using cross-entropy losses on the outputs at all time steps. During testing we consider only the output on the last frame. Input video frames are subsampled by keeping one out of every 5, from an original 25 frames-per-second stream.

2.2. The Old II: 3D ConvNets

3D 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. 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. Also, they seem to preclude the benefits of ImageNet pre-training, and consequently previous work has defined relatively shallow custom archi- tectures

and trained them from scratch. Results on benchmarks have shown promise but have not been competitive with state-of-the-art, making this type of models a good candidate for evaluation on our larger dataset.

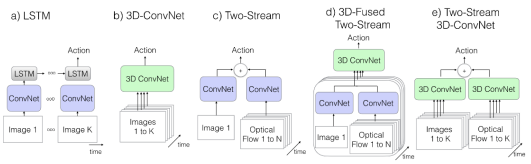

Figure 1. Video architectures considered in this paper. K stands for the total number of frames in a video, whereas N stands for a subset of neighboring frames of the video

The Old III: Two-Stream Networks LSTMs on features from the last layers of ConvNets can model high-level variation, but may not be able to capture fine low-level motion which is critical in many cases. It is also expensive to train as it requires unrolling the network through multiple frames for backpropagation-through-time.

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

With this architecture, they show how 3D ConvNets 1 can benefit from ImageNet 2D ConvNet designs and, option- ally, from their learned parameters. They also adopt a two- stream configuration here that 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.

References

[1] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007. 1

[2] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004. 1