Paper: *Learning Spatiotemporal Features with 3D Convolutional Networks*

Nirajan Koirala

CSC 9010_001

Summary

In this paper, authors propose to learn spatio-temporal features using deep 3D ConvNet to tackle the motion modeling which was not suitable using image based deep features. These type of 3D ConvNets were proposed before however, their work exploits 3D ConvNets in the context of large-scale supervised training datasets and modern deep architectures to achieve the best performance on different types of video analysis tasks. Their network C3D is generic, compact, simple and efficient for video learning tasks. In section 2, they discuss about some previously used methods like iDT and its shortcomings like highly computationally intensive and intractability on large datasets. In section 3, they explain in detail the basic operations of 3D ConvNets and analyze different architectures for 3D ConvNets empirically and further elaborate on the training methods for large scale datasets. 3D convolution is able to preserve the temporal information of the input signals resulting in an output volume. They find that using small receptive fields of 3x3 convolution kernels with deeper architecture yields best results after studying 2D ConvNet. Hence, they fix the spatial receptive field to 3x3 and vary only the temporal depth of the 3D convolutional kernel in their model's architecture. They provide the general notations and common network settings description further in the section. After varying the network's architecture and experimenting with homogeneous and varying temporal depth, they found that learning capacity of different models they've used are comparable and the differences in number of parameters do not affect the results of their architecture search. They provide the network's description like size of convolution filters and pooling layers in the section 3.2. First dataset they've used is Sports-1M dataset and it consists of 1.1 million sports videos containing 487 sports categories. C3D network is trained using SGD with minibatch size of 30 examples. They provide the classification results of their network where C3D network yields an accuracy of 84.4% outperforming DeepVideo's network.

In section 4, they evaluate C3D features on UCF101 dataset which consists of 12,320 videos of 101 human action categories. For iDT, they used the bag-of-work representaion with a codebook size of 5000 for each feature channel of iDT which are trajectories, HOG, HOF, MBHx and MBHy. They are able to achieve best performance using RGB only and RGB+ features. For the ASLAN dataset, they try to predict if a given pair of videos belong to the same or different action. This task they focus on is different from action recognition as the task focuses on predicting action similarity not the actual action label on a set of never seen before actions. Using the similar setup used in previous works, they are able to significantly outperform the state-of-art methods using C3D. For the scene and object recognition task, they evaluate C3D on two benchmarks: YPENN and Maryland. C3D is able to outperform both of the benchmarks by 10% and 1.9% respectively.

Hence, the C3D network learns by optimizing salient motions while optical flow encodes all movement. It is a high performance method which allows for real-time applications and the features are generalizable which allows transfer learning as well. But the input clips are too short for some applications and the resolution is also very low. 3D ConvNets are superior to 2D ConvNets for video tasks because they model appearance and motion simultaneously. C3D with a linear classifier can outperform or approach current best methods on video analysis benchmarks and while pairing them with SVM, they can perform very well in several video classification tasks.

Strengths

- The methods used are very practical and easy to follow.
- Various different types of datasets are used and their experiments are wide ranged.
- Good amount of figures and tables are provided.

Weaknesses

- The organization of figures and tables could have been made a little better.

Points of Confusion

- What is spatial and temporal jittering?
- Varying network architecture subsection of section 3.1 is a little confusing.

Discussion Questions

- How does the fully-connected layers alter performance?
- Would one be able to perform video summarizations using these methods?
- While combining C3D with iDT for performance boost, what can be observed for other type of descriptors or derived information?