T-CNN BASED CONTEXTUAL ACTION RECOGNITION

Venkatakiran S, Dr. S.Venkatanarayanan

Research Scholar, SSSUTMS, SEHORE, MP

Research Guide, SSSUTMS, SEHORE, MP

Abstract

Deep learning has made significant strides in solving some of the sub-problems, there are still many problems lacking satisfactory solutions, especially in real-world applications. Action and gesture recognition have been studied for a while within the fields of computer vision and pattern recognition and substantial progress has been reported for both tasks in the last two decades. Recently, deep learning has irrupted in these fields achieving outstanding results and outperforming "non-deep" state-of-the-art methods. We propose a T-CNN framework that can directly detect a 3D tube enclosing a moving object in a video segment by extending the faster R-CNN framework. A Tube CNN inside is proposed to predict the abjectness of each candidate tube and location parameters specifying the bounding tube.

Keyword: T-CNN, 3D-IoU, R-CNN

Introduction

Deep Learning is a field of Machine Learning specializing in statistical models called Deep Neural Networks. These models can learn complex hierarchical representations that correspond to multiple levels of abstractions. This is done through the use of multiple layers of nonlinear processing units, called neurons, to transform data, where each layer takes the previous layers as input. This creates a flow of information, from the input through the network to the output. The tube space is a very high-dimensional space. It is not possible and not necessary to consider every possible tube. Consider a short video segment with T frames (e.g.8 frames in an $FPS = 30$ video), an object's trajectory is usually very smooth and nearly linear in such a short time period. These quasi-linear tubelets may di↵er in size, direction or speed, however, they all live on a low dimensional manifold with limited degrees of freedom. Because of the quasi-linearity of objects' short trajectories, we use straight anchor tubes as the initial candidates. A naive way to construct a 3D tube is to have the same bounding box positions in all frames[1.2.3], which we call "stationary tubes" Ts , because they correspond to non-moving objects.

Tube Classification Module

Based on the shared feature maps *V*, *M* objectness scores are predicted in the classification module for *M* anchor tubes at each feature map location by a convolutional layer with kernel size 3 *x* 3 (acting on *K* = 256 feature maps). Essentially the objectness score for each anchor is determined from a 3 *x* 3 *x K* feature tensor. This module can be viewed as a fully convolutional network. During training, tube overlapping, i.e.3D intersection over union (3D-IoU) between the anchor tubes and ground truth tubes are computed. Anchor tubes with high 3D-IoU will be selected as positive proposals and assigned label +1, whereas anchor tubes with low 3D-IoU scores (partially overlapped) will be assigned label #1 and the remaining anchor tubes (including pure background) will be ignored. The classification module is trained with the cross-entropy classification loss *LclsT P N* with respect to their ground truth label.

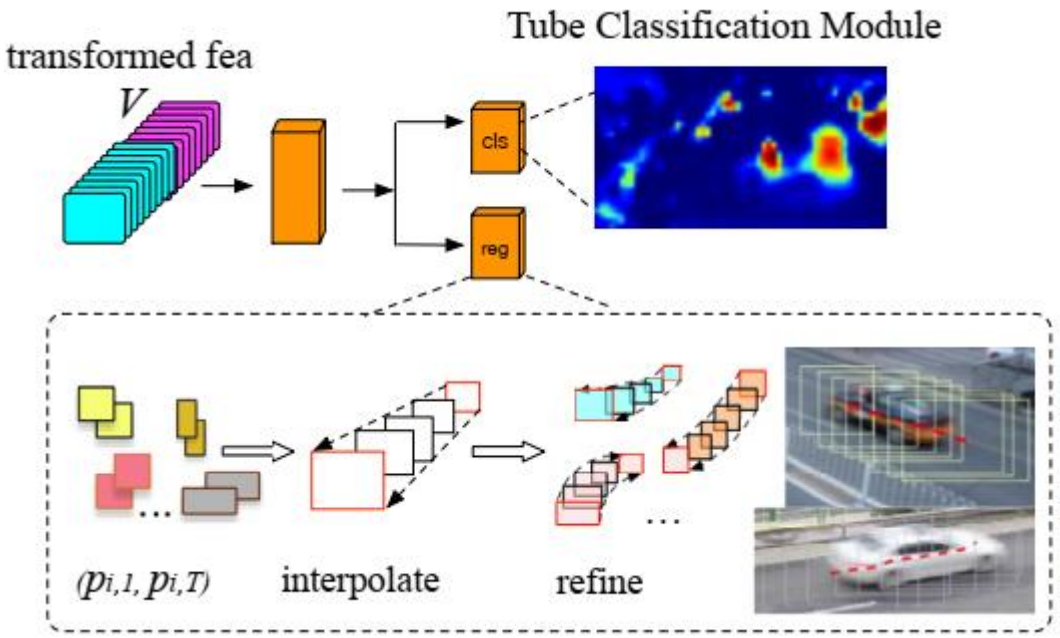

Figure 1. Tube Regression Module

Each tube proposal is given an objectness score predicted by the Tube Classification Module and the tube position offsets predicted by the Tube Offset Regression Module.

Tube Offset Regression Module

Anchor tubes will be ranked based on their objectness scores. For tubes with higher scores, the o↵sets between the corner positions of the tubes and ground truth positions are computed as the regression targets. The regression module will be trained to generate these regression targets from the input 3 *x* 3 *x K* features.

Given *M* candidate anchor tubes at each feature map location, the o↵sets will be predicted so as to "bend" the straight anchor tubes into a shape closer to the ground truth tubes. Following R-CNN, we use the center position and width and height to parameterize the position of a rectangular bounding box in each frame. The offsets of these parameters between the bounding boxes of all frames in an anchor tube (ST) and those in the ground truth tube (GT) are our 3D tube regression target for this anchor tube. We adopt the parameterization of the 4 coordinates in [6], but similar as [5], we normalize the spatial coordinate by the actual width and height of the video frame, so that the normalized coordinate and hence the 4 parameters are all in the range of [0, 1], which helps the convergence speed. The 3D Tube regression targets for positive anchor tube *i* at frame *t* is defined as:

$$tar_{i,t} = \begin{bmatrix} \Delta X_t^{gt} = \frac{(GT_{center\ x})_t - (ST_{center\ x})_t}{(ST_w)_t} \\ \Delta Y_t^{gt} = \frac{(GT_{center\ y})_t - (ST_{center\ y})_t}{(ST_h)_t} \\ \Delta W_t^{gt} = log\frac{(GT_w)_t}{(ST_w)_t} \\ \Delta H_t^{gt} = log\frac{(GT_h)_t}{(ST_h)_t} \end{bmatrix}$$

By learning to regress to these targets, the system can derive the refined locations for all anchor tubes that have high overlap with ground truth bounding tubes. We have explored two ways to wire the tube o↵set regression module: (1) directly predicting o↵sets of all frames and (2) utilizing linear interpolation.

The 3D-Convolutional Dual-Stream (3D-DS)

This approach is very similar to the Dual-Stream architecture. It uses one VGG- 16 CNN for still frames and a network for optical flow. The difference comes in the optical flow network. Taking inspiration from ST004, we proposed a 3Dconvolutional optical flow network in the hopes of capturing more temporal features when compared to the DSN. The rest of the architecture is exactly the same as the Dual-Stream and uses a bi-directional LSTM network with a softmax to produce the final scores for a sequence.

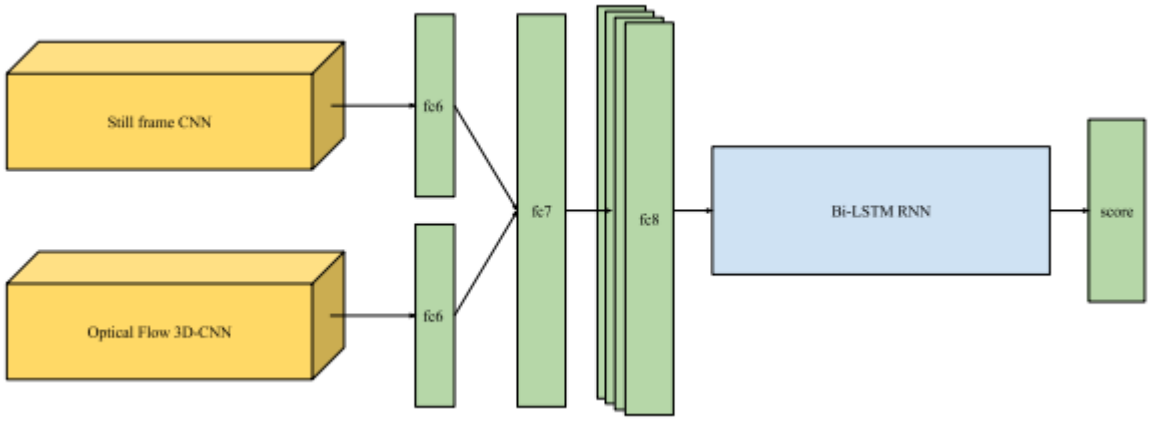

Figure 2. The 3D-Convolutional Dual-Stream Network

The Multi-Stream 3D-Convolutional Network (MS3D)

This architecture is heavily inspired by the architecture described in ST007. It consists of four CNNs, where two of them are taking still frames of video as input and two are using optical flow. The two different types of network both contain one network for handling full frames and one network for cropped action regions using bounding boxes around individual fish. This allows for motion from both individual fish and the shoal of fish to be captured explicitly. The Optical Flow networks are also inspired by ST004 and use 3D-CNNs to capture even more temporal information. The four networks are then fused using two fully connected layers before sequences of activations are fed into a bi-directional LSTM network which calculates the final score using a softmax.

Conclusion

The T-CNN performed very well in classifying images. This is usually good performance of the T-CNN requires a large number of training patterns this is usually true for pictures. Recently, a new set of action recognition data has been proposed. that is, UCF101, which contains 13320 videos. We would like to investigate T-CNN and the features extracted from these large action ID records. A complete video sequence in which covariance (brownian) and local spatiotemporal features (eg spatiotemporal) are presented. Points of interest or dense trajectories), since such a representation can be complementary Presentation / Function of the Fisher Vector Bag.

References

1. Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle and A de A Araujo. Bossa: Extended bow formalism for image classification. In Image Processing (ICIP), 2011 18th IEEE International Conference on, pages 2909–2912. IEEE, 2011
2. Slawomir Bak, Guillaume Charpiat, Etienne Corvee, Francois Bremond and Monique Thonnat. Learning to match appearances by correlations in a covariance metric space. In Computer Vision–ECCV 2012, pages 806–820. Springer Berlin Heidelberg, 2012.
3. Ratnesh Kumar, François Bremondet al. Brownian descriptor: a Rich Meta-Feature for Appearance Matching. In WACV: Winter Conference on Applications of Computer Vision, 2013.
4. Prithviraj Banerjee and Ram Nevatia. Learning neighborhood cooccurrence statistics of sparse features for human activity recognition. In Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on, pages 212–217. IEEE, 2011.
5. Herbert Bay, Tinne Tuytelaars and Luc Van Gool. Surf: Speeded up robust features. pages 404–417. Springer, 2006.
6. P. R. Beaudet. Rotationally invariant image operators. In Proceedings of the 4th International Joint Conference on Pattern Recognition, pages 579–583, Kyoto, Japan, November 1978.
7. Yassine Benabbas, Adel Lablack, Nacim Ihaddadene and Chabane Djeraba. Action recognition using direction models of motion. In Pattern Recognition (ICPR), 2010 20th International Conference on, pages 4295–4298.

8. Piotr Bilinski and François Bremond. Evaluation of local descriptors for action recognition in videos. In International Conference on Computer Vision Systems, Sophia Antipolis, France, September 2011.
9. Piotr Bilinski and Francois Bremond. Statistics of Pairwise Co-occurring Local Spatio-Temporal Features for Human Action Recognition. In Computer Vision–ECCV 2012. Workshops and Demonstrations, pages 311–320. Springer Berlin Heidelberg, 2012.
10. Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani and Ronen Basri. Actions as space-time shapes. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, volume 2, pages 1395–1402. IEEE, 2005.