Action Recognition

Large-scale Video Classification with Convolutional Neural Networks
Summary

Harshal Patel
2019pcp5048
Guided By:Dr. Neeta Nain

MNIT

January 27, 2020

Overview

1 Introduction

2 Related Work

3 Approach

4 Large-scale Video Classification with Convolutional Neural Networks

5 Empirical Evaluation

6 Some more related papers

7 References

Action recognition and why is it tough?

- Huge computational cost
- Capturing long context
- No standard benchmark dataset

Problem Definition

- Action recognition task involves the identification of different actions from video clips (a sequence of 2D frames) where the action may or may not be performed throughout the entire duration of the video.
- This seems like a natural extension of image classification tasks to multiple frames and then aggregating the predictions from each frame.

Related Works

- Traditional Approaches -
  1. Quantizing all features using a learned k-means dictionary
  2. 3D-CNN [2]

Approach: Time Information Fusion in CNNs I

1. Single Frame [10]
2. Late Fusion [10]
3. Early Fusion [10]
4. Slow Fusion [10]

1. **Late Fusion:**
   Places two separate single-frame networks with shared parameters a distance of 15 frames apart and then merges the two streams in fully connected layers

Approach: Time Information Fusion in CNNs II

2. **Early Fusion:**
   Combines information across an entire time window immediately on the pixel level. This is implemented by modifying the filters on the first convolutional layer in the single-frame model by extending them to be of size $11 \times 11 \times 3 \times T$ pixels (T=temporal extent) (similar like 3D-CNN)
3. **Slow Fusion:**
   balanced mix between the two approaches that slowly fuses temporal information throughout the network such that higher layers get access to progressively more global information in both spatial and temporal dimensions.

Multiresolution CNNs [1]

- Used to speed up model while retaining performance
  1. Reduce the number of layers and neurons in each layer but it lowers the performance
  2. Fovea and context streams. 9
     - aims to strike a compromise by having two separate streams of processing over two spatial resolutions
     - context stream receives the downsampled frames at half the original spatial resolution: 89 × 89 pixels
     - Fovea stream receives the center: 89 × 89 region

Notably, this design takes advantage of the camera bias present in many online videos, since the object of interest often occupies the center region.

Multiresolution CNN

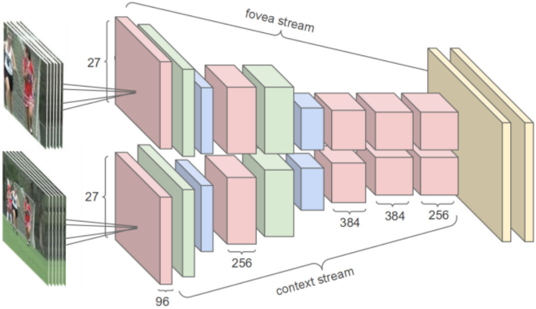

Figure: Multiresolution CNN architecture. Input frames are fed into two separate streams of processing: A context stream that models low-resolution image and a fovea stream that processes high-resolution center crop

Model Architecture

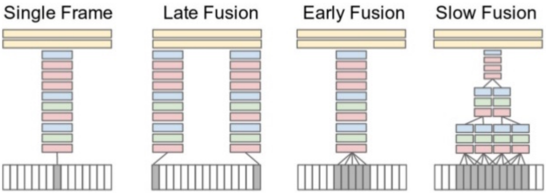

Figure: Approaches for fusing information over temporal dimension through the network. Red, green and blue boxes indicate convolutional, normalization and pooling layers respectively.

- The first layer takes as input an image of predefined size, and has size equal to number of pixels times 3 color channels.
- The last layer outputs the 2k joint coordinates.

Datasets

- Sports-1M dataset - 1million YouTube videos belonging to a taxonomy of 487 classes of sports.
- UCF-101 dataset - 13,320 videos and 101 generic classes

Results

- For slow fusion accuracy is 80.2 % on Sports-1M dataset for top 5 results
- 68.0% accuracy for top 3 result on UCF-101 dataset with transfer learning

Other papers I

1. Two-Stream Convolutional Networks for Action Recognition in Videos [3]
   - Jointly captures spatial relations between part locations and co-occurrence relations between part mixtures, augmenting standard pictorial structure models that encode spatial relations.
2. Long-term Recurrent Convolutional Networks for Visual Recognition and Description [4]
   - Uses LSTM to remember long term temporal information and uses extention of encoder-decoder for video representation , it uses CNN as encoder and LSTM as decoder. The architecture is trained end-to-end with input as RGB or optical flow of 16 frame clips. Final prediction for each clip is the average of predictions across each time step, it gives best result for weighted score of optical flow and RGB inputs.

Other papers II

3. Describing Videos by Exploiting Temporal Structure [5]
   - Introduced a temporal attention mechanism to exploit global temporal structure. Also augments the appearance features with action features that encode local temporal structure. they derived action features from 3-D CNN. Temporal attention mechanism focuses on a small subset of frames while generating description. It uses 3D-CNN-RNN architecture.
4. Show attend and tell:neural image generation with visual attention (for image description) [6]

Other papers III

- Introduces an attention based model that automatically learns to describe the content of images. It describes how we can train this model in a deterministic manner using standard back-propagation techniques and stochastics. Rather than compress an entire image into a static representation, attention allows for important features to dynamically come to the forefront as needed. This is especially useful when there is lot of irrelevant things in image. As the model generates each word,its attention changes to reflect the relevant parts of the image. We can investigate models that can attend to important part of an image while generating its caption.

References I

- Andrej Karpathy and George Toderici and Sanketh Shetty and Thomas Leung and Rahul Sukthankar and Li Fei-Fei *Large-scale Video Classification with Convolutional Neural Networks*. In CVPR, 2014,

- Shuiwang Ji, Wei Xu, Ming Yang, Member, IEEE, and Kai Yu, *3D Convolutional Neural Networks for Human Action Recognition*. in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 35, NO. 1, JANUARY 2013

- Karen Simonyan Andrew Zisserman, *Two-Stream Convolutional Networks for Action Recognition in Videos*. 12 Nov 2014.

References II

- Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell, *Long-term Recurrent Convolutional Networks for Visual Recognition and Descriptions*. May 2016.
- Li Yao et al. *Describing Videos by Exploiting Temporal Structure*.University of montreal, Oct 2015.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*,Apr 2016 .