Better Exploiting OS-CNN for Better Cultural Event Recognition in Still Images

September 15, 2015

1 Team details

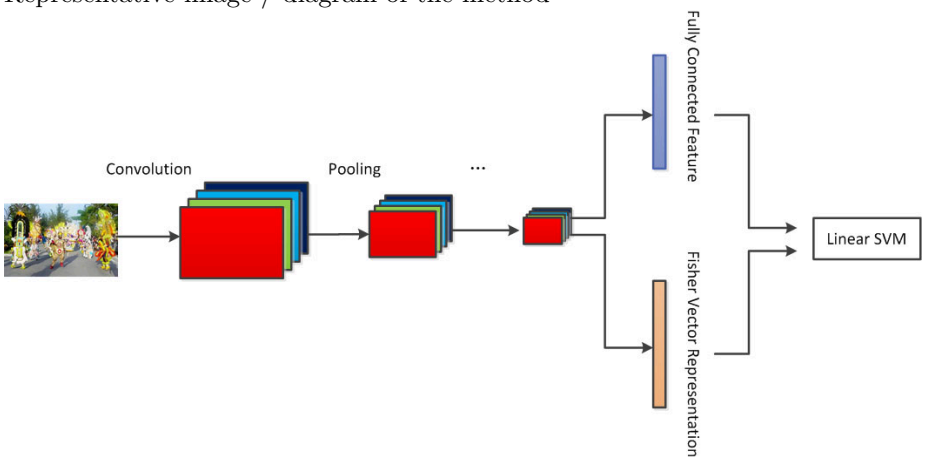
- **Team Name**: MMLAB
- **Team Leader**: Limin Wang
- **Address**: Multimedia Lab, Shenzhen Institutes of Advanced Technology, Shenzhen University Town, Shenzhen, P.R.China.
- **Phone**: 0018613144825084.
- **E-mail**: 07wanglimin@gmail.com
- **Team Members**: Zhe Wang, Sheng Guo, Yu Qiao
- Team website URL (if any)
- **Affiliation**:The Chinese University of Hong Kong. Shenzhen Institute of Advanced Technology.

2 Contribution details

- **Title**: Better Exploiting OS-CNN for Better Cultural Event Recognition in Still Images
- **Final Score**: We achieve mAP: 83.5 on the validation dataset.
- **General Description**: We propose a deep architecture to perform event recognition by extracting visual information from the perspectives of object and scene. Specifically, our proposed OS-CNN is composed of object net and scene net. Based on OS-CNN, we present an effective image representation, by extracting the activations of fully connected layers and convolutional layers. We use average pooling to aggregate the activations of fully connected layers and Fisher vector to encode those of convolutional layers.

- **References**:
  1. L. Wang, Z. Wang, W. Du, and Y. Qiao, Object-Scene Convolutional Neural Networks for Event Recognition in Images, in ChaLearn Looking at People (LAP) workshop, CVPR, 2015.
  2. L. Wang, Y. Qiao, and X. Tang, Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors, in CVPR, 2015.

- Representative image / diagram of the method



3 Data Preprocessing

- Describe features used or data representation model (if any)
  We extract deep learning features with OS-CNN.

- Dimensionality reduction technique applied (if any)
  We use PCA to reduce the dimension of convolutional features.

- Segmentation strategy used (if any)
  There is no segmentation strategy.

- Other techniques/strategy used not included in previous items FOR DATA PREPROCESSING (if any)

4 Classification details

- Classifier or method used to train and validate your results (if any)
  We use SVM as our classifier.

- Large scale strategy (if any)
  We use parallel training of deep CNN to speed up the learning process of our model

- Compositional model used (scene context representation), i.e. pictorial structure (if any)
  There is no compositional model used in our model
- Other technique/strategy used not included in previous items FOR CLASSIFICATION (if any)

5 Global Method Description

- Total method complexity analysis: all stages
  The time of our method mainly depends on the extraction of CNN features. Totallly, it requires about 1s to process an image.
- Which pre-trained or external methods have been used (for any stage, if any)
  We use the pre-trained models on the datasets of ImageNet and Places205 (GoogLeNet and VGGNet)
- Qualitative advantages of the proposed solution
  We treat object and scene as important cues for event recognition, and design a new architecture to exploit both cues for this challenge. Meanwhile, we consider the complementarity among the activations of different CNN layers.
- Results of the comparison to other approaches (if any)
- Novelty degree of the solution and if is has been previously published
  Partially of our method has been published in previous ChaLearn LAP workshop.

6 Other details

- Language and implementation details (including platform, memory, parallelization requirements)
  Cuda, Matlab
- Human effort required for implementation, training and validation?
  It require little efforts to reproduce the results of our method.
- Training/testing expended time?
  Training time: about 2 days. Testing time: about 1 day.
- General comments and impressions of the challenge?