Two-Stream Convolutional Networks for Action Recognition in Videos

Karen Simonyan, Andrew Zisserman
{karen, az}@robots.ox.ac.uk
Visual Geometry Group, University of Oxford, UK

1. OVERVIEW

Motivation
- Deep Convolutional Networks (ConvNets) work very well for image recognition
- It is less clear what is the right deep architecture for video recognition

Main contribution
Two-stream architecture for video classification
- Temporal stream – motion recognition ConvNet
- Spatial stream – appearance recognition ConvNet

2. TWO-STREAM ARCHITECTURE

Predicts action from still images – image classification
Input
- Individual RGB frames
Training
- Leverages large amounts of outside image data by pre-training on ILSVRC (1.2M images, 1000 classes)
- Classification layer re-trained on video frames
Evaluation
- Applied to 25 evenly sampled frames in each clip
- Resulting scores averaged

5. OPTICAL FLOW

- Displacement vector field between a pair of consecutive frames
- Each flow – 2 channels: horizontal & vertical components
- Computed using [Brox et al., ECCV 2004]
- Based on generic assumptions of constancy and smoothness
- Pre-computed on GPU (17fps), JPEG-compressed
- Global (camera) motion compensated by mean flow subtraction

4. SPATIAL STREAM

3. CONVNET LAYER CONFIGURATION

- Video decomposed into spatial & temporal components: still frames & optical flow
- Separate recognition stream for each component
- Streams combined by late fusion of soft-max scores (averaging or linear SVM)
- Most previous approaches: stack frames into a 3-D input volume

6. TEMPORAL STREAM

Multi-task learning to reduce over-fitting
- Video datasets (UCF-101, HMDB-51) are small
- Merging datasets is problematic due to semantic overlap
- Multi-task learning: each dataset defines a separate task (loss)

7. EVALUATION

Video action classification datasets
- UCF-101 (101 class, 13K videos)
- HMDB-51 (51 class, 6.8K videos)

Comparison of individual streams (UCF-101, 1st split, %)

Comparison with the state of the art (mean accuracy over 3 splits, %)

LEARNED FIRST-LAYER CONVOLUTIONAL FILTERS

- Spatial derivatives capture how motion changes in space (generalising hand-crafted features)
- Temporal derivatives capture how motion changes in time

This work was supported by ERC grant VisRec no. 228180 and hardware donation by NVIDIA