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

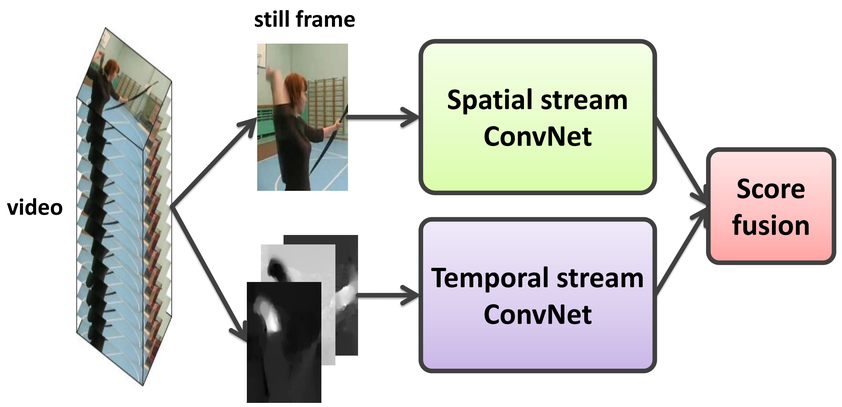
- Deep Convolutional Networks (ConvNets) work very well for image recognition
- It is less clear what is the right deep architecture for *video* recognition

Main contribution

Two-stream architecture for video classification

- Temporal stream – motion recognition ConvNet
- Spatial stream – appearance recognition ConvNet

2. TWO-STREAM ARCHITECTURE



- 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

3. CONVNET LAYER CONFIGURATION

conv1	conv2	conv3	conv4	conv5	full6	full7	full8
7x7x96 stride 2 norm. pool 2x2	5x5x256 stride 2 pool 2x2	3x3x512	3x3x512	3x3x512 pool 2x2	4096 dropout	2048 dropout	softmax

- Used for both spatial & temporal streams
- Similar to [Zeiler & Fergus, arXiv '13] (13.5% top-5 error on ILSVRC)
- 8 weight layers (5 conv. and 3 fully-connected)
- Input crop size: 224x224
- Also used in [Chatfield et al., BMVC '14] (denoted as CNN-M-2048)

4. SPATIAL STREAM

Predicts action from still images – image classification

Input

- Individual RGB frames

Training

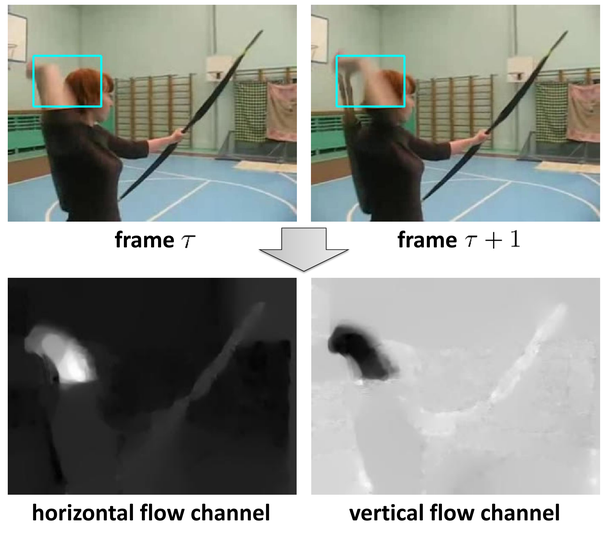
- Leverages large amounts of outside *image* data by pre-training on ILSVRC (1.2M images, 1000 classes)
- Classification layer re-trained on video frames

Evaluation

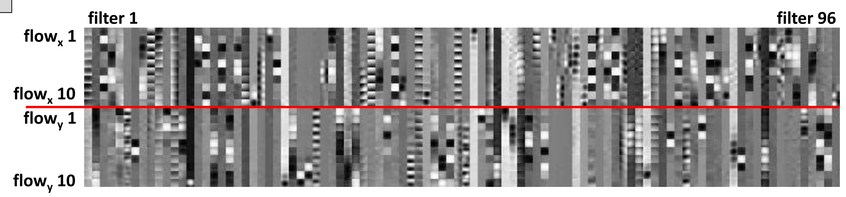
- Applied to 25 evenly sampled frames in each clip
- Resulting scores averaged

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



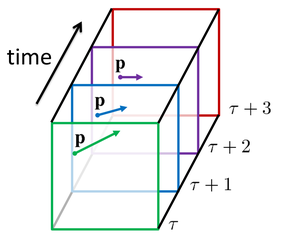
LEARNT FIRST-LAYER CONVOLUTIONAL FILTERS



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

6. TEMPORAL STREAM

Predicts action from motion



Input

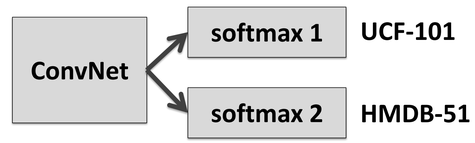
- Explicitly describes motion in video
- Stacked optical flow over several frames

Training

- From scratch with high drop-out (90%)

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)



Evaluation

- Applied to 25 evenly sampled 11-frame fragments in each clip
- Resulting scores averaged

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, %)

Model	UCF-101
Spatial stream, trained from scratch	52.3
Multi-frame input, trained from scratch [Karpathy et al., CVPR '14] , our impl.	56.4
Spatial stream, pre-trained on ILSVRC	72.7
Temporal stream, L=1 stacked flows as input	73.9
Temporal stream, L=5 stacked flows as input	80.4
Temporal stream, L=10 stacked flows as input	81.0

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

Model	UCF-101	HMDB-51
Spatial Stream ConvNet	73.0	40.5
Temporal Stream ConvNet	83.7	54.6
Two-stream ConvNet (SVM fusion)	**88.0**	**59.4**
Spatio-temporal HMAX [Kuehne et al., ICCV '11]	-	22.8
Spatio-temporal ConvNet [Karpathy et al., CVPR '14]	65.4	-
Two-stream ConvNet & LSTM (split 1) [Donahue et al., arXiv '14]	82.9	-
Hand-crafted feat. & Fisher vector [Wang and Schmid, ICCV '13]	85.9	57.2
Hand-crafted feat. & higher-dim encoding [Peng et al., arXiv '14]	**87.9**	61.1
Hand-crafted feat. & deep Fisher encoding [Peng et al., ECCV '14]	-	**66.8**

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