Large-Scale Video Classification with Convolutional Neural Networks [Karpathy et al]

CNN Model

- Treat every video as a bag of short, fixed sized clips
- **Time Information Fusion**
    - To capture temporal information in video, convolutional filters are extended over time.
        - **Single-frame**: no fusion; used as a baseline
        - **Early frame:** entire time window immediately on the pixel level
        - **Late Fusion:** combination of two separate single frame networks with shared parameters
        - **Slow Fusion:** Balanced mix of two approaches. Extend connectivity of all convolutional layers in time and carry out temporal convolutions in addition to spatial convolutions to compute activation functions. Higher layers progressively get access to more global information.

- **Multiresolution CNN**
    - Aim: reduce time complexity
    - Two separate streams of processing over two spatial resolutions:
        - **Context stream:** downsampled frames at half the spatial resolution
        - **Fovea stream:** central region of the video at original resolution

Learning

- Downpour stochastic gradient descent (minibatches of size 32, momentum 0.9,weight decay 0.0005, learning rate initialized to 1e-3)
- Data augmentation and preprocessing to reduce overfitting (crop, resize and flip horizontally wit 50% probability)

Results and datasets

- CNN overfits on UCF-101
- Created a dataset of 1 million YouTube videos and use it for training
- Sports-1M : 63.9 % accuracy (weak labeling)
- Transfer leaning on UCF-101 : Best results 65.4% on fine tuning top 3 layers (train from scratch, fine tune top layer, fine tune top 3 layers, fine tune all layers)
- Best results on sports category

Two Stream Convolutional Networks for Action Recognition in Videos [Simonyan, Zisserman]

CNN Model

- Divide into two streams:
    - Spatial
    - Temporal (optical flow)

- Each stream consists of 5 convolutional, 2 fully connected (with dropout) layers and softmax classifier
- Layers interleaved with rectification (RELU) non-linearities, max-pooling and response normalization
- Given a test video, a fixed number of frames are sampled with equal temporal spacing (25 frames) between them and each class is scored by averaging softmax probrability over sampled frames
- Temporal Convnet uses the following stacking to capture motion information:
    - Optical Flow stacking
    - Trajectory stacking
- Bidirectional optical flow and mean-flow subtraction (crude camera motion compensation)

Training

- Mini batch stochastic gradient descent
- 80K iterations
- Mini-batch size 256, momentum 0.9, learning rate initialized to 10^-2 and decreased uniformly, reinitialized to 10^-3 on 50K iterations, 10^-4 on 70K iterations.
- For fine tuning learning rate is set to 10^-3 after 14K iterations and total of 20K iterations
- Pretraining on ImageNet ILSVRC 2012

Results and datasets

- Overfits UCF-101 and HMDB-51, so multi task learning is employed
- Two softmax layers on top of last fully connected layer (for HMDB-51 and UCF-101), separate loss functions operating on HMDB-51 and UCF-101
- Two stream approach has accuracy of 87.6% on UCF-101 and 57.9% on HMDB-51 matches state-of-the-art