Beyond Short Clips: End-to-End Video-level Learning with Collaborative Memories

Xiuting Yang, Haoqi Fan, Lorenzo Torresani, Larry Davis, Heng Wang

University of Maryland, College Park, Facebook AI, Dartmouth

Motivation
- The standard way of optimizing 3D video models is clip-level training
  - A single short clip is sampled from the full-length video at each iteration
  - The clip-based prediction is optimized w.r.t. the video-level action label
- Limitation of clip-level training
  - Not possible to capture long-range temporal dependencies beyond short clips
  - Video-level label may not be well represented in a brief clip

Coping with GPU Memory Constraint

Batch reduction
- Reduce the batch size $B$ by a factor of $N$: $\hat{B} = \text{round}(\frac{B}{N})$

Multi-iteration
- Unroll the training of $N$ clips into $N$ consecutive iterations

End-to-end Video-level Learning Framework

Our idea: optimize the clip-based model using video-level information collected from the whole video

Multi-clip sampling
- Ensure sufficient temporal coverage of the video

Collaborative memory
- Model dependencies beyond short clips

Video-level supervision
- Joint optimization with a video-level supervision

Experimental
- Video-level learning (with $N > 1$) significantly improves video-level accuracy (2 ~ 3%) and clip-level accuracy
- Our framework generalizes to different backbone architectures and input configurations

Both collaborative memory and end-to-end training contribute to the performance gain
- Our associate memory design can capture cross-clip interaction, while feature gating can prevent over-fitting
- Our approach achieves state-of-the-art results on both action recognition and detection benchmarks