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

Xitong Yang[1], Haoqi Fan[2], Lorenzo Torresani[2,3], Larry Davis[1], Heng Wang[2]

[1]University of Maryland, College Park [2]Facebook AI [3]Dartmouth

CVPR VIRTUAL JUNE 19-25

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

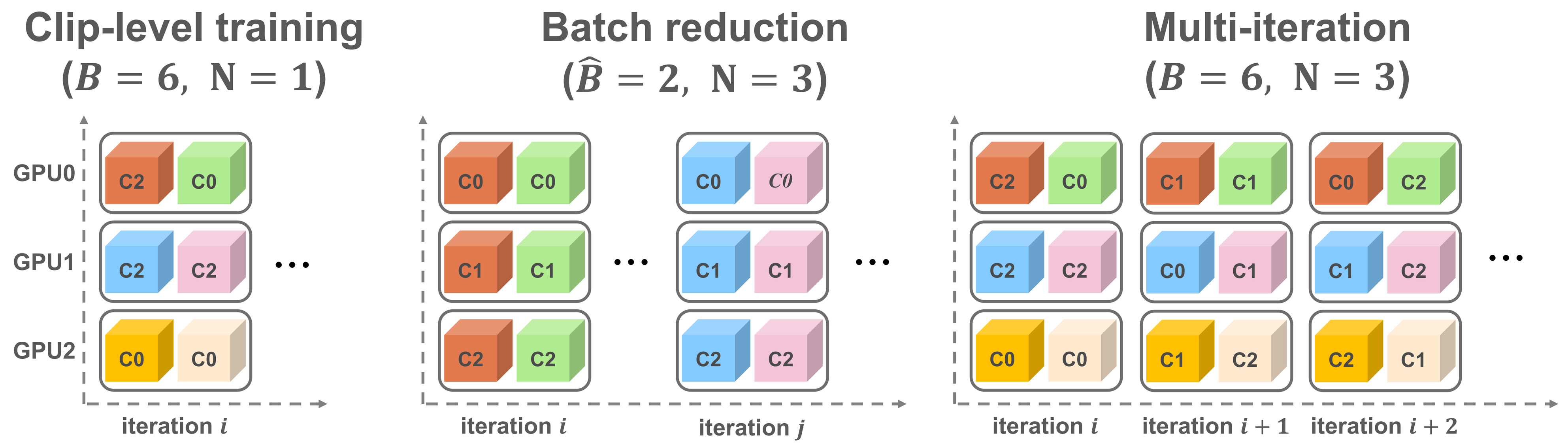
Coping with GPU Memory Constraint

Batch reduction

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

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

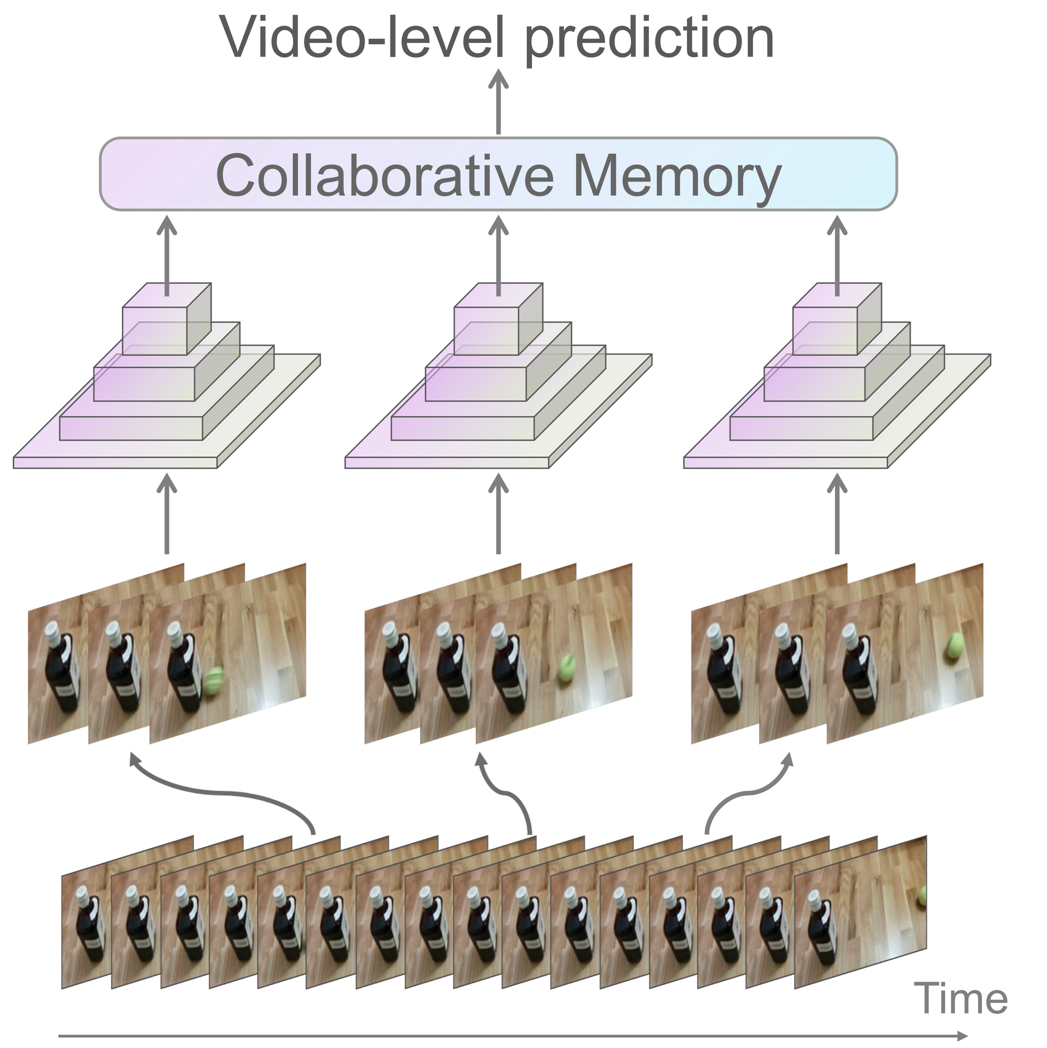
- Ensure sufficient temporal coverage of the video

Collaborative memory

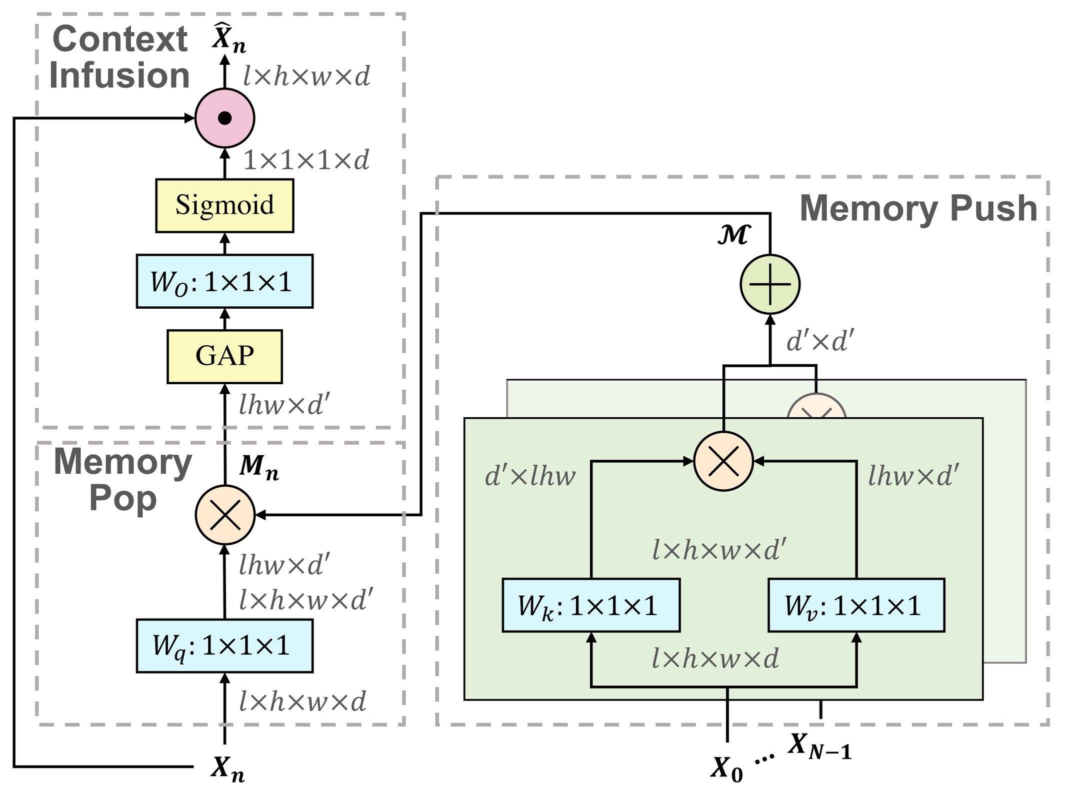
- Model dependencies beyond short clips

Video-level supervision

- Joint optimization with a video-level supervision



Collaborative Memory



Memory interaction

- Memory push

$$\mathcal{M} = Push(\{X_n\}_{n=0}^{N-1}) = \frac{1}{N}\sum_{n=0}^{N-1}(X_n W_k)^T (X_n W_v)$$

- Memory pop

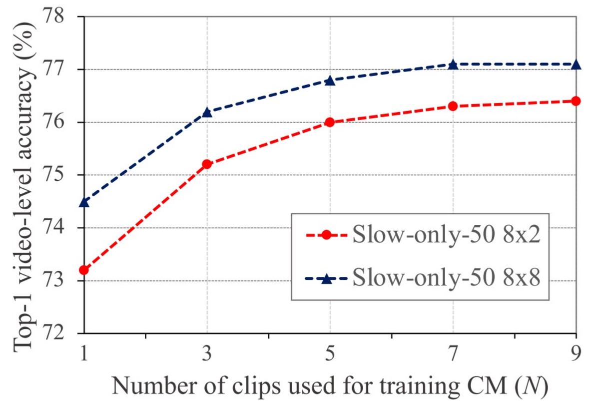
$$M_n = Pop(\mathcal{M}, X_n) = (X_n W_q)\mathcal{M}$$

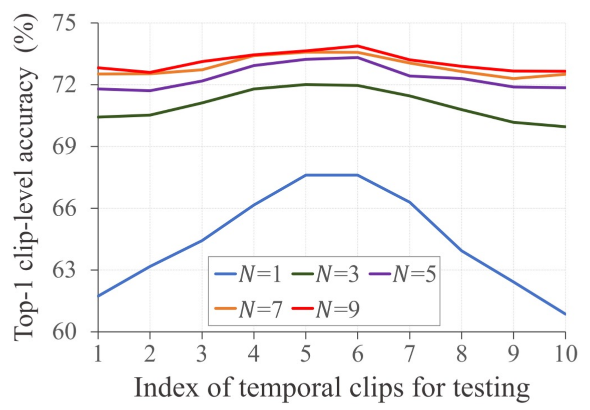
Context infusion

- Feature gating

$$\hat{X}_n = \sigma(Pool(M_n)W_O) \odot X_n + X_n$$

Experiments



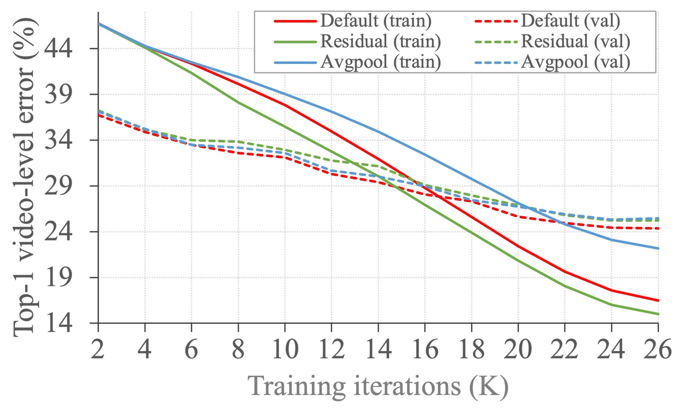


Model	Baseline	Ours	△	FLOPs
Slow-only-50 8×8	74.4	**76.8**	+2.4	1.03×
I3D-50+NL 32×2	74.9	**77.5**	+2.4	1.02×
R(2+1)D-50 16×2	75.7	**78.0**	+2.3	1.01×
SlowFast-50 4×16	75.6	**77.8**	+2.2	1.02×
SlowFast-50 8×8	76.8	**78.9**	+2.1	1.03×

- 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

Multi-clip	Memory	End-to-end	Top-1
	✓	✓	74.5
✓		✓	75.5
✓	✓		75.9
✓	✓	✓	**76.8**

Setting	Associative	Gating	Top-1
w/o memory			75.5
CM (avgpool)		✓	75.8
CM (residual)	✓		76.0
CM (default)	✓	✓	**76.8**



- 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

Methods	Kinetics-400	Kinetics-700	Charades	SSV1
NL I3D + GCN	–	–	39.7	46.1
CorrNet-101	79.2	–	–	53.3
SlowFast-101 + NL*	79.1	70.2	41.3	51.2
Ours (SlowFast-101+NL)	**81.4**	**72.4**	**44.6**	**53.7**

Methods	Extra info.	AVA v2.2
AVSF-101	✓	28.6
AIA (SlowFast-101)	✓	32.3
SlowFast-101*		29.0
Ours (SlowFast-101)		31.6

- Our approach achieves **state-of-the-art** results on both action recognition and detection benchmarks