Dynamic Inference: A New Approach Toward Efficient Video Action Recognition

Wenhao Wu   Dongliang He   Xiao Tan   Shifeng Chen   Yi Yang   Shilei Wen
Task

**Action Recognition:** classify the short clip or untrimmed video into pre-defined class.
Task

Action Recognition: classify the short clip or untrimmed video into pre-defined class.

- More than simply recognizing objects
- Complex person-person interaction & people-object interactions
- Videos bring motions
Motivation

How to Improve the Computation Efficiency of Action Recognition in Videos?
Motivation

How to Improve the Computation Efficiency of Action Recognition in Videos?

The time to process one frame AND the number of processed frames.
Motivation

How to Improve the Computation Efficiency of Action Recognition in Videos?

The time to process one frame AND the number of processed frames.

Direction 1:

Lightweight Base Model

1. 2D Conv + Efficient Temporal Modeling
2. Decompose 3D Conv
3. Network architecture search
4. Others …
How to Improve the Computation Efficiency of Action Recognition in Videos?

The time to process one frame AND the number of processed frames.

**Direction 1:** Lightweight Base Model

1. 2D Conv + Efficient Temporal Modeling
2. Decompose 3D Conv
3. Network architecture search
4. Others …

**Direction 2:** Adaptive Frame Sampler

- Hand-crafted sampler:
  - Uniform sampling,
  - Dense sampling
- Adaptive frame sampler:
  - Adaframe [1], MARL [2], SCSampler [3]

---

Motivation

How to Improve the Computation Efficiency of Action Recognition in Videos?

The time to process one frame AND the number of processed frames.

**Direction 1:** Lightweight Base Model

1. 2D Conv + Efficient Temporal Modeling
2. Decompose 3D Conv
3. Network architecture search
4. Others …

**Direction 2:** Adaptive Frame Sampler

- Hand-crafted sampler: Uniform sampling, Dense sampling
- Adaptive frame sampler: Adaframe [1], MARL [2], SCSampler [3]

**Direction 3:** Dynamic Network Route

- Image recognition:
  - MSDNet [4], SkipNet [5] …
- Video recognition:
  - We try to improve efficiency from dynamic inference viewpoint.

---

Observation

(a) Different “Writing” video instances
Observation

(a) Different “Writing” video instances

Irregular viewpoint

Need varying network capability
Observation

(a) Different “Writing” video instances

(b) “Running” vs. “Long Jump”

Irregular viewpoint

Need varying network capability
Observation

(a) Different “Writing” video instances

Irregular viewpoint

Need varying network capability

(b) “Running” vs. “Long Jump”

Different from “Writing”

Need varying number of frames
Observation

(a) Different “Writing” video instances
   Irregular viewpoint
   Need varying network capability

(b) “Running” vs. “Long Jump”
   Different from “Writing”
   Need varying number of frames

Videos differentiate from each other in terms of their distinguishability.
Approach

(a) Depth-axis dynamic scheme

features  block  sum  classifier
Approach

Top1 is Writing: $0.1 < \text{threshold}_1$

(a) Depth-axis dynamic scheme

features  block  sum  classifier
Approach

Top1 is Writing: $0.1 < \text{threshold}_1$

Top1 is Writing: $0.2 < \text{threshold}_2$

(a) Depth-axis dynamic scheme

features block sum classifier
Approach

Top1 is Writing: 0.1 < threshold_1
Top1 is Writing: 0.2 < threshold_2
Top1 is Writing: 0.5 > threshold_3

Make early prediction!

(a) Depth-axis dynamic scheme

features   block   sum   classifier
Approach

(a) Depth-axis dynamic scheme

(b) Input-axis dynamic scheme

features block sum classifier
Approach

(b) Input-axis dynamic scheme

(c) Input with permutation

features  block  sum  classifier
(a) Depth-axis dynamic scheme
Approach

(a) Depth-axis dynamic scheme

Online temporal shift module \cite{1}

Approach

(a) Depth-axis dynamic scheme

features  block  sum  classifier  shift

(d) Unified
Experiment

Scene-related Datasets

- Kinetics-400
  - 306,245 videos, 400 activity classes
- UCF-101
  - 13,320 videos, 101 classes
- HMDB-51
  - 6,766 videos, 51 classes

Temporal-related Datasets

- Something-Something
  - V1: 108,499 videos, 174 classes
  - V2: 220,847 videos, 174 classes

Evaluation Metrics

- Top-1 precision
- Average FLOPs/Video

FLOPs, which is short for float-point operations
Instantiation: MSDNet-38 \cite{1} & ResNet-50 \cite{2}

(a) Kinetics-400

(b) Something-Something v2


Experiment: Ablation Study

Ablation experimental results with MSDNet backbone on Kinetics-400. “Full Inference” means that, for each video, only the prediction head of the last checkpoint is used.
Experiment: Ablation Study

(a) W/o online temporal shift

FLOPs/Video ~ 53 %

(b) With online temporal shift

FLOPs/Video ~ 50 %
Experiment

(a) UCF-101

- MSDNet-38: Full Inference
- ResNet-50
- ResNet-101

FLOPs/Video \(\sim 70\%\)

(b) HMDB-51

- MSDNet-38: Full Inference
- ResNet-50
- ResNet-101

FLOPs/Video \(\sim 44\%\)
## Comparison with State-of-the-arts

### Kinetics-400

<table>
<thead>
<tr>
<th>Framework</th>
<th>Backbone</th>
<th>Input x # Clips</th>
<th>Prec@1</th>
<th># Params</th>
<th>FLOPs/Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>1SD [1]</td>
<td>3D BN-Inception</td>
<td>[All x 3 x 256 x 256] x 1</td>
<td>70.24</td>
<td>12.7M</td>
<td>544.4G</td>
</tr>
<tr>
<td>S3D [34]</td>
<td>3D BN-Inception</td>
<td>[All x 3 x 256 x 256] x 1</td>
<td>72.20</td>
<td>8.8M</td>
<td>518.6G</td>
</tr>
<tr>
<td>ARTNet with TSN [28]</td>
<td>3D ResNet-18</td>
<td>[16 x 3 x 112 x 112] x 250</td>
<td>69.22</td>
<td>35.2M</td>
<td>592.5G</td>
</tr>
<tr>
<td>MF [2]</td>
<td>-</td>
<td>[16 x 3 x 224 x 224] x 50</td>
<td>72.80</td>
<td>8.0M</td>
<td>555.4G</td>
</tr>
<tr>
<td>ECO [35]</td>
<td>BN-Inception+3D ResNet-18</td>
<td>[16 x 3 x 224 x 224] x 1</td>
<td>69.07</td>
<td>47.5M</td>
<td>64.6G</td>
</tr>
<tr>
<td>R(2+1)D RGB [27]</td>
<td>ResNet-34</td>
<td>[32 x 3 x 112 x 112] x 10</td>
<td>72.00</td>
<td>63.8M</td>
<td>1524G</td>
</tr>
<tr>
<td>Non-local-Ed [31]</td>
<td>ResNet-50</td>
<td>[128 x 3 x 112 x 112] x 10</td>
<td>67.30</td>
<td>35.3M</td>
<td>145.7G</td>
</tr>
<tr>
<td>TSN RGB [40]</td>
<td>BN-Inception</td>
<td>[25 x 3 x 112 x 112] x 10</td>
<td>69.17</td>
<td>10.7M</td>
<td>500.9G</td>
</tr>
<tr>
<td>TSN [20]</td>
<td>ResNet-50</td>
<td>[16 x 3 x 224 x 224] x 1</td>
<td>70.60</td>
<td>24.3M</td>
<td>33G</td>
</tr>
<tr>
<td>StNet [12]</td>
<td>ResNet-50</td>
<td>[16 x 3 x 224 x 224] x 1</td>
<td>72.50</td>
<td>24.3M</td>
<td>64G</td>
</tr>
<tr>
<td>Proposed</td>
<td>MSDNet-38 (Full)</td>
<td>[16 x 3 x 224 x 224] x 1</td>
<td>71.17</td>
<td>62.31M</td>
<td>53G</td>
</tr>
<tr>
<td></td>
<td>MSDNet-38</td>
<td>[11 x 3 x 224 x 224] x 1</td>
<td>71.20</td>
<td>62.31M</td>
<td>26.5G</td>
</tr>
<tr>
<td></td>
<td>ResNet-50</td>
<td>[11 x 3 x 224 x 224] x 1</td>
<td>72.57</td>
<td>29.12M</td>
<td>35G</td>
</tr>
<tr>
<td></td>
<td>ResNet-101</td>
<td>[11 x 3 x 224 x 224] x 1</td>
<td>74.70</td>
<td>48.12M</td>
<td>66G</td>
</tr>
</tbody>
</table>

### UCF-101 & HMDB-51

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>FLOPs</th>
<th>UCF-101</th>
<th>HMDB-51</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARTNet with TSN</td>
<td>3D ResNet-18</td>
<td>5925G</td>
<td>94.3</td>
<td>70.9</td>
</tr>
<tr>
<td>ECO</td>
<td>BNInception+3D</td>
<td>64G</td>
<td>92.8</td>
<td>68.5</td>
</tr>
<tr>
<td>1SD RGB</td>
<td>3D Inception-v1</td>
<td>544G</td>
<td>95.1</td>
<td>74.3</td>
</tr>
<tr>
<td>TSN RGB</td>
<td>BNInception</td>
<td>500G</td>
<td>91.1</td>
<td>-</td>
</tr>
<tr>
<td>TSN $S_{F}$</td>
<td>ResNet-50</td>
<td>33G</td>
<td>91.5</td>
<td>63.2</td>
</tr>
<tr>
<td>TSN $S_{F}$</td>
<td>ResNet-50</td>
<td>64G</td>
<td>91.4</td>
<td>63.6</td>
</tr>
<tr>
<td>TSN $S_{F}$</td>
<td>ResNet-50</td>
<td>33G</td>
<td>94.0</td>
<td>70.3</td>
</tr>
<tr>
<td>StNet</td>
<td>ResNet-50</td>
<td>64G</td>
<td>94.5</td>
<td>70.7</td>
</tr>
<tr>
<td>Proposed</td>
<td>MSDNet-38</td>
<td>53G</td>
<td>93.5</td>
<td>-</td>
</tr>
</tbody>
</table>

### Sth-Sth V1 & V2

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>FLOPs</th>
<th>UCF-101</th>
<th>HMDB-51</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECO$_{SF}$</td>
<td>Kinetics</td>
<td>64G</td>
<td>46.4</td>
<td>46.4</td>
</tr>
<tr>
<td>ECO$_{RP}$ Lite</td>
<td>3D ResNet-18</td>
<td>267G</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1SD</td>
<td>Kinetics</td>
<td>306G</td>
<td>41.6</td>
<td>72.2</td>
</tr>
<tr>
<td>Non-local-13D+GCN</td>
<td>Kinetics</td>
<td>606G</td>
<td>46.1</td>
<td>76.8</td>
</tr>
<tr>
<td>TSN$_{SF}$</td>
<td>Kinetics</td>
<td>33G</td>
<td>19.7</td>
<td>46.6</td>
</tr>
<tr>
<td>TSN$_{SF}$</td>
<td>65G</td>
<td>19.9</td>
<td>47.3</td>
<td>27.8</td>
</tr>
<tr>
<td>TRN Multiscalar</td>
<td>Kinetics</td>
<td>33G</td>
<td>34.4</td>
<td>-</td>
</tr>
<tr>
<td>TRN Two-Stream</td>
<td>ImageNet</td>
<td>42.0</td>
<td>55.5</td>
<td>30.0</td>
</tr>
<tr>
<td>TSM$_{SF}$</td>
<td>Kinetics</td>
<td>33G</td>
<td>43.4</td>
<td>73.2</td>
</tr>
<tr>
<td>TSM$_{SF}$</td>
<td>65G</td>
<td>44.8</td>
<td>74.5</td>
<td>58.2</td>
</tr>
<tr>
<td>Proposed</td>
<td>Kinetics</td>
<td>52.8(G1)/48.0(G2)</td>
<td>46.5</td>
<td>75.2</td>
</tr>
<tr>
<td></td>
<td>ImageNet</td>
<td>38.4(G1)/35.4(G2)</td>
<td>46.5</td>
<td>75.6</td>
</tr>
</tbody>
</table>
Visualization

Playing badminton

Bench pressing

(a) Two video instances which stop at the first checkpoint
(b) Two video instances which stop at the middle checkpoint
Visualization

Trimming trees

Tobogganing

(c) Two video instances which stop at the last checkpoint
Thank you!

Dynamic Inference: A New Approach Toward Efficient Video Action Recognition

Contact: Wenhao Wu

wuwenhao17@mails.ucas.edu.cn

http://whwu95.github.io