Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

João Carreira, Andrew Zisserman
Action recognition

Input: Video sequence

Output: Prediction of the action
Introduction

- Benefits of pre-training on ImageNet
  - Same task different data (classification -> classification)
  - Different task (classification -> segmentation/depth prediction etc.)
- In video domain, benefits of pre-training is an open question
Experimental strategy:

- Reimplement action classification DNNs from the literature
- Analyze their transfer behavior
- Introduce of a new model

Question: Is there a benefit in transfer learning using a large scale video dataset?
Background
The Old I

- **Feature extractor (2D ConvNet)**
  - Pool features from individual frame (Ignoring temporal structure) [15]
- **Add recurrence (LSTM) [5, 34]**
  - Encode state and temporal ordering

**Pros**
- Reuse of image classification networks

**Cons**
- Disjoint/late modeling of spatial and temporal information

![Diagram of LSTM network](image)
The Old II

- 3D ConvNet [14, 15, 28, 29]
  - Frames are stacked in 3rd dimension

**Pros**
- Directly create spatio-temporal representation

**Cons**
- Many more parameters than 2D ConvNets (Harder to train)
- Preclude benefit of ImageNet pre-training
The Old III

- Two stream network [8, 25]
  - RGB (2D ConvNet)
  - Optical Flow (2D ConvNet)

Pros
- Reuse of image classification networks

Cons
- Disjoint/late modeling of spatial and temporal information
The New

Inflating 2D ConvNets into 3D (This paper)

- Reusing structure of well studied 2D ConvNets
- Inflating an additional dimension to kernels
- Bootstrapping parameters from 2D filters

Pros
- Reuse of image classification networks
- Directly create spatio-temporal representation

Cons

Pros

Cons

Images 1 to K

Optical Flow 1 to K

Action

time
The New

- Bootstrapping parameters by satisfying the boring-video fixed point
  - Copy an image to convert it to a “boring-video”
  - The activation from boring-video should be the same as from the original image
  - Achieved by repeating 2D-filters N times and rescale by dividing by N

- Symmetric receptive field might not be optimal
  - Grow too quickly -> conflate edges from different objects
  - Grow too slow -> not capture the entire scene dynamics
**Architectures summary**

Common for all: conv+bn+relu,

<table>
<thead>
<tr>
<th></th>
<th>a) LSTM</th>
<th>b) 3D ConvNet</th>
<th>c) Two-Stream</th>
<th>d) 3D-fused Two-Stream</th>
<th>e) Two-Stream 3D-ConvNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet pretrained (Inception-V1)</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Resolution</td>
<td>224x224</td>
<td>112x112</td>
<td>224x224</td>
<td>224x224</td>
<td>224x224</td>
</tr>
<tr>
<td>Temporal resolution</td>
<td>Sample every fifth frame</td>
<td>16 consecutive frames</td>
<td>*</td>
<td>*</td>
<td>64 consecutive frames</td>
</tr>
</tbody>
</table>

* 5 consecutive frames 10 frames appart + corresponding optical flow frames
<table>
<thead>
<tr>
<th>Method</th>
<th>#Params</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td># Input Frames</td>
<td>Temporal Footprint</td>
</tr>
<tr>
<td>ConvNet+LSTM</td>
<td>9M</td>
<td>25 rgb</td>
<td>5s</td>
</tr>
<tr>
<td>3D-ConvNet</td>
<td>79M</td>
<td>16 rgb</td>
<td>0.64s</td>
</tr>
<tr>
<td>Two-Stream</td>
<td>12M</td>
<td>1 rgb, 10 flow</td>
<td>0.4s</td>
</tr>
<tr>
<td>3D-Fused</td>
<td>39M</td>
<td>5 rgb, 50 flow</td>
<td>2s</td>
</tr>
<tr>
<td>Two-Stream I3D</td>
<td>25M</td>
<td>64 rgb, 64 flow</td>
<td>2.56s</td>
</tr>
</tbody>
</table>

Table 1. Number of parameters and temporal input sizes of the models.
Kinetics dataset [16]

- Covers:
  - Person actions: drawing, drinking
  - Person-Person actions: hugging, kissing
  - Person-Object actions: opening presents, washing dishes
- 400 human action classes
- 400+ clips per class
- miniKinetics: subset of Kinetics:
  - 213 classes
  - Total of 120k clips
## Results 1

<table>
<thead>
<tr>
<th>Architecture</th>
<th>UCF-101</th>
<th></th>
<th>HMDB-51</th>
<th></th>
<th>miniKinetics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RGB</td>
<td>Flow</td>
<td>RGB + Flow</td>
<td>RGB</td>
<td>Flow</td>
<td>RGB + Flow</td>
</tr>
<tr>
<td>(a) LSTM</td>
<td>81.0</td>
<td>-</td>
<td>-</td>
<td>36.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(b) 3D-ConvNet</td>
<td>51.6</td>
<td>-</td>
<td>-</td>
<td>24.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(c) Two-Stream</td>
<td>83.6</td>
<td>85.6</td>
<td>91.2</td>
<td>43.2</td>
<td>56.3</td>
<td>58.3</td>
</tr>
<tr>
<td>(d) 3D-Fused</td>
<td>83.2</td>
<td>85.8</td>
<td>89.3</td>
<td>49.2</td>
<td>55.5</td>
<td>56.8</td>
</tr>
<tr>
<td>(e) Two-Stream I3D</td>
<td>84.5</td>
<td>90.6</td>
<td>93.4</td>
<td>49.8</td>
<td>61.9</td>
<td>66.4</td>
</tr>
</tbody>
</table>
## Results 2

<table>
<thead>
<tr>
<th>Architecture</th>
<th>UCF-101</th>
<th></th>
<th></th>
<th></th>
<th>HMDB-51</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>Fixed</td>
<td>Full-FT</td>
<td>Δ</td>
<td>Original</td>
<td>Fixed</td>
<td>Full-FT</td>
<td>Δ</td>
</tr>
<tr>
<td>(a) LSTM</td>
<td>81.0</td>
<td>81.6</td>
<td>82.1</td>
<td>-6%</td>
<td>36.0</td>
<td>46.6</td>
<td>46.4</td>
<td>-16.7%</td>
</tr>
<tr>
<td>(b) 3D-ConvNet</td>
<td>49.2</td>
<td>76.0</td>
<td>79.9</td>
<td>-60.5%</td>
<td>24.3</td>
<td>47.5</td>
<td>49.4</td>
<td>-33.1%</td>
</tr>
<tr>
<td>(c) Two-Stream</td>
<td>91.2</td>
<td>90.3</td>
<td>91.5</td>
<td>-3.4%</td>
<td>58.3</td>
<td>64.0</td>
<td>58.7</td>
<td>-13.7%</td>
</tr>
<tr>
<td>(d) 3D-Fused</td>
<td>89.3</td>
<td>88.5</td>
<td>90.1</td>
<td>-7.5%</td>
<td>56.8</td>
<td>59.0</td>
<td>61.4</td>
<td>-10.6%</td>
</tr>
<tr>
<td>(e) Two-Stream 13D</td>
<td><strong>93.4</strong></td>
<td><strong>95.7</strong></td>
<td><strong>96.5</strong></td>
<td>-47.0%</td>
<td><strong>66.4</strong></td>
<td><strong>74.3</strong></td>
<td><strong>75.9</strong></td>
<td>-28.3%</td>
</tr>
</tbody>
</table>
Summary

- Demonstrate power of transfer learning in video domain
- Introduce the idea of kernel inflation
- Novel architecture for action recognition
Inflated Inception-V1

Video → 7x7x7 Conv (stride 2) → 1x3x3 Max-Pool (stride 1,2,2) → 1x1x1 Conv → 3x3x3 Conv → 1x3x3 Max-Pool (stride 1,2,2) → Inc. → Rec. Field: 7,11,11

Inc. → Rec. Field: 23,75,75

Inc. → Inc. → Inc. → Inc. → 3x3x3 Max-Pool (stride 2) → Inc. → Rec. Field: 59,219,219

Inc. → 2x2x2 Max-Pool (stride 2) → Inc. → Inc. → 2x7x7 Avg-Pool → 1x1x1 Conv → Inc. → Rec. Field: 99,539,539

Predictions

Inception Module (Inc.)

Next Layer

Concatenation

Inc.

1x1x1 Conv

3x3x3 Conv

3x3x3 Conv

1x1x1 Conv

Previous Layer

3x3x3 Max-Pool
References


