End-to-end Learning of Action Detection from Frame Glimpse in Videos

BIL722 - Advanced Topics in Computer Vision

Ezgi Pekşen Soysal
Hacettepe University
Task: what is the person doing?

Input

Output

Running

Talking
Task: what is the person doing?

Input

Output

- Accuracy
- Efficiency
- Interpretability
Efficient video processing
Efficient video processing
Efficient video processing
Efficient video processing

“Knowing the output or the final state… there is no need to explicitly store many previous states”

Efficient video processing

“Knowing the output or the final state... there is no need to explicitly store many previous states”


Dominant paradigm: sliding windows

Used in all THUMOS challenge action detection entries
[OneVerSch 2014]
[WanQiaTan 2014]
[KarSeiBim 2014]
[YuaPeiNiMouKas 2015]
Efficient video processing

“Knowing the output or the final state… there is no need to explicitly store many previous states”

“Time may be represented in several ways… The intervals between ‘pulses’ need not be equal.”

Our model for efficient action detection

Frame model

Input: A frame

Output

Olga Russakovsky: The human side of computer vision

[YouRusMorFei CVPR'16]
Our model for efficient action detection

Input: A frame

Output: Detection instance [start, end]
Next frame to glimpse

Frame model

Olga Russakovsky: The human side of computer vision

[YouRusMorFei CVPR'16]
Our model for efficient action detection

Output: Detection instance [start, end] Next frame to glimpse

Frame model
Our model for efficient action detection

Frame model

Output:
Detection instance [start, end]
Next frame to glimpse
Our model for efficient action detection

Output:
Detection instance [start, end]
Next frame to glimpse

Recurrent neural network (time information)
Convolutional neural network (frame information)

[t = 0] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
[t = T]
Our model for efficient action detection

Output

Optional output:
Detection **instance** [start, end]
Output:
**Next** frame to glimpse

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

Olga Russakovsky: The human side of computer vision

[YeuRusMorFei CVPR'16]
Training the detection instance output

Training data

Positive video

Negative video
Training the detection instance output

Positive video

Negative video

Training data

Aside:

- effective video annotation
  - [YeuRusJinAndMorFei UnderReview]
  - [LiuRusDenBerFei ImageNetChallenge ‘15]

- weakly supervised detection
  - [YeuRamRusMorFei InPreparation]
Training the detection instance output

Positive video

Training data

<table>
<thead>
<tr>
<th>d₁</th>
<th>d₂</th>
<th>d₃</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g₁</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t₀</td>
<td></td>
<td>tₜ</td>
</tr>
</tbody>
</table>

Detection

<table>
<thead>
<tr>
<th>d₄</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>t₀</td>
</tr>
</tbody>
</table>

Negative video

<table>
<thead>
<tr>
<th>d₄</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>t₀</td>
</tr>
</tbody>
</table>
Training the detection instance output

Positive video

Training data

\[
\begin{array}{c}
\text{t = 0} \\
g_1 \\
\text{t = T}
\end{array}
\quad
\begin{array}{c}
\text{t = 0} \\
g_2 \\
\text{t = T}
\end{array}
\]

Detections

\[
\begin{array}{c}
\text{t = 0} \\
d_1 \\
\text{t = T}
\end{array}
\quad
\begin{array}{c}
\text{t = 0} \\
d_2 \\
\text{t = T}
\end{array}
\quad
\begin{array}{c}
\text{t = 0} \\
d_3 \\
\text{t = T}
\end{array}
\]

Negative video

\[
\begin{array}{c}
\text{t = 0} \\
d_4 \\
\text{t = T}
\end{array}
\]

Reward for detection

\[
\begin{array}{c}
\text{t = 0} \\
\text{t = T}
\end{array}
\]

\[
\begin{array}{c}
\text{t = 0} \\
\text{t = T}
\end{array}
\]

[YouRusMorFei CVPR’16]
Training the detection instance output

Positive video

Training data

\[ g_1 \]
\[ g_2 \]

\( t = 0 \)
\[ d_1 \]
\[ d_2 \]
\[ d_3 \]
\( t = T \)

\( y_1 = 1 \)
\( y_2 = 1 \)
\( y_3 = 2 \)

Detections

Negative video

\( t = 0 \)
\[ d_4 \]
\( t = T \)

\( y_4 = 0 \)

Reward for detection

\( \mathcal{L}(D, G) = \sum_{i} \mathcal{L}_{cls}(d_i, y_i > 0) \)

cross-entropy
classification loss

Olga Russakovsky: The human side of computer vision

[YeuRusMorFei CVPR’16]
Training the detection instance output

\[ \mathcal{L}(D, G) = \sum_i \mathcal{L}_{cls}(d_i, y_i > 0) + \gamma \sum_{i: y_i > 0} \mathcal{L}_{loc}(d_i, g_{y_i}) \]

- \( \mathcal{L}_{cls} \): cross-entropy classification loss
- \( \mathcal{L}_{loc} \): \( L^2 \) distance localization loss
Training the non-differentiable outputs

Training data

\[ t = 0 \quad [\quad ] \quad [\quad ] \quad t = T \]

Detections

\[ t = 0 \quad [\quad ] \quad [\quad ] \quad [\quad ] \quad [\quad ] \quad t = T \]
Training the non-differentiable outputs

(1) whether to predict a detection
(2) where to look next
Training the non-differentiable outputs

Train an policy $\pi_\theta$ for actions (1) and (2) using REINFORCE [Williams 1992]
Training the non-differentiable outputs

Training data

Detections

Model’s action sequence \( \mathbf{a} \)

Frame 1 \( \rightarrow \) Frame 8 \( \rightarrow \) Frame 6 \( \rightarrow \) Frame 15

(1) whether to predict a detection

(2) where to look next

Train an policy \( \pi_{\theta} \) for actions (1) and (2) using REINFORCE [Williams 1992]

Reward for an action sequence \( \mathbf{a} \) : \[ r(\mathbf{a}) = N^+ - \alpha N^- \]
Training the non-differentiable outputs

Training data

Detections

Model’s action sequence \(a\)

Frame 1 | Frame 8 | Frame 6 | Frame 15
------ | ------ | ------ | ------
\(t = 0\) | \(d_1\) | \(d_2\) | \(d_3\) | \(t = T\)
\(t = 0\) | \(\text{good}\) | \(\text{bad}\) | \(\text{bad}\) | \(t = T\)

(1) whether to predict a detection

(2) where to look next

goto frame 8
goto frame 6
goto frame 15

Train an policy \(\pi_\theta\) for actions (1) and (2) using REINFORCE [Williams 1992]

Reward for an action sequence \(a\) : \[ r(a) = N^+ - \alpha N^- \]

Objective: \[ J(\theta) = \sum a p_\theta(a) r(a) \]

Gradient: \[ \nabla J(\theta) = \sum a p_\theta(a) r(a) \nabla \log p_\theta(a) \]

Monte-Carlo approximation: \[ \nabla J(\theta) \approx \frac{1}{K} \sum_{k=1}^{K} r(a^k) \sum_{t=1}^{T} \nabla \log \pi_\theta(a^k_t | M^k_t) \]
## Accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Detection AP at IOU 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>State-of-the-art</td>
</tr>
<tr>
<td>THUMOS 2014</td>
<td>14.4</td>
</tr>
<tr>
<td>ActivityNet sports</td>
<td>33.2</td>
</tr>
<tr>
<td>ActivityNet work</td>
<td>31.1</td>
</tr>
</tbody>
</table>

## Interpretability
### Accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Detection AP at IOU 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>State-of-the-art</td>
</tr>
<tr>
<td>THUMOS 2014</td>
<td>14.4</td>
</tr>
<tr>
<td>ActivityNet sports</td>
<td>33.2</td>
</tr>
<tr>
<td>ActivityNet work</td>
<td>31.1</td>
</tr>
</tbody>
</table>

### Efficiency

Glimpse only 2% of video frames

### Interpretability
### Accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th>State-of-the-art</th>
<th>Our result</th>
</tr>
</thead>
<tbody>
<tr>
<td>THUMOS 2014</td>
<td>14.4</td>
<td><strong>17.1</strong></td>
</tr>
<tr>
<td>ActivityNet sports</td>
<td>33.2</td>
<td><strong>36.7</strong></td>
</tr>
<tr>
<td>ActivityNet work</td>
<td>31.1</td>
<td><strong>39.9</strong></td>
</tr>
</tbody>
</table>

### Efficiency

- **Glimpse only 2% of video frames**

<table>
<thead>
<tr>
<th>Samping</th>
<th>Detection AP at IOU 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>9.3</td>
</tr>
<tr>
<td>Our glimpses</td>
<td><strong>17.1</strong></td>
</tr>
</tbody>
</table>
### Accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th>State-of-the-art</th>
<th>Our result</th>
</tr>
</thead>
<tbody>
<tr>
<td>THUMOS 2014</td>
<td>14.4</td>
<td>17.1</td>
</tr>
<tr>
<td>ActivityNet sports</td>
<td>33.2</td>
<td>36.7</td>
</tr>
<tr>
<td>ActivityNet work</td>
<td>31.1</td>
<td>39.9</td>
</tr>
</tbody>
</table>

### Efficiency

Glimpse only 2% of video frames

<table>
<thead>
<tr>
<th>Sampling</th>
<th>Detection AP at IOU 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>9.3</td>
</tr>
<tr>
<td>Our glimpses</td>
<td>17.1</td>
</tr>
</tbody>
</table>

### Interpretability

- Ground truth
- Detections
- Glimpses
- Frames
The human cost of developing computer vision expertise

Open-world probabilistic human-in-the-loop model

Reinforcement learning for human action detection

[DenRusKraBerBerFei CHI’14]
[DenRusKraSatEtal IJCV’15]
[RusLiFei CVPR’15]
[YeuRusMorFei CVPR’16]
Takeaways

Data is critical

Beyond classification

Resource allocation

Broader context
What next?
Open-world recognition

Long-tail distributions

Challenging objects

Designing open-world evaluation frameworks
Large-scale video analysis

Multi-task models

Formulating the right temporal questions

Visual fluents
Collaborative visual systems

Effective teaching

Understanding human intention

Knowledge acquisition
AI will change the world. Who will change AI?
AI will change the world. Who will change AI?

Stanford Artificial Intelligence Laboratory’s Outreach Summer (SAILORS) program

24 high school girls, 2 weeks
Rigorous AI curriculum emphasizing humanistic applications

http://sailors.stanford.edu
Questions?

AI vision system

Training data
- cars
- dogs
- people

Algorithm

Test data
- Car, Person Building

Strongly supervised:
- RusDenHuabFei ICCV’13
- RusNg CVPR’10
- RusFei ECCVW’10
- RusGupRam UnderReview
- ModVezRusFei CVPR’15
- YeuRusJinAndMorFei UnderReview

Weakly supervised:
- RusLinYuFei ECCV’12
- BeaRusFerFei UnderReview

Olga Russakovsky: The human side of computer vision

http://cs.cmu.edu/~orussako

olgurus@cmu.edu