TimeGate

Efficient Recognition of Long-range Actions

N. Hussein  B. Bejnordi  M. Jain

@UvA  @Qualcomm
Motivation
Actions

human cooking in the kitchen

Videos {oocof.mp4, m1b7n.mp4} from Charades [3]
Actions

Long-range Action

Human Activity [1]
Complex Actions [2]

Short-range Action

Unit-Actions [1]
One-Actions [2]
Long-range Actions

- Long-range actions require a holistic reasoning

Preparing Breakfast

- get
- cook
- put
- wash
Long-range Actions

- Long-range actions require a holistic reasoning.

Preparing Breakfast

- get
- cook
- put
- wash
Long-range comprises unit-actions, some are discriminant
Long-range Actions

• Long-range comprises unit-actions, some are discriminant

Fried Egg

Scrambled Egg

Pancakes
Long-range comprises unit-actions, some are discriminant
Related Works

• Efficient architectures for CNNs
• Efficient architectures for CNNs

\sim 10 \text{ sec.}

\vdots

Efficient CNN

\downarrow

Predictions

\sim 2 \text{ min.}
• Frame-level sampling
Related Works

- Frame-level sampling

Short-range Actions

Fried Egg

- pour oil
- crack egg
- add salt-pepper

Pancakes

- crack egg
- pour milk
Method
We present VideoEpitoma for the efficient recognition of long-range actions.
• **Timestep selector for conditional timestep selection**
• Video classifier considers only selected timesteps
• Timestep selector for conditional timestep selection
• Temporal modeling for context-conditional gating
• Binary gating with the help of relaxation of sigmoid
Timestep Selector

- Clipped sigmoid fits well for timestep selection
• Timestep selector for conditional timestep selection

(a) Timestep Selector

(b) Video Classifier
• Video classifier considers only selected timesteps

(a) Timestep Selector

(b) Video Classifier
Results
Efficiency-Accuracy Tradeoff

- Stand-alone selector improves out-of-box CNN classifiers
• The improvement is regardless of the CNN, and at different time scales.
Better Gating for End-to-end

- In end-to-end, the selector learns better decision to the benefit of the classifier
• Timesteps are selected based on their relevance to their video.
Breakdown of Computational Cost

- The computational cost of the selector is marginal to that of the classifier.

<table>
<thead>
<tr>
<th>Model</th>
<th>Timesteps</th>
<th>FLOPS (G)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LightNet</td>
<td>HeavyNet</td>
<td>LightNet+Gating</td>
</tr>
<tr>
<td>RNet2D</td>
<td>—</td>
<td>16</td>
<td>—</td>
</tr>
<tr>
<td>RNet2D + Ours</td>
<td>16</td>
<td>8</td>
<td>0.94</td>
</tr>
<tr>
<td>SNet3D</td>
<td>—</td>
<td>16</td>
<td>—</td>
</tr>
<tr>
<td>SNet3D + Ours</td>
<td>16</td>
<td>8</td>
<td>0.94</td>
</tr>
<tr>
<td>I3D</td>
<td>—</td>
<td>16</td>
<td>—</td>
</tr>
<tr>
<td>I3D + Ours</td>
<td>16</td>
<td>8</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Thank You!