Video summarisation by classification with deep reinforcement learning

Kaiyang Zhou, Tao Xiang, Andrea Cavallaro

Published in: *British Machine Vision Conference (BMVC) 2018*

Centre for Intelligent Sensing
Queen Mary University of London
What is video summarisation?

**Goal:** to automatically summarize videos into keyframes or key-clips.

We want summaries to be:
- informative
- content-specific
Current video summarisation methods

Unsupervised methods use generic criteria e.g. diversity, representativeness.

1. Feature extraction
2. Clustering
3. Keyframes extraction

Limitations: generic criteria cannot capture content-specific concepts.
Current video summarisation methods

Supervised methods rely on manual annotations.

e.g. scores: \( y = \{0.1, 0.8, 1.0, 0.2, \ldots\} \)

Training

\[
\text{loss} = (y - w^T X)^2
\]

Inference

\[
p = w^T X'
\]

Limitations: labels are costly to collect and prone to be biased.
Our idea: weakly supervised + RL

1. Video-level category labels are descriptive of video content and very easy to obtain.

2. To train a summarisation model by encouraging it to produce summaries maintaining category-related information.
Framework overview

(a) Training classification network.

\[ V_i = \{ x_j \}_{j=1}^T \]

\[ \hat{y}_i \rightarrow \text{Classification loss} \]

(b) Training summarisation network.

\[ V_i = \{ x_j \}_{j=1}^T \]

\[ \text{Summarisation Network} \rightarrow \text{Summary} \rightarrow \text{Classification Network} \rightarrow \hat{y}_i \rightarrow \text{Reward function} \]

\[ \text{Reward} \]

\[ (V_1, y_1 = \text{‘Groom Animal’}) \]

\[ (V_N, y_N = \text{‘Bike Tricks’}) \]

Training videos

\[ (V_1, y_1 = \text{‘Groom Animal’}) \]

\[ (V_N, y_N = \text{‘Bike Tricks’}) \]

Training videos

\[ (V_1, y_1 = \text{‘Groom Animal’}) \]

\[ (V_N, y_N = \text{‘Bike Tricks’}) \]

Training videos
Network architectures

**Classification network**

- **Bidirectional GRU Network**
  - FC
  - $x_1, x_2, \ldots, x_T$

**Summarisation network**

- **Bidirectional GRU Network**
  - FC
  - $x_j, \ldots, x_t, \ldots, x_T$

\[ V(s) + A(s, a) = Q(s, a) \]

$Q(s, a = 1)$: keep frame
$Q(s, a = 0)$: remove frame
Sequential decision making process

\[ t = 1 \]
\[ s_1 = \{x_1, x_2, x_3, x_4, x_5\} \]
\[ s_1 \rightarrow \text{Model} \rightarrow Q(s_1, a_1) \in \mathbb{R}^2 \]
if \( Q(s_1, a_1 = 1) > Q(s_1, a_1 = 0) \): # epsilon-greedy is used in practice
\[ s_2 = \{x_1, x_2, x_3, x_4, x_5\} \] # \( x_1 \) is kept
else:
\[ s_2 = \{x_2, x_3, x_4, x_5\} \] # \( x_1 \) is removed
\[ r_1 = \mathcal{R}(r_1|s_1, a_1, s_2) \]

\[ t = 2 \]
\[ s_2 \rightarrow \text{Model} \rightarrow Q(s_2, a_2) \in \mathbb{R}^2 \]
\[ \cdots \]
\[ \cdots \]
\[ r_2 = \mathcal{R}(r_2|s_2, a_2, s_3) \]
\[ \cdots \]
until \( t = T \) or \( |s_t| < \tau \)


Reward functions

1. Global recognisability reward $r^g_t$

$$r^g_t = \begin{cases} 
+1, & \text{if } \hat{y} = y, \quad \# \text{ summary can be recognised by the expert} \\
-5, & \text{else.}
\end{cases}$$

2. Local relative importance reward $r^l_t$

$$r^l_t = \tanh\left( \frac{\hat{y}^*(s_t) - \hat{y}^*(s_{t+1})}{\eta} \right) + 0.05(1 - a_t)$$

where $\hat{y}^*$ means rank of true category

3. Unsupervised reward $r^u_t$

$$r^u_t = \frac{1}{|\mathcal{Y}||\mathcal{Y} - 1|} \sum_{t \in |\mathcal{Y}|} \sum_{t' \in |\mathcal{Y}|, \ t' \neq t} d(x_t, x_{t'}) + \exp\left(-\frac{1}{T} \sum_{t=1}^{T} \min_{t' \in \mathcal{Y}} \|x_t - x_{t'}\|_2\right)$$

- dissimilarity among selected frames
- reconstruction error
Optimisation with double Q-learning

1. sample experience $e_t = (s_t, a_t, s_{t+1})$ from replay memory $\mathcal{M}$

2. $L = \mathbb{E}_{e_t \sim \mathcal{M}}[(R_t - Q_{\theta}(s_t, a_t))^2]$ 
   \[ \text{s.t. } R_t = r_t + \gamma \max_{a_{t+1}} Q_{\theta}(s_{t+1}, a_{t+1}) \]

3. update model with gradient descent $\theta = \theta - \alpha \nabla_{\theta} L$
## Evaluation: datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th># videos</th>
<th>Length (mins)</th>
<th># categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVSum</td>
<td>50</td>
<td>2-10</td>
<td>10</td>
</tr>
<tr>
<td>CoSum</td>
<td>51</td>
<td>1-12</td>
<td>10</td>
</tr>
</tbody>
</table>

**Categories of TVSum**

1. Changing Vehicle Tire (VT)
2. Getting Vehicle Unstuck (VU)
3. Grooming an Animal (GA)
4. Making Sandwich (MS)
5. Parkour (PK)
6. Parade (PR)
7. Flash Mob Gathering (FM)
8. Beekeeping (BK)
9. Attempting Bike Tricks (BT)
10. Dog Show (DS)

**Categories of CoSum**

11. Base Jumping (BJ)
12. Bike Polo (BP)
13. Eiffel Tower (ET)
14. Excavator River Crossing (ERC)
15. Kids Playing in Leaves (KID)
16. MLB (MLB)
17. NFL (NFL)
18. Notre Dame Cathedral (NDC)
19. Statue of Liberty (SL)
20. Surfing (SURF)
Evaluation: metrics

- **Human summary**
  - True positive
  - False negative

- **Machine summary**
  - True positive
  - False positive

\[
F\text{-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]
# Quantitative results

<table>
<thead>
<tr>
<th>Method</th>
<th>Label</th>
<th>TVSum</th>
<th>CoSum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform sampling</td>
<td>✗</td>
<td>15.5</td>
<td>20.4</td>
</tr>
<tr>
<td>K-medoids</td>
<td>✗</td>
<td>28.8</td>
<td>34.3</td>
</tr>
<tr>
<td>Dictionary selection [4]</td>
<td>✗</td>
<td>42.0</td>
<td>37.2</td>
</tr>
<tr>
<td>Online sparse coding [14]</td>
<td>✗</td>
<td>46.0</td>
<td>-</td>
</tr>
<tr>
<td>Co-archetypal [28]</td>
<td>✗</td>
<td>50.0</td>
<td>-</td>
</tr>
<tr>
<td>GAN [16]</td>
<td>✗</td>
<td>51.7</td>
<td>44.0</td>
</tr>
<tr>
<td>DR-DSN [46]</td>
<td>✗</td>
<td>57.6</td>
<td>47.8</td>
</tr>
<tr>
<td>LSTM [41]</td>
<td>frame-level</td>
<td>54.2</td>
<td>46.5</td>
</tr>
<tr>
<td>GAN [16]</td>
<td>frame-level</td>
<td>56.3</td>
<td>50.2</td>
</tr>
<tr>
<td>DR-DSN [46]</td>
<td>frame-level</td>
<td>58.1</td>
<td>54.3</td>
</tr>
<tr>
<td>Backprop-Grad [21]</td>
<td>video-level</td>
<td>52.7</td>
<td>46.2</td>
</tr>
<tr>
<td>DQSN ($r^g$)</td>
<td>video-level</td>
<td>57.9</td>
<td>50.1</td>
</tr>
<tr>
<td>DQSN ($r^g + r^u$)</td>
<td>video-level</td>
<td>58.1</td>
<td>51.7</td>
</tr>
<tr>
<td>DQSN ($r^g + r^l$)</td>
<td>video-level</td>
<td>58.2</td>
<td>52.0</td>
</tr>
<tr>
<td>DQSN (full model)</td>
<td>video-level</td>
<td>58.6</td>
<td>52.1</td>
</tr>
</tbody>
</table>

Table 1: Summarisation results (%) on TVSum and CoSum. 1\textsuperscript{st}/2\textsuperscript{nd} best in red/blue. Full model means $r^g + r^l + r^u$. 

CIS centre for intelligent sensing

Queen Mary University of London
Analysis of local relative importance reward

Figure 3: Example frames that downgraded (red) / improved (green) the rank of true category in classification when being removed.
Thanks!
Any questions?