Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward

Kaiyang Zhou, Yu Qiao, Tao Xiang

AAAAI 2018
What is video summarization?

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

- We want:
  - Diverse
  - Representative
Application of video summarization
e.g. YouTube video preview
Unsupervised video summarization

Idea: to analyze correlations between frames in feature space

1. Feature extraction  
2. Clustering  
3. Keyframes extraction

Supervised video summarization

Idea: to exploit human labels

scores: $y = \{0.1, 0.8, 1.0, 0.2, \ldots\}$  keyframes: $y = \{0, 1, 1, 0, \ldots\}$

Training

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

Inference

$p = w^T X'$

• Temporal relations are hard to capture by linear models.
Recurrent neural network with supervised learning

Idea: use RNN to capture temporal relations

- Collecting labels here is much more expensive than that of other tasks.
- Labels may not provide good supervision signals. (b/c labels are subjective)

Zhang et al. ECCV 2016.
Main idea
To mimic how humans summarize videos

Is summary diverse and representative?
Diversity-representativeness reward
Diversity reward

\[ R_{\text{div}} = \frac{1}{|Y|(|Y|-1)} \sum_{t \in Y} \sum_{t' \in Y, t' \neq t} d(x_t, x_{t'}) \]

Set of selected frames

Representativeness reward

\[ R_{\text{rep}} = \exp\left(-\frac{1}{T} \sum_{t=1}^{T} \min_{t' \in Y} \|x_t - x_{t'}\|_2\right) \]
Optimization

Reward:

\[ R = R_{\text{div}} + R_{\text{rep}} \]

Objective function:

\[ J(\theta) = \mathbb{E}[R] \]

Approximate gradients via REINFORCE:

\[ \nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} (R_n - b) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t) \]

Inference

Score prediction:
\[
\{p_i\}_{i=1}^T = \text{RNN}(\{x_i\}_{i=1}^T)
\]

Compute clip-level scores:
\[
I(S_k) = \frac{1}{|S_k|} \sum_{i \in S_k} p_i
\]

Select clips (0/1 Knapsack):
\[
\arg \max_{\mu} \sum_k \mu_k I(S_k), \quad \sum_k \mu_k |S_k| \leq \gamma, \quad \mu_k \in \{0, 1\}
\]

Song et al. CVPR 2015.
### Evaluation

#### Metric: \( F-score = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \)

<table>
<thead>
<tr>
<th>Dataset</th>
<th># videos</th>
<th>Length (mins)</th>
<th>Description</th>
<th># annotators per video</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumMe</td>
<td>25</td>
<td>1-6</td>
<td>User videos</td>
<td>15-18</td>
</tr>
<tr>
<td>TVSum</td>
<td>50</td>
<td>2-10</td>
<td>YouTube videos</td>
<td>20</td>
</tr>
</tbody>
</table>

Quantitative Results
Table: Comparison with other unsupervised approaches.

<table>
<thead>
<tr>
<th>Method</th>
<th>SumMe (%)</th>
<th>TVSum (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video-MMR</td>
<td>26.6</td>
<td>-</td>
</tr>
<tr>
<td>Uniform sampling</td>
<td>29.3</td>
<td>15.5</td>
</tr>
<tr>
<td>K-medoids</td>
<td>33.4</td>
<td>28.8</td>
</tr>
<tr>
<td>Vsumm</td>
<td>33.7</td>
<td>-</td>
</tr>
<tr>
<td>Web image</td>
<td>-</td>
<td>36.0</td>
</tr>
<tr>
<td>Dictionary selection</td>
<td>37.8</td>
<td>42.0</td>
</tr>
<tr>
<td>Online sparse coding</td>
<td>-</td>
<td>46.0</td>
</tr>
<tr>
<td>Co-archetypal</td>
<td>-</td>
<td>50.0</td>
</tr>
<tr>
<td>GAN_{dpp}</td>
<td>39.1</td>
<td>51.7</td>
</tr>
<tr>
<td>Ours</td>
<td>41.4</td>
<td>57.6</td>
</tr>
</tbody>
</table>

\{ \uparrow 6\% \} \quad \{ \uparrow 11\% \}
## Quantitative Results

Table: Comparison with other supervised approaches.

<table>
<thead>
<tr>
<th>Method</th>
<th>SumMe (%)</th>
<th>TVSum (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interestingness</td>
<td>39.4</td>
<td>-</td>
</tr>
<tr>
<td>Submodularity</td>
<td>39.7</td>
<td>-</td>
</tr>
<tr>
<td>Summary transfer</td>
<td>40.9</td>
<td>-</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>37.6</td>
<td>54.2</td>
</tr>
<tr>
<td>DPP-LSTM</td>
<td>38.6</td>
<td>54.7</td>
</tr>
<tr>
<td>GAN&lt;sub&gt;sup&lt;/sub&gt;</td>
<td>41.7</td>
<td>56.3</td>
</tr>
<tr>
<td>Ours</td>
<td>41.4</td>
<td>57.6</td>
</tr>
</tbody>
</table>
## Quantitative Results

<table>
<thead>
<tr>
<th>Method</th>
<th>SumMe (%)</th>
<th>TVSum (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interestingness</td>
<td>39.4</td>
<td>-</td>
</tr>
<tr>
<td>Submodularity</td>
<td>39.7</td>
<td>-</td>
</tr>
<tr>
<td>Summary transfer</td>
<td>40.9</td>
<td>-</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>37.6</td>
<td>54.2</td>
</tr>
<tr>
<td>DPP-LSTM</td>
<td>38.6</td>
<td>54.7</td>
</tr>
<tr>
<td>$\text{GAN}_{\text{sup}}$</td>
<td>41.7</td>
<td>56.3</td>
</tr>
<tr>
<td>Ours</td>
<td>41.4</td>
<td>57.6</td>
</tr>
<tr>
<td>Ours (supervised)</td>
<td>42.1</td>
<td>58.1</td>
</tr>
</tbody>
</table>

Table: Comparison with other supervised approaches. For more experiments and details, please see our paper.
Qualitative Results

Video #10 in TVSUM

Manual scores

Prediction by RNN with RL

DR-DSN F-score = 41.9
XCorr = 91.84

Prediction by RNN with supervised learning

$\text{DSN}_{\text{sup}}$ F-score = 41.7
XCorr = 90.13
Summary

1. Proposed a label-free reward, i.e. diversity-representativeness reward.
2. Outperformed/competitive to other unsupervised/supervised ones.
3. Extended the unsupervised method to the supervised version.

Code and data available at: https://github.com/KaiyangZhou/vsumm-reinforce
Thanks!

Any questions?

please feel free to contact me at: k.zhou@qmul.ac.uk