Video summarisation by classification with deep reinforcement learning

Kaiyang Zhou, Tao Xiang, Andrea Cavallaro

Published in:
British Machine Vision Conference (BMVC) 2018

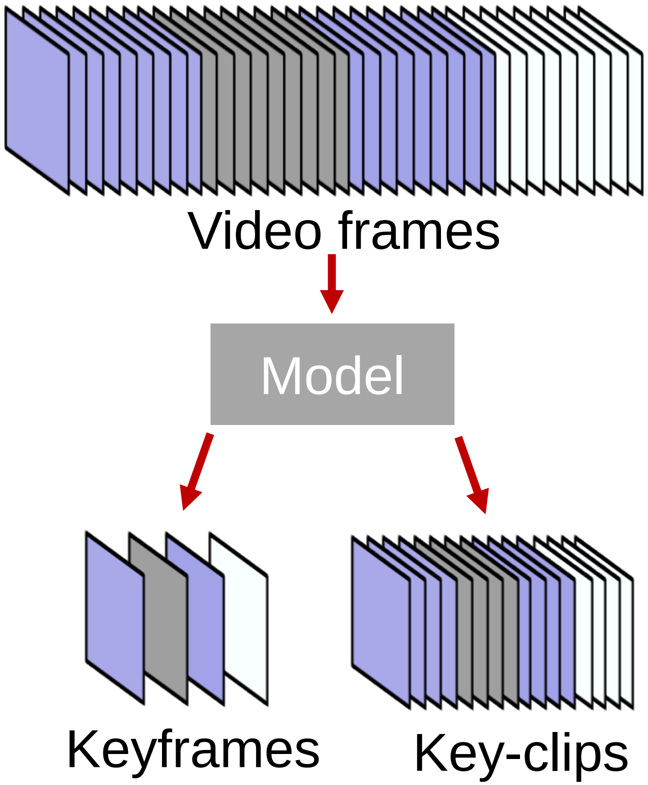
Centre for Intelligent Sensing
Queen Mary University of London

CIS centre for intelligent sensing

Queen Mary University of London

What is video summarisation?

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



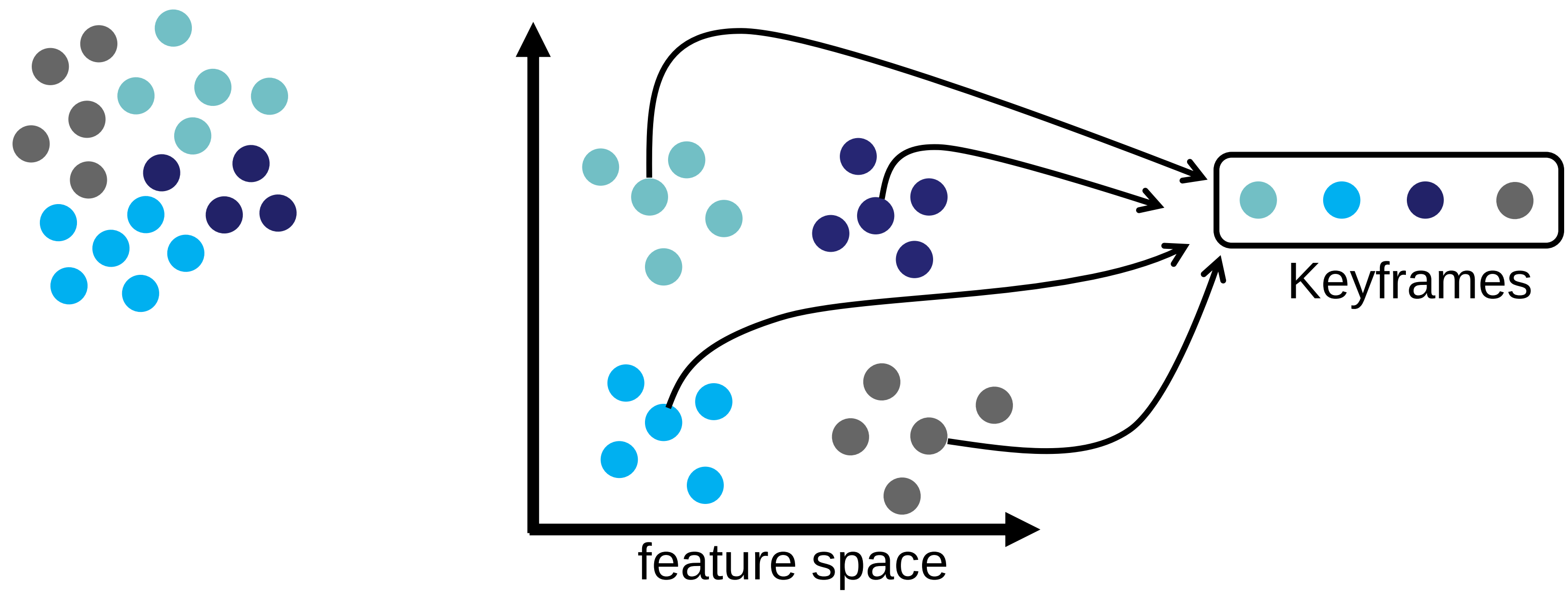
We want summaries to be:

- informative
- content-specific

Current video summarisation methods

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

1. Feature extraction 2. Clustering 3. Keyframes extraction

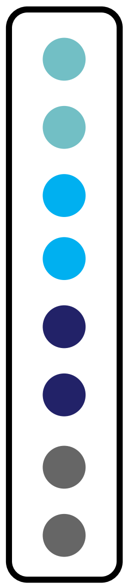


Limitations: generic criteria cannot capture content-specific concepts.

Current video summarisation methods

Supervised methods rely on manual annotations.

e.g. scores: $y = \{0.1, 0.8, 1.0, 0.2, ...\}$



feature vectors

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

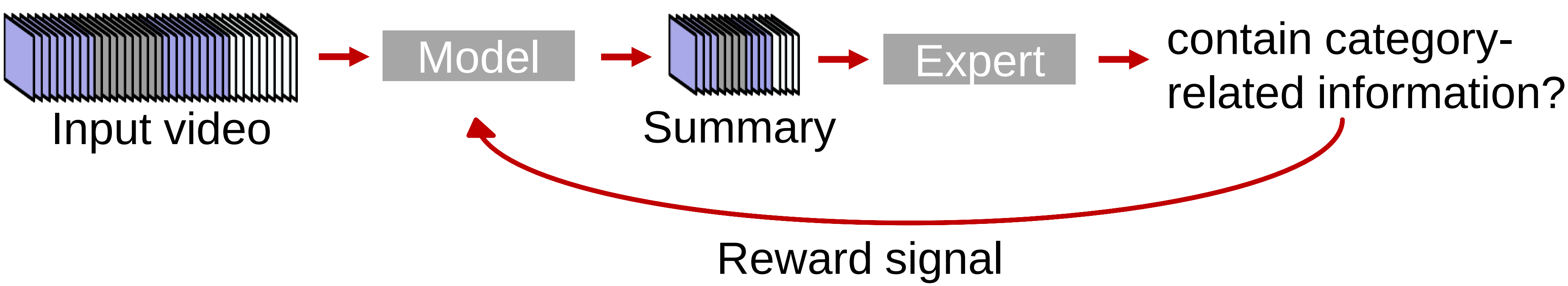
$\xrightarrow{\text{Inference}}$ $p = w^T X'$

Limitations: labels are costly to collect and prone to be biased.

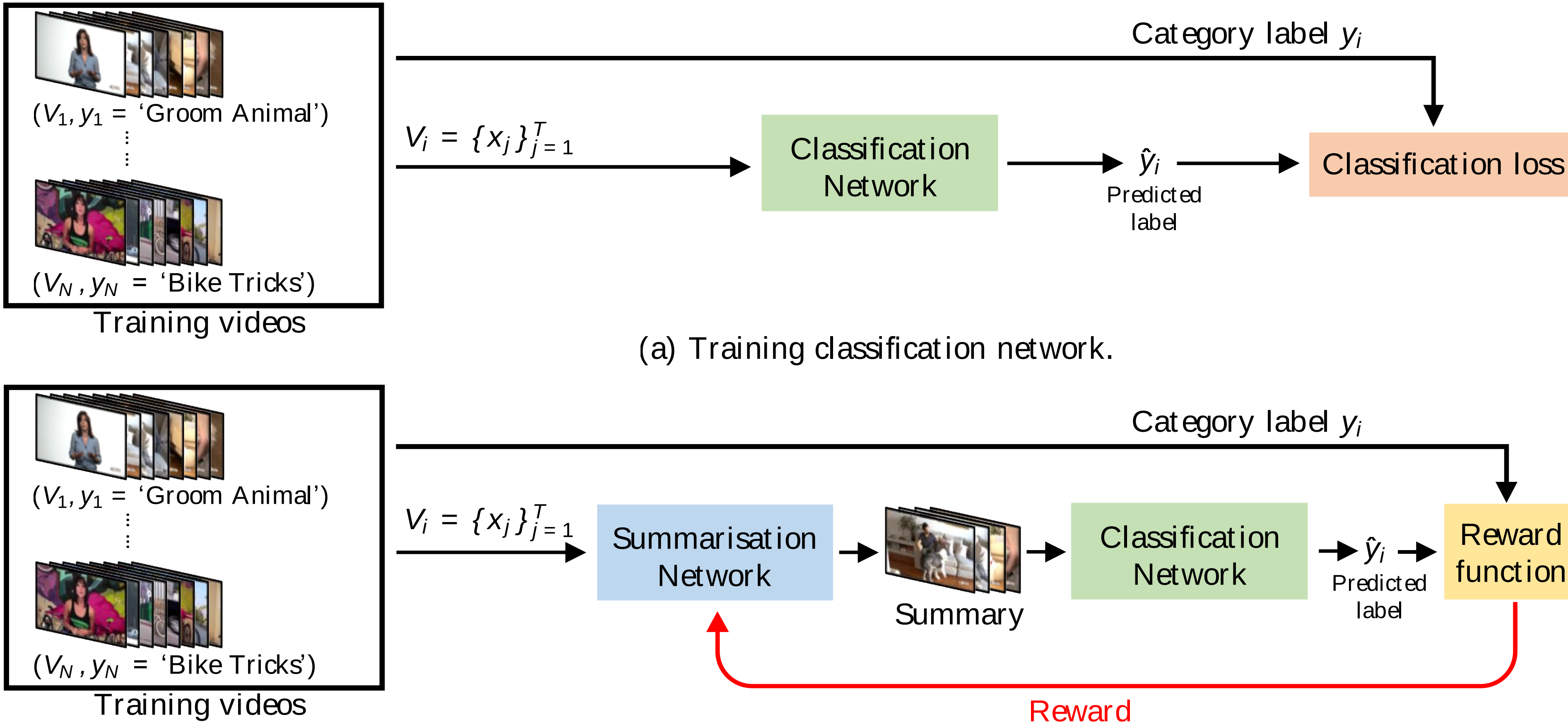
Our idea: weakly supervised + RL

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



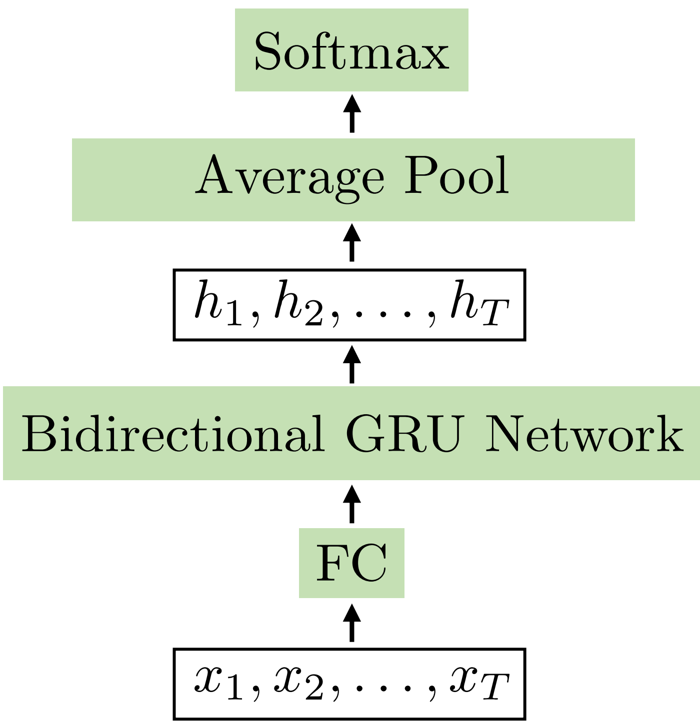
(a) Training classification network.

(b) Training summarisation network.

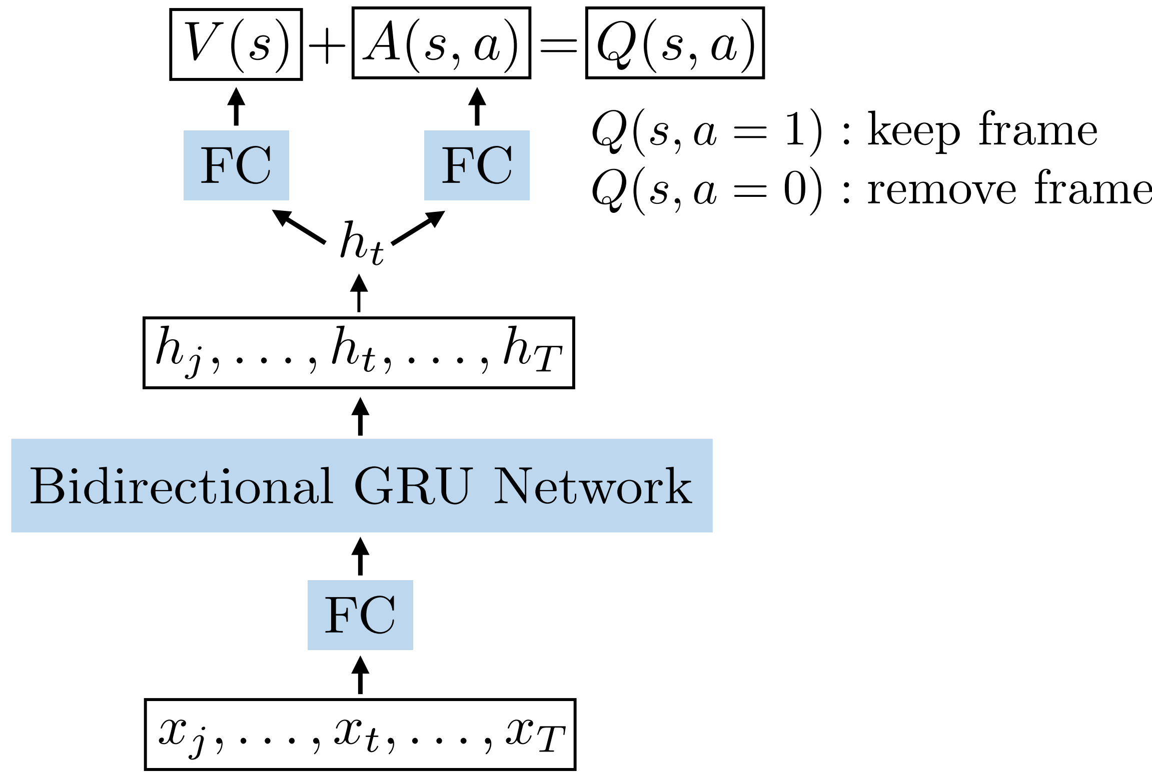
CIS centre for intelligent sensing

Queen Mary University of London

Network architectures



Classification network



Summarisation network

Sequential decision making process

$t = 1$

$s_1 = \{\underline{x_1}, x_2, x_3, x_4, x_5\}$

$s_1 \rightarrow$ **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

$s_2 = \{x_1, \underline{x_2}, x_3, x_4, x_5\}$ # x_1 is kept

else:

$s_2 = \{\underline{x_2}, x_3, x_4, x_5\}$ # x_1 is removed

$r_1 = \mathcal{R}(r_1 | s_1, a_1, s_2)$

$t = 2$

$s_2 \rightarrow$ **Model** $\rightarrow Q(s_2, a_2) \in \mathbb{R}^2$

$\vdots$

$r_2 = \mathcal{R}(r_2 | s_2, a_2, s_3)$

$\vdots$

until $t == T$ or $|s_t| < \tau$

Reward functions

1. Global recognisability reward r_t^g

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

2. Local relative importance reward r_t^l

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

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

3. Unsupervised reward r_t^u

$$r_t^u = \underbrace{\frac{1}{|\mathcal{Y}||\mathcal{Y}-1|} \sum_{t \in |\mathcal{Y}|} \sum_{\substack{t' \in |\mathcal{Y}| \\ t' \neq t}} d(x_t, x_{t'})}_{\text{dissimilarity among selected frames}} + \underbrace{\exp\left(-\frac{1}{T} \sum_{t=1}^{T} \min_{t' \in \mathcal{Y}} ||x_t - x_{t'}||_2\right)}_{\text{reconstruction error}}$$

Optimisation with double Q-learning

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

2. $L = \mathbb{E}_{\{e_t\}\sim\mathcal{M}}[(R_t - Q_\theta(s_t, a_t))^2]$

   s.t. $\quad R_t = r_t + \gamma Q_{\theta^-}(s_{t+1}, \arg\max_{a_{t+1}} Q_\theta(s_{t+1}, a_{t+1}))$

3. update model with gradient descent $\theta = \theta - \alpha \nabla_\theta L$

Evaluation: datasets

Dataset	# videos	Length (mins)	# categories
TVSum	50	2-10	10
CoSum	51	1-12	10

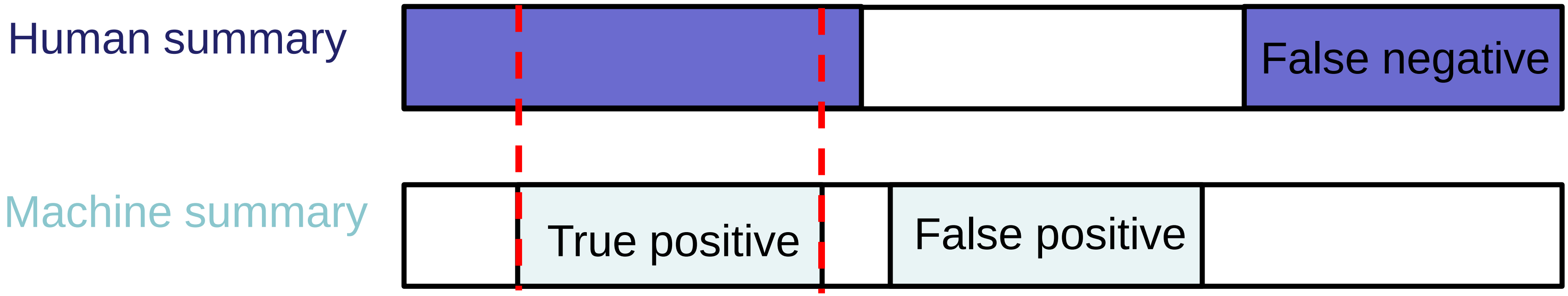
1	Changing Vehicle Tire (VT)
2	Getting Vehicle Unstuck (VU)
3	Grooming an Animal (GA)
4	Making Sandwich (MS)
5	Parkour (PK)
6	Parade (PR)
7	Flash Mob Gatering (FM)
8	BeeKeeping (BK)
9	Attempting Bike Tricks (BT)
10	Dog Show (DS)

Categories of TVSum

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)

Categories of CoSum

Evaluation: metrics



$$\text{F-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Quantitative results

Method	Label	TVSum	CoSum
Uniform sampling	✗	15.5	20.4
K-medoids	✗	28.8	34.3
Dictionary selection [4]	✗	42.0	37.2
Online sparse coding [44]	✗	46.0	-
Co-archetypal [28]	✗	50.0	-
GAN [16]	✗	51.7	44.0
DR-DSN [46]	✗	57.6	47.8
LSTM [41]	frame-level	54.2	46.5
GAN [16]	frame-level	56.3	50.2
DR-DSN [46]	frame-level	58.1	54.3
Backprop-Grad [21]	video-level	52.7	46.2
DQSN (r^g)	video-level	57.9	50.1
DQSN ($r^g + r^u$)	video-level	58.1	51.7
DQSN ($r^g + r^l$)	video-level	58.2	52.0
DQSN (full model)	video-level	58.6	52.1

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

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?

Paper link: https://arxiv.org/abs/1807.03089