Video Summarisation by Classification with Deep Reinforcement Learning

Kaiyang Zhou, Tao Xiang, Andrea Cavallaro

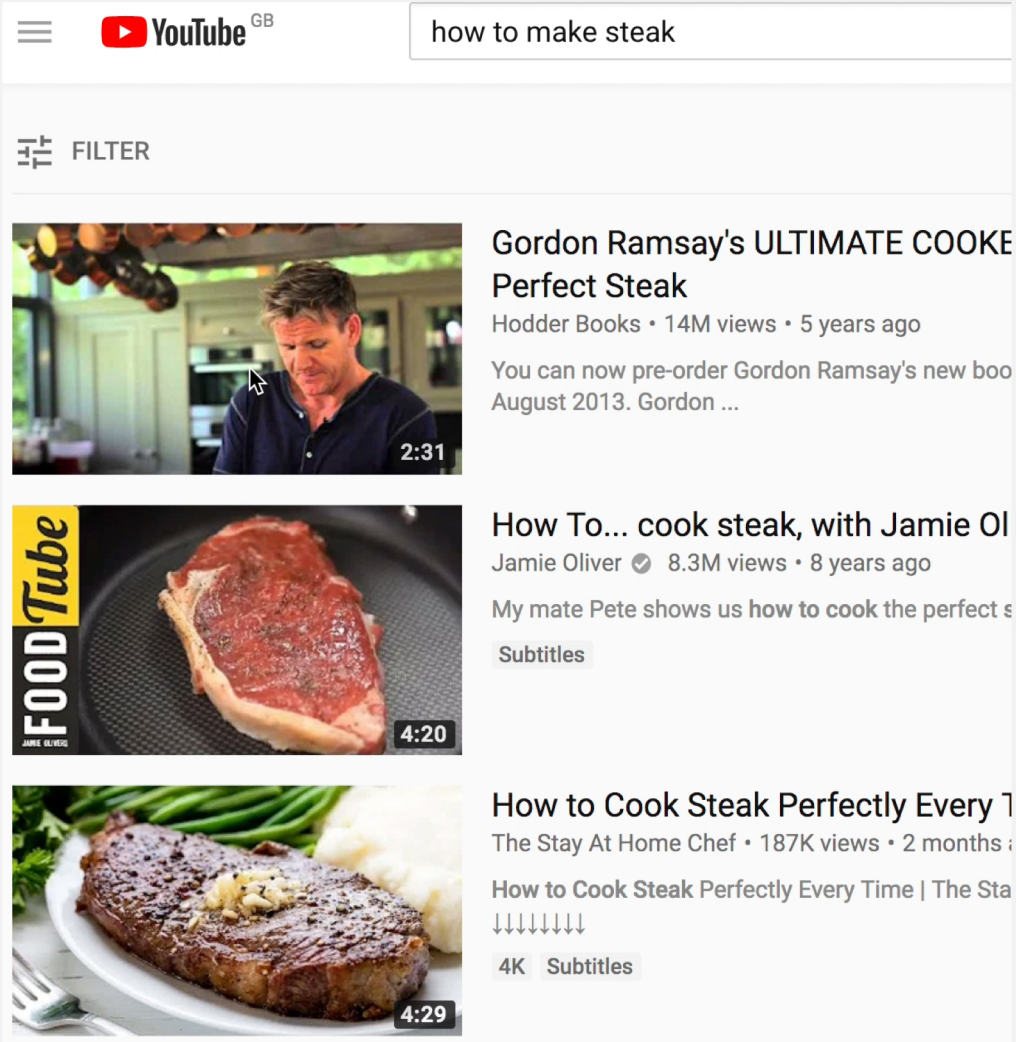
Queen Mary University of London, UK

BMVC 2018

What is video summarisation?

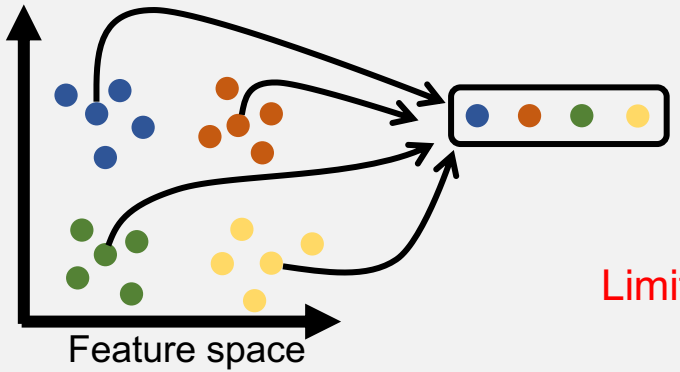
The goal is to learn a model which can automatically summarise videos

Application: YouTube video preview



Most methods are based on *unsupervised* and *supervised* learning.

- Unsupervised methods use heuristic criteria e.g. diversity.



Limitation: criteria are too general.

[Zhao and Xing, CVPR'14; Yang et al., ICCV'15; Mahasseni et al., CVPR'17; Zhou et al., AAAI'18]

- Supervised methods rely on labels to train supervised models.

features

+ labels (e.g. frame importance scores)

Limitation: labels are too expensive to collect and biased.

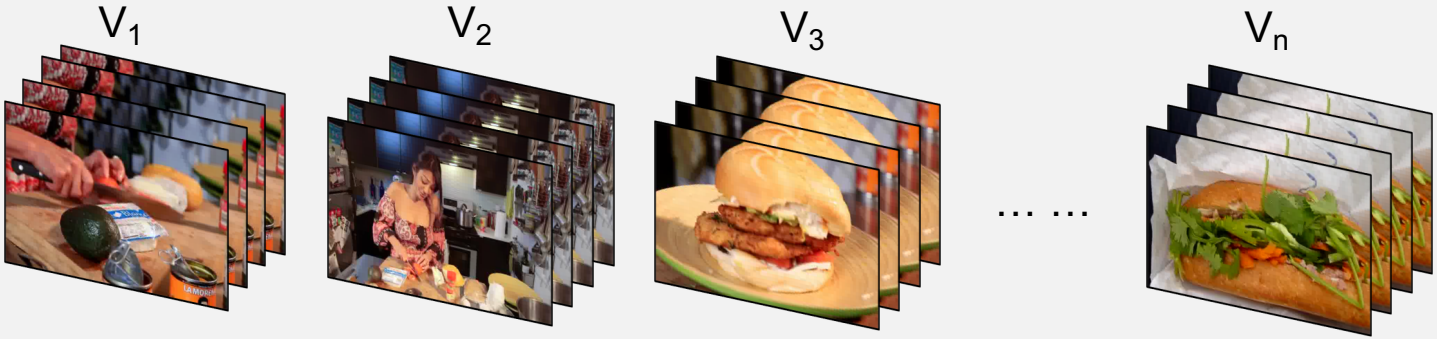
Supervised model

[Gong et al., NIPS'14; Gygli et al., ECCV'14; Zhang et al., ECCV'16; Zhao et al., CVPR'18]

Our idea is to exploit weak labels i.e. video-level categories.

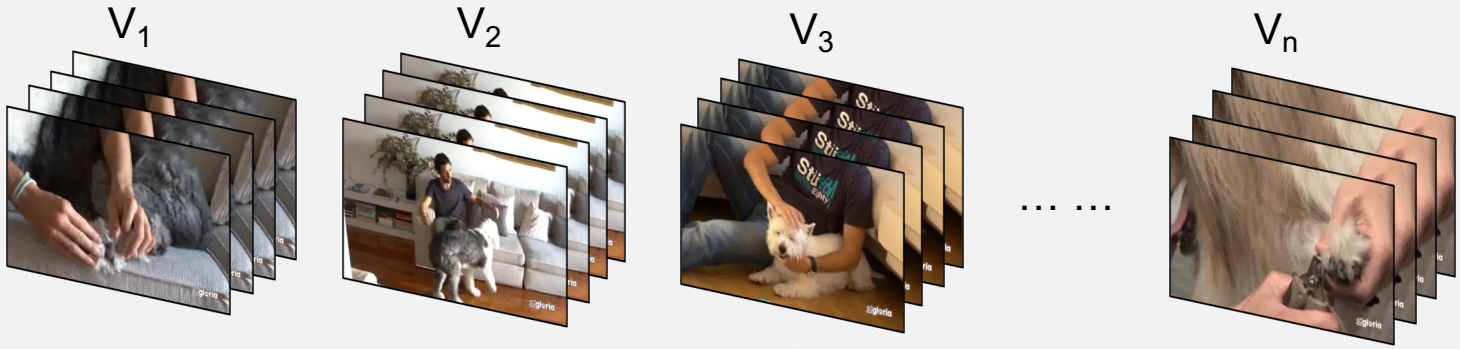
The intuition is that category labels are descriptive of video content and videos with same categories share similarities, for example:

Making sandwich



These videos should contain temporal stages of making sandwich

Groom animal

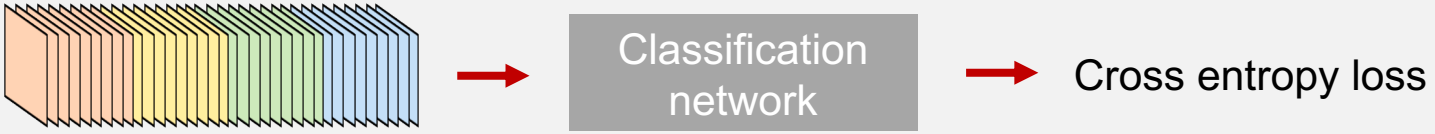


These videos should contain scenes of people working on animals

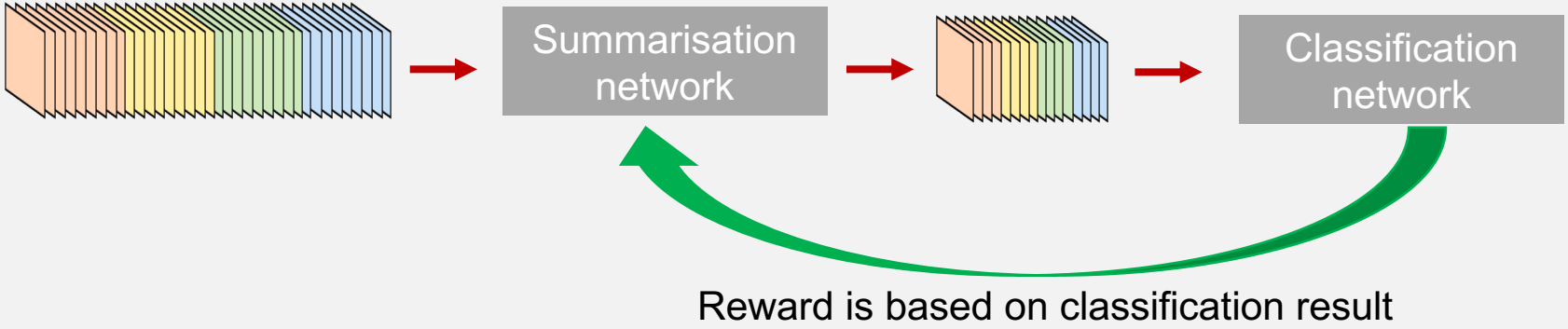
We want to maintain the recognisability of video while removing frames as many as possible

We use RL to train a summarisation agent, which can produce short video summaries that preserve category-specific information.

Stage 1:



Stage 2:



Both networks are bidirectional recurrent networks.

Why use RL:

1. Supervision signal (i.e. classification result) is delayed.
2. Exploration-exploitation strategy of RL can try different combinations of frames.

Markov Decision Process (MDP)

- State: remaining frames in a video sequence, $\{x_t, x_{t+1}, ..., x_T\}$.
- Action: $Q(s_t, a_t = 1)$ for keeping frame, $Q(s_t, a_t = 0)$ for removing frame.
- Transition: $P(s_{t+1}|s_t, a_t)$.
- Reward: $R(s_t, a_t, s_{t+1})$.
- Discount factor: $\gamma = 0.99$.

Optimisation with double Q-learning

- Replay memory: $\mathcal{M} = \{e_1, e_2, ..., e_n\}$ where $e_i = (s_t, a_t, s_{t+1})$.
- ϵ-greedy action selection.
- Minimise regression loss:

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

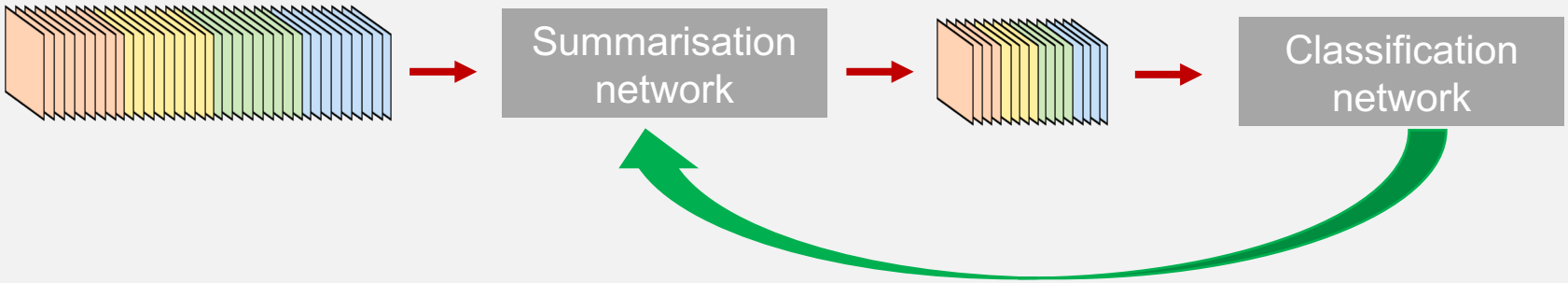
[Mnih et al., NIPSW'13; Hasselt et al., AAAI'16; Zhou et al., AAAI'18]

Reward functions

$$r_t = r_t^g \quad \#\text{global recognisability reward}$$
$$+ r_t^l \quad \#\text{local relative importance reward}$$
$$+ r_t^u \quad \#\text{unsupervised reward}$$

Main contributions

Global recognisability reward



$$r_t^g = \begin{cases} +1, & \text{if} \quad \hat{y} = y, \\ -5 & \text{otherwise}, \end{cases}$$
$$\text{s.t.} \quad t = T.$$

- Give positive rewards if produced summaries are recognisable, otherwise penalise with negative rewards.
- Stronger weights are given for penalty to encourage the summaries to have high recognition accuracy (determined empirically).

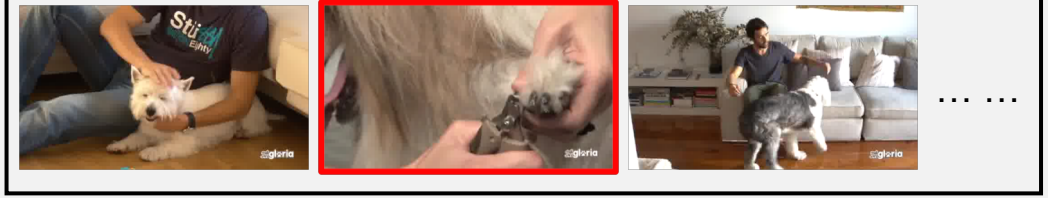
Local relative importance reward

$$r_t^l = 0.05(1 - a_t) + \tanh(\frac{\xi_t - \xi_{t+1}}{\eta}), \quad \text{s.t.} \quad t < T.$$

$$\xi_t \text{ represents rank of true category of } s_t$$

For example,

Current state:

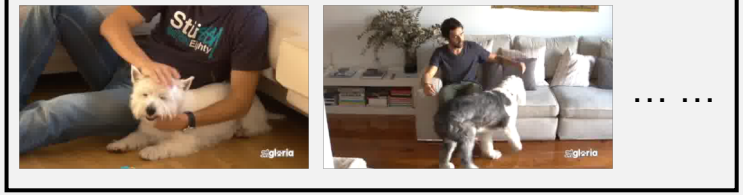


Looking at this frame

1. Groom animal 0.7
2. Dog show 0.28
3. Bike tricks 0.01
4. Parkour 0.01

Current action: Remove frame

Next state:



1. Dog show 0.51
2. Groom animal 0.47
3. Bike tricks 0.01
4. Parkour 0.01

This action should be penalised

Unsupervised reward

Dissimilarity among selected frames

Inverse of reconstruction error

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

$$\text{s.t.} \quad t = T.$$

Indices of selected frames

[Zhou et al., AAAI'18]

Evaluation: datasets

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

[Song et al., CVPR'15; Chu et al., CVPR'15]

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

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

Higher the F-score, closer the machine summary to human summary

Quantitative results

	Method	Label	TVSum	CoSum
Unsupervised	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
Supervised	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
Weakly supervised	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 $r^g + r^l + r^u$.

Quantitative results

	Method	Label	TVSum	CoSum
Unsupervised	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
Supervised	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
Weakly supervised	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

→ Complementarity

→ Local reward is useful

→ Combine all is good

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

Quantitative results

	Method	Label	TVSum	CoSum
Unsupervised	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
Supervised	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
Weakly supervised	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 $r^g + r^l + r^u$.

Quantitative results

	Method	Label	TVSum	CoSum
Unsupervised	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
Supervised	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
Weakly supervised	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 $r^g + r^l + r^u$.

How does local reward help?

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

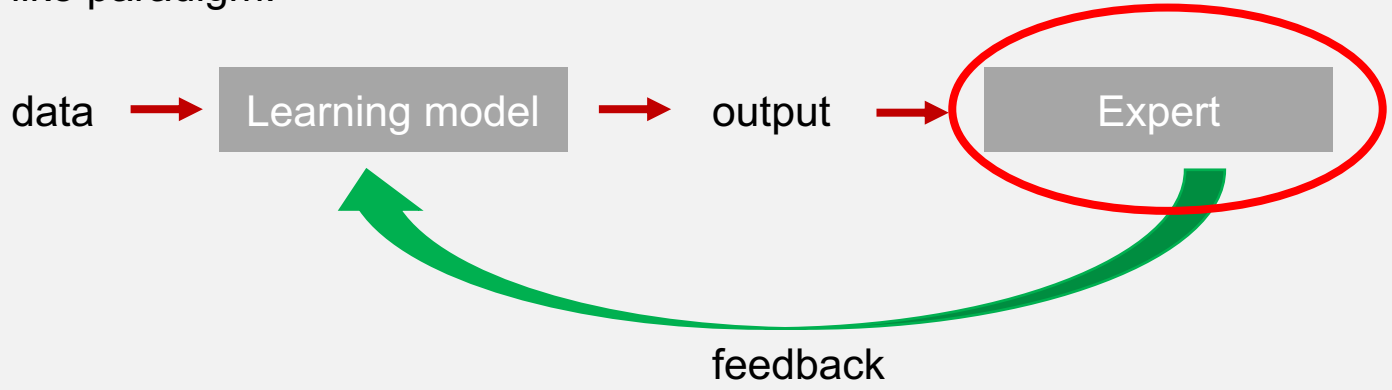
Trick: classification networks needs to be carefully trained.

Summary

We tackled video summarisation using weak labels and RL

Actor-critic-like paradigm:

should be carefully trained so it can distinguish between good and non-good



Thanks
Any questions?

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