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

Полина Святокум

НИУ ВШЭ

05.03.2018

Video summarization

Из видеоряда $V_i = \{v_t\}_1^T$ выбрать подмножество кадров $\mathcal{Y}$, которое будет описывать видео.
Более распространен и успешен supervised подход.

Задача в форме RL

- Unsupervised подход
- Необходимо получать обратный сигнал для обучения
- Награда должна определять идеальное summary

Diversity-representativeness reward

Diversity reward

$$R_{DIV} = \frac{1}{|\mathcal{Y}|(|\mathcal{Y}|-1)} \sum_{t \in \mathcal{Y}} \sum_{t' \in \mathcal{Y} t' \neq t} d(x_t, x_{t'})$$

Dissimilarity function

$$d(x_t, x_{t'}) = 1 - \frac{x_t^T x_{t'}}{||x_t||_2 ||x_{t'}||_2}$$

*****

$$d(x_t, x_{t'}) = 1, \text{ если } |t - t'| > \lambda$$

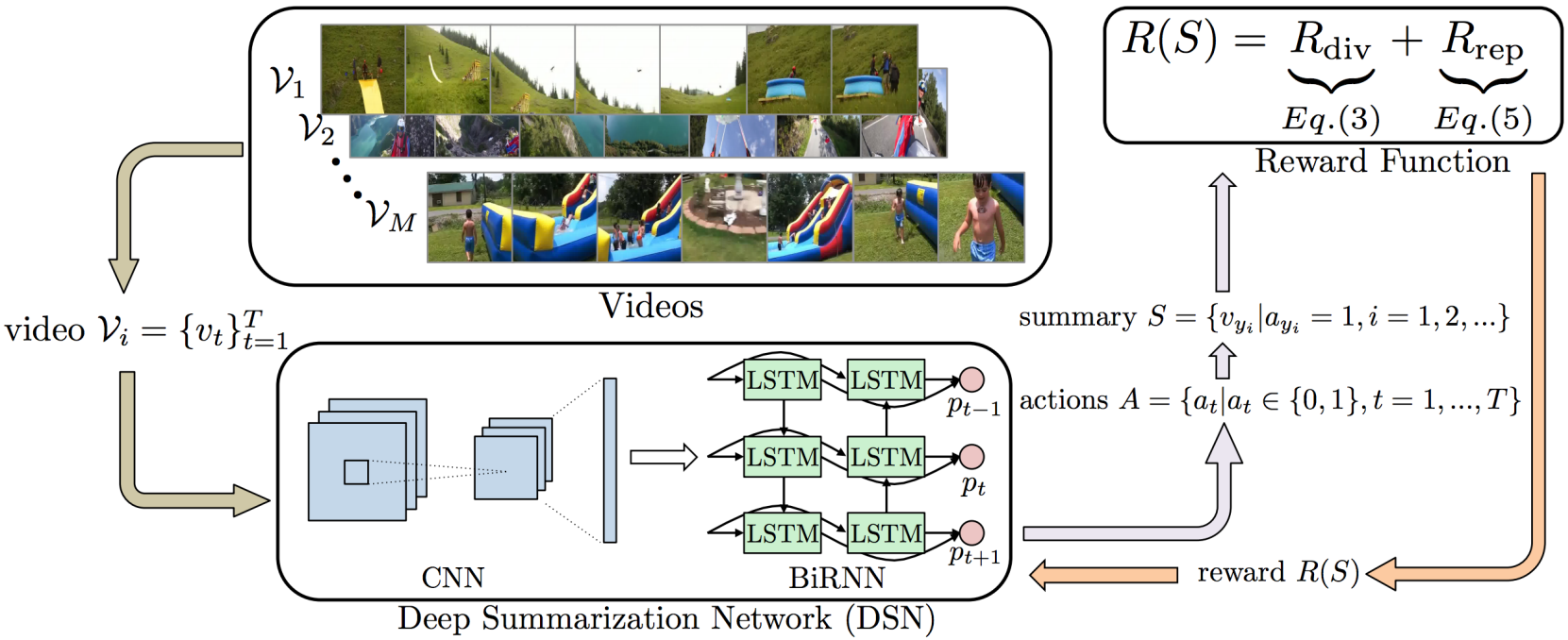
Representativeness reward

$$R_{REP} = \exp(-\frac{1}{T} \sum_{t=1}^{T} \min_{t' \in \mathcal{Y}} ||x_t - x_{t'}||_2)$$

Diversity-representativeness reward

$$R = R_{DIV} + R_{REP}$$

Архитектура сети



- CNN $\rightarrow x_t$ – GoogLeNet обученная на ImageNet
- BiRNN $\rightarrow h_t$
- FC $\rightarrow p_t = \sigma(Wh_t)$
- $a_t \sim Bernoulli(p_t)$

Reinforce

$$J(\theta) = \mathbb{E}_{p_\theta(a_{1:T})}[R(S)]$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{p_\theta(a_{1:T})}\left[R(S)\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|h_t)\right]$$

– многомерное матожидание

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

– большая дисперсия

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

b – baseline, средняя награда

Регуляризации

$L_{percentage} = ||\frac{1}{T}\sum_{t=1}^{T} p_t - \varepsilon||$ – минимизируем, ε – желаемый процент видео в итоговом summary

$L_{weight} = \sum_{i,j} \theta_{i,j}^2$

$$\theta = \theta - \alpha\nabla_\theta(-J + \beta_1 L_{percentage} + \beta_2 L_{weight})$$

Случай supervised learning

Можно добавить информацию из разметки
$\mathcal{Y}* = y_i*$
$L_{MLE} = \sum_{t \in \mathcal{Y}*} \log p(t, \theta)$ – максимизируем Можно совсем не использовать RL и обучать DSN, используя cross-entropy loss (это неэффективно)

Эксперименты

Датасеты

- SumMe – 25 видео, разные темы, от 1 до 6 минут
- TVSum – 50 видео; новости, документальные видео и т.п.; от 2 до 10 минут
- OVP – 50 видео
- YouTube – 39 видео

Асессоры размечали frame-level importance score.
В качестве метрики используется F-score.

$$P = \frac{\text{overlapped duration of A and B}}{\text{duration of A}}, R = \frac{\text{overlapped duration of A and B}}{\text{duration of B}}$$

$$F = 2P \times R/(P + R) \times 100\%$$

Эксперименты

- 5FCV, 80% – обучающая выборка
- 5FCV, 80% + OVP + YouTube– обучающая выборка
- 5FCV, Остальные три датасета – обучающая выборка

Детали реализации

- Используются только 2 кадра в секунду
- $\lambda = 20$
- $\varepsilon = 0.5$
- $N = 5$
- α, β_1, β_2 выбираются по CV

Результаты unsupervised

Table 1: Results (%) of different variants of our method on SumMe and TVSum.

Method	SumMe	TVSum
DSN_{sup}	38.2	54.5
$\text{D-DSN}_{\text{w/o } \lambda}$	39.3	55.7
D-DSN	40.5	56.2
R-DSN	40.7	56.9
DR-DSN	41.4	57.6
$\text{DR-DSN}_{\text{sup}}$	**42.1**	**58.1**

Table 2: Results (%) of unsupervised approaches on SumMe and TVSum. Our DR-DSN performs the best, especially in TVSum where it exhibits a huge advantage over others.

Method	SumMe	TVSum
Video-MMR	26.6	-
Uniform sampling	29.3	15.5
K-medoids	33.4	28.8
Vsumm	33.7	-
Web image	-	36.0
Dictionary selection	37.8	42.0
Online sparse coding	-	46.0
Co-archetypal	-	50.0
GAN_{dpp}	39.1	51.7
DR-DSN	**41.4**	**57.6**

Результаты supervised

Table 3: Results (%) of supervised approaches on SumMe and TVSum. Our DR-DSN$_{sup}$ performs the best.

Method	SumMe	TVSum
Interestingness	39.4	-
Submodularity	39.7	-
Summary transfer	40.9	-
Bi-LSTM	37.6	54.2
DPP-LSTM	38.6	54.7
GAN$_{sup}$	41.7	56.3
DR-DSN$_{sup}$	**42.1**	**58.1**

Table 4: Results (%) of the LSTM-based approaches on SumMe and TVSum in the Canonical (C), Augmented (A) and Transfer (T) settings, respectively.

Method	SumMe			TVSum		
	C	A	T	C	A	T
Bi-LSTM	37.6	41.6	40.7	54.2	57.9	56.9
DPP-LSTM	38.6	42.9	41.8	54.7	59.6	58.7
GAN$_{dpp}$	39.1	43.4	-	51.7	59.5	-
GAN$_{sup}$	41.7	43.6	-	56.3	**61.2**	-
DR-DSN	41.4	42.8	42.4	57.6	58.4	57.8
DR-DSN$_{sup}$	**42.1**	**43.9**	**42.6**	**58.1**	59.8	**58.9**

Литература

Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward
https://arxiv.org/pdf/1801.00054.pdf