Video Summarization Using Reinforcement Learning with Attention

Lakshya
2017eeb1149@iitrpr.ac.in
Indian Institute of Technology Ropar

Abstract

Video summarization techniques aim to facilitate efficient and fast video browsing by generating summaries representing the video content. This paper will formulate video summarization as a parallel decision-making process using a self-attention-based regressor that outputs importance scores for each frame. These scores are then used to generate video summaries. To train our model, we follow a Reinforcement Learning framework that maximizes a diversity-representativeness reward. Since no labels are required, the method is fully unsupervised. Further, we employ a unique rank comparison evaluation method to evaluate the generated summaries. Experimentation reveals that our method results in faster convergence to a higher reward compared to the base method.

Keywords: Reinforcement learning, Attention, Video summarization

1 Introduction

There has been a surge in the development of various video summarization techniques due to the increasing availability of video content online. These approaches comprise of both supervised and unsupervised settings. Recently, RNN models employing LSTM cells and GRU cells have been the main ingredient for these approaches to represent temporally organized information. DR-DSN [10] and DPP-LSTM [9] are some of the most recognized approaches that make use of RNN models. DR-DSN formulates video summarization as a sequential decision-making process using an encoder-decoder format. The encoder is a convolutional neural network used for feature extraction from the video frames, and the decoder is a bidirectional LSTM network responsible for producing probabilities of selecting a frame in summary. VASNet [2] is a supervised approach that employs a self-attention mechanism instead of an RNN model. This results in faster computation compared to RNN approaches when attention span is less than the feature dimensions. They show that the self-attention mechanism performs better than the BiRNN approach in a supervised setting.

Otani et al. [5] conduct a study drawing attention to the drawbacks of utilizing the F1 score metric for comparing summaries, followed by proposing a rank comparison-based approach suited to evaluations where the existence of unique ground truth is not possible.

Our model has been greatly inspired by DR-DSN, where we construct the decoder as a self-attention model inspired by VASNet. To enable stricter comparisons, we do not alter the reward modeling presented in DR-DSN. We further draw comparisons between our method and DR-DSN using both the F1 scores and rank comparison evaluation criteria.

Unsupervised approaches are better suited to this task since the quality of a generated summary is highly subjective and does not enable a unique ground truth label. Furthermore, labeling data for video summaries is a very inefficient and arduous task. Reinforcement Learning aims to optimize an agent's action mechanism by forcing the agent to take actions that result in higher rewards. We carry out the policy gradient updates after the summary generation because we want the evaluation to be based on the full generated summary rather than the summary generated thus far. The temporal order in frame importance scoring is not as crucial for this application as it is for other sequential tasks such as NLP. Hence, we do not include temporal information in the self-attention mechanism.

As per our knowledge, our method is the first to utilize a self-attention based approach in conjunction with reinforcement learning in an unsupervised setting.

2 Methodology

2.1 Computing Features

The videos are broken down into individual frames and each frame is passed through GoogLeNet [7] pretrained on ImageNet dataset [1] to generate feature vectors
$X = \{x_0, \ldots, x_N\}$, $x_i \in \Re^D$.

2.2 Generating Frame Scores

Unlike RNN based networks, Attention methods do not need to deploy complex architectures such as BiLSTM to achieve non-causal behavior. Also, unlike RNN techniques, which suffer from short term memory, attention methods do not face this issue since any time step from the input is readily accessible. The methodology for this segment is precisely as defined in [2].

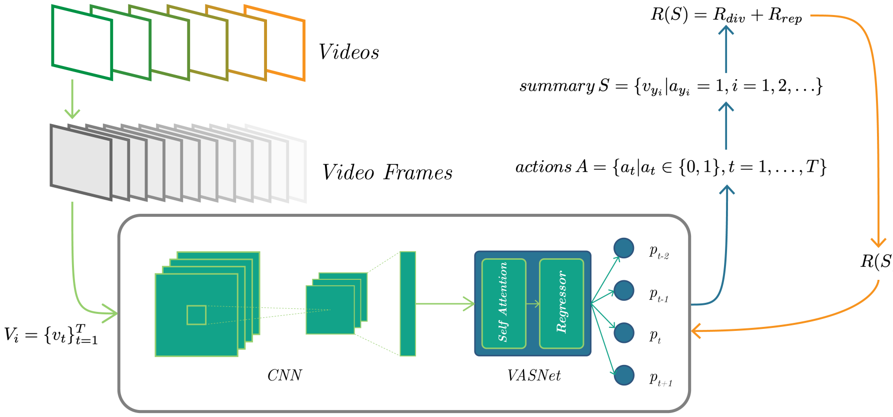
$R(S) = R_{div} + R_{rep}$
Videos
$summary\ S = \{v_{y_i} | a_{y_i} = 1, i = 1, 2, \ldots\}$
Video Frames
$actions\ A = \{a_t | a_t \in \{0, 1\}, t = 1, \ldots, T\}$
Self Attention
Regressor
p_{t-2}
p_{t-1}
p_t
p_{t+1}
$R(S)$
$V_i = \{v_t\}_{t=1}^T$
CNN
VASNet

Figure 1. Training VASNet using Reinforcement Learning. VASNet receives a video V_i and takes action A on which parts of the video are selected as the summary S. The feedback reward $R(S)$ is computed based on the quality of the summary - diversity and representativeness

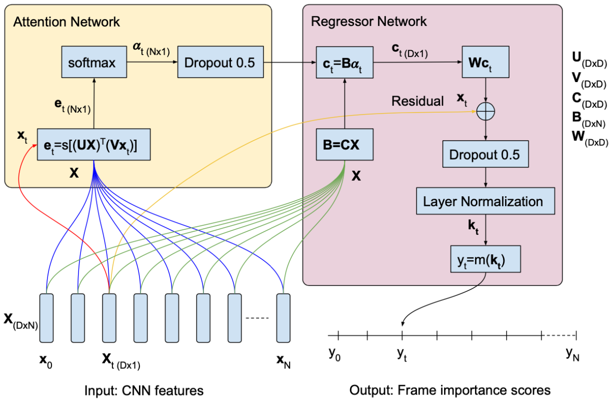

Figure 2. Self-Attention and Regressor modules of VASNet Architecture [2]

Unnormalized self attention weights $e_{t,i}$ and normalized self attention weights $\alpha_{t,i}$

$$e_{t,i} = softmax[(Ux_i)^T(Vx_t)] \quad t = [0, N), i = [0, N) \tag{1}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{N} \exp(e_{t,k})} \tag{2}$$

Transformation C and context vector c_t

$$b_i = Cx_i \tag{3}$$

$$c_t = \sum_{i=1}^{N} \alpha_{t,i} b_i \qquad c_t \in \Re^D \tag{4}$$

$$k_t = norm(dropuout(Wc_t + x_t)) \tag{5}$$

$$y_t = m(k_t) \tag{6}$$

where m is a two layer neural network that performs the frame score regression

2.3 Generating Rewards

During training, the VASNet will receive a reward $R(S)$ for each summary S generated. The objective is to maximize this reward. The reward needs to be modeled such that good summaries result in high reward. Zhou et al.[10] proposed a novel reward function that attempts to qualitatively represent diversity and representativeness of summaries. They argue that good summaries are made up of diverse segments and in those diverse segments, the frames present in the summary are the ones that are the most representative of those segments. Mathematically they formulate the reward as

$$R(S) = R_{div} + R_{rep}$$

where R_{div} is the diversity reward and R_{rep} is the representativeness reward. Let indices of selected frames be represented as $Y = \{y_i | a_{y_i} = 1, i = 1, \ldots, |Y|\}$ These rewards are formulated as:

$$R_{rep} = \exp(-\frac{1}{T} \sum_{t=1}^{T} \min_{t' \in Y} \|x_t - x_{t'}\|_2) \tag{7}$$

$$R_{div} = \frac{1}{|Y|(|Y| - 1)} \sum_{t \in Y} \sum_{\substack{t' \in Y \\ t' \neq t}} d(x_t, x_{t'}) \tag{8}$$

$$d(x_t, x_{t'}) = 1 - \frac{x_t^T x_{t'}}{\|x_t\|_2 \|x_{t'}\|_2} \tag{9}$$

to avoid temporally distant frames from being classified as similar (which is an important aspect of storytelling), $d(x_t, x_{t'})$ is set to 1 if $|t - t'| > \lambda$ where λ controls the degree of temporal difference. This also helps reduce the computational complexity of R_{div}. R_{rep} is formulated as a k-medoids problem.

2.4 Training with Policy Gradient

The goal of the policy network is to learn the policy function π_θ with parameters θ by maximizing the expected rewards

$$J(\theta) = \mathbb{E}_{p_\theta(a_1:T)}[R(S) \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|x_t)] \tag{10}$$

where a_t is the action taken at time step t and x_t is the input feature vector. Using the REINFORCE algorithm [8], we approximate the gradient of the expected reward$J(\theta)$ with respect to θ as

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} (R_n - b) \nabla_\theta \log \pi_\theta(a_t|x_t) \tag{11}$$

where b is the baseline which makes convergence faster by reducing variance. b is computed as the moving average of rewards experienced so far.

2.5 Regularization and Optimization

As per [10], to limit the number of selected frames, a regularization term $L_{percentage}$ is enforced as

$$L_{percentage} = \|\frac{1}{T} \sum_{t=1}^{T} p_t - \epsilon\|^2 \tag{12}$$

where ϵ determines the percentage of frames to be selected. the $l2$ regularization term to prevent overfitting is formulated as

$$L_{weight} = \sum_{i,j} \theta_{i,j}^2 \tag{13}$$

The policy function's parameters are optimized via stochastic gradient descent as

$$\theta = \theta - \alpha \nabla_\theta(-J + \beta_1 L_{percentage} + \beta_2 L_{weight}) \tag{14}$$

where α is the learning rate and β_1, β_2 are hyperparameters. The log probability of actions that lead to an increase in reward is increased while the probability of actions that negatively affected the reward is decreased.

2.6 Summary Generation

KTS [6] is used for temporal segmentation of videos into shots. These shots are assigned scores that are average of frame-level scores for frames in those shots. Shots are selected such that the total score accumulated is maximized with a 15% summary length limit, which is essentially a 0/1 Knapsack problem.

Method	F1 Score on SumMe
DPP-LSTM	0.43
DR-DSN	0.41
Randomized Test	0.41
Human	0.54

Table 1. The max F1 measures for SumMe benchmark as reported in recent works using KTS. It can be noted that random summaries achieve comparable results to the state-of-the-art and even to human annotations. [5]

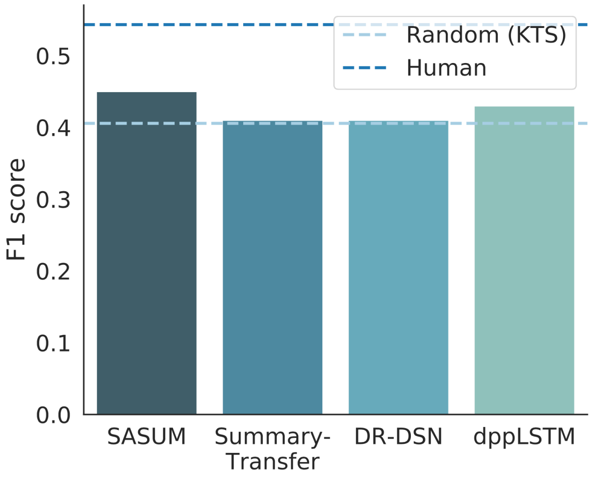

Figure 3. Recently reported F1 scores for methods using KTS segmentation in SumMe. The average score for random summaries with KTS segmentation is rep- resented by a light blue dashed line.[5]

3 Experimentation and Results

We evaluate our method on the SumMe dataset [3]. This dataset comprises 25 diverse user videos, each video ranging from 1 minute to 6 minutes in length and annotated by 15 to 18 humans. We used a 20% holdout set for testing.

3.1 Evaluation Metrics

As per convention and for the sake of uniformity, most methods use the F1 score for analyzing their generated summaries. However, as Otani et al. [5] point out, this method is flawed for this application. Their experimentation shows that most state of the art methods for video summarization score similar to random shot selection. Random selection is supposed to serve as a baseline for these methods, but F1 scores do not serve well for this purpose. Furthermore, since it is not reasonable to expect a singular ground truth for the generated summaries, this problem is further amplified since even human to human score comparisons fare poorly when their F1 scores are compared.

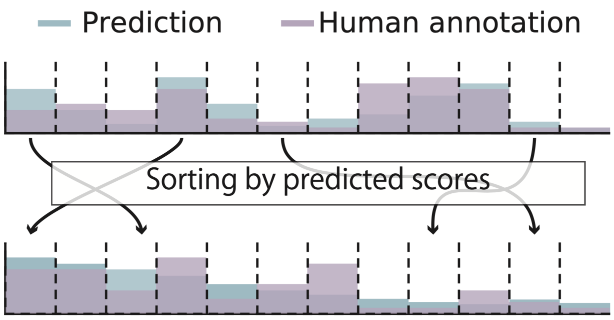

Figure 4. Ordering frame level scores with the monotonically decreasing order of predicted scores [5]

Method	F1	Kendall's τ	
	predicted	predicted	human
DR-DSN	0.395	0.228	0.279
Ours	0.3743	0.225	0.279

Table 2. Average of F1 score on predicted and Kendall's Tau on predicted and human annotated scores compared between our method and DR-DSN

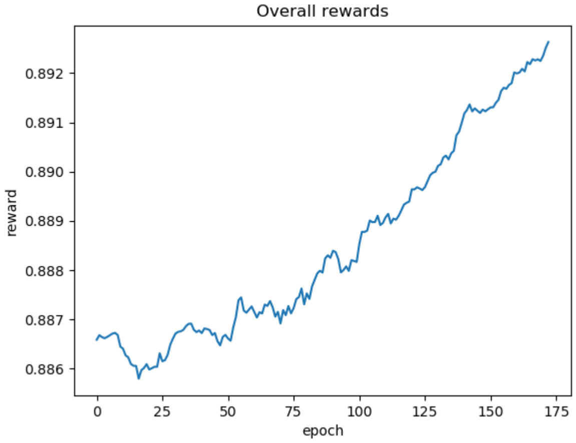

Figure 5. Reward curve for DR-DSN during training for 180 epochs.

To tackle this issue, we use frame score rank order comparison as suggested in [5] in addition to F1 scores. We first arrange the frame scores in monotonically decreasing order of their predicted scores, followed by comparing the ranking of predicted scores in this order to the ranking of human-annotated scores in this order. For this comparison, we employ the usage of Kendall's τ [4]. In the case of multiple human-annotated scores, we take the average of their respective τ values.

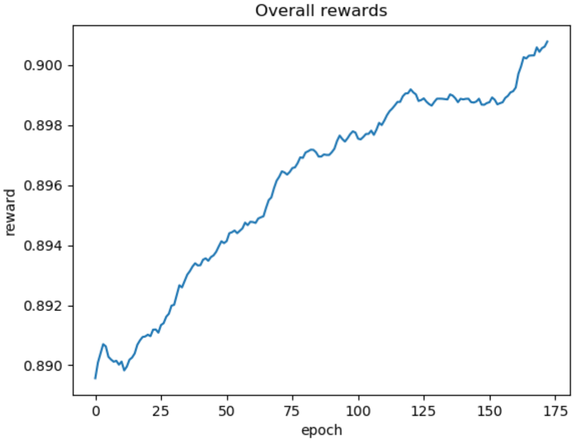

Figure 6. Reward curve for Our method during training for 180 epochs. Faster increase in gained reward by our model when compared to DR-DSN is visible. The final reward is more than that of DR-DSN and the total reward accumulated during training is more than twice that done by DR-DSN

We can observe that Our model trains twice as fast and accumulates twice as much reward compared to DR-DSN. It is interesting to note that our model does not perform better than DR-DSN, even after achieving a greater reward. We hypothesize that this issue occurs due to the imperfect nature of reward modeling.

4 Summary and Future Work

In this work, we developed, as per our knowledge, the first video summarization method that utilizes self-attention in a reinforcement learning framework in an unsupervised setting. We can achieve performance that is comparable to our base paper and can achieve higher rewards. Furthermore, we discuss and employ an alternate form of comparison metric called frame score rank comparison for video summarization methods. We plan to improve the reward modeling in further work and focus on refining the hyperparameters.

The code for implementation and experimentation can be found at the following link: vsumm-reinforce-attention

Acknowledgments

I would like to express my deep gratitude to Dr. Subrahmanyam Murala for providing valuable insights and guidance for this project.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[2] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. Summarizing videos with attention. In *Asian Conference on Computer Vision*. Springer, 39–54.

Method	Parameters	train time (min:sec)	final reward	Δ reward
DR-DSN	2.62605M	16:13	0.8939	0.0070
Ours	7.34823M	08:51	0.9010	0.0146

Table 3. Comparison of various parameters of our model with DR-DSN

[3] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating Summaries from User Videos. In *ECCV*.

[4] Maurice Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* (1938).

[5] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. 2019. Rethinking the evaluation of video summaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7596–7604.

[6] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. 2014. Category-Specific Video Summarization. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 540–555.

[7] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[8] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.

[9] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *European conference on computer vision*. Springer, 766–782.

[10] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2017. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. *arXiv preprint arXiv:1801.00054* (2017).