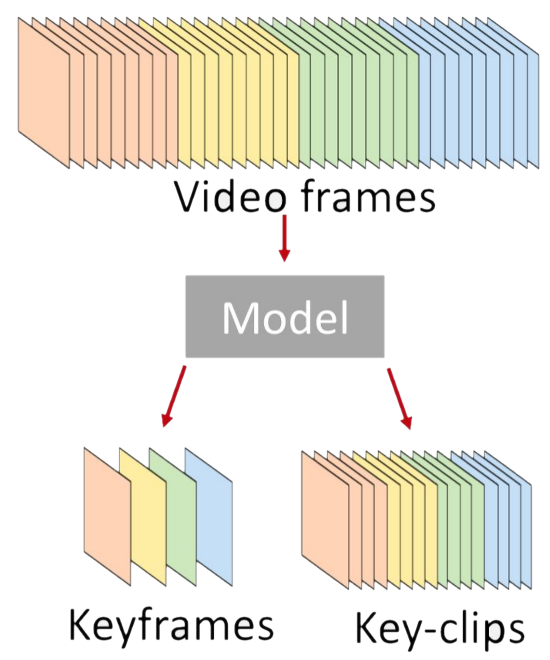
Spoiler-Sensitive Video Summarization With Reinforcement Learning

Justin Chou   Mayank Saxena
Columbia University

Video Summarization Without Spoilers

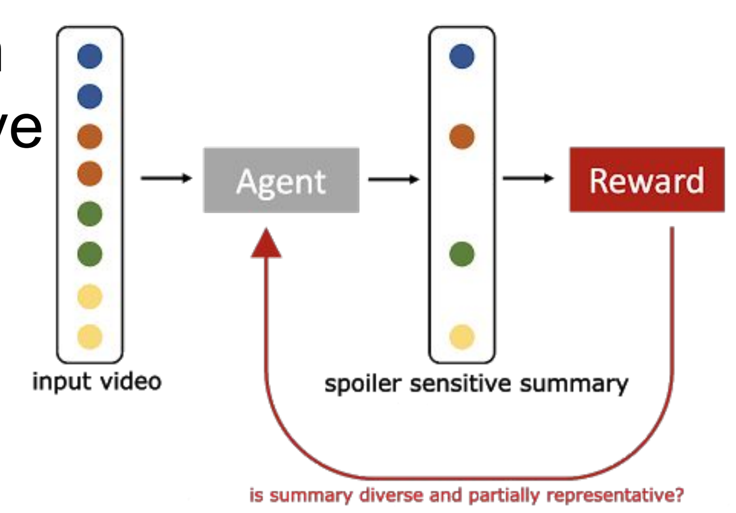
Existing Approaches:

- most industry practices are manual and labor-intensive
- very limited video summarization models that account for spoilers
- spoiler-sensitive unsupervised learning models publicly published for horror films only



Problems with Supervision:

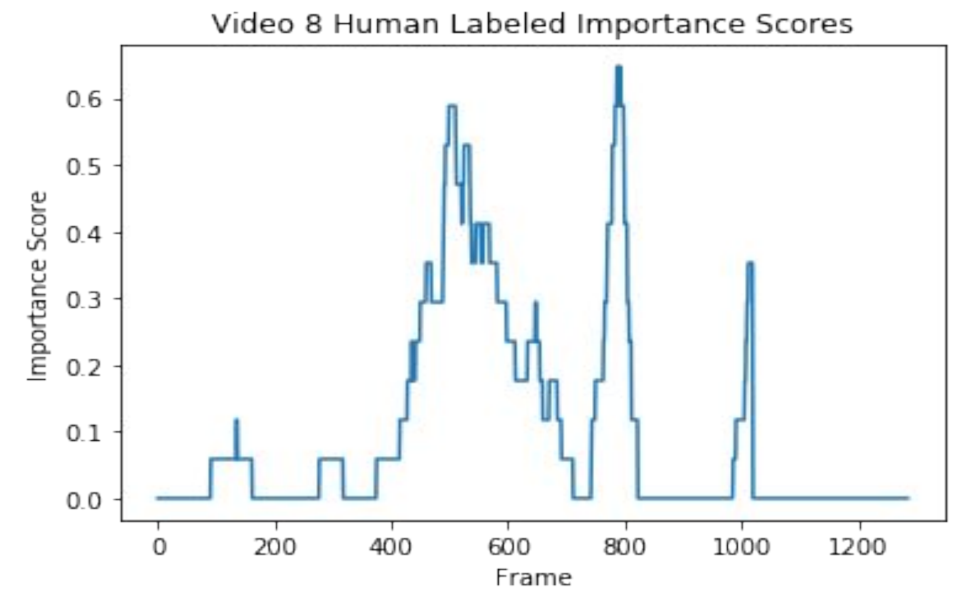
- no single best answer, human labels are costly and subjective
- different genres of video require different models
- human summarization can be mimicked using Markov decision processes

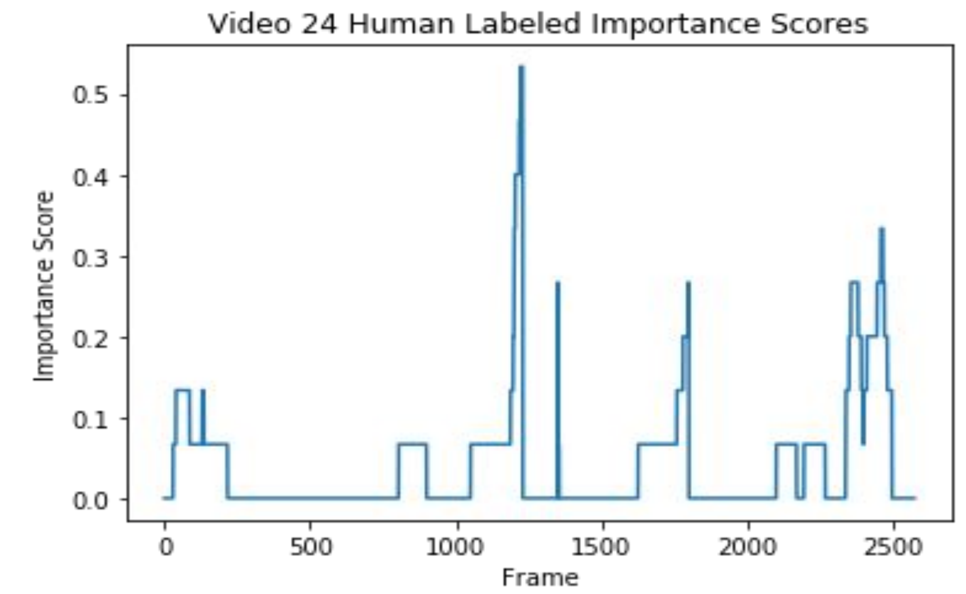


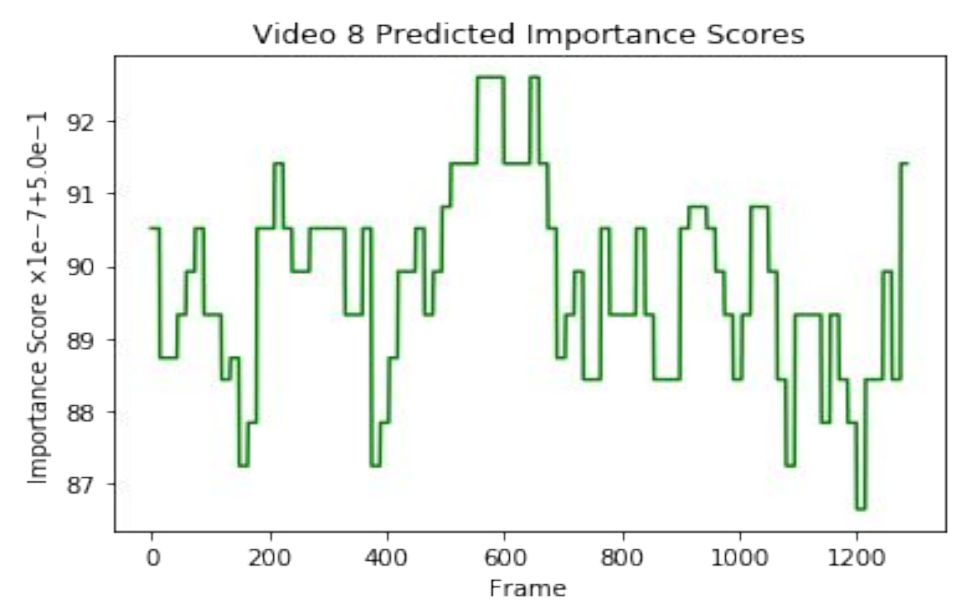
Results

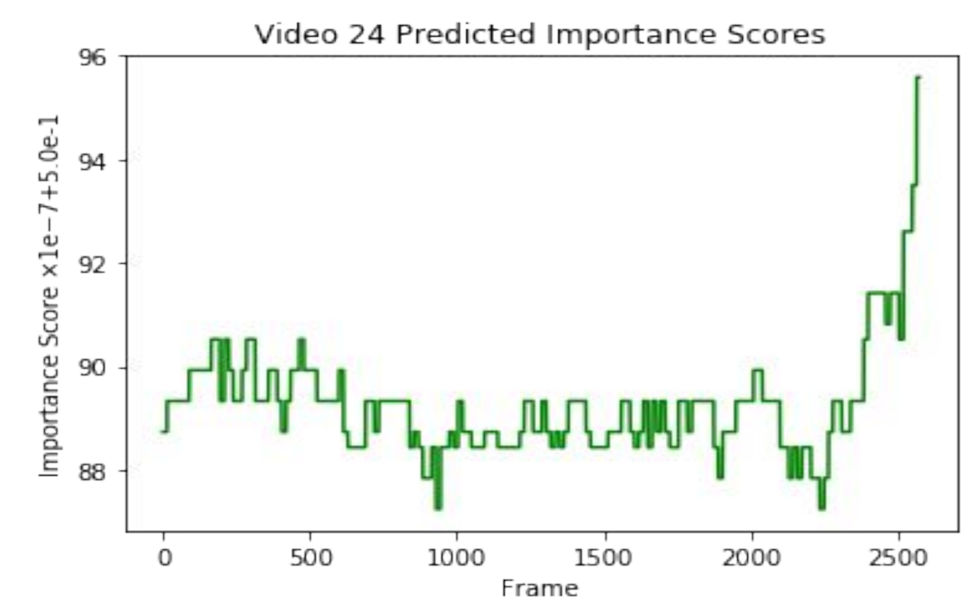
Video	F-Score	
	Ours	DR-DSN
video_8	48.5	12.2
video_24	39.5	23.6
video_17	33.2	13.9
video_2	19.6	0.0
video_12	0.0	0.0

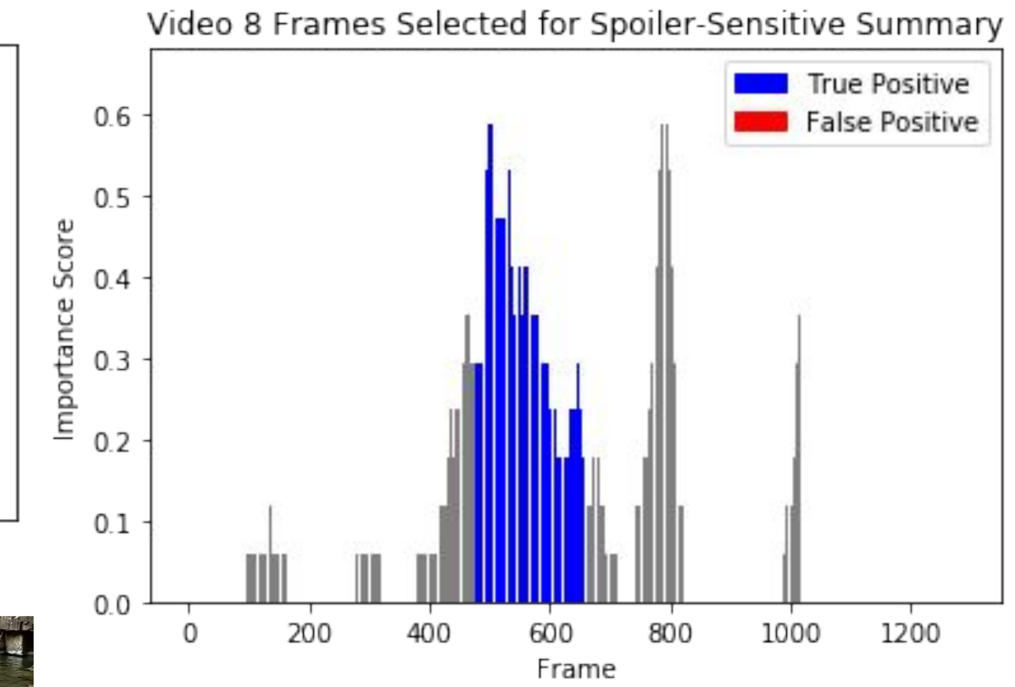
Video	Cross-Correlation	
	Ours	DR-DSN
video_8	58.37	61.94
video_24	48.92	51.58
video_17	314.32	332.66
video_2	244.65	258.67
video_12	45.12	47.74

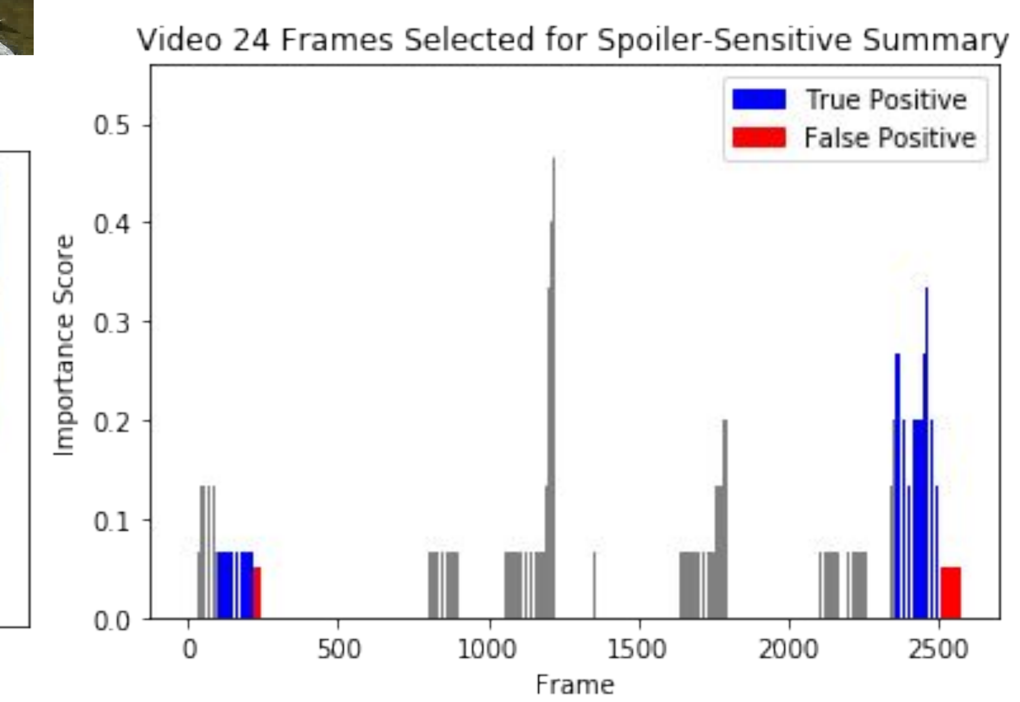








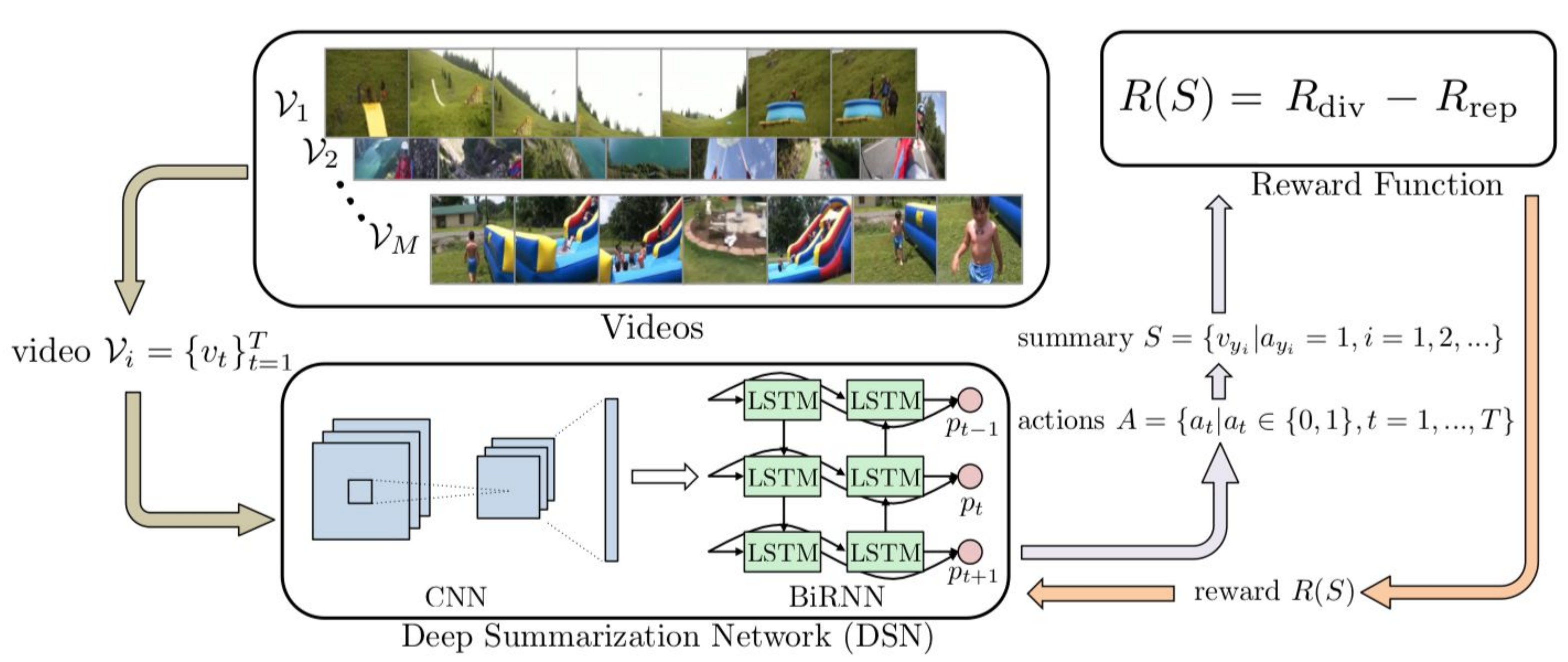




Our Approach

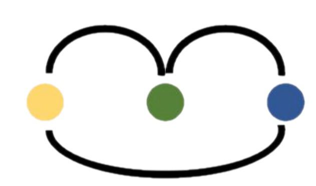
- model video summarization as a decision making process
- develop a deep summarization network to predict probabilities for video frames and make decisions for which frames to include in summary output
- diversity-representativeness reward function to assess video summary and spoiler-sensitivity

Deep Summarization Network

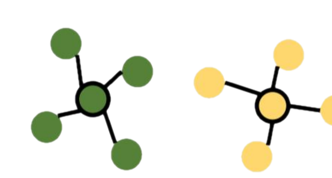


Reward Function $R(S) = R_{\text{div}} - R_{\text{rep}}$

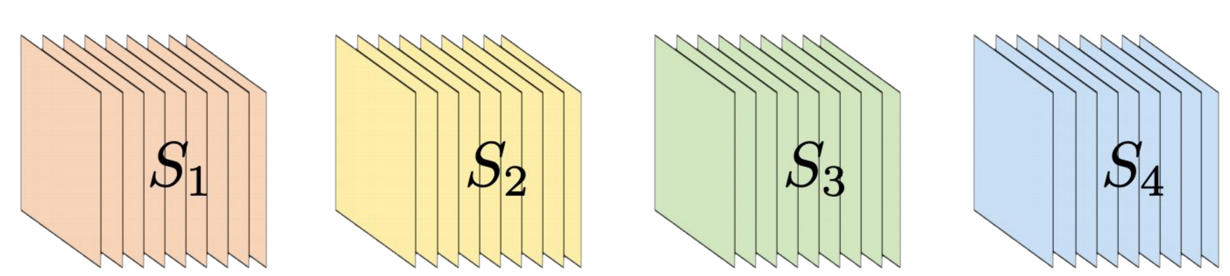
$$R_{\text{div}} = \frac{1}{|\mathcal{Y}|(|\mathcal{Y}|-1)} \sum_{t \in \mathcal{Y}} \sum_{\substack{t' \in \mathcal{Y} \\ t' \neq t}} d(x_t, x_{t'})$$



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



Inference



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

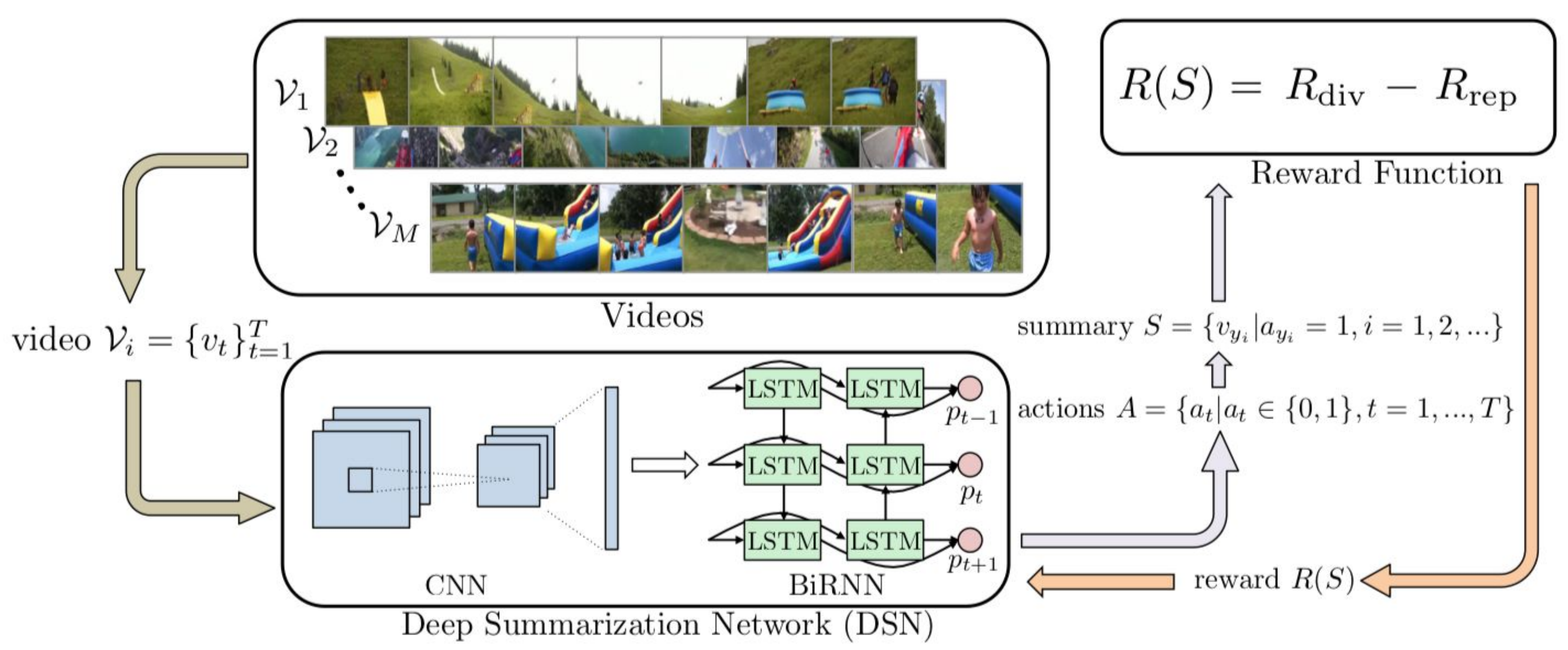
Clip-level scores $I(S_k) = \frac{1}{|S_k|} \sum_{i \in S_k} p_i$

Conclusion and Future Work

- label-free diversity-representativeness reward function is used to train a model for spoiler-sensitive video summarization
- different reward functions may better train the agent to select important frames that do not contain any spoilers
- multimodal signals (eg: audio, chat) can be incorporated for more accurate feature extraction for individual frames
- conduct user studies and experiments to better label positive frames for evaluation of generated summaries
- F-score may be a poor metric for video summarization[2] as state-of-the-art video summarization models only achieve an average F-score of about 40%, explore other metrics like correlation
- application towards generating spoiler free movie trailers

Acknowledgements

[1] Zhou, Kaiyang, Yu Qiao, and Tao Xiang. "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward." *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
[2] Otani, Mayu, et al. "Rethinking the Evaluation of Video Summaries." *arXiv preprint arXiv:1903.11328* (2019).



Reward Function $R(S) = R_{\text{div}} - R_{\text{rep}}$

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

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