Adversarial Unsupervised Video Summarization Augmented with Dictionary Loss

Michail Kaseris, Ioannis Mademlis and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

Horizon 2020 European Union funding for Research & Innovation

Poster 2759

Introduction

- Video data is the dominant and most widespread source of visual information.
- Navigating through vast databases of video content could be cumbersome for the end user as it is time-consuming to find a video that matches their taste.
- A compact *summary* of the original video content must be created in order to meet the demands of indexing enormous amounts of videos.
- The high cost of annotating the training data, the increased risk of overfitting to the specific training videos, as well as the subjective nature of the task, which renders it difficult to obtain a satisfying "ground-truth" summary, have led recently to unsupervised deep neural approaches.
- Unsupervised methods typically consist of LSTMs [4], including autoencoder and discriminator modules, in the context of a unified architecture [6].
- An LSTM *selector network* maps each input video frame representation (typically obtained by a pretrained CNN) onto a normalized scalar importance score, which determines whether it is selected as a key-frame or not.
- Assuming that a good summary must be able to reconstruct the original full/complete input sequence, we add an additional novel *dictionary loss* term during training, which directly penalizes the difference of the fixed-length *summary representation* (final hidden state of the LSTM autoencoder's encoding component) from a similar fixed-length *original sequence representation*.

Baseline architecture description

- In an adversarial framework for unsupervised video summarization by key-frame extraction [6] the generator is typically replaced by a *summarizer* which is fed CNN-derived video frame representations.
- Both the summarizer and the discriminator are LSTM networks.
- The full/original/complete input video sequence is represented by a matrix $\mathbf{X} \in \mathbb{R}^{M \times T}$, where T is the total number of video frames and M the dimensionality of each video frame.
- Each column $\mathbf{x}_t \in \mathbb{R}^M, , t = 1, ..., T$ of $\mathbf{X}$ is a video frame representation, extracted using a pretrained CNN.
- The columns of $\mathbf{X}$ are successively fed to the summarizer, which is composed of three successive LSTM subnetworks, each one unfolding across T time instances: a *selector*, an *encoder* and a *decoder*.
- The selector output is a real vector $\mathbf{s} \in [0, 1]^T$, with each entry of $\mathbf{s}$ reflecting the suitability of the corresponding input video frame as a key-frame.
- Each scalar product $s_t \mathbf{x}_t$ is fed to the encoder, which gradually generates a fixed-length representation of the summary $\mathbf{e} \in \mathbb{R}^H$, where H is the LSTM hidden state dimensionality.
- $\mathbf{e}$ is then fed to the decoder which also unfolds across T time instances.
- The decoder outputs a reconstructed video sequence $\hat{\mathbf{X}} \in \mathbb{R}^{M \times T}$.
- The columns of $\hat{\mathbf{X}}$ are subsequently fed into the discriminator, which is a binary LSTM classifier being optimized to distinguish between original videos ("positive examples") and their summary-based reconstructions ("negative examples").
- During training of the overall architecture, several loss functions are concurrently minimized by different neural components, employed in [1]:
- **Reconstruction loss**: $\mathcal{L}_{recon} = \|\phi(\mathbf{X}) - \phi(\hat{\mathbf{X}})\|_2^2$. $\mathcal{L}_{recon}$ is used to update θ_s, θ_e and θ_d.
- **Original video loss**: $\mathcal{L}_{orig} = (1 - C(\mathbf{X}))^2$, which is the MSE between the original video label (i.e., 1) and the discriminator output for original video input. $\mathcal{L}_{orig}$ is used to update θ_c.
- **Summary loss**: $\mathcal{L}_{sum} = \left(C(\hat{\mathbf{X}})\right)^2$, which is the MSE between the summary label (i.e., 0) and the discriminator output for summary-based reconstructed video input. $\mathcal{L}_{sum}$ is used to update θ_c.
- **Generator loss**: $\mathcal{L}_{gen} = \left(1 - C(\hat{\mathbf{X}})\right)^2$, which is the MSE between the original video label (i.e., 1) and the discriminator output for summary-based reconstructed video input. $\mathcal{L}_{gen}$ is used to update θ_d.
- **Sparsity loss**: $\mathcal{L}_{sparsity} = \|\frac{1}{T}\sum_{t=1}^{T} s_t - \sigma\|_2$ is a diversity-inducing regularizer used to update θ_s. Hyperparameter σ represents the desired percentage of original video frames to be retained in the summary.

Proposed Dictionary Loss

- Building upon the baseline framework [1], the proposed method adds a complementary neural component which is only employed during training, i.e., an LSTM autoencoder that also unfolds across T time instances and consists in a LSTM encoder-decoder architecture.
- It successively receives all original video frame representations $\mathbf{x}_t$ as input, encodes the entire original sequence into a final hidden state $\mathbf{h} \in \mathbb{R}^N$ and subsequently decodes it to approximately reproduce the full original video, where N is the hidden state dimensionality of the parallel autoencoder.
- The reasoning behind the addition of the parallel autoencoder into the overall framework for obtaining a fixed-length representation of the original video, was that the existing $\phi(\mathbf{X})$ which is employed for computing the main reconstruction loss is constructed by the discriminator.
- It is a representation adapted to discriminate between the original input and the summary-based reconstruction that lacks compact information about the original video itself.
- $\mathbf{h}$ is exactly such an original sequence representation, obtained at each iteration of the summarizer training process as the final hidden state of the pretrained parallel encoder.
- The proposed *dictionary loss* $\mathcal{L}_{\text{dict}}$ as an additional training constraint for updating θ_s and θ_e, besides the traditional reconstruction loss, where $\mathcal{L}_{\text{dict}}$ makes use of the vector $\mathbf{h}$ and a common matrix $\mathbf{A}$.
- Our proposed cost function is inspired by the dictionary-of-representatives formulation of unsupervised video key-frame extraction [5].
- Given original input video $\mathbf{X} \in \mathbb{R}^{M \times T}$, the goal is to find an optimal summary matrix $\mathbf{S} \in \mathbb{R}^{M \times C}, C << T$ and a reconstruction coefficient matrix $\mathbf{B} \in \mathbb{R}^{C \times T}$, so that the columns of $\mathbf{S}$ constitute a subset of columns of $\mathbf{X}$ and the following objective is minimized:

$$\min_{\mathbf{S},\mathbf{B}} : \sum(\|\mathbf{X} - \mathbf{SB}\|)_n \tag{1}$$

- In our implementation, the proposed dictionary loss is defined as:

$$\mathcal{L}_{\text{dict}} = \|\mathbf{h} - \mathbf{Ae}\|_2 \tag{2}$$

- Matrix $\mathbf{A}$ transforms the current summary representation to a vector space being simultaneously learnt from all the original videos, therefore $\mathbf{A}$ serves as a global visual dictionary.

Empirical Evaluation

- For this setting, the videos were downsampled to 2 frames per second, the CNN-derived 1024-dimensional video frame representations were extracted from the pool5 layer of a GoogLeNet [8], pretrained on the ImageNet dataset.
- The hidden state of the involved LSTM modules is 500-dimensional.
- Empirical evaluation was conducted using the F-Score metric F and the commonly employed, public datasets SumMe [2] and TVSum [7].

Table 1: F-Score results of unsupervised video summarization methods in two public datasets.

Method	TVSum	SumMe
SUM-GAN-AAE [1]	58.3%	48.9%
vsLSTM [10]	54.2%	37.6%
dppLSTM [10]	54.7%	38.6%
Cycle-SUM [9]	57.6%	41.9%
ACGAN [3]	58.5%	46.0%
Proposed	**59.3%**	**51.0%**

Conclusions

- This paper presented a novel, differentiable loss function inspired by dictionary learning, which is added to the training process of a common adversarial neural video summarization framework.
- The proposed dictionary loss exploits a newly introduced, parallel LSTM autoencoder and biases key-frame selection towards video frames which are collectively able to recreate the original sequence.
- The method surpasses the state-of-the-art when evaluated on two common public relevant datasets, confirming our underlying hypothesis that the reconstructive ability plays a crucial role in key-frame selection.

References

[1] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras. Unsupervised video summarization via attention-driven adversarial learning. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*. Springer, 2020.

[2] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[3] X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan. Unsupervised video summarization with attentive conditional generative adversarial networks. In *Proceedings of the ACM International Conference on Multimedia*, 2019.

[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] I. Mademlis, A. Tefas, and I. Pitas. A salient dictionary learning framework for activity video summarization via key-frame extraction. *Information Sciences*, 432:319–331, 2018.

[6] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. TVSum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] L. Yuan, F. E.H. Tay, P. Li, L. Zhou, and J. Feng. Cycle-SUM: Cycle-consistent adversarial LSTM networks for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[10] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016.

Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911 (AI4Media).