Video semantic segmentation using deep multi-view representation learning

Akrem Sellami, and Salvatore Tabbone

LORIA laboratory

University of Lorraine, FRANCE

Abstract

In this paper, we propose a deep learning model based on deep multi-view representation learning, to address the video object segmentation task. The proposed model emphasizes the importance of the inherent correlation between video frames and incorporates a multi-view representation learning based on deep canonically correlated autoencoders. The multi-view representation learning in our model provides an efficient mechanism for capturing inherent correlations by jointly extracting useful features and learning better representation into a joint feature space, i.e., shared representation. To increase the training data and the learning capacity, we train the proposed model with pairs of video frames, i.e., F_a and F_b. During the segmentation phase, the deep canonically correlated autoencoders model encodes useful features by processing multiple reference frames together, which is used to detect the frequently reappearing. Our model enhances the state-of-the-art deep learning-based methods that mainly focus on learning discriminative foreground representations over appearance and motion. Experimental results over two large benchmarks demonstrate the ability of the proposed method to outperform competitive approaches and to reach good performances, in terms of semantic segmentation.

General context

Video object segmentation

- Extract spatio-temporal regions that correspond to objects moving in the video sequence
- Graph-based approaches
- Deep learning-based approaches

Motivations and goals

Motivations

- Existing models mainly focus on the intra-frame discrimination of primary objects in motion or appearance.
- They ignore the valuable global-occurrence consistency across multiple video frames.
- Recurrent neural networks (RNNs) fail to explore the rich relations, i.e., the high correlation between different video frames, hence do not attain a global perspective.

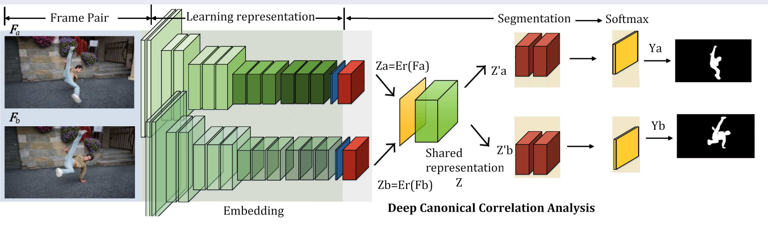
Goals

- Propose a video semantic segmentation model using deep multi-view representation learning to model video semantic segmentation task from a global view
- Capture the rich inherent correlations between all frames
- Improve the segmentation task

Proposed methodology

Multi-view deep representation learning

- Learn a better representation from pairs of frames, i.e, multimodal frames of a video by encoding their useful features in order to capture the inherent correlation between them



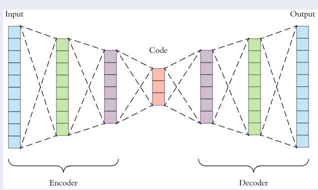
Multi-view representation learning

- Extract useful (relevant) features from multiple input modalities, i.e. pairs of video frames denoted by F_a and F_b, which may be reconstructed.

The AE model

- *Encoder:* Enc_{AE}
- *Bottleneck layer:* $z = Enc_{AE}(F_i)$
- *Decoder:* $Dec_{AE}(Enc_{AE}(F_i)) \approx F_i$
- *Training (MSE):*

$$\frac{1}{n \times m}\sum_{i=1}^{n}\sum_{j=1}^{m}\|Dec_{AE}(Enc_{AE}(F^{i,j})) - F^{i,j}\|^2$$



Multi-view representation learning

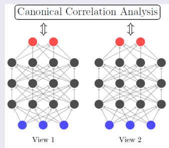
Deep Canonical Correlation analysis (DCCA)

- Find direction vectors $v_j, w_j, j \in \{1, ..., K\}$ that maximize the correlation between the projections $v_j^T Z_a$ and $w_j^T Z_b$ while being minimally redundant:

$$v_j w_j = \underset{v,w}{\arg\max}\, corr(v^T Z_a w^T Z_b)$$

such that $corr(v_j^T Z_a, v_k^T Z_a) = 0, k < j$

$$corr(w_j^T Z_b, w_k^T Z_b) = 0, k < j$$



Multi-view representation learning

Deep Canonically Correlated Autoencoders (DCCAE)

- Optimize the combination of correlation between the learned latent representations (bottleneck layer) and the reconstruction errors of the AEs.

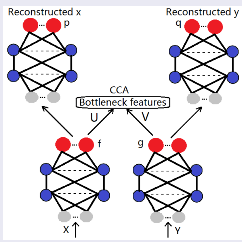
$$\min_{w_f, w_g, w_p, w_q, U, V} -\frac{1}{N} tr(U^T f(X) g(Y)^T V)$$

$$+\frac{\lambda}{N}\sum_{i=1}^{N}(||x_i - p(f(x_i))||^2) + (||y_i - q(g(y_i))||^2)$$

$$\text{s.t.}, U^T(\frac{1}{N} f(X) f(X)^T + r_x I) U = I$$

$$V^T(\frac{1}{N} g(Y) g(Y)^T + r_y I) V = I$$

$$u_i^T f(X) g(Y)^T v_j = 0, \forall i \neq j$$

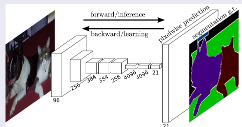


Semantic segmentation

Fully Connected Network (FCN)

- Each layer consists of three-dimensional data of size $H \times W \times C$, where W and H are spatial dimensions, and C is the channel dimension.
- The FCN compute outputs y_{ij} as follows:

$$y_{ij} = f_{ks}(\{x_{si} + \sigma_{s_i, s_j}\} \; 0 < \sigma_i, \sigma_j < k) \tag{1}$$



Experimental results

Data description

- *UAVid dataset:* consists of 30 video sequences. It is composed of 300 images and each of size 3840×2160 or 4096×2160. There are 8 classes that have been selected for semantic segmentation.
- *DAVIS16 dataset:* which consists of 50 videos in total. We select 30 videos for training and 20 for testing.

Evaluation on UAV dataset

IoU scores for different deep learning models

Model	Building	Tree	Clutter	Road	Low Vegetation	Static car	Moving car	Human	mean IoU
FCN-8s	64.3	63.8	33.5	57.6	28.1	8.4	29.1	0.0	35.6
Dilation Net	72.8	66.9	38.5	62.4	34.4	1.2	36.8	0.0	39.1
U-Net	70.7	67.2	36.1	61.9	32.8	11.2	47.5	0.0	40.9
MS-Dilation	74.3	68.1	40.3	63.5	35.5	11.9	42.6	0.0	42.0
Ours	76.1	71.3	43.2	65.7	36.1	12.1	45.8	0.0	43.78

Evaluation on DAVIS16 dataset

Results on the test set

	Method	COSNet [36]	SFL [22]	LMP [37]	FSEG [18]	UOVO [38]	ARP [39]	PDB [25]	Ours
$\mathcal{J}$	Mean	80.5	67.4	70.0	70.7	73.9	76.2	77.2	83.3
	Recall	93.1	81.4	85.0	83.0	88.5	91.1	90.1	98.2
	Decay	4.4	6.2	1.3	1.5	0.6	7.0	0.9	0.1
$\mathcal{F}$	Mean	79.5	66.7	65.9	65.3	68.0	70.6	74.5	80.3
	Recall	89.5	77.1	79.2	73.8	80.6	83.5	84.4	94.3
	Decay	5.0	5.1	2.5	1.8	0.7	7.9	-0.2	0.0
$\mathcal{T}$	Mean	18.4	28.2	57.2	32.8	39.0	39.3	29.1	31.2

- Region similarity $\mathcal{J} = \frac{|M \cap G|}{|M \cup G|}$
- Boundary accuracy $\mathcal{F} = \frac{2 P_c R_c}{P_c + R_c}$
- time stability $\mathcal{T}$