A Self-Paced Fine-Tuning Network for Segmenting Objects in Weakly Labelled Videos

Zhang, Liqiang

May 7,2018

1 Self-Paced Fine-Tuning

With a massive amount of videos can be easily accessed online, an exciting opportunity to learn visual concepts and object models is offered. However, it is hard to directly exploit these online videos in traditional ways because most of online videos are weakly labelled [2]. In order to tackle the aforementioned limitations, author propose a novel self-paced fine-tuning network (SPFTN) in this paper. As shown in Fig. 1, given a group of videos that are weakly labelled as containing common objects from one semantic category, the proposed approach first prepares training data by decomposing these videos into frames and generating segmentation proposals for these frames. [1]

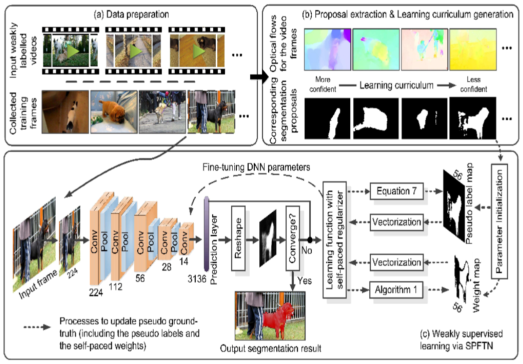
(a) Data preparation
Input weakly labelled videos
Collected training frames
...
(b) Proposal extraction & Learning curriculum generation
Optical flows for the video frames
Corresponding segmentation proposals
More confident
Learning curriculum
Less confident
Fine-tuning DNN parameters
Input frame
224
Conv
Pool
224
112
56
28
14
3136
Prediction layer
Reshape
Converge?
No
Yes
Learning function with self-paced regularizer
Equation 7
Vectorization
Vectorization
Algorithm 1
56
Pseudo label map
Weight map
Parameter Initialization
Processes to update pseudo ground-truth (including the pseudo labels and the self-paced weights)
Output segmentation result
(c) Weakly supervised learning via SPFTN

Figure 1: . The proposed self-paced fine-tuning network-based framework for object segmentation in weakly labelled videos.

2 Objective Function

Given a collection of K video frames $(I_k)_{k=1}^{K}$ extracted from a set of weakly labelled videos from one semantic category, the input dimension of the designed network architecture is set to be 244 ×244. With the input of the video frames and the initial **X** and **Y**, the learning objective gradually discovers confident training samples and use them to fine-tune DNN via mainly minimizing a weighted prediction loss term and a self-paced regularizer:

$$\begin{aligned} \underset{W,Y,V}{min}\ \mathbf{E(W,Y,X)} =& r(\mathbf{W}) + \sum_{k=1}^{K} + f(\mathbf{V};\mathbf{p},\lambda,\gamma,\tau,), \\ & s.t. V \in [0,1]^{d\times K}, \mathbf{p} \in [0,1]^{K}, \\ & \sum_{k} \parallel v_k \parallel_1 \in (0, d\times K), \\ & \sum_{k} \parallel y_k \parallel_1 \in (0, d\times K). \end{aligned} \tag{1}$$

Here r(.) indicates the squared ℓ_2 norm,**W** indicatesthe trainable parameters among the network, $\mathbf{V} = [v_1, v_2 ..., v_K]$ denotes the weight matrix which reflects the self-paced weights for all the pixels of the video frames.

$$\begin{aligned} L(y_k, v_k, &\Phi(I_k \mid \mathbf{W})) = \\ &\sum_{i=1}^{d} v_k^i max(1 - y_k^i \cdot \Phi(I_k \mid \mathbf{W}^i, 0)^2, \end{aligned} \tag{2}$$

Where (2) v_k^i, y_k^i, and $\Phi(I_k \mid \mathbf{W})^i$ indicate the i-th dimension of the weight vector v_k, pseudo label vector y_k, and prediction vector $\Phi(I_k \mid \mathbf{W})^i$, respectively.

3 Model Analysis

Different regularizers	IOU
OURS-GC: OURS w/o group curriculum	0.569
OURS-GC2: OURS w/o the second term in GC	0.564
OURS-GC1: OURS w/o the first term in GC	0.589
OURS with sample diversity term of [13]	0.583
OURS	**0.612**

Table 1: Evaluation of the self-paced regularizers on DAVIS.

The results reported in Table. 1 indicate that each of the regularization terms used in the proposed group curriculum regularizer can benefit the learning procedure [3], while simultaneously using both of them obtains more significant performance gain.

References

[1] Yoshua Bengio, Jér?me Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. *Journal of the American Podiatry Association*, 60(60):6, 2009.

[2] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, page 2261, 2012.

[3] Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu. Predicting eye fixations using convolutional neural networks. In *Computer Vision and Pattern Recognition*, pages 362–370, 2015.