Weakly Supervised Semantic Segmentation Using Web-Crawled Videos

Seunghoon Hong[1,3] **Donghun Yeo**[1] **Suha Kwak**[2] **Honglak Lee**[3] **Bohyung Han**[1]

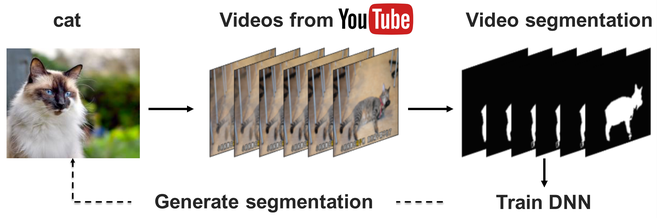
{maga33, hanulbog}@postech.ac.kr skwak@dgist.ac.kr honglak@umich.edu bhhan@postech.ac.kr

1 POSTECH 2 DGIST 3 University of Michigan

CVPR July 21-26 2017 HONOLULU

Motivation

Improving weakly-supervised semantic segmentation by generating synthetic segmentations from **web-crawled videos**



- **Benefits:**
  - ❑ Motion in videos is helpful to distinguish object from background
  - ❑ Videos are collected automatically by web search results
- **Challenges:**
  - ❑ Substantial noises in web-crawled videos
- **Our approach:**

  Exploit both weakly labeled images and videos to compensate segmentation challenge in one data from the other

Step 1. Learning encoder with images

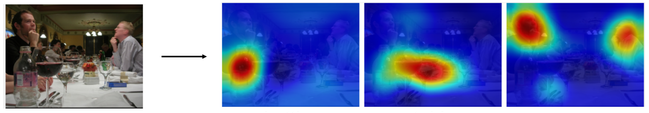
Train classifier and attention model with image-level labels

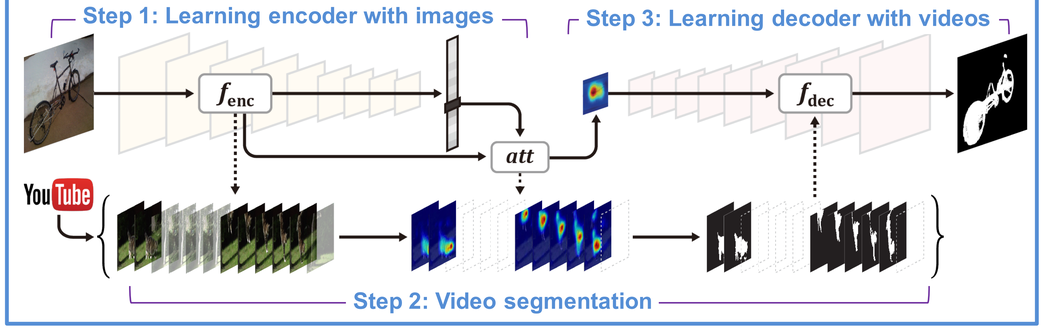
❑ **Classifier**: CNN with global average pooling (GAP) [Zhou et al. 2016]

❑ **Attention**: Class Activation Mapping [Zhou et al. 2016]

$$\alpha^c = F(x) \cdot W \cdot y^c$$

$F(x)$: last conv. output W: weights in fc layer y^c: onehot vector for class c





Step 2. Video segmentation with encoder outputs

Localizing object in a video using encoder outputs

❑ **Temporal localization** by filtering out irrelevant frames based on classification score

❑ **Spatial localization** by computing attention map that discriminates the target from surroundings

Video segmentation by energy minimization on spatio-temporal graph

$$L^* = \underset{L}{\operatorname{argmin}}\, E_u(L) + E_p(L)$$

❑ **Unary potential function**

$$E_u(L) = -\lambda_a \sum_i \log A_i(l_i) \quad \text{superpixel attention}$$

$$-\lambda_m \sum_i \log M_i(l_i) \quad \text{motion likelihood}$$

$$-\lambda_c \sum_i \log C_i(l_i) \quad \text{appearance likelihood}$$

❑ **Pairwise potential function**

$$E_p(L) = \sum_{(i,j)\in\mathcal{E}_s} [l_i \neq l_j] \cdot \phi_s(i,j) \cdot \phi_c(i,j) + \sum_{(i,j)\in\mathcal{E}_t} [l_i \neq l_j] \cdot \phi_t(i,j) \cdot \phi_c(i,j)$$

appearance similarity | spatial similarity | temporal connectivity

Step 3. Learning decoder with the video segmentation results

Train a decoder to map coarse attention map to dense binary mask

❑ **Decoder**: Deconvolution network with shared pooling switch [Noh et al. 2015]

❑ **Benefits:** 1. ***Class attention as input*** allows to ignore objects irrespective of the labeled class.
2. ***Class-agnostic property*** is useful to improve segmentation quality of static objects.

Experiments

Semantic segmentation on PASCAL VOC 2012 dataset

- **Training data:**
  - ❑ Image: PASCAL VOC 2012
  - ❑ Videos: 4.6K YouTube videos collected for 20 PASCAL VOC classes
- **Ablation study**

Method	Video set	mIoU (Val)
MCNN [Tokmakov et al. 2016]	YouTube-Obj.	38.1
Ours	YouTube-Obj.	49.2
	YouTube	**58.1**

❑ Separate training with images and videos improves performance

❑ Collecting more videos improves performance, although obtained videos are noisy and unannotated

- **Comparison to SOA weakly-supervised approaches**

Method	Supervision	mIoU (Val)
SEC [Kolesnikov et al.]	Class label	50.7
What's a Point [Bearman et al. 2016]	Point	46.0
BoxSup [Dai et al. 2015]	Bounding box	62.0
ScribbleSup [Lin et al. 2015]	Scribble	63.1
MCNN [Tokmakov et al. 2016]	Class label + Video	38.1
Ours	**Class label + Video**	**58.1**

❑ Substantial improvement over approaches based on image-level class labels

❑ Competitive performance to approaches based on heavier annotations (point, bounding box)

Video segmentation on YouTube-Object dataset

Unsupervised [2]	Bounding box [9]	**Ours (Class label)**
46.8	56.2	**58.6**

❑ Integrating attention substantially improves segmentation performance over approaches based on naïve motion

❑ Fine-grained attention is sometimes more helpful for segmentation than coarse detection outputs obtained from pre-trained object detector

Project page: http://cvlab.postech.ac.kr/research/weaksup_video