Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation

Tamerlan Tabolov

Department of Computer Science
Higher School Of Economics

April, 2018

Outline

Outline

1. Problem
2. Seed, Expand and Constrain
   - Seeding loss with localization cues
   - Expansion loss with global weighted rank pooling
   - Constrain-to-boundary loss
3. Experiments and results
4. Effect of the different loss terms
5. Conclusion

Problem

- Input: image
- Output: segmentation mask
- Training set contains **weakly** annotated images

Outline

1. Problem
2. Seed, Expand and Constrain
   - Seeding loss with localization cues
   - Expansion loss with global weighted rank pooling
   - Constrain-to-boundary loss
3. Experiments and results
4. Effect of the different loss terms
5. Conclusion

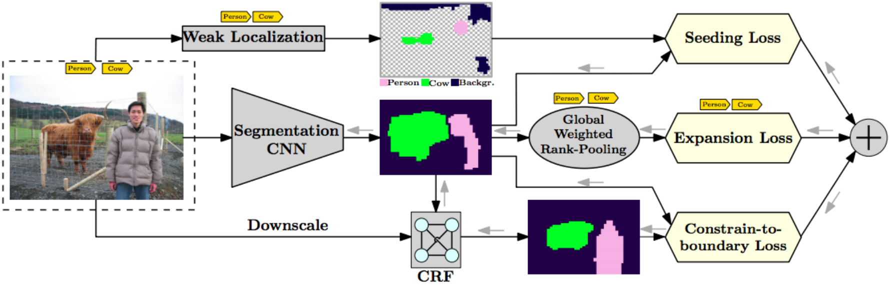
Seed, Expand and Constrain

Use a combination of 3 loss functions:

- Seeding loss to give clues of object locations
- Expansion loss to expand the object seeds to regions of reasonable size and ensure correct classification
- Constrain-to-boundary loss to alleviate the problem of imprecise boundaries

$$\min_{\theta} \sum_{(X,T)\in D} [L_{seed}(f(X;\theta), T) + L_{expand}(f(X;\theta), T) + L_{constrain}(X, f(X;\theta))]$$

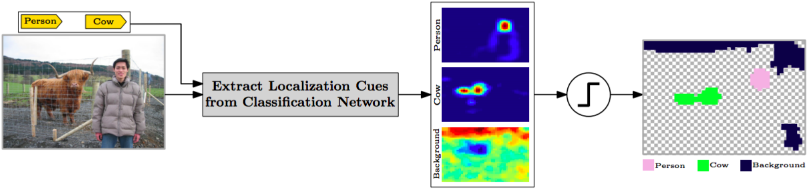
Seed, Expand and Constrain



Seeding loss with localization cues

We can use pretrained deep image classification network to retrieve localization cues!

Let's use them to guide our network



Seeding loss with localization cues

We get the following loss function:

$$-\frac{1}{\sum\limits_{c\in T}|S_c|}\sum_{c\in T}\sum_{u\in S_c}\log f_{u,c}(X)$$

S_c — locations labeled with class c with weak localization procedure

X — image

T — labels for X

$f_{u,c}(X)$ — "probability" that class c is present in location u in X.

Expansion loss with global weighted rank pooling

To ensure that our mask corresponds to image-level labels we can apply standard multi-label classification loss after aggregating classification scores.

Common aggregating methods are:

- Global Max Pooling (GMP)
- Global Average Pooling (GAP)

Expansion loss with global weighted rank pooling

To ensure that our mask corresponds to image-level labels we can apply standard multi-label classification loss after aggregating classification scores.

Common aggregating methods are:

- Global Max Pooling (GMP) BAD! Underestimates
- Global Average Pooling (GAP) BAD! Overestimates

Expansion loss with global weighted rank pooling

Introducing global weighted rank pooling!

$I_c = \{i_1, \dots, i_n\}$ — descending order of prediction scores for $c \in C$: $f_{i_1,c}(x) \geq f_{i_2,c}(x) \geq \dots \geq f_{i_n,c}(x)$ $0 < d_c \leq 1$ — decay parameter.

GWRP:

$$G_c(f(X); d_c) = \frac{1}{Z(d_c)} \sum_{j=1}^{n} (d_c)^{j-1} f_{i_j,c}(X)$$

where $Z(d_c) = \sum_{j=1}^{n} (d_c)^{j-1}$

Expansion loss with global weighted rank pooling

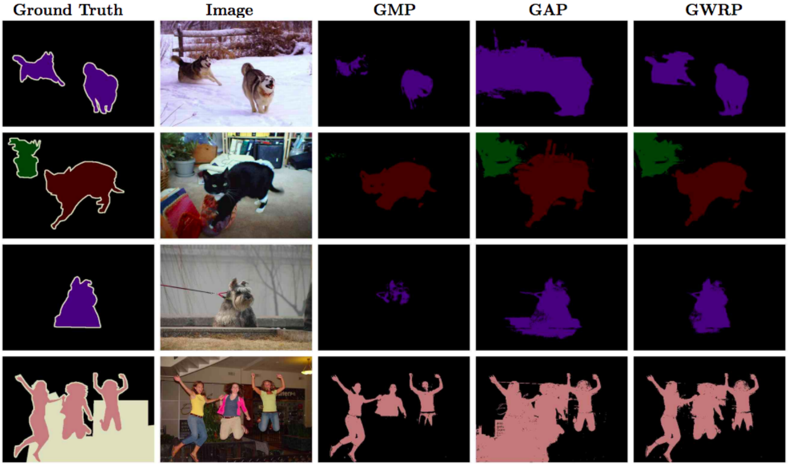
$$G_c(f(X); d_c) = \frac{1}{Z(d_c)} \sum_{j=1}^{n} (d_c)^{j-1} f_{i_j,c}(X)$$

$d_c = 0$: we get GMP

$d_c = 1$: we get GAP

Theoretically we can set d_c individually for each class.

Expansion loss with global weighted rank pooling



Expansion loss with global weighted rank pooling

Finally, the expansion loss:

$$L_{expand}(f(X), T) = -\frac{1}{|T|}\sum_{c \in T} \log G_c(f(X); d_+) -$$

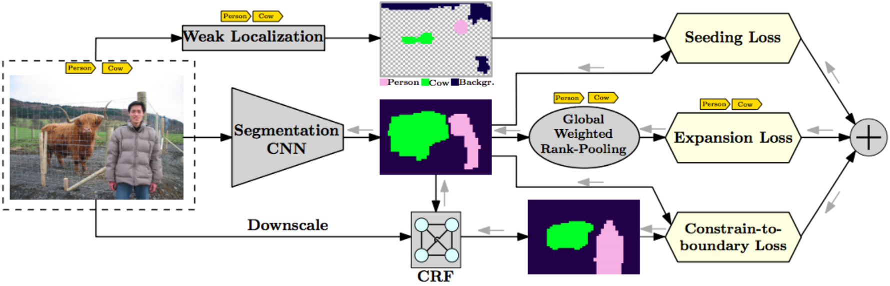
$$\frac{1}{|C' \setminus T|}\sum_{c \in C' \setminus T} \log(1 - G_c(f(X); d_-)) - \log G_{c_{bg}}(f(X), d_{bg})$$

Constrain-to-boundary loss

We also want to penalize NN for producing segmentations that are discontinuous with respect to spatial and color information in the input image Let's construct CRF $Q(X, f(X))$ and define our loss as the mean

KL-divergence between the outputs of the network and the outputs of the CRF:

$$L_{constrain}(X, f(X)) = \frac{1}{n} \sum_{u=1}^{n} \sum_{c \in C} Q_{u,c}(X, f(X)) \log \frac{Q_{u,c}(X, f(X))}{f_{u,c}(X)}$$


Person
Cow
Weak Localization
Person
Cow
Person Cow Backgr.
Segmentation CNN
Person
Cow
Global Weighted Rank-Pooling
Person
Cow
Seeding Loss
Expansion Loss
Constrain-to-boundary Loss
Downscale
CRF

Outline

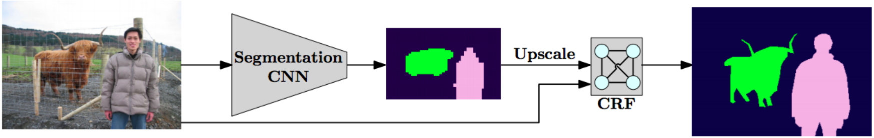
Experiments and results

- Segmentation network: DeepLab-CRF-LargeFOV (Slightly modified VGG)
- Localization network: modified VGG
- Optimization: SGD
- Decay parameters:
  - $d_{-} = 0$
  - $d_{+} = 0.996$
  - $d_{bg} = 0.999$

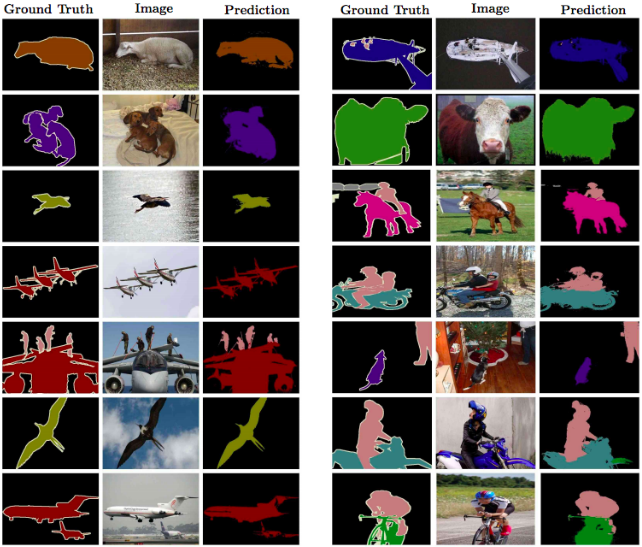
Experiments and results

Our segmentation network produces downscaled images!

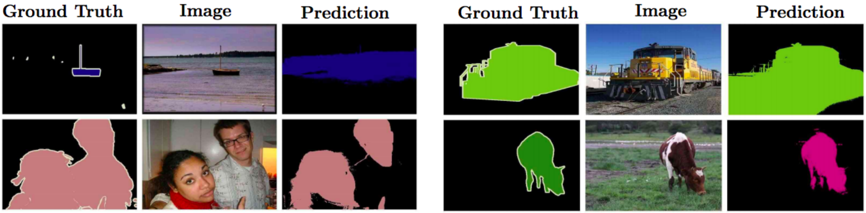
At test time upscale and then apply CRF



Experiments and results



Experiments and results



Experiments and results

PASCAL VOC 2012 *val* set	[4] (Img+Obj)	[14] (stage1)	EM-Adapt (re-impl. of [24])	CCNN [24]	MIL+ILP +SP-sppxl† [26]	SEC (proposed)
background		71.7*	67.2	68.5	77.2	**82.4**
aeroplane		30.7*	29.2	25.5	37.3	**62.9**
bike		**30.5***	17.6	18.0	18.4	26.4
bird		26.3*	28.6	25.4	25.4	**61.6**
boat		20.0*	22.2	20.2	**28.2**	27.6
bottle		24.2*	29.6	36.3	31.9	**38.1**
bus		39.2*	47.0	46.8	41.6	**66.6**
car		33.7*	44.0	47.1	48.1	**62.7**
cat		50.2*	44.2	48.0	50.7	**75.2**
chair		17.1*	14.6	15.8	12.7	**22.1**
cow		29.7*	35.1	37.9	45.7	**53.5**
diningtable		22.5*	24.9	21.0	14.6	**28.3**
dog		41.3*	41.0	44.5	50.9	**65.8**
horse		35.7*	34.8	34.5	44.1	**57.8**
motorbike		43.0*	41.6	46.2	39.2	**62.3**
person		36.0*	32.1	40.7	37.9	**52.5**
plant		29.0*	24.8	30.4	28.3	**32.5**
sheep		34.9*	37.4	36.3	44.0	**62.6**
sofa		23.1*	24.0	22.2	19.6	**32.1**
train		33.2*	38.1	38.8	37.6	**45.4**
tv/monitor		33.2*	31.6	36.9	35.0	**45.3**
average	32.2	33.6*	33.8	35.3	36.6	**50.7**

PASCAL VOC 2012 *test* set	MIL-FCN [25]	CCNN [24]	MIL+ILP +SP-sppxl† [26]	Region score pooling [18]	SEC (proposed)
background		$\approx 71^\ddagger$	74.7	$\approx 74^\ddagger$	**83.5**
aeroplane		24.2	38.8	33.1	**56.4**
bike		19.9	19.8	21.7	**28.5**
bird		26.3	27.5	27.7	**64.1**
boat		18.6	21.7	17.7	**23.6**
bottle		38.1	32.8	38.4	**46.5**
bus		51.7	40.0	55.8	**70.6**
car		42.9	50.1	38.3	**58.5**
cat		48.2	47.1	57.9	**71.3**
chair		15.6	7.2	13.6	**23.2**
cow		37.2	44.8	37.4	**54.0**
diningtable		18.3	15.8	**29.2**	28.0
dog		43.0	49.4	43.9	**68.1**
horse		38.2	47.3	39.1	**62.1**
motorbike		52.2	36.6	52.4	**70.0**
person		40.0	36.4	44.4	**55.0**
plant		33.8	24.3	30.2	**38.4**
sheep		36.0	44.5	48.7	**58.0**
sofa		21.6	21.0	26.4	**39.9**
train		33.4	31.5	31.8	**38.4**
tv/monitor		38.3	41.3	36.3	**48.3**
average	25.7	35.6	35.8	38.0	**51.7**

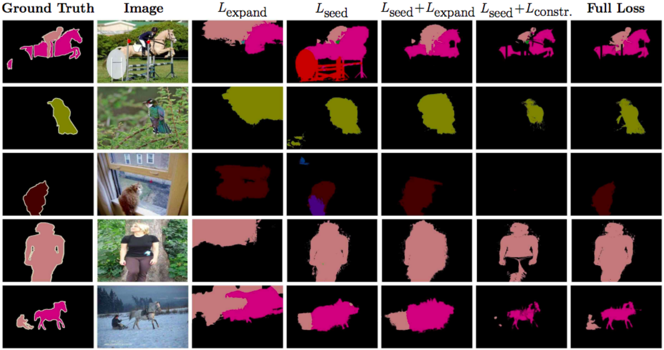
Experiments and results

method	*val*	*test*	comments
DeepLab [6]	67.6	70.3	fully supervised training
STC [42]	49.8*	51.2*	trained on Flickr
TransferNet [12]	52.1	51.2	trained on MS COCO; additional supervision: from segmentation mask of other classes
[4] (1Point)	42.7	–	additional supervision: 1 click per class
[4] (AllPoints-weighted)	43.4	–	additional supervision: 1 click per instance
[4] (squiggle)	49.1	–	additional supervision: 1 squiggle per class
EM-Adapt [23]	38.2	39.6	uses weak labels of multiple image crops
SN_B [41]	41.9	43.2	uses MCG region proposals (see text)
MIP+ILP+SP-seg [26]	42.0	40.6	trained on ImageNet, MCG proposals (see text)
MIL+ILP+SP-bb [26]	37.8	37.0	trained on ImageNet, BING proposals (see text)

(*results from manuscripts that are currently unpublished/not peer-reviewed)

Outline

Effect of the different loss terms



loss function	mIoU (*val*)
L_{expand}	20.6
L_{seed}	45.4
L_{seed} + L_{expand}	44.3
L_{seed} + $L_{constrain}$	50.4
all terms	50.7

Outline

Effect of the different loss terms

- Cool loss function for training segmentation networks when only image-level labels are available
- Outperforms previous state-of-the-art methods by a large margin
- Knowledge about object sizes can improve the segmentation performance