LID Challenge: Weakly Supervised Semantic Segmentation

3d place solution

NoPeopleAllowed: The 3 step approach to weakly supervised semantic segmentation

Mariia Dobko, Ostap Viniavskyi, Oles Dobosevych

UCU & SoftServe team

The Machine Learning Lab at Ukrainian Catholic University, SoftServe

Outline

- Problem description
- Competition
- Approach architecture
  - Step 1. CAM generation via classification
  - Step 2. IRNet for CAM improvements
  - Step 3. Segmentation
- Postprocessing
- Results
- Conclusions

Problem description

Image-level annotations

A key bottleneck in building a DCNN-based segmentation models is that they typically require **pixel level annotated images** during training. Acquiring such data demands an **expensive**, and **time-consuming** effort.
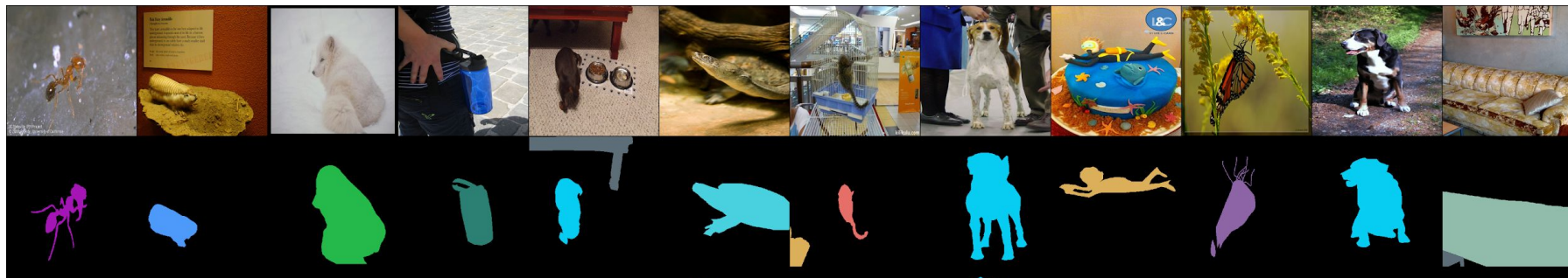
15 times faster to label

> 25 times cheaper
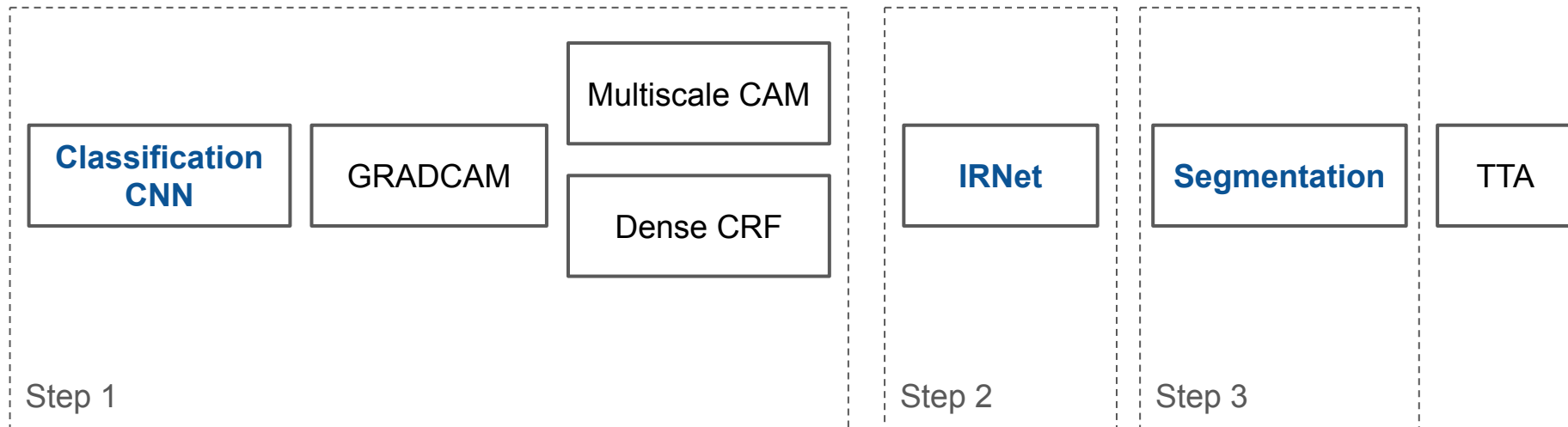0.035$ per image for class,
3.45$ for segmentation

We develop a method that has a **high performance** in segmentation task while also **saves time and expenses** by using only **image-level annotations**.
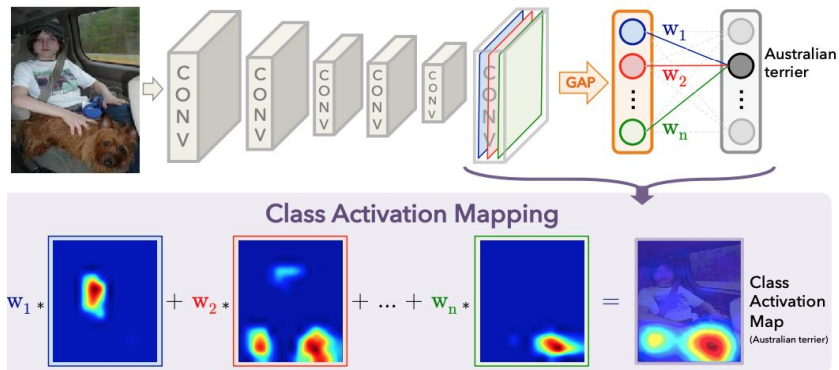
LID Challenge Dataset

- **Multilabel multiclass**
- Pixel-wise labels are provided for validation set only
- **No pixel-wise annotations** are allowed for training

- **200** classes + background
- **456,567** training images
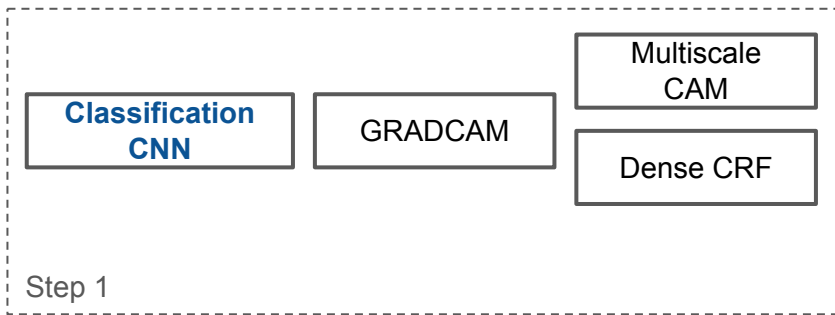 - validation: **4,690**
  - test: **10,000**

Challenges

- **High imbalance** in classes: 'person', 'bird', 'dog'

- **Missing labels**

- Subset of 2014 has better labels for 'person', than the whole dataset

Approach architecture

Step 1. CAM generation via classification

Class Activation Mapping

$w_1 *$ + $w_2 *$ + ... + $w_n *$ = Class Activation Map (Australian terrier)

Input

- 72k - train, 12k validation
- balanced dataset
- no person class

Results

Classification CNN → GRADCAM → Multiscale CAM / Dense CRF
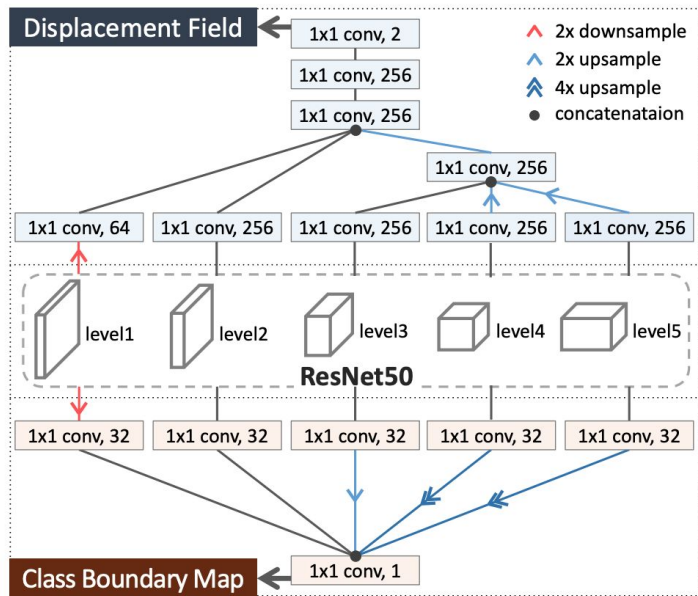
Step 1

Step 2. IRNet for CAM improvements

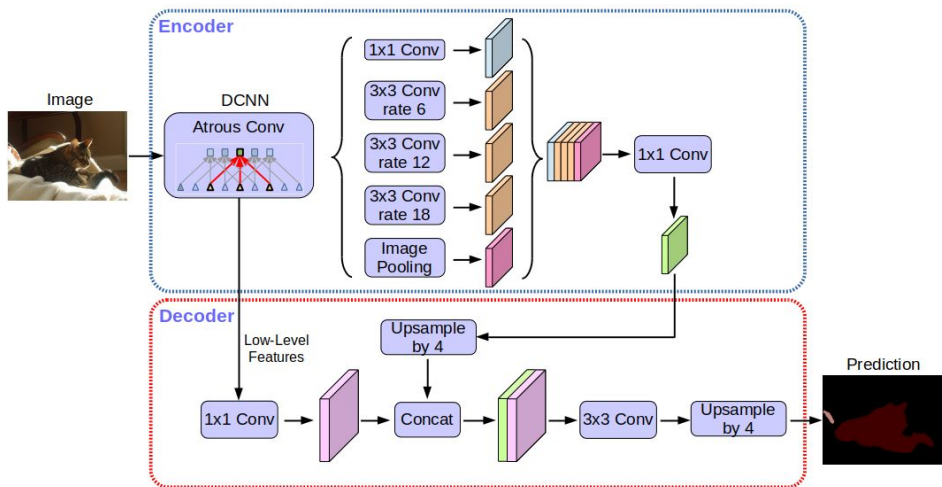Figure 2. Overall architecture of IRNet.

Input

- Select most confident maps
- Threshold CAMs into confident BG, confident FG and unconfident regions

Results

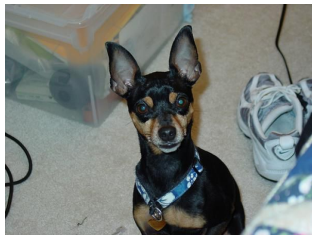# Step 3 - Segmentation

DeepLab v3+

Input

- 352x352 input images
- Strong augmentations
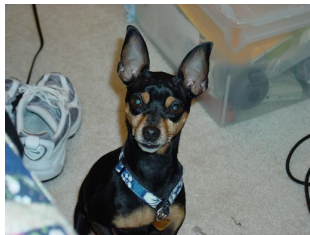- ~42k images for training

Results
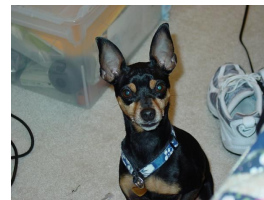
Postprocessing

Image

Horizontal flip

scale=0.5

scale=1

scale=2

TTA

Test Time Augmentations are added after segmentation step. The combination of 2 types of different TTAs, with one having 3 parameters, result in total 6 predictions, which are averaged by mean.

Secret insights

- **VGG** is better for CAM generation as **ResNet** gives artifacts

- **Decrease the output stride** of VGG by removing some of the max pooling operations

- **Confident** and **unconfident** regions for IRNet

- **Multiscale CAM** give a large improvement

- **Dense CRF** doesn't require training, helps to rectify boundaries

- **TTA** after segmentation step drastically improves the results

- Replace stride with dilation in DeepLabv3+ to **decrease the output stride**

Quantitative Results

Model	IRNet threshold	TTA	Person CAM	Mean IoU
DeepLabv3+ encoder: **ResNet50**	**0.3**	No	No	36.65
		Yes		**39.64**
			Yes	**39.80***
	0.5	No	No	37.11
		Yes		39.58
DeepLabv3+ encoder: ResNet101	0.5	No		36.14
		Yes		37.15

Validation set

Experiments with different architectures and parameters on the 3rd step

* wasn't submitted

Quantitative Results

Test set:

DeepLabv3+
+
TTA
(Horizontal Flip,
Multi-scaling)

Rank	Participant team	Mean IoU	Mean accuracy	Pixel accuracy	Last submission at
1	cvl	45.18	59.62	80.46	1 day ago
2	VL-task1	37.73	60.15	82.98	2 days ago
3	UCU & SoftServe	37.34	54.87	83.64	2 days ago
4	IOnlyHaveSevenDays	36.24	68.27	84.10	2 days ago
5	play-njupt	31.90	46.07	82.63	1 month ago
6	xingxiao	29.48	48.66	80.82	1 month ago
7	hagenbreaker	22.50	39.92	77.38	19 days ago
8	go-g0	19.80	38.30	76.21	20 days ago
9	lasthours-try	12.56	24.65	64.35	1 day ago
10	WH-ljs	7.79	16.59	62.52	2 days ago

Open questions

Downsampling was used to balance data $\rightarrow$ **Upsampling** or **combination** of both should be tested

Adding person class labels to the other steps of pipeline $\rightarrow$
   Ability to provide better results for a class which is highly present in data, though severely mislabeled

Mean IoU per class allows to obtain high score even when some classes are skipped $\rightarrow$
            A different metric or combination of metrics should be chosen as a premier for this task

presentation

Thank you for attention!

dobko_m@ucu.edu.ua

viniavskyi@ucu.edu.ua

dobosevych@ucu.edu.ua