greatlearning
Learning for Life

Instance Segmentation

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

greatlearning
Learning for Life

Agenda

- What we have done so far?
- What is Instance Segmentation?
- Model Architecture
- Joint Learning
- Full Scene Inference
- Results

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

What we have done so far?

- **Image Classification**: Classify the main object category within an image.
- **Object Detection**: Identify the object category and locate the position using a bounding box for every known object within an image.
- **Semantic Segmentation**: Identify the object category of each pixel for every known object within an image. **Labels are class-aware.**

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

What is Instance Segmentation?

Instance Segmentation: Identify each object instance of each pixel for every known object within an image. **Labels are instance-aware.**

Instance Segmentation is one level increase in difficulty!!!

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Semantic Seg vs Instance Seg

Input Image

Semantic Segmentation

Instance Segmentation

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Deep Mask

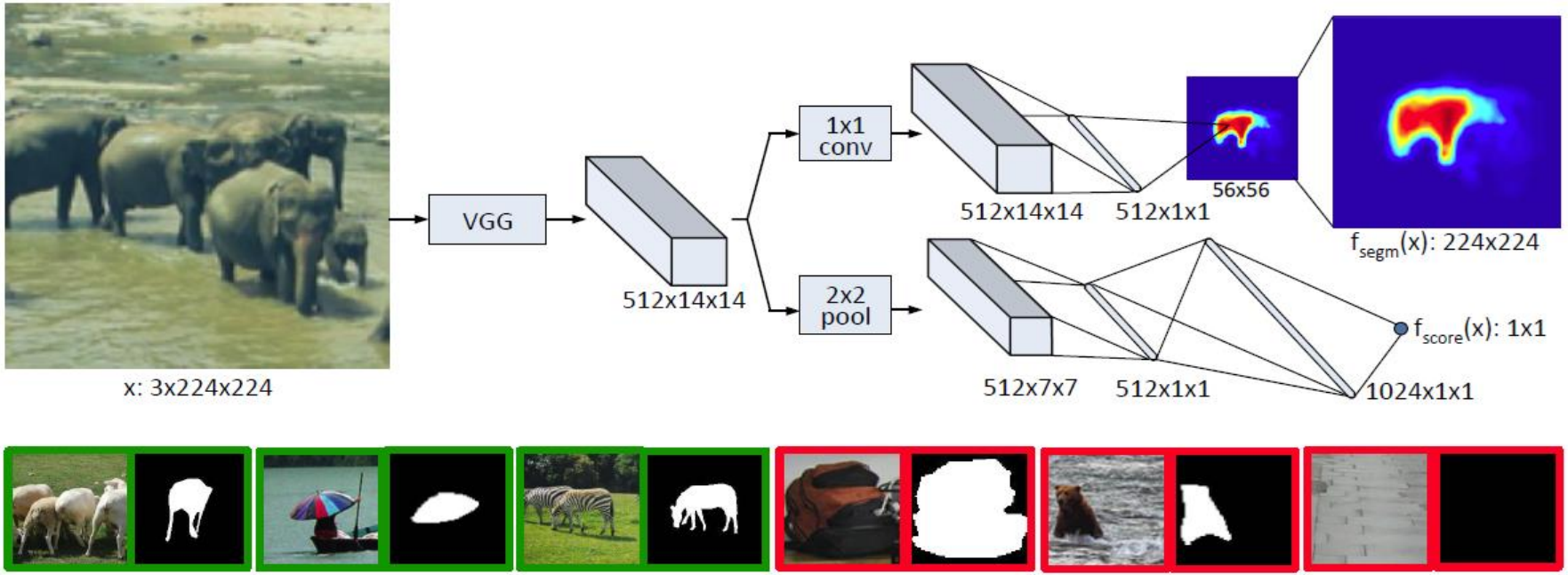
DeepMask is the 2015 NIPS paper with more than 300 citations.

Though it is a paper published in the year of 2015, it is one of the earliest paper using CNN for instance segmentation.

It is worth to know the development of deep-learning-based instance segmentation.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Architecture



Model Architecture (Top), Positive Samples (Green, Left Bottom), Negative Samples (Red, Right Bottom)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Left Bottom: Positive Samples

A label ***yk*=1** is given for *k*-th positive sample. To be a positive sample, two criteria need to be satisfied:

- **The patch contains an object roughly centered in the input patch**.
- **The object is fully contained in the patch and in a given scale range.**

When *yk*=1, the **ground truth mask *mk* has positive values** for the pixels which belong to the **single object** located in the centre of the image patch.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Right Bottom: Negative Samples

Otherwise, a label ***yk*=-1** is given for a negative sample even the object is partially present.

When *yk*=-1, the mask is not used.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Top, Model Architecture: Main Branch

The model as shown above, given the input image patch x, after feature extraction by VGGNet, The fully connected (FC) layers originated in VGGNet are removed.

The last max pooling layer in VGGNet is also removed, thus the output before splitting into two paths are of the size of 1/16 of input.

For example as above, the input is 224×224 (3 is the number of channels in the input image, i.e. RGB), the output at the end of main branch is (224/16)×(224/16) =14×14. (512 is the number of feature maps after convolution.)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

There are two paths after VGGNet:

- The first path is to predict the class-agnostic segmentation mask, i.e.fsegm(x).
- The second path is to assign a score corresponding to how likely the patch is to contain an object, i.e. fscore(x).

Top, First Path: Predicting Segmentation Map

Top, Second Path: Predicting Object Score

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Joint Learning

The loss function is a sum of binary logistic regression losses, one for each location of the segmentation network fsegm(xk) and one for the object score fscore(xk).

$$\mathcal{L}(\theta) = \sum_k \left(\frac{1+y_k}{2w^o h^o} \sum_{ij} \log(1 + e^{-m_k^{ij} f_{segm}^{ij}(x_k)}) + \lambda \log(1 + e^{-y_k f_{score}(x_k)}) \right)$$

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Some more details

- Batch size of 32 is used.
- Pretrained ImageNet model is used.
- There are 75M parameters in total.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Full Scene Inference - Brief

- Multiple max pooling is done on the feature map.
- A pixel shift is performed before each max pooling.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Full Scene Inference - Details

Multiple Locations and Scales

During inference (testing), the model is applied densely at multiple locations with a stride of 16 pixels, and multiple scales from 1/4 to 2 with a step size of square root of 2. This ensures that there is at least one tested image patch that fully contains each object in the image.

Fine Stride Max Pooling

Since the input test image is larger than the training input patch size, we need a corresponding 2D scoring map as an output rather than one single scoring value. An interleaving trick is used before the last max pooling layer for the scoring branch, i.e. the Fine Stride Max Pooling proposed in OverFeat.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Results

MS COCO (Boxes & Segmentation Masks)

80,000 images and a total of nearly 500,000 segmented objects, are used for training. And the first 5000 images of the MS COCO 2014 are used for validated.

PASCAL VOC 2007 (Boxes)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Inference Time

- The inference time in MS COCO is 1.6s per image.
- The inference time in PASCAL VOC 2007 is 1.2s per image.
- Inference time can be further dropped by about 30% by parallelizing all scales in a single batch.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Qualitative Results

DeepMask proposals with highest IoU to the ground truth on selected images from COCO. Missed objects (no matching proposals with IoU > 0.5) are marked with a red outline.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

References

- Paper - [2015 NIPS] [DeepMask] Learning to Segment Object Candidates
- https://towardsdatascience.com/review-deepmask-instance-segmentation-30327a072339

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.