Deep Learning and Segmentation

Lecture by Margrit Betke
CS 585, March 26, 2024
Image Segmentation -- Definition and Tasks

Definition 1:
Segmentation = finding outline of object ("thing") or region ("stuff") in image

Definition 2:
Segmentation = grouping of pixels into regions such that:
• Pixels in each region have a common property
• Pixels in adjacent regions do not share this property
• Exclusive Partitioning: $P_i \cap P_j = \emptyset$, for all $i \neq j$
• Exhaustive Partitioning: Union of $P_i$'s = entire image

Tasks:

**Semantic Segmentation:** Common property: Same “stuff class”

**Instance Segmentation:** Common property: Same “thing class”

**Panoptic Segmentation:** Common property: Either same thing or stuff class
“Semantic” Segmentation = Segmentation

Here: Exclusive & exhaustive partitioning involving 3 object classes:

- All regions with pixels that collectively show bikes are labeled green.
- All regions with pixels of bikers are shown in antique pink.
- All regions background pixels are black.

Model: FCN-8s

Ground truth

Image Credit: Long et al., 2015
“Semantic” Segmentation = Region Segmentation

Model: FCN-8s

Ground truth

Here: Exclusive & exhaustive partitioning involving 3 object classes:

• All regions with pixels that collectively show bikes are labeled green.
• All regions with pixels of bikers are shown in antique pink.
• All regions background pixels are black.

CS 585: Image and Video Computing

Image Credit: Long et al., 2015
SegNet: Encoder-Decoder Architecture for Semantic Segmentation

Badrinayanan et al., 2016
Instance Segmentation = Segmentation of Individual Objects

Phase-contrast microscopy image  Ground truth segmentation  Model segmentation

Image Credit: Yi et al., MIA, 2019
Panoptic Segmentation = Segmentation of regions and objects

Term coined by Kirillov et al., 2018
How can we measure the success of a segmentation model?
Intersection over Union (IoU) or Jaccard Index

Given an object region \( A \), drawn by an expert, and an object region \( B \), determined by the computer, the Jaccard index computes the ratio of the number of pixels common to \( A \) and \( B \) over the number of pixels that are in at least one of the regions: \( |A \cap B| / |A \cup B| \).

Resulting scores range from 0 to 1 with larger values indicating greater similarity between the two regions.
Using a Threshold on the IoU for Classification

Ground truth bounding box: red. Model bounding box: light blue

Image credit: Learnopencv.com
Beware of Annotation Noise

Image credit: Theriault et al., MV, 2012
Beware of Annotation Noise

Image credit: Theriault et al., MV, 2012
ICORD: Intelligent Collection of Redundant Data

Figure 5. An example processed by ICORD involving a cell on a fluorescence microscopy image. ICORD detects in the second round that the outline is sufficiently accurate to be considered a final product ($\tau = 0.83$).

**ICORD Process for Cell Segmentation:**

*Input:* Raw images of cells, quality threshold $\tau$, number of rounds $N$.

1. A single round of crowdsourcing is performed on all cell images. One segmentation is obtained per cell.
2. Crowd segmentations are converted to binary masks, and image and behavioral features are extracted.
3. The prediction model receives the feature vectors and evaluates the quality of each segmentation.
4. For each cell: If the predicted score is higher than threshold $\tau$, the system accepts the annotation (step 7). Otherwise, the annotation is flagged as inaccurate (step 6).
5. Repeat until all cell segmentations are predicted to be accurate or $N$ crowdsourcing rounds have been performed:
   - 6.1 A new round of crowdsourcing is performed on the cell images with annotations flagged as inaccurate.
   - 6.2 Steps 2.-4. are applied to the current segmentation.
6. For any cells still predicted to have inaccurate segmentations, the segmentation among the N collected is chosen that has highest predicted quality.

*Output:* Cell annotations and their predicted quality scores.

Sameki et al., CVPRW 2016
Figure 4. An example processed by ICORD: A phase contrast image of a cell and its segmentations, produced by crowd workers in 5 rounds. In rounds 1–4, the prediction model flagged the segmentations as not sufficiently accurate (quality score below threshold $\tau = 0.75$). In round 5, ICORD predicts that the shown segmentation is accurate (score $> 0.75$) and terminates the processing on this cell. For each round, the Jaccard scores measuring the overlap between expert-drawn and crowd-worker-drawn regions are also displayed (observed and predicted scores only differ by 6 or fewer percentage points).
1. road · sidewalk · parking · rail track
2. person · rider
3. car · truck · bus · on-rails · motorcycle · bicycle · caravan · trailer
4. building · wall · fence · guard rail · bridge · tunnel
5. pole · pole group · traffic sign · traffic light
6. vegetation · terrain
7. sky
8. ground · dynamic · static
UPSNet: Panoptic Segmentation

Runtime speedup 3x over previous work  

Xiong et al., 2019
UPSNet: Panoptic Segmentation

Ground truth | Xiong et al.’s evaluation of Kirillov et al.’s model | Xiong et al., 2019
Mask R-CNN

Extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition.

He et al., ICCV 2017
Backbone Detection Networks used for Segmentation

Faster R-CNN uses a Region Proposal Network (RPN) that shares convolutional features with the Fast R-CNN: Ren et al., NIPS 2015

Fast R-CNN: Girschik, 2015

R-CNN (for “Regions with CNN Features”): Girschik et al., 2014
Domain Adaptive Semantic Segmentation

Wang et al., ICCV 2023

Deep models often generalize poorly to new domains such as different cities or weather in driving scenes. Solution: Domain Transfer

Unsupervised domain adaptation (UDA) allows knowledge transfer from synthetic data (source domain), where pixel-level annotations are more cheaply available, to real-world data (unlabeled target domain).

Extends DAFormer, a Transformer-based model for UDA

Our contribution: A cross-domain attention consistency loss function.
Wang et al., ICCV 2023’s Results

Target Image
Cityscape

Source only
Synthetic

DAFormer

Ours

Ground Truth